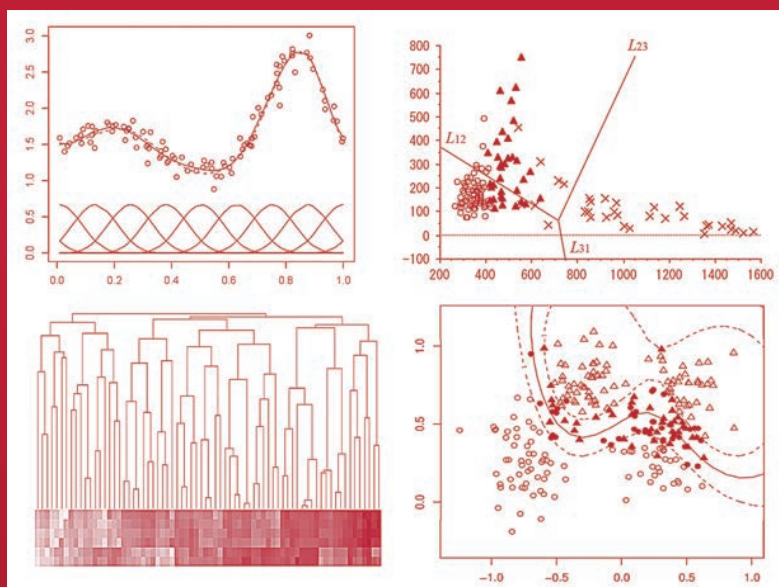


Texts in Statistical Science

# Introduction to Multivariate Analysis

Linear and Nonlinear Modeling



Sadanori Konishi



CRC Press  
Taylor & Francis Group

A CHAPMAN & HALL BOOK

# **Introduction to Multivariate Analysis**

**Linear and Nonlinear Modeling**

## CHAPMAN & HALL/CRC

### Texts in Statistical Science Series

Series Editors

Francesca Dominici, *Harvard School of Public Health, USA*

Julian J. Faraway, *University of Bath, UK*

Martin Tanner, *Northwestern University, USA*

Jim Zidek, *University of British Columbia, Canada*

#### **Statistical Theory: A Concise Introduction**

F. Abramovich and Y. Ritov

#### **Practical Multivariate Analysis, Fifth Edition**

A. Afifi, S. May, and V.A. Clark

#### **Practical Statistics for Medical Research**

D.G. Altman

#### **Interpreting Data: A First Course in Statistics**

A.J.B. Anderson

#### **Introduction to Probability with R**

K. Baclawski

#### **Linear Algebra and Matrix Analysis for Statistics**

S. Banerjee and A. Roy

#### **Statistical Methods for SPC and TQM**

D. Bissell

#### **Bayesian Methods for Data Analysis, Third Edition**

B.P. Carlin and T.A. Louis

#### **Second Edition**

R. Caulcutt

#### **The Analysis of Time Series: An Introduction, Sixth Edition**

C. Chatfield

#### **Introduction to Multivariate Analysis**

C. Chatfield and A.J. Collins

#### **Problem Solving: A Statistician's Guide, Second Edition**

C. Chatfield

#### **Statistics for Technology: A Course in Applied Statistics, Third Edition**

C. Chatfield

#### **Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians**

R. Christensen, W. Johnson, A. Branscum,  
and T.E. Hanson

#### **Modelling Binary Data, Second Edition**

D. Collett

#### **Modelling Survival Data in Medical Research, Second Edition**

D. Collett

#### **Introduction to Statistical Methods for Clinical Trials**

T.D. Cook and D.L. DeMets

#### **Applied Statistics: Principles and Examples**

D.R. Cox and E.J. Snell

#### **Multivariate Survival Analysis and Competing Risks**

M. Crowder

#### **Statistical Analysis of Reliability Data**

M.J. Crowder, A.C. Kimber,  
T.J. Sweeting, and R.L. Smith

#### **An Introduction to Generalized Linear Models, Third Edition**

A.J. Dobson and A.G. Barnett

#### **Nonlinear Time Series: Theory, Methods, and Applications with R Examples**

R. Douc, E. Moulines, and D.S. Stoffer

#### **Introduction to Optimization Methods and Their Applications in Statistics**

B.S. Everitt

#### **Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models**

J.J. Faraway

#### **A Course in Large Sample Theory**

T.S. Ferguson

#### **Multivariate Statistics: A Practical Approach**

B. Flury and H. Riedwyl

#### **Readings in Decision Analysis**

S. French

#### **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition**

D. Gamerman and H.F. Lopes

#### **Bayesian Data Analysis, Third Edition**

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson,  
A. Vehtari, and D.B. Rubin

#### **Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists**

D.J. Hand and C.C. Taylor

**Practical Data Analysis for Designed Practical Longitudinal Data Analysis**

D.J. Hand and M. Crowder

**Logistic Regression Models**

J.M. Hilbe

**Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects**

J.S. Hodges

**Statistics for Epidemiology**

N.P. Jewell

**Stochastic Processes: An Introduction, Second Edition**

P.W. Jones and P. Smith

**The Theory of Linear Models**

B. Jørgensen

**Principles of Uncertainty**

J.B. Kadane

**Graphics for Statistics and Data Analysis with R**

K.J. Keen

**Mathematical Statistics**

K. Knight

**Introduction to Multivariate Analysis: Linear and Nonlinear Modeling**

S. Konishi

**Nonparametric Methods in Statistics with SAS Applications**

O. Korosteleva

**Modeling and Analysis of Stochastic Systems, Second Edition**

V.G. Kulkarni

**Exercises and Solutions in Biostatistical Theory**

L.L. Kupper, B.H. Neelon, and S.M. O'Brien

**Exercises and Solutions in Statistical Theory**

L.L. Kupper, B.H. Neelon, and S.M. O'Brien

**Design and Analysis of Experiments with SAS**

J. Lawson

**A Course in Categorical Data Analysis**

T. Leonard

**Statistics for Accountants**

S. Letchford

**Introduction to the Theory of Statistical Inference**

H. Liero and S. Zwanzig

**Statistical Theory, Fourth Edition**

B.W. Lindgren

**Stationary Stochastic Processes: Theory and Applications**

G. Lindgren

**The BUGS Book: A Practical Introduction to Bayesian Analysis**

D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter

**Introduction to General and Generalized Linear Models**

H. Madsen and P. Thyregod

**Time Series Analysis**

H. Madsen

**Pólya Urn Models**

H. Mahmoud

**Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition**

B.F.J. Manly

**Introduction to Randomized Controlled Clinical Trials, Second Edition**

J.N.S. Matthews

**Statistical Methods in Agriculture and Experimental Biology, Second Edition**

R. Mead, R.N. Curnow, and A.M. Hasted

**Statistics in Engineering: A Practical Approach**

A. V. Metcalfe

**Beyond ANOVA: Basics of Applied Statistics**

R.G. Miller, Jr.

**A Primer on Linear Models**

J.F. Monahan

**Applied Stochastic Modelling, Second Edition**

B.J.T. Morgan

**Elements of Simulation**

B.J.T. Morgan

**Probability: Methods and Measurement**

A. O'Hagan

**Introduction to Statistical Limit Theory**

A.M. Polansky

**Applied Bayesian Forecasting and Time Series Analysis**

A. Pole, M. West, and J. Harrison

**Statistics in Research and Development, Time Series: Modeling, Computation, and Inference**

R. Prado and M. West

**Introduction to Statistical Process Control**

P. Qiu

**Sampling Methodologies with Applications**

P.S.R.S. Rao

**A First Course in Linear Model Theory**

N. Ravishanker and D.K. Dey

**Essential Statistics, Fourth Edition**

D.A.G. Rees

**Stochastic Modeling and Mathematical**

**Statistics: A Text for Statisticians and**

**Quantitative**

F.J. Samaniego

**Statistical Methods for Spatial Data Analysis**

O. Schabenberger and C.A. Gotway

**Large Sample Methods in Statistics**

P.K. Sen and J. da Motta Singer

**Decision Analysis: A Bayesian Approach**

J.Q. Smith

**Analysis of Failure and Survival Data**

P.J. Smith

**Applied Statistics: Handbook of GENSTAT**

**Analyses**

E.J. Snell and H. Simpson

**Applied Nonparametric Statistical Methods,  
Fourth Edition**

P. Sprent and N.C. Smeeton

**Data Driven Statistical Methods**

P. Sprent

**Generalized Linear Mixed Models:**

**Modern Concepts, Methods and Applications**

W. W. Stroup

**Survival Analysis Using S: Analysis of**

**Time-to-Event Data**

M. Tableman and J.S. Kim

**Applied Categorical and Count Data Analysis**

W. Tang, H. He, and X.M. Tu

**Elementary Applications of Probability Theory,**

**Second Edition**

H.C. Tuckwell

**Introduction to Statistical Inference and Its**

**Applications with R**

M.W. Trosset

**Understanding Advanced Statistical Methods**

P.H. Westfall and K.S.S. Henning

**Statistical Process Control: Theory and**

**Practice, Third Edition**

G.B. Wetherill and D.W. Brown

**Generalized Additive Models:**

**An Introduction with R**

S. Wood

**Epidemiology: Study Design and**

**Data Analysis, Third Edition**

M. Woodward

**Experiments**

B.S. Yandell

**Texts in Statistical Science**

# **Introduction to Multivariate Analysis**

**Linear and Nonlinear Modeling**

**Sadanori Konishi**

Chuo University

Tokyo, Japan



**CRC Press**

Taylor & Francis Group

Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group an **informa** business

A CHAPMAN & HALL BOOK

TAHENRYO KEISEKI NYUMON: SENKEI KARA HISENKEI E by Sadanori Konishi © 2010 by Sadanori Konishi

Originally published in Japanese by Iwanami Shoten, Publishers, Tokyo, 2010. This English language edition published in 2014 by Chapman & Hall/CRC, Boca Raton, FL, U.S.A., by arrangement with the author c/o Iwanami Shoten, Publishers, Tokyo.

CRC Press  
Taylor & Francis Group  
6000 Broken Sound Parkway NW, Suite 300  
Boca Raton, FL 33487-2742

© 2014 by Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works  
Version Date: 20140508

International Standard Book Number-13: 978-1-4665-6729-0 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

**Visit the Taylor & Francis Web site at**  
**<http://www.taylorandfrancis.com>**

**and the CRC Press Web site at**  
**<http://www.crcpress.com>**

---

# Contents

---

<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xxi</b>
<b>Preface</b>	<b>xxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Regression Modeling	1
1.1.1 Regression Models	2
1.1.2 Risk Models	4
1.1.3 Model Evaluation and Selection	5
1.2 Classification and Discrimination	7
1.2.1 Discriminant Analysis	7
1.2.2 Bayesian Classification	8
1.2.3 Support Vector Machines	9
1.3 Dimension Reduction	11
1.4 Clustering	11
1.4.1 Hierarchical Clustering Methods	12
1.4.2 Nonhierarchical Clustering Methods	12
<b>2 Linear Regression Models</b>	<b>15</b>
2.1 Relationship between Two Variables	15
2.1.1 Data and Modeling	16
2.1.2 Model Estimation by Least Squares	18
2.1.3 Model Estimation by Maximum Likelihood	19
2.2 Relationships Involving Multiple Variables	22
2.2.1 Data and Models	23
2.2.2 Model Estimation	24
2.2.3 Notes	29
2.2.4 Model Selection	31
2.2.5 Geometric Interpretation	34
2.3 Regularization	36

2.3.1	Ridge Regression	37
2.3.2	Lasso	40
2.3.3	$L_1$ Norm Regularization	44
<b>3</b>	<b>Nonlinear Regression Models</b>	<b>55</b>
3.1	Modeling Phenomena	55
3.1.1	Real Data Examples	57
3.2	Modeling by Basis Functions	58
3.2.1	Splines	59
3.2.2	$B$ -splines	63
3.2.3	Radial Basis Functions	65
3.3	Basis Expansions	67
3.3.1	Basis Function Expansions	68
3.3.2	Model Estimation	68
3.3.3	Model Evaluation and Selection	72
3.4	Regularization	76
3.4.1	Regularized Least Squares	77
3.4.2	Regularized Maximum Likelihood Method	79
3.4.3	Model Evaluation and Selection	81
<b>4</b>	<b>Logistic Regression Models</b>	<b>87</b>
4.1	Risk Prediction Models	87
4.1.1	Modeling for Proportional Data	87
4.1.2	Binary Response Data	91
4.2	Multiple Risk Factor Models	94
4.2.1	Model Estimation	95
4.2.2	Model Evaluation and Selection	98
4.3	Nonlinear Logistic Regression Models	98
4.3.1	Model Estimation	100
4.3.2	Model Evaluation and Selection	101
<b>5</b>	<b>Model Evaluation and Selection</b>	<b>105</b>
5.1	Criteria Based on Prediction Errors	105
5.1.1	Prediction Errors	106
5.1.2	Cross-Validation	108
5.1.3	Mallows' $C_p$	110
5.2	Information Criteria	112
5.2.1	Kullback-Leibler Information	113
5.2.2	Information Criterion AIC	115
5.2.3	Derivation of Information Criteria	121
5.2.4	Multimodel Inference	127

5.3	Bayesian Model Evaluation Criterion	128
5.3.1	Posterior Probability and BIC	128
5.3.2	Derivation of the BIC	130
5.3.3	Bayesian Inference and Model Averaging	132
<b>6</b>	<b>Discriminant Analysis</b>	<b>137</b>
6.1	Fisher's Linear Discriminant Analysis	137
6.1.1	Basic Concept	137
6.1.2	Linear Discriminant Function	141
6.1.3	Summary of Fisher's Linear Discriminant Analysis	144
6.1.4	Prior Probability and Loss	146
6.2	Classification Based on Mahalanobis Distance	148
6.2.1	Two-Class Classification	148
6.2.2	Multiclass Classification	149
6.2.3	Example: Diagnosis of Diabetes	151
6.3	Variable Selection	154
6.3.1	Prediction Errors	154
6.3.2	Bootstrap Estimates of Prediction Errors	156
6.3.3	The .632 Estimator	158
6.3.4	Example: Calcium Oxalate Crystals	160
6.3.5	Stepwise Procedures	162
6.4	Canonical Discriminant Analysis	164
6.4.1	Dimension Reduction by Canonical Discriminant Analysis	164
<b>7</b>	<b>Bayesian Classification</b>	<b>173</b>
7.1	Bayes' Theorem	173
7.2	Classification with Gaussian Distributions	175
7.2.1	Probability Distributions and Likelihood	175
7.2.2	Discriminant Functions	176
7.3	Logistic Regression for Classification	179
7.3.1	Linear Logistic Regression Classifier	179
7.3.2	Nonlinear Logistic Regression Classifier	183
7.3.3	Multiclass Nonlinear Logistic Regression Classifier	187
<b>8</b>	<b>Support Vector Machines</b>	<b>193</b>
8.1	Separating Hyperplane	193
8.1.1	Linear Separability	193
8.1.2	Margin Maximization	196

8.1.3	Quadratic Programming and Dual Problem	198
8.2	Linearly Nonseparable Case	203
8.2.1	Soft Margins	204
8.2.2	From Primal Problem to Dual Problem	208
8.3	From Linear to Nonlinear	212
8.3.1	Mapping to Higher-Dimensional Feature Space	213
8.3.2	Kernel Methods	216
8.3.3	Nonlinear Classification	218
<b>9</b>	<b>Principal Component Analysis</b>	<b>225</b>
9.1	Principal Components	225
9.1.1	Basic Concept	225
9.1.2	Process of Deriving Principal Components and Properties	230
9.1.3	Dimension Reduction and Information Loss	234
9.1.4	Examples	235
9.2	Image Compression and Decompression	239
9.3	Singular Value Decomposition	243
9.4	Kernel Principal Component Analysis	246
9.4.1	Data Centering and Eigenvalue Problem	246
9.4.2	Mapping to a Higher-Dimensional Space	249
9.4.3	Kernel Methods	252
<b>10</b>	<b>Clustering</b>	<b>259</b>
10.1	Hierarchical Clustering	259
10.1.1	Interobject Similarity	260
10.1.2	Intercluster Distance	261
10.1.3	Cluster Formation Process	263
10.1.4	Ward's Method	267
10.2	Nonhierarchical Clustering	270
10.2.1	K-Means Clustering	271
10.2.2	Self-Organizing Map Clustering	273
10.3	Mixture Models for Clustering	275
10.3.1	Mixture Models	275
10.3.2	Model Estimation by EM Algorithm	277
<b>A</b>	<b>Bootstrap Methods</b>	<b>283</b>
A.1	Bootstrap Error Estimation	283
A.2	Regression Models	285
A.3	Bootstrap Model Selection Probability	285

<b>B Lagrange Multipliers</b>	<b>287</b>
B.1 Equality-Constrained Optimization Problem	287
B.2 Inequality-Constrained Optimization Problem	288
B.3 Equality/Inequality-Constrained Optimization	289
<b>C EM Algorithm</b>	<b>293</b>
C.1 General EM Algorithm	293
C.2 EM Algorithm for Mixture Model	294
<b>Bibliography</b>	<b>299</b>
<b>Index</b>	<b>309</b>

This page intentionally left blank

---

# List of Figures

---

1.1	The relation between falling time ( $x$ sec) and falling distance ( $y$ m) of a body.	3
1.2	The measured impact $y$ (in acceleration, g) on the head of a dummy in repeated experimental crashes of a motorcycle with a time lapse of $x$ (msec).	4
1.3	Binary data $\{0, 1\}$ expressing the presence or absence of response in an individual on exposure to various levels of stimulus.	5
1.4	Regression modeling; the specification of models that approximates the structure of a phenomenon, the estimation of their parameters, and the evaluation and selection of estimated models.	6
1.5	The training data of the two classes are completely separable by a hyperplane (left) and the overlapping data of the two classes may not be separable by a hyperplane (right).	10
1.6	Mapping the observed data to a high-dimensional feature space and obtaining a hyperplane that separates the two classes.	10
1.7	72 chemical substances with 6 attached features, classified by clustering on the basis of mutual similarity in substance qualities.	13
2.1	Data obtained by measuring the length of a spring ( $y$ cm) under different weights ( $x$ g).	17
2.2	The relationship between the spring length ( $y$ ) and the weight ( $x$ ).	18
2.3	Linear regression and the predicted values and residuals.	20

- 2.4 (a) Histogram of 80 measured values obtained while repeatedly suspending a load of 25 g and its approximated probability model. (b) The errors (i.e., noise) contained in these measurements in the form of a histogram having its origin at the mean value of the measurements and its approximated error distribution. 21
- 2.5 Geometrical interpretation of the linear regression model  $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ .  $M(X)$  denotes the  $(p + 1)$ -dimensional linear subspace spanned by the  $(p + 1)$   $n$ -dimensional column vectors of the design matrix  $X$ . 35
- 2.6 Ridge estimate (left panel) and lasso estimate (right panel): Ridge estimation shrinks the regression coefficients  $\beta_1, \beta_2$  toward but not exactly to 0 relative to the corresponding least squares estimates  $\hat{\boldsymbol{\beta}}$ , whereas lasso estimates the regression coefficient  $\beta_1$  at exactly 0. 41
- 2.7 The profiles of estimated regression coefficients for different values of the  $L_1$  norm  $= \sum_{i=1}^{13} |\beta_i(\lambda)|$  with  $\lambda$  varying from 6.78 to 0. The axis above indicates the number of nonzero coefficients. 45
- 2.8 The function  $p_\lambda(|\beta_j|)$  (solid line) and its quadratic approximation (dotted line) with the values of  $\beta_j$  along the  $x$  axis, together with the quadratic approximation for a  $\beta_{j0}$  value of 0.15. 48
- 2.9 The relationship between the least squares estimator (dotted line) and three shrinkage estimators (solid lines): (a) hard thresholding, (b) lasso, and (c) SCAD. 50
- 3.1 Left panel: The plot of 104 tree data obtained by measurement of tree trunk girth (inch) and tree weight above ground (kg). Right panel: Fitting a polynomial of degree 2 (solid curve) and a growth curve model (dashed curve). 57
- 3.2 Motorcycle crash trial data ( $n = 133$ ). 59
- 3.3 Fitting third-degree polynomials to the data in the subintervals  $[a, t_1]$ ,  $[t_1, t_2]$ ,  $\dots$ ,  $[t_m, b]$  and smoothly connecting adjacent polynomials at each knot. 60
- 3.4 Functions  $(x - t_i)_+ = \max\{0, x - t_i\}$  and  $(x - t_i)_+^3$  included in the cubic spline given by (3.10). 61
- 3.5 Basis functions: (a)  $\{1, x\}$ ; linear regression, (b) polynomial regression;  $\{1, x, x^2, x^3\}$ , (c) cubic splines, (d) natural cubic splines. 62

3.6	A cubic <i>B</i> -spline basis function connected four different third-order polynomials smoothly at the knots 2, 3, and 4.	63
3.7	Plots of the first-, second-, and third-order <i>B</i> -spline functions. As may be seen in the subintervals bounded by dotted lines, each subinterval is covered (piecewise) by the polynomial order plus one basis function.	65
3.8	A third-order <i>B</i> -spline regression model is fitted to a set of data, generated from $u(x) = \exp\{-x \sin(2\pi x)\} + 0.5 + \varepsilon$ with Gaussian noise. The fitted curve and the true structure are, respectively, represented by the solid line and the dotted line with cubic <i>B</i> -spline bases.	66
3.9	Curve fitting; a nonlinear regression model based on a natural cubic spline basis function and a Gaussian basis function.	70
3.10	Cubic <i>B</i> -spline nonlinear regression models, each with a different number of basis functions (a) 10, (b) 20, (c) 30, (d) 40, fitted to the motorcycle crash experiment data.	73
3.11	The cubic <i>B</i> -spline nonlinear regression model $y = \sum_{j=1}^{13} \hat{w}_j b_j(x)$ . The model is estimated by maximum likelihood and selected the number of basis functions by AIC.	75
3.12	The role of the penalty term: Changing the weight in the second term by the regularization parameter $\gamma$ changes $S_\gamma(\mathbf{w})$ continuously, thus enabling continuous adjustment of the model complexity.	78
3.13	The effect of a smoothing parameter $\lambda$ : The curves are estimated by the regularized maximum likelihood method for various values of $\lambda$ .	82
4.1	Plot of the graduated stimulus levels shown in Table 4.1 along the $x$ axis and the response rate along the $y$ axis.	89
4.2	Logistic functions.	90
4.3	Fitting the logistic regression model to the observed data shown in Table 4.1 for the relation between the stimulus level $x$ and the response rate $y$ .	90
4.4	The data on presence and non-presence of the crystals are plotted along the vertical axis as $y = 0$ for the 44 individuals exhibiting their non-presence and $y = 1$ for the 33 exhibiting their presence. The $x$ axis takes the values of their urine specific gravity.	92

4.5	The fitted logistic regression model for the 77 set of data expressing observed urine specific gravity and presence or non-presence of calcium oxalate crystals.	93
4.6	Plot of post-operative kyphosis occurrence along $Y = 1$ and non-occurrence along $Y = 0$ versus the age ( $x$ ; in months) of 83 patients.	99
4.7	Fitting the polynomial-based nonlinear logisitic regression model to the kyphosis data.	103
5.1	Fitting of 3rd-, 8th-, and 12th-order polynomial models to 15 data points.	107
5.2	Fitting a linear model (dashed line), a 2nd-order polynomial model (solid line), and an 8th-order polynomial model (dotted line) to 20 data.	119
6.1	Projecting the two-dimensional data in Table 6.1 onto the axes $y = x_1$ , $y = x_2$ and $y = w_1x_1 + w_2x_2$ .	139
6.2	Three projection axes (a), (b), and (c) and the distributions of the class $G_1$ and class $G_2$ data when projected on each one.	140
6.3	Fisher's linear discriminant function.	143
6.4	Mahalanobis distance and Euclidean distance.	151
6.5	Plot of 145 training data for a normal class $G_1$ ( $\circ$ ), a chemical diabetes class $G_2$ ( $\blacktriangle$ ), and clinical diabetes class $G_3$ ( $\times$ ).	152
6.6	Linear decision boundaries that separate the normal class $G_1$ , the chemical diabetes class $G_2$ , and the clinical diabetes class $G_3$ .	154
6.7	Plot of the values obtained by projecting the 145 observed data from three classes onto the first two discriminant variables ( $y_1, y_2$ ) in (6.92).	170
7.1	Likelihood of the data: The relative level of occurrence of males 178 cm in height can be determined as $f(178 170, 6^2)$ .	176
7.2	The conditional probability $P(x G_i)$ that gives the relative level of occurrence of data $x$ in each class.	178
7.3	Decision boundary generated by the linear function.	184
7.4	Classification of phenomena exhibiting complex class structures requires a nonlinear discriminant function.	185

7.5	Decision boundary that separates the two classes in the nonlinear logistic regression model based on the Gaussian basis functions.	187
8.1	The training data are completely separable into two classes by a hyperplane (left panel), and in contrast, separation into two classes cannot be obtained by any such linear hyperplane (right panel).	194
8.2	Distance from $\mathbf{x}_0 = (x_{01} \ x_{02})^T$ to the hyperplane $w_1 x_1 + w_2 x_2 + b = \mathbf{w}^T \mathbf{x} + b = 0$ .	196
8.3	Hyperplane ( $H$ ) that separates the two classes, together with two equidistant parallel hyperplanes ( $H_+$ and $H_-$ ) on opposite sides.	197
8.4	Separating hyperplanes with different margins.	198
8.5	Optimum separating hyperplane and support vectors represented by the black solid dots and triangle on the hyperplanes $H_+$ and $H_-$ .	202
8.6	No matter where we draw the hyperplane for separation of the two classes and the accompanying hyperplanes for the margin, some of the data (the black solid dots and triangles) do not satisfy the inequality constraint.	205
8.7	The class $G_1$ data at $(0, 0)$ and $(0, 1)$ do not satisfy the original constraint $x_1 + x_2 - 1 \geq 1$ . We soften this constraint to $x_1 + x_2 - 1 \geq 1 - 2$ for data $(0, 0)$ and $x_1 + x_2 - 1 \geq 1 - 1$ for $(0, 1)$ by subtracting 2 and 1, respectively; each of these data can then satisfy its new inequality constraint equation.	205
8.8	The class $G_2$ data $(1, 1)$ and $(0, 1)$ are unable to satisfy the constraint, but if the restraint is softened to $-(x_1 + x_2 - 1) \geq 1 - 2$ and $-(x_1 + x_2 - 1) \geq 1 - 1$ by subtracting 2 and 1, respectively, each of these data can then satisfy its new inequality constraint equation.	206
8.9	A large margin tends to increase the number of data that intrude into the other class region or into the region between hyperplanes $H_+$ and $H_-$ .	207
8.10	A small margin tends to decrease the number of data that intrude into the other class region or into the region between hyperplanes $H_+$ and $H_-$ .	207
8.11	Support vectors in a linearly nonseparable case: Data corresponding to the Lagrange multipliers such that $0 < \hat{\alpha}_i \leq \lambda$ (the black solid dots and triangles).	211

8.12	Mapping the data of an input space into a higher-dimensional feature space with a nonlinear function.	214
8.13	The separating hyperplane obtained by mapping the two-dimensional data of the input space to the higher-dimensional feature space yields a nonlinear discriminant function in the input space. The black solid data indicate support vectors.	216
8.14	Nonlinear decision boundaries in the input space vary with different values $\sigma$ in the Gaussian kernel; (a) $\sigma = 10$ , (b) $\sigma = 1$ , (c) $\sigma = 0.1$ , and (d) $\sigma = 0.01$ .	221
9.1	Projection onto three different axes, (a), (b), and (c) and the spread of the data.	226
9.2	Eigenvalue problem and the first and second principal components.	230
9.3	Principal components based on the sample correlation matrix and their contributions: The contribution of the first principal component increases with increasing correlation between the two variables.	237
9.4	Two-dimensional view of the 21-dimensional data set, projected onto the first ( $x$ ) and second ( $y$ ) principal components.	239
9.5	Image digitization of a handwritten character.	240
9.6	The images obtained by first digitizing and compressing the leftmost image 7 and then decompressing transmitted data using a successively increasing number of principal components. The number in parentheses shows the cumulative contribution rate in each case.	242
9.7	Mapping the observed data with nonlinear structure to a higher-dimensional feature space, where PCA is performed with linear combinations of variables $z_1, z_2, z_3$ .	250
10.1	Intercluster distances: Single linkage (minimum distance), complete linkage (maximum distance), average linkage, centroid linkage.	262
10.2	Cluster formation process and the corresponding dendrogram based on single linkage when starting from the distance matrix in (10.7).	265

10.3	The dendrograms obtained for a single set of 72 six-dimensional data using three different linkage techniques: single, complete, and centroid linkages. The circled portion of the dendrogram shows a chaining effect.	266
10.4	Fusion-distance monotonicity (left) and fusion-distance inversion (right).	267
10.5	Stepwise cluster formation procedure by Ward's method and the related dendrogram.	271
10.6	Stepwise cluster formation process by $k$ -means.	272
10.7	The competitive layer comprises an array of $m$ nodes. Each node is assigned a different weight vector $\mathbf{w}_j = (w_{j1}, w_{j2}, \dots, w_{jp})^T$ ( $j = 1, 2, \dots, m$ ), and the Euclidean distance of each $p$ -dimensional data to the weight vector is computed.	274
10.8	Histogram based on observed data on the speed of recession from Earth of 82 galaxies scattered in space.	276
10.9	Recession-speed data observed for 82 galaxies are shown on the upper left and in a histogram on the upper right. The lower left and lower right show the models obtained by fitting with two and three normal distributions, respectively.	279

This page intentionally left blank

---

# List of Tables

---

2.1	The length of a spring under different weights.	16
2.2	The $n$ observed data.	17
2.3	Four factors: temperature ( $x_1$ ), pressure ( $x_2$ ), PH ( $x_3$ ), and catalyst quantity ( $x_4$ ), which affect the quantity of product ( $y$ ).	23
2.4	The response $y$ representing the results in $n$ trials, each with a different combination of $p$ predictor variables $x_1, x_2, \dots, x_p$ .	23
2.5	Comparison of the sum of squared residuals ( $\hat{\sigma}^2$ ) divided by the number of observations, maximum log-likelihood $\ell(\hat{\beta})$ , and AIC for each combination of predictor variables.	33
2.6	Comparison of the estimates of regression coefficients by least squares (LS) and lasso $L_1$ .	44
4.1	Stimulus levels and the proportion of individuals responded.	88
5.1	Comparison of the values of RSS, CV, and AIC for fitting the polynomial models of order 1 through 9.	119
6.1	The 23 two-dimensional observed data from the varieties $A$ and $B$ .	138
6.2	Comparison of prediction error estimates for the classification rule constructed by the linear discriminant function.	161
6.3	Variable selection via the apparent error rates (APE).	161

This page intentionally left blank

---

# Preface

---

The aim of statistical science is to develop the methodology and the theory for extracting useful information from data and for reasonable inference to elucidate phenomena with uncertainty in various fields of the natural and social sciences. The data contain information about the random phenomenon under consideration and the objective of statistical analysis is to express this information in an understandable form using statistical procedures. We also make inferences about the unknown aspects of random phenomena and seek an understanding of causal relationships.

Multivariate analysis refers to techniques used to analyze data that arise from multiple variables between which there are some relationships. Multivariate analysis has been widely used for extracting useful information and patterns from multivariate data and for understanding the structure of random phenomena. Techniques would include regression, discriminant analysis, principal component analysis, clustering, etc., and are mainly based on the linearity of observed variables.

In recent years, the wide availability of fast and inexpensive computers enables us to accumulate a huge amount of data with complex structure and/or high-dimensional data. Such data accumulation is also accelerated by the development and proliferation of electronic measurement and instrumentation technologies. Such data sets arise in various fields of science and industry, including bioinformatics, medicine, pharmaceuticals, systems engineering, pattern recognition, earth and environmental sciences, economics, and marketing. Therefore, the effective use of these data sets requires both linear and nonlinear modeling strategies based on the complex structure and/or high-dimensionality of the data in order to perform extraction of useful information, knowledge discovery, prediction, and control of nonlinear phenomena and complex systems.

The aim of this book is to present the basic concepts of various procedures in traditional multivariate analysis and also nonlinear techniques for elucidation of phenomena behind observed multivariate data, focusing primarily on regression modeling, classification and discrimination, dimension reduction, and clustering. Each chapter includes many figures

and illustrative examples to promote a deeper understanding of various techniques in multivariate analysis.

In practice, the need always arises to search through and evaluate a large number of models and from among them select an appropriate model that will work effectively for elucidation of the target phenomena. This book provides comprehensive explanations of the concepts and derivations of the AIC, BIC, and related criteria, together with a wide range of practical examples of model selection and evaluation criteria. In estimating and evaluating models having a large number of predictor variables, the usual methods of separating model estimation and evaluation are inefficient for the selection of factors affecting the outcome of the phenomena. The book also reflects these aspects, providing various regularization methods, including the  $L_1$  norm regularization that gives simultaneous model estimation and variable selection.

The book is written in the hope that, through its fusion of knowledge gained in leading-edge research in statistical multivariate analysis, machine learning, and computer science, it may contribute to the understanding and resolution of problems and challenges in this field of research, and to its further advancement.

This book might be useful as a text for advanced undergraduate and graduate students in statistical sciences, providing a systematic description of both traditional and newer techniques in multivariate analysis and machine learning. In addition, it introduces linear and nonlinear statistical modeling for researchers and practitioners in various scientific disciplines such as industrial and systems engineering, information science, and life science. The basic prerequisites for reading this textbook are knowledge of multivariate calculus and linear algebra, though they are not essential as it includes a self-contained introduction to theoretical results.

This book is basically a translation of a book published in Japanese by Iwanami Publishing Company in 2010. I would like to thank Uichi Yoshida and Nozomi Tsujimura of the Iwanami Publishing Company for giving me the opportunity to translate and publish in English.

I would like to acknowledge with my sincere thanks Yasunori Fujikoshi, Genshiro Kitagawa, and Nariaki Sugiura, from whom I have learned so much about the seminal ideas of statistical modeling. I have been greatly influenced through discussions with Tomohiro Ando, Yuko Araki, Toru Fujii, Seiya Imoto, Mitsunori Kayano, Yoshihiko Maesono, Hiroki Masuda, Nagatomo Nakamura, Yoshiyuki Ninomiya, Ryuei Nishii, Heewon Park, Fumitake Sakaori, Shohei Tateishi, Takahiro Tsuchiya, Masayuki Uchida, Takashi Yanagawa, and Nakahiro Yoshida.

I would also like to express my sincere thanks to Kei Hirose, Shuichi Kawano, Hidetoshi Matsui, and Toshihiro Misumi for reading the manuscript and offering helpful suggestions. David Grubbs patiently encouraged and supported me throughout the final preparation of this book. I express my sincere gratitude to all of these people.

Sadanori Konishi

Tokyo, January 2014

This page intentionally left blank

# Introduction

---

The highly advanced computer systems and progress in electronic measurements and instrumentation technologies have together facilitated the acquisition and accumulation of data with complex structure and/or high-dimensional data in various fields of science and industry. Data sets arise in such areas as genome databases in life science, remote-sensing data from earth-observing satellites, real-time recorded data of motion process in system engineering, high-dimensional data in character recognition, speech recognition, image analysis, etc. Hence, it is desirable to research and develop new statistical data analysis techniques to efficiently extract useful information as well as elucidate patterns behind the data in order to analyze various phenomena and to yield knowledge discovery. Under the circumstances linear and nonlinear multivariate techniques are rapidly developing by fusing the knowledge in statistical science, machine learning, information science, and mathematical science.

The objective of this book is to present the basic concepts of various procedures in the traditional multivariate analysis and also nonlinear techniques for elucidation of phenomena behind the observed multivariate data, using many illustrative examples and figures. In each chapter, starting from an understanding of the traditional multivariate analysis based on the linearity of multivariate observed data, we describe nonlinear techniques, focusing primarily on regression modeling, classification and discrimination, dimension reduction, and clustering.

## 1.1 Regression Modeling

Regression analysis is used to model the relationship between a response variable and several predictor (explanatory) variables. Once a model has been identified, various forms of inferences such as prediction, control, information extraction, knowledge discovery, and risk evaluation can be done within the framework of deductive argument. Thus, the key to solving various real-world problems lies in the development and construction of suitable linear and nonlinear regression modeling.

### 1.1.1 Regression Models

Housing prices vary with land area and floor space, but also with proximity to stations, schools, and supermarkets. The quantity of chemical products is sensitive to temperature, pressure, catalysts, and other factors. In Chapter 2, using *linear regression* models, which provide a method for relating multiple factors to the outcomes of such phenomena, we describe the basic concept of *regression modeling*, including model specification based on data reflecting the phenomena, model estimation of the specified model by least squares or maximum likelihood methods, and model evaluation of the estimated model. Throughout this modeling process, we select a suitable one among competing models.

The volume of extremely high-dimensional data that are observed and entered into databases in biological, genomic, and many other fields of science has grown rapidly in recent years. For such data, the usual methods of separating model estimation and evaluation are ineffectual for the selection of factors affecting the outcome of the phenomena, and thus effective techniques are required to construct models with high reliability and prediction. This created a need for work on modeling and has led, in particular, to the proposal of various regularization methods with an  $L_1$  penalty term (the sum of absolute values of regression coefficients), in addition to the sum of squared errors and log-likelihood functions. A distinctive feature of the proposed methods is their capability for simultaneous model estimation and variable selection. Chapter 2 also describes various regularization methods, including *ridge* regression (Hoerl and Kennard, 1970) and the least absolute shrinkage and selection operator (*lasso*) proposed by Tibshirani (1996), within the framework of linear regression models.

Figure 1.1 shows the results of an experiment performed to investigate the relation between falling time ( $x$  sec) and falling distance ( $y$  m) of a body. The figure suggests that it should be possible to model the relation using a polynomial. There are many phenomena that can be modeled in this way, using polynomial equations, exponential functions, or other specific nonlinear functions to relate the outcome of the phenomenon and the factors influencing that outcome.

Figure 1.2, however, poses new difficulties. It shows the measured impact  $y$  (in acceleration, g) on the head of a dummy in repeated experimental crashes of a motorcycle into a wall, with a time lapse of  $x$  (msec) as measured from the instant of collision (Härdle, 1990). For phenomena with this type of apparently complex nonlinear structure, it is quite difficult to effectively capture the structure by modeling with specific

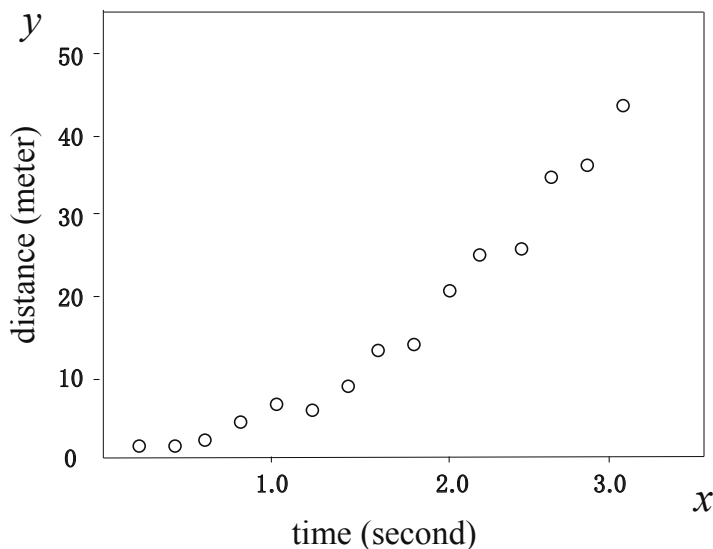


Figure 1.1 *The relation between falling time ( $x$  sec) and falling distance ( $y$  m) of a body.*

nonlinear functions such as polynomial equations and exponential functions.

Chapter 3 discusses *nonlinear regression* modeling for extracting useful information from data containing complex nonlinear structures. It introduces models based on more flexible splines, *B-splines*, and radial basis functions for modeling complex nonlinear structures. These models often serve to ascertain complex nonlinear structures, but their flexibility often prevents their effective function in the estimation of models with the traditional least squares and maximum likelihood methods. In such cases, these estimation methods are replaced by regularized least squares and regularized maximum likelihood methods.

The latter two techniques, which are generally referred to as *regularization* methods, are effectively used to reduce over-fitting of models to data and thus prevent excessive model complexity, and are known to contribute for reducing the variability of the estimated models. This chapter also describes regularization methods within the framework of nonlinear regression modeling.

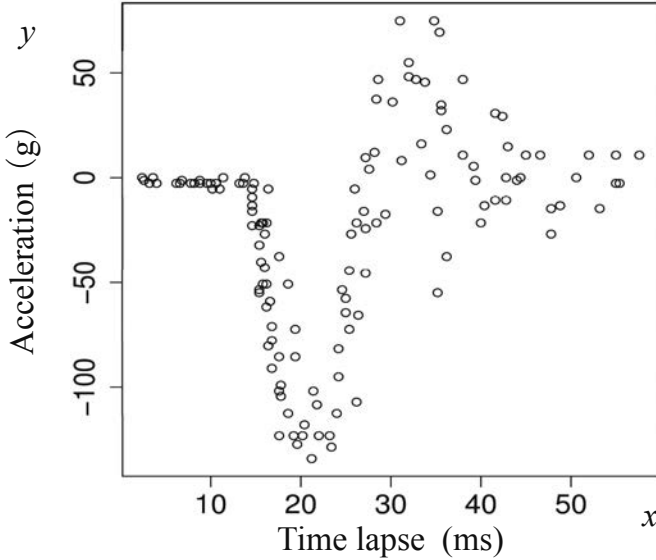


Figure 1.2 *The measured impact  $y$  (in acceleration, g) on the head of a dummy in repeated experimental crashes of a motorcycle with a time lapse of  $x$  (msec).*

### 1.1.2 Risk Models

In today's society marked by complexity and uncertainty, we live in a world exposed by various types of risks. The risk may be associated with occurrences such as traffic accidents, natural disasters such as earthquakes, tsunamis, or typhoons, or development of a lifestyle disease, with transactions such as credit card issuance, or with many other occurrences too numerous to enumerate. It is possible to gauge the magnitude of risk in terms of probability based on past experience and information gained in life in society, but often with only a limited accuracy.

All of this poses the question of how to probabilistically assess unknown risks for a phenomenon using information obtained from data. For example, in searching for the factors that induce a certain disease, the problem is in how to construct a model for assessing the probability of its occurrence based on observed data. The effective probabilistic model for assessing the risk may lead to its future prevention. Through such risk modeling, moreover, it may also be possible to identify important disease-related factors.

Chapter 4 presents an answer to this question, in the form of model-

ing for the risk evaluation, and in particular describes the basic concept of *logistic regression modeling*, together with its extension from linear to nonlinear modeling. This includes models to assess risks based on binary data  $\{0, 1\}$  expressing the presence or absence of response in an individual or object on exposure to various levels of stimulus, as shown in Figure 1.3.

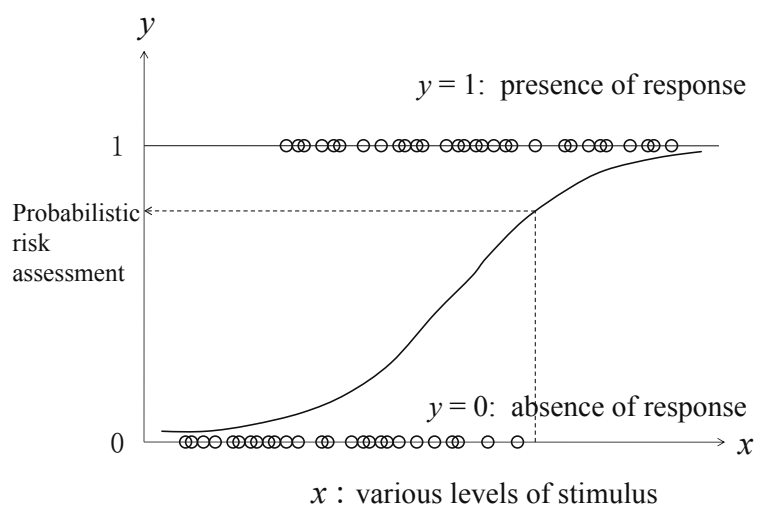


Figure 1.3 *Binary data  $\{0, 1\}$  expressing the presence or absence of response in an individual on exposure to various levels of stimulus.*

1.1.3 *Model Evaluation and Selection*

Figure 1.4 shows a process consisting essentially of the conceptualization of *regression modeling*; the specification of models that approximates the structure of a phenomenon, the estimation of their parameters, and the evaluation and selection of estimated models.

In relation to the data shown in Figure 1.1 for a body dropped from a high position, for example, it is quite natural to consider a polynomial model for the relation between the falling time and falling distance and to carry out polynomial model fitting. This represents the processes of model specification and parameter estimation. For elucidation of this

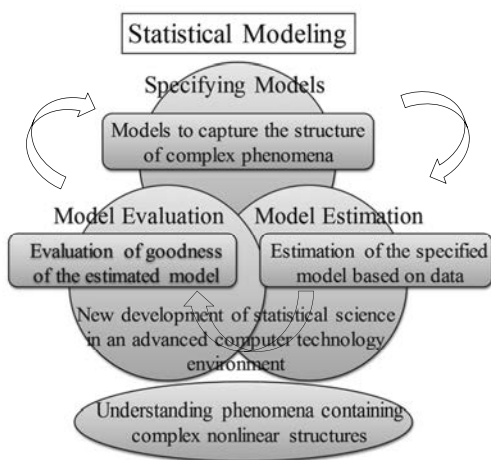


Figure 1.4 *Regression modeling; the specification of models that approximates the structure of a phenomenon, the estimation of their parameters, and the evaluation and selection of estimated models.*

physical phenomenon, however, a question may remain as to the optimum degree of the polynomial model. In the prediction of housing prices with linear regression models, moreover, a key question is what factors to include in the model. Furthermore, in considering nonlinear regression models, one is confronted by the availability of infinite candidate models for complex nonlinear phenomena controlled by smoothing parameters, and the need for selection of models that will appropriately approximate the structures of the phenomena, which is essential for their elucidation.

In this way, the need always arises to search through and evaluate a large number of models and from among them select one that will work effectively for elucidation of the target phenomena, based on the information provided by the data. This is commonly referred to as the *model evaluation and selection problem*.

Chapter 5 focuses on the model evaluation and selection problems, and presents various model selection criteria that are widely used as indicators in the assessment of the *goodness* of a model. It begins with a description of evaluation criteria proposed as estimators of prediction error, and then discusses the AIC (Akaike information criterion) based

on Kullback-Leibler information and the BIC (Bayesian information criterion) derived from a Bayesian view point, together with fundamental concepts that serve as the bases for derivation of these criteria.

The AIC, proposed in 1973 by Hirotugu Akaike, is widely used in various fields of natural and social sciences and has contributed greatly to elucidation, prediction, and control of phenomena. The BIC was proposed in 1978 by Gideon E. Schwarz and is derived based on a Bayesian approach rather than on information theory as with the AIC, but like the AIC it is utilized throughout the world of science and has played a central role in the advancement of modeling. Chapters 2 to 4 of this book show the various forms of expression of the AIC for linear, nonlinear, logistic, and other models, and give examples for model evaluation and selection problems based on the AIC.

Model selection from among candidate models constructed on the basis of data is essentially the selection of a single model that best approximates the data-generated probability structure. In Chapter 5, the discussion is further extended to include the concept of *multimodel inference* (Burnham and Anderson, 2002) in which the inferences are based on model aggregation and utilization of the relative importance of constructed models in terms of their weighted values.

## 1.2 Classification and Discrimination

Classification and discrimination techniques are some of the most widely used statistical tools in various fields of natural and social sciences. The primary aim in discriminant analysis is to assign an individual to one of two or more classes (groups) on the basis of measurements on feature variables. It is designed to construct linear and nonlinear decision boundaries based on a set of training data.

### 1.2.1 Discriminant Analysis

When a preliminary diagnosis concerning the presence or absence of a disease is made on the basis of data from blood chemistry analysis, information contained in the blood relating to the disease is measured, assessed, and acquired in the form of qualitative data. The diagnosis of normality or abnormality is based on multivariate data from several test results. In other words, it is an assessment of whether the person examined is included in a group consisting of normal individuals or a group consisting of individuals who exhibit a disease-related abnormality.

This kind of assessment can be made only if information from test re-