**Article 25fa pilot End User Agreement**

This publication is distributed under the terms of Article 25fa of the Dutch Copyright Act (Auteurswet) with explicit consent by the author. Dutch law entitles the maker of a short scientific work funded either wholly or partially by Dutch public funds to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed under The Association of Universities in the Netherlands (VSNU) ‘Article 25fa implementation’ pilot project. In this pilot research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. Please note that you are not allowed to share this article on other platforms, but can link to it. All rights remain with the author(s) and/or copyrights owner(s) of this work. Any use of the publication or parts of it other than authorised under this licence or copyright law is prohibited. Neither Radboud University nor the authors of this publication are liable for any damage resulting from your (re)use of this publication.

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the Library through email: copyright@ubn.ru.nl, or send a letter to:

University Library  
Radboud University  
Copyright Information Point  
PO Box 9100  
6500 HA Nijmegen

You will be contacted as soon as possible.
The stablespec package for causal discovery on cross-sectional and longitudinal data in R

Ridho Rahmadi, Perry Groot, Tom Heskes

Department of Informatics, Universitas Islam Indonesia, Indonesia
Data Science, ICIS, Radboud University Nijmegen, The Netherlands

Article history:
Received 16 April 2017
Revised 25 September 2017
Accepted 16 October 2017
Available online 17 November 2017

Keywords:
Causal discovery
Structural equation modeling
R package

The R package stablespec is an implementation of our method stable specification search. The method aims at causal discovery on both cross-sectional and longitudinal data through stable specification search in constrained structural equation models.

© 2017 Elsevier B.V. All rights reserved.

Table 1
Software metadata.

<table>
<thead>
<tr>
<th>Software metadata description</th>
<th>Current software version</th>
<th>Permanent link to executables of this version</th>
<th>Legal Software License</th>
<th>Operating System</th>
<th>Installation requirements &amp; dependencies</th>
<th>Link to user manual</th>
<th>Support email for questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(executable) Software metadata description</td>
<td>0.3.0</td>
<td><a href="https://github.com/Neurocomputing/NEUCOM-D-17-00979">https://github.com/Neurocomputing/NEUCOM-D-17-00979</a></td>
<td>MIT License</td>
<td>Linux, OS X, Microsoft Windows</td>
<td>R version 3.1.0 or higher, package dependencies: ggm, matrixcalc, sem, nsga2R, graph, Rgraphviz, methods, polycor, foreach</td>
<td><a href="https://cran.r-project.org/web/packages/stablespec/stablespec.pdf">https://cran.r-project.org/web/packages/stablespec/stablespec.pdf</a></td>
<td><a href="mailto:r.rahmadi@cs.ru.nl">r.rahmadi@cs.ru.nl</a></td>
</tr>
</tbody>
</table>

Table 2
Code metadata.

<table>
<thead>
<tr>
<th>Code metadata description</th>
<th>Current code version</th>
<th>Permanent link to code</th>
<th>Legal Code License</th>
<th>Compilation requirements, operating environments &amp; dependencies</th>
<th>Link to developer documentation/manual</th>
<th>Support email for questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3.0</td>
<td><a href="https://github.com/Neurocomputing/NEUCOM-D-17-00979">https://github.com/Neurocomputing/NEUCOM-D-17-00979</a></td>
<td>MIT License</td>
<td>R version 3.1.0 or higher, package dependencies: ggm, matrixcalc, sem, nsga2R, graph, Rgraphviz, methods, polycor, foreach</td>
<td><a href="https://cran.r-project.org/web/packages/stablespec/stablespec.pdf">https://cran.r-project.org/web/packages/stablespec/stablespec.pdf</a></td>
<td><a href="mailto:r.rahmadi@cs.ru.nl">r.rahmadi@cs.ru.nl</a></td>
</tr>
</tbody>
</table>

* Corresponding author at: Data Science, ICIS, Radboud University Nijmegen, The Netherlands.
E-mail address: r.rahmadi@cs.ru.nl (R. Rahmadi).

https://doi.org/10.1016/j.neucom.2017.10.064
0925-2149/© 2017 Elsevier B.V. All rights reserved.
1. Introduction

Causal modeling aims to understand the underlying mechanisms by which variables in data relate to each other in terms of causal relations. It can also be seen as an attempt to find a generative model [1]. Causal modeling often turns out to be an essential problem in many fields, e.g., [2,3]. In the medical domain, for example, revealing causal relationships may lead to enhancement of clinical practice, e.g., the development of treatment and medication [4]. Stable specification search is a novel causal discovery method based on [5], for cross-sectional data (S3C), and [6], for longitudinal data (S3L). The method is designed to overcome two problems in causal modeling: the issue that the number of possible models is super-exponential in the number of variables and the instability in model selection, i.e., that a slight change in the data can lead to a significant change in the final model. In this paper, we describe the R package stablespec, which contains an implementation of S3C/L.

Our package stablespec attempts to infer the causal structure that best matches a given data set. It implements S3C/L [5,6], which in high-level terms works as follows. S3C/L models causal relations between variables using Structural Equation Models (SEMs) and uses an exploratory approach (i.e., without specifying an initial hypothesis) to search over the model space. S3C/L evaluates models according to two objectives: the model fit and the model complexity. Since both objectives are often conflicting, S3C/L uses a multi-objective optimization approach, called Non-dominated Sorting Genetic Algorithm II (NSGA-II), to search for Pareto optimal models. In addition, in order to deal with the inherent instability of structure estimation from finite data, S3C/L adopts the concept of stability selection using subsampling and selection algorithms [7].

Fig. 1. Plots of edge stability (blue line) and causal path stability (green line is for causal path with length one and red line is for any length) between two variables. The green line between variables 1 and 4 is covered by the blue line as causal path stability with length one and the edge stability have the same value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Implementation and functionalities

The package provides a main function and several support functions. The main function, stablespec, is used for searching optimal model structures and computing stabilities. Parallel computation is facilitated through parallel backend registration. Some additional functions are provided to increase usability, e.g., modelPop for generating random SEM models, repairCyclicModel for repairing a cyclic model so as to be acyclic, plotStability for visualizing the stability of model structures, getModelFitness for scoring SEM models, and dataReshape for reshaping longitudinal data. We add data sets for users to be able to explore stablespec directly without loading external data. Documentation is bundled alongside the package, giving the user detailed guidelines, e.g., each function is accompanied by a running example that users can adopt to their case.

The stablespec package is available at the Comprehensive R Archive Network (CRAN) with MIT license. The package depends on R at least version 3.1.0 and some other R packages, for instance, ggm, sem, nsga2R, polycor, foreach, graph, and Rgraphviz. As the mentioned package dependencies are on both CRAN and Bioconductor, the stablespec can be installed from the R console by typing setRepositories(ind=1:2) and then install.packages("stablespec") in the next line.

3. Experimental result

To demonstrate the package, we consider a data set which describes phenotypic information of children with Attention Deficit Hyperactivity Disorder (ADHD) [8]. The data set consists of 221 subjects, with eight variables as described in Fig. 2. The following example assumes that the package stablespec has been loaded (see the documentation for details on the function arguments). The first two lines are for parallel computation, which requires the packages parallel and doParallel; to compute sequentially, simply remove these lines.

```r
> cl <- makeCluster(detectCores())
> registerDoParallel(cl)
> result <- stablespec(theData=read.csv("ADHD.csv"),
  nSubset=100, nPop=120, longitudinal=FALSE,
  mixture=TRUE,
  consMatrix=matrix(c(2, 1, 3, 1, 4, 1, 5, 1,
                     6, 1, 7, 1, 8, 1), 7, 2, byrow=TRUE),
  toPlot=FALSE)
```

Fig. 1 is an output example using plotStability which shows the stability graphs between some variables. Model complexity is on the x-axis and selection probability on the y-axis. The horizontal line is the boundary of the selection probability \(\pi_{sel}(\text{argument threshold of the function stablespec})\) and the vertical line is the boundary of the model complexity \(\pi_{nc}\) (set to the level of the model complexity at which the minimum average

---

Bayesian Information Criterion (BIC) score is found). The blue line represents the edge stability which constitutes relations between pairs of variables regardless of the direction, while the green and red lines represent the causal path stability with length one and with any length, respectively, which constitute causal relations from variables to other variables. Relevant structures are defined to be those edges and causal paths with a selection probability higher than or equal to \( \pi_{\text{sel}} \) and with a model complexity lower than or equal to \( \pi_{\text{bic}} \). Thus in Fig. 1, the relevant structures are represented by lines that pass through the relevant (top-left) region of the plot. In addition, plotStability returns plots of the aggregated edge stability (Fig. 2b) and the aggregated causal path stability (Fig. 2b). The R scripts to generate the plots in Figures 1, 2a, 2b using plotStability are as follows.

```r
> plotStability(listOfFronts =
result$listOfFronts, stableCausal =
result$causalStab, stableCausal_l1 =
result$causalStab_l1, stableEdge =
result$edgeStab, longitudinal = FALSE)
```

Fig. 2c provides a visualization of the relevant structures (visualization is not part of the package, but left to one's favorite drawing software) obtained through the steps described in [5]. The causal model shown in Fig. 2c is corroborated by studies reported in [9].

4. Conclusions

As an R package, stablespec gives users the flexibility to replicate and extend the algorithms within the R framework. Moreover, comparisons of S3C/L (stablespec) with some other algorithms (R package pcalg and a standalone software TETRAD) show that S3C/L achieve significant improvements over alternative approaches in retrieving causal relations [5,6].

5. Required metadata

5.1. Current executable software version

See Table 1.

5.2. Current code version

See Table 2.

References

variables, activity

Ridho Rahmadi received master degrees in Computer Science (Artificial Intelligence) in 2012 from Czech Technical University in Prague and Johannes Kepler University Linz, Austria. He is currently doing his Ph.D. at the Data Science Group, Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands, with by Tom Heskes and Perry Groot as advisors. He had been a visiting scholar at the Carnegie Mellon University, Pittsburgh, USA. His research interests are in the scope of causal modeling, machine learning, evolutionary algorithm, and measurement error.

Perry Groot received master degrees in Computer Science (Artificial Intelligence) and Mathematics (topology) in 1998 and a Ph.D. degree from the Department of Artificial Intelligence, Vrije University Amsterdam in 2003. Currently, he is a researcher at the Institute for Computing and Information Sciences at the Radboud University Nijmegen, the Netherlands. His research interests include Bayesian reasoning applications, causal modeling, particularly the use of Gaussian processes for preference learning and function optimization.

Tom Heskes is a Professor in Artificial Intelligence, and he leads the Data Science Group at the Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands. He is further affiliated Principal Investigator at the Donders Institute for Brain, Cognition and Behaviour. His research is on Bayesian inference, machine learning, and neural networks. He is involved in several projects that concern applications in, among others, neuroimaging, genomics, and bioinformatics. He has published over 150 peer-reviewed research papers in the above areas. He is currently an Associate Editor of five journals.