Multilevel Cross-dependent Binary Longitudinal Data

Nicoleta Serban[1]
H. Milton Stewart School of Industrial Systems and Engineering
Georgia Institute of Technology
nserban@isye.gatech.edu

Ana-Maria Staicu
Department of Statistics
North Carolina State University
ana-maria_staicu@ncsu.edu

Raymond J. Carroll
Department of Statistics
Texas A&M University
carroll@stat.tamu.edu

We provide insights into new methodology for the analysis of multilevel binary data observed longitudinally, when the repeated longitudinal measurements are correlated. The proposed model is logistic functional regression conditioned on three latent processes describing the within- and between-variability, and describing the cross-dependence of the repeated longitudinal measurements. We estimate the model components without employing mixed-effects modeling but assuming an approximation to the logistic link function. The primary objectives of this paper are to highlight the challenges in the estimation of the model components, to compare two approximations to the logistic regression function, linear and exponential, and to discuss their advantages and limitations. The linear approximation is computationally efficient whereas the exponential approximation applies for rare events functional data. Our methods are inspired by and applied to a scientific experiment on spectral backscatter from long range infrared light detection and ranging (LIDAR) data. The models are general and relevant to many new binary functional data sets, with or without dependence between repeated functional measurements.

Key Words: Binary longitudinal data; Covariogram estimation; Cross-dependent functional data; Functional data analysis; Hierarchical modeling; Mixed models; Multilevel functional data; Principal component estimation.

Short title: Multilevel Binary Longitudinal Data

[1]Correspondent Author

1 Introduction

In the functional data analysis (FDA) literature, it is commonly assumed that the data are a sample of random functions varying smoothly over their observation domain (Ramsay and Silverman, 2005). Limited methodological research has focused on binary-valued functional data. Most of the methods dealing with such data rely on generalized linear mixed models or generalized estimating equations (GEE). Hall et al. (2008) review the existing parametric models and introduce a functional data approach for studying the variability in binary longitudinal data, by means of conditioning on a Gaussian latent process. Their approach uses functional principal component analysis (FPCA) to model the functional latent process and to reconstruct the individual trajectories.

This article is concerned with modeling *cross-dependent binary functional data*. Specifically, for each subject r in the observed data sample, the subject-data are viewed as a set of binary-valued functional observations $\{Y_r(d_{ri}) : i = 1, \ldots, m_r\}$, where a functional observation is a realization of a random function $Y(d_{ri})$ observed at a fixed design point d_{ri}; we denote by $Y(d_{ri}, t)$ the value of the function $Y(d_{ri})$ at time point t or other form of functionality, for example, wavelength. The correlation between two functional observations is assumed to tail off as the distance between their associated design points increases. In this paper, we consider the case when the functional observations $Y_r(d_{ri})$ are observed on a common (moderately) dense grid. The case study presented in the paper is related to the analysis of spectral backscatter from long range infrared light detection and ranging (LIDAR) data, where the goal is to estimate the probability of high spectral backscatter at different wavelengths. Other examples include: (a) distribution of the presence or absence of birds species over time observed at multiple geographical locations, and for multiple species; and (b) product alarms detected and reported to a technical system observed for many products and for a number of systems over time.

Models for cross-dependent multilevel functional data have been proposed by Shi et al. (1996), Rice and Wu (2001), Morris et al. (2003), Baladandayuthapani et al. (2008) and Di et al. (2009) among others. Extending these models to binary-valued random functions is not straightforward because of the nonlinear relationship between the observed functional data and the model components that need to be estimated, and because of the correlation structure between the functional observations corresponding to the same subject.

We propose a nonparametric functional approach for the analysis of cross-dependent binary-valued functional data. Our methodology assumes the existence of an underlying latent bivariate Gaussian process, which can be modeled as in Staicu et al. (2010). Relevant features of the stochastic dependence of the observed data are reflected by the mean and covariance properties of this latent bivariate process. This approach was considered initially by Hall et al. (2008) for modeling binary-valued functional data. The primary contribution is that we suggest estimation methods that accommodate different scenarios for the prevalence of the events - for binary data indexed with 0's and 1's, an event is when we observe $Y_r(d_{ri}) = 1$. Specifically, in the case of non-rare events (relatively balanced number of 0's and 1's), our methods can be viewed as an extension of the estimation techniques of Hall et al. (2008) to multilevel cross-dependent functional data. On the other hand, in the case of rare-events (small number of 1's), different techniques are required, as the former approach provides biased results.

The article is organized as follows. Section 2 introduces the model framework. Section 3 highlights the challenges of the underlying estimation problem. Section 4 provides the estimation procedures under two approximations to the logistic function. We compare the performance of the two approximations in a simulation study in Section 5. Our application is presented in Section 6. Some technical details, additional simulation results and a description of our software implementations are deferred to the Web Appendices.

2 Model Framework

Here we describe the modeling framework and the associated assumptions. Let $Y_{rij} = Y_r(d_{ri}, t_{rij})$ be the value of the response functional observation $Y_r(d_{ri})$ at time point t_{rij}, where d_{ri} denotes the fixed design point associated with this functional observation. Here r indexes the subjects in the sample, $r = 1, \ldots, R$, i indexes the units within the subject at which we observe functional observations, $i = 1, \ldots, m_r$, and j indexes the subunits at which the functional responses are sampled, $j = 1, \ldots, n_{ri}$. In this paper we also use the notation $\mathbf{Y}_{ri} = Y_r(d_{ri})$ to refer to the functional observation for the rth subject at unit d_{ri}. It is assumed that $t_{rij} \in \mathcal{T}$ for some compact interval, for simplicity $\mathcal{T} = [0, 1]$, and $d_{ri} \in \mathcal{D}$ for some open-set domain, for simplicity taken $\mathcal{D} = [0, \infty)^{dim}$, for all r, i and j, where the dimensionality of the spatial domain is $dim \geq 0$. When $dim = 0$, there is no cross-dependence between random functions observed for the same subject. When $dim = 1$, the domain can be time or some other unidimensional functional domain. When $dim = 2$, the domain can be a geographic space, for example.

We assume without loss of generality that $Y_r(d, t)$ are independent realizations of a stochastic process Y observed on the domain $\mathcal{D} \times \mathcal{T}$. A key assumption in our modeling approach is the existence of the latent process X_r such that:

$$E\{Y_r(d_1, t_{11}), \ldots, Y_r(d_m, t_{mn_m})|X_r\} \;=\; \prod_{i=1}^{m} \prod_{j=1}^{n_i} g\{X_r(d_{ri}, t_{ij})\}; \qquad (1)$$

where $0 \leq t_{i1} < \ldots < t_{in_i} \leq 1$, for all i, the function g is a smooth monotone increasing link function. For binary data, the link is $g(x) = \exp(x)/\{1 + \exp(x)\}$, corresponding to the canonical form of the location parameter for the Binomial distribution. To account for the correlation structure of the observed response, we assume a structural decomposition for the latent process X_r into components that exhibit both cross- and within-unit dependence, similar to Staicu et al. (2010). We note that in Staicu et al. (2010), the link function

is $g(x) = x$ corresponding to the canonical form of the location parameter for the normal distribution. Our modeling procedure is more challenging since the relationship between the functional data and the latent process is nonlinear.

More specifically, we assume that for any fixed design point $d_{ri} \in \mathcal{D}$ and $t \in \mathcal{T}$ we have

$$X_r(d_{ri}, t) = \mu(t) + Z_r(t) + \epsilon(d_{ri}, t), \tag{2}$$

where μ is the overall mean function, Z_r is the random deviation from the mean function that is common to all the units, and $\epsilon(d_{ri}, t) = W_{ri}(t) + U_r(d_{ri})$ represents a specific random deviation additively separable in 't' and 'd'. The model structure in equation (2) decomposes the variability in the latent process X_r into between-subject variability $Z_r(t)$ and within-subject variability $\epsilon(d_{ri}, t)$. We assume no dependence across subjects, i.e. $Z_r(t)$ does not depend on the design points $d_{ri}, i = 1, \ldots, m_r$. However, units are cross-dependent; that is, the units are correlated given the subject.

We assume that Z_r and W_{ri} are independent Gaussian processes, with mean zero and covariance functions K_Z and K_W, respectively, defined by $K_Z(t, t') = E\{Z_r(t)Z_r(t')\}$ and $K_W(t, t') = E\{W_{ri}(t)W_{ri}(t')\}$ for $t, t' \in \mathcal{T}$. Also U is a zero-mean second order stationary and isotropic Gaussian process with auto-covariance function $\nu(\cdot)$ defined by $E\{U_r(d)U_r(d')\} = \nu(\|d - d'\|) = \sigma_u^2\rho(\|d - d'\|)$, for any $d, d' \in \mathcal{D}$,where σ_u^2 is the variance, $\rho(\cdot)$ is the auto-correlation function, and $\|\cdot\|$ is the Euclidean distance in $\mathcal{D}$. Furthermore we assume that $\lim \rho(\Delta) = 0$ as $\Delta \to \infty$.

Within this modeling framework, our objective is to estimate the components of the latent structure X_r: the mean and the covariance functions.

3 Model Estimation: Overall Approach

We now discuss the challenges in the estimation of the mean and covariance functions of the latent model components of (2) for binary-valued functional data while considering a simpler

form of it, that is, when $W_{ri}(t) = 0$ and $U_r(d) = 0$. This reduced model was introduced by Hall et al. (2008). In this section, we discuss this reduced model to simplify our derivations motivating various approximation of the link function.

When the functional observations are binary-valued, the mean and covariance functions of the latent structure X_r cannot be estimated using existing methods such as Di et al. (2009) or Staicu et al. (2010), due to the nonlinearity in the link function g. We therefore contrast three different approximations to the logistic function with different applicability.

Linear Approximation: Approximate $g(x)$ using Taylor expansion around zero and truncate the expansion assuming x is small. Therefore, the underlying assumption is that the variation of $X_r(t)$ around its mean function $\mu(t)$ is relatively small or $X_r(t) - \mu(t)$. Such a method is computationally efficient and works well as long as the marginal probabilities $\alpha(t) = \text{pr}\{Y_r(t) = 1\}$ stay away from the endpoints 0 and 1. However, this method fails when the probabilities are very small, the case of rare events. In the Web Appendix A, we provide more insights into this limitation of the linear approximation within a simulation study following the simulation settings of Hall et al. (2008) but under the rare-events case. We find that the simulation of Hall et al. (2008) only validates the use of the linear approximation under the non-rare event case.

Adjusted Exponential Approximation: Employ a Taylor approximation to the link function g by using $\exp(x)/\{1+\exp(x)\} \approx \exp(x)\{1-\exp(x)\}$ assuming that $\exp(x)$ is small. This approximation implies $\exp\{X_r(t) - \mu(t)\}$ small or $\text{pr}\{Y_r(t) = 1\}$ small. The marginal mean function of $Y_r(t)$, $\alpha(t) = E\{Y_r(t)\} = \text{pr}\{Y_r(t) = 1\}$ could therefore be approximated by

$$\widetilde{\alpha}(t) = \exp\{\mu(t) + K_Z(t,t)/2\} - \exp\{2\mu(t) + 2K_Z(t,t)\} \approx \frac{\exp\{\mu(t) + K_Z(t,t)/2\}}{[1 + \exp\{\mu(t) + 3K_Z(t,t)/2\}]}. \quad (3)$$

Exponential Approximation: Drop the second term in the approximation of $\alpha(t)$ in

the adjusted exponential approximation (3), resulting in an exponential approximation to the link function; we denote $\widetilde{\alpha}(t) = \exp\{\mu(t) + K_Z(t,t)/2\}$. This approximation assumes that $\exp(x)$ is very small implying that $\text{pr}\{Y_r(t) = 1\}$ is very small. The exponential approximation is commonly used in the case of biomedical and epidemiological applications with rare events; see for example Piegorsch, et al. (1994), Epstein and Satten (2003), Kwee, et al. (2007) among others. In this paper, we focus on this approximation instead of the adjusted exponential because of the feasibility in deriving estimates for the covariance functions.

Comparison. We compare the linear and exponential approximations in Figure 1 using a simulation study following the setting by Hall et al. (2008) but replacing the mean function with $\mu(t) \longleftarrow \mu(t) - 4$, a rare-events case. Both exponential approximations are clearly more accurate than the linear approximation in estimating $\alpha(t)$. Moreover, in the rare event case, the estimate of μ obtained using the adjusted exponential approximation is very accurate whereas the linear approximation is very biased; and this is because in the rare event case, the influence of the variance function $K_Z(t,t)$ to the marginal mean function $\alpha(t)$ is not negligible. By accounting for the variance function, the two versions of exponential approximations give reliable estimates of the mean function.

The *underlying message* of this comparison is that for binary functional data, different methods need to be employed according to the prevalence of events, and particularly, use the exponential approximation for rare events data.

4 Model Estimation: Procedure

This section details the estimation of the components of the more general model in (2) under the linear and exponential approximations introduced in the previous section. The last subsection particularly contrasts the differences and similarities in the estimation procedure for the two approximations.

4.1 Non-rare events setting: Linear approximation

Assuming that the variation of X_r about its mean is relatively small and using the Taylor expansion of $g\{X_r(d_{ri},t)\} = E\{Y_r(d_{ri},t)|X_r(d_{ri},t)\}$ about $\mu(t)$ as in Hall et al. (2008), we derive the approximation

$$\begin{aligned} g\{X_r(d_{ri},t)\} \approx g\{\mu(t)\} &+ \{Z_r(t)+W_{ri}(t)+U_r(d_{ri})\}g'\{\mu(t)\} \\ &+ (1/2)\{Z_r(t)+W_{ri}(t)+U_r(d_{ri})\}^2 g''\{\mu(t)\}, \end{aligned} \tag{4}$$

where $g'(t) = \partial g(t)/\partial t$, and $g''(t) = \partial^2 g(t)/\partial t^2$. It follows that the marginal probabilities can be approximated by $\text{pr}\{Y_r(d_{ri},t)=1\} \approx g\{\mu(t)\} + \frac{1}{2}\{K_Z(t,t) + K_W(t,t) + \sigma_u^2\}\, g''\{\mu(t)\}$. This approximation, while accurate, may not be easy to apply; following Hall et al. (2008) we consider a simpler alternative, which ignores the second order-term of the approximation (4), and thus uses a linear approximation. The linear approximation of the marginal probability $\alpha(t) = \text{pr}\{Y_r(d_{ri},t)=1\}$ becomes $\alpha(t) \approx g\{\mu(t)\}$. Furthermore, using the linear approximation, we derive approximations for the marginal joint probabilities as

$$\begin{aligned} \text{pr}\{Y_r(d_{ri},t)=1, Y_r(d_{r\ell},t')=1\} \approx\ & g\{\mu(t)\}g\{\mu(t')\} + g'\{\mu(t)\}g'\{\mu(t')\} \\ & \times \{K_Z(t,t') + K_W(t,t')I(i=\ell) + \nu(\|d_{ri}-d_{r\ell}\|)\}, \end{aligned}$$

where $I(i=\ell)$ is the indicator function which equals 1 if $i=\ell$ and 0 otherwise.

Using the approximation of the marginal probabilities we can further derive approximate relationships between the total, within and between covariances of Y_r and the model components in (2) as follows. Denote by $\mathcal{S}_T^Y(t,t') = \text{cov}\{Y_r(d_{ri},t), Y_r(d_{ri},t')\}$ the total covariance of the observed process Y_r, by $\mathcal{S}_B^Y(t,t',\Delta) = \text{cov}\{Y_r(d_{ri},t), Y_r(d_{r\ell},t')$ the between-unit covariance and by $\mathcal{S}_W^Y(t,t',\Delta) = (1/2)E\left[\{Y_r(d_{ri},t) - Y_r(d_{r\ell},t)\}\{Y_r(d_{ri},t') - Y_r(d_{r\ell},t')\}\right]$ the within-unit covariance at time points (t,t') and for units located at distance $\Delta = \|d_{ri} - d_{r\ell}\|$.

$$\mathcal{S}_T^Y(t,t') \approx \{K_Z(t,t') + K_W(t,t') + \sigma_u^2\}g'\{\mu(t)\}g'\{\mu(t')\}; \tag{5}$$

$$\mathcal{S}_B^Y(t,t',\Delta) \approx \{K_Z(t,t') + \nu(\Delta)\}g'\{\mu(t)\}g'\{\mu(t')\}; \quad (6)$$

$$\mathcal{S}_W^Y(t,t',\Delta) \approx \{K_W(t,t') - \nu(\Delta) + \sigma_u^2\}g'\{\mu(t)\}g'\{\mu(t')\}, \quad (7)$$

where $\nu(\Delta) = \text{cov}\{U_r(d'), U_r(d'')\} = \sigma_u^2\rho(\Delta)$ and $\Delta = \|d - d'\|$ is the covariance at lag Δ of the process U_r. These approximations are a natural generalization of the method introduced by Hall et al. (2008) to the multilevel functional model. They are not limited to the logit link function; they hold for any link function for which g' does not vanish and furthermore $\inf_t\{g'(t)\} > 0$. In addition, these equations can be regarded as extensions of the equations (3.4) - (3.6) of Staicu et al. (2010) to the case when the curves are binary-valued.

Equations (5)-(7) provide the intuition behind the road map of the estimation procedure consisting of three steps:

Step 1. Obtain an estimator of the mean function $\mu(t)$ from the linear approximation of the marginal probability;

Step 2. Use equation (7) to estimate the covariogram $\nu(\Delta)$;

Step 3. Estimate the covariance functions K_Z and K_W, using (5)-(7).

We describe the estimation approach for the model components $\mu(t)$, $\nu(\Delta)$, K_Z and K_W in detail in Web Appendix B.

4.2 Rare events setting: Exponential approximation

In this section, we use an exponential approximation for the conditional probability of an event $g\{X_r(d_{ri}, t)\}$. Particularly, we extend the approximation in (3) to the more general model in (2) to approximate the marginal probabilities by

$$\text{pr}\{Y_r(d_{ri},t) = 1\} \approx \frac{\exp\left[\mu(t) + \{K_Z(t,t) + K_W(t,t) + \sigma_u^2\}/2\right]}{1 + \exp\left[\mu(t) + 3\{K_Z(t,t) + K_W(t,t) + \sigma_u^2\}/2\right]}; \quad (8)$$

The approximation in (8) can be used to obtain an approximate estimate for $\mu(t)$ as soon as we have estimates for the covariance functions, $K_Z(t,t)$, $K_W(t,t)$ and σ_u^2. Similarly to the linear approximation, the covariance functions are estimated from an approximate relationship between these functions and the marginal joint probabilities.

The Taylor approximation to the link function from which we derive (8) cannot be used to derive such an approximation for the marginal joint probabilities. For estimating the covariance functions, we instead use the exponential approximation and derive the approximate relationship of the marginal joint probabilities to the covariance functions

$$\text{pr}\{Y_r(d_{ri},t)=1, Y_r(d_{r\ell},t')=1\} \approx \alpha(t)\alpha(t')\exp\left\{K_Z(t,t')+K_W(t,t')I(i=\ell)+\nu(\|d_{ri}-d_{r\ell}\|)\right\}.$$

For this approximation, it is more convenient to work with the marginal second moments $E\{Y_r(d_{ri},t)Y_r(d_{r\ell},t')\}$ than the marginal covariance $\text{cov}\{Y_r(d_{ri},t), Y_r(d_{r\ell},t')\}$.

Before describing the general procedure we introduce some additional notation. Denote by $\mathcal{E}_T^Y(t,t') = E\{Y_r(d_{ri},t)Y_r(d_{ri},t')\}$ the marginal second moment of $Y_r(d_{ri},t)$, and by $\mathcal{E}_B^Y(t,t',\Delta) = E\{Y_r(d_{ri},t)Y_r(d_{r\ell},t')\}$ the between-unit marginal second moment for units which are at distance $\Delta = \|d_{ri}-d_{r\ell}\|$ apart. The quantities $\mathcal{E}_T^Y$ and $\mathcal{E}_Y^B$ correspond to the total covariance, K_T^Y, and between-unit covariance, K_B^Y, respectively, introduced in the context of the linear approximation. Simple algebra yields

$$\mathcal{E}_T^Y(t,t') \approx \alpha(t)\alpha(t')\exp\left\{K_Z(t,t')+K_W(t,t')+\sigma_u^2\right\}; \quad (9)$$

$$\mathcal{E}_B^Y(t,t',\Delta) \approx \alpha(t)\alpha(t')\exp\left\{K_Z(t,t')+\nu(\Delta)\right\}. \quad (10)$$

Equations (8)-(10) provide the intuition behind the road map of the estimation procedure consisting of three steps:

Step 1. Obtain an estimator of the cross-dependence covariance, $\nu(\Delta)$ - this step requires estimation of the marginal mean function $\alpha(t)$ and of $\mathcal{E}_B^Y(t,t',\Delta)$;

Step 2. Use equations (9)-(10) to estimate covariance functions $K_Z(t,t')$ and $K_W(t,t')$;

Step 3. Estimate the mean function $\mu(t)$ using equation (8), by substituting the estimates for all the covariance functions.

We describe the estimation approach for the model components $\mu(t)$, $\nu(\Delta)$, K_Z and K_W in detail in Web Appendix C.

4.3 Linear vs. Exponential Approximation

We conclude this section with a comparison of our derivations for the model component estimates under the two approximations. The primary differences are summarized as follows:

• The estimate of the mean function $\mu(t)$ does not depend on $K_Z(t,t'), K_W(t,t'), \sigma_u^2$ under the linear approximation but it does depend on these components under the exponential approximation.

• Because of the underlying formulas derived under the two approximations, compare equations (5)-(7) to (8)-(10), $K_Z(t,t'), K_W(t,t'), \nu(\Delta)$ and σ_u^2 are estimated using different approximation functions of the within and between covariance functions of Y.

• The estimates of the total, within and between covariance functions of Y differ for the two approximations. For the linear approximation we estimate them assuming a Gaussian process using similar methods as developed by Staicu et al. (2010). In contrast, for the exponential approximation we employ smoothing techniques to estimate these covariance functions. Therefore, the estimation procedure under linear exponential is less computationally expensive than the one under the exponential approximation.

5 Simulation Study

In this section we present a simulation study to assess the finite sample performance of the proposed estimation methodology. The primary objective of the simulation study is to

compare the estimation accuracy under the two approximations for the logistic link function. We focus on the estimation accuracy of the mean function $\mu(t)$ as well as of the covariance functions $K_Z(t,t')$ and $K_W(t,t')$.

Simulation Settings. We generate data following the model in (1) under two settings: (1) non-rare events setting, $\mu(t) = 2\sin(2\pi t)/\sqrt{5}$, and (2) rare events setting $\mu(t) = 2t^2 + 2t - 5$, for $t \in [0,1]$.

For both simulation settings, the latent process is generated using the decomposition (2). The cross-dependence process U_r is assumed Gaussian with covariance function specified by the Matérn correlation function (Matérn, 1986)

$$K_S(\|d_i - d_j\|) = \sigma_u^2 \frac{1}{\Gamma(\rho)} \left(\frac{\phi\|d_i - d_j\|}{2} \right)^{\rho} 2B_\rho(\phi\|d_i - d_j\|),$$

where $\sigma_u^2 = 0.5$, $\phi = 1/4$ and $\rho = 2$. We compare accuracy results for various dimensionality of the domain $\mathcal{D}$, specifically, $dim = 0, 1, 2$. The results in this section are based on $dim = 1$. We include additional simulation results for $dim = 0$ and 2 in Web Appendix C.

The functional processes Z_r and W_{ri} are assumed Gaussian with covariance specified by the covariance functions $K_Z(t,t') = \sum_{k\geq 1} \phi_k^Z(t)\phi_k^Z(t')\lambda_k^Z$ and $K_W(t,t') = \sum_{l\geq 1} \phi_l^W(t)\phi_l^W(t')\lambda_l^W$. We set $\lambda_1^Z = 0.5, \lambda_2^Z = 0.25$, and $\lambda_k^Z = 0$ for $k \geq 3$ and $\lambda_1^W = 0.5, \lambda_2^W = 0.15$ and $\lambda_l^W = 0$ for $l \geq 3$. Also we take $\phi_1^Z(t) = \sqrt{2}\cos(2\pi t)$, $\phi_2^Z(t) = \sqrt{2}\cos(4\pi t)$ and $\phi_1^W(t) = \sqrt{3}(2t-1)$, $\phi_2^W(t) = \sqrt{5}(6t^2 - 6t + 1)$. We simulated the structure of the covariance functions of the processes Z_r and W_{ri} using the Karhunen-Loève (KL) expansion (Karhunen, 1947; Loéve, 1945). Specifically, $\phi_k^Z(t)$ and λ_k^Z for $k \geq 1$ are the eigenfunctions and eigenvalues, respectively, for the KL decomposition of the covariance function $K_Z(t,t')$. Similarly, $\phi_k^W(t)$ and λ_k^W for $k \geq 1$ are the eigenfunctions and eigenvalues, respectively, for the KL decomposition of the covariance function $K_W(t,t')$. Using these covariance structures, the processes Z_r and W_{ri} are non-stationary - more general, but more realistic, modeling assumptions.

We generate 100 data sets following the simulation settings above with a total of $R = 50$ subjects, $M = 50$ units and $N = 25$ sub-units. Estimates of the mean function and covariance functions are obtained for each of the two approximations using the methods introduced in Web Appendices B and C. For visual assessment of the covariance function estimates, K_Z and K_W, we use plots of the eigenfunctions corresponding to their spectral representation.

• Figure 2 shows the estimates for $\mu(t)$ provided by the two approximation methods for all 100 simulations (in grey) and the true mean function (in black). The estimated means for the rare events setting are (negatively) biased when using the linear approximation (Figure 2b) but approximately unbiased under the non-rare events setting (Figure 2a). On the other hand, the estimated means under the rare events setting are unbiasedly estimated when using the exponential approximation (Figure 2d) as compared to the linear approximation (Figure 2b). Therefore, in the rare events setting, the exponential approximation provides considerably more accurate estimates for $\mu(t)$.

• Figures 3 and 4 present the estimated (in grey) and true (in blue) eigenfunctions of the covariance function $K_Z(\cdot, \cdot)$ and $K_W(\cdot, \cdot)$ for the linear and exponential approximations under the rare event case. We also compared the estimated eigenfunctions under the non-rare event setting (see Web Appendix C). We found that in the non-rare events setting, the eigenfunctions are more accurately estimated than under rare events; the improvement is more significant for the between covariance function. This is to be expected, since the accuracy of logistic regression depends on the number of cases.

• Figure 5 presents the estimated cross-correlation function (in grey) in contrast to the true correlation function for the two settings. These estimates are derived using the linear approximation approach. The estimates derived from the exponential approximation approach are similar. As expected, the cross-dependence between the longitudinal binary observations is more accurately estimated for non-rare events than for rare events, regardless

of the method of estimation employed.

• Comparing the estimates for all model components (see Web Appendix C for a table including mean square errors), the most significant improvement in the estimation accuracy is for the mean function $\mu(t)$ and the level-1 covariance function $K_Z(\cdot,\cdot)$ when using the exponential over the linear approximation under the rare-events case.

6 Data Analysis

We consider the analysis of spectral backscatter from long range infrared light detection and ranging (LIDAR) data. The data are described in detail by Carroll et al. (2012), and the estimation of spectral backscatter uses the algorithm of Warren et al. (2008, 2009), but applied to the observed data rather than the deconvolved data. Our main goal is to estimate the probability of high spectral backscatter at different wavelengths.

In the experiment, 30 aerosol clouds are investigated. There were two types of clouds: control clouds that were non-biological in nature and treatment clouds that were biological. For each cloud $r = 1, ..., 30$, functional responses were sampled at $i = 1, ..., 50$ locations d_{ri}, called bursts, these locations being sampled one second apart. Within each location we observe CO_2 laser wavelengths, denoted as $t_{rij} = 1, \ldots, 19$. We then defined high spectral backscatter as being above 0.30, roughly the 90^{th} percentile of all the backscatter data. Thus, $Y_{rij} = Y_r(d_{ri}, t_{rij})$ is the indicator that the estimated backscatter corresponding to the rth cloud, observed at the d_{ri} location/burst and for wavelength t_{rij} is larger than 0.30.

We use the proposed modeling approach and assume that the observed data can be modeled using (1) and (2), with the difference that in (2) the latent mean function accounts for the two groups. Specifically, $\mu(t) = \mu_0(t)I\{G(r) = 0\} + \mu_1(t)I\{G(r) = 1\}$, where $I(\cdot)$ is the indicator function and $G(r)$ denotes the group membership of the cloud r, 0 for the control and 1 for the treatment group. Because a high value of the spectral backscatter

represents a rare event, we use an exponential approximation to estimate the mean and covariance functions of the latent process. The methodology requires slight modifications to account for different group mean functions. For the estimation of the covariance functions K_Z, K_W, and ν we consider first estimating them based on the data of each group; the final estimators are obtained by averaging the group estimators. The group mean estimates $\widehat{\mu}_0(t)$ and $\widehat{\mu}_1(t)$ are based on the estimates of the covariance functions and on the marginal probability estimates.

Figure 6 (a) depicts the estimates of the mean functions of the high spectral backscatter. The two group mean functions have similar shapes, showing two peaks and two dips, but there are differences also. There seems to be a delay between the wavelengths at which the local extremes occur the treatment group and those in the control group. The latent mean function has larger values in the control group than in the treatment group, indicating that the log odds ratio of high spectral backscatter is slightly smaller in the treatment group than in the control group. Interestingly, the estimated group mean functions using the linear approximation (results shown in the Web Appendix D) are very similar, shapewise, to the ones shown in Figure 6(a) but the magnitude of their values is smaller.

Figure 6(b) displays the estimated spatial correlation $\widehat{\nu}(\Delta)/\widehat{\nu}(0)$ as a function of the normalized location/bursts, namely $\Delta_{rij} = |d_{ri} - d_{rj}|/50$. The results indicate that the correlation does not die out rapidly. Specifically, there seems to be high correlation between the spectral backscatter measurements taken within 20 seconds of one another, where the estimated correlation is larger than 0.8. The correlation decreases almost linearly as the measurements are between 20 and 40 seconds apart and it becomes negligible as the measurements are more than 40 seconds apart.

Summaries of the estimated covariances of the two latent processes, $Z_r(\cdot)$ and $W_{ri}(\cdot)$ using the exponential approximation are illustrated in Figure 6(c) and Figure 6(d). The

plots show that the first three eigenfunctions for $\widehat{K}_Z$ and first two eigenfunctions for $\widehat{K}_W$ describe most of the variability (more than 95%) between and within aerosol clouds ($N_Z = 3$ and $N_W = 2$). The interpretation of these results is more challenging. For example, the first estimated eigenfunction provides the following insights. Aerosol clouds with positive scores on the first eigenfunction for the between variability tend to have a log odds ratio of the high spectral backscatter indicator that is smaller than the population average for the first set of wavelengths and larger for the second set of wavelengths.

The figures in Web Appendix D display the eigenfunctions estimated by the linear approximation method. The first two eigenfunctions that explain about 95% of the variability at level 1 are in fact similar only scaled differently. Only the first eigenfunction at level 1 is similar to the corresponding one estimated using the exponential approximation. Moreover, the first functional component at level 2 coincides with the one estimated using the exponential approximation but only one component explains about 95% of the variability at level 2, and therefore, the second component as estimated by the exponential approximation is not uncovered. This suggests that the linear approximation doesn't capture the finer structures of the covariance structures at level 1 and level 2.

The exponential approximation estimated the spatial variance to be $\sigma_u^2 = 0.12$, the eigenvalues at level 1 to be $\lambda_1^Z = 0.84$, $\lambda_2^Z = 0.36$, $\lambda_3^Z = 0.22$ (the first three components explain approximately 95% of the variability) and the the eigenvalues at level 2 to be $\lambda_1^W = 0.19$, $\lambda_2^W = 0.02$ (the first two components explain approximately 98% of the variability). This results indicates that the variability explained at level 1 is roughly 7 times larger than the variability at level 2. This difference is even higher when comparing the estimated eigenvalues from the linear approximation method - the sum of eigenvalues at level 1 is 4.07 as compared to 0.193 at level 2.

7 Discussion

We have provided a means for decomposition, estimation and interpretation for modeling multilevel binary functional data. One challenge is that linearization of the logistic function as proposed by Hall et al. (2008) applies only under a non-rare events setting and it provides biased estimates under the setting of rare events. We therefore developed an estimation procedure based on the exponential approximation to the inverse link function and compared it to the linear approximation to assess their advantages and limitations.

We illustrated the estimation bias due to the linear approximation under the rare events setting in both our simulation study as well as in one motivating case study. The bias is most significant in the estimation of the mean function $\mu(t)$ and the between covariance function $K_Z(t, t')$. Because the estimation of the within covariance and spatial dependence are computationally expensive under the exponential approximation, we therefore recommend using the linear approximation approach for the estimation of the cross-covariance represented by U_{ri} when the number of units M is medium to large.

We highlight here a more difficult setting than the one presented in this paper - multilevel binary-valued functional data where the functional observations are observed sparsely. Hall et al. (2008) motivate their methodology for sparse binary functional data; however, their theoretical results show that the model estimates are biased under this more difficult setting. We conjecture that neither of the two methods apply to sparse binary-valued functional data.

Because our estimation procedure for modeling rare-events longitudinal data is derived assuming the exponential link function, this procedure can also be applied to Poisson count data. Often these data are observed as events at random time points; events are then aggregated within fixed time intervals of equal length. Multilevel count longitudinal data with or without cross-dependence are also common in many applications including sales

retail data where each subject in our notation corresponds to a different product sold at different stores (units) over a period of time (subunits), or neural spike train data where units correspond to a set of neurons of a subject, for which response spike are recorded over a period of time (subunits). However, challenges in these types of applications arise when the counts are sparse, i.e. there are many zero values. Moreover, the Poisson assumption could be restrictive since the mean and variance are often not equal. Investigating functional-based approaches for longitudinal count data is beyond the scope of this paper.

We did not address the problem of predicting individual probability trajectories. The prediction approach by Hall et al. (2008) employs two layers of approximations. First, it uses the normal approximation for the binomial. Second, the expectation of the normal distribution is approximated using a second order linear approximation and the variance using a first order approximation. The first layer of approximation (from binomial to normal) relies on the assumption that $\mathrm{pr}\{Y_r(t) = 1\}$ is away from zero or one, the non-rare case. The second layer of approximation relies on the assumption of non-rare observations as discussed in this paper together with small eigenvalues $\lambda_k^Z, k = 1, \ldots$ and $\lambda_\ell^W, \ell = 1, \ldots$. Based on these observations, we conclude that prediction of the individual probabilities is much more challenging, a research topic by itself.

We also did not address the problem of making inference on the model components of the latent process X_r. Because of the nonlinear relationship between the observed process Y_r and the latent process X_r, there is not a direct way to derive confidence intervals for these model components. When analytical confidence intervals are difficult to derive, often one would resort to sampling techniques. For the nonparametric bootstrap, the predicted trajectories need to be estimated first. For the parametric bootstrap, one could sample from the Gaussian distribution of the latent components Z_r, W_{ri} and U_r to obtain confidence intervals for the predicted trajectories. Moreover, for estimating a bootstrap confidence

interval for $\mu(t)$, one has to re-sample the observed data as well. This last step needs to be performed with cautious due to the dependencies in the observed process Y_r.

Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3-6 are available with this paper at the Biometrics website on Wiley Online Library. In addition to the material referenced in the paper, we also provide the programs implemented in the R statistical software used in the simulation study and the case study of this paper. The software deliverable directory is available as a .zip file with three directories including the R code programs for different values of the dimensionality of the domain of the unit design points.

Acknowledgments

Serban's research was supported by the National Science Foundation Grant CMMI-0954283. Staicu's research was supported by U.S. National Science Foundation grant number DMS-1007466. Carroll's research was supported by the National Cancer Institute Grant R37-CA057030 and in part supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors thank to the referees and associate editor for helpful comments.

References

[1] Baladandayuthapani, V., Mallick, B., Turner, N., Hong, M., Chapkin, R., Lupton, J. and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* 64, 64-73.

[2] Carroll, R. J., Delaigle, A. and Hall, P. (2012). Deconvolution when classifying noisy data involving transformations. *Journal of the American Statistical Association* 107, 1166-1177.

[3] Di, C., Crainiceanu, C. M., Caffo, B. S. and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics* 3, 458-488.

[4] Epstein, M. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* 73, 1316-1329.

[5] Hall, P., Müller, H.-G. and Yao, F. (2008). Modeling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society, Series B* 70, 703-724.

[6] Kwee, L. C., Epstein, M. P., Manatunga, A. K., Duncan, R., Allen, A. S. and Satten, G. A. (2007). Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. *Genetic Epidemiology* 31, 75-90.

[7] Li, Y., Wang, N., Hong, M., Turner, N. D., Lupton, J. R. and Carroll, R. J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *Annals of Statistics*35, 1608-1643.

[8] Matérn, B. (1986). *Spatial Variation*. New York: Springer.

[9] Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association* 98, 573-583.

[10] Piegorsch, W. W., Weinberg, C. R. and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine* 13, 153-162.

[11] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Springer Series in Statistics, NY.

[12] Rice, J. A. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 57, 253-259.

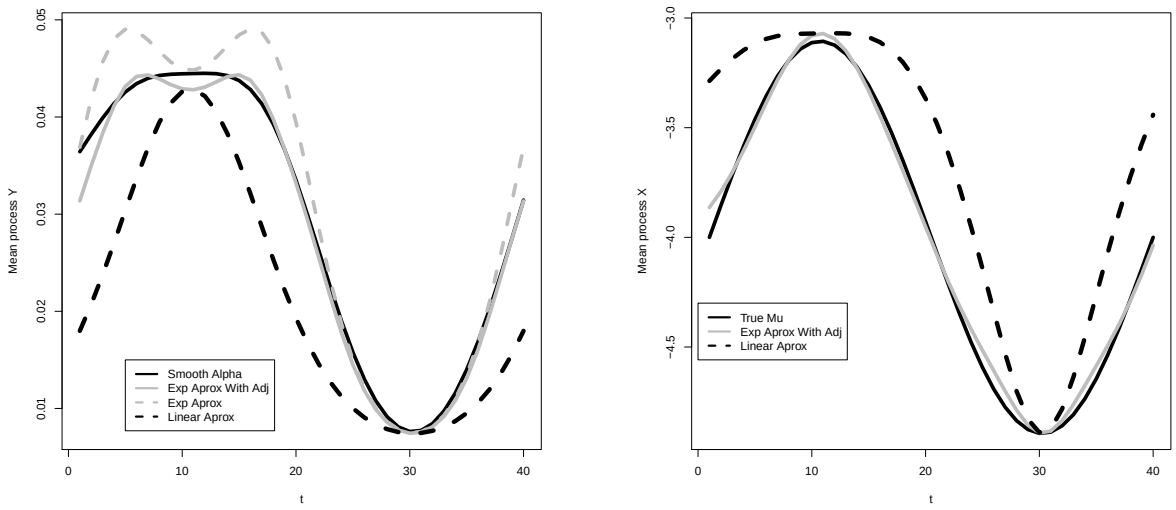
[13] Shi, M., Weiss, R. E., Taylor, J. M. (1996). An analysis of pediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics* 45, 151-163.

[14] Staicu, A.-M., Crainiceanu, C. M. and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics* 11, 177-194.

[15] Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*100, 577-590.

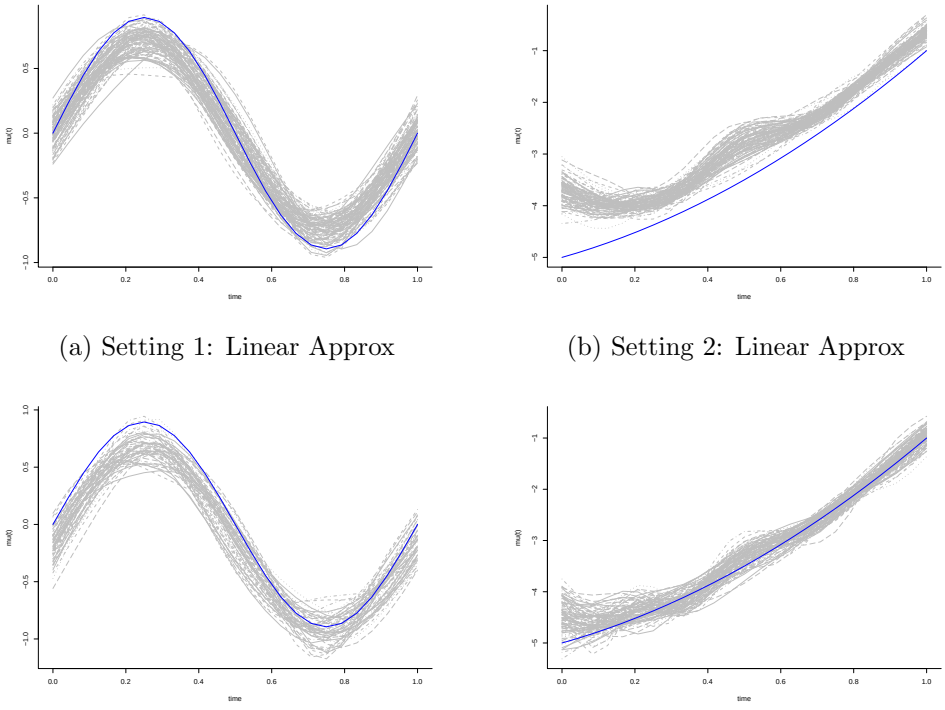
[16] Warren, R. E., Vanderbeek, R. G., Ben-David, A., and Ahl, J. L. (2008). Simultaneous estimation of aerosol cloud concentration and spectral backscatter from multiple-wavelength lidar data. *Applied Optics* 47, 4309-4320.

[17] Warren, R. E., Vanderbeek, R. G., and Ahl, J. L. (2009). Detection and classification of atmospheric aerosols using multi-wavelength LWIR lidar. *Proceedings of SPIE* 7304.



(a) (b)

Figure 1: Rare-events Setting: (a) Approximations of the marginal mean function $\alpha(t) = E\{Y_r(t)\}$ using the true function $\mu(t)$. Depicted are the Monte Carlo estimate $\widetilde{\alpha}(t)$ using a large number of samples, the linear approximation, the exponential approximation, and the adjusted exponential approximation. (b) Approximations of the latent mean function $\mu(t)$. Depicted are the true function $\mu(t)$, the estimates using the linear approximation, and the adjusted exponential approximation.



(a) Setting 1: Linear Approx

(b) Setting 2: Linear Approx

(c) Setting 1: Exponential Approx

(d) Setting 2: Exponential Approx

Figure 2: Mean estimates (in grey) compared to the true mean (solid line) for the two simulation settings under linear and exponential approximations. The top left panel is the result of the linear approximation for non-rare events, while the top right panel is the same approximation for rare events. The bottom left panel is the result of the exponential approximation in the non-rare event case, while bottom right panel is the same approximation in the rare event case. This figure appears in color in the electronic version of this article.

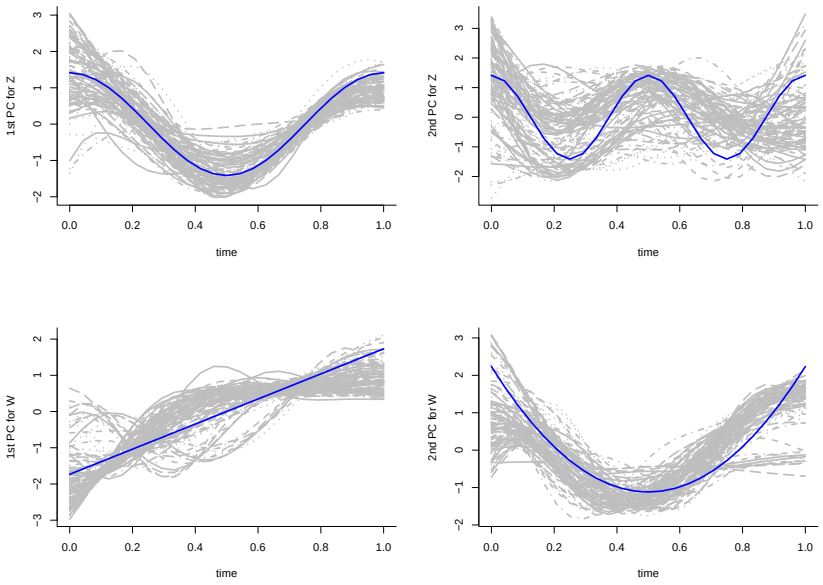


Figure 3: Eigenfunction estimates (in grey) using the linear approximation compared to the true eigenfunctions (solid line) for the rare event simulations (setting 2). The top panels are the eigenfunctions for the covariance $K_Z(t, t')$, while the bottom panels are the eigenfunctions for the covariance $K_W(t, t')$. This figure appears in color in the electronic version of this article.

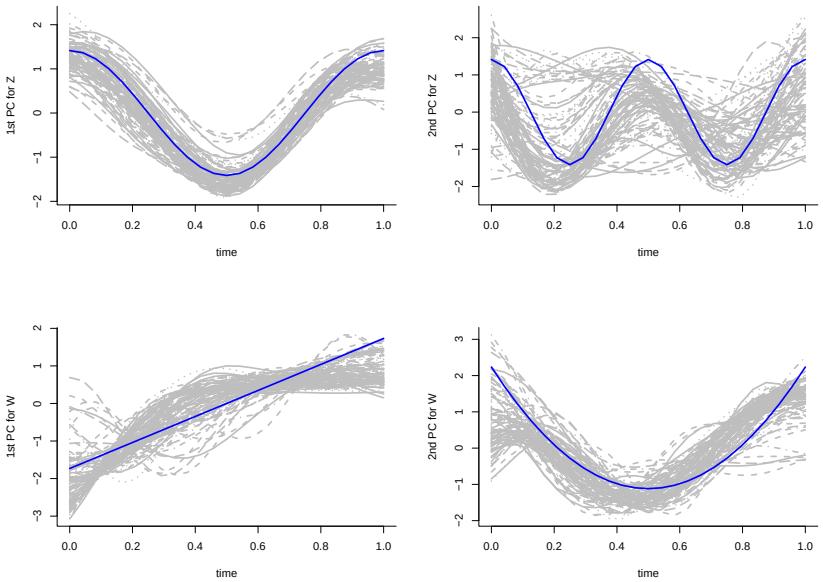
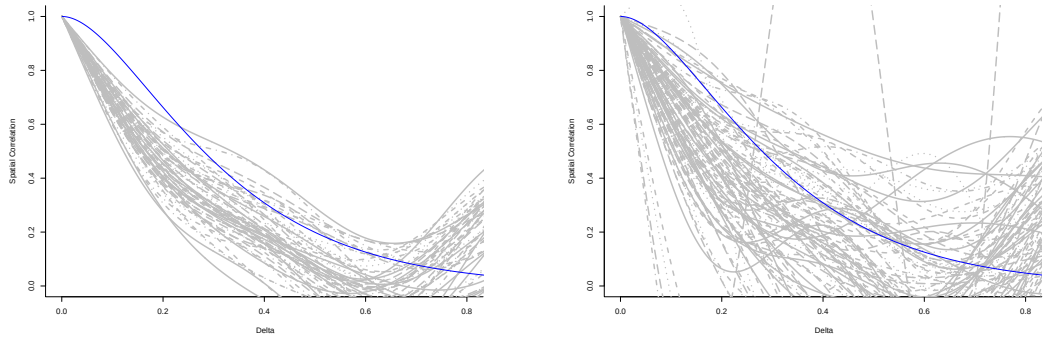
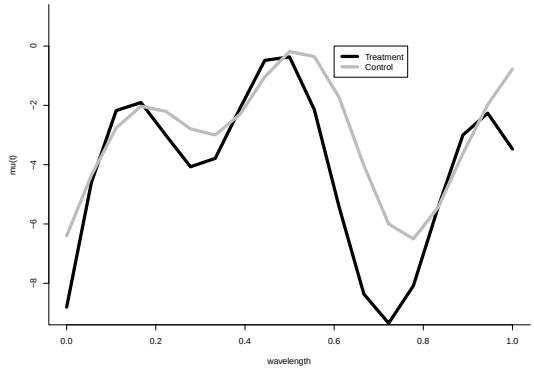


Figure 4: Eigenfunction estimates (in grey) using the exponential approximation compared to the true eigenfunctions (solid line) for the rare event simulations (setting 2). The top panels are the eigenfunctions for the covariance $K_Z(t,t')$, while the bottom panels are the eigenfunctions for the covariance $K_W(t,t')$. This figure appears in color in the electronic version of this article.

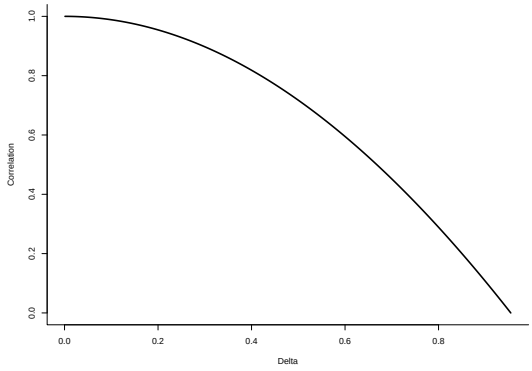


(a) Setting 1 (b) Setting 2

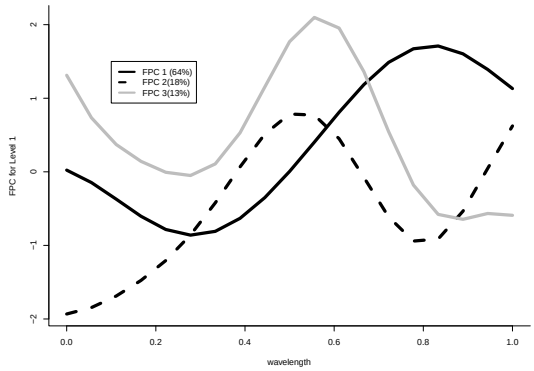
Figure 5: Covariance estimation for the non-rare event case (left panel) and the rare-event case (right panel) estimated using the linear approximation approach and compared to the true function (solid line). This figure appears in color in the electronic version of this article.



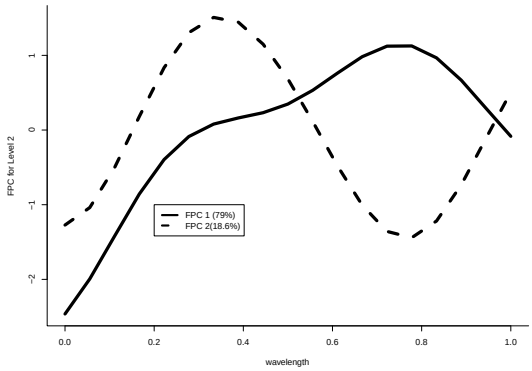
(a) Group mean of the high spectral backscatter



(b) Spatial Correlation



(c) Between-Clouds Covariance: Eigenfunctions



(d) Within-Clouds Covariance: Eigenfunctions

Figure 6: Displayed are: (a) estimated group mean functions of $\mu_g(t)$ for the control group ($g = 0$) and treatment group ($g = 1$), using the exponential approximation; (b) the estimated correlation function varying with the distance between locations/bursts; (c) the first three estimated eigenfunctions of the between-covariance function $K_Z(t, t')$; and (d) the first two estimated eigenfunctions of the within-covariance function $K_W(t, t')$. This figure appears in color in the electronic version of this article.