SCL/CAD replication

BT

2020-01-08

Contents

Dependencies 1
Overview 1
Data 2
Experiment - varying p 2
Results 3

Dependencies

- `MASS`
- `pcalg`
- `biglasso`
- `huge`
- `corpcor`
- `nem`
- `matrixStats`
- `irlba`
- `rsvd`
- `MLmetrics`
- `ROCR`
- `parallel`
- `doParallel`
- `itertools`
- `BiocManager`
- `filehash`
- `R.utils`
- `R.devices`
- `tidyverse`
- `colorspace`

```
sessionInfo()
```

Overview

The goal of this document and code is to (externally) reproduce/replicate the original results in the Noé et al. paper (https://arxiv.org/abs/1905.11506).

The code is a slightly modified and trimmed down version of the original (for the full paper version, see Github repo: https://github.com/unoe/cad). None of the pre- or post-processing steps have been changed. Some function calls have been simplified for better interactability and the overall workflow for a subset of the original experiments has been condensed into this markdown document. File, function and variable names have (largely) been kept as in the original for comparability.

Data

The data structure `data_list` contains all available data points as extracted from the original (raw) csv data files.

Requirement: File `shrna_processed_data.rds` needs to be stored in the "data" folder inside the working directory.

```
#cat('\n load preprocessed shRNA-seq data ... \n')
fid_data <- filePath(getwd(), 'data', 'shrna_processed_data.rds')
data_list <- readRDS(fid_data)
```

Experiment - varying p

```
if(!dir.exists('results')){ dir.create('results') }
```

Note: The following `experiments_*.R` scripts are wrappers for the actual function calls. They also specify certain parameter settings, i.e. the number of repetitions (different subsamples) per setting, the dimensionality p (from 25 to max of about 19000 – *Warning: most methods scale super-exponentially with p and are infeasible to run beyond $p \approx 1000$*).

The first part in some of these scripts is also used to generate the data matrices for each specified parameter setting. The data files are automatically stored in the "data" folder. In principle, this has only to be run once to create the input data for any method call.

Individual results (for each method, setting and repetition) are stored as separate files in the "results" folder. The plotting script then processes the whole directory at the end to combine the results.

Default settings: As defaults, I've set the number of subsampled data sets (i.e. "repetitions") to 3 and restricted the dimensionality to $p = [50, 100]$. In the end, the aim for a full replication of the original results would be 10 repetions and p up to 1000.

Running the original CAD/SCL code

```
#source('experiments/experiments_vary_p_NEW.R')
```

Note: This will take a long time to run for any $p > 200$. Since we are mainly interested in the other methods, this can be skipped.

Running the original GIES code

```
source('experiments/experiments_vary_p_gies.R')
```

To change the GIES function call:

Please modify the function in `run_gies.R` (located in the "R" subfolder in the main directory).

Running the original IDA and PC code

```
source('experiments/experiments_vary_p_LITERATURE.R')
```

To change the IDA and PC function calls:

Please modify the functions `run_ida.R` and `run_pc.R`, respectively (located in the "R" subfolder in the main directory).

Running the original LV-IDA/RFCI code

```
source('experiments/experiments_vary_p_LVIDA.R')
```

To change the LV-IDA/RFCI function call:

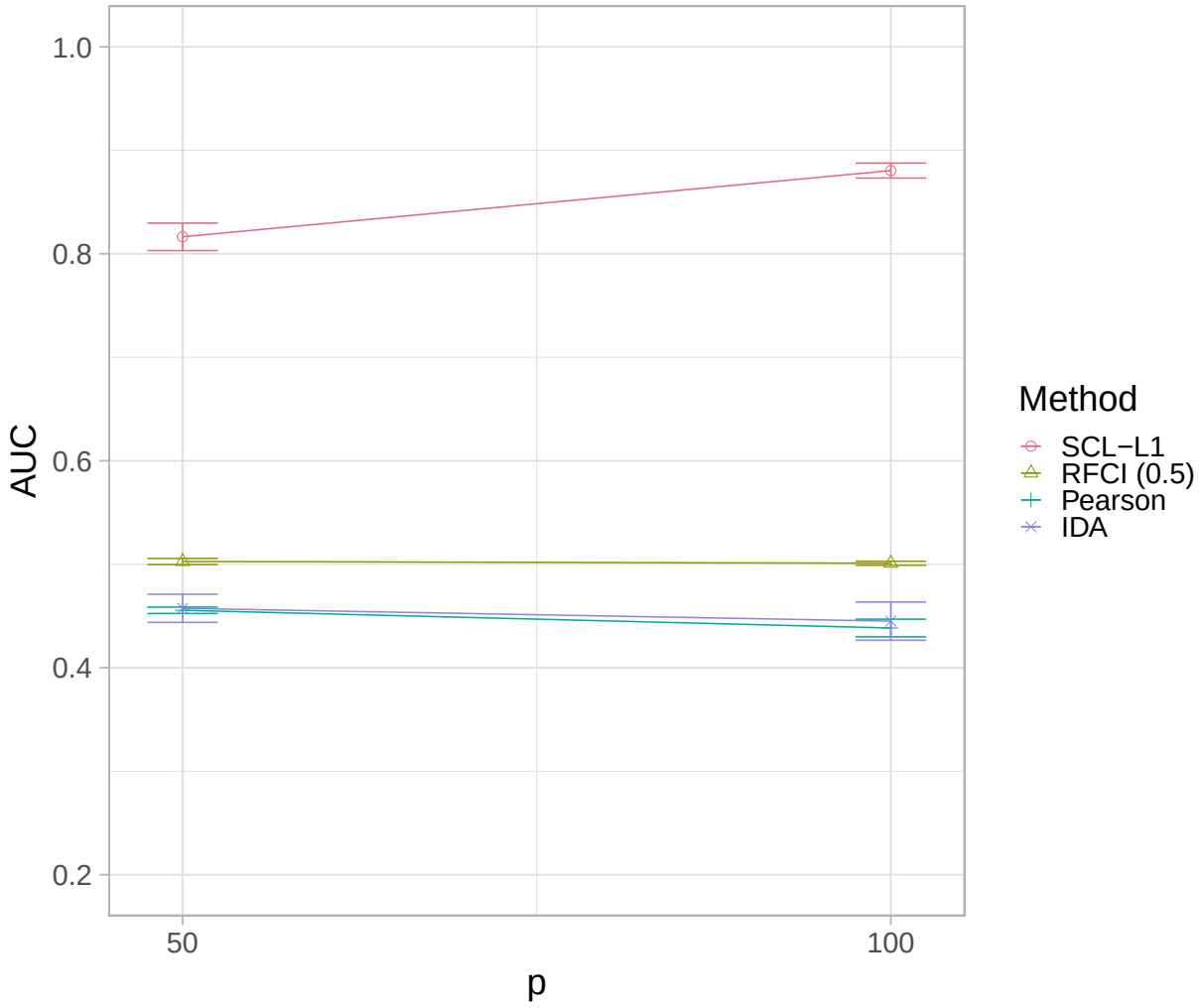
Please modify the function in `run_lvida.R` (located in the "R" subfolder in the main directory).

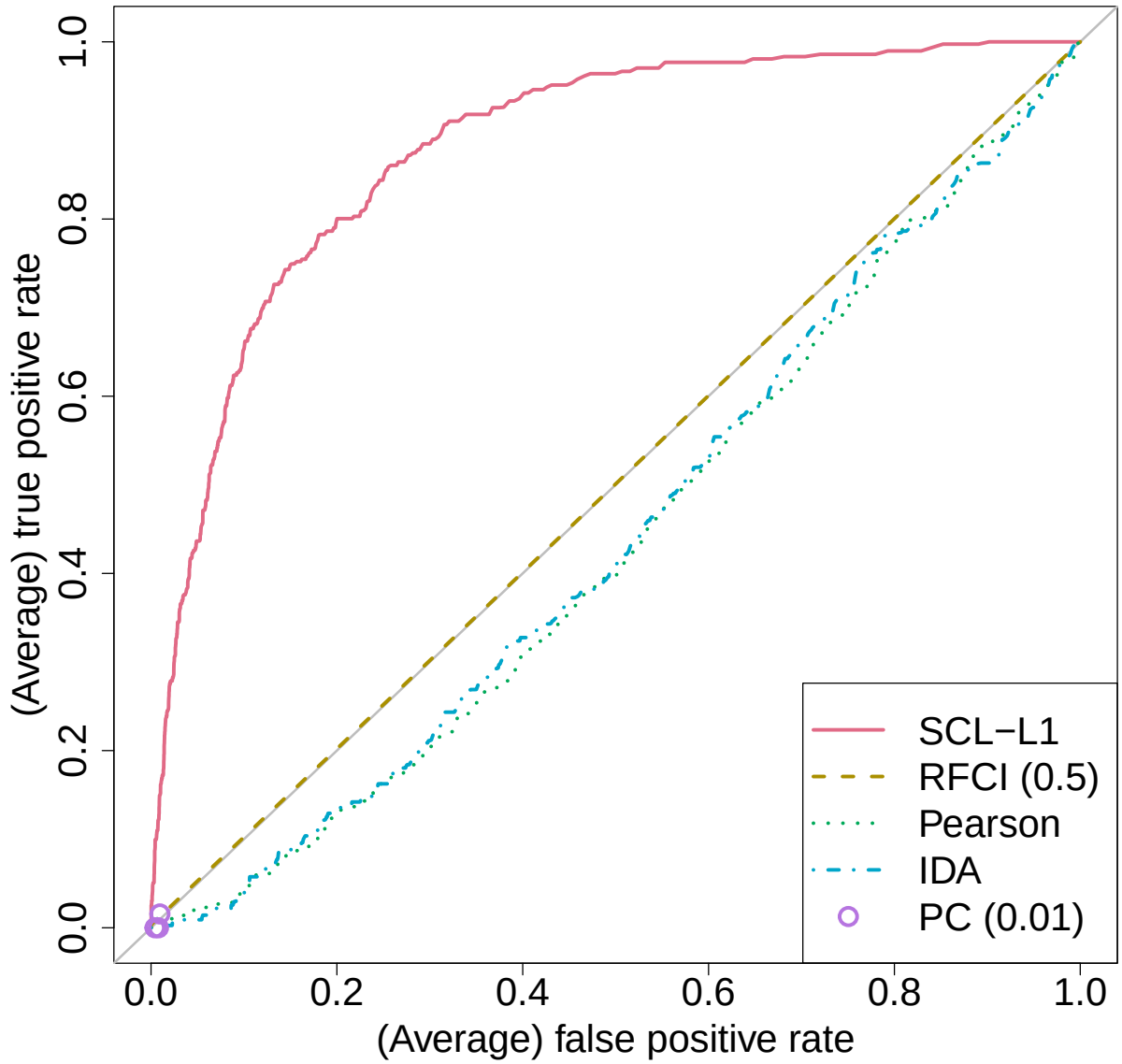
Results

Below are shown some example results for the test settings.

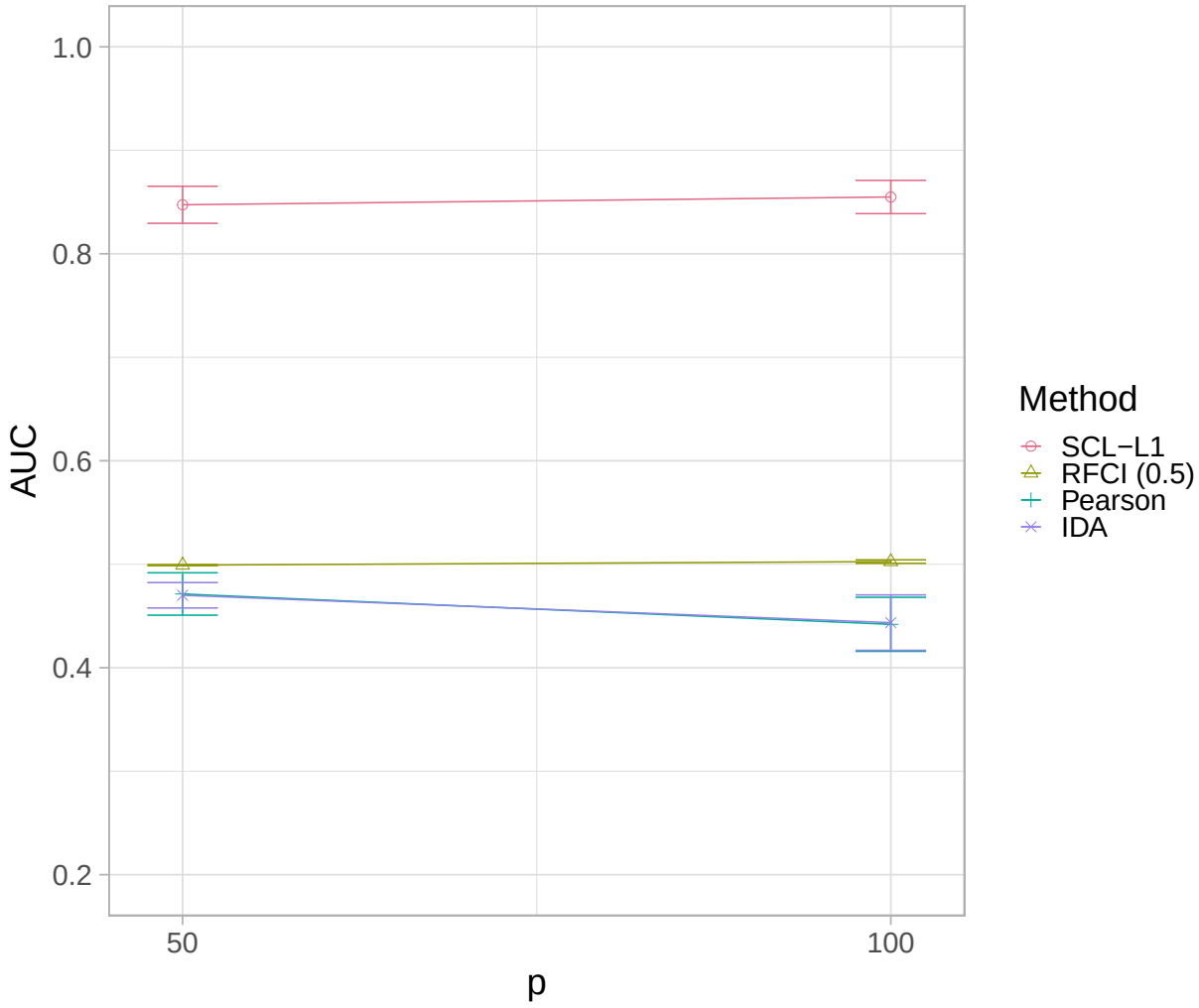
Note: These function calls should also work for any other preliminary, partial or full results. All the files in the "results" folder are automatically taken into account when generating the figures.

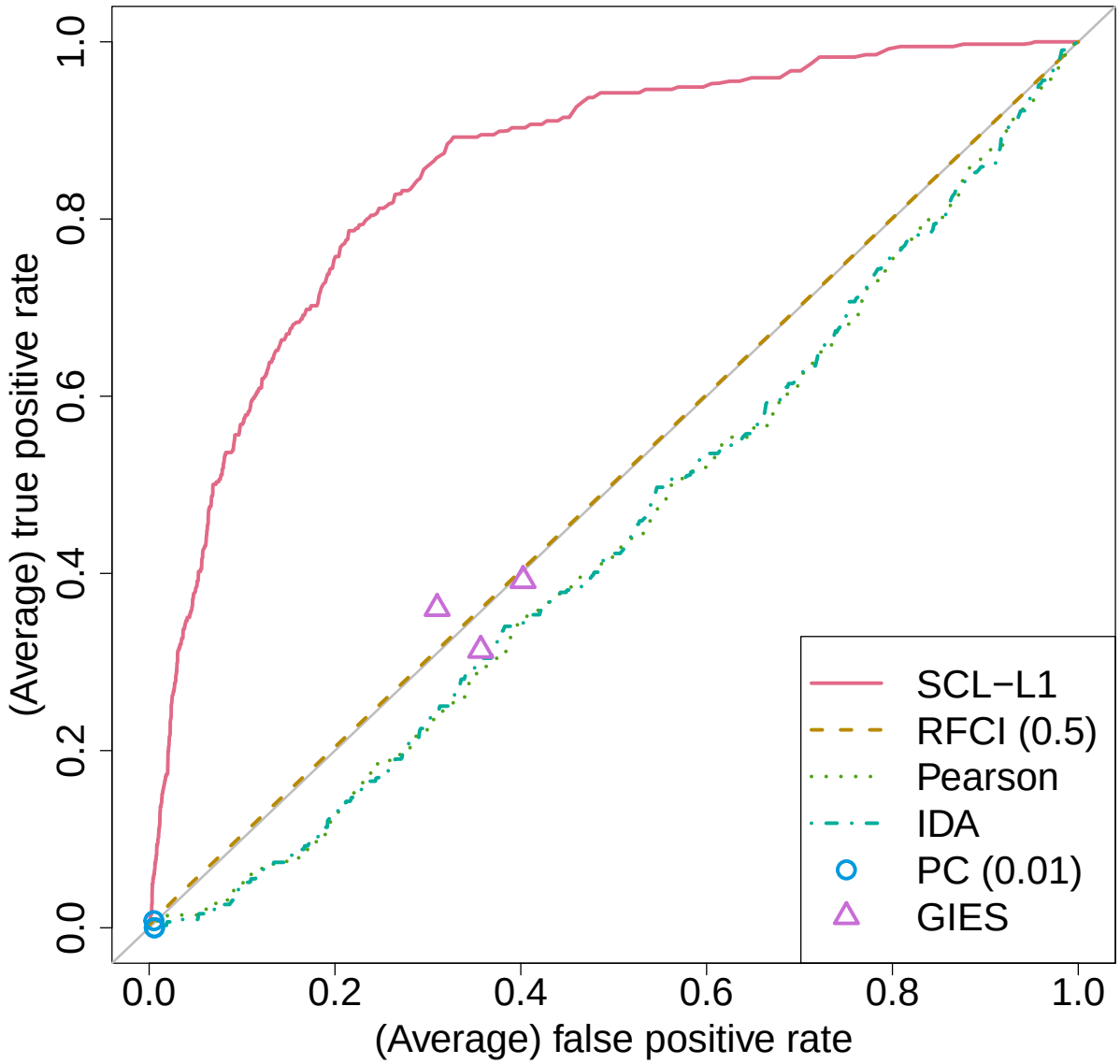
Entry-wise masking



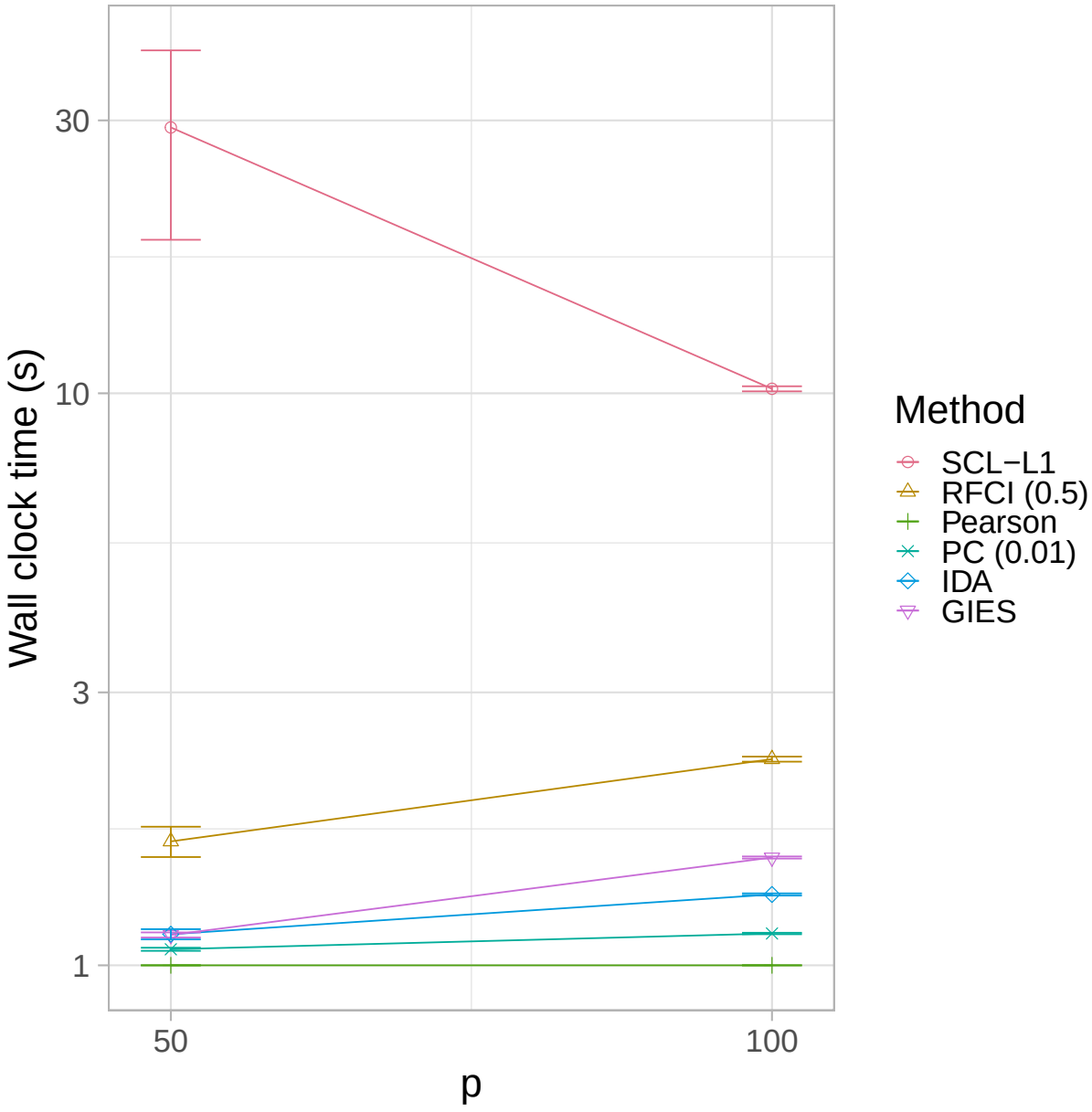

(Average) true positive rate
0.0
0.2
0.4
0.6
0.8
1.0
(Average) false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
SCL−L1
RFCI (0.5)
Pearson
IDA
PC (0.01)

Row-wise masking




(Average) true positive rate
0.0
0.2
0.4
0.6
0.8
1.0
(Average) false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
SCL−L1
RFCI (0.5)
Pearson
IDA
PC (0.01)
GIES

Timings


Wall clock time (s)
30
10
3
1
50
100
p
Method
SCL−L1
RFCI (0.5)
Pearson
PC (0.01)
IDA
GIES