Final Project Report: Feature selection performance of Grace-AKO and BIPnet, recent methods that leverage prior grouping information

Aidan Neher

2023-12-19

## Abstract

This report explores the feature selection performance of Grace-AKO and BIPnet, two recently developed methods leveraging prior grouping information. Effective feature selection is crucial in understanding associations with clinical outcomes and building predictive models. Grace-AKO integrates multiple knockoffs within a network-constrained penalty for controlling the false discovery rate in high-dimensional settings. BIPnet performs Bayesian integrative analysis using multiple omics data and clinical covariates. Both methods present distinct approaches to feature selection and prediction, with GRACE-AKO focusing on FDR control and BIPnet integrating information across multiple datasets. These distinct approaches deserve consideration when choosing an appropriate method for a given context.

## Problem Statement & Background

**Problem Statement:** Effective feature (i.e. variable) selection methods are crucial for understanding which features are associated with clinical outcomes and building efficient predictive models. GRACE-AKO for graph-constrained estimation (Grace) aggregates multiple knockoffs (AKO) with a network-constrained penalty. Bayesian Integrative Analysis and Prediction with network information (BIPnet) and without network information (BIP) performs feature selection and prediction jointly. Both GRACE-AKO and BIPnet are recently proposed methods (2022 and 2023 respectively) and, thus, their feature selection performance deserves further investigation.

**Background:** Grace-AKO and BIPnet leverage prior network information for this purpose. Grace-AKO integrates multiple knockoffs within a network-constrained penalty to control FDR in high-dimensional settings. GRACE-AKO is reported to have better performance than its originally proposed version GRACE in that the aggregation of multiple knockoffs affords increased stability in the results. The documented con of this method is its relatively poor statistical power.

In contrast, BIPnet performs Bayesian integrative analysis and prediction using multiple omics data and allows for clinical covariates. This method is distinct from GRACE-AKO in that it goes about feature selection by way of a Bayesian factor analysis. The factor analysis portion is meant to reduce the dimensionality of the multiple datasets that it aims to integrate inference across. These methods are distinct in that GRACE-AKO provides finite-sample FDR control, whereas BIPnet integrates information across multiple datasets. Thus, we performed a simulation study meant to determine the strengths and weakness of each method. Furthermore, we use elastic net (elnet) as a generally accepted method for feature selection as a benchmark against GRACE-AKO and BIPnet.

**Research Questions:** How do Grace-AKO or BIPnet differ in feature selection in relation to each other and also elnet, a generally accepted method?

## Data Simulation

We iterations, we simulated data to compare GRACE-AKO, GRACE, BIPnet, BIP, and elnet methods. For this simulation study, three parameters were varied: , the number of observations in each dataset, , the total number of features in , and , the number of groups of features in the first dataset considered important to the outcome.

Since BIPnet is an integrative method, we simulate two datasets . Each had n rows representing an observation and features. To start, we simulated i.i.d multivariate normal groups of size 10 features, with 1 representing a primary feature (analogous to a transcription factor) and 9 representing secondary features. Each network’s 10 features had mean zero. The correlation between a primary and each of its 9 secondary features is 0.4, and the correlation between each of the 9 secondary features is represented by a 9×9 compound symmetric matrix with correlation 0.16. These 10 groups were split evenly across the two datasets . With 5 groups of 10 features associated with each dataset, we simulated i.i.d. Normal(0,1) singletons to ensure each dataset was of size by . This corresponds to an overall covariance matrix:

is block diagonal with 10 blocks of size 10, between-block correlation 0 and within each block, the correlation between a primary and each of its 9 secondary features is 0.4, and the correlation between each of the 9 secondary features is represented by a 9×9 compound symmetric submatrix with correlation 0.16.

To simulate the continuous outcome , we firstly defined to be the number of groups important to the outcome . Also, we let of length 10. Then, from , the number of groups important to the outcome in the first dataset, we formed . This associated non-zero repeats with the number of groups considered important in each dataset and zero values with the features considered unimportant. From that, we generated from where is from i.i.d. .

## Methods

On a given dataset, we fit all five methods. GRACE-AKO, GRACE had FDR control fixed at 0.10. GRACE-AKO, GRACE, and elnet had additional hyperparameters set from the best MSE performing model in 10-fold cross-validation. BIPnet and BIP require specification of the hyperparameter , the number of latent factors. This was fixed to , which was the value used by the method’s authors. Additionally, for GRACE-AKO and GRACE, the weights attributed to the connections between groups were fixed to 1. Furthermore, for BIPnet analysis, and were kept as separate matrices. For GRACE-AKO and elastic net, we concatenated these datasets to form .

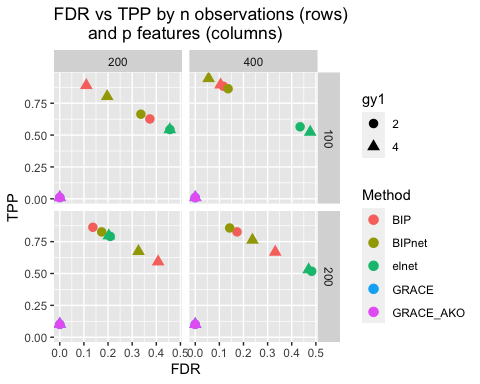
Finally, for each simulation evaluated the following feature selection performance metrics: False Discovery Rate (FDR), True Positive Probability (TPP), and the number of features selected by a given model (n\_features\_selected). These are used to assess differences in feature selection performance.

All analyses will be conducted using the R statistical software, version 4.2.0, and relevant packages.

## Results

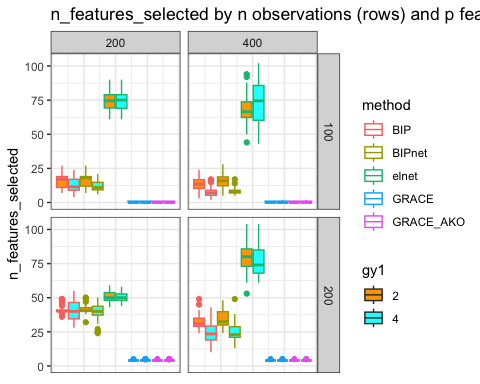
Firstly, we observe that GRACE and GRACE-AKO performed strict FDR control at the cost of a low TPP in these simulated scenarios (Figure 1). Furthermore, we observe that BIP and BIPnet differ across , demonstrating that these methods have a higher TPP and lower FDR when the number of groups relevant to the outcome is spread across the two datasets and not focused in one specific one. Yet, BIP and BIPnet had the highest TPP relative to all other methods when , suggesting that the low-dimensional representation of multiple datsets helped to highlight the shared structure between the datsets.Finally, elnet had a larger FDR and a lower TPP in all combinations of and except for , which suggests sensitivity to a dimensional imbalance.

Figure 1. The mean FDR and mean TPP stratified by n observations (rows) and p features (columns).



Secondly, we notice that elnet selected the most features across all simulated scenarios (Figure 2). This relative feature greediness in these contexts is likely driving its relatively poor TPP and FDR performance in relation to BIP and BIPnet. The number of features selected by BIP and BIPnet is most sensitive to when and .

Figure 2. Number of features selected by each method stratified by n observations (rows) and p features (columns).



## Discussion & Conclusions

Grace-AKO and BIPnet, recently proposed methods that incorporate prior grouping information, show promising results in terms of their ability to control FDR and perform integrative feature selection respectively. However, their use in practice comes with challenges.

Firstly, the a priori definition of grouping structure can be helpful for features selection, but each method goes about including such information in different ways. Grace-AKO accommodates edge weights in networks, unlike BIPnet, which uses binary indicators for whether or not a feature is in the same group as another. For these simulation studies, we set GRACE-AKO (and GRACE)’s weights to 1, but setting uniform Grace-AKO network edge weights might be limiting a strength that this method has over BIPnet, BIP, and elnet. This is especially clear when considering the interpretation of these binary indictors with 1 suggesting that we know two features are grouped together with certainty when in reality we likely do not have such certainty.

On considering the design of this simulation study, defining a plausible covariance matrix for the data simulation that can relate to both methods was difficult. This is especially true since BIPnet is a method that integrates information across multiple datasets. Also, additional parameters were fixed at values that might not be used in practice. For example, for BIP and BIPnet, we defined feature selection as an estimated marginal posterior probability of feature selection in at least one latent factor above 0.5. This threshold might be modified on considering the tradeoff of false discoveries in an applied setting. Another example is in the fixing of the number of latent factors to 4, which is the hyperparameter value used by the authors. The BIP package does not have easy to follow guidance on how to set this parameter or functionality in the package to do so. This is in contrast to the GraceAKO package, which has built-in cross-validation for hyperparameter selection.

In conclusion, BIPnet and GRACE-AKO are promising for feature selection. When finite-sample FDR is of concern, GRACE-AKO is worth exploration. Yet, when integration of multiple datsets for feature selection and prediction is the context, BIPnet is worth investigation. There might have been more similar methods to compare. Yet, comparing 2 seemingly similar on the surface, but, ultimately, quite different methods gave insight into their nuances.

## References

1. Tian, P., Hu, Y., Liu, Z. & Zhang, Y. D. Grace-AKO: a novel and stable knockoff filter for variable selection incorporating gene network structures. BMC Bioinformatics 23, 1–16 (2022).
2. Chekouo, T. & Safo, S. E. Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease. Biostatistics 24, 124–139 (2021).