

Monographs on Statistics and Applied Probability 161

Sufficient Dimension Reduction

Methods and Applications with R

Bing Li



A CHAPMAN & HALL BOOK

Sufficient Dimension Reduction

Methods and Applications with R

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

Editors: F. Bunea, P. Fryzlewicz, R. Henderson, N. Keiding, T. Louis, R. Smith,
and W. Wong

Semialgebraic Statistics and Latent Tree Models

Piotr Zwiernik 146

Inferential Models

Reasoning with Uncertainty

Ryan Martin and Chuanhai Liu 147

Perfect Simulation

Mark L. Huber 148

State-Space Methods for Time Series Analysis

Theory, Applications and Software

*Jose Casals, Alfredo Garcia-Hiernaux, Miguel Jerez, Sonia Sotoca, and
A. Alexandre Trindade 149*

Hidden Markov Models for Time Series

An Introduction Using R, Second Edition

Walter Zucchini, Iain L. MacDonald, and Roland Langrock 150

Joint Modeling of Longitudinal and Time-to-Event Data

Robert M. Elashoff, Gang Li, and Ning Li 151

Multi-State Survival Models for Interval-Censored Data

Ardo van den Hout 152

Generalized Linear Models with Random Effects

Unified Analysis via H-likelihood, Second Edition

Youngjo Lee, John A. Nelder, and Yudi Pawitan 153

Absolute Risk

Methods and Applications in Clinical Management and Public Health

Ruth M. Pfeiffer and Mitchell H. Gail 154

Asymptotic Analysis of Mixed Effects Models

Theory, Applications, and Open Problems

Jiming Jiang 155

Missing and Modified Data in Nonparametric Estimation With R Examples

Sam Efromovich 156

Probabilistic Foundations of Statistical Network Analysis

Harry Crane 157

Multistate Models for the Analysis of Life History Data

Richard J. Cook and Jerald F. Lawless 158

**Nonparametric Models for Longitudinal Data
with Implementation in R**

Colin O. Wu and Xin Tian 159

Multivariate Kernel Smoothing and Its Applications

José E. Chacón and Tarn Duong 160

**Sufficient Dimension Reduction
Methods and Applications with R**

Bing Li 161

For more information about this series please visit:

<https://www.crcpress.com/Chapman--HallCRC-Monographs-on-Statistics--Applied-Probability/book-series/CHMONSTAAPP>



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Sufficient Dimension Reduction

Methods and Applications with R

Bing Li



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

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

© 2018 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

Printed on acid-free paper
Version Date: 20180404

International Standard Book Number-13: 978-1-4987-0447-2 (Hardback)

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/>) 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>

To Yanling



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contents

List of Figures	xiii
List of Tables	xvii
Preface	xix
Author	xxi
1 Preliminaries	1
1.1 Empirical Distribution and Sample Moments	1
1.2 Principal Component Analysis	2
1.3 Generalized Eigenvalue Problem	3
1.4 Multivariate Linear Regression	3
1.5 Generalized Linear Model	5
1.5.1 Exponential Family	5
1.5.2 Generalized Linear Models	6
1.6 Hilbert Space, Linear Manifold, Linear Subspace	8
1.7 Linear Operator and Projection	10
1.8 The Hilbert Space $\mathbb{R}^p(\Sigma)$	11
1.9 Coordinate Representation	12
1.10 Generalized Linear Models under Link Violation	13
2 Dimension Reduction Subspaces	17
2.1 Conditional Independence	17
2.2 Sufficient Dimension Reduction Subspace	21
2.3 Transformation Laws of Central Subspace	24
2.4 Fisher Consistency, Unbiasedness, and Exhaustiveness	25
3 Sliced Inverse Regression	27
3.1 Sliced Inverse Regression: Population-Level Development	27
3.2 Limitation of SIR	30
3.3 Estimation, Algorithm, and R-codes	31
3.4 Application: The Big Mac Index	33

4	Parametric and Kernel Inverse Regression	37
4.1	Parametric Inverse Regression	37
4.2	Algorithm, R Codes, and Application	39
4.3	Relation of PIR with SIR	40
4.4	Relation of PIR with Ordinary Least Squares	42
4.5	Kernel Inverse Regression	42
5	Sliced Average Variance Estimate	47
5.1	Motivation	47
5.2	Constant Conditional Variance Assumption	47
5.3	Sliced Average Variance Estimate	49
5.4	Algorithm and R-code	52
5.5	Relation with SIR	55
5.6	The Issue of Exhaustiveness	56
5.7	SIR-II	58
5.8	Case Study: The Pen Digit Data	60
6	Contour Regression and Directional Regression	63
6.1	Contour Directions and Central Subspace	63
6.2	Contour Regression at the Population Level	65
6.3	Algorithm and R Codes for CR	67
6.4	Exhaustiveness of Contour Regression	69
6.5	Directional Regression	70
6.6	Representation of Λ_{DR} Using Moments	74
6.7	Algorithm and R Codes for DR	76
6.8	Exhaustiveness Relation with SIR and SAVE	77
6.9	Pen Digit Case Study Continued	79
7	Elliptical Distribution and Predictor Transformation	83
7.1	Linear Conditional Mean and Elliptical Distribution	83
7.2	Box-Cox Transformation	88
7.3	Application to the Big Mac Data	92
7.4	Estimating Equations for Handling Non-Ellipticity	94
8	Sufficient Dimension Reduction for Conditional Mean	97
8.1	Central Mean Subspace	97
8.2	Ordinary Least Squares	100
8.3	Principal Hessian Direction	101
8.4	Iterative Hessian Transformation	104
9	Asymptotic Sequential Test for Order Determination	107
9.1	Stochastic Ordering and Von Mises Expansion	107
9.2	Von Mises Expansion and Influence Functions	109
9.3	Influence Functions of Some Statistical Functionals	110
9.4	Random Matrix with Affine Invariant Eigenvalues	112
9.5	Asymptotic Distribution of the Sum of Small Eigenvalues	115

9.6	General Form of the Sequential Tests	117
9.7	Sequential Test for SIR	118
9.8	Sequential Test for PHD	124
9.9	Sequential Test for SAVE	126
9.10	Sequential Test for DR	132
9.11	Applications	139
10	Other Methods for Order Determination	141
10.1	BIC Type Criteria for Order Determination	141
10.2	Bootstrapped Eigenvector Variation	147
10.3	Eigenvalue Magnitude and Eigenvector Variation	150
10.4	Ladle Estimator	152
10.5	Consistency of the Ladle Estimator	156
10.6	Application: Identification of Wine Cultivars	156
11	Forward Regressions for Dimension Reduction	159
11.1	Outer Product of Gradients	160
11.2	Fisher Consistency of Gradient Estimate	163
11.3	Minimum Average Variance Estimate	167
11.4	Refined MAVE and refined OPG	170
11.5	From Central Mean Subspace to Central Subspace	173
11.6	dOPG and Its Refinement	173
11.7	dMAVE and Its Refinement	178
11.8	Ensemble Estimators	180
11.9	Simulation Studies and Applications	184
11.10	Summary	188
12	Nonlinear Sufficient Dimension Reduction	191
12.1	Reproducing Kernel Hilbert Space	192
12.2	Covariance Operators in RKHS	193
12.3	Coordinate Mapping	199
12.4	Coordinate of Covariance Operators	200
12.5	Kernel Principal Component Analysis	202
12.6	Sufficient and Central σ -Field for Nonlinear SDR	204
12.7	Complete Sub σ -Field for Nonlinear SDR	206
12.8	Converting σ -Fields to Function Classes for Estimation	208
13	Generalized Sliced Inverse Regression	211
13.1	Regression Operator	212
13.2	Generalized Sliced Inverse Regression	213
13.3	Exhaustiveness and Completeness	215
13.4	Relative Universality	216
13.5	Implementation of GSIR	217
13.6	Precursors and Variations of GSIR	220
13.7	Generalized Cross Validation for Tuning ε_X and ε_Y	220
13.8	k -Fold Cross Validation for Tuning $\rho_X, \rho_Y, \varepsilon_X, \varepsilon_Y$	223

13.9	Simulation Studies	225
13.10	Applications	227
13.10.1	Pen Digit Data	227
13.10.2	Face Sculpture Data	228
14	Generalized Sliced Average Variance Estimator	233
14.1	Generalized Sliced Average Variance Estimation	233
14.2	Relation with GSIR	237
14.3	Implementation of GSAVE	239
14.4	Simulation Studies and an Application	248
14.5	Relation between Linear and Nonlinear SDR	251
15	The Broad Scope of Sufficient Dimension Reduction	253
15.1	Sufficient Dimension Reduction for Functional Data	253
15.2	Sufficient Dimension Folding for Tensorial Data	256
15.3	Sufficient Dimension Reduction for Grouped Data	259
15.4	Variable Selection via Sufficient Dimension Reduction	260
15.5	Efficient Dimension Reduction	262
15.6	Partial Dimension Reduction for Categorical Predictors	264
15.7	Measurement Error Problem	265
15.8	SDR via Support Vector Machine	267
15.9	SDR for Multivariate Responses	268
	Bibliography	271
	Index	281

List of Figures

1.1	Estimating gradient direction under link violation in Generalized Linear Model.	16
2.1	Sufficient Dimension Reduction subspace.	23
3.1	Illustration of unbiasedness of SIR.	29
3.2	Illustration of limitation of Sliced Inverse Regression.	31
3.3	Scatter plot matrix of ten economic variables in the Big Mac data.	34
3.4	Scatter plot of the response versus the first two SIR predictors.	35
4.1	Big Mac index versus the first PIR predictor.	40
4.2	Big Mac index versus the OLS predictor.	43
4.3	Big Mac index versus the first KIR predictor.	45
5.1	Leading eigenvector $\text{var}(X Y \in J_\ell)$ in the monotone case.	51
5.2	Leading eigenvector of $\text{var}(X Y_\ell)$ in symmetric case.	52
5.3	Y versus first two SAVE predictor for the Big Mac data.	53
5.4	Y versus first two SAVE predictor for the Big Mac data with first predictor removed.	54
5.5	Y versus X_1 and X_2 .	54
5.6	Upper panels: Y versus first two SIR predictors; lower panels: Y versus first two SAVE predictors.	55
5.7	Y versus first two SIR-II predictors for the Big Mac data.	60
5.8	Y versus first two SIR-II predictors for Model (5.8) in Example 5.1.	60
5.9	First three predictors from SIR (upper panel), SAVE (lower-left panel), and SIR-II (lower-right panel).	62
6.1	Illustration of contour directions and central subspace.	64
6.2	Y versus first two CR predictors for the model in Example 5.1.	68
6.3	Reorganization of empirical directions by Directional Regression.	73
6.4	Y versus first two DR predictors for the model in Example 5.1.	77
6.5	Perspective plots for the pen digit case study.	80
6.6	Effect of scale separation by SIR and DR for the pen digit data.	81
6.7	Five handwriting shapes in the pen digit data.	81

7.1	Scatter plot matrix for Box-Cox transformed of the predictor of the Big Mac data.	93
7.2	First SIR predictor for the Box-Cox transformed Big Mac data.	93
8.1	OLS predictor for Big Mac data.	101
8.2	PHD predictor for Big Mac data. Left panel: Y versus first predictor; right panel: Y versus the second predictor.	103
8.3	Perspective plot for the first three PHD predictors for the pen digit data.	103
8.4	Application of IHT to pen digit data.	106
9.1	Boxplots of p-values for sequential test based on SIR as applied to Model (9.18).	123
9.2	Boxplots of p-values for sequential test based on PHD as applied to Model (9.18).	126
9.3	Boxplots of p-values for sequential test based on SAVE as applied to the model in Example 5.1.	132
9.4	Boxplots of p-values for sequential test based on DR for the model in Example 5.1.	136
10.1	Bootstrapped eigenvector variation based on SIR for $k = 1, \dots, 6$, as applied to the model in Example 9.1.	149
10.2	Ladle plot for the model in Example 9.1.	153
10.3	Ladle plot for the wine data.	157
10.4	Comparing the 2-d and 3-d plots to see the effect of the third sufficient predictor in the wind data.	157
11.1	Comparison of OPG, MAVE, rOPG, rMAVE, rdOPG, rdMAVE, reOPG, reMAVE for the model in Example 11.1.	185
11.2	Comparison of OPG, MAVE, rOPG, rMAVE, rdOPG, rdMAVE, reOPG, reMAVE for the model in Example 11.2.	186
11.3	Scatter plot matrix for the abalone data.	187
11.4	Age versus the first two predictors from rOPG and reOPG for the abalone data.	188
12.1	First three linear principal components and first three kernel principal components for the pen digit data.	204
13.1	GSIR for Model I, Scenario A.	226
13.2	First two GSIR predictors for the pen digit data.	228
13.3	Face data representation.	229
13.4	First three GSIR predictors as evaluated at the testing set.	230
13.5	Relation between the true responses and the first three GSIR predictors in the face data.	231
14.1	First five sufficient predictors by GSAVE for the pen digit data.	250

- 15.1 A schematic representation of the functional data collected by EEG
from a subject.

254



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

List of Tables

3.1	Spearman's correlation with the response	35
9.1	p -values for sequential tests applied to the Big Mac data	139
9.2	p -values for sequential tests applied to the pen digit data	140
10.1	Percentage (with symbol % omitted) of correct estimation by BIC	146
10.2	Percentage (with symbol % omitted) of correct estimation by ladle estimator and BIC estimator coupled with the ZMP criterion	155
10.3	p -values for DR-sequential test for the wine data	157
11.1	Types of forward regression estimators	189
13.1	Performance of GSIR (with options $\Lambda_{\text{GSIR}}^{(1)}$ and $\Lambda_{\text{GSIR}}^{(2)}$) under Models I, II, III. Scenarios A, B, C, and four different combinations of sample sizes and dimensions	227
14.1	Comparison of GSAVE and GSIR under Models IV, V, VI and Scenarios A, B, C	249
14.2	Performance of GSAVE under Models I, II, III and Scenarios A, B, C	249



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Preface

Sufficient Dimension Reduction is a rapidly developing research field that has wide applications in regression diagnostics, data visualization, Machine Learning, Genomics, image processing, pattern recognition, and medicine, which often contain a large number of variables. The purpose of the book is to introduce the basic theories and the main methodologies that have been developed in this field, to explore the key technical machineries that have been proven useful for conducting related research, to provide practical and easy-to-use algorithms and computer codes to implement these methodologies, and to survey the recent advances in the frontier of this field, which has grown too vast to be covered in detail in a single book.

Sufficient Dimension Reduction is a powerful tool to extract the core information hidden in the high-dimensional data, for the purpose of classifying or predicting one or several response variables. The extraction of information is based on the notion of sufficiency, which means a set of functions of the predictors provides all the information needed to understand the response, so that the rest of the predictors can be ignored without loss of information. Sufficiency is derived from conditional independence, a statistical concept that plays the central role in this theory.

Sufficient Dimension Reduction is akin to Principal Component Analysis — they both try to organize the variations in the data in an intelligent and interpretable way. However, Principal Component Analysis organizes the variations in the data itself, according to the magnitudes of variations; whereas Sufficient Dimension Reduction organizes the variations in the predictor according to how much they can explain the response variables. Sufficient Dimension Reduction is also akin to variable selection — they both try to reduce the number of variables that predict the response. However, variable selection tries to reduce the number of coordinates in the predicting vector; whereas Sufficient Dimension Reduction tries to reduce the predictor to a few linear combinations, or a few nonlinear functions, of the coordinates. In other words, variable selection reduces the data to achieve sparsity; Sufficient Dimension Reduction reduces the data to achieve low rank.

Sufficient Dimension Reduction has undergone momentous development in recent years, partly due to the increased demands for techniques to process high-dimensional data, a hallmark of our age of Big Data. The heightened development is also propelled by the increased complexity of the data structure. The classical dimension reduction problem proposed in the early 90's was concerned with a single response variable and a vector of continuous predicting variables; it used linear combinations as the sufficient predictors; its objective was to reduce the predictor in the conditional distribution. Since then, Sufficient Dimension Reduction has ex-

panded in many directions. For example, the predictors and the responses can both be functions or vectors of functions; the predictor can be matrix- or tensor-valued; the predictors can have grouped structures, and can be either continuous or categorical. The sufficient predictors are no longer limited to linear functions; it can be a member of a reproducing kernel Hilbert space. The target of reduction is no longer restricted to the whole conditional distribution; they can be the conditional means, conditional quantiles, conditional variances, or other conditional functionals of the response, according to our primary interests in the study.

The book is organized around four main themes. The first three themes belong to linear Sufficient Dimension Reduction: the inverse regression methods, order determination methods and related asymptotic developments, and the forward regression methods. The last theme is nonlinear Sufficient Dimension Reduction.

Specifically, [Chapter 1](#) introduces the preliminary tools that will be used throughout the book, as well as some backgrounds and motivations. [Chapters 2](#) lays out the basic theoretical framework, such as Sufficient Dimension Reduction subspaces and Fisher consistency. [Chapters 3](#) through [6](#) develop a variety of inverse regression estimators, such as the Sliced Inverse Regression, the Parametric and the Kernel Inverse Regression, the Sliced Average Variance Estimate, Contour Regression, and Directional Regression. [Chapter 6](#) discusses the key assumption — the elliptical distribution assumption — that underlies these inverse regression methods. [Chapter 7](#) introduces the dimension reduction framework where the conditional mean is of interest. [Chapters 8](#) and [9](#) cover the order determination methods that determine the number of sufficient predictors to be extracted from the data. [Chapter 10](#), a relatively long chapter, covers the forward regression methods, such as the Outer Product of Gradients, the Minimal Average Variance Estimator, and the Ensemble Estimator. [Chapters 12](#) through [14](#) cover Nonlinear Sufficient Dimension Reduction, which includes the basic theory, the Generalized Sliced Inverse Regression, and the Generalized Sliced Average Variance Estimator. In the last chapter, [Chapter 15](#), we give an overview of the developments that cannot be explored in detail in the previous chapters, which reveals the current scope and trends of this field.

This book grew out of the lecture notes I wrote when I taught such a course in the Spring of 2014 in the Department of Statistics of the Pennsylvania State University, chaired at the time by Professor D. Hunter. I thank the department for giving me such an opportunity and for the stimulating research environment. My work during this period has been supported by the National Science Foundation grants. My special thanks are due to Professor R. D. Cook, whose many inspiring discussions and collaborations during and before the writing of this book have benefited me greatly. I thank Professor X. Yin for reading a large part of the book. I thank my former students, K.-Y. Lee and W. Luo, for helping to collect and organize some computer codes. My other former students, S. Wang, Y. Dong, and A. Artemiou also contributed to the computing codes. I thank Professor B. Sriperumbudur for his useful discussions with me on the reproducing kernel Hilbert space.

Author

Bing Li obtained his Ph.D. from the University of Chicago in 1992. He is a Professor of Statistics at the Pennsylvania State University. His research interests cover Sufficient Dimension Reduction, Statistical Graphical Models, Functional Data Analysis, Machine Learning, Quasilikelihood, Estimating Equations, and Robust Statistics. He is a fellow of the Institute of Mathematical Statistics and a fellow of the American Statistical Association. He is serving as an Associate Editor for *The Annals of Statistics* and the *Journal of the American Statistical Association*.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Preliminaries

1.1	Empirical Distribution and Sample Moments	1
1.2	Principal Component Analysis	2
1.3	Generalized Eigenvalue Problem	3
1.4	Multivariate Linear Regression	3
1.5	Generalized Linear Model	5
1.5.1	Exponential Family	5
1.5.2	Generalized Linear Models	6
1.6	Hilbert Space, Linear Manifold, Linear Subspace	8
1.7	Linear Operator and Projection	10
1.8	The Hilbert Space $\mathbb{R}^p(\Sigma)$	11
1.9	Coordinate Representation	12
1.10	Generalized Linear Models under Link Violation	13

1.1 Empirical Distribution and Sample Moments

Let X be a random vector defined on a probability space (Ω, \mathcal{F}, P) , taking values in a measurable space $(\Omega_X, \mathcal{F}_X)$. Let X_1, \dots, X_n be independent copies of X . We assume Ω_X to be a subset of \mathbb{R}^p , the p -dimensional Euclidean space, and $\mathcal{F}_X = \{\Omega_X \cap B : B \in \mathcal{R}^p\}$, where \mathcal{R}^p is the Borel σ -field on \mathbb{R}^p .

Throughout this book, when there is a sample of n random vectors of p dimension, we always use subscript to indicate subjects, and superscript to indicate components. Thus X_i^k is the k th component of the i th subject. The symbol X_i without a superscript is used to denote the p -dimensional vector $(X_i^1, \dots, X_i^p)^\top$.

The empirical distribution of X based on X_1, \dots, X_n is defined to be the measure on $(\Omega_X, \mathcal{F}_X)$ that assigns n^{-1} mass to each X_i . This measure is denoted by F_n . That is,

$$F_n = n^{-1} \sum_{i=1}^n \delta_{X_i},$$

where δ_{X_i} is a point mass at X_i , defined as the set function

$$\delta_{X_i}(A) = \begin{cases} 1 & \text{if } X_i \in A \\ 0 & \text{if } X_i \notin A \end{cases}.$$

The measure F_n is a random measure, because it depends on the sample X_1, \dots, X_n .

The moments with respect to the measure F_n are called sample moments, and will be indicated by E_n . Thus, for a vector-valued function $f : \Omega_X \rightarrow \mathbb{R}^r$,

$$E_n f(X) = \int f(X) dF_n = n^{-1} \sum_{i=1}^n f(X_i) = n^{-1} \sum_{i=1}^n \begin{pmatrix} f_1(X_i) \\ \vdots \\ f_r(X_i) \end{pmatrix}.$$

The sample covariance matrix and the sample variance matrix can then be defined using E_n , as follows. If $g : \Omega_X \rightarrow \mathbb{R}^r$ is another vector-valued function, then $\text{cov}_n(f(X), g(X))$ is defined as

$$E_n[(f(X) - E_n f(X))(g(X) - E_n g(X))^T],$$

where $(\dots)^T$ denote the transpose of a matrix. The sample variance matrix $\text{var}_n[f(X)]$ is then defined to be the sample covariance matrix between $f(X)$ and $f(X)$; that is,

$$\text{var}_n[f(X)] = \text{cov}_n[f(X), f(X)].$$

1.2 Principal Component Analysis

Suppose X is a random vector in \mathbb{R}^p . The principal components of X are defined to be the set of linear combinations of X that have the largest variances. Thus, at the population level, the first principal component is defined through the following maximization problem:

$$\text{maximize } \text{var}(\alpha^T X) \quad \text{subject to } \|\alpha\| = 1.$$

Let α_1 be the solution to the above problem. Then $\alpha_1^T X$ is called the first principal component at the population level. Let $\Sigma = \text{var}(X)$. Then $\text{var}(\alpha^T X) = \alpha^T \Sigma \alpha$, and so α_1 is the first eigenvector of Σ . Similarly, the k th principal component of X is defined by the problem of

$$\begin{aligned} &\text{maximizing } \alpha^T \Sigma \alpha \\ &\text{subject to } \|\alpha\| = 1, \ell = 1, \dots, k-1, \alpha^T \alpha_\ell = 0. \end{aligned} \tag{1.1}$$

The solution is the k th eigenvector of Σ . The k th principal component at the population level is defined as the random variable $\alpha_k^T X$.

Intuitively, the random variable $\alpha_1^T X$ explains the most variation in X ; $\alpha_2^T X$ explains most variation in X left in the orthogonal complement of α_1 . In this way, we decompose the variations of X sequentially by orthogonal linear combinations.

At the sample level, suppose that X_1, \dots, X_n is an independent and identically distributed (i.i.d.) sample of X . Let $\hat{\Sigma} = \text{var}_n(X)$, and let $\hat{\alpha}_1, \dots, \hat{\alpha}_k$ be the first k eigenvectors of $\hat{\Sigma}$. The first k sample-level principal components of X are

$$\{\hat{\alpha}_\ell^\top X_i : i = 1, \dots, n\}, \quad \ell = 1, \dots, k.$$

1.3 Generalized Eigenvalue Problem

Principal Component Analysis is one of many problems that can be formulated as a generalized eigenvalue problem. Let Σ and Λ be symmetric matrix and Λ be positive definite. The generalized eigenvalue problem is defined by the following iterative optimization problem: at the k th step

$$\begin{aligned} & \text{maximizing} \quad \alpha^\top \Sigma \alpha \\ & \text{subject to} \quad \alpha^\top \Lambda \alpha = 1, \quad \alpha^\top \Lambda \alpha_\ell = 0, \quad \ell = 1, \dots, k-1, \end{aligned} \tag{1.2}$$

where $\alpha_1, \dots, \alpha_{k-1}$ are the maximizers in the previous $k-1$ steps. This is a generalization of problem (1.1) and can be reduced to it by making the transformation $\beta = \Lambda^{1/2} \alpha$. Then this problem becomes

$$\begin{aligned} & \text{maximizing} \quad \beta^\top \Lambda^{-1/2} \Sigma \Lambda^{-1/2} \beta \\ & \text{subject to} \quad \beta^\top \beta = 1, \quad \beta^\top \beta_\ell = 0, \quad \ell = 1, \dots, k-1. \end{aligned}$$

Thus, the solution to problem (1.2) is $\alpha_k = \Lambda^{-1/2} \beta_k$, where β_k is the k th eigenvector of the symmetric matrix $\Lambda^{-1/2} \Sigma \Lambda^{-1/2}$.

We call α_k the k th eigenvector of the generalized eigenvalue problem (Σ, Λ) . We abbreviate the phrase “generalized eigenvalue problem with respect to (Σ, Λ) ” as GEV (Σ, Λ) .

1.4 Multivariate Linear Regression

Let U and V be random vectors in \mathbb{R}^p and \mathbb{R}^q . In multivariate linear regression, at the population level, we are interested in minimizing the least squares criterion

$$E\|U - BV\|^2$$

over all matrices in $\mathbb{R}^{p \times q}$. This problem has an explicit solution, which will be useful in discussing many problems in Sufficient Dimension Reduction.

Henceforth, we will say a random vector V is square integrable if $E\|V\|^2 < \infty$. By the Cauchy-Schwarz inequality, this is true if and only if each component of V has finite second moment. In the following, if A is a positive definite matrix, we write $A > 0$.

Theorem 1.1 *Suppose U and V are square integrable with $E(U) = 0$ and $E(V) = 0$ and $\text{var}(V) > 0$. Then $E\|U - BV\|^2$ is uniquely minimized over $\mathbb{R}^{p \times q}$ by*

$$B^* = E(UV^\top)[E(VV^\top)]^{-1}.$$

PROOF. First, expand $E\|U - BV\|^2$ as

$$\begin{aligned} E\|U - BV\|^2 &= E\|U - B^*V + B^*V - BV\|^2 \\ &= E\|U - B^*V\|^2 + 2\text{tr}E[(U - B^*V)(B^*V - BV)^\top] + E\|B^*V - BV\|^2, \end{aligned} \quad (1.3)$$

where $\text{tr}(\cdots)$ stands for the trace of a matrix. The middle term on the right-hand side is 0, because

$$\begin{aligned} E[(U - B^*V)(B^*V - BV)^\top] &= E[(U - B^*V)V^\top](B^* - B)^\top \\ &= [E(UV^\top) - E(UV^\top)](B^* - B)^\top = 0. \end{aligned}$$

Therefore

$$E\|U - BV\|^2 \geq E\|U - B^*V\|^2$$

for all $B \in \mathbb{R}^{p \times q}$.

To see that the minimizer B^* is unique, we note that if $B \neq B^*$, then the third term on the right-hand side of (1.3) is

$$E\|B^*V - BV\|^2 = \text{tr}[(B^* - B)\text{var}(V)(B^* - B)^\top],$$

which is greater than 0 because $\text{var}(V)$ is positive definite. \square

There are several variations of [Theorem 1.1](#) that will also be useful.

Corollary 1.1 *Suppose U and V are square integrable and $\text{var}(V) > 0$. Then the function $E\|U - a - BV\|^2$ is minimized uniquely by*

$$B^* = \text{cov}(U, V)[\text{var}(V)]^{-1}, \quad a^* = EU - B^*EV.$$

PROOF. Let $U_c = U - E(U)$ and $V_c = V - E(V)$. Then

$$E\|U - a - BV\|^2 = E\|U_c - BV_c\|^2 + \|EU - a - BE(V)\|^2$$

By [Proposition 1.1](#) the first term is minimized at

$$B^* = E(U_c V_c^\top)(E V_c V_c^\top)^{-1} = \text{cov}(U, V)[\text{var}(V)]^{-1}.$$

The second term is 0 if $a^* = E(U) - B^*E(V)$. \square

This result is also applicable if we replace the true distribution of (U, V) by its empirical distribution. Let $(U_1, V_1), \dots, (U_n, V_n)$ be an i.i.d. sample of (U, V) .

Corollary 1.2 *If $\text{var}_n(V) > 0$, then the criterion $E_n\|U - a - BV\|^2$ is uniquely minimized by*

$$\hat{B} = \text{cov}_n(U, V)(\text{var}_n V)^{-1}, \quad \hat{a} = E_n U - \hat{B} E_n V.$$

1.5 Generalized Linear Model

Since one of the first ideas of Sufficient Dimension Reduction stems from a study of Generalized Linear Models under link violation (Li and Duan (1989), Li (1991)), it is helpful to review the basic structure and properties of the Generalized Linear Models. For more information on this topic, see McCullagh and Nelder (1989).

1.5.1 Exponential Family

Let Y be a random variable that takes values in $(\Omega_Y, \mathcal{F}_Y)$. We say that the distribution of Y belongs to an exponential family if the probability density function (p.d.f.) of Y has the form $c(\theta)e^{\theta y}$ with respect to some σ -finite measure ν on Ω_Y . This can be rewritten as

$$e^{\theta y - b(\theta)},$$

where $b(\theta) = -\log c(\theta)$. The moment generating function of Y can be easily computed, as follows:

$$M_Y(t) = \int e^t e^{\theta y - b(\theta)} d\nu(y) = e^{b(t+\theta) - b(\theta)} \int e^{(t+\theta)y - b(t+\theta)} d\nu(y) = e^{b(t+\theta) - b(\theta)}.$$

The cumulant generating function, defined as the natural log of the moment generating function, is then

$$C_Y(\theta) = b(t + \theta) - b(\theta).$$

The derivatives of the cumulant generating function evaluated at $t = 0$ generate cumulants, the first two of which are the mean and the variance:

$$\dot{C}_Y(0) = E_\theta(Y), \quad \ddot{C}_Y(0) = \text{var}_\theta(Y). \quad (1.4)$$

See, for example, McCullagh (1987). It follows that

$$\dot{b}(\theta) = E_\theta(Y), \quad \ddot{b}(\theta) = \text{var}_\theta(Y).$$

From the second equality we see that if $\text{var}_\theta(Y) > 0$ for all θ , then \dot{b} is a monotone increasing function, and therefore its inverse \dot{b}^{-1} is a well defined function. If we denote $E_\theta(Y)$ by μ , then

$$\theta = \dot{b}^{-1}(\mu).$$

Moreover, $\text{var}_\theta(Y)$ can be reexpressed in μ as $\ddot{b}(\dot{b}^{-1}(\mu))$. The function $\ddot{b} \circ \dot{b}^{-1}$ characterizes the mean-variance relation in an exponential family, and is called the *variance function*. We denote the variance function by $V(\mu)$.

1.5.2 Generalized Linear Models

Let X be a random vector in \mathbb{R}^p as defined in [Section 1.1](#). In a Generalized Linear Model we assume that Y is related with X by the conditional density

$$f_{Y|X}(y|x) \propto e^{\theta(x)y - b(\theta(x))}, \quad (1.5)$$

where $\theta(x)$ is a function of x . The regression relation between Y and X is modeled through the link function. Note that

$$\theta(x) = \dot{b}^{-1}(E(Y|x)).$$

We model $E(Y|x)$ by

$$E(Y|x) = \mu(\eta), \quad \eta = \alpha + \beta^\top x,$$

where $\mu(\eta)$ is called the mean function and $\eta = \alpha + \beta^\top x$ is called the the linear predictor or the linear index. Usually, we assume $\mu(\cdot)$ to be one-to-one, and its inverse μ^{-1} is called the link function.

Substituting the relation $\theta(x) = \dot{b}^{-1}(\mu(\alpha + \beta^\top x))$ into the conditional density (1.5), we have

$$f_{Y|X}(y|x; \alpha, \beta) \propto \exp \left\{ (\dot{b}^{-1} \circ \mu)(\alpha + \beta^\top x)y - b((\dot{b}^{-1} \circ \mu)(\alpha + \beta^\top x)) \right\}. \quad (1.6)$$

In Generalized Linear Models, α and β are estimated by maximum likelihood estimation based on the density (1.6). Suppose that $\mathbb{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are a sample of i.i.d. observations on (X, Y) . Then the joint log likelihood is proportional to

$$\begin{aligned} \ell(\alpha, \beta; \mathbb{D}_n) &= E_n \left\{ (\dot{b}^{-1} \circ \mu)(\alpha + \beta^\top X)X - b((\dot{b}^{-1} \circ \mu)(\alpha + \beta^\top X)) \right\} \\ &= E_n \left\{ (\dot{b}^{-1} \circ \mu)(\gamma^\top \tilde{X})Y - b((\dot{b}^{-1} \circ \mu)(\gamma^\top \tilde{X})) \right\}, \end{aligned} \quad (1.7)$$

where

$$\gamma = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} 1 \\ X \end{pmatrix}.$$

Differentiate (1.7) with respect to γ to obtain

$$\partial \ell(\gamma; \mathbb{D}_n) / \partial \gamma = E_n \left\{ \partial [(\dot{b}^{-1} \circ \mu)(\gamma^\top \tilde{X})Y] / \partial \gamma - \partial [b((\dot{b}^{-1} \circ \mu)(\gamma^\top \tilde{X}))] / \partial \gamma \right\}.$$

This function is called the *score function*, and we denote it by $s(\gamma; \mathbb{D}_n)$. The derivatives in the score function are computed by the chain rule:

$$\frac{\partial (\dot{b}^{-1} \circ \mu)(\gamma^\top \tilde{X})}{\partial \gamma} = \frac{\partial \dot{b}^{-1}(\mu)}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \gamma} = \frac{\tilde{X} \dot{\mu}(\gamma^\top \tilde{X})}{\dot{b}(\dot{b}^{-1}(\mu(\gamma^\top \tilde{X})))} = \frac{\tilde{X} \dot{\mu}(\gamma^\top \tilde{X})}{V(\mu(\gamma^\top \tilde{X}))}.$$

Here, $\dot{\mu}(\eta)$ denote the function $\eta \mapsto \partial \mu / \partial \eta$. Similarly,

$$\frac{\partial b((\dot{b}^{-1} \circ \mu)(\gamma^\top \tilde{X}))}{\partial \gamma} = \frac{\partial b(\theta)}{\partial \theta} \bigg|_{\theta = \dot{b}^{-1}(\mu)} \times \frac{\partial \dot{b}^{-1}(\mu)}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \gamma} = \frac{\tilde{X} \dot{\mu}(\gamma^\top \tilde{X}) \mu(\gamma^\top \tilde{X})}{V(\mu(\gamma^\top \tilde{X}))}.$$

Hence the score function is written explicitly as

$$s(\gamma; \mathbb{D}_n) = E_n \left\{ \frac{\tilde{X} \dot{\mu}(\gamma^T \tilde{X}) [Y - \mu(\gamma^T \tilde{X})]}{V(\mu(\gamma^T \tilde{X}))} \right\}.$$

This is completely specified by the mean function μ , which is our regression model, and the mean-variance relation $V(\mu)$, which is determined by the exponential family.

The parameter γ is usually estimated by the maximum likelihood estimation. Under the exponential family assumption, the log likelihood is concave and differentiable. Thus the maximum likelihood estimate can be found by solving the *likelihood equation*

$$s(\gamma; \mathbb{D}_n) = 0.$$

This is usually solved by the Newton-Raphson algorithm, or the Fisher scoring method. See, for example, Section 2.5.1 of McCullagh and Nelder (1989) for details.

The link function that makes $\theta(x) = \gamma^T \tilde{x}$ is called the natural link, or the canonical link. In other words μ has to make $\dot{b}^{-1} \circ \mu$ the identity mapping, which implies $\mu^{-1} = \dot{b}^{-1}$. Under the natural link the conditional density (1.6) reduces to

$$f_{Y|X}(y|x; \gamma) \propto \exp \{ (\gamma^T \tilde{x})y - b(\gamma^T \tilde{x}) \}.$$

The score function reduces to the simple form

$$s(\gamma; \mathbb{D}_n) = E_n [\tilde{X}(Y - \mu(\gamma^T \tilde{X}))].$$

We now illustrate the Generalized Linear Models by two simple examples.

Example 1.1 Suppose $Y \sim \text{Poisson}(\lambda)$. Then

$$f(y; \theta) \propto \lambda^y e^{-\lambda} = e^{y \log \lambda - \lambda} = e^{\theta y - e^\theta}.$$

Here, λ is the conventional parameter of a Poisson distribution, $\theta = \log \lambda$ is the canonical parameter, and the cumulant generating function of Y is

$$C_Y(t) = e^{\theta+t} - e^\theta.$$

From this we see that

$$\dot{b}^{-1}(\mu) = \log \mu, \quad \ddot{b}(\theta) = e^\theta, \quad V(\mu) = \exp(\log(\mu)) = \mu.$$

The natural link function is $\dot{b}^{-1}(\mu) = \log(\mu)$, and the score function is simply

$$E_n[\tilde{X}(Y - e^{\gamma^T \tilde{X}})] = 0.$$

This model is also known as the log linear regression model. □

Example 1.2 Suppose, for a fixed p , Y has a binomial distribution $b(n, p)$, where p is a function of x . That is,

$$f(y) = \binom{n}{x} p^y (1-p)^{n-y} \propto e^{y \log \frac{p}{1-p} + n \log(1-p)}.$$

If we let $\theta = \log[p/(1-p)]$, then $n \log(1-p) = -n \log(1+e^\theta)$. The density $f(y)$ can be rewritten as the canonical form

$$f(y) \propto \exp[\theta y - n \log(1+e^\theta)].$$

Hence

$$b(\theta) = n \log(1+e^\theta), \quad \dot{b}(\theta) = n \frac{e^\theta}{1+e^\theta}, \quad \ddot{b}(\theta) = n \frac{e^\theta}{(1+e^\theta)^2}.$$

It follows that

$$\dot{b}^{-1}(\mu) = \log \frac{\mu/n}{1-\mu/n}, \quad (\ddot{b} \circ \dot{b}^{-1})(\mu) = n(\mu/n)(1-\mu/n).$$

Thus the natural link function is $\log \frac{\mu/n}{1-\mu/n}$, which is called the logit function, and the score function is

$$s(\gamma; \mathbb{D}_n) = E_n \left[\tilde{X} \left(Y - n \frac{e^{\gamma^\top \tilde{X}}}{1 + e^{\gamma^\top \tilde{X}}} \right) \right].$$

This type of Generalized Linear Model is called the logistic regression. □

1.6 Hilbert Space, Linear Manifold, Linear Subspace

The theory of Sufficient Dimension Reduction is geometric in nature, where inner product, orthogonality, and projection play a critical role. In this and the next two sections we bring together some geometric concepts and machineries that will be used repeatedly in this book. When developing these concepts we follow this path:

$$\text{group} \rightarrow \text{Abelian group} \rightarrow \text{vector space} \rightarrow \begin{cases} \text{normed space} \rightarrow \text{Banach space} \\ \text{inner product space} \rightarrow \text{Hilbert space} \end{cases}$$

More information about these topics can be found in Kelley (1955) and Conway (1990).

Let \mathcal{H} be a set. Let $+$ be a mapping from $\mathcal{H} \times \mathcal{H}$ to \mathcal{H} such that the following conditions are satisfied:

1. $+(g_1, g_2), g_3) = +(g_1, +(g_2, g_3))$;
2. there is a member e of \mathcal{H} such that $+(e, g) = +(g, e) = g$ for all $g \in \mathcal{H}$;
3. for each $g \in \mathcal{H}$, there is a member $f \in \mathcal{H}$ such that $+(g, f) = e$.