Causal Factor Analysis

Anonymous Authors[1]

Abstract

1. Introduction

The motivation would be that factor analysis (FA) is relatively familiar (compared to deep generative models) to many researchers who collect data/apply existing algorithms, but FA is not causal (despite often implicitly being treated as such, both in its development and commonly in its applications today) and not as powerful/generalized/applicable as generative deep models. This paper would be geared towards addressing these two shortcomings to provide a useful and theoretically-justified tool to applied scientists.

2. Related Work

Learning Latent Superstructures (Li et al., 2019): They focus on a quite different task, and their method doesn't seem mathematically so helpful to us either. They learn two "levels" of latents: one generates the data, and the other generates the first level. Together, they're a tree-structured Bayesian network, with the leaves being the first level and the rest being part of the superstructure level. The superstructure partitions the leaves into disjoint sets (called facets), and each facet corresponds to a certain way of clustering/partitioning the samples. Their learned structures (especially with the forced tree structure) don't have a clear causal interpretation (and they typically violate our faithfulness conditions), and our learned structure doesn't have an obvious (to me) multidimensional clustering interpretation.

GANs with CI Graphs (Ding et al., 2021): This paper seems quite interesting/helpful for us. They prove that some divergences/metrics (e.g., Wasserstein distance) have the right subadditivity properties so that GAN learning via discriminator networks can be efficiently applied for learning BNs and MRFs—instead of one large discriminitor network over the entire graph, they show you can use multiple simpler ones over local neighborhoods (Markov blankets in BNs and max cliques in MRFs). If we go with this approach (which so far seems like the most promising/straightforward to me), we would basically extend their result from BNs and MRFs to the UDGs that are used in my structure learning method. The local neighborhoods for UDGs should be cliques of a minimal edge clique cover (which is different from the BN and MRF case, but should be upper bounded my the MRF case). So, we'd prove that some metrics also have the right subadditivity properties in the case of UDGs and then we'd use local neighborhood-based discriminators in the usual way with a GAN to learn our generative model, leaving us to figure out some (I think) straightforward details of how we relate the marginal distribution over the observed variables to the marginal over the latent/the joint over the latent and marginal.

CausalGAN (Kocaoglu et al., 2017): This paper could be helpful for our current task depending on our specific implementation/approach. (And it could be helpful for future work in the interventional setting or for different structure learning algorithms.) It focuses on conditional sampling using a known causal graph (at least up to a partial causal order), which may be of use to us if we take the approach of the prev paper and learn our joint distribution as a product of conditional distributions based on local neighborhoods, but the approach of the other paper seems more straightforward and good enough.

CausalVAE (Yang et al., 2021): Similarly to the previous paper, this paper could be helpful for our current task depending on our specific implementation/approach, though they make use of labels representing "additional information", so it's not actually unsupervized learning like in our case. They focus on simultaneously learning a DAG and a generative causal model. Something similar could be interesting for us if we want a continuous optimization version of my currently constraint-based method, though we'd have to work out details of our MeDIL causal model setting as opposed to their more basic causal DAG setting. Their approach also seems to allow easy incorporation of priors, which could be interesting for us.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

3. MeDIL causal models

An intuition for my structure learning algorithm (Measurement dependence inducing latent causal models (Markham & Grosse-Wentrup, 2020).): we're trying to learn the latent sources of our observed data. Imagining the full latent causal DAG, our observed variables are some subset of nodes O such that no descendent in D = de(O) is contained in O, i.e., the intersection of O and D is empty; and the latent variables we learn are the source nodes in this full DAG (so they have no incoming edges); and the structure we learn is which source nodes are ancestors of which observed nodes. This is simpler than the (generally) intractable problem of learning the full latent DAG but still helps answer the question: "what are the primary causes of my observed data?"

3.1. Factorizing MeDIL causal models

Using the parental Markov property, the joint latent and measurement density of a minMCM factorizes according to:

$$f(M, L) = \prod_{i=0}^{m} f(M_i \mid L_{\mathrm{pa}(M_i)}) \prod_{i=0}^{l} f(L_i).$$

For the encoder, the above factorization implies that

$$f(L \mid M) = \prod_{c \in C} f(L_{\mathrm{pa}(M_c)} \mid M_c),$$

where C is the set of connected components of the associated undirected graph (UDG) over M.

For the decoder, we have that

$$f(M \mid L) = \prod_{i=0}^{m} f(M_i \mid L_{\mathrm{pa}(M_i)}).$$

Also note that the minimality assumption of minMCMs implies the factorization

$$f(L) = \prod_{i=0}^{l} L_i.$$

4. Applications/Evaluation

Task: structure recovery w/ latents (metric: SHD); **data**: simulated

Task: missing data imputation (metric: RMSE); **data**: from https://jmlr.org/papers/volume22/20-589/20-589.pdf, also UCI repo

Task: generating new samples (metric: eyeball test); **data**: depends on scale of approach (worst-case, we can generate some simple pixel-level data, e.g. CFL or Helmholtz machine papers)

References

Ding, M., Daskalakis, C., and Feizi, S. Gans with conditional independence graphs: On subadditivity of probability divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 3709–3717. PMLR, 2021. arXiv:2003.00652 [cs.LG].

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv:1709.02023 [cs.LG]*, 2017.

Li, X., Chen, Z., Poon, L. K. M., and Zhang, N. L. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJgNwi09Km. arXiv:1803.05206 [cs.LG].

Markham, A. and Grosse-Wentrup, M. Measurement dependence inducing latent causal models. In Peters, J. and Sontag, D. (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 590–599. PMLR, 03–06 Aug 2020. URL http://proceedings.mlr.press/v124/markham20a.html. arXiv:1910.08778 [stat.ML].

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021. arXiv:2004.08697 [cs.LG].