3. Insufficient

Soham Roy
Martikulation Number: 7028704
soro00002@stud.uni-saarland.de

1 The high level and commonalities

The primary goal of the paper by Wang et. al. Wang and Blei 2019 is to introduce and discuss the concept of the "deconfounder," an alternative method for addressing potential unobserved confounders in observational data, specifically it was shown in the context of movie casting and revenue among other datasets. The movie producer wants to estimate the causal effect of each actor on a movie's revenue, but traditional causal inference methods require strong assumptions and the need to identify, measure, and control for all confounders. Unobserved confounders (like a movie's genre, which could affect both the cast and the revenue) could bias causal estimates. The proposed deconfounder method is essentially a statistical model that attempts to capture the dependence among the actors (the potential causes). It's a latent variable model that, once fitted, can account for unobserved confounders, mitigating the potential bias they could introduce in causal estimates.

This method hinges on two assumptions: that the latent-variable model accurately represents the data which is being tested by a predictive score of how suitable the latent model descibes the causal structures, and that there are no unobserved single-cause confounders (which can't be tested but is a weaker assumption than in traditional methods).

Thus, the paper aims to show how this approach can provide more valid causal inferences under weaker assumptions, making multiple causal inference more feasible than classical causal inference in certain cases.

The main goal of the paper by Kalenpoth et al. Kaltenpoth and Vreeken 2019, is to propose a new method called "CoCa" (Confounded-or-Causal) for causal inference from observational data, specifically for identifying whether two variables X and Y are causally related or if they are likely jointly influenced by an unseen confounding variable Z. Here can be a joint distribution of many random variables (X1,X2...Xm) but Y is a scalar value. The authors aim to make this determination while acknowledging that the common assumption of causal sufficiency (the absence of hidden confounders) is often violated in practice. To overcome this challenge, the paper proposes the use of the Algorithmic Markov Condition (AMC) and the Minimum Description Length (MDL) principle. AMC postulates that the simplest factorization of the joint distribution aligns with the true causal model. The MDL principle, used to approximate the non-computable Kolmogorov complexity, provides a quantitative measure of complexity. The paper also suggests using latent factor modeling to estimate the distribution of the unseen confounder Z.

In comparison to the previous paper, both papers deal with the issue of unobserved confounders in causal inference from observational data. Both propose solutions using latent variable models to approximate these confounders and address how they could possibly relate to the outcome variable (s). However, the methods differ in their approaches. The previous paper proposes the use of a "deconfounder" that's based on the dependency structure of observed causes, while this paper proposes the CoCa method that is based on the Algorithmic Markov Condition and complexity check by Minimum Description Length principle.

Preprint. Under review.

2 Goals

The paper by Wang et al.Wang and Blei 2019 measures the outcome of a regressive variable (a scalar value) like the revenue of a movie as an example shown in the paper while considering that there exists a confounding effect and shows an outcome where the mean square error of the prediction of the outcome of the revenue of a movie is less when their approach assumes a confounding effect. The author compares the result with approaches where it was modeled without any deconfounder and also with the presence of covariates in various combinations. However it was not clear whether the ultimate motive of the author was to get the minimul MSE or just show the effect of the deconfounder is better than one without one. The author here wants to make counterfactual statements of the form "The movie X would have made Y in revenues had A been in its cast" by tapping in the explainability power of the deconfounder. Also the author does not comment anything on the direction of causation as it is already assumed in the problem statement, Their methodology involves disentangling the dependence of the distribution on its assigned causes, rendering these causes independent. and then in successive steps using this deconfounder (z cap) and the converted independent causes to predict the outcome variable using a regressive function while the paper by Kalenpoth et al.Kaltenpoth and Vreeken 2019 is more of a classification problem classifying whether it is the causal or confounding relationship between random variables X and Y and more so if it is causal then it infers the direction of causation between X and Y. Both papers are focused on improving the quality of inferences drawn from data. in paper by Wang et al.Wang and Blei 2019 this manifests as improving prediction accuracy.in the other one, it manifests as correctly classifying relationships as causal or confounding, and correctly identifying the direction of causation. Wang et al.Wang and Blei 2019 does not discuss the direction of causation (as this is not its aim), while Kalenpoth et al.Kaltenpoth and Vreeken 2019 directly addresses this issue.

3 Similarity

Both authors incorporate latent factor models to encapsulate the effect of confounders in their research. Wang et al.Wang and Blei 2019 adopt several approaches, including linear and quadratic models, to align their model family optimally with the cause distribution. They aim for their model family to reflect as much interdependence among the assigned causes as possible, and utilize a latent variable to treat these causes as independent.

After thr fitting phase of the deconfounder , Wang et al. Wang and Blei 2019uses the variable zi which is the expected value of the deconfounder given a configuration of causes and the independent assigned causes, which emerge from this fitting process, are subsequently used to predict the outcome of the target variable. , resulting in a two-tiered approach.

Wang et al. Wang and Blei 2019also implement a significance test between observation and holdout data to assess the predictive power of the latent factor model in depicting the causes. A predictive of 0.1 is regarded as significant which is debatable as this value has been used in these experiments as a one shoe fits all approach. Keeping this aside, They test various models using a trial-and-error strategy to gauge the suitability of the model family. This approach is heavily dependent on finding an appropriate model. Without one, the downstream task of fitting the outcome model using it as a function parameter becomes futile and may even lead to biased results. Given the complexity and interdependence of many assigned variables, devising such a model may prove challenging.

Conversely, Kalenpoth et al.Kaltenpoth and Vreeken 2019 employ standard Probabilistic Principal Component Analysis (PPCA) to model this latent factor variable. Here, the authors use the Minimum Description Length (MDL) principle to approximate the Kolomorov complexity (KLC). If KL(Y|X) + KL(X) or KL(X|Y) + KL(Y) is greater than KL(X|PAx) + KL(Z), where PAx is a subset of Z, we can infer that Z has confounded the relationship. Unlike Wang et al.'sWang and Blei 2019 method, this approach, by comparison between approximations, can operate with any model as a latent factor model. Even if the model fit isn't perfect, it will uniformly impact the causal, anticausal, and confounded directions, eventually simplifying to a comparison of scalars. This significantly decreases the reliance on the model choice.

Despite the potential capability of the deconfounder to handle causes that are causally dependent, the authors advocate for its application only to causes that are not causally dependent, suggesting to omit the dependent ones. This implies that the method might exhibit less robustness or reliability

when causes are causally interlinked. However, it remains unclear how or whether causes are causally dependent on each other, which is precisely the issue that Kalepoth et al.Kaltenpoth and Vreeken 2019 address. Without such clarity, one would need a deep understanding of the causes and the ability to select a subset of causes that are not causally dependent. Kalenpoth et al.Kaltenpoth and Vreeken 2019 it has been stated that the graph satisfoes the causal markov condition but the degree of assocoation between the X s is not known , it will be interesting to know the impact on the confidence scores as a result on the degree of the associativity of the X,s. It has been demonstrated that multi-cause confounders cannot exist, as their presence would violate Pearl's d-separation principle once the Latent Factor Model (LFM) renders all causes independent. In other words, the existence of multi-cause confounders would imply a dependency among the causes. Paper Two also aligns with this concept, inherently applying it through the use of Probabilistic Principal Component Analysis (PPCA). It's argued that the deconfounder somewhat encapsulates multi-cause confounders if they influence more than one cause. However, this approach falls short when addressing single cause confounders (SCC), especially if there are numerous SCC that are indistinguishable in real-life situations. Kalenpoth et al.Kaltenpoth and Vreeken 2019 models a scenario for an SCC where the dimension of X is one, and Z has a direct influence on X. This model works well, but it would be intriguing to explore the effects of an SCC when the dimension of X exceeds one and the SCC impacts one of the variables. In contrast, Wang et al. Wang and Blei 2019does not comment on the dimensions of X and Z, other than noting that a dimension of one disrupts the entire system. Therefore, there must be multiple causes to capture the relationship using the latent variable. All experiments conducted consider X's dimension to be larger than that of Z.

4 Results

Both authors report encouraging results for their respective objectives. Wang et al.Wang and Blei 2019 demonstrate that the application of a deconfounder indeed yiwlds positive effects on the outcome, effectively reducing the mean square error (MSE) of observational data. However, their initial claim of predicting counterfactual statements remains vague, and there's no concrete way to validate it.

In an experiment conducted by Kalenpoth et al.Kaltenpoth and Vreeken 2019, using optical data, the confidence score ranged between -0.15 and 0.15. Although the classification results were mostly correct—successfully predicting the signs on most instances—it's intriguing to consider how other latent models might affect the confidence score. It's reasonable to speculate that more complex latent models might enhance the confidence, as they could better represent relationships and result in a lower Kolmogorov complexity score. It's also possible that model complexity only impacts confounded data, without significantly affecting causal data. Further exploration in this regard would have been beneficial.

Kalenpoth et al.Kaltenpoth and Vreeken 2019 achieve commendable results on synthetic and simulated datasets. Notably, when the dimension of X is smaller than the dimension of Z, the relationship is ambiguous due to the low confidence scored produced by the classifier. Consequently, it's crucial to be aware of the dimensionality of X before fitting a Latent Factor Model (LFM) to ensure optimal capturing of the relationship. While the authors provide empirical support for this notion, additional research or clarification is required as it's a pivotal aspect of their methodology.

On the other hand, Wang et al.Wang and Blei 2019 do not address such considerations in their experiments, where the dimension of X typically exceeds that of Z. They calculate the outcome variable as functions $f(a, z)$ and $f(a, \hat{a}(z))$—where 'a' stands for assigned causes, 'z' for the estimated deconfounder, and $\hat{a}(z)$ for the restructured assigned causes. An important part of their analysis could have been to examine the relationship between the reconstructed and the original causes, potentially using a distance metric or an equivalent measure.

Moreover, interpreting the scores for $f(a, \hat{a}(z))$ is crucial. This function evaluates the outcome by excluding the original assigned causes and using only the reconstructed causes and the deconfounder. It could offer a richer understanding of the outcome-determining process than $f(a, \hat{a}(z))$. The authors cauld have also explored $f(z, a, \hat{a}(z))$, utilizing all three components to potentially decrease the MSE further, hence enhancing predictability. However, the main aim—whether to demonstrate the efficacy of a deconfounder over a non-deconfounded model, or to minimize MSE to its lowest possible value—remains debatable.

A good approach for Wang et al. Wang and Blei 2019would be to find the confounding effect (confidence/strength) using the approach by Kalenpoth et al.Kaltenpoth and Vreeken 2019 and then use their original methodology to infer the outcome instead of blindly applying the deconfounder in every situation, otherwise a possibility where in reality , their is not an effect of a confounder but they are just creating something (using the LFM) that is just delegating the responsibility of the original causes to the deconfounder, and the outcome with this will only be as good as the outcome without the deconfounder but can never exceed !

Across all of Kalenpoth et al.'s Kaltenpoth and Vreeken 2019experiments, a strong correlation emerged between confidence and accuracy, suggesting the credibility of the model when high confidence scores are achieved. However, the inclusion of negative testing, such as creating Z, X, and Y independently to ensure pairwise independence, might have solidified these findings. This would have verified the model's ability to detect the absence of any confounded or causal relationship among variables. Without this examination, the model's performance in such circumstances remains unverified.

5 Conclusion

In conclusion, both the papers by Wang et al. Wang and Blei 2019 and Kalenpoth et al. Kaltenpoth and Vreeken 2019 address the challenge of unobserved confounders in causal inference from observational data, proposing novel methods using latent variable models to approximate these confounders. The papers diverge in their specific methodologies, with the "deconfounder" approach proposed by Wang et al. focusing on dependency structure of observed causes, and the "CoCa" method by Kalenpoth et al. utilizing the Algorithmic Markov Condition and the Minimum Description Length principle.

The goal of the deconfounder method is to predict the outcome variable considering confounding effects, notably reducing the mean square error of observatianal data and enabling counterfactual statements. However, its ultimate objective and handling of single cause confounders remain unclear. On the other hand, the CoCa method successfully classifies relationships as causal or confounding and identifies the direction of causation, but its performance in various dimensions and complexities requires further investigation.

Although both methods achieve encouraging results, each has its limitations, most of which stem from the assumptions they make and their reliance on model choice and other variables (multivariates) does not impede on their process and complicating the relationships. Furthermore, they both fail to address certain aspects of their methodologies, such as interpreting the strength of confounding effect, examining the relationship between reconstructed and original causes, and validating their models in scenarios of independence.

Despite these shortcomings, these studies represent significant advancements in the field of causol inference . It would be intriguing to see how these methods could be combined or improved upon to further enhance the quality of inferences drawn from data, making it possible to better understand and predict the effects of causal factors in a wide range of contexts and more on real-life datasets rather than simulated ones. Future work could also delve deeper into the impact of model choice, the implications of varing dimensions and complexities, and the best practices for validating these novel method.

References

Kaltenpoth, David and Jilles Vreeken (2019). "We are not your real parents: Telling causal from confounded using mdl". In: *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, pp. 199–207.

Wang, Yixin and David M Blei (2019). "The blessings of multiple causes". In: *Journal of the American Statistical Association* 114.528, pp. 1574–1596.