ARE BIBLIOMETRIC CHARACTERISTICS GOOD PREDICTORS OF A PAPER'S CITATION SCORE?

AN APPLICATION OF LEARNING AND VALIDATING CONTINUOUS BNS FOR HIGHLY SKEWED DISTRIBUTIONS

Tina Nane & Anca Hanea

TU Delft

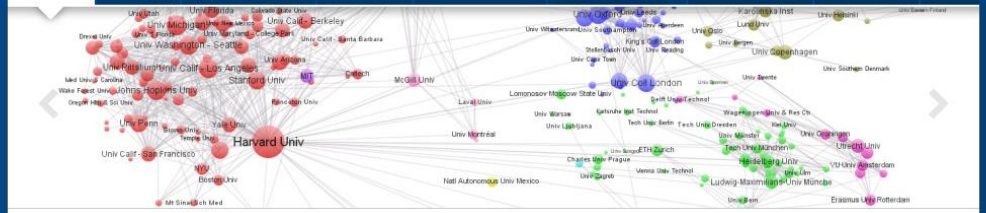
cebra
Centre of Excellence for Biosecurity Risk Analysis

The University of Melbourne

CWTS
Meaningful metrics

Leiden University | CWTS B.V. | Other CWTS sites | About CWTS | About Leiden | Contact

Home | News | Seminars | People | Research | Education | Projects | Services



Centre for Science and Technology Studies

The Centre for Science and Technology Studies (CWTS) studies scientific research and its connections to technology, innovation, and society. Our research, bibliometric and scientometric tools, and evaluation expertise provide a solid basis for supporting research assessment and strategic decision making and for developing science policy.

TU Delft

THE UNIVERSITY OF MELBOURNE

Search | Menu

Collaborative Assessment for Trustworthy Science

the repliCATS project

Publications | our team | Get involved | FAQs | News + Workshops

The repliCATS project aim is to crowdsource evaluations of the credibility of published research in eight social science fields: business research, criminology, economics, education, political science, psychology, public administration, and sociology.

In phase 2, the repliCATS project is about reimagining peer review as a structured deliberation process.

THE repliCATS PROJECT

Tweets from @replicats

repliCATS_project Retweeted

CEBRA
@CEBRA_UoM · Oct 31

Congratulations to @fidlerfm and the team at @replicats (including our own Anca Hanea and Cassie Watts) on winning the @unimelb 2022 Award for Excellence in Interdisciplinary Research!

5

THE UNIVERSITY OF MELBOURNE

Predict mean citation score (*mcs*) of a paper / the average number of citations since publication
(~ future scientific achievement)

"how often a paper is cited should be predicted by the quality and reliability of the science"

"there is no evidence that citations reflect other key dimensions of research quality"

"[…] large enough data sets reveal that there are parameters independent of individual papers' quality that can determine an average citation rate."

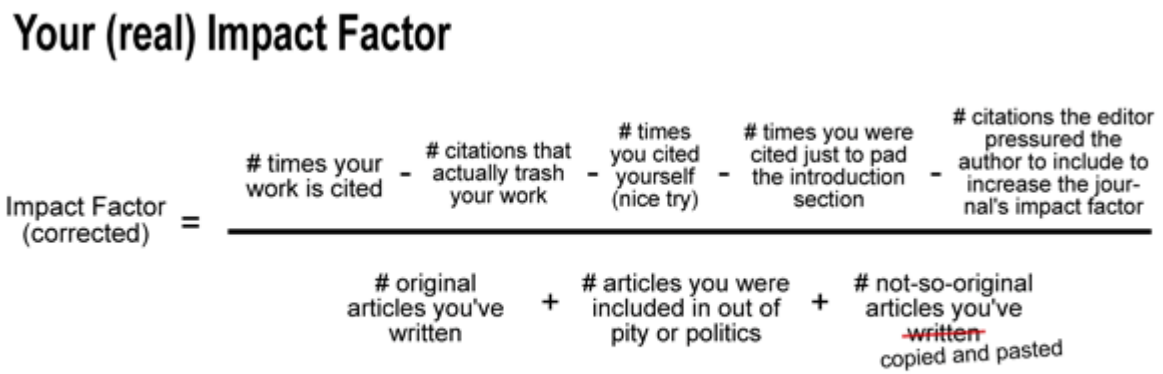
"[…] large enough data sets reveal that there are parameters independent of individual papers' quality that can determine an average citation rate."

Predict *mcs* - papers published 2010-2014, with at least one author @ TU Delft.

Paper

Authors

Journal



Journal

- *js* -- journal citation score = the average citation score of publications in a journal

Paper

- *refs* -- number of references of the paper
- *authors* -- number of authors of the paper

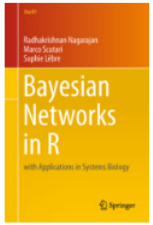
Authors

- *pMax* -- max(p of all authors of the paper)
  - p - number of publications (in Web of Science) of each author of the paper
- *mcsMax* -- max (mcs of all authors of the paper)
  - mcs -- average citation score of an author, average of all citations of all publications
- *pptopMax* -- max(pp_top_10_prop of all authors of the paper)
  - pp_top_10_prop -- percentage of author's publications in top 10% of their field
- *ageMax* -- max (academic age of all authors of the paper)
  - academic age - the difference between the publication year of the paper and the publication year of the first paper of the author

bnlearn - an R package for Bayesian network learning and inference

- Home Page
- Documentation
- Examples
- Research Notes
- Bayesian Network Repository

- About the Author



data & R code

bnlearn is an R package for learning the graphical structure of Bayesian networks, estimate their parameters and perform some useful inference. It was first released in 2007, it has been under continuous development for more than 10 years (and still going strong). To get started and install the latest development snapshot type

```
install.packages("https://www.bnlearn.com/releases/bnlearn_latest.tar.gz", repos = NULL, type = "source")
```

in your R console. (More detailed installation instructions below.)

bnlearn implements the following *constraint-based structure learning algorithms*:

- PC (the *stable* version);
- Grow-Shrink (GS);
- Incremental Association Markov Blanket (IAMB);
- Fast Incremental Association (Fast-IAMB);
- Interleaved Incremental Association (Inter-IAMB);
- Incremental Association with FDR Correction (IAMB-FDR);
- Max-Min Parents & Children (MMPC);
- Semi-Interleaved Hiton-PC (SI-HITON-PC);
- Hybrid Parent

Downloads

current release on CRAN:

4.8.1 [link]

latest snapshot + bugfixes:

4.9-20221107 [link]

From the R Studio CRAN Mirror:

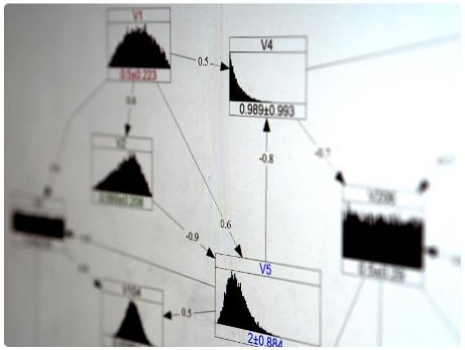
CRAN 4.8.1

downloads 29K/month

LightTwist Software

Software Development

CONTACT US

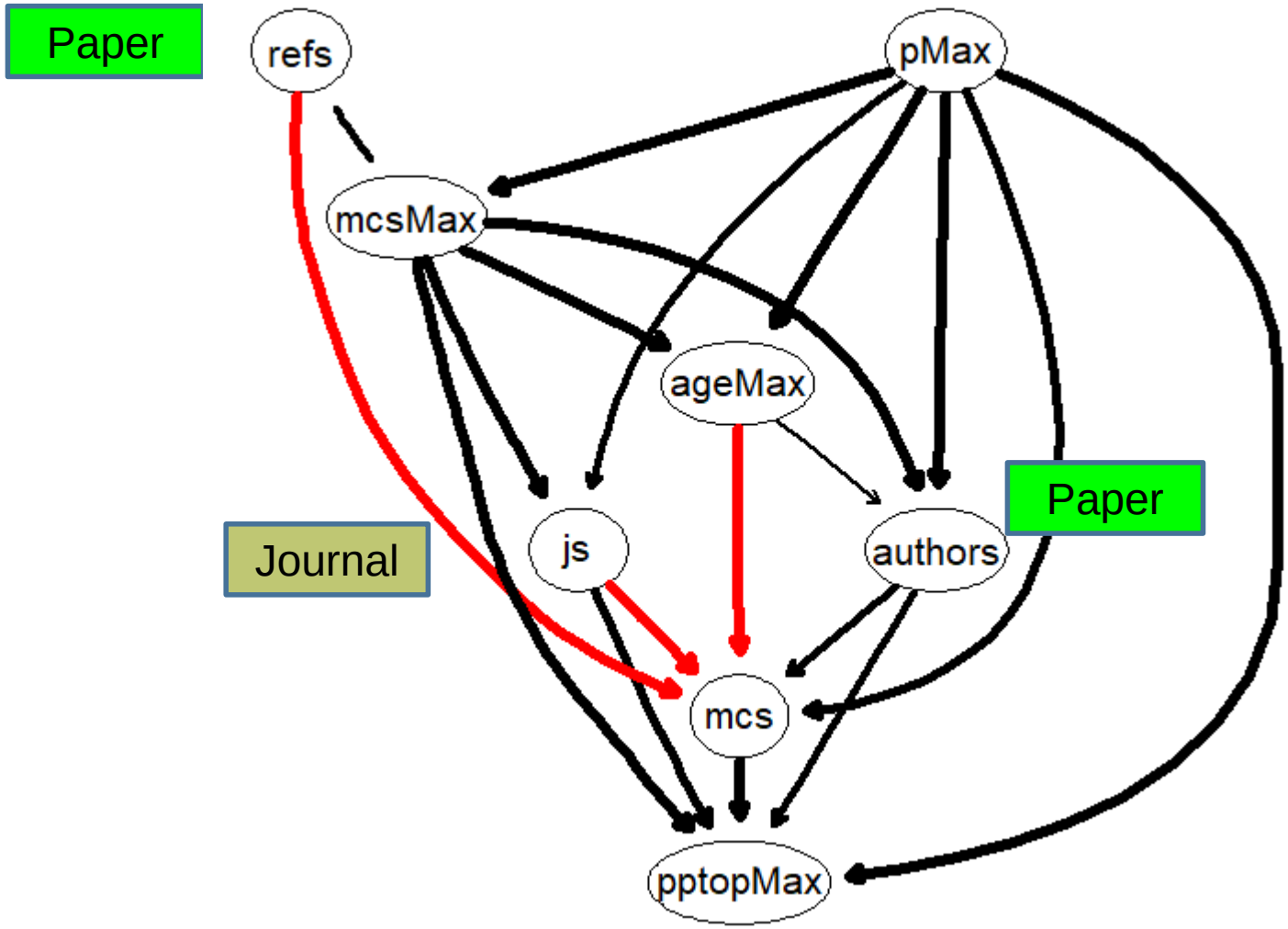


Uninet standalone

Uninet is a standalone uncertainty analysis software package. Its main focus is dependence modelling for high dimensional distributions. Random variables can be coupled using Bayesian networks, vine-copula constructions or dependence trees.

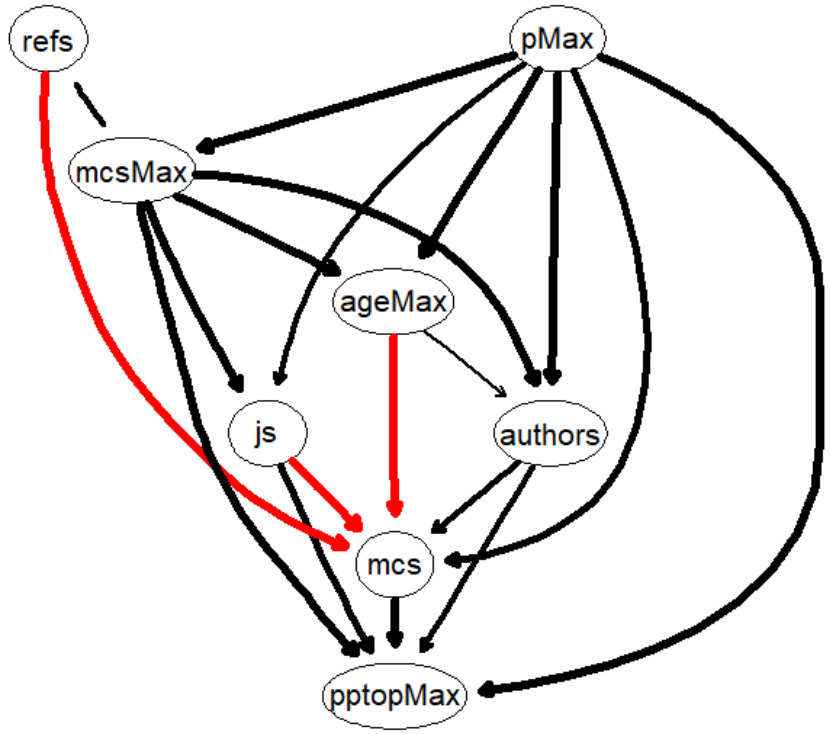
Read the *Uninet help file* describing the software in detail: UninetHelp.pdf (1.4 MB)

Visit the licensing page for details about the Uninet and UninetEngine licences and to find out how to acquire the latest versions.



Max min hill-climbing hybrid algorithm, which combines local learning, constrained-based and search-and-score techniques

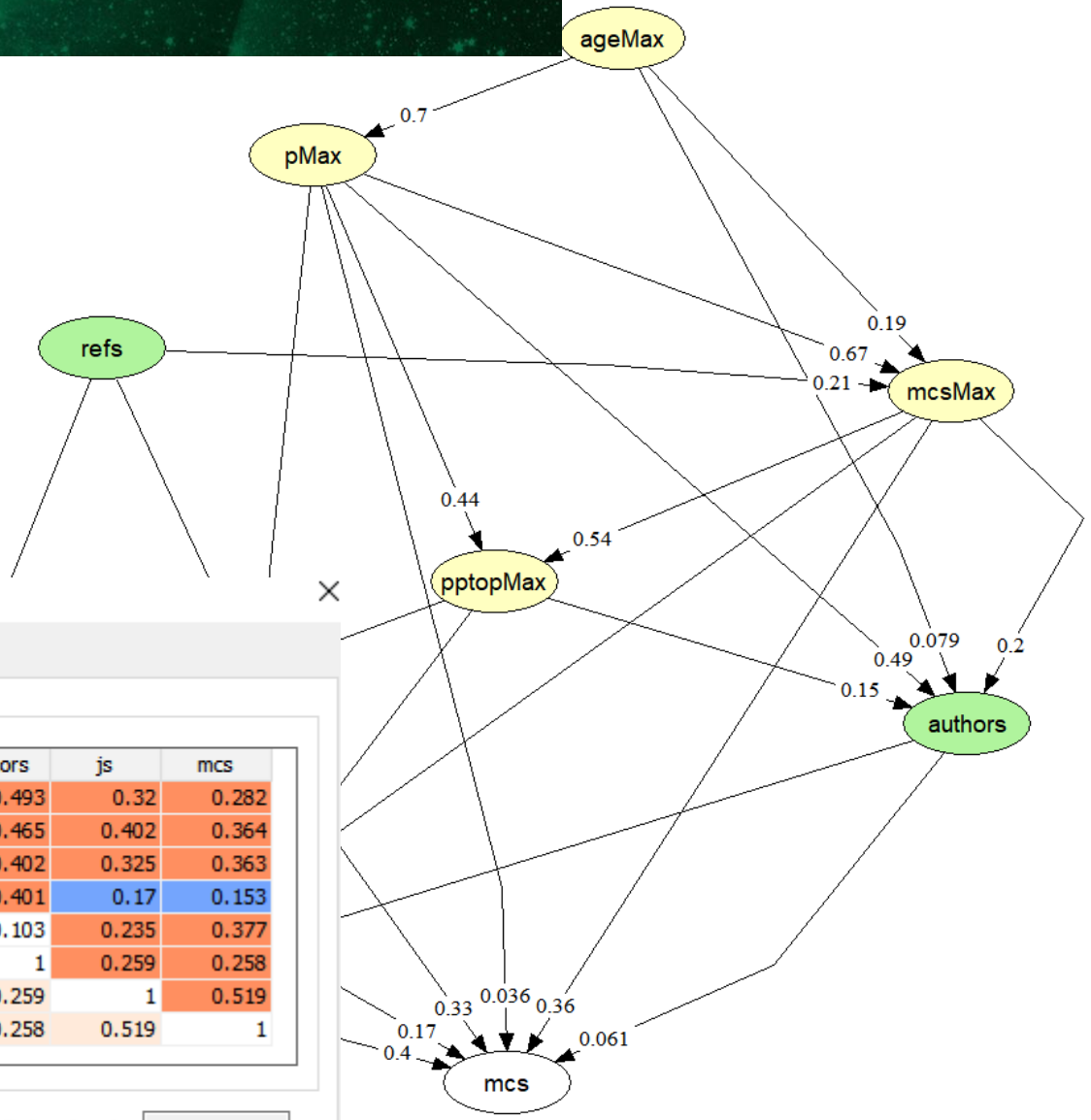
robust



The robustness of the BN structure

- bootstrapped the data 200 times
- BN structures are obtained for each of the bootstrap sample
- strength of each arc is measured by the empirical frequency of that arc over the set of all BNs obtained from each bootstrap sample

Uninet



Correlation Matrices

Bayes net | Empirical normal | Empirical | Determinants

Empirical normal rank correlation matrix

	pMax	mcsMax	pptopMax	ageMax	refs	authors	js	mcs
pMax	1	0.673	0.437	0.695	0.0888	0.493	0.32	0.282
mcsMax	0.673	1	0.655	0.573	0.2	0.465	0.402	0.364
pptopMax	0.437	0.655	1	0.28	0.181	0.402	0.325	0.363
ageMax	0.695	0.573	0.28	1	0.0318	0.401	0.17	0.153
refs	0.0888	0.2	0.181	0.0318	1	0.103	0.235	0.377
authors	0.493	0.465	0.402	0.401	0.103	1	0.259	0.258
js	0.32	0.402	0.325	0.17	0.235	0.259	1	0.519
mcs	0.282	0.364	0.363	0.153	0.377	0.258	0.519	1

Less <<

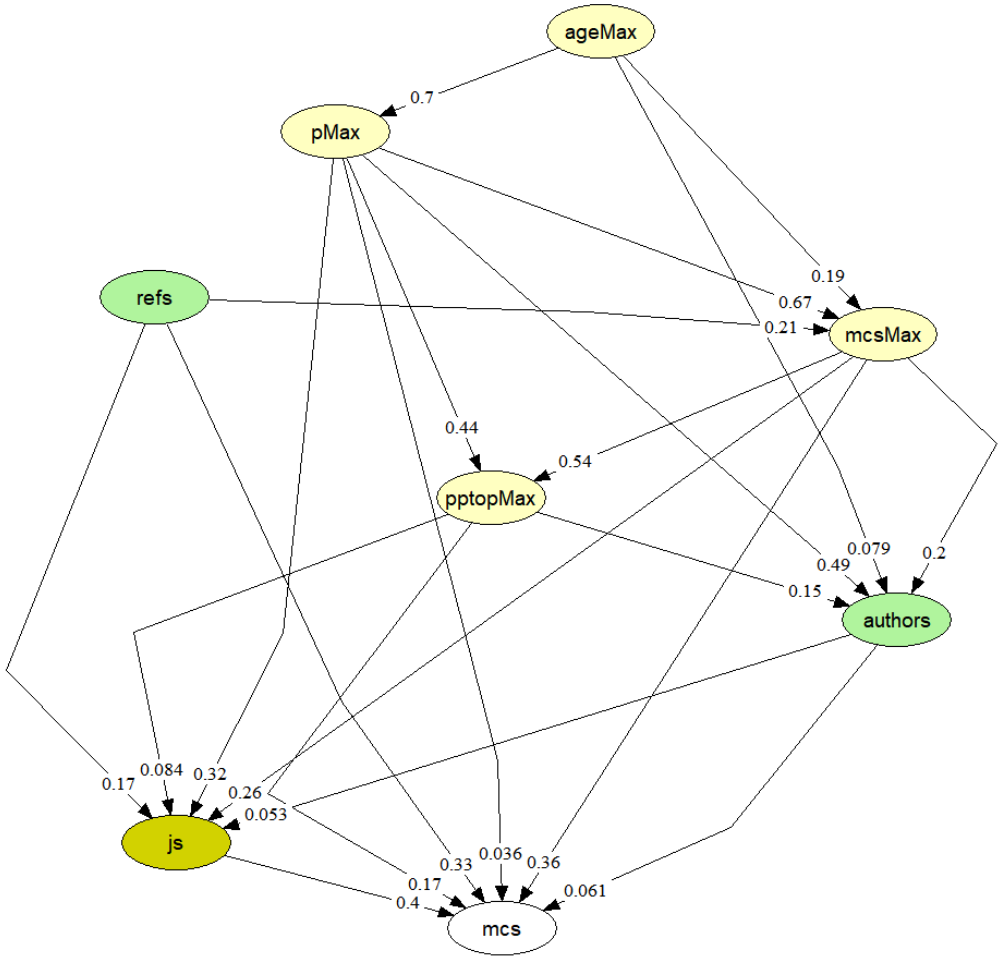
Determinant 0.0491493

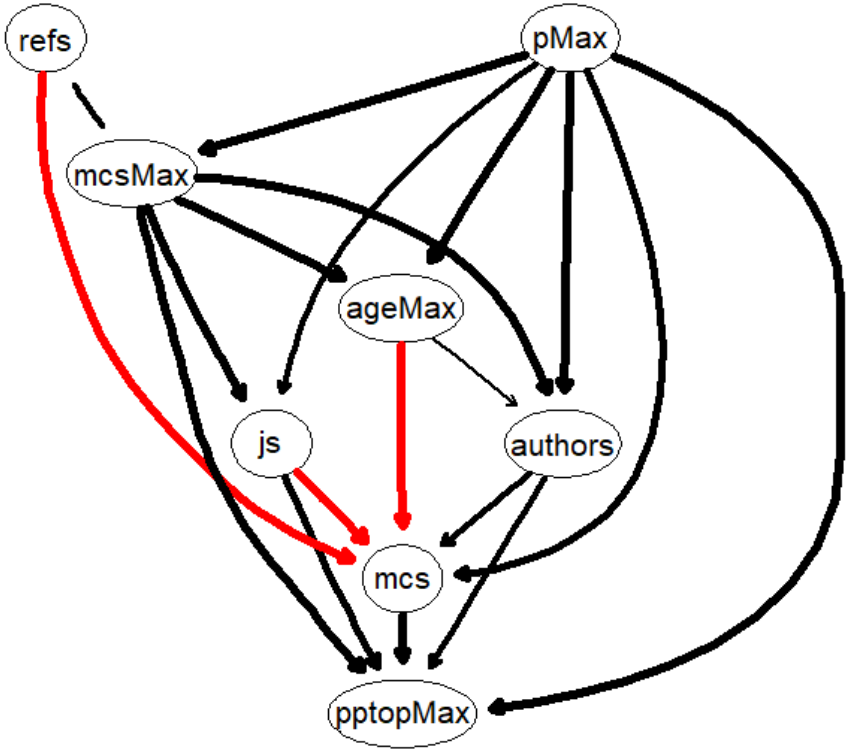
☑ Highlight parent-child correlations

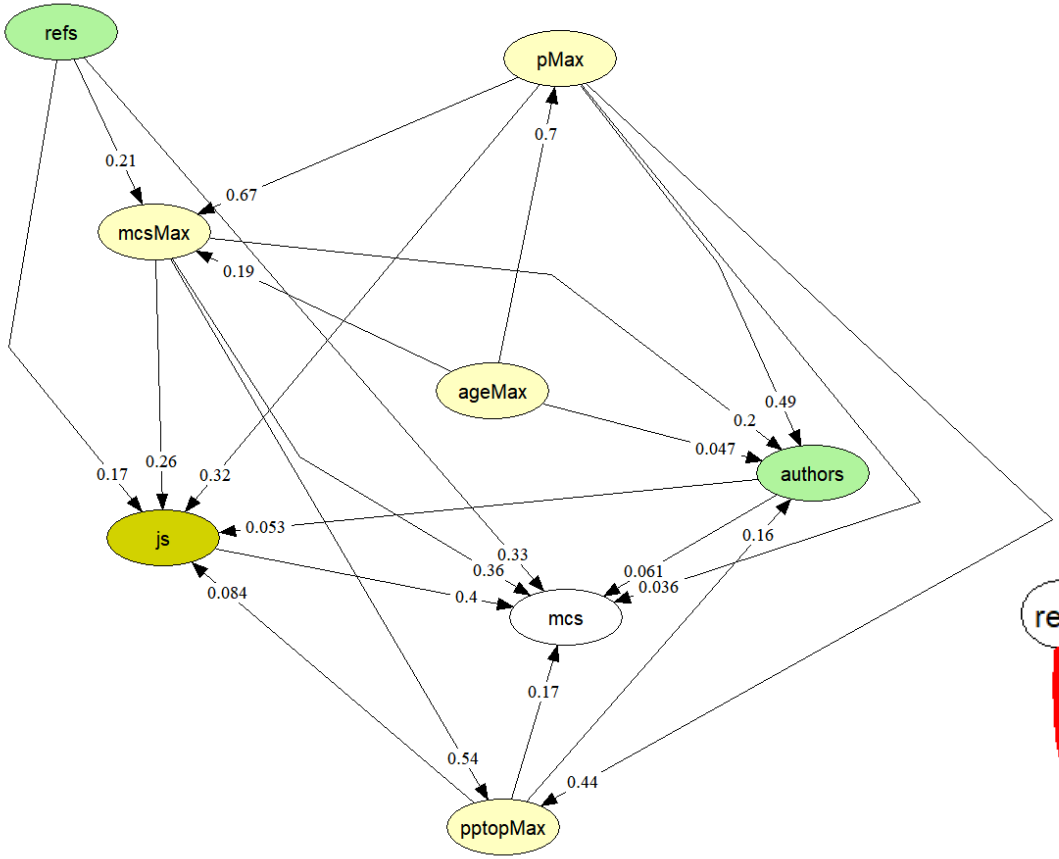
Export

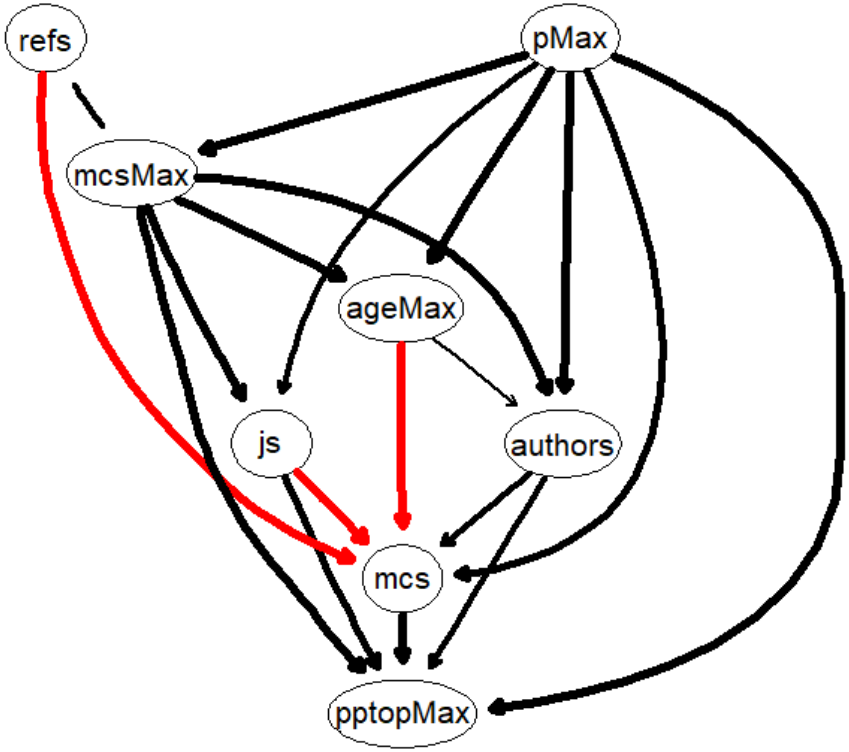
☑ Highlight the "k" highest non parent-child correlations 4

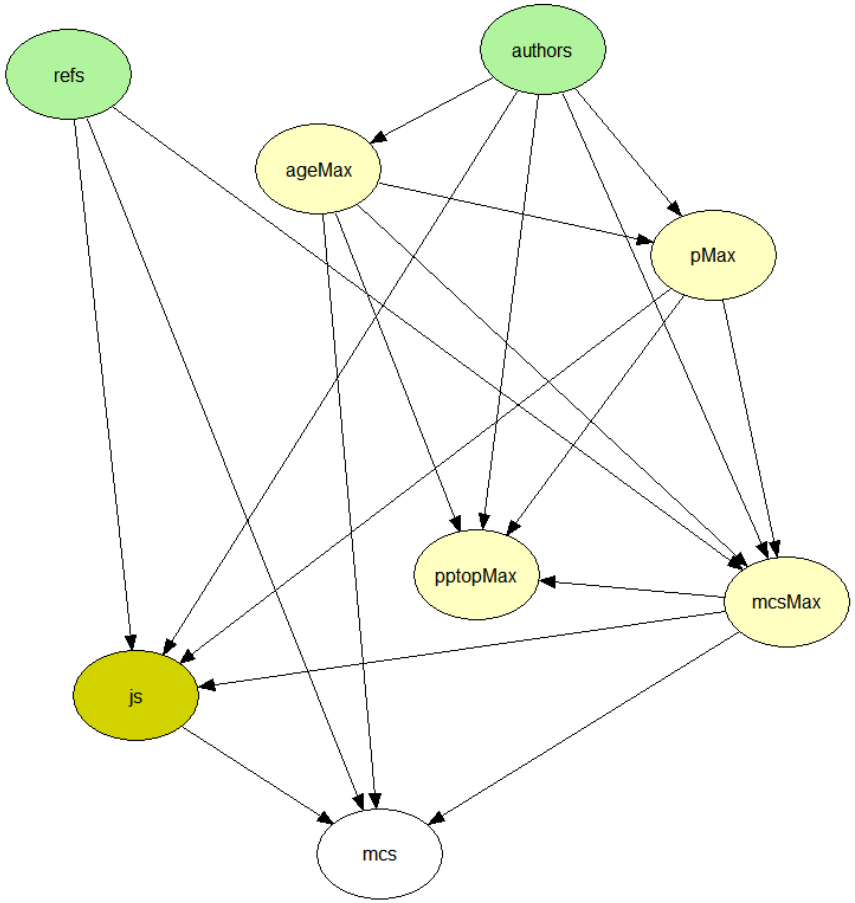
Update

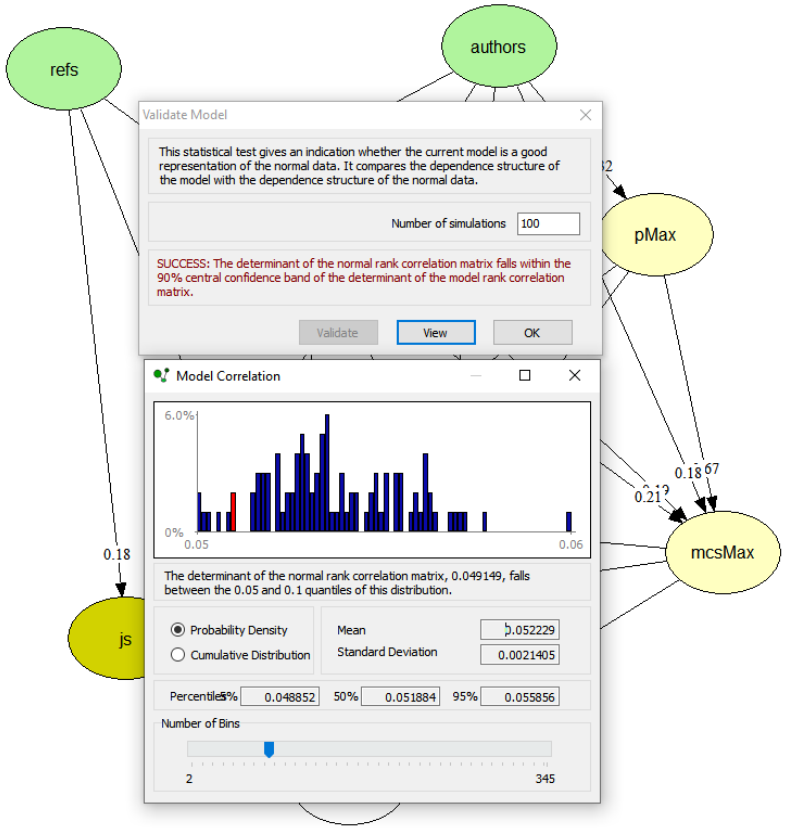

ageMax
pMax
refs
mcsMax
pptopMax
authors
js
mcs
0.7
0.19
0.67
0.21
0.44
0.54
0.49
0.079
0.2
0.15
0.17
0.084
0.32
0.26
0.053
0.33
0.036
0.36
0.17
0.4
0.061

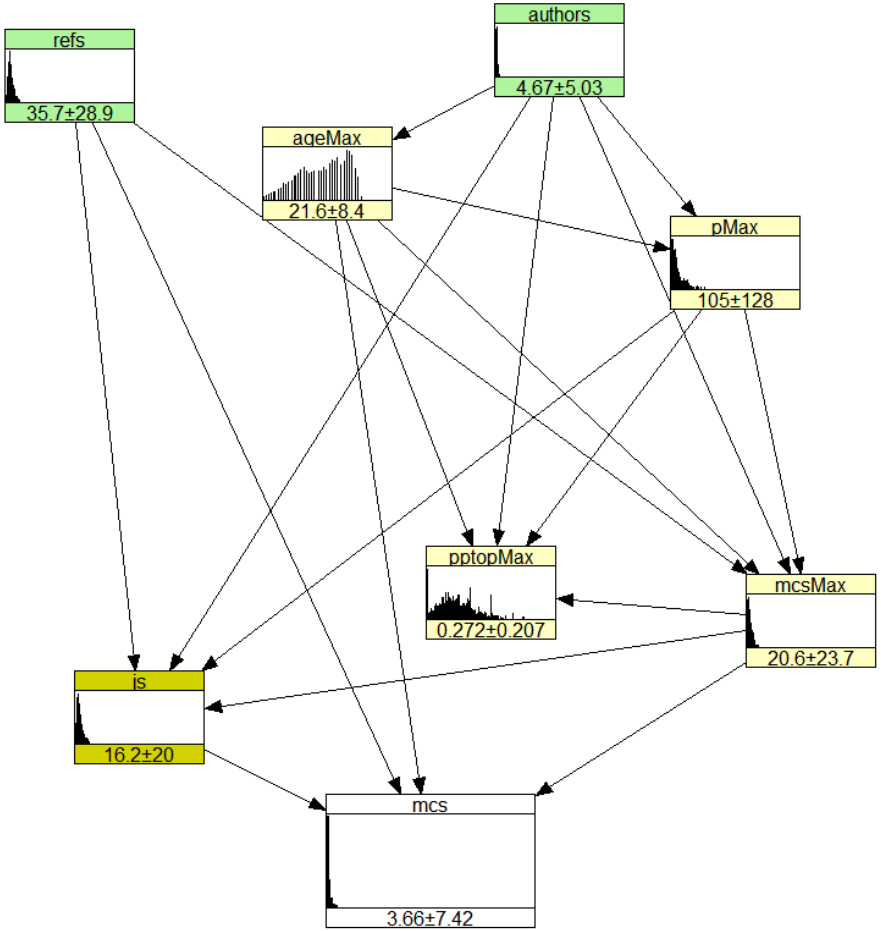

refs
pMax
mcsMax
ageMax
js
authors
mcs
pptopMax

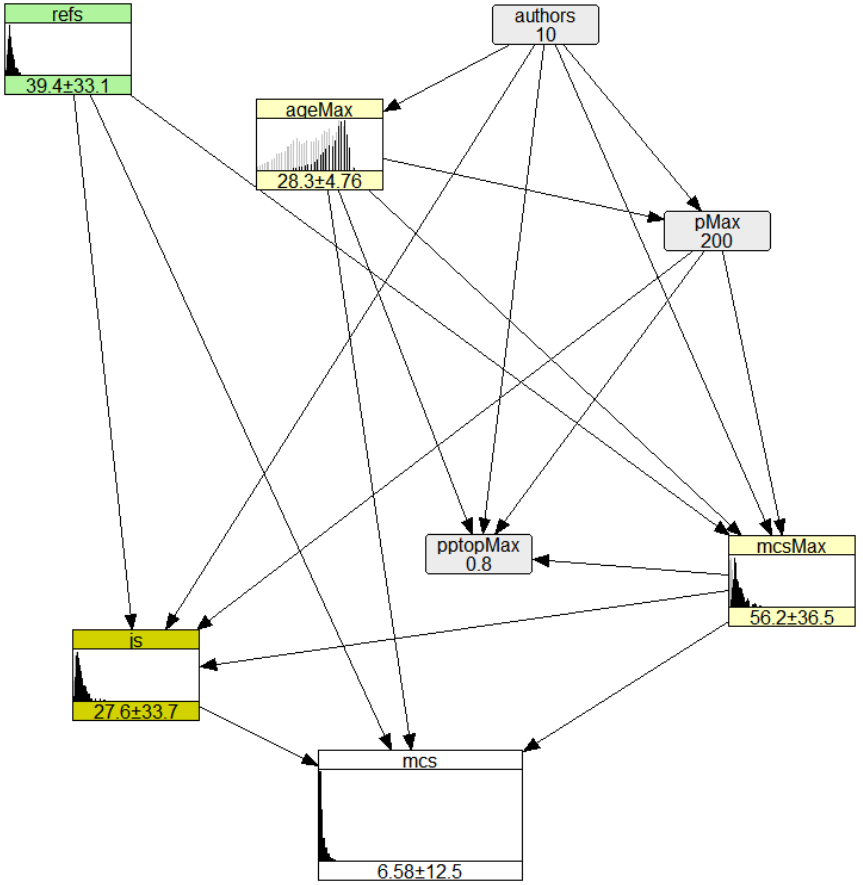

refs
pMax
mcsMax
ageMax
authors
js
mcs
pptopMax
0.21
0.67
0.7
0.19
0.49
0.2
0.047
0.17
0.26
0.32
0.053
0.16
0.33
0.36
0.061
0.036
0.4
0.084
0.17
0.54
0.44

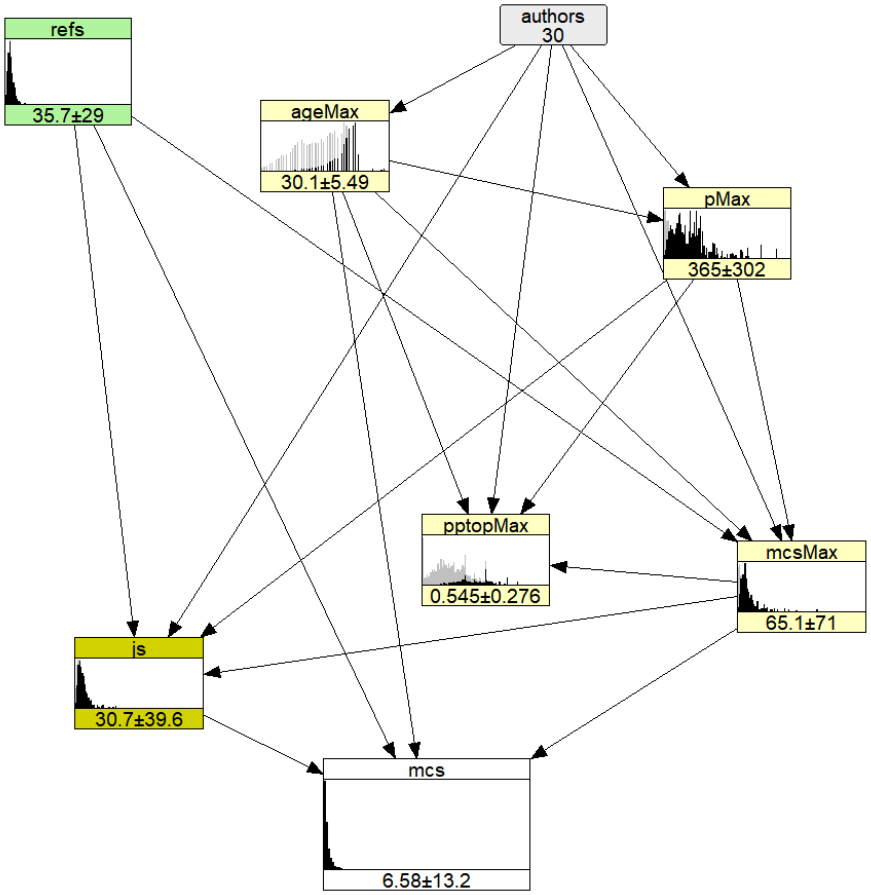

refs
pMax
mcsMax
ageMax
js
authors
mcs
pptopMax

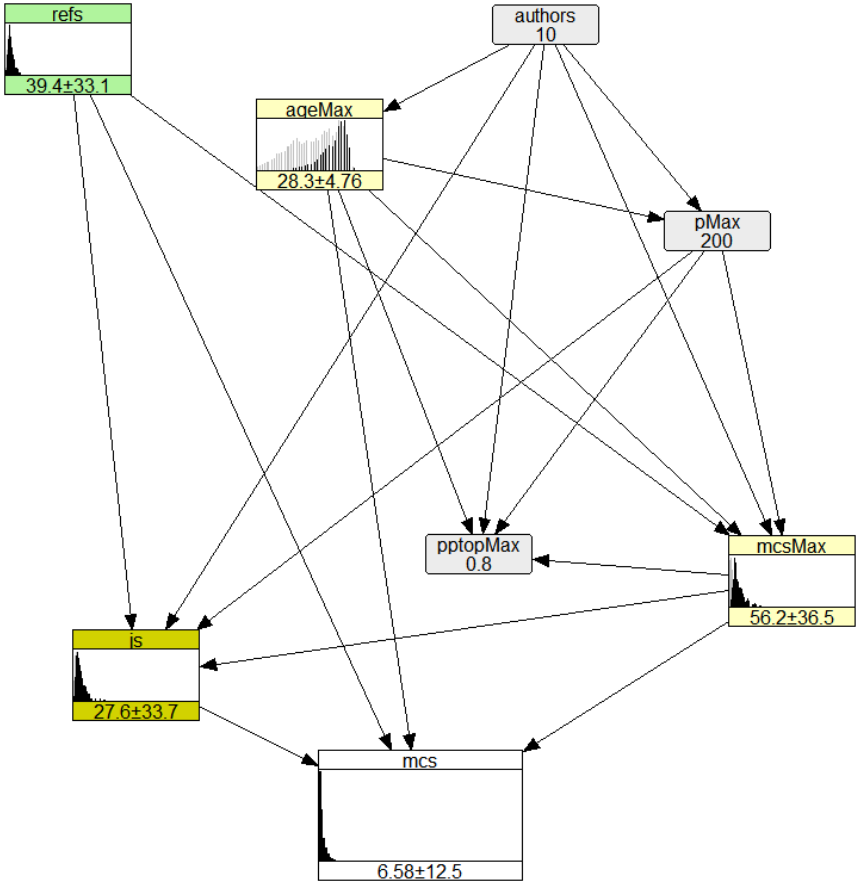

refs
authors
ageMax
pMax
pptopMax
mcsMax
js
mcs

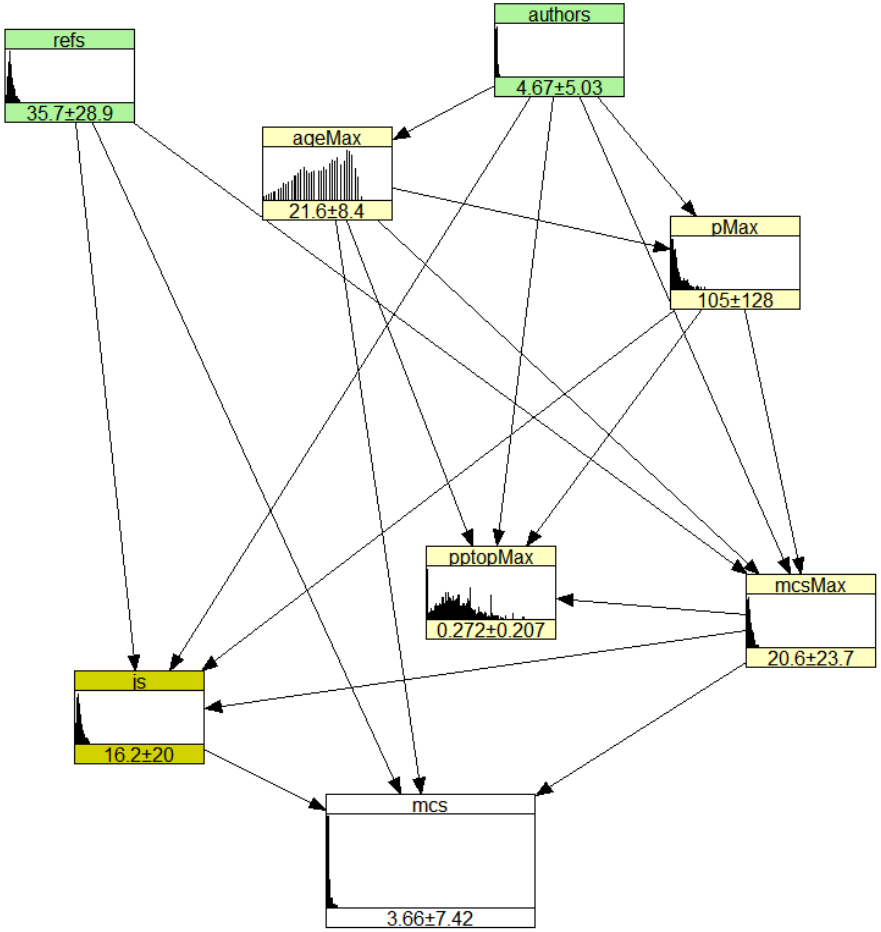

refs
authors
pMax
mcsMax
js
0.18
Validate Model
This statistical test gives an indication whether the current model is a good representation of the normal data. It compares the dependence structure of the model with the dependence structure of the normal data.
Number of simulations
100
SUCCESS: The determinant of the normal rank correlation matrix falls within the 90% central confidence band of the determinant of the model rank correlation matrix.
Validate
View
OK
Model Correlation
6.0%
0%
0.05
0.06
The determinant of the normal rank correlation matrix, 0.049149, falls between the 0.05 and 0.1 quantiles of this distribution.
Probability Density
Cumulative Distribution
Mean
Standard Deviation
0.052229
0.0021405
Percentiles
5%
0.048852
50%
0.051884
95%
0.055856
Number of Bins
2
345

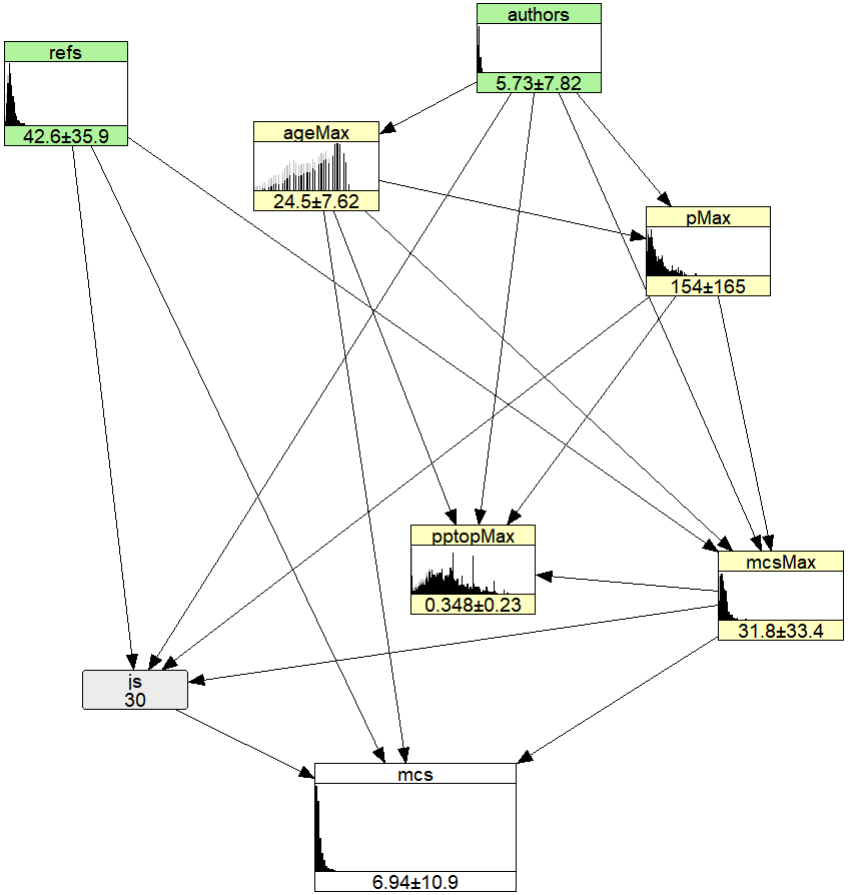

refs
35.7±28.9
authors
4.67±5.03
ageMax
21.6±8.4
pMax
105±128
pptopMax
0.272±0.207
mcsMax
20.6±23.7
js
16.2±20
mcs
3.66±7.42


refs
39.4±33.1
authors
10
ageMax
28.3±4.76
pMax
200
pptopMax
0.8
mcsMax
56.2±36.5
js
27.6±33.7
mcs
6.58±12.5


refs
35.7±29
authors
30
ageMax
30.1±5.49
pMax
365±302
pptopMax
0.545±0.276
mcsMax
65.1±71
js
30.7±39.6
mcs
6.58±13.2


refs
39.4±33.1
authors
10
ageMax
28.3±4.76
pMax
200
pptopMax
0.8
mcsMax
56.2±36.5
js
27.6±33.7
mcs
6.58±12.5


refs
35.7±28.9
authors
4.67±5.03
ageMax
21.6±8.4
pMax
105±128
pptopMax
0.272±0.207
mcsMax
20.6±23.7
js
16.2±20
mcs
3.66±7.42


refs
42.6±35.9
authors
5.73±7.82
ageMax
24.5±7.62
pMax
154±165
pptopMax
0.348±0.23
mcsMax
31.8±33.4
js
30
mcs
6.94±10.9

Computational Statistics (2021) 36:2009–2031
https://doi.org/10.1007/s00180-020-00999-9

ORIGINAL PAPER

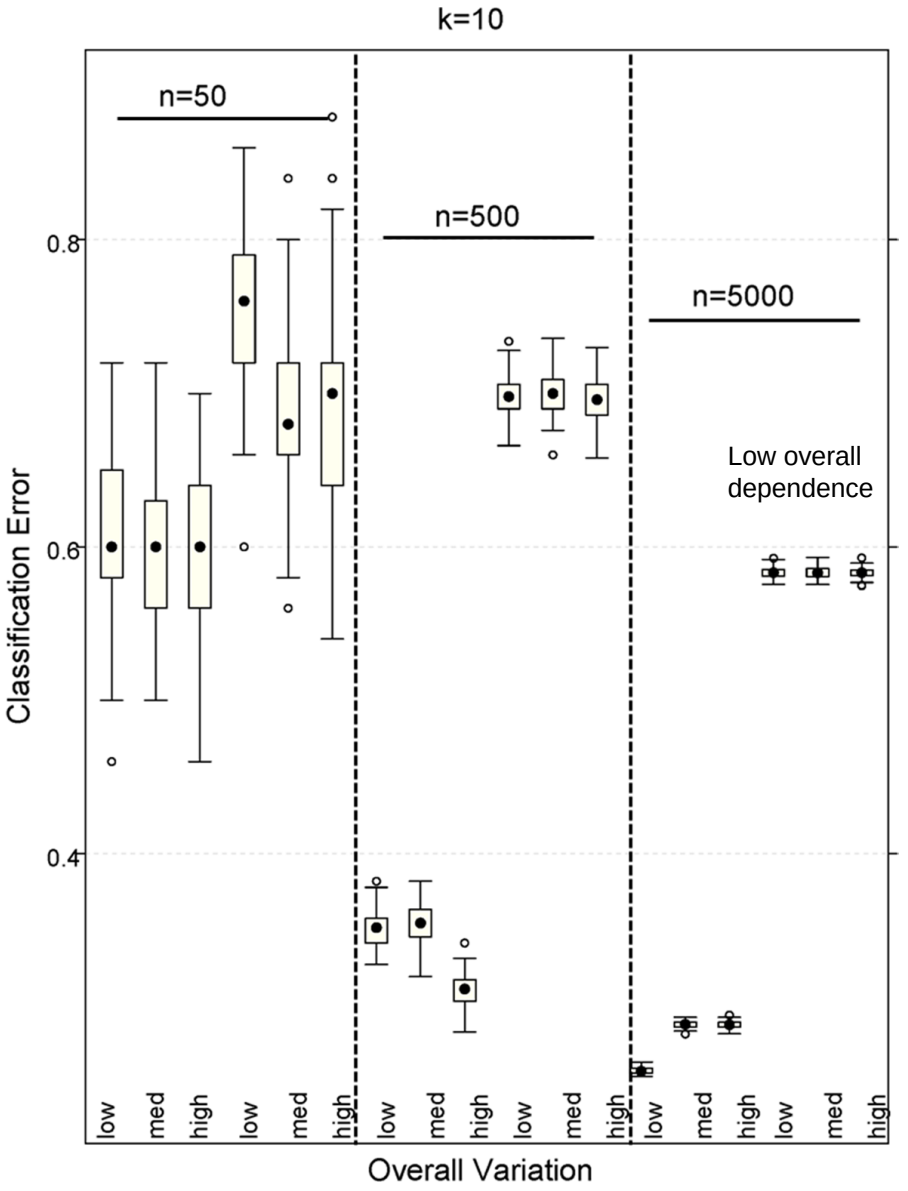
Check for updates

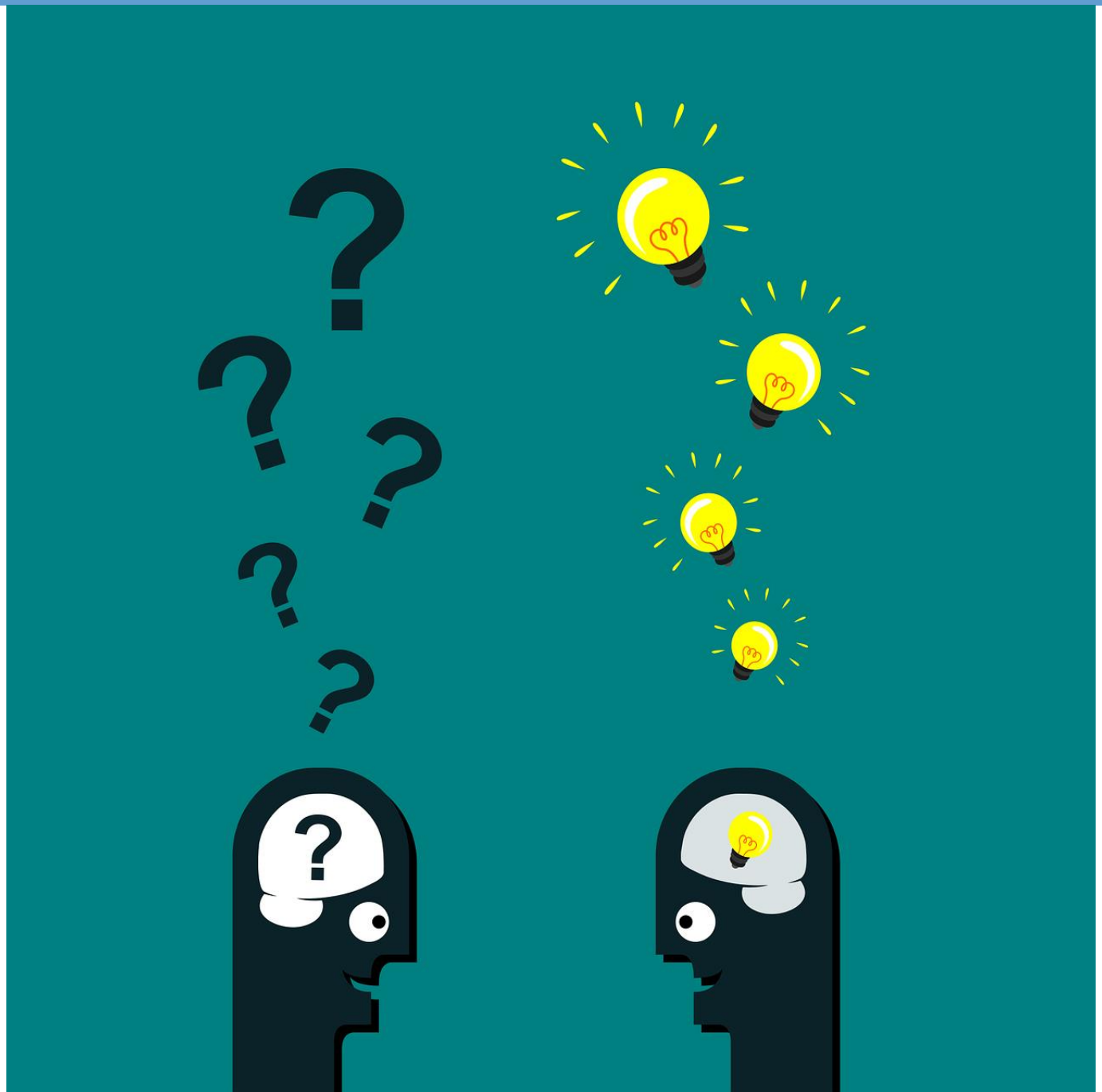
What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?

Bruce G. Marcot[1] · Anca M. Hanea[2]

Received: 8 November 2019 / Accepted: 4 June 2020 / Published online: 13 June 2020
© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2020

- Overall variation CoV = 1.1658
  - between medium and high (closer to high)
- Overall dependence det ~$10^{(-2)}$
  - "medium" dependence
- n = 10725, 8 RV instead of 10
- Classification error 0.624 (std = 0.005)





anca.hanea@unimelb.edu.au & g.f.nane@tudelft.nl

All variables were discretized into 4 states using Haremink's Algorithm [1]

[1] Hartemink AJ (2001) Principled computational methods for the validation and discovery of genetic regulatory networks. PhD Dissertation, Massachusetts Institute of Technology, Cambridge, MA