A smoothing-based goodness-of-fit test of covariance for functional data

Stephanie T. Chen | Luo Xiao | Ana-Maria Staicu

Department of Statistics, North Carolina State University, Raleigh, North Carolina

Correspondence
Stephanie T. Chen, Department of Statistics, North Carolina State University, Raleigh, NC
Email: stchen3@ncsu.edu

1 INTRODUCTION

Functional data have become increasingly common in fields such as medicine, agriculture, and economics. Functional data usually consist of high frequency observations collected at regular intervals, see Ramsay and Silverman (2002, 2005) for an overview of methods and applications. By comparison, longitudinal data typically consist of repeated observations collected at a few time points varying across subjects. In recent years, functional data methods have been successfully extended and applied to longitudinal data (James et al., 2000; Yao et al., 2005). While these methods are more flexible, their estimation and interpretation are more cumbersome than longitudinal methods and require more sampling units or observations for accurate and reliable estimates. Thus, it is natural to question if such flexibility is truly necessary. This article focuses on comparing longitudinal data methods with functional data methods. For example, we consider the case of testing if a simple linear mixed effects model is sufficient for longitudinal data or if a more complex functional data model is required.

This work is motivated by the CD4 cell count dataset from the Multicenter AIDS Cohort Study (Kaslow et al., 1987). CD4 count is a key indicator for AIDS disease progression, and understanding its behavior over time is critical for monitoring HIV+ patients. The dataset is highly sparse, with 5 to 11 irregularly-spaced observations per subject. CD4 counts have been extensively analyzed using longitudinal data methods, e.g., semiparametric and linear random effects models (Taylor et al., 1994; Zeger and Diggle, 1994; Fan and Zhang,
Recently, functional data methods have also been applied to this data (Yao et al., 2005; Goldsmith et al., 2013; Xiao et al., 2018). While the nonparametric functional data methods are highly flexible and better adapt to subject-specific patterns, they are more difficult to implement and interpret compared to the parametric approaches. Therefore it is of interest to test whether the simpler longitudinal methods are sufficient for the data. To the authors’ best knowledge, no formal testing procedure exists for this application.

The inherent difference between functional and traditional longitudinal data methods is in the correlation model between repeated observations. For functional methods, the covariance within a subject is assumed to be smooth with an unknown nonparametric form. The covariance can be estimated by smoothing the sample covariance (Besse and Ramsay, 1986; Yao et al., 2005; Xiao et al., 2018) or constructing a reduced rank approximation by estimating basis functions from smoothed sample curves (James et al., 2000; Peng and Paul, 2009). In contrast, longitudinal data approaches typically assume a simple parametric covariance structure with a few parameters, such as autoregressive or exponential (see Diggle et al. (2002) for an overview), or induced by a random effects model (Laird and Ware, 1982).

Existing work on testing parametric versus nonparametric functions is limited to density and regression functions for the response variable, but has been extended to settings such as semiparametric and functional models; see González-Manteiga and Crujeiras (2013) for a recent review. Hardle and Mammen (1993) propose a smoothing-based goodness-of-fit statistic for regression functions, derive the asymptotic normal distribution, and develop a “wild” bootstrap algorithm for finite samples. Comparisons have also been applied to functional regression for model diagnostics and evaluating assumptions (Chiu and Muller, 2007; Bucher et al., 2011) and testing functional coefficients (Swihart et al., 2014; McLean et al., 2015; Kong et al., 2016). The proposed method is an extension of smoothing-based methods to test the form of the covariance function.

For high-dimensional multivariate data, where observation points are regular and balanced (same for all subjects), a number of methods exist to test an identity or spherical covariance matrix against an unstructured alternative (Ledoit and Ramsay, 1986; Yao et al., 2005; Xiao et al., 2018) or constructing a reduced rank approximation by estimating basis functions from smoothed sample curves (James et al., 2000; Peng and Paul, 2009). In contrast, longitudinal data approaches typically assume a simple parametric covariance structure with a few parameters, such as autoregressive or exponential (see Diggle et al. (2002) for an overview), or induced by a random effects model (Laird and Ware, 1982).

The remainder of this article is organized as follows. Section 2 presents the statistical model and hypothesis test, Section 3 details the proposed test, and Section 4 describes our implementation. Section 5 outlines extensions to general smooth covariance functions. Section 6 presents a simulation study. Section 7 details three applications to diffusion tensor imaging, child growth, and CD4 cell count. Finally, Section 8 summarizes the article and discusses limitations of the proposed test.

2 | STATISTICAL FRAMEWORK

Consider functional or longitudinal data \((t_{ij}, Y_{ij}) \in \mathcal{T} \times \mathbb{R} : i = 1, \ldots, n, j = 1, \ldots, m_i\) where \(i\) denotes the subject index, \(j\) denotes the visit index, and \(Y_{ij}\) is the measurement for the \(i\)-th subject at time \(t_{ij}\). Here, \(n\) is the number of subjects and \(m_i\) the number of observations for the \(i\)-th subject, which can vary across subjects. Assume that \(\mathcal{T} = [a, b]\) is a closed and compact domain. Data are often observed with noise, so we posit the model

\[
Y_{ij} = \mu(t_{ij}) + X_j(t_{ij}) + \epsilon_{ij}.
\]  

(1)

Here \(\mu(t)\) is a smooth mean function, \(X_j\) is a zero-mean Gaussian random function independent between subjects, and \(\epsilon_{ij}\) is Gaussian white noise independently and identically distributed with zero mean and variance \(\sigma^2\), independent of \(X_j\).
Let \( G(t, t') = \text{Cov}[X_i(t), X_i(t')] \) be the covariance function of \( X_i \). Assume that \( G \) is a smooth, positive semidefinite bivariate function defined on \( T^2 \).

We are interested in the form of the covariance, and would like to test the hypothesis that \( G \) has a known parametric form against a general alternative. Motivated by the CD4 dataset, which has previously been fit with a linear random intercept and slope model, we focus on the quadratic polynomial function

\[
G_0(t, t') = \sigma_0^2 + \sigma_0 (t + t') + \sigma_1^2 t^2,
\]

where \((\sigma_0^2, \sigma_0, \sigma_1^2)\) are unknown parameters. Because this covariance is induced by the linear random effects model \( X_i(t) = b_{i0} + b_{1i} t \), where \( b_i = (b_{i0}, b_{1i})^T \) are random effects with zero mean and \( \text{Var}(b_{i0}) = \sigma_0^2 \), \( \text{Var}(b_{1i}) = \sigma_1^2 \), and \( \text{Cov}(b_{0i}, b_{1i}) = \sigma_0 \), testing \( G_0 \) is equivalent to testing if a linear random (or mixed) effects model is sufficient for the data. Note that this is a specific case of the general linear random effects model \( X_i(t) = \sum_{k=1}^K b_{ik} \phi_k(t) \) for random effects \( b_{ik} \) with zero mean and variance \( \sigma_k^2 \) and known functions \( \phi_k(t) \), which has covariance function \( G_0(t, t') = \sum_{k=1}^K \sigma_k^2 \phi_k(t) \phi_k(t') + 2 \sum_{k<k'} \sigma_{kk'} \phi_k(t) \phi_{k'}(t') \), where \( \sigma_{kk'} = \text{Cov}(b_{ik}, b_{i'k'}) \). While we focus on (2), the proposed test can be easily adapted for the more general random effects case or any smooth parametric covariance with finite parameters, as discussed in Section 5. Ideally, scientific or expert knowledge about the underlying process should guide the choice of \( G_0 \). If such information is unavailable, a commonly used and interpretable structure would be preferred.

Formally, the hypothesis test can be written as

\[
H_0 : G(t, t') = G_0(t, t') \quad \text{versus} \quad H_A : G(t, t') \neq G_0(t, t').
\]

Under the null hypothesis, the covariance has a specific parametric form with finite parameters. Under the alternative hypothesis, the covariance function is assumed to be smooth and positive semidefinite. This flexibility may better capture heterogeneity across subjects but is hard to estimate and interpret compared to a parametric model. Therefore, it is desirable to test goodness-of-fit for these two types of models. In the following section, we propose a distance-based goodness-of-fit test for (3) that can be applied to functional data with either a dense common or sparse irregular sampling design.

### 3 SMOOTHING-BASED TEST

We propose a test statistic based on the distance between the covariance functions estimated under the null and alternative hypotheses, respectively. In the remainder of this section, we describe covariance estimation under the null and alternative hypotheses, and then introduce our test statistic. The smooth mean \( \mu(t) \) can be estimated non-parametrically with spline smoothing (Ruppert et al., 2003; Wood, 2003), allowing us to consider only the de-meaned data \( \bar{Y}_{ij} = Y_{ij} - \bar{\mu}(t_{ij}) \) for modeling \( X_i(t_{ij}) + \epsilon_{ij} \). See Section 4 for details.

#### 3.1 Null model

Under the null hypothesis, \( G = G_0 \) is a quadratic polynomial covariance, corresponding to

\[
X_i(t) = b_{i0} + b_{1i} t,
\]

\[
(b_{0i}, b_{1i})^T \sim N \left( 0, \mathbf{V}_0 = \begin{bmatrix} \sigma_0^2 & \sigma_0 \sigma_1 \\ \sigma_0 & \sigma_1^2 \end{bmatrix} \right).
\]

Here, \( X_i(t) \) is a linear random effects model with subject-specific random intercepts and slopes, \( b_{0i} \) and \( b_{1i} \), respectively. Let \( \bar{Y} \) be the \( mn \)-length vector of de-meaned observations for the \( i \)-th subject observed at times \( t_i = (t_{i1}, \ldots, t_{im})^T \), and \( \mathbf{V}_i = [1, t_i]^T \mathbf{V}_0 [1, t_i]^T + \sigma^2 \mathbf{I}_m \) be the corresponding covariance matrix, where \( \mathbf{I}_m \) is a \( m \times m \) identity matrix. Then the unknown parameters in model (4) can be estimated by maximizing the log-likelihood \( \ell'(\mathbf{V}_0, \sigma^2; \bar{Y}, t_i) = \sum_{i=1}^n -\frac{1}{2} \log |\mathbf{V}_i| + \bar{Y}^T \mathbf{V}_i^{-1} \bar{Y} \), where \( |\mathbf{V}_i| \) is the determinant of the matrix \( \mathbf{V}_i \), using an expectation-maximization (EM) or Newton–Raphson algorithm, as outlined in Lindstrom and Bates (1988).

#### 3.2 Alternative model

Under the alternative hypothesis, the covariance function has a smooth, nonparametric form. Approximate \( G_A \) by smoothing the sample covariance using tensor product regression splines as \( G(t, t') = \sum_{h=1}^H \theta_{h0} \mathbf{B}_h(t) \mathbf{B}_h(t') \), where \( \mathbf{B}_h(t) : h = 1, 2, \ldots, H \) are a sequence of cubic B-spline basis functions defined over \( T \) and \( \theta_{h0} \) are coefficients estimated by minimizing the least squares expression

\[
\sum_{i=1}^n \sum_{1 \leq j \neq j' \leq m_i} \left( \hat{\bar{Y}}_{ij} \hat{\bar{Y}}_{ij'} - \sum_{h=1}^H \theta_{h0} \mathbf{B}_h(t_{ij}) \mathbf{B}_h(t_{ij'}) \right)^2,
\]

under the natural symmetry constraint that \( \theta_{h0} = \theta_{h0} \). Denote the estimated alternative covariance as \( \hat{G}_A(t, t') = \sum_{h=1}^H \hat{\theta}_{h0} \mathbf{B}_h(t) \mathbf{B}_h(t') \).

The measurement error, \( \sigma^2 \), in equation (1) can be estimated following Yao et al. (2005) and Goldsmith et al. (2013) by averaging the distance between the diagonals of the raw sample covariance, i.e., \( \bar{\bar{Y}}_{ij}^2 \) for \( 1 \leq j \leq m_i, 1 \leq i \leq n \), and \( \hat{G}_A \). To mitigate boundary effects, only the middle 50% of \( T \) is considered (Staniswalis and Lee, 1998; Yao et al., 2005).
3.3 Test statistic

Using the estimated null and alternative covariances, \( \hat{G}_0 \) and \( \hat{G}_A \), the proposed test statistic is the Hilbert–Schmidt norm distance

\[
T_n = ||\hat{G}_A - \mathcal{K}\hat{G}_0||_{HS},
\]

where \( ||f||_{HS} = \sqrt{\int \int f(t,t')^2 dt dt'} \) for bivariate function \( f \) and \( \mathcal{K}\hat{G}_0 \) is the smoothed null covariance estimate using tensor-product B-splines. That is, replace \( \hat{g}_{ij} \) with \( \hat{g}_{0,i,j} \) in the least squares expression (5) to estimate \( \theta_{0,i,j} \) so \( \mathcal{K}\hat{G}_0(t,t') = \sum_{h,f=1}^{H} \hat{\theta}_{0,h} B_h(t) B_f(t') \). Using the smoothed null eliminates the bias from nonparametric function estimation and is common practice for nonparametric regression tests; see, e.g., Hardle and Mammen (1993). A large \( T_n \) indicates that the null parametric covariance approximates the true covariance poorly. The null distribution of \( T_n \) is difficult to derive as estimation of the alternative is based on second moments of the observed responses. Moreover, even in settings where the null distribution of distance-based test statistic is available, Hardle and Mammen (1993) show that the test statistic converges slowly and recommends bootstrapping instead. In the next section, we propose a wild bootstrap algorithm (Wu, 1996) for the null distribution of \( T_n \) following Hardle and Mammen (1993). Note that one may also consider an empirical version of the proposed test statistic evaluated at the paired time points (see Web Appendix A for an example); we focus on (6) throughout this article.

3.4 Approximate null distribution of \( T_n \) via a wild bootstrap

Denote the \( l \)-th bootstrap sample as \( \{Y_{ij}^{(l)} : i = 1, \ldots, n, j = 1, \ldots, m, t_{ij} \in T \} \), where \( Y_{ij}^{(l)} = \tilde{\mu}(t_{ij}) + X_i^{(l)}(t_{ij}) + \epsilon_{ij}^{(l)} \) for the original time points \( t_{ij} \). Let \( \tilde{\mu}(t) \) be the estimated smooth mean function, \( X_i^{(l)}(t_{ij}) \) be subject trajectories generated from the estimated null model in (4), and \( \epsilon_{ij}^{(l)} \) be simulated residuals using the estimated measurement error in Section 3.2. The test statistic, \( T_n^{(l)} \), can be calculated from the resulting bootstrap sample, and the process is repeated to obtain an approximation of the null distribution of \( T_n \). If the observed statistic is large compared to the null approximation, then reject \( H_0 \). This “wild” bootstrap procedure (Wu, 1986) is outlined in Algorithm 1 and is valid in the regression function setting (Hardle and Mammen, 1993).

4 IMPLEMENTATION

First, estimate the smooth mean \( \mu(t) \) using thin plate regression splines (Wood, 2003) using the gam function in the R package mgcv (Wood, 2017), and subtract from the data. The null model in (4) is a standard random effects model that can be estimated using the lme function in the R package nlme (Pinheiro et al., 2017). For the least squares expression in (5) to smooth the alternative and null covariance estimates, we use \( H = 10 \) cubic B-splines per axis with equally-spaced interior knots. The choice of 10 B-splines balances performance and computational speed, see Web Appendix B for a sensitivity study. While the number of splines needs only be sufficiently large, additional splines may be needed if the data is known or observed to be highly wiggly. Cross-validation or Akaike information criterion (AIC) may be used for a formal selection (see Wood (2003) for discussion).

5 EXTENSIONS

5.1 Smooth covariance

Any smooth parametric covariance function can be tested using the proposed procedure, with modification to the null model and bootstrap algorithm. For example, consider the stationary Gaussian or quadratic exponential covariance function \( G_0(t, t') = \theta e^{-|t-t'|/\delta} \), where \( \delta = |t - t'| \), and \( (\theta, \delta) \) are parameters to be estimated. The null model can be estimated using likelihood-based methods, and bootstrap data generated as \( Y_{ij}^{(l)} = \tilde{\mu}(t_{ij}) + \tilde{\nu}_0^{(l)} z + \epsilon_{ij}^{(l)} \), where \( \tilde{\nu}_0 \) is the estimated mean vector of length \( m \), \( \tilde{\nu}_0^{(l)} \) is the estimated null covariance matrix defined by \( (\tilde{\theta}, \tilde{\delta}) \), and \( X^\frac{1}{2} \) is the square root matrix of the estimated null covariance.
where $X^2 \mathbf{z} = X$. \( z \) is an \( m_i \)-length vector of independent samples from a standard normal distribution, and \( \varepsilon^{(i)}_j \) is an independent vector of residuals from \( N(0, \sigma^2) \).

## 6 | SIMULATION STUDY

We conduct a simulation study to evaluate performance of the proposed bootstrap test and two competing methods, described in Section 6.1, for testing the hypothesis in (3) that the covariance has a quadratic polynomial form. Data are generated as

\[
Y_{ij} = \mu(t_{ij}) + X_i(t_{ij}) + \varepsilon_{ij} \\
X_i(t_{ij}) = b_0 + b_1 t_{ij} + \Delta z_i(t_{ij}),
\]

(7)

for \( i = 1, \ldots, n \) subjects and \( j = 1, \ldots, m_i \) observations per subject. The scalar, \( \Delta \), controls the magnitude of deviation from the null model. The mean, \( \mu(t) \), is set to 0 and the residuals are distributed \( \varepsilon_{ij} \sim N(0, 1) \), independent of \( X_i \). Random intercepts and slopes are sampled from a bivariate normal distribution with zero mean, \( \text{Var}(b) = \text{Var}(b_{ij}) = 1 \) and \( \text{Cov}(b_{ij}, b_{ij}) = -0.5 \), independent of the non-linear function \( z_i \), defined below. The \( t_{ij} \) are observed on a grid of 80 equally spaced points in \([-1, 1]\). If \( m_i = 80 \), the subject is observed at all points and if \( m_i < 80 \), observed time points are uniformly sampled for each subject from the 80 possible points. Tuning parameters are selected as described in Section 4. Consider a factorial combination of the following factors:

1. **Observations per subject** \( (m_i = m) \): (a) \( m = 80 \), (b) \( m = 40 \), (c) \( m = 20 \), (d) \( m = 10 \)
2. **Deviation from the null model**:
   (a) **Quadratic**: \( z_i(t) = b_0 t_i^2, z_i(t) \sim N(0, 1) \)
   (b) **Trigonometric**: \( z_i(t) = \sum_{k=1}^{2} \xi_{ik} \psi_k(t) \),
   \( \{ \psi_1(t), \psi_2(t) \} = \{ \sin(2\pi t), \sin(4\pi t) \} \),
   \( \xi_{ik} \sim N(0, \lambda_k) \), \( \lambda_1 = \lambda_2 = 1 \).

For each factor combination, we use \( L = 1000 \) bootstrap samples per dataset and consider \( n = 100 \) and 500 subjects, and \( n = 50 \) for the \( m = 80 \) setting only. Performance is evaluated in terms of the empirical type I error rate (size) for nominal levels \( \alpha = 0.05 \) and 0.10 based on 5000 simulated datasets, and power at the \( \alpha = 0.05 \) level with 1000 simulated datasets. Results are presented in terms of deviation from the null, defined as \( \Delta^2 \int \text{Var}(z_i(t))/\text{Var}(X_i(t)) dt \).

### 6.1 | Competing methods

As discussed in Section 1, we are unaware of any existing methods for testing covariance that can be applied to all functional or longitudinal data settings. In this subsection, we describe two testing methods that can be applied to specific scenarios of the hypothesis test in (3).

#### 6.1.1 | Direct test

Consider the case where covariance under the alternative hypothesis has a known, parametric form so the null model for \( X_i \) is nested within the alternative model. In essence, test if a more complex covariance better explains the data than the null covariance. For the quadratic polynomial covariance, an alternative may be \( G_A(t, t') = \sigma_0^2 + \sigma_{01}(t + t') + \sigma_1^2 t^2 + \sigma_2^2 t'^2 \).

Then the alternative model can be written as

\[
X_i(t_{ij}) = b_0 + b_1 t_{ij} + b_2 t_{ij}^2
\]

\[
b_i = (b_{0i}, b_{1i}, b_{2i})^T \sim N \left( \begin{bmatrix} \sigma_0^2 & \sigma_{01} & 0 \\ \sigma_{01} & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix} \right),
\]

(8)

Note that this is the model for the quadratic deviation setting in the simulation study. Like the null model, (8) can be estimated using the lme function in the R package nlme (Pinheiro et al., 2017). The hypothesis test is equivalent to testing if \( b_{2i} = 0 \), or \( H_0 : G(t, t') = G_0(t, t') \iff \sigma_2^2 = 0 \) versus \( H_A : G(t, t') \neq G_0(t, t') \iff \sigma_2^2 > 0 \).

Testing zero-value variance components is a non-standard problem because the null hypothesis is on the boundary of the parameter space. Self and Liang (1987) derive the asymptotic null distribution of the likelihood ratio test (LRT) for this setting as a mixture of chi-squared distributions. Crainiceanu and Ruppert (2004) derive the exact finite sample null distribution for the (R)LRT of mixed models with one random effect, and Greven et al. (2008) extend this approach to models with multiple random effects using pseudolikelihood. Because of the limited sample size in our simulation study, we use the finite sample null distribution from Greven et al. (2008), which can be performed efficiently using the exactRLRT function in the R package RLRsim (Schiepl and Bolker, 2016).

#### 6.1.2 | Multivariate test

The Zhong et al. (2017) test for high-dimensional multivariate data can be applied to functional data with a common design. Consider a repeated measures model \( Y_i = \mu + e_i \), where \( Y_i = (Y_{i1}, \ldots, Y_{im})^T \) is a vector of responses, \( \mu \) is a mean vector of length \( m \), and residuals are distributed \( e_i \sim N(0, G) \).

Denote \( \theta_0 \) as the parameter vector defining the covariance matrix under the null hypothesis, \( G_0 \). Let \( G_A \) be the alternative unstructured covariance.

Based on the squared-Frobenius distance between the null and alternative covariances, \( \delta(\theta_0) = \text{tr}(G_A - G_0)^2 \), Zhong et al. (2017) propose the test statistic \( J_n = J_n - J_3 \), where \( J_n \) is an unbiased estimator for \( \delta(\theta_0) \) and \( J_3 \) adjusts for errors in the estimation of \( \theta_0 \).
can be conducted by testing if $\Lambda_\omega$ is significantly larger than 0. With some assumptions on the covariance structure, the asymptotic normal and fixed sample weighted-chi square null distributions can be determined for any parametric covariance, and we provide derivations for the quadratic polynomial covariance in Web Appendix C. In our simulation study, the multivariate test can only be applied to the dense $m = 80$ case, and we use 10,000 samples to approximate the fixed sample null distribution. In Web Appendix D, we also consider performance of the multivariate test in less-ideal settings with small $m$ and unequally-spaced data.

6.1.3 | Limitations of the competing methods

While both competitors utilize test statistics with known null distributions, these tests only apply to limited scenarios. The direct test applies when the alternative is known, parametric, and a superset of the null model. The multivariate test only applies to data with a common design and assumes an unstructured covariance that does not account for smoothness. Thus, the bootstrap test is expected to be more powerful than the multivariate test for testing functional data.

6.2 | Simulation results

Table 1 reports the empirical type I error rates for all three methods. As the multivariate test requires a common sampling design, it can only be applied to the $m = 80$ setting. We report only the fixed-sample weighted chi-squared distribution; results for the asymptotic normal distribution were similar and are presented in Web Appendix D. All three methods have empirical levels close to the nominal levels, although both the bootstrap and direct tests can be slightly conservative for several settings.

Table 1 presents simulation results for the quadratic and trigonometric deviations from the null, by number of observations per subject, $m$. For all methods, power increases with sample size, particularly as data are more densely sampled and the covariance is better estimated. As expected, power depends on how closely the true model matches the specific alternative assumed by the test. The bootstrap test outperforms the multivariate test for all settings because of the more specific form of its alternative, and has higher power than the direct test when the direct test has misspecified the alternative (trigonometric deviation). Conversely, the direct test has higher power when the parametric alternative is correctly specified (quadratic deviation). Both the bootstrap and multivariate tests are better able to detect the trigonometric deviation because the covariance more obviously deviates from the null model. Overall, the bootstrap test performs well in most settings, except when the dataset is small and deviation from the null is small. For example, when $n = 100$ and $m = 10$, the test is underpowered for the quadratic deviation when signal size is small.

In terms of computational speed, the bootstrap test is, unsurprisingly, significantly slower than the competitor methods. A personal laptop with a 2.9 GHz processor took 1–7 min to run a single iteration, compared to 1 and 0.2 s for the direct and multivariate tests, respectively, for the null model with 100 subjects. Reducing the density of inputted data or $L$ number of bootstrap samples can decrease computational time, with some loss of power.

7 | APPLICATIONS

7.1 | Diffusion tensor imaging

We first consider a dataset of diffusion tensor imaging (DTI) of intracranial white matter microstructure with dense, common sampling design for a group of normal and multiple sclerosis patients. Images of the white matter are depicted with tract profiles shown in Figure 2 and available in the R package refund (Goldsmith et al., 2016); see Reich et al. (2010) for study details. Goldsmith et al. (2011) consider this dataset for modeling multiple sclerosis disease status, concluding that inclusion of the tract profile as a functional

| TABLE 1 | Estimated type I error rates for the bootstrap, direct, and multivariate tests at the nominal $\alpha = 0.05$ and $0.10$ levels based on 5000 datasets, by number of subjects ($n$) and observations per subject ($m$). The standard error was 0.003 and 0.004 for $\alpha = 0.05$ and $\alpha = 0.10$, respectively. The multivariate test is applicable for only the dense $m = 80$ setting. |
|---|---|---|---|---|---|---|---|
| $n$ | $m$ | $\alpha = 0.05$ | $\alpha = 0.10$ | $\alpha = 0.05$ | $\alpha = 0.10$ | $\alpha = 0.05$ | $\alpha = 0.10$ |
| 100 | 10 | 0.059 | 0.126 | 0.044 | 0.086 | n/a | n/a |
| 20 | 0.045 | 0.105 | 0.049 | 0.098 | n/a | n/a |
| 40 | 0.042 | 0.093 | 0.049 | 0.102 | n/a | n/a |
| 80 | 0.042 | 0.091 | 0.045 | 0.096 | 0.053 | 0.103 |
| 500 | 10 | 0.047 | 0.105 | 0.049 | 0.096 | n/a | n/a |
| 20 | 0.050 | 0.103 | 0.049 | 0.096 | n/a | n/a |
| 40 | 0.050 | 0.100 | 0.043 | 0.090 | n/a | n/a |
| 80 | 0.044 | 0.093 | 0.046 | 0.096 | 0.053 | 0.105 |
predator improves model performance compared to a subject-specific average of the profile. Note that a subject-specific average is equivalent to the subject-specific intercept in the null model in (4). We evaluate this conclusion formally by testing if a quadratic polynomial covariance is sufficient for modeling the tract profiles, using the bootstrap, multivariate, and direct tests. We focus on the baseline tract profiles of the corpus callosum (CCA), associated with cognitive function, for multiple sclerosis patients, observed on a dense, regular grid of 93 points. After removing subjects with missing observations, the dataset has profiles from 99 subjects, for a total of 9207 observations. Tuning parameters are selected as described in Section 4. The observed test statistic for the bootstrap test is $T_n = 0.071$ corresponding to $p < 0.001$. The direct test yields an RLRT statistic of 1160.6 corresponding to $p < 1 \times 10^{-16}$. The multivariate test yields an observed test statistic of $\Lambda_n = 0.058$, corresponding to $p < 0.001$ for both the weighted chi-squared and asymptotic normal distributions. All three tests support the conclusion that a quadratic polynomial covariance is inadequate for the data, and that a functional method should be used.

7.2 | Child growth measurements

Next, consider the CONTENTS child growth dataset from Lima, Peru (Xiao et al., 2018). The dataset contains irregularly sampled height measurements for 215 children covering 0 to 729 days after birth, for a total of 8839 observations (20–50 observations per subject, observed at different time points), shown in Figure 2. Subject trajectories predicted using functional principal components analysis, shown in Xiao et al. (2018), exhibit curvature not captured by a linear parametric model, suggesting that a functional approach is necessary for the data. We consider this observation formally by testing the quadratic polynomial covariance for the growth data using the bootstrap and direct tests. The observed test statistic for the bootstrap test is $T_n = 494.13$, corresponding to $p = 0.031$, while the RLRT statistic from the direct test is 2205.8, corresponding to $p < 0.001$. Both tests indicate that the parametric quadratic polynomial covariance is not sufficient for the data, and a functional approach should be used instead.
CD4 count data

Last, we consider the motivating example of CD4 cell counts described in Section 1 by conducting a formal test of the quadratic polynomial covariance using the bootstrap and direct tests. The dataset is available in the R package refund (Goldsmith et al., 2016) and includes cell counts from −18 to 52 months since seroconversion; we log-transform the counts to stabilize variability. We consider only subjects with at least 5 observations and who have log-transformed cell counts greater than 4, for a total of 1402 observations from 208 subjects (5–11 observations per subject). The cleaned and log-transformed data are shown in Figure 2.

Because data are sparser than the settings considered in the full study, we conduct a small simulation study to check the size and power of the tests. Simulated data are generated as \( Y_{ij} = X_i(t_{ij}) + \epsilon_{ij} \), where \( X_i(t_{ij}) \) is defined below, \( t_{ij} \) are the time points in the original dataset, and \( \epsilon_{ij} \sim N(0, \hat{\sigma}^2) \), where \( \hat{\sigma}^2 \) is the estimated error variance under the alternative model. The random function \( X_i(t) \) is generated from a multivariate normal distribution with zero-mean and covariance \( \Sigma = (1 - \delta)\hat{\Sigma}_0 + \delta\hat{\Sigma}_A + r\hat{G}_1 \), where \( \hat{\Sigma}_0 \) and \( \hat{\Sigma}_A \) are the estimated covariance matrices from the null and alternative model, respectively, \( \delta \in [0, 1] \) controls the contribution of the null and alternative covariances, and \( \hat{G}_1 \) is the matrix generated from the first three eigenfunctions and eigenvalues of \( (\hat{\Sigma}_A - \hat{\Sigma}_0) \), with magnitude controlled by \( r \geq 0 \). Note that when \( \delta = r = 0 \), \( \Sigma \) is the null covariance, and when \( \delta = 1 \) and \( r = 0 \), \( \Sigma \) is the alternative covariance. To show how power changes with deviation from the null model, let \( \delta = 1 \) when \( r > 0 \). Since the bootstrap test is likely to be underpowered due to sparsity of the data, we also simulate data with double the number of subjects or double the observations per subject. Additional subjects were generated using the same set of observed time points, while additional observations were added by uniformly sampling from the non-observed time points for each subject.

### Table 2

|                | Bootstrap | Direct |
|----------------|-----------|--------|
| \( \alpha = 0.05 \) | \( \alpha = 0.10 \) | \( \alpha = 0.05 \) | \( \alpha = 0.10 \) |
| Standard dataset | 0.057 | 0.113 | 0.046 | 0.010 |
| Double subjects  | 0.056 | 0.108 | 0.047 | 0.095 |
| Double observations | 0.046 | 0.101 | 0.050 | 0.100 |
Table 2 gives the empirical type I error rates based on 5000 simulated datasets, and Figure 3 shows the power from 1000 datasets, for the bootstrap and direct tests. The bootstrap test is underpowered for the true CD4 dataset due to small sample size, and doubling the number of observations per subject resolves this problem.

The observed test statistic for the bootstrap test is \( T_n = 5.025 \) corresponding to \( p = 0.100 \), while the direct test yields an RLRT statistic of 2.704 corresponding to \( p = 0.0428 \). While only the direct test indicates that the quadratic polynomial covariance is not sufficient for the data, Figure 3 shows that the bootstrap test is underpowered, suggesting that a more complex covariance may still be necessary for the data.

8 | CONCLUDING REMARKS

In this article, we propose a smoothing-based goodness-of-fit test of covariance for functional data. We focus on the specific case of testing a quadratic polynomial covariance induced by a linear random intercept and slope model, as motivated by a dataset of CD4 cell counts used to monitor HIV+ patients. Our proposed method can be used to formally test a linear random (or mixed) effects model against a typical functional data approach, and fills a gap in the testing of longitudinal and functional data methods. The proposed bootstrap test can be applied to functional data with either dense common or irregular sampling design, and performs well in simulation studies. Limitations of the method are (a) slow computational speed, and (b) low power for very small datasets with small deviation from the null.

ACKNOWLEDGEMENTS

The authors thank an associate editor and two reviewers for their comments that greatly improved this article, and Dr. Ping-Shou Zhong for providing code for the multivariate test. ST Chen and L Xiao’s research were supported by Grant OPP1148351 and OPP1114097 from the Bill and Melinda Gates Foundation. AM Staicu’s research was supported by NSF grant DMS 1454942 and NIH grants 5P01 CA142538-09 and 2R01MH086633. This work represents the opinions of the researchers and not necessarily that of the granting organizations.

ORCID

Stephanie T. Chen http://orcid.org/0000-0002-0581-4447
Luo Xiao http://orcid.org/0000-0001-8707-0914

REFERENCES

Bai, Z., Jiang, D., Yao, J., and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. Ann Stat 1, 135–141.

Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. Psychometrika 51, 285–311.

Bücher, A., Dette, H., and Wieczorek, G. (2011). Testing model assumptions in functional regression models. J Multivariate Anal 102, 1472–1488.

Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. Ann Stat 39, 2330–2355.

Chiou, J. M. and Müller, H. G. (2007). Diagnostics for functional regression via residual processes. Comput Stat Data Anal 51, 4849–4863.
Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J R Stat Soc* 66, 165–185.

Diggle, P., Haegerty, P., Liang, K. Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford, UK: Oxford University Press.

Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear models with application to longitudinal data. *J R Stat Soc Series B* 62, 303–322.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *Neuromimage* 57, 431–439.

Goldsmith, J., Greven, S., and Crainiceanu, C. M. (2013). Corrected confidence bands for functional data using principal components. *Biometrics* 69, 41–51.

Goldsmith, J., Schiepl, F., Huang, L., Wrobel, J., Gellar, J., Harezlack, J., et al. (2016). refund: Regression with Functional Data. R package version 0.1-16.

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review on goodness-of-fit tests for regression models. *Test* 22, 361–411.

Greven, S., Crainiceanu, C. M., Küchenhoff, and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *J Comput Graph Stat* 17, 870–891.

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann Stat* 21, 1926–1947.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* 87, 587–602.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. (1987). The multicenter AIDS cohort study: Rationale, organization, and selected characteristics of the participants. *Am J Epidemiol* 126, 310–318.

Kong, D., Staicu, A. M., and Maity, A. (2016). Classical testing in functional linear models. *J Nonparametr Stat* 28, 813–838.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann Stat* 30, 1081–1102.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J Am Stat Assoc* 83, 1014–1022.

McLean, M. W., Hooker, G., and Ruppert, D. (2015). Restricted likelihood ratio tests for linearity in scalar-on-function regression. *Stat Comput* 25, 997–1008.

Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of functional principal components from sparse longitudinal data. *J Comput Graph Stat* 18, 995–1015.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., EISPACK, Heisterkamp, S., et al. (2017). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. New York: Springer.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer.

Reich, D. S., Ozturk, A., Calabresi, P. A., and Mori, S. (2010). Automated vs. conventional tractography in multiple sclerosis: Variability and correlation with disability. *Neuroimage* 49, 3047–3056.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Schiepl, F. and Bolker, B. (2016). RLRsim: Exact (Restricted) Likelihood Ratio Tests for Mixed and Additive Models. R package version 3.1-131.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* 82, 605–610.

Swihart, B. J., Goldsmith, J., and Crainiceanu, C. M. (2014). Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics* 56, 483–493.

Taylor, J. M. G., Cumberland, W. G., and Sy, J. P. (1994). A stochastic model for analysis of longitudinal AIDS data. *J Am Stat Assoc* 89, 727–736.

Wood, S. (2003). Thin plate regression splines. *J R Stat Soc Series B* 65, 95–114.

Wood, S. (2017). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version 1.8-23.

Wu, C. J. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat* 14, 1261–1295.

Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. M. (2018). Fast covariance estimation for sparse functional data. *Stat Comput* 28, 511–522.

Yao, F., Müller, H. G., and Wang J. L. (2005). Functional data analysis for sparse longitudinal data. *J Am Stat Assoc* 100, 577–590.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* 50, 689–699.

Zhong, P. S., Lan, W., Song, P. X. K., and Tsai, C. L. (2017). Tests for covariance structures with high-dimensional repeated measurements. *Ann Stat* 45, 1185–1213.

**SUPPORTING INFORMATION**

Web Appendices referenced in Sections 3.3, 4, 6.1.2, 6.2, and 6.3 are available with this article at the Biometrics website on the Wiley Online Library. R code implementing the `bootstrap`, `direct`, and `multivariate` tests for the quadratic polynomial covariance is available with this article at the Biometrics website on the Wiley Online Library.

---

**How to cite this article:** Chen ST, Xiao L, Staicu A-M. A smoothing-based goodness-of-fit test of covariance for functional data. *Biometrics*. 2019;75:562–571. [https://doi.org/10.1111/biom.13005](https://doi.org/10.1111/biom.13005)