ADVANCED MODERN MATHEMATICAL MODELING

ADVANCED MODERN MATHEMATICAL MODELING

-A Data-driven Computational Approach

Samuel S.P. Shen
San Diego State University

WILEY-INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

CONTENTS

FOREWORD

This is primarily a graduate mathematical modeling textbook, first taught at San Diego State University in Fall 2016, but can also be used as a handbook for research in mathematics and statistics applications to natural sciences, engineering, social sciences, and applied mathematics itself. It includes basic skills of applied mathematics consulting, such as dimensional analysis for exploring possible relationships among the relevant variables, R programs for basic statistics, linear algebra, and space-time plotting, and PAMMI (problem identification, abstract of the problem, model formulation, model solution, and interpretation of the modelings results) five steps PAMMI approach to mathematical modeling principles.

The prerequisite for this course are Calculus I-III, one semester of linear algebra, and one semester of ODE (ordinary differential equations), and some basic knowledge of complex variables and infinite series. Thus, students are assumed to know gradient, divergence theorem, vector calculus, partial derivatives, polar coordinates, cylindrical coordinates, spherical coordinates, eigenvalues and eigenvectors of square matrices, dot products, and cross product, and SVD(singular value decomposition) of a rectangular matrix. However, this text still reviews these concepts, definitions, critical formulas and main theorems of the above topics when being used, since some students are rusty with the calculations and have never not fully understood the meaning o the mathematical results anyway.

Topics covered in this book includes matrix models, probability models, regression models, ordinary differential equation models (ODE), partial differential equation (PDE) models, stochastic models, machine learning models, big data models, dimensional analysis, and R programs. R programming is taught in class from beginning. R is free for public download and can be installed easily for either PC or Mac. Computer programming experience is not required to read this book.

This book is an advanced version of my earlier book entitled "Introduction to Modern Mathematical Modeling", which was written for undergraduate students and also a tool book for scientific research and mathematical consulting business. Here, "Modern" means the book being different from the traditional modeling books in the last 50 years. Traditional modeling books focus on deterministic modeling, model setup, and differential equations, and pays little attention to observed data, model validation, intensive computing, and numerical solutions. Our "modern" modeling emphasizes data-driven problems, pays much attention to uncertainties in both observed data and model equations, includes stochastic models and discrete models, care about model interpretation and validation, and blends together mathematics, computing and statistics with at least one specific application domain, such as climate science, artificial intelligence, chemistry, and mechanical engineering. A goal of the Modern Mathematical Modeling class is to enable students to start a applied mathematics consulting business after the one-semester training, since they have learned all the five PAMMI steps and the format of writing consulting reports in the process of writing term papers.

"Advanced" is relative to my earlier book"Introduction to Modern Mathematical Modeling." The "Advanced" book is intended for graduate students and researchers and includes more advanced topics of mathematics, such as PDE's initial-boundary-value problems, nonlinear dynamics and similarity, R graphics and visualization of big space-time, and stochastic models. The spatial scales between microscopic scales to thousands kilometer scales can be between 1.0 [km] and 10^3 and can be modeled by fractal dimensions and nonlinear similarity principles. Expectation and maximization (EM) algorithm is used as an effective way to do optimization when data are incomplete, which is a key algorithm in modern internet surfing designs and in artificial intelligence (AI) and data mining (DM). Cloud computing gives a convenient and easy way to perform supercomputing tasks without accessing supercomputers. Monte Carlo Markov chain (MCMC) effectively models discrete dynamical systems with observed data, compared with the conventional PDE or ODE modelings where the observed data on discrete points and the DE's output on grid boxes often do not match. Stochastic modeling allows nowhere differentiable paths which can be continuous or discontinuous, and hence can be applied to modeling fair stock markets, fractal clouds in the sky, and storm rainfalls.

This book wishes to show that mathematical model can be useful to any discipline and is a key tool to promote effective collaborations among different disciplines for discoveries and method developments. We will show numerous examples of calculus, linear algebra, DEs, complex variables, functional analysis, statistics, big data, signal analysis, computer programming, economics, electric engineering, chemistry, meteorology, oceanography, and other application areas. Through the PAMMI modeling procedures, we emphasize discovery via results finding and interpretation. We also show students how to write short modeling proposals and complete project reports based on mathematical modeling approaches.

–By SS in San Diego, Fall 2016

GLOSSARY

PART I

-DIMENSIONAL ANALYSIS
-R PROGRAMMING
-SPACE-TIME DATA AND THEIR SVD

CHAPTER 1

DIMENSIONAL ANALYSIS –A SHORTCUT TO DISCOVER LAWS OF NATURE

This chapter shows a way to discover law of nature, such as kinetic energy and potential energy exchange, period of simple pendulum, using dimensional analysis, and Newton's second law of motion. The laws may be direct results of the two sides of an equation to have the same dimension or units or the results from reorganizing the dimensions of a given parameter.

—Summary

1.1 Dimensions and units

Length is called the dimension of a line, which is denoted by L. Length can be measured in SI Units: meter, or Imperial Units: feet. The SI is for French words "Systeme international d'unites", i.e., the International System of Units. SI system is also known as the metric system. Its commonly used length units are $m, dm, cm, mm, \mu m, km$; time units: $sec, \mu s$; and mass units: g, kg. The corresponding imperial system are $feet,\ lb,\ sec$. The imperial units is a British system. The SI system was published in 1960, and is now the most popular units system used in science and engineering around the world. The United States is the only major country that is still using the imperial units in engineering, but most science publications in the U.S. have adopted the metric system. The United Kingdom had adopted the metric system in the 1960s.

Any given climate parameter has a dimension and units. The dimension is an intrinsic property of the parameter, such as wind speed as [LT^{-1}], which is applicable to speeds of all kinds of variables, such as the speed of a hurricane eye, speed of California current,

and speed of a river flow. The units, on the other hand, are a description of the dimension with given scales, such as using kilometers to describe horizontal size of the Gulf Stream, but using meters to describe the amplitude of oceanic surface waves. So, a dimension can have many different kinds of units. All the units can be converted to each other with fixed formulas, such as one km = 1,000 meters.

Systematic use of units is very important. Misuse can can have serious consequences. On September 30, 1999, CNN reported that NASA lost a $125 million Mars orbiter because an engineering team mixed the SI system with the imperial system:
`http://www.cnn.com/TECH/space/9909/30/mars.metric.02/`

The units originated with culture and are external properties. The dimension and the laws of nature are intrinsic properties and should be independent of units. $F = ma$ works for both imperial and metric systems. Thus, it is critical that a law of nature is expressed in a single units system, not a mixture of two systems.

1.2 Fundamental physics dimensions: $LMT\theta I$-class

The fundamental dimensions of climate science are the five listed in Table 1.1.

Table 1.1 Fundamental dimensions: $LMT\Theta I$-class

Notation	Meaning	Dimension	Units
$[l]$	Length	L	m
$[m]$	Mass	M	kg
$[t]$	Time	T	sec or s
$[\theta]$	Temperature	Θ	$^\circ K$
$[I]$	Electric current (i.e., the flow flux of electric charges)	I	Amp or A

The dimensions of most other physical quantities can be derived from the above five. For example, speed is the displacement in a unit time and has is dimension LT^{-1}. Table 1.2 shows dimensions of the commonly used climate parameters.

Table 1.2 lists the dimensions of a few commonly used physical quantities.

We normally use square brackets $[\cdot]$ to denote dimension. If v is speed, then [v]= LT^{-1}.

EXAMPLE 1.1

Dimensional analysis of the geo-potential energy: The geo-potential energy of an air mass m at height h is

$$E = mgh. \tag{1.1}$$

Following Table 2, we have

$$[E] = [m][g][h] = M(LT^{-2})L = ML^2T^{-2}. \tag{1.2}$$

The last expression can be further organized into $M(LT^{-1})^2$, hence mass times speed squared, which has a clear physical meaning: kinetic energy $(1/2)mv^2$. This simple analysis links the potential energy and kinetic energy, and helps one to think about the conversion of potential energy into kinetic energy, such as the wind caused by geo-potential differences or air pressure differences.

Table 1.2 Dimensions of derived physical quantities:

	Meaning	Dimension	SI Units
$[v]$	Velocity	LT^{-1}	m/s
$[a]$	Acceleration	LT^{-2}	m/s^2
$[F]$	Force (F=ma)	MLT^{-2}	$N = 1.0kg \cdot m/s^2$
$[\rho]$	Mass density	ML^{-3}	kg/m^3
$[p]$	Pressure (force per area)	$MLT^{-2}L^{-2} = ML^{-1}T^{-2}$	$Pa = N/m^2$
$[E]$	Energy	ML^2T^{-2}	$Joule = 1.0N \cdot m$
$[S]$	Entropy (energy per K)	$ML^2T^{-2}\Theta^{-1}$	$Joule/^\circ K$
$[Q]$	Electric charge	IT	$C = 1.0A \cdot s$
$[E]$	Electric field (force per C)	$NC^{-1} = MLT^{-3}I^{-1}$	v/m
$[B]$	Magnetic field	$N(IL)^{-1} = MT^{-2}I^{-1}$	$T = 1.0kg/(As^2)$
$[\phi]$	Angle	1(dimensionless)	$radian$

Hence, a simple algebraic manipulation of a dimension formula may lead to a new physical phenomenon, such as the phenomena of potential energy and kinetic energy. Mathematical manipulations of such are all regarded as the category of dimensional analysis, which may lead to discoveries of new physical laws and relationships shown in the examples below.

EXAMPLE 1.2

Dimensional analysis of π: The constant π is a ratio of circumference to diameter $\pi = C/D$ for any circle. Thus

$$[\pi] = L/L = 1 \tag{1.3}$$

is dimensionless. π measures the angle of 180° is thus also dimensionless. Any angle can be measured by π or degree and is thus dimensionless, i.e., non-dimensional. The trigonometric functions, logarithmic functions, and exponential functions can only be applied to dimensionless quantities, such as 0.5π, or 2.3, or 1.0. These are pure numbers, but can also be regarded as radians, measuring an angle. However, radian is not a dimension. Of course, the range of trigonometric functions, logarithmic functions is also dimensionless. In the expressions $y = \sin x, y = \ln x, y = \exp(x)$, both x and y are dimensionless. In $\sin(\pi/6) = 0.5$, $\pi/6$ radian is considered dimensionless since radian is dimensionless, and 0.5 is also dimensionless.

Because of this common dimensionless feature of trigonometric functions, logarithmic functions and exponential functions, one may think that these functions should be related. Yes, they are. The exponential function and trigonometric functions are related by

$$e^{i\phi} = \cos\phi + i\sin\phi. \tag{1.4}$$

This is usually called Euler's formula (Leonhard Euler, 1707-1783, Swiss mathematician), illustrated by Fig. 1.1. Physics Nobel laureate Richard Feynman called Euler's equation"the most remarkable formula in mathematics." This equation can help express numerous physical properties, such as wave function in quantum mechanics, homogeneity in universe, water waves, and alternative electric current.

The length of the arc of an interior angle θ and radius r is

$$s = \theta r. \tag{1.5}$$

The dimension of the above equation is

$$[s] = [\theta][r], \tag{1.6}$$

which is

$$L = [\theta]L. \tag{1.7}$$

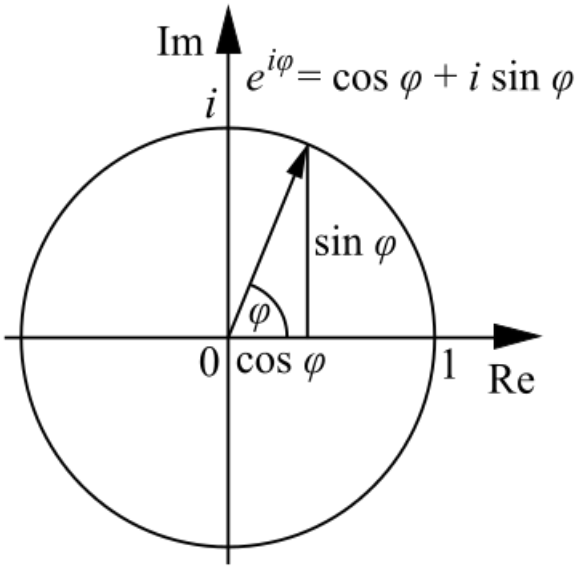

Figure 1.1 Euler's formula.

Hence, $[\theta] = 1$ is dimensionless. This is another way to illustrate that angle is dimensionless, although we customarily use radian or degree to measure an angle. Neither radian nor degree for an angle should be considered dimensional.

The logarithmic function is the inverse function of exponential function. Other trigonometric functions can be derived from cosine and sine functions.

EXAMPLE 1.3

Dimensional analysis of pressure: The pressure's dimension is $[p] = ML^{-1}T^{-2}$ from Table 1.2 and can be re-written as $M(LT^{-1})^2L^{-3}$. Because $M(LT^{-1})^2$ is kinetic energy, $M(LT^{-1})^2L^{-3}$ is thus the kinetic energy per unit volume. This is the definition of pressure from the thermodynamic point of view. An ideal gas' pressure on its container wall is measured by the strength of the gas' kinetic energy per volume.

Thus, the rearrangement of different dimensions can result in very interesting and profound laws of nature. Dimensional analysis provides a powerful tool for discovery. We may say that dimensional analysis is a shortcut for discovery and can simplify experiments that lead to many useful mathematical formulas for climate science, and nature in general.

The pressure dimension can be written in another form

$$[p] = ML^{-1}T^{-2} = (MLT^{-2})L^{-2}. \tag{1.8}$$

This gives another physical meaning. Since LT^{-2} is acceleration, MLT^{-2} is thus ma the force, following Newton's second law. Thus, $(MLT^{-2})L^{-2}$ means the force on a unit area, which is exactly the physical definition of pressure.

According to this definition of pressure, the atmospheric pressure at sea level can be defined as the total gravitational force per unit area acting on the ocean water surface:

$$p = \frac{\int_0^\infty g\rho(z,\phi,\theta,t)dzA}{A}, \tag{1.9}$$

where A is the cross section area of the air mass column from sea level to infinite height, g is the gravitational acceleration, ϕ is latitude, θ is latitude, and t is time. A simplification of the above becomes

$$p(\phi,\theta,t) = g\int_0^\infty \rho(z,\phi,\theta,t)dz. \tag{1.10}$$

This is the precise definition of the sea level pressure (SLP), which is a very important parameter for weather forecasting and is a function of latitude, longitude and time.

1.3 Dimensional analysis for a simple pendulum and sinusoidal waves

Simple pendulum clocks are based on the mechanism of simple pendulum oscillation. For a clock, its most important function is to record time measured as how many oscillation periods. A pendulum period is given by

$$P = 2\pi\sqrt{l/g} \tag{1.11}$$

where l is the length of the string and g is the gravitational constant. This formula can be derived using many methods, including an approach of the second order ordinary differential equation. Here we provide a simple approach via dimensional analysis. A pendulum involves three quantities: mass of the pendulum, length of the string, and the Earth gravity. We may assume that the pendulum's oscillation period depends on these three quantities. To understand the gravity dependence, one can think of an extreme environment: outpace with zero gravity, where the pendulum will not oscillate because of absence of gravity. The period is thus infinity. Similarly one may reasonably conclude that the same pendulum oscillates slower on moon than on Earth, because moon has a smaller gravity force.

Thus, we can assume that the pendulum's period depends on mass, length and gravity, written in the following form:

$$P = \alpha m^a l^b g^c \tag{1.12}$$

with the exponent a, b, c to be determined by using the five fundamental dimensions:

$$[P] = [\alpha][m]^a[l]^b[g]^c = M^a L^b (LT^{-2})^c = M^a L^{b+c} T^{-2c}. \tag{1.13}$$

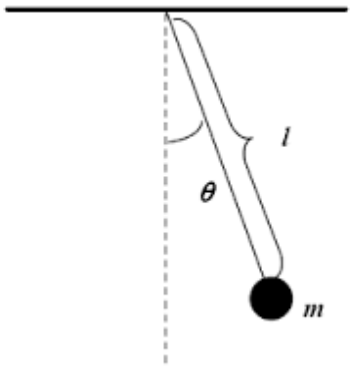

Figure 1.2 Simple pendulum of mass m and length l.

This implies

$$a = 0, \tag{1.14}$$
$$b + c = 0, \tag{1.15}$$
$$-2c = 1. \tag{1.16}$$

These equations have the following solutions

$$c = -1/2, b = 1/2, a = 0. \tag{1.17}$$

The period is proportional to $m^0 l^{1/2} g^{-1/2}$, or

$$P = \alpha m^0 l^{1/2} g^{-1/2} = \alpha\sqrt{l/g}. \tag{1.18}$$

An experiment was conducted in classroom with a string's length equal to 0.88 meters. Two periods were observed with time equal to 3.75 seconds. Substitute this into the above equation:

$$3.75 = 2 \times \alpha\sqrt{0.88/9.8} = 0.60\alpha, \tag{1.19}$$

$$\alpha = 3.75/0.60 = 6.25 = 1.99\pi = \approx 2\pi. \tag{1.20}$$

This is an easy experiment. Since the motion is relatively slow when the string is long enough, with smartphone stopwatch, it is fairly easy to record the time of two or three periods. One can improve the experimental results by making many experiments and use the average results as the final value for α.

1.4 Relationships between magnetic and electric fields

From Table 1.2, the dimension of electric field is $MLT^{-3}I^{-1}$ which can be re-written as $(MT^{-2}I^{-1})(LT^{-1})$, whose first part is magnetic field and second part is velocity. When a conductor moves inside a magnetic field and cut the magnetic lines of force, an electric field can be felt because of the resistance exerted on the conductor, consequently, an electric current is generated and flows through the conductor (see

Fig. 1.3). This is the principle of a power generator. Thus, dimensional analysis helps identify relevant physical quantities and can aid us to find new laws of physics, i.e., mathematical models for the physical quantities.

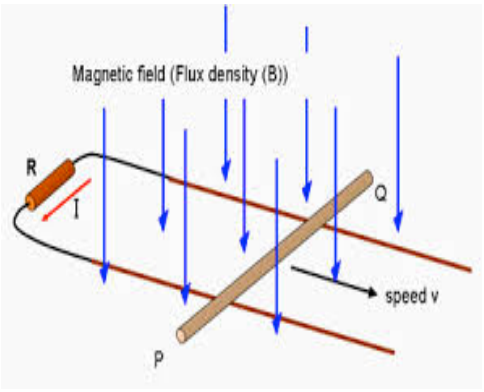

Figure 1.3 Generation of electric current when a conductor moves through a magnetic field and cuts the magnetic lines of force.

In general, a physical quantity X can be written as

$$X = \alpha l^a t^b m^c \theta^d I^e, \tag{1.21}$$

where α is a dimensionless constant. The dimensional analysis equation

$$[X] = \alpha [l]^a [t]^b [m]^c [\theta]^d [I]^e = \alpha L^a T^b M^c \Theta^d I^e \tag{1.22}$$

leads to linear equations for the exponents a, b, c, d, e. Solving these equations, one can obtain values of a, b, c, d, e. Equation (1.21) is then a mathematical model of a physics law, such as $(1/2)gt^2$ for the travelled distance of a free-fall body. Equation (1.22) is a special case of the general Buckingham's Π-theorem, which can be found in more detailed dimensional analysis books (Barenblatt 1987).

Unfortunately, dimensional analysis still cannot determine the value of α, which can be determined by an experiment or other mathematical approaches.

1.5 Dimensional analysis example: state of air

A basic thermodynamics relation is state equation of air that relates pressure p, temperature T, volume V, and the amount of gas n:

$$pV = nRT, \tag{1.23}$$

where R is called the specific gas constant. The dimension of R is

$$\begin{aligned} [R] &= [p][v]/([n][T]) \\ &= (MT^{-2}L^{-1})(L^3/[n])(\Theta^{-1}) \\ &= ML^2T^{-2}\Theta^{-1}[n]^{-1} \\ &= M(L/T)^2\Theta^{-1}[n]^{-1}. \end{aligned} \tag{1.24}$$

The factor $M(L/T)^2$ is mass times velocity and is hence energy. We can use joule as the units of energy, Kelvin degree as the units of temperature, and moles as the units of gas amount. Thus, the gas constant is energy per degree temperature per unit mass. The gas constant R is

$$R = 8.314[JK^{-1}mol^{-1}] \tag{1.25}$$

One mole's mass may be measured by units [g/mol], called molecular mass. Dry air's molecular mass is approximately $28.97[g/mol]$. Thus, in terms of grams, the gas constant R is

$$R = 8.314[JK^{-1}mol^{-1}]/(28.97[JK^{-1}mol^{-1}]) = 0.287[JK^{-1}g^{-1}], \tag{1.26}$$

or

$$R = 287[JK^{-1}kg^{-1}]. \tag{1.27}$$

1.6 Dimensional analysis example: heat diffusion

A point source with Q Joules of heat is initially placed at the mid of a 1-dimensional heat conducting rod. The mid-point is denoted by $x = 0$. The heat will diffuse as time goes, and the temperature $T(x, t)$ at the location x and time t is given by the following formula:

$$T = \frac{Q}{\rho C\sqrt{2\pi Dt}} \exp\left(-\frac{x^2}{2Dt}\right). \tag{1.28}$$

where Q is the heat energy in Joule or other units with dimension $[ML^2T^{-2}]$, ρ is the linear density of the 1-dimensional rod with dimension $[ML^{-1}]$, C is the specific heat which is the energy per unit mass per temperature degree (i.e., the energy needed to heat a unit mass up by one degree) with units $[Joule/(g \times^\circ C)]$ and dimension $[ML^2T^{-2}M^{-1}\Theta^{-1}]$, and D is thermal diffusivity with dimension $[L^2T^{-1}]$. Thus, the dimension of Dt is $[L^2T^{-1}][T] = [L^2]$, which is the same dimension of x^2 inside the exponential function. So, $\frac{x^2}{2Dt}$ is dimensionless, which should be, since exponential, trigonometric and logarithmic functions can only act on dimensionless quantities.

Thermal diffusivity D measures the diffusion rate, a property of the material in which the heat diffuses. Large D means fast heat diffusion. For example, heat diffuses faster in iron than rock.

The dimension of the right hand side is

$$\begin{aligned} & \left[\frac{Q}{\rho C\sqrt{2\pi Dt}}\right] \\ = \quad & \frac{[Q]}{[\rho][C]\sqrt{[Dt]}} \\ = \quad & \frac{[ML^2T^{-2}]}{[ML^{-1}][ML^2T^{-2}M^{-1}\Theta^{-1}]\sqrt{[L^2]}} \\ = \quad & \Theta = [T] \text{ (The dimension of temperature).} \end{aligned} \tag{1.29}$$

This is the same as the dimension of the left hand side T.

1.7 Dimensional analysis for atmospheric Rossby waves

Rossby waves on Earth are very long meanders in high-altitude winds caused by Coriolis force when an atmospheric mass moves in meridional direction. Atmospheric Rossby waves are associated with pressure systems and jet flows and have have weather predictability skills. For example, the arctic region's pressure variation may lead to different jet streams from the mid and high latitude regions, such as the Alaska jet that affects California. The Coriolis force makes the cold air mass move along the equal-pressure curves, or called isobars. This cold air flow along an isobar can help with weather forecasting.

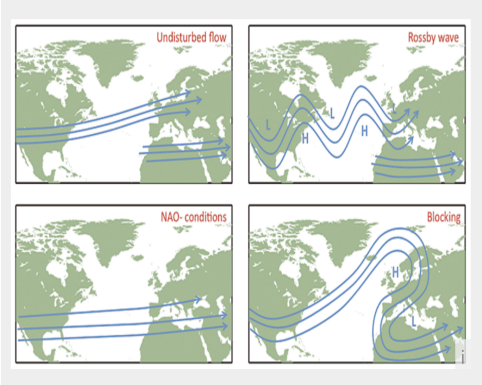

Figure 1.4 Rossby wave and jet stream over mid-latitude zone of the northern hemisphere.

Oceanic Rossby waves, moving mostly in meridional direction too, are gravity waves of thermocline, the interface between the warm upper layer and the cold deeper layer.

The quantitative use of Rossby wave for weather forecasting needs three major parameters of a Rossby wave: wave speed U, wave length L, and Coriolis force given by Coriolis frequency

$$f = 2\Omega \sin \phi \tag{1.30}$$

where $\Omega = 7.3 \times 10^{-5}[radian/sec]$ is the angular frequency of Earth rotation, and ϕ the latitude of the location. The Coriolis force is zero at equator and the largest at poles. The wave speed U can help predict when a Rossby wave can arrive. The wave length L can determine the size of a wave. Coriolis frequency can determine strength of the Earth's rotation effect.

Following the general rule of dimensional analysis, we may form product and quotient from three quantities to form a non-dimensional quantity, which may become an important index to explain physical phenomena. The dimensions of the three quantities are

$$[U] = LT^{-1}, [L] = L, \text{and } [f] = T^{-1}. \tag{1.31}$$

Place them together to form a dimensionless index

$$R_0 = [U]^a[L]^b[f]^c = (LT^{-1})^a L^b (T^{-1})^c = L^{a+b}T^{-a-c}. \tag{1.32}$$

To be dimensionless, we have

$$a + b = 0, \tag{1.33}$$
$$-a - c = 0. \tag{1.34}$$

These two equations yield

$$b = -a, c = -a. \tag{1.35}$$

The simplest and nontrivial solution is $a = 1$, $b = -1$ and $c = -1$, which leads to

$$Ro = U^1 L^{-1} f^{-1}, \quad \text{or} \quad Ro = \frac{U}{Lf}. \tag{1.36}$$

This is the famous Rossby number, which is the unique combination of the three key fundamental quantities: U, L, f. Rossby number (R_0) is geophysical fluid dynamics parameter that measures the Coriolis influence, and was named for Carl-Gustav Arvid Rossby (1898-1957).

Physically, Rossy number can be explained as the ratio of convection to Coriolis force: $v \cdot \nabla v \sim U^2/L$ to $\mathbf{\Omega} \times \mathbf{v} \sim U\Omega$, which can be simplified to

$$\mathrm{Ro} = \frac{U}{Lf}, \tag{1.37}$$

where L is wave length of a Rossby wave, U is the phase velocity,

A large Rossby number in a Rossby wave means that a system is dominated by convection, and Coriolis force is relatively small. For example, tornadoes have a high air flow speed and a large speed gradient, and hence have a large Rossby number, as high as 103. Coriolis force due to Earth's rotation is relatively less important than the centrifugal force in the tornado's fast rotation. A tornado eye has a low pressure. Air pressure has a large gradient from the eye to the tornado's outer boundary. The pressure gradient is balanced by centrifugal forces, which is called cyclostrophic balance. The Coriolis force is negligible. The balance makes the atmosphere moves very fast along the closed isobar curves around the tornado eye, co-centered circles in an ideal case. The air flow direction is perpendicular to the pressure gradient. The circular motion of the air forms a strong vortex with an almost vaccum eye that results in a huge suction and can cause a great destruction.

A small Rossby number means relatively slow convection. A Rossby wave in a large low-pressure system, such as the northern hemispheric jet stream, may have a small Rossby number around 1.0. The Coriolis force plays a critical role in determining the flow. The geophysical force balance is between the Coriolis force and the pressure gradient, which is called geostrophic balance. This balance again makes the air flow along the isobar curves, perpendicular to the pressure gradient.

Here,"strophic" is from Greek, meaning "turning". "Strophe" originally denoted a movement from right to left made by a Greek chorus. "Geo" is also Greek and means Earth. "Geostrophic" in climate science means atmospheric or oceanic wave motion forced by a pressure system and Coriolis force reflecting the Earth's rotation effect.

Another important geophysical parameter is called Rossby parameter which measures the gradient of the Coriolis force in a meridional direction:

$$\beta = \frac{\partial f}{\partial y} = \frac{1}{a}\frac{d}{d\phi}(2\Omega \sin\phi) = \frac{2\Omega \cos\phi}{a} \tag{1.38}$$

where Ω is the angular speed of the Earth's rotation [rad/sec], and a the Earth radius. Here, $dy = d(a\phi) = ad\phi$ is used as a small distance increment in the meridional direction.

The Rossby parameter describes the variation of the Coriolis force with latitude (hence the latitudinal derivative) and does not depend on a weather phenomenon, while Rossby number measures the strength of convection with respect to the Coriolis force and strongly depends on a weather phenomenon.

For a wave, amplitude A is important. The following expression

$$\mathrm{So} = \frac{UA}{Lf} \tag{1.39}$$

has the length dimension. This new parameter is a measure of the strength of the Rossby wave with respect to the wave length.

Another measure of using the depth D of an atmospheric Rossby wave is

$$L_R = \frac{\sqrt{gD}}{f} = \frac{\sqrt{gD}}{2\Omega sin\phi}. \tag{1.40}$$

This one also has the length dimension and measures the radius of the upward and downward circular motion of atmosphere. The Coriolis frequency is approximately equal to $f = 1 \times 10^{-4}[s^{-1}]$ at $\phi = 45^\circ$ latitude. If $D = 4,000[m]$ from land surface, then $L_R = 2,000[km]$ is large. For a very shallow Rossby wave $D = 40[m]$, then $L_R = 200[km]$ is small. L_R is called Rossby radius.

1.8 Estimating the shock wave radius of a nuclear explosion

The instantaneous energy release from a nuclear explosion causes an air shock wave, whose inside pressure is thousands times greater than outside. This pressure difference can push down trees and structures, and tear apart all kinds of objects. If we assume shock wave to be spherical with radius $R(t)$ from the ground zero at t time after the explosion. Given the nuclear energy E, calculate the shock wave radius as a function of time, and hence predict the shock wave's arrival time and prepare for protection.

Shock wave occurring in atmosphere due to the supersonic compression of the air from one side so that the air mass from the side accumulates, cannot escape, builds pressure, develops a large pressure difference with the other side, and hence forms a shock. Two critical elements here are supersonic push and air's compressibility property. Thus, the shock wave radius should be related to density ρ of a compressible air, total energy of the nuclear explosion E, and time t. Because air is light in the scale nuclear explosion, gravity can be negligible. Thus, we assume the following

$$R = \alpha E^a \rho^b t^c. \tag{1.41}$$

The dimension of the above equation is

$$[R] = [\alpha][E]^a[\rho]^b[t]^c, \tag{1.42}$$

which leads to

$$L = 1 \times (ML^2T^{-2})^a(ML^{-3})^bT^c = M^{a+b}L^{2a-3b}T^{-2a+c}. \tag{1.43}$$

The exponents of both sides of this equation should be equal:

$$a + b = 0, \tag{1.44}$$
$$2a - 3b = 1, \tag{1.45}$$
$$-2a + c = 0. \tag{1.46}$$

These three equations have a unique solution

$$a = 1/5, b = -1/5, c = 2/5. \tag{1.47}$$

Therefore,

$$R = \alpha E^{1/5} \rho^{-1/5} t^{2/5}. \tag{1.48}$$

or

$$R = \alpha \left(\frac{E t^2}{\rho} \right)^{1/5}. \tag{1.49}$$

This makes Et^2 a very special term, which is the fifth power of density times length according to dimension equality, meaning the density of the air behind the shock.

Another expression of the shock radius is

$$R = \alpha \left(\frac{E}{\rho} \right)^{1/5} t^{2/5}. \tag{1.50}$$

The log-plot of this $T - t$ relationship is a straight line with slope 2/5:

$$\ln R = \ln \alpha + \ln \left(\frac{E}{\rho} \right)^{1/5} + \frac{2}{5} \ln t. \tag{1.51}$$

Any of the above three formulas can be used to predict the position of the shock wave for a given time, if α is known. Yet, it is not easy to evaluate this α by an experiment since such an experiment is too destructive. A way out is to derive it from mathematical models. Cambridge University fluid dynamicist G. I. Taylor (1886-1975) used a mathematical model and estimated that $\alpha = 1.0$.

Still another way of writing the above equation is

$$E = \frac{R^5 \rho}{t^2}. \tag{1.52}$$

This allows one to estimate the power of an nuclear bomb using news reports on the the shock arrival time at a given location.

If a nuclear bomb test is made underground, seismograph can measure R and t. With the known Earth crest's density, one can then estimate the bomb's power. There are 500 seismograph stations distributed around the world to detect ground-shaking incidents, including earthquakes and nuclear bombs.

One can use similar model to estimate the shock waves caused by supernova explosions (Exploring the X-ray Universe, Seward and Charles 2010).

EXERCISES

1.1 Make a dimensional analysis for the Newton's second law of motion: $F = ma$.

1.2 Design an experiment to demonstrate Newton's second law of motion: $F = ma$, when assuming the mass does not change and observing the data of acceleration and force.

1.3 Make a dimensional analysis for the gravitational force F_g based on law of universal gravitation shown in Fig. 1.5.

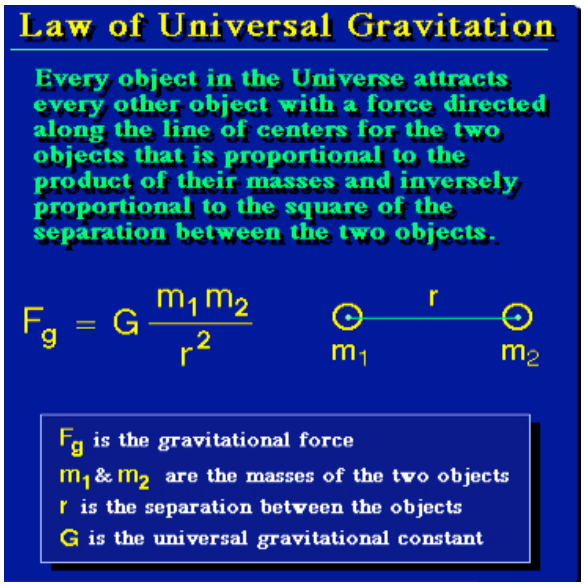

Figure 1.5 Law of universal gravitation.

1.4 Dimensional analysis with experimental data.

a) Make a dimensional analysis for the velocity v and distance h of a free-fall body of mass m.

b) Perform the experiments of free-fall using a coin or any heavy metal or a stone to determine the dimensionless constant for distance. This is a difficult experiment since it happens really fast, and it is very hard to record time.

c) Change the free-fall experiment to free-roll experiment as Galileo did (see the preface of this book). Place a ball on an inclined plate and let it roll down by gravity. The gravity along the plate is now reduced to $g \sin \phi$ where ϕ is the angle between the plate and the flat flow (ideally the tangent plane perpendicular to the Earth's radius). The experiment is now easier since it is easier to record the time. However, the inclined angle should not be too small, which will make friction force non-negligible. Again, one can achieve better accuracy when repeating the experiment many times and using the average result.

1.5 Dimensional analysis example: heat flux from ocean floor to ocean water: Ocean has hydrothermal flow from certain regions of ocean floor through ocean floor crust and

sediments. The heat flux is denoted by q_{cond} in units $[W/cm^2]$. This flux is related to the four critical quantities:

(i) Thermal molecular diffusivity: $D_H[cm^2 sec^{-1}]$,

(ii) Temperature gradient: $\nabla T[K/cm]$,

(iii) Heat capacity of seawater: $C_p[J/(g.K)]$, and

(iv) Density of seawater: $\rho[g/cm^3]$.

Do the following:

a) Find the dimension for each of the five quantities above: $[q_{cond}], [D_H], [\nabla T], [C_p], [\rho]$. Use the given units as a hint.

b) Find how q_{cond} is related to the other four quantities using the following dimensional analysis equation:

$$[q_{cond}] = [D_H]^a [\nabla T]^b [C_p]^c [\rho]^d. \tag{1.53}$$

Use the results in part a) to show that

$$a = b = c = 1. \tag{1.54}$$

This leads to the following equation of heat diffusion from the higher temperature ocean floor to the lower temperature seawater:

$$q_{cond} = D_H \nabla T C_p \rho. \tag{1.55}$$

c) From the units of the right hand side of the above equation:

$$[cm^2 s^{-1}][K/cm][J/(g.K)][g/cm^3], \tag{1.56}$$

show that the units of q_{cond} is W/cm^2.

REFERENCES

1. JG.I. Barenblatt, 1987: Dimensional Analysis, Gordon and Breach Science Publishers, New York, 354pp.
2. F.D. Seward and P.A. Charles, 2010: Exploring the X-rays Universe, 2nd ed., Cambridge University Press, New York, 372pp.

CHAPTER 2

BASICS OF R PROGRAMMING

t is popular in today's mathematical modeling books to use computing tools for complex and tedious algebras so that students can focus on correct usage of the mathematical tools with accurate statement of assumptions and precise interpretation of the results. Among many software packages used in climate community, R's popularity has dramatically increased in the last a few years due to its enormous power of handling big data. We thus choose to include the basics of R for this book. A student who has mastered the R examples used in this book should have sufficient skills to develop R projects independently.

2.1 Download and install R software package

For Windows users, visit the website
`https://cran.r-project.org/bin/windows/base/`
to find the instructions of R program download and installations.

For Mac users, visit
`https://cran.r-project.org/bin/macosx/`

If you experience difficulties, please refer to online resources, Google or Youtube. A recent 3-minute Youtube instruction for R installation for Windows can be found from the following link:

https://www.youtube.com/watch?v=Ohnk9hcxf9M

The same author also has a youtube instruction about R installation for Mac (2 minutes):
https://www.youtube.com/watch?v=uxuuWXU-7UQ

One can choose to install R-Studio instead. Then visit
https://www.rstudio.com/products/rstudio/download/
This site allows to choose Windows, or Mac OS, or Unix.

One can use either R or R-Studio, or both, depending on his interest.

For details about the publicly open access to R-Project, visit
https://www.r-project.org/

The beginners of R would find it very difficult to navigate through this official, formal, detailed, and massive R-Project documentation to learn the program. Fortunately, many excellent tutorials for a quick learn of R programming are available online and in Youtube. One can google around and find a couple of preferred tutorials.

The following section provides R basics useful to this book.

2.2 R Tutorial

2.2.1 R as a smart calculator

R can be used like a smart calculator that allows fancier calculations than those done on regular calculators.

```
1+4
[1] 5
2+pi/4-0.8
[1] 1.985398
x<-1
y<-2
z<-4
t<-2*x^y-z
t
[1] -2
u=2         # "=" sign and "<-" is almost equivalent
v=3         # The text behind the "#" sign is comments
u+v
[1] 5
sin(u*v)    # u*v = 6 is considered radian
[1] -0.2794155
```

2.2.2 Write a function in R

The function command is
`name <- function(var1, var2, ...) expression of the function.`
For example,

```
square <- function(x) x*x
square(4)
[1] 16
fctn <- function(x,y,z) x+y-z/2
fctn(1,2,3)
[1] 1.5
```

2.2.3 Plot with R

R can can plot all kinds of curves, surfaces, statistical plots, and maps. Below are a few examples. For adding labels, ticks, color, and other features to a plot, one can google R plot and find the commands to properly include the desired features.

```
plot(sin, -pi, 2*pi)   #plot the curve of y=sin(x) from -pi to 2 pi
square <- function(x) x*x   #Define a function
plot(square, -3,2)   # Plot the defined function
fctn(1,2,3)
[1] 1.5
 ## Plot a 3D surface
x <- seq(-1, 1, length=100)
Z <- outer(x, x, function(x, y)sqrt(1-x^2-y^2))
#outer (x,y, function) is outer product
persp(x=x, y=x, z=Z, theta=310)
# yields a 3D surface with perspective angle 310 deg

#Another 3D surface plot example
x=seq(0, 300, 10)
y=seq(0, 300, 10)
z=outer(x, y, function(x,y){x^2+y^2})
persp(x, y, z, phi = 10, theta = 45,
    xlab = "X ", ylab = "Y ", main = "3D Surface ")
```

2.2.4 Symbolic calculations by R

People used to think that R can only handle numbers. Actually R can do symbolic calculations, such as finding a derivative. However, up to now R is not the best symbolic calculation tool. One can use WolframAlpha, SymPy, and Yacas for free or use the paid software package Maple or Mathematica. Google symbolic calculation for calculus to find a long list of symbolic calculation software packages, such as `https://en.wikipedia.org/wiki/List_of_computer_algebra_systems`.

```
D(expression(x^2,'x'), 'x')
# Take derivative of x^2 and the answer is 2x
2 * x
```

```
fx= expression(x^2,'x')  #assign a function
D(fx,'x') #differentiate the function with result below
2 * x
fx= expression(x^2*sin(x),'x')
#Change the expression and use the same derivative command
D(fx,'x')
2 * x * sin(x) + x^2 * cos(x)
 fxy = expression(x^2+y^2, 'x','y')
#One can define a function of 2 or more variables
 fxy #This gives the expression of the function in terms of x and y
expression(x^2 + y^2, "x", "y")
D(fxy,'x') #This gives the partial derivative with respect to x: 2 * x
D(fxy,'y') #This gives the partial derivative with respect to y: 2 * y
square = function(x) x^2
integrate (square, 0,1)
#Integrate x^2 from 0 to 1 equals to 1/3 with details below
0.3333333 with absolute error < 3.7e-15
integrate(cos,0,pi/2)
#Integrate cos(x) from 0 to pi/2 equals to 1 with details below
1 with absolute error < 1.1e-14
```

2.2.5 Vectors and matrices

R can handle all kinds of operations vectors and matrices.

```
c(1,6,3,pi,-3) #c() gives a vector and is considered a 4X1 column vector
[1]  1.000000  6.000000  3.000000  3.141593 -3.000000
seq(2,6) #Generate a sequence from 2 to 6
[1] 2 3 4 5 6
seq(1,10,2) # Generate a sequence from 1 to 10 with 2 increment
[1] 1 3 5 7 9
x=c(1,-1,1,-1)
x+1 #1 is added to each element of x
[1] 2 0 2 0
2*x #2 multiplies each element of x
[1]  2 -2  2 -2
x/2 # Each element of x is divided by 2
[1]  0.5 -0.5  0.5 -0.5
y=seq(1,4)
x*y  # This multiplication * multiples each pair of elements
[1]  1 -2  3 -4
x%*%y #This is the dot product of two vectors and yields
     [,1]
[1,]   -2
t(x)  # Transforms x into a row 1X4 vector
     [,1] [,2] [,3] [,4]
[1,]    1   -1    1   -1
t(x)%*%y #This is equivalent to dot product and forms 1X1 matrix
     [,1]
```

```
[1,]   -2
> x%*%t(y) #This column times row yields a 4X4 matrix
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]   -1   -2   -3   -4
[3,]    1    2    3    4
[4,]   -1   -2   -3   -4
my=matrix(y,2,2)
#Convert a vector into a matrix of the same number of elements
#The matrix elements go by column, first column, second, etc
     [,1] [,2]
[1,]    1    3
[2,]    2    4
dim(my)  #find dimensions of a matrix
[1] 2 2
as.vector(my) #Convert a matrix to a vector, again via columns
[1] 1 2 3 4
mx*my #multiplication between each pair of elements
     [,1] [,2]
[1,]    1    3
[2,]   -2   -4
mx/my #division between each pair of elements
     [,1]       [,2]
[1,]  1.0  0.3333333
[2,] -0.5 -0.2500000
mx-2*my
     [,1] [,2]
[1,]   -1   -5
[2,]   -5   -9
mx%*%my #This is the real matrix multiplication in matrix theory
     [,1] [,2]
[1,]    3    7
[2,]   -3   -7
det(my) #determinant
[1] -2
myinv = solve(my) #yields the inverse of a matrix
> myinv
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
> myinv%*%my #verifies the inverse of a matrix
     [,1] [,2]
[1,]    1    0
[2,]    0    1
diag(my) #yields the diagonal vector of a matrix
[1] 1 4
myeig=eigen(my) #yields eigenvalues and unit eigenvectors
myeig
$values
```

```
[1]  5.3722813 -0.3722813
$vectors
           [,1]       [,2]
[1,] -0.5657675 -0.9093767
[2,] -0.8245648  0.4159736
mysvd = svd(my) #SVD decomposition of a matrix M=UDV'
            #SVD can be done for a rectangular matrix of mXn
mysvd
$d
[1] 5.4649857 0.3659662
$u
           [,1]       [,2]
[1,] -0.5760484 -0.8174156
[2,] -0.8174156  0.5760484
$v
           [,1]       [,2]
[1,] -0.4045536  0.9145143
[2,] -0.9145143 -0.4045536

ysol=solve(my,c(1,3))
#solve linear equations matrix %*% x = b
ysol  #solve(matrix, b)
[1]  2.5 -0.5
my%*%ysol #verifies the solution
     [,1]
[1,]    1
[2,]    3
```

2.2.6 Statistics

R was originally designed by statisticians for doing statistics. Thus, R has a comprehensive set of statistics functions. This sub-section gives a few basic commands. More will be described in the statistical modeling chapters.

```
x=rnorm(10) #generate 10 normally distributed numbers
x
 [1]  2.8322260 -1.2187118  0.4690320 -0.2112469  0.1870511
 [6]  0.2275427 -1.2619005  0.2855896  1.7492474 -0.1640900
 mean(x)
[1] 0.289474
var(x)
[1] 1.531215
sd(x)
[1] 1.237423
median(x)
[1] 0.2072969
quantile(x)
        0%        25%        50%        75%       100%
-1.2619005 -0.1994577  0.2072969  0.4231714  2.8322260
```

```
range(x) #yields the min and max of x
[1] -1.261900  2.832226
 max(x)
[1] 2.832226

boxplot(x) #yields the box plot of x
w=rnorm(1000)

summary(rnorm(12)) #statistical summary of the data sequence
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.9250 -0.6068  0.3366  0.2309  1.1840  2.5750

hist(w)
#yields the histogram of 1000 random numbers by normal distribution
```

2.3 Online Tutorials

2.3.1 Youtube tutorial: for true beginners

This is a very good and slow paced 22 minutes youtube tutorial: Chapter 1. An Introduction to R
https://www.youtube.com/watch?v=suVFuGET-0U

2.3.2 Youtube tutorial: for some basic statistical summaries

This is a 9 minutes tutorial by Layth Alwan.
https://www.youtube.com/watch?v=XjOZQN-Nre4

2.3.3 Youtube tutorial: Input data by reading a csv file into R

An excel file can be saved as csv file: xxxx.csv. This 15 minutes youtube shows how to read a csv file into R by Layth Alwan. He also shows linear regression.
https://www.youtube.com/watch?v=QkE8cp0B9gg

R can input all kinds of data files, including xlsx, netCDF, fortran data, and sas data. Some commands are below. One can google to find proper data reading command for your particular data format.

```
mydata <- read.csv("mydata.csv")
# read csv file named "my data.csv"

mydata <- read.table("mydata.txt")
# read text file named "my data.txt"

library(gdata)                    # load gdata package
mydata = read.xls("mydata.xls")   # read an excel file

library(foreign)                  # load the foreign package
mydata = read.mtp("mydata.mtp")   # read from .mtp file
```

```
library(foreign)                  # load the foreign package
mydata = read.spss("myfile", to.data.frame=TRUE)

ff <- tempfile()
cat(file = ff, "123456", "987654", sep = "\n")
read.fortran(ff, c("F2.1","F2.0","I2")) #read a fotran file

library(ncdf)
ncin <- open.ncdf(ncfname)  # open a NetCDF file
lon <- get.var.ncdf(ncin, "lon") #read a netCDF file into R
```

Some libraries are not in the R project anymore. For example,

```
library(ncdf) #The following error message pops up
Error in library(ncdf) : there is no package called ncdf
```

One can then google r data reading netcdf R-project and go to the R-project website. The following can be found.

```
Package ncdf was removed from the CRAN repository.
Formerly available versions can be obtained from the archive.
Archived on 2016-01-11: use 'RNetCDF' or 'ncdf4' instead.
```

This means that one should use RNetCDF, which can be downloaded from internet. Thus, if a library gives an error message, then google the library package, download and install the package, and finally read the data of the particular format.

2.4 More on statistical computing and plotting with R

2.4.1 Introduction

Statistics originated from Latin "status" meaning "state" and is a suite of scientific methods that analyze data and make credible conclusions. Statistical methods are routinely used for climate data, such as calculating the climate normal of precipitation at a weather station, claiming the global warming based on a significant positive linear trend of the surface air temperature (SAT) anomalies, and inferring a significant shift from the lower North Pacific sea level pressure (SLP) state to a higher state. The list of questions such as the above can be infinitely long. The purpose of this chapter is to provide basic concepts and a user manual on the commonly used statistical methods in climate data analysis, so that the users can make credible conclusions with a given error probability.

R-programs will be supplied for examples in this chapter. Users can easily apply these programs and the given formulas in this book for their data analysis needs without prerequisite background of calculus and much statistics. To interpret the statistics results in a meaningful way, domain knowledge of climate science should be very useful when using the statistical concepts and calculations results to state the conclusions from specific climate datasets.

The statistical methods in this chapter focusing on making credible inference about the climate state with a given error probability based on the analysis of climate data, so

that the observed data can form the basis of making objective and reliable conclusions. We will describe a list of statistical indices, such as mean, variance and quantiles, for climate data, and then look into the probability distributions and statistical inferences.

2.4.2 A list of statistical indices for a set of temperature data

Below is data of the global average annual mean temperature anomalies from 1880-2015 (Karl et al. 2015, NOAA GlobalTemp dataset at NCDC `http://www1.ncdc.noaa.gov/pub/data/noaaglobaltemp/operational/`). In the data list, the datum corresponds to 1880 and the last 2015. These 136 years of data are used to illustrate the following statistical concepts: mean, variance, standard deviation, skewness, kurtosis, median, 5th percentile, 95th percentile, and other quantiles. The anomalies are with respect to the 20th century mean, i.e., the 1900-1999 climatology period. The global average of the 20th century mean is 12.7 °C. The 2015 anomaly was 0.65 °C. Thus, the 2015's global average annual mean temperature is 13.4°C.

```
  [1] -0.367918 -0.317154 -0.317069 -0.393357 -0.457649 -0.468707
  [7] -0.451778 -0.498811 -0.403252 -0.353712 -0.577277 -0.504825
 [13] -0.556487 -0.568014 -0.526737 -0.475364 -0.340468 -0.367002
 [19] -0.505967 -0.368630 -0.315155 -0.387099 -0.494861 -0.585158
 [25] -0.663492 -0.535226 -0.457892 -0.617208 -0.684107 -0.672176
 [31] -0.624129 -0.675199 -0.570521 -0.558340 -0.379505 -0.308313
 [37] -0.531023 -0.551480 -0.444860 -0.444257 -0.451256 -0.388185
 [43] -0.469536 -0.455500 -0.489551 -0.385962 -0.305391 -0.393436
 [49] -0.416556 -0.538602 -0.339823 -0.316963 -0.360309 -0.486954
 [55] -0.347795 -0.383147 -0.356958 -0.262097 -0.272009 -0.257514
 [61] -0.152032 -0.050356 -0.095295 -0.088983  0.044418 -0.073264
 [67] -0.251405 -0.297744 -0.296136 -0.303984 -0.405346 -0.255647
 [73] -0.218081 -0.146923 -0.358796 -0.377482 -0.441748 -0.194232
 [79] -0.133076 -0.184608 -0.222896 -0.165795 -0.154384 -0.137509
 [85] -0.393492 -0.322453 -0.267491 -0.257946 -0.274517 -0.151345
 [91] -0.207025 -0.322901 -0.216440 -0.080250 -0.316583 -0.241672
 [97] -0.323398 -0.046098 -0.131010 -0.016080  0.021495  0.057638
[103] -0.061422  0.099061 -0.093873 -0.109097 -0.015374  0.125450
[109]  0.129184  0.050926  0.186128  0.159565  0.010836  0.038629
[115]  0.092131  0.211006  0.074193  0.269107  0.384935  0.194762
[121]  0.177381  0.296912  0.351874  0.363650  0.329436  0.408409
[127]  0.362960  0.360386  0.291370  0.385638  0.453061  0.325297
[133]  0.370861  0.416356  0.491245  0.650217
```

We use R to calculate all the statistical parameters. The data is read as `tmean15`.

```
> mean(tmean15)
[1] -0.2034367
> sd(tmean15)
[1] 0.3038567
> var(tmean15)
[1] 0.09232888
> library(e1071)
```

```
> skewness(tmean15)
[1] 0.7141481
> kurtosis(tmean15)
[1] -0.3712142
> median(tmean15)
[1] -0.29694
> quantile(tmean15,probs= c(0.05,0.25, 0.75, 0.95))
        5%        25%        75%        95%
-0.5792472 -0.4228540 -0.0159035  0.3743795
```

The following R commands can plot the time series of the temperature data with a linear trend (see Fig. 2.1).

```
 > yrtime15=seq(1880,2015)
 > reg8015<-lm(tmean15 ~ yrtime15)
 # Display regression results
> reg8015
Call:
lm(formula = tmean15 ~ yrtime15)
Coefficients:
(Intercept)     yrtime15
 -13.208662     0.006678
 # Plot the temperature time series and its trend line
 > plot(yrtime15,tmean15,xlab="Year",ylab="Temperature deg C",
 main="Global Annual Mean Land and Ocean Surface
 Temperature Anomalies 1880-2015", type="l")
> abline(reg8015, col="red")
> text(1930, 0.4, "Linear temperature trend 0.6678 oC per century",
col="red",cex=1.2)
```

The mathematics formulas for the above statistical parameters are below. Let $x = \{x_1, x_2, \cdots, x_n\}$ be the sampling data for a time series. Then,

$$\text{mean: } \mu(x) = \frac{1}{n}\sum_{k=1}^{n} x_k, \tag{2.1}$$

$$\text{variance by unbiased estimate: } \sigma^2(x) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \mu(x))^2, \tag{2.2}$$

$$\text{standard deviation: } \sigma(x) = (\sigma^2(x))^{1/2}, \tag{2.3}$$

$$\text{skewness: } \gamma_3(x) = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k - \mu(x)}{\sigma}\right)^3, \tag{2.4}$$

$$\text{kurtosis: } \gamma_4(x) = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k - \mu(x)}{\sigma}\right)^4 - 3. \tag{2.5}$$

Mean gives the average of samples. Variance and standard deviation measure the spread of samples. It is large when the samples have a wide spread. Skewness is dimensionless and measures the asymmetry of samples. Zero skewness means symmetric distribution. For example, a normal distribution's skewness is zero. Negative

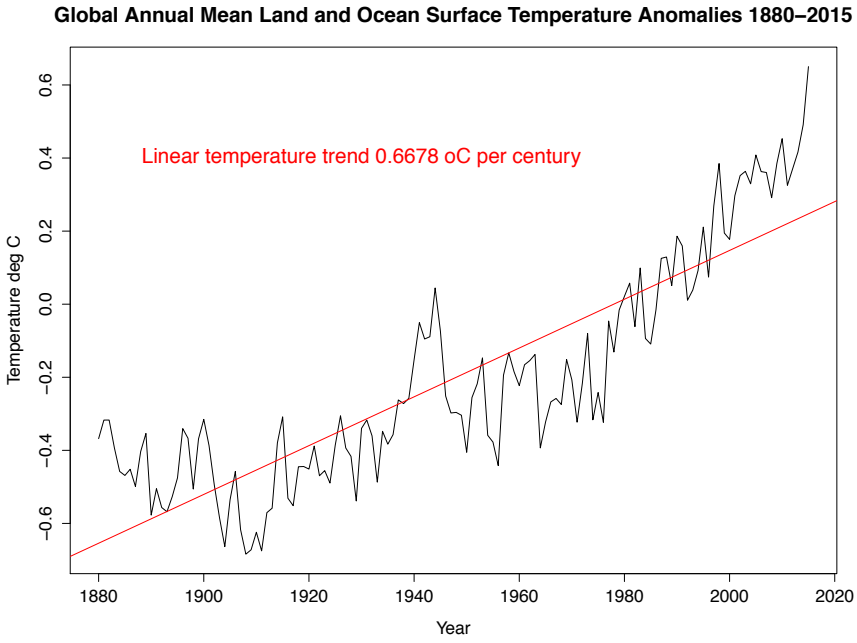

Figure 2.1 Time series of the global average annual mean temperature with respect to 1900-1999 climatology: 12.7 °C.

skewness is skew to the left, meaning that the long distribution tail is on the left. Positive skewness has a long tail on the right. Kurtosis is also dimensionless and measures the peakedness of a distribution. The normal distribution's kurtosis is zero. Positive kurtosis means a high peak at the mean, a slim and tall distribution. This is referred to as leptokurtic. "Lepto" is Greek and means thin or fine. Negative kurtosis means a low peak at the mean, a fat and short distribution, referred to as platykurtic. "Platy" is also Greek and means flat or broad. "Kurtic" and "kurtosis" are Greek and mean peakedness.

For the 136 years of global average annual mean temperature data, the skewness is 0.71, meaning skew to the right with tail to on the right with more extreme high temperatures, as shown in the histogram Fig. 2.2. The kurtosis is -0.37, meaning flatter than a normal distribution also shown by the histogram.

Median is a number sample set such that 50% of the samples are less than the median, and another 50% greater than the median. Sort the samples from the smallest to the largest. The median is the number in the middle. If the number of the samples is even, then median is equal to the mean of the two middle samples.

Quantiles are defined in the same way by sorting. For example, 25-percentile is a sample that 25% of samples are less than this sample. By definition, 75-percentile is larger than 40-percentile. 100-percentile is the largest sample, and 0-percentile is the smallest sample. Usually, people use a box plot to show the typical quantiles. See Fig. 2.3 for the box plot of the 136 years of temperature data.

50-percentile is called median. If the distribution is symmetric, then median is equal to mean. Otherwise they are not. If skew to the right, then mean is on the right

of median: mean greater than median. If skew to the left, then mean is on the left of median: mean less than median. Our 136 years of temperature data are right skewed and have mean equal to -0.2034°C, greater than their median equal to -0.2969°C.

2.4.3 A set of commonly used statistical figures

We will use the 136 years of temperature data and R to illustrate the following commonly used statistical figures: histogram, box plot, scatter plot, qq-plot, and linear regression trend line.

2.4.3.1 Histogram of a set of data

```
h<-hist(tmean15, main="Histogram of 1880-2015 Temperature
Anomalies",xlab="Temperature anomalies") #Plot historgram
xfit<-seq(min(tmean15),max(tmean15), length=30)
areat=diff(h$mids[1:2])*length(tmean15) #Normalization area
yfit<-areat*dnorm(xfit, mean=mean(tmean15), sd=sd(tmean15))
lines(xfit,yfit,col="blue",lwd=2) #Plot the normal fit
```

Figure 2.2 shows the result of the above R histogram commands.

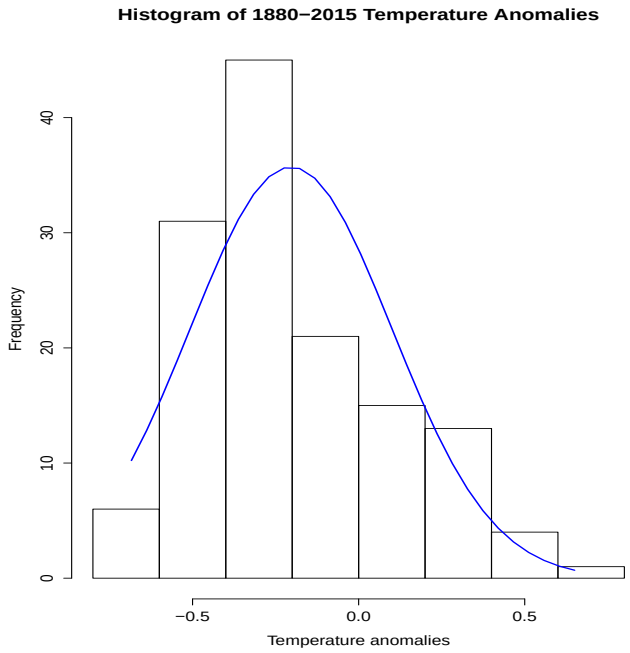

Figure 2.2 Histogram of the global average annual mean temperature anomalies from 1880-2015.

One can also plot the probability density function based on the R's kernel estimate.

```
plot(density(tmean15), main="Kernel estimate
of density",xlab="Temperature") #Kernel estimate density
lines(xfit,dnorm(xfit, mean=mean(tmean15),
sd=sd(tmean15)), col="blue") #Moment estimated normal
```

2.4.3.2 Box plot Figure 2.3 is the box plot of the 136 years of temperature and can be made from the following R command `b=boxplot(tmean15, ylab="Temperature anomalies")`

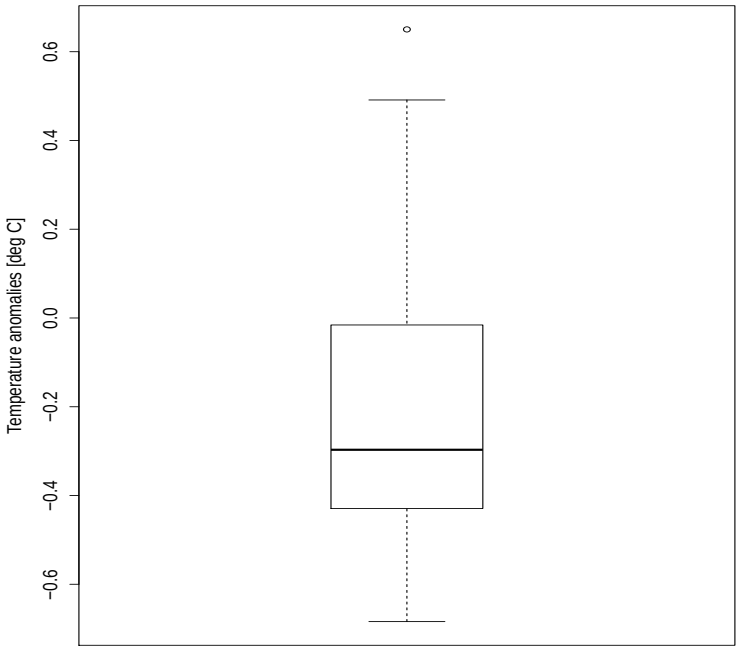

Figure 2.3 Box plot of the global average annual mean temperature anomalies from 1880-2015.

The rectangular box's mid line indicates the level of media, which is -0.30°C. The rectangular box's lower boundary is the first quartile, i.e., 25-percentile. The box's upper boundary is the third quartile, i.e., the 75-percentile. The box's height is the third quartile minus the first quartile, and is called interquartile range (IQR). The upper whisker is the third quartile plus 1.5 IQR. The lower whisker is supposed to be at the first quartile minus 1.5 IQR. However, this whisker is lower than the lower extreme. Thus, the lower whisker takes the lower extreme, which is -0.68 °C. The points outside of the two whiskers are considered outliers. Our dataset has one outlier, which is 0.65 °C. This is the hottest year happened in 2015.

Sometimes, one may need to plot multiple box plots on the same figure, which can be done by R. One can follow an example in R-project document
`http://www.inside-r.org/r-doc/graphics/boxplot`

2.4.3.3 Scatter plot Scatter plot is used to display if two datasets are correlated. We use the southern oscillation index (SOI) and the contiguous United States temperature as an example to describe the scatter plot. The data can be downloaded from
`www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/`
`www.ncdc.noaa.gov/temp-and-precip/`

The following R commands can produce the scatter plot shown in Fig. 2.4.

```
ust=read.csv("USJantemp1951-2016-nohead.csv",header=FALSE)
soi=read.csv("soi-data-nohead.csv", header=FALSE) #Read data
soid=soi[,2] #Take the second column SOI data
soim=matrix(soid,ncol=12,byrow=TRUE)
#Make the SOI into a matrix with each month as a column
soij=soim[,1] #Take the first column for Jan SOI
ustj=ust[,3] #Take the third column: Jan US temp data
plot(soij,ustj,main="Scatter plot between Jan SOI
and US temp",xlab="SOI[dimensionless]",
ylab="US Temp degF", pch=19)
# Plot the scatter plot
soiust=lm(ustj ~ soij) #Linear regression
abline(soiust, col="red") #Linear trend line
```

The correlation between the two datasets is 0. Thus, the slope is also zero.

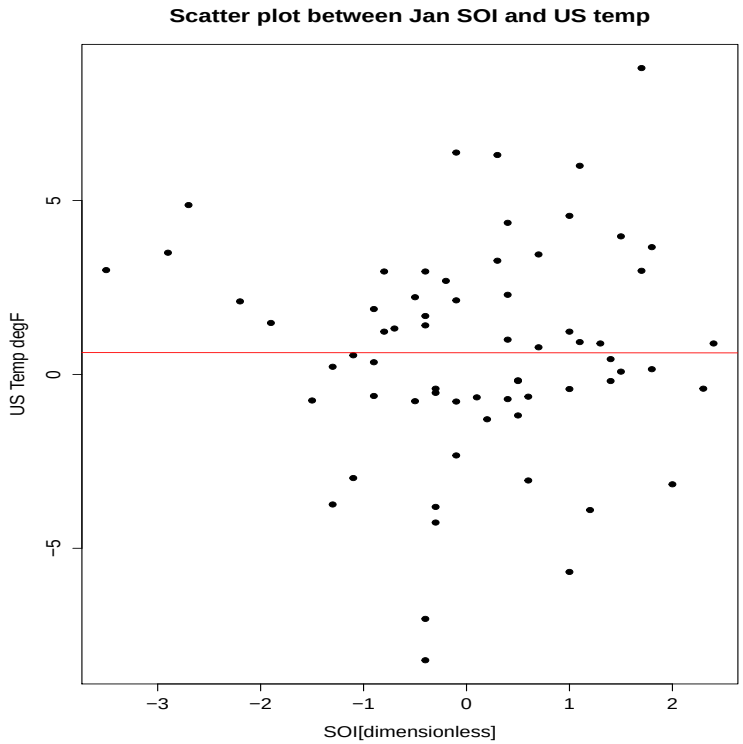

Figure 2.4 Scatter plot of the US January temperature vs. SOI from 1951-2016.

The scatter plot shows that the nearly zero correlation is mainly due to the five negative SOI values, which are El Nino Januarys: 1983 (-3.5), 1992 (-2.9), 1998 (-2.7), 2016 (-2.2), 1958 (-1.9). When these strong El Nino Januarys are removed, then the correlation is 0.2. The slope is then 0.64, compared with 1.0 for perfect correlation.

The R commands to retain the data without the above five El Nino years are below
`soijc=soij[c(1:7,9:32,34:41,43:47,49:65)]`
`ustjc=ustj[c(1:7,9:32,34:41,43:47,49:65)]` With these data, the scatter plot and trend line can be produced in the same way.

We thus may say that the SOI has some prediction skill for the contiguous U.S.' January temperature for the non-El Nino years. This correlation is stronger for particular regions of the U.S. since temperature field over the U.S. is inhomogeneous and is related to the tropical ocean dynamics in different ways. This gives us a hint to find out the prediction skill for an objective field: to plot a scatter plot between the objective field and the field used for prediction. The objective field is called predicant and the field used for prediction is called predictor. A very useful prediction skill is that predictor leads predicant by a time, say a month. Then the scatter plot will be made from the pairs between predictor and predicant data with one month lead. The absolute value of correlation can then be used as a measure of prediction skill. Since 1980s, the U.S. Climate Prediction Center has been using sea surface temperature (SST) and sea level pressure (SLP) as predictors for the U.S. temperature and precipitation via the canonical correlation analysis method. Therefore, before a prediction is made, it is a good idea to examine the prediction skill via scatter plots, which can help identify the best predictors.

However, the scatter plot approach above for maximum correlation is only applicable for linear prediction or weakly nonlinear relationships. Nature can sometimes be very nonlinear, which require more sophisticated assessments of prediction skill, such as neural networks and time-frequency analysis.

2.5 Data matrices and SVD by R

A matrix is a table such as that shown in Fig. 2.5, consisting of N-rows and Y-columns of numbers, which are called elements. Figure 2.5 shows the 10 years' annual precipitation anomalies from 1900-1909 for the 5 lat-lon degree boxes centered at $2.5^\circ E$ for different latitude over the Northern Hemisphere from $2.5^\circ N$ to $72.5^\circ N$.

Lat	Lon	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
2.5	2.5	0.283240	-0.131860	-0.190500	0.160040	-0.878110	0.080356	0.059193	-0.136900	0.200420	0.822600
7.5	2.5	0.172670	0.830550	-0.180350	-0.203630	-0.238590	0.425310	0.002805	0.102780	0.254050	0.516200
12.5	2.5	0.024392	0.152030	-0.034115	-0.062696	-0.192070	0.074360	0.201970	-0.011311	0.035259	0.272010
17.5	2.5	0.006780	0.066783	-0.084581	-0.008636	-0.038109	-0.001092	0.088250	0.011047	0.029358	0.082329
22.5	2.5	0.021162	0.079977	0.020016	-0.022142	-0.027032	0.065704	0.012937	-0.003823	0.032545	0.028636
27.5	2.5	0.049846	0.057413	0.026621	0.019914	-0.002651	0.071242	0.012837	0.001567	0.051857	0.099650
32.5	2.5	0.107740	0.143510	0.061613	0.076137	0.147760	0.137890	-0.074612	0.110300	0.087752	0.126920
37.5	2.5	0.128250	0.211940	0.113010	0.027472	0.183710	0.125550	-0.267500	0.215980	0.007609	0.055573
42.5	2.5	0.158490	0.800950	0.292690	0.172930	0.272010	0.126370	-0.017183	0.184880	0.118980	0.200520
47.5	2.5	-0.112800	0.243130	-0.121630	-0.076247	-0.047231	0.110160	0.080978	-0.091371	0.016172	-0.060487
52.5	2.5	-0.199840	-0.381070	-0.217570	-0.107760	-0.124700	-0.117470	-0.062448	-0.171070	-0.277650	-0.132690
57.5	2.5	-0.076619	-0.515070	0.005342	0.016647	0.137820	0.038041	0.131370	-0.196490	-0.132480	0.014887
62.5	2.5	-0.261760	-0.402600	0.137200	-0.214960	0.249210	0.147550	0.866120	-0.453910	-0.026134	0.053409
67.5	2.5	0.034079	0.223610	0.314090	-0.044832	0.130470	0.201260	0.554170	-0.054434	0.185870	0.308950
72.5	2.5	-0.119680	0.022949	0.004324	-0.050248	0.251330	-0.233080	-1.043800	0.363850	-0.315400	-0.113080

Figure 2.5 Annual precipitation anomalies data of the Northern Hemisphere at longitude $2.5^\circ E$ [Units: mm/day]. The annual total of the anomalies should be multiplied by 365.

Precipitation data [Units: mm/day] at multiple stations and multiple days also form a matrix, normally with stations [marked by station ID] counted in rows and time

[Units: day] counted in columns. The daily minimum surface air temperature (Tmin) data for the same stations and the same period of time form another matrix. In general, the space-time climate data table always forms a matrix. Conventionally, the spatial location determines the row, and the time coordinate determines the column.

Another daily life matrix example is that the ages data of the audience sitting in a movie theater of rows and columns of chairs form a matrix. Their weights form another matrix. Their bank account balance still another, and so on. So, matrix is a data table, and mathematical theories have been developed to study matrices in the 20th century. Computer programs, such

A slightly higher level of matrix concept is the coefficient matrix of a group of linear equations. Solving linear equations is very common in science and engineering. Any numerical solutions of a partial differential equation climate model will have to solve a set of linear equations. The famous Leontief's economic supply-demand balance model is a set of linear equations that can be written in the matrix form.

A simple elementary school kid's problem reads like this: The sum of two brothers' age is 20 years and their difference is 4. What are the ages of the brothers? One can easily guess that the older brother is 12 and the younger one is 8. An eight-year old kid can most likely figure this out. The idea can be extended to a more general form of linear equations.

If we form a set of equations, which would be

$$\begin{aligned} x_1 + x_2 &= 20 \\ x_1 - x_2 &= 4 \end{aligned} \tag{2.6}$$

when x_1 and x_2 stand for the brothers ages.

The matrix form of these equations would be

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.7}$$

which involves three matrices:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 20 \\ 4 \end{bmatrix}. \tag{2.8}$$

Most linear algebra textbooks introduce matrix in this way by describing a linear equation, which is less intuitive for in the field of climate science and policy, which emphasizes data.

The single column n-row matrix is often called an n-dimensional vector.

Although one can easily guess the solution to the matrix equation is $x_1 = 12$ and $x_2 = 8$, a more consistent computing may be done by R using the following commands

```
A<-matrix(seq(1:4),2)
b<-seq(1:2)
A[1,1]=1
A[1,2]=1
A[2,1]=1
A[2,2]=-1
b[1]=20
b[2]=4
```

```
solve(A,b)
#[1] 12  8 This is the result x1=12, and x2=8.
```

This kind of R computer program can solve much more complicated linear equations, such as an equation of 1,000 unknowns rather than 2 in this example.

This solution may be represented as

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \tag{2.9}$$

where A^{-1} is the inverse matrix of $\mathbf{A}$, i.e.,

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{2.10}$$

where

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{2.11}$$

is called identity matrix, which is like value 1.0 in our commonly used real number system.

This chapter will discuss following topics: (i) matrix algebra of addition, subtraction, multiplication and division (i.e., inverse matrix) and linear transform), (ii) space-time decomposition of a space-time climate data matrix, (iii) interpretation of the space-time decomposition using empirical orthogonal functions (EOFs) and principal components (PCs), (iv) matrix application in balancing the mass in chemical reaction equation, and (v) the matrix application in multivariate linear regression.

2.5.1 Matrix algebra and echelon form of a matrix

Addition, subtraction, multiply a matrix by a scalar constant, multiply a matrix by a matrix, a matrix divided by a matrix, and matrix inverse.

2.5.2 Independent row vectors and row echelon form

Example 1:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \tag{2.12}$$

Example 2: 5-deg global precipitation data

```
http://www.ats.ucla.edu/stat/r/library/matrix_alg.htm
```

2.6 Eigenvalues and eigenvectors of a square space matrix

Eigenvalue is a German word, meaning “self-value” or “proper value”. ”Eigen” means ”self”, ”my”.

A square matrix means that the number of rows is equal to the number of columns. The matrix is thus a square. Other matrices may be called rectangular matrices, or tall matrices.

Climate science often considers the covariance or correlation among N stations or grid boxes. The covariance matrix is thus a square matrix. The ij-th element is equal

to the covariance of the i-th station with the j-th station. If Y is the time length, say Y years, of the anomaly data $A_{N\times Y}$, then the covariance matrix is

$$C = AA'/Y. \tag{2.13}$$

EXAMPLE 2.1

```
A=matrix(c(1,-1,2,0,3,1),nrow=2)
A
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]   -1    0    1
covm=(1/(dim(A)[2]))*A%*%t(A)
covm #is the covariance matrix.
          [,1]      [,2]
[1,] 4.6666667 0.6666667
[2,] 0.6666667 0.6666667
u=c(1,-1)
v=covm%*%x
v
     [,1]
[1,]    4
[2,]    0
#u and v are in different directions.
```

In general, a given square matrix C and a given vector x, the product Cx is usually not parallel to x, as shown in the above example. A special case is the matrix's self-vector: w whose Cw is parallel to w, i.e.,

$$Cw = \lambda w, \tag{2.14}$$

where λ is a scalar, which scales w so that the above equation holds and is called eigenvalue, and w is called eigenvector.

R can calculate the eigenvalues and eigenvectors of the above covariance matrix `covm` with a command

```
eigen(covm)
$values
[1] 4.7748518 0.5584816
$vectors
           [,1]       [,2]
[1,] -0.9870875  0.1601822
[2,] -0.1601822 -0.9870875
```

A 2-by-2 order covariance has two eigenvalues, and two eigenvectors: (λ_1, w_1) and (λ_2, w_2):

$$\lambda_1 = 4.7748518, \quad \mathbf{w_1} = \begin{bmatrix} -0.9870875 \\ -0.1601822 \end{bmatrix}, \tag{2.15}$$

$$\lambda_2 = 0.5584816, \quad \mathbf{w_2} = \begin{bmatrix} 0.1601822 \\ -0.9870875 \end{bmatrix}. \tag{2.16}$$

Table 2.1 Space-time data table

	Time 1	Time 2	Time 3	Time 4
Space 1	D11	D12	D13	$D14$
Space 2	D21	D22	D23	$D24$
Space 3	D31	D32	D33	$D34$
Space 4	D41	D42	D43	$D44$
Space 5	D51	D52	D53	$D54$

The eigenvectors are called modes, or empirical orthogonal functions (EOFs). The first a few eigenvectors of large climate covariance matrix often represent some typical climate dynamic patterns, such as El Nino Southern Oscillation (ENSO), North America Oscillation (NAO), and Pacific Decadal Oscillation (PDO).

It is usually that the first mode's components have the same sign, all positive or all negative. The second mode's components have half negative and half positive. Exceptions can happen.

The eigenvalues for a climate covariance matrix are always positive. The sum of all the eigenvalues represent the total variance of the climate system of these N stations.

However, in the climate data analysis, one can find the climate dynamic patterns as eigenvectors more directly from the anomaly data matrix A without computing the covariance matrix C explicitly. This is the singular value decomposition (SVD) approach that separates the space-time anomaly data into space part, time part, and variation energy part. This mathematical law of space-time decomposition is universally applicable to all data we sample in space-time and can help explore the insight physics or science recorded by the data. Efficient computing methods of SVD were extensively researched and developed since 1960s. The leading figure of the field was Gene H. Golub (1932-2007).

2.6.1 An SVD representation model for space-time data

We encounter space-time data every day, such as the air temperature at different locations at different time: the temperature at San Diego in the morning and that at New York at night after your arrival. We may need to examine the precipitation conditions around the world at different days to monitor the agricultural yield. Cellphone companies may need to monitor its market share and its temporal variations at different countries. A doctor may need to monitor a patient's temperature change at different parts: hands, feet, forehead, and mouth. The observed data form a space-time data matrix with the row position corresponding to the spatial location and the column position corresponding to time. See Table 2.1.

Graphically, the space-time data are usually plotted in time series according to each given spatial position, or a spatial map according to each given time. Although these straight forward graphical representation can sometimes provide very useful information for signal detection, such as abnormal conditions indicating a certain decease of a patient, the signals are often buried inside the data and need to be detected by different linear combinations in space and time. Sometimes the data matrix are very big, millions of dimension in either space or time. Then what is the essential information in this big data matrix? Can we distill the most important information and represent

the data in a simpler way but more useful way? A very useful way is the space-time separation. Singular value decomposition (SVD) is a designed for this purpose. SVD decomposes a space-time data matrix into a spatial pattern matrix U, a diagonal energy level matrix D, and a temporal matrix V', i.e., the data matrix A is decomposed into

$$A_{n\times t} = U_{n\times m} D_{m\times m} (V')_{m\times t}. \tag{2.17}$$

where n is the spatial dimension, t is the temporal length, $m = min(n, t)$, and V' the transpose of V.

EXAMPLE 2.2

```
#Develop a 2-by-3 space-time data matrix
A=matrix(c(1,-1,2,0,3,1))
A
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]   -1    0    1
#Perform SVD calculation
msvd=svd(A)
msvd
$d
[1] 3.784779 1.294390
$u
           [,1]       [,2]
[1,] -0.9870875 -0.1601822
[2,] -0.1601822  0.9870875
$v
           [,1]       [,2]
[1,] -0.2184817 -0.8863403
[2,] -0.5216090 -0.2475023
[3,] -0.8247362  0.3913356
#One can verify that A=UDV'.
verim=msvd$u%*%diag(msvd$d)%*%t(msvd$v)
verim
     [,1]         [,2] [,3]
[1,]    1 2.000000e+00    3
[2,]   -1 1.665335e-16    1
round(verim)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]   -1    0    1
#This is the original data matrix A
```

The covariance of the space-time matrix A is a spatial matrix:

$$C = \frac{1}{Y} AA', \tag{2.18}$$

where Y is the number of columns of A and is the time length.

```
covm=(1/(dim(A)[2])*A%*%t(A)
eigcov=eigen(covm)
eigcov
$values
[1] 4.7748518 0.5584816
$vectors
           [,1]       [,2]
[1,] -0.9870875  0.1601822
[2,] -0.1601822 -0.9870875
```

Thus, the covariance matrix' eigenvectors are the same as the SVD eigenvectors of the anomaly matrix. The eigenvalues of covariance matrix and the SVD have following relationship

```
((msvd$d)^2)/(dim(A)[2])=eigcov$values
[1] 4.7748518 0.5584816
```

This is formally stated and proved below.

Theorem 2.1 *The covariance's matrix $C = AA'/Y$'s eigenvectors are the same as the spatial modes:*

$$\mathbf{C}\mathbf{u}_k = \lambda_i \mathbf{u}_k, \ \ (k = 1, 2, \cdots, N), \tag{2.19}$$

and the eigenvalues $lambda_k$ of C and those d_k of A from SVD have the following relationship

$$\lambda_k = d_k^2/Y \ \ (k = 1, 2, \cdots, N), \tag{2.20}$$

where Y is the total time length (i.e., time dimension) of the anomaly data matrix A, and N is the total number of stations for A (i.e., space dimension): $A_{N\times Y}$ and $C_{N\times N}$

Proof:

$$\begin{aligned} A &= UDV' && (2.21)\\ C &= \frac{1}{Y}AA' = \frac{1}{Y}UDV'(UDV')' \\ &= \frac{1}{Y}UDV'(VDU') = \frac{1}{Y}UDDU' = \frac{1}{Y}D^2UU' && (2.22)\\ CU &= \frac{1}{Y}D^2UU'U = \frac{1}{Y}D^2U = \Lambda U && (2.23)\\ \Lambda &= \frac{1}{Y}D^2. && (2.24) \end{aligned}$$

In the above, D^2 is allowed to be moved in front of U because D is a diagonal matrix.

Thus

$$\lambda_k = \frac{d_k^2}{Y} \ \ (k = 1, 2 \cdots, N). \tag{2.25}$$

2.6.2 SVD analysis of Southern Oscillation Index

This section is an SVD approach to construct an optimally weighted Southern Oscillation Index (WSOI).

SOI is an index that monitors ENSO and is the standardized Tahiti (18S, 149W) sea level pressure (SLP) minus that of Darwin (12S, 131E). It measures the SLP difference

between the eastern tropical Pacific and the western tropical Pacific. During a normal year, Darwin's anomaly pressure is lower than that of Tahiti, which maintains the easterlies trade wind and supports the western Pacific warm pool. When the wind reverses, the pressure anomalies have opposite order, leading to the westerlies trade wind and the accumulation of warm water over the eastern tropical Pacific or central tropical Pacific. This is El Nino.

The SLP data of these two stations can be downloaded from
`http://www.cpc.ncep.noaa.gov/data/indices/`

This section uses SVD approach to analyzing the standardized Tahiti and Darwin SLP and developing WSOI.

```
# Read the txt data
Pta<-read.table("~/Desktop/MyDocs/teach/336MathModel-2016SP/
BookMathModeling2016/R-code4MathModelBook/Ch5-SOI/PSTANDtahiti", header=F)
# Remove the first column that is the year
ptamon<-Pta[,seq(2,13)]
#Convert the matrix into a vector according to mon: Jan 1951, Feb 1951, ..., Dec 2015
ptamonv<-c(t(ptamon))
xtime<-seq(1951, 2016-1/12, 1/12)
# Plot the Tahiti standardized SLP anomalies
plot(xtime, ptamonv,type="l",xlab="Year",ylab="Presure",
 main="Standardized Tahiti SLP Anomalies", col="red",
 xlim=range(xtime), ylim=range(ptamonv))
# Do the same for Darwin SLP
Pda<-read.table("~/Desktop/MyDocs/teach/336MathModel-2016SP/
BookMathModeling2016/R-code4MathModelBook/Ch5-SOI/PSTANDdarwin.txt", header=F)
pdamon<-Pda[,seq(2,13)]
pdamonv<-c(t(pdamon))
plot(xtime, pdamonv,type="l",xlab="Year",ylab="Presure",
 main="Standardized Darwin SLP Anomalies", col="blue",
 xlim=range(xtime), ylim=range(pdamonv))
#Plot the SOI index
plot(xtime, ptamonv-pdamonv,type="l",xlab="Year",
ylab="SOI index", col="black",xlim=range(xtime), ylim=c(-4,4), lwd=1)
#Add ticks on top edge of the plot box
axis (3, at=seq(1951,2015,4), labels=seq(1951,2015,4))
#Add ticks on the right edge of the plot box
axis (4, at=seq(-4,4,2), labels=seq(-4,4,2))
# If put a line on a plot, use the command below
lines(xtime,ptamonv-pdamonv,col="black", lwd=1)
```

The accumulative SOI, denoted by CSOI, has a nonlinear trend similar to that of SST over North Atlantic (80W-0, 30-60N). See CPC report on Feb 9, 2016.

```
cnegsoi<--cumsum(ptamonv-pdamonv)
plot(xtime, cnegsoi,type="l",xlab="Year",ylab="Negative CSOI index",
col="black",xlim=range(xtime), ylim=range(cnegsoi), lwd=1)
```

The space-time data matrix of the SLP at Tahiti and Darwin from January 1951-December 2015 can be obtained from

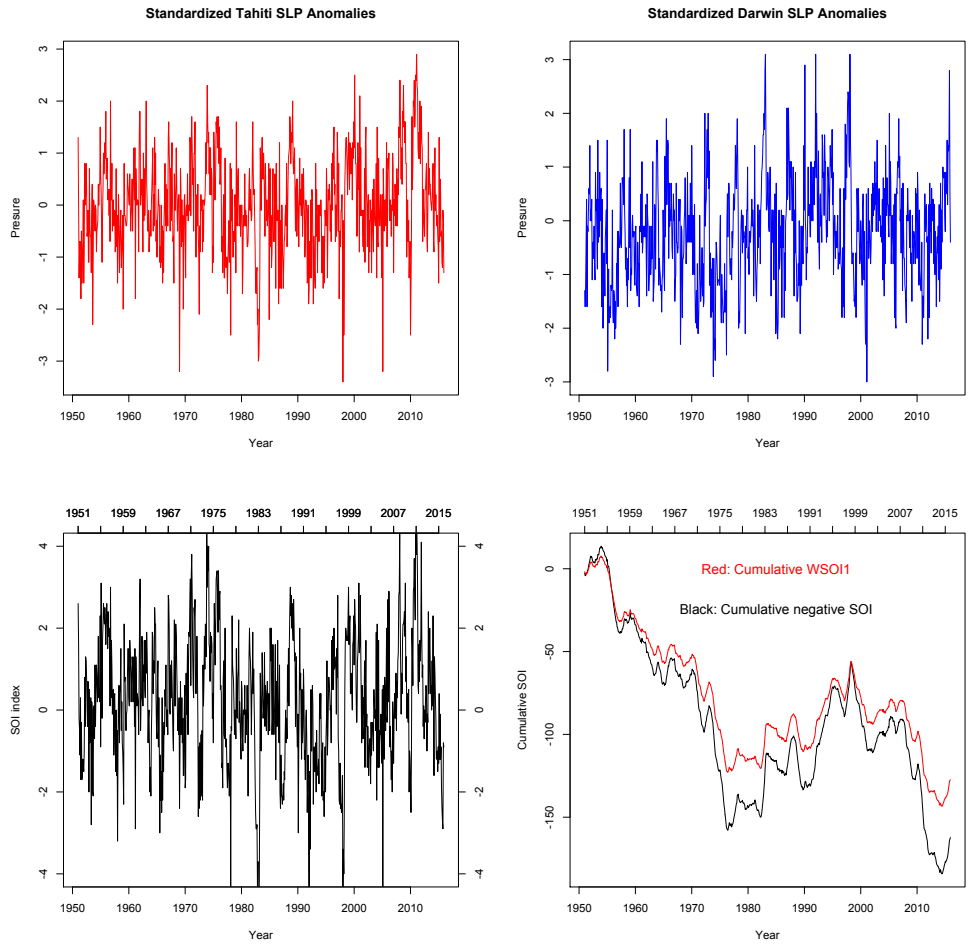

Figure 2.6 Standardized sea level pressure anomalies of Tahiti (up-left panel), that of Darwin (up-right). SOI time series (low-left), and the cumulative of the negative SOI time series (low-right).

```
ptada <-cbind(ptamonv,pdamonv)
```

This is a matrix of two columns: the first column is the Tahiti SLP and the second the Darwin. Because normally the spatial position is by row and time by column, we transpose the matrix `ptada<-t(ptada)` This is the 1951-2015 standardized SLP data for Tahiti and Darwin: 2 rows and 780 columns.

```
dim(ptada)
[1]   2 780
```

Make the SVD space-time separation: `svdptd<-svd(ptada)`
Verify this separation by reconstructing the original space-time data matrix using the SVD results

```
recontd=svdptd$u%*%diag(svdptd$d[1:2])%*%t(svdptd$v)
```

One can verify that `recontd=ptada`.

The spatial matrix U is a 2×2 orthogonal matrix since there are only two points. Each column is an eigenvector of the covariance matrix $C = AA'/t$, where $A_{n\times t}$ is the original data matrix of n spatial dimension and t temporal dimension. These eigenvectors are spatial patterns, called empirical orthogonal function (EOF) in atmospheric sciences. Our U matrix is

```
U=svdptd$u
U
            [,1]      [,2]
[1,] -0.6146784 0.7887779
[2,]  0.7887779 0.6146784
```

The first column is the first spatial mode is $\mathbf{u}_1 = (-0.61, 0.79)$, meaning opposite signs of Tahiti and Darwin, which justifies the SOI index as one pressure minus another. This result further suggests that a better index should be the weighted SOI:

$$WSOI1 = -0.6147P_{Tahiti} + 0.7888P_{Darwin} \tag{2.26}$$

This mode's energy level, i.e., the temporal variance, is $d_1 = 31.35$ given by

```
svdptd$d
[1] 31.34582 22.25421
D=diag(svdptd$d)
D
         [,1]     [,2]
[1,] 31.34582  0.00000
[2,]  0.00000 22.25421
```

which forms the diagonal matrix D in the SVD formula. In the nature, the second eigenvalue is often much smaller than the first, but not this one. The second mode's energy level is $d_2 = 22.25$, equal to 71% of the first energy level.

The second weighted SOI mode, i.e. the second column $\mathbf{u}_2$ of U, is thus

$$WSOI2 = 0.7888P_{Tahiti} + 0.6147P_{Darwin} \tag{2.27}$$

From the SVD formula $A = UDV'$, the above two weighted SOIs are $U'A$:

$$U'A = DV', \tag{2.28}$$

because U is a an orthogonal matrix and $U^{-1} = U'$.

The V matrix is given by

```
V=svdptd$v
V
                [,1]           [,2]
  [1,] -5.820531e-02  1.017018e-02
  [2,] -4.026198e-02 -4.419324e-02
  [3,] -2.743069e-03 -8.276652e-02
  ......
```

The first temporal mode $\mathbf{v}_1$ is the first row of V' and is called the first principal component (PC1). The above formulas imply that

$$\mathbf{v}_1 = WSOI1/d_1 \quad (2.29)$$
$$\mathbf{v}_2 = WSOI2/d_2 \quad (2.30)$$

The two PCs are orthonormal vectors. So are the two EOFs. Thus, the SLP data at Tahiti and Darwin have been decomposed into a set of spatially and temporally orthonormal vectors: EOFs and PCs, together with energy levels.

The WSOIs' standard deviations are d_1 and d_2, reflecting the WSOI's oscillation magnitude and frequency.

We also have the relations

$$d_k PC_k = WSOI_k \ \ (k = 1, 2). \quad (2.31)$$

The two WSOIs are shown in Fig.2.7.

```
%Plot WSOI1
xtime<-seq(1951, 2016-1/12, 1/12)
wsoi1=D[1,1]*t(V)[1,]
plot(xtime, wsoi1,type="l",xlab="Year",ylab="Weighted SOI 1",
col="black",xlim=range(xtime), ylim=range(wsoi1), lwd=1)
axis (3, at=seq(1951,2015,4), labels=seq(1951,2015,4))
%Plot WSOI2
wsoi2=D[2,2]*t(V)[2,]
plot(xtime, wsoi2,type="l",xlab="Year",ylab="Weighted SOI 2",
 col="black",xlim=range(xtime), ylim=c(-2,2), lwd=1)
axis (3, at=seq(1951,2015,4), labels=seq(1951,2015,4))
```

The cumulative WSOIs can be plotted by the following R commands

```
%Plot cumulative WSOI1
cwsoi1=cumsum(wsoi1)
plot(xtime, cwsoi1,type="l",xlab="Year",ylab="Cumulated weighted SOI 1",
col="black",xlim=range(xtime), ylim=range(cwsoi1), lwd=1)
axis (3, at=seq(1951,2015,4), labels=seq(1951,2015,4))
%Plot cumulative WSOI2
cwsoi2=cumsum(wsoi2)
plot(xtime, cwsoi2,type="l",xlab="Year",ylab="Cumulated weighted SOI 2",
col="black",xlim=range(xtime), ylim=range(cwsoi2), lwd=1)
axis (3, at=seq(1951,2015,4), labels=seq(1951,2015,4))
```

The cumulative WSOI1 appears to trace the southern hemisphere's surface air temperature history, according to Jones' data
http://cdiac.ornl.gov/ftp/trends/temp/jonescru/sh.txt
http://cdiac.ornl.gov/trends/temp/jonescru/graphics/glnhsh.png

When the cumulative WSOI decreases, so does the SH surface air temperature from 1951 to 1980. When the cumulative WSOI increases, so does the temperature from the 1980s to the peak 1998. Then cumulative WSOI1 decreases to a plateau from 1998

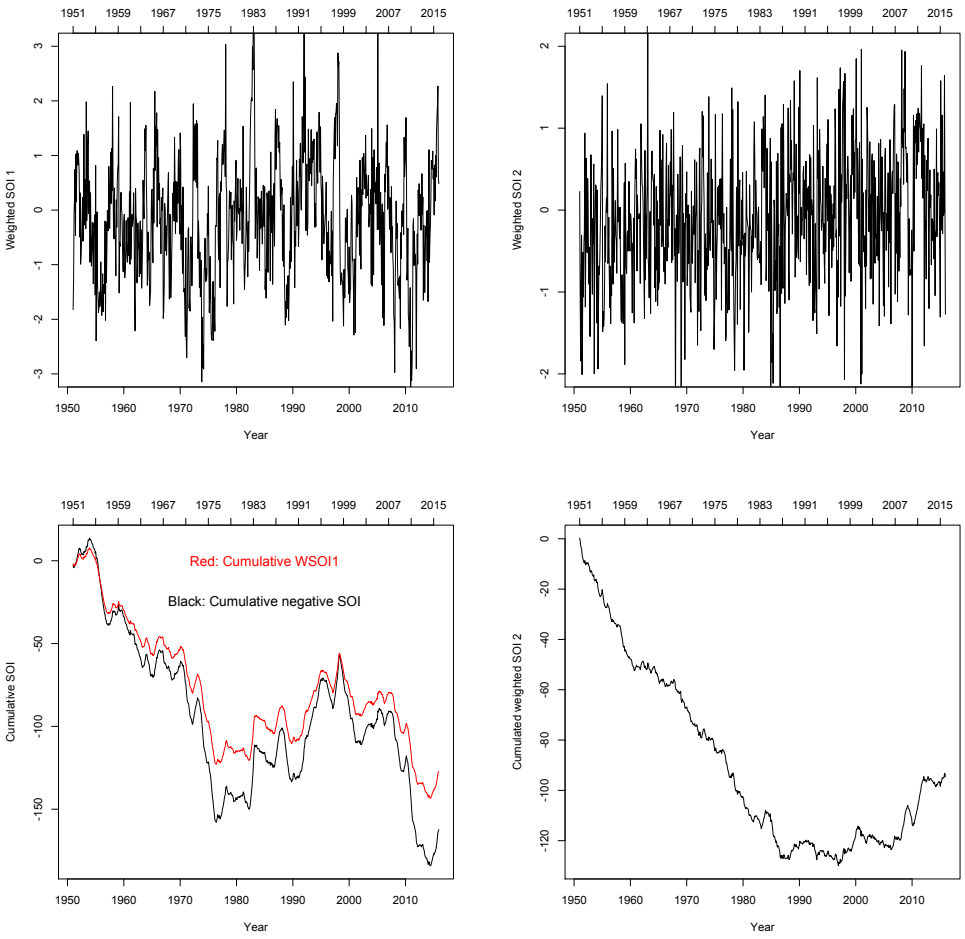

Figure 2.7 Weighted SOI1 (up-left panel), weighted SOI2 (up-right), cumulative WSOI1 (low-left), and cumulative WSOI2 (low-right).

to 2002, another plateau until 2007, then decreases again. This also agrees with the SH surface air temperature trend.

Therefore, SVD results may lead to physical meanings and is a convenient tool to use.

2.7 Visualization of SVD results: EOFs and PCs

We use three examples the visualization of EOFs and PCs from SVD results by `ggplot`.

EXAMPLE 2.3

The space-time data matrix `ptada` of the SLP at Tahiti and Darwin from January 1951-December 2015 has 2 rows and 780 columns.

Based on the previous section's results of the EOF modes, we have the following data and results

```
library(ggplot2)
dat=matrix(c(-18, -12,-149, 131, -0.61, 0.79, 0.79,0.61),nrow=2,
           + dimnames=list(c("Darwin","Tahiti"), c("lat","lon","EOF1","EOF2")))
ft=as.data.frame(dat)
dp=ggplot(ft,aes(lon,lat))
dp1=dp+geom_point(aes(colour=factor(EOF1)),cex=9)
+ xlim(-180,180) + ylim(-90,90)
dp1
```

This plots EOF1 shown in Fig. 2.8.

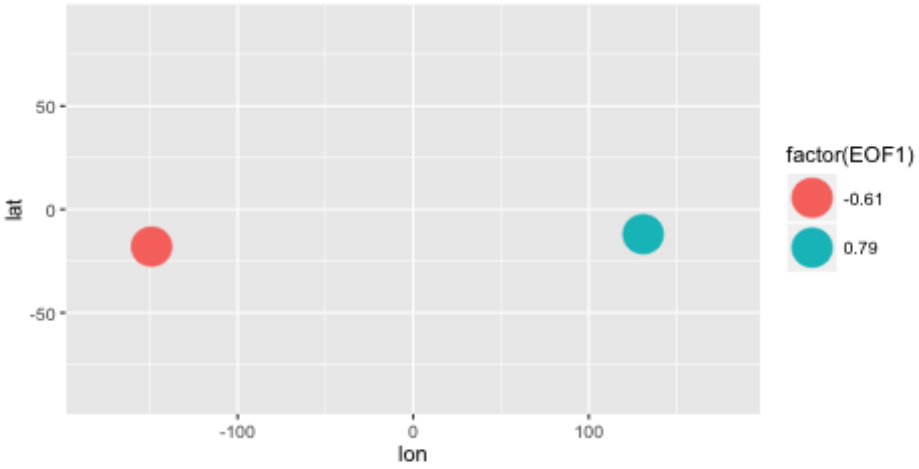

Figure 2.8 EOF1 of the Tahiti and Darwin standardized SLP data.

```
tdsvd<-svd(ptada)
dat=provideDimnames(tdsvd$u,base=list(c("Tahiti", "Darwin"),
            +    c("EOF1","EOF2")))
ft=as.data.frame(dat)
dp=ggplot(ft,aes(lon,lat))
dp1=dp+geom_point(aes(colour=factor(EOF1)),cex=9)
+ xlim(-180,180) + ylim(-90,90)
dp1#Show the plot of EOF1
plot(seq(1951, 2015, length=780), tdsvd$v[1,],
  +  xlab="Year", ylab="WSOI1", col="red",
  + main="PC1 as the weighted SOI")
```

EXAMPLE 2.4

NASA Global Precipitation Climatology Project (GPCP) precipitation data.

EXAMPLE 2.5

NCEP/NCAR Reanalysis data: wind data.

EXAMPLE 2.6

NOAA $5^\circ \times 5^\circ$ latitude-longitude gridded global monthly temperature data since January 1880.

2.8 SVD algorithms and their R codes

Read the following website for more about the SVD R code:

```
https://stat.ethz.ch/R-manual/R-devel/library/base/html/svd.html
```

Read this chapter for the SVD history and mathematics: Cline, A.K., and I.S. Dhillon, 2006: Computation of the singular value decomposition, in Handbook of Linear Algebra, edited by Leslie Hogben, Chapman and Hall/CRC, 45-1-45-13, DOI: 10.1201/9781420010572.ch45.

References and Additional Reading Materials

R2.1 R tutorial by Steve Jost, De Paul University,

```
http://facweb.cs.depaul.edu/sjost/csc423/
```

R2.2 R tutorials by William B. King, Coastal Carolina University,

```
http://ww2.coastal.edu/kingw/statistics/R-tutorials/
```

R2.2a R tutorial: Long and complete R courses.

```
https://www.coursera.org/learn/r-programming
http://swcarpentry.github.io/r-novice-inflammation/
```

R2.3 References for statistics using climate datasets

```
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm
http://empslocal.ex.ac.uk/people/staff/dbs202/cag/courses/MT37C/course-d.pdf
http://www.climate-service-center.de/imperia/md/content/csc/
projekte/csc-report13_englisch_final-mit_umschlag.pdf
```

EXERCISES

2.1 The spatial average of a field, e.g., temperature field $T(\phi, \theta, t)$, on the surface of Earth is defined by the following surface integral

$$\bar{T} = \frac{1}{4\pi R^2} \iint_E T(\phi, \theta, t) dS. \tag{2.32}$$

Here, Earth is regarded approximately as a sphere with radius R, the surface integral's domain of integration E is the Earth's surface, ϕ is latitude, θ longitude, and t is time. In practice, the function $T(\phi, \theta, t)$ is not analytically defined, but the data of $T(\phi, \theta, t)$ is given on a $5^\circ \times 5^\circ$ lat-lon grid. Each grid box is assigned an identity number $i = 1, 2, 3, \cdots, N = 2,592 = 36 \times 72$. Show that the area-weighted approximation, i.e., the Riemann sum of a surface integral, of the above integral is

$$\hat{\bar{T}}(t) = \sum_{i=1}^{N} w_i T(i, t) \tag{2.33}$$

where

$$w_i = \frac{\cos \phi_i}{\sum_{i=1}^{N} \cos \phi_i} \tag{2.34}$$

and ϕ_i is the latitude of the centoid of the grid box i and $T(i, t)$ is the average temperature of the grid box i.

2.2

a) Computed the area-weighted average monthly $2.5^\circ \times 2.5^\circ$ lat-lon grid precipitation anomalies for the Earth surface excluding the polar regions, i.e., only in the latitude bend $(75^\circ S, 75^\circ N)$, from January 1901 -December 2000 from the dataset downloaded from `http://shen.sdsu.edu/press.html`.

b) Plot the monthly time series of the above averages from January 1901-December 2000, and plot a trend line for the data. Discuss your trend results in text.

c) Compute and plot the time series of the monthly precipitation from January 1901 to December 2013 for San Diego County, California, the United States.

2.3

a) Make an SVD analysis for all the January 2.5 deg gridded precipitation data from January 1971 to January 2000.

b) Plot the eigenvalues according to the mode number.

c) Plot the first and second principal components of the above SVD results on the same figure.

d) Use R to verify that

$$V' \times V \tag{2.35}$$

is an expansion of an identify matrix with zeros. What is the dimension of the identity matrix?

e) Use R to verify that

$$U' \times U \tag{2.36}$$

is also an expansion of an identify matrix with zeros. What is the dimension of this identity matrix?

f) Finally use R to verify that

$$UDV' \tag{2.37}$$

can recover the original data matrix for January from 1971 to 2000.

2.4

a) The data of global average annual mean temperature anomalies temperature can be downloaded from

`http://www1.ncdc.noaa.gov/pub/data/noaaglobaltemp/operational/`

It is called "timeseries". Use the data file

`aravg.ann.land_ocean.90S.90N.v4.0.1.201607.asc`

and R to plot the time series of the temperature anomalies from 1880 to 2015.

b) Plot the linear trend line on the same figure and display the trend value in the unit of °C per decade for the following time periods: (i) 1880-2015, (ii) 1901-2000, (iii) 1931-1971, and (iv) 1981-2015.

2.5

a) Computed the area-weighted average monthly $5^\circ \times 5^\circ$ lat-lon grid surface air temperature anomalies for the entire Earth surface from January 1880 to December 2015 using the data that can be downloaded from

`http://www1.ncdc.noaa.gov/pub/data/noaaglobaltemp/operational/gridded/`

Note that this dataset has many missing data. Your area-weighted average is only for the grid boxes that have data.

b) Compare your computed averages with annual mean to the annual data in the previous problem, and describe the differences if there exist any.

c) Plot the monthly time series of the above averages and plot a trend line for the January data. Discuss your trend results in text.

2.6

a) Use the data in the above problem, compute the United States' monthly average temperature time series from January 1880 to December 2015.

b) Plot the above data against time, plot the trend line, and mark the trend in the unit [°C/per decade].

2.7 Let $E = D^2$ be a square of diagonal matrix D in an SVD decomposition $A = UDV'$. Find a relationship between trace of D and that of AA', when $A_{N\times t}$ is the $N \times t$ data matrix and $N < t$.

2.8 Make an SVD analysis for December from 1981-2010 for the NASA Global Precipitation Climatology Project (GPCP) precipitation data:

`http://precip.gsfc.nasa.gov/`

a) Perform SVD and produce matrices U, D, V.

b) Plot PCs based on V's columns and interpret the physical meaning of the result.

c) Plot EOFs based on U's columns and and interpret the physical meaning of the result.

2.9 Do an SVD analysis for a data matrix of your interest.