The Journal of Open Source Software

molic: An R package for multivariate outlier detection in contingency tables

Mads Lindskou[1, 2]

1 Department of Mathematical Sciences, Aalborg University, Denmark **2** Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

DOI: 10.21105/joss.01665

Software

- Review
- Repository
- Archive

Submitted: 16 August 2019
Published: 10 October 2019

Summary

Outlier detection is an important task in statistical analyses. An outlier is a case-specific unit since it may be interpreted as natural extreme noise in some applications, whereas in other applications it may be the most interesting observation. The **molic** package has been written to facilitate the novel outlier detection method in high-dimensional contingency tables (Lindskou, Eriksen, & Tvedebrink, 2019). In other words, the method works for data sets in which all variables are *categorical*, implying that they can only take on a finite set of values (also called *levels*).

The software uses decomposable graphical models (DGMs), where the probability mass function can be associated with an interaction graph, from which conditional independences among the variables can be inferred. This gives a way to investigate the underlying nature of outliers. This is also called *understandability* in the literature. Outlier detection has many applications including areas such as

- Fraud detection
- Medical and public health
- Anomaly detection in text data
- Fault detection (on critical systems)
- Forensic science

The Method

The method can be described by the **outlier test** procedure below. Assume we are interested in whether or not a new observation z is an outlier in some data set D. First an *interaction graph* G is fitted to the variables in D; a decomposable undirected graph that describes the association structure between variables in D. If the assumption that z belongs to D is true, z should be included in D. Denote by D_z the new data set including z. Finally the outlier model M is constructed using G and D_z from which we can query the p-value, p, for the test about z belonging to D. If p is less than some chosen threshold (significance level), say 0.05, z is declared an outlier in D.

```
outlier test ($D$: data, $z$ : new obs.)
    $G :=$ fit_graph(D)
    $D_z := D$ with $z$ included
    $M :=$ fit_outlier(D_ z, G)
    $p :=$ pval(M, deviance(M, z))
    return $p$
end outlier test
```

The `fit_graph` algorithm has three ways of fitting a graph. The `fwd` type (which is default) is an implementation of the efficient step-wise selection procedure (Deshpande, Garofalakis, & Jordan, 2001) used for model selection in decomposable graphs. There is also a backward, `bwd`, type and finally it is also possible to fit a tree interaction graph, i.e. only first order associations.

The `fit_graph` function can be used to explore dependencies between any kind of discrete variables and make statements about conditional dependencies and independencies. A thorough description of the outlier detection method and how to use the software can be found at https://mlindsk.github.io/molic/.

Expert Knowledge

If one has prior knowledge of the underlying nature of the association between variables, this can easily be exploited. One can choose to model only the relationship between variables which have no other associations to any of the remaining variables. This will result in a number of interaction graphs which can then be unified as the union of these graphs. This approach was taken in the example below.

A Use Case in Forensic Science

Recently, advances in DNA sequencing have made it possible to sequence short segments of DNA (< 200 basepairs) including two or more single nucleotide polymorphisms (SNPs). These are called *microhaplotypes* (or microhaps for short) (K. K. Kidd et al., 2014). They have been demonstrated to be well suited for ancestry assessment in the forensic science community. The short distance between SNPs within a microhap implies that recombination among them rarely occurs. Hence, the methodology of T. Tvedebrink, Eriksen, Mogensen, & Morling (2018) can not be used as this assumes mutual independence of the SNPs within a population (corresponding to the null graph with no edges).

In Lindskou et al. (2019) the **molic** package was used to detect outliers in microhap data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). This data contains DNA profiles from five different continental regions (CRs); Europe (EUR), America (AMR), East Asia (EAS), South Asia (SAS) and Africa (AFR).

Consider for example the region SAS as the hypothesized region and all profiles in AFR as profiles to be tested against the hypothesis that their origin is SAS. Two different interaction graphs are used; G which is the result of using the `fit_graph` algorithm with type `fwd` and $G^{\emptyset}$ where all microhap SNPs are assumed to be independent (a graph with no edges). The proportion of profiles from AFR that are outliers in SAS according to the model, is 1 for G and only 0.834 for $G^{\emptyset}$, see Table 1. The outlier test was conducted for all pairs of continental regions. It is seen, that G outperforms $G^{\emptyset}$ in general and the dependency between microhap SNPs cannot be neglected. All tests were conducted with a significance level of 0.05.

The Journal of Open Source Software

H_0 / z	EUR	EAS	AMR	SAS	AFR
EUR	0.054 0.046	1 1	0.191 0.145	0.509 0.231	1 1
EAS	1 1	0.054 0.063	0.994 0.994	0.966 0.980	1 1
AMR	0.778 0.697	1 1	0.095 0.049	0.769 0.565	1 1
SAS	0.922 0.863	1 1	0.710 0.620	0.037 0.047	1 1
AFR	1 0.998	1 1	0.997 0.918	1 0.834	0.101 0.106

▨ G □ $G^{\emptyset}$

Table 1: Performance matrix of outlier tests.

Another model that could have been considered is the saturated model (a complete graph). This is the equivalent of estimating probabilities using the naive frequency counts in the data. For one, it does not (necessarily) capture the biological association between SNPs and second it would, in general, require an enormous amount of data to obtain valid estimates.

References

Deshpande, A., Garofalakis, M., & Jordan, M. I. (2001). Efficient stepwise selection in decomposable models. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 128–135). Morgan Kaufmann Publishers Inc.

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagacé, R., Chang, J., Wootton, S., Haigh, E., et al. (2014). Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Science International: Genetics*, *12*, 215–224. doi:10.1016/j.fsigen.2014.06.014

Lindskou, M., Eriksen, P. S., & Tvedebrink, T. (2019). Outlier detection in contingency tables using decomposable graphical models. *Scandinavian Journal of Statistics*. doi:10.1111/sjos.12407

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., & Morling, N. (2018). Weight of the evidence of genetic investigations of ancestry informative markers. *Theoretical Population Biology*, *120*, 1–10. doi:10.1016/j.tpb.2017.12.004