Assessing Statistical Reliability of LiNGAM via Multiscale Bootstrap

3 authors, including:

Hidetoshi Shimodaira
Kyoto University
116 PUBLICATIONS **14,794** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   multiscale bootstrap View project

Assessing Statistical Reliability of LiNGAM via Multiscale Bootstrap

Yusuke Komatsu[1], Shohei Shimizu[1,2], and Hidetoshi Shimodaira[1]

[1] Dept. Math. and Comp. Sci., Tokyo Institute of Technology, Japan
[2] The Institute of Scientific and Industrial Research, Osaka University, Japan

Abstract. Structural equation models have been widely used to study causal relationships between continuous variables. Recently, a non-Gaussian method called LiNGAM was proposed to discover such causal models and has been extended in various directions. An important problem with LiNGAM is that the results are affected by the random sampling of the data as with any statistical method. Thus, some analysis of the confidence levels should be conducted. A common method to evaluate a confidence level is a bootstrap method. However, a confidence level computed by ordinary bootstrap is known to be biased as a probability-value (p-value) of hypothesis testing. In this paper, we propose a new procedure to apply an advanced bootstrap method called multiscale bootstrap to compute p-values of LiNGAM outputs. The multiscale bootstrap method gives unbiased p-values with asymptotic much higher accuracy. Experiments on artificial data demonstrate the utility of our approach.

Key words: bootstrapping, reliability, structural equation models, independent component analysis, non-Gaussianity

1   Introduction

Structural equation models [1] have been widely applied to analyze causal relationships in many fields. Many methods [2, 3] have been developed to discover such a causal model when no prior knowledge on the network structure is available. Recently, a non-Gaussian method called LiNGAM [4] was proposed. The new method estimates a causal ordering of variables using passive observational data alone. The estimated ordering is correct if the causal relations form a linear structural equation model with non-Gaussian external influence variables *and* the sample size is *infinitely large*. In practice, however, the sample size is finite. The finite sample size induces statistical errors in the estimation, and the estimated ordering may not be right even when the model assumptions are reasonable. Thus, some analysis of the statistical reliability or confidence level of the estimated ordering should be done.

A common procedure to evaluate such a confidence level is statistical hypothesis testing [5]. In statistical testing, one computes a probability-value (p-value) of a hypothesis. The hypothesis is rejected when the p-value is not greater than a pre-specified level of significance, say 5%. Bootstrapping [6] is a well-known

computational method for computing confidence levels when a simple mathematical formula is difficult to derive. It is a resampling method to approximate a random sample by a bootstrap sample that is created by random sampling with replacement from the original single dataset. However, it is known that the naive bootstrapping tends to give overconfidence in wrong hypotheses [7]. Thus, some advanced bootstrap methods to achieve higher accuracy have been proposed [7–10]. Among others, multiscale bootstrapping [9, 10] is much more accurate. In this paper, we propose to apply the multiscale bootstrap to compute confidence levels, *i.e.*, p-values, of variable orderings estimated by LiNGAM.

2 Background

2.1 LiNGAM

In [4], a non-Gaussian variant of structural equation models, which is called LiNGAM, was proposed. Assume that observed data are generated from a process represented graphically by a directed acyclic graph, *i.e.*, DAG. Let us represent this DAG by a $m{\times}m$ adjacency matrix $\mathbf{B}{=}\{b_{ij}\}$ where every b_{ij} represents the connection strength from a variable x_j to another x_i in the DAG, *i.e.*, the *direct* causal effect of x_j on x_i. Moreover, let us denote by $k(i)$ a causal order of variables x_i in the DAG so that no later variable influences any earlier variable. Further, assume that the relations between variables are linear. Then we have

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

where e_i is an external influence variable. All external influences e_i are continuous random variables having *non-Gaussian* distributions with zero means and non-zero variances, and e_i are independent of each other so that there is no unobserved confounding variables [3]. In [4], the LiNGAM model (1) was shown to be identifiable without using any prior knowledge on the network structure. In [4], a discovery algorithm based on independent component analysis (ICA) [11], which is called LiNGAM algorithm, was also proposed to estimate $k(i)$ and b_{ij}.

2.2 Multiscale Bootstrap

Denote by $\mathbf{x}$ a m-dimensional random variable vector and by $\mathbf{X}{=}(\mathbf{x}_1,\cdots,\mathbf{x}_n)$ a random sample of size n from the distribution of $\mathbf{x}$. Further, define a function $f(\mathbf{X})$ so that $f(\mathbf{X}){=}0$ if a hypothesis is rejected and otherwise $f(\mathbf{X}){=}1$. Suppose that we obtain a $m{\times}n$ data matrix $\overline{\mathbf{X}}$ that is generated from $\mathbf{x}$, and the function $f(\overline{\mathbf{X}}){=}1$. Then, it is useful to evaluate how statistically reliable the value of $f(\overline{\mathbf{X}}){=}1$ is since the function could return 0 for another data matrix due to sample fluctuation. In [12], Felesenstein proposed to use bootstrapping [6] to evaluate such reliability. Let us denote by $\mathbf{X}_q^*$ a q-th bootstrap sample of size n^*, which is created by random sampling with replacement from the columns of

$\mathbf{X}$. In ordinary bootstrap, n^* is taken to be n. Then, the bootstrap probability p^{BP} is defined as a frequency that $f(\mathbf{X}^*)=1$:

$$p^{BP} = \frac{1}{Q} \sum_{q=1}^{Q} f\left(\mathbf{X}_q^*\right), \tag{2}$$

where Q is the number of bootstrap replications. A testing procedure was proposed that the hypothesis is rejected if p^{BP} is not greater than a significance level α ($0<\alpha<1$), say 0.05. However, it is known that p^{BP} is *biased* as a p-value [7].

In [9, 10], a class of unbiased p-values was obtained. Let $\mathbf{y}^*$ denote the $\mathbf{y}$ vector corresponding to $\mathbf{X}^*=[\mathbf{x}_1^*, \cdots, \mathbf{x}_{n^*}^*]$. Then the standard deviation of $\mathbf{y}^*$ is proportional to $1/\sqrt{n^*}$, and its value relative to the case $n^*=n$ is called 'scale' of bootstrap resampling; this is defined by $\sigma=\sqrt{n/n^*}$. Then the bootstrap probability p^{BP} in (2) is a function of σ^2, which is denoted by $p_{\sigma^2}^{BP}$ for clarity. The fundamental idea of the multiscale bootstrap [9, 10] is to compute the bootstrap probability $p_{\sigma^2}^{BP}$ with the scale $\sigma^2=-1$, *i.e.*, the bootstrap sample size $n^*=-n$. Of course, it is impossible to set $n^*=-n$. Therefore, one first select several scales $\sigma>0$, computes the bootstrap probability for each of the corresponding bootstrap sample sizes n/σ^2 and extrapolates the bootstrap probabilities to $\sigma^2=-1$, *i.e.*, $n^*=-n$.

We now review a procedure to compute such unbiased p-values. Let us define a bootstrap z-value by

$$z_{\sigma^2} = -\Phi^{-1}\left(p_{\sigma^2}^{BP}\right), \tag{3}$$

where Φ^{-1} is the inverse of the distribution function Φ of the standard Gaussian distribution $N(0,1)$. Further, let us call σz_{σ^2} a normalized bootstrap z-value. Then, consider to model the changes in σz_{σ^2} along the changing the scale σ by a model $\psi(\sigma^2|\boldsymbol{\beta})$, where $\boldsymbol{\beta}=[\beta_0, \cdots, \beta_{h-1}]^T$ is a parameter vector of the model. Two model classes are proposed in [10]:

$$\psi_1^h(\sigma^2|\boldsymbol{\beta}) = \sum_{j=0}^{h-1} \beta_j \sigma^{2j}, \ h \geq 1. \tag{4}$$

$$\psi_2^h(\sigma^2|\boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{h-2} \frac{\beta_j \sigma^{2j}}{1 + \beta_{h-1}(\sigma - 1)}, \ 0 \leq \beta_{h-1} \leq 1, \ h \geq 3. \tag{5}$$

The model (4) is reasonable when the boundary $\partial\mathcal{H}$ is smooth, and the model (5) is preferable when $\mathcal{H}$ is a cone and $\partial\mathcal{H}$ is not smooth. To estimate the models, a number of sets of bootstrap replicates with different scales σ_d ($d=1, \cdots, D$) are first created, and subsequently the bootstrap probability $p_{\sigma_d^2}^{BP}$ for each scale is computed. Note that the bootstrap sample sizes may be different from that of the original dataset. Then, a set of scales and normalized bootstrap z-values $\{\sigma_d, \sigma_d z_{\sigma_d^2}\}$ is obtained. Note that $z_{\sigma_d^2}$ is computed based on $p_{\sigma_d^2}^{BP}$ using (3).

Finally, the model parameter vector $\boldsymbol{\beta}$ are estimated using the set of scales and normalized bootstrap z-values. The maximum likelihood method is applied since $Qp_{\sigma^2}^{BP}$ follows a binomial distribution. A best model $\psi_{best}(\sigma^2|\boldsymbol{\beta})$ is selected using an information criterion AIC [13].

Then, a class of p-values using the best model $\psi_{best}(\sigma^2|\boldsymbol{\beta})$ is derived:

$$
p_h^{MB} = \Phi\left\{ -\sum_{j=0}^{h-1} \frac{(-1-\sigma_0^2)^j}{j!} \frac{\partial^j \psi_{best}(\sigma^2|\hat{\boldsymbol{\beta}})}{\partial(\sigma^2)^j}\Big|_{\sigma_0^2} \right\}, \tag{6}
$$

where σ_0^2 is taken to be unity. The right side of (6) is the first h terms of the Taylor series of the slope of z_{σ^2} at $1/\sigma = 1$, i.e., $\partial z_\sigma^2/\partial(1/\sigma)|_1$, around σ_0^2. Further, p_1^{MB} turns out to be the naive bootstrap probability p^{BP} in (2). The larger h gives an unbiased p-value with asymptotic higher-order accuracy [10]. However, it also makes the maximum likelihood estimation less stable. In practice, h=2 or 3 is often used.

3   A multiscale bootstrap procedure to assessing reliability of LiNGAM

We first define null hypotheses tested. We here focus on the following four types of hypotheses between x_i and x_j $(i \neq j)$, although we can test hypotheses that describe the relations between more than two variables similarly:

1. H_{ij}^+: a hypothesis that x_i is directly caused by x_j, and its connection strength is positive, i.e., $b_{ij}>0$;
2. H_{ij}^-: a hypothesis that x_i is directly caused by x_j, and its connection strength is negative, i.e., $b_{ij}<0$;
3. H_{ji}^+: a hypothesis that x_j is directly caused by x_i, and its connection strength is positive, i.e., $b_{ji}>0$;
4. H_{ji}^-: a hypothesis that x_j is directly caused by x_i, and its connection strength is negative, i.e., $b_{ji}<0$.

We note that the signs of connections strengths are important and interesting in many applications [1] as well as the variable orderings. Further, this way of dividing the space based on the signs and orderings would make the boundaries of the regions be closer to be smooth than solely based on the orderings and help the multiscale bootstrap work better.

We now propose a new procedure to apply Multiscale Bootstrap to LiNGAM, which we call *MB-LiNGAM*:

MB-LiNGAM procedure

1. Select the scales $\sigma_1, \cdots, \sigma_D$ $(D \geq 2)$ so that $n_d^*=n/\sigma_d^2$ is an integer and choose the number of bootstrap replicates Q.
2. Generate Q bootstrap replicates $\mathbf{X}_{q,d}^*$ $(q=1, \cdots, Q)$ for each scale σ_d, i.e., each bootstrap sample size $n_d^*=n/\sigma_d^2$.

3. Perform LiNGAM algorithm to each bootstrap replicate $\mathbf{X}^*_{q,d}$ and compute the bootstrap probabilities $p_d^{BP}(H_{ij}^+)$ and $p_d^{BP}(H_{ij}^-)$ $(i \neq j)$ for each scale σ_d, where $p_d^{BP}(H)$ denotes the bootstrap probability of a hypothesis H for scale σ_d.
4. Compute the multiscale bootstrap p-values $p_h^{MB}(H_{ij}^+)$ and $p_h^{MB}(H_{ij}^-)$ $(i \neq j)$ using the procedure in Section 2.2, more specifically (6), where $p_h^{MB}(H)$ denotes the multiscale bootstrap p-value of H with the order h.

In the simulations below, the ICA part of LiNGAM algorithm is run several times in Step 3. Each time the initial point of the optimization is randomly changed. The set of the estimates that achieves the largest value of an ICA objective function is used in the subsequent steps. It is a common practice to alleviate the effects of possible local maxima.

4 Simulations

We created two LiNGAM models with $m=6$ variables:

$$
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 0 & 0 & 0 \\ b & b & 0 & 0 & 0 & 0 \\ 0 & b & 0 & b & 0 & 0 \\ b & b & b & 0 & b & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}, \tag{7}
$$

where $b=0$ or 0.5, and e_1 and e_2 also followed a Laplace distribution whose mean zero and variance two. We randomly generated 1280 datasets with sample size 1000 under each of the five models. Then we applied MB-LiNGAM procedure in Section 3 to the datasets. The scales σ_d were selected so that they gave integer values of bootstrap sample size and were (approximately) equally-spaced in log-scale between $1/9$ and 9 ($d=1, \cdots, 13$). The number of bootstrap replicates Q was 1000, and the value h for p_h^{MB} was 3.

In Fig. 1, we also show a scatterplot of empirical rejection probabilities by ordinary bootstrap $\mathrm{Prob}\{p^{BP}(H_{32}) < \alpha\}$ and those by multiscale bootstrap $\mathrm{Prob}\{p_3^{MB}(H_{32}) < \alpha\}$ versus significance levels α in the six variable cases. Plots for unbiased tests should be on the diagonal. That is, their rejection probabilities should be equal to the corresponding significance levels. Most of the plots for ordinary bootstrap are far above the diagonal, indicating that ordinary bootstrap gave rather biased p-values and tended to reject reasonable hypotheses much more often than the nominal frequencies or significance levels. In contrast, the plots for multiscale bootstrap are much closer to the diagonal. This showed that multiscale bootstrap provided much better unbiased p-values.

5 Conclusion

We proposed a new procedure to evaluate statistical reliability of LiNGAM. Our procedure gives p-values of variable orderings estimated by LiNGAM and tells which orderings are more reliable.
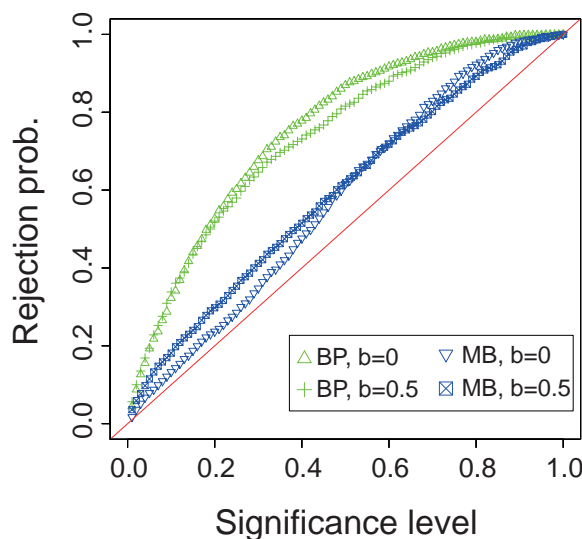
Fig. 1. Scatterplots of empirical rejection probabilities of H_{32}^+ by ordinary bootstrap and those by multiscale bootstrap versus significance levels in the six variable cases.

References

1. Bollen, K.A.: Structural Equations with Latent Variables. Wiley (1989)
2. Pearl, J.: Causality: Models, Reasoning, and Inference. Camb. Univ. Pres. (2000)
3. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. Springer Verlag (1993) (2nd ed. MIT Press 2000).
4. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A linear non-gaussian acyclic model for causal discovery. J. Mach. Lear. Res. **7** (2006) 2003–2030
5. Lehmann, E., Romano, J.: Testing Statistical Hypotheses. Springer (2008)
6. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)
7. Efron, B., Halloran, E., Holmes, S.: Bootstrap confidence levels for phylogenetic trees. In: Proc. Natl. Acad. Sci. USA. (1996) 13429–13434
8. Hall, P.: The bootstrap and Edgeworth expansion. Springer-Verlag (1992)
9. Shimodaira, H.: An approximately unbiased test of phylogenetic tree selection. Systematic Biology **51** (2002) 492–508
10. Shimodaira, H.: Testing regions with nonsmooth boundaries via multiscale bootstrap. J. Statistical Planning and Inference **138** (2008) 1227–1241
11. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. Wiley, New York (2001)
12. Felsenstein, J.: Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39** (1985) 783–791
13. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Control **19**(6) (1974) 716–723