Springer Series in Statistics

Springer Series in Statistics

Alho/Spencer: Statistical Demography and Forecasting.
Andersen/Borgan/Gill/Keiding: Statistical Models Based on Counting Processes.
Atkinson/Riani: Robust Diagnostic Regression Analysis.
Atkinson/Riani/Cerioli: Exploring Multivariate Data with the Forward Search.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Borg/Groenen: Modern Multidimensional Scaling: Theory and Applications, 2nd edition.
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Bucklew: Introduction to Rare Event Simulation.
Cappé/Moulines/Rydén: Inference in Hidden Markov Models.
Chan/Tong: Chaos: A Statistical Perspective.
Chen/Shao/Ibrahim: Monte Carlo Methods in Bayesian Computation.
Coles: An Introduction to Statistical Modeling of Extreme Values.
David/Edwards: Annotated Readings in the History of Statistics.
Devroye/Lugosi: Combinatorial Methods in Density Estimation.
Efromovich: Nonparametric Curve Estimation: Methods, Theory, and Applications.
Eggermont/LaRiccia: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edition.
Fan/Yao: Nonlinear Time Series: Nonparametric and Parametric Methods.
Farebrother: Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume I: Two Crops.
Federer: Statistical Design and Analysis for Intercropping Experiments, Volume II: Three or More Crops.
Ferraty/Vieu: Nonparametric Functional Data Analysis: Models, Theory, Applications, and Implementation
Ghosh/Ramamoorthi: Bayesian Nonparametrics.
Glaz/Naus/Wallenstein: Scan Statistics.
Good: Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition.
Gouriéroux: ARCH Models and Financial Applications.
Gu: Smoothing Spline ANOVA Models.
Györfi/Kohler/Krzyżak/Walk: A Distribution-Free Theory of Nonparametric Regression.
Haberman: Advanced Statistics, Volume I: Description of Populations.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Harrell: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.
Hart: Nonparametric Smoothing and Lack-of-Fit Tests.
Hastie/Tibshirani/Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Hedayat/Sloane/Stufken: Orthogonal Arrays: Theory and Applications.
Heyde: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation.

(continued after index)

Anastasios A. Tsiatis

Semiparametric Theory and Missing Data

Springer

Anastasios A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA
tsiatis@stat.ncsu.edu

Library of Congress Control Number: 2006921164

ISBN-10: 0-387-32448-8
ISBN-13: 978-0387-32448-7

Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springer.com

To
My Mother, Anna
My Wife, Marie
and
My Son, Greg

Preface

Missing data are prevalent in many studies, especially when the studies involve human beings. Not accounting for missing data properly when analyzing data can lead to severe biases. For example, most software packages, by default, delete records for which any data are missing and conduct the so-called "complete-case analysis". In many instances, such an analysis will lead to an incorrect inference. Since the 1980s there has been a serious attempt to understand the underlying issues involved with missing data. In this book, we study the different mechanisms for missing data and some of the different analytic strategies that have been suggested in the literature for dealing with such problems. A special case of missing data includes censored data, which occur frequently in the area of survival analysis. Some discussion of how the missing-data methods that are developed will apply to problems with censored data is also included.

Underlying any missing-data problem is the statistical model for the data if none of the data were missing (i.e., the so-called full-data model). In this book, we take a very general approach to statistical modeling. That is, we consider statistical models where interest focuses on making inference on a finite set of parameters when the statistical model consists of the parameters of interest as well as other nuisance parameters. Unlike most traditional statistical models, where the nuisance parameters are finite-dimensional, we consider the more general problem of infinite-dimensional nuisance parameters. This allows us to develop theory for important statistical methods such as regression models that model the conditional mean of a response variable as a function of covariates without making any additional distributional assumptions on the variables and the proportional hazards regression model for survival data. Models where the parameters of interest are finite-dimensional and the nuisance parameters are infinite-dimensional are called semiparametric models.

The first five chapters of the book consider semiparametric models when there are no missing data. In these chapters, semiparametric models are defined and some of the theoretical developments for estimators of the parameters in these models are reviewed. The semiparametric theory and the properties of the estimators for parameters in semiparametric models are developed from a geometrical perspective. Consequently, in Chapter 2, a quick review of the geometry of Hilbert spaces is given. The geometric ideas are first developed for finite-dimensional parametric models in Chapter 3 and then extended to infinite-dimensional models in Chapters 4 and 5.

A rigorous treatment of semiparametric theory is given in the book *Efficient and Adaptive Estimation for Semiparametric Models* by Bickel et al. (1993). (Johns Hopkins University Press: Baltimore, MD). My experience has been that this book is too advanced for many students in statistics and biostatistics even at the Ph.D. level. The attempt here is to be more expository and heuristic, trying to give an intuition for the basic ideas without going into all the technical details. Although the treatment of this subject is not rigorous, it is not trivial either. Readers should not be frustrated if they don't grasp all the concepts at first reading. This first part of the book that deals only with semiparametric models (absent missing data) and the geometric theory of semiparametrics will be important in its own right. It is a beautiful theory, where the geometric perspective gives a new insight and deeper understanding of statistical models and the properties of estimators for parameters in such models.

The remainder of the book focuses on missing-data methods, building on the semiparametric techniques developed in the earlier chapters. In Chapter 6, a discussion and overview of missing-data mechanisms is given. This includes the definition and motivation for the three most common categories of missingness, namely

- missing completely at random (MCAR)
- missing at random (MAR)
- nonmissing at random (NMAR)

These ideas are extended to the broader class of coarsened data. We show how statistical models for full data can be integrated with missingness or coarsening mechanisms that allow us to derive likelihoods and models for the observed data in the presence of missingness. The geometric ideas for semiparametric full-data models are extended to missing-data models. This treatment will give the reader a deep understanding of the underlying theory for missing and coarsened data. Methods for estimating parameters with missing or coarsened data in as efficient a manner as possible are emphasized. This theory leads naturally to inverse probability weighted complete-case (IPWCC) and augmented inverse probability weighted complete-case (AIPWCC) estimators, which are discussed in great detail in Chapters 7 through 11. As we will see, some of the proposed methods can become computationally challenging if not infeasible. Therefore, in Chapter 12, we give some approximate methods for obtaining more efficient estimators with missing data that are easier to implement. Much of the theory developed in this book is taken from a series of ground-breaking papers by Robins and Rotnitzky (together with colleagues), who developed this elegant semiparametric theory for missing-data problems.

A short discussion on how missing-data semiparametric methods can be applied to estimate causal treatment effects in a point exposure study is given in Chapter 13 to illustrate the broad applicability of these methods. In Chapter 14, the final chapter, we deviate slightly from semiparametric models to discuss some of the theoretical properties of multiple-imputation estimators for finite-dimensional parametric models. However, even here, the theory developed throughout the book will be useful in understanding the properties of such estimators.

Anastasios (Butch) Tsiatis

Contents

Preface vii

1 Introduction to Semiparametric Models 1
1.1 What Is an Infinite-Dimensional Space? 2
1.2 Examples of Semiparametric Models 3
Example 1: Restricted Moment Models 3
Example 2: Proportional Hazards Model 7
Example 3: Nonparametric Model 8
1.3 Semiparametric Estimators 8

2 Hilbert Space for Random Vectors 11
2.1 The Space of Mean-Zero q-dimensional Random Functions 11
The Dimension of the Space of Mean-Zero Random Functions 12
2.2 Hilbert Space 13
2.3 Linear Subspace of a Hilbert Space and the Projection Theorem 14
Projection Theorem for Hilbert Spaces 14
2.4 Some Simple Examples of the Application of the Projection Theorem 15
Example 1: One-Dimensional Random Functions 15
Example 2: q-dimensional Random Functions 16
2.5 Exercises for Chapter 2 19

3 The Geometry of Influence Functions 21
3.1 Super-Efficiency 24
Example Due to Hodges 24
3.2 m-Estimators (Quick Review) 29
Estimating the Asymptotic Variance of an m-Estimator 31
Proof of Theorem 3.2 34

3.3 Geometry of Influence Functions for Parametric Models 38
Constructing Estimators 38
3.4 Efficient Influence Function 42
Asymptotic Variance when Dimension Is Greater than One . . . 43
Geometry of Influence Functions 45
Deriving the Efficient Influence Function 46
3.5 Review of Notation for Parametric Models 49
3.6 Exercises for Chapter 3 50

4 Semiparametric Models 53
4.1 GEE Estimators for the Restricted Moment Model 54
Asymptotic Properties for GEE Estimators 55
Example: Log-linear Model 57
4.2 Parametric Submodels 59
4.3 Influence Functions for Semiparametric
RAL Estimators 61
4.4 Semiparametric Nuisance Tangent Space 63
Tangent Space for Nonparametric Models 68
Partitioning the Hilbert Space 69
4.5 Semiparametric Restricted Moment Model 73
The Space Λ_{2s} 77
The Space Λ_{1s} 79
Influence Functions and the Efficient Influence Function for
the Restricted Moment Model 83
The Efficient Influence Function 85
A Different Representation for the Restricted
Moment Model 87
Existence of a Parametric Submodel for the Arbitrary
Restricted Moment Model 91
4.6 Adaptive Semiparametric Estimators for the Restricted
Moment Model 93
Extensions of the Restricted Moment Model 97
4.7 Exercises for Chapter 4 98

5 Other Examples of Semiparametric Models 101
5.1 Location-Shift Regression Model 101
The Nuisance Tangent Space and Its Orthogonal Complement
for the Location-Shift Regression Model 103
Semiparametric Estimators for β 106
Efficient Score for the Location-Shift Regression Model 107
Locally Efficient Adaptive Estimators 108
Remarks 113
5.2 Proportional Hazards Regression Model with
Censored Data 113
The Nuisance Tangent Space 117

The Space Λ_{2s} Associated with $\lambda_{C|X}(v|x)$ 117
The Space Λ_{1s} Associated with $\lambda(v)$ 119
Finding the Orthogonal Complement of the Nuisance Tangent Space 120
Finding RAL Estimators for β 123
Efficient Estimator 125
5.3 Estimating the Mean in a Nonparametric Model 125
5.4 Estimating Treatment Difference in a Randomized Pretest-Posttest Study or with Covariate Adjustment 126
The Tangent Space and Its Orthogonal Complement 129
5.5 Remarks about Auxiliary Variables 133
5.6 Exercises for Chapter 5 135

6 Models and Methods for Missing Data 137
6.1 Introduction 137
6.2 Likelihood Methods 143
6.3 Imputation 144
Remarks 145
6.4 Inverse Probability Weighted Complete-Case Estimator 146
6.5 Double Robust Estimator 147
6.6 Exercises for Chapter 6 150

7 Missing and Coarsening at Random for Semiparametric Models 151
7.1 Missing and Coarsened Data 151
Missing Data as a Special Case of Coarsening 153
Coarsened-Data Mechanisms 154
7.2 The Density and Likelihood of Coarsened Data 156
Discrete Data 156
Continuous Data 157
Likelihood when Data Are Coarsened at Random 158
Brief Remark on Likelihood Methods 160
Examples of Coarsened-Data Likelihoods 161
7.3 The Geometry of Semiparametric Coarsened-Data Models 163
The Nuisance Tangent Space Associated with the Full-Data Nuisance Parameter and Its Orthogonal Complement .. 166
7.4 Example: Restricted Moment Model with Missing Data by Design 174
The Logistic Regression Model 179
7.5 Recap and Review of Notation 181
7.6 Exercises for Chapter 7 183

8 The Nuisance Tangent Space and Its Orthogonal Complement 185
8.1 Models for Coarsening and Missingness 185
Two Levels of Missingness 185
Monotone and Nonmonotone Coarsening for more than Two Levels 186
8.2 Estimating the Parameters in the Coarsening Model 188
MLE for ψ with Two Levels of Missingness 188
MLE for ψ with Monotone Coarsening 189
8.3 The Nuisance Tangent Space when Coarsening Probabilities Are Modeled 190
8.4 The Space Orthogonal to the Nuisance Tangent Space 192
8.5 Observed-Data Influence Functions 193
8.6 Recap and Review of Notation 195
8.7 Exercises for Chapter 8 196

9 Augmented Inverse Probability Weighted Complete-Case Estimators 199
9.1 Deriving Semiparametric Estimators for β 199
Interesting Fact 206
Estimating the Asymptotic Variance 206
9.2 Additional Results Regarding Monotone Coarsening 207
The Augmentation Space Λ_2 with Monotone Coarsening 207
9.3 Censoring and Its Relationship to Monotone Coarsening 213
Probability of a Complete Case with Censored Data 216
The Augmentation Space, Λ_2, with Censored Data 216
Deriving Estimators with Censored Data 217
9.4 Recap and Review of Notation 218
9.5 Exercises for Chapter 9 220

10 Improving Efficiency and Double Robustness with Coarsened Data 221
10.1 Optimal Observed-Data Influence Function Associated with Full-Data Influence Function 221
10.2 Improving Efficiency with Two Levels of Missingness 225
Finding the Projection onto the Augmentation Space 226
Adaptive Estimation 227
Algorithm for Finding Improved Estimators with Two Levels of Missingness 229
Remarks Regarding Adaptive Estimators 230
Estimating the Asymptotic Variance 233
Double Robustness with Two Levels of Missingness 234

Remarks Regarding Double-Robust Estimators 236
Logistic Regression Example Revisited 236
10.3 Improving Efficiency with Monotone Coarsening 239
Finding the Projection onto the Augmentation Space 239
Adaptive Estimation 243
Double Robustness with Monotone Coarsening 248
Example with Longitudinal Data 251
10.4 Remarks Regarding Right Censoring 254
10.5 Improving Efficiency when Coarsening Is Nonmonotone 255
Finding the Projection onto the Augmentation Space 256
Uniqueness of $\mathcal{M}^{-1}(\cdot)$ 258
Obtaining Improved Estimators with Nonmonotone Coarsening 261
Double Robustness 265
10.6 Recap and Review of Notation 267
10.7 Exercises for Chapter 10 270

11 Locally Efficient Estimators for Coarsened-Data Semiparametric Models 273
Example: Estimating the Mean with Missing Data 275
11.1 The Observed-Data Efficient Score 277
Representation 1 (Likelihood-Based) 277
Representation 2 (AIPWCC-Based) 278
Relationship between the Two Representations 278
$\mathcal{M}^{-1}$ for Monotone Coarsening 282
$\mathcal{M}^{-1}$ with Right Censored Data 284
11.2 Strategy for Obtaining Improved Estimators 285
Example: Restricted Moment Model with Monotone Coarsening 286
Some Brief Remarks Regarding Robustness 290
11.3 Concluding Thoughts 291
11.4 Recap and Review of Notation 292
11.5 Exercises for Chapter 11 293

12 Approximate Methods for Gaining Efficiency 295
12.1 Restricted Class of AIPWCC Estimators 295
12.2 Optimal Restricted (Class 1) Estimators 300
Deriving the Optimal Restricted (Class 1) AIPWCC Estimator 305
Estimating the Asymptotic Variance 307
12.3 Example of an Optimal Restricted (Class 1) Estimator 309
Modeling the Missingness Probabilities 312
12.4 Optimal Restricted (Class 2) Estimators 313

Logistic Regression Example Revisited 319
12.5 Recap and Review of Notation 321
12.6 Exercises for Chapter 12 322

13 Double-Robust Estimator of the Average Causal Treatment Effect 323
13.1 Point Exposure Studies 323
13.2 Randomization and Causality 326
13.3 Observational Studies 327
13.4 Estimating the Average Causal Treatment Effect 328
Regression Modeling 328
13.5 Coarsened-Data Semiparametric Estimators 329
Observed-Data Influence Functions 331
Double Robustness 336
13.6 Exercises for Chapter 13 337

14 Multiple Imputation: A Frequentist Perspective 339
14.1 Full- Versus Observed-Data Information Matrix 342
14.2 Multiple Imputation 344
14.3 Asymptotic Properties of the Multiple-Imputation Estimator 346
Stochastic Equicontinuity 352
14.4 Asymptotic Distribution of the Multiple-Imputation Estimator 354
14.5 Estimating the Asymptotic Variance 362
Consistent Estimator for the Asymptotic Variance 365
14.6 Proper Imputation 366
Asymptotic Distribution of $n^{1/2}(\hat{\beta}^*_n - \beta_0)$ 367
Rubin's Estimator for the Asymptotic Variance 370
Summary 371
14.7 Surrogate Marker Problem Revisited 371
How Do We Sample? 373

References 375

Index 381

1

Introduction to Semiparametric Models

Statistical problems are described using probability models. That is, data are envisioned as realizations of a vector of random variables $Z_1, \ldots, Z_n$, where Z_i itself is a vector of random variables corresponding to the data collected on the i-th individual in a sample of n individuals chosen from some population of interest. We will assume throughout the book that $Z_1, \ldots, Z_n$ are identically and independently distributed (iid) with density belonging to some probability (or statistical model), where a model consists of a class of densities that we believe might have generated the data. The densities in a model are often identified through a set of parameters; i.e., a real-valued vector used to describe the densities in a statistical model. The problem is usually set up in such a way that the value of the parameters or, at the least, the value of some subset of the parameters that describes the density that generates the data, is of importance to the investigator. Much of statistical inference considers how we can learn about this “true” parameter value from a sample of observed data. Models that are described through a vector of a finite number of real values are referred to as finite-dimensional parametric models. For finite-dimensional parametric models, the class of densities can be described as

$$\mathcal{P} = \{p(z, \theta), \theta \in \Omega \subset \mathbb{R}^p\},$$

where the dimension p is some finite positive integer.

For many problems, we are interested in making inference only on a subset of the parameters. Nonetheless, the entire set of parameters is necessary to properly describe the class of possible distributions that may have generated the data. Suppose, for example, we are interested in estimating the mean response of a variable, which we believe follows a normal distribution. Typically, we conduct an experiment where we sample from that distribution and describe the data that result from that experiment as a realization of the random vector

$$Z_1, \ldots, Z_n \text{ assumed iid } N(\mu, \sigma^2);\ \mu \in \mathbb{R}, \sigma^2 > 0; \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ \subset \mathbb{R}^2.$$

Here we are interested in estimating μ, the mean of the distribution, but σ^2, the variance of the distribution, is necessary to properly describe the possible probability distributions that might have generated the data. It is useful to write the parameter θ as $(\beta^T, \eta^T)^T$, where $\beta^{q\times 1}$ (a q-dimensional vector) is the parameter of interest and $\eta^{r\times 1}$ (an r-dimensional vector) is the nuisance parameter. In the previous example, $\beta = \mu$ and $\eta = \sigma^2$. The entire parameter space Ω has dimension $p = q + r$.

In some cases, we may want to consider models where the class of densities is so large that the parameter θ is infinite-dimensional. Examples of this will be given shortly. For such models, we will consider the problem where we are interested in estimating β, which we still take to be finite-dimensional, say q-dimensional. For some problems, it will be natural to partition the parameter θ as (β, η), where β is the q-dimensional parameter of interest and η is the nuisance parameter, which is infinite-dimensional. In other cases, it is more natural to consider the parameter β as the function $\beta(\theta)$. These models are referred to as *semiparametric models* in the literature because, generally, there is both a parametric component β and a nonparametric component η that describe the model. By allowing the space of parameters to be infinite-dimensional, we are putting less restrictions on the probabilistic constraints that our data might have (compared with finite-dimensional parametric models). Therefore, solutions, if they exist and are reasonable, will have greater applicability and robustness.

Because the notion of an infinite-dimensional parameter space is so important in the subsequent development of this book, we start with a short discussion of infinite-dimensional spaces.

1.1 What Is an Infinite-Dimensional Space?

The parameter spaces that we will consider in this book will always be subsets of linear vector spaces. That is, we will consider a parameter space $\Omega \subset \mathcal{S}$, where $\mathcal{S}$ is a linear space. A space $\mathcal{S}$ is a linear space if, for θ_1 and θ_2 elements of $\mathcal{S}$, $a\theta_1 + b\theta_2$ will also be an element of $\mathcal{S}$ for any scalar constants a and b. Such a linear space is finite-dimensional if it can be spanned by a finite number of elements in the space. That is, $\mathcal{S}$ is a finite-dimensional linear space if elements $\theta_1, \ldots, \theta_m$ exist, where m is some finite positive integer such that any element $\theta \in \mathcal{S}$ is equal to some linear combination of $\theta_1, \ldots, \theta_m$; i.e., $\theta = a_1\theta_1 + \ldots + a_m\theta_m$ for some scalar constants $a_1, \ldots, a_m$. The dimension of a finite-dimensional linear space is defined by the minimum number of elements in the space that span the entire space or, equivalently, the number of linearly independent elements that span the entire space, where a set of elements are linearly independent if no element in the set can be written as a linear combination of the other elements. Parameter spaces that are defined in p-dimensional Euclidean spaces are clearly finite-dimensional spaces. A linear space $\mathcal{S}$ that cannot be spanned by any finite set of elements is called an infinite-dimensional parameter space.

An example of an infinite-dimensional linear space is the space of continuous functions defined on the real line. Consider the space $\mathcal{S} = \{f(x), x \in \mathbb{R}\}$ for all continuous functions $f(\cdot)$. Clearly $\mathcal{S}$ is a linear space. In order to show that this space is infinite-dimensional, we must demonstrate that it cannot be spanned by any finite set of elements in $\mathcal{S}$. This can be accomplished by noting that the space $\mathcal{S}$ contains the linear subspaces made up of the class of polynomials of order m; that is, the space $\mathcal{S}_m = \{f(x) = \sum_{j=0}^{m} a_j x^j\}$ for all constants $a_0, \ldots, a_m$. Clearly, the space $\mathcal{S}_m$ is finite-dimensional (i.e., spanned by the elements $x^0, x^1, \ldots, x^m$). In fact, this space is exactly an $m+1$-dimensional linear space since the elements $x^0, \ldots, x^m$ are linearly independent.

Linear independence follows because x^j cannot be written as a linear combination of $x^0, \ldots, x^{j-1}$ for any $j = 1, 2, \ldots$. If it could, then

$$x^j = \sum_{\ell=0}^{j-1} a_\ell x^\ell \text{ for all } x \in \mathbb{R}$$

for some constants $a_0, \ldots, a_{j-1}$. If this were the case, then the derivatives of x^j of all orders would have to be equal to the corresponding derivatives of $\sum_{\ell=0}^{j-1} a_\ell x^\ell$. But the j-th derivative of x^j is equal to $j! \neq 0$, whereas the j-th derivative of $\sum_{\ell=0}^{j-1} a_\ell x^\ell$ is zero, leading to a contradiction and implying that $x^0, \ldots, x^m$ are linearly independent.

Consequently, the space $\mathcal{S}$ cannot be spanned by any finite number, say m elements of $\mathcal{S}$, because, if this were possible, then the space of polynomials of order greater than m could also be spanned by the m elements. But this is impossible since such spaces of polynomials have dimension greater than m. Hence, $\mathcal{S}$ is infinite-dimensional.

From the arguments above, we can easily show that the space of arbitrary densities $p_Z(z)$ for a continuous random variable Z defined on the closed finite interval $[0,1]$ (i.e., the so-called nonparametric model for such a random variable) spans a space that is infinite-dimensional. This follows by noticing that the functions $p_{Zj}(z) = (j+1)^{-1} z^j,\ 0 \leq z \leq 1,\ j = 1, \ldots$ are densities that are linearly independent.

1.2 Examples of Semiparametric Models

Example 1: Restricted Moment Models

A common statistical problem is to model the relationship of a response variable Y (possibly vector-valued) as a function of a vector of covariates X. Throughout, we will use the convention that a vector of random variables Z that is not indexed will correspond to a single observation, whereas $Z_i, i = 1, \ldots, n$ will denote a sample of n iid random vectors. Consider a family of probability distributions for $Z = (Y, X)$ that satisfy the regression relationship

$$E(Y|X) = \mu(X, \beta),$$

where $\mu(X, \beta)$ is a known function of X and the unknown q-dimensional parameter β.

The function $\mu(X, \beta)$ may be linear or nonlinear in β, and it is assumed that β is finite-dimensional. For example, we might consider a linear model where $\mu(X, \beta) = \beta^T X$ or a nonlinear model, such as a log-linear model, where $\mu(X, \beta) = \exp(\beta^T X)$. No other assumptions will be made on the class of probability distributions other than the constraint given by the conditional expectation of Y given X stated above. As we will demonstrate shortly, such models are semiparametric, as they will be defined through an infinite-dimensional parameter space. We will refer to such semiparametric models as "restricted moment models." These models were studied by Chamberlain (1987) and Newey (1988) in the econometrics literature. They were also popularized in the statistics literature by Liang and Zeger (1986).

For illustration, we will take Y to be a one-dimensional random variable that is continuous on the real line. This model can also be written as

$$Y = \mu(X, \beta) + \varepsilon,$$

where $E(\varepsilon|X) = 0$. The data are realizations of $(Y_1, X_1), \ldots, (Y_n, X_n)$ that are iid with density for a single observation given by

$$p_{Y,X}\{y, x; \beta, \eta(\cdot)\},$$

where $\eta(\cdot)$ denotes the infinite-dimensional nuisance parameter function characterizing the joint distribution of ε and X, to be defined shortly. Knowledge of β and the joint distribution of (ε, X) will induce the joint distribution of (Y, X). Since

$$\varepsilon = Y - \mu(X, \beta),$$
$$p_{Y,X}(y, x) = p_{\varepsilon,X}\{y - \mu(x, \beta), x\}.$$

The restricted moment model only makes the assumption that

$$E(\varepsilon|X) = 0.$$

That is, we will allow any joint density $p_{\varepsilon,X}(\varepsilon, x) = p_{\varepsilon|X}(\varepsilon|x) p_X(x)$ such that

$$p_{\varepsilon|X}(\varepsilon|x) \geqslant 0 \text{ for all } \varepsilon, x,$$
$$\int p_{\varepsilon|X}(\varepsilon|x) d\varepsilon = 1 \text{ for all } x,$$
$$\int \varepsilon p_{\varepsilon|X}(\varepsilon|x) d\varepsilon = 0 \text{ for all } x,$$
$$p_X(x) \geq 0 \text{ for all } x,$$
$$\int p_X(x) d\nu_X(x) = 1.$$

Remark 1. When we refer to the density, joint density, or conditional density of one or more random variables, to avoid confusion, we will often index the variables being used as part of the notation. For example, $p_{Y,X}(y, x)$ is the joint density of the random variables Y and X evaluated at the values (y, x). This notation will be suppressed when the variables are obvious. □

Remark 2. We will use the convention that random variables are denoted by capital letters such as Y and X, whereas realizations of those random variables will be denoted by lowercase letters such as y and x. One exception to this is that the random variable corresponding to the error term $Y - \mu(X, \beta)$ is denoted by the Greek lowercase ε. This is in keeping with the usual notation for such error terms used in statistics. The realization of this error term will also be denoted by the Greek lowercase ε. The distinction between the random variable and the realization of the error term will have to be made in the context it is used and should be obvious in most cases. For example, when we refer to $p_{\varepsilon,X}(\varepsilon, x)$, the subscript ε is a random variable and the argument ε inside the parentheses is the realization. □

Remark 3. $\nu_X(x)$ is a dominating measure for which densities for the random vector X are defined. For the most part, we will consider $\nu(\cdot)$ to be the Lebesgue measure for continuous random variables and the counting measure for discrete random variables. The random variable Y and hence ε will be taken to be continuous random variables dominated by Lebesgue measure dy or $d\varepsilon$, respectively. □

Without going into the measure-theoretical technical details, the class of conditional densities for ε given X, such that $E(\varepsilon|X) = 0$, can be constructed through the following steps.

(a) Choose any arbitrary positive function of ε and x (subject to regularity conditions):

$$h^{(0)}(\varepsilon, x) > 0.$$

(b) Normalize this function to be a conditional density:

$$h^{(1)}(\varepsilon, x) = \frac{h^{(0)}(\varepsilon, x)}{\int h^{(0)}(\varepsilon, x) d\varepsilon};$$

i.e.,

$$\int h^{(1)}(\varepsilon, x) d\varepsilon = 1 \text{ for all } x.$$

(c) Center it:
A random variable ε^* whose conditional density, given $X = x$ is
$h^{(1)}(\varepsilon', x) = p(\varepsilon^* = \varepsilon' | X = x)$, has mean

$$\mu(x) = \int \varepsilon' h^{(1)}(\varepsilon', x) d\varepsilon'.$$

In order to construct a random variable ε whose conditional density, given $X = x$, has mean zero, we consider $\varepsilon = \varepsilon^* - \mu(X)$ or $\varepsilon^* = \varepsilon + \mu(X)$. It is clear that $E(\varepsilon|X = x) = E(\varepsilon^*|X = x) - \mu(x) = 0$. Since the transformation from ε to ε^*, given $X = x$, has Jacobian equal to 1, the conditional density of ε given X, defined by $\eta_1(\varepsilon, x)$, is given by

$$\eta_1(\varepsilon, x) = h^{(1)}\left\{\varepsilon + \int \varepsilon h^{(1)}(\varepsilon, x) d\varepsilon, x\right\},$$

which, by construction, satisfies $\int \varepsilon \eta_1(\varepsilon, x) d\varepsilon = 0$ for all x.

Thus, any function $\eta_1(\varepsilon, x)$ constructed as above will satisfy $\eta_1(\varepsilon, x) > 0$,

$$\int \eta_1(\varepsilon, x) d\varepsilon = 1 \text{ for all } x,$$
$$\int \varepsilon \eta_1(\varepsilon, x) d\varepsilon = 0 \text{ for all } x.$$

Since the class of all such conditional densities $\eta_1(\varepsilon, x)$ was derived from arbitrary positive functions $h^{(0)}(\varepsilon, x)$ (subject to regularity conditions), and since the space of positive functions is infinite-dimensional, then the set of such resulting conditional densities is also infinite-dimensional.

Similarly, we can construct densities for X where $p_X(x) = \eta_2(x)$ such that

$$\eta_2(x) > 0,$$
$$\int \eta_2(x) d\nu_X(x) = 1.$$

The set of all such functions $\eta_2(x)$ will also be infinite-dimensional as long as the support of X is infinite.

Therefore, the restricted moment model is characterized by

$$\{\beta, \eta_1(\varepsilon, x), \eta_2(x)\},$$

where $\beta \in \mathbb{R}^q$ is finite-dimensional and $\eta_1(\varepsilon, x) = p_{\varepsilon|X}(\varepsilon|x)$, $\eta_2(x) = p_X(x)$ are infinite-dimensional. Consequently, the joint density of (Y, X) is given by

$$\begin{aligned} p_{Y,X}\{y, x; \beta, \eta_1(\cdot), \eta_2(\cdot)\} &= p_{Y|X}\{y|x; \beta, \eta_1(\cdot)\} p_X\{x; \eta_2(\cdot)\} \\ &= \eta_1\{y - \mu(x, \beta), x\}\eta_2(x). \end{aligned}$$

This is an example of a semiparametric model because the parametrization is through a finite-dimensional parameter of interest $\beta \in \mathbb{R}^q$ and infinite-dimensional nuisance parameters $\{\eta_1(\cdot), \eta_2(\cdot)\}$.

Contrast this semiparametric model with the more traditional parametric model where

$$Y_i = \mu(X_i, \beta) + \varepsilon_i, i = 1, \ldots, n,$$

where ε_i are iid $N(0, \sigma^2)$. That is,

$$p_{Y|X}(y|x; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\frac{\{y-\mu(x,\beta)\}^2}{\sigma^2}\right].$$

This model is much more restrictive than the semiparametric model defined earlier.

Example 2: Proportional Hazards Model

In many biomedical applications, we are often interested in modeling the survival time of individuals as functions of covariates. Let the response variable be the survival time of an individual, denoted by T, whose distribution depends on explanatory variables X. A popular model in survival analysis is Cox's proportional hazards model, which was first introduced in the seminal paper by Cox (1972). This model assumes that the conditional hazard rate, as a function of X, is given by

$$\begin{aligned}\lambda(t|X) &= \lim_{h \to 0} \left\{ \frac{P(t \leq T < t+h | T \geqslant t, X)}{h} \right\} \\ &= \lambda(t) \exp(\beta^T X).\end{aligned}$$

The proportional hazards model is especially convenient when survival times may be right censored, as we will discuss in greater detail in Chapter 5.

Interest often focuses on the finite-dimensional parameter β, as this describes the magnitude of the effect that the covariates have on the survival time. The underlying hazard function $\lambda(t)$ is left unspecified and is considered a nuisance parameter. Since this function can be any positive function in t, subject to some regularity conditions, it, too, is infinite-dimensional. Using the fact that the density of a positive random variable is related to the hazard function through

$$p_T(t) = \lambda(t) \exp\left\{-\int_0^t \lambda(u) du\right\},$$

then the density of a single observation $Z = (T, X)$ is given by

$$p_{T,X}\{t, x; \beta, \lambda(\cdot), \eta_2(\cdot)\} = p_{T|X}\{t|x; \beta, \lambda(\cdot)\}\eta_2(x),$$

where

$$p_{T|X}\{t|x; \beta, \lambda(\cdot)\} = \lambda(t) \exp(\beta^T x) \exp\left\{-\exp(\beta^T x) \int_0^t \lambda(u) du\right\},$$

and exactly as in Example 1, $\eta_2(x)$ is defined as a function of x such that

$$\begin{aligned} \eta_2(x) &\geqslant 0, \\ \int \eta_2(x) d\nu_X(x) &= 1, \end{aligned}$$

for all x. The proportional hazards model has gained a great deal of popularity because it is more flexible than a finite-dimensional parametric model, that assumes that the hazard function for T has a specific functional form in terms of a few parameters; e.g.,

$$\lambda(t, \eta) = \eta \text{ (constant hazard over time – exponential model),}$$

or

$$\lambda(t, \eta) = \eta_1 t^{\eta_2} \text{ (Weibull model).}$$

Example 3: Nonparametric Model

In the two previous examples, the probability models were written in terms of an infinite-dimensional parameter θ, which was partitioned as $\{\beta^T, \eta(\cdot)\}$, where β was the finite-dimensional parameter of interest and $\eta(\cdot)$ was the infinite-dimensional nuisance parameter. We now consider the problem of estimating the moments of a single random variable Z where we put no restriction on the distribution of Z except that the moments of interest exist. That is, we denote the density of Z by $\theta(z)$, where $\theta(z)$ can be any positive function of z such that $\int \theta(z) d\nu_Z(z) = 1$ and any additional restrictions necessary for the moments of interest to exist. Clearly, the class of all $\theta(\cdot)$ is infinite-dimensional as long as the support of Z is infinite. Suppose we were interested in estimating some functional of $\theta(\cdot)$, say $\beta(\theta)$ (for example, the first or second moment $E(Z)$ or $E(Z^2)$, where $\beta(\theta)$ is equal to $\int z\theta(z) d\nu_Z(z)$ or $\int z^2\theta(z) d\nu_Z(z)$, respectively). For such a problem, it is not convenient to try to partition the parameter space in terms of the parameter β of interest and a nuisance parameter but rather to work directly with the functional $\beta(\theta)$.

1.3 Semiparametric Estimators

In a semiparametric model, a semiparametric estimator for β, say $\hat{\beta}_n$, is one that, loosely speaking, has the property that it is consistent and asymptotically normal in the sense that

$$\begin{aligned} (\hat{\beta}_n - \beta) &\xrightarrow{P\{\beta, \eta(\cdot)\}} 0, \\ n^{1/2}(\hat{\beta}_n - \beta) &\xrightarrow{\mathcal{D}\{\beta, \eta(\cdot)\}} N(0, \Sigma^{q \times q}\{\beta, \eta(\cdot)\}), \end{aligned}$$

for all densities "$p\{z, \beta, \eta(\cdot)\}$" within some semiparametric family, where $\xrightarrow{P\{\beta,\eta(\cdot)\}}$ denotes convergence in probability and $\xrightarrow{\mathcal{D}\{\beta,\eta(\cdot)\}}$ denotes convergence in distribution when the density of the random variable Z is $p\{z, \beta, \eta(\cdot)\}$.

We know, for example, that the solution to the linear estimating equations

$$\sum_{i=1}^{n} A^{q\times 1}(X_i, \hat{\beta}_n) \left\{ Y_i - \mu(X_i, \hat{\beta}_n) \right\} = 0^{q\times 1},$$

under suitable regularity conditions, leads to an estimator for β that is consistent and asymptotically normal for the semiparametric restricted moment model of Example 1. In fact, this is the basis for "generalized estimating equations" (GEE) proposed by Liang and Zeger (1986).

The maximum partial likelihood estimator proposed by Cox (1972, 1975) is an example of a semiparametric estimator for β in the proportional hazards model given in Example 2. Also, in Example 3, a semiparametric estimator for the first and second moments is given by $n^{-1}\sum Z_i$ and $n^{-1}\sum Z_i^2$, respectively.

Some issues that arise when studying semiparametric models are:

(i) How do we find semiparametric estimators, or do they even exist?
(ii) How do we find the best estimator among the class of semiparametric estimators?

Both of these problems are difficult. Understanding the geometry of estimators, more specifically the geometry of the *influence function* of estimators, will help us in this regard.

Much of this book will rely heavily on geometric constructions. We will define the influence function of an asymptotically linear estimator and describe the geometry of all possible influence functions for a statistical model. We will start by looking at finite-dimensional parametric models and then generalize the results to the more complicated infinite-dimensional semiparametric models.

Since the geometry that is considered is the geometry of Hilbert spaces, we begin with a quick review of Hilbert spaces, the notion of orthogonality, minimum distance, and how this relates to efficient estimators (i.e., estimators with the smallest asymptotic variance).

2

Hilbert Space for Random Vectors

In this section, we will introduce a Hilbert space without going into much of the technical details. We will focus primarily on the Hilbert space whose elements are random vectors with mean zero and finite variance that will be used throughout the book. For more details about Hilbert spaces, we recommend that the reader study Chapter 3 of Luenberger (1969).

2.1 The Space of Mean-Zero q-dimensional Random Functions

As stated earlier, data are envisioned as realizations of the random vectors $Z_1, Z_2, \ldots, Z_n$, assumed iid. Let Z denote the random vector for a single observation. As always, there is an underlying probability space $(\mathscr{Z}, \mathcal{A}, P)$, where $\mathscr{Z}$ denotes the sample space, $\mathcal{A}$ the corresponding σ-algebra, and P the probability measure. For the time being, we will not consider a statistical model consisting of a family of probability measures, but rather we will assume that P is the true probability measure that generates the realizations of Z.

Consider the space consisting of q-dimensional mean-zero random functions of Z,

$$h : \mathscr{Z} \to \mathbb{R}^q,$$

where $h(Z)$ is measurable and also satisfies

$$\begin{aligned} &\text{(i) } E\{h(Z)\} = 0, \\ &\text{(ii) } E\{h^T(Z)h(Z)\} < \infty. \end{aligned}$$

Since the elements of this space are random functions, when we refer to an element as h, we implicitly mean $h(Z)$. Clearly, the space of all such h that satisfy (i) and (ii) is a linear space. By linear, we mean that if h_1, h_2 are elements of the space, then for any real constants a and b, $ah_1 + bh_2$ also belongs to the space.

In the same way that we consider points in Euclidean space as vectors from the origin, here we will consider the q-dimensional random functions as points in a space. The intuition we have developed in understanding the geometry of two- and three-dimensional Euclidean space will aid us in understanding the geometry of more complex spaces through analogy. The random function

$$h(Z) = 0^{q \times 1}$$

will denote the origin of this space.

The Dimension of the Space of Mean-Zero Random Functions

An element of the linear space defined above is a q-dimensional function of Z. This should not be confused with the dimensionality of the space itself. To illustrate this point more clearly, let us first consider the space of one-dimensional random functions of Z (random variables), where Z is a discrete variable with finite support. Specifically, let Z be allowed to take on one of a finite number of values $z_1, \ldots, z_k$ with positive probabilities $\pi_1, \ldots, \pi_k$, where $\sum_{i=1}^{k} \pi_i = 1$. For such a case, any one-dimensional random function of Z can be defined as $h(Z) = a_1 I(Z = z_1) + \ldots + a_k I(Z = z_k)$ for any real valued constants $a_1, \ldots, a_k$, where $I(\cdot)$ denotes the indicator function. The space of all such random functions is a linear space spanned by the k linearly independent functions $I(Z = z_i), i = 1, \ldots, k$. Hence this space is a k-dimensional linear space. If we put the further constraint that the mean must be zero (i.e., $E\{h(Z)\} = 0$), then this implies that $\sum_{i=1}^{k} a_i \pi_i = 0$, or equivalently that $a_k = -(\sum_{i=1}^{k-1} a_i \pi_i)/\pi_k$. Some simple algebra leads us to conclude that the space of one-dimensional mean-zero random functions of Z is a linear space spanned by the $k-1$ linearly independent functions $\{I(Z = z_i) - \frac{\pi_i}{\pi_k} I(Z = z_k)\}, i = 1, \ldots, k-1$. Hence this space is a $k-1$-dimensional linear space.

Similarly, the space of q-dimensional mean-zero random functions of Z, where Z has finite support at the k values $z_1, \ldots, z_k$, can be shown to be a linear space with dimension $q \times (k-1)$. Clearly, as the number of support points k for the distribution of Z increases, so does the dimension of the linear space of q-dimensional mean-zero random functions of Z.

If the support of the random vector Z is infinite, as would be the case if any element of the random vector Z was a continuous random variable, then the space of measurable functions that make up the Hilbert space will be infinite-dimensional. As we indicated in Section 1.1, the set of one-dimensional continuous functions of Z is infinite-dimensional. Consequently, the set of q-dimensional continuous functions will also be infinite-dimensional. Clearly, the set of q-dimensional measurable functions is a larger class and hence must also be infinite-dimensional.

2.2 Hilbert Space

A Hilbert space, denoted by $\mathcal{H}$, is a complete normed linear vector space equipped with an inner product. As well as being a linear space, a Hilbert space also allows us to consider distance between elements and angles and orthogonality between vectors in the space. This is accomplished by defining an inner product.

Definition 1. Corresponding to each pair of elements h_1, h_2 belonging to a linear vector space $\mathcal{H}$, an inner product, defined by $\langle h_1, h_2 \rangle$, is a function that maps to the real line. That is, $\langle h_1, h_2 \rangle$ is a scalar that satisfies

1. $\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$,
2. $\langle h_1 + h_2, h_3 \rangle = \langle h_1, h_3 \rangle + \langle h_2, h_3 \rangle$, where h_1, h_2, h_3 belong to $\mathcal{H}$,
3. $\langle \lambda h_1, h_2 \rangle = \lambda \langle h_1, h_2 \rangle$ for any scalar constant λ,
4. $\langle h_1, h_1 \rangle \geq 0$ with equality if and only if $h_1 = 0$.

Note 1. In some cases, the function $\langle \cdot, \cdot \rangle$ may satisfy conditions 1–3 above and the first part of condition 4, but $\langle h_1, h_1 \rangle = 0$ may not imply that $h_1 = 0$. In that case, we can still define a Hilbert space by identifying equivalence classes where individual elements in our space correspond to different equivalence classes.

Definition 2. For the linear vector space of q-dimensional measurable random functions with mean zero and finite second moments, we can define the inner product

$$\langle h_1, h_2 \rangle \quad \text{by} \quad E(h_1^T h_2).$$

We shall refer to this inner product as the "covariance inner product."

This definition of inner product clearly satisfies the first three conditions of the definition given above. As for condition 4, we can define an equivalence class where h_1 is equivalent to h_2,

$$h_1 \equiv h_2,$$

if $h_1 = h_2$ a.e. or $P(h_1 \neq h_2) = 0$. In this book, we will generally not concern ourselves with such measure-theoretical subtleties.

Once an inner product is defined, we then define the norm or "length" of any vector (i.e., element of $\mathcal{H}$) (distance from any point $h \in \mathcal{H}$ to the origin) as

$$\|h\| = \langle h, h \rangle^{1/2}.$$

Hilbert spaces also allow us to define orthogonality; that is, $h_1, h_2 \in \mathcal{H}$ are orthogonal if $\langle h_1, h_2 \rangle = 0$.

Remark 1. Technically speaking, the definitions above are those for a pre-Hilbert space. In order to be a Hilbert space, we also need the space to be complete (i.e., every Cauchy sequence has a limit point that belongs to the space). That the space of q-dimensional random functions with mean zero and bounded second moments is complete follows from the L_2-completeness theorem (see Loève 1963, p. 161) and hence is a Hilbert space. □

2.3 Linear Subspace of a Hilbert Space and the Projection Theorem

A space $\mathcal{U} \subset \mathcal{H}$ is a linear subspace if $u_1, u_2 \in \mathcal{U}$ implies that $au_1 + bu_2 \in \mathcal{U}$ for all scalar constants a, b. A linear subspace must contain the origin. This is clear by letting the scalars be $a = b = 0$.

A simple example of a linear subspace is obtained by taking $h_1, \ldots, h_k$ to be arbitrary elements of a Hilbert space. Then the space $a_1h_1 + \cdots + a_kh_k$ for all scalars $(a_1, \ldots, a_k) \in \mathbb{R}^k$ is a linear subspace spanned by $\{h_1, \ldots, h_k\}$.

One of the key results for Hilbert spaces, which we will use repeatedly throughout this book, is given by the projection theorem.

Projection Theorem for Hilbert Spaces

Theorem 2.1. Let $\mathcal{H}$ be a Hilbert space and $\mathcal{U}$ a linear subspace that is closed (i.e., contains all its limit points). Corresponding to any $h \in \mathcal{H}$, there exists a unique $u_0 \in \mathcal{U}$ that is closest to h; that is,

$$\|h - u_0\| \leq \|h - u\| \quad \text{for all} \quad u \in \mathcal{U}.$$

Furthermore, $h - u_0$ is orthogonal to $\mathcal{U}$; that is,

$$\langle h - u_0, u \rangle = 0 \quad \text{for all} \quad u \in \mathcal{U}.$$

We refer to u_0 as the projection of h onto the space $\mathcal{U}$, and this is denoted as $\Pi(h|\mathcal{U})$. Moreover, u_0 is the only element $u \in \mathcal{U}$ such that $h - u$ is orthogonal to $\mathcal{U}$ (see Figure 2.1).

The proof of the projection theorem for arbitrary Hilbert spaces is not much different or more difficult than for a finite-dimensional Euclidean space. The condition that a Hilbert space be complete is necessary to guarantee the existence of the projection. A formal proof can be found in Luenberger (1969, Theorem 2, p. 51). The intuition of orthogonality and distance carries over very nicely from simple Euclidean spaces to more complex Hilbert spaces.

A simple consequence of orthogonality is the Pythagorean theorem, which we state for completeness.

Theorem 2.2. *Pythagorean Theorem*
If h_1 and h_2 are orthogonal elements of the Hilbert space $\mathcal{H}$ (i.e., $\langle h_1, h_2 \rangle = 0$), then

$$\|h_1 + h_2\|^2 = \|h_1\|^2 + \|h_2\|^2.$$

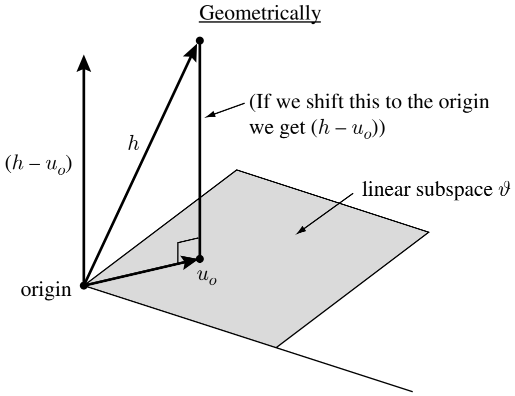

Fig. 2.1. Projection onto a linear subspace

2.4 Some Simple Examples of the Application of the Projection Theorem

Example 1: One-Dimensional Random Functions

Consider the Hilbert space $\mathcal{H}$ of one-dimensional random functions, $h(Z)$, with mean zero and finite variance equipped with the inner product

$$\langle h_1, h_2 \rangle = E(h_1 h_2)$$

for $h_1(Z), h_2(Z) \in \mathcal{H}$. Let $u_1(Z), \ldots, u_k(Z)$ be arbitrary elements of this space and $\mathcal{U}$ be the linear subspace spanned by $\{u_1, \cdots, u_k\}$. That is,

$$\mathcal{U} = \{a^T u; \quad \text{for} \quad a \in \mathbb{R}^k\},$$

where

$$u^{k \times 1} = \begin{pmatrix} u_1 \\ \vdots \\ u_k \end{pmatrix}.$$

The space $\mathcal{U}$ is an example of a finite-dimensional linear subspace since it is spanned by the finite number of elements $u_1(Z), \ldots, u_k(Z)$. This subspace is contained in the infinite-dimensional Hilbert space $\mathcal{H}$. Moreover, if the elements $u_1(Z), \ldots, u_k(Z)$ are linearly independent, then the dimension of $\mathcal{U}$ is identically equal to k.

Let h be an arbitrary element of $\mathcal{H}$. Then the projection of h onto the linear subspace $\mathcal{U}$ is given by the unique element $a_0^T u$ that satisfies

$$\langle h - a_0^T u, a^T u \rangle = 0 \quad \text{for all} \quad a = (a_1, \ldots, a_k)^T \in \mathbb{R}^k,$$

or

$$\sum_{j=1}^{k} a_j \langle h - a_0^T u, u_j \rangle = 0 \quad \text{for all} \;\; a_j, \; j = 1, \ldots, k.$$

Equivalently, $\langle h - a_0^T u, u_j \rangle = 0 \quad$ for all $j = 1, \ldots, k$,

or

$$E\{(h - a_0^T u) u^T\} = 0^{(1 \times k)},$$

or

$$E(hu^T) - a_0^T E(uu^T) = 0^{(1 \times k)}.$$

Any solution of a_0 such that

$$a_0^T E(uu^T) = E(hu^T)$$

would lead to the unique projection $a_0^T u$.

If $E(uu^T)$ is positive definite, and therefore has a unique inverse, then

$$a_0^T = E(hu^T)\{E(uu^T)\}^{-1},$$

in which case the unique projection will be

$$u_0 = a_0^T u = E(hu^T)\{E(uu^T)\}^{-1} u.$$

The norm-squared of this projection is equal to

$$E(hu^T)\{E(uu^T)\}^{-1} E(uh).$$

By the Pythagorean theorem,

$$\begin{aligned} \|h - a_0^T u\|^2 &= E(h - a_0^T u)^2 \\ &= E(h^2) - E(hu^T)\{E(uu^T)\}^{-1} E(uh). \end{aligned}$$

Example 2: q-dimensional Random Functions

Let $\mathcal{H}$ be the Hilbert space of mean-zero q-dimensional measurable random functions with finite second moments equipped with the inner product

$$\langle h_1, h_2 \rangle = E(h_1^T h_2).$$

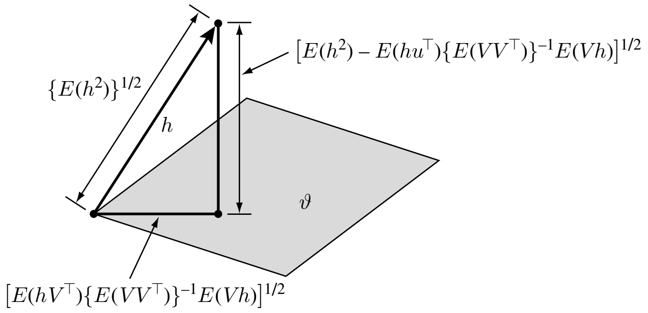
$\{E(h^2)\}^{1/2}$
h
$[E(h^2) - E(hu^\top)\{E(VV^\top)\}^{-1}E(Vh)]^{1/2}$
ϑ
$[E(hV^\top)\{E(VV^\top)\}^{-1}E(Vh)]^{1/2}$

Fig. 2.2. Geometric illustration of the Pythagorean theorem

Let "$v(Z)$" be an r-dimensional random function with mean zero and $E(v^T v) < \infty$. Consider the linear subspace $\mathcal{U}$ spanned by $v(Z)$; that is,

$$\mathcal{U} = \{B^{q\times r} v, \text{ where } B \text{ is any arbitrary } q \times r \text{ matrix of real numbers}\}.$$

The linear subspace $\mathcal{U}$ defined above is a finite-dimensional linear subspace contained in the infinite-dimensional Hilbert space $\mathcal{H}$. If the elements $v_1(Z), \ldots, v_r(Z)$ are linearly independent, then the dimension of $\mathcal{U}$ is $q \times r$. This can easily be seen by noting that $\mathcal{U}$ is spanned by the $q \times r$ linearly independent elements $u_{ij}(Z), i = 1, \ldots, q, j = 1, \ldots, r$, of $\mathcal{H}$, where, for any $i = 1, \ldots, q$, we take the element $u_{ij}^{q\times 1}(Z) \in \mathcal{H}$ to be the q-dimensional function of Z, where all except the i-th element are equal to 0 and the i-th element is equal to $v_j(Z)$ for $j = 1, \ldots, r$.

We now consider the problem of finding the projection of an arbitrary element $h \in \mathcal{H}$ onto $\mathcal{U}$. By the projection theorem, such a projection $B_0 v$ is unique and must satisfy

$$E\{(h - B_0 v)^T B v\} = 0 \quad \text{for all}\, B \in \mathbb{R}^{q\times r}. \tag{2.1}$$

The statement above being true for all B is equivalent to

$$E\{(h - B_0 v) v^T\} = 0^{q\times r} \quad \text{(matrix of all zeros)}. \tag{2.2}$$

To establish (2.2), we write

$$E\{(h - B_0 v)^T B v\} = \sum_i \sum_j B_{ij} E\{(h - B_0 v)_i v_j\}, \tag{2.3}$$

where $(h - B_0 v)_i$ denotes the i-th element of the q-dimensional vector $(h - B_0 v)$, v_j denotes the j-th element of the r-dimensional vector v, and B_{ij} denotes the (i, j)-th element of the matrix B.

If we take $B_{ij} = 1$ for $i = i'$ and $j = j'$, and 0 otherwise, it becomes clear from (2.3) that