Rank-Based Estimation for Generalized Additive Models

by

Hannah Correia

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 6, 2016

Keywords: generalized additive models, smoothing spline, robust estimation

Approved by

Asheber Abebe, Professor of Mathematics and Statistics
F. Stephen Dobson, Professor of Biological Sciences
Bertram Zinner, Associate Professor of Mathematics and Statistics

Abstract

This dissertation focuses on improvement of generalized additive models (GAMs) using rank estimators. First, we introduce estimation of the smoothing functions in GAMs via backfitting in a local scoring algorithm using maximization of the expected log likelihood function with weights. Improvements of GAM estimation have focused on the smoothers used in the local scoring algorithm, but poor prediction for non-Gaussian data motivates the need for robust estimation of GAMs. Rank-based estimation as a robust and efficient alternative to the likelihood-based estimator of GAMs is proposed, and it is shown that rank GAM estimators can be restructured as iteratively reweighted GAM estimators. Simulations further support the use of rank-based GAM estimation for heavy-tailed or contaminated sources of data common in climate studies. Successful application of rank GAM estimation is employed for fisheries data, a field which commonly uses GAMs for their high degree of flexibility in modeling complex systems and could benefit from improved model prediction performance for non-Gaussian data. Cross-validation shows improved prediction performance for rank GAMs over GAMs, and improved adjusted R^2 values highlight the better fit of rank GAMs for the given data.

Acknowledgments

Put text of the acknowledgments here.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to Generalized Additive Models

The generalized linear model (GLM) (McCullagh and Nelder, 1989) is a popular method for modeling the mean of data via a link function $g(\mu) = \eta$ and takes the form

$$\eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{1.1}$$

where $X_1, \ldots, X_p$ are covariates contributing to $\mu = E(Y)$ and Y has an exponential family distribution. The GLM is not always appropriate for given data as it restricts the model to a linear relationship between $g(\mu)$ and the covariates. The generalized additive model (GAM) (Hastie and Tibshirani, 1990) is a more flexible extension of the generalized linear model, with $\beta_0, \beta_1, \ldots, \beta_p$ replaced by nonparametric smooth functions $f_j(\cdot)$ such that

$$\eta = f_0 + \sum_{j=1}^{p} f_j(X_j) \tag{1.2}$$

1.1 GAM Estimation

Similar to GLM, where estimation of the coefficient vector $\boldsymbol{\beta}$ is informative, estimation of the smooth functions $f_1(\cdot), \ldots, f_p(\cdot)$ in GAM is required to effectively estimate the model. This was achieved via local scoring by Hastie and Tibshirani (1990) in which the smoothers are estimated individually using a local scoring algorithm.

Assume Y has a density $f_Y(y, \theta, \phi)$ belonging to the exponential family with θ location parameter and ϕ scale parameter

$$f_Y(y, \theta, \phi) = \exp\left\{\frac{y\theta - a(\theta)}{b(\phi)} + c(y - \phi)\right\}$$

In GLM of the form given in (1.1) where $g(\mu) = \eta$, $E(Y|\mathbf{X}) = \mu$ is related to θ by $\mu = a'(\theta)$. Since estimation of μ is of primary interest, an interative Fisher scoring procedure can be used to find the maximum likelihood estimate of $\hat{\boldsymbol{\beta}}$. For a given $\hat{\eta}$ with fitted $\hat{\mu}$, form an adjusted variable

$$z = \hat{\eta} + (y - \hat{\mu}) \left(\frac{d\eta}{d\mu} \right) \tag{1.3}$$

Define the weights as

$$W = \left(\frac{d\mu}{d\eta} \right)^2 V^{-1} \tag{1.4}$$

where V is variance of Y at $\mu = \hat{\mu}$. The estimate $\hat{\boldsymbol{\beta}}$ is derived by regressing z on $1, x_1, \ldots, x_p$ using the given weights. Then a new $\hat{\eta}$, $\hat{\mu}$, and new adjusted variable z can be calculated, and a new $\hat{\boldsymbol{\beta}}$ is computed with weights W. This weighted least squares process is iterated until the deviance

$$D(y, \hat{\mu}) = 2[\log L(y) - \log L(\hat{\mu})]$$

is sufficiently small. This iterative process was shown to be identical to the Fisher scoring procedure by (Nelder and Wedderburn, 1972).

The scoring procedure for additive models follows similarly from the iterative Fisher scoring procedure, where the smooth functions $f_j(\cdot)$ are estimated. Assume the same conditions as given above for the GLM case, except η is as given in (1.2). To estimate $f(\cdot)$, we must smooth the adjusted variable z on X for each $\{f_j(\cdot)\}_{j=1}^p$. This can be achieved through a variety of scatterplot smoothers including running lines smoothers, kernel estimation, or smoothing splines (Wahba, 1990).

The local scoring procedure is a generalized technique for estimating smooth functions using maximization of the expected log likelihood. Consider (1.2) in the case of $p = 1$, where Y_is are independent and Y has a distribution belonging to the exponential family. Say the conditional density of Y given $X = x$ is $h(y, \eta(x))$. Then $\hat{\eta}(\cdot)$ is chosen to maximize the expected log likelihood function $E\left[l(\eta(X), Y)\right]$. If $\eta(x)$ is a nonparametric function, then differentiating the expected log likelihood with respect to η yields $E\left[dl/d\eta|x)_{\hat{\eta}(x)}\right] = 0$

assuming interchangeability of expectation and differentiation. For an initial $\eta^0(x)$, we can derive an improved estimate of η using Taylor series expansion:

$$\eta^1(x) = \eta^0(x) - \frac{E(dl/d\eta|x)}{E(d^2l/d\eta^2|x)} = E\left[\eta^0(x) - \frac{\partial l/\partial\eta}{E\left[\partial^2 l/\partial\eta^2|x\right]}\middle| x\right] \tag{1.5}$$

which becomes the local scoring update for backfitting in the local scoring procedure. Since we are considering the exponential family of densities, we can simplify (1.5) by calculating $dl/d\eta = (y-\mu)V^{-1}(d\mu/d\eta)$ and $E(d^2l/d\eta^2|x) = -(d\mu/d\eta)^2V^{-1}$ to get

$$\eta^1(x) = E[\eta^0(x) + (Y-\mu)(d\eta/d\mu)|x]$$

which can be thought of as a smooth of the adjusted variable (1.3)

$$\eta^1(x) = \boldsymbol{\delta}[\eta^0(x) + (y-\mu)(d\eta/d\mu)] \tag{1.6}$$

with the smoother $\boldsymbol{\delta}$ and weights $(d\mu/d\eta)^2V^{-1}$.

For the case of $p > 1$ of (1.2) where $E[Y|\mathbf{X}] = g(\mu) = \eta$ and $E[f_j(X_j)] = 0$ for all j, the exponential family local scoring update is

$$\eta^1(x) = E[\eta^0(\mathbf{x}) + (Y-\mu)(d\eta/d\mu)|\mathbf{x}] \tag{1.7}$$

Each $f_j(\cdot)$ can be estimated iteratively using framework of the backfitting algorithm by Friedman and Stuetzle (1981) to give the general local scoring algorithm proposed by Hastie and Tibshirani (1990):

Step $m = 0$: $f_0 = g(E(y))$, $f_1^0(\cdot) = f_2^0(\cdot) = \ldots = f_p^0(\cdot) = 0$

Step $m = m + 1$:

i. Use local scoring update given by (1.7) to create adjusted variable $Z = \eta^{m-1} + (Y - \mu^{m-1})(\partial\eta/\partial\mu^{m-1})$ where $\eta^{m-1} = f_0 + \sum_{j=1}^p f_j^{m-1}(X_j)$

ii. Define weights, $W = (\partial\eta/\partial\eta^{m-1})^2 V^{-1}$

iii. For $j = 1$ to p: Fit additive model to adjusted variable using weights W via backfitting to obtain $\hat{f}_j^m(\cdot)$ and model η^m.

Until: $E(D(Y, \mu^m)$ no longer decreases.

A major focus of improvements to GAM estimation has been on the scatterplot smoothers in (1.6). Hastie and Tibshirani (1990) focused on running line smoothers.

(Wahba, 1990) smoothing splines, elegant but computationally expensive

Penalized least squares for estimating f_0 (? Wood, 2002, "GAMs with integrated model selection using penalized regression splines and applications to environmental modelling")

1.2 Contribution

The first part of this dissertation is concerned with improving the estimation of GAMs using rank estimation. It will be shown that the proposed rank GAM estimation method produces better fits over a range of data types with some marginal tradeoff in computational expense.

The second part is an application of the rank GAM estimation method used to improve GAM fit on fisheries data with two- and three-dimensional covariates. It will be shown that the rank GAM method produces better smoothing estimates and therefore better fits for several types of models through higher adjusted R^2 values. Cross-validation errors are reduced for the rank GAM method when compared to GAM estimation, showing improvement of prediction capability when using rank GAM estimation.

Chapter 2

Rank-Based GAMs

2.1 Introduction

Classical GAM estimation involves penalized likelihood or penalized least squares (Wood, 2006). It is well-known that such approaches are sensitive to outliers and departures from underlying distributions (eg. normal errors). To accommodate a wide range of data generating phenomena, robust approaches have been proposed recently. These include Alimadad and Salibian-Barrera (2011), Croux et al. (2012), and Wong et al. (2014). These are all M-type estimators. We are interested in R-estimators as defined in Jaeckel (1972) and Hettmansperger and McKean (2011). As discussed in Draper (1988), both M and R estimators provide robust fits with no clear winner. While both M and R are location invariant, only R-estimators are scale invariant. This makes them very attractive for estimation in complex model settings. A drawback of both M and R estimation is that they are computationally expensive. For the linear regression model, Sievers and Abebe (2004) gave an approach that uses iterative least squares fitting to obtain R-estimators. Recently, Miakonkana and Abebe (2014) extended this to generalized linear models. Wong et al. (2014) derived a computationally efficient M-estimator of the GAM model, again using iterative fitting of GAMs via penalized least squares. In this chapter, we propose a rank estimator of GAMs and develop an efficient iterative computational algorithm. Our method, which we call the *iterated regularized rank quasi-likelihood (IRRQL)* procedure, depends on ranking of Pearson residuals to account for the mean-variance dependence in GAMs.

2.2 Rank-Based Estimation

Suppose we have a linear regression model $Y_i = \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n$. The vector of model residuals is given by $\mathbf{z}(\boldsymbol{\beta})$ with ith component $z_i(\boldsymbol{\beta}) = Y_i - \mathbf{x}_i^T\boldsymbol{\beta}$. Jaeckel (1972) proposed to estimate the regression slope parameter $\boldsymbol{\beta}$ by minimizing

$$\|\mathbf{z}(\boldsymbol{\beta})\|_\varphi \equiv \sum_{i=1}^{n} \varphi\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right) z_i(\boldsymbol{\beta}) \,, \tag{2.1}$$

where $\phi : (0,1) \to \mathbb{R}$ is a nondecreasing score function such that $\int \phi = 0$ and $\int \phi^2 = 1$ and $R(\cdot)$ is the rank function. He showed that this produces a regression estimator that is equivalent to the rank score estimator given by Jurečková (1971). He also showed that the quantity $\|\cdot\|_\varphi$ is a convex pseudo-norm on $\mathbb{R}^n$. Because $\|\cdot\|_\varphi$ is a pseudo-norm, it is invariant to constant translations; hence, it cannot be used to estimate the intercept α. To see this, consider the simplest case of the linear score function $\phi(u) = \sqrt{12}(u - 1/2)$ resulting in the so-called Wilcoxon pseudo-norm. In this case, it is easy to observe that minimizing (2.1) is equivalent to minimizing

$$\sum_{i<j} |z_i(\boldsymbol{\beta}) - z_j(\boldsymbol{\beta})| \,.$$

For this and other general discussion regarding the use of (2.1) in the linear model, one is referred to the monograph Hettmansperger and McKean (2011).

For the linear regression, it has been shown that the estimator resulting from the Wilcoxon estimation is robust in the presence of outliers and heavy tailed error distributions. It is also very efficient. For instance, it achieves 95.5% relative efficiency versus the least squares estimator when the underlying error distribution is normal and the relative efficiency is much higher for distributions with tails heavier than the normal. The worst case relative efficiency is 86.4% for symmetric error distributions. So, there is much appeal to using $\|\cdot\|_\varphi$ for estimation purposes. The inference also extends to hypothesis testing

(Hettmansperger and McKean, 2011). For example, we can easily define drop, Wald, or score tests for testing significance of model parameters.

In recent years, the method has been employed for models other than the linear model. Bindele and Abebe (2015) studied rank estimation of semiparametric models with responses missing at random. They showed that the rank estimator remains robust and efficient with efficiency improving relative to standard imputation methods when a large proportion of the responses are missing. The Wilcoxon (and its weighted versions) have been used for estimation of general nonlinear regression (Abebe and McKean, 2013), generalized linear models (Miakonkana and Abebe, 2014), varying coefficient models (Wang et al., 2009), and functional regression (Denhere and Bindele, 2016) among others.

A parallel development involves the signed-rank method. This is essentially a rank weighted L_1 norm and thus can provide estimate of the intercept. The estimates of parameters found using the signed-rank method are the same as those using the rank method. However, distributional results for the signed-rank method require the symmetry of the error distribution, reducing its appeal. Nevertheless, the method has been successfully used for obtaining robust and efficient estimates for linear models (Hössjer, 1994), nonlinear models (Bindele and Abebe, 2012), and nonlinear models with multidimensionally indexed parameters (Nguelifack et al., 2015).

Some of the development has been facilitated by the iterative reweighted least squares procedure given by Sievers and Abebe (2004). This greatly simplifies the computation of rank regression coefficients even for complex models (Abebe et al., 2016).

2.3 Rank-Based GAM Estimation

For obtaining the rank estimator of GAMs, we will use a penalized version of the rank quasi-score function given in Miakonkana and Abebe (2014). The responses $\{Y_i\}_{i=1}^n$ are assumed to be independent and follow a distribution from the exponential family with expectation μ_i and variance $V(\mu_i)$. To simplify our discussion and theoretical development, we

will consider the simple $p = 1$ version of the GAM model (1.2) given by

$$g(\mu_i) = f(x_i)$$

as well as the linear (Wilcoxon) score function $\sqrt{12}(u - .5)$.

Taking a set of prespecified basis functions $\mathbf{b} = (b_1(\cdot), \ldots, b_m(\cdot))'$, the function f is assumed to have a representation

$$f(x_i; \boldsymbol{\theta}) = \sum_{j=1}^{m} b_j(x_i)\theta_j \equiv \mathbf{b}_i^T \boldsymbol{\theta} \tag{2.2}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ is a vector of basis coefficients and we suppress x_i in $\mathbf{b}$.

Ignoring the extra scale parameter, we define the Pearson residuals as

$$z_i(\boldsymbol{\theta}) = \frac{Y_i - \mu_i}{\sqrt{V(\mu_i)}} \, .$$

The rank quasi-likelihood function is then

$$\boldsymbol{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ \frac{R(z_i(\boldsymbol{\theta}))}{n+1} - \frac{1}{2} \right\} \frac{\partial \mu_i / \partial \boldsymbol{\theta}}{\sqrt{V(\mu_i)}} \, . \tag{2.3}$$

By taking $h \equiv g^{-1}$, we have $\mu_i = h(\mathbf{b}_i^T \boldsymbol{\theta})$ and $\partial \mu_i / \partial \boldsymbol{\theta} = h'(\mathbf{b}_i^T \boldsymbol{\theta}) \mathbf{b}(x_i)$ and

$$\boldsymbol{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ \frac{R(z_i(\boldsymbol{\theta}))}{n+1} - \frac{1}{2} \right\} \frac{h'(\mathbf{b}_i^T \boldsymbol{\theta}) \mathbf{b}_i}{\sqrt{V(h(\mathbf{b}_i^T \boldsymbol{\theta}))}} \, .$$

Theoretically, the rank estimator of $\boldsymbol{\theta}$ is found by solving $\boldsymbol{\ell}(\boldsymbol{\theta}) = \mathbf{0}$. However, for the estimation of GAMs, we will need to impose a smoothness penalty. Thus we define the *regularized rank quasi-likelihood (RRQL)* function and solve

$$\boldsymbol{\ell}_\lambda(\boldsymbol{\theta}) \equiv \boldsymbol{\ell}(\boldsymbol{\theta}) + 2\mathbf{S}_{\lambda_n} \boldsymbol{\theta} = \mathbf{0} \, , \tag{2.4}$$

where $\mathbf{S}_{\lambda_n} = \lambda_n \mathbf{D}$, $\lambda_n > 0$ is a smoothing parameter and $\mathbf{D}$ is an $m \times m$ penalty matrix. We let $\widetilde{\boldsymbol{\theta}}_n$ represent the zero of the RRQL function; that is $\widetilde{\boldsymbol{\theta}}_n$ solves $\boldsymbol{\ell}_{\lambda_n}(\boldsymbol{\theta}) = \mathbf{0}$.

However, finding a direct solution of $\boldsymbol{\ell}_{\lambda}(\boldsymbol{\theta}) = \mathbf{0}$ is difficult. Below we will define an iterative scheme to approximate $\widetilde{\boldsymbol{\theta}}_n$. To that end, define the pseudo-Pearson 'residuals'

$$z_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \frac{Y_i - h(\mathbf{b}_i^T \boldsymbol{\theta})}{\sqrt{V(h(\mathbf{b}_i^T \boldsymbol{\theta}^*))}}$$

and define the corresponding rank estimator as the minimizer of $\|\mathbf{z}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\|_w$, where

$$\|\mathbf{z}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\|_w = \sum_{i=1}^{n} \left\{ \frac{R(z_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*))}{n+1} - \frac{1}{2} \right\} z_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \ .$$

Using the IRLS scheme of Sievers and Abebe (2004), this can be represented as

$$\|\mathbf{z}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\|_w = \sum_{i=1}^{n} w_i(\boldsymbol{\theta}) \frac{(Y_i - h(\mathbf{b}_i^T \boldsymbol{\theta}))^2}{V(h(\mathbf{b}_i^T \boldsymbol{\theta}^*)}$$

where, letting $m = \text{med}\{Y_i - h(\mathbf{b}_i^T \boldsymbol{\theta})\}$, the weights are defines as

$$w_i(\boldsymbol{\theta}) = \begin{cases} \frac{\frac{R(z_i(\boldsymbol{\theta}, \boldsymbol{\theta}^*))}{n+1} - \frac{1}{2}}{Y_i - h(\mathbf{b}_i^T \boldsymbol{\theta}) - m} & \text{if } Y_i - h(\mathbf{b}_i^T \boldsymbol{\theta}) - m \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Taking the weights w_i at a different value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}'$, and taking the derivative of $\|\cdot\|_w$ with respect to $\boldsymbol{\theta}$ we obtain the approximate rank score function

$$2 \sum_{i=1}^{n} w_i(\boldsymbol{\theta}') \frac{(Y_i - h(\mathbf{b}_i^T \boldsymbol{\theta}))}{V(h(\mathbf{b}_i^T \boldsymbol{\theta}^*)} h'(\mathbf{b}_i^T \boldsymbol{\theta}) \mathbf{b}_i$$

Following Wedderburn (1974), if we now take $\boldsymbol{\theta}^* = \boldsymbol{\theta}$, then we get the weighted quasi-likelihood function

$$\boldsymbol{\ell}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{i=1}^{n} w_i(\boldsymbol{\theta}') \left\{ \frac{Y_i - h(\mathbf{b}_i^T \boldsymbol{\theta})}{V(h(\mathbf{b}_i^T \boldsymbol{\theta}))} \right\} h'(\mathbf{b}_i^T \boldsymbol{\theta}) \mathbf{b}_i = \sum_{i=1}^{n} w_i(\boldsymbol{\theta}') \frac{(Y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\theta}}$$

which is exactly a weighted form of the classical GLM quasi-likelihood function.

We can now define the penalized quasi-likelihood as

$$\boldsymbol{\ell}_{\lambda_n}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \boldsymbol{\ell}(\boldsymbol{\theta}, \boldsymbol{\theta}') + \mathbf{S}_{\lambda_n} \boldsymbol{\theta} \ .$$

Now, suppose we have a suitable initial estimator of $\boldsymbol{\theta}$, say $\widehat{\boldsymbol{\theta}}_n^{(0)}$. This can be the classical penalized likelihood estimator. For $k = 1, 2, \ldots$, we define $\widehat{\boldsymbol{\theta}}_n^{(k)}$ as a solution of the *iterated regularized rank quasi-likelihood (IRRQL)* function

$$\boldsymbol{\ell}_{\lambda_n}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{(k-1)}) = \mathbf{0} \ ,$$

which can be computed by iteratively solving a weighted GAM estimating equation. For the unpenalized version $\boldsymbol{\ell}_0(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{(k-1)})$, Miakonkana and Abebe (2014) showed that the iteration converges to the solution of $\boldsymbol{\ell}_0(\boldsymbol{\theta})$.

Let $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$ and the $n \times m$ coefficient matrix be

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_n^T \end{bmatrix} \ .$$

Note that we have two rank estimators of $\boldsymbol{\theta}$: $\widetilde{\boldsymbol{\theta}}_n$ which solves the RRQL and $\widehat{\boldsymbol{\theta}}_n^{(k)}$, $k = 1, 2, \ldots$ given $\widehat{\boldsymbol{\theta}}_n^{(0)}$ which solves the IRRQL. We can correspondingly define two rank-based GAM

estimators of $\mathbf{f}$ using equation (2.2). We define these as

$$\widetilde{\mathbf{f}}_n = \mathbf{B}\widetilde{\boldsymbol{\theta}}_n \tag{2.5}$$

and

$$\widehat{\mathbf{f}}_n^{(k)} = \mathbf{B}\widehat{\boldsymbol{\theta}}_n^{(k)} \ , \quad k = 1, 2, \dots \ . \tag{2.6}$$

Note that for a given k, $\widehat{\mathbf{f}}_n^{(k)}$ is just a regular weighted GAM estimator. Thus, its asymptotic properties are well understood and are part of the standard GAM literature (cf. Hastie and Tibshirani, 1990; Wood, 2006). Theorem 2.1 below gives consistency of $\widehat{\mathbf{f}}_n^{(k)}$. However, we need to understand whether $\widehat{\mathbf{f}}_n^{(k)}$ provides a good approximation of the 'true' rank estimator $\widetilde{\mathbf{f}}_n$. Below we give conditions under which $\widehat{\mathbf{f}}_n^{(k)}$ gives a valid approximation of $\widetilde{\mathbf{f}}_n$. The theorem following the conditions gives the asymptotic equivalence of $\widehat{\mathbf{f}}_n^{(k)}$ and $\widetilde{\mathbf{f}}_n$. Before giving the conditions, we note that when using the iteratively reweighted least squares (IRLS) approach of fitting GAMs, there is a reproducing kernel Hilbert space (RKHS) representation $\mathbf{f}^T\Gamma^{1/2}\mathbf{R}\Gamma^{1/2}\mathbf{f}$ of the penalty function $\lambda_n\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}$, where Γ is the IRLS weight (Wong et al., 2014). In this set up, the residual smoother matrix for GAM estimation is $\mathbf{H}_{\lambda_n} = (\mathbf{I} + 2\lambda_n\mathbf{R})^{-1}$, where $\mathbf{R}$ is the reproducing kernel.

We assume the following conditions hold:

(A1) The function f is bounded; that is, $\sup_{-\infty<t<\infty} |f(t)| < \infty$.

(A2) Let $\mathscr{F}$ be the space of all f's. We assume that $\mathscr{F}$ is a reproducing kernel Hilbert space.

(A3) Let $\mathscr{C}_\alpha = \{f \in \mathscr{F} : \|f\|_{\mathscr{F}} \leq \alpha\}$ for some constant α. We assume that $\mathscr{C}_\alpha$ is compact with respect to L_2 norm.

(A4) Let d_n be the maximum diagonal element of $\mathbf{H}_{\lambda_n}$. We assume that $\lambda_n/n \to 0$ and $d_n \to 0$ as $n \to \infty$. Moreover, $tr(\mathbf{H}_{\lambda_n})/\lambda_n < K < \infty$.

Theorem 2.1 *Under (A1) – (A4), for* $k \in \mathbb{N}$,

$$n^{-1}E\{\|\widehat{\mathbf{f}}_n^{(k)} - \mathbf{f}\|^2\} \ \rightarrow \ 0$$

as $n \rightarrow \infty$

Theorem 2.2 *Under (A1) – (A4), for* $k \in \mathbb{N}$,

$$\frac{\|\widehat{\mathbf{f}}_n^{(k)} - \widetilde{\mathbf{f}}_n\|}{E\{\|\widehat{\mathbf{f}}_n^{(k)} - \mathbf{f}\|\}} \xrightarrow{\mathcal{P}} 0$$

as $n \rightarrow \infty$

The proof of Theorem 2.1 can be constructed in a straightforward manner following Hastie and Tibshirani (1990) with residuals $\mathbf{z}$ replaced by $\mathbf{W}^{1/2}\mathbf{z}$, where $\mathbf{W} = diag(w_1, \ldots, w_n)$. The proof of Theorem 2.2 is similar to the one given by Wong et al. (2014) for M-estimators.

There are certain practical considerations that need attention. The first is the degrees of freedom of the estimation problem. The RKHS literature defines the effective degrees of freedom as $tr(\mathbf{H}_{\lambda_n})$. So, (A4) specifies a balance between the effective degrees of freedom and the smoothing parameter. We still need a way to estimate the smoothing parameter λ_n. In this thesis, we employ generalized cross-validation to select the parameter λ_n. This is the most common approach in the literature (Wood, 2006). Finally, one may question the value of fully iterating k. If the one step estimator gives us consistent estimators (Theorems 2.1 and 2.2), then why do we need to iterate more than once? This was answered in Sievers and Abebe (2004) and Miakonkana and Abebe (2014) where using fixed-point theory it was established that as $k \rightarrow \infty$ the IRLS rank estimator converges to the true rank estimator for finite samples. In our notation, this is

$$\lim_{k \rightarrow \infty} \widehat{\mathbf{f}}_n^{(k)} = \widetilde{\mathbf{f}}_n$$

for n fixed.

2.4 Simulations

The following simulations reflect how the proposed rank GAM estimators performs against GAM estimators and least absolute deviation (LAD) estimators in specific settings for finite samples. All simulations were performed on $n = 100$ samples and repeated for 1000 iterations. The simple model

$$y_j = f(x_j, z_j) + \varepsilon_i \ , \qquad j = 1, \ldots, n \ .$$

In the simulation, f is generated as

$$f(x_j, z_j) = (\pi^{s_x} s_z)(1.2) \exp\left(-\frac{(x_j - 0.2)^2}{s_x^2} - \frac{(z_j - 0.3)^2}{s_z^2}\right) + (0.8) \exp\left(-\frac{(x_j - 0.7)^2}{s_x^2} - \frac{(z_j - 0.8)^2}{s_z^2}\right)$$

where $s_x = 0.3$ and $s_z = 0.4$; x and z are 100 random deviates generated from a continuous uniform distribution with range $[0, 1]$. The correlation between two observations a distance r apart is $\exp(-(r/d)^2)$. The relative efficiencies of the rank GAM and LAD estimators as compared to GAM estimators were obtained by

$$\mathrm{RE}(\mathrm{R}, \mathrm{LS}) = \frac{\sum_{j=1}^n (f_j - \widehat{f}_{\mathrm{LS},j})^2}{\sum_{j=1}^n (f_j - \widehat{f}_{\mathrm{R},j})^2} \quad \text{and} \quad \mathrm{RE}(\mathrm{LAD}, \mathrm{LS}) = \frac{\sum_{j=1}^n (f_j - \widehat{f}_{\mathrm{LS},j})^2}{\sum_{j=1}^n (f_j - \widehat{f}_{\mathrm{LAD},j})^2} \ ,$$

respectively. Here with $\widehat{f}_{\mathrm{LS}}$ the function f estimated using the classical likelihood method; that is, penalized least squares. Similarly, $\widehat{f}_{\mathrm{R}}$ and $\widehat{f}_{\mathrm{LAD}}$ represent the fitted values using the rank GAM and LAD methods, respectively. The R^2 values of the GAM, rank GAM, and LAD models were calculated as a function of the correlations

$$R^2 = \rho^2(f, \widehat{f}_a) = \frac{\left[\sum_{j=1}^n (\widehat{f}_{a,j} - \overline{\widehat{f}}_a)(f_j - \bar{f})\right]^2}{\sum_{j=1}^n (\widehat{f}_{a,j} - \overline{\widehat{f}}_a)^2 \sum_{j=1}^n (f_j - \bar{f})^2}$$

with $\overline{\widehat{f}}_a$, a = LS, R, LAD, being the model predictions and f being the true function. We can obtaining both the LAD and R estimators using the weighted GAM approach where we use the score function $\phi = \text{sgn}(u)$ for LAD and $\phi = \sqrt{3}u$ for R estimators in the weight function.

The first simulation involved testing the performance of GAM estimators in the presence of heavy-tailed error distributions. To simulate this, the errors ε were randomly generated from Student's t distributions with increasing degrees of freedom e^k where k is taken from 1 to 5 in steps of .5. The correlation between two observations a distance r apart is $\exp(-(r/d)^2)$ with $d = 0.1$. The GAM, LAD model, and rank GAM were fit for the given data and estimates derived for each. The relative efficiencies and R^2 values were then calculated as described above and plotted for logarithms of the corresponding degrees of freedom, $\log(df)$ (Figure 2.1). The left panel of Figure 2.1 shows that both LAD and rank GAM estimators are more efficient than the GAM (LS) estimator when the error distribution is heavy-tailed. As expected the efficiency drops as the tails of the distribution approach the tails of the normal distribution. However, the loss in efficiency is less than 5% for the rank GAM method. This is in line with the theoretical asymptotic relative efficiency values for the Wilcoxon procedure. The LAD estimator is generally less efficient than the rank GAM estimator. The right panel of Figure 2.1 shows that all methods provide improved fit as the tails of the error distribution approach $N(0,1)$ tails with the rank GAM giving slightly better fit for heavy tails and GAM giving slightly better fit for tails approaching $N(0,1)$.

The second simulation tested GAM estimation performance when there are outliers in the measured response. To simulate this, we generated random errors ε drawn from a contaminated normal distribution. The contaminated normal distribution is defined by creating a normal-normal Huber contaminated distribution as

$$CN(\delta, \sigma) = (1 - \delta)N(0, 1) + \delta N(0, \sigma^2) \ ,$$

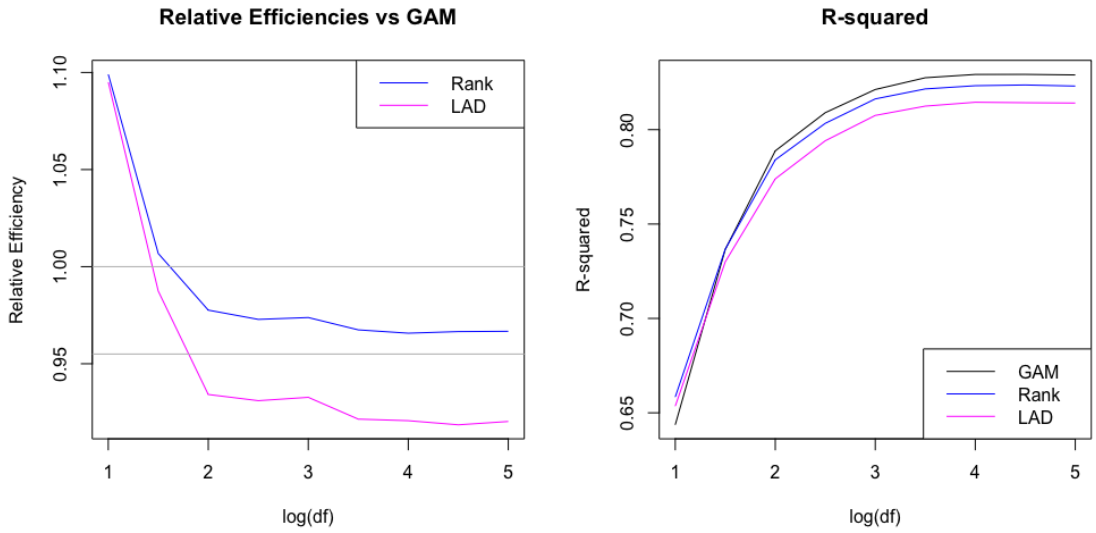


Figure 2.1: Heavy-tailed distribution relative efficiencies and R^2 values for increasing $\log(df)$.

where $\delta \in [0, 1]$ and $\sigma > 0$. This means the errors are drawn from the $N(0, 1)$ distribution with probability $1 - \delta$ and from the $N(0, \sigma^2)$ distribution with probability δ. To simulate this, we generate a random variate B from the Bernoulli distribution with probability of success $1 - \delta$; so, $B = 1$ with probability $1 - \delta$ and $B = 0$ with probability δ. Contaminated normal errors $\varepsilon \sim CN(\delta, \sigma)$ are then generated as

$$\varepsilon \sim B * X_1 + (1 - B) * X_2$$

where $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, \sigma^2)$. For our simulation experiment we took $\sigma = 3$ and δ taken from 0 to 0.35 in steps of .05 ($\mathcal{B}_1$). Once again, the correlation between two observations a distance r apart is $\exp(-(r/d)^2)$ with $d = 0.1$. Again the GAM, LAD model, and rank GAM were fit for the given data and relative efficiencies and R^2 calculated. These values were plotted against the proportion of contamination in Figure 2.2. The left panel shows that both LAD and rank GAM estimators were more efficient than GAM estimators for greater than five percent contamination, with rank GAM producing most efficient estimates for contamination ranging from five to fifteen percent. The gain in efficiency, however, asymptotes for high levels of contamination. This is generally the case when it

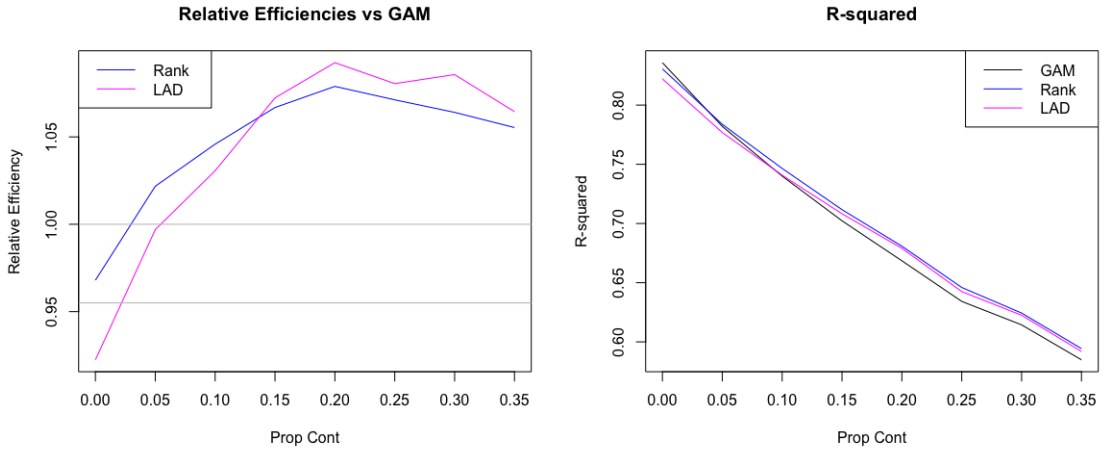


Figure 2.2: Relative efficiencies and R^2 values for increasing proportions of contamination of the normal distribution.

becomes difficult to distinguish between the 'real' and 'contaminating' distributions. The right panel of Figure 2.2 shows that rank GAM produces marginally better fitting models for contamination greater than five percent.

The performance of GAM estimators under various levels of spatial clustering was analyzed in the third simulation. The errors ε were drawn from a normal distribution centered at zero with standard deviation of one. The correlation between two observations a distance r apart is $\exp(-(r/d)^2)$ with d taken from 0.1 to 0.4 in steps of .05 for varying correlation structure. These represent weak clustering (almost independence) to string clustering of the spatial data. This also means the number of clusters in the data decreases with increasing d. The GAM (LS), LAD model, and rank GAM were fit for the given data and relative efficiencies and R^2 calculated as before; these were plotted against increasing correlation between observations (Figure 2.3). The left panel shows that the rank GAM estimator was more efficient than LAD and both are less efficient than GAM when the errors have low spatial correlation. However, the loss in efficiency for rank GAM in comparison to GAM ranged from 3% for low spatial correlation and 0% for high spatial correlation. The right panel of Figure 2.2 shows that rank GAM produced marginally poorer fitting models than GAM for low spatial correlation and virtually the same fit when the spatial correlation was

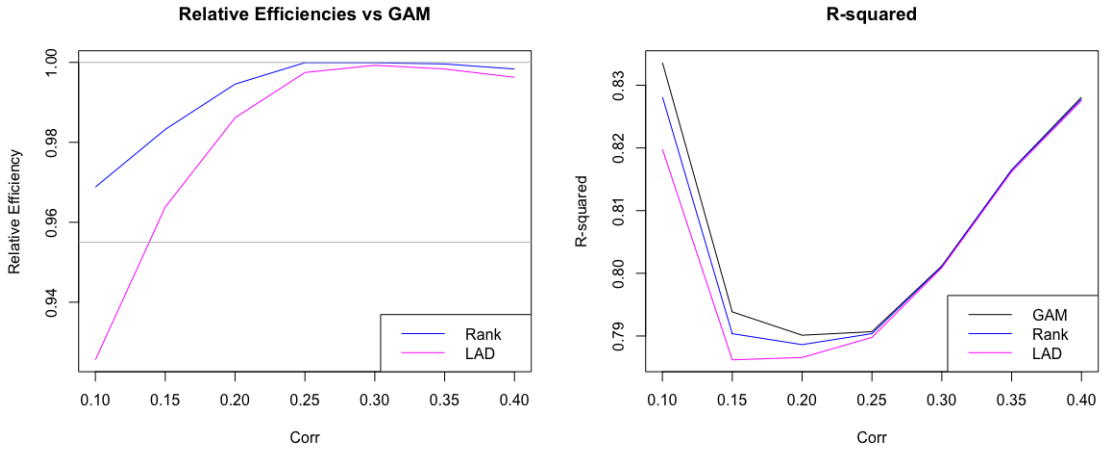


Figure 2.3: Relative efficiencies and R^2 values for increasing spatial correlations in data.

large. An interesting observation is that all methods provided poorer fit as the correlation increased from $d = .10$ to $d = .20$ but the fits improved when correlation increased further. The progression from $d = .10$ to $d = .40$ represents progression from almost independent spatial data structure to several small spatial clusters to a few large spatial clusters. It appears that the worst fits are obtained when the data are derived from several small spatial clusters.

2.5 Conclusions

This chapter proposes and studies rank-based estimators of generalized additive models. This provides a viable alternative to the usual likelihood based estimator of GAMs. Our estimation algorithm is simple. Our reformulation of rank estimators of GAMs as iteratively reweighted penalized least squares estimators, we manage to (1) take advantage of existing weighted GAM theory to establish the theoretical properties of the proposed estimator and (2) take advantage of existing software (eg. `mgcv` in R) to fit the models. in particular, our estimation procedure proceeds by performing repeated weighted GAM fits until convergence conditions are met. We evaluated the relative change in fits to establish convergence.

Our simulation experiments show that the proposed rank GAM estimation method outperforms GAM and LAD for data derived from processes that are heavy-tailed or contaminated. This is fairly common in climate studies and investigators often depend on simplifying the problem so that they can apply simple nonparametric tests such as the Wilcoxon rank-sum test. However, such approaches are not easy to apply for high-dimensional data with complex underlying structure. Thus, the proposed method gives a practical approach for studying problems where classical fitting of GAMs is inefficient.

Chapter 3

Rank GAM Applications

3.1 Introduction

Studies have revealed potential amplified warming effects in the northern latitudes ($> 60°$N) relative to overall global warming trends (Serreze and Francis, 2006) (Holland and Bitz, 2003). This trend has driven changes in Pacific marine systems and is predicted to affect future fish diversity and populations (Brander, 2007) (Cheung et al., 2013). Current fisheries research focuses on single species or predator-prey population mapping and prediction, however fisheries management research and ecologists now need better modeling techniques to analyze the responses of fishes to climate shifts in order to improve prediction and management.

Fisheries data are typically correlated over space and time (i.e. spatio-temporal data) and the relationship between variables is complex. Standard GLMs are usually not sufficient to describe such a complex system. The smoothing of each covariate with individual smoothing functions in GAMs are advantageous for such systems by assuming unknown nonparametric relationships between the covariates and the response. GAM can also be extended to correlated spatio-temporal data with relative ease (Fang and Chan, 2015) and can parse out major sources of variability when several models are fit and compared.

The generalized additive model has been considerably employed in the analysis of fisheries data over the past decade, and extensions for spatial and spatio-temporal data using tensor smoothers have been explored more recently. These models typically use spatial components such as latitude, longitude, and depth to analyze movement and changes in fish populations over two- or three-dimensional space by including factors such as sea surface temperature (SST) as changing over space and time.

Due to the non-Gaussian nature of ecological data and lack of robust estimation for these types of data, inferences and predictions using GAMs are typically difficult to achieve, particularly as the dimensions and number of covariates increases. It will be shown that the rank GAM estimation method described previously will produce more robust estimates for ecological data and can produce better fitting models.

3.2 Data

The data for this study was collated from two data sets provided by the National Oceanic and Atmospheric Administration (NOAA) for download to the public. Of primary use was the annual longline survey data (1979-present) of the Marine Ecology and Stock Assessment (MESA) Program conducted by the Auke Bay Laboratories in Alaska (AFSC, 2015). The MESA Program has performed longline surveys independently since 1979, dropping baited lines at specific locations ("stations") off the coast of Alaska to collect information of groundfish species from longline sets at specific stations 30-50 km apart all along the coast of Alaska. Seven major groundfish species are surveyed by the Alaska Fisheries Science Center (AFSC): giant grenadier, Pacific cod, Pacific halibut, rougheye rockfish, shortraker rockfish, shortspine thornyhead, and sablefish. The AFSC records weights and number of fish per species collected at each location. Each station is surveyed once a year, every year using longline sets consisting of 80 skates weighted and sunk to the sea floor 150-1000 meters off the coast. Sampling occurs on one day for each station. The longline sets are set out in the morning and left for a minimum of three hours before collection begins. If conditions are predicted to exceed, 10-foot seas and 30-knot winds, the longline sets are not deployed that day as the fish are likely to drop off the line, creating a sampling bias.

Information collected on the MESA sampling runs record the number of species collected at each location. A catch per unit effort (CPUE) is also calculated by the AFSC for each species at each station from the total number of fish caught divided by the total number of hooks deployed each day, therefore the CPUE is a more standardized measure of catch at

each location. MESA data were acquired for the years of 1979 through 2013. There is a high proportion of zero data that indicates no catch for the day and therefore a possible absence of fish at that location.

Global sea surface temperature (SST) readings were obtained from the National Centers for Environmental Information (formerly the National Climate Data Center) (NOAA, 2015). The data are interpolated and constructed from satellites, buoys, and ships at 1/4° latitude-longitude grids using a method by Richard W. Reynolds at the National Centers for Environmental Prediction. Since it was necessary to retrieve data from the early 1980s, only infrared satellite sensors known as Advanced Very High Resolution Radiometer (AVHRR) were available beginning in 1981. Each day contains four variables: daily SST, SST anomaly, estimated error standard deviation of analyzed SST, and sea ice concentration. The daily SST from AVHRR was compared to surface buoy measures of SST at various locations and found to be a very accurate and more complete record of SST across space. An average SST for each 1/4° latitude-longitude grid was calculated per year. Since deriving a mean SST by year flattens out most within year variability, a yearly coefficient of variance for SST was also calculated as

$$c_v = \frac{\sigma}{\mu}$$

at each latitude-longitude pairing, with σ being the yearly standard deviation and μ the yearly mean of SST. The coefficient of variance is a ratio of the standard deviation to the mean and includes information about the variability of the data.

In this study, we focused on two larger and more economically important fish: sablefish and Pacific cod. The fisheries data were matched with SST data from 1981 to 2013. There are 2344 observations for the sablefish and 1949 observations for Pacific cod. The CPUE measures for Pacific cod were log-transformed, while sablefish CPUE was sufficiently normal to use an identity link function in the GAM and rank GAM models.

3.3 Methods

The main goal of this analysis was to compare GAMs with the proposed rank GAMs and explore the relationship between the catch numbers and spatial, temporal, and environmental factors for both species of fish. The chosen dependent variable was catch per unit effort (CPUE) of all fish of each species caught at one station on one measurement day. A variation of log transformation was performed on CPUE for Pacific cod to induce normality, as previous analyses revealed high occurrences of low or zero CPUE. The models included three main factors that are theorized to affect population dynamics: location in the form of latitude-longitude pairs, time in the form of years, and environment in the form of yearly SST coefficient of variance ($SSTc_v$). All analyses were performed using the free statistical software R and include use of the `mgcv` package (Wood, 2006).

Three main questions were addressed in this analysis: (i) How does CPUE change over location and time? This was addressed by fitting a spatial and spatio-temporal model for each species. (ii) Does SST contribute to variation in CPUE? A spatio-temporal with environmental model was fit assuming SST changes over space and time and considering the relationship between SST and CPUE. (iii) Do the proposed rank GAMs produce better estimates of the smoothing functions in the tested models? Cross-validation was performed on the GAMs and rank GAMs and the mean of cross-validation errors compared.

3.3.1 Models

In this study, we will consider three formulations to model CPUE: spatial, spatio-temporal, and spatio-temporal with environmental.

Spatial formulation

The spatial model assumes that distribution of Pacific cod and sablefish along the studied region is only affected by location, and that CPUE variations are due to interannual

changes in stock size:

$$Y_{t,(u,v)} = s_1(u, v) + \varepsilon_{t,(u,v)} \tag{3.1}$$

where Y is the CPUE, t is time in years from 1981-2013, u is latitude, v is longitude, and s_1 is a 2D smoothing function. For Pacific cod, $Y_{t,(u,v)}$ is the natural log of CPUE at (u, v) in t. For sablefish, $Y_{t,(u,v)}$ is CPUE at (u, v) in t.

Spatio-temporal formulation

For the spatio-temporal model, time is included into a three-dimensional tensor smoothing function which allow species distribution to smoothly and simultaneously change over location and time.

$$Y_{(u,v,t)} = z_1(u, v, t) + \varepsilon_{(u,v,t)} \tag{3.2}$$

where z_1 is a three-dimensional tensor smoothing function; CPUE, time, and location are defined as in the spatial formulation.

Spatio-temporal with environmental formulation

This model allows the effect of the environmental variable SST to also smoothly change over location and time by using a tensor smoothing function:

$$Y_{(u,v,t)} = z_1(u, v, t) + z_2(u, v, t) \cdot SST_{(u,v,t)} + \varepsilon_{(u,v,t)} \tag{3.3}$$

where z_1 and z_2 are three-dimensional tensor smoothing functions; CPUE, time, and location are defined as in the spatial and spatio-temporal formulations.

3.3.2 Cross-validation

Cross-validation was performed on the GAM and rank GAM version of each of the above formulations. The data was randomly split into ten roughly equal sets $F_1, \ldots, F_k$, $k = 10$. Consider training on (x_j, y_j), $j \notin F_{10}$, and validating on (x_j, y_j), $j \in F_{10}$. For each

estimation technique, an estimated $\hat{f}_a^{-k}$, $a = \text{GAM, R}$, is computed on the training sets and the total error is recorded on the validation set

$$e_k(a) = \sum_{j \in F_k} (y_j - \hat{f}_a^{-k}(x_j))^2$$

For each estimation technique $a = \text{GAM, R}$, the average error was computed over all k folds

$$CV(a) = \frac{1}{n} \sum_{k=1}^{K} e_k(a) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in F_k} (y_j - \hat{f}_a^{-k}(x_j))^2$$

The method a that minimizes CV is considered the best method for minimizing prediction error and therefore the optimal method of prediction for the given model.

3.3.3 Adjusted R^2 and Effect Size

For the GAM and rank GAM versions of each formulation with p predictors, an adjusted R^2 value was calculated using R^2 as defined previously:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1} = 1 - \frac{\frac{1}{n-p-1} \sum_{j=1}^{n} (f_j - \hat{f}_j)^2}{\frac{1}{n-1} \sum_{j=1}^{n} (f_j - \bar{f})^2}$$

An R^2 value explains the goodness-of-fit of the model, but becomes inflated with the addition of multiple predictors regardless of their significant contributions to the model. The adjusted R^2 value compensates for the number of predictors in the model and therefore produces a more accurate representation of the percentage of data explained by the model.

The effect size of predictor A, a measure of the contribution of that predictor to the response, was calculated using Cohen's f^2 value

$$f^2 = \frac{R^2_f - R^2_0}{1 - R^2_f}$$

which is a function of the R^2 values of the model including predictor A (R_f^2) and with predictor A removed from the model (R_0^2). The Cohen's f^2 value can therefore be considered a ratio of the proportion of variance explained by predictor A to the proportion of unexplained variance. A very large Cohen's f^2 value, $f^2 >> 1$ indicates that the proportion of variance explained by predictor A is much greater than the proportion of variance not explained by the full model.

3.4 Results

3.4.1 Sablefish

The three models fit for sablefish CPUE with intercept estimates, standard errors, and p-values are given in Table 3.1. It can be seen that cross-validation prediction error (CV Pred Error) is reduced for the rank GAM estimation method of all three models. Adjusted R^2 values increase for the rank GAM versions, showing rank GAM produces a better fit model than GAM estimation. Note the model fit improves greatly when using a spatio-temporal model for sablefish CPUE over a purely spatial model, regardless of estimation method used. Since adjusted R^2 only increases if the added predictor contributes significantly to the model, it can be seen from the adjusted R^2 values of the spatio-temporal model and the spatio-temporal with environmental model (SpT-Enviro) that although SST may contribute significantly to the model, the effect of SST on sablefish CPUE is likely small.

The effect size of location and year for the spatio-temporal model fit for sablefish CPUE is given in Table 3.2. While the effects for location and year increase when going from GAM estimation to rank GAM estimation, the ratio of location to year effect sizes remains similar between the two estimation methods, indicating consistency of rank GAM estimation for two-dimensional and three-dimensional predictors.

The change in sablefish CPUE over space modeled using rank GAM estimation is illustrated in Figure 3.1. The highest sablefish CPUE counts are found in the Gulf of Alaska region and taper off toward the Bering Sea.

Table 3.1: Three models of sablefish CPUE tested: (1) spatial, (2) spatio-temporal and (3) spatial and environmental. Estimated intercepts, statistical significance, adjusted R^2, and cross-validation prediction error are shown for each method.

Spatial Model	$Y_{(u,v)} = \beta_0 + s_1(u,v) + \varepsilon_{(u,v)}$				
Method	Intercept	Std. Error	Pr($>\|t\|$)	Adj. R^2	CV Pred Error
GAM	7.100	.042	0.000	.678	1.258
Rank GAM	7.004	.037	0.000	.737	1.239
Spatio-Temporal Model	$Y_{(u,v,t)} = \beta_0 + z_1(u,v,t) + \varepsilon_{(u,v,t)}$				
Method	Intercept	Std. Error	Pr($>\|t\|$)	Adj. R^2	CV Pred Error
GAM	7.100	.032	0.000	.817	.902
Rank GAM	7.034	.026	0.000	.875	.856
SpT-Enviro Model	$Y_{(u,v,t)} = \beta_0 + z_1(u,v,t) + z_2(u,v,t) \cdot SST_{(u,v,t)} + \varepsilon_{(u,v,t)}$				
Method	Intercept	Std. Error	Pr($>\|t\|$)	Adj. R^2	CV Pred Error
GAM	7.680	.245	0.000	.844	.893
Rank GAM	7.423	.187	0.000	.906	.843

Table 3.2: Cohen's f^2 values of location and time for the GAM and rank GAM versions of the spatio-temporal model fit for sablefish CPUE

Spatio-Temporal Model	$Y_{(u,v,t)} = \beta_0 + z_1(u,v,t) + \varepsilon_{(u,v,t)}$		
Method	Location	Year	Ratio
GAM	4.496	0.803	5.597
Rank GAM	6.940	1.169	5.938

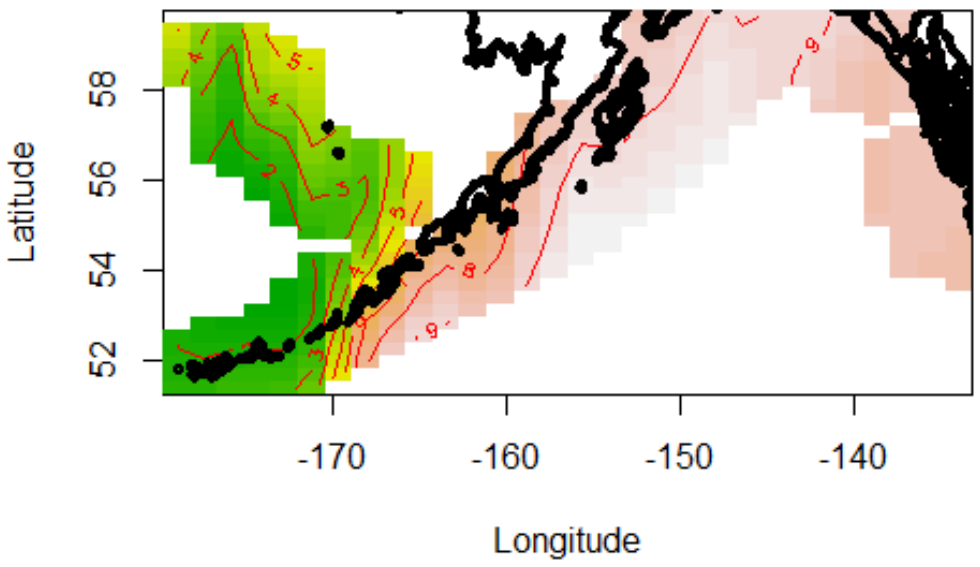


Figure 3.1: Spatial change in CPUE for sablefish. Dark green indicates low values of CPUE; pink to white indicates high values of CPUE.

3.4.2 Pacific cod

The three models fit for Pacific cod CPUE with intercept estimates, standard errors, and p-values are given in Table 3.3. As in the sablefish models, cross-validation prediction error (CV Pred Error) is reduced for the rank GAM estimation method of all three models. Adjusted R^2 values increase for the rank GAM versions, showing rank GAM produces a better fit model than GAM estimation. Again, the adjusted R^2 improves greatly when using a spatio-temporal model for Pacific cod CPUE over a purely spatial model, regardless of estimation method used. As before, it can be seen from the adjusted R^2 values of the spatio-temporal model and the spatio-temporal with environmental model (SpT-Enviro) that although SST may contribute significantly to the model, the effect of SST on Pacific cod CPUE is again likely small.

The effect size of location and year for the spatio-temporal model fit for Pacific cod CPUE is given in Table 3.2. While the effects for location and year increase considerably

Table 3.3: Three models of Pacific cod CPUE tested: (1) spatial, (2) spatio-temporal and (3) spatial and environmental. Estimated intercepts, statistical significance, adjusted R^2, and cross-validation prediction error are shown for each method.

Spatial Model	$Y_{(u,v)} = \beta_0 + s_1(u,v) + \varepsilon_{(u,v)}$				
Method	Intercept	Std. Error	Pr(> |t|)	Adj. R^2	CV Pred Error
GAM	1.086	.014	0.000	.598	.323
Rank GAM	1.056	.011	0.000	.683	.306
Spatio-temporal Model	$Y_{(u,v,t)} = \beta_0 + z_1(u,v,t) + \varepsilon_{(u,v,t)}$				
Method	Intercept	Std. Error	Pr(> |t|)	Adj. R^2	CV Pred Error
GAM	1.086	.009	0.000	.817	.188
Rank GAM	1.067	.006	0.000	.904	.166
SpT-Enviro Model	$Y_{(u,v)} = \beta_0 + z_1(u,v,t) + z_2(u,v,t) \cdot SST_{(u,v,t)} + \varepsilon_{(u,v)}$				
Method	Intercept	Std. Error	Pr(> |t|)	Adj. R^2	CV Pred Error
GAM	1.228	.073	0.000	.832	.183
Rank GAM	1.195	.049	0.000	.915	.168

Table 3.4: Cohen's f^2 values of location and time for the GAM and rank GAM versions of the spatio-temporal model fit for Pacific cod CPUE

Spatio-Temporal Model	$Y_{(u,v,t)} = \beta_0 + z_1(u,v,t) + \varepsilon_{(u,v,t)}$		
Method	Location	Year	Ratio
GAM	3.319	1.264	2.626
Rank GAM	6.518	2.408	2.707

when going from GAM estimation to rank GAM estimation, the ratio of location to year effect sizes remains stable between the two estimation methods, indicating consistency.

The change in Pacific cod CPUE over space is illustrated in Figure 3.2. Highest CPUEs are concentrated along the coast of Alaska in the gulf region.

3.5 Discussion

Fitting GAMs is a common method employed in fisheries research to model population changes over space and time, and the method proved its simplicity to deploy on the Alaskan fisheries data. The spatial models show that both sablefish and Pacific cod are caught more easily in the Gulf of Alaska region (Figure 3.1 and 3.2). Static images only partially capture

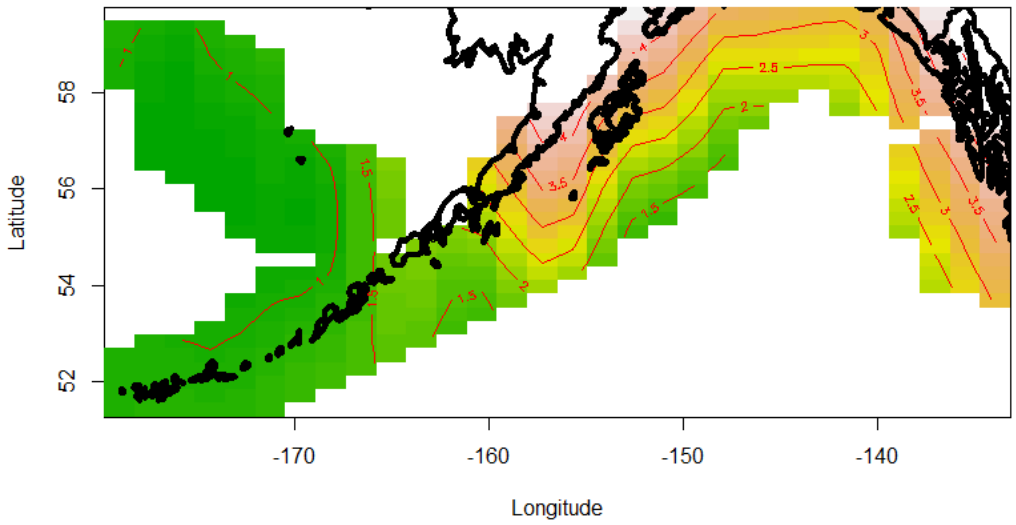
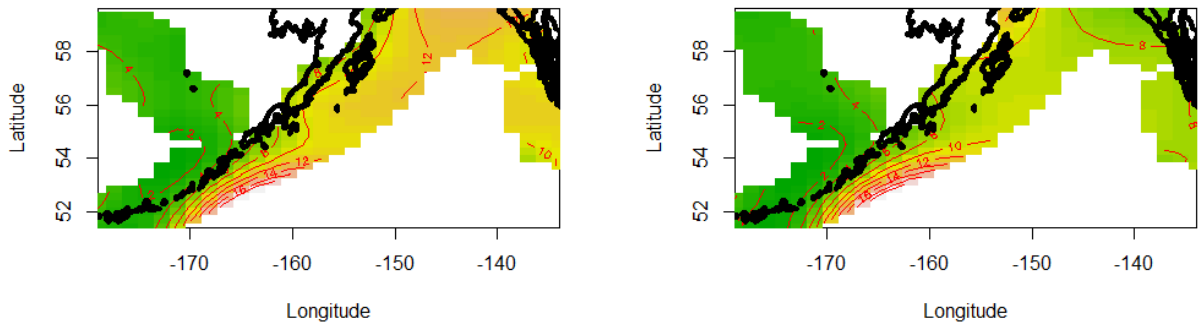


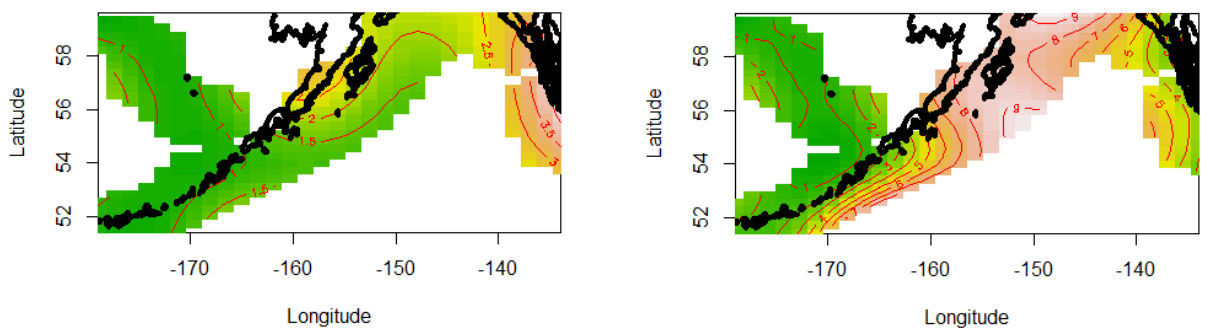
Figure 3.2: Spatial change in CPUE for Pacific cod. Dark green indicates low CPUE; pink to white indicates high CPUE.

the change in CPUE over space and time. In Figure 3.3, the CPUE of sablefish is shown for 1990 and 2008 where climate events in the northern Pacific Ocean were similar. Time-lapsed images for all 33 years shows CPUE fluctuating from low to high several times over the years. Figure 3.4 shows increased CPUE in a larger portion of the Gulf of Alaska in 2008 for the Pacific cod.



(a) 1990 (b) 2008

Figure 3.3: CPUE for sablefish in 1990 and 2008.



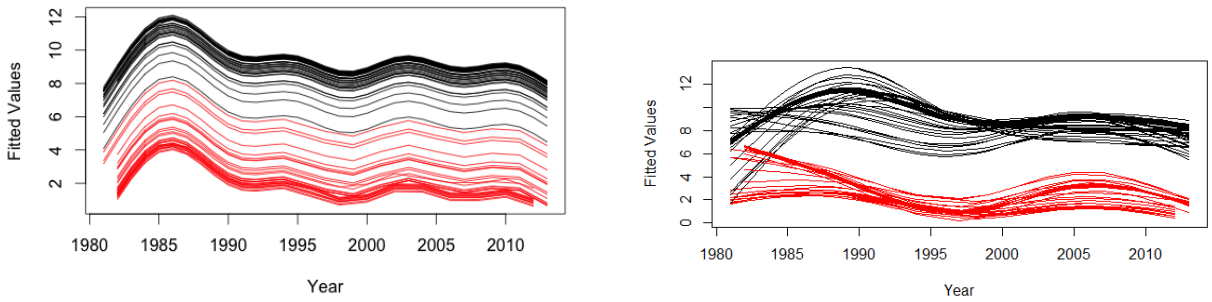
(a) 1990 (b) 2008

Figure 3.4: CPUE for Pacific cod in 1990 and 2008.

The spatio-temporal model is an as-yet uncommon way of modeling CPUE, by considering CPUE to change over location and time simultaneously. As an exercise, the model

$$Y_{t,(u,v)} = \beta_0 + s_1(u,v) + g_1(t) + \varepsilon_{t,(u,v)} \tag{3.4}$$

was fit and compared to the model given in equation (3.2) for sablefish CPUE. Using the three-dimensional tensor models allowed for plotting CPUE change over time for each location point, thereby gaining a better understanding of the "ease of catch" of each species and whether locations share certain CPUE functional patterns. Each station's fitted CPUE values were plotted against time in years, which produced the plots in Figure 3.5. The first model produces results showing the same pattern of CPUE change over time for all stations at differing levels. The spatio-temporal model produces different patterns of CPUE over time for each station, and a grouping of stations with common patterns become apparent. This highlights patterns not visible when CPUE is reduced to two-dimentional space covariates and lends support for incorporating time to covariates where relevant. This spatio-temporal modeling approach is not yet common for fisheries data; typical models consider space separately from time in two-dimentional latitude-longitude pairs or three-dimentional latitude-longitude-depth triples. This grouping trend is apparent in the six other species of



(a) Time as additive effect

(b) Spatio-temporal formulation

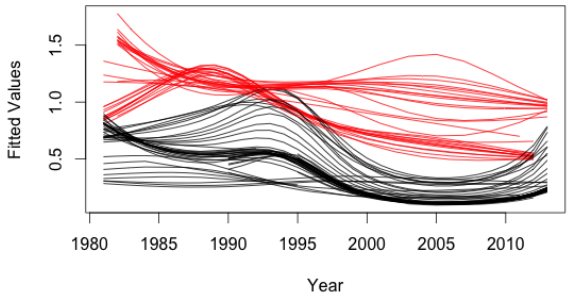
Figure 3.5: Sablefish CPUE modeled using 2D and 3D smoothers.

groundfish surveyed by the MESA project (Figure 3.6), which indicates a station-location effect that may be of future research interest.

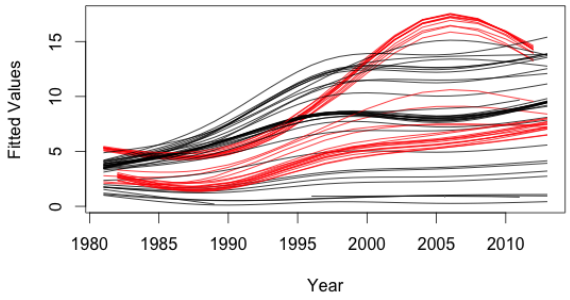
The effect size values given in Tables 3.2 and 3.4 indicate a higher effect size for location than year. Based on Cohen's suggested levels at which to describe an effect, these values would be considered very high. The problem with measuring effect size on these spatio-temporal GAM models is concerned with the model structure. The full model for our effect size calculations was given in equation (3.2). Standard considerations for measuring location effect size dictates the removal of the location variables from the model to produce a reduced model

$$Y_t = \beta_0 + g_1(t) + \varepsilon_t$$

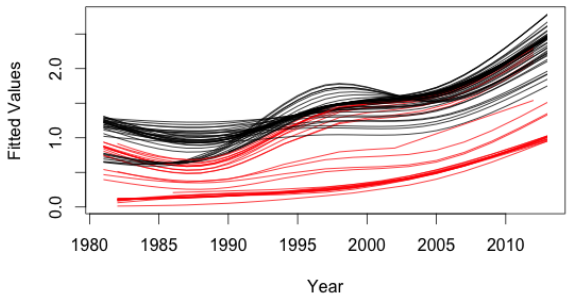
which considers CPUE as only a function of time with g_1 being a one-dimensional smoothing function. Similarly, measuring year effect size would necessitate a reduced model of the form given in equation (3.1). These reduced models have changed the type of function being estimated for the covariate of interest, therefore changing the type of model being compared to the full model. The other issue is in the full model itself. Assuming the reduced models given above, another natural full model to calculate effect size could be as given in equation (3.4). Since this would have R^2 values different from the original full model, different Cohen's f^2 values would be produced for the same reduced models. This highlights a weakness in



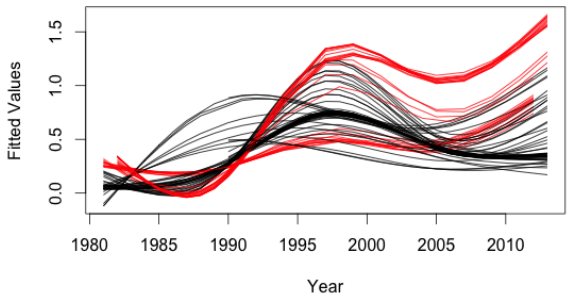
(a) Pacific cod



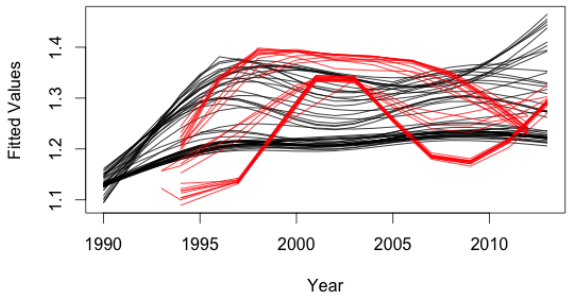
(b) Giant grenadier



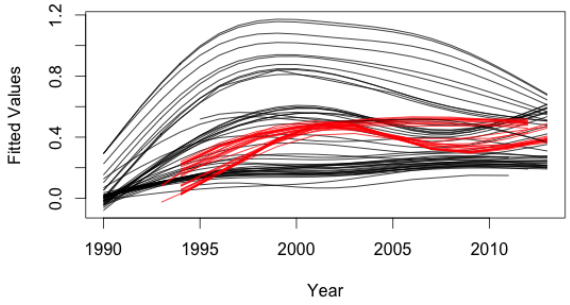
(c) Shortspine thornyhead



(d) Pacific halibut



(e) Rougheye rockfish



(f) Shortraker rockfish

Figure 3.6: CPUE for each of six groundfish species modeled using the spatio-temporal formulation.

Cohen's f^2 that a best-fit model should be properly defined when discussing effect size and also shows an inherent issue with using Cohen's f^2 as a measure of effect size for GAMs and rank GAMs since the smoothed functions can change dimentions from full to reduced models.

Bathythermographic and hydrochemical data are available from buoys within the MESA survey area that capture temperature variations with depth and chemical analysis of the surrounding waters. Further work can focus on interpolating buoy data for the fish catch locations using SST from the satellite data to match SST from the buoys for interpolation, since the buoy SST and satellite SST have been found to be highly linearly correlated. This will provide further predictors of climate to include in the models and better determine what kind of effect climate may have on groundfish species. Only two species were modeled in this analysis, however the modeling of the four other species would be approached similarly and could be easily extended in future work.

Chapter 4

Discussion on Rank GAMs and Future Work

This dissertation focused on improvement of GAM estimation by using rank estimators to increase robustness and efficiency, particularly for non-Gaussian data common in ecology and climatology. The rank GAM estimation method was shown to be a reformulation of GAM likelihood estimation using iterative reweighting, which allows for the use of well-defined GAM theory to assert the asymptotic properties and consistency of rank GAM estimation under certain conditions. In the presence of heavy-tailed error distributions or spatial clustering, rank GAM estimation lost little efficiency and produced better fit in cases of heavy-tailed errors. Rank GAM estimation outperforms GAM estimation in efficiency and fit for errors with contamination ranging from five to fifteen percent, illustrating the advantage of rank GAM estimation for responses with outliers. Since rank GAM estimation was expected to perform better for non-Gaussian data, several models were fit and compared using GAM and rank GAM estimation for a fisheries dataset. Improved fit and reduced cross-validation prediction errors were consistently found for the rank estimation of GAMs fit for the data. Inclusion of time as a dimension by using tensor functions highlighted response patterns more representative of the spatial variation expected in the data.

One issue with GAMs being used for environmental data is that the models can become overly complex for data with a large number of covariates. Future plans of adding buoy data to the fisheries dataset will complicate the fitting of GAMs for this data and necessitate a different approach to avoid the 'curse of dimensionality'. Additive index models, in which the response is related to a vector of predictors

$$Y = h_1(\alpha_1'\mathbf{x}) + h_2(\alpha_2'\mathbf{x}) + \ldots + h_p(\alpha_p'\mathbf{x})$$

with projection indices $\alpha_1, \alpha_2, \ldots, \alpha_p$ and nonparametric ridge functions $h_1, h_2, \ldots, h_p$ would allow for this type of flexibility. Rank estimation of the single index regression model, a variety of the additive index model, is currently being established. Application of rank single-index models to the new data would be of interest for furthering modeling and prediction of fisheries data with complex oceanographic covariates.

Bibliography

Abebe, A. and McKean, J. W. (2013). Weighted Wilcoxon estimators in nonlinear regression. *Aust. N. Z. J. Stat.*, 55(4):401–420.

Abebe, A., McKean, J. W., Kloke, J. D., and Bilgic, Y. (2016). *Iterated reweighted rank-based estimates for gee models*. to appear.

AFSC (2015). Noaa mesa: Longline survey. Online. Accessed: 2014-04-14.

Alimadad, A. and Salibian-Barrera, M. (2011). An outlier-robust fit for generalized additive models with applications to disease outbreak detection. *J. Amer. Statist. Assoc.*, 106(494):719–731.

Bindele, H. F. and Abebe, A. (2012). Bounded influence nonlinear signed-rank regression. *Canad. J. Statist.*, 40(1):172–189.

Bindele, H. F. and Abebe, A. (2015). Semi-parametric rank regression with missing responses. *J. Multivariate Anal.*, 142:117–132.

Brander, K. M. (2007). Global fish production and climate change. *Proceedings of the National Academy of Sciences*, 104(50):19709–19714.

Cheung, W. W. L., Sarmiento, J. L., Dunne, J., Frolicher, T. L., Lam, V. W. Y., Deng Palomares, M. L., Watson, R., and Pauly, D. (2013). Shrinking of fishes exacerbates impacts of global ocean changes on marine ecosystems. *Nature Clim. Change*, 3(3):254–258.

Croux, C., Gijbels, I., and Prosdocimi, I. (2012). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*, 68(1):31–44.

Denhere, M. and Bindele, H. F. (2016). Rank estimation for the functional linear model. *Journal of Applied Statistics*, pages 1–17.

Draper, D. (1988). Rank based robust analysis of linear models. I. Exposition and review. *Statist. Sci.*, 3(2):239–271. With comments and a rejoinder by the author.

Fang, X. and Chan, K.-S. (2015). Additive models with spatio-temporal data. *Environ. Ecol. Stat.*, 22(1):61–86.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76(376):817–823.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London.

Hettmansperger, T. P. and McKean, J. W. (2011). *Robust nonparametric statistical methods*, volume 119 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, second edition.

Holland, M. M. and Bitz, C. M. (2003). Polar amplification of climate change in coupled models. *Climate Dynamics*, 21(3):221–232.

Hössjer, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Amer. Statist. Assoc.*, 89(425):149–158.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, 43:1449–1458.

Jurečková, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.*, 42:1328–1338.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London. Second edition.

Miakonkana, G.-v. M. and Abebe, A. (2014). Iterative rank estimation for generalized linear models. *J. Statist. Plann. Inference*, 151/152:60–72.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nguelifack, B. M., Kwessi, E. A., and Abebe, A. (2015). Generalised signed-rank estimation for nonlinear models with multidimensional indices. *J. Nonparametr. Stat.*, 27(2):215–228.

NOAA (2015). National data buoy center - optimum interpolation sea surface temperature. Online. Accessed: 2015-09-18.

Serreze, M. C. and Francis, J. A. (2006). The arctic amplification debate. *Climatic Change*, 76(3):241–264.

Sievers, G. L. and Abebe, A. (2004). Rank estimation of regression coefficients using iterated reweighted least squares. *J. Stat. Comput. Simul.*, 74(11):821–831.

Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wang, L., Kai, B., and Li, R. (2009). Local rank inference for varying coefficient models. *J. Amer. Statist. Assoc.*, 104(488):1631–1645.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439–447.

Wong, R. K. W., Yao, F., and Lee, T. C. M. (2014). Robust estimation for generalized additive models. *J. Comput. Graph. Statist.*, 23(1):270–289.

Wood, S. N. (2006). *Generalized additive models*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL. An introduction with ßfR.