

Appendix C

Description of Data Sets

In this appendix, we describe the various data sets used in the book. All data sets are downloaded with the accompanying software. We provide the data sets in MATLAB binary format (MAT-files).

abrasion

These data are the results from an experiment measuring three variables for 30 rubber specimens [Davies, 1957; Cleveland and Devlin, 1988]: tensile strength, hardness and abrasion loss. The abrasion loss is the amount of material abraded per unit of energy. Tensile strength measures the force required to break a specimen (per unit area). Hardness is the rebound height of an indenter dropped onto the specimen. Abrasion loss is measured in g/hp-hour; tensile strength is measured in kg/cm²; and the unit for hardness is ° Shore. The intent of the experiment was to determine the relationship between abrasion loss with respect to hardness and tensile strength.

animal

This data set contains the brain weights and body weights of several types of animals [Crile and Quiring, 1940]. According to biologists, the relationship between these two variables is interesting, since the ratio of brain weight to body weight is a measure of intelligence [Becker, Cleveland and Wilks, 1987]. The MAT-file contains variables: **AnimalName**, **BodyWeight**, and **BrainWeight**.

BPM data sets

There are several BPM data sets included with the text. We have **iradbpm**, **ochiaibpm**, **matchbpm** and **L1bpm**. Each data file contains the interpoint distance matrix for 503 documents and an array of class labels, as described in [Chapter 1](#). These data can be reduced using ISOMAP before applying other analyses.

calibrat

This data set reflects the relationship between radioactivity counts (**counts**) to hormone level for 14 immunoassay calibration values (**tsh**). The original

source of the data is Tiede and Pagano [1979], and we downloaded them from the website for Simonoff [1996]: www.stern.nyu.edu/SOR/SmoothMeth.

cereal

These data were obtained from ratings of eight brands of cereal [Chakrapani and Ehrenberg, 1981; Venables and Ripley, 1994]. The **cereal** file contains a matrix where each row corresponds to an observation and each column represents one of the variables or the percent agreement to statements about the cereal. The statements are: comes back to, tastes nice, popular with all the family, very easy to digest, nourishing, natural flavor, reasonably priced, a lot of food value, stays crispy in milk, helps to keep you fit, fun for children to eat. It also contains a cell array of strings (**labs**) for the type of cereal.

environmental

This file contains data comprising 111 measurements of four variables. These include **Ozone** (PPB), **SolarRadiation** (Langleys), **Temperature** (Fahrenheit), and **WindSpeed** (MPH). These were initially examined in Bruntz, et al. [1974], where the intent was to study the mechanisms that might lead to air pollution.

ethanol

A single-cylinder engine was run with either ethanol or indolene. This data set contains 110 measurements of compression ratio (**Compression**), equivalence ratio (**Equivalence**), and NO_x in the exhaust (**NOx**). The goal was to understand how NO_x depends on the compression and equivalence ratios [Brinkman, 1981; Cleveland and Devlin, 1988].

example96

This is the data used in Example 9.6 to illustrate the box-percentile plots.

example104

This loads the data used in Examples 10.10 and 10.12 to illustrate parallel coordinate plots. It is a subset of the BPM data that was reduced using ISOMAP as outlined in Example 10.4.

forearm

These data consist of 140 measurements of the length in inches of the forearm of adult males [Hand, et al., 1994; Pearson and Lee, 1903].

galaxy

The **galaxy** data set contains measurements of the velocities of the spiral galaxy NGC 7531. The array **EastWest** contains the velocities in the east-west direction, covering around 135 arc sec. The array **NorthSouth** contains the velocities in the north-south direction, covering approximately 200 arc sec. The measurements were taken at the Cerro Tololo Inter-American Observatory in July and October of 1981 [Buta, 1987].

geyser

These data represent the waiting times (in minutes) between eruptions of the Old Faithful geyser at Yellowstone National Park [Hand, et al., 1994; Scott, 1992].

hamster

This data set contains measurements of organ weights for hamsters with congenital heart failure [Becker and Cleveland, 1991]. The organs are heart, kidney, liver, lung, spleen and testes.

iris

The **iris** data were collected by Anderson [1935] and were analyzed by Fisher [1936] (and many statisticians since then!). The data set consists of 150 observations containing four measurements based on the petals and sepals of three species of iris. The three species are: *Iris setosa*, *Iris virginica*, and *Iris versicolor*. When the **iris** data file is loaded, you get three 50×4 matrices, one corresponding to each species.

leukemia

The **leukemia** data set is described in detail in [Chapter 1](#). It measures the gene expression levels of patients with acute leukemia.

lsiex

This file contains the term-document matrix used in Example 2.3.

lungA, lungB

The **lung** data set is another one that measures gene expression levels. Here the classes correspond to various types of lung cancer.

oronsay

The **oronsay** data set consists of particle size measurements. It is described in Chapter 1. The data can be classified according to the sampling site as well as the type (beach, dune, midden).

playfair

This data set is described in Cleveland [1993] and Tufte [1983], and it is based on William Playfair's (1801) published displays of demographic and economic data. The **playfair** data set consists of the 22 observations representing the populations (thousands) of cities at the end of the 1700s and the diameters of the circles Playfair used to encode the population information. This MAT-file also includes a cell array containing the names of the cities.

pollen

This data set was generated for a data analysis competition at the 1986 Joint Meetings of the American Statistical Association. It contains 3848

observations, each with five fictitious variables: ridge, nub, crack, weight, and density. The data contain several interesting features and structures. See Becker, et al. [1986] and Slomka [1986] for information and results on the analysis of these artificial data.

posse

The **posse** file contains several data sets generated for simulation studies in Posse [1995b]. These data sets are called **croix** (a cross), **struct2** (an L-shape), **boite** (a donut), **groupe** (four clusters), **curve** (two curved groups), and **spiral** (a spiral). Each data set has 400 observations in 8-D. These data can be used in PPEDA and other data tours.

salmon

The **salmon** data set was downloaded from the website for the book by Simonoff [1996]: www.stern.nyu.edu/SOR/SmoothMeth. The MAT-file contains 28 observations in a 2-D matrix. The first column represents the size (in thousands of fish) of the annual spawning stock of Sockeye salmon along the Skeena River from 1940 to 1967. The second column represents the number of new catchable-size fish or recruits, again in thousands of fish.

scurve

This file contains data randomly generated from an S-curve manifold. See Example 3.5 for more information.

singer

This file contains several variables representing the height in inches of singers in the New York Choral Society [Cleveland, 1993; Chambers, et al., 1983]. There are four voice parts: sopranos, altos, tenors, and basses. The sopranos and altos are women, and the tenors and basses are men.

skulls

These data were taken from Cox and Cox [2000]. The data originally came from a paper by Fawcett [1901], where they detailed measurements and statistics of skulls belonging to the Naqada race in Upper Egypt. The **skulls** file contains an array called **skullsdata** for forty observations, 18 of which are female and 22 are male. The variables are greatest length, breadth, height, auricular height, circumference above the superciliary ridges, sagittal circumference, cross-circumference, upper face height, nasal breadth, nasal height, cephalic index, and ratio of height to length.

software

This file contains data collected on software inspections. The variables are normalized by the size of the inspection (the number of pages or SLOC – single lines of code). The file **software.mat** contains the preparation time in minutes (**prepage**, **prepsloc**), the total work hours in minutes for the

meeting (**mtgsloc**), and the number of defects found (**defpage**, **defsloc**). A more detailed description can be found in [Chapter 1](#).

spam

These data were downloaded from the UCI Machine Learning Repository:

<http://www.ics.uci.edu/~mlearn/MLRepository/.html>

Anyone who uses email understands the problem of spam, which is unsolicited email, commercial or otherwise. For example, spam can be chain letters, pornography, advertisements, foreign money-making schemes, etc. This data set came from Hewlett-Packard Labs and was generated in 1999. The **spam** data set consists of 58 variables: 57 continuous and one class label. If an observation is labeled class 1, then it is considered to be spam. If it is of class 0, then it is not considered spam. The first 48 attributes represent the percentage of words in the email that match some specified word corresponding to spam or not spam. There are an additional six variables that specify the percentage of characters in the email that match a specified character. Others refer to attributes relating to uninterrupted sequences of capital letters. More information on the attributes is available at the above internet link. One can use these data to build classifiers that will discriminate between spam and non-spam emails. In this application, a low false positive rate (classifying an email as spam when it is not) is very important.

sparrow

These data are taken from Manly [1994]. They represent some measurements taken on sparrows collected after a storm on February 1, 1898. Eight morphological characteristics and the weight were measured for each bird, five of which are provided in this data set. These are on female sparrows only. The variables are total length, alar extent, length of beak and head, length of humerus, and length of keel of sternum. All lengths are in millimeters. The first 21 of these birds survived, and the rest died.

swissroll

This data file contains a set of variables randomly generated from the Swiss roll manifold with a hole in it. It also has the data in the reduced space (from ISOMAP and HLL) that was used in Example 3.5.

votfraud

These data represent the Democratic over Republican pluralities of voting machine and absentee votes for 22 Philadelphia County elections. The variable **machine** is the Democratic plurality in machine votes, and the variable **absentee** is the Democratic plurality in absentee votes [Simonoff, 1996].

yeast

The **yeast** data set is described in [Chapter 1](#). It contains the gene expression levels over two cell cycles and five phases.