Generalised Linear Stochastic Blockmodelling and Inference in Multi-Subject Networks

Dragana M. Pavlovic[1], Emma K. Towlson[3], Soroosh Afyoni[2], Petra E. Vértes[4], Edward T. Bullmore[4,5], Thomas E. Nichols[1,2]

[1,2]University of Warwick, Dept. of Statistics & Warwick Manufacturing Group, Coventry, UK; [3]University of Cambridge, Dept. of Physics, Cavendish Laboratory, Cambridge, UK; [4]University of Cambridge, Brain Mapping Unit, Dept. of Psychiatry, Cambridge, UK; [5]GlaxoSmithKline, Clinical Unit Cambridge, Addenbrooke's Hospital, Cambridge, UK.

Research Problem

There is a great interest in models that can decompose brain functional or structural networks into Q sub-groups (blocks) of functionally similar nodes. However, such models are suitable only for a single network analysis and their application in multi-subject networks poses several unresolved problems including:

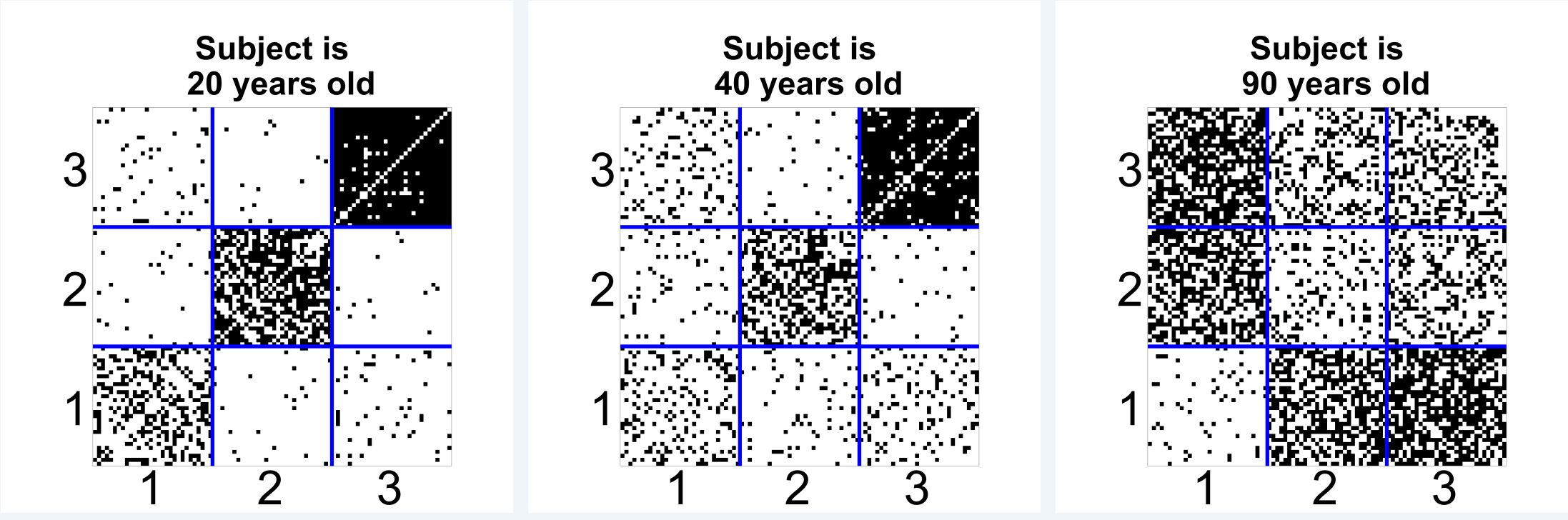
1. How we can estimate a common network decomposition in multi-subject data, while accounting for between subject variability?
2. How we can use such network decomposition to infer differences between populations (e.g., patients vs. controls), or effects of covariates?

In this work, we address these problems by developing a stochastic block model (SBM) for multi-subject binary network data, that includes a logistic regression model within each block and block-to-block relationships. While others [4] have network regression approaches, they have been based on edge-varying covariates for a single network, instead of subject-varying for multi-subject data.

Contributions

The SBM [1] models edges as homogeneous Bernoulli random variables within and between blocks of nodes, estimating the number of blocks and their composition. We extend the SBM to account multi-subject data, while allowing for additional variability according to a logistic regression model. The key details are as follows.

1. The GL-SBM estimates a common network decomposition but allows for subject-wise variability in edge occurrence. In the illustration below, the variability between subjects is explained by an age covariate.



2. The GL-SBM allows us to construct tests with p-values from asymptotic theory (Wald test) or resampling procedures (Permutation test). Thus, for example, we can ask if there is a significant age effect in block (1,1)?
3. To ensure a good behaviour of the tests, when edge-counts are very rare or saturated, we use a Firth estimator [2].
4. We propose a two-stage estimation algorithm which combines the variational approximation and the Newton-Raphson optimisation.

GL-SBM

The main parameters in the GL-SBM are: $\boldsymbol{\alpha}$ (proportions of nodes in each of the Q blocks) and regression coefficients $\boldsymbol{\beta}_{ql}$, for $q,l = 1,\dots,Q$, each of which is a vector of length P. For adjacency matrix $X_k = ((X_{ijk}))$ and covariate values d_k, for subjects $k = 1,\dots,K$, the model is

$$\boldsymbol{Z}_i \sim Categorical(Q,\boldsymbol{\alpha}) \quad (1)$$

$$X_{ijk}|Z_{iq}=1,Z_{jl}=1 \sim Bernoulli(\pi_{qlk}) \quad (2)$$

$$\log\left(\frac{\pi_{qlk}}{1-\pi_{qlk}}\right) = \boldsymbol{d}_k^\top \boldsymbol{\beta}_{ql}, \quad (3)$$

where $\boldsymbol{Z}_i$ is the latent block-indicator variable of node i, and $pi_q lk$ are the subject-specific edge rates for block (q,l). The estimation is based on the two-stage algorithm, which combines the variational approximation and Newton-Raphson algorithm with Firth regularisation on $\boldsymbol{\beta}_{ql}$, and estimate of Q is found with the Integrated Classification Likelihood. (See details in be [5, 6]).

Data

We consider a multi-subject study with 13 controls and 12 patients with schizophrenia [3]. The individual functional networks were derived from the resting-state fMRI time series (297 nodes) and, at scale 2 of discrete wavelet transform (0.06-0.125 Hz). Correlations were transformed to Fisher-Z scores and threshold at 5% FDR, producing a binary network for each subject.

We consider covariates of age, premorbid IQ and per-subject motion in the scanner, as well as a patient/control effect. The covariates are column-wise assigned into the design matrix $\boldsymbol{D}$, so that the first two columns are group intercepts. Also, $\boldsymbol{D}$ is centred about the mean covariate values.

Simulations Settings and Results

We simulated data for $K = 10$ subjects, with networks of 50, 100 and 500 nodes. We set $Q = 3$ and study the effect of block sizes under three proportion designs: Balanced ($\alpha = (0.33,0.33,0.33)$), Mildly Unbalanced ($\alpha = (0.6,0.3,0.1)$) and Unbalanced ($\alpha = (0.7,0.3,0.1)$). Each network is simulated according to the connectivity rates PI1-8 (see Fig. 1 (a)). Also, we consider the cases when there is no age effect ($\beta_{ql} = 0$) and decreasing age effect ($\beta_{ql} = -0.025$). We use the notation $n50\_0$ to indicate network with 50 nodes and no age effect while $n50\_0025$ indicate network with 50 nodes and age effect of -0.025. For each combination of parameters, we generated 1000 network realisations. Except for nearly or totally unidentifiable block structure (PI1-2 and PI5), the recovery of true block structure was excellent (Fig. 1 (b)), as was the control of false positives (Fig. 1 (c) and (d)).

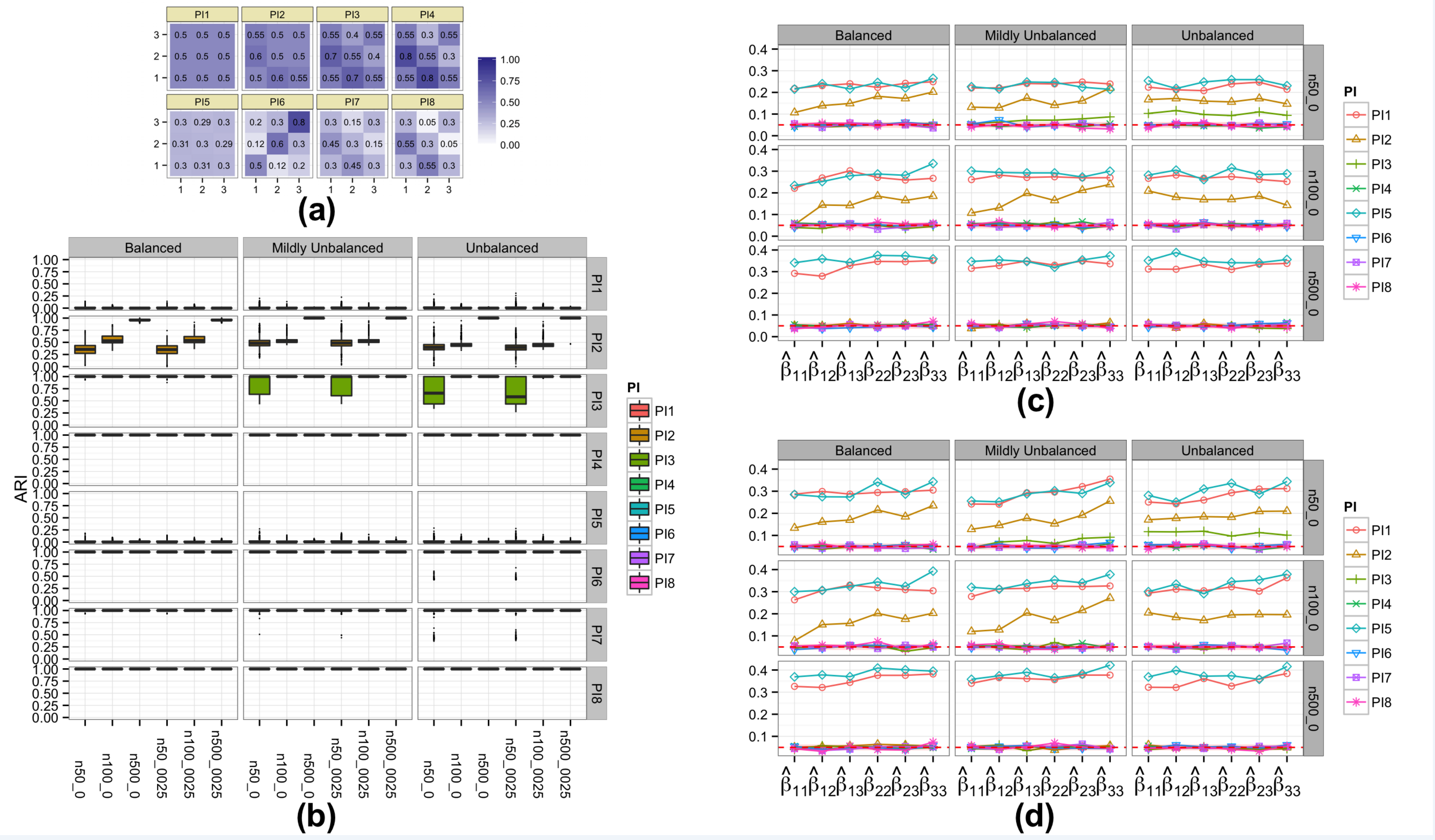


Figure 1: **(a)** Design of connectivity structures PI1-8. **(b)** Recovery of true node assignments measured with Adjusted Rand Index (ARI). **(c)** Wald test-False Positive Rates (FPR) for $\hat{\boldsymbol{\beta}}_{ql}$. **(d)** Permutation test - FPR for $\hat{\boldsymbol{\beta}}_{ql}$.

Real Data Results

The GL-SBM discovered well-known resting-state networks (Fig. 2 (a)), as well as strong patent/control and age effects (Fig. 2 (b)&(c)).

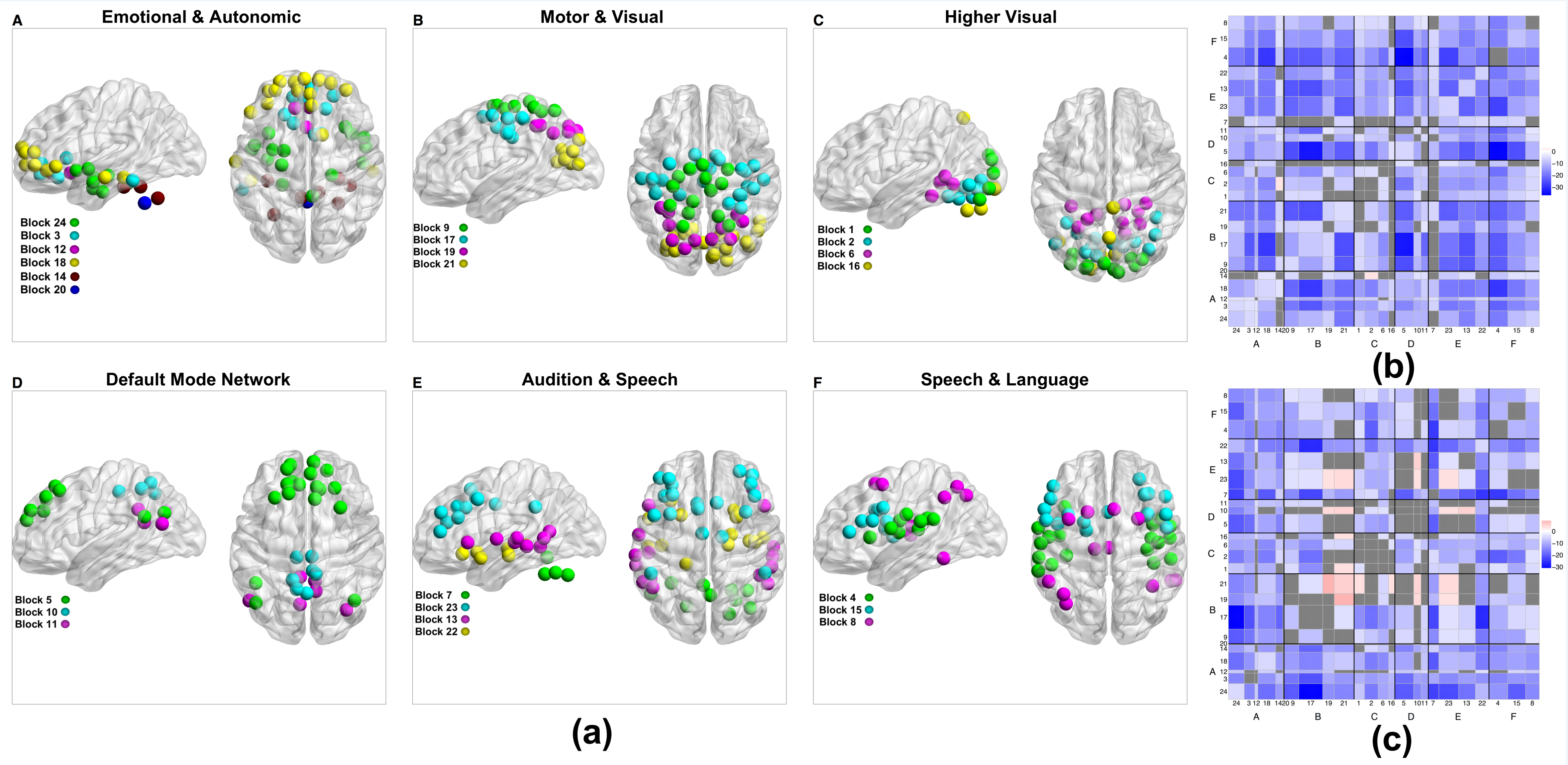


Figure 2: **(a)** Anatomical locations of individual blocks. **(b)** Bonferroni thresholded (5%) Wald score image of the intercepts (Patients vs. Controls). **(c)** Bonferroni thresholded (5%) Wald score image of common age effect.

Conclusions

We have developed a novel stochastic block regression model for multi-subject binary network data. In real data applications, the GL-SBM identified anatomically and functionally plausible blocks, as well as differences in connectivity between patients and controls and their variation with age (`http://warwick.ac.uk/tenichols/ohbm`).

References

[1] Daudin, Picard, Robin *A mixture model for random graphs*, Statistics and computing, (2008).

[2] Firth, David. *Bias reduction of maximum likelihood estimates*, Biometrika 80.1 (1993): 27-38.

[3] Lynall, Mary-Ellen, et al. *Functional connectivity and brain networks in schizophrenia* The Journal of Neuroscience 30.28 (2010): 9477-9487.

[4] Mariadassou, Mahendra, Stéphane Robin, and Corinne Vacher. *Uncovering latent structure in valued graphs: a variational approach*. The Annals of Applied Statistics 4.2 (2010): 715-742.

[5] Pavlovic, Vértes, Bullmore, Schafer, and Nichols, *Stochastic blockmodelling of the modules and core of the Caenorhabditis elegans connectome*, PLOS ONE (2014)

[6] Pavlovic, D. M., Vértes, P. E., Towlson E. K., Afyouni S., Bullmore, E. T. & Nichols, T. E. (in submission). *Stochastic Blockmodelling and Inference in Multi-Subject Networks with Mixture Models*, Computational Statistics and Data Analysis (2015)