A Brief Discussion of Telling Causal from Confounded for Continuous Observational Data

Xiao (Ariel) Liang

LIACS, Leiden University
x.liang.2@umail.leidenuniv.nl

Abstract. Suppose observations of $(\mathbf{X}, Y)$ are continuous-valued and correlated. Two methods of judging whether $\mathbf{X}$ causes Y or they are jointly caused by unobserved sources are briefly discussed and compared.

Keywords: Causal · Confounded · Observational · MDL · PPCA

1 Introduction

Causal inferences from passively observed data have been an open problem in statistical science for decades. The major challenges arise from the inexhaustible possibilities of hidden confounders that might affect both sides of random variables from which we intend to infer causality. Obvious examples include ages causing grey hairs and raising death rates at the same time, and sun exposures confounding the use of sun blocks and chances of skin cancer. In practice, unfortunately, many factors are either unknown or immeasurable, and therefore the core assumption in causal inference, causal sufficiency [10], seldom holds. Statisticians have been approaching this problem with a number of novel methods, such as stratified covariate balancing [1] and latent factor modelling [9].

Meanwhile, drawing causal conclusions from observational data has increasingly attracted attention in the research field of machine learning. In addition to determining the direction of a causal relationship [3], distinguishing causality from confoundedness is also a topic that extensive works have been focusing on. From a more practical perspective, computer scientists strive to circumvent unrealistic assumptions and reduce computational effort.

In this article, I briefly discuss two models — the first inspired by Independent Component Analysis (ICA) and the second the Minimum Description Length (MDL) principle — that consider continuous random variables and compute scores for decision-making. In both settings, n simple random samples of $\mathbf{X} = (X_1, ..., X_m)$ and Y are observed, where m could be any positive integer, while Y is required to be a scalar. The main task for both models is to draw a conclusion based on the observations over $(\mathbf{X}, Y)$, whether $\mathbf{X}$ is the true cause of Y or they are jointly caused by a continuous-valued confounder, $\mathbf{Z} = (Z_1, ..., Z_k)$. Note that $\mathbf{Z}$ is not observed. For this task, the ICA-based model constructs an indicator for the strength of confounding, whereas the MDL-based model

explicitly scores both hypotheses. Then limitations and advantages of each model will be analyzed.

In Section 2 and Section 3, respectively, I outline the decision-making procedures with the two models. Notations and details that are harmless to the nature of modelling are particularly lined up to ease the comparison. The reason that the MDL-based approach is pragmatically more favorable is explained in Section 4. Finally, I conclude with Section 5.

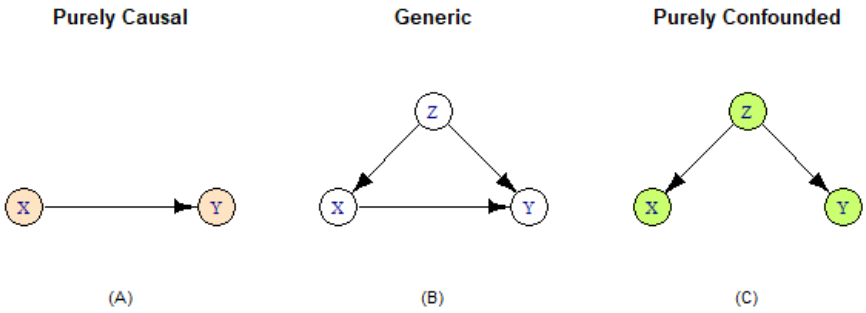

Fig. 1. Three scenarios of the relationships among $\mathbf{X}$, Y, and $\mathbf{Z}$.

2 The ICA-based Model

In 2018, Janzing and Scholkopf proposed to compute a score from the empirical covariance matrices to communicate the strength of confounding [6]. The higher the score, the more likely the presence of a hidden, common cause. Whether the score is large enough to signify the dominance of confoundedness is determined with a hypothesis test.

2.1 Basic Setup

Suppose one is given n samples from the joint distribution $P(\mathbf{X}, Y)$ over two statistically dependent, continuous variables, $\mathbf{X} = (X_1, ..., X_m)$, of arbitrary dimensions, and a scalar, Y. Because information about the confounder is unavailable, it is reasonable to assume independent, multiple confounding sources, which can be formalized as

$$\mathbf{Z} = (Z_1, ..., Z_k) \sim \mathrm{N}(0, \mathbf{I}_{k\times k}) \tag{1}$$

where $k \geq m$; Using structural equations [9], the generic scenario of $\mathbf{Z}$ related to $(\mathbf{X}, Y)$, shown in Fig.1(B), can be formulated as

$$\mathbf{X}_{m\times 1} = \mathbf{M}_{m\times k}\mathbf{Z}_{k\times 1} \tag{2}$$

$$Y = (\mathbf{a}_{m\times 1})^T \mathbf{X}_{m\times 1} + (\mathbf{c}_{k\times 1})^T \mathbf{Z}_{k\times 1} \tag{3}$$

where $\mathbf{M}, \mathbf{a}$, and $\mathbf{c}$ are all continuous-valued and control the linear relationships among $\mathbf{Z}$ and $(\mathbf{X}, Y)$. When $\mathbf{a} = 0$, $\mathbf{X}$ does not influence Y directly, and the correlation between them is completely generated by their common cause, $\mathbf{Z}$; in other words, it is represented by Fig.1(C). Similarly, Fig.1(A) represents $\mathbf{c} = 0$.

Since

$$\mathbf{\Sigma_{XX}} = \mathrm{Cov}(\mathbf{MZ}, \mathbf{MZ}) = \mathbf{MM^T}$$
$$\mathbf{\Sigma_{X}}_{Y} = \mathrm{Cov}(\mathbf{MZ}, (\mathbf{a^T M + c^T})\mathbf{Z}) = \mathbf{M(c + M^T a)}$$

the estimated regression coefficients, approximating the direct linear effect of $\mathbf{X}$ upon Y, can be further calculated by

$$\breve{\mathbf{a}} = \mathbf{\Sigma_{XX}}^{-1}\mathbf{\Sigma_{X}}_{Y} = \mathbf{a} + \mathbf{M^{-T}c}$$

where $\mathbf{M^{-T}}$ denotes the transpose of the pseudo-inverse of $\mathbf{M}$. It follows that the relative deviation of $\breve{\mathbf{a}}$ from $\mathbf{a}$ measures the confounding influence.

Definition 1. *The strength of confounding, β, is continuous and quantifies how likely the statistical dependence between $\mathbf{X}$ and Y is due to confoundedness.*

$$\beta = \frac{\|\breve{\mathbf{a}} - \mathbf{a}\|^2}{\|\mathbf{a}\|^2 + \|\breve{\mathbf{a}} - \mathbf{a}\|^2} \in [0, 1]$$

The strength of confounding, defined by Janzing and Scholkopf [5], equals zero if and only if $\mathbf{c} = 0$, i.e. when the purely causal scenario is true, and equals one if and only if $\mathbf{a} = 0$, i.e. when $\mathbf{X}$ and Y are purely confounded.

Naturally, one seeks to estimate β from $(\mathbf{X}, Y)$ and use $\hat{\beta}$ to determine by hypothesis test whether the strength of confounding is significant.

2.2 Asymptotic Properties

First, it can be proved by the Law of Large Numbers that when m is sufficiently large, and assuming standard Gaussian distribution for the coefficient vectors,

$$\beta \approx \frac{\tau(\Sigma_{\mathbf{XX}}^{-1})\sigma_c^2}{\tau(\Sigma_{\mathbf{XX}}^{-1})\sigma_c^2 + \sigma_a^2} = \frac{\tau(\Sigma_{\mathbf{XX}}^{-1})\theta}{\tau(\Sigma_{\mathbf{XX}}^{-1})\theta + 1} \tag{4}$$

where $\tau(\cdot) = \frac{1}{m} \cdot tr(\cdot)$ denotes the normalized trace of a matrix, and $\theta = \sigma_c^2/\sigma_a^2$ may be found by maximizing the log density function

$$\log[p_\theta(\frac{\breve{a}}{\|\breve{a}\|})] = \frac{\log[\det(\mathbf{R}_\theta)] - m \cdot \log[\langle \frac{\breve{a}}{\|\breve{a}\|}, \mathbf{R}_\theta{}^{-1} \frac{\breve{a}}{\|\breve{a}\|} \rangle]}{2} \tag{5}$$

$$\mathbf{R}_\theta = \mathbf{I} + \theta \cdot \mathbf{\Sigma_{XX}^{-1}} \tag{6}$$

in which $\langle\, , \rangle$ denotes the inner product of two vectors. In this way, one manages to approximate β from the empirical covariance matrices $\widehat{\mathbf{\Sigma_{XX}}}$ and $\widehat{\mathbf{\Sigma_{X}}_{Y}}$.

Secondly, under the same assumptions, the null hypothesis $\mathrm{H}_0 : \beta = 0$ is equivalent to $\mathcal{H}_0 : \mathcal{T} = 0$, under which the test statistic $\mathcal{T}$ obeys a mixed χ^2 - distribution approximately.

$$\begin{aligned} \mathcal{T}(\frac{\breve{a}}{\|\breve{a}\|}) &= \frac{1}{\sqrt{m}}[\langle \frac{\breve{a}}{\|\breve{a}\|}, \mathbf{\Sigma_{XX}^{-1}} \frac{\breve{a}}{\|\breve{a}\|} \rangle - \tau(\mathbf{\Sigma_{XX}^{-1}})] \\ &\approx \frac{1}{\sqrt{m}}[\sum_{j=1}^{j=m} a_j^2 s_j - \tau(\mathbf{\Sigma_{XX}^{-1}})] \end{aligned} \tag{7}$$

where $\breve{\mathbf{a}}/\|\breve{\mathbf{a}}\| = (a_1, ..., a_m)^T$ and $\{s_j\}_{j=1}^{j=m}$ denote the eigenvalues of $\mathbf{\Sigma_{XX}^{-1}}$.

2.3 Confounding Strength and Hypothesis Test

With n simple random samples over $(X_1, ..., X_m, Y)$, one can follow the procedure below and estimate the strength of confounding for them. The larger the value of $\hat{\beta}$, to greater extent one should trust the existence of hidden confounder.

Procedure 1:
Estimating the Strength of Confounding, β

Inputs: $(\mathbf{X}, Y)$
Output: $\hat{\beta}$

1. Compute the empirical covariance matrices: $\widehat{\mathbf{\Sigma_{XX}}}$ and $\widehat{\mathbf{\Sigma_{X}}_Y}$
2. Calculate $\widehat{\breve{\mathbf{a}}}$ by $\widehat{\breve{\mathbf{a}}} = \widehat{\mathbf{\Sigma_{XX}}}^{-1} \widehat{\mathbf{\Sigma_{X}}_Y}$
3. Find $\hat{\theta}$ that maximizes (5) combined with (6)
4. Estimate β by plugging $\widehat{\mathbf{\Sigma_{XX}}}$ and $\hat{\theta}$ into (4)

The absence of such a threshold that explicitly judges the significance of the deviation of β from 0 calls for hypothesis test. If significance is detected, one opts for the hypothesis that an unobserved confounder $\mathbf{Z}$ dominates the correlation between $\mathbf{X}$ and Y; otherwise the correlation is still presumed to imply causality.

Procedure 2:
Hypothesis Test $\mathcal{H}_0 : \mathcal{T} = 0$

Inputs: $\widehat{\mathbf{\Sigma_{XX}}}$, $\widehat{\breve{\mathbf{a}}}$, prescribed significance level α
Output: a decision regarding $\mathcal{H}_0$

1. Compute $\mathcal{T}^\star$ by plugging $\widehat{\mathbf{\Sigma_{XX}}}$, $\widehat{\breve{\mathbf{a}}}$, and m into (7)

2. Calculate the p-value, $p^\star = \mathrm{P}(\mathcal{T} \geq \mathcal{T}^\star)$

3. **If** $p^\star > \alpha$, reject the null hypothesis $\mathcal{H}_0$
   else do not reject $\mathcal{H}_0$

3 The MDL-based Model

Kaltenpoth and Vreeken developed the CoCa (short for Confounded-or-Causal) model in 2019, which scores each of the two hypotheses that are to be chosen from and advocates the one with the lower score [7]. The normalized difference between the two scores is interpreted as the confidence in this decision.

3.1 Kolmogorov Complexity and MDL

The Minimum Description Length (MDL) principle is a concrete realization of Kolmogorov complexity, turning the philosophy of finding truth by simplicity into a computable strategy. Applied to the main task in this article, it involves describing the observations with hypotheses. It is postulated that the description with the true cause, together with the true cause itself, can be encoded in shorter length than the combination with any other hypotheses.

Let $\mathcal{L}_{co}$ and $\mathcal{L}_{ca}$ be the encoding length, respectively, under the purely confounded hypothesis, $\mathcal{H}_{co}$, and the purely causal hypothesis, $\mathcal{H}_{ca}$. The postulate then narrows down to

$$\begin{cases} \mathcal{L}_{ca} < \mathcal{L}_{co}, \text{ if } \mathcal{H}_{ca} \text{ is true} \\ \mathcal{L}_{ca} > \mathcal{L}_{co}, \text{ if } \mathcal{H}_{co} \text{ is true} \end{cases}$$

Definition 2. *The one-part MDL, or refined MDL, calculates the combined length of the hypothesis and the description of data under that hypothesis.*

$$\mathcal{L}(\,\cdot\,|\mathcal{H}) = -\log \int_{H \in \mathcal{H}} P(\,\cdot\,|H) dP(H)$$

where $P(H)$ is a prior on $\mathcal{H}$, a class of arbitrary cases under the hypothesis.

There are two motivations for adopting the one-part MDL, rather than the practically more prevalent, two-part MDL. First, mathematical analysis is more convenient. In particular, it has been proved to be asymptotically mini-max optimal [2]. Secondly, the two-part MDL is unfeasible for Probabilistic Principal Component Analysis (Probabilistic PCA) by which hypotheses are parameterized in this model. As each H corresponds to a set of specific value(s) taken by normally distributed variable(s), using $L(\,\cdot\,, H) = L(H) + L(\,\cdot\,|H)$ will involve infinite, arbitrary choices of H.

3.2 Probabilistic Principal Component Analysis

The CoCa model also assumes (1) for the hidden confounder, except that $k \leq m$ is required. This is because the effect between variables is not only assumed to be linear but also modelled by Probabilistic PCA, one of the well studied techniques of latent factor modelling [9]. Similar to classical PCA, the linear relationship is formulated as a mapping to a subspace of lower dimensions with the least information loss, only assumptions of Gaussian distributions are added for the unobserved. Intuitively speaking, if $\mathbf{Z}$ causes $\mathbf{X}$, it is frequently the case that the knowledge contained in $\mathbf{X}_{m\times 1}$ may be represented by $\mathbf{Z}_{k\times 1}$ more succinctly; hence $k \leq m$.

Under $\mathcal{H}_{co}$, $\mathbf{Z}$ causes $\mathbf{X}$ and Y separately. Integrating (1) with

$$\mathbf{W}_{k\times(m+1)} = [W_1, ..., W_{m+1}];\ W_j \sim \mathrm{N}(0, \sigma_W^2 \mathbf{I}_{k\times k}) \tag{8}$$

$$\begin{bmatrix}\mathbf{X}_{m\times 1}\\ Y_{1\times 1}\end{bmatrix} = \begin{bmatrix}\mathbf{W}_{(X)}^T\\ \mathbf{W}_{(Y)}^T\end{bmatrix}\mathbf{Z} + \begin{bmatrix}\epsilon_1\\ \vdots\\ \epsilon_{m+1}\end{bmatrix} = \mathbf{W}^{\mathbf{T}}\mathbf{Z} + \epsilon;\ \epsilon \sim \mathrm{N}(0, \mathbf{\Sigma}_\epsilon)$$

the CoCa model induces the joint distribution of $(\mathbf{X}, Y)$ conditional on arbitrary $\mathbf{Z}$ and $\mathbf{W}$ as

$$\begin{bmatrix}\mathbf{X}_{m\times 1}\\ Y_{1\times 1}\end{bmatrix} |\mathbf{Z},\ \mathbf{W} \sim \mathrm{N}(\mathbf{W}^{\mathbf{T}}\mathbf{Z}, \begin{bmatrix}\sigma_x^2\mathbf{I}_{m\times m} & \\ & \sigma_y^2\end{bmatrix}) \tag{9}$$

and scores the purely confounded hypothesis by

$$\begin{aligned}\mathcal{L}_{co}(\mathbf{X}, Y) &= -\log \int p(\mathbf{X}, Y|\mathbf{Z}, \mathbf{W}) p(\mathbf{Z}) p(\mathbf{W})\, d\mathbf{W} d\mathbf{Z}\\ &\approx -\log[\frac{1}{N}\sum_{i=1}^{i=N} p(\mathbf{X}, Y|\hat{\mathbf{Z}}_i, \hat{\mathbf{W}}_i)]\end{aligned} \tag{10}$$

where the integral is approximated by a sum, and N samples of $(\hat{\mathbf{Z}}_i, \hat{\mathbf{W}}_i)$ are drawn from (1) and (8).

In order to line up the computation, Gaussian distribution is also assumed for $\mathbf{X}$ when it presumably causes Y. Then the assumptions include

$$\begin{aligned}\mathbf{X} = (X_1, ..., X_m);\ X_j &\sim \mathrm{N}(0, \sigma_x^2\mathbf{I})\\ \mathbf{w}_{m\times 1} &\sim \mathrm{N}(0, \sigma_w^2\mathbf{I}_{m\times m})\\ Y = \mathbf{w}^{\mathbf{T}}\mathbf{X} + \eta;\ \eta &\sim N(0, \sigma_\eta^2)\end{aligned} \tag{11}$$

$$Y|\mathbf{X}, \mathbf{w} \sim \mathrm{N}(\mathbf{w}^{\mathbf{T}}\mathbf{X}, \sigma_y^2) \tag{12}$$

Drawing N samples of $\hat{\mathbf{w}}_i$ from (11), the CoCa model scores $\mathcal{H}_{ca}$ by

$$\begin{aligned}\mathcal{L}_{ca}(\mathbf{X}, Y) &= -\log[\mathrm{P}(\mathbf{X}) \int \mathrm{P}(Y|\mathbf{X}, \mathbf{w})\mathrm{P}(\mathbf{w})\, d\mathbf{w}]\\ &\approx -\log[\frac{\mathrm{P}(\mathbf{X})}{N}\sum_{i=1}^{i=N}\mathrm{P}(Y|\mathbf{X}, \hat{\mathbf{w}}_\mathbf{i})]\end{aligned} \tag{13}$$

3.3 Scores and Confidence

Given n observations of $(\mathbf{X}, Y)$, one can make a choice between $\mathcal{H}_{co}$ and $\mathcal{H}_{ca}$ with explicitly quantified confidence:

$$\mathcal{C}_{\mathbf{X}Y} = \frac{\mathcal{L}_{co}(\mathbf{X}, Y) - \mathcal{L}_{ca}(\mathbf{X}, Y)}{\max\{\mathcal{L}_{co}(\mathbf{X}, Y), \mathcal{L}_{ca}(\mathbf{X}, Y)\}} \tag{14}$$

The greater the absolute value of $\mathcal{C}_{\mathbf{X}Y}$, the more confident one is about the decision. Simulations show that the confidence score strongly correlates with the accuracy of conclusions.

Procedure 3:
Choosing Hypothesis: Confounded versus Causal

Inputs: observations of $(\mathbf{X}, Y)$, N
Output: $\mathcal{H}_{co}$ or $\mathcal{H}_{ca}$, $\mathcal{C}_{\mathbf{X}Y}$

1. Draw N simple random samples of $(\hat{\mathbf{Z}}_i, \hat{\mathbf{W}}_i)$ from (1) and (8)
2. Compute $\mathcal{L}_{co}(\mathbf{X}, Y)$ by (9) and (10)
3. Draw N simple random samples of $\hat{\mathbf{w}}_i$ from (11)
4. Calculate $\mathcal{L}_{ca}(\mathbf{X}, Y)$ by (12) and (13)
5. **If** $\mathcal{L}_{co}(\mathbf{X}, Y) < \mathcal{L}_{ca}(\mathbf{X}, Y)$, choose $\mathcal{H}_{co}$
   else choose $\mathcal{H}_{ca}$
6. Quantify the confidence associated with the decision by (14)

---

4 Discussion

Both models presume linear forms of effect between variables and support the decision-making by quantifying confoundedness founded in the observations. However, they are applicable to different situations.

4.1 Limitations of the Confounding Strength

The ICA-based model requires the dimension of $\mathbf{X} = (X_1, ..., X_m)$ to be sufficiently large. Synthetic data suggest that the asymptotic properties in 2.2 start to behave when $m = 10$, and perform very well when $m = 20$ [6]. If $m < 10$ in the real-world problem, therefore, the CoCa model serves as a better tool since it allows m to be any positive integers. In addition, although the two models posit

$k \geq m$ and $k \leq m$ respectively, it is barely relevant to the choice of model unless one has prior knowledge about the number of hidden confounding source(s).

Another major drawback of the ICA approach is that both overfitting and confounding can pull up the value of β by driving $\check{\mathbf{a}}$ towards the eigenspace of $\mathbf{\Sigma_{XX}}$ associated with small eigenvalues [4]. To rule out the former and evaluate the latter by $\hat{\beta}$, sample sizes of the order 10 000 are already necessary for $m = 10$ according to simulation results [6]. Hence the confounding strength is only useful for massive observations.

Lastly, like the presumption of innocence in court, $\mathcal{H}_0$ has an advantage over $\mathcal{H}_1$ by default in the hypothesis test. On one hand, $\mathbf{X}$ is presumed to cause Y until compelling evidence is found against it; on the other hand, not rejecting $\mathcal{H}_0$ is far from denying the presence of $\mathbf{Z}$. A decision based on the hypothesis test is thereby biased in nature. Yet the CoCa model cleverly avoids this problem by assessing two hypotheses in equal position.

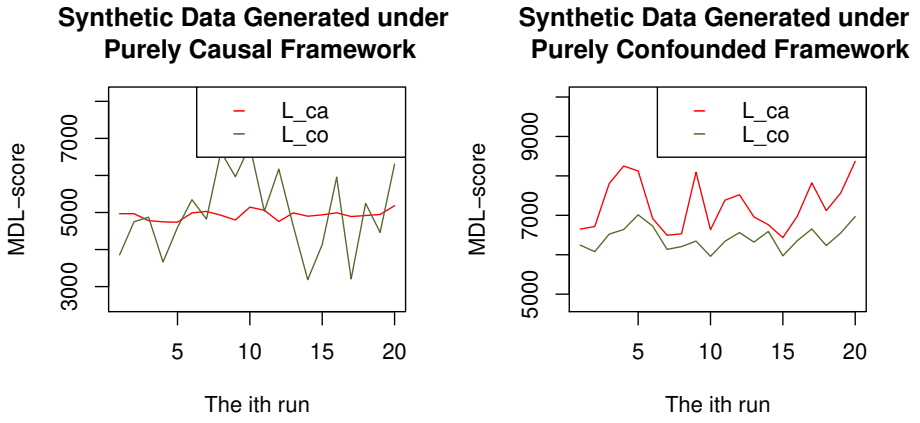

Fig. 2. When $m = k = 6$, and 500 samples of $\mathbf{X}$ are generated from the bimodal distribution, $Beta(0.5, 0.5)$, $\mathcal{L}_{co}$ fluctuates so drastically that in about half of the 20 simulations, the MDL approach would miss the causality between $\mathbf{X}$ and Y and leads to the wrong conclusion (left), although it performs well on $(\mathbf{X}, Y)$ that are purely confounded by a *Beta* source (right).[1]

4.2 Limitations of the CoCa Model

The MDL-based model is preferable in general due to fewer restrictions on data and greater power of decision-making. Nonetheless, two limitations of it are worth noticing.

[1] https://github.com/PawinData/CoCa/blob/master/Beta.R

When the Gaussian assumption on $\mathbf{X}$ is strongly violated in the observations, (12) becomes unsuitable. Simulations show that the CoCa model possesses a degree of robustness regarding $\mathbf{X}$ generated from the Laplace, the log-normal, and the uniform distribution [7]. Still, multimodality and extreme skewedness might invalidate the model. For instance, the MDL approach frequently fails to make correct decisions with simulated bimodal data as illustrated by Fig.2.

Moreover, the CoCa model focuses on the extremes of a spectrum and overlooks the scenarios that $\mathbf{X}$ and $\mathbf{Z}$ jointly cause Y, which the confounding strength takes into account. For every possibility represented by Fig.1(B), a binary classification is carried out by judging whether it is "closer" to Fig.1(A) or Fig.1(C). Albeit convenient, the information loss could be critical if the presence of confounders itself is of interest. For example, have temperatures, precipitation, and emission of gases caused the formation of atmospheric particulate matters?[2] While the MDL approach can merely infer confoundedness with a very low confidence score ($\mathcal{C}_{\mathbf{X}Y} \approx -0.0183$), the confounding strength, β, is sufficiently close to 1; intuitively speaking, the quantity of confoundedness is already significant, yet not considerably larger than the quantity of causality. Nevertheless, meteorologists need to explore other potential contributors.[3]

5 Conclusion

Given continuous-valued observations over statistically dependent $(\mathbf{X}, Y)$, two models attempt to tell whether their relationship is causal or confounded by quantifying the confoundedness found in the data. While the ICA-based approach is more suitable for high-dimensional $\mathbf{X}$ and conservative inference, the CoCa model excels at flexibility and making unbiased decisions. The MDL principle can be combined with more advanced techniques of latent factor modelling to deal with nonlinear cases.

[2] http://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data
[3] https://github.com/PawinData/CoCa/blob/master/AIR.Rmd

Appendix

In CoCa model, Kaltenpoth and Vreeken proposed a highly flexible and promising methodology of integrating the MDL principle with latent factor modelling. By explicitly formulating the distribution of the unobserved $\mathbf{Z}$, the likelihood of the observed conditional on the prior may be derived, and the MDL scores computed. However, the CoCa model is unable to handle the nonlinear relationships among $(\mathbf{X}, Y)$ and $\mathbf{Z}$.

One of ideas is to replace Probabilistic PCA in CoCa model with a nonlinear version of it — Gaussian Process Latent Variable Models (GP-LVM) [8]. Relationships among variables would be modelled by Gaussian Processes, which are mappings from a latent space to the observed-data space with the inner-product kernel substituted with a covariance function that allows for nonlinear properties.

For a semester-long research project, I plan to find a real-world dataset in which nonlinear relationships are more appropriate, construct a covariance function with corresponding nonlinear properties, and test the extension of $\mathcal{L}_{ca}(\mathbf{X}, Y)$ and $\mathcal{L}_{co}(\mathbf{X}, Y)$ on both the real and the simulated data.

References

1. Alemi, F., Elrafey A., Avramovic I.: Covariate Balancing through Naturally Occurring Strata. Health Services Research 53(1) (2016)
2. Grunwald, P. D.: The Minimum Description Length Principle. MIT Press. Boston (2007)
3. Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., Scholkopf, B.: Nonlinear causal discovery with additive noise models. NIPS, 689–696 (2009)
4. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. 1st edition, John Wiley Sons. New York (2001)
5. Janzing, D., Scholkopf, B.: Detecting confounding in multivariate linear models. Journal of Causal Inference 6(1), (2017) doi:10.1515/jci-2017-0013
6. Janzing, D., Scholkopf, B.: Detecting non-causal artifacts in multivariate linear regression models. JMLR, 2245–2253 (2018)
7. Kaltenpoth, D., Vreeken, J.: We Are Not Your Real Parents: Telling Causal from Confounded using MDL. In: Proceedings of the 2019 SIAM International Conference on Data Mining, pp. 199–207. SIAM, Calgary (2019)
8. Lawrence, N.: Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. JMLR, 6:1783–1816 (2005)
9. Loehlin, J. C.: Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis. Psychology Press, Hove UK (1998)
10. Reichenbach, H.: The Direction of Time. Dover (1956)