

# A Simulation Study Comparing Knot Selection Methods With Equally Spaced Knots in a Penalized Regression Spline

Eduardo L. Montoya<sup>1</sup>, Nehemias Ulloa<sup>1</sup> & Victoria Miller<sup>1</sup>

<sup>1</sup> Department of Mathematics, California State University, Bakersfield, USA

Correspondence: Eduardo L. Montoya, Department of Mathematics, California State University, Bakersfield 93311, USA. Tel: 1-661-654-6838. E-mail: emontoya2@csu.edu

Received: May 16, 2014 Accepted: June 15, 2014 Online Published: June 26, 2014

doi:10.5539/ijsp.v3n3p96 URL: <http://dx.doi.org/10.5539/ijsp.v3n3p96>

## Abstract

Penalized regression splines are a commonly used method to estimate complex non-linear relationships between two variables. The fit of a penalized regression spline to the data depends on the number of knots, knot placement, and the value of the smoothing parameter. In this paper, we use a simulation study to compare knot selection methods with equidistant knots in a penalized regression spline model. We found that one method generally performed better than others. The results provide guidance in selecting the number of equidistant knots in a penalized regression spline model.

**Keywords:** knots selection, smoothing parameter, penalized regression splines

## 1. Introduction

Consider the general regression model with a single explanatory variable that takes the form

$$y_i = g(x_i) + \varepsilon_i, \text{ for } i = 1, \dots, n, \quad (1)$$

where  $x_i \in [a, b]$ ,  $y_i$  is a response variable,  $x_i$  is a covariate,  $g(x)$  is the regression function dependent on the covariate,  $n$  is the number of observations, and  $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$  for all  $i$ . The goal with model (1) is to estimate the function  $g(x)$ , which describes the relationship between  $x$  and  $y$  in the presence of uncertainty. There are various approaches for estimating the true underlying curve,  $g(x)$  for all  $x \in [a, b]$ , depending on what we know about  $g(x)$ .

Non-parametric methods do not assume a parametric form for  $g(x)$  in model (1), but instead the estimation of  $g(x)$  is driven by the data. For example, to obtain a cubic penalized regression spline estimate of  $g(x)$ , one may find  $\widehat{g}(x)$  to minimize the following criterion

$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int_a^b (g''(x))^2 dx \quad (2)$$

where  $\lambda$  is called the smoothing parameter and  $a = x_1 < \dots < x_n = b$ . To minimize fitting criterion (2), a penalized regression spline approximates  $g(x)$  by a linear combination of basis functions. The number of basis functions that approximate  $g(x)$  depends on the number of knots  $K$ . The cubic penalized regression spline estimate obtained by minimizing fitting criterion (2) is dependent on the value of  $\lambda$ ,  $K$ , and the location of the knots. There are various families of basis functions that may be used in a penalized regression spline. To confine the study to a reasonable length, we only consider a truncated power basis for illustration in Section 2.

When a knot is placed at each data point, the penalized regression spline is referred to as a smoothing spline. Using smoothing splines, Lee (2003) compares, among others, Akaike's information criterion corrected, Cross-Validation, Generalized Cross-Validation, and Mallows' Cp criterion to select the smoothing parameter. The location of the knots may also be chosen to control the shape of the curve estimate. Stone et al. (1997) and references cited therein discuss various stepwise selection procedures to select the number of knots and their location in a regression spline setting. In a penalized regression spline setting, Spiriti et al. (2013) proposes two random search algorithms for selecting the knot locations.

Using a penalized regression spline, Ruppert (2002) compares the performance of two different algorithms (Myopic and Full Search) to select the number of knots when placing the knots at quantiles of the  $x_i$ 's with generalized cross validation for smoothing parameter selection and found that both algorithms perform well. Ruppert et al. (2003) further expands on this discussion and provides a default choice for the number of knots when placing the knots at quantiles of the  $x_i$ 's. Wand and Ormerod (2008) also place the knots at quantile locations using a similar default rule as Ruppert et al. (2003) to select the number of knots in a penalized regression spline. Using equidistant knot locations, Eilers and Marx (2010) compare two types of penalized regression splines. They demonstrated that a large number of equally spaced knots outperformed quantile spacing of the knots in their investigation, but an algorithm for determining what is large was not discussed. See Wand (2000) for a discussion on some of these procedures and others in a regression and penalized regression spline setting.

The term “mixed model spline” has been used by some to indicate a spline smooth estimated via the mixed model methodology. Welham et al. (2007) provide an overview of smoothing splines and penalized regression splines in a linear mixed model setting. For quantile knot placement, they suggest using the default choice, Myopic algorithm, or Full Search algorithm proposed by Ruppert et al. (2003), but the performance of these methods in their setting were not investigated. Kauermann and Opsomer (2011) provide a likelihood based approach to select the number of knots using quantile knot spacing in a mixed model spline. Software implementation and other examples of mixed model splines are discussed in Ngo and Wand (2004).

In this article, we build upon the analysis of Ruppert (2002) and Ruppert et al. (2003) in regards to selecting the number of knots in a penalized regression spline; a notable difference is that we use equally spaced knots and several smoothing parameter selection methods under various simulation settings. Our aim is to determine if one method for choosing the number of equally spaced knots outperforms another, thereby providing a preferred approach for selecting the number of equally spaced knots. To our knowledge, the performance of these knot selection approaches when using equidistant knots have not been examined previously. In Section 2, we review a penalized regression spline with a truncated power basis. We henceforth refer to a penalized regression spline with a truncated power basis as a TSM. In Section 3, we present several commonly used criteria for selecting the smoothing parameter  $\lambda$  in a penalized regression spline. In Section 4, we review three different methods presented in Ruppert et al. (2003) to select the number of knots. In Section 5, we present a simulation study to examine the performance of the knot selection methods with equally spaced knots. We conclude with discussion in Section 6.

## 2. A Regression Spline With a Truncated Power Basis

We use a cubic penalized regression spline with a truncated power basis to estimate  $g(x)$  in model (1). Key references for regression splines with a truncated power basis include Hastie et al. (2001) and Ruppert et al. (2003). Using a cubic penalized regression spline with a truncated power basis, we model  $g(x)$  as

$$g(x) = \beta_0 + \beta_1 x + \dots + \beta_3 x^3 + \sum_{j=1}^K \beta_{3j} (x - k_j)_+^3.$$

where  $k_j$  is the  $j^{\text{th}}$  knot, and  $(x - k_j)_+ = (x - k_j)$  if  $x \geq k_j$  or zero if  $x < k_j$ . Assuming that the observation points  $x_i$  are unique, an estimate of  $g(x)$  can be found to minimize (2) if  $\lambda$  and  $K$  are held constant. To illustrate, we follow a matrix-based approach as described in Ruppert et al. (2003), among others. Let  $\mathbf{y} = [y_1 \dots y_n]^T$ ,  $\boldsymbol{\beta} = [\beta_1 \dots \beta_3 \beta_{31} \dots \beta_{pK}]^T$ ,  $\boldsymbol{\varepsilon} = [\varepsilon_1 \dots \varepsilon_n]^T$ , and

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \dots & x_1^3 & (x_1 - k_1)_+^3 & \dots & (x_1 - k_K)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_n & \dots & x_n^3 & (x_n - k_1)_+^3 & \dots & (x_n - k_K)_+^3 \end{bmatrix}.$$

Using this notation, the TSM may be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where  $\mathbf{y}$  is a column vector,  $\mathbf{X}$  is a  $n \times (4 + K)$  matrix. To estimate  $\boldsymbol{\beta}$ , one may use the estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{R})^{-1} \mathbf{X}^T \mathbf{y},$$

where

$$\mathbf{R} = \begin{bmatrix} \mathbf{0}_{4 \times 4} & \mathbf{0}_{4 \times K} \\ \mathbf{0}_{K \times 4} & \mathbf{I}_{K \times K} \end{bmatrix}$$

and  $\mathbf{I}_{K \times K}$  is a  $K \times K$  identity matrix. To increase numerical stability when using a truncated power basis in our simulation study in Section 5, we replace  $\lambda \mathbf{R}$  in (2) by  $10^{-10} \mathbf{R} + \lambda \mathbf{R}$  (Ruppert, 2002). This results in a penalty lower bound of  $10^{-10}$  instead of 0. Using the estimate from (2), we obtain

$$\widehat{g}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \dots + \widehat{\beta}_3 x^3 + \sum_{j=1}^K \widehat{\beta}_{3j} (x - k_j)_+^3.$$

This estimate depends on the value of  $\lambda$  and  $K$ .

### 3. Selecting the Smoothing Parameter

In this section, we discuss various data driven criteria for selecting  $\lambda$  in fitting criterion (2) conditional on the value of  $K$ : Akaike's information criterion corrected (*AICc*), Bayesian information criterion (*BIC*), Cross-Validation (*CV*) and Generalized Cross-Validation (*GCV*) criterion. Each of these data driven criteria provides an approach to select the value of  $\lambda$  conditional on the number of knots, and each criterion is a function of  $\lambda$ . Selection of the value of  $K$  is addressed in Section 4.

For the *AICc*, *BIC*, *CV*, and *GCV* criteria, the value of  $\lambda$  that gives the minimum value of the criteria is taken to be a good value for the smoothing parameter conditional on the value of  $K$ . Each of these methods is dependent on the sum of squares error,

$$SSE(\lambda) = \sum_{i=1}^n (y_i - \widehat{g}(x_i))^2,$$

as well as the effective degrees of freedom of  $\widehat{g}(x)$ ,  $df_\lambda = \text{tr}(\mathbf{S}_\lambda)$  as defined in Buja et al. (1989), where

$$\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{R})^{-1} \mathbf{X}^\top.$$

The matrix  $\mathbf{S}_\lambda$  is known as the smoother matrix. The more complex or "wiggly" the estimate of  $g(x)$ , the higher its  $df_\lambda$ .

The method of CV recycles the data by using training and test samples. Specifically, for each  $i = 1, \dots, n$ , we obtain the estimate of  $g(x)$  that minimizes

$$\sum_{l=1, l \neq i}^n (y_l - g(x_l))^2 + \lambda \int_a^b (g''(x))^2 dx$$

conditional on the value of  $\lambda$ . Denote the minimizing solution based on all the data except the  $i^{\text{th}}$  observation by  $\widehat{g}_{-i}(x_i)$ . This process is repeated for each observation. There are  $n$  cases that one can delete; therefore, for each  $\lambda$  the cross-validation score is defined as

$$CV(\lambda) = \sum_{i=1}^n \{y_i - \widehat{g}_{-i}(x_i)\}^2.$$

Because of the implementation, the CV method is also known as the leave-one-out method. The value of  $\lambda$  that gives the minimum CV score is taken to be a good choice for the smoothing parameter. If a smoother matrix exists, then the CV formula (Silverman, 1985) may be expressed as

$$CV(\lambda) = \sum_{i=1}^n \left( \frac{y_i - \widehat{g}(x_i)}{1 - S_{ii}} \right)^2,$$

where  $S_{ii}$  denotes the  $i^{\text{th}}$  diagonal element of  $\mathbf{S}_\lambda$  and  $\widehat{g}(x_i)$  is the estimate that minimizes (2). Using this formula we only have to fit the data once for each  $\lambda$ .

Akaike's information criterion corrected, *AICc*, was introduced by Hurvich et al. (1998) because the commonly used Akaike's information criterion (Eubank, 1999) may have a tendency to over-fit the curve estimate for small samples. The *AICc* criterion may be expressed as

$$AICc(\lambda) = \ln(n^{-1} SSE(\lambda)) + \frac{2(df_\lambda + 1)}{n - df_\lambda - 2}.$$

The BIC criterion (Schwarz, 1978) may be defined as,

$$BIC(\lambda) = \ln(n^{-1}SSE(\lambda)) + \frac{\log(n)}{n}df_{\lambda},$$

which is similar to AICc but penalizes a model fit with a larger  $df_{\lambda}$  more strongly than the AICc for large  $n$ .

Developed by Craven and Wahba (1979), GCV may be defined by

$$GCV(\lambda) = n^{-1} \frac{SSE(\lambda)}{(1 - n^{-1}df_{\lambda})^2}.$$

Each of these smoothing parameter selection methods consists of trading off the complexity of an estimate of  $g(x)$  against how well the model fits the data. Using an adequate number of knots, the size of the smoothing parameter controls the influence of the penalty in (2).

#### 4. Knot Selection Methods

We discuss three methods given in Ruppert et al. (2003) to select the number of knots  $K$ . These methods were presented using the quantiles of the  $x_i$ 's as the knot locations with  $GCV$  for smoothing parameter selection. In Section 6, we study these methods using equally spaced knots with several smoothing parameter selection methods. The three methods to select the number knots  $K$  are the fixed selection method, Myopic algorithm, and Full Search algorithm. When using fitting criterion (2) selecting too many knots or too few knots may lead to over-fitting or under-fitting, respectively. The intent of these knot selection methods is to provide an adequate number of knots to allow a flexible enough fit when controlling the smoothness of the curve estimate with a smoothing parameter.

##### 4.1 Fixed Selection Method

The fixed selection approach sets the number of knots to

$$K = \min\left(\frac{1}{4} \times \text{number of unique } x_i\text{'s}, 35\right).$$

This is referred to as the default choice by Ruppert et al. (2003).

##### 4.2 Myopic Algorithm

The Myopic algorithm is an algorithm that uses a smoothing parameter selection method. For illustration, we will use the  $AICc$  criterion. Begin with a set of candidate values for the number of knots  $K$ . In this study, we use candidate values similar to those suggested in Ruppert et al. (2003). Specifically, we set  $\{K_1, K_2, K_3, K_4, K_5, K_6\} = \{5, 10, 20, 40, 80, 160\}$ , assuming  $160 \leq n$ . If  $160 > n$ , then reduce the maximum of the set of candidate knot values to  $n$ . For  $w = 1, \dots, 6$ , proceed as follows:

- 1) Fit the model with  $AICc$  to select  $\lambda$  using  $K_w$  knots. Let  $\lambda_1$  denote the smoothing parameter chosen by  $AICc$ .
- 2) Fit the model with  $AICc$  to select  $\lambda$  using  $K_{w+1}$  knots. Let  $\lambda_2$  denote the smoothing parameter chosen by  $AICc$ .
- 3) If  $AICc(\lambda_2) > 0.98AICc(\lambda_1)$ , then stop and use the number of knots that corresponds to  $\min(AICc(\lambda_1), AICc(\lambda_2))$ . If  $AICc(\lambda_2) \leq 0.98AICc(\lambda_1)$ , then repeat steps 1-3 for  $w = 2, 3, 4, 5, 6$  until the inequality in step 3 is satisfied or when  $w = 6$ . If  $w = 6$ , then  $K_6$  knots is assumed to be best.

##### 4.3 Full Search Algorithm

The Full Search algorithm is similar to the Myopic algorithm, but it iterates through the list of possible knots. Begin with a set of candidate values for the number of knots,  $\{K_1, K_2, K_3, K_4, K_5, K_6\} = \{5, 10, 20, 40, 80, 160\}$ . For  $w = 1, \dots, 6$ , proceed as follows:

- 1) Fit the model with  $AICc$  to select  $\lambda$  using  $K_w$  knots for  $w = 1, \dots, 6$ . Let  $\lambda_w$  denote the smoothing parameter chosen by  $AICc$  for  $w = 1, \dots, 6$ .
- 2) For  $w = 1, \dots, 6$ , select the value of  $AICc(\lambda_w)$  that is the smallest, and then use  $K_w$  knots.

Although we used  $AICc$  to illustrate the Myopic and Full Search algorithm, one may use any of the smoothing parameter selection methods discussed in Section 3. More details regarding these knot selection methods may be found in Ruppert (2002) and Ruppert et al. (2003).

## 5. A Simulation Study

In this section, we perform a simulation study using 12 different combinations for selecting  $\lambda$  and  $K$  to compare the knot selection methods when using equally spaced knots. Recall the data-driven smoothing parameter selection approaches discussed in Section 3:  $AICc$ ,  $BIC$ ,  $CV$ , and  $GCV$ . We use each of these smoothing parameter selection methods combined with a knot selection method (fixed knot selection method (FM), Myopic algorithm knot selection method (MA), or Full Search algorithm knot selection method (FA)) to select the number of equally spaced knots  $K$ . For comparison, we also consider a smoothing spline estimate by placing a knot at each data point, and we denote this approach as the SMS method. Note that any penalized regression or smoothing spline estimator  $\widehat{g}(x)$  will always depend on the chosen smoothing parameter  $\lambda$  and the number of knots  $K$ , but we suppress the notation of this dependence for better readability.

To compare the performance of an estimator  $\widehat{g}(x)$  to the true function  $g(x)$ , we use the median of a measure used by Lee (2003) to compare smoothing parameter selection methods for smoothing splines. The measure is defined as

$$D(g, \widehat{g}) = \frac{\sum_{i=1}^n (\widehat{g}(x_i) - g(x_i))^2}{\sum_{i=1}^n (\widehat{g}_{\lambda'}(x_i) - g(x_i))^2}, \quad (3)$$

such that the estimate  $\widehat{g}_{\lambda'}(x_i)$  depends on the smoothing parameter  $\lambda'$  chosen to minimize

$$\sum_{i=1}^n (\widehat{g}_{\lambda'}(x_i) - g(x_i))^2.$$

The closer the median of (3) is to 1, the better we consider the estimator  $\widehat{g}(x)$  relative to the true function  $g(x)$ . To obtain an analytical expression of the median of (3), we require the sampling distribution of the estimator  $\widehat{g}(x)$ . However, the sampling distribution will not be analytically derivable due to the smoothing parameter selection. Therefore, for each method of choosing  $\lambda$  and  $K$ , we estimate the median of (3) via simulation using

$$SED = \text{Median}(D(g, \widehat{g}_l)), \quad (4)$$

where the subscript  $l$  in  $\widehat{g}_l$  denotes the corresponding estimate of  $g(x)$  for the  $l^{\text{th}}$  simulated data set.

Using different simulation settings, (4) is obtained with  $L = 1000$  simulated data sets under model (1). For the simulations, we use three different true functions  $g(x)$ , each with three different sample sizes, as previously used in a simulation study by Wang and Wahba (1995) to study the performance of confidence intervals for smoothing splines. We denote the three forms of true function  $g(x)$  as  $g_1(x)$ ,  $g_2(x)$ , and  $g_3(x)$ , and they are defined as

$$\begin{aligned} g_1(x) &= \frac{1}{3} \text{Beta}_{20,5}(x) + \frac{1}{3} \text{Beta}_{12,12}(x) + \frac{1}{3} \text{Beta}_{7,30}(x), \\ g_2(x) &= \frac{1}{3} \text{Beta}_{10,5}(x) + \frac{1}{3} \text{Beta}_{7,7}(x) + \frac{1}{3} \text{Beta}_{5,10}(x), \text{ and} \\ g_3(x) &= \frac{6}{10} \text{Beta}_{30,17}(x) + \frac{4}{10} \text{Beta}_{3,11}(x) \end{aligned}$$

where

$$\text{Beta}_{c,d}(x) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} x^{c-1} (1-x)^{d-1} \text{ for } x \in [0, 1].$$

These curves are presented based on the ascending order of their signal to noise ratio (SNR),  $SD(g)/\sigma$ , as defined in Luo and Wahba (1997).

We use R (R Core Team, 2012) to carry out the simulations and to write a function to fit the TSM with any of the smoothing parameter selection methods described in Section 3, combined with a knot selection method described in Section 4. For each of  $L = 1000$  simulated datasets, (4) is computed for each combination of  $g(x) \in \{g_1(x), g_2(x), g_3(x)\}$ ,  $n \in \{32, 64, 128\}$ , and  $\sigma \in \{.3, .5\}$  under each combination of the smoothing parameter and knot selection methods previously discussed when using equidistant knot locations. A set of observation locations,  $\{x_1, \dots, x_n\}$ , for this simulation study are generated randomly from a uniform distribution on the  $(0, 1)$  interval. As an illustration, Figure 1 shows a single simulated data set for each respective sample ( $n = 32, 64$ , or  $128$ ) along with its corresponding true curve ( $g_1(x)$ ,  $g_2(x)$ , or  $g_3(x)$ ) when  $\sigma = .3$ .

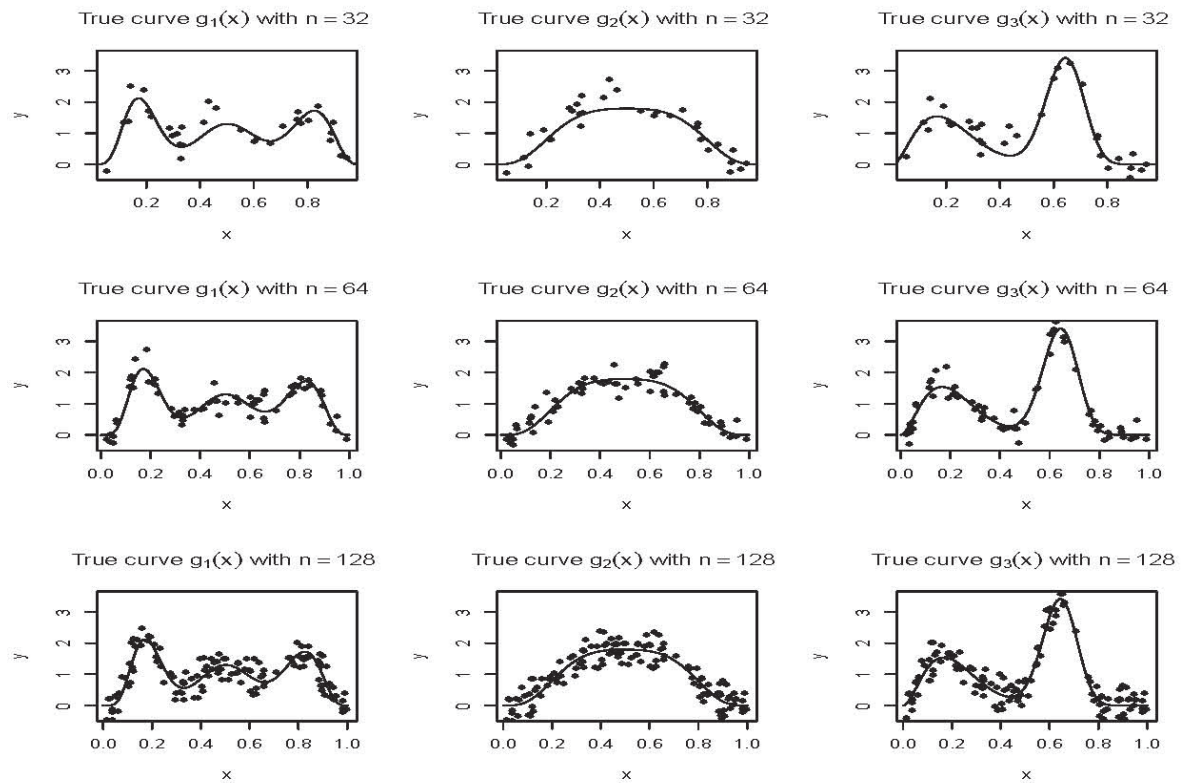


Figure 1. Top row:  $g_1(x)$  (left panel),  $g_2(x)$  (center panel) and  $g_3(x)$  (right panel) and a corresponding single simulated data set (one of the  $l$  simulated datasets) of size  $n = 32$  when using  $\sigma = .3$ . Second and third rows: analogous for  $n = 64$  and  $n = 128$  for a single simulated dataset

The simulated median estimates given in (4) for each knot selection method under the various simulation settings when the true curve is  $g_1(x)$ ,  $g_2(x)$ , and  $g_3(x)$  are provided in Tables 1, 2, and 3, respectively.

Table 1. The SED, when using each smoothing parameter selection criterion, under each knot selection method, for true curve  $g_1(x)$

	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$
	<i>AICc</i>		<i>BIC</i>		<i>CV</i>		<i>GCV</i>	
	FM							
$n = 32$	1.116	1.150	1.086	1.147	1.359	1.202	1.076	1.118
$n=64$	1.070	1.086	1.108	1.146	1.081	1.109	1.072	1.086
$n = 128$	1.063	1.065	1.142	1.146	1.074	1.081	1.076	1.082
	MA							
$n = 32$	1.218	1.224	1.226	1.235	1.667	1.258	1.171	1.222
$n=64$	1.128	1.118	1.153	1.170	1.127	1.138	1.130	1.121
$n = 128$	1.097	1.104	1.193	1.186	1.112	1.122	1.104	1.118
	FA							
$n = 32$	1.185	1.217	1.207	1.252	1.580	1.258	1.167	1.227
$n=64$	1.128	1.113	1.153	1.160	1.127	1.139	1.133	1.123
$n = 128$	1.098	1.107	1.191	1.18	1.113	1.122	1.109	1.118
	SMS							
$n = 32$	1.128	1.188	1.148	1.234	1.525	1.262	1.131	1.171
$n=64$	1.068	1.089	1.116	1.149	1.083	1.122	1.085	1.101
$n = 128$	1.077	1.074	1.157	1.148	1.090	1.091	1.092	1.095



Table 2. The SED, when using each smoothing parameter selection criterion, under each knot selection method, for true curve  $g_2(x)$ 

	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$
	<i>AICc</i>		<i>BIC</i>		<i>CV</i>		<i>GCV</i>	
FM								
$n = 32$	1.202	1.088	1.214	1.094	1.209	1.111	1.234	1.122
$n = 64$	1.210	1.240	1.428	1.291	1.208	1.237	1.224	1.250
$n = 128$	1.154	1.157	1.338	1.423	1.163	1.166	1.175	1.160
MA								
$n = 32$	1.241	1.107	1.253	1.109	1.229	1.139	1.276	1.155
$n = 64$	1.227	1.256	1.347	1.323	1.239	1.254	1.241	1.262
$n = 128$	1.185	1.222	1.373	1.449	1.188	1.226	1.188	1.227
FA								
$n = 32$	1.241	1.110	1.263	1.115	1.237	1.146	1.287	1.174
$n = 64$	1.233	1.259	1.354	1.323	1.242	1.264	1.254	1.271
$n = 128$	1.196	1.232	1.375	1.449	1.207	1.243	1.204	1.241
SMS								
$n = 32$	1.193	1.091	1.216	1.095	1.194	1.112	1.244	1.123
$n = 64$	1.164	1.222	1.339	1.291	1.166	1.226	1.185	1.238
$n = 128$	1.143	1.161	1.344	1.406	1.148	1.164	1.157	1.161

Table 3. The SED, when using each smoothing parameter selection criterion, under each knot selection method, for true curve  $g_3(x)$ 

	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$
	<i>AICc</i>		<i>BIC</i>		<i>CV</i>		<i>GCV</i>	
FM								
$n = 32$	1.078	1.127	1.057	1.081	1.098	1.118	1.045	1.067
$n = 64$	1.067	1.078	1.114	1.117	1.119	1.123	1.066	1.086
$n = 128$	1.057	1.062	1.153	1.198	1.063	1.073	1.063	1.067
MA								
$n = 32$	1.129	1.141	1.134	1.173	1.239	1.221	1.123	1.135
$n = 64$	1.133	1.111	1.18	1.134	1.195	1.145	1.132	1.112
$n = 128$	1.096	1.111	1.279	1.219	1.11	1.122	1.093	1.112
FA								
$n = 32$	1.129	1.140	1.141	1.177	1.240	1.228	1.124	1.137
$n = 64$	1.132	1.110	1.179	1.134	1.196	1.145	1.133	1.113
$n = 128$	1.097	1.112	1.274	1.218	1.103	1.127	1.094	1.117
SMS								
$n = 32$	1.133	1.152	1.105	1.161	1.309	1.265	1.095	1.122
$n = 64$	1.071	1.079	1.125	1.118	1.125	1.127	1.078	1.093
$n = 128$	1.060	1.072	1.176	1.200	1.070	1.088	1.066	1.080

When the true curve is  $g_1(x)$ , the *FM* provides the lowest SED when compared to the *MA* and *FA* across all simulation settings. In addition, the *FM* provides a lower SED than the *SMS* for all but one simulation setting. Under  $g_2(x)$ , the *FM* provides the lowest SED when compared to the *MA* and *FA* for all but one simulation setting. When comparing the *FM* to the *SMS*, the *FM* tended to outperform the *SMS* for  $n = 32$  and  $n = 128$ , but not for  $n = 64$ . When the true curve is  $g_3(x)$ , the *FM* provides the lowest SED for all cases when compared to the *MA* and *FA*. The *FM* also provides a lower SED than the *SMS* for all cases. The *FM* clearly outperforms both the *MA* and *FA* for almost all circumstances in our study in terms of (4). In addition, the *FM* provides a lower SED than the *SMS* for most cases.

Boxplots of the replications of  $D(g, \hat{g}_l)$  for  $l = 1, \dots, 1000$  on the log scale for each simulation setting when  $\sigma = .3$  and  $\sigma = .5$  are shown in Figures 2 and 3, respectively. Note that the results under *GCV* with the *FM* tend to, more often than not, exhibit less variability than the other smoothing parameter selection methods with the *FM* in terms of (4). While the differences among the SED's for the each simulation setting may not seem large, we conduct

all pairwise comparisons of the medians using a Kruskal-Wallis test as described in Siegel and Castellan (1988) at significance level .05 to compare the  $FM$ ,  $MA$ , and  $FA$ . This approach performs Wilcoxon rank-sum tests on all possible pairwise comparisons of the  $FM$ ,  $MA$ , and  $FA$  under a given smoothing parameter selection method, true curve, sample size, and noise level. Furthermore, we compare the  $FM$  with the  $SMS$  using a Wilcoxon rank-sum test. The result of all comparisons when  $\sigma = .3$  and  $\sigma = .5$  are summarized in Figures 2 and 3, respectively.

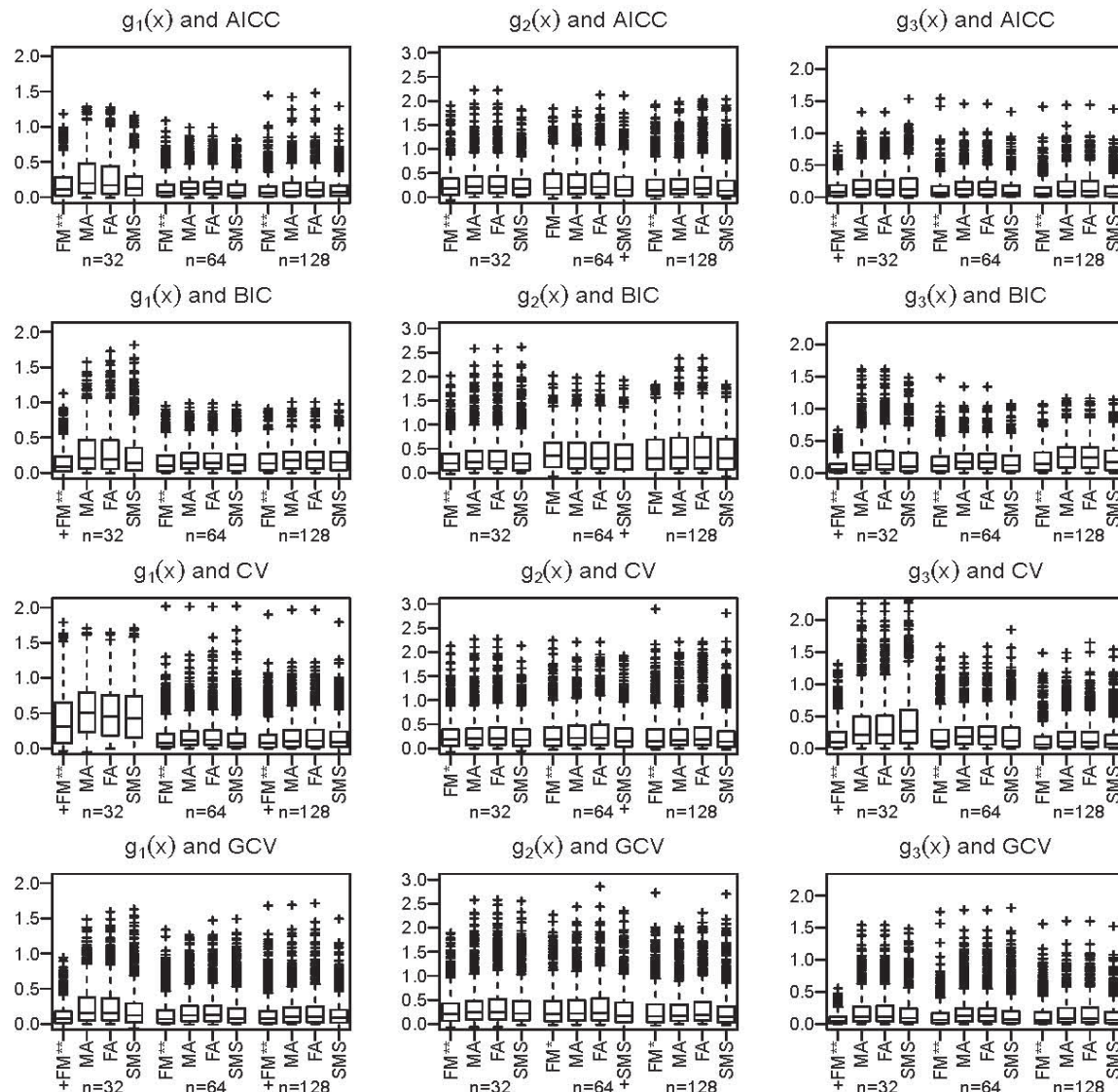
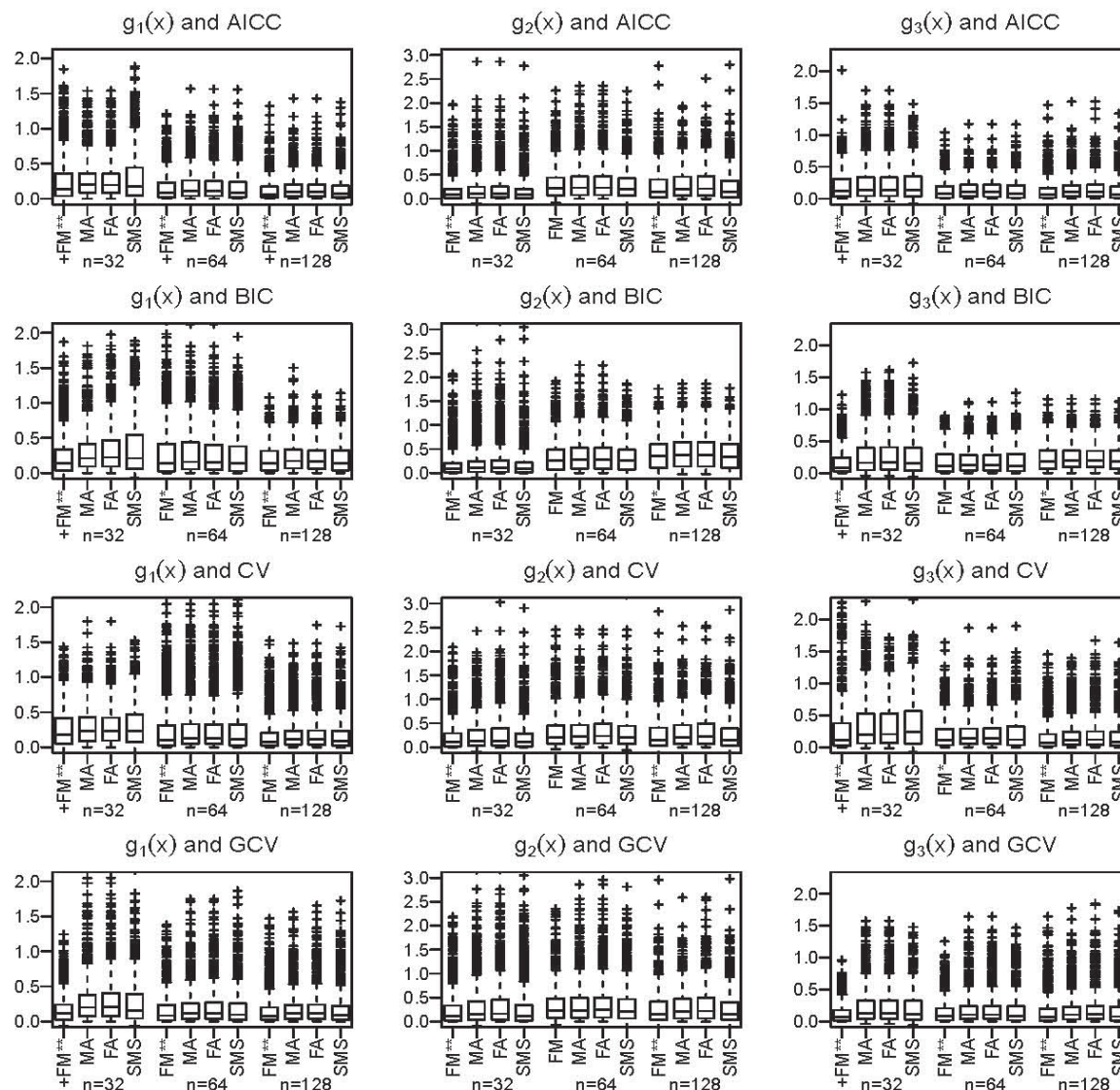


Figure 2. Replications of  $D(\hat{g}_l, g)$  on the log scale, for  $l = 1, \dots, 1000$  when  $\sigma = .3$  under different simulation settings. Top left corner panel: Under each sample size, the four box plots contain the replications when the true curve is  $g_1(x)$  when  $AICC$  is used for smoothing parameter section for the  $FM$ ,  $MA$ ,  $FA$ , and  $SMS$ , respectively.

An  $FM^{**}$  denotes that the median for the  $FM$  was significantly lower than the median of the  $MA$  and  $FA$ . An  $FM^*$  denotes that the median for the  $FM$  was significantly lower than the  $MA$  or  $FA$ . An  $FM$  without a  $*$  or  $**$  indicates that the median for  $FM$  was not significantly different from median of the  $MA$  or  $FA$ . An  $+FM$  denotes that the median for the  $FM$  was significantly lower than the median of the  $SMS$ . An  $FM$  without a  $+$  preceding it indicates that the median for  $FM$  was not significantly different from median of the  $SMS$ . An  $+SMS$  denotes that the median for the  $SMS$  was significantly lower than the median of the  $FM$ . Remaining panels: Analogous top left corner panel for the true curve and smoothing parameter selection method shown in the title of the panel



Figure 3. Analogous to Figure 2 when  $\sigma = .5$ 

Under  $\sigma = .3$ , the tests indicate that the *FM* has a lower SED than the *MA* and *FM* under the true curves  $g_1(x)$  and  $g_3(x)$  for a given sample size and smoothing parameter selection method. Under  $g_2(x)$ , we find the *FM* has an SED that is lower or no different than the SED under the *MA* and *FA*. When comparing the *FM* and *SMS*, the results indicate that the *FM* performed just as well or better for most cases. For  $\sigma = .5$ , the results show the *FM* performs better or not significantly different than the *MA* and *FA* across all settings. Furthermore, the results indicate that the *FM* performed just as well or better than the *SMS* for all cases.

Tables 4, 5, and 6 show the median number of knots selected, when using each smoothing parameter selection criterion, under each knot selection method, for true curve  $g_1(x)$ ,  $g_2(x)$ , and  $g_3(x)$ , respectively. In addition to the *FM* providing favorable results based on (4), the *FM* also uses less knots than the *SMS*. This shows that the *FM* is more computationally attractive in terms of model fitting. Tables 4, 5, and 6 also provide the median computational time, in terms of elapsed time in seconds with our computing resources, when using each smoothing parameter selection criterion, under each knot selection method, for true curve  $g_1(x)$ ,  $g_2(x)$ , and  $g_3(x)$ , respectively. The results show that the *FM* has a lower computational time than the *MA*, *FA*, and *SMS*. The difference in time between the methods increases as the sample size increases. Boxplots of the smoothing parameter selected on the log scale for each simulated data when  $\sigma = .3$  and  $\sigma = .5$  are shown in Figures 4 and 5, respectively. Under a given true curve and smoothing parameter selection method, the selected smoothing parameters exhibit similar

variability for a given sample size. We also note that the variation in the smoothing parameters selected tends to be smaller for the larger sample sizes. This shows that when there is more information, the smoothing parameter is less variable.

Table 4. The median number of knots selected, when using each smoothing parameter selection criterion, under each knot selection method, for true curve  $g_1(x)$ . The corresponding median computing time is shown in parentheses below the median number of knots selected

		$\sigma = .3$ $\sigma = .5$		$\sigma = .3$ $\sigma = .5$		$\sigma = .3$ $\sigma = .5$		$\sigma = .3$ $\sigma = .5$	
		<i>AICc</i>		<i>BIC</i>		<i>CV</i>		<i>GCV</i>	
FM	$n = 32$	8.000 (0.009)	8.000 (0.008)	8.000 (0.009)	8.000 (0.008)	8.000 (0.008)	8.000 (0.008)	8.000 (0.009)	8.00 (0.008)
	$n = 64$	16.000 (0.025)	16.000 (0.019)	16.000 (0.04)	16.000 (0.018)	16.000 (0.018)	16.000 (0.018)	16.000 (0.046)	16.000 (0.023)
	$n = 128$	32.000 (0.061)	32.000 (0.062)	32.000 (0.065)	32.000 (0.063)	32.000 (0.064)	32.000 (0.064)	32.000 (0.066)	32.000 (0.063)
	$n = 32$	11.251 (0.016)	7.539 (0.016)	13.093 (0.015)	8.742 (0.015)	9.920 (0.018)	8.057 (0.017)	15.624 (0.016)	11.400 (0.015)
	$n = 64$	18.800 (0.063)	15.504 (0.054)	17.990 (0.054)	13.675 (0.047)	20.078 (0.066)	17.183 (0.056)	21.828 (0.065)	18.892 (0.054)
	$n = 128$	21.260 (0.197)	17.590 (0.184)	19.610 (0.188)	16.840 (0.179)	26.028 (0.213)	20.800 (0.206)	26.210 (0.203)	21.140 (0.191)
MA	$n = 32$	14.756 (0.054)	9.853 (0.05)	17.09 (0.047)	12.155 (0.041)	13.874 (0.053)	10.640 (0.049)	18.149 (0.051)	14.148 (0.049)
	$n = 64$	27.388 (0.111)	27.253 (0.093)	27.622 (0.085)	26.264 (0.075)	27.528 (0.18)	26.199 (0.088)	28.492 (0.179)	28.277 (0.088)
	$n = 128$	42.650 (1.031)	39.374 (0.932)	42.350 (0.909)	38.742 (0.807)	41.960 (1.007)	39.220 (0.943)	43.344 (0.98)	39.768 (0.934)
	$n = 32$	32.000 (0.028)	32.000 (0.028)	32.000 (0.031)	32.000 (0.028)	32.000 (0.031)	32.000 (0.029)	32.000 (0.031)	32.000 (0.029)
	$n = 64$	64.000 (0.107)	64.000 (0.094)	64.000 (0.099)	64.000 (0.094)	64.000 (0.116)	64.000 (0.111)	64.000 (0.108)	64.000 (0.101)
	$n = 128$	128.000 (0.699)	128.000 (0.616)	128.000 (0.607)	128.000 (0.542)	128.000 (0.662)	128.000 (0.627)	128.000 (0.647)	128.000 (0.626)
SMS	$n = 32$	32.000 (0.028)	32.000 (0.028)	32.000 (0.031)	32.000 (0.028)	32.000 (0.031)	32.000 (0.029)	32.000 (0.031)	32.000 (0.029)
	$n = 64$	64.000 (0.107)	64.000 (0.094)	64.000 (0.099)	64.000 (0.094)	64.000 (0.116)	64.000 (0.111)	64.000 (0.108)	64.000 (0.101)
	$n = 128$	128.000 (0.699)	128.000 (0.616)	128.000 (0.607)	128.000 (0.542)	128.000 (0.662)	128.000 (0.627)	128.000 (0.647)	128.000 (0.626)

Table 5. Analogous to Table 4 for true curve  $g_2(x)$ 

		$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$
		<i>AICc</i>		<i>BIC</i>		<i>CV</i>		<i>GCV</i>	
FM	$n = 32$	8.000 (0.01)	8.000 (0.018)	8.000 (0.01)	8.000 (0.017)	8.000 (0.009)	8.000 (0.012)	8.000 (0.009)	8.00 (0.011)
	$n = 64$	16.000 (0.02)	16.000 (0.037)	16.000 (0.021)	16.000 (0.035)	16.000 (0.02)	16.000 (0.034)	16.000 (0.019)	16.000 (0.034)
	$n = 128$	32.000 (0.068)	32.000 (0.136)	32.000 (0.07)	32.000 (0.131)	32.000 (0.068)	32.000 (0.107)	32.000 (0.064)	32.000 (0.079)
	$n = 32$	6.64 (0.016)	6.055 (0.016)	6.798 (0.015)	6.185 (0.016)	6.883 (0.015)	6.427 (0.015)	7.571 (0.015)	7.041 (0.015)
	$n = 64$	7.36 (0.022)	7.155 (0.021)	7.400 (0.019)	6.905 (0.02)	7.900 (0.02)	7.450 (0.02)	8.228 (0.02)	7.670 (0.02)
	$n = 128$	7.080 (0.186)	7.235 (0.186)	6.750 (0.18)	7.56 (0.188)	7.750 (0.185)	7.685 (0.185)	7.705 (0.183)	7.645 (0.199)
FA	$n = 32$	9.452 (0.046)	7.884 (0.046)	9.798 (0.044)	7.815 (0.042)	9.271 (0.045)	8.452 (0.044)	10.684 (0.045)	9.473 (0.044)
	$n = 64$	13.189 (0.037)	12.862 (0.035)	13.789 (0.031)	11.545 (0.028)	14.124 (0.033)	13.459 (0.03)	14.822 (0.033)	13.976 (0.031)
	$n = 128$	15.96 (0.803)	16.175 (0.798)	12.564 (0.792)	17.908 (0.74)	16.364 (0.81)	16.889 (0.788)	16.873 (0.806)	16.984 (0.792)
	$n = 32$	32.000 (0.033)	32.000 (0.06)	32.000 (0.032)	32.000 (0.057)	32.000 (0.031)	32.000 (0.038)	32.000 (0.029)	32.000 (0.034)
	$n = 64$	64.000 (0.097)	64.000 (0.097)	64.000 (0.095)	64.000 (0.105)	64.000 (0.097)	64.000 (0.102)	64.000 (0.097)	64.000 (0.101)
	$n = 128$	128.000 (0.522)	128.000 (0.522)	128.000 (0.515)	128.000 (0.49)	128.000 (0.53)	128.000 (0.516)	128.000 (0.528)	128.000 (0.52)

Table 6. Analogous to Table 4 for true curve  $g_3(x)$ 

		$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$	$\sigma = .3$	$\sigma = .5$
		<i>AICc</i>		<i>BIC</i>		<i>CV</i>		<i>GCV</i>	
FM	$n = 32$	8.000 (0.009)	8.000 (0.009)	8.000 (0.009)	8.000 (0.008)	8.000 (0.008)	8.000 (0.008)	8.000 (0.009)	8.00 (0.009)
	$n = 64$	16.000 (0.044)	16.000 (0.043)	16.000 (0.047)	16.000 (0.044)	16.000 (0.047)	16.000 (0.046)	16.000 (0.048)	16.000 (0.045)
	$n = 128$	32.000 (0.061)	32.000 (0.061)	32.000 (0.066)	32.000 (0.064)	32.000 (0.066)	32.000 (0.067)	32.000 (0.066)	32.000 (0.064)
	$n = 32$	13.382 (0.016)	13.364 (0.018)	15.402 (0.015)	14.847 (0.017)	15.513 (0.018)	14.891 (0.019)	16.302 (0.017)	15.846 (0.017)
	$n = 64$	18.510 (0.062)	15.490 (0.064)	16.674 (0.056)	14.660 (0.058)	19.824 (0.066)	17.648 (0.063)	22.046 (0.111)	19.044 (0.064)
	$n = 128$	21.280 (0.198)	16.490 (0.209)	17.490 (0.192)	15.020 (0.202)	25.03 (0.217)	19.72 (0.223)	25.05 (0.204)	20.090 (0.205]
FA	$n = 32$	15.252 (0.051)	15.929 (0.049)	17.508 (0.047)	17.449 (0.045)	16.350 (0.05)	16.765 (0.049)	17.410 (0.05)	17.688 (0.049)
	$n = 64$	26.226 (0.104)	23.156 (0.094)	23.782 (0.086)	21.534 (0.08)	26.428 (0.174)	24.956 (0.09)	28.660 (0.176)	25.622 (0.09)
	$n = 128$	42.972 (0.961)	33.89 (0.928)	35.820 (0.89)	29.966 (0.859)	44.002 (0.971)	36.656 (0.944)	44.124 (0.969)	35.748 (0.94)
	$n = 32$	32.000 (0.027)	32.000 (0.027)	32.000 (0.029)	32.000 (0.029)	32.000 (0.03)	32.000 (0.03)	32.000 (0.029)	32.000 (0.029)
	$n = 64$	64.000 (0.103)	64.000 (0.106)	64.000 (0.1)	64.000 (0.105)	64.000 (0.117)	64.000 (0.117)	64.000 (0.108)	64.000 (0.11)
	$n = 128$	128.000 (0.632)	128.000 (0.621)	128.000 (0.598)	128.000 (0.569)	128.000 (0.647)	128.000 (0.632)	128.000 (0.635)	128.000 (0.623)

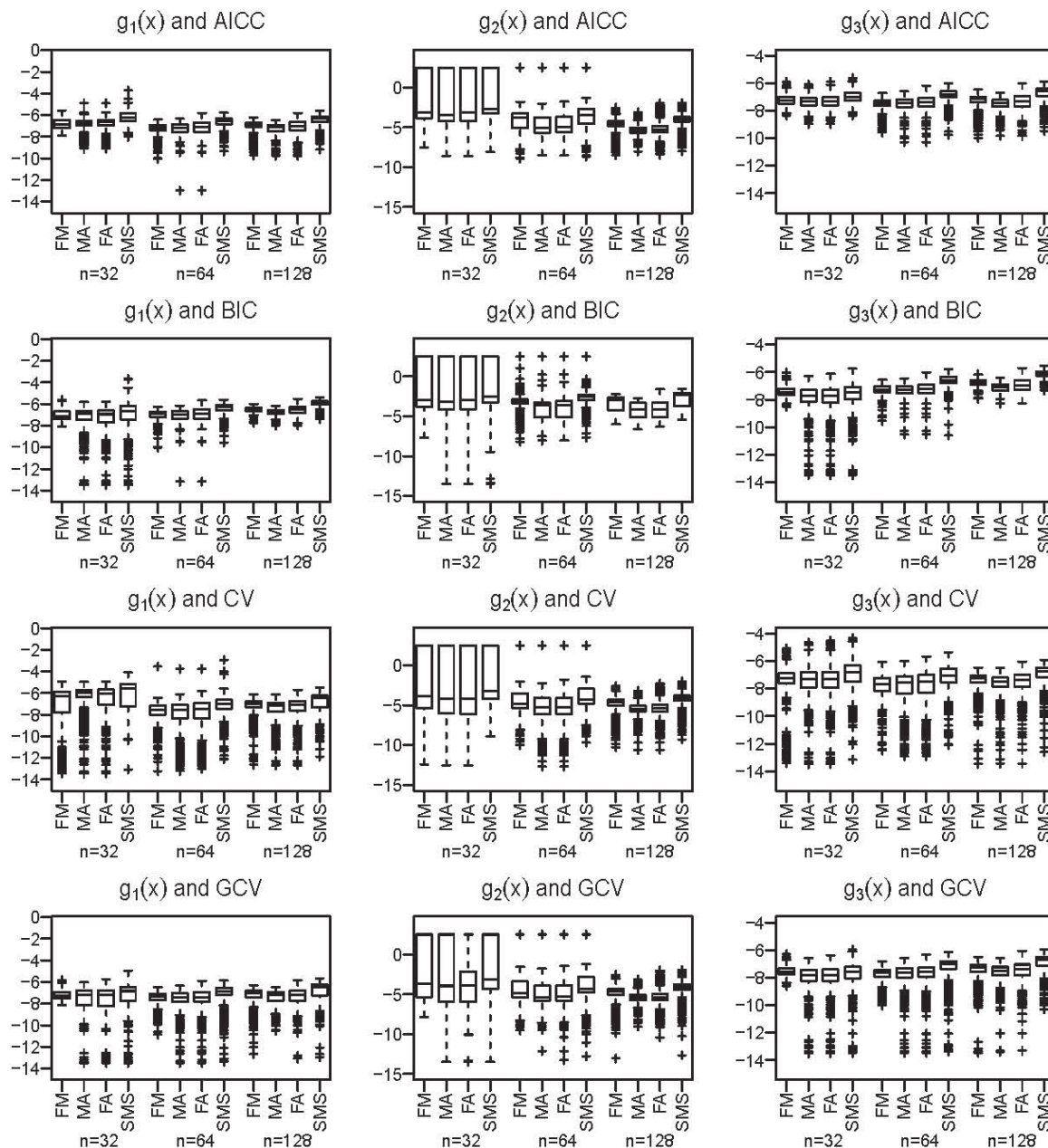
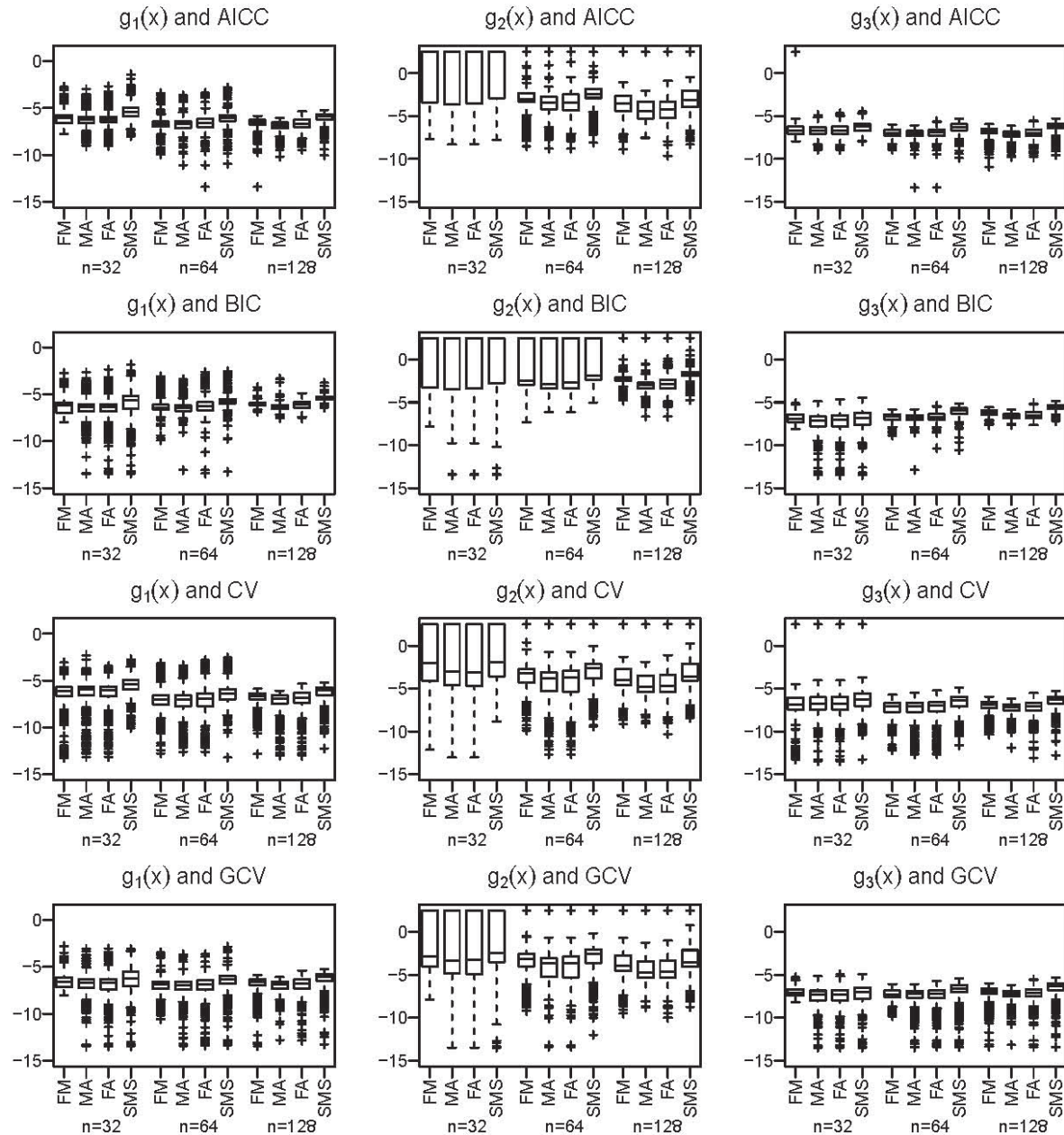


Figure 4. The smoothing parameter  $\lambda$  selected on the log scale for each simulated data set, when  $\sigma = .3$  under different simulation settings. Top left corner panel: Under each sample size, the three box plots contain the smoothing parameter selected when the true curve is  $g_1(x)$  when AICC is used for smoothing parameter selection for the FM, MA, FA, and SMS, respectively. Remaining panels: Analogous top left corner panel for the true curve and smoothing parameter selection method shown in the title of the panel



Figure 5. Analogous to Figure 4 when  $\sigma = .5$ 

## 6. Conclusions

We compared the performance of three knot selection methods using the median of (3) to provide guidance in selecting the number of equidistant knots in a penalized regression spline model. In addition, for comparison, we also considered a smoothing spline estimate by placing a knot at each data point. One should keep in mind that our findings arise from a simulation study. However, under our simulation settings, the *FM* tended to perform just as well or better than the *MA*, *FA*, and *SMS* across all smoothing parameter selection methods considered. The boxplots given in Figures 2 and 3 show that *GCV* with the *FM* generally exhibits lower variability than the other smoothing parameter selection methods with the *FM* in terms of (4). Since the smoothing parameter selection methods considered have the same computational complexity, *GCV* should be preferred. Overall, our results suggest that when using equidistant knots in a penalized regression spline, the default rule provided by the *FM* with *GCV* should be used.



## Acknowledgements

We thank the editor and two reviewers for helpful comments and suggestions. This research was supported in part by NSF grant HRD-1137774.

## References

- Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2), 453-510. <http://dx.doi.org/10.1214/aos/1176347115>
- Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4), 377-403. <http://dx.doi.org/10.1007/BF01404567>
- Eilers, P. H. C., & Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6), 637-653. <http://dx.doi.org/10.1002/wics.125>
- Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed.). Marcel Dekker, Inc.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *Elements of Statistical Learning* (1st ed.). Springer. <http://dx.doi.org/10.1007/978-0-387-21606-5>
- Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, 60, 271-293. <http://dx.doi.org/10.1111/1467-9868.00125>
- Kauermann, G., & Opsomer, J. D. (2011). Data-driven selection of the spline dimension in penalized spline regression. *Biometrika*, 98(1), 225-230. <http://dx.doi.org/10.1093/biomet/asq081>
- Lee, T. C. M. (2003). Smoothing parameter selection for smoothing splines: A simulation study. *Computational Statistics & Data Analysis*, 42(1), 139-148. [http://dx.doi.org/10.1016/S0167-9473\(02\)00159-7](http://dx.doi.org/10.1016/S0167-9473(02)00159-7)
- Luo, Z., & Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association*, 92(437), 107-116. <http://dx.doi.org/10.1080/01621459.1997.10473607>
- Ngo, L., & Wand, M. P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, 9(1), 1-54.
- R Core Team. (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4), 735-757. <http://dx.doi.org/10.1198/106186002853>
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression* (1st ed.). Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511755453>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464. <http://dx.doi.org/10.1214/aos/1176344136>
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill.
- Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, 47, 1-52.
- Spiriti, S., Eubank, R., Smith, P. W., & Young, D. (2013). Knot selection for least-squares and penalized splines. *Journal of Statistical Computation and Simulation*, 83(6), 1020-1036. <http://dx.doi.org/10.1080/00949655.2011.647317>
- Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25(4), 1371-1470. <http://dx.doi.org/10.1214/aos/1031594728>
- Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics*, 15(4), 443-462. <http://dx.doi.org/10.1007/s001800000047>
- Wand, M. P., & Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalised splines. *Australian & New Zealand Journal of Statistics*, 50(2), 179-198. <http://dx.doi.org/10.1111/j.1467-842X.2008.00507.x>
- Wang, Y., & Wahba, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison

to Bayesian confidence intervals. *Journal of Statistical Computation and Simulation*, 51(2-4), 263-279.  
<http://dx.doi.org/10.1080/00949659508811637>

Welham, S. J., Cullis, B. R., Kenward, M. G., & Thompson, R. (2007). A comparison of mixed model splines for curve fitting. *Australian & New Zealand Journal of Statistics*, 49(1), 1-23.  
<http://dx.doi.org/10.1111/j.1467-842X.2006.00454.x>

### Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).