User guide for the MIIC algorithm command line version

Nadir Sella,[1,2] Louis Verny,[1,2,3] Séverine Affeldt,[1,2,4]
Hervé Isambert,[1,2,†]

[1]Institut Curie, PSL Research University, CNRS, UMR168, 26 rue dUlm, 75005 Paris, France
[2]Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France

[3] Current address: CENOVA SAS, 100 Avenue Charles De Gaulle, 92200 Neuilly Sur Seine
[4] Current address: ICAN, UPMC, Paris, France

[†]To whom correspondence should be addressed; E-mail: herve.isambert@curie.fr

The main folder contains the scripts and source code for the reconstruction of networks starting from observation data. The structure is the following:

The directories are organized as follow:

```
/
├── common
│   ├── miic.R
│   ├── gmPlot.R
│   └── gmSummary.R
├── sharedLib
├── data
│   └── some data input/output...
├── miic
│   └── all miic executables
└── doc
    └── documentation
```

Package requirements

To launch the miic.R script you need to have R installed on your machine, along with some packages that are available in the CRAN repository.

Rpackages MASS, getopt, plotrix, methods, igraph, ppcor, bnlearn

Calling the inference methods with `miic`

You can call the inference methods through the *miic.R* script.

Overview

main *~/common/miic.R*

lib *~/common/lib/...*

Arguments *(mandatory: *)*

-i * file path of the input dataset[1]

-o * directory path for the output of the inference method[2]

-d steps to perform ('1,2,3,4' or '1,2' or '1,3' *etc...*)
default: '1,2,3,4'[3]

-p parameters for the inference method (see the following subsections). The value expected here is of type character: '$param_1$:$value_1$,$param_2$:$value_2$ *etc...*'

-t file path to the true edges; used during the *summary* step[4]

-s file path to the category order used during the *summary* step[5]

-x file path to the blackbox file containing the edges to be removed at the beginning of the reconstruction step.[6]

[1]The input dataset should be a tab separated table, with column names but no row names. Missing values should be indicated with *NA*. Each column corresponds to a categorized variable and each row to one sample.

[2]To prevent from overwriting existing results, if the output directory already exists, the skeleton inference step returns a message and stops.

[3](1) skeleton, (2) orientation, (3) summary, (4) plot

[4]The true edges file has two space-separated columns. Each line corresponds to one edge. The orientation is $col1 \rightarrow col2$.

[5]The orientation is $col1 \rightarrow col2$. This files provides 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).

[6]It must be formatted as a two-column file, `Node1 Node2`, with a field separator between them.

-l file path to the layout of each vertex; used during the *plot* step[7]

-c if given, edges will be filtered according to their confidence ratio. It needs two parameters, described in . To use a different threshold without shuffling the data a second time, an identical command line with a different confidence ratio threshold can be used. The program will keep the edges satisfying the new confidence ratio, using the previously calculated mutual information

-v if given, detailed information is given in the log file and command shell

-n if given, the graphml file is created in the output folder[8]

An example:
Rscript ***miic.R -i*** *../data/asiaDataset.txt* ***-o*** *../data/asiaNetwork* ***-m*** *miic* ***-p*** *cpx:nml,efn:-1,lat:yes,prg:yes* ***-c*** *csh:100,ccr:0.01*

When calling the available inference methods with *miic.R*, the 'p' option can be used to indicate the chosen parameters. The value expected for this option is of type character: '$param_1$:$value_1$,$param_2$:$value_2$ *etc...*'. The possible $param_i$ and $value_i$ for each method are detailed in the following subsections.

Option '-p' for `miic`

cpx formula used to compute the complexity term ['mdl'[9] or 'nml'[10]]
default: nml (*Ex.: -p '...,cpx:mdl,...'*)

lat should the network be reconstructed under the hypothesis that some variables might not be observed? ['yes' or 'no']
default: no (*Ex.: -p ...,lat:yes,...*)

prg should the network be oriented using the propagation rule? ['yes' or 'no']
default: yes (*Ex.: -p ...,prg:yes,...*)

efn number of uncorrelated samples
default: number of rows of the input dataset (*Ex.: -p ...,efn:1000,...*)

A '-p' example: *-p cpx:mdl,efn:1000*

[7]The layout file has three tab separated columns, the first column being optional. Each line corresponds to the (x,y) coordinates of each vertex. The first column can contain the label of the vertex as indicated in the colnames of the input dataset table. The order in which the coordinates are given also corresponds to the order of the colnames of the input dataset table.

[8]If the layout file is provided, the network will be stored in a xgmml file format, that allows the node positioning in the Cytoscape tool. We have noticed that in this case edges are note coloured, we will soon try to solve it.

[9]Maximum description length

[10]Normalized Maximum Likelihood

Option '-c' for `miic`

csh number of random shuffling of the input dataset, in order to get the random mutual information between `miic` inferred edges (*Ex.: -c ...,csh:10,...*)

ccr confidence ratio used as a threshold for filtering the edges. (*Ex.: -c ...,ccr:0.01,...*)

A '-c' example: *-c csh:100,ccr:0.01*

Viewing inferred networks

The inferred networks can be viewed in pdf format,automatically generated with igraph (http://igraph.org/) or manipulated for better display using cytoscape (http://www.cytoscape.org). The files are located in the following directories:

- Unfiltered network: the pdf and graphml files are located in the output directory set by the -o entry in the command line
- Filtered networks (using -c option with `miic`): the pdf and graphml file are located in the subdirectory 'shuffle_[cshValue]/filtered_network_[ccrValue]', which can be found in the output directory set by the -o entry in the command line

In order to visualize the network in a correct, pleasant and interactive way, we recommend the utilization of Cytoscape tool, version 3.1.0 or later (`http://www.cytoscape.org/`). Cytoscape is available for Windows, Linux and OsX.
In order to load the network you have to go through the following steps:

1. Import the network: File⇒Import⇒Network⇒File, and select the graphml file
2. Import the style: File⇒Import⇒Styles, and select the miic_style.xml file present in this folder
3. Select the loaded style: under the "Style" panel present in "Control Panel" select miic_style