Topics in Statistical Analysis and Interpretation of Geophysical Data Sets

Sergey Kravtsov[1]

Atmo 500 Lecture Notes, Fall 2013

[1] These notes rely heavily, in presentation of selected topics, on the work of Von Mises (1964), Press et al. (1994), and Prof. D. Hartmann's lecture notes. Some parts of the present notes are in fact direct duplication, or minor rewording of the above cited texts. These notes should not, therefore, be considered as the original presentation; rather, the material from different textbooks has been compiled here by the author in a specific order, as well as augmented by the author's comments. In principle, no other text, beside these notes, is required for the students to fully understand the material and succeed in this class.

Contents

Chapter 1

Introductory Notes

1.1 Preliminary Remarks

1.1.1 Purpose of the course

The purpose of this course is to introduce basic statistical concepts and develop a working knowledge of a number of statistical methods currently used for analysis, interpretation and modeling of weather/climate-related (observed and model-generated) data sets. The presentation will be based on a balance between mathematical rigor in derivation of various statistical techniques and the necessity to cover a fairly large (although by no means complete) set of analysis methods. A particular attention will be paid to the question of how to choose and apply (an) appropriate statistical method(s) depending on the nature of the phenomenon under consideration. Each topic covered in the lecture presentations will be complemented by exercises using synthetic and real data sets in practical classes.

1.1.2 Outline of the course

In the remainder of Chapter 1, we will introduce the object of our analysis and define some important dynamical (e.g., signal and noise), as well as statistical (e.g., probability distributions) concepts. The statistical techniques we will review further are designed to:

- recognize dominant and possibly *predictable* patterns of natural and forced climate variability (*signals*) in the background of unpredictable *noise* (see Section 1.2.3) and infer physical mechanisms associated with these signals
- validate numerical climate models by comparing characteristics of observed and simulated data sets
- use the signals identified by *descriptive statistics* methods (first two items above) for the purpose of constructing a predictive statistical model to forecast the climate subsystem's behavior in the future (*inferential statistics*)

Climate variability has an inherently nonlinear character. Numerous instabilities and interactions within the climate system impose natural limits on predictability; in particular, a common property of climate models (and climate) is *sensitive dependence* on initial conditions: small perturbations of the latter result, in the long run, to large changes in the subsequent climate evolution. Typical time scales on which such discrepancies happen are related to how fast a certain climate subsystem "forgets" initial conditions. In general, we will concentrate on detection of large-scale, low-frequency climate signals in the presence of smaller-scale, higher-frequency noise. Our purpose will thus be to identify the features of climate evolution that are not entirely unpredictable. In general, we will look for certain *statistically significant* relationships between two or more climatic variables (for example, time correlation). The degree of statistical significance can be evaluated by assuming that the variables are not related and working out the chances of observing the value of, say, correlation, like the one we have obtained from our data sample. If the chance is not large, we can reject our null hypothesis of no relationship between the variables and can even hope that there is indeed some relationship. The problem of hypothesis testing is dealt with in Chapter 2. Various tests of statistical significance described in this chapter are used throughout the remainder of the course.

Chapter 3 introduces linear statistical models or regression models, which are useful for exploring the relationships between climatic variables and can be used for climate prediction, as well as for certain types of statistical significance testing. The data analysis techniques described in this chapter are also a necessary introduction to the matrix methods discussed in Chapter 4.

One of the applications of matrix methods is data compression, by which the high-dimensional data set is replaced by a small number of modes (spatial patterns and the associated time series). Chapters 5 and 6 will deal with the analysis of data sets so reduced using two complementary statistical descriptions of low-frequency climatic variability, namely *episodic* description, in which one looks for recurrent and/or persistent states and transitions between them (Chapter 5) and *oscillatory* description in frequency domain designed to identify periodicities in the time series under consideration (Chapter 6). The latter chapter will also discuss some aspects of time series filtering. Finally, Chapter 7 will conclude the course with a summary and outlook.

1.1.3 Textbooks and online sources of material

These notes rely heavily, in presentation of selected topics, on the work of Von Mises (1964), Press et al. (1994), and Prof. D. Hartmann's lecture notes (see the reference below). Some parts of the present notes are in fact direct duplication, or minor rewording of the above cited texts. These notes should not, therefore, be considered as the original presentation; rather, the material from different textbooks has been compiled here by the author in a specific order, as well as augmented by the author's comments. In principle, no other text, beside these notes, is required for the students to fully understand the material and succeed in this class. $\boxed{!}$

The classical text in probability theory and statistics is Von Mises (1964). Wilks (1995) and von Storch and Zwiers (1999) discuss in depth applications of various statistical methods to problems in atmospheric and climate science. A good summary of basic statistics, linear matrix operations, spectral analysis and regression techniques can also be found in *Numerical Recipes* (Press et al. 1994). There also exists a number of online statistical texts (lecture notes, online courses, statistical manuals). Here are a few examples:

- http://www.statsoft.com/textbook: Online statistics textbook
- http://www.atmos.washington.edu/∼dennis/: Go to ATMS 552 and click on "Class Notes" to get to Lecture notes of Prof. Dennis Hartmann

Finally, MATLAB's statistics toolbox contains a brief description and illustration of its intrinsic functions and analysis methods.

1.1.4 Statistical software

Practical exercises will be done mostly using MATLAB. No prior MATLAB experience is, however, necessary. When working with very large data sets, MATLAB's memory limitations can become a problem. In such cases, some combination of MATLAB and FORTRAN can be of help. Press et al. (1994) provide the reader with many useful source codes for performing various types of statistical analysis and explanations thereof; the codes themselves are available online (both in FORTRAN and C). Finally, advanced spectral analysis methods will be illustrated using MTM–SSA toolkit available from “http://www.atmos.ucla.edu/tcd/ssa/.”

1.2 General properties of climatic data sets

1.2.1 Representation of data in the form of two-dimensional matrices

Our analysis will deal, in most cases, with *long multivariate time series* of climatic fields; for example, the data set can consist of daily values of a single *variable*, say geopotential height, on a regular grid in space, produced from an integration of a numerical climate model, or, alternatively, of irregularly spaced station values of observed geopotential height. Space can be either one- (e.g., several latitudinal locations), two- (e.g., longitude–latitude grid), or three-dimensional (longitude, latitude, height/pressure), but in each case we will usually string variables to form a big one-dimensional vector. For example, if we have observations at I longitudes, J latitudes and K height/pressure levels, the resulting data vector $\mathbf{x} \equiv \{x_m\}\Big|_{m=1}^{M}$ will have a dimension $M = I \times J \times K$, where M is the total number of different locations. Suppose we have N observations of this data vector. This data set can thus be represented as a two-dimensional matrix, which we will call the *input data matrix* $\mathbf{X} \equiv \{x_{n,m}\}$, where $1 \leq n \leq N$ and $1 \leq m \leq M$:

$$\mathbf{X} \equiv \begin{pmatrix} x_{1\,1} & x_{1\,2} & \cdots & x_{1\,M} \\ x_{2\,1} & x_{2\,2} & \cdots & x_{2\,M} \\ \multicolumn{4}{c}{\dots\dots\dots\dots\dots\dots} \\ x_{N\,1} & x_{N\,2} & \cdots & x_{N\,M} \end{pmatrix} . \tag{1.1}$$

! Depending on application, one can design the input data matrix in several ways:

- A *space–time* array, as in the example above, consists of values of a single variable at M locations taken at N different times
- A *parameter–time* array is represented by values of M different variables (geopotential height, sea-surface temperature, etc.) measured at a single location at N different times
- A *parameter–space* array will be composed of values of M variables taken at N different locations at a single time

In this course, we will restrict ourselves with data sets written in space–time and parameter–time form, so that the first dimension will always be time dimension. One can also construct, in this case, an extended input data matrix of two or more space–time or parameter–time arrays by column augmentation.

Example 1.1 *Suppose that we want to study the effect of tropical climate variability onto the low-frequency component of mid-latitude atmospheric flow using 50 years of reanalyzed observations. A possible way to set up the input data matrix for such an exercise could be as follows. We first take a set of* $N = 365 \times 50/10 = 1825$ *consecutive ten-day averages of (i) 700-mb geopotential height (*Z_{700}*) anomalies on a regular* $5^\circ \times 5^\circ$ *grid in the* $30^\circ N$– $60^\circ N$ *belt (*$M_1 = 36 \times 6 = 216$ *data points) and (ii) sea-surface temperature (SST) anomalies on a regular* $5^\circ \times 5^\circ$ *grid in the* $10^\circ S$– $10^\circ N$ *belt (*$M_2 \sim \mathcal{O}(70) < 36 \times 4 = 144$ *data points, since some of the points are over land). "Anomaly" means that we have removed, at each spatial point and for each variable, this variable's time averaged value; each anomaly field has thus a zero time mean; one also says in this case that each time series has been* centered. *Since we are considering the relationship between two fields that have different units (meters and degrees), we also have to form dimensionless fields (* nondimensionalize *time series); for example, we can divide each value in a given time series by some quantity that measures the amplitude of this field's variability (see chapter 4 for further discussion). We might also want to remove* seasonal cycle *from our time series, since we would like to study the intrinsic dynamics of relationship between climatic signals in middle latitudes and tropics, rather than detect correlations caused by external forcing. The two resulting data matrices* $\mathbf{X}^{(1)}$ *(*Z_{700}*) and* $\mathbf{X}^{(2)}$ *(SST) have thus dimensions* $N \times M_1$ *and* $N \times M_2$*, respectively. One can now form a new, single data matrix* $\mathbf{X} = \mathbf{X}^{(1)} \sqcup \mathbf{X}^{(2)}$ *of dimension* $N \times M$*, where* $M = M_1 + M_2 \sim \mathcal{O}(300)$*,* $N = 1825$*, and*

$$\mathbf{X} \equiv \left(\begin{array}{cccccccc} x^{(1)}_{1\,1} & x^{(1)}_{1\,2} & \cdots & x^{(1)}_{1\,M_1} & x^{(2)}_{1\,1} & x^{(2)}_{1\,2} & \cdots & x^{(2)}_{1\,M_2} \\ x^{(1)}_{2\,1} & x^{(1)}_{2\,2} & \cdots & x^{(1)}_{2\,M_1} & x^{(2)}_{2\,1} & x^{(2)}_{2\,2} & \cdots & x^{(2)}_{2\,M_2} \\ \multicolumn{8}{c}{\dotfill} \\ x^{(1)}_{N\,1} & x^{(1)}_{N\,2} & \cdots & x^{(1)}_{N\,M_1} & x^{(2)}_{N\,1} & x^{(2)}_{N\,2} & \cdots & x^{(2)}_{N\,M_2} \end{array} \right). \tag{1.2}$$

1.2.2 Structure versus sampling

In climate science, one is usually interested in spatial patterns of a given field or relationships between two or more different fields; realizations, or *sampling* of such space- and/or parameter-domain *structures* occur in time domain. One task of the statistical analysis is to identify robust structures (e.g., recurrent or persistent spatial patterns) that are statistically significant (reproducible). Statistical significance means that if we were, for example, to conduct our analysis using two independent data samples (say, using first half and second half of available time series), we would get the same structures. In order to obtain such robust solutions, the number of *degrees of freedom* in the domain of sampling must be much larger than that in the domain of structure. One possible measure of whether the analysis is likely to give statistically significant results is the aspect ratio M/N of the input data matrix; one would want the aspect ratio to be as small as possible. In *Example* 1.1, $M/N \approx 1/6$.

In reality, however, the aspect ratio is not a very good indicator of robustness, since

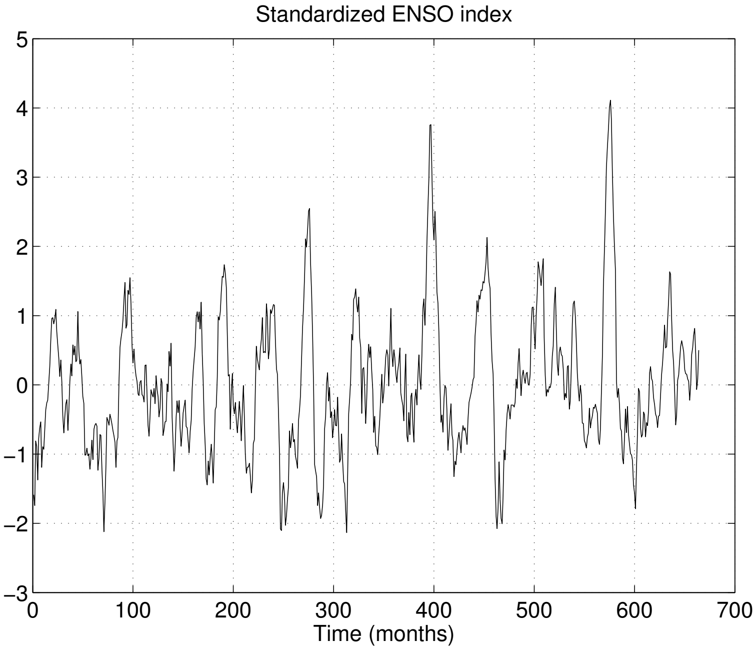


Figure 1.1: Standardized Niño-3 index time series. Seasonal cycle removed.

geophysical data sets are typically characterized by strong *autocorrelation* in both space and time domain. Returning to *Example* 1.1, it turns out that one can predict the evolution of the tropical SST (see Fig. 1.1) up to a couple of months into the future knowing the past two months' values, so the *effective* number of temporal degrees of freedom for a 50-yr-long SST record is approximately 300. This would indicate that the aspect ratio will be close to one, and not equal to 1/6 as we have inferred before. On the other hand, one can also show that only a few patterns dominate low-frequency variability of both SST and Z_{700} fields, so that the effective aspect ratio could be in fact rather small, on the order of $\mathcal{O}(10)/300 \approx 1/30$.

1.2.3 Signal versus noise

Climatic variability is comprised of phenomena with time scales anywhere from days to millenia and spatial scales ranging from 100 km (e.g., hurricanes, ocean currents and eddies) to global scale (glacial-to-interglacial climate transitions, global warming). In general, the processes that operate on larger time scales tend to involve structures with larger spatial scales as well. If one imagines a hypothetical input data matrix that describes all aspects of climate variability, this matrix would have a huge number of degrees of freedom and would virtually be impossible to analyze. Instead, one always concentrates on a climate system's subset, which is governed, presumably, by specific dynamics. The choice of climate subsystem for a phenomenon of interest is based on a physical intuition about the nature of this phenomenon.

Climate variability can be characterized as either *intrinsic* to a given subset of climate system, or *forced* by external sources, whose dynamics is not considered explicitly. An obvious example of external source of climate variability is solar forcing: seasonal changes in the amount of incoming solar radiation cause nearly periodic modulations in most climatic variables. In contrast to these externally forced variations, intrinsic variability arises as a result of complex interactions between various subcomponents of a given climate subsystem, and is typically characterized by a large degree of irregularity.

The notion of a *signal* is closely related to the concept of *predictability*: if the dynamics of a certain climate subsystem is well understood, it means, among other things, that one can assess how much we can say about its future evolution given the knowledge of past evolution; in particular, how far into the future we can make useful predictions. For example, synoptic meteorologists study the behavior of individual storms and can produce successful weather forecasts up to a few days into the future. The behavior of individual storms here is the signal that is being studied. The signal is typically characterized by a particular spatial and time scale. The phenomena that have shorter time scales and smaller spatial scales than those of the signal and are unpredictable on a time scale of a signal are considered as *noise*. In the example above, individual clouds within a storm can be thought of as noise. Therefore, the (subjective) definitions of signal and noise above depend on the dynamics considered and time scale of interest: "What's one person's signal is another person's noise."

Due to nonlinear nature of climate system, it is impossible to "decompose" climate evolution into a set of signals governed by separate dynamics: the climate variability involves interaction between processes on a wide range of time scales. The climatic data sets are thus mixtures of signal and noise; furthermore, they are typically characterized by a very low *signal-to-noise ratio*. An important task of statistical data analysis is to help identify signals in geophysical time series and use this information to (i) develop physical understanding of the phenomena of interest; and (ii) establish predictability limits associated with this phenomenon.

1.3 Elementary statistical concepts

Observed climatic quantities are generally not exact and are always subject to measurement errors (due to *instrumental noise*). Even if the data set is produced by an integration of a numerical model and contains no measurement errors *per se*, the (large-scale, low-frequency) signal of interest is typically contaminated by noise due to irregular, chaotic character of higher-frequency variability. Furthermore, the process under consideration can itself have random features either due to its intrinsic nonlinear dynamics, or due to interactions with high-frequency transients. Therefore, identification of a climate signal and its prediction, and often times the most natural description of the signal itself, are best of all formulated in probabilistic terms.

1.3.1 Probability distributions. Events. Statistical independence

Let's call a *collection*, or *population* $\{x_n\}$, $1 \leq n < \infty$, an infinite sequence of observations of some quantity x, which can attain either discrete or continuous set of finite values. In discrete case of I possible outcomes, we can define an *event* as an occurrence of a given value of $x = x^{(i)}$, $1 < i \leq I$. Suppose that among first N elements of the sequence $\{x_n\}$, the event $x^{(i)}$ occurs $N^{(i)}(N)$ times. The probability p_i of an event $x^{(i)}$ is then given by

$$p_i \equiv \lim_{N\to\infty} \frac{N^{(i)}(N)}{N}, \tag{1.3}$$

provided the limit in (1.3) exists[1]. The sum of event probabilities over all events $\sum\limits_{i=1}^{I} p_i = 1$. In other words, the probability of observing each time any one of I possible values of x is 100%. The set $(p_1, p_2, \ldots, p_I)$ is called the *discrete probability distribution* of a collective $\{x_n\}$.

In an analogous fashion, continuous distributions are characterized by the *probability density function* (PDF) $p(x)$. The probability of observing, in a given experiment, an event

[1] A more rigorous definition of probability would require, in addition to the existence of limiting frequency p_i, that our infinite sequence also satisfied the condition of randomness (Von Mises 1964), that is p_i's independence of *place selection*. An example of place selection would be to take only even or odd elements of a primary sequence.

“the value of x belongs to the interval $[a,\, b]$” is

$$P(a \le x \le b) = \int\limits_a^b p(x)\, dx; \quad \int\limits_{-\infty}^{\infty} p(x)\, dx = 1. \tag{1.4}$$

Another useful quantity is the so-called *cumulative distribution function* (c.d.f) $P(\xi)$, which is defined as the probability of obtaining the value of x that is smaller than a given value ξ:

$$P(\xi) = \int\limits_{-\infty}^{\xi} p(x)\, dx; \quad P(-\infty) = 0; \quad P(\infty) = 1. \tag{1.5}$$

The c.d.f. is increasing monotonically from the value of 0 at $-\infty$ to the value of 1 at ∞ (the probability of observing a finite value of x is 100%). The c.d.f. in the case of a discrete distribution is a step function increasing from zero to one in a number of finite jumps.

$\boxed{!}$ Multivariate distributions can be constructed in an analogous way. For example, if we are given two-dimensional collective of pairs $\{x_n,\, y_n\}$, $1 \le n < \infty$, in which x and y take a continuous set of finite values, the probability of observing an event “$x \in [a,\, b]$, $y \in [c,\, d]$” is

$$P(a \le x \le b;\, c \le y \le d) = \int\limits_a^b \int\limits_c^d p(x,y)\, dx\, dy; \quad \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} p(x,y)\, dx\, dy = 1, \tag{1.6}$$

where $p(x,y)$ is the two-dimensional PDF[2].

$\boxed{!}$ In the case of observational data, as in (1.1) and (1.2), the *sample distributions* (that is, distributions computed as in (1.4), (1.6), but based on a finite number of data points) are necessarily discontinuous. We can, however, approximate an observed distribution as a continuous one by, for example, interpolation. The distribution so obtained can be compared with some other (known) observed or theoretical distribution; this comparison might point to interesting dynamical features of the phenomenon under consideration. Useful methods for estimating a sample’s PDF are discussed in Chapter 5.

Conditional probability. Suppose that we are given two events E_1 and E_2, whose probabilities are $P(E_1)$ and $P(E_2)$, respectively. Events could be, for example, E_1 = “the value

[2] Note that if we were to “scramble” or *repartition* two-dimensional data, the $(x_n,\, y_n)$ pairs should be preserved, otherwise the PDF of a new collective will not, in general, be equal to original PDF. In other words, we cannot reshuffle $\{x_n\}$ and $\{y_n\}$ independently without affecting the two-dimensional PDF, unless x and y are *statistically independent* (probability of a certain value of x does not depend on the value of y and vice versa [see below]).

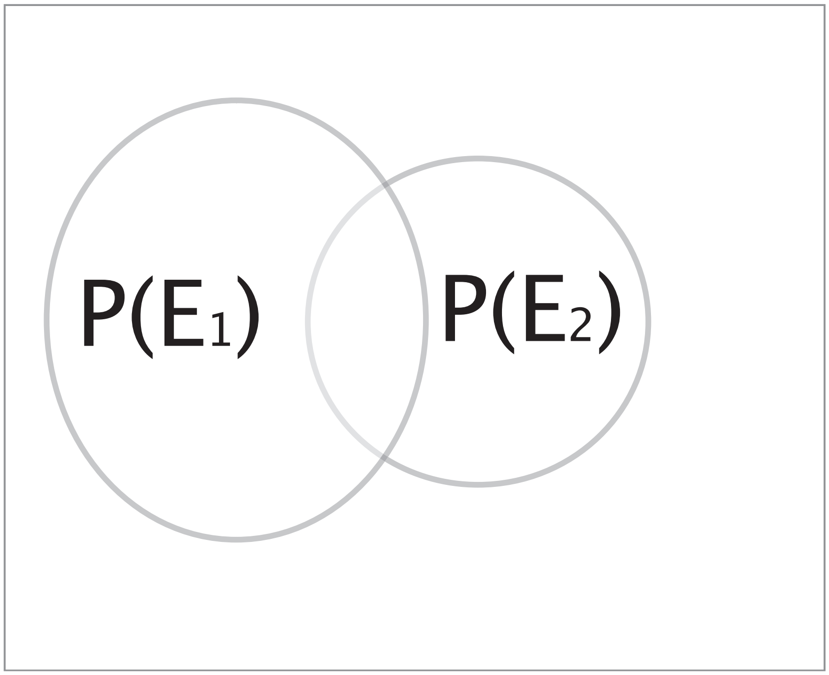


Figure 1.2: Venn diagram: the area within a rectangle represents the total probability of one, while the area inside the two ellipses — probabilities of the events E_1 and E_2.

of some ENSO index in January exceeds a certain threshold (see Fig. 1.1)," and E_2 = "the value of the same index in February exceeds some other threshold." The probability that E_1 $[E_2]$ will not occur is $1 - P(E_1)$ $[1 - P(E_2)]$. Let's call $P(E_1) \cap P(E_2)$ the probability that both events will occur (*intersection of probabilities*), and $P(E_1) \cup P(E_2)$ the probability that at least one of the events (E_1 or E_2) will occur (*union of probabilities*). As seen from Fig. 1.2, the union of the two probabilities is

$$P(E_1) \cup P(E_2) = P(E_1) + P(E_2) - P(E_1) \cap P(E_2), \tag{1.7}$$

since in adding the two events' areas the intersection gets counted twice and, therefore, must be subtracted. The intersection of mutually exclusive events is zero.

An important statistical concept is the one of *conditional probability* $P(E_2 \,|\, E_1)$, that is the probability that E_2 will occur given that E_1 has occurred. From Fig. 1.2 , this quantity equals to the ratio of the intersection's area to the E_1-ellipse area:

$$P(E_2 \,|\, E_1) = \frac{P(E_1) \cap P(E_2)}{P(E_1)}. \tag{1.8}$$

Rearranging (1.8), we have

$$P(E_1) \cap P(E_2) = P(E_2 \,|\, E_1) \cdot P(E_1) = P(E_1 \,|\, E_2) \cdot P(E_2). \tag{1.9}$$

The latter formula represents *multiplicative law of probability.* If the two events are independent $[P(E_2 \,|\, E_1) = P(E_2)]$, it follows from (1.9) that

$$P(E_1) \cap P(E_2) = P(E_1) \cdot P(E_2). \tag{1.10}$$

Example 1.2 *If the probability of getting heads (tails) on a coin flip is* 0.5 *and the flips are independent of one another, the probability of getting heads (tails) N times in a row is* 0.5^N*; the probability thus decreases with N exponentially*[3]. *An alternative example of the case in which the events are likely to be highly dependent is the ENSO index example above: while the probability of having the index exceed the threshold in February could be low, the conditional probability of this event's occurrence given the threshold has been exceeded in January can be rather high, close to* 1.

1.3.2 Fundamental statistical quantities

Mean value. Variance of a distribution

Suppose that we have a sample $\{x_n\}$, $1 \le n \le N$ from a collective with a one-dimensional discrete distribution, in which x can take values $x^{(1)}$, $x^{(2)}$, ..., $x^{(K)}$, and that each of $x^{(k)}$ has occurred N_k times. The average value of x is then given by

$$\frac{1}{N}\sum_{k=1}^{K} N_k x^{(k)} = x^{(1)}\frac{N_1}{N} + x^{(2)}\frac{N_2}{N} + \ldots + x^{(K)}\frac{N_K}{N}. \tag{1.11}$$

Taking the limit $N \to \infty$ and introducing probabilities $p_k = p(x^{(k)}) = \lim\limits_{N\to\infty} N_k(N)/N$, we obtain the *mean value* θ of the distribution under consideration:

$$\theta = \sum_{k=1}^{K} x^{(k)} p_k = \sum_{k=1}^{K} x^{(k)} p(x^{(k)}). \tag{1.12a}$$

[3] In this example, we have derived from a collective (an infinite number of outcomes of a coin flipping) with a discrete distribution p_1((heads) = 0.5, p_2(tails) = 0.5, a different collective, in which we consider an infinite number of N-flip sequences and define the two possible events to be "all N flips are heads" and "at least one in N flips is tails." We then computed the probability distribution in this new collective. *The general task of probability calculus is to compute the probability distribution in derived collectives from the given distributions in the collectives from which they have been derived.*

In the case of continuos distribution with the probability density function $p(x)$ the mean value is given by

$$\theta = \int_{-\infty}^{\infty} xp(x)dx. \tag{1.12b}$$

The *variance* σ^2 of the distribution, which characterizes the spread of x-values around their mean, is defined as

$$\sigma^2 = \sum_{k=1}^{K} (x^{(k)} - \theta)^2 p(x^{(k)}) \tag{1.13a}$$

in the discrete case and as

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \theta)^2 p(x)dx. \tag{1.13b}$$

for the continuous distribution. The quantity σ is called the *standard deviation.*

Expectation relative to a distribution. Moments of a distribution

Mean value and variance of a distribution are special cases of *functionals* (that is functions of functions) that can be derived relative to a given distribution. Let $f(x)$ be a function, defined for all $x^{(k)}$ of a discrete distribution and for all x of a continuous distribution; in the latter case we also assume that $f(x)$ is continuous in the domain $p(x) > 0$. The *expectation* $E\{f\}$ of f relative to the distribution under consideration for a discrete and continuous distribution are given by

$$E\{f\} = \sum_{k=1}^{K} f(x^{(k)}) p(x^{(k)}) = \int_{-\infty}^{\infty} f(x)dP(x) \tag{1.14a}$$

and

$$E\{f\} = \int_{-\infty}^{\infty} f(x)p(x)dx = \int_{-\infty}^{\infty} f(x)dP(x), \tag{1.14b}$$

respectively. Here $P(x)$ is the cumulative distribution function (see Section 1.3.1).

Using the definitions above, the mean value and variance of a distribution can be written as

$$\theta = E\{x\}, \quad \sigma^2 = E\{(x-\theta)^2\} = E\{(x - E\{x\})^2\}. \tag{1.15}$$

To further characterize a distribution, we also introduce moments of a distribution $P(x)$ about some quantity c. The moment of order r, $M_c^{(r)}$, is defined as

$$M_c^{(r)} = \int\limits_{-\infty}^{\infty} (x-c)^r \, dP(x). \tag{1.16}$$

The mean of a distribution is thus $\theta = M_0^{(1)}$, and the variance is $\sigma^2 = M_\theta^{(2)}$.

Higher order moments taken about the mean are usually nondimensionalized by an appropriate power of the standard deviation. The *skewness* $m_3 = M_\theta^{(3)}/\sigma^3$ measures the degree of asymmetry of the distribution about the mean. Positive skewness corresponds to a distribution with a longer tail on a positive side of the mean and vice versa. The *kurtosis* $m_4 = M_\theta^{(4)}/\sigma^4$ is similar to variance in that it measures the spread of a distribution about the mean.

Median and mode

The *median* x_{med} of a probability distribution function $p(x)$ is the value of x for which larger and smaller values of x are equally probable:

$$\int\limits_{-\infty}^{x_{\text{med}}} p(x)\, dx = \int\limits_{x_{\text{med}}}^{\infty} p(x)\, dx. \tag{1.17}$$

The median of a distribution can be estimated from a finite sample $\{x_n\}$, $1 \le n \le N$, with N being odd, as the value of x_i which has equal numbers of values above and below it, or as the mean of two central values if N is even.

The *mode* of a probability distribution function $p(x)$ is the value of x where $p(x)$ takes the maximum value. If a distribution has two relative maxima, one says that this distribution is *bimodal.* Bimodal or multi-modal distributions may arise in nonlinear systems characterized by the presence of multiple *attractors*; for example, multiple steady states (stable or unstable). Inferring the structure of the observed PDFs may thus provide useful information about the dynamical properties of (known or unknown) underlying equations (see Chapter 5).

1.3.3 Distributions in more than one dimension

We now consider the case of a two-dimensional distribution; the extension to a general case of I-dimensional distribution is analogous. The mean $(\theta,\, \phi)$ of a two-dimensional collective $\{x_n,\, y_n\}$ is given by

$$\theta = \sum_{k=1}^{K} \sum_{l=1}^{L} x^{(k)} p(x^{(k)},\, y^{(l)}); \quad \phi = \sum_{k=1}^{K} \sum_{l=1}^{L} y^{(l)} p(x^{(k)},\, y^{(l)}). \tag{1.18a}$$

Here x and y were assumed to attain a discrete set of values $x^{(1)}$, $x^{(2)}$, $\ldots$, $x^{(K)}$ and $y^{(1)}$, $y^{(2)}$, $\ldots$, $y^{(L)}$, respectively; $p(x^{(k)},\, y^{(l)})$ is probability of the event $(x_n,\, y_n) = (x^{(k)},\, y^{(l)})$. For a continuous distribution with the probability density function $p(x)$ the expressions for θ and ϕ are

$$\theta = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} x p(x,\, y)\, dx\, dy; \quad \phi = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} y p(x,\, y)\, dx\, dy. \tag{1.18b}$$

The object which is analogous to the variance in I dimensions is a $I \times I$ matrix $C_{i,j}$, $1 \le i,\, j \le I$, called the *covariance matrix.* The covariance matrix is symmetric, that is $C_{i\,j} = C_{j\,i}$. For our two-dimensional example in discrete and continuous cases, the components of the covariance matrix are written as

$$\begin{aligned} C_{11} &= \sum_{k=1}^{K} \sum_{l=1}^{L} (x^{(k)} - \theta)^2 p(x^{(k)},\, y^{(l)}); \quad C_{22} = \sum_{k=1}^{K} \sum_{l=1}^{L} (y^{(l)} - \phi)^2 p(x^{(k)},\, y^{(l)}); \\ C_{21} &= C_{12} = \sum_{k=1}^{K} \sum_{l=1}^{L} (x^{(k)} - \theta)(y^{(l)} - \phi) p(x^{(k)},\, y^{(l)}), \end{aligned} \tag{1.19a}$$

$$\begin{aligned} C_{11} &= \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} (x-\theta)^2 p(x,\, y)\, dx\, dy; \quad C_{22} = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} (y-\phi)^2 p(x,\, y)\, dx\, dy; \\ C_{21} &= C_{12} = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} (x-\theta)(y-\phi) p(x,\, y)\, dx\, dy, \end{aligned} \tag{1.19b}$$

respectively.

References

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1994: *Numerical Recipes.* 2-nd edition. Cambridge University Press, 994 pp.

Von Mises, R., 1964: *Mathematical Theory of Probability and Statistics.* Academic Press, New York.

Von Storch, H., and F. Zwiers, 1999: *Statistical Analysis in Climate Reserach.* Cambridge University Press, Cambridge, United Kingdom, 484pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* International Geophysics Series, v. 59), Academic Press, San Diego, 467pp.

Chapter 2

Statistical Inference and Hypothesis Testing

In this chapter, we will consider a set of statistical data $\{x_n\}$, $1 \le n \le N_x$. The data set can represent, for example, N_x observations of some physical quantity x (e.g., temperature, amount of daily precipitation), in which case each of x_n may take continuous values. In other cases, only a set of discrete x-values can be realized. The latter situation typically occurs when we derive a new sample from our original sample of data by counting certain “events” (e.g., if the amount of daily precipitation for a given day exceeds a certain threshold, we tabulate this event as “1,” while in the opposite case we assign to that day the number “0”). We want to infer from our data set some probabilistic information, *viz.* how well can the set of $\{x_n\}$ be described in terms of some known statistical distribution? In other cases, we will consider an additional data sample $\{y_n\}$, $1 \le n \le N_y$, and seek to establish the “sameness” or “differentness” of two data sets. For example, we are analyzing output from two different climate models and would like to know if both of them produce the same time-mean state or if there are significant differences between the two models' climates.

2.1 The average and dispersion of a data sample

The *average* value a and the *dispersion* s^2 of the sample $\{x_n\}$, $1 \le n \le N$ are defined as

$$a = \frac{1}{N}\sum_{n=1}^{N} x_n; \quad s^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - a)^2 = \frac{1}{N}\sum_{n=1}^{N} x_n^2 - a^2. \tag{2.1}$$

The quantities a and s^2 should not be confused with the corresponding characteristics (that is, the mean and the variance) of infinite samples (infinite samples are called collectives or populations; see Sections 1.3.1 and 1.3.2 of Chapter 1).

We assume that each observation x_n is randomly taken from a population characterized by a cumulative distribution function $P_n(x)$ (probability of obtaining x_n-value smaller or equal to x).

$\boxed{!}$ It is often assumed that the c.d.f is the same for all observations ($P_n(x) = P(x)$), but in some cases it is natural to consider more general case of a finite set of c.d.f.'s. For example, monthly Niño-3 index time series exhibits a strong periodic modulation associated with seasonal forcing; the distributions of this index values for a given month are thus likely to have different means, and possibly different variances as well.

The mean values and variances of the n-th theoretical distribution are given by

$$\theta_n = \int\limits_{-\infty}^{\infty} x dP_n(x); \quad \sigma_n^2 = \int\limits_{-\infty}^{\infty} (x-\theta_n)^2 dP_n(x) = \int\limits_{-\infty}^{\infty} x^2 dP_n(x) - \theta_n^2. \tag{2.2}$$

Let us also assume that our observations are independent, so that the probability of obtaining a set of values $x_1,\, x_2,\, \ldots,\, x_N$ has a distribution function with the element $dP_1(x_1)$, $dP_2(x_2)$, $\ldots$, $dP_N(x_N)$. The expectation E of any function of N variables $F(x_1,\, x_2,\, \ldots,\, x_N)$ with respect to this distribution is

$$E\{F(x_1,\, x_2,\, \ldots,\, x_N)\} = \int\int \ldots \int F(x_1,\, x_2,\, \ldots,\, x_N)\, dP_1(x_1)\, dP_2(x_2)\, \ldots\, dP_N(x_N), \tag{2.3}$$

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

If F depends on one variable only, that is $F = f(x_1)$, the expectation is found as

$$E\{f(x_1)\} = \int f(x_1)\, dP_1(x_1) \int dP_2(x_2) \ldots \int dP_N(x_N) = \int f(x_1)\, dP_1(x_1), \tag{2.4}$$

since the integrals $\int dP_n(x_n)$ are all equal to one. Similarly, for a product $F = f(x_1)g(x_2)$, the expectation is

$$E\{f(x_1)g(x_2)\} = \int f(x_1)\, dP_1(x_1) \int g(x_2)\, dP_2(x_2). \tag{2.5}$$

In addition, for any two constants c_1 and c_2, and any two functions F and G

$$E\{c_1F + c_2G\} = c_1E\{F\} + c_2E\{G\}. \tag{2.6}$$

Let us now compute expectations of the average a and dispersion s^2, as defined in (2.1). First, taking either $F = f(x_n) = x_n$ or $F = x_n^2$ and using (2.4), (2.5), we get

$$E\{x_n\} = \int x\, dP_n = \theta_n; \quad \int x^2\, dP_n = \sigma_n^2 + \theta_n^2, \tag{2.7a}$$

$$E\{x_l x_k\} = \theta_l \theta_k; \quad l \neq k. \tag{2.7b}$$

Now, due to (2.6), we find

$$E\{a\} = \frac{1}{N}\sum_{n=1}^{N} E\{x_n\} = \frac{1}{N}\sum_{n=1}^{N} \theta_n. \tag{2.8}$$

The expectation of the average of the sample $\{x_n\}$, $1 \leq n \leq N$ *is the average of the mean values* θ_n *of the individual distributions* $P_n(x)$. One also says that the average a defined by (2.1) is an *unbiased* estimate of the true mean.

In an analogous fashion, we get the following expression for the expectation of dispersion:

$$E\{s^2\} = \frac{1}{N}\sum_{n=1}^{N} E\{x_n^2\} - E\{a^2\} = \frac{1}{N}\sum_{n=1}^{N}(\sigma_n^2 + \theta_n^2) - \frac{1}{N^2}E\left\{\left(\sum_{n=1}^{N} x_n\right)^2\right\}.$$

Since

$$\left(\sum_{n=1}^{N} x_n\right)^2 = \sum_{n=1}^{N} x_n^2 + 2\sum_{l<k}^{1\ldots N} x_l x_k,$$

we get using (2.7a) and (2.7b)

$$E\left\{\left(\sum_{n=1}^{N} x_n\right)^2\right\} = \sum_{n=1}^{N}(\sigma_n^2 + \theta_n^2) + 2\sum_{l<k}^{1\ldots N}\theta_l\theta_k = \sum_{n=1}^{N}\sigma_n^2 + \left(\sum_{n=1}^{N}\theta_n\right)^2.$$

Substituting this expression into the equation above, we find

$$E\{s^2\} = \left(\frac{1}{N} - \frac{1}{N^2}\right)\sum_{n=1}^{N}\sigma_n^2 + \frac{1}{N}\sum_{n=1}^{N}\theta_n^2 - \left(\frac{1}{N}\sum_{n=1}^{N}\theta_n\right)^2.$$

The sum of the latter two terms is equal to the dispersion of N quantities θ_n:

$$\frac{1}{N}\sum_{n=1}^{N}\theta_n^2 - \theta^2 = \frac{1}{N}\sum_{n=1}^{N}(\theta_n - \theta)^2; \quad \theta = \frac{1}{N}\sum_{n=1}^{N}\theta_n.$$

The final expression for $E\{s^2\}$ is thus

$$E\{s^2\} = \frac{N-1}{N}\frac{1}{N}\sum_{n=1}^{N}\sigma_n^2 + \frac{1}{N}\sum_{n=1}^{N}(\theta_n-\theta)^2; \quad \theta = \frac{1}{N}\sum_{n=1}^{N}\theta_n. \tag{2.9}$$

The expectation of the dispersion s^2 of the sample $\{x_n\}$, $1 \le n \le N$ equals $(N-1)/N$ times the average of the variances σ_n^2 plus the dispersion of the mean values θ_n of the individual distributions $P_n(x)$.

In the case in which all $P_n(x)$ are equal or, at least, have the same values of θ and σ^2, (2.8) and (2.9) become

$$E\{a\} = \theta; \quad E\{s^2\} = \frac{N-1}{N}\sigma^2. \tag{2.10}$$

We can rewrite the second expression above in the form

$$\sigma^2 = E\left\{\frac{N}{N-1}s^2\right\} = E\left\{\frac{1}{N-1}\sum_{n=1}^{N}(x_n-a)^2\right\}. \tag{2.11}$$

Equation (2.11) thus gives an unbiased estimate of the true variance.

If the probability distributions $P_n(x)$ are known, one can compute the expectations of a and s^2 according to (2.8) and (2.9). If these expectations are not close to the values of a and s^2 derived from an available data sample, one can say that the hypothesis that the probability distributions $P_n(x)$ are underlying the data has been rejected. However, it is unclear how close the theoretical and observed values should be for us to reject our hypothesis. An answer to this question can be obtained by computing not only the expectations, but also the variances of a and s^2.

From (1.15), we have

$$\mathrm{Var}\{a\} = E\{(a-E\{a\})^2\} = E\{a^2\} - (E\{a\})^2. \tag{2.12}$$

We have already computed in our previous calculations $E\{a^2\}$, which equals to

$$E\{a^2\} = \frac{1}{N^2}E\left\{\left(\sum_{n=1}^{N}x_n\right)^2\right\} = \frac{1}{N^2}\sum_{n=1}^{N}\sigma_n^2 + \left(\frac{1}{N}\sum_{n=1}^{N}\theta_n\right)^2.$$

The last term in the above equation is simply $(E\{a\})^2$, and we end up with the following expression for $\mathrm{Var}\{a\}$:

$$\mathrm{Var}\{a\} = \frac{1}{N^2}\sum_{n=1}^{N}\sigma_n^2. \tag{2.13}$$

In the case of equal $P_n(x)$, (2.13) gives

$$\mathrm{Var}\{a\} = \sigma^2/N. \tag{2.14}$$

Let us now also compute, in the case of equal distributions, the variance of s^2. Calling τ^4 the moment of fourth order with respect to the mean

$$\tau^4 = \int\limits_{-\infty}^{\infty} (x-\theta)^4 \, dP(x) \tag{2.15}$$

and denoting $x'_n \equiv x_n - \theta$, we get

$$E\{x'_n\} = 0; \quad E\{{x'_n}^2\} = \sigma^2; \quad E\{{x'_n}^4\} = \tau^4; \quad E\{x'^2_l {x'_k}^2\} = \sigma^4, \quad l \neq k. \tag{2.16}$$

In addition, the expectation vanishes for all products which contain at least one variable in the first power, e.g. $x'_1 x'_2$, $x'_1 {x'_2}^2$, etc., due to (2.5) and the first formula (2.16).

By definition

$$\mathrm{Var}\{s^2\} = E\{s^4\} - \left(E\{s^2\}\right)^2. \tag{2.17}$$

The expression for s^2 in terms of the x' variables is

$$s^2 = \frac{1}{N}\sum_{n=1}^{N} {x'_n}^2 - \frac{1}{N^2}\left(\sum_{n=1}^{N} x'_n\right)^2 = \frac{N-1}{N^2}\sum_{n=1}^{N} {x'_n}^2 - \frac{2}{N^2}\sum_{l<k}^{1\ldots N} x'_l x'_k.$$

Note that in forming s^4, the product of the last two sums in the last expression only contains terms with one of the variables in the first power, whose expectations vanish. Also, the products of the two terms like $x'_l x'_k$ with $l \neq k$ have the expectation zero. Therefore,

$$E\{s^4\} = \frac{(N-1)^2}{N^4}\left[\sum_{n=1}^{N} E\{{x'_n}^4\} + 2\sum_{l<k}^{1\ldots N} E\{x'^2_l {x'_k}^2\}\right] + \frac{4}{N^4}\sum_{l<k}^{1\ldots N} x'^2_l {x'_k}^2$$

Substituting the expressions (2.16) into the above formula and noting that the number of terms with $l < k$ is $N(N-1)/2$, we get

$$E\{s^4\} = \frac{(N-1)^2}{N^3}\tau^4 + \left[\frac{N-1)^3}{N^3} + 2\frac{N-1}{N^3}\right]\sigma^4.$$

Combining this expression, as well as (2.10), with (2.17), we get

$$\mathrm{Var}\{s^2\} = \frac{N-1}{N^3}\left[(N-1)\tau^4 - (N-3)\sigma^4\right] \approx \frac{1}{N}(\tau^4 - \sigma^4), \tag{2.18}$$

with the last expression valid in the limit of large N.

In the case of a sample of size N drawn from the population whose distribution has a mean value θ and variance σ^2, the sample's average a and dispersion s^2 are given by

$$a = \theta \pm \frac{\sigma}{\sqrt{N}}; \quad s^2 = \frac{N-1}{N}\sigma^2 \pm \sqrt{\frac{\tau^4 - \sigma^4}{N}}. \tag{2.19}$$

Note that the standard deviations of both the average and the dispersion tend to zero as the sample size increases, but do so very slowly, at the $1/\sqrt{N}$ rate.

Exercise 1. Suppose we are given two samples $\{x_n\}$, $1 \le n \le N_\mathrm{x}$ and $\{y_n\}$, $1 \le n \le N_\mathrm{y}$. Each of $\{x_n\}$ is drawn from a distribution with a known mean $\theta_{\mathrm{x},n}$ and variance $\sigma^2_{\mathrm{x},n}$, while these quantities for each of $\{y_n\}$ are $\theta_{\mathrm{y},n}$ and $\sigma^2_{\mathrm{y},n}$, respectively. Consider a combined sample $\{z_n\}$, $1 \le n \le N_\mathrm{z}$ of the size $N_\mathrm{z} = N_\mathrm{x} + N_\mathrm{y}$, in which first N_x elements are the corresponding values of $\{x_n\}$ multiplied by $N_\mathrm{z}/N_\mathrm{x}$, while the remaining N_y elements are the values of $\{y_n\}$ multiplied by $-N_\mathrm{z}/N_\mathrm{y}$. Compute the expectation and the variance of the average of $\{z_n\}$. Express these quantities via expectations and variances of the average of original samples $\{x_n\}$ and $\{y_n\}$.

2.2 Central limit theorem. Normal distribution

A very important theorem in probability calculus is the **Central Limit Theorem**, a version of which due to Liapounoff we will formulate here without proof (see Von Mises 1964):

Theorem 1 *Consider a sample of independent chance variables $\{x_n\}$, $1 \le n \le N$ associated with distributions $P_n(x)$. Let θ_n, σ_n^2 be the mean and variance of the n-th distribution, respectively, and $M_n^{(k)}$ the absolute moment of order k about the mean. If individual variances σ_n^2 are all bounded and*

$$\lim_{N\to\infty} \frac{\sum\limits_{n=1}^{N} M_n^{(k)}}{\left(\sqrt{\sum\limits_{n=1}^{N} \sigma_n^2}\right)^k} \to 0, \quad \text{for} \quad \text{some} \quad k > 2, \tag{2.20}$$

then the probability density function $p(X;\,\theta,\,\sigma)$ of the sum $X \equiv \sum\limits_{n=1}^{N} x_n$, tends, for $N \to \infty$, to

$$p(X;\,\theta,\,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left[\frac{X-\theta}{\sigma}\right]^2\right\}, \tag{2.21}$$

where $\theta = \sum\limits_{n=1}^{N} \theta_n$, $\sigma^2 = \sum\limits_{n=1}^{N} \sigma_n^2$.

The two-parameter distribution (2.21) is called *Gaussian* or *normal* distribution with mean θ and variance σ^2. Simply put, the central limit theorem states that the distribution of the sum of independent chance variables drawn from **any** "sufficiently good" distributions will tend to normal as the size of a sample increases. Assuming the individual variances have the same order of magnitude, the sum in the denominator of (2.20) increases as N^k; therefore, for the individual distributions to be "good," it is sufficient that one of their moments of order 3 or higher be bounded.

Implications of the central limit theorem. Returning to climate system or, for that matter, to any nonlinear system with many degrees of freedom, and given the central limit theorem, we shouldn't be surprised to find out that most observables are distributed normally about their means, or can be transformed in such a way that they become normally distributed. If $\mathbf{X}$ is the climate-state vector, $\overline{\mathbf{X}}$ its time mean, and $\mathbf{x} = \mathbf{X} - \overline{\mathbf{X}}$ the vector of anomalies, then the evolution of $\mathbf{x}$ is expressed as

$$\dot{\mathbf{x}} = \mathbf{L}\mathbf{x} + \mathbf{N}(\mathbf{x}). \tag{2.22}$$

Here the dot denotes time derivative, $\mathbf{L}$ is a linear operator, and $\mathbf{N}$ represents nonlinear terms; both $\mathbf{L}$ and $\mathbf{N}$ may be functions of $\overline{\mathbf{X}}$. Even if the exact form of Eq. (2.22) were known, it would contain a very large number of degrees of freedom, so that its direct numerical integration would not be feasible due to insufficient computer power.

A common approach to solving Eq. (2.22) in practice is based on assuming scale separation. In this case, the full climate-variable vector $\mathbf{x}$ is represented as the sum of a climate "signal" $\mathbf{x}_{\mathrm{S}}$ and a "noise" $\mathbf{x}'_{\mathrm{N}}$:

$$\mathbf{x} = \mathbf{x}_{\mathrm{S}} + \mathbf{x}'_{\mathrm{N}}, \tag{2.23}$$

where the noise field is typically characterized by smaller scales in both space and time. Upon substituting the decomposition (2.23) into Eq. (2.22) and omitting the subscripts, the latter becomes:

$$\dot{\mathbf{x}} = \mathbf{L}\mathbf{x} + \mathbf{N}(\mathbf{x}) + \mathbf{R}(\mathbf{x},\,\mathbf{x}'). \tag{2.24}$$

In order to obtain a closed form of the reduced dynamics equation (2.24), one has to make assumptions about the term $\mathbf{R}(\mathbf{x}, \mathbf{x}')$. A closure of this Reynolds-stress term is used in many climate GCMs: one assumes that small-scale, high-frequency transients — due to instabilities of the large-scale, low-frequency flow — act on the latter as a linear diffusion that merely flattens, on long time scales, spatial gradients of the large-scale field; the corresponding eddy diffusivities are estimated from available data by trial-and-error. It is widely recognized, however, that the underlying assumption in this "eddy-diffusion" closure does not generally hold.

Another possible assumption is that the residual term $\mathbf{R}(\mathbf{x}, \mathbf{x}')$ in (2.24) depends only on the "fast" variables $\mathbf{x}'$: $\mathbf{R}(\mathbf{x}, \mathbf{x}') = \mathbf{R}(\mathbf{x}')$ [this is also not true in general, and used here for illustrative purposes only]. The equation governing evolution of $\mathbf{x}'$ also has the form of (2.22); the mathematical structure of linear and nonlinear operators in this equation is such that the "fast" subsystem is typically characterized by numerous instabilities which determine, in particular, the "memory" of this subsystem: once again, one measure of the memory is how fast the system "forgets" initial conditions. If we assume that the memory is short enough, then the "fast" component will be represented by a set of independent random fields. The derivation of the effective reduced dynamics equation (2.24) with forcing $\mathbf{R}(\mathbf{x}, \mathbf{x}') = \mathbf{R}(\mathbf{x}')$ usually involves time averaging on the slow time scale over a large number (on the order of the ratio of slow and fast time scales) of independent random realizations of $\mathbf{R}(\mathbf{x}')$: according to the central limit theorem, therefore, $\mathbf{R}(\mathbf{x}')$ can be modeled as the normally distributed random noise[1].

Note that if the "slow" nonlinear operator $\mathbf{N}(\mathbf{x}) = 0$, and noise forcing is normally distributed, the PDF of the signal will also be normally distributed. Therefore, tracking deviations from Gaussianity in the distribution of observed large-scale low-frequency fields can point to a nonlinear origin of the observed variability, which may in turn be associated with an increased climate predictability. Suppose, for example, that unforced version of (2.24) has a stable steady state. The system's trajectory will then tend to pause in the vicinity of this state and one can use this information to improve predictions (for example, make a skillful medium-range forecast). Analysis techniques for tracking, interpreting, and utilizing deviations from Gaussianity in observed data sets will be considered in greater detail in Chapter 5.

Increased predictability may also be associated with the presence of a preferred period in the "slow" climate subsystem: such oscillations may be nonlinear (intrinsic variability) or

[1] For the system with quadratic nonlinearities exhibiting significant time scale separation, one can derive a rigorous dynamical formulation of a reduced-order model, which involves cubic and quadratic nonlinearities, as well as additive and multiplicative noise.

linear. In the latter case, the oscillations are typically damped (do not occur in the absence of external forcing) and excited at the expense of energy supplied by noise. Detection of low-frequency oscillations in otherwise noisy time series will be a subject of Chapter 6.

2.3 Comparing means using normal distribution

2.3.1 Standard normal distribution

The cumulative distribution function $P(x;\, \theta,\, \sigma)$ associated with the normal distribution (2.21) is

$$P(X;\, \theta,\, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{X} \exp\left\{ -\frac{1}{2} \left[\frac{\xi - \theta}{\sigma} \right]^2 \right\} d\xi, \tag{2.25}$$

or, for standardized variable

$$z \equiv \frac{X - \theta}{\sigma} \tag{2.26}$$

the standard normal distribution is

$$P(z;\, 0,\, 1) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{z} \exp\left\{ -\frac{1}{2}\xi^2 \right\} d\xi. \tag{2.27}$$

The probability that a normally distributed variable falls within one standard deviation of its mean value is given by

$$P(-1 \leq z \leq 1) = \frac{1}{\sqrt{2\pi}} \int\limits_{-1}^{1} \exp\left\{ -\frac{1}{2}\xi^2 \right\} d\xi = 68.27\%, \tag{2.28}$$

and similarly for two and three standard deviations

$$P(-2 \leq z \leq 2) = \frac{1}{\sqrt{2\pi}} \int\limits_{-2}^{2} \exp\left\{ -\frac{1}{2}\xi^2 \right\} d\xi = 95.45\%,$$

$$P(-3 \leq z \leq 3) = \frac{1}{\sqrt{2\pi}} \int\limits_{-3}^{3} \exp\left\{ -\frac{1}{2}\xi^2 \right\} d\xi = 99.73\%.$$

Thus, there is only 4.55% probability that a normally distributed variable will deviate from its mean by more than two standard deviations. This probability is the two-tailed probability (both negative and positive values of the variable are considered). The probability that a normal variable will *exceed* its mean by more than two standard deviations is only half of that, which equals to 2.275% (see also Fig. 2.1).

2.3.2 Mean of a large sample

According the the central limit theorem, the average $\bar{x} \equiv \sum_{n=1}^{N} x_n$ of a very large sample $\{x_n\}$, $1 \leq n \leq N$; $N \to \infty$, will be distributed normally; the corresponding distribution parameters, that is the mean and the variance being given by (2.8), (2.10) and (2.13), (2.14), respectively[2]. The standard variable used to compare a sample mean to the true mean

$$z \equiv \frac{\bar{x} - \theta}{\sigma/\sqrt{N}} \tag{2.29}$$

has the standard normal distribution (2.27). The formula (2.29) defines the so-called z *statistic.*

If our variable is normally distributed, than the probability α of observing the value of z greater than some specified value z_α can be found as the area under the standard normal PDF (Fig. 2.1). Note that normal distribution is symmetric; therefore, the value of $z_{-\alpha}$ for which only α-fraction of realizations is expected to have $z < z_{-\alpha}$ is $z_{-\alpha} = z_\alpha$. The quantity $2(1-\alpha) \times 100\%$ is the so-called *confidence level* or *significance level.* For example, there is 95% probability that sampled z statistic falls in the interval

$$-z_{0.025} < \frac{\bar{x} - \theta}{\sigma/\sqrt{N}} < z_{0.025}. \tag{2.30}$$

Therefore, the true mean is expected to lie in the interval

$$\bar{x} - z_{0.025}\frac{\sigma}{\sqrt{N}} < \theta < \bar{x} + z_{0.025}\frac{\sigma}{\sqrt{N}} \tag{2.31}$$

[2] Suppose that we are drawing our samples from some population generally characterized by non-Gaussian PDF. The thought experiment we perform is thus as follows. We first generate a sample of N random numbers, whose distribution is given, and record the average value of this sample. We then generate another sample of the same size, and compute its average and so on. The procedure is repeated M times. If we now plot the PDF of the resulting sample of M average values, it will be normally distributed with the appropriate values of parameters θ and σ.

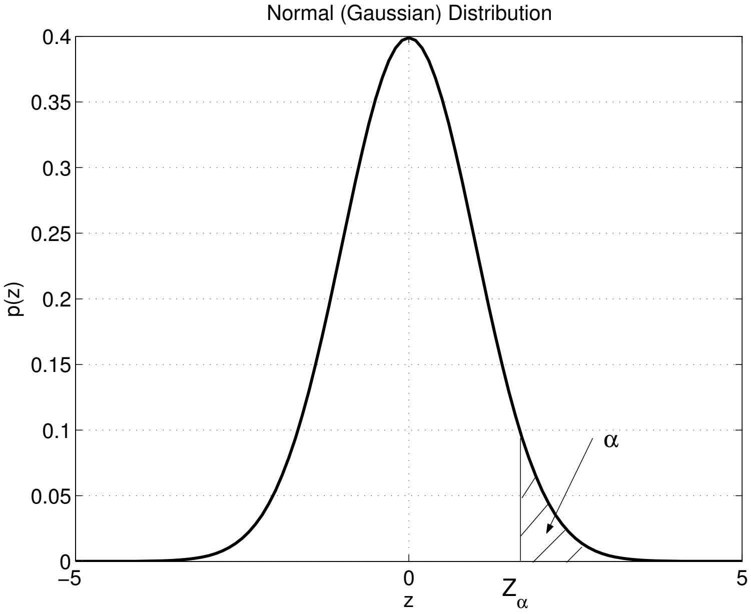


Figure 2.1: Normal PDF and significance testing: the value of z statistic z_α corresponds to the area α of a region bounded by the abscissa, vertical line at z_α and the standard normal PDF curve; this area equals the fraction of observed z values expected to exceed z_α. For example, $z_{0.15865} = 1$; here $\alpha = (1 - 0.6827)/2 = 0.15865$ [see (2.28)].

with 95% confidence[3] [compare with the first equation (2.19)]. If $\bar{x}$ is such that at least one of (2.31) is not satisfied, our *null hypothesis* that the underlying distribution has mean θ and variance σ^2 is *rejected* at the 95% confidence level[4].

Now, suppose we have two samples of data, of sizes N_1 and N_2, the corresponding variances being σ_1^2 and σ_2^2, and we expect that the difference between sample means $\Delta\bar{x} \equiv \bar{x}_1 - \bar{x}_2$ is Δ (often assumed to be zero in practice). We would like to know if our assumptions about σ's and $\Delta\bar{x}$ are correct. The standardized variable that provides a significance test

[3] Note that the sample mean θ is assumed to be known and constant, so that (2.31) in fact provides estimates on the sample's $\bar{x}$: $\theta - z_{0.025}\frac{\sigma}{\sqrt{N}} < \bar{x} < \theta + z_{0.025}\frac{\sigma}{\sqrt{N}}$ [(2.31) can be interpreted as an estimator of a true mean, but only in a special narrow sense — see Sections 2.6.3 and 2.7.5].

[4] Note that we did not have to assume anything about the PDFs of individual observations in our sample (for example, we did not have to assume that the data is Gaussian). All we need is for the sample to be large, then, according to the central limit theorem, the distribution of the sample mean will be normal.

for the differences between means is (see Exercise 1)

$$z = \frac{\Delta\bar{x} - \Delta}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}. \tag{2.32}$$

2.3.3 Small samples: Student's t distribution

One problem with respect to the results of the preceding section is that both the true mean and the true variance were assumed to be known *a priory.* In reality, we are typically given the sample average a and the sample dispersion s^2 based on some sample of finite size N. If N is very large (in practice larger than 30–50), these estimated values approach the true values (see Section 2.1), while the distribution of the mean approaches normal distribution (Section 2.2)[5].

The Student's t statistic is defined by substituting, in (2.29), the expected values of the mean and variance according to (2.19):

$$t \equiv \frac{\bar{x} - a}{\hat{s}/\sqrt{N}} = \frac{\bar{x} - a}{s/\sqrt{N-1}}; \quad \hat{s} = s\sqrt{\frac{N}{N-1}}. \tag{2.33}$$

Theorem 2 *If we draw a sample of size N from a normally distributed population, the values of* t *statistic (2.33) are distributed with the following probability density $f(t)$:*

$$f(t) = \frac{f_0(\nu)}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{(\nu+1)}{2}}}, \tag{2.34}$$

where $\nu = N - 1$ is the number of degrees of freedom and $f_0(\nu)$ is a constant that depends on ν and makes the area under the curve $f(t)$ equal to unity.

! The Student's t distribution is thus merely the probability density you expect to get when you draw a sample of finite size from a normally distributed population. If we have a small sample, therefore, drawn from a population that is not normally distributed, the mean of this sample won't in general be t-distributed. As the sample size increases, however, the distribution of the sample mean will tend to Student's (and normal) distribution irrespective of distributions of individual data entries.

[5] The condition of large N is usually fulfilled in climatic data.

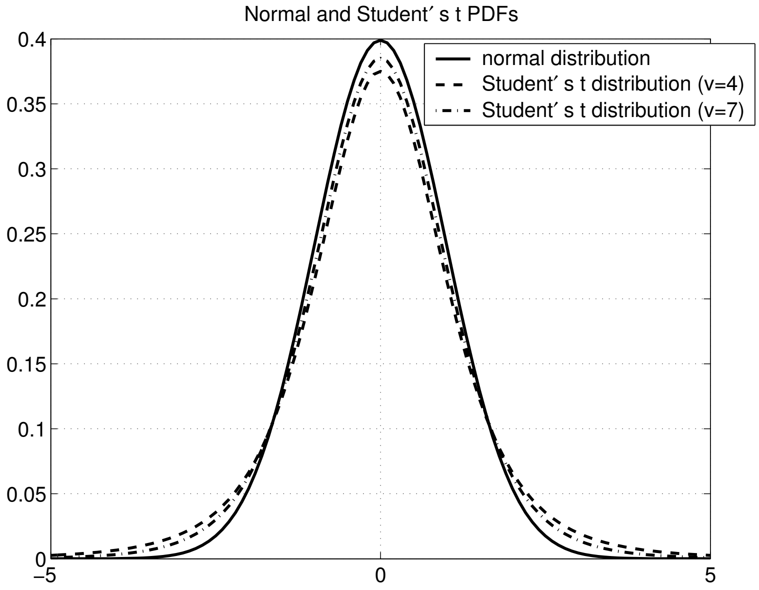


Figure 2.2: Student's t distribution approaches normal distribution as the sample size (and, therefore, number of degrees of freedom) increases.

The plots of $f(t)$ for $\nu = 4$ and $\nu = 7$ are shown in Fig. 2.2. Note that the tails of $f(t)$ are longer than in the corresponding normal distribution. As the sample size n and number of degrees of freedom $\nu = N - 1$ increases, the Student's t distribution tends to normal distribution. Because of the latter property, **there is no reason to use the normal distribution in preference to Student's t in testing statistical significance**.

The relevant statistic for measuring the significance of a difference of means $\Delta\bar{x}$ (relative to some expected value Δ) between two samples of different sizes N_1, N_2 and dispersions s_1^2, s_2^2 is [cf. (2.32)]:

$$t = \frac{\Delta\bar{x} - \Delta}{\sqrt{\frac{\hat{s}_1^2}{N_1} + \frac{\hat{s}_2^2}{N_2}}}; \quad \hat{s}_1 = s_1\sqrt{\frac{N_1}{N_1 - 1}}; \quad \hat{s}_2 = s_2\sqrt{\frac{N_2}{N_2 - 1}}. \tag{2.35}$$

The statistic above is distributed *approximately* as Student's t (2.34) with the number of degrees of freedom ν given by

$$\nu \approx \frac{\left[\frac{\hat{s}_1^2}{N_1} + \frac{\hat{s}_2^2}{N_2}\right]^2}{\frac{[\hat{s}_1/N_1]^2}{N_1-1} + \frac{[\hat{s}_2/N_2]^2}{N_2-1}}. \tag{2.36}$$

Example 2.1 (Due to D. Hartmann). *In a sample of ten winters the mean January temperature is $42°F$ and the standard deviation is $5°F$. What are the 95% confidence limits on the true mean January temperature?*

- *Desired confidence level is* 95%.
- *The null hypothesis is that the true mean is between $42 \pm \Delta T$. The alternative is that it is outside of this region.*
- *We will use the t statistics.*
- *The critical region is $|t| < t_{0.025}$, which for $\nu = N - 1 = 9$ is $|t| < 2.26$. From (2.33) we have*
$$\bar{T} - 2.26\frac{s}{\sqrt{N-1}} < \theta < \bar{T} + 2.26\frac{s}{\sqrt{N-1}}.$$
- *Plugging in the numbers we get $38.23 < \theta < 45.77$. We have 95% certainty that the true mean lies between these values.* If we had a guess about what the true mean was, *we could say whether the data would allow us to reject this null hypothesis at the significance level stated.*

Exercise 2. What would be the 95.45% confidence limits if we wrongly used z statistic in the example above?

2.4 Binomial distribution

2.4.1 The problem of repeated trials (Bernoulli)

Before we move further to consider the tests of whether two distributions have significantly different variances, we will need to discuss a fundamental problem which has numerous applications in both probability theory and statistical analysis of data.

Bernoulli problem. *Suppose we have a set of n independent trials; the outcome of each trial is either "success" or "failure," with a probability of a success being p and that of a failure $q = 1 - p$. What is the probability $p_n(k)$ of having exactly k successes?*

Example 2.2 *Consider extreme events in the Niño-3 index time series (see Fig. 1.1). To do so, we can form a new time series, of the size of the original time series according to the following rule: the value of the new time series at a given time equals 1 (success) if the index exceeds 1 standard deviation from its average value; otherwise, the new time series takes the value of 0 (failure). We thus end up with a sequence of zeros and ones, and estimate the value of p (or q) as ratio of the number of successes (or failures) to the total number of points in the time series. If all the events are independent, their distribution is given by the solution of the Bernoulli problem (see also Exercise 5).*

Once again, to compute the probability, we consider a population of M n-dimensional sets of independent trials. We should then count the number of sets $M_k(M)$ in which we have exactly k successes, and take the limit $\lim\limits_{M\to\infty} M_k(M)/M$. Take, for example, $n = 3$. The possible ($2^3 = 8$) outcomes of our trials are 000, 001, 010, 100, 011, 101, 110, 111. Since the events are independent, the probabilities $p_3(0)$, $p_3(1)$, $p_3(2)$, $p_3(3)$ of obtaining 0, 1, 2, and 3 successes, respectively, can be found as $p_3(0) = q^3$, $p_3(1) = 3pq^2$, $p_3(2) = 3p^2q$, $p_3(3) = p^3$. Note that $p_3(0) + p_3(1) + p_3(2) + p_3(3) = (p + q)^3 = 1$.

Higher-dimensional cases are considered in an analogous fashion: $p_n(k)$ has a general form of the sum of terms $q^k p^{n-k}$; the coefficients of this sum represent the number of possibilities to place individual objects (ones) on n places. Since these binomial coefficients are given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{2.37}$$

the solution to Bernoulli problem is

$$p_n(k) = \binom{n}{k} p^k q^{n-k}, \tag{2.38}$$

a binomial distribution (see an example in Fig. 2.3).

2.4.2 Mean value and variance of the binomial distribution

The moments of an arbitrary order for binomial distribution can be found using an identity

$$(q + pt)^n = \sum_k p_n(k) t^k. \tag{2.39}$$

Differentiating this equation with respect to t gives

$$n(q + pt)^{n-1} p = p_n(1) + 2p_n(2)t + 3p_n(3)t^2 + \ \ldots + np_n(n)t^{n-1}, \tag{2.40}$$

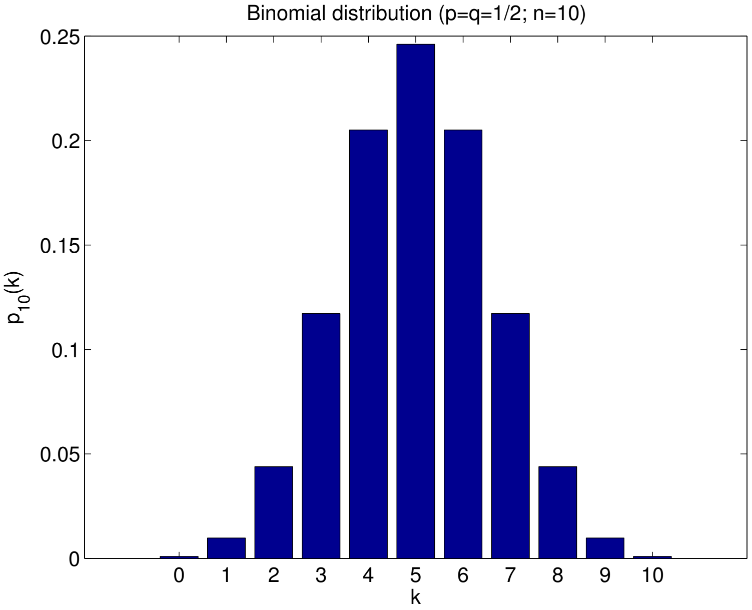


Figure 2.3: Binomial distribution $p_n(k)$ for $p = q = 1/2$ and $n = 10$.

which results, for $t = 1$, in the expression for the mean of the binomial distribution:

$$a_n \equiv E_n(k) \equiv \sum_k kp_n(k) = np. \tag{2.41}$$

[The left-hand side of (2.40) for $t = 1$ just equals nq, since $p + q = 1$, while the right-hand side is $\sum_k kp_n(k)$, which, by definition (1.12a), is the mean of distribution.]

Exercise 3. Show that the variance s_n^2 of the binomial distribution is given by

$$s_n^2 = npq. \tag{2.42}$$

Hint: differentiate (2.40) with respect to t and set $t = 1$, then use (1.12a), (1.13a) and an identity $p + q = 1$.

2.4.3 Normal approximation to binomial

The calculation of probabilities related to binomial distribution becomes tedious as the sample becomes large. Fortunately, we do not have to do this calculation, since we have the central limit theorem at our disposal (see Section 2.2). Note that the binomial problem can be reformulated as follows: find the distribution of the sum k of n independent random variables taking values 1 and 0 from a discrete distribution $p(1) = p$, $p(0) = q = 1 - p$ (since zeros do not contribute to this sum, we are indeed counting the number of ones, or successes). The expectation and the variance of the individual terms in this sum are equal to p and pq, respectively[6] [cf. (2.41) and (2.42)]. According to the central limit theorem, therefore, the statistic

$$\frac{k - np}{\sqrt{np(1-p)}}$$

has the standard normal distribution (2.27) as $n \to \infty$.

Example 2.3 (Due to D. Hartmann) *An earthquake forecaster has forecast 200 earthquakes. How many times in 200 trials must s/he be successful so we can say with* 95% *certainty that the forecasts have a nonzero skill?*

The null hypothesis H_0 is that forecasts have no skill (probability of success and failure are equal $p = q = 1/2$) and the confidence level is 0.05, or 95%*. The number of forecasts s^* that we want to find is thus given, according to (2.38), by*

$$P(s > s^* |\, H_0) = 0.05 = \sum_{i=s^*}^{200} \binom{200}{s} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{200-i}.$$

Solving this equation for s^* is extremely tedious. However, we can convert this to the problem

$$P(s > s^* |\, H_0) = P\left(\frac{s - np}{\sqrt{np(1-p)}} > \frac{s^* - np}{\sqrt{np(1-p)}}\right) = P\left(z > \frac{s^* - np}{\sqrt{np(1-p)}}\right) = 0.05,$$

where z has a standard normal distribution, for which $P(z > 1.64) = 0.05$. Our solution is, therefore,

$$\frac{s - np}{\sqrt{np(1-p)}} > 1.64; \quad s > 112.$$

So to pass a no-skill test on a sample of this size, the forecaster must be right 56% *of the time. This level of skill, while significantly different from zero, may not be practically useful.*

[6] Can you show this?

Exercise 4. Solve the above problem for the case of 20, rather than 200 trials.

Exercise 5. Consider Example 2.2: the total number of observations is 664 (slightly more than 55 years of monthly observations), the number of extreme events is 100. In the first half of the record there are 40 events, while there are 60 events in the second half of the record. Is this increase in the ENSO occurrences during the past 27 years statistically significant at the 95.54% significance level? Use normal approximation to the binomial distribution. Can we believe this answer?

Exercise 6. In the above example, we now define extreme events differently by binning the data using non-overlapping three-month box-car averages. The resulting time series has 220 points; there are now 34 events that exceed one standard deviation: 13 events in the first half of the time series and 21 events in the second half. Is this difference significant? *Hint:* Assume the events in both parts of the record come from the same binomial distribution; then compute probabilities of getting less than 14 and more than 20 events out of 110.

2.4.4 Non-parametric statistical tests

Binomial distribution can also be used to perform the so called *non-parametric* statistical tests, in which we do not to assume that the data's PDF is known. A good illustration of this approach is the Signs Test.

Suppose that we have paired data (x_i, y_i). We want to know if there is a shift in mean location from set x_i to set y_i. We know that the data are unlikely to be normally distributed and we don't want to assume that they are. We pose the statistical problem in terms of the two data set's medians: the null hypothesis is that the medians of the sets are identical; the alternative is that they are not. These statements can be written in terms of a probability $P(y_i > x_i)$ as

$$H_0 : \quad P(y_i > x_i) = 0.5; \quad H_1 : \quad P(y_i > x_i) \neq 0.5.$$

We next replace each pair with a signed integer equal to one according to the following rule:

$$y_i > x_i \longrightarrow + \qquad y_i < x_i \longrightarrow -$$

If the median of the two data sets are the same, the plus and minus signs should be equally probable, so that the + and − correspond to binomially distributed "success" and "failure." The probability of getting a certain number of + and − signs can thus be calculated using (2.38) with $p = q = 1/2$.

Example 2.4 (D. Hartmann) **Cloud seeding experiment.** *Ten pairs of very similar 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 resulted in the following table:*

Table 2.1: Cloud seeding experiment

Cloud Pair	Precipitation (untreated)	Precipitation (treated)	$y_i > x_i$?
1	10	12	+
2	6	8	+
3	15	10	−
4	3	7	+
5	5	6	+
6	14	4	−
7	12	14	+
8	2	8	+
9	17	29	+
10	8	10	+

There are thus eight pluses and two minuses. Is this statistically significant at the 95% level, so that we can say that the median values of the two samples are different? The chances of getting eight successes in ten trials are

$$P(k \geq 8) = \sum_{k=8}^{10} \binom{10}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{10-k} = 0.055,$$

$$P(k \leq 2) = \sum_{k=0}^{2} \binom{10}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{10-k} = 0.055.$$

Since our null hypothesis assumes random character in our data sets, it does not distinguish between positive and negative shifts of the median and we have to add up the two probabilities *above in the two-sided test. We get $P = 0.11$, which fails 95% confidence test. The effect (if any) appears not to be very pronounced; if we still want to investigate whether there is an effect, however small it is, we would have to perform more of cloud seeding experiments to make our data sample bigger.*

Exercise 7 (Bootstrap method). Another important nonparametric statistical technique is the *bootstrap method*. It belongs to the family of *Monte Carlo* methods (see Section 2.7.1 and Chapter 3), which involve generating a large number of synthetic realizations of a given data set using a statistical predictive model (this model is in turn derived from the data set under consideration). Constructing such a model relies on some assumptions about the data set. The model can be distribution-based (in this case, the PDF of the data set is estimated from the data, then sets of random numbers are drawn from this probability distribution) or trajectory-based (given past evolution of our variable and an estimate of the noise uncertainty, we predict the value of this variable at the next time); see, once again, Section 2.7.1 and Chapter 3 for further detail. If one does not know enough about the physical process underlying a given data set and/or nature/ditribution of errors (either measurement errors or those associated with the dynamical noise), one uses the bootstrap method, which actually belongs to the class of the distribution-based methods: it views the actual data set as a discrete probability distribution consisting of the delta functions at measured values.

Return to the cloud-seeding experiment Example 2.4. We have ten observations of the rainfall amount from treated and untreated clouds. Let us consider, for example, the results from the treated-cloud experiments and generate synthetic data sets of ten figures by the following procedure. Using a random-number generator producing a uniform random deviate that lies in the interval [0, 9] (take the one that produces deviates in the interval [0, 1] and multiply the numbers it produces by 9), rounding off to the nearest integer and adding 1 to the result, get a random integer number in the interval [1, 10]. Take the observation of precipitation from the corresponding experiment. If our random number is 3, for example, the first entry of our synthetic data set would be 48 (see Table 2.1). Repeating these drawing 10 times, we get a synthetic sample, whose average will in general be different from the original sample average since some of the data points will be duplicated (for instance, label "3" may occur more than once), while others not included. If we continue this procedure 1000 times, we will get 1000 estimates of the mean precipitation that fell out of the untreated clouds. After sorting the estimates in the ascending order, we can say that, for example, 95% of the precipitation data lies within the values associated with the 25-th and 975-th estimates. We can now do the same procedure with the rainfall data collected from seeded clouds and see if the 95% confidence intervals for the two tests overlap or not. Alternatively, we can sort the differences between the synthetic estimates and see directly if the observed difference in precipitation averages exceeds "synthetic" 95% level. *Do this estimation and see if the conclusion will be consistent with that of the Signs Test.*

There is a variety of distribution-free tests; examples include the “Wilcoxon signed rank test” and the “Wilcoxon–Mann–Whitney test”; see, for example, Mendenhall (1990) or class notes by Profs. D. B. Stephenson and R. E. Benestad (http://www.gfi.uib.no/∼nilsg/kurs/notes/node54.html).

2.5 Poisson distribution

2.5.1 Rare events

Consider the Liapounoff condition (2.20) under which the result (2.21) of the central limit theorem 1 is valid. The denominator of (2.20) is the sum of the individual variances. In order for this condition to be fulfilled, therefore, we need this sum to increase fast enough as the size of the sample increases. In the case of the Bernoulli problem (see Section 2.4.1), the denominator of (2.20) is given by (2.42):

$$s_n^2 = npq.$$

Suppose that p is very small: $p \to 0$, $q \to 1$; then, as $n \to \infty$, the sum of the variances will increase with n very slowly, and the normal approximation to the binomial distribution (2.38)

$$p_n(k) \equiv \binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left\{-\frac{(x-np)^2}{2npq}\right\}$$

will only hold if n is truly large so that $np \to \infty$. Small value of p means that the corresponding events can be characterized as rare events. A better-than-normal (that is, valid for smaller n) approximation to the binomial distribution for the case of rare events has been obtained by Poisson (1837).

2.5.2 Derivation of the Poisson law

Let us introduce, in the Bernoulli formula, the mean value $a \equiv np$ and rewrite this formula in the following form:

$$\begin{aligned} p_n(k) &= \frac{n(n-1)\,\ldots\,(n-k+1)}{k!}\left(1-\frac{a}{n}\right)^{n-k}\left(\frac{a}{n}\right)^k \\ &= \frac{a^k}{k!}\left(1-\frac{a}{n}\right)^n \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\,\ldots\,\left(1-\frac{k-1}{n}\right)}{\left(1-\frac{a}{n}\right)^k}. \end{aligned} \qquad (2.43)$$

If n increases, while a (NB! $p \to 0$) and k are kept constant, the latter fraction in (2.43) tends to unity, while

$$\lim_{n\to\infty}\left(1-\frac{a}{n}\right)^n = \mathrm{e}^{-a}.$$

The approximation to the Bernoulli formula (2.38) valid for large n and constant $a = pn$ and k is

$$p_\infty(k) = \lim_{\substack{n\to\infty\\ k=\mathrm{const},\, a\equiv np=\mathrm{const}}} p_n(k) = \frac{a^k \mathrm{e}^{-a}}{k!}. \tag{2.44}$$

The formula (2.44) gives a distribution $p_\infty(k)$, which is known as *Poisson distribution.* It is easily seen that

$$\sum_{k=0}^{\infty} p_\infty(k) = \mathrm{e}^{-a}\sum_{k=0}^{\infty}\frac{a^k}{k!} = \mathrm{e}^{-a}\cdot\mathrm{e}^{a} = 1. \tag{2.45}$$

The expectation of the Poisson-distributed variable is

$$E\{k\} = \sum_{k=0}^{\infty} k\, p_\infty(k) = a\mathrm{e}^{-a}\sum_{k=1}^{\infty}\frac{a^{k-1}}{(k-1)!} = a\mathrm{e}^{-a}\sum_{k=0}^{\infty}\frac{a^k}{k!} = a\mathrm{e}^{-a}\cdot\mathrm{e}^{a} = a. \tag{2.46}$$

Exercise 8. Show that the variance Var$\{k\}$ of a Poisson-distributed variable k is given by

$$\mathrm{Var}\{k\} \equiv E\{k^2\} - (E\{k\})^2 = a. \tag{2.47}$$

Hint: Compute first $E\{k(k-1)\}$.

Both the expectation and the variance of the Poisson-distributed variable equal to the distribution parameter a. The sum of independent Poisson random variables is also Poisson distributed with the parameter equal to the sum of the individual parameters[7]. According to the central limit theorem, on the other hand, the distribution of this sum tends to Gaussian distribution. The latter two facts combined mean that the Poisson distribution tends to Gaussian as $a \to \infty$ (in practice, it is essentially Gaussian at $a \approx 100$). An example of Poisson distribution is shown in Fig. 2.4.

[7] We are counting a large number of a small-probability events. If we have two *independent* sequences of events, with the sizes and individual event probabilities $n^{(1)}$, $n^{(2)}$ and $p^{(1)}$, $p^{(2)}$, respectively, we can form a new sequence of events with $n = n^{(1)}+ n^{(2)}$ and $p = [p^{(1)}n^{(1)} + p^{(2)}n^{(2)}]/n$. The mean value associated with the new sample will then be $a = np = a^{(1)} + a^{(2)}$, where $a^{(1)} = n^{(1)}p^{(1)}$ and $a^{(2)} = n^{(2)}p^{(2)}$ are these values for the two samples under consideration. The total number of events in the new sample will thus be Poisson-distributed with a parameter a.

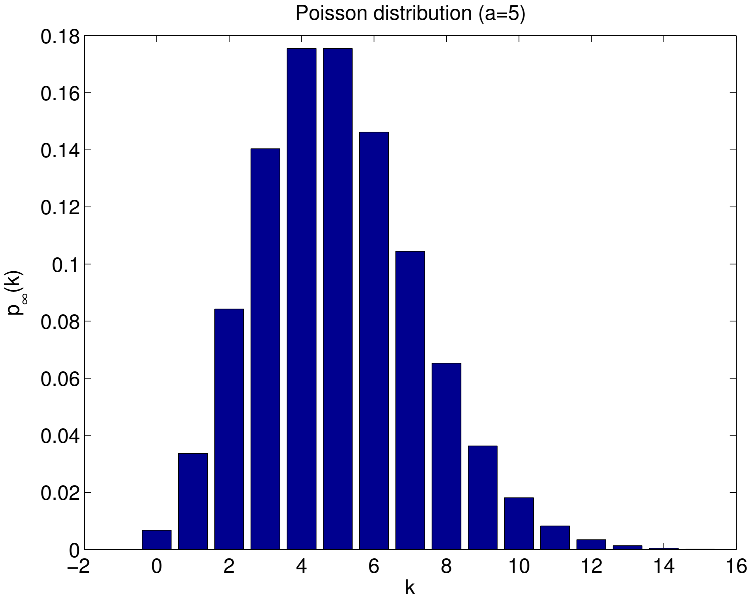


Figure 2.4: Poisson distribution $p_\infty(k)$ with $a = 5$.

2.5.3 Discussion and examples

Poisson vs. Gauss. We have seen that the Poisson distribution converges towards Gaussian as the number of rare events becomes large. However, this convergence is not uniform when measured in terms of fractional accuracy (ratio of the Poisson PDF to the Gaussian PDF with the same mean and standard deviation): the farther k is from its expected value, the poorer the fractional accuracy is. The Gaussian distribution always predicts that "tail" events are much less likely than they actually (by Poisson) are. It means that if we are interested in the extreme events that have a large magnitude, we have to have a larger sample size than in the case of, say, intermediate-amplitude events to achieve the same fractional accuracy of a Gaussian fit. In relatively short observational records, the normal distribution is, therefore, often rather poorly realized (for example, the events that exceed 3 standard deviations happen much more frequently than 0.27% of the time, as in the Gaussian case). Such cases are considered in the field of *robust statistics*, which we will briefly discuss in Chapter 3.

Brownian motion. A classical example of Poisson distribution in a physical system is the motion of small suspended particles (dust etc.) in liquid or gas. *In order to check random* *nature of this phenomenon*, one may count the number of particles that are present at a certain instant of time in some small portion of space occupied by the liquid, repeating these observations many times. The results are then plotted in terms of the number of occasions at which a certain number of particles was observed, divided by the total number of observations.

The number n of particles present in the fluid is very large. Since the space in which the observations are conducted is small compared to the total volume of fluid, the individual probability p for a particle to be found in this space is very small (this can be estimated as a ratio of the controlled volume to the total volume of fluid), but the expected number np of particles observed at a given time remains moderate. Thus the conditions under which Poisson formula solves approximately Bernoulli problem are fulfilled.

In order to apply the theory, one has first to estimate the value of $np = a$, the expected number of particles in the volume of observation. If the number of observations m is large, the expected number of particles in the control volume is approximately equal to the total number of particles M observed in all experiments divided by the number of experiments. The probability to find k particles in an experiment is then given by $p_\infty(k) = \mathrm{e}^{-a}a^k/k!$, while $mp_\infty(k)$ is the expected number of those cases in which k particles were counted. These expectations are the quantities to be compared with the observed frequencies. If, for example, in $m = 500$ observations a total of 1500 particles have been counted, then $500\mathrm{e}^{-3}3^3/2! = 112$ would be an expected number of observations with $k = 2$.

Persistent climatic states. Another example of the application of the Poisson distribution is the description of anomalously persistent climatic states. In this case, the events are defined as the occurrences in which the location of the tip of climate-state vector belongs to a certain (small) portion of the systems phase space (Fig. 2.5).

This control area is usually chosen to correspond to the phase-space region characterized by an enhanced PDF (see Chapter 5); such regions are presumably associated with recurrent and/or persistent events, whose knowledge may be useful for climate prediction. Analysis of the recurrent/persistent states may also point to the dynamical mechanisms which lead to enhanced climate predictability. It is therefore important to have an objective criterion that would allow one to decide whether a region of the phase space really stands out compared to other regions, either in terms of anomalous slow-down of the state-vector trajectory there (persistent states), or in terms of the frequency with which the control region is visited by the state vector (recurrent states).

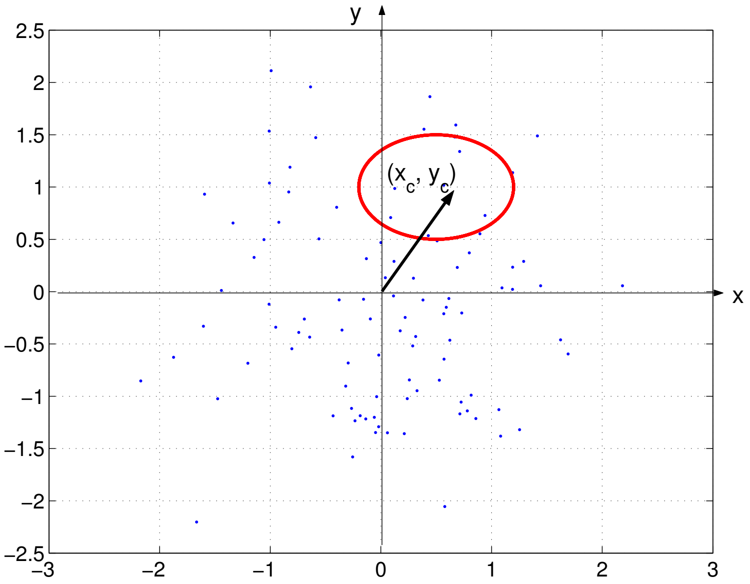


Figure 2.5: Schematic diagram of the climate-state vector evolution: the climate state is represented by a position of the tip of the vector in the system's phase space; climatological state is placed at the origin.

A possible strategy to accomplish the latter goal is the following. Once again, we first assume that the events are independent and estimate the probability p of a single event as a fraction of time the system spends in the region of interest. For a daily time series of length n, for example, we count the number of days in which the climate state was within our controlled volume in the phase space, and divide it by the total number of days in the time series. Provided the probability is small, but there is a lot of points in the time series, we might expect that the events are Poisson-distributed. This means that if we have divided our long time series into m shorter intervals (which should still be long enough to accommodate a large number of events for the Poisson asymptotic to be valid) and counted the total number of events within each interval, we expect that $M_k = m\mathrm{e}^{-pn/m}(pn/m)^k/k!$ intervals will contain k events, since the Poisson distribution parameter is $a = pn/m$ (probability p of a single event times the length $N = n/m$ of an interval).

The quantity M_k should be compared with the actual number $M_{k,1}$ of intervals containing k events. The confidence limits on M_k [the range of M_k within which a majority (say 95%) of M_k values based on samples of finite (but large) size N, drawn from population with discrete distribution $p(1) = p$; $p(0) = 1 - p$ (p is small), is expected to lie] can in principle be estimated analytically (for example, if the total number n of observations is large, mathematically speaking infinite, the observed and expected *distributions* of M_k can be compared using the so-called χ^2 *test*; see Section 2.6). In practice, however, the confidence limits are very easily determined numerically using *Monte Carlo* procedure (see Exercise 7) by generating many (typically 1000) surrogate sets (of size n) of random sequences of zeros and ones (probability of "one" equals p)[8], and computing their actual $M_k^{(s)}$ just as for the data set under consideration; 1000 estimates of this quantity so obtained are sorted in the ascending order. The upper/lower 95% confidence limits are then the 50-th/950-th values of $M_k^{(s)}$. Further discussion and examples of Monte Carlo significance tests can be found in Section 2.7.1, as well as in Chapters 3, 5 and 6.

The differences between our observed and theoretical distributions can exceed 95% confidence limits in some region of k-values. For example, the tail of the observed distribution can be significantly longer than that associated with Poisson distribution. This might mean that the events are probably not completely random and are most likely characterized by anomalously large persistence or recurrence [to distinguish between these two possibilities one has to analyze distribution of the length of "runs" (sequences of consecutive "ones")].

Now, let us choose another region of the phase space that *contains the same number of events (observations)* and repeat this procedure to find $M_{k,2}$. We can now study differences between $M_{k,1}$ and $M_{k,2}$ and estimate the statistical significance of these differences using confidence limits based on a null hypothesis that both sets of observations came from the same Poisson distribution[9].

2.5.4 Exponential and Gamma distributions

The Poisson distribution is intimately connected with the exponential distribution. Let us change notation $pn \longrightarrow \lambda t$, and consider Poisson distribution in which the average number of event occurrences per unit time (λ) is constant (in the formula above t stands for time). What is the distribution of the amount of time between events?

Let T be the amount of time until the first occurrence. The probability of no events in the time interval $[0,\, t]$ is $p_\infty(0) = \mathrm{e}^{-\lambda t}$ [see (2.44)]. By definition of T this also means that $P(T > t) = \mathrm{e}^{-\lambda t}$. The answer to our problem is thus given by $P(T \leq t) = 1 - P(T > t) = 1 - \mathrm{e}^{-\lambda t}$. This is a cumulative distribution function of the amount of time between

[8] A way to do so is to generate a set of random variables uniformly distributed in the interval [0, 1] and then assign to each individual entry a value of "one" if the random variable is less than p and "zero" otherwise.

[9] The tests of persistence will be considered in greater detail in Chapter 5.

events. The associated probability density function is just the derivative of this expression with respect to t:

$$f(t) = \lambda \mathrm{e}^{-\lambda t}, \tag{2.48}$$

the *exponential distribution*. The expectation value of the time between two events is thus $\int\limits_0^\infty \lambda \mathrm{e}^{-\lambda t} t\, dt = 1/\lambda$ (the integral is calculated using integration by parts), and the variance is $1/\lambda^2$ (Can you show this?).

The exponential distribution is a special case of a two-parameter *Gamma distribution*, whose PDF is given by

$$f(x|\, a,\, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} \mathrm{e}^{-\frac{x}{b}}, \tag{2.49}$$

where

$$\Gamma(a) = \int\limits_0^\infty t^{a-1} \mathrm{e}^{-t}\, dt$$

is the gamma function. It can be shown that when a is large, the gamma distribution closely approximates a normal distribution.

The gamma distribution has density only for positive real numbers. It can thus be used for a description of data which are not symmetrically distributed with respect to their values; for example rainfall data. Another special case of gamma distribution is the χ^2 distribution (see also Section 2.5.3) which gets special attention because of its importance in normal sampling theory.

2.6 χ^2 distribution

2.6.1 Checking a known distribution: χ^2 test

Suppose we are given binned data; for example, we have a temperature time series and group the events into K specified ranges of temperature (see Example 2.2). Let us call n_k the number of events which belong to k-th bin. We would like to check whether the observed probabilities of events n_k/n; $n = \sum_k n_k$ are consistent with some specified distribution p_k; see discussion in Section 2.5.3. The natural approach to this problem is to consider a test function of the form

$$F(n_1, n_2, \ldots, n_{k-1}) = \sum_{k=1}^{K} \lambda_k \left(\frac{n_k}{n} - p_k\right)^2, \tag{2.50}$$

where λ_k are some positive weights. The function F depends on $k-1$ variables only, since one of the variables can be expressed in terms of the others ($\sum\limits_k n_k = n$). We thus proceed by computing the expectation and variance of F under the assumption (and also a null hypothesis) that observations are drawn from a known population with probabilities $p_1, p_2, \ldots, p_K$, and then compare $E\{F\} \pm \sqrt{\mathrm{Var}\{F\}}$ with the observed value of F. The large observed value of F would indicate that the null hypothesis is rather unlikely.

The expectation $E\{F\}$ is the sum of expectations of individual terms in the sum (2.50). Consider the events in k-th bin only: the probability of occurrence of such an event is p_k, while the probability of non-occurrence is obviously $1-p_k$. The expectation of n_k is thus nq_k [see (2.41)], while the variance of n_k [that is, expectation of $(n_k - np_k)^2$] is $np_k(1-p_k)$ [see (2.42)]. Plugging the latter expression into (2.50), we get

$$E\{F\} = \frac{1}{n}\sum_{k=1}^{K} \lambda_k p_k (1-p_k). \tag{2.51}$$

If we choose

$$\lambda_k = \frac{n}{p_k}, \tag{2.52}$$

then the function F is denoted by χ^2 and called Chi-square:

$$\chi^2 = \frac{1}{n}\sum_{k=1}^{K} \frac{(n_k - np_k)^2}{p_k} = \frac{1}{n}\sum_{k=1}^{K} \frac{n_k^2}{p_k} - n. \tag{2.53}$$

Substituting λ_k from (2.52) into (2.51), we find that

$$E\{\chi^2\} = \frac{1}{n}\sum_{k=1}^{K} n(1-p_k) = \sum_{k=1}^{K} (1-p_k) = k-1. \tag{2.54}$$

The expectation of the χ^2 is independent of the specific values of the probabilities $p_1, p_2, \ldots, p_K$ and equal to the number of degrees of freedom; the latter equals the number of bins minus one. The simplicity of (2.54) and its independence of the underlying probability distribution indicate that the choice of λ_k according to (2.52) is a good one. This choice gives to the deviation squares $(n_k - np_k)^2$ the weights that are inversely proportional to the k-th event's expected frequency p_k. Therefore, deviations in the central portion and tails of the distribution are estimated with comparable fractional accuracy.

The variance of χ^2 can be computed in a similar fashion, as $E\{(\chi^2)^2\} - (E\{\chi^2\})^2$, and our program for checking whether observations are likely to come from a given distribution can be fulfilled. In practice, however, we usually deal with the situation in which the total number of observations n is large. In such cases, it turns out to be possible to find the *distribution of Chi-square*, that is the function that describes the probability of $a \leq \chi^2 \leq b$, for an arbitrary interval $[a,\, b]$.

2.6.2 Derivation of the χ^2 distribution for an infinite sample

In order to compute the distribution of χ^2, we first need to find the probability $P(n_1,\, n_2,\, \dots\, ,\, n_{K-1})$ of observing a set of numbers $n_1,\, n_2,\, \dots\, ,\, n_K$ (corresponding to events with probabilities $p_1,\, p_2,\, \dots\, ,\, p_K$) out of $n = \sum\limits_{k=1}^{K} n_k$ trials. The quantity $P(n_1,\, n_2,\, \dots\, ,\, n_{K-1})$ is given by the product of the probability of any individual arrangement of results including n_1 observations that fall in bin 1, n_2 observations that fall in bin 2 etc., and the number of such arrangements. The former term in this product is thus

$$p_1^{n_1}\, p_2^{n_2}\, \dots\, p_K^{n_K}\,,$$

while the number of arrangements indicates how many possible ways there are to select, out of n places, n_1 as belonging to one group $\binom{n}{n_1}$, then out of the remaining $n - n_1$ places n_2 for the second group $\binom{n-n_1}{n_2}$ and so forth. This number is thus equal to

$$\frac{n!}{n_1!\, n_2!\, \dots\, n_K!}.$$

The probability $P(n_1,\, n_2,\, \dots\, ,\, n_{K-1})$ is therefore given by

$$P(n_1,\, n_2,\, \dots\, ,\, n_{K-1}) = \frac{n!}{n_1!\, n_2!\, \dots\, n_K!} p_1^{n_1}\, p_2^{n_2}\, \dots\, p_K^{n_K}\,. \tag{2.55}$$

The probability distribution (2.55) is known as the *polynomial distribution*; it reduces to binomial formula (2.38) in a special case of $K = 2$.

It can be shown that as $n \to \infty$, the polynomial distribution (2.55) asymptotes the expression

$$P(n_1,\, n_2,\, \dots\, ,\, n_{K-1}) \approx \frac{\mathrm{e}^{-\chi^2/2}}{\sqrt{(2\pi n)^{K-1} p_1\, p_2\, \dots\, p_K}}, \tag{2.56}$$

with χ^2 given by (2.53). In order to derive (2.56), one should use the asymptotic expression ($n \to \infty$) for the factorial $n! = n^n e^{-n}\sqrt{2\pi n}$ and substitute it in place of all factorials appearing in (2.55). Computing then $\log P$, taking the limit $n \to \infty$, and using the definition of χ^2 (2.53) leads to (2.56).

The result (2.56) means that *the polynomial distribution becomes normal as n increases indefinitely; the probability has a constant value for all sets $n_1, n_2, \ldots, n_K$ to which one and the same value of χ^2 belongs.* Using this result, we can now compute the probability $P(a \le \chi^2 \le b)$. For this, let us introduce, in place of n_k, normalized variables

$$u_k = \frac{n_k - np_k}{\sqrt{np_k}}, \quad \text{or} \quad n_k = np_k\left(1 + \frac{u_k}{\sqrt{np_k}}\right). \tag{2.57}$$

The expression (2.53) for the χ^2 now becomes

$$\chi^2 = u_1^2 + u_2^2 + \ldots + u_K^2. \tag{2.58}$$

Consider first a special case of $K = 3$. Here

$$\chi^2 = u_1^2 + u_2^2 + \left(\frac{u_1\sqrt{p_1} + u_2\sqrt{p_2}}{\sqrt{p_3}}\right)^2.$$

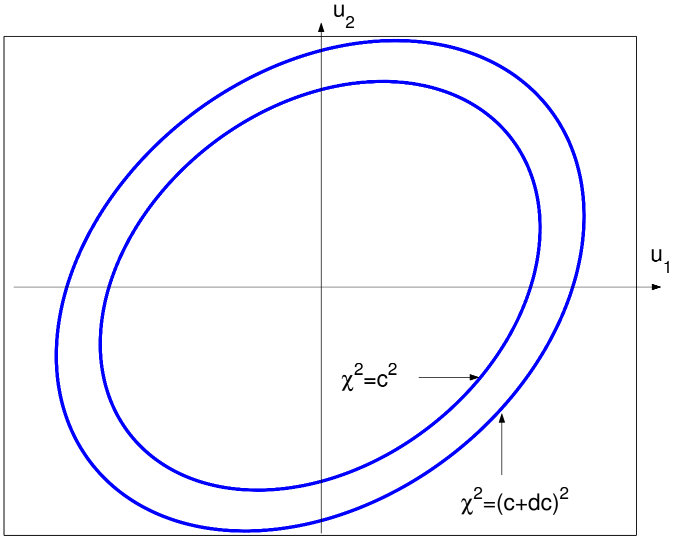


Figure 2.6: Schematic of constant χ^2 curves for $K = 3$.

In a (u_1, u_2) coordinate system, the locus of the points $\chi^2 = \text{const} = c^2$ is an ellipse (Fig. 2.6). The ellipses for different values of c all have the same orientation, while the size of their axes is proportional to c. The probability of χ^2 lying in the interval c^2 to $(c+dc)^2$ is given by the sum of all P-values that correspond to points with coordinates (u_1, u_2) within the annular arcs shown in Fig. 2.6. All these P are equal (except for infinitesimal differences) since they belong to essentially the same χ^2 value [see (2.56)] $\chi^2 = c^2$, while the number of points within the annular arcs tends to the area between the two arcs as $n \to \infty$. The latter area is equal to the length of the arc (this length is in turn proportional to c) and the distance dc. We thus have, for a three-dimensional case,

$$P\{c^2 \leq \chi^2 \leq (c+dc)^2\} = \text{const} \cdot \mathrm{e}^{-c^2/2} c dc \quad (K=3). \tag{2.59}$$

If we consider the case with $K = 4$, the surfaces of constant χ^2 in three dimensions will be ellipsoids, whose surface area will replace the length of the curve in the analysis above. Since the surface increases with the second power of linear dimensions, we would have to replace the factor c in (2.59) by c^2, and, in general case of K dimensions, by c^{K-2}. The expression (2.59) is replaced, in the latter case, by

$$P\{c^2 \leq \chi^2 \leq (c+dc)^2\} = \text{const} \cdot \mathrm{e}^{-c^2/2} c^{K-2} dc. \tag{2.60}$$

Changing the variable $c^2 \to x$, we have $(c+dc)^2 = c^2 + 2c\,dc = x + dx$ and $c^{K-2}dc = \frac{1}{2}c^{K-3}dx = \frac{1}{2}\sqrt{x}^{K-3}dx$, so that

$$P\{x \leq \chi^2 \leq x + dx\} = \text{const} \cdot \mathrm{e}^{-x/2} x^{(K-3)/2} dx. \tag{2.61}$$

The probability density of χ^2 is, therefore,

$$f_k(x) = \text{const} \cdot \mathrm{e}^{-x/2} x^{(K-3)/2}, \tag{2.62}$$

where the constant is found so that $\int\limits_0^\infty f_k(x)\,dx = 1$.

We arrive at the following result: *The probability density of χ^2 for infinite n is independent of the original probabilities p_1, p_2, ... , p_K and is given by*

$$f_K(x) = \frac{\mathrm{e}^{-x/2}}{2\Gamma(\frac{K-1}{2})} \left(\frac{x}{2}\right)^{(K-3)/2}, \tag{2.63}$$

where

$$\Gamma(a) = \int\limits_0^\infty t^{a-1}\mathrm{e}^{-t}\,dt$$

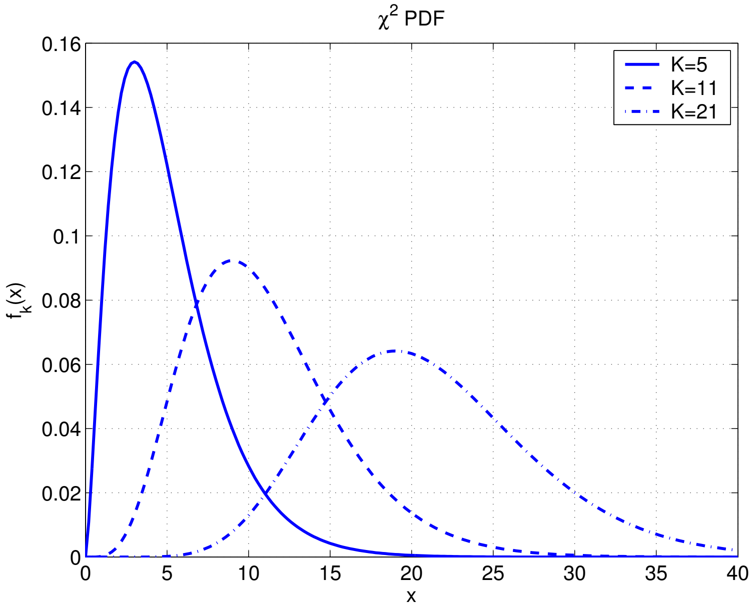


Figure 2.7: χ^2 PDF $f_K(x)$ for different values of K.

is the gamma function. The distribution (2.63) is known as the *Chi-square distribution with* $(K-1)$ *degrees of freedom.* Examples of $f_K(x)$ for different K are presented in Fig. 2.7.

It can easily be shown that the expectation and variance of χ^2 are equal to $(K-1)$ and $2(K-1)$, respectively:

$$\chi^2 = K - 1 \pm \sqrt{2K-2}. \tag{2.64}$$

! Note that $(K-1)$ is the expectation of χ^2 for whatever n [see (2.54)], while the variance is equal to $2(K-1)$ only for infinite n; the complete expression for variance (not shown) includes the term which is inversely proportional to n.

2.6.3 Normal sampling theory: Tests of variance

Just as for the Poisson-distributed random variables (see footnote at the end of Section 2.5.2), *the probability of the sum of two independent χ^2-distributed random variables is also χ^2-distributed, the number of degrees of freedom of the resulting distribution being equal to the sum of these numbers for two original distributions.* This immediately follows from the definition of χ^2 function (2.53): since the two sets of events are assumed to be independent, we can combine them and re-scale individual probabilities to accommodate a larger sample size and more bins (total number of possible events will be the sum of these numbers for the two original distributions). We can then take the limit of $n \to \infty$, as in the section 2.6.2, and arrive at the χ^2 distribution based on the new sample, which will have an appropriate number of degrees of freedom, as stated above[10].

Consider now random variable z having the standard normal distribution. What will be the distribution of z^2?

$$\begin{aligned} P(a < z^2 < b) &= 2P(\sqrt{a} < z < \sqrt{b}) = \frac{2}{\sqrt{2\pi}} \int\limits_{\sqrt{a}}^{\sqrt{b}} \mathrm{e}^{-\frac{z^2}{2}}\, dz = \\ & \frac{2}{\sqrt{2\pi}} \int\limits_{a}^{b} \frac{\mathrm{e}^{-\frac{x}{2}}}{2\sqrt{x}}\, dx = \int\limits_{a}^{b} \frac{1}{2\sqrt{\pi}} \mathrm{e}^{-\frac{x}{2}} \left(\frac{x}{2}\right)^{-\frac{1}{2}} dx. \end{aligned}$$

Noting that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and comparing the above expression with (2.63) for $K = 2$, we find that *the square of a random variable having the standard normal distribution has a χ^2 distribution with one degree of freedom.*

Now, using the latter property in combination with the summation property of the χ^2-distributed variables, we arrive at the following result: *The sum of squares of ν independent random* normal *variables of zero mean and unit variance has a χ^2 distribution with ν degrees of freedom.*

Testing for significance using χ^2. Suppose we want to know whether sample variances are truly different. One way to do so will be to consider a null hypothesis that both samples come from a normal distribution with zero mean and the same variance σ (we thus first subtract from each sample its respective average value). If we are given a set of N such

[10] From this property and the central limit theorem, it immediately follows that the χ^2 distribution tends to normal as the number of degrees of freedom tends to infinity, *viz.* if x is χ^2-distributed with $K-1$ degrees of freedom and $K \to \infty$, then the variable

$$z = \frac{x - (K-1)}{\sqrt{2(K-1)}}$$

has a standard normal distribution (2.27) [the normal distribution with zero mean and unit variance].

observations, and s is the sample standard deviation, then the statistic

$$\chi^2 = (N-1)\frac{s^2}{\sigma^2} \tag{2.65}$$

is χ^2 distributed with $\nu = N-1$ degrees of freedom[11].

The χ^2 distribution is not symmetric, so that, for example, the 95% confidence limits on the true variance will be

$$\frac{(N-1)s^2}{\chi^2_{0.975}} < \sigma^2 < \frac{(N-1)s^2}{\chi^2_{0.025}}. \tag{2.66}$$

The expression above means that if we had very many samples of size N drawn from a normal population with zero mean and variance σ, we would only expect 2.5% of the samples to have variances $s^2 > s^2_{0.975} = \sigma^2\chi^2_{0.975}/(N-1)$ and another 2.5% to have variances $s^2 < s^2_{0.025} = \sigma^2\chi^2_{0.025}/(N-1)$.

! Note, however, that the inverse statement of the type "*if a sample of size N has the dispersion s^2, then the true variance of the underlying probability distribution lies within limits given by (2.66) with a probability of* 95%" is a false one (this also refers to other tests using confidence limits methodology; see (2.31) and Example 2.1)[12]. The pitfall is that if we are looking at multiple observations of sequences of size N, each of these sequences is characterized by *its own* value of s^2, which may lead to very different limits for σ^2 in (2.66). Our statements about σ lying within these limits (*not the limits characterizing this particular sample, but the limits that are changing from one realization to the other!*) will be true in 95% of the cases.

Suppose that we have two samples of sizes N_1 and N_2, with sample variances s_1^2 and s_2^2. If now $s^2_{1,\,0.025} > s^2_{2,\,0.975}$ or $s^2_{2,\,0.025} > s^2_{1,\,0.975}$ then we can say that the null hypothesis about the samples coming from the same normally distributed population with zero mean can be rejected at the 95% confidence level. Note that the unknown value of σ^2 conveniently drops out of the above inequalities. We consider next the test that allows one to make statements about whether two samples' variances are different (given, once again, the null hypothesis that both samples are drawn from the same normally distributed population) by studying the ratio of sample variances.

F-test for significantly different variances. The F distribution is naturally related to the χ^2 distribution. If s_1^2 and s_2^2 are sample variances of two independent random samples of

[11] The number of degrees of freedom is the number of independent samples N minus the number of parameters in the statistic which must be estimated. For example, in the t statistic, true mean must be estimated based on N independent samples of data, so that the number of degrees of freedom $\nu = N-1$. Similarly, for the χ^2 statistic, we need to estimate true variance and, once again, $\nu = N-1$.

[12] See also discussion in Section 2.7.5.

size N_1 and N_2, drawn from the same normal population with zero mean, then the statistic

$$F(\nu_1,\, \nu_2) = \frac{s_1^2}{s_2^2} = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \tag{2.67}$$

is F-distributed; $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$ are the numerator and denominator degrees of freedom, respectively. We won't list here the formula for the F distribution's PDF; an example of F distribution is however shown in Fig. 2.8. The values of F statistic that are much larger or much smaller than one indicate significant differences in terms of two samples' variances, and the confidence levels can be assigned in a usual fashion using F-distribution c.d.f.

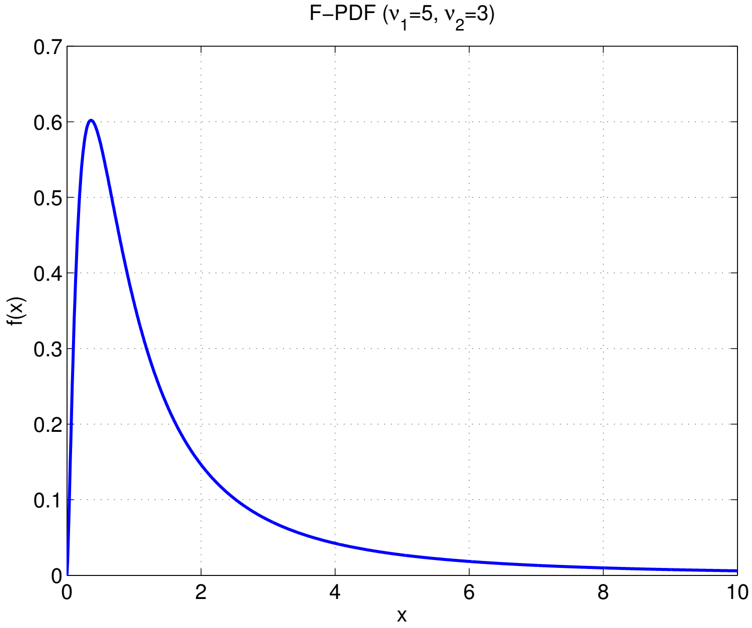


Figure 2.8: F-PDF $f(x)$ for $\nu_1 = 5$ and $\nu_2 = 3$.

The F statistic is very useful in analysis of regression (see Chapter 3) and in testing the significance of spectral peaks (see Chapter 6).

2.7 Statistical inference

2.7.1 Hypothesis testing procedure. Monte Carlo testing

The analysis of statistical data as outlined in preceding sections has proceeded as follows:

(i) from n observations $x_1, x_2, \ldots, x_n$ the value of a function (*statistic*) $F(x_1, x_2, \ldots, x_n)$ was derived (for example, t or χ^2); then

(ii) the distribution of F was computed subject to some assumptions (*null hypothesis*) about the observations (e.g., in Student's t-test, all observations are assumed to be independent and to come from a normally distributed population with known mean θ and unknown variance); and, finally,

(iii) the observed and theoretically computed F-values were compared with the purpose of falsifying (*rejecting*) the null hypothesis.

$\boxed{!}$ Note that if a statistic falls in a *reasonable* part of the distribution, it does not mean that the hypothesis has been "verified" or "proved." The hypothesis can, however, be substantiated by ruling out, statistically, a whole set of competing hypotheses.

Statistical significance testing thus consists of five steps *which should be followed in order* (D. Hartmann; see Example 2.1):

- State the significance level
- State the null hypothesis H_0 and its alternative H_1
- State the statistic used
- State the critical region
- Evaluate the statistic and state the conclusion

Significance level. The acceptable level of uncertainty people usually choose is 95%, in which case there is a 5% chance of accepting the null hypothesis wrongly – a *type II error* (*type I error* is when the correct null hypothesis is rejected).

Null hypothesis. Construction of the null hypothesis and its alternative is critical to the meaning of statistical significance testing — one must 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 conventional wisdom or a zero information conclusion, while its alternative is an "interesting" conclusion that follows directly and uniquely from the rejection of the null hypothesis. Typical examples of H_0 and H_1 hypotheses follow[13]:

(1) H_0: The means of two samples are equal (Chapter 2)
H_1: The means of two samples are different

(2) H_0: The correlation coefficient between two samples is zero (see Chapter 3)
H_1: There is a nonzero correlation between the two samples

(3) H_0: The variance associated with the leading EOF is less than or equal to that associated with the second EOF (see Chapter 4)
H_1: The variance of the leading EOF exceeds that of the second EOF

(4) H_0: The estimated probability density at a certain point of the phase space is less than or equal to that associated with a linear stochastic process (see Chapter 5)
H_1: The PDF at this point exceeds the "linear" PDF

(5) H_0: The variance at a certain period is less than or equal to that of the red-noise background spectrum (see Chapter 6)
H_1: The variance at this period exceeds the red-noise background

Parameter estimation and Monte Carlo tests. Note that we are always comparing statistical characteristics of a given data set with those of a hypothetical data set which is assumed to be drawn from a population with some known properties. In fact, it is only very rarely that one has knowledge of the parameters of underlying distributions, so we always have to *estimate* required parameters[14]. For example, we divide a long time series (say, a set of daily values of temperature at a certain location) into shorter intervals and ask whether one of these intervals is characterized by truly larger values of temperature compared to other parts of the time series. In this case, it might be appropriate to assume that the data

[13] In all cases it is assumed, in addition, that the sample(s) consist(s) of a number of *statistically independent* realizations (see Section 2.7.2); if the sample(s) is(are) small, one may also want to use normality assumption or revert to nonparametric tests — these additional items implicitly enter every statistical significance test.

[14] The goal of the most general inference problem is to make quantitative probabilistic statements about the distribution of a random variable given an observed finite sample; see Section 2.7.4.

set is drawn from a normally-distributed population and estimate the expected mean and variance of this normal distribution from the entire (long) sample's average and dispersion.

The latter estimates turn out to also be the *maximum likelihood estimates* (MLE) for the true mean and true variance given a finite sample of normally distributed data. To obtain MLE we seek, given an assumed distribution (Gaussian in our present case), to maximize the likelihood function. The likelihood function has the same form as the normal probability density function (2.21), but the roles of the variables are reversed. For the PDF, the parameters θ and σ are known constants and the variable is X. For the likelihood function, the sample values (the X's) are already observed. So they are the fixed constants, while the unknown parameters play the role of the variables. MLE involves calculating the values of the parameters that are associated with the highest likelihood given the particular set of data (see Section 2.7.5 and Chapter 3 for further detail).

In our latter example the assumption of Gaussian distribution might have been a reasonable one. What if instead of a temperature data set, we consider a rainfall data set? Since the latter set has, by definition, only positive values which, in addition, are not distributed symmetrically about their mean, Gaussian distribution may not be a proper one to use for the description of this data set. One can use instead the Gamma distribution (2.49) of Section 2.5.4, and compute MLE estimates of its parameters a and b.

In other cases, it is easier to compactly represent the data not in terms of the probability distribution function, but rather in terms of a model that predicts the system's future evolution given the knowledge of the system's history and an estimate of the intrinsic noise. For example, we have a gridded multidimensional data set, for which PDF estimation is not feasible (even if we new what type of PDF to use). A possible solution is to reduce the dimension of the data set by Empirical Orthogonal Function (EOF) analysis [also referred to as Principle Component Analysis (PCA)] (Chapter 4) and then apply a multiple linear regression (MLR) to connect the reduced-state vector's time derivative with the state itself and compute the parameters of this model's stochastic forcing (Chapter 3).

As soon as we have associated our data set with some distribution or with some model, we can in fact estimate statistical significance using Monte Carlo methods (see discussion in Exercise 7 of this chapter, as well as Chapter 3). For the present example, in which we would like to establish if the mean of a particular subsample is different from the full sample's mean, we would generate many (typically 1000) surrogate data sets of the size of our original subsample by either drawing these surrogate subsamples from the population with our estimated PDF or performing multiple integrations of our stochastic model; the mean of each surrogate subsample would then be computed and stored. The 1000 surrogate mean values so obtained should be sorted in the ascending order; we then assign 95% probability to the event that the mean of our actual subsample should be confined by the values of 25-th and 975-th sorted surrogate means, provided the actual subsample is indeed drawn from the distribution we have modeled; otherwise, our null hypothesis gets rejected.

In Monte Carlo significance testing, step (ii) of the hypothesis testing procedure is performed numerically, rather than analytically, while steps (i) and (iii) remain the same. Equivalently, the null hypothesis now involves construction of a distribution-based or trajectory-based model and subsequent multiple integrations of this model to determine the boundaries of the critical region.

2.7.2 Degrees of freedom

We have already discussed the issue of degrees of freedom (see Sections 1.2.2, 2.3.3, and 2.6). This number is formally the number of independent measurements of the quantity or event of interest that is included in the sample. It is sometimes difficult to assess the number of independent realizations that we have in our sample, since the answer may depend on the time and space scale of the phenomenon of interest: in other words, geophysical data sets are typically characterized by a large spatial and temporal *correlation* (see Chapters 3 and 4). An example is given in Section 1.2.2.

There is a number of techniques to estimate the number of degrees of freedom in a sample of spatiotemporal data (Leith 1973; Bretherton et al. 1999). We will discuss these methods after reviewing some background material in regression (Chapter 3), matrix methods (Chapter 4) and time series analysis (Chapter 6).

As a side note, it can be mentioned that the issue of degrees of freedom is safely avoided in many cases that use trajectory-based Monte-Carlo simulations as a part of statistical significance estimation (see Section 2.7.1). This happens because the spatial and temporal correlations within the data set are typically explicitly allowed for in the process of constructing the model that mimics the data set (and in this model itself; these statements will become clearer when we will have considered the data modeling strategies in Chapter 3). !

2.7.3 *A priori* and *a posteriori* significance tests

Another concept in significance testing that is often a source of confusion has to do with the concept of *a priori* and *a posteriori* statistical significance. Let us illustrate this concept with the following example (once again, due to D. Hartmann).

Example 2.5 (A posteriori problem) *Suppose we want to test whether there is any day in December during which it rains more in Seattle than in any other December day. We use December rainfall data for the past 120 years. The daily precipitation amounts are to a good approximation uncorrelated from day to day, so we actually have 120 independent data points for each day in December. Furthermore, the standard deviations for each day (computed over 120 available points) are similar, so we can use the grand mean standard deviation (computed over* 120×31 *available data points) for our statistical significance testing. Our problem is thus to compare the mean for each day with the grand mean for all the days to see if any day stands out.*

We find that the mean for December 15 exceeds the grand mean of all days in the month sufficiently to pass 99% *confidence level (Can you say what would be the difference between the two means in units of the standard deviation?). Does this mean that there must be actual dynamical reasons for the December 15 being the rainiest day of December?*

Let's see. Suppose our desired significance level is 99%. Our null hypothesis is that all the days are independent and drawn from the population with the same mean and standard deviation. The probability $p(31)$ that mean precipitation (over 120 years) for *none* of the 31 days will exceed a certain threshold value is thus $p(31) = 99\%$. How do we compute this threshold value? Let's call $p(1)$ the probability that the precipitation for a *single* day out of 31 will not exceed our threshold. Since precipitation amounts for all days are independent and are assumed to be drawn from the same population, we have $0.99 = p(31) = \{p(1)\}^{31}$ [see (1.10) and Example 1.2], or $p(1) = \sqrt[31]{0.99} \approx 0.999677$. This result means that *in order to ensure that one of the days really stands out of the rest of days in terms of precipitation with the probability of* 99%, *the mean precipitation for this day must exceed the threshold corresponding to* 99.9677% *significance level* (What, in this case, would be the difference between the mean for this day and the grand mean over all days in terms of the standard deviation?). By comparison, if the threshold value for a single-day precipitation is chosen to correspond to the 99% significance level, the actual probability that this day is special is $0.99^{31} \approx 0.73$ or 73%, which is not a very high chance by usual standards. The probability that December 15 stands out is, therefore, not too impressive.

In the example above, we had no *a priori* reason to assume that December 15 is special. Therefore, to estimate the chance that each of the 31 days of December is represented by an independent sample drawn from the same population, we had to take the probability of one event exceeding the criterion, 99%, and raise it to the power equal to the number of independent chances (31) we have given the events to exceed this probability — this is called *a posteriori* statistical analysis. But what if we really had a reason to assume that December 15 is special? Let us say, we got to know that aliens have been seeding the clouds in Seattle on December 15 for the past 120 years and our theoretical calculations predict the seeding should have a significant effect? — In this case, *a priori* significance testing is appropriate and can be used in support of our theory.

2.7.4 General inference problem. Bayes problem. Bayes theorem

General inference problem. Let us come back to a discussion in Section 2.7.1 and take up on the issue of the true parameters of the probability distribution (which presumably underlies our finite sample of data) being actually unknown. Consider, for simplicity, the case of one unknown parameter; for example, we have a sample of size n — $(x_1, x_2, \dots, x_n)$ — which we assume to come from a normally distributed population with known variance σ and unknown mean θ. Let us call x our sample's average. The conditional probability density $p_n(x \mid \theta)$ (whose integrals over some interval of $x^{(1)} \le x \le x^{(2)}$ represent probability of the sample mean x to lie within the interval $[x^{(1)}, x^{(2)}]$) given the value of θ is thus

$$p_n(x \mid \theta) = \frac{1}{\sigma}\sqrt{\frac{n}{2\pi}} \exp\left\{-\frac{n}{2}\left[\frac{x-\theta}{\sigma}\right]^2\right\}. \tag{2.68}$$

The general inference problem is formulated as follows: *Given the function $p_n(x \mid \theta)$ and an observed value of x (NB! both $p_n(x \mid \theta)$ and x are based on n observations), find, for each interval T, the chance $Q_n(T)$ that θ falls in T:*

$$Q_n(T) = P\{\theta \in T\}. \tag{2.69}$$

Bayes problem. In order to find the answer to the inference problem, one has to realize that this answer must depend not only on the observed value of x and conditional probability $p_n(x \mid \theta)$, but also on the function $p_0(\theta)$, which is called *a priori chance* or *overall chance* of an θ value. The quantity $p_0(\theta)$ defines the probability that our object of experimentation (subsequently subjected to n trials) is indeed characterized by the distribution with the parameter θ. For example, we take daily-mean temperature samples of size $n = 100$; in order to solve the inference problem, we have to know what the overall distribution of θ (true mean of random temperature variable) is. In other words, we allow for the fact that different samples of temperature might come from the distributions having different values of θ: we characterize this by counting hypothetical number of cases $N_k(N)$ in which our randomly chosen sample would have a value of θ in the interval $\theta \pm d\theta$, divided by the total number of cases N in the limit $N \to \infty$ and call this number (in the continuous case taking also the limit $d\theta \to 0$) $p_0(\theta)$. If we do know a priori chance of θ — $p_0(\theta)$, then an observation of the sample mean will give us additional information so that we could say more about the values of θ by computing *a posteriori* chance of θ. **Of course, the terms *a priori* and *a posteriori* here have nothing to do with the concepts outlined in Section 2.7.3**: they just refer to *different* probabilities of an θ value — namely, (i) the one in the absence of (prior to) observations (a priori chance); and (ii) the one after the observed value of x is available (a posteriori chance).

The necessity of knowing $p_0(\theta)$ was first recognized by Thomas Bayes (1763) and the problem formulated above is also known as **Bayes problem**. Let us derive the solution of Bayes problem in the case of discrete events. Denote events as E_1, E_2, ... , E_K and assume that the events (i) have positive probabilities; (ii) are mutually exclusive; and (iii) define all possibilities (sum of events' probabilities is equal to one). This is illustrated in Venn diagram of Fig. 2.9 for $K = 4$ (the rectangle is assumed to have a unit area).

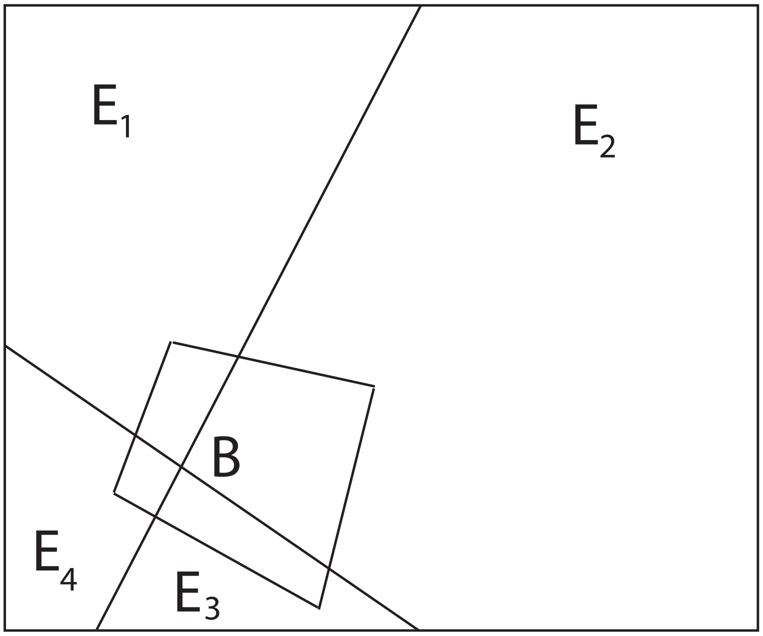


Figure 2.9: Illustration to the solution of the Bayes problem (see text).

If, more specifically, the four events refer to probabilities of four possible values of θ, then the corresponding discrete distribution is our $P_0(\theta)$ — an *a priori* chance of θ. Now, consider an event B, which is also defined on the same set of events (that is, contains a statement about the value of θ). What is the conditional probability of E_i, given B has occurred? In the basic example of the present section, the event B is "the sample average has a certain value." The answer to the latter question, as easily seen from Fig. 2.9, is

$$P(E_i \,|\, B) = \frac{P(B \,|\, E_i)P(E_i)}{\sum\limits_{k=1}^{K} P(B \,|\, E_k)P(E_k)}, \tag{2.70}$$

since the conditional probability of E_k given B equals to the (area of) intersection of E_k and B, divided by the area of B. Note that *a posteriori* chance of θ (provided that we interpret events in terms of possible values of θ) given by (2.70) can be distributed very differently from the *a priori* chance; this distribution will depend on the shape and location of B in the diagram.

Bayes theorem. Returning to the continuous case and using notations of the general inference problem (2.69), we write the solution of this problem as

$$Q_n(T) \equiv P(\theta \in T) = \frac{\int_{(T)} p_n(x \,|\, \theta) p_0(\theta)\, d\theta}{\int p_n(x \,|\, \theta) p_0(\theta)\, d\theta}, \tag{2.71}$$

where the integral in the denominator is extended over all values of θ for which $p_0(\theta)$ is different from zero. If the latter (*a priori*) density is assumed to be constant, the solution becomes

$$Q'_n(T) = \frac{\int_{(T)} p_n(x \,|\, \theta)\, d\theta}{\int p_n(x \,|\, \theta)\, d\theta}, \quad \text{if} \quad p_0(\theta) = \text{const.} \tag{2.72}$$

It is fairly easy to prove (see Von Mises 1964) that under certain (not very restrictive) conditions on $p_0(\theta)$, $Q'_n(T) \to Q_n(T)$ as $n \to \infty$: *The inferred chance $Q_n(T)$ approaches, with increasing n, the value $Q'_n(T)$ which holds for* $p_0 = \text{const}$.

The crucial condition under which the above result is valid has to do with the property of $p_n(x \,|\, \theta)$ *condensing* as $n \to \infty$, that is, for a fixed x, the density $p_n(x \,|\, \theta)$ becomes more and more confined in a neighborhood of some point θ_x as n becomes larger and larger. For example, the distribution (2.68) condenses at the point $\theta_x = x$. Now, let us define interval T as a neighborhood of our conditional probability's condensation point $p_n(x \,|\, \theta)$. As $n \to \infty$, the integrals in the numerator and denominator of (2.71) tend to the same value (since the integration outside the condensation region does not contribute, increasingly with n, to the value of both integrals) and $Q_n(T) \to 1$ as $n \to \infty$ no matter how small neighborhood of θ_x we have chosen. For the same reasons, $Q'_n(T) \to 1$ as $n \to \infty$, and, therefore, $Q_n(T) \to Q'_n(T)$.

The result that $Q_n(T) \to 1$ as $n \to \infty$ when applied to the Bernoulli problem [repeated alternatives with (now unknown) probabilities of "success" p and "failure" $q = 1 - p$; see Section 2.4.1] can be proven under even less restrictive conditions [$p_0(p)$ must be bounded, continuous and be nonzero at the point $p = \theta$, where $\theta = n_1/n$; n_1 is a number of successes in n trials]. In this case it is also known as **Bayes theorem**:

Theorem 3 (Bayes theorem) *The chance, inferred from n trials with $n_1 = \theta n$ successes for the fact that the probability of a single success p lies in the interval $\theta - d\theta < p < \theta + d\theta$ tends toward unity as as n increases indefinitely, no matter how small $d\theta$ is.*

The above results imply, in general, that as a number of trials (sample size) becomes large, inferences about the underlying distribution parameter(s) θ (θ may also be a vector) can be made from the sample averages, without the knowledge and irrespective of a priori probability $p_0(\theta)$. On the other hand, *no inferences can be made from a small number of observations unless something is known about the a priori probability $p_0(\theta)$.* Example: suppose that we know, for a given region T and the region outside of T, which we denote by $\bar{T}$, that the $\min\{p_0(\theta \in T)\} = m$ and $\max\{p_0(\theta \in \bar{T})\} = M$. We write

$$\frac{1}{Q_n(T)} = \frac{\int_{(T)} \cdots + \int_{(\bar{T})} \cdots}{\int_{(T)} \cdots} \leq 1 + \frac{MQ'_n(\bar{T})}{mQ'_n(T)},$$

or, since $Q'_n(T) + Q'_n(\bar{T}) = 1$,

$$Q_n(T) \geq \frac{Q'_n(T)}{Q'_n(T) + \frac{M}{m}[1 - Q'_n(T)]}. \tag{2.73}$$

In particular, if $M \leq m$, then $Q_n(T) \geq Q'_n(T)$.

2.7.5 Re-examination of the method of confidence intervals

How to reconcile the statement of the preceding section (about impossibility of inference from a small number of observations without the knowledge of a priori distribution of parameter(s) of interest) with the hypothesis testing procedure, in which one talks about pre-specified probability of a certain parameter to have a value in a certain range? There is nothing in the hypothesis testing that restricts the samples to be large enough to avoid the influence of a priori distribution (we might want to track the number of degrees of freedom in t test and χ^2 test, but we do not have to *assume* that this number is large).

We have already made cautionary notes about interpretation of confidence intervals as statements about the distribution parameters in Sections 2.3.2 and 2.6.3[15]. Let us now look in more detail into how the confidence limits in the hypothesis testing can be used to formulate statements about *unknown* distribution parameters; we will see that these statements do not really depend on the a priori distribution of these parameters (a good thing!), while the high success chance in the method of confidence intervals is reached at the expense of freedom in formulating the contentions about the parameter lying within an interval of values (in particular, it will turn out that we have really no control in specifying this interval — not a very good thing, in principle).

Let us refer to Fig. 2.10. In this section, we, once again, call θ the parameter on which the distribution of the quantity x depends. We thus assume that the conditional distribution $p_n(x\,|\,\theta)$ is known, while the overall distribution $p_0(\theta)$ is *unknown.* The chance density for the occurrence of definite x and definite θ is $p_n(x\,|\,\theta)p_0(\theta)$. The total range of possible x- and θ-values is indicated in Fig. 2.10 as the rectangle $ABCD$. Consider within this rectangle some domain β. If we conduct an infinite sequence of experiments using n trials or observations involving n independent samples, the outcome of each experiment is

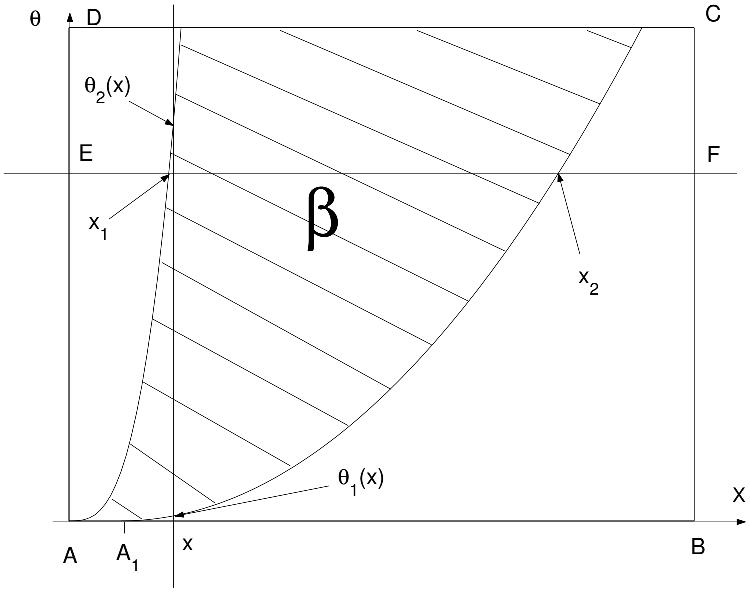


Figure 2.10: Method of confidence intervals (see text). Reproduced from Von Mises (1964).

[15] No paradox arises, however, if we interpret hypothesis tests as the statements about the sample quantities (averages, dispersion etc.) given the hypothesis that the sample comes from a distribution with *known* parameters.

represented by a point in $(x,\,\theta)$-plane. The limiting frequency of this point falling into the region β is

$$P(\beta) = \int\int_{(\beta)} p_n(x \mid \theta) p_0(\theta)\, dx\, d\theta. \tag{2.74}$$

The quantity $P(\beta)$ can, in general, be computed if $p_0(\theta)$ is known. There exists, however, a special region within $ABCD$, whose $P(\beta)$ can be found *independently of any knowledge or assumptions about $p_0(\theta)$.* On a straight line EF parallel to x-axis (fixed θ), the integral $\int\limits_E^F p_n(x \mid \theta)\, dx$ has, by definition, the value of one. Given the quantity $\alpha < 1$, we can, therefore, find some smaller interval from $x_1(\theta)$ to $x_2(\theta)$, for which

$$\int\limits_{x_1(\theta)}^{x_2(\theta)} p_n(x \mid \theta)\, dx = \alpha. \tag{2.75}$$

The locus of the points $x_1(\theta)$ and $x_2(\theta)$ for all θ define two curves: we choose the former curve (associated with x_1) to start at point A (Fig. 2.10), while the latter curve to end in C and call the region between these two curves α-belt β_α. Substituting (2.75) into (2.74), and noting that $\int\limits_A^D p_0(\theta)\, d\theta = 1$, we find that $P(\beta_\alpha) = \alpha$.

Thus, *for any prescribed $\alpha < 1$, an α-belt β_α can be found for which the chance $P(\beta_\alpha)$ has the value α [that is, if we conduct, once again, a series of experiments (each involving n trials), a fraction α of these experiments will result in $(x,\,\theta)$-values belonging to β_α, as the number of experiments increases indefinitely].* How can we use the belt to formulate statements about θ? Suppose that in a single set of n experiments, we have observed a certain value of x. Let us draw, in Fig. 2.10, a vertical line with the abscissa x; this line will intersect the boundaries of the belt in two points with ordinates $\theta_1(x)$ and $\theta_2(x)$. The statement "θ lies between $\theta_1(x)$ and $\theta_2(x)$" is equivalent to the statement "$(x,\,\theta)$-point belongs to the belt β."

Therefore, *if, in a series of experiment (each experiment consists of n observations, from which we derive the quantity x), we pronounce, following each experiment, the contention that θ lies in the interval $\theta_1(x)$ to $\theta_2(x)$, where $\theta_1(x)$ is the smallest and $\theta_2(x)$ is the largest θ-value in the α-belt with abscissa x, we have the chance α of being right.* In other words, if $\alpha = 0.9$, and we process the results of a large number of experiments, our statements about θ belonging to the interval $\theta_1(x)$ to $\theta_2(x)$ (the value of x is an outcome of a given experiment) will be right in 90% of the cases. We can thus make inference statements, whose chance of success will be as high as we wish. Note, however, that, as advertised in the beginning of this section, the downside of the high success rate in the method of confidence intervals is the fact that the interval $[\theta_1(x),\, \theta_2(x)]$ is not pre-specified: it depends on x and is thus changing from one experiment to the other. The inequalities restricting the range of parameter in the hypothesis testing [(2.31), Example 2.1, (2.66)] should be understood in this narrow sense only.

$\boxed{!}$

We have seen that the method of confidence intervals is in some sense a restricted form of parameter inference. Let us also comment here on the maximum likelihood estimation (MLE; see Section 2.7), in which one is concerned with obtaining the "best" estimate of θ given an observed value of x. The quantity $p_n(x,\, \theta)p_0(\theta)$ is proportional to the a posteriori chance of θ. Therefore, the value of θ, which *makes this product maximum* will be correct (in the long run, after many experiments are completed and documented) in more of the cases in which this particular value of x has been observed, than any other θ value. The problem is, once again, that a priori chance of θ, $p_0(\theta)$ is unknown, so what is traditionally referred to as the maximum likelihood estimate of θ is computed under the assumption $p_0(\theta) = \text{const}$ and is thus defined by the equation

$$\frac{\partial}{\partial\theta} p_n(x,\, \theta) = 0.$$

If x is the average of n observations and if θ is the theoretical mean value of x, the function $p_n(x,\, \theta)$ shows a property of condensation as n increases: the values of $p_n(x,\, \theta)$ out of immediate neighborhood of θ become very small and negligible [see (2.19) of Section 2.1]. Furthermore, in this case, the a posteriori distribution Q_n becomes more and more independent of $p_0(\theta)$. These properties make the MLE estimate of the population's mean θ, derived based on the sample's average x, the one that has, approximately, the greatest chance to be correct. Further discussion and examples of maximum likelihood estimation will be given in Chapters 3, 5, and 6.

Let us connect the ideas developed in this section to the normal sampling significance testing procedures. Namely, we consider two examples covering the cases described in Sections 2.3.2 and 2.3.3.

Confidence intervals on the mean of a normally distributed population inferred from a finite sample: Case of known variance. Consider the problem of determining the true mean θ given a finite sample's average value x. We will assume that the sample of size n is drawn from a normal distribution with a known variance σ; x is distributed according to (2.68). The range of possible $(x,\, \theta)$-values covers, in this case, the whole $(x,\, \theta)$-plane (see Fig. 2.11).

The α-belt β_α in this case is the strip limited by two parallels to the bisectrix of the axes with the half width $OA = \xi$ determined by the following equation:

$$\alpha = \Theta(\sqrt{\frac{n}{2}}\frac{\xi}{\sigma}), \tag{2.76}$$

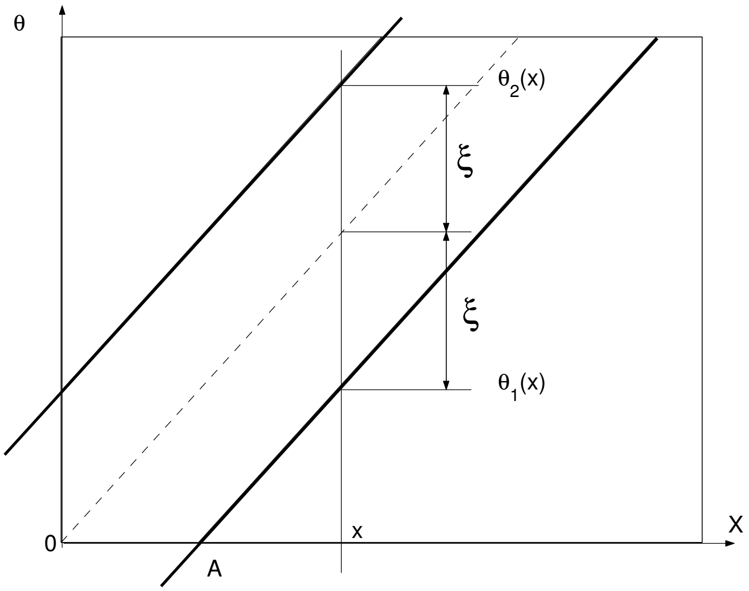


Figure 2.11: Confidence intervals based on a sample from a normally distributed population with a known variance. Reproduced from Von Mises (1964).

where

$$\Theta(u) = \frac{2}{\sqrt{\pi}} \int\limits_0^u \mathrm{e}^{-u^2}\, du \tag{2.77}$$

is the *probability integral.* It is easily seen that the equations (2.76) and (2.77) are equivalent to the condition (2.75) defining the α-belt β_α, provided $p_n(x\,|\,\theta)$ is given by (2.68). GIven a sample's average x, the limits in θ are (see Fig. 2.11) $\theta_1(x) = x - \xi$ and $\theta_2(x) = x + \xi$.

Therefore, *if N independent samples of size n, drawn from a normally distributed population with unknown mean θ and known variance σ are considered, and N sample averages $x^{(k)}$, $1 \le k \le N$ are computed, then the contention that the true mean θ lies between $x^{(k)} - \xi$ and $x^{(k)} + \xi$, where ξ computed from (2.76), will be correct, in the long run ($N \to \infty$) in αN out of N cases.*

Variance unknown: Student's t test. Consider the previous example, but drop an unrealistic assumption that σ is fixed and known. We thus end up with two unknown parameters θ and σ, and would like to make statements about the true mean θ given an observation of a sample's average x and dispersion s^2.

To do so, we have to generalize definition of the α-belt to include multivariate case. The conditional probability $p_n(x,\, s \,|\, \theta,\, \sigma)$ is now defined in a four-dimensional space. A region in this space can be defined by means of a function $F(x,\, s,\, \theta,\, \sigma)$; namely, let us define region β as the collection of points for which some function $F(x,\, s,\, \theta,\, \sigma) < 0$. The F defining the β_α is chosen so that for each pair of constant θ and σ

$$\int\int_{(\beta:\, F<0)} p_n(x,\, s \,|\, \theta,\, \sigma)\, dx\, ds = \alpha; \quad \alpha < 1. \tag{2.78}$$

Given (2.78), it is straightforward to show that the chance of a point $x,\, s,\, \theta,\, \sigma$ falling in β_α is equal to α [the double integral (2.74) for the case of one unknown parameter is substituted by a quadruple integral in the present case]. Given the observed values of x and s, the inequality $F(x,\, s,\, \theta,\, \sigma) < 0$ thus gives an estimate of the ranges of θ and σ that will be correct, if multiple samples with their respective $(x,\, s)$-pairs are considered, in α-fraction of all cases.

The joint distribution of the sample's average x and dispersion s^2 under the assumption that the sample is drawn from a normal population with the mean θ and variance σ^2 can be shown (see Von Mises 1964) to equal to

$$p_n(x,\, s \,|\, \theta,\, \sigma) = \text{const} \cdot \mathrm{e}^{-n[s^2+(x-\theta)^2]/2\sigma^2} s^{n-2}, \tag{2.79}$$

while the appropriate choice of F is

$$F(x,\, s,\, \theta,\, \sigma) = (n-1)\left(\frac{x-\theta}{s}\right)^2 - t_\alpha^2 \equiv t^2 - t_\alpha^2, \tag{2.80}$$

where t is the Student's t-ratio [see (2.33) of Section 2.3.3)], while t_α is found from

$$\alpha = \int\limits_{-t_\alpha}^{t_\alpha} f(t)\, dt. \tag{2.81}$$

Here $f(t)$ is the PDF of Student's t distribution with $n-1$ degrees of freedom — Eq. (2.34) of Section 2.3.3.

The points in $(x,\, s)$-plane for which F determined by (2.80) is less than zero fill the sector (see Fig. 2.12) between the two straight lines AB and AC which intersect the x-axis in $x = \theta$ and form with the vertical AD the angle ϕ: $\tan\phi = t_\alpha/\sqrt{n-1}$. If t_α is given by (2.81), the probability of the point $(x,\, s,\, \theta,\, \sigma)$ falling in this sector (limiting frequency of samples with the appropriate values of $x,\, s,\, \theta,\, \sigma$) is equal to α.

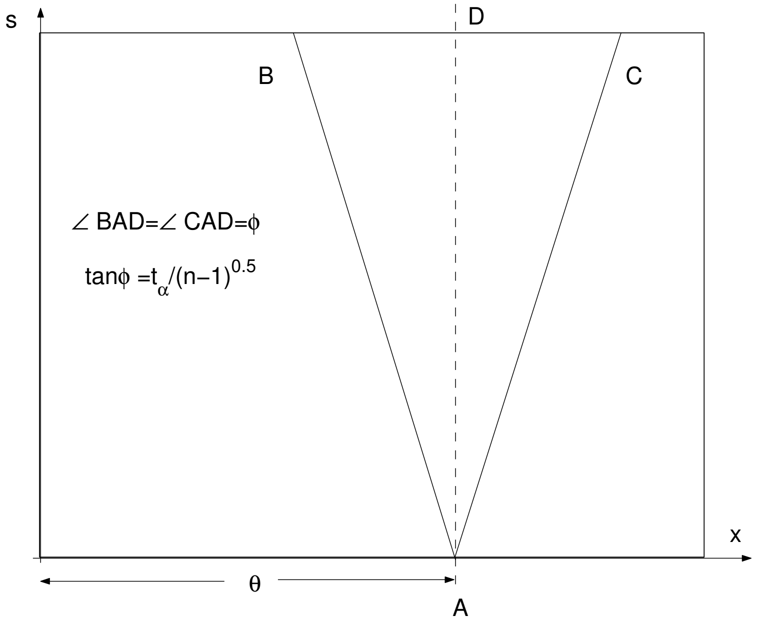


Figure 2.12: Confidence intervals based on a sample from a normally distributed population with unknown variance (Student's t test). Reproduced from Von Mises (1964).

Therefore, in a series of N observed samples of size n, we expect that in αN cases (as $N \to \infty$) the following inequality is valid:

$$(n-1)\left(\frac{x-\theta}{s}\right)^2 \leq t_\alpha^2 \quad \text{or} \quad x - \frac{t_\alpha}{\sqrt{n-1}}s \leq \theta \leq x + \frac{t_\alpha}{\sqrt{n-1}}s. \tag{2.82}$$

The latter equation is identical to the one used in Example 2.1.

2.7.6 Concluding remarks

The present chapter has provided fundamental concepts and ideas pertaining to the probability theory and statistical data analysis. The emphasis has been on the *descriptive* statistics, which explores the properties of given data sets without trying to associate them with a certain *model*. We were mainly concerned with establishing how similar of different two or more sets of data are (are the sample means different? are the sample variances different? are the distributions from which the data sets are presumably drawn different or not? etc.) These questions have been answered using the methodology of *hypothesis testing*, which provides answers in terms of the probability of a certain *event* (e.g., the means of two samples are different) to happen, given some assumptions about the actual distributions underlying the data.

Geophysical data sets are typically characterized by a fairly large number of degrees of freedom (that is, the number of independent measurements), which has profound consequences with respect to their statistical properties. In particular, the importance of *normal (Gaussian)* distribution becomes apparent due to the *central limit theorem*; furthermore, large sample size plays an important role in the problem of *parameter inference*: in this limit, statements about the parameters of underlying distribution can be made based on the value of sample averages, without any a priori information. We have thus been able to introduce the subject of *inferential* statistics: a suite of statistical analysis techniques and procedures that are *model-dependent* (see example of the nonparametric bootstrap method and general description of Monte Carlo procedures for significance testing).

In the following chapters, we will use the techniques described presently in combination with linear (matrix methods: multiple regression, PC analysis) and nonlinear (PDF estimation and cluster analysis) methods to both detect and model potentially predictable *signals* on the background of random (unpredictable) *noise*.

References

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate*, **12**, 1990–2009.

Huff, D., 1954: *How to Lie with Statistics.* Norton and Co., New York, 142pp.

Knight, K., 2000: *Mathematical Statistics. Texts in Statistical Science.* Chapman and Hall/CRC. 481pp.

Larson, R. L., and M. L. Marx, 1986: *An Introduction to Mathematical Statistics and its Applications.* 2nd edition, Prentice–Hall, Englewood Cliffs, N. J., 630pp.

Leith, C. E., 1973: The standard error of time-averaged estimates of climatic means. *J. Appl. Meteorol.*, **12**, 1066–1069.

Mendenhall, W., D. D. Wackerly, and R. L. Sheaffer, 1990: *Mathematical Statistics with Applications.* PWS–Kent, Boston, 688pp.

Panofsky, H. A., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology.* Pennsylvania State University, University Park, 224pp.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1994: *Numerical Recipes.* 2-nd edition. Cambridge University Press, 994 pp.

Von Mises, R., 1964: *Mathematical Theory of Probability and Statistics.* Academic Press, New York.

Von Storch, H., and F. Zwiers, 1999: *Statistical Analysis in Climate Reserach.* Cambridge University Press, Cambridge, United Kingdom, 484pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* International Geophysics Series, v. 59), Academic Press, San Diego, 467pp.

Zwiers, F. W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate*, **8**, 336–351.

Chapter 3

Regression and Theory of Correlation. Modeling of Data

Suppose that we are given a set $\{y_n\}$ $(1 \leq n \leq N)$ of N observations of some quantity y and that each observation is associated with some value of the independent variable x. Consider first the case in which the values $\{x_n\}$ $(1 \leq n \leq N)$ are assumed to be known *exactly*, while each observation y_n is susceptible to random measurement errors; these errors are characterized by the standard deviation σ_n. For example, we are processing a set of simultaneous observations ($\{y_n\}$) of an air pollutant's concentration at N stations, which are located at points with specified coordinates (in one-dimensional case — $\{x_n\}$).

We would like to represent this data set economically by fitting it to a model that relates the *predictand* or *response variable* y to *predictor variable* x and depends on J adjustable parameters a_j $(1 \leq j \leq J)$

$$y = \hat{y}(x;\, a_1,\, a_2,\, \ldots,\, a_J). \tag{3.1}$$

The algebraic form of the model (3.1) is assumed to be known and the problem of finding appropriate values of parameters a_m is referred to as the *regression problem.* Solving the regression problem will enable us, in particular, to infer the values of y for any given x (for example, interpolate or extrapolate irregularly spaced observations onto a regular grid etc.).

A general way of solving the regression problem (3.1) is to design a *merit function* that measures the agreement between the data and the model given a particular set of model parameters. The parameters are then adjusted to achieve a minimum of the merit function, yielding *best-fit parameters.* One of the most widely used choices of the merit function has the quadratic form, resulting in the so called **least-squares fit**:

$$\text{Find } \{a_1,\, a_2,\, \ldots,\, a_J\} \longrightarrow \text{minimize} \sum_{n=1}^{N} [y_n - \hat{y}(x_n;\, a_1,\, a_2,\, \ldots,\, a_J)]^2. \tag{3.2}$$

We thus want to find a set of model parameters that minimizes the sum of the squared differences between the data and our assumed parametric dependence $\hat{y}(x)$.

3.1 Least squares as a maximum likelihood estimator. Chi-square fitting

Well, we have solved (3.2) and obtained our "best-fit" parameters. What are the uncertainties associated with these estimates? How do we know whether our least-squares fit (3.2) is a "good" one or not? In general, what is the connection between the regression problem and probabilistic aspect of data analysis? In answering the above questions, one has to acknowledge the fact that data are in general not exact: they are either subject to measurement errors, or, in the model-generated data, to natural predictability limits rooted in the climate system's nonlinear dynamics. In the latter case, useful deterministic (predictable over a relatively long time scale) relations between two or more variables (say, persistent large-scale flow patterns) are typically masked by shorter-time-scale smaller-spatial-scale variability, which can be treated, on the long time scale of interest, as random noise.

Maximum likelihood estimation. Suppose that the set $\{\Delta y_n\}$ $(1 \le n \le N)$ of normalized deviations $\Delta y_n \equiv (y_n - \hat{y})/\sigma_n$ of each measurement y_n from the "true" model $\hat{y}(x_n;\, a_1,\, a_2,\, \ldots,\, a_J)$[1] is a sample of size N drawn from a population having the standard normal distribution. The probability dP_n of the n-th measurement to fall within an infinitesimal interval of length dy containing y_n is therefore given by the area of a shaded strip in Fig. 3.1:

$$\begin{aligned} dP_n = p_n(y_n)\, dy &= \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{y_n - \hat{y}(x_n;\, a_1,\, a_2,\, \ldots,\, a_J)}{\sigma_n}\right)^2\right] dy \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\Delta y_n^2\right] dy_n', \end{aligned} \tag{3.3}$$

where $dy_n' \equiv dy/\sigma_n$. Since all measurements are independent, the increment of the probabil-

[1] That is, the model with the "correct" parameters $a_1,\, a_2,\, \ldots,\, a_J$.

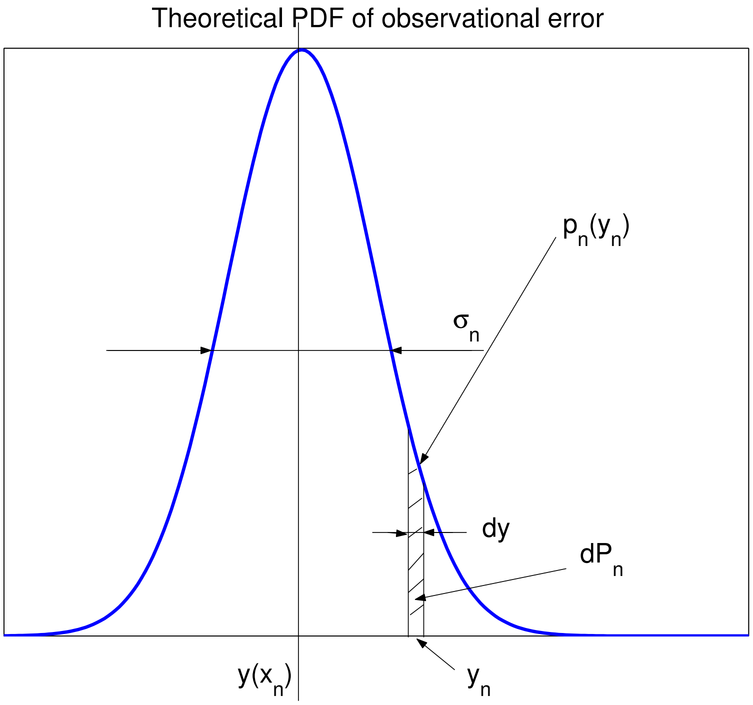


Figure 3.1: Hypothesized distribution of observational error (see text for details).

ity $\mathbf{dP}$ of the entire data set to occur, that is the probability that each of N measurements will fall within the distance dy of its actual observed value y_n $(1 \le n \le N)$, is

$$\begin{aligned}\mathbf{dP} = dP_1\;\ldots,\;dP_N \;&=\; \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{y_n - \hat{y}(x_n;\,a_1,\,a_2,\,\ldots,\,a_J)}{\sigma_n}\right)^2\right] dy \\ &=\; \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\Delta y_n^2\right] dy'_n. \qquad (3.4)\end{aligned}$$

We are interested in finding the set of parameters $\{a_n\}$ $(1 \le n \le N)$, which maximizes the probability (3.4) of our data set to occur. In other words, we are looking for the set of parameters, whose *likelihood* is maximized given our observed data. Maximizing (3.4) is equivalent to maximizing its logarithm, which is, in turn, equivalent to minimizing the negative of its logarithm. Taking the natural logarithm of (3.4) and multiplying the resulting sum by -1, we find that the maximum likelihood estimate of our parameters is

$$\{a_1,\, a_2,\, \ldots,\, a_J\} \longrightarrow \text{minimize} \sum_{n=1}^{N} \Delta y_n^2 \equiv \sum_{n=1}^{N} \frac{[y_n - \hat{y}(x_n;\, a_1,\, a_2,\, \ldots,\, a_J)]^2}{\sigma_n^2}, \tag{3.5}$$

since the other terms in the some are constant and do not depend on our adjustable parameters. Comparing (3.5) and (3.2) we see that least-squares fitting is a maximum likelihood estimation of the fitted parameters *provided* that the measurement errors are (i) independent; (ii) normally distributed; and (iii) have the same standard uncertainty ($\sigma_n = \sigma$).

Chi-square fitting. In a general case of non-equal σ_n's, the merit function to be minimized is given by (3.5) and called *chi-square*:

$$\chi^2 \equiv \sum_{n=1}^{N} \Delta y_n^2 \equiv \sum_{n=1}^{N} \frac{[y_n - \hat{y}(x_n;\, a_1,\, a_2,\, \ldots,\, a_J)]^2}{\sigma_n^2}. \tag{3.6}$$

If our normalized observational errors $\Delta y_n \equiv [y_n - \hat{y}(x_n)]/\sigma_n$ have the standard normal distribution, the quantity χ^2 in (3.6) has the χ^2 distribution with N degrees of freedom (see Section 2.6.3), hence its name.

After we have adjusted our parameters to minimize the value of χ^2, the individual terms in the sum (3.6) are no longer all independent, since in the process of computing the best-fit parameters, we have imposed M additional constraints. The minimum of a quadratic functional (3.6) is achieved if the parameters $\mathbf{a}$ satisfy the system of equations $\partial\chi^2(a_1,\, \ldots\,,\, a_J)/\partial a_m = 0$ $(1 \le j \le J)$, or

$$0 = \sum_{n=1}^{N} \left(\frac{y_n - \hat{y}(x_n)}{\sigma_n}\right) \left(\frac{1}{\sigma_n} \frac{\partial \hat{y}(x_n;\, a_1,\, a_2,\, \ldots,\, a_J)}{\partial a_j}\right); \quad j = 1,\, \ldots\,,\, J. \tag{3.7}$$

Call $\hat{\mathbf{a}} \equiv (\hat{a}_1,\, \ldots\,,\, \hat{a}_J)$ the solution of (3.7) [the set of our best-fit parameters]. Defining the quantities $Y'_{n,\,j} \equiv \frac{1}{\sigma_n}\frac{\partial \hat{y}(x_n;\,\hat{\mathbf{a}})}{\partial a_j}$, where the partial derivatives are computed for $\mathbf{a} = \hat{\mathbf{a}}$, we see that the normalized observational errors $\Delta y_n \equiv [y_n - \hat{y}(x_n)]/\sigma_n$ satisfy J constraints

$$0 = \sum_{n=1}^{N} \Delta y_n Y'_{n,\,j}; \quad j = 1,\, \ldots\,,\, J. \tag{3.8}$$

Since $Y'_{n,\,j}$ is just the set of known numbers, we see that any set of J of the Δy_n's can be linearly expressed in terms of other $N - J$ Δy_n's; we are thus left with $N - J$ independent measurements only. It turns out that **for the models that are *linear* in $\mathbf{a} \equiv (a_1, \ldots, a_J)$, the probability distribution of the function χ^2 *at its minimum* $\hat{\mathbf{a}}$ is the χ^2 distribution with $\mathbf{N - J}$ degrees of freedom**[2]**.**

The latter property gives us means to estimate the *goodness of fit*, by computing the expected fraction of cases in which the sum of squares of $N - J$ standard normal random variables will exceed our observed χ^2, that is, the probability of exceeding the value of χ^2 by chance (see Section 2.6.3). Small values of this probability (say 1%) indicate in general that our χ^2 is unlikely to be large simply due to unfortunate sampling, but rather one of the following three possibilities takes place:

- the model is wrong and can be statistically rejected (see more discussion of this possibility below);
- the estimates of the measurement errors σ_n were wrong (the errors are in fact larger than was stated, so that the χ^2 sum is in fact smaller); conversely, if the measurement errors have been overestimated, the fit might appear to be unrealistically good;
- the measurement errors are not normally distributed. Non-normal distributions are typically characterized by longer tails — that is, they generate a larger fraction of points with large deviations from the mean (see, for example, Section 2.5.3); therefore, given the same standard deviation, the sum of squares of random variables drawn from such distributions will tend to be larger than the sum of the same number of normally distributed variables. The subject of *Robust Statistics* (see the discussion at the end of Section 3.2) deals with cases in which the normal model is a bad approximation. If we know how measurement errors are actually distributed, it is possible to generate synthetic data set via *Monte Carlo* simulations (Section 2.7.1); this will also provide a way to estimate uncertainties of the estimated model parameters (Section 3.6.1).

Exercise 9. How do you expect the value of χ^2 to change (increase, decrease, stay the same) if the measurements in a data set are not independent (compared to the case of the same number of independent measurements)? How would the expected spread of χ^2 values change?

The first possibility from the list above relates to the case in which we are assuming a fixed parametric form of some deterministic relation between our two variables and

[2] This is in general not true for models that are nonlinear in $\mathbf{a}$.

estimate the likelihood of this assumption given the data subjected to measurement errors; the negative result here indicates that the assumed deterministic relation is unlikely. Such formulation is typical for statistical analysis of engineering problems. In geophysics (in particular, in meteorological and oceanographic applications), the relation between two variables is rarely purely deterministic due to presence of instabilities and dynamical noise which masks the signal (useful relationship between the variables) of interest. Furthermore, the amplitude of the noise is typically as large as or larger than the amplitude of the signal. In this situation, the "dynamical" uncertainties (rather than the "measurement uncertainties") are not known in advance. A way to proceed under such circumstances is to assume that all data points have a certain fixed uncertainty $\sigma_n = \sigma$ and also to *assume a good fit*! The model parameters are then estimated by minimizing χ^2 (in which σ is now constant, so that the problem reduces to the standard least-squares fit) and finally, the standard uncertainty is estimated as

$$\sigma^2 = \sum_{n=1}^{N} [y_n - \hat{y}(x_n)]^2/(N - J). \tag{3.9}$$

We can then try to answer a question of how reasonable our model is by using this estimate of uncertainty to fit a different sample of paired data $\{x_n, y_n\}$ (for example, from a different time segment of a numerical [dynamical] model integration) to our [statistical] model, while comparing the estimated parameters of this fit and our original fit (we can now do so, since given the input data uncertainty σ and our new χ^2 fit, we can estimate the uncertainty of our output fitted parameters).

Taking this procedure one step further, we can divide available sample into two arbitrary segments, estimate the uncertainty as above by assuming a good fit on one of the sub-samples (*train our model*) and then try to predict the other sub-sample using this model (*validate our model*). A measure of the model performance (for example, correlation coefficient between the model prediction and actual data; see below) gives an estimate (in some controlled fashion related to the data uncertainty) of how good the model is. Now, the division into training and validation intervals can be done in many possible ways and our measure of the model's goodness can be ensemble averaged (this will reduce the chances that the model performs too well or too poorly due to a particular random sampling). This procedure is called *cross-validation.* We will come back to the problem of statistical forecasting in Section 3.7.

3.2 Fitting data to a straight line. Theory of correlation

Fitting data to a straight line. A didactic example, which is, however, of frequent practical application as well, is fitting a set of N data points $\{x_n,\, y_n\}$ $(1 \le n \le N)$ to a straight line

$$y = bx + a, \tag{3.10}$$

were a and b are unknown coefficients that need to be estimated in an optimal way using available data. This problem is often referred to as *linear regression.* Once again, each x_n is assumed to be known exactly, while each "measurement" y_n associated with x_n has a known standard error σ_n [the quantity $(y_n - a - bx_n)/\sigma_n$ is thus assumed to be a random variable drawn from the standard normal distribution].

The χ^2 merit function (3.6) in this case is given by

$$\chi^2 = \sum_{n=1}^{N} \left(\frac{y_n - a - bx_n}{\sigma_n} \right)^2 . \tag{3.11}$$

If the measurements are indeed normally distributed, than minimizing the expression above will give the maximum likelihood estimate of our linear model's parameters; *otherwise, we will just end up with the straight-line fit that minimizes the weighted distance between this line and our set of points — not necessarily a useless estimate!* To achieve the minimum of χ^2, the parameters a and b must satisfy the following equations [cf. (3.7)]:

$$\begin{aligned} 0 &= \frac{\partial \chi^2}{\partial a} = -2 \sum_{n=1}^{N} \frac{1}{\sigma_n} \frac{y_n - a - bx_n}{\sigma_n}, \\ 0 &= \frac{\partial \chi^2}{\partial b} = -2 \sum_{n=1}^{N} \frac{x_n}{\sigma_n} \frac{y_n - a - bx_n}{\sigma_n}. \end{aligned} \tag{3.12}$$

Expressions (3.12) state that normalized measurement errors $\Delta y_n \equiv (y_n - a - bx_n)/\sigma_n$, upon our adjusting the parameters a and b, are subject to two *linear* constraints, with coefficients $Y'_{n,m}$ $(1 \le n \le N,\ 1 \le m \le 2)$ $[Y'_{n,1} = 1/\sigma_n,\ Y'_{n,2} = x_n/\sigma_n]$ that *do not depend on* a *and* b [cf. (3.8)]. This property of parametric regression models with linear dependence on parameters (the property of independence of additional constraints on the values of fitted parameters) enables one to derive the theoretical distribution of χ^2 — the χ^2 distribution with $N - 2$ degrees of freedom for the two-parameter case of the present section, and with $N - M$ degrees of freedom for a general case of M parameters. For nonlinear models, $Y'_{n,m}$ will depend on parameters and this result is not valid (although in practice the χ^2 distribution is not too bad an assumption even for models that are not linear in their parameters).

Let us denote, for any data set $\{z_n\}$ $1 \leq n \leq N$ the quantity

$$\bar{z} \equiv \frac{\sum\limits_{n=1}^{N} z_n/\sigma_n^2}{\sum\limits_{n=1}^{N} 1/\sigma_n^2}. \tag{3.13a}$$

If $\sigma_n = \text{const}$, $\bar{z}$ in (3.13a) represents the sample's average value. If we are given two data sets $\{x_n\}$ $1 \leq n \leq N$ and $\{y_n\}$ $1 \leq n \leq N$, we can also define, in addition to $\bar{x}$ and $\bar{y}$, the quantities $\overline{x^2}$, $\overline{y^2}$ and $\overline{xy}$ in an analogous way:

$$\overline{x^2} \equiv \frac{\sum\limits_{n=1}^{N} x_n^2/\sigma_n^2}{\sum\limits_{n=1}^{N} 1/\sigma_n^2}, \quad \overline{y^2} \equiv \frac{\sum\limits_{n=1}^{N} y_n^2/\sigma_n^2}{\sum\limits_{n=1}^{N} 1/\sigma_n^2}, \quad \overline{xy} \equiv \frac{\sum\limits_{n=1}^{N} x_n y_n/\sigma_n^2}{\sum\limits_{n=1}^{N} 1/\sigma_n^2} \tag{3.13b}$$

With these notations, the system (3.12) can be written as:

$$\begin{aligned} a + b\bar{x} &= \bar{y} \\ a\bar{x} + b\overline{x^2} &= \overline{xy}. \end{aligned} \tag{3.14}$$

The solution of (3.14) is

$$\begin{aligned} \Delta &\equiv \overline{x^2} - (\bar{x})^2 \\ b &= (\overline{xy} - \bar{x}\bar{y})/\Delta \\ a &= \bar{y} - b\bar{x} = (\overline{x^2}\bar{y} - \bar{x}\overline{xy})/\Delta, \end{aligned} \tag{3.15}$$

and these are the expressions for our best-fit parameters.

Exercise 10. What is the minimum value of the merit functional χ^2? Define the deviations from our weighted averages (3.13a), (3.13b), or *anomalies* x'_n and y'_n, as

$$x'_n = x_n - \bar{x} \text{ and } y'_n = y_n - \bar{y}; \quad 1 \leq n \leq N. \tag{3.16a}$$

Express the quantity $y^*_n \equiv y_n - \hat{y} \equiv y_n - a - bx_n$ via x'_n, y'_n, and b, and show that if (3.12) are satisfied, then

$$\chi^2 \equiv \sum_{n=1}^{N} \frac{y_n^{*2}}{\sigma_n^2} = \sum_{n=1}^{N} \frac{y_n'^2}{\sigma_n^2} - b^2 \sum_{n=1}^{N} \frac{x_n'^2}{\sigma_n^2}. \tag{3.16b}$$

The goodness-of-fit can now be assessed by computing the probability of obtaining, by chance, the value of χ^2 larger (that is, *worse*) than our estimated value, assuming that χ^2 is indeed χ^2 distributed with $N-2$ degrees of freedom.

Let us now compute the standard uncertainty of our estimated parameters. The expressions (3.15) can also be written in the following form:

$$a = \sum_{n=1}^{N} \alpha_n y_n; \quad b = \sum_{n=1}^{N} \beta_n y_n, \tag{3.17}$$

where

$$\alpha_n \equiv \frac{1}{\sum\limits_{n=1}^{N} 1/\sigma_n^2} \frac{(\overline{x^2} - \bar{x}x_n)/\sigma_n^2}{\Delta}; \quad \beta_n \equiv \frac{1}{\sum\limits_{n=1}^{N} 1/\sigma_n^2} \frac{(x_n - \bar{x})/\sigma_n^2}{\Delta}. \tag{3.18}$$

The variances of the coefficients a and b, σ_{a}^2 and σ_{b}^2, respectively, are expressed through the variances of individual observations $\mathrm{Var}\{y_n\} \equiv \sigma_n^2$ as[3]

$$\sigma_{\mathrm{a}}^2 = \sum_{n=1}^{N} \alpha_n^2 \sigma_n^2; \quad \sigma_{\mathrm{b}}^2 = \sum_{n=1}^{N} \beta_n^2 \sigma_n^2. \tag{3.19}$$

Substituting (3.19) into (3.18) and using definitions (3.13a), (3.13b) for $\bar{x}$ and $\overline{x^2}$, as well as the first of equations (3.15) for Δ, we obtain

$$\sigma_{\mathrm{a}}^2 = \frac{1}{\sum\limits_{n=1}^{N} 1/\sigma_n^2} \frac{\overline{x^2}}{\Delta}; \quad \sigma_{\mathrm{b}}^2 = \frac{1}{\sum\limits_{n=1}^{N} 1/\sigma_n^2} \frac{1}{\Delta}. \tag{3.20}$$

Looking at the expressions for the regression coefficients in the form (3.17) we see that even if the individual observations y_n are independent, the coefficients a and b cannot in general be considered as independent random variables, since they are but different linear combinations of all y_n's. To measure the degree of *linear association* between two random variables (in the present case a and b), we can introduce the *covariance* [see (1.19b) of Section 1.3.3] $\mathrm{Cov}\{a,\, b\}$:

$$\mathrm{Cov}\{a,\, b\} \equiv E\{(a - E\{a\})(b - E\{b\})\} = E\{ab\} - E\{a\}E\{b\}. \tag{3.21}$$

[3] The derivation is analogous to that for the variance of the sample's average (Section 2.1 and *Exercise 1*).

The covariance between a and b can be shown to be [derivation is using (3.21), but otherwise is completely analogous to that for (3.19)]

$$\mathrm{Cov}\{a,\, b\} = \sum_{n=1}^{N} \alpha_n \beta_n \sigma_n^2, \tag{3.22}$$

which, upon substitution of expressions (3.18) for α_n and β_n, becomes

$$\mathrm{Cov}\{a,\, b\} = -\frac{1}{\sum\limits_{n=1}^{N} 1/\sigma_n^2} \frac{\bar{x}}{\Delta}. \tag{3.23}$$

Exercise 11. Derive the expressions (3.19), (3.20), (3.22), and (3.23).

! The *coefficient of correlation* r_{ab} between a and b is defined as

$$r_{\mathrm{ab}} \equiv \frac{\mathrm{Cov}\{a,\, b\}}{\sqrt{\mathrm{Var}\{a\}\mathrm{Var}\{b\}}} = -\frac{\bar{x}}{\sqrt{\overline{x^2}}}. \tag{3.24}$$

The correlation coefficient is the number taking values from -1 to 1. A positive value of r_{ab} indicates that the errors in a and b are correlated (are likely to have the same sign), while the opposite is true for the negative value of r_{ab}, in which case the errors in a and b are *anticorrelated* (are likely to have the opposite signs). Zero correlation coefficient signals that the errors in a and b are *linearly independent*. We will return to the sampling theory of correlation in Section 3.3.

Theory of correlation between two finite samples. We can define a measure of linear association between two finite data samples $\{x_n\}$ $1 \le n \le N$ and $\{y_n\}$ $1 \le n \le N$ — the *correlation coefficient* — in the same way we have just introduced the correlation coefficient between two linearly dependent random populations, but using sample averages in place of expectation integrals. The correlation coefficient so defined is intimately connected with the problem of linear regression in the following way.

We first rearrange the formula (3.16b) by dividing both its right- and left-hand sides by $\sum\limits_{n=1}^{N} 1/\sigma_n^2$ and using definitions (3.13b) as

$$\overline{y'^2} = \overline{y^{*2}} + b^2\overline{x'^2}, \tag{3.25a}$$

or, after division by $\overline{y'^2}$,

$$1 = \frac{\overline{y^{*2}}}{\overline{y'^2}} + b^2\frac{\overline{x'^2}}{\overline{y'^2}}. \tag{3.25b}$$

The expression (3.15) for b, written in terms of anomalies x' and y' [see (3.16a)], becomes

$$b = \overline{x'y'}/\overline{x'^2}. \tag{3.26}$$

Substituting (3.26) into (3.25b), we get

$$1 = \frac{\overline{y^{*2}}}{\overline{y'^2}} + \frac{(\overline{x'y'})^2}{\overline{x'^2}\,\overline{y'^2}} = \frac{\overline{y^{*2}}}{\overline{y'^2}} + r^2, \tag{3.27}$$

where we have defined the quantity r as

$$r \equiv \frac{\overline{x'y'}}{\sqrt{\overline{x'^2}}\sqrt{\overline{y'^2}}}. \tag{3.28}$$

Consider now the special case, in which all standard uncertainties are assumed to be equal: $\sigma_n = \sigma$; the χ^2 fit then reduces to the ordinary least-squares fit. In this case, the anomalies $x' \equiv x - \bar{x}$ and $y' \equiv y - \bar{y}$ are defined with respect to the sample average $\bar{x}$ and $\bar{y}$, while the quantities $\overline{x'^2}$, $\overline{y'^2}$, and $\overline{x'y'}$ are entries of the *dispersion* matrix $\mathbf{D}$, which is a finite-sample analog of the covariance matrix based on a population of a random-variable pair:

$$\mathbf{D} \equiv \begin{pmatrix} \overline{x'^2} & \overline{x'y'} \\ \overline{x'y'} & \overline{y'^2} \end{pmatrix}. \tag{3.29}$$

The dispersion matrix is a finite-sample analog of the covariance matrix based on a two-dimensional random population [see (1.19a) and (1.19b)]. The quantity $\overline{y'^2}$ is the dispersion of y, while $\overline{y^{*2}}$ is the square of the so-called *root-mean-square error*, or *r.m.s. error* of our least-squares fit. The ratio $\overline{y^{*2}}/\overline{y'^2}$ in (3.27) thus measures *the fraction of dispersion* unexplained *by a least-squares fit.* Since

[fraction of explained dispersion] + [fraction of unexplained dispersion] = 1,

the square of *correlation coefficient* r defined by (3.28) measures the fraction of dispersion explained by a linear least-squares fit between two variables. This fraction is naturally less than unity unless x and y are exactly linearly related; therefore, $r^2 \leq 1$ or $-1 \leq r \leq 1$.

For example, if the correlation coefficient is equal to $r = 0.5$, the dispersion of the data set $\hat{y}_n \equiv a + bx_n$ $(1 \leq n \leq N)$, where a and b are the best-fit parameters based on N pairs of $\{x_n,\, y_n\}$, is only equal to 25% of the dispersion of the original set $\{y_n\}$. Thus, 75% of the original set's dispersion remains "unexplained" by our least-squares fit. The normalized r.m.s. error is, therefore, equal to $\sqrt{0.75} \approx 0.87$. In other words, only 13% reduction in the r.m.s. error of y due to hypothesized linear dependence between x and y results from the correlation coefficient of 0.5. Consider the following table:

r	**r.m.s. error**
0.98	20%
0.90	43%
0.80	60%
0.50	87%
0.30	96%

As this table illustrates, large value of correlation coefficient does not necessarily mean that the statistically significant linear association between y and x can be exploited to forecast the value of the variable y given the knowledge of the variable x. In other words, given enough data, we may be able to show that the true correlation coefficient exceeds 0.3 at 99% confidence level (see Section 3.3), but this correlation, according to the table above, is useless for forecasting, reducing r.m.s. error by 4% only!

Exercise 12. In the least-squares fit ($\sigma_n = \sigma$), show that the correlation between $\epsilon \equiv \{y_n - a - bx_n\}$ $(1 \le n \le N)$ and $x \equiv \{x_n\}$ $(1 \le n \le N)$ (a and b are the best-fit parameters for N pairs $\{x_n, y_n\}$) is zero. Prove also that the correlation $r_{\epsilon \mathrm{y}}$ between ϵ and $y \equiv \{y_n\}$ $(1 \le n \le N)$ satisfies the relation $r^2_{\epsilon \mathrm{y}} + r^2_{\mathrm{xy}} = 1$, where r_{xy} is the correlation between x and y.

! The correlation coefficient is often used as a measure of whether two data sets are related via cause-and-effect relationship or not. When doing so, one has to realize that the following possibilities might take place:

- Zero or small correlation coefficient does not necessarily mean that the two variables are not related. The variables may:
  - be related nonlinearly (see Fig. 3.2d). For example, if the true relationship is $y = x^2$ and data is sampled evenly with respect to $x = 0$, then the linear correlation coefficient is zero.
  - be in quadrature with each other. For example, meridional wind and geopotential are approximately uncorrelated along latitudes even though the winds are very well approximated as the derivative of the geopotential (by geostrophy) — one says that the meridional wind is *in quadrature* with the geopotential [if the geopotential $\Phi(x) \sim \sin(x)$, then $v(x) \sim \partial\Phi/\partial x \sim \cos(x) = \sin(x + \pi/2) \sim \Phi(x + \pi/2)$].
- Large correlations may occur if two dynamically unrelated variables are both correlated with the third variable. Example: most geophysical variables are correlated with seasonal cycle. This correlation does not mean that the reason for the cold Arctic during Northern Hemisphere's winter lies in anomalously warm Antarctica. High anti-correlation in this case might be regarded as spurious from the standpoint of trying to find real relationship between two variables that might lead to physical insight or be useful in prediction.
- Be mindful of other possibilities illustrated in Figs. 3.2b,c: in the former case, the data contains an outlier and we'd better use *robust regression* techniques (see the discussion at the end of the present

section), while the latter case was created by drawing the values of y from the normal distribution with the mean of -0.5 for any $x < 0$ and with the mean of $+0.5$ for $x > 0$; the deterministic relationship between y and x (contaminated by noise) thus has a step-function character.

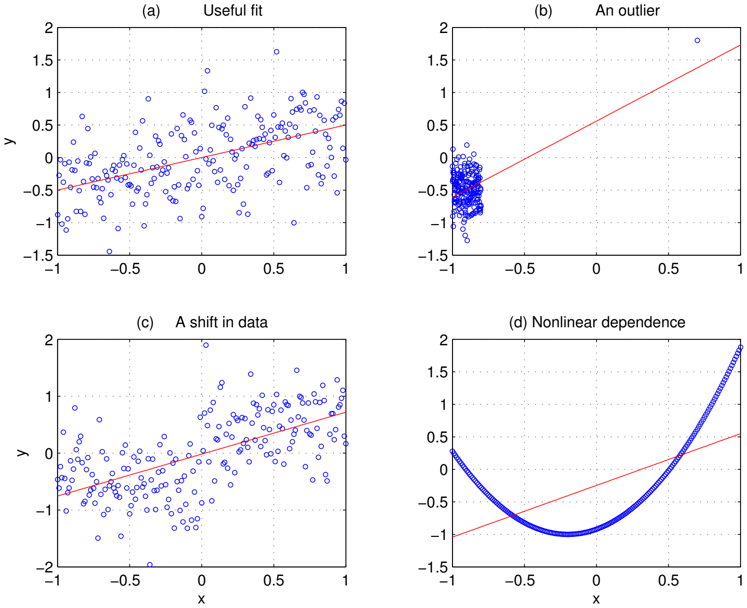


Figure 3.2: Examples of **linear regression lines — in all cases a linear correlation of y with x is 0.5**: (a) a useful fit; (b) an outlier; (c) a shift in data (no actual linear trend present); (d) y is exactly related to x via parabolic (that is, not linear) expression.

Straight-line fit for a data set with errors in both dimensions. In the above discussion, we have assumed that one of the variables (x) is precisely known, while the other one (y) is contaminated by measurement errors. In reality, it often happens that both variables are subject to errors. The simplest way to proceed in this case is to treat one of the variables as an independent variable, that is, ignore the associated measurement errors, and apply the standard linear regression. Note that the outcome of such a procedure depends on which of the variables is assumed to be independent: the two possible regression lines obtained by regressing y on x and x on y are the same only if the data are exactly collinear. $\boxed{!}$

The task of fitting the straight-line model (3.10) to the data is considerably harder in the case we want to consistently allow for the fact that both variables are known to within some finite precision. Each quantity $y_n - a - bx_n \equiv (y_{\mathrm{d},\,n} - a - bx_{\mathrm{d},\,n}) + (y_{\mathrm{r},\,n} - bx_{\mathrm{r},\,n})$ is assumed to be the sum of two parts, of which $(y_{\mathrm{d},\,n} - a - bx_{\mathrm{d},\,n})$ represents the true deterministic relation between y and x (and is equal to zero for a perfect fit), while the remainder is random and consists of the errors due to uncertainties in both x and y. The latter two random components are assumed to be statistically independent, so that the expression for the variance of $y_n - a - bx_n$ is

$$\mathrm{Var}\{y_n - a - bx_n\} = \mathrm{Var}\{y_n\} + b^2\mathrm{Var}\{x_n\}; \tag{3.30}$$

note that if x_n is known exactly ($\mathrm{Var}\{x_n\} = 0$), then $\mathrm{Var}\{y_n - a - bx_n\} = \mathrm{Var}\{y_n\}$. The quantity χ^2 given by

$$\chi^2 = \sum_{n=1}^{N} \frac{(y_n - a - bx_n)^2}{\sigma^2_{\mathrm{y},n} + b^2\sigma^2_{\mathrm{x},n}}, \tag{3.31}$$

where $\sigma^2_{\mathrm{y},n}$, $\sigma^2_{\mathrm{x},n}$ are variances of y and x measurements, is thus χ^2-distributed as the sum of N random variables normalized by their respective standard deviations (if N is large, the possible non-gaussianity of individual errors does not matter for the validity of the latter statement; otherwise, the individual errors are implicitly assumed to be Gaussian-distributed — in which case adjusting a and b to minimize (3.31) also gives the maximum likelihood estimate for these parameters.)An extra difficulty, which we encounter in trying to minimize (3.31) is that the function $\partial\chi^2/\partial b$ is nonlinear in b and its solution is more challenging to find. A useful geometrical interpretation of χ^2 given by (3.31) is the one in terms of the dispersion in the direction of smallest χ^2 between each data point and the line with slope b ("direction of minimum variance" in Fig. 3.3). In a particular case $\sigma^2_{\mathrm{y},n} = \sigma^2_{\mathrm{x},n} = \mathrm{const}$, minimizing χ^2 is equivalent to minimizing the perpendicular distance between the data points from the line in a two-dimensional space. This problem is solved by the so-called empirical orthogonal function (EOF) analysis, also known as the principal component analysis (PCA); see Chapter 4. This method finds an orthogonal rotation of the (x, y) coordinate system $x' = x\cos\theta - y\sin\theta$; $y' = x\sin\theta + y\cos\theta$, where θ is the angle of rotation, that aligns the new x-axis with the direction of maximum variance and the new y axis with the perpendicular direction of minimum variance (see Fig. 3.3). Posed in this way, the regression problem is related to finding eigenvalues and eigenvectors of the data's dispersion matrix.

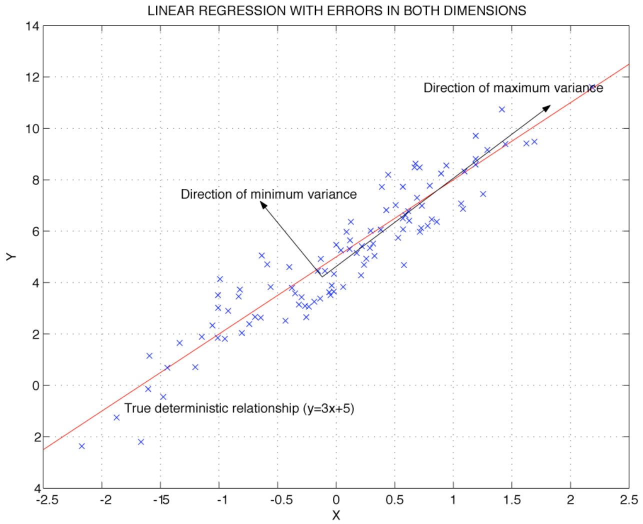
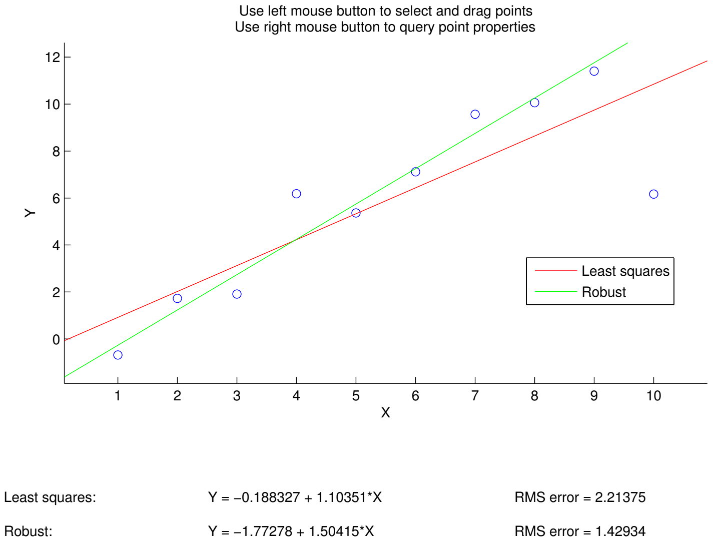


Figure 3.3: Straight-line fit with errors in both coordinates (see text for details).

Robust estimation. Figure 3.3(b) presents an example of the case in which the standard linear least squares procedure does not work because of the presence of an outlier in the data. Mathematically speaking, the reason for this failure is due to the assumption of a Gaussian distribution of errors, implicit in the above procedure, being violated. If we know that the data are strongly non-normal, it is desirable to use *robust techniques* of statistical analysis. The term "robust" is used here in the sense of being less sensitive to the small departures from the idealized assumptions about the probability distributions which underlie the data, than the technique based on the assumption of Gaussianity. In Fig. 3.3(b), just one oulier point changes the linear fit dramatically. A similar example is presented in Fig. 3.4. There is an outlier point at $x \approx 10$, which results in the slope of the least-squares straight-line fit (red line) to be underestimated. A possible way of making least-squares estimation more robust is to use χ^2 fitting by assigning the outlier points smaller weights in the merit functional than to the points in the central portion of a sample distribution. In the iterative technique called the *robust regression*, this assignment is done iteratively, by first computing the ordinary least-squares fit, then searching for the outliers with respect to the fitted line and assigning to those points smaller weights for the subsequent χ^2 fit. The outliers are then re-defined with respect to this new fit and the procedure is repeated until convergence of the fitted line's slope and intercept.

!


Use left mouse button to select and drag points
Use right mouse button to query point properties

```
Least squares:          Y = −0.188327 + 1.10351*X          RMS error = 2.21375

Robust:                 Y = −1.77278 + 1.50415*X           RMS error = 1.42934
```

Figure 3.4: Robust regression [produced by MATLAB's command "robustdemo"] (see text for details).

There is a number of other parametric and nonparametric robust techniques. The former assume some kind of non-Gaussian distribution, typically with longer tails (e.g., two-sided exponential), and derive the maximum likelihood estimators in a fashion similar to the χ^2-fit derivation of Section 3.1. Nonparametric methods seek to maximize some measure of association between two data sets without *a priori* assumptions about the underlying probability distributions. Yet another technique, *Kalman filtering*, produces "best estimates" of a signal in the presence of noise, by an optimal online processing of incoming raw measurements in a way that accounts for slow changes both in the signal and in the noise (error) covariance. Related *data assimilation* methods combining numerical models and observational data are currently used for operational weather forecasting.

3.3 Sampling theory of correlation

Suppose we have computed the correlation coefficient r between the components of a paired data set $\{x_n,\, y_n\}$ $(1 \le n \le N)$ using (3.28). How do we decide if this value of correlation coefficient is statistically significant? As in any statistical significance testing procedure, we assume that our paired data set is just a sample of size N of independent random variables drawn from a known two-dimensional distribution, whose covariance matrix (see 1.19a or 1.19b) is diagonal. The latter condition means, in other words, that x and y are assumed to be truly uncorrelated. We then would like to compute the distribution of such finite samples' correlation coefficient and check whether the observed correlation falls within the appropriate critical region or not. Below we list, without proof, several statements pertaining to the distribution of the finite paired sample's correlation coefficient.

The first statement is that if our theoretical distribution is sufficiently "good" (tails fall off to zero sufficiently rapidly), the sample size N is large ($N > 500$, according to Press et al. 1994), and the true correlation coefficient $\rho = 0$, then a finite sample's correlation coefficient r has a Gaussian distribution with mean zero and variance $1/N$. Once again, the theoretical distribution in the case of large N need not necessarily be two-dimensional Gaussian, or *binormal* —

$$p(x,\, y) \sim \exp\{-\frac{1}{2}(a_{11}x^2 - 2a_{12}xy + a_{22}y^2)\} \tag{3.32}$$

with $a_{12} = 0$ [a_{11}, a_{22} and a_{12} are arbitrary constants, and the theoretical correlation coefficient ρ between two random variables defined by distribution (3.32) is $\rho = -a_{12}/\sqrt{a_{11}a_{22}}$] — for the above statement about the distribution of r to be true.

All further statements ASSUME that the underlying theoretical PDF is the binormal one — (3.32), but DO NOT assume, in return, that N is large. For example, to test the null hypothesis of zero correlation, one makes use of the fact that the quantity

$$t = r\sqrt{\frac{N-2}{1-r^2}} \tag{3.33}$$

has the Student's t-distribution with $\nu = N-2$ degrees of freedom (and, of course, asymptotes Gaussian distribution with mean zero and variance $1/N$ as $N \to \infty$).

If the true correlation coefficient is not expected to be zero, the significance testing relies on the so-called *Fisher's z-transformation*, which converts the (asymmetricallydistributed) r into the variable z which is normally distributed:

$$z = \frac{1}{2} \ln \left\{ \frac{1+r}{1-r} \right\}, \tag{3.34}$$

with the mean μ_z

$$\mu_z = \frac{1}{2} \left[\ln \left\{ \frac{1+\rho}{1-\rho} \right\} + \frac{\rho}{N-1} \right] \tag{3.35}$$

and standard deviation σ_z

$$\sigma_z \approx \frac{1}{\sqrt{N-3}}. \tag{3.36}$$

Example 3.1 *Let us take $N = 21$ and $r = 0.8$ and find 95% confidence limits*[4] *on ρ. The value of z given by (3.34) is 1.0986, and the 95% confidence limits on a true value of μ_z are*

$$z - 1.96\sigma_z < \mu_z < z + 1.96\sigma_z \text{ or } 0.6366 < \mu_z < 1.5606,$$

where we have used the expression (3.35) for σ_z and applied two-sided test with $z_{0.025} = 1.96$.

To convert this to the statement about correlation coefficient, we make use of the fact that N is sufficiently large, so we neglect for simplicity the second term on the right-hand side of the expression (3.35) for μ_z. This gives

$$\rho \approx \tanh(\mu_z),$$

yielding the 95% confidence interval on ρ to be $0.56 < \rho < 0.92$.

The above procedure can also be used to assess statistical significance of the difference between the correlation coefficients $r_1 - r_2$ based on samples of sizes N_1 and N_2: the statistic

$$z = \frac{z_1 - z_2 - (\mu_{z_1} - \mu_{z_2})}{\sigma_{z_1 - z_2}}, \tag{3.37}$$

in which $\sigma^2_{z_1-z_2} \equiv \sigma^2_{z_1} + \sigma^2_{z_2}$ and the quantities z's, μ_z's and σ_z's are given by expressions (3.34), (3.35), and (3.36) estimated using the data for our first and second sample, has the standard normal distribution.

[4] Recall, however, that the method of confidence intervals gives an estimate of a true parameter in a particular sense; see Section 2.7.5.

3.4 Autocorrelation

3.4.1 Autocorrelation function

Given a continuous function $x(t)$ of an independent variable t, defined on the interval $[t_1,\, t_2]$, the *autocovariance function* $\phi(\tau),\ \tau \geq 0$ is

$$\phi(\tau) \equiv \frac{1}{t_2 - t_1 - \tau} \int\limits_{t_1}^{t_2-\tau} x'(t)x'(t+\tau)\, dt, \tag{3.38}$$

where the perturbation x' with respect to the average $\bar{x}$ is given by

$$\begin{aligned} x'(t) &\equiv x(t) - \bar{x}, \\ \bar{x} &\equiv \frac{1}{t_2 - t_1} \int\limits_{t_1}^{t_2} x(t)\, dt. \end{aligned} \tag{3.39}$$

In the discrete case, in which x is defined at equally spaced points t_1, $t_2 = t_1 + \Delta t$, $t_3 = t_2 + \Delta t$, $\ldots$, $t_N = t_{N-1} + \Delta t$, we can calculate the autocovariance at lag time $L\Delta t$ as

$$\phi(L\Delta t) \equiv \frac{1}{N - L - 1} \sum_{k=1}^{N-L} x'_k x'_{k+L} \equiv \overline{x'_k x'_{k+L}};\ \ L = 0,\, 1,\, 2, \ldots \tag{3.40}$$

and, of course,

$$\begin{aligned} x'_k &\equiv x_k - \bar{x}, \\ \bar{x} &\equiv \frac{1}{N} \sum_{k=1}^{N} x_k. \end{aligned} \tag{3.41}$$

If t is interpreted as time and $x(t)$ is a time series of a variable of interest, then the autocovariance is the covariance of the variable with itself at some other time, measured by a time lag (or lead) τ. The autocovariance at lag zero is thus equal to the variance (dispersion) of the variable: $\phi(0) = \overline{x'^2}$.

We can also think of autocovariance as of the characteristic of a random process; our time series $x(t)$ is then assumed to be but a single realization of this process and the autocovariance based on this time series — the sample autocovariance — is an approximation to the true autocovariance. The process is the law by which, knowing the value of the variable at a given initial time t_1, we can estimate the value of this variable at some later time t_2, while the random nature of the process assures that any two such realizations of our system's trajectory will be different. Suppose we have performed M long simulations of our trajectory from the same initial conditions $x(t = 0) = x_0$ and we have obtained M estimates $x_{\text{in}}^{(m)}$ of our variable at time t_1 and M estimates $x_{\text{f}}^{(m)}$ of our variable at time $t_2 = t_1 + \tau$, $\tau \geq 0$ (t_1 is assumed to be much longer than our system's internal memory). Let us define now the quantities

$$\begin{aligned}
\bar{x}_{\text{in}} &\equiv \lim_{M\to\infty} \sum_{m=1}^{M} x_{\text{in}}^{(m)}/M, \quad \bar{x}_{\text{f}} \equiv \lim_{M\to\infty} \sum_{m=1}^{M} x_{\text{f}}^{(m)}/M \\
x_{\text{in}}^{\prime(m)} &\equiv x_{\text{in}}^{(m)} - \bar{x}_{\text{in}}, \quad x_{\text{f}}^{\prime(m)} \equiv x_{\text{f}}^{(m)} - \bar{x}_{\text{f}}, \\
\Phi(t_1,\, t_2) &\equiv \lim_{M\to\infty} \sum_{m=1}^{M} x_{\text{in}}^{\prime(m)} x_{\text{f}}^{\prime(m)}/M. \qquad (3.42)
\end{aligned}$$

For the *stationary process*, $\bar{x}_{\text{in}} = \bar{x}_{\text{f}} = \bar{x}$ and $\Phi(t_1,\, t_2) = \Phi(\tau)$. The latter quantity is the covariance function of a stationary process, which only depends on a time interval τ and does not depend on the choice of initial point t_1. **In a time series analysis, it is assumed, in general, that the time series is stationary (that is, the underlying process is stationary); this implies, in particular, that one needs to remove any trends from the time series prior to the analysis (see Section 3.4.3).**

Let us now come back to the definition (3.40). The covariance at lag Δt, for example, is obtained by first forming the anomalies x' with respect to the sample's average; then the covariance is estimated as the sum $x_2 x_1 + x_3 x_2 + \ldots + x_N x_{N-1}$ divided by $N - 2$. This estimate would coincide with (3.42) for $t_2 = t_1 + \Delta t$ in the limit $N \to \infty$ *if we have assumed the underlying process is stationary* [in this case the pairs $(x_k,\, x_{k+1})$ can be viewed as independent pairs separated by Δt]. The assumption of stationarity also implies that the autocovariance function is symmetric $\phi(\tau) = \phi(-\tau)$:

$$\phi(-L\Delta t) = \phi(L\Delta t). \qquad (3.43)$$

Normalized autocovariance

$$r(\tau) = r(-\tau) \equiv \phi(\tau)/\phi(0) \qquad (3.44)$$

is called the *autocorrelation*.

3.4.2 Red noise and white noise

An important example of a stationary random process is the so-called *red-noise* process, which is defined as

$$x(t) = ax(t-\Delta t) + (1-a^2)^{1/2}\epsilon(t), \tag{3.45}$$

where $0 \le a \le 1$ is the parameter measuring the degree to which the memory of the previous state is retained, ϵ is a random number drawn, at every time step, from the standard normal distribution with mean zero and unit standard deviation and Δt is the time interval between two consecutive data points. The process (3.45) is characterized by $\bar{x} = 0$ and $\overline{x'^2} = 1$.

What is the autocorrelation of a red-noise process? Multiply both sides of (3.45) by $x(t-\Delta t)$ and ensemble average the resulting expression (recall that we have denoted ensemble averaging operation by the overbar):

$$\begin{aligned}
\overline{x(t-\Delta t)x(t)} &= a\overline{x(t-\Delta t)x(t-\Delta t)} + (1-a^2)^{1/2}\overline{x(t-\Delta t)\epsilon(t)} \\
&= a\cdot 1 + (1-a^2)^{1/2}\cdot 0; \quad \text{therefore} \\
\overline{x(t-\Delta t)x(t)} &\equiv \rho(\tau = \Delta t) = a.
\end{aligned}$$

The autocorrelation of the process (3.45) at lag Δt is equal to a. Let us now express $x(t+\Delta t)$ via $x(t-\Delta t)$:

$$\begin{aligned}
x(t+\Delta t) &= ax(t) + (1-a^2)^{1/2}\epsilon(t) \\
&= a^2x(t-\Delta t) + a(1-a^2)^{1/2}\epsilon(t) + (1-a^2)^{1/2}\epsilon(t-\Delta t).
\end{aligned}$$

Multiplying the above expression by $x(t-\Delta t)$ and ensemble averaging yields

$$\begin{aligned}
\overline{x(t-\Delta t)x(t+\Delta t)} &= a^2\overline{x(t-\Delta t)x(t-\Delta t)} + a(1-a^2)^{1/2}\overline{x(t-\Delta t)\epsilon(t)} \\
&+ (1-a^2)^{1/2}\overline{x(t-\Delta t)\epsilon(t-\Delta t)} \\
&= a^2\cdot 1 + a(1-a^2)^{1/2}\cdot 0 + (1-a^2)^{1/2}\cdot 0; \quad \text{therefore} \\
\overline{x(t-\Delta t)x(t+\Delta t)} &\equiv \rho(\tau = 2\Delta t) = a^2 = \rho(\tau = \Delta t)^2.
\end{aligned}$$

The autocorrelation of a red-noise process at lag $2\Delta t$ is equal to the autocorrelation at lag Δt squared. By induction

$$\rho(\tau = n\Delta t) = \rho(\tau = \Delta t)^n = a^n. \tag{3.46}$$

The function which satisfies the above property is the exponential function, so the autocorrelation function of a red-noise process (3.45) is

$$\rho(\tau) = \exp\{-|\tau|/T\}, \quad T \equiv -\Delta t/\ln a. \tag{3.47}$$

The red noise process is often used as a null hypothesis about the nature of observed geophysical time series. The system's slow dynamical processes are characterized by the "memory" parameter a, while the fast processes supply the energy to the low-frequency subsystem via stochastic excitation. In a special case $a = 0$ our data series becomes that of independent random numbers — *white noise*; this system has no memory of the past state. The autocorrelation function of the white noise is the delta function $\delta(\tau)$.

Examples of autocorrelation function for various combinations of signal and noise are shown in Fig. 3.5.

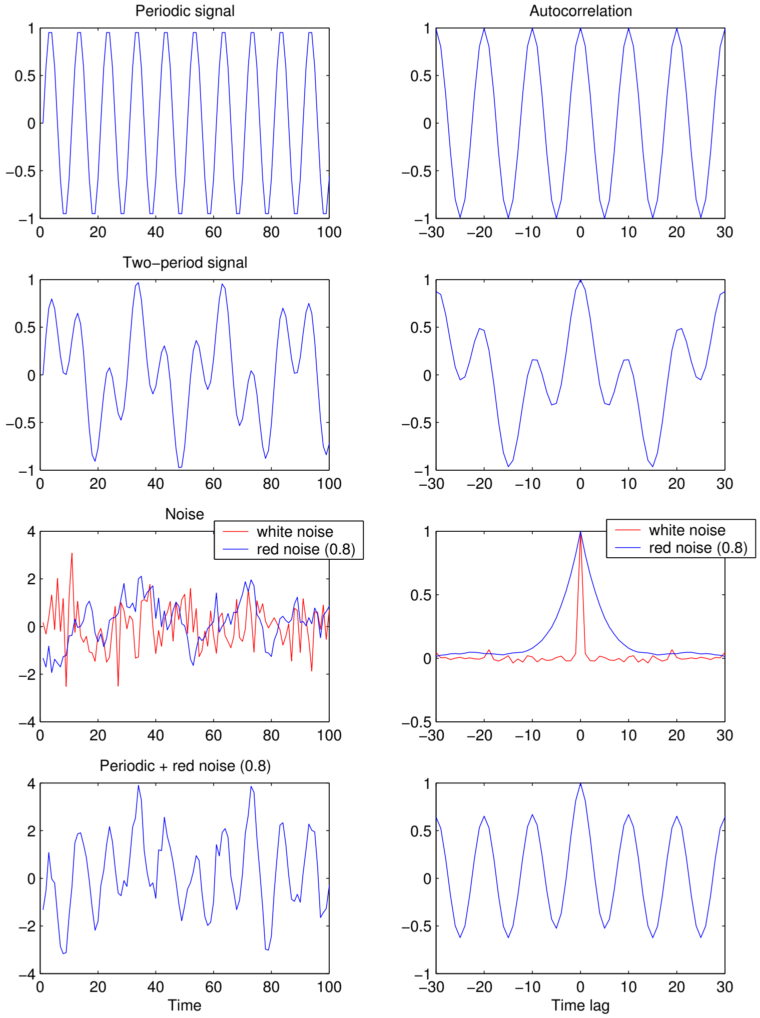


Figure 3.5: Examples of autocorrelation function.

3.4.3 How to estimate the number of degrees of freedom in a time series?

The autocorrelation of the time series is something that can be used to estimate the number of effective degrees of freedom N^* in this time series (denote the length of the time series by N: $N^* \leq N$). We will list here two estimates of the number of degrees of freedom in the time series: a more conservative one due to Leith (1973) and an alternative, less conservative estimate by Bretherton et al. (1999). Both are based on the sample's autocorrelation r at lag Δt, where Δt is the sampling interval. The Leith's expression reads as

$$N^* = \frac{N\Delta t}{2T}, \tag{3.48}$$

where T is the time interval over which the autocorrelation drops to $1/e$. In other words, the number of degrees of freedom in the time series is equal to half of the number of e-folding time scales. For a red-noise process, T is uniquely defined by the value of lag-1 autocorrelation $r(\Delta t)$; see (3.47). The expression (3.48) then becomes

$$\frac{N^*}{N} = -\frac{1}{2}\ln[r(\Delta t)], \tag{3.49}$$

where we have substituted the true value of lag-1 autocorrelation by its sample estimate $r(\Delta t)$. Note that (3.49) results in a meaningless $N^*/N > 1$ for $r(\Delta t) < 0.16$, in which case N^*/N must be set to unity (the two consecutive points in the time series are uncorrelated and, therefore, we have N independent samples).

Bretherton et al. (1999) have suggested that a more accurate estimate of the number of degrees of freedom (particularly for variance and covariance analysis) is

$$\frac{N^*}{N} = \frac{1 - r(\Delta t)^2}{1 + r(\Delta t)^2}. \tag{3.50}$$

The two dependencies (3.49) and (3.50) are shown in Fig. 3.6. We see that the Bretherton et al.'s formula allows about twice as many degrees of freedom as the Leith's formula when the autocorrelation is large.

3.4.4 Linear trend. Testing for trends

We now recall that the autocorrelation, at least when used as above to estimate the number of degrees of freedom in the time series, has been derived under the assumption that the

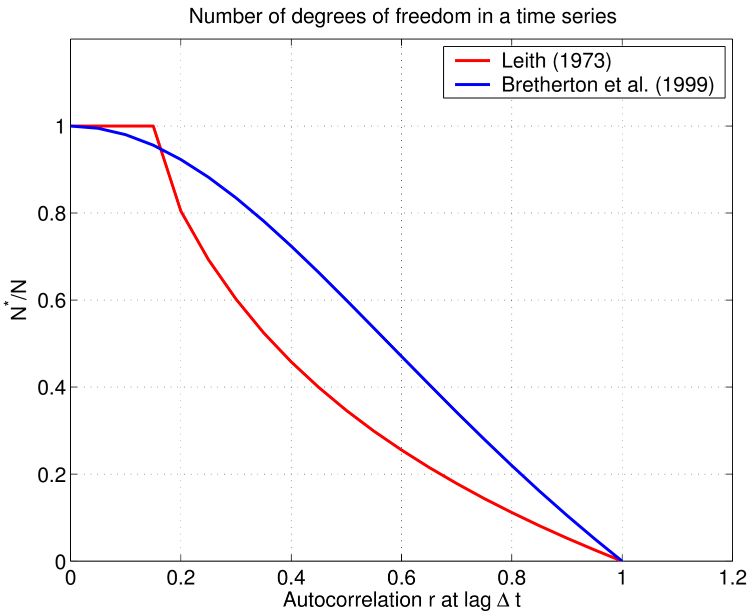


Figure 3.6: Number of degrees of freedom in a time series as a function of lag-1 autocorrelation.

underlying process is stationary. Given a time series, therefore, it is important to remove any trends. Linear trends can be removed by correlating the time series with the time axis, as in Section 3.2. Detrended time series' lag-1 autocorrelation can then be computed as in (3.40), (3.44) and the effective number of degrees of freedom estimated by either of (3.49) or (3.50). In practice, it is enough to restrict oneself to such linear detrending for the purposes of estimating the number of effective degrees of freedom in a time series.

After the number of degrees of freedom in the time series has been estimated and the time series is to be subjected to further statistical analysis, it is desirable to reduce the number of points in the time series to match the inferred number of degrees of freedom by, for example, binning the data in some way or another (removing or averaging out extra points will not, by definition, reduce the information content of the data, since the latter points are too strongly correlated with the neighbors; in fact, retaining dependent points *deteriorates* the accuracy of regression analysis, for example, — see Section 3.5). The statistical significance of the trend in data can then be estimated by any type of parameteric or nonparameteric technique. We will consider below examples of application of t-test, signs test and bootstrap method to estimate statistical significance of global warming.

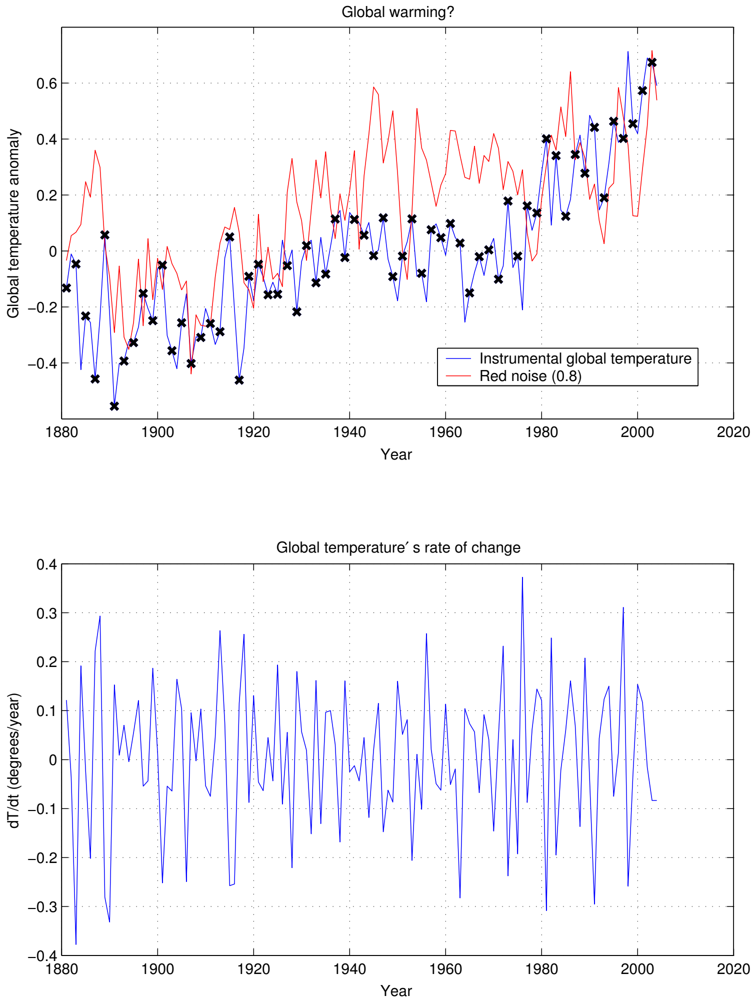


Figure 3.7: Testing for trends: global warming (see text for details).

Exercise 13. Consider annual-mean data for the global temperature anomaly (1881–2004; Fig. 3.7). The instrumental record is shown in blue (top panel) and appears to indicate that the global temperature increased by about 1 degree during the past one hundred years. Is that the result of human-induced CO_2 increase in the atmosphere? The red curve in the same plot is that of a *stationary* red-noise sample with the same dispersion as the instrumental data record, lag-1 autocorrelation of 0.8 and zero mean. This sample is characterized by temperature anomalies of about -0.4 in the beginning of the century and by those of about $+0.6$ in the end of the century; furthermore, the level and overall time scales of variability in this time series are similar to those of the instrumental record. This illustrates that, at the least, there is a nonzero probability that the apparent global temperature trend is just due to particular random sampling and has nothing to do with increased CO_2 in the atmosphere.

Exercise 13 (continued). There is a number of ways to see if this trend is statistically significant [assuming that all we have is the data sample above — in practice we may also try to employ historical data sets (proxy data) and GCM modeling resources to argue for or against the global warming occurrence]. The steps could be:

(1) Detrend the time series and compute its lag-1 autocorrelation: this results in the value of $r(1\,\text{year}) \approx 0.5$, so that the number of effective degrees of freedom is (Bretherton et al. 1999) $N^* \approx N(1-0.25)/(1+0.25) = 0.6N \approx 75$, since $N = 124$. Let us define three possible time series: (i) the original annual data ($N_1 = 124$), (ii) two-year non-overlapping box-car averages [the first point is just the average between the first and second point of original time series, the second point is the average between the third and fourth points of the original time series etc. — equivalently, we may just consider every second point of our time series (this sampling is shown by black x-signs in Fig. 3.7); the resulting time series has $N_2 = 62$ points], and (iii) four-year box-car averages ($N_3 = 31$). We can apply the same types of analysis to all three cases and compare the results. Note, however, that the case (ii) is the optimal one, since N_2 is closest to N^*; the case (i) clearly overestimates the number of independent samples in our time series, while the case (iii) loses too much useful information, which may result in an unnecessarily reduced statistical significance.

(2) Form the series of time derivatives $T_{n+1} - T_n$ [example for the case (i) is shown in the bottom panel of Fig. 3.7]. To this series, we can apply: (a) t-test or (b) bootstrap estimation to see if the average time derivative (whose observed value will be equal, in fact, to $(T_N - T_1)/(N-1)$ — a measure of the slope of the temperature time series) is significantly different from zero; (c) signs test to determine if the median of the time-derivative set is significantly different from zero.

(3) We can also proceed by computing the least-squares fit to the temperature time series and estimating weather its slope b is significantly different from zero. The statistic

$$t = b\sqrt{\frac{\overline{x'^2}}{(\frac{\chi^2}{N})/(N-2)}} \tag{3.51}$$

is t-distributed with $N-2$ degrees of freedom (provided the observations are independent). The above formula becomes (3.33) for the case $\overline{y'^2} = \overline{x'^2} = 1$, in which $b = r$ and $\chi^2/N = 1-r^2$ (can you see why this is true?)

(4) The trend in Fig. 3.7 seems to be nonuniform, with the steepest warming after 1970. Another way to estimate the significance of warming would be to compare the average temperatures in 1881–1970 and 1971–2004 using t-test for the difference in means (setting the expected difference to zero, of course).

Exercise 13 (continued). There are still other ways to test for trends: for example, we could fit a red-noise process to the observed time series and do Monte-Carlo simulations to estimate a large number of synthetic trends and compare these trends with the observed one and so on.

Are the results of (2)–(4) consistent? How can we interpret discrepancies? What can we say about inferred causes of warming — Is it likely to be a linear response to increased CO_2 in the atmosphere? How probable it is that the warming is just a statistical hoax?

3.5 Multiple linear regression. General linear least squares

Suppose now that we have more than one predictor variable. For example we measure temperature z_n (response variable) at a set of coordinate points $\{x_n, y_n\}$ $(1 \leq n \leq N)$, and we would like to determine an optimal linear fit

$$z = a_0 + a_1 x + a_2 y. \tag{3.52}$$

Generalizing to the case of N observations $x_j^{(n)}$ of an arbitrary number J of predictors x_j, and N observations $y^{(n)}$ of a response variable y, the problem is to find a set of best-fit parameters $a_0, a_1, a_2, \ldots, a_J$ for the linear model

$$y = a_0 + a_1 x_1 + \ldots + a_J x_J. \tag{3.53}$$

This problem is known as *multiple linear regression*, "linear," since the model (3.53) is linear in its parameters. The dependence on predictor variables need not be linear, however. Consider, once again, an example of one predictor variable x and construct the model $y(x)$ as a linear combination of any number J of specified functions $X_j(x)$ (*basis functions*). The functions could be $X_0(x) = 1$, $X_1(x) = x$, $X_2(x) = x^2$, $\ldots$, $X_J(x) = x^J$, in which case

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_J x^J. \tag{3.54}$$

The model (3.54) is known as *response surface model*, and the associated regression problem is called *polynomial regression* (quadratic model for the case $J = 2$). The general form of a generalized regression model is

$$y(x) = \sum_{j=1}^{J} a_j X_j(x), \tag{3.55}$$

and the problem of optimal fitting a set of a_j $(1 \le j \le J)$ to the model (3.55) given observed series $\{x^{(n)},\, y^{(n)}\}$ $(1 \le n \le N)$ is called the *general linear least squares* problem. Note that mathematically the general linear least squares problem (3.55) is equivalent to multiple linear regression (3.53) [due to linear dependence on model parameters], but the former attempts to model nonlinear relationship between predictor and response variables.

3.5.1 Statement of the problem

The general linear least squares problem is solved by minimizing the χ^2 merit functional, defined now as

$$\chi^2 = \sum_{n=1}^{N} \left[\frac{y^{(n)} - \sum_{j=1}^{J} a_j X_j(x^{(n)})}{\sigma_n} \right]^2, \tag{3.56}$$

where σ_n is the measurement error (standard uncertainty) of the n-th data point. Let $\mathbf{X} \equiv \{x_{nj}\}$ be an $N \times J$ matrix whose components x_{nj} are given by

$$x_{nj} \equiv \frac{X_j(x^{(n)})}{\sigma_n}. \tag{3.57a}$$

Define also an N-component vector $\tilde{\mathbf{y}} \equiv \{\tilde{y}_n\}$ and a J-component vector of parameters $\mathbf{a} \equiv \{a_j\}$:

$$\tilde{y}_n \equiv \frac{y^{(n)}}{\sigma_n}, \quad \mathbf{a} \equiv \{a_j\}. \tag{3.57b}$$

The matrix $\mathbf{X}$ is called the *design matrix* of the fitting problem; this matrix, as well as the response-variable and parameter vectors are schematically shown below:

$$\mathbf{X} \equiv \begin{pmatrix} \frac{X_1(x^{(1)})}{\sigma_1} & \frac{X_2(x^{(1)})}{\sigma_1} & \dots & \frac{X_J(x^{(1)})}{\sigma_1} \\ \frac{X_1(x^{(2)})}{\sigma_2} & \frac{X_2(x^{(2)})}{\sigma_2} & \dots & \frac{X_J(x^{(2)})}{\sigma_2} \\ \dots & \dots & \dots & \dots \\ \frac{X_1(x^{(N)})}{\sigma_N} & \frac{X_2(x^{(N)})}{\sigma_N} & \dots & \frac{X_J(x^{(N)})}{\sigma_N} \end{pmatrix} \quad \tilde{\mathbf{y}} \equiv \begin{pmatrix} \frac{y^{(1)}}{\sigma_1} \\ \frac{y^{(2)}}{\sigma_2} \\ \dots \\ \frac{y^{(N)}}{\sigma_N} \end{pmatrix} \tag{3.58}$$

$$\mathbf{a} \equiv (\quad a_1 \quad a_2 \quad \dots \quad a_J \quad).$$

Review of vectors and matrices. Let us now define a few useful matrix–vector operations. $\boxed{!}$

The *scalar product* $\mathbf{a} \cdot \mathbf{b}$ of the two vectors $\mathbf{a}$ and $\mathbf{b}$ of the same dimension J is the number

$$\mathbf{a} \cdot \mathbf{b} \equiv \sum_{j=1}^{J} a_j b_j. \tag{3.59}$$

The *matrix product* of the two matrices $\mathbf{A}$ and $\mathbf{B}$, with dimensions $N \times J$ and $J \times M$ (NB! inner dimensions of the two matrices must agree!) can be defined as the $N \times M$ matrix $\mathbf{C}$, whose elements c_{nm} are given by

$$\mathbf{C} \equiv \mathbf{A} \cdot \mathbf{B} : \quad c_{nm} \equiv \mathbf{a_n} \cdot \mathbf{b_m} \equiv \sum_{j=1}^{J} a_{nj} b_{jm} \tag{3.60}$$

— the scalar product of the n-th row of $\mathbf{A}$ and m-th column of $\mathbf{B}$. In case $M = 1$ the above notation defines the product of an $N \times J$ matrix onto a column vector of dimension $J \times 1$; the result is the column vector of dimension $N \times 1$.

The *sum(difference)* $\mathbf{c} = \mathbf{a} \pm \mathbf{b}$ of two vectors $\mathbf{a}$ and $\mathbf{b}$ (of the same dimension) is the vector whose components are equal to the sum (difference) of the respective components of $\mathbf{a}$ and $\mathbf{b}$.

The *length* of the vector is defined as

$$|\mathbf{a}| \equiv \sqrt{\mathbf{a} \cdot \mathbf{a}}. \tag{3.61}$$

Two vectors $\mathbf{a}$ and $\mathbf{b}$ (of the same dimension) are called *orthogonal* if their scalar product is zero:

$$\mathbf{a} \cdot \mathbf{b} = 0. \tag{3.62}$$

Orthogonal vectors of unit length are called *orthonormal.* N orthonormal vectors of length N form a *basis.* Example: the columns (or, equivalently, rows) of the identity matrix $\mathbf{I}$, whose diagonal elements are all equal to one, and all others are equal to zero (the identity matrix is a special case of a diagonal matrix, whose off-diagonal elements are all zero) —

$$\mathbf{I} \equiv \left(\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \multicolumn{4}{c}{\dots\dots\dots\dots} \\ 0 & 0 & \cdots & 1 \end{array} \right) \tag{3.63}$$

— form a basis.

The *inverse* $\mathbf{A}^{-1}$ of a square ($N \times N$) matrix $\mathbf{A}$ is the matrix which satisfies the relation

$$\mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{I}. \tag{3.64}$$

The notation (3.64) denotes a linear system of $N \times N$ equations for $N \times N$ unknown elements of $\mathbf{A}^{-1}$. This system only has a solution if all of its $N \times N$ equations are linearly independent (that is, none of the equations can be represented as a linear combination of others; otherwise the number of unknowns exceeds the number of independent equations and the system is *underdetermined*). If the latter independency condition is not satisfied, then this linear system of equations, as well as the matrix $\mathbf{A}$ are called *singular.* The inverse of a diagonal matrix is also diagonal matrix, each diagonal element of which is the inverse of the corresponding diagonal element of the original matrix. If the inverse $\mathbf{A}^{-1}$ of a square matrix $\mathbf{A} \equiv \{a_{nm}\}$ is equal to the matrix *transpose* $\mathbf{A}^{\mathrm{T}} \equiv \{a_{mn}\}$ (the transpose of a matrix is the matrix in which the rows are the columns of the original matrix and vice versa) —

$$\mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}}, \tag{3.65}$$

then the matrix A is called *orthogonal.* It can be shown that the rows (and columns) of orthogonal matrix form a basis.

Using definitions (3.58), and treating $\tilde{\mathbf{y}}$ and $\mathbf{a}$ as column vectors, the merit functional (3.56) can be written in vector notation as

$$\chi^2 = |\tilde{\mathbf{y}} - \mathbf{X} \cdot \mathbf{a}|^2, \tag{3.66}$$

so that our fitting problem becomes:

$$\text{find} \quad \mathbf{a} \quad \text{which} \quad \text{minimizes} \quad |\tilde{\mathbf{y}} - \mathbf{X} \cdot \mathbf{a}|. \tag{3.67}$$

3.5.2 Solution by use of normal equations

Basic formalism. Minimum of the quadratic form (3.56) is achieved if the parameters $\mathbf{a}$ satisfy the system of linear equations obtained by setting the partial derivatives of χ^2 with respect to each parameter to zero:

$$0 = \sum_{n=1}^{N} \frac{1}{\sigma_n^2} \left[y^{(n)} - \sum_{j=1}^{J} a_j X_j(x^{(n)}) \right] X_k(x^{(n)}) \quad k = 1, \ldots, J, \tag{3.68}$$

or, equivalently, as a matrix equation

$$\sum_{j=1}^{J} \xi_{kj} a_j = \eta_k, \tag{3.69}$$

where

$$\xi_{kj} \equiv \sum_{n=1}^{N} \frac{X_j(x^{(n)}) X_k(x^{(n)})}{\sigma_n^2} \qquad \Longleftrightarrow \qquad [\xi] \equiv \mathbf{X}^{\mathrm{T}} \cdot \mathbf{X} \tag{3.70}$$

is a $J \times J$ matrix and

$$\eta_k \equiv \sum_{n=1}^{N} \frac{y^{(n)} X_k(x^{(n)})}{\sigma_n^2} \qquad \Longleftrightarrow \qquad [\eta] \equiv \mathbf{X}^{\mathrm{T}} \cdot \tilde{\mathbf{y}} \tag{3.71}$$

a vector of length J. In matrix notation, the latter equation is

$$[\xi] \cdot \mathbf{a} = [\eta] \qquad \Longleftrightarrow \qquad (\mathbf{X}^{\mathrm{T}} \cdot \mathbf{X}) \cdot \mathbf{a} = \mathbf{X}^{\mathrm{T}} \cdot \tilde{\mathbf{y}}. \tag{3.72}$$

Equivalent formulations (3.68), (3.69), and (3.72) are called the *normal equations* of the least-squares problem.

The solution to the system of normal equation, the set of best-fit parameters, is given by

$$\mathbf{a} = [\xi]^{-1}[\eta] \quad \Longleftrightarrow \quad a_j = \sum_{k=1}^{J} [\xi]^{-1}_{jk}\eta_k = \sum_{k=1}^{J} c_{jk} \sum_{n=1}^{N} \frac{y^{(n)} X_k(x^{(n)})}{\sigma_n^2}, \tag{3.73}$$

where we have defined the matrix $\mathbf{C} \equiv [\xi]^{-1}$ and used definition (3.71) for η_k. Interchanging the order of summation, we get

$$a_j = \sum_{n=1}^{N} \alpha_j^{(n)} y^{(n)}, \quad \text{where} \quad \alpha_j^{(n)} \equiv \sum_{k=1}^{J} c_{jk} \frac{y^{(n)} X_k(x^{(n)})}{\sigma_n^2} \tag{3.74}$$

— the expressions for the best-fit parameters are linear in $y^{(n)}$ [compare with (3.17)]. Therefore, the variance of the parameters [see (3.19)] is

$$\mathrm{Var}\{a_j\} \equiv \sigma^2(a_j) = \sum_{n=1}^{N} (\alpha_j^{(n)})^2 \mathrm{Var}\{y^{(n)}\} = \sum_{n=1}^{N} (\alpha_j^{(n)})^2 \sigma_n^2 =$$

$$\sum_{n=1}^{N} \sum_{k=1}^{J} c_{jk} \frac{X_k(x^{(n)})}{\sigma_n^2} \sum_{l=1}^{J} c_{jl} \frac{X_l(x^{(n)})}{\sigma_n^2} \sigma_n^2 = \sum_{k=1}^{J} \sum_{l=1}^{J} c_{jk} c_{jl} \left[\sum_{n=1}^{N} \frac{X_k(x^{(n)}) X_l(x^{(n)})}{\sigma_n^2} \right]. \tag{3.75}$$

The last term in square brackets is just the element $[\xi]_{kl}$ of the matrix $[\xi]$. Since $\mathbf{C} \equiv [\xi]^{-1}$, then convoluting (applying matrix product) by summing over either k or l will result in the identity matrix, while the remaining summation (over l or k) will be the product of the matrix $\mathbf{C}$ with the identity matrix, resulting in

$$\sigma^2(a_j) = c_{jj} \tag{3.76}$$

— diagonal elements of $\mathbf{C}$ are the variances (squared standard uncertainties) of the best-fit parameters. Similarly, off-diagonal elements of $\mathbf{C}$ are covariances of the best-fit parameters; see (3.21), (3.22).

Finally, the goodness-of-fit can be estimating by ranking the "observed" value of χ^2 [estimated from (3.66) with $\mathbf{a}$ given by (3.73)] with respect to χ^2 distribution with $N - M$ degrees of freedom.

How many variables to use? Let us look at the regression problem above from a slightly different perspective. Once again, given a series of length N of predictand $y^{(n)}$ and those of a set of predictors $x_j^{(n)}$ $(1 \le j \le J,\ 1 \le n \le N)$ [predictors can represent different variables, as in multiple linear regression, or can be specified different functions of a single variable, as in general linear least-squares fit], we would like to parametrize y as an optimal linear function of $\mathbf{x} \equiv \{x_j\}$. Suppose that our "measurement errors" are unknown (for example, we get our time series from a simulation of a climate model and we hypothesize that the "signal" is our linear relation between predictand and predictors, but this relation is contaminated by noise — typical for climatic time series). In this case, we might want to center our raw series by removing their respective averages and scale them by their respective standard deviations (square root of dispersion)[5]:

$$x_j^{(n)*} \equiv \frac{x_j^{(n)} - \overline{x}_j}{s_{x_j}}, \quad y^{(n)*} \equiv \frac{y^{(n)} - \overline{y}}{s_y}. \tag{3.77}$$

In the following, we will drop the stars that indicate standardized variables, for convenience. With the above rescaling, the normal equations (3.69) become

$$\sum_{j=1}^{J} r_{kj} a_j = r_k, \tag{3.78}$$

where r_{kj} is the correlation coefficient between x_k and x_j and r_k is that between x_k and y.

Consider now the special case of $J = 2$. The solution to (3.78) is

$$a_1 = \frac{r_1 - r_{12} r_2}{1 - r_{12}^2}; \quad a_2 = \frac{r_2 - r_{12} r_1}{1 - r_{12}^2}. \tag{3.79}$$

As in the one-variable case of Section 3.2, the total dispersion $\overline{y'^2}$ of predictand can be represented as the sum of "explained" $[\chi^2/N \equiv \overline{(y - \hat{y})^2}]$ and "unexplained" $[\overline{(\hat{y} - \overline{y})^2}]$ dispersion (where, of course, $\hat{y} \equiv a_1 x_1 + a_2 x_2$). Rearranging this expression in the following way

$$\chi^2 = N\overline{y'^2}(1 - R^2)$$

defines the *multiple correlation coefficient* R [whose square is the fraction of explained dispersion; compare with (3.27)], which can be shown to be equal to

$$R^2 = \frac{r_1^2 + r_2^2 - 2 r_1 r_2 r_{12}}{1 - r_{12}^2}. \tag{3.80}$$

From the above, it becomes clear that adding a second predictor to a linear regression model is only justified if $R^2 > r_1^2$, since only in this case we would "explain" more of the

[5] In case some of the predictors and predictand have different units, this rescaling might be our only reasonable choice of fitting strategy.

variability by our model. The *minimal useful correlation* r_2^* between a predictand y and an additional predictor x_2 can thus be defined to accommodate the latter condition:

$$|R| > |r_1| \quad \text{if} \quad |r_2| > |r_2^*| \equiv |r_1 r_{12}|. \tag{3.81}$$

It can easily be checked that substituting $r_2^* = r_1 r_{12}$ into (3.80) results in $R^2 = r_1^2$, so that including the second predictor has no influence on the explained dispersion; in other words, the second predictor does not contribute at all to reducing the χ^2.

Stability of multiple linear regression. Similar considerations apply when considering the third predictor and so on. In general, we need to pick a set of *largely uncorrelated* (nearly orthogonal) predictor variables so that each of them is as highly correlated with the response variable as possible in order to achieve a statistically significant (reproducible on a number of independent samples) fit. If our additional predictor variable has a low correlation with the response variable and/or high correlation with existing predictors, its inclusion is not justified according to (3.81). In the latter case of a high correlation between one or more predictor variables, the linear system of normal equations (3.78) becomes nearly singular, which has a detrimental effect on the linear fit.

Consider the expressions (3.79) for the linear fit coefficients in the case $J = 2$, for example. If x_1 is perfectly correlated with x_2 ($r_{12} = 1$), then $r_2 = r_1$ and the expressions for a_1 and a_2 are of the type $0/0$ (we cannot fit a meaningful plane if we are only given a data on a line). If x_1 and x_2 are nearly (but not perfectly) correlated, then the coefficients are still a ratio of two very small numbers, and are thus unstable (will most definitely change from one independent sample to another). This is also reflected in the fact that the variance of the coefficients, given by $[r_{kj}]^{-1}$, will be large (since the inverse of a nearly singular matrix will contain large elements).

For the above reasons, adding more predictors to be used in a linear regression problem generally lowers the statistical significance of the "fit" to the data points, and the less likely the same estimate of regression parameters will be obtained based on an independent data sample. An objective way to choose an optimal number of predictors, or *regularize* nearly singular regression problem involves using singular value decomposition (SVD) of the design matrix.

3.5.3 Review of Singular Value Decomposition (SVD)

SVD methods are based on the following theorem of linear algebra (proof is beyond the scope of these notes). Any $N \times J$ matrix $\mathbf{X}$, whose number of rows N is greater than or equal to its number of columns J can be factored into (represented as the product of) three matrices: (i) $N \times J$ matrix $\mathbf{U}$, which is *column-orthogonal* ($\mathbf{U}^{\mathrm{T}}\cdot\mathbf{U} = \mathbf{I}_{J\times J}$); (ii) diagonal matrix $\mathbf{W}$ with positive or zero elements (these elements are called the *singular values*); and (iii) the transpose of a $J \times J$ orthogonal matrix ($\mathbf{V}^{\mathrm{T}}\cdot\mathbf{V} = \mathbf{I}_{J\times J}$):

$$\begin{pmatrix} \\ \\ \\ & \mathbf{X} & \\ \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ \\ & \mathbf{U} & \\ \\ \\ \\ \end{pmatrix} \cdot \begin{pmatrix} w_1 & & & \\ & w_2 & \cdots & \\ & & \cdots & \\ & & \cdots & \\ & & & w_J \end{pmatrix} \cdot \begin{pmatrix} \\ & \mathbf{V}^{\mathrm{T}} & \\ \\ \end{pmatrix}, \tag{3.82}$$

where

$$\begin{pmatrix} \\ & \mathbf{U}^{\mathrm{T}} & \\ \\ \end{pmatrix} \cdot \begin{pmatrix} \\ \\ \\ & \mathbf{U} & \\ \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ & \mathbf{V}^{\mathrm{T}} & \\ \\ \end{pmatrix} \cdot \begin{pmatrix} \\ & \mathbf{V} & \\ \\ \end{pmatrix}$$

$$= \begin{pmatrix} \\ & \mathbf{I} & \\ \\ \end{pmatrix} \tag{3.83a}$$

The latter orthogonality conditions can also be written in a component form using Kronecker-delta notation (δ_{ij} is unity if $i = j$ and zero otherwise):

$$\sum_{n=1}^{N} U_{ni}U_{nj} = \sum_{k=1}^{J} V_{ki}V_{kj} = \delta_{ij}, \quad 1 \le (i,\, j) \le J. \tag{3.83b}$$

The SVD decomposition can also be carried out for $N < J$, in which case the singular values w_j for all $j > N$ are equal to zero, and the corresponding columns of $\mathbf{U}$ are also zero; naturally, orthogonality conditions (3.83b) hold only for $i, j \leq N$.

The SVD decomposition can be done no matter how singular the matrix $\mathbf{X}$ is. SVD decomposition is unique up to (i) an arbitrary simultaneous permutation (re-ordering) of columns of $\mathbf{U}$, diagonal elements of $\mathbf{W}$ and columns of $\mathbf{V}$ (that is, *rows* of $\mathbf{V}^{\mathrm{T}}$); or (ii) forming arbitrary linear combinations of any columns of $\mathbf{U}$ and $\mathbf{V}$ (and scaling so that their lengths remain to be unity) whose corresponding elements of $\mathbf{W}$ happen to be exactly equal (that is, if any pair of such columns is substituted by linear combinations defined above, the matrix multiplication of SVD components so modified will also give the original matrix).

SVD of a square matrix. If $\mathbf{X}$ is $J \times J$ square matrix, then $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{V}$ all have dimensions $J \times J$. Let us compute the inverse of $\mathbf{X}$ — the matrix $\mathbf{X}^{-1}$, in terms of $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{V}$. The inverses of orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ are equal to their transposes, while the inverse of a diagonal matrix $\mathbf{W}$ is the diagonal matrix whose elements are the reciprocals of the elements w_j. From (3.82) and (3.83a) it then follows that

$$\mathbf{X}^{-1} = \mathbf{V} \cdot [\mathrm{diag}\ (1/w_j)] \cdot \mathbf{U}^{\mathrm{T}} \tag{3.84}$$

[multiply the decomposition (3.82) on the left by the right-hand side of (3.84) and use (3.83a) to show that the result is the identity matrix]. Therefore, according to (3.84), the matrix $\mathbf{X}$ is singular if one or more of its singular values are zero. If these values are nonzero, but small, the matrix is nearly singular, or *ill-conditioned*: this is measured by the *condition number*, which is the ration of the largest of the w_j to the smallest of the w_j.

SVD is very useful in diagnosing the solvability of linear systems of equations of the form

$$\mathbf{X} \cdot \mathbf{a} = \mathbf{y}, \tag{3.85}$$

where $\mathbf{X}$ is a matrix $J \times J$, while $\mathbf{a}$ and $\mathbf{y}$ are vectors of dimension J. In case $\mathbf{X}$ is non-singular (Fig. 3.8a), the above equation defines linear mapping of an original vector space into the one of the same dimension, with vector $\mathbf{a}$ mapped into a vector $\mathbf{y}$. However, if $\mathbf{X}$ is singular, it maps a vector space into the one with a lower dimension (Fig. 3.8b), for example, two-dimensional plane into a one-dimensional line (that is, a 2-D vector into a point!). The latter subspace is called the *range* of $\mathbf{X}$ (since it can be "reached" by applying transformation $\mathbf{X}$ to the original space defined by all possible $\mathbf{a}$'s). The dimension of this subspace (the number of linearly independent vectors that can be found in it) is called the

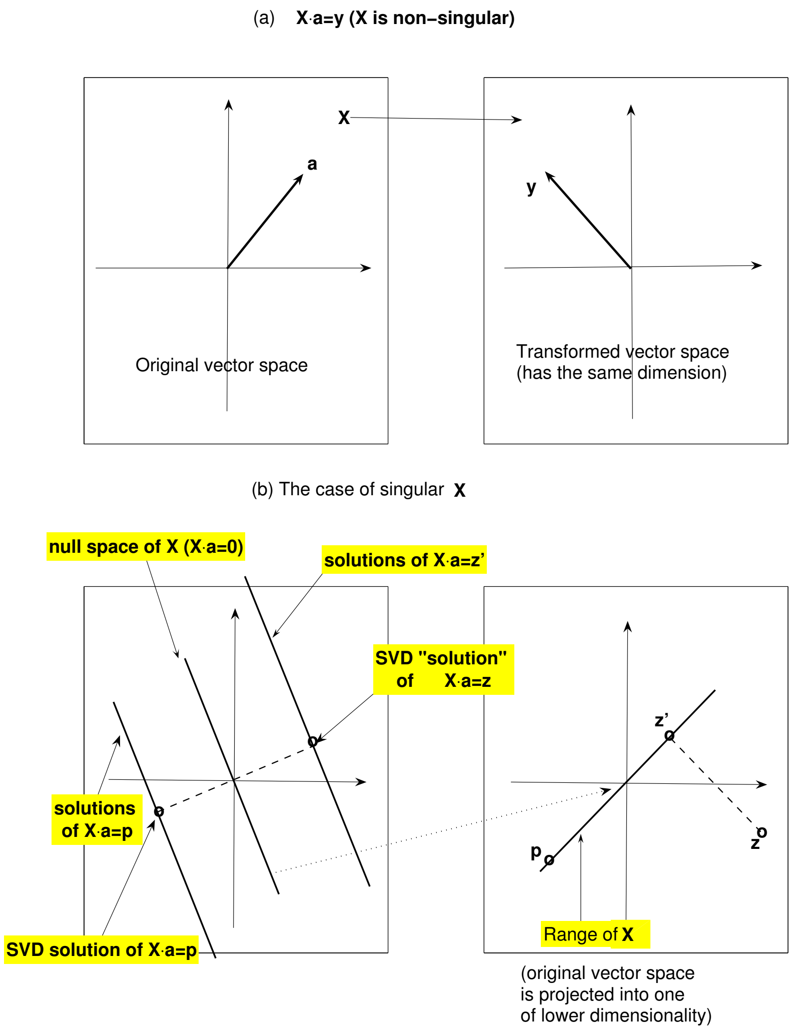


Figure 3.8: Solution of linear systems using SVD (see text for details).

rank of **X**. The rank of a non-singular $J \times J$ matrix is equal to J. The rank of a singular $J \times J$ matrix is less than J. The *nullspace* of **X** is the subspace of the original space that is mapped to zero, and the dimension of the nullspace is called the *nullity* of **X**. The nullity of a non-singular matrix is zero. For an arbitrary $J \times J$ matrix **nullity plus rank equals J**.

The utility of SVD is in that it explicitly constructs orthonormal bases for the nullspace and range of a singular matrix; in particular, the columns of **U** corresponding to non-zero singular values are an orthonormal set of basis vectors that span the range, while the columns of $\mathbf{V}$ corresponding to zero singular values form an orthonormal basis for the nullspace. The latter property means that SVD automatically provides the solution of a homogeneous problem (3.85), with $\mathbf{y} = \mathbf{0}$.

Consider now the case of a singular $\mathbf{X}$ and $\mathbf{y} \neq \mathbf{0}$ and compute the quantity

$$\mathbf{a} = \mathbf{V} \cdot [\text{diag } (1/w_j)] \cdot (\mathbf{U}^{\mathrm{T}} \cdot \mathbf{y}), \tag{3.86}$$

where, **if $\mathbf{w_j} = \mathbf{0}$, we need to replace $\mathbf{1/w_j}$ by zero!** The following statements apply (see Fig. 3.8b):

(i) if $\mathbf{y} = \mathbf{p}$ is in the range of $\mathbf{X}$, then (3.86) gives the vector solution of (3.85) with the smallest length $|\mathbf{a}|$ [that is, from an infinite number of possible solutions (infinite, since we can add to our solution any linear combination of vectors from the nullspace of $\mathbf{X}$), it picks the one closest to zero];

(ii) if $\mathbf{y} = \mathbf{z}$ is outside of the range of $\mathbf{X}$, then the solution (3.86) is the same as (i) for $\mathbf{y} = \mathbf{z}'$, where $\mathbf{z}'$ is the point from the range of $\mathbf{X}$ closest to $\mathbf{z}$.

Both cases can be written in the form of a single statement: **the solution (3.86) finds**

$$\mathbf{a} \quad \textbf{that} \quad \textbf{minimizes} \quad r \equiv |\mathbf{X} \cdot \mathbf{a} - \mathbf{y}|. \tag{3.87}$$

SVD for more equations than unknowns. The above results generalize to the case of overdetermined system of linear equations:

$$\left(\begin{array}{c} \\ \\ \\ \mathbf{X} \\ \\ \\ \\ \end{array} \right) \cdot \left(\begin{array}{c} \\ \mathbf{a} \\ \\ \end{array} \right) = \left(\begin{array}{c} \\ \\ \\ \mathbf{y} \\ \\ \\ \\ \end{array} \right) \tag{3.88}$$

Here $\mathbf{X}$ is an $N \times J$ matrix and the vectors $\mathbf{a}$ and $\mathbf{y}$ have dimensions of J and N, respectively. Given the singular value decomposition (3.82) of $\mathbf{X}$, the solution of (3.88) which minimizes

r defined by (3.87), is given by (3.86):

$$\begin{pmatrix} \\ \mathbf{a} \\ \\ \end{pmatrix} = \begin{pmatrix} & & \\ & \mathbf{V} & \\ & & \end{pmatrix} \cdot \begin{pmatrix} w_1^{-1} & & & \\ & w_2^{-1} & \cdots & \\ & & \cdots & \\ & & \cdots & \\ & & & w_J^{-1} \end{pmatrix} \cdot \begin{pmatrix} & & & \\ & & \mathbf{U}^{\mathrm{T}} & \\ & & & \end{pmatrix} \cdot \begin{pmatrix} \\ \\ \mathbf{y} \\ \\ \\ \end{pmatrix} \tag{3.89}$$

3.5.4 Solution by use of SVD. Dealing with collinearity

Let us now come back to our χ^2 fitting problem (3.67), whose solution is given, via the SVD decomposition (3.82) of the $N \times J$ design matrix $\mathbf{X}$, by (3.89), with $\mathbf{y} = \hat{\mathbf{y}}$. Let the vectors $\mathbf{U}_{(j)}$ $(1 \le j \le J)$ be the columns of $\mathbf{U}$ (each such vector has the length $N \ge J$), and $\mathbf{V}_{(j)}$ $(1 \le j \le J)$ be the columns of $\mathbf{V}$ (each such vector has the length J). The solution (3.89) can then be written in the form:

$$\mathbf{a} = \sum_{j=1}^{J} \left(\frac{\mathbf{U}_{(j)} \cdot \hat{\mathbf{y}}}{w_j} \right) \mathbf{V}_{(j)}. \tag{3.90}$$

One can show that the standard uncertainties (standard deviations) of the estimated parameters are given, for the k-th component of $\mathbf{a}$, by

$$\sigma^2(a_k) = \sum_{j=1}^{J} \frac{1}{w_j^2} \mathbf{V}^2_{(j),k} = \sum_{j=1}^{J} \left(\frac{V_{kj}}{w_j} \right)^2, \tag{3.91a}$$

while the covariance between a_k and a_m is

$$\mathrm{Cov}(a_k,\, a_m) = \sum_{j=1}^{J} \left(\frac{V_{kj} V_{mj}}{w_j^2} \right). \tag{3.91b}$$

The above estimates of parameter uncertainties must be identical with (3.76), that is, variances and covariances of the parameters are the elements of the $(\mathbf{X}^{\mathrm{T}} \cdot \mathbf{X})^{-1}$.

Exercise 14. Substitute SVD decomposition of $\mathbf{X} \equiv \mathbf{U} \cdot \mathbf{W} \cdot \mathbf{V}^{\mathrm{T}}$ into $(\mathbf{X}^{\mathrm{T}} \cdot \mathbf{X})^{-1}$ and show that (3.91a), (3.91b) result. *Hint.* If $\mathbf{A}$ is an $N \times M$- and $\mathbf{B}$ is an $M \times K$-matrix (so that the product $\mathbf{A} \cdot \mathbf{B}$ is defined), then $(\mathbf{A} \cdot \mathbf{B})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}} \cdot \mathbf{A}^{\mathrm{T}}$.

We have seen in Section 3.5.2 that employing an additional predictor variable that happens to be highly correlated with one of the previously used predictors makes the design matrix nearly singular, resulting in possible instability of the multiple regression procedure. It is often not obvious if a certain predictor will be detrimental for the MLR, because its high correlation might be with some linear combination of previously used predictors, rather than with just one of them, with the same result of making the design matrix nearly singular. The presence of hidden linear dependencies between two or more of predictor variables is called *collinearity* or *multiple collinearity.*

What SVD does is in fact forming *orthogonal* linear combinations of predictors, whose contributions to reducing χ^2 are proportional to the associated singular values w_j. If some singular values are small (the condition number of the design matrix is large), a way to *regularize* nearly-singular regression problem is to edit these singular values, by replacing the corresponding factors $1/w_j$ in (3.90), (3.91a), and (3.91b) with zeros. This procedure of editing small singular values thus: (i) reduces uncertainty (and increases statistical significance) of estimated parameters; and (ii) produces nearly-minimal χ^2 by throwing away only those linear combinations of predictor variables that contribute little to reducing χ^2. This is called *principle component regression*, due to association of the SVD with the *principle component analysis* (also known as *empirical orthogonal function* (EOF) analysis; Chapter 4) — eigenanalysis of $\mathbf{X} \cdot \mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}} \cdot \mathbf{X}$.

Aside from the principle component regression, there are multitudes of regularization methods that deal with the problem of collinearity — from a naive *stepwise* regression, which tries out different linear combinations of predictor variables, ending up, iteratively, with the optimal set of predictors — to fairly sophisticated ones, such as *partial least-squares* (PLS) procedure. The latter method uses the principle component regularization (which only employs the information inherent in the design matrix; recall that the latter matrix is based on the *predictor* variables) to define the initial basis of orthogonal predictors, but then seeks linear combinations of basis vectors (or "*rotates*" principal components) in a way to ensure that rotated variables are maximally correlated with the response variable (predictand). The optimal number of initial principal components retained is determined by cross-validation (see Section 3.7), in which the regression model based (or "*trained*") on a part of the data set, is used to predict (or "*validated upon*") the remaining part of the data set, for a number of possible repartitions of the data set into training and validation segments.

3.6 Confidence limits on estimated model parameters

Denote the vector of "true" parameters of a linear regression model by $\mathbf{a}_{\text{true}}$; that is, we assume that there exists a true relationship

$$y = a_{\text{true},\,1}x_1 + a_{\text{true},\,2}x_2 + \ldots + a_{\text{true},\,\text{J}}x_J \tag{3.92}$$

between the observed variables y and $\mathbf{x} \equiv (x_1,\, x_2,\, \ldots\, ,\, x_J)$. We estimate the parameters $\mathbf{a}$ by applying multiple linear regression to N independent measurements of predictors $\mathbf{x}$ and predictand y. This procedure gives us a set of estimated parameters $\mathbf{a}_{(0)}$, which is in general different from $\mathbf{a}_{\text{true}}$ due to inherent unpredictability associated with randomness of "measurement" errors. If we had another realization of our observational data set and repeated the above analysis, we would end up with yet another estimate of parameters $\mathbf{a}_{(1)}$, and so on. Infinite number of samples of size N would supply us with the probability distribution of $\mathbf{a}_{(k)}$ (the mean of which would necessarily be equal to $\mathbf{a}_{\text{true}}$).

3.6.1 Monte Carlo simulations of synthetic data sets

Of course, we do not have an access to the infinite number of realizations of $\mathbf{a}_{(k)}$ — we just have one data set of size N, and one estimate of the parameters $\mathbf{a}_{(0)}$. However, if we have a guess about the process that produced our data set, we can generate an arbitrary number of synthetic realizations of this data set and estimate the distribution of the parameters about their "synthetic true mean" $\mathbf{a}_{(0)}$ computing, for each synthetic realization, its own set of parameters $\mathbf{a}_{(k)}$. If the way in which random errors enter the "experiment" and data analysis does not vary rapidly as a function of $\mathbf{a}_{\text{true}}$, our synthetic *Monte Carlo simulations* provide a numerical estimate of the distribution of $\mathbf{\Delta a} \equiv \mathbf{a}_{(k)} - \mathbf{a}_{\text{true}}$, from which we can make probabilistic statements about our estimated parameters; for example: "Is the slope of the χ^2-fit based on the 1900–1950 global temperature data set different from the one based on 1951–2000 portion of the global temperature record?" See *Exercise 15* of Section 3.7 for an example of Monte-Carlo-simulation-based analysis.

3.6.2 Constant chi-square boundaries as confidence limits

Let us summarize the general linear least-squares solution derivation in a slightly different way, following Box et al. (1994). The problem is, once again, given the expression

$$\hat{\mathbf{y}} = \mathbf{X} \cdot \mathbf{a} + \mathbf{e}, \tag{3.93}$$

where $\mathbf{X}$ is the design matrix of N weighted observations of predictor-variable vector of dimension J, and $\hat{\mathbf{y}}$ is the vector of N observations of the response variable, find the vector $\hat{\mathbf{a}}$ which minimizes the residual vector (or "*unexplained variance*") $\mathbf{e}$ (of dimension N):

$$S(\mathbf{a}) \equiv \mathbf{e}^{\mathrm{T}}\cdot\mathbf{e} = (\hat{\mathbf{y}} - \mathbf{X}\cdot\mathbf{a})^{\mathrm{T}}\cdot(\hat{\mathbf{y}} - \mathbf{X}\cdot\mathbf{a}); \tag{3.94}$$

Note that the quantity $S(\mathbf{a})$ is identical to what we have previously called the χ^2 merit functional.

Plugging the decomposition

$$\hat{\mathbf{y}} - \mathbf{X}\cdot\mathbf{a} = \hat{\mathbf{y}} - \mathbf{X}\cdot\hat{\mathbf{a}} - \mathbf{X}\cdot(\mathbf{a} - \hat{\mathbf{a}})$$

into (3.94) and choosing

$$(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})\cdot\hat{\mathbf{a}} = \mathbf{X}^{\mathrm{T}}\cdot\hat{\mathbf{y}} \tag{3.95}$$

(normal equations!), results in the following expression

$$S(\mathbf{a}) = S(\hat{\mathbf{a}}) + (\mathbf{a} - \hat{\mathbf{a}})^{\mathrm{T}}\cdot\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}\cdot(\mathbf{a} - \hat{\mathbf{a}}), \tag{3.96}$$

vectors $\hat{\mathbf{y}}-\mathbf{X}\cdot\hat{\mathbf{a}}$ and $\mathbf{X}\cdot(\mathbf{a}-\hat{\mathbf{a}})$ being orthogonal. The last term in (3.96) is a positive-definite quadratic form; it thus follows that the minimum of $S(\mathbf{a})$ is achieved at $\mathbf{a} = \hat{\mathbf{a}}$, defined by the normal equations (3.95) — yielding the same result we have previously derived by differentiating the merit functional with respect to regression parameters.

If the measurement errors are normally distributed, one can derive analytical distributions for the quadratic forms $S(\hat{\mathbf{a}})$ and $(\mathbf{a}-\hat{\mathbf{a}})^{\mathrm{T}}\cdot\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}\cdot(\mathbf{a}-\hat{\mathbf{a}})$. As we have mentioned before, the former is given by the χ^2 distribution with $N - J$ degrees of freedom. It turns out that the latter form is also χ^2-distributed, but with J degrees of freedom. These two properties allow us to use constant χ^2 boundaries as the confidence limits on the estimated model parameters. In fact, it is more convenient to use the statistic

$$\frac{(\mathbf{a} - \hat{\mathbf{a}})^{\mathrm{T}}\cdot\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}\cdot(\mathbf{a} - \hat{\mathbf{a}})}{S(\hat{\mathbf{a}})}\frac{N - J}{J}, \tag{3.97}$$

which is distributed as $F(J, N - J)$ [see (2.67)]. In particular, the inequality

$$\frac{(\mathbf{a} - \hat{\mathbf{a}})^{\mathrm{T}}\cdot\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X}\cdot(\mathbf{a} - \hat{\mathbf{a}})}{S(\hat{\mathbf{a}})}\frac{N - J}{J} \le F_{\alpha}(J, N - J) \tag{3.98}$$

defines $1 - \alpha$ confidence region for $\mathbf{a}$.

3.6.3 Confidence limits from SVD

The expression (3.98) for $J = 1$ defines an interval, for $J = 2$ — an ellipse, for $J = 3$ — an ellipsoid and so on. When the solution of the regression problem is written in terms of the SVD decomposition $\mathbf{X} = \mathbf{U} \cdot \mathbf{W} \cdot \mathbf{V}^{\mathrm{T}}$ of the design matrix $\mathbf{X}$, the above geometrical objects are given by the expression

$$w_1^2(\mathbf{V}_{(1)}\cdot\mathbf{\Delta a})^2 + \ldots + w_J^2(\mathbf{V}_{(J)}\cdot\mathbf{\Delta a})^2 = S(\hat{\mathbf{a}})F_\alpha(J, N-J)\frac{J}{N-J}, \tag{3.99}$$

where $\mathbf{\Delta a} \equiv \mathbf{a} - \hat{\mathbf{a}}$, and $\mathbf{V}_{(j)}$ is j-th column of $\mathbf{V}$ — this means that the columns of $\mathbf{V}$ are orthonormal vectors aligned with principal axes of J-dimensional ellipsoid defining $1 - \alpha$ confidence region for the estimated regression parameters.

3.7 Regression models as a means of forecasting

Forecast skill and rms error. Climatology, persistence, and damped persistence forecasts. Consider a forecast model that produces a large number of forecasts x_{f} of a quantity-of-interest x. For example, we initialize our model at some time $t = t_0$ using an observed value of $x_0 = x(t_0)$ and integrate it for $\tau = t_1 - t_0$ to get our forecast of the value of x at $t = t_1$: $x_{\mathrm{f}}(t_1)$. This procedure results in the forecast time series $x_{\mathrm{f}}(t)$, which should be compared with the actual observed evolution of $x(t)$ in order to make statements about how skillful our forecast model is; in particular, the term *forecast skill* relates to the correlation r between these two series. Another measure of how well the model performs in terms of forecasting is the root-mean-square (rms) error of our forecast time series relative to actual data:

$$\epsilon = \sqrt{\overline{(x - x_{\mathrm{f}})^2}}, \tag{3.100}$$

where the overbar denotes the time average.

The skill and rms errors are related. Suppose that our model is able to reproduce climatological statistics (the first two moments of "true" climate variability), as measured by $\bar{x}$, and $\overline{x'^2}$ (as before, the prime denotes the anomaly, or the deviation from the time average):

$$\bar{x}_{\mathrm{f}} = \bar{x}; \quad \overline{x_{\mathrm{f}}'^2} = \overline{x'^2}. \tag{3.101}$$

It then follows that

$$\epsilon^2 = \overline{(x' - x_{\mathrm{f}}')^2} = \overline{x'^2} - 2\overline{x'x_{\mathrm{f}}'} + \overline{x_{\mathrm{f}}'^2} = 2(\overline{x'^2} - \overline{x'x_{\mathrm{f}}'}). \tag{3.102a}$$

or, dividing through by $\overline{x'^2}$,

$$\frac{\epsilon^2}{\overline{x'^2}} = 2(1-r). \tag{3.102b}$$

The model has no skill if the anomaly time series x' and $x'_{\rm f}$ are uncorrelated ($r = 0$), in which case $\epsilon^2 = 2\overline{x'^2}$: the squared rms error is twice that of *climatological forecast*, in which $x_{\rm f}(t)$ is set to $\bar{x}$.

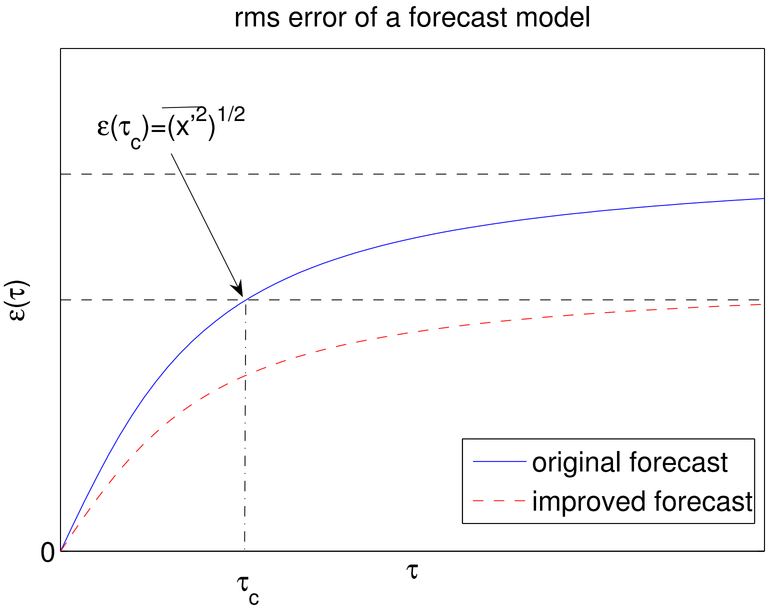


Figure 3.9: Verification of forecast models (see text for details).

One can thus judge how good the forecast model is by comparing its skill with that of climatological forecast (see Fig. 3.9). Here the forecasts are made, using a given model, from a number of observed initial conditions $x(t_0)$ (for a series of different t_0), to predict, for each t_0, the value of $x(t_0+\tau)$. Figure 3.9 shows the rms distance between the actual time series and hypothetical forecasts for different values of τ. As $\tau \to \infty$, the model's skill deteriorates and the rms distance tends to $\sqrt{2\overline{x'^2}}$, but before this happens, the model's rms curve passes, at $\tau = \tau_0$, the climatological forecast's rms, $\sqrt{\overline{x'^2}}$. The simple forecast $x_{\rm f}$ can be made better (denote the improved forecast by $\hat{x}_{\rm f}$), that is superior to climatology $\bar{x}$ for all τ, by using the regression model of the form

$$\hat{x}_{\rm f} = a x_{\rm f} + (1-a)\bar{x}, \tag{3.103}$$

and estimating a, for a given τ, by minimizing

$$\hat{\epsilon}^2 \equiv \overline{(x - x_\mathrm{f})^2}, \tag{3.104}$$

which results in

$$a = r(\tau) = \frac{\overline{x' x'_f}}{\overline{x'^2}} \tag{3.105}$$

(can you show this?). The dependency $\hat{\epsilon}(\tau)$ is plotted as a dashed curve in Fig. 3.9.

Another benchmark that the forecast models are usually compared against is the *persistence* forecast, in which $x_\mathrm{f}(t + \tau) = x(t)$. Obviously, the persistence forecast will be better than climatological forecast for short τ and worse than climatological forecast as τ becomes large. The persistence forecast improved according to (3.103), (3.105) is called the *damped persistence* forecast. **If we have a forecast model, we need to show that this model outperforms the damped persistence forecast in order to claim a useful skill.**

Statistical prediction and red noise. The linear regression techniques described in this chapter can be used for constructing an entirely data-based, predictive model of an observed phenomenon, by using past values of the observed variable to predict its future values. Consider, as an example, the anomaly time series (that is, average has been removed) of some quantity x and construct the model governed by

$$\hat{x}(t + \Delta t) = ax(t) + bx(t - \Delta t), \tag{3.106}$$

in which we are trying to predict the value of the variable at time $t+\Delta t$ using the information at the present time t and one time step into the past $t - \Delta t$. The model parameters a and b are obtained by minimizing rms distance between x and $\hat{x}$. Recall that in order to improve forecasting using the second predictor $x(t - \Delta t)$, compared with the one-predictor $(x(t))$ model, the correlation of this predictor with the response variable must exceed the minimum useful correlation (3.81). For the model (3.106), this is given by

$$|\overline{x(t + \Delta t)x(t - \Delta t)}| \geq |\overline{x(t + \Delta t)x(t)} \cdot \overline{x(t)x(t - \Delta t)}|/\overline{x'^2}, \tag{3.107a}$$

or

$$|r(2\Delta t)| \geq [r(\Delta t)]^2. \tag{3.107b}$$

If our time series is a *red-noise* process (Section 3.4.2), then the equality sign is realized in (3.107b) [compare with (3.46)], so that the value at two lags previous to now contributes exactly the minimum useful correlation, and there is no point in using a second predictor in this case. Our forecast skill as a function of the forecast period τ will be given by the autocorrelation (3.47) and will thus be equivalent to persistence forecast (since the autocorrelation can be computed by shifting a given time series by τ — we thus assume that $x(t+\tau)$ will be the same as $x(t)$).

Inverse stochastic models. Parametric linear least-squares can be used to construct more general forms of statistical forecast models to predict the evolution of a state vector $\mathbf{x}$ (of dimension I) describing some sub-component of the climate system. Consider an example of a quadratic regression model

$$dx_i = (\mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^{(0)}\mathbf{x} + {c_i}^{(0)})dt + d{r_i}^{(0)}\ 1 \le i \le I. \tag{3.108}$$

The matrices $\mathbf{A}_i$, the rows $\mathbf{b}_i^{(0)}$ of the matrix $\mathbf{B}^{(0)}$ and the components ${c_i}^{(0)}$ of the vector $\mathbf{c}^{(0)}$, as well as the components ${r_i}^{(0)}$ of the residual forcing $\mathbf{r}^{(0)}$, are determined by least-squares. The residual forcing is now considered as the part of the model rather than just an estimate of model errors. This noise models unresolved processes and is essential in energizing large-scale low-frequency variability we would like to model.

Our "observations" are typically not quite independent: the stochastic forcing $\mathbf{r}^{(0)}$ in Eq. (3.108) typically involves serial correlations and might also depend on the modeled process $\mathbf{x}$. One possible way of dealing with this problem is to include an additional model level to express the time increments $d\mathbf{r}^{(0)}$ (equivalent, in numerical practice, to the time derivative of the residual forcing $\mathbf{r}^{(0)}$) as a linear function of an extended state vector $[\mathbf{x},\, \mathbf{r}^{(0)}] \equiv (\mathbf{x}^{\mathrm{T}},\, \mathbf{r}^{(0)\mathrm{T}})^{\mathrm{T}}$, and estimate this level's residual forcing $\mathbf{r}^{(1)}$. The linear dependence is used since the non-Gaussian statistics of the data has already been captured by the first nonlinear level. More (linear) levels are being added in the same way, until the $(L+1)$-th level's residual $\mathbf{r}^{(L+1)}$ becomes white in time, and its lag-0 correlation matrix converges to a constant matrix:

$$\begin{array}{rcl}
dx_i & = & (\mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^{(0)}\mathbf{x} + c_i^{(0)})\, dt + r_i^{(0)}\, dt, \\
d{r_i}^{(0)} & = & \mathbf{b}_i^{(1)}[\mathbf{x},\, \mathbf{r}^{(0)}]dt + {r_i}^{(1)}\, dt, \\
d{r_i}^{(1)} & = & \mathbf{b}_i^{(2)}[\mathbf{x},\, \mathbf{r}^{(0)},\, \mathbf{r}^{(1)}]dt + {r_i}^{(2)}\, dt, \\
 & \dots & \\
d{r_i}^{(L)} & = & \mathbf{b}_i^{(L)}[\mathbf{x},\, \mathbf{r}^{(0)},\, \mathbf{r}^{(1)},\, \dots,\, \mathbf{r}^{(L)}]dt + d{r_i}^{(L+1)};\ 1 \le i \le I.
\end{array} \tag{3.109}$$

The convergence of this procedure is guaranteed since, with each additional level $l \ge 1$, we are accounting for additional time-lag information, thereby squeezing out any time correlations from the residual forcing.

In practice, we approximate the increments dx_i, $d{r_i}^{(l)}$ as

$$dx_i = x_i^{j+1} - x_i^j, \qquad d{r_i}^{(l)} = {r_i}^{(l),j+1} - {r_i}^{(l),j},\ 1 \le l \le L, \tag{3.110}$$

where j is the time index, while dt is assumed to be equal to the data set's sampling interval; without loss of generality, we use $dt = 1$. The last-level residual's $d{r_i}^{(L+1)}$ covariance matrix is estimated directly from its multivariate time series; in subsequent integrations of the inverse model, this forcing is approximated as a spatially correlated white noise.

One can in principle rewrite the multi-level system (3.109) as a single equation that involves time-lagged values of x_i and ${r_i}^{(l)}$; the resulting construct is equivalent to a multivariate version of autoregressive–moving average (ARMA) model (Box et al. 1994), except for the nonlinear dependence on x_i that we allow here, and which is not present in standard ARMA models. Even for a standard, linear model, though, the way we estimate the coefficients of this model by successive introduction of additional levels is algorithmically simple, numerically efficient and dynamically transparent. The system (3.109) describes a wide class of nonlinear, non-Gaussian processes in a fashion that explicitly accounts for the modeled process $\mathbf{x}$ feeding back on the noise statistics.

The optimal number of state-vector components in Eq. (3.109) is assessed in practice using Monte-Carlo simulations: in these cross-validation tests, the inverse model is trained on one segment of the available data and is then used to estimate the properties of the model evolution during the validation interval. The measure used to assess the statistical model's performance depends on the purpose at hand: If the model is to be used for prediction, the forecast skill, quantified by the correlation between the forecast and observed fields or the root-mean-square (rms) distance between the two is an appropriate measure of model performance; in the more theoretical applications below, it is the statistical characteristics of the observed and modeled evolution, such as PDFs of model variables (see Chapter 5) and their power spectra (Chapter 6).

We can test this procedure and learn how to apply regression techniques and concepts discussed in the present chapter by doing the following

Exercise 15. Consider the monthly Niño-3 index time series $x(t)$ (Fig. 1.1) [seasonal cycle has been removed]. Call $\Delta x(t)$ the time series of differences between two consecutive values of this index.

- Fit a linear regression model to express $\Delta x(t)$ via $x(t)$. Plot the rms distance and forecast skill of this model as a function of $\tau(= 1, 2, \dots, 12$ months).

- Form now the series of differences between consecutive $\Delta x(t)$ and construct a two-level regression model. How do the skill and rms error of this model compare to those of our first model?

- Continue on adding levels in the same way. How does the skill and rms error change?

- Consider the case of polynomial predictors 1, x, x^2, ..., x^J and fit such a polynomial regression model to predict Δx for $J = 1, 2, 3, 4, 5$. Compute forecast skill and rms errors of these models and compare them with previous models' values. Do you encounter instabilities in any of your integrations?

- Add to polynomial regression models above more linear levels, as before, and repeat the analysis. Compare across all models.

- Seasonal cycle (ENSO is known to be largely locked to the seasonal cycle). Include, at the first level of each regression model, two more predictors $\cos(2\pi t/T)$ and $\sin(2\pi t/T)$ ($T = 12$ months). Repeat the analysis for each regression model you have constructed. Plot the skill and rms error of each model as a function of the calendar month.

- Enter cross validation: divide the time series into several intervals (10–11-year long). Throw away the data from one of the intervals and train the regression models above on the remaining data. Use this model to forecast the variability in the omitted time segment. Repeat this procedure with all pairs of training/validation periods. Plot the cross-validated skills and rms errors for each model and compare them with your previous *hindcasts* (predictions without cross validation, in which training and validation intervals coincide).

How would you go about estimating the uncertainties of your models' coefficients using Monte Carlo integrations? Compute the uncertainties of the coefficients in two of the above cases.

References

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate*, **12**, 1990–2009.

Box, G. E. P., G.M. Jenkins, and G.C. Reinsel, 1994: *Time Series Analysis, Forecasting and Control.* Prentice Hall, Englewood Cliffs, NJ, 3rd edition, 592pp.

Da Costa, E., and R. Vautard, 1997: A qualitative realistic low-order model of the extratropical low-frequency variability built from long records of potential vorticity. *J. Atmos. Sci.*, **54**, 1064–1084.

Daoud, W. Z., J. D. W. Kahl, and J. K. Ghorai, 2003: On the synoptic-scale Lagrangian autocorrelation function. *J. Appl. Meteor.*, **42**, 318–323.

DelSole, T., 1996: Can quasigeostrophic turbulence be modeled stochastically? *J. Atmos. Sci.*, **53**, 1617–1633.

DelSole, T., 2000: A fundamental limitation of Markov models. *J. Atmos. Sci.*, **57**, 2158–2168.

Dillon, W. R., and M. Goldstein, 1984: *Multivariate Analysis: Methods and Applications.* Wiley and Sons, 587pp.

Draper, N. R., and H. Smith, 1966: *Applied Regression Analysis.* Wiley and Sons, New York, 407pp.

Hand, D., H. Mannila, and P. Smyth, 2001: *Principles of Data Mining.* MIT Press, Cambridge, MA, 546 pp.

Höskuldsson, A., 1996: *Prediction Methods in Science and Technology.* Thor Publishing, Denmark.

Hsieh, W. W., and B. Tang, 1998: Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Am. Meteorol. Soc.*, **79**, 1855–1870.

Huff, D., 1954: *How to Lie with Statistics.* Norton and Co., New York, 142pp.

Johnson, S. D., D. S. Battisti, and E. S. Sarachik, 2000: Empirically derived Markov models and prediction of tropical Pacific sea surface temperature anomalies. *J. Climate*, **13**, 3–17.

Kondrashov, D., S. Kravtsov, A. W. Robertson, and M. Ghil, 2005: A hierarchy of data-based ENSO models. *J. Climate*, accepted.

Larson, R. L., and M. L. Marx, 1986: *An Introduction to Mathematical Statistics and its Applications.* 2nd edition, Prentice–Hall, Englewood Cliffs, N. J., 630pp.

Leith, C. E., 1973: The standard error of time-averaged estimates of climatic means. *J. Appl. Meteorol.*, **12**, 1066–1069.

McCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models.* Chapman and Hall, 511 pp.

Navone, H. D., and H. A. Ceccatto, 1994: Predicting Indian monsoon rainfall–a neural network approach. *Clim. Dyn.*, **10**, 305–312.

Noble, B., and J. W. Daniel, 1988: *Applied Linear Algebra.* Englewood Cliffs, Prentice-Hall, 521pp.

Panofsky, H. A., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology.* Pennsylvania State University, University Park, 224pp.

Penland, C., 1989: Random forcing and forecasting using principal oscillation pattern analysis. *Mon. Wea. Rev.*, **117**, 2165–2185.

Penland, C., 1996: A stochastic model of Indo-Pacific sea-surface temperature anomalies. *Physica D*, **98**, 534–558.

Penland, C., and M. Ghil, 1993: Forecasting Northern Hemisphere 700-mb geopotential height anomalies using empirical normal modes. *Mon. Wea. Rev.*, **121**, 2355–2372.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea-surface temperature anomalies. *J. Climate*, **8**, 1999–2024.

Penland, C., and L. Matrosova, 1998: Prediction of tropical Atlantic sea-surface temperatures using linear inverse modeling. *J. Climate*, **11**, 483–496.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1994: *Numerical Recipes.* 2-nd edition. Cambridge University Press, 994 pp.

Roebber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas Jr., 2003: Improving snowfall forecasting by diagnosing snow density. *Weather and Forecasting*, **18**, 264–287.

Spiegel, M. R., 1961:*Statistics.* Schaum's Outline Series in Mathematics, New York, McGraw Hill, 359pp.

Strang, G., 1988: *Linear Algebra and Its Applications.* 3rd edition, Harcourt Brace, 505pp.

Von Mises, R., 1964: *Mathematical Theory of Probability and Statistics.* Academic Press, New York.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the Geopotential Height Field during the Northern Hemisphere winter. *Mon. Wea. Rev.*, **109**, 784–812.

Wetherill, G. B., 1986: *Regression Analysis with Applications.* Chapman and Hall, 311 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* (International Geophysics Series, v. 59), Academic Press, San Diego, 467pp.

Winkler, C. R., M. Newman, and P. D. Sardeshmukh, 2001: A linear model of wintertime low-frequency variability. Part I: Formulation and forecast skill. *J. Climate*, **14**, 4474–4494.

Wold, S., A. Ruhe, H. Wold, and W. J. Dunn III, 1984: The collinearity problem in linear regression: The Partial Least Square approach to generalized inverses. *SIAM J. Sci. Stat. Comp.*, **5**, 735–743.

Yuval, and W. W. Hsieh, 2002: The impact of time-averaging on the detectability of nonlinear empirical relations. *Q. J. R. Meteorol. Soc.*, **128**, 1609–1622.

Zwiers, F. W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate*, **8**, 336–351.

Chapter 4

Matrix Methods for Analysis of Structure in Data Sets

4.1 Introduction to matrix methods

4.2 Empirical Orthogonal Function (EOF)/Principal Component (PCA) Analysis

4.2.1 Introduction to EOF analysis

4.2.2 EOFs as efficient representations of data sets

4.2.3 Manipulation of EOFs and PCs

4.2.4 Scaling and display of EOFs and PCs

4.2.5 EOF analysis via SVD of the input data matrix

4.2.6 Statistical significance of EOFs

4.2.7 Interpretation of EOFs. How large should the domain size be?

4.2.8 Rotation of EOFs

4.2.9 Variations and applications of EOF analysis

4.3 Maximum Covariance Analysis (MCA) and Canonical Correlation Analysis (CCA)

4.3.1 MCA formalism

4.3.2 Scaling and display of singular vectors

4.3.3 Statistical significance of MCA analysis

4.3.4 MCA analysis of unrelated fields

4.3.5 Criticisms of MCA Analysis

4.3.6 Canonical Correlation Analysis

4.3.7 Applications of MCA and CCA Analyses

References

Barnett, T. P., and R. W. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825–1850.

Branstator, G., 1987: A striking example of the atmospheric leading traveling pattern. *J. Atmos. Sci.*, **44**, 2310–2323.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data sets. *J. Climate*, **5**, 541–560.

Cheng, X. H., and J. M. Wallace, 1993: Analysis of the northern-hemisphere wintertime 500-hPa height field spatial patterns. *J. Atmos. Sci.*, **50**, 2674–2696.

Cherry, S., 1996: Singular value decomposition analysis and canonical correlation analysis. *J. Climate*, **9**, 2003-2009.

Cherry, S., 1997: Some comments on singular value decomposition analysis. *J. Climate*, **10**, 1759-1761.

D'Andrea, F., 2002: Extratropical low-frequency variability as a low-dimensional problem. Part II: Stationarity and stability of large-scale equilibria. *Q. J. R. Meteorol. Soc.*, **128**, 1059–1073.

D'Andrea, F., and R. Vautard, 2001: Extratropical low-frequency variability as a low-dimensional problem. Part I: A simplified model. *Q. J. R. Meteorol. Soc.*, **127**, 1357–1374.

Deser, C., 2000: On the teleconnectivity of the "Arctic Oscillation." *Geophys. Res. Lett.*, **27**, 779–782.

Dillon, W. R., and M. Goldstein, 1984: *Multivariate Analysis: Methods and Applications.* Wiley and Sons, 587pp.

Farrell, B. F., and P. J. Ioannou, 1993: Stochastic forcing of the linearized Navier-Stokes equations. *Phys. Fluids A*, **5**, 2600–2609.

Farrell, B. F., and P. J. Ioannou, 1995: Stochastic dynamics of the midlatitude atmospheric jet. *J. Atmos. Sci.*, **52**, 1642–1656.

Fraedrich, K., C. Ziehmann, and F. Sielmann, 1995: Estimates of spatial degrees of freedom. *J. Climate*, **8**, 361–369

Franzke, C., A. J. Majda, and E. Vanden-Eijnden, 2005: Low-order stochastic mode reduction for a realistic barotropic model climate. *J. Atmos. Sci.*, **62**, in press.

Horel, J. D., 1981: A rotated principal component analysis of the interannual variability of the Northern Hemisphere 500-mb height field. *Mon. Wea. Rev.*, **109**, 2080–2092.

Horel, J. D., 1984: Complex principal component analysis: Theories and examples. *J. Appl. Meteorol.*, **23**, 1660–1673.

Hu, Q., 1997: On the uniqueness of the singular value decomposition in meteorological applications. *J. Climate*, **10**, 1762-1766.

Jin, S.-X., and H. von Storch, 1990: Predicting the state of the Southern Oscillation using principal oscillation pattern analysis. *J. Climate*, **3**, 1316–1329.

Jolliffe, I. T., 2002: *Principal Component Analysis.* Springer, 2nd edition, 487pp.

Kutzbach, J. E., 1967: Empirical eigenvectors of sea-level pressure, surface temperature, and precipitation complexes over North America. it J. Appl. Meteorol., **6**, 791–802.

Morrison, D. F., 1976: *Multivariate Statistical Methods.* McGraw–Hill.

Mundt, M. D., and J. E. Hart, 1994: Secondary instability, EOF reduction, and the transition to baroclinic chaos. *Physica D*, **78**, 65–92.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699–706.

North, G. R., 1984: Empirical orthogonal functions and normal modes. *J. Atmos. Sci.*, **41**, 879–887.

Newman, M., and P. D. Sardeshmukh, 1995: A caveat concerning singular value decomposition. *J. Climate*, **8**, 352–360.

Overland, J. E., and R. W. Preisendorfer, 1982: A significance test for principal components applied to cyclone climatology. *Mon. Wea. Rev.*, **110**, 1–4.

Preisendorfer, R. W., 1988: *Principal Component Analysis in Meteorology and Oceanography.* Elsevier, New York, 425 pp.

Prohalska, J., 1976: A technique for analyzing the linear relationships between two meteorological fields. *Mon. Wea. Rev.*, **104**, 1345–1353.

Richman, M. B., 1986: Rotation of principal components. *J. Climatology*, **6**, 293–335.

Rinne, J., and V. Karhila, 1975: A spectral barotropic model in horizontal empirical orthogonal functions. *Quart. J. Roy. Meteor. Soc.*, **101**, 365–382.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**, 853–871.

Schubert, S. D., 1985: A statistical-dynamical study of empirically determined modes of atmospheric variability. *J. Atmos. Sci.*, **42**, 3–17.

Selten, F. M., 1995: An efficient description of the dynamics of the barotropic flow. *J. Atmos. Sci.*, **52**, 915–936.

Selten, F. M., 1997: Baroclinic empirical orthogonal functions as basis functions in an atmospheric model. *J. Atmos. Sci.*, **54**, 2100–2114.

Sirovich, L., and J. D. Rodriguez, 1987: Coherent structures and chaos – a model problem. *Phys. Lett.*, **120**, 211–214.

Strang, G., 1988: *Linear Algebra and Its Applications.* 3rd edition, Harcourt Brace, 505pp.

Vimont, D., 2002: The seasonal footprinting mechanism in the Pacific: Implications for ENSO. *J. Climate*.

Von Storch, H., and F. Zwiers, 1999: *Statistical Analysis in Climate Reserach.* Cambridge University Press, Cambridge, United Kingdom, 484pp.

Von Storch, H, G. Bürger, R. Schnur, and J.-S. von Storch, 1995: Principal oscillation patterns: A review. *J. Climate*, **8**, 377–400.

Wallace, J. M., 2000: North Atlantic Oscillation/annular mode: Two paradigms — one phenomenon. *Quart. J. Roy. Meteor. Soc.*, **126**, 791–805.

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**, 561–576.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* International Geophysics Series, v. 59), Academic Press, San Diego, 467pp.

Chapter 5

Probability Density Estimation. Compositing. Cluster Analysis

Chapter 6

Spectral Methods for Time Series Analysis. Filtering of Time Series

Chapter 7

Recapitulation