Investigation of MMPC and MMHC Algorithms on Microscopy Data

Cameron R. Wolff

Supervised by Lauren M. Sanders, Ryan T. Scott, and Robert J. Reynolds

September 21, 2023

Abstract

In the following experiments, microscopy data from the OSDR GLDS datasets is run through the MMPC and MMHC bayesian network learning algorithms. The goal of the following experiments is to derive knowledge from an empirical dataset, specifically in regards to deriving correlations between astronaut health risks. The value of such graphs would be in creating a new metric to measure expert generated astronaut health DAGs against.

Algorithms

MMPC

The Min Max Parents and Children (MMPC) algorithm is a constraint-based algorithm to estimate the underlying structure of a Directed Acyclic Graph (DAG). MMPC is a local learning algorithm that starts by selecting a random variable T from the dataset. MMPC then identifies parents and children of T using Pearson's Correlation Conditional Independence test. This process is then repeated by substituting every variable from the dataset until the entire skeleton of the Bayesian Network (BN) is learned.

MMHC

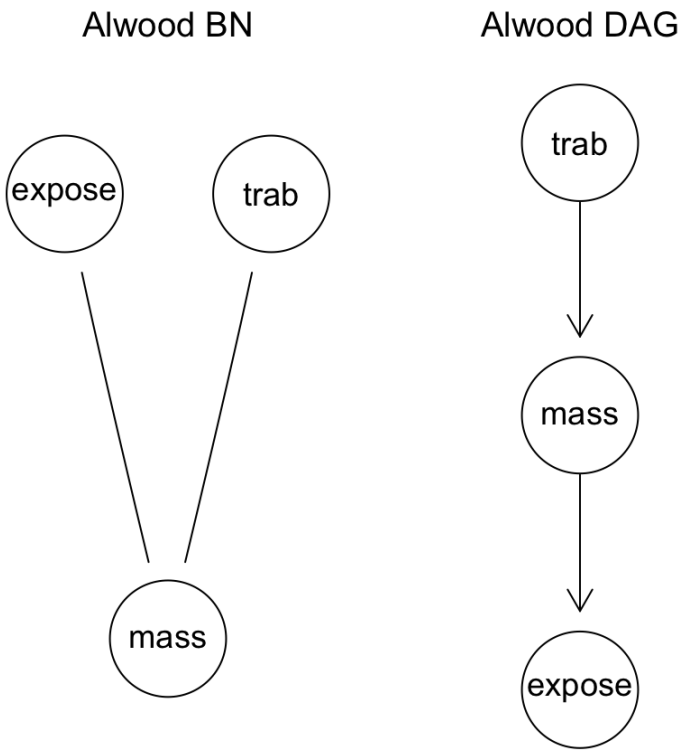
The Min Max Hill Climb (MMHC) algorithm is a 2 Phase Restricted Maximization (hybrid algorithm) and an extension of the MMPC algorithm. First, the MMPC algorithm is run to construct a BN skeleton. Then, a Greedy Hill-Climbing Search (HC) is performed to add directionality. HC Starts by considering an empty graph, then recursively attempts to make additions, deletions, or direction reversals that lead to the largest BDeu score. The HC in MMHC is improved over a traditional HC because it only considers an edge if it was discovered by MMPC.

Experiments on Small Datasets

The following experiments were run on small datasets with between 40 and 200 rows. On each graph, the type of graph and dataset used is noted, with BNs being undirected bayesian networks generated by the MMPC algorithm, and DAGs being directed acyclic graphs generated by the MMHC algorithm. Modifications to the dataset are noted on the graph

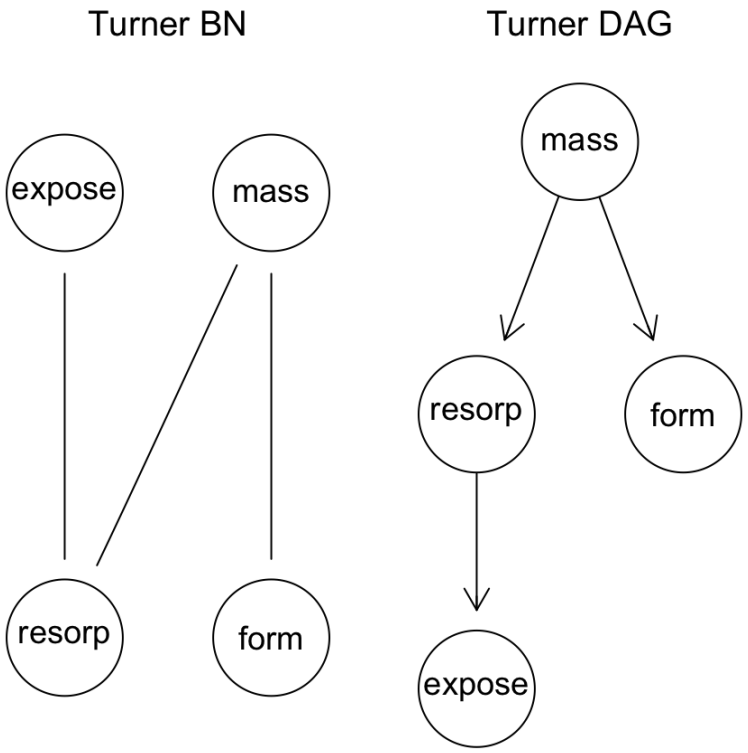
Allwood

The results from the Allwood dataset were identical between the MMPC and MMHC algorithms. Both algorithms deduced that the trabecular feature was correlated to mass which was correlated to exposure. However, MMHC's addition of directionality incorrectly put exposure as an effect, when effect was a parameter of the experiment.



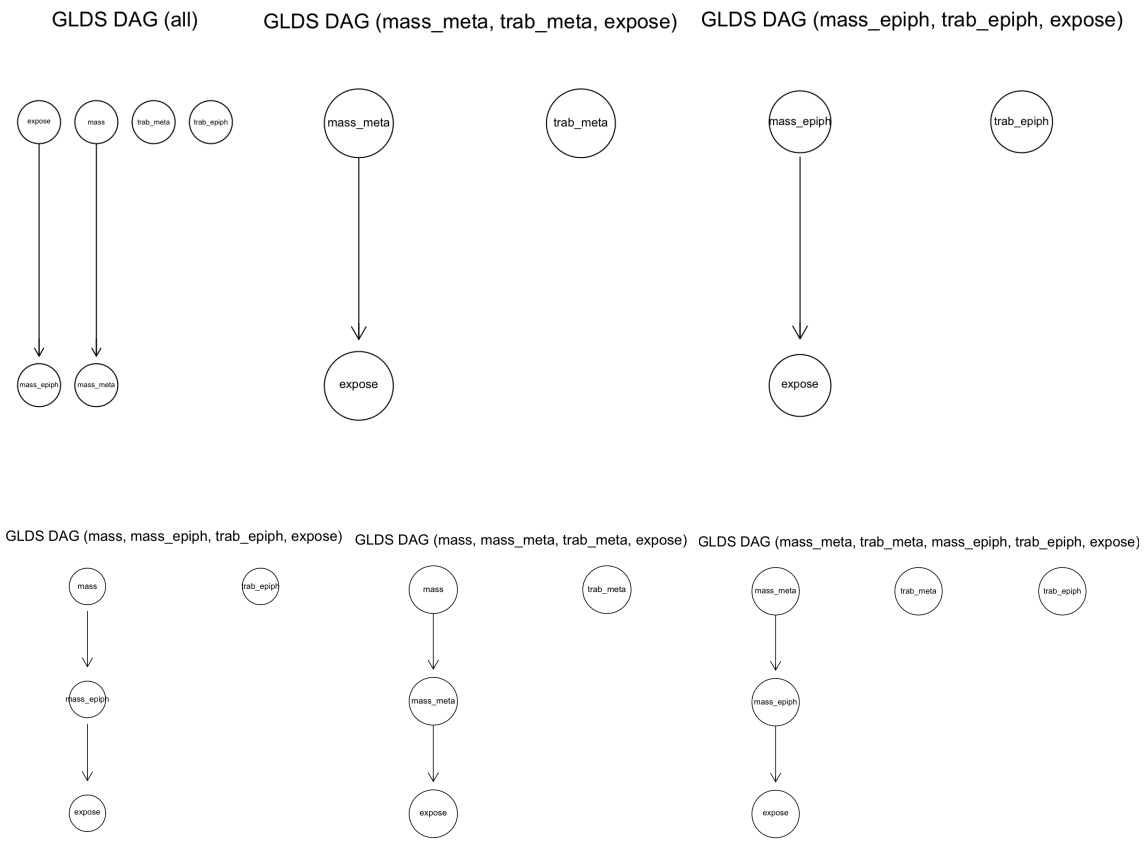
Turner

For the Turner dataset, results were identical between the MMPC and MMHC output. However, similar to the Allwood experiment MMHC makes connections in incorrect directions, stating resorption is the cause of exposure, rather than the opposite.



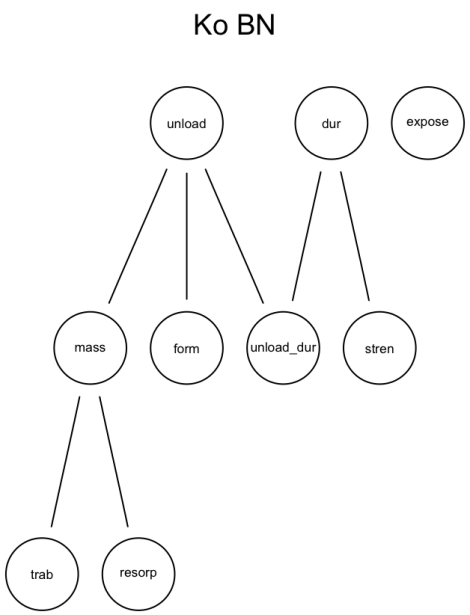
GLDS

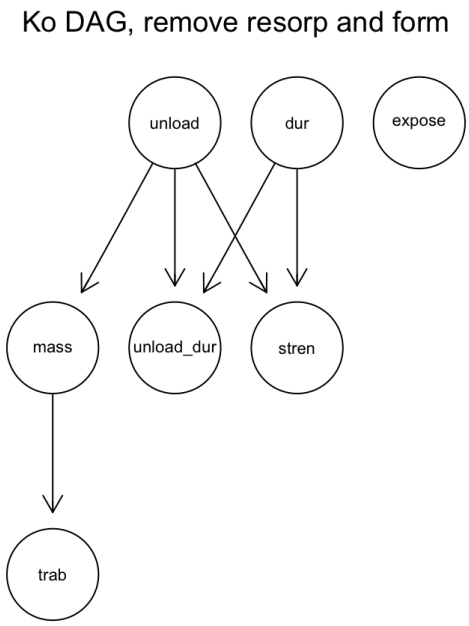
For the original full (all) dataset run on the GLDS dataset, the output from MMPC and MMHC were identical aside from directionality. The following experiments show similar results, with little difference between the two algorithms. Each past the (all) example is a different subset of the dataset, with the specific subset noted on the visualization. Interesting, in all cases but the full dataset exposure is noted as a cause of one of the mass features. In the (all) case, exposure is noted as the cause of mass epiphysis, however where MMPC draws correlation between mass and mass metaphysis, MMHC derives incorrect causality from mass to mass metaphysis.

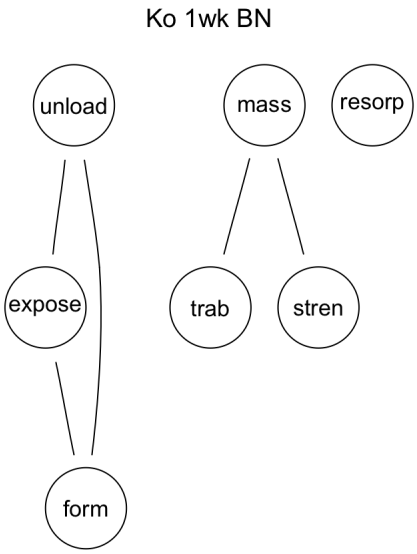


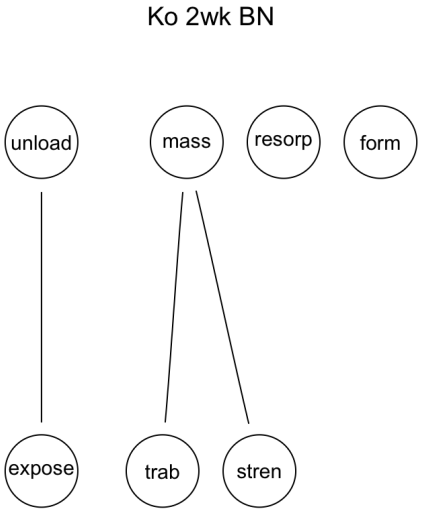
Ko

For the following experiments, while the full dataset was given to the MMPC algorithm, modifications had to be made for the MMHC dataset, as MMHC imposes some constraints as to the types of data it will accept. Along with this modification, other subsets are taken at different time intervals and fed to both algorithms, each with varying levels of similarity and difference between the two algorithms.


Ko BN
unload
dur
expose
mass
form
unload_dur
stren
trab
resorp


Ko DAG, remove resorp and form
unload
dur
expose
mass
unload_dur
stren
trab


Ko 1wk BN
unload
mass
resorp
expose
trab
stren
form


Ko 2wk BN
unload
mass
resorp
form
expose
trab
stren

Experiment on Large Dataset

Procedure

This experiment ran using the Costes 2021 dataset from GLDS-366. Repeated data entry columns 29 and 30 were omitted from the dataset, along with duplicate rows and rows containing NA data. The chromosome column was one-hot encoded to allow for variables to be correlated to specific chromosomes. The dataset had a dimension of 56,129 rows and 53 columns after preprocessing.

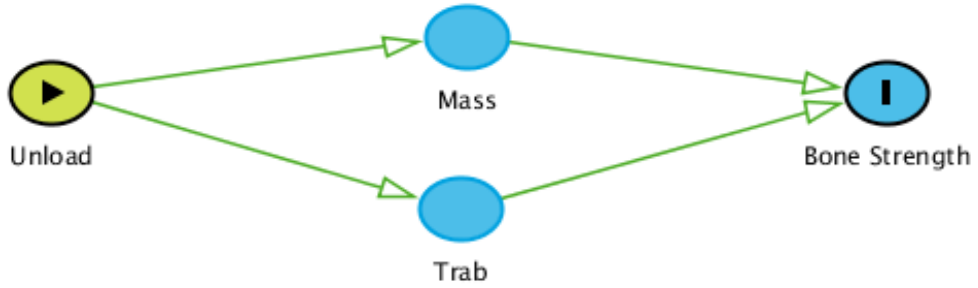
Output

After multiple attempts to run the MMPC and MMHC algorithms on the Costes dataset, the algorithms were unable to process the volumes of data. This process was attempted multiple times, with the longest training session taking over 78 hours. It is unknown whether this failure to process the dataset was a failure of the algorithms to handle a dataset of this magnitude, or a result of the hardware running the algorithm. All algorithms were run on a Ryzen 5 5600X CPU with 32GB of RAM

Experiment on Ko for Comparability

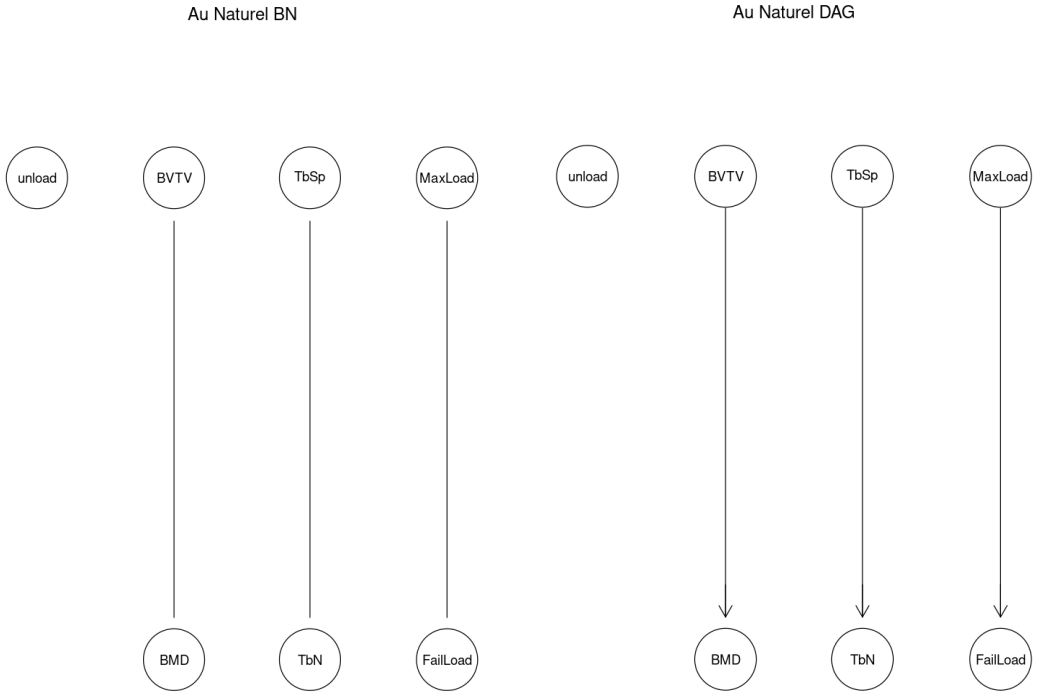
Procedure

For this experiment, only the main (n=140) file of the Ko dataset was used, since the "histo" half is too limited, and not a correct measurement. The Ko dataset was subset to only include observations that were either 2 weeks or 4 weeks in duration, then subset again to include only the animals that were PWB100 (no tail suspension/controls) or PWB20 (80% unloaded). Finally, unload was changed to a boolean value, 0 for the PWB100 group, and 1 for the PWB20 group. The dataset was then subset into 2 groups: "Au naturel", (unload, BVTV, BMD, TbSp, TbN, MaxLoad, and FailLoad), and "PCA-synthetic" (unload, PCA from BVTV and BMD, PCA from TbSp and TbN, and PCA from MaxLoad and FailLoad). Ideal results for the Au Naturel scenario would mean a parallel structure, with BVTV and BMD would both get arrows from unload and both have arrows to MaxLoad and FailLoad. For the the PCA-synthetic scenario, the ideal DAG would looks as such:

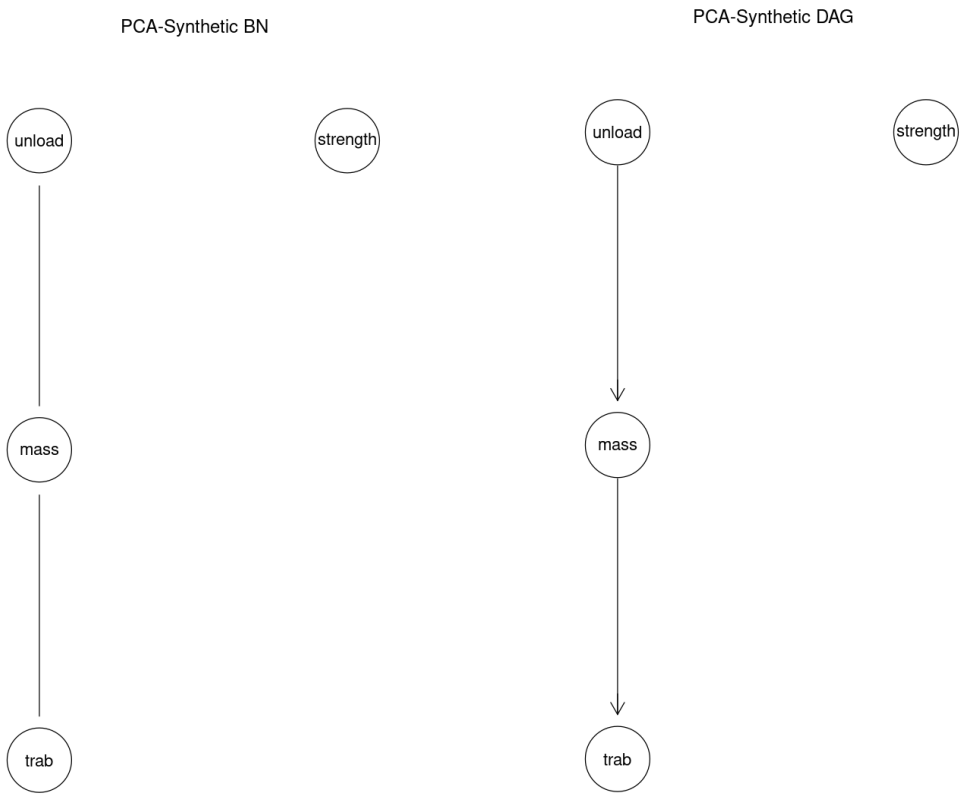


Output

The results for the "Au Naturel" data subset are as follows for the MMPC and MMHC algorithms:



The MMPC and MMHC have identical outputs, with the MMHC algorithm simply adding directionality to the output of MMCP. For the "PCA-synthetic" data subset, the results are as follows for MMPC and MMHC:



Much like with the “Au-Naturel” results, the output for “PCA-synthetic” is identical for both MMPC and MMHC aside from directionality.

Results

Experiments on Small Datasets

For the small datasets, the resulting graphs showed little similarity to the physical expectations of these experiments. Not only this, but the MMHC algorithm performs worse than the MMPC algorithm, adding directionality in the incorrect direction. This indicates that while these two algorithms manage to successfully create relations based on correlation, the MMHC algorithm is unable to correctly make relations on causality that reflect reality.

Experiment on Large Dataset

For the large dataset, the MMPC and MMHC algorithms were unable to handle the volume of data. While this could be attributed to the hardware these algorithms were run on, due to the recursive nature of these algorithms performing an intensive parents and children search for every feature in the dataset, this is most likely a limitation of the algorithm. Results may differ if the same tests were performed on significantly more powerful hardware.

Experiment on Large Dataset

This experiment differed from the previous two, in that there existed a concrete measure of success and failure. For the “Au Naturel” dataset, the results differed substantially from the baseline DAG. MMPC found relations and MMHC found causality between similar features, for example stating trabecular separation was the cause of trabecular number, and the max load was the cause of the fail load. These results do not make physical sense, and instead reflect correlation rather than causality. For the “PCA-Synthetic” dataset, the output does generate one correct relation, stating unload is correlated to mass. However, the algorithms fail to recognize any other relations, and importantly fail to correlate a change in strength to any other variables.

Discussion

In these experiments, it was found that when the MMPC and MMHC algorithms are run on small microscopy datasets, the BNs and DAGs generated by the algorithms poorly represent the real relations. Most results show the algorithms generating relations that would be physically impossible, and that represent correlation rather than causality. One possible cause of this is the general method MMPC and MMHC use for deriving structure. Because all relations are derived using probability, either testing correlation between related nodes or minimizing a probability

heuristic, the algorithms do not have an accurate notion of causality. Code for these experiments can be found at github.com/dag-ml/MMPC-MMHC-Microscopy.

References

Ko 2020, https://osdr.nasa.gov/bio/repo/data/studies/OSD-477
Ko 2020, https://osdr.nasa.gov/bio/repo/data/studies/OSD-608
Keune 2016, https://osdr.nasa.gov/bio/repo/data/studies/OSD-310
Dubeé 2016, https://osdr.nasa.gov/bio/repo/data/studies/OSD-489
Keune 2015, https://osdr.nasa.gov/bio/repo/data/studies/OSD-351
Costes 2021, https://osdr.nasa.gov/bio/repo/data/studies/OSD-366
mmhc, Time and sample efficient discovery of Markov blankets and direct causal relations
mmpc, The max-min hill-climbing Bayesian network structure learning algorithm
mmpc bnlearn, Max-Min Parents and Children (MMPC) learning algorithm
mmhc bnlearn, Max-Min Hill Climbing (MMHC) learning algorithm
bn class, BNLearn's Bayesian Network Representation