

國立交通大學

生醫工程研究所

碩士論文

貝氏架構下部份最小平方法

Bayesian-based Partial Least Squares Method



研究生：張書豪

指導教授：蕭子健

中華民國九十八年十月

貝氏架構下部份最小平方法

Bayesian-based Partial Least Squares Method

研究生：張書豪

Student : Shu-Hao Chang

指導教授：蕭子健

Advisor : Tzu-Chien Hsiao

國立交通大學
生醫工程研究所
碩士論文



Submitted to Institute of Biomedical Engineering
College of Computer Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

Oct 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年十月

貝氏架構下部份最小平方法

研究生：張書豪

指導教授：蕭子健

國立交通大學

生醫工程研究所



本論文的目的是在於建構一種分析法則，是一種以機率為基礎的多變數分析方法。此新的學習法則稱之貝氏架構下部份最小平方法，綜合了廣泛應用在生物訊號量測與分析的多變數方法中的部份最小平方法以及正則化的優點，並且導入貝氏分析的觀點，即使資料在有雜訊的情況下，可避免過度配適的現象，得到較好的估算結果。

在模擬數據分析部份，貝氏架構下部份最小平方法用來分析二種不同的波形，另外，也提出了一假設，我們考慮資料分佈為高斯分佈與一般分佈是否會造成整體分析效能的不同，利用正切函數來針對資料進行轉換，並以均方根誤差及相關係數來做為判定的標準說明貝氏架構下部份最小平方法可得到較好的結果。得到一具有雜訊消除的分析方法，並於未來將之應用於生醫訊號量測分析上。

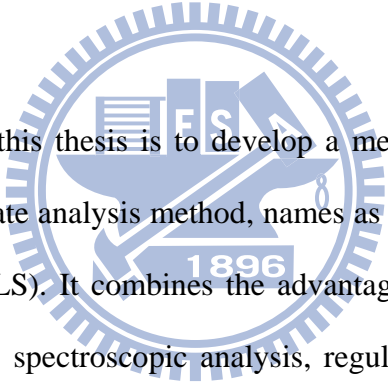
Bayesian-based Partial Least Squares Method

Student : Shu-Hao Chang

Advisor : Tzu-Chien Hsiao

Institute of Biomedical Engineering
National Chiao Tung University

Abstract



The main purpose of this thesis is to develop a method of analyzing. It is the probability-based multivariate analysis method, names as Bayesian-based partial least squares (Bayesian-based PLS). It combines the advantages of PLS which is widely used method in biomedical spectroscopic analysis, regularization technique and the Bayesian analysis to provide an efficient procedure to avoid the circumstance of overfitting and attain better results when calibrating under noisy data.

In the simulated experiments, Bayesian-based PLS is applied to analyze two different kinds of simulated waves. Besides, we also make an assumption to consider data with Gaussian distribution and uniform distribution. We examine these two cases to know which is better for analyzed results. The tangent function is used for transfer function. According to estimated standard of root mean square error and correlation coefficient, proving that Bayesian-based PLS has better analyzed performance. In the future, we will apply the proposed method which is able to reduce noise signal to Bio-signal measurement and analysis.

Acknowledgement

First of all, I would like to express my sincere appreciation to my advisor, Dr. TC Hsiao, for his helpful guidance, suggestions and supports. I had gained more experience and better logistic thinking for researching throughout my Master degree. Under his guidance, he shows how those multivariate analysis methods work and gives me a way to learn to treat and analyze the problem. For his advice, the thesis had been completed successfully. I would also like to thank Professor Wei-Ching Wang, Chih-Yu Wang, Kar-Kin Zao and Jiun-Hung Lin for reviewing my thesis and providing many useful comments on my research.

In the past of two years, I would express my gratitude to all the members in the laboratory of 704, including VBM and MIP. Thanks for their supports, suggestions and encouragements. Because of their companies, I could do my research patiently and strongly. The research assistant of VBM, HL Yu, she helps me not only my research but also everything in my daily life. Many thanks for your help. My classmates and roommates, CH Hsu, HY Hsu, YC Huang, CW Kao, KY Chen, TC Li and YI Hsieh, thanks for your assistance and encouragement every single day of my life in Hsinchu. All of you accompany with me to finish my Master degree. I will always remember all of you and the life here.

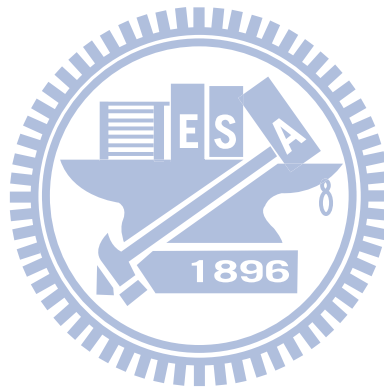
Finally, thanks my family for their understanding, supports, encouragements and the greatest love. Thanks all of my friends to give me a lot of cares and inspirations. I think I am so lucky to be you friend. Just a call or a message would let me more powerful to keep on fighting.

Life is sometimes tough ; however, there is nothing to defeat us with loves of you. Thank you for your loves.

Contents

Chinese abstract.....	i
Abstract.....	ii
Acknowledgement.....	iii
Contents.....	iv
List of Figures.....	vi
List of Tables.....	ix
Chapter 1. Introduction.....	1
1.1. Motivation.....	1
1.2. Literature study.....	2
1.3. Related work.....	6
1.4. Thesis Organization.....	8
Chapter 2. Materials and Methods.....	9
2.1. Least Squares (LS).....	9
2.2. Partial Least Squares (PLS).....	11
2.3. Bayesian analysis.....	15
2.4. Bayesian Regularization.....	21
Chapter 3. Bayesian-based PLS.....	22
3.1. Data preprocessing.....	22
3.2. Bayesian-based PLS algorithm.....	25
Chapter 4. Experiments and Results.....	28
4.1. Illustration.....	28
4.1.1. Synthesized simulation data.....	28
4.1.2. Criterion of estimation.....	29
4.1.3. Conditional training.....	31

4.2. Simulation data	33
4.2.1. Sigmoid function.....	33
4.2.2. Gaussian-base spectrum.....	38
4.2.3. Preprocessing.....	43
Chapter 5. Discussion.....	46
Chapter 6. Conclusions and Future works.....	48
6.1. Conclusions.....	48
6.2. Future works.....	48
References.....	49



List of Figures

Figure 1.1 The concept of inductive inference.....	3
Figure 1.2 Research tracing diagram.....	5
Figure 1.3 Trade-off curve.....	7
Figure 2.1 Two-layer architecture of LS method.....	10
Figure 2.2 The computational procedure of PLS.....	12
Figure 2.3 Three-layer architecture of PLS method.....	13
Figure 2.4 PLS learning flow chart.....	14
Figure 2.5 Data modeling process.....	16
Figure 2.6 Why Bayes embodies Occam's razor.....	17
Figure 2.7 The Occam factor.....	19
Figure 3.1 Tangent sigmoid function.....	23
Figure 3.2 The flow chart of data preprocessing.....	24
Figure 4.1 A sketch map of correlation coefficient.....	29
Figure 4.2 Root mean square error.....	30
Figure 4.3 Self-calibration and self-prediction (SCSP).....	32
Figure 4.4 Cross validation (CV).....	32
Figure 4.5 Noisy data (points) and sigmoid function (curve).....	33
Figure 4.6 RMSE as a function of N/S ratio under SCSP (PRLS).....	34
Figure 4.7 RMSE as a function of N/S ratio under SCSP (Bayesian-based PLS)...	34
Figure 4.8 RMSE as a function of N/S ratio under SCSP (PLS).....	35
Figure 4.9 Correlation coefficient as a function of N/S ratio under SCSP (PRLS).	35
Figure 4.10 Correlation coefficient as a function of N/S ratio under SCSP (Bayesian-based PLS).....	36
Figure 4.11 Correlation coefficient as a function of N/S ratio under SCSP (PLS)..	36

Figure 4.12 Prediction error sum of squares (PRESS) under CV.....	37
Figure 4.13 The linear combination of two Gaussian functions with different mean and standard deviation.....	38
Figure 4.14 The training data set of Gaussian-base spectrum.....	39
Figure 4.15 RMSE as a function of N/S ratio under SCSP (PRLS).....	39
Figure 4.16 RMSE as a function of N/S ratio under SCSP (Bayesian-based PLS).	40
Figure 4.17 RMSE as a function of N/S ratio under SCSP (PLS).....	40
Figure 4.18 Correlation coefficient as a function of N/S ratio under SCSP (PRLS).....	41
Figure 4.19 Correlation coefficient as a function of N/S ratio under SCSP (Bayesian-based PLS).....	41
Figure 4.20 Correlation coefficient as a function of N/S ratio under SCSP (PLS)..	42
Figure 4.21 Prediction error sum of squares (PRESS) under CV.....	42
Figure 4.22 The original training data set X.....	44
Figure 4.23 The new training data set X'.....	44
Figure 4.24 RMSE as a function of FWHM under SCSP.....	45
Figure 4.25 Correlation coefficient as a function of FWHM under SCSP.....	45
Figure 5.1 Where is the best solution.....	46
Figure 5.2 Trade-off curves of Bayesian-based PLS and PLS.....	47

Chapter 1

Introduction

1.1 Motivation

In 1998, Hsiao et al. proposed a similar conceptual architecture of Partial Least Squares (PLS) and Backpropagation Networks (BPN) [12]. This is a first time to compare the training procedure and investigate the physical meaning of BPN from PLS. Although PLS can be treated as special solution of BPN and be also used as initial weights for BPN in 2003 [1], the adaptive and momentum properties of BPN are still unclear from PLS. The over-fitting problem is not solved at BPN training. Regularization technique is one kind of methods to deal with over-fitting and under-fitting problem. In 2008, Chang et al applied the regularization technique to construct the PLS and proposed a novel method, Partial Regularized Least Square (PRLS), to noise reduction application [3]. To go a step further, I would like to discuss the different regularized methods with different input data distribution. If it's possible, I would also like to make a fundamental proof in mathematics between the two different scales of total squared error and vary of weighting coefficients respectively.

In this thesis, the regularization concept by multiply regularized parameter λ will be adopted to put these two criterion together for finding the appropriate relation. The Bayes' rule is also applied to evaluate the evidence for finding the best choice of regularized parameter λ [4]. In order to compare proposed PRLS in 2008, a probability-based analysis method by combining PLS is named as Bayesian-based PLS.

1.2 Literature study

Multivariate analysis methods are successfully applied to signal processing, widely used in many fields including spectrum analysis [5], bio-signal process [6], image processing [7], and etc. In general, the learning procedure of multivariate analysis methods can be separated into three types, i.e. deterministic, iterated, and hybrid algorithms. The method with deterministic algorithms is also called regressor model. The widely used regressors are Least Squares (LS), Principal Component Analysis (PCA) [8] and Partial Least Squares (PLS) [9]. The method with iterated algorithms is usually adopted in Artificial Neural Network (ANN). Multi-Layer Perceptron (MLP) which is the most practical model in ANN is typically used in supervised learning problems [10]. The results of multivariate analysis methods obtain from regressor model is deterministic, but the results obtain from ANN model is iterative to get the optimal solution.

However it has a main drawback which PCA lacks for information about which principal components are important for desired output and how many components are needed to compress the input data. Oja [8], [11] proposed PCA to reduce the dimension of input data by K-L transformation. PLS is a calibrated regression in common use and it can compress the input data and solve the main drawback of PCA. But PLS estimation suffers from overfitting is more serious than PCA [8]. Hsiao proposed a novel concept to combine the advantages of deterministic and iterated algorithms, i.e. PLS and BPN respectively [12], It's the first time to treat PLS as a three-layer ANN structure, prove the PLS as a special solution of general delta rule (GDR), and investigate the weights meaning Hsiao also adopted the PLS results as weights initialized method of BPN to get the near global minimum [1]. This novel hybrid algorithm of multivariate analysis method is PLS-BPN. The results of

PLS-BPN show that it's fast converge into a near global minimum than PLS and BPN. The research tracing diagram will be illustrated in Figure 1.2.

Figure 1.1 illustrates a brief concept of data modeling process. The process is started by gathering data and creating models to specify the data that we operate. It includes two levels of inference. The first level is model fitting; we fit each model to the data. In this level, the task is to infer what the free parameters of each model might be given the data. The second level is model comparison; we assign preferences to the alternative models. After these two levels of inference, we can have some useful information to make decision.

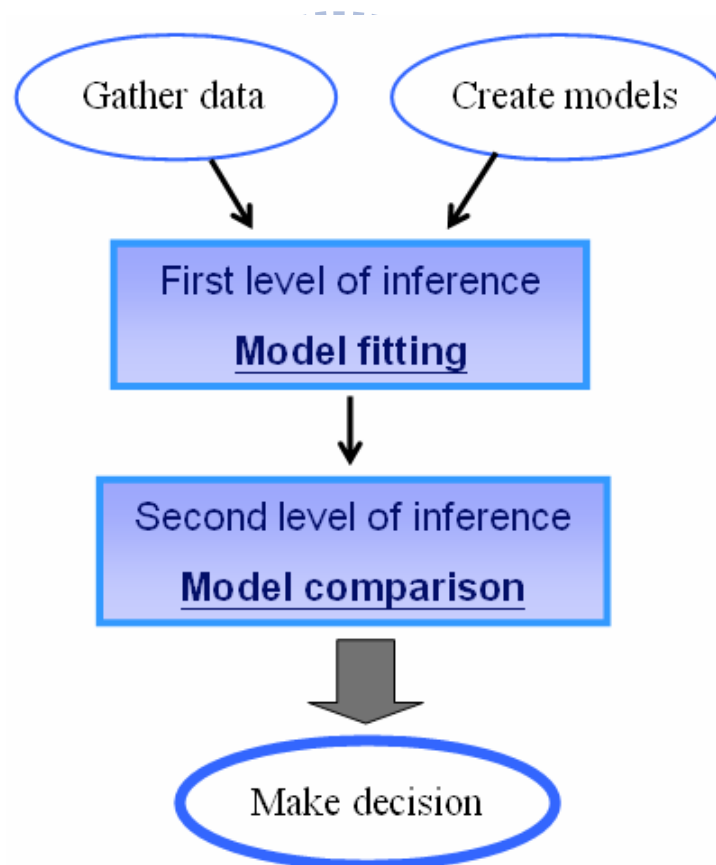


Figure 1.1 The concept of inductive inference

Chen [13] proposed Orthogonal Least Squares (OLS) based on radial basis function network (RBFN) also suffered from the same situation. By adopting the regularization technique, Chen [2] also constructed Regularized Orthogonal Least Squares (ROLS) to solve the problem of overfitting. Chang [3] regard PLS as three layer network in order to add regularization term into the structure. Following the architecture of ROLS method, we modify the PLS by combining the advantages of regularization to establish a novel calibrated model, names Partial Regularized Least Squares (PRLS). And we apply PRLS to analyze the data for noise reduction application. We improve the accuracy better than PLS under influence of noisy training data.

In recently research, we further consider the concept of Bayes' rule in our study [4]. In probability, Bayes' rule shows how one conditional probability, such as the probability of a hypothesis given observed evidence, depends on its inverse; here it means the probability of that evidence given the hypothesis. It is common to think of Bayes' rule in terms of updating our belief about a hypothesis A in the light of new evidence B. Specially, the posterior probability $P(A | B)$ is calculated by multiplying the prior probability $P(A)$ by the likelihood probability $P(B | A)$ that event B will occur if event A is true. The formula of Bayes' rule is shown as below :

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad (1-1)$$

Here $P(A | B)$ denotes the posterior probability, $P(B | A)$ is likelihood probability, $P(A)$ is prior probability and $P(B)$ is the evidence probability.

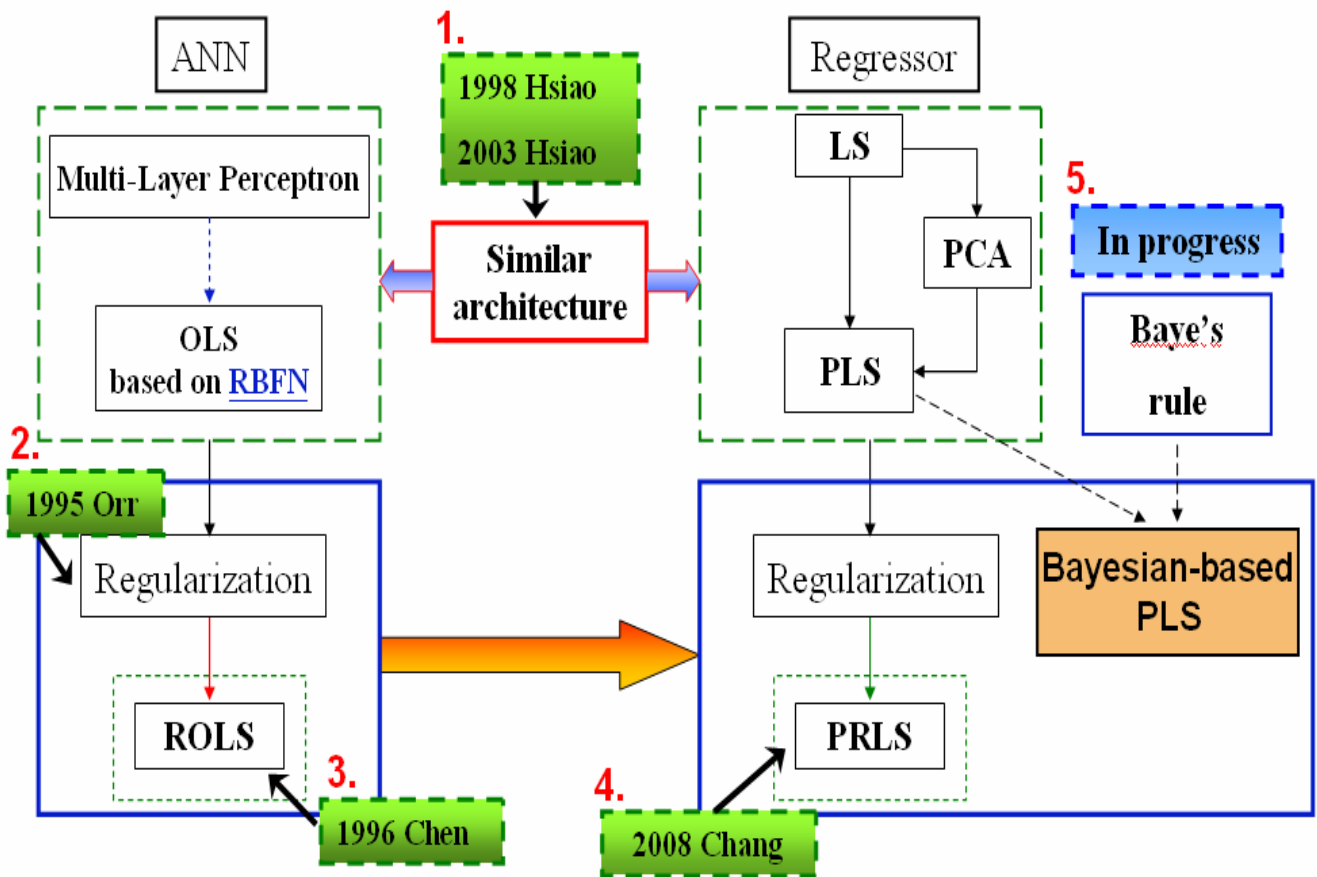


Figure 1.2 Research tracing diagram

1.3 Related work

1.3.1 Regularization

In many fields of mathematics, regularization technique has been used to solve ill-posed problem or avoid over-fitting problem [2] [15]. In regularization technique, the error function is minimized which depends on the network weights as well as the fit error [15]. In the recent study (Orr 1993), it has applied zero-order regularization technique to construct RBF networks. The zero-order regularization is equivalent to simple weight-decaying in gradient descent method for MLP neural network [16]. A theoretical reason for regularization is that it makes an effort to impose Occam's razor on the solution. From a Bayesian point of view, many regularization techniques correspond to imposing certain prior distributions on model parameters.

However, zero-order regularization, though dominated by better methods, demonstrates most of the basic ideas that are used in inverse problem theory. In general, let us define $A[\mathbf{u}] > 0$ and $B[\mathbf{u}] > 0$ be two positive functionals of \mathbf{u} , so we can try to determine \mathbf{u} by either :

Minimize : $A[\mathbf{u}]$ or $B[\mathbf{u}]$

The first, A , measures something like the agreement of a model to the data (e.g., χ^2), denote the agreement between data and solution, or "sharpness" of mapping between true and estimated solution. And B measures something like "smoothness" of the desired solution, means the smoothness or stability of the solution.

In summary, regularization is Lagrange multiplier equation combines with a quadratic constraint to minimize the weighted sum $A[\mathbf{u}] + \lambda B[\mathbf{u}]$ and lead to a adequate solution for \mathbf{u} . Here, λ is the regularized parameter. The constant λ adjudicates a delicate compromise between the two subjects.

Figure 1.3 illustrates the trade-off curve between agreement A and smoothness B . Almost all inverse problem methods involve a trade-off between two optimizations. So, we want to select an appropriate parameter to control the trade-off curve and find the best solutions from all achievable solutions have shown as below.

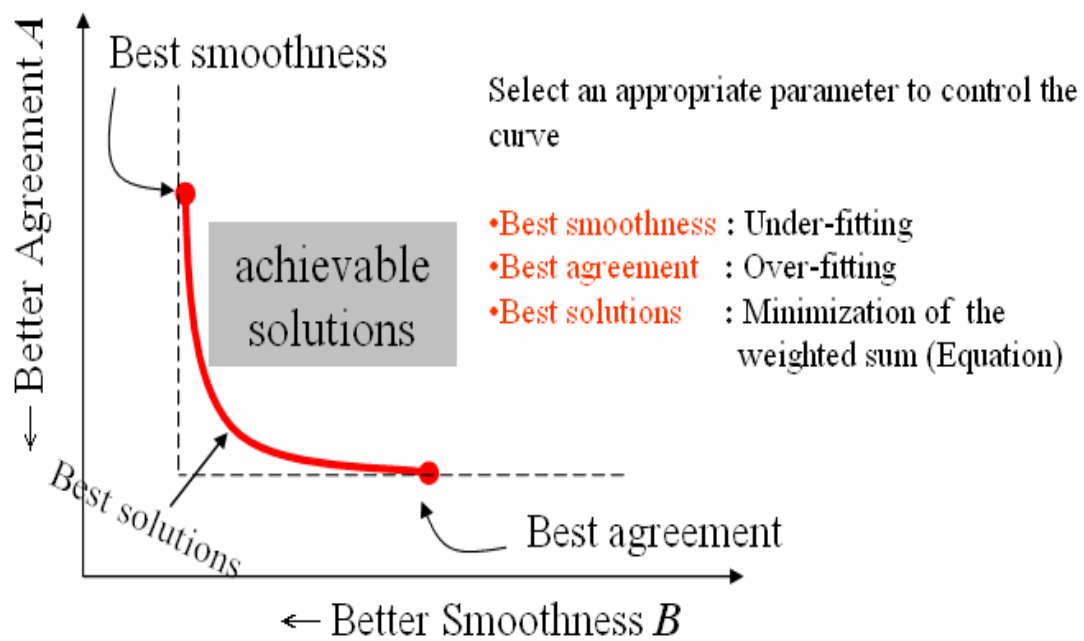


Figure 1.3 Trade-off curve [14]

1.4 Thesis Organization

The structure of the thesis is described as follow. The first chapter gives an introduction and the motivation for my research. Next section, in chapter 2 we depict some calibration models, Bayesian analysis, and Bayesian regularization in my study. In chapter 3, we will make some discussion between Bayesian regularization and PLS. Later, we propose a novel calibration model, names as Bayesian-based PLS, by combining PLS with the concept of Bayes' rule and regularization technique. Chapter 4 shows the simulation experiment results. Then, we will make some discussions in chapter 5 Conclusions and future works are listed in the final chapter.



Chapter 2

Materials and Methods

2.1 Least Squares (LS)

The least squares (LS) method is used to approximate the parameters and find the best fitting curve to fit the given data. Classic LS regression has minimum sum of squared residuals between data set and estimation. Suppose the linear model is given by $f(x_i) = a_0 + x_{i1}a_1 + x_{i2}a_2 + \dots + x_{im}a_m, i = 1, 2, \dots, n$. The LS method use this model to approximate the given set of data. And the sum of squared error (SSE) is calculated as below :

$$SSE = \sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n (y_i - (a_0 + x_{i1}a_1 + x_{i2}a_2 + \dots + x_{im}a_m))^2 \quad (2-1)$$

and we get the partial differential equations for each a_j , the derivation is :

$$\frac{SSE}{\partial a_j} = 2 \sum_{i=1}^n (y_i - (a_0 + x_{i1}a_1 + x_{i2}a_2 + \dots + x_{im}a_m))(-x_{ij}) = 0 \quad (2-2)$$

where $j = 1, 2, \dots, m$

We also can illustrate LS method to a two-layer ANN architecture shown as Figure 2.1. And we transform the data set to matrix form. Then matrix \mathbf{X} represents the input data $\mathbf{X} = [x_1 \ x_2 \ x_3 \ \dots \ x_m]$; $x_m = [x_{1m} \ x_{2m} \ x_{3m} \ \dots \ x_{nm}]$, real output $\mathbf{Y} = [y_1 \ y_2 \ y_3 \ \dots \ y_n]^T$ and weight coefficient $\mathbf{a} = [a_1 \ a_2 \ a_3 \ \dots \ a_m]^T$.

The LS procedure in matrix form is defined as :

$$\mathbf{Y} = \mathbf{Xa} + \boldsymbol{\varepsilon} \quad (2-3)$$

We calculate the weighting coefficients due to (2-3).

$$\mathbf{X}^T \mathbf{Y} \approx \mathbf{X}^T \mathbf{X} \mathbf{a} \quad (2-4)$$

$$\mathbf{a} \approx (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Y}) \quad (2-5)$$

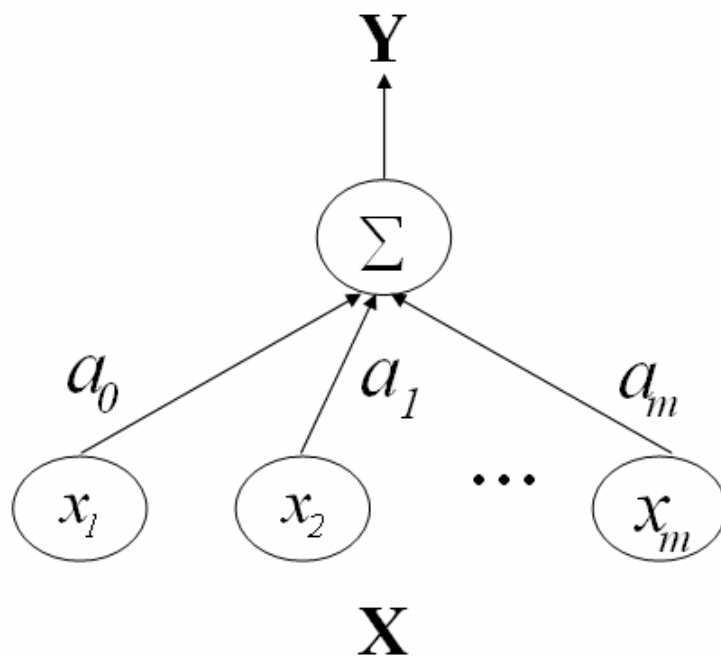


Figure 2.1 Two-layer architecture of LS method

2.2 Partial Least Squares (PLS)

PLS is a method which the most widely used in biomedical spectroscopic analysis. It is a popular technique that generalizes and combines features from principal component analysis (PCA) and multiple regressions. The purpose of PLS is to predict or analyze a set of dependent variables from a set of independent variables or predictors. PLS regression is mainly useful when we have to predict a set of dependent variables from a large set of independent variables. It is used to find the fundamental relations between two matrices (\mathbf{X} and \mathbf{Y}), i.e. a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the \mathbf{X} space that explains the maximum multidimensional variance direction in the \mathbf{Y} space.

We will illustrate the general underlying model of multivariate PLS as follow and show you the architecture of multivariate system if we treat PLS as a three-layer ANN network.

The independent variable matrix $\mathbf{X}_{n \times m}$ decomposed into matrix $\mathbf{T}_{n \times a}$ with corresponding weighting matrix $\mathbf{P}_{a \times m}$ and dependent variable matrix $\mathbf{Y}_{n \times 1}$ can be decomposed into matrix $\mathbf{T}_{n \times a}$ with corresponding weighting matrix $\mathbf{Q}_{a \times 1}$. The mathematic form is represented as follows :

$$\begin{aligned}\mathbf{X}_{n \times m} &= \mathbf{X}^{(1)} + \mathbf{X}^{(2)} + \dots + \mathbf{X}^{(a)} + \mathbf{E} \\ &= \mathbf{t}_1 \mathbf{p}_1 + \mathbf{t}_2 \mathbf{p}_2 + \dots + \mathbf{t}_a \mathbf{p}_a + \mathbf{E} \\ &= \mathbf{T}_{n \times a} \mathbf{P}_{a \times m} + \mathbf{E}\end{aligned}\tag{2-6}$$

$$\begin{aligned}
\mathbf{Y}_{n \times 1} &= \mathbf{Y}^{(1)} + \mathbf{Y}^{(2)} + \dots + \mathbf{Y}^{(a)} + \mathbf{F} \\
&= \mathbf{t}_1 \mathbf{q}_1 + \mathbf{t}_2 \mathbf{q}_2 + \dots + \mathbf{t}_a \mathbf{q}_a + \mathbf{F} \\
&= \mathbf{T}_{n \times a} \mathbf{Q}_{a \times 1} + \mathbf{F}
\end{aligned} \tag{2-7}$$

From the formula (2-6) and (2-7) above, we also can illustrate the mathematic relation for computing PLS in Figure 2.2. It shows the regression steps how PLS decomposed.

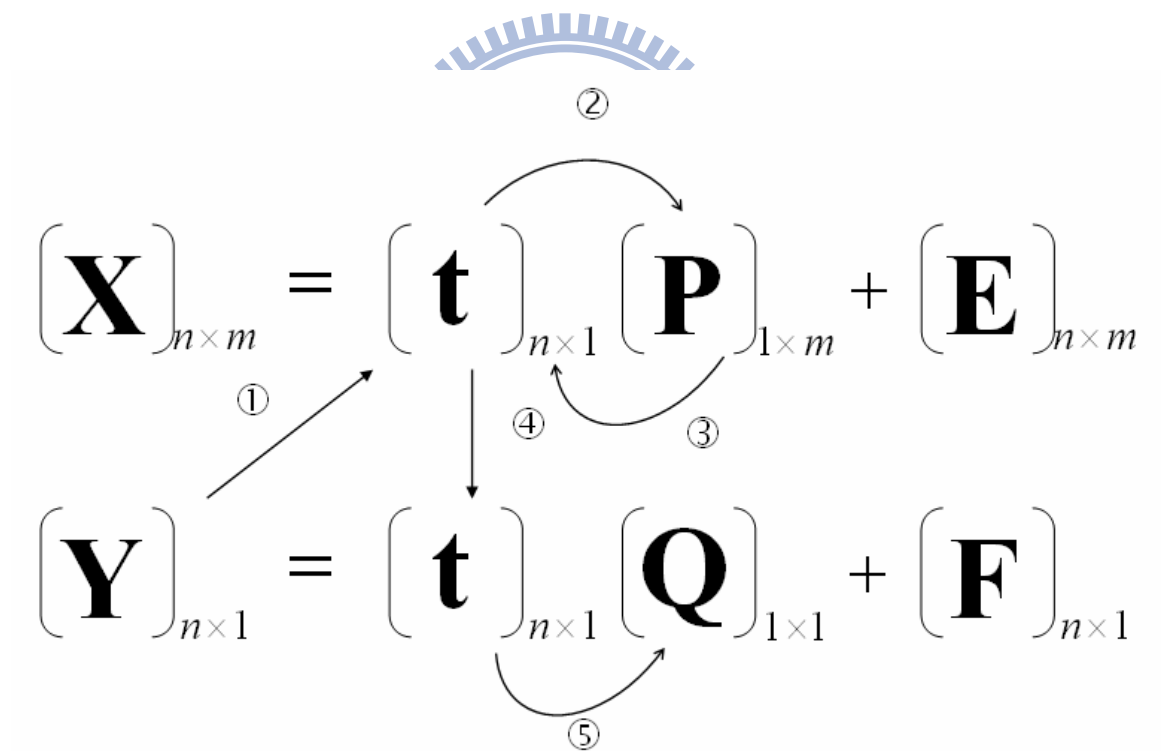


Figure 2.2 The computational procedure of PLS

After derivative, we exactly find out the residual matrix $\|\mathbf{E}_{n \times m}\|$ and $\|\mathbf{F}_{n \times 1}\|$ are minimized through the course of decomposing the matrix \mathbf{X} and \mathbf{Y} . When computational iteration equation to a ($a \leq n$) or the residual small than a minimum, PLS procedure would terminate.

Ham [17] and Hsiao [12] bring up an idea which regards PLS as one kind of artificial neural networks. In the purpose, transformation between independent and dependent variables can be represented as three-layer ANN architecture. It is shown as Figure 2.3. And the PLS learning procedure will be illustrated in Figure 2.4.

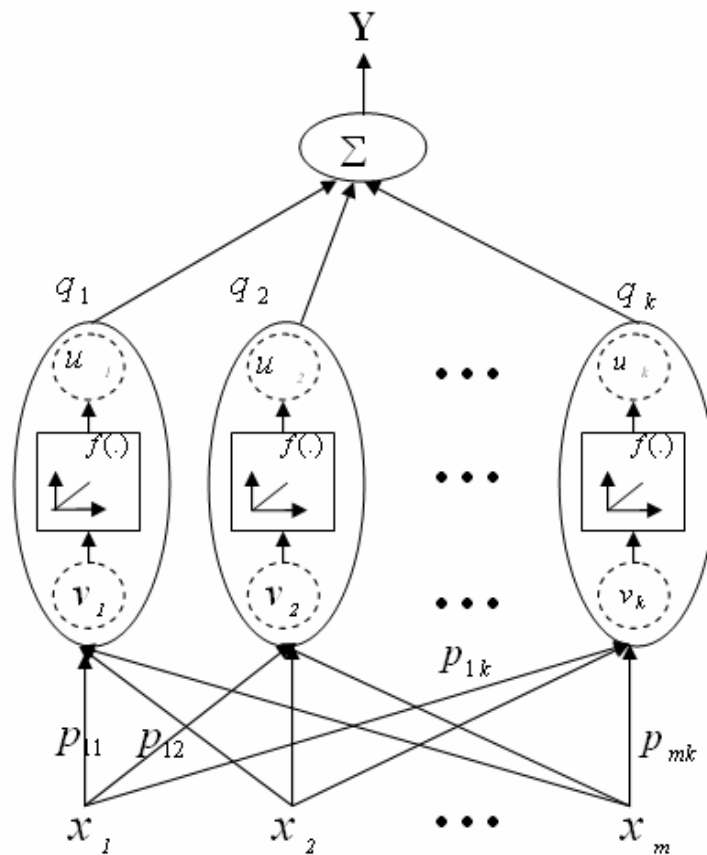


Figure 2.3 Three-layer architecture of PLS method

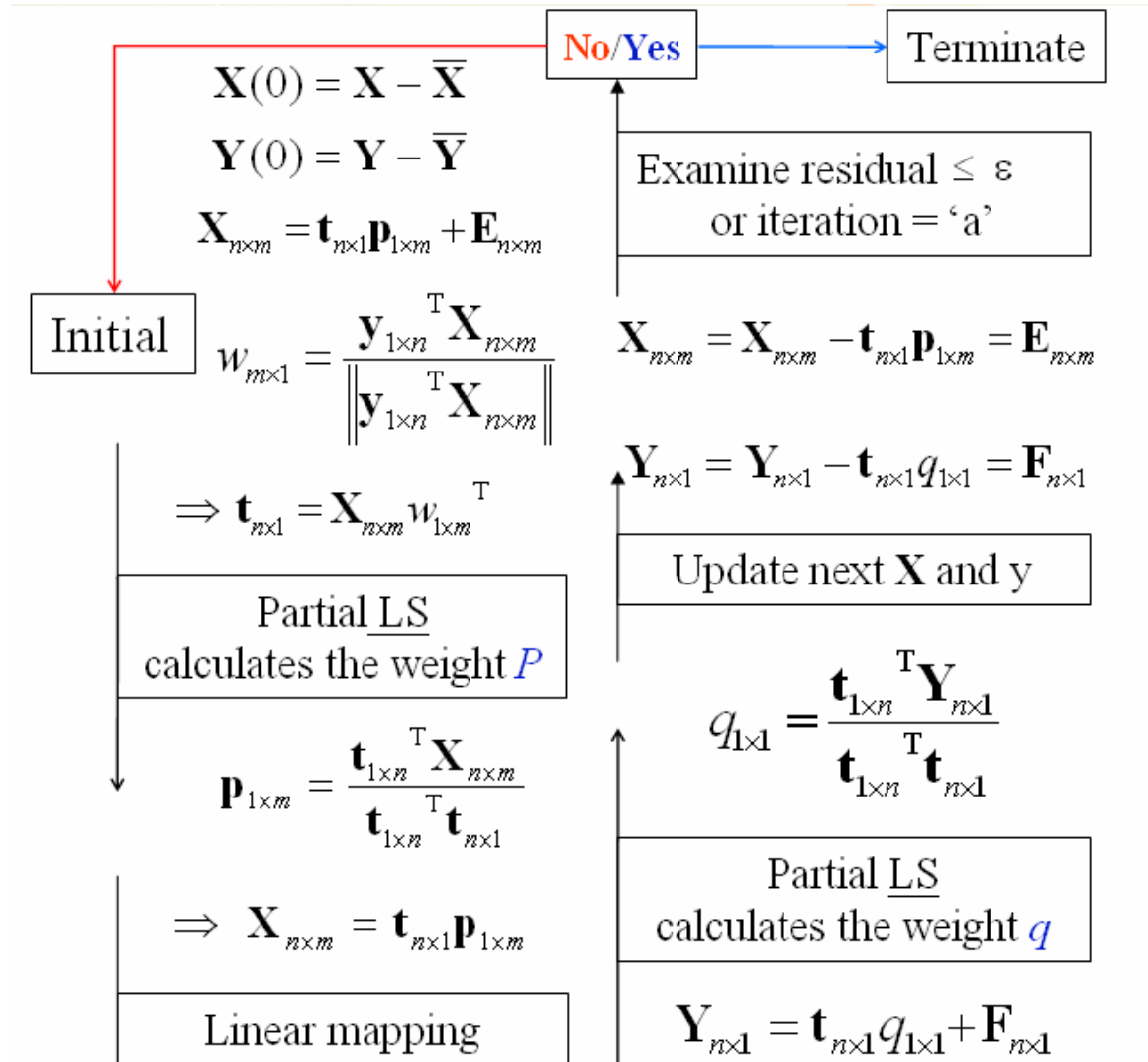


Figure 2.4 PLS learning flow chart

2.3 Bayesian analysis

Bayesian refers to methods in probability and statistics named after the Reverend Thomas Bayes. Bayesian methods for inductive inference were first developed in detail early this century by the Cambridge geophysicist, Sir Harold Jeffreys [18]. Bayesian inference is the statistical inference in which evidence or observations are used to update or to newly calculate the probability that a hypothesis might be true. Bayesian inference uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and calculates a numerical estimate of the degree of belief in the hypothesis after evidence has been observed. The fundamental concept of Bayesian analysis is that the plausibilities of alternative hypotheses are represented by probabilities, and inference is performed by evaluating those probabilities.

In David J.C. Mackay proposed paper [4], the Bayesian approach to regularization and model-comparison is clarified by studying the inference problem of interpolating noisy data. The concepts and methods described are quite general and can be applied to many other data modelling problems.

In his study, we can examine the posterior probability distribution to set the regularized constants. The way in which Bayes infers the values of regularized constants and noise levels has an elegant interpretation in terms of the effective number of parameters determined by the data set.

Two levels of inference are involved in the task of data modelling. Figure 2.5 will show you where Bayesian inference fits into the data modelling process and illustrate an abstraction of the part of the scientific process in which data is collected and modelled. At the first level of inference, we assume that one of the models we created is true, then we fit the model to the data. And the second level of inference is the model comparison. The two double-framed boxes denote the two steps which

involve inference. However, Bayes' rule can only be used in these two steps. Bayes' rule may be used to find the most probable parameter values and the error bars on these parameters. The second inference task requires a quantitative Occam's razor to penalise overcomplex models. Bayes can assign objective preferences to the alternative models in a way that automatically and quantitatively embodies Occam's razor [18][19]. Complex models are automatically self-penalizing under Bayes' rule.

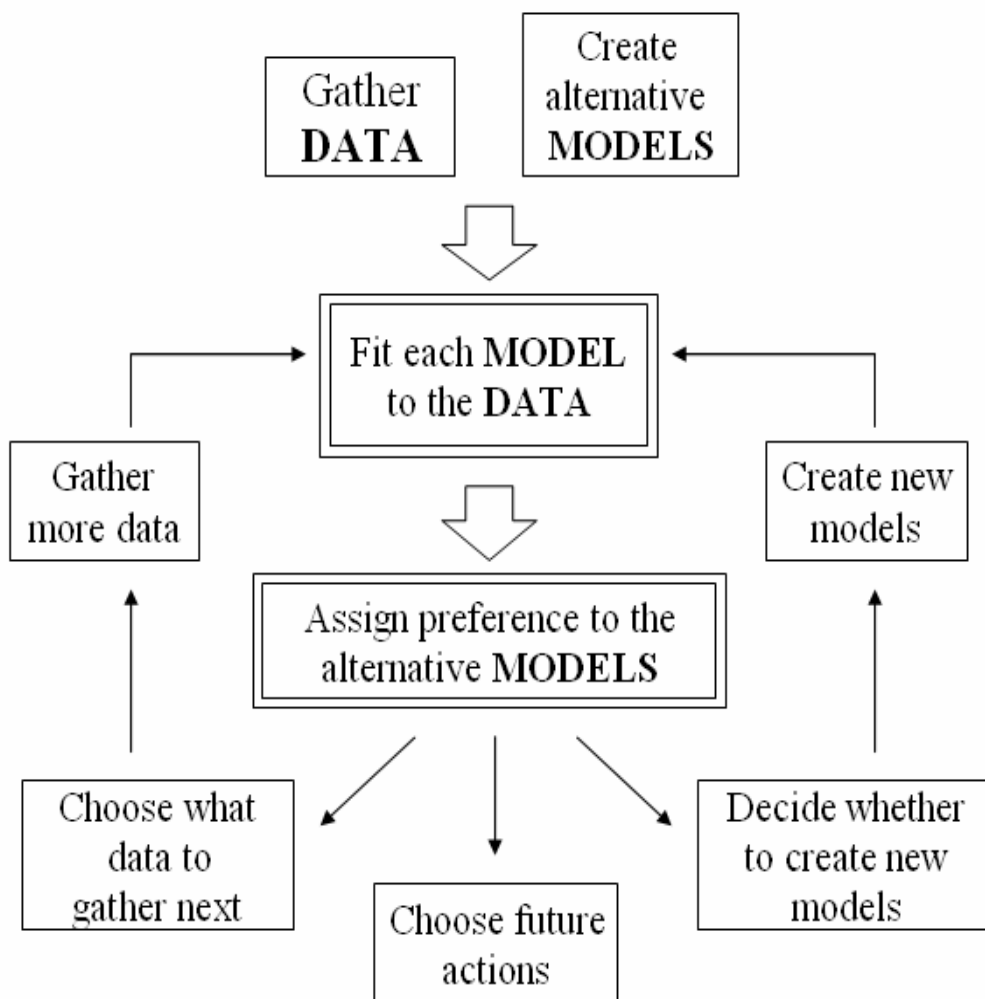


Figure 2.5 Data modeling process [4]

Model comparison is a difficult task because it's not possible simply to find the best model that fits the data set. Occam's razor is the principle that states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis. A problem should be stated in its basic and simplest form.

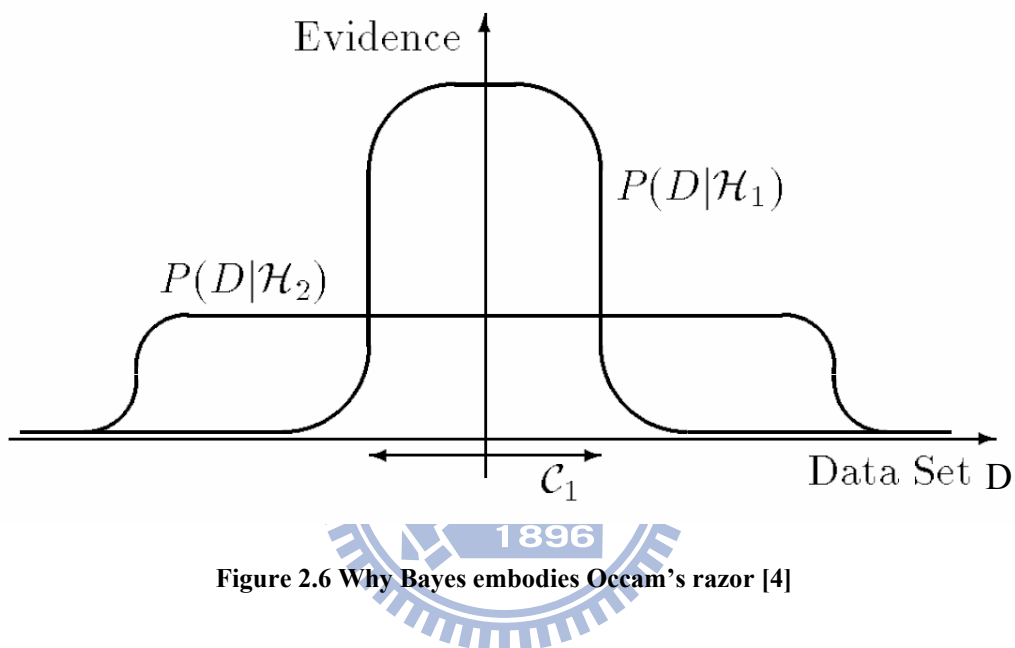


Figure 2.6 Why Bayes embodies Occam's razor [4]

The Figure 2.6 shows the intuition for why complex models penalized. Bayes' rule rewards models according to how well they predict actual data. These predictions are quantified by a normalized probability distribution on data sets D and this probability, $P(D|H_i)$, is known as the evidence for H_i . A simple model H_1 makes only a limited range of predictions, $P(D|H_1)$; a more powerful model H_2 that has more free parameters than H_1 , is able to predict a larger variety of data sets. However, this means that H_2 can not predict the data sets in region C_1 as strongly as H_1 . Assume that the two models have been assigned the equal prior probabilities. Then if the data set falls in region C_1 , the less powerful model H_1 will be the more probable than to the model H_2 .

Let us write down the Bayes' rule for the two levels of inference so that we can examine explicitly how Bayesian model comparison works.

Model fitting: At the first level of inference, we assume that one model H_i is true, we infer what the model's parameter w might be given the data D . Using Bayes' rule, the posterior probability of the parameter w is :

$$P(w | D, H_i) = \frac{P(D | w, H_i)P(w | H_i)}{P(D | H_i)} \quad (2-8)$$

And we also can rewrite this formula in words :

$$\text{Posterior} = \frac{\text{likelihood} \times \text{Prior}}{\text{Evidence}}$$

Model comparison : At the second level of inference, we infer which model is the most sensible give the data. And the posterior probability for each model is defined as :

$$P(H_i | D) \propto P(D | H_i)P(H_i) \quad (2-9)$$

Assuming that we have no reason to assign strongly differing priors $P(H_i)$ to the alternative models, models H_i are ranked by evaluating the evidence. New models are compared with previous by evaluating the evidence for them.

Let us explicitly study the evidence to gain insight into how the Bayesian Occam's razor works. The evidence is the normalizing constant for equation (2-8) :

$$P(D | H_i) = \int P(D | w, H_i)P(w | H_i)dw \quad (2-10)$$

Figure 2.7 shows the quantities that determine the Occam factor for hypothesis H_i having a single parameter w . The dotted line that represented the prior distribution for the parameter has width $\Delta^0 w$. The solid line that represented the posterior distribution has a single peak at w_{MP} with characteristic width Δw . The Occam factor is $\frac{\Delta w}{\Delta^0 w}$.

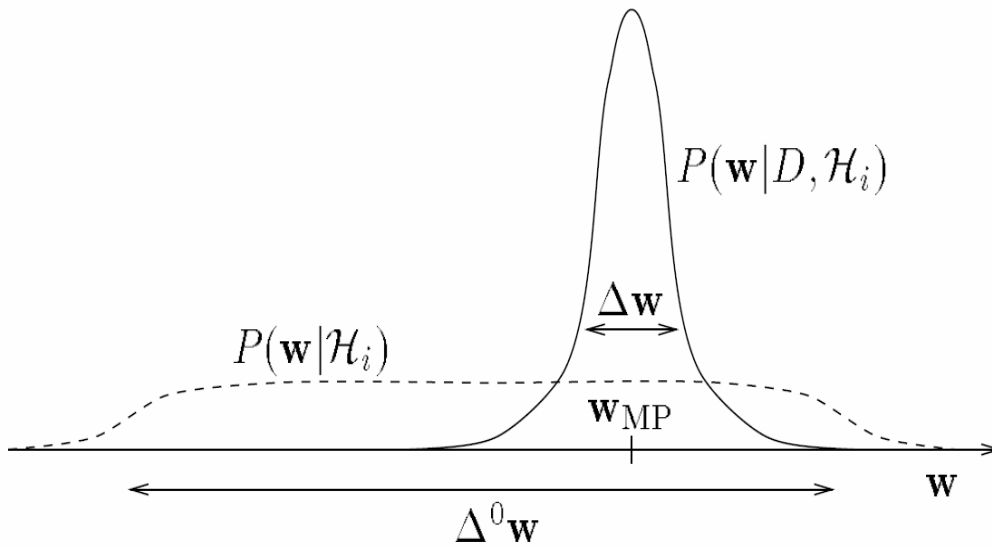


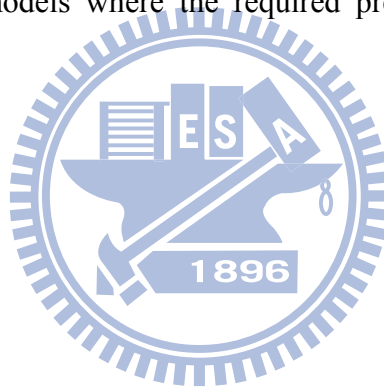
Figure 2.7 The Occam factor [4]

Therefore the evidence is evaluated by taking the best fit likelihood and multiplying it by the Occam factor.

$$P(D | H_i) \cong \underbrace{P(D | w_{MP}, H_i)}_{\text{Best fit likelihood}} \underbrace{P(w_{MP} | H_i) \Delta w}_{\text{Occam factor}} \quad (2-11)$$

Evidence \cong Best fit likelihood Occam factor

The quantity Δw is the posterior uncertainty in w . Imagine for simplicity that the prior $P(w|H_i)$ is uniform on some large interval $\Delta^0 w$, representing the range of values of w that H_i thought possible before the data arrived. Then $P(w_{\text{MP}}|H_i) = \frac{1}{\Delta^0 w}$. The log of the Occam factor can be interpreted as the amount of information we gain about the model when the data arrive. Comparison of evidence, $P(D|H_i)$, provides a purely objective way to rank hypotheses. Evaluation of evidence is an extension of maximum likelihood model selection : multiply the best fit likelihood by the Occam factor. No more computationally difficult than finding the best fit parameters. The Occam factor automatically penalizes a model which requires fine tuning of its parameters. It promotes models where the required precision of its parameters is coarse.



2.4 Bayesian regularization

In this section, we will introduce the Bayesian regularization and examine the probability distribution to set the regularized parameter λ . The selection of regularized parameter λ is the key concept of our method. So we adopt Bayesian analysis to infer the optimal value of λ . To infer from the data what value λ should have, we evaluate some probability distribution.

As mentioned earlier, we add the concept of regularized technique to PLS and rewrite the new error criterion as :

$$E_e = \mathbf{e}^T \mathbf{e} + \lambda \mathbf{q}^T \mathbf{q}, \quad \lambda \geq 0 \quad (2-12)$$

Where $\mathbf{e}^T \mathbf{e}$ means the total sum of squared error and $\mathbf{q}^T \mathbf{q}$ is the weighting vector which infers the output directly. However, original PLS calibration reduces the total error as far as possible but if there has noisy signal (outlier) in the training data, the prediction may fit to the noisy data. So the predicted accuracy will be poor for the unseen data. Vary of weighting coefficients $\mathbf{q}^T \mathbf{q}$ controls the covariance for two variables.

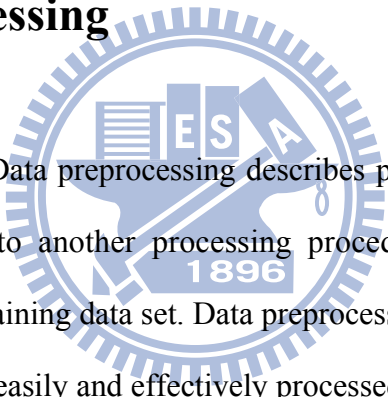
As shown in section 1.3, Figure 1-2 gives a briefly interpretation for error criterion and regularized parameter λ . The regularized parameter λ is added to the term to make the calibration curve smooth without oscillating. The Bayesian-based PLS keeps the balance between smoothness of curve and accuracy in calibration phase.

Chapter 3

Bayesian-based PLS

In this chapter, we establish a novel analyzed method, Bayesian-based PLS, by applying Bayesian approach to PLS method. An elegant approach to the selection of the regularization parameter is to adopt Bayesian interpretation and evaluate the evidence probability to find the best value of regularization parameter. The evidence procedure we adopt is to calculate the probability $P(D | \alpha, \beta, H)$.

3.1 Data preprocessing



In computer science, Data preprocessing describes processing performed on the raw data to transform it to another processing procedure. The result after data preprocessing is the final training data set. Data preprocessing transforms the data to a new type that will be more easily and effectively processed for the purpose of the user. There are many different widely used methods and techniques for data preprocessing, including sampling, cleaning, normalization, transformation, denoising, feature extraction and selection, etc. The sampling is the process of selecting a representative subset from a large population of data. The transformation is usually applied so that the data appear to more closely meet the assumptions of a statistical inference procedure that is to be applied. The denoising is the method that eliminate the noise from the source data. The normalization, which organizes data more efficient to access and more normal, which typically means conforming to some regularity or rule, or returning from some state of abnormality and feature extraction is a process of dimensionality reduction. It projects a data set with higher dimensionality onto a

smaller number of dimensions. As such it is useful for data visualization, since a complex data set can be effectively visualized when it is reduced to smaller dimensions.

In our study, we make the assumption for considering data preprocessing procedure into our method. Transformation is adopted to make the training data set to another form. Here the tangent sigmoid function is selected and shown in Figure 3.1.

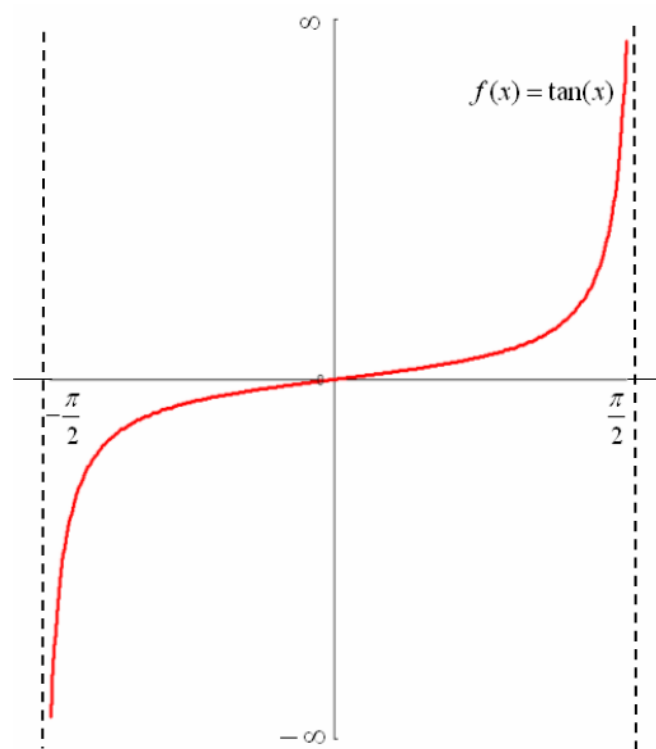


Figure 3.1 Tangent sigmoid function

For constructing our three-layer architecture of Bayesian-based PLS, the data preprocessing procedure had been added between input and hidden layer and the flow chart has shown in Figure 3.2. In Figure 3.2, the original data \mathbf{X} was generated with uniform distribution. After tangent function transferring, we obtain new form of data \mathbf{X}' . The probability density function (PDF) of \mathbf{X}' is Gaussian distribution or normal distribution.

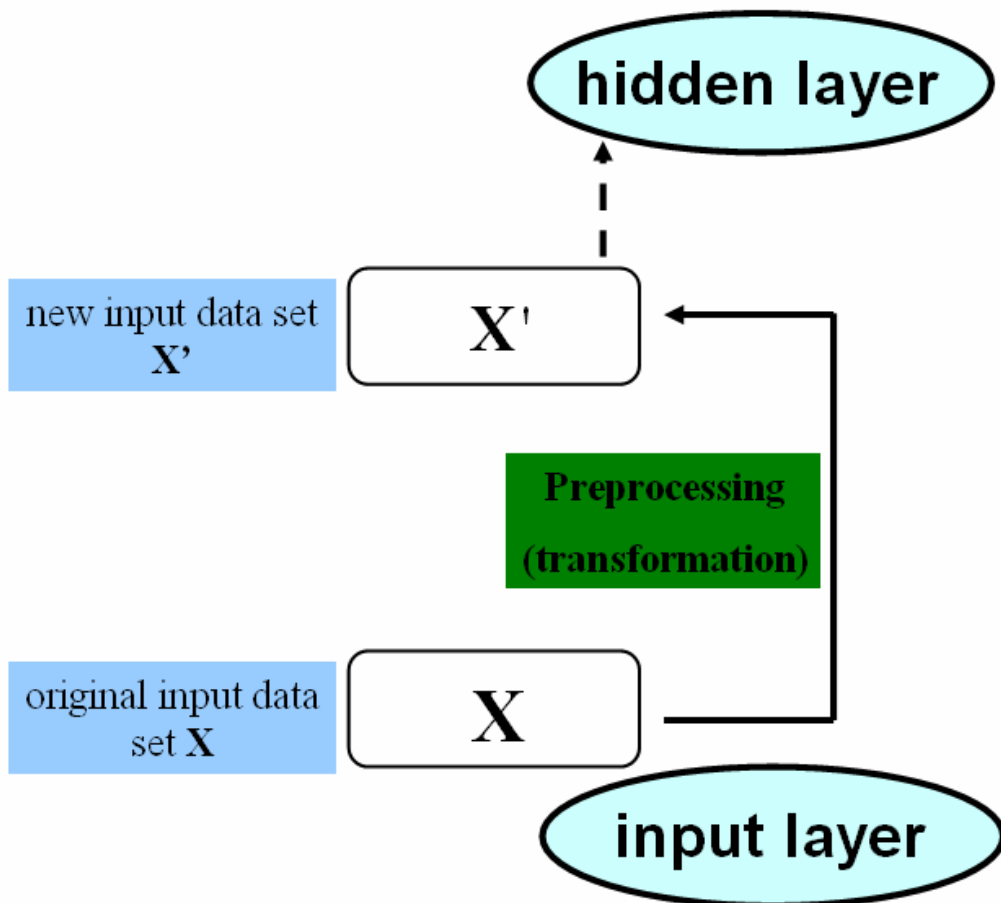


Figure 3.2 The flow chart of data preprocessing

3.2 Bayesian-based PLS algorithm

The assumption we made is to consider different kinds of probability density function and try to find out the relation between maximum covariance and minimum sum of squared error for them. In original PLS three-layer ANN architecture, it searches the maximum variance between input and hidden layer and finds the minimum sum of total squared error between hidden and output layer. Here we define the total data misfit function as :

$$M = \alpha E_D + \beta E_W \quad (3-1)$$

E_D is the residual squared error function and E_W is commonly referred to as a regularizing function. In order to find the optimal value for λ , the regularized parameter, we apply Bayesian interpretation and calculate the best value for λ by using the evidence procedure, an approximate Bayesian scheme reviewed in Mackay [4]. Using Bayes' rule, we get the posterior probability of the parameter w is the formula (2-8). Now that we want to evaluate the evidence find the value of λ . We define the probability of the data given the parameter w is :

$$P(D | w, H_i) = \frac{\exp(-\alpha E_D)}{Z_D(\alpha)} = \frac{\exp(-\frac{1}{2} \sum_{i=1}^n e_i^2)}{(2\pi\sigma_e^2)^{n/2}} \quad (3-2)$$

and a prior probability on the parameter w is :

$$P(w | H_i) = \frac{\exp(-\beta E_W)}{Z_W(\beta)} = \frac{\exp(-\frac{1}{2} \sum_{j=1}^k w_j^2)}{(2\pi\sigma_w^2)^{k/2}} \quad (3-3)$$

And if α and β are known, then the posterior probability of the parameter w is :

$$P(w | D, H_i) = \frac{\exp(-M(w))}{Z_M(\alpha, \beta)} = \frac{\exp(-\alpha E_D - \beta E_W)}{Z_D(\alpha) Z_W(\beta) P(D | H_i)} \quad (3-4)$$

From above formulas and given some hypotheses, we evaluate the evidence for α, β .

We have the formula (3-5) as :

$$P(D | H_i) = P(D | \alpha, \beta, H_i) = \frac{Z_M(\alpha, \beta)}{Z_W(\alpha)Z_D(\beta)} \quad (3-5)$$

Thus we can write the log evidence for α and β as :

$$\log P(D | \alpha, \beta, H_i) = -(\alpha E_D^{MP} + \beta E_W^{MP}) - \frac{1}{2} \log \det(\mathbf{A}) + \frac{k}{2} \log \alpha + \frac{N}{2} \log \beta - \frac{N}{2} \log 2\pi \quad (3-6)$$

Then we differentiate the log evidence, from (3-6), with respect to α and β so as to find the condition that is satisfied at the maximum. We can obtain derivation for differentiating with respect to α and β from formula (3-6).

$$\frac{d}{d\alpha} \log P(D | \alpha, \beta, H_i) = -E_D^{MP} - \frac{1}{2} \text{Trace}(\mathbf{A}^{-1} \mathbf{B}) + \frac{n}{2} \frac{1}{\alpha} = 0$$

First, we differentiate the log evidence with respect to α and get, setting the derivation to zero :

$$2\alpha E_D = n - \alpha \text{Trace}(\mathbf{A}^{-1} \mathbf{B}) = n - \gamma \quad (3-7)$$

And then, we differentiate with respect to β .

$$\frac{d}{d\beta} \log P(D | \alpha, \beta, H_i) = -E_W^{MP} - \frac{1}{2} \text{Trace}(\mathbf{A}^{-1}) + \frac{k}{2} \frac{1}{\beta} = 0$$

We obtain the following condition for the most probable value of β :

$$2\beta E_W^{MP} = k - \beta \text{Trace}(\mathbf{A}^{-1}) = \sum_{j=1}^k \frac{\delta_j}{\delta_j + \beta} = \gamma \quad (3-8)$$

According to (3-7) and (3-8), we rewrite the new error criterion as :

$$M = \alpha E_D + \beta E_W \approx \mathbf{e}^T \mathbf{e} + \lambda \mathbf{q}^T \mathbf{q} \quad (3-9)$$

Following that, we can find the correct value to use for λ by given an initial value of $\lambda (\lambda \geq 0)$ in the following iterative procedure. The value of λ is updated by the formula as follows :

$$\lambda_{k+1} = \frac{\gamma_k}{n - \gamma_k} \frac{\mathbf{e}^T \mathbf{e}}{\mathbf{q}^T \mathbf{q}} \quad (3-10)$$

where

$$\gamma_k = \sum_{j=1}^k \frac{\mathbf{t}_j^T \mathbf{t}_j}{\mathbf{t}_j^T \mathbf{t}_j + \lambda_j} \quad (3-11)$$



Chapter 4

Experiments and Results

In this chapter, we will demonstrate the simulation experiment results including sigmoid function, Gaussian-based spectrum and data preprocessing. In our simulation data experiments, the results shows the analyzed performance are nearly between Bayesian-based PLS and PRLS and they all have better performance than original PLS.

4.1 Illustration

4.1.1 Synthesized simulation data

In simulation data calculation, we use synthesize testing data with noise to examine the efficiency of our method. We add the noise generated by Gaussian probability density function with zero mean and set the value of standard deviation, so as to vary the level of noise. The noise to signal (N/S) ratio is also used to set up a standard of the variation. Given a signal data set $signal_i$ and Gaussian noise data set $noise_i$ with zero mean, $1 \leq i \leq n$.

The mean of signal and noise data set are :

$$\begin{aligned}\mu_S &= \frac{\sum_{i=1}^n signal_i}{n} \\ \mu_N &= \frac{\sum_{i=1}^n noise_i}{n}\end{aligned}\tag{4-1}$$

The variance of the signal and noise data set are :

$$Var(signal) = \frac{\sum_{i=1}^n (signal_i - \mu_s)^2}{n}$$

$$Var(noise) = \frac{\sum_{i=1}^n (noise_i - \mu_N)^2}{n} \quad (4-2)$$

The noise to signal (N/S) ratio is

$$N/S \text{ ratio} = \frac{\sqrt{Var(noise)}}{\sqrt{Var(signal)}} \quad (4-3)$$

4.1.2 Criterion of estimation

Here we use two familiar estimators to verify the performance of our method. One of them is root mean square error (RMSE). RMSE is one of many ways to quantify the amount by which an estimator differs from the true value of the quantity being estimated like as a loss function. The other one is correlation coefficient which indicates the strength and direction of a linear relationship between two variables. The correlation coefficient is a value between 0 and 1. It is a measure of how well trends in the predicted values follow trends in past actual values. Following, we illustrate the concept of correlation coefficient in Figure 4.1.

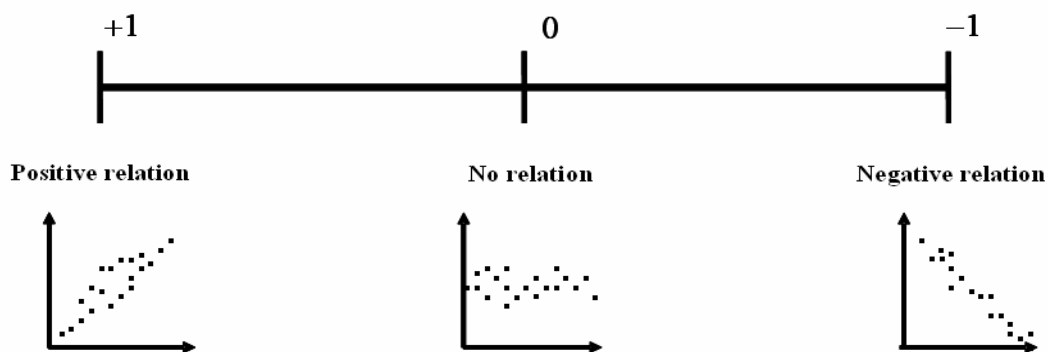
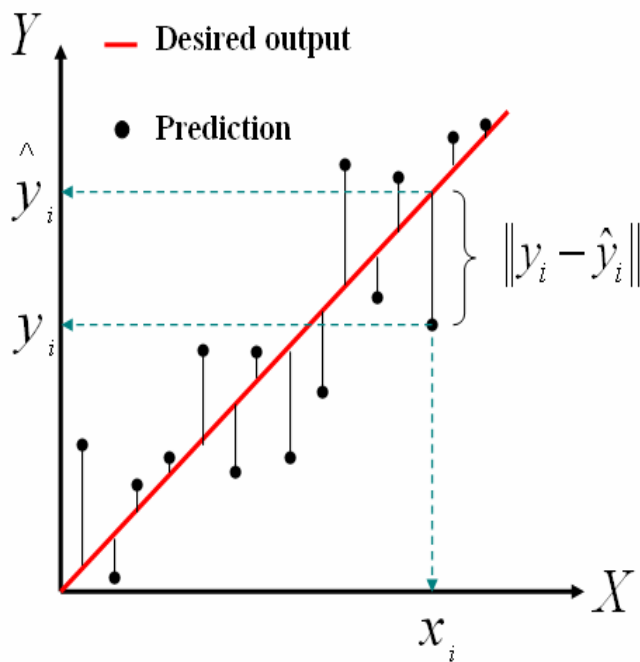


Figure 4.1 A sketch map of correlation coefficient

Given a set of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, the formula for computing the correlation coefficient is given by :

$$r = \frac{1}{n-1} \sum \left(\frac{\mathbf{X} - \bar{\mathbf{X}}}{S_x} \right) \left(\frac{\mathbf{Y} - \bar{\mathbf{Y}}}{S_y} \right) \quad (4-4)$$



$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

To acquire accurate prediction, we hope that RMSE value minimize as far as possible.

Figure 4.2 Root mean square error

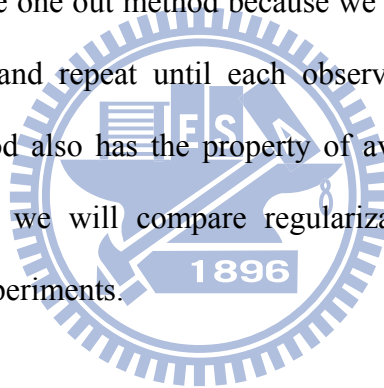
Figure 4.2 shows that the main concept of RMSE is to calculate the average of the distance between prediction and desired output data. To acquire accurate prediction, we hope that RMSE minimizes as far as possible.

4.1.3 Conditional training

Here we also calibrate the training data in different conditions : (1) self-calibration and self-prediction (SCSP) and (2) cross validation (CV). In order to understand easily what is difference between SCSP and CV. We use diagrams to illustrate. Figure 4.3 shows the principle of SCSP and Figure 4.4 shows CV.

SCSP is a traditional training mode and the training data set is also prediction data set. Usually the result of SCSP is ideal if there is no noise hidden in the source data. However data usually goes along with noise and SCSP would be influenced by hidden information so that results may not necessarily meet to desire.

CV is also called leave one out method because we select a validation data from original training data set and repeat until each observation in the set is used as validation data. The method also has the property of avoiding overfitting but costs heavy computation. Next, we will compare regularization technique and CV in simulation and real data experiments.



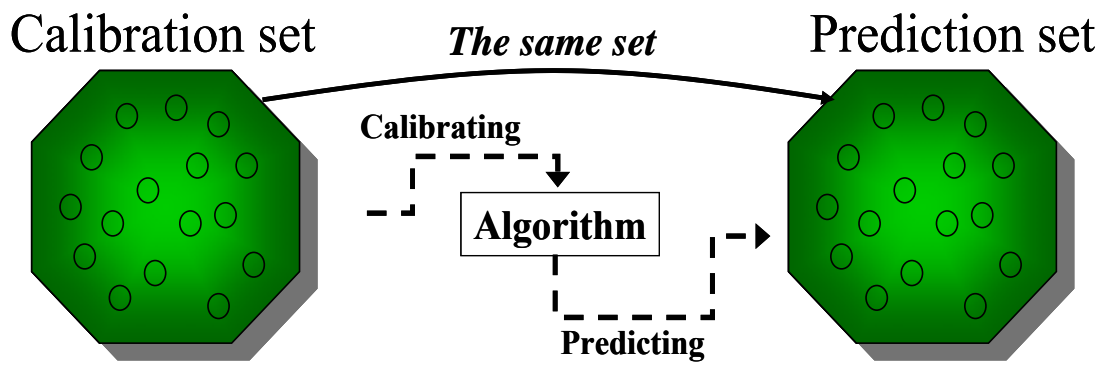


Figure 4.3 Self-calibration and self-prediction (SCSP)

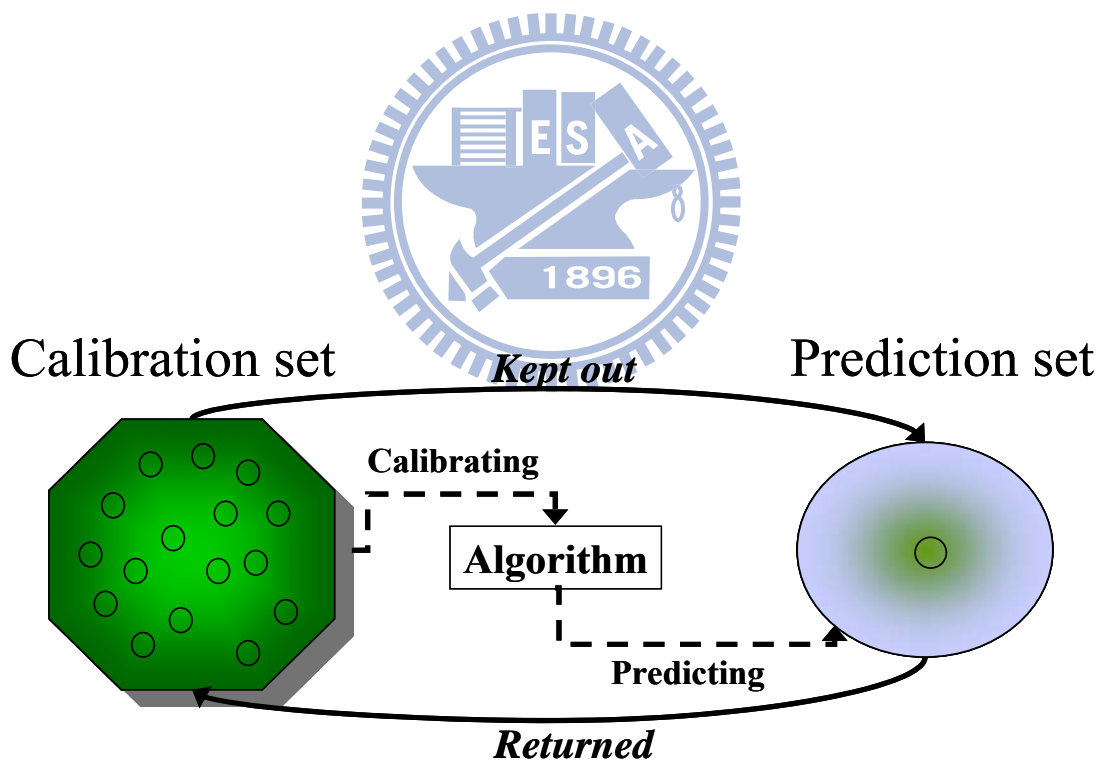


Figure 4.4 Cross validation (CV)

4.2 Simulation data

In this section, we will generate sigmoid function, Gaussian-based spectrum data and preprocessing procedure under SCSP and CV condition. We calibrate these different kinds of data set. After predicting, we apply the criterion of estimation to examine which one is better among PLS, PRLS and Bayesian-based PLS methods.

4.2.1 Sigmoid function

In this simulation, we use hybrid of sine and cosine function to examine.

$$f(x_i) = a_i \sin(x_i) + b_i \cos(x_i), \quad 0 \leq x \leq 2\pi \quad (4-5)$$

The training data were generated from $f(x_i) + \varepsilon_i$, where x_i has take from the uniform distribution in $(0, 2\pi)$ and the noise ε_i had a Gaussian distribution with zero mean. The training data and the sigmoid function $f(x_i)$ are plotted in Figure 4.5. The training data is highly ill-conditioned.

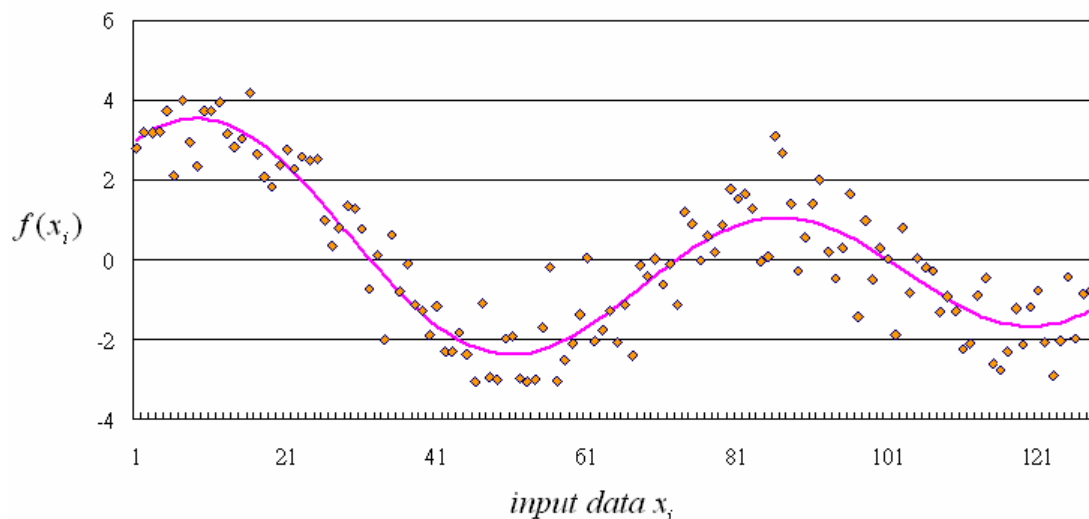


Figure 4.5 Noisy data (points) and sigmoid function (curve)

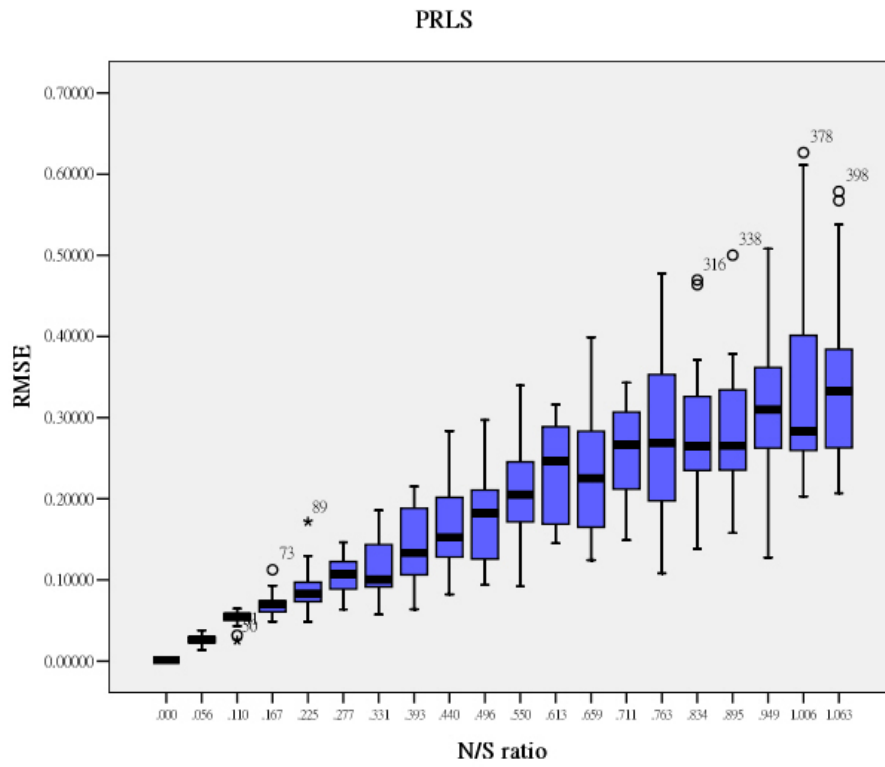


Figure 4.6 RMSE as a function of N/S ratio under SCSP (PRLS)

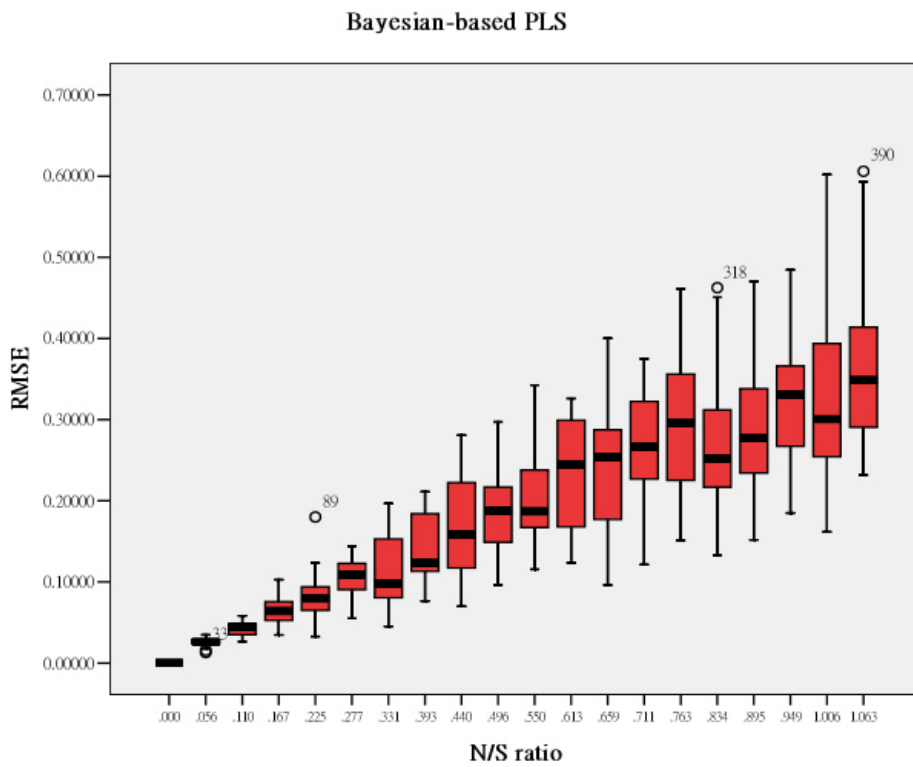


Figure 4.7 RMSE as a function of N/S ratio under SCSP (Bayesian-based PLS)

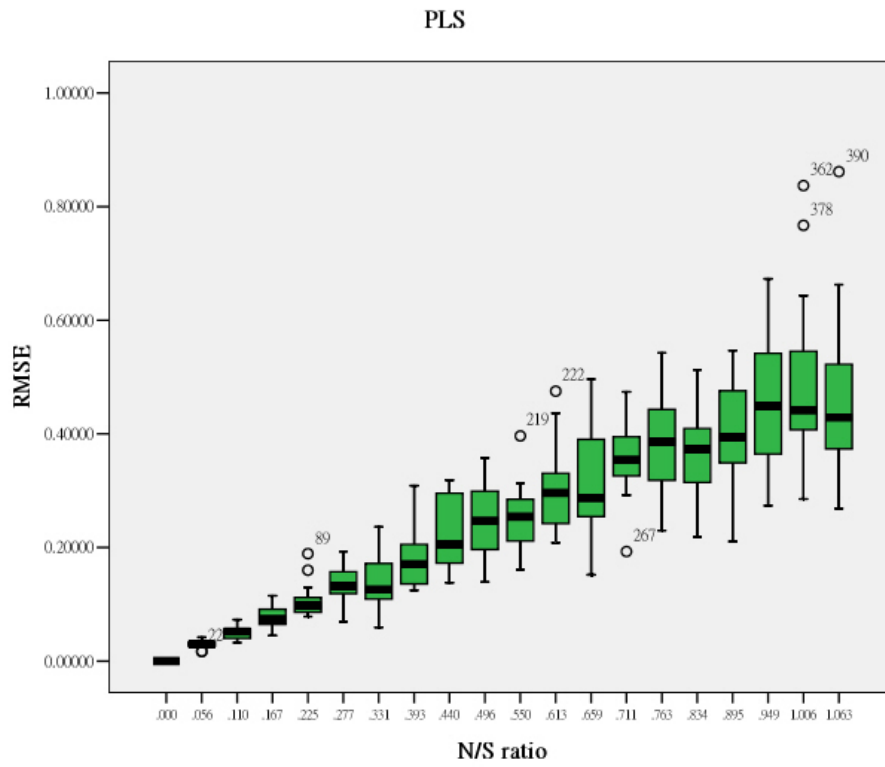


Figure 4.8 RMSE as a function of N/S ratio under SCSP (PLS)

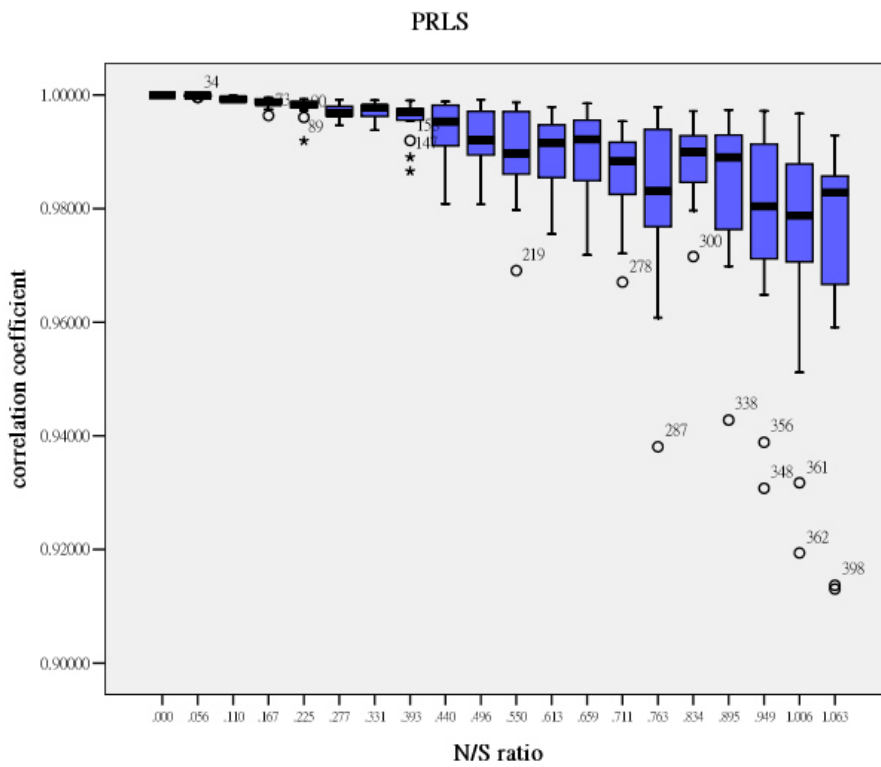


Figure 4.9 Correlation coefficient as a function of N/S ratio under SCSP (PRLS)

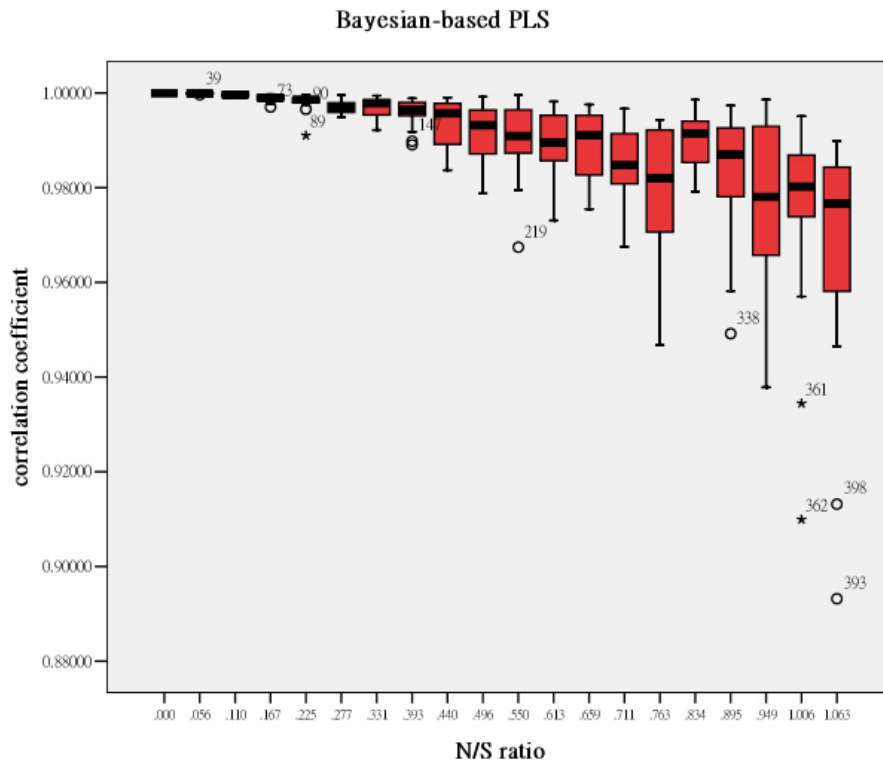


Figure 4.10 Correlation coefficient as a function of N/S ratio under SCSP (Bayesian-based PLS)

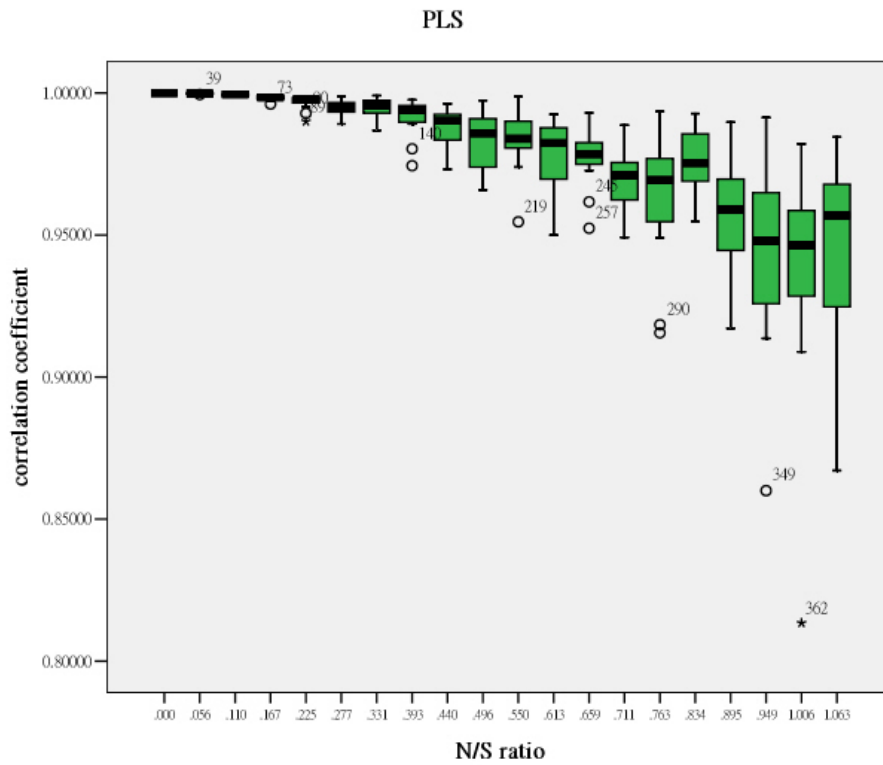


Figure 4.11 Correlation coefficient as a function of N/S ratio under SCSP (PLS)

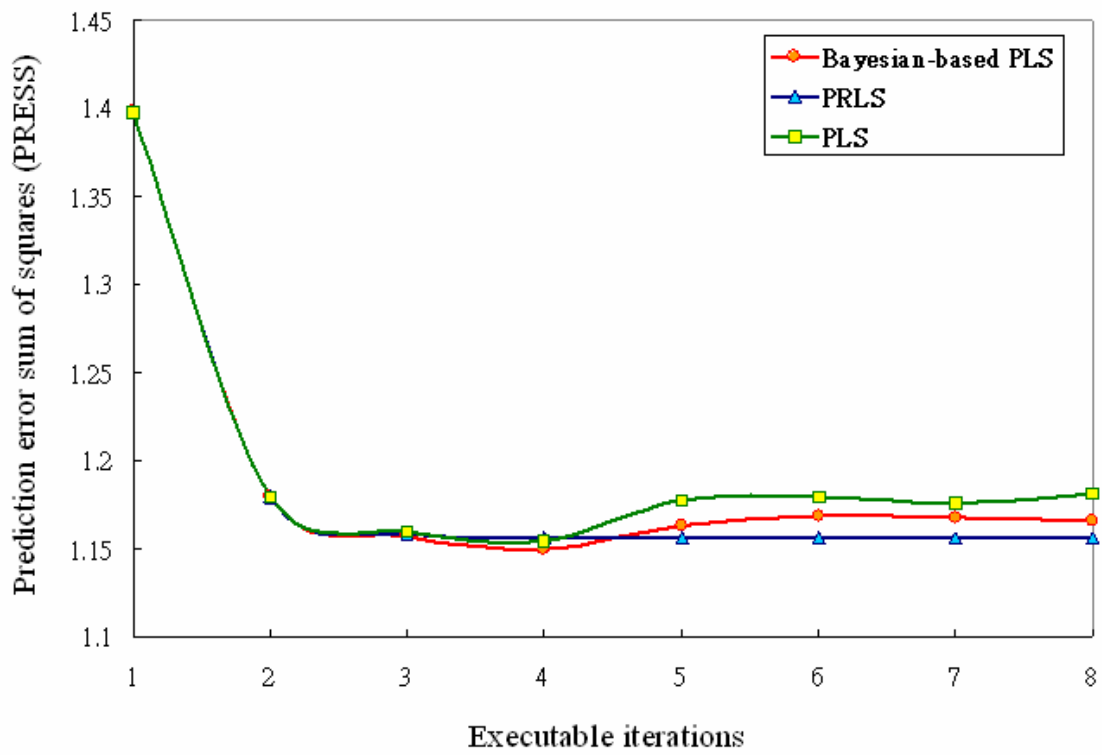


Figure 4.12 Prediction error sum of squares (PRESS) under CV

4.2.2 Gaussian-based spectrum

We would like to generate two Gaussian functions $g(x)$ with mean = 510 and the standard deviation = 15, $h(x)$ with mean = 540 and the standard deviation = 10. $f(x)$ is the linear combination of $g(x)$ and $h(x)$ plotted in Figure 4.13.

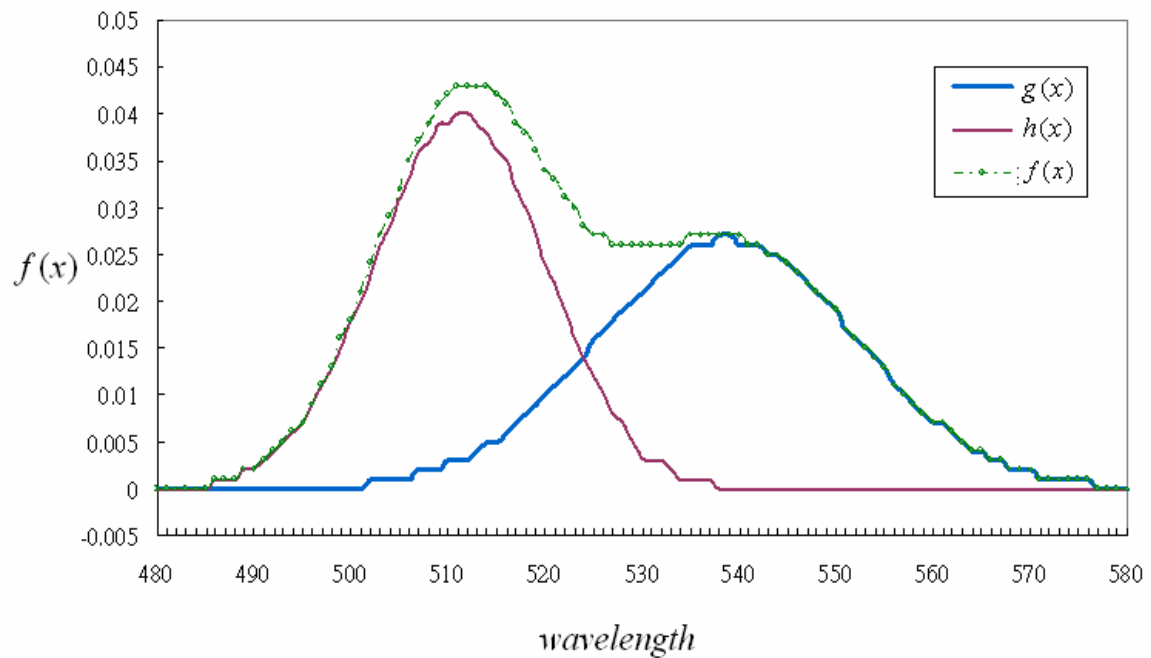


Figure 4.13 The linear combination of two Gaussian functions with different mean and standard deviation.

The training data set $X_i + \varepsilon$ can be generated by linear combination of $g(x)$ and $h(x)$ with Gaussian noise ε . The training data set will be represented in Figure 4.14.

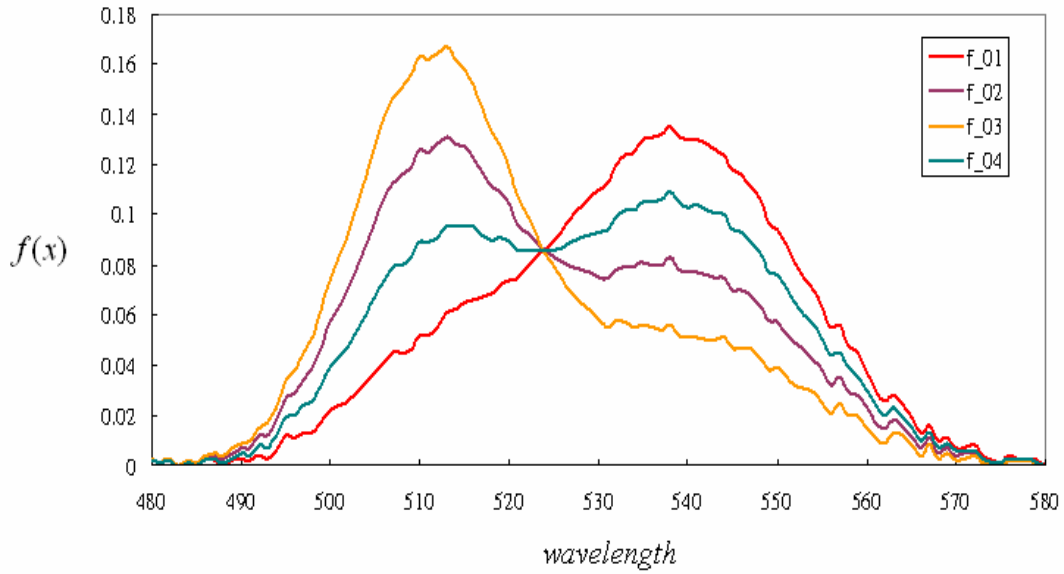
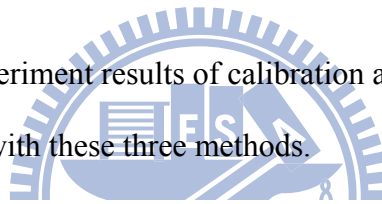


Figure 4.14 The training data set of Gaussian-base spectrum

Next, we will show the experiment results of calibration as follows. We can compare the analyzed performance with these three methods.



PRLS

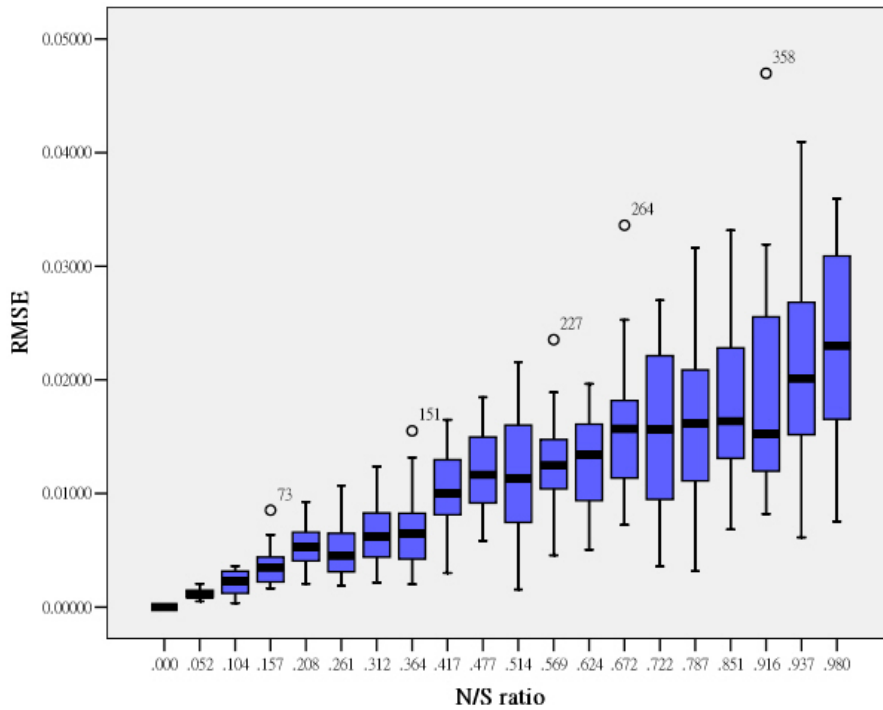


Figure 4.15 RMSE as a function of N/S ratio under SCSP (PRLS)

Bayesian-based PLS

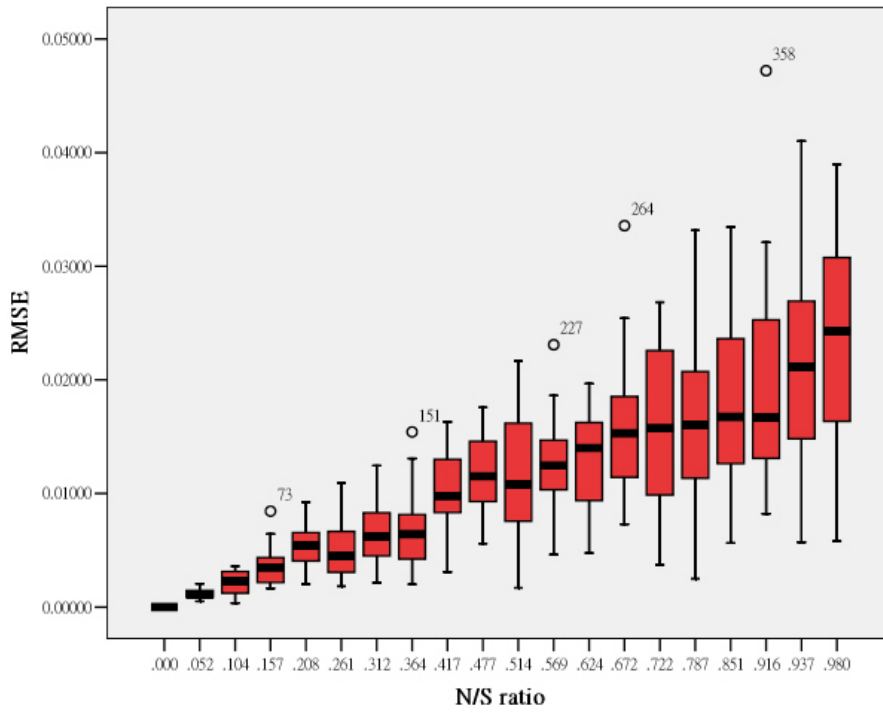


Figure 4.16 RMSE as a function of N/S ratio under SCSP (Bayesian-based PLS)

PLS

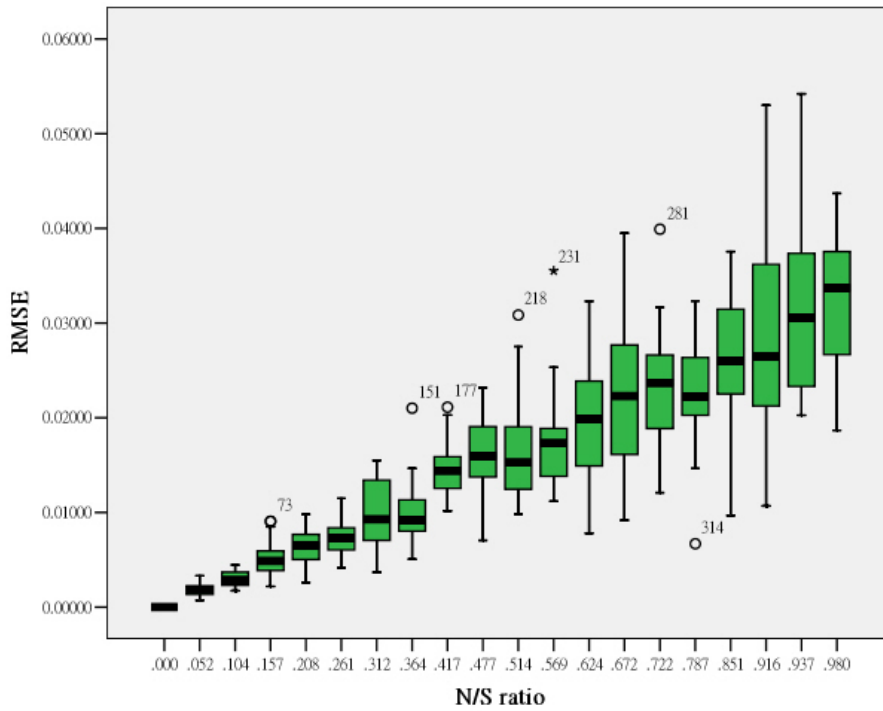


Figure 4.17 RMSE as a function of N/S ratio under SCSP (PLS)

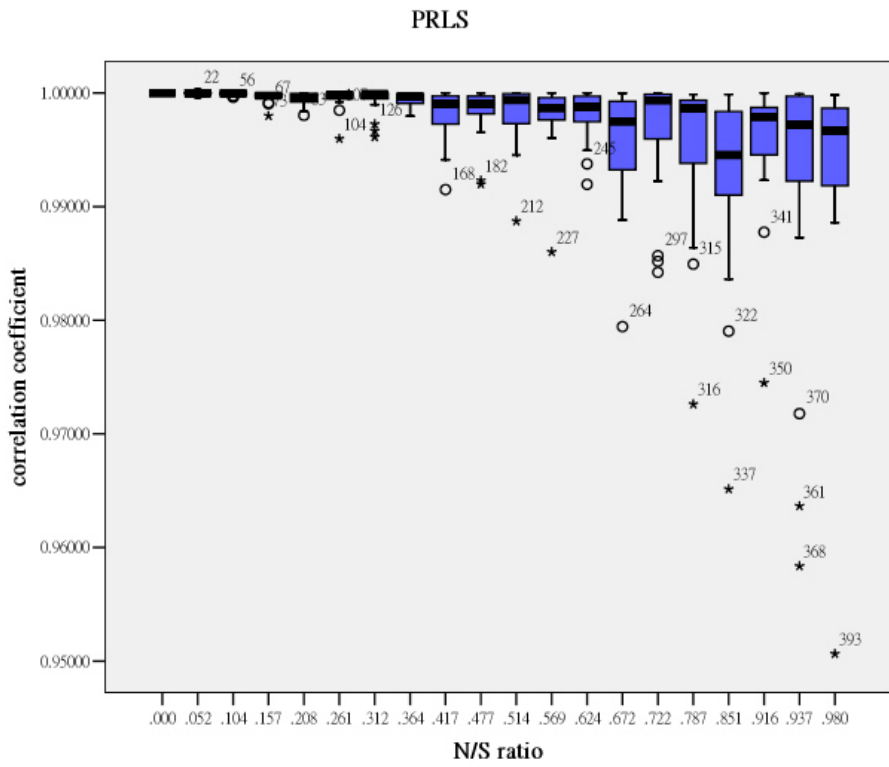


Figure 4.18 Correlation coefficient as a function of N/S ratio under SCSP (PRLS)

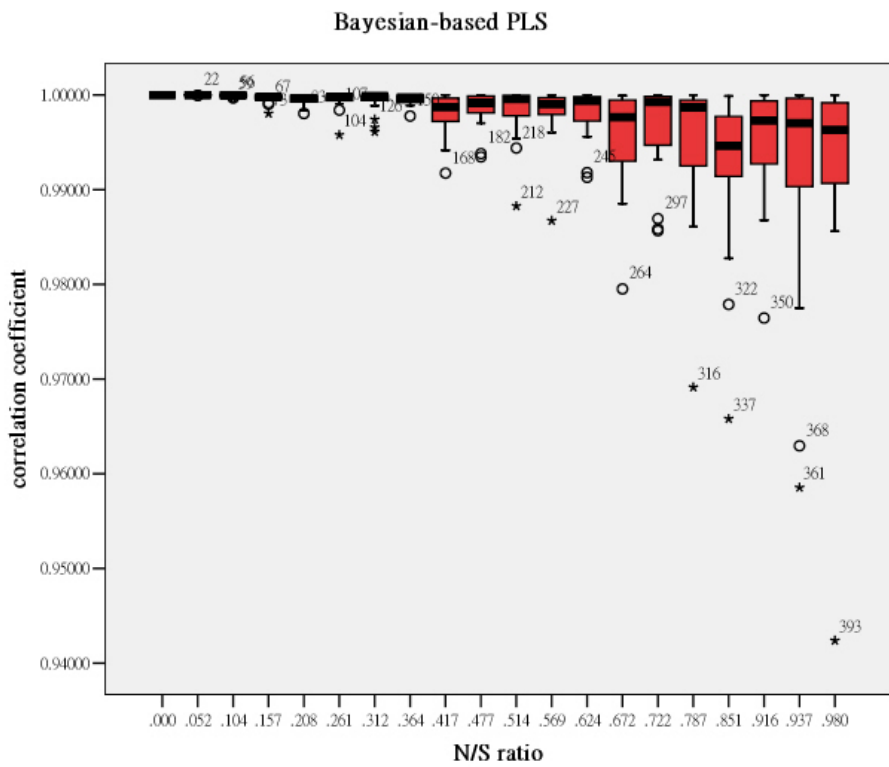


Figure 4.19 Correlation coefficient as a function of N/S ratio under SCSP (Bayesian-based PLS)

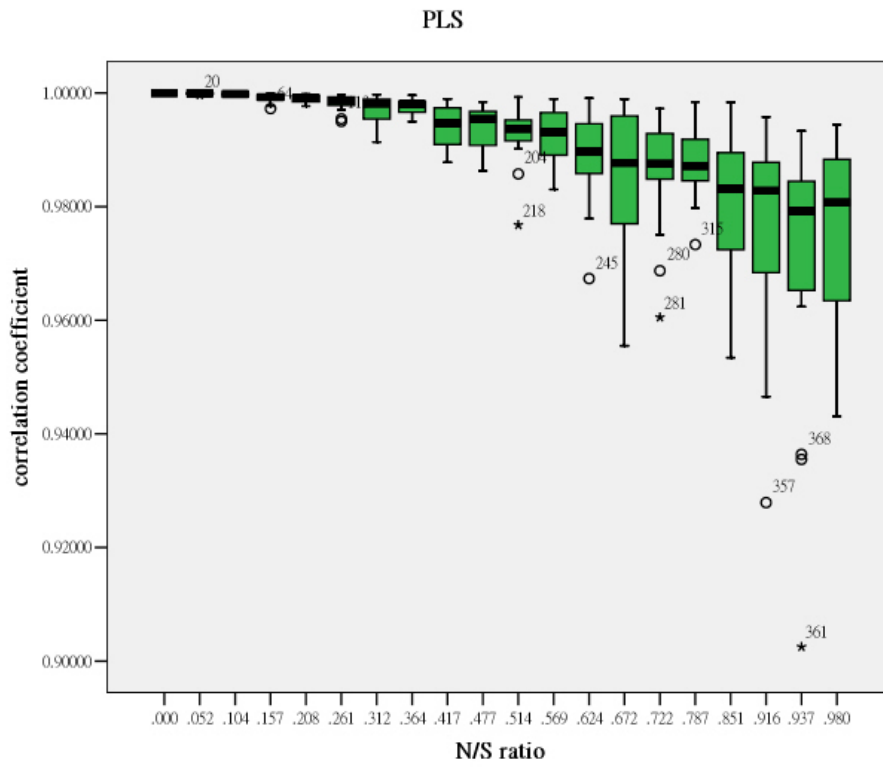


Figure 4.20 Correlation coefficient as a function of N/S ratio under SCSP (PLS)

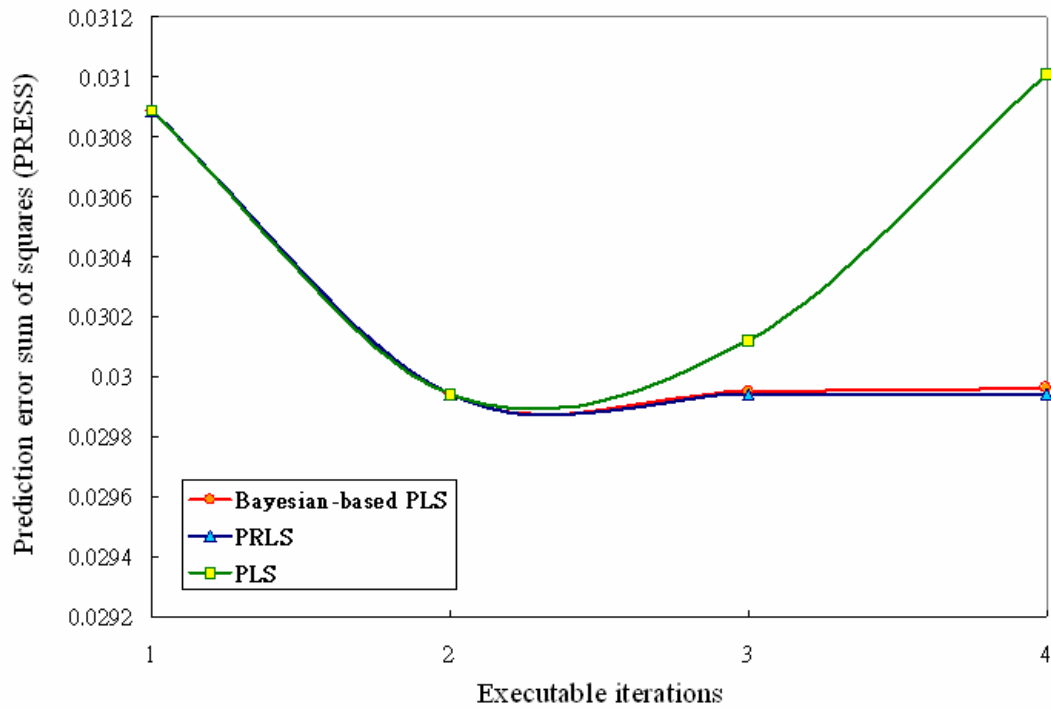


Figure 4.21 Prediction error sum of squares (PRESS) under CV

From our experiment results, we can find out both Bayesian-based PLS and PRLS have better performance than PLS, and the analyzed result of Bayesian-based PLS is nearly to PRLS whether the prediction is under SCSP condition. The Figure 4.12 shows the CV result, we can realize that Bayesian-based PLS is not better than PRLS. We think that the result might be influenced by the selection of prior and the data we simulate.

4.2.3 Preprocessing

In this part, we generate original training data set taken from the uniform distribution in $(-1, 1)$. Then we use tangent sigmoid as a function to transfer the original training data set to a new one which is Gaussian-based. We make an assumption about whether the training data set is center distribution or more uniform will make better analyzed performance. They will give some illustration in following figures.

Figure 4.22 and 4.23 represent the original training data set X and the new training data set X' which is transferred by tangent function respectively.

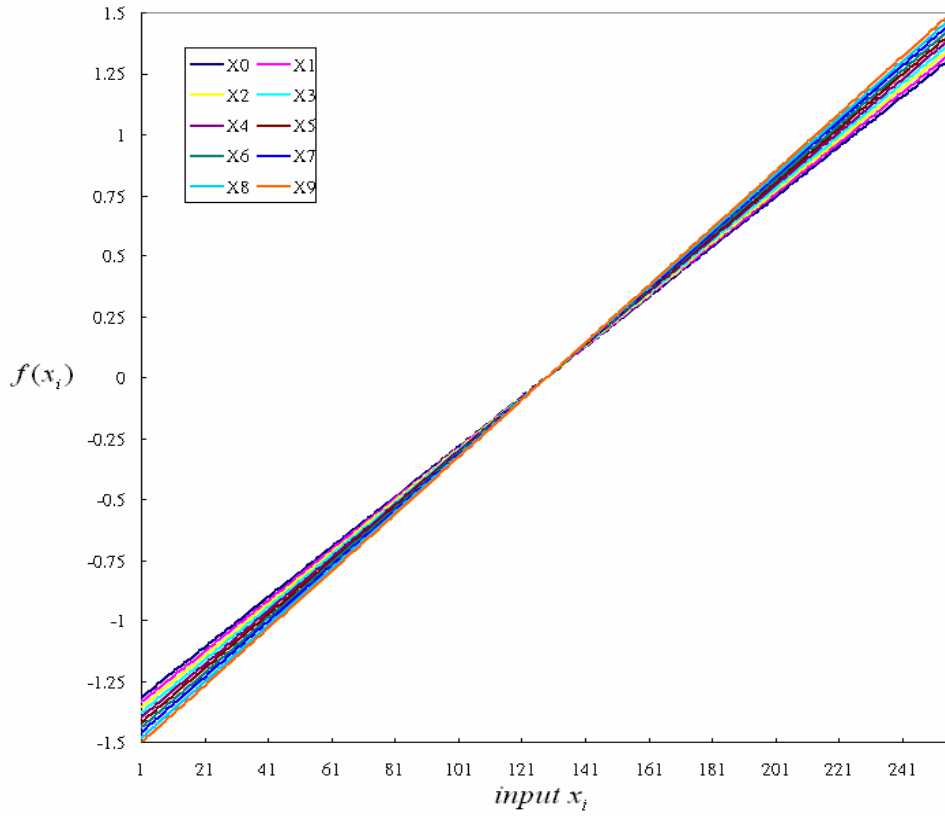


Figure 4.22 The original training data set X

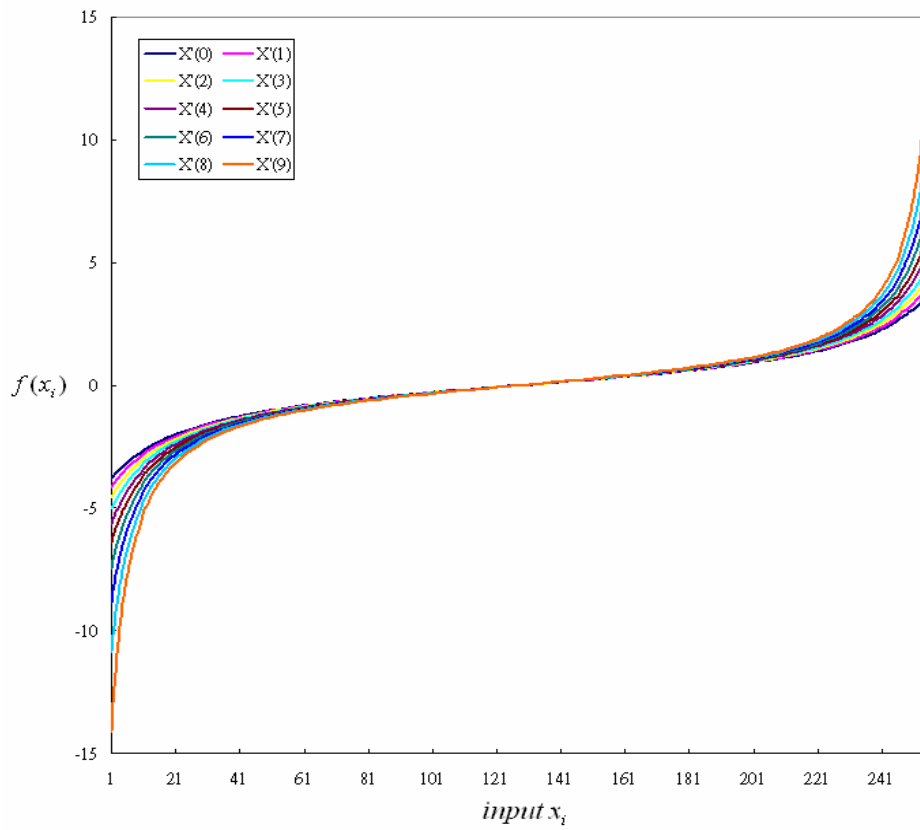


Figure 4.23 The new training data set X'

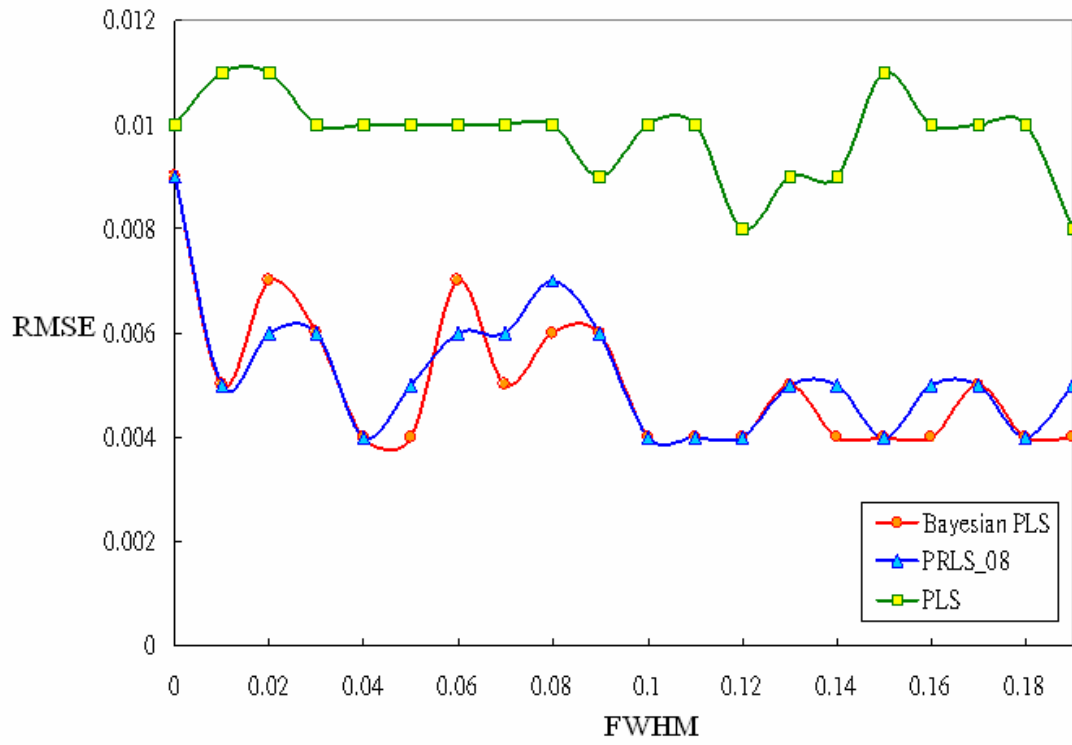


Figure 4.24 RMSE as a function of FWHM under SCSP

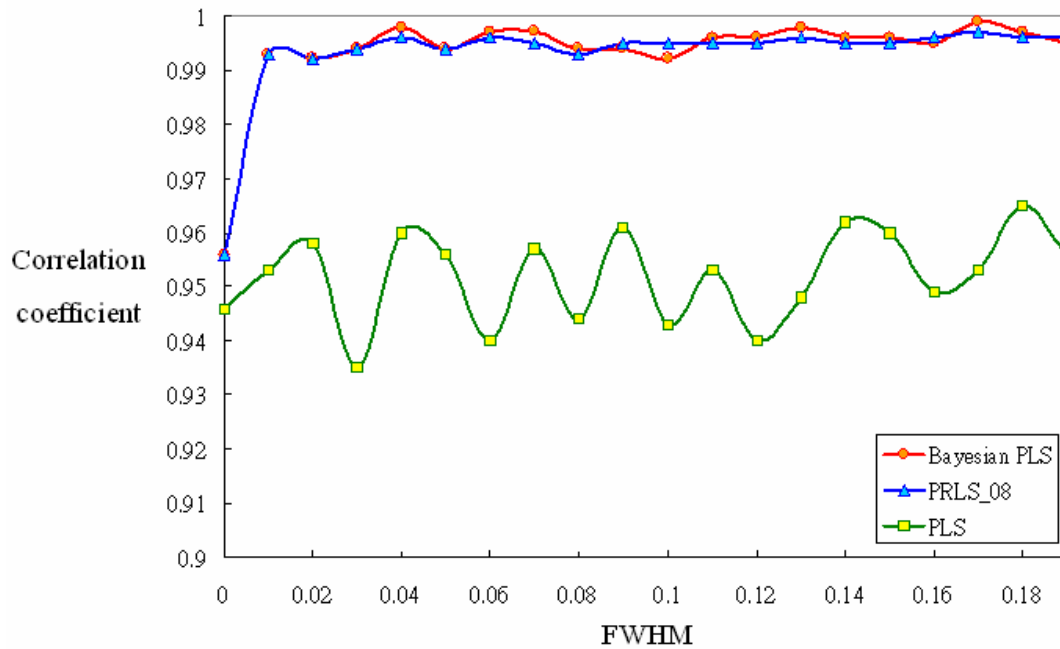


Figure 4.25 Correlation coefficient as a function of FWHM under SCSP

Chapter 5

Discussion

For regularization concept, almost all inverse problem methods involve a trade-off between two optimizations : agreement between data and solution, and smoothness of the solution. We define that the unconstrained minimum of agreement and the unconstrained minimum of smoothness is the best solution. Figure 5.1 will give you a brief thought about that. Here, we have a question for how to define or find out the location of the best solution between “Best smoothness” line and “Best agreement” line.

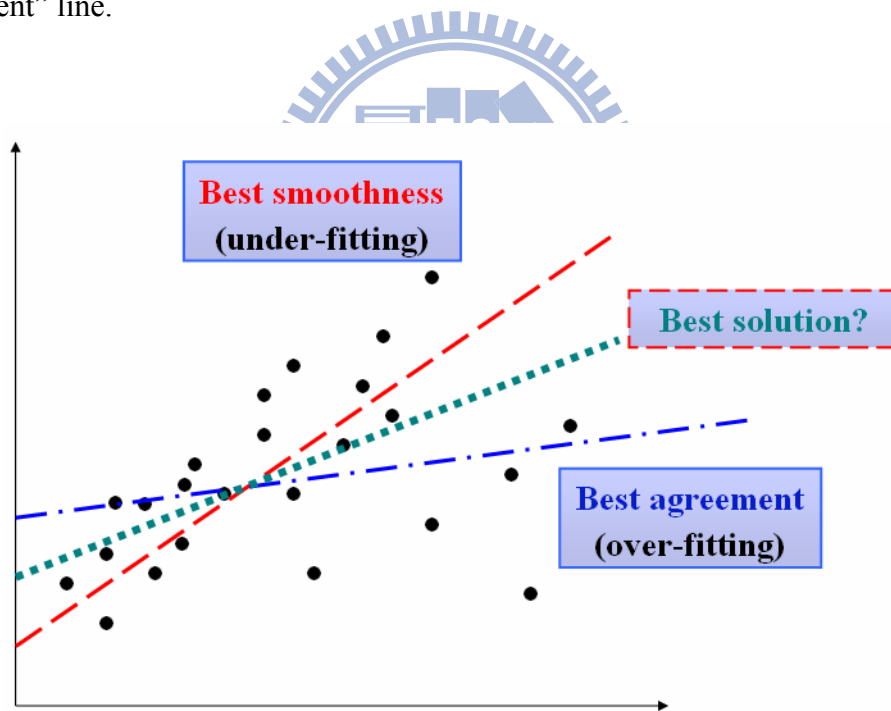


Figure 5.1 Where is the best solution

The estimated criterion RMSE and correlation coefficient would involve a trade-off relationship. In our data experiment results, we hope the RMSE is low and correlation coefficient is high to verify our proposed method. So, we need some

verification to explain this problem. We make a assumption that our proposed method and PLS may have different curves as shown in Figure 5.2. In further study, we will have a fundamental proof for this issue.

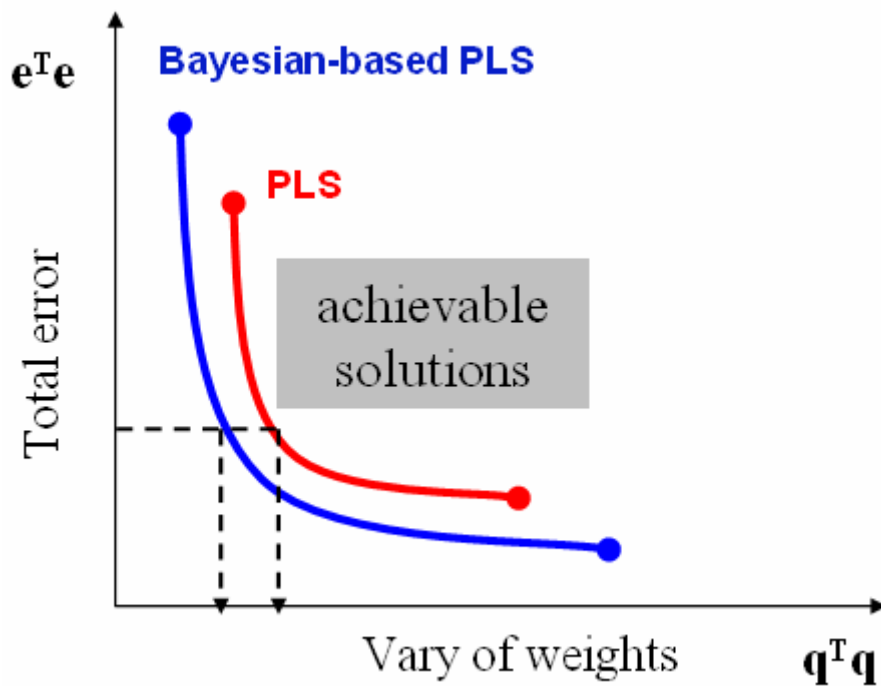


Figure 5.2 Trade-off curves of Bayesian-based PLS and PLS

The preprocessing result, we transfer the original data set to Gaussian form to examine whether the performance is better or not. We make different widths for FWHM to verify our proposed method. But we could obviously find out the hypothesis for data preprocessing doesn't accomplish to our expectation. The results after preprocessing might be influenced by the limitation of tangent function. The data after tangent function transferring may be divergent so that the analyzed results would be affected for this reason.

The local and global minimum problem is another issue we concern. We would like to find the best solution to approximate nearly global minimum.

Chapter 6

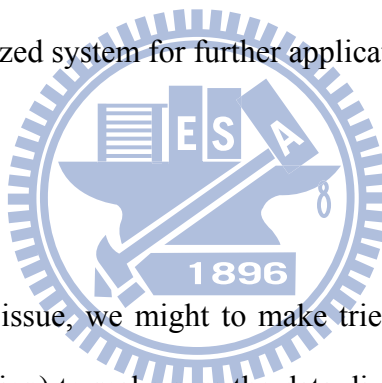
Conclusions and Future works

6.1 Conclusions

We have established a probability based analyzed method which combines the advantages of regularization and the properties of PLS for a novel calibration model. The proposed method, Bayesian-based PLS, is able to reduce the noise signal hidden in the training data. And it has better analyzed results than original PLS method when training data accompanying noise signal during calibration phase. So we can apply our method to on-line analyzed system for further application.

6.2 Future works

In data preprocessing issue, we might to make tries for other kinds of transfer function (e.g., arcsine function) to make sure the data divergent problem and improve the limitation of transformation accuracy to obtain better performance for further study. The track of best solution between the agreement and smoothness is our next objective to achieve. Then, we also consider to make the results approximated to the global minimum so that we can apply the proposed method for weights initialization of backpropagation network. There still have another issue we have to take into account. The selection of appropriate prior would probably affect the analyzed result. So we need to make a study about the prior probability to make sure that we don't have a bad or wrong one.



References

- [1] Hsiao TC, Lin CW, Chiang HH, “*Partial least squares algorithm for weights initialization of the back-propagation network*”, *Neurocomputing*, vol. 50, pp. 237-247, 2003.
- [2] Chen S, Chng ES, Alkadhimi K, “*Regularized orthogonal least squares algorithm for constructing radial basis function networks*”, *International Journal of Control*, vol. 64, pp. 829-837, 1996.
- [3] Chang SH, Chiou YJ, Yu C, Lin CW, Hsiao TC, “*A Novel Multivariate Analysis Method with Noise Reduction*”, 4th European Congress for Medical and Biomedical Engineering, 2008.
- [4] MacKay DJC, “*Bayesian interpolation*”, *Neural Computation*, vol. 4, pp. 415-447, 1992.
- [5] Bhandare P, Mendelson Y, Peura RA, Janatsch G, Kruse-Jarres JD, Marbach R, Heise HM, “*Multivariate determination of glucose in whole blood using partial least-squares and artificial neural networks based on mid-infrared spectroscopy*”, *Applied Spectroscopy*, vol. 47, pp. 1214-1221, 1993.
- [6] Möcks J, Verleger R, “*Multivariate methods in biosignal analysis: application of principal component analysis to event-related*”, *Techniques in the behavioral and neural sciences*, vol. 5, pp. 399-458, 1991.
- [7] Castellanos G, Delgado E, Daza G, Sanchez LG, Suarez JF, “*Feature Selection in Pathology Detection using Hybrid Multidimensional Analysis*”, *Proceedings of International Conference of EMBS*, pp. 5950-5953, 2006.
- [8] Oja E, “*A simplified neuron model as a principal component analyzer*”, *Journal of Mathematics and Biology*, vol. 15, pp. 267-273, 1982.
- [9] Harald M, Tormod N, “*Multivariate Calibration*”, 2nd Edition, John Wiley & Sons, Great Britain, 1996.

- [10] Huang KY, “*Neural Networks and Pattern Recognition*”, 2nd Edition, 維科圖書有限公司, 2003.
- [11] Oja E, Karhunen J, “*Recursive construction of Karhunen-Loeve expansions for pattern recognition purposes*”, Proceedings of 5th Int. Conf. on Pattern Recognition, pp. 1215-1218, 1980.
- [12] Hsiao TC, Lin CW, Zeng MT, Chiang Kenny HH, “*The Implementation of Partial Least Squares with Artificial Neural Network Architecture*”, 20th Annual International Conference of the IEEE Engineering in Medicine Biology Society, vol. 3, pp. 1341-1343, 1998.
- [13] Chen S, Cowan CFN, Grant PM, “*Orthogonal least squares learning algorithm for radial basis function networks*”, IEEE Transactions on Neural Networks, vol. 2, pp. 302-309, 1991.
- [14] Press HW, Vetterling WT, Teukolsky SA, Flannery BP, “*Numerical Recipes in C: the art of scientific computing*”, 2nd Edition, Cambridge University Press, 1993.
- [15] Orr MJL, “*Regularization in the selection of radial basis function centers*”, Neural Computation, vol. 7, pp. 606-623, 1995.
- [16] Hertz J, Krough A, Palmer R, “*Introduction to the Theory of Neural Computation*”, Redwood city, California, USA, Addison-Wesley, 1991.
- [17] Ham FM, Kostanic I, “*A Neural Network Architecture for Partial Least Squares Regression with Supervised Adaptive Modular Hebbian Learning*”, Neural, Parallel, Scientific Computation, vol. 6, pp. 35-72, 1998.
- [18] Jeffreys H, “*Theory of Probability*”, Oxford University Press, 1939.
- [19] Gull SF, “*Bayesian inductive inference and maximum entropy*”, Maximum Entropy and Bayesian Methods in Science and Engineering, vol. 1, pp. 53-74, 1988.