PROBABILISTIC MODELS BASED ON EXPERIMENTAL OBSERVATIONS USING SPARSE BAYES METHODOLOGY

BY

MINSEO KIM

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Adviser:

Associate Professor Junho Song

ABSTRACT

The quality of people's lives depends on safe and reliable infrastructure. However, there exist various types of uncertainties that may influence performance of structures, which could cause unexpected failures. Therefore, it is important to quantify the risk of such failures through systematic treatment of uncertainties and make a risk-informed decision. As an effort to predict uncertain performance of structural elements based on experimental observations, the principle of Bayesian inference has been often used. In this study, the recently proposed Sparse Bayes method is reviewed and tested by use of an experimental database of the shear strengths of reinforced concrete beams without stirrups. The performance of the Sparse Bayes method is demonstrated through comparison with existing methods such as least-square method and penalized least-square method. The Sparse Bayes method is further developed to identify a few representative points in the parameter space by grouping the relevant vectors identified by the method using the *k*-means clustering algorithm. The study confirms that the Sparse Bayes method has wide applicability, ability to achieve an optimal fitting, and efficiency. Therefore, the method can potentially provide a useful tool to develop powerful probabilistic models for various problems based on experimental observations.

Dedicated to Sangbum Kim and Yeojoo Nam

ACKNOWLEDGEMENTS

I would like to thank my adviser Professor Junho Song for his time and efforts to help me learn and overcome lots of hardships throughout my graduate life in University of Illinois at Urbana-Champaign. He has not only been a great adviser to guide me choosing the topic to study and write a thesis for the research but also has been a supporter and mentor to encourage me when I failed on circumstances and taught me with life lessons that I will always remember. I sincerely thank Professor Junho Song for being my adviser and wish him best luck on everything. Last but not the least, I would like to thank my mom and dad for their support and encouragement.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This study reviews the concepts of Sparse Bayesian learning and the relevance vector machine (Tipping 2001) and tests the method as a potential tool for developing strength models of structural elements based on experimental observations. The method is further developed to gain important insight about the given phenomenon, which may be useful for future engineering research and design practice.

1.1 Motivation

Nowadays various engineering communities recognize significant uncertainties in phenomenon and mathematical models, as well as their critical impacts on the actual performance of engineering systems. As an effort to estimate uncertain parameters in a mathematical model effectively using available experimental data, the principle of Bayesian inference has been studied by many researchers (Neal 1996, Berger 1985). In particular, theories of Bayesian inference were recently used to achieve optimal level of fitting during a regression analysis, i.e. avoiding over- and under-fitting (Tipping 2004). For example, Tipping (2004) introduced a sparsity concept to Bayesian inference to overcome computational challenges in full numerical integration calculations, which are needed to achieve the optimal level of fitting. In this study, the Sparse Bayesian method (Tipping 2001) is reviewed and used to develop strength models of reinforced concrete beams using an experimental database to test the applicability of the Sparse Bayes method to structural engineering problems and to explore further development of the method to gain insight of a given phenomenon from experimental observations and to help decision-making process in engineering designs.

1.2 Objectives and Scope

The primary objective of this study is to test the Sparse Bayes methodology as a tool for developing statistical model of structural engineering phenomena. This study also aims at providing practical guidelines and decision frameworks for using the Sparse Bayes methodology in predicting strengths of structural elements and designing appropriately. A preliminary study is also performed to explore the possibility of using the Sparse Bayes methodology to gain important insights in civil engineering problems.

1.3 Organization

Following this introductory chapter, Chapter 2 provides an in-depth review on the Sparse Bayes methodology with focus on theories of Bayesian Inference. The chapter also describes the MATLAB® codes created by Mike Tipping for the Sparse Bayes methodology (Tipping 2009), which were used in this study. In Chapter 3, three different statistical methods – Least-Square, Penalized Least-Square, Bayesian Inference are reviewed and compared using the experimental database of the shear strengths of reinforced concrete beams with no stirrups (Reineck *et al*. 2003), which was used by Song *et al*. (2010) for the development of another Bayesian inference methodology. The performance of the Sparse Bayes method is discussed in detail through this comparison. Chapter 4 describes the further analysis of the database introduced in previous chapter using the Sparse Bayes methodology. Using the relevant vector concept, influential parameters and critical domains in the design space are identified for the shear strengths of reinforced beams, as by-products of the Sparse Bayes methodology. Finally, a summary of the major findings and suggestions for future study are presented in Chapter 5.

CHAPTER 2

REVIEW OF SPARSE BAYES METHODOLOGY

This chapter reviews the Sparse Bayes concept and the Sparse Bayes methodology for optimal statistical inference based on Tipping (2001).

2.1 Sparse Bayesian Model

Sparse Bayesian model (Tipping 2001) is a generalized "linear" regression model in which Bayesian inference is used with a specific prior over the parameters to give a sparsity. Herein, the term "linear" means the model is a linear function of general basis or kernel functions.

When N observations are available, the inputs of the model, i.e. variables used to predict a target value in the model, are set up as $\mathbf{X} = \{\boldsymbol{x}_1\ \boldsymbol{x}_2 \ldots \boldsymbol{x}_N\}$ and the corresponding target vectors are $\mathbf{t}=\{t_1\ \ t_2\ \ldots\ t_N\}^{\mathrm{T}}$. The "model" $y(\mathbf{x})$ is a linear function of M adjustable parameters $\mathbf{w} = \{w_1, w_2, \ldots, w_M\}$, i.e.

$$y(\mathbf{x};\mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) \tag{2.1}$$

where $\phi_m(\mathbf{x})$ is a basis function of the model. The "prior" distribution, i.e. the distribution of the model parameter $\mathbf{w}$ before the objective information from the observations is incorporated during a Bayesian inference, is defined as follows:

$$p(\mathbf{w} \mid \alpha_1,\ldots,\alpha_M) = \prod_{m=1}^{M}\left[(2\pi)^{-1/2}\alpha_m^{1/2}\exp\left\{-\frac{1}{2}\alpha_m\omega_m^2\right\}\right] \tag{2.2}$$

where α_m is a hyperparameter representing the uncertainty in the estimated weight w_m, $m=1,\ldots,M$. These hyperparmeters control the inverse variance of each weight as shown in the equation. However the model $\boldsymbol{p}(\mathbf{w}|\boldsymbol{\alpha}_1,\ldots,\alpha_M)$ is still a Gaussian density function, which does not give a sparse model (Tipping 2004). So, hyperpriors are introduced over all α_m to find the basis terms whose weights α_m are peaked around zero.

When uniform scale priors are used for example, each hyperparameter controls the individual weight. Using such hyperpriors in the Bayesian inference, one identifies the weight terms whose posterior density concentrates at zero. The associated inputs can be considered irrelevant and can be removed from the linear model $y(\mathbf{x};\mathbf{w})$ to achieve a sparse model. Non-zero weighed vectors are called 'relevant' vectors, which indicate the locations in the parameter domain relevant enough to have basis functions. This is the main feature of the Sparse Bayes mechanism.

After a prior is defined as described above, the posterior distribution over hyperparameters α_m is obtained based on the observations $\mathbf{t}$. The posterior distribution over weights is given as:

$$p(\mathbf{w},\alpha,\sigma^2 \mid \mathbf{t}) = \frac{p(\mathbf{t} \mid \mathbf{w},\alpha,\sigma^2)\,p(\mathbf{w},\alpha,\sigma^2)}{p(\mathbf{t})} \tag{2.3}$$

To facilitate the calculation, the posterior distribution in (2.3) can be alternatively computed as follows (Tipping 2001).

$$p(\mathbf{w},\alpha,\sigma^2 \mid \mathbf{t}) = p(\mathbf{w} \mid \mathbf{t},\alpha,\sigma^2)\,p(\alpha,\sigma^2 \mid \mathbf{t}) \tag{2.4}$$

where $p(\mathbf{w} \mid t,\alpha,\sigma^2)$ is the 'weight posterior' distribution. This posterior distribution defined over weights is Gaussian and calculated as follows:

$$p(\mathbf{w} \mid \mathbf{t},\alpha,\sigma^2) = \frac{p(\mathbf{t} \mid \mathbf{w},\sigma^2)\,p(\mathbf{w} \mid \alpha)}{p(\mathbf{t} \mid \alpha,\sigma^2)} \tag{2.5}$$

In detail, the Gaussian posterior distribution in (2.5) is given as

$$p(\mathbf{w} \mid \mathbf{t}, \alpha, \sigma^2) = (2\pi)^{-(N+1)/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{w}-\mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{w}-\mu)\right\}$$
$$\text{where } \Sigma = \left(\sigma^{-2}\Phi^{\mathrm{T}}\Phi + \mathbf{A}\right)^{-1} \text{ and } \mu = \sigma^{-2}\Sigma\Phi^{\mathrm{T}}\mathbf{t} \tag{2.6}$$

where $\mathbf{A}$ is a diagonal matrix of hyperparameters, i.e. $\mathbf{A} = diag(\alpha_1, ..., \alpha_M)$ and Φ is a set of basis functions, i.e. $\Phi = [\Phi_{mn}] = [\phi_m(\mathbf{x}_n)]$.

Once the posterior distribution is constructed, one way to estimate the parameters is to find the values that maximize the posterior distribution in (2.5). This is termed "maximum *a posteriori*" (MAP) estimation and considered as "short cut" because one concerns only about where the posterior distribution becomes maximum and ignores the variability and actual distribution of the unknown coefficients.

In order to find the optimal level of fitting, the marginal likelihood is first calculated by integrating out the weight terms:

$$\begin{aligned} p(\mathbf{t} \mid \alpha, \sigma^2) &= \int p(\mathbf{t} \mid \mathbf{w}, \sigma^2) p(\mathbf{w} \mid \alpha) d\mathbf{w} \\ &= (2\pi)^{-N/2} \left|\sigma^2\mathbf{I} + \alpha^{-1}\Phi\Phi^{\mathrm{T}}\right|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\sigma^2\mathbf{I} + \alpha^{-1}\Phi\Phi^{\mathrm{T}}\right)^{-1}\mathbf{t}\right\} \end{aligned} \tag{2.7}$$

"Full Bayesian" approach aims to find α and σ^2 that maximize the marginal likelihood in (2.7), denoted by α_{MP} and σ^2_{MP}, respectively. When α_i is "infinite" or very large, the uncertainty in the coefficient is negligible, so the corresponding basis function ϕ_i is retained. If α_i does not reach infinity (or does not blow up), ϕ_i basis function can be deleted or α_i is re-estimated. The following re-estimation formula is derived from differentiating $\log p(\mathbf{t} \mid \alpha, \sigma^2)$ with respect to α and σ^2 (Tipping 2001):

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} \text{ and } (\sigma^2)^{new} = \frac{\|\mathbf{t} - \Phi\mu\|^2}{N - \sum_i \gamma_i}$$
$$\gamma_i = 1 - \alpha_i \Sigma_{ii} \tag{2.8}$$

where γ_i is a measure of well-determinedness of parameter $\mathbf{w}$ and its range is $\gamma_i \in [0,1]$ (Tipping 2001). This process is repeated until convergence. *SparseBayes* MATLAB® software package developed by Mike Tipping (Tipping 2009) can perform these complicated re-estimations and were used for the numerical examples in this study.

2.2 Summary of Learning Algorithm Procedure

The inference algorithm of the Sparse Bayes method described in Section 2.1 is illustrated by a flow chart in Figure 2.1.

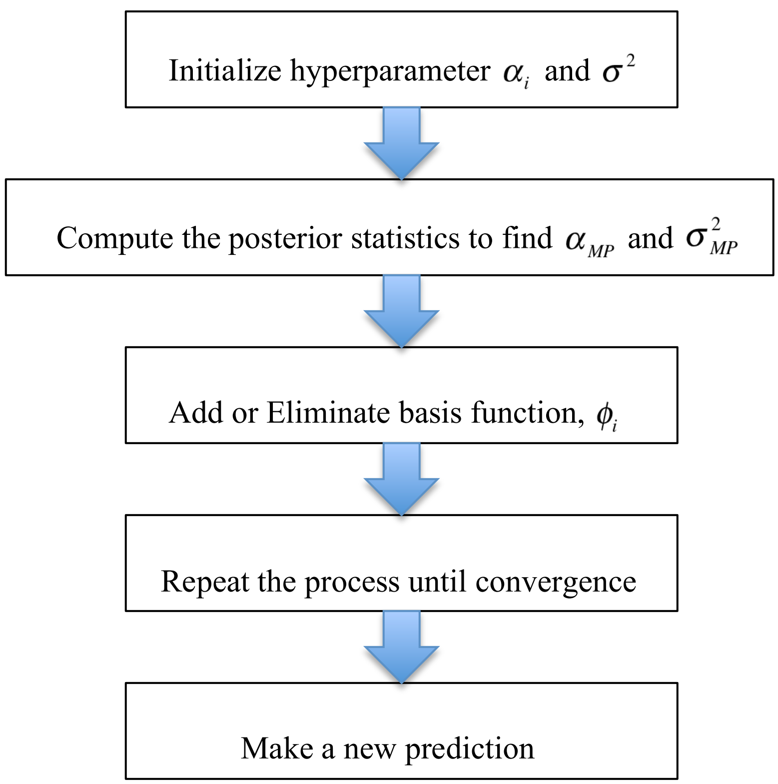

Figure 2.1. Learning Algorithm Procedure

The SparseBayesDemo MATLAB® code illustrates regression and classification models in one and two dimensions with user-chosen Gaussian or Bernoulli likelihood function models. Using a sparse subset of the possible basis functions, the data is derived from the generative model. At the MATLAB® prompt, user can choose to write one of two likelihoods functions, its dimension, and a noise. The example of Gaussian model with one dimension is written at the prompt such as ‘SparseBayesDemo(‘Gaussian’,1)’ and the graphical outcome is yielded as below:

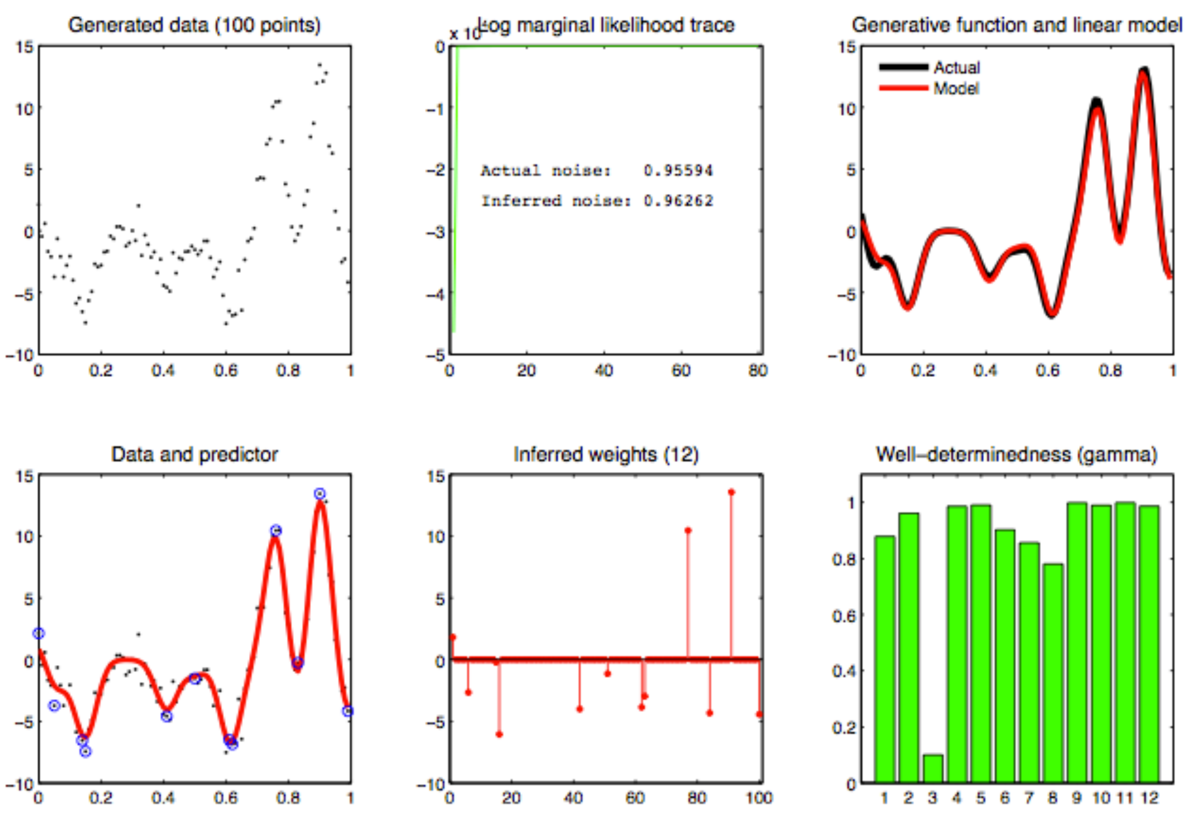

Figure 2.2. Graphical result of Gaussian model from *SparseBayesDemo* (Tipping 2009)

The result of Sparse Bayes gives six graphical outcomes which demonstrates the analysis made by Sparse Bayes MATLAB® code. Figure 2 is the example in which 100 data points are randomly generated. 'Generated data' graph shows all the data points and the relevant vectors are indicated on 'Data and Predictor' graph with circled dots. From 'Inferred weights' graph, one can see that 12 relevant vectors are chosen from Sparse Bayes algorithm, and the weight for each vector is shown as a straight line. The larger magnitude of the weight indicates higher correlation of the chosen vector with the predictive model. 'Well-determinedness' graph is the bar chart of the gamma value in (2.8) calculated for each relevant vector. The gamma value is in a range from zero to one and can be interpreted as a measure of how 'well-determined' its vector is by the data (MacKay 1992). If the parameter fits the data, the gamma value will be close to one. From the Sparse Bayes software by Mike Tipping, the analysis of model estimation became easy to visualize and understand the functionality for Sparse Bayesian models.

2.3 Advantage of Sparse Bayes Methodology

The Sparse Bayes methodology utilizes the marginalized likelihood, which allows us to make an optimal prediction on hyperparameters despite varying parameters and unknown validation data. While Bayesian inference itself requires a tedious integration procedure, Sparse Bayes methodology eventually sets many weights to zero and predict the model in an efficient way, i.e. produces a model with fewer number of relatively important basis functions ("relevant vectors"). Sparsity in the estimated function $y(\mathbf{x})$ reduces the complexity and increases the speed in calculations for predictions. Moreover, Sparse Bayesian methodology gives an insight regarding important parameters in the models and important regions in the parameter space, which will be further investigated in Chapter 4. In next chapter, three different inference methodologies are applied to a dataset for the purposes of test and comparison.

CHAPTER 3

TESTS AND APPLICATIONS OF SPARSE BAYES USING EXPERIMENTAL DATA

In this chapter, the performance of the Sparse Bayes method is tested and compared with those by traditional inference methods such as least-square method and penalized least-square method using an experimental database of the shear strengths of reinforced concrete beams.

3.1 Experimental Database of Shear Strengths of Reinforced Concrete Beams

In this chapter, the experimental database by Reineck *et al.* (2003) is used to test the Sparse Bayes method. Reineck *et al.* (2003) developed a large database for the shear strengths of reinforced concrete (RC) beams without shear reinforcement. With a broad scope of structural parameters available, shear strengths are observed from 439 RC specimens. From the original database by Reineck *et al*. (2003), some data points such as light-weight concrete and the tests in which maximum aggregate size is unavailable are removed (Song *et al*. 2010). As a result, 398 tests are used for this study.

Figure 3.1, adopted from Song *et al*. (2010), shows the distributions of parameters influencing the shear strength over the reduced database. Around 300 tests members have rectangular cross-sections while the others have T-shaped beams. It is also observed that 230 beams in the database have compressive strength of a concrete lower than 40 MPa, and most of the specimens have the effective depth between 200 to 300 mm. The figure gives a comprehensive understanding of the shear strengths distribution over a wide range of parameter values. Using this database, probabilistic shear strengths models are developed by three different methodologies: Sparse Bayes method, least-square method and penalized least-square method. First, these methods are briefly reviewed in the following sections.

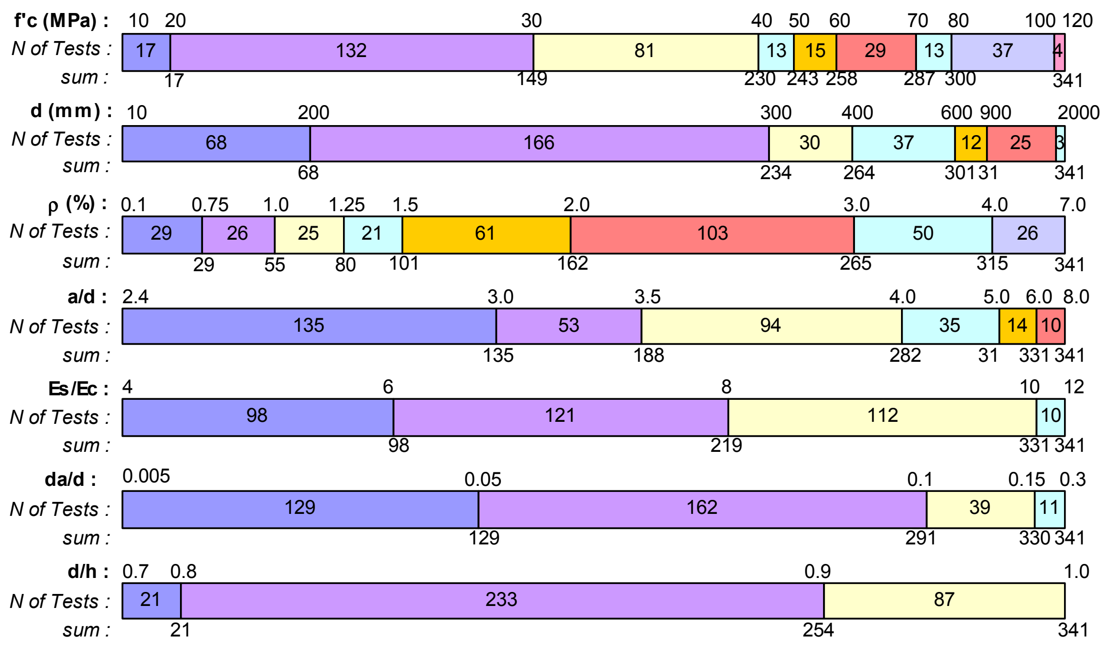

Figure 3.1. Distributions of parameter values of RC beams in shear database (Song *et al.*, 2010)

3.2 Least-Square Methodology

The Least-Square (LS) method is one of the most basic approaches in estimating model parameters based on given data. From a linear model $y(\mathbf{x};\mathbf{w})$ in (2.1), the least-square method finds the values of $\mathbf{w}$ that minimize the error measure

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left[t_n - \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}_n)\right]^2 \tag{3.1}$$

which is proportional to the sum of the squared prediction errors of the model. The weights giving the least square error are derived as

$$\mathbf{w}_{LS} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \tag{3.2}$$

The LS method is straightforward to use, but the method has a risk of "over-fitting" the model with the training data (Tipping 2004).

3.3 Penalized Least-Square Methodology

To reduce the risk of developing an over-fitted model, in the Penalized Least-Square (PLS) method, one adds a weight penalty term to the error measure. The error measure of the PLS method is thus

$$\hat{E}(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_w(\mathbf{w}) \tag{3.3}$$

where $E_D(\mathbf{w})$ is the error measure in (3.1) used for the LS method, λ is a regularization parameter introduced to achieve an optimal level of fitting, and $E_w(\mathbf{w})$ represents the scale of the weights, i.e.

$$E_w(\mathbf{w}) = \frac{1}{2}\sum_{m=1}^{M} w_m^2 \tag{3.4}$$

From (3.3) and (3.4), it is clear that the larger value of λ is used, the smaller weight values will be achieved during the minimization, which leads to a smooth and under-fitted model. The regularization parameter is often determined through validation using additional dataset that is not used for training purpose.

3.4 Comparison of Three Methodologies

Since the PLS method requires a separate dataset for determining the regularization parameter, i.e. validation dataset, the shear strength database is divided into three different sets; a training set (for model construction), a validation set (for determining λ), and a test dataset (for measuring the error). In this study, one eighth of the entire dataset of 398 shear strength data is used as the training set while another one eighth is used as a validation set. The remaining data in the set are used for testing purpose. Using a random number generator in MATLAB®, these three sets are arbitrarily determined. As a result, approximately fifty data are in a group of training and validation sets respectively, and around three hundred data are in a test set.

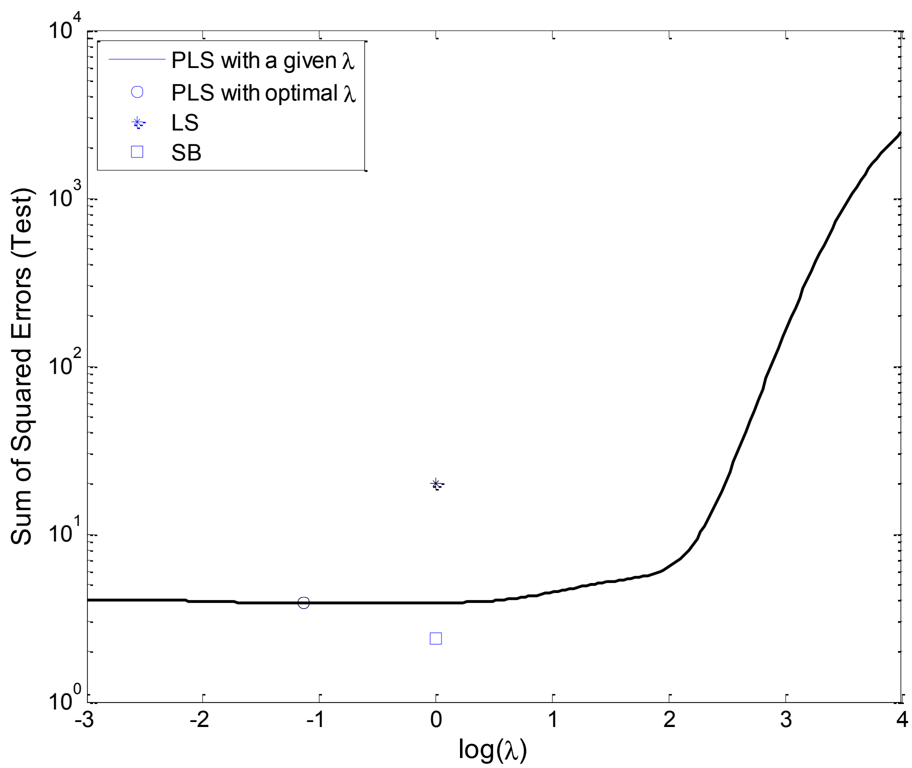

Figure 3.2. Comparison of three different methodologies

The markers in Figure 3.2 show the sums of the squared errors by LS, PLS and SB methods calculated for the test data set. Since PLS requires an additional validation data set while the other two methods not, for a fair comparison, LS and SB methods used both training and validation sets for the training purpose. The curve in the plot is the performance of the PLS model when the regularization parameter is arbitrarily given instead of being determined by minimization of the error in (3.3).

The LS shows the largest sum of squared errors for the test set because the model is over-fitted with the set used for training, and thus fails to represent the general shear strength mechanism shown in the bigger data set. The PLS model with the optimal λ value found from the validation data set performs significantly better than the LS model. This result indicates that the validation set helps reduce the risk of over-fitting. Finally, the plot shows that the Sparse Bayes methodology further reduces the error in the test set even though the same number of data points is used for the model construction and no efforts for validation were needed. The test, validation, and training sets are randomly generated for numerous times to check the reliability of the results. And the SB methodology consistently shows the best result during the process and similar trends are shown from the three methodologies with randomly selected sets. This result shows that the SB method has a good potential as an efficient and accurate statistical inference tool. Future research is needed to explore the full potential of the method in developing statistical models for various civil engineering applications.

CHAPTER 4

FURTHER DEVELOPMENT OF SPARSE BAYES METHODOLOGY

In this chapter, Sparse Bayesian methodology is applied to the database demonstrated in Chapter 2. Influential parameters are distinguished in determining the shear strengths of the reinforced beams and further analysis was made using the relevant vector concept.

4.1 Review of Existing Shear Strength Models

Song *et al*. (2010) used the shear strengths experimental database by Reineck *et al*. (2003) to develop probabilistic shear strength models using a Bayesian methodology (Gardoni *et al*. 2002). The size of the database was reduced to 398 to exclude the data in which the maximum aggregate size is not available. Song *et al*. (2010) discussed shear failure mechanisms in reinforced concrete beams to identify critical parameters influencing shear strengths of RC members without transverse reinforcement. To this end, existing shear strength models were also reviewed. Table 1, adopted from Song *et al.* (2010), summarizes existing shear strength models reviewed and further developed based on the database in their study. Most shear design codes use empirical models to predict shear strengths as seen in Table 4.1. Most of the shear strength models share six parameters $\{b_w, \rho, h, d, a/d, f_c'\}$, which are critical parameters that influence the shear strengths of reinforced concrete beams without stirrups.

Model	Equation
ACI 11-3 (ACI 2002)	$V_c = \frac{1}{6}\sqrt{f_c'}b_w d$
ACI 11-5 (ACI 2002)	$V_c = \left(0.158\sqrt{f_c'} + 17\rho\frac{V_u d}{M_u}\right)b_w d \le 0.3\sqrt{f_c'}b_w d$ where $V_u d / M_u \le 1.0$
Eurocode draft (2003)	$V_c = 0.12k(100\rho f_c')^{1/3} b_w d$ where $k = 1+\sqrt{200/d} \le 2.0$, $\rho \le 0.02$
Tureyen & Frosch (2003)	$V_c = \frac{5}{12}\sqrt{f_c'}b_w c$ where $c = kd$, $k = \sqrt{2\rho n + (\rho n)^2} - \rho n$ and $n = E_s / E_c$
Zsutty (1971)	$V_c = 2.2\left(f_c'\rho\frac{d}{a}\right)^{1/3} b_w d$
Okamura & Higai (1980)	$V_c = 0.2\frac{(100\rho)^{1/3}}{(d/1000)^{1/4}}(f_c')^{1/3}\left(0.75+\frac{1.40}{a/d}\right)^{1/3} b_w d$
Bazant and Yu (2005)	$V_c = 1.1044\cdot\rho^{3/8} b_w\left(1+\frac{d}{a}\right)\sqrt{\frac{f_c' d_0 d}{1+d_0/d}}$ where $d_0 = \kappa(f_c')^{-2/3}$ and $\kappa = 693.7623\sqrt{d_a}$
Russo et al. (2005)	$V_c = 1.13\,\xi\left[\rho^{0.4}(f_c')^{0.39} + 0.5\rho^{0.83} f_y^{0.89}\left(\frac{a}{d}\right)^{-1.2-0.45(a/d)}\right]b_w d$ where $\xi = \frac{1+\sqrt{5.08/d_a}}{\sqrt{1+d/(25d_a)}}$

Notations: V_c (N): shear strength; f_c' (MPa): concrete compressive strength; b_w (mm): web width; d (mm): effective depth; $\rho = A_s / b_w d$: longitudinal reinforcement ratio in which A_s is the amount (area) of longitudinal reinforcement; V_u (N): factored shear force; M_u (N·mm): factored moment; $E_s = 2.0\times10^5$ (MPa): elastic modulus of reinforcement; $E_c = 4700\sqrt{f_c'}$ (MPa): elastic modulus of concrete; a (mm): shear span length; d_a (mm): the maximum aggregate size; and f_y (MPa): the yielding strength of the longitudinal reinforcement.

Table 4.1. Existing shear strength models (Song e*t al*. 2010)

To further investigate the influences of these parameters, in this study, probabilistic shear strength models are developed using the Sparse Bayes methodology. The same reduced database by Reineck *et al* (2003) and Sparse Bayes MATLAB® code (Tipping 2009) are used for the model development. Various combinations of the six critical parameters are used as model parameters.

4.2 Parameters Influencing Shear Strength and their Application to Sparse Bayes

Sparse Bayes is used to estimate the number of relevant vectors, correlations and errors in calculating the main function of shear strength for nine combinations of the six critical model parameters. The natural logarithm is applied to each parameter value before the model development to achieve homoscedasticity of the model (Song *et al*. 2010). Table 4.2 shows nine possible combinations of the six influencing parameters considered in this study. Table 4.3 shows the results of the model construction using the Sparse Bayes method in terms of the number of relevant vectors, sum of the squared errors, maximum error, and covariance of the errors.

	$b_w \cdot d$	f_c'	ρ	a/d	E_s/E_c	d_d/d	d/h
Case 1	o	o	o	x	x	x	x
Case 2	o	o	o	o	x	x	x
Case 3	o	o	o	o	o	x	x
Case 4	o	o	o	o	x	o	x
Case 5	o	o	o	o	x	x	o
Case 6	o	o	o	o	o	o	x
Case 7	o	o	o	o	o	x	o
Case 8	o	o	o	o	x	o	o
Case 9	o	o	o	o	o	o	o

Table 4.2. Nine combinations of critical parameters

	Combinations of Parameters	# of Relevant Vectors	Sum of Squared Errors	Maximum Error	c.o.v. of Errors
Case 1	$b_w d,\ f_c',\ \rho$	88	11.872	0.402	0.196
Case 2	$b_w d,\ f_c',\ \rho,\ a/d$	102	12.983	0.642	0.209
Case 3	$b_w d,\ f_c',\ \rho,\ a/d,\ E_s/E_c$	96	12.983	0.458	0.207
Case 4	$b_w d,\ f_c',\ \rho,\ a/d,\ d_a/d$	113	16.728	1.083	0.270
Case 5	$b_w d,\ f_c',\ \rho,\ a/d,\ d/h$	101	9.289	0.269	0.176
Case 6	$b_w d,\ f_c',\ \rho,\ a/d,\ E_s/E_c,\ d_a/d$	118	13.977	0.670	0.246
Case 7	$b_w d,\ f_c',\ \rho,\ a/d,\ E_s/E_c,\ d/h$	92	10.823	0.271	0.206
Case 8	$b_w d,\ f_c',\ \rho,\ a/d,\ d/h,\ d_a/d$	114	11.528	0.623	0.225
Case 9	$b_w d,\ f_c',\ \rho,\ a/d,\ E_s/E_c,\ d_a/d,\ d/h$	107	12.918	0.575	0.235

Table 4.3. Results of development of probabilistic strength models by Sparse Bayes method

4.3 Analysis of Results

From the results in Table 4.3, one can identify which combinations of influencing parameters are the most reasonable choices by comparing the errors and the number of relevant vectors chosen from the Sparse Bayes method. In terms of the number of relevant vectors, Case 1, Case 3 and Case 7 are showing better results. Requiring fewer number of relevant vector means the model captures relatively important domains in the parameter space more effectively. Case 5 and Case 7 show superior results in terms of sum of the squared errors, which indicates that the models have smaller errors in average. These two cases demonstrate smaller maximum errors as well, which means the models do not have subdomain where the errors are relatively extreme. The coefficients of variation (c.o.v.) of errors are similar for all cases. Therefore, Case 5, i.e. strength models defined in terms of $\{b_w d, f_c', \rho, a/d, d/h\}$, stands out as the best combination of critical parameters.

Additionally, it is observed that the number of parameters used in the Sparse Bayes method does not have a strong correlation with the number of relevant vectors. The relevant vectors chosen out of 398 data points are in a range between eighty and a hundred. Too small number of relevant vectors can be risky and less reliable. Too many relevant vectors may not be helpful in identifying important subdomains using the Sparse Bayes methodology. Therefore, it is desirable to choose models that have reasonable numbers of relevant vectors in such an investigation. In the next section, the identified relevant vectors are further investigated to gain insight of the shear strengths of reinforced concrete beams.

4.4 Further Investigation of Identified Relevant Vectors

First, the relevant vectors identified by the Sparse Bayes method are grouped by use of the *k*-means clustering algorithm in order to discover the patterns and correlations of relatively important subdomains in the parameter space.

The *k*-means algorithm is one of the popular unsupervised learning algorithms (MacQueen 1967). K-means classify a given data set for a pre-determined number of clusters, *k*.

The algorithm finds clusters such that the centroids of the identified clusters are placed far away from each other. Figure 4.1 illustrates a typical k-means algorithm.

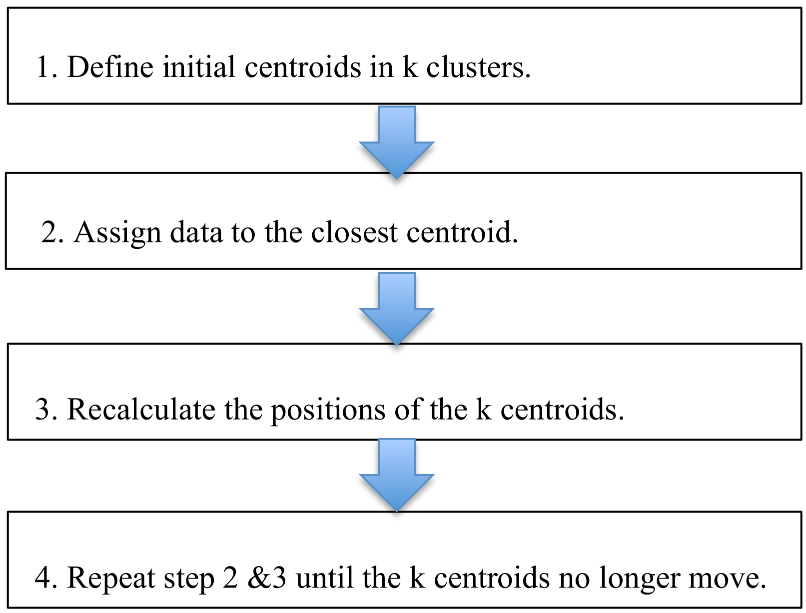

Figure 4.1. Example of *k*-clustering Algorithm

This algorithm aims at minimizing the following objective function, which is the sum of the squared distance between the center of the cluster and the data point in the cluster:

$$F = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left\| x_i^j - c_j \right\|^2 \tag{4.1}$$

where x_i^j is the *i*-th data point in the j-th cluster, $i=1,\ldots,n_j$, and $j = 1,\ldots,k$, and c_j is the centroid of the *j*-th cluster with n_j data points. The *k*-means algorithm is simple and easy, but one needs to choose the number of clusters before the clustering.

The *k*-means clustering is performed for 101 relevant vectors from Case 5. The relevant vectors identified by the Sparse Bayes using the combination of five parameters,

$b_w d$, f_c', ρ, a/d, d/h, i.e. Case 5 are used as an example. As a reasonable number of clusters, *k*, the study tries 4, 5 and 6. The coordinates at the identified centroids, and their mean, variance and coefficient of variation are summarized in the tables below. The order of the c.o.v.'s (from the highest to the lowest) is also shown to compare that in which parameter the clusters are the most well-separated.

k=4	$b_w d$	f_c'	ρ	a/d	d/h
Cluster 1	0.579	0.885	0.560	0.650	0.428
Cluster 2	0.627	0.885	0.117	0.568	0.364
Cluster 3	0.776	0.875	0.430	0.592	0.294
Cluster 4	0.507	0.966	0.854	0.607	0.895
Mean	0.622	0.903	0.490	0.604	0.495
Variance	0.114	0.0427	0.306	0.034	0.272
c.o.v.	0.183	0.0473	0.623	0.057	0.549
c.o.v. order (high to low)	3	5	1	4	2

Table 4.4. Results of *k*-means clustering of relevant vectors (*k*=4)

k=5	$b_w d$	f_c'	ρ	a/d	d/h
Cluster 1	0.842	0.853	0.874	0.567	0.265
Cluster 2	0.578	0.883	0.583	0.644	0.429
Cluster 3	0.507	0.966	0.854	0.607	0.895
Cluster 4	0.756	0.892	0.246	0.584	0.301
Cluster 5	0.603	0.881	0.170	0.590	0.381
Mean	0.657	0.895	0.545	0.598	0.454
Variance	0.137	0.0425	0.330	0.029	0.255
c.o.v.	0.209	0.0475	0.605	0.048	0.561
c.o.v. order (high to low)	3	5	1	4	2

Table 4.5. Results of *k*-means clustering of relevant vectors (*k*=5)

k=6	$b_w d$	f_c'	ρ	a / d	d / h
Cluster 1	0.756	0.892	0.246	0.584	0.301
Cluster 2	0.638	0.864	0.436	0.787	0.371
Cluster 3	0.558	0.891	0.617	0.582	0.443
Cluster 4	0.507	0.966	0.854	0.607	0.895
Cluster 5	0.842	0.853	0.874	0.567	0.265
Cluster 6	0.595	0.883	0.141	0.565	0.390
Mean	0.649	0.891	0.528	0.615	0.444
Variance	0.127	0.0399	0.307	0.084	0.230
c.o.v.	0.195	0.0448	0.582	0.137	0.517
c.o.v. order (high to low)	3	5	1	4	2

Table 4.6. Results of *k*-means clustering of relevant vectors (*k*=6)

It is observed that the c.o.v.'s of the centroid coordinates are not sensitive to the chosen number of clusters, and thus for the tested range, i.e. 4, 5 and 6, the optimal number of clusters needs to be determined based on the dissimilarities of the cases represented by the centroids in terms of their shear failure mechanisms. It is also noted that the order of the c.o.v.'s between the parameters is consistent while ρ, the longitudinal reinforcement ratio stays the highest. This implies that the reinforcement ratio is the most influential parameter that separates the clusters and thus considered as a relatively important parameter to be considered during a parametric study.

4.5 Observations

With chosen set of combinations of five parameters $\{b_w d, f_c', \rho, a / d, d / h\}$ from previous section, two parameters are selectively chosen and drawn in a scattered graph. Each color represents a different cluster, and the centroid of each cluster is marked as an asterisk. Below is a graph of the reinforcement ratio, ρ in y-axis with four other parameters in x-axis.

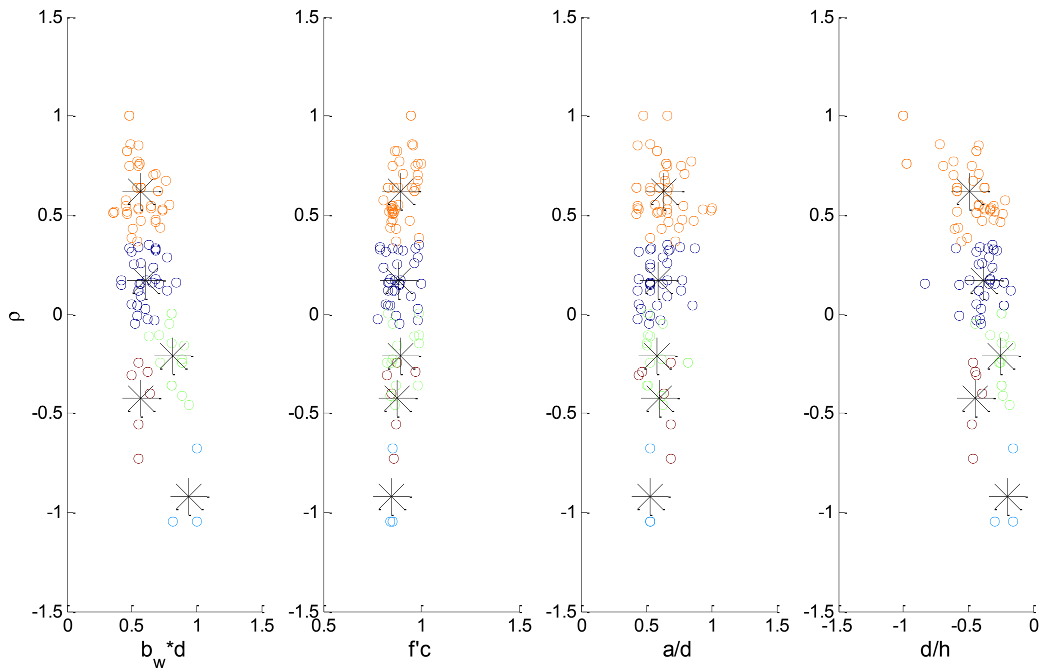

Figure 4.2. Distribution of clusters along ρ and four other parameters (natural logarithms applied to all parameters)

From Figure 4.2, it is distinguishable that five clusters are clearly distributed along the axis of ρ. This graphical result proves that the reinforcement ratio, ρ is an influential parameter that separates the clusters well. The second highest coefficient of variance during a parametric study in Section 4.4 is d/h. Therefore a graph of d/h in y-axis with three parameters, $b_w d$, f_c', a/d, in x-axis is shown in Figure 4.3. Unlike Figure 4.2, Figure 4.3 does not show a good separation between clusters along the axis of d/h.

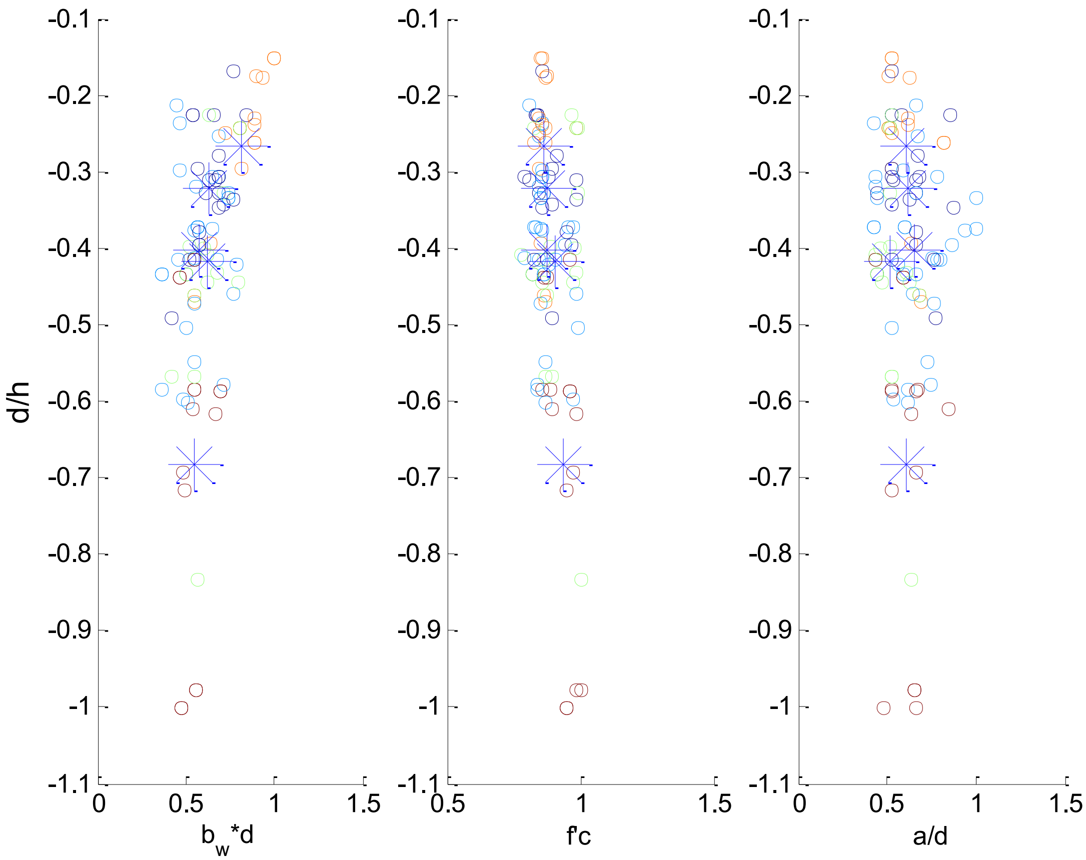

Figure 4.3. Distribution of clusters along d / h and three other parameters (natural logarithm applied to all parameters)

CHAPTER 5

CONCLUSIONS

This study investigated the Sparse Bayes methodology by Mike Tipping and tested the method as a statistical inference tool for civil engineering data. The method was further developed to identify important parameters and subdomains in the parameter space.

There are many parameters that contribute to the shear strengths of reinforced concrete beams. Among those variables, the combination of $\{b_w d, f_c', \rho, a/d, d/h\}$ was identified as the most influential set that can predict the shear strength model with Sparse Bayes methodology. From extensive amount of given data, only one fourth was chosen to develop the model without additional validation dataset. The model was able to predict the shear strength model accurately for the remaining test data set. This is highly effective way of predicting a phenomenon and the Sparse Bayes methodology was proved to be a useful tool for developing statistical model for structural engineering problems.

Furthermore, the further development using the *k*-means algorithm revealed that the most influential parameter in the shear strength model of reinforced concrete beam is the reinforcement ratio, which gives out the highest correlation of coefficient when *k*-means clustering algorithm was performed, and well separated the identified clusters.

This study shows the applicability of using the Sparse Bayes method to structural engineering problems and gives an insight to aid in decision-making process for future engineering research and projects.

REFERENCES

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis*. Springer, second edition. New York, NY.

Gardoni, P., Der Kiureghian, A. and Mosalam, K.M. 2002. Probabilistic capacity models and fragility estimates for reinforced concrete columns based on experimental observations. *J. Eng. Mech.* ASCE.128(10), 1024-1038.

MacKay, D.J.C. 1992. Bayesian interpolation. *Neural Computation*. 4(3):415 447.

MacQueen, J.B. 1967. Some Methods for classification and Analysis of Multivariate Observations: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press.

Neal, R.M. 1996. *Bayesian Learning for Neural Networks*. Springer, New York, NY.

Reineck, K.H., Kuchma, D.A., Kim, K.S. and Marx, S. 2003. Shear database for reinforced concrete members without shear reinforcement. *ACI Struct. J.,* 100(2), 240-249.

Song, J., Kang, W.H., Kim, K.S. and Jung S. 2010. Probabilistic shear strength models for reinforced concrete beams without shear reinforcement. *Structural Engineering and Mechanics*. Vol. 34, No 1, 15-38.

Tipping, M.E. 2001. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* 1:211-244.

Tipping, M.E. 2004. Bayesian Inference: An Introduction to principles and practice in machine learning. *Microsoft Research.*

Tipping, M.E. 2009. An Efficient Matlab Implementation of the Sparse Bayesian Modelling Algorithm, Version 2.0. *Vector Anomality.* www.vectoranomaly.com