DISSERTATION

NONPARAMETRIC FUNCTION SMOOTHING: FIDUCIAL INFERENCE OF FREE KNOT SPLINES AND ECOLOGICAL APPLICATIONS

Submitted by

Derek L. Sonderegger

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2010

COLORADO STATE UNIVERSITY

April 5, 2010

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY DEREK L. SONDEREGGER ENTITLED NONPARAMETRIC FUNCTION SMOOTHING: FIDUCIAL INFERENCE OF FREE KNOT SPLINES AND ECOLOGICAL APPLICATIONS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Barry R. Noon

Hari Iyer

Advisor: Haonan Wang

Co-Advisor: Jan Hannig

Department Chair: F. Jay Breidt

ABSTRACT OF DISSERTATION

NONPARAMETRIC FUNCTION SMOOTHING: FIDUCIAL INFERENCE OF FREE KNOT SPLINES AND ECOLOGICAL APPLICATIONS

Nonparametric function estimation has proven to be a useful tool for applied statisticians. Classic techniques such as locally weighted regression and smoothing splines are being used in a variety of circumstances to address questions at the forefront of ecological theory.

We first examine an ecological threshold problem and define a threshold as where the derivative of the estimated functions changes states (negative, possibly zero, or positive) and present a graphical method that examines the state changes across a wide interval of smoothing levels. We apply this method to macro-invertabrate data from the Arkansas River.

Next we investigate a measurement error model and a generalization of the commonly used regression calibration method whereby a nonparametric function is used instead of a linear function. We present a simulation study to assess the effectiveness of the method and apply the method to a water quality monitoring data set.

The possibility of defining thresholds as knot point locations in smoothing splines led to the investigation of the fiducial distribution of free-knot splines. After introducing the theory behind fiducial inference, we then derive conditions sufficient to for asymptotic normality of the multivariate fiducial density. We then derive the fiducial density for an arbitrary degree spline with an arbitrary number of knot points. We

then show that free-knot splines of degree 3 or greater satisfy the asymptotic normality conditions. Finally we conduct a simulation study to assess quality of the fiducial solution compared to three other commonly used methods.

Derek L. Sonderegger
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Spring 2010

ACKNOWLEDGEMENTS

Without the help of my committee members, this dissertation would not have been possible. Jan Hannig has been a wonderful instructor, advisor, research partner and friend. Haonan Wang was instrumental in suggesting the subject area and has helped me to see that there are research opportunities in even the most applied areas. Barry Noon and the Noon Lab have helped keep me focused on problems with ecological significance. I have enjoyed the opportunity to offer statistical help while learning as much ecology as possible. Will Clements has been a wonderful co-author and a pleasure to work with.

The CSU PRIMES program was invaluable for providing the opportunity to meet people in ecology and develop my professional relationships. The funding allowed me to explore my applied research areas more deeply. My funding has come from a variety of sources and I would like to acknowledge Evironmental Protection Agency (EPA) STAR Program, grant #R832441, the National Science Foundation (NSF), IGERT grant #DGE-0221595, and NSF grant #DMS- 0706761.

I would like to thank my parents for their love and support throughout my long college career. At times it must have seemed like I would never graduate. In particular I would like to thank my father who encouraged me to take math courses for fun as an undergraduate. I wouldn't be where I am if it weren't for that suggestion.

Finally I want to thank Aubrey for her patience, understanding, and unwavering support. She has been a rock of consistency as I've dealt with the euphoria and depression that comes with any non-trivial challenge.

DEDICATION

To Aubrey and our future together.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION TO NONPARAMETRIC SMOOTHING

1.1 Introduction to locally weighted polynomials

Locally weighted polynomials were first introduced by Cleveland (1979) and have been successfully used in many application areas. Researchers are often interested in finding a functional relationship between a random variable Y and some predictive variable x. If we assume that $Y = f(x) + \epsilon$, then knowing that any continuous curve can be accurately represented by a polynomial over short regions leads us to restrict our attention to a small region around a point of interest, say x^*. Using points "close" to x^*, we use the standard linear model methods to fit a degree p polynomial to the data and use that polynomial to estimate $f(x^*)$. We denote the polynomial as $g(x|\hat{\boldsymbol{\alpha}})$.

To address the question of what is meant by "close", we weight the data points using a kernel function $K([x_i - x^*]h^{-1})$ where h is a parameter that controls how close a data point must be to x^* to have a large impact on the regression. The function $K(\cdot)$ is usually a symetric positive function with $K(x)$ decreasing as $|x|$ increases. The most common choice for $K(\cdot)$ is the standard normal density.

Once the data weights have been calculated, the weight matrix is

$$\boldsymbol{W}_{x^*} = \text{diag}\left\{K\left(\frac{x_1 - x^*}{b}\right), \ldots, K\left(\frac{x_n - x^*}{b}\right)\right\}$$

and the regression polynomial has coefficients $\hat{\boldsymbol{\alpha}}_{x^*} = \left(\boldsymbol{X}^T\boldsymbol{W}_{x^*}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}_{x^*}\boldsymbol{y}$. Taking $\hat{f}(x) = g(x|\hat{\boldsymbol{\alpha}})$ we obtain a function estimate.

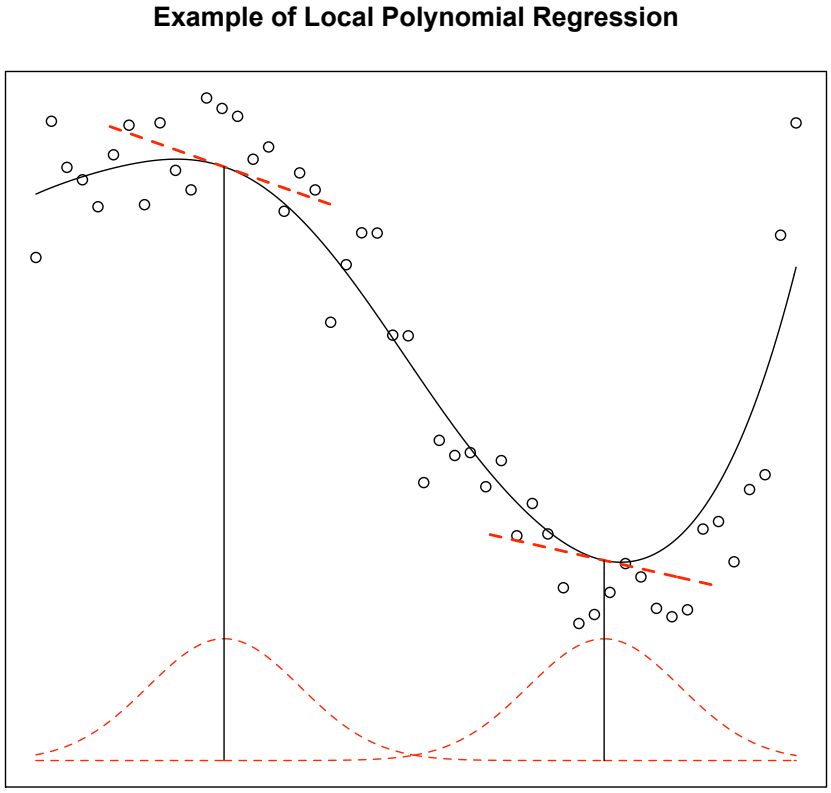

Figure 1.1.1: An illustration of fitting a local polynomial regression.

The selection of the bandwidth parameter can be chosen by cross-validation and or by the problem context. For example, if we are given yearly rainfall data and we are interested in decadal osscilation, a bandwidth of $h = 5$ to 10 would capture that trend better than a smaller bandwidth.

1.2 Introduction to smoothing splines

Given a interval of $\mathbb{R}$ say $[a, b]$ and a knot point $t \in [a, b]$, a spline on $[a, b]$ is defined by polynomials on $[a, t]$ and $[t, b]$ but a smoothness condition is enforced at $x = t$. In general, if degree p polynomials are fit to the two sections, then the $(p-1)$th derivative should exist at $x = t$.

1.2.1 Truncated polynomial basis

We consider a degree p polynomial spline with κ knotpoints t_k and $m + \kappa$ coefficients. Let $\boldsymbol{t}$ be the vector of knot points and $\boldsymbol{\alpha}$ be the vector of coefficients and $\boldsymbol{\theta} = \left\{\boldsymbol{\alpha}^T, \boldsymbol{t}^T\right\}^T$. The spline can be written using many different basis functions, but here we consider the piecewise polynomial definition:

$$g(x_i|\boldsymbol{\theta}) = \sum_{j=0}^{p} \alpha_j x_i^j + \sum_{k=1}^{\kappa} \alpha_{p+k} \left(x_i - t_k\right)_+^p$$

where

$$(u)_+ = \begin{cases} 0 & \text{if } u < 0 \\ u & \text{otherwise} \end{cases}$$

is the trunction operator which has higher precendence than exponentiation, ie $(-1)_+^2 = \left[(-1)_+\right]^2 = 0$.

With known knot points, fitting the vector of coefficients $\boldsymbol{\alpha}$ is straightforward using standard linear models methods. Let $\boldsymbol{X}$ be the matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \dots & x_1^p & (x_1 - t_1)_+^p & \dots & (x_1 - t_\kappa)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^p & (x_n - t_1)_+^p & \dots & (x_n - t_\kappa)_+^p \end{bmatrix}$$

then $\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

1.2.2 B-spline basis

B-splines are an appealing spline basis that are commonly used in numerical applications because the basis vectors are close to orthogonal. The downside is that the coefficient interpretation is not as obvious as in the truncated polynomial case. The recursion based definition that is typically used was first given by deBoor 1972. The order 1 (degree 0) B-splines are

$$B_j^1(x) = \begin{cases} 1 & \text{if } t_j \le x < t_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

The general formula for a B-spline of arbitrary order m is

$$B_j^m(x) = \omega_j^m(x) B_j^{m-1}(x) + (1 - \omega_{j+1}^m(x)) B_{j+1}^{m-1}(x) \tag{1.2.1}$$

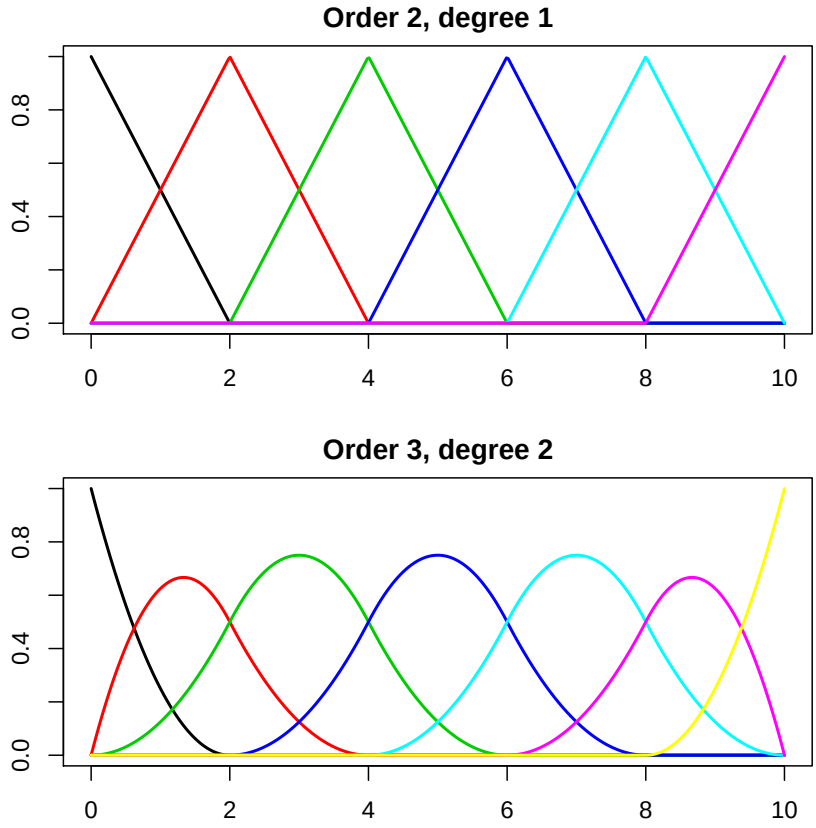

Figure 1.2.1: B-spline basis functions

where

$$\omega_j^m(x) = \frac{x - t_j}{t_{j+m-1} - t_j}. \tag{1.2.2}$$

1.2.3 Penalized Splines

If the knot point locations are known, then standard linear model theory applies. Unfortunately the question of knot point location is not trivial. Typically too many knot points are used and a penalty term is added during the minimiazation of the squared error. This idea can be found in Parker and Rice 1985, O'Sullivan 1986) and the penalized model commonly used today was given by Eilers and Marx (1996) is often called P-splines. The b-spline basis functions are regression coefficients are selected by minimizing

$$S = \sum_{i=1}^{n} \{y_i - g(x_i|\boldsymbol{\alpha})\}^2 + \lambda \sum_{j=3}^{p+\kappa} \left(\Delta^2 \alpha_j\right)$$

where $\Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$. The asymptotic theory addressed in the case of increasing knots by Hall and Opsomer (2005).

1.2.4 Free Knot selection

The case where the knot points are not fixed, and are to be estimated is often called *free-knot splines*. The asymptotic behavior maximum likelihood solution is given by Stone and Huang (2002). Mao and Zhao (2003) discuss the free-knot splines and the give confidence intervals for the parameters. They note that the likelihood surface often has many saddle points and local maxima and therefore care must be taken when finding the global maximum. They start the numerical optimization program with a variety of initial knot point locations to try to avoid local maxima. The optimization routine they used was the routine DBSVLS from software vender International Mathematical and Statistical Libraries.

The Bayesian solution is to make the knot point locations be a regular model parameter and the standard Bayesian method works quite well. DiMatteo et al. (2001) add a model selection step for deciding on the optimal number of knot points by using reversable-jump Markov Chain Monte Carlo and can therefore give a posterior probability of a particular number of knots. Software is available from Robert Kass's website http://lib.stat.cmu.edu/~kass/bars/bars.html.

Chapter 2

USING SIZER TO DETECT THRESHOLDS IN ECOLOGICAL DATA

2.1 Preface

This chapter contains work that was done in collaberation with Haonan Wang, Will Clements and Barry Noon and has been published elsewhere (Sonderegger et al., 2008). My contribution to the work was in several parts. First, I contributed a solid understanding of nonparametric function estimation. Secondly, I was solely responsible the software development of the SiZer package in R. Finally I was responsible for synthesizing the information from each discipline and brought all the ideas together into the resulting paper.

In the field of ecology, collaboration is common and a necessary part of the advancement of the field (Ewel, 2001). As the field becomes more quantitative, there is a need 'outsource' the understanding of the finer technical details in statistical theory to statisticians interested in ecology. This creates a professional niche space that I intend to work in. The following work should not be viewed as my ecological knowledge, but rather as a reflection of my ability to in an interdisciplinary setting.

2.2 Introduction

Theoretical and empirical studies suggest that some ecosystems may show abrupt, non-linear changes in one or more response variables in response to environmental drivers (May 1977; Connell and Sousa 1983; Knowlton 1992; Estes and Duggins 1995; Groffman et al. 2006). Shifts to alternative stable states have been reported in a variety of ecosystems, including lakes, coral reefs, deserts, and oceans

(Scheffer et al., 2001). These shifts can be triggered by natural disturbance, such as fire or flooding, or anthropogenic factors, such as climate change, nutrient accumulation, exotic species, and toxic chemicals. Although communities may recover from natural disturbance through successional processes, human-induced disturbances are often unprecedented and move ecological systems to novel, alternative states (Holling 1986; Folke et al. 2002). In addition, if ecosystems are chronically stressed due to natural or anthropogenic disturbances, such systems may move to alternative states that remain stable, even when the stressors are removed (eg Carpenter 2001; Scheffer et al. 2001; van Nes et al. 2002; Scheffer and Carpenter 2003).

One can consider thresholds as ecological non-linearities, where substantial changes in an ecological state variable are a consequence of small, continuous changes in an independent (stressor) variable (Muradian 2001). The point or region at which rapid change initially occurs defines the threshold. Near this point, small changes in stressor intensity produce large effects on state variables. Unfortunately, there can be an inherent arbitrariness to the threshold concept, because it does not take into account whether the change in the value of the state variable is ecologically relevant. Statistical models have been developed in other disciplines, to detect breakpoints in non-linear response functions, but it is not always clear which models are appropriate for a particular ecological dataset.

Here, we demonstrate a method by Chaudhuri and Marron (1999) that makes few model assumptions and is therefore suitable for a broad range of ecological problems. Their method, Significant Zero crossings (SiZer), applies a non-parametric smoother to the stressor–response data, and then examines the derivatives of the smoothed curve to identify the existence of a threshold. To illustrate this method, we consider benthic macroinvertebrate data collected on the Arkansas River, a metal-polluted stream in Colorado. We use SiZer to examine the nature of the threshold(s) and to select between two competing threshold models. We then use SiZer in a multivariate setting by examining the first axis of the canonical discriminant analysis for the same

dataset. Similar to principal components analysis, this axis is the linear combination of the 20 dominate taxa, that contains the most yearly variation.

2.3 Data

The Arkansas River (Figure 1) is located in the southern Rocky Mountain ecoregion of Colorado. Mining operations in this watershed have had a major impact since the mid-1800s, when gold was discovered near Leadville, Colorado. Concentrations of heavy metals, particularly cadmium (Cd), copper (Cu), and zinc (Zn), are greatly elevated downstream from Leadville and often exceed acutely toxic levels (Clements, 2004a). Over the past 18 years (1989–2006), physiochemical characteristics, habitat quality, heavy metal concentrations, and the responses of macroinvertebrate communities were quantified at several stations in the Upper Arkansas River Basin. In 1993, 4 years after this research program began, state and federal agencies initiated a large-scale restoration program designed to improve water quality in the Arkansas River. To quantify recovery, we examined temporal changes in the abundance of metal-sensitive mayflies (Ephemeroptera: Heptageniidae) collected in the fall from 1989 to 2004. During each of the 18 years, five replicate samples were taken. The mayfly counts were transformed (square root) to stabilize the variance. Recovery was defined as the threshold where mayfly abundance became asymptotic. While the study was primarily concerned with the effect of heavy-metal pollution, this paper uses time as the independent variable for clearer illustration of the method.

2.4 SiZer approach and derivative definition of thresholds

An intuitive way of defining a threshold for a state variable that is a continuous function of an environmental driver is to consider where the function's derivatives change significantly. Non-parametric smoothers provide a method for finding a smooth response function that is data driven and requires only weak assumptions. Smoothing splines (Green and Silverman 1993; Wahba 1975), LOESS (Cleveland and

Figure 2.3.1: The Arkansas River after restoration efforts, 200 meters downstream of the confluence with the California Gulch.

Devlin, 1988), and locally weighted polynomial regression (Fan and Gijbels, 1996), are well known. These techniques result in an estimated smooth response function, the estimated derivative(s), and confidence intervals (CI) for the functions and derivatives. The SiZer methodology can be implemented using any of these techniques, but we have restricted our discussion to locally weighted polynomials. All SiZer CIs in this chapter are reported at the 95% level, based on Hannig and Marron (2006) row-wise intervals.

In the Arkansas River data, both a piecewise linear (PL; Barrowman and Myers 2000; Toms and Lesperance 2003) or bent-cable (BC; Chiu et al. 2006) model would fit the data. Both models assume a linear relationship with a single threshold. The difference is that the PL model assumes an abrupt transition between the linear sections, while the BC model assumes a quadratic bend connecting the two linear pieces. The PL model is a simple case of the BC model where the half-width of the bend is zero.

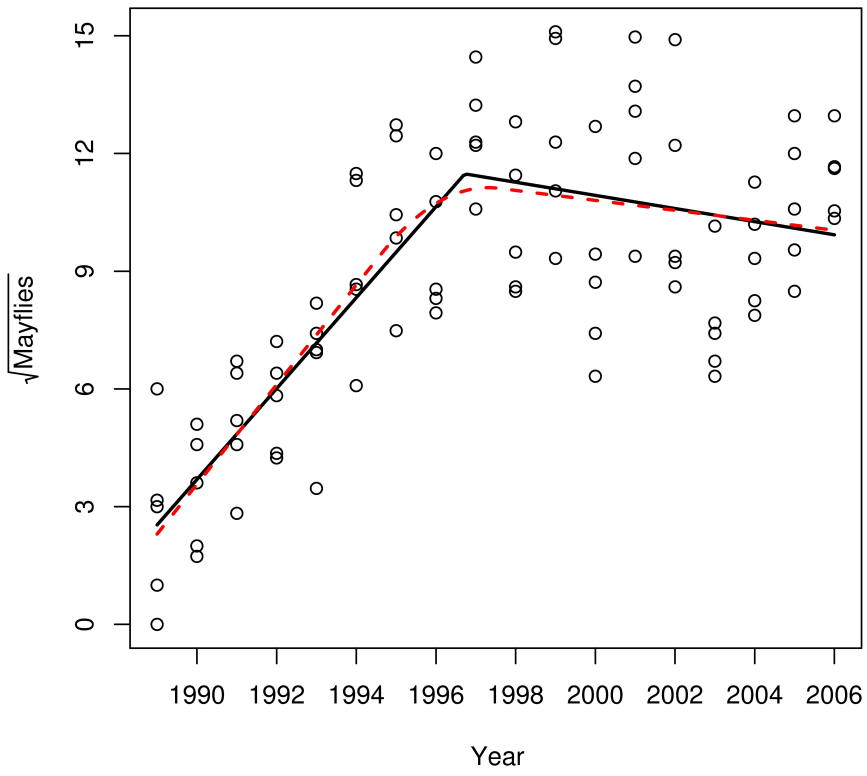

Figure 2.4.1: The Arkansas River mayfly data fit by the piecewise linear (black, solid, threshold = 1996.7, CI = (1994.6, 1997.5)) and bent-cable (red, dashed, threshold = 1996.1, CI = (1989.0, 2000.0)) models.

Traditional model selection methods, such as Akaike's information criterion (Burnham and Anderson, 2002) do not decisively rate one model over the other. The difference between AIC values is 1.82 in favor of the PL model. However, an inverted likelihood ratio test (Seber and Wild, 1989) resulted in a 95% confidence interval for the half-width of the quadratic bend that does not contain zero$(0.04, 5.52)$. Point estimates of the threshold were similar for the PL and BC models $(1996.7, 1996.1)$, but 95% bootstrap confidence intervals for the threshold were $(1994.6, 1997.5)$ and $(1989.0, 2000.0)$ respectively.

Using a non-parametric smoother, every point along the independent axis can be classified into one of three states: the estimated slope is positive (ie the CI of the first derivative contains only positive values), possibly zero (the CI contains zero), or negative (the CI contains only negative values). Each point could be similarly classified using the estimated second (or higher order) derivative.

Many interesting relationships can be found by examining the state changes of the derivatives. By noting how many times the state of the first derivative changes, inference about where the true relationship is increasing or decreasing can be made. The Arkansas River data show that abundance of mayflies is clearly increasing, then flattens out and seems to decrease slightly near $x = 2002$. The second derivative contains information about the curvature of the data. At small x values, the second derivative could be zero, indicating that there is little or no curvature. However, between 1994 and 2000, the second derivative is significantly negative, indicating that the function is concave down and providing support in the data for the BC model.

There is no mathematical reason to partition the curve by where the derivative is different than zero. A similar procedure could be implemented to partition the x-axis into segments that have a derivative different than 5, for example. However, the choice to partition on $\hat{f}'(x) = 0$ is appropriate in many ecological problems in which the increase or decrease of the response variable is of interest. Moreover, detecting a

change in the rate of increase (or decrease) can be made by second derivative, because a change of rate causes curvature.

Given a small number of models that are thought to quantify a researcher's beliefs and hypotheses about a system, traditional model selection methods such as AIC, AICc, Mallows Cp, and goodness-of-fit tests (Kutner et al., 2005) fail to directly explain why one model is selected over another. By using a derivative-based method in the process of model selection, the researcher can investigate how the data support one model over the other, or what assumptions in a model are being violated. The non-parametric implementation allows the researcher to examine a broad range of questions, including the number of thresholds in a system.

2.5 Smoothing bandwidths

One complication of the derivative approach is the estimation of the smoothed function and its derivative(s). Most non-parametric smoothing algorithms, including smoothing splines and locally weighted polynomial regression, have a tuning parameter that controls the smoothness of the resulting curve. By manipulating this parameter, the resulting smoothing function can range from a simple linear regression to perfectly (over)fitting the data. There are several methods for selecting the tuning parameter (Fan and Gijbels 1995; Ruppert et al. 1995; Hengartner et al. 2002), but none are uniformly superior.

SiZer, as proposed and implemented by Chaudhuri and Marron (1999) and Hannig and Marron (2006), uses the idea of locally weighted polynomial regression (Fan and Gijbels 1996; Loader 1999). When the weight function (also called a kernel function) is the normal density curve, the level of smoothing is controlled by the standard deviation of the kernel. For a given tuning parameter h, and a given point x_0, $\hat{f}'(x_0)$ is obtained by weighting the data points according to a normal curve centered at x_0, with a standard deviation $\sigma = h$. This means that data close to x_0 (eg within $\pm h$) have a large influence over the smoothing function, data between h and $2h$ away are

less influential, and data farther than $2h$ from x_0 have only a slight influence. In locally weighted polynomial regression, the tuning parameter h is the width parameter of the kernel function and is commonly referred to as the bandwidth. Other common choices for the kernel function include the uniform density and triangle density functions. Here, we use the normal kernel.

The novel aspect of SiZer (Chaudhuri and Marron, 1999) is that it considers all reasonable bandwidth values and exploits the notion that different values provide different information about the data. The SiZer approach explores how the derivative changes along the independent axis, as well as across the range of bandwidth values, and displays this information in one image (Figure 3). To read the SiZer map, first notice that the y-axis represents the bandwidth parameter h, displayed in units of $\log(h)$ for visual clarity. Wherever the map is blue, the derivative is significantly increasing; wherever it is purple, the derivative is possibly zero, and wherever it is red, the derivative is significantly decreasing. At very small bandwidths, $\hat{f}'(x_0)$ is influenced by a small number of data points, and gray areas in the SiZer map indicate that the estimated effective sample size (Chaudhuri and Marron, 1999) is less than 5. The white lines give a visual representation of the size of the bandwidth. The horizontal distance between the lines is drawn to be $2h$, indicating the effective width of the locally weighted polynomial.

To demonstrate the effect of bandwidth on the smoothing function, Figure 3 displays Arkansas River data with three different choices of bandwidth h, each highlighted by the horizontal black line in the adjacent SiZer maps. The top row of graphs represents a smoothing parameter that is too large ($h = 7$) and has over-smoothed the data and fails to detect the transition from an increasing to a flat (or possibly decreasing) function. At this scale of view, the first derivative SiZer row is completely blue, suggesting an increasing function with no threshold. The second derivative is negative (red), indicating that there is downward curvature in the function. At an intermediate level of smoothing ($h = 2$), the smoother captures the initial increasing

section and transition to a flat function, as indicated by the first derivative SiZer map. The second derivative map shows that the function is reasonably linear, except for a region of concavity in the middle and a second region of convexity near 2004. The piecewise linear and bent-cable models presented in Figure 2 capture features that are visible at this scale. At a very low level of smoothing ($h = 0.75$), at any given point, the smoother is being influenced by a very small number of data points (estimated effective sample size ˜15). Consequently, the power of testing if the derivative(s) are not equal to zero is low. In this study, researchers were not particularly interested in annual variation, but wanted to detect trends occurring over multiple years associated with recovery of this system; therefore, bandwidths $h > 1$, $(\log_{10} h > 0)$ should be considered. After considering each of these bandwidths, particularly h near 2, the data support the bent-cable model over the piecewise linear model, because of the curvature near the threshold at the intermediate bandwidths. However, the SiZer analysis suggests considering a model with two threshold points to account for the decrease near 2003.

Because the SiZer map contains information at many different scales, there is seldom a "best" bandwidth to examine. Therefore, we recommend an evaluation of the derivative at different resolutions of the data. Just as, when viewing a tree from a distance, at large bandwidths only gross features are discernible, so, as the observer gets closer to the tree (ie as the bandwidth decreases), the overall pattern cannot be seen, but smaller features come into focus. Only by examining the function across a range of bandwidths can a researcher gain a clear understanding of the data.

SiZer cannot, however, always estimate the location of the threshold. Because $\hat{f}$ is calculated from nearby values, if f has a threshold at $x = \alpha$, then $\hat{f}$ is not necessarily first affected by the threshold at $x = \alpha$. Furthermore, where $\hat{f}$ is affected by the threshold changes with the bandwidth. This phenomenon can be seen in the Arkansas River example. The threshold from an increasing to a flat function drifts from near 1995 to 2004 as the bandwidth increases.

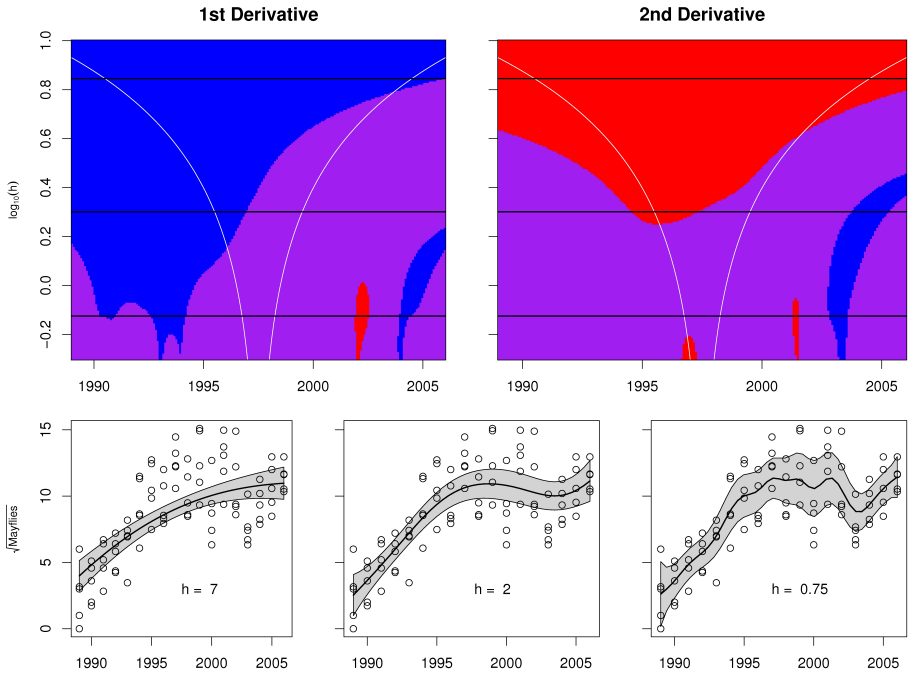

Figure 2.5.1: SiZer maps of the Arkansas River data and associated smoothing functions at three different bandwidths. SiZer maps categorize the derivative as positive (blue), negative (red), or possibly zero (purple). The black lines in the SiZer maps show bandwidth parameter corresponding to the three smoothing functions. The bandwidth $h = 7$ is clearly over-smoothing the data and does not capture the flatness (or decrease) in the second half of the data. The bandwidth $h = 0.75$ is under-smoothing the data and is being affected by random perturbations in the data.

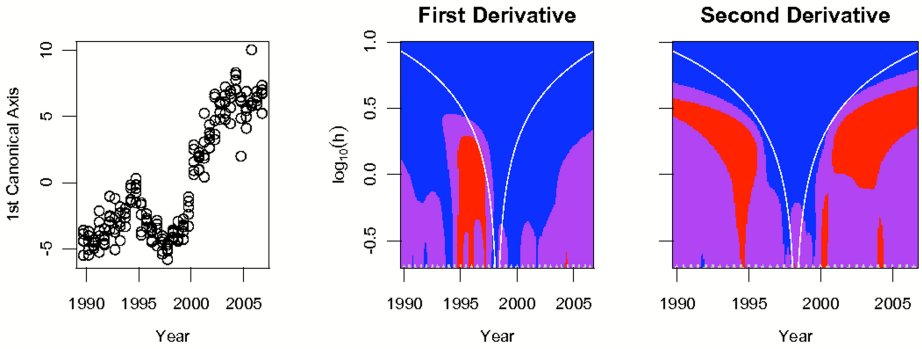

Figure 2.6.1: A scatterplot of the first axis of a canonical discriminate analysis of the Arkansas River data versus time, along with its first and second derivative SiZer maps.

2.6 Using SiZer to identify multiple thresholds

Multivariate analysis of macroinvertebrate data collected from the Arkansas River provided an opportunity to investigate ecological thresholds in community composition over the duration of the monitoring project. In this example, canonical discriminant analysis was used to examine differences among years, based on abundance of the 20 dominant taxa. A threshold response in this example represents an abrupt shift in community composition from one year to the next. Ignoring the issue of non-independence of observations in multivariate space, we applied SiZer to the first canonical axis, which explained 58.2% of the total variation (Figure 4). The first derivative shows a generally increasing function, but there is a sharp decreasing trend between 1995 and 1997. These results reflect macroinvertebrate community responses to changes in water quality from 1989 to 2006. Heavy metal concentrations declined from 1989–1994, increased abruptly in 1995 and 1996, and then declined again as a result of ongoing restoration in the Arkansas River (Clements, 2004a).

The second derivative from the SiZer plot shows two distinct thresholds. The first, near 1996, is a change from concave down to concave up. The second, near 2000, is a change from concave up to concave down. At intermediate smoothing levels near 1996, the 95% CI goes from being less than zero, to containing zero, to being greater

than zero. For the second threshold near 2000, which is a result of macroinvertebrate community recovery after improvements in water quality, the 95% CI changes from being greater than zero, to containing zero, to being less than zero.

2.7 Discussion

Fitting mathematical models to observed data is difficult, in part because of the uncertainty in model selection. Traditional model selection methods tend to encourage examination of vast numbers of models. Burnham and Anderson (2002) address this issue by differentiating exploratory studies from confirmatory ones. When this is not possible, inference must be made carefully to avoid inflated Type I errors (including a variable in the model when, in fact, it has no effect on the response), due to picking the "best" model for the data. As presented here, SiZer is most naturally used in exploratory studies, but if sufficient data are available, part of the data may be used for exploratory model selection and the other part used for inference. Burnham and Anderson (2002) also strongly advocate only examining models that have sound scientific explanations. Since SiZer encourages the practitioner to create an appropriate response function from the SiZer map, the ecological justification and empirical evidence for a particular form of the response function can be combined. This should result in small numbers of models, to be used in subsequent model selection steps or model averaging. SiZer also separates the question of statistical significance from ecological significance, by showing the statistically significant features at each bandwidth and then allowing the researcher to decide which features are important.

Fitting threshold models is particularly difficult, because researchers often assume that the existence and number of thresholds is known. For example, a piecewise linear analysis will always find a single threshold, regardless of whether the true functional relationship contains no threshold or multiple thresholds. SiZer can provide insight into the number of thresholds and general form of the relationship. Unfortunately, SiZer cannot provide estimates and confidence regions for the thresholds it

detects. SiZer can only be used to select a model; a model fitting procedure such as maximum likelihood estimation must be used subsequently. Furthermore, by definition, SiZer can only address thresholds in the context of changes in state of the derivative. Uniform or gradual changes that lead to irreversible state changes are not detectable by SiZer.

The mathematics that SiZer employs can be readily extended to multiple dimensions by using the multidimensional gradient rather than the one-dimensional derivative (Godtliebsen et al., 2002); however, SiZer's main strength, its graphical presentation, cannot be easily extended past two dimensions. Covariates can be accounted for in an additive fashion, by using SiZer on the nonparametric portion of a generalized additive model (GAM). Finally, one direction for future research is to extend SiZer to work with local quantile regression instead of local mean regression. Pointwise estimates of the quantile function should not be difficult to obtain, but appropriate row-wise confidence intervals might be.

A Matlab implementation of SiZer has been made available by S Marron (www.stat.unc.edu/faculty/marron/marron_software.html). An R package of SiZer, along with code for the piecewise linear and bent-cable models, is available on the Comprehensive R Archive Network (www.cran.r-project.org).

Chapter 3

EFFECTS OF MEASUREMENT ERROR ON THE STRENGTH OF CONCENTRATION-RESPONSE RELATIONSHIPS IN AQUATIC TOXICOLOGY

3.1 Preface

This chapter contains work that was done in collaboration with Haonan Wang, Hwang Yao, and Will Clements and has been published elsewhere (Sonderegger et al., 2009). Much of the initial statistical initial was done by Hwang Yao (Yao, 2008), but was not presentable to non-statisticians. My contribution was first to consider what simulations would be of most interest to ecologists and second to apply the method to field data and present the statistical and ecological reasons for modeling the Fall and Spring CCU values in the manner that we did. Finally I was responsible for combining the statistical and ecological information for the resulting paper.

3.2 Introduction

In ecological risk assessment, it is common to take a measurement of contaminant concentration and use that single observation to represent exposure that an organism would experience over some duration of time. This is a legitimate practice if there is very little temporal or spatial variation in contaminant concentrations, however, that is rarely the case. While spatial variation in contaminant concentrations resulting from patchy distributions in the field may be quantified by taking replicate samples, experimental designs rarely account for significant temporal variation. Because

of natural variation in stream discharge, temperature, pH and other physicochemical characteristics, water contaminant concentrations often show significant temporal variation. Clements (2004a) reported significant seasonal variation in heavy metal concentrations associated with stream discharge. Researchers measuring diel (24 h) cycles of heavy metals have reported that concentrations of Zn can increase by a factor of five from afternoon minimum values to early morning maximum levels (Nimick et al., 2003). While every attempt can be made to collect water quality samples at the same time every year, natural fluctuations will cause the samples to be taken at different points of the daily and annual cycles.

Measurement error is a commonly studied topic in statistics (Fuller, 1987; Carroll et al., 2006; Cheng and van Ness, 1999). The most well known result is that for simple linear regression, uncertainty in measuring the predictor variable leads to a substantial underestimation of the relationship between the predictor and response variable. Consequently, if measurement errors are present but unaccounted for in a statistical model, the resulting inference will be less likely to detect an association between contaminants and responses and the magnitude of the association between contaminants and responses will typically be underestimated.

Nimick et al. (2003) recommends modifying traditional field sampling methods in order to account for measurement error but this presupposes that the error mechanism is known and that the increased sampling is practical. Yuan (2007) provides a *post-hoc* method of including measurement error in data analysis. For the method we describe, the sampling design is created with measurement error in mind, and the additional fieldwork is not too burdensome.

Regression calibration is a common method for addressing measurement error and it relies on using an auxiliary covariate that is measured without error to estimate the covariate with error. Cai et al. (2000) introduced the idea of using a nonparametric smoother to do this estimation and Yao (2008) first applied the methods to this system.

3.3 Methods

Data used for this analysis were collected from the Arkansas River, a metal-polluted stream located in central Colorado. Elevated concentrations of heavy metals (Cd, Cu, and Zn) were consistently reported downstream of Leadville, CO and frequently exceeded acutely toxic levels (Clements, 2004a). From 1989 to 2004 water chemistry, habitat quality, and abundance of macroinvertebrates were measured at several monitoring stations upstream and downstream from metal sources. In 1991, after two years of monitoring, state and federal agencies began a comprehensive restoration effort to improve water quality in the river.

The goal of this long-term study was to use the relationship between metal concentration and macroinvertebrate community structure to assess effectiveness of remediation and resulting improvements in water quality. Because the stream receives a mixture of heavy metals (Cd, Cu, Zn), we used cumulative criterion units (CCU) to quantify metal contamination (Clements et al., 2000). CCU is defined as the ratio of the measured metal concentration to the hardness adjusted chronic criterion concentration, summed for each metal. In this research, the abundance of metal-sensitive mayflies (Ephemeroptera:Heptageniidae) was used as an indicator of stream health. We used a square-root transformation of mayfly abundances to stabilize the variance. In this paper we report data from station AR1 (Clements, 2004a) collected from 1989-2004. During the spring and fall of each year, 5 replicate macroinvertebrate samples were collected along with a single water sample for analysis of heavy metals. Because metal concentrations were represented by a single water sample on each sampling occasion, and because those measurements show considerable diel and seasonal variability (Clements, 2004b), significant measurement error may exist in these data.

Because of seasonal variation in metal contamination and invertebrate colonization from an upstream source, spring and fall mayfly densities show surprisingly little

correlation. To more clearly demonstrate the method used, in this paper we analyze only the fall mayfly samples. Spring CCU values tend to be higher and are more variable than the fall values and that any yearly averaging necessitates a statistical modification similar to the method being proposed.

Due to the remediation efforts above the Arkansas River, we expected a long-term decreasing trend in CCU, but the exact shape of this relationship is unknown. We modeled this relationship in two ways. First, we examined only the fall CCU values and fit a non-parametric function with no constraints on its shape. Second, created a model that incorporates both the spring and fall water quality measurements and our knowledge that a strong remediation effort occurred in the summer of 1991. The remediation is modeled by allowing a discontinuity in the smoothing function, but unfortunately this leaves only 4 data points to estimate the curve before remediation. Due to this data scarcity, we restrict our smoothing function to a flat line over this region.

We refer to the regression of mayfly counts on the raw CCU values as the *naive* estimator of the slope parameter, the regression onto the smoothed fall CCU values as the *de-noised* estimator of the slope parameter, and the regression onto the smoothed CCU values where we used both fall and spring CCU values and allow a discontinuity in 1991 as the *sophisticated de-noised* model.

3.4 Mathematical Details

Consider the simple regression model where

$$Y = \alpha + \beta x + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \tag{3.4.1}$$

Suppose that we observe the data $\{X_i, Y_i\}$where $X_i = x_i + \delta_i$ and $\delta_i \sim N(0, \tau^2)$ for i in $\{1, \ldots, n\}$. In addition, auxiliary information (such as time) is available such that $x = g(t)$. It is reasonable to use the observed data $\{X_i, t_i\}$ to estimate the true values

of the covariate $\{x_i\}$. The observed $\{Y_i\}$ can then be regressed onto these de-noised estimates $\{\hat{x}_i\}$.

Regression calibration is typically done by performing a linear regression of $\{X_i\}$ onto$\{t_i\}$ in order to estimate $\{x_i\}$, but more sophisticated modeling methods could also be used. Nonparametric smoothers are often more convenient because the researcher does not have worry about imposing a functional form on the relationship, only restrictions on the continuity or smoothness. Because the relationship between $\{X_i\}$ and $\{t_i\}$ is typically not of interest, the difficulty of interpretation of the nonparametric smoother is not an issue. The smoother could also incorporate system knowledge to impose physical constrains on the prediction (e.g. the smoothed values must be positive, the relationship form is known over an interval but is unknown over the rest).

The two most common approaches to finding the de-noised or 'smoothed' version of $\{X_i\}$ using nonparametric function estimation are kernel estimation (Fan and Gijbels, 1996) and regression splines (Green and Silverman, 1993; Ruppert et al., 2003). Although both approaches are appropriate, preliminary analyses showed similar results and due to computational advantages, we restrict our discussion to regression splines.

Cai et al. (2000) introduced the methodology of regressing the response onto the smoothed predictor using a wavelet smoother and Cui et al. (2002) extended these ideas to the kernel regression smoother. Both papers demonstrated the asymptotic normality and consistency of the slope parameter of the de-noised variable.

A researcher cannot simply use the de-noised version of an explanatory variable in subsequent analysis and inference without adjustment. While it is possible to derive the asymptotic distribution of $\hat{\beta}$ in certain instances, in general, bootstrap methods are easily used to calculate desired confidence intervals (CI). Because observations were collected at specific time points, the simple method of re-sampling the vectors $\{X_i, Y_i, t_i\}$ does not work. Instead, a bootstrap sample is created by independently

Parameters			Ignoring Measurement Error			De-noising Procedure		
σ	δ	N	Bias	Length	Coverage	Bias	Length	Coverage
0.2	0.02	30	−0.006	0.480	0.952	−0.002	0.441	0.934
		60	−0.002	0.340	0.950	0.003	0.326	0.939
		120	−0.005	0.240	0.951	0.000	0.235	0.946
	0.2	30	−0.281	0.538	0.442	0.004	0.671	0.943
		60	−0.293	0.372	0.125	0.003	0.476	0.940
		120	−0.301	0.262	0.004	0.006	0.340	0.946
	0.4	30	−0.611	0.488	0.002	0.030	1.631	0.940
		60	−0.628	0.329	0.000	0.010	0.885	0.944
		120	−0.635	0.228	0.000	0.003	0.568	0.939

Table 3.1: Simulation results comparing the naive estimator that ignores measurement error with the proposed de-noising procedure for a variety of parameter combinations. The bias column is difference between the mean estimate and the true parameter value. Length is the average length of the resulting 95% CI. Coverage is the proportion of simulations whose 95% CI contained the true parameter value.

re-sampling estimates of the measurement errors $d_i = X_i - \hat{x}_i$ and process errors $e_i = Y_i - \hat{y}_i$ and then adding those errors to the estimated values. To be explicit, for the ith observation of the bootstrap sample, two indices are randomly selected (say j, k) and the bootstrap observation is $\left\{\hat{X}_i + d_j, \hat{Y}_i + e_k, t_i\right\}$.

3.5 Simulation study

To demonstrate the effect of ignoring measurement error, we examine three examples of the measurement error model 3.4.1 where $\alpha = 0$, $\beta = 1$, and $g(t) = (1-t)^2$ for $t \in [0, 1]$. In the first case δ is small (attenuation factor $\lambda \approx 0.81$) reflecting an instance where measurement error should not have a substantial effect. The second case shows $\delta = \sigma$ ($\lambda \approx 0.29$) and the third case has $\delta = 2\sigma$ ($\lambda \approx 0.17$). We ran 2000 simulations for each case. For the de-noising procedure, each simulation inference was based on 500 bootstrap samples. The output of this simulation is shown in Table 1.

When measurement error was the same or greater than the response error, the de-noising procedure was clearly superior to the naive estimator. The bias of the naive estimator was quite large and the confidence interval coverage rates (percent of CIs that contain the true parameter value) were far from the desired 95% rate.

	$\hat{\beta}$	95% CI
Naive Estimator	-2.14	$(-3.21, -1.07)$
De-noised Estimator	-2.88	$(-4.06, -1.95)$
Sophisticated De-noised	-2.45	$(-3.74, -1.55)$

Table 3.2: Slope parameter estimates and corresponding 95% confidence intervals for the naive versus the de-noised estimators for the Arkansas River field data.

The de-noising procedure handled the measurement error reasonably well in that the observed bias is quite small. The observed coverage rates were close to the desired 95% level.

In the case where the measurement error standard deviation δ was small, the de-noising procedure did not provide any benefit over standard linear model procedure; however, the procedure did not perform substantially worse. Neither procedure had appreciable bias, the average lengths of the 95% CIs were roughly equivalent, and coverage rates were close to the desired 95%.

3.6 Results and Discussion

Metal concentrations (as CCU) decreased over time as a result of remediation activities in the Arkansas River (Figs. 1 and 2). Standard errors and confidence intervals for the naive estimator were based on assumed asymptotic normality of the error terms. Confidence intervals for the both de-noised estimators were based on $n = 10000$ bootstrap samples. The confidence interval lengths for the naive and de-noised estimators are similar, indicating the small loss of power associated with using the more complicated estimator (Table 2). The most important difference is that the naive estimator has a much smaller magnitude than either of the de-noised estimators. The boundaries of the CI for the de-noised estimator have a substantially larger magnitude than those of the naive estimator. These results indicate that by ignoring measurement error, scientists risk underestimating the relationship between the abundance of metal-sensitive mayflies and heavy metal pollution.

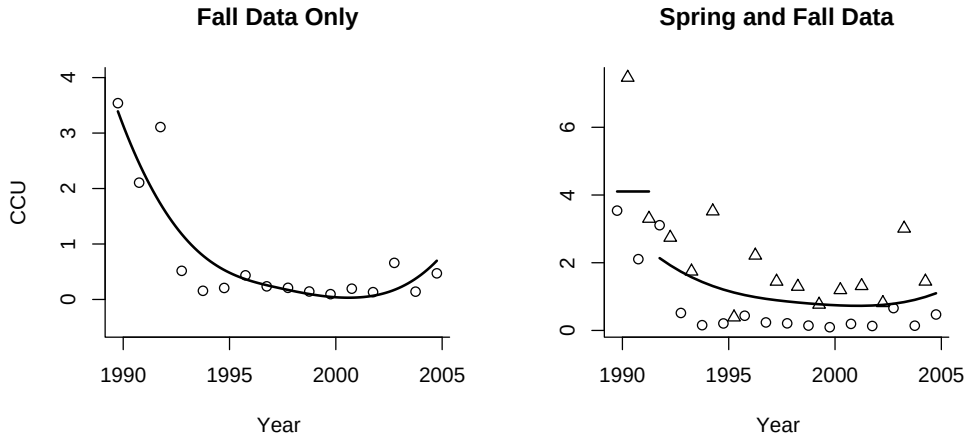

Figure 3.6.1: Left: CCU versus time along with the smoothing function. Right: CCU versus time for both spring (triangles) and fall (circles) data. The smoothing function was forced to be flat until the remediation and then a smoother was fit to the remaining data.

3.7 Conclusions

The relationship between chemical concentrations and biological responses is an integral component of ecological risk assessment. Thus, any factor that consistently affects the nature of this relationship has the potential to fundamentally alter our understanding of how chemicals impact ecosystems. Nimick et al. (2003) suggested that because of temporal variation in contaminant concentrations, it might be necessary to modify traditional field sampling protocols in aquatic ecosystems. We agree with this recommendation, but feel the potential effects of unmeasured temporal variation may be considerably more insidious. Our results suggest that temporal variation in contaminant concentrations introduces significant bias into the concentration-response relationship. This bias typically results in an underestimation of the strength of the relationship between contaminants and biological responses. Field data from the Arkansas River showed that the slope estimates ($\hat{\beta}$) of the relationship between abundance of mayflies and metal concentration increased in magnitude by approximately 14-34% when we accounted for measurement error. While this systematic bias is relatively small if the measurement error is small compared to the overall variability, it

increases as measurement errors increase. Our simulation results indicated that the naive estimator consistently provided more biased estimates of slope parameters and misleading CIs as the amount of measurement error increased. While the lengths of the CIs for the de-noised estimator were longer in high measurement error cases, the estimator was effectively unbiased and provided CIs that contained the true parameter value at the desired 95% rate.

Although this research has only illustrated the 1-dimensional case, this procedure can easily be extended to the multivariate case in several ways. First, equation (1) could include other covariates that do not have measurement error. Second, the smoothed variable could be a function of 2 or more auxiliary variables. Third, the auxiliary variable could be used to smooth several covariates.

The de-noising procedure's success is based on the smoother being a consistent estimator of $\{x_i\}$. For both the spline and local polynomial regression smoothers, all that is necessary is for the measurement errors to have mean 0. If that is not the case, an appropriate adjustment to the modeling of $g(t)$ must be made.

The procedure suggested here is applicable in a large number of situations and is relatively easy to implement. There appears to be little cost in inferential power when measurement error is small, and reduces bias in parameter estimates when measurement error is moderate to large. As such, there is little reason not to use such a procedure in situations where appropriate covariates are available.

Chapter 4

FIDUCIAL INFERENCE FOR FREE-KNOT SPLINES

4.1 Background

We are interested in the nonparametric regression model

$$y_i = g(x_i) + \sigma\epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} N(0,1)$ and $\sigma > 0$ is unknown. There are several ways to estimate $g(\cdot)$, including local polynomial fitting (Fan and Gijbels, 1996), wavelets (Picard and Tribouley, 2000) and spline models (Wahba, 1975). We will assume that $g(\cdot)$ is in the class of spline functions of order m (degree $p = m - 1$). This assumption is not burdensome because every continuous function on the closed interval $[a, b]$ can be approximated arbitrarily well by a spline of order m, provided enough knot points are allowed (Schumaker, 2007).

We consider a degree p polynomial spline with κ knotpoints t_k and $m + \kappa$ coefficients. Let $\boldsymbol{t}$ be the vector of knot points and $\boldsymbol{\alpha}$ be the vector of coefficients and $\boldsymbol{\theta} = \left\{\boldsymbol{\alpha}^T, \boldsymbol{t}^T\right\}^T$. The spline can be written using many different basis functions, but here we consider the piecewise polynomial definition:

$$g(x_i|\boldsymbol{\theta}) = \sum_{j=0}^{p} \alpha_j x_i^j + \sum_{k=1}^{\kappa} \alpha_{p+k} \left(x_i - t_k\right)_+^p$$

where

$$(u)_+ = \begin{cases} 0 & \text{if } u < 0 \\ u & \text{otherwise} \end{cases}$$

is the truncation operator which has higher precedence than exponentiation, ie $(-1)_+^2 = \left[(-1)_+\right]^2 = 0$.

A spline can be thought of as function that is a polynomial of degree p in between adjacent knot points but enforces a smoothness constraint between adjacent polynomial sections.

For model identifiability, we assume that the knot points are distinct and not on the boundaries. Furthermore we assume that the coefficients associated with a knot point are not zero, ie the knot point is a real change.

4.1.1 Spline History

Spline models are an extremely popular method for estimating $g(\cdot)$. The mathematical literature on splines exploded after the 1960's (Schumaker, 2007). Of particular interest was the recurrent relationships that define the numerically stable B-spline basis (de Boor, 1972). Wahba (1975) introduces smoothing splines by including a large number of knot locations and finding the spline function $g(x|\hat{\boldsymbol{\theta}})$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left(g(x_i|\hat{\boldsymbol{\theta}}) - y_i\right)^2 + \lambda\int\left(g^{(m)}(x|\hat{\boldsymbol{\theta}})\right)^2 dx$$

This is the typical sum of squares estimator with a penalty term for extreme changes in the m^{th} derivative. Wahba recommends selecting the value of λ using the cross-validation method. There is a connection between the smoothing parameter λ and the model degrees of freedom and many smoothing spline software packages allow a user to specify the model degrees of freedom or use generalized cross-validation to select it.

This solution is quite practical, but lacks interpretability. Because the knot placement is selected by the user and the number of knots is often large, how the individual regression coefficients affect the resulting function is not necessarily obvious. Further, the knot location can be the primary research interest. For example, a knot point might be regarded as a change point.

Spline models where the knot locations are estimated in the same manner as the regression coefficients are known as free-knot splines. DiMatteo et al. (2001) introduced a Bayesian solution to the problem along with a convenient software package

(Wallstrom et al., 2007). Their solution uses a uniform prior on the knot point locations, a unit-information Normal prior for the regression coefficients and the improper prior σ^{-1} for the standard deviation. Unfortunately the software only allows for degree 3 polynomials and the users ability to control how many knot points are allowed is limited.

The maximum likelihood solution to the problem was given in Mao and Zhao (2003). However the process of exploring the likelihood surface is non-trivial because of the large number of local maximums (Jupp, 1978). (Mao and Zhao, 2003) address this by examining a large number of starting points for the knot locations and using a proprietary function from the International Mathematical and Statistical Libraries. They use the asymptotic normality of the MLE to calculate confidence intervals for the function $f(x)$. They did not address the quality of a confidence interval for a knot point.

4.1.2 Fiducial History

Fisher (1930) first introduced his idea of fiducial inference in order to address what he felt was the major shortcoming of Bayesian inference. His goal was to invent a posterior-like distribution without the need for a prior distribution. He did not succeed in developing a general theory for finding these fiducial distributions and his idea was met with extreme skepticism. In the 1990's generalized confidence intervals (Weerahandi, 1993) were found to have very good small sample properties and Hannig et al. (2006b) showed the connection between generalized confidence intervals and Fisher's fiducial inference. Hannig (2009) went on to develop a general theory for developing fiducial solutions which has been used in a variety of contexts (Hannig et al., 2006a; E et al., 2008).

The general framework of fiducial inference assumes that the n observed data can be written in a structural equation $\mathbb{X} = \mathbf{G}(\xi, \mathbb{U})$, where ξ is a p length vector of parameters, and $\mathbb{U}$ is a random vector of with a completely known distribution.

Setting $\mathbb{X}_0 = (X_1, \ldots, X_p)$, $\mathbb{X}_c = (X_{p+1}, \ldots, X_n)$, $\mathbb{U}_0 = (U_1, \ldots, U_p)$ and $\mathbb{U}_c = (U_{p+1}, \ldots, U_n)$ the structural equation can be factorized as

$$\mathbb{X}_0 = \mathbf{G_0}(\xi, \mathbb{U}_0) \quad \text{and} \quad \mathbb{X}_c = \mathbf{G_c}(\xi, \mathbb{U}_c).$$

Assuming that for each $\xi \in \Xi$ that $\mathbf{G}_0(\xi, \cdot)$ and $\mathbf{G}_c(\xi, \cdot)$ are one-to-one and differentiable and that $\mathbf{G}_0(\xi, \cdot)$ also invertible, then Hannig (2009) showed that the generalized fiducial distribution is

$$r(\xi|\mathbf{x}) = \frac{f_{\mathbf{x}}(x|\xi) J_0(x_0, \xi)}{\int_\Xi f_{\mathbf{x}}(x|\xi') J_0(x_0, \xi') d\xi'}$$

where

$$J_0(x_0, \xi) = \left| \frac{det\left(\frac{d}{d\xi}\mathbf{G}_0^{-1}(x_0, \xi)\right)}{det\left(\frac{d}{dx_0}\mathbf{G}_0^{-1}(x_0, \xi)\right)} \right|$$

and $f_{\mathbf{x}}(x|\xi)$ is the density function. Since the selection to use the first p data coordinates to use in $\mathbf{G}_0$, we could select any p coordinates that satisfy the one-to-one, differentiable, and invertible conditions. Hannig (2009) suggests letting J be the average of all possible values of J_0 and using

$$r(\xi|\mathbf{x}) = \frac{f_{\mathbf{x}}(x|\xi) J(x_0, \xi)}{\int_\Xi f_{\mathbf{x}}(x|\xi') J(x_0, \xi') d\xi'}$$

This distribution is similar to a Bayesian posterior distribution with the Jacobian taking the role of the prior. It is not surprising the the integral in the denominator is often impossible to calculate and we often must turn to the same numerical methods that Bayesians do.

4.2 Asymptotic properties

Estimators often have an asymptotic normal distribution and fiducial estimators do as well. Conditions A0-A6 in the appendix are sufficient to prove that the maximum likelihood estimators to have an asymptotic normal distribution (Lehmann and Casella, 1998).

Theorem 1. *Under assumption A0-A6, the maximum likelihood estimators* $\hat{\boldsymbol{\xi}}_n$ *are consistent and* $\sqrt{n}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi})$ *is asymptotically normal with mean* $\mathbf{0}$ *and covariance matrix* $[I(\boldsymbol{\xi})]^{-1}$ *where* $I(\boldsymbol{\xi})$ *is the Fisher information matrix.*

Conditions found in Ghosh and Ramamoorthi (2003) give sufficient conditions for the Bayesian posterior distribution to be asymptotically normal. Hannig (2009) gives sufficient conditions for the asymptotic normality of the fiducial distribution in the univariate parameter case. The extension to the multiparameter case is straightforward. Let $\mathcal{R}_{\boldsymbol{\xi}}$ be an observation from the fiducial distribution $r(\boldsymbol{\xi}|\boldsymbol{x})$ and denote the density of $s = \sqrt{n}\left(\mathcal{R}_{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}}_n\right)$ by $\pi^*(\boldsymbol{\xi}, \boldsymbol{x})$.

Theorem 2. *Under assumptions A0-A6, B1-B2, and C1-C2*

$$\int_{\mathbb{R}^p} \left| \pi^*(\boldsymbol{s}, \boldsymbol{x}) - \frac{\sqrt{det\,|I(\boldsymbol{\xi}_0)|}}{\sqrt{2\pi}} e^{-\boldsymbol{s}^T I(\boldsymbol{\xi}_0)\mathbf{s}/2} \right| d\boldsymbol{s} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0 \tag{4.2.1}$$

Proof. Assume without loss of generality that $\Xi = \mathbb{R}^p$. We denote $J_n(\boldsymbol{x}_n, \boldsymbol{\xi})$ as the average of all possible Jacobians over a sample of size n and $\pi(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}_0} J_0(\boldsymbol{x}, \boldsymbol{\xi})$. Assumption C2 and the uniform strong law of large numbers for U-statistics imply that $J_n(\boldsymbol{x}, \boldsymbol{\xi}) \overset{a.s.}{\to} \pi(\boldsymbol{\xi})$ uniformly in $\boldsymbol{\xi} \in \bar{B}(\boldsymbol{\xi}_0, \delta)$ and that $\pi(\boldsymbol{\xi})$ is continuous. Therefore

$$\sup_{\boldsymbol{\xi} \in \bar{B}(\boldsymbol{\xi}_0, \delta)} |J_n(x_n, \boldsymbol{\xi}) - \pi(\boldsymbol{\xi})| \to 0 \;\; P_{\boldsymbol{\xi}_0}\, a.s.$$

We now follow the proof of the univariate case. Let

$$\begin{aligned}
\pi^*(\boldsymbol{s}, \boldsymbol{x}) &= \frac{J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) f\left(\boldsymbol{x}_n | \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right)}{\int_{\mathbb{R}^p} J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) f\left(\boldsymbol{x}_n | \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) d\boldsymbol{t}} \\
&= \frac{J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right)\right]}{\int_{\mathbb{R}^p} J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right)\right] d\boldsymbol{t}} \\
&= \frac{J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right]}{\int_{\mathbb{R}^p} J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t}}
\end{aligned}$$

and just as Ghosh and Ramamoorthi (2003), we let $H = -\frac{1}{n}\frac{\partial}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}} L_n\left(\hat{\boldsymbol{\xi}}_n\right)$ and we notice that $H \to I\left(\boldsymbol{\xi}_0\right)$ *a.s.* $P_{\boldsymbol{\xi}_0}$. It will be sufficient to prove

$$\int_{\mathbb{R}^p} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] - \pi\left(\boldsymbol{\xi}_0\right) \exp\left[\frac{-\boldsymbol{t}^T I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}}{2}\right] \right| d\boldsymbol{t} \overset{P_{\boldsymbol{\xi}_0}}{\to} 0 \quad (4.2.2)$$

Let t_i represent the ith component of vector $\boldsymbol{t}$. By Taylor's Theorem, we can compute

$$\begin{aligned} L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) &= L_n\left(\hat{\boldsymbol{\xi}}_n\right) + \sum_{i=1}^{p}\left(\frac{t_i}{\sqrt{n}}\right)\frac{\partial}{\partial \xi_i} L_n\left(\hat{\boldsymbol{\xi}}_n\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\frac{t_i t_j}{\left(\sqrt{n}\right)^2}\frac{\partial}{\partial \xi_i \partial \xi_j} L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right) \\ &\quad + \frac{1}{6}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\frac{t_i t_j t_k}{\left(\sqrt{n}\right)^3}\frac{\partial}{\partial \xi_i \partial \xi_j \partial \xi_k} L_n\left(\boldsymbol{\xi}'\right)\right) \\ &= L_n\left(\hat{\boldsymbol{\xi}}_n\right) - \frac{\boldsymbol{t}^T H \boldsymbol{t}}{2} + R_n \end{aligned}$$

for some $\boldsymbol{\xi}' \in \left[\hat{\boldsymbol{\xi}}_n, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right]$. Notice that $R_n = O_p\left(\|\boldsymbol{t}\|/n^{3/2}\right)$.

Given any $0 < \delta < \delta_0$ and $c > 0$, we break $\mathbb{R}^p$ into three regions:

$$\begin{aligned} A_1 &= \left\{\boldsymbol{t} : \ \|\boldsymbol{t}\| < c \log\sqrt{n}\right\} \\ A_2 &= \left\{\boldsymbol{t} : \ c\log\sqrt{n} < \|\boldsymbol{t}\| < \delta\sqrt{n}\right\} \\ A_3 &= \left\{\boldsymbol{t} : \ \delta\sqrt{n} < \|\boldsymbol{t}\|\right\} \end{aligned}$$

On $A_1 \cup A_2$ we compute

$$\int_{A_1 \cup A_2} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right.$$
$$\left. -\pi\left(\boldsymbol{\xi}_0\right) \exp\left[-\frac{1}{2}\boldsymbol{t}' I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}\right] \right| d\boldsymbol{t}$$

$$\leq \int_{A_1 \cup A_2} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \right|$$
$$\cdot \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t}$$
$$+ \int_{A_1 \cup A_2} \left| \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right.$$
$$\left. -\pi\left(\boldsymbol{\xi}_0\right) \exp\left[-\frac{1}{2}\boldsymbol{t}' I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}\right] \right| d\boldsymbol{t}$$

Since $\pi(\cdot)$ is a proper prior on $A_1 \cup A_2$, then the second term goes to 0 by the Bayesian Bernstein-von Mises theorem. Next we notice that

$$\int_{A_1 \cup A_2} \left| J_n\left(x, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \right|$$
$$\cdot \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t}$$
$$\leq \sup_{\boldsymbol{t} \in A_1 \cup A_2} \left| J_n\left(x, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \right|$$
$$\cdot \int_{A_1 \cup A_2} \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t}$$

Since $\sqrt{n}\left(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0\right) \xrightarrow{\mathcal{D}} N\left(0, I\left(\boldsymbol{\xi}_0\right)^{-1}\right)$, then

$$P_{\boldsymbol{\xi}_0}\left[\left\{\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n};\ \boldsymbol{t} \in A_1 \cup A_2\right\} \subset B\left(\boldsymbol{\xi}_0, \delta_0\right)\right] \to 1.$$

Furthermore, since $L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right) = -\frac{\boldsymbol{t}^T H \boldsymbol{t}}{2} + R_n$ then the integral converges in probability to 1. Since $\max_{\mathbf{t} \in A_1 \cup A_2} \|\boldsymbol{t}/\sqrt{n}\| \leq \delta$ and $J_n \to \pi$, then the term $\to 0$ in probability.

Next we turn to

$$\begin{aligned}
&\int_{A_3} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right. \\
&\qquad \left. -\pi\left(\boldsymbol{\xi}_0\right) \exp\left[\frac{-\boldsymbol{t}^T I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}}{2}\right] \right| d\boldsymbol{t} \\
\leq\ & \int_{A_3} J_n\left(\boldsymbol{x}_i, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\
&+ \int_{A_3} \pi\left(\boldsymbol{\xi}_0\right) \exp\left[\frac{-\boldsymbol{t}^T I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}}{2}\right] d\boldsymbol{t}
\end{aligned}$$

The second integral goes to 0 in $P_{\boldsymbol{\xi}_0}$ probability because $\min_{A_3} \|\boldsymbol{t}\| \to \infty$. As for the first integral,

$$\begin{aligned}
&\int_{A_3} J_n\left(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\
=\ & \frac{1}{n}\sum_{i=1}^{n} \int_{A_3} J\left(\boldsymbol{x}_i, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\
=\ & \frac{1}{n}\sum_{i=1}^{n} \int_{A_3} J\left(\boldsymbol{x}_i, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) f\left(x_i | \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \\
&\qquad \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right) - \log f\left(x_i | \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right)\right] d\boldsymbol{t}
\end{aligned}$$

Because $J(\cdot)$ is a probability measure, then so is $J(\cdot) f(\cdot)$. Assumption C1 assures that the exponent goes to $-\infty$ and therefore the integral converges to 0 in probability.

Having shown 4.2.2, we now follow Ghosh and Ramamoorthi (2003) and let

$$C_n = \int_{\mathbb{R}^p} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right| d\boldsymbol{t}$$

then the main result to be proved 4.2.1 becomes

$$\begin{aligned}
C_n^{-1} \left\{ \int_{\mathbb{R}^p} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right. \right. \cdot \\
\left. \left. -C_n \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}} e^{-\boldsymbol{s}^T I(\boldsymbol{\xi}_0)\mathbf{s}/2} \right| \right\} d\boldsymbol{s} \ \overset{P_{\boldsymbol{\xi}_0}}{\to}\ 0 \quad (4.2.3)
\end{aligned}$$

Because

$$\begin{aligned}\int_{\mathbb{R}^p} J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \exp\left[-\frac{\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] d\boldsymbol{s} &= J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \int_{\mathbb{R}^p} \exp\left[-\frac{\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] d\boldsymbol{s} \\ &= J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \frac{\sqrt{2\pi}}{\sqrt{\det(H)}} \\ &\overset{a.s.}{\to} \pi\left(\boldsymbol{\xi}_0\right) \sqrt{\frac{2\pi}{\det\left(I\left(\boldsymbol{\xi}_0\right)\right)}}\end{aligned}$$

and 4.2.2 imply that $C_n \overset{P}{\to} \pi\left(\boldsymbol{\xi}_0\right) \sqrt{\frac{2\pi}{\det(I(\boldsymbol{\xi}_0))}}$ it is enough to show that the integral in 4.2.3 goes to 0 in probability. This integral is less than $I_1 + I_2$ where

$$\begin{aligned}I_1 = \int_{\mathbb{R}^P} \Bigg| & J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \\ & - J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \exp\left[\frac{-\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] \Bigg| \, d\boldsymbol{s}\end{aligned}$$

and

$$I_2 = \int_{\mathbb{R}^P} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \exp\left[\frac{-\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] - C_n \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}} e^{-\boldsymbol{s}^T I(\boldsymbol{\xi}_0)\mathbf{s}/2} \right| d\boldsymbol{s}.$$

Equation 4.2.2 shows that $I_1 \to 0$ in probability and I_2 is

$$\begin{aligned}I_2 &= \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) - C_n \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}} \right| \int_{\mathbb{R}^P} \exp\left[\frac{-\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] d\boldsymbol{s} \\ &\overset{P}{\to} 0\end{aligned}$$

because $J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \overset{P}{\to} \pi\left(\boldsymbol{\xi}_0\right)$ and $C_n \overset{P}{\to} \pi\left(\boldsymbol{\xi}_0\right) \sqrt{\frac{2\pi}{\det(I(\boldsymbol{\xi}_0))}}$. □

4.3 Fiducial free-knot splines

We first define

$$g(x_i|\hat{\boldsymbol{\theta}}) = \sum_{j=0}^{p} \hat{\alpha}_j x_i^j + \sum_{k=1}^{\kappa} \hat{\alpha}_{p+k} \left(x_i - \hat{t}_k\right)_+^p$$

and let $\boldsymbol{\xi} = \left\{\boldsymbol{\theta}^T, \sigma^2\right\}^T$. To derive the form of the Jacobian we first recognize that

$$\boldsymbol{G}_0^{-1}\left(y_i, \boldsymbol{\xi}\right) = \frac{1}{\sigma}\left(y_i - g(x_i|\boldsymbol{\theta})\right)$$

and therefore

$$\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial \boldsymbol{\alpha}} = -\frac{1}{\sigma}\left(1, x_i, \ldots, x_i^p, (x_i - t_1)_+^p, \ldots, (x_i - t_\kappa)_+^p\right)$$

$$\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial \boldsymbol{t}} = \frac{p}{\sigma}\left(\alpha_{p+1}(x_i - t_1)_+^{p-1}, \ldots, \alpha_{p+\kappa}(x_i - t_\kappa)_+^{p-1}\right)$$

$$\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial \sigma^2} = -\frac{1}{2\sigma^3}(y_i - g(x_i|\boldsymbol{\theta}))$$

$$\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial y_i} = \frac{1}{\sigma}$$

Using these results, then for any selection of data points that satisfies the necessary criteria, say $\boldsymbol{y}_0 = \{y_{(1)}, \ldots, y_{(l)}\}$ where $l = p + \kappa + 2$, the Jacobian is

$$J_0(\boldsymbol{y}_0, \boldsymbol{\xi}) = \left| \frac{1}{\sigma^2} p^\kappa \left[\prod_{j=1}^{\kappa} \alpha_{p+\kappa} \right] \det \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{\alpha}} & \boldsymbol{B}_{\boldsymbol{t}} & \boldsymbol{B}_{\sigma^2} \end{bmatrix} \right|$$

where

$$\boldsymbol{B}_{\boldsymbol{\alpha}} = \begin{bmatrix} 1 & x_{(1)} & \ldots & x_{(1)}^p & (x_{(1)} - t_1)_+^p & \ldots & (x_{(1)} - t_\kappa)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{(l)} & \ldots & x_{(l)}^p & (x_{(l)} - t_1)_+^p & \ldots & (x_{(l)} - t_\kappa)_+^p \end{bmatrix},$$

$$\boldsymbol{B}_{\boldsymbol{t}} = \begin{bmatrix} (x_{(1)} - t_1)_+^{p-1} & \ldots & (x_{(1)} - t_\kappa)_+^{p-1} \\ \vdots & \ddots & \vdots \\ p(x_{(l)} - t_1) & \ldots & (x_{(l)} - t_\kappa)_+^{p-1} \end{bmatrix},$$

and

$$\boldsymbol{B}_{\sigma^2} = \begin{bmatrix} -\frac{1}{2}(y_{(1)} - g(x_{(1)}|\boldsymbol{\theta})) \\ \vdots \\ -\frac{1}{2}(y_{(l)} - g(x_{(l)}|\boldsymbol{\theta})) \end{bmatrix}.$$

However the question of what sets of indices satisfy the one-to-one, and invertability requirements is not obvious. We opted to only consider the sets of indices that include at least two observations from each inter-knot region. Since calculating the average Jacobian value of all possible sets of indices was infeasible, we take a random sample of suitable sets of indices and average the resulting Jacobian values.

In order to examine the fiducial distribution, we turn to Markov Chain Monte Carlo methods. The approach that we took was to take the current knot locations and add a random normal deviate. Then treating the knot points as known constants,

the fiducial distribution of the regression coefficients is known and we can sample from that distribution. Likewise once the knots points and regression coefficients are known, the fiducial distribution of the variance component is known and can be sampled from. Together these steps create a proposed value $\boldsymbol{\xi}^*$ from the previous $\boldsymbol{\xi}$ and the density function for selecting $\boldsymbol{\xi}^*$ given $\boldsymbol{\xi}$ is denoted $p\left(\boldsymbol{\xi}^*|\boldsymbol{\xi}\right)$. If the ratio

$$\frac{f\left(\boldsymbol{y}|\boldsymbol{\xi}^*\right)}{f\left(\boldsymbol{y}|\boldsymbol{\xi}\right)}\frac{p\left(\boldsymbol{\xi}|\boldsymbol{\xi}^*\right)}{p\left(\boldsymbol{\xi}^*|\boldsymbol{\xi}\right)}$$

is greater than a $Uniform(0,1)$ random deviate, we accept the proposed value as the next value in the Markov chain.

Theorem 3. *Let $\boldsymbol{\theta}$ be the parameters of a free-knot spline of degree 3 or greater with truncated polynomial basis functions. Define $\boldsymbol{\xi} = (\sigma^2, \boldsymbol{\theta})$. Let $\pi^*\left(\boldsymbol{\xi}, \boldsymbol{y}\right)$ be the fiducial distribution of $\mathcal{R}_{\boldsymbol{\xi}}$. Then*

$$\int_{\mathbb{R}^p}\left|\pi^*\left(\boldsymbol{s}, \boldsymbol{y}\right) - \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}}e^{-\boldsymbol{s}^T I(\boldsymbol{\xi}_0)\mathbf{s}/2}\right| d\boldsymbol{s} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0$$

Proof. It suffices to show that the free-knot spline satisfies assumptions A0-A6, B1-B2, C1-C2. These are shown in appendix A. □

4.4 Simulation Study

First we consider the single knot case with order $m = 4$ and examine the confidence intervals for the knot point. We consider the spline

$$Y_i = 0 + 1.2\left(x_i\right) - 3\left(x_i^2\right) + 1.4\left(x_i^3\right) + 4.8\left(x_i - \frac{1}{2}\right)_+^3 + \sigma\epsilon_i$$

on $x \in [0,1]$ and n is the sample size. For each parameter combination, we created 1000 data sets and for each data set calculated the the minimum confidence level that would capture the true knot point value and can therefore consider the actual coverage rate versus the nominal coverage rate for any confidence level.

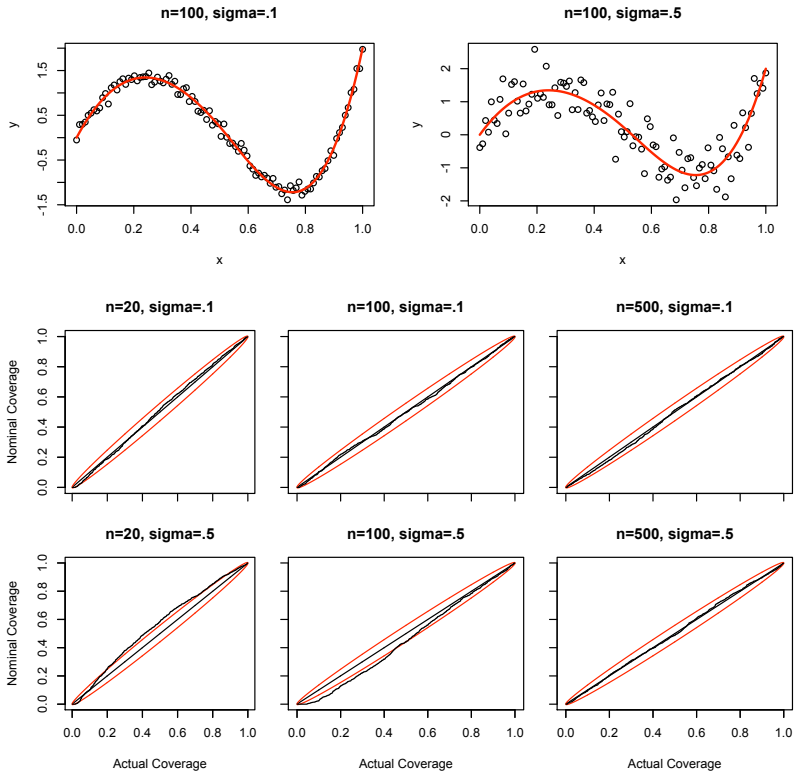

Figure 4.4.1: Truncated polynomial basis results. The top row gives two examples of the data sets fit. Three sample sizes $(20, 100, 500)$ and two error standard deviations $(\sigma = 0.1, 0.5)$ were considered $N = 1000$ simulations were run for each combination. The bottom rows are plots of the actual coverage rates versus nominal coverage rates for the knot point location. The closer to the $y = x$, the better performance of the confidence interval. The red bands give the area of natural fluctuation due to randomness.

The truncated polynomial basis appears to work well for large sample sizes, but in the case of a small sample size and large variability, the knot point confidence interval was a little more liberal than can be explained by simulation variability.

Unfortunately the truncated polynomial basis has drastic numerical deficiencies. Because changing the coefficient on the basis function x^p drastically affects the function value on all sections on which the spline is defined, this basis is unsuitable for serious computation. For example, the variance of $\mathcal{R}_{\alpha_i}$ increases with increasing values of i.

B-splines provide a more numerically stable set of basis functions but are harder to work with algebraically due to their recursive nature. The derivatives can be calculated but become increasingly more cumbersome to write in closed form. We performed the same simulation study using the b-spline basis.

The confidence interval lengths for this simulation are quite interesting. In the small variance case, the interval lengths decrease as expected, but the lengths appear to stay the same in the large variance case. Investigating individual model fits shows that the models fit the data quite well. The large interval widths reflect extreme flexibility of high order splines and the large amount of data relative to the error variance necessary to accurately estimate the knot point.

In the final simulation we only consider the b-spline basis. We consider the case of multiple knots ($\kappa = 3$), and the spline with $\boldsymbol{\alpha} = \{0, 2, 1, -3, 4, -3, 1\}$ and $\boldsymbol{t} = \{0.3, 0.5, 0.8\}$. Since there are 11 parameters in the model, it is unsurprising that the confidence intervals remain large in the high variance case, however the coverage rates tend to be acceptable. The extreme flexibility of the high order splines continues to make the knot point confidence intervals quite wide.

Finally we consider fitting a spline to a data generated by a non-spline function. For this example we consider $y_i = \Phi(x_i) + \sigma\epsilon_i$ where $\epsilon_i \sim N(0, 1)$. We fit these data with a degree 1 spline with 2 knot points. The spline that minimizes the integrated squared difference has knot points at ± 1.2.

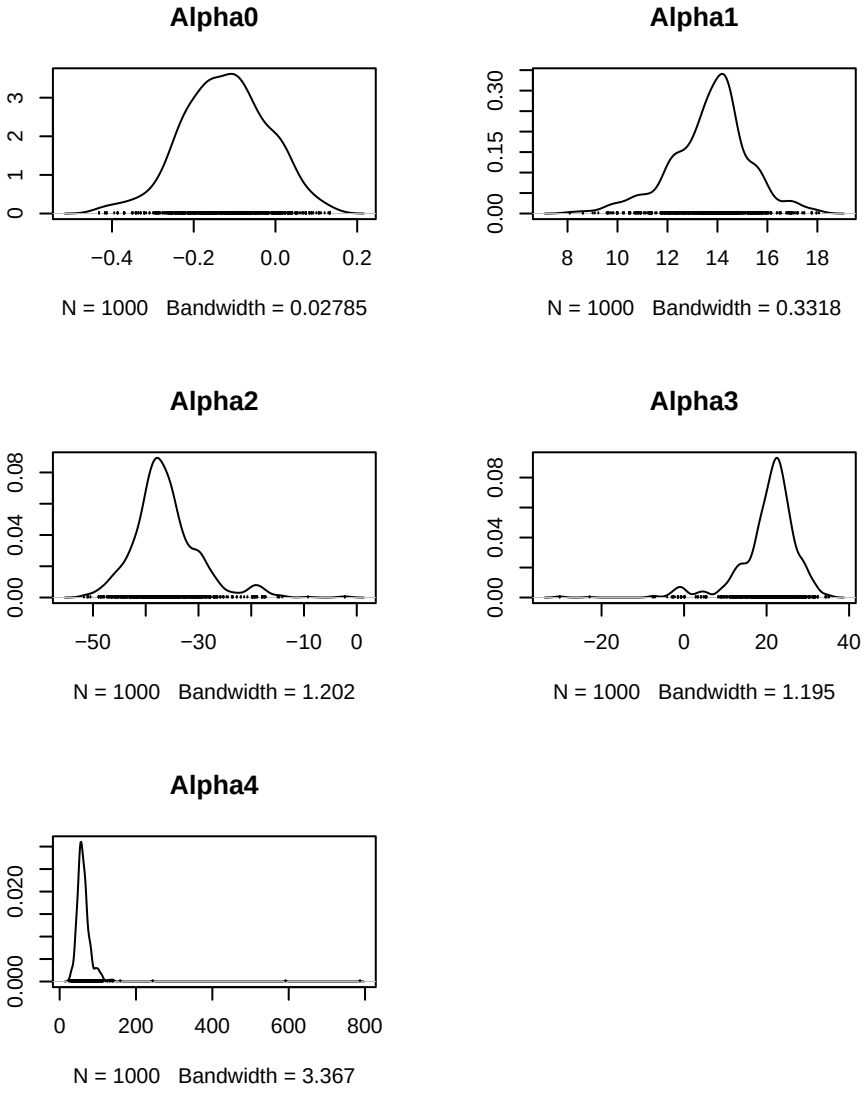

Figure 4.4.2: An example of increased variance of the distribution of $\mathcal{R}_{\alpha_i}$ as i increases for truncated polynomial basis functions.

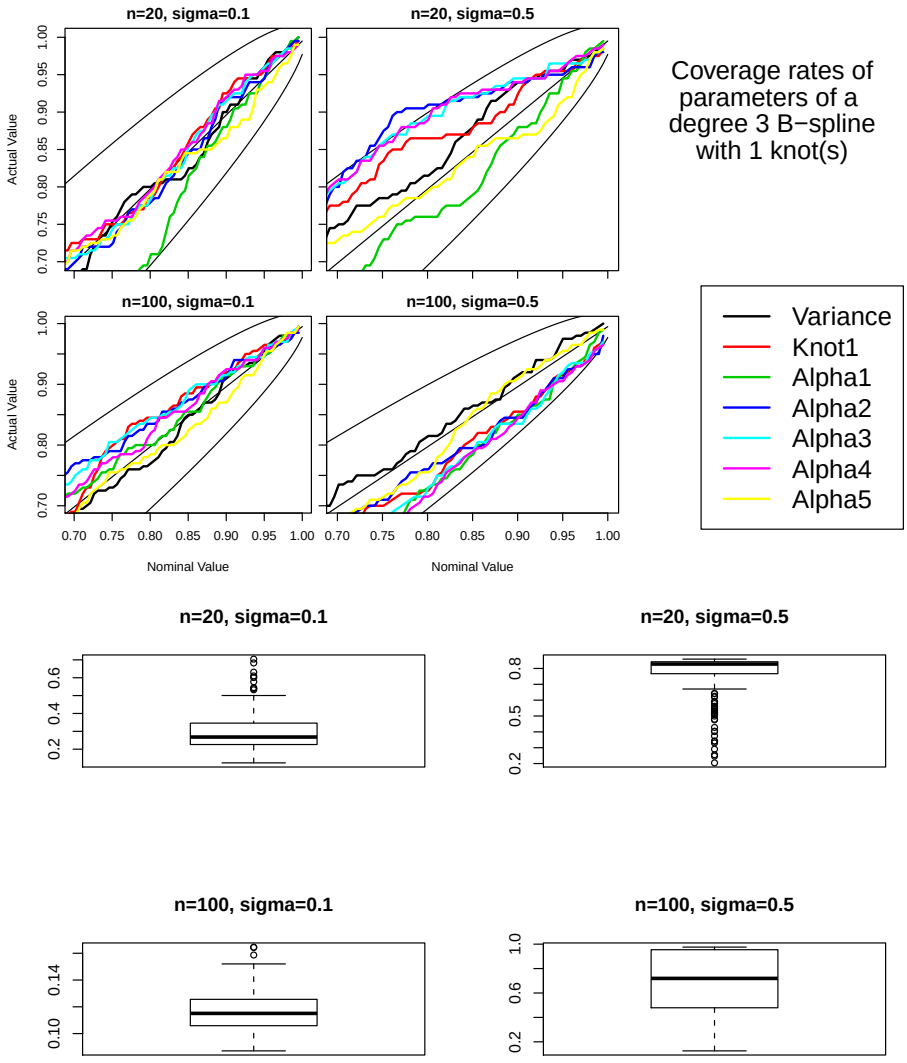

Figure 4.4.3: Two sample sizes $(20, 100)$ and two error standard deviations $(\sigma = 0.1, 0.5)$ were considered and $N = 200$ simulations were run for each combination.

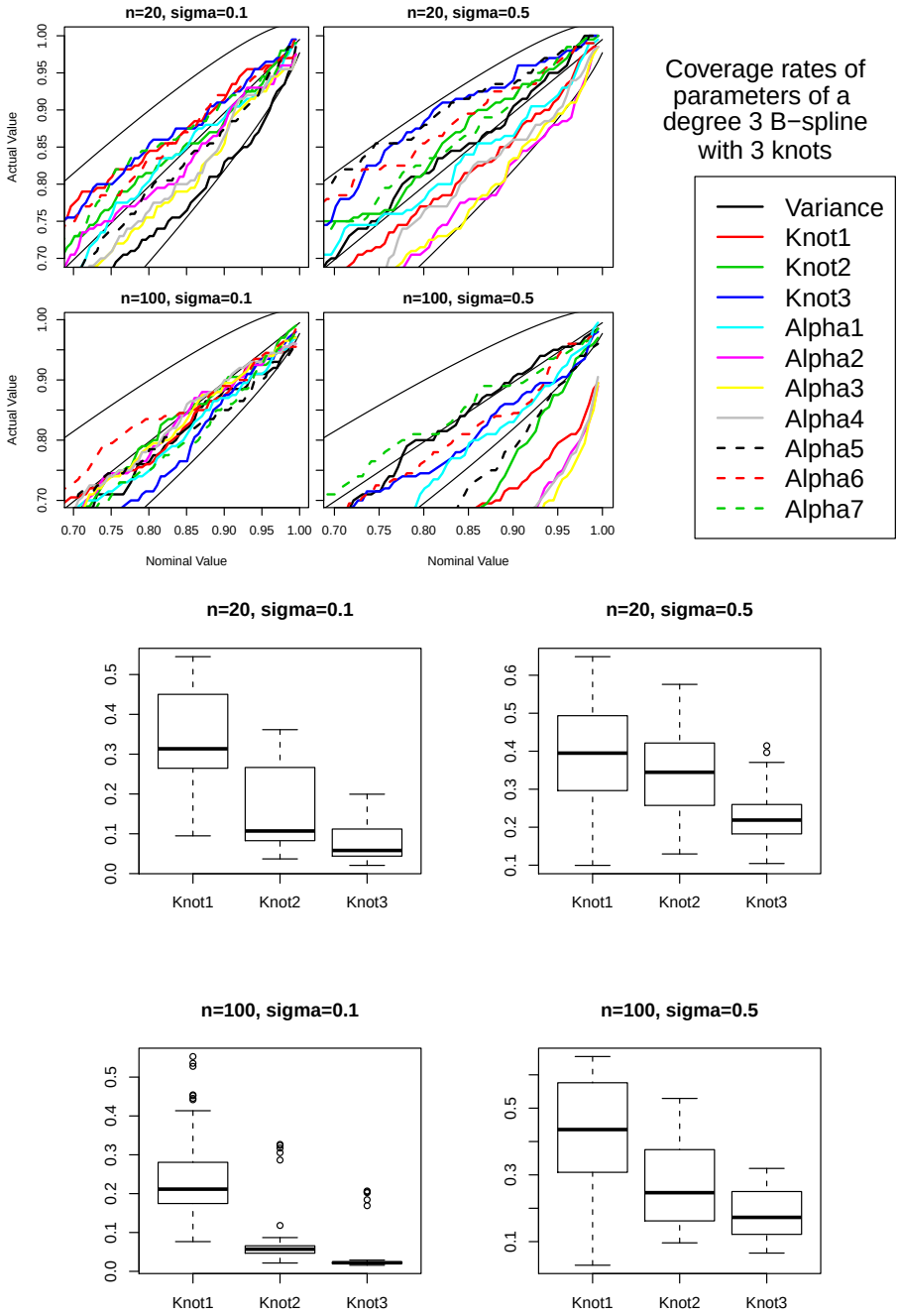

Figure 4.4.4: Two sample sizes $(20, 100)$ and two error standard deviations $(\sigma = 0.1, 0.5)$ were considered and $N = 200$ simulations were run for each combination.

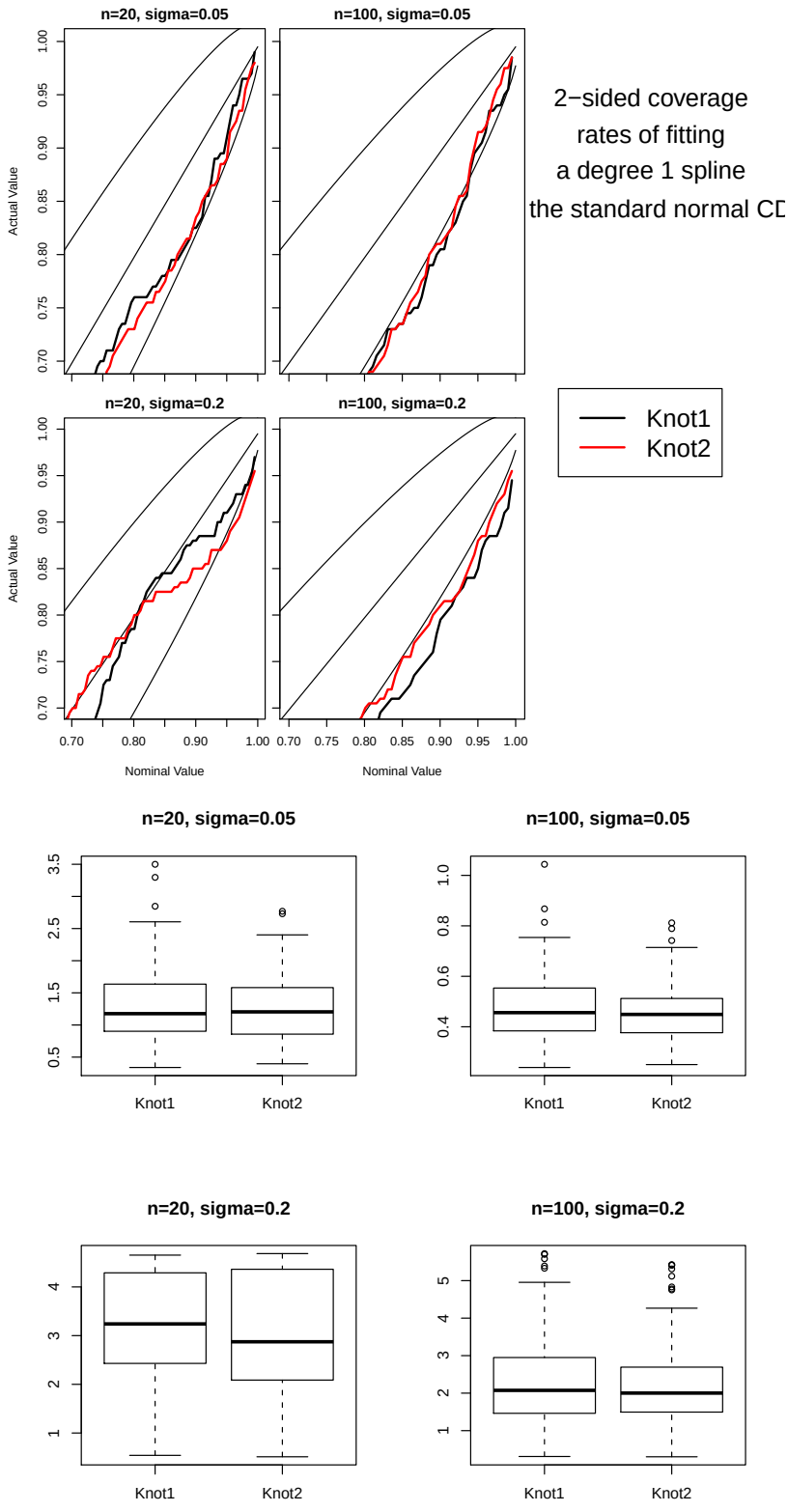

Figure 4.4.5: Two sample sizes $(20, 100)$ and two error standard deviations $(\sigma = 0.1, 0.5)$ were considered and $N = 200$ simulations were run for each combination.

At both variance levels the confidence intervals are converging towards zero indicating that the fiducial spline is converging to the optimal spline. The bias of the coverage rate is not surprising given that the data were not generated from the model we are fitting.

4.5 Selecting the number of Knots

The number of knot points is often not known and must be estimated. Mao and Zhao (2003) addressed this question using a model selection procedure. They used a modified criteria to pick the model with the number of knots that minimized

$$GCV(\kappa) = \frac{\sum \left\{ y_i - \hat{f}(x_i) \right\}^2}{\{n - (2\kappa + 4)\}^2 / n}$$

and this criteria worked slightly better than using AIC value of Akaike (1973) and was comparable to using the small sample AICc value of Hurvich and Tsai (1989). We also used AIC, AICc, MDL, BIC, GCV and DIC to select the optimal number of knot points and also note the heuristic behavior of the MCMC algorithm in cases of too many or too few knots. It should be noted that since the various information criterion are derived for the maximum likelihood solutions and as such, do not have a strong theoretical justification.

There are several behaviors of the MCMC algorithm that are indicative of having selected the wrong number of knots. If the selected number of knots is too small, convergence of the algorithm is often very slow. For example if the true number of knot points is 2 and only 1 is fit, then the trace of the Markov chain shows the knot point jumping from one knot to the other. If the selected number of knots is too large, convergence is also slowed but the additional knots tend to cluster towards ends of range of x-values.

In investigating AIC, it is not clear whether to calculate the AIC value at each chain step and then use the mean of these AIC values (we denote this as Mean AIC)

	$n = 20$					$n = 100$					$n = 500$				
	κ selected					κ selected					κ selected				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Mean AIC	0	48	44	0	8	0	18	82	0	0	0	0	98	2	0
AIC of Mean	0	47	46	0	7	0	14	85	1	0	0	0	98	1	1
Mean AICc	0	100	0	0	0	0	0	0	100	0	0	100	0	0	0
AICc of Mean	0	98	2	0	0	0	19	80	1	0	0	0	100	0	0
Mean GCV	0	0	100	0	0	0	0	0	100	0	0	100	0	0	0
GCV of Mean	0	64	35	0	1	0	16	83	1	0	0	0	98	1	1
Mean MDL	0	0	100	0	0	0	100	0	0	0	0	100	0	0	0
MDL of Mean	0	55	42	0	3	0	40	60	0	0	0	0	100	0	0
Mean BIC	0	100	0	0	0	0	100	0	0	0	0	100	0	0	0
BIC of Mean	0	58	39	0	3	0	42	58	0	0	0	0	100	0	0
DIC	0	0	100	0	0	0	0	0	100	0	0	0	0	0	100

Table 4.1: For each sample size $N = 100$ data sets were created and analyzed assuming different numbers of knot points.

or to calculate the mean value of fiducial distribution for each parameter and use the parameter's point estimate to calculate an AIC value (we denote this as AIC of Mean). We repeat this naming convention for AICc, BIC, GCV, and MDL as well.

We compare these 11 approaches in a simulation study (table 1). The true model for the data was the same as in the multiple knot study (*i.e.* $\kappa = 3$) and the error standard deviation was $\sigma = 0.2$. For each of 3 sample sizes ($n \in \{20, 100, 500\}$) $N = 100$ data sets were generated and then analyzed 5 separate times, each assuming a different number of knots ($\kappa \in \{1, \ldots, 5\}$).

The first thing to notice is that calculating a criterion value at each step of the chain and then using the average value tends to select a model with too few knots this bias gets worse as the sample size increases. Of the methods that calculate the criterion value using the parameter point estimates, the AIC method selected the correct number of knots the most often at the $n = 20, 100$ levels. DIC selected the correct number of knots at a small sample size, but overestimates the number of knots at larger sample sizes.

4.6 Discussion

The fiducial solution to the free-knot spline problem is asymptotically correct and performs well in simulation studies. The number of knots can be selected via a model selection procedure such as AIC, AICc, BIC or MDL.

Unfortunately the MCMC method of evaluating the fiducial density is computer intensive. For every proposed step of the chain, the Jacobian must be estimated, which requires a large number of determinants to be calculated. In the Jacobian calculation, most of the matrix columns have a linear dependence on $\boldsymbol{\alpha}$ and thus the can be factored out, but the column $\boldsymbol{B}_{\sigma^2}$ depend on both $\boldsymbol{\alpha}$ and $\boldsymbol{t}$ in a non-linear fashion.

If the variance parameter was a known quantity, the dimension of the parameter space that the Jacobian depends on would be greatly reduced. In this case, there is considerable computation savings to be had by creating a grid of points in $[x_{min}, x_{max}]^{\kappa}$ and evaluate the Jacobian at each point. Because the Jacobian function is continuous in $\boldsymbol{t}$, a linear interpolation of near-by grid points will suffice to estimate the Jacobian value at any given set of knots. Further computational savings can be had by only calculating the Jacobian at a particular grid point once the value is requested. In this manner long chains could be calculated in an efficient manner.

Using a consistent estimator of σ^2 in the likelihood and Jacobian calculations would lead to computational savings but at a cost of generally under-estimating the variability of the other parameters. A simulation study of this method should be undertaken to examine how strong this effect is.

The question of model selection is not yet complete. While AIC, AICc, BIC, or MDL provide a method of selecting the number of knot points, a fully fiducial solutions would be desirable. Hannig and Lee (2009) used ideas similar to the reversible-jump MCMC methods for model selection in wavelet regression and a similar methodology should be applicable.

The final question to be resolved is to compare the fiducial method to the competing Bayesian and maximum likelihood methods. The comparison with the Bayesian method BARS (Wallstrom et al., 2007) is not a fair comparison because BARS focuses on the resulting curve values and not the knot point locations or number of knots. Once the model selection question is solved in the fiducial method, a simulation study to compare the predicted values of the fiducial method versus the Bayesian method will be undertaken.

4.6.1 Software

Software to preform the fiducial knot selection is available via Comprehensive R Archive Network (CRAN) as the package `fiducialSplines`. It is my belief that in order to influence the methodology of applied statisticians and researchers in other fields, the statistical tools must be made available in a format that is easy to use and acquire.

The package is has routines for fitting the fiducial spline model, assessing convergence of the Markov chain, calculating statistics from the fiducial distribution, and comparing splines of different order and number of knots.

- `fiducial.spline` calculates several chains and uses the Gelman and Rubin convergence diagnostic Gelman et al. (2003) to assess convergence.
- `fiducial.spline.simple` calculates one chain with a set length.
- `plot`, `confint`, `predict`, `AIC`, `BIC` methods are implemented.
  - `plot` displays the chain path and can be used to assess convergence.
  - `confint` gives either confidence intervals for the parameter values or a confidence band of the spline at a given set of x-values. All confidence intervals are calculated element-wise with no correction for creating many (or an infinite) number of intervals.

  - `predict` gives median spline values for a given set of x-values.
  - `AIC` calculates the AIC (Akaike, 1973) value for each element of the chain.

- `display` plots the input data along with the predicted function along with a confidence band.
- `lengthen`, `merge`, `trim` provide methods to easily manipulate chains during an interactive procedure of assessing chain convergence.

4.7 Appendix A

We start with several assumptions. The assumptions A0-A6 are sufficient for the maximum likelihood estimate to converge asymptotically to a normal distribution and can be found in Lehmann and Casella (1998) as 6.3 (A0)-(A2) and 6.5 (A)-(D). The assumption B2 shows that the Jacobian converges to a prior (Hannig 2009) and B1 is the assumption necessary for the Bayesian solution to converges to that of the MLE Ghosh and Ramamoorthi (Theorem 1.4.1).

4.7.1 Assumptions

4.7.1.1 Conditions for asymptotic normality of the MLE

(A0) The distributions $P_{\boldsymbol{\xi}}$ are distinct.

(A1) The set $\{x : f(x|\boldsymbol{\xi}) > 0\}$ is independent of the choice of $\boldsymbol{\xi}$.

(A2) The data $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ are iid with probability density $f(\cdot|\boldsymbol{\xi})$.

(A3) There exists an open neighborhood about the true parameter value $\boldsymbol{\xi}_0$ such that all third partial derivatives $(\partial^3/\partial\xi_i\partial\xi_j\partial\xi_k)\, f(\boldsymbol{x}|\boldsymbol{\xi})$ exist in the neighborhood, denoted by $B(\boldsymbol{\xi}_0, \delta)$.

(A4) The first and second derivatives of $L(\boldsymbol{\xi}, x) = \log f(x|\boldsymbol{\xi})$ satisfy

$$E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial \xi_j} L(\boldsymbol{\xi}, x)\right] = 0$$

and

$$\begin{aligned} I_{j,k}(\boldsymbol{\xi}) &= E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial \xi_j} L(\boldsymbol{\xi}, x) \cdot \frac{\partial}{\partial \xi_k} L(\boldsymbol{\xi}, x)\right] \\ &= -E_{\boldsymbol{\xi}}\left[\frac{\partial^2}{\partial \xi_j \partial \xi_k} L(\boldsymbol{\xi}, x)\right]. \end{aligned}$$

(A5) The information matrix $I(\boldsymbol{\xi})$ is positive definite for all $\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \delta)$

(A6) There exists functions $M_{jkl}(\boldsymbol{x})$ such that

$$\sup_{\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \delta)} \left| \frac{\partial^3}{\partial \xi_j \partial \xi_k \partial \xi_l} L(\boldsymbol{\xi}, x) \right| \leq M_{j,k,l}(x) \quad \text{and} \quad E_{\boldsymbol{\xi}_0} M_{j,k,l}(x) < \infty$$

4.7.1.2 Conditions for the Bayesian posterior distribution to be close to that of the MLE.

Let $\pi(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}_0} J_0(X_0, \boldsymbol{\xi})$ and $L_n(\boldsymbol{\xi}) = \sum L(\boldsymbol{\xi}, X_i)$

(B1) For any $\delta > 0$ there exists $\epsilon > 0$ such that

$$P_{\boldsymbol{\xi}_0}\left\{ \sup_{\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)} \frac{1}{n} \left(L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0) \right) \leq -\epsilon \right\} \to 1$$

(B2) $\pi(\boldsymbol{\xi})$ is positive at $\boldsymbol{\xi}_0$

4.7.1.3 Conditions for showing that the fiducial distribution is close to the Bayesian posterior

(C1) For any $\delta > 0$

$$\inf_{\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)} \frac{\min_{i=1\ldots n} L(\boldsymbol{\xi}, X_i)}{|L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)|} \xrightarrow{P_{\boldsymbol{\xi}_0}} 0$$

(C2) Let $\pi(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}_0} J_0(X_0, \boldsymbol{\xi})$. The Jacobian function $J(\boldsymbol{X}, \boldsymbol{\xi}) \overset{a.s.}{\to} \pi(\boldsymbol{\xi})$ uniformly on compacts in $\boldsymbol{\xi}$. In the single variable case, this reduces to $J(\boldsymbol{X}, \xi)$ is continuous in ξ, $\pi(\xi)$ is finite and $\pi(\xi_0) > 0$, and for some δ_0

$$E_{\xi_0}\left(\sup_{\xi \in B(\xi_0, \delta)} J_0(\boldsymbol{X}, \xi) \right) < \infty.$$

In the multivariate case, we follow Yeo and Johnson (2001). Let

$$J_j\left(x_1,\ldots,x_j;\boldsymbol{\xi}\right)=E_{\boldsymbol{\xi}_0}\left[J_0\left(x_1,\ldots,x_j,X_{j+1},\ldots,X_k;\boldsymbol{\xi}\right)\right].$$

(C2.a) There exists a integrable and symmetric functions $g\left(x_1,\ldots,x_j\right)$ and compact space $\bar{B}\left(\boldsymbol{\xi}_0,\delta\right)$ such that for $\boldsymbol{\xi}\in\bar{B}\left(\boldsymbol{\xi}_0,\delta\right)$ then $\left|J_j\left(x_1,\ldots,x_j;\boldsymbol{\xi}\right)\right|\leq g\left(x_1,\ldots,x_j\right)$ for $j=1,\ldots,k$.

(C2.b) There exists a sequence of measurable sets S_M^k such that

$$P\left(\mathbb{R}^k-\cup_{M=1}^{\infty}S_M^k\right)=0$$

(C2.c) For each M and for all $j\in 1,\ldots,k$, $J_j\left(x_1,\ldots,x_j;\boldsymbol{\xi}\right)$ is equicontinuous in $\boldsymbol{\xi}$ for $\{x_1,\ldots,x_j\}\in S_M^j$ where $S_M^k=S_M^jS_M^{k-j}$.

4.7.2 Proof of assumptions for free-knot splines using a truncated polynomial basis.

We now consider the free-knot spline case. Suppose we are interested in a p degree (order $m=p+1$) polynomial spline with κ knot points, $\boldsymbol{t}=\{t_1,\ldots,t_\kappa\}^T$ where $t_k\in(a+\delta,b-\delta)$ and $|t_i-t_j|\leq\delta$ for $i\neq j$ and some $\delta>0$. Furthermore, we assume that the data points $\{x_i,y_i\}$ are such that the x_i values are equally spaced along a grid in $[a,b]$. This assumption could be relaxed to an assumption about the rate at which data points are added to any arbitrary region must be proportional to the length of the region, but for simplicity we use equally spaced data.

Denote the truncated polynomial spline basis functions as

$$\begin{aligned}N(x,\boldsymbol{t}) &= \{N_1(x,\boldsymbol{t}),\ldots,N_{\kappa+m}(x,\boldsymbol{t})\}^T\\ &= \{1,x,\ldots,x^p,(x-t_1)_+^p,\ldots,(x-t_\kappa)_+^p\}^T\end{aligned}$$

and let $y_i=N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}+\sigma\epsilon_i$ where $\epsilon_i\overset{iid}{\sim}N(0,1)$ and thus the density function is

$$f(y,\boldsymbol{\xi})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]$$

where $\boldsymbol{\xi}=\{\boldsymbol{t},\boldsymbol{\alpha},\sigma^2\}$ and the log-likelihood function is

$$L(\boldsymbol{\xi},y)=\frac{1}{2}\log 2\pi-\frac{1}{2}\log\sigma^2-\frac{1}{2\sigma^2}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2$$

4.7.2.1 Assumptions A0-A4

Assumptions A0-A2 are satisfied. We now consider assumption A3 and A4. We note that if $p \geq 4$ then the necessary three continuous derivatives exist and now examine the derivatives. Let $\boldsymbol{\theta} = \{\boldsymbol{t}, \boldsymbol{\alpha}\}$ and thus

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y)\right] &= E_{\boldsymbol{\xi}}\left[-\frac{1}{2\sigma^2}2\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right] \\ &= -\frac{1}{2\sigma^2}2\left(E_{\boldsymbol{\xi}}\left[y\right]-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},\boldsymbol{y})\right] &= E_{\boldsymbol{\xi}}\left[-\frac{1}{2\sigma^2}+\frac{1}{2\left(\sigma^2\right)^2}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right] \\ &= -\frac{1}{2\sigma^2}+\frac{1}{2\left(\sigma^2\right)^2}\left(\sigma^2\right) \\ &= 0. \end{aligned}$$

Next we consider information matrix. First we consider the $\boldsymbol{\theta}$ terms.

$$\begin{aligned} &E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},\boldsymbol{y})\frac{\partial}{\partial\theta_k}L(\boldsymbol{\xi},\boldsymbol{y})\right] \\ &= E_{\boldsymbol{\xi}}\left[\frac{1}{\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right] \\ &= \frac{1}{\sigma^4}E_{\boldsymbol{\xi}}\left[\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \\ &= \frac{1}{\sigma^2}\left(\frac{\partial}{\partial\theta_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \end{aligned}$$

The j,k partials for the second derivative are

$$\begin{aligned} &\frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\boldsymbol{\xi},y) \\ &= \frac{\partial}{\partial\theta_j}\left[-\frac{1}{2\sigma^2}2\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right] \\ &= \frac{\partial}{\partial\theta_j}\left[-\frac{1}{\sigma^2}\left(-y_i\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)+N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right)\right] \\ &= -\frac{1}{\sigma^2}\left[-y\frac{\partial^2}{\partial\theta_j\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}+\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right. \\ &\qquad\left.+N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\frac{\partial^2}{\partial\theta_j\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right] \end{aligned}$$

which have expectation

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\boldsymbol{\xi},y)\right] &= -\frac{1}{\sigma^2}\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \\ &= -E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\theta_k}L(\boldsymbol{\xi},\boldsymbol{y})\right] \end{aligned}$$

as necessary. Next we consider

$$\begin{aligned} &E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right] \\ &= E_{\boldsymbol{\xi}}\left[\frac{1}{\sigma^2}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\left[-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]\right] \\ &= E_{\boldsymbol{\xi}}\left[-\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}+\frac{1}{2\sigma^6}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^3\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right] \\ &= 0 \end{aligned}$$

which is equal to

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j\partial\sigma^2}L(\boldsymbol{\xi},\boldsymbol{y})\right] &= E_{\boldsymbol{\xi}}\left[\frac{2}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right] \\ &= 0. \end{aligned}$$

Finally

$$\begin{aligned} &E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right] \\ &= E_{\boldsymbol{\xi}}\left[\left\{-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right\}\left\{-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right\}\right] \\ &= E_{\boldsymbol{\xi}}\left[\frac{1}{4\sigma^4}-\frac{2}{4\sigma^6}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2+\frac{1}{4\sigma^8}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^4\right] \\ &= \frac{1}{4\sigma_0^4}-\frac{2}{4\sigma_0^6}\sigma_0^2+\frac{1}{4\sigma_0^8}3\sigma_0^4 \\ &= \frac{2}{4\sigma_0^4} \end{aligned}$$

which is equal to

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\sigma^2\partial\sigma^2}L(\boldsymbol{\xi},y)\right] &= E_{\boldsymbol{\xi}}\left[\frac{1}{2}\sigma^{-4}-\frac{2}{2}\sigma^{-6}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right] \\ &= \frac{1}{2}\sigma^4-\frac{2}{2}\sigma^4. \end{aligned}$$

Therefore the interchange of integration and differentiation is justified.

4.7.2.2 Assumptions A5

To address whether the information matrix is positive definite, we notice that since $E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right]>0$ and $E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right]=0$, we only need to be concerned with the sub-matrix

$$\begin{aligned} I_{j,k}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y_i)\,\frac{\partial}{\partial\theta_k}L(\boldsymbol{\xi},y_i)\right] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right). \end{aligned}$$

where the σ^{-2} term can be ignored because it doesn't affect the positive definiteness. First we note

$$\frac{\partial}{\partial t_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha} = -p\,(x_i-t_j)_+^{p-1}\,\alpha_{p+j+1}$$

$$\frac{\partial}{\partial\alpha_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha} = N_j(x_i,\boldsymbol{t}).$$

If we let

$$X=\begin{bmatrix} N_1(x_1,\boldsymbol{t}) & \cdots & N_{m+\kappa}(x_1,\boldsymbol{t}) & \frac{\partial}{\partial t_1}N(x_1,\boldsymbol{t})^T\boldsymbol{\alpha} & \cdots & \frac{\partial}{\partial t_\kappa}N(x_1,\boldsymbol{t})^T\boldsymbol{\alpha} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ N_1(x_n,\boldsymbol{t}) & \cdots & N_{m+\kappa}(x_n,\boldsymbol{t}) & \frac{\partial}{\partial t_1}N(x_n,\boldsymbol{t})^T\boldsymbol{\alpha} & \cdots & \frac{\partial}{\partial t_\kappa}N(x_n,\boldsymbol{t})^T\boldsymbol{\alpha} \end{bmatrix}$$

then $I(\boldsymbol{\theta})=X^TX$. Then $I(\boldsymbol{\theta})$ is positive definite if the columns of X are linearly independent. This is true under the assumptions that $t_j\neq t_k$ and that $\alpha_{m+j}\neq 0$.

4.7.2.3 Assumptions A6

We next consider a bound on the third partial derivatives. We start with the derivatives of the basis functions.

$$\frac{\partial^2}{\partial t_j\partial t_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}=0 \qquad \text{if } j\neq k$$

$$\frac{\partial^2}{\partial t_j\partial t_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha} = p(p-1)\,(x-t_j)_+^{p-2}\,\alpha_{p+j+1}$$

$$\frac{\partial^2}{\partial\alpha_j\partial\alpha_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}=0$$

$$\frac{\partial^2}{\partial t_j \partial \alpha_{p+j+1}} N(x,\boldsymbol{t})^T \boldsymbol{\alpha} = -p\,(x-t_j)_+^{p-1}$$

$$\frac{\partial^3}{\partial t_j \partial t_j \partial t_j} N(x,\boldsymbol{t})^T \boldsymbol{\alpha} = -p(p-1)(p-2)\,(x-t_j)_+^{p-3}\,\alpha_{p+j+1}$$

$$\frac{\partial^3}{\partial t_j \partial t_j \partial \alpha_{p+j+1}} N(x,\boldsymbol{t})^T \boldsymbol{\alpha} = p(p-1)\,(x-t_j)_+^{p-2}$$

Since x is an element of a compact set, then for $\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \delta)$ all of the above partials are bounded as is $N(x,\boldsymbol{t})^T\boldsymbol{\alpha}$. Therefore

$$\begin{aligned}
&\frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} L(\boldsymbol{\xi}, x) \\
&= -\frac{1}{\sigma^2}\left[-y\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} + \left(\frac{\partial^2}{\partial\theta_j\partial\theta_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial^2}{\partial\theta_l} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right. \\
&\quad + \left(\frac{\partial^2}{\partial\theta_j\partial\theta_l} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial^2}{\partial\theta_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \\
&\quad + \left(\frac{\partial^2}{\partial\theta_l\partial\theta_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial^2}{\partial\theta_j} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \\
&\quad \left. + N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\left(\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right]
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\sigma^2} L(\boldsymbol{\xi}, x) \\
&= \frac{1}{\sigma^4}\left[-y\frac{\partial^2}{\partial\theta_j\partial\theta_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} + \left(\frac{\partial}{\partial\theta_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_j} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right. \\
&\qquad \left. + N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\frac{\partial^2}{\partial\theta_j\partial\theta_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right]
\end{aligned}$$

and

$$\frac{\partial^3}{\partial\theta_j \partial\sigma^2 \partial\sigma^2} L(\boldsymbol{\xi}, y) = -\frac{2}{\sigma^6}\left(y - N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_j} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)$$

and

$$\frac{\partial^3}{\partial\sigma^2 \partial\sigma^2 \partial\sigma^2} L(\boldsymbol{\xi}, \boldsymbol{y}) = -\frac{1}{\sigma^6} + \frac{3}{\sigma^8}\left(y - N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2$$

are also bounded $\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \delta)$ since $\sigma_0^2 > 0$ by assumption. The expectation of the bounds also clearly exists.

4.7.2.4 Lemmas

To show that the remaining assumptions are satisfied, we first examine the behavior of

$$g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i) = N(x_i, \boldsymbol{t}_0)^T \boldsymbol{\alpha}_0 - N(x_i, \boldsymbol{t})^T \boldsymbol{\alpha}.$$

Notice that for x_i chosen on a uniform grid over $[a, b]$ then

$$\frac{1}{n} \sum_{i=1}^{n} \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 \;\; \rightarrow \;\; \frac{1}{b-a} \int_a^b \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x)\right)^2 \, dx.$$

Furthermore we notice that $g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x\right)$ is also a spline. The sum of the two splines is also a spline. Consider the degree p case of $g\left(x|\boldsymbol{\alpha}, t\right) + g\left(x|\boldsymbol{\alpha}^*, t^*\right)$ where $t < t^*$. Then the sum is a spline with knot points $\{t, t^*\}$ and whose first $p+1$ coefficients are $\boldsymbol{\alpha} + \boldsymbol{\alpha}^*$ and last two coefficients are $\left\{\alpha_{p+1}, \alpha^*_{p+1}\right\}$.

At this point we also notice

$$\begin{aligned}
E\left[n^{-1} \sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right) \epsilon_i\right] &= n^{-1} \sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right) E\left[\epsilon_i\right] \\
&= 0 \\
\\
V\left[n^{-1} \sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right) \epsilon_i\right] &= n^{-2} V\left[\sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right) \epsilon_i\right] \\
&= n^{-2} \sum V\left[g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right) \epsilon_i\right] \\
&= n^{-2} \sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)^2 V\left[\epsilon_i\right] \\
&= n^{-2} \sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)^2 \\
&\rightarrow 0
\end{aligned}$$

and that $\sum \epsilon_i^2 \sim \chi^2_n$ and thus $n^{-1} \sum \epsilon_i^2$ converges in probability to the constant 1. Therefore, by the SLLN,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right) + \sigma_0 \epsilon_i\right]^2 &= \frac{1}{n} \sum_{i=1}^{n} \left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right)\right]^2 + \frac{2\sigma_0}{n} \sum_{i=1}^{n} \epsilon_i g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right) + \frac{\sigma_0^2}{n} \sum_{i=1}^{n} \epsilon_i^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right)\right]^2 + O_p\left(n^{-1}\right) + \frac{\sigma_0^2}{n} \sum_{i=1}^{n} \epsilon_i^2 \\
&\stackrel{a.s.}{\rightarrow} \frac{1}{b-a} \int_a^b \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x)\right)^2 \, dx + \sigma_0^2.
\end{aligned}$$

Lemma 4. *Given a degree p polynomial $g(x|\boldsymbol{\alpha})$ on $[a,b]$ with coefficients $\boldsymbol{\alpha}$, then $\exists\, \lambda_{n,m}, \lambda_{n,M} > 0$ such that $||\boldsymbol{\alpha}||^2\lambda_{n,m}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left[g(x_i|\boldsymbol{\alpha})\right]^2 \leq ||\boldsymbol{\alpha}||^2\lambda_{n,M}^2$.*

Proof. If $\boldsymbol{\alpha} = \mathbf{0}$, then $g\left(x|\boldsymbol{\alpha}\right) = 0$ and the result is obvious. If $g\left(x|\boldsymbol{\alpha}\right)$ is a polynomial with at least one non-zero coefficient, it therefore cannot be identically zero on $[a,b]$ and therefore for $n > p$ then $\frac{1}{n}\sum\left[g(x_i|\boldsymbol{\alpha})\right]^2 > 0$ since the polynomial can only have at most p zeros. We notice that

$$\begin{aligned}
\int_a^b \left[g(x|\boldsymbol{\alpha})\right]^2 dx &= \int_a^b \left[\sum_{i=0}^{p}\alpha_i^2 x^{2i} + 2\sum_{i=0}^{p-1}\sum_{j=i+1}^{p}\alpha_i\alpha_j x^{i+j}\right] dx \\
&= \sum_{i=0}^{p}\frac{\alpha_i^2}{i+1}x^{2i+1} + 2\sum_{i=0}^{p-1}\sum_{j=i+1}^{p}\frac{\alpha_i\alpha_j}{i+j+1}x^{i+j+1}\Bigg|_{x=a}^{b} \\
&= \boldsymbol{\alpha}^T X \boldsymbol{\alpha}
\end{aligned}$$

where the matrix $\boldsymbol{X}$ has i,j element $\left(b^{i+j} - a^{i+j}\right)/(i+j)$. Since $\int_a^b \left[g(x|\boldsymbol{\alpha})\right]^2 dx > 0$ for all $\boldsymbol{\alpha}$ then the matrix X must be positive definite. Next we notice that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left[g(x_i|\boldsymbol{\alpha})\right]^2 &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}^T\boldsymbol{X}_i\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\left(\frac{1}{n}\sum\boldsymbol{X}_i\right)\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\boldsymbol{X}_n\boldsymbol{\alpha}
\end{aligned}$$

and therefore $\boldsymbol{X}_n \to \boldsymbol{X}$ and therefore, denoting the eigen-values of $\boldsymbol{X}_n$ as $\boldsymbol{\lambda}_n$ and the eigenvalues of $\boldsymbol{X}$ as $\boldsymbol{\lambda}$, we have $\boldsymbol{\lambda}_n \to \boldsymbol{\lambda}$

Letting $\lambda_{n,m}$ and $\lambda_{n,M}$ be the minimum and maximum eigen-values of $\boldsymbol{X}_n$ be the largest, then $\lambda_{n,m}^2\left\|\boldsymbol{\alpha}\right\|^2 \leq \frac{1}{n}\sum\left[g(x|\boldsymbol{\alpha})\right]^2 \leq \lambda_{n,M}^2\left\|\boldsymbol{\alpha}\right\|^2$. □

The values $\lambda_{n,m}, \lambda_{n,M}$ depend on the interval that the polynomial is integrated/summed over and that if $a = b$, then the integral is zero. In the following lemmas, we assume that there is some minimal distance between two knot-points and between a knot-point and the boundary values a, b.

Lemma 5. *Given a degree p spline $g(x|\boldsymbol{\theta})$ with κ knot points on $[a,b]$, let $\tau = (|a| \vee |b|)^{\kappa}$. Then $\forall\ \delta > 2\tau$, $\exists\ \lambda_n > 0$ such that if $\|\boldsymbol{\theta}\| > \delta$ then $\frac{1}{n}\sum \left[g(x_i|\boldsymbol{\theta})\right]^2 > (\delta^2 + \tau^2)\,\lambda_n$.*

Proof. Notice that $||\boldsymbol{\theta}||^2 > \delta^2 > 4\tau^2$ implies $||\boldsymbol{\alpha}||^2 > \delta^2 - \tau^2$. First we consider the case of $\kappa = 1$. If $\alpha_0^2 + \cdots + \alpha_p^2 > (\delta^2 + \tau^2)/9$ then $\frac{1}{n}\sum \left[g(x_i|\boldsymbol{\theta})\right]^2 1_{[a,t]}(x_i) > \lambda_n (\delta^2 + \tau^2)$ for some $\lambda_n > 0$. If $\alpha_0^2 + \cdots + \alpha_p^2 \leq (\delta^2 + \tau^2)/9$ then $\alpha_{p+1}^2 \geq 3(\delta^2 + \tau^2)/4$. Therefore $(\alpha_p + \alpha_{p+1})$, the coefficient of the x^p term of the polynomial on $[t_1, b]$ is

$$\begin{aligned} \|\alpha_p + \alpha_{p+1}\|^2 &> \|\alpha_{p+1}\|^2 - \|\alpha_p\|^2 \\ &> \frac{3(\delta^2 + \tau^2)}{4} - \frac{(\delta^2 + \tau^2)}{4} \\ &> \frac{1}{2}\left(\delta^2 + \tau^2\right) \end{aligned}$$

and thus the squared norm of the coefficients of the polynomial on $[t_1, b]$ must also be greater than $\frac{1}{2}(\delta^2 + \tau^2)$ and thus $\frac{1}{n}\sum \left[g(x_i|\boldsymbol{\theta})\right]^2 1_{[t,b]}(x_i) > \lambda_n (\delta^2 + \tau^2)$ for some $\lambda_n > 0$. The proof for multiple knots is similar, only examining all $\kappa + 1$ polynomial sections for one with coefficients with squared norm larger than some fraction of $(\delta^2 + \tau^2)$. □

Lemma 6. *For all $\delta > 0$, there exists $\lambda_n > 0$ such that for all $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \delta)$ then $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 > \lambda_n \delta$.*

Proof. By the previous lemma, for all $\Delta > 2\tau$ there exists $\exists\, \Lambda_n > 0$ such that for all $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \Delta)$ then $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 > \Lambda_n \Delta$. We now consider the region

$$\mathcal{C} = \text{closure}\left[B(\boldsymbol{\theta}_0, \Delta) \setminus B(\boldsymbol{\theta}_0, \delta)\right]$$

Assume to the contrary that there exists $\delta > 0$ such that $\forall\, \lambda_n > 0$, $\exists\, \boldsymbol{\theta} \in \mathcal{C}$ such that $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 \leq \lambda_n \delta$ and we will seek a contradiction. By the negation, there exists a sequence $\boldsymbol{\theta}_n \in \mathcal{C}$ such that $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 \to 0$. But since $\boldsymbol{\theta}_n$ is in a compact space, there exists a sub-sequence $\boldsymbol{\theta}_{n_k}$ that converges to $\boldsymbol{\theta}_\infty \in \mathcal{C}$ and $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 = 0$. But since $\boldsymbol{\theta}_0 \notin \mathcal{C}$ this is a contradiction. □

Corollary 7. *There exists λ such that for any $\delta > 0$ and $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \delta)$*

$$\frac{1}{n}\sum_{i=1}^{n}\left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right) + \sigma_0\epsilon_i\right]^2 \quad \geq \quad \lambda_n^2\delta^2 + O_p\left(n^{-1/2}\right) + \sigma_0^2.$$

We now focus our attention on the ratio of the maximum value of a polynomial and its integral.

Lemma 8. *Given a degree p polynomial $g\left(x|\boldsymbol{\alpha}\right)$ on $[a, b]$, then*

$$\frac{\max_{i\in\{1,\ldots,n\}}\left[g\left(x_i|\boldsymbol{\alpha}\right)\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[g\left(x_i|\boldsymbol{\alpha}\right)\right]^2 dx} \leq \frac{\lambda_M^2}{\lambda_{n,m}^2} \rightarrow \frac{\lambda_M^2}{\lambda_m^2}$$

for some $\lambda_M, \lambda_m > 0$.

Proof. Since we can write $\left[g\left(x|\boldsymbol{\alpha}\right)\right]^2 = \boldsymbol{\alpha}^T W_x \boldsymbol{\alpha}$ for some non-negative definite matrix W_x which has a maximum eigen-value $\lambda_{M,x}$, and because the the maximum eigen-value is a continuous function in x, let $\lambda_M = \sup \lambda_{M,x}$. Then the maximum of $\left[g\left(x|\boldsymbol{\alpha}\right)\right]^2$ over $x \in [a, b]$ is less than $\left\|\boldsymbol{\alpha}\right\|^2 \lambda_M^2$. The denominator is bounded from below by $\left\|\boldsymbol{\alpha}\right\|^2 \lambda_{n,m}^2$. □

Lemma 9. *Given a degree p spline $g\left(x|\boldsymbol{\theta}\right)$ on $[a, b]$, then*

$$\frac{\max\left[g\left(x|\boldsymbol{\theta}\right)\right]^2}{\int_a^b\left[g\left(x|\boldsymbol{\theta}\right)\right]^2 dx} \leq \frac{\lambda_M^2}{\lambda_m^2}$$

for some $\lambda_M, \lambda_m > 0$.

Proof. Since a degree p spline is a degree p polynomial on different regions defined by the knot-points, and because the integral over the whole interval $[a, b]$ is greater than the integral over the regions defined by the knot-points, we can use the previous lemma on each section and then chose the largest ratio. □

Lemma 10. *Given a degree p spline $g\left(x|\boldsymbol{\theta}\right)$ on $[a, b]$ then*

$$\frac{n^{-1/2}\max_i\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)\right]^2}{n^{-1}\sum_{i=1}^{n}\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)\right]^2} = O_p\left(1\right) \tag{4.7.1}$$

uniformly over $\boldsymbol{\theta}$.

Proof. Notice

$$\begin{aligned}
\frac{n^{-1/2}\max_i\left[\epsilon_i\sigma_0+g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0+g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2} &\leq \frac{2n^{-1/2}\max_i\left[\epsilon_i^2\sigma_0^2\right]+2n^{-1/2}\max_i\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0+g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2} \\
&= \frac{2\sigma_0^2n^{-1/2}\max_i\epsilon_i^2+\max_i\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0+g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2} \\
&= \frac{O_p\left(\frac{\log n}{\sqrt{n}}\right)+\max_i\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0+g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2}
\end{aligned}$$

and since $n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0+g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2 \xrightarrow{P} \frac{1}{b-a}\int_a^b\left(g(\boldsymbol{\theta}_0,\boldsymbol{\theta},x)\right)^2\,dx+\sigma_0^2$, and lemma 8 bounds the ratio of the terms that involve $\boldsymbol{\theta}$, this ratio is bounded in probability uniformly over $\boldsymbol{\theta}$. □

4.7.2.5 Assumptions B1

Returning to assumption B1, we now consider $\boldsymbol{\xi}\notin B(\boldsymbol{\xi}_0,\delta)$ and

$$\begin{aligned}
L_n\left(\boldsymbol{\xi}\right) &= \sum\log\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-1}{2\sigma}\sum\left(y_i-N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]\right\} \\
&= -\frac{n}{2}\log\left(2\pi\right)-n\log\sigma-\frac{1}{2\sigma}\sum\left[y_i-N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right]^2 \\
&= -\frac{n}{2}\log\left(2\pi\right)-n\log\sigma-\frac{1}{2\sigma}\sum\left[N(x_i,\boldsymbol{t}_0)^T\boldsymbol{\alpha}_0+\sigma_0\epsilon_i-N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right]^2 \\
&= -\frac{n}{2}\log\left(2\pi\right)-n\log\sigma-\frac{1}{2\sigma}\sum\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)+\sigma_0\epsilon_i\right]^2
\end{aligned}$$

and therefore

$$\begin{aligned}
&\frac{1}{n}\left(L_n(\boldsymbol{\xi})-L_n(\boldsymbol{\xi}_0)\right) \\
&= -\log\sigma-\frac{1}{2n\sigma^2}\sum\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)+\sigma_0\epsilon_i\right]^2+\log\sigma_0+\frac{1}{2n\sigma_0}\sum\left[g\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0,x_i\right)+\sigma_0\epsilon_i\right]^2 \\
&= \log\frac{\sigma_0}{\sigma}-\frac{1}{2n\sigma^2}\sum\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)+\sigma_0\epsilon_i\right]^2+\frac{1}{2n\sigma_0^2}\sum\left[\sigma_0\epsilon_i\right]^2 \\
&= \log\frac{\sigma_0}{\sigma}-\frac{\left(\lambda_n\left(\boldsymbol{\theta},\boldsymbol{\theta}_0\right)\right)^2}{2\sigma^2}-\frac{\sigma_0^2}{2\sigma^2}+\frac{1}{2n}\sum\left[\epsilon_i\right]^2
\end{aligned}$$

where

$$\left[\lambda_n\left(\boldsymbol{\theta},\boldsymbol{\theta}_0\right)\right]^2 = \frac{1}{n}\sum\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)+\sigma_0\epsilon_i\right]^2-\sigma_0^2$$

which converges in probability to $\frac{1}{b-a}\int_a^b \left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x\right)\right]^2 dx$. The function goes to $-\infty$ as $\sigma \to 0$ and $\sigma \to \infty$. Taking the derivative

$$\frac{d}{d\sigma}\left[\log\frac{\sigma_0}{\sigma} - \frac{1}{2\sigma^2}\left[\left(\lambda_n\right)^2 + \sigma_0^2\right] + \frac{1}{2n}\sum \epsilon_i^2\right] = -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left[\left(\lambda_n\right)^2 + \sigma_0^2\right]$$

and setting it equal to zero yields a single critical point of at $\sigma^2 = \left[\left(\lambda_n\right)^2 + \sigma_0^2\right]$ which results in a maximum of

$$\log\left(\frac{\sigma_0}{\sqrt{\left(\lambda_n\right)^2 + \sigma_0^2}}\right) - \frac{1}{2} + \frac{1}{2}n^{-1}\sum \epsilon_i^2 \tag{4.7.2}$$

which bounded away from zero in probability for $\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)$

4.7.2.6 Assumption C1

Assumption C1 is

$$\inf_{\boldsymbol{\xi}\notin B(\boldsymbol{\xi}_0,\delta)} \frac{\min_{i=1\ldots n} L(\boldsymbol{\xi}, X_i)}{\left|L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right|} \xrightarrow{P_{\boldsymbol{\xi}_0}} 0$$

First notice

$$\begin{aligned} L(\boldsymbol{\xi}, Y_i) &= -\frac{1}{2}\log\left(2\pi\right) - \log\sigma - \frac{1}{2\sigma^2}\left(Y_i - N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2 \\ &= -\frac{1}{2}\log\left(2\pi\right) - \log\sigma - \frac{1}{2\sigma^2}\left(\epsilon_i\sigma_0 + N(x_i,\boldsymbol{t}_0)^T\boldsymbol{\alpha}_0 - N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2 \\ &= -\frac{1}{2}\log\left(2\pi\right) - \log\sigma - \frac{1}{2\sigma^2}\left(\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0,\boldsymbol{\theta},x_i)\right)^2 \end{aligned}$$

and we consider $\mathcal{C} = \{\boldsymbol{\xi} : \boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0,\delta)\}$. Define

$$\begin{aligned} f_n\left(\boldsymbol{\xi}\right) &= \frac{\min\ L\left(\boldsymbol{\xi}, Y_i\right)}{\left|L_n\left(\boldsymbol{\xi}\right) - L_n\left(\boldsymbol{\xi}_0\right)\right|} \\ &= \frac{-\frac{1}{2}\log\left(2\pi\right) - \log\sigma - \frac{1}{2\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0,\boldsymbol{\theta},x_i)\right]^2}{n\cdot\frac{1}{n}\left|L_n\left(\boldsymbol{\xi}\right) - L_n\left(\boldsymbol{\xi}_0\right)\right|} \end{aligned}$$

and notice that the denominator is bounded away from 0 by 4.7.2.

$$\begin{aligned}
f_n(\boldsymbol{\xi}) &= \frac{-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{-n\cdot\frac{1}{n}\left(L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right)} \\
&= \frac{\frac{1}{\sqrt{n}}\left[-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2\right]}{-\sqrt{n}\cdot\frac{1}{n}\left[n\log\frac{\sigma_0}{\sigma} - \frac{1}{2\sigma^2}\sum\left[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i\right]^2 + \frac{1}{2}\sum\epsilon_i^2\right]} \\
&= \frac{1}{\sqrt{n}}\cdot\frac{-\frac{1}{2\sqrt{n}}\log(2\pi) - \frac{1}{\sqrt{n}}\log\sigma - \frac{1}{2\sqrt{n}\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{-\log\frac{\sigma_0}{\sigma} + \frac{1}{2n\sigma^2}\sum\left[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i\right]^2 - \frac{1}{2n}\sum\epsilon_i^2} \\
&= \frac{1}{\sqrt{n}}\left[\frac{-\frac{1}{2\sqrt{n}}\log(2\pi)}{-\log\frac{\sigma_0}{\sigma} + \frac{1}{2n\sigma^2}\sum\left[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i\right]^2 - \frac{1}{2n}\sum\epsilon_i^2} + \right. \\
&\qquad \left. \frac{-\frac{1}{\sqrt{n}}\log\sigma - \frac{1}{2\sqrt{n}\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{-\log\frac{\sigma_0}{\sigma} + \frac{1}{2n\sigma^2}\sum\left[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i\right]^2 - \frac{1}{2n}\sum\epsilon_i^2}\right]
\end{aligned}$$

We consider the infimums of the terms inside the brackets separately.

For the first term, since the denominator is bounded in probability above 0 uniformly in $\boldsymbol{\theta}$, and the numerator goes to zero, the infimum of the first term goes to 0 in probability.

The second term is uniformly bounded over $\boldsymbol{\theta}$ by lemma 9. Notice that the numerator is

$$\begin{aligned}
&-\frac{1}{\sqrt{n}}\log\sigma - \frac{1}{2\sqrt{n}\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2 \\
&\geq -\frac{1}{\sqrt{n}}\log\sigma - \frac{\max\left[\epsilon_i\sigma_0\right]^2}{\sqrt{n}\sigma^2} - \frac{\max\left[g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{\sqrt{n}\sigma^2} \\
&= -\frac{1}{\sqrt{n}}\log\sigma - \frac{\sigma_0^2\, O_p(\log\, n)}{\sqrt{n}\sigma^2} - \frac{\max\left[g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{\sqrt{n}\sigma^2} \\
&\geq \frac{-\log n}{\sqrt{n}}\log\sigma - \frac{\sigma_0^2\, O_p(\log\, n)}{\sqrt{n}\sigma^2} - \frac{\max\left[g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{\sqrt{n}\sigma^2}
\end{aligned}$$

and all three terms of the numerator converge to 0 for every σ. Therefore for $\sigma \in [0, d]$ for some large d, the infimum converges to 0. For $\sigma > d$, the $\log\sigma$ terms dominate and the infimum occurs at $\sigma = d$ which also converges to 0. Therefore

$$\inf_{\boldsymbol{\xi}\notin B(\boldsymbol{\xi}_0,\delta)} \frac{\min L(\boldsymbol{\xi}, Y_i)}{\left|L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right|} \xrightarrow{P} 0.$$

4.7.2.7 Assumptions C2

Finally we turn our attention to the Jacobian. Recall that the Jacobian is

$$J_0\left(\boldsymbol{y}_0, \boldsymbol{\xi}\right) = \left| \frac{1}{\sigma^2} p^{\kappa} \left[\prod_{j=1}^{\kappa} \alpha_{p+\kappa} \right] \det \left[\begin{array}{ccc} \boldsymbol{B}_{\boldsymbol{\alpha}} & \boldsymbol{B}_{\boldsymbol{t}} & \boldsymbol{B}_{\sigma^2} \end{array} \right] \right|$$

where

$$\boldsymbol{B}_{\boldsymbol{\alpha}} = \left[\begin{array}{ccccccc} 1 & x_{(1)} & \dots & x_{(1)}^p & (x_{(1)} - t_1)_+^p & \dots & (x_{(1)} - t_\kappa)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{(l)} & \dots & x_{(l)}^p & (x_{(l)} - t_1)_+^p & \dots & (x_{(l)} - t_\kappa)_+^p \end{array} \right],$$

$$\boldsymbol{B}_{\boldsymbol{t}} = \left[\begin{array}{ccc} \left(x_{(1)} - t_1\right)_+^{p-1} & \dots & \left(x_{(1)} - t_\kappa\right)_+^{p-1} \\ \vdots & \ddots & \vdots \\ p\left(x_{(l)} - t_1\right) & \dots & \left(x_{(l)} - t_\kappa\right)_+^{p-1} \end{array} \right],$$

and

$$\boldsymbol{B}_{\sigma^2} = \left[\begin{array}{c} -\frac{1}{2}\left(y_{(1)} - g(x_{(1)}|\boldsymbol{\theta})\right) \\ \vdots \\ -\frac{1}{2}\left(y_{(l)} - g(x_{(l)}|\boldsymbol{\theta})\right) \end{array} \right].$$

Following the notation of Yeo and Johnson, we suppress parenthesis and 0 subscripts. We consider the $\boldsymbol{\xi}$ in compact space $\bar{B}(\boldsymbol{\xi}_0, \delta)$. We notice that for $\delta < \sigma^{-2}$ that $J(\boldsymbol{y}; \boldsymbol{\xi}) \leq \delta^{\kappa+1} p^{\kappa} g(\boldsymbol{y})$ for some $g(\boldsymbol{y})$ because $\boldsymbol{B}_{\boldsymbol{\alpha}}$ and $\boldsymbol{B}_{\boldsymbol{t}}$ are functions of $\boldsymbol{x}, \boldsymbol{t}$ which are bounded.

We let S_M^l be the unit square in $\mathbb{R}^l$ of radius M.

Finally we notice that $J_j(y_1, \dots, y_j; \boldsymbol{\xi}) = E\left[J\left(y_1, \dots, y_j, Y_{j+1}, \dots, Y_l; \boldsymbol{\xi}\right)\right]$ is a polynomial in $\boldsymbol{\theta}$ scaled by σ^2, which is equicontinuous on compacts of $\boldsymbol{\xi}$ where σ is bounded away from 0.

Chapter 5

A GENERALIZATION OF PIECEWISE LINEAR MODELS: FREE-KNOT SPLINES AND FIDUCIAL INFERENCE

5.1 Introduction

Changes in environmental drivers can cause some ecosystems to show sudden non-linear changes (May 1977; Connell and Sousa 1983; Knowlton 1992; Estes and Duggins 1995; Groffman et al. 2006) and modeling these events is an important research area. Toms and Lesperance (2003) suggest modeling ecological thresholds using a piecewise linear model. As its name suggests, linear functions are fit on adjacent regions of the independent axis but constrained so as the resulting function is continuous. Often referred to as the 'broken stick' model, the piecewise linear model is the simplest case of free-knot splines. This approach has been successfully used in a variety of studies (Homan et al. 2004; Walsh et al. 2005; Carbone et al. 2007; Kinupp and Magnusson 2005).

Toms and Lesperance (2003) suggest three methods for calculating confidence intervals for the threshold. Asymptotic normality, inverted F-test, and bootstrapping were investigated and the authors suggested that the inverted F-test was better in large sample cases and bootstrapping was better for small sample sizes. We further investigate the performance of the latter two tests along with the Bayesian and fiducial solutions to the free-knot spline problem.

We believe that there is also a second type of threshold that is often of interest. A change in the second derivative from increasing to decreasing or vice-verse can be

seen in a sigmoid curve. The piecewise linear model is not applicable here, but a free-knot spline of degree two is.

5.2 Simulations

We performed a simulation study with two levels of variance, three levels of sample size, and two locations of the change point for both the piece-wise linear and sigmoid cases.

We first consider coverage rates of the various methods. In the 'coverage plots' presented, the x-axis denotes the desired confidence level and the y-axis is the observed coverage rate in the experiment. If the observed coverage rate is below the equivalence line ($y = x$), then the method is considered *liberal* and if the observed rate is above the equivalence line then the method is *conservative*. Ideally a method would lie exactly on the equivalence line but a conservative method is more preferable to a liberal because claiming a 95% coverage rate when in truth the coverage rate is less is a more serious error than having the true coverage rate being larger than claimed. The only complaint against a conservative method is that the lengths of confidence intervals are larger than necessary to achieve the desired confidence level.

5.2.1 Centered Knot Point

We first consider the case with the true knot point in the center of the independent axis. At larger sample sizes all the methods performed reasonably, however the bootstrap method was abnormally liberal compared to the other methods at small sample sizes and small variance. At a small sample size and large variability all the methods had issues but the Fiducial and Bayesian methods performed better than the bootstrap and inverse-F methods in the piecewise-linear case. In the cases where the bootstrap method was not significantly liberal, the median confidence interval length was longer than the other methods.

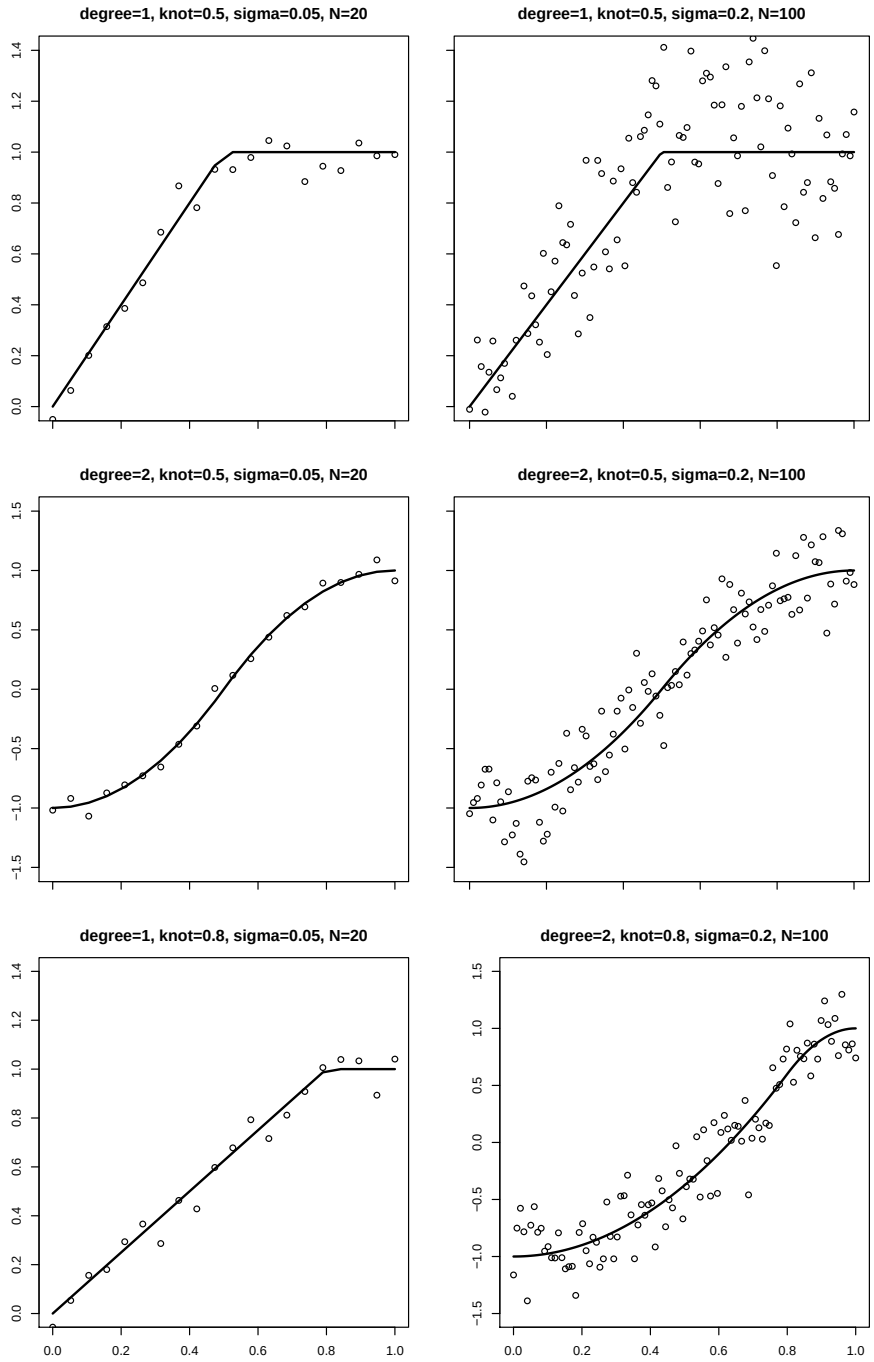

Figure 5.2.1: Examples of the simulated datasets.

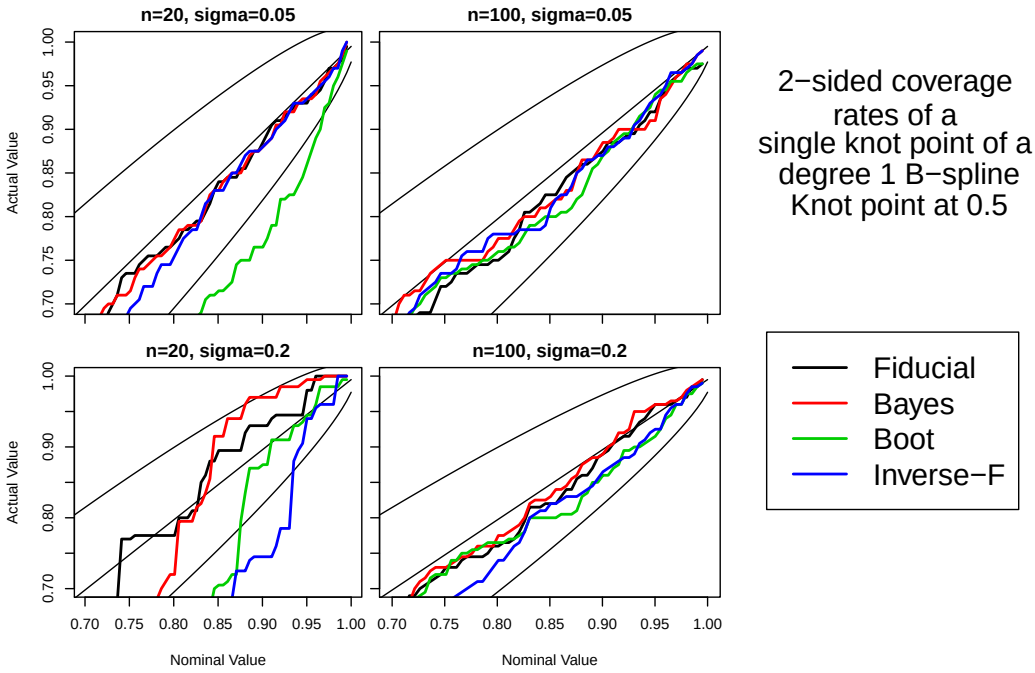

Figure 5.2.2: Coverage plots for piecewise linear, center knot point simulation. Each graph is a plot of the nominal confidence level versus the observed coverage rate. The thin straight line is the equivalence line and the curved thin black lines give the region of coverage rates that could reasonably occur by chance.

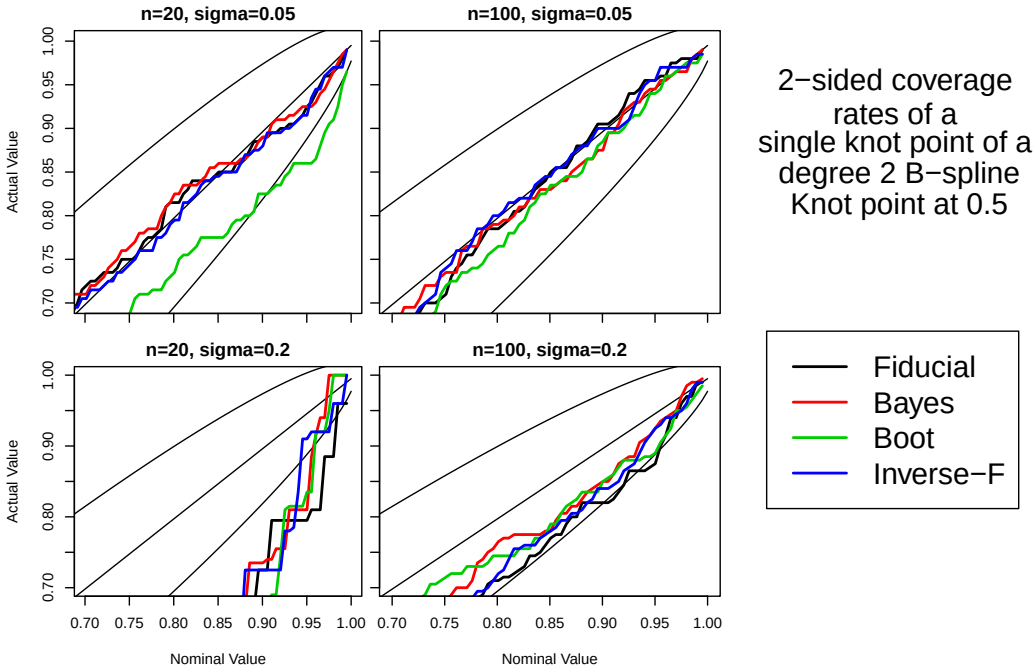

Figure 5.2.3: Coverage plots for the sigmoid curve with center knot point.

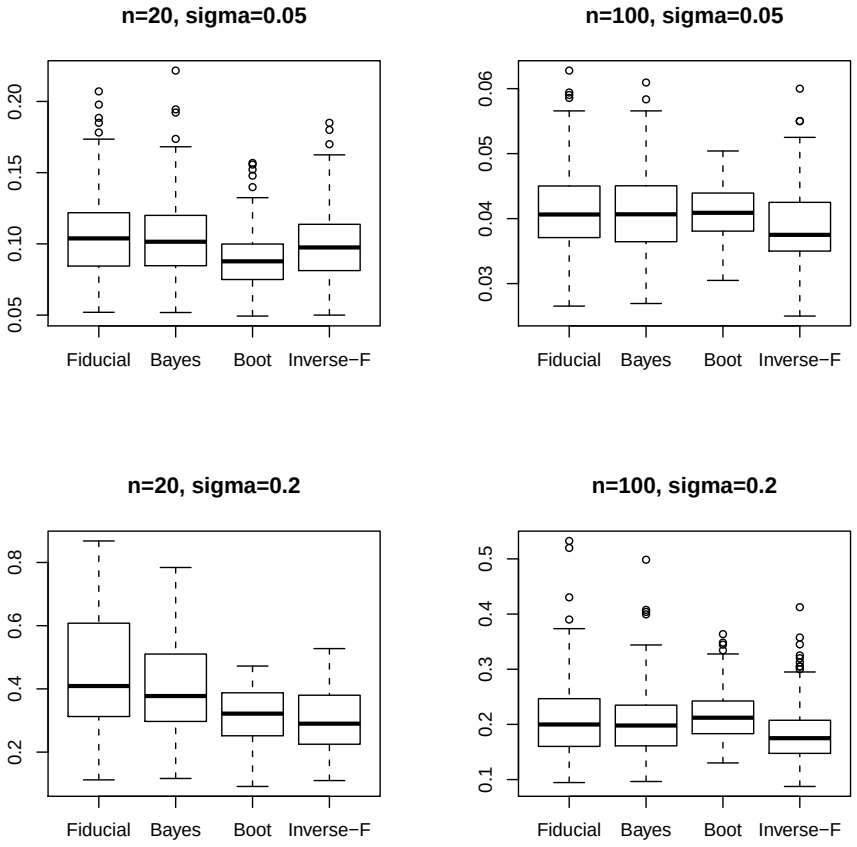

Figure 5.2.4: Confidence interval lengths for piecewise-linear with center knot point.

Next we compare the mean length of a confidence interval for each method. In general the bootstrap method has smaller variability than the other methods, and in when the bootstrap method was liberal, the confidence interval length was shorter than other methods. Among the other three methods, the Bayesian and fiducial methods tend to have a similar mean and variance for the confidence interval length. The inverted-F test performed well with slightly smaller confidence interval widths than the fiducial or Bayesian.

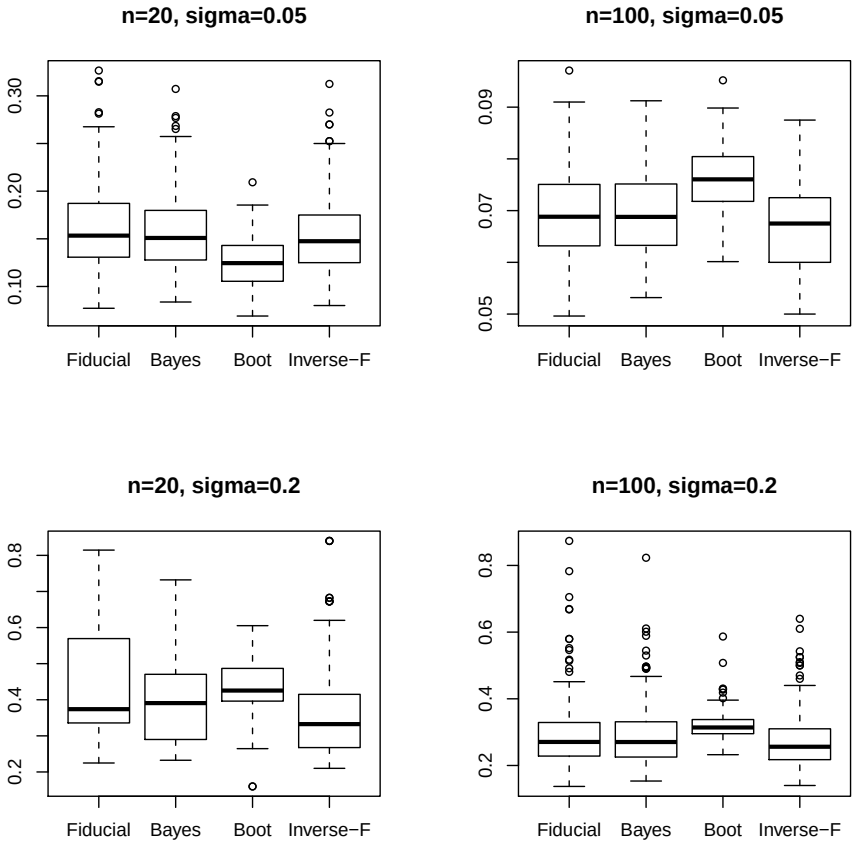

Figure 5.2.5: Confidence interval lengths for sigmoid with center knot point.

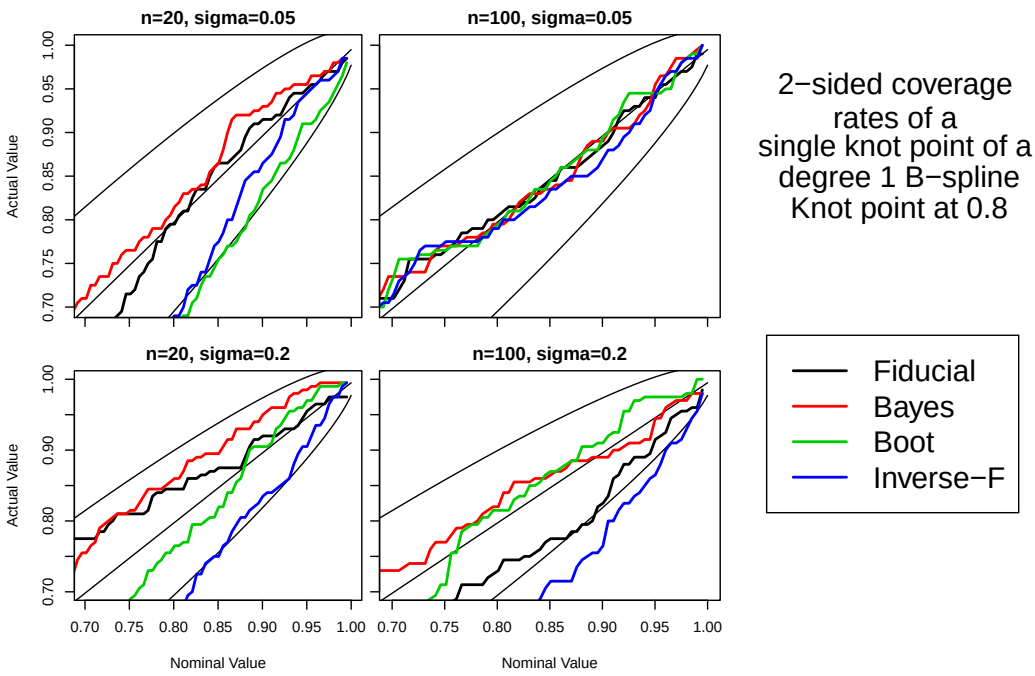

Figure 5.2.6: Coverage plots for piecewise linear, edge knot point simulation

5.2.2 Edge Knot Point

The next set of simulations is where the knot point is not in the center of the data but rather off to one edge. This is a harder problem because of unequal sample sizes before and after the knot point.

The bootstrap method performed quiet well considering how poorly it performed in the centered knot case. The Bayesian method consistently had conservative coverage rates, but had larger interval lengths than the fiducial in the degree 2 case.

The fiducial method had good coverage rates in all but the large sample, large variance edge knot case. The Bayesian method consistently had coverage rates were either indistinguishable from the nominal value or was at least no worse than the other methods. The Bayesian and fiducial methods generally had similar coverage rates.

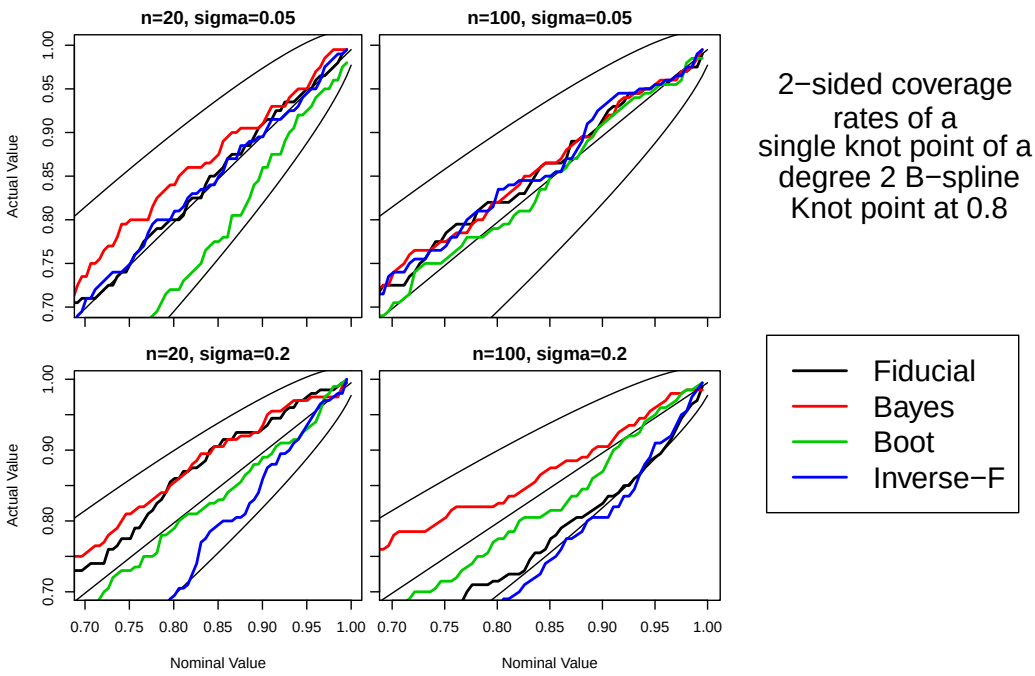

Figure 5.2.7: Coverage plots for the sigmoid curve with edge knot point.

5.3 Conclusions

The simulation study shows that the Bayesian method is superior to the bootstrap and inverse-F test. The bootstrap and inverse-F test both had instances where the observed converge rates were significantly liberal. When the two methods were not liberal, the interval lengths were generally comparable to the Bayesian and fiducial methods. The fiducial method had liberal coverage rates in the large variance large sample edge knot case one and the reason is unclear. Further research into the selection of data points for the Jacobian calculation will hopefully reveal the problem.

While the Bayesian and fiducial methods do have a significant computation overhead compared to the inverse-F test, the more accurate coverage rates are sufficient to justify using their use in free-knot spline problems.

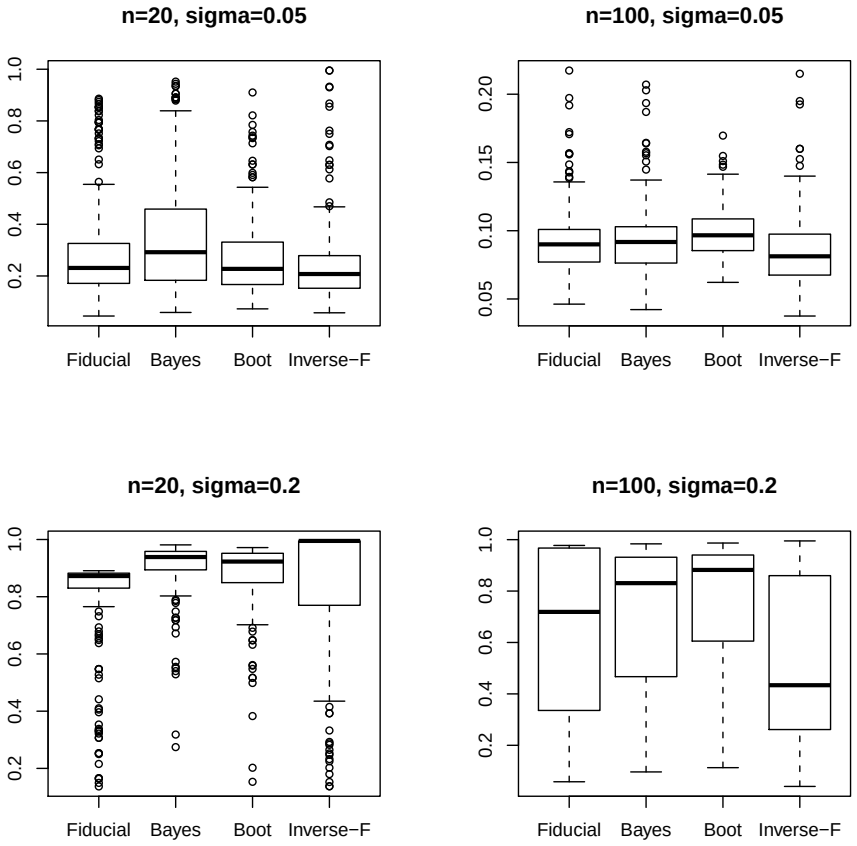

Figure 5.2.8: Confidence interval lengths for piecewise-linear with edge knot point.

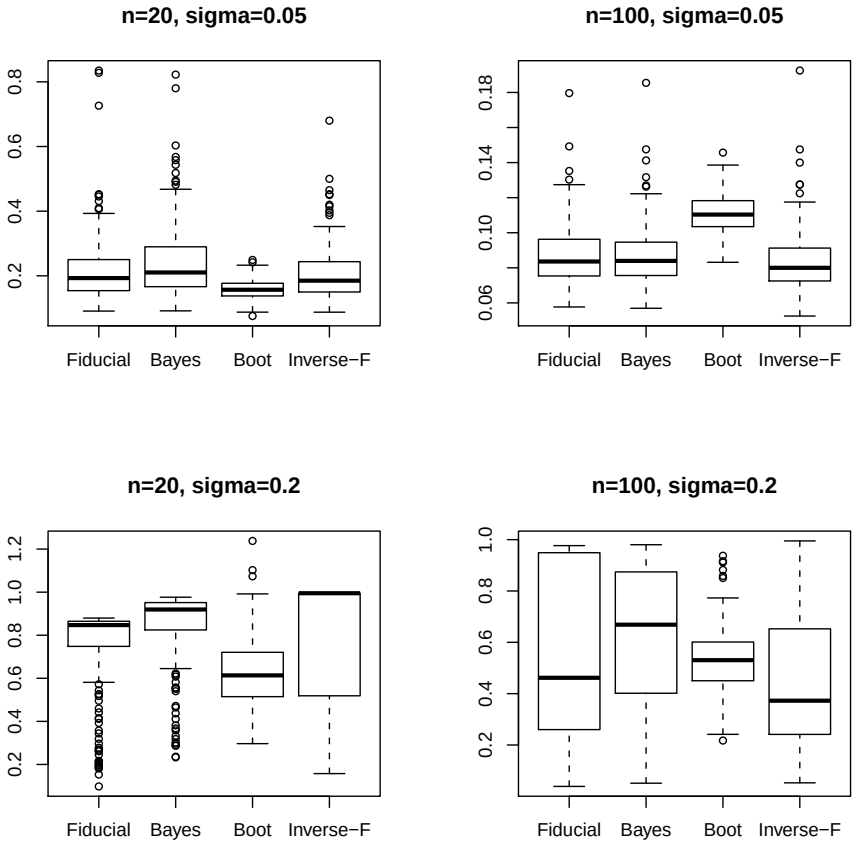

Figure 5.2.9: Confidence interval lengths for sigmoid with edge knot point.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Akademiai Kiado, Budapest.

Barrowman, N. J. and Myers, R. A. (2000). Still more spawner-recruitment curves; the hockey stick and its generalizations. *Candian Journal of Fisheries and Aquatic Sciences*, 57:665–676.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag, New York, 2 edition.

Cai, Z., Naik, P. A., and Tsai, C. L. (2000). De-noised least squares estimators: an application to estimating advertisment effectiveness. *Statistica Sinica*, 10:1231–1243.

Carbone, C., Teacher, A., and Rowcliffe, J. M. (2007). The costs of carnivory. *PLOS Biology*, 5(2).

Carpenter, S. R. (2001). Alternative states of ecosystems: evidence and its implications for environmental decisions. In Press, M. C., Huntley, N. J., and Levin, S., editors, *Ecology: Achievement and Challenge*, chapter 17, pages 357–386. Cambridge University Press.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman & Hall/CRC.

Chaudhuri, P. and Marron, J. S. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823.

Cheng, C.-L. and van Ness, J. W. (1999). *Statistical Regression with Measurement Error*. Oxford University Press Inc., New York.

Chiu, G. S., Lockhart, R., and Routledge, R. (2006). Bent-cable regression theory and applications. *Journal of the American Statistical Association*, 101(474):542–553.

Clements, W., Carlisle, D., Lazorchak, J., and Johnson, P. (2000). Heavy metals structure benthic communities in Colorado mountain streams. *Ecological Applications*, 10(2):626–638.

Clements, W. H. (2004a). Small-scale experiments support causal relationships between metal contamination and macroinvertebrate community responses. *Ecological Applications*, 14:954–967.

Clements, W. H. (2004b). Unpublished observations.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Connell, J. H. and Sousa, W. P. (1983). On the evidence needed to judge ecological stability or persistence. *The American Naturalist*, 121(6):789–824.

Cui, H., He, X., and Zhu, L. (2002). On regression estimators with de-noised variables. *Statista Sinica*, 12(4):1191–1205.

de Boor, C. (1972). On calculating with b-splines. *Journal of Approximation Theory*, 6:50–62.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071.

E, L., Hannig, J., and Iyer, H. (2008). Fiducial intervals for variance components in an un- balanced two-component normal mixed linear model. *Journal of the American Statistical Association*, 103:854–865.

Estes, J. A. and Duggins, D. O. (1995). Sea otters and kelp forests in alaska: generality and variation in a community ecological paradigm. *Ecological Monographs*, 65(1):75–100.

Ewel, K. C. (2001). Natural resource management: the need for interdisciplinary collaboration. *Ecosystems*, 4(8):716–722.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting - variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society Series B-Methodological*, 57(2):371–394.

Fan, J. and Gijbels, I. (1996). *Local polynomial modeling and its applications*. Chapman & Hall, London.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, xxvi:528–535.

Folke, C., Carpenter, S., Elmqvist, T., Gunderson, L. H., Holling, C., and Walker, B. (2002). Resilience and sustainable developement: building adaptive capacity in a world of transformations. *AMBIO*, 31(5):437–440.

Fuller, W. (1987). *Measurement Error Models*. John Wiley and Sons, Inc.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Spring-Verlag.

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2002). Significance in scale space for bivariate density estimation. *Journal of Computional and Graphical Statistics*, 11(1):1–21.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall.

Groffman, P. M., Baron, J. S., Blett, T., Gold, A. J., Goodman, I., Gunderson, L. H., Levinson, B. M., Palmer, M. A., Paerl, H. W., Peterson, G. D., Poff, N. L., Rejeski, D. W., Reynolds, J. F., Turner, M. G., Weathers, K. C., and Wiens, J. (2006). Ecological thresholds: the key to successful environmental management or an important concept with no practical application? *Ecosystems*, 9(1):1–13.

Hall, P. and Opsomer, J. D. (2005). Theory for penelized spline regression. *Biometrika*, 92(1):105–118.

Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19:491–544.

Hannig, J., E, L., Abdel-Karim, E., and Iyer, H. (2006a). Simultaneous fiducial generalized confidence intervals for ratios of means of lognormal distributions. *Statistics*, 35(2):261–269.

Hannig, J., Iyer, H., and Patterson, P. (2006b). Fiducial generalized confidence intervals. *Journal of the American Statistical Association*, 101(473):254–269.

Hannig, J. and Lee, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika*.

Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for sizer. *Journal of the American Statistical Association*, 101(474):484–499.

Hengartner, N. W., Wegkamp, M. W., and Matzer-Lober, E. (2002). Bandwidth selection for local linear regression smoothers. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64:791–804.

Holling, H. (1986). A classification of trivariate regression-models. *Biometrical Journal*, 28(7):783–790.

Homan, R. N., Windmiller, B. S., and Reed, J. M. (2004). Critical thresholds associated with habitat loss for two vernal pool-breeding amphibians. *Ecological Applications*, 14(5):1547–1553.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Jupp, D. L. B. (1978). Approximation to data by splines with free knots. *SIAM Journal on Numerical Analysis*, 15(2):328–343.

Kinupp, V. F. and Magnusson, W. E. (2005). Spatial patterns in the understory shrub genus psychotria in central amazonia: effects of distance and topography. *Journal of Tropical Ecology*, 21(4):363–374.

Knowlton, N. (1992). Thresholds and multiple stable states in coral-reef community dynamics. *American Zoologist*, 32(6):674–682.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied linear statistical models*. McGraw-Hill/Irwin, New York, NY, USA, 5 edition.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer, New York.

Loader, C. (1999). *Local regression and likelihood*. Springer-Verlag, New York.

Mao, W. and Zhao, L. H. (2003). Free-knot polynomial splines with confidence intervals. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 65(4):901–919.

May, R. M. (1977). Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature*, 269(5628):471–477.

Muradian, R. (2001). Ecological thresholds: a survey. *Ecological Economics*, 38(1):7–24.

Nimick, D. A., Gammons, C. H., Cleasby, T. E., Madison, J. P., Skaar, D., and Brick, C. M. (2003). Diel cycles in dissolved metal concentrations in streams: Occurrence and possible causes. *Water Resourses Research*, 39(9).

Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Annals of Statistics*, 28(1):298–335.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press, New York, New York.

Scheffer, M. and Carpenter, S. R. (2003). Catastrophic regime shifts in ecosystems: linking theory to observation. *Trends in Ecology & Evolution*, 18(12):648–656.

Scheffer, M., Straile, D., van Nes, E. H., and Hosper, H. (2001). Climatic warming causes regime shifts in lake food webs. *Limnology and Oceanography*, 46(7):1780–1783.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, New York, NY, USA, third edition.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear regression*. John Wiley and Sons, New York, New York, USA.

Sonderegger, D. L., Wang, H., Clements, W. H., and Noon, B. R. (2008). Using sizer to detect thresholds in ecological data. *Frontiers in Ecology and the Environment*.

Sonderegger, D. L., Wang, H., Huang, Y., and Clements, W. H. (2009). Effects of measurement error on the strength of concentration-response relationships in aquatic toxicology. *Ecotoxicology*, 18(7):824–828.

Stone, C. J. and Huang, J. Z. (2002). Free knot splines in concave extended linear modeling. *Journal of Statistical Planning and Inference*, 108:219–253.

Toms, J. D. and Lesperance, M. L. (2003). Piecewise regression: a tool for identifying ecological thresholds. *Ecology*, 84(8):2034–2041.

van Nes, E. H., Scheffer, M., van den Berg, M. S., and Coops, H. (2002). Dominance of charophytes in eutrophic shallow lakes-when should we expect it to be an alternative stable state? *Aquatic Botany*, 72(3–4):275–296.

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische Mathematik*, 24(5):383–393.

Wallstrom, G., Liebner, J., and Kass, R. E. (2007). An implementation of bayesian adaptive regression splines (bars) in c with s and r wrappers. *Journal of Statistical Software*, 26(1):1–21.

Walsh, C. J., Fletcher, T. D., and Ladson, A. R. (2005). Stream restoration in urban catchments through redesigning stormwater systems: looking to the catchment to save the stream. *Journal of the North American Benthological Society*, 24(3):690–705.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, 88(423):899–905.

Yao, H. (2008). Using measurement error model to assess the effects of metal pollution to species in arkansas river. Master's thesis, Colorado State University.

Yuan, L. L. (2007). Effects of measurement error on inferences of environmental conditions. *Journal of the North American Benthological Society*, 26(1):152–163.