UCGE Reports
Number 20327

Department of Geomatics Engineering

A Study on Canonical Expansion of Random Processes with Applications in Estimation Problems

(URL: http://www.geomatics.ucalgary.ca/graduatetheses)

by

Zhan Zhang

January 2011

UNIVERSITY OF CALGARY

A Study on Canonical Expansion of Random Processes with Applications in Estimation Problems

by

Zhan Zhang

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF GEOMATICS ENGINEERING

CALGARY, ALBERTA

JANUARY, 2011

Abstract

Canonical expansion is an effective tool of studying the second-order random process by decomposing the process into an orthogonal expansion based on the information of the second moment. In essence, it is one of those techniques which can be categorised under the theory of orthogonal functions. The current study is devoted to applying this technique in the optimal estimation of random process according to the principle of Minimum Mean Square Error (MMSE), by constructing both the optimal linear and non-linear operators through the canonical expansion. The whole theory of canonical expansion is grounded on the theories of linear integral equation and linear algebra. The principle of MMSE results in the Wiener-Hopf equation for the linear estimation, and the regression operator in the non-linear case. Both the estimators can be constructed by the principal components that are generated through the canonical expansion. Numerical experiments show that such a method can give very accurate results for estimations of different time series. Also, the relation and comparison between the linear and non-linear operators are revealed through those numerical examples, in which the noise models are all Gaussian processes.

Acknowledgements

Firstly, I feel gratitude to my respectful supervisor Dr. J.A.R. Blais, for his help, support and guidance through the past years of supervision. His knowledge and decades of commitment and dedication to the geoscience will always be my inspiration. Thank-you as well to Dr. Wolfgang Keller, of the Universität Stuttgart, for clarifications on his work which is related to the present research.

Also, I would like to acknowledge my friends in the department Hang (Terry) Liu, Peng Xie, Feng Xu and Xing (Bob) Zhao, for their various help and aid to me during the completion of this work. Especially, sincere thanks go to my congenial friend Sina Taghvakish, for his help, assistance and discussions with me through the past months.

Finally, I would like to thank Dr. Michael Sideris, Dr. Xin Wang, and Dr. Paul Binding from the Department of Mathematics and Statistics, for serving on the Examining Committee.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Acronyms and Symbols

Acronyms

CDF	Cumulative Density Function
EOF	Empirical Orthogonal Function
FT	Fourier Transform
G-M	Gauss-Markov
JB	Jordan Block
K-L	Karhunen-Loève
LS	Least-Squares
MMSE	Minimum Mean Square Errors
PC	Principal Components
PCA	Principal Components Analysis
PDF	Probability Density Function
PSD	Power Spectrum Density
RW	Random Walk
SD	Spectral Decomposition
SVD	Singular Value Decomposition
WGN	White Gaussian Noise
W-H	Wiener-Hopf

Symbols

$(x,\ y)$	Stochastic Inner Product (Second Order Moment between Two Stochastic Processes x and y)
$\|\bullet\|$	Norm in Hilbert Space
A	Matrix of Principal Components
A_L	Matrix of Principal Components, with Lower Triangular Structure
$a_i(t)$	Principal Component (continuous case)
$c_i(s)$	Coefficients in Galerkin's Form
D	Diagonal Matrix of Variances of Uncorrelated Basis Variables
D_i	Variance of Each Uncorrelated Basis Variable
$E[X]$	Mathematical Expectation of X
e	Error of Estimation
e_i	Orthonormal Basis in Hilbert Space
$g^{(k)}(s,t)$	Weighting Function
h_{ik}	$\widetilde{\mathrm{A}}^{(k)}\xi_i$
I	Identity Matrix
K	Variance-Covariance Matrix of Process X
$K_X(t,t')$	Covariance Function of Process X
$K_{XY}(t,t')$	Cross-covariance Function of Two Processes X and Y
k	Intensity of White Noise
L	Linear Space of Estimators / Random Variables
$l\left(W(s), \widetilde{\mathrm{A}}Z(t)\right)$	Loss Function of a System

m_X	Mean Value of Process X
$\min(t, t')$	Minimum Function
$N(0,1)$	Standard Normal Distribution
q	$e^{-\beta\tau}$
Q_L	Lower Triangular and Hankel's Form of q
R	Uniform number on [0,1]
$\Re^H$	Space of Random Functions with a Hilbert Space Structure
s	Time of Estimation
T	Time Period of Observation
t	Time of Observation
U_i	Random Variable
V_i	Uncorrelated Basis Variable
$\vec{V}$	Vector of Uncorrelated Basis Variables
$W(s)$	True Signal
X	Data Matrix of a Vector of Random Processes
$X^0(t)$	Centred Random Function
$x_j(t)$	Optimal Coordinate Function
Z_v	Inner Product of Observation and Principal Component
$Z(t)$	Observation
$\widetilde{\mathrm{A}}$	Optimal Operator
$\widetilde{\mathrm{A}}^{(k)}$	Component in Linear Form of Operator
α_{ik}	Inner Product of $a_i(\tau)$ and $\xi_k(\tau)$
$\widetilde{\mathrm{B}}$	Arbitrary Operator
$\delta(t-s)$	Dirac Function
δ_{ij}	Kronecker Delta
H_i	Inner Product of Z_v and α_{vi}
η_i	Coefficient in Linear Regression of Signal
Λ	Diagonal Matrix of Eigenvalues of Variance-Covariance Matrix
λ_v	Eigenvalues of Variance-Covariance Matrix
μ_{ij}	Second Moment of Two Random Variables
ξ_i	Coefficient in Linear Regression of Observation
$\rho(s,t)$	Kernel of Linear Integral Operator
Σ	Matrix of Singular Values
σ^2	Variance of Noise
τ	Time / Time Lag
Φ	Matrix of Eigenvectors of Variance-Covariance Matrix
$\Phi(x)$	Gaussian Cumulative Density Function
$\varphi(s)$	Eigenfunctions of Covariance Function
χ_i	Coefficients in Linear Form of Operator
Ω	Matrix of Optimal Coordinate Functions

Chapter One: INTRODUCTION

The canonical expansion is a technique that entails an orthogonal decomposition of a centred stochastic (or random) process into a linear combination of uncorrelated random variables with deterministic functions as their coefficients. These coefficients contain the second moment information (covariance function) of the whole process. This method may have a better-known name as Karhunen-Loève (K-L) theorem, yet it will be explained later that the canonical expansion is more general than this K-L theorem. In addition, one variation of this approach is the famous Principal Components Analysis (PCA), which is a statistical tool involving a reduction of dimensionality by transforming a large correlated data set into only several uncorrelated principal components (Joilliffe, 2002). The canonical expansion stands accepted nowadays as an important technique in science and engineering.

In a wider sense, this technique ought to be regarded as one of the treatments when studying a random process, and from this point of view, it is parallel to the idea of Fourier transform and the concept of power spectrum of a time series, since both techniques entail one characterisation of the stochastic process. On the other hand, mathematically speaking, this expansion approach is nothing else but a particular case of the theory of orthogonal functions (other techniques such as Fourier Analysis, Spherical Harmonics and Wavelets Analysis are also special cases of this more general theory). As a result, the canonical expansion could be fairly deemed as the ‘Fourier Analysis’ of stochastic processes.

Among various applications of this method (especially the PCA), the present work will be mainly devoted to the usage of canonical expansion in both the Wiener-Kolmogorov theory of optimal linear filtering, and in a broader class the non-linear optimal filtering theory considered by Pugachev and Andreyev. The essence of such optimal filtering theories is to design an operating system to provide the best solution of each particular estimation problem, or in other words, to minimise or maximise the criterion that is chosen for optimality (Andreyev, 1969). For example, for a time series contaminated by some random noise, it is our goal to recover this useful signal as best as possible, according to some criterion of optimality, such as that the mean squares of estimation errors should be minimised. In this sense, the optimal filtering is a *synthesis* problem, namely, one has to strive to develop a systematic procedure as the realisation of a theoretically optimum design, out of all the other permissible operations, in accordance with the same chosen principle.

Following this idea, it will be demonstrated in the current thesis (a) how to obtain the optimal operators, primarily based on the principle of minimum mean square errors (MMSE), via the method of canonical expansion; and (b) how to apply this theory in different estimation problems in geodynamics and geomatics engineering. In general, we will see that with the help of canonical expansion (or principal components), it is possible to extract the probabilistic information of noise, and to apply it in the construction of either a linear Wiener-Kolmogorov or an optimal non-linear estimator. In the linear case, a *weighting function* (or matrix) will be produced from the canonical expansion, which acts as a linear transformation of the observation to the optimal estimation; while in the

latter case, the expansion form will be utilised to calculate the conditional probability density function (PDF) of the useful signal given an observation, in order to formulate the corresponding conditional expectation as the operator itself.

Before we commence our study, it does no harm to confine ourselves to the clarification of some terminologies that we will be continuously adopting throughout this thesis, so as not to arouse any inconsistency, ambiguity or misunderstanding of the material:

1. There is no differentiation between the names random function, random process, stochastic process or time series (it is to be noted that, as in some literatures, random process and stochastic process may refer to somewhat different meanings), they all represent a random-valued function whose argument is an evolving parameter (time).
2. The random process that is considered throughout the thesis is the second moment process, i.e., a full knowledge of the first and second moment information characterises the whole process.
3. The *covariance function*, which characterises the covariance (or correlation) between any pair of random variables in the process, is defined as:

$$K_X(t,t') = E\left[\left(X(t)-m_X\right)\left(X(t')-m_X\right)\right] \tag{1.1}$$

where m_X denotes the mean value of the stochastic process $X(t)$.

This is sometimes called the *auto-covariance function* in the literature, however, the term covariance function is preferably adopted in the present work.

When characterising the covariance between any pair of variables of two

processes, the *cross-covariance function* is used:

$$K_{XY}(t,t') = E\left[\left(X(t) - m_X\right)\left(Y(t') - m_Y\right)\right] \tag{1.2}$$

4. The theory presented in this work will focus solely on the case of one-dimensional random processes with only one parameter. The entire theory can be generalised into the case of multi-dimensional processes (or random vectors) without any apparent difficulty.
5. All the random variables and random functions treated in this thesis are real-valued.

1.1 Background and Literature Review

As we have mentioned before, the present work deals with the optimal estimation via the canonical expansion technique. As a result, it is expedient to provide an overview of the history of canonical expansion (or Karhunen-Loève expansion) and the PCA technique, followed by a brief survey of the chronological development of optimal estimation of stochastic processes in the past century. Only in such a way can we acquire a full picture of the background of the topic that we are about to develop in the next chapters.

1.1.1 History of Canonical Expansion and Principal Components Analysis

The earliest version of canonical expansion is in fact a monumental discovery in statistics. In 1901, Pearson published his pioneering work in curve fitting, in finding the best line representing a collection of correlated data, which marks the beginning of a century's study on the approach of Principal Components Analysis (Pearson, 1901; Joilliffe, 2002). The other acknowledged pioneer is Hotelling, whose idea was to generate

a ‘fundamental set of independent variables’ to determine the original variables in a data set. His method is to use Lagrange multipliers to transform it into a problem of eigenvalues (Joilliffe, 2002). There are also a number of scientists who have made important contributions to this technique; for a more thorough and complete account on this approach, one may consult Joilliffe (2002) for more details.

In addition, the PCA is closely related to a matrix decomposition technique called Singular Value Decomposition (SVD), as a means of obtaining the principal components directly from the observation matrix. It was originated by some eminent mathematicians such as Jordan, Beltrami, Sylvester, Schmidt and Weyl, through their immortal works both in bilinear forms and the algebraic theory of integral equations (Stewart, 1992). The study of its practical computation still remains one of the most vivid and fruitful topics in linear algebra today.

On the other hand, separately from the ideas of Pearson and Hotelling from the point of view of statistics, the same topic was approached to by Karhunen from a more analytical perspective in his 1947 paper ‘*Über lineare Methoden in der Wahrscheinlichkeitsrechunung*’ (literally, ‘On linear methods in the Calculus of Probability’), although it was claimed that a special form had already been obtained by Kolmogorov in 1942 (Pugachev, 1965). In this type of work, the Karhunen-Loève expansion was established based on the analytical theory of random functions with a close connection to the theory of orthogonal functions. Pugachev (1965) summarised these works into a systematic theory, for which he coined the name of canonical

expansion. In Pugachev's work, the K-L expansion is integrated into a more general form of expansion similar to the idea of Fourier expansion of an analytical function, which is optimal in a least-squares (LS) sense.

Nevertheless, it was not until the last decades that this technique became fairly popular in studying stochastic processes, due to the boom of high-speed computers that could handle the large amount of computations. Nowadays, this method has been adopted for various purposes: data compression, pattern recognition, system identification and control problems, to name a few. Especially in geophysics and geodynamics, besides the name PCA, it is also known as Empirical Orthogonal Function (EOF) Analysis. It is primarily used for extraction of the modes (which contain the probabilistic information) of a dynamic model from its observation, as a basis for prediction or interpretation purposes. In such studies, the EOFs are normally four-dimensional data-driven vectors in a spatial-temporal stochastic field (North, 1984).

1.1.2 Development of the Theory of Statistically Optimal Estimation

The modern theory of statistically optimal estimation was initiated by Wiener, Khintchine and Kolmogorov circa 1940. In its primary stage, the theory was closely related to the spectral theory (or correlation theory) of stationary random functions, developed independently by Wiener and Khintchine (Khintchine, 1934). Wiener and Kolmogorov developed their own optimal filters respectively from continuous time domain and discrete time domain analyses of stationary time series (Gelb et al., 1974). An important extension was made by Zadeh and Ragazzini on extending the signal to a random

function with a non-random polynomial in a finite domain, and a form of Wiener-Hopf equation was also developed (Zadeh and Ragazzini, 1950). Booton (1952) also provided a solution to the Wiener-Hopf equation when the input random function is non-stationary. The most famous technique was introduced by Kalman in 1960, on designing a recursive procedure ‘based on state-space, time domain formulations’ (Gelb et al., 1974). This technique is nowadays widely accepted as a standard treatment in filtering problems, and a tremendous amount of literature has been devoted to the studies on this topic.

Apart from Kalman filtering, another approach based on the Karhunen-Loève expansion was also developed by a number of mathematicians and scientists in order to explore the optimal estimation theory. Davis (1952) used this technique to solve the problem of optimal prediction in the non-stationary case. Extensive studies of applying this method in optimal estimation problems were carried out by Pugachev in a series of papers around 1960s, which marked an important progress in this direction. The method was also studied around the same time by Parzen (1963), who attempted to unify this approach with other techniques at that time under a more general treatment containing the former ones as special cases. Although less popular, this approach is still being studied nowadays in filtering and prediction. For some of these contributions, one may refer to Kudritskii (2000) and Shaikin (2007), and for a more historical and even philosophical overview, one may consult Dougherty (2009).

1.2 Thesis Objectives

The principal objectives of the thesis are to establish a more solid foundation for the method of canonical expansion, to construct both the linear and non-linear optimal operators by the principal components, and to show the application of such an approach in the MMSE estimation (Wiener-Hopf estimator) of some particular problems that are commonly seen in the field of geomatics engineering or geodynamics by numerical experiments. More specifically, the following points will be addressed:

1. to provide a comprehensive survey of the analytical theory of canonical expansion from different fields of mathematics, especially from the view of functional analysis and the theory of linear integral equations;
2. to apply the canonical expansion to formulate the optimal operator according to the principle of MMSE, by solving the Wiener-Hopf equation (linear case) and constructing the regression operator (non-linear case); and
3. to research the performance and computability of both operators in selected numerical examples, with different Gaussian noise models.

In all, it is the intention to enrich the repertoire of current approaches for studying the estimation problems of time series in geodynamics, geomatics engineering or geophysics. Moreover, the author hopes that the solutions offered in the proposed examples will serve as classical paradigms for the advanced research of more elegant problems, using this particular approach.

1.3 Thesis Outline

In Chapter two, the general theory of the canonical expansion will be established. Various terminologies used in the analytical theory of canonical expansion will be explained, such as coordinate functions, principal components, etc. Two approaches for finding the PCs will be investigated: the first way depends on the triangular decomposition of a matrix and a recursive generation; the second is the famous Karhunen-Loève theorem. Examples of canonical expansion of some common random processes involving the use of both approaches will be provided. Furthermore, we will bring this method under the scheme of functional analysis, by comparing it with Fourier and wavelets analysis, which also share the same structure of a Hilbert space.

Next, we will apply the method to constructing the optimal estimators. Chapter Three mainly deals with the linear estimation under the principle of MMSE. Initially, the principle of MMSE will be presented using the theory of Hilbert space. Following this idea, the MMSE principle will lead to Wiener-Hopf (W-H) equation containing the sought optimal linear operator. The rest of the chapter will be devoted to providing a detailed algorithm on how to obtain this linear operator, together with its error, by solving the W-H equation using Galerkin's method and principal components.

Chapter Four is a subsequent chapter on formulation of non-linear operator according to the same MMSE principle. The operator in this case is the regression of the signal on the observation. In parallel, a thorough description will be offered on how to obtain the regression using the principal components when the noise model is Gaussian. In addition,

a generalisation of this method for designing an optimal operator according to any principle of optimality will be outlined.

Chapter Five deals with the application of the foregoing theory to estimation problems in geophysics or geomatics engineering. Three examples will be provided, and in each case, either the linear and non-linear operators will be established, or their computability will be discussed. These examples are firstly intended to demonstrate how to apply the proposed procedure of estimation in the previous two chapters in each specific simulated data model, and also to investigate and compare the performance and effectiveness of both linear and non-linear operators in each case.

Chapter Six serves as a summary of the entire thesis. In addition, comments and recommendations, based on the conclusions from previous chapters, will be presented in order to cover the whole spectrum of topics and analyses throughout the thesis, regarding this particular technique.

Chapter Two: GENERAL THEORY OF CANONICAL EXPANSIONS OF RANDOM FUNCTIONS

2.1 Canonical Expansion of One-dimensional Random Functions

Let us consider now a centred random function of one parameter $X^0(t)$, i.e., a random function with its expectation subtracted. The idea of canonical expansion is to try to express this centred process by a linear form of uncorrelated random variables with zero mean values:

$$X^0(t) = \sum_i x_i(t)\, V_i \qquad (2.1.1)$$

Here, V_i stand for the uncorrelated variables, the set of which are called the *basis.* $x_i(t)$ are some deterministic functions, known as *coordinate functions*. Under appropriate mathematical conditions, such expansion can be understood as the projection of the centred random process onto the basis of uncorrelated zero-mean variables.

2.1.1 Coordinate Functions

From the assumption, we know that the vector $\vec{V} = \{V_1,\ V_2,\ \cdots\}^T$ is a basis of uncorrelated variables of zero means, then we can write:

$$E[V_i] = 0 \quad E[V_i\, V_j] = 0 \ \text{ when } i \neq j, \quad \text{and} \quad Var[V_i] = E[V_i^2] = D_i \qquad (2.1.2)$$

If we multiply by V_j on both sides of (2.1.1), and then calculate the expectation, we get:

$$E[X^0(t)\, V_j] = \sum_i E[V_i\, V_j]\, x_i(t)$$

By the conditions (2.1.2), we then can obtain our formula for the coordinate functions

$$x_j(t) = \frac{1}{D_j} E\left[X^0(t)\, V_j \right] \tag{2.1.3}$$

We can also prove that, by substituting (2.1.3) into (2.1.1), and calculate the remainder of any other possible representation of the form (2.1.1), this set of functions (2.1.3) gives the best approximation to the original centred random function in a least-squares (LS) sense. Hence, we can call those functions, satisfying the condition (2.1.3), the *optimal coordinate functions*. Nevertheless, such an optimal form is always hard to find in practice, and it is customary to use some sub-optimal functions (in the LS sense) to approximate them.

2.1.2 Construction of the Basis and the Principal Components

Having established the formula (2.1.3), we are ready to consider in this section a primary question: how to construct the basis $\vec{V}$, namely, those uncorrelated variables, from the random process itself.

Let $\overrightarrow{X^0(t_h)} = \left\{ X^0(t_1), \quad X^0(t_2), \quad \dots \quad X^0(t_n) \right\}^T$ represent an n-dimensional centred variable of a random process for different instants. The idea of the construction is to produce *n* linear forms of this set of values and use them as our basis $\vec{V}$. We can write this relation in matrix notation as follows:

$$\begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \bullet \begin{pmatrix} X^0(t_1) \\ \vdots \\ X^0(t_n) \end{pmatrix} \quad \text{or} \quad \vec{V} = A \cdot \overrightarrow{X^0(t_h)} \tag{2.1.4}$$

This matrix of coefficients is generally known as the *principal components* (PC) (Jolliffe, 2002), and it serves as a linear transformation of any aggregate of the random variables in the stochastic process into a set of uncorrelated variables.

Next, the task is to further determine the property of the yet arbitrary matrix A. This can be done by directly using the conditions (2.1.2), and after a simple calculation, we have an important result:

$$A \cdot K \cdot A^T = diag(D_1, \cdots D_n) = D \tag{2.1.5}$$

where K is the variance-covariance matrix of vector random variables $\overrightarrow{X^0(t_h)}$, and D is a diagonal matrix with the variances of each component of the basis (2.1.4) in a descending order on the main diagonal. This fact suggests that the symmetric, positive-definite real matrix K is similar to a diagonal matrix, and A is a non-singular matrix which could be determined in infinitely many ways. We will be discussing two possible choices of A in the next sections.

Consequently, we are able to determine the optimal coordinate function (2.1.3). By substituting (2.1.4) into (2.1.3), we have:

$$D \cdot \Omega = A \cdot K \tag{2.1.6}$$

where $$\Omega = \begin{pmatrix} x_1(t_1) & \cdots & x_1(t_n) \\ \vdots & \ddots & \vdots \\ x_n(t_1) & \cdots & x_n(t_n) \end{pmatrix}$$

Right-multiplying A^T on both sides of (2.1.6), and considering (2.1.5):

$$D \cdot \Omega \cdot A^T = A \cdot K \cdot A^T = D$$

which implies:

$$A \cdot \Omega^T = I \tag{2.1.7}$$

This important relation is called the *bi-orthogonality* of the optimal coordinate functions Ω and the principal components A. For a non-singular matrix A, this means that $\Omega^T = A^{-1}$. Correspondingly, from (2.1.5), we can express the variance-covariance matrix in terms of the optimal coordinate functions:

$$K = \Omega^T \cdot D \cdot \Omega \tag{2.1.8}$$

2.1.3 Integral Representations (continuous case)

All the results that we have obtained are in the discrete formulation based on the information on the variance-covariance matrix of the distinct points of the stochastic process. Yet, if the analytical form of covariance function is known to us a priori, then all the formulations should yield continuous functions, and the construction of the basis of uncorrelated variables is simply a generalisation of (2.1.4) to the integral case:

$$V_i = \int_T a_i(t) X^0(t) dt \tag{2.1.4a}$$

Consistently, the optimal coordinate functions in (2.1.6), the condition of bi-orthogonality (2.1.7) and the variance-covariance matrix expansion (2.1.9) can be brought into the integral forms. We put them here as follows:

$$x_j(t) = \frac{1}{D_j} \int_T a_j(s) K_X(t, s) ds \tag{2.1.6a}$$

$$\int_T a_i(t)\, x_j(t)\, dt = \delta_{ij} \tag{2.1.7a}$$

$$K_X(t, t') = \sum_i D_i\, x_i(t)\, x_i(t') \tag{2.1.8a}$$

Similarly, according to the theorem in last section, if the covariance function can be expanded as (2.1.8a), together with bi-orthogonality (2.1.7a), then the process can be expanded into the canonical form (2.1.1) as well. For some other necessary and sufficient conditions of existence of canonical expansion, one may refer to Dougherty, 2009. Also, regarding (2.1.7a), it is always possible to find the two bi-orthogonal systems if either system is composed of linearly independent functions (see Kantorovich & Akilov, 1964). We will use these formulae in this section later for calculating the canonical expansions of white noise, 1st order Gauss-Markov process and random walk in Section 2.4.

2.1.4 Recursive Formulae of Canonical Expansion

As we said earlier, there are many different ways of choosing the principal components in A . One of them is called *LDU Decomposition* (an important variant of the *LU Decomposition*; see Watkins, 2002). It involves a decomposition of a matrix into the product of three matrices: unit lower-triangular, diagonal and unit upper-triangular ones. By this LDU Decomposition, relation (2.1.5) can also be written as:

$$A_L \cdot K \cdot A_L^T = D \tag{2.1.9}$$

where $A_L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 1 \end{pmatrix}$

So (2.1.4) in this case has become:

$$\begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 1 \end{pmatrix} \bullet \begin{pmatrix} X^0(t_1) \\ X^0(t_2) \\ \vdots \\ X^0(t_n) \end{pmatrix} \quad \text{or} \quad \vec{V} = A_L \cdot \overrightarrow{X^0(t_h)} \qquad (2.1.10)$$

Because of the bi-orthogonality condition (2.1.7), we can reverse this equation, and express the observation points of the process in terms of the basis $\{V_i\}$:

$$\begin{pmatrix} X^0(t_1) \\ X^0(t_2) \\ \vdots \\ X^0(t_n) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ x_1(t_2) & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_n) & x_2(t_n) & \cdots & 1 \end{pmatrix} \bullet \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{pmatrix} \quad \text{or} \quad \overrightarrow{X^0(t_h)} = \Omega_L^T \cdot \vec{V} \qquad (2.1.11)$$

It is worth noting that, unlike the Karhunen-Loève expansion which we will be discussing soon, there is no reduction in dimensionality involved in this LDU technique: apparently, as the observation points increase one by one, we have to introduce every time a new variable from the basis $\{V_i\}$, that is to say, the total number of the uncorrelated variables should be as the same as that of the points in the process.

The transformation (2.1.11) in fact can help us develop a series of *recursive formulae* to calculate the variances D_j and coordinate functions $x_j(t)$: one can assume $V_1 = X^0(t_1)$ and D_1 is equal to the variance of this variable, and all the remaining terms can be determined directly by the definitions of variance and optimal coordinate function (2.1.3) in a recursive manner:

$$D_j = K_X(t_j, t_j) - \sum_{h=1}^{j-1} D_h \, x_h(t_j)^2 \tag{2.1.12}$$

$$x_j(t) = \frac{1}{D_j}\left[K_X(t, t_j) - \sum_{h=1}^{j-1} D_h \, x_h(t) \, x_h(t_j) \right] \tag{2.1.13}$$

We will use these formulae to provide a canonical expansion for the important first-order Gauss-Markov process in Section 2.4, where we can discover that such expansion can actually take a very brief form, even though the dimensionality is not reduced.

2.2 Canonical Expansion and Karhunen-Loève Theorem

The most common way of finding the matrix A of principal components is the well-known Karhunen-Loève (K-L) theorem, which is a widely used technique in various branches of science and engineering. In fact, the optimal coordinate function provided by (2.1.3) is more of a theoretical expression, generally hard to compute, and we therefore need to conceive some practical method to find a new sequence of coordinate functions which can be regarded as approximations to those optimal ones. This is the key idea lying behind the K-L theorem from the perspective of the theory of canonical expansion. In this section, we will examine this theorem and its relation with those results that we have arrived at in the foregoing sections.

Consider a *symmetric integral equation* in which the covariance function is the symmetric kernel (Schmidt, 1907; Blais, 1988):

$$\int_a^b K(t, s)\varphi(s)\,ds = \lambda\varphi(t) \tag{2.2.1}$$

we call the solution to this equation $\varphi(s)$ the *eigenfunction* of the *kernel* $K(s,t)$ belonging to the *eigenvalues* λ, subject to the condition:

$$\int_a^b \varphi_\mu(s)\,\varphi_\nu(s)\,ds = \delta_{\mu\nu} \tag{2.2.2}$$

$\varphi_\mu(s)$ and $\varphi_\nu(s)$ are eigenfunctions corresponding to different eigenvalues λ_μ and λ_ν. The eigenfunctions $\{\varphi_\nu(s)\}$ constitute a complete normalised system of the kernel function which is the covariance function in our context. From *Mercer's theorem* (although the result had been provided by Schmidt in his 1907 paper already), the covariance function in this case can be decomposed into a uniformly and absolutely convergent series of eigenfunctions:

$$K(t,s) = \sum_\nu \lambda_\nu\,\varphi_\nu(t)\,\varphi_\nu(s) \tag{2.2.3}$$

with the arranged descending order of positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots > 0$. Furthermore, once (2.2.3) holds true, then the stochastic process (as long as it is continuous and differentiable) can be expanded as (Schmidt, 1907):

$$X^0(t) = \sum_\nu \varphi_\nu(t)\,V_\nu \tag{2.2.4}$$

$$\text{with} \quad V_\nu = \int_T \varphi_\nu(s)\,X^0(s)\,ds \tag{2.2.5}$$

We can readily tell that (2.2.4) and (2.2.5) are exactly the canonical expansion (2.1.1) and the construction of uncorrelated basis (2.1.4a) when $a_\nu(t) = x_\nu(t) = \varphi_\nu(t)$. The expansion (2.2.4) is the formal expression of Karhunen-Loève theorem (Blais, 1988).

The discrete version (i.e., the matrix form) of the Mercer's theorem (2.2.3) is the *spectral decomposition* (SD) of the variance-covariance matrix K_X:

$$K_X = \Phi \cdot \Lambda \cdot \Phi^T \tag{2.2.6}$$

where Φ is the matrix of eigenvectors of K_X, and Λ is the diagonal matrix of the eigenvalues of K_X arranged in a descending order. Especially, for real-valued random variables, K_X is a real symmetric matrix, Φ is an orthogonal matrix:

$$\Phi \cdot \Phi^T = I \tag{2.2.7}$$

This result is in accordance with that we have obtained by the theory of linear integral equations (See (2.2.2)).

2.3 Further Discussion from the Standpoint of Functional Analysis

Suppose that random variables are represented as points in a linear space in which we define the inner product as their second-order moment:

$$\langle X, Y \rangle = E[XY] \tag{2.3.1}$$

This linear space of random variables with (2.3.1) as the inner product is a Hilbert space (Dougherty, 1999). We will generalise this idea to the case of stochastic processes.

Let all the stochastic processes (or random functions) constitute a linear space $\Re$, in which each random function is regarded as a point. We define the *inner product* of any two points $x = X(t)$ and $y = Y(t)$ as their second-order moment:

$$(x,\ y) = E[X(t)\,Y(t')] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\,y\,f(x, y; t,\ t')\,dx\,dy \tag{2.3.2}$$

where $f(x, y; t, t')$ is the joint PDF of any pair of random variables belonging to the processes $X(t)$ and $Y(t')$ respectively. It is expedient to note that, more generally, (2.3.2) should be defined for the complex-valued random functions as $E\left[X(t)\overline{Y(t')}\right]$; however, due to the fact that throughout the thesis no complex-valued process will be considered, for consistency of the symbols, we will restrict our discussion to the real-valued case.

To differentiate from the usual definition of inner product, we will call (2.3.2) the *stochastic inner product*, specifically referring to the second-order moment between any two random processes. Moreover, if both $X(t)$ and $Y(t')$ are centred processes, the stochastic inner product (2.3.2) is simply the *covariance function* (in the discrete case, it will be the *variance-covariance matrix*).

With the definition of (2.3.2), it can be verified the following properties hold true (in general, these properties can be extended for complex-valued stochastic processes):

$$\begin{aligned} &1)\quad (x,\ y) = (y,\ x);\\ &2)\quad (ax_1 + bx_2,\ y) = a(x_1,\ y) + b(x_2,\ y);\\ &3)\quad (x,\ x) \geq 0;\quad (x,\ x) = 0 \ \textit{iff}\ x = 0. \end{aligned} \tag{2.3.3}$$

With the stochastic inner product (2.3.2) and these properties, the space $\mathfrak{R}$ of the random functions has the structure of a Hilbert space, and we write $\mathfrak{R}^H$ denoting this fact (the completeness and separability of this space also need to be verified, however, this is beyond the scope of the thesis. For a simple discussion, see Dougherty, 1999).

If we normalise our basis vector $\vec{V} = \{V_1,\ V_2,\ \cdots\}^T$ by dividing each variable by the square root of its variance, we get the following condition:

$$E[e_i] = 0\,, \quad E[e_i\, e_j] = \delta_{ij}\,, \qquad \text{where} \quad e_i = \frac{1}{\sqrt{D_i}} V_i \tag{2.3.4}$$

It is easily seen that the new basis $\vec{e} = \{e_1,\ e_2,\ \cdots\}^T$ is an *orthonormal basis*, and the optimal coordinate functions given by (2.1.3), according to (2.3.2), can be written as:

$$x_j(t) = \frac{1}{D_j} E[X^0(t)\, V_j] = \frac{1}{\sqrt{D_j}} (x^0,\ e_j) \tag{2.3.5}$$

Correspondingly, the canonical expansion (2.1.1) can be arranged as:

$$X^0(t) = \sum_i x_i(t)\, V_i = \sum_i (x^0,\ e_i) \frac{V_v}{\sqrt{D_v}} = \sum_i (x^0,\ e_i)\, e_i \tag{2.3.6}$$

Thus, (2.3.5) and (2.3.6) suggest that the optimal coordinate functions are the *Fourier coefficients* with respect to the orthonormal basis (2.3.4). In other words, the canonical expansion is the Fourier expansion of a random process in the space with the stochastic inner product (2.3.2), so it is the best approximation in the LS sense. This conclusion agrees with what we stated in Section 2.1.1.

Table 2-1 presents a comparison among the classical Fourier analysis, Haar wavelets and the current method of canonical expansion. They all have the same structure of a Hilbert space: while both the Fourier and Wavelets analysis emphasise the representation of functions, the canonical expansion adopts the exactly parallel idea in the treatment of

second-order random processes. From this perspective, we can comprehend the method of canonical expansion as the Fourier analysis of the stochastic process.

	Fourier Analysis	Haar Wavelet	Canonical Expansion
Elements	Continuous or finite jump functions with period 2π	Continuous functions over $[0,1]$	2^{nd} order random process
Inner Product	$\int_0^{2\pi} f(t)\,g(t)\,dt$	$\int_0^1 f(t)\,g(t)\,dt$	$E\left[X(t)\,Y(t')\right]$
Orthonormal Basis	$\left\{\frac{1}{\sqrt{2\pi}}e^{ivt}\right\}$ $v=0,\pm1,\pm2,\ldots$	$\left\{\chi_n^{(k)}(s)\right\}$ $k=1,2,\ \ldots 2^{n-1}$	$\left\{\frac{V_v}{\sqrt{D_v}}\right\}$ $v=1,2,\ \ldots$
Fourier Coefficients	$\frac{1}{2\pi}\int_0^{2\pi} f(t)\cdot e^{-ivt}\,dt$	$\int_0^1 f(t)\,\chi_n^{(k)}(t)\,dt$	$x_j(t)=\frac{1}{D_j}E\left[X^0(t)\,V_j\right]$

Table 2-1. Comparison of Fourier Analysis, Haar Wavelets and Canonical Expansion, from the Structure of Hilbert Space

2.4 Examples of Expansion

Now, we will use our formulae which have been obtained so far to generate the canonical expansions of several random processes, including white noise, 1^{st} order Gauss-Markov (G-M) process and random walk (or Wiener process), since most random phenomena in geomatics engineering or geodynamics can be modeled by these processes. Once we have the expansion forms, the principal components are immediately known.

2.4.1 White Noise

The white noise $W(t)$ is a pure theoretical concept, describing a process with uncorrelated values and infinite dispersions and a flat spectrum. Its mathematical expectation and covariance function are defined as:

$$E_W(t) \equiv 0\,, \qquad K_W(t,t') = k \cdot \delta(t-t') \tag{2.4.1}$$

k is the constant intensity of the white noise, and $\delta(t-t')$ is Dirac's delta function.

We will now expand the white noise on the interval $[0,T]$ into the canonical form. The condition of bi-orthogonality (2.1.7a) are satisfied by the functions:

$$a_v(t) = e^{i\omega_v t}\,, \qquad x_v(t) = \frac{1}{T}\, e^{i\omega_v t} \tag{2.4.2}$$

where $i = \sqrt{-1}\,, \quad \omega_v = \dfrac{2\pi v}{T}, \qquad v = 0, \pm 1, \pm 2, \ldots$

and from (2.1.6a), we can determine the variances of each variable in the basis:

$$D_v = \frac{k\int_T e^{i\omega_v s}\,\delta(s-t)\,ds}{\dfrac{1}{T}e^{i\omega_v t}} = kT \tag{2.4.3}$$

Therefore, the canonical expansion of the white noise process and its covariance function over the region $[0,T]$ are given by:

$$W(t) = \frac{1}{T}\sum_v e^{i\omega_v t}\, V_v \tag{2.4.4}$$

$$K_W(t,t') = \frac{k}{T}\sum_v e^{i\omega_v (t-t')} \tag{2.4.5}$$

Therefore, it can be readily seen that if the process is a white noise, its principal components are nothing but the Fourier basis –- sine and cosine functions. Also, for the white noise with (2.4.1), according to Khintchine's theorem, its *power spectral density* (PSD) is a constant value. By (2.4.3), the variances of uncorrelated basis for white noise are the same constant kT , which is proportional to the PSD.

2.4.2 First-order Gauss-Markov Process

The first-order Gauss-Markov (G-M) model depicts a process whose joint distribution of any number of variables is a Gaussian one and whose conditional distribution of any variable given a set of variables of previous instants depends solely on the variable preceding it. Its covariance function has the form:

$$K_X(t,t') = \sigma^2 e^{-\beta|t-t'|} \tag{2.4.6}$$

where σ^2 is the variance of each variable of the process , β is a parameter which is to be estimated from a large sample data. From this covariance function, it is possible to construct two kinds of PCs by utilising different approaches provided in previous sections, namely, the *K-L theorem* and the *LDU decomposition*.

2.4.2.1 Canonical expansion using Karhunen-Loève theorem

Firstly, we bring the covariance function (2.4.6) into the equation (2.2.1), and get:

$$\lambda_v \varphi_v(t) = \int_0^T \sigma^2 e^{-\beta|t-s|} \varphi_v(s) ds \tag{2.4.7}$$

The solution to this equation are the sine functions of the form (Pugachev, 1965):

$$\varphi_v(t)=\sqrt{\frac{2}{T+\lambda_v}}\sin\left[\omega_v\left(t-\frac{T}{2}\right)+\frac{v\pi}{2}\right]\ , \qquad v=1,2,\ldots \tag{2.4.8}$$

where the eigenvalues are $\lambda_v=\sigma^2\frac{2\beta}{\beta^2+\omega_v^2}$, $v=1,2,\ldots$

and $\{\omega_v\}$ are the positive roots of this equation with an ascending order:

$$\tan\omega T=-\frac{2\beta\omega}{\beta^2-\omega^2} \tag{2.4.9}$$

Next, based on the bi-orthogonality (2.1.7a) and the eigenfunction (2.4.8), we can write our coordinate functions as:

$$x_v(t)=\frac{2}{T+\lambda_v}\sin\left[\omega_v\left(t-\frac{T}{2}\right)+\frac{v\pi}{2}\right] \qquad v=1,2,3,\ldots \tag{2.4.10}$$

The variances of each uncorrelated variable of the basis are:

$$D_v=\frac{\sigma^2}{2}\lambda_v(T+\lambda_v) \tag{2.4.11}$$

Finally, the expansions of the 1st order G-M process (centred) and its covariance functions are:

$$X^0(t)=\sum_v\sin\left[\omega_v\left(t-\frac{T}{2}\right)+\frac{v\pi}{2}\right]\frac{2V_v}{T+\lambda_v} \qquad v=1,2,3,\ldots \tag{2.4.12}$$

$$K_X(t,t')=2\sum_v\frac{\lambda_v}{T+\lambda_v}\sin\left[\omega_v\left(t-\frac{T}{2}\right)+\frac{v\pi}{2}\right]\sin\left[\omega_v\left(t'-\frac{T}{2}\right)+\frac{v\pi}{2}\right] \tag{2.4.13}$$

2.4.2.2 *Canonical expansion for equidistant observation*

It is possible to avoid the complexity of the foregoing computation in the K-L expansion by utilising the recursive relations we have built up in Section 2.1.3. The procedure is allegedly due to Akimov (Pugachev, 1965).

Supposing that the observations are given at $n+1$ equidistant points $t_i = (i-1)\tau$, we start with the assumption that $V_1 = X^0(t_1)$ and $Var[V_1] = \sigma^2$, and the first coordinate function is given by (2.1.13) in Section 2.1.4:

$$x_1(t) = e^{-\beta|t|} \tag{2.4.14}$$

For brevity, we denote $q = e^{-\beta\tau}$, and after successively applying the formulae (2.1.12) and (2.1.13), the rest of the variances and coordinate functions are given by:

$$D_j = \sigma^2(1-q^2) \qquad j = 2, 3, \ldots \tag{2.4.15}$$

$$x_j(t) = \frac{1}{1-q^2}\left[e^{-\beta|t-(j-1)\tau|} - qe^{-\beta|t-(j-2)\tau|}\right] \qquad j = 2, 3, \ldots \tag{2.4.16}$$

Consequently, the coordinate function for each observation point is:

$$x_j(t_i) = q^{i-j} \tag{2.4.17}$$

The canonical expansion of the $n+1$ observational points reads:

$$\begin{pmatrix} X^0(t_1) \\ X^0(t_2) \\ X^0(t_3) \\ \vdots \\ X^0(t_{n+1}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ q & 1 & 0 & \cdots & 0 \\ q^2 & q & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q^n & q^{n-1} & q^{n-2} & \cdots & 1 \end{pmatrix} \bullet \begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ \vdots \\ V_{n+1} \end{pmatrix}$$

$$\text{or} \quad \overrightarrow{X^0(t_h)} = Q_L \cdot \vec{V} \tag{2.4.18}$$

The form of Q_L has the structure of both a lower triangular form and *Hankel's form* (Kowalewski, 1909). From (2.1.10), the PCs are given by the inverse of the coefficient matrix in this expression (Pugachev, 1965):

$$A_L = Q_L^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -q & 1 & \cdots & 0 & 0 \\ 0 & -q & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -q & 1 \end{pmatrix} \tag{2.4.19}$$

We have thus discovered that the principal components, in the case of 1st order G-M process, can be selected as a *lower Jordan block* (Gantmacher, 1959).

2.4.3 Random Walk

Random Walk (RW) process characterises a random function whose derivative is white noise (Nassar, 2003), viz.:

$$\frac{dX(t)}{dt} = W(t), \qquad \text{with} \quad X(0) = 0 \tag{2.4.20}$$

The solution is straightforward:

$$X(t) = \int_0^t W(s)\,ds \tag{2.4.21}$$

As before, in order to obtain the expansion form of the RW process, firstly we need to find its covariance function. From (2.4.21), we can readily tell that the mathematical expectation of a RW process is identically zero; and recalling the covariance function of white noise defined by (2.4.1), we use (2.4.20) to calculate the covariance function for RW process as:

$$K_X(t, t') = k \cdot \min\left(t,\ t'\right) \tag{2.4.22}$$

It is worthy to note that unlike the previous two cases, this covariance function is no longer stationary but a linear function of the time, where the variances of the variables in the RW process increases in the same order as the time, as it evolves.

Now we turn to the expansion of such a covariance function on a given interval. It is helpful to write $(s-T, s)$ instead of $(0, T)$ for the sake of convenience of computation.

In order to get the exact expression for the expansion, we have to transform the covariance function (2.4.22) into a form similar to that of the 1st order G-M process (2.4.6), by introducing an auxiliary variable (Pugachev, 1965):

$$u = \frac{1}{2a}\ln\left(\frac{t}{s-T}\right), \qquad u \in [0,\ U], \quad \text{where} \quad U = \frac{1}{2a}\ln\left(\frac{s}{s-T}\right) \tag{2.4.23}$$

where a is an arbitrary positive constant.

By such introduction, the covariance function (2.4.22) is transformed to:

$$K_X(u, u') = k\, f(u)\, f(u)\, e^{-a|u-u'|} \tag{2.4.24}$$

where $f(u) = \sqrt{t} = \sqrt{s-T}\ e^{au}$

Hence, the covariance function of the RW process has been interpreted in terms of the auxiliary variable (2.4.23), possessing a form similar to that of the 1st order G-M process (2.4.6). It is then quite natural to assume that the coordinate functions and consequently the canonical expansion of RW process should resemble formulae (2.4.10) and (2.4.12). Having brought back the original variable t, we have determined the expansion of the RW process on the interval $(s-T, s)$ as

$$X(t)=\sum_{v}\sin\left[\omega_v\alpha(t)+\frac{v\pi}{2}\right]\frac{2\sqrt{t}}{U+\lambda_v}V_v\ ,\qquad v=1,2,3,\ldots \tag{2.4.25}$$

where $\alpha(t)=\frac{1}{2a}\ln\left[\frac{t}{\sqrt{(s-T)s}}\right]$, $\lambda_v=k\,\frac{2a}{a^2+\omega_v^2}$

Similar to (2.4.9), $\{\omega_v\}$ are the positive roots of the equation:

$$\tan(\omega U)=-\frac{2a\omega}{a^2-\omega^2} \tag{2.4.26}$$

The variances of the uncorrelated variables are given by:

$$D_v=\frac{1}{2}\lambda_v(U+\lambda_v) \tag{2.4.27}$$

Accordingly, the expansion of the covariance function has the form:

$$K_X(t,t')=2\sum_{v}\frac{\lambda_v}{U+\lambda_v}\sin\left[\omega_v\alpha(t)+\frac{v\pi}{2}\right]\sin\left[\omega_v\alpha(t')+\frac{v\pi}{2}\right]\sqrt{t\,t'} \tag{2.4.28}$$

The meaning of this example lies in the fact that even if the process is expressed as a relation of a first-order differential equation, it is still possible to obtain the analytical form of its canonical expansion and correspondingly the principal components.

2.4.4 Further Remarks

Those three examples above are the most common processes (especially the noise models) that we usually deal with in practice. This means that, in most practical problems we can obtain a canonical form, and consequently its PCs, for the considered process. As we will demonstrate in the next two chapters, this information is critical to formalise the linear and non-linear filters for estimation purposes. Hence, as long as we have a

complete knowledge of the noise model and its parameters, we are able to tackle the estimation problems from the perspective of the canonical expansions.

It is also possible, that in reality, the actual noise is a combination of several uncorrelated processes. For example, it could be a sum of white noise and 1st order G-M process, or of several G-M processes with different parameters. If this be the case, our covariance function is definitely a linear combination of these uncorrelated processes:

$$K_X(t,t') = \sum_i b_i \; K_i(t,t') \tag{2.4.29}$$

The PCs of this mixed process should be a linear combination of the PCs for each component covariance function $K_i(t,t')$, once the parameters for each process and the coefficients b_i are known. Otherwise, we should estimate the empirical variance-covariance matrix from a large sample data of this combined process, and perform the spectral decomposition (SD) (2.2.6) to obtain the PCs.

2.5 Comments on Canonical Expansion of Vector Random Functions – Singular Value Decomposition

Finally, we will give some comments on the canonical expansion of a vector of random functions $\{X_1(t),\ X_2(t),\ \cdots,\ X_M(t)\}^T$, since in many practical problems it is necessary to consider several random processes simultaneously rather than a single one. In this case, the foregoing theory of canonical expansion holds true for each component of the random vector, so that the vector can be projected on the basis of uncorrelated variables:

$$\begin{pmatrix} X_1(t) \\ X_2(t) \\ \vdots \\ X_M(t) \end{pmatrix} = \begin{pmatrix} x_{11}(t) & x_{12}(t) & \dots & x_{1N}(t) \\ x_{21}(t) & x_{22}(t) & \dots & x_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1}(t) & x_{M2}(t) & \dots & x_{MN}(t) \end{pmatrix} \bullet \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{pmatrix} \tag{2.5.1}$$

The principal components of the random vector can therefore be found through the condition of bi-orthogonality (2.1.7) for every entry in the matrix in (2.5.1). All the discussion in the previous sections are still valid here.

In practice, the stationary random vector is usually given in terms of discrete observations, and this results in an $M \times N$ data matrix. The PCs can be algebraically found by the approach of Singular Value Decomposition (SVD) of this data matrix:

$$X = U \cdot \Sigma \cdot V^T \tag{2.5.2}$$

where X is the centred data matrix, Σ is an $M \times N$ diagonal matrix, U and V are orthogonal (or unitary) matrices. The SVD is related to spectral decomposition (SD) of the variance-covariance matrix in the following manner:

$$X \cdot X^T = U \cdot \Sigma \cdot V^T \cdot V \cdot \Sigma^T \cdot U^T = U \cdot (\Sigma \cdot \Sigma^T) \cdot U^T \tag{2.5.3}$$

Therefore, as is shown by this relation, the column vectors of U are PCs of the random vector, and the singular values of the matrix Σ are the square roots of the eigenvalues of the variance-covariance matrix (Blais, 2010).

Chapter Three: DETERMINATION OF THE LINEAR OPTIMAL ESTIMATOR BY THE CANONICAL EXPANSION METHOD

In estimation problems, it is always the central task to optimally recover the true signal from the observations in which the signal is contaminated by some random noise. In the parlance of mathematics, this is equivalent to defining an operator on the observations, such that the result of this operation is as close as possible to the true signal:

$$e(s,\ t) = W(s) - \widetilde{\mathrm{A}}Z(t) = \min \tag{3.1}$$

Here, the true signal is represented by $W(s)$, which is a random function. $\widetilde{\mathrm{A}}Z(t)$ stands for an operator $\widetilde{\mathrm{A}}$ operating on the observation $Z(t)$, and we will call it the *estimator*. Throughout the thesis, we will solely consider the *additive model* of observation, i.e., a second-order random noise superimposed on the real signal.

The criterion (3.1) encompasses all the practical problems of *extrapolation (prediction), interpolation, filtering (smoothing)* and *detection*, depending on the intersection of the aggregate of observational points t over a period $[a, a+T]$, and that of instants s for the signal. Based on this idea, we will do a classification of these problems, according to Middleton (1963) and Parzen (1963), in Table 3-1, of which the first column enlists various kinds of estimation problems, while the second column shows the relation between the fore-mentioned two aggregates of instants.

Therefore, for various kinds of optimal estimation problems, they can always be unified into the study of equation (3.1). In the present chapter, we will adopt the principle of

ESTIMATION PROBLEMS	**RELATION BETWEEN $\{s\}$ AND OBSERVATION PERIOD $[a,\ a+T]$**
Extrapolation	$\{s < a\} \cup \{s > a+T\}$
Prediction	$\{s > a+T\}$
Filtering / Detection / Interpolation	$s \in [a,\ a+T]$
Filtering & Prediction	$\{s \in [a,\ a+T]\} \cup \{s > a+T\}$

Table 3-1. Extrapolation, Prediction, Filtering, Interpolation and Detection, with their Attributes

minimum mean square errors (MMSE) as the criterion of the optimality, to approach the estimation problem when the sought operator $\widetilde{A}$ is linear. In this case, the principle of MMSE will lead us to the famous *Wiener-Hopf (W-H) equation*, and the solution of this W-H equation will produce an optimal system (the operator $\widetilde{A}$) of estimation whose input is observation. There are various approaches to solve this equation under different assumptions, for instance, Sanso and Sideris (1997) tackled this equation by the method of Fourier Transform (FT), so that a parallel optimal linear system can be established in the frequency domain. Here, we will solve this equation adopting *Galerkin's form* of solution by the method of canonical expansion (Pugachev, 1965). A similar approach of solving this equation using Galerkin's method is proposed by Keller (2000), where he provides the solution by selecting the Haar wavelets as the basis functions for the

Galerkin form. This idea is essentially parallel to what we are about to develop in next few sections, yet we will use the principal components (or *Karhunen-Loève basis*, Keller, 2004) as the basis functions, and a linear regression model for the signal.

As Blais (1979) suggests when considering the LS estimation, it is more appropriate to introduce a more abstract functional model for the purpose of mathematical analysis. We will therefore demonstrate the problems of optimal estimation under the scheme of functional analysis and the theory of operators.

3.1 Principle of MMSE Estimation

In Section 2.3, we have noted that the linear space of all random processes, with the stochastic inner product defined by the second order moment (2.3.1), is a Hilbert space. Since the Hilbert space is a *normed space*, the distance between any two points can then be measured in terms of this inner product. In addition, let L be a linear subspace of $\mathfrak{R}^H$, consisting of those random functions which are potentially the estimators of the process $w = W(s)$. The MMSE estimation problem is then reduced to finding the best estimator $\hat{w} = \hat{W}(s)$ within L which has the closest distance to w (remember that w is not necessarily in L). From the projection theorem of Hilbert space, we know that each element in $\mathfrak{R}^H$ has a unique representation (Collatz, 1964; Kantorovich & Akilov, 1964; Lusternik & Sobolev, 1974):

$$w = \hat{w} + e, \qquad \text{where } \hat{w} \in L \text{ and } e \perp L \tag{3.1.1}$$

Here, we call $\hat{w}$ the projection of w on the subspace L. Therefore, the condition of the optimality is equivalent to the perpendicular condition in (3.1.1):

$$\left(e,\ h\right)=\left(w-\hat{w},\ h\right)=0\,,\qquad \forall\ h\in L \tag{3.1.2}$$

Now, for an observation $z=Z(t)$ in $\Re^H$, it is then necessary to find the optimal operator $\tilde{\mathrm{A}}$ that carries z to $\hat{w}$, and (3.1.2) implies:

$$\left(w-\tilde{\mathrm{A}}z,\ \tilde{\mathrm{B}}z\right)=0 \tag{3.1.3}$$

where $\tilde{\mathrm{B}}$ is any operator which takes the observation z to another estimator in L. Hence, any operator $\tilde{\mathrm{A}}$ which satisfies the condition (3.1.3) is the optimal one which we are looking for, according to the MMSE criterion. This relation can be adequately illustrated by Figure 3-1.

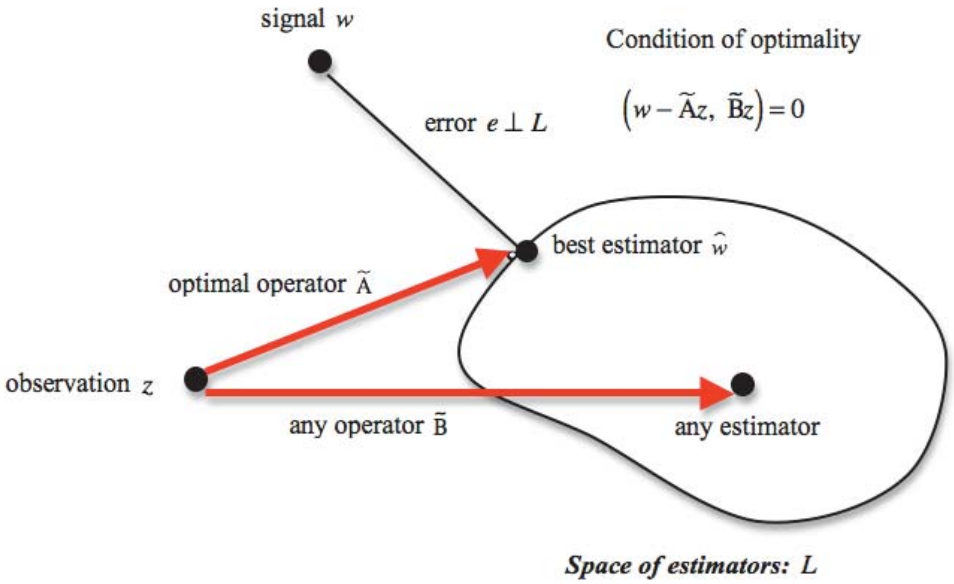

Figure 3-1. Principle of Minimum Mean Square Estimation using the Projection Theorem of Hilbert Space of Stochastic Processes

Hence, under the principle of MMSE, we have built up the optimality condition for any operator using the projection theorem of Hilbert space, so far independent of any statistical property of either the signal or the observation. Throughout this chapter, we will study the formulation of this optimal operator when it is linear; the discussion of the case when it is non-linear will be found in the next chapter.

3.2 Optimal Linear Operator and Wiener-Hopf equation

In this section, we will formalise the expression of the linear optimal operator $\widetilde{\mathrm{A}}$, when both the signal and the observation can be linearised. Such an estimator is called a *Wiener filter*. Under these assumptions, it will be seen that the sought operator is a linear combination of linear integral operators. In order to find their kernels, we will use the method of canonical expansion (or the principal components). In the following, we will derive a systematic approach to complete the solution of this problem (Pugachev, 1965).

3.2.1 Wiener-Hopf Equation and Wiener Filter

It is already seen from the principle of MMSE estimation that any operator which gives the best estimation given an observation must obey the condition (3.1.3). Keeping in mind that all the operators are now linear, which means that the operation of mathematical expectation can be interchanged with that of linear operators, we can do such rearrangement of (3.1.3) using the basic properties of the second-order moment:

$$\begin{aligned}&\left(\widetilde{\mathrm{A}}z - w,\ \widetilde{\mathrm{B}}z\right) = \left(\widetilde{\mathrm{A}}z,\ \widetilde{\mathrm{B}}z\right) - \left(w,\ \widetilde{\mathrm{B}}z\right) = \widetilde{\mathrm{B}}^{*}\left(\widetilde{\mathrm{A}}z,\ z\right) - \widetilde{\mathrm{B}}^{*}\left(w,\ z\right) = 0 \\ &\Rightarrow \widetilde{\mathrm{B}}^{*}\left\{\left(\widetilde{\mathrm{A}}z,\ z\right) - \left(w,\ z\right)\right\} = 0\end{aligned} \tag{3.2.1}$$

The necessary condition of the vanishing of the left side of this equation, for it should be satisfied for any operator (since $\tilde{\mathrm{B}}$ is an arbitrary operator), is given by:

$$\tilde{\mathrm{A}}(z,\ z) = (w,\ z) \tag{3.2.2}$$

This is the famous *Wiener-Hopf equation* (Blais, 1988; this is also coined with the name *Wiener-Kolmogorov equation*; see Keller 2000 & 2004). Any operator satisfying this equation is called a *Wiener filter*, which minimises the mean-square errors (see Keller, 2004). Next, in order to solve (3.2.2), we will perform two stages of linearisation: the first on signals w and z, and the second on the operator $\tilde{\mathrm{A}}$ (Pugachev, 1965). In this chapter, we will assume that $\tilde{\mathrm{A}}$ is represented by a linear integral operator, so the estimated signal $\widehat{W}(s)$ will be the result of this operator applied to the observation $Z(t)$:

$$\widehat{W}(s) = \tilde{\mathrm{A}}Z(t) = \int_T \rho(s,t)Z(t)\,dt \tag{3.2.3}$$

where $\rho(s,t)$ stands for the weighting function of this integral operator.

3.2.1.1 Linear regression of signals

Let $L = span\left\{\ U_1,\ \ U_2\ \ \ldots\ \ U_n\ \right\}$ be a subspace spanned by random variables $\{U_i\}$, so any element in L can be written as a linear combination of these elements. We use the *projection theorem* to represent both the observation z and the signal itself w as a sum of their projections on the spanned space L and the elements perpendicular to L:

$$\begin{aligned} z &= \sum_i^n \xi_i\, U_i + x \\ w &= \sum_i^n \eta_i\, U_i + y \end{aligned} \tag{3.2.4}$$

where $(x,\ U_i)=0$ and $(y,\ U_i)=0$

We can always assume that x and y are some zero-means processes without loss of generality. In the parlance of statistics, these are simply the *linear regression models* of the signal and observation. With the aid of these decompositions, and using the condition of orthogonality, we calculate the W-H equation (3.2.2) as:

$$\begin{aligned}
&\widetilde{\mathrm{A}}(z,\ z)=(w,\ z)\\
&\Rightarrow \widetilde{\mathrm{A}}\left(\sum_i^n \xi_i\, U_i + x,\ \sum_j^n \xi_j\, U_j + x,\right)=\left(\sum_i^n \eta_i\, U_i + y + y,\ \sum_j^n \xi_j\, U_j + x\right)\\
&\Rightarrow \sum_i^n\sum_j^n \widetilde{\mathrm{A}}\xi_i\,\xi_j\left(U_i,\ U_j\right)+\widetilde{\mathrm{A}}(x,\ x)=\sum_i^n\sum_j^n \eta_i\,\xi_j\left(U_i,\ U_j\right)+(y,\ x)\\
&\Rightarrow \widetilde{\mathrm{A}}(x,\ x)=\sum_i^n\sum_j^n(\eta_i-\widetilde{\mathrm{A}}\xi_i)\,\mu_{ij}\,\xi_j+(y,\ x)
\end{aligned} \tag{3.2.5}$$

where for convenience, the second moment $\left(U_i,\ U_j\right)$ is denoted by μ_{ij}. By setting $\chi_j=\sum_i^n \mu_{ij}(\eta_i-\widetilde{\mathrm{A}}\xi_i)$, the W-H equation is now transformed into an equivalent set of $n+1$ equations of n unknown parameters and an unknown operator:

$$\begin{cases}
\widetilde{\mathrm{A}}(x,\ x)=\sum_j^n \chi_j\,\xi_j+(y,\ x)\\
\chi_j=\sum_i^n \mu_{ij}(\eta_i-\widetilde{\mathrm{A}}\xi_i)
\end{cases}
\qquad i,j=1,2,\ \ldots\ n \tag{3.2.6}$$

3.2.1.2 Linearisation of operator

The structure of the first equation in (3.2.6) implies a linearisation of the operator:

$$\widetilde{\mathrm{A}} = \widetilde{\mathrm{A}}^{(0)} + \sum_{k=1}^{n} \chi_k \, \widetilde{\mathrm{A}}^{(k)} \tag{3.2.7}$$

Having brought this form into the first equation of (3.2.6), we can write:

$$\left(\widetilde{\mathrm{A}}^{(0)} + \sum_{k=1}^{n} \chi_k \, \widetilde{\mathrm{A}}^{(k)} \right)(x,\ x) = \sum_{j}^{n} \chi_j \, \xi_j + (y,\ x)$$

$$\Rightarrow \widetilde{\mathrm{A}}^{(0)}(x,\ x) + \sum_{k=1}^{n} \chi_k \, \widetilde{\mathrm{A}}^{(k)}(x,\ x) = (y,\ x) + \sum_{j}^{n} \chi_j \, \xi_j$$

Comparing both sides of the identity, and equating the corresponding terms, the first equation of (3.2.6) has now transformed into a system of $n+1$ equations of $n+1$ unknown linear operators:

$$\begin{cases} \widetilde{\mathrm{A}}^{(0)}(x,\ x) = (y,\ x) \\ \widetilde{\mathrm{A}}^{(k)}(x,\ x) = \xi_k \qquad k = 1, 2, \ldots, n \end{cases} \tag{3.2.8}$$

These equations share the common structure of *Fredholm equation of the first kind*, and we will return to their solutions in the next section.

We have yet to fix those parameters $\{\chi_k\}$ from the remaining n equations in (3.2.6). For brevity, we write this system in matrix form:

$$\begin{pmatrix} \chi_1 \\ \vdots \\ \chi_n \end{pmatrix} = \begin{pmatrix} \mu_{11} & \cdots & \mu_{n1} \\ \vdots & \ddots & \vdots \\ \mu_{1n} & \cdots & \mu_{nn} \end{pmatrix} \bullet \begin{pmatrix} \eta_1 - \widetilde{\mathrm{A}}\xi_1 \\ \vdots \\ \eta_n - \widetilde{\mathrm{A}}\xi_n \end{pmatrix} \tag{3.2.9}$$

where the matrix $\{\mu_{ij}\}$ is the matrix of the second moment of any pair of variables U_i and U_j. Especially, if $\{U_i\}$ are centred variables, the matrix $\{\mu_{ij}\}$ is the *covariance matrix* of the variables $\{U_i\}$.

Substitute the operator $\widetilde{\mathrm{A}}$ in this equation by (3.2.7). For each element $\widetilde{\mathrm{A}}\xi_i$, we have:

$$\widetilde{\mathrm{A}}\xi_i = \left(\widetilde{\mathrm{A}}^{(0)} + \sum_{k=1}^{n} \chi_k \widetilde{\mathrm{A}}^{(k)}\right)\xi_i = \widetilde{\mathrm{A}}^{(0)}\xi_i + \sum_{k=1}^{n} \chi_k \left(\widetilde{\mathrm{A}}^{(k)}\xi_i\right) = h_{i0} + \sum_{k=1}^{n} \chi_k h_{ik}$$

where for brevity we write $h_{ik} = \widetilde{\mathrm{A}}^{(k)}\xi_i$. In matrix form, this is:

$$\begin{pmatrix} \widetilde{\mathrm{A}}\xi_1 \\ \vdots \\ \widetilde{\mathrm{A}}\xi_n \end{pmatrix} = \begin{pmatrix} h_{10} \\ \vdots \\ h_{n0} \end{pmatrix} + \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nn} \end{pmatrix} \bullet \begin{pmatrix} \chi_1 \\ \vdots \\ \chi_n \end{pmatrix} \tag{3.2.10}$$

In addition, we can prove that the matrix $\{h_{ik}\}$ is *symmetric*. In fact, $\widetilde{\mathrm{A}}^{(k)}$ is a linear operator, and consider the second equation in (3.2.8), we have:

$$h_{ik} = \widetilde{\mathrm{A}}^{(k)}\xi_i = \widetilde{\mathrm{A}}^{(k)}\left(\widetilde{\mathrm{A}}^{(i)}(x,\ x)\right) = \widetilde{\mathrm{A}}^{(i)}\left(\widetilde{\mathrm{A}}^{(k)}(x,\ x)\right) = \widetilde{\mathrm{A}}^{(i)}\xi_k = h_{ki} \tag{3.2.11}$$

Therefore, (3.2.9) becomes:

$$\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} = \begin{pmatrix} h_{10} \\ \vdots \\ h_{n0} \end{pmatrix} + \left\{ \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nn} \end{pmatrix} + \begin{pmatrix} \mu_{11} & \cdots & \mu_{n1} \\ \vdots & \ddots & \vdots \\ \mu_{1n} & \cdots & \mu_{nn} \end{pmatrix}^{-1} \right\} \bullet \begin{pmatrix} \chi_1 \\ \vdots \\ \chi_n \end{pmatrix} \tag{3.2.12}$$

Thus, by solving this system of linear equations, we can find the parameters $\{\chi_k\}$. These parameters, together with those operators in the Fredholm equations (3.2.8), determine the linearised operator (3.2.7) as the solution of the W-H equation.

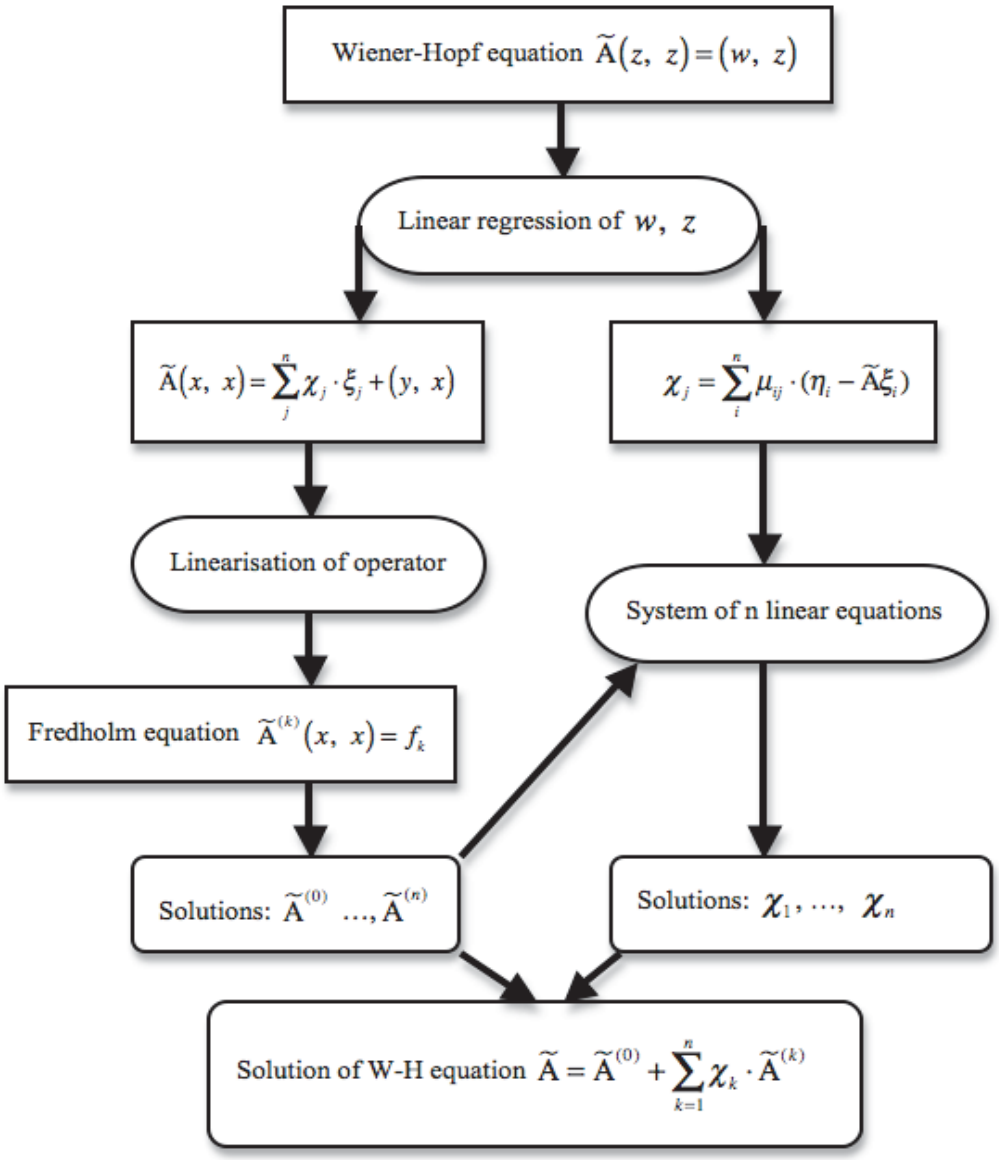

Figure 3-2. Algorithm of the Solution to the Wiener-Hopf Equation by Linearisation

To sum up this algorithm, Figure 3-2 is a flowchart showing the introduced procedure for solving the Wiener-Hopf equation. As we can see, it requires a linearisation of the signal first, the result of which is a linear system of one operator and n unknown parameters. Next, a further linearisation on the operator is needed, whose number should agree with that of linear terms in signals by the previous linearisation, so that they could be

transformed into a number of Fredholm equations of the first kind. The solutions to these equations can be sent back to the original linear system, under which the n unknown parameters can be solved. Finally, the linear combination of these parameters and the solutions of Fredholm equation is the solution to the W-H equation.

3.2.2 Solution of Fredholm Equation Using Galerkin's Method and Canonical Expansion

Now the only task is to acquire the sequence of linear operators $\left\{\widetilde{\mathrm{A}}^{(k)}\right\}$ through the solution of the Fredholm equations (3.2.8). All equations in (3.2.8) can be unified in the following form:

$$\widetilde{\mathrm{A}}^{(k)}\left(x,\ x\right) = f_k \qquad k = 0,1,\ldots,n \tag{3.2.13}$$

To solve this equation, we will use *Galerkin's method* (Kantorovich & Akilov, 1964; Keller, 2004) and borrow some results that we have already obtained in Chapter Two on canonical expansion.

3.2.2.1 Galerkin's method and canonical expansion

If $\left\{\widetilde{\mathrm{A}}^{(k)}\right\}$ are linear integral operators, then each equation of the system (3.2.13) is of the following form:

$$\int_T \rho(s,t)\,K_X(t,\tau)dt = f(s,\tau) \tag{3.2.14}$$

Hence, our sequence $\left\{\widetilde{\mathrm{A}}^{(k)}\right\}$ is completely determined by the function $\rho(s,t)$, which is the solution to be sought of this integral equation with the symmetric kernel $K_X(t,\tau)$.

The idea of *Galerkin's method*, generally speaking, consists in writing the solution as a linear combination of elements from a complete orthogonal system, in other words, we look for a local solution in a linear subspace instead of the entire Hilbert space (Keller, 2004). In our context, we will select the *principal components* $\{a_i(t)\}$ of the noise model x as the basis of Galerkin's solution; thus, the solution of (3.2.14) is written as:

$$\rho(s,t) = \sum_i c_i(s)\, a_i(t) \tag{3.2.15}$$

where $\{a_i(t)\}$ are PCs satisfying the condition of bi-orthogonality (2.1.7a) with the optimal coordinate functions $\{x_j(t)\}$ in canonical expansion of the noise $X(t)$:

$$\int_T a_i(t)\, x_j(t)\, dt = \delta_{ij} \tag{2.1.8a}$$

By introducing (3.2.15), our Fredholm equation (3.2.14) has now become:

$$\begin{aligned} &\int_T \sum_i c_i(s) a_i(t) K_X(t,\tau) dt = f(s,\tau) \\ \Rightarrow &\sum_i c_i(s) \left(\int_T a_i(t) K_X(t,\tau)\, dt \right) = f(s,\tau) \\ \Rightarrow &\sum_i D_i\, c_i(s)\, x_i(\tau) = f(s,\tau) \end{aligned} \tag{3.2.16}$$

In such a way, we have obtained a canonical expansion for the function $f(s,\tau)$ in terms of the coordinate functions.

On multiplying the function $a_j(\tau)$ on both sides of (3.2.16), and taking the integral, we have:

$$\int_T \sum_i D_i\, c_i(s)\, a_j(\tau) x_i(\tau) d\tau = \int_T a_j(\tau) f(s,\tau) d\tau$$

Applying the bi-orthogonality equation (2.1.7a), we then have an expression for $c_j(s)$, which is very similar to the expression for the coordinate function (2.1.6a) :

$$c_j(s) = \frac{1}{D_j}\int_T a_j(\tau) f(s,\tau) d\tau \tag{3.2.17}$$

Finally, if we substitute these $c_i(s)$ in (3.2.15), the solution to the Fredholm equation of the first kind (3.2.14) using Galerkin's method is then of the form:

$$\rho(s,t) = \int_T \left(\sum_i \frac{a_i(t) a_i(\tau)}{D_i} \right) f(s,\tau) d\tau \tag{3.2.18}$$

This is sometimes described as *Green's* or the *inverse kernel* applied to the function $f(s,\tau)$ (Blais, 1988).

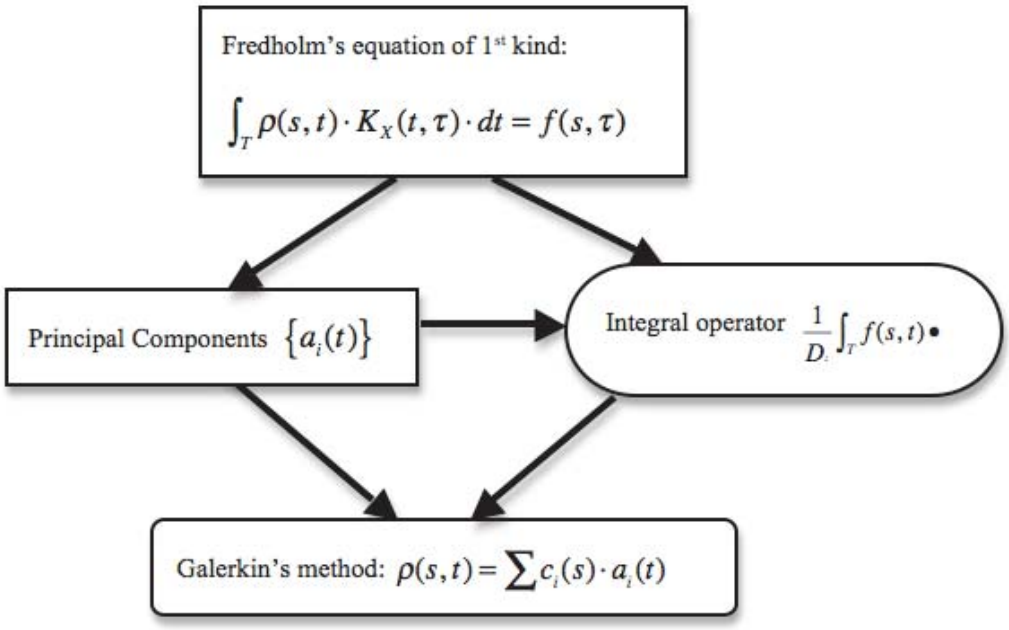

Figure 3-3 Procedure of solving Fredholm's Equation of First Kind by Galerkin's Method using the Principal Components

Figure 3-3 illustrates the algorithm of this method. From the known covariance function of the noise, we can generate its principal components first, and use them as the basis in the linear solution; on the other hand, we will use the known function on the right hand side of the equation as the kernel of an integral operator, and apply them on those PCs to obtain the coefficients for the linear form, which is our desired solution.

3.2.2.2 General solution of Wiener-Hopf equation

Now we come back to the system (3.2.8). For the first equation of $\widetilde{\mathrm{A}}^{(0)}$, it is only necessary to replace the function $f(s,\tau)$ by the cross-covariance function $K_{YX}(s,\tau)$:

$$\int_T g^{(0)}(s,t)K_X(t,\tau)dt = K_{YX}(s,\tau) \tag{3.2.19}$$

and because of (3.2.13), the solution can be written as:

$$g^{(0)}(s,t) = \sum_i y_i(s)a_i(t) \tag{3.2.20}$$

$$\text{where } y_i(s) = \frac{1}{D_i}\int_T a_i(\tau)K_{YX}(s,\tau)d\tau \tag{3.2.21}$$

Analogously, for the rest of equations of $\widetilde{\mathrm{A}}^{(k)}$, we only need to set $f(s,\tau) = \xi_k(\tau)$, and in this case, the functions $g^{(k)}(s,t)$ correspondingly contain only one parameter t, i.e., $g^{(k)}(s,t) = g^{(k)}(t)$. The equation (3.2.14) here takes the form:

$$\int_T g^{(k)}(t)K_X(t,\tau)dt = \xi_k(\tau) \tag{3.2.22}$$

then we get Galerkin's solution:

$$g^{(k)}(t) = \sum_i \alpha_{ik}\, a_i(t) \tag{3.2.23}$$

where $\alpha_{ik} = \frac{1}{D_i}\int_T a_i(\tau)\xi_k(\tau)d\tau$ (3.2.24)

Alternatively, if we write the above result in a matrix form, we have:

$$\begin{pmatrix} g^{(0)} \\ g^{(1)} \\ \vdots \\ g^{(n)} \end{pmatrix} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \\ \alpha_{11} & \alpha_{21} & \cdots & \alpha_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1n} & \alpha_{2n} & \cdots & \alpha_{nn} \end{pmatrix} \bullet \begin{pmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_n(t) \end{pmatrix} \quad (3.2.25)$$

In practice, the observation is given at discrete points, so (3.2.25) will take the matrix form, which we will call the *Galerkin matrix*.

Recall the symmetric matrix composed by $h_{ik} = \widetilde{\mathrm{A}}^{(k)}\xi_i$ in (3.2.10) and (3.2.11), since now $\widetilde{\mathrm{A}}^{(k)}$ takes the form of an integral operator, we have:

$$h_{ik} = \int_T g^{(k)}(t)\xi_i(t)dt \quad (3.2.26)$$

This symbol h_{ik} simply means the linear integral operator with $g^{(k)}(s,t)$ as the kernel applying on each coefficient of linear regression form of the signal.

Taking (3.2.23) and (3.2.24) into consideration, we can rearrange (3.2.26) as:

$$h_{ik} = \int_T \sum_j \alpha_{jk} a_j(t)\xi_i(t)dt = \sum_j \alpha_{jk}\left(\int_T a_j(t)\xi_i(t)dt\right) = \sum_j D_j \alpha_{jk}\alpha_{ji}$$

or, as a matrix:

$$\begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nn} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{n1} \\ \vdots & \ddots & \vdots \\ \alpha_{1n} & \cdots & \alpha_{nn} \end{pmatrix} \bullet \begin{pmatrix} D_1 & & 0 \\ & \ddots & \\ 0 & & D_n \end{pmatrix} \bullet \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix} \quad (3.2.27)$$

Similarly, we have the expression for $\{h_{i0}\}$:

$$\begin{pmatrix} h_{10} \\ \vdots \\ h_{n0} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{n1} \\ \vdots & \ddots & \vdots \\ \alpha_{1n} & \cdots & \alpha_{nn} \end{pmatrix} \bullet \begin{pmatrix} D_1 & & 0 \\ & \ddots & \\ 0 & & D_n \end{pmatrix} \bullet \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{3.2.28}$$

and in such a way, we have finally acquired the expression for the system (3.2.12):

$$\begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix} - \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{n1} \\ \vdots & \ddots & \vdots \\ \alpha_{1n} & \cdots & \alpha_{nn} \end{pmatrix} \bullet \begin{pmatrix} D_1 & & 0 \\ & \ddots & \\ 0 & & D_n \end{pmatrix} \bullet \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \left\{ \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{n1} \\ \vdots & \ddots & \vdots \\ \alpha_{1n} & \cdots & \alpha_{nn} \end{pmatrix} \bullet \begin{pmatrix} D_1 & & 0 \\ & \ddots & \\ 0 & & D_n \end{pmatrix} \bullet \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix} + \begin{pmatrix} \mu_{11} & \cdots & \mu_{n1} \\ \vdots & \ddots & \vdots \\ \mu_{1n} & \cdots & \mu_{nn} \end{pmatrix}^{-1} \right\} \bullet \begin{pmatrix} \chi_1 \\ \vdots \\ \chi_n \end{pmatrix}$$

(3.2.29)

The solution of this linear system, together with the matrix form (3.2.10) of the coefficients, fully determines the solution to the Wiener-Hopf equation, given by:

$$g(s,t) = g^{(0)}(s,t) + \sum_{k=1}^{n} \chi_k(s)\, g^{(k)}(t) \tag{3.2.30}$$

3.2.3 Quality of the Solution

To fulfill the study of this Wiener filter problem, it is still necessary to determine the quality of this linear operator, viz. the solution of the Wiener-Hopf equation. To do this, we will measure the mean-square error of the operator, or in the language of Hilbert space, the distance from the true signal to the W-H type estimator in its estimator space, because the mean-square error is simply the square of the norm of this error. Again, using the properties of inner product and linear operator (Lusternik & Sobolev, 1974), we get:

$$\begin{aligned} \|e\|^2 &= \left(\widetilde{\mathrm{A}}z - w,\ \widetilde{\mathrm{A}}z - w\right) \\ &= \widetilde{\mathrm{A}}\left[\widetilde{\mathrm{A}}\left(z,\ z\right) - \left(w,\ z\right)\right] - \widetilde{\mathrm{A}}\left(w,\ z\right) + \left(w,\ w\right) \end{aligned} \tag{3.2.29}$$

Due to the fact that the expression inside the square brackets is nothing but the W-H equation, the first term vanishes identically, accordingly:

$$\|e\|^2 = \left(w,\ w\right) - \widetilde{\mathrm{A}}\left(w,\ z\right) \tag{3.2.30}$$

similar to the calculation (3.2.4), we can obtain a system of the same kind:

$$\begin{cases} \|e\|^2 = \left(y,\ y\right) - \widetilde{\mathrm{A}}\left(y,\ x\right) + \sum_j^n \chi_j \eta_j \\ \chi_j = \sum_i^n \mu_{ij}(\eta_i - \widetilde{\mathrm{A}}\xi_i) \end{cases} \tag{3.2.31}$$

Since our operator has been decomposed into the linear form (3.2.6), and for the consistency of the symbols, we set $\left(y,\ x\right) = \xi_0$, recalling the notation $h_{ik} = \widetilde{\mathrm{A}}^{(k)}\xi_i$ that was defined previously, as well the fact that the matrix $\left\{h_{ik}\right\}$ is symmetric and $\eta_i - h_{i0} = p_i$, the mean square error has then reduced to:

$$\|e\|^2 = \left(y,\ y\right) - h_{00} + \sum_i^n \chi_i p_i \tag{3.2.32}$$

or, using the matrix relation (3.2.10) in Section 3.2.1 , we can equally write:

$$\|e\|^2 = \left(y,\ y\right) - h_{00} + \begin{pmatrix} \chi_1 \\ \vdots \\ \chi_n \end{pmatrix}^T \bullet \left\{ \begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nn} \end{pmatrix} + \begin{pmatrix} \mu_{11} & \cdots & \mu_{n1} \\ \vdots & \ddots & \vdots \\ \mu_{1n} & \cdots & \mu_{nn} \end{pmatrix}^{-1} \right\} \bullet \begin{pmatrix} \chi_1 \\ \vdots \\ \chi_n \end{pmatrix} \tag{3.2.33}$$

Chapter Four: DETERMINATION OF THE OPTIMAL NON-LINEAR ESTIMATOR BY THE CANONICAL EXPANSION METHOD

In the last chapter, we demonstrated how to find the optimal linear operator subject to the principle of MMSE. Nevertheless, in some practical problems, it is not the linear operator which gives the best solution to the equation (3.1), but the non-linear one. In terms of the theory of operators, the set of linear operators only constitutes a linear subspace of all permissible operators that can be applied to observation. Hence, the task in this chapter is to extend from this linear subspace to the whole space of operators in which to find the optimal operator, and to provide a systematic procedure of calculating this operator using the canonical expansion and principal components, particularly when the noise obeys a Gaussian distribution. Moreover, the same procedure can be generalised to acquire the optimal operator according to any other criterion of optimality along with MMSE. This will also be discussed in general in the present chapter.

4.1 Optimal Non-linear Operator under the MMSE Principle

In Section 3.1, we have demonstrated that the condition of optimality on the operator can be established from the projection theorem in Hilbert space, namely:

$$\left(w - \widetilde{\mathrm{A}}z,\ \widetilde{\mathrm{B}}z\right) = 0 \qquad (\forall \widetilde{\mathrm{B}}) \tag{4.1.1}$$

In the last chapter, we have shown that if $\widetilde{\mathrm{A}}$ is a linear operator, this condition will produce the Wiener-Hopf equation. Nevertheless, if $\widetilde{\mathrm{A}}$ is a non-linear one, that computation will no longer be valid, since the operator cannot be integrated inside the

inner product. We therefore require a different approach to arrive at a general solution to this equation in the non-linear case.

Remembering the inner product is defined as the second moment, the condition (4.1.1) is actually equivalent to:

$$E\left[\left\{\tilde{\mathrm{A}}Z(t)-W(s)\right\}\tilde{\mathrm{B}}Z(\tau)\right]=0 \quad (\forall\tilde{\mathrm{B}}) \tag{4.1.2}$$

Using the identity that the mathematical expectation is equal to the expectation of the *conditional expectation* given a realisation, we get:

$$E\left[\left\{\tilde{\mathrm{A}}Z(t)-W(s)\right\}\tilde{\mathrm{B}}Z(\tau)\right]=E\left[E\left[\left\{\tilde{\mathrm{A}}Z(t)-W(s)\right\}\tilde{\mathrm{B}}Z(\tau)\middle|\ Z(t)\right]\right]=0 \tag{4.1.3}$$

Using the basic properties of mathematical expectation, the inner conditional expectation on the right hand side can be rearranged in this way:

$$\begin{aligned}&E\left[\left\{\tilde{\mathrm{A}}Z(t)-W(s)\right\}\tilde{\mathrm{B}}Z(\tau)\middle|\ Z(t)\right]=\tilde{\mathrm{B}}Z(\tau)E\left[\left\{\tilde{\mathrm{A}}Z(t)-W(s)\right\}\middle|\ Z(t)\right]\\&=\tilde{\mathrm{B}}Z(\tau)\left\{\tilde{\mathrm{A}}Z(t)-E\left[W(s)\middle|\ Z(t)\right]\right\}\end{aligned} \tag{4.1.4}$$

The optimality condition for the non-linear operator (4.1.2) has now become:

$$E\left[\tilde{\mathrm{B}}Z(\tau)\left\{\tilde{\mathrm{A}}Z(t)-E\left[W(s)\middle|\ Z(t)\right]\right\}\right]=0 \tag{4.1.5}$$

Since this condition should be satisfied for any operator $\tilde{\mathrm{B}}$, this suggests the following relation must hold true as the necessary condition:

$$\tilde{\mathrm{A}}Z(t)=E\left[W(s)\middle|\ Z(t)\right] \tag{4.1.6}$$

It is worth pointing out that this condition is also a sufficient condition for the optimal operator (Pugachev, 1965). Thus, from the projection theorem of Hilbert space, we have deduced this well-known fact: the non-linear estimator that gives the best estimation of

the signal, among all the possible estimators and according to the MMSE, is the conditional mathematical expectation given an observation, which is normally called the *regression* of the signal on the observation (or *a-posteriori mathematical expectation*; see Andreyev, 1969). In the next section, we will use the method of canonical expansion to obtain an analytical expression for the regression (4.1.6), when the noise is Gaussian.

4.2 Determination of the Non-linear Operator by the Canonical Expansion

Roughly speaking, there are two possibilities of finding an analytical form for the regression operator (4.1.6). The first way is to reduce this non-linear operator to a combination of linear ones (if it is reducible), and for each linear sub-operator, we can apply the foregoing theory of construction of W-H type linear estimators using PCs. For a more general scheme, we can conceive an algorithm of direct computation from the definition of regression, which is the goal of this section. The procedure described in this section is due to Pugachev (1965) and Andreyev (1969).

4.2.1 Determination of Regression when the Noise is Gaussian

Assuming again that we write the observation as a linear regression model:

$$Z(t) = \sum_{i}^{n} \xi_i(t)\, U_i + X(t) \tag{4.2.1}$$

and the Gaussian noise has the form of canonical expansion:

$$X(t) = \sum_{j} x_j(t)\, V_j \tag{3.2.13}$$

We introduce a new variable as the inner product of PC and observation:

$$Z_v = \int_T a_v(t) Z(t) dt \tag{4.2.2}$$

As before, the $a_v(t)$ and $x_i(t)$ are subject to the bi-orthogonality condition (2.1.7a). Recalling expression (3.2.23), we calculate:

$$Z_v = \int_T a_v(t) \left(\sum_i^n \xi_i(t)\, U_i + \sum_j x_j(t)\, V_j \right) dt = D_v \sum_i^n \alpha_{vi}\, U_i + V_v \tag{4.2.3}$$

The reason why we have introduced this auxiliary variable is that in this way the observation $Z(t)$ can be decomposed in a linear form which is exhaustively determined by the new variables, with its coordinate functions being those $x_v(t)$:

$$Z(t) = \sum_v x_v(t)\, Z_v \tag{4.2.4}$$

Therefore, all the probabilistic properties of the observation can be measured by fully characterising the properties of these newly generated variables.

Recall that the regression is formally equal to:

$$E\left[W(s) \middle| Z(t) \right] = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} W(s) f\left(U_1,\, U_2,\, \dots \middle| z_1,\, z_2,\, \dots \right) dU_1 dU_2 \dots \tag{4.2.5}$$

where the signal $W(s)$ takes the linear form $W(s) = \sum_i^n \xi_i(t)\, U_i$, or more generally speaking, the signal can take any form depending on these variables $\{U_i\}$. To determine the conditional PDF $f\left(U_1,\, U_2,\, \dots \middle| z_1,\, z_2,\, \dots \right)$ in (4.2.5), i.e., the *a-posteriori density* of $\{U_i\}$, we use *Bayes' formula*:

$$f\left(U_1,\, U_2,\, \ldots \middle| z_1,\, z_2,\, \ldots\right) = \frac{f\left(Z_1,\, Z_2,\, \ldots \middle| u_1,\, u_2,\, \ldots\right) f\left(u_1,\, u_2,\, \ldots\right)}{\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f\left(Z_1,\, Z_2,\, \ldots \middle| u_1,\, u_2,\, \ldots\right) f\left(u_1,\, u_2,\, \ldots\right) du_1 du_2 \ldots} \tag{4.2.6}$$

If the noise is Gaussian, it means the conditional PDF $f\left(Z_v \middle| u_1,\, u_2,\, \ldots\right)$ is Gaussian too, because each variable $\left\{Z_v \middle| u\right\}$ should be deemed as the result of the linear transformation of Gaussian variables. Also, from (4.2.3), we can easily obtain the mean value and variance of each variable $\left\{Z_v \middle| u\right\}$:

$$E\left[Z_v \middle| u\right] = D_v \sum_i^n \alpha_{vi}\, u_i \;, \quad \text{and} \quad Var\left[Z_v \middle| u\right] = D_v \tag{4.2.7}$$

Thus, for each single variable $\left\{Z_v \middle| u\right\}$, its PDF is given by:

$$f\left(Z_v \middle| u\right) = \frac{1}{\sqrt{2\pi D_v}} \exp\left\{ - \frac{\left(Z_v - D_v \sum_i^n \alpha_{vi}\, u_i \right)^2}{2 D_v} \right\} \tag{4.2.8}$$

Furthermore, these variables are uncorrelated, and in the context of Gaussian variables, it is equivalent to saying that they are independent. Correspondingly, the joint distribution of the conditional variables can be decomposed into the product of each distribution, and taking (4.2.3) and (4.2.7) into consideration, the Bayes formula reads

$$f\left(U \middle| z_1,\, z_2,\, \ldots\right) = \frac{\prod_v f\left(Z_v \middle| u\right) f\left(u\right)}{\int_{-\infty}^{+\infty} \prod_v f\left(Z_v \middle| u\right) f\left(u\right) du} = \frac{\exp\left\{ \sum_i \mathrm{H}_i\, u_i - \frac{1}{2} \sum_{i,j} h_{ij}\, u_i\, u_j \right\} f\left(u\right)}{\int_{-\infty}^{+\infty} \exp\left\{ \sum_i \mathrm{H}_i\, u_i - \frac{1}{2} \sum_{i,j} h_{ij}\, u_i\, u_j \right\} f\left(u\right) du} \tag{4.2.9}$$

where for simplicity, h_{ij} and H_i are defined as:

$$h_{ij} = \sum_v D_v \alpha_{vi}\alpha_{vj} \qquad \text{and} \quad \mathrm{H}_i = \sum_v Z_v \alpha_{vi} \tag{4.2.10}$$

Thus, the regression in general is determined by

$$E\left[W(s)\middle| Z(t)\right] = \frac{\int_{-\infty}^{+\infty} W(s,u)\exp\left\{\sum_i \mathrm{H}_i u_i - \frac{1}{2}\sum_{i,j} h_{ij} u_i u_j\right\} f(u)du}{\int_{-\infty}^{+\infty} \exp\left\{\sum_i \mathrm{H}_i u_i - \frac{1}{2}\sum_{i,j} h_{ij} u_i u_j\right\} f(u)du} \tag{4.2.11}$$

Now it is expedient to summarise the above results as the following procedure of calculating the regression using the approach of canonical expansion:

- **Step I**: Use the PCs to transform the observation into a sequence of random variables $\{Z_v\}$ defined by (4.2.2), to set up a one-to-one correspondence between these variables and the observation.
- **Step II**: Find the conditional PDF of the variables $\{U_v\}$ (signal-dependent) given $\{Z_v\}$ using Bayes' formula.
- **Step III**: Find the conditional expectation.

4.2.2 Quality of the Regression Estimator

As in the case of linear operator, we need to determine the quality of this regression estimator. Again, the mean-square error is measured by the norm:

$$\|\varepsilon\|^2 = \left(w - \widetilde{\mathrm{A}}z,\ w - \widetilde{\mathrm{A}}z\right) = \left(w - \widetilde{\mathrm{A}}z,\ w\right) - \left(w - \widetilde{\mathrm{A}}z,\ \widetilde{\mathrm{A}}z\right) \tag{4.2.12}$$

The second term on the right hand side of this equation vanishes identically, for this is the condition of optimality (3.1.3). Thus, using the iterated expectation:

$$\begin{aligned} \|\varepsilon\|^2 &= \left(w - \tilde{\mathrm{A}}z,\ w\right) = E\left[\left\{W(s) - E\left[W(s)\middle| Z(t)\right]\right\}W(s)\right] \\ &= E\left[E\left[\left\{W(s)^2 - W(s)\cdot E\left[W(s)\middle| Z(t)\right]\right\}\middle|\ Z(t)\right]\right] \\ &= E\left[E\left[W(s)^2\middle|\ Z(t)\right] - E\left[W(s)\middle| Z(t)\right]^2\right] \\ &= E\left[Var\left[W(s)\middle| Z(t)\right]\right] \end{aligned} \tag{4.2.13}$$

This essential condition implies that under the principle of MMSE, unlike the optimal linear operator (Wiener type), our optimal non-linear operator is a stable one in the sense that for any given realisation of the random process, the minimum mean-square error remains an invariant.

4.3 Further Remarks on the Determination of the Optimal Operator under Various Principles

So far, we have confined ourselves only to the MMSE estimation, and have formulated both the linear and non-linear estimators under this criterion. Yet, MMSE is not a unique principle of optimality in practice: for different purposes, we may equally establish other criteria for determining the optimal operator (or system). For instance, in a signal detection problem, we normally use Neyman-Pearson's lemma as the optimal decision of a signal being present or absent. Then comes the question of how to adjust the foregoing theory to formulating the optimal operator according to various criteria. Pugachev (1965) and Andreyev (1969) both propose a thorough discussion in this direction. Although it is beyond the scope of this thesis, we will sketch out a rough outline in this section.

Recall that the principle of MMSE is:

$$\| e \|^2 = E\left[\left(W(s) - \widetilde{\mathrm{A}}Z(t)\right)^2\right] = \min \tag{4.3.1}$$

Here, if we regard the mean-square error as a special form of a more general *loss function (or cost function)*, the principle of optimality can thus be generalised to:

$$\| e \|^2 = E\left[l\left(W(s), \widetilde{\mathrm{A}}Z(t)\right)\right] = \min \tag{4.3.2}$$

where $l\left(W(s), \widetilde{\mathrm{A}}Z(t)\right)$ is the loss function.

Similar to our treatment of regression operator, this condition can be rearranged as:

$$\| e \|^2 = E\left[E\left[l\left(W(s), \widetilde{\mathrm{A}}Z(t)\right)\middle| Z(t)\right]\right] = \min \tag{4.3.3}$$

since this condition should be satisfied for all the realisations, we finally have:

$$E\left[l\left(W(s), \widetilde{\mathrm{A}}Z(t)\right)\middle| Z(t)\right] = \min \tag{4.3.4}$$

This is the condition of determining the optimal operator according to any criterion. This is called the minimum *conditional risk*. As long as we are able to construct the loss function, the optimal operator (either linear or non-linear) $\widetilde{\mathrm{A}}$ can be found from this condition.

Therefore, we can carry out an analogous procedure to that of calculation of the non-linear regression estimator. The only difference is that the regression is the conditional expectation of the signal given any realisation, while in the present case of conditional risk, the signal is substituted by a loss function inside the operation of conditional expectation, i.e., we have to minimise the integral:

$$E\left[l\left(W(s), \widetilde{\mathrm{A}}Z(t)\right) \middle| Z(t) \right] =$$
$$\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} l\left(W(s, U_1, U_2, \ldots), \widetilde{\mathrm{A}}\left(s, Z_1, Z_2, \ldots\right)\right) f\left(U_1, U_2, \ldots \middle| Z_1, Z_2, \ldots\right) dU_1 dU_2 \ldots$$

(4.3.5)

$\{U_v\}$ are the variables that the signal depends on, and $\{Z_v\}$ are those variables generated from the principal components, which uniquely determine the observation. In such a way, defining an optimal operator according to any criterion can be generally solved by the method of canonical expansion.

We can conclude that, for each particular estimation problem, the primary task is to set up the most appropriate principle of optimality, so that the corresponding loss function can be found. Once this is done, the optimal operator that is being sought can be acquired by minimising the conditional risk, with the aid of the principal components and the Bayes formula. In a manner of speaking, this proposed procedure should have a more fundamental meaning than the linear approach provided in Chapter Three, because it is directly derived from the loss function according to the selected criterion of optimality, and in this sense, the W-H type linear estimator is only a special case of such procedure. For an extensive discussion on different principles of optimality, one should consult Pugachev (1965) and Andreyev (1969) for more details.

Chapter Five: APPLICATIONS

The foregoing ideas of adopting the canonical expansion for the MMSE estimation will in this chapter be applied in solving three particular problems. It is an attempt to show that this expansion approach could be utilised in investigating a collection of estimation problems in geophysics or geomatics engineering, as well as to provide numerical tests in order to examine the performance or computability of the linear and non-linear operators that we have established in the previous two chapters based on the principal components. The selected examples encompass filtering, prediction and extrapolation, all of which are commonly dealt with and well-studied in the existing literature of past decades (although they are the same problem in essence). All of the noise models are of simple type, either a white noise or a first-order Gauss-Markov process, merely for illustration purposes; and without loss of generality, all the processes considered are of zero mean.

5.1 Random Constant Reconstruction

The first example is the reproduction of a random signal which is a constant in time. This random constant is a uniformly distributed variable over some interval. The model of this problem then takes the form:

$$\begin{aligned} Z(t) &= U + X(t) \\ W(s) &= U \end{aligned} \tag{5.1.1}$$

where $Z(t)$ denotes the observation and $W(s)$ denotes the signal. $X(t)$ is the noise and U a uniformly distributed variable over $[-c, +c]$.

In electrical engineering, this problem is widely known as *DC level of a signal* (Kay, 1999). It can also be understood as a first moment estimation problem. The reason why we adopt this simple example is to show, using the previously constructed estimators, the fact that the linear operator sometimes gives the suboptimal estimation result which necessitates to formulate a non-linear operator to produce the optimal outcome. When it comes to the situation that the noise is a more complicated process than WGN, one can readily follow the exact procedure described here but with a substitution of PCs in the noise model. One of the applications of this model can also be found in the analysis of inertial sensor random errors. In Strapdown Inertial Navigation System (SINS) errors, this constant value is called *random bias* (Nassar, 2003). Thus, the proposed method can well be applied in this problem of random bias estimation (where the associated Gaussian noise model is usually a 1st order Gauss-Markov process; see Nassar 2003).

5.1.1 Methodology

The most typical and simplest estimator for this model (5.1.1) under a WGN is the *sample mean*, given by:

$$\hat{A} = \frac{1}{N}\sum_{n=1}^{N} Z(n) \tag{5.1.2}$$

The quality of this estimator is measured by its variance:

$$Var\left[\hat{A}\right] = \frac{\sigma^2}{N} \tag{5.1.3}$$

We will then compare the performances of both the optimal linear operator (the solution of Wiener-Hopf equation) and the optimal non-linear operator (the regression estimator),

with this unbiased sample mean estimator, to see how accurate the estimated values are in each case are compared to the true value.

5.1.1.1 *Optimal linear operator*

In Chapter Three, we have proved that the best linear operator must satisfy the Wiener-Hopf equation (3.2.2). The solution of the W-H equation in the present case reduces to:

$$g(s,t) = \chi_1(s)\, g^{(1)}(t) \tag{5.1.4}$$

where $g^{(1)}(s,t)$ is the solution of the Fredholm equation of first kind:

$$\int_T g^{(1)}(t)\, K_X(t,\tau)dt = 1 \tag{5.1.5}$$

By Galerkin's method, the solution has the form:

$$g^{(1)}(t) = \sum_\nu \alpha_{\nu 1}\, a_\nu(t) \tag{5.1.6}$$

$$\text{with} \quad \alpha_{\nu 1} = \frac{1}{D_\nu}\int_T a_\nu(t)dt \tag{5.1.7}$$

To determine the parameter χ_1 in (5.1.4), we need to solve the system (3.2.10), which in this case consists of only one variable:

$$\chi_1 = \frac{c^2}{c^2 h_{11} + 3}\ , \qquad \text{where} \quad h_{11} = \sum_\nu D_\nu\, \alpha_{\nu 1}^2 \tag{5.1.8}$$

Thus, the weighting function (5.1.4) of the optimal linear estimator has the form:

$$g(s,t) = \frac{c^2}{c^2 h_{11} + 3}\left(\sum_\nu \alpha_{\nu 1}\, a_\nu(t)\right) \tag{5.1.9}$$

and the estimator in this case is simply a biased mean, given by:

$$\widehat{W}(s) = \widetilde{\mathrm{A}}Z(t) = \int_T g(s,t)Z(t)\,dt \tag{5.1.10}$$

The mean-square error is determined by (3.2.33):

$$\| e \|^2 = \chi_1^2 (h_{11} + \frac{3}{c^2}) = \frac{c^2}{c^2 h_{11} + 3} = \chi_1 \tag{5.1.11}$$

5.1.1.2 Optimal non-linear operator

As is stated in Section 4.2, the optimal non-linear operator under the MMSE criterion is the regression of the signal on the observation. Since the signal is nothing but a random variable U_1, our estimator (4.2.5) becomes:

$$E\left[U_1 \middle| Z(t)\right] = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} U_1 f\left(U_1 \middle| z_1,\ z_2,\ \ldots\right) dU_1 \tag{5.1.12}$$

and the variables $\{Z_\nu\}$ are given by (4.2.3):

$$Z_\nu = D_\nu \alpha_{\nu 1} U_1 + V_\nu \tag{5.1.13}$$

Hence, the estimator (5.1.12), after the transformation by (5.1.13), takes the form which corresponds to (4.2.10):

$$E\left[U_1 \middle| Z(t)\right] = \frac{\int_{-c}^{+c} u_1 \exp\left\{\mathrm{H}_1 u_1 - \frac{1}{2} h_{11} u_1^2\right\} du_1}{\int_{-c}^{+c} \exp\left\{\mathrm{H}_1 u_1 - \frac{1}{2} h_{11} u_1^2\right\} du_1} \tag{5.1.14}$$

where $\mathrm{H}_1 = \sum_\nu Z_\nu \alpha_{\nu 1}$ and $h_{11} = \sum_\nu D_\nu \alpha_{\nu 1}^2$.

The calculation of (5.1.14) involves the computations of the two integrals of transcendental functions on both the numerator and the denominator. The result will give the following form of the regression estimator (5.1.14):

$$E\left[U_1 \middle| Z(t)\right] = \frac{\mathrm{H}_1}{h_{11}} - \frac{1}{\sqrt{2\pi h_{11}}} \frac{\exp\left\{-\frac{1}{2}\left(c\sqrt{h_{11}} - \frac{\mathrm{H}_1}{\sqrt{h_{11}}}\right)^2\right\} - \exp\left\{-\frac{1}{2}\left(c\sqrt{h_{11}} + \frac{\mathrm{H}_1}{\sqrt{h_{11}}}\right)^2\right\}}{\Phi\left(c\sqrt{h_{11}} - \frac{\mathrm{H}_1}{\sqrt{h_{11}}}\right) + \Phi\left(c\sqrt{h_{11}} + \frac{\mathrm{H}_1}{\sqrt{h_{11}}}\right)}$$

where $\Phi(x)$ is a Gaussian CDF: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left\{-\frac{u^2}{2}\right\} du$. (5.1.15)

Finally, we will determine the error of this estimator. As is shown in the last chapter, the mean square error in this case is given by the expectation of the conditional variance:

$$\|\varepsilon\|^2 = E\left[Var\left[W(s)\middle| Z(t)\right]\right] \tag{4.2.11}$$

which is currently equal to the following expression:

$$\|\varepsilon\|^2 = E\left[E\left[U_1^2\middle| Z(t)\right] - E\left[U_1\middle| Z(t)\right]^2\right] \tag{5.1.16}$$

As we can observe, the second term inside the outer square brackets on the right-hand side of (5.1.16) is simply the square value of non-linear estimator (5.1.15), the first term can also be calculated in a similar way as we did when calculating (5.1.14). Thus, this proves our conclusion of (4.2.12), that the mean-square error of the non-linear estimator is a constant for any given observation.

5.1.2 Noise Model

The fore-mentioned results in Section 5.1.1 apply to every noise model, as long as the PCs of noise can be found. In this example, the noise model is a WGN; therefore, for the convenience of computation, we need to further obtain the practical forms for both

operators in the case of WGN. As is shown in Section 2.4, the principal components of WGN coincide with the sine and cosine functions, which are apparently the Fourier basis:

$$a_v(t) = e^{i\omega_v t} \tag{5.1.17}$$

To calculate the linear operator, the sequence $\{\alpha_{v1}\}$ is determined by (5.1.7):

$$\alpha_{v1} = \frac{1}{\sigma^2 T}\int_T e^{i\omega_v t}\,dt = \begin{cases} 1/\sigma^2 & v = 0 \\ 0 & v = \pm 1,\ \pm 2, \ldots \end{cases} \tag{5.1.18}$$

Consequently, for the optimal linear operator, the weighting function becomes:

$$g(s,t) = \frac{1}{\sigma^2}\frac{c^2}{c^2 h_{11} + 3}\ , \qquad \text{where } h_{11} = D_0\,\alpha_{01}^2 = \frac{T}{\sigma^2} \tag{5.1.19}$$

The linear estimator (5.1.10) in the WGN case is equal to:

$$\tilde{\mathrm{A}}Z(t) = \frac{1}{\sigma^2}\frac{c^2}{c^2 h_{11} + 3}\sum_{n=1}^{N} Z[n]\Delta t \tag{5.1.20}$$

Thus, the weighting function (5.1.20) in this case actually gives a *biased mean*, and it becomes smaller when the noise variance increases.

To calculate the non-linear operator in the case of WGN, firstly we need to determine the auxiliary variable $\{Z_v\}$, which according to (5.1.18) consists of only one variable Z_0:

$$Z_0 = \int_T a_0(t)Z(t)dt = \int_T Z(t)\,dt = \sum_{n=1}^{N} Z[n]\Delta t \tag{5.1.21}$$

Thus, H_1 and h_{11} in this case are consequently given by:

$$\mathrm{H}_1 = Z_0\ \alpha_{01} = \frac{Z_0}{\sigma^2} \quad \text{and} \quad h_{11} = \frac{T}{\sigma^2} \tag{5.1.22}$$

and consequently the non-linear optimal estimator (5.1.15) in the case of WGN reads:

$$E\left[U_1 \middle| Z(t)\right] = \frac{Z_0}{T} - \frac{\sigma}{\sqrt{2\pi T}} \frac{\exp\left\{-\frac{1}{2}\left(c\sqrt{\frac{T}{\sigma^2}} - \frac{Z_0}{\sigma\sqrt{T}}\right)^2\right\} - \exp\left\{-\frac{1}{2}\left(c\sqrt{\frac{T}{\sigma^2}} + \frac{Z_0}{\sigma\sqrt{T}}\right)^2\right\}}{\Phi\left(c\sqrt{\frac{T}{\sigma^2}} - \frac{Z_0}{\sigma\sqrt{T}}\right) + \Phi\left(c\sqrt{\frac{T}{\sigma^2}} + \frac{Z_0}{\sigma\sqrt{T}}\right)} \tag{5.1.23}$$

This formula implies that the non-linear estimator, in the case of a WGN, is a non-linear function of the sample mean.

5.1.3 Simulation Results and Performance Comparison

5.1.3.1 Simulation

The first step is to generate the random constant signal. This can be done by generating a uniformly distributed number R over the region $[0,1]$, and transforming this number to the interval $[-c,+c]$:

$$U = c\,(2R-1) \tag{5.1.24}$$

Then, we generate the independent normally distributed variables with zero mean and variance σ^2, and add it to the just generated uniform variable U. The observation points are chosen at the equidistant in the interval $[0,T]$, with $T = 10s$ and $\Delta t = 0.01s$. Figure 5-1 shows one realisation of the true signal (the random variable U) and the noisy signal, for $c = 5$ and $\sigma^2 = 1.5$.

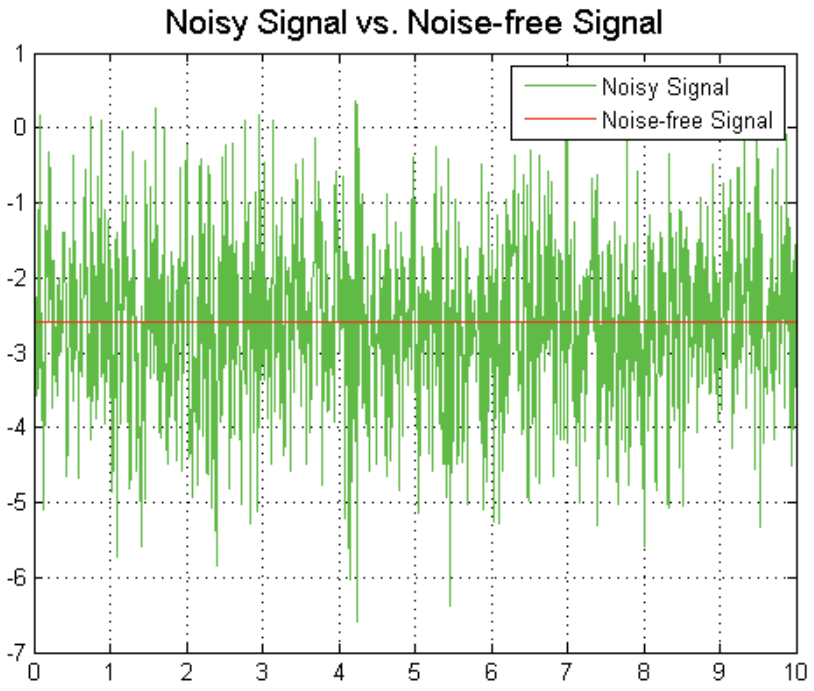

Figure 5-1. One Realisation of the Random Constant Signal and the Noisy Signal as a result of a superimposed White Gaussian Noise N(0, 1.5)

5.1.3.2 Analysis of result

Now, we will examine the performances of the three above formulated estimators in this random constant reconstruction case, to see how accurately each one behaves:

- Sample Mean Estimator (5.1.2)
- W-H type Linear Estimator (5.1.20)
- Regression (non-linear) Estimator (5.1.23)

Table 5-1 shows the result of estimation by the three operators, respectively, for 10 realisations of reconstructing the uniform variable (the 'True Value' in Table 5-1) over $[-5, +5]$, and Table 5-2 provides the corresponding errors of these realisations.

No.	TRUE VALUE	SAMPLE MEAN	LINEAR ESTIMATOR	NON-LINEAR ESTIMATOR
1	3.002805	3.017968	3.634516	3.020986
2	-3.581137	-3.573213	-4.303193	-3.576606
3	-0.782387	-0.776547	-0.935190	-0.777324
4	4.157355	4.131457	4.975482	4.122621
5	2.922073	2.886663	3.476386	2.889549
6	4.592924	4.587459	5.524641	4.488145
7	1.557407	1.554408	1.871962	1.555963
8	-4.642883	-4.589386	-5.526962	-4.489389
9	3.491293	3.537060	4.259653	3.540469
10	1.787352	1.814817	2.185570	1.816632

Table 5-1. Ten Realisations of Estimations using Sample Mean, W-H Type Linear Operator & Regression Estimator, when $\sigma^2 = 1.5$

From these two tables, we can draw the conclusion that, the linear estimator is apparently not optimal in this case, leading to rather high errors when used for estimation. The optimal estimators in this case are the sample mean and the non-linear a-posteriori estimator. They share the same level of optimality by both giving a very close approximation to the true value, especially for the noise with a small variance (in our case, $\sigma^2 = 1.5$), although it seems that there is a relative instability in the errors of the non-linear estimator. Hence, we can assert that the overall optimal estimator in this case should be a non-linear one. Nevertheless, when the variance of the noise grows larger, both the performances of the sample mean estimator and the regression estimator should weaken, where the performance of the regression one is more affected by this increase

than that of the sample mean. In general, we can claim that for a relatively small variance of the noise, the non-linear estimation achieves a very accurate estimation.

No.	SAMPLE MEAN	LINEAR ESTIMATOR	NON-LINEAR ESTIMATOR
1	2.2993e-4	0.3991	3.3055e-4
2	6.2779e-5	0.5214	2.0528e-5
3	3.4107e-5	0.0233	2.5640e-5
4	6.7074e-4	0.6693	12.0650e-4
5	1.2539e-3	0.3073	1.0578e-3
6	5.5732e-5	0.8644	0.0114017
7	8.9924e-6	0.0989	2.0861e-6
8	2.8620e-3	0.7816	23.5605e-3
9	2.0945e-3	0.5904	2.4182e-3
10	7.5434e-4	0.1586	8.5732e-4

Table 5-2. Experimental Square Errors of the Three Estimators, $\sigma^2 = 1.5$.

In this simple case of illustration, we have observed a very high performance of the non-linear estimation using the a-posteriori expectation given the observation, in spite of the high complexity in the calculation of the estimator and its errors, involving the computation of some improper integrals of transcendental functions. Moreover, if the signal consists of more linear terms which obey other distribution laws, and the noise is Gaussian, we can carry out the same procedure to calculate the optimal regression estimator, as long as those improper integrals can be estimated. Yet, when the dimension of signal becomes too large (i.e., a large number of random variables appear in the

signal), this computational load grows up significantly, which makes the calculation of this non-linear operator fairly unfeasible. We will see this case in next section.

5.2 Prediction of Stationary Signal with Non-stationary Noise

As is mentioned at the beginning of Chapter Three, Keller (2000) has proposed a method of solving the Wiener-Hopf equation in the case of a non-stationary noise, using the approach of Haar wavelets. In his paper, he chooses White Gaussian Noise (WGN) with piecewise constant variance added on a stationary signal as the numerical illustration. This problem can be equally solved using the method of canonical expansion, and we will use his example for the demonstration in this section. Firstly, we will simulate an observation by combining the constructed signal and the piecewise noise; and then, we will provide the general form for the optimal linear operator, and investigate the computability of the non-linear operator in this particular case; finally, we will utilise the linear operator to perform the prediction, and see the effect.

5.2.1 Signal and Noise Construction

The stationary signal under study has a covariance function of the form (Keller, 2000):

$$K_W(t,t') = \frac{1}{1+\left[\alpha(t-t')\right]^2} \tag{5.2.1}$$

The validity of the selection of a covariance function of this form is simply that its Fourier Transform, which is the PSD, is always positive. Thus, the stationary process with the covariance (5.2.1) has the real physical meaning in practice. Also, the noise is

selected as a piecewise constant non-stationary WGN with different variances $\sigma_1^2 = 4.0$ and $\sigma_2^2 = 1.0$ in two consecutive equal periods of observation.

Next, we need to construct the signal from this covariance function. This is done by the following steps (Keller, 2000):

- Computation of the covariance matrix from the covariance function (5.2.1)
- *Cholesky decomposition* of the covariance matrix:

$$K = L \cdot L^T \tag{5.2.2}$$

- Signal construction as a linear form:

$$\overrightarrow{W(t)} = L \cdot \vec{u} \tag{5.2.3}$$

where $\vec{u}$ is a vector of uniformly distributed independent variables over the region $\left[-\frac{1}{2},\ \frac{1}{2}\right]$ with unit variance.

It can be proved that the signal defined in (5.2.3) has the covariance (5.2.1):

$$E\left[W(t_i)\,W(t_j)\right] = E\left[\sum_p l_{ip}\,u_p \sum_q l_{jq}\,u_q\right] = \sum_p l_{ip}\,l_{jp}$$

or $K_W = L \cdot L^T = K$ (5.2.4)

Ultimately, we only need to add the piecewise WGN on the constructed signal (5.2.3) to produce the noisy signal. Figure 5-2 depicts such realisation, with 512 samples. The first 256 points in signal are mingled with a WGN with $\sigma_1^2 = 4.0$, while a relatively weaker noise with $\sigma_2^2 = 1.0$ is superimposed on the rest of samples. The value of α is chosen as 0.08 (the smaller this value, the smoother the signal is).

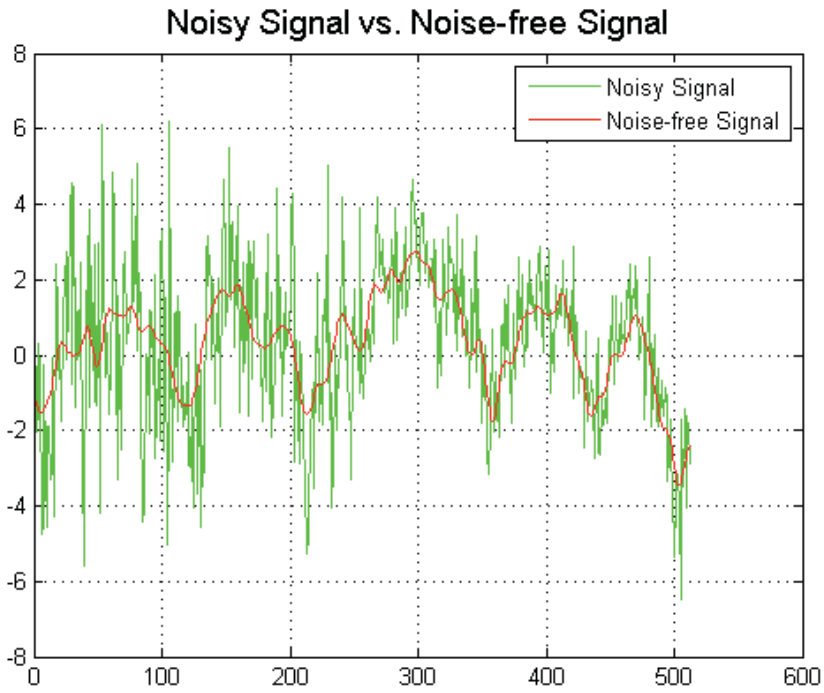

Figure 5-2. Noisy Signal vs. Noise-free Signal, with piecewise Noise $\sigma_1^2 = 4.0$ and $\sigma_2^2 = 1.0$, 512 Samples

5.2.2 Methodology

In this numerical example, we will use the optimal linear operator (predictor) to reconstruct the true signal due to its simplicity, though it is suboptimal out of all possible operators. It will be shown later that the optimal operator (non-linear) entails a relatively high complexity in calculation, which makes it not quite feasible for practical use.

5.2.2.1 Optimal linear operator

The signal is given by (5.2.3) as a linear form, so it is possible to construct a Wiener filter from the W-H equation, with the weighting function:

$$g(s,t)=\sum_{k=1}^{N}\chi_k(s)\,g^{(k)}(t) \tag{5.2.5}$$

N is the total number of sample points. Since the observation is discrete at equidistant points, the component weighting functions $g^{(k)}(s,t)$ in (5.2.5) are determined in the matrix form, and we call this the *Galerkin matrix*:

$$\begin{pmatrix} g^{(1)}(t_1) & g^{(1)}(t_2) & \dots & g^{(1)}(t_N) \\ g^{(2)}(t_1) & g^{(2)}(t_2) & \dots & g^{(2)}(t_N) \\ \vdots & \vdots & \ddots & \vdots \\ g^{(N)}(t_1) & g^{(N)}(t_2) & \dots & g^{(N)}(t_N) \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{21} & \dots & \alpha_{N1} \\ \alpha_{12} & \alpha_{22} & \dots & \alpha_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1N} & \alpha_{2N} & \dots & \alpha_{NN} \end{pmatrix} \bullet \begin{pmatrix} a_1(t_1) & a_1(t_2) & \dots & a_1(t_N) \\ a_2(t_1) & a_2(t_2) & \dots & a_2(t_N) \\ \vdots & \vdots & \ddots & \vdots \\ a_N(t_1) & a_N(t_2) & \dots & a_N(t_N) \end{pmatrix} \tag{5.2.6}$$

As we can see, each entry in the Galerkin matrix is in such way determined by Galerkin's form of the solution of the Fredholm equation. As a result, the weighting function (5.2.5) should be written as a matrix product form of Galerkin matrix and the coefficients $\{\chi_k\}$, which we will call the *weighting matrix* of optimal linear predictor:

$$\begin{pmatrix} g(t_1,t_1) & g(t_1,t_2) & \dots & g(t_1,t_N) \\ g(t_2,t_1) & g(t_2,t_2) & \dots & g(t_2,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ g(t_N,t_1) & g(t_N,t_2) & \dots & g(t_N,t_N) \end{pmatrix} = \begin{pmatrix} \chi_1(t_1) & \chi_2(t_1) & \dots & \chi_N(t_1) \\ \chi_1(t_2) & \chi_2(t_2) & \dots & \chi_N(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(t_N) & \chi_2(t_N) & \dots & \chi_N(t_N) \end{pmatrix} \bullet \begin{pmatrix} g^{(1)}(t_1) & g^{(1)}(t_2) & \dots & g^{(1)}(t_N) \\ g^{(2)}(t_1) & g^{(2)}(t_2) & \dots & g^{(2)}(t_N) \\ \vdots & \vdots & \ddots & \vdots \\ g^{(N)}(t_1) & g^{(N)}(t_2) & \dots & g^{(N)}(t_N) \end{pmatrix} \tag{5.2.7}$$

The linear Wiener filter, as is obtained from W-H equation, is now equal to:

$$\widetilde{\mathrm{A}}Z(t)=\begin{pmatrix} g(t_1,t_1) & g(t_1,t_2) & \dots & g(t_1,t_N) \\ g(t_2,t_1) & g(t_2,t_2) & \dots & g(t_2,t_N) \\ \vdots & \vdots & \ddots & \vdots \\ g(t_N,t_1) & g(t_N,t_2) & \dots & g(t_N,t_N) \end{pmatrix} \bullet \begin{pmatrix} Z(t_1) \\ Z(t_2) \\ \vdots \\ Z(t_N) \end{pmatrix} \tag{5.2.8}$$

where the vector $\overrightarrow{Z(t)}$ are observations.

5.2.2.2 Computability of optimal non-linear operator

Generally, the non-linear operator is not feasible in this case despite its superior performance. We will explain the reason here as we formulate this operator. The construction of non-linear regression operator should follow the 3-step procedure which has been outlined in Section 4.2.1. The first step is to build a one-to-one correspondence between the observation and a new set of random variables $\{Z_v\}$ whose total number is equal to that of observation points, defined by the following relation:

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_N \end{pmatrix} = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & D_N \end{pmatrix} \bullet \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1N} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha_{N1} & \alpha_{N2} & \cdots & \alpha_{NN} \end{pmatrix} \bullet \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix} + \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{pmatrix} \tag{5.2.9}$$

Next, we will need Bayes' formula to obtain the form of conditional density function:

$$f\left(U_1,\, U_2,\, \dots \,|z_1,\, z_2,\, \dots\right) = \frac{\exp\left\{\sum_i \mathrm{H}_i u_i - \frac{1}{2}\sum_{i,j} h_{ij} u_i u_j\right\}}{\int_{-\frac{1}{2}}^{\frac{1}{2}} \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} \exp\left\{\sum_i \mathrm{H}_i u_i - \frac{1}{2}\sum_{i,j} h_{ij} u_i u_j\right\} du_1 \dots du_N} \tag{5.2.10}$$

In order to calculate this conditional PDF, we need to find the value of the multi-dimensional integral in the denominator. However, since the number of terms in the linear form of the signal is equal to the number of the sample points, the integral involves a very high dimension and the integration of an exponential function of a linear form plus a bilinear form. Such complex computations can only be done by complicated substitutions of variables, which require a lengthy and heavy calculation in practice. Therefore, we will not develop the non-linear operator in this case.

5.2.3 Noise Model

Next, we need to find the principal components, and subsequently find those related quantities that are essential to the calculation of Galerkin matrix (5.2.6) and consequently the weighting matrix (5.2.7). Since the noise is a WGN in the observation interval $[0,\ T]$, the canonical expansion has the form:

$$W(t) = \frac{1}{T}\sum_{v} e^{i\omega_v t} V_v \tag{2.2.4}$$

and the PCs are given by $a_v(t) = e^{i\omega_v t}$, yet with different variances of basis variables V_v:

$$D_v = \begin{cases} \sigma_1^2 \dfrac{T}{2} & , \quad \left[0,\ \dfrac{T}{2}\right] \\ \sigma_2^2 \dfrac{T}{2} & , \quad \left[\dfrac{T}{2},\ T\right] \end{cases} \tag{5.2.11}$$

Hence, we have to perform the prediction in $\left[0,\ \dfrac{T}{2}\right]$ and $\left[\dfrac{T}{2},\ T\right]$ respectively, i.e., we have to find the weighting matrices of the predictor on each region of variances.

For such a construction of the predictor, firstly, we need to find the coefficient matrix $\{\alpha_{vr}\}$, which is the result of an integral operator with the r^{th} coefficient in the linearised signal as the kernel, applied to the v^{th} principal component (see (3.2.22)). In this case, to each region there corresponds a particular matrix, whose entry is given by:

$$\alpha_{vr} = \begin{cases} \dfrac{2}{\sigma_1^2 T}\int_0^{\frac{T}{2}} a_v(\tau)\, l_r(\tau)\, d\tau & \left[0,\ \dfrac{T}{2}\right] \\ \dfrac{2}{\sigma_1^2 T}\int_{\frac{T}{2}}^{T} a_v(\tau)\, l_r(\tau)\, d\tau & \left[\dfrac{T}{2},\ T\right] \end{cases} \tag{5.2.12}$$

where $l_r(\tau)$ stands for the r^{th} column of the matrix L.

Therefore, the two matrices of (5.2.12) in each half of the interval are arranged as:

$$\begin{pmatrix} \alpha_{11}^{(1)} & \cdots & \alpha_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ \alpha_{N1}^{(1)} & \cdots & \alpha_{NN}^{(1)} \end{pmatrix} = \begin{pmatrix} \sigma_1^2\dfrac{T}{2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_1^2\dfrac{T}{2} \end{pmatrix}^{-1} \bullet \begin{pmatrix} a_1(t_1) & \cdots & a_1(t_{\frac{N}{2}}) \\ \vdots & \ddots & \vdots \\ a_N(t_1) & \cdots & a_N(t_{\frac{N}{2}}) \end{pmatrix} \bullet \begin{pmatrix} l_{11} & \cdots & l_{1N} \\ \vdots & \ddots & \vdots \\ l_{\frac{N}{2},1} & \cdots & l_{\frac{N}{2},N} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_{11}^{(2)} & \cdots & \alpha_{1N}^{(2)} \\ \vdots & \ddots & \vdots \\ \alpha_{N1}^{(2)} & \cdots & \alpha_{NN}^{(2)} \end{pmatrix} = \begin{pmatrix} \sigma_2^2\dfrac{T}{2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_2^2\dfrac{T}{2} \end{pmatrix}^{-1} \bullet \begin{pmatrix} a_1(t_{\frac{N}{2}+1}) & \cdots & a_1(t_N) \\ \vdots & \ddots & \vdots \\ a_N(t_{\frac{N}{2}+1}) & \cdots & a_N(t_N) \end{pmatrix} \bullet \begin{pmatrix} l_{\frac{N}{2}+1,1} & \cdots & l_{\frac{N}{2}+1,N} \\ \vdots & \ddots & \vdots \\ l_{N,1} & \cdots & l_{N,N} \end{pmatrix}$$

(5.2.13)

Thus Galerkin matrices of both halves are accordingly given by:

$$\begin{pmatrix} g_1^{(1)}(t_1) & \cdots & g_1^{(1)}(t_{\frac{N}{2}}) \\ \vdots & \ddots & \vdots \\ g_1^{(N)}(t_1) & \cdots & g_1^{(N)}(t_{\frac{N}{2}}) \end{pmatrix} = \begin{pmatrix} \alpha_{11}^{(1)} & \cdots & \alpha_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ \alpha_{N1}^{(1)} & \cdots & \alpha_{NN}^{(1)} \end{pmatrix}^T \bullet \begin{pmatrix} a_1(t_1) & \cdots & a_1(t_{\frac{N}{2}}) \\ \vdots & \ddots & \vdots \\ a_N(t_1) & \cdots & a_N(t_{\frac{N}{2}}) \end{pmatrix}$$

$$\begin{pmatrix} g_2^{(1)}(t_{\frac{N}{2}+1}) & \cdots & g_2^{(1)}(t_N) \\ \vdots & \ddots & \vdots \\ g_2^{(N)}(t_{\frac{N}{2}+1}) & \cdots & g_2^{(N)}(t_N) \end{pmatrix} = \begin{pmatrix} \alpha_{11}^{(2)} & \cdots & \alpha_{1N}^{(2)} \\ \vdots & \ddots & \vdots \\ \alpha_{N1}^{(2)} & \cdots & \alpha_{NN}^{(2)} \end{pmatrix}^T \bullet \begin{pmatrix} a_1(t_{\frac{N}{2}+1}) & \cdots & a_1(t_N) \\ \vdots & \ddots & \vdots \\ a_N(t_{\frac{N}{2}+1}) & \cdots & a_N(t_N) \end{pmatrix}$$

(5.2.14)

where the subscripts in g_1 and g_2 refer to the regions $\left[0,\ \frac{T}{2}\right]$ and $\left[\frac{T}{2},\ T\right]$.

We still need to determine the coefficient matrix $\left\{\chi_i(t_j)\right\}$ for calculating the weighting matrix (5.2.7). To this end, we will first calculate the elements $\left\{h_{\mu\nu}\right\}$:

$$\begin{pmatrix} h_{11}^{(1)} & \cdots & h_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ h_{N1}^{(1)} & \cdots & h_{NN}^{(1)} \end{pmatrix} = \begin{pmatrix} \alpha_{11}^{(1)} & \cdots & \alpha_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ \alpha_{N1}^{(1)} & \cdots & \alpha_{NN}^{(1)} \end{pmatrix}^T \bullet \begin{pmatrix} \sigma_1^2 \frac{T}{2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_1^2 \frac{T}{2} \end{pmatrix} \bullet \begin{pmatrix} \alpha_{11}^{(1)} & \cdots & \alpha_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ \alpha_{N1}^{(1)} & \cdots & \alpha_{NN}^{(1)} \end{pmatrix}$$

(5.2.15)

for the first region, and the elements $\left\{h_{\mu\nu}^{(2)}\right\}$ for the second region will be calculated in the same manner.

Thus, the coefficient matrices $\left\{\chi_i^{(1)}(t_j)\right\}$ and $\left\{\chi_i^{(2)}(t_j)\right\}$ for $\left[0,\ \frac{T}{2}\right]$ and $\left[\frac{T}{2},\ T\right]$ are respectively obtained as:

$$\begin{pmatrix} \chi_1^{(1)}(t_1) & \cdots & \chi_N^{(1)}(t_1) \\ \vdots & \ddots & \vdots \\ \chi_1^{(1)}(t_{\frac{N}{2}}) & \cdots & \chi_N^{(1)}(t_{\frac{N}{2}}) \end{pmatrix} = \left(\begin{pmatrix} h_{11}^{(1)}+1 & \cdots & h_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ h_{N1}^{(1)} & \cdots & h_{NN}^{(1)}+1 \end{pmatrix}^{-1} \bullet \begin{pmatrix} l_{11} & \cdots & l_{1N} \\ \vdots & \ddots & \vdots \\ l_{\frac{N}{2},1} & \cdots & l_{\frac{N}{2},N} \end{pmatrix}^T \right)^T$$

$$\begin{pmatrix} \chi_1^{(2)}(t_{\frac{N}{2}+1}) & \cdots & \chi_N^{(2)}(t_{\frac{N}{2}+1}) \\ \vdots & \ddots & \vdots \\ \chi_1^{(2)}(t_N) & \cdots & \chi_N^{(2)}(t_N) \end{pmatrix} = \left(\begin{pmatrix} h_{11}^{(2)}+1 & \cdots & h_{1N}^{(2)} \\ \vdots & \ddots & \vdots \\ h_{N1}^{(2)} & \cdots & h_{NN}^{(2)}+1 \end{pmatrix}^{-1} \bullet \begin{pmatrix} l_{\frac{N}{2}+1,1} & \cdots & l_{\frac{N}{2}+1,N} \\ \vdots & \ddots & \vdots \\ l_{N,1} & \cdots & l_{N,N} \end{pmatrix}^T \right)^T$$

(5.2.16)

It is to be noted that the superscripts in $h^{(1)}$, $h^{(2)}$ and $\chi^{(1)}$, $\chi^{(2)}$ in (5.2.16) again indicate the regions to which these matrices belong. Finally, the weighting matrices of our optimal linear operator in this case read respectively as:

$$\begin{pmatrix} g_1(t_1,t_1) & \cdots & g_1(t_1,t_{\frac{N}{2}}) \\ \vdots & \ddots & \vdots \\ g_1(t_{\frac{N}{2}},t_1) & \cdots & g_1(t_{\frac{N}{2}},t_{\frac{N}{2}}) \end{pmatrix} = \begin{pmatrix} \chi_1^{(1)}(t_1) & \cdots & \chi_N^{(1)}(t_1) \\ \vdots & \ddots & \vdots \\ \chi_1^{(1)}(t_{\frac{N}{2}}) & \cdots & \chi_N^{(1)}(t_{\frac{N}{2}}) \end{pmatrix} \bullet \begin{pmatrix} g_1^{(1)}(t_1) & \cdots & g_1^{(1)}(t_{\frac{N}{2}}) \\ \vdots & \ddots & \vdots \\ g_1^{(N)}(t_1) & \cdots & g_1^{(N)}(t_{\frac{N}{2}}) \end{pmatrix}$$

$$\begin{pmatrix} g_2(t_{\frac{N}{2}+1},t_{\frac{N}{2}+1}) & \cdots & g_2(t_{\frac{N}{2}+1},t_N) \\ \vdots & \ddots & \vdots \\ g_2(t_N,t_{\frac{N}{2}+1}) & \cdots & g_2(t_N,t_N) \end{pmatrix} = \begin{pmatrix} \chi_1^{(2)}(t_{\frac{N}{2}+1}) & \cdots & \chi_N^{(2)}(t_{\frac{N}{2}+1}) \\ \vdots & \ddots & \vdots \\ \chi_1^{(2)}(t_N) & \cdots & \chi_N^{(2)}(t_N) \end{pmatrix} \bullet \begin{pmatrix} g_2^{(1)}(t_{\frac{N}{2}+1}) & \cdots & g_2^{(1)}(t_N) \\ \vdots & \ddots & \vdots \\ g_2^{(N)}(t_{\frac{N}{2}+1}) & \cdots & g_2^{(N)}(t_N) \end{pmatrix}$$

(5.2.17)

Finally, the theoretical mean square errors of this operator on $\left[0,\ \frac{T}{2}\right]$ and $\left[\frac{T}{2},\ T\right]$ are respectively measured by those elements on the principal diagonal of the following matrix products, according to (3.2.33):

$$\| e_1 \|^2 = \begin{pmatrix} \chi_1^{(1)}(t_1) & \cdots & \chi_N^{(1)}(t_1) \\ \vdots & \ddots & \vdots \\ \chi_1^{(1)}(t_{\frac{N}{2}}) & \cdots & \chi_N^{(1)}(t_{\frac{N}{2}}) \end{pmatrix} \bullet \begin{pmatrix} h_{11}^{(1)}+1 & \cdots & h_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ h_{N1}^{(1)} & \cdots & h_{NN}^{(1)}+1 \end{pmatrix} \bullet \begin{pmatrix} \chi_1^{(1)}(t_1) & \cdots & \chi_N^{(1)}(t_1) \\ \vdots & \ddots & \vdots \\ \chi_1^{(1)}(t_{\frac{N}{2}}) & \cdots & \chi_N^{(1)}(t_{\frac{N}{2}}) \end{pmatrix}^T$$

$$\| e_2 \|^2 = \begin{pmatrix} \chi_1^{(2)}(t_{\frac{N}{2}+1}) & \cdots & \chi_N^{(2)}(t_{\frac{N}{2}+1}) \\ \vdots & \ddots & \vdots \\ \chi_1^{(2)}(t_N) & \cdots & \chi_N^{(2)}(t_N) \end{pmatrix} \bullet \begin{pmatrix} h_{11}^{(2)}+1 & \cdots & h_{1N}^{(2)} \\ \vdots & \ddots & \vdots \\ h_{N1}^{(2)} & \cdots & h_{NN}^{(2)}+1 \end{pmatrix} \bullet \begin{pmatrix} \chi_1^{(2)}(t_{\frac{N}{2}+1}) & \cdots & \chi_N^{(2)}(t_{\frac{N}{2}+1}) \\ \vdots & \ddots & \vdots \\ \chi_1^{(2)}(t_N) & \cdots & \chi_N^{(2)}(t_N) \end{pmatrix}^T$$

(5.2.18)

5.2.4 Analysis of Result

Figure 5-3 shows the constructed signal using this Wiener type filter and the original signal which has been depicted in Figure 5-2. It is to be borne in mind that this linear

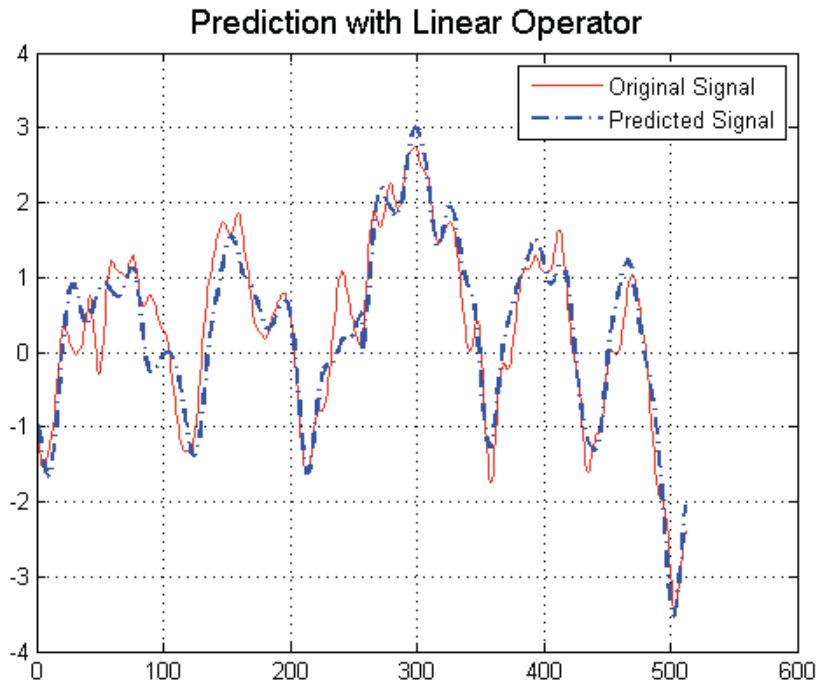

Figure 5-3. Reconstructed Signal using Linear Operator and its Original Signal, N=512, from piecewise Noise: $\sigma_1^2 = 4.0$, $\sigma_2^2 = 1.0$

predictor is theoretically suboptimal in the present case, as was mentioned earlier. As we can see, the tendency of the variation of the original signal is fully recovered by the

reconstructed one with a fairly satisfactory performance, although the effect of prediction is not that impressive. Some subtle variations of the signal cannot be reflected by this operator, particularly in the first half region with noise variance $\sigma_1^2 = 4$. Consequently, we can claim that the ability of prediction is relatively weaker in the region with stronger noise.

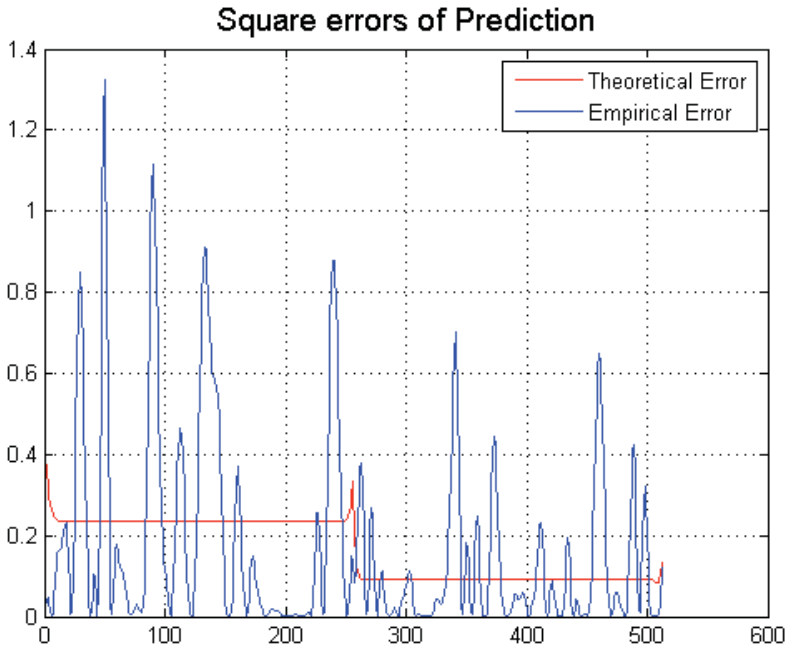

Figure 5-4. Mean-Square Errors of the Prediction, with both Theoretical and Empirical Values, N=512, $\sigma_1^2 = 4.0$ and $\sigma_2^2 = 1.0$

Figure 5-4 is an illustration of the theoretical mean-square errors obtained from (5.2.18), together with the empirical errors produced from observation. Most of the theoretical errors in the first half region (with $\sigma_1^2 = 4.0$) are around 0.23, and the other half are about 0.09; also, the biggest empirical square error in the first region climbs up over 1.3, while this value is only about 0.7 in the lower variance region. Additionally, the empirical

prediction errors suffer from a more violent fluctuation with bigger noise. All of these observations imply that the linear operator is more sensitive to the noise with higher intensity, resulting in poorer prediction. In addition, it may be worth mentioning an interesting fact, that at the boundary of both regions, the theoretical error (red line in Figure 5-4) tends to be blown up to a certain amount, normally known as Gibbs' phenomenon (Blais, 1988).

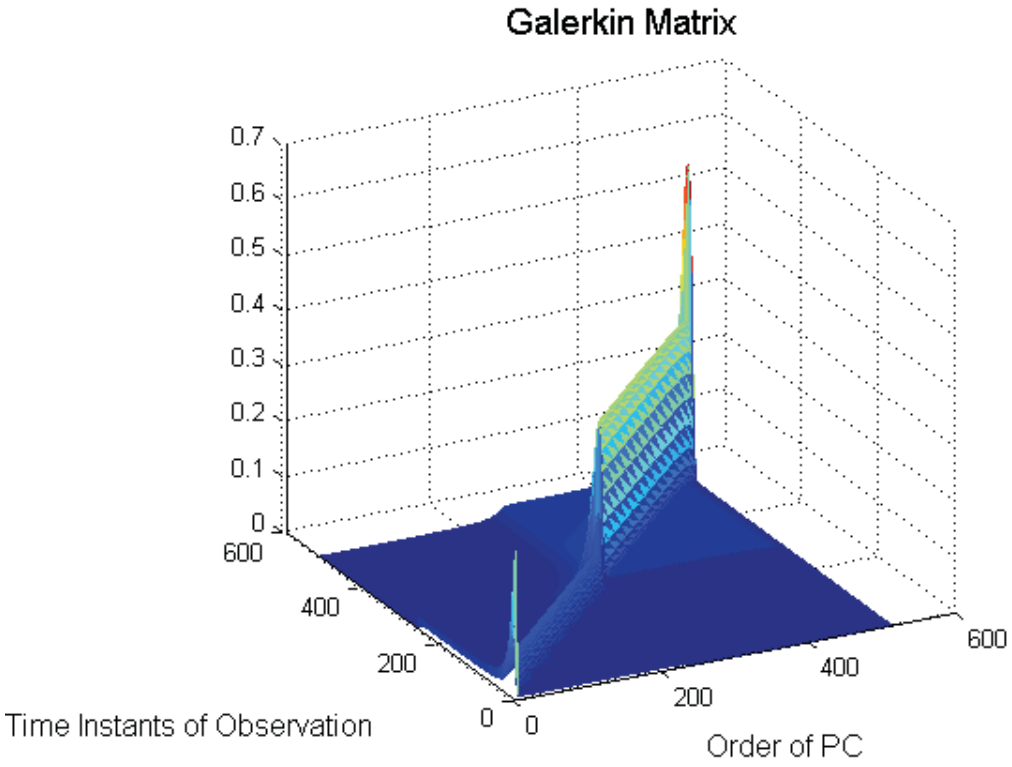

Figure 5-5. 3D Visualisation of Galerkin Matrix

Figure 5-5 provides a 3D visualisation of the Galerkin matrix, composed of the two matrices in (5.2.14). Basically it shows the magnitudes of each entry of Galerkin matrix, so that we can explore the effect of weighting for the observation at each instant by each order of PC (or Fourier basis). One axis corresponds to the observation instants, while the

other indicates the order of PCs. Firstly, we can see a ‘diagonal effect’ of this Galerkin matrix, which suggests that every PC should be used for weighting the observation points; and for each point, only a few PCs should contribute dominantly to such weighting, or say, each PC should have a major effect on only a small collection of observation points. Also, we can clearly see a difference of this weighting effect between the two half regions with different noise intensities, where the Galerkin form takes significantly larger values in the region with smaller noise intensity. Comparing with the Galerkin matrix of Haar wavelets (Keller, 2000), the Galerkin matrix of PCs have a smoother look, this may be due to the fact that the Haar wavelets are more of a discontinuous form, while the PCs are actually harmonic functions. Finally, Gibbs’ phenomenon which appeared in the theoretical mean-square errors (red line in Figure 5-4) can also be observed here in the visualisation of Galerkin matrix.

It is expedient to mention that if we had not used the piecewise prediction according to the individual variances of each region, we would have been led to a very bad outcome of prediction. For example, if we were to combine the information of PCs in two regions together and carry out the prediction as a whole, the result of such prediction would be rather poor. What is more, the bigger the contrast between variances within piecewise noise, the more disastrous the result of prediction is for the lower intensity region.

5.2.5 Concluding Remarks

There are several points that should be brought to our attention regarding this example:

i. There is no reduction of dimensionality here in this example, for in the case of white noise, the PCs agree with the Fourier basis. Also, the number of PCs should be chosen the same as the dimension of the data vector.

ii. If the noise has different intensities in parts of the observation period, as in the present case, to attain the best result, the prediction has to be performed separately for each part.

iii. The quality of the linear operator weakens as the noise variance increases, with a larger mean value, and higher variances, of prediction error.

iv. One always has to face a trade-off when finding the most suitable operator in each particular case. There is always the alternative between a suboptimal linear operator with considerable ease of calculation, and an optimal non-linear operator requiring a lot of complicated computations. One has to adjust the selection according to the objective.

In this example, the signal is a linear form whose number of terms is the same as that of the total samples. The computation of the non-linear regression operator requires an enormous effort in both quantity and complexity. Instead, the optimal linear operator only requires several multiplications of matrices.

5.3 Extrapolation of a Signal which is a Linear Function of Time

As the last example in the chapter, we will visit the extrapolation problem of a random signal which is a linear function of time. The model this time is:

$$\begin{aligned} Z(t) &= U_1 + U_2 t + X(t) \\ W(s) &= U_1 + U_2 s \end{aligned} \tag{5.3.1}$$

Here, stands for those instants within the observation period, while s represents the instant at which the value of signal is to be extrapolated. U_1 and U_2 are two independent zero-mean Gaussian variables, $X(t)$ is a centred 1st order Gauss-Markov process.

This problem can be regarded as a particular case of a more general one, i.e., the extrapolation of a useful random signal which is a polynomial of time:

$$W(s) = U_1 + U_2\, s + U_3\, s^2 + \cdots \tag{5.3.2}$$

U_i are either random variables with known distribution or unknown parameters. One of the applications of such problem in geophysics or geodynamics could be the extrapolation of the *secular drift* in *polar motion*, for the time series of polar motion can be regarded as a stochastic process (Keller, 2004). There are five constituents in the observed polar motion (both in the decomposed X and Y components): annual prograde and retrograde, Chandler wobble, a smooth secular drift and a noise (Gibert and Le Mouel, 1998). After separating and removing the annual prograde and retrograde, along with the Chandler wobble (such procedures can be done by wavelet analysis; see Keller 2004), there will only be the secular drift and the noise involved in the time series of polar motion, thus the drift could be adequately approximated by a linear time function (or any degree of polynomial) in order to predict its future trend.

5.3.1 Methodology

Our goal is to use the method of canonical expansion to build the optimal linear operator for such an extrapolation problem. In fact, the optimal non-linear operator in this case

reduces to the linear one, as we will prove soon in this section. Thus, we will formulate the Wiener-Hopf type optimal linear operator according to the scheme in Chapter Two. Also, we will not only establish our linear estimator by using the K-L method to produce the PCs, but also employ the Akimov's construction of the PCs, introduced in Section 2.4, to investigate the effect of extrapolation using these two choices of PCs.

5.3.1.1 Optimal linear operator

In the present case, the signal depends on two independent random variables. So according to our theory, the optimal linear operator takes the form of a combination of two weighting functions:

$$g(s,t) = \chi_1(s)\cdot g^{(1)}(t) + \chi_2(s)\cdot g^{(2)}(t) \tag{5.3.3}$$

with $g^{(1)}(t)$ and $g^{(2)}(t)$ as the solution of the following two integral equations:

$$\begin{cases} \int_T g^{(1)}(t)\,K_X(t,\tau)\,d\tau = 1 \\ \int_T g^{(2)}(t)\,K_X(t,\tau)\,d\tau = t \end{cases} \tag{5.3.4}$$

where the covariance function is that of the 1st order Gauss-Markov process:

$$K_X(t,\tau) = \sigma^2 e^{-\beta|t-\tau|} \tag{2.4.6}$$

Both solutions are provided by Galerkin's method in terms of principal components:

$$\begin{cases} g^{(1)}(t) = \sum_v \alpha_{v1}\,a_v(t) \\ g^{(2)}(t) = \sum_v \alpha_{v2}\,a_v(t) \end{cases}, \qquad \text{where} \begin{cases} \alpha_{v1} = \dfrac{1}{D_v}\int_T a_v(t)\,dt \\ \alpha_{v2} = \dfrac{1}{D_v}\int_T a_v(t)\,t\,dt \end{cases} \tag{5.3.5}$$

The coefficients (Galerkin's matrix) can be found according to (3.2.10):

$$\begin{pmatrix} 1 \\ s \end{pmatrix} = \begin{pmatrix} h_{11} + \dfrac{1}{\mu_{11}} & h_{12} \\ h_{21} & h_{22} + \dfrac{1}{\mu_{22}} \end{pmatrix} \bullet \begin{pmatrix} \chi_1(s) \\ \chi_2(s) \end{pmatrix} \tag{5.3.6}$$

Again, s is the point at which the value of the linear time function is to be extrapolated. Ultimately, we have our optimal linear estimator, from the Wiener-Hopf equation, in this case of extrapolation of a linear time function:

$$\widetilde{\mathrm{A}}Z(t) = \int_T \left(\chi_1(s)\, g^{(1)}(t) + \chi_2(s)\, g^{(2)}(t) \right) Z(t)\, dt \tag{5.3.7}$$

The mean square error of this estimator, for every time instant, is given by:

$$\| e(s) \|^2 = \begin{pmatrix} \chi_1(s) \\ \chi_2(s) \end{pmatrix}^T \bullet \begin{pmatrix} h_{11} + \dfrac{1}{\mu_{11}} & h_{12} \\ h_{21} & h_{22} + \dfrac{1}{\mu_{22}} \end{pmatrix} \bullet \begin{pmatrix} \chi_1(s) \\ \chi_2(s) \end{pmatrix} \tag{5.3.8}$$

5.3.1.2 Optimal non-linear operator

In this case, the optimal non-linear operator happens to coincide with the linear one, for the variables in both signal and noise are Gaussian. In fact, this well-known fact can be proved by the same approach of using the PCs, provided in Section 4.2.

Since the observation can be transformed into a combination of variables Z_ν defined by (4.2.2), the conditional expectation operator (4.1.6) now takes the form:

$$E\left[W(s) \middle| Z(t) \right] = E\left[W(s) \middle| Z_1, Z_2, \ldots, Z_N \right] \tag{5.3.9}$$

In particular, as we have mentioned in Chapter Four, if both the signal and noise are Gaussian, $\{Z_v\}$ are also Gaussian, then the right-hand side of (5.3.9) can be written as:

$$E\left[W(s)\middle| Z_1, Z_2, \ldots, Z_N\right] = \sum_v \frac{k_v}{D_v} Z_v \tag{5.3.10}$$

Here, k_v denotes the covariance between the signal $W(s)$ and the generated variable Z_v. Therefore, we have arrived at the conclusion that the optimal regression operator in the Gaussian case is actually a linear operator (in general, it should be a non-homogeneous linear operator if the signal is not centred, see Pugachev, 1965).

5.3.2 Noise Model

For the 1st order G-M process, we have proposed two methods of finding its principal components. The first is based on the K-L expansion theorem, and derived from the solution of the integral equation (2.4.7) directly; an alternative way is provided in Section 2.4.2, where the PCs are arranged as Jordan block. It is to be remembered that, due to its nature of recursion, there is no reduction of dimensionality in the latter case. We will avail ourselves of these two approaches, in this section, to establish the PCs and the forms of the linear operator.

5.3.2.1 Karhunen-Loève basis

As just mentioned, the PCs are obtained by solving the integral equation (2.4.7) over the interval $[0, T]$, and this is already done in Section 2.4, where the result is given by (Section 2.4.2.1):

$$a_v(t) = \sin\left[\omega_v\left(t - \frac{T}{2}\right) + \frac{v\pi}{2}\right] \qquad v = 1, 2, \ldots \tag{5.3.11}$$

where ω_v are roots in ascending order of the following equation:

$$\tan(\omega T) = -\frac{2\beta\omega}{\beta^2 - \omega^2} \tag{2.4.9}$$

Yet, the solution to this equation is not straightforward; it has to be solved approximately by a numerical method such as *Newton's method*. To avoid such heavy computation, in spite of its exactness, we will use the *spectral decomposition* (or *eigen-decomposition*) of the variance-covariance matrix to obtain the eigenvalues and eigenvectors in order to form the PC, i.e. we obtain the matrices Φ and Λ from the following matrix equation:

$$K_X = \Phi \cdot \Lambda \cdot \Phi^T \tag{2.2.6}$$

Each column vector of Φ stands for a PC. It is to be noted that, the principal component can be chosen arbitrarily subject to only a constant; in other words, we can multiply any number to a PC. In accordance with our solution in Section 2.4.2, the PCs (5.3.11) in the present case are selected as a scalar product of the eigenvectors of Φ:

$$a_v(t_i) = \sqrt{\frac{T + \lambda_v}{2}}\,\varphi_v(t_i) \tag{5.3.11a}$$

Subsequently, instead of the eigenvalues of variance-covariance matrix K, the corresponding variances of each variables in the basis of canonical expansion, arranged in a descending order $D_1 \geq D_2 \geq D_3 \geq \cdots > 0$, are given by:

$$D_v = \frac{\sigma^2}{2}\lambda_v(T + \lambda_v) \tag{2.4.11}$$

Once we have the PCs (5.3.11a) and the variances (2.4.11), we can calculate the optimal linear operator by using the principal components defined in K-L sense, as the procedure we have outlined earlier. The matrix $\{\alpha_{vr}\}$ in this case is expressed as:

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \vdots & \vdots \\ \alpha_{M1} & \alpha_{M2} \end{pmatrix} = \begin{pmatrix} D_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_M \end{pmatrix}^{-1} \bullet \begin{pmatrix} a_1(t_1) & \dots & a_1(t_N) \\ \vdots & \ddots & \vdots \\ a_M(t_1) & \cdots & a_M(t_N) \end{pmatrix} \bullet \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix} \tag{5.3.12}$$

where M is the degree of the canonical expansion. Subsequently, the Galerkin's matrix now takes the form of a two-dimensional vector:

$$\begin{pmatrix} g^{(1)}(t_1) & g^{(2)}(t_1) \\ g^{(1)}(t_2) & g^{(2)}(t_2) \\ \vdots & \vdots \\ g^{(1)}(t_N) & g^{(2)}(t_N) \end{pmatrix}^T = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \vdots & \vdots \\ \alpha_{M1} & \alpha_{M2} \end{pmatrix}^T \bullet \begin{pmatrix} a_1(t_1) & \dots & a_1(t_N) \\ \vdots & \ddots & \vdots \\ a_M(t_1) & \cdots & a_M(t_N) \end{pmatrix} \tag{5.3.13}$$

To complete the solution of the Fredholm equation, we only need the following matrix to be inserted to into (5.3.6) to calculate the coefficients $\chi_1(s)$ and $\chi_2(s)$:

$$\begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \vdots & \vdots \\ \alpha_{M1} & \alpha_{M2} \end{pmatrix}^T \bullet \begin{pmatrix} D_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_M \end{pmatrix} \bullet \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \vdots & \vdots \\ \alpha_{M1} & \alpha_{M2} \end{pmatrix} \tag{5.3.14}$$

5.3.2.2 Jordan block – Akimov's method

Alternatively, for equidistant observation points, PCs can also be obtained by the method of Akimov, where each PC is given by each row of the following Jordan block (Section 2.4.2.2):

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -q & 1 & \cdots & 0 & 0 \\ 0 & -q & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -q & 1 \end{pmatrix} \quad \text{where } q = e^{-\beta\tau} \tag{2.4.19}$$

All the remaining quantities that are essential to constructing the optimal linear estimator can be obtained in the same way as (5.3.12) – (5.3.14), with only a substitution of the matrix of the PC in the K-L sense by (2.4.19).

5.3.3 Simulation of First-order Gauss-Markov Process

Unlike white noise, there is no direct way of generating the 1st order G-M process, and we have to invent a system or shaping filter (Nassar, 2003) of doing so. To this end, we propose a discrete recursive equation (Gelb, 1974; Blais, 1988) for the generation:

$$x_{k+1} = qx_k + w_k \quad \text{where } q = e^{-\beta\tau} \text{ and } w_k \sim N(0, \sigma_W^2) \tag{5.3.15}$$

It is easily seen that if the first input x_0 is a Gaussian variable, then all the subsequent variables are Gaussian; also, the value of the present variable only depends on that of the one which precedes it, thus it clearly satisfies a first-order Markov property. Once the sequence is generated, the second central moments of this stationary process should be estimated from the simulation by:

$$K_X(n\tau) = \frac{1}{N-n}\sum_{i=1}^{N-n}\left(x(t_{i+n}) - \widehat{m_X}\right)\left(x(t_i) - \widehat{m_X}\right) \tag{5.3.16}$$

where $\widehat{m_X}$ is the mean value of the sample data, and $n\tau$ is the time lag between any pair of samples. Thus, the estimated variance of the process is consequently given by:

$$\widehat{\sigma_X^2} = K_X(0) = \frac{1}{N}\sum_{i=1}^{N}\left(x(t_i) - \widehat{m_X}\right)^2 \tag{5.3.17}$$

In practice, the value of β is not known a priori (it is the reciprocal of correlation time of process), and it has to be estimated as well from the sample covariance $\widehat{K_X}(n\tau)$ by (5.3.16) for a sequence of time lags using a very large data sample.

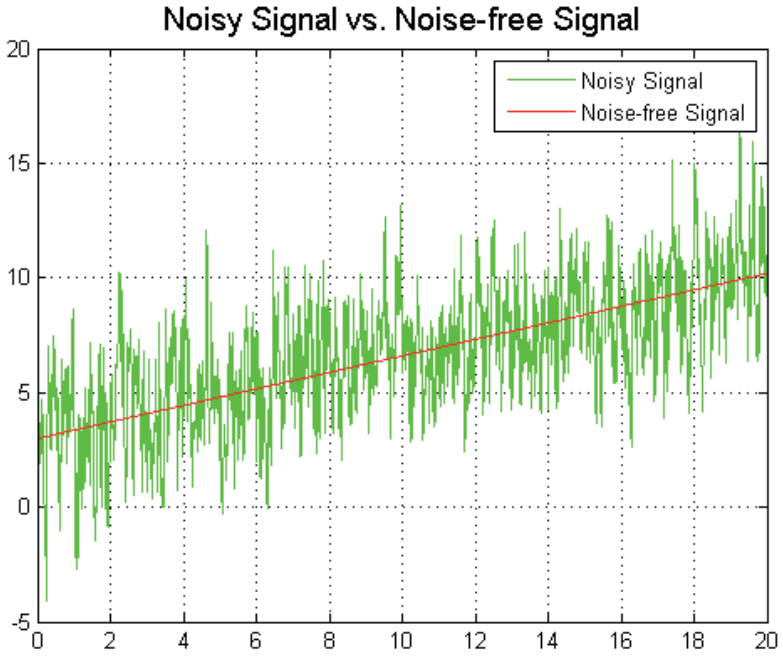

Figure 5-6. One Realisation of Random Linear Signal and the Observation of this Signal plus the Simulated 1st Order Gauss-Markov Process

Figure 5-6 shows one realisation of the true signal and the noisy signal where the constructed 1st order G-M process by (5.3.15) is added upon the original signal. The observation period is chosen as $[0, T]$, with $T = 20$ and $\Delta t = 0.01$. U_1 and U_2 are Gaussian variables satisfying respectively $N(0, 3)$ and $N(0, 1)$. The parameter β of the

G-M noise is selected as 25 (this means the correlation time of the process is 0.04), and the sample variance $\widehat{\sigma_X^2}$ is estimated by (5.3.17) as 4.61499.

5.3.4 Analysis of Result

Initially, we will address the question: how many principal components are needed in the weighting functions or Galerkin's solution (5.3.4) so that the linear filter could give a close approximation to the true signal, or equivalently speaking, how many terms in the K-L expansion of the G-M process are sufficient to give a close approximation to the process itself ? As we can see, Figure 5-7 depicts the extrapolated signal of the original

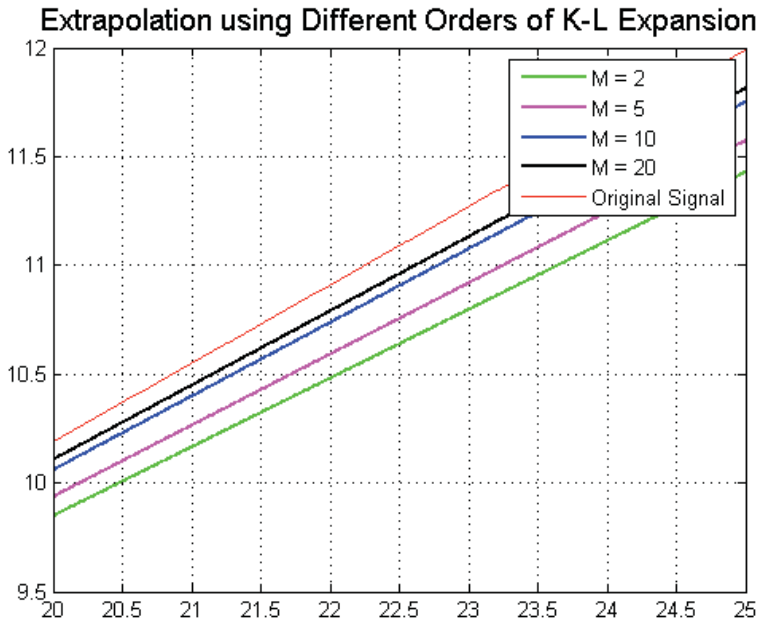

Figure 5-7. Extrapolation using Different Orders of Karhunen-Loève Expansion. Orders of the PCs: M = 2, 5, 10 & 20

signal in Figure 5-6 on the interval $[T, T+S]$, with $S = 5$, for different orders of principal components. We can see that, the more PCs (corresponding to the number of

first terms in the canonical expansion of the G-M noise) we use to construct the linear estimator, the better approximation we obtain to the real random signal. Both $M = 10$ and $M = 20$ (using the first 10 and 20 PCs) give a very close extrapolation to the real signal, with the errors in both cases remaining under 0.05.

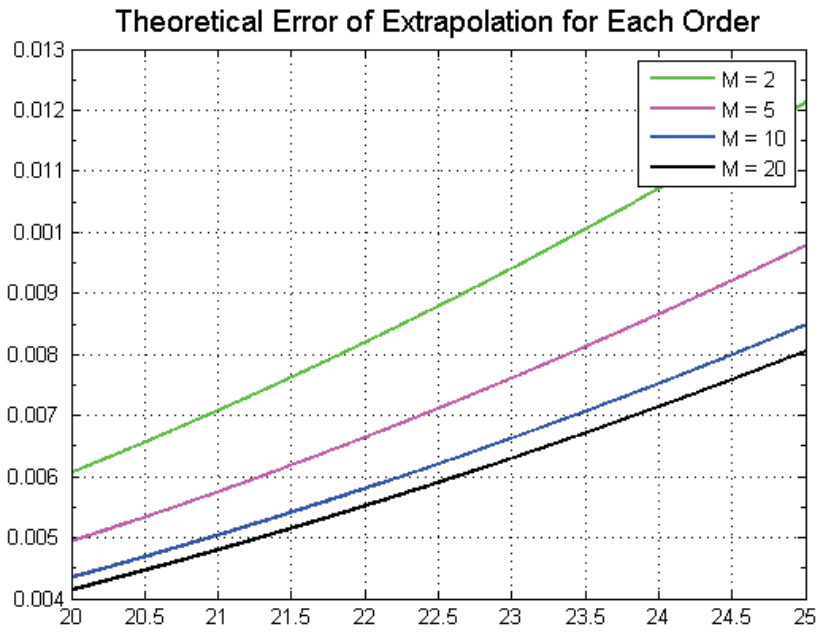

Figure 5-8. Theoretical Minimum Mean Square Errors of Optimal Linear Estimation for Each Order of PC

Figure 5-8 provides the theoretical mean square errors of estimation for each order of the PC, derived from (5.3.9), in order to evaluate the optimality of these estimators. We can observe that, the further the time instant, at which the signal is to be extrapolated, stays away from the observation period, the bigger the extrapolation error is. This is quite a natural result, since apparently the observation should produce more accurate result for nearer points. Moreover, a lateral comparison between those lines suggests that, the smaller the number of PCs (less order of expansion) is, the larger this error will become

as the instant moves along. For example, at the last instant $s = 25$, the error of extrapolation for $M = 20$ still stays under 0.008, but that of $M = 2$ has already climbed up over 0.012. Based on this fact, we can claim that, for a reasonable number of PCs of selection, its curve of errors can be fairly approximated by a linear function with very small slope and intercept, and as a result, we can have a very good result of extrapolation.

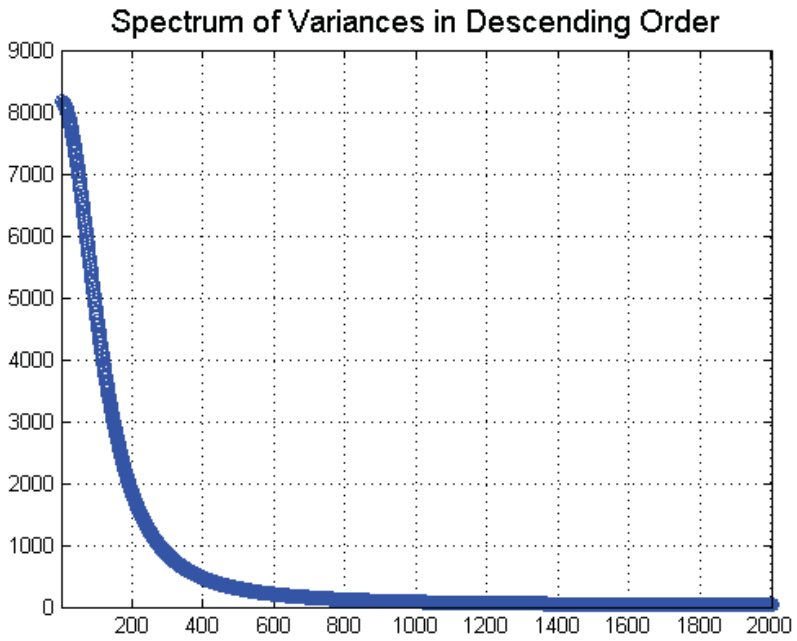

Figure 5-9. Continuous Spectrum of Variances of Basis Variables for K-L Expansion, arranged in a Descending Order

Figure 5-9 shows the continuous spectrum of variances defined by (2.4.11) of basis variables of the expansion. They are arranged in a descending order, with roughly the first 200 values account for most of the variability of the observation data, though according to our result shown in Figure 5-7, the first few terms are sufficient for the estimation purpose. In this regard, Table 5-3 shows the percentage of sum of those

variances D_v, corresponding to each order of PCs in Figure 5-7, over the trace of the diagonal matrix D_v:

$$Percentage = \frac{\sum_{v=1}^{M} D_v}{trace(diag(D_v))} \tag{5.3.20}$$

ORDER OF EXPANSION	$M = 2$	$M = 5$	$M = 10$	$M = 20$
VARIANCE PERCENTAGES	1%	3%	6%	12%

Table 5-3. Percentage of Variances for Each Order of Expansion

We can see that, the first 20 terms of PCs would account for only 12% of the total variances, yet it has already given a very close extrapolation to the original signal. This fact illustrates the difference between a *spectral decomposition* (SD) and *singular value decomposition* (SVD). In SVD, the data matrix refers to a time series for a random vector, and the SVD reduces the dimensionality of this matrix to only a small number of singular values which contain most of the variances of the matrix (a very high percentage in all); but in SD, we are simply dealing with one time series, and although the dimensionality is as well reduced significantly by this decomposition for the estimation purpose, it can retain only a small amount of variation of the whole data sequence. In another turn of phrase, the SVD algorithm provides a *discrete* spectrum of data set in which several terms can claim most of the second moments, while the SD produces a *continuous* spectrum in which the second moments are comparatively averaged in these

values in a descending order (Figure 5-9). Nevertheless, this small percentage will not affect the excellent performance of estimation at all.

Finally, we will compare the performance of extrapolation using two different choices of PCs: one is derived from the K-L expansion, as is used earlier; the other is obtained by Akimov's method, in which the PCs are arranged in a Jordan Block (JB) (See Section 5.3.2.2). Figure 5-10 shows the result of two extrapolations for another realisation of our random linear signal: one is by K-L expansion when $M = 20$, the other is by this Jordan block selection; the variance of the G-M noise this time is around 5.23. We can see that, both choices of PCs produce very accurate extrapolation of the original signal over the region $[T, T+S]$.

To compare the optimality of these two extrapolations, again, we need the theoretical mean-square errors of both optimal linear operators. Figure 5-11 describes the minimum mean-square errors of these two operators, obtained from (5.3.9). Obviously, despite the result in Figures 5-10, the extrapolation using K-L expansion with 20 PCs has a significantly lower mean errors than the PCs with JB, with all the errors of JB within the range of 0.08 and 0.16 while those with K-L remain a level of 0.01. This means in general (for many realisations), the selection of K-L expansion should give better results of extrapolation than those of PCs with JB structure. However, those PCs with JB are very easy and straightforward to compute, comparing to the spectral decomposition needed in K-L expansion.

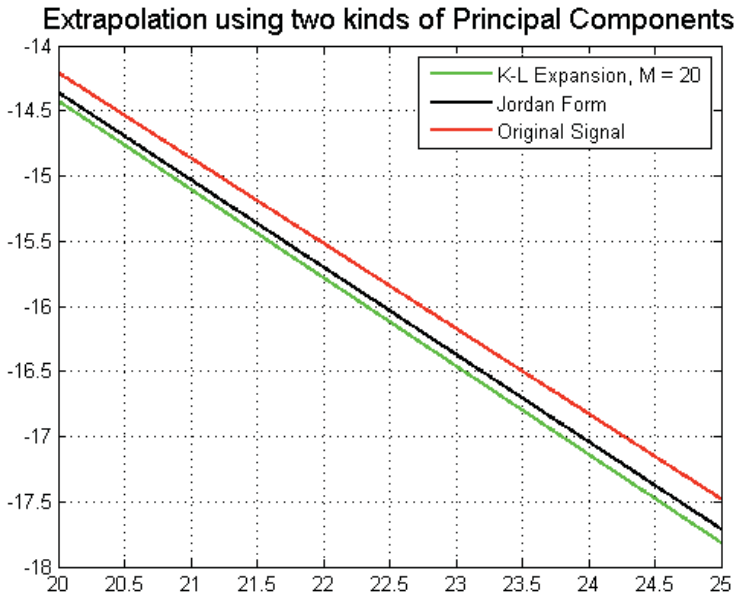

Figure 5-10. Extrapolation using both the K-L Expansion with M=20 and Akilov's Selection of the PCs having a Jordan Block Structure

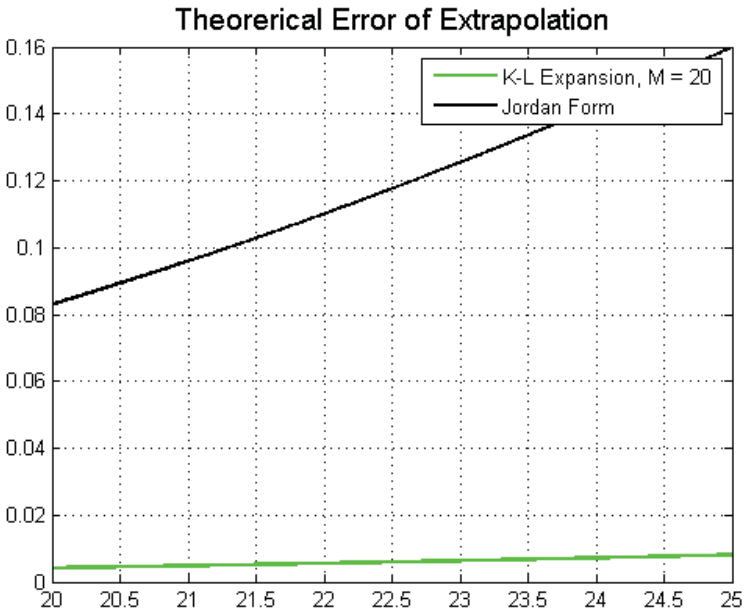

Figure 5-11. Theoretical Minimum Mean Square Errors of Optimal Linear Estimation using Two Choices of the PCs (K-L and JB)

Chapter Six: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

6.1 Summary

In the present work, we have carried out a detailed study on the fundamental theory of canonical expansions, and its applications in both linear and non-linear estimation under MMSE principle. Also, we have examined the performances of both operators in different cases of signal and noise models. From the theoretical exploration in Chapter Two, we have seen that this theory is simply one class of the more universal theory of orthogonal functions, and it is bound up with the theories of linear algebra and integral equations (or with the eigenvalue and eigenvector problems in both theories). In consequence, the Karhunen-Loève expansion (or PCA) is an important special case of the canonical expansion, although it is always possible to obtain a different manner of expansion. We have also seen that, theoretically, almost all the common random processes are capable of such expansions, so the techniques proposed in the later chapters can be adequately applied to every noise model in practice.

In Chapters Three and Four, the optimal operator according to the MMSE has been formulated by using the projection theorem in Hilbert space. If the operator is linear, then the form of this operator depends on the solution of the Wiener-Hopf equation, which we have solved by Galerkin's method using the principal components. The treatment differs from the linear case when the operator is not linear. In this case the optimal operator is the regression operator. We have also shown how to formulate the conditional expectation using PCs when the noise model obeys the Gaussian distribution law.

Subsequently in Chapter Five, numerical tests have been given in order to investigate the performance and computability of both operators for different signal forms and noise models. The first case shows that the non-linear operator is in general the optimal one compared to the linear one. The second example illustrates that the non-linear operator becomes practically not really feasible when the dimension of the signal is too large, and instead we have to adopt the linear one in spite of its sub-optimality; the third one is a case where the linear and non-linear operators coincide with each other when the variables in both the signal and noise are Gaussian. As we have seen, the operators using the canonical expansion can attain accurate results in each experiment.

In all, the current work attempts to initiate a study in geomatics engineering on the method of canonical expansions and its application in the optimal estimation of time series, introduced by a group of scientists from the Russian school (e.g. Kolmogorov, Pugachev and Andreyev). Normally, the method of Fourier Transform (FT) is the typical treatment of the stationary time series by studying the PSD in the frequency domain. Therefore, the canonical expansion should be considered as a parallel approach to the method of FT when studying stationary processes. However, for non-stationary processes, the canonical expansion is more general, since as long as the PCs can be extracted from the covariance model, the optimal estimation can be performed in the same manner as we have developed in the thesis, without any assumption of the usual stationarity. As we have seen, in this thesis, we have successfully used this technique to solve the Wiener-Hopf equation and obtained the optimal linear estimation for a few time series. Also, the contribution includes a formulation of the non-linear regression, which is

the optimal operator overall. This opens a new direction of applications of such an approach in various problems not only in the MMSE estimation problems, but also in other optimal estimation problems under different criteria. Theoretically speaking, any operation or system obtained from this non-linear formulation should be the best estimator among all the possible operations and systems. This normally requires a combination of the method of canonical expansion with other numerical techniques.

6.2 Conclusions and Recommendations

Generally speaking, the canonical expansion is an effective approach of using the principal components derived from the noise model to filter out the unwanted noise in the observed time series, so that the useful signal can be extracted with the minimum square errors. It can be used in a variety of estimation problems in geomatics engineering and geodynamics: filtering, prediction and detection. Finally, we draw the following conclusions and recommendations, based on the results achieved in the present study:

1. Canonical expansion is an approach of linearisation of a random process into a combination of orthogonal elements, which contain all information about the second moment of the process.
2. In essence, the selection of principal components depends on the solution A of the matrix equation of similarity (or, equivalently, the corresponding linear integral equation in the continuous case):

   $$A \cdot K \cdot A^T = D$$

   In the thesis, we have investigated two solutions:

1) K-L expansion: the matrix A is the matrix of eigenvectors of the variance-covariance matrix K, correspondingly, D contains all its eigenvalues;
2) LDU decomposition: the matrix A is selected as a lower triangular matrix, so that every PC can be generated recursively.

Both selections entail a decomposition of the random process in a manner of canonical expansion.

3. Principal components of the noise model, generated by canonical expansion, can be adequately applied in constructing Galerkin's form of the solution of a Wiener-Hopf type optimal linear operator in MMSE estimation problems. Both the above mentioned choices of PCs give good performances of this operator.
4. The proposed linear operator can be utilised in filtering out both stationary and non-stationary noise.
5. The optimal linear operator is not always the best choice of MMSE estimation. In general, the one defined by the a-posteriori conditional expectation (or regression) offers the best solution to the estimation problem, and the linear one is only a special case of this non-linear one. Principal components can also be used in formulating this non-linear operator, and its performance is quite convincing.
6. However, the calculation of this non-linear regression operator sometimes involves great complexity. As a result, in practice, for the purpose of simplicity, it is at times expedient to adopt a sub-optimal linear operator at the cost of optimality.

7. A few models of problems that are frequently dealt with in geomatics engineering and geodynamics have been brought forward in order to examine the performances of the estimators in each case.
8. The proposed method of non-linear estimation using principal components is not restricted to the MMSE case only; on the contrary, it can be used in a much wider range of optimal estimation problems. MMSE is only a special case of a more general criterion of optimality with which a well-defined loss function is associated. In such manner, canonical expansion should be considered as one of the effective approaches in every branch of estimation theory.

References

Andreyev, N.I. (1969), 'Correlation Theory of Statistically Optimal Systems', W.B.Saunders Company

Blais, J.A.R. (1979), 'Least-Squares Block Adjustment of Stereoscopic Models and Error Analysis', reprint edition of *the author's PhD thesis in University of New Brunswick*, The University of Calgary Publication 30001

Blais, J.A.R. (1988), 'Estimation and Spectral Analysis', University of Calgary Press

Blais, J.A.R. (2010), 'Least Squares for Practitioners' (Review Article), *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, Vol.2010, Article ID 508092, 19 Pages

Booton, R.C. (1952), 'An Optimization Theory for Time-Varying Linear Systems with Nonstationary Statistical Inputs', *Proceedings of the IRE*, Vol. 40, Issue 8, pp.977-981

Collatz, L. (1964), 'Funktionalanalysis und Numerische Mathematik', *Die Grundlehren der Mathematischen Wissenschaften*, Band 120, Springer-Verlag, Berlin/Göttingen/Heidelberg

Davis, R.C. (1952), 'On the Theory of Prediction of Nonstationary Stochastic Processes', *Journal of Applied Physics*, Vol. 23, Issue 9, pp.1047-1053

Dougherty, E.R. (1999), 'Random Processes for Image and Signal Processing.' *SPIE/IEEE Series Imaging Science & Engineering*, SPIE/IEEE Press

Dougherty, E.R. (2009), 'Epistemology and the Role of Mathematics in Translational Science', *Tampere International Center for Signal Processing, TICSP Series* No.47, pp. 60-68, 2009

Gantmacher, F.R. (1959), 'The Theory of Matrices', Vol.1, Chelsea Publishing Company, New York, N.Y.

Gelb, A. et al. (1974), 'Applied Optimal Estimation', The MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Gibert, D. and J-L. Le Mouel (2008), 'Inversion of Polar Motion Data: Chandler Wobble, Phase Jumps and Geomagnetic Jerks', *Journal of Geophysical Research*, Vol.113, B10405, 2008

Jolliffe, I.T. (2002), 'Principal Component Analysis', 2nd edition, Springer Series in Statistics

Kantorovich, L.V. & Akilov, G.P. (1964), 'Functional Analysis in Normed Spaces', *Pergamon Press Ltd. International Series of Monographs in Pure and Applied Mathematics*

Kay, S.M. (1993), 'Fundamentals of Statistical Signal Processing. Vol. I Estimation Theory', Prentice Hall PTR, Upper Saddle River, New Jersey

Keller, W. (2000), 'A Wavelet Approach to Non-Stationary Collocation', in *Schwarz, K.P. (Editor), Geodesy Beyond 2000. The Challenge of the First Decade. International Association of Geodesy Symposia*, Vol. 121, Springer-Verlag, Berlin-Heidelberg 2000 , pp.208-214

Keller, W. (2004), 'Wavelets in Geodesy and Geodynamics', Walter de Gruyter, Berlin

Khintchine (1934), 'Korrelationstheorie der stationären stochastischen Prozesse', *Mathematische Annalen*, Band 109, pp.604-615

Kowalewski, G. (1909), 'Einfürung in die Determinatentheorie, einschließlich der unendlichen und Fredholmschen Determinaten', Leipzig, Verlag und Veit &

Comp.

Kudritskii, V.D. (2000), 'Optimal Linear Filtering and Extrapolation of Realizations of a Vector Random Function Observed with Errors', *Cybernetics and Systems Analysis*, Vol.36, No.4, 2000, pp.579-584

Lusternik, L.A. & Sobolev, V.J. (1974), 'Elements of Functional Analysis', Hindustan Publishing Corpn. (I) Delhi & John Wiley & Sons, Inc. New York

Middleton, D. (1963), 'Optimization Problems in Statistical Communication Theory', Chapter 7 in *Bellman, R.E. (Editor) 'Mathematical Optimization Techniques'*, University of California Press, Berkeley and Los Angeles, California

Nassar, S. (2003), 'Improving the Inertial Navigation System (INS) Error Model for INS and INS/DGPS Applications', *PhD Dissertation*, Department of Geomatics Engineering, University of Calgary, UCGE Reports, No. 20183

North, G.R. (1984), 'Empirical Orthogonal Functions and Normal Modes', *Journal of Atmospheric Science*, Vol. 41, Issue 5, pp.879-887

Parzen, E. (1963), 'A New Approach to the Synthesis of Optimal Smoothing and Prediction Systems', Chapter 5 in *Bellman, R.E. (Editor) 'Mathematical Optimization Techniques'*, University of California Press, Berkeley and Los Angeles, California

Pearson, K. (1901), 'On Lines and Planes of Closest Fit to Systems of Points in Space', *Philosophical Magazine* 2: 559-572

Pugachev, V.S. (1965), 'Theory of Random Functions, and its application to control problems', *Pergamon Press Ltd. International Series of Monographs on Automation and Automatic Control*, Vol.5

Sanso, F. & Sideris, M. G. (1997), 'On the Similarities and Differences between Systems Theory and Least-Squares Collocation in Physical Geodesy', *Bollettino di Geodesia e Scienze Affini*, Anno LVI, N.2, 1997, pp.175-206

Schmidt, E. (1907), 'Zur Theorie der Linearen und Nichtlinearen Integralgleichungen. I Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener.' *Mathematische Annalen*, Band 63, pp.433-476

Shaikin, M.E. (2007), 'Orthogonal Expansions of the Nonlinear Functionals of the Gaussian Process and Their Application to Statistical Analysis', *Automation and Remote control*, 2007, Vol.68, No.3, pp.446-455

Stewart, G.W. (1992), 'On the Early History of the Singular Value Decomposition', *SIAM Review* 35 (4): 551-566

Watkins, D.S. (2002), 'Fundamentals of Matrix Computations', 2nd Edition, *Pure and Applied Mathematics, A Wiley-Interscience Series of Texts, Monographs and Tracts*, John Wiley & Sons, Inc. New York

Zadeh, L.A. & Ragazzini, J.R. (1950), 'An Extension of Wiener's Theory of Prediction', *Journal of Applied Physics*, Vol. 21, Issue 7, pp.645-655