Xu *Appl Inform (2017) 4:17*
https://doi.org/10.1186/s40535-017-0046-1

Applied Informatics

METHODOLOGY Open Access

Star-causality and factor analysis: old stories and new perspectives

Lei Xu[1,2*]

*Correspondence:
lxu@cs.sjtu.edu.cn;
lxu@cse.cuhk.edu.hk
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
Full list of author information is available at the end of the article

Abstract

Advances in causal discovery from data are becoming a widespread topic in machine learning these recent years. In this paper, studies on conditional independence-based causality are briefly reviewed along a line of observable two-variable, three-variable, star decomposable, and tree decomposable, as well as their relationship to factor analysis. Then, developments along this line are further addressed from three perspectives with a number of issues, especially on learning approximate star decomposable, and tree decomposable, as well as their generalisations to block star-causality analysis on factor analysis and block tree decomposable analysis on linear causal model.

Keywords: Causal discovery, Factor analysis, Tree decomposable, Block star-causality

Background

From the view of probability, "two events X_1 and X_2 have no relationship" is described as two random variables X_1 and X_2 that are independent in their corresponding joint distribution $p(x_1, x_2) = p(x_1)p(x_2)$. Away from this end, X_1 and X_2 must get some dependence, which can be one of the different types. In most of the existing big data mining efforts, what considered is correlation. Without correlation $E[X_1X_2] = E[X_1]E[X_2]$ or $\mathrm{Cov}[X_1X_2] = E[X_1X_2] - E[X_1]E[X_2] = 0$ means that there is no dependence of the second order between two events, but they may be still dependent in one of higher order types.

Particularly, the correlation $E[X_1X_2]$ is symmetric, while there also exists some relationship that is asymmetric and even more interesting. One example is the causality, i.e., the occurrence of the event X_1 causes occurrence of the event X_2, but not inversely. But an asymmetric relationship is not necessarily a causal relation. One example is a regression $E(x_2|x_1)$ that is widely considered in data analysing studies. However, a regression does not necessarily represent a causal relation.

In fact, whether X_1 and X_2 have a causal relation also depends on their environment W, which was first made precise by the common cause principle of Reichenbach (1956). This principle makes it possible to infer causal relation from statistical relation. Specifically, it follows from the non-correlation

$$E[X_1, X_2|W] = E[X_1|W]E[X_2|W] \tag{1}$$

Springer Open

or the conditional independence

$$p(x_1, x_2|w) = p(x_1|w)p(x_2|w) \tag{2}$$

that we can infer that there must exist one of the three causal relationships $X_1 \leftarrow W \rightarrow X_2$, $X_1 \rightarrow W \rightarrow X_2$, $X_1 \leftarrow W \leftarrow X_2$, though we can not identify specifically which one. We may identify Eq. (1) or even Eq. (2) from samples of variables x_1, x_2, w when they are binary variables. However, it becomes increasingly difficult when the variables take multiple values or even continuous values, for which a kernel-based approach has been proposed to deal with such a task in Fukumizu et al. (2008). Even worse, the environment typically consists of a set of features $W_1, \ldots, W_k$, which makes the task become even much more difficult. Alternatively, the Rubin Causal Model was first proposed in 1974 by Rubin and subsequently studied for many years (Rubin and Rubin 2011), which considers the so-called average causal effect (ACE) by computing $E[X_2|X_1, W]$ or its differences with X_1, W taking different values.

Pearl (1986) has shown that the following decomposable distribution

$$p(x_1, x_2, x_3, w) = p(w)p(x_1|w)p(x_2|w)p(x_3|w) \tag{3}$$

of dichotomous variables x_1, x_2, x_3, w can be identified by examining whether the observable three-variable distribution

$$p(x_1, x_2, x_3) = \sum_w p(x_1, x_2, x_3, w) \tag{4}$$

satisfies a necessary and sufficient condition on seven joint-occurrence probabilities of one, two, and three dichotomous variables, where these joint-occurrence probabilities are estimated from samples of x_1, x_2, x_3. Moreover, a necessary but not sufficient condition for $p(x_1, x_2, x_3)$ to be star-decomposable (as illustrated in Fig. 1a, b and to be further described in "Methods") is that all correlation coefficients $\rho_{ji},\ i,j \in \{1,2,3\}$ obey the following triangle inequalities:

$$\rho_{jk} \geq \ \rho_{ji}\rho_{ik}, \quad \text{with} \quad \rho_{jk}\rho_{ik}\rho_{ji} \geq 0, \quad i \neq j \neq k. \tag{5}$$

Furthermore, for a tree-decomposable distribution (as illustrated in Fig. 1c and to be further described in "Methods") of dichotomous variables, it is also shown in Pearl (1986) that the topology of this tree can be uncovered uniquely from the observed correlation coefficients between pairs of variables, based on the following TETRAD conditions (Spearman 1904; Anderson and Rubin 1956):

$$T_e^{(ijkl)} = \rho_{ij}\rho_{kl} - \rho_{il}\rho_{jk} = 0, \quad i \neq j \neq k \neq l. \tag{6}$$

Subsequently, Xu (1986) and Xu and Pearl (1987) further proceeded to study the distribution Eq. (3) of Gaussian variables x_1, x_2, x_3, w with three new results as follows:

1. The analysing tool used in Pearl (1986) stems from Eqs. (3) and (4) on dichotomous variables (i.e., Eq. 24 in Pearl 1986) that considers the products of conditional independence indirectly in a linear mixture, led to a set of constraint equations that are

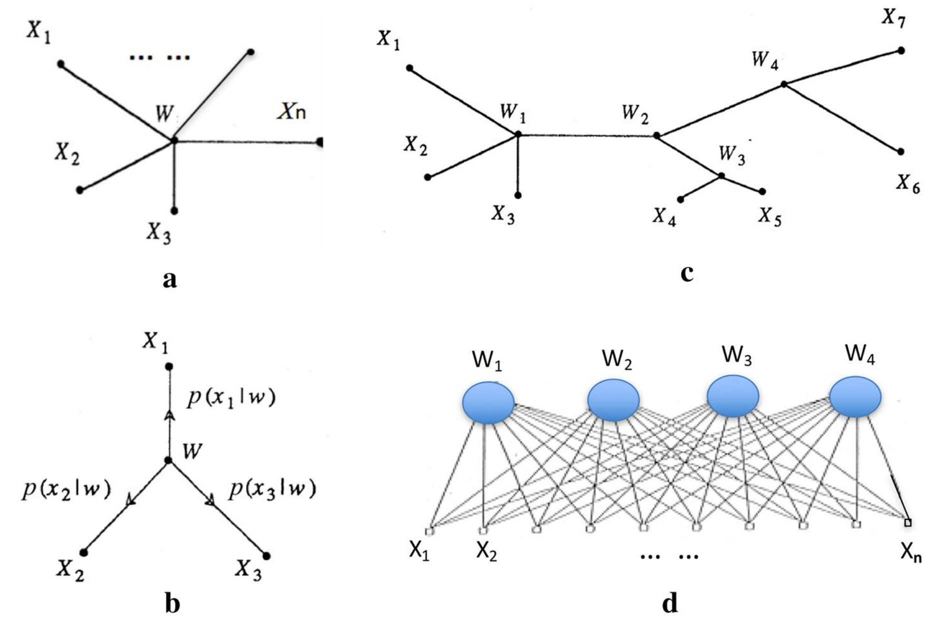


Fig. 1 Star-causality and factor analysis: old stories. **a** Star-decomposable. **b** Triplet star-decomposable. **c** Tree topology and tree-decomposable distribution. **d** Factor analysis model

solved to get a necessary and sufficient condition. Differently, a new tool is suggested in Xu (1986) and Xu and Pearl (1987), which stems from

$$p(x_1, x_2, x_3|w) = p(x_1|w)p(x_2|w)p(x_3|w) \tag{7}$$

that directly considers the product of conditional independence for inferring the star structure or topology of causality, and subsequently identifies the parameters of the involved distributions by

$$p(x_1, x_2, x_3) = \int p(x_1, x_2, x_3, w)\, \mathrm{d}w. \tag{8}$$

2. Instead of following Pearl (1986) that considers join probabilities to form constraint equations from Eq. (4), the equation by Eq. (7) is turned into one or a number of equations on different orders of statistics. Particularly, for Eq. (7) with Gaussian variables x_1, x_2, x_3, w, the block decomposition of covariance matrix (Gigi 1977) is adopted with equalities and inequalities on the second orders of statistics as constraints, which are further simplified into Eq. (5).
3. Specifically, the necessary and sufficient condition for $p(x_1, x_2, x_3)$ of Gaussian variables to be star-decomposable is simply that the triangle inequalities by Eq. (5), i.e., the star-causality by Eq. (3) and the latent structure by Eq. (4) can be recovered from merely the second order statistics, i.e., correlation coefficients $\rho_{ji},\ i,j \in \{x_1, x_2, x_3\}$.

When all the variables are Gaussians, the latent structure by

$$p(x_1, x_2, \ldots, x_n) = \int p(x_1, x_2, \ldots, x_n, w)\, \mathrm{d}w \tag{9}$$

with the star-causality by

$$p(x_1, x_2, \ldots, x_n) = \int p(x_1, x_2, \ldots, x_n, w)\, \mathrm{d}w \tag{10}$$

is actually equivalent to the classical factor analysis with only one factor. Pioneered by Spearman (1904), whether the factor analysis model (as illustrated in Fig. 1d and to be further described in the next section) is identifiable has been a classical topic for more than 100 years, from perspectives that are more or less similar to constraints on the second-order statistics obtained from Eq. (9). The well-known TETRAD equations or differences were discovered already in Spearman (1904) and have been used for constructing casual structures not just in Pearl (1986) but also by others (Spirtes and Glymour 2000; Bartholomew 1995; Bollen and Ting 2000). Moreover, Theorem 4.2 in Anderson and Rubin (1956) also gave a necessary and sufficient condition for identifying whether a covariance matrix can be the one of a factor analysis model with one factor and three observation variables, which is actually equivalent to Eq. (5) but expressed in a different format.

Methods

Following Pearl (1986), the following decomposition of a joint distribution

$$p(x_1, \ldots, x_n | w) = \prod_{i=1}^{n} p(x_i | w) \tag{11}$$

is called star-decomposable distribution, as illustrated in Fig. 1a, and particularly triplet star-decomposable in Fig. 1b. Also, w acts as a common cause that emits to affect the observable variables $x_1, \ldots, x_k$; we use star-causality to name such a simple but important casual structure. A typical tree-causality is in a tree structure, as illustrated in Fig. 1c. Moreover, we say that a distribution $p(x_1, \ldots, x_k)$ is tree-decomposable if it is the marginal of a distribution $p(x_1, \ldots, x_n; w_1, \ldots, w_m), m \le n-2$ that supports a tree-structured, such that $W_1, \ldots, W_m$ correspond to the internal nodes of a tree and $x_1, \ldots, x_n$ to its leaves.

We further push forward developments of discovering causality along the line of Xu (1986) and Xu and Pearl (1987) from three perspectives.

First, the causal tree constructing procedure proposed in Pearl (1986), and also adopted in Xu (1986) and Xu and Pearl (1987), may be improved by the following three considerations:

(a) In that procedure, constructing causal tree is made via joining triplets by checking the TETRAD equations by Eq. (6) while triplets were detected by the triangle inequalities by Eq. (5). However, Pearl (1986) pointed out that TETRAD equalities are unlikely to be satisfied forever in practice because we often have only sample estimates of the correlation coefficients. Though it was also tried in Pearl (1986) to decide the 4-tuple topology on the basis of the permutation of indices that minimises the difference $T_e^{(ijkl)}$, experiments found that the structure which evolves from such a method is very sensitive to inaccuracies in the estimates of the correlation coefficients. Here, we suggest to consider TETRAD equalities by minimising the difference $T_{\rm e}^{(ijkl)}$ subject to the constraints by Eq. (5).

(b) Not limited to consider triplet star-decomposable, star-causality in Fig. 1a may also consider in the same line of Xu (1986) and Xu and Pearl (1987), while the necessary and sufficient condition for star-decomposable is not just satisfying the triangle inequalities by Eq. (5) but also $0.5n(n-1)-n$ equalities, which is equivalent to Theorem 4.2 in Anderson and Rubin (1956) for the identifiability of a covariance matrix to be the one of the factor analysis models with one factor in general. In other words, the consideration above can be extended to a general case in a similar way.

(c) Moreover, we may also combine an edge removing procedure as used in the well-known PC algorithm (Spirtes and Glymour 1993, 2000) by which the link between two nodes is removed by testing the independence between them conditioning on the rest nodes. This checking also relates to inaccuracies in the estimates of correlation coefficients, for which we may consider to add in minimising $T_{\rm e}^{(ijkl)}$ subject to the constraints by Eq. (5). Second, in addition to the above improvements, we proceed to a new method. The existing procedure is featured by making testing based on the set of correlation coefficients between observable variables, while the new method first estimates another set of correlation coefficients between observable variables and latent variables, and then makes testing based on both the sets. Specifically, we propose the following two suggestions:

(d) Equations (11) and (12) in Xu and Pearl (1987) were derived from Eq. (11) and are rewritten below:

$$\begin{aligned} &T_{\rm e}^{(ijw)} = 0,\ i \neq j, \quad \text{and} \quad B_{\rm e}^{(iw)} > 0, \quad \forall i \\ &\text{where} \quad T_{\rm e}^{(ijw)} = \sigma_{ij} - \frac{\sigma_{iw}\sigma_{jw}}{\sigma_{ww}}, \quad \text{and} \quad B_{\rm e}^{(iw)} = \sigma_{ii} - \frac{\sigma_{iw}^2}{\sigma_{ww}}. \end{aligned} \tag{12}$$

Constructing star-causality can be made by learning $\sigma_{iw}, \forall i$ and $\sigma_{ww} > 0$ (or simply setting $\sigma_{ww} = 1$) by the following constrained optimisation

$$\max \sum_i [B_{\rm e}^{(iw)} - \rho R^{(iw)}], \text{ s.t. } T_{\rm e}^{(ijw)} = 0, \quad \forall i \neq j, \tag{13}$$

which may have different implementations, e.g., by the Lagrange method. Also, sparse learning is added via the term

$$R^{(iw)} = \sum_i |\sigma_{iw}|, \tag{14}$$

which prefers to push σ_{iw} towards zero in order to reduce a false or unreliable relation, where ρ is a coefficient that controls the strength. $R^{(iw)}$ has no action if we simply set $\rho = 0$ while a large action when $\rho > 0$ gets a large value. After learning, we test whether this star-causality is justified via testing $T_{\rm e}^{(ijw)} = 0,\ \forall i \neq j$ or with help of some sum $\sum T_{\rm e}^{(ijw)}$ as a statistics.

(e) Once a star-causality is made, the latent node can be treated in a way similar to observable nodes, such that a new star-causality can be constructed from a combination of observable nodes and learned latent nodes. Hence, constructing tree-decomposable causality can be made from star-causality in at least two manners. First, a tree-decomposable structure can be grown up from a star-causality by gradually learning and testing newly added observable nodes and latent nodes. Second, constructing a number of star-causality structures in parallel, and then combining them to form a tree-decomposable structure with help of some composition of the above learning and testing.

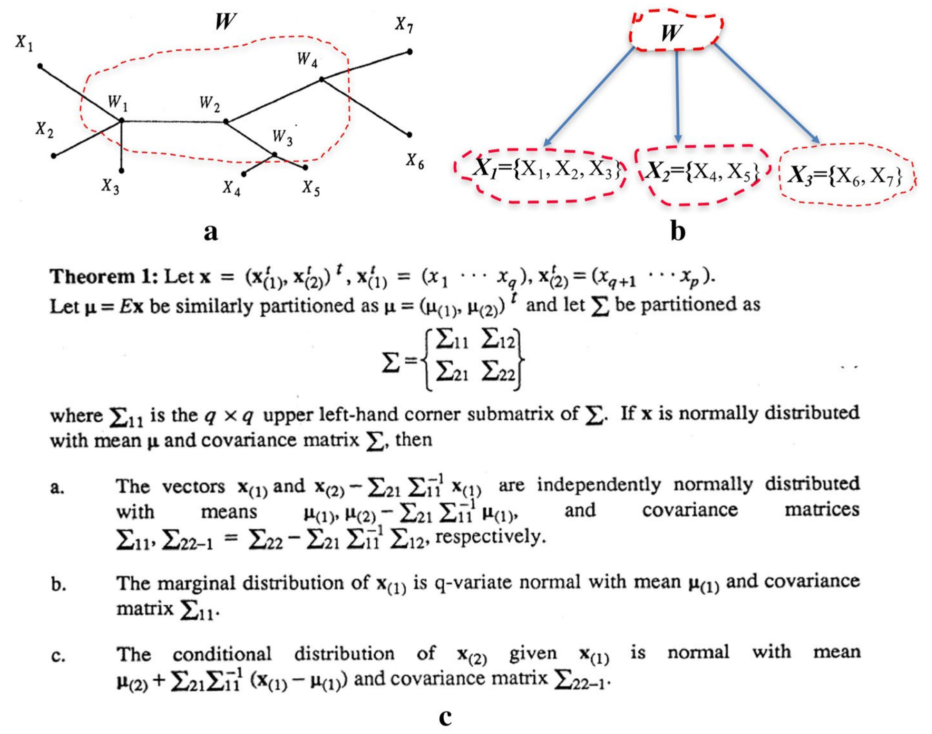
(f) The above studies may be further extended to consider non-Gaussian variables in a two-stage approach. At the first stage, the topology of the star-causality and even generally tree-decomposable causality can be obtained from the correlation coefficients. At the second stage, the conditional probabilities and the marginal probabilities of each latent node can be estimated from Eq. (9) in a way similar to that in Xu (1986) and Xu and Pearl (1987). Specifically, each link can be still a linear equation and the conditional distribution $p(x_i|w)$ or $p(x_i|w_k)$ is still Gaussian, while each inner node w or w_k can even come from a non-Gaussian distribution. Moreover, we may also obtain constraint equations of higher order statistics from Eq. (11). Third, beyond causality between variables, we further proceed to considering causality between sets or blocks of variables. Lumping latent factors $\{W_k\}$ into one vector factor $\mathbf{W}$, the factor analysis model in Fig. 1d may be turned into a block star-decomposable structure still in the format of Fig. 1a, with w in Eq. (10) simply replaced by $\mathbf{W} = [W_1, \ldots, W_k]^{\mathrm{T}}$. In the sequel, we address further details.

(g) In a way similar to that adopted in Xu (1986) and Xu and Pearl (1987), we may obtain a necessary and sufficient condition for such a star-decomposable based on Theorem 1 given in Fig. 2c. For the block star-decomposable problem in Fig. 2a, this is equivalent to that the solution $\Sigma_{XW}, \Sigma_{WW}, D$ of the following matrix equation is unique:

$$\Sigma_{XX} - \Sigma_{XW}\Sigma_{WW}^{-1}\Sigma_{XW}^{\mathrm{T}} = D, \tag{15}$$

where $\Sigma_{XX} = [\sigma_{x_ix_j}]$ is the covariance matrix of the vector $X = [x_1, \ldots, x_n]^{\mathrm{T}}$, $\Sigma_{XW} = [\sigma_{x_iw_k}]$ is the covariance matrix between the vector X and the lumped latent vector W, and $D = \mathrm{diag}[d_1, \ldots, d_n], d_j > 0, \forall j$ is a diagonal matrix. Getting such a unique solution is generally difficult, but possible when Σ_{XW}, Σ_{WW} have some particular structures. A typical example is $\Sigma_{WW} = \mathrm{diag}[\sigma_{w_1w_1}, \ldots \sigma_{w_mw_m}]$, which equivalently leads to getting the necessary and sufficient condition for identifying the factor analysis model $\mathbf{X} = A\mathbf{W} + \mu + \varepsilon$ with a diagonal covariance matrix of ε, e.g., Theorem 4.1 in Anderson and Rubin (1956). Additionally, from the same motivation as getting Eq. (12) we can get

$$\begin{aligned} T_{\mathrm{e}}^{(ijw)} &= \sigma_{x_ix_j} - \sum_k \frac{\sigma_{x_iw_k}\sigma_{x_jw_k}}{\sigma_{w_kw_k}}, \\ B_{\mathrm{e}}^{(iw)} &= \sigma_{x_ix_i} - \sum_k \frac{\sigma_{x_iw_k}^2}{\sigma_{w_kw_k}}. \end{aligned} \tag{16}$$



Theorem 1: Let $\mathbf{x} = (\mathbf{x}_{(1)}^t, \mathbf{x}_{(2)}^t)^t$, $\mathbf{x}_{(1)}^t = (x_1 \cdots x_q)$, $\mathbf{x}_{(2)}^t = (x_{q+1} \cdots x_p)$.
Let $\boldsymbol{\mu} = E\mathbf{x}$ be similarly partitioned as $\boldsymbol{\mu} = (\boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)})^t$ and let Σ be partitioned as

$$\Sigma = \begin{Bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{Bmatrix}$$

where Σ_{11} is the $q \times q$ upper left-hand corner submatrix of Σ. If $\mathbf{x}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then

a. The vectors $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{x}_{(1)}$ are independently normally distributed with means $\boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_{(1)}$, and covariance matrices $\Sigma_{11}, \Sigma_{22-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, respectively.

b. The marginal distribution of $\mathbf{x}_{(1)}$ is q-variate normal with mean $\boldsymbol{\mu}_{(1)}$ and covariance matrix Σ_{11}.

c. The conditional distribution of $\mathbf{x}_{(2)}$ given $\mathbf{x}_{(1)}$ is normal with mean $\boldsymbol{\mu}_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})$ and covariance matrix Σ_{22-1}.

Fig. 2 Star-causality and factor analysis: new perspectives. **a** Lumping factors into a vector $\mathbf{W} = [W_1, \ldots, W_4]$. **b** Triplet star-decomposable in blocks. **c** Block decomposition of covariance matrix [see pp 51–53 in Gigi (1977)].

Then, constructing block star-causality can be made by learning $\sigma_{x_i w_k}, \forall i, k$ and $\sigma_{w_i w_i} > 0, \forall i$ (or simply setting each $\sigma_{w_i w_i} = 1$) again by the constrained optimisation Eq. (13) with $R^{(iw)} = \sum_i |\sigma_{iw_k}|$ and the subsequent testing.

(h) Similar to what addressed in the above e and d, constructing tree-decomposable causality can be made from star-causality. Moreover, a tree decomposable structure Fig. 2a can be turned into not only a problem of star-decomposable causality in Fig. 1a by lumping latent factors $\{W_k\}$ into one vector factor but also a problem of triplet star-decomposable causality in blocks as illustrated in Fig. 2b. Then, we get Σ_{XW} in a block structure, which increases the chance that Eq. (15) becomes uniquely solved. Considering variables of W in some structure, we may also extend this line to study a linear causal structure (Zhang et al. 2017; Shimizu et al. 2011) with both latent variables and loading variables in structures. Also, discovering causality may be further made within each subsets of variables. In other words, we may discovery causality on multiple levels of a hierarchy in a top-down manner, or even trading off the top-down manner with the bottom-up manner addressed in the first two perspectives.

Concluding remarks

Considers minimising the TETRAD differences by $T_e^{(ijkl)}$ subject to the constraints by Eq. (5) may motivate a new road for learning a causal model from samples to approximate tree decomposable causality. Instead of making tests based on the set of correlation coefficients between observable variables, as typically made in the existing procedure, we first perform the optimisation by Eq. (13) to estimate another set of correlation coefficients between observable variables and latent variables, and then make tests based on both the sets. We may further proceed along this road to make block star-causality analysis on factor analysis and block tree decomposable analysis on linear causal model.

Author details
[1] Centre for Cognitive Machines and Computational Health (CMaCH), The School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, SEIEE Building 3, 800 Dongchuan Road, Minhang District, 200240 Shanghai, China. [2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China.

Acknowledgements
This work was supported by the Zhi-Yuan chair professorship start-up grant (WF220103010) from Shanghai Jiao Tong University.

Competing interests
The author declares that they have no competing interests.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 29 November 2017 Accepted: 5 December 2017
Published online: 12 December 2017

References

Anderson TW, Rubin H (1956) Statistical inference in factor analysis. Proc Third Berkeley Symp Math Stat Probab 5:111–150

Bartholomew DJ (1995) Spearman and the origin and development of factor analysis. Br J Math Stat Psychol 48(2):211–220

Bollen KA, Ting K-F (2000) A tetrad test for causal indicators. Psychol Methods 5(1):3

Fukumizu K, Gretton A, Sun X, Schölkopf B (2008) Kernel measures of conditional dependence. In: Advances in neural information processing systems. pp 489–496

Gigi NC (1977) Multivariate statistical inference. Academic Press, New York

Pearl J (1986) Fusion, propagation, and structuring in belief networks. Artif intell 29(3):241–288

Reichenbach H (1956) The direction of time. University of California Press, Berkeley

Rubin DB, Rubin JL (2011) Causal model. In: International encyclopedia of statistical science. pp 1263–1265. Springer, Berlin

Shimizu S, Inazumi T, Sogawa Y, Hyvärinen A, Kawahara Y, Washio T, Hoyer PO, Bollen K (2011) Directlingam: a direct method for learning a linear non-Gaussian structural equation model. J Mach Learn Res 12:1225–1248

Spearman C (1904) "General intelligence," objectively determined and measured. Am J Psychol 15(2):201–292

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search. MIT Press, Cambridge

Spirtes PG, Glymour C (1993) Causation, prediction and search. In: Lecture notes in statistics, vol 81. Springer, Berlin

Xu L (1986) Investigation on signal reconstruction, search technique, and pattern recognition. Ph.D. Dissertation, Tsinghua University

Xu L, Pearl J (1987) Structuring causal tree models with continuous variables. In: Proceedings of the 3rd annual conference on uncertainty in artificial intelligence. pp 170–179

Zhang K, Gong M, Ramsey J, Batmanghelich K, Spirtes P, Glymour C (2017) Causal discovery in the presence of measurement error: identifiability conditions. arXiv preprint arXiv:1706.03768