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

Ying Jin Andrew Leroux

March 07, 2023

Abstract:

Biomedical investigators are often interested in predicting future observations of subjects based on their historical data, referred to as dynamic prediction. Traditional methods are often limited in flexibility and computationally intensive, especially with non-Gaussian data. To address these issues, we propose a novel method for dynamic prediction based on Generalized Functional Principal Component Analysis (FPCA). Assume the observed outcome follows an exponential family distribution parameterized by a latent Gaussian function, the proposed method consists of the following steps: 1) Bin the data across functional domain into small, equal-length intervals; 2) Fit local generalized mixed models at every bin to estimate individual latent functions; 3) Fit FPCA model to smooth latent functions and 4) Obtain estimates of subject-specific PC scores using partial observations and recover the unobserved part on the binned grid. Our simulation study showed the proposed method achieved significantly better out-of-sample predictive performance compared to existing methods with much shorter computation time, thus has the potential to be widely applicable to large datasets.

Introduction

- Overview of dynamic prediction methods

Prediction of repeated measures has been a problem of interest in the biomedical field. Typically, such 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. Traditionally, repeated measures have 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; Rizopoulos 2022). These methods, while allowing for correlation between repeated measures, are limited in terms of flexibility of correlation structure and the ability to handle out-of-sample prediction. Therefore, one may turn to functional mixed effect models when measures are dense across the domain. Such methods accommodates more flexible correlation structure by modeling subject-specific random effects as a function, but often cause nontrivial computational burden. A feasible approach to

address this issue is non-parametric smoothing (Scheipl et al. 2014), 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. In this project, we will focus on the prediction of non-Gaussian outcomes (e.g. binary or count) from a random-intercept only model, with no covariates considered. In other words, we aim to propose a new fast, scalable method for dynamic prediction of discrete function tracks based only on past observations using functional mixed effect model with fPCA smoothing.

- Dynamic prediction with functional methods

Research on dynamic prediction of functional outcomes 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.

- fPCA and GFPCA (Leroux et al. n.d.)

Unlike FPCA on Gaussian data, 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 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*.

Method

- Need better notation system
- Change to general exponential family
- Add section of de-bias score: conditional on subject. Need repeat simulation to demonstrate

Result

- Repeat simulation: repeat 10-100 times

- Different set-up:

a. Different eigenfunctions: with or without periodicity (start with the current one)
b. Outcome: binary or count
c. Sample size: start with N=500
d. Grid density: start with J=1000
e. Bin width, overlap or not: start with non-overlap, bin width = 10 (100 bins)

- Real 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.