**Title**

Hello everyone! I’m Aidan, a 2nd year PhD student working with Mark Fiecas, Thierry Chekouo, and Sandra Safo. Today I’ll share my reading of Thierry and Sandra’s paper *Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease* and how this work relates to my current research.

**What will I share?**

Today I’ll introduce Thierry and Sandra’s paper, providing context as to the “why” behind the work. Afterward, I’ll discuss the model they’ve developed, touch on their simulation studies, application results, and conclusions. Lastly, I’ll discuss my research interests and my current project as an extension of their method.

**Bayesian Factor Analysis**

[**Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors1**](https://en.wikipedia.org/wiki/Factor_analysis)**.**[**It is a process of grouping variables you’ve recorded into categories called factors2**](https://www.indeed.com/career-advice/career-development/factor-analysis)**.**[**This can help you reduce the total number of variables involved in a statistical study, which can make evaluating the data easier2**](https://www.indeed.com/career-advice/career-development/factor-analysis)**.**[**Statisticians take these unobserved variables and study whether they could be common factors behind observed outputs in a data set3**](https://www.masterclass.com/articles/factor-analysis)**.**

**Introduction: Bayesian, explicit modeling of factors shared across data for prediction**

Disease risk is complex and generally involves genetic, biological, and environmental factors. This approach jointly associates data from multiple sources, accounts for clinical covariate, and predicts an outcome. In addition to prediction, variable selection is a priority since doing so allows for interpretability. Furthermore, as a Bayesian method, prior functional knowledge can be incorporated into any analysis.

Frequentist methods exist for joint association and prediction. The authors mention a few… Luo et al. (2016) and Safo et al. and (2021) combined canonical correlation analysis (CCA) with classification/ regression methods. Safo et al. (2021) proposed a joint association and classification method that combines CCA and linear discriminant analysis and uses the normalized Laplacian of a graph to smooth the rows of the discriminant vectors from each omics data.

This method jointly associates multiple data and predicts a clinical outcome in a Bayesian framework. There are Bayesian methods for integrating omics data and predicting a clinical outcome (Wang and others, 2012; Chekouo and others, 2015, 2017). *This method’s novelty is particularly its explicit modeling of the factors that are shared/ common across data types*. *Such an approach* *supports interpretability, variable selection, and prediction.*

**Introduction: Problem-driven method development**

Problem-driven method development is highlighted. I’m appreciative of this because I’m someone oriented toward solving real-world problems. Also, I’m new to biostatistics as a discipline and sometimes struggle with how I might develop my own methods someday.

Here, the authors analyze data from Emory University and Georgia Tech Predictive Health Institute study, which began in 2005. They are aimed at identifying genetic variants, genes and pathways contributing to healthy adult 10-year atherosclerosis cardiovascular disease (ASCVD) risk for a few reasons:

* Cardiovascular diseases (CVD) are the leading cause of death in the U.S. (AHA, 2016).
* Nearly half of the U.S. population will have some form of CVD by 2035, costing the economy an estimated $2 billion/day in medical costs (AHA, 2016).
* Research suggests environmental risk factors for ASCVD (e.g., age, gender, hypertension) account for only half of all cases of ASCVD (Bartels and others, 2012).

The results might be investigated as candidate biomarkers/ interrogated for novel therapy approaches.

As an aside, today, I’m working with Thierry, Sandra, and Mark Fiecas, and another application is motivating extension of this method.

**Model Formulation: Factor analysis relates different data types.**

* *Primary aim is to combine multiple data types and incorporate clinical covariates, a response variable, and external grouping information.*
* Omics data…
* Outcome…

**Model Formulation: Choosing active components and important features.**

* Component- and variable-level latent binary variables…

**Model Formulation: Incorporating grouping/ pathway information.**

* Data source… Authors grouped data at two levels: SNPs into nearby genes, and genes into pathways. To group SNPs into genes, authors identified genes close to SNPs using the R package bioMart. Authors grouped genes into pathways via a network analysis on Ingenuity Pathway Analysis (IPA), a software program for analyzing gene expression patterns using a built-in scientific literature-based database (according to IPA Ingenuity Web Site, www.ingenuity.com).
* Method… ***A picture containing text, font, screenshot, document

  Description automatically generated***global shrinkage hyperparameter?
* ***A picture containing text, font, screenshot, algebra

  Description automatically generated***

**Model Formulation: It’s important to step back…**

* Hierarchical model sketch ☺

**Posterior Inference: MCMC mixture**

* For each MCMC, fit from 4 starting points and feature/ group selection marginal posterior probabilities correlation > 0.97 indicated appropriateness

**Prediction: Posterior median used for posterior predictive.**

**Simulation Studies:**

* Scenarios…
* Results…

**Application Results:**

**Conclusions/ Questions**

* 1
* 2
* 3

**Introduction to my research interests**

My research interests have evolved over the years but have a consistent theme: a desire to understand complex systems and distill learnings.

Last semester, I had the opportunity to explore with Mark Fiecas as my mentor thanks to the Stat Gen Training Program. On reading Robert Tibshirani’s research group’s paper *Cooperative learning for Multiview analysis*, I was attracted to Multiview learning methods, [methods that learn shared patterns across multiple distinct feature sets/ data sources**1**](https://www.sciencedirect.com/science/article/pii/S0925231221004768). Then I asked for reading suggestions from Sandra Safo and discovered “Multiview learning” [is also known as data fusion, data integration](https://www.sciencedirect.com/science/article/pii/S1566253516302032), or can be associated with multi-omics analyses. Also last semester, I became familiar with -omics technologies, using PLINK to perform a GWAS, and applied to the NSF’s Graduate Research Fellowship Program, which I learned I did not get. But that’s okay because I’m appreciative of the learning experience applying to an NSF fellowship and considering how my research might unfold in the next three years.

**Why? Extension to a longitudinal context**

In my first year, through an RA, I became familiar with the Adolescent Brain and Cognitive Development (ABCD) Study. The ABCD Study is an ongoing longitudinal research study involving data collection at 21 sites across the United States. [It is the largest long-term study of brain development and child health in the US**3**](https://en.wikipedia.org/wiki/ABCD_Study). [The study explores the environmental, social, genetic, and biological factors that affect brain and cognitive development, behavior, and health**4**](https://www.nimh.nih.gov/research/research-funded-by-nimh/research-initiatives/adolescent-brain-cognitive-developmentsm-study-abcd-studyr). There are many data types available, so Mark and I wondered how we might apply Multiview learning in this longitudinal context.

Mark expressed the thought of developing a longitudinal data analysis method to Thierry Chekouo and turned-out Thierry had already been considering such a project. Thierry and Sandra Safo had extended data integration methods for variable selection and prediction to a Bayesian framework in their recently published *Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease*, which included an R package [BIPnet](https://github.com/chekouo/BIPnet). Thierry, Sandra, Mark, and I first met to discuss extending this method to be capable of longitudinal analysis with application to the ABCD Study early this semester.

**Exploratory Analyses: Adolescent Brain Cognitive Development (ABCD) Study®**

Since wanting to extend BIPnet longitudinally, we have started analysis pipeline development, exploratory data analyses, and model formulation. Also, I have spent time getting familiar with C++ and the Rcpp package.

For better or for worse, the COVID-19 disrupted the ABCD Study’s data collection. Neuroimaging data is collected every two years, and the pandemic started between baseline and time point 2 (T2). In our exploratory analyses, we observed substantial neuroimaging missingness at T2. Consequently, we restricted our sample to subjects with neuroimaging data at both time points. Doing so leaves us with ZZZ subjects/ ZZZZ subjects enrolled. There is missingness in the covariates/ response variables we are including but plan to impute this data. Note, when I say neuroimaging data, for now, I mean the two structural MRI metrics Surface Area and Cortical Thickness for 34 regions x2 hemispheres x2 metrics = 128 features.

**Model Formulation: Early thoughts**

In a Bayesian analysis, we ultimately want the joint posterior distribution, so we can perform variable selection/prediction. While I still need to work through the notation and think through the details, here I am presenting the model at a high level. This represents … One complication of the ABCD Study is that data is collected across 21 sites and subjects might share the same family. The variables … represent random intercepts to account for nested effects, but we need to interpret/resolve error terms in relation to these random effects in the outcome equation.

**Ponderings/ Next Steps**

Presently, I have a few questions and next steps I’m pondering.

Firstly, is the combination of imaging data, covariates, and clinical outcomes biologically meaningful? How might we incorporate prior information? At Saonli and Mark’s talk at the Human in the Data Spotlight Series last Thursday, I was reminded of the importance of including a subject matter expert in the conversation, so I’m wondering how, as a student, I might build relationships with non-statistics collaborators.

Secondly, are two time points sufficient for considering the proposed method’s efficacy longitudinally? I’m unsure at present. I have more to read on longitudinal analysis. Perhaps we can include additional data as it’s released or might consider adding/ switching to other data.

Thirdly, how might I implement the method so it’s scalable? Sandra mentioned her general preference for frequentist methods thanks to their relative computational efficiency, but Bayesian methods offer the chance to incorporate prior information. I’m wondering how to architect the software so adding additional omics data doesn’t increase the runtime beyond practicality.

Regardless, I’m hoping to submit this methods and applied work to this fall’s ENAR student paper competition.

Finally, I’d like to acknowledge Mark, Sandra, and Thierry for their mentorship, and the NIH for making this work possible. Also, thanks to this Stat Gen training program, its organizers Saonli, Tianzhong, and Wei Pan, the biostats faculty generally, fellow students, family, and friends for their support.