On Differentiable Bayesian Causal Structure Learning

Simon Rittel[1,2] **Sebastian Tschiatschek**[1]

[1]Faculty of Computer Science, University of Vienna, Vienna, Austria
[2]UniVie Doctoral School Computer Science , University of Vienna , Vienna, Austria

Abstract

This extended abstract reviews differentiable Bayesian causal structure learning (CSL) and discusses why recent works on Bayesian causal discovery published in top-tier conference do not yet meet important desiderata. In particular, we advocate against the current trend of global regularization via prior terms.

Introduction Approximately 30 years have passed since seminal research papers [Heckerman et al., 1994, Heckerman, 1995, Friedman and Koller, 2000] provided the impetus for the research on Bayesian approaches to learning the structure of probabilistic graphical models. With the discovery of identifiable semi-parametric models [Shimizu et al., 2006, Hoyer et al., 2008, Zhang and Hyvärinen, 2009, Loh and Bühlmann, 2014, Peters et al., 2014] and the combination with interventional data, learning of Bayesian networks became gradually a causal problem.

Bayesian formulation In our analysis, we focus on differentiable Bayesian approaches to causal structure learning (CSL) that learn a generative model for both the causal graph $\boldsymbol{G}$ and the data $\mathbf{X} \coloneqq \{\boldsymbol{X}^{(n)}\}_{n=1}^{N}$ without the restriction to discrete random variables nor linear relationships:

$$p(\boldsymbol{G}, \mathbf{X}) = p(\boldsymbol{G}) \prod_{n=1}^{N} p(\boldsymbol{X}^{(n)} | \boldsymbol{G}) \,, \tag{1}$$

where $\boldsymbol{X}^{(n)}$ are i.i.d. according to the underlying *Functional Causal Model* (FCM)[1]. The posterior over the graph is proportional to the joint probability in Equation 1[2]:

$$p(\boldsymbol{G}|\mathbf{X}) = \frac{p(\boldsymbol{G}, \mathbf{X})}{\sum_{\boldsymbol{G}} p(\boldsymbol{G}, \mathbf{X})} \propto p(\boldsymbol{G}, \mathbf{X}) \,. \tag{2}$$

[1]FCM, DAGs & CSL are properly introduced in Appendix A.
[2]See Appendix B for the full model including parameter uncertainty.

The Bayesian formulation allows models to express uncertainty in the prediction of the causal structure.
In recent years the number of publications at top-tier conferences that follow this line of research and explicitly claim a Bayesian formulation of CSL have increased significantly. From the authors' viewpoint, a substantial number of these works rather resemble point predictors with regularization than truly Bayesian posteriors for the causal graph. The contribution of this work on Bayesian CSL is to review general desiderata, to address common shortcomings in recent models, and to share new insights hopefully encouraging further research on the outlined issues.

Probability distribution over DAGs For acyclic FCMs the support of both, the prior and posterior distribution, should be restricted to *Directed Acyclic Graphs* (DAGs)[1]. Non-negative, differentiable constraint functions $h(\boldsymbol{G})$ allow relaxation of the combinatorial search problem to a continuous program [Zheng et al., 2018, Yu et al., 2019, Bello et al., 2022]. The more cycles a graph has, the greater $h(\boldsymbol{G}) \geq 0$, with equality only for acyclic graphs. Lorch et al. [2021] propose a Gibbs prior in which such constraint function is incorporated via an exponential term with a prefactor λ that is annealed over the training:

$$p(\boldsymbol{G}) \propto \exp\big(-\lambda h(\boldsymbol{G})\big) \,. \tag{3}$$

Increasing λ decreases the probability mass of any weighted adjacency matrix with cycles. For a sufficiently high λ, the posterior in Equation 2 is almost exclusively concentrated on acyclic graphs. We argue that due to the interference with the likelihood that scales with N in Equation 1, the regularization strength λ and therefore the prior should depend on the sample size N which is rather untypical for Bayesian priors. Moreover, when modeling the probabilities of edges in $\boldsymbol{G}$ independently, the prior regularizes them equally via $h(\boldsymbol{G})$ to avoid any cycle. Consequently, it locks the distribution to some ordering over the nodes. To overcome this limitation, Lorch et al. [2021] model the posterior by a particle representation. Otherwise, the acyclicity enforced via a Gibbs factor rather limits any method to a point estimator

Accepted for the 9th Causal Inference Workshop at the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024).

for the most likely ordering in the causal graph [Annadani et al., 2021, Geffner et al., 2022, Lorch et al., 2022, Ashman et al., 2023].
Another line of research considers the restriction of the posterior distribution to DAGs via modeling a permuted, upper triangular adjacency [Cundy et al., 2021, Charpentier et al., 2022, Annadani et al., 2023] or restricting possible edges iteratively [Deleu et al., 2022]. With the notable exception of DPM-DAG [Rittel and Tschiatschek, 2023], these approaches do not allow to specify probabilistic knowledge on particular causal relations in the prior. The novelty of DPM-DAG consists in an ordering-based, probabilistic model that empowers domain experts to build a consistent prior over DAGs by iteratively specifying marginal probabilities over direct causal edges. In contrast to [Cundy et al., 2021], the authors of DPM-DAG mask a full adjacency matrix to ensure acyclicity and that the modeled graph parameters correspond to the same causal relation under different permutations which is a necessary requirement for going beyond point estimation Rittel and Tschiatschek [2023]. The corresponding probabilistic mask captures the idea of an ordering-based search that allows to partially mediate the super-exponentially space of DAGs [Friedman and Koller, 2000, Teyssier and Koller, 2005].
Modeling both the prior and posterior with the same distribution facilitates simple sequential Bayesian updates and continual learning. Note that both explicitly require modeling a distribution over parameters to update their uncertainty from previously observed data correctly, cf. Appendix B.

Sparsity of the causal graph For supergraphs of the true underlying causal graph the likelihood in Equation 1 can only increase [Koller and Friedman, 2009]. Occam's razor advocates to favor simpler models, i.e. ones with fewer edges that yield the same likelihood, and motivates sparsity regularization of the predicted causal graphs. Sparsity-favoring priors are reported to be beneficial compared to uniform priors that assign a high probability on complex structures, i.e. denser graphs [Eggeling et al., 2019].
Following Lorch et al. [2021], several recent works on differentiable Bayesian CSL also introduce sparsity over l_1 or l_2 norms in an exponential term in the prior [Annadani et al., 2021, Tigas et al., 2022, Hägele et al., 2023, Geffner et al., 2022, Ashman et al., 2023]. Such a term was initially motivated to express prior knowledge about the expected number of edges. While the logarithm of the modeled joint distribution renders it equivalent to an ordinary, additional regularization term, we argue that the formulation within the prior is misguiding for two reasons.
Firstly, domain experts barely have such knowledge that applies equally to all variables. Since modeled variables are distinct, labeled entities, experts rather have subjective beliefs about particular causal relations among them. Secondly and more importantly, the likelihood term requires the regularization of the number of edges, hence, its strength should scale proportional to the number of observed samples N (cf. Equation 1 and Appendix C). We agree with [Eggeling et al., 2019] that a prior depending on N as the 'data prior' [Pensar et al., 2016] or a Horseshoe prior [Piironen and Vehtari, 2017] is not a prior in the Bayesian sense, but differ in our conclusion that regularization has to be paired with the likelihood term.
Even in the wider field of differentiable CSL which is not limited to only Bayesian approaches, the often used l_1 regularization provides a bias in contrast to the l_0 penalty typically applied in discrete search approaches [Bhattacharya et al., 2021, Ng et al., 2024]. An alternative to sparsity regularization is pruning a learned graph. Thresholding edge probabilities or weights both require setting a hyperparameter that plays a decisive role for the final results of CSL [Ng et al., 2024].

Tuning any prefactor for the strength of the regularization or thresholding remains necessary, though challenging, since the contribution of parents can be nonlinear and differ in its scale. To the best of the authors' knowledge, there's no agreed consensus on hyperparameter tuning yet. Due to the nature of the problem, hyperparameter training on a validation set generated from the same causal graph does not help. We explicitly stress that using data that originate from other causal graphs is not substantiated. While it does yield improvements on metrics for synthetic data sets by hinting information about the distribution of their underlying random graphs, it should not help in learning a single graph. We sincerely believe that this rather veiled hyperparameter training hinders scientific progress in the field of CSL, since the comparability of different methods is subverted.
For synthetic data sets researchers have access to the true causal graph, hence, we argue that any hyperparameter choice for sparsity regularization shall be conclusively substantiated. To compare different methods for CSL, it seems best practice to evaluate all method using the same sparsity regularization, thresholding or pruning if applicable.

Conclusion This review discusses some limitations of recent works on Bayesian CSL. In particular, we argue that from a Bayesian viewpoint the observation model has to be regularized towards sparser graphs. The qualitative analysis shows that neither sparsity nor acyclicity should be enforced as additional terms in a Gibbs prior and calls for further research on adaptive regularization, thresholding, and pruning. Moreover, the expressivity of probabilistic models whose support is limited to DAGs needs further investigation.
In closing, research on Bayesian CSL requires more attention on non-particle models that can approximate different FCMs simultaneously, i.e. that can process the information of different sampled graphs. Due to the outlined shortcomings, we conclude that most Bayesian CSL algorithms that learn a generative model over observed data alongside the causal graph rather resemble point estimators.

Acknowledgements

This work was partially funded by the Federal Ministry of Education, Science and Research (BMBWF) of Austria within the interdisciplinary project "Digitize! Computational Social Science in the Digital and Social Transformation".

References

Yashas Annadani, Jonas Rothfuss, Alexandre Lacoste, Nino Scherrer, Anirudh Goyal, Yoshua Bengio, and Stefan Bauer. Variational causal networks: Approximate Bayesian inference over causal structures. *KDD Workshop on Bayesian causal inference for real world interactive systems*, 2021.

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. BayesDAG: Gradient-based posterior inference for causal discovery. In *Advances in Neural Information Processing System*, volume 36, 2023.

Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal reasoning in the presence of latent confounders via neural ADMG learning. In *International Conference on Learning Representations*. Openreview, 2023.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2314–2322. PMLR, 2021.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014.

Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable DAG sampling. In *International Conference on Learning Representations*. Openreview, 2022.

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for Bayesian causal discovery. In *Advances in Neural Information Processing Systems*, volume 34, pages 7095–7110, 2021.

Tristan Deleu, António Góis, Chris Chinenye Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 518–528. PMLR, 2022.

Ralf Eggeling, Jussi Viinikka, Aleksis Vuoksenmaa, and Mikko Koivisto. On structure priors for learning Bayesian networks. In *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1687–1695. PMLR, 2019.

Nir Friedman and Daphne Koller. Being Bayesian about network structure. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, pages 201–210. Morgan Kaufmann, 2000.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Deep end-to-end causal inference. *NeurIPS Workshop on Causality for Real-world Impact*, 2022.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. BaCaDI: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1411–1436. PMLR, 2023.

David Heckerman. A Bayesian approach to learning causal networks. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, pages 285–295. Morgan Kaufmann, 1995.

David Heckerman, Dan Geiger, and David Maxwell Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence*, pages 293–301. Morgan Kaufmann, 1994.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pages 689–696, 2008.

Marcus Kaiser and Maksim Sipos. Unsuitability of NOTEARS for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters*, 54(3): 1587–1595, 2022.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1): 3065–3105, 2014.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable Bayesian structure learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 24111–24123, 2021.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ignavier Ng, Biwei Huang, and Kun Zhang. Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 71–105. PMLR, 2024.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

Johan Pensar, Henrik J. Nyman, Jarno Lintusaari, and Jukka Corander. The role of local partial independence in learning of bayesian networks. *International Journal of Approximate Reasoning*, 69:91–105, 2016.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. In *Journal of Machine Learning Research*, volume 15(1), pages 2009–2053. JMLR, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press, 2017.

Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 905–913. PMLR, 2017.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems*, volume 34, pages 27772–27784, 2021.

Alexander G. Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, and Sebastian Weichwald. A scale-invariant sorting criterion to find a causal order in additive noise models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Simon Rittel and Sebastian Tschiatschek. Specifying prior beliefs over DAGs in deep Bayesian causal structure learning. In *European Conference on Artificial Intelligence*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1962–1969. IOS Press, 2023.

Jonas Seng, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Learning large DAGs is harder than you think: Many losses are minimal for the wrong DAG. In *International Conference on Learning Representations*. Openreview, 2023.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. In *Journal of Machine Learning Research*, volume 7, pages 2003–2030. JMLR, 2006.

Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 548–549. AUAI Press, 2005.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? Experimental design for causal models at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 2019.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press, 2009.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, pages 9492–9503, 2018.

On Differentiable Bayesian Causal Structure Learning (Supplementary Material)

Simon Rittel[1,2] **Sebastian Tschiatschek**[1]

[1]Faculty of Computer Science, University of Vienna, Vienna, Austria
[2]UniVie Doctoral School Computer Science , University of Vienna , Vienna, Austria

A PRELIMINARIES FOR CAUSAL MODELING

A *Functional Causal Model* (FCM)[1] is a triple $\mathcal{M} \coloneqq \{\mathbf{X}_d, (\epsilon_d, P_{\epsilon_d}), f_d\}_d^D$ of a set of endogenous variables $\mathbf{X}$, a set of exogenous noise variables ϵ with joint probability distribution P_ϵ and a set of deterministic functions $\boldsymbol{f}$, one to generate each random endogenous variable X_d as a function of the other endogenous ones denoted as $\mathbf{X}_{\sim d}$ and its exogenous noise ε_d:

$$\forall d : \mathrm{X}_d \coloneqq f_d(\mathbf{X}_{\sim d}, \varepsilon_d)\,. \tag{4}$$

For this analysis we assume the absence of selection bias or latent confounders, i.e. causal sufficiency. The structure induced by the direct functional dependencies is often restricted to be acyclic such that it can be represented by a *Directed Acyclic Graph* (DAG) or equivalently by its adjacency matrix $\boldsymbol{G} \in \{0,1\}^{D\times D}$ with a one-to-one correspondence between random variables and nodes. The d-th column of $\boldsymbol{G}$ then encodes the parents $\mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_d)$ of a node/random variable X_d, i.e. the subset of $\mathbf{X}_{\sim d}$ that have a direct influence on X_d over f_d and an edge directed at X_d in the causal DAG $\boldsymbol{G}$. The task of *Causal Structure Learning* (CSL), also known as causal discovery, is to identify the underlying causal graph from observed data $\mathbf{X}$. When approximating the generally nonlinear functions $\boldsymbol{f}$ by some function model $\hat{\boldsymbol{f}}$ parameterized by $\boldsymbol{\theta}$, the adjacency matrix $\boldsymbol{G}$ can be used as a mask for $\mathbf{X}$:

$$\mathrm{X}_d \approx \hat{f}_d(\mathbf{G}\mathrm{X}, \varepsilon_d)\,. \tag{5}$$

B PARAMETER UNCERTAINTY

Due to the exogenous, random noise ϵ, a FCM is by definition inherently probabilistic. This aleatoric uncertainty can be captured by the observation model $p(\boldsymbol{X}|\boldsymbol{G})$. In addition, the finite size N of the observed data set $\mathbf{X}$ introduces epistemic uncertainty over the DAG $\mathbf{G}$ as well as over the underlying deterministic parameters of the structural functions. Denoting the random functional parameters (and their realizations) by $\boldsymbol{\Theta}$, the full generative model including all three sets of random variable, $\mathbf{G}$, $\mathbf{X}$ and $\boldsymbol{\Theta}$, is depicted in Figure 1a and reads

$$p(\boldsymbol{G}, \boldsymbol{\Theta}, \mathbf{X}) = p(\boldsymbol{G})\, p(\boldsymbol{\Theta}|\boldsymbol{G})\, p(\mathbf{X}|\boldsymbol{G}, \boldsymbol{\Theta})\,. \tag{6}$$

For its marginal probability over the causal graph $\mathbf{G}$ and observed data $\mathbf{X}$ introduced in Equation 1, the expectation value over $p(\boldsymbol{\Theta}|\boldsymbol{G})$ can be upper-bounded by its maximum likelihood estimate $\boldsymbol{\theta}^*_{\boldsymbol{G}}$:

$$\begin{aligned} p(\boldsymbol{G}, \mathbf{X}) &= \int p(\boldsymbol{G}, \boldsymbol{\Theta}, \mathbf{X})\ \mathrm{d}\boldsymbol{\Theta} && (7)\\ &= p(\boldsymbol{G}) \int p(\boldsymbol{\Theta}|\boldsymbol{G})\, p(\mathbf{X}|\boldsymbol{G}, \boldsymbol{\Theta})\ \mathrm{d}\boldsymbol{\Theta} && (8)\\ &\leq p(\boldsymbol{G})\ \ p_{\boldsymbol{\theta}^*_{\boldsymbol{G}}}(\mathbf{X}|\boldsymbol{G})\,, && (9)\\ &\text{where } \boldsymbol{\theta}^*_{\boldsymbol{G}} \coloneqq \arg\max_{\boldsymbol{\Theta}} p(\mathbf{X}|\boldsymbol{G}, \boldsymbol{\Theta})\,. \end{aligned}$$

[1]Also known as *Structural Causal Model* (SCM) or *Structural Equations Models* (SEM), we avoid the latter term, since equations are typically considered to be bidirectional.

Accepted for the 9th Causal Inference Workshop at the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024).

(a) Global model

(b) Local model

Figure 1: Graphical models for Bayesian CSL. The unobserved exogenous noise $\boldsymbol{\epsilon}$ is included for clarity, but typically modeled by the observational distribution in Equation 12. *(a)* At a global scale the influence of the causal graph $\mathbf{G}$ on parameters $\boldsymbol{\Theta}$ and observed random variables $\mathbf{X}$ is emphasized. *(b)* At a local scale the modularity of the FCM becomes evident. Here, $\mathrm{Pa}_{\boldsymbol{G}}$ denotes the parents of an observed random variable X_d according to some fixed causal graph $\boldsymbol{G}$.

Note that the parameters $\boldsymbol{\theta}^*_{\boldsymbol{G}}$ still depends on a particular graph $\boldsymbol{G}$, otherwise CSL is limited to point estimation. Hence, for such reduced model a different set of parameters governing the generation of some $\mathrm{X}_d \in \mathrm{X}$ has to be modeled for each graph. It is worth mentioning that even then only upper, but no lower bounds to the marginal probability $p(\boldsymbol{G}, \mathbf{X})$ and evidence $p(\mathbf{X})$ are obtained, yielding an over-confident estimation.

C OBSERVATION MODEL

The principle of independent causal mechanism Pearl [2009], Peters et al. [2017] —already incorporated by the definition of the FCM in Appendix A— motivates two standard assumptions on the distribution over the FCM parameters $\boldsymbol{\Theta}$ that allows to relax the dependence on the complete causal graph Heckerman et al. [1994], Friedman and Koller [2000].

(1) Parameter modularity states that the parameters for the structural function of an observed random variable $\boldsymbol{\Theta}_d$ depend only on their corresponding variable X_d, its exogenous noise ϵ_d and its parents $\mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_d)$, but not on any other observed random variables:

$$\boldsymbol{\Theta}_d \perp\!\!\!\perp \mathrm{X}_{\sim d} \mid \mathrm{X}_d, \epsilon_d, \mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_d)\,. \tag{10}$$

(2) Global parameter independence postulates that two sets of FCM parameters for different variables, $\boldsymbol{\Theta}_i$ and $\boldsymbol{\Theta}_j$, are independent given the parents of their corresponding observed random variables X_i and X_j:

$$\boldsymbol{\Theta}_i \perp\!\!\!\perp \boldsymbol{\Theta}_j \mid \mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_i) \cup \mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_j)\,. \tag{11}$$

Under these assumptions, the FCM parameters can be interpreted as mutually independent, exogenous noise variables similar to ϵ. Given global parameter independence (Equation 11) and parameter modularity (Equation 10), the observation model for Equation 1 (including parameter uncertainty) is depicted in Figure 1b and can be split dimensionwise into likelihood terms that are only coupled by the causal graph $\boldsymbol{G}$:

$$p(\boldsymbol{X}^{(n)} | \boldsymbol{G}, \boldsymbol{\Theta}) = \prod_{d=1}^{D} p\left(\mathrm{X}_d^{(n)} \mid \mathrm{Pa}_{\boldsymbol{G}}^{(n)}(\mathrm{X}_d), \boldsymbol{\Theta}_d\right). \tag{12}$$

Recall from Equation 1 that the joint distribution over the observed data set $\mathbf{X}$ factorizes as a product over independent samples $\boldsymbol{X}^{(n)}$. This illustrates that for large sample sizes the likelihood term $p(\mathbf{X}|\boldsymbol{G})$ dominates any prior $p(\boldsymbol{G})$ with support for all DAGs (c.f. paragraph on sparsity). In the absence of interventions, semi-parametric assumption on the structural functions of the FCM enable identification of the causal relations Shimizu et al. [2006], Hoyer et al. [2008], Zhang and Hyvärinen [2009], Loh and Bühlmann [2014], Peters et al. [2014]. In order to learn them by maximization of the likelihood, the observation model has to be correctly specified, i.e. for additive (post-)nonlinear models the distributions of the exogenous, but unobserved noise variables has to be known Bühlmann et al. [2014], Reisach et al. [2021], Kaiser and Sipos [2022], Seng et al. [2023]. Consequently, data normalization can shatter guarantees for structural identifiability of the causal graph which poses an open problem and active research direction Reisach et al. [2023].