4B1-R-2-4

Bootstrapping confidence intervals in linear non-Gaussian causal model

Kittitat Thamvitayakul　清水　昌平　鷲尾　隆　田代　竜也
Shohei Shimizu　Takashi Washio　Tatsuya Tashiro

大阪大学産業科学研究所
The Institute of Scientific and Industrial Research (ISIR) Osaka University

We consider a problem of finding significant connection strengths of variables in a linear non-Gaussian causal model called LiNGAM. In our previous work, bootstrapping confidence intervals of connection strengths were simultaneously computed in order to test their statistical significance. However, such a naive approach raises the multiple comparison problem which many directed edges are likely to be falsely found significant. Therefore, in this study, we tested two representative techniques of multiple testing correction approaches, the Bonferroni correction and Mandel's approach, then evaluated their performance. We found that both the Bonferroni correction and Mandel's approach are able to control the familywise error rate of the confidence intervals to be less than the significance level in LiNGAM.

1. Introduction

In causal analysis, confidence interval is used to determine whether an edge in the directed acyclic graph is the significant edge or not. To construct a confidence interval by using standard methods, an appropriate transformation or any other background knowledge is required [1]. However, in order to construct this confidence interval more simply, the bootstrap technique, which is one of resampling methods, could be used.

In previous work on LiNGAM [2], the bootstrap confidence interval was used to determine the significance of the elements in the adjacency matrix estimated by LiNGAM. However, due to the fact that those elements in the matrix contain the relations between a number of variables, the confidence intervals with appropriate range could not be constructed.

In the present study, we tried to use a multiple testing approach in order to construct a more appropriate confidence interval for each element in the adjacency matrix. We decided to use two multiple testing approaches, the Bonferroni correction [3], and the recently developed Mandel's approach [4], in this study.

2. Background

2.1 LiNGAM

In [5], LiNGAM is a model used for exploratory causal analysis. It is assumed that the data are generated from a process represented graphically by a directed acyclic graph, or DAG. Let b_{ij} be the connection strength from a variable x_j to x_i, if b_{ij} is non-zero then it means that there is a directed edge from variable x_j to x_i, and let $k(i)$ be a causal order of x_i in DAG so that no later variable determines or has a directed path on any earlier variable. Without loss of generality, each observed variable x_i is assumed to have zero mean. Then we have

$$x_i = \sum_{k(j)<k(i)} b_{ij}x_j + e_i, \tag{1}$$

where e_i are external influences that are continuous variables having *non-Gaussian* distributions with zero means and non-zero variances and are mutually independent. The independence assumption between e_i means that there is no latent confounding variable. The model (1) is rewritten in matrix form as

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}, \tag{2}$$

where the connection strength matrix or adjacency matrix $\mathbf{B}$ collects b_{ij} and the vectors $\mathbf{x}$ and $\mathbf{e}$ are p-dimensional vectors collecting x_i and e_i respectively, and $\mathbf{B}$ could be permuted to be lower triangular with all zeros on the diagonal by simultaneous equal row and column permutations according to a causal ordering $k(i)$.

The modeling purpose by using this model is to estimate the connection strength matrix $\mathbf{B}$ based on data $\mathbf{x}$ only. Note that since the external influences e_i have *non-Gaussian* distributions and are mutually independent, the model (2) is known to be identifiable.

2.2 DirectLiNGAM

In [2], a direct estimation method called DirectLiNGAM was proposed. DirectLiNGAM estimates causal orders one by one and eventually a causal ordering of all the variables. An exogenous variable is a variable with no parent and the corresponding row of $\mathbf{B}$ has all zeros. It can be at the top of such a causal ordering that makes $\mathbf{B}$ lower triangular with zeros on the diagonal.

In DirectLiNGAM, we continue removing the effect of the exogenous variable from the other variables by regressing it out. This procedure is iterated until all the variables are ordered. The point is how we can find an exogenous variable. The following lemma of [2] shows how it is possible:

連絡先: Kittitat Thamvitayakul, 大阪大学産業科学研究所, 567-0047 大阪府茨木市美穂ヶ丘８－１, kittitat@ar.sanken.osaka-u.ac.jp

Lemma 1 *Assume that the input data* $\boldsymbol{x}$ *strictly follows LiNGAM, that is, the model (2) with non-Gaussian external influences. This means that we assume that all the model assumptions are met and the sample size is infinite. Denote by* $r_i^{(j)}$ *the residual when* x_i *is regressed on* x_j*:* $r_i^{(j)} = x_i - \frac{cov(x_i,x_j)}{var(x_j)} x_j (i \neq j)$*. Then a variable* x_i *is exogenous if and only if* x_j *is independent of its residuals* $r_i^{(j)}$ *for all* $i \neq j$*.*

To evaluate independence between a variable x_j and its residuals $r_i^{(j)} (i \neq j)$, we first evaluate pairwise independence between the variable and each of the residuals using a kernel-based estimator of mutual information called KGV [6], which we denote by $KGV(x_i, r_i^{(j)})$, and subsequently compute the sum of the pairwise independence measures over the residuals. The non-negative estimator KGV asymptotically goes to zero if and only if the variables are independent [6]. Thus we obtain the following statistic to evaluate independence between a variable x_j and its residuals $r_i^{(j)}$:

$$T_{kernel}(x_j; U) = \sum_{i \in U, i \neq j} KGV(x_j, r_i^{(j)}), \tag{3}$$

where U denotes the set of the subscripts of variables x_i, that is, $U = \{1, \ldots, p\}$.

2.3 Multiple comparison approaches

Multiple comparison problem is a problem which occurs when one considers a group of hypotheses simultaneously. This problem increases the probability of rejecting even one of the true null hypotheses, known as familywise error rate or FWER, to exceed the groupwise significance level or α. Many statistical techniques have been used to develop in order to correct this problem. In this subsection, we mention two representative techniques of multiple comparison approaches: the Bonferroni correction [3] and Mandel's approach [4].

2.3.1 Bonferroni correction

The Bonferroni correction allows many confidence intervals to be constructed while assuring the groupwise significance level is still maintained [3].

In case of LiNGAM, if we have p variables then the size of adjacency matrix is $p \times p$. In other words, we construct $p^2 - p$ confidence intervals for the matrix $\mathbf{B}$ *excluding* the elements on the diagonal line. In order to construct confidence intervals for $p^2 - p$ elements with groupwise significance level $1 - \alpha$, one could construct each confidence interval with coefficient $1 - \frac{\alpha}{p^2 - p}$. Then for *each* b_{ij} of the adjacency matrix, we have

$$P\left(L_{ij} \leq b_{ij} \leq U_{ij}\right) \geq 1 - \frac{\alpha}{p^2 - p}, \tag{4}$$

where b_{ij} is each element in $p \times p$ adjacency matrix $\mathbf{B}$, and L_{ij} and U_{ij} are, respectively, the confidence lower bound and upper bound of the element b_{ij} calculated by the bootstrap percentile method, that is, L_{ij} is the element at $\left(\frac{1}{2} \cdot \frac{\alpha}{p^2 - p}\right)^{th}$ percentile and U_{ij} is the element at $100 \cdot \left(1 - \frac{1}{2} \cdot \frac{\alpha}{p^2 - p}\right)^{th}$ percentile in the ordered N replications of *each* b_{ij} generated by using the bootstrap method.

2.3.2 Mandel's approach

In [4], Mandel proposed an algorithm which is used to construct simultaneous $(1 - \alpha)$-bootstrap confidence intervals given only data $\mathbf{X}$. Note that each sample of $\mathbf{X}$ consists of p variables. When applying this algorithm to LiNGAM, we could find the upper limit of the confidence interval by constructing $p^2 - p$ confidence intervals with a groupwise significance level of $1 - \alpha$ as follows:

1. We generated N bootstrap samples from the original data. For each sample, estimated matrix $\mathbf{B}_k = [b_{ij}^k]$ where b_{ij}^k is the element at row i column j in $p \times p$ adjacency matrix estimated by LiNGAM at k^{th} sample.
2. For each coordinate i, j, we ordered N bootstrap estimates according to their values and denoted them by $b_{ij}^{(1)} < \cdots < b_{ij}^{(r(i,j,k))} < \cdots < b_{ij}^{(N)}$. Let $r(i,j,k)$ be the sequence order of b_{ij}^k and $b_{ij}^{(r(i,j,k))}$ be the value corresponding to b_{ij}^k.
3. We defined $\rho(k) = \max_{i,j}(r(i,j,k))$ to be the largest order of k^{th} sample.
4. We ordered $\rho(1), ..., \rho(N)$ according to their values then picked up the order at $100 \cdot \left(1 - \frac{\alpha}{2}\right)^{th}$ percentile, defined as $\rho_{1-\alpha/2}$.
5. We took the upper limits of the confidence interval of each b_{ij} to be $b_{ij}^{(\rho_{1-\alpha/2})}$.

By construction, at most $\alpha/2$ of the bootstrap estimates have a coordinate with value larger than the upper limit of the confidence interval [4].

The lower limit of the confidence interval is also constructed in the same way. Then the probability when all $p^2 - p$ elements fall in their own confidence interval is

$$P\left(\bigcap_{i,j,i \neq j} \{L_{ij} \leq b_{ij} \leq U_{ij}\}\right) \geq 1 - \alpha. \tag{5}$$

3. Simulations

In this study, we performed two experiments with simulated data. Both experiments consist of 5000 trials. In each trial, we generated datasets with dimension $p = 4$ and sample size $n = 1000$ then we constructed the confidence intervals of the estimated matrix $\mathbf{B}$ by using the naive approach, Bonferroni correction, and Mandel's approach.

The null hypothesis in this case is that each element b_{ij} in the adjacency matrix $\mathbf{B}$ is zero. Therefore, the FWER here is the probability when at least one of these null hypotheses is erroneously rejected.

In the first experiment, we constructed $p \times p$ adjacency matrix with all zeros which means that each variable is independent of the others. However, in the second experiment, we constructed $p \times p$ adjacency matrix with some non-zero elements. We replaced each non-zero element in the matrix by a value randomly chosen from the interval $[-1.5, -0.5] \cup [0.5, 1.5]$ and selected variances of the external influences e_i from the interval [1,3] as in [2]. Steps of the experiment are:

1. We set the adjacency matrix $\mathbf{B}$ as a zero matrix in the first experiment, and set the adjacency matrix $\mathbf{B}$ as a matrix with some non-zero elements in the second experiment.

2. We generated simulation data with sample size n by independently drawing the external influence variables e_i from various 18 non-Gaussian distributions used in [6] including super- and sub-Gaussian distributions and symmetric and asymmetric distributions. Then the values of the observed variables x_i were generated according to the LiNGAM model (2).

3. We iterated bootstrap sampling from the generated data for N replications (in this simulation, $N = 2000$) and then we calculated an estimated matrix $\mathbf{B}$ by using a direct method [2] for LiNGAM in each time of bootstrapping.

4. We constructed confidence intervals of every element in matrix $\mathbf{B}$ (*except* the elements in diagonal line) with $\alpha = 0.05$ by using multiple comparison approaches and then checked whether zero falls inside the confidence interval of each element in matrix $\mathbf{B}$ or not.

5. We repeated step 2 to 4 for 5000 times.

6. Let precision P be the ratio of the number of correctly estimated nonzeros to the total number of estimated nonzeros, recall R be the ratio of number of correctly estimated nonzeros to the total number of real nonzeros. We calculated P, R, FWER, and computational time of each approach.

The results are shown in Table 1. In the first experiment, there is no nonzero element in matrix $\mathbf{B}$, for this reason, all values of P and R in this experiment become N/A. From this table, we can see that the familywise error rates, or FWER, of the naive method in both experiments are greater than α. Meanwhile, with the use of multiple comparison approaches, those of the Bonferroni correction and Mandel's approach are quite similar to each other and, more importantly, become less than α.

However, in the step 2 of the Mandel's approach, a sequence order of each bootstrap estimate is calculated by using a linear search, which usually takes much longer computational time as compared to the other two approaches in both experiments. Therefore, we could say that the Bonferroni correction is the most appropriate approach in this experiment for the practical usage.

Simulations		Naive	Bonferroni	Mandel
1^{st} Experiment	P	N/A	N/A	N/A
	R	N/A	N/A	N/A
	FWER	0.065	0.004	0.003
	Time	3.832	3.516	2821.1
2^{nd} Experiment	P	0.964	0.996	0.997
	R	0.865	0.826	0.805
	FWER	0.205	0.030	0.030
	Time	3.716	3.434	2849.3

Table 1: Precision, Recall, Familywise Error Rate, and computational time (sec.) of each method in both experiments with 5000 times

4. Conclusions

From this study, we could say that both Bonferroni correction and Mandel's approach are able to control the familywise error rate of the confidence intervals to be less than the significance level in LiNGAM. Besides, the Bonferroni correction could be considered as the most appropriate approach in this experiment for the practical usage because of its short computational time.

References

[1] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.

[2] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, April 2011.

[3] Y. Hochberg and A.C. Tamhane. *Multiple comparison procedures.* Wiley, 1987.

[4] M. Mandel and R. A. Betensky. Simultaneous confidence intervals based on the percentile bootstrap approach. *Computational Statistics & Data Analysis*, 52:2158–2165, 2008.

[5] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[6] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.