stats

MDPI

Article

Cross-Validation, Information Theory, or Maximum Likelihood? A Comparison of Tuning Methods for Penalized Splines

Lauren N. Berry [1,2,†] **and Nathaniel E. Helwig** [1,2,*]

[1] Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA; berry478@umn.edu
[2] School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA
* Correspondence: helwig@umn.edu; Tel.: +1-612-624-8363
† Current address: Department of Statistics, Grand Valley State University, Allendale, MI 49401, USA.

Abstract: Functional data analysis techniques, such as penalized splines, have become common tools used in a variety of applied research settings. Penalized spline estimators are frequently used in applied research to estimate unknown functions from noisy data. The success of these estimators depends on choosing a tuning parameter that provides the correct balance between fitting and smoothing the data. Several different smoothing parameter selection methods have been proposed for choosing a reasonable tuning parameter. The proposed methods generally fall into one of three categories: cross-validation methods, information theoretic methods, or maximum likelihood methods. Despite the well-known importance of selecting an ideal smoothing parameter, there is little agreement in the literature regarding which method(s) should be considered when analyzing real data. In this paper, we address this issue by exploring the practical performance of six popular tuning methods under a variety of simulated and real data situations. Our results reveal that maximum likelihood methods outperform the popular cross-validation methods in most situations—especially in the presence of correlated errors. Furthermore, our results reveal that the maximum likelihood methods perform well even when the errors are non-Gaussian and/or heteroscedastic. For real data applications, we recommend comparing results using cross-validation and maximum likelihood tuning methods, given that these methods tend to perform similarly (differently) when the model is correctly (incorrectly) specified.

Keywords: functional data analysis; nonparametric regression; regularization; smoothing

Citation: Berry, L.N.; Helwig, N.E. Cross-Validation, Information Theory, or Maximum Likelihood? A Comparison of Tuning Methods for Penalized Splines. *Stats* **2021**, *4*, 701–724. https://doi.org/10.3390/stats4030042

Academic Editor: Manuel Oviedo de la Fuente

Received: 30 July 2021
Accepted: 28 August 2021
Published: 2 September 2021

1. Introduction

Functional data analysis (FDA) considers the analysis of data that are (noisy) realizations of a functional process [1–3]. Such data can be found in many fields [4,5] and are becoming more common in the biomedical and social sciences, e.g., in the form of ecological momentary assessments [6,7] collected using smart phone apps. Most FDA techniques can be interpreted as functional extensions of standard methods used in applied statistics. For example, the nonparametric regression model considered in this paper can be interpreted as a functional extension of the simple model $Y = \mu + \epsilon$, which is assumed for a one sample location test. A fundamental aspect of many FDA applications is choosing a method to smooth the (noisy) functional data, and splines are one of the most popular smoothing methods used in applications of FDA [2,4]. Note that spline smoothers assume a nonparametric regression model of the form $Y_i = \eta(X_i) + \epsilon_i$, where $\eta(\cdot)$ is the unknown mean function.

Nonparametric regression models are frequently used in applied research to estimate unknown functional relationships between variables (e.g., see [8–14]). Unlike parametric regression models, which assume that the functional relationship between variables has a known form that depends on unknown parameters, nonparametric regression models do not assume that the form of the relationship between variables is known [15,16].

Instead of assuming a particular functional form for the relationship, nonparametric regression models attempt to learn the form of the functional relationship from a sample of noisy data. As a result, nonparametric regression models are a type of statistical learning (e.g., see [17]), given that the collected data enable the researcher to discover functional forms that describe relations between variables. The overarching goal of nonparametric regression is to estimate a function that fits the data well while still maintaining a parsimonious (i.e., smooth) estimate of the functional relationship.

Penalized splines are a popular approach for estimating unknown functional relationships from noisy data. Note that penalized splines are a form of generalized ridge regression [18], where a quadratic smoothness penalty is added to the ordinary least squares loss function. The influence of the smoothness penalty is controlled by a nonnegative smoothing (or tuning) parameter $\lambda \geq 0$, which controls the trade-off between fitting the data well and obtaining a smooth estimate. This paper focuses on the Gaussian-type response, so the fit to the data is measured by the ordinary least squares loss function. More generally, penalized splines can be viewed as a form of penalized likelihood estimation [19], where the goal is to find the function that minimizes

$$-\frac{1}{n}\text{Log-Likelihood} + \lambda\text{Penalty}, \tag{1}$$

where the first term quantifies the fidelity to the data (with n denoting the sample size) and the second term quantifies the (lack of) parsimony of the estimate.

As the smoothing parameter $\lambda \to 0$, the log-likelihood term dominates the loss functional, which causes the estimator to converge to the maximum likelihood estimator. In contrast, as $\lambda \to \infty$, the penalty term dominates the loss functional, which causes the estimator to converge to a "perfectly smooth" estimator (later defined). When working with finite samples of noisy data, it is desirable to select a $\lambda \in (0, \infty)$ that provides an ideal balance between fitting and smoothing the data. If the signal to noise level is relatively large, it may be possible to manually select a reasonable value of λ via visual inspection. However, for reliable and valid smoothing parameter selection across multiple noise levels, some automated method for selecting λ should be preferred. A variety of different methods have been proposed for automatically selecting an ideal value of λ for a given sample of data. However, there is no general consensus as to which method should be preferred for general situations.

In this paper, we compare six popular tuning methods that can be categorized into three distinct groups: (i) cross-validation based methods, which include ordinary cross-validation (OCV) and generalized cross-validation (GCV); (ii) information theoretic methods, which include an information criterion (AIC) and the Bayesian information criterion (BIC); and (iii) maximum likelihood methods, which include standard (ML) and restricted maximum likelihood estimation (REML). The cross-validation tuning methods are often the default choice for smoothing parameter selection, e.g., the popular `smooth.spline` function in R [20] offers both the GCV (default) and OCV tuning options. Despite the popularity of the OCV and GCV, it is known that these tuning criteria can breakdown when the model error terms are correlated [21,22]. In such situations, these CV criteria tend to under-smooth the data (i.e., chooses a λ that is too small) because the structure in the error terms is perceived to be part of the mean structure.

When a researcher has a priori knowledge that the errors are correlated, it is possible to incorporate that knowledge into the estimation problem (e.g., see [23–25]). However, in most real data applications, the researcher lacks prior knowledge about the nature of the error correlation structure, so it is not possible to incorporate such information into the estimation process. One (naive) option would be to fit a penalized spline and then to inspect the model residuals in an attempt to learn about the error correlation structure. However, it has been shown that estimating the correlation structure from residuals is difficult, given that the residuals often look uncorrelated even when the error terms are correlated [22]. This is due to the (previously mentioned) issue that the error correlation is

often absorbed into the estimate of the mean function when using popular tuning methods such as the OCV and GCV. Consequently, when the error terms may be correlated, some robust alternative tuning approach is needed.

Past research has shown that the cross-validation and maximum likelihood tuning methods have several common features [26]; however, certain tuning criteria may be more robust than the OCV and GCV in the presence of misspecified error structures. Specifically, Krivobokova and Kauermann [27] showed that the REML tuning criterion should be expected to outperform the OCV, GCV, and AIC (with respect to mean function recovery) when the errors are autocorrelated, and Lee [28] found that the AIC is more robust than the cross-validation criteria when the errors have non-constant variance. However, these findings focused on the situation when the errors follow a Gaussian distribution. To the best of our knowledge, no study has thoroughly compared the various tuning criteria under a wide variety of different combinations of error variance, error correlation, and error distribution. In this study, we explore how the distributional properties of the error terms affect not only the mean function recovery (which has been the focus in past studies) but also the standard errors used for inference about the unknown function (which has been largely ignored in past works).

The remainder of this paper is organized as follows. Section 2 provides some background about estimation and inference for smoothing splines. Section 3 presents the six smoothing parameter selection methods: OCV, GCV, AIC, BIC, ML, and REML. Section 4 conducts a thorough simulation study comparing the performance of the tuning methods under a variety of different data generating conditions. Section 5 demonstrates that the different tuning methods can produce noteworthy differences when analyzing real data. Finally, Section 6 discusses the important conclusions drawn from the current work as well as future research directions related to robust estimation and inference for penalized splines.

2. Theoretical Background

2.1. Smoothing Spline Definition

Let $\{(x_i, y_i)\}_{i=1}^n$ denote a set of n observations, where $y_i \in \mathbb{R}$ is the real-valued response variable for the i^{th} observation and $x_i \in [a, b]$ is the predictor variable for the i^{th} observation. Note that the predictor is assumed to be bounded, and we can assume that $a = 0$ and $b = 1$ without loss of generality. Consider a nonparametric regression model of the form

$$y_i = \eta(x_i) + \epsilon_i, \tag{2}$$

where $\eta(\cdot)$ is the unknown smooth function relating x_i to y_i, and ϵ_i is the error term for the i^{th} observation. In standard applications of nonparametric regression, the errors are assumed to be independent and identically distributed realizations of a continuous random variable with mean zero, i.e., $\epsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$. Note that this implies that the unknown function $\eta(\cdot)$ determines the conditional mean of Y given X. If the errors are correlated and/or heteroscedastic, then $\eta(\cdot)$ still determines the conditional mean of Y given X, as long as the error terms satisfy $E(\epsilon_i) = 0$.

Given a smoothing parameter $\lambda \geq 0$, the m^{th} order polynomial smoothing spline estimator η_λ is the minimizer of a penalized least squares functional, i.e.,

$$\eta_\lambda = \min_{\eta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \eta(x_i))^2 + \lambda J_m(\eta) \tag{3}$$

where $J_m(\eta) = \int_0^1 |\eta^{(m)}(z)|^2 dz$ is the penalty functional, with $\eta^{(m)}(\cdot)$ denoting the m^{th} derivative of $\eta(\cdot)$, and $\mathcal{H} = \{\eta : J_m(\eta) < \infty\}$ denotes the space of functions with square integrable m^{th} derivatives. The function space $\mathcal{H}$ can be decomposed into two orthogonal subspaces such as $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{\eta : J_m(\eta) = 0\}$ is the null space and $\mathcal{H}_1 = \{\eta : 0 < J_m(\eta) < \infty\}$ is the contrast space. The implies that $\eta = \eta_0 + \eta_1$, where $\eta_0 \in \mathcal{H}_0$ and $\eta_1 \in \mathcal{H}_1$. The null space is spanned by the m basis functions $\phi_j(x) = x^j$ for $j = 0, \ldots, m-1$. As the smoothing parameter $\lambda \to \infty$, the estimator η_λ is suppressed to

its null space representation η_0. For example, setting $m = 2$ produces a cubic smoothing spline, where the estimator η_λ approaches a linear function as $\lambda \to \infty$.

2.2. Representation and Computation

The Kimeldorf–Wahba representer theorem [29] reveals that the minimizer of the penalized least squares functional in Equation (3) can be written as

$$\eta_\lambda(x) = \sum_{j=0}^{m-1} \beta_j \phi_j(x) + \sum_{k=1}^{r} \gamma_k \kappa(x, x_k^*) \tag{4}$$

where $\{\phi_j\}_{j=0}^{m-1}$ are known functions that span the null space, $\kappa(\cdot,\cdot)$ is the reproducing kernel of the contrast space, and $\{x_k^*\}_{k=1}^{r}$ are the selected knots. Note that the representer theorem uses all observed design points as knots (i.e., $r = n$ and $x_i^* = x_i$); however, it is possible to obtain good approximations using $r \ll n$ selected design points as knots [30]. For practical computation, note that the reproducing kernel function has the form

$$\kappa(x, y) = \psi_m(x)\psi_m(y) + (-1)^{m-1}\psi_{2m}(|x - y|), \tag{5}$$

where ψ_m are Bernoulli polynomials [16,31]. Using the Kimeldorf–Wahba representation in Equation (4), the function estimation reduces to the estimation of the unknown coefficient vectors $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{m-1})^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)^\top$.

The Kimeldorf–Wahba representer theorem implies that the penalized least squares functional in Equation (3) can be written as

$$\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}\|^2 + \lambda \boldsymbol{\gamma}^\top \mathbf{Q} \boldsymbol{\gamma}, \tag{6}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is the response vector, $\mathbf{X} = [\phi_j(x_i)]$ is the null space basis function matrix, $\mathbf{Z} = [\kappa(x_i, x_k^*)]$ is the contrast space basis function matrix, and $\mathbf{Q} = [\kappa(x_j^*, x_k^*)]$ is the penalty matrix. Given λ, the optimal coefficient estimates have the form

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_\lambda \\ \hat{\boldsymbol{\gamma}}_\lambda \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{X} & \mathbf{Z}^\top\mathbf{Z} + n\lambda\mathbf{Q} \end{bmatrix}^\dagger \begin{bmatrix} \mathbf{X}^\top \\ \mathbf{Z}^\top \end{bmatrix} \mathbf{y}, \tag{7}$$

where $\mathbf{A}^\dagger$ denotes the Moore–Penrose pseudoinverse of $\mathbf{A}$ [32,33]. Note that the coefficient estimates in Equation (7) are unique if the selected knots satisfy $0 \le x_1^* < \cdots x_r^* < 1$, in which case the matrix $\mathbf{Z}$ is a full column rank (assuming that $r < n$).

In nonparametric regression, we are not typically interested in the values of the coefficients. Instead, we are interested in the fitted values, which have the form

$$\hat{\mathbf{y}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda + \mathbf{Z}\hat{\boldsymbol{\gamma}}_\lambda = \mathbf{S}_\lambda \mathbf{y}, \tag{8}$$

where

$$\mathbf{S}_\lambda = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{X} & \mathbf{Z}^\top\mathbf{Z} + n\lambda\mathbf{Q} \end{bmatrix}^\dagger \begin{bmatrix} \mathbf{X}^\top \\ \mathbf{Z}^\top \end{bmatrix} \tag{9}$$

is the "smoothing matrix", which is the nonparametric regression analogue of the "hat matrix" in linear models. Note that the fitted values and smoothing matrix are subscripted with λ, given that the coefficient estimates (and, consequently, $\hat{\mathbf{y}}_\lambda$ and $\mathbf{S}_\lambda$) change as a function of the smoothing parameter. The trace of the smoothing matrix is denoted by $\nu_\lambda = \mathrm{tr}(\mathbf{S}_\lambda)$, which is often referred to as the effective degrees of freedom of the function estimate.

2.3. Bayesian Inference

It is well-known that the smoothing spline estimator η_λ can be interpreted as a Bayesian estimate of a Gaussian process [34,35]. To arrive at the Bayesian interpretation, first define $\eta_0(x) = \boldsymbol{\phi}^\top \boldsymbol{\beta}$, where $\boldsymbol{\phi}^\top = (\phi_0(x), \ldots, \phi_{m-1}(x))$ is the null space basis

functions evaluated at x, and define $\eta_1(x) = \boldsymbol{\kappa}^\top \boldsymbol{\gamma}$, where $\boldsymbol{\kappa}^\top = (\kappa(x, x_1^*), \ldots \kappa(x, x_r^*))$ is the contrast space reproducing the kernel function evaluated at x and the selected knots. Next, assume the prior distributions $\boldsymbol{\beta} \sim N(\mathbf{0}, \tau^2\mathbf{I})$, and $\boldsymbol{\gamma} \sim N(\mathbf{0}, \theta^2\mathbf{Q}^\dagger)$, where $\theta^2 = \frac{\sigma^2}{n\lambda}$, and assume that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are independent of one another and are independent of the ϵ_i terms. Defining $\boldsymbol{\alpha} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ as the combined coefficient vector, the prior assumptions imply that $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\alpha)$, where $\boldsymbol{\Sigma}_\alpha = \text{bdiag}(\tau^2\mathbf{I}, \theta^2\mathbf{Q}^\dagger)$ is a block diagonal covariance matrix. Assuming that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, the prior distributions imply that the (unconditional) distribution of the response vector is $\mathbf{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_y)$, where $\boldsymbol{\Sigma}_y = \tau^2\mathbf{X}\mathbf{X}^\top + \theta^2\mathbf{Z}\mathbf{Q}^\dagger\mathbf{Z}^\top + \sigma^2\mathbf{I}$. This also implies that the covariance between $\mathbf{y}$ and $\boldsymbol{\alpha}$ has the form $\boldsymbol{\Sigma}_{y\alpha} = \left[\tau^2\mathbf{X}, \theta^2\mathbf{Z}\mathbf{Q}^\dagger\right]$.

Given these assumptions, the posterior distribution of $\boldsymbol{\alpha}$ given $\mathbf{y}$ is multivariate normal $(\boldsymbol{\alpha}|\mathbf{y}) \sim N(\boldsymbol{\mu}_{\alpha|y}, \boldsymbol{\Sigma}_{\alpha|y})$ with mean vector and covariance matrix

$$\begin{aligned} \boldsymbol{\mu}_{\alpha|y} &= \boldsymbol{\Sigma}_{y\alpha}^\top \boldsymbol{\Sigma}_y^{-1}\mathbf{y} \\ \boldsymbol{\Sigma}_{\alpha|y} &= \boldsymbol{\Sigma}_\alpha - \boldsymbol{\Sigma}_{y\alpha}^\top \boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{y\alpha}, \end{aligned} \tag{10}$$

which is a well-known property of the multivariate Gaussian distribution (e.g., see [36]). Defining $\theta^2 = \frac{\sigma^2}{n\lambda}$, it can be shown that, as $\tau^2 \to \infty$, we have the relations

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\alpha|y} &= \lim_{\tau^2 \to \infty} \boldsymbol{\mu}_{\alpha|y} = (\mathbf{M}^\top\mathbf{M} + n\lambda\mathbf{Q}^*)^\dagger \mathbf{M}^\top\mathbf{y} \\ \hat{\boldsymbol{\Sigma}}_{\alpha|y} &= \lim_{\tau^2 \to \infty} \boldsymbol{\Sigma}_{\alpha|y} = \sigma^2(\mathbf{M}^\top\mathbf{M} + n\lambda\mathbf{Q}^*)^\dagger \end{aligned} \tag{11}$$

where $\mathbf{M} = [\mathbf{X}, \mathbf{Z}]$ is the model matrix and $\mathbf{Q}^* = \text{bdiag}(\mathbf{0}, \mathbf{Q})$ is a block diagonal penalty matrix where the zeros correspond to the $\mathbf{X}$ portion of $\mathbf{M}$. Note that $\hat{\boldsymbol{\mu}}_{\alpha|y}$ is the coefficient estimates from Equation (7) and that $\hat{\boldsymbol{\Sigma}}_{\alpha|y}$ is the inner portion of the smoothing matrix from Equation (9). Thus, the smoothing spline estimator can be interpreted as a posterior mean estimator under the specified prior distribution assumptions.

The Bayesian interpretation of a smoothing spline can be useful for assessing the uncertainty of the predictions from a fit smoothing spline. First, note that the model predictions can be written as $\hat{\eta}_\lambda(x) = \boldsymbol{\phi}^\top\hat{\boldsymbol{\beta}}_\lambda + \boldsymbol{\kappa}^\top\hat{\boldsymbol{\gamma}}_\lambda = \boldsymbol{\xi}^\top\hat{\boldsymbol{\alpha}}_\lambda$ where $\boldsymbol{\xi}^\top = [\boldsymbol{\phi}^\top, \boldsymbol{\kappa}^\top]$ and $\hat{\boldsymbol{\alpha}}_\lambda = \hat{\boldsymbol{\mu}}_{\alpha|y}$. Using the results in Equation (11), the posterior distribution of $\eta(x)$ given $\mathbf{y}$ is univariate normal $(\eta(x)|\mathbf{y}) \sim N(\mu_{\eta(x)|y}, \sigma^2_{\eta(x)|y})$, where the posterior mean is $\mu_{\eta(x)|y} = \hat{\eta}_\lambda(x) = \boldsymbol{\xi}^\top\hat{\boldsymbol{\mu}}_{\alpha|y}$ and the posterior variance is $\sigma^2_{\eta(x)|y} = \boldsymbol{\xi}^\top\hat{\boldsymbol{\Sigma}}_{\alpha|y}\boldsymbol{\xi}$. This implies that the $100(1-\alpha)\%$ Bayesian "confidence interval" for $\eta(x)$ has the form

$$\hat{\eta}_\lambda(x) \pm Z_{1-\alpha/2}\sigma_{\eta(x)|y}, \tag{12}$$

where $Z_{1-\alpha/2}$ denotes the standard normal critical value that cuts off $\alpha/2$ in the upper tail. When the model assumptions are reasonable, the Bayesian confidence intervals tend to have "across the function" coverage properties, such that the $100(1-\alpha)\%$ confidence interval is expected to contain about $100(1-\alpha)\%$ of the true $\eta(x)$ values [34,35].

3. Tuning Methods

3.1. Cross-Validation Methods

Ordinary cross-validation (OCV), which is also referred to as leave-one-out cross-validation, is a special case of k-fold cross-validation where k (the number of folds) is equal to n (the sample size). This means that each observation has a turn being the "test set" or "test observation" since only one observation is held at a time. The use of OCV for model selection and model assessment was independently discovered by Allen [37] and Stone [38]

in the context of regression. Wahba and Wold [39] later suggested the use of OCV when fitting smoothing spline models. The OCV method seeks to find the λ that minimizes

$$\text{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \eta_\lambda^{[i]}(x_i)\right)^2, \tag{13}$$

where $\eta_\lambda^{[i]}(x_i)$ is the function that minimizes the penalized least squares functional, leaving out the i^{th} data pair (x_i, y_i). More specifically,

$$\eta_\lambda^{[i]} = \min_{\eta\in\mathcal{H}} \frac{1}{n}\sum_{j=1, j\neq i}^{n}(y_j - \eta(x_j))^2 + \lambda J_m(\eta) \tag{14}$$

is the minimizer of the leave-one-out version of the penalized least squares functional.

The form of the OCV given in Equation (13) suggests that evaluating the OCV criterion for a given λ requires fitting the model n different times (once for each x_i). Fortunately, the leave-one-out function evaluation can be written in terms of the solution fit to the full sample of data, such as

$$\eta_\lambda^{[i]}(x_i) = \frac{\eta_\lambda(x_i) - s_{ii(\lambda)}y_i}{1 - s_{ii(\lambda)}}, \tag{15}$$

where $s_{ii(\lambda)}$ is the i^{th} diagonal element of $\mathbf{S}_\lambda$ [16]. This implies that the OCV criterion can be evaluated for a given λ via a single fitting of the model. Plugging the form of $\eta_\lambda^{[i]}(x_i)$ into the OCV criterion in Equation (13) produces

$$\text{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \eta_\lambda(x_i)}{1 - s_{ii(\lambda)}}\right)^2, \tag{16}$$

which is the computational form of the OCV criterion.

The computational form of the OCV in Equation (16) reveals that the OCV can be interpreted as a weighted least squares criterion, where the weights are defined $(1 - s_{ii(\lambda)})^{-2}$. The leverage scores satisfy $s_{ii(\lambda)} \in (0, 1)$ so each observation can have a notably different amount of influence on the OCV criterion. To equalize the influence of the observations on the smoothing parameter selection, the generalized cross-validation (GCV) criterion of Craven and Wahba [40] replaces the leverage scores with their average value. More specifically, the GCV method seeks to find the λ that minimizes

$$\text{GCV}(\lambda) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \eta_\lambda(x_i))^2}{(1 - \nu_\lambda/n)^2}, \tag{17}$$

where $\nu_\lambda = \text{tr}(\mathbf{S}_\lambda)$ is the effective degrees of freedom of the estimator η_λ. The GCV criterion is typically preferred over the OCV criterion, especially when there are replicate x_i scores in the sample. Furthermore, assuming that $\epsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$, the GCV is known to have desirable asymptotic properties (e.g., see [41]).

3.2. Information Theory Methods

The information theoretic approaches for selecting λ require more assumptions than are required by the cross-validation based methods. More specifically, the information theory methods assume that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, which makes it possible to explicitly define the likelihood of the generated data (under the assumption that $\eta(x)$ is an unknown constant given x). The assumption of iid Gaussian error terms implies that the distribution of the response vector is $\mathbf{y} \sim N(\boldsymbol{\eta}, \sigma^2\mathbf{I})$, where the mean vector is $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$. Given a sample

of n independent observations and assuming that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, the log-likelihood function has the form

$$l(\lambda, \sigma^2) = -\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \eta_\lambda(x_i))^2 + n\log(\sigma^2) + n\log(2\pi)\right] \tag{18}$$

which depends on λ and the error variance σ^2. The maximum likelihood estimate of the error variance has the form $\hat{\sigma}^2_\lambda = \frac{1}{n}\sum_{i=1}^{n}(y_i - \eta_\lambda(x_i))^2$, and plugging this into the log-likelihood function has the form

$$\tilde{l}(\lambda) = l\left(\lambda, \hat{\sigma}^2_\lambda\right) = -\frac{1}{2}\left[n + n\log(\hat{\sigma}^2_\lambda) + n\log(2\pi)\right], \tag{19}$$

which only depends on λ through $\hat{\sigma}^2_\lambda$.

An information criterion (AIC) was proposed by Akaike [42] to compare a set of candidate models, with the goal being to select the model that loses the least amount of information about the (unknown) true data generating process. The AIC method for selecting λ involves selecting the λ that minimizes

$$\text{AIC}(\lambda) = -2\tilde{l}(\lambda) + 2\nu_\lambda, \tag{20}$$

where $\nu_\lambda = \text{tr}(\mathbf{S}_\lambda)$ is the effective degrees of freedom of the function estimate. Note that it is possible to use other degrees of freedom estimates for η_λ; however, we prefer the ν_λ estimate given that this estimate is used for the GCV criterion. The Bayesian information criterion (BIC) proposed by Schwarz [43] has the form

$$\text{BIC}(\lambda) = -2\tilde{l}(\lambda) + \log(n)\nu_\lambda, \tag{21}$$

which is similar to the AIC but uses a different weight on the penalty. Assuming that $n \geq 8$, the BIC penalty weight of $\log(n)$ is larger than the AIC penalty weight of 2, which implies that the BIC will tends to select larger values of λ (i.e., smoother models).

3.3. Maximum Likelihood Methods

The maximum likelihood approaches for selecting λ exploit the computational relationship between penalized splines and linear mixed effects models [24,44,45]. This approach uses similar arguments to the Bayesian confidence intervals with the exception that the null space coefficients are treated as fixed effects. More specifically, assume that $\boldsymbol{\gamma} \sim N(\mathbf{0}, \frac{\sigma^2}{n\lambda}\mathbf{Q}^\dagger)$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and assume that $\boldsymbol{\gamma}$ is independent of $\epsilon_i \; \forall i$. This implies that the response vector is $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}_\lambda)$, where the null space representation contains the "fixed effects" and the contrast space representation contains the "random effects", with covariance matrix proportional to $\boldsymbol{\Sigma}_\lambda = \frac{1}{n\lambda}\mathbf{Z}\mathbf{Q}^\dagger\mathbf{Z}^\top + \mathbf{I}$. Given a sample of n independent observations, the log-likelihood function has the form

$$L(\lambda, \sigma^2) = -\frac{1}{2}\left[\sigma^{-2}\mathbf{r}^\top\boldsymbol{\Sigma}_\lambda^{-1}\mathbf{r} + \log(|\boldsymbol{\Sigma}_\lambda|) + n\log(\sigma^2) + n\log(2\pi)\right], \tag{22}$$

where $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. The maximum likelihood estimate for σ^2 has the form $\hat{\sigma}^2_{\lambda(\text{ML})} = \frac{1}{n}\mathbf{r}^\top\boldsymbol{\Sigma}_\lambda^{-1}\mathbf{r}$, and plugging this into the log-likelihood function produces

$$\text{ML}(\lambda) = L\left(\lambda, \hat{\sigma}^2_{\lambda(\text{ML})}\right) = -\frac{1}{2}\left[n + \log(|\boldsymbol{\Sigma}_\lambda|) + n\log(\mathbf{r}^\top\boldsymbol{\Sigma}_\lambda^{-1}\mathbf{r}) + n\log(2\pi/n)\right], \tag{23}$$

which is the (profile) maximum likelihood criterion for selecting λ.

Restricted maximum likelihood (REML) estimation takes into account the reduction in degrees of freedom due to estimating the m null space coefficients [46]. The REML log-likelihood function has the form

$$R(\lambda, \sigma^2) = L(\lambda, \sigma^2) - \frac{1}{2}\Big[\log(|\mathbf{X}^\top \boldsymbol{\Sigma}_\lambda^{-1}\mathbf{X}|) - m\log(2\pi\sigma^2)\Big], \tag{24}$$

and the REML estimate for σ^2 has the form $\hat{\sigma}^2_{\lambda(\text{REML})} = \frac{1}{n-m}\mathbf{r}^\top \boldsymbol{\Sigma}_\lambda^{-1}\mathbf{r}$. Plugging this error variance estimate into the log-likelihood function produces the (profile) REML criterion for selecting λ, which has the form

$$\text{REML}(\lambda) = -\frac{1}{2}\Big[\tilde{n} + \log(|\boldsymbol{\Sigma}_\lambda|) + \tilde{n}\log(\mathbf{r}^\top \boldsymbol{\Sigma}_\lambda^{-1}\mathbf{r}) + \tilde{n}\log(2\pi/\tilde{n}) + \log(|\mathbf{X}^\top \boldsymbol{\Sigma}_\lambda^{-1}\mathbf{X}|)\Big], \tag{25}$$

where $\tilde{n} = n - m$ is the degrees of freedom of $\hat{\sigma}^2_{\lambda(\text{REML})}$. Note that the REML criterion is generally preferred over the ML criterion for variance component estimation, particularly when the sample size is small to moderate.

4. Simulation Study

4.1. Simulation Design

To investigate the performance of the tuning parameter selection methods discussed in the previous section, we designed a simulation study that compares the methods under a variety of different data generating conditions. More specifically, we designed a fully crossed simulation study that compares the performance of the tuning methods under all combinations of five different design factors: the function smoothness (three levels: later defined), the error standard deviation (three levels: constant, increasing, and parabolic), the error correlation (three levels: $\rho \in \{0, 1/3, 2/3\}$), the error distribution (three levels: normal, t_5, and uniform), and the sample size (four levels: $n \in \{50, 100, 200, 400\}$). For each combination of data generating conditions, the predictor scores were defined as $x_i = \frac{i-1}{n-1}$ for $i = 1, \ldots, n$.

The three levels of the function smoothness are depicted in Figure 1 (left) and are from the simulation studies of Wahba [34]. The three levels of the error standard deviation are depicted in Figure 1 (right) and are defined as $\sigma(x) = 1$ (constant), $\sigma(x) = x + 1/2$ (increasing), and $\sigma(x) = 4(x - 1/2)^2 + 1/2$ (parabolic). To generate correlated errors, we use an autoregressive process of order one, i.e., AR(1) process, so that $\text{Cor}(x_i, x_j) = \rho^{|i-j|}$ for all i, j. The generation of correlated multivariate normal (or t) data is straightforward given that the AR(1) correlation structure can be incorporated into the covariance matrix. To generate correlated uniform data, we use the method proposed by Falk [47], which produces uniformly distributed data with desired correlations.

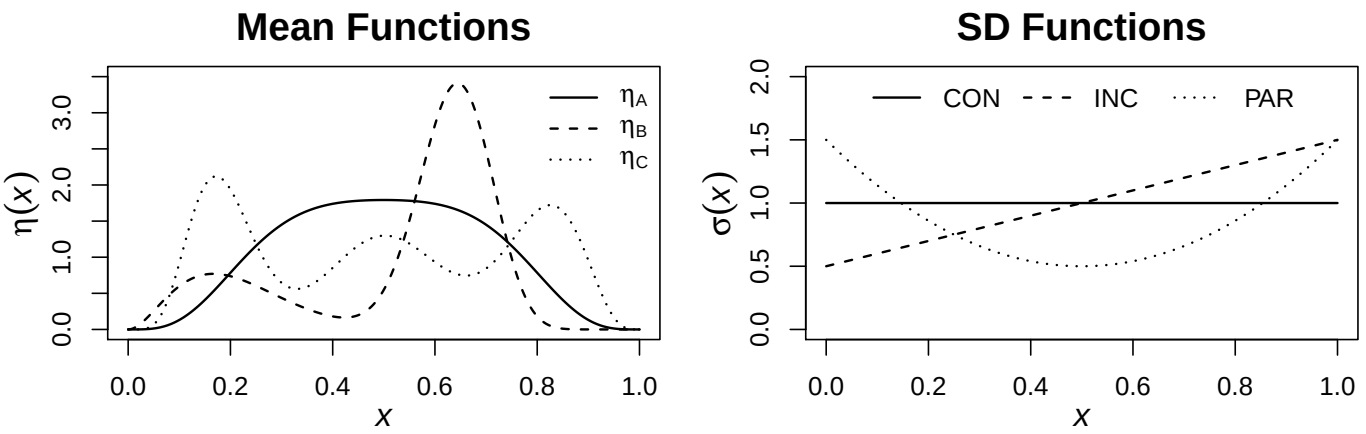

Figure 1. Simulation design functions. (**Left**): the three mean functions from Grace Wahba [34]. (**Right**): the three standard deviation functions: constant (CON), increasing (INC), and parabolic (PAR).

4.2. Simulation Analyses

For each of the 324 levels of the simulation design (3 η × 3 σ × 3 ρ × 3 F_e × 4 n), we generated data according to the model in Equation (2). Within each cell of the simulation design, we generated $R = 1000$ independent replications of the data. For each sample of generated data, we fit the model using the six smoothing parameter selection methods discussed in the previous section. The models were fit in R [20] via the `ss()` function in the **npreg** package [48] using $r = 20$ knots placed evenly across the range of the x_i scores. To evaluate the performance of the different tuning methods, we calculated the root mean squared error (RMSE) of the estimator, i.e.,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\eta(x_i) - \hat{\eta}_\lambda(x_i))^2} \tag{26}$$

so smaller values of RMSE indicate better recovery of the true mean function. We also calculated the coverage of the 95% Bayesian confidence interval for each tuning method

$$\text{Coverage} = \frac{1}{n}\sum_{i=1}^{n} I\{a(\hat{\eta}_\lambda(x_i)) \leq \eta(x_i) \leq b(\hat{\eta}_\lambda(x_i))\} \tag{27}$$

where $I\{\cdot\}$ denotes an indicator function, $a(\hat{\eta}_\lambda(x_i)) = \hat{\eta}_\lambda(x_i) - 1.96\hat{\sigma}_\lambda\sqrt{s_{ii(\lambda)}}$ is the lower bound for the 95% Bayesian CI, and $b(\hat{\eta}_\lambda(x_i)) = \hat{\eta}_\lambda(x_i) + 1.96\hat{\sigma}_\lambda\sqrt{s_{ii(\lambda)}}$ is the upper bound for the 95% Bayesian CI. Note that, when evaluated at the n design points, the estimated posterior variance has the form $\hat{\sigma}^2_{\eta(x_i)|y} = \hat{\sigma}^2_\lambda s_{ii(\lambda)}$.

4.3. Simulation Results

Figure 2 displays the RMSE for each tuning method in each combination of data generating function η, autocorrelation ρ, and sample size n when the errors are Gaussian and homoscedastic. The results reveal that, for all combinations of η and ρ, all of the methods tend to result in better function recovery (i.e., smaller RMSE) as n increases, which was expected. The interesting finding is that the maximum likelihood-based methods (REML and ML) tend to produce RMSE values that are similar to or smaller than the RMSE values produced by the cross-validation-based methods (OCV and GCV) and the information theory-based methods (AIC and BIC). The only exception is that the cross-validation methods tend to outperform the maximum likelihood methods when all three of the following conditions are true: (i) the sample size is small, (ii) the mean function is rather jagged, and (iii) the errors are independent. When $\rho > 0$, the maximum likelihood methods produce substantially smaller RMSE values compared with the other methods. This effect exists for all combinations of η and n, but the RMSE difference diminishes as the function roughness and/or the sample size increases.

Figure 3 displays the coverage of the 95% Bayesian confidence intervals for each tuning method in each combination of data generating function η, autocorrelation ρ, and sample size n when the errors are Gaussian and homoscedastic. The results reveal that the performance of the Bayesian confidence intervals depends on the combination of the function smoothness, the error autocorrelation, the sample size, and the chosen tuning method. When there is no autocorrelation present, all of the tuning methods except the BIC tend to produce better coverage rates (i.e., closer to the nominal level) as the sample size n increases, regardless of the function smoothness. Furthermore, when there is no autocorrelation present, the maximum likelihood methods and the BIC result in noteworthy under-coverage when the function is jagged and the sample size is small. When there exists moderate autocorrelation in the errors (i.e., $\rho = 1/3$), all of the tuning methods result in noteworthy under-coverage, with the maximum likelihood methods performing better for smooth functions and the cross-validation methods performing better for more jagged functions. When the autocorrelation is larger (i.e., $\rho = 2/3$), all of the methods have similarly poor performance.

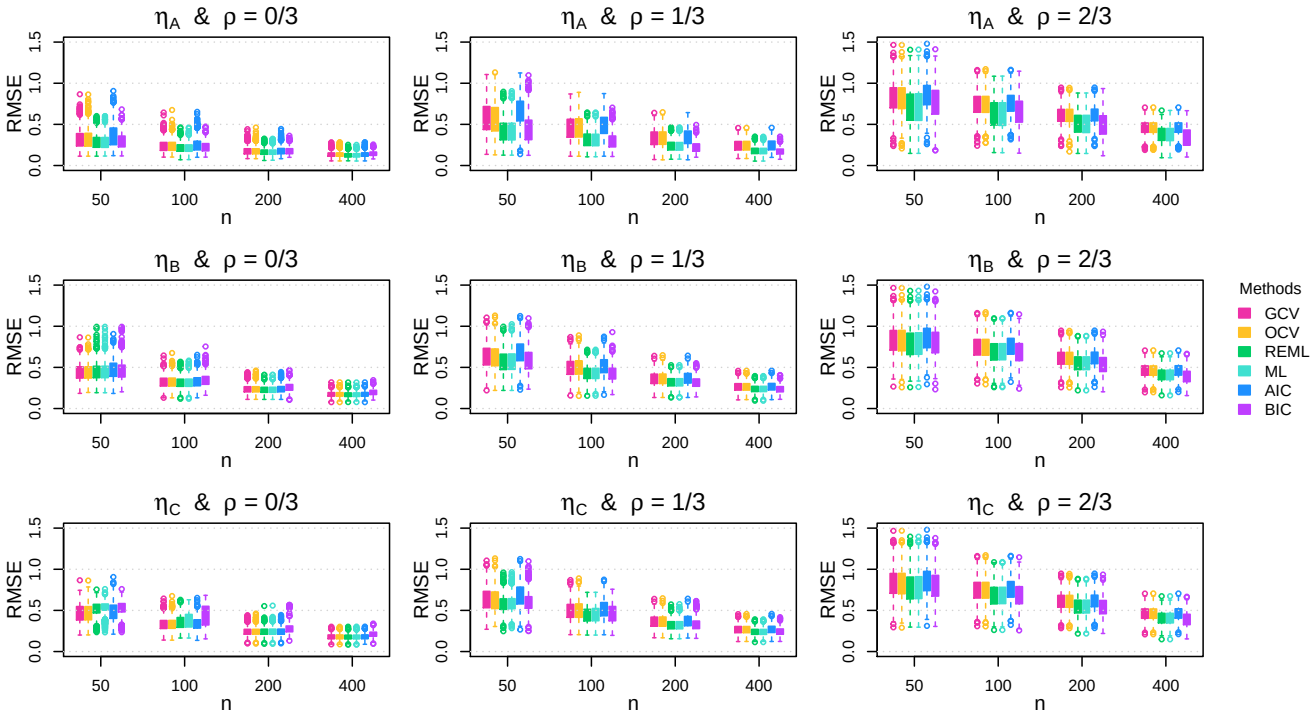

Figure 2. Simulation RMSE results. Boxplots of the RMSE across the 1000 simulation replications for homoscedastic Gaussian errors. The rows show the results for the mean functions, and the columns show the results for the autocorrelation parameters.

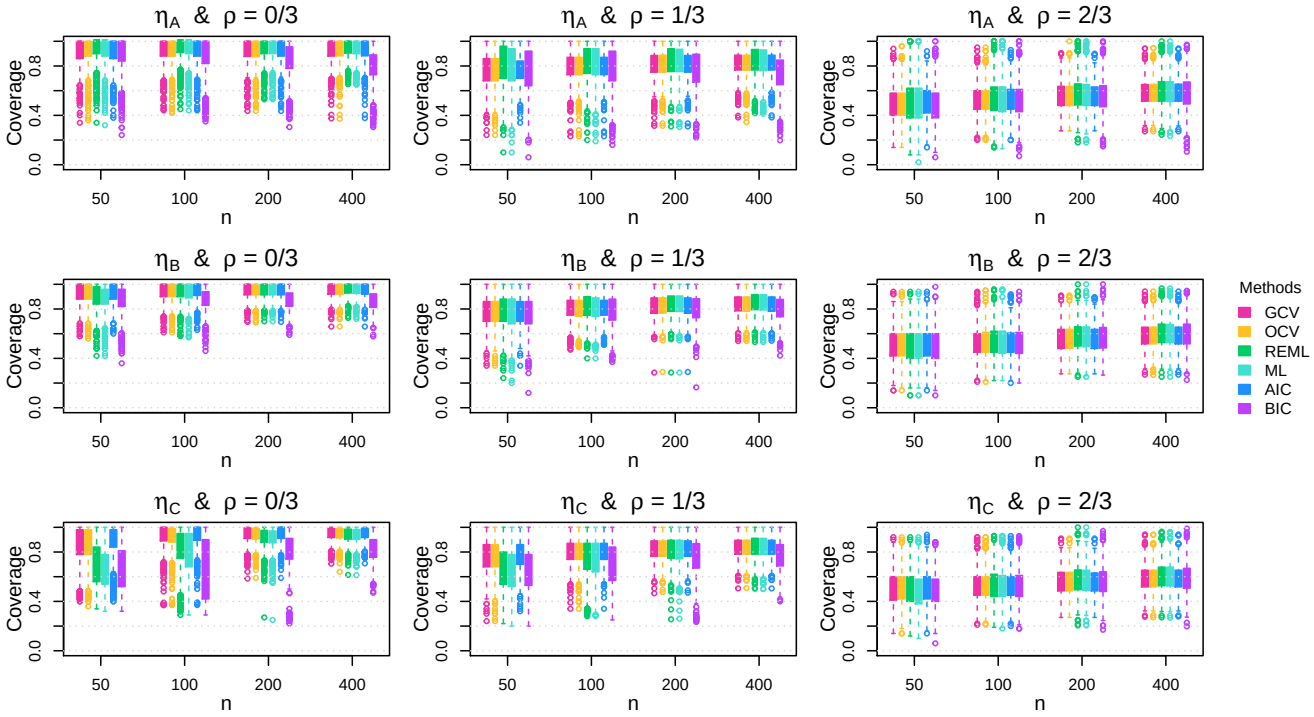

Figure 3. Simulation coverage results. Boxplots of the coverage rate for the 95% Bayesian confidence intervals across the 1000 simulation replications for homoscedastic Gaussian errors. The rows show the results for the mean functions, and the columns show the results for the autocorrelation parameters.

Our discussion of the results focused on a subset of the simulation conditions, which are sufficient for understanding the primary findings of the simulation study. Specifically, we focused on the results when the errors are Gaussian with constant variance, whereas

the results for non-Gaussian and heteroscedastic errors are presented in Appendix A (Gaussian), Appendix B (t_5), and Appendix C (uniform). Interestingly, we found that the RMSE and coverage results were quite similar for the non-Gaussian and heteroscedastic cases. In other words, the primary conclusions that were made about the results in Figures 2 and 3 also apply to the results for non-Gaussian distributions with non-constant variance. When the errors followed a multivariate t_5 distribution, the RMSE values tended to be a bit larger for all methods, which is not surprising. However, the primary conclusions regarding the effect of autocorrelation remained the same. It is rather interesting to note that the REML criterion tends to perform relatively well—especially in the presence of autocorrelation in the errors—even when the errors do not follow a Gaussian distribution. Consequently, the REML criterion seems to be a reasonable default tuning criterion as long as the sample size is not too small.

5. Real Data Examples

5.1. Global Warming Example

Our first example uses global land–ocean temperature data, which are freely available from NASA's Goddard Institute for Space Studies [49,50]. The dataset contains the global land–ocean temperature index from the years 1880 to 2020. Note that the global land–ocean temperature index is the change in global surface temperatures (in degree Celsius) relative to the 1951–1980 average temperatures. Positive values indicate that the average temperature for a given year was above the average temperature for the years 1951–1980, and negative values indicate that the average temperature for a given year was below the average temperature for the years 1951–1980. In our example, we compare the results of the trend estimate using the GCV and REML criteria for selecting the tuning parameter of the smoothing spline. We use the `.nknots.smspl` function in R [20] to select the number of knots, which results in the selection of 76 knots. For both tuning methods, we use the same 76 knot values to fit the cubic smoothing spline.

The results are plotted in Figure 4, which reveals that the GCV and REML tuning methods produce drastically different pictures of the temperature change across time. Although both methods show that the global land–ocean temperature index has increased over time (particularly since the 1970s), the GCV and REML solutions tell notably different stories about the nature of the increase. The GCV solution has an estimated degrees of freedom of $\hat{v}_\lambda = 47.52$ and suggests that the temperature index is rather volatile from year to year. In contrast, the REML solution has an estimated degrees of freedom of $\hat{v}_\lambda = 11.43$ and suggests that the temperature index changes in a rather smooth fashion from year to year. Based on the simulation results, it seems likely that the GCV criterion is under-smoothing the data, which may have some noteworthy autocorrelation in the error structure. Consequently, we contend that the REML solution should be preferred for interpreting the nature of the yearly changes in the global land–ocean temperature index.

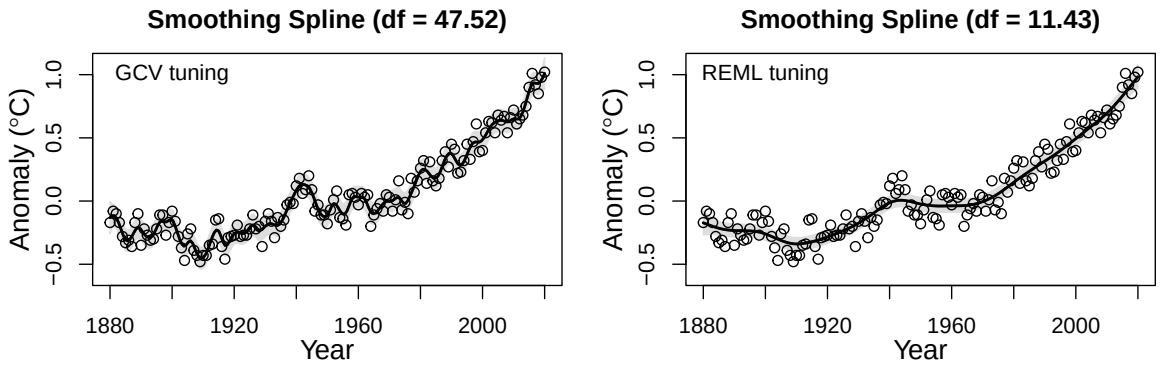

Figure 4. Global warming results. Smoothing spline solution for temperature data using GCV tuning (**left**) and REML tuning (**right**). Created using the `ss()` function in the **npreg** R package [48].

5.2. Motorcycle Accident Example

Our second example uses acceleration data from a simulated motorcycle accident. This dataset was first considered by Silverman [51] and has since been popularized via its inclusion in the popular **MASS** R package (see `mcycle`, [52]). The dataset contains the head acceleration (in g) as a function of time (in milliseconds) for $n = 133$ points of simulated data. The data are meant to simulate the acceleration curve of the head after a motorcycle accident and were simulated for the purpose of evaluating motorcycle helmets. Due to the simulation procedure, the resulting data are noisy realizations of the true acceleration curve, so the data need to be smoothed in order to estimate the expected head acceleration as a function of time. As in the previous example, (i) we compare the results of the curve estimate using the GCV and REML criteria for selecting the tuning parameter of the smoothing spline and (ii) we use the `.nknots.smspl` function in R to select the number of knots, which resulted in the selection of 61 knots. For both tuning methods, we use the same 61 knot values to fit the cubic smoothing spline.

The results are plotted in Figure 5, which reveals that the GCV and REML tuning methods produce rather similar estimates in this case. The GCV criterion $\hat{v}_\lambda = 12.21$ selected a slightly smaller degree of freedom estimate compared with the REML solution $\hat{v}_\lambda = 13.86$. However, from a practical perspective, the two tuning methods result in smoothing spline estimates that produce the same interpretation of the acceleration curve. The estimated acceleration curve reveals that the head experiences negative acceleration (15–30 ms) followed by a rebound effect of positive acceleration (30–40 ms) before loosely stabilizing around zero (40+ ms). Finally, it is worth noting that the data in this example violate the homogeneity of variance assumption, which is required for the Bayesian confidence intervals. Therefore, although the GCV and REML tuning criteria may provide satisfactory estimates of the acceleration curve, the Bayesian confidence intervals do not provide satisfactory estimates of uncertainty—in this case, they suggest too much uncertainty in the estimate from 0 to 15 ms.

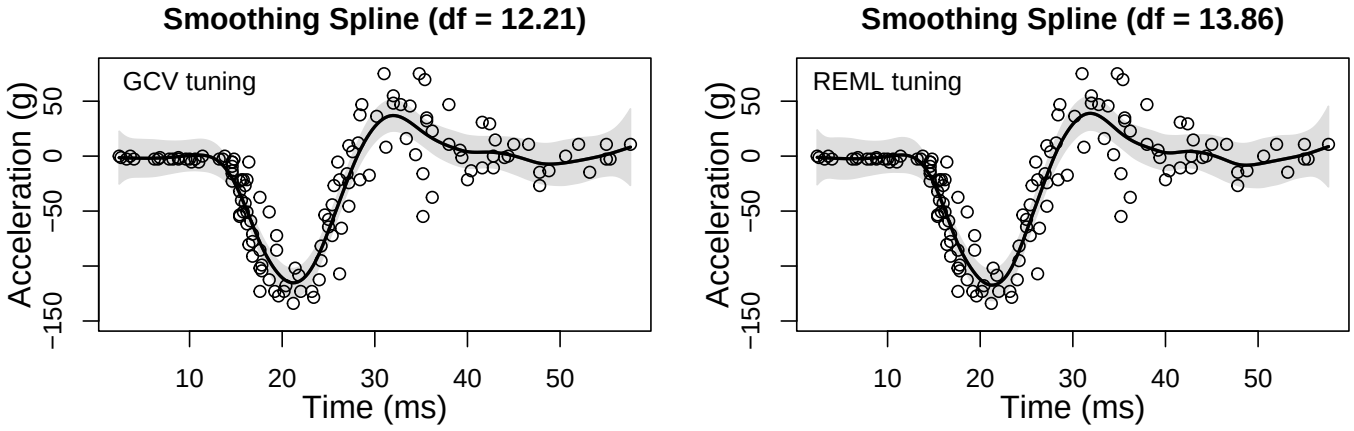

Figure 5. Motorcycle accident results. Smoothing spline solution for motorcycle data using GCV tuning (**left**) and REML tuning (**right**). Created using the `ss()` function in the **npreg** R package [48].

6. Discussion

6.1. Overview

Due to their unique combination of flexibility and interpretability, smoothing splines are frequently used to understand functional relationships in applied research. Unlike standard parametric regression methods (which make strict assumptions) and standard machine learning methods (which produce black-box predictions), smoothing splines are able (i) to learn functional forms and (ii) to produce insightful visualizations. To provide a valid estimate of unknown functional relationships, the smoothing spline estimator requires selecting a smoothing parameter that provides the "correct" balance

between fitting and smoothing the data. Specifically, the success of a smoothing spline depends on choosing the tuning parameter λ that satisfies a Goldilocks phenomenon: if λ is too small, the estimator has too much variance, and if λ is too large, the estimator has too much bias. Despite the well-known importance of selecting the "correct" λ, there is little agreement in the literature regarding which method should be used. Tuning methods can produce different results [26,27] so the choice of tuning method matters.

To address this issue, we explored the relative performance of six popular methods used to select the smoothing parameter λ. Unlike previous studies on this topic, (i) we compared a diverse collection of tuning methods, which included cross-validation, information theoretic, and maximum likelihood methods; (ii) we designed an extensive simulation study that evaluated each method's performance under a variety of data generating conditions; and (iii) we assessed the performance of the methods with respect to both function estimation and statistical inference. Furthermore, we used both simulated and real data examples to demonstrate the substantial differences that different smoothing methods can have on the solution. As we elaborate in the following paragraphs, the primary take-home message from our work is that any real data application should compare the results using both the GCV and REML smoothing parameter selection criteria—which is rarely performed in practice.

6.2. Summary of Results

Our simulation results replicate several important findings and provide novel insights about the performance of different tuning methods for smoothing splines. The finding that common tuning methods (i.e., OCV, GCV, and AIC) can breakdown in the presence of autocorrelated errors replicates several past studies on the topic [21,22]. Furthermore, our finding that the REML and ML tuning criteria are relatively robust in the presence of autocorrelated errors supports the theoretical results of Krivobokova and Kauermann [27]. In addition to replicating these known results, our simulation produced several important and novel findings: (i) the performance of the Bayesian confidence intervals deteriorates as the degree of autocorrelation increases; (ii) the cross-validation tuning methods only show an advantage over REML/ML when n is small, η is rough, and $\rho = 0$; and (iii) the superior performance of the REML/ML tuning methods persists even when the errors are non-Gaussian and/or heteroscedastic.

Our real data results demonstrate the important role that the smoothing parameter selection method plays in understanding functional relationships from a fit smoothing spline. Using the GCV versus the REML criterion, a researcher would arrive at a remarkably different interpretation of the global warming trends. Since these are real data, we cannot be sure whether the GCV or REML solution is closer to the truth. However, in this case, it seems likely that the GCV criterion has capitalized on autocorrelation in the error terms, which manifests itself as a part of the mean structure. In contrast, the REML solution seems to provide a rather parsimonious and intuitive estimate of the global warming trends, which agrees with visual intuition about the nature of the trends across time. Of course, the specifics of the global warming example do not generalize to other datasets, e.g., the two tuning methods performed similarly for the motorcycle data. However, this example illustrates how (i) the GCV criterion can under-smooth data and (ii) the REML criterion can overcome this under-smoothing issue.

6.3. Limitations and Future Directions

This paper only considers the nonparametric regression model in Equation (2), where the unknown function $\eta(\cdot)$ describes the conditional mean of Y given X. Accordingly, this paper only compares tuning methods that are applicable to the penalized least squares problem in Equation (3). Although quite general, the model in Equation (2) can be extended in a variety of different ways. For example, in a generalized nonparametric regression model, the function $\eta(\cdot)$ describes the conditional mean of an exponential family response variable as a function of a predictor [16]. As another example, in nonparametric quantile regression, the function $\eta(\cdot)$ describes the condi-

tional quantile of a response variable as a function of a predictor [53]. Furthermore, penalized splines can be incorporated into other types of nonparametric estimators, e.g., M-estimators [54]. These extensions require different estimation and tuning methods, so the results in this paper cannot be generalized to such extensions. Thus, future work is needed to determine which tuning methods should be preferred when penalized splines are used for various FDA extensions of the simple nonparametric regression model in Equation (2).

6.4. Conclusions

When using smoothing splines for real data analysis, it seems that many researchers do not give much thought to the smoothing parameter selection method. This is likely because many software implementations of smoothing splines do not emphasize the importance of the tuning method. Furthermore, most softwares only implement one or two tuning methods, so researchers rarely have the option to explore a multitude of tuning methods. For example, the `smooth.spline()` function in R [20] only offers the OCV and GCV tuning methods, and most users seem to (purposefully or unwittingly) use the default GCV tuning method. It is important to note that the GCV method is also the default in the **mgcv** R package [55], the **bigsplines** R package [56], and the **npreg** R package [48]; however, these packages offer more tuning options. For a flexible alternative to R's `smooth.spline()` function, we recommend the `ss()` function from the **npreg** package, given that it has nearly identical syntax to the `smooth.spline()` function and makes it possible to easily compare the results using multiple tuning criteria.

Author Contributions: Conceptualization, L.N.B. and N.E.H.; methodology, L.N.B. and N.E.H.; software, N.E.H.; validation, L.N.B. and N.E.H.; formal analysis, L.N.B. and N.E.H.; investigation, L.N.B. and N.E.H.; resources, L.N.B. and N.E.H.; data curation, L.N.B. and N.E.H.; writing—original draft preparation, L.N.B.; writing—review and editing, L.N.B. and N.E.H.; visualization, L.N.B. and N.E.H.; supervision, N.E.H.; project administration, N.E.H.; funding acquisition, N.E.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the following National Institutes of Health (NIH) grants: R01EY030890, R01MH115046, U01DA046413, and R01MH112583.

Data Availability Statement: The supporting materials published with this paper include the data and R code needed to reproduce the simulation and real data results. The global temperature anomaly data were obtained from https://data.giss.nasa.gov/gistemp/ accessed on 17 May 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OCV	Ordinary Cross-Validation
GCV	Generalized Cross-Validation
AIC	An Information Criterion
BIC	Bayesian Information Criterion
ML	Maximum Likelihood
REML	Restricted Maximum Likelihood
RMSE	Root Mean Squared Error

Appendix A. Supplementary Results for Gaussian Errors

Appendix A.1. RMSE Results

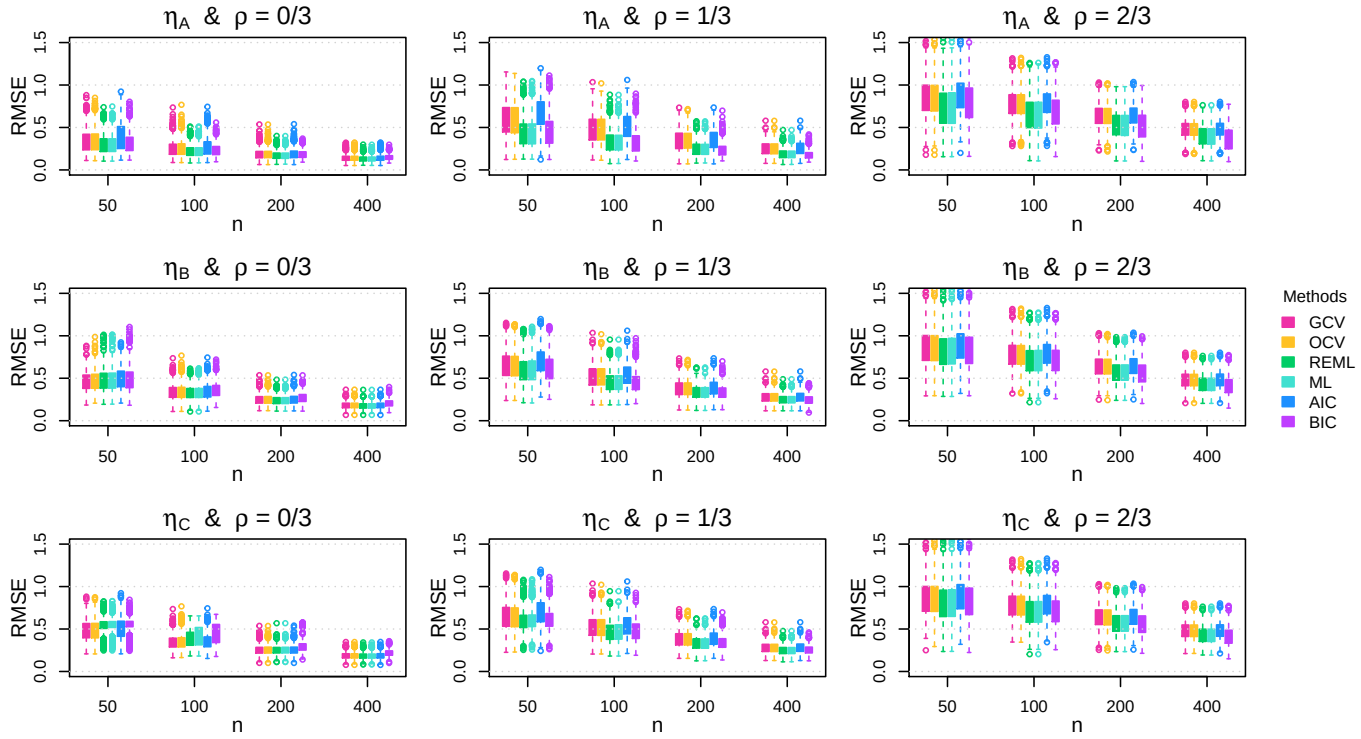

Figure A1. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for Gaussian errors with increasing standard deviation.

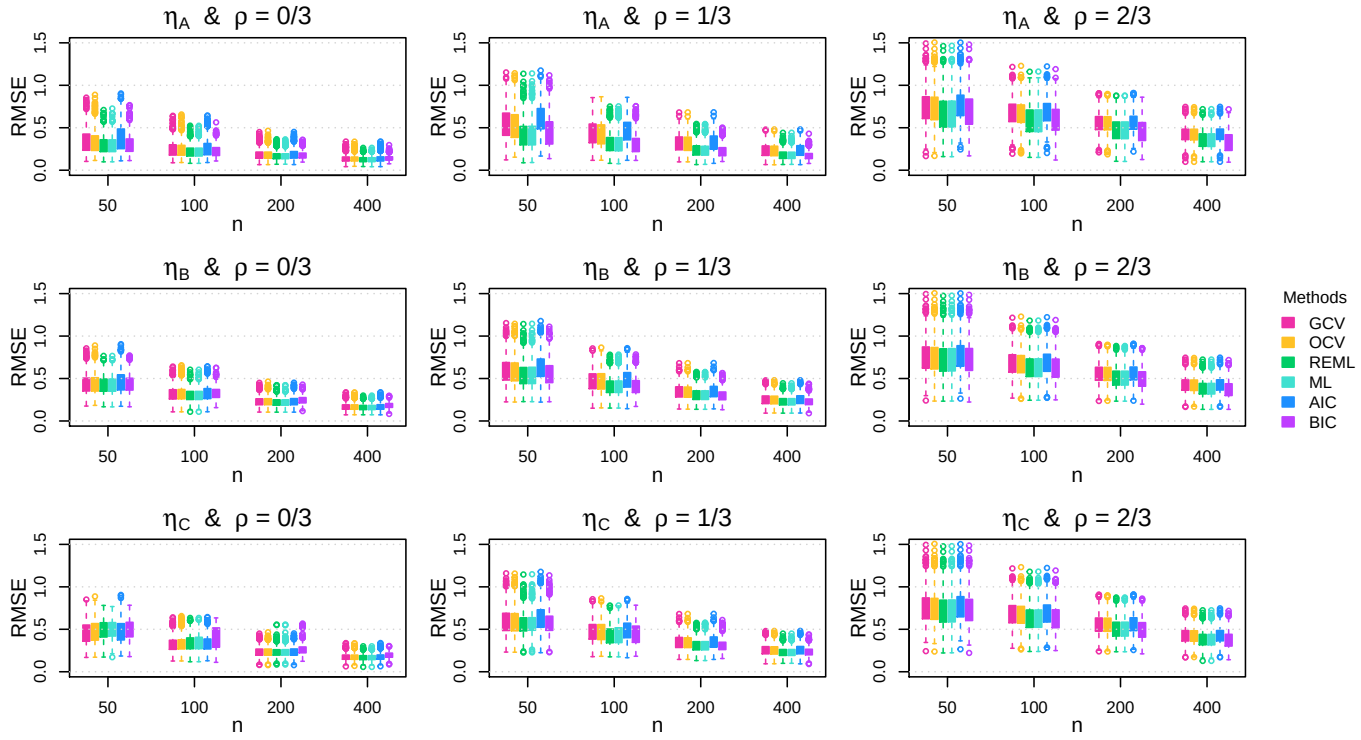

Figure A2. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for Gaussian errors with parabolic standard deviation.

Appendix A.2. Coverage Results

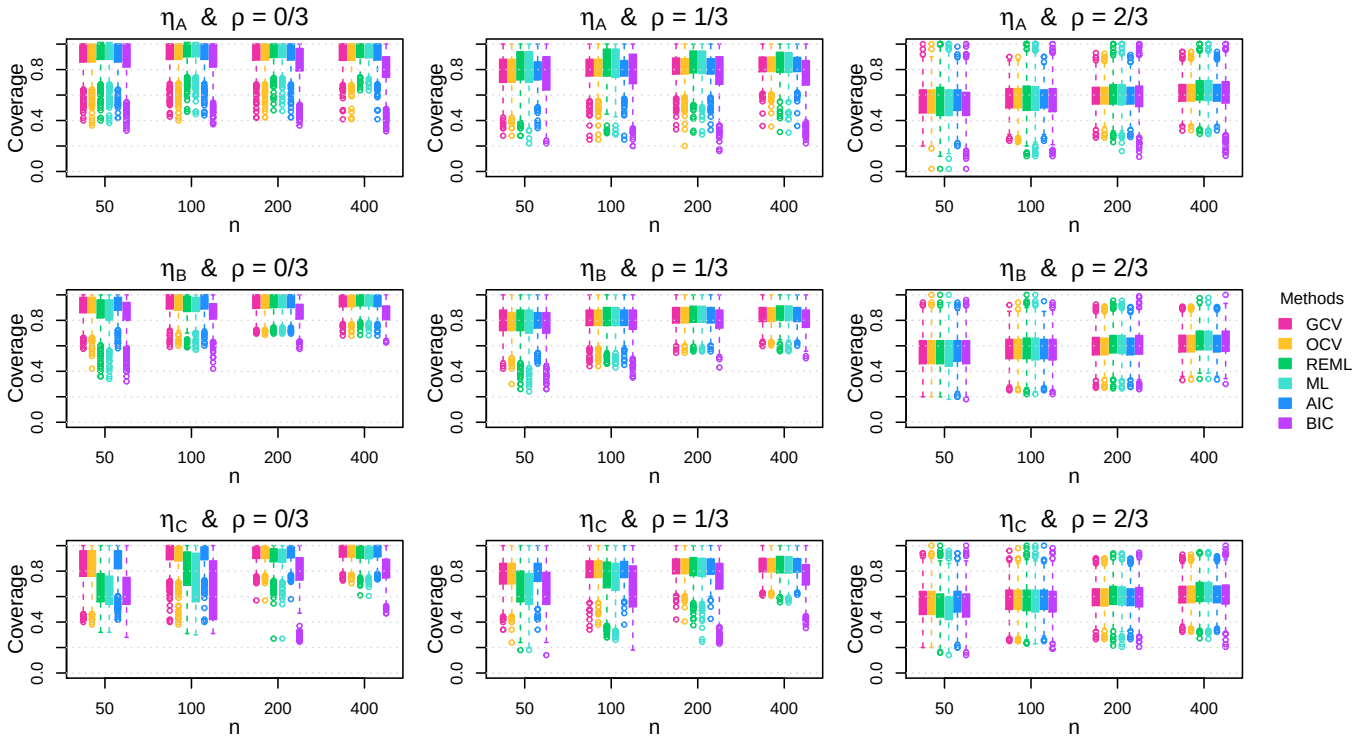

Figure A3. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for Gaussian errors with increasing standard deviation.

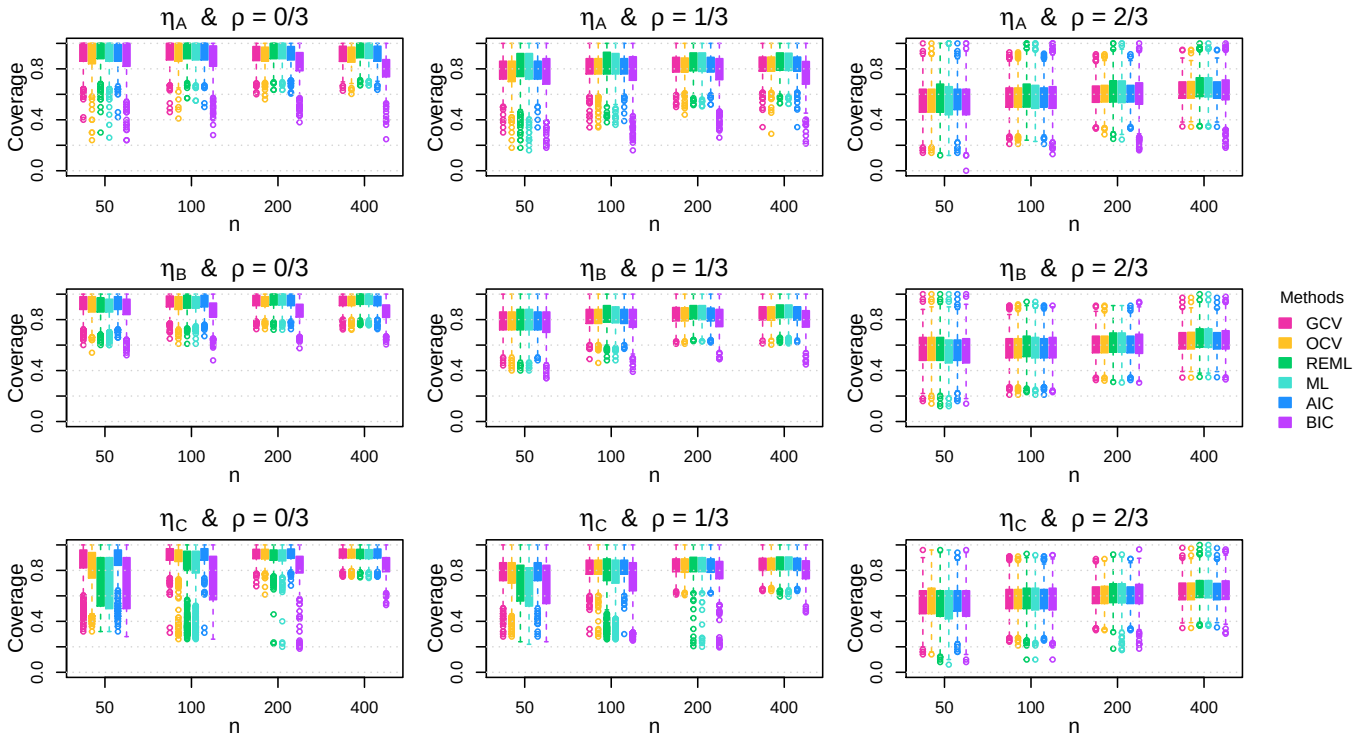

Figure A4. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for Gaussian errors with parabolic standard deviation.

Appendix B. Supplementary Results for Multivariate t_5 Errors

Appendix B.1. RMSE Results

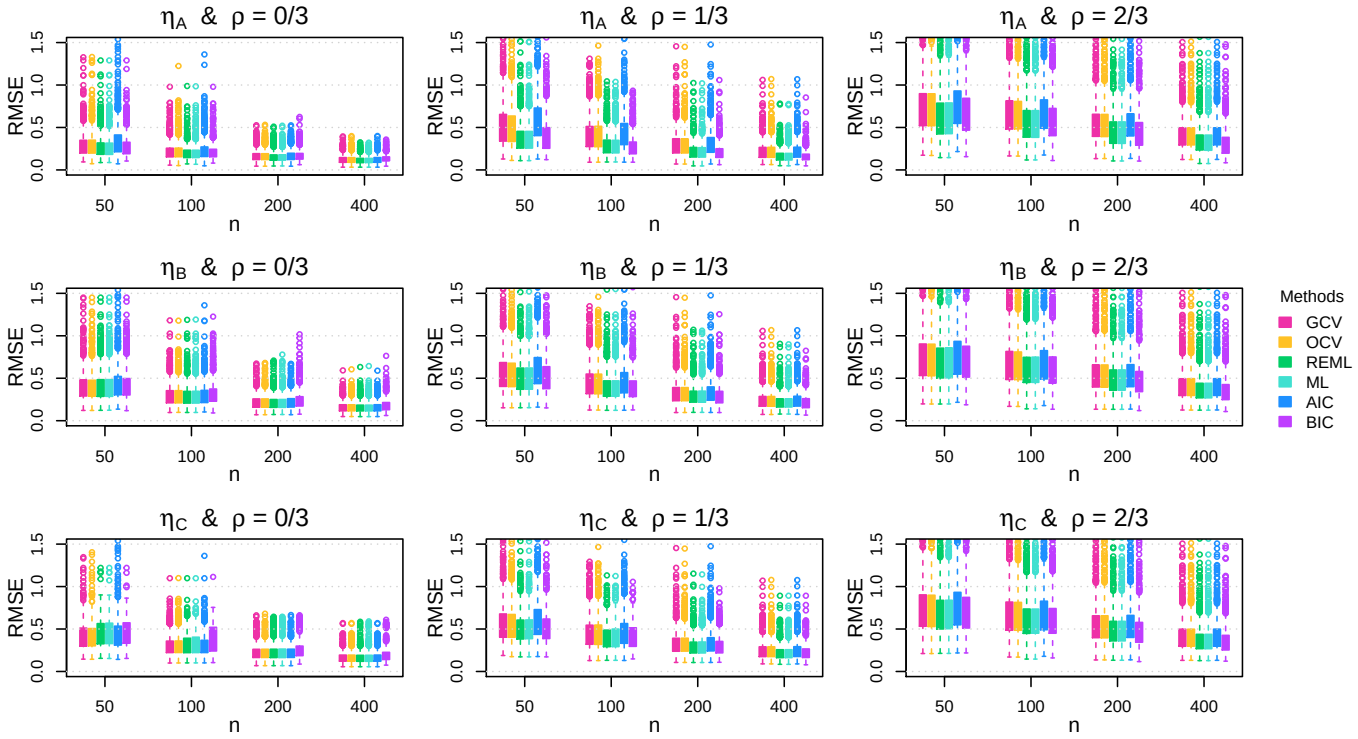

Figure A5. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for multivariate t_5 errors with constant standard deviation.

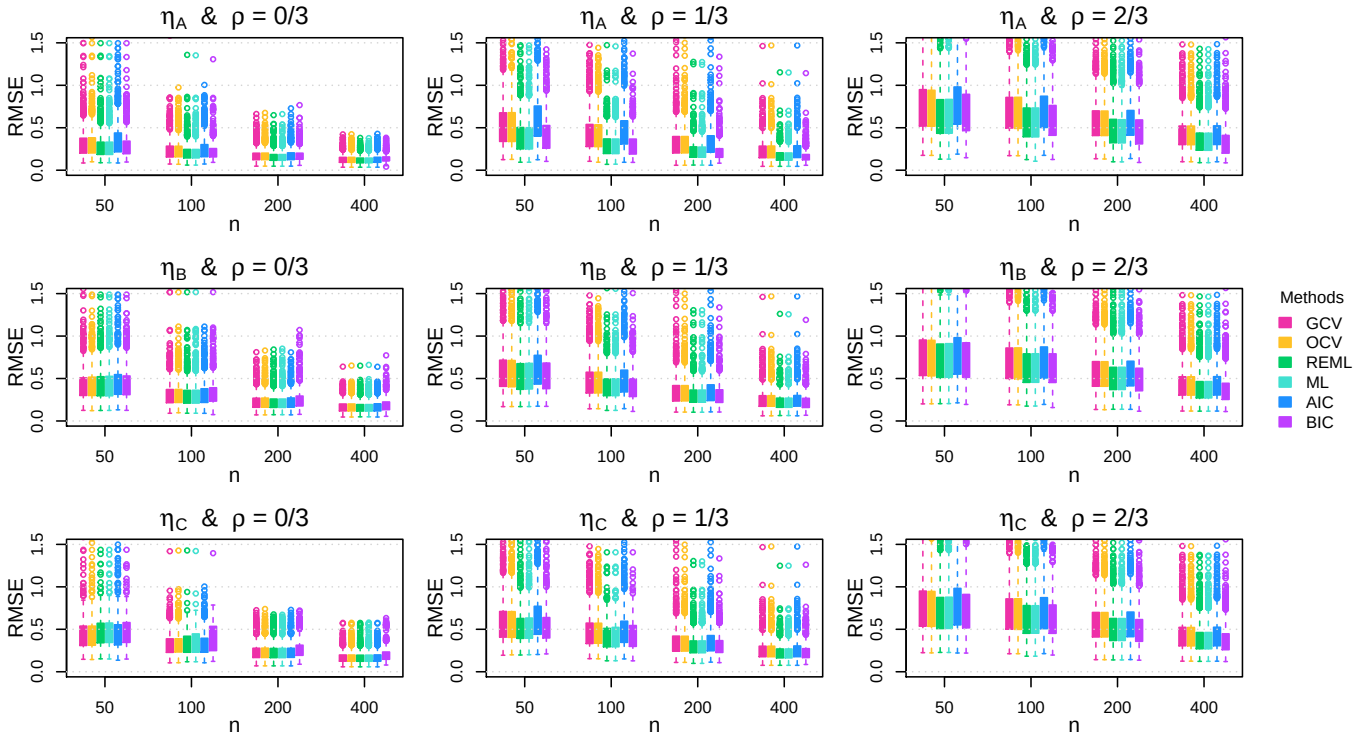

Figure A6. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for multivariate t_5 errors with increasing standard deviation.

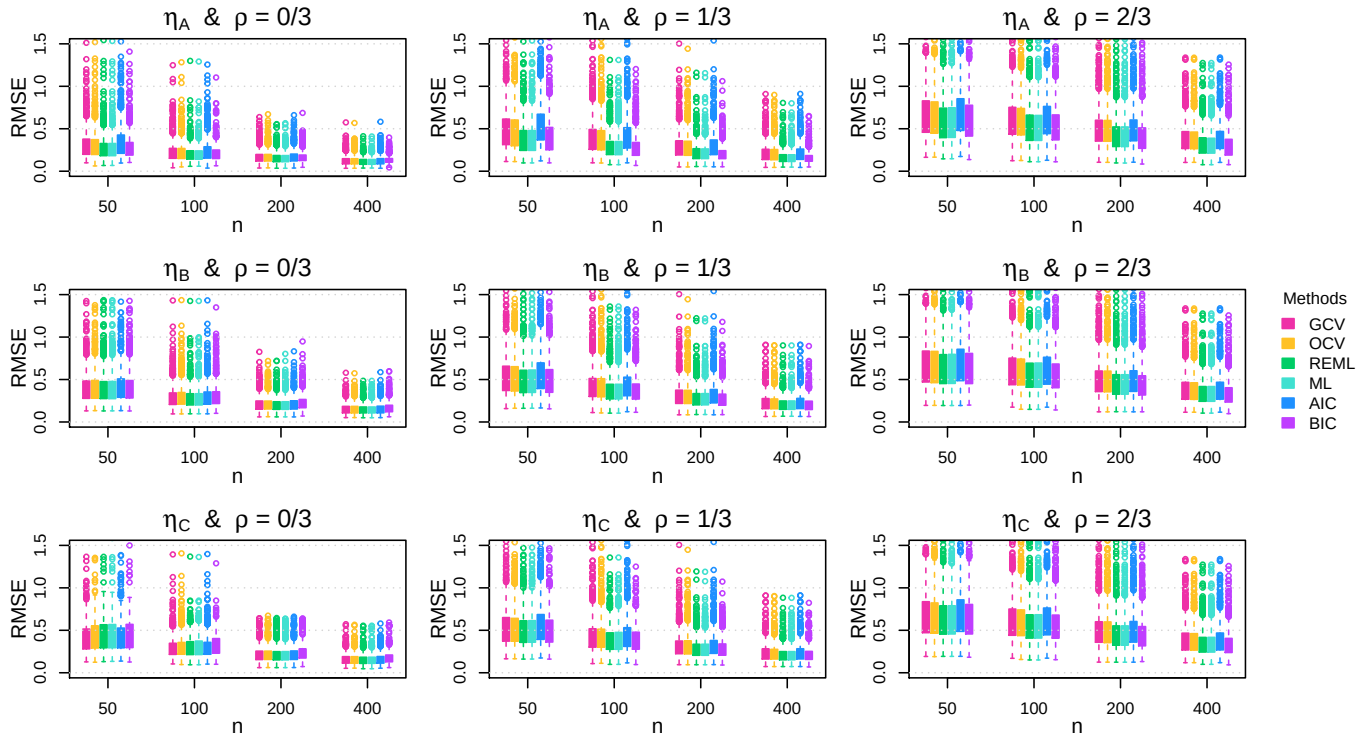

Figure A7. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for multivariate t_5 errors with parabolic standard deviation.

Appendix B.2. Coverage Results

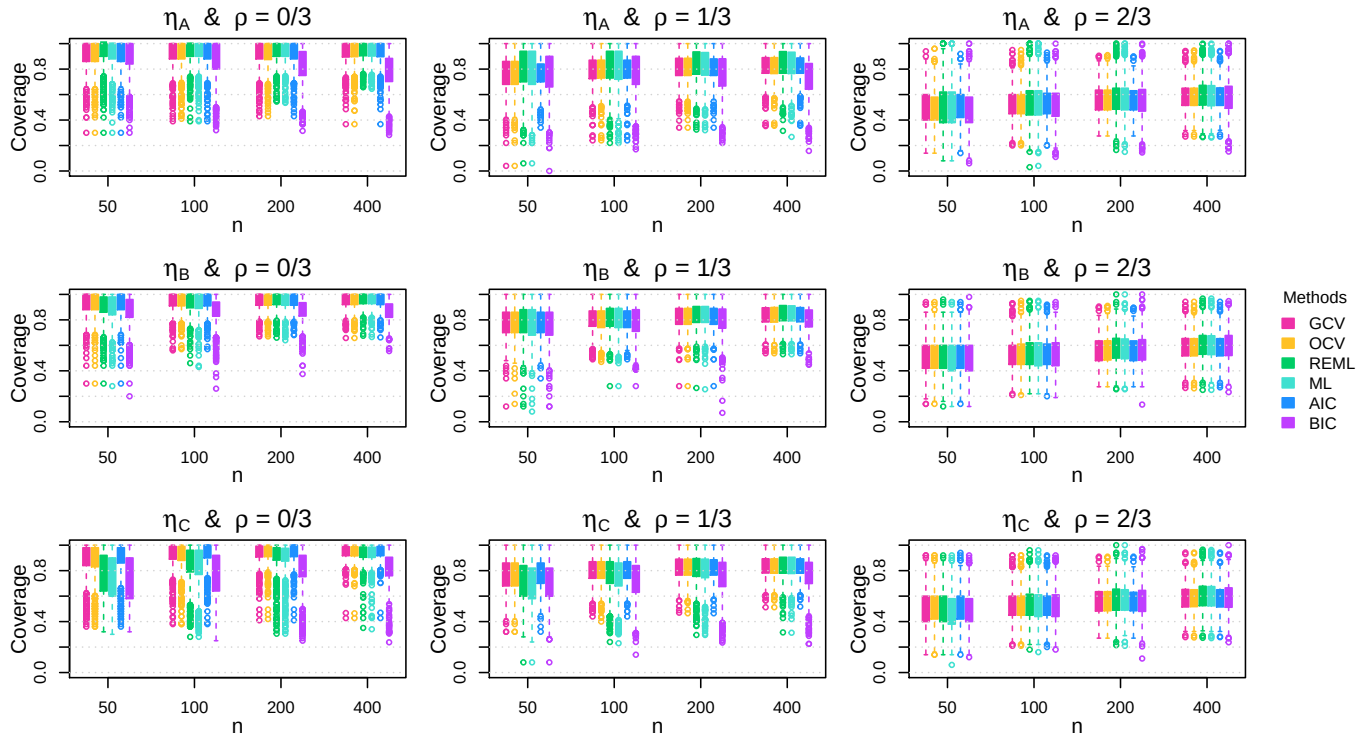

Figure A8. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for multivariate t_5 errors with constant standard deviation.

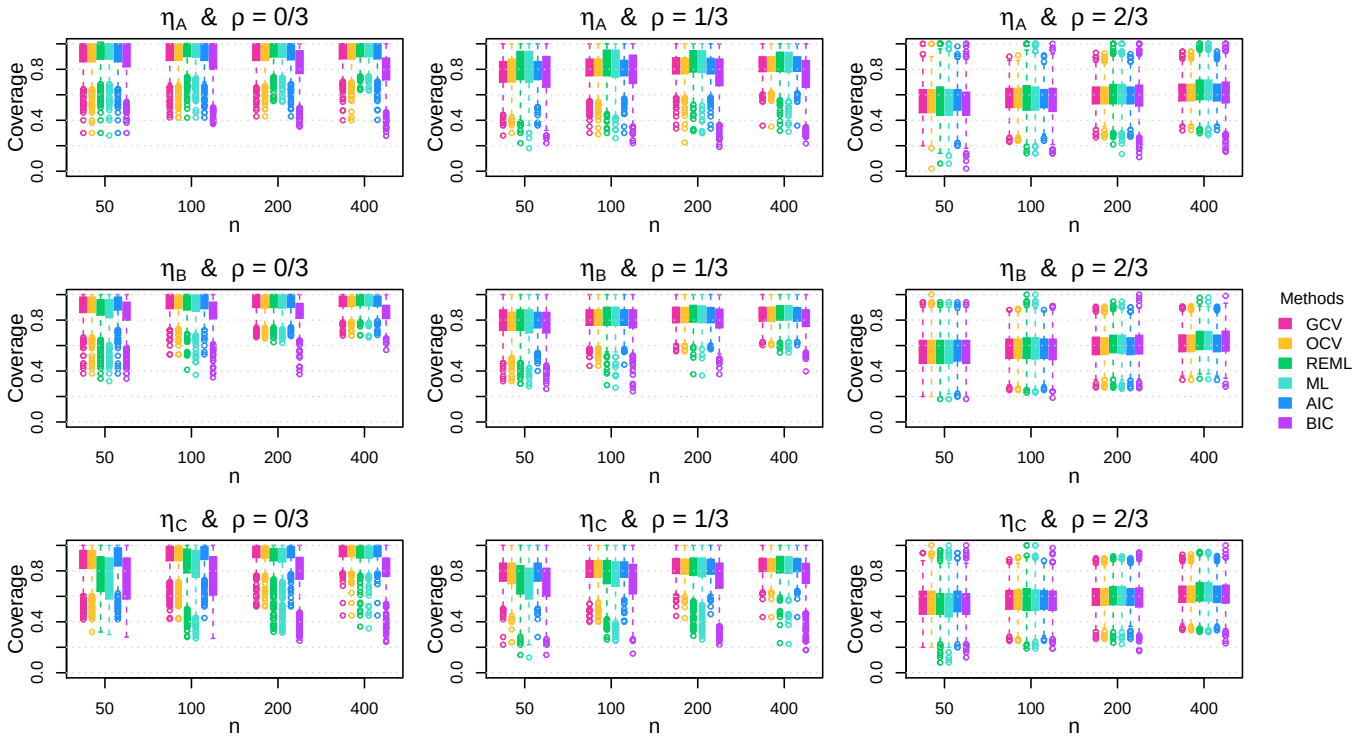

Figure A9. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for multivariate t_5 errors with increasing standard deviation.

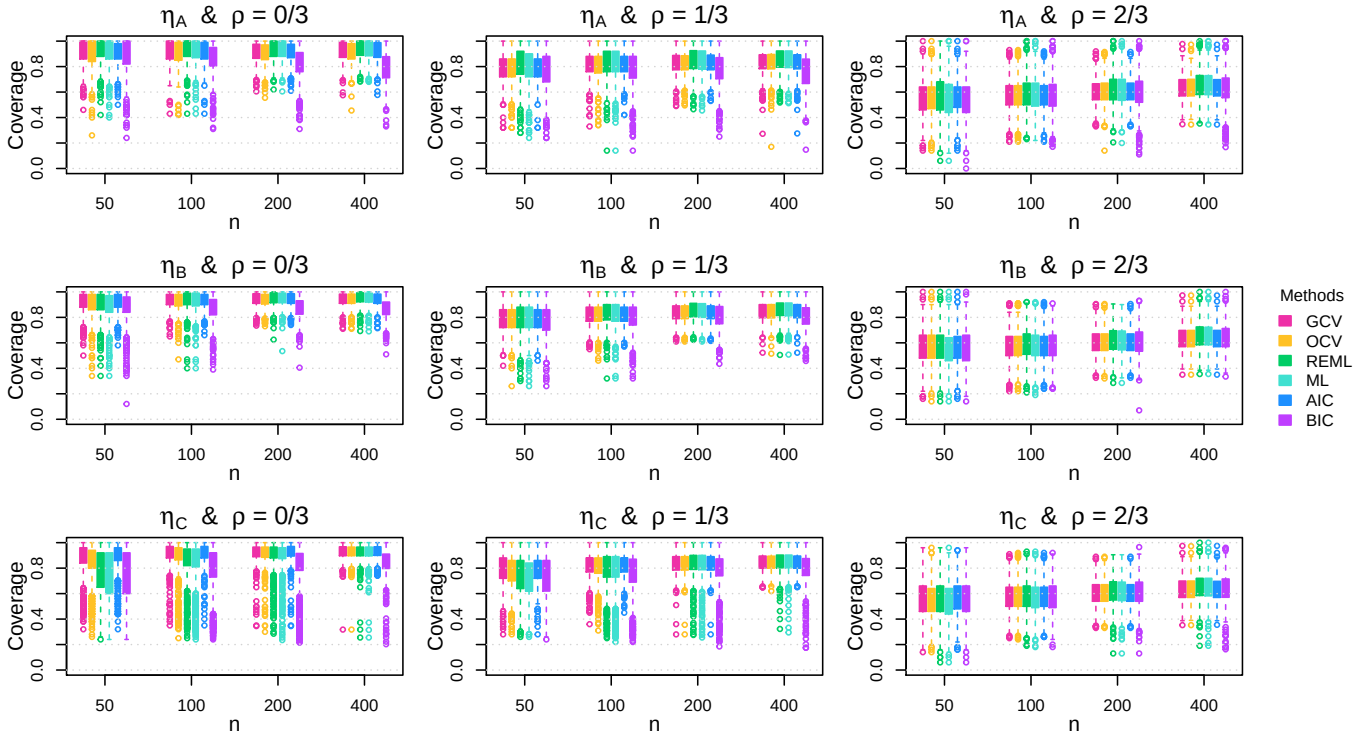

Figure A10. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for multivariate t_5 errors with parabolic standard deviation.

Appendix C. Supplementary Results for Uniform Errors

Appendix C.1. RMSE Results

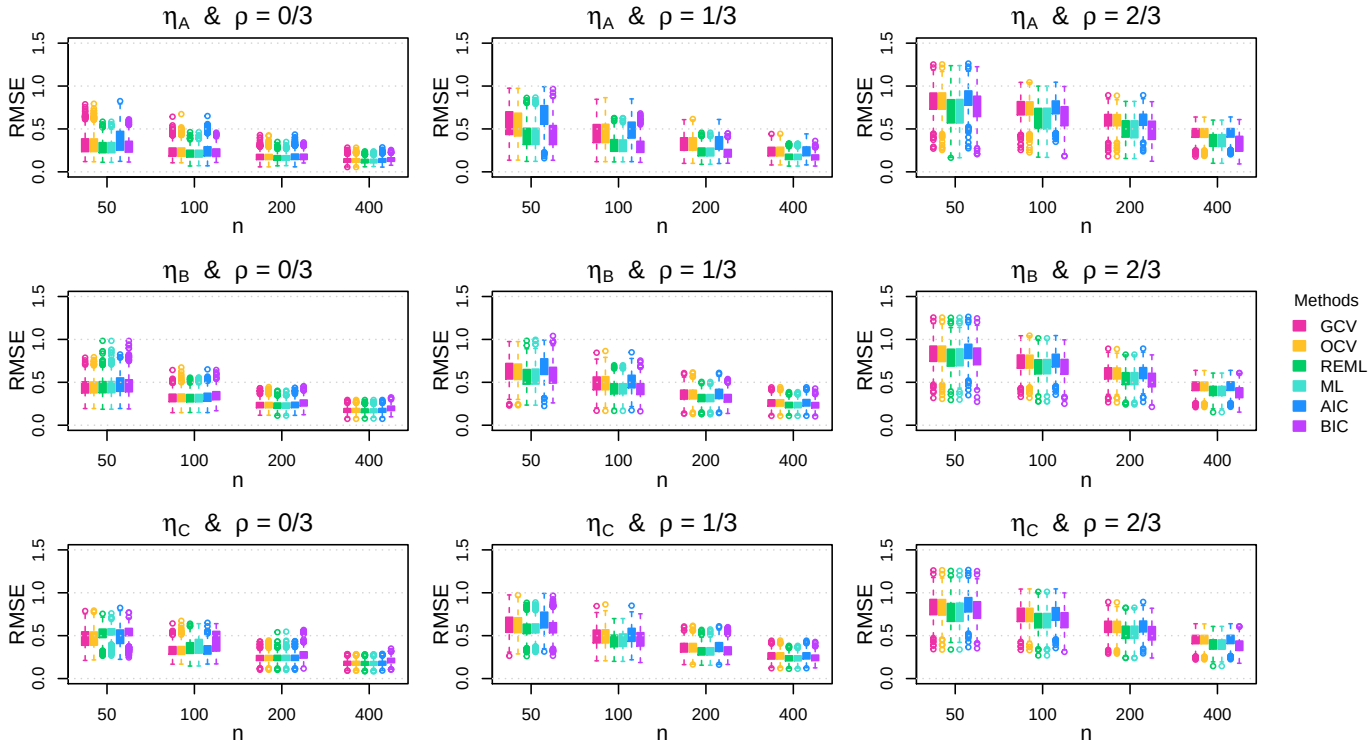

Figure A11. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for uniform errors with constant standard deviation.

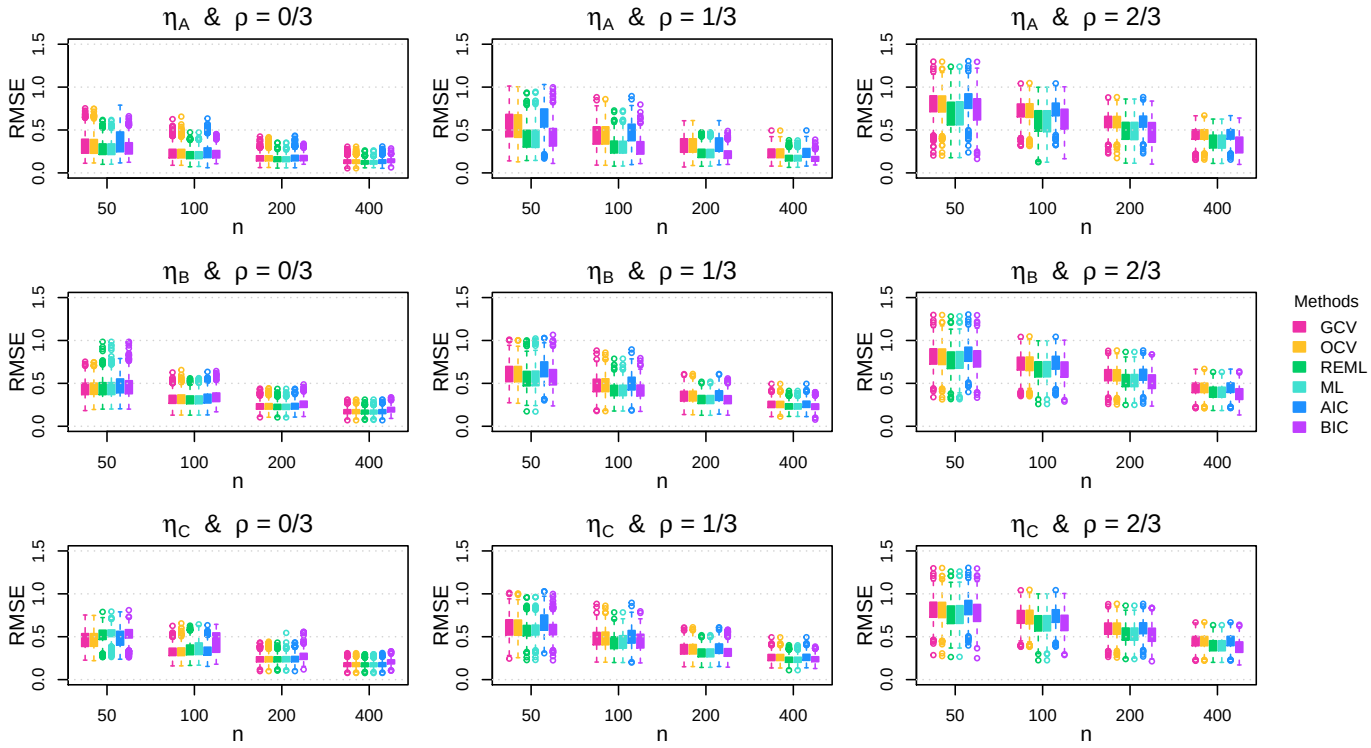

Figure A12. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for uniform errors with increasing standard deviation.

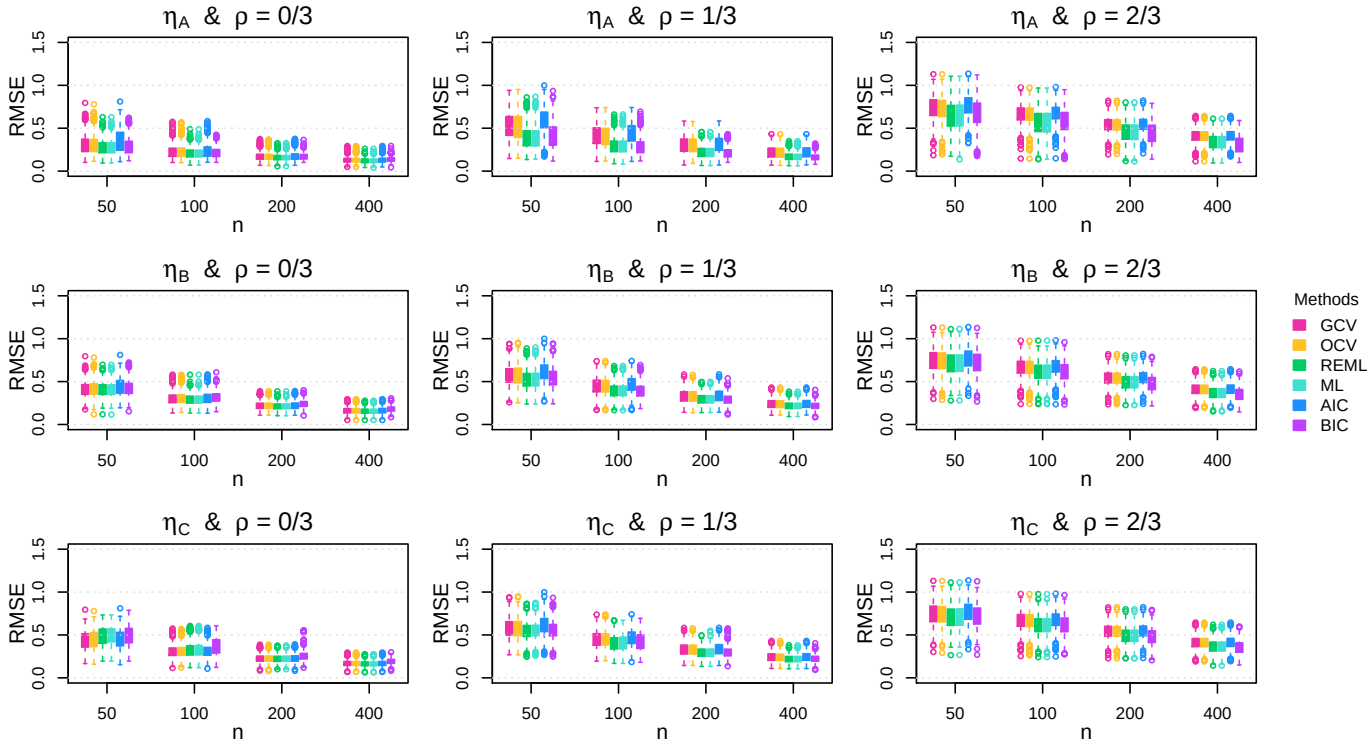

Figure A13. Boxplots of the root mean squared error (RMSE) across the 1000 simulation replications for uniform errors with parabolic standard deviation.

Appendix C.2. Coverage Results

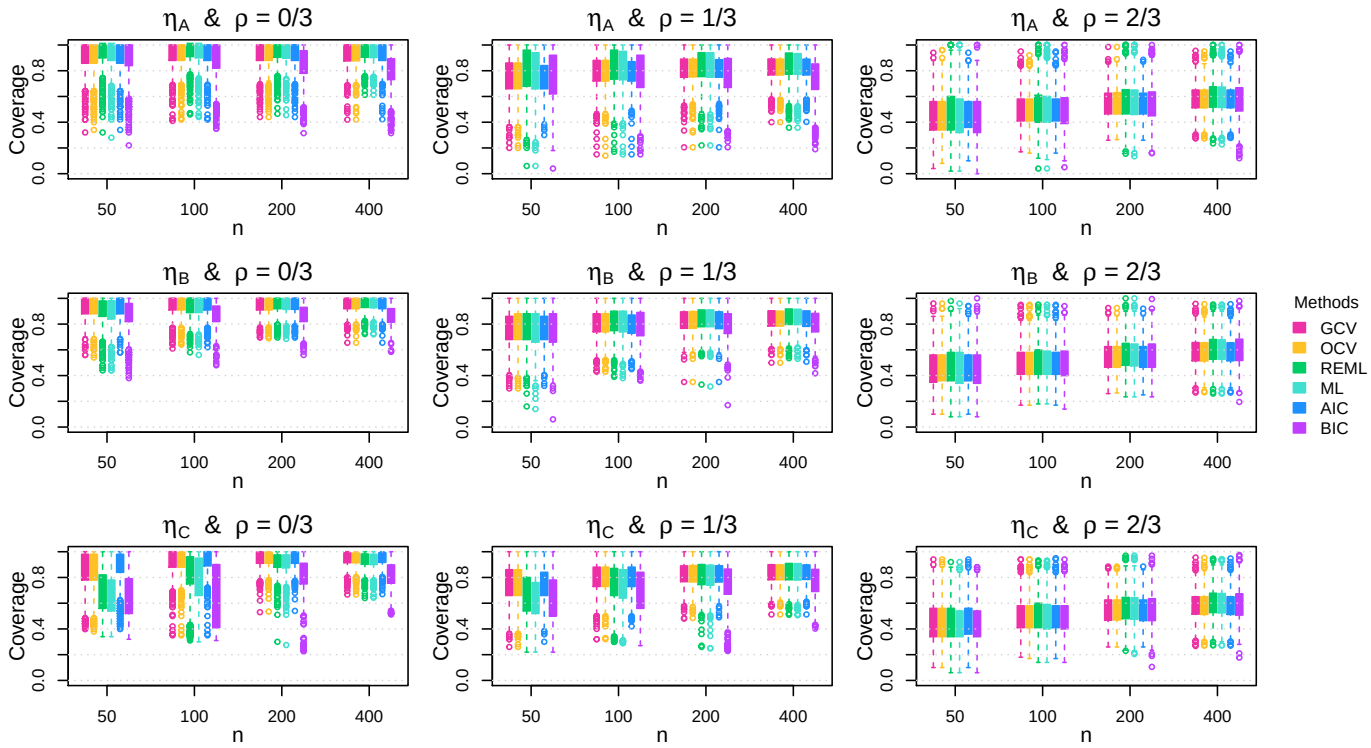

Figure A14. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for uniform errors with constant standard deviation.

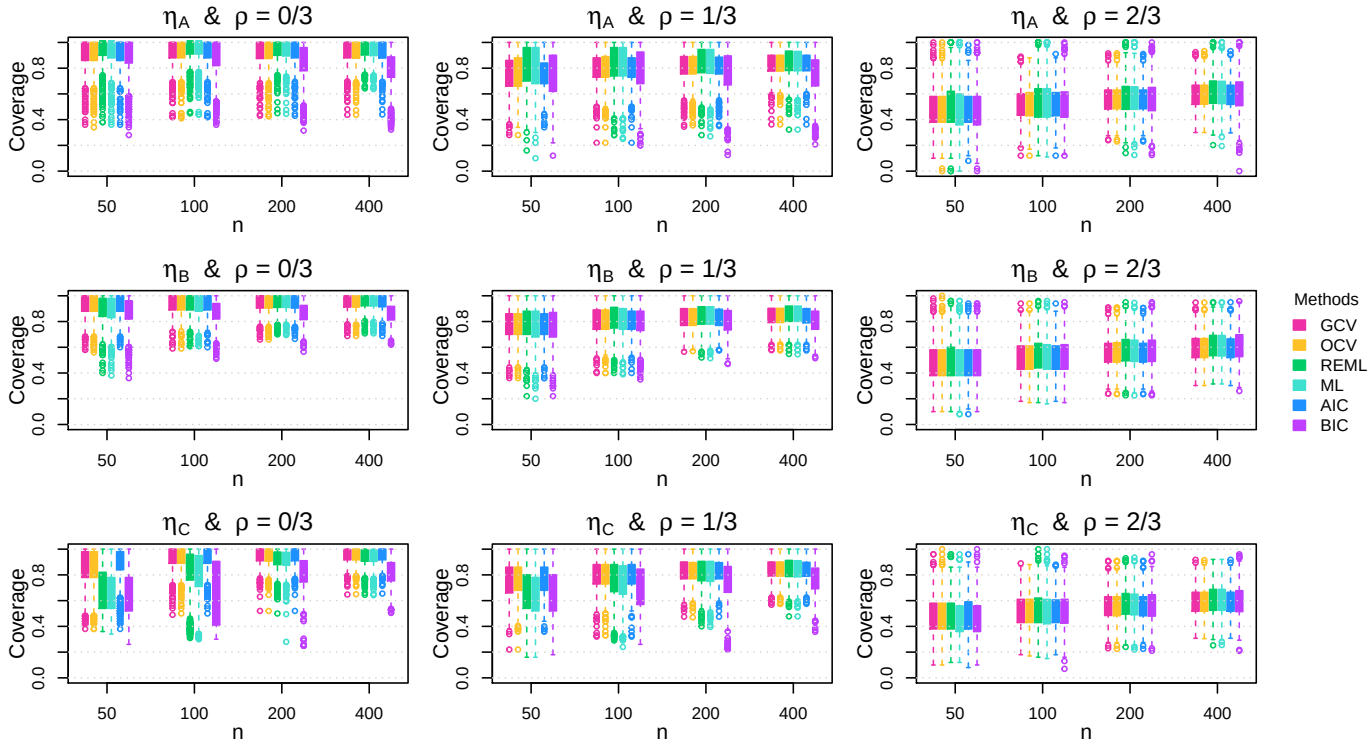

Figure A15. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for uniform errors with increasing standard deviation.

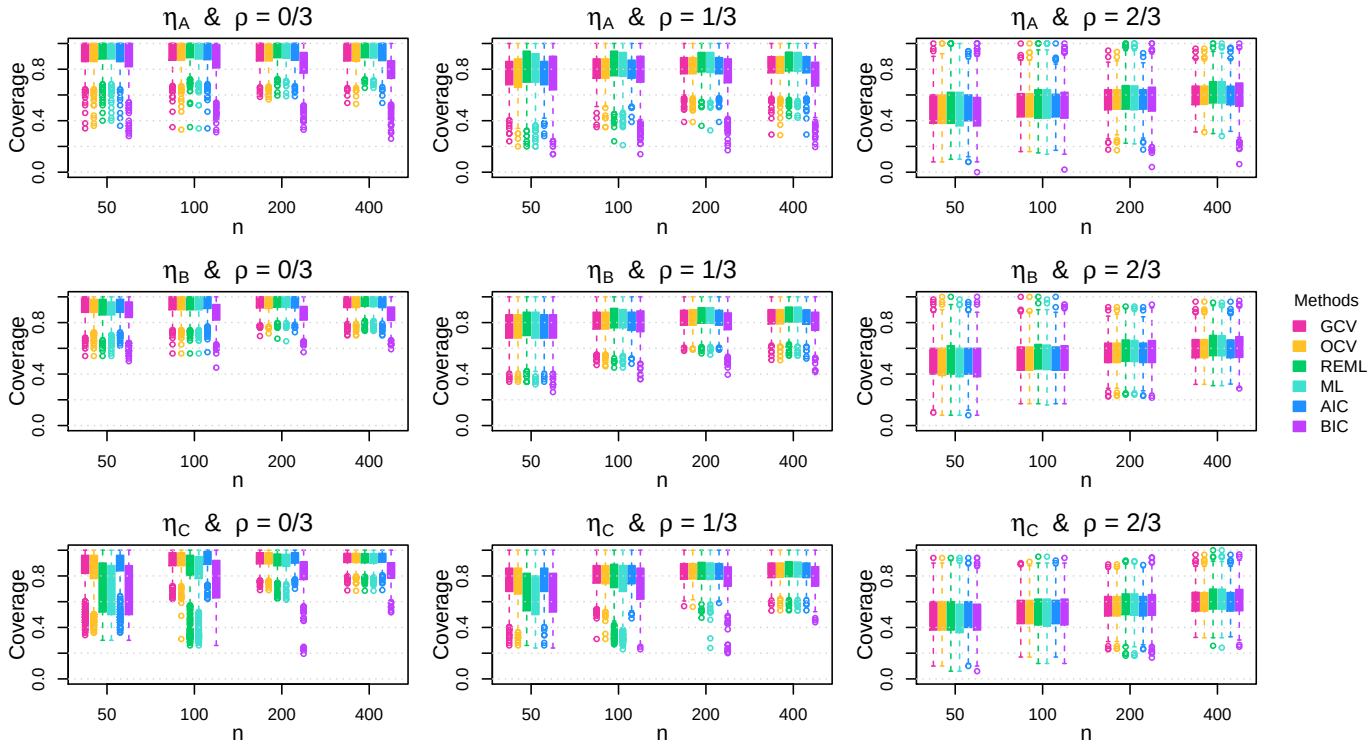

Figure A16. Boxplots of the coverage rate for the 95% Bayesian confidence interval across the 1000 simulation replications for uniform errors with parabolic standard deviation.

References

1. Ramsay, J.O.; Silverman, B.W. *Applied Functional Data Analysis*; Springer: New York, NY, USA, 2002.
2. Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*, 2nd ed.; Springer: New York, NY, USA, 2005.
3. Ramsay, J.O.; Hooker, G.; Graves, S. *Functional Data Analysis with R and MATLAB*; Springer: New York, NY, USA, 2009.
4. Ullah, S.; Finch, C.F. Applications of functional data analysis: A systematic review. *BMC Med. Res. Methodol.* **2013**, *13*, 43. [CrossRef]
5. Wang, J.L.; Chiou, J.M.; Müller, H.G. Functional Data Analysis. *Annu. Rev. Stat. Its Appl.* **2016**, *3*, 257–295. [CrossRef]
6. Stone, A.A.; Shiffman, S. Ecological momentary assessment (EMA) in behavorial medicine. *Ann. Behav. Med.* **1994**, *16*, 199–202. [CrossRef]
7. Shiffman, S.; Stone, A.A.; Hufford, M.R. Ecological Momentary Assessment. *Annu. Rev. Clin. Psychol.* **2008**, *4*, 1–32. [CrossRef]
8. Helwig, N.E.; Gao, Y.; Wang, S.; Ma, P. Analyzing spatiotemporal trends in social media data via smoothing spline analysis of variance. *Spat. Stat.* **2015**, *14*, 491–504. [CrossRef]
9. Helwig, N.E.; Shorter, K.A.; Hsiao-Wecksler, E.T.; Ma, P. Smoothing spline analysis of variance models: A new tool for the analysis of cyclic biomechaniacal data. *J. Biomech.* **2016**, *49*, 3216–3222. [CrossRef]
10. Helwig, N.E.; Sohre, N.E.; Ruprecht, M.R.; Guy, S.J.; Lyford-Pike, S. Dynamic properties of successful smiles. *PLoS ONE* **2017**, *12*, e0179708. [CrossRef] [PubMed]
11. Helwig, N.E.; Ruprecht, M.R. Age, gender, and self-esteem: A sociocultural look through a nonparametric lens. *Arch. Sci. Psychol.* **2017**, *5*, 19–31. [CrossRef]
12. Lawrence, R.L.; Sessions, W.C.; Jensen, M.C.; Staker, J.L.; Eid, A.; Breighner, R.; Helwig, N.E.; Braman, J.P.; Ludewig, P.M. The effect of glenohumeral plane of elevation on supraspinatus subacromial proximity. *J. Biomech.* **2018**, *79*, 147–154. [CrossRef] [PubMed]
13. Almquist, Z.W.; Helwig, N.E.; You, Y. Connecting Continuum of Care point-in-time homeless counts to United States Census areal units. *Math. Popul. Stud.* **2020**, *27*, 46–58. [CrossRef]
14. Hammell, A.E.; Helwig, N.E.; Kaczkurkin, A.N.; Sponheim, S.R.; Lissek, S. The temporal course of over-generalized conditioned threat expectancies in posttraumatic stress disorder. *Behav. Res. Ther.* **2020**, *124*, 103513. [CrossRef] [PubMed]
15. Helwig, N.E. Regression with ordered predictors via ordinal smoothing splines. *Front. Appl. Math. Stat.* **2017**, *3*, 15. [CrossRef]
16. Helwig, N.E. Multiple and Generalized Nonparametric Regression. In *SAGE Research Methods Foundations*; Atkinson, P., Delamont, S., Cernat, A., Sakshaug, J.W., Williams, R.A., Eds.; SAGE: Thousand Oaks, CA, USA, 2020. [CrossRef]
17. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*; Springer: New York, NY, USA, 2013. [CrossRef]
18. Hoerl, A.; Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
19. Gu, C.; Kim, Y.J. Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Stat.* **2002**, *30*, 619–628. [CrossRef]
20. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing; R Version 4.1.0.; R Core Team: Vienna, Austria, 2021.
21. Altman, N.S. Kernel smoothing of data with correlated errors. *J. Am. Stat. Assoc.* **1990**, *85*, 749–759. [CrossRef]
22. Opsomer, J.; Wang, Y.; Yang, Y. Nonparametric regression with correlated errors. *Stat. Sci.* **2001**, *16*, 134–153. [CrossRef]
23. Wang, Y. Mixed effects smoothing spline analysis of variance. *J. R. Stat. Soc. Ser. B* **1998**, *60*, 159–174. [CrossRef]
24. Wang, Y. Smoothing spline models with correlated random errors. *J. Am. Stat. Assoc.* **1998**, *93*, 341–348. [CrossRef]
25. Zhang, D.; Lin, X.; Raz, J.; Sowers, M. Semiparametric stochastic mixed models for longitudinal data. *J. Am. Stat. Assoc.* **1998**, *93*, 710–719. [CrossRef]
26. Reiss, P.T.; Ogden, R.T. Smoothing parameter selection for a class of semiparametric linear models. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 505–523. [CrossRef]
27. Krivobokova, T.; Kauermann, G. A note on penalized spline smoothing with correlated errors. *J. Am. Stat. Assoc.* **2007**, *102*, 1328–1337. [CrossRef]
28. Lee, T.C.M. Smoothing parameter selection for smoothing splines: A simulation study. *Comput. Stat. Data Anal.* **2003**, *42*, 139–148. [CrossRef]
29. Kimeldorf, G.; Wahba, G. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **1971**, *33*, 82–95. [CrossRef]
30. Kim, Y.J.; Gu, C. Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. R. Stat. Soc. Ser. B* **2004**, *66*, 337–356. [CrossRef]
31. Gu, C. *Smoothing Spline ANOVA Models*, 2nd ed.; Springer: New York, NY, USA, 2013. [CrossRef]
32. Moore, E.H. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.* **1920**, *26*, 394–395. [CrossRef]
33. Penrose, R. A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc.* **1955**, *51*, 406–413. [CrossRef]
34. Wahba, G. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Stat. Soc. Ser. B* **1983**, *45*, 133–150. [CrossRef]
35. Nychka, D. Bayesian confidence intervals for smoothing splines. *J. Am. Stat. Assoc.* **1988**, *83*, 1134–1143. [CrossRef]
36. Kalpić, D.; Hlupić, N. Multivariate Normal Distributions. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 907–910. [CrossRef]

37. Allen, D.M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127. [CrossRef]
38. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B (Methodol.)* **1974**, *36*, 111–133. [CrossRef]
39. Wahba, G.; Wold, S. A completely automatic French curve: Fitting spline functions by cross validation. *Commun. Stat.* **1975**, *4*, 1–17. [CrossRef]
40. Craven, P.; Wahba, G. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **1979**, *31*, 377–403. [CrossRef]
41. Li, K.C. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Stat.* **1987**, *15*, 958–975. [CrossRef]
42. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
43. Schwarz, G.E. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
44. Wahba, G. A comparison of GCV and GML for choosing the smoothing parameters in the generalized spline smoothing problem. *Ann. Stat.* **1985**, *4*, 1378–1402. [CrossRef]
45. Ruppert, D.; Wand, M.P.; Carroll, R.J. *Semiparametric Regression*; Cambridge University Press: Cambridge, UK, 2003.
46. Patterson, H.D.; Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **1971**, *58*, 545–554. [CrossRef]
47. Falk, M. A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Commun. Stat.-Simul. Comput.* **1999**, *28*, 785–791. [CrossRef]
48. Helwig, N.E. *npreg: Nonparametric Regression via Smoothing Splines*; R Package Version 1.0-6; The Comprehensive R Archive Network. 2021. Available online: https://cran.r-project.org/package=npreg (accessed on 22 August 2021).
49. GISTEMP Team. *GISS Surface Temperature Analysis (GISTEMP)*; Dataset Version 4; NASA Goddard Institute for Space Studies. Available online https://data.giss.nasa.gov/gistemp/ (accessed on 17 May 2021).
50. Lenssen, N.; Schmidt, G.; Hansen, J.; Menne, M.; Persin, A.; Ruedy, R.; Zyss, D. Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.* **2019**, *124*, 6307–6326. [CrossRef]
51. Silverman, B.W. Aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. R. Stat. Soc. Ser. B* **1985**, *47*, 1–52. [CrossRef]
52. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002; ISBN 0-387-95457-0.
53. Koenker, R.; Ng, P.; Portnoy, S. Quantile smoothing splines. *Biometrika* **1994**, *81*, 673–680. [CrossRef]
54. Li, G.Y.; Shi, P.; Li, G. Global convergence rates of B-spline M-estimators in nonparametric regression. *Stat. Sin.* **1995**, *5*, 303–318.
55. Wood, S.N. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*; R Package Version 1.8-35; The Comprehensive R Archive Network. 2021. Available online: https://cran.r-project.org/package=mgcv (accessed on 22 August 2021).
56. Helwig, N.E. *bigsplines: Smoothing Splines for Large Samples*; R Package Version 1.1-1; The Comprehensive R Archive Network. 2018. Available online: https://cran.r-project.org/package=bigsplines (accessed on 22 August 2021).