**Meeting attendees.**

Xia Jiang,  Garrett Barber

**Meeting time**

9:00 am – 10:00 am, Dec 06, 2022

**Meeting agenda.**

1. Reviewed progress last week.
2. Dr. Jiang gave feedback to Garrett’s work.
3. Dr. Jiang answered Garrett’s questions.

**Progress made in the past week.**

**Issues/Questions and Comments**

Dr. Jiang’s feedback to Garrett’s work per her email reply.

Hello,

I am glad you made progress. See some of my comments below in yellow highlight. I’d like to review your progress in detail and give my feedback following your report tomorrow morning during our ZOOM meeting.

# MBIL

MBIL (Markov Blanket and Interactive risk factor learner) is an algorithm that utilizes Bayesian networks and information theory to determine direct and interactive risk factor for metastatic breast cancer (MBC). (See papers for details and citations.)

## Install

MBIL can be installed from PyPI:

`pip install mbil-py`

Has the installation instruction been tested? Does it truly work?

### For all examples shown:

This is the data:

|B|         C|           D|           F|            E|

|-|--|---|---|---|

|2|         1|           0|           1|           0|

|3|         1|           1|           2|           1|

|3|         0|           1|           1|           0|

|3|         0|           1|           1|           1|

|2|         1|           2|           2|           0|

These are the set variables:  No these are not variable, rather, they are parameters. Please consider add explanation of each of the parameters and value ranges.

`alpha = 4`

`target = "E"`

`top = 20`

`max\_single\_predictors = 20`

`max\_interaction\_predictors = 20`

`max\_size\_interaction = 3`

`threshold = 0.05`

`maximum\_number\_of\_parents=7`

These are the basic score and search objects:

`score\_test\_obj = mbilscore.mbilscore(dataset\_df=dataset\_df, target=target, alpha=alpha)`

`search\_test\_object = mbilsearch.mbilsearch(threshold=threshold,

                                           max\_single\_predictors= max\_single\_predictors,

                                           max\_interaction\_predictors=max\_interaction\_predictors,

                                           max\_size\_interaction= max\_size\_interaction,

                                           dataset\_df = dataset\_df,

                                           alpha = alpha,

                                           target = target)`

## MBILScore functions and their uses:

### mbilscore.calculate\_score

Parameters: an mbilscore object and subset\_size

Return value: a hashmap storing the results

Calculates the BDeu score for all possible subsets (maybe P(data | DAG))

The BDeu score is a score that measures the probability of the data given the directed acyclic graph using a parameter alpha to represent prior equivalent sample size.

Example:

`scores = score\_test\_obj.calculate\_score(top = top, subset\_size = 2)`.   Is top supposed to be an mbilscore object based on your description rather than a number (20)?

`print(scores)`

Output:

`[("['B', 'C']", -3.753417975251508), ("['B', 'F']", -4.158883083359674), ("['B', 'D']", -4.382026634673884), ("['C', 'D']", -4.382026634673884), ("['D', 'F']", -4.382026634673884), ("['C', 'F']", -4.85203026391962)]`

### mbilscore.calculate\_information\_gain

Parameters: an mbilscore object and subset\_size

Return value: a hashmap storing the results

Calculates the information gain for all possible subsets

Information gain is the expected reduction in entropy of a variable conditional on a seperate variable

Example:

`ig\_scores = score\_test\_obj.calculate\_information\_gain(top = top, subset\_size = 2)`.  )`.   Is top supposed to be an mbilscore object based on your description rather than a number (20)?

`print(ig\_scores)`

Output:

`[("['D', 'F']", 0.5709505944546686), ("['B', 'C']", 0.5709505944546684), ("['B', 'F']", 0.5709505944546684), ("['C', 'D']", 0.5709505944546684), ("['B', 'D']", 0.4199730940219749), ("['C', 'F']", 0.17095059445466854)]`

## MBILSearch functions and their uses:

### mbilsearch.get\_single\_predictors\_score

Parameters: an mbilsearch object

Return value: a list of all direct risk predictors and their corresponding Bayesian score as a float

Calculates the Bayesian score for every direct risk predictor

Example:

When creating an mbilsearch object the function is called when initializing the variable `single\_list\_score`

Thus the variable can be printed as so

`print(search\_test\_object.single\_list\_score)`

Output:

`[('B', -3.5835189384561104)]`.     Shouldn’t it be a list of single predictors? In your example, it should include B, C, D, and F each separatedly?

### mbilsearch.get\_interaction\_predictors\_score

Parameters: an mbilsearch object

Return value: a list of the top interactive risk predictors and their corresponding score as a float

Calculates the score between all interactions of predictors

Example:

When creating an mbilsearch object the function is called when initializing the variable `interaction\_list\_score`

Thus the variable can be printed as so

`print(mbilsearch.interaction\_list\_score)`

Output:

`[("['B', 'C']", -3.753417975251508), ("['B', 'F']", -4.158883083359674), ("['C', 'D']", -4.382026634673884), ("['D', 'F']", -4.382026634673884), ("['C', 'F']", -4.85203026391962)]`

**Questions from Garrett and Dr. Jiang Answers**

What functions should I be defining?

- mbilscore functions? mbilsearch functions? scores functions?

We are not trying to define single score or search functions, rather we try to define a function that generates a list of single predictors, and a function that generates a list of interactive (epistatic or multi-dimensional) predictors, and a function that generates a list of direct and interactive causal predictors. If we don’t have these functions in our current python version, we will need to add them. If we do, we could just rename them to reflect closely what they are doing. In addition, we will have other assistant (intermediate) score and search functions defined and developed along the way.

Organizational structure of Readme?

- What should the order be? Description, Functions, examples?

Yes this order is good with me. I suggest we could do more examples. For example, in term of the example datasets, we could use another dataset that is real and/or simulated. We could also show example of learning 3-D interactive predictor. You could also compare the results of learning using our interactive risk factors learner and using other traditional methods. The comparison methods could include the following causal learning methods as showing in the table 2. These methods can be obtained freely via CCD ( Pitt Center for Causal Discovery). I will send you more datasets (done! There is a real dataset and a simulated dataset now in the github repo).

**Table 2.** Methods for Aim 1.



How in depth do you want the definitions of the functions?

As general as possible. You can assume that the user does not know much about our work before he/she learns from your readme. You can assume the user is like you when you first start to work on this.

- How much prior knowledge should I assume?

Not much prior knowledge is required to understand your readme.

Any aspects/functions of the program you want me to emphasize?

Garrett thinks it is already answer through our discussions today.

(Maybe) Any visuals that have already been created available for readme?

The figure that I provided is a good example for illustrating BDeu score. Please feel free to generate more figures (visuals) as you feel needed for other functions. I will provide more if I come across more that I think suitable. Be creative!

**Ongoing tasks that cover more than a week**

**Specific tasks for the coming week (the original task assignment for two weeks)**

1. Write a readme file for the python version of MBIL.
2. Explanation in details in your own word of the MBIL algorithm.
3. For each function of the MBIL-py provides, write an example to help users to understand the function and how they can use the function. Using figures is encouraged.
4. Learn from other existing readme files (such as the one provided by Shap https://github.com/slundberg/shap#readme) as to how to write a good readme file.
5. Other than the ideas that I give in (1) and (2), be creative. …….
6. Deadline: 2022.12.13

**Less urgent tasks**