Dynamic Prediction of Non-Guassian Outcome with fast Generalized Functional Principal Analysis

Ying Jin Andrew Leroux

March 17, 2023

Introduction

Biomedical investigators are often interested in predicting future observations of subjects based on their historical data, typically referred to as dynamic prediction. Traditionally, this type of data has been modeled using marginal models (generalized estimating equations) or conditional models (mixed effect models) (Laird and Ware 1982; LIANG and ZEGER 1986; Lindstrom and Bates 1990; "Nonlinear models for repeated measurement data" 2003), and predictions are made based on the correlation between repeated measures from the same subject, and/or covariates that can be either fixed or time-varying. However, these methods are limited in terms of flexibility of correlation structure and the ability to handle out-of-sample prediction. When sample size is large or the density of repeated measures is high, they also tend to cause severe computational burden (Rizopoulos 2022).

To address these problems, one may turn to functional mixed effect models instead when measures are dense across the domain. Such methods accommodate more flexible correlation structure by modeling subject-specific random effects as a function, and non-parametric smoothing (Scheipl et al. 2014) can be incorporated to speed up the computation, such as spline basis functions or eigenfunctions from functional principal component analysis (fPCA). The introduction of basis functions also makes out-of-sample prediction more straightforward. Instead of estimating subject-specific random effects of new observations, we can simply estimate coefficients/loadings on the basis function used for smoothing. Existing research on dynamic prediction with functional data analysis methods has been focusing on continuous/Gaussian outcomes, modelling subject-specific random effects with FPCA (Chiou 2012; Goldberg et al. 2014; Shang 2017). Kraus (2015) has used this approach to predict missing observations in partially observed function tracks, and Delaigle and Hall (2016) achieved similar goals using Markov Chains. While methods mentioned above used only partial observations for prediction with an intercept-only model, Leroux et al. (2018) proposed Functional Concurrent Regression (FCR) framework which can incorporate the effect of subject-specific predictors. However, little extension was made on prediction of non-Gaussian functions, such as binary and count outcomes.

Unfortunately, fewer papers have focused on its extension to non-Gaussian data, such as series of binary or count outcomes. Existing methods also tend to be very computationally intensive. For example, Chen et al. (2013) proposed approaches to fit marginal functional models that is compatible to multi-level, generalized outcomes. Goldsmith et al. (2015) established a model framework that takes into account the fixed effect of time-invariant covariates, with parameters estimated with Bayesian method in *Stan*. Gertheiss et al. (2016) identified bias introduced by directly applying FPCA methods to generalized functions, and proposed to address this problem using a two-stage, joint estimation strategy. Linde (2009) used an adapted Bayesian variational algorithm for FPCA of binary and count data. In terms of implementation, Wrobel et al. (2019) proposed a fast, efficient way to fit GFPCA on binary data using EM algorithm, accompanied by the an open source R package *registr*.

In this paper, we aim to develop a fast, scalable method for dynamic prediction of discrete function tracks based on functional mixed effect model with fPCA smoothing. Section 2 presents the procedure of the proposed method. In Section 3, we illustrate the performance and efficiency of our proposed method in a simulation study. In Section 4, we apply this method to a real-world dataset. Section 5 presents a discussion of davantages and limitation of the proposed method.

Method

The observed data for a single subject i is (t, $Y_i(t)$), where t consists of dense, discrete points along the functional domain, and $Y_i(t)$ is the non-Gaussian outcome observed at t. We assume that the outcome $Y_i(t)$ can be characterized by a latent continuous function $\eta_i(t)$. That is, at a specific t, $Y_i(t)$ follows a exponential family distribution such that:

$$g[E(Y_i(t))] = \eta_i(t) = \beta_0(t) + b_i(t)$$

where g is a appropriate link function, $\beta_0(t)$ is the population mean of latent function, and $b_i(t)$ is a subjects-level random effect function follows a zero-mean Gaussian process.

While $b_i(t)$ is not observed, it can be approximated under the FPCA framework $b_i(t) = \sum_{k=1}^{K} \xi_{ik}\phi_{ik}(t) + \epsilon_i(t)$. Here $\phi_{ik}(t)$ is a set of orthogonal eigenfunctions that explains the most variation in $b_i(t)$, and ξ_{ik} are subject-specific PC scores (or loadings) on each eigenfunction. Additionally, ξ_{ik} are mutually independent obver both subject (i) and eigenfunctions (k). That is, each ξ_{ik} follows normal distribution $N(0, \lambda_k)$ where λ_k is the kth eigenvalues: $\int \phi_k^2(t)dt = \lambda_k$. $\epsilon_i(t)$ here is a residual function that accounts for the variation not explained by the first K eigenfunctions from FPCA. We assume it follows a zero-mean Gaussian process. At a specific point t, $\epsilon_i(t) \sim N(0, \sigma^2)$.

Based on the problem set up above, we propose the following algorithm for PFCA on the unobserved latent process $\eta_i(t)$:

1. Bin the observed outcomes in to small, non-overlapping, equal length intervals. We hereafter index the bins by their midpoints s.
2. Fit a local Generalized Mixed Model at every bin. Specifically, at bin s, we fit $g[E(Y_i(t))] = \beta_0(s) + b_i(s)$ for all t in bin s. From this series of models we can get estiamtes of population mean $\hat{\beta}_0(s)$ and subject-level random effect $\hat{b}_i(s)$, thus estimate of the individual latent functions at every bin: $\hat{\eta}_i(s) = \hat{\beta}_0(s) + \hat{b}_i(s)$.
3. Fit FPCA on the estimated latent functions $\hat{\eta}_i(s)$, and obtain estimates of basis functions $\mathbf{\Phi} = \{\phi_1(s), ..., \phi_k(s)\}$, eigenvalues $\hat{\lambda}_1 ... \hat{\lambda}_k$, population mean $\hat{\beta}_0(s)$ and residual variance $\hat{\sigma}^2$.

4. With components extracted above, calculate the maximum likelihood estimate (MLE) of the subject-specific PC scores $\hat{\xi}_{ik}$ of new samples based on their partially observed data. Then the value of latent functions at unobserved points can be estimated as $\hat{\eta}_i(s) = \hat{\beta}_0(s) + \sum_{k=1}^{K} \hat{\xi}_{ik}\phi_k(s)$

Following the algorithm above, predictions of individual latent functions are made on the binned grid based on partially observed non-Gaussian functions tracks of new subjects. Since the bins are set up to be small in length, the binned grid would still be dense enough. However, in situations where we need predictions on the original, un-binned grid instead, linear interpolation turns out to be a fast, convenient way with good performance for prediction at points between bins.

Precision of prediction, usually quantified by the variance of prediction error $Var(\hat{\eta}_i(s) - \eta_i(s))$, is also straightforward under this framework. In step 4 we calculated the MLE of $\hat{\xi}_{ik}$. Based on likelihood theory, its variance can be estimated with observed information $I(\hat{\xi}_{ik})$, which is essentially the second derivative of likelihood at $\hat{\xi}_{ik}$. Therefore, the variation of prediction interval is:

$$Var(\hat{\eta}_i(s) - \eta_i(s)) = \mathbf{\Phi}(s)I(\hat{\boldsymbol{\xi}}_i)\mathbf{\Phi}^T(s) + \hat{\sigma_\epsilon}^2$$

Where $\mathbf{\Phi}(s) = (\phi_1(s)...\phi_K(s))$ and $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i1}, ..., \hat{\xi}_{iK})$.

Simulation

In this section, we illustrate the predictive performance and computational efficiency of the proposed method through a simulation study. We simulated 50 datasets, each with 500 subjects. For every subject, we generate 1000 binary outcomes $Y_i(t) \in (0, 1)$ across functional domain $t \in [0, 1]$, where the distribution of outcome is characterized by a continuous latent function. The data generation mechanism can be expressed as follows:

$$\begin{aligned}
&Y_i(t) \sim Binomial(\frac{exp(\eta_i(t))}{1 + exp(\eta_i(t))}) \\
&\eta_i(t) = f_0(t) + \xi_{i1}sin(2\pi t) + \xi_{i2}cos(2\pi t) + \xi_{i3}sin(4\pi t) + \xi_{i4}cos(4\pi t)
\end{aligned}$$

In this simulation, we set $f_0(t) = 0$. ξ_{ik} are mutually independent normal random variables $\xi_{ik} \sim N(0, \gamma_k)$. Here we set the values of γ_k to be 0.5^{k-1}, $k \in (1, ..., 4)$. In addition, for simplicity of presentation, we generate data on a regular grid, which means observations points are equally distributed across $[0, 1]$ and are the same for all subjects.

We use two metrics to evaluate the out-of-sample predictive performance: integrated squared error (ISE) and Area-Under-the-Receiver-Operator-Curve (AUC). ISE assess the prediction accucay of latent continuous function. It is evaluated on the binned grid at midpoints of each unobserved bin. If the entire functional domian has S bins, but we have observations up to the mth bin, then ISE is defined as $\frac{1}{N}\sum_{i=1}^{N}\sum_{s=m+1}^{S}(\hat{\eta}_i(s) - \eta_i(s))^2$. The second metric, AUC, focuses on evaluation of prediction of the binary outcome. Since the binary outcomes are generated on the original, un-binned grid, we evaluated AUC on this grid as well and estimated values of latent functions between bins with linear interpolation.

As a reference method, we compare our method to Generalized Linear Mixed Models using Adaptive Gaussian Quadrature (GLMMadaptive). This is one of the fastest existing method developed for dynamic prediction of repeated generalized outcomes. Just like many mixed models, this method is very limited in terms of flexibility. For example, the model used for prediction of our simulated datset would simply be an linear model with one covariate indicating observation time: $g(E(Y_i)) = \beta_0 + \beta_1 t + b_{i0} + b_{i1} t$. While the flexibility of this mixed model can be increased using spline functions, the dimension of spline basis is also restricted by computational ability, and is unfeasible to implement under the scale of our simulated data or the complexity of our proposed method.

The average ISE and AUC across all simulation is presented in Table 1, and Figure 1 presents the predicted function curves of four randomly-drawn subjects from the first simulation. The prediction is conditioning on different length of observed track (specifically, with observations up to t = 200, 400, 600, 800 respectively), and evaluation is made on equal-length time windows on the unobserved tracks following the maximum observation time. As Table 1 reveals, the fGFPCA outperformed GLMMadaptive under every scenario, which was expected since GLMMadaptive can only accomodate simple model structures. In addition, fGFPCA also takes less computation time. For one simulated dataset, fGFPCA spent 4.94 minutes on model fitting and out-of-sample prediction, while GLMMadaptive took 5.11 minutes.

Because of the flexibility of our proposed model framework, the accuracy of prediction at specific time points would improve with more observed data. This is revealed in Table 1 as ISE decreases and AUC increases with maximum observed time (left-to-right), also in Figure 1 as predicted curves get closer with longer observed track. However the same tendency is not observed with GLMMadaptive models, as a result of restricted model flexibility. A linear model that fits well to a specific part of latent function can fit very badly to the following parts, especially when the underlying latent function has cyclic patterns. Therefore, more observations do not necessarily make the model more predictive.

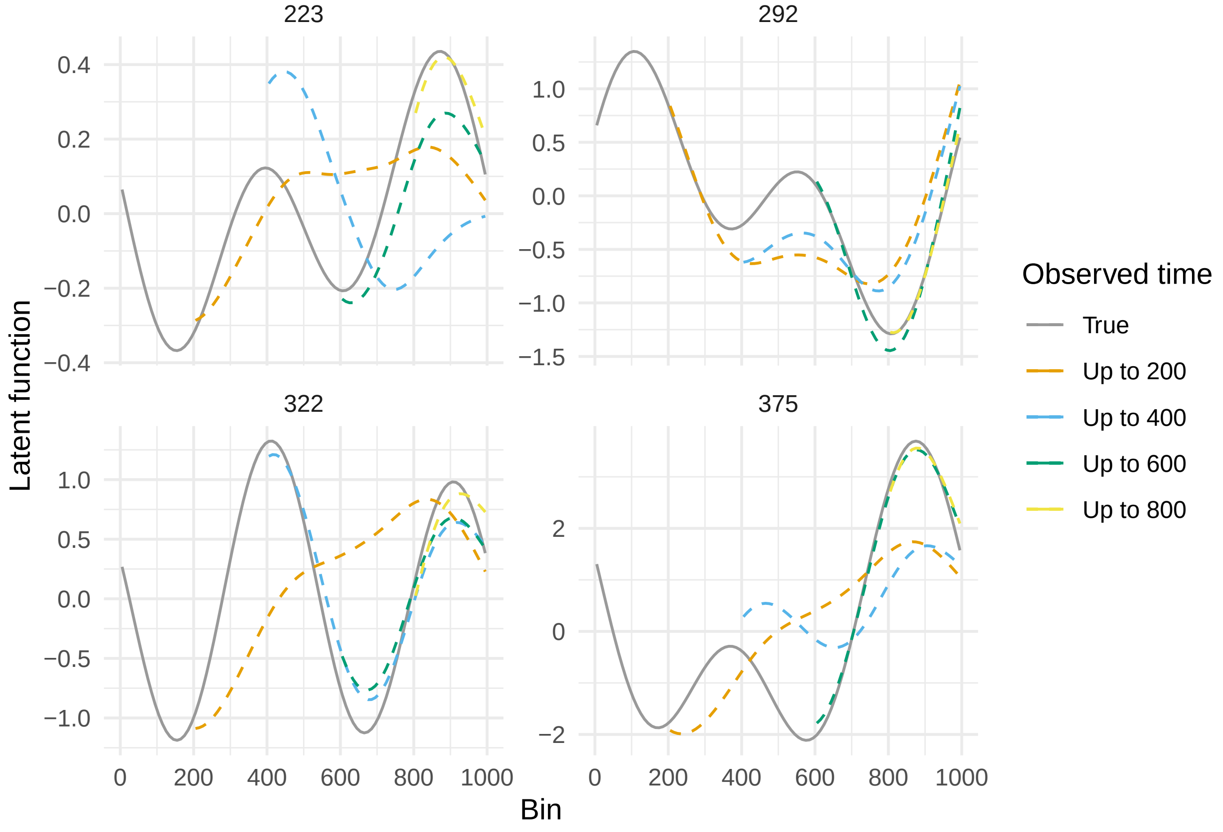

Figure 1: Predicted track of four randomly selected subjects. The grey dash line indicates the true latent continuous function. The dashed lines indicated predicted latent function tracks, and color indicates different observations time.

Table 1: Predictive performance of fGFPCA and GLMM adaptive on the simulated datasets. ISE and AUC are average values across all 50 simulations.

	Maximum observed time							
	fGFPCA				GLMMadaptive			
	200	400	600	800	200	400	600	800
Prediction time window								
ISE								
(200, 400]	15.29				38.20			
(400, 600]	18.68	8.53			28.59	26.46		
(600, 800]	22.53	5.51	1.65		31.32	27.97	27.40	
(800, 1000]	11.34	8.72	1.90	1.30	55.19	46.89	58.57	59.25
AUC								
(200, 400]	0.74				0.59			
(400, 600]	0.66	0.73			0.52	0.59		
(600, 800]	0.71	0.79	0.80		0.67	0.70	0.69	
(800, 1000]	0.74	0.75	0.78	0.78	0.52	0.56	0.53	0.57

Data application

Discussion

- Grid
- Score bias: cannot demonstrate without repeat simulation

References

Chen, H., Wang, Y., Paik, M. cho, and Choi, H. A. (2013), "A marginal approach to reduced-rank penalized spline smoothing with application to multilevel functional data," *J Am Stat Assoc.*, 108, 1216–1229. https://doi.org/10.1080/01621459.2013.826134.

Chiou, J.-M. (2012), "Dynamical functional prediction and classification, with application to traffic flow prediction," *The Annals of Applied Statistics*, Institute of Mathematical Statistics, 6, 1588–1614. https://doi.org/10.1214/12-AOAS595.

Delaigle, A., and Hall, P. (2016), "Approximating fragmented functional data by segments of markov chains," *Biometrika*, 103, 779–799. https://doi.org/10.1093/biomet/asw040.

Gertheiss, J., Goldsmith, J., and Staicu, A. (2016), "A note on modeling sparse exponential-family functional response curves," *Comput Stat Data Anal*, 105, 46–52. https://doi.org/10.1016/j.csda.2016.07.010.

Goldberg, Y., Ritov, Y., and Mandelbaum, A. (2014), "Predicting the continuation of a function with applications to call center data," *Journal of Statistical Planning and Inference*, 147, 53–65. https://doi.org/https://doi.org/10.1016/j.jspi.2013.11.006.

Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015), "Generalized multilevel function-on-scalar regression and principal component analysis," *Biometrics*, 71, 344–53. https://doi.org/10.1111/biom.12278.

Hall, P., Müller, H.-G., and Yao, F. (2008), "Modelling sparse generalized longitudinal observations with latent gaussian processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 703–723. https://doi.org/https://doi.org/10.1111/j.1467-9868.2008.00656.x.

Kraus, D. (2015), "Components and completion of partially observed functional data," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, [Royal Statistical Society, Wiley], 77, 777–801.

Laird, N. M., and Ware, J. H. (1982), "Random-effects models for longitudinal data," *Biometrics*, [Wiley, International Biometric Society], 38, 963–974.

Leroux, A., Crainiceanu, C. M., and Wrobel, J. (n.d.). "Fast generalized functional principal component analysis."

Leroux, A., Xiao, L., Crainiceanu, C., and Checkley, W. (2018), "Dynamic prediction in functional concurrent regression with an application to child growth," *Statistics in medicine*, 37, 1376–1388.

LIANG, K.-Y., and ZEGER, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22. https://doi.org/10.1093/biomet/73.1.13.

Linde, van der (2009), "A bayesian latent variable approach to functional principal components analysi with binary and count data," *A StA Adv Stat Anal*, 307–333. https://doi.org/10.1007/s10182-009-0113-6.

Lindstrom, M. J., and Bates, D. M. (1990), "Nonlinear mixed effects models for repeated measures data," *Biometrics*, [Wiley, International Biometric Society], 46, 673–687.

"Nonlinear models for repeated measurement data: An overview and update" (2003), [International Biometric Society, Springer], 8, 387–419.

Rizopoulos, D. (2022), *GLMMadaptive: Generalized linear mixed models using adaptive gaussian quadrature*.

Scheipl, F., Staicu, A.-M., and Greven, S. (2014), "Functional additive mixed models," *J Comput Graph Stat*, 24, 447–501. https://doi.org/10.1080/10618600.2014.901914.

Shang, H. L. (2017), "Functional time series forecasting with dynamic updating: An application to intraday particulate matter concentration," *Econometrics and Statistics*, 1, 184–200. https://doi.org/https://doi.org/10.1016/j.ecosta.2016.08.004.

Suresh, K., Taylor, J. M. G., Spratt, D. E., Daignault, S., and Tsodikov, A. (2017), "Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model," *Biom J*, 59, 1277–1300. https://doi.org/10.1002/bimj.201600235.

Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019), "Registration for exponential family functional data," *Biometrics*, 75, 48–57. https://doi.org/https://doi.org/10.1111/biom.12963.