Inference and Prediction in Large Dimensions

Denis Bosq, Delphine Blanke

Université Pierre et Marie Curie - Paris 6
Paris, France

John Wiley & Sons, Ltd

This work is in the Wiley-Dunod Series co-published between Dunod and John Wiley & Sons, Ltd.
Inference and Prediction in Large Dimensions
WILEY SERIES IN PROBABILITY AND STATISTICS

established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors
David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher,
Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, Sanford Weisberg,
David W. Scott, Adrian F. M. Smith

Editors Emeriti
Vic Barnett, J. Stuart Hunter, David G. Kendall, Jozef L. Teugels

A complete list of the titles in this series appears at the end of this volume.
Inference and Prediction in Large Dimensions

Denis Bosq, Delphine Blanke

Université Pierre et Marie Curie - Paris 6
Paris, France

This work is in the Wiley-Dunod Series co-published between Dunod and John Wiley & Sons, Ltd.
Contents

List of abbreviations ix

Introduction 1

Part I Statistical Prediction Theory 5

1 Statistical prediction 7
  1.1 Filtering ........................................ 7
  1.2 Some examples ................................... 8
  1.3 The prediction model ............................ 9
  1.4 P-sufficient statistics .......................... 11
  1.5 Optimal predictors ............................. 15
  1.6 Efficient predictors ............................ 21
  1.7 Loss functions and empirical predictors ....... 28
     1.7.1 Loss function .............................. 29
     1.7.2 Location parameters ........................ 30
     1.7.3 Bayesian predictors ......................... 31
     1.7.4 Linear predictors ........................... 32
  1.8 Multidimensional prediction ..................... 33

2 Asymptotic prediction 41
  2.1 Introduction .................................... 41
  2.2 The basic problem ................................ 41
  2.3 Parametric prediction for stochastic processes 43
  2.4 Predicting some common processes ............... 47
  2.5 Equivalent risks .................................. 54
  2.6 Prediction for small time lags .................... 56
  2.7 Prediction for large time lags ..................... 58
Part II Inference by Projection 61

3 Estimation by adaptive projection 63
  3.1 Introduction 63
  3.2 A class of functional parameters 63
  3.3 Oracle 66
  3.4 Parametric rate 68
  3.5 Nonparametric rates 72
  3.6 Rate in uniform norm 79
  3.7 Adaptive projection 81
    3.7.1 Behaviour of truncation index 82
    3.7.2 Superoptimal rate 85
    3.7.3 The general case 88
    3.7.4 Discussion and implementation 92
  3.8 Adaptive estimation in continuous time 92

4 Functional tests of fit 97
  4.1 Generalized chi-square tests 97
  4.2 Tests based on linear estimators 101
    4.2.1 Consistency of the test 105
    4.2.2 Application 106
  4.3 Efficiency of functional tests of fit 107
    4.3.1 Adjacent hypotheses 107
    4.3.2 Bahadur efficiency 110
  4.4 Tests based on the uniform norm 111
  4.5 Extensions. Testing regression 113
  4.6 Functional tests for stochastic processes 115

5 Prediction by projection 117
  5.1 A class of nonparametric predictors 117
  5.2 Guilbart spaces 121
  5.3 Predicting the conditional distribution 122
  5.4 Predicting the conditional distribution function 124

Part III Inference by Kernels 131

6 Kernel method in discrete time 133
  6.1 Presentation of the method 133
  6.2 Kernel estimation in the i.i.d. case 135
  6.3 Density estimation in the dependent case 138
    6.3.1 Mean-square error and asymptotic normality 138
    6.3.2 Almost sure convergence 140
  6.4 Regression estimation in the dependent case 148
    6.4.1 Framework and notations 148
CONTENTS

6.4.2 Pointwise convergence ................................. 150
6.4.3 Uniform convergence .................................. 153
6.5 Nonparametric prediction by kernel. ....................... 157
   6.5.1 Prediction for a stationary Markov process
       of order $k$ ....................................... 157
   6.5.2 Prediction for general processes ..................... 160

7 Kernel method in continuous time ............................. 163
   7.1 Optimal and superoptimal rates for density estimation .... 163
       7.1.1 The optimal framework .......................... 164
       7.1.2 The superoptimal case ......................... 167
   7.2 From optimal to superoptimal rates ..................... 170
       7.2.1 Intermediate rates ............................ 170
       7.2.2 Classes of processes and examples ............... 172
       7.2.3 Mean-square convergence ........................ 173
       7.2.4 Almost sure convergence ......................... 177
       7.2.5 An adaptive approach .......................... 180
   7.3 Regression estimation ................................ 181
       7.3.1 Pointwise almost sure convergence ............... 182
       7.3.2 Uniform almost sure convergence ................ 184
   7.4 Nonparametric prediction by kernel ...................... 186

8 Kernel method from sampled data ............................ 189
   8.1 Density estimation .................................. 190
       8.1.1 High rate sampling ............................ 190
       8.1.2 Adequate sampling schemes ..................... 193
   8.2 Regression estimation ................................ 198
   8.3 Numerical studies .................................. 201

Part IV Local Time ............................................. 207

9 The empirical density ....................................... 209
   9.1 Introduction ....................................... 209
   9.2 Occupation density .................................. 209
   9.3 The empirical density estimator ....................... 212
       9.3.1 Recursivity .................................. 213
       9.3.2 Invariance .................................. 213
   9.4 Empirical density estimator consistency ................ 214
   9.5 Rates of convergence ................................ 217
   9.6 Approximation of empirical density by common
       density estimators ................................. 220
Part V Linear Processes in High Dimensions

10 Functional linear processes
10.1 Modelling in large dimensions
10.2 Projection over linearly closed spaces
10.3 Wold decomposition and linear processes in Hilbert spaces
10.4 Moving average processes in Hilbert spaces
10.5 Autoregressive processes in Hilbert spaces
10.6 Autoregressive processes in Banach spaces

11 Estimation and prediction of functional linear processes
11.1 Introduction
11.2 Estimation of the mean of a functional linear process
11.3 Estimation of autocovariance operators
11.3.1 The space $S$
11.3.2 Estimation of $C_0$
11.3.3 Estimation of the eigenelements of $C_0$
11.3.4 Estimation of cross-autocovariance operators
11.4 Prediction of autoregressive Hilbertian processes
11.5 Estimation and prediction of ARC processes
11.5.1 Estimation of autocovariance
11.5.2 Sampled data
11.5.3 Estimation of $\rho$ and prediction

Appendix
A.1 Measure and probability
A.2 Random variables
A.3 Function spaces
A.4 Common function spaces
A.5 Operators on Hilbert spaces
A.6 Functional random variables
A.7 Conditional expectation
A.8 Conditional expectation in function spaces
A.9 Stochastic processes
A.10 Stationary processes and Wold decomposition
A.11 Stochastic integral and diffusion processes
A.12 Markov processes
A.13 Stochastic convergences and limit theorems
A.14 Strongly mixing processes
A.15 Some other mixing coefficients
A.16 Inequalities of exponential type

Bibliography

Index
### List of abbreviations

*...*  
see the Appendix

■  
end of a proof

⋄  
end of an example, a remark

a.e.  
amost everywhere

cadlag  
continuous on the right and limit on the left

iff  
if and only if

i.i.d. r.v.  
independent and identically distributed random variables

A(i)–(iii)  
conditions A(i) to A(iii)

A(i),(iii)  
conditions A(i) and A(iii)

AR  
AutoRegressive process

ARH (ARB)  
AutoRegressive Hilbertian (Banach) process

BUE (BUP)  
Best Unbiased Estimator (Predictor)

EDE  
Empirical Density Estimator

EUE (EUP)  
Efficient Unbiased Estimator (Predictor)

FLP  
Functional Linear Process

GSM  
Geometrically Strongly Mixing process

LCS  
Linearly Closed Space

LIL  
Law of the Iterated Logarithm

LPH(WS)  
Linear Process in $H$ (in the Wide Sense)

MA  
Moving Average process

MAH  
Moving Average Hilbertian process

MLE  
Maximum Likelihood Estimator

MWS  
Markovian process in the Wide Sense

OU  
Ornstein–Uhlenbeck process

QPE  
Quadratic Prediction Error

WSH  
Weakly Stationary Hilbertian process

$\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$  
sets of natural numbers, integers, rational numbers, real numbers, complex numbers.

$[a, b]$, $]a, b[$  
closed interval, open interval

$\bar{A}$, $\overline{A}$, $A^c$  
interior, closure and complement of $A$

$\sigma(\cdots)$  
$\sigma$-algebra generated by $\cdots$

$\mathbbm{1}_A$  
indicator of $A$: $\mathbbm{1}_A(x) = 1$, $x \in A$; $\mathbbm{1}_A(x) = 0$, $x \in A^c$
\[ \ln x \] logarithm of \( x \)

\[ [x] \] integer part (floor) of \( x \)

\( x_+, x_- \) \( x_+ = \max(0, x), x_- = \max(0, -x) \)

\( f \otimes g \) for all \((x, y), (f \otimes g)(x, y) = f(x)g(y)\)

\( u_n \sim v_n \) \( u_n/v_n \to 1 \)

\( u_n \simeq v_n \) there exist constants \( c_1 \) and \( c_2 \) such that, for large enough \( n \)

\( 0 < c_1 v_n < u_n < c_2 v_n \)

\( u_n = o(v_n) \) \( u_n/v_n \to 0 \)

\( u_n = \mathcal{O}(v_n) \) there is a \( c > 0 \) such that \( u_n \leq c v_n, n \geq 1 \)

\( X \perp Y \) the random variables \( X \) and \( Y \) are stochastically independent

\( X \overset{d}{=} Y \) the random variables \( X \) and \( Y \) have the same distribution

\( \overset{d}{\to}, \overset{w}{\to} \) convergence in distribution, weak convergence

\( \overset{p}{\to}, \overset{a.s.}{\to} \) convergence in probability, almost sure convergence
Introduction

The purpose of the present work is to investigate inference and statistical prediction when the data and (or) the unknown parameter have large or infinite dimension.

The data in question may be curves, possibly interpolated from discrete observations, or sequences associated with a large number of items.

The parameters are functions (distribution function, density, regression, spectral density, ...) or operators (covariance or correlation operators, ...).

Grenander’s well-known book ‘Abstract Inference’ was devoted to such topics. However our aim is rather different: whereas Grenander studies a maximum likelihood-type method (the ‘sieves’), we must often use nonparametric instruments.

More precisely we focus on adaptive projection and kernels methods and, since prediction is our main goal, we study these methods for correlated data.

Statistical prediction theory

In the first part we provide some elements of statistical prediction theory.

A priori a prediction problem is similar to an estimation problem since it deals with ‘approximation’ of a nonobserved random variable by using an observed one. Despite this analogy, prediction theory is rarely developed from a statistical point of view.

In Chapter 1 we study properties of the prediction model: sufficient statistics for prediction, unbiased predictors, Rao–Blackwell Theorem and optimality, Cramér-Rao type inequalities, efficiency and extended exponential model, and extensions to function spaces. Various applications to prediction of discrete or continuous time processes are given.

Chapter 2 deals with asymptotics. We present the Blackwell algorithm for prediction of 0-1 sequences; and results concerning pointwise convergence and limit in distribution for predictors associated with estimators of the unknown parameter. In addition we briefly present prediction for small and large time lags.


Inference by projection

The second part considers statistical models dependent on an infinite-dimensional functional parameter, say \( \varphi \). The main assumption is that \( \varphi \) is the limit of a sequence \( (\varphi_k) \) of finite-dimensional parameters that are the expectation of known functions of data. This is the case for density, spectral density and covariance operators, among others.

For this type of parameter one may construct genuine estimators using empirical means, the main problem being to determine a suitable truncation index \( k_n \) which depends on the size \( n \) of the sample.

In a Hilbertian context these estimators may be interpreted as ‘projection estimators’. Now, the choice of \( k_n \) depends on the norm of \( ||\varphi - \varphi_k||, \ k \geq 1 \). Since this norm is, in general, unknown, we study the behaviour of an adaptive version of the projection estimator. It can be shown that this estimator reaches a parametric rate on a subclass dense in the parameter space and the optimal rate, up to a logarithm, elsewhere (Chapter 3).

Chapter 4 is devoted to tests of fit based on projection estimators. If the functional parameter is density, these tests appear as extensions of the \( \chi^2 \)-test, the latter being associated with a histogram. Various properties of these functional tests are considered: asymptotic level and power under local hypotheses, asymptotic efficiency with respect to the Neyman–Pearson test, and Bahadur efficiency.

Finally, Chapter 5 deals with a class of nonparametric predictors based on regression projection estimators. The general framework of the study allows application of the results to approximation of the conditional distribution of the future, given the past. As an application we construct prediction intervals.

Inference by kernels

Part three deals with the popular kernel method. This consists of regularization of the empirical measure by convolution.

Chapter 6 presents the method in a discrete time context. Since our final goal is prediction we concentrate on the case where data are correlated. Optimal rates of convergence for density and regression estimators are given, with application to prediction of Markov and more general processes.

Continuous time is considered in Chapter 7. On one hand, one may obtain optimal (minimax) rates similar to those of the discrete case; on the other hand, irregularity of sample paths provides additional information which leads to ‘superoptimal’ rates. This phenomenon occurs if, for example, one observes a stationary diffusion process. In a multidimensional context the situation is rather intricate and various rates may appear, in particular for multidimensional Gaussian or diffusion processes. Finally, if regularity of sample paths is unknown, it is possible to construct adaptive estimators that reach optimal rates, up to a logarithm.
Now, in practice, observation of a continuous time process with irregular sample path is somewhat difficult. Thus, in general, the Statistician works with sampled data. Chapter 8 examines that situation. We especially investigate high rate sampling. We show that, for density estimation, there exists an optimal sampling rate which depends on regularity and dimension of sample paths. Various simulations confirm the theoretical results. Analogous results are obtained for regression estimation.

Local time

In this part, we consider the special case where the empirical measure of a continuous time process has a density with respect to Lebesgue measure. This ‘empirical density’ (or ‘local time’) does exist in many cases, in particular for real stationary diffusion processes. It defines an unbiased density estimator, which reaches the superoptimal rate and has various nice asymptotic properties. Moreover, the projection and kernel density estimators are approximations of local time.

Linear processes in high dimensions

Let $\xi = (\xi_t, t \in \mathbb{R})$ be a real continuous time process. One may cut $\xi$ into pieces by setting $X_n = (\xi_{n\delta+t}, 0 \leq t \leq \delta)$, $n \in \mathbb{Z}$ ($\delta > 0$). So $\xi$ may be interpreted as a sequence of random variables with values in some function space. This representation is a motivation for studying discrete time processes in such a space.

In Chapter 10 we consider linear processes in Hilbert spaces. We use a definition based on Wold decomposition which is more general than the classical one. Then, having in mind statistical applications, we focus on autoregressive and moving average processes in Hilbert spaces. Some extensions to Banach space are also presented.

Finally, Chapter 11 deals with inference for functional linear processes. Estimation of the mean is studied in detail. To estimate the covariance operator of an autoregressive process we use the fact that the empirical covariance operator is the empirical mean of as $S$-valued autoregressive process, where $S$ denotes the space of Hilbert–Schmidt operators. Empirical cross-covariance operators enjoy similar properties. The above considerations lead to estimation of the autocorrelation operator, allowing us to construct a suitable statistical predictor.

General comments

Prerequisites for this book are knowledge of probability based on measure theory, a basic knowledge of stochastic processes, and classical theory of statistics. Some reminders appear in the appendix.

A large number of results presented in this book are new or very recent. Some of them are improvements of results that appear in Bosq (1998, 2000, 2005).
For the references we have had to make a difficult choice among thousands of papers devoted to the various topics developed in this book.

Acknowledgments

Part of the material presented in this book is based on our own published work. We are greatly indebted to our collaborators and colleagues from whom we learned a great deal. We also thank J. Pavlevski, S. Barclay, W. Hunter, S. Kiddle, A. Bourguignon and E. d’Engenières for their kind encouragements and support, as well as S. Lightfoot for the preparation of the book. Finally, we especially acknowledge C. Lendrem for her very careful rereading.
Part I
Statistical Prediction Theory
1

Statistical prediction

1.1 Filtering

Filtering is searching information provided by observed events on nonobserved events. These events are assumed to be associated with random experiments.

To describe such a problem, one may define a *probability space* $(\Omega, \mathcal{A}, \mathbb{P})$ and two sub-$\sigma$-algebras of $\mathcal{A}$, say $\mathcal{B}$ and $\mathcal{C}$, respectively observed and non-observed. ‘$\mathcal{B}$-observed’ means that, for all $B \in \mathcal{B}$, the observer knows whether $B$ occurs or not. Information provided by $\mathcal{B}$ on $\mathcal{C}$ may be quantified by $P^B$, the *conditional probability* with respect to $\mathcal{B}$. In general the statistician does not know $P^B$.

In the following, we are mainly interested in prediction (or forecasting). This signifies that $\mathcal{B}$ is associated with the past and $\mathcal{C}$ with the future. In practice one tries to predict a $\mathcal{C}$-measurable *random variable* $Y$ from a $\mathcal{B}$-measurable random variable $X$.

If $X$ is partly controllable, one can replace prediction by foresight, which consists in preparing for the future by constructing scenarios and then by selecting the most favourable option. It is then possible to make a plan for the future. We refer to Kerstin (2003) for a discussion about these notions.

Predictions can also sometimes modify the future. For example the publication of economic forecasts may change the behaviour of the economic agents involved and, therefore, influences future data. This phenomenon is discussed in Armstrong (2001), among others.
In this book we don’t deal with this kind of problem, but we focus on statistical prediction, that is, prediction of future from a time series (i.e. a family of data indexed by time, see figure 1.1).\(^1\)

### 1.2 Some examples

We now state some prediction problems. The following examples show that various situations can arise.

**Example 1.1 (The Blackwell’s problem)**

The purpose is to predict occurrence or nonoccurrence of an event associated with an experiment when \(n\) repetitions of this experiment are available. Let \((x_n, n \geq 1)\) be the sequence of results \((x_n = 1\) if the \(n\)th event occurs, \(= 0\) if not\)) and note by \(p_n = p_n(x_1, \ldots, x_n)\) any predictor of \(x_{n+1}\). The problem is to choose a ‘good’ \(p_n\). Note that the sequence \((x_n)\) may be deterministic or random.

**Example 1.2 (Forecasting a discrete time process)**

Let \((X_n, n \geq 1)\) be a real square integrable ‘random process’. One observes \(X = (X_1, \ldots, X_n)\) and intends to predict \(Y = X_{n+h} (h \geq 1); h\) is called the forecasting horizon.

\(^1\)Provided by the NOAA-CIRES Climate Diagnostics Center: http://www.cdc.noaa.gov
This scheme corresponds to many practical situations: prediction of temperature, unemployment rate, foreign exchange rate, …

Example 1.3 (Forecasting a continuous time process)
Let \( (X_t, t \in \mathbb{R}_+) \) be a real square integrable *continuous time random process*. One wants to predict \( Y = X_{t+h} (h > 0) \) from the observed piece of sample path \( X = (X_t, 0 \leq t \leq T) \) or \( X = (X_{t_1}, \ldots, X_{t_n}) \).

The difference with the above discrete time scheme is the possibility of considering small horizon forecasting (i.e. \( h \to 0(+) \)).

Classical applications are: prediction of electricity consumption, evolution of market prices during a day’s trading, prediction of the *counting process* associated with a *point process*, …

Example 1.4 (Predicting curves)
If the data are curves one may interpret them as realizations of random variables with values in a suitable function space. Example 1.3 can be rewritten in this new framework.

This kind of model appears to be useful when one wishes to predict the future evolution of a quantity during a full time interval. For example, electricity consumption for a whole day or variations of an electrocardiogram during one minute.

Example 1.5 (Prediction of functionals)
In the previous examples one may predict functionals linked with the future of the observed process. For instance:

- The conditional density \( f_Y^X \) or the conditional distribution \( P_Y^X \) of \( Y \) given \( X \).
- The next crossing at a given level, that is
  \[
  Y_t = \min\{\tau : \tau > n, X_{\tau} = x\},
  \]
  and, more generally, the next visit to a given Borel set,
  \[
  Y_B = \min\{\tau : \tau > n, X_{\tau} \in B\}.
  \]

Notice that the forecasting horizon is not defined for these visit questions.
- Finally it is interesting to construct prediction intervals of the form \( P(Y \in [a(X), b(X)]) \).

We now specify the prediction model, beginning with the real case.

1.3 The prediction model
Let \( (\Omega, \mathcal{A}, P_\theta, \theta \in \Theta) \) be a statistical model, where \( \mathcal{A} \) is the \( \sigma \)-algebra of events on \( \Omega \) and \( (P_\theta) \) a family of probability measures on \( \mathcal{A} \) indexed by the unknown
parameter $\theta$. $B = \sigma(X)$ is the observed $*\sigma$-algebra* and $C$ the nonobserved one. $X$ takes its values in some $*\text{measurable space}* (E_0, B_0)$.

*A priori* one intends to predict a $C$-measurable real random variable $Y$. Actually, for reasons that are specified below, we extend the problem by considering prediction of $g(X, Y, \theta) \in \bigcap_{\theta \in \Theta} L^2(P_{\theta})$, where $g$ is known and $Y$ nonobserved.

If $g$ only depends on $Y$ we will say that one deals with *pure prediction*, if it depends only on $\theta$ it is an *estimation problem*; finally, if $g$ is only a function of $X$ the topic is *approximation* (at least if $g$ is difficult to compute!). The other cases are *mixed*. So, prediction theory appears as an extension of estimation theory.

Now a *statistical predictor* of $g(X, Y, \theta)$ is a known real measurable function of $X$, say $p(X)$. Note that the $*\text{conditional expectation}* E_{\theta, g}$ is not in general a statistical predictor. In the following we assume that $p(X) \in \bigcap_{\theta \in \Theta} L^2(P_{\theta})$, unless otherwise stated.

In order to evaluate the accuracy of $p$ one may use the *quadratic prediction error* (QPE) defined as

$$R_{\theta}(p, g) = E_{\theta}(p(X) - g(X, Y, \theta))^2, \theta \in \Theta.$$  

This *risk function* induces the *preference relation*

$$p_1 \prec p_2 \iff R_{\theta}(p_1, g) \leq R_{\theta}(p_2, g), \theta \in \Theta. \quad (1.1)$$

If (1.1) is satisfied we will say that ‘the predictor $p_1$ is preferable to the predictor $p_2$ for predicting $g$’, and write $p_1 \prec p_2 \ (g)$.

The QPE is popular and easy to handle, it has withstood the critics because it is difficult to find a good substitute. However, some other preference relations will be considered in Section 1.7 of this chapter.

Now, let $\mathcal{P}_G$ be the class of statistical predictors $p$ such that $p(X) \in G = \bigcap_{\theta \in \Theta} G_\theta$ where $G_\theta$ is some closed linear space of $L^2(P_{\theta})$, with $*\text{orthogonal projector}* \Pi^\theta, \theta \in \Theta$. The following lemma shows that a prediction problem is, in general, mixed.

**Lemma 1.1** *(Decomposition of the QPE)*

*If $p \in \mathcal{P}_G$, its QPE has decomposition*

$$E_{\theta}(p - g)^2 = E_{\theta}(p - \Pi^\theta g)^2 + E_{\theta}(\Pi^\theta g - g)^2, \theta \in \Theta. \quad (1.2)$$

**Hence, $p_1 \prec p_2$ for predicting $g$ if and only if $p_1 \prec p_2$ for predicting $\Pi^\theta g$.**

**Proof:**

Decomposition (1.2) is a straightforward application of the Pythagoras theorem.
Therefore

\[ E_\theta(p_1 - g)^2 \leq E_\theta(p_2 - g)^2 \iff E_\theta(p_1 - \Pi^\theta g)^2 \leq E_\theta(p_2 - \Pi^\theta g)^2, \theta \in \Theta. \]

Lemma 1.1 is simple but crucial: one must focus on the statistical prediction error \( E_\theta(p - \Pi^\theta g)^2 \), since the probabilistic prediction error \( E_\theta(\Pi^\theta g - g)^2 \) is not controllable by the statistician.

Thus, predicting \( g(X, Y, \theta) \) or predicting \( \Pi^\theta g(X, Y, \theta) \) is the same activity. In particular, to predict \( Y \) is equivalent to predicting \( \Pi^\theta Y \); this shows that a nondegenerated prediction problem is mixed.

**Example 1.6**
Suppose that \( X_1, \ldots, X_n \) are observed temperatures. One wants to know if \( X_{n+1} \) is going to exceed some threshold \( \theta \); then \( g = \mathbb{1}_{X_{n+1} > \theta} \) and the problem is mixed.

Let us now give some classical examples of spaces \( G \).

**Example 1.7**
If \( G = \cap_{\theta \in \Theta} L^2(\Omega, \mathcal{B}, P^\theta) \), then \( \Pi^\theta \) is the conditional expectation given \( \mathcal{B} \).

**Example 1.8**
Suppose that \( X = (X_1, \ldots, X_n) \) where \( X_i \in L^2(\theta_i), \ 1 \leq i \leq n \) and \( G = \text{sp} \{1, X_1, \ldots, X_n\} \) does not depend on \( \theta \). Then \( \Pi^\theta g \) is the affine regression of \( g \) on \( X \).

**Example 1.9**
If \( X = (X_1, \ldots, X_n) \) with \( X_i \in L^{2k}(\theta_i), \ 1 \leq i \leq n, \ (k \geq 1) \) and \( G = \text{sp}\{1, X_i, \ldots, X_i^k, 1 \leq i \leq n\} \) does not depend on \( \theta \). Then \( \Pi^\theta g \) is a polynomial regression:

\[ \Pi^\theta g(X_1, \ldots, X_n) = a_0 + \sum_{i=1}^{n} \sum_{j=1}^{k} b_{ji} X_i^j, \]

where \( a_0 \) and \( (b_{ji}) \) only depend on \( \theta \).

**1.4 P-sufficient statistics**
As in estimation theory, sufficient statistics play a significant role in prediction. The definition is more restrictive.
Definition 1.1
A statistic \( S(X) \) is said to be sufficient for predicting \( g(X, Y, \theta) \) (or P-sufficient) if

(i) \( S(X) \) is sufficient in the usual sense, i.e. the conditional distribution \( P_{S(X)}^X \) of \( X \) with respect to \( S(X) \) does not depend on \( \theta \).

(ii) For all \( \theta, X \) and \( g(X, Y, \theta) \) are *conditionally independent* given \( S(X) \).

Condition (ii) means that

\[
PS_{\theta}(X \in B, g \in C) = P_{\theta}(X \in B)P_{\theta}^{S(X)}(g \in C),
\]

\( \theta \in \Theta, B \in B_{\mathbb{R}}, C \in B_{\mathbb{R}} \).

Note that, if \( g(X, Y, \theta) = g(S(X), \theta) \), (ii) is automatically satisfied. Moreover one may show (see Ash and Gardner 1975, p. 188) that (ii) is equivalent to

(ii)' \( P_{\theta}^X = P_{\theta}^{S(X)} \), \( \theta \in \Theta \).

We now give a statement which connects sufficiency and P-sufficiency.

Theorem 1.1
Suppose that \( (X, Y) \) has a strictly positive density \( f_{\theta}(x, y) \) with respect to a *\( \sigma \)-finite* measure \( \mu \otimes \nu \). Then

(a) If \( S(X) \) is P-sufficient for predicting \( Y \), \( (S(X), Y) \) is sufficient in the statistical model associated with \( (X, Y) \).

(b) Conversely if \( f_{\theta}(x, y) = h_{\theta}(S(x), y)c(x)d(y) \), then \( S(X) \) is P-sufficient for predicting \( Y \).

Proof:

(a) Consider the decomposition

\[
f_{\theta}(x, y) = f_{\theta}(x)f_{\theta}(y|x),
\]

where \( f_{\theta}(\cdot) \) is the density of \( X \) and \( f_{\theta}(\cdot|x) \) the conditional density of \( Y \) given \( X = x \). If \( S(X) \) is P-sufficient, the factorization theorem (see Lehmann and Casella 1998, p. 35) yields \( f_{\theta} = \varphi_{\theta}(S(x))\psi(x) \), where \( \psi \) does not depend on \( \theta \).

Now (ii)' entails \( f_{\theta}(y|x) = \gamma_{\theta}(y|S(x)) \) where \( \gamma_{\theta}(\cdot|x) \) is the conditional density of \( Y \) given \( S(X) = \cdot \).

Finally \( f_{\theta}(x, y) = \varphi_{\theta}(S(x))\gamma_{\theta}(y|S(x))\psi(x) \) and the factorization theorem gives sufficiency of \( (S(X), Y) \) in the model associated with \( (X, Y) \).

(b) Conversely, the relation

\[
f_{\theta}(x, y) = h_{\theta}(S(x), y)c(x)d(y)
\]
may be replaced by \( f_\theta(x, y) = h_\theta(S(x), y) \) by substituting \( \mu \) and \( \nu \) for \( c \cdot \mu \) and \( d \cdot \nu \) respectively.

This implies

\[
f_\theta(x) = \int h_\theta(S(x), y) \, d\nu(y) := H_\theta(S(x)),
\]

thus, \( S(X) \) is sufficient, and since

\[
f_\theta(y|x) = \frac{h_\theta(S(x), y)}{H_\theta(S(x))},
\]

(ii)' holds, hence \( S(X) \) is \( P \)-sufficient.

We now give some examples and counterexamples concerning \( P \)-sufficiency.

**Example 1.10**

Let \( X_1, \ldots, X_{n+1} \) be independent random variables with common density \( \theta e^{-\theta x} I_{x>0} \) \( (\theta > 0) \). Set \( X = (X_1, \ldots, X_n) \) and \( Y = X_{n+1} - X_n \). Then \( \sum_{i=1}^n X_i \) is sufficient for \( X \) but \( (\sum_{i=1}^n X_i, Y) \) is not sufficient for \( (X, Y) \).

This shows that, in Theorem 1.1 (a), \( P \)-sufficiency of \( S(X) \) cannot be replaced by sufficiency.

**Example 1.11**

Let \( (X_1, X_2) \) and \( Y \) be such that the density of \( (X, Y) \) is

\[
f_\theta(x_1, x_2, y) = \theta^{3/2} e^{-\theta(x_1+y)} e^{-\frac{x_1^2}{2}} e^{-\frac{x_2^2}{2}} I_{x_1>0, x_2>0, y>0}, (\theta > 0).
\]

Then \( (X_1, Y) \) is sufficient in the model \( (X, Y) \) but \( X_1 \) is not sufficient in the model \( (X_1, X_2) \).

This shows that, in Theorem 1.1 (b), the result is not valid if the special form \( f_\theta(x, y) = h_\theta(S(x), y)c(x)d(y) \) does not hold.

**Example 1.12** *(discrete time Gaussian process)*

Let \( (X_i, i \geq 1) \) be a "Gaussian process", where \( X_i \overset{d}{=} \mathcal{N}(\theta, 1), i \geq 1 (\theta \in \mathbb{R}) \).

Suppose that the covariance matrix \( C_n \) of \( X = (X_1, \ldots, X_n) \) is known and invertible for each \( n \), and set

\[
C_{n+1} = \begin{bmatrix}
C_n & \gamma_{n+1} \\
\gamma_{n+1} & 1
\end{bmatrix}.
\]
Then, in order to predict $Y = X_{n+1}$, the statistic

$$(C_n^{-1}u_nX, C_n^{-1}X_{n+1})$$

where $u_n = (1, \ldots, 1)'$, is $P$-sufficient. The statistic $C_n^{-1}u_nX$ is sufficient but not $P$-sufficient. \hfill \Box

**Example 1.13 (Poisson process)**

Let $N = (N_t, t \geq 0)$ be a homogeneous *Poisson process* with intensity $\lambda$, and observed over $[0, T]$. Then $N_T$ is sufficient. Now, since $N$ is a *Markov process* the $\sigma$-algebras $\sigma(N_s, s \leq T)$ and $\sigma(N_s, s > T)$ are independent given $N_T$. Hence $N_T$ is $P$-sufficient for predicting $N_{T+h}$ ($h > 0$). \hfill \Box

**Example 1.14 (Ornstein–Uhlenbeck process)**

Consider an *Ornstein–Uhlenbeck process* (OU) defined as

$$X_t = \int_{-\infty}^{t} e^{-\theta(t-s)} dW(s), t \in \mathbb{R}, (\theta > 0),$$

where $W = (W_t, t \in \mathbb{R})$ is a *standard bilateral Wiener process* and the integral is taken in *Ito* sense. $(X_t)$ is a zero-mean stationary Gaussian Markov process. Here the observed variable is $X(T) = (X_t, 0 \leq t \leq T)$.

The likelihood of $X(T)$ with respect to the distribution of $W(T) = (W_t, 0 \leq t \leq T)$ in the space $C([0, T])$ of continuous real functions defined on $[0, T]$ with uniform norm is

$$L(X(T), \theta) = \exp \left(-\frac{\theta}{2} (X_T^2 - X_0^2 - T) - \frac{\theta^2}{2} \int_0^T X_t^2 dt \right). \quad (1.3)$$

(See Liptser and Shiraev 2001). Then, the factorization theorem yields sufficiency of the statistics $(X_0^2, X_T^2, \int_0^T X_t^2 dt)$. Consequently $Z_T = (X_0, X_T, \int_0^T X_t^2 dt)$ is also sufficient.

Now, since $(X_t)$ is Markovian, we have

$$\sigma(X(T)) \subseteq \sigma(X_{T+h})|\sigma(X_T),$$

then

$$P_{\theta, X_{T+h}}^{Z_T} = P_{\theta, X_{T+h}}^{X_T} = P_{\theta, X_{T+h}}^{X(T)}$$

hence (ii)' holds and $Z_T$ is $P$-sufficient for predicting $X_{T+h}$. \hfill \Box

The next statement shows that one may use $P$-sufficient statistics to improve a predictor. It is a Rao–Blackwell type theorem (See Lehmann 1991, p. 47).
Theorem 1.2 (Rao–Blackwell theorem)
Let $S(X)$ be a $P$-sufficient statistic for $g(X,Y,\theta)$ and $p(X)$ a statistical predictor of $g$. Then $E^{S(X)}p(X)$ is preferable to $p(X)$ for predicting $g$.

Proof: $P^{S(X)}_{\theta,X}$ does not depend on $\theta$, thus $q(X) = E^{S(X)}p(X)$ is again a statistical predictor. Furthermore

$$E_\theta(p - g)^2 = E_\theta(p - q)^2 + E_\theta(q - g)^2 + 2E_\theta((p - q)(q - g)).$$

(1.4)

Now, by definition of conditional expectation $E_\theta((p - q)q) = 0$ and, from condition (ii) in Definition 1.1,

$$E_\theta[(p - q)g] = E_\theta[E_\theta^x(p - q)g] = E_\theta^x(p - q)E_\theta^x(g) = 0$$

since $E_\theta^x(p) = q$; thus (1.4) gives $E_\theta(p - g)^2 \geq E_\theta(q - g)^2$.

Note that, if (ii) does not hold, the result remains valid provided $E_\theta[(p - q)g] = 0$, $\theta \in \Theta$.

Finally it is noteworthy that, if $S(X)$ is $P$-sufficient to predict $g$ it is also $P$-sufficient to predict $E_\theta^{S(X)}(g) = E_\theta^x(g)$; actually this conditional expectation is a function of $S(X)$, hence $X \perp E_\theta^{S(X)}(g)|S(X)$.

1.5 Optimal predictors
A statistical predictor $p_\theta$ is said to be optimal in the family $P$ of predictors of $g$ if $p_\theta \prec p, p \in P$

that is

$$E_\theta(p_\theta - g)^2 \leq E_\theta(p - g)^2; \quad \theta \in \Theta, p \in P.$$ 

Notice that, in the family of all square integrable predictors, such a predictor does not exist as soon as $E_\theta^x(g)$ is not constant in $\theta$. Indeed, $p_1(X) = E_\theta^x(g)$ is optimal at $\theta_1$ when $p_2(X) = E_{\theta_2}^x(g)$ is optimal at $\theta_2$ which is impossible if $E_{\theta_1}(E_{\theta_1}^x(g) - E_{\theta_2}^x(g))^2 \neq 0; i = 1, 2$.

Thus, in order to construct an optimal predictor, it is necessary to restrict $P$. For example one may consider only unbiased predictors.

Definition 1.2
A predictor $p(X)$ of $g(X,Y,\theta)$ is said to be unbiased if

$$E_\theta(p(X)) = E_\theta(g(X,Y,\theta)), \theta \in \Theta.$$ 

(1.5)

Condition (1.5) means that $p$ is an unbiased estimator of the parameter $E_\theta(g)$. 

OPTIMAL PREDICTORS
Example 1.15 (Autoregressive process of order 1)
Consider an autoregressive process of order 1 (AR(1)) defined as

\[ X_n = \sum_{j=0}^{\infty} \theta^j \varepsilon_{n-j}, n \in \mathbb{Z} \tag{1.6} \]

where \((\varepsilon_n, n \in \mathbb{Z})\) is strong white noise (i.e. a sequence of i.i.d. random variables such that \(0 < \mathbb{E}\varepsilon_n^2 = \sigma^2 < \infty\) and \(\mathbb{E}\varepsilon_n = 0\)) and \(0 < |\theta| < 1\). The series converges in mean square and almost surely.

From (1.6) it follows that

\[ X_n = \theta X_{n-1} + \varepsilon_n, n \in \mathbb{Z}, \]

then

\[ \mathbb{E}_{\theta, \sigma^2}(X_{n+1}^{i \leq n}) = \theta X_n. \]

For convenience we suppose that \(\sigma^2 = 1 - \theta^2\), hence \(\mathbb{E}_{\theta, \sigma^2}(X_n^2) = 1\). Now a natural unbiased estimator of \(\theta\), based on \(X_1, \ldots, X_n\) \((n > 1)\) is

\[ \hat{\theta}_n = \frac{1}{n-1} \sum_{i=1}^{n-1} X_i X_{i+1}, \]

hence a predictor of \(X_{n+1}\) is defined as

\[ \hat{X}_{n+1} = \hat{\theta}_n X_n. \]

Now we have

\[ \mathbb{E}_\theta(\hat{X}_{n+1}) = \theta^2 \frac{1 - \theta^{n-1}}{(1 - \theta)(n - 1)} \mathbb{E}_\theta(X_0^3), \]

thus \(\hat{X}_{n+1}\) is unbiased if and only if (iff)

\[ \mathbb{E}_\theta(X_0^3) = 0, 0 < |\theta| < 1. \]

We now give an extension of the classical Lehmann–Scheffé theorem (see Lehmann 1991, p. 88).

First recall that a statistic \(S\) is said to be complete if

\[ \mathbb{E}_\theta(U) = 0, \theta \in \Theta \text{ and } U = \varphi(S) \Rightarrow U = 0, P_\theta \text{ a.s. for all } \theta. \]

Theorem 1.3 (Lehmann–Scheffé theorem)
If \(S\) is a complete \(P\)-sufficient statistic for \(g\) and \(p\) is an unbiased predictor of \(g\), then \(\mathbb{E}^S(p)\) is the unique optimal unbiased predictor of \(g\) \((P_\theta \text{ a.s. for all } \theta)\).
PROOF:
From Theorem 1.2 any optimal unbiased predictor of $g$ is a function of $S$. Thus $E^S(p)$ is a candidate since $E_\theta[E^S p] = E_\theta(p) = g$. But completeness of $S$ entails uniqueness of an unbiased predictor of $g$ as a function of $S$, hence $E^S(p)$ is optimal. □

Example 1.13 (continued)
$N_T$ is a complete $P$-sufficient statistic for $N_{T+h}$, the consequently unbiased predictor

$$p(N_T) = \frac{T + h}{T} N_T$$

is optimal for predicting $N_{T+h}$. Its quadratic error is

$$E_\lambda(p(N_T) - N_{T+h})^2 = \lambda h \left(1 + \frac{h}{T}\right).$$

It is also optimal unbiased for predicting

$$E_\lambda^{N_T}(N_{T+h}) = \lambda h + N_T$$

with quadratic error

$$E_\lambda(p(N_T) - E_\lambda^{N_T}(N_{T+h}))^2 = \frac{\lambda h^2}{T}. \quad \Box$$

The following statement gives a condition for optimality of an unbiased predictor.

Theorem 1.4
Set $\mathcal{U} = \{U(X) : E_\theta U^2(X) < \infty, E_\theta U(X) = 0; \theta \in \Theta\}$. Then an unbiased predictor $p$ of $g$ is optimal iff

$$E_\theta[(p - g)U] = 0; U \in \mathcal{U}, \theta \in \Theta. \quad (1.7)$$

PROOF:
If $p$ is optimal, we set

$$q = p + \alpha U, U \in \mathcal{U}, \alpha \in \mathbb{R},$$

then

$$E_\theta(p + \alpha U - g)^2 \geq E_\theta(p - g)^2,$$

therefore

$$\alpha^2 E_\theta U^2 + 2\alpha E_\theta((p - g)U) \geq 0, U \in \mathcal{U}, \alpha \in \mathbb{R}$$

which is possible only if (1.7) holds.
Conversely, if \( p \) satisfies (1.7) and \( p' \) denotes another unbiased predictor, then 
\( p' - p \in \mathcal{U} \), therefore
\[
E_\theta[(p - g)(p' - p)] = 0,
\]
which implies
\[
E_\theta[(p' - g)^2 - (p - g)^2] = E_\theta(p'^2 - p^2) - 2E_\theta[(p' - p)g]
\]
\[
= E_\theta(p'^2 - p^2) - 2E_\theta[(p' - p)p]
\]
\[
= E_\theta(p' - p)^2 \geq 0,
\]
thus \( p \) is preferable to \( p' \).

Note that such a predictor is unique. Actually, if \( p' \) is another optimal unbiased predictor, (1.7) yields
\[
E_\theta((p' - p)U) = 0, U \in \mathcal{U}, \theta \in \Theta
\]
which shows that \( p' - p \) is an optimal unbiased predictor of 0. But 0 is an optimal unbiased predictor of 0, with quadratic error 0, thus \( E_\theta(p' - p)^2 = 0, \theta \in \Theta \) hence \( p' = p, P_\theta \text{ a.s. for all } \theta \).

Now, since an unbiased predictor of \( g \) is an unbiased estimator of \( E_\theta g \), it is natural to ask whether the best unbiased estimator (BUE) of \( E_\theta g \) and the best unbiased predictor (BUP) of \( g \) coincide or not.

The next theorem gives an answer to this question.

**Theorem 1.5**

*The BUE of \( E_\theta g \) and the BUP of \( g \) coincide iff*
\[
E_\theta(gU) = 0, U \in \mathcal{U}, \theta \in \Theta.
\] (1.8)

**Proof:**
First suppose that (1.8) holds. If \( p \) is the BUE of \( E_\theta g \), it is also the BUP of \( E_\theta g \), then Theorem 1.4 implies that, for all \( U \in \mathcal{U} \) and all \( \theta \in \Theta \), we have
\[
E_\theta((p - E_\theta g)U) = 0,
\] (1.9)
therefore \( E_\theta(pU) = 0 \), and from (1.8) it follows that
\[
E_\theta[(p - g)U] = 0
\]
then (1.7) holds and \( p \) is the BUP of \( g \).