Project Proposal

Renat Sergazinov, Gozde Sert, Isaac Ray

Introduction

Gaussian graphical models allow for particularly simple learning as all the dependence structure is essentially summarized in the associated covariance matrix. In a recent paper by Dasgupta et al. (2022), the authors have shown that learning different covariance structure based on the extraneous covariates strictly improves on the results from the previous methods. However, experiments conducted in Helwig et al. (2022) demonstrate that despite outperforming the other methods in sensitivity, the proposed approach suffers from high false positive rates.

In this project, the goal is to investigate potential fixes for the high false positive rate issue. First, we summarize the approach proposed in Dasgupta et al. (2022) and then discuss the potential fixes. Let us introduce the following notation: $X \in R^{n \times p}$ is the data matrix, $Z \in R^{n \times k}$ is the matrix of extraneous covariates. The algorithm then roughly follows:

1. Fix $l \in \{1, \dots, n\}$.
2. Attach weights $w_l(z_i) = w(Z_{l,:}, Z_{i,:})$ to each $i \in \{1, \dots, n\}$, where $Z_{l,:} \in R^k$ is a vector of extraneous covariates and $w(Z_{l,:}, Z_{i,:})$ is a valid distance measure.
3. Fix $j \in \{1, \dots, p\}$.
4. Given the weights above, perform a weighted regression of $X_{:,j}$ onto $X_{:,-j}$ and extract the coefficient vector $\beta_j^l \in R^{p-1}$.
5. Repeat steps 1-4 for $l \in \{1, \dots, n\}$ and $j \in 1, \dots, p$.
6. Use collection of vectors $\{\beta_j^l\}_{j=1}^n$ to estimate the covariance matrix (undirected conditional dependency) among covariates $(X_{l,1}, \dots, X_{l,p})$ for an individual $l \in \{1, \dots, n\}$.

Potential Fixes: Clustering

We note that for step 4 usually a Bayesian regression with a spike-and-slab prior is used, which requires variational approximation. We hypothesize that the high false positive rate could be partially caused by re-using the same prior inclusion probability, π, for all the individuals $l \in \{1, \dots, n\}$. In this project, one of the ideas is to investigate the effect of using different prior inclusion probabilities for different individuals. The outline of our approach is as follows:

1. First, assume we have an oracle for both the number of different distributions from which the observed $Z_{l,:}$'s are drawn, a mapping for which $Z_{l,:}$'s came from which distribution, as well as what the prior inclusion probabilities should be as a function of the mapping. Given this oracle, we can choose perfectly π_l for each individual based on the corresponding extraneous covariates $Z_{l,:}$. Does the corresponding change in the posterior inclusion probability lead to fewer false positives?
2. If the above is effective at reducing false positives, can we effectively and efficiently learn from the data:
    1. Can we cluster $Z_{l,:}$ and learn number of clusters at the same time (perhaphs through CRP)?
    2. What are good choices for the prior inclusion probabilities given the number of clusters and cluster assignment?

Potential Fixes: Data Scaling

One observation that we potentially want to exploit is that in the package developed by Helwig et al. (2022) none of the covariate matrices are scaled. What will happen if we try various rescalings of X and or Z before doing inference? Maybe false positives are related to different covariates being on very different scales but with the same global prior inclusion probability. This idea is parallel to the one proposed above but deals with the same problem just from a different angle.

Potential Fixes: Tree Extension

In the Lie et al. (2010) paper, the CART approach to partitioning the covariate space might be effective if a more flexible model is used (rather than trees built from binary decision stumps, perhaps trees made from multivariate spanning decision trees).