Published in final edited form as:
Stat Med. 2014 November 30; 33(27): 4825–4840. doi:10.1002/sim.6241.

Functional Linear Models for Zero-Inflated Count Data with Application to Modeling Hospitalizations in Patients on Dialysis

Damla Şentürk[a,†,*], **Lorien S. Dalrymple**[b], and **Danh V. Nguyen**[c,d]

[a]Department of Biostatistics, University of California, Los Angeles

[b]Division of Nephrology, Department of Medicine, University of California, Davis

[c]Department of Medicine, University of California, Irvine

[d]Institute for Clinical and Translational Science, University of California, Irvine

Summary

We propose functional linear models for zero-inflated count data with a focus on the functional hurdle and functional zero-inflated Poisson (ZIP) models. While the hurdle model assumes the counts come from a mixture of a degenerate distribution at zero and a zero-truncated Poisson distribution, the ZIP model considers a mixture of a degenerate distribution at zero and a standard Poisson distribution. We extend the generalized functional linear model framework with a functional predictor and multiple cross-sectional predictors to model counts generated by a mixture distribution. We propose an estimation procedure for functional hurdle and ZIP models, called penalized reconstruction (PR), geared towards error-prone and sparsely observed longitudinal functional predictors. The approach relies on dimension reduction and pooling of information across subjects involving basis expansions and penalized maximum likelihood techniques. The developed functional hurdle model is applied to modeling hospitalizations within the first two years from initiation of dialysis, with a high percentage of zeros, in the Comprehensive Dialysis Study participants. Hospitalization counts are modeled as a function of sparse longitudinal measurements of serum albumin concentrations, patient demographics and comorbidities. Simulation studies are used to study finite sample properties of the proposed method and include comparisons with an adaptation of standard principal components regression (PCR).

Keywords

functional data analysis; end stage renal disease; hurdle model; sparse longitudinal design; United States Renal Data System; zero-inflated Poisson model

1 Introduction

Functional data analysis has rapidly expanded in recent years, providing a framework for analysis of data which are curves or functions [1]. Of particular interest are regression

*Correspondence to: Damla Şentürk, Department of Biostatistics, University of California, Los Angeles, CA 90095.
†dsenturk@ucla.edu

models for outcomes with a functional covariate/predictor that varies over time (t) (e.g., [2]–[6], among others). Generalized functional linear models (GFLM; [2]) have been proposed to relate a generalized scalar outcome, Y, to a functional predictor, $X(t)$, along with cross-sectional covariates, $Z = (Z_1, \ldots, Z_p)$,

$$g(\mu) = \beta_0 + \int \beta(t) X(t) dt + \sum_{r=1}^{p} \alpha_r Z_r, \quad (1)$$

where $\mu = E\{Y|X(t), Z\}$. The regression coefficient function, $\beta(t)$, can be interpreted as a weight function, capturing the variation over the support (t) of the functional predictor that is associated with the outcome.

For count outcomes, such as the number of hospitalizations during an observation period, the standard Poisson distribution does not fit well when there are excess zeros. For example, in the Comprehensive Dialysis Study (CDS; [7]), which is used to illustrate the proposed methods (in Section 5), 53% of patients were not hospitalized during the study follow-up period. Useful models for counts with excess zeros include the hurdle model [8]–[11] and the zero-inflated Poisson (ZIP) model [12]. As we will detail in the proposed functional hurdle model specification, in Section 2, each patient on dialysis has an individual-specific probability of having a hospitalization that depends on the functional predictor $X(t)$ and baseline covariates Z_r; thus, a binomial probability model governs the binary outcome process of having zero vs. a positive number of hospitalizations. Once the realization is positive, the hurdle is crossed and the conditional distribution of positive counts of the number of hospitalizations is modeled as a zero-truncated Poisson distribution. Given all patients on dialysis are at some level for risk of hospitalization, the functional hurdle model is conceptually preferable; however, the ZIP model is particularly relevant in other applications where the counts arise from mixing a standard Poisson process with a degenerate process at zero, producing “structural” zeros (e.g., abstainers in drug abuse, smoking or sexual behavioral studies).

In this work, we propose functional hurdle and ZIP models to relate a zero-inflated count outcome (number of hospitalizations) to a functional predictor, specifically, albumin concentrations sampled during the first two years of dialysis, a life-sustaining treatment for individuals with end-stage renal disease. Albumin concentrations are associated with hospitalizations and death in patients on dialysis ([13], Chapter 3). Our proposed functional regression models for zero-inflated count data are also aimed towards sparsely sampled functional predictors, which are common in longitudinal data. For example, in the CDS, serum collection was scheduled quarterly within the first two years after the start of dialysis; however, albumin measurements were sparse with the total number of measurements per subject ranging from 1 to 5, due to incomplete serum collection.

We note that common estimation techniques proposed for GFLMs rely on basis expansion of the functional predictor $X(t)$ and coefficient function $\beta(t)$ for dimension reduction, followed by least squares, maximum likelihood or penalized maximum likelihood estimation [14]–[16]. Cardot and Sarda [17] and Long [18] propose spline basis for expanding $X(t)$ and $\beta(t)$ followed by penalized maximum likelihood for densely observed functional data. Müller

and Stadtmüller [19] expand both $X(t)$ and $\beta(t)$ on the functional principal components basis of $X(t)$, followed by weighted least squares estimation. We will refer to this approach as functional principal components regression (PCR). Goldsmith et al. [20] argue that the few functional principal components chosen in applications may not be an adequate basis choice to expand $\beta(t)$, since $\beta(t)$ may not lie in the space spanned by the principal components functions of the functional predictor. Hence, they propose to expand the functional predictor on its functional principal components basis, and to model the functional regression function as penalized splines. They suggest the use of a large number of basis functions in both expansions and introduce regularization by using restricted maximum likelihood in an associated mixed effects model for the choice of the smoothing parameter. As is used in other functional regression models, their proposal is geared towards sparsely sampled functional predictor processes observed with additive measurement error via the use of functional principal components analysis [21]–[25].

We propose an estimation procedure for functional hurdle and ZIP models with a sparsely observed functional predictor process, similar to the measurements of albumin concentration in the CDS, potentially observed with measurement error. These are novel developments, particularly within the context of sparse designs, because sparse data are commonly encountered in longitudinal studies, where repeated measurements are available on each subject for a small total number of measurements (i.e., infrequent) at a set of subject-specific time points (i.e., irregular). Using spline basis directly in the expansions of the predictor process and regression coefficient function, as proposed by Cardot and Sarda [17] and Long [18], is not feasible in sparse data applications. Hence, our proposed estimation procedure adds a *reconstruction* step which makes spline basis expansion feasible. The proposed penalized reconstruction (PR) method begins by reconstructing the sparse longitudinal measurements on the predictor process on a dense grid via functional principal components analysis. The regression functions are then expanded on spline basis and coefficients in the expansion are estimated via penalized maximum likelihood using the reconstructed functional predictor. After basis expansions, Goldsmith et al. [20] induce regularization by using random coefficients and carrying out estimation in an associated generalized mixed effects model. We consider penalized likelihood estimation instead, because it extends to hurdle and ZIP modeling for zero-inflated counts more conveniently and avoids the computational challenges of fitting a hurdle or a ZIP model with a large number of random effects.

Specification of the functional hurdle and ZIP models are proposed in Section 2. Section 3 details the proposed estimation method (PR) for functional hurdle and ZIP models, as well as its applicability to generalized functional linear models (1). Because the estimation machinery developed in this paper is applicable for a generalized outcome, such as a binary outcome, in the GFLM model (1), we unified the presentation of the proposed estimation approach so that it is applicable to the GFLM generally. For comparison, we also describe an extension of PCR estimation in Section 3. Simulation studies examining the relative efficacy of the proposed estimation procedure and an extension of PCR are described in Section 4. We illustrate the proposed method with the aforementioned CDS data, where we utilized the functional hurdle model to examine the relationship between hospitalization and

a functional covariate, serum albumin concentration, together with baseline covariates (Section 5). We conclude with a brief discussion in Section 6.

2 Functional Hurdle and ZIP Models for Zero-Inflated Count Data

We introduce the functional hurdle and ZIP models for zero-inflated count data. We begin with the functional hurdle model; the functional ZIP model development will proceed similarly. The hurdle process models a count response, Y_i, as arising from a point mass at zero and a zero-truncated Poisson process. More precisely, the proposed hurdle distribution for modeling a count response, Y_i, that depends on a functional predictor $X_i(t)$, in addition to baseline predictors Z_{ri}, is

$$\Pr\{Y_i=y_i|X_i(t),Z_{ri}\}=\begin{cases}1-p_i, & \text{for } y_i=0,\\ p_i\frac{e^{-\lambda_i}\lambda_i^{y_i}}{(1-e^{-\lambda_i})y_i!}, & \text{for } y_i>0\end{cases}, \quad (2)$$

where the probability of having a positive count, namely $p_i = \Pr\{Y_i > 0|X_i(t), Z_{ri}\}$ (i.e., the parameter in the binary process), and the Poisson process rate λ_i of the positive counts (i.e., the parameter of the zero-truncated Poisson process) are modeled simultaneously. Choices of link functions, $g_1(\cdot)$ and $g_2(\cdot)$, are needed to connect the functional predictor $X_i(t)$ and the baseline covariates Z_{ri} to the Bernoulli probability (p_i) and Poisson rate (λ_i). More specifically,

$$\begin{aligned} g_1(p_i)&=\beta_0+\int\beta(t)X_i(t)dt+\sum_{r=1}^{p}\alpha_r Z_{ri}, \quad \text{and}\\ g_2(\lambda_i)&=\gamma_0+\int\gamma(t)X_i(t)dt+\sum_{r=1}^{p}\zeta_r Z_{ri}, \end{aligned} \quad (3)$$

where $g_1(p_i) = \log\{p_i/(1 - p_i)\}$ is the common logistic link function and and $g_2(\cdot) = \log(\lambda_i)$ the log-link function. The above formulation of the functional hurdle model (2–3) considers the same functional and baseline predictors in modeling the binary and zero-truncated Poisson processes for our application; however, the model can easily be extended to include different predictors for the two processes.

As introduced in Section 1, the key quantities of interest in (3), with respect to the functional predictor $X(t)$, are $\beta(t)$ and $\gamma(t)$. In the CDS data application, the regression coefficient function $\beta(t)$ in (3) captures variation during the time regions of the albumin trajectory after the initiation of dialysis that is associated with the likelihood of having zero vs. a positive number of hospitalizations. Similarly, $\gamma(t)$ highlights the time regions that contribute to the count (or rate) of hospitalizations. Also, note that the different coefficients, namely $\{\alpha_r\}$ and $\{\zeta_r\}$, allow flexibility for accommodating the fact that potentially different baseline predictors are associated with binary and zero-truncated Poisson processes in (3): for example, a prior history of congestive heart failure may be associated with having zero vs. a positive number of hospitalizations, but may not be associated with the zero-truncated Poisson count process for the subsequent count (or rate) of hospitalization.

In contrast to the hurdle model, the ZIP regression model assumes that the counts arise from mixing a standard Poisson process with a degenerate process at zero; thus, leading to observed excess zeros beyond the expected amount under a standard Poisson distribution alone. This mixture with a degenerate distribution at zero implies that a subset of observations originates from a subpopulation (or state) that can only have zero counts (known as "structural" zeros); the remaining zeros come from the Poisson subpopulation. The ZIP model was originally proposed by Lambert [12] for modeling counts of defective components in manufacturing processes, where structural zeros correspond to observations from a perfect manufacturing state that produces only perfect components. Other ZIP applications where subpopulations of structural zeros can be conceptualized include abstainers in drug abuse or sexual behavioral studies (e.g., [26]). ZIP regression models have been applied in a myriad of other applications, including abundance of rare species ([27], [28]), horticulture ([29]) and public health ([30], [31]) among others. With a functional predictor, analogous to the functional hurdle model (2–3), we propose a functional ZIP model given by

$$\Pr\{Y_i=y_i|X_i(t),Z_{ri}\}=\begin{cases} p_i e^{-\lambda_i}+(1-p_i), & \text{for } y_i=0, \\ p_i\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, & \text{for } y_i>0 \end{cases}, \quad (4)$$

where p_i and λ_i are related to the functional predictor $X_i(t)$ and baseline covariates Z_{ri} via suitable link functions, as given in (3). In contrast to the functional hurdle model, $\beta(t)$ in (3) for the functional ZIP model weighs the time regions of the functional predictor associated with the probability of arising from the Poisson subpopulation (p_i or the probability of the perfect state $1 - p_i$, depending on the model parametrization). $\gamma(t)$ weighs the time regions of $X(t)$ which influence the counts in the Poisson count subpopulation.

3 Estimation

We propose an estimation procedure called penalized reconstruction (Section 3.1) for generalized functional linear, hurdle and ZIP models. For comparison, we also describe a standard principal components regression (PCR) adaption for the proposed models. Both estimation procedures include a dimension reduction step, achieved by basis expansion, followed by penalized maximum likelihood or maximum likelihood estimation. The penalized reconstruction (PR) approach begins with a reconstruction of the functional predictor process, $X(t)$, based on sparse longitudinal data. Finally, the regression function, $\beta(t)$, is expanded on spline basis functions.

3.1 Penalized Reconstruction

Step 0: Reconstruction of the Predictor Trajectories from Sparse Longitudinal Data—The observed predictor trajectories $X_i(t)$ for $i = 1, \ldots, n$ subjects in model (3), are assumed to be square integrable realizations of the random smooth process X. Sparse data (e.g., infrequent and irregular data), as illustrated by the CDS data outlined in the Introduction section, is characterized by subject-specific random observation times $T_{ij} \in [0, T]$, for $j = 1, \ldots, N_i$ and a small total number of repeated measurements N_i. We also assume

additive measurement error on the functional predictor, i.e., $X_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$, where ε_{ij} are i.i.d. measurement errors with mean zero and finite variance.

Reconstruction of the predictor trajectories is based on the Karhunen-Loéve expansion for the observed process for subject i,

$$X_{ij}=\mu_X(T_{ij})+\sum_{v=1}^{\infty}\xi_{iv}\psi_v(T_{ij})+\varepsilon_{ij}, \quad (5)$$

where ξ_{iv} is the vth functional principal component score playing the role of random effects with $E(\xi_v) = 0$ and $\text{var}(\xi_v) = \rho_v$, and $\psi_v(t)$ is the vth eigenfunction. Quantities in (5), namely $\mu_X(t)$, ξ_{iv} and $\tau_v(t)$, will be obtained based on estimation of the moments of the underlying smooth random process X; specifically, the mean function $\mu_X(t)$ and auto-covariance function, denoted $G_{XX}(s, t)$. Moments are obtained via smoothing, which pools information from all subjects. The mean function estimate, $\hat{\mu_X}(t)$, is obtained by smoothing the aggregated data (T_{ij}, X_{ij}) for $i = 1, \ldots, n$, $j = 1, \ldots, N_i$ with local linear fitting. Next, the raw auto-covariances are computed as $G_{XX,i}(T_{ij}, T_{ij'}) = \{X_{ij} - \hat{\mu_X}(T_{ij})\}\{X_{ij'} - \hat{\mu_X}(T_{ij'})\}$. These raw estimates are fed into a two dimensional local least squares algorithm to obtain the final smooth estimates $\hat{G}_{XX}$. The effects of the additive measurement error can be eliminated by excluding the diagonal raw auto-covariance elements $G_{XX,i}(T_{ij}, T_{ij})$, $i = 1, \ldots, n$ and $k = 1, \ldots, N_i$ in the two-dimensional smoothing step. In addition, the non-negative definiteness of the estimated auto-covariance matrix can be guaranteed by excluding the negative estimates of the eigenvalues and corresponding eigenfunctions from the functional principal component decomposition of the auto-covariance operator, $\hat{G}_{XX}$. For details, the reader is referred to Şentürk and Müller [6]. For a computationally efficient bandwidth choice in the proposed one- and two-dimensional smoothing, we adopt the generalized cross-validation algorithm of Liu and Müller [32].

Once the moments are estimated, the eigenfunctions $\tau_v(t)$ are estimated through a functional principal component step applied to the discretization of the smooth auto-covariance estimator $\hat{G}_{XX}$. Subject-specific eigen-scores ξ_{iv} are recovered using Gaussian assumptions on all eigen-scores and measurement error, based on the conditional expectation $E(\xi_{iv}|U_i, N_i, T_i)$, where U_i is the $N_i \times 1$ observation vector $U_i \equiv (X_{i1}, \ldots, X_{iN_i})^T$ with $X_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$ and N_i and $T_i = (T_{i1}, \ldots, T_{iN_i})$ are the total number of repeated measurements and the vector of observation time points for subject i, respectively. (The reader is referred to [22], [33], [6] and [24] where the explicit expression of $\hat{\xi_{iv}}$ and its derivation is provided.) Based on (5), the functional predictor is reconstructed as $\tilde{X}_i(t_j) \equiv \hat{\mu}_X(t_j)+\sum_{v=1}^{K}\hat{\xi}_{iv}\hat{\psi}_v(t_j)$; we do this for an equidistant dense grid of time points t_j, $j = 1, \ldots, N$. The number K of eigen-components included can be chosen by various criteria; we utilize the fraction of variance explained, similar to Crainiceanu, Staicu and Di [5].

Step 1: Dimension Reduction via Basis Expansion—Next we consider the expansion of the coefficient function $\beta(t)$ for the functional predictor in a generalized functional linear model (1), functional hurdle model (3) or functional ZIP model using a set

of spline basis functions. More specifically, we consider $\beta(t) \approx \sum_{u=1}^{K_b} b_u \phi_u(t)$, where $\varphi_u(t)$ is taken to be truncated power series spline basis. Thus, $\beta(t) \approx b_1 + b_2 t + b_3 t^2 + \sum_{u=4}^{K_b} b_u (t-\kappa_u)_+^2$ with $\{\kappa_u\}_{u=4}^{K_b}$ representing the knots. The exact number of basis functions used K_b is not important as long as K_b is large. In applications we take $K_b = 20$. The position of the knots in the truncated power spline basis is also not particularly important and is taken at the quantiles of the distribution of t_j. Expansion for $\gamma(t)$ in a functional hurdle (3) or ZIP model follows similarly to the above expansion for $\beta(t)$. That is, $\gamma(t) \approx \sum_{\vartheta=1}^{K_a} a_\vartheta \theta_\vartheta(t)$, where $\theta_\vartheta(t)$ is also the truncated power series spline basis.

Using the above basis expansion, combined with the reconstructed predictor process $\tilde{X}_i(t)$,

$$\int X_i(t)\beta(t)dt \approx \int \tilde{X}_i(t)\beta(t)dt \approx \sum_{u=1}^{K_b} b_u \int \tilde{X}_i(t)\phi_u(t)dt = W_i b, \quad (6)$$

where $b = (b_1, \ldots, b_{K_b})^T$ and W_i is a $1 \times K_b$ vector with the uth entry equal to $\int \tilde{X}_i(t)\varphi_u(t)dt$. For sparse longitudinal data, direct estimation of $\int X_i(t)\varphi_u(t)dt$ is not feasible, since there is only a small number of total measurements over time available per subject. Hence, the proposed PR approach addresses this challenge by first reconstructing $X_i(t)$ on a dense grid of time points so that W_i can be feasibly estimated, via approximating the integral $\int \tilde{X}_i(t)\varphi_u(t)dt$. The expansion for $\gamma(t)$ proceeds similarly as for $\beta(t)$ above; thus, we have, analogous to (6) $\int \tilde{X}_i(t)\gamma(t)dt = W_i a$, where $a = (a_1, \ldots, a_{K_a})^T$ and now the uth entry of W_i is $\int \tilde{X}_i(t)\theta_u(t)dt$, $u = 1, \ldots, K_a$.

Using the expansion in (6), the generalized functional linear model (1) can be approximated by a generalized linear model:

$$g(\mu_i) = \beta_0 + \int \beta(t)X_i(t)dt + \sum_{r=1}^{p} \alpha_r Z_{ri} \approx \beta_0 + W_i b + \sum_{r=1}^{p} \alpha_r Z_{ri}. \quad (7)$$

That is, $g(\mu_i) \approx (1_n, W, Z)(\beta_0, b, a)^T$ with $n \times (1 + K_b + p)$ design matrix, $(1_n, W, Z)$, and parameter vector $(\beta_0, b, a)^T$, where $b = (b_1, \ldots, b_{K_b})^T$, $a = (a_1, \ldots, a_p)^T$, 1_n is a $n \times 1$ vector of ones, $n \times K_b$ matrix $W = (W_1, \ldots, W_n)^T$, $n \times p$ matrix of predictors $Z = (Z_1, \ldots, Z_p)$, and $Z_r = (Z_{r1}, \ldots, Z_{rn})^T$.

The functional hurdle and ZIP models for zero-inflated counts given in (3) and (4), respectively, can be similarly expanded and approximated by

$$g_1(p_i) \approx \beta_0 + W_i b + \sum_{r=1}^{p} \alpha_r Z_{ri}, \quad \text{and} \quad (8)$$

$$g_2(\lambda_i) \approx \gamma_0 + W_i a + \sum_{r=1}^{p} \zeta_r Z_{ri}. \quad (9)$$

The design matrix, $(1_n, W, Z)$, and parameter vector, $(\beta_0, b, a)^T$, in the binary part (8) are defined analogously as detailed above for the generalized functional linear model. Similarly, for the zero-truncated Poisson part (9), the design matrix is $(1_n, W, Z)$ and parameter vector is $(\gamma_0, a, \zeta)^T$, where $a = (a_1, \ldots, a_{K_a})^T$ and $\zeta = (\zeta_1, \ldots, \zeta_p)^T$.

Step 2: Penalized Maximum Likelihood—We propose to estimate the coefficient functions, $\beta(t)$ and $\gamma(t)$, using the induced generalized linear model, as well as the the induced hurdle and ZIP models, with estimated W_i by penalized maximum likelihood. A common and effective penalization approach used in functional data analysis to regularize smoothness of the regression functions is to penalize the second derivative of the regression functions. First, for the generalized linear model given in (1) and (7), we estimate the parameter vector (β_0, b, a) by maximizing the penalized log-likelihood

$$\ell_{\mathrm{GLM}}(\beta_0, b, \alpha, \hat{W}, Z) \equiv \ell(\beta_0, b, \alpha, \hat{W}, Z) - \delta \int \{\beta^{(2)}(t)\}^2 dt$$

where $\ell(\cdot)$ is an appropriately chosen log-likelihood corresponding to a generalized outcome (e.g. Bernoulli, Poisson etc. form an exponential family) and δ is the regularization parameter. Thus, using the basis expansion of $\beta(t)$ as described in step 1, we rewrite the second (penalty) term above as

$$\delta \int \{\beta^{(2)}(t)\}^2 dt = \delta \sum_{u=1}^{K_b} \sum_{u'=1}^{K_b} b_u b_{u'} \int \phi_u''(t) \phi_{u'}''(t) dt = \delta b' R b$$

where R is the matrix with $(u, u')th$ entry is equal to $\int \phi_u''(t) \phi_{u'}''(t) dt$. Hence, the penalized likelihood to be maximized is $\ell_{\mathrm{GLM}}(\beta_0, b, a, \hat{W}, Z) = \ell(\beta_0, b, a, \hat{W}, Z) - \delta b' R b$.

Next, we derive the penalized likelihood for the functional hurdle model described by (3), (8) and (9). For this, we estimate the parameter vectors (β_0, b, a) and (γ_0, a, ζ) by maximizing the penalized likelihood, denoted by

$$\ell_{\mathrm{Hurdle}}(\beta_0, b, \alpha, \gamma_0, a, \zeta, \hat{W}, Z) \equiv \Delta_{\mathrm{bin.}}(\beta_0, b, \alpha, \hat{W}, Z) + \Delta_{\mathrm{pos.}}(\gamma_0, a, \zeta, \hat{W}, Z)$$

where, similar to the generalized linear model case above,

$$\Delta_{\mathrm{bin.}}(\beta_0, b, \alpha, \hat{W}, Z) = \ell_{\mathrm{zero}}(\beta_0, b, \alpha, \hat{W}, Z) - \delta_1 b' R b, \quad \text{and}$$
$$\Delta_{\mathrm{pos.}}(\gamma_0, a, \zeta, \hat{W}, Z) = \ell_{\mathrm{pos.}}(\gamma_0, a, \zeta, \hat{W}, Z) - \delta_2 a' R a.$$

$\ell_{\mathrm{bin.}}(\beta_0, b, a, \hat{W}, Z)$ and $\ell_{\mathrm{pos.}}(\gamma_0, a, \zeta, \hat{W}, Z)$ correspond to the log-likelihood portions for the binary and zero-truncated Poisson processes, respectively, of the hurdle distribution (2) for a count response Y_i; which through direct calculations yield $\ell_{\mathrm{bin.}}(\beta_0, b, a, \hat{W}, Z) = \Sigma_{Y_i=0} \log\{1 - p_i(\beta_0, b, a, \hat{W}, Z)\} + \Sigma_{Y_i 0} \log\{p_i(\beta_0, b, a, \hat{W}, Z)\}$ and $\ell_{\mathrm{pos.}}(\gamma_0, a, \zeta, \hat{W}, Z) =$

$\Sigma_{Y_i0}[Y_i\log\{\lambda_i(\gamma_0, a, \zeta, \hat{W}, Z)\} - \lambda_i(\gamma_0, a, \zeta, \hat{W}, Z) - \log\{1-e^{-\lambda_i(\gamma_0,a,\zeta,\hat{W},Z)}\} - \log(Y_i!)]$. Note that because the likelihood $\ell_{\text{Hurdle}}(\cdot)$ factors into separate log-likelihood terms for (β_0, b, α) and (γ_0, a, ζ), namely $\Delta_{\text{bin.}}$ and $\Delta_{\text{pos.}}$, they can be efficiently maximized separately. Maximization of $\Delta_{\text{bin.}}$ represents fitting a penalized logistic regression to binary data $\{Y_i\mathbb{I}_{Y_i=0}, \mathbb{I}_{Y_i>0}\}_{i=1}^n$ and maximization of $\Delta_{\text{pos.}}$ represents fitting a penalized zero-truncated Poisson model to the positive counts $\{Y_i\mathbb{I}_{Y_i>0}\}_{i=1}^n$, where $\mathbb{I}_E$ denotes the indicator function for event A.

For fitting the functional ZIP model (4), the penalized functional ZIP log-likelihood is

$$\ell_{\text{ZIP}}(\beta_0, b, \alpha, \gamma_0, a, \zeta, \hat{W}, Z) \equiv \ell(\beta_0, b, \alpha, \gamma_0, a, \zeta, \hat{W}, Z) - \delta_1 b' Rb - \delta_2 a' Ra,$$

where

$$\begin{aligned}\ell(\beta_0, b, \alpha, \gamma_0, a, \zeta, \hat{W}, Z) = &\sum_{Y_i=0} \log[p_i(\beta_0, b, \alpha, \hat{W}, Z)e^{-\lambda_i(\gamma_0,a,\zeta,\hat{W},Z)} + \{1-p_i(\beta_0, b, \alpha, \hat{W}, Z)\}] \\ &+ \sum_{Y_i>0} [\log\{p_i(\beta_0, b, \alpha, \hat{W}, Z)\} - \lambda_i(\gamma_0, a, \zeta, \hat{W}, Z) \\ &+ Y_i\log\{\lambda_i(\gamma_0, a, \zeta, \hat{W}, Z)\} - \log(Y_i!)].\end{aligned}$$

The functional ZIP (log−) likelihood, $\ell_{\text{ZIP}}(\cdot)$, does not separate and needs to be maximized jointly to simultaneously estimate all model parameters, $(\beta_0, b, \alpha, \gamma_0, a, \zeta)$. We note that for simplicity of notation, the functional ZIP model is represented using the same model parameters notation as those used in the functional hurdle model above.

The regularization parameter d in the functional generalized linear model is chosen by multi-fold cross-validation (CV). The CV error is normalized with the variance of the prediction:

$$CV(\delta) = \sum_j \sum_{i=1}^{n_j} \left\{ \frac{Y_i - \hat{\mu}_i^{-j}}{\sqrt{V(\hat{\mu}_i^{-j})}} \right\}^2 \quad (10)$$

where $\hat{\mu}_i^{-j}$ denotes the mean predicted value for the ith subject in left-out group j. The regularization parameters δ_1 and δ_2 in the functional hurdle model are chosen by two separate one-dimensional grid searches with $\hat{\mu_i} = e^{\hat{\eta i}}/(1 + e^{\hat{\eta i}})$, $V(\hat{\mu_i}) = \hat{\mu_i}(1 - \hat{\mu_i})$, and $\hat{\eta}_i = \hat{\beta}_0 + \hat{W}_i\hat{b} + \sum_{r=1}^p \hat{\alpha}_r Z_{ri}$ for the logistic part; similarly, for the Poisson part $\hat{\mu_i} = \hat{\lambda_i}/(1 - e^{-\hat{\lambda_i}})$, $\hat{\lambda_i} = e^{\hat{\eta i}}$, $V(\hat{\mu}_i) = (\hat{\lambda}_i + \hat{\lambda}_i^2)/(1-e^{-\hat{\lambda}_i}) - \{\hat{\lambda}_i/(1-e^{-\hat{\lambda}_i})\}^2$, and $\hat{\eta}_i = \hat{\gamma}_0 + \hat{W}_i\hat{a} + \sum_{r=1}^p \hat{\zeta}_r Z_{ri}$ These separate regularizations for the functional hurdle model components are applicable since the likelihood, $\ell_{\text{Hurdle}}(\cdot)$, factors. However, for the functional ZIP model, a two dimensional grid search is needed for the choice of δ_1 and δ_2 where $\hat{\mu_i} = \hat{p_i}\hat{\lambda_i}$, $V(\hat{\mu}_i) = \hat{p}_i(\hat{\lambda}_i + \hat{\lambda}_i^2) - (\hat{p}_i\hat{\lambda}_i)^2$, $\hat{p_i} = e^{\hat{\eta 1 i}}/(1 + e^{\hat{\eta 1 i}})$, $\hat{\lambda_i} = e^{\hat{\eta 2 i}}$, $\hat{\eta}_{1i} = \hat{\beta}_0 + \hat{W}_i\hat{b} + \sum_{r=1}^p \hat{\alpha}_r Z_{ri}$ and $\hat{\eta}_{2i} = \hat{\gamma}_0 + \hat{W}_i\hat{a} + \sum_{r=1}^p \hat{\zeta}_r Z_{ri}$. Recall that W_i is estimated directly via approximation of the

integral $\int \tilde{X}_i(t)\varphi_u(t)dt$, where $\tilde{X}_i(t)$ for cross-validation is reconstructed based on the population mean and eigenfunctions estimated with subjects in the *j*th group excluded.

3.2 Standard Principal Components Regression

Step 1: Dimension Reduction via Basis Expansions—For subsequent comparative simulation studies with the proposed method PR, we also adapt the standard principal components regression (PCR). Briefly, PCR expands both the longitudinal predictor $X(t)$ and the functional coefficient $\beta(t)$ on a few PC functions of $X(t)$, denoted by $\psi_v(t)$:

$$X_i(t) \approx \mu_X(t) + \sum_{v=1}^{K} \xi_{iv}\psi_v(t), \quad \beta(t) \approx \sum_{v=1}^{K} b_v\psi_v(t), \quad (11)$$

where $\xi_{iv} = \int\{X_i(t) - \mu_X(t)\}\psi_v(t)dt$ and the number of PCs, K, is typically small (e.g., two to three components). In applications, we utilize the fraction of variance explained in choosing K. For the functional hurdle or ZIP models $\gamma(t) \approx \sum_{v=1}^{K} a_v\psi_v(t)$. Using (11),

$$\int X_i(t)\beta(t)dt \approx \sum_{v=1}^{K} b_v \left[\sum_{v=1}^{K}\{\xi_{iv} + \int \mu_X(t)\psi_v(t)dt\}\right] = (\xi_i^{\mathrm{T}} + \tilde{\mu})b \quad (12)$$

where $\xi_i = (\xi_{i1}, \ldots, \xi_{iK})^{\mathrm{T}}$, $b = (b_1, \ldots, b_K)^{\mathrm{T}}$ and $\tilde{\mu}$ is a $1 \times K$ vector with the vth entry equal to $\int \mu_X(t)\psi_v(t)dt$. Eigenscores ξ_i and $\tilde{\mu}$ are estimated based on the functional PCs expansions outlined in Section 3.1 step 0. Using similar basis expansions as in (11) for the coefficient function $\gamma(t)$, we have $\int X_i(t)\gamma(t)dt \approx (\xi_i^{\mathrm{T}} + \tilde{\mu})a$. The generalized functional linear model reduces to $g(\mu_i) = \beta_0 + \int X_i(t)\beta(t)dt + \sum_{r=1}^{p}\alpha_r Z_{ri} \approx \beta_0 + (\xi_i^{\mathrm{T}} + \tilde{\mu})b + \sum_{r=1}^{p}\alpha_r Z_{ri}$ and the functional hurdle model reduces to $g_1(p_i) \approx \beta_0 + (\xi_i^{\mathrm{T}} + \tilde{\mu})b + \sum_{r=1}^{p}\alpha_r Z_{ri}$ and $g_2(\lambda_i) \approx \gamma_0 + (\xi_i^{\mathrm{T}} + \tilde{\mu})a + \sum_{r=1}^{p}\zeta_r Z_{ri}$.

Step 2: Maximum Likelihood—As the number of functional PCs basis functions used is small, PCR uses maximum likelihood without penalization, where the likelihoods for the corresponding models are as outlined in Section 3.1 step 2.

4 Simulation Studies

4.1 Simulation Design

We carry out three simulation studies to evaluate the performance of the proposed estimation algorithm, PR, and also to compare its performance to the more conventional approach of PCR for 1) generalized functional linear models, 2) functional hurdle and 3) ZIP models under sparse and denser longitudinal designs. Moderate and larger sample sizes of $n = 200$ and $n = 400$ are used. In all three studies, estimation of the moments of the predictor processes including the bivariate smoothing procedures and the choice of appropriate bandwidths in the functional principal components decompositions are carried out with the publicly available software package PACE (http://anson.ucdavis.edu/~ntyang/PACE; [22], [33], [4]). The number of functional principal components basis functions used in the

expansions for the PCR and the reconstruction step of PR are chosen by the fraction of variance explained. The number of components are selected that explain at least 90% and 85% of the variation in the longitudinal predictor for sparse and denser longitudinal designs, respectively, similar to [20]. This criterion typically selects one to three functional principal components basis functions, where the most common value selected is two.

In all three simulation studies, the covariate process X is generated according to $X_i(t)=\mu_X(t)+\sum_{v=1}^{10}\xi_{iv}\psi_v(t)$, where the functional principal component scores ξ_{iv} are simulated from independent normals with means zero and variances equal to $5/v^2$, $\mu_X(t) = 2\sin(\pi t/2)$, $\psi_v(t)=\sqrt{2}\sin(\pi vt)$ for $0 \le t \le 1$. The predictor trajectories are assumed to be observed with measurement error. They are simulated independently from a Gaussian distribution with zero mean and variance equal to 0.7. The cross-sectional covariate Z_i is simulated from a Gaussian distribution with zero mean and variance equal to 2. The number of repeated measurements for $n = 200$ and 400 subjects are chosen randomly from [1, 15] and [15, 30] with equal probabilities for sparse and denser designs, respectively. The observation times T_{ij} for each subject are randomly selected for the longitudinal covariate from the time interval [0, 1]. Reported results for all simulations (Section 4.2 below) are based on 200 Monte-Carlo runs. Further details of the three simulation cases are as follows.

Generalized functional linear model—The regression parameters are $\beta_0(t) = 1$, $\beta(t) = -t^2$ and $\alpha = 0.8$. The response variable Y_i are simulated from a Bernoulli distribution with mean $p_i = E\{Y_i|X_i(t), Z_i\} = g^{-1}\{\beta_0 + \int\beta(t)X_i(t) + \alpha Z_i\}$, where $g^{-1}(\eta_i) = e^{\eta_i}/(1 + e^{\eta_i})$.

Functional hurdle model—The regression parameters are $\beta_0(t) = -1$, $\gamma_0 = 1$, $\beta(t) = t$, $\gamma(t) = t^2/2$, $\alpha = 0.5$ and $\zeta = 0.5$. The functional hurdle model is simulated from

$$p\{Y_i=y_i|X_i(t),Z_i\}=\begin{cases}1-p_i & \text{for } y_i=0,\\ \frac{p_i}{1-e^{-\lambda_i}}\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right) & \text{for } y_i>0\end{cases}$$

where $g_1(p_i) = \beta_0 + \int\beta(t)X_i(t)dt + \alpha Z_i$ and $g_2(\lambda_i) = \gamma_0 + \int\gamma(t)X_i(t)dt + \zeta Z_i$ with a logistic link function $g_1(\cdot)$ and a log-link $g_2(\cdot)$.

Functional ZIP model—The regression parameters are taken to be $\beta_0(t) = -1$, $\gamma_0 = 1$, $\beta(t) = t$, $\gamma(t) = t^2/2$, $\alpha = 0.5$ and $\zeta = 0.5$. The functional ZIP model is simulated from

$$p(Y_i=y_i|X_i(t),Z_i)=\begin{cases}p_ie^{-\lambda_i}+(1-p_i) & \text{for } y_i=0,\\ p_i\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right) & \text{for } y_i>0,\end{cases}$$

where p_i and λ_i are related to the predictors in the same way as described for the functional hurdle model.

Preliminary simulation studies were carried out to select the optimal regularization parameters for PR at each of the three simulation settings. Results from the main simulation

studies reported in Section 4.2 use the independently selected regularization values from the preliminary simulation studies and are summarized in the Appendix. To study the performance of the proposed estimation method for the regression parameters and the regression function, we use relative mean squared deviation error (ME):

$$\mathrm{ME}_{\beta}=\frac{\int\{\beta(t)-\hat{\beta}(t)\}^2 dt}{\int\beta^2(t)dt}, \quad \mathrm{ME}_{\beta_0}=\frac{(\beta_0-\hat{\beta}_0)^2}{\beta_0^2} \quad \text{and} \quad \mathrm{ME}_{\alpha}=\frac{(\alpha-\hat{\alpha})^2}{\alpha^2}.$$

For functional hurdle and ZIP models, ME_{γ}, ME_{γ_0} and ME_{ζ} are defined similarly.

4.2 Simulation Results

The cross-sectional medians and the 5% and 95% cross-sectional percentiles of the estimated regression functions $\beta(t)$ and $\gamma(t)$ from PR are given in Figure 1 for the functional hurdle model at $n = 200$ sparse design. The medians from PCR are also given. For PR, the median regression function estimates track the true regression functions. The (median) PCR estimates deviate more from the true regression functions relative to the proposed estimates.

The performance of the two methods with respect to the relative mean squared deviation error (ME) are summarized in more details in Tables 1, 2 and 3 from the generalized functional linear model, functional hurdle and functional ZIP models, respectively. More specifically, the median, 25% and 75% percentiles of ME are provided for PR and PCR over all three simulation cases (generalized functional linear model, functional hurdle model and functional ZIP model). There are several conclusions that can be drawn. First, the proposed PR method provides considerable efficiency gains over the standard PCR in all simulations and consistently for each sample size and design setting, especially for the regression functions $\beta(t)$ and $\gamma(t)$ compared to other constant model parameters β_0, a, γ_0 and ζ. The gain of the proposed approach over PCR is most substantial in estimation of the binary part of the functional ZIP model; ME for the constant model parameters are large for PCR. Poor relative performance of PCR is due to its sensitivity to the number of principal components selected, since it does not include any regularization, unlike PR. As Goldsmith et al. [20] also point out, the few principal components used in applications of the PCR method may be a poor basis choice for the regression functions if they do not lie in the space spanned by the relatively few principal components. This is also displayed in Figure 1 where the linear and quadratic underlying regression functions are not represented well by the lower dimensional estimated sinusoidal principal components functions. As expected ME becomes smaller for denser designs and for larger sample sizes.

5 Application to Data from the Comprehensive Dialysis Study

The Comprehensive Dialysis Study [7] is a prospective cohort study of end-stage renal disease patients where longitudinal serum samples within the first two years of dialysis were collected on 266 patients who newly initiated dialysis in the US between 2005 and 2007. To illustrate the proposed methods, we model hospitalization counts within the first two years of dialysis as a function of serum albumin concentrations, which is a sparsely sampled longitudinal predictor process. Baseline covariates of interest include age at the initiation of

dialysis, gender, body mass index and comorbidities (diabetes, peripheral vascular disease and congestive heart failure). Hospitalization data on CDS study participants were obtained from United States Renal Data System (USRDS). For our modeling, we used serum albumin concentrations and the number of hospitalizations between 100 and 550 days after the initiation of dialysis, since the minimum time to first serum collection was approximately 3.4 months and most serum measurements were collected between [100, 550] days after the initiation of dialysis. Hence the analysis cohort consists of 228 patients, where the number of longitudinal serum albumin concentration measurements range from 1 to 5 per subject and the number of hospitalizations range from 0 to 16 with 53% zeros. We choose to focus on the functional hurdle model for the CDS data since all patients are conceptually at some positive risk of hospitalization and it is difficult to conceptualize a subpopulation of dialysis patients from a “perfect state” who have no chance of being hospitalized. We use the proposed functional hurdle model to jointly model the binary process of having zero vs. a positive number of hospitalizations and the number/count of hospitalizations for those patients who have at least one hospitalization.

Longitudinal albumin trajectories are given in Figure 2 (c), where there seems to be no substantive trend in time for the mean albumin concentration. Estimated regression coefficients from the functional hurdle model PR fit are given in Table 4 and estimated regression functions from both the binary and zero-truncated Poisson parts are displayed in Figure 3, along with ±2 bootstrap error bands. One functional principal component explaining 95.3% of the variation is selected in the reconstruction step. The regularization parameters δ_1 and δ_2 for the binary and zero-truncated Poisson parts are chosen to be 1.5 and .5, respectively, by multi-fold cross-validation. Bootstrap error bands are reported based on 200 bootstrap samples generated via resampling from subjects with replacement and using the same regularization parameters as selected in the original data set.

Among the cross-sectional covariates, baseline age at the initiation of dialysis is the only variable with an all positive error band for the binary process of having zero vs. a positive number of hospitalization(s). Older patients at the initiation of dialysis have a trending association with a higher probability of having a positive count of hospitalizations. Error bands for the regression function for effects of longitudinal albumin concentrations have sections that are away from zero for the binary process, but not on the zero-truncated Poisson process of the positive number of hospitalization counts. We note that although the CDS data provide a clear conceptual illustration of proposed methods, the sample size in the CDS study is a limitation for especially highlighting associations with multiple comorbidities, where the zero-truncated Poisson part of the hurdle model is fitted with even a smaller sample size (about half of the original sample size) since it considers only those patients with a positive hospitalization count. Nevertheless, the functional hurdle model still highlights trending associations with the binary part of the model. These associations are not detected with a Poisson functional linear model fit to the data (results are omitted). Figure 2 (a) and (b) display the distributions of the linear predictor (η_i) from the binary part of the functional hurdle model fit and average albumin concentration levels respectively, for patients with (front) and without (back) a positive count of hospitalizations. The proposed functional hurdle model provides a mild separation for the two cohorts of patients with and

without hospitalizations, where lower average albumin concentrations are associated with a higher probability of having at least one hospitalization. This is consistent with prior studies associating lower serum albumin levels with a decline in glomerular filtration rate, a marker for advancing renal failure ([34]).

Results from the functional hurdle model are also consistent with cross-sectional model fits to the data, summarized in Table 5, where a cross-sectional hurdle model is used to regress the number of hospitalizations on albumin levels averaged throughout the entire study period along with baseline age, gender, body mass index and comorbidities. Age and average albumin levels are found significant for the binary process but not in the zero-truncated Poisson part supporting the results from the functional hurdle model. Congestive heart failure is also found to be significantly associated with a higher probability of having at least one hospitalization in the cross-sectional hurdle model. Even though the functional and cross-sectional hurdle models agree on the major trends in the data, the functional hurdle model provides further insights to the relative associations of albumin concentrations from different time periods after initiation of dialysis with the probability of having at least one hospitalization. Note that even with the large bootstrap error bands, mainly due to small sample size, the decreasing trend in the estimated regression function $\beta(t)$ in the binary part of the functional hurdle model (Figure 3 (a)) suggests stronger associations with albumin concentrations when patients have been on dialysis for a longer period of time and a higher probability of having at least one hospitalization within two years of dialysis initiation. This illustrates an important feature of the functional hurdle or ZIP models; they offer additional insights over their cross-sectional counterparts through the regression functions that highlight time regions in the support of the predictor process that are associated with the binary and count parts of the models.

Finally note that in the current analysis, the hospitalization count (response) and albumin concentrations (functional predictor) are obtained over the same time domain. This is because most hospitalizations occur during the first 2 years after the initiation of dialysis and in our application the functional predictor is also recorded over this same time period. Hence regression relations should be strictly interpreted as an association. However, our proposed model may be applied to clinical data where the predictor domain (time period) precedes the time period where the counts (outcome) are measured. For example other studies may consider the association between functional markers/predictors of hospitalization (whether using serum albumin, a marker of kidney function or C-reactive protein, a marker of inflammation etc.) in the year prior to dialysis/ESRD with hospitalization counts within two years after initiation of dialysis in Chronic Kidney Disease (CKD) stage 3 and 4 patients. Such results would be informative to explore clinical functional predictors/markers during CKD management that potentially may be associated with future reduced hospitalization risk (e.g., during the first 1–2 years of dialysis).

6 Discussion

Functional hurdle and ZIP models are proposed as extensions of generalized functional linear models for modeling mixture distributions of counts with excess zeros. The proposed models retain interpretations of a functional linear model through the functional regression

coefficients over their cross-sectional hurdle and ZIP counterparts, while accommodating a mixture distribution for the response with excess zeros. The proposed estimation approach PR, as well as the extension of the standard PCR approach considered are all designed to handle sparse longitudinal data observed with measurement error. This is due to the functional principal components decompositions utilized in both algorithms that pool information across all subjects in addressing sparsity issues and use smoothing to adjust for measurement error ([6], [24], [22], [33]). While functional principal components analysis provides a parsimonious framework in the analysis of a single functional process, in regression contexts PCR may not provide adequate representation of the regression functions of interest especially when they do not lie in the lower dimensional space spanned by the few principal components selected in applications. Instead, the proposed approach that uses larger numbers of basis functions coupled with penalization leads to substantial improvements in efficiency. These current findings are consistent with prior discussion in the literature in the context of generalized functional linear models ([20]).

Note that we also explored a second estimation approach, called dual penalized expansion (DPE) which relies on similar ideas as used in the approach by Goldsmith et al. [20] of decoupling the basis expansions of the predictor process and the regression coefficient function; that is, separate basis functions, namely, functional principal components basis and spline basis, are used in the expansions, respectively. Since DPE yielded very similar finite sample properties to the proposed approach, we now defer the development of DPE to our online supporting information file. Supporting information documentation also contains comparisons between finite sample performance of PR, PCR and DPE via simulations.

Goldsmith et al. ([35]) point to two separate sources of uncertainty in analysis of functional trajectories, the model-based uncertainty and uncertainty in the functional principal components decomposition. The bootstrap confidence intervals account for the uncertainty in the functional principal components decomposition but do not account for the model-based uncertainty. Goldsmith et al. ([35]) are able to combine these two sources of uncertainty using iterated expectation and variance formulas in the context of forming prediction/confidence intervals for subject-specific curves based on a single functional process. Extensions to build valid inference procedures for the mean or regression functions in a functional regression model and the proposed functional ZIP and hurdle models and studying their finite sample properties are open problems.

Another important direction for future research is functional regression modeling with follow-up data truncated by death. Truncation by death is common in studies on geriatric populations; the loss of person-days due to truncation by death is about 5.03% of the total follow-up time for all patients in the CDS. Prior studies proposed careful analysis of targets of inference for longitudinal data truncated by death in the context of generalized linear models ([36], [37]) and generalized varying coefficient models ([38]); extensions of these work to functional regression is an important open problem. Finally our current work develops in great detail models for a single longitudinal predictor. Extensions to multiple longitudinal predictors would utilize similar ideas of dimension reduction and penalization. For identifiability issues in cases of higher dimensional longitudinal predictors, we refer readers to the recent work of Scheipl and Greven [39].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank two referees and the Associate Editor for their constructive comments that improved the manuscript. This publication was made possible by grants R01 DK092232 and K23 DK093584 from the National Institute of Diabetes and Digestive and Kidney Diseases and by grant UL1 TR000153 from the National Center for Advancing Translational Sciences. The interpretation and reporting of the data presented here are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the United States government.

References

1. Ramsay, J.; Silverman, B. Functional Data Analysis. New York: Springer; 2005.
2. James G. Generalized linear models with functional predictors. Journal of the Royal Statistical Society, Ser. B. 2002; 64:411–432.
3. James G, Silverman B. Functional adaptive model estimation. Journal of the American Statistical Association. 2005; 100:565–576.
4. Müller H. Functional modeling and classification of longitudinal data. Scandinavian Journal of Statistics. 2005; 32:223–240.
5. Crainiceanu C, Staicu A, Di C. Generalized multilevel functional regression. Journal of the American Statistical Association. 2009; 104:1550–1561. [PubMed: 20625442]
6. Şentürk D, Müller HG. Functional varying coefficient models for longitudinal data. Journal of the American Statistical Association. 2010; 105:1256–1264.
7. Kutner NG, Johansen KL, Kaysen GA, Pederson S, Chen SC, Agodoa LY, Eggers PW, Chertow GM. The comprehensive dialysis study (CDS): a USRDS special study. Clinical Journal of the American Society of Nephrology. 2009; 4:645–650. [PubMed: 19261814]
8. Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986; 33:341–365.
9. King G. Event count models for international relations: Generalizations and applications. International Studies Quarterly. 1989; 33:123–147.
10. Heilbron, D. SIMS Technical Report 9. Department of Epidemiology and Biostatistics, University of California; San Francisco: 1989. Generalized linear models for altered zero probabilities and overdispersion in count data.
11. Heilbron D. Zero-altered and other regression models for count data with added zeros. Biometrical Journal. 1994; 36:531–547.
12. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34:1–14.
13. United States Renal Data System. USRDS 2012 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; Bethesda, MD: 2012.
14. Yuan M, Cai T. A reproducing kernel Hilbert space approach to functional linear regression. The Annals of Statistics. 2010; 38:3412–3444.
15. James G, Wang J, Zhu J. Functional linear regression thats interpretable. The Annals of Statistics. 2009; 37:2083–2108.
16. Reiss P, Ogden R. Functional principal component regression and functional partial least squares. Journal of the American Statistical Association. 2007; 102:984–996.
17. Cardot H, Sarda P. Estimation in generalized linear model for functional data via penalized likelihood. Journal of Multivariate Analysis. 2005; 92:24–41.
18. Long Q. A note on generalized functional linear model and its application. Journal of Statistical Planning and Inference. 2012; 142:2599–2606. [PubMed: 22711973]
19. Müller H, Stadtmüller U. Generalized functional linear models. The Annals of Statistics. 2005; 33:774–805.

20. Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. Penalized functional regression. Journal of Computational and Graphical Statistics. 2011; 20:830–851. [PubMed: 22368438]
21. Yao F, Müller HG, Clifford AJ, Dueker SR, Follett J, Lin Y, Buchholz B, Vogel JS. Shrinkage estimation for functional principal component scores, with application to the population kinetics of plasma folate. Biometrics. 2003; 59:676–685. [PubMed: 14601769]
22. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association. 2005a; 100:577–590.
23. James G, Hastie TJ, Sugar CA. Principal component models for sparse functional data. Biometrika. 2000; 87:587–602.
24. Şentürk D, Nguyen DV. Varying coefficient models for sparse noise-contaminated longitudinal data. Statistica Sinica. 2011; 21:1831–1856.
25. Şentürk D, Dalrymple LS, Mohammed SM, Kaysen GA, Nguyen DV. Modeling time-varying effects with generalized and unsynchronized longitudinal data. Statistics in Medicine. 2013; 32:2971–2987. [PubMed: 23335196]
26. Buu A, Johnson NJ, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. Statistics in Medicine. 2010; 30:2326–2340. [PubMed: 21563207]
27. Welsh A, Cunningham R, Donnelly C, Lindenmayer D. Modeling the abundance of rare species statistical-models for counts with extra zeros. Ecological Modelling. 1996; 88:297–308.
28. Chiogna M, Gaetan C. Semiparametric zero-inflated Poisson models with application to animal abundance studies. Environmetrics. 2007; 18:303–314.
29. Hall DB. Zero-inflated Poisson and Binomial regression with random effects: A case study. Biometrics. 2000; 56:1030–1039. [PubMed: 11129458]
30. Yau KKW, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Statistics in Medicine. 2001; 2:2907–2920. [PubMed: 11568948]
31. Lam KF, Xue H, Cheung YB. Semiparametric analysis of zero-inflated count data. Biometrics. 2006; 62:996–1003. [PubMed: 17156273]
32. Liu, B.; Müller, HG. Functional data analysis for sparse auction data. Wiley & Sons; New York: 2008. p. 269-290.
33. Yao F, Müller HG, Wang JL. Functional linear regression analysis for longitudinal data. Annals of Statistics. 2005b; 3:2873–2903.
34. Kaysen GA. Serum albumin concentration in dialysis patients: why does it remain resistant to therapy? Kidney Int Suppl. 2003; 87:92–98.
35. Goldsmith J, Gereven S, Crainiceanu C. Corrected confidence bands for functional data using principal components. Biometrics. 2012; 69:41–51. [PubMed: 23003003]
36. Kurland BF, Heagerty PJ. Directly parameterized regression conditioning on being alive: Analysis of longitudinal data truncated by deaths. Biostatistics. 2005; 6:241–258. [PubMed: 15772103]
37. Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. Statistical Science. 2009; 24:211–222. [PubMed: 20119502]
38. Estes JP, Nguyen DV, Dalrymple LS, Mu Y, Şentürk D. Cardiovascular event risk dynamics over time in older patients on dialysis: A generalized multiple-index varying coefficient model approach. Biometrics. 2014 in press.
39. Scheipl, F.; Greven, S. Technical report 125. Vol. 2012. Department of Statistics, University of Munich; 2012. Identifiability in penalized function-on-function regression models.

Appendix

Independent (separate) preliminary simulation studies were carried out to select the optimal regularization parameters for PR used for the main simulation studies of Section 4. Median d values minimizing CV error (10) across multiple runs for the generalized functional linear

model fits is 0.01 for both the sparse and denser designs at $n = 200$ and $n = 400$. The selected (δ_1, δ_2) regularization pairs for the functional hurdle model fits in the sparse design are (.1, .05) and (.05, .05), and they are (.075, .05) and (.05, .05) for the denser design at $n = 200$ and $n = 400$, respectively. For the functional ZIP model, the values in the sparse and denser designs are both (.1, .05) for $n = 200$; they are (.1, .075) and (.1, .05) for $n = 400$.

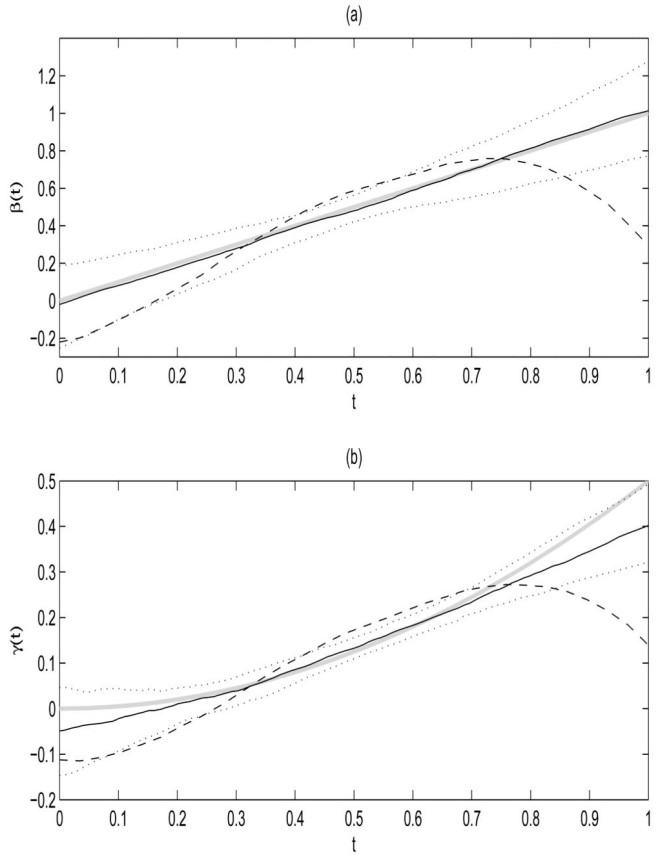

Figure 1.

The cross-sectional medians (thick gray) and the 5% and 95% cross-sectional percentiles (dotted) of the estimated regression functions (a) $\beta(t)$ and (b) $\gamma(t)$ (solid) from PR over 200 simulation runs for the functional hurdle model at $n = 200$ sparse design. Medians from PCR (dashed) are also given.

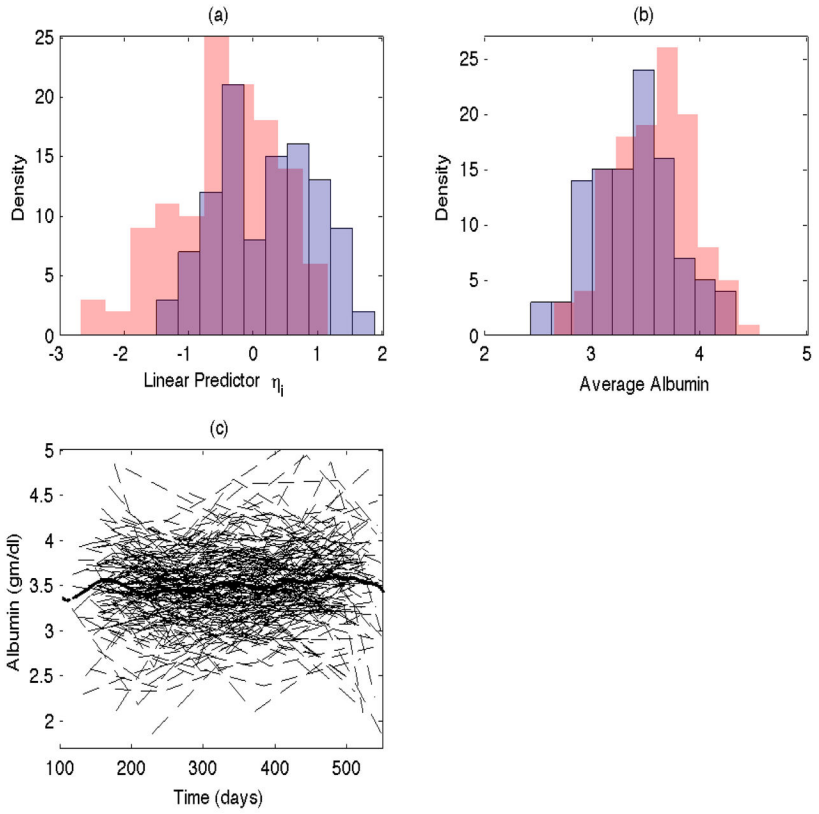

Figure 2.

(a) Histogram of the linear predictor η_i for patients who had no hospitalization (back in pink) and for patients who had a positive count of hospitalizations (front in purple), (b) histogram of average albumin level in the observation period for patients who had no hospitalization (back in pink) and for patients who had a positive count of hospitalizations (front in purple), (c) observed individual trajectories (dashed) and the smoothed estimate of the cross-sectional mean functions (thick solid) for longitudinal albumin concentrations.

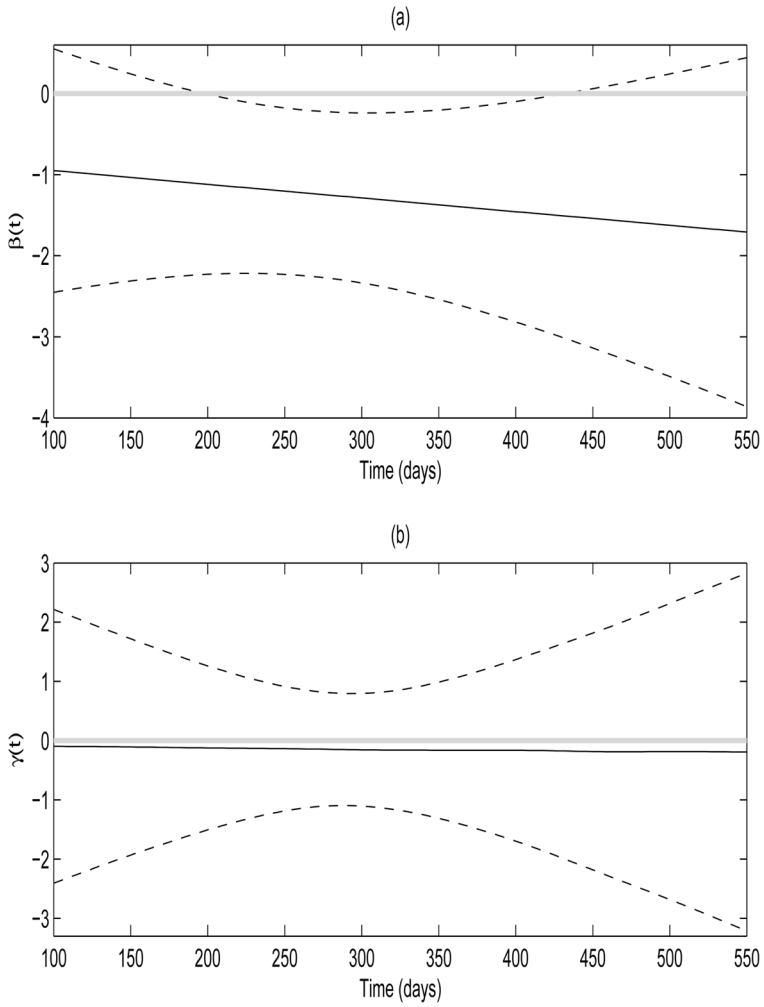

Figure 3.

Estimated regression functions from the (a) binary and (b) zero-truncated Poisson parts of the functional hurdle model displaying the effects of albumin concentration on hospitalization counts. ±2 bootstrap error bands are given dashed, while a horizontal line at zero is given in gray for ease of interpretation.

Table 1

Simulation results for generalized functional linear model

Median, mean and squared deviation error (ME) reported for all model parameters for different estimation techniques in fitting a generalized functional linear model over different sparsity levels of the longitudinal predictor and at different sample sizes in 200 Monte Carlo runs.

Design	n	Median	25%	75%	Median	25%	75%	Median	25%	75%
			ME_{β}			ME_{β_0}			ME_{α}	
Sparse										
PR	200	.253	.105	.481	.041	.008	.113	.017	.003	.047
PCR	200	.448	.325	.650	.044	.010	.116	.018	.003	.046
PR	400	.116	.061	.222	.018	.003	.063	.008	.003	.017
PCR	400	.330	.266	.415	.024	.004	.078	.007	.003	.018
Denser										
PR	200	.182	.090	.371	.034	.009	.091	.015	.004	.039
PCR	200	.372	.300	.541	.043	.008	.127	.015	.004	.041
PR	400	.110	.051	.215	.017	.003	.054	.006	.001	.019
PCR	400	.299	.259	.383	.024	.007	.070	.006	.001	.019

PR: penalized reconstruction; PCR: principal components regression.

Table 2

Simulation results for functional hurdle model

Median, mean, squared deviation error (ME) reported for all model parameters for different estimation techniques in fitting a functional hurdle model over different sparsity levels of the longitudinal predictor and at different sample sizes in 200 Monte Carlo runs.

Design	n	Median	25%	75%	Median	25%	75%	Median	25%	75%
			ME_{β}			ME_{β_0}			ME_{α}	
Binary										
Sparse										
PR	200	.085	.032	.256	.044	.018	.119	.023	.006	.091
PCR	200	.243	.174	.386	.068	.015	.149	.024	.006	.090
PR	400	.051	.022	.109	.022	.005	.049	.017	.005	.044
PCR	400	.195	.159	.246	.029	.004	.073	.018	.004	.045
Denser										
PR	200	.087	.031	.190	.038	.012	.098	.030	.007	.093
PCR	200	.230	.171	.351	.057	.012	.143	.031	.007	.092
PR	400	.045	.019	.112	.017	.004	.052	.017	.005	.039
PCR	400	.185	.157	.243	.027	.006	.073	.017	.004	.038
			ME_{γ}			ME_{γ_0}			ME_{ζ}	
Zero-truncated Poisson										
Sparse										
PR	200	.088	.049	.210	.007	.002	.019	.002	<.001	.005
PCR	200	.308	.251	.419	.017	.004	.047	.003	.001	.009
PR	400	.077	.033	.174	.004	.001	.011	.001	<.001	.002
PCR	400	.273	.243	.326	.012	.005	.025	.001	<.001	.003
Denser										
PR	200	.063	.026	.128	.004	.001	.012	.001	<.001	.004
PCR	200	.282	.242	.378	.011	.003	.031	.002	<.001	.006

Design	n	Median	25%	75%	Median	25%	75%	Median	25%	75%
			ME_{β}			ME_{β_0}			ME_{α}	
PR	400	.056	.025	.138	.003	.001	.008	.001	<.001	.002
PCR	400	.269	.243	.312	.007	.001	.020	.001	<.001	.003

PR: penalized reconstruction; PCR: principal components regression.

Table 3

Simulation results for functional ZIP model

Median, mean, squared deviation error (ME) reported for all model parameters for different estimation techniques in fitting a functional ZIP model over different sparsity levels of the longitudinal predictor and at different sample sizes in 200 Monte Carlo runs.

Design	n	Median	25%	75%	Median	25%	75%	Median	25%	75%
			ME_{β}			ME_{β_0}			ME_{α}	
Binary										
Sparse										
PR	200	.113	.049	.247	.038	.010	.130	.039	.009	.108
PCR	200	1.11	.566	2.30	.788	.164	1.27	.850	.243	1.78
PR	400	.060	.023	.136	.021	.005	.078	.020	.005	.063
PCR	400	1.24	.684	1.99	1.01	.708	1.39	1.16	.645	1.78
Denser										
PR	200	.093	.045	.176	.043	.009	.156	.033	.009	.098
PCR	200	1.07	.545	2.21	.823	.206	1.196	.984	.279	1.93
PR	400	.047	.020	.091	.024	.006	.068	.021	.006	.053
PCR	400	1.14	.720	1.69	.926	.383	1.22	1.07	.470	2.06
			ME_{γ}			$ME_{\gamma 0}$			ME_{ζ}	
Poisson										
Sparse										
PR	200	.078	.038	.137	.005	.001	.017	.001	<.001	.004
PCR	200	.461	.303	.857	.081	.016	.200	.020	.005	.049
PR	400	.065	.026	.107	.006	.002	.013	.001	<.001	.002
PCR	400	.387	.284	.679	.118	.045	.241	.028	.009	.055
Denser										
PR	200	.057	.027	.124	.004	.001	.012	.001	<.001	.004
PCR	200	.395	.282	.735	.085	.013	.207	.020	.003	.050

Design	n	Median	25%	75%	Median	25%	75%	Median	25%	75%
			ME_{β}			ME_{β_0}			ME_{α}	
PR	400	.048	.020	.083	.002	<.001	.006	<.001	< .001	.002
PCR	400	.361	.276	.689	.113	.032	.226	.026	.006	.049

PR: penalized reconstruction; PCR: principal components regression.

Table 4

Functional hurdle model results for the baseline covariates

Estimated coefficients for the baseline variables from the functional hurdle model fits using longitudinal albumin concentrations within two years of initiation of dialysis. Lower and upper bounds are reported from ±2 bootstrap error bands.

	Binary			Zero-truncated Poisson		
Variable	**Estimate**	**Lower B.**	**Upper B.**	**Estimate**	**Lower B.**	**Upper B.**
Intercept	1.526	−2.659	5.711	4.797	0.780	8.815
Age	0.032	0.007	0.057	−0.015	−0.035	0.005
Body mass index	0.002	−0.045	0.049	0.002	−0.029	0.033
Diabetes	0.204	−0.498	0.906	0.060	−0.429	0.549
Peripheral vascular disease	0.263	−0.507	1.033	0.444	−0.043	0.931
Congestive heart failure	0.695	−0.006	1.396	0.241	−0.205	0.687
Gender	−0.542	−1.142	0.058	0.395	−0.048	0.838

Table 5

Cross-sectional hurdle model results

Fitted values from a cross-sectional hurdle model for number of hospitalizations along with p-values for their significance. Model fits contain the average albumin values throughout the study (Avg-alb).

	Binary		Zero-truncated Poisson	
Variable	**Estimate**	**P-value**	**Estimate**	**P-value**
Intercept	1.866	0.291	1.385	0.010
Age	0.029	0.012	−0.002	0.636
Body mass index	−0.004	0.829	0.001	0.860
Diabetes	0.198	0.541	−0.069	0.489
Peripheral vascular disease	0.260	0.431	−0.079	0.458
Congestive heart failure	0.720	0.034	−0.162	0.146
Gender	−0.509	0.088	−0.022	0.809
Avg-alb	−1.087	0.007	−0.063	0.594