Neuro-Causal Factor Analysis

Alex Markham [1] **Mingyu Liu** [2] **Bryon Aragam** [2] **Liam Solus** [1]

Abstract

We revisit nonlinear factor analysis from a comparatively new perspective given by advancements in causal discovery and deep learning, introducing a framework for *Neuro-Causal Factor Analysis (NCFA)*. Our approach is fully nonparametric: It identifies factors via latent causal discovery methods and then uses a variational autoencoder (VAE) that is constrained to abide by the Markov factorization of the distribution with respect to the learned graph. We evaluate NCFA on real and synthetic data sets, finding that it performs comparably to standard VAEs on data reconstruction tasks but with the advantages of sparser architecture, lower model complexity, and causal interpretability. Unlike traditional factor analysis methods, our NCFA method allows learning and reasoning about the latent factors underlying observed data from a justifiably causal perspective, even when the relations between factors and measurements are highly nonlinear.

1. Introduction

Since its development over a century ago, factor analysis (FA) (Spearman, 1904)[1]has been applied in many scientific fields, including genomics, computational biology (Pournara & Wernisch, 2007; Velten et al., 2022), economics (Forni & Reichlin, 1998; Ludvigson & Ng, 2007), sociology (Bollen, 2012) and many others. The goal of FA is to offer explanations of variability among correlated (more generally, dependent) observables via (potentially) fewer latent variables that capture the degree to which the observables in the system vary jointly. For the sake of identifiability, it is common to assume linearity, although in practice it is well-known that many problems exhibit complex nonlinear latent structures. With the rise of nonparametric deep generative models that allow representing highly nonlinear relationships between dependent observables, one might hope to combine the best of both worlds.

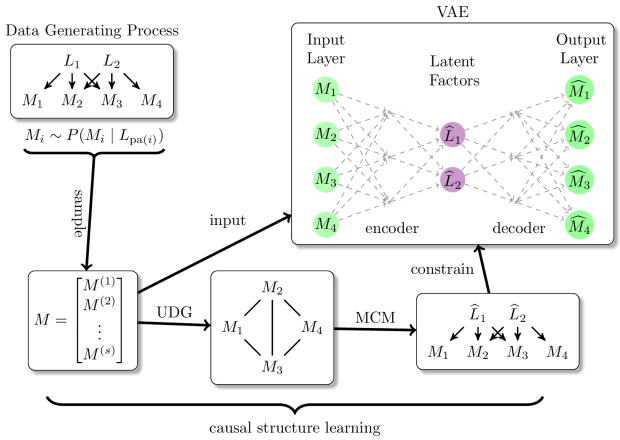


Figure 1: Pipeline for learning a neuro-causal factor model. Given a sample from a suitable data generating process, NCFA estimates a causal structure, which it then uses to constrain a VAE that it trains on the sample, resulting in a causally-interpretable deep generative model.

Moreover, within applications such as those listed above, FA is considered useful because the learned factors (latents) offer a possible interpretation of relevant observed correlations. Many applied FA studies provide an interpretation of the learned factors based on the observed variables whose joint correlation they encode. A natural tendency when trying to interpret these factors is to assume they reflect possible common causes linking observed variables. However, the models used in such studies are not necessarily built with causality in mind. Collectively, these considerations purport a need for a framework for nonlinear *causal* factor analysis that combines identifiability with flexibility that use modern advances in deep generative models and causality.

To this end, we propose *Neuro-Causal Factor Analysis (NCFA)*, augmenting classic FA on both fronts by leveraging advancements of the last few decades, including (i)

[1]Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden [2]University of Chicago, Chicago, USA. Correspondence to: Alex Markham <alex.markham@causal.dev>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative Modeling. *The 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

[1]We would like to briefly draw the reader's attention to and repudiate the historical context within which factor analysis and related methods were originally developed (e.g. Saini, 2019; Crenshaw et al., 1995; Stubblefield, 2007).

Algorithm 1 Neuro-Causal Factor Analysis (NCFA)

input sample S of measurement variables M,
significance level α,
latent degrees of freedom λ;

1: Estimate undirected dependence graph $\mathcal{U}$, via pairwise marginal independence tests with threshold α;

2: Identify minimum edge clique cover $\mathcal{C}$ of $\mathcal{U}$ and construct corresponding minimum MCM graph $\mathcal{G}$;

3: Assign the remaining $\lambda - |\mathcal{C}|$ latents to the cliques in $\mathcal{C}$ to produce the NCFA-graph $\widetilde{\mathcal{G}}$;

4: Estimate functions $f_{[n]}$ using a VAE constrained by $\widetilde{\mathcal{G}}$, with residual measurement errors ϵ;

output neuro-causal factor model $\langle \widetilde{\mathcal{G}}, f_{[n]}, \epsilon \rangle$, with NCFA graph $\widetilde{\mathcal{G}}$, loading functions $f_{[n]}$, and residual measurement errors ϵ;

causal discovery (Spirtes et al., 2000; Pearl, 2009) and (ii) deep generative models, such as variational autoencoders (VAEs) (Kingma & Welling, 2014). Our main contribution is the NCFA framework (Figure 1) for modeling such situations so as to yield causally interpretable, identifiable FA models with the flexibility and data replication capabilities afforded by deep generative models. In addition to elaborating the general theoretical framework by which NCFA uses *Reichenbach's common cause principle* (Reichenbach, 1956, p. 157) to structure the latent space and architecture of a VAE, we provide an algorithm and open source implementation for inference and prediction with NCFA models. We demonstrate on both synthetic and real data that NCFA performs comparably to baseline deep generative models, while also offering the possibility of justifiably causal interpretations of learned factors.

2. NCFA Models

Consider a collection of jointly distributed measurement variables $(M_1, \ldots, M_n)$ for which we assume that all dependences are explained by the existence of a latent common cause of the measured variables; i.e., that no M_i and M_j share a direct casual relation. If we were able to observe these latent confounders and condition upon them $M_1, \ldots, M_n$ would become mutually independent. Hence, the only causal structure encoded via conditional independence in the observed distribution is contained in their marginal independence structure, which can be encoded in an undirected graph:

Definition 2.1. The *unconditional dependence graph* (UDG) for the jointly distributed random variables $(M_1, \ldots, M_n)$ is the undirected graph $\mathcal{U}$ with node set $[n] = \{1, \ldots, n\}$ and edge set

$$E = \{i - j : M_i \not\perp\!\!\!\perp M_j\}.$$

To recover a causal interpretation of the relations that hold among the measurement variables we extend a UDG graph to a *(minimum) MCM graph*. Following the principle of Occam's Razor, we would like to explain the observed dependences in $(M_1, \ldots, M_n)$ in the simplest possible way, i.e., using the fewest possible latents to serve as the common causes of the measurement variables that exhibit dependence. To do so, we identify a *minimum edge clique cover* of the UDG $\mathcal{U}$, which is a collection $\mathcal{C} = \{C_1, \ldots, C_K\}$ of *cliques* (i.e., complete subgraphs of $\mathcal{U}$) such that for every $i - j \in E$ the pair i, j is contained in at least one clique in $\mathcal{C}$ and there exists no set of cliques with this property that has cardinality smaller than $|\mathcal{C}|$.

Definition 2.2. Let $\mathcal{U}$ be an undirected graph with minimum edge clique cover $\mathcal{C} = \{C_1, \ldots, C_K\}$. The *(minimum) MCM graph* $\mathcal{G}$ for $\mathcal{U}$ and $\mathcal{C}$ is the DAG with vertices $[n] \cup L$ where $L = \{\ell_1, \ldots, \ell_K\}$ and edge set

$$E = \{\ell_i \to j : j \in C_i \, \forall i = 1, \ldots, K\}.$$

We call $|L|$ the number of *causal degrees of freedom*.

An example of a UDG and a corresponding MCM graph is presented in Figure 3 in the appendix. Minimum MCM graphs were originally defined in the context of *MeDIL causal models* (Markham & Grosse-Wentrup, 2020). A summary of this theory is given in Appendix B.

Since we assumed all marginal dependencies in $(M_1, \ldots, M_n)$ are explainable by the existence of a latent common cause, then the observed distribution $(M_1, \ldots, M_n)$ is realizable as the marginal distribution of $(M_1, \ldots, M_n)$ in the joint distribution $(M_1, \ldots, M_n, L_1, \ldots, L_K)$ that is Markov to the DAG $\mathcal{G}$, where L_i is the random variable represented by the node ℓ_i in $\mathcal{G}$. From a factor analysis perspective, the latents $L_1, \ldots, L_K$ are the factors to be inferred.

2.1. NCFA graphs and Variational Autoencoders

The construction of a MCM graph from a UDG should also feel familiar from the perspective of VAEs. A classical VAE architecture assumes no marginal independencies among the observed variables and hence begins with a UDG that is the complete graph. The associated minimum MCM graph would then have a single latent variable as a parent common to all observed variables, however, the standard practice is to add multiple independent latent variables as parents of all observed variables in the system in order to afford us flexibility in how we model the latent space of the distribution. To mimic this practice while preserving the causal structure inherent in a minimum MCM graph $\mathcal{G} = \langle [n] \cup L, E \rangle$,

we replace each ℓ_i with a set of independent latent nodes $\mathcal{L}_i = \{\ell_{i,1}, \dots, \ell_{i,k_i}\}$, for some $k_i \geq 1$, each with the same connectivity (i.e. children) as ℓ_i. The number of latents to be assigned to each clique in this fashion is a parameter called the *latent degrees of freedom*: $\lambda = \sum_{C_i \in \mathcal{C}} k_i$. We call the resulting graph a *NCFA-graph* of $\mathcal{G}$ with λ latent degrees of freedom.

Definition 2.3. Let $\mathcal{G}$ be a minimum MCM graph for the UDG $\mathcal{U} = \langle [n], E \rangle$ and the minimum edge clique cover $\mathcal{C} = \{C_1, \dots, C_K\}$ of $\mathcal{U}$. A *NCFA graph* of $\mathcal{G}$ with λ latent degrees of freedom is a graph $\widetilde{\mathcal{G}}$ with node set $[n] \cup \widetilde{\mathcal{L}}$ and edge set $\widetilde{E}$ where

$$\widetilde{\mathcal{L}} = \mathcal{L}_1 \cup \dots \cup \mathcal{L}_K$$

for $\mathcal{L}_i = \{\ell_{i,1}, \dots, \ell_{i,k_i}\}$, $k_i \geq 1 \ \ \forall i \in [K]$, and

$$\widetilde{E} = \{\ell_{i,j} \to j : \forall j \in C_i, \forall i \in [K]\}.$$

Each node $\ell_{i,j}$ represents a latent variable $Z_{i,j}$. Since the latent nodes in $\mathcal{L}_i$ all have the same connectivity as the single latent ℓ_i their joint distribution $f(L_i) = \prod_{j=1}^{k_i} f(Z_{i,j})$ represents the common cause of the nodes in C_i, which was previously only represented by ℓ_i in $\mathcal{G}$. The factors to be inferred from a factor analysis perspective are now the random vectors $L_1, \dots, L_K$ with $L_i = (Z_{i,1}, \dots, Z_{i,k_i})$, which still have the causal interpretation afforded by the minimum MCM graph. However, the multiple latents provide us flexibility to model the effects of the causal factors.

Definition 2.4. A *NCFA model* is a joint distribution $(M_1, \dots, M_n)$ for which there is a NCFA-graph $\widetilde{\mathcal{G}} = \langle [n] \cup \widetilde{\mathcal{L}}, \widetilde{E} \rangle$ and functions $f_1, \dots, f_n$ for which $M_i := f_i(Z_{\mathrm{pa}_{\widetilde{G}}(i)})$ for all $i \in [n]$.

When modeling a distribution via a NCFA model, the functions f_i are treated as unknowns to be inferred via a deep generative model such as a VAE. The encoder maps the observations into the latent space as the joint posterior distribution $f(Z|M_1, \dots, M_n)$ where Z denotes the joint distribution of the $Z_{i,j}$, and the decoder maps from the latent space into the observational space according to the factorization

$$f(M_1, \dots, M_n|Z) = \prod_{i=1}^{n} f(M_i|Z_{\mathrm{pa}_{\widetilde{G}}(i)}).$$

The joint distribution of the latent space is $f(Z) = \prod_{i=1}^{K} f(L_i)$; i.e., it is a product of the (joint) distributions we have specified to represent each of the latent causes in the minimum MCM model $\mathcal{G}$ for $\mathcal{U}$. Following training of the VAE, the model may be used to generate predictions on observables.

2.2. Identifiability of minimum MCM graphs and ECC-model Equivalence

While there exist equivalence classes of minimum MCM graphs containing multiple elements, there also exist classes that are singletons; in other words, there exist undirected graphs (UDGs) with a unique minimum edge clique cover. For such UDGs, the minimum MCM graph is identifiable.

Theorem 2.5. *Suppose that the data-generating distribution is Markov to a minimum MCM graph $\mathcal{G}$ such that:*

1. *The UDG $\mathcal{U}$ for $\mathcal{G}$ admits a unique minimum edge clique cover, and*
2. $M_i \perp\!\!\!\perp M_j \iff i - j \notin E^{\mathcal{U}}$.

Then $\mathcal{G}$ is identifiable.

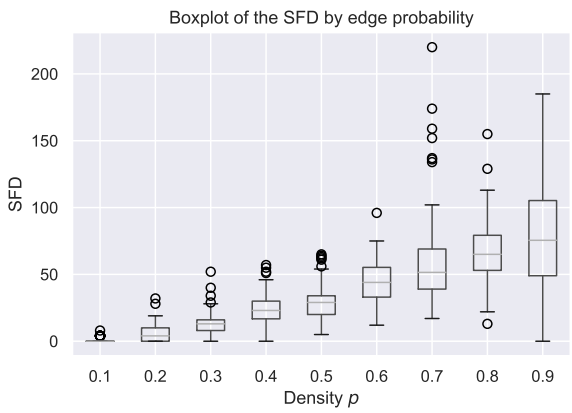
The identifiability result in Theorem 2.5 applies to models that are of practical interest. For instance, it follows from the results in (Deligeorgaki et al., 2022) that whenever a UDG admits a minimum edge clique cover such that each clique contains a pure measurement variable (e.g. as in Donoho & Stodden, 2003; Arora et al., 2012; Bing et al., 2020; Moran et al., 2023), the minimum edge clique cover is unique. Hence, the MCM graph is identifiable.

3. Neuro-Causal Factor Analysis

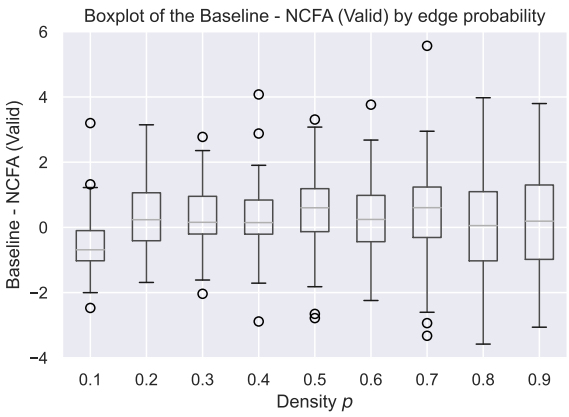
We now present our main contribution, the Neuro-Causal Factor Analysis (NCFA) algorithm, given in Algorithm 1. The NCFA algorithm runs by the logic described in Section 2: namely, it infers a UDG from data, identifies a minimum edge clique cover $\mathcal{C} = \{C_1, \dots, C_K\}$ for $\mathcal{U}$, builds the corresponding NCFA-graph $\mathcal{G}$ with λ latent degrees of freedom and then trains a VAE according to the functional relationships specified by $\widetilde{\mathcal{G}}$.

To estimate the UDG, marginal independence tests are performed. Starting with the complete graph, the edge $i - j$ is removed whenever M_i and M_j are deemed independent, i.e. according to a test with statistics such as distance-covariance (Székely et al., 2007; Markham et al., 2022) or Chatterjee's coefficient (Chatterjee, 2021; Lin & Han, 2022). A minimum edge clique cover is then identified for the estimated UDG $\widehat{\mathcal{U}}$. In general, this is an NP-hard problem, however there are both exact algorithms that work well for small graphs and heuristic algorithms that scale to large graphs (Gramm et al., 2009; Conte et al., 2020; Ullah, 2022).

Once a minimum edge clique cover is identified, the corresponding NCFA graph with λ latent degrees of freedom is constructed. Here, we ensure that at every clique in the minimum edge clique cover of $\widehat{\mathcal{U}}$ is assigned at least one latent variable. The remaining $\lambda - K$ latents are then distributed uniformly over the cliques. In this implementation



(a)



(b)

Figure 2: Results of NCFA on synthetic data sets from randomly generated graphs.

	samp size	n	α	λ	$\lvert L \rvert$	Training-Δ	Validation-Δ
MNIST	42000	784	0.05	153664	1	-0.00475	-0.04814
TCGA	632	1000	0.05	250000	8129	0.11488	0.11865

Table 1: Results of NCFA on two real data sets with default parameter values.

of NCFA, we set default $\lambda = \lfloor n^2/4 \rfloor$, a known upper bound on the number of cliques in a minimum edge clique cover of a graph on n nodes (Erdős et al., 1966). Finally, a VAE for the functional relations specified by the NCFA-graph is trained. One could, in principle, alternatively use any deep generative model. See Appendix D for further details.

Since NCFA constructs its model via the MCM graph $\widehat{\mathcal{U}}$, the estimated factors (i.e., joint distributions) $f(L_i)$ in the factorization of the latent distribution represent the distributions for the primary causes of the measurement variables to which the latent nodes in $\mathcal{L}_i$ are connected. This yields a factor analysis model in which the latent factors can justifiably be causally interpreted. Furthermore, while each latent variable $\ell_{i,j}$ is assigned a Gaussian prior in the VAE, by assigning $\mathcal{L}_i = \{\ell_{i,1}, \ldots, \ell_{i,k_i}\}$ latents to each clique C_i, instead of a single latent ℓ_i, each causal latent in the minimum MCM graph is modeled as a mixture distribution which can be arbitrarily non-Gaussian. Hence, the estimated factors have both a causal interpretation while additionally being as nonlinear as necessary.

4. Applications on Synthetic and Real Data

We now present some results applying NCFA to synthetic and real data sets, observing that the performance of NCFA is competitive with classical VAEs while additionally offering a nonlinear, causally interpretable factor model. We summarize some key results here, while deferring details to the appendices.

We first tested NCFA on synthetic data, compared to both a ground truth causal model and a baseline VAE, shown in Figure 2. Results are grouped according to edge density of the generating UDG, shown along the x-axis. Figure 2a contains box plots of distance between the true MCM causal structure and that learned by NCFA (lower is better). Here, distance between MCM graphs is measured using the *Structural Frobenius Difference* (SFD), which is a modification of the more common Structural Hamming Distance (SHD) for graphs with possibly different numbers of nodes (see Appendix E for more details on SFD and its relation to SHD). Figure 2b contains box plots of Validation-Δ, the difference between the final validation loss of the baseline VAE and that of NCFA (higher is better).

We also tested NCFA on two real datasets, MNIST and TCGA, comparing its performance to a baseline VAE. In both cases, there is no ground truth causal graph, so we focus on VAE metrics as a benchmark. We report the results in Table 1, with additional details in the appendix. For MNIST, sample size is much larger than number of measurement variables n, but this is not true of TCGA. When run using default settings for α, λ, we see that NCFA achieves comparable training and validation to the baseline VAE, demonstrating that it learns reasonable constraints (i.e. causal relations) as well as its ability to scale well to high-dimensional settings. In fact, for TCGA the training and validation losses are lower for NCFA, suggesting that incor-

porating the causal structure learned by NCFA improved model performance. Curiously, for MNIST, the minimum MCM graph consisted of just a single latent, suggesting the causal structure in this dataset is limited, which matches expectations. This does not mean that there are not multiple, interpretable latents to be discovered as is well-documented in the literature, but perhaps that these latents do not have a strong causal interpretation.

References

Adachi, K. Factor analysis: Latent variable, matrix decomposition, and constrained uniqueness formulations. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11 (3):e1458, 2019.

Arora, S., Ge, R., and Moitra, A. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pp. 1–10. IEEE, 2012.

Bing, X., Bunea, F., Ning, Y., and Wegkamp, M. Adaptive estimation in structured factor models with applications to overlapping clustering. *Annals of Statistics*, 48(4), 2020.

Bollen, K. A. Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38:37–72, 2012.

Chatterjee, S. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

Chatterjee, S. A survey of some recent developments in measures of association. *arXiv preprint arXiv:2211.04702*, 2022.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Conte, A., Grossi, R., and Marino, A. Large-scale clique cover of real-world networks. *Information and Computation*, 270:104464, Feb 2020. ISSN 0890-5401. doi: 10.1016/j.ic.2019.104464. URL http://dx.doi.org/10.1016/j.ic.2019.104464.

Crenshaw, K., Gotanda, N., Peller, G., and Thomas, K. *Critical race theory: The Key Writings that formed the Movement*. The New Press, 1995.

Deligeorgaki, D., Markham, A., Misra, P., and Solus, L. Combinatorial and algebraic perspectives on the marginal independence structure of bayesian networks. *arXiv preprint arXiv:2210.00822*, 2022.

Donoho, D. and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 16, 2003.

Erdős, P., Goodman, A. W., and Pósa, L. The representation of a graph by set intersections. *Canadian Journal of Mathematics*, 18:106–112, 1966.

Forni, M. and Reichlin, L. Let's get real: a factor analytical approach to disaggregated business cycle dynamics. *The Review of Economic Studies*, 65(3):453–473, 1998.

Gilbert, E. N. Random Graphs. *The Annals of Mathematical Statistics*, 30(4):1141 – 1144, 1959. doi: 10.1214/aoms/1177706098. URL https://doi.org/10.1214/aoms/1177706098.

Gramm, J., Guo, J., Hüffner, F., and Niedermeier, R. Data reduction and exact algorithms for clique cover. *ACM Journal of Experimental Algorithmics*, 13, Feb 2009. ISSN 1084-6654. doi: 10.1145/1412228.1412236. URL http://dx.doi.org/10.1145/1412228.1412236.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5): 411–430, 2000.

Jolliffe, I. T. *Principal component analysis for special types of data*. Springer, 2002.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6114 [stat.ML].

Lin, Z. and Han, F. On boosting the power of Chatterjee's rank correlation. *Biometrika*, 08 2022. ISSN 1464-3510. doi: 10.1093/biomet/asac048. URL https://doi.org/10.1093/biomet/asac048. asac048.

Ludvigson, S. C. and Ng, S. The empirical risk–return relation: A factor analysis approach. *Journal of financial economics*, 83(1):171–222, 2007.

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].

Markham, A., Das, R., and Grosse-Wentrup, M. A distance covariance-based kernel for nonlinear causal clustering in heterogeneous populations. In *Conference on Causal Learning and Reasoning*, pp. 542–558. PMLR, 2022.

Moran, G. E., Sridhar, D., Wang, Y., and Blei, D. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2023.

Mulaik, S. A. *Foundations of factor analysis*. CRC press, 2009.

Ning, L. and Georgiou, T. T. Sparse factor analysis via likelihood and l 1-regularization. In *2011 50th ieee conference on decision and control and european control conference*, pp. 5188–5192. IEEE, 2011.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572, 1901.

Pournara, I. and Wernisch, L. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC bioinformatics*, 8:1–20, 2007.

Ramos-Carreño, C. and Torrecilla, J. L. dcor: Distance correlation and energy statistics in python. *SoftwareX*, 22: 101326, 2023.

Reichenbach, H. *The direction of time*. University of California Press, 1956.

Saini, A. *Superior: the return of race science*. Beacon Press, 2019.

Spearman, C. General ability, objectively determined and measured. *American Journal of Psychology*, 15:201, 1904.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Stubblefield, A. "Beyond the pale": Tainted whiteness, cognitive disability, and eugenic sterilization. *Hypatia*, 22(2):162–181, 2007.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, pp. 2769–2794, 2007.

Trendafilov, N. T., Fontanella, S., and Adachi, K. Sparse exploratory factor analysis. *Psychometrika*, 82:778–794, 2017.

Ullah, A. Computing clique cover with structural parameterization. *arXiv preprint arXiv:2208.12438*, 2022.

Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., and Stegle, O. Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nature methods*, 19(2): 179–186, 2022.

Yamamoto, M., Hirose, K., and Nagata, H. Graphical tool of sparse factor analysis. *Behaviormetrika*, 44:229–250, 2017.

A. Linear Factor Models

We now give a brief introduction to (linear) factor analysis, focusing on the key terms and mathematical ideas[2] that we connect to latent causal discovery and deep generative models, but for a more in-depth introduction and discussion about FA, see Mulaik (2009).

Definition A.1. A *factor model* represents a random (row) vector $\mathbf{M} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ consisting of n *measurement variables* as a linear transformation of a standard jointly normal random vector $\mathbf{L} \sim \mathcal{N}(\mathbf{0}, I_K)$ of $K < n$ *latent factors* via *factor loading weights* $W \in \mathbb{R}^{K \times n}$ plus a jointly normal random vector of n *error terms* $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, D)$, where $D \in \mathbb{R}_+^{n \times n}$ is a diagonal matrix, using the equation

$$\mathbf{M} = \mathbf{L}W + \boldsymbol{\epsilon}.$$

Given a sample $M \in \mathbb{R}^{s,n} \sim \mathbf{M}$ and assuming that $\mathbf{L}$ and $\boldsymbol{\epsilon}$ are probabilistically independent, the factor model can be estimated (Adachi, 2019) from the empirical covariance matrix $\widehat{\Sigma} = \frac{1}{s} M^\top M$ by finding $\widehat{W}$ and $\widehat{D}$ that minimize the squared Frobenius norm

$$\|\widehat{\Sigma} - \widehat{W}^\top \widehat{W} - \widehat{D}\|_F^2.$$

Such a solution is unique only up to orthogonal transformations of $\widehat{W}$, and so without further (e.g., in our case, causal) assumptions, finding a solution does not always warrant a meaningful interpretation of the resulting factor model. This unidentifiability especially poses a problem in *exploratory* FA, where there is no prior knowledge about $\widehat{\Sigma}, \widehat{W}$ or $\widehat{D}$, but less so in *confirmatory* FA, where experts can incorporate domain knowledge to constrain and interpret the possible solutions as well as test specific hypotheses.

Additionally, there are possibilities for either restricting or relaxing the FA model and methods of solving it, including closely related methods like PCA (Pearson, 1901; Hotelling, 1933; Jolliffe, 2002), ICA (Comon, 1994; Hyvärinen & Oja, 2000), and many others beyond our scope. Notably, compared to other related work, sparse FA (Ning & Georgiou, 2011; Trendafilov et al., 2017; Yamamoto et al., 2017), which penalizes $\widehat{W}$ according to the number of nonzero entries, produces solutions more closely related to those we find with NCFA. The two main differences between sparse FA and NCFA are that (i) rather than explicitly penalizing the solution to encourage sparsity, NCFA simply learns a causal structure that exhibits a structure typically sought in sparse FA, and (ii) like most FA methods, sparse FA still assumes linearity and Gaussianity, whereas NCFA can be highly nonlinear and nonparametric.

B. MeDIL Causal Models

The NCFA models used in this paper are a subfamily of the models known as MeDIL causal models, originally introduced in (Markham & Grosse-Wentrup, 2020). For contextualization purposes, we give a brief description of MeDIL causal models here.

A *MCM graph* is a triple $\mathcal{G} = \langle M, L, E \rangle$ where $M = \{1, \ldots, n\}$ and $L = \{\ell_1, \ldots, \ell_K\}$ are disjoint sets of vertices corresponding, respectively, to observed variables and latent variables, and E is the collection of directed edges between the nodes. In a MCM graph we require that the nodes in M are all *sinks* (i.e., have out-degree 0) and that the edges E are such that $\mathcal{G}$ is a *directed acyclic graph* (DAG). A distribution belongs to the *Measurement Dependence Inducing Latent (MeDIL) Causal Model* if it factors according to $\mathcal{G}$; i.e., the probability density function (or probability mass function) satisfies

$$f(x_1, \ldots, x_n, x_{\ell_1}, \ldots, x_{\ell_K}) = \prod_{i=1}^{n} f(x_i | x_{\mathrm{pa}_{\mathcal{G}}(i)}) \prod_{j=1}^{K} f(x_{\ell_j} | x_{\mathrm{pa}_{\mathcal{G}}(\ell_i)}).$$

Since we do not observe the latent variables, but only the observed variables, $X_1, \ldots, X_n$ the distributions of interest factorizes as

$$f(x_1, \ldots, x_n) = \int_{\mathcal{X}_L} \prod_{i=1}^{n} f(x_i | x_{\mathrm{pa}_{\mathcal{G}}(i)}) \prod_{j=1}^{K} f(x_{\ell_j} | x_{\mathrm{pa}_{\mathcal{G}}(\ell_i)}) dx_L = \prod_{i=1}^{n} \varphi_i(x_i). \tag{1}$$

[2] In case of conflicting notational conventions, e.g., L to denote a loading matrix in FA literature versus denoting a set of latent variables in the causal graphical model literature, we favor the latter.

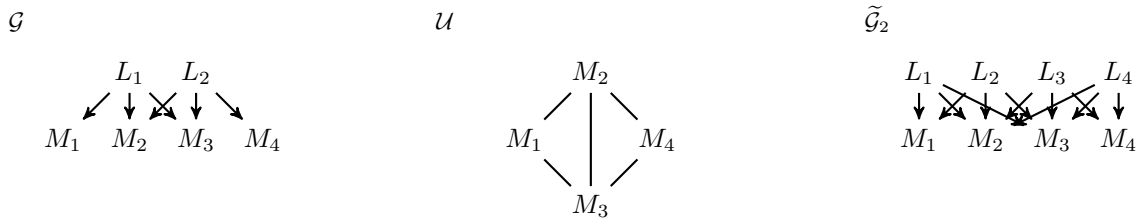


Figure 3: A UDG $\mathcal{U}$ on measurement variables M_1, M_2, M_3, M_4 and a corresponding minimum MCM graph $\mathcal{G}$ for the minimum edge clique cover $\mathcal{C} = \{C_1 = \{M_1, M_2, M_3\}, C_2 = \{M_2, M_3, M_4\}\}$. $\widetilde{\mathcal{G}}$ is the NCFA-graph for $\mathcal{G}$ with $\lambda = 4$ latent degrees of freedom. Note that all three graphs encode the exact same set of (marginal) independencies among the measurement variables M_1, M_2, M_3, M_4.

Note that since all observed variables are sink nodes in $\mathcal{G}$, integration over the latents yields a factorization of the distribution into a product of potential functions, one for each observed variable X_i that depends only on x_i. The MeDIL model $\mathcal{M}(\mathcal{G})$ consists of all observable distributions that arise according to this factorization.

In general, MCM graphs may have a complex directed acyclic structure over the latents that induce correlations amongst the observables. However, since we only consider the observed variables, we may reduce to a minimum MCM representation of the distribution. In a *minimum MCM graph*, we further assume that the latent variables are *source nodes* (i.e., have in-degree 0) and out-degree at least 1. Hence, a *minimum MeDIL model*; i.e., a MeDIL model that factorizes as in equation (1) explains the associations amongst the observed variables in the simplest possible way: via a collection of independent latents. These independent latents can be thought of as source nodes in the original MCM graph G, and the associated minimum MCM graph as the graph produced after we marginalize out all non-source latent variables. While learning the complex causal structure on the latents in a general MCM graph may be intractable, one can apply existing methods to estimate the simpler minimum MCM structure. Intuitively, one can think of learning the minimum MCM graph of a model as identification of the "primary causes" of the associations between the observed variables.

It follows from the factorization of a distribution according to a minimum MCM graph that the only conditional independence constraints amongst the observed variables are marginal independence constraints. These, in turn, encode the condition that two nodes do not have a shared (latent) parent. From the causal perspective, we see that two nodes in the system are marginally independent if and only if they share a latent common cause. Hence, minimum MeDIL models are the natural representation of Reichenbach's Common Cause Principle.

It follows that a natural representation of the model using only the observed variables is via an undirected graph $\mathcal{U} = \langle M, E \rangle$, where $i - j \notin E$ if and only if i and j are independent in the joint distribution of the observed variables. We call this graph the *unconditional dependence graph* (UDG) of the model. A *minimum edge clique cover* of the undirected graph U is a collection of cliques $\mathcal{C} = \{C_1, \ldots, C_K\}$ for which every pair i, j satisfying $i - j \in E$ is contained in at least one clique in $\mathcal{C}$ and there exists no set of cliques with this property that has cardinality smaller than $|\mathcal{C}|$. Since a UDG depends only on the observed variables, it is possible to learn a UDG from the available data via pairwise marginal independence tests. A minimum MCM graph that captures the marginal independence structure encoded via the UDG is then the minimum MCM graph where we have one latent ℓ_i for each clique $C_i \in \mathcal{C}$ and ℓ_i has a directed arrow to each node $j \in C_i$ (and no other adjacencies).

To learn the assignment of latents producing a minimum MCM graph from a UDG $\mathcal{U}$, we must learn a minimum edge clique cover of $\mathcal{U}$. The problem of identifying a minimum edge clique cover of an undirected graph is NP-hard. There are exact algorithms (Gramm et al., 2009; Ullah, 2022) showing it to be NP-complete and fixed-parameter tractable. However, there are also polynomial-time heuristic algorithms (Conte et al., 2020) that can efficiently handle very large graphs.

C. Identifiability of minimum MCM graphs: Examples and Counterexamples

An example of a minimum MCM graph that cannot be identified because it admits multiple edge clique covers is the graph:

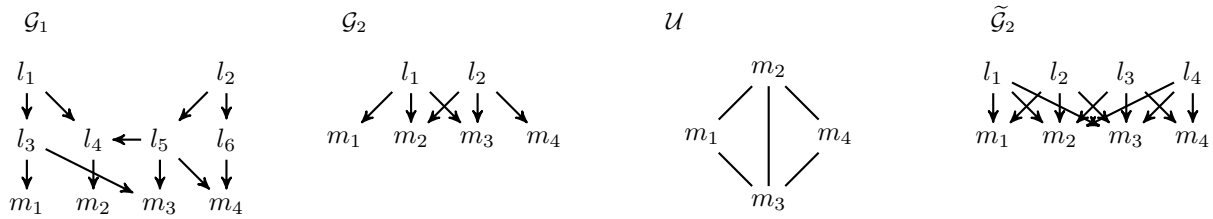
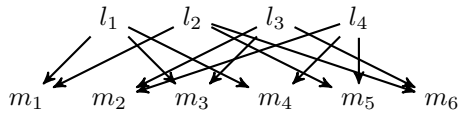


Figure 4: A MCM graph $\mathcal{G}_1$ with observed variables m_1, m_2, m_3, m_4, a minimum MCM graph $\mathcal{G}_2$ representing the same set of observable distributions (M_1, M_2, M_3, M_4) and the associated UDG $\mathcal{U}$. The (unique) minimum edge clique cover of $\mathcal{U}$ is $\mathcal{C} = \{\{m_1, m_2, m_3\}, \{m_2, m_3, m_4\}\}$. Hence, $\mathcal{G}_2$ has exactly one latent variable for each of these two cliques, connected to exactly the nodes in a single clique in $\mathcal{C}$. The NCFA-graph $\widetilde{\mathcal{G}}_2$ is an augmentation of $\mathcal{G}_2$ for $\lambda = 4$ latent degrees of freedom (see Section 2). All four graphs encode a single marginal independence constraint, $M_1 \perp\!\!\!\perp M_4$, and hence represent the same family of observed distributions.

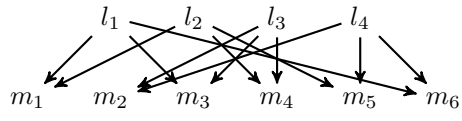
$\mathcal{G}$



.

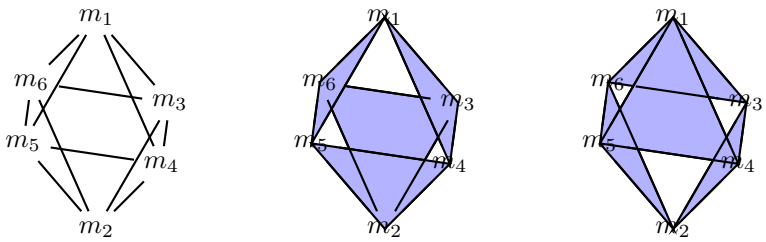
The UDG of $\mathcal{G}$ is the edge graph $\mathcal{U}$ of the octahedron which admits exactly two minimum edge clique covers: $\mathcal{C}_1 = \{\{1,3,4\},\{1,5,6\},\{2,3,6\},\{2,4,5\}\}$ and $\mathcal{C}_2 = \{\{1,3,6\},\{1,4,5\},\{2,3,4\},\{2,5,6\}\}$. $\mathcal{C}_1$ recovers the minimum MCM graph $\mathcal{G}$, whereas $\mathcal{C}_2$ produces the minimum MCM graph $\mathcal{H}$, depicted below:

$\mathcal{H}$



.

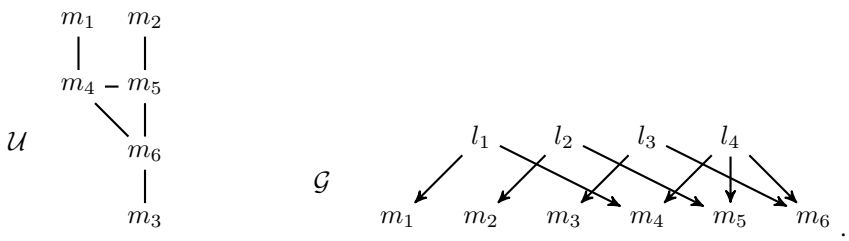
This issue arises due to the spherical structure of the UDG $\mathcal{U}$, depicted on the left below. In the remaining two figures, the each shaded triangle represents the children of a latent node (i.e., a clique in a minimum edge clique cover) of a possible MCM graph with associated UDG $\mathcal{U}$. The middle figure represents the minimum MCM graph $\mathcal{G}$ and the rightmost figure represents the alternative $\mathcal{H}$.



The spherical structure of this graph creates symmetry allowing for multiple minimum edge clique covers. In particular, this gives an example to where we lack identifiability and Theorem 2.5 does not apply. Interestingly, we see that this lack of identifability can be "worst possible" in the sense that if the true minimum MCM graph is $\mathcal{G}$, but the wrong edge clique cover is chosen, we could learn $\mathcal{H}$, which captures none of the true latents and instead specifies four completely incorrect latents. Future work, possibly incorporating interventional knowledge, to address such identifiability concerns would be of interest.

On the other hand, graphs admitting a *pure measurement variable* in each clique of a minimum edge clique cover will have a unique minimum edge clique cover, making the resulting minimum MCM graph identifiable (see Theorem 2.5). From a graphical perspective, we call this condition the *1-pure child constraint* since it insists that each latent node ℓ_i in the minimum MCM graph has (at least) one child m_i which has no other parents than ℓ_i (i.e., $\ell_i \to m_i$ is an edge of $\mathcal{G}$ and m_i has no other adjacencies in $\mathcal{G}$). However, the 1-pure-child constraint is not necessary for a UDG to have a unique minimum edge clique cover and hence for identifiability to hold (according to Theorem 2.5). In particular, the following UDG $\mathcal{U}$ has

the unique minimum edge clique cover $\mathcal{C} = \{\{1,4\},\{2,5\},\{3,6\},\{4,5,6\}\}$ and hence the corresponding minimum MCM graph $\mathcal{G}$ is identifiable:



Note here that the latent variables ℓ_1, ℓ_2 and ℓ_3 each have a pure child (m_1, m_2 and m_3, respectively). However, the latent ℓ_4 does not. Hence $\mathcal{G}$ is identifiable but does not satisfy the 1-pure-child constraint.

D. NCFA Implementation

We provide a Python implementation (including thorough documentation) of the NCFA algorithm (Algorithm 1) as well as scripts for reproducing all of our results, released as a free/libre software package: `https://after.review`.

Our implementation makes use of the following Python packages: `NumPy` (Harris et al., 2020), `PyTorch` (Paszke et al., 2019), `dcor` (Ramos-Carreño & Torrecilla, 2023), and `xicorrelation` (`https://github.com/jettify/xicorrelation`). Additionally, we provide a Python wrapper of the Java `ECC` package (`https://github.com/Pronte/ECC`).

Further details can be found in our code and documentation, but we summarize the most important implementation details in the following:

- **Marginal independence testing (Step 1, Algorithm 1)**: We use statistical hypothesis tests of independence (either using `dcor` for distance covariance based tests or `xicorrelation` for Chatterjee's coefficient based tests) for specified threshold value α, with the p-values being computed based on the asymptotic theory as opposed to with permutation tests.
- **Edge clique cover (Step 2, Algorithm 1)**: We use the default heuristic solver of the `ECC` package, except when making the 1-pure-child assumption, in which case the corresponding learned UDG $\widehat{\mathcal{U}}$ has equal intersection and independence numbers, a condition known (Deligeorgaki et al., 2022) to allow for a subcubic time exact solver, which we implement ourself (see package documentation for further details).
- **Assignment of latent degrees of freedom according to λ (Step 3, Algorithm 1)**: First, each of the K cliques from the edge clique cover is assigned one latent; then, the remaining $\lambda - K$ latents are distributed equally over the cliques.
- **VAE architecture (Step 4, Algorithm 1)**: The learned UDG $\widehat{\mathcal{U}}$ induces an NCFA model with NCFA graph $\widehat{\mathcal{G}}$ (cf. Section 2.1). The NCFA graph encodes the structural connections between the latents and observables as a biadjacency matrix, which is used to mask the connections in the decoder of the VAE. In this way, each observable only depends on the latent it is connected to in the learned NCFA graph $\widehat{\mathcal{G}}$ (cf. Definition 2.4). The decoder structure is given by a single layer of the SparseLinear network with a masked linear activation function. The model was trained using the AdamW optimizer with a learning rate of 1e-5 and each time run for 200 epochs.

Except for the last experiment, no hyperparameter tuning was performed, and instead default, reasonable choices are used (e.g. $\alpha = 0.05$ and $\lambda = \lfloor n^2/4 \rfloor$). We anticipate improvements are possible with careful hyperparameter tuning. All experiments were run on Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz. Total combined runtime was about 50 hours.

E. Evaluation Metrics

NCFA faces a trade-off between causal constraints and expressivity: an unconstrained, fully connected VAE ignores this structure, and has free reign to fit the data arbitrarily, at the cost of interpretability and potentially *a*causal relationships (e.g. spurious correlations). The additional structure offered by the minimum MCM graph in NCFA brings in causal structure

and interpretation, but can hamper training if the structure is incorrect. Of course, when the causal structure is correct, there should be no significant loss in expressivity. Thus, ideally we will see no significant degradation in the loss, which is an indicator of structural fidelity. We measure this with the metric Δ which is the difference between the loss of an unconstrained, baseline VAE and the NCFA loss. On synthetic data where we know the causal ground truth, we can also directly measure structural fidelity using graph comparison metrics. See Appendix E for detailed definitions of our metrics.

Distance measures Unlike classical VAE methods, NCFA has a causal discovery step in which it infers the UDG upon which the VAE in the model is based. The structural Hamming distance (SHD) is commonly used to quantify the difference between graph structures and is simply defined to be the number of edges that appear in one graph but not the other. SHD, however, is not applicable in our case as (i) it is defined for graphs having the same number of vertices, whereas for a fixed finite sample from a set of measurement variables, differently estimated minimum MCMs may have differing numbers of latent variables, and (ii) even if we compute, e.g., the SHD between the undirected graphs $\mathcal{U}, \mathcal{U}'$ (which *do* have the same number of vertices), this aligns poorly with intuitions about distance between their respective generating MCM graphs $\mathcal{G}, \mathcal{G}'$. To remedy this, we introduce the following distance:

Definition E.1. The *structural Frobenius* (or *medil*) *distance* between biadjacency matrices $B_1 \in \{0,1\}^{K_1,n}$ and $B_2 \in \{0,1\}^{K_2,n}$, corresponding to minimum MCM graphs $\mathcal{G}_1, \mathcal{G}_2$ each with $n = |M|$ measurement variables and K_i latents, respectively, for $i = 1, 2$, is defined as

$$\text{SFD}(B_1, B_2) \coloneqq \|B_1^\top B_1 - B_2^\top B_2\|_F^2.$$

Note that, by virtue of being defined using the Frobenius norm, the medil distance is a proper distance metric. The intuition behind this definition is that we transform a given minimum MCM structure into a weighted undirected graph, allowing it to be easily compared to other minimum MCM structures over the same set of measurement variables. Via this transformation, the weights keep track of how many latent parents each pair of measurement variables have in common as the (i, j)-th entry in the matrix $B_1^\top B_1$ records the number of colliderless paths between measurement variables i and j in the graph. For our graphs, two nodes admit such a path between them if and only if they are connected by a latent, and this connection adds exactly 1 to the (i, j)-th matrix entry. Hence, these matrices will be equal if and only if the MeDIL models are identical (for a consistent ordering of measurement variables) (Deligeorgaki et al., 2022, Section 2.2). Figure 5 provides an example comparing the SFD between minimum MCMs and the SHD between their corresponding (unweighted) undirected graphs.

VAE loss function The training and validation losses reported here and in the main paper are standard ELBO metrics (e.g. Kingma & Welling, 2014) computed using a 70/30 training/validation split.

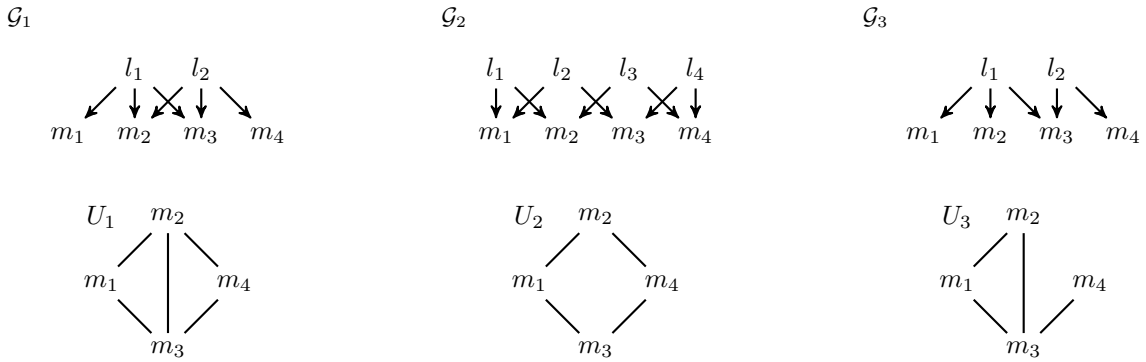
F. Simulations: Specifications and Additional Results

Data Generation We generated 10 Erdős-Rényi random undirected graphs (Gilbert, 1959) on 10 nodes for each density value $p \in \{0.1, 0.2, \ldots, 0.9\}$. These are actual density values, not expectations, i.e., for each density value we sample uniformly from graphs with $p\binom{10}{2}$ edges. These random graphs constitute UDGs, so we compute their minimum edge clique covers to determine the corresponding MCM structure. Importantly, while this approach lets us sample from the entire space of minimum MCM structures, it only lets us directly specify the graph density of the UDG over measurement variables and not the number of latent causal factors or the density of the MCM structure.

Using the ground truth MCM structures as generated above, we construct a linear factor model (Definition A.1) with factor loading weights drawn from the interval $[-2, -0.5) \cup (0.5, 2]$ and standard normal error terms, from which we draw 10 different data sets (per edge density p) containing 1000 observations each, totaling 90 different ground truth causal structures and 900 different data sets.

Model Hyperparameters For each data set, we trained a NCFA model with $\alpha = 0.05$, $\lambda = 25$ (which includes estimating the causal structure) as well as two baselines: (i) a classic fully connected VAE with the same number of latent degrees of freedom, and (ii) a NCFA model given the true causal structure and with the latent degrees of freedom λ being set equal to the true number of causal latent factors $K = |L|$.

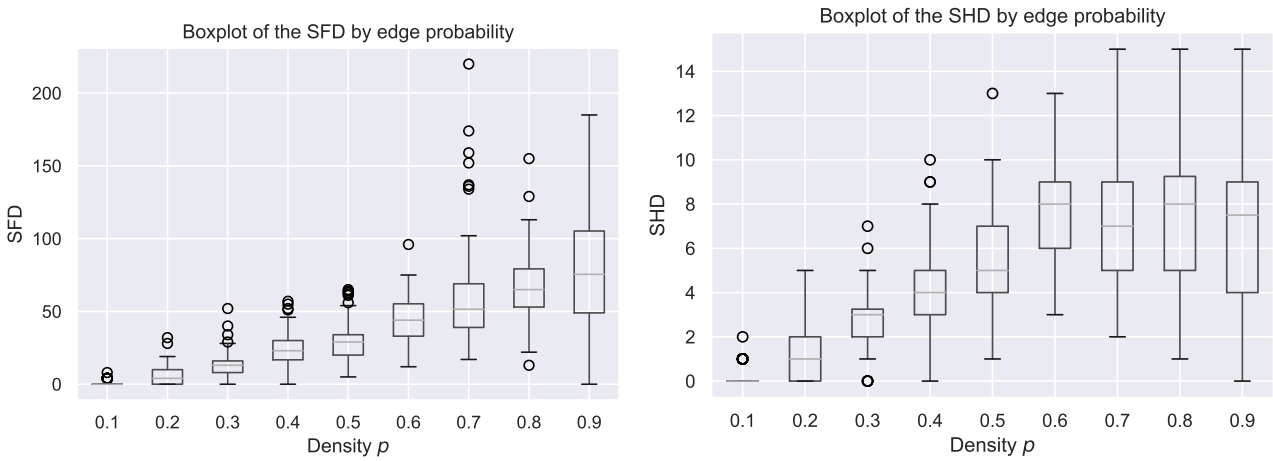
Additional Results As is commonly seen in causal discovery tasks, NCFA recovers causal structure well in the sparse setting but increasingly less so in denser settings. Causal discovery is notoriously difficult, especially in the small-sample



$$B_1 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \qquad B_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad B_3 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

	$\mathcal{G}_1$ and $\mathcal{G}_2$	$\mathcal{G}_1$ and $\mathcal{G}_3$	$\mathcal{G}_2$ and $\mathcal{G}_3$
structural Hamming distance	1	1	2
medil distance	10	5	7

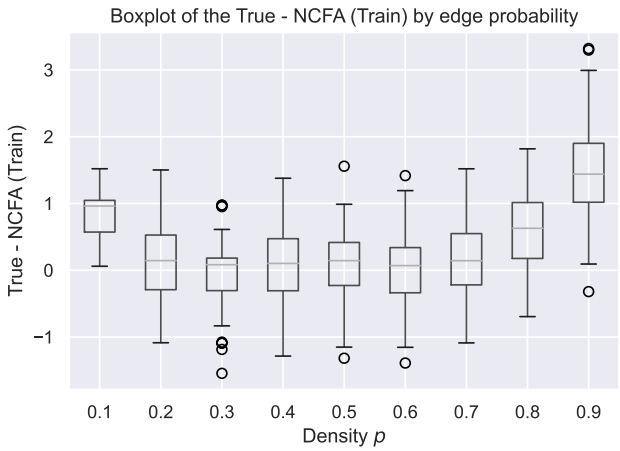
Figure 5: Pairwise comparison of structural Hamming distance and medil distance on three causal factor structures, $\mathcal{G}_1, \mathcal{G}_2$, and $\mathcal{G}_3$. For each structure $\mathcal{G}_i$, we also show its associated undirected graph U_i and its biadjacency matrix B_i.



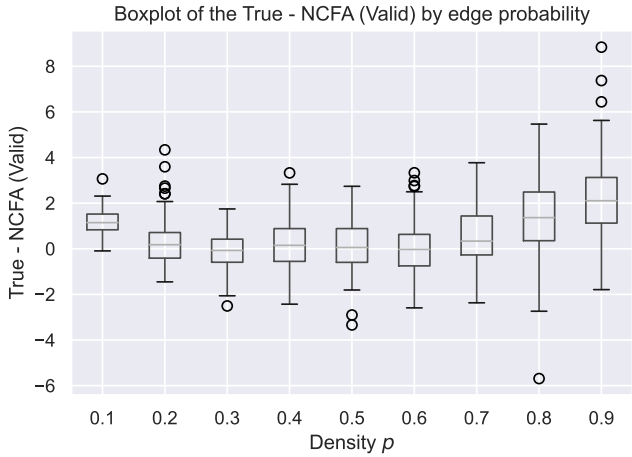
(a) (b)

Figure 6: Results of NCFA on synthetic data sets from randomly generated graphs, in terms of causal structure learning: (a) shows structural Frobenius distance (SFD) between learned (biadjacency matrix representation of) causal structures and the ground truth; (b) shows structural Hamming distance (SHD) between (undirected graph representation of) learned causal structures and the ground truth.

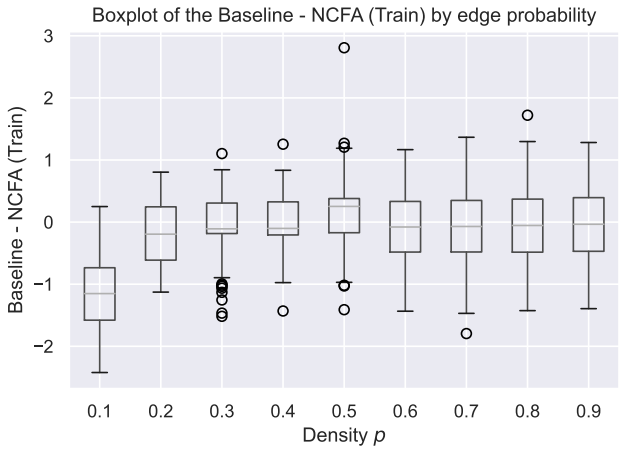
regime, but NCFA benefits from only needing to perform *marginal* independence tests (so the conditioning set is always empty). In terms of performance as a generative model, we see that NCFA generally improves the validation loss compared to the baseline VAE since the median loss difference is above 0 for all edge densities except for $p = 0.1$, even as the true graph density increases. This indicates both that the causal structure provides helpful constraints in the NCFA pipeline and that NCFA is robust in the face of moderate misestimation of the causal structure.



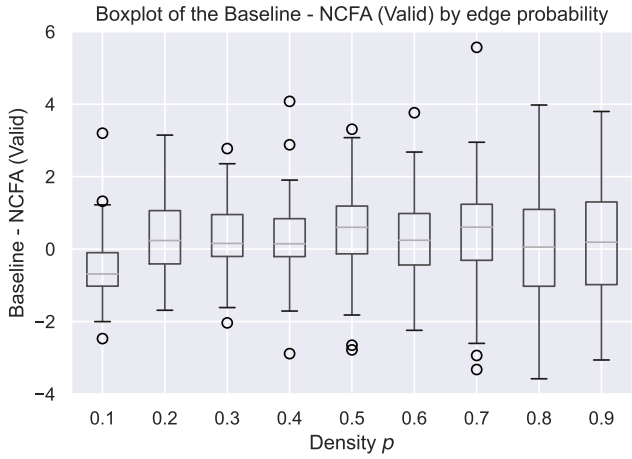
(a) Training-Δ-True



(b) Validation-Δ-True



(c) Training-Δ-Baseline VAE



(d) Validation-Δ-Baseline VAE

Figure 7: Results of NCFA on synthetic data sets from randomly generated graphs, in terms of final loss of the trained deep generative models: (a) shows Training-Δ-True, the difference of training loss between the learned NCFA model and the NCFA model when given the ground truth causal structure; (b) shows Validation-Δ-True, difference between NCFA and ground truth validation losses; (c) shows Training-Δ-Baseline, difference between NCFA and Baseline VAE training losses; (d) shows Validation-Δ-Baseline, difference between NCFA and Baseline validation losses.

Figure 6 shows the distances (both SFD and SHD) between the learned and true causal structures, in the form of a box plot for each different edge density p, demonstrating better performance (in terms of causal structure learning) for sparser graphs. However, in terms of performance as a generative model, Figures 7 and 8 (respectively showing boxplot summary of the Δ losses and averaged Δ losses for different generating edge densities) demonstrate that NCFA generally achieves training loss comparable to the baseline VAE while improving the validation loss, and at the same time improving both training and validation loss compared to the ground truth model (owing to NCFA's more expressive VAE architecture and higher λ). Together, these results indicate both that the causal structure provides helpful constraints in the NCFA pipeline and that NCFA is robust in the face of moderate misestimation of the causal structure.

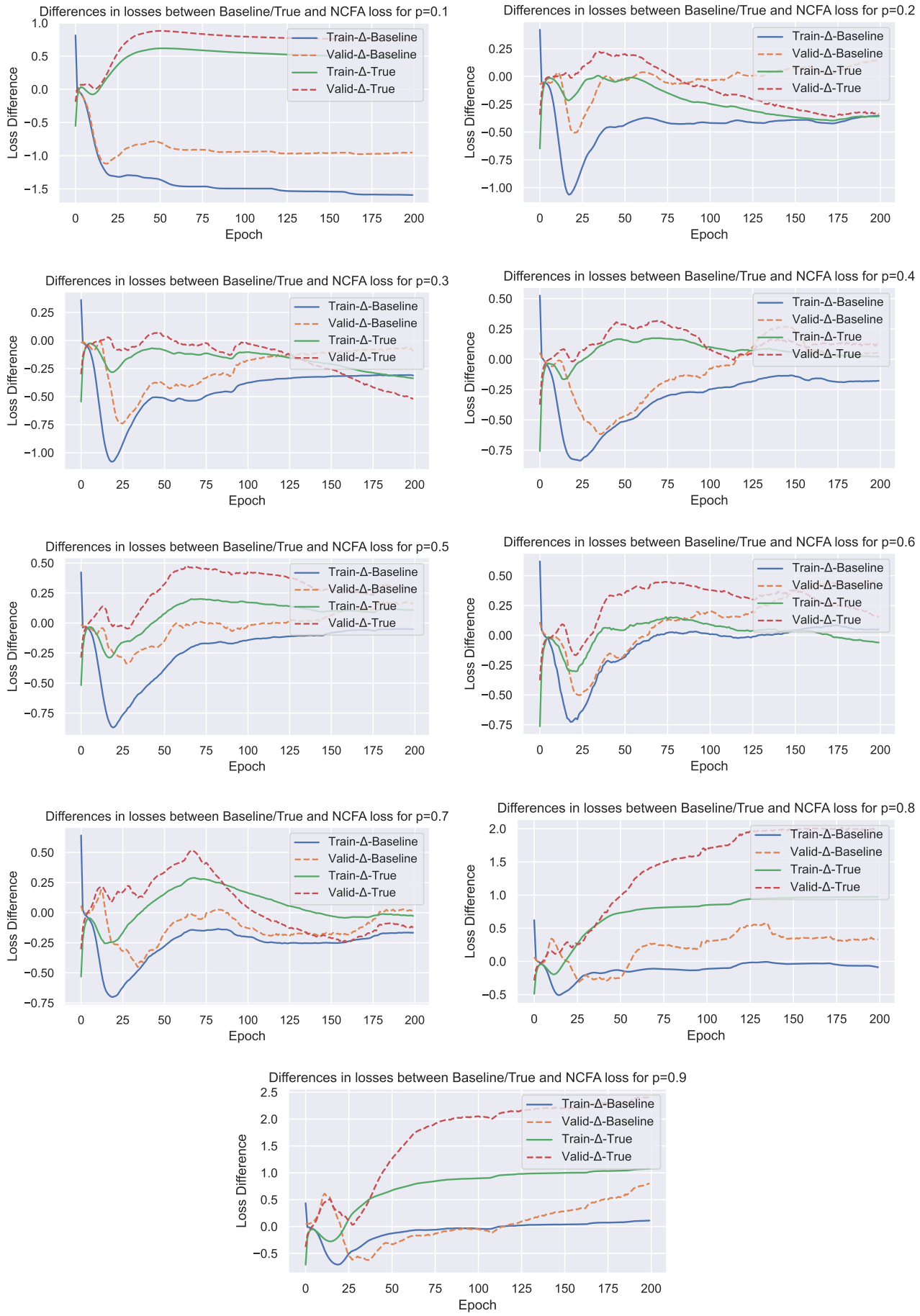


Figure 8: Δ curves for NCFA versus baseline and ground truth on synthetic data sets from randomly generated graphs,

	samp size	n	α	λ	$\lvert L \rvert$	Training-Δ	Validation-Δ
MNIST	42000	784	0.05	153664	1	-0.00475	-0.04814
TCGA	632	1000	0.05	250000	8129	0.11488	0.11865
MNIST	42000	784	0.001	7800	560	-76.682	-74.163
TCGA	632	1000	0.05	10000	969	-78.721	-68.117

Table 2: Results of NCFA on two real data sets: the first two rows show results for default parameter values, while the second two show results of tuned parameters, decreasing α for MNIST and λ for both; n reports the number of measurement variables, $|L|$ the number of learned causal latents, while Training- and Validation-Δ report the differences between NCFA and baseline VAE losses.

G. Real Data Analysis: Specifications and Additional Results

Model Hyperparameters As noted in Section 4, we used default settings for the first run on MNIST and TCGA (first and second rows of Table 1, and Figures 9a and 9b). For the second run on each (third and fourth rows of Table 1, and Figures 9c and 9d), we tuned the hyperparameters slightly, changing $\alpha = 0.001$ and using distance covariance based tests[3] on MNIST to encourage greater sparsity in the learned NCFA model. Furthermore, for the second run we made the 1-pure-child assumption for both MNIST and TCGA, allowing us to greatly reduce λ and make use of a polynomial time exact minimum edge clique cover solver (described in Appendix D).

Additional Results On both datasets, the default λ and maximum allowed $|L| < \lambda$ were quite large, so we also ran experiments under the 1-pure-child assumption (second two rows of Table 2), which guarantees that $|L| \leq n$, allowing us to safely reduce λ from $\lfloor n^2/4 \rfloor$ to, e.g., $10n$. Additionally, we decreased α to 0.001 for MNIST, taking advantage of the large sample size and encouraging NCFA to learn a sparser structure. However, based on the training and validation differences, NCFA failed to converge properly compared to the baseline VAE. In the case of MNIST, we attribute this to it arguably being a data set without causally meaningful sparse latents. For TCGA, the performance of NCFA without the 1-pure-child assumption yielded a better performance than the baseline VAE. Hence, the decrease in performance of NCFA under this constraint could suggest that the true causal structure of TCGA simply does not abide by the 1-pure-child assumption. Collectively, these results suggest that NCFA with default parameter specifications appears to yield competitive, if not improved, performance over baseline VAE models that successfully incorporate causal structure when it is present to be learned. When NCFA has free reign to learn whatever causal structure (when it exists, as in TCGA) can be gleaned from the data, it appears to benefit training. However, the second round of experiments suggest that one should take care when adjusting the algorithm to fit a specified causal structure, such as the 1-pure-child constraint, as forcing possibly nonexistent causal structure into the model may be detrimental to the models predictive capabilities. This is in line with the observation at the start of Section 4 that one risks hampering training when the causal structure is misspecified.

Figure 9 shows the Δ curves corresponding to the runs described in Section 4 and Table 1, demonstrating that NCFA with default parameter specifications yields competitive, if not improved, performance over baseline VAE models. When NCFA has free reign (as in the default run) to learn whatever causal structure (when it exists, as in TCGA) can be gleaned from the data, it appears to benefit training. However, the second round of experiments suggest that one should take care when adjusting the algorithm to fit a specified causal structure, such as the 1-pure-child constraint, as forcing possibly nonexistent causal structure into the model may be detrimental to the models predictive capabilities.

[3] Instead of Chaterjee's coefficient based tests—it is known that distance covariance can be more powerful when one is concerned with finding independence, whereas Chatterjee's coefficient may be preferred for measuring strength of dependence (Chatterjee, 2022; Lin & Han, 2022).

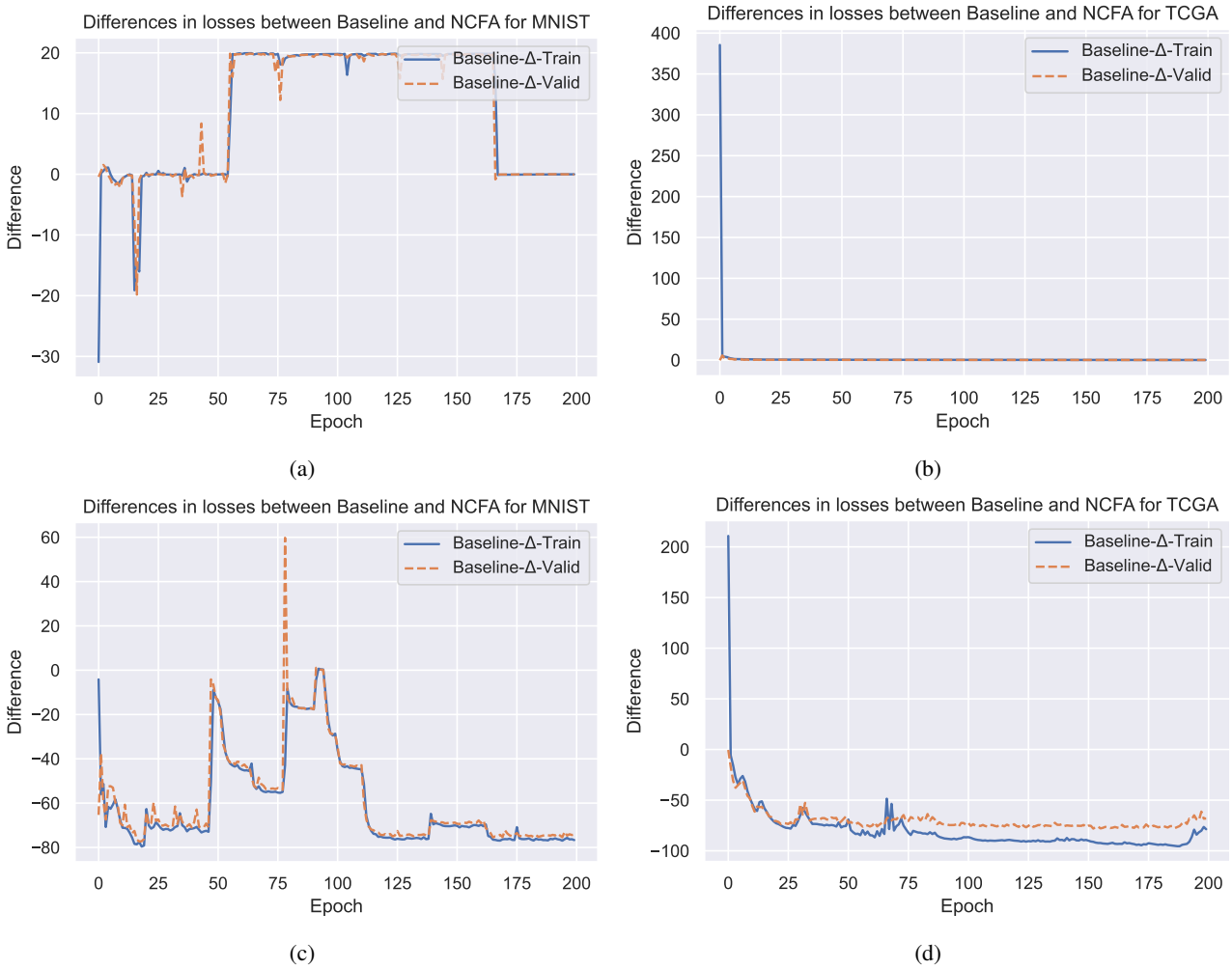


Figure 9: Δ curves for NCFA versus baseline VAE on MNIST and TCGA data sets. (a) and (b) show results of NCFA with default settings (on MNIST and TCGA respectively), while (c) and (d) show results of NCFA with additional 1-pure-child assumption (for both) and more sensitive independence tests (for MNIST).