Statistics in Interdisciplinary Research: Methods, Tools and Inspirations

Ying Jin

Deparment of Biostatistics and Informatics
University of Colorado Anschutz Medical Campus

June 26th 2024

Introduction

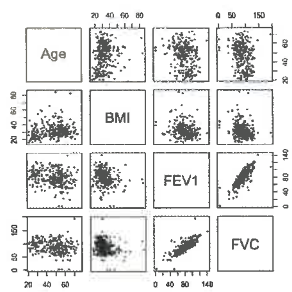
- Experience
  - PhD candidate in Biostatistics, CU Anschutz
    - Research Assistant in the Center of Design of Innovative Analysis (CIDA) and the Pulmonary Translational Core (PTraC)
    - Research Assistant on R01 grant NS060910: *Statistical methods for longitudinal multivariate neuroimaging biomarkers*
  - Master of Statistics, Columbia University
    - Research Assistant at the Division on Substance Use Disorders
- Research interest:
  - Developing tools to facilitate interdisciplinary communication
  - Methodological development inspired by collaborative research
    - Predictive modeling
    - Time-to-even outcomes
    - Repeated measures across time/space

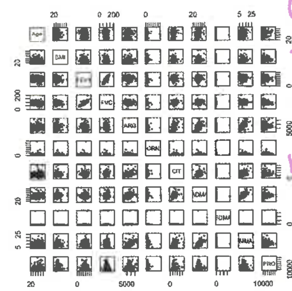
Context-Driven Interactive Visualization with VisX: an innovative tool to facilitate interdisciplinary communication

Background

I like this framing - it helps understand why your research matters

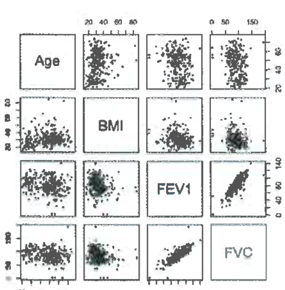
- Interpretation of the relationship between variables can be challenging
  - A large number of variables, i.e., multidimensional data

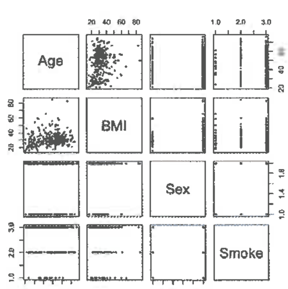




Background

- Interpretation of the relationship between variables can be challenging
  - Different types of variables, e.g. continuous, categorical, ordinal...





Method: VisX

Is this short for something?

- Challenge: too many variables
- Method: spatial encoding
  - Represent variables as points on a 2D surface
  - Represent correlation/association as distance between points
  - Strongly correlated variables are clustered together
- CMD scale (Classical Multidimensional Scaling) (Gower, 1966)
  - Observed covariate $\boldsymbol{X} = \{x_{ip}\}, i = 1...N, p = 1...P$
  - Develop a **dissimilarity measure** between covariates: d_{lp}
    e.g. Euclidean distance: $d_{lp}^2 = \sum_{i=1}^{N}(x_{il} - x_{ip})^2$
  - Dissimilarity matrix: $\boldsymbol{D} = \{d_{lp}\}, l, p = 1...P$
  - Project $\boldsymbol{D}$ onto a 2-D space. (e.g. PCA)

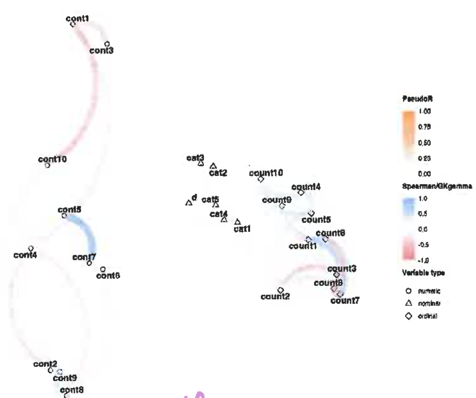
Method: VisX

- Challenge: different types of variables
- Method: a "comprehensive" dissimilarity matrix:
  - Type of association d_{lp} depends on the types of x_l and x_p
    - Continuous variables: pearson/spearman correlation
    - Categorical variables: Psuedo R^2
    - Ordinal variables: rank-based measures
  - d_{lp} needs to be on the same scale

Method: VisX

- Challenge: too much information
- Method: a "comprehensive" visualization scheme
  - Shape of points - variable type
  - Color scale of edge - association type
  - Color, transparency and thickness of edge - strength of association
  - Let's make it an interactive Shiny App!

Results: Simulated Data



are

- Variables of the same type is clustered together
- Correlation exist between categorical and count variables

Results: Obesity and Asthma study

A rich, multi-institutional dataset (Holguin et al., 2013)

- Demographics and medical history
- L-arginine metabolites: L-arginine, ADMA, cirtrulline, ornthine and proline
- Polyamines: putrisicne, spermine and spermidine
- Lung function: FEV1 and FVC
- Asthma: Asthma Control Score (ACQ) and Asthma Quality Of Life Questionnaires Score (AQLQ)
- Healthcare utilization: use of emergency room (ER) and/or steroid

Results: Obesity and Asthma study

some other points/clusters seem closer together. why are these 3 most important?

maybe animations that point to these clusters?

asthma?

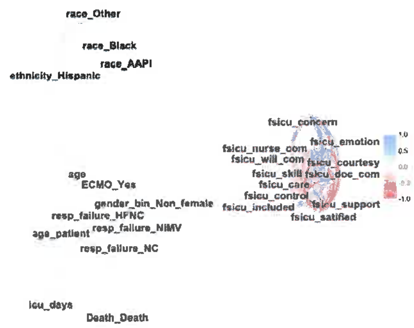
- Variables are clustered by group
- Lung function is highly correlated with age, sex and smoking status
- BMI is moderately correlated with L-arginines
- Smoking status is moderately correlated with polyamines

Result: Family satisfaction with intensive care unit

Multi-site dataset of: (Amass et al., 2022)

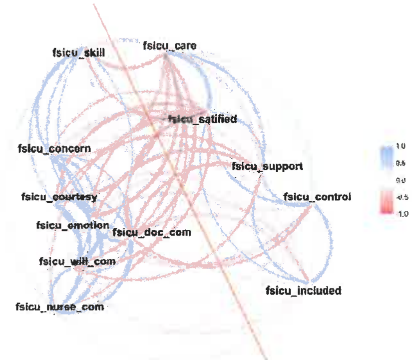
- Family satisitaction with ICU (FS-ICU)
  - 12 questions covering various aspects of ICU service
  - Score from 1 to 5
- Demographic information and the medical history

Result: Family satisfaction with intensive care unit



- All the FS-ICU questions are clustered together
- They do not seem to be corerlated with other variables

Result: Family satisfaction with intensive care unit



- Two clear clusters negatively correlated with each other
- And idea why? any

Discussion

VisX can be used to...

- Visualize mixed-type multidimensional dataset in a comprehensive way
- Facilitate detection and interpretation of data structure
- Engage domain experts in the analysis procedure
- For large datasets, real-time feedback can be slow
- Always looking for interesting datasets!

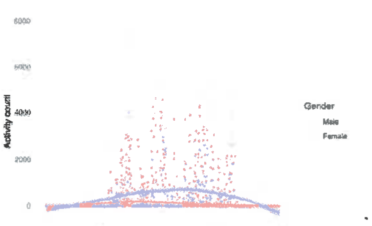
is this future directions?

is this a limitation?

Dynamic Prediction of Generalized Functional Data: Inspiration from Minute-by-Minute Activity Indicator

Background

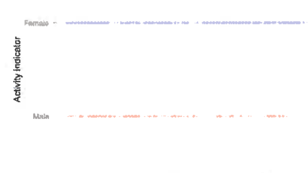
- Technology development has made the collection and storage of **dense** repeated measures possible
  - Accelerometer data
  - Daily weigh-in
  - Pixel intensity and its derivatives



- Functional data
  - High density
  - Complex underlying pattern

Background

- Technology development has made the collection and storage of **dense** repeated measures possible
  - Accelerometer data
  - Daily weigh-in
  - Pixel intensity and its derivatives



- Generalized functional data
  - Functional data with discrete value
  - Often preferred for better interpretation

Background

- The availability of such datasets has motivated healthcare practitioners to ask new questions:
  - The effect of "shape" on health outcomes
  - Integration of data from different devices
  - Prediction of future development based on historical records
- Inspired by these questions, functional data analysis (FDA) was born

was this an existing method? => reference
or one you created?

Functional Data Analysis

- Unit of observation:
  - A series of measurements $Y(t_j)$
  - Collected over a dense grid $j = 1 \dots J$
  - Along the study domain $t \in T$
- Theory framework
  - Conceptualize $Y(t_j)$ as discrete realization of a function $Y(t)$
  - Assume $Y(t)$ can be characterized by a continuous latent function $\eta(t)$
    - ⋆ Continuous: $E(Y(t)) = \eta(t)$
    - ⋆ Generalized: $g(E(Y(t))) = \eta(t)$

FDA in Dynamic Prediction

of?

- Predicting the future development based on historical record
  - Activity pattern
  - Child growth
  - Location/shape of lesions
- Desirable features
  - Highly individualized
  - Temporal updates
- Challenges
  - Dimensionality and complexity
  - Out-of-sample prediction

FDA in Dynamic Prediction

- Challenge: dimensionality and complexity
  - Mixed model with unstructured correlation
  - $\frac{J(J-1)}{2}$ correlation coefficient to estimate
- Method: Generalized Functional Principal Component Analysis (GFPCA)
  - Functional extension of PCA

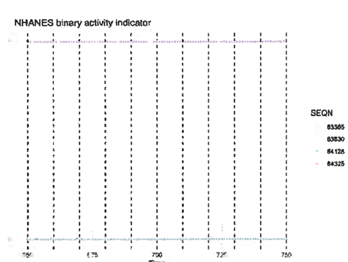
$$g(E(Y_i(t))) = \eta_i(t) = f_0(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t)$$

  - ξ_{ik} are mutually independent scores/loadings. $\xi_{ik} \sim N(0, \lambda_k)$
  - We only need to estimate $K(J+2)$ parameters!

FDA in Dynamic Prediction

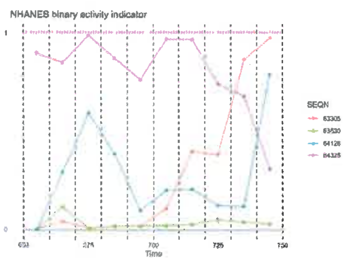
- Challenge: slow implementation for very large datasets
- Method: fast implementation of GFPCA (fGFPCA)
  - Fast implementation of FPCA exists for **Continuous** outcomes (e.g., FACE by Xiao et al. (2016))
  - Estimate $\eta(t)$, but on a slightly sparser grid

FDA in Dynamic Prediction



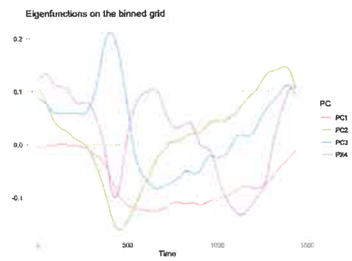
- Bin the observed outcomes in to small, non-overlapping, equal length bins

FDA in Dynamic Prediction



- Bin the observed outcomes in to small, non-overlapping, equal length bins.
- Fit a local, intercept-only generalized linear mixed model at every bin

FDA in dynamic Prediction



- Fit FPCA on the estimated latent functions $\eta(t)$ to obtain estimates
  - Eigenfunctions $\hat{\phi}_K$
  - Variance of scores $\hat{\lambda}_K$
  - Population mean $\hat{f}_0$

FDA in Dynamic Prediction

- Challenge: out-of-sample prediction
- Method:
  - Maximum Likelihood Estimation (MLE):
    - For a new subject with partially observed track, at any t_j beyond observation,

$$\hat{\eta}(t_j) = \hat{f}_0(t_j) + \sum_{k=1}^{K} \hat{\xi}_k \hat{\phi}_k(t_j)$$

  - Bayes theorem
    - Prior distribution: $\xi_k \sim N(0, \hat{\lambda}_k)$
    - Posterior distribution:

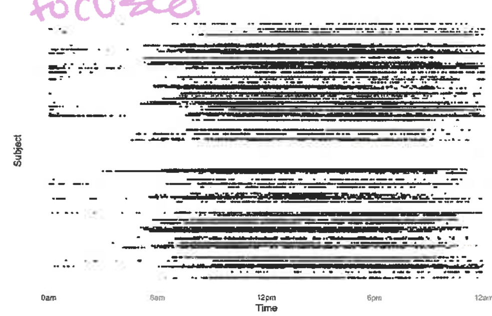
$$P(Y(t_j)|\xi) = l(\xi) = \sum log(h(Y(t_j))) + \eta(t_j)T(Y(t_j)) - log(A[\eta(t_j)])$$

- Use spline basis to project prediction to the original grid

Predicting NHANES Binary Activity Indicator

- The National Health and Nutrition Examination Survey (NHANES)
  - A large, stratified, multistage survey conducted by the Centers for Disease Control (CDC)
  - Represent the non-institutionalized US population
  - We focuse on the minute-level activity indicator

focuzed



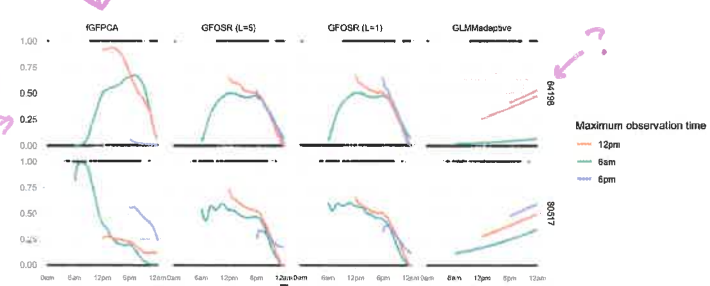
Predicting NHANES Binary Activity Indicator

- Proposed method: fGFPCA
- Reference methods:
  - GLMM using Adaptive Gaussian Quadrature (GLMMadaptive), with a random slope for time
  - Generalized Function on Scalar Regression (GFOSR)
- Evaluation metrics
  - Area-Under-the-Receiver-Operator-Curve (AUC)

Results: Individual Predicted Tracks

what's this axis?

?



- fGFPCA can accomodate much greater flexibility, thus more consistent with the activity pattern
- Prediction of fGFPCA updates as extra data is collected

Result: AUC (%)

	Maximum observation time											
	fGFPCA			GFOSR (L=5)			GFOSR (L=1)			GLMMadaptive		
Window	6am	12pm	6pm	6am	12pm	6pm	6am	12pm	6pm	6am	12pm	6pm
6am-12pm	**70.3**			68.9			68.4			58.1		
12pm-6pm	**53.6**	**70.9**		51.9	69.9		52.0	65.4		53.2	70.1	
6pm-12am	**71.6**	**67.9**	**77.3**	67.7	67.8	72.3	67.7	67.6	70.8	51.4	56.5	62.6

- fGFPCA outperforms the reference methods in all cases, espeicially
  - as observed track extends
  - when prediction window is far from observed track
- Computation time is similar between fGFPCA and GLMMadaptive, but the former is much more flexible

Simulation Study

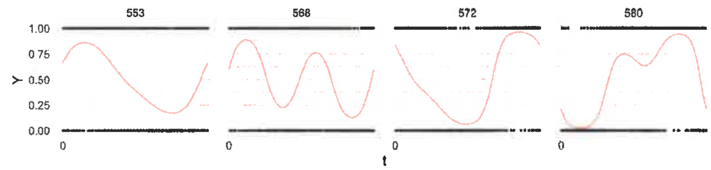
- Simulation set-up:

$$Y_i(t) \sim Bernoulli(\frac{exp(\eta_i(t))}{1+exp(\eta_i(t))})$$

$$\eta_i(t) = f_0(t) + \xi_{i1}\sqrt{2}sin(2\pi t) + \xi_{i2}\sqrt{2}cos(2\pi t) + \xi_{i3}\sqrt{2}sin(4\pi t) + \xi_{i4}\sqrt{2}cos(4\pi t)$$

$$f_0(t) = 0, \quad \xi_{ik} \sim N(0, 0.5^{k-1}), \quad k \in \{1,2,3,4\}$$

- Complex non-linear underlying pattern



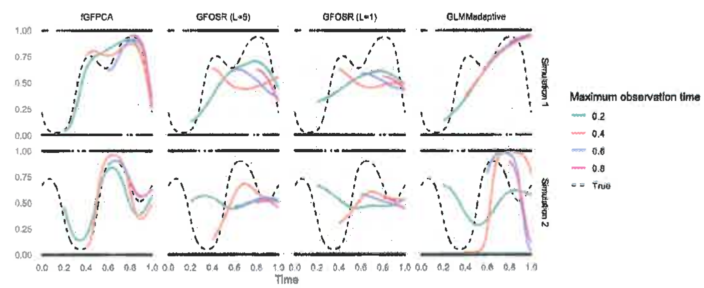
Simulation Study

	Simulation 1	Simulation 2
Training size	500	100
Test size	100	
Number of measurements per subject	1000	
Number of simulated datasets	500	
Random effects in GLMMadaptvie	Linear	Spline basis
Number of observations used in GFOSR	L=1 or L = 5	

axis?

- In simulation 2, we increase the complexity of GLMMadaptive at the expense of training size
- In addition to AUC, we also use Integrated Squared Error (ISE) for performance evluation

Results: Individual Predicted Tracks



- fGFPCA shows advantage in flexibility, accuracy and efficiency

Results: Simulation 1

	Maximum observation time															
	fGFPCA				GFOSR (L=5)				GFOSR (L=1)				GLMMadaptive			
Window	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
ISE																
(0.2, 0.4]	**146**				275				363				388			
(0.4, 0.6]	**184**	**75**			277	220			287	263			292	270		
(0.6, 0.8]	**218**	**49**	**16**		322	373	325		386	411	370		316	283	278	
(0.8, 1.0]	**109**	**78**	**18**	**12**	291	318	351	334	329	341	354	347	563	478	598	600
AUC (%)																
(0.2, 0.4]	**74.8**				68.6				62.4				59.1			
(0.4, 0.6]	**66.4**	**73.4**			56.3	63.0			54.3	59.0			52.4	60.6		
(0.6, 0.8]	**71.5**	**79.0**	**80.3**		66.9	62.8	67.6		60.4	57.7	61.5		66.9	69.4	68.7	
(0.8, 1.0]	**74.0**	**75.5**	**78.1**	**78.4**	62.6	60.6	55.2	58.4	58.8	56.4	53.7	55.1	51.4	55.6	52.6	56.4

- The advantages of fGFPCA seem greater compared to the NHANES data application
- Possibly due to complex underlying pattern

Results: Simulation 2

↳ explain change (more flexible)

	Maximum observation time															
	fGFPCA				GFOSR (L=5)				GFOSR (L=1)				GLMMadaptive			
Window	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
ISE																
(0.2, 0.4]	**150.6**				286.9				367.3				374.8			
(0.4, 0.6]	**188**	**77**			287	227			291	266			268	463		
(0.6, 0.8]	**224**	**52**	**17**		329	375	327		387	408	388		320	287	234	
(0.8, 1.0]	**112**	**81**	**20**	**13**	301	325	354	335	332	343	355	348	228	470	332	132
AUC (%)																
(0.2, 0.4]	**74.6**				67.9				62.0				67.2			
(0.4, 0.6]	**66.2**	**73.4**			54.9	62.7			53.6	59.0			62.2	67.1		
(0.6, 0.8]	**71.0**	**78.7**	**80.1**		65.6	61.7	66.8		59.8	57.4	61.3		68.9	69.8	73.1	
(0.8, 1.0]	**74.0**	**75.4**	**78.1**	**78.4**	61.1	58.8	55.1	57.8	58.0	55.6	53.6	55.1	68.1	62.9	67.9	74.3

- The advantages of fGFPCA sustained after increasing the flexiblity of GLMM adaptive

Discussion

- fGFPCA can accommodate more flexible correlation structure between repeated measure
- Compared to mixed models, fGFPCA reduced time spent on model fitting while achieving much better predictive performance
- However, extension to other data structure is at work:
  - Multi-level functions
  - Multi-variate functions

References

J C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3/4): 325–338, 1966. doi: https://doi.org/10.2307/2333639.

Fernando Holguin, Suzy A A Comhair, Stanley L Hazen, Robert W Powers, Sumita S Khatri, Eugene R Bleecker, William W Busse, William J Calhoun, Mario Castro, Anne M Fitzpatrick, Benjamin Gaston, Elliot Israel, Nizar N Jarjour, Wendy C Moore, Stephen P Peters, W Gerald Teague, Kian Fan Chung, Serpil C Erzurum, and Sally E Wenzel. An association between l-arginine/asymmetric dimethyl arginine balance, obesity, and the age of asthma onset phenotype. *American Journal of Respiratory and Critical Care Medicine*, 187(2):153–9, 2013. doi: 10.1164/rccm.201207-1270OC.

Timothy Amass, Lauren Jodi Van Scoy, May Hua, Melanie Ambler, Priscilla Armstrong, Matthew R. Baldwin, Rachelle Bernacki, Mansoor D. Burhani, Jennifer Chiurco, Zara Cooper, Hope Cruse, Nicholas Csikesz, Ruth A. Engelberg, Laura D. Fonseca, Karin Halvorson, Rachel Hammer, Joanna Heywood, Sarah Hochendoner Duda, Jin Huang, Ying Jin, Laura Johnson, Masami Tabata-Kelly, Emma Kerr, Trevor Lane, Melissa Lee, Keely Likosky, Donald McGuirl, Tijana Milinic, Marc Moss, Elizabeth Nielsen, Ryan Peterson, Sara J. Puckey, Olivia Rea, Sarah Rhoads, Christina Sheu, Wendy Tong, Pamela D. Witt, James Wykowski, Stephanie Yu, Renee D. Stapleton, and J. Randall Curtis. Stress-Related Disorders of Family Members of Patients Admitted to the Intensive Care Unit With COVID-19. *JAMA Internal Medicine*, 182(6):624–633, 06 2022. ISSN 2168-6106. doi: 10.1001/jamainternmed.2022.1118. URL https://doi.org/10.1001/jamainternmed.2022.1118.

Luo Xiao, Vadim Zipunnikov, David Ruppert, and Ciprian Crainiceanu. Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, 26(1):409–421, 2016. doi: https://doi.org/10.1007/s11222-014-9485-x.

Add in a final conclusion slide

- re-iterate your research experience and goals
- an opportunity to highlight why you are the best fit for this role!