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,

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

Copyright © ${Date}. ${Publisher}. All rights reserved.

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.

Copyright © ${Date}. ${Publisher}. All rights reserved.

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

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

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 kth step

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

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

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

Thus, the solution to problem (1.2) is $\alpha_k = \Lambda^{-1/2} \beta_k$, where β_k is the 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*

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

Copyright © ${Date}. ${Publisher}. All rights reserved.

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*

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

Copyright © ${Date}. ${Publisher}. All rights reserved.

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:

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

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 $V(\mu)$.

Copyright © ${Date}. ${Publisher}. All rights reserved.

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,

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

Copyright © ${Date}. ${Publisher}. All rights reserved.

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. □

Copyright © ${Date}. ${Publisher}. All rights reserved.

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:

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

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

Let $\mathscr{H}$ 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 $+(g,f) = e$.

Copyright © ${Date}. ${Publisher}. All rights reserved.

The pair $(\mathscr{H},+)$ of the set $\mathscr{H}$ and the operation $+$ is called a group. A group is an Abelian group or commutative group if it satisfies the additional condition

4. $+(g,f) = +(f,g)$ for all $f,g \in \mathscr{H}$.

Usually, we write $+(g,f)$ as $g+f$, write f in statement 3 as $-g$, and write the e in statement 2 as 0. We call it the zero element of $\mathscr{H}$.

Now suppose $(\mathscr{H},+)$ is an Abelian group, and suppose there is a mapping $\cdot$ from $\mathbb{R}\times\mathscr{H}$ to $\mathscr{H}$ such that

5. for any $a,b \in \mathbb{R}$ and $f \in \mathscr{H}$, $\cdot(a,\cdot(b,f)) = \cdot(ab,f)$
6. for any $a \in \mathbb{R}$ and $f,g \in \mathscr{H}$, $\cdot(a,f+g) = \cdot(a,f)+\cdot(a,g)$.
7. for any $a,b \in \mathbb{R}$, $f \in \mathscr{H}$, $\cdot(a+b,f) = \cdot(a,f)+\cdot(b,f)$
8. for any $f \in \mathscr{H}$, $\cdot(1,f) = f$.

Usually, we write $\cdot(a,f)$ as $a\cdot f$ or simply af. An Abelian group $(\mathscr{H},+)$, together with the mapping $\cdot : \mathbb{R}\times\mathscr{H}\to\mathscr{H}$ that satisfies the above conditions, is called a vector space, and is denoted by $(\mathscr{H},+,\cdot)$. For simplicity, we will just say $\mathscr{H}$ is a vector space, without writing $+,\cdot$ explicitly.

For a vector space $\mathscr{H}$, if there is a mapping $u : \mathscr{H}\times\mathscr{H}\to\mathbb{R}$ that satisfies the following conditions:

9. for any $f,g \in \mathscr{H}$, $u(f,g) = u(g,f)$
10. for any $f,g,h \in \mathscr{H}$, $u(f+g,h) = u(f,h)+u(g,h)$
11. for any $a \in \mathbb{R}$, $f,g \in \mathscr{H}$, $u(af,g) = au(f,g)$,

then u is called a semi-inner product, and $(\mathscr{H},u)$ is called a semi-inner product space. Usually, we write $u(f,g)$ as $\langle f,g\rangle$. If, in addition,

12. for any $f \in \mathscr{H}$, $u(f,f) = 0$ implies f is the 0 element of $\mathscr{H}$,

then we call u an inner product and $(\mathscr{H},u)$ an inner product space.

Suppose $\mathscr{H}$ is a vector space. If there is a mapping $\rho : \mathscr{H}\to\mathbb{R}$ satisfies the following conditions

9′. for any $a \in \mathbb{R}$, $f \in \mathscr{H}$, $\rho(af) = |a|\rho(f)$

10′. for any $f,g \in \mathscr{H}$, $\rho(f+g) \le \rho(f)+\rho(g)$

11′. for any $f \in \mathscr{H}$, $\rho(f) = 0$ implies that f is the 0 element in $\mathscr{H}$.

Then we call ρ a norm in $\mathscr{H}$ and $(\mathscr{H},\rho)$ a normed space. Usually, we write $\rho(f)$ as $\|f\|$.

Suppose $(\mathscr{H},u)$ is an inner product space, then it can be shown that the mapping

$$\mathscr{H}\to\mathbb{R}, \quad f\mapsto\langle f,f\rangle^{1/2}$$

is a norm in $\mathscr{H}$. So an inner product space is a special normed space with its norm defined by $\|f\| = \langle f,f\rangle^{1/2}$.

A sequence $\{f_n\}$ in a normed space $(\mathscr{H},\|\cdot\|)$ is called a Cauchy sequence if, for any $\varepsilon > 0$, there is an m such that for all $n_1,n_2 > m$, $\|f_{n_1}-f_{n_2}\| < \varepsilon$. A normed space $(\mathscr{H},\|\cdot\|)$ is said to be complete if every Cauchy sequence in $\mathscr{H}$ converges to a member of $\mathscr{H}$. That is, there is $f \in \mathscr{H}$ such that $\|f_n - f\| \to 0$. A complete normed

Copyright © ${Date}. ${Publisher}. All rights reserved.

space is called a Banach space. If an inner product space $(\mathscr{H}, \langle\cdot,\cdot\rangle)$ is complete in terms of the norm $\|f\| = \langle f, f\rangle^{1/2}$, then it is called a Hilbert space.

Suppose $\mathscr{H}$ is a vector space. A subset $\mathscr{S} \subseteq \mathscr{H}$ is called a linear manifold in $\mathscr{H}$ if

1. for any $f, g \in \mathscr{S}$, $f + g \in \mathscr{S}$
2. for $a \in \mathbb{R}$, $f \in \mathscr{S}$, $af \in \mathscr{S}$.

If $\mathscr{S}$ is a linear manifold in $\mathscr{H}$ and $\mathscr{S}$ is closed, then $\mathscr{S}$ is a linear subspace of $\mathscr{H}$.

1.7 Linear Operator and Projection

In later parts of the book, on those topics related to functional data and kernel mapping, we will frequently employ linear operators in Hilbert spaces and their coordinate representations.

Let $\mathscr{H}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$. A linear operator is a mapping $T : \mathscr{H} \to \mathscr{H}$ such that

1. for any $f, g \in \mathscr{H}$, $T(f+g) = T(f) + T(g)$
2. for any $a \in \mathbb{R}$, $T(af) = aT(f)$.

A linear operator T is said to be idempotent if $T^2 = T$. That is, for any $f \in \mathscr{H}$, $T(T(f)) = T(f)$. A linear operator T is self adjoint if, for any $f, g \in \mathscr{H}$,

$$\langle f, Tg\rangle = \langle Tf, g\rangle.$$

If a linear operator $P : \mathscr{H} \to \mathscr{H}$ is both idempotent and self adjoint, then it is called a projection.

If T is a linear operator, then $\ker(T)$, the kernel of T, is the set $\{f \in \mathscr{H} : Tf = 0\}$. It is easy to show that $\ker(T)$ is a linear manifold; it can also be shown that $\ker(T)$ is closed. Therefore $\ker(T)$ is a linear subspace of $\mathscr{H}$. The symbol $\mathrm{ran}\,(T)$ stands for the range of T, which is the set $\{Tf : f \in \mathscr{H}\}$. It is also easy to verify that $\mathrm{ran}\,(T)$ is a linear manifold in $\mathscr{H}$, but it may not be closed, and consequently $\mathrm{ran}\,(T)$ may not be a subspace of $\mathscr{H}$. We use $\overline{\mathrm{ran}}\,T$ to denote the closure of $\mathrm{ran}\,(T)$, which is always a subspace of $\mathscr{H}$. The range of a projection P is always closed, implying $\mathrm{ran}\,(P) = \overline{\mathrm{ran}}\,(P)$.

Sometimes we call a projection P the projection on to the subspace $\mathscr{S}$, where $\mathscr{S}$ is $\mathrm{ran}\,(P)$. Conversely, for any subspace $\mathscr{S}$ of $\mathscr{H}$, there is a unique projection P, whose range is $\mathscr{S}$.

Let $\mathscr{S}$ be a subspace of $\mathscr{H}$ and let f be a member of $\mathscr{H}$. Then there is a unique member of $\mathscr{S}$, say f^*, such that $f - f^* \perp \mathscr{S}$. Furthermore, it can be shown

$$f - f^* \perp \mathscr{S} \Leftrightarrow \|f - f^*\| \leq \|f - g\| \quad \text{for all } g \in \mathscr{S}.$$

The unique existence of f^* defines the mapping

$$\mathscr{H} \to \mathscr{S}, \quad f \mapsto f^*.$$

It can be shown that this mapping is precisely the mapping $P_{\mathscr{S}}$.

Copyright © ${Date}. ${Publisher}. All rights reserved.

1.8 The Hilbert Space $\mathbb{R}^p(\Sigma)$

Now let us consider the special case of $\mathscr{H} = \mathbb{R}^p$. A member v of $\mathbb{R}^p$ is the vector $(v_1, \ldots, v_p)^{\mathsf{T}}$. For $u, v \in \mathbb{R}^p$, define

$$u + v = (u_1 + v_1, \ldots, u_p + v_p)^{\mathsf{T}}.$$

Then $(\mathbb{R}^p, +)$ is an Abelian group. For $v \in \mathbb{R}^p$, $a \in \mathbb{R}$, define

$$\cdot(a, v) = av = (av_1, \ldots, av_p)^{\mathsf{T}}.$$

Then $(\mathbb{R}^p, +, \cdot)$ is a vector space. Let $\Sigma \in \mathbb{R}^{p \times p}$ be a positive definite matrix. For $u, v \in \mathbb{R}^p$, define

$$\langle u, v \rangle = u^{\mathsf{T}} \Sigma v.$$

Then $(\mathbb{R}^p, +, \cdot, \langle \cdot, \cdot \rangle)$ is an inner product space. It is true that any finite-dimensional normed space is complete. Therefore $(\mathbb{R}^p, +, \cdot, \langle \cdot, \cdot \rangle)$ is also a Hilbert space. We will abbreviate this Hilbert space by the simple $\mathbb{R}^p(\Sigma)$.

Let $u_1, \ldots, u_m$ be a collection of vectors in $\mathbb{R}^p$. Let $\mathscr{S}$ be the set

$$\operatorname{span}(u_1, \ldots, u_n) = \{c_1 u_1 + \cdots + c_m u_m : c_1, \ldots, c_m \in \mathbb{R}\}.$$

It is easy to see that this is a linear manifold in $\mathbb{R}^p$. Because this linear manifold has finite dimension, it is also closed. Thus it is a subspace of $\mathbb{R}^p$. We call this subspace the linear span of $u_1, \ldots, u_m$, and write it as $\operatorname{span}(u_1, \ldots, u_m)$. Conversely, any linear subspace of $\mathscr{S}$ is a linear span of a set of vectors in $\mathbb{R}^p$. The projection on to a subspace of the Euclidean space $\mathbb{R}^p$ can be expressed explicitly. Let $\mathscr{S} = \operatorname{span}\{u_1, \ldots, u_m\}$, and let B denote the matrix $(u_1, \ldots u_m)$. We use $P_{\mathscr{S}}$ to denote the projection on to the subspace $\mathscr{S}$. Let

$$P_B(\Sigma) = B(B^{\mathsf{T}} \Sigma B)^{-1} B^{\mathsf{T}} \Sigma.$$

Proposition 1.1 *The linear operator*

$$P_{\mathscr{S}}: \quad \mathbb{R}^p \to \mathbb{R}^p, \quad v \mapsto P_B(\Sigma) v$$

is a projection.

PROOF. We first note that

$$P_B(\Sigma) P_B(\Sigma) = B(B^{\mathsf{T}} \Sigma B)^{-1} (B^{\mathsf{T}} \Sigma B)(B^{\mathsf{T}} \Sigma B)^{-1} B^{\mathsf{T}} \Sigma = P_B(\Sigma).$$

Hence, for any $v \in \mathbb{R}^p(\Sigma)$, $P_{\mathscr{S}}^2(v) = [P_B(\Sigma)]^2 v = P_B(\Sigma) v = P_{\mathscr{S}}(v)$. Thus $P_{\mathscr{S}}$ is idempotent. Moreover, for any $u, v \in \mathbb{R}^p$,

$$\begin{aligned} \langle P_{\mathscr{S}}(u), v \rangle &= (P_B(\Sigma) u)^{\mathsf{T}} \Sigma v \\ &= u^{\mathsf{T}} \Sigma B (B^{\mathsf{T}} \Sigma B)^{-1} B^{\mathsf{T}} \Sigma v \\ &= u^{\mathsf{T}} \Sigma [B (B^{\mathsf{T}} \Sigma B)^{-1} B^{\mathsf{T}} \Sigma] v \\ &= \langle u, P_{\mathscr{S}}(v) \rangle. \end{aligned}$$

Thus $P_{\mathscr{S}}$ is self-adjoint. □

Copyright © ${Date}. ${Publisher}. All rights reserved.

1.9 Coordinate Representation

To implement kernel related methods for in Chapters 12 through 14, we need to use the coordinate representations of a function or a linear operator in finite-dimensional Hilbert spaces. Our notations are adopted from Horn and Johnson (1985). Let $\mathscr{H}$ be a finite-dimensional Hilbert space with spanning system $\mathscr{B} = \{b_1, \ldots, b_m\}$. Then any member f of $\mathscr{H}$ can be written as $c_1 b_1 + \cdots + c_m b_m$, where $c_1, \ldots, c_m \in \mathbb{R}$. The vector $(c_1, \ldots, c_m)^{\mathsf{T}}$ is called the coordinate of f relative to the spanning system $\mathscr{B}$. In the cases where $b_1, \ldots, b_m$ are linearly dependent, a member of $\mathscr{H}$ can have many coordinate representations, but this does not concern us because the function it represents is unique. The coordinate of f relative to a spanning system $\mathscr{B}$ is written as $[f]_{\mathscr{B}}$. Thus we can write $f = [f]_{\mathscr{B}}^{\mathsf{T}} b_{1:m}$, where $b_{1:m} = (b_1, \ldots, b_m)^{\mathsf{T}}$.

The matrix of inner products $\{\langle b_i, b_j\rangle_{\mathscr{H}} : i, j = 1, \ldots, m\}$ is called the Gram matrix of $\mathscr{B}$, and is written as $G_{\mathscr{B}}$. We can represent the inner product between two members of $\mathscr{H}$ using their coordinates and the Gram matrix as follows

$$\langle f, g\rangle_{\mathscr{H}} = \sum_{i=1}^{m} \sum_{j=1}^{m} ([f]_{\mathscr{B}})_i ([g]_{\mathscr{B}})_j \langle b_i, b_j\rangle_{\mathscr{H}} = [f]_{\mathscr{B}}^{\mathsf{T}} G_{\mathscr{B}} [g]_{\mathscr{B}},$$

where, for example, $([f]_{\mathscr{B}})_i$ stands for the ith component of the vector $[f]_{\mathscr{B}}$.

Let $\mathscr{H}_1$ and $\mathscr{H}_2$ be two finite dimensional Hilbert spaces spanned by the subsets $\mathscr{B}_1 = \{b_1^{(1)}, \ldots, b_{m_1}^{(1)}\} \subseteq \mathscr{H}_1$ and $\mathscr{B}_2 = \{b_1^{(2)}, \ldots, b_{m_2}^{(2)}\} \subseteq \mathscr{H}_2$, respectively. Let $A : \mathscr{H}_1 \to \mathscr{H}_2$ be a linear operator and f be a member of $\mathscr{H}_1$. Then Af is a member of $\mathscr{H}_2$ and its coordinate representation relative to $\mathscr{B}_2$ is obtained by the following calculation:

$$\begin{aligned} Af =& A\left(\sum_{i=1}^{m_1} ([f]_{\mathscr{B}_1})_i b_i^{(1)}\right) \\ =& \sum_{i=1}^{m_1} ([f]_{\mathscr{B}_1})_i A b_i^{(1)} \\ =& \sum_{i=1}^{m_1} ([f]_{\mathscr{B}_1})_i \sum_{j=1}^{m_2} ([A b_i^{(1)}]_{\mathscr{B}_2})_j b_j^{(2)}. \end{aligned}$$

We see that the coordinate of Af relative to $\mathscr{B}_2$ is simply the vector

$$\left\{ \sum_{i=1}^{m_1} ([f]_{\mathscr{B}_1})_i ([A b_i^{(1)}]_{\mathscr{B}_2})_j : j = 1, \ldots, m_2 \right\}.$$

Motivated by this relation, we write the matrix

$$\begin{pmatrix} ([A b_1^{(1)}]_{\mathscr{B}_2})_1 & \cdots & ([A b_{m_1}^{(1)}]_{\mathscr{B}_2})_1 \\ \vdots & & \vdots \\ ([A b_1^{(1)}]_{\mathscr{B}_2})_{m_2} & \cdots & ([A b_{m_1}^{(1)}]_{\mathscr{B}_2})_{m_2} \end{pmatrix}$$

as ${}_{\mathscr{B}_2}[A]_{\mathscr{B}_1}$, and call it the coordinate representation of the linear operator A. Using this notation we can conveniently write

$$[Af]_{\mathscr{B}_2} = ({}_{\mathscr{B}_2}[A]_{\mathscr{B}_1})[f]_{\mathscr{B}_1}.$$

Copyright © ${Date}. ${Publisher}. All rights reserved.

Carrying the logic in this notation further, let $\mathscr{H}_3$ be a third finite-dimensional Hilbert space with spanning system $\mathscr{B}_3 = \{b_1^{(3)}, \ldots, b_{m_3}^{(3)}\}$. Let $A_1 : \mathscr{H}_1 \to \mathscr{H}_2$ and $A_2 : \mathscr{H}_2 \to \mathscr{H}_3$ be linear operators, and let f be a member of $\mathscr{H}_1$. Then

$$\begin{aligned} {}_{\mathscr{B}_3}[A_2 A_1 f]_{\mathscr{B}_1} &= {}_{\mathscr{B}_3}[A_2(A_1 f)]_{\mathscr{B}_1} \\ &= ({}_{\mathscr{B}_3}[A_2]_{\mathscr{B}_2})[A_1 f]_{\mathscr{B}_1} \\ &= ({}_{\mathscr{B}_3}[A_2]_{\mathscr{B}_2})({}_{\mathscr{B}_2}[A_1]_{\mathscr{B}_1})([f]_{\mathscr{B}_1}). \end{aligned}$$

In the meantime, we have

$${}_{\mathscr{B}_3}[(A_2 A_1) f]_{\mathscr{B}_1} = ({}_{\mathscr{B}_3}[A_2 A_1]_{\mathscr{B}_1})[f]_{\mathscr{B}_1}.$$

Comparing these two equations we have

$${}_{\mathscr{B}_3}[A_2 A_1]_{\mathscr{B}_1} = ({}_{\mathscr{B}_3}[A_2]_{\mathscr{B}_2})({}_{\mathscr{B}_2}[A_1]_{\mathscr{B}_1}).$$

1.10 Generalized Linear Models under Link Violation

As a prelude to Sufficient Dimension Reduction, we present a case study of a property of the Generalized Linear Model when the link function μ^{-1} (or equivalently, μ) is misspecified. This property was discovered by Li and Duan (1989), and is a starting point (and arguably the starting point) of Sufficient Dimension Reduction. This property states that the maximum likelihood estimate of α, β is Fisher consistent even if the link function is misspecified, provided there is some symmetry in X.

Recall that, under the canonical link, the log likelihood for the Generalized Linear Model is $(\alpha + \beta^{\mathsf{T}} X)Y - b(\alpha + \beta^{\mathsf{T}} X)$, where $\dot{b}$ is a monotone increasing function, which means that b is a convex function. The maximum likelihood estimation pertains to maximizing $E_n[(\alpha + \beta^{\mathsf{T}} X)Y - b(\alpha + \beta^{\mathsf{T}} X)]$. At the population level, we maximize the function

$$E[(\alpha + \beta^{\mathsf{T}} X)Y - b(\alpha + \beta^{\mathsf{T}} X)].$$

Let (α_0, β_0) represent the true values of (α, β). Then, by the well known theory of maximum likelihood estimation (see, for example, Lehmann and Casella (1998a), Theorem 3.2), the above function is uniquely maximized at (α_0, β_0). What is interesting is that, as shown in Li and Duan (1989), even when the function b is misspecified, the maximizer of the above function (with incorrectly specified b) still gives the correct direction of β_0, as long as X has linear conditional expectation.

Specifically, let us replace b by c, an arbitrary convex function. Equivalently, one can regard this replacement as replacing the true mean function μ by an arbitrary monotone function. That is, let

$$R(\alpha, \beta) = E[(\alpha + \beta^{\mathsf{T}} X)Y - c(\alpha + \beta^{\mathsf{T}} X)]. \tag{1.8}$$

The main result of Li and Duan (1989) is that, if (α_1, β_1) minimizes $R(\alpha, \beta)$, then,

Copyright © ${Date}. ${Publisher}. All rights reserved.

under the condition that $E(X|\beta_0^{\mathsf{T}}X)$ is a linear function of $\beta_0^{\mathsf{T}}X$, β_1 is proportional to β_0. Since $R_0(\alpha,\beta)$ is the expectation of the true log likelihood, (α_0,β_0) is the true parameter. Hence β_1 is proportional to the true parameter β_0 regardless of whether the link function is correctly specified, provided that X has a linear conditional mean given $\beta_0^{\mathsf{T}}X$. Since this assumption will appear frequently in this book, we give it a formal definition.

Assumption 1.1 *We say that X satisfies the linear conditional mean assumption with respect to β if $E(X|\beta^{\mathsf{T}}X)$ is a linear function of $\beta^{\mathsf{T}}X$.*

It can be shown that if this condition is satisfied for all $\beta \in \mathbb{R}^p$, then X has an elliptically contoured distribution, and vice versa. See Eaton (1986). The consequence of the linear conditional mean assumption is that $E(X|\beta^{\mathsf{T}}X)$ can be expressed as the projection in $\mathbb{R}^p(\Sigma)$.

Lemma 1.1 *Suppose $\beta^{\mathsf{T}}\Sigma\beta > 0$ and X satisfies Assumption 1.1, then*

$$E(X - EX|\beta^{\mathsf{T}}X) = P_\beta^{\mathsf{T}}(\Sigma)(X - EX).$$

PROOF. Denote EX by μ. Because $E(X|\beta^{\mathsf{T}}X)$ is linear in $\beta^{\mathsf{T}}X$, we have

$$E(X|\beta^{\mathsf{T}}X) = c + D\beta^{\mathsf{T}}X$$

where $c \in \mathbb{R}^p$ and $D \in \mathbb{R}^{p\times d}$. Take unconditional expectation on both sides to obtain

$$\mu = c + D\beta^{\mathsf{T}}\mu.$$

Substituting this into the previous line, we have $E(Z|\beta^{\mathsf{T}}Z) = D\beta^{\mathsf{T}}Z$. Multiply both sides of this equation from the right-hand side by $Z^{\mathsf{T}}\beta$, to obtain

$$E(Z|\beta^{\mathsf{T}}Z)Z^{\mathsf{T}}\beta = D\beta^{\mathsf{T}}ZZ^{\mathsf{T}}\beta.$$

Taking unconditional expectation on both sides, we have $\Sigma\beta = D\beta^{\mathsf{T}}\Sigma\beta$, which implies the desired relation. □

We now prove the result of Li and Duan (1989).

Proposition 1.2 *Suppose the following conditions hold.*

1. *The probability density function of $Y|X$ is proportional to*

$$\exp\{(\alpha_0 + \beta_0^{\mathsf{T}}x)y - b(\alpha_0 + \beta_0^{\mathsf{T}}x)\};$$

2. *Assumption 1.1 is satisfied for β_0;*
3. *$\beta^{\mathsf{T}}\Sigma\beta > 0$;*
4. *c is a strictly convex function; $(\alpha + \beta^{\mathsf{T}}X)Y - c(\alpha + \beta^{\mathsf{T}}X)$ is integrable for all $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$.*

If (α_1,β_1) is the maximizer of $R_1(\alpha,\beta)$, then $\beta_1 \propto \beta_0$.

Copyright © ${Date}. ${Publisher}. All rights reserved.

PROOF. Condition 1 implies

$$E(Y|X) = \mu(\alpha_0 + \beta_0^\top X) = \dot{b}(\alpha_0 + \beta_0^\top X).$$

This also implies $E(Y|X) = E(Y|\beta_0^\top X)$. By definition,

$$R(\alpha,\beta) = E[(\alpha + \beta^\top X)Y] - E[c(\alpha + \beta^\top X)].$$

The first term on the right-hand side is

$$\begin{aligned} E[(\alpha+\beta^\top X)Y] &= E[(\alpha+\beta^\top X)E(Y|X)] \\ &= E[(\alpha+\beta^\top X)E(Y|\beta_0^\top X)] \\ &= E[E(\alpha+\beta^\top X|\beta_0^\top X)Y] \\ &= E[a + b^\top E(X|\beta_0^\top X)Y]. \end{aligned}$$

By Lemma 1.1,

$$E(X|\beta_0^\top X) = P_{\beta_0}^\top(\Sigma)(X-\mu) + \mu.$$

Hence

$$\begin{aligned} E[(\alpha+\beta^\top X)Y] &= E[\alpha + \beta^\top P_{\beta_0}^\top(\Sigma)(X-\mu) + \beta^\top \mu)Y] \\ &= E[\alpha + \beta^\top P_{\beta_0}^\top(\Sigma)X - \beta^\top P_{\beta_0}^\top(\Sigma)\mu + \beta^\top \mu)Y] \\ &= E[\alpha + \beta^\top \mu - \beta^\top P_{\beta_0}^\top(\Sigma)\mu + \beta^\top P_{\beta_0}^\top(\Sigma)X)Y] \\ &= E[\gamma + \delta\beta_0^\top X)Y], \end{aligned}$$

where $\gamma = \alpha + \beta^\top \mu - \beta^\top P_{\beta_0}^\top(\Sigma)\mu$, $\delta = \beta^\top \Sigma\beta_0/(\beta_0^\top \Sigma\beta_0)$. Meanwhile, by Jensen's inequality,

$$\begin{aligned} E[c(\alpha+\beta^\top X)] &= E[E(c(\alpha+\beta^\top X)|\beta_0^\top X)] \\ &\geq E[c(E(\alpha+\beta^\top X)|\beta_0^\top X)] \\ &= E[c(\gamma + \delta\beta_0^\top X)]. \end{aligned}$$

So $R(\alpha,\beta) \leq R(\gamma,\delta\beta_0)$ for all $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^p$. Because c is strictly convex, R has a unique maximizer, leading to $\beta_1 = \delta\beta_0$. □

To gain more insight into this somewhat surprising result, in Figure 1.1 we plot a surface S, representing $E(Y|X) = E(Y|\beta^\top X)$, and a plain P, representing the least-squares estimate of the surface S. The surface varies only in one direction, which corresponds to the direction of β. We see that, as long as the distribution of X is symmetric about the direction in which the surface varies (as indicated by the dotted rectangle at the bottom), the gradient of the plain shares the same direction with the gradient of the surface. Thus, if we are interested in the direction, but not the magnitude, of the gradient, then we need not have an accurate estimate of the surface, because a simple model such as the linear model provides a reasonably good estimate of the gradient direction, even though it is a poor estimate of the surface.

Copyright © ${Date}. ${Publisher}. All rights reserved.

This result means that we can borrow the symmetry in the distribution of X to estimate the direction of the gradient using a rough estimate of the surface. This is significant because, especially in a high-dimensional setting, it is not easy to estimate a surface accurately as it involves high-dimensional smoothing, which causes what is known as "the curse of dimensionality". However, fitting a least-squares plane does not involve any smoothing, and thus it provides a reasonably good estimate of the gradient direction in the high-dimensional setting. The same can be said of any other simple parametric regression model, such as the quadratic regression model. This is one of the most important principles of Sufficient Dimension Reduction, and leads to many useful estimators described in the subsequent chapters.

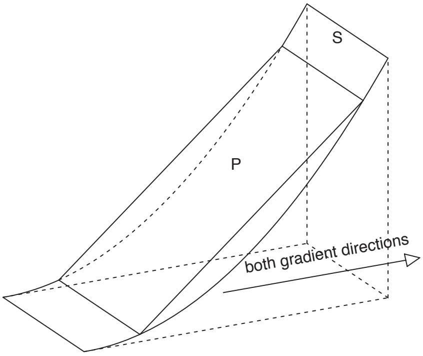

Figure 1.1 *Estimating gradient direction under link violation in Generalized Linear Model.*

Copyright © ${Date}. ${Publisher}. All rights reserved.