Project Results

Isaac Ray, Renat Sergazinov, Gözde Sert

Background

Problem

Despite this model demonstrating superior sensitivity to true dependence relations than competing methods, it suffers from lower specificity. Compared with competing methods from the `mgm` (Haslbeck and Waldorp 2020) and `JGL` (Danaher 2018)packages, the specificity gets substantially worse as the number of features increases. In the case of analyzing gene expression data, this could lead to worse outcomes than a less sensitive and more specific model since the cost of carrying out experiments which show a lack of a predicted relation may be very expensive. Further, validating a true but relatively weak relationship may not be desirable considering the cost. Ideally we want to **increase the specificity of the model without substantially hurting the model's sensitivity and speed**.

Attempted Solutions

We'll look at a variety of different modifications to the algorithm in order of least to most extensive changes necessary, and use the `covdepGE` package(Helwig et al. 2022) in order to generate data for simulation studies. For some of these solutions, the package's functions will be used without modification and changes will occur outside of the inference algorithm. Otherwise, any changes to the functions will be explicitly noted.

In general, we expect that the underlying cause of the specificity issue is due to having a common prior inclusion probability π across every spike-and-slab regression being performed despite the varying values of Z (and potentially X). We may expect that for certain values of Z, we have a different belief about whether variables in X are related. We'll approach this from 2 angles; first by trying to modify our variables and covariates in such a way as to make a common π a more appropriate choice, and then by modifying the algorithm to allow for multiple π values to be specified either a priori or as a function $\pi(X, Z)$ through something like clustering.

Feature Scaling Changes

The first approach will be to use a different or additional approach to feature scaling on X and/or Z in order to try and make a singular prior inclusion probability more appropriate.

Existing Feature Scaling through Normalization

Currently, the default behavior in the `covdepGE` function is to perform a columnwise Z-score Normalization on Z and a columnwise 0 centering on X. For brevity we'll denote this procedure by "normalization". The baseline performance under this scheme is given below. All experiments were run under 4 different setups each having different values for p and n, and data simulated using the `generateData` function. To assess sensitivity and specificity, we'll examine the number of false positives per sample and number of false negatives per sample across 100 replications of each simulation setup. So, in all cases lower numbers are desirable. First, we'll look at the baseline performance of the existing function with no changes to the default behavior.

p	n	Baseline Mean (sd)
5	90	0.34 (0.54)
15	90	0.82 (0.94)
25	150	1.29 (0.98)
50	150	4.36 (2.22)

Table 1: False **positives** per sample - Normalized Z, Centered X

p	n	Baseline Mean (sd)
5	90	0.99 (0.53)
15	90	1.41 (0.69)
25	150	0.91 (0.38)
50	150	1.19 (0.48)

Table 2: False **negatives** per sample - Normalized Z, Centered X

Max-Min Feature Scaling

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.34 (0.54)	0.81 (0.63)
15	90	0.82 (0.94)	4.08 (1.44)
25	150	1.29 (0.98)	11.89 (2.5)
50	150	4.36 (2.22)	19.86 (3.59)

Table 3: False **positives** per sample - Max/Min Scaled Z

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.99 (0.53)	1.14 (0.49)
15	90	1.41 (0.69)	2.35 (0.78)
25	150	0.91 (0.38)	1.71 (0.7)
50	150	1.19 (0.48)	2.42 (0.71)

Table 4: False **negatives** per sample - Max/Min Scaled Z

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.34 (0.54)	0 (0)
15	90	0.82 (0.94)	0 (0)
25	150	1.29 (0.98)	0 (0)
50	150	4.36 (2.22)	0 (0)

Table 5: False **positives** per sample - Max/Min Scaled X

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.99 (0.53)	4.67 (0)
15	90	1.41 (0.69)	4.67 (0)
25	150	0.91 (0.38)	4.67 (0)
50	150	1.19 (0.48)	4.67 (0)

Table 6: False **negatives** per sample - Max/Min Scaled X

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.34 (0.54)	0 (0)
15	90	0.82 (0.94)	0 (0)
25	150	1.29 (0.98)	0 (0)
50	150	4.36 (2.22)	0 (0)

Table 7: False **positives** per sample - Max/Min Scaled X and Z

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.99 (0.53)	4.67 (0)
15	90	1.41 (0.69)	4.67 (0)
25	150	0.91 (0.38)	4.67 (0)
50	150	1.19 (0.48)	4.67 (0)

Table 8: False **negatives** per sample - Max/Min Scaled X and Z

Max-Min + Normalization Feature Scaling

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.34 (0.54)	0.24 (0.47)
15	90	0.82 (0.94)	0.63 (0.83)
25	150	1.29 (0.98)	1.34 (0.99)
50	150	4.36 (2.22)	4.08 (1.89)

Table 9: False **positives** per sample - Max/Min Scaled then Normalized Z, Centered X

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.99 (0.53)	1.03 (0.45)
15	90	1.41 (0.69)	1.53 (0.65)
25	150	0.91 (0.38)	0.95 (0.46)
50	150	1.19 (0.48)	1.15 (0.44)

Table 10: False **negatives** per sample - Max/Min Scaled then Normalized Z, Centered X

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.34 (0.54)	0.26 (0.47)
15	90	0.82 (0.94)	0.61 (0.77)
25	150	1.29 (0.98)	1.34 (0.97)
50	150	4.36 (2.22)	4.07 (1.84)

Table 11: False **positives** per sample - Normalized Z, Max/Min Scaled then Centered X

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.99 (0.53)	1 (0.47)
15	90	1.41 (0.69)	1.48 (0.63)
25	150	0.91 (0.38)	0.92 (0.44)
50	150	1.19 (0.48)	1.12 (0.43)

Table 12: False **negatives** per sample - Normalized Z, Max/Min Scaled then Centered X

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.34 (0.54)	0.24 (0.44)
15	90	0.82 (0.94)	0.61 (0.77)
25	150	1.29 (0.98)	1.35 (0.98)
50	150	4.36 (2.22)	4.07 (1.82)

Table 13: False **positives** per sample - Max/Min Scaled then Centered X and then Normalized Z

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.99 (0.53)	1.02 (0.46)
15	90	1.41 (0.69)	1.47 (0.65)
25	150	0.91 (0.38)	0.92 (0.44)
50	150	1.19 (0.48)	1.11 (0.43)

Table 14: False **negatives** per sample - Max/Min Scaled then Centered X and then Normalized Z

Multiple Prior Inclusion Probabilities

Oracle or Informative Prior

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.34 (0.54)	0.45 (0.62)
15	90	0.82 (0.94)	1.25 (0.97)
25	150	1.29 (0.98)	3.24 (1.39)
50	150	4.36 (2.22)	9.16 (2.42)

Table 15: False **positives** per sample - Multiple PIP with Oracle Clustering

p	n	Baseline Mean (sd)	Mean (sd)
5	90	0.99 (0.53)	1.09 (0.58)
15	90	1.41 (0.69)	1.53 (0.66)
25	150	0.91 (0.38)	0.99 (0.42)
50	150	1.19 (0.48)	1.01 (0.36)

Table 16: False **negatives** per sample - Multiple PIP with Oracle Clustering

Prior Inclusion through Covariate Clustering

p	n	clusts	Baseline Mean (sd)	Mean (sd)
5	90	2	0.34 (0.54)	0.47 (0.66)
5	90	3	0.34 (0.54)	0.47 (0.68)
5	90	6	0.34 (0.54)	0.48 (0.64)
15	90	2	0.82 (0.94)	1.07 (0.98)
15	90	3	0.82 (0.94)	1.21 (0.99)
15	90	6	0.82 (0.94)	1.41 (1)
25	150	2	1.29 (0.98)	2.71 (1.37)
25	150	3	1.29 (0.98)	3.2 (1.32)
25	150	6	1.29 (0.98)	3.54 (1.36)
50	150	2	4.36 (2.22)	7.76 (2.3)
50	150	3	4.36 (2.22)	8.78 (2.43)
50	150	6	4.36 (2.22)	10.14 (2.46)

Table 17: False **positives** per sample - Multiple PIP with Hierarchical Clustering

p	n	clusts	Baseline Mean (sd)	Mean (sd)
5	90	2	0.99 (0.53)	1.08 (0.61)
5	90	3	0.99 (0.53)	1.09 (0.6)
5	90	6	0.99 (0.53)	1.09 (0.61)
15	90	2	1.41 (0.69)	1.53 (0.69)
15	90	3	1.41 (0.69)	1.51 (0.67)
15	90	6	1.41 (0.69)	1.52 (0.67)
25	150	2	0.91 (0.38)	1 (0.42)
25	150	3	0.91 (0.38)	0.99 (0.43)
25	150	6	0.91 (0.38)	0.99 (0.42)
50	150	2	1.19 (0.48)	1.06 (0.37)
50	150	3	1.19 (0.48)	1.04 (0.37)
50	150	6	1.19 (0.48)	1 (0.37)

Table 18: False **negatives** per sample - Multiple PIP with Hierarchical Clustering

Prior Inclusion at the Individual Level

Danaher, Patrick. 2018. "JGL: Performs the Joint Graphical Lasso for Sparse Inverse Covariance Estimation on Multiple Classes." https://CRAN.R-project.org/package=JGL.

Haslbeck, Jonas M. B., and Lourens J. Waldorp. 2020. "{Mgm}: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data" 93. https://doi.org/10.18637/jss.v093.i08.

Helwig, Jacob, Sutanoy Dasgupta, Peng Zhao, Bani Mallick, and Debdeep Pati. 2022. "covdepGE: Covariate Dependent Graph Estimation." https://CRAN.R-project.org/package=covdepGE.