Two-Stage Variance Approximation for Heteroscedastic Causal Discovery

Anonymous authors
Paper under double-blind review

Abstract

Learning causal structures from observational data is difficult when noise variances are unequal or depend on parent values (heteroscedasticity). We propose a two-stage framework that decouples structure learning from variance estimation. Instead of modeling full variance functions, we use a variance-matrix approximation: node-wise constant variances expanded across samples, refined by a small, bounded per-sample correction. We show that, under heteroscedastic causal models, the optimal constant surrogate equals the expected conditional variance, and the residual approximation error is a scale-invariant gap. We develop practical centralized and federated algorithms using stabilizers, including variance clipping and a progressive variance floor. Extensive empirical studies on both synthetic and real-world data show that our proposed approach discovers more plausible causal structures than competing baselines.

1 Introduction

Inferring cause–effect structure from observational data is a long-standing goal in fields ranging from biomedicine and climate science to economics and machine learning. A directed acyclic graph (DAG) provides a compact representation of causal relations, supporting tasks such as mechanistic interpretation and counterfactual reasoning (Upadhyaya et al., 2023; Piccininni et al., 2020; Kleinberg & Hripcsak, 2011; Huang et al., 2021; Ebert-Uphoff & Deng, 2012; Lopez et al., 2022). Over the years, numerous approaches for discovering a DAG from data have been developed. These include *constraint-based methods* that use conditional independence (CI) tests to prune and orient edges (Spirtes & Glymour, 1991; Spirtes et al., 2000), *score-based methods* that search the graph space for maximizing a score (Chickering, 2002; Huang et al., 2018), *order-based* methods that reduce the search to ordering permutations (Lin et al., 2024; Rolland et al., 2022; Montagna et al., 2023), and *continuous optimization methods* that relax the discrete DAG search into a smooth optimization (Zheng et al., 2018; Lachapelle et al., 2019; Ng et al., 2020). Each family comes with trade-offs in statistical efficiency, identifiability assumptions, and computational scalability.

A central modeling choice across all these methods is how to treat the *noise*. A majority of classical literature assumes homoscedasticity, i.e., noise variables with equal or constant variance, either explicitly or implicitly via the choice of loss function or regularization. In practice, however, real data seldom conform to constant noise levels (Merz et al., 2021; Kersting et al., 2007). In many domains, variability differs across variables (independent but non-identically distributed, i.n.i.d.) or depends on the covariate context, including the parent values of a node in the causal graph (a heteroscedastic regime). These departures from homoscedasticity matter both statistically and computationally. Statistically, they can create asymmetries that aid identification under functional-causal assumptions (Shimizu et al., 2006; Zhang et al., 2012; Yin et al., 2024; Kikuchi, 2022). Computationally, heteroscedasticity complicates the optimization landscape by coupling residual magnitudes with variance parameters inside the likelihood, increasing non-convexity and sensitivity to initialization.

Despite these challenges, we focus on optimization-based DAG learning methods for their attractive properties. Optimization-based approaches seamlessly integrate with modern machine learning toolchains, enabling gradient-based updates and automatic differentiation, and tend to scale better to high-dimensional problems

than discrete searches. Their flexibility allows incorporating additional modeling assumptions (nonlinear mechanisms, sparse or ordering priors, etc.) within a unified optimization framework (Zheng et al., 2018; Lachapelle et al., 2019; Kyono et al., 2020). For these reasons, continuous optimization has become a central paradigm in causal discovery, especially in domains where data are large, distributed, or heterogeneous. However, their performance hinges on how well the underlying statistical assumptions (about noise) align with reality and with the optimization procedure. When the noise variance is also treated as a learnable function (as needed for fully heteroscedastic), the objective can become poorly conditioned. In particular, a flexible variance network introduces directions in the loss landscape that can drive the objective arbitrarily low by shrinking variances to zero in tandem with residuals. This is exacerbated in small-sample regimes or if the variance function class is misspecified, leading to unstable training.

In privacy-sensitive domains like health informatics, data are distributed across multiple institutions (clients) that cannot share raw data, and requires federated learning (FL) to estimate a global causal graph by coordinating local computations. FL brings an additional layer of complexity and constraint to the causal discovery problem (Gao et al., 2023; Ng & Zhang, 2022; Li et al., 2024; Ye et al., 2024; Wang et al., 2025b). The setting introduces heterogeneous data distributions (each client may have a different data distribution, even different variable sets) and limits the amount of information exchanged. Variance modeling under FL is especially brittle: a high-capacity variance model not only requires ample data to avoid overfitting but also increases communication overhead and could destabilize global optimization if client updates diverge due to local noise idiosyncrasies.

We argue that, in many practical settings, the bottleneck is the cost and instability of learning a full parent-to-variance mapping. Rather than directly modeling each node's conditional variance function with high complexity, we propose to focus on the quantity that a continuous DAG optimizer actually needs: the sample–variable variance matrix that weights residuals in the likelihood. At a high level, we treat variance modeling as a form of regularization and conditioning for structure learning objective. The variance proxy should be statistically reasonable, simple to estimate, and stable enough to support the DAG optimization.

Concretely, we study differentiable DAG learning under two common heterogeneous-noise regimes: (a) an *i.n.i.d. noise model* where noice variances differ across nodes but are constant for any given, and (2) a fully *heteroscedastic noise model* where each node's variance can vary with its parents. Our goal is to clarify *when* a sample variance surrogates suffices and *how* they can be incorporated into existing DAG learning frameworks. We introduce a two-stage approach that first learns the causal structure assuming fixed per-node variance surrogates, then refines the variance estimates – alternating. We develop a novel *variance-matrix approximation* strategy that provides such surrogates with provable properties, and we tailor it to both centralized and federated settings.

Contributions. (1) We propose and motivate the use of node-wise constant variance proxies with lightweight correction as surrogates for heteroscedastic noise. This approach drastically simplifies variance estimation in the DAG learning task.

(2) We provide a theoretical characterization of the approximation gap under heteroscedastic models, showing that the best constant surrogate for a node is its expected conditional variance, and that the remaining variance-only error is a scale-invariant difference between arithmetic and geometric means.

(3) Building on this idea, we design a two-stage DAG learning procedure (centralized and federated) that integrates these variance surrogates. Our framework leverages existing differentiable DAG learners for the structure update (Stage I) and introduces a new Stage II for variance re-estimation via closed-form updates and minor neural adjustments. We incorporate practical stabilizers, including variance clipping and a gradually tightening lower bound on variances, to avoid the numerical pathologies that arise when naively learning variances jointly with structure.

(4) Empirically, we conduct extensive experiments on both synthetic and real-world data to demonstrate that our approach achieves state-of-the-art performance under heteroscedastic noise, while maintaining competitive performance under homoscedastic settings.

Related Work. Causal structure learning spans constraint-based, score-based, and continuous optimization approaches, with NOTEARS and its extensions popularizing differentiable acyclicity. However, most

assume equal-variance noise, which can mislead in practice. Recent studies exploit heteroscedasticity for identifiability, and federated frameworks address distributed settings but still mainly rely on homoscedastic SEMs. A full review is provided in Appendix A.1.

2 Preliminaries

We consider a structural causal model over a set of d observed random variables $X = (X_1, X_2, \cdots, X_d) \in \mathbb{R}^d$ associated with nodes in a directed acyclic graph (DAG) $\mathcal{G}$. Let $\mathbf{B}^\star \in \{0,1\}^{d\times d}$ denote the ground-truth adjacency matrix of $\mathcal{G}$. Each variable is generated by structural function of its parents plus a noise term:

1) In the **independent but non-identically distributed (i.n.i.d.) noise model**, often called a linear SEM with unequal variances, we have for node j:

$$X_j = f_j(\mathrm{pa}(X_j)) + E_j, \quad j = 1, \cdots, d,$$

where $\mathrm{pa}(X_j)$ denotes the set of parent variables of X_j in $\mathcal{G}$. The noise terms E_j are mutually independent, mean zero, and have variances that may *differ by node*. The classical homoscedastic model is a special case where $\mathrm{Var}(E_1) = \mathrm{Var}(E_2) = \cdots = \mathrm{Var}(E_d)$; here we allow $\mathrm{Var}(E_j) = \sigma_j^2$ to be distinct for each j. These variances are constant per node, independent of parent values, but heterogeneous across nodes.

2) In the **heteroscedastic causal model (HCM)**, the noise variance can vary with the parent values (Immer et al., 2023; Duong & Nguyen, 2023):

$$X_j = f_j(\mathrm{pa}(X_j)) + g_j(\mathrm{pa}(X_j))Z_j,\ j = 1, \cdots, d, \tag{1}$$

where Z_j are mutually independent random variables, $f_j(\cdot)$ and $g_j(\cdot)$ are the underlying structural causal function and noise function (conditional standard deviation function) with $g_j(\cdot) > 0$. Thus, considering Z_j are standard normal variables, the conditional distribution under HCM is $p(X_j|\mathrm{pa}(X_j)) \sim \mathcal{N}\left(f_j(\mathrm{pa}(X_j)), g_j(\mathrm{pa}(X_j))^2\right)$. equation 1 is sometimes referred to as a Heteroscedastic Noise Model (HNM) in the literature (Strobl & Lasko, 2023) when $g_j(\cdot)$ models the conditional mean absolute deviation.

Under either model, after drawing n samples we have a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. The data is input to our structure learning algorithm. The presence of heteroscedasticity introduces an instance-specific variance matrix $\mathbf{\Sigma}^\star \in \mathbb{R}^{n\times d}$, where each entry $\mathbf{\Sigma}^\star_{i,j}$ is the variance of X_j in sample i. In an i.n.i.d. noise model, $\mathbf{\Sigma}^\star$ has constant columns (all entries in column j equal σ_j^2). Under a fully heteroscedastic model, the columns of $\mathbf{\Sigma}^\star$ can vary row by row according to $g_j(\mathrm{pa}(X_{i,j}))^2$. We do not observe $\mathbf{\Sigma}^\star$, but our goal is to learn the graph structure $\mathbf{B}^*$ and ideally also form some estimate of the variance structure.

Differentiable DAG Learning Objective. We adopt a continuous optimization approach to learn the DAG. We parameterize each variable's conditional expectation $f_j(\mathrm{pa}(X_j))$ by a neural network (or linear model) $F_j(\cdot; \mathbf{B}, \mathbf{W}_F)$, where $\mathbf{B}$ is an adjacency matrix with real-valued weights during optimization and $\mathbf{W}_F$ denotes the parameters of the functions. Similarly, for the heteroscedastic case, we can introduce a network $G_j(\cdot; B, W_G)$ to represent the variance function $g_j(\cdot)$, with parameters $\mathbf{W}_G$. In practice, $\mathbf{B}$ will be constrained to represent an acyclic graph (via a differentiable acyclicity penalty), and $\mathbf{W}_F, \mathbf{W}_G$ adjust the specific functional mappings. For a given sample i, let $r_{i,j} = \mathbf{X}_{i,j} - F_j(\mathbf{X}_i; \mathbf{B}, \mathbf{W}_F)$ be the residual for variable j, and let $\sigma_{ij}^2 := G_j(\mathbf{X}_i; \mathbf{B}, \mathbf{W}_G)$ be the predicted noise variance for $\mathbf{X}_{i,j}$ (with $G_j(\cdot)$ outputting a positive value). If we assume the noise Z_j in (2) is Gaussian, then the log-likelihood of sample i (up to constants) includes a term $\frac{1}{2}\left(r_{ij}^2/\sigma_{ij}^2 + \log \sigma_{ij}^2\right)$ for each variable j. Summing over all samples and variables, the heteroscedastic Gaussian negative log-likelihood (hNLL) objective can be written as:

$$\min_{\mathbf{B}, \mathbf{W}_F, \mathbf{W}_G} \mathcal{L}_{\mathrm{hNLL}}(\mathbf{X}) = \sum_i^n \sum_j^d \left(\frac{1}{2\sigma_{ij}^2} r_{ij}^2 + \frac{1}{2} \log \sigma_{ij}^2 \right) \text{ s.t. } h(\mathbf{B}) = 0. \tag{2}$$

Here $h(\mathbf{B}) = 0$ is the differentiable acyclicity constraint of Zheng et al. (2018), with $h(\mathbf{B}) = \mathrm{tr}(e^{\mathbf{B}\circ\mathbf{B}}) - d = 0$ ensuring $\mathbf{B}$ (viewed as weighted adjacency) has no cycles. In practice, one augments the objective with a penalty or Lagrange multipliers to enforce $h(\mathbf{B}) \approx 0$.

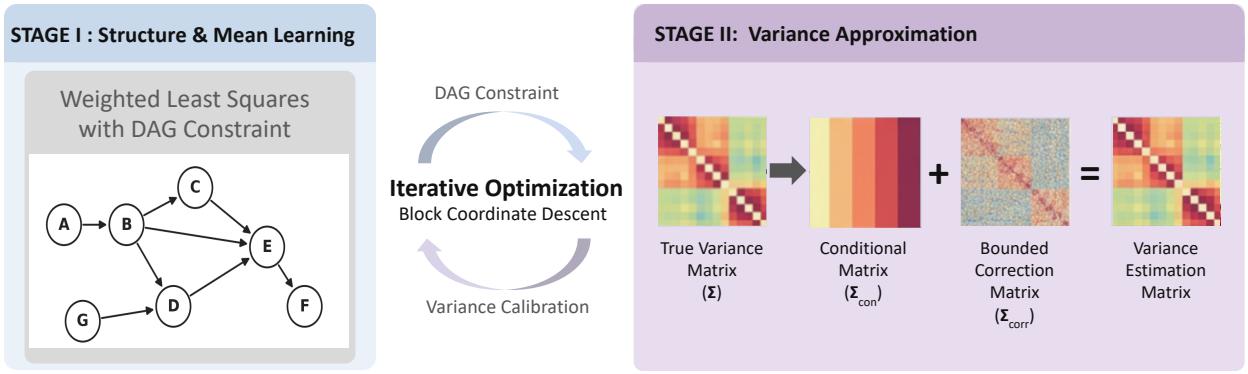


Figure 1: two-stage causal Structure Learning with variance Matrix Approximation (SL-MA): Alternating two-stage optimization for causal structure learning, where structure and mean estimation under fixed variance is coupled with a stable variance approximation that decomposes the noise matrix into a node-wise constant component and a bounded heteroscedastic correction.

Challenges of Heteroscedastic DAG Learning. Directly optimizing hNLL is challenging for two main reasons: (a). The residual terms r_{ij}^2 and the variance terms σ_{ij}^2 are coupled inside the summation. This coupling makes the objective highly non-convex and unstable. (b). Approximating a complex g_j with a finite-capacity model G_j can cause misspecification: the learned σ_{ij}^2 might have an incorrect range or functional form, complicating optimization under limited data. In FL settings, these issues are amplified by the fact that each client has fewer samples, and the variance models might overfit or diverge across clients.

In summary, heteroscedastic noise offers new information for causal discovery but also poses new optimization difficulties. We address these issues by breaking the problem into two stages and introducing a well-chosen variance surrogate. The key idea is to separate the learning of structure (the f_j functions and $\mathbf{B}$) from the learning of noise variances (g_j or σ values), and to constrain the latter with a simplified approximation that is easier to estimate and optimize.

3 Proposed Method

While it is in principle feasible to jointly optimize the residual and variance terms, direct gradient-based training on the full heteroscedastic objective often suffers from severe numerical instabilities (e.g., driving $\sigma_{ij}^2 \to r_{ij}^2 \to 0$). A detailed discussion and illustrative example are provided in Appendix A.2. In practice, overly flexible variance models may exploit the likelihood objective in unintended ways, particularly under finite-sample conditions. To mitigate these issues, we adopt a two-stage decomposition strategy and incorporate explicit safeguards in variance estimation. An overview of the framework is illustrated in Figure 1.

3.1 Centralized Two-Stage Prototype

We first present our approach in the centralized setting and then extend it to the federated case, where the decentralized formulation arises naturally. Particularly, we approximate the instance-wise variance matrix as a node-wise constant component plus a bounded, lightweight correction that captures heteroscedastic variation without introducing variance-collapse instabilities.

The method can be interpreted as an iterative block coordinate descent (Tseng, 2001; Bertsekas & Tsitsiklis, 2015) over two parameter blocks: (i) the structure and mean parameters $(\mathbf{B}, \mathbf{W}_F)$, and (ii) the variance parameters, which we approximate with a surrogate. Accordingly, the learning procedure is decomposed into two interleaved stages:

Stage I (Structure learning under fixed variance): In this stage, we estimate the causal structure by learning the DAG $\mathbf{B}$ and conditional mean functions F_j, assuming fixed variance values $\tilde{\sigma}_j^2$ for each node. This reduces to DAG learning with homoscedastic noise, essentially a weighted least-squares problem subject

to an acyclicity constraint (since minimizing $\sum_{i,j} \frac{1}{2\tilde{\sigma}_j^2} r_{ij}^2 + \text{const}$ corresponds to weighted least squares). Existing differentiable DAG learning methods (e.g., NOTEARS, GOLEM) can be directly applied with a loss reweighted by $\tilde{\sigma}_j^2$.

Stage II (Variance learning given structure): For this stage, we hold the structure $\mathbf{B}$ and $\mathbf{W}_F$ fixed and update our estimates of the noise variances for each node. In equation 2, with $\mathbf{B}, \mathbf{W}_F$ fixed, the optimal σ_j^2 (under the assumption that it is constant across i for node j) corresponds to the average of r_{ij}^2 over i. Accordingly, Stage II will set $\tilde{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^n r_{ij}^2$ for each j. We will derive this more formally later.

The two stages are alternated iteratively: Stage II updates the variance estimates, Stage I re-estimates the structure with the new weights, and this process repeats until convergence.

Challenges: The design of Stage II has two challenges:

1) Modeling challenges. The true noise functions $g_j(\cdot)$ are unknown and potentially highly complex. Approximating them with neural networks, as in prior work, poses a risk of misspecification: the learned $\hat{g}_j$ may produce unbounded outputs or incorrect functional forms, and a highly flexible $\hat{g}_j$ could overfit random fluctuations in limited data. This simplification avoids directly fitting complex functions, but it raises the open question of how to reintroduce controlled flexibility to capture heteroscedasticity without restoring full model complexity.

2) Optimization challenge. The two-stage schedule alone, while avoiding simultaneous updates, does not automatically fix all issues. The space of variance parameters needs some constraints (like $\sigma_{ij} \geq \epsilon$ or a regularization) to avoid the trivial collapse. However, it also faces the challenges of the misspecified variance range. Another optimization consideration is the interaction between stages: even if each stage individually guarantees a non-increasing loss, the combined process could get stuck in a suboptimal configuration.

Given these challenges, we pose the guiding question: *Can we learn (or closely approximate) the truth variance matrix $\mathbf{\Sigma}^\star$ in a way that is friendly to both modeling and optimization, and with quantifiable approximation error?*

3.2 Two-Stage Causal Structure Learning with Variance Matrix Approximation

What does variance matrix $\mathbf{\Sigma}^\star$ look like? Given data $\mathbf{X}$, the entry $\mathbf{\Sigma}^\star_{ij} = g_j(\text{pa}(\mathbf{X}_{ij}))$. Due to the independence of each row and column, i.e., independence between $\{g_j\}$ and no specific assumption regarding each g_j, we cannot expect either a low rank matrix nor a matrix with other common structures.

The only prior knowledge available for estimating or constructing $\mathbf{\Sigma}^\star$ is that columns of the matrix are independent (by assumptions in the i.n.i.d. noise model and HCM), while entries within each column share a common functional dependence. For instance, if the noise magnitude at the j-th node is parameterized by a sigmoid function of its parent values, the variances within that column are jointly shaped by the shrinking behavior of the sigmoid function, then the entries in that column are expected to vary in a coordinated way due to the function's shrinking effect.

Beyond these observations, $\mathbf{\Sigma}^\star$ could be very complex if g_j are complex. Our approach is to *approximate* $\mathbf{\Sigma}^\star$ *with the sum of two structured matrices*: **(a) A conditioning matrix $\mathbf{\Sigma}^{\text{c}} \in \mathbb{R}^{n\times d}$**, constructed under the i.n.i.d. noise assumption, where each node is assigned a constant variance. Concretely, all entries within a given column share the same value, and the true underlying information reduces to a vector $\mathbf{\Sigma}^{\text{c}} \in \mathbb{R}^d$. **(b) A correction matrix $\mathbf{\Sigma}^{\text{h}} \in \mathbb{R}^{n\times d}$**, compensates row-wise heterogeneity. It is an optional refinement and provides a practical way to compensate for the fact that the "pooled distribution" is heavier-tailed than a single Gaussian with the same variance.

3.2.1 Variance Matrix Approximation

This subsection illustrates the strategies for approximating the above two matrices. We begin with the approximation in a centralized manner, then extend it to the FL setting.

A conditioning matrix. By assuming node-wise constant variances. The objective reduces to

$$\min_{\mathbf{B},\mathbf{W}_F,\{\sigma_j\}_{j\in[d]}} \mathcal{L}_{\text{hNLL}}(\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{d}\left(\frac{1}{2\sigma_j^2}r_{ij}^2 + \frac{1}{2}\log\sigma_j^2\right), \text{ s.t. } h(\mathbf{B}) = 0,$$

where $r_{ij} := \mathbf{X}_{i,j} - F_j(\mathbf{X}_i; \mathbf{B}, \mathbf{W}_F)$.

Regarding the variance learning of Stage-II, the maximum likelihood estimation for each variance admits a closed form: $\tilde{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} r_{ij}^2$ (by taking the gradient of $\mathcal{L}_{\text{hNLL}}$ w.r.t. $\{\sigma_j\}_{j\in[d]}$ and setting it as zero, we obtain the closed-form solution). Accordingly, Stage-II constructs $\mathbf{\Sigma}^{\text{c}}$ by vertically expanding the vector:

$$\arg\min_{\{\sigma_j\}_{j\in[d]}} \mathcal{L}_{\text{hNLL}}(\mathbf{X}; \mathbf{B}, \mathbf{W}_F) = \left[\frac{1}{n}\sum_{i}^{n} r_{i,1}^2, \cdots, \frac{1}{n}\sum_{i}^{n} r_{i,d}^2\right]^\top, \tag{3}$$

where $\mathbf{B}, \mathbf{W}_F$ are fixed. And, $\widetilde{\mathbf{\Sigma}}^{\text{c}} = \text{Expand}([\widetilde{\sigma}_1^2, \cdots, \widetilde{\sigma}_d^2]^\top)$ where $\text{Expand}(\cdot): \mathbb{R}^{1\times d} \to \mathbb{R}^{n\times d}$ replicates the row vector across n samples.

A correction matrix. While the constant-variance estimator $\mathbf{\Sigma}^{\text{c}}$ provides a simple and convex closed-form solution, it may insufficiently capture row-wise heterogeneity in the true variance matrix $\mathbf{\Sigma}^\star$. To allow for flexible refinement, we introduce an additive correction mechanism that encourages adjustments at Stage-II:

$$\min_{\mathbf{W}_H} \mathcal{L}_{\text{hNLL}}(\mathbf{X}; \mathbf{B}, \mathbf{W}_F, \widetilde{\mathbf{\Sigma}}^{\text{c}}) = \sum_{i=1}^{n}\sum_{j=1}^{d}\Big(\frac{1}{2(\widetilde{\mathbf{\Sigma}}_{ij}^{\text{c}} + h_j(\mathbf{X}_i \odot \mathbf{B}_j))} r_{ij}^2 + \frac{1}{2}\log(\widetilde{\mathbf{\Sigma}}_{ij}^{\text{c}} + h_j(\mathbf{X}_i \odot \mathbf{B}_j))\Big), \tag{4}$$

where $H(\cdot) := \{h_j(\cdot)\}_{j\in[d]}$ denotes a set of node-wise functions, parameterized by $\mathbf{W}_H$, and the term $\mathbf{X}_i \odot \mathbf{B}_j$ represents the parent values of j-th node at i-th observation. Thus, the correction matrix $\mathbf{\Sigma}^{\text{h}}$ is defined over $\{h_j(\cdot)\}_{j\in[d]}$ with corresponding data and adjacency matrix. It is worth noting that to ensure a positive output range of the final variance, i.e., $\widetilde{\mathbf{\Sigma}}_{ij}^{\text{c}} + h_j(\mathbf{X}_i \odot \mathbf{B}_j)$, the output range of $\{h_j(\cdot)\}_{j\in[d]}$ should be constrained. For instance,

$$\widetilde{\mathbf{\Sigma}}_{ij}^{\text{h}} \equiv h_j(\cdot) = \rho\widetilde{\mathbf{\Sigma}}_{ij}^{\text{c}}\text{Tanh}\left(\bar{h}_j(\mathbf{X}_i \odot \mathbf{B}_j)\right), \tag{5}$$

where ρ is a predefined scalar.

3.2.2 Two-stage Integration

With Stage-I (general causal structure learning under fixed variance) and the proposed Stage-II (variance matrix learning) in place, we design strategies to jointly incorporate both stages. Specifically, we introduce: (i) a variance lower-bound annealing strategy, progressively decreasing the variance floor during training to stabilize optimization; and (ii) a stopping criterion based on $\mathcal{L}_{\text{NLL}}(\cdot)$ rather than $\mathcal{L}_{\text{hNLL}}(\cdot)$, motivated by the fact that $\mathcal{L}_{\text{hNLL}}(\cdot)$ may decrease arbitrarily by neglecting the role of variance, whereas $\mathcal{L}_{\text{NLL}}(\cdot)$ avoids this issue. Algorithm 1 outlines the resulting procedure.

3.2.3 Extension to Federated Learning

In a federated scenario, we have K clients each holding local data $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k\times d}$. We assume all clients observe the same set of variables $\mathbf{X}_1, \ldots, X_d$. Our goal is to learn a single global DAG that explains data across all clients. We adopt a synchronous federated optimization model akin to FedAvg for continuous structure learning (Gao et al., 2023; Ng & Zhang, 2022): clients compute local updates, which are then aggregated by a server to update global parameters.

We extend our two-stage design to the federated setting: in *Stage I*, clients perform local structure learning under a shared variance matrix and the server aggregates global parameters; in *Stage II*, clients estimate local variances and corrections, which are then combined to update the global variance model. This iterative procedure yields synchronized updates of both structure and variance across clients. Full algorithmic details are provided in Appendix A.3 (Algorithm 2).

Algorithm 1 two-stage causal Structure Learning with variance Matrix Approximation (SL-MA)

Input: data $\mathbf{X} \in \mathbb{R}^{n \times d}$, function $F(; \mathbf{B}, \mathbf{W}_F)$, tolerance ϵ (1e-3), max iteration T (20); scalar ρ (0.5), variance bound $\underline{\sigma}^2, \overline{\sigma}^2$ (1e-2, 1), decay strategy (linear decay).
Initialize: $\widetilde{\mathbf{\Sigma}} = \overline{\sigma}^2 \mathbf{1}_{n \times d}$ and $t = 0$.
while $|\ell_t - \ell_{t-1}| < \epsilon$ and $t < T$ **do**
    ## Stage-I:
    Initialize training hyper-parameters
    $\widetilde{\mathbf{B}}, \widetilde{\mathbf{W}}_F \leftarrow$ Structure learning under fixed: $\widetilde{\mathbf{\Sigma}}$
    ## Stage-II:
    $\widetilde{\mathbf{\Sigma}}^c \xleftarrow{\text{equation 3}} \widetilde{\mathbf{B}}, \widetilde{\mathbf{W}}_F$
    $\text{var}_{\min} = \text{Decay strategy}(t+1, T, \underline{\sigma}^2, \overline{\sigma}^2)$
    $\widetilde{\mathbf{\Sigma}}^c = \text{clip}(\widetilde{\mathbf{\Sigma}}^c, \text{var}_{\min}, \overline{\sigma}^2)$
    $\widetilde{\mathbf{\Sigma}}^h \xleftarrow{\text{equation 5}} \widetilde{H}(\cdot) \xleftarrow{\text{equation 4}} \widetilde{\mathbf{B}}, \widetilde{\mathbf{W}}_F, \widetilde{\mathbf{\Sigma}}^c$
    $\overline{\sigma}^2 = \max(\widetilde{\mathbf{\Sigma}}^c)$
    $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Sigma}}^c + \widetilde{\mathbf{\Sigma}}^h$
    ## For stopping mechanism:
    $\ell_t = \mathcal{L}_{\text{NLL}}(\mathbf{X}; \widetilde{\mathbf{B}}, \widetilde{\mathbf{W}}_F)$
    t = t + 1
end whilereturn $\widetilde{\mathbf{B}}$ corresponding to lowest ℓ_t.

4 Analyses of Noise-Wise Constant Variances Approximation

Approximation Error. We present Proposition 4 to quantitatively characterize the approximation error in the training objective that arises when the conditioning matrix $\mathbf{\Sigma}^c$ is used to approximate the ground-truth matrix $\mathbf{\Sigma}^\star$. In particular, the quantity ξ_j^{var} defined in Proposition 4 measures the node-wise contribution of variance misspecification within the KL divergence term $\mathbb{E}\left[-\log p_{\text{app}(X)}\right] = \mathbb{E}\left[-\log p_{\text{true}(X)}\right] + \text{KL}(p_{\text{true}} \| p_{\text{app}})$. Proposition 4 shows that the node-wise error ξ_j^{var} corresponds to the logarithmic AM–GM gap of conditional variances, which depends on relative fluctuations and is invariant to global noise scaling.

Under the heteroscedastic model equation 1: (i) The optimal constant variance is $(\sigma_j^2)^\star = \mathbb{E}[g_j(\text{pa}(X_j))^2]$. (ii) The irreducible variance gap is

$$\xi_j^{\text{var}} = \tfrac{1}{2}\left(\log \mathbb{E}[g_j(\text{pa}(X_j))^2] - \mathbb{E}[\log g_j(\text{pa}(X_j))^2]\right),$$

which vanishes iff $g_j(\text{pa}(X_j))^2$ is constant.

Mitigating hNLL Pathologies. Direct heteroscedastic NLL optimization is unstable since mean and variance networks can jointly drive $r_{ij}^2, \sigma_{ij}^2 \to 0$, yielding degenerate solutions (Section 3, Appendix A.2). We stabilize training by: (i) *Node-wise initialization*: closed-form variance surrogates $\widetilde{\sigma}_j^2 = \frac{1}{n}\sum_i r_{ij}^2$; (ii) *variance lower-bound annealing strategy*: a decaying lower bound preventing collapse; (iii) *Lightweight correction*: $\mathbf{\Sigma}^{\text{h}}$ introduces limited flexibility for row-wise heterogeneity. These elements make hNLL optimization stable while preserving heteroscedastic benefits.

5 Experiments

5.1 Setup

Noise Mode. We focus on linear structural causal functions f_j and evaluate robustness primarily under different specifications of the noise function g_j (conditional standard deviation).

To evaluate robustness under diverse heteroscedastic regimes, we consider four canonical noise settings. These settings differ in whether the noise variance is constant across all variables, node-specific but fixed, or dynamically dependent on the parent nodes.

1. **Equal Variance (EV).** All variables share the same homoscedastic Gaussian noise: $E_{ij} \sim \mathcal{N}(0, \sigma^2), \quad \forall i, j$. This is the simplest regime and serves as the classical benchmark for methods that assume homoscedasticity.

2. **Node-wise Variances (NV).** Each node is assigned its own variance, drawn from Uniform$(0.1, 1)$, which is held constant across samples: $E_{ij} \sim \mathcal{N}(0, \sigma_j^2), \quad \sigma_j \sim \text{Uniform}(0.1, 1)$. This setting breaks the equal-variance assumption while preserving per-node homoscedasticity. It corresponds to the classical i.n.i.d. model and provides a controlled intermediate between EV and fully input-dependent heteroscedastic noise.

3. **HCM with *sigmoid* (Lin et al., 2024; Duong & Nguyen, 2023) .** The variance of each node is a deterministic function of its parents, following the Gaussian variance parameterization framework (Duong & Nguyen, 2023):

$$g_j(\text{pa}(X_{:,j})) := \exp\left(\tfrac{1}{2}\log\left(\tfrac{1}{2}\exp\left(\sum_{p\in\text{pa}(X_{:,j})} m_j(X_{:,p})\right)\right)\right).$$

   In the sigmoid case, $m_j(x) = 1/(1 + e^{-x})$, yielding smooth bounded noise scaling as a nonlinear function of parent inputs.

4. **HCM with *mixture* (Duong & Nguyen, 2023) .** To model more diverse heteroscedasticity, we extend the above by randomly sampling m_j from a function family $\mathcal{F} = \{\text{linear}, \sin, \log, \text{sigmoid}, \text{square}\}$. Each function captures qualitatively different nonlinear transformations:

   - *Linear*: normalized affine maps with random coefficients, introducing directionality;
   - *Sin*: $\sin(2\pi x)$, producing periodic variance fluctuations;
   - *Log*: $\log(1 + x - \min(x))$, compressing heavy-tailed inputs;
   - *Sigmoid*: bounded nonlinear mapping, similar to saturating activations in neural nets;
   - *Square*: variance amplification via quadratic scaling.

   At each evaluation, one function is chosen uniformly at random and applied to the parents before aggregation and positivity transformation. This stochastic construction induces a diverse family of noise functions, mimicking realistic environments where the heteroscedastic mechanism itself may switch across functional forms.

Although the positivity transformation originates from Gaussian variance parameterization, its role is more general: it enforces positivity of the scale while allowing input-dependent mappings. Thus, EV and NV represent homoscedastic baselines (global vs. node-specific), whereas the two HCM variants represent genuinely heteroscedastic regimes of increasing complexity.

Data and Metrics. We evaluate our method on both synthetic and real-world datasets, including a clinical Alzheimer's disease case study (Section 6). For the synthetic setting, ground-truth DAGs are generated from Erdős–Rényi (ER) graphs ERDdS & R&wi (1959) with expected degree 2 and $d = 20$, and observations are sampled from corresponding linear SEMs under different noise regimes. We assess performance in centralized and federated settings, with main results reported in Table 1 and Figures 2–3, and details in Appendix B.

We report standard metrics Chickering (2002); Zheng et al. (2018): *Structural Hamming Distance (SHD)*, which counts edge differences from the ground truth, *SHD at the Completed Partially Directed Acyclic Graphs level (SHD-CPDAG)*, and *True Positive Rate (TPR)*, measuring the proportion of correctly recovered edges.

Baselines. We compare against representative approaches for causal structure learning: (a) Continuous optimization methods NOTEARS Zheng et al. (2018) and GOLEM Ng et al. (2020). (b) Ordering-based heuristic VarSort Reisach et al. (2021), which infers causal orderings from marginal variances. (c) Constraint-based HOST Duong & Nguyen (2023), which adapts conditional independence testing to heteroscedastic settings; and (d) ICDH Yin et al. (2024), a two-stage framework alternating between noise-function approximation

and structure update. We reimplement ICDH under linear SEMs with neural approximators, including a global MLP variant (ICDH-MLP) and a node-specific variant (ICDH-Nodewise).

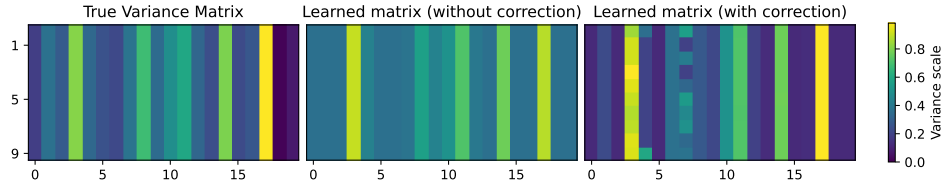
Proposed Method. Building on NOTEARS as the Stage I backbone, we introduce a second stage with either (i) a conditioning matrix (Proposed$^{[-]}$) or (ii) conditioning plus a correction matrix (Proposed$^{[+]}$), where corrections $\bar{h}_j(\cdot)$ are parameterized by node-specific MLPs. This addresses noise misspecification while preserving interpretability, with full details in Section 3. In federated settings, we also compare against NOTEARS-ADMM Ng & Zhang (2022) and FedDAG Gao et al. (2023), which extend acyclicity-constrained learning to distributed environments.

Table 1: Centralized evaluation under linear structural causal function with varying sample sizes (n) and noise modes (degree 2, $d = 20$). SHD: Structural Hamming Distance (lower is better). TPR: True Positive Rate (%). Particularly, $^{[-]}$ and $^{[+]}$ indicate the proposed method, Algorithm 1, without and with the correction matrix, respectively.

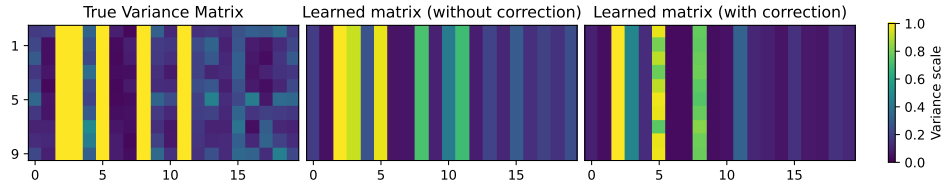
Noise mode	Method	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
		SHD	TPR	SHD	TPR	SHD	TPR	SHD	TPR
EV	NOTEARS	7±0	88.3±2.4	8±3	86.7±3.1	5±2	90.8±4.2	5±2	90.0±4.1
	GOLEM-EV	5±3	88.3±4.2	7±3	86.7±3.1	3±0	92.5±0.0	6±3	88.3±4.2
	HOST	25±4	92.5±3.5	10±7	94.2±3.1	4±4	96.7±3.1	3±4	96.7±3.1
	VarSort	40±6	93.3±1.2	23±6	91.7±3.1	13±2	92.5±2.0	9±4	94.2±1.2
	ICDH Nodewise MLP	10±2	89.2±2.4	5±2	90.8±3.1	4±0	92.5±2.0	5±2	92.5±3.5
	ICDH MLP	24±2	69.2±7.2	19±2	68.3±6.2	9±2	83.3±2.4	13±5	76.7±6.2
	Proposed method$^{[-]}$	6±0	89.2±1.2	8±3	87.5±2.0	4±3	91.7±4.7	4±2	91.7±4.2
	Proposed method$^{[+]}$	8±0	88.3±3.1	8±3	86.7±3.1	6±2	90.8±4.2	5±2	90.8±3.1
NV	NOTEARS	27±3	50.0±7.4	24±3	54.2±8.2	21±5	60.8±10.3	21±4	58.3±9.6
	GOLEM-NV	30±9	42.5±14.7	26±10	46.7±15.5	28±8	44.2±11.8	23±8	50.8±15.0
	HOST	30±5	39.2±15.5	27±3	46.7±9.6	25±4	46.7±13.3	24±4	47.5±15.4
	VarSort	36±5	39.2±6.6	27±12	49.2±17.1	24±6	54.2±8.5	26±4	50.0±8.2
	ICDH Nodewise MLP	26±5	56.7±6.6	21±4	62.5±10.6	25±2	58.3±5.9	21±4	63.3±11.2
	ICDH MLP	22±4	74.2±5.1	19±20	66.7±8.5	14±4	75.8±5.9	14±1	75.8±4.7
	Proposed method$^{[-]}$	17±5	66.7±8.2	13±0	77.5±5.4	13±6	76.7±10.1	17±4	70.8±4.2
	Proposed method$^{[+]}$	19±7	65.0±12.7	15±4	73.3±7.7	11±2	76.7±4.2	12±2	75.0±6.1
g_j with sigmoid	NOTEARS	24±2	60.8±6.6	24±4	61.7±8.2	28±2	55.8±5.1	29±5	56.7±3.1
	GOLEM-NV	27±7	45.8±18.5	30±6	45.8±11.2	24±10	58.3±18.3	25±4	57.5±10.8
	HOST	28±7	49.2±13.6	24±4	55.8±19.6	27±4	50.8±15.5	26±4	53.3±12.5
	VarSort	29±11	48.3±14.3	26±6	51.7±8.2	31±5	45.8±6.6	24±0	53.3±2.4
	ICDH Nodewise MLP	32±2	58.3±3.1	23±4	65.0±2.0	26±4	60.8±10.1	25±7	63.3±10.1
	ICDH MLP	30±6	65.0±12.2	33±4	58.3±14.3	25±7	70.8±2.4	20±9	75.8±8.2
	Proposed method$^{[-]}$	7±1	90.0±3.5	11±7	85.8±8.5	9±7	89.2±8.2	6±6	90.0±7.4
	Proposed method$^{[+]}$	11±4	90.0±4.1	4±3	93.3±2.4	5±4	94.2±5.1	3±1	94.2±3.1
g_j with mixture	NOTEARS	26±3	61.7±4.2	26±6	57.5±5.4	17±6	74.2±9.6	20±0	69.2±5.9
	GOLEM-NV	29±5	43.3±6.6	31±1	45.8±9.4	29±6	46.7±17.0	27±7	46.7±13.3
	HOST	29±7	44.2±9.6	26±2	49.2±10.1	27±4	52.5±15.9	26±5	53.3±13.6
	VarSort	36±1	39.2±2.4	37±7	38.3±8.2	23±7	55.8±9.6	21±1	58.3±1.2
	ICDH Nodewise MLP	22±2	69.2±2.4	20±3	66.7±4.2	20±1	73.3±5.1	18±4	75.8±2.4
	ICDH MLP	33±15	65.8±11.2	28±7	62.5±8.2	35±12	69.2±12.5	33±13	64.2±3.1
	Proposed method$^{[-]}$	18±3	75.8±5.1	16±6	75.8±8.2	15±7	79.2±12.3	16±6	77.5±6.1
	Proposed method$^{[+]}$	19±4	77.5±9.4	16±4	80.8±7.2	16±6	78.3±11.6	17±5	74.2±8.2

5.2 Results

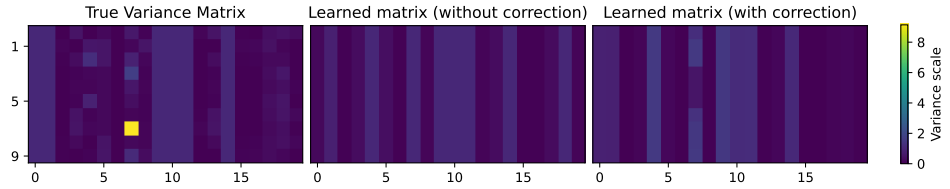
Centralized Performance across Noise Regimes. Table 1 reports centralized results (degree 2, $d = 20$) under four noise regimes, focusing on a representative subset of baselines and evaluating *SHD* and TPR.



(a) Noise mode - NV

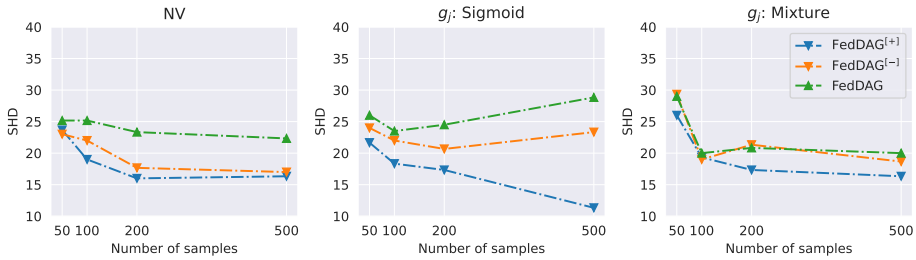


(b) Noise mode - g_i with sigmoid

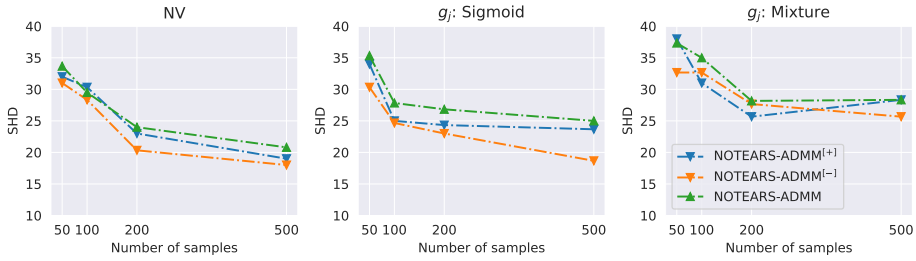


(c) Noise mode - g_i with mixture.

Figure 2: Comparison of the ground-truth variance matrix and the variance matrices learned by the proposed methods under three noise regimes, using the same setting as Table 1 with 100 samples. Each heatmap displays $d = 20$ nodes (x-axis) across 10 randomly selected observations (y-axis), with color representing the variance magnitude.

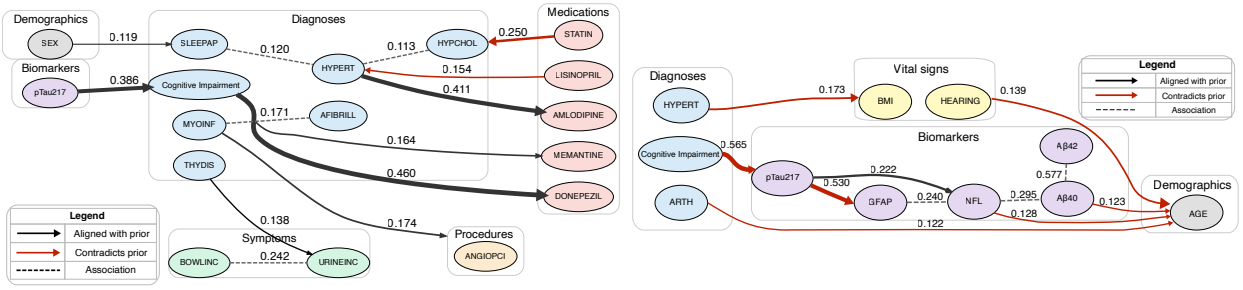


(a) FedDAG vs. FedDAG$^{[-]\setminus[+]}$.



(b) NOTEARS-ADMM vs. NOTEARS-ADMM$^{[-]\setminus[+]}$.

Figure 3: SHD comparison between baseline federated methods and their variants augmented with the proposed method (Algorithm 2): baseline$^{[-]}$ (conditioning matrix only) and baseline$^{[+]}$ (conditioning plus correction matrix).



(a) Proposed method$^{[-]}$ (b) NOTEARS

Figure 4: Causal diagrams learned from IADRC data. (a) Proposed method$^{[-]}$; (b) NOTEARS. Black solid arrows denote edges aligned with prior, red solid arrows contradict prior (directions retained as learned), and dashed lines indicate associations (statistical dependencies present in the data, not interpreted as causal). Edge labels are learned weights and line width scales with weight.

In the Equal Variance (EV) setting, homoscedastic methods such as NOTEARS achieve low SHD and high TPR, and our method remains competitive, indicating that the proposed variance-matrix approximation does not sacrifice performance when the equal-variance assumption holds. However, when variances differ, classical methods degrade substantially. In the *Node-wise Variance (NV)* regime, NOTEARS and ICDH exhibits much higher SHD with TPR below 65%, whereas our method reduces SHD and keeps TPR close to 75%, highlighting the benefit of node-specific variance modeling.

Under heteroscedastic regimes with nonlinear noise, the performance gap further widens. In the *sigmoid* setting, representative baselines typically yield SHD above 25 with TPR below 60%, while our method achieves SHD reduction near $5 \sim 10$ and TPR above 85%. The *mixture* setting is the most challenging, where baseline performance becomes less stable across sample sizes; in contrast, our method maintains SHD under 20 with TPR around 75–80%, demonstrating robustness to heterogeneous noise.

For completeness, we present additional results for $d \in 10, 50$ in Tables 3, 4 (Appendix B). We also conduct independent experiment and report CPDAG-SHD and the corresponding TPR in Table 2 (Appendix B). Overall, although classical methods remain competitive under homoscedastic noise, their performance degrades markedly in the presence of heteroscedasticity. By explicitly modeling variance, our method achieves consistent improvements across regimes and remains robust to both node-wise and input-dependent heteroscedastic noise, despite the lack of full identifiability guarantees.

Verification of the Learned Variance Matrix. Figure 2 visualizes the proposed variance–matrix approximation under the same setting as Table 1 with 100 samples. The conditioning matrix alone reproduces the NV column-constant structure, while the correction adds only minor refinements. For sigmoid-based noise, the column-wise pattern is preserved, though the correction does not fully capture ground-truth heterogeneity. In the mixture case, the heavier-tailed global structure is retained. Overall, the conditioning matrix offers a stable baseline, whereas modeling fine-grained row-wise effects remains challenging.

Performance Comparison under FL Setting. Figure 3 shows the effect of adding the proposed conditioning-matrix approximation to FedDAG and NOTEARS-ADMM. Each baseline has two variants: baseline$^{[-]}$ (conditioning only) and baseline$^{[+]}$ (conditioning with correction). Across sample sizes and noise regimes, the augmented variants typically improve or remain competitive, with the largest SHD gains seen for FedDAG under simpler NV and sigmoid modes. For the mixture regime, improvements persist but are less pronounced. As a caveat, in FL, the best iteration was chosen using the ground-truth DAG rather than training loss (unlike the centralized case). Appendix B.3, Figure 5, shows loss-based selection, where usually only one augmented variant outperforms the baseline. This limitation underscores that reliable stopping criteria remain an open challenge.

Effectiveness of variance lower-bound annealing strategy. In addition, we present a simplified comparison to partially justify the effectiveness of variance lower-bound annealing strategy. Under the same experimental setting as in Table 1, the table 5 in the Appendix B.2 shows that removing this stabilizing strategy generally leads to degraded performance.

6 Case Study

IADRC Data. We conduct a case study using multi-modal clinical data curated by the Indiana Alzheimer's Disease Research Center (IADRC), an NIA-funded Alzheimer's Disease Research Center that contributes to the National Alzheimer's Coordinating Center (NACC). The cohort comprises $N = 1,364$ participants with $4,870$ visit-level records spanning demographics, diagnoses, vital signs, medications, and cognitive assessments, plus $7,553$ biomarker measurements. Fluid, plasma, and genetic biomarkers are represented as preprocessed quantitative features, including amyloid-β 1–42 (Aβ42), amyloid-β 1–40 (Aβ40), total tau (T-tau), neurofilament light chain (NFL), phosphorylated tau at threonine 181 (pTau181), phosphorylated tau at threonine 217 (pTau217), glial fibrillary acidic protein (GFAP), and apolipoprotein E (APOE) genotype, rather than raw imaging or assay files.

6.1 Experimental Analysis

We compare our approach (Proposed method$^{[-]}$) with NOTEARS in the case study. Figure 4 provides learned causal diagrams for both methods. In general, our approach consistently yields compact graphs with directionally plausible relations (e.g., pTau217 $\rightarrow$ Cognitive Impairment $\rightarrow$ Donepezil/Memantine) Mielke et al. (2020); Wang et al. (2025a); Birks & Harvey (2018); Merative, Micromedex (2025), where Cognitive Impairment is a clinician-rated category including dementia, all MCI subtypes, and "impaired, not MCI," as categorized in the NACC UDS Form Besser et al. (2018). pTau217 is a strong biomarker for Alzheimer's disease and cognitive impairment Mielke et al. (2020); Wang et al. (2025a), and Donepezil/Memantine are guideline-recommended symptomatic treatments approved by FDA Birks & Harvey (2018); Merative, Micromedex (2025). Furthermore, our method avoids implausible arrows into exogenous covariates such as AGE Hernán & Robins (2010). In contrast, NOTEARS often produces unstable structures and biologically questionable directions. For example, it places AGE as a child of biomarkers (Aβ40/NFL $\rightarrow$ AGE) Khalil et al. (2020), inverts the biomarker–outcome link (Cognitive Impairment $\rightarrow$ pTau217).

We attribute this performance gap primarily to objective–data mismatch in noise modeling. In its standard least-squares formulation, NOTEARS is most aligned with a linear-Gaussian SEM under homoscedastic noise and is primarily designed for continuous variables. In our clinical setting with mixed variable types (continuous biomarkers, binary indicators, and ordinal covariates), such assumptions are likely violated, which can degrade estimation and make edge orientation more sensitive, consistent with the less accurate/less stable structures produced by NOTEARS in our case study. By contrast, our method explicitly targets heteroscedasticity via a two-stage variance approximation and stabilization (e.g., variance flooring/correction), which improves structural recovery in this setting. Overall, these results suggest that accounting for heteroscedastic noise is important for causal discovery on real-world clinical data.

7 Conclusion

This work addressed the challenge of causal structure learning under non-equal noise variances, a regime that frequently arises in practice but is poorly handled by existing optimization-based DAG learners. We showed that simple but principled variance–matrix approximations can stabilize optimization and improve robustness in causal structure learning under heterogeneous noise. Our framework consistently outperforms or refines equal-variance baselines in centralized and federated settings, and yields biologically plausible graphs in real data. These results highlight that carefully designed variance modeling serves as a crucial mechanism for enhancing the reliability and robustness of optimization-based causal discovery in heterogeneous, real-world environments.

References

Dimitri Bertsekas and John Tsitsiklis. *Parallel and distributed computation: numerical methods*. Athena Scientific, 2015.

Lilah Besser, Walter Kukull, David S Knopman, Helena Chui, Douglas Galasko, Sandra Weintraub, Gregory Jicha, Cynthia Carlsson, Jeffrey Burns, Joseph Quinn, et al. Version 3 of the national alzheimer's coordinating center's uniform data set. *Alzheimer Disease & Associated Disorders*, 32(4):351–358, 2018.

Jacqueline S Birks and Richard J Harvey. Donepezil for dementia due to alzheimer's disease. *Cochrane Database of systematic reviews*, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Bao Duong and Thin Nguyen. Heteroscedastic causal structure learning. *arXiv preprint arXiv:2307.07973*, 2023.

Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665, 2012.

P ERDdS and A R&wi. On random graphs i. *Publ. math. debrecen*, 6(290-297):18, 1959.

Enmao Gao, Jinzhu Chen, Li Shen, Tianyang Liu, Mingming Gong, and Howard Bondell. Feddag: Federated dag structure learning. In *Transactions on Machine Learning Research*, 2023.

Zhimin Guo, Xiangyu Li, Xiaoxiao Wu, and Qiang Zhou. Fedecd: Federated causal discovery from heterogeneous data. In *Advances in Neural Information Processing Systems*, 2024.

Kuan He, Zhihan Zhang, Bryon Aragam, and Jingtian Liang. Distributionally robust causal structure learning under heterogeneous environments. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28185–28198, 2021.

Miguel A Hernán and James M Robins. Causal inference, 2010.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1551–1560, 2018.

Yiyi Huang, Matthäus Kleindessner, Alexey Munishkin, Debvrat Varshney, Pei Guo, and Jianwu Wang. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in big Data*, 4:642182, 2021.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pp. 14316–14332. PMLR, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, pp. 393–400, 2007.

Michael Khalil, Lukas Pirpamer, Edith Hofer, Margarete M Voortman, Christian Barro, David Leppert, Pascal Benkert, Stefan Ropele, Christian Enzinger, Franz Fazekas, et al. Serum neurofilament light levels in normal aging and their association with morphologic brain changes. *Nature communications*, 11(1):812, 2020.

Go Kikuchi. Differentiable causal discovery under heteroscedastic noise. In *International Conference on Neural Information Processing*, pp. 284–295, 2022.

Samantha Kleinberg and George Hripcsak. A review of causal inference for biomedical informatics. *Journal of biomedical informatics*, 44(6):1102–1112, 2011.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: Regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Loka Li, Ignavier Ng, Gongxu Luo, Biwei Huang, Guangyi Chen, Tongliang Liu, Bin Gu, and Kun Zhang. Federated causal discovery from heterogeneous data. *arXiv preprint arXiv:2402.13241*, 2024.

Yingyu Lin, Yuxing Huang, Wenqin Liu, Haoran Deng, Ignavier Ng, Kun Zhang, Mingming Gong, Yi-An Ma, and Biwei Huang. A skewness-based criterion for addressing heteroscedastic noise in causal discovery. *arXiv preprint arXiv:2410.06407*, 2024.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35:19290–19303, 2022.

Merative, Micromedex. Memantine and donepezil (oral route). Mayo Clinic, 10 2025. URL `https://www.mayoclinic.org/drugs-supplements/memantine-and-donepezil-oral-route/description/drg-20137323`.

Bruno Merz, Günter Blöschl, Sergiy Vorogushyn, Francesco Dottori, Jeroen CJH Aerts, Paul Bates, Miriam Bertola, Matthias Kemter, Heidi Kreibich, Upmanu Lall, et al. Causes, impacts and patterns of disastrous river floods. *Nature Reviews Earth & Environment*, 2(9):592–609, 2021.

Michelle M Mielke, Jeremiah A Aakre, Alicia Algeciras-Schimnic, Nicholas Proctor, Mary M Machulda, David S Knopman, Clifford R Jack Jr, Ronald C Petersen, and Jeffrey L Dage. Comparison of cerebrospinal fluid phosphorylated tau 181 and 217 for cognitive progression: Biomarkers (non-neuroimaging): Longitudinal and prognostic biomarker studies. *Alzheimer's & Dementia*, 16:e040503, 2020.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pp. 726–751. PMLR, 2023.

Ignavier Ng and Kun Zhang. Towards federated bayesian network structure learning with continuous optimization. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

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

Marco Piccininni, Stefan Konigorski, Jessica L Rohmann, and Tobias Kurth. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. *BMC medical research methodology*, 20(1):179, 2020.

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

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

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

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 1991.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 72:102099, 2023.

Natasa Tagasovska, Hugo Chavez, Yu Cui, and David Lopez-Paz. Distinguishing cause from effect using quantiles. In *International Conference on Machine Learning*, pp. 9302–9312, 2020.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Pulakesh Upadhyaya, Kai Zhang, Can Li, Xiaoqian Jiang, Yejin Kim, et al. Scalable causal structure learning: Scoping review of traditional and deep learning algorithms and new opportunities in biomedicine. *JMIR Medical Informatics*, 11(1):e38266, 2023.

Yue Wang, Tianshu Zhu, Qian Cheng, Xiaolin Cui, Pengfei Zhang, Zhiming Lu, and Alzheimer's Disease Neuroimaging Initiative (ADNI)*. Predicting brain health in community-dwelling elderly populations by integrating gaussian mixture model and plasma biomarkers. *Journal of Alzheimer's Disease Reports*, 9: 25424823251331110, 2025a.

Yunxia Wang, CAO Fuyuan, Kui Yu, and Jiye Liang. Federated causal structure learning with non-identical variable sets. In *Forty-second International Conference on Machine Learning*, 2025b.

Sijie Xu, Omar Mian, Alexander Marx, and Jilles Vreeken. Inferring cause and effect in the presence of heteroscedastic noise. In *International Conference on Machine Learning*, pp. 24811–24825, 2022.

Qian Ye, Arash A Amini, and Qiang Zhou. Communication-efficient federated structural learning under heterogeneity. *arXiv preprint arXiv:2210.15450*, 2022.

Qiaoling Ye, Arash A Amini, and Qing Zhou. Federated learning of generalized linear causal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6623–6636, 2024.

Naiyu Yin, Tian Gao, Yue Yu, and Qiang Ji. Effective causal discovery under identifiable heteroscedastic noise model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163, 2019.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

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*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *Conference on Uncertainty in Artificial Intelligence*, pp. 156–166, 2020.

A Appendix

LLM Usage Disclosure

A large language model (ChatGPT, GPT-5, OpenAI) was used solely to polish the writing of this manuscript, including grammar correction and stylistic refinement.

A.1 Related Work

Causal structure learning spans constraint-based, score-based, and continuous optimization approaches. Classical methods, such as PC/FCI use conditional independence tests to recover Markov equivalence classes (Spirtes & Glymour, 1991; Spirtes et al., 2000); score-based search (e.g., GES) optimizes penalized likelihoods over graphs (Chickering, 2002); and identifiability via non-Gaussian noise is exemplified by LiNGAM (Shimizu et al., 2006). Recent continuous methods formulate acyclicity via differentiable constraints, exemplified by NOTEARS and its extensions to nonlinear mechanisms and likelihood-based objectives (DAG-GNN, GraN-DAG, GOLEM, NOTEARS-MLP), but they largely assume equal-variance additive noise, which can be misleading (e.g., "varsortability" artifacts) in benchmarks (Zheng et al., 2018; Ng et al., 2020; Yu et al., 2019; Lachapelle et al., 2019; Zheng et al., 2020; Reisach et al., 2021; Pearl et al., 2000; Koller & Friedman, 2009; Peters et al., 2017).

Heteroscedasticity provides asymmetries that can aid identification: bivariate methods leverage quantile-based directionality (Tagasovska et al., 2020); partitioned-variance regression (Xu et al., 2022) and location–scale identifiability results (Immer et al., 2023) formalize when variance signals suffice. Multivariate advances include HOST (heteroscedastic CI testing with Gaussian mechanisms), ICDH (alternating structure and variance-function fitting), and a differentiable mean–variance framework (Duong & Nguyen, 2023; Yin et al., 2024; Kikuchi, 2022). Connections to variance networks and aleatoric uncertainty further motivate modeling input-dependent noise Nix & Weigend (1994); Kendall & Gal (2017). However, most methods either impose restrictive forms or struggle to scale.

In federated settings, recent work separates global structure from local mechanisms (FedDAG), stabilizes non-*i.i.d.* optimization via proximal terms (FedCausal), exploits environment shifts (DA RING), or adapts independence-change principles (Gao et al., 2023; Ye et al., 2022; He et al., 2021; Guo et al., 2024). Yet these frameworks generally inherit equal-variance SEM assumptions; explicit heteroscedastic modeling in FL remains underexplored.

A.2 Discussion: What if we optimize the heteroskedastic (profiled) likelihood freely?

With the per-sample hNLL

$$\mathcal{L}_{hNLL}(\mathbf{X}) = \sum_{i,j} \left[\frac{1}{2\sigma_{ij}^2} r_{ij}^2 + \frac{1}{2} \log \sigma_{ij}^2 \right],$$

where r_{ij} is the residual of variable j at sample i. If the per-entry variances σ_{ij}^2 are optimized freely, each summand

$$\phi(\sigma_{ij}^2) := \frac{1}{2\sigma_{ij}^2} r_{ij}^2 + \frac{1}{2} \log \sigma_{ij}^2$$

the pointwise minimizer is $\tilde{\sigma}_{ij}^2 = r_{ij}^2$, which, plugged back, yields the minimized value

$$\phi(\tilde{\sigma}_{ij}^2) = \frac{1}{2}(1 + \log r_{ij}^2).$$

Hence, as $|r_{ij}| \downarrow 0$ we obtain $\phi(\tilde{\sigma}_{ij}^2) \to -\infty \quad \text{because} \quad \log r_{ij}^2 \to -\infty$.

Degeneracy: If the model class for r_{ij} can (nearly) interpolate the training residuals at the observed $\mathbf{X}_i$, the full NLL becomes unbounded below. This creates a *variance-collapse* failure mode: the optimizer simultaneously drives $r_{ij} \to 0$ and $\sigma_{ij}^2 \to r_{ij}^2$, pushing the objective to $-\infty$. I.e., the optimizer can "cheat" by driving the variance network and the mean network as $\sigma_{ij}^2 \to r_{ij}^2 \to 0$.

This leads to "cherry-pick one term". Once per-entry variances are free Optimizing the sum can be dominated by "winning" on a single (i, j) pair: drive one residual toward zero while letting the variance tracker follow it ($\sigma_{ij}^2 = r_{ij}^2$) and the entire objective dives to $-\infty$, regardless of what happens elsewhere. In other words, gradient descent may preferentially "optimize one term" instead of improving the overall fit.

This leads to numerical instability for gradient methods. Parameterize $\sigma_{ij} > 0$ directly. For one term

$$\psi(\sigma) = \frac{r^2}{2\sigma^2} + \log \sigma.$$

At the stationary point $\sigma^2 = r^2$, curvature is $\psi''(\sigma) = \frac{2}{r^2}$. Thus, as $|r| \downarrow 0$ the landscape becomes extremely sharp (ill-conditioned). Moreover, gradients w.r.t. structural parameters inherit factors, such as r/σ^2; when $\sigma \approx |r| \downarrow$, these explode, causing numerical instability and erratic updates.

A.3 Proposed Method: FL Version Algorithm

Algorithm 2 two-stage Federated causal Structure Learning with variance Matrix Approximation (FSL-MA)

Input: K clients with corresponding distributed data $\{\mathbf{X}_k\}_{k\in[K]}$, functions $\{F_k(;\mathbf{B}_k, \mathbf{W}_{F,k})\}_{k\in[K]}$, tolerance ϵ (1e-3), max iteration T (20); scalar ρ (0.5), variance bound $\underline{\sigma}^2, \overline{\sigma}^2$ (1e-2, 1), decay strategy (linear decay).
Initialize: $\widetilde{\mathbf{\Sigma}}_k = \overline{\sigma}^2 \mathbf{1}_{n\times d}$ and $t = 0$.
while $|\ell_t - \ell_{t-1}| < \epsilon$ or $t < T$ **do**
    ## Stage-I:
    Initialize FL training hyper-parameters
    $\widetilde{\mathbf{W}}_F, \widetilde{\mathbf{B}} \leftarrow$ Federated structure learning under locally fixed: $\widetilde{\mathbf{\Sigma}}_k$
    Broadcast $\widetilde{\mathbf{W}}_F, \widetilde{\mathbf{B}}$ to $\{\widetilde{\mathbf{W}}_{F,k}\}_{k\in[K]}, \widetilde{\mathbf{B}}_k$
    ## Stage-II:
    for each client k **do**
        $\widetilde{\mathbf{\Sigma}}_k^c \xleftarrow{\text{equation 3}} \widetilde{\mathbf{B}}_k, \widetilde{\mathbf{W}}_{F,k}$
    end for
    $\widetilde{\mathbf{\Sigma}}^c = \text{Aggregate}(\{\widetilde{\mathbf{\Sigma}}_k^c\}_{k\in[K]})^{[a]}$
    $\text{var}_{\min} = \text{Decay strategy}(t+1, T, \underline{\sigma}^2, \overline{\sigma}^2)$
    Broadcast $\widetilde{\mathbf{\Sigma}}^c = \text{clip}(\widetilde{\mathbf{\Sigma}}^c, \text{var}_{\min}, \overline{\sigma}^2)$
    for each client k **do**
        $\widetilde{H}(\cdot)_k \xleftarrow{\text{equation 4}} \widetilde{\mathbf{B}}_k, \widetilde{\mathbf{W}}_{F,k}, \widetilde{\mathbf{\Sigma}}^c$
    end for
    Broadcast $\widetilde{H}(\cdot) = \text{Aggregate}(\{\widetilde{H}(\cdot)_k\}_{k\in[K]})$
    for each client k **do**
        $\widetilde{\mathbf{\Sigma}}_k^h \xleftarrow{\text{equation 5}} \widetilde{H}(\cdot)$
        $\widetilde{\mathbf{\Sigma}}_k = \widetilde{\mathbf{\Sigma}}^c + \widetilde{\mathbf{\Sigma}}_k^h$
    end for
    ## For stopping mechanism:
    $\ell_t = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{\text{NLL}}(\mathbf{X}_k; \widetilde{\mathbf{B}}_k, \widetilde{\mathbf{W}}_{F,k})$
    t = t + 1
end whilereturn $\widetilde{\mathbf{B}}$ corresponding to lowest ℓ_t.

[a] True underlying information of $\mathbf{\Sigma}^c \in \mathbb{R}^{n\times d}$ reduces to a vector $\mathbf{\Sigma}^c \in \mathbb{R}^d$, hence dim n is adaptable for matrix $\mathbf{\Sigma}^c$ related operation.

A.4 Proof of section 4

Under the heteroscedastic model equation 1: (i) The optimal constant variance is $(\sigma_j^2)^\star = \mathbb{E}[g_j(\mathrm{pa}(X_j))^2]$. (ii) The irreducible variance gap is

$$\xi_j^{\mathrm{var}} = \tfrac{1}{2}\big(\log \mathbb{E}[g_j(\mathrm{pa}(X_j))^2] - \mathbb{E}[\log g_j(\mathrm{pa}(X_j))^2]\big)\,,$$

which vanishes iff $g_j(\mathrm{pa}(X_j))^2$ is constant.

Proof. Given the true (node-wise) distribution and approximation of distribution:

True: $X_j \sim \mathcal{N}\left(f_j(\mathrm{pa}(X_j)), g_j(\mathrm{pa}(X_j))^2\right)$; Approximation $X_j \sim \mathcal{N}\left(\tilde{f}_j(\mathrm{pa}(X_j)), \sigma_j^2\right)$

We first simplify the notation as

True: $X_j \sim \mathcal{N}(\mu_j(X), v_j(X))$; Approximation: $X_j \sim \mathcal{N}(\tilde{\mu}_j(X), \sigma_j^2)$.

The per-node population error of using a constant variance is formulated as the expected conditional KL:

$$\xi_j(\sigma_j^2, \tilde{\mu}_j) := \mathbb{E}\left[\mathrm{KL}(\mathcal{N}(\mu_j(X), v_j(X)) \| \mathcal{N}(\tilde{\mu}_j(X), \sigma_j^2))\right]$$

Using the closed form for Gaussian–Gaussian KL,

$$\mathrm{KL} = \frac{1}{2}\left(\frac{v_j(X)}{\sigma_j^2} + \frac{(\tilde{\mu}_j(X) - \mu_j(X))^2}{\sigma_j^2} - 1 + \log\frac{\sigma_j^2}{v_j(X)}\right).$$

Taking expectation over X,

$$\xi_j(\sigma_j^2, \tilde{\mu}_j) = \frac{1}{2}\left(\frac{\mathbb{E}\left[v_j(X)\right]}{\sigma_j^2} - 1 + \log\sigma_j^2 - \mathbb{E}\left[\log v_j(X)\right]\right) + \frac{1}{2\sigma_j^2}\mathbb{E}\left[(\tilde{\mu}_j(X) - \mu_j(X))^2\right],$$

where (a). the first bracket is the error purely from variance misspecification, and (b). the second term is the error from mean misspecification.

By minimize ξ_j over σ_j^2, we have

- Optimal constant variance: $(\sigma_j^2)^\star = \mathbb{E}\left[v_j(X)\right]$.
- Irreducible (best possible) variance-only gap in objective of training: $\xi_j^{\mathrm{var}} = \frac{1}{2}\left(\log \mathbb{E}\left[v_j\right] - \mathbb{E}\left[\log v_j\right]\right)$.

Substituting $g_j(\mathrm{pa}(X_j))^2$ back concludes the proof. □

B Additional Experiments

B.1 Centralized Performance with SHD-CPDAG.

Table 2 reports centralized results (degree 2, $d = 20$) under four noise regimes, focusing on a representative subset of baselines and evaluating *SHD-CPDAG* and TPR.

B.2 Centralized Performance with Different Number of Node.

We extend the set of baselines evaluated in Table 1 and report results in terms of SHD and TPR. We additionally evaluate the centralized performance of the proposed method, using the same configuration as Table 1 except for the following modifications: Table 3 adopts $d = 10$, Table 4 adopts $d = 50$.

Table 2: Centralized evaluation under linear structural causal function with varying sample sizes (n) and noise modes (degree 2, $d = 20$). SHD-C denotes SHD at the CPDAG level, shortened from SHD-CPDAG for table compactness; lower is better. TPR: True Positive Rate (%). Particularly, $^{[-]}$ and $^{[+]}$ indicate the proposed method, Algorithm 1, without and with the correction matrix, respectively.

Noise mode	Method	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
		SHD-C	TPR	SHD-C	TPR	SHD-C	TPR	SHD-C	TPR
EV	NOTEARS	9.0±3.7	87.5±4.1	2.7±2.5	94.2±5.1	7.7±3.7	85.8±6.2	3.3±1.7	93.3±1.2
	ICDH Nodewise MLP	9.7±0.5	90.8±2.4	5.3±2.9	90.8±3.1	4.3±2.1	92.5±2.0	4.7±2.9	92.5±3.5
	Proposed method$^{[+]}$	7.3±1.7	89.2±2.4	2.3±2.1	94.2±5.1	5.7±3.9	89.2±6.6	3.3±1.7	93.3±1.2
NV	NOTEARS	37.3±6.1	54.2±8.2	24.3±7.8	57.5±12.2	24.7±3.9	53.3±7.7	22.0±8.5	58.3±13.1
	ICDH Nodewise MLP	26.0±5.9	55.0±12.4	25.0±4.2	58.3±7.7	26.0±2.4	58.3±5.9	23.0±7.3	62.5±13.4
	Proposed method$^{[+]}$	21.3±5.8	73.3±8.5	16.7±4.9	78.3±6.2	15.0±2.8	74.2±4.2	14.0±0.8	75.0±4.1
g_j with sigmoid	NOTEARS	24.3±2.5	60.8±5.1	27.0±3.7	56.7±1.2	28.3±1.7	53.3±2.4	29.7±3.7	56.7±3.1
	ICDH Nodewise MLP	28.7±2.5	55.0±2.0	28.3±3.1	57.5±2.0	29.7±2.6	54.2±4.2	25.7±3.4	61.7±4.2
	Proposed method$^{[+]}$	17.3±7.6	82.5±9.4	7.3±7.6	91.7±6.6	7.3±7.6	93.3±6.2	7.7±10.1	90.8±7.7
g_j with mixture	NOTEARS	28.3±4.2	61.7±4.2	25.3±7.4	60.0±7.4	16.7±6.9	74.2±9.6	21.3±2.1	69.2±5.9
	ICDH Nodewise MLP	24.0±3.3	69.2±2.4	20.7±5.0	66.7±7.7	21.7±2.9	73.3±5.1	18.7±5.3	73.3±3.1
	Proposed method$^{[+]}$	19.0±1.6	81.2±1.8	14.3±4.6	80.6±3.5	14.5±3.5	82.2±5.3	15.5±3.5	78.8±5.3

Table 3: Centralized evaluation (10 nodes, ER) under linear structural causal function with varying sample sizes (n). SHD: Structural Hamming Distance (lower is better). TPR: True Positive Rate (%).

Noise mode	Method	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
		SHD	TPR	SHD	TPR	SHD	TPR	SHD	TPR
g_j with mixture	NOTEARS	11±5	63.3±9.4	13±2	55.0±10.8	13±8	53.3±24.9	9±5	63.3±20.1
	GOLEM-NV	9±4	75.0±8.2	13±3	56.7±6.2	10±6	66.7±1.7	6±2	75.0±8.2
	HOST	12±6	68.3±17.0	11±4	71.7±15.5	10±1	76.7±8.5	11±2	83.3±4.7
	VarSort	12±3	53.3±9.4	11±2	56.7±6.2	11±1	66.7±4.7	10±5	58.3±13.1
	ICDH Nodewise MLP	8±3	73.3±4.7	14±2	53.3±11.8	12±8	55.0±22.7	11±7	63.3±25.9
	ICDH MLP	13±6	58.3±22.5	14±4	55.0±14.1	13±5	48.3±19.3	16±4	56.7±14.3
	Proposed method$^{[-]}$	6±5	81.7±8.5	11±3	68.3±10.3	10±6	65.0±18.7	11±6	70.0±20.4
	Proposed method$^{[+]}$	7±6	80.0±14.7	7±0	80.0±4.1	6±3	75.0±10.8	7±6	80.0±17.8

Table 4: Centralized evaluation (50 nodes, ER) under linear structural causal function with varying sample sizes (n). SHD: Structural Hamming Distance (lower is better). TPR: True Positive Rate (%).

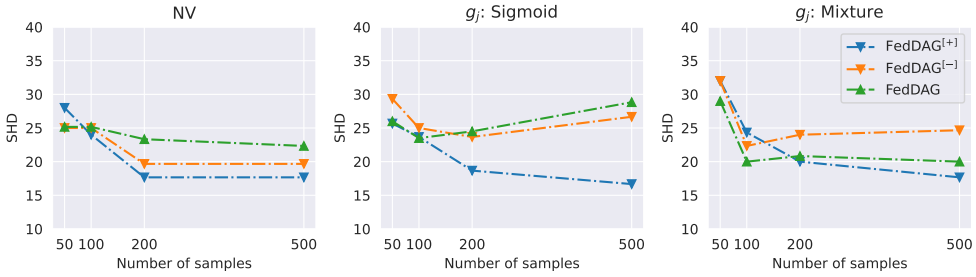
Noise mode	Method	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
		SHD	TPR	SHD	TPR	SHD	TPR	SHD	TPR
g_j with mixture	NOTEARS	60±9	69.3±3.3	42±15	77.3±6.5	52±16	71.7±3.3	49±17	72.3±6.2
	GOLEM-NV	87±16	59.3±10.3	68±9	66.7±5.2	80±10	57.7±4.9	85±14	58.0±3.7
	HOST	177±24	32.3±2.9	95±16	49.3±5.3	77±14	60.3±5.7	82±0	61.7±1.7
	VarSort	104±6	32.3±3.3	100±9	34.7±4.9	86±7	41.7±3.1	84±5	52.0±2.9
	ICDH Nodewise MLP	62±11	71.0±2.9	48±8	78.3±3.4	41±1	78.3±0.9	63±8	69.3±3.3
	ICDH MLP	105±10	45.0±4.1	68±35	60.7±19.4	44±12	72.7±8.2	48±23	74.0±11.0
	Proposed method$^{[-]}$	60±11	73.7±6.9	40±14	82.0±6.4	38±3	81.0±2.2	45±7	77.3±4.0
	Proposed method$^{[+]}$	69±3	73.3±5.9	35±13	82.7±6.2	39±5	79.3±1.2	42±14	78.0±7.0

Table 5: Centralized evaluation (20 nodes, ER) under linear structural causal function with varying sample sizes (n). SHD: Structural Hamming Distance (lower is better).

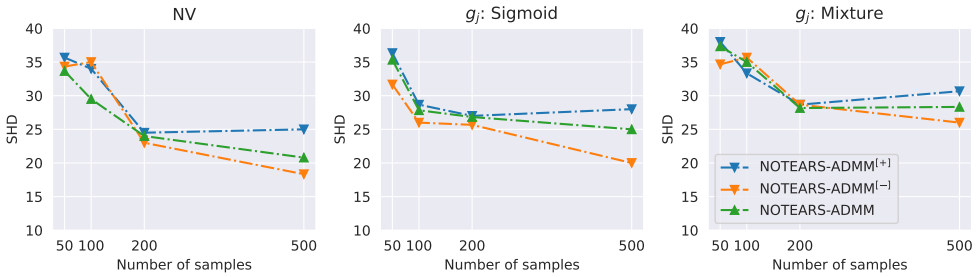
Noise mode	Method	SHD			
		$n = 50$	$n = 100$	$n = 200$	$n = 500$
g_j with mixture	Proposed method$^{[-]}$	7±1	11±7	9±7	6±6
	Proposed method$^{[-]}$ w.o. "progressive variance flooring"	18.3±8.2	8.7±5.4	10.3±9.9	13.0±8.6

B.3 Performance Comparison under FL Setting

Similar to Figure 3 in the main content, Figure 5 evaluates the impact of incorporating the proposed conditioning-matrix approximation into two federated baselines, NOTEARS-ADMM and FedDAG, where the best iterations are selected using loss values as in the centralized experiments.



(a) FedDAG vs. augmented FedDAG.



(b) NOTEARS-ADMM vs. augmented NOTEARS-ADMM.

Figure 5: SHD comparison between baseline federated methods and their variants augmented with the proposed method: baseline$^{[-]}$ (conditioning matrix only) and baseline$^{[+]}$ (conditioning plus correction matrix). Results are reported under three noise modes: {Noise-wise constant Variances (NV), g_j with sigmoid, g_j with mixture}.