Short review for
"Latent Network Estimation and Variable Selection for Compositional Data Via Variational EM"

Jiaxin Hu

This work proposed a Bayesian method to simultaneously select the covariates with large effects to the objects (e.g. microbes, genes) and estimate the network interactions among the objects. Compared with previous works, proposed method aims for the count data and tackles two problems in one shot.

The framework assumes there is a latent variable Z_i that impacts the prior distribution of the count data X_i for samples $i \in [n]$. Specifically, the prior distribution of Z_i is

$$Z_i|B_0, B, M_i, \Omega \sim \mathcal{N}(B_0 + M_i B, \Omega^{-1}),$$

where B is the effect of covariates M_i to Z_i and Ω inverse covariance matrix among the columns of Z. The sparsity in B facilities the feature selection in M and the Ω can be interpreted as the network interactions among the columns in Z, which corresponding to the target objects; i.e., columns in X. The prior distributions for the parameters B_0, B, Ω and the prior of X_i related to Z_i are carefully proposed. A variational inference based EM algorithm is also proposed for the computation.

Simulation results indicate that the variable selection results of proposed method is not overly sensitive to parameters in the priors and the accuracy of network estimation is improved due to the robust selection and the simultaneous procedure.

Comments by me:

1. As the paper argues, the reason for the network improvements comes from the robust feature selection and the simultaneous algorithm. It is unknown whether the proposed method would improve the estimation when no covariate or no informative covariate is available.

2. The interpretation of the network interactions are pretty important for biological studies. I think it is not easy to interpret the network Ω as the gene/microbe correlation or regulatory network due to the complex hierarchical structure from Z to X. Also, the meaning of the latent variable Z is hard to interpret.

3. The dimensions of simulations and real microbe data are relatively small compared with the data sets in the genetics studies. Also, the number of covariates (e.g. SNP) may be much larger than the number of network objects (e.g. genes), but such cases are not included in the proposed method. The computation for such high dimensional cases is also challenging.

References