High Dimensional Latent Gaussian Copula Model for Mixed Data in Imaging Genetics

Aiying Zhang[1], Jian Fang[1], Vince D. Calhoun[2], Yu-Ping Wang[1]
[1] Department of Biomedical Engineering, Tulane University
[2]The Mind Research Network, University of New Mexico

Motivation

Schizophrenia (SZ) is a chronic and severe mental disorder that affects how a person thinks, feels, and behaves. Studies have shown that many complex mental disorders including SZ are highly correlated with genetic variants. Currently, it is possible to collect data on structural, functional brain images (fMRI), as well as genetic factors (e.g. SNPs) for the same subject. However, the integration of neuroimaging and genetic biomarkers for comprehensive understandings of mental illnesses has become a daunting challenge. Our goal is therefore to find significant associations between specific brain regions and genetic variants.

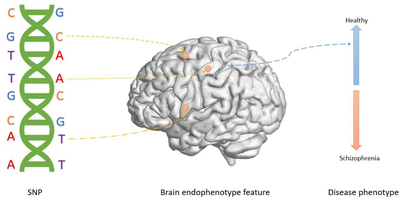


Fig.1. The relationship among SNP, brain connectivity and the disease

In statistics, this corresponds to a mixed data modeling problem. The image data is always continuous, while the SNPs data is multinomial with value {0,1,2}. Thus, we propose a latent Gaussian copula model for mixed data containing multinomial components, which fills a vacancy of graphical modeling in imaging genetics. The performance of the proposed method is firstly assessed through simulation studies. Then it is validated with real fMRI and SNP data collected by the Mind Clinical Imaging Consortium (MCIC) for a schizophrenia study.

Methods

Assumption:
The discrete values are obtained by discretizing a latent continuous variable at some unknown cutoffs.
Definition:

- Gaussian copula model (nonparanormal model) [1]:

A random vector $\boldsymbol{X} = (X_1, X_2, \dots, X_p)'$ is sampled from the Gaussian copula model (nonparanormal model), i.e. $\boldsymbol{X} \sim NPN(0, \boldsymbol{\Sigma}, f)$, if and only if there exists a set of monotonically increasing transformation $f = \{f_j\}_{j=1}^{p}$, satisfying $f(\boldsymbol{X}) = \left(f(X_1), f(X_2), \dots, f(X_p)\right)' \sim N_p(0, \boldsymbol{\Sigma})$ with $diag(\boldsymbol{\Sigma}) = \boldsymbol{I}_p$.

- Latent Gaussian copula model (LNPN):

Let $\boldsymbol{X} = (\boldsymbol{X^1}, \boldsymbol{X^2})$ be a p-dimensional random vector, where $\boldsymbol{X^1}$ is a p_1-dimension multinomial vector and $\boldsymbol{X^2}$ is a p_2-dimension continuous vector, $p_1 + p_2 = p$. We say $\boldsymbol{X} \sim LNPN(0, \boldsymbol{\Sigma}, \mathrm{f}, \boldsymbol{C})$, if there exists a p_1 dimension random vector $\boldsymbol{Z^1} = (Z_1, Z_2, \dots, Z_{p_1})$ such that $\boldsymbol{Z} := (\boldsymbol{Z^1}, \boldsymbol{X^2}) \sim NPN(0, \boldsymbol{\Sigma}, f)$ and $X_j = \sum_{l=1}^{l} I(Z_j > C_{jl}), \forall\, j = 1,2, \dots, p_1$, where $\boldsymbol{C}_{p_1 \times L} = (\boldsymbol{C_1}, \boldsymbol{C_2}, \dots, \boldsymbol{C}_{p1})'$ is a matrix of constants.

❖ Noted:
Since $\boldsymbol{Z^1}$ is unobserved, The cutoff matrix $\boldsymbol{C}$ cannot be estimated. But $\Delta_{jl} := f(C_{jl})$ is identifiable. So we use $LNPN(\boldsymbol{\Sigma}, \boldsymbol{\Delta}, f)$ to denote the model.

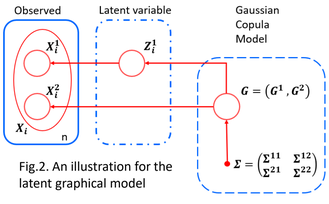


Fig.2. An illustration for the latent graphical model

Graph structure:

- After the transformation, we can estimate the latent correlation matrix $\boldsymbol{\Sigma}$, which is generated by Gaussian distributed variables.
- Under the LNPN, the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ identifies the conditional independence among $\boldsymbol{Z}$.
- Since the data in imaging genetics are always high dimensional, we adopted the ψ-learning method proposed by Liang et. al [2] to estimate $\boldsymbol{\Omega}$.

Theoretical results

Conditions:

a) There exists a constant $\delta > 0$, such that $|\Sigma_{jk}| \le 1 - \delta, \forall\, j \ne k = 1, \dots, p$.
b) There exists a constant $M > 0$, such that $|\boldsymbol{\Delta_j}|_\infty \le M, \forall j = 1, \dots, p$.

Theorem 1. *Under conditions (a) and (b), for any $t > 0$, we have*

- $P\left(\sup_{1 \le j,k \le p_1} |\hat{R}_{jk} - \Sigma_{jk}| > t\right) \le 2\exp\left(-\frac{nt^2}{8L_2^2}\right) + 4\exp\left\{-\frac{n\pi t^2}{64L^4 L_1^2 L_2^2}\right\} + 4\exp\left\{-\frac{2nM^2}{L_1^2}\right\}$
- $P\left(\sup_{1 \le j \le p_1, p_1+1 \le k \le p} |\hat{R}_{jk} - \Sigma_{jk}| > t\right) \le 2p_1p_2\exp\left\{-\frac{nt^2}{8L_3^2}\right\} + 2p_1p_2\exp\left\{-\frac{n\pi t^2}{48L^2 L_1^2 L_3^2}\right\} + 2p_1p_2\exp\left\{-\frac{2nM^2}{L_1^2}\right\}$
- $\sup_{p_1+1 \le j,k \le p} |\hat{R}_{jk} - \Sigma_{jk}| \le 2.45\pi\sqrt{log p_2/n}$

where L_1, L_2, L_3 are positive constants.

Corollary 1. *Under assumptions (a) and (b), with probability greater than $1 - p^{-1}$, we have*

$$\sup_{1 \le j,k \le p} |\hat{R}_{jk} - \Sigma_{jk}| \le C\sqrt{\log p\, /n}$$

where C is a constant independent of (n, p).

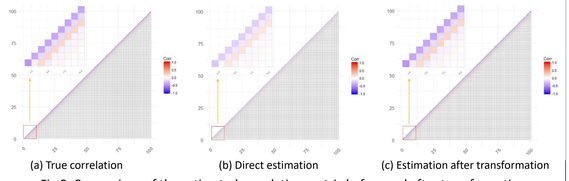
Simulation Results

Simulation Setting

- Simulate an autoregressive process of order two. Set $L = 2$.
- Consider the following two data generating scenarios:
  i. (Discrete case) Simulate $\boldsymbol{X} = (X_1, X_2, \dots, X_p)'$, where $X_j = I(Z_j > C_{j1}) + I(Z_j > C_{j2}), \forall j = 1,2, \dots, p$ and $\boldsymbol{Z} \sim N(0, \boldsymbol{\Sigma})$.
  ii. (Mixed case) Simulate $\boldsymbol{X} = (X_1, X_2, \dots, X_p)'$, where $X_j = I(Z_j > C_{j1}) + I(Z_j > C_{j2})$, for $j = 2i - 1, i = 1,2, \dots, p/2$, $\boldsymbol{Z} \sim N(0, \boldsymbol{\Sigma})$ and $X_j = Z_j$, for $j = 2i, i = 1,2, \dots, p/2$.
- Run100 datasets independently.

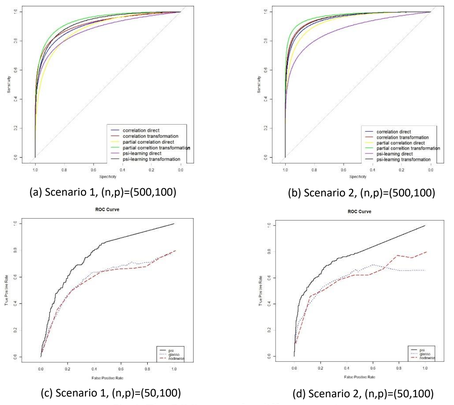
Simulation Result

- Used Pearson correlation, partial correlation and ψ-correlation.
- Direct estimation v.s. estimation after our proposed transformation.
- For high dimensional case, we compared ψ-learning method with gLasso [3] and nodewise regression [4] methods.



(a) True correlation (b) Direct estimation (c) Estimation after transformation

Fig.3. Comparison of the estimated correlation matrix before and after transformation



(a) Scenario 1, (n,p)=(500,100) (b) Scenario 2, (n,p)=(500,100)

(c) Scenario 1, (n,p)=(50,100) (d) Scenario 2, (n,p)=(50,100)

Fig.4. The ROC curves for different settings

Schizophrenia Study

- Use fMRI and SNPs data collected by MCIC.
- 183 subjects: 79 SZ patients (age: 34 ± 11, 22 females)
  104 healthy controls (age: 32 ± 11, 44 females)
- Follow the same preprocessing procedures as in Lin et al. [5],
- Implemented a multiple t-test between case and control for SNP and set significance level $p = 0.0001$.
- 116 regions of interests (ROIs) and 248 SNPs for analysis.

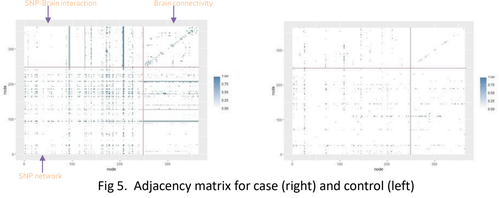


Fig 5. Adjacency matrix for case (right) and control (left)

Table 1. Identified aberrant SNP interactions

SNP Index (gene name)	SNP Index (gene name)
rs2253907*	rs10917813
rs2523638*	rs2278683 (DARS*)
rs11136067	rs12351317
rs1898413 (RORA-AS1*)	rs17702722
rs11635975 (RORA-AS1*)	rs10880976 (LOC100288798*)

Note: • denotes the gene has been reported to be related to SZ.

Table 2. Top 10 significant pairs of aberrant SNP-ROI interaction

SNP INDEX	ROI	P-VALUE (fdr corrected)
rs11635975	Olfactory cortex (R)	8.303586e-37
rs2523638	Parahippocampus (R)	1.498625e-35
rs2253907	Putamen(R)	3.819558e-35
rs11635975	Hippocampus(R)	4.945131e-35
rs1898413	Superior temporal gyrus(L)	4.945131e-35
rs1898413	Cingulate gyrus, mid part (L)	4.945131e-35
rs2523638	Angular gyrus (R)	1.270561e-34
rs2523638	Cerebellum 4 5(R)	8.795459e-33
rs2523638	Superior frontal gyrus, orbital(L)	8.795459e-33
rs2523638	Cerebellum 3(L)	8.795459e-33

Conclusions

- We developed a latent graphical copula model for mixed data that contains both continuous and discrete types.
- The simulation results demonstrate that our method is stable over different settings and performs well in detecting graph structures.
- The imaging genomics study identified novel interactions between ROIs and genes, which helps our understanding of genomic mechanisms underlying SZ disease.

Acknowledge

The work is funded by NIH (R01GM109068, R01MH104680, R01MH107354), and NSF (#1539067).

Contact

Aiying Zhang
Tulane University
Email: azhang4@tulane.edu
Phone:(352)281-9166
Wechat:

Major References

1. H. Liu, J. D. Lafferty, and L. A. Wasserman, "The nonparanormal: semiparametric estimation of high dimensional undirected graphs," J. Mach. Learn. Res., vol. 10, pp. 2295–2328, 2009.
2. F. Liang, Q. Song, and P. Qiu, "An equivalent measure of partial correlation coefficients for high dimensional gaussian graphical models," J. Amer. Statist. Assoc., vol.110, pp. 1248–1265, 2015.
3. M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," Biometrika, vol. 94, pp.19–35, 2007.
4. N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the lasso," Annals of Statistics, vol. 34, pp. 1436–1462, 2006.
5. D. Lin, V. D. Calhoun, and Y.-P. Wang, "Correspondence between fmri and snp data by group sparse canonical correlation analysis.," Medical image analysis, vol. 18(6), pp. 891–902, 2014.