Bayesian Learning for Earthquake Engineering Applications and Structural Health Monitoring

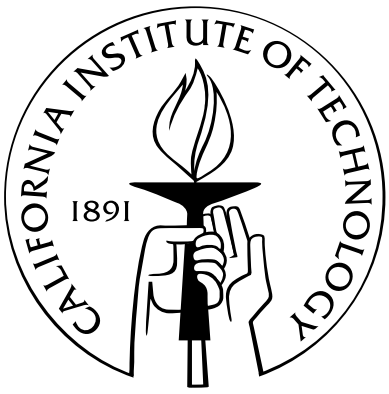
Thesis by

Chang Kook Oh

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2007

(Defended September 17, 2007)

Acknowledgements

I am deeply grateful to God for his sincere guidance through the past five years at Caltech and for help and consolation with which I have overcome difficulties. I hope and pray this work would be beneficial in humble ways to help people and to protect their lives and structures from dangerous earthquakes and damage.

I would like to thank my advisor, Dr. Jim Beck, with all my heart. With his guidance, patience, and especially astonishing insight, it has been possible for me to make academic achievements. For the past five years, he has been an advisor, a researcher, a mentor, and a role model to me. I appreciate this greatly.

I also would like to express my gratitude to my committee members, Dr. John Hall, Dr. Tom Heaton, Dr. Keith Porter, and Dr. G. Ravichandran for their warm encouragement and especially the contributions to make me grow as a researcher, a scientist, and an engineer.

I would like to thank my friends, inside and outside of Caltech. First, I thank John, Case, Swami, Jeff, and Matt. They are colleague researchers who came to Caltech CEE earlier than me and helped me especially at my first year. It was valuable to me to study with Judy, Jing, Li, John, Tomo, Kunihiko, Alex, Daniel, and Joseph. We studied together, shared knowledge and idea, and formed plans for our future. Because of you, the time I spent at Caltech was more precious. I also thank a lot of Korean colleagues, especially, Inchan, Jinyoo, Sanggeun, Byungjoon, Wonjin, Chihoon, Taewan, and Shinchul. I also thank L.A. Onnuri Church fellows, especially, Paster Jinso Yoo, Paster Byungjoo Song, Wonchun and Yongwoo in Choir. I will never forget the prayers, hospitality, and warm encouragement you have showed constantly. Many thanks to my colleague friend, Junho, Hyuck, Seungdae, and Hyunwoo, who

inspired me not to lose a vision for future.

It is a privilege to have precious friends like Junho, Jongwook, and Sukjin. Thanks for your friendship which will last forever.

Above all, I would like to thank my parents, parents-in-law, and a brother, Changmin for their everlasting love, prayer and support. They are always with me in a moment of joys and sorrows. I could not have accomplished this without you.

Last but definitely not least, I deeply thank my beloved wife, Heejo, for her dedication, unfailing encouragement, belief in my dream, and timeless love. Thank you.

Abstract

Parallel to significant advances in sensor hardware, there have been recent developments of sophisticated methods for quantitative assessment of measured data that explicitly deal with all of the involved uncertainties, including inevitable measurement errors. The existence of these uncertainties often causes numerical instabilities in inverse problems that make them ill-conditioned.

The Bayesian methodology is known to provide an efficient way to alleviate this ill-conditioning by incorporating the prior term for regularization of the inverse problem, and to provide probabilistic results which are meaningful for decision making.

In this work, the Bayesian methodology is applied to inverse problems in earthquake engineering and especially to structural health monitoring. The proposed methodology of Bayesian learning using automatic relevance determination (ARD) prior, including its kernel version called the Relevance Vector Machine, is presented and applied to earthquake early warning, earthquake ground motion attenuation estimation, and structural health monitoring, using either a Bayesian classification or regression approach.

The classification and regression are both performed in three phases: (1) Phase I (feature extraction phase): Determine which features from the data to use in a training dataset; (2) Phase II (training phase): Identify the unknown parameters defining a model by using a training dataset; and (3) Phase III (prediction phase): Predict the results based on the features from new data.

This work focuses on the advantages of making probabilistic predictions obtained by Bayesian methods to deal with all uncertainties and the good characteristics of the proposed method in terms of computationally efficient training, and, especially,

prediction that make it suitable for real-time operation. It is shown that sparseness (using only smaller number of basis function terms) is produced in the regression equations and classification separating boundary by using the ARD prior along with Bayesian model class selection to select the most probable (plausible) model class based on the data. This model class selection procedure automatically produces optimal regularization of the problem at hand, making it well-conditioned.

Several applications of the proposed Bayesian learning methodology are presented. First, automatic near-source and far-source classification of incoming ground motion signals is treated and the Bayesian learning method is used to determine which ground motion features are optimal for this classification. Second, a probabilistic earthquake attenuation model for peak ground acceleration is identified using selected optimal features, especially taking a non-linearly involved parameter into consideration. It is shown that the Bayesian learning method can be utilized to estimate not only linear coefficients but also a non-linearly involved parameter to provide an estimate for an unknown parameter in the kernel basis functions for Relevance Vector Machine. Third, the proposed method is extended to a general case of regression problems with vector outputs and applied to structural health monitoring applications. It is concluded that the proposed vector output RVM shows promise for estimating damage locations and their severities from change of modal properties such as natural frequencies and mode shapes.

Contents

Acknowledgements **iii**

Abstract **v**

1 Introduction **1**

1.1 Motivation 1

1.2 Bayesian Methodology 3

1.3 Background on Structural Health Monitoring 4

1.3.1 Pattern Classification Applied to Structural Health Monitoring 6

1.3.2 Regression Procedure Applied to Structural Health Monitoring 7

1.4 Organization 7

2 Bayesian Learning Using Automatic Relevance Determination Prior and Relevance Vector Machine **9**

2.1 Bayesian Learning Method Using Various Priors 10

2.1.1 Least-Squares Estimation using Linear Model 10

2.1.2 Bayesian Inference 12

2.2 Bayesian Learning Using Automatic Relevance Determination Prior: I–Regression 17

2.2.1 Training Phase in Regression 18

2.2.2 Prediction Phase in Regression 21

2.3 Bayesian Learning Using Automatic Relevance Determination Prior: II–Classification 22

2.3.1 Training Phase in Classification 22

2.3.2 Prediction Phase in Classification 26
2.4 Relevance Vector Machine 26
2.4.1 Kernel Methods 27
2.4.2 Relevance Vector Machine Regression 28
2.4.2.1 Training Phase 28
2.4.2.2 Prediction Phase 31
2.4.3 Relevance Vector Machine Classification 31
2.4.3.1 Training Phase 31
2.4.3.2 Prediction Phase 33
2.4.4 Relevance Vector Machine Regression for Vector Outputs . . . 33
2.4.4.1 Training Phase 33
2.4.4.2 Prediction Phase 37
2.4.5 Illustrative Examples 38
2.4.5.1 RVM Regression: Sinc Function Estimation 38
2.4.5.2 RVM Classification: Ripley's 2-D Gaussian-mixture Dataset 40

3 Near-source and Far-source Classification for Earthquake Early Warning 43
3.1 Near-source and Far-source Classification 44
3.1.1 Earthquake Data 44
3.1.2 Separating Boundary Model 45
3.2 Comparison of Results 46
3.2.1 Function for Separating Boundary 46
3.2.2 Leave-One-Out Cross-Validation 48
3.2.3 Posterior Probability of Each Model Class 49
3.2.4 Effect of Prior 51
3.3 Conclusions 52

4 Ground Motion Attenuation Relations (Ground Motion Prediction Equations) using Regression 54

4.1 Estimation of Earthquake Ground Motion 55
4.1.1 Earthquake Data 55
4.1.2 Boore-Joyner Attenuation Model 55
4.1.3 Training Phase for PGA Estimation 56
4.1.4 Posterior Robust Predictive Probability Distribution for PGA 57
4.1.5 Estimation of a Non-linearly Involved Parameter 58
4.1.5.1 Inclusion of the Fictitious Depth Defining Model Class 58
4.1.5.2 Estimating the Fictitious Depth by Stochastic Simulation 58
4.2 Comparison of Results 60
4.3 Conclusions 61

5 Structural Health Monitoring 69
5.1 Illustrative Examples for Structural Health Monitoring 69
5.1.1 RVM Classification for SHM 69
5.1.1.1 Planar Shear Building Models 69
5.1.1.2 Bridge Models 70
5.1.1.3 Classification Results 71
5.1.2 RVM Regression with Vector Outputs 75
5.1.2.1 Step 1: Detecting Damage Locations 77
5.1.2.2 Step 2: Assessing Damage Severity 78
5.2 IASC-ASCE Structural Health Monitoring Benchmarks 80
5.2.1 IASC-ASCE Benchmark Structure 80
5.2.2 Identification of Modal Parameters 81
5.2.3 Damage Cases and Damage Patterns 85
5.2.4 Damage Cases 1 – 3 86
5.2.4.1 Training Phase 86
5.2.4.2 Prediction Phase 88
5.2.5 Damage Cases 4 – 5 96
5.2.5.1 Training Phase 97

5.2.5.2 Prediction Phase 99
5.2.5.3 Identification of Damage Locations 99
5.2.5.4 Assessment of Damage Severities 100
5.3 Conclusions 107

6 Concluding Remarks and Future Work **109**

A Posterior PDF and evidence by Using Bayes' Theorem **120**

B Laplace Approximation **123**

C Bayesian Model Class Selection **126**
C.1 Hyperparameter Optimization 126
C.2 Noise Variance Optimization 128

List of Figures

2.1 Illustration of Ockham's Razor (Mackay, 1992b). 17
2.2 Configuration of Sigmoid Function. 24
2.3 RVM and SVM Regression Based on Dataset of 100 Points Generated from sinc Function with Gaussian Noise of Mean 0 and Standard Deviation 0.01. 39
2.4 RVM and SVM Classification Based on Dataset of 250 Points Drawn Randomly from Two Mixtures of Two 2-D Gaussians. (The RVs and SVs that control the decision boundaries are shown as blue circles. P_1 and P_0 denote the probabilities of labels $y = 1$ and $y = 0$, respectively.) 42

4.1 Estimated Values of Regression Coefficients and Samples of Fictitious Depth during MCMC Algorithm. 59
4.2 Mean and Standard Deviation of Regression Coefficients and the Generated Fictitious Depth during MCMC Algorithm. 59

5.1 Configuration of 2-D Truss and 3-D Frame Bridge Structures. 70
5.2 RVM and SVM Classification Using the Dataset of the Scaled Fundamental Mode Shape Simulated from the 3-story Building. 10% and 3% noise are added to the fundamental mode shape and fundamental frequency, respectively. 72
5.3 Steel Frame Scaled Model Structure Used for Benchmarks. (taken from *http://mase.wustl.edu/wusceel/asce.shm/structure.htm.*) 82
5.4 Configuration of Benchmark Structure. 83

5.5 Sample Correlations between the First-Floor Reference Channel and the Measured Accelerations at the First to Fourth Floor, respectively. . . . 89
5.6 Identified Mode Shapes for Case 1, Pattern 0, 1, and 2. 91
5.7 Identified Mode Shapes for Case 2, Pattern 0, 1, and 2. 92
5.8 Identified Mode Shapes for Case 3, Pattern 0, 1, and 2. 93
5.9 Floor Plan for Benchmark Structure. 98
5.10 Correlations between Measurements at Reference Channel and Measured Accelerations at Node 11, 13, 15, and 17, respectively. 100
5.11 Identified Mode Shapes for Case 4, Pattern 0, 1, 2, 3, and 4. 102
5.12 Identified Mode Shapes for Case 5, Pattern 0, 1, 2, 3, 4, 5, and 6. . . . 103

List of Tables

2.1 Comparison between RVM and SVM for Sinc Function Estimation. . . 40

2.2 Comparison between RVM and SVM for Gaussian-mixture Dataset. . . 41

3.1 Number of Near-source and Far-source Records in Earthquake Dataset Used for Classification. ([a] moment magnitude M_w is cited from Havard CMT solution and [b] listed fault models are utilized to classify near-source and far-source station.) 45

3.2 Eight Extracted Features. 45

3.3 Coefficients for Optimal Separating Boundary Function for Each Model Class. ([a] N_i is the number of parameters used for each model. [b]– means the corresponding parameters are not considered for each model. [c]0 means the corresponding parameters are automatically pruned during training.) 48

3.4 Prior Covariance Matrix for Each Model Class. ([a]diag means diagonal matrix with the diagonal elements following.) 48

3.5 Classification Results for Earthquake Database Using Three Different Model Classes. (Bold values represent the least misclassification rate.) 48

3.6 Misclassification Rates Based on Leave-One-Out Cross-Validation. . . . 49

3.7 Posterior Probability Calculation for Bayesian Model Class Selection. ([a]ln Ockham, ln Likelihood, and ln Evidence are natural logarithms of the Ockham factor, likelihood, and evidence, respectively. [b]Probability is calculated from the evidence on the basis that the $\mathcal{M}_i$ $(i = 1, 2, 3)$ are equally probable a priori.) 51

3.8 Components of ln Ockham Factor in (3.11). 51

4.1 Definition of Site Classes. (G_B or G_C is 1 if a site is classified in class B or C, respectively, and 0 otherwise.) 56
4.2 Comparison Results for Parameter Estimation. 60
4.3 Earthquake Records Used. 62

5.1 Material Properties and Constants for a 3-D Frame Bridge. 71
5.2 Comparison between RVM and SVM for 3-story Building Example. . . 71
5.3 Comparison between RVM and SVM for 3-story Building Example. . . 73
5.4 Comparison between Scaled Fundamental Mode Shape and the First Ritz Vector Applied to Various SHM Examples by RVM. ("Errors" is misclassification rate.) 74
5.5 Identify Single Damage of (a) 20% and (b) 80% Stiffness Reduction in Each Story. (Bold numbers correspond to actual damage locations.) . . 79
5.6 Identify Damage of 20% Stiffness Reduction in Both x_1/x_2 Stories. (Bold numbers correspond to actual damage locations.) 79
5.7 Identify Damage in the $2^{nd}/3^{rd}$ Stories of r_2/r_3% Stiffness Reduction. (Bold numbers correspond to actual damage locations.) 79
5.8 Estimated Damage Severities. 80
5.9 Damage Cases and Patterns in Detail. 85
5.10 Extracted Modal Frequencies for Cases 1–5. 90
5.11 Stiffness Loss Predictions for Each Candidate Feature (Damage Case 1). 94
5.12 Stiffness Loss Predictions for Each Candidate Feature (Damage Case 2). 95
5.13 Stiffness Loss Predictions for Each Candidate Feature (Damage Case 3). 95
5.14 Identified Damage for Damage Case 4 and 5 using Damage Signature. . 101
5.15 Actual Stiffness Loss for Damage Case 5. (Damage Patterns 1 to 4 are also applied to Damage Case 4.) 104
5.16 Predicted Stiffness Loss for Damage Case 4. 105
5.17 Predicted Stiffness Loss for Damage Case 5. 106

Chapter 1

Introduction

1.1 Motivation

In recent years, there have been many advancements in sensor development, including digital, wireless sensor units and sensor networks, which enable the effective collection of a large amount of data useful for earthquake engineering as well as structural health monitoring. These new technologies have facilitated rapid acquisition of measurements which are useful for comprehensive assessment of post-disaster damage. They also provide invaluable information to improve the scientific understanding about the earthquake itself, as well as the dynamic responses of structures subject to earthquake shaking. The improvements in sensing and communication technology are also making it possible to install dense networks of sensors.

In parallel to these hardware developments, methodologies for the quantitative assessment of measured data have been developed, although much less attention has been paid to developing methodologies that can deal with all of the uncertainties involved and provide appropriate probabilistic predictions. Also, the existence of uncertainties due to measurement errors and especially modeling errors, and the lack of sufficient data, can cause inverse problems to be ill-conditioned.

In recent years, sophisticated data processing algorithms have been developed by computer scientists and statisticians working on statistical learning theory (also named "machine learning") such as Neural Networks, Support Vector Machine (SVM), Relevance Vector Machine (RVM), and so on. This machine learning is a subfield of

artificial intelligence and has the goal of using a computer to learn some relationship between input and output based on a training dataset. In particular, SVM is a recently-developed, powerful, state-of-the-art technique for regression and classification (Burges, 1998; Schölkopf and Smola, 2002; Vapnik, 1998).

There are two main types of machine learning problems: classification and regression. Classification can be defined as the act of identifying a separating boundary that separates different-class data in a feature space by using a given training dataset, and then deciding the category to which new data belongs by using that separating boundary. Regression is to infer a mathematical relationship between inputs and the corresponding outputs based on given dataset, which is usually done by prescribing a parameterized mathematical form and then estimating the parameters.

SVM is a machine learning algorithm that has been used as an efficient tool in bioinformatics, computer science, and ,to a much lesser extent, civil engineering. In classification, SVM determines the separating boundaries between classes by maximizing the *margin*, which is the distance between two different classes when the training data is mapped into the transformed feature space, while also minimizing the misclassification error. Similarly when using SVM in regression, the parameters of a pre-defined function are estimated with the consideration of regularization (this term has the same function form as the margin in classification), while also minimizing the error using a so-called ϵ-insensitive loss function. SVM has various advantages, such as (1) solving a *convex* optimization problem during training which guarantees a global optimum instead of a local one, (2) faster convergence in training than most other pattern recognition methods, for example, neural networks (Ding and Dubchak, 2001), and (3) efficient operation when predicting a result for new input data.

One disadvantage of SVM is that it is not a probabilistic method and so it does not explicitly quantify the uncertainties involved. A probabilistic treatment of learning from data and making predictions is recommended so that the uncertain errors caused by modeling and measurements can be explicitly addressed. This consideration motivated our use of Bayesian learning with an automatic relevance determination prior and an application of this, called the Relevance Vector Machine (RVM), which was

recently introduced as a Bayesian learning method using the same form of kernel basis functions as SVM (Tipping, 2000 and 2001; Tipping and Faul, 2003). It overcomes some disadvantages of SVM, such as:

(1) no explicit treatment of uncertainty in the predictions

(2) a relatively larger number of kernels required (increasing approximately linearly with the number of training data)

(3) waste of data and computational effort to estimate a trade-off parameter by cross-validation.

1.2 Bayesian Methodology

Most engineering problems can be divided into two main categories: forward problems and inverse problems. Forward problems compute the outputs using known mathematical models of systems for given inputs, while inverse problems infer mathematical models of unknown systems based on measured inputs and outputs. It is well known that the forward problem can usually be set up to be well-posed while the inverse problem is often inherently ill-posed so that a solution is non-unique, or may not even exist. Regularization theory was suggested to alleviate this ill-posedness in inverse problems by Tikhonov (Groetsch, 1984) and it has been proven to provide satisfactory results (e.g., Beck et al., 1985; Lee et al., 1999; Park et al., 2001). For non-Bayesian methods, another penalty term is usually added with a parameter to adjust its trade-off with datafit errors.

The Bayesian methodology provides logical and quantitative rules to treat inverse problems based on measurements or observations and any prior knowledge that is available.

Probabilistic inference is performed by using Bayes' theorem:

$$\mathrm{P}(hypothesis|data, I) \propto \mathrm{P}(data|hypothesis, I) \times \mathrm{P}(hypothesis|I)$$

$$\text{posterior} \quad \propto \quad \text{likelihood} \quad \times \quad \text{prior}$$

where I stands for background information. It was mentioned that (Sivia, 1996):

> *The power of Bayesian theorem lies in the fact that it relates the quantity of interest, the probability that the hypothesis is true given the data, to the term that we have a better chance of being able to assign, the probability that we would have observed the measured data if the hypothesis was true.*

In the Bayesian methodology, the prior can be used to provide regularization of ill-posed problems.

In this dissertation, it is demonstrated that Bayesian learning using an automatic relevance determination prior is effective for regularization against ill-conditioning and for avoiding over-fitting of data by using model class selection by applying the approach on earthquake engineering problems and on structural health monitoring.

1.3 Background on Structural Health Monitoring

Structural Health Monitoring (SHM) is the implementation of a damage detection and assessment strategy to aerospace, mechanical, and civil structures, based on monitoring sensor signals. The SHM process includes obtaining measured data on structural dynamic responses from an array of installed sensors over a certain period of time, extracting the damage sensitive features from the dataset and then analyzing the features to determine the current health states. The goal of SHM is to provide reliable information about the integrity of the structure after extreme events such as earthquakes, as well as investigating serviceability which might be impacted by aging or environmental effects (Brownjohn et al., 2004).

Two systematic ways to perform the damage assessment are by pattern classification or by regression based on the measured data, where features extracted from the measurements are used either to classify which damage state the structure is in (including the possibility that it is undamaged) or to predict damage locations and severity, respectively.

This pattern classification and regression belong to what is called *supervised learn-*

ing by computer scientists, in which the given training dataset consists of either the extracted features along with the corresponding labels for the damage states of interest (classification) or the measured inputs and outputs as continuous variables (regression). On the other hand, when the training dataset is given without any labels, it is called *unsupervised learning*. In this latter case, either the dataset is gathered into groups based on similarity between them, which is called clustering, or the probability distribution is determined by density estimation from the dataset. In SHM, unsupervised learning can be applied to a dataset not containing features from the damaged structure to identify the presence of the damage alone, while supervised learning uses data from an undamaged and possibly damaged structure to quantify the damage severities as well as to identify the locations of damage. Supervised learning methods are selected in this dissertation to obtain more comprehensive damage assessment results.

However, there is a difficulty in extracting a dataset from not only undamaged structures but also possibly damaged ones, since different damage scenarios can not be imposed on real structures of interest. To overcome this difficulty, model-based damage detection methods may be utilized to obtain damage sensitive features from a finite element (F.E.) model by selecting various possible damage scenarios for a structure under inspection, assuming that the F.E. model gives a good representation of the behavior of the real structure.

Needless to say, constructing a good structural model is essential in model-based methods and this can be performed by model-updating methods (e.g., Beck and Katafygiotis, 1998). The merit of model-based methods is that it is possible to obtain quantitative information about damage, such as damage indication, location, and severity. In this thesis, model-based supervised-learning damage assessment is viewed as consisting of four steps:

(1) Construct an updated baseline F.E. model of the real structure.

(2) Generate training data by imposing different damage patterns on the F.E. structural model and extracting the corresponding damage sensitive features.

(3) Apply the Bayesian learning method to this training data to develop an algorithm

to identify damage locations that can be applied to data from the real structure.
(4) Also use this Bayesian learning method to develop a procedure to estimate damage severities that can be applied to real data.

To extract appropriate features, usually dynamic characteristics such as estimated modal parameters are utilized, based on the basic premise that structural characteristics such as the stiffness, mass, or energy dissipation properties of a system are changed by damage, and these in turn alter the dynamic characteristics of the system.

1.3.1 Pattern Classification Applied to Structural Health Monitoring

Pattern classification is used to classify features based on information extracted from measurements or observations. The classified patterns are usually groups of features considered to have the same properties that are of interest.

There are three main phases in pattern classification: feature extraction, training, and prediction phases. The main objective of pattern classification is to determine the separating decision boundaries between data having different properties. These boundaries could be so complicated as to classify all training data completely with no misclassification of the data or so simple as to give many misclassifications. Since the main concern is to make accurate predictions of the label for new data which is not included in the training data, it is important to avoid either over-fitting of the data caused by an over-complicated decision boundary or under-fitting of the data by an over-simplified one. It will be shown in this work how Bayesian learning with an automatic relevance determination prior can achieve a decision boundary of optimal form for SHM applications.

1.3.2 Regression Procedure Applied to Structural Health Monitoring

A regression approach to SHM may be accomplished by utilizing a Bayesian learning methodology to update the probability distribution over the unknown model parameters of a regression model based on given data, using an automatic relevance determination prior to eliminate irrelevant terms. This regression approach undergoes the same three steps as pattern classification: feature extraction, training, and prediction phases. One advantage of a regression approach to SHM compared with classification is that it does not require as much computational effort to generate a large amount of training data for different damage scenarios, since unlike classification, it is unnecessary to cover for each possible damage location and severity.

The main objective in SHM using a regression approach is to estimate the most plausible regression model which relates the input features and the corresponding output (damage locations and severities) based on the training dataset, and then to perform a damage assessment using the regression model with new data (features extracted in real-time from sensor signals).

1.4 Organization

This dissertation presents the detailed mathematical background and procedures for newly-developed Bayesian learning methods that use an automatic relevance determination prior, including the Relevance Vector Machine. These methods are then applied to some earthquake engineering problems and to structural health monitoring. For SHM, an extended version of RVM is presented to effectively deal with vector outputs. The capabilities of the proposed methodology for each application are demonstrated.

Chapter 2 describes the detailed mathematical procedures that are utilized for regression and classification problems.

Chapter 3 presents an application of the proposed method using an appropriate

dataset for earthquake early warning, so that incoming seismic signals can be automatically classified in real time as near-source or far-source. Chapter 4 deals with the estimation of earthquake ground motion attenuation equations using a strong-motion database.

In Chapter 5, structural health monitoring applications are presented. For this, an enhanced algorithm is presented and applied to illustrative examples, including the IASC-ASCE SHM benchmarks.

Chapter 2

Bayesian Learning Using Automatic Relevance Determination Prior and Relevance Vector Machine

Bayesian Learning is performed in three phases in both regression and classification:

- Phase I (Feature Extraction Phase): This phase distills a small number of regression variables or features from a large set of data that are thought to explain or help predict quantities of interest in case of regression, or characterize each class of interest in the data in case of classification.
- Phase II (Training Phase): This phase identifies a mathematical relationship between input regression variables or features and quantities of interest in regression, or a separating boundary based on extracted features for classification, usually using some form of regularization during the identification.
- Phase III (Prediction Phase): In this phase, a prediction is made using the estimated regression equation or separating boundary from the previous phase to decide what is the expected response corresponding to new data in regression ,or which is the appropriate class for new data in classification.

Bayesian methods for regression and classification problems have the advantage that they make probabilistic predictions for the responses corresponding to inputs or

for the class that corresponds to a given feature vector (rather than giving only a point estimate in regression or a possibly misleading yes/no answer in classification). These predictions are based on a rigorous Bayesian learning procedure that rests on the axioms of probability. The essential ingredients are a set of predictive probability models involving a parameterized regression function or separating boundary function, and a probability model (the prior) over this set. The prior can be pragmatically chosen by the user to regularize the ill-conditioned problem of identifying a pre-defined mathematical form of regression equation or a boundary that separates the classes in the feature vector space. In the absence of such regularization, the training phase will usually lead to "over-fitting" of the data, so that the generalization beyond the training data in the prediction phase will perform poorly.

In this chapter, the mathematical procedure for the novel methodology of *Bayesian learning using automatic relevance determination (ARD) prior* and the special case of the Relevance Vector Machine (RVM) is introduced, and the extension of RVM to vector-valued outputs is investigated.

2.1 Bayesian Learning Method Using Various Priors

In this section, the advantages of a Bayesian learning methodology using various priors are presented and compared with those of some non-Bayesian methods. For simplicity, the focus in this introductory section is on regression problems involving the prediction of some scalar quantity y.

2.1.1 Least-Squares Estimation using Linear Model

We start with the familiar linear (with respect to its parameters) regression equation of the form:

$$f(\underline{x}|\underline{\theta}) = \underline{\tau}(\underline{x})^T \underline{\theta} = \sum_{j=1}^{m} \theta_j \cdot g_j(\underline{x}) + \theta_0 \tag{2.1}$$

where $\underline{x} \in \mathbb{R}^n$ is the selected input (regression variables), $\underline{\tau}(\underline{x}) \in \mathbb{R}^{m+1} = [1,\ \underline{g}(\underline{x})^T]^T$, $\underline{g}(\underline{x})$ is the vector of chosen linear or nonlinear basis functions, and the unknown parameters $\underline{\theta} \in \mathbf{\Theta} \subset \mathbb{R}^{m+1}$ define a specific predictive model within a set of possible models defined by $\mathbf{\Theta}$.

Let $\mathcal{D}_N = \{(\underline{x}_i, y_i) : i = 1, ..., N\} = (\mathbf{X}, \underline{y})$ denote the data for identifying the model. Then, least-squares estimation (LSE) is performed by minimizing the sum of squares error with respect to $\underline{\theta} \in \mathbf{\Theta}$:

$$E_1(\underline{\theta}) = \|\underline{y} - \underline{f}(\mathbf{X}|\underline{\theta})\|^2 = \sum_{i=1}^{N}(y_i - f(\underline{x}_i|\underline{\theta}))^2 \tag{2.2}$$

to get an estimate $\hat{\underline{\theta}}$ of the parameter vector.

In case of a linear regression model, an analytical solution can be obtained by taking first and second derivatives of (2.2). For a nonlinear model, however, it is not usually possible or feasible to obtain an analytical solution. In such a situation, the solution must be sought numerically using a nonlinear optimization algorithm. It is, however, well-known that least-squares estimation alone may lead to poor generalization due to over-fitting of given data (i.e., the prediction error for new data is poor). Furthermore, many least-squares problems are ill-conditioned, that is, there are many least-squares solutions, or at least many near-optimal solutions.

For a non-Bayesian approach, an ill-conditioned problem can be regularized by adding a term:

$$E_2(\underline{\theta}) = \|\underline{\theta}\|^2 = \sum_{i=0}^{m} \theta_i^2 \tag{2.3}$$

to $E_1(\underline{\theta})$ in (2.2) that penalizes large values of the θ_i's. This penalty term is suggested by Tikhonov to alleviate ill-conditioning in inverse problems (Groetsch, 1984). Then, the unknown parameter vector $\underline{\theta}$ is estimated by minimizing the objective function:

$$E(\underline{\theta}) = C_1 E_1(\underline{\theta}) + C_2 E_2(\underline{\theta}) \tag{2.4}$$

where $C_1(> 0)$ and $C_2(> 0)$ allow a trade-off between the fit to the data and the size

of the θ_i's.

The cost function is usually defined as:

$$E(\underline{\theta}) = \frac{1}{2}E_1(\underline{\theta}) + \frac{\lambda}{2}E_2(\underline{\theta}) \tag{2.5}$$

The trade-off parameter λ in (2.5) is called as "Regularization Factor", which controls the complexity of the regression equation and avoids over-fitting; as $\lambda \to 0$, the smaller regularization effect causes oscillations of the regression equation by ill-conditioning, but on the other hand, the regression equation becomes flat as $\lambda \to \infty$. Intermediate values of λ control the smoothness of the regression function.

The Support Vector Machine (SVM), one of the state-of-the-art machine learning techniques, adopts this penalty term $E_2(\underline{\theta})$ with $C_1 = C$(trade-off parameter) and $C_2 = \frac{1}{2}$; C is determined by cross-validation (Vapnik, 1998). In 10-fold cross-validation, for example, the training dataset is divided into 10 subsets of (approximately) equal size and the algorithm is trained 10 times, each time leaving out one of the subsets from training and then using the omitted subset to determine the optimal value of C which satisfies a certain prediction-error criterion, for example, minimizing the sum of prediction errors by using the omitted subset as the prediction dataset. This procedure, however, means that the corresponding algorithm requires large amount of computational effort during the training phase.

As an another way to avoid over-fitting in non-Bayesian method, one can use a different penalty term; for example, proportional to m, the number of adjustable parameters in the model, as in Akaike's Information Criterion (AIC) (Akaike, 1974). AIC selects the model with the lowest AIC value which gives a balance between the fit to the data and the number of parameters.

2.1.2 Bayesian Inference

Unlike LSE, Bayesian inference gives a probabilistic description of the unknown parameters rather than a point estimate. Additionally, it provides a unifying framework to control model complexity, i.e., to overcome the over-fitting problem, and it allows

the modeling uncertainties for the parameters to be integrated out when making predictions by using marginalization, i.e., by taking all likely values into consideration during integration.

To account for the fact that no model gives prefect predictions, the regression equation is embedded in a probability model by introducing an uncertain prediction error ϵ, so that the quantity y to be predicted using $\underline{x} \in \mathbb{R}^n$ is given by

$$y = f(\underline{x}|\underline{\theta}) + \sigma\epsilon \tag{2.6}$$

where ϵ is a Gaussian variable with zero mean and unit variance. This choice of the probability model for the prediction error is motivated by Jaynes' Principle of Maximum Entropy (Jaynes, 2003). Equation (2.6) defines a probability model

$$p(y|\underline{x}, \underline{\theta}, \sigma^2) = \mathcal{N}(y|f(\underline{x}|\underline{\theta}), \sigma^2) \tag{2.7}$$

i.e., y is conditionally Gaussian with mean $f(\underline{x}|\underline{\theta})$ and variance σ^2.

Bayesian inference using Bayes' theorem gives the posterior for the unknown parameter $\underline{\theta}$ in terms of the likelihood and prior:

$$p(\underline{\theta}, \sigma^2|\mathcal{D}_N) = \frac{p(\mathcal{D}_N|\underline{\theta}, \sigma^2)p(\underline{\theta})p(\sigma^2)}{p(\mathcal{D}_N)} \propto p(\mathcal{D}_N|\underline{\theta}, \sigma^2)p(\underline{\theta})p(\sigma^2) \tag{2.8}$$

The likelihood $p(\mathcal{D}_N|\underline{\theta}, \sigma^2)$ in (2.8) measures how well the parameters $\underline{\theta}$ and σ^2 predict the observed data $\mathcal{D}_N$ and it can be expressed as a Gaussian distribution based on the probability model in (2.7):

$$\begin{aligned} p(\mathcal{D}_N|\underline{\theta}, \sigma^2) &= \prod_{i=1}^{N} p(y_i|\underline{x}_i, \underline{\theta}, \sigma^2) \\ &= (2\pi\sigma^2)^{-N/2} \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f(\underline{x}_i|\underline{\theta}))^2\Big] \end{aligned} \tag{2.9}$$

The prior $p(\underline{\theta})$ can be chosen to reflect the analyst's uncertainty about the value of parameter $\underline{\theta}$ and a reasonably flexible choice is a Gaussian distribution parameterized

by $\underline{\alpha}$ as follows:

$$p(\underline{\theta}|\underline{\alpha}) = \mathcal{N}(\underline{\theta}|\underline{0}, \mathbf{A}^{-1}) = (2\pi)^{-(m+1)/2}|\mathbf{A}|^{1/2}\exp\Big[-\frac{1}{2}\underline{\theta}^T\mathbf{A}\underline{\theta}\Big] \tag{2.10}$$

where $\mathbf{A} = \text{diag}(\alpha_0, \alpha_1, ..., \alpha_m) \in \mathbb{R}^{(m+1)\times(m+1)}$ and $\underline{\alpha}$ is called a hyperparameter, since it parameterizes the prior distribution of the parameters. The Bayesian approach with this Gaussian prior can be categorized into three cases according to the constraints on the values of the hyperparameter vector $\underline{\alpha}$:

C1. Non-informative prior where all $\alpha_i \to 0$; in this case, if the focus is only on the most probable values of $\underline{\theta}$ and σ^2, it is equivalent to Maximum Likelihood Estimation (MLE);

C2. Gaussian prior with the same variance for each parameter, i.e., $\alpha_i = \alpha$ for $i = 0, ..., m$, leading to a regularization method known as *Ridge Regression*;

C3. Automatic Relevance Determination (ARD) prior with all α_i independent.

When the number of parameter becomes large, i.e., m increases, C_1 gives an ill-conditioned problem, C_2 introduces regularization to give better conditioning, but C_3 not only allows regularization but also controls model complexity by automatically selecting the relevant regression terms to give sparsity; i.e., ARD prior selects only a small number of relevant basis expansion terms by automatically pruning others (Mackay, 1994; Tipping, 2000).

C1: Non-informative Prior

The non-informative or uniform prior treats all possible parameter values as being equally plausible a priori. Therefore, the posterior state of knowledge is influenced only by the data through the likelihood. If only the most probable values of the parameters are examined, they are given by maximizing the likelihood to give the MLE values $\hat{\underline{\theta}}$ and $\hat{\sigma}$. This implies that $\hat{\underline{\theta}} = \arg\min E_1(\underline{\theta})$ where $E_1(\underline{\theta})$ is defined in (2.2) and $\hat{\sigma}^2 = \frac{1}{N}E_1(\hat{\underline{\theta}})$. Actually, MLE gives the same solution as LSE when

the non-informative prior is utilized with the Gaussian prediction error ϵ, which can lead to ill-conditioning. It is noted that classical MLE is not equivalent to the full Bayesian approach with a uniform prior because the latter gives a posterior PDF which shows how plausible each of the possible parameter values are, which is characteristic of the Bayesian approach; this allows, for example, the modeling uncertainty for the parameters to be integrated out when making predictions. This procedure of marginalization is presented later.

C2: Gaussian Prior with Equal Variances

If we assume the same variance for all parameters θ_i (i.e., $\alpha_0 = \alpha_1 = ... = \alpha_m = \alpha$), the resulting Gaussian prior has a form

$$p(\underline{\theta}|\alpha) = \mathcal{N}(\underline{\theta}|\underline{0}, \alpha^{-1}) = \Big(\frac{\alpha}{2\pi}\Big)^{(m+1)/2} \exp\Big[-\frac{\alpha}{2}\sum_{i=0}^{m}\theta_i^2\Big] \tag{2.11}$$

Also the posterior PDF $p(\underline{\theta}|\mathcal{D}_N, \sigma^2, \underline{\alpha})$ is given by Bayes' theorem:

$$p(\underline{\theta}|\mathcal{D}_N, \sigma^2, \underline{\alpha}) = \frac{p(\mathcal{D}_N|\underline{\theta}, \sigma^2)p(\underline{\theta}|\underline{\alpha})}{p(\mathcal{D}_N|\sigma^2, \underline{\alpha})} \propto p(\mathcal{D}_N|\underline{\theta}, \sigma^2)p(\underline{\theta}|\underline{\alpha}) \tag{2.12}$$

Thus, taking the natural log of (2.9), (2.11), and (2.12) (ignoring additive constant terms) leads to:

$$\begin{aligned} \ln p(\underline{\theta}|\mathcal{D}_N, \sigma^2, \underline{\alpha}) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f(\underline{x}_i|\underline{\theta}))^2 - \frac{\alpha}{2}\sum_{i=0}^{m}\theta_i^2 \\ &= -\frac{1}{2\sigma^2}E_1(\underline{\theta}) - \frac{\alpha}{2}E_2(\underline{\theta}) \end{aligned} \tag{2.13}$$

Therefore, for σ^2 fixed, the objective function in regularized LSE, $E(\underline{\theta})$ in (2.4), corresponds to the log of the posterior PDF $p(\underline{\theta}|\mathcal{D}_N, \sigma^2, \alpha)$, and so the regularized least-squares estimates correspond to the most probable values of $\underline{\theta}$, given σ^2 and α. However, in the Bayesian approach, both σ^2 and α can also be estimated directly from the data by maximizing the evidence $p(\mathcal{D}_N|\sigma^2, \alpha)$, as shown in Section 2.2.

The Gaussian prior in (2.11) controls ill-conditioning by automatically providing

regularization when estimating the unknown parameters through the penalty term $E_2(\underline{\theta})$ in (2.13). All the available dataset can be devoted for training since the trade-off parameters σ^2 and α can be estimated directly without using cross-validation by using model class selection (Beck and Yuen, 2004).

This concept of model class selection is consistent with *Ockham's razor*, which is the principle of preferring a simpler model unless there are compelling reasons for a more complex model. Suppose there are two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for comparison and the relative plausibility between these two models needs to be estimated in the light of data. Bayes' Theorem yields the relative probability between the two models:

$$\frac{P(\mathcal{M}_1|\mathcal{D}_N)}{P(\mathcal{M}_2|\mathcal{D}_N)} = \frac{p(\mathcal{D}_N|\mathcal{M}_1)P(\mathcal{M}_1)}{p(\mathcal{D}_N|\mathcal{M}_2)P(\mathcal{M}_2)} \tag{2.14}$$

If the two models are treated as equally plausible a priori, then the second ratio is unity and the first ratio embodies Ockham's razor and is called the Ockham factor. Since a simpler model $\mathcal{M}_1$ tends to make a more precise prediction, while a complex model $\mathcal{M}_2$ is capable of making a great variety of predictions, the situation is shown schematically in Figure 2.1 which shows how $\mathcal{M}_1$ is more probable if the data is very likely based on $\mathcal{M}_1$ (large Ockham factor in (2.14)) whereas $\mathcal{M}_2$ is only more probable if the data is very unlikely based on $\mathcal{M}_1$ (small Ockham factor in (2.14)) (Mackay, 1992b, 1995). Now consider σ_1^2 and α_1 in (2.13) as defining model class $\mathcal{M}_1$ and σ_2^2 and α_2 defining $\mathcal{M}_2$, then the most probable model class can be selected using (2.14) where $p(\mathcal{D}_N|\mathcal{M}_i) = p(\mathcal{D}_N|\sigma_i^2, \alpha_i) = \int p(\mathcal{D}_N|\underline{\theta}, \sigma_i^2)p(\underline{\theta}|\alpha_i)d\underline{\theta}$. In general, the parameters σ^2 and α can be chosen to maximize $p(\mathcal{D}_N|\sigma^2, \alpha)$.

C3: Automatic Relevance Determination Prior

In regression problems, some of the input variables will have a strong influence on the prediction of the output variable while others may be irrelevant. Automatic Relevance Determination (ARD) prior automatically suppresses irrelevant input variables by pruning them out. This is done by introducing an independent variance α_i^{-1} for each

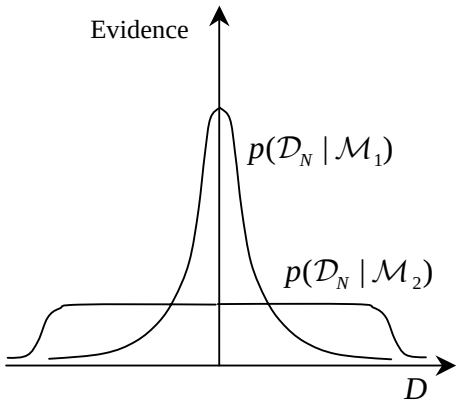


Figure 2.1: Illustration of Ockham's Razor (Mackay, 1992b).

parameter θ_i in the prior. The resulting prior has the same form as (2.10):

$$p(\underline{\theta}|\underline{\alpha}) = \mathcal{N}(\underline{\theta}|\underline{0}, \mathbf{A}^{-1}) = (2\pi)^{-(m+1)/2}|\mathbf{A}|^{1/2} \exp\Big[-\frac{1}{2}\underline{\theta}^T \mathbf{A}\underline{\theta}\Big]$$

As shown in the remaining sections of Chapter 2, this one-to-one correspondence of the hyperparameter vector $\underline{\alpha}$ to the parameter vector $\underline{\theta}$ makes Bayesian learning using ARD prior effective in practice in yielding sparsity (i.e., utilizing only a small number of input or, equivalently, relevant basis expansion terms, by automatically pruning others; this occurs because during estimation, some $\alpha_i \to \infty \Rightarrow \theta_i \to 0$, which results in the irrelevant input terms being pruned). In the next two sections, the focus is on this ARD prior and its use in regression and classification problems.

2.2 Bayesian Learning Using Automatic Relevance Determination Prior: I–Regression

In this section, the detailed mathematical procedure of Bayesian learning for regression using the ARD prior is presented. The classification counterpart will be presented in Section 2.3.

2.2.1 Training Phase in Regression

Let the most probable value of the output vector $y \in \mathbb{R}$ be related to an input vector (predictor variables) $\underline{x} \in \mathbb{R}^n$ by a chosen regression function $f(\underline{x}|\underline{\theta})$ when the model parameter vector $\underline{\theta}$ is specified. This function is embedded in a probability model by introducing an uncertain prediction error to account for the fact that no model gives prefect predictions, so:

$$y = f(\underline{x}|\underline{\theta}) + \sigma\epsilon \tag{2.15}$$

which defines the probability model:

$$p(y|\underline{x}, \underline{\theta}, \sigma^2) = \mathcal{N}(y|f(\underline{x}|\underline{\theta}), \sigma^2) \tag{2.16}$$

where $f(\underline{x}|\underline{\theta})$ is defined as a linear combination of the $g_j(\underline{x})$, j^{th} component of $\underline{g}(\underline{x})$ with a parameter θ_j so that:

$$f(\underline{x}|\underline{\theta}) = \underline{\tau}(\underline{x})^T\underline{\theta} = \sum_{j=1}^{m} \theta_j \cdot g_j(\underline{x}) + \theta_0 \tag{2.17}$$

where $\underline{\tau}(\underline{x}) = [1,\ \underline{g}(\underline{x})^T]^T$ and each $g_i(\underline{x})$ is a linear or nonlinear basis function.

Let $\mathcal{D}_N = \{(\underline{x}_i, y_i) : i = 1, ..., N\} = (\mathbf{X}, \underline{y})$ denote the data with input (predictor variables) $\underline{x}_i \in \mathbb{R}^n$. Based on a Gaussian white-noise model for the prediction errors ϵ_i (which is a maximum entropy probability distribution for given mean and variance (Jaynes, 2003)), the ϵ_i are modelled independently and identically distributed as $\mathcal{N}(\underline{0}, \sigma^2)$. Thus, the likelihood for the given dataset is

$$\begin{aligned} p(\mathcal{D}_N|\underline{\theta}, \sigma^2) &= \prod_{i=1}^{N} p(y_i|\underline{x}_i, \underline{\theta}, \sigma^2) \\ &= \mathcal{N}(\mathbf{\Phi}\underline{\theta}, \sigma^2\mathbf{I}) \\ &= (2\pi\sigma^2)^{-N/2} \exp\Big[-\frac{1}{2\sigma^2}\|\underline{y} - \mathbf{\Phi}\underline{\theta}\|^2 \Big] \end{aligned} \tag{2.18}$$

where $\mathbf{\Phi} = [\underline{\tau}(\underline{x}_1), ..., \underline{\tau}(\underline{x}_N)]^T \in \mathbb{R}^{N\times(m+1)}$, and $\underline{\theta} = [\theta_0, \theta_1, ..., \theta_m]^T$.

Define the ARD prior PDF for $\underline{\theta}$ to be

$$\begin{aligned} p(\underline{\theta}|\underline{\alpha},\sigma^2) &= \mathcal{N}(\underline{0},\mathbf{A}^{-1}(\underline{\alpha})) \\ &= (2\pi)^{-\frac{m+1}{2}}|\mathbf{A}(\underline{\alpha})|^{\frac{1}{2}}\exp\Big\{-\frac{1}{2}\underline{\theta}^T\mathbf{A}(\underline{\alpha})\underline{\theta}\Big\} \end{aligned} \tag{2.19}$$

i.e., $\underline{\theta}$ is Gaussian with mean $\underline{0}$ and covariance matrix $\mathbf{A}^{-1}(\underline{\alpha})$, as before, and independent of the prediction-error variance σ^2, which is taken as fixed here; its estimation is discussed later in this section.

Then the posterior PDF for the unknown parameters $\underline{\theta}$ can be calculated via Bayes' Theorem by

$$\begin{aligned} p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha},\sigma^2) &= \frac{p(\mathcal{D}_N|\underline{\theta},\underline{\alpha},\sigma^2)p(\underline{\theta}|\underline{\alpha},\sigma^2)}{p(\mathcal{D}_N|\underline{\alpha},\sigma^2)} \\ &= (2\pi)^{-\frac{m+1}{2}}|\hat{\mathbf{\Sigma}}|^{-1/2}\exp\Big[-\frac{1}{2}(\underline{\theta}-\hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta}-\hat{\underline{\theta}})\Big] \end{aligned} \tag{2.20}$$

where $\hat{\mathbf{\Sigma}} = (\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}+\mathbf{A})^{-1}$ and $\hat{\underline{\theta}} = \sigma^{-2}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\underline{y}$.

In the next step, Bayesian model class selection is used to select the most plausible hyperparameter $\underline{\alpha} \in \mathcal{A}$ and σ^2 (e.g., Beck and Yuen, 2004). The model class $\mathcal{M}$ is defined as a mathematical structure assumed for $p(y|\underline{x},\underline{\theta},\sigma^2)$ *and* the prior PDF $p(\underline{\theta}|\sigma^2,\alpha)$. The most probable model class $\mathcal{M}(\hat{\underline{\alpha}},\hat{\sigma}^2)$ based on data $\mathcal{D}_N$ is given by finding $\hat{\underline{\alpha}}$ and $\hat{\sigma}^2$ that maximizes $p(\underline{\alpha},\sigma^2|\mathcal{D}_N) \propto p(\mathcal{D}_N|\underline{\alpha},\sigma^2)p(\underline{\alpha},\sigma^2)$. If a uniform prior on $\underline{\alpha}$ and σ^2 is considered, then it is equivalent to the maximization of the evidence for $\underline{\alpha}$ and σ^2, $p(\mathcal{D}_N|\underline{\alpha},\sigma^2)$, which is equivalent to the maximization of $\ln p(\mathcal{D}_N|\underline{\alpha},\sigma^2)$ given by:

$$\begin{aligned} \mathcal{L}(\underline{\alpha},\sigma^2) &= \ln p(\mathcal{D}_N|\underline{\alpha},\sigma^2) = \ln\int_{\mathbb{R}^{m+1}} p(\mathcal{D}_N|\underline{\theta},\sigma^2)p(\underline{\theta}|\underline{\alpha})d\underline{\theta} \\ &= -\frac{1}{2}\Big[N\ln 2\pi + \ln|\sigma^2\mathbf{I}+\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T| + \underline{y}^T(\sigma^2\mathbf{I}+\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T)^{-1}\underline{y}\Big] \\ &= -\frac{1}{2}\Big[N\ln 2\pi + \ln|\mathbf{C}| + \underline{y}^T\mathbf{C}^{-1}\underline{y}\Big] \end{aligned} \tag{2.21}$$

where $\mathbf{C}(\underline{\alpha},\sigma^2) = \sigma^2\mathbf{I}+\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T$ and $\mathbf{A}(\underline{\alpha})$ is defined as before. Details of this

derivation are given in Appendix A.

The maximization of $\mathcal{L}(\underline{\alpha}, \sigma^2)$ is performed using an iterative procedure as follows. To determine α_i given the latest values of the other α_j's ($j \neq i$), $\mathbf{C}$ can be re-written as the sum of two terms, one that is related to α_i and another that is not, as follows:

$$\begin{aligned} \mathbf{C} &= \sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T = \sigma^2\mathbf{I} + \sum_m \alpha_m^{-1}\underline{\tau}_m\underline{\tau}_m^T \\ &= \sigma^2\mathbf{I} + \sum_{m\neq i} \alpha_m^{-1}\underline{\tau}_m\underline{\tau}_m^T + \alpha_i^{-1}\underline{\tau}_i\underline{\tau}_i^T = \mathbf{C}_{-i} + \alpha_i^{-1}\underline{\tau}_i\underline{\tau}_i^T \end{aligned} \tag{2.22}$$

where $\mathbf{C}_{-i}$ is the covariance matrix $\mathbf{C}$ with the components of $\underline{\tau}_i = \underline{\tau}(\underline{x}_i)$ removed.

By using the matrix determinant and inverse identities,

$$|\mathbf{C}| = |\mathbf{C}_{-i}||1 + \alpha_i^{-1}\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i| \tag{2.23}$$

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\underline{\tau}_i\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i} \tag{2.24}$$

$\mathcal{L}(\underline{\alpha}, \sigma^2)$ becomes

$$\begin{aligned} \mathcal{L}(\underline{\alpha}, \sigma^2) &= -\frac{1}{2}\Big[N\ln 2\pi + \ln|\mathbf{C}_{-i}| + \underline{y}^T\mathbf{C}_{-i}^{-1}\underline{y} \\ &\quad - \ln\alpha_i + \ln(\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i) - \frac{(\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{y})^2}{\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i}\Big] \\ &= \mathcal{L}(\alpha_{-i}, \sigma^2) + \frac{1}{2}\Big[\ln\alpha_i - \ln(\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i) + \frac{(\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{y})^2}{\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i}\Big] \\ &= \mathcal{L}(\alpha_{-i}, \sigma^2) + l(\alpha_i, \sigma^2) \end{aligned} \tag{2.25}$$

Therefore, the term related with α_i is isolated in $l(\alpha_i, \sigma^2)$.

By setting the partial derivative of (2.25) with respect to α_i to zero, the value that maximizes $\mathcal{L}(\underline{\alpha}, \sigma^2)$ is found. This Bayesian model class selection procedure gives:

$$\frac{\partial\mathcal{L}(\underline{\alpha}, \sigma^2)}{\partial\alpha_i} = \frac{1}{\alpha_i} - \frac{1}{\alpha_i + S_i} - \frac{Q_i^2}{\alpha_i + S_i} = 0 \tag{2.26}$$

which leads to:

$$\hat{\alpha}_i = \begin{cases} \infty, & \text{if } Q_i^2 \leq S_i \\ \frac{S_i^2}{Q_i^2 - S_i}, & \text{if } Q_i^2 > S_i \end{cases} \tag{2.27}$$

where $Q_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{y}$ and $S_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{\tau}_i$ with $\mathbf{C} = \sigma^2 \mathbf{I} + \mathbf{\Phi} \mathbf{A}^{-1} \mathbf{\Phi}^T$. In practice, many of the α_is approach infinity during this iterative process, and so the corresponding θ_is approach to zero, thereby pruning the corresponding term of $\underline{\tau}(\underline{x})$ from the regression function. In the end, only the terms of $\underline{\tau}(\underline{x})$, corresponding to a small number of the finite α_is, are retained. The detailed procedure of this Bayesian model class selection is described in Appendix C.

To summarize: starting with an initial estimate of $\underline{\hat{\alpha}}$, $\hat{\alpha}_i$ is iteratively calculated from (2.27) for each $i = 1, ..., N$, always utilizing the latest estimates for the α_j to evaluate $\mathbf{C}(\underline{\alpha})$, and this procedure is continued until it converges to $\underline{\hat{\alpha}}$. For those $\hat{\alpha}_i$ that approach infinity, there is a pruning of the corresponding basis function vectors $\tau_i(\underline{x})$ since $\hat{\alpha}_i \to \infty \Rightarrow \hat{\theta}_i \to 0$. Thus, only the basis function term $\tau_j(\underline{x})$ that have $\hat{\alpha}_j$ finite are used in determining the regression equation.

It is shown in Appendix C.2 that the noise variance σ^2 can be re-estimated by:

$$\hat{\sigma}^2 = \frac{\|\underline{y} - \mathbf{\Phi} \underline{\hat{\theta}}\|^2}{N - m + \sum_i \hat{\alpha}_i \hat{\Sigma}_{ii}} \tag{2.28}$$

after each new iteration of the $\underline{\alpha}$ optimization.

2.2.2 Prediction Phase in Regression

Based on the results from the previous subsection, prediction is performed as follows. We want to probabilistically estimate a new and unknown output $\tilde{y}$ based on a given input variable $\underline{\tilde{x}}$ and the dataset for training $\mathcal{D}_N$. The desired probability distribution is given by marginalization followed by the product rule (or, equivalently, the Theorem

of Total Probability):

$$\begin{aligned} p(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N) &= \int p(\tilde{y}, \underline{\theta}, \underline{\alpha}, \sigma^2|\underline{\tilde{x}}, \mathcal{D}_N) d\underline{\theta} d\underline{\alpha} d\sigma^2 \\ &= \int p(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N, \underline{\theta}, \sigma^2) p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}, \sigma^2) p(\underline{\alpha}, \sigma^2|\mathcal{D}_N) d\underline{\theta} d\underline{\alpha} d\sigma^2 \end{aligned} \quad (2.29)$$

This robust predictive PDF takes into account all possible uncertainties for the parameters $\underline{\theta}$, $\underline{\alpha}$, and σ^2. This marginalization to obtain the robust predictive PDF is a great advantage of the Bayesian approach over other methods since it means no information is lost during parameter estimation. Using Laplace's asymptotic approximation for large N (Beck and Katafygiotis, 1998; See Appendix B for details):

$$p(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N) \cong \int p(\tilde{y}|\underline{\tilde{x}}, \underline{\theta}, \hat{\sigma}^2) p(\underline{\theta}|\mathcal{D}_N, \underline{\hat{\alpha}}, \hat{\sigma}^2) d\underline{\theta} = \mathcal{N}(\tilde{y}|y_*, \sigma_*^2) \quad (2.30)$$

where $\underline{\hat{\alpha}}$, $\hat{\sigma}^2$ are the most probable values for $\underline{\alpha}$, σ^2, respectively, based on data $\mathcal{D}_N$, derived as in Section 2.2.1, $y_* = \underline{\hat{\theta}}^T \tau(\underline{\tilde{x}})$ and $\sigma_*^2 = \hat{\sigma}^2 + \tau(\underline{\tilde{x}})^T \hat{\mathbf{\Sigma}} \tau(\underline{\tilde{x}})$. In (2.30) $p(\underline{\theta}|\mathcal{D}_N, \underline{\hat{\alpha}}, \hat{\sigma}^2)$ is the posterior distribution given in (2.20).

2.3 Bayesian Learning Using Automatic Relevance Determination Prior: II–Classification

2.3.1 Training Phase in Classification

In classification, the data $\mathcal{D}_N = \{(\underline{x}_i, y_i) : i = 1, ..., N\} = (\mathbf{X}, \underline{y})$ consists of features (predictor variables) $\underline{x}_i \in \mathbb{R}^n$ and labels $y_i \in \{0, 1\}$ (for example, $y_i = 0$ for the data in one class, $y_i = 1$ for the data in another class).

Suppose that the function characterizing the separating boundary between the two classes is taken as a linear combination of some prescribed basis functions $g_j(\underline{x})$ with unknown coefficients $\underline{\theta} = \{\theta_0,\ \theta_1,\ ...,\ \theta_m\} \in \mathbb{R}^{m+1}$:

$$f(\underline{x}|\underline{\theta}) = \underline{\tau}(\underline{x})^T \underline{\theta} = \sum_{j=1}^{m} \theta_j g_j(\underline{x}) + \theta_0 \quad (2.31)$$

where $\underline{\tau}(\underline{x}) = [1,\ \underline{g}(\underline{x})^T]^T \in \mathbb{R}^{m+1}$ is the vector of chosen linear or non-linear basis function $\underline{g}(\underline{x})$ of features $\underline{x} = \{x_1, ..., x_m\}^T$. The separating boundary function $f(\underline{x}|\underline{\theta})$ is also called the discriminant function. For a known parameter vector $\underline{\theta}$, the separating boundary between the different classes is defined as $f(\underline{x}|\underline{\theta}) = 0$, and probabilistic predictions of the class label $y \in \{0, 1\}$ corresponding to extracted features $\underline{x}$ will be based on the probability model:

$$\begin{aligned} P(y|\underline{x}, \underline{\theta}) &= \begin{cases} \phi(f(\underline{x}|\underline{\theta})), & \text{if } y = 1 \\ 1 - \phi(f(\underline{x}|\underline{\theta})), & \text{if } y = 0 \end{cases} \\ &= \phi(f(\underline{x}|\underline{\theta}))^y \{1 - \phi(f(\underline{x}|\underline{\theta}))\}^{1-y} \end{aligned} \tag{2.32}$$

where $\phi(\cdot) \in [0, 1]$ is the monotonically increasing sigmoid function on $\mathbb{R}$ defined by $\phi(x) = \frac{1}{1+e^{-x}}$ so $\lim_{x\to\infty} \phi(x) = 1$, $\lim_{x\to-\infty} \phi(x) = 0$, and $\phi(x) + \phi(-x) = 1$ (Figure 2.2). Thus, when $f(\underline{x}|\underline{\theta})$ is large and positive (respectively, negative), the probability is near 1 that $\underline{x}$ corresponds to an instance of the $y = 1$ (respectively, $y = 0$) class. Of course, since (2.31) is just a model for the separating boundary, there are no true values of $\underline{\theta}$ to be "estimated", but we can learn about how plausible its various values are by Bayesian updating using the data $\mathcal{D}_N$.

Based on the predictive probability model (2.32), the likelihood $P(\mathcal{D}_N|\underline{\theta})$ is:

$$\begin{aligned} P(\mathcal{D}_N|\underline{\theta}) &= \prod_{i=1}^{N} P(y_i|\underline{x}_i, \underline{\theta}) \\ &= \prod_{i=1}^{N} \phi(f(\underline{x}_i|\underline{\theta}))^{y_i} \{1 - \phi(f(\underline{x}_i|\underline{\theta}))\}^{1-y_i} \end{aligned} \tag{2.33}$$

which measures how well the predictive probability model defined by $\underline{\theta}$ predicts the actual data.

The ARD prior $p(\underline{\theta}|\underline{\alpha})$ is defined identically with the regression case:

$$p(\underline{\theta}|\underline{\alpha}) = (2\pi)^{-\frac{m+1}{2}} |\mathbf{A}(\underline{\alpha})|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\underline{\theta}^T \mathbf{A}(\underline{\alpha})\underline{\theta} \Big\} \tag{2.34}$$

which provides a means of regularizing and encouraging sparsity during the learning

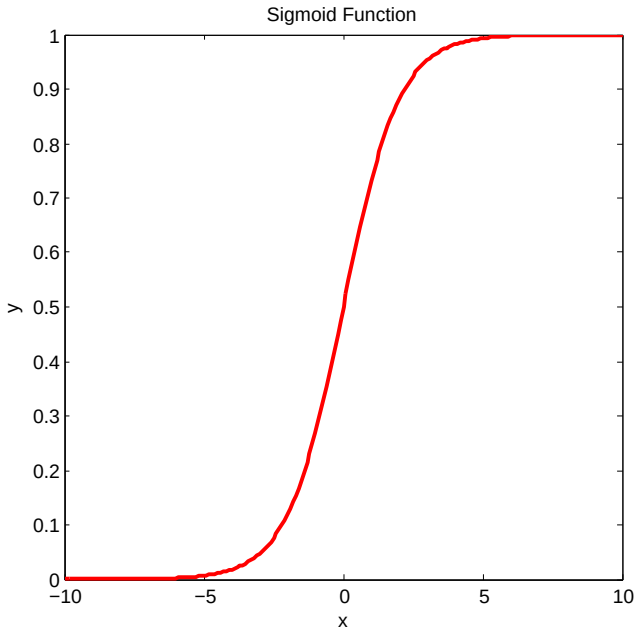


Figure 2.2: Configuration of Sigmoid Function.

process (Mackay, 1994).

The main difference between regression and classification is that one can no longer do analytical integration with respect to $\underline{\theta}$, so one instead constructs a Gaussian approximation of the posterior $p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})$ based on Laplace's asymptotic approximation (Beck and Katafygiotis, 1998; Mackay, 1992). This is achieved by making a quadratic approximation of the log-posterior around the most probable value, $\hat{\underline{\theta}}$, given by maximization of the posterior PDF, leading to a Gaussian distribution with mean $\hat{\underline{\theta}}$ and covariance matrix $\hat{\mathbf{\Sigma}}$ which is the inverse of the negative of the Hessian matrix of the log-posterior.

The detailed procedure for the Laplace approximation is as follows:

(1) For a given value of $\underline{\alpha}$, the log-posterior constructed from (2.33) and (2.34) (ig-

noring irrelevant additive terms that depend only on $\underline{\alpha}$) is:

$$\begin{aligned}\ln[p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha})] &= \sum_{n=1}^{N}\ln[P(y_n|\underline{\theta},\underline{x}_n)]+\ln[p(\underline{\theta}|\underline{\alpha})]\\ &= \sum_{n=1}^{N}\Big[y_n\cdot\ln\phi_n(\underline{\theta})+(1-y_n)\cdot\ln\{1-\phi_n(\underline{\theta})\}\Big]\\ &\quad -\frac{1}{2}\underline{\theta}^T\mathbf{A}(\underline{\alpha})\underline{\theta}\end{aligned} \tag{2.35}$$

where $\mathbf{A}(\underline{\alpha})=\mathrm{diag}(\alpha_0,\alpha_1,...,\alpha_N)$ and $\phi_n(\underline{\theta})=\phi(f(\underline{x}_n|\underline{\theta}))$. By using an iterative procedure based on a second-order Newton method (or any other optimization method), the most probable values $\hat{\underline{\theta}}(\underline{\alpha})$ are estimated by maximizing $\ln[p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha})]$.
(2) The inverse covariance matrix is $\hat{\mathbf{\Sigma}}^{-1}(\underline{\alpha})=-\nabla_{\underline{\theta}}\nabla_{\underline{\theta}}\ln p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha})$ evaluated at $\hat{\underline{\theta}}(\underline{\alpha})$ and the resulting Gaussian approximation of the posterior distribution is:

$$p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha})\cong(2\pi)^{-(m+1)/2}|\hat{\mathbf{\Sigma}}|^{-1/2}\ \exp\Big\{-\frac{1}{2}(\underline{\theta}-\hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta}-\hat{\underline{\theta}})\Big\} \tag{2.36}$$

where

$\hat{\mathbf{\Sigma}}(\underline{\alpha})=(\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi}+\mathbf{A})^{-1}\in\mathbb{R}^{(m+1)\times(m+1)}$: covariance matrix for $\underline{\theta}$, given $\underline{\alpha}$

$\hat{\underline{\theta}}(\underline{\alpha})=\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\mathbf{B}\hat{\underline{y}}(\underline{\alpha})$: the most probable value of parameter $\underline{\theta}$, given $\underline{\alpha}$

$\hat{\underline{y}}(\underline{\alpha})=\mathbf{\Phi}\hat{\underline{\theta}}+\mathbf{B}^{-1}(\underline{y}-\phi(\mathbf{\Phi}\hat{\underline{\theta}}))\in\mathbb{R}^N$

$\mathbf{B}(\underline{\alpha})=\mathrm{diag}(\beta_1,...,\beta_N)\in\mathbb{R}^{N\times N}$ with $\beta_n(\underline{\alpha})=\phi_n(\hat{\underline{\theta}})(1-\phi_n(\hat{\underline{\theta}}))$

$\mathbf{\Phi}=[\underline{\tau}_1,...,\underline{\tau}_N]^T\in\mathbb{R}^{N\times(m+1)}$

$\underline{\tau}_n=\underline{\tau}(\underline{x}_n)=[1,\ \underline{g}(\underline{x})^T]^T\in\mathbb{R}^{m+1}$.

The posterior in (2.35) contains all that is known about the parameters $\underline{\theta}$ based on the assumed model class $\mathcal{M}(\underline{\alpha})$ and the data $\mathcal{D}_N$.

The Bayesian model class selection procedure is applied as in the regression case leading to the same equation as in (2.27):

$$\hat{\alpha}_i=\begin{cases}\infty, & \text{if } Q_i^2\le S_i\\ \frac{S_i^2}{Q_i^2-S_i}, & \text{if } Q_i^2>S_i\end{cases} \tag{2.37}$$

where $Q_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{y}$ and $S_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{\tau}_i$, but with different $\mathbf{C} = \mathbf{B}^{-1} + \mathbf{\Phi} \mathbf{A}^{-1} \mathbf{\Phi}^T$.

In summary, the training phase in the classification case follows an identical procedure to the regression case presented in Section 2.2.1 except that:
(1) A sigmoid function is adopted to construct the probability model in (2.32) and hence it appears in the likelihood function.
(2) No prediction error variance σ^2 is required.
(3) The Laplace asymptotic approximation is utilized for estimating a Gaussian approximation for the posterior $p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})$.
(4) $\mathbf{C} = \mathbf{B}^{-1} + \mathbf{\Phi} \mathbf{A}^{-1} \mathbf{\Phi}^T$ is used instead of $\mathbf{C} = \sigma^2 \mathbf{I} + \mathbf{\Phi} \mathbf{A}^{-1} \mathbf{\Phi}^T$ in the Bayesian model class selection procedure for selecting $\underline{\alpha}$.

2.3.2 Prediction Phase in Classification

Predictive probability $P(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N)$ for the unknown label $\tilde{y}$ with the corresponding new feature $\underline{\tilde{x}}$ is given similarly as before by the Theorem of Total Probability and by using Laplace's asymptotic approximation twice:

$$\begin{aligned} P(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N) &= \int P(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N, \underline{\theta}) p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}) p(\underline{\alpha}|\mathcal{D}_N) d\underline{\theta} d\underline{\alpha} \\ &\cong \int P(\tilde{y}|\underline{\tilde{x}}, \underline{\theta}) p(\underline{\theta}|\mathcal{D}_N, \underline{\hat{\alpha}}) d\underline{\theta} \\ &\cong P(\tilde{y}|\underline{\tilde{x}}, \underline{\hat{\theta}}(\underline{\hat{\alpha}})) \end{aligned} \tag{2.38}$$

where $\underline{\hat{\alpha}}$, $\underline{\hat{\theta}}$ are the most probable values for $\underline{\alpha}$, $\underline{\theta}$, respectively, based on data $\mathcal{D}_N$, and $P(\tilde{y}|\underline{\tilde{x}}, \underline{\hat{\theta}}(\underline{\hat{\alpha}}))$ is given by (2.32). To deal with $p(\underline{\theta}|\mathcal{D}_N, \underline{\hat{\alpha}})$, the approximate form of the posterior distribution given in (2.36) is utilized, consistent with Laplace's asymptotic approximation for large N.

2.4 Relevance Vector Machine

In the previous sections, the Bayesian learning method using the ARD prior for regression and classification problems is investigated. However, there is still the issue of

choosing the basis functions in (2.17) and (2.31). Moreover, instead of dealing with a scalar-valued regression output, generalization to vector-valued outputs is important for some applications, such as structural health monitoring. In this section, the Bayesian learning method using the ARD prior is applied to regression and classification problems including vector-valued regression, by using kernel functions centered on each data point as the basis functions.

2.4.1 Kernel Methods

The choice of basis functions in $f(\underline{x}|\underline{\theta})$ can sometimes be based on theoretical considerations but often this is lacking and so we can use a data-based approach where the basis functions are chosen as a kernel function centered on each data point (Schölkopf and Smola, 2002), then the ARD prior can be used to automatically remove the "irrelevant" kernel terms.

The Relevance Vector Machine (RVM) is a kernel version of Bayesian learning using the ARD prior:

$$f(\underline{x}|\underline{\theta}) = \sum_{i=1}^{N} \theta_i \cdot k(\underline{x}, \underline{x}_i) + \theta_0 \tag{2.39}$$

where $\underline{\theta} = [\theta_0, \theta_1, ..., \theta_{N+1}]^T$. This equation is (2.17) or (2.31) with $g_i(\underline{x}) = k(\underline{x}, \underline{x}_i)$. Using kernel functions, however, increases the number m of parameters, i.e., dimension of $\underline{\theta}$, since $m = N$ (m and N are the number of basis terms and training data, respectively) and this is likely to cause ill-conditioning when estimating $\underline{\theta}$. This is why the ARD prior is used in the RVM to greatly reduce the number of kernel basis functions.

2.4.2 Relevance Vector Machine Regression

2.4.2.1 Training Phase

A probability model is introduced as before in Section 2.2.1:

$$\begin{aligned} y &= f(\underline{x}|\underline{\theta}) + \sigma\epsilon \\ &= \Big[\sum_{j=1}^{N} \theta_j \cdot k(\underline{x}, \underline{x}_j) + \theta_0\Big] + \sigma\epsilon \end{aligned} \tag{2.40}$$

where $\underline{\theta} = [\theta_0, \theta_1, ..., \theta_N]^T \in \mathbb{R}^{N+1}$. Note that the dimension of $\underline{\theta}$, i.e., the number of unknown parameters that need to be estimated, increases up to $N+1$ when N is the number of the given data.

The likelihood for a given dataset $\mathcal{D}_N = \{(\underline{x}_i, y_i) : i = 1, ..., N\} = (\mathbf{X}, \underline{y})$ also has the same form as in (2.18):

$$\begin{aligned} p(\mathcal{D}_N|\underline{\theta}, \sigma^2) &= \mathcal{N}(\mathbf{\Phi}\underline{\theta}, \sigma^2\mathbf{I}) \\ &= (2\pi\sigma^2)^{-N/2} \exp\Big[-\frac{1}{2\sigma^2}\|\underline{y} - \mathbf{\Phi}\underline{\theta}\|^2\Big] \end{aligned} \tag{2.41}$$

where $\mathbf{\Phi} = [\underline{\tau}(\underline{x}_1), ..., \underline{\tau}(\underline{x}_N)]^T \in \mathbb{R}^{N\times(N+1)}$, and $\underline{\tau}(\underline{x}_i) = [1, k(\underline{x}_1, \underline{x}_i), ..., k(\underline{x}_N, \underline{x}_i)]^T \in \mathbb{R}^{N+1}$.

Using the same ARD prior as in (2.19) with mean $\underline{0}$ and covariance matrix $\mathbf{A}^{-1}(\underline{\alpha})$,

$$p(\underline{\theta}|\underline{\alpha}, \sigma^2) = (2\pi)^{-\frac{N+1}{2}} |\mathbf{A}(\underline{\alpha})|^{\frac{1}{2}} \exp\Big\{-\frac{1}{2}\underline{\theta}^T\mathbf{A}(\underline{\alpha})\underline{\theta}\Big\} \tag{2.42}$$

the posterior PDF for the unknown parameters $\underline{\theta}$ can be calculated via Bayes' Theorem:

$$\begin{aligned} p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}, \sigma^2) &= \frac{p(\mathcal{D}_N|\underline{\theta}, \underline{\alpha}, \sigma^2)p(\underline{\theta}|\underline{\alpha}, \sigma^2)}{p(\mathcal{D}_N|\underline{\alpha}, \sigma^2)} \\ &= (2\pi)^{-\frac{N+1}{2}} |\hat{\mathbf{\Sigma}}|^{-1/2} \exp\Big[-\frac{1}{2}(\underline{\theta} - \hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \hat{\underline{\theta}})\Big] \end{aligned} \tag{2.43}$$

where $\hat{\mathbf{\Sigma}} = (\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{A})^{-1}$ and $\hat{\underline{\theta}} = \sigma^{-2}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\underline{y}$.

The detailed Bayesian model class selection procedure is similar to Section 2.2. The maximum of the log evidence is determined by finding the stationary points of $\mathcal{L}(\underline{\alpha}, \sigma^2)$ with respect to each α_i and σ^2. To find $\hat{\alpha}_i$:

$$\frac{\partial \mathcal{L}(\underline{\alpha}, \sigma^2)}{\partial \alpha_i} = \frac{1}{\alpha_i} - \frac{1}{\alpha_i + S_i} - \frac{Q_i^2}{\alpha_i + S_i} = 0 \tag{2.44}$$

which leads to:

$$\hat{\alpha}_i = \begin{cases} \infty, & \text{if } Q_i^2 \leq S_i \\ \frac{S_i^2}{Q_i^2 - S_i}, & \text{if } Q_i^2 > S_i \end{cases} \tag{2.45}$$

where $Q_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{y}$ and $S_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{\tau}_i$ with $\mathbf{C} = \sigma^2 \mathbf{I} + \mathbf{\Phi} \mathbf{A}^{-1} \mathbf{\Phi}^T$. If any of the α_i's approach infinity during this iterative process, then the corresponding θ_i approaches zero, thereby pruning the corresponding kernel $k(\underline{x}, \underline{x}_i)$ from the regression function. In practice (refer to the applications in the following Chapters), most of the kernel basis function terms are pruned out during the training phase; the terms which are not pruned are called *Relevance Vectors* (RVs). Since the number of RVs is quite small compared with the number of data, a sparse regression model is obtained.

Since both Q_i and S_i depend on the α_i, (2.45) does not give an explicit solution, but instead requires an iterative procedure. For the iterative procedure to determine the $\hat{\alpha}_i$, the "bottom-up" technique can be used in contrary to the "top-down" method presented in Section 2.2 and 2.3. This algorithm is described in Tipping and Faul (2003) and is summarized below. While the "top-down" algorithm starts with all basis functions included and prunes most of them out in the training procedure, the "bottom-up" algorithm starts with no basis functions included and starts to add relevant ones. The advantages of using the bottom-up algorithm are:

(1) This technique significantly reduces the training time;

(2) It prevents ill-conditioning that can occur during inversion of the Hessian matrix in the training phase.

The procedure for the "bottom-up" approach is as follows (Tipping and Faul, 2003). Starting with an initial estimate of $\underline{\hat{\alpha}}$, $\hat{\alpha}_i$ is iteratively calculated for each

$i = 0, ..., N$ by going through the following steps:

(1) Initialize σ^2, for example, $\frac{1}{10}$ of the sample variance of $\underline{y}$.

(2) Set all $\underline{\alpha}_{-i}$ to infinity and calculate each α_i for $i = 0, 1, ..., N$. α_i can be calculated from (2.45) using a single basis vector $\underline{\tau}_i$.

$$\alpha_i = \frac{S_i^2}{Q_i^2 - S_i} = \frac{\|\underline{\tau}_i\|^2}{\|\underline{\tau}_i^T \underline{y}\|^2 / \|\underline{\tau}_i\|^2 - \sigma^2} \tag{2.46}$$

since $\mathbf{C}_{-i}^{-1} = \sigma^{-2}$ in the first iteration.

(3) Compute $\hat{\mathbf{\Sigma}} = (\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{A})^{-1}$ and $\hat{\underline{\theta}} = \sigma^{-2}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\underline{y}$. These are initially scalars, and then in later iterations they are computed using the basis functions so far included.

(4) Compute S_m and Q_m for all $m = \{0, 1, ..., N\}$ using:

$$\begin{aligned}
S_m &= \frac{\alpha_m s_m}{\alpha_m - s_m} \\
Q_m &= \frac{\alpha_m q_m}{\alpha_m - s_m} \\
s_m &= \sigma^{-2}\underline{\tau}(\underline{x}_m)^T\underline{\tau}(\underline{x}_m) - \sigma^{-4}\underline{\tau}(\underline{x}_m)^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\underline{\tau}(\underline{x}_m) \\
q_m &= \sigma^{-2}\underline{\tau}(\underline{x}_m)^T\underline{y} - \sigma^{-4}\underline{\tau}(\underline{x}_m)^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\underline{y}
\end{aligned}$$

(5) Select the basis function $\underline{\tau}(\underline{x}_i)$ and the corresponding hyperparameter α_i^{new} that maximizes $\mathcal{L}(\underline{\alpha}, \sigma^2)$ and then update hyperparameter α_i, as follows:

- If $Q_i^2 > S_i$ and $\alpha_i < \infty$, then update $\alpha_i = \alpha_i^{new} = S_i^2/(Q_i^2 - S_i)$.
- If $Q_i^2 > S_i$ and $\alpha_i = \infty$, then add $\underline{\tau}(\underline{x}_i)$ and update $\alpha_i = \alpha_i^{new} = S_i^2/(Q_i^2 - S_i)$.
- If $Q_i^2 \leq S_i$, then prune $\underline{\tau}(\underline{x}_i)$ and set $\alpha_i = \alpha_i^{new} = \infty$.

(6) Recompute the variance $(\sigma^2)^{new} = (\|\underline{y} - \mathbf{\Phi}\hat{\underline{\theta}}\|^2)/(N - \sum_i \gamma_i)$, $\hat{\mathbf{\Sigma}}$, and then $\hat{\underline{\theta}}$ where $\gamma_i = 1 - \alpha_i\mathbf{\Sigma}_{ii}$ and $\mathbf{\Sigma}_{ii}$ is the i^{th} diagonal element of $\mathbf{\Sigma}$ computed with current $\underline{\alpha}$ and σ^2.

(7) Repeat the above procedure (from 3 to 6) until it converges.

2.4.2.2 Prediction Phase

Based on the results from the previous section, a probabilistic prediction for the unknown response $\tilde{y}$ is calculated from new input $\underline{\tilde{x}}$ and the dataset $\mathcal{D}_N$ using the Theorem of Total Probability to include all possible uncertainties of the parameters:

$$p(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N) = \mathcal{N}(\tilde{y}|y_*, \sigma_*^2)$$

where $\hat{\underline{\alpha}}$, $\hat{\sigma}^2$ are the most probable values for $\underline{\alpha}$, σ^2, respectively, based on data $\mathcal{D}_N$, $y_* = \hat{\underline{\theta}}^T \tau(\underline{\tilde{x}})$, and $\sigma_*^2 = \hat{\sigma}^2 + \tau(\underline{\tilde{x}})^T \mathbf{\Sigma} \tau(\underline{\tilde{x}})$, as before.

2.4.3 Relevance Vector Machine Classification

2.4.3.1 Training Phase

The RVM classification also uses Bayesian learning with the ARD prior. For the available dataset $\mathcal{D}_N = \{(\underline{x}_i, y_i) : i = 1, ..., N\} = (\mathbf{X}, \underline{y})$ with a predictive probability model $P(y|\underline{x}, \underline{\theta})$, data $\mathcal{D}_N$ is used to update the prior PDF $p(\underline{\theta}|\mathcal{M}(\underline{\alpha}))$ to $p(\underline{\theta}|\mathcal{D}_N, \mathcal{M}(\underline{\alpha}))$ via Bayes' Theorem, where $\mathcal{M}(\underline{\alpha})$ denotes the *model class* (i.e., mathematical structure assumed for $P(y|\underline{x}, \underline{\theta})$ *and* the prior PDF $p(\underline{\theta}|\mathcal{M}(\underline{\alpha}))$).

Using a monotonically increasing sigmoid function $\phi(\cdot) \in [0, 1]$ defined as before, the predictive probability model is:

$$P(y|\underline{x}, \underline{\theta}) = \phi(f(\underline{x}|\underline{\theta})^y \{1 - \phi(f(\underline{x}|\underline{\theta}))\}^{1-y} \tag{2.47}$$

where label $y \in \{0, 1\}$, $\underline{\theta} \in \mathbb{R}^{N+1}$ and $k(\underline{x}, \underline{x}_i)$ is a specified kernel function. Note that

$$f(\underline{x}|\underline{\theta}) = \sum_{j=1}^{N} \theta_j k(\underline{x}, \underline{x}_j) + \theta_0$$

and if $f(\underline{x}|\underline{\theta}) = 0$, $P(y = 1|\underline{x}, \underline{\theta}) = P(y = 0|\underline{x}, \underline{\theta}) = 0.5$, so $\sum_{i=1}^{N} \theta_i k(\underline{x}, \underline{x}_i) + \theta_0 = 0$

defines the boundary surface in the feature space between those feature vectors that imply label $y = 1$ is more likely and those that imply $y = 0$ is more likely. Note also from (2.47) that $P(y = 0|\underline{x}, \underline{\theta}) + P(y = 1|\underline{x}, \underline{\theta}) = 1$.

Assume $\mathcal{D}_N$ consists of independent samples, then the likelihood function is:

$$\begin{aligned} P(\mathcal{D}_N|\underline{\theta}) &= \prod_{i=1}^{N} P(y_i|\underline{x}_i, \underline{\theta}) \\ &= \prod_{i=1}^{N} \phi(f(\underline{x}_i|\underline{\theta}))^{y_i}\{1 - \phi(f(\underline{x}_i|\underline{\theta}))\}^{1-y_i} \end{aligned} \tag{2.48}$$

Once again, define the ARD prior PDF to be Gaussian with mean $\underline{0}$ and covariance matrix $\mathbf{A}^{-1}(\underline{\alpha}) = \text{diag}(\alpha_0^{-1}, \alpha_1^{-1}, ..., \alpha_N^{-1})$, i.e., $p(\underline{\theta}|\underline{\alpha}) = \mathcal{N}(\underline{0}, \mathbf{A}^{-1}(\underline{\alpha}))$. Note that each hyperparameter $\underline{\alpha} \in \mathcal{A} \subset \mathbb{R}^{N+1}$ defines a model class $\mathcal{M}(\underline{\alpha})$ where each predictive model in the class is defined by specifying $\underline{\theta} \in \mathbf{\Theta}$ in (2.47), independent of the hyperparameter $\underline{\alpha}$, which only serves to specify the prior for the model class.

The RVM classification uses Laplace's asymptotic approximation for the posterior, $p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})$ (MacKay, 1992; Beck and Katafygiotis, 1998; Appendix B), which leads to:

$$p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}) \cong (2\pi)^{-(N+1)/2}|\hat{\mathbf{\Sigma}}|^{-1/2} \exp\Big\{ -\frac{1}{2}(\underline{\theta} - \underline{\hat{\theta}})^T \hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \underline{\hat{\theta}})\Big\} \tag{2.49}$$

where $\hat{\mathbf{\Sigma}}(\underline{\alpha}) = (\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A})^{-1} \in \mathbb{R}^{(N+1)\times(N+1)}$,

$\underline{\hat{\theta}}(\underline{\alpha}) = \hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\mathbf{B}\underline{\hat{y}}$,

$\underline{\hat{y}}(\underline{\alpha}) = \mathbf{\Phi}\underline{\hat{\theta}} + \mathbf{B}^{-1}(\underline{y} - \phi(\mathbf{\Phi}\underline{\hat{\theta}})) \in \mathbb{R}^N$,

$\mathbf{A}(\underline{\alpha}) \in \mathbb{R}^{(N+1)\times(N+1)}$,

$\mathbf{B}(\underline{\alpha}) = \text{diag}(\beta_1, ..., \beta_N) \in \mathbb{R}^{N\times N}$,

$\beta_n(\underline{\alpha}) = \phi(f(\underline{x}_n|\underline{\theta}))\{1 - \phi(f(\underline{x}_n|\underline{\theta}))\}$,

$\mathbf{\Phi} = [\underline{\tau}(\underline{x}_1), ..., \underline{\tau}(\underline{x}_N)]^T \in \mathbb{R}^{N\times(N+1)}$,

$\underline{\tau}(\underline{x}_n) = [1, k(\underline{x}_n, \underline{x}_1), ..., k(\underline{x}_n, \underline{x}_N)]^T \in \mathbb{R}^{N+1}$.

In the next step of the RVM, Bayesian model class selection is used to optimize the hyperparameter $\underline{\alpha} \in \mathcal{A}$ as detailed in Appendix C. The most plausible hyperparameter

$\underline{\alpha}$ is:

$$\hat{\alpha}_i = \begin{cases} \infty, & \text{if } Q_i^2 \leq S_i \\ \frac{S_i^2}{Q_i^2 - S_i}, & \text{if } Q_i^2 > S_i \end{cases} \tag{2.50}$$

where $Q_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{y}$ and $S_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{\tau}_i$ with $\mathbf{C} = \mathbf{B}^{-1} + \mathbf{\Phi} \mathbf{A}^{-1} \mathbf{\Phi}^T$.

To estimate this $\hat{\underline{\alpha}}$, the bottom-up algorithm explained for RVM regression is also utilized in order to prevent ill-conditioning which may occur during the training phase.

2.4.3.2 Prediction Phase

For the unknown label $\tilde{y}$ corresponding to new feature $\underline{\tilde{x}}$, the desired predictive probability is given by the Theorem of Total Probability:

$$P(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N) \;\; \cong \;\; P(\tilde{y}|\underline{\tilde{x}}, \underline{\hat{\theta}}(\underline{\hat{\alpha}})) \tag{2.51}$$

where $\underline{\hat{\alpha}}$, $\underline{\hat{\theta}}$ are the most probable values for $\underline{\alpha}$, $\underline{\theta}$, respectively, based on data $\mathcal{D}_N$, and $P(\tilde{y}|\underline{\tilde{x}}, \underline{\hat{\theta}}(\underline{\hat{\alpha}}))$is given by (2.32).

2.4.4 Relevance Vector Machine Regression for Vector Outputs

2.4.4.1 Training Phase

The original RVM algorithm was presented only for a scalar output (Tipping, 2001). In this section, the procedure for training and prediction using an expanded RVM methodology that is applicable to vector outputs is presented (Thayananthan, 2005).

Let $\underline{f}(\underline{x}|\underline{\theta})$ denote the chosen regression function relating the feature vector $\underline{x} \in \mathbb{R}^L$ to the most probable value of the output vector $\underline{y} \in \mathbb{R}^M$ (for example, damage location or severity index vector in structural health monitoring), when the model parameter vector $\underline{\theta}$ is specified. This function is embedded in a probability model by introducing an uncertain prediction error to account for the fact that no model gives prefect

predictions, so:

$$\underline{y} = \underline{f}(\underline{x}|\underline{\theta}) + \underline{\epsilon} \tag{2.52}$$

where $\underline{\epsilon}$ is modeled as a Gaussian vector with zero mean and covariance matrix $\mathbf{\Omega} = \text{diag}(\sigma_1^2, ..., \sigma_M^2)$. This choice of the probability model for the prediction error is motivated by Jaynes' Principle of Maximum Entropy (Jaynes, 2003). It gives a Gaussian (Normal) predictive probability model (PDF) for the output vector $\underline{y}$:

$$p(\underline{y}|\underline{x}, \underline{\theta}, \underline{\sigma}^2) = \mathcal{N}(\underline{y}|\underline{f}(\underline{x}|\underline{\theta}), \mathbf{\Omega}) \tag{2.53}$$

where $\underline{\sigma}^2 = [\sigma_1^2, ..., \sigma_M^2]^T$ and $\mathbf{\Omega} = \text{diag}(\underline{\sigma}^2)$.

Let $\mathcal{D}_N = \{(\underline{x}_i, \underline{y}_i) : i = 1, ..., N\}$ denote a training dataset where $\underline{x}_i \in \mathbb{R}^L$ is the i^{th} example of the feature vector and $\underline{y}_i \in \mathbb{R}^M$ is the corresponding vector output. As before, the regression function is expressed in terms of a kernel basis expansion where the i^{th} kernel function is centered at data point $\underline{x}_i$ (we use Gaussian radial basis functions in the examples later) so that for the m^{th} component of the regression function $\underline{f}$:

$$f_m(\underline{x}|\underline{\theta}) = \theta_{m0} + \sum_{i=1}^{N} \theta_{mi} k(\underline{x}, \underline{x}_i) = \underline{\tau}(\underline{x})^T \underline{\theta}_m \tag{2.54}$$

where $\underline{\theta}_m = [\theta_{m0}, \theta_{m1}, ..., \theta_{mN}]^T \in \mathbb{R}^{N+1}$, $\underline{\theta} = [\underline{\theta}_1^T, ..., \underline{\theta}_M^T]^T \in \mathbb{R}^{(N+1)M}$, and $\underline{\tau}(\underline{x}) = [1, k(\underline{x}, \underline{x}_1), ..., k(\underline{x}, \underline{x}_N)]^T \in \mathbb{R}^{N+1}$ for $m = 1, ..., M$. Using Bayes' Theorem to incorporate the information from the data $\mathcal{D}_N$ leads to the posterior PDF for the model parameter vector $\underline{\theta}$:

$$p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}, \underline{\sigma}^2) = \frac{p(\mathcal{D}_N|\underline{\theta}, \underline{\sigma}^2) p(\underline{\theta}|\underline{\alpha})}{p(\mathcal{D}_N|\underline{\alpha}, \underline{\sigma}^2)} \tag{2.55}$$

where $\underline{\alpha} = [\alpha_0, ..., \alpha_N]^T$ contains hyperparameters that control the prior for $\underline{\theta}$.

The components of $\underline{y}$ are independent (since $\mathbf{\Omega}$ is diagonal) so the likelihood function can be expressed as a product of Gaussians for each output component:

$$p(\mathcal{D}_N|\underline{\theta}, \underline{\sigma}^2) = \prod_{m=1}^{M} \mathcal{N}(\underline{\nu}_m|\mathbf{\Phi}\underline{\theta}_m, \sigma_m^2 \mathbf{I}) \tag{2.56}$$

where $\underline{\nu}_m = [(\underline{y}_1)_m, ..., (\underline{y}_N)_m]^T$ and $\mathbf{\Phi} = [\underline{\tau}(\underline{x}_1), ..., \underline{\tau}(\underline{x}_N)]^T \in \mathbb{R}^{N\times(N+1)}$. The prior is:

$$p(\underline{\theta}|\underline{\alpha}) = \prod_{m=1}^{M} \mathcal{N}(\underline{\theta}_m|\underline{0}, \mathbf{A}^{-1}) \tag{2.57}$$

where matrix $\mathbf{A} = \text{diag}(\alpha_0, ..., \alpha_N)$. The posterior PDF for $\underline{\theta}$ is:

$$p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}, \underline{\sigma}^2) \propto \prod_{m=1}^{M} \mathcal{N}(\underline{\theta}_m|\hat{\underline{\theta}}_m, \mathbf{\Sigma}_m) \tag{2.58}$$

where $\hat{\underline{\theta}}_m = \sigma_m^{-2}\mathbf{\Sigma}_m\mathbf{\Phi}^T\underline{\nu}_m$ and $\mathbf{\Sigma}_m = (\sigma_m^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{A})^{-1}$ give the most probable value of $\underline{\theta}_m$ and its covariance matrix, respectively.

In the next step, Bayesian model class selection is used to select the most probable hyperparameters $\hat{\underline{\alpha}}$ and variances $\hat{\underline{\sigma}}^2$ based on data $\mathcal{D}_N$. If we take a uniform prior probability distribution over all possible model classes defined by $\underline{\alpha}$ and $\underline{\sigma}^2$ then by applying Bayes' Theorem at the model class level, the most probable model class is the one that maximizes the log evidence (Beck and Yuen, 2004), which is given by (Tipping and Faul, 2003):

$$\begin{aligned}
\mathcal{L}(\underline{\alpha}, \underline{\sigma}^2) &= \ln p(\mathcal{D}_N|\underline{\alpha}, \underline{\sigma}^2) \\
&= \ln \int p(\mathcal{D}_N|\underline{\theta}, \underline{\sigma}^2) p(\underline{\theta}|\underline{\alpha}) d\underline{\theta} \\
&= -\frac{1}{2}\sum_{m=1}^{M}\Big[N\ln 2\pi + \ln|\mathbf{C}_m| + \underline{\nu}_m^T\mathbf{C}_m^{-1}\underline{\nu}_m\Big] \\
&= \mathcal{L}(\underline{\alpha}_{-i}, \underline{\sigma}^2) + \sum_{m=1}^{M}\Big[\ln\alpha_i - \ln(\alpha_i + S_{mi}) + \frac{Q_{mi}^2}{\alpha_i + S_{mi}}\Big] \\
&= \mathcal{L}(\underline{\alpha}_{-i}, \underline{\sigma}^2) + l(\alpha_i, \underline{\sigma}^2)
\end{aligned} \tag{2.59}$$

where $\mathcal{L}(\underline{\alpha}_{-i}, \underline{\sigma}^2)$ is the evidence with $\underline{\tau}_i = \underline{\tau}(\underline{x}_i)$ excluded, $\mathbf{C}_m = \sigma_m^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T$, $S_{mi} = \frac{\alpha_i s_{mi}}{\alpha_i - s_{mi}}$ and $Q_{mi} = \frac{\alpha_i q_{mi}}{\alpha_i - S_{mi}}$ and s_{mi} and q_{mi} are calculated using the Woodbury

identity as follows:

$$s_{mi} = \sigma_m^{-2} \underline{\tau}_i^T \underline{\tau}_i - \sigma_m^{-4} \underline{\tau}_i^T \mathbf{\Phi} \mathbf{\Sigma}_m \mathbf{\Phi}^T \underline{\tau}_i$$
$$q_{mi} = \sigma_m^{-2} \underline{\tau}_i^T \underline{\nu}_m - \sigma_m^{-4} \underline{\tau}_i^T \mathbf{\Phi} \mathbf{\Sigma}_m \mathbf{\Phi}^T \underline{\nu}_m$$

The maximum of the log evidence is determined by iteratively maximizing $l(\alpha_i, \underline{\sigma}^2)$ with respect to α_i and σ_m^2. For example, to find $\hat{\alpha}_i$:

$$\frac{\partial \mathcal{L}(\underline{\alpha}, \underline{\sigma}^2)}{\partial \alpha_i} = \sum_{m=1}^{M} \left\{ \frac{1}{\alpha_i} - \frac{1}{\alpha_i + S_{mi}} - \frac{Q_{mi}^2}{\alpha_i + S_{mi}} \right\} = 0 \tag{2.60}$$

and similarly, an expression can be found for $\hat{\underline{\sigma}}_m^2$. Since S_{mi} and Q_{mi} in $l(\alpha_i, \underline{\sigma}^2)$ depend on all of the α_is (since $\mathbf{\Sigma}_m$ does), an iterative algorithm explained in detail in Appendix C is applied to solve these equations for each $\hat{\alpha}_i$ and $\hat{\sigma}_m^2$. In practice, many of the α_is approach infinity during this iterative process, and so the corresponding θ_{mi} for each $m = 1, ..., M$ are set to zero (they have zero mean and zero variance), thereby pruning the corresponding kernel $k(\underline{x}, \underline{x}_i)$ from the regression function. In the end, only the kernels corresponding to a small number of the $\underline{x}_i$ are retained (as before, these are called the relevance vectors).

That training algorithm for RVM for vector outputs is summarized as follows (Thayananthan, 2005):

(1) Initialize σ_m^2 as $\frac{1}{10}$ of the sample variance of $\underline{\nu}_m$, for $m = 1, ...M$.

(2) Compute $\hat{\mathbf{\Sigma}}_m$ and $\hat{\underline{\theta}}_m$ for $m = 1, ...M$ from:

$$\hat{\mathbf{\Sigma}}_m = (\sigma_m^{-2} \mathbf{\Phi}^T \mathbf{\Phi} + \mathbf{A})^{-1}$$
$$\hat{\underline{\theta}}_m = \sigma_m^{-2} \hat{\mathbf{\Sigma}}_m \mathbf{\Phi}^T \underline{\nu}_m$$

(3) Compute S_{mi} and Q_{mi} for all $m = 1, ..., M$ and $i = 0, ..., N$ using:

$$
\begin{aligned}
S_{mi} &= \frac{\alpha_i s_{mi}}{\alpha_i - s_{mi}} \\
Q_{mi} &= \frac{\alpha_i q_{mi}}{\alpha_i - s_{mi}} \\
s_{mi} &= \sigma_m^{-2} \underline{\tau}(\underline{x}_i)^T \underline{\tau}(\underline{x}_i) - \sigma_m^{-4} \underline{\tau}(\underline{x}_i)^T \mathbf{\Phi} \mathbf{\Sigma}_m \mathbf{\Phi}^T \underline{\tau}(\underline{x}_i) \\
q_{mi} &= \sigma_m^{-2} \underline{\tau}(\underline{x}_i)^T \underline{\nu}_m - \sigma_m^{-4} \underline{\tau}(\underline{x}_i)^T \mathbf{\Phi} \mathbf{\Sigma}_m \mathbf{\Phi}^T \underline{\nu}_m
\end{aligned}
$$

(4) Select the basis function $\underline{\tau}(\underline{x}_i)$ and the corresponding hyperparameter α_i^{new} that maximizes $\mathcal{L}(\underline{\alpha}, \underline{\sigma}^2)$:

$$
\begin{aligned}
\alpha_i^{new} &= \arg\max_{\alpha_i} l(\alpha_i, \underline{\sigma}^2) \\
m &= \arg\max_{i} \; l(\alpha_i^{new}, \underline{\sigma}^2)
\end{aligned}
$$

where $l(\alpha_i, \sigma^2)$ is defined (implicitly) by (2.59).

- If $\alpha_m = \infty$ and $\alpha_m^{new} < \infty$, then add $\underline{\tau}(\underline{x}_m)$ and update $\alpha_m = \alpha_m^{new}$.
- If $\alpha_m < \infty$ and $\alpha_m^{new} = \infty$, then remove $\underline{\tau}(\underline{x}_m)$ and update $\alpha_m = \infty$.
- If $\alpha_m < \infty$ and $\alpha_m^{new} < \infty$, then update $\alpha_m = \alpha_m^{new}$

(5) Recompute the variance σ_m^2, then $\hat{\mathbf{\Sigma}}_m$, and $\hat{\underline{\theta}}_m$, using:

$$
\hat{\sigma}_m^2 = \frac{||\underline{\nu}_m - \mathbf{\Phi} \hat{\underline{\theta}}_m||^2}{M - \sum_{i=1}^{M} \gamma_i}
$$

where $\gamma_i = 1 - \alpha_i \Sigma_{ii}$ and Σ_{ii} is the i^{th} diagonal elements of Σ.

(6) Repeat the above procedure (from 2 to 5) until it converges.

2.4.4.2 Prediction Phase

In the prediction phase, we make predictions for the output (damage index vector) corresponding to a new feature vector $\underline{\tilde{x}}$ based on the robust posterior predictive

probability distribution as follows. The robust predictive probability for the corresponding output vector $\underline{\tilde{y}}$ based on the most probable model class is given by the Theorem of Total Probability:

$$\begin{aligned} p(\underline{\tilde{y}}|\underline{\tilde{x}}, \mathcal{D}_N, \underline{\hat{\alpha}}, \underline{\hat{\sigma}}^2) &= \int p(\underline{\tilde{y}}|\underline{\tilde{x}}, \underline{\theta}, \underline{\hat{\sigma}}^2) p(\underline{\theta}|\mathcal{D}_N, \underline{\hat{\alpha}}, \underline{\hat{\sigma}}^2) d\underline{\theta} \\ &= \mathcal{N}(\underline{\tilde{y}}|\underline{y}_*, \mathbf{\Omega}_*) \end{aligned} \tag{2.61}$$

where $\underline{y}_* = [y_{1*}, ..., y_{M*}] \in \mathbb{R}^M$, $y_{m*} = \underline{\hat{\theta}}_m^T \underline{\tau}(\underline{\tilde{x}})$, $\underline{\theta} = [\underline{\theta}_1^T, ..., \underline{\theta}_M^T]^T \in \mathbb{R}^{(N+1)M}$, $\mathbf{\Omega}_* = \text{diag}(\sigma_{1*}^2, ..., \sigma_{M*}^2) \in \mathbb{R}^{M \times M}$, $\sigma_{m*}^2 = \hat{\sigma}_m^2 + \underline{\tau}(\underline{\tilde{x}})^T \hat{\mathbf{\Sigma}}_m \underline{\tau}(\underline{\tilde{x}})$ and $\hat{\mathbf{\Sigma}}_m = (\hat{\sigma}_m^{-2} \mathbf{\Phi}^T \mathbf{\Phi} + \hat{\mathbf{A}})^{-1}$.

2.4.5 Illustrative Examples

2.4.5.1 RVM Regression: Sinc Function Estimation

As an illustrative example for RVM regression, the function $\text{sinc}(x) = \sin(x)/x$, is chosen. 100 uniformly-spaced samples of $\underline{x} \in [-10, 10]$ are generated with Gaussian noise of mean 0 and standard deviation 0.01 (sinc function and generated samples shown as red dashed line and red dots, respectively, in Figure 2.3).

The Gaussian kernel, $k(\underline{x}_m, \underline{x}_n) = \exp\left(-\frac{\|\underline{x}_m - \underline{x}_n\|^2}{2^2} \right)$ with width = 2 is used for both RVM and SVM where the SVM results using the same dataset are also presented for the purpose of comparison. The solid blue lines in Figure 2.3 are the estimated regression functions by RVM and SVM. The blue circles around some of the data specify Relevance Vectors (RVs) or Support Vectors (SVs); these RVs or SVs control the identified regression function based on the 100 data points.

The trade-off parameter C for SVM is estimated by 5-fold cross-validation and is given in Table 2.1. In 5-fold cross-validation, the training dataset is divided into 5 subsets of (approximately) equal size and SVM is trained 5 times, each time leaving out one of the subsets from training and then using the omitted subset to determine the value of C which satisfies a certain error criterion i.e., to minimize the prediction error for the omitted subset. It is shown in Figure 2.3 that the RVM algorithm provides a satisfactory regression function using only 8 RVs, a much smaller number of

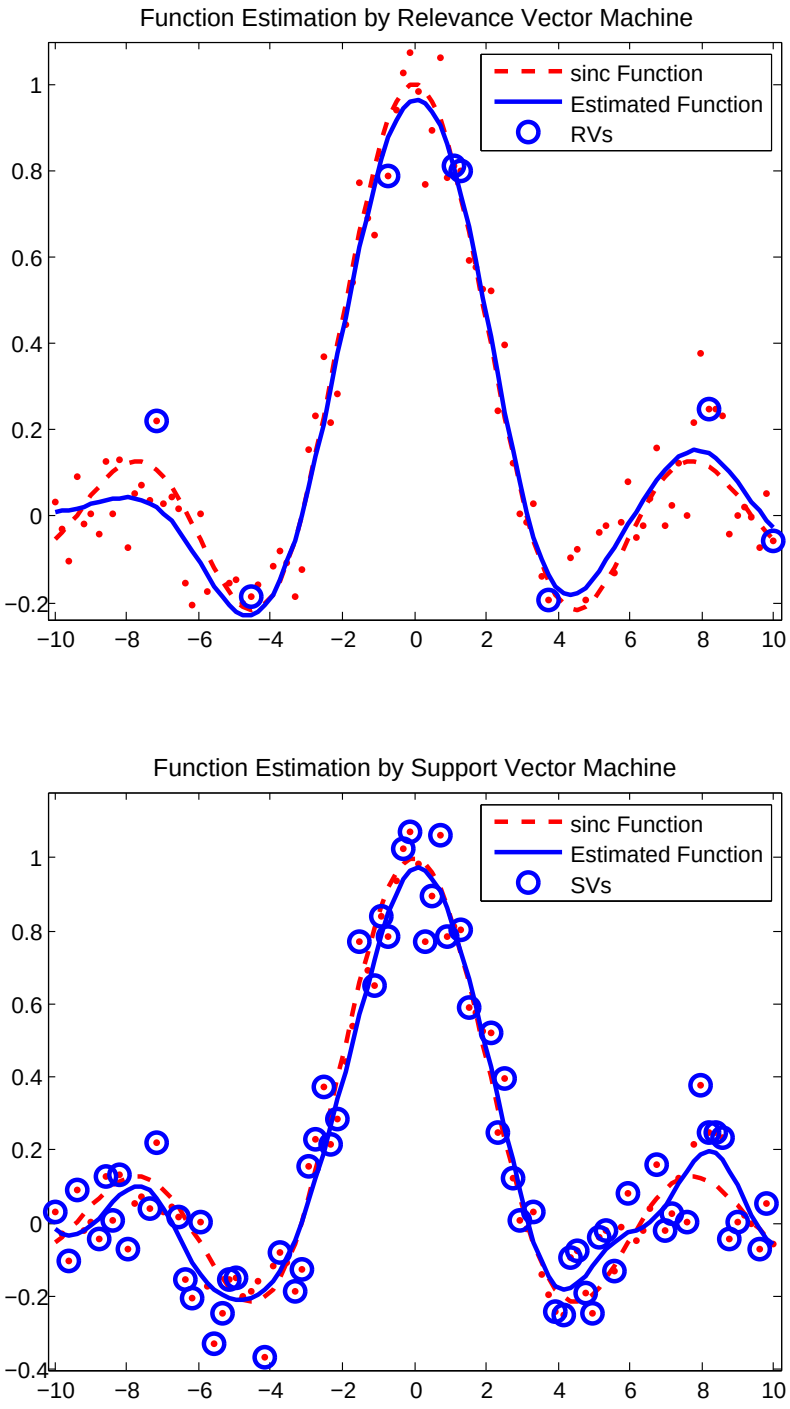


Figure 2.3: RVM and SVM Regression Based on Dataset of 100 Points Generated from sinc Function with Gaussian Noise of Mean 0 and Standard Deviation 0.01.

terms compared with 67 SVs for the case of SVM.

	C	No. of RVs or SVs	Error (RMS)
RVM	N/A	8	0.039
SVM	1.62	67	0.047

Table 2.1: Comparison between RVM and SVM for Sinc Function Estimation.

100 test data generated from the same sinc function are used for quantitative comparison between RVM and SVM and the prediction is performed. The root-mean-square (RMS) errors between the data and the RVM and SVM predictions are computed for each case as shown in Table 2.1; the smaller RMS error and much smaller number of terms in the regression function suggest that the RVM is a more effective regression method.

2.4.5.2 RVM Classification: Ripley's 2-D Gaussian-mixture Dataset

As a simple example of an RVM application for classification, it is applied to randomly generated datasets from two mixtures of two equally-weighted 2-D Gaussian PDFs (equal variances of 0.03) due to Ripley (1996). Data are labelled with $y = 1$ (blue dots) if from the mixture of two Gaussian PDFs with means at $\underline{x} = (-0.3, 0.7)$ and $(0.4, 0.7)$, and labelled with $y = 0$ (red crosses) if from the mixture of the two Gaussian PDFs with means at $\underline{x} = (-0.7, 0.3)$ and $(0.3, 0.3)$. The results by SVM using the same dataset are also presented for the purpose of comparison. The Gaussian kernel, $k(\underline{x}_m, \underline{x}_n) = \exp\left(-\frac{\|\underline{x}_m - \underline{x}_n\|^2}{0.5^2}\right)$ is selected for both RVM and SVM, and the results from using 250 data points with labels y_n and feature vectors $\underline{x}_n = (x_{1n}, x_{2n})$, $n = 1, ..., 250$, are shown in Figure 2.4, where the principal decision boundary separating the two labelled datasets is plotted as a solid blue line for RVM and SVM. The dashed and dotted lines alongside the solid line in the RVM classification represent more conservative decision boundaries explained later. The blue circles around some of the data specify Relevance Vectors (RVs) or Support Vectors (SVs). They are the data points controlling the decision boundary based on the 250 data points. The parameter C given in Table 2.2 represents the trade-off between model complexity

and misclassification in SVM and it is determined by using 10-fold cross-validation (Vapnik, 1998). Figure 2.4 demonstrates that the RVM algorithm provides a satisfactory decision boundary by using only 7 RVs (i.e., 7 Gaussian kernel terms), which is quite small compared with 93 SVs for the case of SVM. Note that large x_1 and x_2 correspond to a high probability for label $y = 1$ in RVM.

	C	No. of RVs or SVs	Misclassification rate(%)	
			$P_1 = P_0 = 0.5$	$P_1 = 0.4$ and $P_0 = 0.4$
RVM	N/A	7	9.90	6.80
SVM	1.32	93	9.90	N/A

Table 2.2: Comparison between RVM and SVM for Gaussian-mixture Dataset.

For quantitative comparison, 1,000 test data are generated from the same Gaussian-mixture distribution with known labels, and prediction is performed to investigate the respective misclassification rates based on the trained RVM and SVM algorithms. The misclassification rates are given in Table 2.2 as the percentage misclassified by being on the wrong side of the principal decision boundary (which corresponds to a probability = 0.5 threshold in the RVM). For example, if 99 test data are misclassified (i.e., given a wrong label), then the error is $\frac{99}{N} \times 100 = 9.9\%$, where N=1,000 is the total number of test data. As is shown in Table 2.2, the misclassification rate for both methods is the same when using the principal decision boundary. The misclassification rate with RVM drops to 6.80% if two conservative decision boundaries corresponding to $P_1 = 0.4$ and $P_0 = 0.4$ are used; e.g., a data point with actual label $y = 0$ is misclassified only if it appears above the contour for $P_0 = 0.4$. Such probabilistic decision making is possible only for RVM.

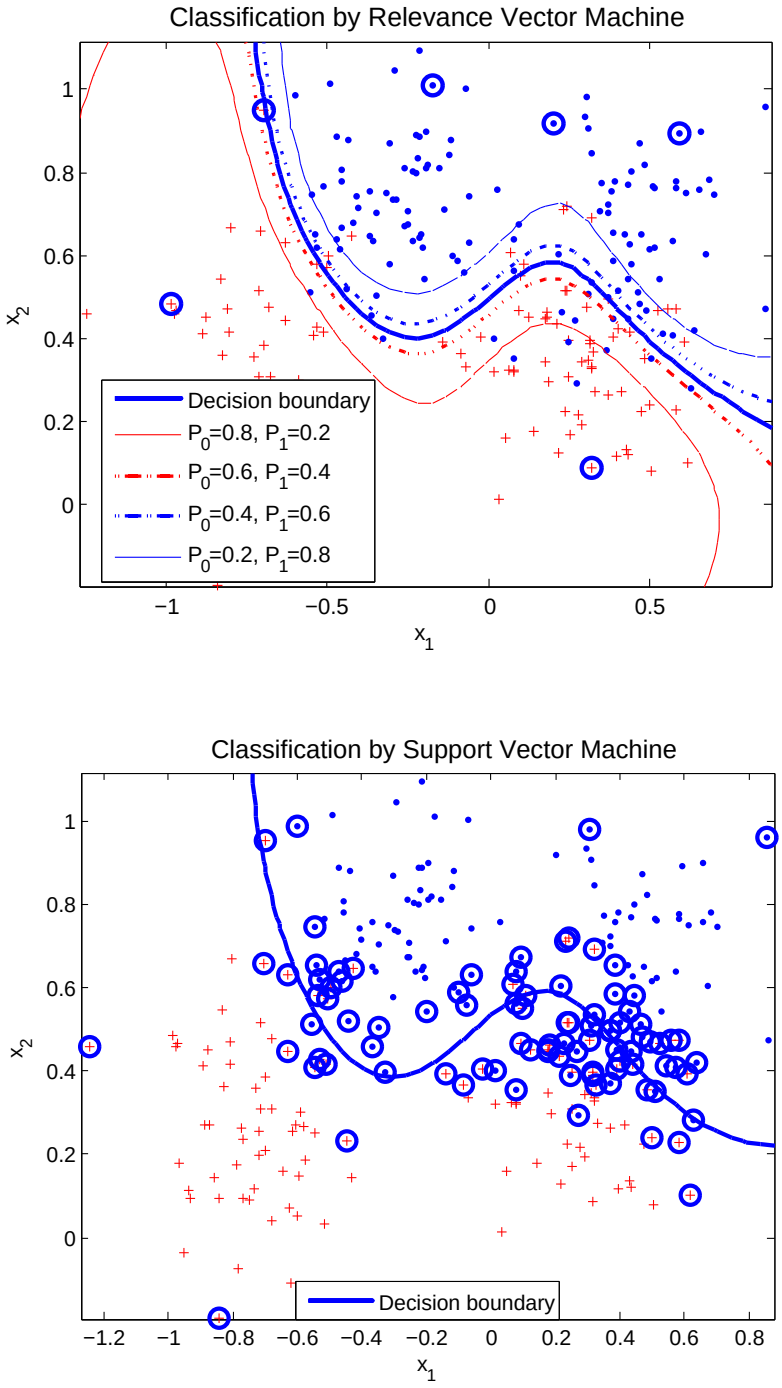


Figure 2.4: RVM and SVM Classification Based on Dataset of 250 Points Drawn Randomly from Two Mixtures of Two 2-D Gaussians. (The RVs and SVs that control the decision boundaries are shown as blue circles. P_1 and P_0 denote the probabilities of labels $y = 1$ and $y = 0$, respectively.)

Chapter 3

Near-source and Far-source Classification for Earthquake Early Warning

Since an earthquake is an abrupt event that comes without much warning, there is increasing research interest in automated seismic early warning systems that can take rapid actions to mitigate damage and loss before the onset of the damaging ground shaking at a facility (Allen and Kanamori, 2003; Cua, 2005). The basic principle in seismic early warning is that an automated and reliable system may allow time for taking mitigation measures because the speed of transmitted signals (about $300,000\ km/s$) to the system computer from the seismic network sensors that detect the onset of the event is much faster than that of the most damaging S-waves (about 3.5 km/s).

The Virtual Seismologist (VS) method was recently developed for an early warning system (Cua, 2005), which can estimate the location of the epicenter and the magnitude within a few seconds after the detection of the P-waves near the causative fault. This VS method, however, currently works for moderate earthquakes of magnitude less than about 6.5 because it assumes a point-source model for the rupture (Cua, 2005). To construct a seismic early warning system dealing with larger earthquakes, knowledge of the fault geometry is essential, and an important ingredient in establishing the extent of the rupturing fault is to be able to classify the station into near-source and far-source (Yamada et al., 2007).

In this Chapter, automatic near-source and far-source classification of incoming ground motion signals is presented, and the Bayesian learning method using an ARD prior determines which ground motion features are optimal for this classification. In Section 3.1.1, the earthquake dataset used for training is described, and the prediction phase of the Bayesian learning method using the ARD prior is presented in Section 3.1.2. In Section 3.2, the classification results obtained by the proposed method are presented and compared with those from a previous related study (Yamada et al., 2007).

3.1 Near-source and Far-source Classification

3.1.1 Earthquake Data

The same dataset used previously by Yamada et al. (2007) is utilized to allow comparison of the results. This dataset consists of 695 strong-motion records from 9 earthquakes of magnitude greater than 6.0: Imperial Valley (1979), Loma Prieta (1989), Landers (1992), Northridge (1994), Hyogoken-Nanbu (1995), Izmit (1999), Chi-Chi (1999), Denali (2002) and Niigataken-Chuetsu(2004). If the station is located less than 10 km from the fault rupture, the corresponding records are categorized as near-source (NS) and far-source (FS) otherwise. Only stations with fault distances less than 200 km are included, since otherwise the ground motion amplitudes are small, resulting in a low signal-to-noise ratio. In Table 3.1, the utilized number of NS and FS records for each earthquake is listed.

The eight ground motion features listed in Table 3.2 were extracted from each of the 695 records by Yamada (2007) after suitable signal processing of the accelerograms. The $\log_{10}$ values of these features are combined into a vector $\underline{g}(\underline{x}) \in \mathbb{R}^8$:

$$\begin{aligned} \underline{g}(\underline{x}) = \; [\;& \log_{10} H_j, \; \log_{10} Z_j, \; \log_{10} H_a, \; \log_{10} Z_a, \; \log_{10} H_v, \; \log_{10} Z_v, \\ & \log_{10} H_d, \; \log_{10} Z_d]^T \end{aligned} \tag{3.1}$$

Earthquake	$M_w{}^a$	NS	FS	Total	Fault Model^b
Imperial Valley (1979)	6.5	14	20	34	Hartzell and Heaton (1983)
Loma Prieta (1989)	6.9	8	39	47	Wald et al. (1991)
Landers (1992)	7.3	1	112	113	Wald and Heaton (1994)
Northridge (1994)	6.6	17	138	155	Wald et al. (1996)
Hyogoken-Nanbu (1995)	6.9	4	14	18	Wald (1996)
Izmit (1999)	7.6	4	13	17	Sekiguchi and Iwata (2002)
Chi-Chi (1999)	7.6	42	172	214	Ji et al. (2003)
Denali (2002)	7.8	1	29	30	Tsuboi et al. (2003)
Niigataken-Chuetsu (2004)	6.6	9	58	67	Honda et al. (2005)
Total		100	595	695	

Table 3.1: Number of Near-source and Far-source Records in Earthquake Dataset Used for Classification. (^{a} moment magnitude M_w is cited from Havard CMT solution and ^{b} listed fault models are utilized to classify near-source and far-source station.)

Ground Motion Feature	Unit
Horizontal Peak Ground Jerk (H_j)	(cm/s^3)
Vertical Peak Ground Jerk (Z_j)	(cm/s^3)
Horizontal Peak Ground Acceleration (H_a)	(cm/s^2)
Vertical Peak Ground Acceleration (Z_a)	(cm/s^2)
Horizontal Peak Ground Velocity (H_v)	(cm/s)
Vertical Peak Ground Velocity (Z_v)	(cm/s)
Horizontal Peak Ground Displacement (H_d)	(cm)
Vertical Peak Ground Jerk (Z_d)	(cm)

Table 3.2: Eight Extracted Features.

where H and Z mean the peak horizontal and vertical components and j, a, v and d stand for jerk, acceleration, velocity, and displacement, respectively. The same dataset of feature vectors is also used in Yamada et al. (2007) where a Bayesian classification scheme was also applied, but an automatic relevance determination prior was not used to select which of these features were most relevant for near-source versus far-source classification.

3.1.2 Separating Boundary Model

For NS and FS classification, the extracted features $\underline{x} \in \mathbb{R}^8$ defined in Table 3.2 are used with label $y \in \{0, 1\}$ ($y = 0$ for far-source data, $y = 1$ for near-source data), and the theory for Bayesian learning for classification that is presented in Section

2.3 is applied to the earthquake dataset with the number of data points $N = 695$. The separating boundary between the two classes is taken as a linear combination of the logarithms of the features $\underline{x} = \{x_1, ..., x_8\}^T$ with unknown coefficients $\underline{\theta} = \{\theta_0,\ \theta_1,\ ...,\ \theta_8\}^T \in \mathbb{R}^9$:

$$f(\underline{x}|\underline{\theta}) = \sum_{j=1}^{8} \theta_j g_j(\underline{x}) + \theta_0 \tag{3.2}$$

The separating boundary is defined by $f(\underline{x}|\underline{\theta}) = 0$, and probabilistic predictions of the class label $y \in \{0, 1\}$ corresponding to extracted features $\underline{x}$ are based on the probability model:

$$P(y|\underline{x}, \underline{\theta}) = \phi(f(\underline{x}|\underline{\theta}))^y \{1 - \phi(f(\underline{x}|\underline{\theta}))\}^{1-y} \tag{3.3}$$

where the sigmoid function $\phi(\cdot) \in [0, 1]$ is defined as before in Section 2.3.

3.2 Comparison of Results

3.2.1 Function for Separating Boundary

In a previous study that used a fixed prior (instead of the ARD prior), the three-parameter model given in (3.4) was found to give the optimal separating boundary function based on the earthquake dataset described in Section 3.1.1 (Yamada et al., 2007):

$$\mathcal{M}_1 :\ f(\underline{x}|\hat{\underline{\theta}}) = 6.046 \log_{10} Z_a + 7.885 \log_{10} H_v - 27.091 \tag{3.4}$$

This corresponds to a model class, denoted $\mathcal{M}_1$ here, that was selected by finding the most probable model class among 255 ($=2^8 - 1$) models consisting of all possible combinations of the 8 features in Table 3.2 and using a fixed Gaussian prior $p(\underline{\theta}|\mathcal{M})$ for each model class $\mathcal{M}$. For $\mathcal{M}_1$, the Gaussian prior was selected to have the same standard deviation of $\alpha_i^{-\frac{1}{2}} = 100$ for each coefficient θ_i. The misclassification rates for $\mathcal{M}_1$ are 22.00% and 2.02% for the NS and FS data, respectively.

The proposed method of Bayesian learning is first applied here to a model class

with the same features as in (3.4) but using the ARD prior which has an *independent* variance α_i^{-1} for each θ_i $(i = 0, 1, 2)$. Model class selection is used to determine the optimal model class as described before (the corresponding optimal prior variances are given later). The optimal boundary function for this model class $\mathcal{M}_2$ is given in (3.5):

$$\mathcal{M}_2 : \; f(\underline{x}|\hat{\underline{\theta}}) = 6.129 \log_{10} Z_a + 7.484 \log_{10} H_v - 26.588 \tag{3.5}$$

The corresponding misclassification rates are 23.00% and 2.02% for NS and FS data, respectively. Note that $\mathcal{M}_2$ is the most probable model class based on the earthquake dataset that is restricted to contain only Z_a and H_j in the boundary function; it does not necessarily optimize misclassification rates.

Based on the misclassification rates, it could be concluded that the difference in performance between the two three-parameter models (3.4) and (3.5) is negligible. However, it is shown later that $\mathcal{M}_1$ is actually less probable than $\mathcal{M}_2$, based on the data.

Finally, the proposed methodology of Bayesian learning using the ARD prior is applied to models containing all 8 features in Table 3.2. This produces a five-parameter model class $\mathcal{M}_3$ whose optimal separating boundary function is:

$$\begin{aligned} \mathcal{M}_3 : \; f(\underline{x}|\hat{\underline{\theta}}) &= 2.055 \log_{10} H_j + 5.350 \log_{10} Z_a + 4.630 \log_{10} H_v \\ &+ 1.972 \log_{10} H_d - 30.982 \end{aligned} \tag{3.6}$$

Note that for $\mathcal{M}_2$, the Bayesian learning algorithm is restricted to have no more than $\log_{10} Z_a$ and $\log_{10} H_v$, the features that are used in the previous model class $\mathcal{M}_1$, while $\mathcal{M}_3$ actually selects 2 additional features (giving a total of 4 features selected from a potential of 8 by automatically pruning irrelevant features). The corresponding misclassification rates for $\mathcal{M}_3$ are 18.00% and 1.85% for NS and FS data, respectively, much smaller than those for $\mathcal{M}_1$ and $\mathcal{M}_2$.

The coefficients for the optimal separating boundaries, the optimal prior variances and the corresponding classification results for each model class are summarized in

$\mathcal{M}$	^aN_i	1	H_j	Z_j	H_a	Z_a	H_v	Z_v	H_d	Z_d
$\mathcal{M}_1$	3	-27.091	^{b}–	–	–	6.046	7.885	–	–	–
$\mathcal{M}_2$	3	-26.588	–	–	–	6.129	7.484	–	–	–
$\mathcal{M}_3$	5	-30.982	2.055	^{c}0	0	5.350	4.630	0	1.972	0

Table 3.3: Coefficients for Optimal Separating Boundary Function for Each Model Class. (^aN_i is the number of parameters used for each model. ^{b}– means the corresponding parameters are not considered for each model. ^{c}0 means the corresponding parameters are automatically pruned during training.)

$\mathcal{M}$	Prior Covariance Matrix
$\mathcal{M}_1$	diag$^a(100^2, 100^2, 100^2)$
$\mathcal{M}_2$	diag$(26.76^2, 6.19^2, 7.57^2)$
$\mathcal{M}_3$	diag$(31.23^2, 2.25^2, 5.48^2, 4.97^2, 2.20^2)$

Table 3.4: Prior Covariance Matrix for Each Model Class. (^{a}diag means diagonal matrix with the diagonal elements following.)

	Actual Class	Predicted Class		Total Observations
		Near-source	Far-source	
$\mathcal{M}_1$	Near-source	78(78.00%)	22(22.00%)	100
	Far-source	12(2.02%)	583(97.98%)	595
	Total Predictions	90	605	695
$\mathcal{M}_2$	Near-source	77(77.00%)	23(23.00%)	100
	Far-source	12(2.02%)	583(97.98%)	595
	Total Predictions	89	606	695
$\mathcal{M}_3$	Near-source	82(82.00%)	**18(18.00%)**	100
	Far-source	**11(1.85%)**	584(98.15%)	595
	Total Predictions	93	602	695

Table 3.5: Classification Results for Earthquake Database Using Three Different Model Classes. (Bold values represent the least misclassification rate.)

Tables 3.3, 3.4, and 3.5, respectively. The performance of these three model classes are next examined by leave-one-out cross-validation and then by calculating their probability based on the earthquake data $\mathcal{D}_N$.

3.2.2 Leave-One-Out Cross-Validation

Table 3.5 shows the classification results for models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ using all 695 records in the earthquake dataset, both for training and predicting the labels. As shown in this table, $\mathcal{M}_3$ outperforms the two other models on the basis of smaller

misclassification rates. For another check on the performance of these three models for predicting the class, leave-one-out cross-validation (LOOCV) is performed.

LOOCV, as the name implies, takes one data point at a time from the whole dataset and then a prediction is made based on the optimal separating boundary determined from the remaining data. This procedure is repeated until each data point has been compared with the prediction (taken here as the class with the higher predictive probability). Actually, LOOCV is equivalent to K-fold cross-validation where K(= 695 here) is equal to the number of data in the original dataset. Note that LOOCV is commonly used in Tikhonov regularization to select the regularizing parameter, but this is handled automatically in the Bayesian approach presented here.

The results of LOOCV for each model class are presented in Table 3.6. Based on the misclassification rate, which is the ratio of the number of misclassified data to the total number of data, classification model $\mathcal{M}_3$ shows a better performance.

Model	Prediction Error
$\mathcal{M}_1$	36/695 (5.18%)
$\mathcal{M}_2$	37/695 (5.32%)
$\mathcal{M}_3$	**31/695 (4.46%)**

Table 3.6: Misclassification Rates Based on Leave-One-Out Cross-Validation.

3.2.3 Posterior Probability of Each Model Class

In this section the posterior probability of each model class in the set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\}$ is computed based on the dataset $\mathcal{D}_N$ of 695 records:

$$\begin{aligned} P(\mathcal{M}_i|\mathcal{D}_N, \mathcal{M}) &= \frac{P(\mathcal{D}_N|\mathcal{M}_i)P(\mathcal{M}_i|\mathcal{M})}{P(\mathcal{D}_N|\mathcal{M})} \\ &= \frac{P(\mathcal{D}_N|\mathcal{M}_i)P(\mathcal{M}_i|\mathcal{M})}{\sum_{i=1}^{I} P(\mathcal{D}_N|\mathcal{M}_i)P(\mathcal{M}_i|\mathcal{M})} \end{aligned} \tag{3.7}$$

where $P(\mathcal{D}_N|\mathcal{M}_i)$ is the evidence for $\mathcal{M}_i$, $P(\mathcal{M}_i|\mathcal{M})$ is the prior reflecting the initial choice of the probability of each model class in set $\mathcal{M}$, and the denominator $P(\mathcal{D}_N|\mathcal{M})$ is a normalizing constant. Assigning equal prior probability to each model class, the

posterior probability of each model class is proportional to its evidence, that is:

$$P(\mathcal{M}_i|\mathcal{D}_N, \mathcal{M}) \propto P(\mathcal{D}_N|\mathcal{M}_i) \tag{3.8}$$

Using the Theorem of Total Probability, the evidence is calculated from:

$$P(\mathcal{D}_N|\mathcal{M}_i) = \int P(\mathcal{D}_N|\underline{\theta}_i, \mathcal{M}_i) p(\underline{\theta}_i|\mathcal{M}_i) d\underline{\theta}_i \tag{3.9}$$

This is the average value of the likelihood weighted by the prior over all possible values of the parameters $\underline{\theta}_i$. For a large number of data, an asymptotic approximation can be applied to the integral in (3.9) (Beck and Yuen, 2004):

$$P(\mathcal{D}_N|\mathcal{M}_i) \cong P(\mathcal{D}_N|\hat{\underline{\theta}}_i, \mathcal{M}_i) \frac{2\pi^{N_i/2} p(\hat{\underline{\theta}}_i|\mathcal{M}_i)}{\sqrt{|H(\hat{\underline{\theta}}_i)|}} \tag{3.10}$$

where $\hat{\underline{\theta}}_i$ is the most probable value of $\underline{\theta}_i$ and N_i is the number of parameters in model class $\mathcal{M}_i$. The first factor in (3.10) is the likelihood, and the remaining factors together are the *Ockham factor*. This Ockham factor penalizes more complex models. The Hessian matrix $H(\underline{\theta}_i)$ in (3.10) is given by the same expression as for $\hat{\Sigma}^{-1}(\underline{\alpha})$ after (2.49) where each variance α_i^{-1} is given in Table 3.4. The posterior probabilities for each of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are presented in Table 3.7, which shows that $\mathcal{M}_3$ is much more probable than $\mathcal{M}_1$ and $\mathcal{M}_2$ based on the dataset $\mathcal{D}_N$.

There is a nice information-theoretic interpretation (Beck and Yuen, 2004, Muto and Beck, 2007) of the log evidence that shows that it consists of the difference between a datafit term (the posterior mean of the log likelihood function for the model class) and a relative entropy term which quantifies the amount of information (in the sense of Shannon) extracted from the data by the model class. It is the latter term that prevents over-fitting to the data and which leads to an automatic Principle of Model Parsimony (Beck and Yuen, 2004) when Bayesian updating is performed over a set of model classes, as done here. This information-theoretic interpretation is evident from the asymptotic approximation (3.10) for large N which shows that the log evidence

$\mathcal{M}$	ln Ockham[a]	ln Likelihood[a]	ln Evidence[a]	Probability[b]
$\mathcal{M}_1$	-15	-81	-96	0.00
$\mathcal{M}_2$	-10	-79	-89	0.11
$\mathcal{M}_3$	-12	-75	-87	0.89

Table 3.7: Posterior Probability Calculation for Bayesian Model Class Selection. ([a]ln Ockham, ln Likelihood, and ln Evidence are natural logarithms of the Ockham factor, likelihood, and evidence, respectively. [b]Probability is calculated from the evidence on the basis that the $\mathcal{M}_i$ $(i = 1, 2, 3)$ are equally probable a priori.)

is approximated by the sum of the log likelihood of the most probable model in the model class and the log Ockham factor, which is an asymptotic approximation for the negative of the relative entropy. This is how it was originally discovered (Beck and Yuen, 2004) but more recently it has been proved for the general case (Muto and Beck, 2007).

3.2.4 Effect of Prior

It seems obvious that the likelihood calculated by using the more complex model $\mathcal{M}_3$ compared with $\mathcal{M}_1$ should be larger. However, the ARD prior leads to a larger Ockham factor for $\mathcal{M}_3$ than $\mathcal{M}_1$, resulting in the higher evidence and higher posterior probability for the more complex model class. Evidently, this difference in the Ockham factor comes mostly from the different priors, as shown by the breakdown of the log Ockham factor into its three terms in Table 3.8, where:

$$\ln Ockham = \frac{N_i}{2}\ln(2\pi) + \ln p(\hat{\underline{\theta}}_i|\mathcal{M}_i) - \frac{1}{2}\ln|H(\hat{\underline{\theta}}_i)| \tag{3.11}$$

$\mathcal{M}$	$\frac{N_i}{2}\ln(2\pi)$	$\ln p(\hat{\underline{\theta}}_i\|\mathcal{M}_i)$	$-\frac{1}{2}\ln\|H(\hat{\underline{\theta}}_i)\|$	ln Ockham
$\mathcal{M}_1$	2.7568	-16.6140	-1.5514	-15.4085
$\mathcal{M}_2$	2.7568	-11.3634	-1.6298	-10.2363
$\mathcal{M}_3$	4.5947	-15.1611	-1.7572	-12.3237

Table 3.8: Components of ln Ockham Factor in (3.11).

3.3 Conclusions

A novel methodology of Bayesian learning using the automatic relevance determination (ARD) prior is applied to classify measured earthquake motions into near-source and far-source. The extracted features in the training dataset correspond to the $\log_{10}$ values of peak jerk, acceleration, velocity, and displacement in the horizontal and vertical directions from 695 earthquake records, and these data are used with Bayesian learning to establish a separating boundary in the feature space. The ARD prior plays an important role by promoting sparsity when selecting the important features (i.e., by utilizing only a small number of relevant features after automatically pruning the remaining features). The selected best separating boundary for classification of seismic signals into near-source and far-source is:

$$\begin{aligned} f(\underline{x}_i|\hat{\underline{\theta}}) &= 2.055\log_{10} H_j + 5.350\log_{10} Z_a + 4.630\log_{10} H_v \\ &+ 1.972\log_{10} H_d - 30.982 \end{aligned} \tag{3.12}$$

where H_j, Z_a, H_v and H_d are the horizontal jerk, vertical acceleration, horizontal velocity, and horizontal displacement, respectively, of the ground motion record. Based on (3.12), the probability for new data with features $\underline{\tilde{x}}$ to be classified as near-source ($\tilde{y} = 1$) or far-source ($\tilde{y} = 0$) is:

$$P(\tilde{y} = 1|\underline{\tilde{x}}, \hat{\underline{\theta}}) = \frac{1}{1 + \exp(-f(\underline{\tilde{x}}|\hat{\underline{\theta}}))} \tag{3.13}$$

$$P(\tilde{y} = 0|\underline{\tilde{x}}, \hat{\underline{\theta}}) = 1 - P(\tilde{y} = 1|\underline{\tilde{x}}, \hat{\underline{\theta}}) \tag{3.14}$$

In view of the results so far achieved, it can be concluded that it is beneficial to use the proposed Bayesian learning using the ARD prior because it leads to:

- higher correct classification rates (equivalent to a lower misclassification rate) (see Table 3.5);
- better generalization performance, as demonstrated by the leave-one-out cross-validation results (see Table 3.6);

- the most probable model class based on the calculated posterior probability (see Table 3.7).

The proposed method is readily applied to real-time analysis of recorded seismic ground motions for near-source and far-source classification, since the only calculations involved are those implied by (3.12) to (3.14).

Chapter 4

Ground Motion Attenuation Relations (Ground Motion Prediction Equations) using Regression

Modeling the attenuation of ground shaking intensity measures such as peak ground acceleration (PGA) or response spectral ordinates is essential for seismic hazard analysis. The attenuation model can be used for earthquake-resistant design purposes, as well as for the inversion problem which deals with estimating the size and location of an earthquake event.

Least-squares regression analysis that minimizes the Euclidean norm of the errors between model and data has often been performed to estimate the unknown parameters in a prescribed mathematical form for the attenuation equation (e.g., Boore et al., 1993 and 1997). However, it is well known that prediction errors may be larger than the data-fit errors due to over-fitting of the data; a more complex model with more unknown parameters may fit the given data better, but it may result in poor future predictions. For a well-known ground motion attenuation equation (Boore et al., 1997), Bayesian model class selection was performed and it is concluded that all the proposed input variables for that equation may not be necessary (Muto, 2006).

In this chapter, the procedure of Bayesian Learning using the ARD prior is presented to identify a probabilistic attenuation model for PGA from recorded ground motions and the results are compared to a previously-developed stochastic simulation

method for identifying a predictive attenuation equation for PGA.

4.1 Estimation of Earthquake Ground Motion

4.1.1 Earthquake Data

The earthquake database to identify the probabilistic attenuation model is presented in Table 4.3 at the end of this chapter. Data collected from 271 strong motion records in 20 earthquakes are utilized (Boore et al., 1997). For peak accelerations, the geometric mean value of two orthogonal horizontal components is used rather than their maximum, since it represents a more stable peak acceleration parameter (Campbell, 1981). The magnitudes of the earthquakes are equal to or greater than 5.0, representing events which are of most concern in earthquake-resistant design.

4.1.2 Boore-Joyner Attenuation Model

There is a well-known regression equation for the peak ground acceleration from an earthquake (e.g., Boore, et al., 1993):

$$\log(PGA) = b_1 + b_2(M-6) + b_3(M-6)^2 + b_4 R + b_5 \log R + b_6 G_B + b_7 G_C + \sigma\epsilon \quad (4.1)$$

where M is the magnitude of an earthquake, $R = \sqrt{d^2 + h^2}$ (called the fault distance), d is the closest horizontal distance in km from the site to a point on the surface that lies directly above the rupture, h is a fictitious depth parameter introduced to be representative of a regional event, G_B, $G_C \in \{0, 1\}$ are binary soil classification parameters in Table 4.1, and ϵ is the uncertain prediction error which is modeled as a Gaussian variable with mean zero and standard deviation unity. Boore et al. (1993) suggested $h = 5.57$ for California earthquakes. The model parameters are σ, h and $\underline{b} = \{b_1, ..., b_7\}^T$.

In Muto (2006), Bayesian Model Class Selection was performed to find out the most probable model class by using several methods to evaluate the necessary quan-

tities: Laplace's asymptotic approximation and three stochastic simulation methods: Gibbs sampler, Metropolis-Hastings Algorithm, and the newly developed Transient Markov Chain Monte Carlo Method (Ching and Chen, 2006). For h, the fictitious depth, which is non-linearly involved in the regression equation, the most plausible value was estimated by maximizing its posterior probability distribution based on the data, instead of using the given value of h by Boore el al. (1993).

Site Class	Range of Shear Velocity
A	greater than 750 m/s
B	360 m/s to 750 m/s
C	180 m/s to 360 m/s
D	less than 180 m/s

Table 4.1: Definition of Site Classes. (G_B or G_C is 1 if a site is classified in class B or C, respectively, and 0 otherwise.)

4.1.3 Training Phase for PGA Estimation

Let us assume first that h is given, then the parameters and the function form can be defined as

$$\begin{aligned} y &= \log(PGA) \\ &= b_1 + b_2(M-6) + b_3(M-6)^2 + b_4 R + b_5 \log R + b_6(G_B) + b_7(G_C) + \sigma\epsilon \\ &= \underline{\tau}(\underline{x}, h)^T \underline{b} + \sigma\epsilon = f(\underline{x}|\underline{b}, h) + \sigma\epsilon \end{aligned} \tag{4.2}$$

where $\underline{x} = \{M, d, G_B, G_C\}^T$, $\underline{b} = \{b_1, b_2, b_3, b_4, b_5, b_6, b_7\}^T$, $\underline{\tau}(\underline{x}, h) = [1, M-6, (M-6)^2, R(h), \log R(h), (G_B), (G_C)]^T$, and $R = \sqrt{d^2 + h^2}$. Thus, y is Gaussian, $p(y|\underline{x}, \underline{b}, \sigma^2, h) = \mathcal{N}(y|f(\underline{x}|\underline{b}, h), \sigma^2)$, given $\underline{b}$, σ^2, and h.

The likelihood for the given earthquake dataset $\mathcal{D}_N = \{(\underline{x}_i, y_i) : i = 1, ..., N\} = (\mathbf{X}, \underline{y})$ with $\underline{x}_i = \{M_i, d_i, (G_B)_i, (G_C)_i\}^T \in \mathbb{R}^4$ and $y_i = \{\log(PGA)_i\} \in \mathbb{R}$ is:

$$p(\mathcal{D}_N|\underline{b}, \sigma^2, h) = \mathcal{N}(\mathbf{\Phi}\underline{b}, \sigma^2\mathbf{I}) = (2\pi\sigma^2)^{-N/2} \exp\Big[-\frac{1}{2\sigma^2}\|\underline{y} - \mathbf{\Phi}\underline{b}\|^2\Big] \tag{4.3}$$

where $\mathbf{\Phi}(h) = [\underline{\tau}(\underline{x}_1, h), ..., \underline{\tau}(\underline{x}_N, h)]^T \in \mathbb{R}^{N\times 7}$. Define the ARD prior PDF as before,

that is, $p(\underline{b}|\underline{\alpha},\sigma^2,h) = p(\underline{b}|\underline{\alpha}) = \mathcal{N}(\underline{0},\mathbf{A}^{-1}(\underline{\alpha}))$ so $\underline{b}$ is Gaussian with mean $\underline{0}$ and covariance matrix $\mathbf{A}^{-1}(\underline{\alpha})$.

Then the posterior PDF for the unknown parameters $\underline{b}$ can be calculated via Bayes' theorem:

$$\begin{aligned} p(\underline{b}|\mathcal{D}_N,\underline{\alpha},\sigma^2,h) &= \frac{p(\mathcal{D}_N|\underline{b},\underline{\alpha},\sigma^2,h)p(\underline{b}|\underline{\alpha},\sigma^2,h)}{p(\mathcal{D}_N|\underline{\alpha},\sigma^2,h)} \\ &= (2\pi)^{-7/2}|\hat{\mathbf{\Sigma}}|^{-1/2}\exp\Big[-\frac{1}{2}(\underline{b}-\hat{\underline{b}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{b}-\hat{\underline{b}})\Big] \end{aligned} \tag{4.4}$$

where $\hat{\mathbf{\Sigma}} = (\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}+\mathbf{A})^{-1}$, $\hat{\underline{b}} = \sigma^{-2}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\underline{y}$, and $\hat{\sigma}^2 = \frac{\|\underline{y}-\mathbf{\Phi}\hat{\underline{b}}\|^2}{N-\Sigma_i\gamma_i}$ as before in Section 2.2.1. Note that $\hat{\mathbf{\Sigma}}$ and $\hat{\underline{b}}$ depend on $\underline{\alpha}$, σ^2, and h.

4.1.4 Posterior Robust Predictive Probability Distribution for PGA

Let $\tilde{y}$ denote the unknown $\log(PGA)$ calculated from $\underline{\tilde{x}} = \{\tilde{M},\tilde{d},\tilde{G}_B\tilde{G}_C\}^T$, then the desired posterior robust predictive probability distribution for PGA is given by the Theorem of Total Probability where Laplace's approximation is used to evaluate the integral over $\underline{\alpha}$, σ^2, and h (i.e., $\underline{\alpha}$, σ^2, and h define the model class for y which covers all possible values of $\underline{b}$, where the latter is treated analytically since it is Gaussian given $\underline{\alpha}$, σ^2, and h):

$$\begin{aligned} p(\tilde{y}|\underline{\tilde{x}},\mathcal{D}_N) &= \int p(\tilde{y},\underline{b},\underline{\alpha},\sigma^2,h|\underline{\tilde{x}},\mathcal{D}_N)d\underline{b}d\underline{\alpha}d\sigma^2dh \\ &= \int p(\tilde{y}|\underline{\tilde{x}},\mathcal{D}_N,\underline{b},\sigma^2,h)p(\underline{b}|\mathcal{D}_N,\underline{\alpha},\sigma^2,h)p(\underline{\alpha},\sigma^2,h|\mathcal{D}_N)d\underline{b}d\underline{\alpha}d\sigma^2dh \\ &\cong \int p(\tilde{y}|\underline{\tilde{x}},\underline{b},\hat{\sigma}^2,\hat{h})p(\underline{b}|\mathcal{D}_N,\hat{\underline{\alpha}},\hat{\sigma}^2,\hat{h})d\underline{b} \\ &= \mathcal{N}(\tilde{y}|y_*,\sigma_*^2) \end{aligned} \tag{4.5}$$

where $y_* = \hat{\underline{b}}^T\underline{\tau}(\underline{\tilde{x}},\hat{h})$, $\sigma_*^2 = \hat{\sigma}^2 + \underline{\tau}(\underline{\tilde{x}},\hat{h})^T\hat{\mathbf{\Sigma}}\underline{\tau}(\underline{\tilde{x}},\hat{h})$ and $\hat{\underline{\alpha}}$, $\hat{\sigma}^2$, and $\hat{h}$ are the most plausible values which maximize the evidence $p(\mathcal{D}_N|\underline{\alpha},\sigma^2,h)$. In (4.5), $\hat{\underline{\alpha}}$ and $\hat{\sigma}^2$ are determined as in Section 2.2 where a non-informative prior on the parameters

defining the model class is chosen. Then the most plausible fictitious depth $\hat{h}$ can be determined by maximizing the evidence of h:

$$h = \arg\max_{h} \; p(\mathcal{D}_N|\underline{\hat{\alpha}}(h), \hat{\sigma}^2(h), h) \tag{4.6}$$

4.1.5 Estimation of a Non-linearly Involved Parameter

4.1.5.1 Inclusion of the Fictitious Depth Defining Model Class

As described in the previous Section 4.1.4, a non-linearly involved parameter h can be estimated by considering it to define model class along with $\underline{\alpha}$ and σ^2. The estimated regression coefficients $\underline{b}$, σ, and h are in Table 4.2, Case 2.

4.1.5.2 Estimating the Fictitious Depth by Stochastic Simulation

The Metropolis-Hastings algorithm (a Markov Chain Monte Carlo (MCMC) simulation method (Martinez and Martinez, 2002)) is also applied for estimating the fictitious depth, h, instead of using Bayesian Model Class Selection, as above.

This method consists of two steps. In the first step, the Bayesian learning algorithm using the ARD prior is applied to estimate the regression coefficients $\underline{b}$ for given h; then the MCMC algorithm was applied to generate samples from the posterior PDF $p(h|\underline{b}, \mathcal{D}_N, \underline{\hat{\alpha}}, \hat{\sigma}^2)$ based on σ^2 and the regression coefficients $\underline{b}$ estimated in the previous step. A lognormal distribution for h is used as a proposal distribution in the Metropolis-Hastings algorithm with the mean and standard deviation estimated from the depths for 12 earthquakes.

The generated samples of fictitious depth h and the corresponding estimated regression coefficients $\underline{b}$ is shown in Figure 4.1 as a function of the number of generated samples of h. The moving averages for each parameter are also shown in Figure 4.2 with dotted line representing the standard deviation. The burn-in period for MCMC is about 30 samples. The plot shows around 300 samples are enough for obtaining the converged value of h, and for the comparison in the next section, the mean of 500 samples of h is taken. The estimated regression coefficients and h are in Table

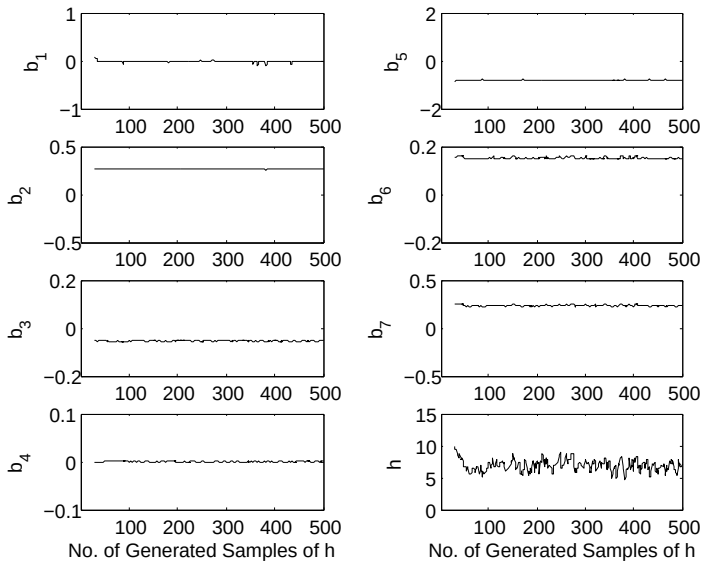


Figure 4.1: Estimated Values of Regression Coefficients and Samples of Fictitious Depth during MCMC Algorithm.

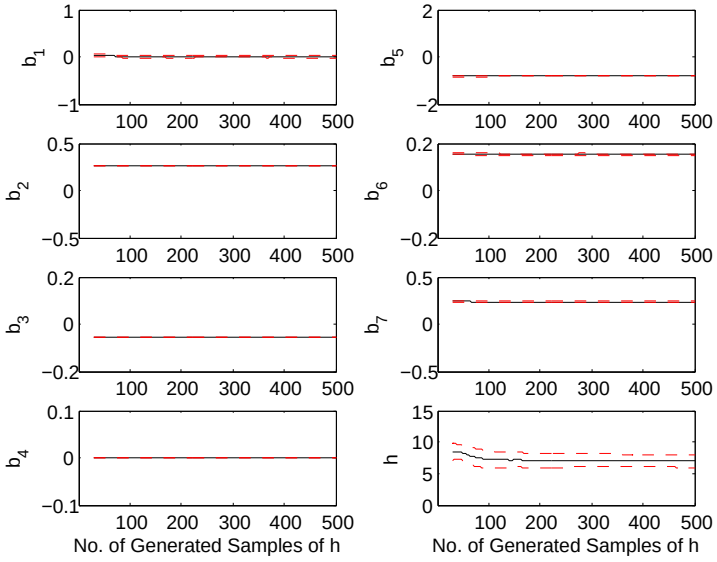


Figure 4.2: Mean and Standard Deviation of Regression Coefficients and the Generated Fictitious Depth during MCMC Algorithm.

4.2, Case 3. Compared with Case 1 (Bayesian Model Class Selection by Maximizing Evidence) and even Case 2 (result in the previous section), the proposed method is proven as effective in estimating a non-linearly involved parameter h.

4.2 Comparison of Results

In this section, the results from the various methods are compared. Table 4.2 shows the comparison results. Case 1 shows the mean parameter estimates by utilizing Transitional Markov Chain Monte Carlo simulation (Muto, 2006) and a fixed prior on the model parameters. The results for Case 2 and 3 are obtained by using (4.6) and MCMC for h, respectively, along with Bayesian learning using the ARD prior for the regression coefficients. Dashes (i.e., –) for b_1, b_3, and b_4 in Case 1 represent that corresponding coefficients are not considered for that method and zeros for b_1 and b_4 (Case 2 and 3) stand for pruning of the corresponding parameters by the ARD prior, not excluding them beforehand.

Case	b_1	b_2	b_3	b_4	b_5	b_6	b_7	h	σ
1	–	0.230	–	–	-0.834	0.164	0.256	6.78	0.197
2	0.000	0.260	-0.053	0.000	-0.804	0.150	0.231	6.43	0.194
3	0.000	0.260	-0.054	0.000	-0.810	0.152	0.233	6.98	0.194

Table 4.2: Comparison Results for Parameter Estimation.

The Bayesian learning method using the ARD prior is efficient in computational cost. For Case 1, the most probable model class is selected using Bayesian model class selection from considering all $127 = {}_7C_1 + {}_7C_2 + {}_7C_3 + {}_7C_4 + {}_7C_5 + {}_7C_6 + {}_7C_7 = 2^7 - 1$ possibilities. The proposed method, however, automatically selects the most plausible model class based on the given data, which results in smaller errors in shorter time (or results in much less computational effort). The improvement in efficiency comes from being able to use a continuous-variable optimization algorithm on $\underline{\alpha}$, i.e., to perform continuous model class selection, rather than a discrete optimization over all 127 possible model classes. Therefore, it is beneficial to select continuous model class selection, especially using the ARD prior which assigns one hyperparameter per

input that controls the contribution of that corresponding input.

In contrast to Case 1, the results from the proposed method show the existence of an additional term with coefficient b_3 (see Table 4.2). The small *negative* values of b_3 seems to suggest that some saturation of PGA is expected with magnitude, but more research is needed with a dataset that contains more large earthquake recordings. This term has been adopted to provide better fits for longer period ground motions for the 5% response spectral ordinates where saturation with magnitude appears to be more pronounced (Boore et al., 1997).

4.3 Conclusions

The Bayesian learning method using the ARD prior is shown to be an efficient tool for estimating unknown parameters in the peak ground attenuation equation; more generally, this work suggests that it should be a reliable and promising estimation tool for function estimation from data.

The proposed method has several advantages:

(1) It performs automatic model class selection by optimizing the hyperparameters, which reduces computational effort greatly compared with studying all 127 model classes as done in discrete optimization over all model classes in previous work.

(2) It provides robust probabilistic estimation by considering all forms of uncertainty.

There has been much previous work on the estimation of earthquake ground motion attenuation equations; recently the next generation attenuation relationships have been studied (*http://peer.berkeley.edu/nga/index.html*). In this situation, the Bayesian learning method using the ARD prior should prove to be an effective tool, having the ability to choose the most plausible model based on an earthquake dataset. In this approach, one selects a flexible model containing all terms thought to influence the ground shaking at a site, and then lets the algorithm automatically prune irrelevant terms by using the ARD prior and the earthquake dataset.

Table 4.3: Earthquake Records Used.

Earthquake	Magnitude(M)	Distance(d)	Site Class	PA_H1	PA_H2
Imperial Valley 1	7.00	12.0	C	0.359	0.224
Kern County	7.40	42.0	B	0.196	0.177
Kern County	7.40	85.0	B	0.135	0.090
Kern County	7.40	109.0	B	0.054	0.048
Kern County	7.40	107.0	C	0.062	0.044
Daly City	5.30	8.0	A	0.127	0.105
Parkfield	6.10	16.1	B	0.411	0.282
Parkfield	6.10	17.3	B	0.072	0.066
Parkfield	6.10	6.6	C	0.509	
Parkfield	6.10	9.3	C	0.467	0.403
Parkfield	6.10	13.0	C	0.279	0.276
Borrego Mountain	6.60	45.0	C	0.142	0.061
San Fernando	6.60	17.0	B	0.374	0.288
San Fernando	6.60	25.7	B	0.114	0.103
San Fernando	6.60	60.7	B	0.057	0.047
San Fernando	6.60	19.6	C	0.200	0.159
Sitka	7.70	45.0	A	0.110	0.090
Managua	6.20	5.0	C	0.390	0.330
Point Mugu	5.60	16.0	C	0.130	0.080
Hollister	5.20	17.0	A	0.011	0.008
Hollister	5.20	8.0	B	0.120	0.050
Hollister	5.20	10.0	B	0.140	0.100
Hollister	5.20	10.0	C	0.170	0.100
Santa Barbara	5.87	0.0	B	0.210	0.100
Santa Barbara	5.87	11.0	B	0.390	0.240
Santa Barbara	5.87	14.0	B	0.280	0.240
St. Elias	7.60	25.4	B	0.160	0.110
Coyote Lake	5.80	9.1	A	0.127	0.100
Coyote Lake	5.80	1.2	B	0.419	0.344
Coyote Lake	5.80	17.9	B	0.110	0.090
Coyote Lake	5.80	19.2	B	0.120	0.080
Coyote Lake	5.80	30.0	B	0.044	0.040
Coyote Lake	5.80	3.7	C	0.257	0.236
Coyote Lake	5.80	5.3	C	0.267	0.260
Coyote Lake	5.80	7.4	C	0.263	0.196
Imperial Valley 2	6.50	14.0	B	0.200	0.110
Imperial Valley 2	6.50	23.5	B	0.167	0.149
Imperial Valley 2	6.50	26.0	B	0.210	0.120
Imperial Valley 2	6.50	0.5	C	0.320	0.300

Earthquake	Magnitude(M)	Distance(d)	Site Class	PA_H1	PA_H2
Imperial Valley 2	6.50	0.6	C	0.520	0.360
Imperial Valley 2	6.50	1.3	C	0.720	0.450
Imperial Valley 2	6.50	1.4	C	0.316	0.240
Imperial Valley 2	6.50	2.6	C	0.810	0.660
Imperial Valley 2	6.50	3.8	C	0.640	0.500
Imperial Valley 2	6.50	4.0	C	0.560	0.400
Imperial Valley 2	6.50	5.1	C	0.510	0.370
Imperial Valley 2	6.50	6.2	C	0.400	0.270
Imperial Valley 2	6.50	6.8	C	0.610	0.380
Imperial Valley 2	6.50	7.5	C	0.260	0.220
Imperial Valley 2	6.50	7.6	C	0.240	0.240
Imperial Valley 2	6.50	8.4	C	0.459	0.311
Imperial Valley 2	6.50	8.5	C	0.230	0.200
Imperial Valley 2	6.50	8.5	C	0.220	0.170
Imperial Valley 2	6.50	10.6	C	0.280	0.220
Imperial Valley 2	6.50	12.6	C	0.380	0.380
Imperial Valley 2	6.50	12.9	C	0.310	
Imperial Valley 2	6.50	15.0	C	0.110	0.080
Imperial Valley 2	6.50	16.0	C	0.430	0.330
Imperial Valley 2	6.50	17.7	C	0.267	0.263
Imperial Valley 2	6.50	18.0	C	0.150	0.110
Imperial Valley 2	6.50	22.0	C	0.150	0.150
Imperial Valley 2	6.50	22.0	C	0.150	0.120
Imperial Valley 2	6.50	23.0	C	0.130	0.086
Imperial Valley 2	6.50	23.2	C	0.188	0.149
Imperial Valley 2	6.50	32.0	C	0.066	0.049
Imperial Valley 2	6.50	32.7	C	0.349	0.235
Imperial Valley 2	6.50	36.0	C	0.100	0.070
Imperial Valley 2	6.50	43.5	C	0.163	0.122
Imperial Valley 2	6.50	49.0	C	0.140	0.110
Imperial Valley 2	6.50	60.0	C	0.049	0.043
Livermore Valley 1	5.80	20.8	B	0.045	0.010
Livermore Valley 1	5.80	33.1	B	0.056	0.050
Livermore Valley 1	5.80	40.3	B	0.065	0.060
Livermore Valley 1	5.80	15.7	C	0.154	0.060
Livermore Valley 1	5.80	16.7	C	0.052	0.040
Livermore Valley 1	5.80	28.5	C	0.086	0.050
Livermore Valley 2	5.50	10.1	B	0.267	0.190
Livermore Valley 2	5.50	26.5	B	0.026	0.030

Earthquake	Magnitude(M)	Distance(d)	Site Class	PA_H1	PA_H2
Livermore Valley 2	5.50	29.0	B	0.039	
Livermore Valley 2	5.50	30.9	B	0.112	0.050
Livermore Valley 2	5.50	37.8	B	0.065	0.040
Livermore Valley 2	5.50	4.0	C	0.259	0.220
Livermore Valley 2	5.50	17.7	C	0.275	0.090
Livermore Valley 2	5.50	22.5	C	0.058	0.040
Livermore Valley 2	5.50	48.3	A	0.026	0.020
Horse Canyon	5.30	5.8	A	0.123	0.088
Horse Canyon	5.30	12.0	A	0.133	0.118
Horse Canyon	5.30	12.1	A	0.073	0.067
Horse Canyon	5.30	36.1	A	0.111	0.084
Horse Canyon	5.30	20.6	B	0.097	0.076
Horse Canyon	5.30	20.6	B	0.096	0.096
Horse Canyon	5.30	25.3	B	0.181	0.114
Horse Canyon	5.30	36.3	B	0.110	0.094
Horse Canyon	5.30	41.4	B	0.040	0.320
Horse Canyon	5.30	43.6	B	0.047	0.044
Horse Canyon	5.30	44.4	B	0.022	0.017
Horse Canyon	5.30	35.9	C	0.082	0.050
Horse Canyon	5.30	38.5	C	0.094	0.060
Horse Canyon	5.30	46.1	C	0.057	0.046
Horse Canyon	5.30	47.1	C	0.080	0.062
Loma Prieta	6.92	10.5	A	0.500	0.430
Loma Prieta	6.92	29.9	A	0.060	0.040
Loma Prieta	6.92	32.5	A	0.090	0.070
Loma Prieta	6.92	42.7	A	0.070	0.070
Loma Prieta	6.92	67.6	A	0.110	0.060
Loma Prieta	6.92	69.0	A	0.040	0.060
Loma Prieta	6.92	72.6	A	0.110	0.070
Loma Prieta	6.92	77.2	A	0.080	0.070
Loma Prieta	6.92	78.5	A	0.090	0.080
Loma Prieta	6.92	79.5	A	0.060	0.030
Loma Prieta	6.92	80.5	A	0.050	0.060
Loma Prieta	6.92	0.0	B	0.500	0.640
Loma Prieta	6.92	10.9	B	0.370	0.330
Loma Prieta	6.92	11.7	B	0.340	0.530
Loma Prieta	6.92	12.0	B	0.330	0.260
Loma Prieta	6.92	12.3	B	0.250	0.280
Loma Prieta	6.92	12.5	B	0.440	0.470

Earthquake	Magnitude(M)	Distance(d)	Site Class	PA_H1	PA_H2
Loma Prieta	6.92	13.2	B	0.280	0.270
Loma Prieta	6.92	19.9	B	0.170	0.130
Loma Prieta	6.92	20.0	B	0.250	0.260
Loma Prieta	6.92	21.7	B	0.190	0.170
Loma Prieta	6.92	34.1	B	0.070	0.070
Loma Prieta	6.92	36.1	B	0.130	0.080
Loma Prieta	6.92	38.7	B	0.080	0.080
Loma Prieta	6.92	42.0	B	0.110	0.130
Loma Prieta	6.92	46.4	B	0.110	0.120
Loma Prieta	6.92	46.5	B	0.090	0.160
Loma Prieta	6.92	46.6	B	0.090	0.100
Loma Prieta	6.92	48.7	B	0.100	0.110
Loma Prieta	6.92	49.9	B	0.070	0.100
Loma Prieta	6.92	53.0	B	0.060	0.090
Loma Prieta	6.92	53.7	B	0.070	0.070
Loma Prieta	6.92	56.0	B	0.080	0.080
Loma Prieta	6.92	57.7	B	0.160	0.160
Loma Prieta	6.92	58.7	B	0.060	0.060
Loma Prieta	6.92	75.9	B	0.120	0.100
Loma Prieta	6.92	77.6	B	0.050	0.060
Loma Prieta	6.92	8.6	C	0.470	0.540
Loma Prieta	6.92	12.1	C	0.330	0.370
Loma Prieta	6.92	14.0	C	0.370	0.550
Loma Prieta	6.92	15.8	C	0.220	0.420
Loma Prieta	6.92	24.3	C	0.330	0.230
Loma Prieta	6.92	25.4	C	0.290	0.270
Loma Prieta	6.92	27.0	C	0.160	0.170
Loma Prieta	6.92	27.5	C	0.220	0.190
Loma Prieta	6.92	27.8	C	0.230	0.250
Loma Prieta	6.92	29.3	C	0.110	0.130
Loma Prieta	6.92	31.4	C	0.100	0.140
Loma Prieta	6.92	31.4	C	0.120	0.090
Loma Prieta	6.92	34.8	C	0.200	0.210
Loma Prieta	6.92	35.0	C	0.290	0.190
Loma Prieta	6.92	42.4	C	0.150	0.200
Loma Prieta	6.92	50.9	C	0.170	0.160
Loma Prieta	6.92	56.3	C	0.140	0.180
Loma Prieta	6.92	61.6	C	0.080	0.090
Loma Prieta	6.92	63.2	C	0.330	0.240

Earthquake	Magnitude(M)	Distance(d)	Site Class	PA_H1	PA_H2
Loma Prieta	6.92	67.3	C	0.100	0.130
Loma Prieta	6.92	68.8	C	0.040	0.040
Loma Prieta	6.92	75.2	C	0.260	0.200
Loma Prieta	6.92	76.3	C	0.200	0.260
Loma Prieta	6.92	78.6	C	0.050	0.050
Loma Prieta	6.92	78.8	C	0.290	0.270
Loma Prieta	6.92	80.5	C	0.080	0.080
Petrolia	7.10	1.9	A	0.210	0.180
Petrolia	7.10	9.8	B	0.480	0.320
Petrolia	7.10	12.3	B	0.390	0.550
Petrolia	7.10	13.7	B	0.120	0.120
Petrolia	7.10	14.6	B	0.280	0.320
Petrolia	7.10	17.6	B	0.260	0.260
Petrolia	7.10	23.9	B	0.180	0.140
Petrolia	7.10	32.6	B	0.180	0.240
Petrolia	7.10	0.0	C	0.690	0.620
Petrolia	7.10	10.0	C	0.300	0.370
Petrolia	7.10	27.8	C	0.200	0.150
Petrolia	7.10	35.8	C	0.170	0.160
Landers	7.30	2.1	A	0.880	0.630
Landers	7.30	27.6	A	0.120	0.120
Landers	7.30	37.6	A	0.060	0.050
Landers	7.30	41.9	A	0.090	0.070
Landers	7.30	51.3	A	0.060	0.050
Landers	7.30	56.2	A	0.030	0.030
Landers	7.30	60.1	A	0.050	0.060
Landers	7.30	60.8	A	0.040	0.050
Landers	7.30	68.2	A	0.042	0.053
Landers	7.30	68.3	A	0.150	0.120
Landers	7.30	69.7	A	0.060	0.080
Landers	7.30	70.2	A	0.030	0.020
Landers	7.30	78.0	A	0.030	0.040
Landers	7.30	78.7	A	0.140	0.130
Landers	7.30	83.7	A	0.050	0.060
Landers	7.30	86.0	A	0.040	0.050
Landers	7.30	89.0	A	0.030	0.030
Landers	7.30	89.4	A	0.040	0.040
Landers	7.30	89.4	A	0.050	0.040
Landers	7.30	93.3	A	0.040	0.040

Earthquake	Magnitude(M)	Distance(d)	Site Class	PA_H1	PA_H2
Landers	7.30	95.9	A	0.050	0.030
Landers	7.30	97.4	A	0.020	0.020
Landers	7.30	97.6	A	0.080	0.080
Landers	7.30	99.4	A	0.040	0.050
Landers	7.30	100.1	A	0.050	0.050
Landers	7.30	104.8	A	0.030	0.030
Landers	7.30	112.2	A	0.040	0.060
Landers	7.30	117.9	A	0.030	0.040
Landers	7.30	11.3	B	0.290	0.280
Landers	7.30	17.7	B	0.207	0.188
Landers	7.30	22.5	B	0.170	0.150
Landers	7.30	22.8	B	0.430	0.280
Landers	7.30	25.8	B	0.220	0.220
Landers	7.30	27.7	B	0.136	0.134
Landers	7.30	27.8	B	0.137	0.087
Landers	7.30	37.7	B	0.150	0.140
Landers	7.30	45.4	B	0.180	0.170
Landers	7.30	45.4	B	0.100	0.120
Landers	7.30	57.0	B	0.130	0.140
Landers	7.30	57.8	B	0.040	0.060
Landers	7.30	59.5	B	0.050	0.070
Landers	7.30	61.7	B	0.070	0.050
Landers	7.30	62.4	B	0.080	0.090
Landers	7.30	62.6	B	0.060	0.060
Landers	7.30	64.1	B	0.080	0.080
Landers	7.30	65.0	B	0.120	0.110
Landers	7.30	65.6	B	0.050	0.050
Landers	7.30	66.9	B	0.060	0.060
Landers	7.30	71.9	B	0.080	0.090
Landers	7.30	74.8	B	0.040	0.050
Landers	7.30	76.0	B	0.120	0.120
Landers	7.30	76.1	B	0.080	0.090
Landers	7.30	77.5	B	0.100	0.090
Landers	7.30	79.0	B	0.050	0.080
Landers	7.30	79.4	B	0.050	0.030
Landers	7.30	80.6	B	0.060	0.060
Landers	7.30	81.2	B	0.100	0.110
Landers	7.30	83.7	B	0.060	
Landers	7.30	84.7	B	0.060	0.060

Earthquake	Magnitude(M)	Distance(d)	Site Class	PA_H1	PA_H2
Landers	7.30	85.7	B	0.080	0.080
Landers	7.30	86.4	B	0.080	0.090
Landers	7.30	88.3	B	0.110	0.110
Landers	7.30	89.6	B	0.070	0.050
Landers	7.30	92.4	B	0.090	0.130
Landers	7.30	93.1	B	0.080	0.120
Landers	7.30	95.0	B	0.050	0.050
Landers	7.30	96.2	B	0.040	0.050
Landers	7.30	99.2	B	0.100	0.050
Landers	7.30	101.7	B	0.120	0.070
Landers	7.30	105.7	B	0.040	0.050
Landers	7.30	107.4	B	0.080	0.080
Landers	7.30	116.1	B	0.070	0.050
Landers	7.30	118.2	B	0.060	0.080
Landers	7.30	26.3	C	0.250	0.150
Landers	7.30	36.7	C	0.090	0.090
Landers	7.30	37.7	C	0.120	0.100
Landers	7.30	49.6	C	0.130	0.290
Landers	7.30	52.6	C	0.100	0.080
Landers	7.30	54.9	C	0.120	0.100
Landers	7.30	65.5	C	0.050	0.050
Landers	7.30	66.8	C	0.050	0.070
Landers	7.30	69.1	C	0.100	0.090
Landers	7.30	72.6	C	0.090	0.070
Landers	7.30	72.7	C	0.120	0.100
Landers	7.30	77.5	C	0.120	0.120
Landers	7.30	79.0	C	0.060	0.070
Landers	7.30	79.6	C	0.220	0.110
Landers	7.30	79.9	C	0.090	0.080
Landers	7.30	86.8	C	0.140	0.110
Landers	7.30	87.3	C	0.050	0.050
Landers	7.30	98.7	C	0.080	0.070
Landers	7.30	98.7	C	0.060	0.100
Landers	7.30	105.6	C	0.130	0.150
Landers	7.30	106.2	C	0.040	0.030
Landers	7.30	115.3	C	0.090	0.080
Landers	7.30	117.6	C	0.050	0.070

Chapter 5

Structural Health Monitoring

5.1 Illustrative Examples for Structural Health Monitoring

5.1.1 RVM Classification for SHM

In this section, the application of RVM classification for SHM is investigated using several simple structural systems.

5.1.1.1 Planar Shear Building Models

The first examples are 3-, 4-, and 5-story planar shear-building models. The models are used to generate 400 training data corresponding to either noisy scaled fundamental mode shapes (dividing by the square of the fundamental frequency) or noisy first Ritz vectors, that is, two datasets are generated to use as features. The lumped floor masses are $m = 100$ kips/g $= 0.2591$ kip sec^2/in and the interstory stiffnesses are $k = 31.56$ kips/in. The natural frequencies of the undamaged 3-, 4-, and 5-story buildings are 0.78, 0.61, and 0.50 Hz, respectively.

Damage is imposed as a 20% stiffness reduction to the first story of each building and three levels of Gaussian noise of mean zero and standard deviations 3%, 5%, and 10% are added to the extracted mode shapes and Ritz vectors and 3% is added to the natural frequencies. Then the scaled fundamental mode shapes and first Ritz vectors are prepared as feature vectors.

5.1.1.2 Bridge Models

Figure 5.1 shows the configuration of a 2-D truss and a 3-D frame representing simple bridge models. For both the truss and bridge training datasets, 400 9-D feature vectors are extracted from a finite element model of each structure. For each structure, the feature vectors are either the noisy fundamental scaled mode shapes or the noisy first force-dependent Ritz vectors, and they are simulated for undamaged and 20% damaged structures with different levels of simulated measurement noise added. Damage is imposed as a 20% stiffness reduction to the element labeled 1 of each bridge models.

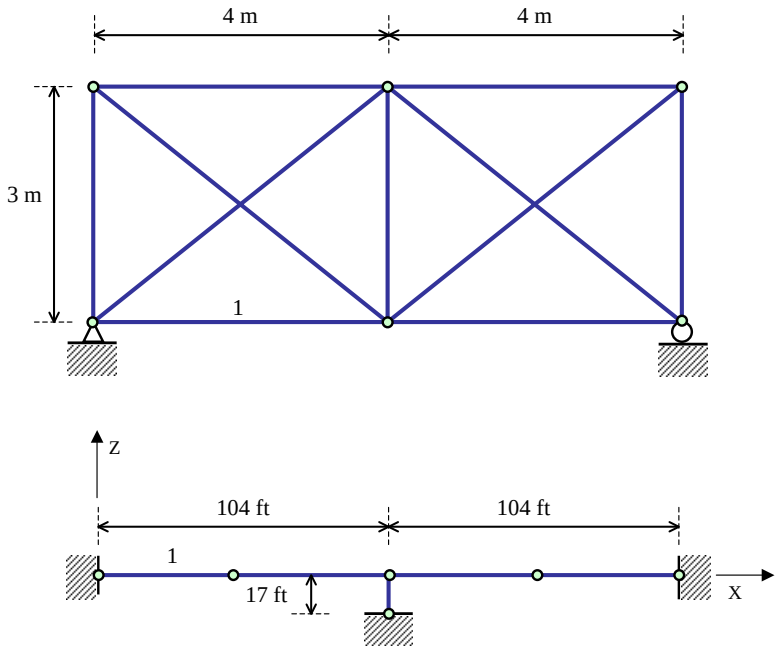


Figure 5.1: Configuration of 2-D Truss and 3-D Frame Bridge Structures.

For the 2-D truss bridge model, density (ρ), Young's modulus (E), and cross-sectional area (A) are 7850 kg/m^3, 220 GPa, and 16.5 cm^2, respectively, and the values along with units used as material properties and constants for the 3-D frame bridge are listed in Table 5.1. The fundamental frequencies of the 2-D truss and 3-D frame models are 67.49 and 7.57 Hz, respectively.

		Values	Units
ρ		4.658e-3	(kips sec^2/ft^4)
E		518400	(kips/ft^2)
I_z	(Deck)	130.23	(ft^4)
I_z	(Column)	34.92	(ft^4)
I_y	(Deck)	4976.85	(ft^4)
I_y	(Column)	34.92	(ft^4)
GJ	(Deck)	1.55e+8	(kips ft^2)
GJ	(Column)	1.55e+7	(kips ft^2)
A	(Deck)	51.66	(ft^2)
A	(Column)	22.35	(ft^2)

Table 5.1: Material Properties and Constants for a 3-D Frame Bridge.

5.1.1.3 Classification Results

Figure 5.2 provides SHM results for the planar 3-story shear building using the scaled mode shapes. The 3-D feature vectors are viewed at the same azimuth and elevation for both RVM and SVM. The separating decision boundary, shown as a gray-colored plane, is estimated from each RVM and SVM method and separates the training dataset. Datasets generated from undamaged and 20% damaged 3-story shear buildings are plotted as red crosses and blue dots, respectively.

	C	No. of RVs or SVs	Misclassification rate (%)		Execution Time(sec)
			$P_1 = P_0 = 0.5$	$P_1 = 0.4$ and $P_0 = 0.4$	
RVM	N/A	2	1.50	1.25	4.70
SVM	1.0156	93	1.75	N/A	14.14

Table 5.2: Comparison between RVM and SVM for 3-story Building Example.

The performance of RVM is slightly better than that of SVM in the sense of the misclassification rate, as shown in Table 5.2, but the number of controlling feature vectors, i.e., the number of Relevance Vectors (RVs) and Support Vectors (SVs), differs substantially: 2 RVs for RVM and 93 SVs for SVM. Another specific characterization of RVM is that the RVs are located further from the decision boundary than SVs, which may explain the higher sparsity (i.e., fewer kernel terms defining the separating boundary) inherent in RVM. In other words, SVM uses data near the separating boundary to estimate it, while only a few representative data play the important role

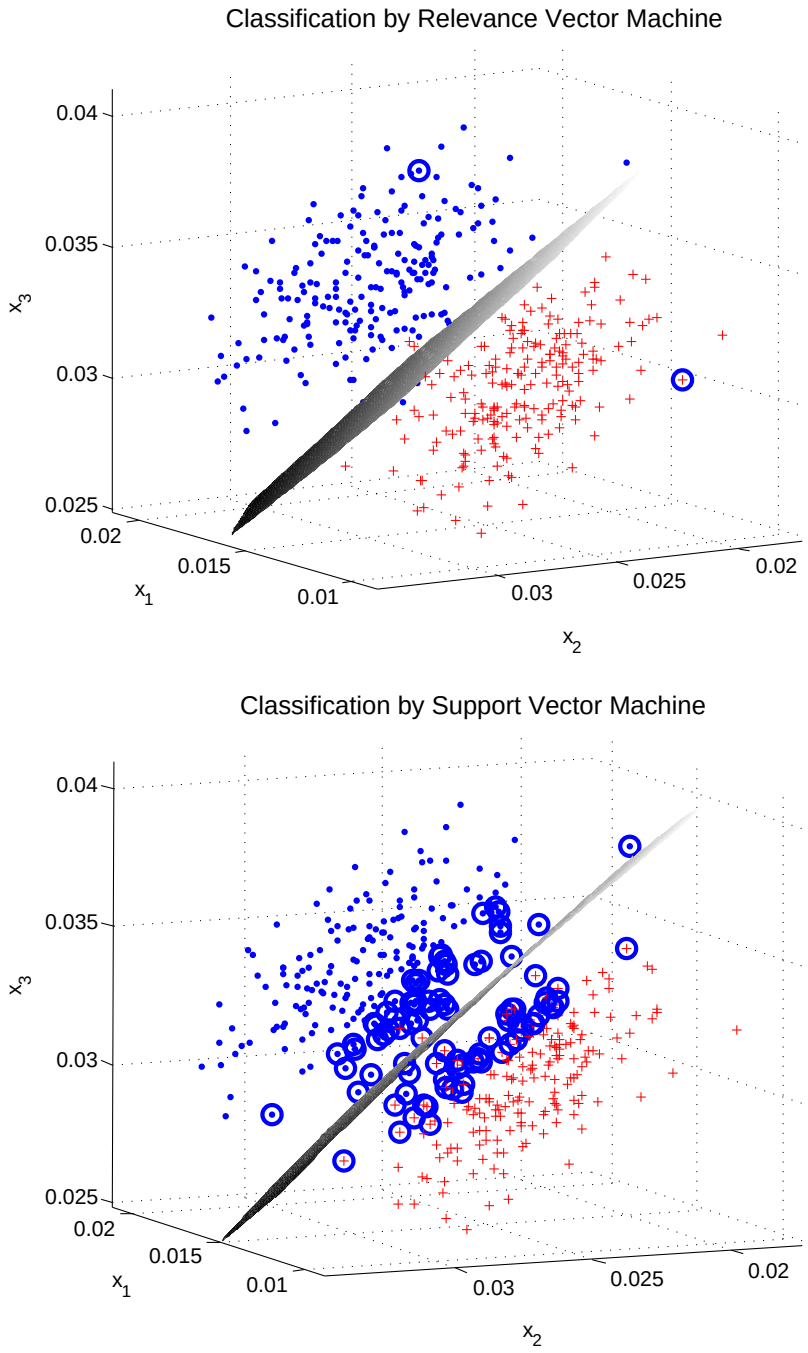


Figure 5.2: RVM and SVM Classification Using the Dataset of the Scaled Fundamental Mode Shape Simulated from the 3-story Building. 10% and 3% noise are added to the fundamental mode shape and fundamental frequency, respectively.

	Imposed Damage (%)	Misclassification rate (%)	
		RVM	SVM
3-Story Building	5	6.0	5.5
	25	0.5	0.5

Table 5.3: Comparison between RVM and SVM for 3-story Building Example.

of defining the separating boundary in RVM. In SVM, 10-fold cross-validation is also employed for deciding the trade-off parameter C; this parameter provides trade-off between data-fit errors and the complexity of separating boundaries (classification), or estimated regression equations (regression). Table 5.2 shows the detailed results, including the value of C for SVM and the reduction of RVM misclassification rate from 1.50% to 1.25% when more conservative decision boundaries corresponding to $P_1 = 0.4$ and $P_0 = 0.4$ are considered for RVM. The elapsed times for executing the RVM and SVM programs are also shown. As mentioned before, SVM takes longer because of the cross-validation, which requires multiple processing of the training data.

New monitoring data of the scaled mode shape are generated for two cases: 5% and 25% stiffness reduction in the first story of the damaged 3-story building, then the RVM and SVM trained by the undamaged (0%) and 20% damaged datasets are applied. The 5% and 25% damaged states are expected to be classified close to the 0% and 20% categories, respectively. Table 5.3 shows the results of classification: the performance of each method becomes better as the damage severity increases. The higher misclassification rate in predicting the 5% damaged state is expected because the damage features do not differ much in the initial damage stages. The 6% error percentage, however, is a promising result considering that 10% level of noise is added to generate the scaled mode shapes and only first mode information is used.

Table 5.4 shows the comparison results of RVM using the scaled fundamental mode shape and the first Ritz vector as features to check their relative performance, since it is often claimed that the first Ritz vector is a better feature than the fundamental mode shape for damage detection. The smaller misclassification rate between each case is written in bold and it shows that the scaled fundamental mode shape is

	Dim. of features	Noise level(%)	Scaled mode shape		Ritz vector	
			No. of RVs	Errors(%)	No. of RVs	Errors(%)
3-Story Building	3	3	2	**0.25**	2	5.00
		5	2	**1.00**	4	17.50
		10	2	**1.50**	5	31.00
4-Story Building	4	3	2	**0.25**	3	2.00
		5	3	**2.50**	3	14.00
		10	3	**3.25**	6	27.50
5-Story Building	5	3	3	**0.25**	3	1.00
		5	4	**0.75**	5	11.00
		10	4	**7.25**	11	25.50
Truss	9	3	5	**0.50**	2	10.50
		5	8	**1.00**	2	14.00
		10	16	**7.25**	6	19.50
Bridge	9	3	7	**1.50**	5	3.00
		5	12	**6.00**	4	15.00
		10	20	**12.00**	7	31.00

Table 5.4: Comparison between Scaled Fundamental Mode Shape and the First Ritz Vector Applied to Various SHM Examples by RVM. ("Errors" is misclassification rate.)

consistently more damage sensitive.

Some remarks concerning a pattern classification approach for SHM are warranted. As shown above, binary classification (two-labeled dataset) using RVM is effective in providing SHM results with a small amount of misclassification (considering the amount of noise). Moreover, the Bayesian classification methodology allows probabilistic predictions, i.e., predictions with the degree of belief, so that users can make conscious decisions on structural safety, operational loss, or potential catastrophe caused by the decisions. For example, if one user wants to operate a SHM system for a nuclear power plant especially to reduce missed alarm, (i.e., a SHM system does not warn the user even if damage exists), decision boundaries with a higher probability to an undamaged state should be selected in order to reduce the unchecked defects. On the other hand, frequent false alarm of a smoke detector would let users turn off the device and this also can be prevented by choosing decision boundaries with a lower probability to a smoke-free state. However, as the structure becomes complicated, the size of the generated dataset to cover all damage cases may become very large.

Suppose that 100 training data need to be generated from a 3-story shear building model for each damage location and each different damage severity. If the number of possible damage severities is 3 (for example, 25%, 50%, and 75% stiffness reduction) and the number of simultaneous damage locations is 3 (i.e., damage can occur at each floor simultaneously), then the total number of training dataset is 11,800 ($= ({}_3C_0 + {}_3C_1 3 + {}_3C_2 3^2 + {}_3C_3 3^3) \times 100$); this total number of training dataset becomes 101,400 for a 5-story building. To reduce the size of the dataset, we can shift to a Bayesian regression approach for SHM; then the total number of data is reduced to 118 and 1,014 for 3-story and 5-story buildings, respectively. Another advantage of a regression approach is that it is possible to estimate damage severities as continuous quantities instead of discrete damage states; for example, instead of generating the training dataset with 3 different damage severities as before, the trained algorithm with a smaller number of possible damage severities (say, 25% and 75% stiffness reduction) is expected to provide the prediction for 50% stiffness reduction. This can reduce the total number of training data to 27 and 243 for 3-story and 5-story buildings, respectively. However, to apply RVM regression for SHM, it is necessary to extend the RVM algorithm to treat vector outputs. In the next section, an extended RVM regression approach for SHM is investigated.

5.1.2 RVM Regression with Vector Outputs

There has been a lot of research on what to use as damage-informative features for SHM: modal properties such as natural frequencies and mode shapes, or the changes in these quantities (Ching and Beck, 2004; Vanik et al., 2000; Yuen et al., 2004); the ratios between changes of the measured eigenvalues (Cawley and Adams, 1979); Ritz vectors, or the changes in them (Cao and Zimmerman, 1997; Sohn and Law, 2001; Lam et al., 2006); damage signature, defined as the ratio of the change of eigenvectors to the change of eigenvalues (Yuen and Lam, 2006); mode shapes scaled by the inverse square of their natural frequencies (Oh and Beck, 2006); and so on. The goal has been to search for features that are sensitive to structural damage but

insensitive to modeling and/or measurement errors.

In this section, a two-step approach is performed to detect and assess structural damage (Yuen and Lam, 2006). The first step is to identify any damage locations. Damage signatures defined as the ratios of the change in N_m mode shape components to the change of a reference eigenvalue (e.g., fundamental frequency squared) are utilized as inputs to the RVM and a damage location index vector is chosen as output:

$$\begin{aligned} &\text{Step 1 Input : Damage Signature from j}^{\text{th}} \text{ mode } (\text{DS}_j) = \frac{\Delta\phi_j}{\Delta\omega_1^2},\ j = 1, ..., N_m \\ &\text{Step 1 Output : } \underline{L} = [L_1, ..., L_{N_L}]^T \end{aligned} \tag{5.1}$$

where N_L is the number of possible damage locations and L_i has value of 1 if damage exists at the i^{th} location and 0 otherwise. In the second step, utilizing the identified damage locations from the first step, damage severity is estimated by using the changes in mode shape components and natural frequencies from the undamaged (baseline) structure as input, and a damage severity index vector as output:

$$\begin{aligned} &\text{Step 2 Input : } [\Delta\phi_j, \Delta f],\ j = 1, ..., N_m \\ &\text{Step 2 output : } \underline{E} = [E_1, ..., E_{N_E}]^T \end{aligned} \tag{5.2}$$

where N_E is the number of damage locations identified from the first step and E_i has values between zero and one representing the fractional stiffness reduction at corresponding structural elements.

Systematic methodologies for pattern recognition or regression, such as ANN, SVM, and RVM, are classified as supervised learning methods since the damage states for a training dataset are assumed to be known a priori. This condition can not be satisfied with real data since it is not possible to induce various damage states in an existing structure. Supervised learning, however, enables the SHM results to provide more information regarding damage severity and location. In this study, we implement a supervised learning approach by using a finite-element (FE) model of the structure to generate the training dataset by introducing various damage patterns

in the FE model. However, the FE model used for training is not the same as the one representing the actual structure in the testing phase; realistic modeling error is reflected in the difference between the actual structural system and the FE model used to generate a training dataset. Also, a certain amount of noise is added to the simulated acceleration time histories to encourage robustness against measurement errors during the operating phase.

A five-story shear building is chosen which has previously been used to study the applicability of ANN for estimating the damage locations and severity (Yuen and Lam, 2006). A similar study is performed here using the vector output RVM. However, for more realism, modeling errors are introduced in this study when generating a training dataset by using 95% of the floor mass and 90% of the interstory stiffness when constructing the FE model from which mode shapes and natural frequencies are calculated for various damage states. (Yuen and Lam (2006) did not consider modeling error when generating a training dataset).

Training consists of two steps as explained before. In the first step, damage signatures and indices (5.1) are used as inputs and outputs, respectively. After training a RVM to find the damage locations, another RVM is trained to estimate damage severities using changes in the modal parameters (here fundamental frequency and mode shape as inputs and damage severity indices (5.2) as outputs).

5.1.2.1 Step 1: Detecting Damage Locations

When training the RVM to find the damaged stories, the total number of possible damage cases is 32 (including an undamaged case) because multiple stories may be damaged. For each of these damage cases, a 50% stiffness reduction is imposed for each damaged story to generate the feature vectors from the training FE structural model. The RVM trained to find damage locations is then implemented by using the mode shapes and natural frequencies from simulated dynamic data from the other FE model that represents the actual shear building. To provide a severe test, only the fundamental frequency and mode shape are extracted from noisy data obtained by adding 5% root-mean-square discrete white-noise to simulated acceleration time

histories to reflect the measurement noise.

The damage cases used in this prediction phase are as follows:

(1) Single damage in each story with 20% stiffness reduction (see results in Table 5.5(a)).

(2) Single damage in each story with 80% stiffness reduction (see results in Table 5.5(b)).

(3) Damage in two stories with 20% stiffness reduction in each (see results in Table 5.6).

(4) Damage in the 2^{nd} and 3^{rd} stories with selected stiffness reductions r_2% and r_3%, respectively (see results in Table 5.7).

In these tables, RVM output values near one indicate the location of damage, so a threshold of 0.5 was taken to indicate that the story is damaged. Note that datasets for prediction are not included in the training dataset. From Table 5.5 to Table 5.7, it can be concluded that the RVM works well for identifying damage locations with a performance comparable with the previous study using ANN (see the results in Yuen and Lam (2006)). In two instances of Damage Case 3, 1/3 and 3/4, the RVM indicates the top story is damaged when it is not; this is not so serious because this should be assigned a low damage severity in Step 2. On the other hand, in Damage Cases 1 and 2, story 4 damage is missed, and in Damage Case 4, damage is missed in the 50/10, 50/20, 50/80 and 50/90 cases, which is a more serious failure of the damage location approach.

5.1.2.2 Step 2: Assessing Damage Severity

After locating the potentially damaged structural members, the severity of the damage is estimated. Two different cases are considered when training and predicting for this purpose:

(1) The fundamental frequency and mode shape are used together and are generated from FE model with a single damage at the 2nd story. These features are generated by imposing 10%, 20%, 30%, 40%, 50%, 60%, and 70% stiffness reduction to the corresponding story. For the prediction, the stiffness at the 2nd story is reduced by

Story	(a) Damage Case 1					(b) Damage Case 2				
	1	2	3	4	5	1	2	3	4	5
1	**1.07**	-0.10	-0.19	-0.02	0.00	**1.06**	-0.43	-0.01	0.01	0.00
2	0.27	**0.75**	0.17	-0.04	0.00	0.10	**0.57**	0.17	0.00	0.00
3	0.12	0.11	**0.88**	0.15	0.00	0.03	-0.09	**1.22**	-0.02	0.00
4	-0.28	-0.32	0.42	**0.98**	0.00	-0.06	-0.16	0.14	**0.87**	0.00
5	0.42	0.11	0.42	0.31	**0.40**	0.18	0.10	0.14	0.27	**0.40**

Table 5.5: Identify Single Damage of (a) 20% and (b) 80% Stiffness Reduction in Each Story. (Bold numbers correspond to actual damage locations.)

Damage Case 3	Story				
	1	2	3	4	5
1/2	**0.81**	**0.96**	0.11	-0.05	-0.02
1/3	**0.97**	0.33	**1.33**	-0.32	0.58
1/4	**1.12**	0.33	0.20	**1.21**	0.30
1/5	**1.07**	0.29	0.23	-0.36	**1.30**
2/3	-0.12	**0.94**	**0.99**	-0.22	0.16
2/4	-0.20	**1.03**	0.01	**0.79**	0.25
2/5	-0.24	**0.86**	0.03	-0.23	**1.12**
3/4	-0.06	0.03	**1.00**	**0.64**	0.52
3/5	-0.05	0.02	**0.83**	-0.07	**0.72**
4/5	-0.01	-0.01	0.08	**0.79**	**0.52**

Table 5.6: Identify Damage of 20% Stiffness Reduction in Both x_1/x_2 Stories. (Bold numbers correspond to actual damage locations.)

Damage Case 4	Story				
r_2/r_3	1	2	3	4	5
50/10	-0.02	**0.97**	**0.25**	-0.18	0.01
50/20	-0.05	**1.01**	**0.42**	-0.18	0.05
50/30	-0.05	**1.03**	**0.64**	-0.26	0.15
50/40	-0.07	**1.01**	**0.74**	-0.15	-0.02
50/50	-0.07	**0.97**	**1.02**	-0.08	0.04
50/60	-0.02	**0.85**	**1.25**	-0.08	0.03
50/70	0.24	**0.88**	**1.36**	-0.13	-0.18
50/80	-0.28	**0.42**	**0.98**	-0.03	0.22
50/90	-0.14	**0.13**	**0.50**	-0.10	0.21

Table 5.7: Identify Damage in the $2^{nd}/3^{rd}$ Stories of r_2/r_3% Stiffness Reduction. (Bold numbers correspond to actual damage locations.)

	Damage Case	Story				
		1	2	3	4	5
Damaged	1	0.00	**0.20**	0.00	0.00	0.00
	2	0.00	**0.50**	**0.30**	0.00	0.00
Predicted	1	0.00	**0.19**	0.00	0.00	0.00
	2	0.00	**0.51**	**0.32**	0.00	0.00

Table 5.8: Estimated Damage Severities.

20%.

(2) Similarly to the single-damage case, 10% to 70% stiffness reductions are imposed on the 2nd and 3rd stories. The same features as in (1) are obtained from the FE model for training. For the prediction, damage is imposed as a 50% and 30% stiffness reduction at the 2nd and 3rd stories, respectively.

Note that in the second step, the damage locations are assumed known. The predicted damage severities shown in Table 5.8 are very close to the exact values: 19.42% for the first case and 50.51% and 32.19% for the 2nd and 3rd stories for the second case. The performance is comparable to that of the study using ANN for the same example (see the results in Yuen and Lam (2006)). We conclude that the proposed vector output RVM successfully estimates both the damage locations and the damage severity for this simple illustrative example (with minor exceptions as shown in Table 5.7).

5.2 IASC-ASCE Structural Health Monitoring Benchmarks

5.2.1 IASC-ASCE Benchmark Structure

The IASC-ASCE benchmarks were developed for researchers to apply numerous SHM studies to a common structure for a common objective, thereby providing a platform for side-by-side comparison of SHM methods. The series of benchmarks were initiated by the joint IASC-ASCE Task Group on SHM (*http://mase.wustl.edu/wusceel /asce.shm/benchmarks.htm*). There were two phases, each consisting of simulated and

test data benchmarks; only the Phase I simulated-data benchmarks are considered here. (See the special issue of the *Journal of Engineering Mechanics*, January 2004, devoted to these benchmarks.)

As the first step, the Task Group constructed an analytical structural model based on an existing steel frame scaled-model structure located at the University of British Columbia and shown in Figure 5.3. This structure consists of 4-story, 2-bay by 2-bay braced frame with 2.5 $m\times$ 2.5 m floor dimensions and 0.9 m height per story. Element properties for this model are detailed in Johnson et al. (2000, 2004). Two finite element (FE) models with 12 and 120 degrees of freedom (DOF) were constructed to simulate the acceleration time histories measured in the X (strong) and Y (weak) directions at the location shown with red dots in Figure 5.4. The details on the simulation cases with damage patterns are summarized in Table 5.9.

5.2.2 Identification of Modal Parameters

In the first step of damage detection, the lower-mode mode shapes are extracted from the measured time histories using a modal identification program called MODE-ID (Beck, 1996). This program uses a parametric time domain technique to estimate the modal parameters efficiently by using a non-linear least-squares method based on a linear dynamic model and the measure-of-fit objective function defined by:

$$J(\psi) \equiv \sum_{i=1}^{N} ||y(i\Delta t) - q(i\Delta t, f, \psi)||^2 \tag{5.3}$$

where ψ is the vector of modal parameters, y, $q \in \mathbb{R}^{N_0}$ are the measured and model responses, N_0 is the number of output channels, f is the measured inputs, and N is the number of sampled data points.

For ambient vibrations or, more generally, vibrations caused by unknown inputs, MODE-ID is not immediately applicable because it requires the input time histories to compute the model response. In this case, MODE-ID takes advantages of the fact that the theoretical cross-correlation functions of the response for a system satisfying

Figure 5.3: Steel Frame Scaled Model Structure Used for Benchmarks. (taken from *http://mase.wustl.edu/wusceel/asce.shm/structure.htm.*)

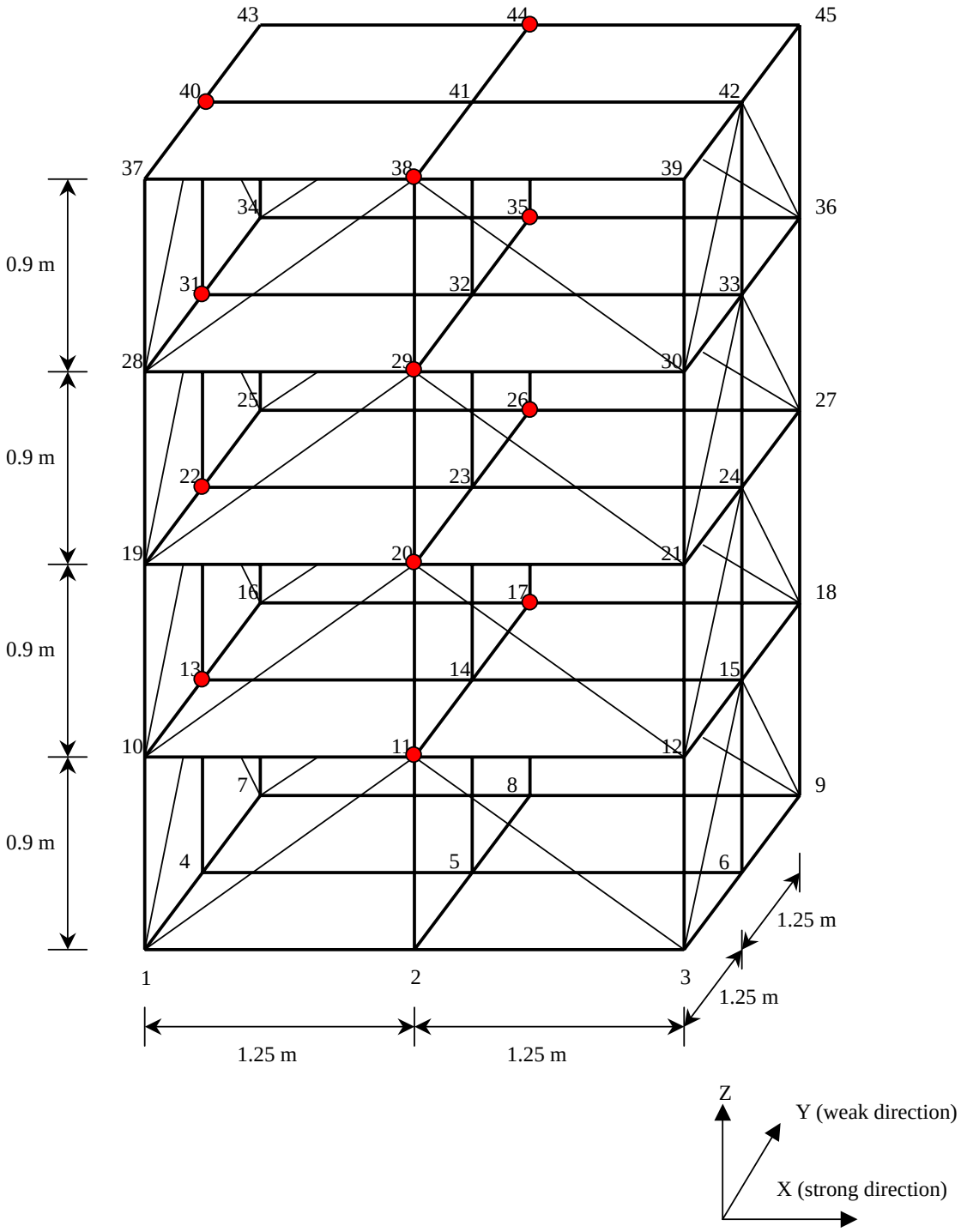


Figure 5.4: Configuration of Benchmark Structure.

the equation of motion,

$$M\ddot{x}(t) + C\dot{x}(t) + Kx(t) = f(t) \tag{5.4}$$

where x, $f \in \mathbb{R}^n$ and M, C, and K are mass, damping and stiffness matrices, respectively, satisfy the equations of motion for free vibrations if the inputs are modeled as temporally uncorrelated and stationary (Beck et al., 1994; Beck, 1996). Thus, sample cross-correlation functions are formed from the measured responses and given to MODE-ID as free vibrations, where the time lag τ serves as a pseudo vibration time.

The cross-correlation between two stationary signals, $x_i(t)$ and $x_r(t)$, is defined as

$$\begin{aligned} R_{ir}(\tau) &\equiv E[x_i(t) \cdot x_r(t+\tau)] \\ &\cong \frac{1}{T}\sum_{i=1}^{T}[x_i(t) \cdot x_r(t+\tau)] \end{aligned} \tag{5.5}$$

where $x_i(t)$ and $x_r(t)$ are the acceleration measurements at the i^{th} channel and at a reference channel, respectively. From this, one can construct a time history (with respect to τ) consisting of these correlations for all channels as a column vector for each time lag $\tau = 0, 1, ...$

$$[R_{1r}(\tau)R_{2r}(\tau)...R_{nr}(\tau)]^T \tag{5.6}$$

where n is the total number of output channels. This vector time history serves as the measured output response for MODE-ID for its free-vibration mode of operation. The same procedure can also be applied to accelerations, which most vibration sensors measure (Beck et al., 1994). MODE-ID estimates the natural frequency, damping ratio, and mode shape components at the observed degrees of freedom for a specified number of modes of vibration.

5.2.3 Damage Cases and Damage Patterns

The RVM approach to SHM is applied to simpler cases (damage cases 1 – 3) and then extended to more realistic ones (damage cases 4 – 5) of the Benchmark. For all damage cases, simplified models, such as shear building models using lumped masses with reduced DOFs, are used to generate the training dataset in order to reflect modeling error. The simulated dataset from the IASC-ASCE benchmark simulation phase I is used for the prediction phase (*http://mase.wustl.edu/wusceel/asce.shm/analyt_1.htm*). All of the damage cases and damage patterns are listed in Table 5.9. Damage patterns are defined as (Johnson et al., 2004):

1) No stiffness in the braces of the first story (i.e., the braces still contribute mass, but provide no resistance within the structure)

2) No stiffness in any of the braces of the first and third stories

3) No stiffness in one brace in the first story (north brace on the west face of the structure)

4) No stiffness in one brace in the first story (north brace on the west face) and in one brace in the third story (west brace on the north face)

Description	Case				
	1	2	3	4	5
12 DOF Model	O		O	O	
120 DOF Model		O			O
Symmetric Mass	O	O	O		
Asymmetric Mass				O	O
Ambient Excitation	O	O			
Shaker on Roof			O	O	O
Damage Patterns: Remove Followings					
1) All Braces in 1^{st} Story	O	O	O	O	O
2) All Braces in 1^{st} & 3^{rd} Stories	O	O	O	O	O
3) One Brace in 1^{st} Story				O	O
4) One Brace in 1^{st} & 3^{rd} Stories				O	O
5) 4) & Loosen Floor Beam at 1^{st} Level					O
6) 2/3 Stiffness in One Brace at 1^{st} Story					O

Table 5.9: Damage Cases and Patterns in Detail.

5) The same as damage pattern 4) but with the north floor beam at the first level on the west face of the structure (i.e., the beam from (2.5m, 0, 0.9m) to (2.5m, 1.25m, 0.9m)) partially unscrewed from the northwest column at (2.5m, 0, 0.9m) consequently, the beam-column connection there can only transmit forces and cannot sustain any bending moments

6) Two thirds stiffness (i.e., a one-third stiffness loss) in one brace in the first story (the same brace damaged in pattern 3: the north brace on the west face).

5.2.4 Damage Cases 1 – 3

5.2.4.1 Training Phase

To generate the training dataset, a four-story shear building model with lumped masses is utilized. This lumped mass model uses the so-called "nominal" mass matrix to construct the diagonal mass matrix of $M = \text{diag}\{3242,\ 2652,\ 2652,\ 1809\}$ kg (Yuen et al., 2004). The interstory stiffnesses are parameterized with stiffness parameter θ_i scaling the i^{th} story stiffness $k_i^u = 68.1$MN/m so that the damage is represented as a fraction of the undamaged stiffness:

$$k_i^{pd} = \theta_i k_i^u$$

where k_i^u and k_i^{pd} stand for the stiffness in the undamaged and possibly damaged states of the i^{th} story, respectively.

The generated feature vectors consist of mode shape changes and modal frequency changes, since the changes of modal parameters are considered to be more insensitive to modeling errors than the parameter values themselves (Lam et al., 2006). Several candidates are tested to investigate their performance or sensitivity to damage in the prediction phase.

Using Δ to denote the difference between the potentially damaged and undamaged values, the eight candidates are prepared for both training and prediction phases:

(a) $\underline{x}_i^a = [\Delta\left(\frac{\underline{\phi}_1}{\omega_1^2}\right)\ \Delta\left(\frac{\underline{\phi}_2}{\omega_2^2}\right)]_i$ with the component of $\underline{\phi}_j$ at roof = 1.

(b) $\underline{x}_i^b = [\Delta\left(\frac{\underline{\phi}_1}{\omega_1^2}\right)\ \Delta\left(\frac{\underline{\phi}_2}{\omega_2^2}\right)]_i$ with $||\underline{\phi}_j||^2 = 1$.

(c) $\underline{x}_i^c = [\frac{\Delta\underline{\phi}_1}{\Delta\omega_1^2}\ \frac{\Delta\underline{\phi}_2}{\Delta\omega_2^2}]_i$ with the component of $\underline{\phi}_j$ at roof $= 1$.

(d) $\underline{x}_i^d = [\frac{\Delta\underline{\phi}_1}{\Delta\omega_1^2}\ \frac{\Delta\underline{\phi}_2}{\Delta\omega_2^2}]_i$ with $||\underline{\phi}_j||^2 = 1$.

(e) $\underline{x}_i^e = [\Delta\underline{\phi}_1\ \Delta f_1\ \Delta\underline{\phi}_2\ \Delta f_2]_i$ with the component of $\underline{\phi}_j$ at roof $= 1$.

(f) $\underline{x}_i^f = [\Delta\underline{\phi}_1\ \Delta f_1\ \Delta\underline{\phi}_2\ \Delta f_2]_i$ with $||\underline{\phi}_j||^2 = 1$.

(g) $\underline{x}_i^g = [\Delta\underline{\phi}_1\ \frac{\Delta f_1}{f_1^u}\ \Delta\underline{\phi}_2\ \frac{\Delta f_2}{f_2^u}]_i$ with the component of $\underline{\phi}_j$ at roof $= 1$.

(h) $\underline{x}_i^h = [\Delta\underline{\phi}_1\ \frac{\Delta f_1}{f_1^u}\ \Delta\underline{\phi}_2\ \frac{\Delta f_2}{f_2^u}]_i$ with $||\underline{\phi}_j||^2 = 1$.

for $i = 1, ..., N$, and where $\underline{\phi}_j$ and ω_j or f_j are j^{th} mode shape and corresponding modal frequency with $\omega_j = 2\pi f_j$ for $j = 1, 2$. Note that x_i^c and x_i^d have the same form of damage signature as defined earlier and x_i^g and x_i^h are considered normalized frequencies.

For training, eight levels of stiffness losses are imposed, i.e., 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80% of each undamaged element stiffness, and then features are simulated from the mass and stiffness matrices via eigen-analysis with the restriction that the damage can occur at most two different locations at the same time. The total number of training vectors is 417 including undamaged data:

$$\text{Total No. of Data} = {}_4\mathbf{C}_0 + {}_4\mathbf{C}_1 8^1 + {}_4\mathbf{C}_2 8^2 = 417$$

where the combinatorial factor is defined as ${}_n\mathbf{C}_r = \frac{n!}{(n-r)!r!}$.

Only the mode shapes and modal frequencies of the first two modes are selected to compose the feature vectors in the training dataset. The results presented later in Tables 5.11, 5.12, and 5.13 show that damage can be identified and assessed satisfactorily with these modal parameters.

5.2.4.2 Prediction Phase

The dataset for prediction is extracted from the simulated benchmark structure data using the specified values of the damping coefficient, time step Δt, and noise level: 0.01, 0.004 sec, and 10%, respectively. The first 10 sec of data was eliminated to exclude the transient response time history and the remaining 20 sec was taken as stationary response. As noted previously, the input time histories that were utilized to simulate the ambient vibrations are not used here, since they are unknown in practice. Note that a comparison of the unknown input case and the known input case shows that the estimates for the stiffness parameters have a significantly larger coefficient of variation in the former case (Yuen et al., 2004).

To extract the modal parameters with unknown input, the sample correlation functions are evaluated according to the procedure explained in Section 5.2.2. The acceleration measurements at the lowest floor (i.e., at the first floor in this study) are selected as the reference channel to correlate with others. Since the first and second mode shapes in each direction are utilized for all damage cases from 1 to 5, possible nodes for a higher mode can be excluded at the first-floor level. The roof acceleration also could be used for the reference channel, but it is known that the relative contributions of the higher modes is greater at lower floors than that at the roof (Beck et al., 1994).

The sample correlations between the first-floor reference channel and the channels at each of the floors in the weak directions are shown in Figure 5.5. Only 1 sec is shown, since only this segment is used in MODE-ID. The extracted modal frequencies are listed in Table 5.10 and the mode shapes for each damage case and pattern are shown in Figures 5.6, 5.7, and 5.8. After extracting the modal parameters using MODE-ID, an investigation is performed to determine which combinations of features are more sensitive to damage.

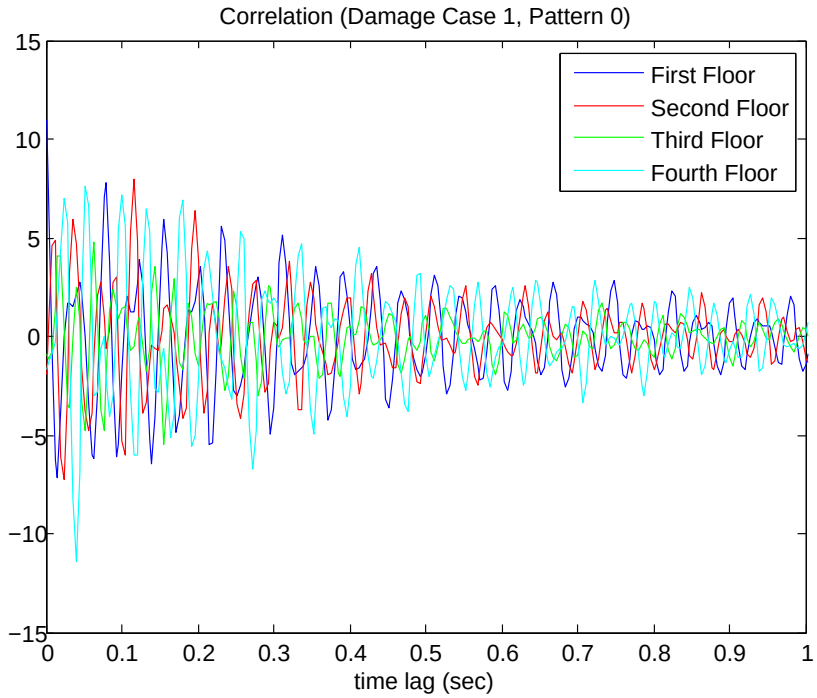


Figure 5.5: Sample Correlations between the First-Floor Reference Channel and the Measured Accelerations at the First to Fourth Floor, respectively.

Damage Case	Damage Pattern	Frequencies (Hz)			
		W1	S1	W2	S2
1	0	9.35	25.57	38.79	48.00
	1	6.26	21.48	37.52	47.91
	2	5.94	14.96	36.29	41.48
2	0	8.54	23.44	36.75	46.98
	1	5.54	19.52	35.40	46.76
	2	4.91	12.45	34.61	38.73
3	0	9.49	25.54	38.58	48.55
	1	6.31	21.34	37.59	47.18
	2	5.83	14.86	36.34	41.15
4	0	9.25	11.58	25.21	31.66
	1	6.17	9.81	21.26	28.62
	2	5.77	9.37	14.79	24.76
	3	8.76	11.58	24.39	31.66
	4	8.76	11.42	24.39	30.81
5	0	8.50	9.04	23.13	25.58
	1	5.43	7.33	18.99	22.52
	2	4.90	6.60	12.28	17.60
	3	7.98	9.03	22.34	25.58
	4	7.98	8.76	22.32	24.78
	5	7.92	8.76	22.28	24.78
	6	8.35	9.03	22.86	25.58

Table 5.10: Extracted Modal Frequencies for Cases 1–5.

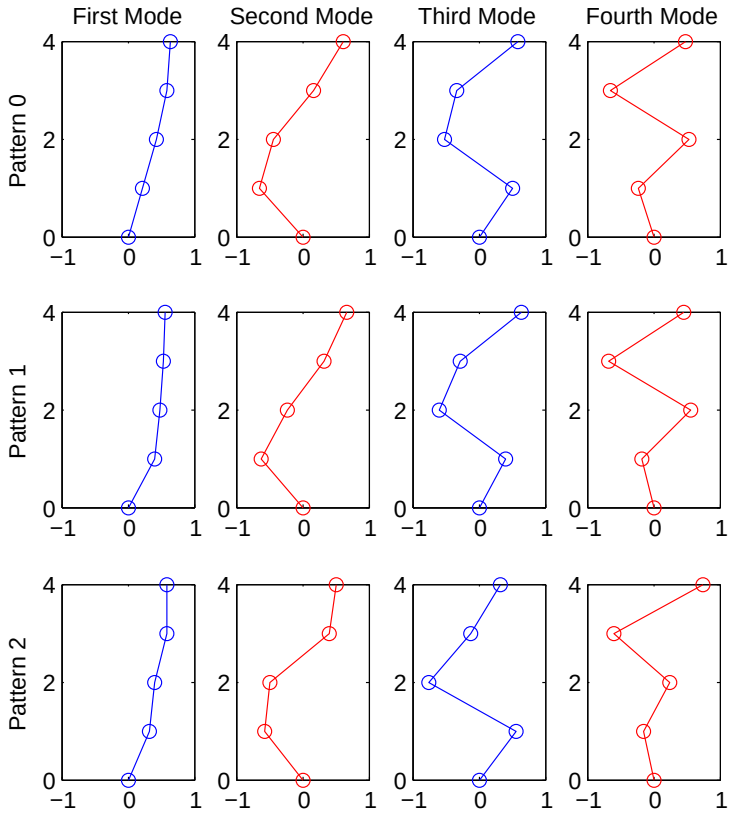


Figure 5.6: Identified Mode Shapes for Case 1, Pattern 0, 1, and 2.

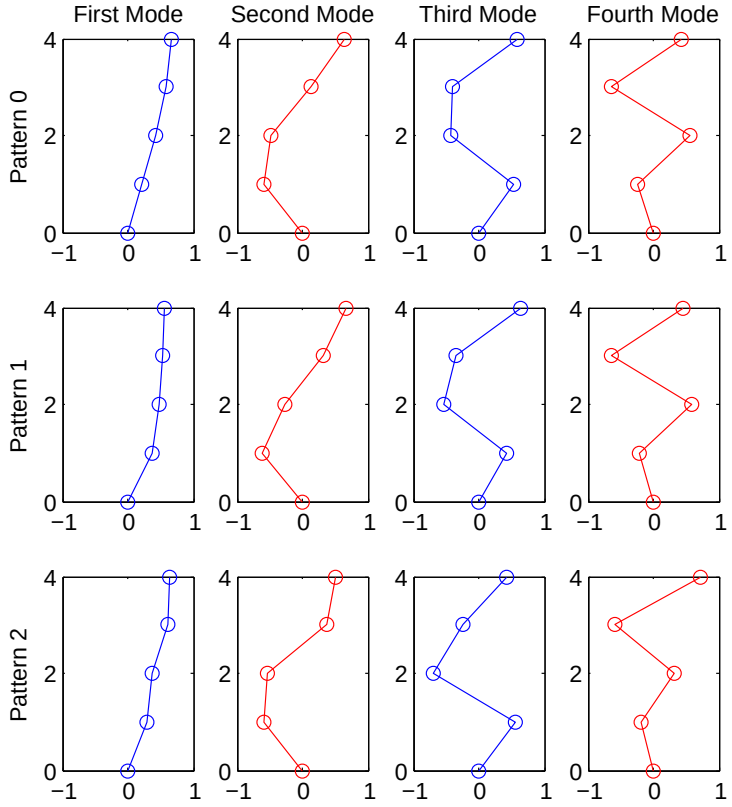


Figure 5.7: Identified Mode Shapes for Case 2, Pattern 0, 1, and 2.

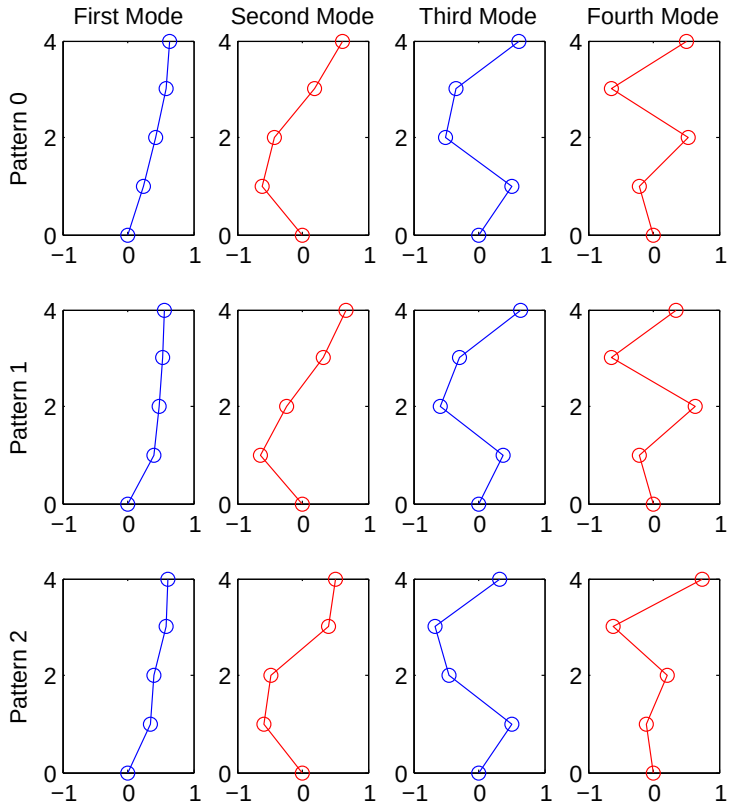


Figure 5.8: Identified Mode Shapes for Case 3, Pattern 0, 1, and 2.

The identified modal parameters in the formats defined earlier in (a) through (h) are utilized in the one-step procedure to identify the damage locations and to estimate the damage severity: this one-step procedure can identify the damage locations and damage severities simultaneously, while the two-step procedure presented before assesses the damage severities in the second-step only at the damage locations identified in the first-step. Damage patterns 0, 1, and 2 for each damage case 1 through 3 represent the undamaged state, single damage with all braces in the 1^{st} story removed and multiple damage with all braces in the 1^{st} and the 3^{rd} story removed, respectively. The stiffness for damage pattern 1 and 2 corresponds to 29% of the undamaged stiffness, i.e., 71% stiffness reduction of the undamaged stiffness. The predictions for each damage case and pattern are summarized in Tables 5.11, 5.12, and 5.13. The stiffness parameters in the "Damaged" row correspond to the actual stiffness reduction (71%), calculated based on a shear building model.

	Used Features	Damage Case	Damage Pattern	Story 1	2	3	4
Damaged		1	1	**0.71**	0.00	0.00	0.00
			2	**0.71**	0.00	**0.71**	0.00
Predicted	$\underline{x}^a$	1	1	**0.71**	0.08	0.00	0.02
			2	**0.70**	0.00	**0.65**	0.00
	$\underline{x}^b$	1	1	**0.70**	0.06	0.00	0.09
			2	**0.69**	0.00	**0.68**	0.03
	$\underline{x}^c$	1	1	**1.56**	0.00	0.21	0.34
			2	**0.92**	0.06	**0.73**	0.05
	$\underline{x}^d$	1	1	**0.53**	0.00	0.00	0.07
			2	**0.74**	0.00	**0.74**	0.00
	$\underline{x}^e$	1	1	**0.74**	0.00	0.03	0.04
			2	**0.71**	0.00	**0.69**	0.05
	$\underline{x}^f$	1	1	**0.71**	0.01	0.00	0.02
			2	**0.71**	0.00	**0.61**	0.02
	$\underline{x}^g$	1	1	**0.72**	0.00	0.00	0.03
			2	**0.71**	0.00	**0.67**	0.02
	$\underline{x}^h$	1	1	**0.74**	0.00	0.00	0.03
			2	**0.73**	0.00	**0.70**	0.02

Table 5.11: Stiffness Loss Predictions for Each Candidate Feature (Damage Case 1).

	Used Features	Damage Case	Damage Pattern	Story 1	2	3	4
Damaged		2	1	**0.71**	0.00	0.00	0.00
			2	**0.71**	0.00	**0.71**	0.00
Predicted	$\underline{x}^a$	2	1	**0.77**	0.09	0.00	0.12
			2	**0.72**	0.00	**0.71**	0.05
	$\underline{x}^b$	2	1	**0.78**	0.04	0.00	0.22
			2	**0.75**	0.05	**0.98**	0.00
	$\underline{x}^c$	2	1	**2.29**	0.00	0.21	0.55
			2	**0.39**	0.06	**0.50**	0.00
	$\underline{x}^d$	2	1	**0.33**	0.00	0.00	0.00
			2	**0.31**	0.00	**0.35**	0.00
	$\underline{x}^e$	2	1	**0.75**	0.00	0.01	0.00
			2	**0.72**	0.02	**0.76**	0.01
	$\underline{x}^f$	2	1	**0.73**	0.00	0.00	0.00
			2	**0.71**	0.00	**0.66**	0.00
	$\underline{x}^g$	2	1	**0.75**	0.04	0.00	0.00
			2	**0.74**	0.05	**0.78**	0.02
	$\underline{x}^h$	2	1	**0.75**	0.00	0.00	0.00
			2	**0.79**	0.01	**0.78**	0.00

Table 5.12: Stiffness Loss Predictions for Each Candidate Feature (Damage Case 2).

	Used Features	Damage Case	Damage Pattern	Story 1	2	3	4
Damaged		3	1	**0.71**	0.00	0.00	0.00
			2	**0.71**	0.00	**0.71**	0.00
Predicted	$\underline{x}^a$	3	1	**0.71**	0.08	0.00	0.00
			2	**0.67**	0.00	**0.69**	0.04
	$\underline{x}^b$	3	1	**0.69**	0.06	0.00	0.17
			2	**0.71**	0.00	**0.79**	0.00
	$\underline{x}^c$	3	1	**1.40**	0.00	0.39	0.21
			2	**0.76**	0.18	**0.62**	0.00
	$\underline{x}^d$	3	1	**0.44**	0.02	0.00	0.00
			2	**0.99**	0.04	**1.02**	0.00
	$\underline{x}^e$	3	1	**0.73**	0.04	0.03	0.05
			2	**0.74**	0.03	**0.71**	0.03
	$\underline{x}^f$	3	1	**0.70**	0.03	0.00	0.04
			2	**0.71**	0.00	**0.62**	0.01
	$\underline{x}^g$	3	1	**0.72**	0.04	0.00	0.04
			2	**0.72**	0.00	**0.69**	0.00
	$\underline{x}^h$	3	1	**0.74**	0.01	0.00	0.05
			2	**0.75**	0.00	**0.72**	0.00

Table 5.13: Stiffness Loss Predictions for Each Candidate Feature (Damage Case 3).

The results are best for feature $\underline{x}^e$, which outperforms slightly $\underline{x}^g$ and $\underline{x}^h$, based on the root mean-square errors calculated using the real stiffness reduction. For damage cases 4 and 5, therefore, features having the same form as $\underline{x}^e$ are used for SHM.

The analysis results show that the proposed algorithm applied to the IASC-ASCE benchmarks provides reliable results for damage cases 1, 2, and 3.

5.2.5 Damage Cases 4 – 5

Damage case 4 – 5 are based on a three-dimensional 12-DOF shear building model for training. The main difference with the previous damage cases 1 – 3 is that damage cases 4 – 5 can locate damage in a face of building, not just in a story.

On the other hand, there is a problem with using the same procedure as in damage cases 1 – 3; that is, where eight levels of stiffness reductions are imposed on the structural elements and then the modal features are extracted from a 12-DOF FE model and used as the dataset to train the RVM algorithm. This problem is due to the number of possible scenarios which increases enormously and causes computational difficulties:

$$\text{No. of Possible Damage Scenarios} = \sum_{i=0}^{N_L} {}_{N_{TL}}\mathbf{C}_i \; N_{DL}^i$$

where N_L, N_{TL}, and N_{DL} represent the number of simultaneous damage locations, the number of all possible damage locations, and the number of damage levels, respectively, and $\mathbf{C}$ is the combinatorial factor defined earlier. For example, the total number of training data becomes $\sum_{i=0}^{4} {}_{16}\mathbf{C}_i 3^i = 163,669$ when one considers three different damage levels, such as 25%, 50%, and 75%, occurring at 4 different locations among a total of 16 locations at the same time; thus, the largest matrix in the Bayesian learning method is $\mathbf{\Phi} \in \mathbb{R}^{163669 \times 163670}$. Although these problems caused by the large training dataset can be handled using a parallel computing capability, using expert knowledge before training the algorithm to recognize the critical points where damage is most likely to occur may be used to reduce the number of cases significantly.

Another strategy, and the one chosen here, is to use the two-step approach applied

to the five-story shear building model in Section 5.1.2. As shown for the five-story shear building, a damage signature is first computed to give information on the damage locations. Since the damage index $\underline{L}$ in (5.1) has a value between 0 and 1 (corresponding to undamaged and damaged cases, respectively), we judge that damage is likely if the damage index exceeds 0.6. This value of 0.6 is somewhat ad-hoc, but the second step that estimates damage severity should correct for any potential damage locations shown in the first step that do not correspond to actual damage. This whole procedure is an effort to reduce the number of training data, and the following results show that it is useful from this point of view. Details are presented in the following sections.

5.2.5.1 Training Phase

In these cases, a three-dimensional 12-DOF shear building model is employed to generate a training dataset. Damage can be identified by utilizing stiffness parameters θ_{ij} for each story i and face j for $i = 1, 2, 3, 4$ and $j = +x, -x, +y, -y$ faces (see Figure 5.9 for details):

$$k_{ij}^{pd} = \theta_{ij} k_{ij}^{u}$$

where the k_{ij}^{u} are computed from the benchmark structure with an assumption of shear building model:

$$k_{i,+x}^{u} = k_{i,-x}^{u} = 34.0 MN/m$$
$$k_{i,+y}^{u} = k_{i,-y}^{u} = 53.5 MN/m$$

In Figure 5.9, the floor plan is shown with the shear center ($\bar{x}_i$,$\bar{y}_i$) at the i^{th} floor calculated by

$$\bar{x}_i = \frac{a(k_{i,+x} - k_{i,-x})}{2(k_{i,+x} + k_{i,-x})}$$
$$\bar{y}_i = \frac{a(k_{i,+y} - k_{i,-y})}{2(k_{i,+y} + k_{i,-y})}$$

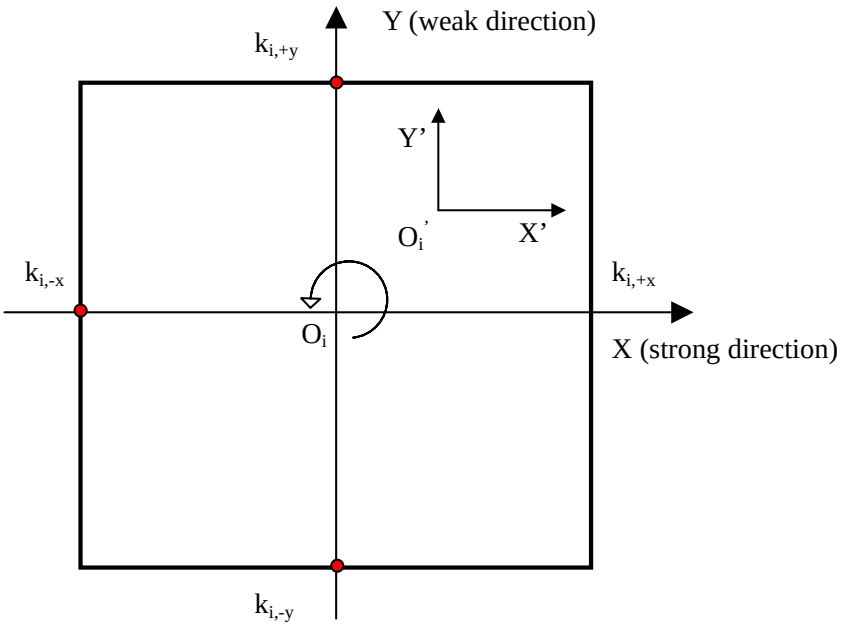


Figure 5.9: Floor Plan for Benchmark Structure.

where a is the width of floor ($a = 2.5$ m here).

The local stiffness matrix with respect to shear center O_i' can be computed (Yuen et al., 2004):

$$\mathbf{K}_i' = \begin{bmatrix} k_{ix} & 0 & 0 & -k_{ix} & 0 & 0 \\ 0 & k_{iy} & 0 & 0 & -k_{iy} & 0 \\ 0 & 0 & k_{it} & 0 & 0 & -k_{it} \\ -k_{ix} & 0 & 0 & k_{ix} & 0 & 0 \\ 0 & -k_{iy} & 0 & 0 & k_{iy} & 0 \\ 0 & 0 & -k_{it} & 0 & 0 & k_{it} \end{bmatrix} \tag{5.7}$$

where

$$k_{ix} = k_{i,+y} + k_{i,-y}$$

$$k_{iy} = k_{i,+x} + k_{i,-x}$$

$$k_{it} = (\frac{a}{2} - \bar{x}_i)^2 k_{i,+x} + (\frac{a}{2} - \bar{y}_i)^2 k_{i,+y} + (\frac{a}{2} + \bar{x}_i)^2 k_{i,-x} + (\frac{a}{2} + \bar{y}_i)^2 k_{i,-y}$$

The local stiffness matrix is transformed with respect to geometric center O_i via matrix $\bar{\mathbf{T}}$:

$$\mathbf{K}_i = \bar{\mathbf{T}}_i^T \mathbf{K}_i' \bar{\mathbf{T}}_i \tag{5.8}$$

where

$$\bar{\mathbf{T}}_i = \begin{bmatrix} \mathbf{T}_i^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_i^{-1} \end{bmatrix} \text{ with } \mathbf{T}_i = \begin{bmatrix} 1 & 0 & \bar{y}_i \\ 0 & 1 & -\bar{x}_i \\ 0 & 0 & 1 \end{bmatrix}$$

These constructed local stiffness matrices are assembled and used to simulate the training dataset consisting of modal frequencies and corresponding mode shapes measured at $-x, +y, -y$ faces (shown as red dots in Figure 5.9).

5.2.5.2 Prediction Phase

As before, the sample correlation functions are computed using the acceleration time histories at the reference channel and at the channels located on each face per floor. Acceleration time histories are generated from the full structural model for the benchmark phase I using the specified values of the damping, Δt, and noise level: 0.01, 0.004 sec, and 10%, respectively. Figure 5.10 shows the correlation functions at nodes 11, 13, 15, and 17 located on the first floor (see Figure 5.4). The extracted modal frequencies are summarized in Table 5.10 and the corresponding mode shapes are shown in Figures 5.11 and 5.12 (the third mode is the fundamental torsional mode which received no, or very little, excitation). The mode shapes at $-X$, $+Y$, and $-Y$ faces (i.e., the faces which are perpendicular to the $-X$, $+Y$, and $-Y$ axes, respectively) are shown for each damage pattern. Using these features, damage identification and damage severity estimation are performed.

5.2.5.3 Identification of Damage Locations

To develop the training dataset consisting of the damage signatures, 50% stiffness loss is assigned to the stiffness in each (weak and strong) direction as before, with the damage index set to 1 for damaged elements; both elements that contribute to

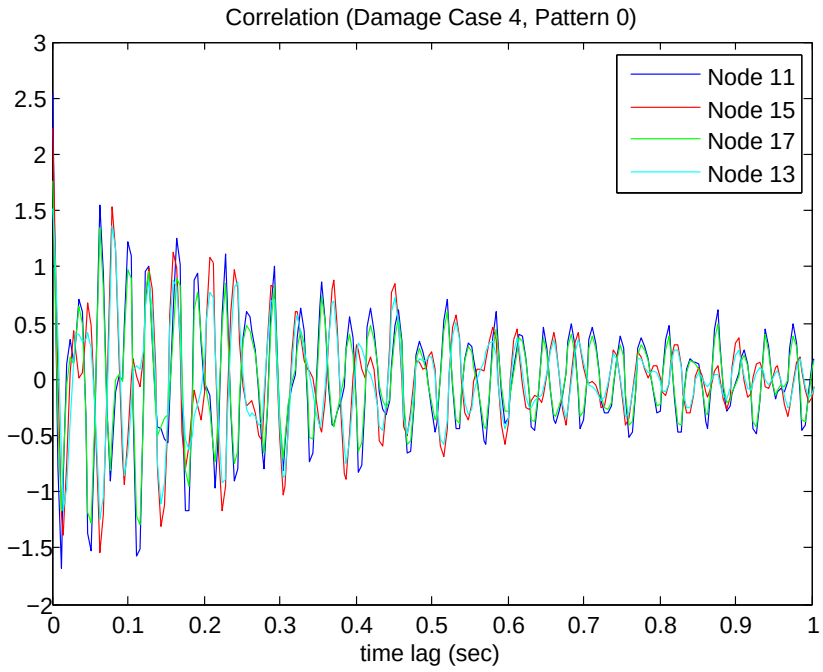


Figure 5.10: Correlations between Measurements at Reference Channel and Measured Accelerations at Node 11, 13, 15, and 17, respectively.

the stiffness in the corresponding direction have their stiffness reduced (for example, elements in the $+X$ and $-X$ faces when assigning damage in the weak direction). The prediction results are summarized in Table 5.14 and one can see that all actually damaged elements are included when using the criteria of 0.6 for the damage index threshold. Note that in this case, the stiffnesses are constrained to be the same in each direction, i.e., $\theta_{i,+y} = \theta_{i,-y}$ and $\theta_{i,+x} = \theta_{i,-x}$. All suspicious locations to be considered as possibly damaged are shown bolded and underlined in the table.

5.2.5.4 Assessment of Damage Severities

After identifying the potential damage locations, the damage severities are estimated using another trained RVM. Two levels of damage, 30% and 70%, are used at the locations obtained in the first step in order to generate the training dataset, and the prediction is performed using the features extracted with the imposed damage as shown in Table 5.15. As is shown in Table 5.16 and 5.17 for damage case 4 and 5, respectively, this two-step approach can successfully identify damage locations and

Damage Case	Damage Pattern	Story 1		Story 2		Story 3		Story 4	
		$\theta_{i,y}$	$\theta_{i,x}$	$\theta_{i,y}$	$\theta_{i,x}$	$\theta_{i,y}$	$\theta_{i,x}$	$\theta_{i,y}$	$\theta_{i,x}$
4	1	**1.48**	**2.05**	0.19	0.49	0.23	0.44	0.20	0.34
	2	**1.16**	**1.53**	-0.12	0.43	**0.76**	**0.85**	0.29	0.32
	3	-1.94	**0.95**	0.42	0.37	**1.93**	0.39	**2.59**	0.07
	4	0.06	**1.24**	0.57	**0.95**	**1.23**	0.59	0.35	0.28
5	1	**1.82**	**2.26**	0.19	0.43	0.11	0.44	-0.02	0.32
	2	**0.97**	**1.42**	0.42	0.33	**0.79**	**0.90**	0.07	0.26
	3	**3.60**	**0.90**	-1.29	0.43	**1.06**	0.48	-2.59	2.62
	4	-0.22	**1.27**	0.19	0.56	**1.63**	0.57	0.08	0.35
	5	-0.23	**1.30**	0.21	0.56	**1.63**	0.56	0.08	0.36
	6	**5.10**	**0.79**	**3.09**	-0.19	**1.38**	**0.71**	-7.01	-0.85

Table 5.14: Identified Damage for Damage Case 4 and 5 using Damage Signature.

their severities by applying the steps sequentially.

In damage patterns 5 and 6 (only for damage case 5), one brace in the 1^{st} story and the 1^{st}, 3^{rd} stories are removed with a loosened floor beam in the 1^{st} story for both patterns, respectively (refer to Table 5.9). Comparing the results from damage pattern 4 with those from damage pattern 5, the effect of loosing the floor connection is negligible as was stated in Yuen et al. (2004).

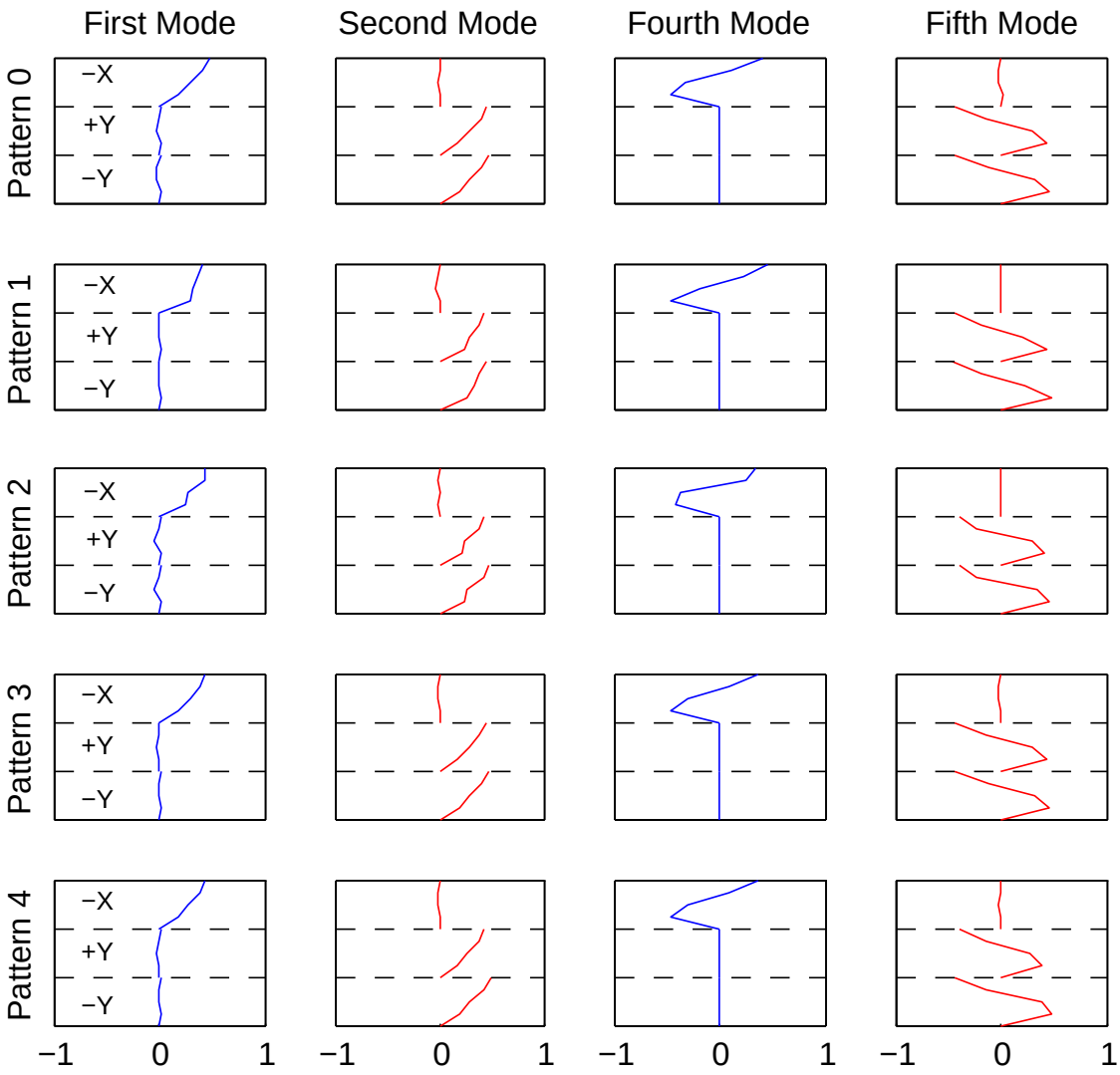


Figure 5.11: Identified Mode Shapes for Case 4, Pattern 0, 1, 2, 3, and 4.

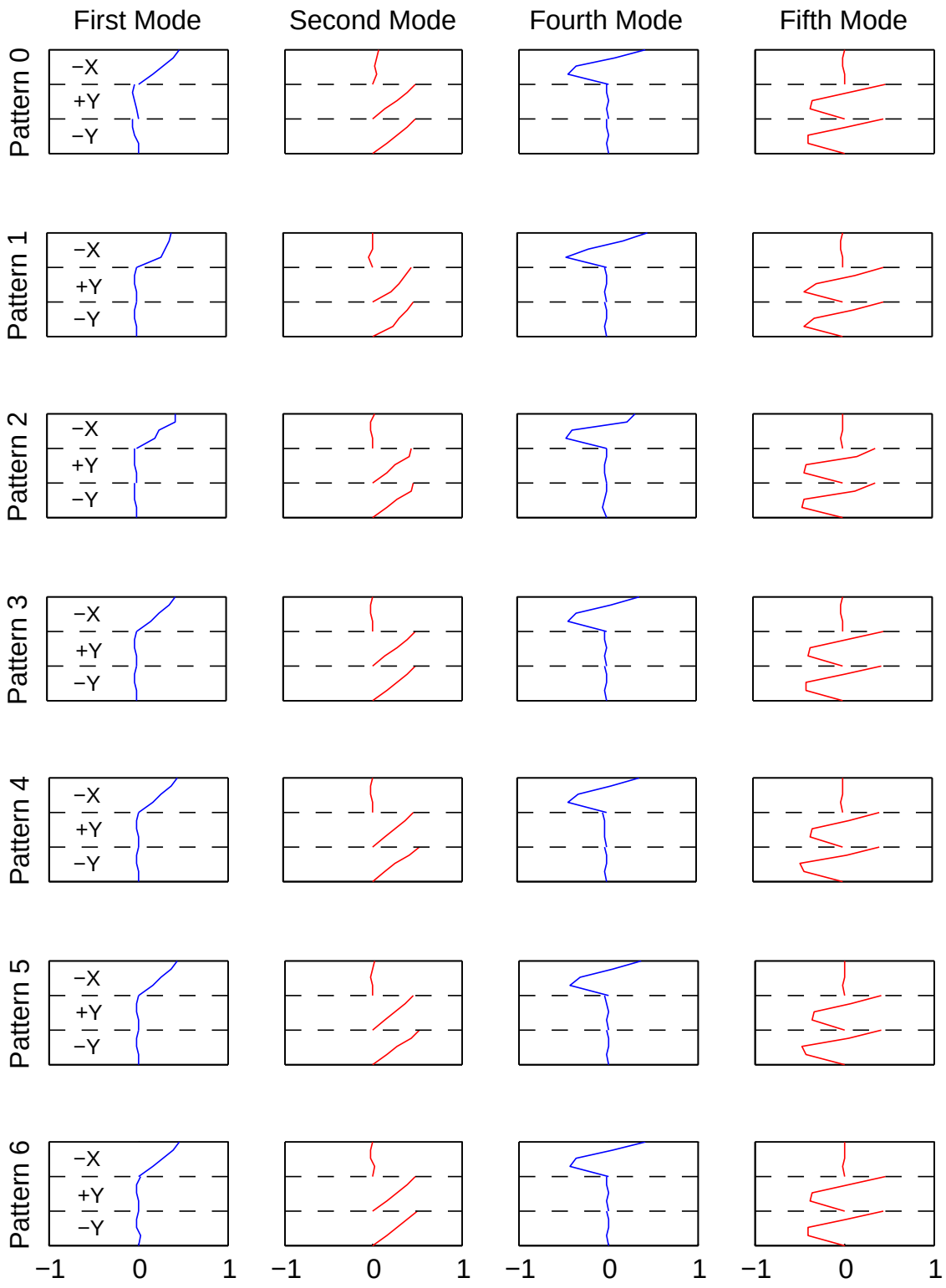


Figure 5.12: Identified Mode Shapes for Case 5, Pattern 0, 1, 2, 3, 4, 5, and 6.

Floor	Damage Pattern	$\theta_{i,-y}$	$\theta_{i,+x}$	$\theta_{i,+y}$	$\theta_{i,-x}$
1	1	**0.45**	**0.71**	**0.45**	**0.71**
	2	**0.45**	**0.71**	**0.45**	**0.71**
	3	0.00	**0.36**	0.00	0.00
	4	0.00	**0.36**	0.00	0.00
	5	0.00	**0.36**	0.00	0.00
	6	0.00	**0.12**	0.00	0.00
2	1	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.00
3	1	0.00	0.00	0.00	0.00
	2	**0.45**	**0.71**	**0.45**	**0.71**
	3	0.00	0.00	0.00	0.00
	4	**0.23**	0.00	0.03	0.00
	5	**0.23**	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.00
4	1	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.00

Table 5.15: Actual Stiffness Loss for Damage Case 5. (Damage Patterns 1 to 4 are also applied to Damage Case 4.)

Floor	Damage Pattern	$\theta_{i,-y}$	$\theta_{i,+x}$	$\theta_{i,+y}$	$\theta_{i,-x}$
1	1	**0.43**	**0.70**	**0.44**	**0.72**
	2	**0.42**	**0.69**	**0.43**	**0.71**
	3	0.00	**0.37**	0.00	0.00
	4	0.00	**0.30**	0.00	0.02
2	1	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.02
3	1	0.00	0.00	0.00	0.00
	2	**0.46**	**0.72**	**0.42**	**0.69**
	3	0.03	0.00	0.02	0.00
	4	**0.18**	0.00	0.03	0.00
4	1	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	3	0.02	0.00	0.00	0.00
	4	0.03	0.00	0.00	0.00

Table 5.16: Predicted Stiffness Loss for Damage Case 4.

Floor	Damage Pattern	$\theta_{i,-y}$	$\theta_{i,+x}$	$\theta_{i,+y}$	$\theta_{i,-x}$
1	1	**0.44**	**0.71**	**0.43**	**0.73**
	2	**0.43**	**0.71**	**0.42**	**0.70**
	3	0.03	**0.41**	0.02	0.07
	4	0.00	**0.36**	0.00	0.00
	5	0.00	**0.37**	0.00	0.00
	6	0.00	**0.21**	0.01	0.03
2	1	0.00	0.00	0.00	0.00
	2	0.00	0.01	0.00	0.00
	3	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00
	5	0.00	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.00
3	1	0.00	0.00	0.00	0.00
	2	**0.46**	**0.70**	**0.43**	**0.73**
	3	0.00	0.00	0.02	0.00
	4	**0.22**	0.00	0.02	0.00
	5	**0.22**	0.00	0.07	0.00
	6	0.04	0.00	0.02	0.00
4	1	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
	3	0.02	0.00	0.00	0.00
	4	0.08	0.00	0.03	0.00
	5	0.08	0.00	0.03	0.00
	6	0.09	0.00	0.00	0.00

Table 5.17: Predicted Stiffness Loss for Damage Case 5.

5.3 Conclusions

Parallel to the development of sensor technology, there is a need for efficient and reliable data processing tools for structural health monitoring. For SHM which deals with a large amount of sensor data with intrinsic errors, the Bayesian updating methodology is a powerful way to make probabilistic-based decisions, since it quantitatively treats all the uncertainties involved. In particular, Bayesian learning using RVM has significant potential for a systematic SHM methodology. Its advantages include:
(1) Quantitative procedure for meaningful probabilistic decision-making that explicitly treats all uncertainties rather than a deterministic result with no indication of how much confidence should be attached to the prediction.
(2) Automatic determination of the trade-off between the fit to the data and model complexity (i.e., no need to perform cross-validation).
(3) Clear use of a prior distribution that automatically prunes irrelevant kernel terms, which results in greatly reducing the number of parameters that must be estimated.
(4) Very efficient in the operating phase compared with the common SHM approach based on structural model updating (e.g., Ching and Beck, 2004)

In the classification application presented in this study, only a binary classification case is explored for SHM. In future work, an investigation will be made of multi-classification cases which deal with multiple classes of damage (i.e., different levels and locations).

A regression method based on the vector output RVM is introduced to determine the damage locations and severity from changes in the identified modal parameters. RVM automatically selects the most probable model class to provide the best predictions for damage assessments by maximizing the evidence for the model class based on the regularizing ARD prior. In regression problems, once the RVM is trained, it is efficient for on-line SHM based on extracting the selected feature vector from dynamic data. From the five-story shear building model and the IASC-ASCE benchmark structure examples, we conclude that the proposed vector output RVM shows promise for estimating both the damage locations and the damage severities from

changes in the structure's dynamic characteristics, such as its modal parameters, when the SHM is performed in two steps: first identify damage locations and then estimate damage severities.

Chapter 6

Concluding Remarks and Future Work

Recent achievements in modern sensing technology, such as wireless and digital sensor development, enables the collection of large amounts of data that contain less noise with less expense. These technological improvements necessitate accompanying development of sophisticated data analyzing methodologies. In this dissertation, a novel Bayesian learning method using an automatic relevance determination prior is demonstrated and it is extended to perform classification and regression with vector outputs.

In contrast to non-Bayesian methods, the proposed Bayesian methodology is shown to provide a probabilistic interpretation of the results with the consideration of all possible uncertainties, and with regularization to alleviate ill-conditioning and data over-fitting in inverse problems. Furthermore, in the prediction phase, the proposed methods work in real-time to render it of significant value for on-line earthquake early warning systems as well as structural health monitoring systems.

In Chapter 3, an application to earthquake early warning systems is demonstrated. For an abrupt calamity such as earthquakes, the availability of a warning system is of great value in reducing the loss of human lives and the operational loss of civil structures. By being able to classify measured accelerograms in real-time into near-source and far-source, a first step in an earthquake early warning system for large earthquakes is accomplished. As is shown, this newly-introduced method is capable

of providing a robust classification result, which lowers the misclassification rates and prediction errors by assigning an independent prior variance to each parameter. The proposed method can effectively adjust the trade-off between data-fit errors and model complexity by virtue of the ARD prior, which is shown to provide smaller prediction errors as judged by leave-one-out cross-validation and the calculated evidence for the model.

In Chapter 4, ground motion attenuation equations are estimated by using a regression model based on the Boore-Joyner attenuation model. The obtained result is also compared with a previous one using the Bayesian method with non-informative prior. The proposed method using the ARD prior yields the most probable model with the inclusion of one additional term of $(M-6)^2$ compared with the previous model. Another focus of this chapter is on the estimation of a non-linearly involved parameter h (the fictitious depth) which is estimated by model class selection and by stochastic simulation using a Markov Chain Monte Carlo simulation method. The same method for estimating a non-linearly involved parameter can also be applied for the width estimation of the radial basis function kernels in the Relevance Vector Machine.

The applications described in Chapter 3 and 4 show that the Bayesian learning method using the ARD prior plays an important role in a feature selection algorithm when the features extracted from measurements or observation are utilized as inputs via a sum of basis functions with unknown parameters as coefficients. Using model class selection to find the optimal hyperparameters (variances) in the prior, some of the coefficients become zero (Gaussian with zero mean and zero variance), thereby pruning out terms that prove to be irrelevant for predictions, as determined from the data. Therefore, one can choose the strategy to just let the algorithm sort out the relevant terms by initially including all seemingly relevant terms.

In Chapter 5, structural health monitoring (SHM) applications are investigated using the so-called Relevance Vector Machine (RVM). RVM is an extended version of the Bayesian learning method using the ARD prior which incorporates kernel basis functions for classification and regression. RVM classification is first applied to

SHM problems using simulated data from various FE models with different levels of noise added. The proposed methodology is able to provide good classification results for damage detection along with a quantification of the degree of belief in the result via associated probabilities. This probabilistic interpretation for damage detection has an advantage of great value in that it helps to give an importance ranking when inspecting a possibly damaged structure, that is, those locations with high probability of damage might be inspected first. However, the classification approach has a disadvantage that the number of training data may be too many which makes the training phase computationally expensive as the structure becomes more complicated (refer to Section 5.1.1.3). Although this problem can be handled using parallel computing capability, an effort to reduce the number of cases, such as using an expert's prior knowledge to recognize the structurally weak points, is valuable in making the classification approach more practical, rather than installing parallel computers.

Another strategy is to use RVM regression. Most previous work on RVM applied to regression problems has used a scalar output, as in the original theory, but in this study, RVM is applied using vector outputs to examine its effectiveness; this is an ongoing research topic in machine learning in order to make RVM a good tool for general regression problems. A two-step method to identify potentially-damaged locations first and then to estimate the corresponding damage severities only at the identified structural elements is presented and illustrated using a 5-story shear-building model and a 4-story IASCE-ASCE benchmark model using the phase I simulated data. It is shown that the two combined procedures are complementary to each other and they provide good estimates of the damage locations and associated severities.

Future work will include the extension of the classification method from a binary to a multi-class case, which can then be applied to soil liquefaction problems, fragility function estimation problems in performance-based earthquake engineering and SHM using real datasets. The regression method may be further applied as an effective regression tool to develop ground motion attenuation equations using the strong motion records in the PEER NGA database (*http://peer.berkeley.edu/nga/*). It can be also applied to the IASC-ASCE SHM benchmark simulation phase II

and experimental phase II, as well as to a bridge health monitoring benchmark (*http://people.cecs.ucf.edu/catbas/*).

In conclusion, Bayesian learning using the ARD prior and its kernel implementation in RVM for vector output regression are powerful tools to make robust predictions considering all possible inherent uncertainties and to provide a probabilistic interpretation helpful when making decisions. By using model class selection and regularization simultaneously, the proposed method makes on-line tasks from earthquake early warning to SHM possible, so that these tasks can be operated efficiently in real time.

Bibliography

Akaike, H. (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, 19 (6), 716–723.

Allen., R.M., and Kanamori, H. (2003). "The Potential for Earthquake Early Warning in Southern California." *Science 300*, 786–789.

Beck, J.L., May, B.S., and Polidori, D.C. (1994). "Determination of Modal Parameters from Ambient Vibration Data for Structural Health Monitoring." *Proceedings of the First World Conference on Structural Control*, Los Angeles, California, U.S.A.

Beck, J.L. (1996). "System Identification Methods applied to Measured Seismic Response." *Proceedings of the Eleventh World Conference on Earthquake Engineering*, Acapulco, Mexico.

Beck, J.L., and Katafygiotis, L.S. (1998). "Updating Models and Their Uncertainties: Bayesian Statistical Framework." *Journal of Engineering Mechanics*, 124, 455–461.

Beck, J.L., and Yuen, K.V. (2004). "Model Selection using Response Measurements: a Bayesian Probabilistic Approach." *Journal of Engineering Mechanics*, 130, 192–203.

Beck, J.V., Blackwell, B., and St. Clair, C.R. (1985). "Inverse Heat Conduction: Ill-posed Problems." *Wiley*, New York.

Bishop, C.M., and Tipping, M.E. (2003). *Advances in Learning Theory: Methods, Models and Applications.*, ISO Press, Amsterdam.

Boore, D.M., Joyner, W.B., and Fumal, T.E. (1993). "Estimation of Response Spectra and Peak Accelerations from Western North American Earthquakes: An Interim Report." *U.S. Geol. Surv. Open-File Rept. 93–509.*

Boore, D.M., Joyner, W.B., and Fumal, T.E. (1997). "Equations for Estimating Horizontal Response Spectra and Peak Acceleration from Western North American Earthquakes: A Summary of Recent Work." *Seismological Research Letters*, 68(1), 128–153.

Brownjohn, J., Tjin, S., Tan, G., and Tan, B. (2004). "Structural Health Monitoring Paradigm for Civil Infrastructures.", 1^{st} *FIG International Symposium on Engineering Surveys for Construction Works and Structural Engineering.*

Burges, C.J.C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition.", *Data Mining and Knowledge Discovery*, 2(2), 121–167.

Campbell, K.W. (1981). "Near-Source Attenuation of Peak Horizontal Acceleration." *Bulletin of Seismological Society of America*, 71(6), 2039–2070.

Cao, T.T., and Zimmerman, D.C. (1997). "A Procedure to Extract Ritz Vectors from Dynamic Testing Data." *Proceedings of the 15th International Modal Analysis Conference*, Orlando, FL, 2, 1036–1042.

Cawley, P., and Adams, R.D. (1979). "The Location of Defects in Structures from Measurements of Natural Frequencies." *Journal of Vibration and Acoustics*, 14(2), 49–57.

Ching, J., and Beck, J.L. (2004). "Bayesian Analysis of the Phase II IASC-ASCE Structural Health Monitoring Experimental Benchmark Data." *Journal of Engineering Mechanics*, 130(10), 1233–1244.

Ching, J., and Chen, Y.-J. (2006). "Transitional Markov Chain Monte Carlo Method for Bayesian Model Updating, Model Class Selection and Model Averaging." *Journal of Engineering Mechanics*, 133(7), 816–831.

Cua, G.B. (2005). "Creating the Virtual Seismologist: Developments in Ground Motion Characterization and Seismic Early Warning." *Ph. D. Thesis in Civil Engineering*, California Institute of Technology.

Ding, C.H.Q., and Dubchak, I. (2001). "Multi-class Protein Fold Recognition using Support Vector Machines and Neural Networks.", *Bioinfomatics*, 17(4), 349–358.

Duda, R.O., Hart, P.E., and Stork, D.G. (2000). "Pattern Classification." *Wiley-interscience*, New York.

Faul, A.C., and Tipping, M.E. (2002). "Analysis of Sparse Bayesian Learning." *Advances in Neural Information Processing Systems 14* , 383–389.

Groetsch, C.W. (1984). "The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind." *Pitman Advanced Publishing*, Boston.

Hartzell, S., and Heaton, T. (1983). "Inversion of Strong Ground Motion and Teleseismic Waveform Data for the Fault Rupture History of the 1979 Imperial Valley, California, earthquake." *Bull. Seism. Soc. Am.*, 73, 1553–1583.

Jaynes, E.T. (2003). Probability Theory: The Logic of Science. *University Press*, Cambridge.

Ji, C., Helmberger, D.V., Wald, D.J., and Ma, K.F. (2003). "Slip History and Dynamic Implication of 1999 Chi-Chi Earthquake." *J. Geophys. Res.*, 108(B9).

Johnson, E.A., Lam, H.F., Katafygiotis, L., and Beck, J.L. (2000). "A Benchmark Problem for Structural Health Monitoring and Damage Detection." *Proceedings of the* 14^{th} *ASCE Engineering Mechanics Conference*, Austin, Texas.

Johnson, E.A., Lam, H.F., Katafygiotis, L., and Beck, J.L. (2004). "Phase I IASC - ASCE Structural Health Monitoring Benchmark Problem Using Simulated Data." *Journal of Engineering Mechanics*, 130(3), 3–15.

Joyner, W.B., and Boore, D.M. (1981). "Peak Horizontal Acceleration and Velocity from Strong-Motion Records including Records from the 1979 Imperial Valley, California, earthquake." *Bull. Seim. Soc. Am.*, 71(6), 2011–2038.

Joyner, W.B., and Boore, D.M. (1982). "Prediction of Earthquake Response Spectra." *U.S. Geol. Surv. Open-File Rept. 82-977.*

Lam, H.F., Yuen, K.V., and Beck, J.L. (2006). "Structural Health Monitoring via Measured Ritz Vectors Utilizing Artificial Neural Networks." *Computer-Aided Civil and Infrastructure Engineering*, 21(4), 232–241.

Lee, H.S., Kim, Y.H., Park, C.J., and Park, H.W. (1999). "A New Spatial Regularization Scheme for the Identification of Geometric Shapes of Inclusions in Finite Bodies." *International Journal for Numerical Methods in Engineering*, 46(7), 973–992.

Mackay, D.J.C. (1992a). "The Evidence Framework Applied to Classification Networks." *Neural Computation*, 4, 720–736.

Mackay, D.J.C. (1992b). "Bayesian Interpolation." *Neural Computation*, 4(3), 415–447.

Mackay, D.J.C. (1994). "Bayesian Non-linear Modelling for the Prediction Competition." *ASHRAE Transactions*, V.100 Pt. 2, 1053–1062.

Mackay, D.J.C. (1995). "Probable Networks and Plausible Predictions - a Review of Practical Bayesian Methods for Supervised Neural Networks." *Network: Computation in Neural Systems*, 6(3), 469–505.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). "Multivariate Analysis." *Probability and Mathematical Statistics*, Academic Press.

Martinez, W.L., and Martinez, A.R. (2002). "Computational Statistics Handbook with Matlab." *CRC Press*, Florida.

Muto, M.M. (2006). "Application of Stochastic Simulation Methods to System Identification." *Ph. D. Thesis*, California Institute of Technology.

Muto, M.M., and Beck, J.L. (2007). "Bayesian Updating and Model Class Selection for Hysteretic Structural Models using Stochastic Simulation." *J. Vib. Control*, (in press).

Oh, C.K., and Beck, J.L. (2006). "Sparse Bayesian Learning for Structural Health Monitoring." *Proceedings of the 4th World Conference on Structural Control and Monitoring*, San Diego, CA.

Oh, C.K., Beck, J.L., and Yamada, M. (2007). "Bayesian Learning using Automatic Relevance Determination Prior for Earthquake Early Warning." *Journal of Engineering Mechanics*, (under review).

Oh, C.K., and Beck J.L. (2007). "Relevance Vector Machine Regression Applied to Structural Health Monitoring." *Procedings of the 3rd International Conference on Structural Health Monitoring of Intelligent Infrastructure*, Vancouver, Canada, Nov. 2007.

Park, H.W., Shin, S.B., and Lee, H.S. (2001). "Determination of an Optimal Regularization Factor in System Identification with Tikhonov Function for Linear Elastic Continua." *International Journal for Numerical Methods in Engineering*, 51(10), 1211–1230.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.*, Cambridge University Press, Cambridge.

Sekiguchi, H., and Iwata, T. (2002). "Rupture Process of the 1999 Kocaeli, Turkey, Earthquake Estimated from Strong-motion Wave Forms." *Bull. Seism. Soc. Am.*, 92, 300–311.

Schölkopf, B., and Smola, A.J. (2002). *Probability, Reliability and Statistical Methods in Engineering Design.*, Wiley, New York.

Sivia, D.S. (2000) "Data Analysis: A Bayesian Tutorial." *Clarendon Press*, Oxford.

Sohn, H., and Law, K.H. (2001) "Damage Diagnosis Using Experimental Ritz Vectors." *Journal of Engineering Mechanics*, 127(11), 1184–1193.

Thayananthan, A. (2005). "Template-based Pose Estimation and Tracking of 3D Hand Motion." *Ph. D. Thesis in Engineering*, University of Cambridge.

Tipping, M.E. (2000). "The Relevance Vector Machine." *Advances in Neural Information Processing Systems 12*, 652–658.

Tipping, M.E. (2001). "Sparse Bayesian Learning and the Relevance Vector Machine." *Journal of Machine Learning Research*, 1, 211–244.

Tipping, M.E., and Faul, A.C. (2003). "Fast Marginal Likelihood Maximisation for Sparse Bayesian Models." *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, Key West, FL, Jan 3–6.

Tipping, M.E. (2004). "Bayesian Inference: An Introduction to Principles and Practice in Machine Learning." *Advanced Lectures on Machine Learning*, 41–62, Springer.

Tsuboi, S., Komatitsch, D., Ji, C., and Tromp, J. (2003). "Broadband Modeling of the 2002 Denali Fault Earthquake on the Earth Simulator." *Phys. Earth Planet. Interiors*, 139, 305–312.

Vanik, M.W., Beck, J.L. and Au, S.K. (2000). "Bayesian Probabilistic Approach to Structural Health Monitoring." *Journal of Engineering Mechanics*, 126, 738–745.

Vapnik, V.N. (1998). "Statistical Learning Theory." *Wiley*, New York.

Wald, D.J., Heaton, T., and Helmberger, D.V. (1991). "Rupture Model of the 1989 Loma Prieta Earthquake from the Inversion of Strong Motion and Broadband Teleseismic Data." *Bull. Seism. Soc. Am.*, 81, 1540–1572.

Wald, D.J., and Heaton, T. (2004). "Spatial and Temporal Distribution of Slip for the 1992 Landers, California, earthquake." *Bull. Seism. Soc. Am.*, 84, 668–691.

Wald, D.J., Heaton, T., and Hudnut, K.W. (1996). "A Dislocation Model of the 1994 Northridge, California, Earthquake Determined from Strong-motion, GPS, and Leveling-line Data." *Bull. Seism. Soc. Am.*, 86, 49–70.

Wald, D.J. (1996). "Slip History of the 1995 Kobe, Japan, Earthquake Determined from Strong Motion, Teleseismic, and Geodetic Data." *J. Phys. Earth*, 44, 489–503.

Yamada, M., Heaton, T., and Beck, J.L. (2007). "Early Warning Systems for Large Earthquakes: Near-source versus Far-source Classification." *Bull. Seism. Soc. Am.*, (accepted for publication).

Yuen, K.V., Au, S.K., and Beck, J.L. (2004). "Two-Stage Structural Health Monitoring Approach for Phase I Benchmark Studies." *Journal of Engineering Mechanics*, 130, 16–33.

Yuen, K.V., and Lam, H.F. (2006). "On the Complexity of Artificial Neural Networks for Smart Structure Monitoring." *Engineering Structures*, 28(7), 977–984.

Appendix A

Posterior PDF and evidence by Using Bayes' Theorem

In this section, the detailed derivations for posterior PDF and evidence is presented for the regression problem. For classification, the same procedure is utilized leading to similar results which will be explained at the end. According to Bayes' theorem, the posterior can be expressed by likelihood, prior, and evidence:

$$\begin{aligned} \text{posterior} &= \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \\ p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}, \sigma^2) &= \frac{p(\mathcal{D}_N|\underline{\theta}, \underline{\alpha}, \sigma^2) \times p(\underline{\theta}|\underline{\alpha}, \sigma^2)}{p(\mathcal{D}_N|\underline{\alpha}, \sigma^2)} \end{aligned} \tag{A.1}$$

The likelihood and prior are defined as:

$$\begin{aligned} p(\mathcal{D}_N|\underline{\theta}, \sigma^2) &= \mathcal{N}(\mathbf{\Phi}\underline{\theta}, \sigma^2\mathbf{I}) \\ &= (2\pi\sigma^2)^{-N/2} \exp\Big[-\frac{1}{2\sigma^2}\|\underline{y} - \mathbf{\Phi}\underline{\theta}\|^2 \Big] \end{aligned}$$

$$\begin{aligned} p(\underline{\theta}|\underline{\alpha}, \sigma^2) &= \mathcal{N}(\underline{0}, \mathbf{A}^{-1}(\underline{\alpha})) \\ &= (2\pi)^{-\frac{N+1}{2}} |\mathbf{A}(\underline{\alpha})|^{\frac{1}{2}} \exp\Big[-\frac{1}{2}\underline{\theta}^T \mathbf{A}(\underline{\alpha})\underline{\theta} \Big] \end{aligned}$$

where $\mathbf{\Phi} = [\underline{\tau}(\underline{x}_1), ..., \underline{\tau}(\underline{x}_N)]^T \in \mathbb{R}^{N\times(N+1)}$, $\underline{\tau}(\underline{x}_i) = [1, k(\underline{x}_i, \underline{x}_1), ..., k(\underline{x}_i, \underline{x}_N)]^T$, and $\underline{\theta} = [\theta_0, \theta_1, ..., \theta_N]^T$.

In this section, the derivation of posterior PDF and evidence is presented based

on Bayes' theorem. These can be performed by expanding the known right-hand side of likelihood and prior in terms of $\underline{\theta}$, rather than using Bayes' theorem (A.1) directly.

$$\begin{aligned} p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}, \sigma^2) \times p(\mathcal{D}_N|\underline{\alpha}, \sigma^2) &= p(\mathcal{D}_N|\underline{\theta}, \underline{\alpha}, \sigma^2) \times p(\underline{\theta}|\underline{\alpha}, \sigma^2) \\ &= \underbrace{(2\pi\sigma^2)^{-\frac{N}{2}}(2\pi)^{-\frac{N+1}{2}}|\mathbf{A}(\underline{\alpha})|^{\frac{1}{2}}}_{<1>} \times \\ &\quad \underbrace{\exp\Big[-\frac{1}{2\sigma^2}||\underline{y}-\mathbf{\Phi}\underline{\theta}||^2 - \frac{1}{2}\underline{\theta}^T\mathbf{A}(\underline{\alpha})\underline{\theta}\Big]}_{<2>} \end{aligned} \quad \text{(A.2)}$$

$< 1 >$ in (A.2) can be transformed,

$$\begin{aligned} < 1 > &= (2\pi\sigma^2)^{-\frac{N}{2}}(2\pi)^{-\frac{N+1}{2}}|\mathbf{A}(\underline{\alpha})|^{\frac{1}{2}} \\ &= (2\pi)^{-\frac{N}{2}}(2\pi)^{-\frac{N+1}{2}}((\sigma^2)^{-N}|\mathbf{A}(\underline{\alpha})|)^{\frac{1}{2}} \\ &= (2\pi)^{-\frac{N}{2}}(2\pi)^{-\frac{N+1}{2}}((\sigma^2)^{-N}|\mathbf{A}(\underline{\alpha})|)^{\frac{1}{2}} \\ &= (2\pi)^{-\frac{N}{2}}(2\pi)^{-\frac{N+1}{2}}|\mathbf{C}|^{-\frac{1}{2}}|\hat{\mathbf{\Sigma}}|^{-\frac{1}{2}} \end{aligned} \quad \text{(A.3)}$$

using determinant identity,

$$\begin{aligned} |\mathbf{C}| &= |\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}(\underline{\alpha})^{-1}\mathbf{\Phi}^T| \\ &= |\mathbf{A}(\underline{\alpha})|^{-1}|\sigma^2\mathbf{I}||\mathbf{A}(\underline{\alpha}) + \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}| \\ &= |\mathbf{A}(\underline{\alpha})|^{-1}|\sigma^2\mathbf{I}||\mathbf{\Sigma}|^{-1} \\ &= (\sigma^2)^N|\mathbf{A}(\underline{\alpha})|^{-1}|\mathbf{\Sigma}|^{-1} \end{aligned}$$

$< 2 >$ in (A.2) can be transformed,

$$\begin{aligned}
< 2 > &= -\frac{1}{2\sigma^2}||\underline{y} - \mathbf{\Phi}\underline{\theta}||^2 - \frac{1}{2}\underline{\theta}^T\mathbf{A}(\underline{\alpha})\underline{\theta} \\
&= -\frac{1}{2}\Big[\underline{\theta}^T(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{A}(\underline{\alpha}))\underline{\theta} - 2\sigma^{-2}\underline{y}^T\mathbf{\Phi}\underline{\theta} + \sigma^{-2}\underline{y}^T\underline{y}\Big] \\
&= -\frac{1}{2}\Big[(\underline{\theta} - \hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \hat{\underline{\theta}}) - \hat{\underline{\theta}}^T\hat{\mathbf{\Sigma}}^{-1}\hat{\underline{\theta}} + \sigma^{-2}\underline{y}^T\underline{y}\Big] \\
&\quad \text{since, } 2\hat{\underline{\theta}}^T\hat{\mathbf{\Sigma}}^{-1}\underline{\theta} - 2\sigma^{-2}\underline{y}^T\mathbf{\Phi}\underline{\theta} = 0 \\
&= -\frac{1}{2}\Big[(\underline{\theta} - \hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \hat{\underline{\theta}}) + \underline{y}^T(\sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{\Phi}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\sigma^{-2})\underline{y}\Big] \\
&= -\frac{1}{2}\Big[(\underline{\theta} - \hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \hat{\underline{\theta}}) + \underline{y}^T(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}(\underline{\alpha})^{-1}\mathbf{\Phi}^T)^{-1}\underline{y}\Big] \\
&= -\frac{1}{2}\Big[(\underline{\theta} - \hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \hat{\underline{\theta}}) + \underline{y}^T\mathbf{C}^{-1}\underline{y}\Big] \qquad \text{(A.4)}
\end{aligned}$$

using Woodbury inversion identity,

$$\sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{\Phi}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\sigma^{-2} = (\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}(\underline{\alpha})^{-1}\mathbf{\Phi}^T)^{-1}$$

Combining (A.3) and (A.4) provides expressions on posterior PDF and evidence given by:

$$\begin{aligned}
p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}, \sigma^2) &= (2\pi)^{-\frac{N+1}{2}}|\hat{\mathbf{\Sigma}}|^{-1/2}\exp\Big[-\frac{1}{2}(\underline{\theta} - \hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \hat{\underline{\theta}})\Big] \\
p(\mathcal{D}_N|\underline{\alpha}, \sigma^2) &= (2\pi)^{-\frac{N}{2}}|\mathbf{C}|^{-\frac{1}{2}}\exp\Big[-\frac{1}{2}\underline{y}^T\mathbf{C}^{-1}\underline{y}\Big]
\end{aligned}$$

where $\hat{\mathbf{\Sigma}}$, $\hat{\underline{\theta}}$, and $\mathbf{C}$ are defined as before.

For classification problem, substitute $\mathbf{C} = \mathbf{B}^{-1} + \mathbf{\Phi}\mathbf{A}(\underline{\alpha})^{-1}\mathbf{\Phi}^T$ with $\mathbf{B} = \text{diag}(\phi_1(\underline{\theta})\{1 - \phi_1(\underline{\theta})\}, ..., \phi_N(\underline{\theta})\{1 - \phi_N(\underline{\theta})\})$.

Appendix B

Laplace Approximation

In classification, Laplace approximation is applied to posterior PDF. In this section, the detailed mathematical procedure is presented. This Laplace approximation, as explained before, is a quadratic approximation of the ln-posterior around the most probable value, $\hat{\underline{\theta}}$, given by maximization of the posterior PDF, leading to a Gaussian distribution with mean $\hat{\underline{\theta}}$ and covariance matrix $\hat{\mathbf{\Sigma}}$ which is the inverse of the negative of the Hessian matrix of the ln-posterior.

For a given value of $\underline{\alpha}$, the log-posterior is:

$$\begin{aligned} \ln[p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})] &= \sum_{n=1}^{N} \ln[P(y_n|\underline{\theta}, \underline{x}_n)] + \ln[p(\underline{\theta}|\underline{\alpha})] \\ &= \sum_{n=1}^{N} \Big[y_n \cdot \ln \phi_n(\underline{\theta}) + (1 - y_n) \cdot \ln\{1 - \phi_n(\underline{\theta})\} \Big] \\ &\quad -\frac{1}{2}\underline{\theta}^T \mathbf{A}(\underline{\alpha})\underline{\theta} \end{aligned} \tag{B.1}$$

where $\mathbf{A}(\underline{\alpha}) = \text{diag}(\alpha_0, \alpha_1, ..., \alpha_N)$ and $\phi_n(\underline{\theta}) = \phi(f(\underline{x}_n|\underline{\theta}))$.

The most probable values $\hat{\underline{\theta}}(\underline{\alpha})$ are estimated by equating the first derivative of

(B.1) to zero:

$$\begin{aligned}\frac{\partial \ln[p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha})]}{\partial\theta_j} &= \sum_{n=1}^{N}\Big[y_n\cdot\frac{\partial\ln\phi_n(\underline{\theta})}{\partial\theta_j}+(1-y_n)\cdot\frac{\partial\ln\{1-\phi_n(\underline{\theta})\}}{\partial\theta_j}\Big]-\alpha_j\theta_j \\ &= \sum_{n=1}^{N}\Big[\underline{\tau}(\underline{x}_n)_j\Big\{y_n\cdot(1-\phi_n(\underline{\theta}))-(1-y_n)\cdot\phi_n(\underline{\theta})\Big\}\Big]-\alpha_j\theta_j \\ &= \sum_{n=1}^{N}\Big[\underline{\tau}(\underline{x}_n)_j\Big\{y_n-\phi_n(\underline{\theta})\Big\}\Big]-\alpha_j\theta_j=0 \qquad \text{(B.2)}\end{aligned}$$

where $\underline{\tau}(\underline{x}_n)_j$ is the jth element of $\underline{\tau}(\underline{x}_n)$ defined as $\underline{\tau}(\underline{x}_n)=[1,k(\underline{x}_n,\underline{x}_1),...,k(\underline{x}_n,\underline{x}_N)]^T\in\mathbb{R}^{N+1}$.

Equation (B.2) gives

$$\mathbf{\Phi}^T(\underline{y}-\underline{\psi})=\mathbf{A}(\underline{\alpha})\underline{\theta} \qquad \text{(B.3)}$$

where $\mathbf{\Phi}=[\underline{\tau}_1,...,\underline{\tau}_N]^T\in\mathbb{R}^{N\times(N+1)}$, and $\underline{\psi}=[\phi_1(\underline{\theta}),...,\phi_N(\underline{\theta})]^T\in\mathbb{R}^N$.

Because of the non-linearity of $\underline{\theta}$ (in $\underline{\psi}$ at left-hand-side), use Taylor expansion. Then,

$$\begin{aligned}\phi_n(\underline{\theta}) &= \phi_n(\hat{\underline{\theta}})+\sum_{j=0}^{N}\frac{\partial\phi_n(\underline{\theta})}{\partial\theta_j}(\theta_j-\hat{\theta}_j)+\text{higher order term} \\ &\cong \phi_n(\hat{\underline{\theta}})+\sum_{j=0}^{N}\phi_n(\underline{\theta})\{1-\phi_n(\underline{\theta})\}\underline{\tau}(\underline{x}_n)_j(\theta_j-\hat{\theta}_j) \qquad \text{(B.4)}\end{aligned}$$

which gives

$$\underline{\psi}\cong\hat{\underline{\psi}}+\mathbf{B}\mathbf{\Phi}(\underline{\theta}-\hat{\underline{\theta}}) \qquad \text{(B.5)}$$

where $\mathbf{B}=\text{diag}(\phi_1(\underline{\theta})\{1-\phi_1(\underline{\theta})\},...,\phi_N(\underline{\theta})\{1-\phi_N(\underline{\theta})\})$.

Substituting (B.5) into (B.3) gives,

$$\begin{aligned} & \mathbf{\Phi}^T(\underline{y} - \underline{\psi}) = \mathbf{A}(\underline{\alpha})\underline{\theta} \\ \rightarrow \quad & \mathbf{\Phi}^T(\underline{y} - \hat{\underline{\psi}} - \mathbf{B}\mathbf{\Phi}(\underline{\theta} - \hat{\underline{\theta}})) = \mathbf{A}(\underline{\alpha})\underline{\theta} \\ \rightarrow \quad & \underline{\theta} = (\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A}(\underline{\alpha}))^{-1}\mathbf{\Phi}^T(\underline{y} - \hat{\underline{\psi}} + \mathbf{B}\mathbf{\Phi}\hat{\underline{\theta}}) \\ \rightarrow \quad & \underline{\theta} = \hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\mathbf{B}\hat{\underline{y}} \end{aligned} \tag{B.6}$$

where $\hat{\mathbf{\Sigma}} = (\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A}(\underline{\alpha}))^{-1}$ proved shortly and $\hat{\underline{y}} = \mathbf{\Phi}^T(\mathbf{B}^{-1}(\underline{y} - \hat{\underline{\psi}}) + \mathbf{\Phi}\hat{\underline{\theta}}) \in \mathbb{R}^N$.

The inverse covariance matrix is $\hat{\mathbf{\Sigma}}^{-1}(\underline{\alpha}) = -\nabla_{\underline{\theta}}\nabla_{\underline{\theta}} \ln p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})$ evaluated at $\hat{\underline{\theta}}(\underline{\alpha})$:

$$\begin{aligned} \frac{\partial^2 \ln[p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})]}{\partial\theta_j\partial\theta_k} &= -\sum_{n=1}^{N} \underline{\tau}(\underline{x}_n)_j \frac{\partial\phi_n(\underline{\theta})}{\partial\theta_k} - \alpha_k \\ &= -\sum_{n=1}^{N} \underline{\tau}(\underline{x}_n)_j \cdot \phi_n(\underline{\theta})\{1 - \phi_n(\underline{\theta})\} \cdot \underline{\tau}(\underline{x}_n)_k - \alpha_k \end{aligned} \tag{B.7}$$

which gives,

$$\hat{\mathbf{\Sigma}}^{-1}(\underline{\alpha}) = -\nabla_{\underline{\theta}}\nabla_{\underline{\theta}} \ln p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}) = (\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A}(\underline{\alpha})) \tag{B.8}$$

Appendix C

Bayesian Model Class Selection

C.1 Hyperparameter Optimization

In this section, detailed procedure for optimizing hyperparameters, in other words, Bayesian model class selection is presented. In this study, the most plausible hyperparameter $\underline{\alpha} \in \mathcal{A}$ and σ^2 (in regression only) based on data $\mathcal{D}_N$ is selected by finding $\hat{\underline{\alpha}}$ and $\hat{\sigma}^2$ that maximizes $p(\mathcal{D}_N|\underline{\alpha}, \sigma^2)$, equivalently $\mathcal{L}(\underline{\alpha}, \sigma^2) = \ln p(\mathcal{D}_N|\underline{\alpha}, \sigma^2)$, considering a uniform prior on $\underline{\alpha}$ and σ^2 as explained in Chapter 2.

$$
\begin{aligned}
\mathcal{L}(\underline{\alpha}, \sigma^2) &= \ln p(\mathcal{D}_N|\underline{\alpha}, \sigma^2) \\
&= \ln \int_{-\infty}^{\infty} p(\mathcal{D}_N|\underline{\theta}, \sigma^2) p(\underline{\theta}|\underline{\alpha}) d\underline{\theta} \\
&= -\frac{1}{2}\Big[N \ln 2\pi + \ln|\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T| + \underline{y}^T(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T)^{-1}\underline{y}\Big] \\
&= -\frac{1}{2}\Big[N \ln 2\pi + \ln|\mathbf{C}| + \underline{y}^T\mathbf{C}^{-1}\underline{y}\Big] \\
&= -\frac{1}{2}\Big[N \ln 2\pi + \ln|\mathbf{C}_{-i}| + \underline{y}^T\mathbf{C}_{-i}^{-1}\underline{y} \\
&\quad - \ln\alpha_i + \ln(\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i) - \frac{(\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{y})^2}{\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i}\Big] \\
&= \mathcal{L}(\alpha_{-i}, \sigma^2) + \frac{1}{2}\Big[\ln\alpha_i - \ln(\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i) + \frac{(\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{y})^2}{\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i}\Big] \\
&= \mathcal{L}(\alpha_{-i}, \sigma^2) + \frac{1}{2}\Big[\ln\alpha_i - \ln(\alpha_i + S_i) + \frac{Q_i^2}{\alpha_i + Si}\Big] \\
&= \mathcal{L}(\alpha_{-i}, \sigma^2) + l(\alpha_i, \sigma^2)
\end{aligned}
\tag{C.1}
$$

where $S_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{\tau}_i$ and $Q_i = \underline{\tau}_i^T \mathbf{C}_{-i}^{-1} \underline{\tau}_i$.

The hyperparameters $\underline{\hat{\alpha}}$ and $\hat{\sigma}^2$ to maximize (C.1) can be estimated analytically by equating the derivative of (C.1) with zero. Since only $l(\alpha_i, \sigma^2)$ is related with α_i,

$$\begin{aligned} \frac{\partial \mathcal{L}(\underline{\alpha}, \sigma^2)}{\partial \alpha_i} &= \frac{\partial l(\alpha_i, \sigma^2)}{\partial \alpha_i} \\ &= \frac{1}{2}\Big[\frac{1}{\alpha_i} - \frac{1}{\alpha_i + S_i} - \frac{Q_i^2}{(\alpha_i + S_i)^2}\Big] \\ &= \frac{\alpha_i^{-1} S_i^2 - (Q_i^2 - S_i)}{2(\alpha_i + S_i)^2} \end{aligned} \tag{C.2}$$

Equating (C.2) $= 0$ gives two stationary points such as:

$$\hat{\alpha}_i = \begin{cases} \infty \\ \frac{S_i^2}{Q_i^2 - S_i} \end{cases} \tag{C.3}$$

with the constraint of $Q_i^2 > S_i$ for α_i should be positive as a variance.

The second derivative of (C.1) provides more information on the nature around the two stationary solutions in (C.3). Differentiation of (C.1) with respect to α_i once more gives:

$$\frac{\partial^2 \mathcal{L}(\underline{\alpha}, \sigma^2)}{\partial \alpha_i^2} = \frac{-\alpha_i^{-2} S_i^2 (\alpha_i + S_i)^2 - 2(\alpha_i + S_i)[\alpha_i^{-1} S_i^2 - (Q_i^2 - S_i)]}{2(\alpha_i + S_i)^4} \tag{C.4}$$

For $\alpha_i = \frac{S_i^2}{Q_i^2 - S_i}$,

$$\frac{\partial^2 \mathcal{L}(\underline{\alpha}, \sigma^2)}{\partial \alpha_i^2} = \frac{-S_i^2}{a \alpha_i^2 (\alpha_i + S_i)^2} \tag{C.5}$$

Since this is always negative, $\mathcal{L}(\underline{\alpha}, \sigma^2)$ in (C.1) has a maximum, with the constraint of $Q_i^2 > S_i$.

For $\alpha_i = \infty$, not only the second derivative in (C.4), but also further derivatives give zero. The sign of the first derivative in (C.2), however, depends on $-(Q_i^2 - S_i)$, such as:

- If $Q_i > S_i$, then the first derivative is negative which leads that $\mathcal{L}(\underline{\alpha}, \sigma^2)$ has a maximum at $\alpha_i = \frac{S_i^2}{Q_i^2 - S_i}$.

- If $Q_i < S_i$, then the first derivative is positive, meaning that $\mathcal{L}(\underline{\alpha}, \sigma^2)$ has a maximum at $\alpha_i = \infty$.

- If $Q_i = S_i$, $\mathcal{L}(\underline{\alpha}, \sigma^2)$ has a maximum at $\alpha_i = \infty$, since two stationary points in (C.3) becomes identical.

Therefore, the optimized hyperparameters for Bayesian model class selection can be summarized as (Faul and Tipping, 2002):

$$\hat{\alpha}_i = \begin{cases} \infty, & \text{if } Q_i^2 \leq S_i \\ \frac{S_i^2}{Q_i^2 - S_i}, & \text{if } Q_i^2 > S_i \end{cases} \tag{C.6}$$

C.2 Noise Variance Optimization

For the noise variance σ^2, a re-estimation equation can be derived as follows. Using the determinant identity (C.7) (Mardia et al., 1979) and Woodbury inversion identity (C.8),

$$|\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T| = |\mathbf{A}^{-1}||\sigma^2\mathbf{I}||\mathbf{A} + \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}| \tag{C.7}$$

$$(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T)^{-1} = \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{\Phi}(\mathbf{A} + \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\sigma^{-2} \tag{C.8}$$

with

$$|\sigma^2\mathbf{I}| = (\sigma^2)^N|\mathbf{I}| = (\sigma^2)^N \tag{C.9}$$

$\mathcal{L}(\underline{\alpha}, \sigma^2)$ in (C.1) becomes,

$$\begin{aligned} \mathcal{L}(\underline{\alpha}, \sigma^2) &= -\frac{1}{2}\Big[N\ln 2\pi + \ln|\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T| + \underline{y}^T(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T)^{-1}\underline{y}\Big] \\ &= -\frac{1}{2}\Big[N\ln 2\pi - \ln|\mathbf{A}| + N\ln\sigma^2 + \ln|\mathbf{A} + \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}| \\ &\quad + \sigma^{-2}\underline{y}^T\underline{y} - \sigma^{-4}\underline{y}^T\mathbf{\Phi}(\mathbf{A} + \sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\underline{y}\Big] \\ &= -\frac{1}{2}\Big[N\ln 2\pi - \ln|\mathbf{A}| + N\ln\sigma^2 - \ln|\hat{\mathbf{\Sigma}}| + \sigma^{-2}\underline{y}^T(\underline{y} - \mathbf{\Phi}\hat{\underline{\theta}})\Big] \\ &= -\frac{1}{2}\Big[N\ln 2\pi - \ln|\mathbf{A}| + N\ln\sigma^2 - \ln|\hat{\mathbf{\Sigma}}| + \sigma^{-2}||\underline{y} - \mathbf{\Phi}\hat{\underline{\theta}}||^2 + \hat{\underline{\theta}}^T\mathbf{A}\hat{\underline{\theta}}\Big] \end{aligned} \tag{C.10}$$

using

$$\begin{aligned}
\sigma^{-2}\underline{y}^T(\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}) &= \sigma^{-2}(\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}+\mathbf{\Phi}\hat{\underline{\theta}})^T(\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}) \\
&= \sigma^{-2}||\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}||^2+\sigma^{-2}\underline{y}^T\mathbf{\Phi}\hat{\underline{\theta}}-\sigma^{-2}\hat{\underline{\theta}}^T\mathbf{\Phi}^T\mathbf{\Phi}\hat{\underline{\theta}} \\
&= \sigma^{-2}||\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}||^2+(\sigma^{-2}\underline{y}^T\mathbf{\Phi}\mathbf{\Sigma}^T)\mathbf{\Sigma}^{-1}\hat{\underline{\theta}}-\sigma^{-2}\hat{\underline{\theta}}^T\mathbf{\Phi}^T\mathbf{\Phi}\hat{\underline{\theta}} \\
&= \sigma^{-2}||\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}||^2+\hat{\underline{\theta}}^T\mathbf{\Sigma}^{-1}\hat{\underline{\theta}}-\sigma^{-2}\hat{\underline{\theta}}^T\mathbf{\Phi}^T\mathbf{\Phi}\hat{\underline{\theta}} \\
&= \sigma^{-2}||\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}||^2+\hat{\underline{\theta}}^T\mathbf{A}\hat{\underline{\theta}}
\end{aligned}$$

where $\hat{\mathbf{\Sigma}}=(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}+\mathbf{A})^{-1}$ and $\hat{\underline{\theta}}=\sigma^{-2}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\underline{y}$.

The derivative of (C.10) with respect to σ^{-2} (for simplicity, differentiate with respect to σ^{-2} instead of σ^2) is (Tipping, 2001):

$$\begin{aligned}
\frac{\partial\mathcal{L}(\underline{\alpha},\sigma^2)}{\partial\sigma^{-2}} &= \frac{\partial\mathcal{L}_1(\sigma^2)}{\partial\sigma^{-2}} \\
&= \frac{1}{2}\Big[N\sigma^2-||\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}||^2-\mathrm{tr}(\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{\Phi})\Big] \\
&= \frac{1}{2}\Big[N\sigma^2-||\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}||^2-\sigma^2\sum_i\gamma_i\Big] \qquad \text{(C.11)}
\end{aligned}$$

using

$$\begin{aligned}
&\mathbf{\Sigma}^{-1}=A+\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} \\
&\rightarrow \mathrm{tr}(\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{\Phi})=\sigma^2\mathrm{tr}(\mathbf{I}-\mathbf{\Sigma}\mathbf{A})=\sigma^2\sum_i\gamma_i \qquad \text{(C.12)}
\end{aligned}$$

Equating (C.11) to zero gives an equation for updated $\hat{\sigma}^2$ after each iteration:

$$(\hat{\sigma}^2)=\frac{\|\underline{y}-\mathbf{\Phi}\hat{\underline{\theta}}\|^2}{N-\sum_i\gamma_i} \qquad \text{(C.13)}$$

where $\gamma_i=1-\alpha_i\mathbf{\Sigma}_{ii}$ and $\mathbf{\Sigma}_{ii}$ is the ith diagonal element of $\mathbf{\Sigma}$ computed with current $\underline{\alpha}$ and σ^2.

For the classification problem, substitute $\mathbf{C}=\mathbf{B}^{-1}+\mathbf{\Phi}\mathbf{A}(\underline{\alpha})^{-1}\mathbf{\Phi}^T$ with $\mathbf{B}=\mathrm{diag}(\phi_1(\underline{\theta})\{1-\phi_1(\underline{\theta})\},...,\phi_N(\underline{\theta})\{1-\phi_N(\underline{\theta})\})$ and $\hat{\mathbf{\Sigma}}^{-1}=(\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi}+\mathbf{A}(\underline{\alpha}))$ with no

noise variance.