MIIC online user Guide

Isambert lab - Institut Curie

February 22, 2021

1 Workbench: reconstructing your network using MIIC online

MIIC online server [1] aims at reconstructing causal, non-causal or mixed networks between the variables without an *a priori* choice on the type of reconstructed network. The workbench page allows to launch a job using the parameters, described in the following section.

1.1 Basic Settings *(mandatory: *)*

A section listing the parameters used to perform a network reconstruction.

Job name* : The name of your job[1]

Email (optional) : Your email address[2]

Dataset* : The input dataset[3] you want to analyze. It should be a table with comma, semicolon, tab, pipe or colon, as field separators, without sample ID, and with variable names specified as column names or row names. Theses variable names are used to label the nodes of the reconstructed network. Missing values are allowed in the dataset and should be indicated with *NA* in the dataset table. Each variable should be categorical or quantitative discrete.

Variable names* : Indicates whether the variable names are located in the first row or the first column of the dataset table.

1.2 Algorithm advanced parameters

An optional section allowing to specify various parameters of the MIIC algorithm and MIIC online display.

Neff : Effective number of independent samples (**Neff** $<$ **N**). The default value, -1, implies that all samples are taken as independent observations, *i.e.* **Neff** $=$ **N**.

Seed : Value used to initialize the pseudorandom number generator in the random sampling phase which is performed only if an effective number of samples is set, *i.e.* $0 <$ **Neff** $<$ **N**.

Complexity : Complexity criterion to take into account finite size effects from **N** or **Neff** samples for the network reconstruction. Complexity criterion: Normalized Maximum Likelihood (NML) *versus* Minimum Description Length / Bayesian Information Criterion (MDL / BIC). Default value: NML

[1] Necessary to retrieve your results in the Results page
[2] If provided, a notification email is sent when the job is completed
[3] Dataset maximum size: 500 variables and 200 MB in size

var_names	var_type	levels_increasing_order	group
Age	1		Clinical baseline
Menopausal status	0	Premenopausal,Postmenopausal	Clinical baseline
City	0		Patient information

Orientation : Is the orientation of v-structures ($\searrow\swarrow$) enabled? Default: YES

Propagation : Should orientations be propagated downstream of v-structures ($\searrow\swarrow\rightarrow\rightarrow$)? Default: YES

Latent : When enabled, this parameter allows to detect the effects of unobserved (latent) common causes on the relationships between observed variables, represented by bidirected edges, *i.e.* $\longleftrightarrow$. Default: NO

1.3 Supplementary files

In this section, it is possible to upload optional supplementary files that give specifications about various aspects of `MIIC online` reconstruction. To see how they must be formatted, please take a look at the downloadable online example files.

True edges : An optional file allowing to evaluate the performances of `MIIC online` reconstruction against a known Directed Acyclic Graph (DAG)[4]. The returned performance measures are 'Precision', 'Recall' and 'F1-score'.

Network layout : An optional file specifying node positions in the 2D representation of the network, containing an x y coordinate pair for each node (separated with a separator).[5] It is possible to directly upload a network layout file (".json") that have been saved and downloaded in the network interactive visualization page. Note: when you modify the layout of a network in the interactive visualization page and you save it, the modified layout will be stored in the web server, meaning that a new access to the page will show the new layout and the old layout will be overwritten.

Category order : An optional file providing information about how to consider the different states of categorical variables. It will be used to compute the signs of the edges (using spearman correlation coefficient) by ranking the levels of each variable according to the order given in the file. This file is necessary (except for numerical variables) to obtain edge colors corresponding to the signs of their partial correlations (positive in red, negative in blue). If it is not possible or desirable to order the states of some variables, the column "levels_increasing_order" can be left empty for these variables. The edges involving those variables are then colored in gray in the reconstructed network. (NB: in this case, the field separator is still needed between the node name and the empty "levels_increasing_order" cell in the category order file). A fourth column named "group" can be added to regroup variables in different sets. Variables in the same group will be coloured with the same colour. An example of this file with all 4 columns is provided below.

Excluded edges : An optional file containing any prior knowledge about edges that should be excluded in the reconstructed network. It should be formatted as a two-column file, `Node1 Node2`, with a field separator between them.

[4]The reference DAG should be provided as a two-column table, without column names, where each row corresponds to an edge, with the first column including a source node and the second column a target node (see example).

[5]The nodes are considered in the same order as in the input dataset, unless an optional first column is added, specifying the name of each node.

1.4 Skeleton cut

This step aims at filtering the least robust edges in the network skeleton based on an edge-specific confidence ratio, computed by comparing the probability to discard an edge before and after random permutations of the input dataset. Smaller confidence ratios imply more reliable edges.

Confidence filtering : On/Off toggle switch specifying whether edge-specific confidence ratios should be evaluated and used to filter the least robust edges. The corresponding summary file contains the confidence value for each edge, and a column specifying if the edge has been filtered out at this step. Default: Off

Number of shufflings : number of random permutation shufflings of the input dataset to estimate the edge-specific confidence ratios. Default value: 100.

Confidence threshold : Threshold above which corresponding edges are discarded from the network skeleton. Default value: 0.01.

1.5 Orientation cut

This step aims at filtering the orientations that are evaluated by the algorithm based on V-structures.
This threshold is used when deducing the type of an edge tip (head/tail) from the probability of orientation. For a given edge tip, and denoted the probability of it being a head as p, the orientation is accepted if $\frac{1-p}{p} < threshold$. 0 results in rejecting all orientations, 1 in accepting all orientations.

2 Waiting Page: `miic` is working to get your results...

This page sums up the details of the submitted job and provides information about its current state (enqueued, skeleton initialization, skeleton iteration, post-processing). The parameters of the job are available through the button "View job details". It is also possible to bookmark and save the web page. If cookies are enabled, the job is also saved in the **Results** page, avoiding the user to wait on this page until the job is finished (if an e-mail is not provided) and to perform the saving process manually. Once the job is finished, the user is automatically redirected to the job results page. If the job takes more than 30 seconds, the server offers the possibility to provide an email address, if it was not already provided, and asks if the user prefers to stay on the waiting page or go to the page where all user jobs are stored (if cookies are enabled). If one or more jobs are not finished, this page is updated every minute, providing an updated global jobs status. The possibility to provide an e-mail is always available in this section through the button "Register mail ".

3 Results Page: Here is your network !

This page displays your network and a lot more. The upper section sums up information about the job, whereas the bottom section provides the results in the form of a series of tabs described below.

Graph : An interactive visualization of the network reconstructed by `MIIC online`, where nodes can be moved with a simple drag and drop in order to get a more personalized view the network. If a layout file is provided as supplementary file, it is applied to the visualization, while the force directed layout is applied otherwise. Red edges represent positive correlations

while blue ones represent negative correlations. Edges colored in gray are present for categorical variables with no information provided on the category order file, since in such case it is not possible to evaluate the sign of the correlation. Edge thickness is reported accordingly to edge strength, whose quantity is reported in the "Summary" tab, in the "log_confidence" or "info_shifted" column.

Summary : A table containing detailed information about the inferred and discarded XY edges. A tool allows to sort and filter each column according to its values.

- **x**: name of X node
- **y**: name of Y node (Note: edges are sorted in alphabetical order).
- **type**: inferred (P), discarded (N) edge. If the true edge file is also uploaded, P becomes TP (true positive) or FP (false positive), and N becomes TN (true negative) of FN (false negative). This information refers to correct or erroneous edges with respect to the network skeleton (*i.e.* without edge orientations). The oriented network prediction accuracy is evaluated comparing the inferred network with the CPDAG (Completed Partially Oriented Acyclic Graph) obtained from the provided DAG.
- **ai**: contains the comma separated list of the accepted contributing nodes, $\{A_i\}$.
- **info**: the mutual information between X and Y multiplied by the number of samples for which the values for variables X,Y and $\{A_i\}$ are available (*i.e.* without 'NA'), *i.e.* $NI(X;Y|\{A_i\})$
- **info_cond**: the remaining information between X and Y after conditioning on $\{A_i\}$ multiplied by the number of samples for which the values for variables X,Y and $\{A_i\}$ are available (*i.e.* without 'NA'), *i.e.* $NI(X;Y|\{A_i\})$
- **cplx**: the NML or MDL complexity value $k_{X;Y|\{A_i\}}$
- **Nxy_ai**: the number of samples for which the values for variables X,Y and $\{A_i\}$ are available (*i.e.* without 'NA')
- **Info_shifted**: the evaluation of the remained marginal mutual information minus =the complexity term; $-logP_{XY} = NI(X;Y|\{A_i\}) - k_{X;Y|\{A_i\}} = NI'(X;Y\{A_i\})$ value. It is also a way to quantify the strength of the edge XY. (Hint: Below a few units (1-3) the confidence on the predicted edge remains low, while above 10 the predicted edges can be considered as reliable.)
- **infOrt**: the orientation of the edge XY. It is the same value as in the adjacency matrix at row x and column y
- **trueOrt**: the orientation of the edge XY present in the true edge file (if the true edge file is provided)
- **isOrtOk**: information about the consistency of the inferred graph's orientations with the reference graph given (if the true edge file is provided)
  - $\Rightarrow$ Y: the orientation is consistent
  - $\Rightarrow$ N: the orientation is not consistent [6]
- **sign**: sign of the partial correlation for edge XY conditioned on $\{A_i\}$
- **partial_correlation**: value of the partial correlation coefficient for edge XY conditioned on $\{A_i\}$
- **IsCausal**: An edge can be called being causal if:

[6] with the CPDAG computed using the provided true graph.

a. it is part of a v-structure
b. the origin node if at the tip of another v-structure
c. the edge is not forming v-structures with the incoming edges of the origin node

- **proba**: shows the probability for orienting this edge, computed from the V-structure with the strongest absolute three-point information.
- **confidence**: if the confidence cut option is enabled, it is the confidence ratio C_{XY} between the probability to remove edge XY using the actual dataset and the mean probability to do the same averaged over multiple randomly permutated datasets. The lower C_{XY}, the higher the confidence on the XY edge. This value can be used to retain only high confidence edges in the predicted network.

Probabilities : This tab lists the orientation probabilities of the two edges of all unshielded triples of the reconstructed network: [node$_1$] $p_1 - p_2$ [mid-node] $p_3 - p_4$ [node$_2$], where p_i are the endpoint arrow head orientation probability of each edge: *e.g.* $p_i > 0.5$ corresponds to an arrow head, $p_i < 0.5$ corresponds to an arrow tail ($p_i = 0.5$ is undefined or corresponds to an undirected edge). For each open triplet, we evaluate the 3-point information ($NI3$ value), which reports the strength of the signature of causality. More negative values correspond to more reliable v-structures.

Variable plots : Distribution plots are available for each variable in the input data. Bar-plots are shown for each categorical variable and histograms for every continuous variable (with am additional density curve).

Data dictionary : A table showing the given category order file loaded in the workbench page.

Download : A tab allowing to select and download the files generated by `MIIC online` and presented in the results page, along with the input observation dataset. Two supplementary downloadable files not present in the results page are the non-oriented and oriented adjacency matrices. Adjacency matrices are square matrices used to represent the inferred graph, as the matrix entries indicate whether pairs of vertices are adjacent or not in the graph. The matrix can be read as a (row, column) set of couples where the row represents the source node, X, and the column the target node, Y. Since `MIIC` can reconstruct mixed networks (with directed, undirected and bi-directed edges), five labels are needed to describe adjacencies:

- 0: means that X and Y are not adjacent
- 1: means that the XY edge is undirected as $X - Y$
- 2: means that the XY edge is directed as $X \rightarrow Y$
- -2: means that the XY edge is directed as $X \leftarrow Y$
- 6: means that the XY edge is bidirected as $X \leftrightarrow Y$

Note that the oriented adjacency matrix is not symmetric by construction.

References

[1] Sella N, Verny L, Uguzzoni G, Affeldt S, Isambert H. "MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data", Bioinformatics 34(13):2311-2313 (2018).