Second Place	
Authors	Kristin Mara (kmara08@winona.edu) Department of Mathematics & Statistics, Winona State University Rosie Roessel (ror213@lehigh.edu) Department of Mathematics, Lehigh University Samantha Meadows (meado1sl@cmich.edu) Department of Mathematics, Central Michigan University
Title	Semiparametric Regression for Measurement of Parts Data
Type of Project	REU Project
Instructor Sponsor	Chin-I Cheng (cheng3c@cmich.edu) Department of Mathematics, Central Michigan University
Abstract	One of the approaches to model the smooth function in a nonparametric model is to approximate it by adopting adequate basis functions. In this research project, we will approximate the smooth function by a truncated polynomial basis with degree 2, which contains the polynomial basis and the splines constructed by knots. After we fix the number of knots, the function is estimated by well-known methods such as ordinary least squares, penalized spline regression and linear mixed model. We propose our version of Bayesian penalized spline, which provides comparable results. The prior distribution is chosen to be "objective" so it will minimize the influence to the posterior distribution and maintain the advantages Bayesian statistics provided. The non-informative Jeffreys prior is adopted for the polynomial basis and the variance component in the model. The prior for the splines constructed by knots is elicited from the penalty term in the penalized likelihood. To ensure the posterior distributions are proper, we have to use an informative prior on the smoothing parameter. To achieve the goal of having an "objective" prior for the smoothing parameter, we will use the effective df_{fit} (degrees of freedom for fit) to determine the hyperparameters in the prior distribution. After we fit a nonparametric model, we look at a semiparametric model. This semiparametric model combines our nonparametric model with a categorical variable. We use the Akaike Information Criterion (AIC) to compare those methods proposed through a simulated data set and a manufacturing data set.

Semiparametric Regression for Measurement of Parts Data

Abstract

One of the approaches to model the smooth function in a nonparametric model is to approximate it by adopting adequate basis functions. In this research project, we will approximate the smooth function by a truncated polynomial basis with degree 2, which contains the polynomial basis and the splines constructed by knots. After we fix the number of knots, the function is estimated by well-known methods such as ordinary least squares, penalized spline regression and linear mixed model. We propose our version of Bayesian penalized spline, which provides comparable results. The prior distribution is chosen to be "objective" so it will minimize the influence to the posterior distribution and maintain the advantages Bayesian statistics provided. The non-informative Jeffreys prior is adopted for the polynomial basis and the variance component in the model. The prior for the splines constructed by knots is elicited from the penalty term in the penalized likelihood. To ensure the posterior distributions are proper, we have to use an informative prior on the smoothing parameter. To achieve the goal of having an "objective" prior for the smoothing parameter, we will use the effective df_{fit} (degrees of freedom for fit) to determine the hyperparameters in the prior distribution. After we fit a nonparametric model, we look at a semiparametric model. This semiparametric model combines our nonparametric model with a categorical variable. We use the Akaike Information Criterion (AIC) to compare those methods proposed through a simulated data set and a manufacturing data set.

1 Introduction

For our project, we are looking at different regression methods and how they perform compared to each other. Regression is widely used for prediction and to help find relationships between different variables. The three types of regression methods we are looking at are ordinary least squares, penalized spline, and linear mixed model regression. After fitting these models, we want to see if we can use Bayesian estimation to find a model that provides comparable results.

Before we began our study, we had to choose if we were going to use nonparametric, semiparametric, or parametric regression to model our data set. In nonparametric regression, the data is not required to fit any particular structure, which is useful when we know little or nothing about the true function. A parametric model assumes that our data follows a probability distribution and can be modeled by specific structure. Semiparametric regression has a combination of parametric and nonparametric parts. In this type of model, you must specify the structure of the parametric term, which is usually assumed linearly related to your response.

The rest of the paper's layout is as follows: section 2 discuss the nonparametric model for a continuous variable estimated by ordinary least square regression, penalized spline and linear mixed model. Section 3 describe the semiparametric model and its estimations based on ordinary least square regression, penalized spline and linear mixed model. Section 4 propose our version of Bayesian penalized spline model. Section 5 and 6 evaluates the approaches using simulation example and manufacturing data, respectively. We summarize our conclusion in section 7.

2 Nonparametric Model

First, we started with the nonparametric model with a continuous variable x,

$$y_i = f(x_i) + \varepsilon_i, \tag{1}$$

where y is the response variable that we are interested in. The function $f(x)$ is a smooth function which can be linear but does not have to be. The error terms, denoted ε_i, follow a normal distribution with a mean of zero and a variance of σ^2. Now we define

$$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)', \tag{2}$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'. \tag{3}$$

Therefore, $\boldsymbol{y}$ is an $n \times 1$ vector, meaning that we have n data points. We will be using a truncated power basis to approximate $f(x)$ (Ruppert, Wand, & Carroll, 2003), which is

$$f(x_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \sum_{r=1}^{k} \beta_{p+r}(x_i - K_r)_+^p,$$

where p is the degree of the polynomial basis, k is the number of knots in the model, and K is the value at which each knot is placed; so K_r would be the value of the r^{th} knot. A knot is a point on the graph where two line segments meet, but they do not necessarily form a straight line. These knots allow the regression curve to fit your data better.

To simplify our equation we can use matrix notation, making it

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

where

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p+k})' \tag{5}$$

is a $(p + k + 1) \times 1$ vector of regression coefficients and $\boldsymbol{X}$ is an $n \times (p + k + 1)$ matrix of a predictor variable and its knots defined as

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^p & (x_1 - K_1)_+^p & \cdots & (x_1 - K_k)_+^p \\ 1 & x_2 & \cdots & x_2^p & (x_2 - K_1)_+^p & \cdots & (x_2 - K_k)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p & (x_n - K_1)_+^p & \cdots & (x_n - K_k)_+^p \end{pmatrix}. \tag{6}$$

In this research project, we will be looking at ordinary least squares, penalized spline, linear mixed models, and Bayesian penalized spline regression. We will first be going through the basic set up for these methods, then we will look at the methods on simulated data, and finally we will be looking at all the methods on our manufacturing data.

2.1 Ordinary Least Squares Regression

The first method we will look at is the ordinary least squares regression. This model finds the function that fits the data best while minimizing the distance between the data points and the function. The model to fit is equation (4), where $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ are defined by equations (2), (6), (5), and (3). To find $\hat{\boldsymbol{\beta}}$, the estimate of $\boldsymbol{\beta}$, we want to minimize

$$\left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \right\|^2,$$

giving us

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{7}$$

Once we have $\hat{\boldsymbol{\beta}}$, we can compute $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and obtained $R^2_{adj.}$ to see if our model is a good fit for the data, where

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

$$R^2_{adj.} = 1 - \left(\frac{n-1}{n-(p+k+1)}\right)\left(1 - R^2\right).$$

In this project, we used the $R^2_{adj.}$ to select the number of knots. The models, which vary in number of knots, with the highest $R^2_{adj.}$ would be selected.

In this regression method, the df_{fit} (degrees of freedom for fit), or number of fitted parameters, depends on the trace of the hat matrix.

$$tr(\boldsymbol{H}) = tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = tr((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X})$$

$$= tr(\boldsymbol{I}_{df_{fit}}) = df_{fit} = p + k + 1$$

where the hat matrix $\boldsymbol{H}$ is defined by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'. \tag{8}$$

2.2 Penalized Spline Regression

The penalized spline regression method penalizes over-fits of your data, because a model that is over-fit will not necessarily do well when predicting. This model is to minimize

$$\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 + \lambda\boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta}. \tag{9}$$

where $\boldsymbol{D} = \begin{bmatrix} 0_{(p+1)\times(p+1)} & 0_{(p+1)\times k} \\ 0_{k\times(p+1)} & I_{k\times k} \end{bmatrix} = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix}$, and λ is the smoothing parameter (Ruppert et al., 2003).

When λ is small, the model will be rough or bumpy, and as λ gets larger the model starts to look more like a straight line.

The estimate for $\boldsymbol{\beta}$ in equation (9) is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{10}$$

To find $\hat{\boldsymbol{\beta}}$, we estimate λ by generalized cross validation (GCV). The computation of GCV depends on the smoothing matrix, which for penalized spline is

$$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}'. \tag{11}$$

The df_{fit} is based on the trace of the smoothing matrix, which can be manipulated by the selection of λ.

$$df_{fit} = tr(\boldsymbol{S}_\lambda) = tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}') = tr((\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}'\boldsymbol{X}).$$

If $\lambda \to 0$, then $tr(\boldsymbol{S}_\lambda) = p + k + 1$.

If $\lambda \to \infty$, then $tr(\boldsymbol{S}_\lambda) = p + 1$.

GCV is based on the average of the diagonals of the smoothing matrix,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{S}_{\lambda,ii} = \frac{1}{n} tr(\boldsymbol{S}_\lambda) = \frac{1}{n} df_{fit}$$

where $\boldsymbol{S}_{\lambda,ii}$ denotes the i^{th} diagonal of the smoothing matrix. Therefore,

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\left(1 - \frac{1}{n} df_{fit}\right)^2}. \tag{12}$$

We always want to minimize equation (12), and the λ associated with this minimum is the λ that we want to use in our model. Once we have our λ we can go back and find our $\hat{\boldsymbol{\beta}}$ in equation (10).

2.3 Linear Mixed Model Regression

The third regression method we looked at was linear mixed model regression, which incorporates both fixed effects along with random effects in the model. This method is frequently used when dealing with repeated measures data. The linear mixed model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_x + \boldsymbol{Z}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{13}$$

where $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are defined in equations (2) and (3). We now have $\boldsymbol{X}_1$ and $\boldsymbol{\beta}_x$ looking at the polynomial basis part of the model, and $\boldsymbol{Z}$ and $\boldsymbol{\mu}$ are looking at the spline constructed by knots part of the model. In this case $\boldsymbol{X}_1$ is an $n \times (p+1)$ matrix, $\boldsymbol{Z}$ is an $n \times k$ matrix, $\boldsymbol{\beta}_x$ is a $(p+1) \times 1$ vector, and $\boldsymbol{\mu}$ is a $k \times 1$ vector. Both $\boldsymbol{\beta}_x$ and $\boldsymbol{\mu}$ are vectors of regression coefficients. The $\boldsymbol{X}_1$, $\boldsymbol{Z}$, $\boldsymbol{\beta}_x$ and $\boldsymbol{\mu}$ are defined as,

$$\boldsymbol{X}_1 = \begin{pmatrix} 1 & x_1 & \cdots & x_1^p \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p \end{pmatrix}, \tag{14}$$

$$\boldsymbol{Z} = \begin{pmatrix} (x_1 - K_1)_+^p & (x_1 - K_2)_+^p & \cdots & (x_1 - K_k)_+^p \\ (x_2 - K_1)_+^p & (x_2 - K_2)_+^p & \cdots & (x_2 - K_k)_+^p \\ \vdots & \vdots & \ddots & \vdots \\ (x_n - K_1)_+^p & (x_n - K_2)_+^p & \cdots & (x_n - K_k)_+^p \end{pmatrix}, \tag{15}$$

$$\boldsymbol{\beta}_x = (\beta_0, \beta_1, \cdots, \beta_p)', \tag{16}$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_k)'. \tag{17}$$

We are trying to minimize the penalized likelihood:

$$\left\|\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_x - \boldsymbol{Z}\boldsymbol{\mu}\right\|^2 + \lambda\left\|\boldsymbol{\mu}\right\|^2.$$

Thus, we can find $\hat{\boldsymbol{\beta}}_x$ and $\hat{\boldsymbol{\mu}}$ by

$$\hat{\boldsymbol{\beta}}_x = (\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \tag{18}$$

$$\hat{\boldsymbol{\mu}} = \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_x), \tag{19}$$

where $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{Z}' + \lambda\boldsymbol{I}$ (Ruppert et al., 2003).

Once again in order to find $\hat{\boldsymbol{\beta}}_x$ and $\hat{\boldsymbol{\mu}}$ from equations (18) and (19), we have to find λ by minimizing equation (12).

3 Semiparametric Model

After adding a categorical variable x_2 to the nonparametric model in equation (1), the semiparametric model is

$$y_i = f(x_i) + \beta_2 x_{2i} + \varepsilon_i,$$

where $f(x)$ is again approximated by the truncated polynomial basis. The categorical variable will have $j+1$ levels, so we will need to use j dummy variables in the model. Dummy variables are vectors of zeros and ones, with the zeros representing when a certain level of the categorical variable is absent and the ones indicating presence. Our model is now

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \textstyle\sum_{r=1}^{k} \beta_{p+r}(x_i - K_r)_+^p + \textstyle\sum_{l=1}^{j} \beta_{p+k+l} M_{il} + \varepsilon_i$$

where M_l represents the different dummy variables, and j is the number of dummy variables in the model.

In matrix notation we have

$$\boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_x + \boldsymbol{X}_2 \boldsymbol{\beta}_{dum} + \boldsymbol{Z}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{20}$$

where $\boldsymbol{X}_1, \boldsymbol{Z}, \boldsymbol{\beta}_x$ and $\boldsymbol{\mu}$ are defined in equations (14), (15), (16), and (17),

$$\boldsymbol{X}_2 = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1j} \\ M_{21} & M_{22} & \cdots & M_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nj} \end{pmatrix},$$

$$\text{and } \boldsymbol{\beta}_{dum} = (\beta_{p+k+1}, \cdots, \beta_{p+k+j})'.$$

The estimates of $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^{'}, \boldsymbol{\beta}_{dum}^{'}, \boldsymbol{\mu}^{'})'$ using **ordinary least squares** is equation (7) with $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Z})$.

In **penalized spline regression**, the estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^{'}, \boldsymbol{\beta}_{dum}^{'}, \boldsymbol{\mu}^{'})'$ is obtained by minimizing expression (9), where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Z})$ and the λ that is associated with the minimum GCV.

The **linear mixed model** is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}_1 \boldsymbol{\beta}_x + \boldsymbol{X}_2 \boldsymbol{\beta}_{dum} + \boldsymbol{Z}\boldsymbol{\mu} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\mu} + \boldsymbol{\varepsilon} \end{aligned}$$

where the $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_x', \boldsymbol{\beta}_{dum}')'$. The $\boldsymbol{Z}$ and $\boldsymbol{\mu}$ were defined in equations (15) and (17). The $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ will be estimated by equations (18) and (19) with $\boldsymbol{X}_1$ replaced by $\boldsymbol{X}$ and $\boldsymbol{\beta}_x$ replaced by $\boldsymbol{\beta}$. The required λ is obtained by minimizing GCV.

4 Bayesian Penalized Spline Regression

4.1 Bayesian Statistics

Another way to model the data is to use Bayesian statistics. Bayesian statistics express the uncertainty of an unknown parameter θ in terms of probability. In Bayesian statistics, you start with a prior $\pi(\theta)$, which is based on the current knowledge you have. You then use your observed data, which follows a likelihood distribution $f(y|\theta)$ to “update” your knowledge. This “updated” knowledge is called the posterior distribution $f(\theta|y)$, which can be represented by Bayes Theorem:

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int\limits_{\mathbb{R}} f(y|\theta)\pi(\theta)d\theta}, \text{ where } \mathbb{R} \text{ is the support of } \theta,$$

$$p(\theta|y) \propto f(y|\theta)\pi(\theta),$$

so the posterior distribution $p(\theta|y)$ is proportional to the product of the likelihood and the prior.

4.2 Nonparametric Model with One Continuous Variable

4.2.1 Prior Selection

Using a Bayesian model, all the parameters are considered to be random. Because of this we need to have a prior distribution for each parameter. In some situations there is information we can use from past experiments, from which we can make an informative prior for the parameter. In other situations very little is known about a parameter so a non-informative prior is formed. These non-informative priors are almost always improper which means that $\int\limits_{\mathbb{R}} f(\theta)d\theta \neq 1$, so they have to be chosen carefully to guarantee proper posterior distributions.

In this study, we are aiming to use "objective" priors, which use minimum prior information to derive the posterior distribution. For $\boldsymbol{\beta}_x$ and the variance of $\boldsymbol{\varepsilon}$, σ^2, in equation (13), we used non-informative Jeffreys prior

$$\begin{aligned} \pi(\boldsymbol{\beta}_x) &\propto 1, \\ \pi(\sigma^2) &\propto \frac{1}{\sigma^2}. \end{aligned}$$

We then looked at finding our prior for $\boldsymbol{\mu}$. Notice that minimizing

$$\left\|\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_x - \boldsymbol{Z}\boldsymbol{\mu}\right\|^2 + \lambda \left\|\boldsymbol{\mu}\right\|^2$$

is the same as maximizing

$$\tfrac{-1}{2\sigma^2} \left\|\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_x - \boldsymbol{Z}\boldsymbol{\mu}\right\|^2 - \tfrac{\lambda}{2\sigma^2} \left\|\boldsymbol{\mu}\right\|^2.$$

Looking at the above penalized likelihood, we can see that the penalty term looks like a normally distributed for $\boldsymbol{\mu}$ with a mean of 0 and a variance of $\frac{\sigma^2}{\lambda}$ (Kimeldorf & Wahba, 1971). Using this, we found

$$\pi(\boldsymbol{\mu}) \sim N\left(0, \tfrac{\sigma^2}{\lambda}\boldsymbol{I}_k\right).$$

Finally we needed our prior for λ. We needed this prior to be a proper prior so that we would end up with a proper posterior distribution (Sun & Speckman, 2008). Thus we proposed

$$\pi(\lambda) \sim Gamma\left(\frac{1}{2}, 2b\right), \tag{21}$$

where $\pi(\lambda) = \frac{1}{(2b)^{1/2}\Gamma(1/2)}\lambda^{-1/2}e^{-\lambda/2b}$.

We will discuss how we chose our b in the "Determining Hyperparameter b" subsection.

4.2.2 Posterior Derived

After assigning priors for each parameter, we will have to derive the posterior distributions,

$$f_3(\theta_1|y) = \int_{\theta_2} \cdots \int_{\theta_w} f_2(\theta_1|y, \theta_2, \cdots, \theta_w) d\theta_2 \cdots d\theta_w.$$

To avoid messy integration to get our posterior, we will generate random samples from the full conditionals through Gibbs sampler (Gelman, Carlin, Stern, & Rubin, 2004). The random samples created will form the posterior distributions.

Full conditional:

$$f_2(\theta_1|y, \theta_2, \cdots, \theta_w) \propto f_1(y|\theta_1, \cdots, \theta_w)\pi_1(\theta_1)\pi_2(\theta_2)\cdots\pi_w(\theta_w).$$

To improve the mix of Gibbs sampler between $\boldsymbol{\beta}_x$ and $\boldsymbol{\mu}$, we adopted block sampling (Gelman et al., 2004). Both $\boldsymbol{\beta}_x$ and $\boldsymbol{\mu}$ will be sampled simultaneously through multivariate normal distribution. This approach allows for less iterations due to rapid convergence. Using the priors from section 4.2.1 and setting $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{Z})$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_x', \boldsymbol{\mu}')'$, the full conditionals are:

$$\boldsymbol{\beta}|\sigma^2, \lambda, \boldsymbol{y} \sim N\left((\lambda\boldsymbol{D} + \boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, (\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} + \frac{\lambda}{\sigma^2}\boldsymbol{D})^{-1}\right) \tag{22}$$

$$\lambda|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y} \sim Gamma\left(\frac{k+1}{2}, \left(\frac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta} + \frac{1}{2b}\right)^{-1}\right) \tag{23}$$

$$\sigma^2|\boldsymbol{\beta}, \lambda, \boldsymbol{y} \sim Inv.Gamma\left(\frac{n+k}{2}, \frac{1}{2}\left((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta}\right)\right) \tag{24}$$

4.2.3 Determining Hyperparameter b

To determine the hyperparameter b for $\pi(\lambda)$ in equation (21), we start with choosing a prior df_{fit} that we think our data should have. The smoothing matrix for Bayesian penalized spline is $\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}'$. Then

$$df_{fit} = tr(\boldsymbol{S}_\lambda) = tr\left(\left(\boldsymbol{I} + \lambda\boldsymbol{D}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)^{-1}\right) = tr\left((\boldsymbol{I} + \lambda\boldsymbol{v})^{-1}\right) = \sum_{i=1}^{p+k+1} \frac{1}{1 + \lambda v_i} \tag{25}$$

where $\boldsymbol{v}$ is a $(p+k+1) \times 1$ vector of the eigenvalues for $\boldsymbol{D}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (Hastie, Tibshirani, & Friedman, 2001).

For a given df_{fit}, we can solve for λ in equation (25). We take this $\hat{\lambda}$ as the median for the prior $\pi(\lambda)$, a skewed distribution, to specify b. The prior $\pi(\lambda)$ can be considered as a scaled χ^2 distribution with 1 df. The median of $\pi(\lambda)$ is $b\chi^2_{(1)} = .455b$. We can solve for b by $\hat{\lambda} = .455b$.

4.3 Semiparametric Model with One Continuous and One Discrete Variable

The Bayesian penalized spline for the semiparametric model in equation (20) minimizes

$$\left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_2\boldsymbol{\beta}_{dum}\right|\right| + \lambda\boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta}, \tag{26}$$

$$\text{where } \boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{Z}),\ \boldsymbol{\beta} = (\boldsymbol{\beta}_x', \boldsymbol{\mu}')' \text{ and } \boldsymbol{D} = \begin{bmatrix} 0_{(p+1)\times(p+1)} & 0_{(p+1)\times k} \\ 0_{k\times(p+1)} & I_{k\times k} \end{bmatrix}.$$

We proposed both a non-informative and an informative prior for $\boldsymbol{\beta}_{dum}$.

4.3.1 Non-informative Prior for $\boldsymbol{\beta}_{dum}$

Our non-informative prior for $\boldsymbol{\beta}_{dum}$ is

$$\pi(\boldsymbol{\beta}_{dum}) \propto 1.$$

Using this non-informative prior, we got the following full conditionals,

where $\boldsymbol{X}_* = (\boldsymbol{X}_2, \boldsymbol{X})$, $\boldsymbol{\beta}_* = (\boldsymbol{\beta}'_{dum}, \boldsymbol{\beta}')'$, and $\boldsymbol{D}_* = \begin{bmatrix} 0_{(p+j+1)\times(p+j+1)} & 0_{(p+j+1)\times k} \\ 0_{k\times(p+j+1)} & I_{k\times k} \end{bmatrix}$.

$$\boldsymbol{\beta}_*|\sigma^2, \lambda, \boldsymbol{y} \sim N\left((\lambda \boldsymbol{D}_* + \boldsymbol{X}'_*\boldsymbol{X}_*)^{-1}\boldsymbol{X}'_*\boldsymbol{y}, (\tfrac{1}{\sigma^2}\boldsymbol{X}'_*\boldsymbol{X}_* + \tfrac{\lambda}{\sigma^2}\boldsymbol{D}_*)^{-1}\right),$$

$$\lambda|\boldsymbol{\beta}_*, \sigma^2, \boldsymbol{y} \sim Gamma\left(\tfrac{k+1}{2}, \left(\tfrac{1}{\sigma^2}\boldsymbol{\beta}'_*\boldsymbol{D}_*\boldsymbol{\beta}_* + \tfrac{1}{2b}\right)^{-1}\right),$$

$$\sigma^2|\boldsymbol{\beta}_*, \lambda, \boldsymbol{y} \sim Inv.Gamma\left(\tfrac{n+k}{2}, \tfrac{1}{2}\left((\boldsymbol{y} - \boldsymbol{X}_*\boldsymbol{\beta}_*)'(\boldsymbol{y} - \boldsymbol{X}_*\boldsymbol{\beta}_*) + \lambda\boldsymbol{\beta}'_*\boldsymbol{D}_*\boldsymbol{\beta}_*\right)\right).$$

4.3.2 Informative Prior for $\boldsymbol{\beta}_{dum}$

Our informative prior for $\boldsymbol{\beta}_{dum}$ is

$$\pi(\boldsymbol{\beta}_{dum}) \sim N\left(0, \tfrac{\sigma^2}{\lambda_1}\boldsymbol{I}_j\right), \text{ where } \pi(\lambda_1) \sim Gamma\left(\tfrac{1}{2}, 2\right).$$

From the penalty term in equation (26), we elicit the prior for $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}) \sim N\left(0, \tfrac{\sigma^2}{\lambda}\boldsymbol{D}^{-1}\right).$$

And from section 4.2.1

$$\pi(\lambda) \sim Gamma\left(\tfrac{1}{2}, 2b\right) \text{ and } \pi(\sigma^2) \propto \tfrac{1}{\sigma^2}.$$

Using these priors, we had the following full conditionals, where

$$\lambda_1|\lambda, \boldsymbol{\beta}, \boldsymbol{\beta}_{dum}, \sigma^2, \boldsymbol{y} \sim Gamma\left(\tfrac{j+1}{2}, \left(\tfrac{1}{\sigma^2}\boldsymbol{\beta}'_{dum}\boldsymbol{\beta}_{dum} + \tfrac{1}{2}\right)^{-1}\right)$$

$$\lambda|\lambda_1, \boldsymbol{\beta}, \boldsymbol{\beta}_{dum}, \sigma^2, \boldsymbol{y} \sim Gamma\left(\tfrac{k+1}{2}, \left(\tfrac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta} + \tfrac{1}{2b}\right)^{-1}\right)$$

$$\boldsymbol{\beta}|\lambda_1, \lambda, \boldsymbol{\beta}_{dum}, \sigma^2, \boldsymbol{y} \sim N\left((\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}_2\boldsymbol{\beta}_{dum}), (\tfrac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} + \tfrac{\lambda}{\sigma^2}\boldsymbol{D})^{-1}\right)$$

$$\boldsymbol{\beta}_{dum}|\lambda_1, \lambda, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y} \sim N\left((\boldsymbol{X}'_2\boldsymbol{X}_2 + \lambda_1\boldsymbol{I}_j)^{-1}(\boldsymbol{X}'_2\boldsymbol{y} - \boldsymbol{X}'_2\boldsymbol{X}\boldsymbol{\beta}), (\tfrac{1}{\sigma^2}\boldsymbol{X}'_2\boldsymbol{X}_2 + \tfrac{\lambda_1}{\sigma^2}\boldsymbol{I}_j)^{-1}\right)$$

$$\sigma^2|\lambda_1, \lambda, \boldsymbol{\beta}, \boldsymbol{\beta}_{dum}, \boldsymbol{y} \sim Inv.Gamma\left(\tfrac{n+k+j}{2}, \tfrac{1}{2}\left(\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_2\boldsymbol{\beta}_{dum}\right\|^2 + \lambda\boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta} + \lambda_1\left\|\boldsymbol{\beta}_{dum}\right\|^2\right)\right)$$

4.4 Model Selection

We compare all the models using the Akaike Information Criterion (AIC). The model with the lowest AIC will be selected.

$$AIC(\lambda) = log\left(\sum_{i=1}^{n}(y_i - \hat{y_i})^2\right) + \frac{2df_{fit}}{n} \tag{27}$$

To compute the AIC for each of our approaches, we need the df_{fit}, so we must know the smoothing matrix for the approach.

The smoothing matrices for the ordinary least squares and penalized spline methods are defined in equations (8) and (11).

The linear mixed model smoothing matrix is

$$\begin{gathered}\boldsymbol{S}_\lambda = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1} + \boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1} \\ -\boldsymbol{Z}\boldsymbol{Z}'\Sigma^{-1}\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{\Sigma}^{-1} \\ \text{where } \boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{Z}' + \lambda\boldsymbol{I}.\end{gathered}$$

The Bayesian penalized spline smoothing matrix is

$$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}'.$$

While the penalized spline and Bayesian penalized spline smoothing matrices appear identical, the Bayesian approach uses the posterior mean for λ, where the penalized spline approach uses the λ we found from minimizing the GCV in equation (12).

5 Simulation Study

5.1 Nonparametric Model with One Continuous Variable

To gain understanding of the different models, we used R to create a simulated data set for a nonparametric model,

$$y_i = f(time_i) + \varepsilon_i = 1 + 3 * sin(2 * time_i * \pi - \pi) + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, 4)$ with *time* being specified by twenty even increments between zero and one. This function for y generated 20 normal random samples, which is plotted in figure 1.

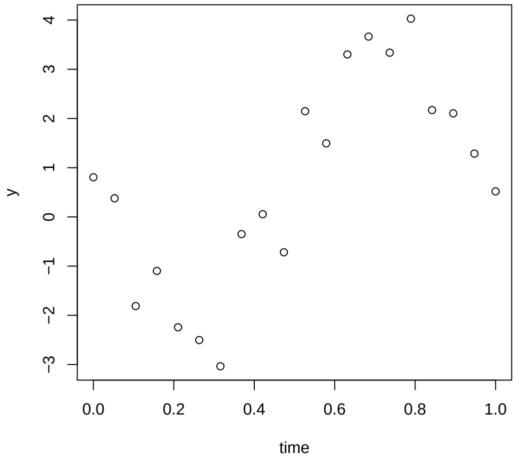

Figure 1: Simulated data points

Figure 1 suggested that we set the polynomial basis power $p = 2$ with the number of knots $k = 3$. This should be flexible enough to fit the data with possible curvature.

5.1.1 Ordinary Least Squares Model

The ordinary least squares regression is estimated by equation (7) with determined structure of $p = 2$ and $k = 3$.

5.1.2 Penalized Spline Model

The penalized spline method is more flexible than the ordinary least squares method. It will smooth the rough fits by ordinary least square if necessary. The estimate by penalized spine is in equation (10). We ran a loop of 100 evenly-spaced λ between 0 and .001. To determine the best λ, we checked to see which λ in this loop produced the lowest GCV. Using the corresponding λ, we can find the df_{fit}, the $\hat{\boldsymbol{\beta}}$, and thus our fit. Our minimum GCV was 14.7001, whose $\lambda = 4.4444 \times 10^{-4}$ and $df_{fit} = 4.7667$.

5.1.3 Linear Mixed Model

Linear mixed models consider the polynomial basis as the fixed effect and the spline constructed by knots as the random effect. The $\hat{\boldsymbol{\beta}}_x$ in equation (18) and $\hat{\boldsymbol{\mu}}$ in equation (19) are estimated with $\lambda = 4.4444 \times 10^{-4}$ which minimized the GCV.

5.1.4 Bayesian Penalized Spline Model

The simulated data set plotted in figure 1, suggested that we would want to fit at least a cubic function to our model. We choose the df_{fit} to be around 4 to 5 to provide sufficient flexibility to the fit. Following the discussion in "Determining Hyperparameter b", several prior df_{fit} are considered and the corresponding λ and b are listed in table 1. We examined the fits based on the b listed in table 1 and found that the fits are pretty robust to the choice of b. We choose the b= .0006. We then use R to compute 10,000 iterations to generate random samples using Gibbs sampler (Gelman et al., 2004) for $\boldsymbol{\beta}$, λ, and σ^2 from their full conditionals in equations (22), (23) and (24). We then applied a burn-in, which chopped off the first 2,000 iterations, getting rid of the influence of our initial values for the parameters. The convergence was rapid as showed in figure 2. Those random samples were used to create the posterior distributions. Doing so, we were able to determine the posterior means of both $\hat{\boldsymbol{\beta}}_x$ and $\hat{\boldsymbol{\mu}}$ to compute a fit.

Prior df_{fit}	λ	b
3.5	.0089	.0196
4	.0023	.0051
4.5	.0008	.0017
5	.0003	.0006

Table 1: Different λ and b values based on chosen prior df_{fit}

5.1.5 Model Comparison for Simulated Data with *time* only

Finally, we can compare the ordinary least squares, penalized spline, linear mixed model, and Bayesian penalized spline fits. Penalized spline and linear mixed model are using their best λ fits, for which is the same value in this case, so the two fits are the same. We can see this in figure 3. The fit by penalized spline was based on penalizing the splines constructed by knots in equation (9), which is similar to the linear mixed model that has the splines constructed by knots fit as random effect. Therefore, there is no surprise that both fits provided similar results. Then, we computed the AIC using equation (27) for approaches proposed.

For this simulated data, we should use the Bayesian penalized spline approach since it has the lowest AIC showed in table 2. Note that the penalized spline and linear mixed model approaches gave the same AIC due to the same fits provided.

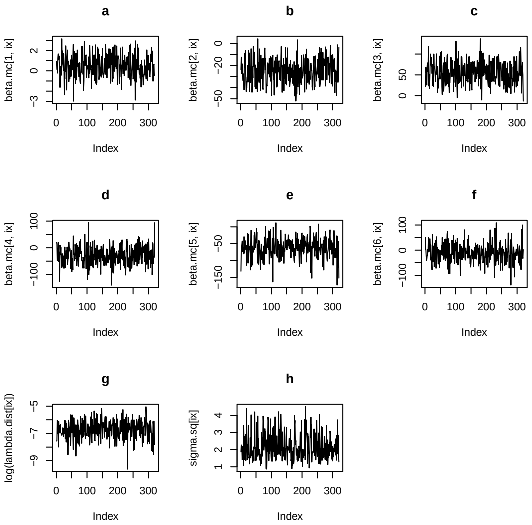

Figure 2: Trace plots of (a): β_1, (b): β_2, (c): β_3, (d): μ_1, (e): μ_2, (f): μ_3, (g):λ, and (h): σ^2

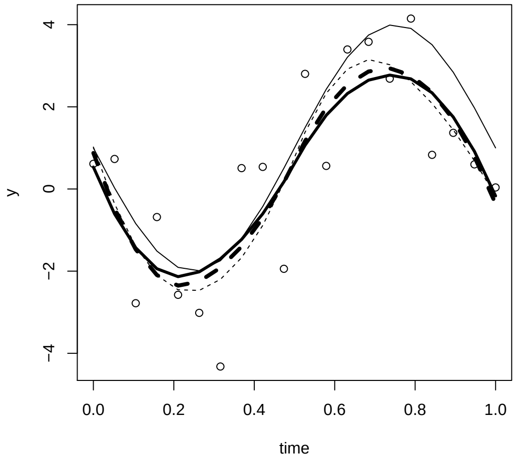

Figure 3: Plot of $f(x)$ used to generate the data (solid), ordinary least squares (dashed), penalized spline (thick dashed line), linear mixed model (thick dashed line), and Bayesian penalized spline (thick solid)

Approach	AIC
Ordinary least squares	4.031754
Penalized spline	3.937102
Linear mixed model	3.937102
Bayesian penalized spline	3.807775

Table 2: AIC for the different methods

If the Bayesian penalized spline yields a similar result to the penalized spline and linear mixed model methods, then why would we go through all the trouble to use it? First, if prior information is available, the Bayesian framework will provide a way to incorporate this information. Further, using the frequentist approach, we estimated λ using the minimum GCV. This gave us a single value estimate for the λ. In the Bayesian approach, we get an entire distribution for λ, which will easily extend the statistical inference to hypothesis testing. As a result, the Bayesian approach is not as susceptible to outlier values, like those in the simulated data when *time* is .3 and when *time* is .8.

5.2 Semiparametric Model with a Continuous and a Discrete Variable

We used R to simulate a data set that looked at adding a categorical variable with three levels, making our model semiparametric as in equation (20). They are different only in intercept. Let

$$\begin{gathered}
y_{1i} = f(time_i, op_{1i}) + \varepsilon_i = 1 + 3 * sin(2 * time_i * \pi - \pi) + \varepsilon_i, \\
y_{2i} = f(time_i, op_{2i}) + \varepsilon_i = 2 + 3 * sin(2 * time_i * \pi - \pi) + \varepsilon_i, \\
y_{3i} = f(time_i, op_{3i}) + \varepsilon_i = 3 + 3 * sin(2 * time_i * \pi - \pi) + \varepsilon_i, \\
\text{and } \boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3)
\end{gathered}$$

where $\varepsilon_i \sim N(0,1)$, and *time* is specified by twenty even increments between zero and one. The response for the first operator is given in $\boldsymbol{y}_1$, for the second operator the response is in $\boldsymbol{y}_2$, and for the third operator the response is in $\boldsymbol{y}_3$. This function for $\boldsymbol{y}$ generated 60 normal random samples. The simulated data is in figure 4.

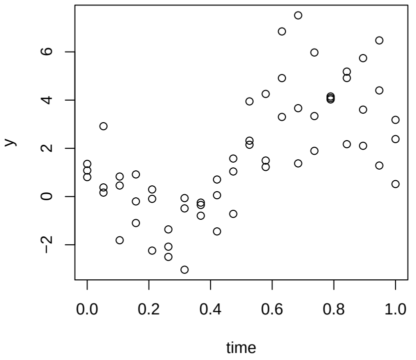

Figure 4: Simulated data points with three operators

We estimate the fit based on approaches, **ordinary least square**, **penalized spline** and **linear mixed model** discussed in section 3 and **Bayesian penalized spline model** in section 4. We again used three knots for all the models so that we can compare the results.

The estimate of $\boldsymbol{\beta}$ in **ordinary least squares** is in equation (7), where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Z})$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}'_x, \boldsymbol{\beta}'_{dum}, \boldsymbol{\mu}')'$. For **penalized spline**, the λ that minimized GCV is $\lambda = 2.2222 \times 10^{-4}$. In **linear mixed model**, the minimum GCV gave $\lambda = 2.23 \times 10^{-4}$, which allowed us to compute both our $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ to determine a fit.

For **Bayesian penalized spline**, we have to specify the hyperparameter b for the prior $\pi(\lambda)$. We looked at having 4 to 5 prior df_{fit}, before including the operator effects. Now with three operators included, we are going to add 3 more df, giving us the prior $df_{fit} = 7$, giving us $\lambda = .000095$ and $b = .00021$. Both non-informative and informative priors for $\boldsymbol{\beta}_{dum}$ were examined.

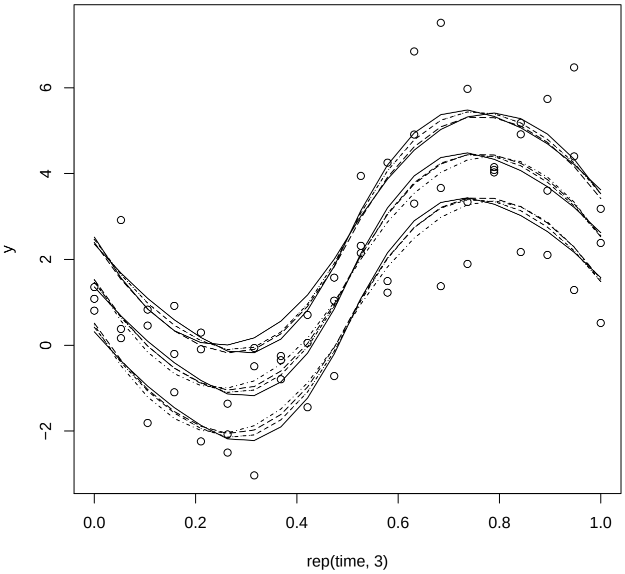

Figure 5: Plot of ordinary least squares (solid), penalized spline (dashed), linear mixed model (dotted), non-informative Bayesian penalized spline (dot-dashed), and informative Bayesian penalized spline (long dashed).

Approach	AIC
Ordinary least squares	4.541667
Penalized spline	4.536191
Linear mixed model	4.536197
Bayesian penalized spline (non-informative)	4.548832
Bayesian penalized spline (informative)	4.512195

Table 3: AIC for the different methods

5.2.1 Model Comparison for Simulated Data with *time* and *operators*

Although the Bayesian penalized spline with the informative prior model gave the lowest AIC (as indicated in table 3), all of the AIC values are close. This tells us that all of the models would provide a similar fit for the data as showed in figure 5.

6 Manufacturing Data

This new data concerns parts called “004 profiles,” which are produced by stamping out metal material through a machine. Some parts of the machine become worn out after a certain period of time, resulting in the “004 profiles” being out of specification. In the “004 profiles” data there are three variables. There is “profile,” which is the measurement taken; “time,” which is the time at which the measurement was taken; and there is also “operator,” which is the ID of the operator that took the measurement. We considered “profile” to be our response variable, “time” was our continuous predictor variable and “operator” is a categorical predictor variable. Figure 6 is a plot of the raw data.

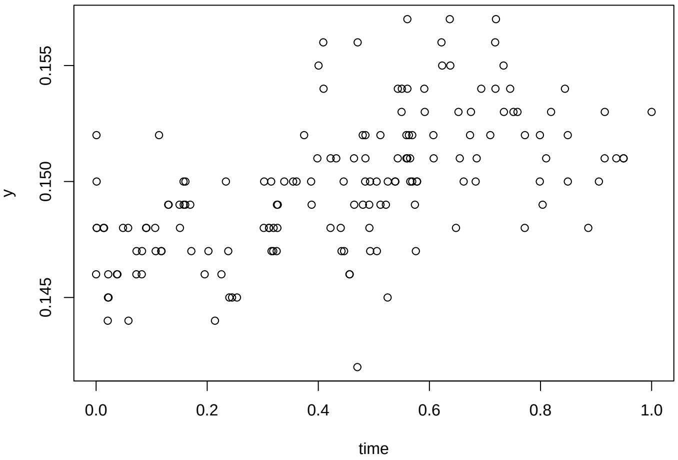

Figure 6: Manufacturing data points

Our first step was to choose the number of knots for our model, which was done using the ordinary least squares method, and selecting the number of knots associated with the greatest $R^2_{adj.}$ value. This resulted in us choosing to implement 5 knots in our model.

Knots	R^2_{adj}
1	.3753
2	.3781
3	.3718
4	.3743
5	.4024
6	.3921

Table 4: $R^2_{adj.}$ values for the different number of knots in ordinary least square

6.1 Model the Variable "*time*" only

The nonparametric model is proposed to model the variable "*time*",

$$y = f(time) + \epsilon.$$

Using the methods laid out in section 2.1, we first tested the ordinary least squares. Next we looked at the penalized spline approach described in section 2.2. We found our minimum GCV = .00089982, with corresponding $df_{fit} = 7.669$ and $\lambda = 2.020202 \times 10^{-5}$. Moving on to the linear mixed model approach described in section 2.3, we found that our best $\lambda = 2.118182 \times 10^{-5}$, based on a minimum GCV = .00089985 with df_{fit}=7.656. Since the λ values are so close to zero, both the penalized spline and the linear mixed models have comparable fits to that of the ordinary least squares model.

Lastly, we implemented the Bayesian penalized spline approach which required us to specify the hyperparameter b. We tried several prior df_{fit} listed in table 5.

Prior df_{fit}	λ	b
6.5	.000225	.000495
7	.000097	.000213
7.5	.000035	.000077
7.66	.000021	.000046

Table 5: Different λ and b values based on chosen prior df_{fit}

We chose our prior $df_{fit} = 7.66$ because our penalized spline and linear mixed model approaches both had approximately this value. Thus, our b=.000046. Again, we could have chosen a different prior df_{fit} since the posterior df_{fit} is robust to the choice of b, and still have ended up with similar results.

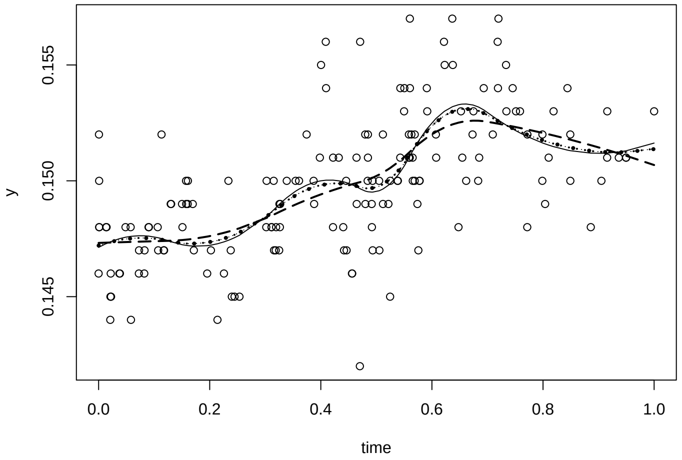

Figure 7: Plot of ordinary least squares (solid), penalized spline with $\lambda = 2.020202 \times 10^{-5}$ (dotted), linear mixed model with $\lambda = 2.118182 \times 10^{-5}$ (thick dotted), and Bayesian penalized spline (dashed) where b=.000046

Once we had all of our fits, we used the AIC to compare them.

Approach	AIC
Ordinary least squares	-7.014243
Penalized spline	-7.015655
Linear mixed model	-7.015612
Bayesian penalized spline	-6.998666

Table 6: AIC for the different methods

In table 6 we can see that the penalized spline approach had the lowest AIC. Even though this is the lowest value, all of the values are close so they will give about the same fit for our manufacturing data set of “004 profiles.” These fits are presented in figure 7.

6.2 Model Variables “*time*” and “*operators*”

In addition to having *time* as our independent variable, we also use the operators that took the measurements as our independent categorical variable. In this data set there were 18 different operators, so our categorical variable will have 18 levels with 17 dummy variables incorporated in our equation (20).

Using the methods laid out in section 3, we first tested the ordinary least squares approach. Next we looked at the penalized spline approach. We found our minimum GCV = .00099725, with corresponding $df_{fit} = 24.61611$ and $\lambda = 2.020202 \times 10^{-5}$. Moving on to the linear mixed model approach, we found that our best $\lambda = 2.118182 \times 10^{-5}$, based on a minimum GCV = .0009972932 with df_{fit}=24.60107.

We then looked at the Bayesian penalized spline with a semiparametric model as described in section 4.3, with both an informative and a non-informative prior for β_{dum} using block sampling. For the non-informative prior, we will be using the full conditionals from the “Non-informative Prior” in subsection 4.3.1 and for the informative prior, we will be using the full conditionals from the “Informative Prior” in subsection 4.3.2. Both of these methods need us to first find b. Several prior df_{fit} is proposed and the corresponding b is provided in table 7.

Prior df_{fit}	λ	b
23	.000390	.000858
23.5	.000179	.000394
24	.000079	.000174
24.5	.000028	.00062
24.6	.000021	.000047

Table 7: Different λ and b values based on chosen prior df_{fit}

We chose our prior $df_{fit} = 24.6$ because our penalized spline and linear mixed model approaches both had approximately this value. Thus, our b=.000047. We can see the fit for the non-informative prior in figure 8, and the fit for the informative prior in figure 9. We can see from both figure 8 and 9 that we have 18 lines. These lines represent the different operators. Some of the lines are longer because some operators took more measurements than others.

Approach	AIC
Ordinary least squares	-6.935337
Penalized spline	-6.936578
Linear mixed model	-6.936503
Bayesian penalized spline (non-informative)	-6.929622
Bayesian penalized spline (informative)	-6.929623

Table 8: AIC for the different methods

Once we had all of our fits, we used the AIC to make the grand comparison. Again, they are all pretty close so they will give about the same fit for our manufacturing data set of “004 profiles.”

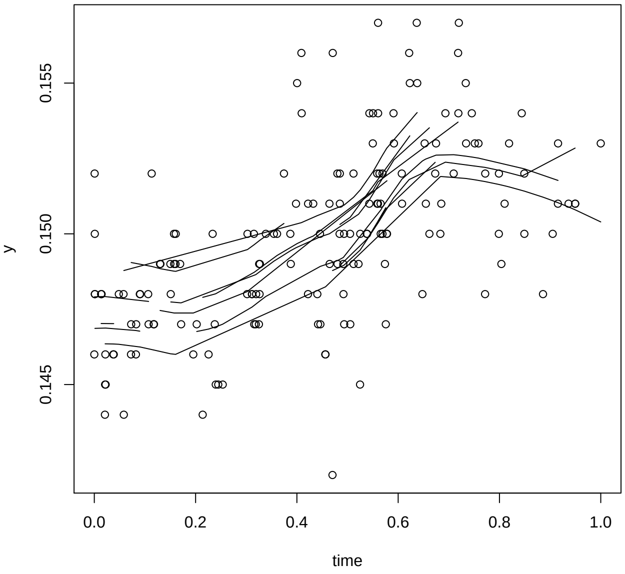

Figure 8: Plot of non-informative Bayesian penalized spline fit where b=.000047 with 18 operators

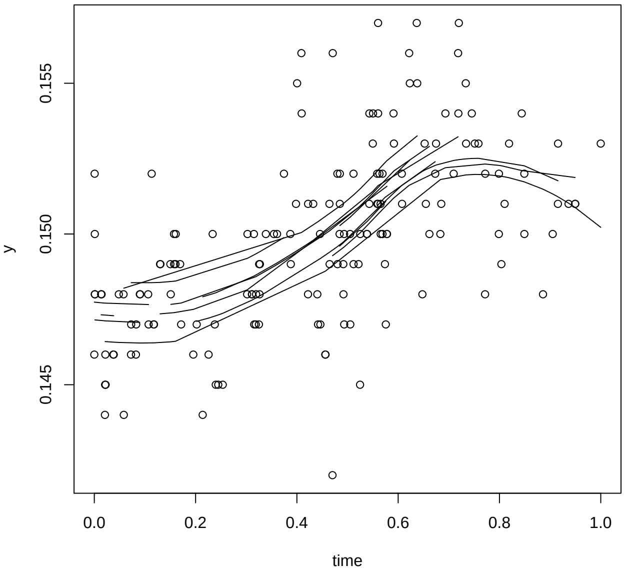

Figure 9: Plot of informative Bayesian penalized spline fit where b=.000047 with 18 operators

7 Conclusion

Based on the simulated data for a nonparametric model in section (5.1), we found that the Bayesian penalized spline method was slightly better than the ordinary least squares, penalized spline, and linear mixed model methods. We believe this is due to the fact that we incorporated prior information into this method, which helped to improve our estimation.

We also found that when dealing with a semiparametric model for our simulated data shown in section (5.2), all of the methods gave approximately the same result. Looking between the non-informative and the informative prior for β_{dum}, we also saw very similar results. We believe that this is because of our small λ value, which caused the variance $\frac{\sigma^2}{\lambda}$ for our informative prior for β_{dum} to be very large. This in turn produced an almost uniform distribution, which was how our non-informative prior was classified, thus they gave very similar results. Therefore, the data provided sufficient information to overcome the influence of the priors selected.

Even though Bayesian penalized spline yields a similar result to the penalized spline and linear mixed model methods, it offers the framework to incorporate prior information when available. The posterior distribution obtained for λ provided the information for statistical inference while the frequestist's estimates based on minimum GCV only provided the point estimate.

In the future, there are several extensions of this project that can be done. We have worked with nonparametric and semiparametric models, but we could also work with an additive model, which combines two continuous functions. Further, in our work, we used a truncated polynomial basis for our calculations. We could try other basis, like the radial basis function that takes the absolute value of the knots, to see the effects on the estimation. Finally, because we have only worked with predictions, we could use hypothesis testing to assess the data in this fashion.

References

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis.* CRC Press LLC.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction: with 200 full-color illustrations.* Springer-Verlag Inc.

Kimeldorf, G. S., & Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Application*, *33*, 82-94.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression.* Cambridge UP.

Sun, D., & Speckman, P. L. (2008). Bayesian hierarchical linear mixed models for additive smoothing splines. *Annals of the Institute of Statistical Mathematics*, *60*(3), 499–517.