Monographs on Statistics and Applied Probability 161

Sufficient Dimension Reduction

Methods and Applications with R

Bing Li

CPD CERTIFIED
The CPD Certification Service

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Sufficient Dimension Reduction

Methods and Applications with R

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

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

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

Printed on acid-free paper
Version Date: 20180404

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

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

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

To Yanling

Contents

List of Figures

List of Tables

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 expanded 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.

State College *Bing Li*

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*.

Chapter 1

Preliminaries

1.1 Empirical Distribution and Sample Moments

Let X be a random vector defined on a probability space $(\Omega, \mathscr{F}, P)$, taking values in a measurable space $(\Omega_X, \mathscr{F}_X)$. Let $X_1, \ldots, X_n$ be independent copies of X. We assume Ω_X to be a subset of $\mathbb{R}^p$, the p-dimensional Euclidean space, and $\mathscr{F}_X = \{\Omega_X \cap B : B \in \mathscr{R}^p\}$, where $\mathscr{R}^p$ 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 kth component of the ith subject. The symbol X_i without a superscript is used to denote the p-dimensional vector $(X_i^1, \ldots, X_i^p)^\mathsf{T}$.

The empirical distribution of X based on $X_1, \ldots, X_n$ is defined to be the measure on $(\Omega_X, \mathscr{F}_X)$ 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 $X_1, \ldots, X_n$.

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 $f : \Omega_X \to \mathbb{R}^r$,

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

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

$$E_n[(f(X) - E_n f(X))(g(X) - E_n g(X))^{\mathsf{T}}],$$

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

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

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:

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

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

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

The solution is the kth eigenvector of Σ. The kth principal component at the population level is defined as the random variable $\alpha_k^{\mathsf{T}} X$.

Intuitively, the random variable $\alpha_1^{\mathsf{T}} X$ explains the most variation in X; $\alpha_2^{\mathsf{T}} X$ 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 $X_1, \ldots, X_n$ is an independent and identically distributed (i.i.d.) sample of X. Let $\hat{\Sigma} = \text{var}_n(X)$, and let $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ be the first k eigenvectors of $\hat{\Sigma}$. The first k sample-level principal components of X are

$$\{\hat{\alpha}_\ell^\mathsf{T} X_i : i = 1, \ldots, n\}, \quad \ell = 1, \ldots, k.$$

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 kth step

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

where $\alpha_1, \ldots, \alpha_{k-1}$ 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

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

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

We call α_k the kth eigenvector of the generalized eigenvalue problem (Σ, Λ). We abbreviate the phrase "generalized eigenvalue problem with respect to (Σ, Λ)" as $\text{GEV}(\Sigma, \Lambda)$.

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*

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

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

$$\begin{aligned} E\|U-BV\|^2 &= E\|U-B^*V+B^*V-BV\|^2 \\ &= E\|U-B^*V\|^2+2\mathrm{tr}E[(U-B^*V)(B^*V-BV)^{\mathsf{T}}]+E\|B^*V-BV\|^2, \end{aligned} \tag{1.3}$$

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

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

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

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

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

There are several variations of Theorem 1.1 that will also be useful.

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

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

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

$$B^* = E(U_cV_c^{\mathsf{T}})(EV_cV_c^{\mathsf{T}})^{-1} = \mathrm{cov}(U,V)[\mathrm{var}(V)]^{-1}.$$

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

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

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

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

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 $(\Omega_Y, \mathscr{F}_Y)$. 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:

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

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

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

From the second equality we see that if $\mathrm{var}_\theta(Y) > 0$ 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, $\mathrm{var}_\theta(Y)$ 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

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

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

$$E(Y|x) = \mu(\eta), \quad \eta = \alpha + \beta^{\mathsf{T}} x,$$

where $\mu(\eta)$ is called the mean function and $\eta = \alpha + \beta^{\mathsf{T}} x$ 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 $\theta(x) = \dot{b}^{-1}(\mu(\alpha + \beta^{\mathsf{T}} x))$ into the conditional density (1.5), we have

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

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

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

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

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

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:

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

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

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

Hence the score function is written explicitly as

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

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 $\theta(x) = \gamma^{\mathsf{T}}\tilde{x}$ 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

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

The score function reduces to the simple form

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

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

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

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

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

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:

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

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

Let $\mathscr{H}$ be a set. Let $+$ be a mapping from $\mathscr{H} \times \mathscr{H}$ to $\mathscr{H}$ 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 $\mathscr{H}$ such that $+(e,g) = +(g,e) = g$ for all $g \in \mathscr{H}$;
3. for each $g \in \mathscr{H}$, there is a member $f \in \mathscr{H}$ such that $+(g,f) = e$.