stats

MDPI

Article

Robust Permutation Tests for Penalized Splines

Nathaniel E. Helwig [1,2]

[1] Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA; helwig@umn.edu; Tel.: +1-612-624-8363
[2] School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

Abstract: Penalized splines are frequently used in applied research for understanding functional relationships between variables. In most applications, statistical inference for penalized splines is conducted using the random effects or Bayesian interpretation of a smoothing spline. These interpretations can be used to assess the uncertainty of the fitted values and the estimated component functions. However, statistical tests about the nature of the function are more difficult, because such tests often involve testing a null hypothesis that a variance component is equal to zero. Furthermore, valid statistical inference using the random effects or Bayesian interpretation depends on the validity of the utilized parametric assumptions. To overcome these limitations, I propose a flexible and robust permutation testing framework for inference with penalized splines. The proposed approach can be used to test omnibus hypotheses about functional relationships, as well as more flexible hypotheses about conditional relationships. I establish the conditions under which the methods will produce exact results, as well as the asymptotic behavior of the various permutation tests. Additionally, I present extensive simulation results to demonstrate the robustness and superiority of the proposed approach compared to commonly used methods.

Keywords: generalized ridge regression; nonparametric methods; penalized least squares; randomization tests; smoothing and nonparametric regression

Citation: Helwig, N.E. Robust Permutation Tests for Penalized Splines. *Stats* **2022**, *5*, 916–933. https://doi.org/10.3390/stats5030053

Academic Editor: Wei Zhu

Received: 18 August 2022
Accepted: 11 September 2022
Published: 16 September 2022

1. Introduction

1.1. Penalized Spline Prevalence

Penalized splines are used within modern multiple and generalized nonparametric regression frameworks [1,2], such as the smoothing spline analysis of variance models [3–5] and generalized additive models [6–8], to discover unknown functional relationships between a response variable Y and a collection of predictors $X_1, \ldots, X_d$. Penalized splines and their variants have been applied to understand functional relationships in data from a variety of different disciplines. For example, recent applications include the use of penalized splines to model spatiotemporal patterns in US homelessness [9], biomechanical analysis of locomotion data [10,11], fear learning curves in veterans with PTSD [12], functional properties of successful smiles [13], self-esteem trajectories across the lifespan [14], and spatiotemporal trends in social media [15].

In addition to being a commonly used tool in applied research, penalized splines are also frequently the focus of theoretical and computational statistics research. For example, there has been recent interest in fitting penalized spline models with ordinal predictors and responses [16,17], mixed-effects models with penalized spline components [18–20], efficient approximation for large samples [21,22], and efficient algorithms for fitting penalized spline models to big data [23,24]. Furthermore, given the frequent usage of penalized splines for the analysis of real world data, there has been a growing interest in the robustness of penalized spline tuning and inference methods under model misspecification [25]. Recently, there has also been the development of alternative spline penalization approaches for nonparametric function estimation from noisy data [26].

1.2. Penalized Spline Definition

Consider a multiple nonparametric regression model [1,2] of the form

$$Y_i = \eta(\boldsymbol{X}_i) + \epsilon_i \tag{1}$$

where Y_i is the i-th realization of the response variable $Y \in \mathbb{R}$, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top$ is the i-th realization of the predictor variable $\boldsymbol{X} \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, and ϵ_i is the i-th realization of the error term $\epsilon \in \mathbb{R}$. The error terms are assumed to satisfy $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma_i^2$ with $\sigma_i^2 < \infty$ denoting the i-th observation's error variance. To estimate η, it is typical to minimize the penalized least squares (PLS) functional

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \eta(\boldsymbol{X}_i))^2 + \lambda J(\eta)$$

where $J(\cdot)$ is the penalty functional, and $\lambda \geq 0$ is the smoothing parameter, which controls the balance between fitting and smoothing.

Given λ, the Kimeldorf-Wahba representer theorem [27] reveals that the minimizer of the PLS functional has the form

$$\eta_\lambda(\boldsymbol{X}) = a + \sum_{v=1}^{m-1} b_v N_v(\boldsymbol{X}) + \sum_{i=1}^{r} c_i R(\boldsymbol{X}, \boldsymbol{X}_i^*)$$

where $\{N_v\}_{v=0}^{m-1}$ are known functions that span the null space $\mathcal{H}_0 = \{\eta : J(\eta) = 0\}$ with $N_0(\boldsymbol{X}) = 1$, the symmetric and bivariate function $R(\cdot,\cdot)$ is the known reproducing kernel of the contrast space $\mathcal{H}_1 = \{\eta : J(\eta) < \infty\}$, the collection of predictor scores $\{\boldsymbol{X}_i^*\}_{i=1}^r$ are the selected spline knots, $a \in \mathbb{R}$ is the unknown intercept, and $\mathbf{b} = (b_1, \ldots, b_{m-1})^\top$ and $\mathbf{c} = (c_1, \ldots, c_r)^\top$ are the unknown basis function coefficient vectors. The representer theorem uses all design points as knots (i.e., $r = n$ and $\boldsymbol{X}_i^* = \boldsymbol{X}_i$), but reasonable approximations can be obtained using $r < n$ knots [21,22,28].

1.3. Penalized Spline Estimation

Using the representer theorem, the nonparametric regression model can be written as

$$Y_i = a + \boldsymbol{N}_i^\top \mathbf{b} + \boldsymbol{R}_i^\top \mathbf{c} + \epsilon_i$$

where $\boldsymbol{N}_i^\top = (N_1(\boldsymbol{X}_i), \ldots, N_{m-1}(\boldsymbol{X}_i))$ is the null space basis function vector for the i-th observation, and $\boldsymbol{R}_i^\top = (R(\boldsymbol{X}_i, \boldsymbol{X}_1^*), \ldots, R(\boldsymbol{X}_i, \boldsymbol{X}_r^*))$ is the reproducing kernel function evaluated at $\boldsymbol{X}_i$ and the selected knots. Let $Y_i^c = Y_i - \bar{Y}$ denote the centered response and let $\boldsymbol{N}_i^c = \boldsymbol{N}_i - \bar{\boldsymbol{N}}$ and $\boldsymbol{R}_i^c = \boldsymbol{R}_i - \bar{\boldsymbol{R}}$ denote the centered basis functions, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $\bar{\boldsymbol{N}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{N}_i$, and $\bar{\boldsymbol{R}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{R}_i$. Then, the PLS functional can be rewritten as

$$\frac{1}{n}\|\boldsymbol{Y}_c - \mathbf{N}_c\mathbf{b} - \mathbf{R}_c\mathbf{c}\|^2 + \lambda \mathbf{c}^\top \mathbf{Q}\mathbf{c}$$

where $\boldsymbol{Y}_c = (Y_1^c, \ldots, Y_n^c)^\top$, $\mathbf{N}_c = [\boldsymbol{N}_1^c, \ldots, \boldsymbol{N}_n^c]^\top$ and $\mathbf{R}_c = [\boldsymbol{R}_1^c, \ldots, \boldsymbol{R}_n^c]^\top$ are the mean centered basis function matrices, and $\mathbf{Q} = [R(\boldsymbol{X}_i^*, \boldsymbol{X}_j^*)]$ is the penalty matrix.

Let $\mathbf{d} = (b_1, \ldots, b_{m-1}, c_1, \ldots, c_r)^\top$ denote the combined basis function coefficient vector, and let $\mathbf{K}_c = (\mathbf{N}_c, \mathbf{R}_c)$ denote the combined (centered) basis function matrix. Given the smoothing parameter λ, it is well known that the basis function coefficients that minimize the PLS functional can be written as

$$\hat{\mathbf{d}}_\lambda = \left(\mathbf{K}_c^\top \mathbf{K}_c + n\lambda \mathbf{Q}^*\right)^\dagger \mathbf{K}_c^\top \boldsymbol{Y}_c$$

where $\mathbf{Q}^* = \text{bdiag}(\mathbf{0}_{m-1}, \mathbf{Q})$ is the block diagonal penalty matrix, and $(\cdot)^\dagger$ denotes the Moore-Penrose pseudo-inverse [29,30]. The least squares estimate of the intercept term can

then be written as $\hat{a} = \bar{Y} - \bar{\mathbf{K}}^\top \hat{\mathbf{d}}_\lambda$, where $\bar{\mathbf{K}}^\top = (\bar{\mathbf{N}}^\top, \bar{\mathbf{R}}^\top)$. The fitted values have the form $\hat{\boldsymbol{\eta}}_\lambda = \hat{a} + \hat{\boldsymbol{\eta}}_{0\lambda} + \hat{\boldsymbol{\eta}}_{1\lambda}$, where $\hat{\boldsymbol{\eta}}_{0\lambda} = \mathbf{N}_c \hat{\mathbf{b}}_\lambda$ is the (non-constant) null space contribution, and $\hat{\boldsymbol{\eta}}_{1\lambda} = \mathbf{R}_c \hat{\mathbf{c}}_\lambda$ is the contrast space contribution with $\hat{\mathbf{d}}_\lambda^\top = (\hat{\mathbf{b}}_\lambda^\top, \hat{\mathbf{c}}_\lambda^\top)$.

1.4. Bayesian Interpretation

Given the estimated function $\hat{\eta}_\lambda$, the statistical inference about the unknown function η is often conducted using the Bayesian interpretation of a smoothing spline [31,32]. This approach assumes that $\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1$, where the null space function $\boldsymbol{\eta}_0$ has a vague prior, and the contrast space function $\boldsymbol{\eta}_1$ has a Gaussian process prior with a mean zero and covariance matrix proportional to $\mathbf{Q}^\dagger$. Given these prior assumptions, it can be demonstrated that the posterior distribution of η given $\boldsymbol{Y}$ is multivariate normal with mean vector $\hat{\boldsymbol{\eta}}_\lambda$ and covariance matrix $\left(\mathbf{K}^\top \mathbf{K} + n\lambda \mathbf{Q}^*\right)^\dagger$, where $\mathbf{K} = (\mathbf{1}_n, \mathbf{N}, \mathbf{R})$. When the smoothing parameter λ is chosen via the generalized cross-validation (GCV) criterion [33], confidence intervals formed using the Bayesian covariance matrix tend to have an "across-the-function" coverage property [32]. See [25] for an investigation of the Bayesian CI coverage properties.

The Bayesian confidence intervals can be used to assess the uncertainty of the fitted values and the component functions, i.e., the main and interaction effects of the d predictors [34,35]. However, testing hypotheses about the nature of the function η is a more difficult problem. Various different approaches have been proposed for testing hypotheses about penalized smoothers [36–43]. However, these methods are primarily designed for testing specific hypotheses (e.g., $H_0 : \eta \in \mathcal{H}_0$) under the assumption of homoscedastic Gaussian errors. The exception is Wood's (2013a) approach, which is designed to be more general, but the validity of Wood's approach depends on having a correctly specified parametric (exponential family) distribution. As a result, the generalizability of the existing inference methods is limited, and the validity of these methods is suspect when the parametric assumptions are questionable.

1.5. Proposed Approach

When working with real data, it can be difficult to determine whether or not the parametric assumptions that are required for valid hypothesis testing are reasonably met. Furthermore, if the error terms are non-Gaussian and/or heteroscedastic, the proposed hypothesis tests may produce substantially misleading inferential results. To overcome this important practical issue, I propose a flexible permutation testing framework for robust inference in nonparametric regression models. The proposed approach extends recent advances in robust permutation tests for linear models [44] to generalized ridge regression (GRR) and penalized smoothing problems. As I demonstrate in the following sections, the proposed framework can be used for overall (omnibus) tests about functional relationships, as well as more specific (conditional) tests.

The remainder of this paper is organized as follows: Section 2 develops the foundations and theory for omnibus permutation tests using GRR estimators, Section 3 extends the permutation testing framework to conditional tests of effects, Section 4 presents extensive simulation results to validate the theoretical results derived in the previous sections, and Section 5 discusses how the proposed framework can be flexibly adapted for testing a variety hypotheses about semi- and non-parametric regression models.

2. Omnibus Regression Tests

2.1. Model and Estimation

Given an independent sample of n observations, consider the linear regression model

$$Y_i = \alpha + \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i \tag{2}$$

for $i \in \{1, \ldots, n\}$, where Y_i is the i-th realization of the response variable $Y \in \mathbb{R}$, $\boldsymbol{X}_i^\top = (X_{i1}, \ldots, X_{ip})$ is the i-th realization of the predictor vector $\boldsymbol{X}^\top = (X_1, \ldots, X_p) \in \mathbb{R}^p$, and ϵ_i is the i-th realization of the error term $\epsilon \in \mathbb{R}$. The error terms are assumed to satisfy

$E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma_i^2$ with $\sigma_i^2 < \infty$ denoting the i-th observation's (finite) error variance. Without a loss of generality, we can assume that the response and predictors are centered, which implies that the intercept α can be dropped from the model for estimation purposes. Let $Y_i^c = Y_i - \bar{Y}$ and $\boldsymbol{X}_i^c = \boldsymbol{X}_i - \bar{\boldsymbol{X}}$ denote the mean centered response and predictor vector for the i-th observation, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$.

To estimate the coefficients in $\boldsymbol{\beta}$, consider minimizing the GRR loss function [45]

$$\frac{1}{n}(\boldsymbol{Y}_c - \mathbf{X}_c\boldsymbol{\beta})^\top(\boldsymbol{Y}_c - \mathbf{X}_c\boldsymbol{\beta}) + \boldsymbol{\beta}^\top\boldsymbol{\Delta}\boldsymbol{\beta} \tag{3}$$

where $\boldsymbol{Y}_c = (Y_1^c, \ldots, Y_n^c)^\top$ is the centered response vector, $\mathbf{X}_c = [\boldsymbol{X}_1^c, \ldots, \boldsymbol{X}_n^c]^\top$ is the centered design matrix, and $\boldsymbol{\Delta}$ is a $p \times p$ symmetric and positive semi-definite penalty matrix. The coefficients that minimize the GRR problem in Equation (3) have the form

$$\hat{\boldsymbol{\beta}}_\Delta = \left(\frac{1}{n}\mathbf{X}_c^\top\mathbf{X}_c + \boldsymbol{\Delta}\right)^{-1}\left(\frac{1}{n}\mathbf{X}_c^\top\boldsymbol{Y}_c\right) \tag{4}$$

which is subscripted to emphasize that the estimated coefficients depend on the penalty matrix $\boldsymbol{\Delta}$. Given the estimated slope vector $\hat{\boldsymbol{\beta}}_\Delta$, the least squares estimate of the intercept has the form $\hat{\alpha} = \bar{Y} - \bar{\boldsymbol{X}}^\top\hat{\boldsymbol{\beta}}_\Delta$.

2.2. Asymptotic Distributions

Consider the linear regression model from Equation (2) with the assumptions:

A1. $Y_i = \alpha + \boldsymbol{X}_i^\top\boldsymbol{\beta} + \epsilon_i$ with $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma_i^2 < \infty$ for $i = 1, \ldots, n$

A2. $(Y_i, \boldsymbol{X}_i)$ are iid from a distribution satisfying $E(\epsilon_i\boldsymbol{X}_i) = 0$

A3. $\boldsymbol{\Sigma}_X = E\big((\boldsymbol{X}_i - \boldsymbol{\mu}_X)(\boldsymbol{X}_i - \boldsymbol{\mu}_X)^\top\big)$ and $\boldsymbol{\Omega}_X = E\big(\epsilon_i^2(\boldsymbol{X}_i - \boldsymbol{\mu}_X)(\boldsymbol{X}_i - \boldsymbol{\mu}_X)^\top\big)$ are nonsingular, where $\boldsymbol{\mu}_X = E(\boldsymbol{X}_i)$, and $\frac{1}{n}\mathbf{X}_c^\top\mathbf{X}_c + \boldsymbol{\Delta}$ is almost surely invertible. In the fixed predictors case, the mean vector is $\boldsymbol{\mu}_X = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ and the the covariance matrix terms are defined as $\boldsymbol{\Sigma}_X = \frac{1}{n}\mathbf{X}_c^\top\mathbf{X}_c$ and $\boldsymbol{\Omega}_X = \frac{1}{n}\mathbf{X}_c^\top\boldsymbol{\Psi}\mathbf{X}_c$, where $\boldsymbol{\Psi} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.

Given assumptions A1–A3, the GRR estimator provides an estimate of $\boldsymbol{\beta}_\Delta = \boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\sigma}_{XY}$, where $\boldsymbol{\Sigma}_{X\Delta} = \boldsymbol{\Sigma}_X + \boldsymbol{\Delta}$ and $\boldsymbol{\sigma}_{XY} = E((\boldsymbol{X}_i - \boldsymbol{\mu}_X)(Y_i - \mu_Y))$ is the covariance between $\boldsymbol{X}_i$ and Y_i. Note that $\boldsymbol{\beta}_\Delta = \boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Sigma}_X\boldsymbol{\beta}$, where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\sigma}_{XY}$ is estimated by the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\sigma}}_{XY}$ with $\hat{\boldsymbol{\Sigma}}_X = \frac{1}{n}\mathbf{X}_c^\top\mathbf{X}_c$ and $\hat{\boldsymbol{\sigma}}_{XY} = \frac{1}{n}\mathbf{X}_c^\top\boldsymbol{Y}_c$. This implies that the GRR estimator can be written as $\hat{\boldsymbol{\beta}}_\Delta = \hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}\hat{\boldsymbol{\Sigma}}_X\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\Sigma}}_{X\Delta} = \hat{\boldsymbol{\Sigma}}_X + \boldsymbol{\Delta}$.

Lemma 1. *Given assumptions A1–A3, the GRR estimator $\hat{\boldsymbol{\beta}}_\Delta$ from Equation (4) is asymptotically normal with mean vector $\boldsymbol{\beta}_\Delta$ and covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_{X\Delta}^{-1}$, i.e.,*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_\Delta - \boldsymbol{\beta}_\Delta\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_{X\Delta}^{-1}\right)$$

as $n \to \infty$, where the notation $\xrightarrow{d}$ denotes convergence in distribution.

Lemma 1 can be proved using results of White [46], who demonstrated that the OLS estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal with a mean vector $\boldsymbol{\beta}$ and covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_X^{-1}$ under assumptions A1–A3. Noting that $\hat{\boldsymbol{\beta}}_\Delta = \hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}\hat{\boldsymbol{\Sigma}}_X\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_\Delta = \boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Sigma}_X\boldsymbol{\beta}$ completes the proof of Lemma 1, given that $\hat{\boldsymbol{\Sigma}}_X$ is a consistent estimator of $\boldsymbol{\Sigma}_X$. See Appendix A.1 for details.

Lemma 2. *Consider the linear model in Equation (2) with the assumptions $\boldsymbol{\beta} \sim N(\mathbf{0}_p, \frac{\sigma^2}{n}\boldsymbol{\Delta}^{-1})$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and suppose that $\boldsymbol{\beta}$ and ϵ_i are independent of one another. Under these assumptions, the posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ is multivariate normal with mean vector $\hat{\boldsymbol{\beta}}_\Delta$ and covariance matrix $\frac{\sigma^2}{n}\hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}$. The asymptotic mean vector is $\boldsymbol{\beta}_\Delta$ and the asymptotic covariance matrix is $\frac{\sigma^2}{n}\boldsymbol{\Sigma}_{X\Delta}^{-1}$.*

Lemma 2 can be proved by using the results of Henderson [47,48], who derived the covariance matrix of the best linear unbiased estimator (BLUE) and the best linear unbiased predictor (BLUP) in linear mixed models (e.g., see [49]). The asymptotic mean vector and covariance matrix result from the facts that $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\boldsymbol{\sigma}}_{XY}$ are consistent estimators of $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\sigma}_{XY}$, respectively. See Appendix A.2 for details.

2.3. Test Statistics

Consider the linear model in Equation (2), and suppose that we want to test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}_p$ versus the alternative hypothesis $H_1 : \boldsymbol{\beta} \neq \mathbf{0}_p$, where the notation $\mathbf{0}_p$ denotes a $p \times 1$ vector of zeros. If $\boldsymbol{\beta} \sim N(\mathbf{0}_p, \frac{\sigma^2}{n}\boldsymbol{\Delta}^{-1})$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ are independent of one another, one may consider using the F statistic

$$F = \frac{n}{p\hat{\sigma}^2}\hat{\boldsymbol{\beta}}_\Delta^\top \hat{\boldsymbol{\Sigma}}_{X\Delta}\hat{\boldsymbol{\beta}}_\Delta \tag{5}$$

where $\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^n \hat{\epsilon}_i^2$ is the estimated error variance, and $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ is the residual with $\hat{Y}_i = \hat{\alpha} + \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_\Delta$ denoting the fitted value for the i-th observation. Note that if H_0 is true and $\boldsymbol{\beta}$ is fixed, then the F statistic would approach an F distribution with degrees of freedom parameters p and $n - p - 1$ as $\boldsymbol{\Delta} \to \mathbf{0}$. However, under the assumptions of Lemma 2, the F statistic in Equation (5) will not follow an $F_{p,n-p-1}$ distribution under H_0, and it may produce asymptotically invalid results when used in a permutation test.

The assumption $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ may be questionable in many real data situations, where the error terms may be non-Gaussian and/or heteroscedastic. If the error terms are heteroscedastic, i.e., if $E(\epsilon_i^2) = \sigma_i^2$, then the F statistic may not produce valid results even in a permutation test (see [44,50,51]). Note that even if $\epsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$ for some non-Gaussian distribution, the F statistic may produce asymptotically invalid results when used in a permutation test. Instead, consider the Wald test statistic

$$W = n\hat{\boldsymbol{\beta}}_\Delta^\top \left[\hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}\hat{\boldsymbol{\Omega}}_X\hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}\right]^{-1}\hat{\boldsymbol{\beta}}_\Delta \tag{6}$$

where $\hat{\boldsymbol{\Omega}}_X = \frac{1}{n}\mathbf{X}_c^\top \mathbf{D}_Y^2 \mathbf{X}_c$ with $\mathbf{D}_Y = \text{diag}(Y_1^c, \ldots, Y_n^c)$. Under assumptions A1–A3, the W statistic asymptotically follows a χ_p^2 distribution when H_0 is true, which is a result of Lemma 1 and the consistency of the estimators $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\boldsymbol{\Omega}}_X$ under H_0.

2.4. Permutation Inference

Let the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ denote some permutation of the integers $\{1, \ldots, n\}$, and define $\boldsymbol{Y}_\pi = (Y_{\pi_1}^c, \ldots, Y_{\pi_n}^c)$ to be the permuted (and centered) response vector using the permutation vector $\boldsymbol{\pi}$. Furthermore, let $F(Y_\pi, X)$ and $W(Y_\pi, X)$ denote the test statistics from Equations (5) and (6) calculated using the permuted response vector $\boldsymbol{Y}_\pi$. When $\boldsymbol{X}_i$ is independent of ϵ_i, the permutation test conducted using $W(Y_\pi, X)$ will be exact, given that $\boldsymbol{\Omega}_X = E(\sigma_i^2)\boldsymbol{\Sigma}_X$ when $\boldsymbol{X}_i$ and ϵ_i are independent. However, the permutation test using $F(Y_\pi, X)$ is not guaranteed to be exact or asymptotically valid, given that the asymptotic sampling distribution of $F(Y, X)$ may not be the same as the permutation distribution (see [44]). When there is dependence between $\boldsymbol{X}_i$ and ϵ_i, the permutation test conducted using $F(Y_\pi, X)$ will be inexact and asymptotically invalid, whereas the permutation test conducted using $W(Y_\pi, X)$ will be inexact and asymptotically valid. Define the additional assumption A4: $E(Y_i^4) < \infty$ and $E(X_{ij}^4) < \infty$ for all j.

Theorem 1. *Consider the linear model with assumptions A1–A4. When $\boldsymbol{\beta} = \mathbf{0}_p$, the permutation distribution of $W(Y_\pi, X)$ converges to a χ_p^2 distribution as $n \to \infty$.*

Theorem 1 can be proved by combining the results in Lemma 1 with the results in Theorem 3.1 of DiCiccio and Romano [44], who derived the asymptotic nature of the

permutation distribution $W(Y_\pi, X)$ for the OLS estimator $\hat{\boldsymbol{\beta}}$. Note that Theorem 1 reveals that a permutation test using $W(Y_\pi, X)$ will produce asymptotically level α rejection rates when the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}_p$ is true.

Let the vector $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ denote a resigning vector where $\psi_i \in \{-1, 1\}\ \forall i$, and define $\boldsymbol{Y}_\psi = (\psi_1 Y_1^c, \ldots, \psi_n Y_n^c)$ to be the resigned (and centered) response vector using the resigning vector $\boldsymbol{\psi}$. Furthermore, let $F(Y_\psi, X)$ and $W(Y_\psi, X)$ denote the test statistics from Equations (5) and (6) calculated using the resigned response vector $\boldsymbol{Y}_\psi$. When $\boldsymbol{X}_i$ is independent of ϵ_i and the errors are symmetric, the permutation test conducted using $W(Y_\psi, X)$ will be exact, but the permutation test using $F(Y_\psi, X)$ may be invalid. When $\boldsymbol{X}_i \not\perp\!\!\!\perp \epsilon_i$, the permutation test conducted using $F(Y_\psi, X)$ will be inexact and asymptotically invalid, whereas the permutation test conducted using $W(Y_\psi, X)$ will be inexact and asymptotically valid, regardless of whether or not errors are symmetric.

Theorem 2. *Consider the linear model with assumptions A1–A4. When $\boldsymbol{\beta} = \mathbf{0}_p$, the permutation distribution of $W(Y_\psi, X)$ converges to a χ^2_p distribution as $n \to \infty$.*

Theorem 2 can be proved by combining the results in Lemma 1 with the results in Theorem 3.2 of DiCiccio and Romano [44], who derived the asymptotic nature of the permutation distribution $W(Y_\psi, X)$ for the OLS estimator $\hat{\boldsymbol{\beta}}$. Note that Theorem 2 reveals that a permutation test using $W(Y_\psi, X)$ will produce asymptotically level α rejection rates when the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}_p$ is true.

Corollary 1. *Define $\boldsymbol{Y}_{\pi\psi} = (\psi_1 Y^c_{\pi_1}, \ldots, \psi_n Y^c_{\pi_n})$ to be the permuted and resigned (centered) response vector using the permutation vector $\boldsymbol{\pi}$ and resigning vector $\boldsymbol{\psi}$. Consider the linear model with assumptions A1–A4. When $\boldsymbol{\beta} = \mathbf{0}_p$, the permutation distribution of $W(Y_{\pi\psi}, X)$ converges to a χ^2_p distribution as $n \to \infty$, where $W(Y_{\pi\psi}, X)$ is the test statistic in Equation (6), calculated using the permuted and resigned response vector $\boldsymbol{Y}_{\pi\psi}$.*

Corollary 1 follows directly from the results of Theorems 1 and 2.

3. Conditional Regression Tests

3.1. Model and Estimation

Given an independent sample of n observations, consider the linear regression model

$$Y_i = \alpha + \boldsymbol{X}_i^\top \boldsymbol{\beta} + \boldsymbol{Z}_i^\top \boldsymbol{\gamma} + \epsilon_i \tag{7}$$

for $i \in \{1, \ldots, n\}$, where Y_i is the i-th realization of the response variable $Y \in \mathbb{R}$, $\boldsymbol{X}_i^\top = (X_{i1}, \ldots, X_{ip})$ and $\boldsymbol{Z}_i^\top = (Z_{i1}, \ldots, Z_{iq})$ are the i-th realizations of the predictor vectors $\boldsymbol{X}^\top = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and $\boldsymbol{Z}^\top = (Z_1, \ldots, Z_q) \in \mathbb{R}^q$, and ϵ_i is the i-th realization of the error term $\epsilon \in \mathbb{R}$. The predictor variables are assumed to be partitioned into two sets, such that $\boldsymbol{X}$ contains the variables of interest for inference purposes, and $\boldsymbol{Z}$ contains the covariates that will be conditioned on. Let $\boldsymbol{M}_i^\top = (\boldsymbol{X}_i^\top, \boldsymbol{Z}_i^\top)$ denote the combined predictor vector, and let $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)^\top$ denote the combined coefficient vector. Furthermore, let $\boldsymbol{M}_i^c = \boldsymbol{M}_i - \bar{\boldsymbol{M}}$ denote the mean centered (combined) predictor vector, where $\bar{\boldsymbol{M}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{M}_i$ is the sample average of the combined predictor vector.

To estimate the coefficients in $\boldsymbol{\theta}$, consider minimizing the GRR loss function

$$\frac{1}{n}(\boldsymbol{Y}_c - \mathbf{M}_c\boldsymbol{\theta})^\top(\boldsymbol{Y}_c - \mathbf{M}_c\boldsymbol{\theta}) + \boldsymbol{\theta}^\top\boldsymbol{\Delta}\boldsymbol{\theta} \tag{8}$$

where $\mathbf{M}_c = [\boldsymbol{M}_1^c, \ldots, \boldsymbol{M}_n^c]^\top$ is the centered design matrix, and $\boldsymbol{\Delta}$ is an $r \times r$ symmetric and positive semi-definite penalty matrix (where $r = p + q$ is the total number of slope coefficients). The coefficients that minimize Equation (8) can be written as

$$\hat{\boldsymbol{\theta}}_\Delta = \left(\frac{1}{n}\mathbf{M}_c^\top \mathbf{M}_c + \boldsymbol{\Delta}\right)^{-1}\left(\frac{1}{n}\mathbf{M}_c^\top \boldsymbol{Y}_c\right) \tag{9}$$

which is subscripted to emphasize that the estimated coefficients depend on the penalty matrix $\boldsymbol{\Delta}$. Given the estimated slope vector $\hat{\boldsymbol{\theta}}_\Delta$, the least squares estimate of the intercept has the form $\hat{\alpha} = \bar{Y} - \bar{\boldsymbol{M}}^\top \hat{\boldsymbol{\theta}}_\Delta$.

3.2. Asymptotic Distributions

Consider the linear regression model from Equation (7) with the assumptions:

B1. $Y_i = \alpha + \boldsymbol{M}_i^\top \boldsymbol{\theta} + \epsilon_i$ with $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma_i^2 < \infty$ for $i = 1, \ldots, n$

B2. $(Y_i, \boldsymbol{M}_i)$ are iid from a distribution satisfying $E(\epsilon_i | \boldsymbol{M}_i) = 0$

B3. $\boldsymbol{\Sigma}_M = E\big((\boldsymbol{M}_i - \boldsymbol{\mu}_M)(\boldsymbol{M}_i - \boldsymbol{\mu}_M)^\top\big)$ and $\boldsymbol{\Omega}_M = E\big(\epsilon_i^2(\boldsymbol{M}_i - \boldsymbol{\mu}_M)(\boldsymbol{M}_i - \boldsymbol{\mu}_M)^\top\big)$ are non-singular, where $\boldsymbol{\mu}_M = E(\boldsymbol{M}_i)$, and $\frac{1}{n}\mathbf{M}_c^\top \mathbf{M}_c + \boldsymbol{\Delta}$ is almost surely invertible. In the fixed predictors case, the mean vector is $\boldsymbol{\mu}_M = \frac{1}{n}\sum_{i=1}^n \boldsymbol{M}_i$ and the covariance matrix terms are defined as $\boldsymbol{\Sigma}_M = \frac{1}{n}\mathbf{M}_c^\top \mathbf{M}_c$ and $\boldsymbol{\Omega}_M = \frac{1}{n}\mathbf{M}_c^\top \boldsymbol{\Psi} \mathbf{M}_c$, where $\boldsymbol{\Psi} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.

Given assumptions B1–B3, the GRR estimator provides an estimate of $\boldsymbol{\theta}_\Delta = \boldsymbol{\Sigma}_{M\Delta}^{-1}\boldsymbol{\sigma}_{MY}$, where $\boldsymbol{\Sigma}_{M\Delta} = \boldsymbol{\Sigma}_M + \boldsymbol{\Delta}$ and $\boldsymbol{\sigma}_{MY} = E((\boldsymbol{M}_i - \boldsymbol{\mu}_M)(Y_i - \mu_Y))$ is the covariance between $\boldsymbol{M}_i$ and Y_i. Note that $\boldsymbol{\theta}_\Delta = \boldsymbol{\Sigma}_{M\Delta}^{-1}\boldsymbol{\Sigma}_M\boldsymbol{\theta}$, where $\boldsymbol{\theta} = \boldsymbol{\Sigma}_M^{-1}\boldsymbol{\sigma}_{MY}$ is estimated by the OLS estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\Sigma}}_M^{-1}\hat{\boldsymbol{\sigma}}_{MY}$ with $\hat{\boldsymbol{\Sigma}}_M = \frac{1}{n}\mathbf{M}_c^\top \mathbf{M}_c$ and $\hat{\boldsymbol{\sigma}}_{MY} = \frac{1}{n}\mathbf{M}_c^\top \boldsymbol{Y}_c$. This implies that the GRR estimator can be written as $\hat{\boldsymbol{\theta}}_\Delta = \hat{\boldsymbol{\Sigma}}_{M\Delta}^{-1}\hat{\boldsymbol{\Sigma}}_M\hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\Sigma}}_{M\Delta} = \hat{\boldsymbol{\Sigma}}_M + \boldsymbol{\Delta}$.

Lemma 3. *Given assumptions B1–B3, the GRR estimator $\hat{\boldsymbol{\theta}}_\Delta$ from Equation (9) is asymptotically normal with mean vector $\boldsymbol{\theta}_\Delta$ and covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}_{M\Delta}^{-1}\boldsymbol{\Omega}_M\boldsymbol{\Sigma}_{M\Delta}^{-1}$, i.e.,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_\Delta - \boldsymbol{\theta}_\Delta) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}_{M\Delta}^{-1}\boldsymbol{\Omega}_M\boldsymbol{\Sigma}_{M\Delta}^{-1}\right)$$

as $n \to \infty$, where the notation $\xrightarrow{d}$ denotes convergence in distribution.

Lemma 4. *Consider the linear model in Equation (7) with the assumptions $\boldsymbol{\theta} \sim N(\mathbf{0}_r, \frac{\sigma^2}{n}\boldsymbol{\Delta}^{-1})$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and suppose that $\boldsymbol{\theta}$ and ϵ_i are independent of one another. Under these assumptions, the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$ is multivariate normal with mean vector $\hat{\boldsymbol{\theta}}_\Delta$ and covariance matrix $\frac{\sigma^2}{n}\hat{\boldsymbol{\Sigma}}_{M\Delta}^{-1}$. The asymptotic mean vector is $\boldsymbol{\theta}_\Delta$ and the asymptotic covariance matrix is $\frac{\sigma^2}{n}\boldsymbol{\Sigma}_{M\Delta}^{-1}$.*

Note that Lemmas 3 and 4 can be proved using analogues of the results that were used to prove Lemmas 1 and 2. Specifically, for Lemma 3, we can use a direct analogue of the proof for Lemma 1 with $\boldsymbol{\theta}_\Delta$ replacing $\boldsymbol{\beta}_\Delta$ and the matrices $\boldsymbol{\Omega}_M$ and $\boldsymbol{\Sigma}_{M\Delta}$ replacing the matrices $\boldsymbol{\Omega}_X$ and $\boldsymbol{\Sigma}_{X\Delta}$. For Lemma 4, we can use a direct analogue of the proof for Lemma 2 with $\boldsymbol{\theta}_\Delta$ replacing $\boldsymbol{\beta}_\Delta$ and $\boldsymbol{\Sigma}_{M\Delta}$ replacing $\boldsymbol{\Sigma}_{X\Delta}$. For both cases, we need to replace the unpenalized coefficient vector $\boldsymbol{\beta}$ with the unpenalized coefficient vector $\boldsymbol{\theta}$.

3.3. Test Statistics

Consider the linear model in Equation (7), and suppose that we want to test the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}_p$ versus the alternative hypothesis $H_1 : \boldsymbol{\beta} \neq \mathbf{0}_p$. This is the same null hypothesis that was considered in the previous section, but now the nuisance effects $\boldsymbol{Z}_i^\top \boldsymbol{\gamma}$ are included in the model (i.e., conditioned on) while testing the significance of the $\boldsymbol{\beta}$ vector. Assuming that $\boldsymbol{\theta} \sim N(\mathbf{0}_r, \frac{\sigma^2}{n}\boldsymbol{\Delta}^{-1})$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ are independent of one another, we could use the F test statistic

$$F = \frac{n}{p}\hat{\boldsymbol{\beta}}_\Delta^\top\left(\hat{\sigma}^2\mathbf{S}\hat{\boldsymbol{\Sigma}}_{M\Delta}^{-1}\mathbf{S}^\top\right)^{-1}\hat{\boldsymbol{\beta}}_\Delta \tag{10}$$

where $\mathbf{S} = [\mathbf{I}_p, \mathbf{0}_{p\times q}]$ is a $p \times r$ selection matrix such that $\mathbf{S}\boldsymbol{\theta}_\Delta = \boldsymbol{\beta}_\Delta$ (note that $\mathbf{I}_p$ is the $p \times p$ identity matrix and $\mathbf{0}_{p\times q}$ is a $p \times q$ matrix of zeros), $\hat{\sigma}^2 = \frac{1}{n-r-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2$ is the estimated error variance, and $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ is the residual with $\hat{Y}_i = \hat{\alpha} + \boldsymbol{M}_i^\top\hat{\boldsymbol{\theta}}_\Delta$ denoting the fitted value for the i-th observation.

When H_0 is true and the assumptions in Lemma 4 are met, the F statistic approaches an F distribution with degrees of freedom parameters p and $n-r-1$ as $\boldsymbol{\Delta} \to \mathbf{0}$. For non-zero penalties, the F statistic in Equation (10) will not follow an $F_{p,n-r-1}$ distribution, and may produce asymptotically invalid results when used in a permutation test, especially when the error terms are heteroscedastic (see [44,50,51]). In such cases, the Wald test statistic should be preferred

$$W = n\hat{\boldsymbol{\beta}}_\Delta^\top\left[\mathbf{S}\hat{\boldsymbol{\Sigma}}_{M\Delta}^{-1}\hat{\boldsymbol{\Omega}}_M\hat{\boldsymbol{\Sigma}}_{M\Delta}^{-1}\mathbf{S}^\top\right]^{-1}\hat{\boldsymbol{\beta}}_\Delta \tag{11}$$

where $\hat{\boldsymbol{\Omega}}_M = \frac{1}{n}\mathbf{M}_c^\top\mathbf{D}_{\hat{\epsilon}}^2\mathbf{M}_c$ with $\mathbf{D}_{\hat{\epsilon}} = \text{diag}(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$. Under assumptions B1–B3, the W statistic asymptotically follows a χ_p^2 distribution when H_0 is true, which is a result of Lemma 3 (and the consistency of the estimators $\hat{\boldsymbol{\Sigma}}_M$ and $\hat{\boldsymbol{\Omega}}_M$).

3.4. Permutation Inference

Table 1 depicts eight different permutation methods that have been proposed for testing the significance of regression coefficients in the presence of nuisance parameters. The eight methods can be split into three different groups: (i) methods that permute the rows of $\mathbf{X}$ [52–54], (ii) methods that permute $\boldsymbol{Y}$ with $\mathbf{Z}$ included in the model [55–57], and (iii) methods that permute $\boldsymbol{Y}$ after partialling out $\mathbf{Z}$ [58–60]. All of these methods were originally proposed for use with the OLS estimator $\hat{\boldsymbol{\theta}}$ and the F test statistic. Recent works have incorporated the use of the robust W test statistic with these permutation methods [44,50,51]. However, these authors only considered a theoretical analysis of the DS and FL permutation methods, and no previous works seem to have studied these methods using the GRR estimators from Equations (4) and (9).

Table 1. Permutation methods for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}_p$ in the presence of the nuisance parameters $\boldsymbol{\gamma}$. The intercept α is excluded from each model for notational simplicity.

Code	Method	Permutation Method
DS	Draper-Stoneman (1966)	$Y = PX\beta + Z\gamma + \epsilon$
OS	O'Gorman-Smith (2005/8)	$Y = PR_ZX\beta + Z\gamma + \epsilon$
MA	Manly (1986)	$PY = X\beta + Z\gamma + \epsilon$
FL	Freedman-Lane (1983)	$(H_Z + PR_Z)Y = X\beta + Z\gamma + \epsilon$
TB	ter Braak (1992)	$(H_M + PR_M)Y = X\beta + Z\gamma + \epsilon$
SW	Still-White (1981)	$PR_ZY = X\beta + \epsilon$
KC	Kennedy-Cade (1996)	$PR_ZY = R_ZX\beta + \epsilon$
HJ	Huh-Jhun (2001)	$PQ'R_ZY = Q'R_ZX\beta + \epsilon$

Notes. P is a permutation matrix. $R_Z = I - H_Z$ where $H_Z = Z(Z^\top Z + n\Delta_Z)^{-1}Z^\top$ is the hat matrix with Z in the model. $R_M = I - H_M$ where $H_M = M(M^\top M + n\Delta)^{-1}M^\top$ is the hat matrix with $M = (X, Z)$ in the model. For the Huh-Jhun method, $R_Z = QQ^\top$ where the columns of Q are mutually orthogonal.

To understand the motivation of the various permutation methods in Table 1, assume that the penalty matrix is a block diagonal such as $\boldsymbol{\Delta} = \text{bdiag}(\boldsymbol{\Delta}_X, \boldsymbol{\Delta}_Z)$, where $\boldsymbol{\Delta}_X$ and $\boldsymbol{\Delta}_Z$ denote the penalty matrices for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. Using the well-known form for the inverse of a block matrix [61–64], the coefficient estimates from Equation (9) have the form

$$\hat{\boldsymbol{\beta}}_\Delta = \left(\mathbf{X}_c^\top\mathbf{R}_Z\mathbf{X}_c + n\boldsymbol{\Delta}_X\right)^{-1}\mathbf{X}_c^\top\mathbf{R}_Z\boldsymbol{Y}_c$$
$$\hat{\boldsymbol{\gamma}}_\Delta = \left(\mathbf{Z}_c^\top\mathbf{Z}_c + n\boldsymbol{\Delta}_Z\right)^{-1}\mathbf{Z}_c^\top\left(\boldsymbol{Y}_c - \mathbf{X}_c\hat{\boldsymbol{\beta}}_\Delta\right)$$

where $\mathbf{R}_Z = \mathbf{I}_n - \mathbf{Z}_c(\mathbf{Z}_c^\top \mathbf{Z}_c + n\boldsymbol{\Delta}_Z)^{-1}\mathbf{Z}_c^\top$ is the residual forming matrix for the model that only includes the nuisance effects in the model, i.e., $\boldsymbol{Y} = \alpha + \boldsymbol{Z}^\top\gamma + \epsilon$. This implies that the (centered) fitted values can be written as $\hat{\boldsymbol{Y}}_c = \mathbf{X}_c\hat{\boldsymbol{\beta}}_\Delta + \mathbf{Z}_c\hat{\gamma}_\Delta = \mathbf{R}_Z\mathbf{X}_c\hat{\boldsymbol{\beta}}_\Delta + \mathbf{H}_Z\boldsymbol{Y}_c$, where $\mathbf{H}_Z = \mathbf{Z}_c\left(\mathbf{Z}_c^\top\mathbf{Z}_c + n\boldsymbol{\Delta}_Z\right)^{-1}\mathbf{Z}_c^\top$ is the hat matrix for the linear model that only includes the nuisance effects. Thus, when $\boldsymbol{\Delta}_Z = \mathbf{0}_{q\times q}$, all of the permutation methods except SW will produce the same observed F statistic (when the permutation matrix is $\mathbf{P} = \mathbf{I}_n$).

Consider the additional assumption that the response and predictor variables have finite fourth moments, i.e., B4: $E(Y_i^4) < \infty$, $E(X_{ij}^4) < \infty\ \forall j$, and $E(Z_{ik}^4) < \infty\ \forall k$. Assume B1–B4 and that the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}_p$ is true. Using the W test statistic from Equation (11), the following can be said about the finite sample and asymptotic properties of the various permutation methods in Table 1: the DS method is exact when $\boldsymbol{X} \perp\!\!\!\perp (Y, \boldsymbol{Z})$ and asymptotically valid otherwise; the MA method is exact when $Y \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{Z})$ and asymptotically valid otherwise; the SW method is inexact and asymptotically valid only when $E\left((\boldsymbol{X} - \boldsymbol{\mu}_X)(\boldsymbol{Z} - \boldsymbol{\mu}_Z)^\top\right) = \mathbf{0}_{p\times q}$; the other five methods (OS, FL, TB, KC, HJ) are inexact and asymptotically valid.

The asymptotic behaviors of the DS and FL methods were proved by DiCiccio and Romano [44]. The asymptotic validity of the OS method can be proved using a similar result as used for the DS method, given that $\frac{1}{n}\mathbf{X}_c^\top\mathbf{R}_Z\mathbf{X}_c$ is a consistent estimator of $\boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_{XZ}\boldsymbol{\Sigma}_{Z\Delta}^{-1}\boldsymbol{\Sigma}_{ZX}$ and $\frac{1}{n}\mathbf{X}_c^\top\mathbf{R}_Z\boldsymbol{Y}_c$ is a consistent estimator of $\boldsymbol{\sigma}_{XY} - \boldsymbol{\Sigma}_{XZ}\boldsymbol{\Sigma}_{Z\Delta}^{-1}\boldsymbol{\sigma}_{ZY}$. The asymptotic validity of the MA method can also be proved using a similar result as used for the DS method. The asymptotic validity of the TB method can be proved using a similar result as used for the FL method, given that $\frac{1}{n}\mathbf{M}_c^\top\mathbf{R}_M\mathbf{M}_c$ is a consistent estimator of $\boldsymbol{\Sigma}_M - \boldsymbol{\Sigma}_M\boldsymbol{\Sigma}_{M\Delta}^{-1}\boldsymbol{\Sigma}_M$ and $\frac{1}{n}\mathbf{M}_c^\top\mathbf{R}_M\boldsymbol{Y}_c$ is a consistent estimator of $\boldsymbol{\sigma}_{MY} - \boldsymbol{\Sigma}_M\boldsymbol{\Sigma}_{M\Delta}^{-1}\boldsymbol{\sigma}_{MY}$. Finally, note that the KC and HJ methods are asymptotically equivalent, and the SW method is asymptotically equivalent to the KC and HJ methods when $\boldsymbol{X}$ and $\boldsymbol{Z}$ are uncorrelated. The asymptotic validity of the KC and HJ methods follow from the results in Theorem 1, given that these methods permute the response after partialling out the nuisance effects. It is important to note that if $\boldsymbol{X}$ and $\boldsymbol{Z}$ are correlated, the SW method will produce asymptotically invalid results because $\boldsymbol{Z}$ is partialled out of $\boldsymbol{Y}$ but not $\boldsymbol{X}$.

4. Simulation Studies

4.1. Simulation A

The first simulation study was designed to explore the validity of the claims made about the omnibus permutation tests discussed in Section 2. Simulation A was a fully crossed design that manipulated three design factors: (i) the data generating distribution (three levels: left skew normal, standard normal, right skew normal), (ii) the error standard deviation (three levels: constant, increasing, and parabolic), and (iii) the sample size (five levels: $n \in \{10, 25, 50, 100, 200\}$), see Figure 1. For each of the 45 combinations of simulation design parameters (3 F_ϵ × 3 σ × 5 n), I generated 10,000 independent copies of the data from the model in Equation (1) with $X_i = (i-1)/(n-1)$ and $\eta(X_i) = 0$ (so that $Y_i = \epsilon_i$). For each generated sample of data, the *ss()* function in the **npreg** R package [65] was used to fit a cubic smoothing spline with $r = 5$ knots, which were placed at the quantiles of the predictor scores. The generalized cross-validation (GCV) criterion [33] was used to select the smoothing parameter λ.

Ten different inference methods (six permutation tests and four parametric tests) were used to test the null hypothesis of no functional relationship. The six permutation tests were formed using the two test statistics (F and W) combined with the three permutation methods discussed in Section 2: permute Y, sign-flip Y, and both permuting and sign flipping. The permutation tests were implemented using the *np.reg.test()* function in the **nptest** R package [66] using the default number of resamples (the default uses $R = \min(R_0, 9999)$ resamples, where R_0 is the number of elements of the exact null distribution: $R_0 = n!$ for the permute method, $R_0 = 2^n$ for the sign-flip method, and $R_0 = n!2^n$ for the combined method). The first two parametric tests were formed by comparing the F statistic from Equation (5) to an $F_{p,n-p-1}$ distribution, and the W statistic from Equation (6) to a χ_p^2

distribution. The other two parametric tests were the F tests that are implemented by the **mgcv** R package [67] and the **npreg** R package [65]. Note that these F tests compare the F statistic from Equation (5) to an $F_{\nu,n-\nu-1}$ distribution, where ν is an estimate of the effective degrees of freedom. The **npreg** package defines ν as the trace of the smoothing matrix, whereas the **mgcv** package uses a more complex estimate of ν (see [42]).

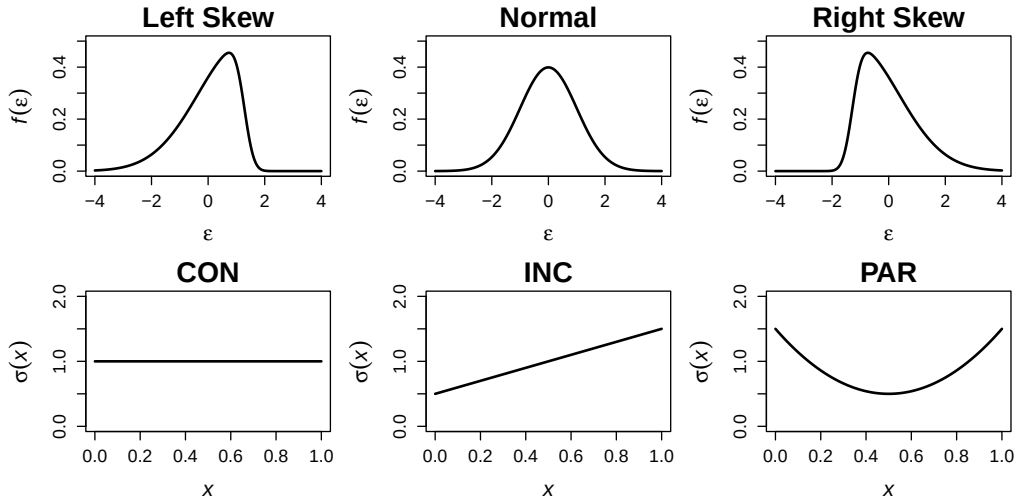

Figure 1. Simulation A Design. The top row shows the three data generating distributions, which each have mean $\mu = 0$ and variance $\sigma^2 = 1$. The bottom row shows the three error standard deviations: CON is constant error standard deviation with $\sigma(x) = 1$, INC is increasing the error standard deviation with $\sigma(x) = 1/2 + x$, and PAR is a parabolic error standard deviation with $\sigma(x) = 1/2 + 4(x - 1/2)^2$. All nine combinations (= 3 $f(\epsilon)$ distributions $\times$ 3 $\sigma(x)$ patterns) of these data generating factors were explored in the simulation.

Figure 2 displays the type I error rate for each inference method in each combination of Simulation A conditions. All of the inference methods perform similarly across the three data generating distributions, so the following discussions of the results apply to each of the three data generating distributions. When using the W test statistic, the three permutation methods produced accurate type I error rates for all combinations of data generating conditions (see red square, blue circle, and green triangle), and the parametric χ^2_p approximation produced asymptotically accurate results (see purple plus sign). When using the F test statistic, the permutation methods produced inflated type I error rates (see orange x, yellow diamond, and brown upside-down triangle), and the parametric $F_{p,n-p-1}$ approximation produced inconsistent results across the different error standard deviation conditions (see pink square with x). The parametric $F_{\nu,n-\nu-1}$ tests implemented by the **mgcv** and **npreg** packages also produce inconsistent results across the different error standard deviation conditions, such that the type I error rate is inflated in the increasing and parabolic conditions (see gray asterisk and black diamond with plus sign).

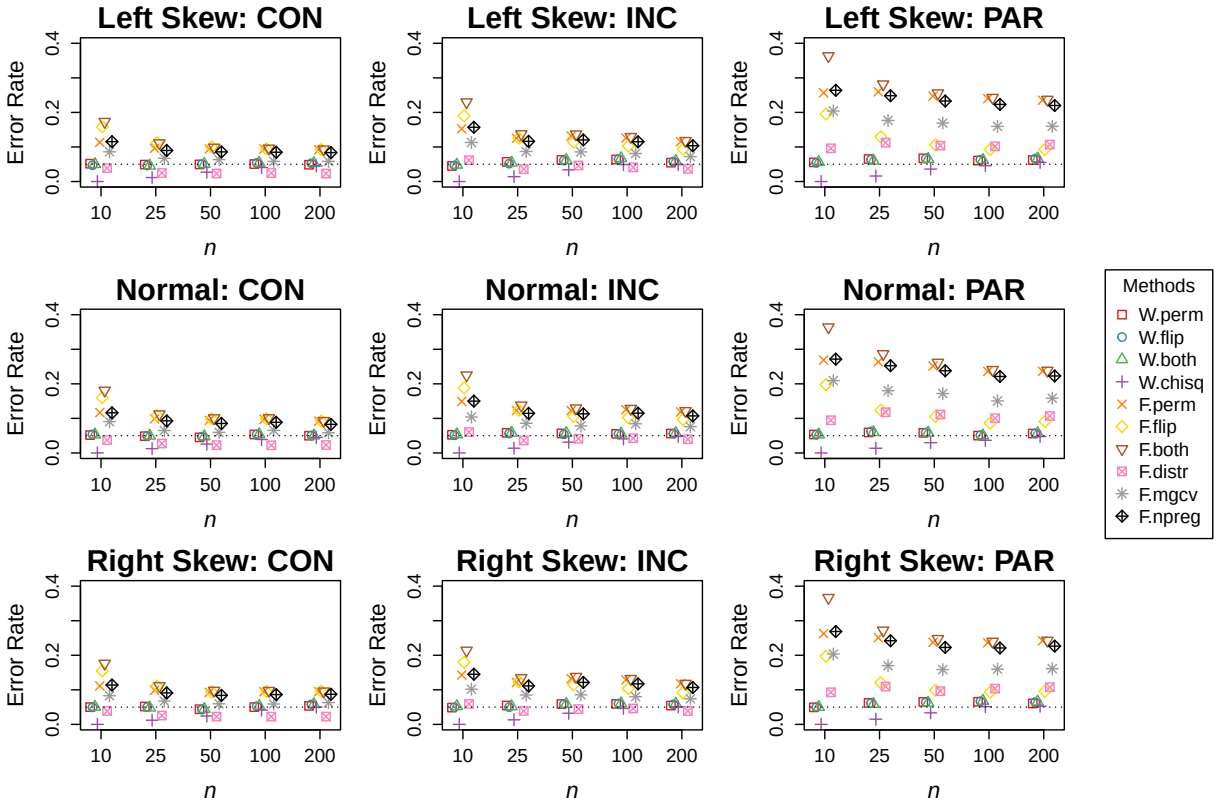

Figure 2. Simulation A Results. Within each subplot, the Type I error rate is plotted for each inference method at each sample size. The rows denote the three different data generating distributions, and the columns denote the three different error standard deviation functions. The nominal $\alpha = 0.05$ rate is denoted with a dotted line.

4.2. Simulation B

The second simulation study was designed to explore the validity of the claims made about the conditional permutation tests discussed in Section 3. Simulation B was a fully crossed design that manipulated two design factors: (i) the error standard deviation (three levels: constant, increasing, and parabolic), and (ii) the sample size (five levels: $n \in \{10, 25, 50, 100, 200\}$). Given that the results in Simulation A did not noticeably differ across the three data generating distributions, errors were generated from a normal distribution throughout Simulation B. For each of the 15 combinations of data generating parameters ($3\ \sigma \times 5\ n$), I generated 10,000 independent copies of the data from the model in Equation (1) with $X_i = (i-1)/(n-1)$ and $\eta(X_i) = X_i$ (unlike Simulation A, the data generating mean function now includes a linear effect). As in the previous simulation, the *ss()* function in the **npreg** R package [65] was used to fit a cubic smoothing spline with $r = 5$ knots, and the GCV criterion was used to select the smoothing parameter.

The null hypothesis of a linear relationship was tested using a total of 19 inference methods: 16 permutation tests and 3 parametric tests. The 16 permutations tests were formed using the two test statistics (F and W) combined with the eight permutation methods in Table 1. As in Simulation A, the permutation tests were implemented using the *np.reg.test()* function in the **nptest** R package [66] using the default number of resamples (i.e., $R = 9999$) to form the permutation distribution. The first two parametric tests were formed by comparing the F statistic from Equation (10) to an $F_{p,n-r-1}$ distribution, and the W statistic from Equation (11) to a χ^2_p distribution. The other parametric test is the F test that is implemented by the **npreg** package, which compares the F statistic from Equation (10) to an $F_{v-m,n-v-1}$ distribution. Note that the "cardinal" spline parameterization used in the **mgcv** R package does not separate the linear and non-linear portions of the function; thus, a test of linearity is not possible using this package.

Figure 3 displays the type I error rate for each inference method in each combination of Simulation B conditions. When the errors have constant variance, using the F statistic produces (i) inflated type I error rates using all permutation methods, (ii) deflated type I error rates using the parametric test, and (iii) asymptotically accurate error rates using the **npreg** test. In contrast, when using the W test statistic, all of the inference methods except SW produce asymptotically accurate error rates when the errors have constant variance. As expected, the F statistic produces asymptotically invalid results when the errors have non-constant variance, with the performance being the worst in the parabolic condition. In contrast, the W statistic produced asymptotically valid type I error rates when the errors have non-constant variance (using all methods except SW).

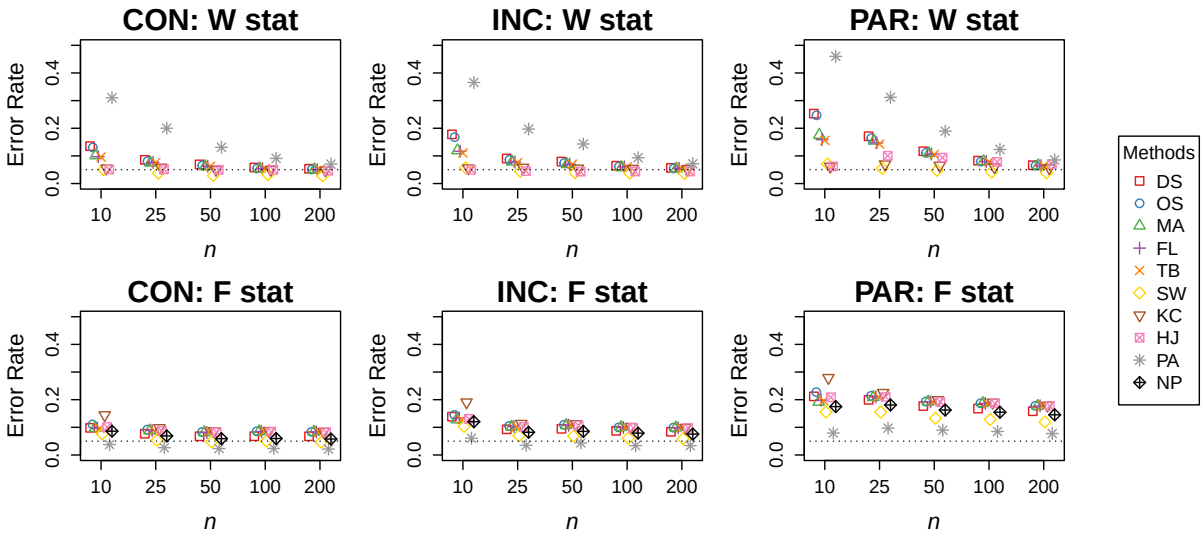

Figure 3. Simulation B results. Within each subplot, the Type I error rate is plotted for each inference method at each sample size. The rows denote the two different test statistics (W and F), and the columns denote the three different error standard deviation functions. The nominal $\alpha = 0.05$ rate is denoted with a dotted line.

5. Discussion

5.1. Summary of Findings

Penalized splines are frequently used in applied research for understanding functional relationships between variables. Although there has been a considerable body of work on the estimation and computation of penalized splines, there have been relatively few papers on statistical inference about the nature of the functional relation. In most applications, the statistical inference for penalized splines is conducted using the random effects or Bayesian interpretation of a smoothing spline. These inferential frameworks rely on parametric assumptions about the unknown function and error terms, i.e., that η is a Gaussian process and ϵ_i are iid Gaussian variables. Even when these parametric assumptions are met, valid statistical inference can be challenging due to the need for a reasonable estimate of the degrees of freedom of the penalized spline estimate. Furthermore, when these parametric assumptions are incorrect (e.g., due to heteroscedastic and/or non-Gaussian errors), the standard inferential tools can produce substantially misleading results.

In this paper, I developed a flexible and robust permutation testing framework for inference using penalized splines. Unlike the existing methods for statistical inference with penalized splines, the proposed methods can provide asymptotically valid inferential results under a variety of data generating situations. Furthermore, unlike a majority of the existing methods for hypothesis testing with penalized splines, the proposed approach can be flexibly adapted to test a variety of standard and non-standard hypotheses about the nature of the functional relationships between variables. The omnibus test in Section 2 is frequently of interest in practical applications, but has been ignored by most inferential

tests for penalized splines (e.g., see [42]). The conditional test in Section 3 can be used to test the classic hypothesis $\eta \in \mathcal{H}_0$ (which has been considered in many papers), as well as the more specialized hypotheses about the main and/or interaction effects of predictors.

The simulation results in Section 4 clearly demonstrate the benefits of the proposed approach over standard methods used for inference with penalized splines. In particular, the simulation results demonstrate that classic (parametric) methods that rely on an F test statistic can produce substantially inflated type I error rates when the error terms are heteroscedastic. Furthermore, the simulation results demonstrate that the F test statistic can even produce inaccurate results when used in a permutation test. In contrast, the permutation tests using the robust W test statistic can produce exact results for the omnibus tests and asymptotically valid results for the conditional tests—even when the errors depend on the predictor(s). Moreover, the χ^2_p approximation using the robust W statistic produced asymptotically valid results that were reasonably close to the nominal $\alpha = 0.05$ rate with only $n = 100$ (for the omnibus test) or $n = 200$ (for the conditional test).

5.2. Future Directions

Although the simulation studies only explored the omnibus and conditional tests with a single predictor, the theoretical results in Sections 2 and 3 can be readily applied to penalized spline models with multiple predictors. With $d > 1$ predictors, the number of possible hypotheses that could be tested increases, given that it is possible to test omnibus or conditional tests about any of the main or interaction effects. The simulation results for a single predictor revealed that the permutation methods of Kennedy and Cade [59] and Huh and Jhun [60] performed the best for conditional tests with a single predictor, so I hypothesize that these procedures will be ideal for tests of main and/or interaction effects with $d > 1$ predictors. However, future theoretical and Monte Carlo simulation work is needed to determine which permutation strategies should be preferred for conditional tests of main and/or interaction effects of multiple predictors.

The theoretical results in Sections 2 and 3 were developed for the classic formulation of penalized splines, which solves a GRR problem to obtain the function estimate. Note that the penalized spline GRR problem is a special case of the elastic net penalty [68] that is used in the recently proposed kernel eigenvector smoothing and selection operator (kesso) regression method [26]. Although the theorems derived in Sections 2 and 3 assume a ridge penalty, it is straightforward to develop extensions of these formulas for elastic net penalties used in the kesso regression. Specifically, the vector version of the coefficients, which is given after Equation (8) in [26], can be used to express the kesso coefficients as a modification of the least squares coefficients. However, future Monte Carlo research is needed to explore the accuracy of the various permutation strategies when using elastic net penalties to fit penalized spline regression models.

Finally, it is worth noting that this paper only considered the GCV tuning method, which is one of several smoothing parameter selection methods available in the *ss()* function. Recent work has demonstrated that the different tuning methods tend to produce similar results across a wide variety of data generating conditions—including non-normal and/or heteroscedastic errors (see [25]). However, it was noted that the maximum likelihood based tuning method tended to perform (i) worse than the cross-validation based methods when the sample size was small and (ii) better than the cross-validation based methods when the errors were correlated. Given these past findings, I would not expect the simulation results in this paper to significantly change if a different tuning criterion were used instead of the GCV. However, future research should explore the performance of the proposed permutation testing approach using different combinations of tuning methods and permutation strategies to analyze data from a wide variety of data-generating conditions.

Funding: This research was funded by the following National Institutes of Health (NIH) grants: R01EY030890, R01MH115046, U01DA046413, and R43AG074740.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The supporting information contains the following files:

simA	R script for reproducing simulation A (.R file)
simB	R script for reproducing simulation B (.R file)
sstest	R function for omnibus and conditional smoothing spline permutation tests (.R file)

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BLUE	Best Linear Unbiased Estimator
BLUP	Best Linear Unbiased Predictor
GCV	Generalized Cross-Validation
GRR	Generalized Ridge Regression
OLS	Ordinary Least Squares
PLS	Penalized Least Squares

Appendix A. Proofs

Appendix A.1. Proof of Lemma 1

To prove Lemma 1, one needs to demonstrate that under assumptions A1-A3, the OLS estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\sigma}}_{XY}$ is asymptotically normal with mean vector $\boldsymbol{\beta} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\sigma}_{XY}$ and covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_X^{-1}$. In other words, one needs to demonstrate that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_X^{-1}\right)$$

where $\cdot \xrightarrow{d} \cdot$ denotes that the random vector on the lefthand side converges in distribution to the probability distribution specified on the righthand side. This result, which is originally due to White [46], can be derived using basic rules of expectation and covariance operators in combination with a multivariate central limit theorem.

First, under assumptions A1-A3, note that the expectation of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E\left(\hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\sigma}}_{XY}\right) \\ &= E\left((\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right) \\ &= E\left((\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top\mathbf{X}_c\boldsymbol{\beta} + (\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top\boldsymbol{\epsilon}\right) \\ &= \boldsymbol{\beta} \end{aligned}$$

where the second line is due to the fact that $\mathbf{X}_c^\top\mathbf{Y}_c = \mathbf{X}_c^\top\mathbf{Y}$, the third line is due to the facts that $\mathbf{X}_c^\top\mathbf{1}_n = \mathbf{0}_p$ and $\mathbf{X}_c^\top\mathbf{X} = \mathbf{X}_c^\top\mathbf{X}_c$ by definition, and the fourth line is due to the fact that $E(\mathbf{X}_c^\top\boldsymbol{\epsilon}) = \mathbf{0}_p$ by assumption A2.

Second, under assumptions A1–A3, note that the covariance matrix of $\hat{\boldsymbol{\beta}}$ satisfies

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}\left(\boldsymbol{\beta} + (\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top\boldsymbol{\epsilon}\right) \\ &= \mathrm{Cov}\left((\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top\boldsymbol{\epsilon}\right) \\ &= \mathrm{E}\left((\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\mathbf{X}_c^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\mathbf{X}_c(\mathbf{X}_c^\top\mathbf{X}_c)^{-1}\right) \\ &\asymp \frac{1}{n}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_X^{-1} \end{aligned}$$

where the first line uses the previous result from the expectation derivation, the second line uses the fact that $\boldsymbol{\beta}$ is a constant vector, the third line uses the fact that $E\left((\mathbf{X}_c^\top \mathbf{X}_c)^{-1}\mathbf{X}_c^\top \boldsymbol{\epsilon}\right) = \mathbf{0}_p$ by assumption A2, and the fourth line uses the definitions of $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Omega}_X$ given in assumption A3. Note that the notation $\asymp$ should be read as 'is asymptotically equal to'.

Thus, under assumptions A1-A3, the OLS coefficient estimates have mean vector $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and asymptotic covariance matrix $\text{Cov}(\hat{\boldsymbol{\beta}}) \asymp \frac{1}{n}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_X^{-1}$. The asymptotic multivariate normality of $\hat{\boldsymbol{\beta}}$ results from applying the multivariate central limit theorem using the consistent estimators $\hat{\boldsymbol{\Sigma}}_X = \frac{1}{n}\mathbf{X}_c^\top \mathbf{X}_c$ and $\hat{\boldsymbol{\Omega}}_X = \frac{1}{n}\mathbf{X}_c^\top \text{diag}(\hat{\boldsymbol{\epsilon}})\mathbf{X}_c$ in place of the unknown asymptotic covariance matrix components $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Omega}_X$.

Finally, note that the asymptotic expectation of $\hat{\boldsymbol{\beta}}_\Delta$ has the form

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_\Delta) &\asymp \boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Sigma}_X E(\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Sigma}_X\boldsymbol{\beta} \\ &= \boldsymbol{\beta}_\Delta \end{aligned}$$

and the asymptotic covariance of $\hat{\boldsymbol{\beta}}_\Delta$ has the form

$$\begin{aligned} \text{Cov}(\hat{\boldsymbol{\beta}}_\Delta) &\asymp \boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Sigma}_X \text{Cov}(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}_X\boldsymbol{\Sigma}_{X\Delta}^{-1} \\ &= \frac{1}{n}\boldsymbol{\Sigma}_{X\Delta}^{-1}\boldsymbol{\Omega}_X\boldsymbol{\Sigma}_{X\Delta}^{-1} \end{aligned}$$

given that $\hat{\boldsymbol{\beta}}_\Delta = \hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}\hat{\boldsymbol{\Sigma}}_X\hat{\boldsymbol{\beta}}$, which completes the proof.

Appendix A.2. Proof of Lemma 2

To prove Lemma 2, ones needs to demonstrate that under assumptions A1–A3 and the prior distribution assumptions, the posterior distribution of the coefficients is a multivariate normal with mean vector $\hat{\boldsymbol{\beta}}_\Delta = \hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}\hat{\boldsymbol{\sigma}}_{XY}$ and covariance matrix $\frac{\sigma^2}{n}\hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}$. In other words, one needs to demonstrate that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_\Delta - \boldsymbol{\beta}\right) \sim N\left(\mathbf{0}, \sigma^2\hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}\right)$$

where the sign of the term on the lefthand side has been flipped to emphasize the similarity in form to the result in Appendix A.1. This result, which was first given by Henderson [47,48], can be proven using a classic result of multivariate normal theory (e.g., see [69]).

Consider a partitioned vector $\boldsymbol{Z}$ that has a multivariate normal distribution, i.e.,

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{bmatrix}\right).$$

The posterior distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \mathbf{y}$ is multivariate normal, i.e., $(\boldsymbol{X}|\boldsymbol{Y} = \mathbf{y}) \sim N(\boldsymbol{\mu}_{X|Y}, \boldsymbol{\Sigma}_{X|Y})$, where the posterior mean vector and covariance matrix have the form

$$\begin{aligned} \boldsymbol{\mu}_{X|Y} &= \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}(\mathbf{y} - \boldsymbol{\mu}_Y) \\ \boldsymbol{\Sigma}_{X|Y} &= \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} \end{aligned}$$

In this case, $\boldsymbol{Z}^\top = (\boldsymbol{\beta}^\top, \boldsymbol{Y}^\top)$ where $\boldsymbol{Y} = \alpha\mathbf{1}_n + \mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is the response vector. Note that the prior mean vectors are $\boldsymbol{\mu}_\beta = \mathbf{0}_p$ and $\boldsymbol{\mu}_Y = \alpha\mathbf{1}_n$, and the prior covariance matrices are $\boldsymbol{\Sigma}_{\beta\beta} = \frac{\sigma^2}{n}\boldsymbol{\Delta}^{-1}$ and $\boldsymbol{\Sigma}_{YY} = \sigma^2(\frac{1}{n}\mathbf{X}_c\boldsymbol{\Delta}^{-1}\mathbf{X}_c^\top + \mathbf{I}_n)$, and the covariance between $\boldsymbol{\beta}$ and $\boldsymbol{Y}$ is $\boldsymbol{\Sigma}_{\beta Y} = \frac{\sigma^2}{n}\boldsymbol{\Delta}^{-1}\mathbf{X}_c^\top$. Thus, the posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ has mean vector

$$\begin{aligned}
\boldsymbol{\mu}_{\beta|Y} &= \boldsymbol{\mu}_{\beta} + \boldsymbol{\Sigma}_{\beta Y}\boldsymbol{\Sigma}_{YY}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{Y}) \\
&= \frac{1}{n}\boldsymbol{\Delta}^{-1}\mathbf{X}_c^{\top}\left(\frac{1}{n}\mathbf{X}_c\boldsymbol{\Delta}^{-1}\mathbf{X}_c^{\top} + \mathbf{I}_n\right)^{-1}\boldsymbol{Y}_c \\
&= \frac{1}{n}\boldsymbol{\Delta}^{-1}\mathbf{X}_c^{\top}\left(\mathbf{I}_n - \mathbf{X}_c\left(\mathbf{X}_c^{\top}\mathbf{X}_c + n\boldsymbol{\Delta}\right)^{-1}\mathbf{X}_c^{\top}\right)\boldsymbol{Y}_c \\
&= \left(\mathbf{X}_c^{\top}\mathbf{X}_c + n\boldsymbol{\Delta}\right)^{-1}\mathbf{X}_c^{\top}\boldsymbol{Y}_c
\end{aligned}$$

where the second line plugs in the definitions of the parameters, the third line uses Equation (17) of Henderson and Searle [70] to more conveniently write the matrix inverse, and the last line results from straightforward algebraic simplification of the third line. Similarly, the posterior covariance matrix of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ has the form

$$\begin{aligned}
\boldsymbol{\Sigma}_{\beta|Y} &= \boldsymbol{\Sigma}_{\beta\beta} - \boldsymbol{\Sigma}_{\beta Y}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{Y\beta} \\
&= \frac{\sigma^2}{n}\left(\boldsymbol{\Delta}^{-1} - \frac{1}{n}\boldsymbol{\Delta}^{-1}\mathbf{X}_c^{\top}\left(\frac{1}{n}\mathbf{X}_c\boldsymbol{\Delta}^{-1}\mathbf{X}_c^{\top} + \mathbf{I}_n\right)^{-1}\mathbf{X}_c\boldsymbol{\Delta}^{-1}\right) \\
&= \frac{\sigma^2}{n}\left(\boldsymbol{\Delta}^{-1} - \frac{1}{n}\boldsymbol{\Delta}^{-1}\mathbf{X}_c^{\top}\left(\mathbf{I}_n - \mathbf{X}_c\left(\mathbf{X}_c^{\top}\mathbf{X}_c + n\boldsymbol{\Delta}\right)^{-1}\mathbf{X}_c^{\top}\right)\mathbf{X}_c\boldsymbol{\Delta}^{-1}\right) \\
&= \sigma^2\left(\mathbf{X}_c^{\top}\mathbf{X}_c + n\boldsymbol{\Delta}\right)^{-1}
\end{aligned}$$

where the second line plugs in the definitions of the parameters, the third line plugs in the convenient representation of $\boldsymbol{\Sigma}_{YY}^{-1}$ from Equation (17) of Henderson and Searle [70], and the final line results from the straightforward algebraic manipulation of the third line. Noting that $\boldsymbol{\mu}_{\beta|Y} = \hat{\boldsymbol{\beta}}_{\Delta}$ and $\boldsymbol{\Sigma}_{\beta|Y} = \frac{\sigma^2}{n}\hat{\boldsymbol{\Sigma}}_{X\Delta}^{-1}$ completes the proof.

References

1. Fox, J. *Quantitative Applications in the Social Sciences: Multiple and Generalized Nonparametric Regression*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2000. [CrossRef]
2. 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 Publications, Inc.: London, England, 2020. [CrossRef]
3. Wahba, G. *Spline Models for Observational Data*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1990.
4. Wang, Y. *Smoothing Splines: Methods and Applications*; CRC Press: Boca Raton, FL, USA, 2011.
5. Gu, C. *Smoothing Spline ANOVA Models*, 2nd ed.; Springer: New York, NY, USA, 2013. [CrossRef]
6. Hastie, T.; Tibshirani, R. *Generalized Additive Models*; Chapman and Hall/CRC: New York, NY, USA, 1990.
7. Ruppert, D.; Wand, M.P.; Carroll, R.J. *Semiparametric Regression*; Cambridge University Press: Cambridge, UK, 2003.
8. Wood, S.N. *Generalized Additive Models: An Introduction with R*, 2nd ed.; Chapman & Hall: Boca Raton, FL, USA, 2017.
9. 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]
10. Kage, C.C.; Helwig, N.E.; Ellingson, A.M. Normative cervical spine kinematics of a circumduction task. *J. Electromyogr. Kinesiol.* **2021**, *61*, 102591. [CrossRef]
11. 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. doi: 10.1016/j.jbiomech.2016.07.035. [CrossRef] [PubMed]
12. 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]
13. 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]
14. 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]
15. 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]
16. Helwig, N.E. Regression with ordered predictors via ordinal smoothing splines. *Front. Appl. Math. Stat.* **2017**, *3*, 1–13. [CrossRef]
17. Gu, C. Nonparametric regression with ordinal responses. *Stat* **2021**, *10*, e365. [CrossRef]
18. Gu, C.; Ma, P. Optimal smoothing in nonparametric mixed-effect models. *Ann. Stat.* **2005**, *33*, 1357–1379. [CrossRef]

19. Gu, C.; Ma, P. Generalized Nonparametric Mixed-Effect Models: Computation and Smoothing Parameter Selection. *J. Comput. Graph. Stat.* **2005**, *14*, 485–504. [CrossRef]
20. Helwig, N.E. Efficient estimation of variance components in nonparametric mixed-effects models with large samples. *Stat. Comput.* **2016**, *26*, 1319–1336. [CrossRef]
21. 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]
22. Gu, C.; Kim, Y.J. Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Stat.* **2002**, *30*, 619–628. [CrossRef]
23. Helwig, N.E.; Ma, P. Fast and stable multiple smoothing parameter selection in smoothing spline analysis of variance models with large samples. *J. Comput. Graph. Stat.* **2015**, *24*, 715–732. [CrossRef]
24. Helwig, N.E.; Ma, P. Smoothing spline ANOVA for super-large samples: Scalable computation via rounding parameters. *Stat. Interface* **2016**, *9*, 433–444. [CrossRef]
25. 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. [CrossRef]
26. Helwig, N.E. Spectrally sparse nonparametric regression via elastic net regularized smoothers. *J. Comput. Graph. Stat.* **2021**, *30*, 182–191. [CrossRef]
27. Kimeldorf, G.; Wahba, G. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **1971**, *33*, 82–95. [CrossRef]
28. Ma, P.; Huang, J.; Zhang, N. Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika* **2015**, *102*, 631–645. [CrossRef]
29. Moore, E.H. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.* **1920**, *26*, 394–395. [CrossRef]
30. Penrose, R. A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc.* **1955**, *51*, 406–413. [CrossRef]
31. Wahba, G. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Stat. Soc. Ser. B* **1983**, *45*, 133–150. [CrossRef]
32. Nychka, D. Bayesian confidence intervals for smoothing splines. *J. Am. Stat. Assoc.* **1988**, *83*, 1134–1143. [CrossRef]
33. 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]
34. Gu, C.; Wahba, G. Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J. Comput. Graph. Stat.* **1993**, *2*, 97–117.
35. Marra, G.; Wood, S.N. Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Stat.* **2012**, *39*, 53–74. [CrossRef]
36. Cox, D.; Koh, E.; Wahba, G.; Yandell, B.S. Testing the (Parametric) Null Model Hypothesis in (Semiparametric) Partial and Generalized Spline Models. *Ann. Stat.* **1988**, *16*, 113–119. [CrossRef]
37. Zhang, D.; Lin, X. Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **2003**, *4*, 57–74. [CrossRef]
38. Liu, A.; Wang, Y. Hypothesis testing in smoothing spline models. *J. Stat. Comput. Simul.* **2004**, *74*, 581–597. [CrossRef]
39. Crainiceanu, C.; Ruppert, D.; Claeskens, G.; Wand, M.P. Exact likelihood ratio tests for penalised splines. *Biometrika* **2005**, *92*, 91–103. [CrossRef]
40. Scheipl, F.; Greven, S.; Küchenhoff, H. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput. Stat. Data Anal.* **2008**, *52*, 3283–3299. [CrossRef]
41. Nummi, T.; Pan, J.; Siren, T.; Liu, K. Testing for Cubic Smoothing Splines under Dependent Data. *Biometrics* **2011**, *67*, 871–875. [CrossRef] [PubMed]
42. Wood, S.N. On *p*-values for smooth components of an extended generalized additive model. *Biometrika* **2013**, *100*, 221–228. [CrossRef]
43. Wood, S.N. A simple test for random effects in regression models. *Biometrika* **2013**, *100*, 1005–1010. [CrossRef]
44. DiCiccio, C.J.; Romano, J.P. Robust Permutation Tests For Correlation And Regression Coefficients. *J. Am. Stat. Assoc.* **2017**, *112*, 1211–1220. [CrossRef]
45. Hoerl, A.; Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
46. White, H. A Heteroscedasticity-Consistent Covariance Matrix and a Direct Test for Heteroscedasticity. *Econometrica* **1980**, *48*, 817–838. [CrossRef]
47. Henderson, C.R. Estimation of genetic parameters (abstract). *Ann. Math. Stat.* **1950**, *21*, 309–310.
48. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **1975**, *31*, 423–447. [CrossRef]
49. Robinson, G.K. That BLUP is a Good Thing: The Estimation of Random Effects. *Stat. Sci.* **1991**, *6*, 15–32. [CrossRef]
50. Helwig, N.E. Robust nonparametric tests of general linear model coefficients: A comparison of permutation methods and test statistics. *NeuroImage* **2019**, *201*, 116030. [CrossRef] [PubMed]
51. Helwig, N.E. Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging. *WIREs Comput. Stat.* **2019**, *2*, e1457. [CrossRef]
52. Draper, N.R.; Stoneman, D.M. Testing for the Inclusion of Variables in Linear Regression by a Randomisation Technique. *Technometrics* **1966**, *8*, 695–699. [CrossRef]

53. O'Gorman, T.W. The Performance of Randomization Tests that Use Permutations of Independent Variables. *Commun. Stat. Simul. Comput.* **2005**, *34*, 895–908. [CrossRef]
54. Nichols, T.E.; Ridgway, G.R.; Webster, M.G.; Smith, S.M. GLM permutation: nonparametric inference for arbitrary general linear models. *NeuroImage* **2008**, *41*, S72.
55. Manly, B. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. *Res. Popul. Ecol.* **1986**, *28*, 201–218. [CrossRef]
56. Freedman, D.; Lane, D. A Nonstochastic Interpretation of Reported Significance Levels. *J. Bus. Econ. Stat.* **1983**, *1*, 292–298. [CrossRef]
57. ter Braak, C.J.F. Permutation Versus Bootstrap Significance Tests in Multiple Regression and ANOVA. In *Bootstrapping and Related Techniques. Lecture Notes in Economics and Mathematical Systems,* ; Jöckel, K.H., Rothe, G., Sendler, W., Eds.; Springer: Berlin/Heidelberg, Germany, 1992; Volume 376, pp. 79–86.
58. Still, A.W.; White, A.P. The approximate randomization test as an alternative to the *F* test in analysis of variance. *Br. J. Math. Stat. Psychol.* **1981**, *34*, 243–252. [CrossRef]
59. Kennedy, P.E.; Cade, B.S. Randomization tests for multiple regression. *Commun. Stat. Simul. Comput.* **1996**, *25*, 923–936. [CrossRef]
60. Huh, M.H.; Jhun, M. Random Permutation Testing in Multiple Linear Regression. *Commun. Stat. Theory Methods* **2001**, *30*, 2023–2032. [CrossRef]
61. Schur, J. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *J. FüR Die Reine Und Angew. Math.* **1917**, *1917*, 205–232. [CrossRef]
62. Hotelling, H. Further Points on Matrix Calculation and Simultaneous Equations. *Ann. Math. Stat.* **1943**, *14*, 440–441. [CrossRef]
63. Hotelling, H. Some New Methods in Matrix Calculation. *Ann. Math. Stat.* **1943**, *14*, 1–34. [CrossRef]
64. Duncan, W.J. Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices). *Lond. Edinb. Dublin Philos. Mag. J. Sci. Seventh Ser.* **1944**, *35*, 660–670. [CrossRef]
65. Helwig, N.E. *npreg: Nonparametric Regression via Smoothing Splines*; R Package Version 1.0-9; R Foundation for Statistical Computing, Vienna, Austria, 2022. https://cran.r-project.org/package=npreg
66. Helwig, N.E. *nptest: Nonparametric Tests*; R Package Version 1.0-3; R Foundation for Statistical Computing, Vienna, Austria, 2021. https://cran.r-project.org/package=nptest
67. Wood, S.N. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation and GAMMs by REML/PQL*; R Package Version 1.8-40; R Foundation for Statistical Computing, Vienna, Austria, 2022. https://cran.r-project.org/package=mgcv
68. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
69. 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]
70. Henderson, H.V.; Searle, S.R. On deriving the inverse of a sum of matrices. *SIAM Rev.* **1981**, *23*, 53–60. [CrossRef]