MARKOV CHAIN MONTE CARLO METHODS FOR REGRESSION SPLINES WITH A PENALIZED ACCEPTANCE RATIO

David Keith Stamps
M.A., Statistics, May, 1978, University of Missouri - Columbia
B.A., Mathematics, May, 1976, University of Missouri - Columbia

Submitted to the Graduate School of the

UNIVERSITY OF MISSOURI - ST. LOUIS

In partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

August, 2006

ADVISORY COMMITTEE
Haiyan Cai, Ph.D.
Chairperson

Charles Chui, Ph.D.
Ronald Dotzel, Ph.D.
Qingtang Jiang, Ph.D.

July 20, 2006

Acknowledgements

Special thanks to Dr. Cai for his tremendous support and encouragement throughtout the course of this project. I not only learned much, but enjoed the process. I look forward to our continued friendship.
Thanks also to Drs. Chui, Jiang, and Dotzel for suggestions on the project and writing.

This research is dedicated to my wife, Vicki, and my daughter, Jessica, who always believe in me more than I do myself.

Contents

1 Introduction **6**

2 Literature Review **12**
2.1 Model Selection Methodologies 13
2.2 Knot Selection Methodologies 23

3 Markov Chain Monte Carlo Methods **38**
3.1 Markov Chains 40
3.2 MCMC Algorithms 47
3.3 Gibbs Sampling 50
3.4 Metropolis-Hastings Algorithms 51

4 Spline Regression with Penalty Function **55**
4.1 Prior Distributions for Parameters 55
4.2 Penalty Function Description 57
4.3 Spline Regression 61
4.4 Hierchical Structure of the Model 66
4.5 Specification of Prior Distributions for Parameters 70
4.6 Penalty Function 75

5 MCMC Algorithm for Spline Regression **87**
5.1 Proposed Knot Move 89
5.2 Proposed Knot Addition or Deletion 92
5.3 Update of B-Spline Coefficients 100
5.4 Update of Variance Term 103
5.5 Function Estimation 105

6 Simulation Results **108**

7 Conclusion **130**

List of Figures

1 Estimates for the model (No penalty) 113
2 Estimates for the model (AIC penalty) 114
3 Estimates for the model (BIC penalty) 115
4 Estimates for the model (Normal(5,2) penalty) 116
5 Estimates for the model (Second-difference penalty) . . . 118
6 Distribution of number of knots (No penalty) 119
7 Distribution of number of knots (Normal(5,2) penalty) . . 120
8 Confidence Intervals for Estimated Curve 121
9 Estimates for Quadratic Generating Function (No Penalty) 125
10 Estimates for Quadratic Generating Function (Normal(5,2) Penalty) 126
11 Estimates for Sinusoidal Generating Function (No Penalty) 127
12 Estimates for Sinusoidal Generating Function (Normal(5,2) Penalty) 128

List of Tables

1 Summary of Results for Five-Knot Spline 122
2 Summary of Prediction Evaluation for Five-Knot Spline . . 123
3 Summary of DIC for Five-Knot Spline 123
4 Summary of Prediction Evaluation for Five-Knot Spline . . 124
5 Summary of Prediction Evaluation for Quadratic Function 125
6 Summary of Prediction Evaluation for Sinusoidal Function 129

Abstract

An increasingly popular method for fitting complex models, particularly with a hierchical structure involves the use of Markov Chain Monte Carlo simulation. Within a Bayesian framework, two major strategies for the construction of these Markov chains are prominent. These two strategies are Gibbs sampling and Metropolis-Hastings methods. Also, recent research in the area of MCMC methods has witnessed the emergence of modeling efforts which permit the movement of the chain across models of varying dimensions. Because the Markov chain, if properly constructed, converges to the joint posterior distribution of the parameters to be estimated, Bayesian averaging of the iterations in the chain, once approximate convergence has been realized, is an attractive option for producing a final estimated function. With the transdimensional methodology, this Bayesian averaging process takes place across these models of differing dimensions. The purpose of this research is to incorporate a penalty function as an integral component of the transition kernel of the Markov

Chain to impose desired constraints on the final model. The class of functions used to model data in this process are cubic splines on a finite closed interval. Furthermore, the knots for the spline function are allowed to change over the course of the Markov chain, so that the final Bayesian averaging process takes place, not only across models of varying dimensions, but also across models with differing knot locations. The primary penalty function that is investigated, in its logarithm, is a quadratic function of the number of knots, imposing larger penalties as the number of knots increases. It is shown that this penalty function actually induces a prior distribution on the number of knots which is proportional to a Normal distribution. It is also shown that this penalty function can be written as a penalized Kullback-Leibler distance measure, where the penalty is an increasing linear function of the number of knots and can be chosen in such a way to achieve a desired mean and variance of the Normal prior distribution.

However, this penalty function strategy is general and can be applied to influence the final estimation in areas such as smoothness and knot

spread. The performance of the methodology is compared with results using no penalty and standard penalty functions such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). This is done by evaluation of prediction errors for data sets which are independent of the modeling process.

1 Introduction

Linear models historically have provided a theoretical framework in which a response variable is considered to be a function of independent, or predictor variables. Sometimes these predictors are referred to as explanatory variables, as they presumably possess some plausible association with, but not necessarily causal, relationship with the response variable. In the most common approach, the mean of the response variable is assumed to be a linear combination of the predictor variables, where the coefficients of the predictors are estimated by means of a least-squares process. An individual observation is assumed to be a sampled observation from a normal distribution with mean calculated from this linear combination plus a random term, which is normally distributed with mean 0 and a constant variance across all levels of the independent variables. Thus, the expression for the linear model is:

$$y = \beta_0 + \beta_1 * x_1 + \ldots + \beta_p * x_p + \epsilon \tag{1}$$

or, in matrix form,

$$Y = X * \beta + \epsilon \tag{2}$$

The presence of the random error term in the model is a recognition of the fact that uncertainty in parameter estimates exists and that some measure of this uncertainty is assumed to be simple random noise. That concept includes the fact, however, that some of this noise is not truly random noise, but lack of fit due to real explanatory variables which have not been included in the model. Overriding the entire process, however, is uncertainty regarding the nature of the model assumptions, including misspecification of the functional relationship itself. This is nothing more than the admission that no model perfectly reflects reality. Even well-specified models may suffer from the ability to incorporate only a finite number of predictors and only a finite number of observations can be collected. Omission of useful predictors as well as inclusion of questionable ones can contribute to model misspecification. The inclusion of variables which exhibit multicollinearity is evidence of a model that is overspecified. Such problems can contribute to biased estimates of the parameters and/or poor fit to the data. Violations of statistical assumptions, such as non-normality of the error terms or non-constant variance of these er-

ror terms (heteroscedasticity), may result in similar problems, but can sometimes be treated through techniques such as data transformations. These difficulties can be diagnosed through the use of residual plots, Q-Q plots, and other tests, but application of the appropriate remedy can be a matter of the researcher's insight and experience.

Much of the appeal of this linear model stems from the ease with which estimation is possible and the nature of the statistical properties of the estimates, such as consistency and unbiasedness. Confidence intervals and hypothesis tests are also straightforward from the parametric assumptions of the model. More recent efforts have enabled models to easily accomodate non-normal error terms through the use of generalized linear models. These models connect the mean of the dependent variable to the predictors through a link function. These models have provided a convenient framework for estimation when the response variable is Binomial, Exponential, Poisson, or follows some other distribution belonging to the Exponential family of distributions.

Even greater generality and flexibility have been achieved through more

robust techniques which impose fewer constraints on the model. These approaches may attempt to address such issues as local oscillations in the response variable and the smoothness of the response surface. Various robust regression options are available for capturing the local behavior of a response function, while some type of penalty is generally imposed on the likelihood function in order to prevent parameter estimates which produce departures from smoothness in this function. A frequent approach is to penalize the square of the second derivative of the response function. In theory, then the objective is to minimize the following quantity:

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda * \int f''(x)^2 dx \tag{3}$$

It is known that the natural cubic spline is the function, f, which minimizes this quantity among the class of continuously second-differentiable functions. The use of spline functions can be particulary helpful for approximating the behavior of data without imposing assumptions which can be questionable.

The flexibility inherent in the use of splines proves to be a two-edged sword. A spline function can significantly reduce the number of predictors

in the model, but the placement of the knots within the domain of the spline argument remains an issue.

The difficulty with the use of spline functions, which is the topic of this research, is related to the uncertainty the researcher generally has, knowing not only the number of knots, where the behavior of the response variable may change, but the location of those knots. Much effort has been exerted to address the problem by developing an algorithm which performs the analysis in stages. Frequently, the number of knots and their locations are determined in the first stage and, once that is accomplished, estimation of the response function proceeds conditioned on these values. The determination of the knot locations is done so as to optimize some criterion, such as a generalized cross-validation test or possibly using some sequential process involving the addition and removal of candidate knots one at a time. The results of the models fit in this sequential process are compared before fixing the knot vector. The purpose of this research is to explore and extend recent efforts to incorporate the uncertainty which exists for both the number and location of the knots for the spline us-

ing a Bayesian averaging estimation process. The averaging of models is computed over a sample from a Markov Chain Monte Carlo (MCMC) simulation. The major contribution from this research will be to illustrate how a penalty function approach can be implemented as an integral component in the construction of an MCMC algorithm. It is clear from penalized estimation that the use of such penalty functions when applied to the likelihood function is equivalent to inducing a prior distribution on the set of models and/or parameter values which are under consideration.

2 Literature Review

The use of spline functions, while providing broad flexibility for model-building, like any other approach, is hampered by some weaknesses. Its flexibility lies in the choice of the number and placement of knots for the model, as well as the degree of the spline function. Not only can the shape of the spline be influenced by these parameters, but duplication of knots can enable the model to accomodate discontinuities in selected derivatives of the spline function, or even the actual function itself. In addition, the type of functions which comprises the basis for the overall spline function can offer certain advantages. For example, the use of a B-spline basis results in a set of basis functions whose coefficients have only local influence over the shape of the estimated curve. This allows the insertion or deletion of knots by the researcher without a resulting global change in the estimated curve. Another set of basis functions, the trunctated power functions, can provide the ability to impose constraints on the magnitudes of model coefficients which penalize departures from various features of smoothness. Additional details regarding this capabil-

ity will be discussed in this chapter.

2.1 Model Selection Methodologies

Before exploring spline functions in detail, the topic of model selection is of interest. Generally, this refers to the selection of a single model from a set of candidate models. Some criterion is optimized across the candidate models, with the selected model providing the optimal value among this set. Once the model selection decision has been made, straighforward estimation is performed for the selected model.

Some model selection methods will now be reviewed. These would include the traditional stepwise procedures in linear models which sequentially include and exclude predictors based on a straighforward F-test, conditioned on the independent variables currently assumed to be in the model. These are well known and need no discussion. In addition, some additional approaches include: training/test set analysis, generalized cross-validation, stepwise procedures, and information theory approaches. It needs to be said that when model selection is used, there is

the inevitable uncertainty regarding the correctness of the model selected. This has lead to what have termed "model-averaging" approaches. In fact, that approach is fundamental to this research.

At the most crude level, one could simply select the model with the smallest sum of squared error terms. This is unacceptable, however, because of the inherent bias in using the same data for modeling and evaluation purposes. Moreover, given a sequence of nested candidate models, the inclusion of an additional predictor will always reduce SSE.

The first approach involves the estimation of parameters from a training set of data. This method is not particularly sophisticated. Only a subset of the data which have been collected is used for the estimation of parameter values, through the methods appropriate for the model. Once these estimates are obtained, forecasted values are calculated for each of the observations in the "test" data set. This keeps the evaluation of model fit independent of parameter estimation. However, it does forfeit some of the observed data which could be used to obtain potentially more reliable estimates. If the data set is sufficiently large, it may be that a fairly

large percentage of the data points may be used for modeling purposes and still retain a reasonable amount for testing. Some measure, such as mean-squared-error of the forecasts, may be used to compare competing models.

Independent validation of parameter estimates with this type of approach provides some safeguard against overfitting. A model with many parameters which suffers from overfitting to a training data set may well perform more poorly, in comparison with competing models with fewer parameters, if some of the predictors in the larger model prove to be spurious. Rather than measuring a true effect, these spurious predictors may only be capturing unusual behavior in the function, or random noise.

A similar method for assessing the viability of a model is the use of cross-validation measures. One difference from the previous approach is that here, no observation is strictly considered to be training data or test data. If there are n observations, then consider the following:

Let $(\mathrm{x}_i, \mathrm{y}_i) = \mathrm{i}^{th}$ observation

f_{-i} = model esimated from data omitting i^{th} observation

$$\text{CV} = \sum_{i=1}^{n}(y_i - f_{-i}(x_i))^2$$

In essence, CV is a mean-squared-error type of statistic. However, in the case of each squared residual, the forecasted value for x_i is not based on data which include this observation. Values of CV can be compared for all models under consideration.

There are then finally those types of validation measures which are based on information theory. Boltzmann's concept of generalized entropy is dicussed by Akaike [1] and is revelant to physics and thermodynamics. Kullback and Leibler [2] introduced the notion of an information distance, based on this notion of entropy, which is useful for the comparison of competing models in statistics. Suppose that a set of data has been collected, having been generated from an underlying "true" model, designated by f, which is unknown. Also, consider a model specification that is under consideration, g_1. The Kullback-Leiber distance (K-L distance) from g_1 to f is defined as:

$$d_{KL} = \int \ln\left(\frac{f(x)}{g(x|\underline{\theta})}\right) f(x)dx \tag{4}$$

where $\underline{\theta}$ is the vector of parameters under model g_1 .

While f is unknown (and may only be a finite approximation to reality, with potentially an infinite number of parameters), the difference between g_1 and some other candidate model, g_2, could be written in the following manner:

$$\begin{aligned} I(f, g_1) - I(f, g_2) &= \int \ln\left(\frac{f(x)}{g_1(x|\underline{\theta})}\right) f(x)dx - \int \ln\left(\frac{f(x)}{g_2(x|\underline{\theta})}\right) f(x)dx \\ &= E_f[\log f(x) - \log g_1(x|\underline{\theta})] - E_f[\log f(x) - \log\left(g_2(x|\underline{\theta})\right] \\ &= E_f[\log\left(g_2(x|\underline{\theta})\right] - E_f \log g_1(x|\underline{\theta})] \end{aligned}$$

When actual modeling is in view, of course, $\underline{\theta}$ must be replaced by estimates calculated from the actual data. This, again, results in values for these parameters which include a measure of uncertainty and differ from the values which would actually minimize the distance function. Rather than selecting a model based on minimized actual K-L distance, it is prudent to make this determination based on expected K-L distance over the set of candidate models. Akaike [1] showed that the maximized log-likelihood is, on average, an overesetimate of this measure. Akaike

[1] showed that this bias is approximately equal to k, the number of free parameters which must be estimated in the model. Given a set of candidate models which can be of varying dimensions and possibly from different sets of statistical families of distributions, any two models can be compared. In fact, for any model, g, we can write:

$$I(f, g) = Constant - E_f[log(g(x)|\underline{\theta}))]$$

or, using estimated values, we get:

$$I(f, g) = Constant - (L(\hat{\underline{\theta}}|\underline{y})) - k$$

where L is the log-likelihood function given the data, $\underline{y}$. Because the constant is independent of the model, then the model which maximizes the penalized log-likelihood function, L - k, is the model of choice. Thus, knowledge of the true underlying model, regardless of the number of parameters is unnecessary for the sake of comparison of competing models. This is true whether the true model is a member of the candidate set or not.

It is important to realize that the bias correction produced by Akaike is an asymptotic (sample) result. Thus, to the degree that, in practice,

the sample size of the data is small, the bias term, k, must be modified. Research designed to identify and quantify this modification is due to Hurvich and Tsai [3]. As the ratio of the number of parameters estimated to sample size becomes significant, the accuracy of Taylor series expansions used in the derivation of AIC suffers. The final result shows that, for small samples, the unbiased K-L distance can be written as:

$$(L(\underline{\hat{\theta}}|\underline{y})) - ((k * n)/(n - (k + 1)))$$

This clearly simplifies to the AIC as n $\rightarrow \infty$.

It should also be noted that AIC, and other information-based model selection criteria, are fundamentally frequentist in philosophy. That is, the rationale which motivates the development of the results is that the data alone, apart from any prior belief system of the researcher, are responsible for the outcome. Bayesian approaches incorporate prior beliefs or information as part of the mathematical development.

Although not strictly motivated by the concept of information theory, the Schwarz Information Criterion (SIC) is another prominent approach to the problem of model selection. Also known as the Bayesian Informa-

tion Criterion (BIC), due to the work of Schwarz [4], this measure is an approximation to the Bayes factor commonly used when comparison of competing models is performed. The Bayes factor is nothing more than the ratio of the posterior densities of two competing models, given the observed data. An understanding of the development of the BIC starts with Bayes' Theorem:

$$\begin{aligned} P(A|B) &= \frac{P(A \bigcap B)}{P(B)} \\ &= \frac{P(B|A) * P(A)}{P(B)} \end{aligned} \tag{5}$$

In terms of continuous density functions, the theorem is:

$$\begin{aligned} f(y|x) &= \frac{f(x|y) * f(y)}{f(x)} \\ &= \frac{f(x|y) * f(y)}{\int f(x|y) * f(y) dy} \end{aligned}$$

for all continuous random variables, x and y.

The application of Bayes' Theorem to the issue of model selection proceeds as follows. Suppose there is a set, $\boldsymbol{M}$, of candidate models and a set of prior probabilities associated with each of these models. For the derivation of BIC, all candidate models are assumed to have equal prior

probability. The objective is, then, to compute the probability for each model given the observed data. Following Bayes' theorem, this posterior probability can be written as:

$$f(m|y) = \frac{f(y|m) * f(m)}{f(y)} \tag{6}$$

Because $\underline{y}$ is the realized data set, the denominator in the above expression is independent of m, so that we may write:

$$\begin{aligned} f(m|\underline{y}) &\propto f(\underline{y}|m) * f(m) \\ &= f(m) * \int f(\underline{y}|\underline{\theta}_m, m) * f(\underline{\theta}_m|m) d\underline{\theta}_m \end{aligned}$$

This makes it necessary to integrate out the nuisance parameter, $\underline{\theta}_m$, over the parameter space associated with model m. Under the assumption of equal model prior probabilities, for the purpose of comparison of any two candidate models, the leading term, f(m), can be ignored. To approximate the integral, a Taylor series expansion of f($\underline{y}|\underline{\theta}_m$), centered at the maximum-likelihood estimate, $\widetilde{\underline{\theta}}_m$, of the parameter vector, $\underline{\theta}_m$, is performed. This results in the final expression for the BIC, assuming again that the sample size is large:

$$\ln f(x,m) = L(\underline{y}|\widetilde{\underline{\theta}}_m) - ((k/2) * \ln(n)) \tag{7}$$

In terms of implementation, the resulting mathematical expressions for AIC and BIC look very similar, namely:

$$AIC = \ln(f(\underline{y}|\hat{\underline{\Theta}}) - k$$

$$BIC = \ln(f(\underline{y}|\hat{\underline{\Theta}}) - ((1/2) * \ln(n) * k)$$

The two expressions differ by a constant multiplied by the dimension of the model considered. Also, to be observed is the fact that the use of BIC results in the selection of models of smaller dimension. The choice of which penalty term to use may well be determined by the philosophy of modeling on the part of the practitioner. Or, a desire for parsimony as opposed to overfitting may drive the decision. Yet, the similarity of the final mathematical expression suggests the potential for some type of unifying understanding of the two, and also the possibility of attempting to leverage the best features of both.

2.2 Knot Selection Methodologies

The method of model selection can become increasingly complex, particulary when the added task of knot selection for the spline model is part of the process. The general methodologies which have just been described facilitate this effort, but additional tools are useful, almost essential, when the potential set of possible knot locations for a spline becomes prohibitively large. It can become computationally difficult to compute every possible model and the values of AIC and/or BIC for each.

Before exploring some of the analytical approaches to knot selection, first it is not surprising that practical considerations may not permit complete freedom in the choice of knots for the model. Using time as the spline variable, the structure of a business calendar, for example, may only allow changes in a financial model at defined points in time, such as the beginning of a new fiscal year. The depreciation in value of a financial asset, such as an automobile, may experience changing patterns as it ages. The automobile generally experiences rapid depreciation early in its product life and this rate decreases as both time and distance driven increase.

This value may experience significant change points which are, at least, in part, the result of customer perception. These change points might occur when the automobile odometer reading crosses certain thresholds, such as 10,000 miles or the value at which a warranty provision expires. The actual change in value may best be represented by a jump discontinuity, but such a jump may not be practically permissible. The terms of a lease or purchase contract may rigidly govern where change points can occur in modeling the value of an asset. A similar phenomenon could occur in the realm of the escalation of insurance premiums. Clearly, policy renewal dates and/or birthdates of the policyholder, may control when premium increases may be implemented due to changes in mortality or morbidity rates due to age. It is also true that even when thresholds such as these are enforced, the number of such thresholds for a given model may be limited. This can aid communication of the model to colleagues in the company and the industry. It can also facilitate the marketability of products to consumers, even if some goodness-of-fit is sacrificed.

These types of practical constraints may actually turn out to be a hid-

den blessing. Even though the set of candidate knots may be severely limited by such constraints, they can at the same time reduce the problem to a manageable level. While limiting the flexibility of the spline model by not allowing complete freedom of knot selection, the set of candidate knots is kept at a reasonable size and knots are likely to reside at, or close to, nice values.

Given ample freedom for placement of the spline knots, it is possible that a plot of the dependent variable versus the independent variable (underlying the spline function) may provide sufficient insight for the researcher to make informed conjectures about knot locations. The density of data at potential knots may temper this decision, particularly if the desired number of knots is small. However, often data are sufficiently ambiguous due to other predictor variables and random noise, so that the decision is unclear.

The majority of my review of the literature will focus on free-knot splines. Free-knot splines here are understood to refer to models which rely upon some data-driven methodology for the selection of the knot loca-

tions, and possibly the number of knots as well. However, some mention of regression splines and P-splines (penalized regression splines) should be made. Regression splines operate with a given set of knots and compute parameter estimates by least-squares minimization. The actual estimates will depend on the basis functions used for the spline space. Before discussing specific details regarding competing methodologies for model selection (knot selection), we will initially assume that the knots in the model have been chosen and now are fixed. The focus, at this point, is to examine the control that the practitioner has over the shape of the estimated spline function given this fixed set of knots.

Unrestricted regression splines possess certain inherent vulnerabilities, leading to potentially undesirable characteristics of the estimated function. If the fixed set of knots is small, fit to the data may be sacrificed. If a dense set of knots is used, the risk of overfitting and lack of control of the total variation in this function exist. Overfitting to a single data set can also result in a poor fit to a second data set from the same underlying population. A popular mediating solution is to select a dense set

of knots and impose a penalty on the spline coefficients which prevents serious departure from some measure of smoothness.

Some of the primary work for fitting splines to data with an underlying penalty function is due to Eilers and Marx [5]. Generally, P-splines place knots at a large number of locations, perhaps uniformly spaced over the spline argument, or at uniformly spaced quantiles of this variable. Or, knots can be placed at actual values of the spline argument found in the data. The approach of Eilers and Marx is to employ a set of B-splines as basis functions for the spline, along with a collection of difference penalties on the estimated coefficients of these B-splines. Work done by O'Sullivan [6] had effectively done this by defining the following objective function:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{k}\hat{c_j} * \beta_j(t_i))^2 + \lambda \int \sum_{j=1}^{k}(\hat{c_j} * \beta_j''(t))^2 dt \tag{8}$$

Eilers and Marx constructed a modified penalty function, using differences (of unspecified order) of estimated B-spline coefficients:

$$S = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k}(\hat{c_j} * \beta_j(t_i))^2 + \lambda * \sum_{l=k+1}^{n}(\Delta^k \hat{c_l})^2 \tag{9}$$

where k is some higher-order finite difference on the B-spline coefficients. The sum of these differences should provide a good approximation to the

integral which serves as the penalty in O'Sullivan's work. Clearly, it is the intuitive discrete analogue of the integral. Notice that Eilers and Marx extended the penalty function beyond the work of O'Sullivan to allow the inclusion of any order of differences in the penalty. The question of the value of λ, the smoothing parameter, which should be used, is part of the discussion in Eilers and Marx. When λ=0, the problem reduces to the ordinary least-squares problem. As λ becomes very large, the estimated function approaches a polynomial of degreee k-1. The recommendation of Eilers and Marx is the use of cross-validation statistics for choosing the optimal value of the smoothing parameter.

A worthwhile achievement in this approach is the ease with which the penalty function can be incorporated into the traditional least-squares equations. If D_k represents a k x k matrix of penalties, then the penalized least-squares system is:

$$X'y = (X'X + \lambda D_k' D_k)\beta \tag{10}$$

Eilers and Marx demonstrate that the penalty function used by O'Sullivan (1986, 1988), although similar to their own, produces a more complex set

of equations to solve in order to minimize the objective function.

Ruppert and Carroll [7] extended the idea of penalty splines to include the notion of a spatially adaptive penalty function. Their research, like that of Eilers and Marx, fits a spline of some fixed degree, say k, but with a different set of basis functions spanning the same spline space. In their approach, these basis functions are composed of monomials $\{x^0, x^1, x^2, \ldots, x^k\}$, plus a set of "truncated power functions," of the form $\{(x\text{-}t_0)_+^k, (x\text{-}t_1)_+^k, \ldots, (x\text{-}t_j)_+^k\}$. Again, the set of knots is fixed, but is a smaller set than that used by Eilers and Marx. Eilers and Marx chose a set of equally-spaced knots, but Ruppert and Carroll employ a smaller set of knots at equally-spaced quantiles of the spline variable. While the number of knots for the spatially adaptive method is fixed, Ruppert and Carroll do recommend an algorithm for their selection prior to the modeling process.

As opposed to the global penalty approach, Ruppert and Carroll propose a penalty function which is spatially heterogeneous, permitting the penalty, and therefore, the smoothness, of the estimated spline to vary

across the knots. They suggest that a linear spline be fit to the natural logarithm of the penalty function at a subset of the knots. So, a subset, say m, of the total set of j knots, is selected to serve as knots (change-points) for this penalty function. Thus, the estimated function looks as follows:

$$f(x) = \beta_0 + \beta_1 * x + \beta_2 * x^2 + \ldots + \beta_k * x^k + \sum_{i=1}^{j} \alpha(t_i) * \beta_{k+i} * (x - t_i)_+^k \quad (11)$$

where the estimates are such that the objective function,

$$S(\beta) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \sum_{i=1}^{j} \alpha(t_i) * \beta_{k+i}^2$$

is minimized.

The advantage in defining the spanning functions to include the truncated power functions lies in the freedom to construct the penalty function, α, so that changes are initiated at each of the designated knots. In the final analysis, the approach advocated by Ruppert and Carroll grants the model the ability to accomodate heterogeneity in the variability of the response function. If the function is known to oscillate more rapidly in certain ranges of the spline, this is recognized by reducing the penalty function in this local neighborhood.

The estimation procedure assumes that the design matrix is defined in the following manner:

$$X_i = i^{th} row$$

$$= \{1, x_i, \ldots, x_i^k, (x_i - t_1)_+^k, \ldots, (x_i - t_j)_+^k\}$$

and we have a penalty matrix which is a diagonal matrix with zeroes comprising the first k+1 diagonal elements and the remaining j diagonal entries are equal to $\alpha(t_i)$, i=1,2, ...,j. The estimator of the parameter vector, β, then is:

$$\beta(\alpha) = (X'X + D(\alpha))^{-1} * X'Y \tag{12}$$

where α is selected by a generalized cross-validation (GCV) criterion.

Shifting attention to knot selection, a variety of ideas have been entertained. One of the simplest of these can be attributed to Friedman and Silverman [8]. The method discussed here is one which fits piecewise linear functions to data. Without addressing the advisability of this family of candidate models, the knot selection procedure is much like many other stepwise selection methods.

A sequence of knot selection decisions is performed, choosing the knot

at each step which minimizes the average squared residual (ASR). In actual practice, the set of candidate knots must be limited to some finite set. Otherwise, the process of selection cannot be completed. The authors limit this set to the realized values of the spline variable in the data. The main purpose is to allow adequate flexibility in the linear spline where the data points are dense. The sequence proceeds by placing the first knot at the candidate knot which minimizes ASR. Continuing in this manner, an additional knot is added at each step which minimizes this same criterion, assuming that previously selected knots are kept in the model. At the end of the process, that model chosen at one of the steps in the process which minimizes a generalized cross-validation (GCV) statistic is chosen as the final model. The ordinary "one-at-a-time" cross-validation measure is computed by averaging the squared-error (residual) for the i^{th} observation based on the remaining n-1 sample points. This can be written as:

$$CV = \frac{1}{n} * \sum_{i}^{n} (y_i - f_i(x_i))^2$$

or,

$$CV = \frac{1}{n} * \sum_{i}^{n} \frac{(y_i - \overline{y}_i(x_i))^2}{(1 - h_{\lambda_i})^2}$$

Here, λ_i represents the ith diagonal element of the smoother matrix, H, as defined by $\overline{y} = \text{Hy}$. The GCV statistic is a computationally advantageous generalization of this concept which replaces λ_i by its average value.

Another stepwise methodology is discussed by Stone, Hansen, Koopenberg, and Truong (1995) in the context of extended linear models. The procedure is more complicated than that of Friedman and Silverman. The underlying concept is to add knots sequentially from a minimum number of knots until some prescribed maximum is reached, and then deleting a knot step-by-step until the original minimum value is reached. During the addition steps, a Rao statistic is employed for decision-making, while a Wald statistic is the measure used for knot deletion decisions. At any stage of the addition phase of the process, suppose the current set of knot subintervals is (assuming left and right endpoint of a and b respectively):

$$\{ (a, t_1), (t_1, t_2), (t_2, t_3), \ldots, (t_{k-1}, t_k), (t_k, b) \}$$

and defining potential knots for inclusion at the quartiles in these subintervals, a Rao statistic identifies an optimal knot within each subinterval. Recall that, in general, the Rao and Wald statistics are defined in the following manner:

Let $\underline{\theta}$ be a parameter vector to be estimated. The score function is defined as:

$$U_i(\theta) = \frac{\partial}{\partial \theta_i}\left(\ln L(\theta_i|\underline{y}) \right) \tag{13}$$

where L is the likelihood function and $\underline{y}$ the observed data. Also, let

$$I_{ij} = -E_{\underline{\theta}}\left[\frac{\partial^2}{\partial \theta_i \theta_j} \ln (\underline{\theta}|y) \right]$$

be the information matrix. Then the Rao statistic is defined as:

$$R = U(\underline{\theta}_0)' * I(\underline{\theta}_0)^{-1} * U(\underline{\theta}_0) \tag{14}$$

And the Wald statistic is defined as:

$$W = (\tilde{\underline{\theta}} - \underline{\theta}_0)' * I(\tilde{\underline{\theta}}) * (\tilde{\underline{\theta}} - \underline{\theta}_0) \tag{15}$$

where $\underline{\theta}_0$ is a set of values for the parameter vector which are to be tested and $\underline{\tilde{\theta}}$ is the set of actual parameter estimates calculated from the data.

As the knot deletion stage progresses, at each step, the least significant current knot is deleted, using the Wald statistic. This is repeated until only three knots are left. At the end of the entire procedure, the models which were selected at each step are then compared according to the AIC criterion. After this selection, some additional refinement selects a final new candidate knot from each subinterval. One of the candidates is then selected as the new knot and the dimension of the model is increased by one.

A different philosophy of knot selection undergirds the work done by Lindstrom [9]. Rather than relying upon model fit as the criterion for selection, certain constraints are actually placed upon the knot vector itself. Lindstrom's research builds upon the research done by Jupp [10], where he addressed what he termed the problem of "lethargy" in free-knot spline modeling. Lethary refers to the tendency, when the knots are parameters to be estimated, for model-fitting algorithms, to become

trapped in some local neighborhood of the multi-dimensional space of possible knot locations. That is, if the number of knots is fixed at some value, say k, then a k-dimensional cube exists for knot values. When lethargy occurs, there may be some k-dimensional neighborhood, possibly on or near an edge of the simplex, where a local optimal point exists. The algorithm may search in this neighborhood for a solution and be unable to escape. This local optimal point may not be the global solution and the algorithm finally fails to locate the proper overall knot vector which is optimal.

Solution to the lethargy problem is achieved by Jupp through the use of a transformation on the knot vector. That transformation is defined in the following manner:

$$\text{Let a} = \gamma_0 <= \gamma_1 <= \gamma_2 <=, \ldots, \gamma_k <= \gamma_{k+1} = b$$

and define:

$$\text{h}_i = \frac{(\gamma_i - \gamma_{i-1})}{b-a} \quad \text{i=1, } \ldots \text{ ,k}$$

The basic idea is to penalize knot vectors where knots coalesce, either by duplication or by their close proximity. The estimator of the spline

coefficients is that set of values, $c_1, \ldots, c_k$, which minimizes the penalized residual sum-of-squares, where the penalty function has the form:

$$J = \left(\frac{(p-1)}{[\ln(P(\gamma^0(k))]} \right) * \ln(P(\gamma)) + 1$$

$$\text{where } \ln[P(\gamma)] \text{ - } \sum_{i=1}^{k+1} \ln((k+1) * h_i)$$

and p, γ^0(k) are constants set by the practitioner.

3 Markov Chain Monte Carlo Methods

Undergirding the methodology that is central to this research is the mathematics of Bayesian analysis. As previously mentioned, the calculation of Bayesian probabilities and density functions is the well-known elementary result from probability theory, Bayes' theorem (5).

Bayesian analysis, generally, allows the practitioner to incorporate his/her own prior beliefs, if any, regarding true parameter values. In fact, by specifying an appropriately defined prior distribution for the parameters in the model, some of the constrained estimation techniques discussed earlier can be achieved through a Bayesian approach. One which will be discussed is the penalizing of spline coefficients to prevent overfitting. Of great significance to this research is the growing recent interest and development of modeling approaches, using creative tools, called Markov Chain Monte Carlo (MCMC) methods. These methods can be highly imaginative and have the advantage of being capable of providing estimates for parameters which arise from very complex model specifications. This includes many cases where the parameters which are defined by the model follow some

type of hierchical structure. A particular strength of these MCMC methods is the fact that, when desired, they are also capable of recognizing the uncertainty that accompanies model specification itself. A primary way of incorporating this uncertainty in the final parameter estimates is the use of Bayesian model averaging. Model averaging, not unique to Bayesian analysis, is the calculation of a parameter's final estimate by averaging the estimate of that parameter over a number of models which differ in specification. The Bayesian averaging [17] which is undertaken is the averaging, not of parameter estimates, but of estimated function values, where the estimated function is averaged over models with differing sets of spline knots. Greater attention to the details of this modeling concept will be given later.

3.1 Markov Chains

Before introducing the present research, a review of the necessary mechanism for implementation of the modeling process, Markov chains, should be done. Markov chains are defined by a stochastic process in which a sequence, indexed by the positive integers (or indexed by an uncountable set, in some cases) of random variables has a particular property of dependence. It will suffice for our research to assume that the Markov chains of interest can be considered to be sequences of random variables. Specifically, define this sequence of random variables as:

$$\{\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \dots\}$$

The dependent relationship which exists among these random variables can be written as:

$$f(X_k|X_1, X_2, X_{k-1}) = f(X_k|X_{k-1}) \tag{16}$$

Thus, the conditional density of any particular random variable, X_k, in the sequence, given previous history of the chain only depends on the value of the previous one, X_{k-1}. So, the history of the sequence prior to X_{k-1} has no impact on the density of X_k. One may have a time series

which exhibits Markov Chain properties. A simple example of a Markov Chain is a random walk. Here, a person may be positioned at the origin of the number line at the start of the chain and then flip a coin. If a head is the outcome, the person steps one unit in the positive direction to +1. If tails, the person steps in the opposite direction to -1. The process continues with the same coin flip performed at each stage and the person advancing or retreating one step from the present position on the number line. Clearly, the probability function governing the position at the next step in the chain depends only on the present position, no matter the sequence of steps which preceded the arrival of the process at its current state.

Associated with any Markov chain is the set of outcomes which may result at any stage in the chain. This set of outcomes is known as the state space, S . It may consist of a finite, countable, or uncountable number of points. Thus, as the chain progresses from any step to the next, there is a transition from one state to another. Consequently, given the state space, there exists a probability density governing the likelihood that the

chain moves from any state to any other state. Although the state space may be discrete or continuous, for the sake of simplicity, it will currently be assumed to be discrete, possibly countable. It is customary to define a transition matrix, P, whose elements are these transition probabilities. Specifically, the ijth element of the matrix, p_{ij}, defines the probability that the chain moves to state j in the next step given that the chain is currently in state i. Clearly, any element on the diagonal represents the probability that the chain will remain in its current state.

There are some obvious conditions that must hold for a matrix P to function as a transition matrix. The sum of the entries in each row must be equal to 1. These is simply the sum of the transitional probabilities to the next state, given the current state of the chain. In addition, given the property of dependency that defines a Markov chain, and assumming the chain is homogeneous (that the transition matrix, P, is invariant over the chain), it is a straightforward algebraic exercise to show that given the current state, i, of the chain, the probability that the chain lands in state j after exactly n iterations of the chain is: $\quad P^n = \sum_z P^k * P^{n-k.}$

for $0 \leq \mathrm{k} \leq \mathrm{n}$. These are known as the Chapman-Kolgomorov equations.

With this background, states of the Markov chain can be designated by such titles as recurrent, positive recurrent, transient, absorbing, all of which embody the likelihood that the chain will return from its current state, i, to state i at some later iteration of the sequence. These properties have much to say regarding convergence properties of the chain. Of primary interest will be whether the chain converges to a limiting distribution vector. A distribution vector of the Markov chain is a vector of probabilities that the chain is in each of its possible states at a point in the sequence. To be more clear, suppose the state space of a Markov chain is defined by: $S = \{\mathrm{s}_1, \mathrm{s}_2, \ldots, \mathrm{s}_m\}$, as an example. Then, define the distribution vector, $\mathrm{p}_k = \{\mathrm{p}_{k,1}, \mathrm{p}_{k,2}, \ldots, \mathrm{p}_{k,m}\}$, to be the vector of probabilities that the chain is in each of the m possible states at the k^{th} iteration of the chain. If a limiting distribution exists for the Markov chain, designation by π, then we have:

$$\lim_{i \to \infty} x_i = \pi = (\pi_1, \pi_2, \ldots, \pi_m)$$

To begin, if i and j are two states of a Markov chain, and there exists some integer, k, such that $p_{ij}^k > 0$ there also exists l such that $p_{ji}^l > 0$, then these two states are said to communicate. If all pairs of states in S communicate, the state space, and the chain, are said to be irreducible. Also, a state, y, of the state space, is called recurrent, if starting in y, then the probability that the chain returns to state y in the future, ρ_{yy}, is equal to 1. If the expected time, $E(T_y)$, in number of iterations, until the state revisits state y has finite expected value, additionally, the state is called positive recurrent. If the mean of this return time is unbounded, then the state is called null recurrent. A state which has a positive probability that the chain never returns to that state is called a transient state. Summarizing, states are classified as:

1) ρ_{yy}	$=$	1	Recurrent
2) $E(T_y)$	$<$	∞	Positive Recurrent
3) $E(T_y)$	$=$	∞	Null Recurrent
4) ρ_{yy}	$<$	1	Transient
5) p_{yy}	$=$	1	Absorbing

As we come to the issue of limiting distributions for Markov chains, one more important property needs to be discussed. This is the notion of periodicity. The period of a state, $y \ \epsilon \ S$, is defined as the greatest common divisor of the following set: [11] :

$$\mathrm{d}_y \quad = \quad \{n \geq 1 : \ \ P^n(y,y) \ \ > 0\}$$

If $\mathrm{d}_y = 1$, then the state y is said to be aperiodic. If two states communicate, then they have the same period. Thus, if a Markov chain is irreducible, all of its state have the same period. In this case, also, if an individual state is aperiodic, then the entire chain is said to be aperiodic. An aperiodic chain which has all positive recurrent states is called ergodic. Having established some terminology, then the following limiting theorem, found in Gamerman [11], is valid:

If a Markov chain is irreducible, positive recurrent, and aperiodic, then:

$$\lim_{n\to\infty} P^n(x,y) = \pi(y) \quad \forall x, y \ \ \epsilon S \tag{17}$$

Finally, it is often desirable that the steps of a Markov chain satisfy a special condition known as the detailed balance equation. This condition is instrumental in the strategic construction of a type of MCMC

algorithm (Metropolis-Hastings algorithms [12]). This equation, which is fundamental to this research, is:

$$\pi(x) \times P(x, y) = \pi(y) \times P(y, x) \tag{18}$$

The objective of Monte Carlo Markov Chains is to construct a procedure which simulates a chain whose limiting distribution, π, is the posterior distribution of the parameter vector. Although the discussion thus far has assumed that a discrete state space is in view, with the corresponding transition matrix, P, the treatment from now on will be concerned with continuous state spaces (the possible values for the set of parameters). With this in mind, the transition matrix is replaced with the concept of a transition kernel, P, which is a probability density function for proposed new parameter values given their current values. The goal is now to design, for a given modeling effort, an algorithm which has as its limiting distribution, π, and then as the n $\rightarrow \infty$ (the number of chain iterations), then the steps in the chain can be assumed to approximate a random sample from π. This permits the practitioner to simulate a random sample from extremely complex, multi-dimensional parameter spaces, where

the necessary analytical techniques may well be intractible.

Again, Markov chains which satisfy the detailed balance equation for some probability distribution, π, prove to be extremely useful. Of primary interest for the current research is the fact that when this property holds for an irreducible chain, then that chain is postive recurrent, and has limiting distribution, π (3.1). By summing both sides of this equation over all states, x, we obtain:

$$\sum_x \pi(x) * P(x,y) = \sum_x \pi(y) * P(y,x) = \pi(y) \tag{19}$$

which defines a stationary distribution, π, for the chain.

3.2 MCMC Algorithms

The need for Markov Chain Monte Carlo (MCMC) schemes is driven by the inability in many situations to draw random samples directly from the probability distribution of a set of parameters. This is often due to the multi-dimensional and/or hierchical structure of parameter vectors in a modeling problem, making the necessary evaluation of multidimensional

integrals intractible. As has been stated previously, Bayesian analysis presumes that a parameter vector, $\underline{\theta}$, is to be estimated, and that a prior distribution exists for this vector, and that the observed data, $\underline{y}$, are available. Restating the posterior density, by Bayes' theorem:

$$\frac{f(\underline{\theta}|\underline{y})f(\underline{\theta})}{\int \cdots \int f(\underline{y}|\underline{\theta}) * f(\underline{\theta})d\underline{\theta}} \tag{20}$$

It is clear that the calculation of the denominator, except under the most simple probability distributions, will be prohibitive, even numerically.

At times, this problem may be avoided when the prior distribution for $\underline{\theta}$ and the distribution from which are data are selected, are conjugate. The concept of conjugacy is that the prior distribution for the parameter vector, $\underline{\theta}$, and the data are related mathematically in such a way that the posterior distribution for $\underline{\theta}$ belongs to the same family as the prior. This may allow straightforward estimation of the parameters.

For example, suppose that we wish to estimate the mean, p, of a Bernoulli process. A random sample of n observations is selected and the mean of the set of 0's and 1's is calculated. In Bayesian analysis, it

is commonly the case that a prior distribution for p is defined from the family of Beta distributions. Here, the random variable, p, is assumed to have the following probability density function (pdf):

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)*\Gamma(\beta)} * p^{(\alpha-1)} * (1-p)^{(\beta-1)} \quad 0 < p < 1 \tag{21}$$

If Y represents the number of successess observed in n sampled observations of the Bernoulli process, then using the Binomial distribution:

$$P(Y=y) = \binom{n}{y} * p^y * (1-p)^{n-y} \quad y = 0, 1, \ldots, n \tag{22}$$

The resulting product of f(y) and f(y|θ) results in:

$$f(p|y) = \binom{n}{y} * p^y * (1-p)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)*\Gamma(\beta)} * p^{y+\alpha-1} \times (1-p)^{n-y-\beta-1}$$

$$\Rightarrow f(\theta|y) \propto p^{y+\alpha-1} \times (1-p)^{n-y-\beta-1} \tag{23}$$

Observation of the form of this expression indicates that it is proportional to a Beta density, like the prior, but with parameters y+α, and n-y-β. This is characteristic of Bayesian analysis where the complex denominator of the posterior distribution, which is constant for a given sample, need not be evaluated.

More closely allied with the spline problem is the conjugacy relationship

which often exists in regression modeling in a Bayesian framework. The parameters of interest, the regression coefficients and the scale parameters, are often assumed to follow a normal and Gamma prior distribution respectively. Here, however, under the assumption of normal error terms, the joint posterior of these parameters is not conjugate with the normal error terms. The conditional distribution of each parameter, given the other, however, does have this type of conjugacy. Discussion of this conjugacy can be found in Gamerman [11]. This type of conjugacy is important for some of the MCMC theory, particulary the Gibbs sampling methodology.

3.3 Gibbs Sampling

Gibbs sampling provides one of the major approaches for implementing MCMC simulations. Here, it is assumed that we wish to sample $\pi(\underline{\theta})$, where $\underline{\theta}$ is a vector of parameters, $\underline{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$. Each θ_i may be a vector or a scalar. If we also define θ_{-i} as the vector which is identical to $\underline{\theta}$, but excludes θ_i, then consider all the full conditional distributions,

$\pi(\theta_i|\theta_{-i})$ to be known and that is possible to randomly sample from each of these. This is an advantage since it is more than likely impossible to sample directly from the joint density of $\underline{\theta}$.

Gibbs sampling consists of sampling sequentially from each of the k conditional distributions, updating θ_i at the ith step in the sequence. After a sufficient number of iterations of this sequence, then it may be assumed that the result at the end of a sequence is a single random sample from the joint distribution of the k parameters. It is recognized that this conclusion is approximately true and convergence of the chain needs to be monitored. The applicability of the Gibbs sampling approach to the Bayesian problem should be evident.

3.4 Metropolis-Hastings Algorithms

An innovative method for constructing chains for MCMC estimation is the class of Metropolis-Hastings algorithms [12] [13]. In these methodologies, the posterior density of the parameters for the model is coupled with a proposal distribution, q, to provide the machinery to implement

the simulation. If we assume that the current state of the chain (current parameter estimates) is $\underline{\theta}$, the proposal distribution is a probability distribution which governs the potential movement of the Markov chain in the next step to a new set of parameter estimates, $\underline{\theta}'$. What is interesting about Metropolis-Hastings algorithms is the ability to tailor, to some degree, the form of the proposal distribution, and the fact that the proposed set of new values for the parameter estimates may, or may not, be accepted as the new values. That is, there is a positive probability that the chain will remain in its current state, when the proposed new values are not accepted. Instead, they are accepted according to the value of a certain acceptance ratio. The acceptance probability is defined as:

$$\alpha(\underline{\theta}, \underline{\theta}^{'}) = min\left(1, \frac{f(\underline{\theta}^{'})|\underline{y}) * q(\underline{\theta}^{'}, \underline{\theta})}{f(\underline{\theta})|\underline{y}) * q(\underline{\theta}, \underline{\theta}^{'})}\right) \tag{24}$$

where q($\underline{\theta}$,$\underline{\theta}^{'}$) represents the proposal density of moving to state $\underline{\theta}'$, given that the current state is $\underline{\theta}$.

In this manner, the product of the proposal density, q($\underline{\theta}$,$\underline{\theta}'$) and $\pi(\underline{\theta},\underline{\theta}')$, functions as the transition kernel (transition matrix in the purely discrete

case). Or, writing

$$P(\underline{\theta}, \underline{\theta'}) = q(\underline{\theta}, \underline{\theta'}) * \alpha(\underline{\theta}, \underline{\theta'})$$

the detailed balance equation is satisfied, with:

$$\pi(\underline{\theta}) * P(\underline{\theta}, \underline{\theta'}) = \pi(\underline{\theta'}) * P(\underline{\theta}, \underline{\theta'})$$

with f($\underline{\theta}$) as the limiting distribtuion, π.

These types of Metropolis-Hastings schemes can be quite flexible, allowing the practitioner to construct the proposal density, q, to achieve efficiency. In addition, capability exists to construct algorithms in which a Gibbs-style update of parameters, where a subset of the parameters is updated at any given iteration, making the specification of the proposal much simpler. The only requirement that must be met for the proposal density, because the detailed balance equation is satisfied, is irreducibility, which essentially implies that the entire parameter space (multi-dimensional) must be capable of being scanned over the life of the chain.

Further work in Metropolis-Hastings proceudres has been done by Green [14]. Green extended the concept of these algorithms to transdimensional

parameter spaces. In this context, the number of parameters in the model is not fixed, even while the chain is progressing; so that, transitions in the chain potentially involve movements from a parameter space with k_1 parameters to one with k_2 parameters, where $k_1 \neq k_2$. This occurs when the proposal density actually involves the addition or deletion of parameters in a transition.

4 Spline Regression with Penalty Function

Free-knot splines, as previously defined, refer to that class of spline functions which serve as models, and for which an algorithm based on the data selects the number of knots and their locations. Several of the methods discussed have as their goal the selection of a single model based on a single set of knots. The spline function is then fit, generally by least-squares, using an appropriate set of basis functions using these knots. Some attention has been given to the notion of model averaging, where a single model is not in view, but this will be discussed in much greater detail in our research. Model averaging is central to this research because the set of knots which is used in the estimation process changes over the course of the Markov Chain Monte Carlo simulation which will be implemented.

4.1 Prior Distributions for Parameters

As previously discussed, the Bayesian framework for this research requires the specification of a set of priors for the set of parameters which accom-

pany the model. The model which will be employed is hierchical, particularly due to the fact that the number of knots which defines the model at any specific iteration in the Markov Chain governs the number of knot locations and the number of B-spline coefficients (all of which collectively, together with the intercept term and the variance term comprise the parameter vector). This actually results in some prior distributions in the model which are conditional.

The prior distributions for various parameters can take any form, but the practitioner may have insight which make practical sense (satisfying real-life constraints). However, certain statistical distributions, because of their functional form and/or conjugacy properties, tend to be prominent in the literature and in practice. One of the beauties of the MCMC methodology, in which exact evaluation of posterior densities is not required, is that the range of options for these priors is essentially limitless. One can specify whatever functional form for these priors which is desirable, whether the result is a commonly recognizable statistical distribution or not. One can also specify a prior distribution, only up to a constant,

if the necessary integration to force the total probability to sum to 1 is intractable.

4.2 Penalty Function Description

While the application of a penalty function to model selection and modification is a common technique is parametric modeling approaches, it may seem foreign to MCMC methodology. It is possible to penalize departures from smoothness, as is done in parametric modeling, by implementing appropriate prior distributions on the model coefficients. However, the same can be done by penalizing the likelihood function as an integral part of the MCMC algorithm at each step in the chain. It turns out that the use of a penalty function in the acceptance ratio for the Metropolis-Hastings procedure induces an assumed prior distribution on the parameter set (or a subset of these parameters). What is an advantage in this alternative method is the ability to customize the penalty without knowing the closed form of the prior distribution which is induced by this penalty function. It is my intention to investigate one specific penalty function to show that it

naturally leads to a common prior distribution for the number of knots in the model, but I will also address some other options and illustrate some results from these briefly. This demonstrates that the penalty function methodology of this research achieves a flexibility, by allowing the user to impose whatever constraints on the parameters that may be desirable, without forcing the user to resort necessarily to the standard set of prior distributions in practice.

The unique aspect of our work is the incorporation of a penalty function which is implemented as a component in the acceptance ratio of a Metropolis-Hastings procedure. A discrete prior distribution for the number of knots, which is proportional to a Normal density is the focus of this research. The mean and variance of this Normal prior density can be selected to influence the final estimated spline curve. It will be shown that this prior distribution is equivalent to using a penalized version of the likelihood function (or, equivalently, a penalized acceptance ratio) which is a linear combination of the well-known penalties, AIC [1] and BIC [4]. In fact, a normal prior (with the appropriate mean and variance),

it will be demonstrated, can always be written in a manner equivalent to this penalty function (up to a proportionality constant) The penalty function varies according to the number of knots in each candidate model. These two penalty functions represent two broad, contrasting philosophies regarding model selection. The AIC penalty represents a frequentist perspective, while BIC incorporates Bayesian concepts. The AIC is actually based on expected Kullback-Leibler (K-L) distance and BIC is a measure of the posterior probability that a given candidate model is correct under the assumption that all candidate models are equally likely. The properties possessed by this prior distribution will also be of interest. Also to be discussed will be the implications for the model fit and a comparison with results obtain without a penalty and the use of AIC only and BIC separately as penalty functions.

Splines provide a great deal of flexibility, depending on the priorities of the modeler. Greater fidelity to the observed data can always be achieved by increasing the number of knots, but at the expense of a less parsimonious model specification and potential poor fit to a new set of data

drawn from the same underlying distribution. The model is simply guilty of overfitting and adjusting to what is, in reality, noise. This problem may be alleviated through the use of a penalty function of the coefficients to ensure a more smooth function. Fewer knots will yield a poorer fit, but is clearly easier to communicate. Even discontinuities in the function, or some order of derivatives of the function, is possible through knot duplication. Knot duplication is an extremely promising area of research toward which we anticipate making some future contributions. Note that it is possible to approximate splines of lower order with our present algorithm by specifying knots which are very close to each other, but this is, of course, different from exact duplication. Some minor modifications in the MCMC algorithm contained in this research would be required to successfully implement a procedure which allows actual duplication of knot locations.

The thrust of this research will be to achieve some measure of compromise between parsimony and goodness-of-fit. The objective will be to arrive at an estimated curve which lies between the one bounded by the

curves which would be obtained by use of the AIC and BIC individually in the same MCMC algorithm. The curve will inherit some of the strengths and drawbacks of both approaches, but will serve well in numerous contexts.

4.3 Spline Regression

The problem that I will consider can be described in the following manner. Suppose that a random sample of observations is available, y = {y_1, y_2, . . . , y_n}, where y_i can be expressed as a smooth function of an independent variable, t:

$$y_i = f(t_i) + \epsilon_i \tag{25}$$

Here, the final term is the usual error term, ϵ_i, for a general linear model. In the current research model, it will be assumed that: $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i, ϵ_j are independent for i≠j. It may be useful to think of the variable, t, as a measure of time, thus making the function relationship a time-series model. The knots in the model be viewed as discrete time points where change in the series may be occurring.

The degree of spline which will be employed is the cubic spline, a common choice for this type of modeling. A cubic spline is a continuous piecewise polymonial, where the cubic polynomial is unique between each pair of contiguous knots. It will be useful for our purpose to define, for a given knot vector: $\underline{t}$: a=$t_0 < t_1 < \ldots < t_k$=b, the space, $S_{\underline{t},m} = \{f(t) \in C^{m-2}[a,b]: f_{|[t_{i-1},t_i]}\}$. $C^{m-2}[a,b]$ refers to the subset of all splines on [a,b] which are (m-2)-times continously differentiable at each point in this closed interval. As we will be considering cubic splines (m=4), this refers to those cubic splines with continuous second derivatives on [a,b]. Because of the closed interval, only continuity of the right derivatives at a, and the left derivatives at b is required. This follows the definition as outlined in Chui [15].

Interpolating splines are of some interest as background here. Given a set of data:

$$\{(y_1, t_1), (y_2, t_2), \ldots, (y_n, t_n)\}, \tag{26}$$

a spline is fit to the data using the values of t from the data as knots, and forcing the fitted spline to agree with the actual data at these knots.

Interpolation alone is not sufficient to ensure a unique spline for this purpose. So, the knot vector is extended, by duplicating the left and right endpoints m-1 times. So, the extended knot vector would appear as:

$$\underline{t} : a = t_{-m+1} = \ldots t_0 < t_1 < \ldots < t_n = \ldots = t_{n+m-1} = b$$

If these splines are cubic, this results in the ordinary cubic spline fit. Without an extension of the knot vector, a unique result can be achieved by imposing restrictions on the derivatives in the first and last subintervals of the closed interval [a,b].

Starting with this background, the following functional relationship will be assumed in the model for our research:

$$y = f(t)$$

where:

$$f(t) = c_0 + \sum_{i=1}^{k+4} B_i(t),$$

where B_i(t) (i = 1, 2, . . . , k+4) will be the basis functions employed in the model. This is done by using B-splines as these basis functions, where B_i

is the B-spline identified with the (i-3)rd knot and B_1, B_2, and B_3 will refer to those B-splines which are associated with the extended knot vector at the left endpoint, a. This is in keeping with the extended knot vector definition described above for the case of cubic splines. Although there are only k interior knots, instead of k+3, there k+4 B-spline coefficients because of the default knot at the left endpoint, a.

A description of B-spline functions on a closed interval can be found in Chui [15]. One of the primary advantages of these basis functions for our problem is the ease with which the MCMC algorithm can be constructed to define proposal transitions from the current state of the model to the next involving either the addition or deletion of a knot. This results in a change not only in the number of basis functions, but their interpretation for our research. The key advantage of these basis functions that will aid our effort is the property that these B-spline functions have compact support. In this modeling effort, this means that each B-spline function impacts only a subset of the fixed interval [a,b]. The support of B-spline function, B_i, will be $[t_{i-3}, t_i]$. So, $B_i(t)=0$ for $t \notin [t_{i-3}, t_i]$.

Thus, in the execution of the MCMC simulation, adding or deleting a knot can be done without affecting the current estimated function, f(t), outside this range. This results in the changing of only a few of the estimated coefficients of the B-splines in the model. It should be noted that special treatment is necessary for B-splines at the endpoints (Chui) [15].

The research contained in this paper, as previously indicated, makes no attempt to select an optimal set of knots for the estimated cubic spline. Instead, as the MCMC simulation progresses, various iterations of the Markov chain may reside in a parameter state with different knots, and even differing numbers of knots. The MCMC algorithm continues executing and eventually reaches, approximately, the limiting distribution, for the chain, which in this problem is the posterior density of the complete set of parameters to be estimated in the model. The set of parameters consists of the number of knots, the knot locations given this number of knots, the coefficients of the B-splines used in the model, and the individual observation-level variance term. Thus, the limiting distribution actually consists of a transdimensional space of parameters. Once it has

been determined that the chain has reached equilibrium, then the parameter estimates at each iteration beyond this point can be considered to be an approximate random sample from the limiting distribution. The final estimated function is actually a Bayesian averaging process of the estimated functional value at each point t in the interval [a,b]. This is a crucial point to remember because as the various iterations are drawn from parameter spaces of differing dimensions, the actual interpretation of parameters to be estimated can be different. So, we will finally not be interested in any type of estimated value of a particular B-spline coefficient, for example, because the estimation of these coefficients and consequently the estimation of f itself is not based on a fixed set of knot locations. The notion of Bayesian model averaging is discussed by Draper [17].

4.4 Hierchical Structure of the Model

The major categories of MCMC algorithms have been described in the review of the current literature. Much of the work that will be done in this paper is an extension of work done by Biller [16]. The methodology will be

to construct a Metropolis-Hastings algorithm which allows the addition and deletion of knots, thus resulting in movement across parameter spaces of varying dimensions, and using the theory introduced for this type of process by Green [14]. The model will have to adhere to a hierchical structure, in which the number of knots is the initial parameter value which must be known. Let us call the number of knots in the model, k. The possible values of k need to be defined. Now, the endpoints of the fixed interval for the model will be assumed to be fixed for the problem. So, the set of k knots that we are referring to will be in the open interval (a,b). These will be referred to as interior knots. So, once the number of knots, k, is determined, the next level in the hierarchy of parameters is the set of knot locations. These knot locations will be designated by the knot vector $\underline{t}^{(k)} = \{t_1, t_2, \ldots, t_k\}$, where $a < t_1 < t_2 < \ldots < t_k < b$. Following this level, the B-spline functions can be defined, and the coefficients for these basis functions are then parameters to be estimated. The set of coefficients for the model will be designated by $\underline{c}^{(k)} = \{c_1, c_2, \ldots, c_k\}$. Regardless of the number of knots, the error term is included

at each step in the procedure. If we allow $\underline{\theta}$ to represent the entire set of parameters, then we may write:

$$\underline{\theta}^{(k)} = \{k, \underline{t}^{(k)}, \underline{c}^{(k)}, \sigma^2\}$$

In fact, it is useful to think of the entire parameter space, over all possible dimensions, as a countable (in our case, finite) union of parameter spaces, each of the form: {k} x {$\underline{t}^{(k)}$, $\underline{c}^{(k)}$, σ^2 }. From here on, we will understand $\underline{\theta}^{(k)}$ to be defined as we have done so. If we write the likelihood function as: P($\underline{y} \mid \underline{\theta}$), then hearkening back to the discussion earlier regarding Bayes' theorem, the posterior density of the parameter vector, $\underline{\theta}$, follows the following relationship (where the prior distribution for the individual parameters will be identified by a capital P):

$$f(\theta^k|\underline{y}) \propto P(k) * P(\underline{t}^{(k)}|k) * P(\underline{c}^{(k)}|k, \underline{t}^{(k)}) * P(\sigma^2) * P(\underline{y}|k, \underline{t}^{(k)}, \underline{c}^{(k)}, \sigma^2) \tag{27}$$

The prior distributions are:

P(k)	Number of knots
P($\underline{t}^{(k)}$|k)	Knot locations (given k)
P($\underline{c}^{(k)}$|k,$\underline{t}^{(k)}$)	Spline coefficients (given k,$\underline{t}^{(k)}$)

$P(\sigma^2)$ Individual observation variance term

The hierchical structure of the model is mirrored in the definition of the joint prior distribution of the parameter set. Thus, when a set of initial values is to be randomly assigned to these parameters to initiate the execution of the MCMC simulation, these values must be sequentially selected in accordance with this hierchical structure (first the number of knots is initialized, then the knot locations given k, and finally, the values of the B-spline coefficients, which is conditioned on k, but independent of $\underline{t}^{(k)}$). The variance parameter can be intialized independently of all of these.

As indicated, a Metropolis-Hastings style MCMC algorithm will be constructed for this research. It will largely follow the type of algorithm described in the work of Biller [16], with some modifications, and with the addition of a penalty function employed in the acceptance ratio. The type of approach employed follows the reversible jump MCMC (RJMCMC) methodology which was proposed by Green [14], and briefly referred to earlier. An RJMCMC algorithm forms a transition kernel for the Markov

Chain, which permits the traversing of the entire parameter space, including potential transitions from one iteration in the chain to the next which result in a change in the dimension of the model. This technique must not only allow such transitions, but at the same time, must simultaneously satisfy the detailed balance equation.

4.5 Specification of Prior Distributions for Parameters

Discussion of the prior distributions for the various parameters in the model once again will largely emulate that of Biller [16], but less latitude is given to the prior for the number of knots than in that work. First of all, the prior distribution which is used for the number of knots inevitably is governed by the range of values which are deemed appropriate by the practitioner. As we have defined the spline function to automatically have knots at the endpoints of the fixed interval, [a,b], this implies that at least two interior knots must be selected from (a,b) in order for a legitimately defined cubic spline to be fit to the data. Also, the knots added to create the extended knot vector (for ordinary cubic splines) are not involved in

the value of k, the number of interior knots. Some value for the maximum number of possible knots permissible for the model, k_{max}, must be fixed by the practitioner. For example, I decided to fix this value at 40, as this is similar to the value used by Eilers and Marx [5] in their work on P-splines, although in this work, the dimension of the model is fixed at this value.

It is prudent to set the value of k_{max} reasonably large, whether models of dimensions this large are highly likely under the posterior density or not. We will assemble a set of candidate knots, *K,* which has k_{max} points, all of which lie in (a,b). All of these points will serve as potential locations for the knots in the spline function, regardless of the number of knots in the model.

It is assumed here that the number of interior knots will be an integer in the range from 2 to k_{max}. So, there are k_{max}-1 distinct values for k, to which the prior distribution, P(k), will assign prior probabilities. The two types of obvious possible prior distributions for k, discussed by Biller [16], are the discrete uniform and the truncated Poisson (with a suitable

value of the intensity parameter, λ, and which is normalized to sum to 1). The discrete uniform prior simply assigns equal prior probabilities to all values of k (from 2 through k_{max}). It is vital for the purposes of this research, due to the nature of the penalty function we will employ, that the discrete uniform prior be used. Otherwise, as we discuss later in the paper, the prior distribution for the number of knots, k, induced by the penalty function, will be multiplied by a value which is not constant for k. The identification of the penalty function with certain familiar probability distributions for the prior would be impossible. However, nothing about the MCMC would be invalidated by the use of a Poisson prior.

The prior distribution for the knot locations, which is conditioned on the number of knots in the model, will also assign equal prior probabilities to all possible knot vectors (with k interior knots) which are selected from the candidate set, K. So, if we designate this set in the following way:

$$K = \{u_1, u_2, \ldots, u_{k_{max}}\}$$

where each $u_i \in$ (a,b), the distinct values for u_i must be specified. We will space these equidistantly across the open interval (a,b). An alterna-

tive would be to place them at equidistant percentiles of the variable, t. However, duplicate values of t could potentially introduce discontinuities in the spline f(t) or its derivatives. It might also be undesirable to place knots in close proximity.

Now, suppose that the number of interior knots in the model is k. Given that there are k_{max} possible knot locations, the prior distribution for these locations that assigns equal probabilities to all the potential knot vectors is:

$$P(\underline{t}^{(k)})|k) = \frac{1}{\binom{k_{max}}{k}} \tag{28}$$

For the set of B-spline coefficients, the prior distribution will follow a multivariate normal distribution. This type of prior is described by Gamerman [18]. It is assumed that for these parameters, $\underline{c}^{(k)}$, along with the intercept term, we have (recall that a model with k knots has k+4 B-spline coefficients):

$$\underline{c}^{(k)} \sim MVN(\underline{0}_{(k+5)}, \sigma_0^2 * I_{k+5}) \tag{29}$$

Note: This implies that the prior for the set of B-spline coefficients assumes that they are independent random variables. They are also assumed to have equal variances, regardless of the dimension of the model. In fact, the dimensions of the multivariate mean vector and the variance/covariance matrix of the prior depends on the number of knots which has significant implications for implementation of the MCMC algorithm. This will be disucssed later in this paper. Also, assuming that there is no particular desire to impose limitations on the magnitude of the coefficients, the value of σ^2 will be relatively (in light of the order of magnitude of the dependent variable) large.

Finally, Gamerman [11], in discussing the use of prior distributions at length, refers to the common use of the Inverse Gamma distribution as a suitable prior for the variance. For normally distributed response data, with mean μ and variance σ^2, use of this prior, along with a normal prior for μ, produces a conditional conjugacy. This accounts for its common use. In the research here, the prior for σ^2 will follow an Inverse Gamma

distribution, the notation being:

$$\sigma^2 \sim IG(1, S_0), \text{where } \mathrm{S}_0 \text{ is a value set by the practitioner} \qquad (30)$$

In summarization, the priors are defined as follows:

$$\begin{aligned}
\mathrm{k} &\sim \text{Discrete Uniform} \quad (\mathrm{p}=1/(\mathrm{k}_{max}\text{-}1) \text{ for all k}) \\
\underline{\mathrm{t}}^{(k)} &\sim \text{Discrete Uniform} \quad (\mathrm{p}=1/\tbinom{k_{max}}{k}) \\
\underline{\mathrm{c}}^{(k)} &\sim \mathrm{MVN}(\underline{0}_{k+5},\ \sigma^2 * \mathrm{I}_{k+5}) \\
\sigma^2 &\sim \mathrm{IG}(1,\mathrm{S}_0)
\end{aligned}$$

4.6 Penalty Function

The primary thrust of the current research is the introduction of a specific penalty function into the Markov Chain Monte Carlo (MCMC) algorithm which generates the estimated cubic spline function fit to a set of observed data. This penalty function is an additional component of the acceptance ratio integral to the Metropolis-Hastings type of MCMC procedures. Some brief attention will be given to the fact that other penalty functions can be introduced to the acceptance ratio in a similar fashion, but for the purpose of the simulation work contained in this paper, most

of the focus will be on this specific penalty function.

The motivation of the penalty function is to strike a balance between models with minimal parameters, which are parsimonious in their functional form, but may sacrifice fit to the data, and models with larger numbers of parameters which fit the specific data under study, but sacrifice parsimony and risk overfitting. Overfitting is the phenomenon in which the number of model parameters is large enough (close to the number of observations) so that the fit to these data is nearly perfect, but may inadvertantly pick up noise in the parameter estimates, and result in potentially poor fit to a second data set generated by the same process.

The penalty function, which will be denoted by R(k), is a function of k, the number of knots in the model. By defining a penalty function which is quadratic in k, the outcome is a penalty function which places the greatest likelihood at a specified value (by the practitioner), and the least likelihood at the extremes. This penalizes models in which the number of knots differs significantly from the mean value. A particular quadratic penalty function is employed, which is a convex combination of the Akaike Infor-

mation Criterion (AIC) and the Bayesian Information Criterion (BIC), weighted by a linear weighting function (a function of k). The form of the penalty function is (where ν is constant such that $0 \leq \nu \leq 1$):

$$\ln(R(k)) = \lambda(k) * [(\nu * \ln(AIC)) + ((1-\nu) * \ln(BIC))] \tag{31}$$

Note that λ(k) is not a constant, but is a function of k itself. The function that will be used is a linear function of k. The linear function is selected so that the mean and variance of the prior for k have values specified by the practitioner. Recalling these two commonly used penalty functions:

$$\ln(AIC) = (-1/2 * \ln(n)) * k \tag{32}$$

$$\ln(BIC) = -k \tag{33}$$

The function, λ(k), will be defined after certain results have been demonstrated. Initially, let us simply define λ(k) as γ_1*k + γ_2. The penalty function can clearly be seen to produce a quadratic function of k

when the terms are multiplied. The result is:

$$\ln R(k) = (\gamma_1 * k + \gamma_2)(\nu * (1) + ((1-\nu) * (\ln n/2))) * (-k)$$

$$= -((\nu + (1-\nu)(\ln n/2)) * (\gamma_1 * k^2)) - ((\nu + (1-\nu)(\ln n/2)) * (\gamma_2 * k)) \tag{34}$$

First, a preliminary result is useful.

Theorem 1. *Any penalty function of the form, Q(k)=exp(c*(-k)), or* $\ln(Q(k))$ *= -c*k, induces a prior distribution on the number of knots which is Exponential with mean* $\frac{1}{c}$ *.*

Proof. This is straightforward. Knowing that the exponential density function is:

$$\text{f(k)} = \beta^*\exp(\text{-}\beta^*\text{k}) \text{ for k} > 0,$$

Then, clearly Q(k) $\propto$ f(k), where the constant β is independent of k. Because the constant cancels out in the acceptance ratio for the Metropolis-Hastings acceptance ratio, this constant can be ignored.

□

A natural corollary then is:

Lemma 1. *The use of AIC and BIC as penalty functions for k induces*

exponential prior densities on the number of knots with means $\frac{1}{\ln n/2}$ *and 1, respectively.*

Proof. The result is clear because $\ln AIC$ = [-$\ln n/2$]*k and $\ln BIC$ = -1*k. □

The exponential density function is known to possess the "memory-less" property where, F(t+s | t) = F(s), ∀ t,s > 0. This can be seen for penalty functions such as AIC and BIC in the fact that the ratio of the penalty function at k and k+1 is constant for all k. Thus, the penalty function itself obviously is not constant for all k, but the penalty ratio (for models having dimensions which differ by one) is constant. This is, in fact, the ratio employed in the MCMC process. It is also evident that the ratio $\left(\frac{R(k+1)}{R(k)}\right)$ is less when AIC is employed, making the acceptance probabilities less for BIC than AIC.

Now, consider the penalty function, R(k), in our methodology, where it will be shown that this constant ratio is no longer the case. Since R(k) is an exponential function whose exponent is a quadratic function of k, it is then possible to prove that R(k) is actually proportional to a normal

density function at the integer values of k=2, 3, . . . , k_{max}.

Theorem 2. *The penalty function, R(k), produces a prior distribution for the number of knots, k, which is approximately equal to a normal prior distribution with mean* $\frac{\gamma_2}{\gamma_1/2}$ *and variance* $\frac{1}{[(2*\nu+(1-\nu)*(\ln n)]*\gamma_1}$.

Proof.

The penalty function, R(k), is:

$$\mathrm{R(k)} = \exp(\text{-}((\nu + (1\text{-}\nu)(\ln n/2))*(\gamma_1*\mathrm{k}^2))\text{-}\ [(\nu + (1\text{-}\nu)*(\ln n/2))*(\gamma_2*\mathrm{k}))]$$

$$= \exp(\text{-}[(\nu + (1\text{-}\nu)*(\ln n/2))*(\gamma_1)\ ][\mathrm{k}^2 + \tfrac{\gamma_2}{\gamma_1}\ *\mathrm{k}])$$

Designating the constant, $2*[((\nu+(1\text{-}\nu^*(\ln 2/2))*\gamma_1]$, by $(\sigma_p)^2$, we can write the exponent as:

$$= \text{-C} * (\ \tfrac{1}{\sigma_p^2}) * (\mathrm{k}^2 + \tfrac{\gamma_2}{\gamma_1}*\mathrm{k}) + (\tfrac{\gamma_2}{2*\gamma_1})^2 \text{ - } (\tfrac{\gamma_2}{2*\gamma_1})^2$$

This is nothing more than completing the square, resulting in:

$$= \text{-C} * (\ \tfrac{1}{2*\sigma_p}^2) * (\mathrm{k}\text{-}\ \tfrac{\gamma_2}{\gamma_1/2})^{\ 2}$$

This, being merely the exponent of the penalty function, R(k), is proportional to a normal density function. The constant, C, is simply a multiplicative constant not involving k, so that it may be absorbed into the proportionality constant. In fact, it becomes evident by simple mul-

tiplication that:

$$\begin{aligned} \mathrm{C} &= [\nu + (1\text{-}\nu)^*(\ln n/2)]^*\gamma_1 \\ &= [\nu + (1\text{-}\nu)^*(\ln n/2)]^*2^*(\gamma_1/2) \end{aligned}$$

The mean and variance of the normal density are evident from the exponent. □

Because the prior distribution for the number of knots assumed in our MCMC procedure is discrete uniform, then multiplying the prior by this penalty function results in an expression which is proportional to the normal density function at integer values with mean and variance specified above. So, the application of this penalty function in the MCMC algorithm is then essentially equivalent to using this normal prior distribution for the number of knots, k. This is true due to the fact that the area under the normal curve between j-1 and j+1 for any integer, j, can be approximated by Simpson's rule.

It is straightforward to observe that any normal prior distribution for the number of knots, k, $\exp(a_2{}^*k^2+a_1{}^*k+a_0)$, will result in a procedure which is equivalent to imposing a penalty function on the number of knots

of the form:

$$\ln Q(k) = ((\gamma_1 + \gamma_2 * k) * (\nu * AIC + (1 - \nu) * BIC)) \qquad (35)$$

Theorem 3. *Any normal prior distribution for the number of knots can be approximated by the use of a penalty function of the form Q(k).*

Proof. Consider a penalty function, Q(k), of the form:

$$Q(k) = exp((\gamma_1 + \gamma_2 * k) * (\nu * AIC + (1 - \nu) * BIC))$$

where γ_1, γ_2, and ν are constants. If we wish for the normal prior to have mean μ and variance σ^2, this can easily be accomplished by arbitrarily specifying one of these three constants and solving for the other two so that the mean and variance of the normal density, as calculated in the previous theorem are achieved. If the value of ν is set arbitrarily (preferably between 0 and 1), then the appropriate values of γ_1 and γ_2 are:

$$\gamma_1 = \frac{1}{\sigma^2}[\nu + (1 - \nu) * (\ln n)] \qquad \text{and}$$

$$\gamma_2 = \text{-}\ 2^*\mu^*\gamma_1. \qquad \square$$

An insightful observation can be made regarding the use of the penalty

function, R(k), versus AIC or BIC. It will be the approach in this paper to set the value of the constant ν in this function to 1. This results in the simplified expression:

$$\ln\left(R(k) = (\gamma_1 * k + \gamma_2) * (-k)\right.$$

Setting this equal to the corresponding exponent for a normal distribution with the desired mean μ and variance σ^2, we have:

$$(-\gamma_1 * k^2) - \gamma_2 * k = -\left(\frac{1}{\sigma^2}\right) * (k^2 - 2\mu * k)$$

Setting coefficients equal, we find that:

$$\gamma_1 = \frac{1}{2 * \sigma^2}$$
$$\gamma_2 = \frac{-\mu}{\sigma^2}$$

Thus, to achieve an induced normal prior with mean μ and variance σ^2, one acceptable penalty function is:

$$\left[\left(\frac{1}{2 * \sigma^2} * k\right) - \left(\frac{-\mu}{\sigma^2}\right)\right] * AIC$$
$$= \left[\left(\frac{1}{2 * \sigma^2} * k\right) - \left(\frac{-\mu}{\sigma^2}\right)\right] * (-k)$$

Some insight can be gained regarding the anticipated behavior of the size of the model if one examines the expression for $\ln R(k+1)$ - $\ln R(k)$. Substituting these values into $\ln(R(k))$ yields:

$$[-(2k+1) * \gamma_1] - \gamma_2$$

When the values of the BIC penalty are similarly evaluated for k+1 and k, we simply obtain -$\ln n/2$ for this difference. It can be easily shown by simple algebra that the expression in front of -k is greater than $\frac{\ln n}{2}$ when k > μ + (σ^2*$\ln n$) - 1. At values of k which exceed the right-hand side of this inequality, we can anticipate our procedure to be more likely to reject the model of higher dimension than would be the case when the BIC penalty is employed.

This can provide some assistance in designing the penalty function in such a way that a final model is realized which is effectively smaller (in terms of number of effective parameters) than that realized under the BIC penalty function. It is immediately apparent that, generally, this penalty function will impose higher penalties on models with larger numbers of knots, but lesser penalties for those models with few parameters. It can

then be expected to produce results representative of models with moderate numbers of knots. Whatever priorities may be involved, the choice of model dimension can be incorporated into the values of μ and σ^2.

Other probability density functions can serve as the prior distribution for the number of knots through this penalty function approach. For example, the Gamma distribution has density function:

$$f(k) = \frac{1}{\Gamma(\alpha) * \beta^{\alpha-1}} * k^{\alpha-1} * e^{-\frac{k}{\beta}}$$

for k > 0 with parameters $\alpha > 0$ and $\beta > 0$. One can construct the penalty function by ignoring the constants which do not involve k. The penalty function is then:

$$Q(k) = -1 * k^{\alpha-1} * e^{-\frac{k}{\beta}} \tag{36}$$

$$\ln(Q(k)) = -\left[(\alpha - 1) * \ln k - \frac{k}{\beta}\right] \tag{37}$$

This would again make the use of $\ln(Q(k))$ practical when penalizing the log-likelihood function. The same could also be done with the Geometric probability distribution. Here, the probability of k knots would be defined as:

$$P(k) = -(p * (1-p)^k)$$

Here, p represents a number between 0 and 1, which is the probability of success in a Bernoulli trial. For the purpose of the penalty function, it is nothing more than a parameter which regulates the mean of the prior for k. The penalty function would have the form (ignoring the constant, p):

$$\ln(Q(k)) = k * \ln(1 - p)$$

Interestingly, one can once again use this as a prior for the number of knots, k, and induce both the AIC and BIC, by setting the parameter, p, to the appropriate value. For example, if p is set to (1 - e^{-1}), then $\ln(Q(k))$ = k * (-1) = -k = AIC. Note: This observation can be generally applied to distributions belonging to the exponential family, because AIC and BIC are both linear functions of k.

5 MCMC Algorithm for Spline Regression

What follows is a discussion of how the transition kernel in Markov Chain Monte Carlo procedures is implemented. The algoritm, as stated, consists of a sequence of different move types, each of which consists of its own proposal, which is accepted or rejected (according the Metropolis-Hastings acceptance probability), followed by the next move proposal. The notations that will prevail throughout the discussion are:

$P(\underline{y} \mid \underline{\theta})$ = likelihood of observed data given parameter vector $\underline{\theta}$

$P(\underline{\theta})$ = joint prior distribution of parameter vector
(with the hierchical structure)

$q(\underline{\theta},\underline{\theta'})$ = Proposal density for transition from current
parameter vector $\underline{\theta}$ to proposal parameter vector $\underline{\theta'}$

$J(\underline{\theta},\underline{\theta'})$ = Jacobian for transformation from parameter space
$\underline{\theta}$ to parameter space for $\underline{\theta'}$

$R(\underline{\theta})$ = Penalty function for the model with parameter vector $\underline{\theta}$

Before discussing the move types involved in the MCMC algorithm, those move types are defined as follows:

1. Movement of an active knot to an inactive knot location
2. Addition of a currently inactive knot or deletion of a currently active knot
3. Update of B-spline coefficients
4. Update of individual observation variance term, σ^2

This sequence is executed at each step in the chain (in this order, although this is not necessary). So, it is transparent that a Gibbs-style process, in which blocked parameters are updated one block at a time, is occurring within each iteration of the algorithm. However, the acceptance or rejection of any individual proposed transition in any of the above steps is a Metropolis-Hastings procedure where the acceptance ratio is the deciding factor whether the proposed transition is actually accepted. It may be that transitions for some of the four types of moves are accepted, while others are rejected.

5.1 Proposed Knot Move

In the course of the MCMC algorithm, an individual iteration of the chain a proposal to move to an existing active knot to an inactive knot location may or may not occur. Such a proposal occurs when $k < k_{max}$. In this step. only a change in knot location is proposed, while k and $\underline{c}^{(k)}$ are left unchanged. The proposal is also defined in such a way that not all currently active knot locations may be eligible to be moved. This restriction is imposed to exploit the local support properties of the B-spline functions.

For a given active knot, say t_j, let m_j represent the number of inactive knots, u (from the candidate set K), that satisfy:

$t_{j-1} < u < t_{j+1}$. So, in words, u is an inactive candidate knot that lies between the two active knots that are contiguous to t_j. So, the move proposal is designed so that a minimal number of B-splines are disturbed. In fact, only five B-splines are affected.

Now, from the set of current knots, it is possible that only some are movable. In this case, for an active knot, t_i, which is not movable, there

exist no inactive knots which fall in the interval (t_{i-1}, t_{i+1}). Note: if M_k (the set of movable knots) = $\emptyset$, then proceed to step 2 of the MCMC algorithm.

However, when $M_k \neq \emptyset$, then suppose there are $N_m \geq k$ movable knots. The proposed knot move consists of selecting one of these movable knots at random, say t_j, followed by randomly selecting one of the m_j candidate knots, say t_{j^*}, which are eligible given that t_j was chosen. All this results in the following expression for the joint probability of selecting t_j and t_{j^*}:

$$q(t_j, t_{j^*}) = \frac{1}{m_j} \times \frac{1}{N_m} \tag{38}$$

This is the transition kernel which is generated at this step of the process. Recalling the detailed balance equation, it will be necessary to calculate the corresponding reverse move (moving active knot t_{j^*} to inactive knot t_j). To understand how this would affect the transition kernel, consider the following diagram:

$$t_1 < \ldots t_{j-1} < t_{j^*} < t_j < t_{j+1}$$

In the reverse move, t_{j^*} is an active knot, and t_j is not. The contiguous knots for t_{j^*} are t_{j-1} and t_{j+1} and the number of potential move positions

for t_{j^*} is once again m_j since the inactive knots between t_{j-1} and t_{j+1} are identical with our previous discussion (with the exception that the previous t_j replaces t_{j^*}). Now, referring to the current state of the parameter vector as $\underline{\theta}$ and the proposed parameter vector as $\underline{\theta^*}$, where the parameter vector is the entire set of parameters, but the only change from $\underline{\theta}$ to $\underline{\theta^*}$ is the change of one knot position. Then the calculation of the transition probability yields:

$$\alpha(\underline{\theta}, \underline{\theta^*}) = min\left(1, \frac{P(\underline{y}|\underline{\theta^*}) * q(\underline{\theta^*}, \underline{\theta})}{P(\underline{y}|\underline{\theta}) * q(\underline{\theta}, \underline{\theta^*})}\right) = \frac{N_m}{N_{m^*}} \tag{39}$$

Now, it may be asked, if t_{j^*} simply replaces t_j in the interval (t_{j-1},t_{j+1}) as an active knot, what circumstances should lead to $N_m \neq N_{m^*}$? This would occur under an example such as the following. Consider the following sequence of active and inactive knots (the t's are active, the u's are candidate knots which are inactive):

$$t_{j-2} < t_{j-1} < u_1 < u_2 < t_j < u_3 < u_4 < u_5 < t_{j+1}$$

Suppose u_1 is selected as the proposed new knot location. Then the new proposed knot vector looks like:

$$t_{j-2} < t_{j-1} < t_j < u_1 < u_2 < u_3 < u_4 < u_5 < t_{j+1}$$

where the knot which starts as u_1 becomes t_j and the current active knot, t_j, becomes the inactive knot, u_3. Other obvious relabeling occurs as well. Note: there is no change in model dimension under this type of move proposal. So, the penalty function which is to be implemented has no effect on the acceptance probability ratio.

5.2 Proposed Knot Addition or Deletion

This step is actually one step, which for any given individual proposal, will be either a proposed knot addition or deletion. Initially, the decision whether to propose the addition or deletion is made at random (assuming that both proposals are possible). If only two interior knots are active, a knot addition will automatically be proposed. If all candidate knots are active, a knot deletion proposal will occur with probability 1. As we

proceed with the discussion of this step, it must be borne in mind that if a knot addition is the proposal under consideration, then the reverse step (which must be understood correctly to properly evaluate the acceptance ratio) is a knot deletion. Suppose that the current set of active knots is:

$$\underline{t}^{(k)} = \{a = t_0, t_1, t_2, \ldots, t_k, t_{k+1} = b\}$$

If a knot addition is to be proposed, then one of the k_{max}-k inactive knots is selected at random for this. Thus, the proposed knot vector now consists on k+1 interior knots (plus the endpoints a and b). This necessarily introduces a potential new knot location that must lie between two currently active knots, t_i and t_{i+1}. One can easily see that by introducing a new knot into the model, not only is this new location a parameter, but an additional B-spline basis function must be added so that the spline space for [a,b] of dimension k+5 can be spanned. So, the transition kernel will consist of the density of selecting one of the inactive knots for inclusion, plus an auxiliary uniform[0,1] random variable, v, which is introduced to maintain the dimension matching described by Green [14], and which follows the methodology of Biller [16].

Let us denote the proposed knot vector with the star superscript, $\underline{t}^{*(k+1)}$. So, given the proposed placement of the new active knot, this vector can be specified as:

$$\underline{t}^{*(k+1)} = \{a = t_0^* < t_1^* < \ldots < t_i^* < t_{i+1}^* < t_{i+2}^* \ldots < t_{k+1}^* < t_{k+2}^* = b\}$$

To keep the bookkeeping as clear as possible, here:

t_j^*	$= t_j$	for j=1, 2, . . . , i
t_{i+1}^*	$= u$,	where u is the inactive knot proposed for inclusion
t_j^*	$= t_{j-1}$	for j=i+2, . . . , k+1

Because of the transdimensional nature of this proposed move type, when an additional knot location is proposed to be added to the active set of knots, this implies that an additional B-spline function must be added to the parameter set. So that the necessary computation of the Metropolis-Hastings acceptance ratio can be computed, the auxiliary random variable, v, is generated. This is just a uniform random number on the interval [0,1], which enables a bijection to be constructed between the joint current parameter space (dimension=2k+5) and proposal (new knot, u, and auxiliary variable, v: dimension=2, making a total of 2k+7) and the

new parameter space (dimension=k+5+k+1+1=2k+7). This dimension matching will also introduce an additional term into the acceptance ratio, the Jacobian of this bijection between parameter spaces. Including this Jacobian term (but ignoring the penalty function R($\underline{\theta}$)) now, the acceptance ratio now has the form:

$$\alpha(\underline{\theta}, \underline{\theta^*}) = min\left(1, \frac{P(\underline{y}|\underline{\theta^*}) * P(\underline{\theta^*})q(\underline{\theta^*}, \underline{\theta}) * J}{P(\underline{y}|\underline{\theta}) * P(\underline{\theta})q(\underline{\theta}, \underline{\theta^*})}\right) \tag{40}$$

The form of each of the terms needs to be explained. In order to understand these, the transition that is being proposed from vector of current B-spline coefficients $\underline{c}$ to $\underline{c^*}$ is, in reality, a mapping (recalling the uniform variate, v):

$$\phi : (\underline{c}, v) \mapsto \underline{c^*} \tag{41}$$

First, assuming that a new active knot is to be added, this knot is selected at random from the available k_{max} - k inactive knots. Thus, this proposal probability is clearly:

$$q(t^*) = \frac{1}{k_{max} - k} \tag{42}$$

The random variable, v, has density function, q(v)=1,

for $0 < v < 1$. This random variable is used to map the set of current k+4

B-spline coefficients to a new set of (k+1)+4 coefficients. Following the procedure employed by Biller [16], three B-spline coefficients are modified in a manner which mimics rules used for updating B-spline coefficients in the work of Lyche and Strom [19]. Denoting the new set of proposed coeffients by $\underline{c^*}$, these rules are as follows:

1. c^*_{i+1} = (v*c_i) + ((1-v)*c_{i+1})
2. c^*_i = c_i - (r*c^*_{i+1})
3. c^*_{i+2} = c_{i+1} - ((1-r)∗ c^*_{i+1})

r is defined as: $\frac{t^*-t_i}{t_{i+1}-t_i}$

which is simply the relative position of the new knot, t^*, in the interval, (t_i, t_{i+1}). Along with these modified coefficients, $c^*_j = c_j$ for j=1,2, . . . , i-1 and $c^*_j = c_{j-1}$ for j=i+3, . . . , k+1. Effectively, there are only three coefficients which are modified, while others may experience a change in subscript without modifying their impact on the estimated function. Thus, the proposed new set of parameter values includes a new knot lo-

cation and a new set of B-spline coefficients. This transformation, ϕ, requires the computation of its associated Jacobian. Recall that the Jacobian of a transformation is the determinant of the matrix of partial derivatives of the current parameters (including the random variable, v) and the transformed parameters. The generation of v provides us with the necessary square matrix (of dimension k+5 x k+5). However, because virtually all of the B-spline coefficients remained unchanged under this transformation, this Jacobian matrix is comprised of block submatrices along the diagonal and the computation of the determinant is greatly simplied. This is another great advantage of the use of the B-splines as the basis for the spline.

Given that there is only a real change in the 3x3 set of variables described above, the value of the determinant is nothing more than a matter of calculating the determinant of this 3x3 submatrix.

Here is the appearance of the 3x3 matrix of partial derivatives which result from the above transformation:

$$\begin{pmatrix} \frac{\partial c_i^*}{\partial c_i} & \frac{\partial c_{i+1}^*}{\partial c_i} & \frac{\partial c_{i+2}^*}{\partial c_i} \\ \frac{\partial c_i^*}{\partial c_{i+1}} & \frac{\partial c_{i+1}^*}{\partial c_{i+1}} & \frac{\partial c_{i+2}^*}{\partial c_{i+1}} \\ \frac{\partial c_i^*}{\partial c_v} & \frac{\partial c_{i+1}^*}{\partial c_v} & \frac{\partial c_{i+2}^*}{\partial c_v} \end{pmatrix} \tag{43}$$

Supplying the actual expressions for the partial derivatives results in:

$$\begin{Bmatrix} 1 - rv & v & v * (r - 1) \\ vr - 1 & 1 - v & (1 - v) * r + v \\ r * (c_{i+1} - c_i & c_i - c_{i+1} & (1 - r) * (c_{i+1} - c_i) \end{Bmatrix} \tag{44}$$

Evaluation of the determinant of this matrix yields $|J| = |c_{i+1} - c_i|$. Thus, when we substitute this into the acceptance probability, α, the ratio simplifies to (and including the penalty function as well):

$$f(\underline{\theta^*}|\underline{y}) = \frac{P(\underline{y}|\underline{\theta^*}) * P(\underline{\theta^*}) * q(\underline{\theta^*}, \underline{\theta}) * J * R(k+1)}{P(\underline{y}|\underline{\theta}) * P(\underline{\theta}) * q(\underline{\theta}, \underline{\theta^*}) * R(k)} \tag{45}$$

This complicated looking expression simplifies greatly, when, in calculating the ratio, we recognize that many parameters remain unchanged in the proposal:

$$\frac{P(\underline{\theta}^*)}{P(\underline{\theta})} = \frac{P(k+1)}{P(k)} * \frac{\binom{k_{max}}{k+1}}{\binom{k_{max}}{k}} * \frac{P(\underline{t}^*)}{P(\underline{t})} * \frac{P(\underline{c}^*)}{P(\underline{c})}$$

$$= \frac{\binom{k_{max}}{k+1}}{\binom{k_{max}}{k}} * (2\pi * \sigma^2)^{-1/2} * exp\left[\frac{1}{2 * \sigma^2} * (\underline{c}'\underline{c} - \underline{c}^{*\prime}\underline{c}^*)\right]$$

$$= \frac{k+1}{k_{max} - k} * (2\pi * \sigma^2)^{-1/2} * exp\left[\frac{1}{2 * \sigma^2} * (\underline{c}'\underline{c} - \underline{c}^{*\prime}\underline{c}^*)\right]$$

The penalty function is the final component for the evalation of the acceptance ratio.

$$R(k) = (exp(\gamma_1 * k^2 + \gamma_2)) * (-k)$$

This, then, is the acceptance ratio used when a knot addition is proposed. When the reverse proposal is under consideration (deletion of the knot t^* in our discussion), the acceptance ratio will simply be the reciprocal of this expression. A pair of comments are in order. When a knot deletion is proposed, no auxiliary uniform random variable is needed

because the current location of the knot which is to be deleted will correspond deterministically with a specific knot addition proposal, including a fixed value for v. Also, it must be recalled that when the current active knot vector has either the minimum or maximum number of knots, then the probability that a knot addition or deletion is proposed is now no longer 1/2, but 0 or 1 accordingly.

5.3 Update of B-Spline Coefficients

It proves to be very advantageous to separate the step in which knots are added or deleted from the step in which the B-spline coefficients are updated. Attempting a joint proposal of both the proposed knot change and an update of the complete set of B-spline coefficients would seriously complicate the computation of the Jacobian of the transformation. The methodology that will be used in the process here is the weighted least-squares proposal contained in the work of Gamerman [18]. The approach in the fitting of generalized linear models is employed where Fisher scoring is used to iteratively estimate the model's parameters. Here, in the proposal of a new set of parameter values, the posterior density of the

B-spline coefficients given the data and other fixed parameter values is maximized, and the estimates of these, along with the estimated variance/covariance matrix for these is used as the mean and variance of the multivariate normal proposal distribution. Specifically, suppose that the current values for the B-spline coefficients are given by $\underline{c}^{(k)}$, assuming that the current number of knots in the model is k (plus the intercept term). Denote by $\underline{c}^{*(k)}$, the proposed values for the model (with the same k knots).

According to Gamerman [18], the proposed set of new values for these coefficients is randomly selected from a multivariate normal with the following mean and variance:

$$\underline{c}^{*(k)} = (\Sigma^{-1} + X'W(\underline{c}^{(k)})X)^{-1}(X'W(\underline{c}^{(k)}\underline{y} * \underline{c}^{(k)})^{-1} \tag{46}$$

In this expression, X is the design matrix comprised of the values of each of the k+4 B-spline functions at the value for the spline variable, plus the intercept term. W represents a diagonal matrix of weights in

the iterative process, where the diagonal elements are each equal to the reciprocal of the current variance term (σ^2). The matrix, Σ, is the prior (k+5)x(k+5) variance/covariance matrix for the joint prior distribution for the B-spline coefficients and the intercept term, and the vector, $\underline{y}$, is the data vector. Thus, the proposal is one step of the usual iterative weighted least-squares approach to fitting a generalized linear model.

As with the other proposal types, the reverse proposal can be defined, where it is assumed that the model would potentially transition from $\underline{c}^{*(k)}$ to $\underline{c}^{(k)}$. When the final decision is considered, whether to accept the move to $\underline{c}^{*(k)}$ as the new B-spline coefficients, the acceptance probability depends merely on the ratio of posterior densities,

$$\frac{P(\underline{y}|k,\underline{t}^{*(k)},\underline{c}^{*(k)},\sigma^2)}{P(\underline{y}|k,\underline{t}^{(k)},\underline{c}^{(k)},\sigma^2)}$$

times the ratio of the proposal densities,

$$\frac{q(\underline{c}^{*(k)},\underline{c}^{(k)})}{q(\underline{c}^{(k)},\underline{c}^{*(k)})}.$$

5.4 Update of Variance Term

The final component which is eligible for a proposed new estimated value is the individual observation level error term, σ^2. Recall that this parameter was assigned a prior Inverse Gamma distribution, denoted by IG(1, S_0). It may be noted at this point, that when σ^2 has this density, then the scale parameter (under normally distributed data) is $1/\sigma^2$, which will then have a Gamma distribution, of the form, Gamma(1,S_0). Because the Gamma distribution (Gamma(α,β)) has a mean of $\frac{\alpha}{\beta}$, this leads to the result that the prior distribution for the scale parameter has mean $1/S_0$, or, equivalently, that the prior for the variance, σ^2, has mean S_0. This can be viewed, intuitively, as a prior sample of size 1, with a prior value for the "error sum of squares" of S_0.

Given the conditional conjugacy that exists when the mean of the data has a normal prior and the scale parameter has a Gamma prior, it follows

that the posterior density for the scale parameter, given fixed values for the data and other parameters, likewise follows a Gamma distribution. In fact, to verify this fact, designating the scale paramter by $\phi = \frac{1}{\sigma^2}$:

$$P(\phi|\underline{y}, k, \underline{t}, \underline{c}) \propto P(\underline{y}|k, \underline{t}, \underline{c}, \phi) * P(k) * P(\underline{t}|k) * P(\underline{c}|k, \underline{t}) * P(\phi)$$

Because the only terms on the right-hand side that involve ϕ are the likelihood function and P(ϕ), then this proportionality statement can be written in the simpler form:

$$P(\phi|\underline{y}, k, \underline{t}, \underline{c}) \propto P(\underline{y}|k, \underline{t}, \underline{c}, \phi) * P(\phi) \tag{47}$$

Or,

$$P(\phi|\underline{y}, k, \underline{t}, \underline{c}) \propto exp(-(S_1/2) * \phi) * exp(-S_0) \tag{48}$$

Here, S_1 is the error sum of squares for the current model fit. But, including the appropriate expression for the likelihood, we have:

$$P(\phi|\underline{y}, k, \underline{t}, \underline{c}) \propto (\phi)^{-n/2} * exp(-S_1 * \phi) * \frac{S_0}{2} * exp(-S_0/2) \tag{49}$$

Combining terms as required yields:

$$P(\phi|\underline{y}, k, \underline{t}, \underline{c}) \propto (\phi)^{-(n+1)/2} * exp(-(S_0 + S_1)/2) \tag{50}$$

Thus, the posterior density of ϕ is clearly Gamma((n+1)/2,(S_0+S_1)/2)), so that the posterior of the variance σ^2, is Inverse Gamma with mean ((S_0+S_1)/2)/(n+1). In this manner, it is seen that the posterior density has a mean which is infuenced to a greater degree for smaller samples. The MCMC algorithm at this point randomly selects a proposed new value for the estimated variance from this posterior distribution (which depends on the values of the other parameters in the model) and computes the appropriate acceptance probability.

5.5 Function Estimation

The estimated curve from the algorithm as a whole is determined by averaging the estimates of the value of f(t), at any specific value for the independent variable, t, for a sufficiently large number of sample observations from the limiting distribution, π. Some determination can be made that the limiting distribution has been approximately realized during the course of the chain. In a discussion of various options for making the determination, Gamerman [18] offers some suggestions. Because of the changing dimension of the algorithm in this research, even under the

limiting distribution, stability of the B-spline coefficients in this mature part of the chain in not a helpful diagnostic. More meaningful diagnostic approaches rely upon whatever metrics remain invariant in their interpretation across chain iterations. In this research, estimated functional values will suffice for this purpose.

When the limiting distribution has been reached, it is useful to divide the observations into batches and compute averages for each of the batches at each unique value of t. Once these batch averages have been calculated, the averages of the batches (at each t), can also be averaged to arrive at a final estimated value for f(t). This batch methodology facilitates the estimation of the standard error of this estimate. It is also prudent to form batches from observations which are separated by some large number of iterations, perhaps 50. This can be helpful for overcoming dependency between estimated values of iterations which are close. It is straightforward to demonstrate, given a fixed set of candidate knots, that the final estimated curve, f(t), remains a cubic spline function, with knots located at the union of all knots which occur in at least one of the

iterations which contribute to the final calculation.

6 Simulation Results

In order to evaluate the performance of the penalty function, R(k), several data sets were generated from underlying smooth functions, together with an additive random noise term. For each instance of a test function, f, a training set was generated for the purpose of model estimation. Then five additional test data sets were generated for each function, f, for the purpose of testing prediction accuracy. Functions with differing numbers of local extremum points were used to evaluate the robust nature of the penalty function. The set of generating functions will be designated by $f_i(t)$. They were defined as follows:

f_1 = Cubic spline on [0,10] with interior interior knots at

$\{2.3, 3.8, 5.2, 7.1, 8.6\}$

f_2 = Cubic spline on [0,10] with interior interior knots at

$\{2.0, 4.6, 5.9, 6.6, 8.1\}$

$f_3 = -1.5 + \frac{1}{2.75} * t^2$

$f_4 = 8.2 + \sin(3\pi * t)$

A series of runs of the MCMC algorithm was conducted for each test function using various penalty functions. Not all penalty functions were used for each test function, but here is a summary of those employed:

SUMMARY OF PENALTY FUNCTIONS

1 NO PENALTY

2 AIC PENALTY

3 BIC PENALTY

4 R(k), μ=20, σ^2=25

5 R(k), μ=10, σ^2=9

6 R(k), μ=5, σ^2=2

7 Eilers and Marx penalty (second differences of B-spline coefficients)

The MCMC algorithm was coded in a sequence of Matlab programs. For each run, it was determined to execute 50,000 iterations of the Markov chain, with the first 25,000 steps considered to be the burn-in period. Then, the remaining 25,000 steps were taken as a random sample from the limiting distribution of the chain. These 25,000 were divided into 50

batches of 500 iterations, with every other batch omitted from the estimation process. This was done to more accurately approximate independence of batches. Average functional estimates were obtained for each of the 25 batches of 500 iterations. In addition to the test data sets that were used to measure prediction error for each experiment, a measure called the Deviance Information Criterion (DIC) was also calculated for each model. For the two most frequently used model selection criteria, AIC and BIC, the calculation of these statistics is evaluated at the maximum-likelihood estimates of the parameters. Because the MCMC technique samples from the joint posterior distribution of these parameters, the estimates at any given iteration of the chain may differ from these values. The DIC is described by Spiegenhalter, Best, Carlin, and van der Linde [20]. In the fitting of generalized linear models, the deviance is defined as:

$$D = -2 * \ln(\underline{y}|\underline{\theta})$$

The DIC is then defined as $\overline{D} + \mathrm{p}_D$, where p_D is a measure of the effective number of parameters. This recognizes that the number of parameters is not fixed in the Bayesian averaging process. $\overline{D}$ can be considered to be

the expected value of the deviance and can be readily calculated as the Markov Chain progresses. It can be calculated as $\overline{D}$, or the estimated average value for this statistic. The effective number of parameters is: p_D = $\overline{D}$ - D($\overline{\theta}$). The second term is simply the value of the deviance when the average values of the estimated function for the sampled observations are used to compute the deviance statistic. Simplifying yields:

$$DIC = \overline{D} + p_D \tag{51}$$

Smaller values of the DIC statistic are indicative of a better fit, corrected for the effective number of parameters. It needs to be admitted that the DIC is an asymptotic statistic which assumes that the joint posterior density function for the parameters is Multivariate Normal, which is not the case in our research. We will calculate it for a rough guideline of fit, however. We will now track the performance of the various penalty approaches for these test functions.

CASE STUDY 1

f_1, n=100, $\sigma^2 = 1.5$, Intercept term=4.0

B-spline coefficients: 1.3, 3.8, -7.5, 9.2, -8.1, 6.7, -8.8, 4.7, 7.5

Values for the independent variable, t, were set at intervals of .1 starting at .1. The final sample point was generated at t=9.95 (rather than 10.0).

The value for the dependent variable was randomly generated using these parameter values, one at each of the specified values of t. A table is given below showing a summary of the results of the modeling effort. It is clear that the use of no penalty function produces the largest number of knots across the Bayesian averaging process, but both the AIC and BIC produce similar results.

The following graph (Figure 1) shows the results with no penalty function. The graph shows the actual data (circle), the underlying function (black), f_1, and the estimated values for the function over the range [0,10] for t (blue).

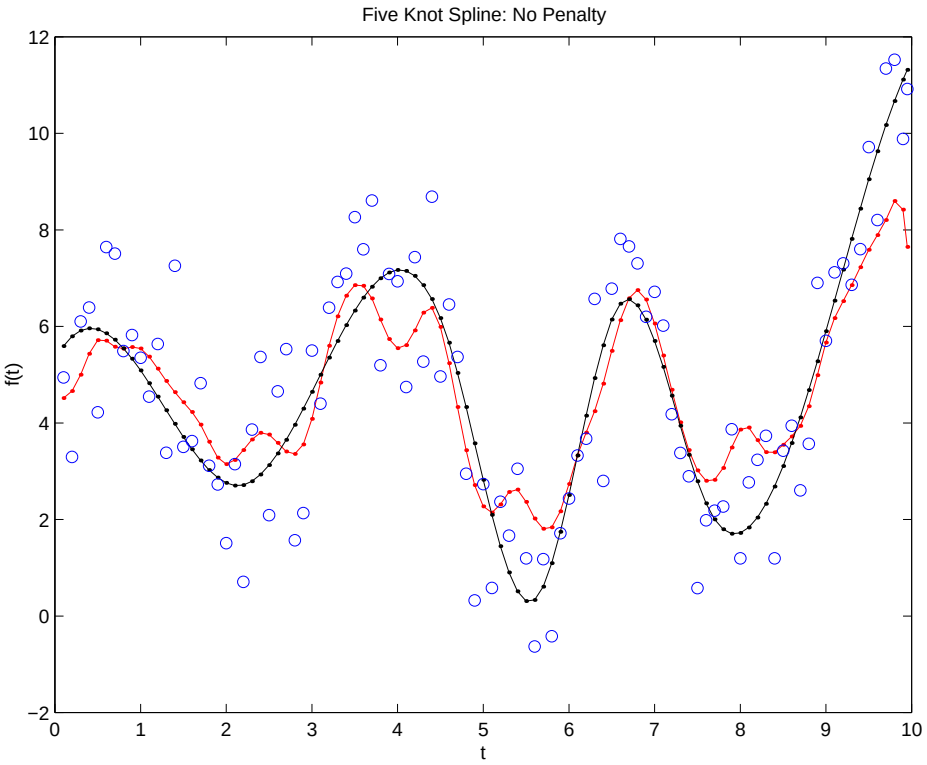

Figure 1: Estimates for the model (No penalty)

It is evident that the estimated function follows the pattern of the data. It will also be clear from results shown later that this approach is greedy in terms of the number of knots selected for the model. The vast majority of the time, the number of knots selected once convergence of the chain has been reached is the maximum value. The graph in Figure 2 below shows the results with the AIC penalty function.

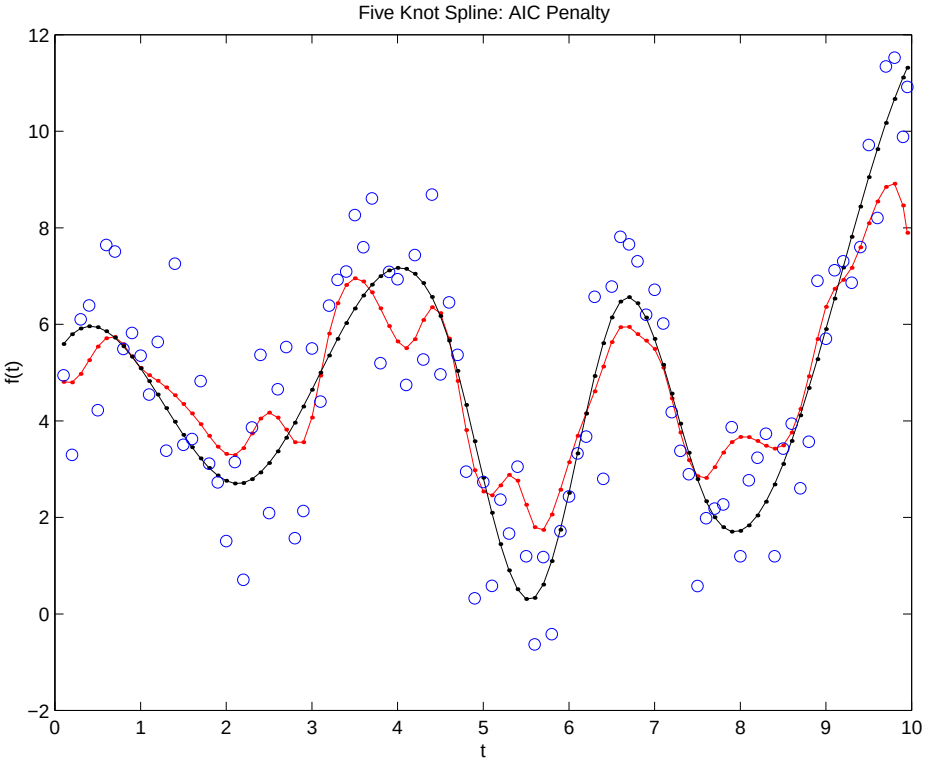

Figure 2: Estimates for the model (AIC penalty)

Although a penalty function has been applied here, it is still the case that the number of knots selected generally is very near the maximum. In fact, the same turns out to be true when the BIC penalty function is employed. This is due to the fact that, although a larger penalty is applied

to models with more knots, the individual transition from one step in the Markov chain to the next only involves a difference in the log-penalties of the competing models of 1. Figure 3 illustrates the results obtained for the BIC penalty function.

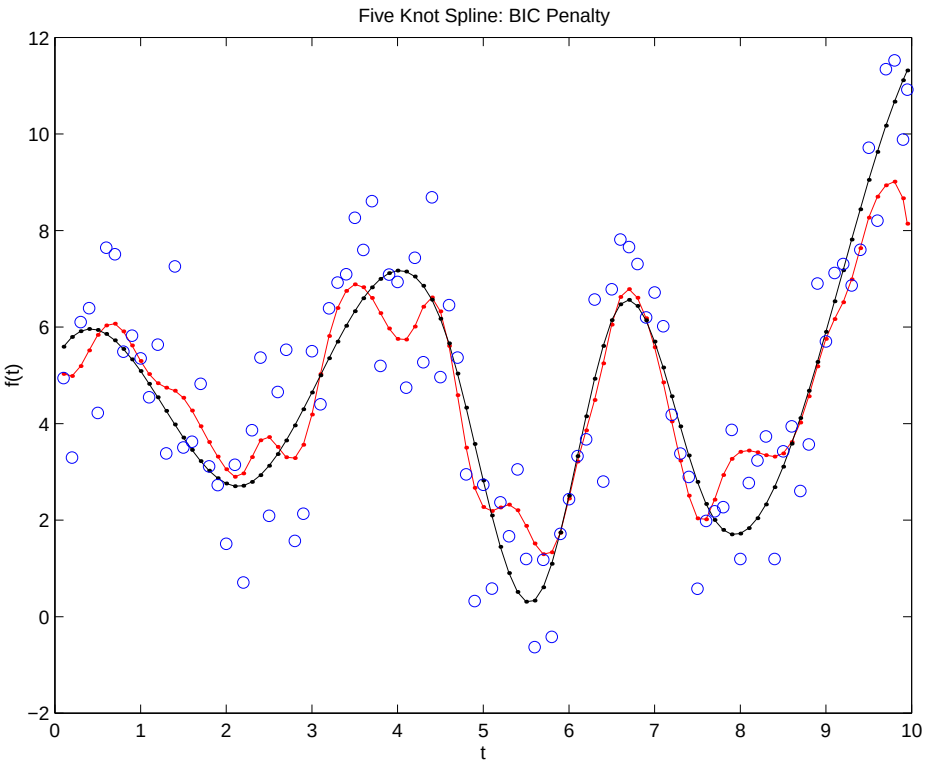

Figure 3: Estimates for the model (BIC penalty)

In order to examine the viability of the proposed penalty function, R(k), a variety of mean/variance pairs were tested. As noted above, for Penalty 4, μ was set equal to 20 and σ^2 to 25. For Penalty 5, these values were 8 and 9, respectively. Neither of these values produced results that were considered significantly better (in terms of prediction error) to the first three options. The final one listed above shows better performance for Test 1, however. In this case, Penalty 6 is a penalty function which induces a normal prior with mean 5 and variance 2. Under this scenario, the following results were obtained: It is visibly evident that this penalty

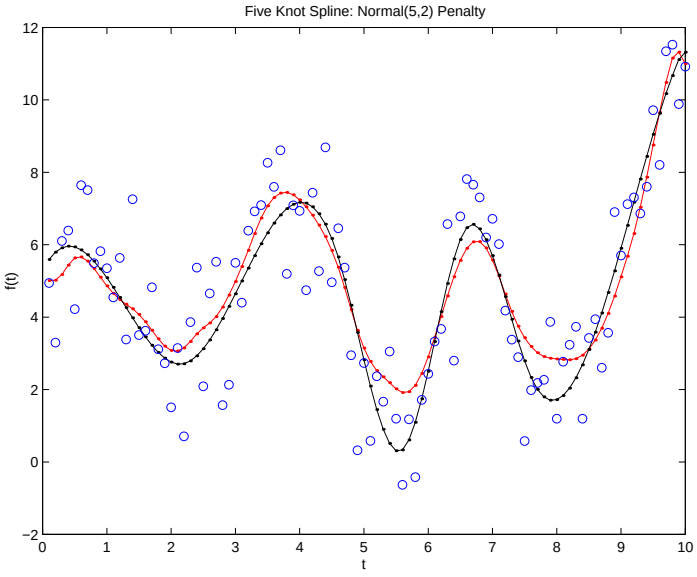

Figure 4: Estimates for the model (Normal(5,2) penalty)

approach does not exhibit as many local minor oscillations as the previous approaches. Table 1 indicates that the average number of knots used in the sampled iterations of the Markov chain is smaller for this penalty function. This raises the hope that this option will avoid the overfitting that is likely incorporated in those methods using no penalty, the AIC, or the BIC methodologies. Finally, to demonstrate that another type of penalty function, based on smoothness rather than dimension, approach 7 listed above, was run. In this modeling approach, the exponential of the sum of the squared second-differences of the B-spline coefficients was used as the penalty function (resembling the Eilers/Marx model selection philosophy). It is possible to specify a prior on the spline coefficients to accomplish this approach, but the basis functions must be defined as the appropriate monomials plus the truncated power functions. By specifying a prior distribution for these truncated power functions with a small variance, large magnitudes for these second differences can be penalized as desired. The use of the penalty function is a simpler method for implementation of the same concept. The graph showing the results obtained

is found in Figure 5.

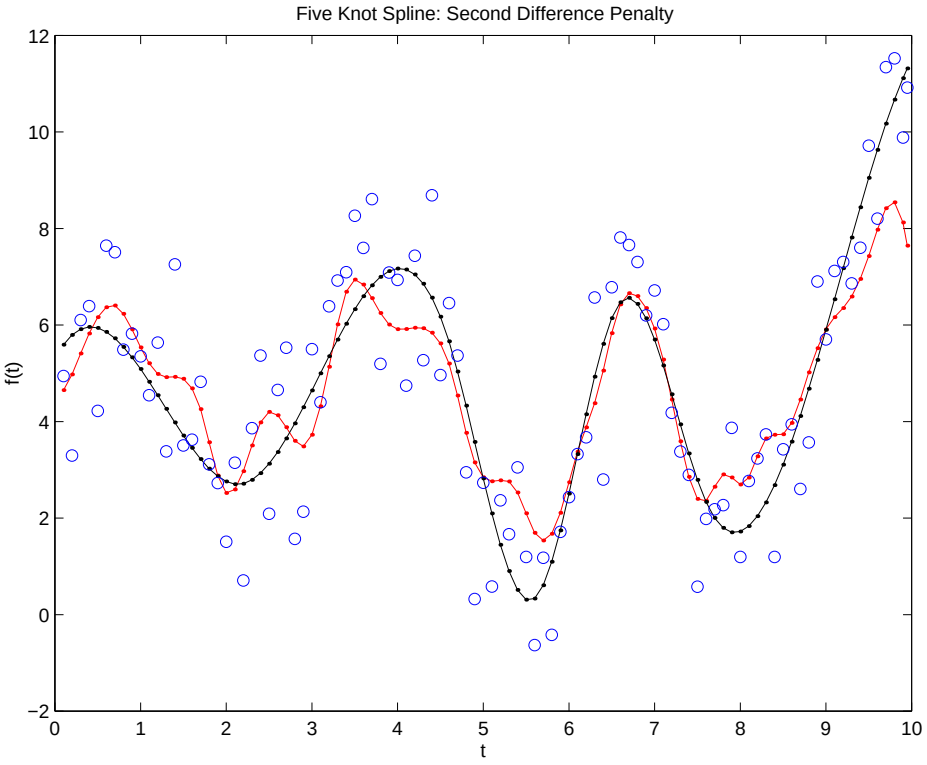

Figure 5: Estimates for the model (Second-difference penalty)

A display of the distribution of the number of knots for the approach with no penalty and the approach with the Normal(5,2) penalty is shown below for contrast. Clearly, the use of no penalty results in the incorpora-

tion of all candidate knots in the vast majority of the sampled iterations of the simulation. The penalty function tempers this result considerably and allows for appropriate smoothing.

In addition, figure 8 shows the results when a bootstrap sample of

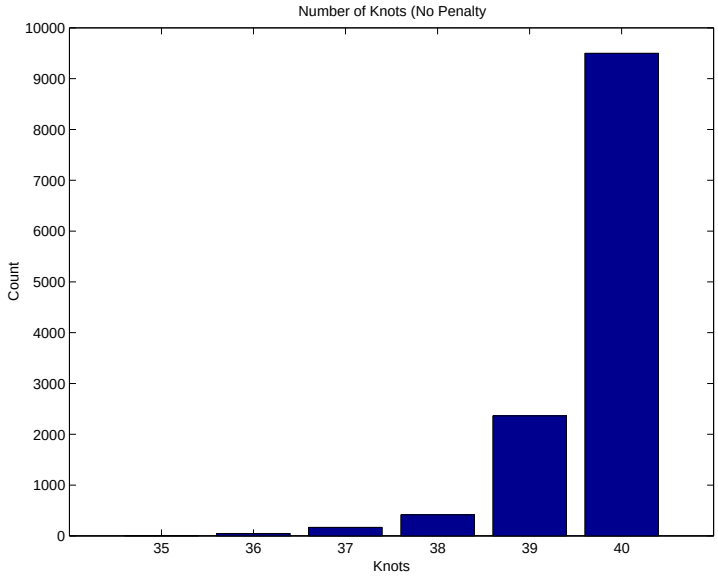

Figure 6: Distribution of number of knots (No penalty)

1000 observations from the these cases was used to generate 90 percent pointwise confidence intervals for the estimated functional values at t ϵ {.1, .2, . . . , 9.9, 9.95 } for the No Penalty approach and the Normal(5,2) penalty function. Neither approach seems to have a clear advantage.

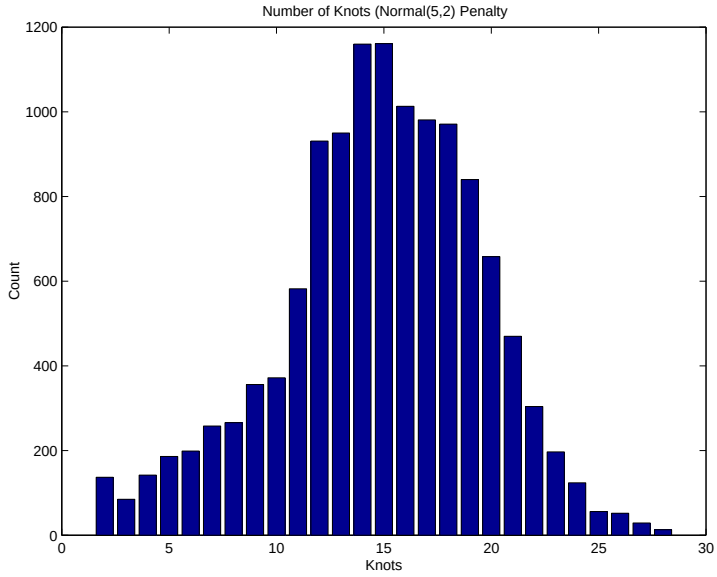

Figure 7: Distribution of number of knots (Normal(5,2) penalty)

A summary of the results from these models for f_1 is displayed in Table 1.

The only penalty function which makes any substantial difference in the dimension of the model which is fit is the Normal prior with mean 5 and variance 2. In terms of local shape of the final estimated curve, it is then only option which has any promise of escaping the peril of overfitting. In order to investigate this issue, five additional data sets of 100 observations at the same values of t were generated and the sum of squared errors computed for each model for each of these data sets. This permits

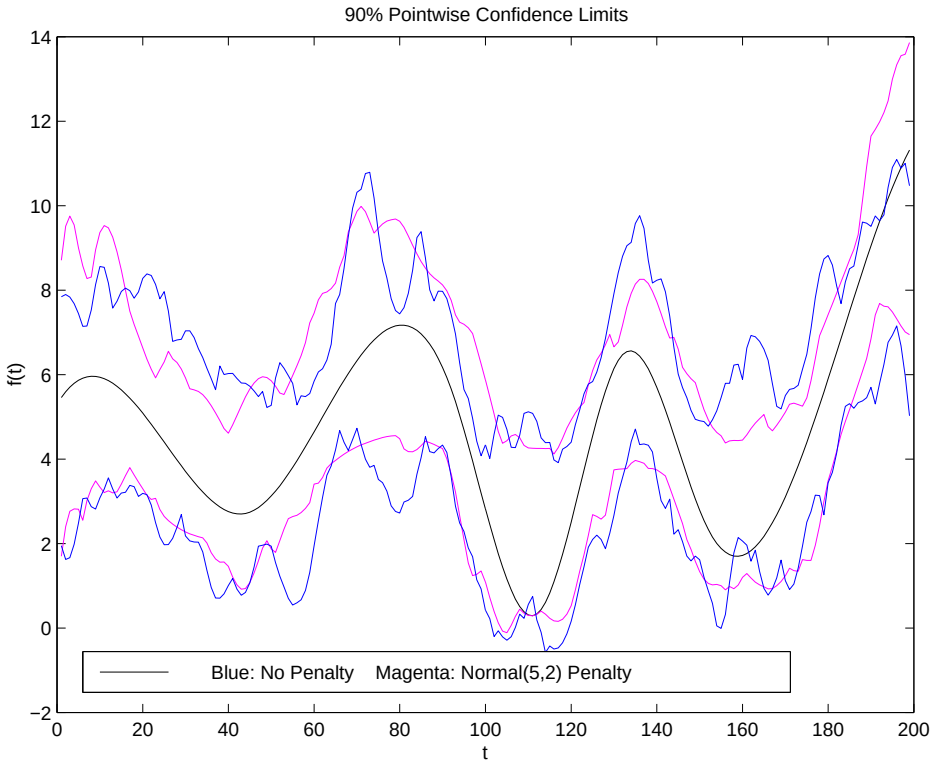

Figure 8: Confidence Intervals for Estimated Curve

us to observe how well each approach performs against data generated independently of the modeling algorithm. The investigation indicates that the Normal(5,2) penalty does, in fact, perform better, in terms of prediction accuracy. The Normal(5,2) prior induced by the proposed penalty function, R(k), has the smallest prediction error for each of the five data sets. The results are summarized below in Table 2:

PENALTY	SSE	AVG KNOTS
None	110.5	39.7
AIC	181.9	39.4
BIC	152.5	38.8
Normal(20,25)	177.6	39.4
Normal(10,9)	172.1	38.5
Normal(8,9)	162.1	38.0
Normal(5,2)	177.0	14.8
Sec. Diff.	177.4	39.6

Table 1: Summary of Results for Five-Knot Spline

The DIC statistics for each of these approaches were:

There is strong evidence that the Normal(5,2) penalty function is a promising approach for this particular underlying function. The values of γ_1 and γ_2, as defined earlier are: γ_1=.25 and γ_2=-2.5.

The second generating function that was used, f_2, was also a spline with five knots, defined on the same closed interval [0,10]. The knots are placed close to the locations for f_1. In addition, the B-spline coefficients are identical to f_1, so that we can determine whether a slight perturbation in the generating function would alter the type of results we obtain. All modeling scenarios were not run, but the results resemble those of the

PENALTY	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
None	300.4	328.0	281.4	303.4	307.9
AIC	291.1	310.3	277.9	306.2	294.5
BIC	261.2	270.6	256.8	276.0	266.0
Normal(20,25)	292.6	315.3	275.2	307.4	305.4
Normal(10,9)	297.0	325.4	268.9	295.7	294.2
Normal(8,9)	271.7	282.5	279.9	285.7	274.6
Normal(5,2)	254.0	254.2	264.3	260.2	259.6
Sec. Diff.	307.0	318.2	274.1	286.2	295.8

Table 2: Summary of Prediction Evaluation for Five-Knot Spline

NONE	AIC	BIC	Normal(20,25)	Normal(10,9)	Normal(8,9)	Normal(5,2)	Sec. Diff
10.07	8.86	9.16	8.92	8.94	9.01	7.53	8.91

Table 3: Summary of DIC for Five-Knot Spline

first spline. The parameters for Case Study 2 are:

CASE STUDY 2

f_1, n=100, $\sigma^2 = 1.5$, Intercept term=4.0

B-spline coefficients: 1.3, 3.8, -7.5, 9.2, -8.1, 6.7, -8.8, 4.7, 7.5

Table 4 is the summary of the results:

In addition, the DIC statistics are: No Penalty: DIC=10.14, Normal(5,2)

PENALTY	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
None	356.9	358.6	327.1	337.8	317.3
Normal(5,2)	349.7	369.0	291.9	295.8	289.6

Table 4: Summary of Prediction Evaluation for Five-Knot Spline

Penalty: DIC=8.84

To evaluate how well the proposed penalty function does with a generating function without any local extremum points (in the closed interval), the quadratic function designated as f_3 above was tested. So, this case study is defined as:

CASE STUDY 3

f_3 = -1.5 + $\frac{1}{2.75}$*t^2 = , n=100, σ^2=1.5,

Initially, graphs for the results for runs with no penalty and the Normal(5,2) penalty are displayed below (Figures 9 and 10):

Table 5 shows the prediction error summary and reveals that R(k) does not have a decisive edge in this case. The DIC statistics are: No

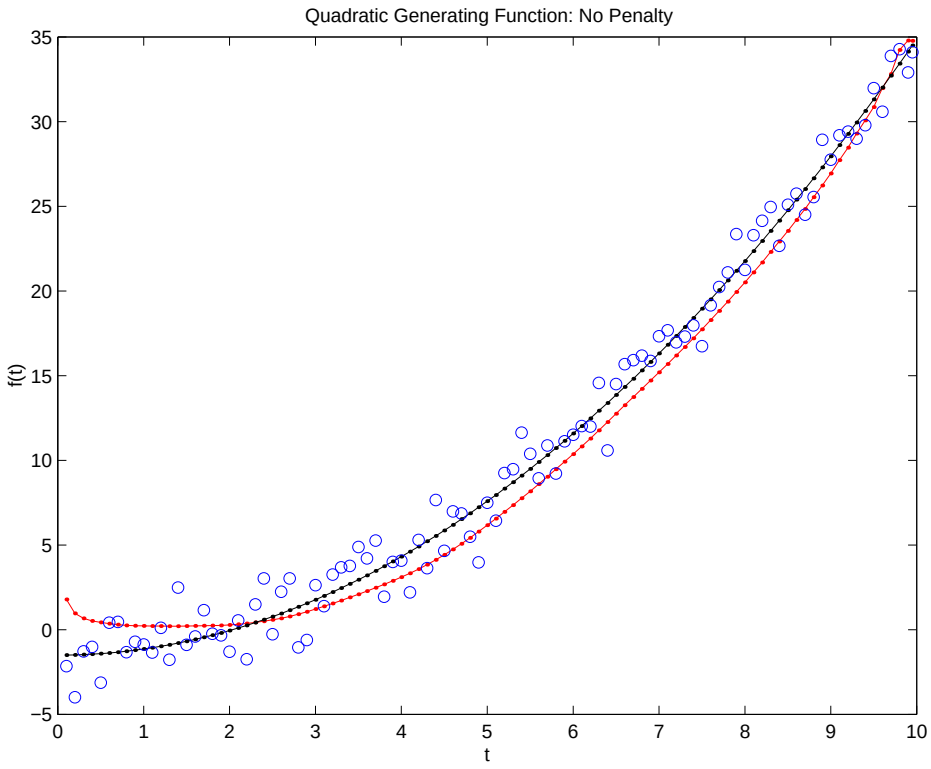

Figure 9: Estimates for Quadratic Generating Function (No Penalty)

penalty: DIC=-1.15, Normal(5,2) penalty: DIC=-1.00. So, the proposed penalty function does not show any improvement here either.

PENALTY	Data set 1	Data set 2	Data set 3	Data set 4	Data set 5
None	1140.3	348.6	372.7	248.7	321.3
Normal(5,2)	1106.8	351.3	371.7	255.5	292.4

Table 5: Summary of Prediction Evaluation for Quadratic Function

Finally, a sinusoidal generating function was used to create a sample

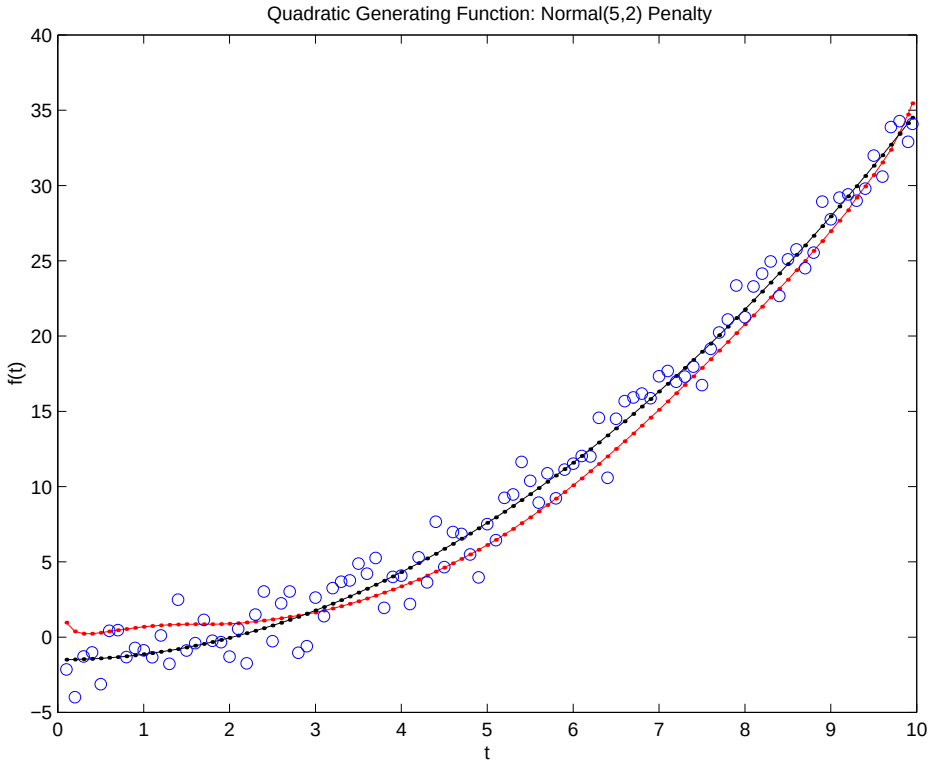

Figure 10: Estimates for Quadratic Generating Function (Normal(5,2) Penalty)

of 100 observations. This function is specified as f_4 above on the interval [0,10]. To challenge the method, an outlier was introduced at t=5.7 by adding 5.0 to the dependent variable. Thus, case study 4 is defined in the following way:

CASE STUDY 4

$f_3 = 8.2 + \sin 3\pi * t = $, n=100, σ^2=2.0,

The results from the modeling process are shown for the competing

approaches below.

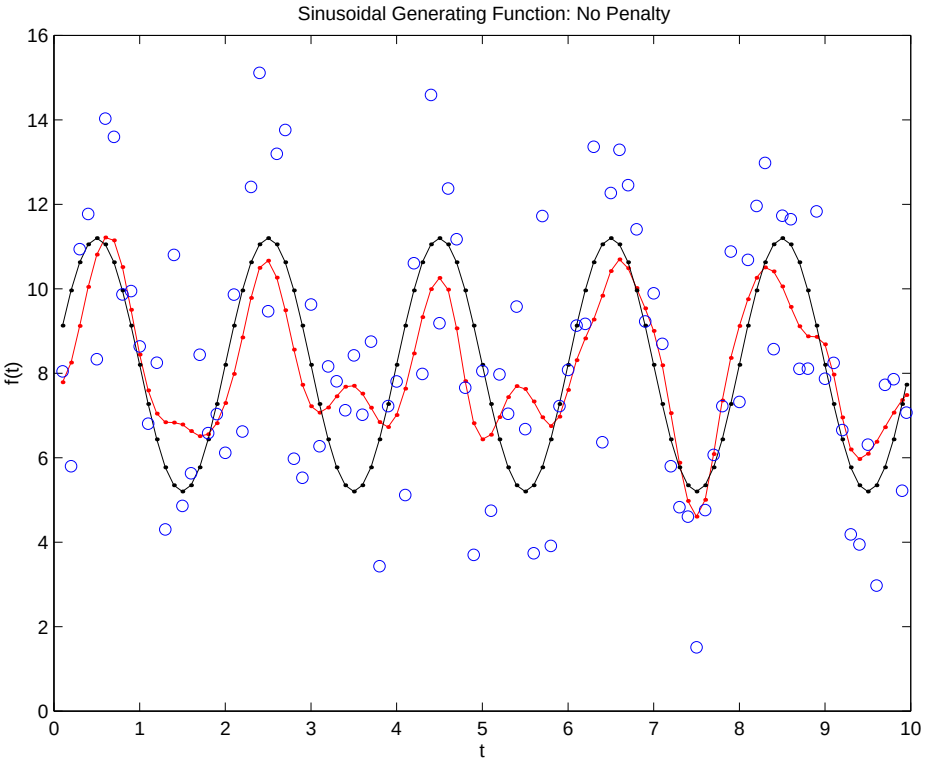

Figure 11: Estimates for Sinusoidal Generating Function (No Penalty)

It is evident that the use of the penalty function dampens the effect of the outlier. Once again, five additional data sets were generated using this underlying sinusoidal function. The outlier was omitted to examine

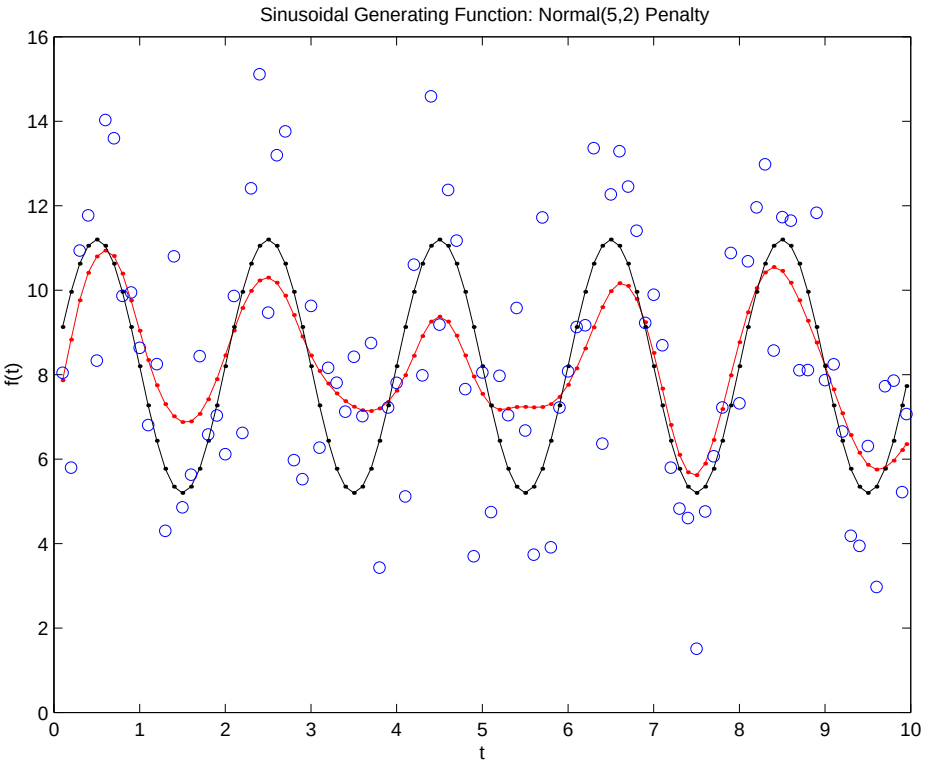

Figure 12: Estimates for Sinusoidal Generating Function (Normal(5,2) Penalty)

how well the penalty function mirrors the behavior of the true underlying function. Table 6 shows the prediction error summary and reveals that the R(k) has a slight edge for this experiment. The DIC statistics for the modeling approaches are: No Penalty: DIC=1.92, Normal(5,2) Penalty: DIC=-.85.

PENALTY	Data set 1	Data set 2	Data set 3	Data set 4	Data set 5
None	634.1	681.4	771.1	721.0	779.6
Normal(5,2)	648.4	710.7	725.9	664.5	719.3

Table 6: Summary of Prediction Evaluation for Sinusoidal Function

7 Conclusion

Markov Chains provide an extremely powerful framework for complex model formulation. Hierchical models, with the added capability of averaging estimates across models of varying dimensions, permit the practitioner to incorporate uncertainty about the model specification, in addition to the traditional uncertainty attached to the parameter estimates themselves. Those models which fit the sampled data better are reflected in the higher frequency that they occur in the sample selected from the mature Markov chain simulation. In this manner, these models contribute more to the final estimation of the model. Those model candidates which are less likely are not entirely excluded, but make relatively minor contributions to this final estimation. This increases the safeguard against the distortion in estimation due to the possible presence of outliers.

The use of appropriately designed penalty functions can be easily implemented using the design in this paper and give greater weight to the types of models that are preferable. The penalty function induces a prior distribution on some or all of the parameters in the model and need not

belong to any well-known class of distribution functions. More research is possible to determine whether a best penalty function can be found for certain classes of underlying generating functions. Also, investigation of the capabilities of this approach where the underlying dependent variable for the model arises from any probability distribution within the Exponential family may be of interest. Only Normally distributed responses have been considered in this paper. Finally, given that splines have been used for modeling purposes in this work, the incorporation of repeated knots in the knot vector will be the subject of future work. The option will permit the extension of the techniques used in this research to lower order splines in parts of the estimated functions, as well as possible discontinuities in this function.

References

[1] Akaike, H., A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC 19**, 716.723.

[2] Kullback, S. and Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86.

[3] Hurvich, C. M. and Tsai, C. L. Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

[4] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

[5] Eilers, P. H. C. and Marx, B. D. Flexible smoothing using B-splines and penalized likelihood. *Statistical Science*, **11**(2): 89-121.

[6] O'Sullivan, F., A Statistical Perspective on Ill-posed Inverse Problems. *Statistical Science*, **1** 502-527.

[7] D. Ruppert and R. Carroll, Spatially Adaptive Penalties for Spline Fitting. *Australian and New Zealand Journal of Statistics*, 1999, **42**, 205-224.

[8] B. Silverman and J. Friedman, Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, 1989, Vol. 31, No. 1.

[9] M. Lindstrom, Penalized Estimation of Free-Knot Splines. *Journal of Computational and Graphical Statistics*, 1999, Vol. 8, Issue 2, 333-352.

[10] D. L. B. Jupp, Approximation to Data by Splines with Free Knots. *SIAM Journal of Numerical Analysis*, 15, 328-343.

[11] D. Gamerman, *Markov Chain Monte Carlo Stochastic simulation for Bayesian Inference*, 1997, Chapman and Hall.

[12] N. Metropolis, A. W. Rosenbluth, M. N., A. H. Teller and E. Teller, Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 1953, **21**, 1097-92.

[13] W. K. Hasting, W. K., Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970, **57**, 97-109.

[14] Peter Green, Reversible jump Markov chain Monte carlo computation and Bayesian model determination. *Biometrika*, 1995, **82**, pp. 711-32.

[15] Charles Chui, Wavelets: A Mathematical Tool for Signal Analysis *SIAM: Monographs on Mathematical Modeling and Computation*, 1997, p. 142.

[16] C. Biller, Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models. *Journal of Computational and Graphical Statistics*, 2000, Vol. 9, Issue 1, 122-140.

[17] D. Draper, Assessment and Propagation of Model Uncertainty, *Journal of the Royal Statistical Society, Series B*, 1995 **57,** 45-70.

[18] D. Gamerman, Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 1997, 7, 57-68.

[19] T. Lyche and K. Strom, Knot Insertion for Natural Splines, *Annals of Numerical Mathematics*, 1996, 3: 221-246.

[20] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, Bayesian Measures of Model Complexity and Fit, *Journal of the Royal Statistical Society, Series B*, 2002 64, **583-640.**

-