Joint modelling of binary and continuous measurements in large health surveys and its application to network analysis, frailty, and mortality in NHANES 1999-2010

Debangan Dey[1], Irina Gaynanova[2], Vadim Zipunnikov[1]
Johns Hopkins Bloomberg School of Public Health[1], Texas A&M University[2]

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Abstract

- Network analysis has rapidly gained popularity in neuroimaging, genomics and other scientific domains. However, a little has been done to adapt network analysis to heterogeneous measurements collected by national health surveys and biobanks.
- This is primarily due to a lack of understanding on how to jointly model multiple comorbidities, health deficits, and health biomarkers, often recorded via binary and continuous measurements. Our approach adapts a recently proposed semiparametric gaussian copula[1] that estimates a latent correlation structure of mixed type (binary and continuous) random vectors through rank-based procedure.
- After estimating joint distribution of latent continuous variables, we build a network by applying sparse inverse covariance estimation method (Graphical Lasso) to control for number of connections. We choose the optimum penalizing parameter in a way to ensure the stability of the network.
- Extending further, we propose a novel solution to jointly model outcome and predictors, impute missing data, perform dimension reduction and do prediction both on the latent and observed space. The key advantage of this approach is the combination of mixed data-type under a uniform modelling framework.
- We demonstrate this method on 47 binary and continuous variables typically included in Frailty Index (FI). Using latent principal components and network connectivities, a few weighted versions of FI are developed and compared in predicting 5-year mortality in National Health and Nutrition Examination Survey.

Framework

Definition 1:
We define a random vector $X = (X_1, \dots, X_p)^t \sim \text{NPN}\,(0, \Sigma, f)$ if there exists a set of monotonic increasing functions $f = (f_1, f_2, \dots, f_p)$ such that - $f(X) = \left(f_1(X_1), f_2(X_2), \dots, f_p(X_p)\right)^t \sim N(0, \Sigma)$ with $\Sigma_{jj} = 1 \; \forall \; 0 \le j \le p_1 + p_2 = p$. NPN stands for Non-paranormal distribution defined by Liu et al[1].

Definition 2:
Suppose we have an observed vector of variables $X = (X_b, X_c)^t$, where X_b represents $p_1$dimensional binary random variables and X_c represents p_2 dimensional continuous random variables. We say, $X \sim LNPN(0, \Sigma, f, C)$ if there exists a set of latent variables Z_b and a vector of constants $C = \left(C_1, C_2, \dots, C_{p_1}\right)^t$ such that $X_{bj} = I\,(Z_{bj} > C_j)$ for $j = 1, \dots, p_1$ and $Z = (X_c, Z_b) \sim NPN\,(0, \Sigma, f)$. LNPN stands for latent non-paranormal distribution.

We use a semiparametric Gaussian Copula framework to jointly model binary (clinical) and continuous (lab) measurements and recover latent underlying structures. For i-th subject, we observe the vector $\mathbf{X_i} = (\boldsymbol{X_{ib}}, \boldsymbol{X_{ic}})$, where $\boldsymbol{X_{ib}}$ represents p_1-dimensional binary measurements and $\boldsymbol{X_{ic}}$ represents p_2-dimensional continuous measurements. We assume $\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n} \sim LNPN(0, \Sigma, f, C)$ and we have latent unobserved variables $\boldsymbol{Z_1}, \boldsymbol{Z_2}, \dots, \boldsymbol{Z_n} \sim NPN(0, \Sigma, f)$ as defined above.

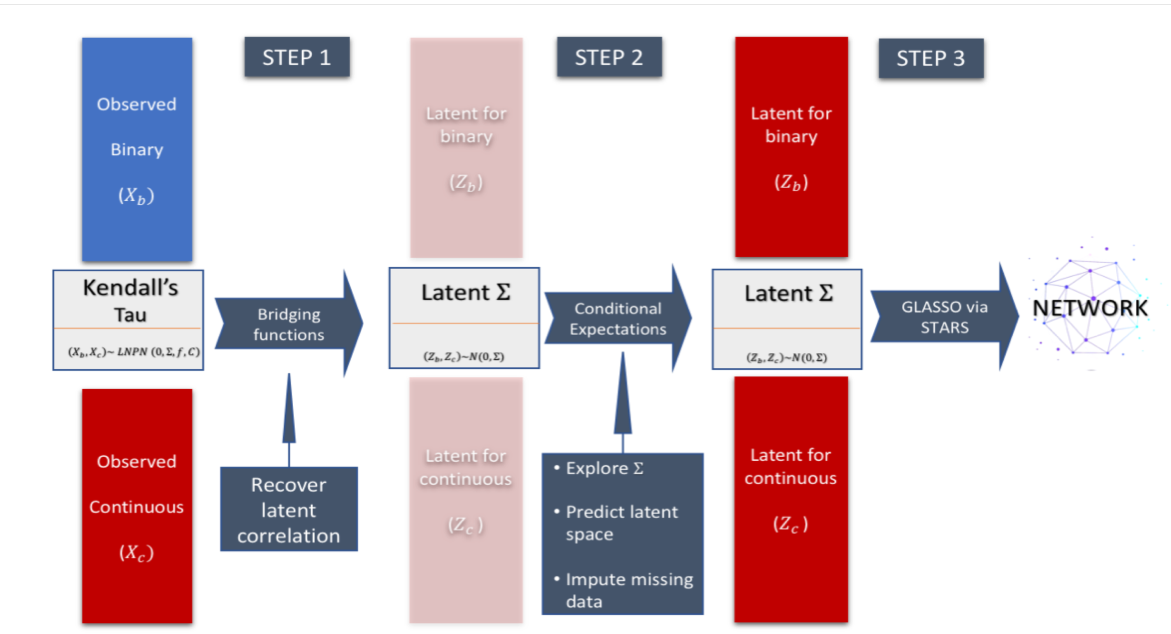


Table 1: Comparison
(Y = outcome, X = predictor, $\hat{L}$ =predicted latent variable)

	(Traditional models)	(Joint normal models)
1. Joint dependence structure	a) No clear way to think, apart from getting a non-parametric estimate of sample covariance matrix (like using Kendall's Tau or Spearman's Rank Correlation). b) No clear method for dimension reduction except PCA.	a) We can use $\hat{\Sigma}_{XX}$ or more broadly $\hat{\Sigma}$ to define and visualize the dependence structure among covariates and also with outcome included. b) We can do dimension reduction of covariates after finding PC loadings of $\hat{\Sigma}_{XX}$ and we can later compute principal scores on the latent space.
2. Individual associations	a) We can do logistic regression of Yon individual components of X one by one. b) Measures of fit: AUC or residual deviance.	a) We can use the specific elements of $\hat{\Sigma}_{YX}$ to denote the association of Y with corresponding variable. b) Measures of fit: Latent R-square ($\hat{\Sigma}_{YX}[k]^2$) for k-th covariate.
3. Global association	a) Global logistic regression model. b) Measures of fit: AUC or residual deviance.	a) We can define latent β as $\beta_L = \hat{\Sigma}_{YX}\Sigma_{XX}^{-1}$ and use it as a coefficient on $\hat{L}$ to fit global association model. =(or) We can fit global logistic regression of Y on $\hat{L}$. b) Latent R-square: $\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ (or) AUC or residual deviance from doing logistic regression of Y on $\hat{L}$.
4. Missing data imputation	No clear way, except we impute by corresponding means or take only complete cases.	We do imputation naturally as a process of finding and computing $\hat{L}$.
5. Prediction on new data	Prediction after fitting Yon X in a traditional logistic regression model.	Compute $\hat{L}$ from new vector of X using population level assumptions. Use one of the fitted models in (3.) to predict Y.

Table 2: Figure terminology

	Details	Method
Latent coefficient	Latent, uses joint dependency of measurements	Perform linear regression on latent space and obtain latent coefficient as $\beta_L = \hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1}$
Logistic coefficient	Observed, uses joint dependency of measurements	Obtained from logistic regression of Y on X after mean-imputing and scaling X.
Latent individual correlation	Latent, uses measurements one-by-one	For each covariate, get the latent correlation with outcome (mortality), i.e. the vector $\hat{\Sigma}_{YX}$.
Column norm of latent covariance matrix	Latent, uses joint dependency of measurements	Euclidean norm of columns of $\hat{\Sigma}_{XX}$
Column norm of latent precision matrix	Latent, uses joint dependency of measurements	Euclidean norm of columns of $\hat{\Sigma}_{XX}^{-1}$.
AUC	Observed, uses measurements one-by-one	AUC from fitting models of logistic regression of Y on a single measurement.
p-value	Observed, uses measurements one-by-one	-$\log_{10}$(p-value) from fitting models of logistic regression of Y on a single measurement.
Mutual Information	Observed, uses measurements one-by-one	Mutual information of Y with every component of X.
Variance explained	Latent, uses measurements one-by-one	The latent R^2 value ($\hat{\Sigma}_{YX}^2$).

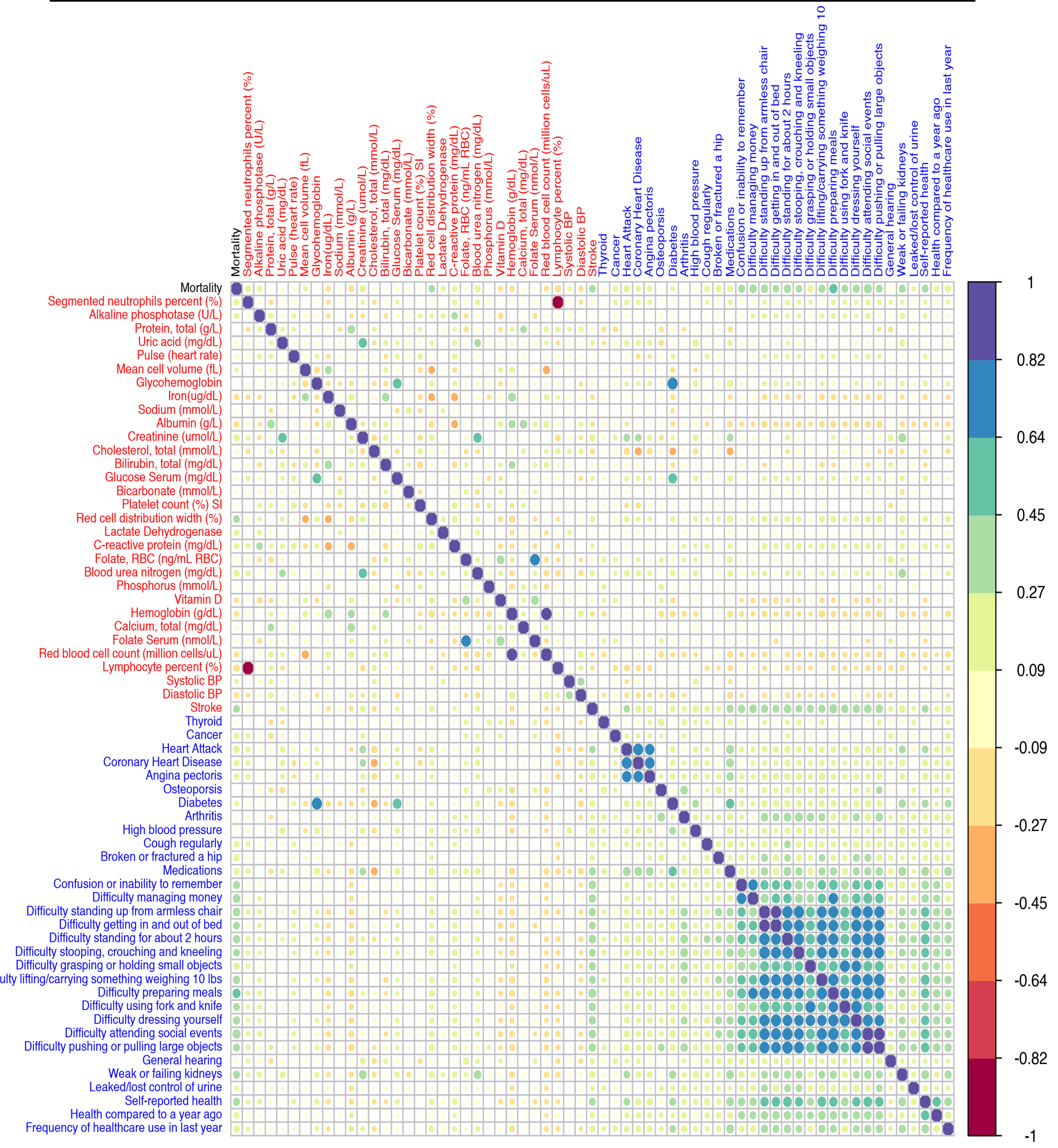
Joint modeling of outcome and predictor

- A binary (0/1) outcome of interest Y_i (5 year mortality status).
- We treat (Y_i, X_i) as our new vector of interest and conduct steps 1, 2, 3 to jointly model the outcome and the covariates (binary and continuous measurements).
- Instead of imposing assumptions on the distribution of $Y|X$, we assume that jointly $(Y, X) \sim LNPN(0, \Sigma, f, C)$. As a consequence, we can get the estimate of latent correlation matrix –

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{YY} & \hat{\Sigma}_{YX} \\ \hat{\Sigma}_{XY} & \hat{\Sigma}_{XX} \end{pmatrix}$$

We compare how our approach differs from traditional modelling approaches in Table 1.

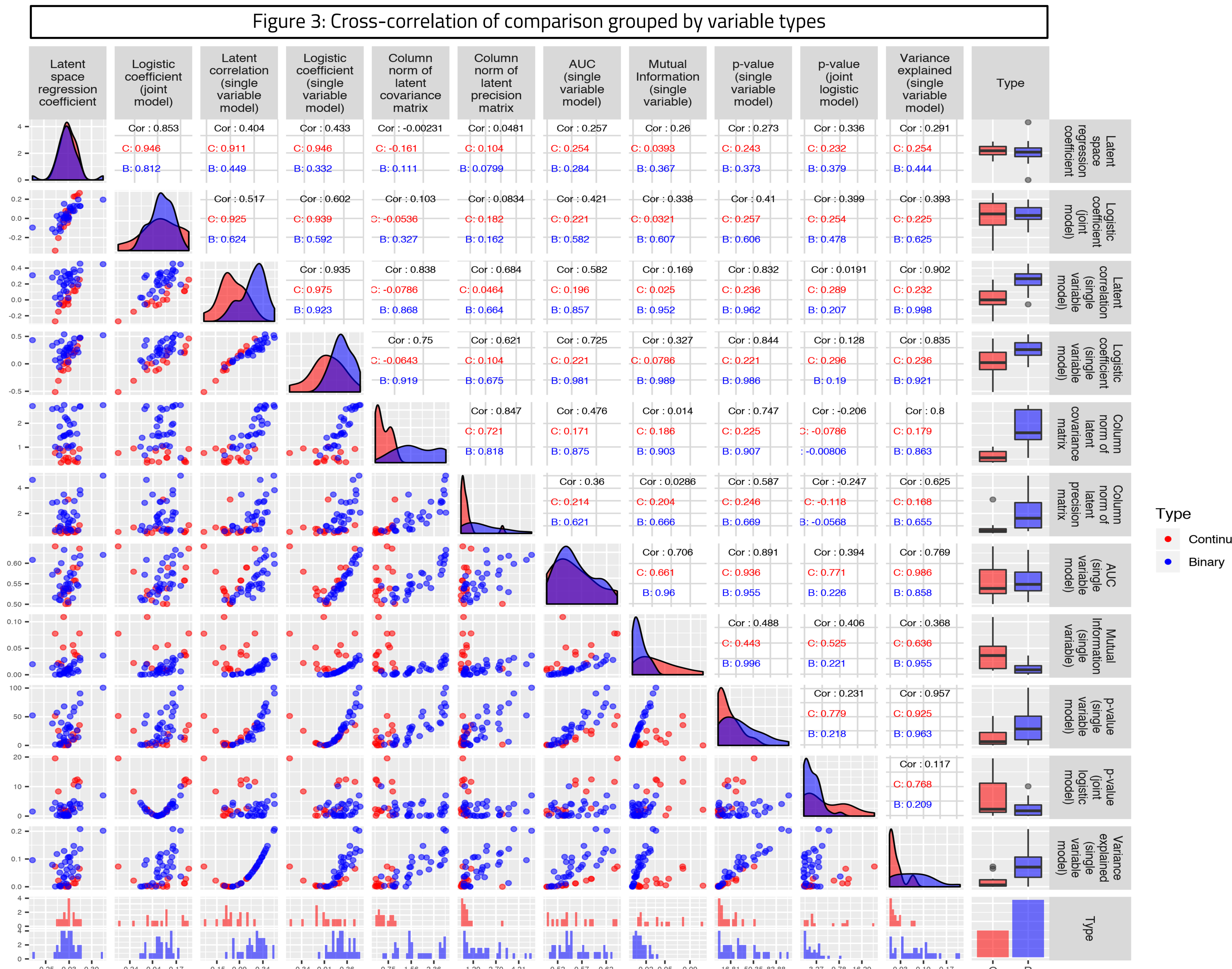
Figure 1: Heatmap of latent correlation



From the network, we group the frailty variables into three classes –
(i) Direct *(directly connected to mortality node),*
(ii) Indirect *(connected with a node which is connected to mortality),*
(iii) No *(not in class (i) and (ii)).*

We consider a list of informative measures and diagnostics listed in **Table 2** and the number of connections of a node, group them into connection categories (**Fig. 3**) and variable types (**Fig. 4**), visualize them in a scatterplot matrix with rank correlation values printed in the upper half.

Figure 3: Cross-correlation of comparison grouped by variable types



Results

Data:

- We combined data on 30 continuous measurements (lab) and 32 binary measurements (clinical) from 1999-2010 cohorts of National Health and Nutrition Examination Survey (NHANES) for everyone aged 60+.
- We removed subjects with more than 20% missing information and out of the remaining, 8947 subjects satisfied our criteria of 5-year follow-up on mortality.
- To tackle multicollinearity, we chose 15 continuous variables out of 30 in a forward selection way of explaining most variability.

Below we visualize the latent dependence structure of the deficits through the heatmap of the estimated latent correlation matrix (Fig. 1) and the network estimated through the graphical lasso approach (using STARS criteria for stability.

Figure 2: Network of deficits

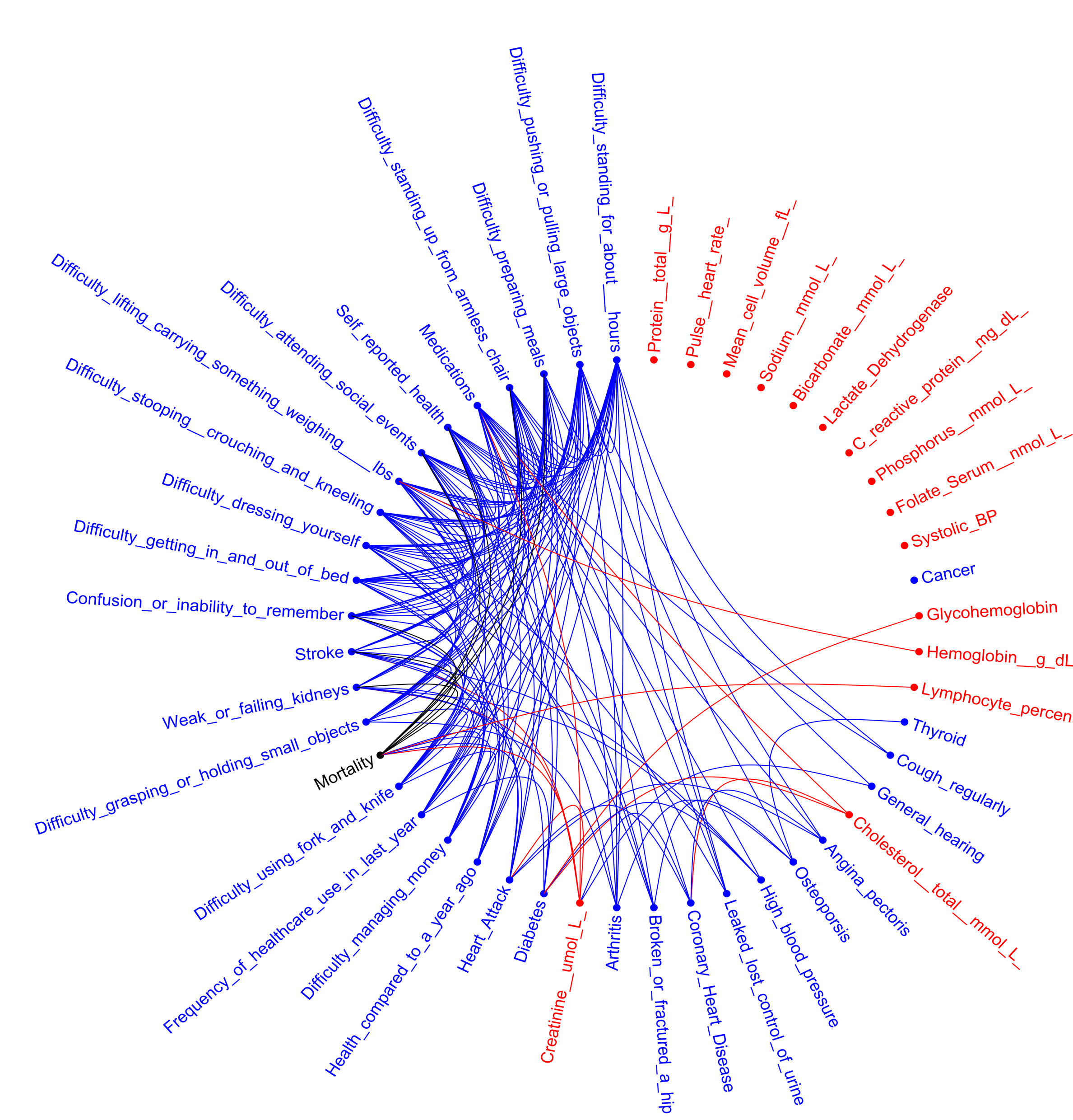
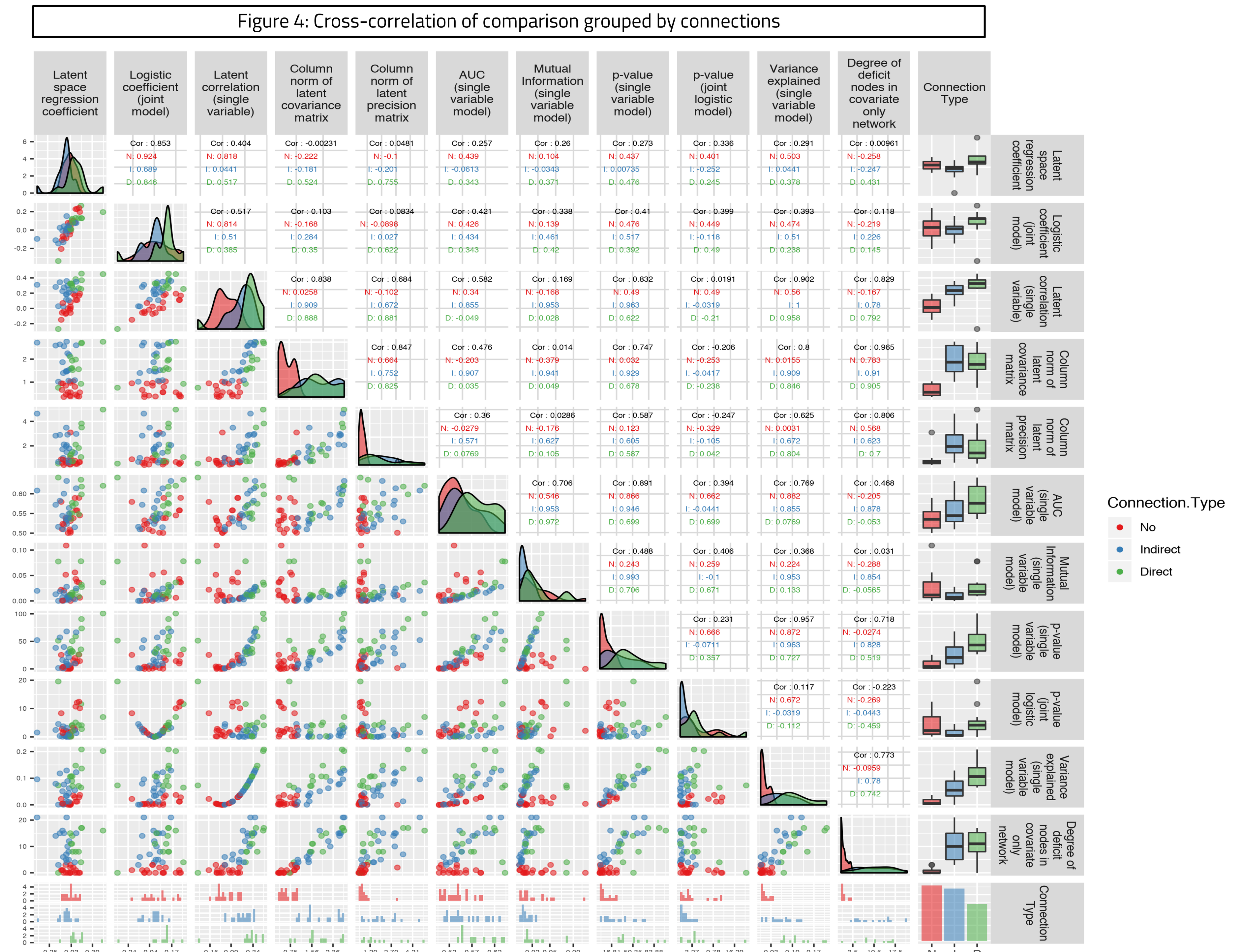


Figure 4: Cross-correlation of comparison grouped by connections



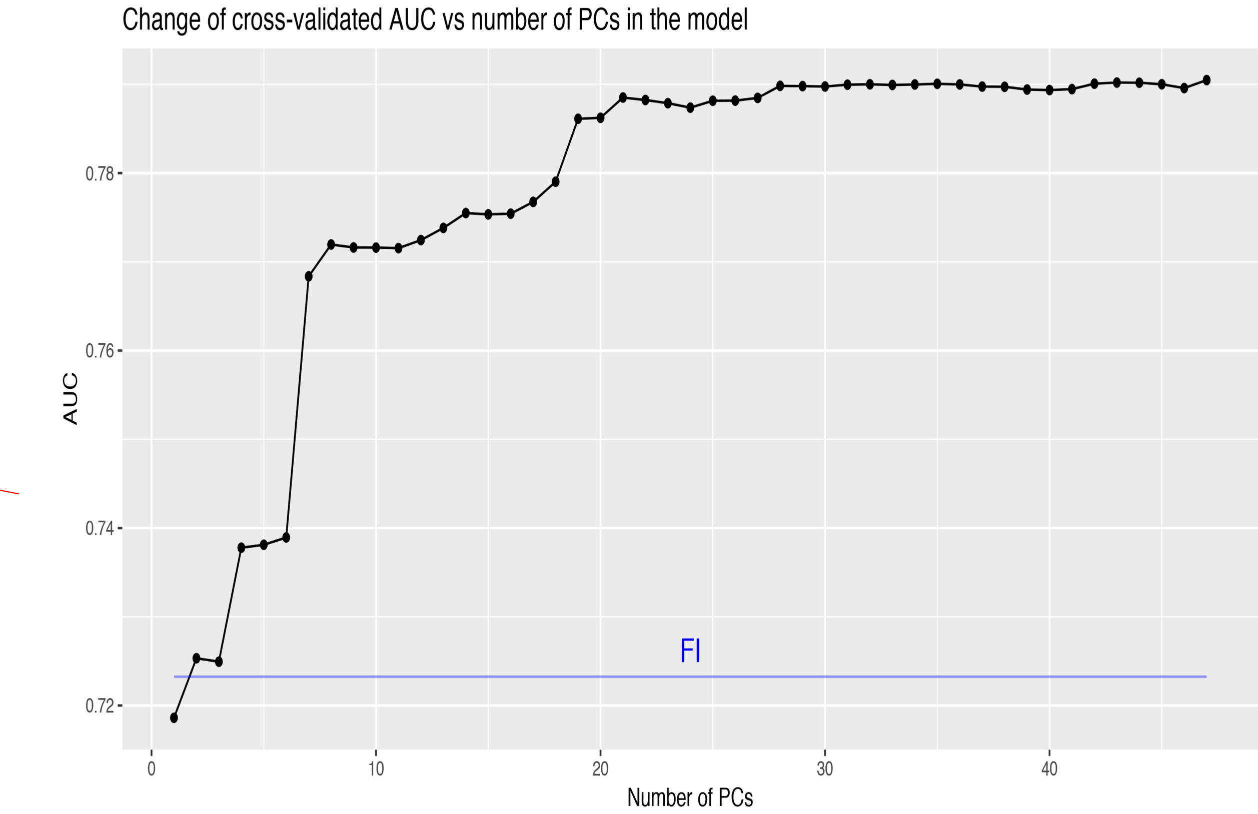
Conclusion and discussion

A. Insights from joint modelling
- A strong correlation between individual latent correlation with the coefficient of a single variable logistic regression model.
- A strong linear association between column norms of $\hat{\Sigma}_{XX}$ and the measurement specific correlation with outcome (mortality), specifically for binary (clinical) measurements.
- The higher the column norm of a specific measurement, the more we expect it to be related to other measurements and thus more relevant to the outcome.
- A strong correlation between the degree of a node in the network and latent correlation with mortality can substantiate the hypothesis that the more connected is the node to other nodes, the more related will it be to mortality. Farell et al [2] based their simulations on a similar hypothesis and analyzed simulated network scenarios in frailty deficits.
- This gives us understanding of how intra-dependency of covariates can influence association with outcome and how to filter out redundancy.

B. Improvement of Frailty Index (FI)
- **Current approach:** First binarize the continuous measurements (lab) to get clinical deficit indicators then define ***Frailty Index (FI)*** as the proportion of deficits accumulated among a set of measured binary deficits.
- **Our approach:** Do principal component analysis on the latent correlation matrix and define principal scores based on the predicted latent measurements. ***(Modified FI)***

  Advantages – Takes into account the dependence among measurements and leads us to the direction of the most variability among subjects.

Figure 5: AUC comparison of Latent PC vs FI



References

1. Fan, Jianqing, Han Liu, Yang Ning, and Hui Zou. "High dimensional semiparametric latent graphical model for mixed data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79, no. 2 (2017): 405-421.

2. Farrell, Spencer G., Arnold B. Mitnitski, Olga Theou, Kenneth Rockwood, and Andrew D. Rutenberg. "Probing the network structure of health deficits in human aging." *Physical Review E*98, no. 3 (2018): 032302.