

**DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES**

**THE USE OF SPLINE, BAYESIAN SPLINE AND
PENALIZED BAYESIAN SPLINE REGRESSION
FOR MODELING**

**by
Mahmut Sami ERDOĞAN**

**July, 2013
İZMİR**

**MODELLEME İÇİN SPLAYN, BAYESYEN
SPLAYN VE CEZALANDIRILMIŞ BAYESYEN
SPLAYN REGRESYON KULLANIMI**

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for
the Degree of Master of Science in Statistics**

**by
Mahmut Sami ERDOĞAN**

**July, 2013
İZMİR**

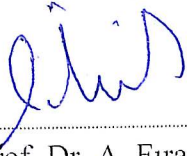
M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “THE USE OF SPLINE, BAYESIAN SPLINE AND PENALIZED BAYESIAN SPLINE REGRESSION FOR MODELING” completed by MAHMUT SAMİ ERDOĞAN under supervision of ASSOC. PROF. DR. ÖZLEM EGE ORUÇ and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



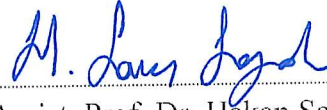
Assoc. Prof. Dr. Özlem EGE ORUÇ

Supervisor



Assist. Prof. Dr. A. Fırat ÖZDEMİR

(Jury Member)



Assist. Prof. Dr. Hakan Savaş SAZAK

(Jury Member)



Prof. Dr. Ayşe OKUR

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

Foremost, I wish to express my sincere gratitude to my advisor Assoc. Prof. Özlem EGE ORUÇ for the continuous support of my thesis study and research for her patience, motivation, enthusiasm and immense knowledge.

Also, I would like to thank to Asst. Prof. Neslihan DEMİREL and Asst. Prof. Zeynep Filiz EREN DOĞU for their positive comments and contributions, and my colleagues for their devoted support.

I would also like to express my deepest gratitude to my beloved family who have never given up their support spiritually throughout my life.

Mahmut Sami ERDOĞAN

THE USE OF SPLINE, BAYESIAN SPLINE AND PENALIZED BAYESIAN SPLINE REGRESSION FOR MODELING

ABSTRACT

The nonparametric regression methods which are called spline, penalized spline and Bayesian spline bring great advantages such as not depending on the fixed model and flexibility in modeling. In particular, penalized spline regression uses the idea of nonparametric spline smoothing and it is in fact just a generalization of smoothing splines that should allow more flexibility in a choice of the spline model, the basis functions, and the penalty. In this study, distribution graph of ratios of export to import in Turkey is modeled using the nonparametric regression methods that are spline and Bayesian spline regression. For both methods, the knot sequence coincides with the end points of the interval. The results of these regression models are compared and interpreted. Then, we focus on a penalized spline regression with Bayesian perspective on the same data set and the smoothing for a variety of lambda values is performed. In addition, the contribution of a prior distribution is explained to determine the smoothing parameter. Then, we propose a new smoothing parameter by using the amount of information contained in the normal distribution. It has been observed that this parameter is very sensitive against small changes. This result denotes that the proposed smoothing parameter we obtained is appropriate for using in the penalized Bayesian spline regression applications.

Keywords: Spline function, bayesian spline regression, penalized bayesian spline regression, mcmc, smoothing parameter.

MODELLEME İÇİN SPLAYN, BAYESYEN SPLAYN VE CEZALANDIRILMIŞ BAYESYEN SPLAYN REGRESYON KULLANIMI

ÖZ

Splayn, cezalandırılmış splayn ve Bayesyen splayn olarak adlandırılan parametrik olmayan regresyon yöntemleri modellemede esneklik ve sabit bir modele bağlı olmamak gibi büyük avantajlar sağlar. Cezalandırılmış splayn regresyon, parametrik olmayan splayn düzeltme düşüncesini kullanır. Bu regresyon aslında splayn düzeltme genelleştirilmesidir ve splayn modelin, temel fonksiyonlarının ve cezanın seçiminde daha fazla esnekliğe izin verir. Bu çalışmada, Türkiye’de ihracatın ithalatı karşılama oranlarının dağılım grafiği parametrik olmayan regresyon yöntemleri; splayn ve Bayesyen splayn regresyon kullanılarak modellenmiştir. Her iki yöntem için, düğüm noktaları aralıkların uç noktaları ile aynı alınmıştır. Bu regresyon modellerinin sonuçları karşılaştırılmış ve yorumlanmıştır. Daha sonra aynı veri seti üzerinde Bayesyen perspektif ile cezalandırılmış splayn regresyona uygulanmış ve çeşitli lambda değerleri için düzeltme gerçekleştirilmiştir. Ek olarak, düzeltme parametresini belirlemede önsel dağılımın katkısı açıklanmıştır. Ayrıca, normal dağılımın bilgi içeriği miktarını kullanarak yeni bir düzeltme parametresi önerilmiştir. Bu parametrenin küçük değişiklikler karşısında çok hassas olduğu gözlemlenmiştir. Bu sonuç; önerilen düzeltme parametresinin cezalandırılmış Bayesyen splayn regresyon uygulamalarında kullanılmak için uygun olduğunu göstermiştir.

Anahtar Kelimeler: Spline fonksiyonu, bayesyen splayn regresyon, cezalandırılmış bayesyen splayn regresyon, mcmc, düzeltme parametresi.

CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGMENTS	iii
ABSTRACT.....	iv
ÖZ	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
CHAPTER ONE-INTRODUCTION	1
CHAPTER TWO-BAYESIAN APPROACH.....	4
2.1 Basic Concepts of the Bayesian Approach.....	5
2.2 Bayesian Inference	6
2.3 Prior Distributions and Their Selection.....	7
2.3.1 Noninformative Priors	9
2.3.2 Informative Priors	10
2.3.3 Conjugate Priors	11
2.3.4 Some Basic Bayesian Models.....	12
CHAPTER THREE-BAYESIAN COMPUTATION	17
3.1 Bayesian Central Limit Theorem	17
3.2 Markov Chain Monte Carlo Method	18
3.2.1 Gibbs Sampling	19
3.2.2 Metropolis Hasting Algorithm.....	22
CHAPTER FOUR-BAYESIAN AND SPLINE REGRESSION	24
4.1 Bayesian Regression.....	24

4.1.1 Bayesian Regression Model	26
4.2 Spline Regression	28
4.2.1 Penalized Spline Regression.....	30
CHAPTER FIVE-APPLICATION	33
CHAPTER SIX-CONCLUSION	39
REFERENCES	41
APPENDIX- THE CODES OF PROGRAMS.....	44

LIST OF FIGURES

	Page
Figure 4.1 A least-squares spline fit to ratios of export to import data using the manually-selected knots. The knots used were 18, 50, 59.	29
Figure 4.2 A least-squares spline fit to ratios of export to import data using the manually-selected knots. The knots used were 6, 12, 18, 30, 36, 50, 54, 59.	30
Figure 5.1 Plots for regression assumptions.	34

LIST OF TABLES

	Page
Table 2.1 Differences between Frequentist and Bayesian approach.....	5
Table 2.2 Jeffreys priors.....	10
Table 2.3 Conjugate priors	11
Table 5.1 The results of Spline Regression.....	33
Table 5.2 The results of Bayesian Spline Regression.....	35
Table 5.3 The values of AIC and BIC for spline regression and Bayesian spline regression	36
Table 5.4 Penalized Bayesian Spline Regression Model for different λ	37
Table 5.5 Penalized Bayesian Spline Regression Model for different λ^*	37

CHAPTER ONE

INTRODUCTION

There are basically two different philosophical approaches in the science of statistics. The classical (Frequentist) approach and Bayesian approach. The classical approach shows parallelism with the deductive method, while Bayesian approach shows parallelism with the inductive method. These approaches constructs alternatives to each other to explicating of axioms in the science of statistics and examining many topics and concepts.

The basis of Bayesian methods is based on Bayes Theorem. This theorem proposed on the purpose of calculation of posterior probabilities using the prior probabilities by the British mathematician Thomas Bayes in the 18th century. Bayesian methods were not used too much in the past years due to difficulty of its theory and implementation. However, with the improving technology in the recent years, an important step was taken with respect to calculations. In this way, many of the statistical concepts are interpreted differently and handled with Bayesian approach.

Bayesian methods have an important place in statistical inference. The main difference between the Classical approach and Bayesian approach is the parameter forms of thinking on inference. Parameter is considered as a random variable which has a probability distribution in Bayesian approach. Accordingly, prior distribution is determined for unknown parameter and posterior distribution of parameter is obtained by it combined with existing data. Briefly, all the inference procedures related to parameter are made based on posterior distribution in Bayesian analysis. In the classical approach, parameter is seen as a fixed unknown. Parameter estimation is calculated only on the basis of the existing data. Hence due to fact that the parameter itself is not a result of repetition of the real experiments, existence of probability distribution is unthinkable.

Bayesian methods, nowadays, used to many application areas such as finance, biostatistics, econometrics. Some of the scientists engaged in studies using these methods in recent years are as follows. Geweke used these methods in the field of econometrics in his study in 1999. Carlin and Louis applied these methods to empirical Bayes in 2000. O'Hagan (1995), Berger and Pericchi (1996) and Berger (1998) benefited from Bayesian methods about model selection. Gersch and Kitagawa (1995) with West and Harrison (1997) used these methods in time series analysis. Dey and Sinha (1993) studied using these methods about reliability and survival analysis.

In this study, Bayesian methods which used in spline regression analysis which developed based on regression analysis will be examined. Bayesian regression spline and spline regression results will be interpreted by comparing with an application made on a real data set. The basic concepts of Bayesian approach in detail described in the second section which follows the back of introduction section in study. In the third section, Markov Chain Monte Carlo which is a kind of convergence method used to obtain posterior distributions examined and Monte Carlo Integration, Gibbs sampling, Metropolis and Metropolis Hastings algorithms described one by one. In the fourth section, Bayesian Regression is discussed in summary and its theoretical background is described. Also Spline regression, Penalized Spline Regression and smoothing parameter term are examined within the outline. Finally, a new definition for the smoothing parameter is done in Bayesian framework.

In fifth section, which is a part of application, first, spline regression analysis applied to data of export/import rate obtained from Turkish Statistical Institute for the state in which the numbers and positions of the knots are known. Bayesian spline regression has been applied using WinBUGS programming for the same data and it has been interpreted by comparing these results. Then Penalized Spline Regression is examined with Bayesian approach and models are established for the different values of the smoothing parameter which obtained using prior distributions. The founded model for large λ value is gradually shown to approaches to simple regression.

Lastly, the performance of new smoothing parameter is investigated and is made as a proposal performing the application.

CHAPTER TWO

BAYESIAN APPROACH

When the historical development of statistical inference is examined, three main approaches are encountered. These are the Bayesian approach, the classical approach and the likelihood based approach. The Bayesian approach, which has been developed from the Bayes Theorem, introduced into the literature by Thomas Bayes in 1761, is known to be influential on the statistical inference methods from the end of the 18th century and until the mid-20th century. The classical approach was offered by Laplace (1764) at the same period, and later it was developed and introduced into the literature by Neyman and Pearson. After these approaches, the likelihood based approach developed by Fisher has brought a new dimension to statistical inference. Scholars who adopted the classical and likelihood based approaches exhibited a critical attitude towards Bayesian approach, since they thought that it was an objective method. Due to this attitude, the difficulty of its theory and its implementation, Bayesian approach could not be used in statistical inference for many years. However, there have been significant improvements in calculations due to the developing technology in recent years. Thus, many statistical concepts are being reconsidered and interpreted from a different perspective using the Bayesian approach.

In other approaches, which were developed independently of the Bayesian approach, the concepts and methods defined for inference are totally different. In Bayesian methods, inferences are made depending on the present, priori knowledge. This dependence on subjectivity is one of the most prominent criticisms to Bayesian methods. Proponents of the classical approach criticize the priori knowledge due to departure from objectivity. Proponents of the Bayesian approach, on the other hand, argue that some knowledge from the past could not be neglected and the objective information obtained from the data and the priori knowledge should be incorporated; and also they think that too much hypotheses in the classical approach would yield to deceptive and misleading results. According to Bayesian statisticians, lack of flexibility in the classical approach is another negative treat.

The major difference between the classical approach and the Bayesian approach is the way they think of the parameter while performing inferences. In the Bayesian approach the parameter is thought as a stochastic variable with a probability distribution. In this respect, for the predictor of the parameter, prior probability distribution is determined. It is combined with the present data and the posterior probability distribution of the parameter predictor is obtained. To summarize, all inference operations related to the parameter is done based on the posterior distribution in the Bayesian approach. In the classical approach, on the other hand, the parameter is considered as a constant. Parameter prediction is only calculated based on the present data at hand. Therefore, since the parameter itself is not the result of the repeated actual trials, it cannot be argued that there is a probability distribution. The differences between the Bayesian and the classical approaches are presented as a table below:

Table 2.1 Differences between Frequentist and Bayesian approach

Concept	Bayesian	Frequentist
θ	Random	Fixed but unknown
$\hat{\theta}$	Fixed	Random
Randomness	Subjective	Sampling
Distribution of interest	Posterior	Sampling Distribution

2.1 Basic Concepts of the Bayesian Approach

In this chapter, the Bayes theorem, which is the basis of the Bayesian approach, and the basic concepts used in the Bayesian analysis will be introduced.

Bayes Theorem: Let us consider the B_1, B_2, \dots, B_k discrete events set, whose combination gives the sample space, S . If A is an event defined in the sample space, S , then

$$P(B_i|A) = \frac{P(A \cap B_i)}{\sum_{i=1}^k P(A \cap B_i)} = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^k P(B_i)P(A|B_i)}, i = 1, \dots, k \quad (2.1)$$

It can be shown $\sum_{i=1}^k P(B_i|A) = 1$. $P(B_i)$, $i = 1, \dots, k$ probabilities are called the prior probabilities. $P(B_i|A)$, $i = 1, \dots, k$ probabilities are called the posterior probabilities. These are the probabilities after knowing the results of the experiment.

2.2 Bayesian Inference

In Bayesian approach, where the probability distribution of the parameter to be predicted is considered as a random variable; let us consider that θ indicates the unknown parameter vector and y indicates the observed value. We can write the equation below, with reference to the Bayes theorem:

$$P(\theta|y) = \frac{P(\theta, y)}{P(y)} = \frac{f(y|\theta)P(\theta)}{P(y)} \quad (2.2)$$

$f(\theta|y)$ in Equation (2.2) expresses the likelihood function, $P(\theta)$ expresses the prior distribution. As for $P(y)$, it expresses the marginal likelihood and is represented as below:

$$P(y) = \int f(y|\theta)P(\theta)d\theta \quad (2.3)$$

$P(y)$, which is the marginal likelihood, is called as the normalizing constant in the literature, and at the same time it ensures that the integral of the posterior distribution result equal to 1. Also, in order to obtain Bayesian inferences, some integrals should be solved to obtain the posterior distribution as it is seen in the formula in Equation 2.3.

As there is no any expression about θ in the normalizing constant, this coefficient is a constant independent of the θ parameter. As the distribution of θ is to be obtained in Equation 2.2, and since normalizing constant is constant value

independent of θ , the equation in 2.2 can be rewritten as below, using a proportionality expression:

$$P(\theta|y) \propto f(y|\theta)P(\theta) \quad (2.4)$$

Equation (2.4) can be explained as the product of prior distribution and the likelihood function is proportional to the posterior distribution (Carlin & Louis, 2009). All inferences and calculations about the θ parameter are done using the $P(\theta|y)$ posterior distribution. In Bayesian statistics, a posterior distribution obtained from an analysis, can be used as a prior distribution for the next analysis.

Prior distribution has some parameters. These parameters are called as hyperparameter. The parameter values of a hyperparameter can be undetermined or unknown. In that case, a distribution is assigned to the hyperparameter and this distribution is included in the analysis. The distribution of the hyperparameter is called as the hyperprior. The Bayesian models, in which the hyperprior distribution is used, are called the hierarchical Bayesian models. The basic representation of Bayesian hierarchical models is as follows:

$$P(\theta, \beta|y) \propto f(y|\theta)P(\theta|\beta)P(\beta) \quad (2.5)$$

β in Equation 2.5 represents the hyperparameter. When the equation is examined, it is seen that the hyperparameter is not included in the probability function. The reason for this is that β does not influence the observed values directly, but does via θ . Therefore, β hyperparameter is not generally included in the probability functions belonging to hierarchical models (Ntzoufras, 2009).

2.3 Prior Distributions and Their Selection

In Bayesian approaches, a distribution is determined for the θ parameter, depending on the prior knowledge. These knowledge can be the personal beliefs of the researcher or expert opinions, or they can be obtained from previous studies. The

researcher reflects the prior knowledge into the analysis and combines it with the data and obtains the posterior distribution.

As it is emphasized in the previous sections, the major objection to Bayesian approaches is the prior distributions since they disrupt objectivity. Selection of a prior distribution appropriate for the study to be conducted may cancel all these objections. Different posterior distributions can be obtained by using different prior distributions for the same data. Thus, the selection of the prior distribution is one of the most important issues in Bayesian approach. A misspecified prior distribution may have a negative effect on inference (Beaumont & Rannala, 2004). Therefore, it would be useful to select a prior distribution after examining the previously used prior distribution for the research subject at hand.

The size of the data is another important issue in Bayesian approaches. The increase in the number of data may decrease the dominance of the prior distribution in obtaining the posterior distribution, and the likelihood function may become dominant. In that case, results similar to the ones in the classical approaches could be drawn. The distribution pattern of the likelihood function is another issue to be considered. If the likelihood function is sharp and the prior distribution is more oblate, than the contribution of the prior distribution to the posterior distribution would not be much (Box & Tiao, 1973).

In previous years, the applicability of the Bayes theorem was an issue to be considered while choosing the prior distribution. For instance, since the cases where the size of θ increased caused integrals that were impossible to solve, they prevented the posterior distribution to be obtained. In that case, the prior distribution which was necessary to obtain the posterior distribution was used. However, in recent years, these kinds of hinders have come to an end due to the methods developed under the name of Markov Chain Monte Carlo. Now integrals that are impossible to solve analytically can be easily solved with these methods. Thus the limitation of the prior distribution to easily obtain the posterior distribution disappeared.

In Bayesian models, it may not always be possible to determine the posterior distribution with various calculation methods after determining the prior distribution. The researcher, in such cases, may prefer a prior distribution which would enable obtaining the posterior distribution more easily. Sometimes, prior knowledge could not be trusted and a data-driven analysis may be required, or sometimes the researcher wants to include the powerful knowledge at hand into the analysis. For these reasons, prior distributions are categorized among themselves. There are three different distribution types in the literature. These are conjugate priors, informative priors and noninformative priors.

2.3.1 Noninformative Priors

If there is no any information about the θ parameter to be predicted, or the information at hand is not trusted, or the posterior distribution is wanted to be obtained as a result of a data-driven inference, the prior distribution to be used is called the noninformative prior distribution.

With the use of these priors, the influence of the prior distribution on the posterior distribution is at minimum. The results obtained by using the noninformative priors are expected to be similar to the results obtained with the classical approach. The reason for this is that in this approach the inference is made with only the information obtained from the data. The most widely used noninformative priors in the literature are the uniform prior and Jeffreys prior.

Uniform Prior: The uniform prior can be listed among the most widely used noninformative priors. Bayes and Laplace argued that when nothing is known about the θ parameter, $P(\theta)$ prior distribution should be uniformly distributed and all the possible results of θ should be the same. This is also known as the principle of insufficient reason (Syversveen, 1998).

Jeffreys Prior: Jeffreys offered this noninformative prior distribution, called after him, in 1961. The Fisher Information Matrix is used while obtaining this prior distribution. Jeffreys prior distribution can also be given an example to the improper

prior distributions; because it is not a probability function or a probability density function. However, the posterior distributions obtained by using these priors are probability functions or probability density functions. The table below presents some Jeffreys priors.

Table 2.2 Jeffreys priors

Likelihood	Parameters	Priors
Normal (When σ^2 is known)	$\theta = \mu$	1
Normal (When θ is known)	$\theta = \sigma$	σ^{-1}
Normal	$\theta = \mu, \sigma$	σ^{-1}
Bernoulli	$\theta = p$	$(pq)^{-1/2}$
Normal	$\theta = \lambda$	$(\lambda)^{-1/2}$

Another definition for prior distributions is made according to being proper or improper. If the determined prior distribution is not a probability function or a probability density function, this prior distribution is called as the improper prior distribution. There is no requirement for a proper prior distribution to obtain the posterior distribution. Priors (improper) which are not probability functions or probability density functions can be used in the analyses. However, even though these priors are used, it is a requirement for the posterior distribution to result as a probability function or a probability density function. Use of improper priors may result in improper posterior distributions. Therefore, they should be used carefully.

2.3.2 Informative Priors

Informative priors enable the researchers to incorporate their prior knowledge into the analysis. Information obtained from previous studies can be given as an example.

However, even though there is some information, it may be difficult to express these with a distribution. Also, use of informative priors, contrary to noninformative priors, seriously influences the posterior distribution. Therefore, one should be extremely sensitive in selecting the informative priors.

2.3.3 Conjugate Priors

If the prior distribution and the posterior distribution determined for the θ parameter is the same, these are called the conjugate priors. Analyses in which the posterior distribution is normal when the prior distribution is normal too, or the posterior distribution is inverse gamma when the prior distribution is inverse gamma, can be given example to conjugate priors. Conjugate priors are useful since they enable obtaining the posterior distributions in closed form. However, conjugate priors should be used carefully; because these priors show very specific prior knowledge. Some conjugate prior distributions are given in the table below:

Table 2.3 Conjugate priors

Likelihood	Prior Distribution	Posterior Distribution
Normal (When σ^2 is known)	Normal	Normal
Normal (When θ is known)	Inverse Gamma	Inverse Gamma
Poisson	Gamma	Gamma
Exponential	Gamma	Gamma
Uniform	Pareto	Pareto
Bernoulli	Beta	Beta
Binomial	Beta	Beta

2.3.4 Some Basic Bayesian Models

In terms of being an example for obtaining the posterior distribution, some basic Bayesian models for situations with a given likelihood function and a prior distribution, are discussed below.

Normal & Normal Model: The Bayesian model has two basic steps. These are the specification of the $Y|\theta \sim f(y|\theta)$ likelihood function and the $\theta \sim P(\theta)$ prior distribution. The simplest Bayesian analysis is the one in which the prior distribution is known (Carlin & Louis, 2009).

The situation, $(N(\theta, \sigma^2))$, in which the data is normally distributed with the θ average and σ^2 variance, will be examined. When the distribution of θ , in case σ^2 is known, is to be obtained, there is no need to assign the prior distribution to variance, since the prior distribution is only assigned to the unknown. Accordingly, the likelihood function can be written as below:

$$f(y|\theta) = \prod_{i=1}^n f(y_i|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum(y-\theta)^2}{2\sigma^2}} \quad (2.6)$$

Here, let us assume that the prior distribution specified for θ is normal, too. Let θ distribute normally with μ and τ^2 hyperparameters $(N(\mu, \tau^2))$. Since μ and τ^2 are the parameters of the prior distribution, they show the hyperparameters and it is assumed that they are known. The form of the prior distribution for θ can be expressed as below:

$$p(\theta) = \left(\frac{1}{\sqrt{2\pi}\tau}\right) e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \quad (2.7)$$

Using these information, the posterior distribution can be obtained as in Equation 2.8:

$$P(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{m(y)} = \frac{\left(\frac{1}{\sqrt{2\pi}\tau}\right) e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum(y-\theta)^2}{2\sigma^2}}}{\int \left(\frac{1}{\sqrt{2\pi}\tau}\right) e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum(y-\theta)^2}{2\sigma^2}} d\theta} \quad (2.8)$$

As it was mentioned in the previous section, since we are interested only in the distribution of θ , the expression which do not include θ can be removed and a proportional expression can be used. In that case, the equation below is obtained:

$$P(\theta|y) \propto e^{-\frac{(\theta-\mu)^2}{2\tau^2}} e^{-\frac{\sum(y-\theta)^2}{2\sigma^2}} \quad (2.9)$$

In order to obtain the posterior distribution from Equation 2.9, \bar{y} is inserted into and subtracted from the expression $\sum(y - \theta)^2$ and the equations below are obtained:

$$\sum (y - \theta)^2 = \sum (y - \bar{y} + \bar{y} - \theta)^2 \quad (2.10)$$

$$\sum (y - \theta)^2 = \sum (y - \bar{y})^2 + n(\theta - \bar{y})^2 \quad (2.11)$$

$P(\theta|y)$ can be rewritten proportionally as below:

$$P(\theta|y) \propto e^{-\frac{(\theta-\mu)^2}{2\tau^2}} e^{-\frac{\sum(y-\bar{y})^2+n(\theta-\bar{y})^2}{2\sigma^2}} \quad (2.12)$$

The posterior distribution can be obtained by subtracting the expression independent from θ from Equation 2.12, as below:

$$P(\theta|y) \propto e^{-\frac{(\theta-\mu)^2}{2\tau^2}} e^{-\frac{n(\theta-\bar{y})^2}{2\sigma^2}} \quad (2.13)$$

$$P(\theta|y) \propto e^{-\frac{1}{2}\left(\frac{\theta^2(n\tau^2+\sigma^2)-2\theta(n\bar{y}\tau^2+\mu\sigma^2)}{\sigma^2\tau^2}\right)} \quad (2.14)$$

$$P(\theta|y) \propto e^{-\frac{1}{2} \left(\frac{(\theta - \frac{\mu\sigma^2 + n\tau^2\bar{y}}{n\tau^2 + \sigma^2})^2}{\frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}} \right)} \quad (2.15)$$

As it is seen in Equation 2.15, the final form of the posterior distribution is a normal distribution with a $\frac{\mu\sigma^2 + n\tau^2\bar{y}}{n\tau^2 + \sigma^2}$ average and a $\frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$ variance ($P(\theta|y) \sim N\left(\frac{\mu\sigma^2 + n\tau^2\bar{y}}{n\tau^2 + \sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}\right)$). As it can be understood from this example, when the prior distribution is specified as normal distribution, the posterior distribution is also distributed normally. When the parameters of the posterior distribution are examined, σ^2 being greater than τ^2 means that prior information is more precise. The increase in the value of τ^2 causes the prior distribution to lose its influence gradually. In case of $\tau^2 \rightarrow \infty$, the results converge to the classical approach.

Another issue to be addressed in Bayesian approach is the concept of precision. This concept has a direct relation with variance. Precision is expressed as 1/variance. As it can be seen from this expression, there is an inverse proportion between precision and variance. A decrease in variance approximates precision to ∞ . The variance of the prior distribution and the data can be expressed with the precision concept. Let $\bar{p} = \frac{1}{\tau^2}$ indicate the precision of the prior distribution, and $p = \frac{1}{\sigma^2/n}$ indicate the precision of the sample. According to this information, the average and the variance of the posterior distribution can be written in terms of precision. Variance of the posterior distribution has the form $\frac{1}{\frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}}$. This expression can be simplified as $V(\theta|y) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$. When the expression is examined, it is seen that the precision of the posterior distribution is the sum of the precision of the prior distribution and the precision of the sample.

Normal & Inverse Gamma Model: When variance is known and when the prior is specified as normal distribution for the average, the posterior distribution in Equation 2.15 is obtained. Let us consider the opposite of the situation in this section. Let us obtain the posterior distribution when the average is known but the variance is

not known. Let us assume that the likelihood function is distributed normally with the parameters μ and σ^2 and the prior distribution of the variance has the $IG(a_0, \beta_0)$ inverse gamma distribution with a_0 and β_0 parameters.

Accordingly, the posterior distribution can be written as the equation below:

$$p(\sigma^2|y, \mu) \propto p(y|\mu, \sigma^2)p(\sigma^2) \quad (2.16)$$

$$p(\sigma^2|y, \mu) \propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \left(\frac{\beta_0^{a_0}}{\Gamma(a_0)} \theta^{-(a_0+1)} \exp\left(-\frac{\beta_0}{\sigma^2}\right)\right) \quad (2.17)$$

Since we are interested in the distribution of σ^2 , expression independent of σ^2 can be dropped from Equation 2.17.

$$p(\sigma^2|y, \mu) \propto \prod_{i=1}^n (\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \theta^{-(a_0+1)} \exp\left(-\frac{\beta_0}{\theta}\right) \quad (2.18)$$

$$p(\sigma^2|y, \mu) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right) (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{\beta_0}{\sigma^2}\right) \quad (2.19)$$

The posterior distribution of σ^2 can be found as below by making some mathematical corrections on Equation 2.19.

$$p(\sigma^2|y, \mu) \propto (\sigma^2)^{-(a_0 + \frac{n}{2} + 1)} \exp\left(-\left[\frac{\beta_0}{\sigma^2} + \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right]\right) \quad (2.20)$$

$$p(\sigma^2|y, \mu) \propto (\sigma^2)^{-(a_0 + \frac{n}{2} + 1)} \exp\left(-\left[\frac{2\beta_0 + 2\left(\frac{\sum_{i=1}^n (y_i - \mu)^2}{2}\right)}{2\sigma^2}\right]\right) \quad (2.21)$$

$$p(\sigma^2|y, \mu) \propto (\sigma^2)^{-(a_0 + \frac{n}{2} + 1)} \exp\left(-\left[\frac{\beta_0 + \left(\frac{\sum_{i=1}^n (y_i - \mu)^2}{2}\right)}{\sigma^2}\right]\right) \quad (2.22)$$

When we specify the prior distribution as inverse gamma, the posterior distribution is also obtained as inverse gamma, as it can be seen in Equation 2.22. The parameters of the posterior distribution a_1 and β_1 are found $a_0 + \frac{n}{2}$ and $\beta_0 + \left(\frac{\sum_{i=1}^n (y_i - \mu)^2}{2}\right)$ respectively. After all these operations, the posterior distribution will be inverse gamma again for the situations in which the data is distributed normally and the prior distribution pertaining to the variance is specified as inverse gamma.

CHAPTER THREE

BAYESIAN COMPUTATION

When complex problems are encountered in Bayesian approaches, it may not be possible to obtain the posterior distribution due to the inability to take the required integrals. The increase in the size of the θ parameter is another factor affecting the insolubility. It becomes difficult to obtain marginal posterior distributions of the parameters as the size increases, and generally these cannot be expressed in mathematical form. In recent years, methods known as the Markov Chain Monte Carlo, enable the use of Bayesian approaches in complex problems. This section will discuss the Markov Chain Monte Carlo (MCMC), which is the most widely used, a asymptotic approach and a stochastic simulation method known as the Bayesian central limit theorem.

3.1 Bayesian Central Limit Theorem

If the number of the observation in the data set is very large, the likelihood will be quite peaked, and small changes in the prior will have little effect on the resulting posterior distribution. In this condition, the following theorem, which is called Bayesian Central Limit Theorem, shows that the posterior distribution $P(\theta|x)$ will be approximately normal.

Theorem: $X_1, X_2 \dots X_n$ be a random sample from the distribution $f_i(x_i|\theta)$ and thus likelihood function is $f(x|\theta) = \prod_{i=1}^n f_i(x_i|\theta)$. Suppose the prior $\pi(\hat{\theta})$ and $f(x|\theta)$ are positive and $\hat{\theta}^\pi$ the posterior mode of θ . Then the posterior distribution $P(\theta|x)$ for large n can be approximated by a normal distribution having mean equal to posterior mode ($\hat{\theta}^\pi$), and covariance matrix equal to minus the inverse Hessian matrix $[(I^\pi(x))]^{-1}$ of the log posterior evaluated at the mode, $p(\theta|x) \sim N(\hat{\theta}^\pi, I(\hat{\theta}^\pi)^{-1})$. Hessian matrix $[(I^\pi(x))]^{-1}$ is generalized observed Fisher information matrix for θ . This matrix is given in equation (3.1).

$$I_{ij}^{\pi}(x) = - \left[\frac{\delta^2}{\delta\theta_i \delta\theta_j} \log f(x|\theta)\pi(\theta) \right]_{(\theta=\hat{\theta}^{\pi})} \quad (3.1)$$

3.2 Markov Chain Monte Carlo Methods

In the past, while the Bayesian technique has always been powerful, it has not always been practical. Initially Bayesian analysis was generally limited to problems involving a very small set of statistical distributions to describe the prior information and the likelihood of the observed data. These so-called “conjugate” distributions have the property that when the prior distribution and the likelihood function for the data are combined with Bayes theorem, the posterior distribution is of the same type as the prior but with updated parameters. If the analysis involved “conjugate” distributions, the posterior distribution could be derived analytically. However, computational advances have made it possible to evaluate complex Bayesian models by using numerical approximation and simulation techniques. This has increased the range of problems and sophistication of analyses now accessible to Bayesian techniques far beyond those limited to that small set of statistical distributions accessible previously. One of these computational techniques applies Markov Chain Monte Carlo (MCMC), which is essentially Monte Carlo integration using Markov chains, simulation.

The Monte Carlo method is based on a simple idea: one can learn anything about a posterior distribution by repeatedly drawing from it and empirically summarizing those draws. For instance, we might be interested in computing the posterior expected value, which can be done analytically by computing a high dimensional integral:

$$E[\theta|x] = \int_{\theta} \theta f(\theta|x) d\theta \quad (3.2)$$

If we were able to produce a random sequence of K draws $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(K)}$ from $f(\theta|x)$, we can approximate the posterior expected value by taking the average of these draws:

$$E[\theta|x] = \int_{\theta} \theta f(\theta|x) d\theta \approx \frac{1}{K} \sum_{k=1}^K \theta^k \quad (3.3)$$

The precision of the estimate depends solely on the quality of the algorithm employed, and the number of draws taken from the posterior distribution what all of these methods have in common is that they serve to compute high-dimensional integrals using simulation. A great deal of work in numerical analysis is devoted to understanding the properties of algorithms; for such a discussion of commonly used methods in Bayesian statistics, see Tierney (1994).

To use the Monte Carlo method to summarize posterior distributions, it is necessary to have algorithms that are well-suited to producing draws from commonly found posterior distributions. Two algorithms, the Gibbs sampling and Metropolis-Hastings algorithms, have proven to be very useful for applied Bayesian work. Both of these algorithms are MCMC methods, which mean that the sequences of $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(K)}$ draws are dependent; each draw $\theta^{(K+1)}$ depends only on the previous draw $\theta^{(K)}$. The sequence of draws thus forms a Markov chain. Algorithms are constructed such that the Markov chain converges to the posterior density (its steady state) regardless of the starting values. The most commonly usages of the MCMC algorithms are presented in this section.

3.2.1 Gibbs Sampling

The Gibbs sampler (Geman & Geman, 1984) has its origins in image processing. It is thus somewhat ironic that the powerful machinery of MCMC methods had essentially no impact on the field of statistics until rather recently.

The Gibbs sampler is a special case of Metropolis-Hastings sampling wherein the random value is always accepted. The task remains to specify how to construct a Markov Chain whose values converge to the posterior distribution. The key to the Gibbs sampler is that one only considers univariate conditional distributions. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms. Thus, one simulates n random variables sequentially

from the n univariate conditionals rather than generating a single n -dimensional vector in a single pass using the full joint distribution.

Suppose that our parameter vector θ has m components, making our target distribution $f(\theta_1, \theta_2, \dots, \theta_K | x)$. To use the Gibbs sampler, one begins by choosing starting values $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_K^{(0)}$ (these are usually chosen near the posterior mode or the maximum likelihood estimates). One then repeats, for $T = 1, \dots, t$ iterations (making sure to store the sequence of draws at each iteration):

Draw $\theta_1^{(t)}$ from $p(\theta_1 | \theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_K^{(t-1)}, x)$
 Draw $\theta_2^{(t)}$ from $p(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_K^{(t-1)}, x)$
 Draw $\theta_3^{(t)}$ from $p(\theta_3 | \theta_1^{(t)}, \theta_2^{(t)}, \theta_4^{(t-1)}, \dots, \theta_K^{(t-1)}, x)$
 \vdots
 Draw $\theta_K^{(t)}$ from $p(\theta_K | \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_{K-1}^{(t)}, x)$

Repeating this process t times, generates a Gibbs sequence of length t . To obtain the desired total of m sample points, one samples the chain (i) after a sufficient burn-in to removal the effects of the initial sampling values and (ii) at set time points following the burn-in. The Gibbs sequence converges to a stationary distribution that is independent of the starting values, and by construction this stationary distribution is the target distribution we are trying to simulate (Tierney, 1994).

To illustrate the Gibbs sampling algorithm in practice, example shows sampling from a Poisson/Gamma hierarchical model, where $Y_i | \theta_i \sim \text{Poisson}(\theta_i t_i)$, $\theta_i \sim G(\alpha, \beta)$ and $\beta \sim IG(c, d)$ respectively. Here let us assume that t_i and the hyperparameters c and d are known. The mathematical form of these distributions is as below:

$$f(y_i | \theta_i) = \frac{e^{-(\theta_i t_i)} (\theta_i t_i)^{y_i}}{y_i!}, y_i \geq 0, \theta_i > 0 \quad (3.4)$$

$$g(\theta_i|\beta) = \frac{\theta_i^{\alpha-1} e^{-\theta_i/\beta}}{\Gamma(\alpha)\beta^\alpha}, \alpha > 0, \beta > 0 \quad (3.5)$$

$$h(\beta) = \frac{e^{-1/(\beta d)}}{\Gamma(c)d^c\beta^{c+1}}, c > 0, d > 0 \quad (3.6)$$

The gamma prior distribution and poisson likelihood function are conjugate with the inverse gamma hyperprior and the gamma prior distribution. Here the aim is to obtain the marginal posterior distribution of θ_i using these priors. A close form could not be obtained for the marginal posterior distribution of θ_i . However, the full conditioned distributions of θ_i and β can be easily found using the Gibbs sampling (Carlin & Louis, 2009).

The full conditioned distribution distributions of θ_i can be obtained as below:

$$p(\theta_i|\theta_{j \neq i}, \beta, y) \propto f(y_i|\theta_i)g(\theta_i|\beta) \quad (3.7)$$

$$p(\theta_i|\theta_{j \neq i}, \beta, y) \propto \theta_i^{y_i+\alpha-1} e^{-\theta_i(t_i+1/\beta)} \quad (3.8)$$

$$p(\theta_i|\theta_{j \neq i}, \beta, y) \propto G(\theta_i|y_i + \alpha, (t_i + 1/\beta)^{-1}) \quad (3.9)$$

Similarly, the full conditioned distribution for β can be obtained as below:

$$p(\beta|\{\theta_i\}, y) \propto \left[\prod_{i=1}^k g(\theta_i|\beta) \right] h(\beta) \quad (3.10)$$

$$p(\beta|\{\theta_i\}, y) \propto \left[\prod_{i=1}^k \frac{e^{-\theta_i/\beta}}{\beta^\alpha} \right] \frac{e^{-1/(\beta d)}}{\beta^{c+1}} \quad (3.11)$$

$$p(\beta|\{\theta_i\}, y) \propto \frac{e^{-\frac{1}{\beta}(\sum_{i=1}^k \theta_i + \frac{1}{d})}}{\beta^{k\alpha+c+1}} \quad (3.12)$$

$$p(\beta|\{\theta_i\}, y) \propto IG(\beta|k\alpha + c, (\sum_{i=1}^k \theta_i + 1/d)^{-1}) \quad (3.13)$$

In the Equations 3.16 and 3.20 above, full conditioned distributions for θ_i and β were obtained in two forms. Using conjugate priors and selecting a hierarchical structure eased the operations.

When a conjugate prior is not selected in Bayesian methods, the full conditioned distributions may not transform into a known distribution. In such a case, it is more appropriate to use another MCMC algorithm, the Metropolis Hasting Algorithm.

3.2.2 Metropolis Hasting Algorithm

Another algorithm that enjoys common use in applied Bayesian statistics is the Metropolis-Hastings algorithm, first introduced by Metropolis et al. (1953) and generalized by Hastings (1979). It is also the case that the Gibbs sampling algorithm is a special case of the Metropolis-Hastings algorithm.

Metropolis algorithm includes a unnormalized posterior distribution $h(\theta)$ and a proposal distribution. In order to implement the Metropolis algorithm, first the proposal distribution $q(\theta^*|\theta^{(t-1)})$ should be specified. It is assumed that the proposal distribution in the Metropolis algorithm is symmetrical. M-H algorithm does not require such an assumption. If the proposal distribution is not well chosen, all of the candidate views are rejected and the chain remains stuck in certain points for the great proportion of the time. Therefore, the selected proposal distribution and the MCMC expression should be formed more carefully by confirming the tendency of the algorithm (Koop, 2003).

Contrary to the Gibbs sampling, the candidate point is not always accepted in the Metropolis algorithm. An initial value is specified for each parameter, and the algorithm is continued until convergence is obtained. Algorithm steps are as below:

Step 1: Initial values $\theta^{(0)}$ are specified.

Step 2: $\alpha = \frac{h(\theta^*)}{h(\theta^{t-1})}$ probability is computed. Since it is not a requirement for the proposal distribution to be symmetrical in M-H algorithm ($q(\theta^{t-1}|\theta^*) \neq q(\theta^*|\theta^{t-1})$), probability is computed as below:

$$\alpha = \frac{h(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}{h(\boldsymbol{\theta}^{t-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})} \quad (3.14)$$

Step 3: If $\alpha \geq 1$ then the candidate point is accepted and expressed as $\boldsymbol{\theta}^* = \boldsymbol{\theta}^t$; later, another candidate point is selected. If for the selected candidate point, than this point is accepted with the probability; if not it is rejected with the probability. If the proposal distribution is not well chosen, all the candidate views are rejected and the chain remains stuck in certain points for the great proportion of the time. Therefore, the selected proposal distribution and the MCMC expression should be formed more carefully by confirming the tendency of the algorithm (Koop, 2003).

With the algorithm above, a Markov chain is formed. In this chain, each simulation value is only linked to the previous value. After reaching the required iteration number, the convergence is obtained. Therefore, the desired posterior distribution is obtained.

CHAPTER FOUR

BAYESIAN AND SPLINE REGRESSION

Regression analysis is one of the most widely used statistical tools. Nowadays, this analysis is carried out with different alternative approaches. The most commonly used alternative approaches in the literature are called Bayesian methods and Splines. In this chapter provides a brief summary of the Bayesian and Spline regression methods.

4.1 Bayesian Regression

Regression analysis is used to answer questions about how one variable depends on the level of one or more other variables. Recently, this analysis is carried out with different alternative approaches. Bayesian methods are one of these approaches. In some situations there is an advantage of being Bayesian when fitting a regression model. These situations where it pays to be Bayes include:

- When there is prior information about the regression coefficients.
- When we are interested in estimating functions of regression coefficients.
- When the regression model is non-linear.
- When the distribution of the errors is non-normal.
- When we have repeated measurements on some sample units.

In this chapter provides a brief summary of the Bayesian regression methods.

In the usual multiple regression problem, we are interested in describing the variation in a dependent (response) variable y in terms of k independent (predictor) variables x_1, x_2, \dots, x_k . We describe the mean value of y_i , the response for the i th individual, as

$$E(y_i|\beta, X) = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}, \quad i = 1, \dots, n \quad (4.1)$$

where x_1, x_2, \dots, x_k are the independent values for the i th individual and $\beta_1, \beta_2, \dots, \beta_k$ are unknown regression parameters. The $\{y_i\}$ are assumed to be conditionally independent given values of the parameters and the independent variables. In the ordinary linear regression setting, we assume equal variances, where $\text{var}(y_i|\theta, X) = \sigma^2$. Finally, we assume that the errors $e_i = y_i - E(y_i|\beta, X)$ are independent and normally distributed with mean 0 and variance σ^2 [$e_i \sim N(0, \sigma^2)$]. Also, error terms are independent of each other.

The usual regression models can be easily formulated within a Bayesian framework. Bayesian methods can be used for any probability distribution. Methods presented in section (4.1) come from the Bayesian theory for normally distributed random variables.

In the classical regression, the distribution of X is assumed to provide no information about the conditional distribution of Y given X (Gelman et al, 2004) but in Bayesian regression, the distribution of the independent variables is included the likelihood function. For this reason regarding distribution of the independent variable is eliminated with proportional expression. As a result, Bayesian regression does not deal with the distribution of the independent variable. The mathematical presentation of this situation is given below.

Let ψ denote the parameter vector of X . If prior distribution is independent, $\beta_1, \beta_2, \dots, \beta_k, \sigma^2$ and ψ , we can write this equation,

$$p(\psi, \beta, \sigma^2) = p(\psi)p(\beta, \sigma^2) \quad (4.2)$$

Then, posterior distributions can be divided two factors,

$$p(\psi, \beta, \sigma^2|X, y) = p(\psi|X)p(\beta, \sigma^2|X, y) \quad (4.3)$$

Since the distribution of the independent variables is included the likelihood function, we can write proportional equation is given below.

$$p(\beta, \sigma^2 | X, y) \propto p(\beta, \sigma^2) p(y | X, \beta, \sigma^2) \quad (4.4)$$

A similar result is obtained the distribution of independent variable.

4.1.1 Bayesian Regression Model

In linear regression, the observations consist of a response variable in a vector y and one or more predictor variables in a matrix X . The parameters are the regression coefficients β and the error variance of the fitted model, σ^2 . The model that relates observations and parameters is written:

$$y | \beta, \sigma^2, X \sim N(X\beta, \sigma^2 I) \quad (4.5)$$

The matrix notation of this model is,

$$y_i = X\beta + \varepsilon \quad (4.6)$$

$$P(y | \beta, \sigma^2, X) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)\right\} \quad (4.7)$$

Bayesian regression analysis begins with a prior distribution. Since a noninformative prior distribution assigns the same probability to each possible value of the parameters, it is most commonly used in linear regression. A noninformative prior distribution that is commonly used for linear regression is

$$p(\beta, \sigma) \propto \frac{1}{\sigma} \quad (4.8)$$

Using the likelihood function and prior distributions which are obtained from equations (4.7) and (4.8), we achieved the posterior distribution of β given σ^2 .

$$p(\beta, \sigma | y, X) \propto \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2} \left[vs^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]\right\} \quad (4.9)$$

where $vs^2 = (y - X\beta)'(y - X\beta)$ and $v = n - k$.

The marginal posterior probability distribution of β which is derived by integrating the posterior distribution of β given σ^2 over all possible values of σ^2 .

$$p(\beta|y) = \int_0^{\infty} p(\beta, \sigma|y) d\sigma = \int_0^{\infty} p(\beta|\sigma, y) p(\sigma|y) d\sigma \quad (4.10)$$

$$p(\beta|y) \propto \left[1 + \frac{1}{v} (\beta - \hat{\beta})' \frac{X'X}{s^2} (\beta - \hat{\beta}) \right]^{-\frac{(k+v)}{2}} \quad (4.11)$$

Equation (4.11) is written $(\beta|y) \sim \text{Multivariate Student } t(n - k, \hat{\beta}, s^2)$. The multivariate Student t distribution has three parameters, the degrees of freedom $(n - k)$, the mean $\hat{\beta}$, and the scale factor s^2 .

A similar process can follow for σ^2 . The marginal posterior distribution of σ^2 (i.e. the integral over all possible values of β of the joint distribution of β and σ^2) is

$$p(\sigma|y) = \int_{-\infty}^{\infty} p(\beta, \sigma|y) d\beta \quad (4.12)$$

$$p(\sigma|y) \propto \frac{1}{\sigma^{v+1}} \exp\left(-\frac{vs^2}{2\sigma^2}\right) \quad (4.13)$$

Equation (4.13) is written $\sigma^2|y \sim \text{Inverse } \chi^2(n - k, s^2)$ and it says that the probability distribution of σ^2 given y follows an inverse χ^2 distribution.

The other purpose of regression analysis is prediction. Let \tilde{X} denote the matrix of independent variables and \tilde{y} denote the values of the dependent variable. The predictive distribution, \tilde{y} , given a new set of predictors \tilde{X} has mean

$$E(\tilde{y}|y) = \tilde{X}\beta \quad (4.14)$$

The marginal posterior distribution of the variance of this prediction is

$$\text{Var}(\tilde{y}|\sigma^2, y) = (I + \tilde{X}V_{\beta}\tilde{X}')\sigma^2 \quad (4.15)$$

where I is the identity matrix. This variance formula has two components, $I\sigma^2$ for sampling variance of the new observations and $\tilde{X}V_{\beta}\tilde{X}'\sigma^2$ for uncertainty about β . The marginal posterior distribution of \tilde{y} given y is

$$p(\tilde{y}|y) = \int p(\tilde{y}|\beta, \sigma^2)p(\beta, \sigma^2|y)d\beta d\sigma^2 \quad (4.16)$$

Equation (4.16) is written $p(y_p|y) \sim \text{Multivariate Student } t [n - k, \tilde{X}\hat{\beta}, (I + \tilde{X}V_{\beta}\tilde{X}')s^2]$

4.2 Spline Regression

Spline Regression is one of the most popular and powerful techniques in nonparametric regression. Spline regression models have been used in many fields such as operation, econometrics, medicine and agriculture. Effective results are obtained with the application of spline regression on datum which is not explained by linear and high degree regression.

Regression models in which the function changes at one or more points along the range of the predictor are called splines, or piecewise polynomials, and the location of these shifts are called knots. If the knots are fixed by the analyst, then splines can be fitted quite easily with the regression procedure. A spline model is hypothesized when the analyst expects that the relationship between the predictor and the response variable is altered at some value or values along the range of the predictor. The shift at the knot points could involve a change in the form of the relationship, such as a shift from a linear to a quadratic relationship, the addition or subtraction of a constant to all predicted response values to the right of the knot, or simply a change in the slope, acceleration, etc. of the regression function. The general form of the spline regression model is described below.

$$m(x_i) = \beta_0 + \beta_1x + \dots + \beta_px^p + \sum_{k=1}^K b_k(x - t_k)_+^p \quad (4.17)$$

where t_k are fixed and known knots, K are the number of knots, $\beta_0, \beta_1, \dots, \beta_p, b_1, b_2, \dots, b_K$ are unknown regression coefficients in the model. Also, p indicate the degree of spline regression model and $(x - t_k)_+^p$ statement is included as basis function in the model. An important characteristic of function $(x - t_k)_+^p$ is that equal to 0 value as minimum and it is positive definite. If the value of independent variables smaller than knot value; the value of function will be equal to 0. Otherwise, if the value of independent variables greater than knot value, the value of function will be equal to the degree of p th of the value of independent variable minus knot value.

We have been assuming that the knots are known. In general, they are unknown, and spline regression problem can be formulated as an ordinary regression problem with a transformed predictor, it is possible to apply variable selection techniques such as back-ward selection to choose a set of knots. The usual approach is to start with a set of knots located at a subset of the order statistics of the predictor. Then backward selection is applied, using the truncated power basis form of the model. Each time a basis functions eliminated, the corresponding knot is eliminated. Once the knots are fixed, spline regression is a parametric regression. Figure 4.1 and Figure 4.2 exhibit an example of a least-squares spline with manually-selected knots, applied to a data set consisting of the ratios of exports to imports.

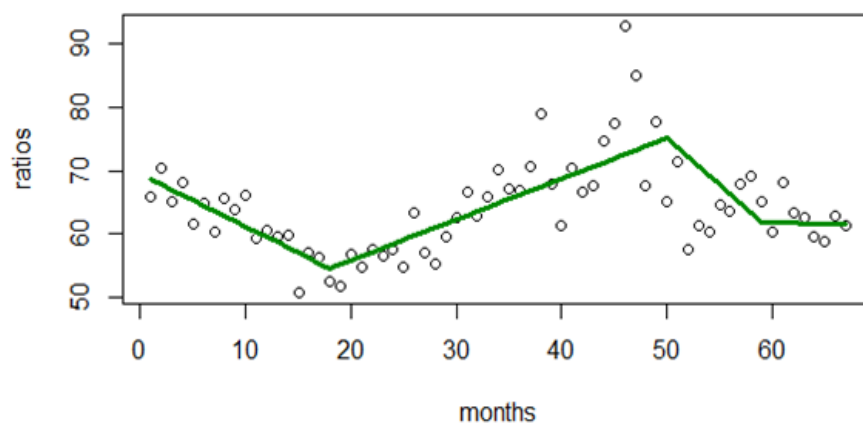


Figure 4.1 A least-squares spline fit to ratios of export to import data using the manually-selected knots. The knots used were 18, 50, 59.

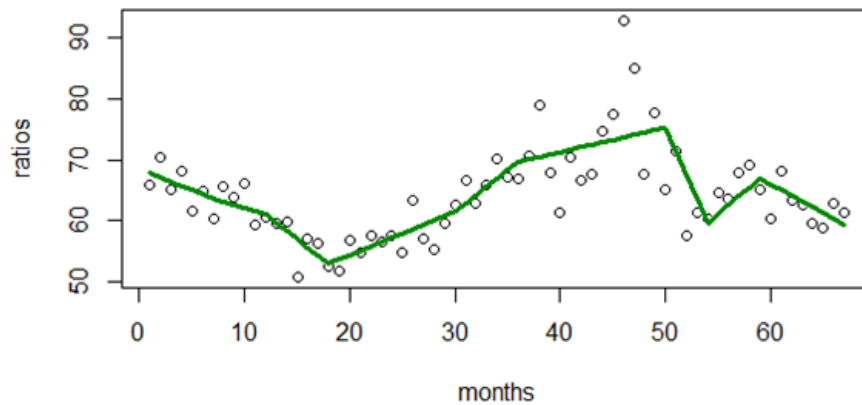


Figure 4.2 A least-squares spline fit to ratios of export to import data using the manually-selected knots. The knots used were 6, 12, 18, 30, 36, 50, 54, 59.

A substantial improvement can be obtained by manually selecting additional knots. Adjusting the knot that was already there improves the fit as well. As it is seen from figures, the same data set is examined with different number of knots. The first and second figure, respectively, 3 and 8 knots are determined. In this way, appropriate data set for spline regression is interested; analysis can be made on the data set by determining knots via researchers. Scatterplots encountered in the daily life, selection of knots location and determination of the number of knots is very difficult. Because locations and number of knots not always clearly apparent. In such circumstances, researchers can constitute more than one model. Then, which model is better to be decided by making comparisons between models. There is a set of criteria that can be used in decision making. Some of these criteria F statistic, R-Squared, Adjusted R-Squared, whether regression coefficients are statistically significant can be listed in the form.

4.2.1 Penalized Spline Regression

Penalized spline regression models are a popular statistical tool for curve fitting problems due to their flexibility and computational efficiency. It is a nonparametric regression technique that relies on principles of statistical theory to minimize the possibility of overfitting (Keele, 2008). The basic idea behind penalized regression methods is to quantify the notion of roughness of a curve through a suitable penalty functional and then to pose the estimation problem in a way that makes explicit the

necessary compromise between bias and variability in curve fitting. Spline regression needs to choose the number of knots and their positions but estimation is sensitive to this choice. Penalized spline regression uses a penalization parameter (λ), which is related to the fluctuations of the regression function, to reduce the impact of this choice. Consider the regression model;

$$y_i = m(x_i) + \varepsilon_i \quad (4.18)$$

where $m(\cdot)$ is a smooth function which is defined as,

$$m(x) = \beta_0 + \beta_1 X + \sum_{j=1}^q \beta_{ij}(x - K_{ij})_+ \quad (4.19)$$

The aim of the regression analysis to estimate the regression function f , where $E(Y|X) = m(x)$. Here, we directly solve for the function f that minimizes the following objective function, a penalized version of the least squares objective:

$$\sum_{i=1}^n \{y_i - m(x_i)\}^2 + \frac{1}{\lambda} \theta^T D \theta \quad (4.20)$$

where $\theta = (\beta_0, \beta_1, \beta_{i1}, \beta_{i2}, \beta_{iq})$ is the vector of unknown regression coefficients. The first term captures the fit to the data, while the second penalizes curvature. Here, λ is the smoothing parameter, the selection of the λ smooth parameter is of great importance in penalized spline regression. The case $\lambda = 0$ corresponds to the unconstrained case. Increasing the value of λ downweights the influence of the knots and gives a less rough fit. If we take λ to be very large, then the effect of the knots diminishes and the least-squares line is approached. There exist some methods for choosing λ and the knot locations from the data.

In equation (4.21), D is a known positive semi-definite penalty matrix. It is defined as follows;

$$D = \begin{bmatrix} 0_{(p+1) \times (p+1)} & 0_{(p+1) \times K} \\ 0_{K \times (p+1)} & \Sigma^{-1} \end{bmatrix} \quad (4.21)$$

In Bayesian approach to avoid overfitting, we penalize the b 's by assuming that the coefficients of $(x - t_k)_+^p$ are normally distributed random variables with mean 0 and variance σ_b^2 to be estimated. (Gimenez et al, 2009) This is the reason why this approach is referred to as penalized splines (Ruppert et al, 2003). The selection of the smooth parameter $\lambda = \tau_\varepsilon/\tau_\beta = \sigma_b^2/\sigma_\varepsilon^2$ is of great importance in penalized Bayesian spline regression. The small value of λ corresponds oversmoothing. The large value of λ corresponds undersmoothing. In this study, we proposed a new smoothing parameter using the information content of normal distribution in Bayesian framework. Under the assumption of the coefficients of basis functions are normally distributed, the new smoothing parameter is defined as the ratio of the information content of normal distribution, $(\lambda^* = \log_2[\sigma_b(2\pi e)^{1/2}] / \log_2[\sigma_\varepsilon(2\pi e)^{1/2}])$. The performance of this parameter will be investigated in the application chapter.

CHAPTER FIVE APPLICATION

This chapter presents an application to compare the performance of spline, Bayesian spline and penalized Bayesian spline models. Our aim is to compare the performance of three different models in terms of their value of coefficient of determination. The models are illustrated with an application to ratios of export to import data set given in Turkish Statistical Institute (TÜİK). These data consist of sixty-seven month periods (May 2007 to November 2012). The independent variable and dependent variable are defined respectively as month and the ratio of export to import.

Then spline regression was applied for this data set. In this data set we specified four interior knots given by (17, 49, 53, 57) and the degree of the spline is one. Using the R code and then uses least squares to construct the regression model for ratios of export to import data set. The results of spline regression are given in Table 5.1.

Table 5.2 The results of Spline Regression

```

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.5783    2.0850  33.850 < 2e-16 ***
x            -1.0270    0.1634  -6.285 3.86e-08 ***
b1           1.7940    0.2119   8.466 6.96e-12 ***
b2          -5.7076    0.8489  -6.723 6.91e-09 ***
b3           7.2988    1.6698   4.371 4.89e-05 ***
b4          -3.1246    1.2499  -2.500  0.0151 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.475 on 61 degrees of freedom
Multiple R-squared:  0.6691, Adjusted R-squared:  0.6419
F-statistic: 24.66 on 5 and 61 DF, p-value: 1.698e-13

```

From the Table 5.1; intercept, the coefficient of independent variable and coefficient of the basis functions in the model were obtained. All of these coefficients are statistically significant. According the value of F-statistic, the model is valid. Coefficient of determination (R-Squared) for this model is obtained as

0.6691. Thereafter we investigated the data to satisfy the regression assumptions and were obtained from following graphs.

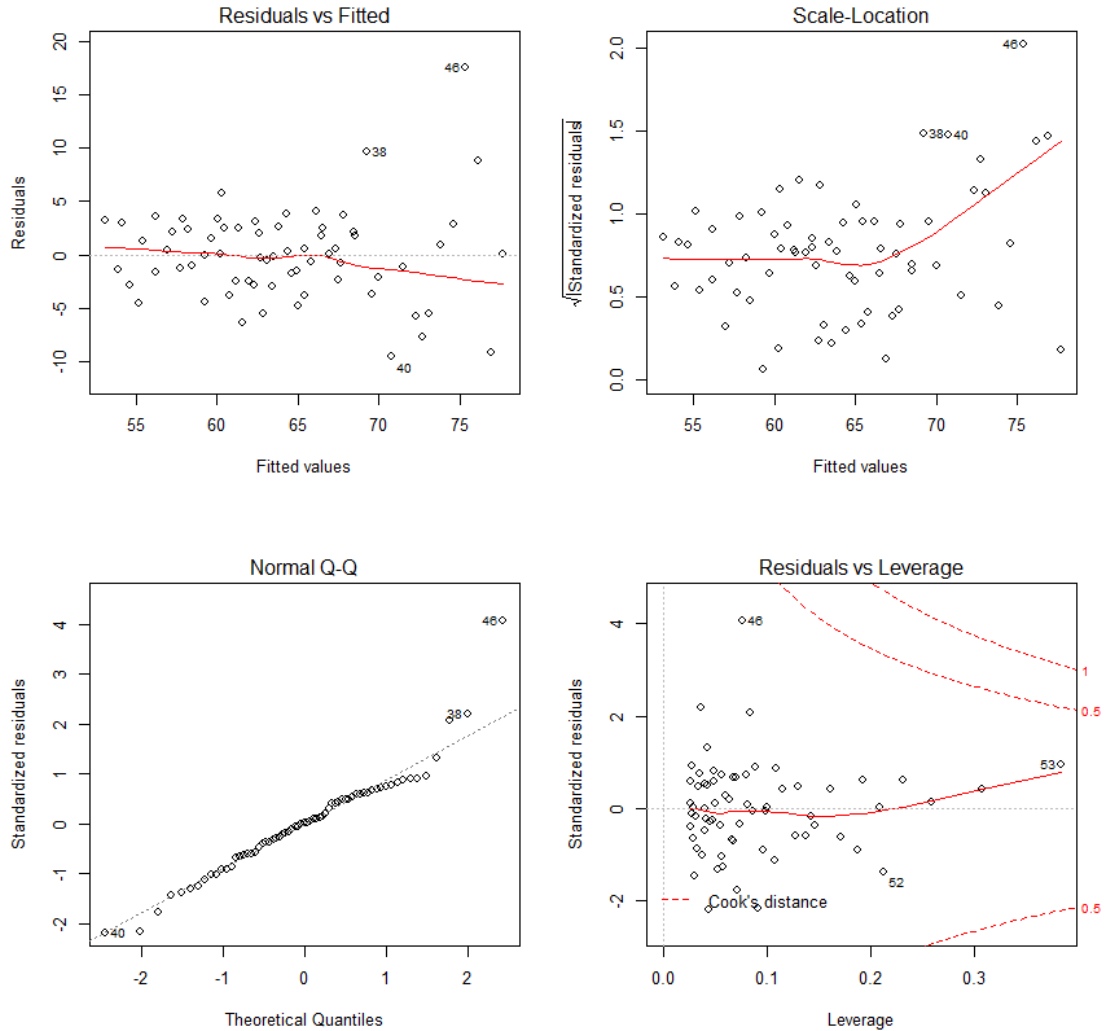


Figure 5.1 Plots for regression assumptions

We saw that all assumptions were satisfied out of correlated residuals. Since we are interested in nonparametric regression techniques, we assume the residuals are uncorrelated.

Then, we applied Bayesian spline regression analysis for same data set. Prior distributions are determined for each parameter which is considered as random variables in the model. Parameters and it's a prior distributions are summarized in equation (5.1).

$$\left\{ \begin{array}{l} y_i = \beta_0 + \beta_1 x + \sum_{k=1}^4 b_k (x - t_k)_+ + \varepsilon_i \\ \beta_0, \beta_1 \sim N(0, 10^6) \\ b_k \sim N(0, 10^6) \\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \end{array} \right. \quad (5.1)$$

The different prior distribution can be selected for variance in the literature. In this analysis, the distribution of precision parameter $\tau_\varepsilon = \sigma_\varepsilon^{-2}$ is taken as gamma distribution. Using the WinBUGS code and then least squares to construct the Bayesian spline regression model for ratios of export to import data set. There are three different stages of WinBUGS program. These are, writing code for interest model, loading data and creating the initial values for the parameters respectively. The burn-in period, which was used to eliminate the effect of the initial values, was consisted 2000 iterations in this example. The WinBUGS code of this application is given by in appendix. The results of Bayesian spline regression is given in Table 5.2.

Table 5.2 The results of Bayesian Spline Regression

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta[0]	70.57	1.831	0.0129	66.93	70.57	74.16	2001	20000
beta[1]	-1.027	0.1436	0.001075	-1.307	-1.028	-0.7419	2001	20000
b[1]	1.795	0.1865	0.001448	1.427	1.796	2.159	2001	20000
b[2]	-5.703	0.7493	0.005345	-7.175	-5.704	-4.227	2001	20000
b[3]	7.285	1.471	0.01035	4.343	7.279	10.16	2001	20000
b[4]	-3.113	1.097	0.007845	-5.294	-3.111	-0.9428	2001	20000
sigmae[ps]	3.924	0.3148	0.002527	3.364	3.903	4.602	2001	20000

The values related to the posterior distribution such as posterior mean, posterior median, MC error, 2.5% and 97.5% quantiles were obtained. MC error is used to decide the parameters convergence or not. If this value is smaller than 0.05 we can decide parameter convergence. MC values of all parameters in Bayesian Spline Regression model is smaller than 0.05. So we decided that parameters of the models convergence. The R-Squared of the model was obtained 0.6697.

When we compared the two regression models, both models shown similar characteristics. The coefficients of β and b parameter vectors were very similar and the coefficients of determination of two models were obtained the same. But, the standard errors of parameter estimations of Bayesian spline regression were smaller

than spline regression models. For this reason, we conclude that Bayesian spline regression model parameter estimation is more reliable than spline regression model.

Akaike information criterion (AIC), (Akaike, 1973) and Bayesian information criterion (BIC), (Schwarz, 1978) are the two most popular information criteria in the literature. These information criteria are often used for model selection and variable selection in Bayesian analysis. To investigate this further we computed the value of the AIC and BIC for the spline regression model and for the Bayesian spline regression model. The results are presented in Table 5.3. The results show that the spline regression model provides a better fit to the data in terms of lower AIC and BIC.

Table 5.3 The values of AIC and BIC for spline regression and Bayesian spline regression

Model	AIC	BIC
Spline Regression	398.637	414.070
Bayesian Spline Regression	406.726	422.159

Penalized spline regression models are a popular statistical tool for curve fitting problems due to their flexibility and computational efficiency. For this reason, Bayesian penalized spline regression analysis was applied for the same data set. The penalty term, $\lambda = \sigma_b^2 / \sigma_\varepsilon^2$ which restricts fluctuations of $\hat{\eta}$ was added to the Bayesian spline model. The coefficient of determination and regression coefficients of this model were obtained for different penalty terms $1/\lambda$. The results are given in Table 5.4.

Table 5.4 Penalized Bayesian Spline Regression Model for different λ

Parameter	$1/\lambda=0,85$	$1/\lambda=2,25$	$1/\lambda=17.1$	$1/\lambda=267,6$
b_1	1.736	1.689	1.47	0.7639
b_2	-4.583	-4.239	-2.456	-0.755
b_3	5.518	4.311	1.352	-0.1455
b_4	-2.003	-1.302	-0.0028	-0.0057
σ_ε^2	20.912	21.669	58.339	357.588
σ_b^2	24.581	9.61	3.411	1.336
R^2	0.612	0.575	0.458	0.211

From the Table 5.4, we observe that the coefficients of basis functions decrease as the penalty term $1/\lambda$ increase. Also, the coefficient of determination of the model gradually diminishes. Another point is that if $1/\lambda$ is large, the effect of the knots diminishes and the model approaches to the least-squares line.

We calculated the coefficient of determination and regression coefficients for penalized Bayesian regression model to investigate the performance of the new smoothing parameter. The results are given in Table 5.5.

Table 5.5 Penalized Bayesian Spline Regression Model for different λ^*

Parameter	$1/\lambda^*=1$	$1/\lambda^*=1.164$	$1/\lambda^*=1.695$	$1/\lambda^*=2.787$
b_1	1.726	1.692	1.472	0.7641
b_2	-4.737	-4.212	-2.465	-0.7558
b_3	5.293	4.257	1.364	-0.1454
b_4	-1.876	-1.273	-0.006	-0.0057
σ_ε^2	21.169	21.734	58.125	356.454
σ_b^2	21.603	9.437	3.433	1.336
R^2	0.604	0.578	0.460	0.212

According to Table 5.5, small changes in λ^* have made drastic changes in smoothing of the model. So, we conclude that λ^* is more sensitive than λ . If the amount of

information contained of the distribution of basis functions increases, the value of $1/\lambda^*$ decreases. It corresponds under smoothing. If the information contained of the distribution of error term decreases, the value of $1/\lambda^*$ increases. This situation corresponds oversmoothing.

CHAPTER SIX

CONCLUSION

This thesis has been mainly motivated by the increased research activity in applied and methodological aspects of the nonparametric regression approach. We presented the three most common nonparametric regression models, which are called spline, Bayesian spline and penalized Bayesian spline, discussing advantages and disadvantages of them representations. In addition, we proposed a new smoothing parameter using the information content of normal distribution for penalized Bayesian spline regression model. The data application included in Chapter 5 concerned ratios of export to import data set given in Turkish Statistical Institute (TÜİK). These data consist of sixty-seven month periods (May 2007 to November 2012). Application is used to compare the performance of the regression models to that of the splines and different penalty terms.

When we compared the spline and Bayesian spline regression models, both models show similar characteristics. The coefficients of β and b parameter vectors were very similar and the coefficients of determination of two models were obtained same. But, the standard errors of parameter estimations of Bayesian spline regression were smaller than spline regression models. For this reason, we conclude that Bayesian spline regression model parameter estimation is more reliable than spline regression model. AIC and BIC are often used model selection and variable selection in Bayesian analysis. To investigate this further we computed the value of the AIC and BIC for the spline regression model and for the Bayesian spline regression model. The results show that the spline regression model provides a better fit to the data in terms of lower AIC. For this reason, classical spline regression is more preferable for this data set.

We also compared penalized Bayesian spline models using different penalty terms. The different models on the same data set have been set up using different value of λ . From the results, we observe that the coefficients of basis functions decrease as the penalty term $1/\lambda$ increase. Also, the coefficient of determination of

the model gradually diminishes. Another point is that If $1/\lambda$ is large, then the effect of the knots diminishes and the model approaches to the least-squares line. The selection of the smooth parameter $\lambda = \tau_\varepsilon/\tau_\beta = \sigma_b^2/\sigma_\varepsilon^2$ is of great importance in penalized Bayesian spline regression. The small value of λ corresponds oversmoothing. The large value of corresponds undersmoothing.

In addition, we proposed a new smoothing parameter using the information content of normal distribution. Under the assumption of the coefficients of basis functions are normally distributed, the new smoothing parameter ($\lambda^* = \log_2[\sigma_b(2\pi e)^{1/2}]/\log_2[\sigma_\varepsilon(2\pi e)^{1/2}]$) is defined as the ratio of the information content of normal distribution. According to results, small changes in λ^* have made drastic changes in smoothing of the model. So, we conclude that λ^* is more sensitive than traditional smoothing parameter (λ). If the amount of information contained of the distribution of basis functions increases, the value of $1/\lambda^*$ decreases. It corresponds undersmoothing. If the information contained of the distribution of error term decreases, the value of $1/\lambda^*$ increases. This situation corresponds oversmoothing. We conclude that the proposed smoothing parameter (λ^*) provides a better insight into the different levels of penalization terms that imposed the smoothing for spline curve. This can be useful for prior distribution inflection within a Bayesian inference framework. Also the proposed smoothing parameter performs smoothing as a parallel to known smoothing parameter in the literature and show similar characteristics. Accordingly, different smoothing parameters subject to the random variable can be identified.

REFERENCES

- Akaike, H., (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium Information Theory* , 267-281.
- Beaumont, M. A., & Rannala, B., (2004). The bayesian revolution in genetics. *Nature Reviews Genetics*, 5, 251-261.
- Berger, J. O., & Pericchi, L. R., (1996). The intrinsic bayes factor for model selection and prediction, *Journal of the American Statistical Association*, 91, 109–122.
- Box, G. E. P., & Tiao C. G., (1973). *Bayesian inference in statistical analysis*. London: Addison-Wesley.
- Carlin, B. P., & Louis, T. A., (2000). *Bayes and empirical Bayes methods for data analysis* (2nd edition), USA: Chapman and Hall.
- Carlin, B. P., & Louis, T. A., (2009). *Bayesian methods for data analysis* (3rd edition), USA: Chapman and Hall.
- Ciprian, M. C., Ruppert, D., & Wand, M. P., (2005). Bayesian analysis for penalized spline regression using winbugs, *Journal of Statistical Software*, 14, 1-24
- Ciprian, M. C., Ruppert, D., Carroll, R. J., Joshi, A. & Goodner, B., (2007). Spatially adaptive bayesian penalized splines with heteroscedastic errors, *Journal of Computational and Graphical Statistics*, 16, 265-288
- Demirhan, H., (2004). *Logaritmik doğrusal modellerde parametrelerin ve beklenen göze sıklıklarının bayesci kestirimi*. Master of Science Thesis. Hacettepe University.
- Dey, D. K., & Sinha, D., (1999). Bayesian model determination in lifetime data analysis, *Brazilian Journal of Probability and Statistics*, 2, 1-19.

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B., (2004). *Bayesian data analysis* (2nd edition), USA: Chapman and Hall.
- Geman, S., & Geman, D., (1984). *Stochastic relaxation, gibbs distribution and bayesian restoration of images*. Retrieved January 5, 2013, from <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4767596>.
- Gersch, W., & Kitagawa, G., (1995). *Smoothness priors analysis of quasiperiodic time series*. Retrieved December 5, 2013, from <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=540868>.
- Gimenez, O., Bonner, S. J., King, R., Parker, R. A., Brooks, S. P., Jamieson, L. E et al. (2009). WinBUGS for population ecologists: Bayesian modeling using Markov Chain Monte Carlo methods, *Modeling Demographic Processes In Marked Populations Environmental and Ecological Statistics*, 3, 883-915.
- O'Hagan, A., (1995). Fractional bayes factors for models comparison, *Journal of the Royal Statistical Society*, 57, 99-138.
- Hastings, W. K., (1979). Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, 57, 97-109.
- Ibrahim, J. G., & Chen, M. H., & Sinha, D., (2000). Power prior distributions for regression models, *Statistical Science*, 15, 45-60.
- Koop, G., (2003). *Bayesian econometrics*, USA, John Wiley & Sons Inc.
- Keele, L., (2008). *Semiparametric regression for the social sciences*, United Kingdom: Wiley & Sons, Ltd

- Metropolis, N., & Rosenbluth, A. W., & Rosenbluth, M. N., & Teller, A. H., & Teller, A., (1953). Equation of state calculations by fast computing machines, *The Journal of Chemical Physics*, 21, 1087–1092.
- Ntzoufras, I., (2009). *Bayesian modeling using winbugs*, John Wiley & Sons.
- O'Hagan, A., (1995). Fractional bayes factors for models comparison, *Journal of the Royal Statistical Society*, 57, 99-138.
- Schwarz, G., (1978). Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.
- Syversveen, A. R., (1998). *Noninformative Bayesian Priors. Interpretation And Problems With Construction And Applications*, USA, Cambridge University.
- Tierney, L., (1994). Markov chains for exploring posterior distributions, development and communication, *Annals of Statistics*, 22, 1701-1762.
- West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models biology* (2nd edition), New York: Springer.

APPENDIX

THE CODES OF PROGRAMS

Appendix 1: The codes of Classical Spline Regression in R programming.

```
x=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67)
```

```
y=c(65.9,70.3,65.1,68.2,61.6,64.8,60.4,65.5,63.9,66.1,59.3,60.6,59.4,59.8,50.7,57.1,56.3,52.5,51.8,56.7,54.6,57.4,56.5,57.5,54.8,63.4,57,55.2,59.5,62.6,66.5,62.9,65.9,70.2,67,66.9,70.6,78.9,67.9,61.2,70.4,66.6,67.6,74.7,77.5,92.9,85,67.7,77.8,65.1,71.5,57.4,61.3,60.4,64.6,63.5,67.9,69.1,65.1,60.3,68.1,63.3,62.5,59.5,58.7,62.9,61.2)
```

```
plot(x,y)
```

```
x17 <- ( x - 17 )
```

```
x17[ x17<0 ] <- 0
```

```
x49 <- ( x - 49 )
```

```
x49 [ x49<0 ] <- 0
```

```
x53 <- ( x - 53 )
```

```
x53[ x53<0 ] <- 0
```

```
x57 <- ( x - 57 )
```

```
x57[ x57<0 ] <- 0
```

```
fit <- lm( y ~ x + x17 + x49 + x53 + x57 )
```

```
print( summary( fit ) )
```

```
AIC(fit)
```

```
BIC(fit)
```

Appendix 2: The codes of Bayesian Spline Regression in WinBUGS programming

```

model{
C<-90.0
      pi<-3.141593
for (i in 1:n)
{response[i]~dnorm(m[i],taueps)
m[i]<-inprod(beta[],X[i,1])+inprod(b[],Z[i,1])
log.like[i] <- -0.5*log(2*pi)-0.5*log(sigmaeps)-0.5*(response[i]-m[i])*(response[i]-
m[i])/sigmaeps

like[i] <- exp( log.like[i] )
}
dm <- 7

      Deviance <- -2*sum(log.like[1:n])

      AIC <- Deviance + dm*2
      BIC <- Deviance + dm*log(n)
      L <- prod( like[1:n] )

for (k in 1:nknots){b[k]~dnorm(0,1.0E-6)}
for (l in 1:degree+1){beta[l]~dnorm(0,1.0E-6)}
taueps~dgamma(1.0E+1,1.0E-6);
sigmaeps<-1/(taueps);
sigma<-sqrt(sigmaeps);
for (i in 1:n)
{for (l in 1:degree+1){X[i,l]<-pow(covariate[i],l-1)}}
for (i in 1:n)
{for (k in 1:nknots)
{u[i,k]<-(covariate[i]-knot[k])*step(covariate[i]-knot[k])
Z[i,k]<-pow(u[i,k],degree)}}
for (i in 1:n) {
numerator[i] <- (m[i] - mean(response[]))*(m[i] -
mean(response[]))
denominator[i] <- (response[i] -
mean(response[]))*(response[i] - mean(response[]))
}
R2 <- sum(numerator[]) / sum(denominator[])
for (i in 1:n)
{}
}

list(n=67,nknots=4,degree=1,
knot=c(17,49,53,57),covariate = c(
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,

```

```
32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67),
  response =
c(65.9,70.3,65.1,68.2,61.6,64.8,60.4,65.5,63.9,66.1,59.3,60.6,59.4,59.8,50.7,57.1,56.3,52.5,51.8,56.7,54.6,57.4,56.5,57.5,54.8,63.4,57,55.2,59.5,62.6,66.5,62.9,65.9,70.2,67,66.9,70.6,78.9,67.9,61.2,70.4,66.6,67.6,74.7,77.5,92.9,85,67.7,77.8,65.1,71.5,57.4,61.3,60.4,64.6,63.5,67.9,69.1,65.1,60.3,68.1,63.3,62.5,59.5,58.7,62.9,61.2))

list(beta=c(0,0), b=c(0,0,0,0),
sigmaeps=0.1)
```