HESSIAN RANK CONSTRAINT FOR LEARNING STRUCTURE OF NONLINEAR LATENT VARIABLE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncovering latent variables and their causal relations from observed data is a fundamental yet challenging problem. Existing methods often rely on restrictive assumptions, such as linear relations or invertible mixing functions. To better address this problem under general nonlinear mixing procedures, we propose a condition called the cross-Hessian Rank Constraint (HRC), which serves as a primitive rank-based tool for nonlinear latent causal discovery. In particular, we show that a rank-based property arises from the cross-Hessian of the observed-data log-density in the nonlinear case, revealing information about the latent variables, and reduces to the Tetrad constraints in the linear Gaussian case. More specifically, when two groups of observed variables are d-separated by a lower-dimensional set of latent variables, their cross-Hessian admits an approximately low-rank factorization. Under non-degeneracy and the singular-value gap assumptions, its thresholded rank identifies the dimension of these latent variables. The approximation is accurate when the conditional-score remainder is small under the posterior latent distribution. As a downstream application, we instantiate HRC in the pure one-factor measurement model for locating latent variables and recovering their causal structure up to Markov equivalence class. Experiments on synthetic and real-world datasets support the theoretical results.

1 INTRODUCTION

Discovering causal structure in the presence of causally related latent confounders is one of the central and most challenging problems in causal discovery. Classical approaches, such as FCI and its variants (Spirtes et al., 2000; Pearl, 2009; Zhang, 2008; Colombo et al., 2012; Akbari et al., 2021; Mokhtarian et al., 2025), use conditional independence information to recover aspects of the causal graph under latent confounding. However, these methods do not determine the number or the locations of latent variables, which has motivated a growing line of work on causal discovery among latent variables (Silva et al., 2006; Kummerfeld & Ramsey, 2016).

Recovering causal structure among latent variables is typically studied in settings where observed variables are leaves in the graph and are not directly adjacent. Within this setting, existing approaches rely on additional structural or distributional assumptions and exploit a variety of statistical signatures, including rank constraints (Silva et al., 2006; Kummerfeld & Ramsey, 2016; Huang et al., 2022; Xie et al., 2022; Dong et al., 2024), higher-order moments (Shimizu et al., 2009; Cai et al., 2019; Xie et al., 2020; Adams et al., 2021; Chen et al., 2022), matrix decomposition (Anandkumar et al., 2013; Chen et al., 2025), copula-based formulations (Cui et al., 2018), multi-domain information (Zeng et al., 2021; Li et al., 2023; Sturma et al., 2023; Zhang et al., 2024), and mixture-based oracles (Kivva et al., 2021; 2022). Despite their methodological diversity, most of these methods still rely on restrictive assumptions, such as linear measurement relations from latent to observed variables or discrete latent variables. These assumptions may be hard to satisfy in real-world scenarios, where observed variables are often generated through general nonlinear processes. Please refer to Appendix A for further discussion of related work on causal discovery in the presence of latent variables.

Extending causal discovery among latent variables to the general nonlinear setting remains relatively underexplored, and most of existing works focus on a special class of hierarchical latent-structure discovery problems. In this direction, Kong et al. (2023) identify hierarchical latent variables under a specific nonlinear latent variable model and then recover the structure among them. However, they

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

assume that latent variables are deterministic functions of measured variables. This assumption is often violated, for instance, if each measured variable has its own noise. More recently, Prashant et al. (2025) study the nonlinear latent hierarchical model from a different perspective. Specifically, they leverage the rank of the Jacobian of the conditional expectation mapping between groups of observed variables to infer the size of separating latent sets and recover the latent structures. Nevertheless, they assume the existence of a function of the measured variables with the same dimensionality as the latent variables that make them conditionally independent. Although these works mark important steps, they primarily focus on a specific class of hierarchical latent-structure discovery problems and rely on restrictive assumptions. For causal discovery among latent variables under general nonlinear mixing processes, a more primitive tool is still urgently needed.

Accordingly, in this paper, we propose a primitive tool for causal discovery among latent variables under general nonlinear mixing procedures. In linear latent variable models, Tetrad constraints characterize rank deficiencies in cross-covariance matrices (Huang et al., 2022). Interestingly, we show that an analogous rank-based property also arises in the nonlinear setting. Specifically, if two groups of observed variables are d-separated by a lower-dimensional set of latent variables, then their cross-Hessian is close to a matrix whose rank is controlled by the latent dimension. Under non-degeneracy and a singular-value gap, the corresponding thresholded rank identifies this dimension. The assumption constrains conditional-score remainders rather than the nonlinear mixing functions themselves. Taken together, these observations lead to the cross-Hessian Rank Constraint (**HRC**), a primitive rank-based criterion for nonlinear latent causal discovery.

As a concrete algorithmic application, we instantiate HRC in the standard pure one-factor measurement setting and incorporate it as the rank criterion within FOFC for latent-variable location and within PC-style search for learning the latent causal structure up to Markov equivalence (Kummerfeld & Ramsey, 2016; Silva et al., 2006; Dong et al., 2024). Experiments on synthetic data support our theoretical results, and the FRED-MD experiment recovers an economically interpretable latent structure.

2 PRELIMINARY

Denote by $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ a directed acyclic graph (DAG), where $\mathcal{V} := \mathcal{L} \cup \mathcal{X}$ consists of $n + m$ nodes. $\mathcal{L} := \{L_i\}_{i=1}^{n}$ denotes the set of n latent nodes, and $\mathcal{X} := \{X_j\}_{j=1}^{m}$ denotes the set of m observed nodes.[1] The edge set $\mathcal{E}$ represents causal relations, and ϵ_{V_i} denotes the independent noise associated with node V_i. These variables are generated via the following nonlinear latent variable model:

$$L_i = f_{L_i}(\mathrm{Pa}(L_i), \epsilon_{L_i}), \quad X_j = f_{X_j}(\mathrm{Pa}(X_j), \epsilon_{X_j}), \tag{1}$$

where f_{L_i} and f_{X_j} denote the general nonlinear functions and $\mathrm{Pa}(\cdot)$ denotes the set of parent variables of a given node in $\mathcal{G}$. Note that $\mathrm{Pa}(L_i) \subseteq \mathcal{L}$ and $\mathrm{Pa}(X_j) \subseteq \mathcal{V}$. Thus, latent variables are allowed to cause observed variables, and causal relations among observed variables are also allowed, but observed variables cannot be ancestors of latent variables. Such a data generation process provides a general formulation of the latent-variable causal models commonly considered in prior works (Silva et al., 2006; Kummerfeld & Ramsey, 2016), while allowing a general nonlinear process. Uppercase letters denote random variables or vectors and lowercase letters their realizations; P and p denote distributions and densities, respectively. See Appendix B for further notation. For disjoint $A, B \subseteq \mathcal{X}$, let $p_{A,B}$ be their marginal density under $P_{A,B}$. Its log-density cross-Hessian is

$$\mathbf{H}_{A,B} := \frac{\partial^2 \log p_{A,B}}{\partial A \, \partial B^\top}. \tag{2}$$

Here, derivatives are taken with respect to the corresponding arguments of the density. The notation $\mathrm{rank}(\mathbf{H}_{A,B})$ refers to the rank of this matrix-valued function. For any matrix-valued function $\mathbf{M}$ and nonnegative threshold function τ, define its thresholded-rank function by $\mathrm{rank}_\tau(\mathbf{M}) := \#\{j : \sigma_j(\mathbf{M}) > \tau\}$, where j ranges over the valid singular-value indices. A scalar τ is interpreted as a constant function. Statements involving $\sigma_j(\mathbf{M})$ or $\mathrm{rank}_\tau(\mathbf{M})$ are understood pointwise almost surely under the relevant observed distribution. Based on the aforementioned data generation process, our goal is twofold. First, we seek a criterion from the observed distribution that reflects the d-separation by latent variables under general nonlinear processes. Second, we use this criterion for causal

[1] Throughout the paper, we use "observed variables" and "measured variables" interchangeably to refer to variables directly available in the data.

discovery among latent variables. More specifically, we identify the locations of latent variables, i.e., which groups of observed variables are associated with the same latent variables, and then recover the causal relations among the latent variables up to the Markov equivalence class.

3 Cross-Hessian Rank Constraint

In this section, we develop the cross-Hessian Rank Constraint (HRC). Theorem 1 shows that latent d-separation yields an approximately low-rank cross-Hessian whose deviation is controlled by the conditional-score remainders. Under non-degeneracy and a singular-value gap, Theorem 2 identifies the minimum latent dimension through thresholded rank. We then give concrete smoothness conditions controlling the approximation error and estimate the cross-Hessian from observed variables using kernel density estimation (KDE).

3.1 Identifiability of the Dimensions of Latent Variables

In the linear setting, d-separation by latent variables induces rank constraints on cross-covariance matrices, which can reveal latent dimension (Spearman, 1928; Silva et al., 2006; Sullivant et al., 2010). For instance, if a single latent variable d-separates two observed groups A and B, then $\mathrm{rank}(\Sigma_{A,B}) \leq 1$. Because covariance captures only linear associations, this method does not directly extend to nonlinear measurement processes. We establish an analogous robust constraint using the cross-Hessian of the observed marginal log-density. If a single latent variable d-separates A and B, the second singular value of $\mathbf{H}_{A,B}$ is controlled by the conditional-score approximation error. Under the non-degeneracy and signal-gap conditions introduced later, thresholding at this error level yields rank one. Figure 1 illustrates this setting with L_1 d-separating $A = \{X_1, X_2\}$ and $B = \{X_3, X_4\}$. We first state two assumptions.

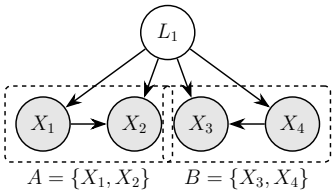


Figure 1: A causal graph where the latent variable L_1 d-separates two observed groups, $A = \{X_1, X_2\}$ and $B = \{X_3, X_4\}$

Assumption 1 (Smoothness and Integrability). *For fixed disjoint observed variables A, B and latent variables $\mathcal{Z} \subseteq \mathcal{L}$. Let Z denote its associated random vector. The densities $p_{A,B,Z}$, $p_{A,B}$, $p_{A|Z}$, and $p_{B|Z}$ are positive and twice continuously differentiable in their observed arguments, and differentiation under the relevant integrals is valid up to second order. Moreover,* $\mathrm{Cov}(Z \mid A, B)$ *exists and is finite $P_{A,B}$-almost surely.*

Discussion: These are standard regularity conditions for differentiating latent-variable marginal likelihoods (Louis, 1982). Specifically, positivity and twice continuous differentiability ensure that the relevant log-densities and cross-Hessian are well defined. Differentiation under the integral justifies passing derivatives through the latent marginalization, and finiteness of the posterior covariance makes the covariance factorization well defined.

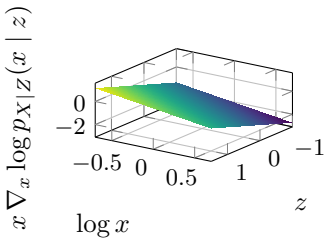


Figure 2: Nonlinear measurement with an affine conditional score.

Assumption 2 (Approximate Affine Derivatives). *Let A and B be disjoint groups of observed variables and let $Z \in \mathbb{R}^d$ be associated with a set of latent variables between them. Define the conditional scores $\mathbf{s}_A(a, z) := \nabla_a \log p_{A|Z}(a \mid z)$ and $\mathbf{s}_B(b, z) := \nabla_b \log p_{B|Z}(b \mid z)$. They admit the decompositions*

$$\begin{aligned} \mathbf{s}_A(a, z) &= \mathbf{u}_A(a) + \mathbf{V}_A(a)z + \mathbf{r}_A(a, z), \\ \mathbf{s}_B(b, z) &= \mathbf{u}_B(b) + \mathbf{V}_B(b)z + \mathbf{r}_B(b, z), \end{aligned} \tag{3}$$

where $\mathbf{u}_A, \mathbf{u}_B$ are intercept terms, $\mathbf{V}_A(a) \in \mathbb{R}^{|A|\times d}$ and $\mathbf{V}_B(b) \in \mathbb{R}^{|B|\times d}$ are slope matrices, and $\mathbf{r}_A, \mathbf{r}_B$ are remainder terms. Writing $Q_{a,b} := P_{Z|A=a,B=b}$ and $\bar{\mathbf{r}}_G := \mathbf{r}_G - \mathbb{E}_{Q_{a,b}}[\mathbf{r}_G]$, there exist prescribed nonnegative tolerance functions $\varepsilon_A(a, b)$ and $\varepsilon_B(a, b)$ such that

$$\left(\mathbb{E}_{Q_{a,b}} \|\bar{\mathbf{r}}_G\|^2\right)^{1/2} \leq \varepsilon_G(a, b), \qquad G \in \{A, B\}. \tag{4}$$

Discussion: Assumption 2 constrains the conditional scores rather than the measurement functions. One natural affine representation is obtained from a first-order Taylor expansion in z around a reference point z_0. In this representation, $\mathbf{V}_A(a) = \nabla_z \mathbf{s}_A(a, z_0)$ describes the first-order response of the conditional score to latent variation, while $\mathbf{r}_A$ collects the remaining higher-order variation; the same interpretation applies to group B. The posterior-centered bound controls precisely the part of this remainder that enters the cross-Hessian covariance. Thus, ε_A and ε_B quantify departures from affine conditional-score behavior. Figure 2 gives a nonlinear zero-remainder example: although $X = \exp(Z + \epsilon)$ is nonlinear in Z, its conditional score is affine in z. Further multivariate and iterated nonlinear constructions are provided in Appendix E.2.

Theorem 1 (**Approximate Cross-Hessian Rank under d-Separation**). *Let A, B be disjoint observed groups and let $Z \in \mathbb{R}^d$ be associated with latent variables that d-separate them. Under Assumptions 1 and 2, define*

$$\mathbf{H}^{\text{aff}}_{A,B} := \mathbf{V}_A(A)\,\mathrm{Cov}(Z \mid A, B)\mathbf{V}_B(B)^\top \tag{5}$$

and let $\varepsilon_{A,B} := \nu_Z(\|\mathbf{V}_A\|_2\varepsilon_B + \|\mathbf{V}_B\|_2\varepsilon_A) + \varepsilon_A\varepsilon_B$, where $\nu_Z := (\mathbb{E}[\|Z - \mathbb{E}[Z \mid A, B]\|^2 \mid A, B])^{1/2}$. Then the cross-Hessian satisfies $\mathbf{H}_{A,B} = \mathbf{H}^{\text{aff}}_{A,B} + \mathbf{E}_{A,B}$ with $\|\mathbf{E}_{A,B}\|_2 \le \varepsilon_{A,B}$. Consequently, $\sigma_j(\mathbf{H}_{A,B}) \le \varepsilon_{A,B}$ for every valid singular-value index $j > d$. Equivalently,

$$\mathrm{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) \le d. \tag{6}$$

Proof sketch and discussion. The full proof is in Appendix D.1. Latent d-separation gives $\mathbf{H}_{A,B} = \mathrm{Cov}(\mathbf{s}_A(A, Z), \mathbf{s}_B(B, Z) \mid A, B)$. Substituting Assumption 2 yields the affine benchmark in Equation (5) plus

$$\mathbf{E}_{A,B} = \mathbf{V}_A\,\mathrm{Cov}(Z, \mathbf{r}_B \mid A, B) + \mathrm{Cov}(\mathbf{r}_A, Z \mid A, B)\mathbf{V}_B^\top + \mathrm{Cov}(\mathbf{r}_A, \mathbf{r}_B \mid A, B). \tag{7}$$

Cauchy–Schwarz gives the stated bound $\|\mathbf{E}_{A,B}\|_2 \le \varepsilon_{A,B}$. Since $\mathrm{rank}(\mathbf{H}^{\text{aff}}_{A,B}) \le d$, Weyl's inequality gives Equation (6). Hence d-separation implies approximate rank at most d; the exact rank bound is recovered when the remainders vanish. Within-group causal relations are also allowed.

Assumption 3 (**Non-degeneracy**). *For disjoint groups A and B of observed variables, let $\mathcal{Z}^*$ be a minimum-cardinality set of latent variables that d-separates them, with $d^* := |\mathcal{Z}^*|$ and associated random vector $Z^* \in \mathbb{R}^{d^*}$. Under $P_{A,B}$, almost surely, (i) $\mathrm{Cov}(Z^* \mid A, B)$ is positive definite; and (ii) $\mathbf{V}_A(A)$ and $\mathbf{V}_B(B)$ have full column rank d^*.*

Assumption 3 ensures that no rank is lost in the affine benchmark associated with a minimum set of latent variables that d-separates A and B. When $d^* = 0$, the empty set d-separates A and B, and Assumption 3 imposes no additional condition. Condition (i) requires the posterior covariance of Z^* to retain variation along every latent direction, while condition (ii) requires each observed group to be locally informative about every such direction. In a linear Gaussian measurement model, $\mathbf{V}_A = \Omega_A^{-1}\Lambda_A$ and $\mathbf{V}_B = \Omega_B^{-1}\Lambda_B$, where Ω_A and Ω_B denote the conditional residual covariance matrices in groups A and B. Thus, condition (ii) is equivalent to requiring both loading matrices to have full column rank. When full rank is feasible, violations form a measure-zero algebraic subset of the unrestricted loading-parameter space, paralleling the generic exclusion of rank-deficient parameterizations in linear rank-faithfulness results (Sullivant et al., 2010; Huang et al., 2022; Dong et al., 2024). Please refer to Appendix E.1 for more discussion.

Theorem 2 (**Robust Hessian Rank Constraint**). *Let A and B be disjoint groups of observed variables, and let $\mathcal{Z}^*$ be a minimum-cardinality set of latent variables that d-separates them. Write $d^* := |\mathcal{Z}^*|$. Suppose Assumptions 1–3 hold with respect to $\mathcal{Z}^*$. If $d^* \ge 1$, suppose*

$$\sigma_{d^*}(\mathbf{H}^{\text{aff}}_{A,B}) > 2\varepsilon_{A,B}. \tag{8}$$

No signal-gap condition is required when $d^* = 0$. Then

$$\mathrm{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) = d^*. \tag{9}$$

Proof sketch and discussion. If $d^* = 0$, d-separation by the empty set gives $A \perp\!\!\!\perp B$, hence $\mathbf{H}_{A,B} = 0$ and the thresholded rank is zero. For $d^* \ge 1$, Assumption 3 gives $\mathrm{rank}(\mathbf{H}^{\text{aff}}_{A,B}) = d^*$. The decomposition in Theorem 1 and Weyl's inequality imply that all singular values beyond d^* are at most $\varepsilon_{A,B}$, while Equation (8) keeps the first d^* singular values above this threshold. The thresholded rank is therefore d^*. Please refer to Appendix D.3 for the complete proof. In the linear Gaussian case, the conditional scores are exactly affine, so $\varepsilon_{A,B} = 0$; see Appendix D.2.

Corollary 2.1. ***(Linear Gaussian Special Case)*** *Let A and B be disjoint observed groups. Let $\mathcal{Z}^*$ be a minimum latent d-separator between them, write $d^* = |\mathcal{Z}^*|$, and let $Z^* \in \mathbb{R}^{d^*}$ be its associated random vector. Suppose the full structural causal model is linear Gaussian, the joint distribution of (A, B) is non-degenerate, and Assumptions 1 and 3 hold. Then*

$$\operatorname{rank}(\mathbf{H}_{A,B}) = d^*.$$

3.2 Robustness Analysis of the Cross-Hessian Rank Constraint

The following result uses smoothness and posterior concentration to bound the conditional-score remainder and the resulting cross-Hessian approximation error.

Theorem 3 (**Approximate Cross-Hessian Factorization**). *For A, B, Z as in Theorem 1, fix $y = (y_A^\top, y_B^\top)^\top$ with $p_{A,B}(y) > 0$. Under Assumption 1, suppose that, for some $z_0 \in \mathbb{R}^d$ and $\delta > 0$:*

- ***(Lipschitz Continuity.)*** *The conditional scores $\mathbf{s}_A(y_A, \cdot)$ and $\mathbf{s}_B(y_B, \cdot)$ are twice continuously differentiable on $\mathcal{B}(z_0, \delta)$, and their componentwise Hessians are L_s-Lipschitz on this ball.*
- ***(Posterior Concentration.)*** *The conditional distribution $P_{Z|A=y_A,B=y_B}$ is supported on $\mathcal{B}(z_0, \delta)$.*

Let $\mathbf{V}_A = \nabla_z \mathbf{s}_A(y_A, z_0)$ and $\mathbf{V}_B = \nabla_z \mathbf{s}_B(y_B, z_0)$ be the corresponding score Jacobians. Let $\widetilde{\mathbf{C}}_A$ and $\widetilde{\mathbf{C}}_B$ collect the componentwise curvature bounds $\|\nabla_z^2 s_i(\cdot, z_0)\|_2 + \delta L_s/3$ for the respective conditional scores. Let $\mathbf{H}^{\text{aff}}_{A,B}(y) := \mathbf{V}_A \operatorname{Cov}(Z \mid A = y_A, B = y_B)\mathbf{V}_B^\top$ denote the affine Taylor benchmark. Then

$$\left\|\mathbf{H}_{A,B}(y) - \mathbf{H}^{\text{aff}}_{A,B}(y)\right\|_2 \le \tfrac{1}{2}\delta^3\Big(\|\mathbf{V}_A\|_2\|\widetilde{\mathbf{C}}_B\|_2 + \|\mathbf{V}_B\|_2\|\widetilde{\mathbf{C}}_A\|_2\Big) + \tfrac{1}{4}\delta^4\|\widetilde{\mathbf{C}}_A\|_2\|\widetilde{\mathbf{C}}_B\|_2. \tag{10}$$

Discussion. The proof is provided in Appendix D.5. Theorem 3 supplies explicit sufficient conditions for the small-remainder requirement in Assumption 2. Its bound decreases when the posterior latent values are more concentrated and the conditional scores have lower local curvature. The result therefore explains when the approximate HRC characterization is accurate; it does not prescribe how the cross-Hessian is estimated. More experiments in Appendix G.3 examine this result.

3.3 Hessian Estimation and Rank Recovery

We estimate the cross-Hessian rank from observed data in two steps. First, we estimate the cross-Hessian matrix. Second, we recover its rank by thresholding the singular values of the estimated matrix. For the first step, write $Y = (A^\top, B^\top)^\top \in \mathbb{R}^s$, where $s = |A| + |B|$, and let $Y^{(1)}, \ldots, Y^{(N)}$ be the corresponding i.i.d. marginal samples. Using kernel density estimation (KDE) (Chen, 2017) with kernel $\mathcal{K}$ and bandwidth $h > 0$, we estimate $p_{A,B}$ and its cross-Hessian by

$$\widehat{\mathbf{H}}_{A,B}(y) = \left[\nabla_y^2 \log \widehat{p}_h(y)\right]_{A,B} = \frac{\nabla^2_{y_A,y_B}\widehat{p}_h(y)}{\widehat{p}_h(y)} - \frac{\nabla_{y_A}\widehat{p}_h(y)\,\nabla_{y_B}\widehat{p}_h(y)^\top}{\widehat{p}_h(y)^2}, \quad \widehat{p}_h(y) = \frac{1}{Nh^s}\sum_{i=1}^{N}\mathcal{K}\left(\frac{y - Y^{(i)}}{h}\right). \tag{11}$$

The derivation of the estimated cross-Hessian in Equation (11) is provided in Appendix D.4.

For the second step, finite-sample estimation error can make singular values that are zero for the true cross-Hessian appear slightly positive. We therefore compute the singular values of $\widehat{\mathbf{H}}_{A,B}(y)$, retain only those larger than a threshold τ, and use their number as the estimated rank. We therefore use a bootstrap to calibrate a numerical threshold in practice. For each bootstrap replicate $b = 1, \ldots, B_{\text{boot}}$, we resample $\{Y^{(i)}\}_{i=1}^N$ with replacement, recompute the marginal KDE cross-Hessian $\widehat{\mathbf{H}}^{(b)}_{A,B}(y)$ using Equation (11), and have

$$e_b = \left\|\widehat{\mathbf{H}}^{(b)}_{A,B}(y) - \widehat{\mathbf{H}}_{A,B}(y)\right\|_2. \tag{12}$$

Let τ_{boot} be the q_{boot}-th empirical quantile of $\{e_b\}_{b=1}^{B_{\text{boot}}}$. Let $\sigma_1(\cdot)$ denote the largest singular value, and let $\rho \in (0, 1)$ be a prespecified, dimensionless relative tolerance held fixed throughout a given analysis. We define the relative numerical threshold $\tau_{\text{rel}} = \rho\,\sigma_1(\widehat{\mathbf{H}}_{A,B}(y))$. This relative threshold treats singular values below a ρ fraction of the largest one as numerically zero. We then have:

$$\tau = \max\{\tau_{\text{boot}}, \tau_{\text{rel}}\}, \quad \widehat{r}_\tau = \#\left\{j : \sigma_j(\widehat{\mathbf{H}}_{A,B}(y)) > \tau\right\}. \tag{13}$$

The resulting pointwise rank estimate counts every singular value above τ as nonzero and treats all remaining singular values as numerical zero. Given the estimated cross-Hessian, we determine its

rank by thresholding its singular values. Bootstrap resampling quantifies the finite-sample variability of the complete KDE cross-Hessian estimator: each resample produces a new matrix estimate, and an empirical quantile of their spectral-norm deviations gives the bootstrap threshold τ_{boot}. We combine it with the relative cutoff $\tau_{\text{rel}} = \rho\,\sigma_1(\widehat{\mathbf{H}}_{A,B}(y))$ and use $\tau = \max\{\tau_{\text{boot}}, \tau_{\text{rel}}\}$, counting every estimated singular value above τ as nonzero. This yields a data-adaptive finite-sample rank estimate while excluding numerically negligible directions. The threshold τ is a practical finite-sample calibration. Appendix D.6 establishes rank recovery for the corresponding ideal rule when the threshold is a valid upper bound on the total approximation and estimation error. For empirical HRC decisions, we estimate the rank at a finite set of interior points sampled from the observed data after excluding low-density regions and report their modal value. Evaluation-point selection and rank aggregation are detailed in Appendix D.7.

4 Application in Learning the Structure among Latent Variables

Although HRC is not restricted to pure measurement models, an end-to-end procedure requires observed groups that can be associated with individual latent variables. We therefore instantiate HRC in the standard pure one-factor measurement setting, using its rank constraints first to locate latent variables and then to test conditional independence among them. The resulting FOFC+PC procedure is an illustrative application rather than a restriction of HRC. We now use the aforementioned results for causal discovery in nonlinear measurement models [2]. Following the standard setting of one-factor measurement models (Kummerfeld & Ramsey, 2016), we consider a pure measurement model in which each observed variable has at most one latent parent, has no observed parents, and has no correlated measurement error with other observed variables. For a better understanding, the definition of pure indicator, pure cluster, and key latent is shown as:

Definition 1 (**Pure Indicator, Pure Cluster, and Key Latent**). *An observed variable X_i is called a pure indicator of a latent variable L_k if, in the measurement part of the graph, L_k is the only latent parent of X_i, X_i has no observed parents, and X_i shares no correlated measurement error with other observed variables. A set $C_k \subseteq \mathcal{X}$ is called a pure cluster of L_k if all variables in C_k are pure indicators of L_k and are not indicators of any other latent variable. In this case, L_k is called the key latent of C_k, since it is the common latent source underlying the cluster.*

Under this setting, we first locate latent variables by clustering indicators that share the same one-dimensional key latent, using the thresholded HRC rank-one constraint in place of covariance rank. We then use pure indicators as surrogates to recover the latent causal structure up to Markov equivalence. Thus, HRC serves as a building block for existing algorithms rather than a new algorithm.

4.1 Locating Latent Variables via HRC Clustering

The first task is latent-variable location, namely identifying which observed variables are associated with the same key latent. We follow the Find One Factor Clusters (FOFC in short) framework (Kummerfeld & Ramsey, 2016) and adapt it to the nonlinear setting using the HRC constraint. At a high level, the algorithm searches over triples and quartets of observed variables. A candidate triple is retained as pure only when adding another measured variable produces the rank pattern expected from a one-factor measurement structure; otherwise, the triple is rejected as mixed. In our version, this rank pattern is tested by HRC: a quartet is accepted as a nonlinear one-factor quartet if all three pairwise partitions have thresholded cross-Hessian rank one. Retained triples are then merged into maximal clusters, each of which is interpreted as the location of one latent variable.

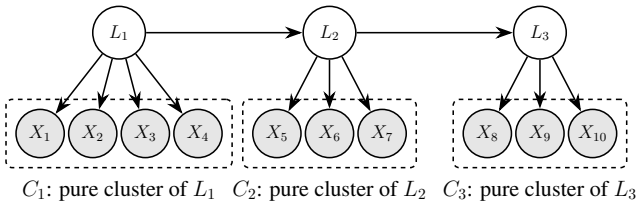


Figure 3: Illustration of HRC-based latent location and latent causal discovery. The observed variables form three pure measurement clusters C_1, C_2, C_3, corresponding to the key latents L_1, L_2, L_3, respectively. The latent variables may be causally related, e.g., $L_1 \to L_2 \to L_3$.

[2] Please refer to Appendix C for the definition of measurement model.

To make this procedure concrete, we further provide an example in Figure 3. The observed variables form three pure clusters: $C_1 = \{X_1, X_2, X_3, X_4\}$ for L_1, $C_2 = \{X_5, X_6, X_7\}$ for L_2, and $C_3 = \{X_8, X_9, X_{10}\}$ for L_3, while the latent variables may be causally related, e.g., $L_1 \to L_2 \to L_3$. For example, the triple $\{X_1, X_2, X_3\}$ is retained because adding X_4 forms an effective rank-one quartet explained by the single key latent L_1; similarly, triples inside C_2 and C_3 are retained and then merged into their corresponding maximal clusters. Theorem 2 supplies these thresholded rank-one constraints under its approximation and signal-gap conditions.

Assumption 4 (**Quartet Separation**). *For every quartet queried by FOFC, Theorem 2 applies to all three pairwise partitions. Moreover, every mixed quartet has at least one partition for which the minimum set of latent variables that d-separates the two groups has cardinality different from one.*

Discussion. This is the HRC counterpart of the standard separability conditions used to establish the correctness of FOFC (Kummerfeld & Ramsey, 2016). Intuitively, a pure one-factor quartet has rank one for all three pairwise partitions, whereas a mixed quartet is revealed by at least one partition whose separating latent dimension is not one.

Proposition 1 (**Correctness of HRC-based FOFC under Quartet Separation** (Kummerfeld & Ramsey, 2016)). *Under the population conditions for FOFC and Assumption 4, HRC-based FOFC returns the maximal pure one-factor measurement clusters with at least three observed variables.*

Discussion. Under Assumption 4, HRC provides the same quartet decisions required by FOFC, while the subsequent cluster construction is unchanged. The proof is given in Appendix D.8.

4.2 Learning Causal Structure among Latent Variables

After the first stage, each recovered cluster C_k provides measured indicators associated with a key latent L_k. The second task is to learn the causal structure among these latent variables. We use pure indicators as observed surrogates for the corresponding latent variables. To test latent conditional independence, we construct two observed groups from these indicators and check whether their thresholded cross-Hessian rank equals the dimension of the conditioning latent variables.

In general, to test whether two latent variables L_i and L_j are conditionally independent given a latent set $Q \subseteq \mathcal{L} \setminus \{L_i, L_j\}$ with $|Q| = q$, we construct two disjoint observed groups A and B from their measured indicators. Specifically, A contains one indicator of L_i, B contains one indicator of L_j, and both A and B contain one indicator for each latent variable in Q. Thus, each conditioning latent is represented on both sides of the partition, while the two endpoint latents are represented on different sides. Keeping the original rank-based procedure, we replace the covariance-rank decision with the HRC decision by checking whether the thresholded rank of $\mathbf{H}_{A,B}$ equals q.

For example, in Figure 3, the recovered clusters C_1, C_2, C_3 locate the key latents L_1, L_2, L_3, respectively. To test whether $L_1 \perp\!\!\!\perp L_3 \mid L_2$, we choose measured variables from the corresponding clusters and split them into two observed groups. Specifically, we may set $A = \{X_1, X_5\}$ and $B = \{X_8, X_6\}$, where X_1 and X_8 are indicators of the endpoint latents L_1 and L_3, while X_5 and X_6 are two indicators of the conditioning latent L_2 placed on different sides of the partition. We then determine whether the thresholded rank of $\mathbf{H}_{A,B}$ equals one, corresponding to the one-dimensional conditioning set $\{L_2\}$. The validity of this surrogate rank test is formalized as follows.

Theorem 4 (**Latent Conditional Independence via HRC**). *Assume a correctly identified pure nonlinear one-factor measurement model whose latent distribution is Markov and faithful to its latent DAG. Let L_i and L_j be distinct latent variables, and let $Q \subseteq \mathcal{L} \setminus \{L_i, L_j\}$ with $|Q| = q$. Suppose every latent variable in Q has at least two pure indicators. Construct disjoint observed groups A and B so that A contains one pure indicator of L_i, B contains one pure indicator of L_j, and each latent in Q is represented by one pure indicator in each group. Suppose Assumptions 1–3 hold for every HRC query involved and, whenever $q \geq 1$, the signal-gap condition in Equation (8) also holds. Then*

$$L_i \perp\!\!\!\perp L_j \mid Q \quad \Longleftrightarrow \quad \mathrm{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) = q. \tag{14}$$

The proof is in Appendix D.9. This theorem provides a conditional-independence test for latent variables using observed surrogate groups. The test uses the thresholded cross-Hessian rank, with the approximation tolerance and signal gap inherited from Theorem 2. We incorporate this decision rule into a PC-style constraint-based search over the recovered latent variables, which we call HRC-PC. Its oracle correctness guarantee is stated below.

Table 1: Results under different functional forms.

	HRC		Prashant et al. [2025]		Kong et al. [2023]		Silva et al. [2006]		Kummerfeld and Ramsey [2016]	
	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓
Mixed	**0.985** ± 0.052	**0.093** ± 0.314	0.437 ± 0.326	2.290 ± 0.907	0.274 ± 0.269	2.953 ± 1.011	0.224 ± 0.284	3.363 ± 1.276	0.120 ± 0.244	3.060 ± 0.852
Sin	**0.966** ± 0.088	**0.193** ± 0.507	0.410 ± 0.318	2.387 ± 0.856	0.304 ± 0.248	2.827 ± 0.941	0.102 ± 0.196	3.593 ± 1.012	0.068 ± 0.196	3.150 ± 0.810
Sinh	**0.996** ± 0.026	**0.023** ± 0.151	0.448 ± 0.314	2.293 ± 0.866	0.237 ± 0.263	3.163 ± 0.980	0.112 ± 0.217	3.626 ± 1.028	0.110 ± 0.235	3.107 ± 0.934
Softplus	**0.974** ± 0.083	**0.167** ± 0.469	0.466 ± 0.321	2.217 ± 0.905	0.194 ± 0.265	3.293 ± 0.954	0.422 ± 0.337	2.826 ± 1.511	0.101 ± 0.231	3.153 ± 0.883
Tanh	**0.950** ± 0.105	**0.303** ± 0.632	0.449 ± 0.324	2.267 ± 0.912	0.307 ± 0.252	2.803 ± 0.956	0.206 ± 0.308	3.526 ± 1.488	0.107 ± 0.234	3.063 ± 0.899
Average	**0.974** ± 0.071	**0.156** ± 0.414	0.442 ± 0.320	2.291 ± 0.889	0.263 ± 0.259	3.008 ± 0.968	0.213 ± 0.297	3.387 ± 1.313	0.101 ± 0.228	3.107 ± 0.876

Table 2: Results under different latent dimensions.

	HRC		Prashant et al. [2025]		Kong et al. [2023]		Silva et al. [2006]		Kummerfeld and Ramsey [2016]	
Dimension	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓	F1↑	SHD↓
4	**0.962** ± 0.099	**0.203** ± 0.493	0.487 ± 0.259	2.844 ± 1.505	0.502 ± 0.238	2.404 ± 1.053	0.214 ± 0.297	3.387 ± 1.313	0.254 ± 0.362	3.500 ± 1.743
6	**0.952** ± 0.085	**0.469** ± 0.790	0.417 ± 0.176	5.333 ± 1.155	0.513 ± 0.159	4.716 ± 0.958	0.216 ± 0.220	5.546 ± 1.648	0.210 ± 0.091	5.328 ± 0.165
8	**0.937** ± 0.100	**0.972** ± 1.021	0.264 ± 0.161	8.688 ± 1.899	0.184 ± 0.130	8.732 ± 1.244	0.204 ± 0.186	8.807 ± 2.853	0.191 ± 0.061	8.452 ± 0.337
10	**0.923** ± 0.110	**1.214** ± 1.542	0.154 ± 0.105	12.538 ± 1.866	0.073 ± 0.098	11.084 ± 1.307	0.163 ± 0.138	11.276 ± 3.176	0.127 ± 0.048	11.036 ± 0.433

Theorem 5 (Correctness of HRC-PC). *Suppose HRC-based FOFC correctly recovers the target latent variables and their pure measurement clusters. Assume that these latent variables are causally sufficient and that their distribution satisfies the causal Markov and faithfulness conditions with respect to DAG of the latent variables. If every HRC query made by HRC-PC agrees with the corresponding latent conditional-independence relation, then HRC-PC returns the Markov equivalence class of the DAG of the Latent variables.*

Discussion. The proof is given in Appendix D.10. Theorem 5 establishes the correctness of the graph-search stage when the required HRC decisions are correct; it does not assert consistency of the particular bootstrap and relative-threshold calibration used in the experiments.

5 Experiments

5.1 Synthetic Experiment

Setup. We evaluate HRC on synthetic nonlinear latent variable models with pure measurement variables. Each observed variable is generated from its latent parent through a nonlinear function. We consider five different nonlinear functions: $\sin$, $\sinh$, $\tanh$, softplus, and a mixed setting that assigns them independently across observed variables. We vary the latent dimensionality over $\{4, 6, 8, 10\}$. For each dimension, we use 10 random latent graphs and 10 seeds per graph. We consider sample sizes $N \in \{200, 1000, 2000, \ldots, 9000\}$. For latent variable location, we use F1 for evaluation. For latent causal discovery, we evaluate recovery of the latent Markov equivalence class using F1 and structural Hamming distance (SHD). We report the mean and standard deviation over all graph structures and random-seed combinations. Please refer to Appendix G.1 for details.

Baselines. For the latent-variable location task, all methods use the FOFC framework; our method replaces the original covariance-rank constraint with the proposed Hessian Rank Constraint, while the baseline keeps the classical covariance-rank test. For latent causal discovery, we compare with Prashant et al. (2025); Kong et al. (2023); Kummerfeld & Ramsey (2016) and Silva et al. (2006). These methods represent recent nonlinear latent causal discovery approaches as well as classical covariance-rank-based methods for linear latent variable models and one-factor measurement models.

Results. Figure 4 reports the clustering results for latent-variable location. Since both methods use the same FOFC framework, the comparison reflects the effect of the rank constraint. HRC achieves higher clustering F1 than the baseline, suggesting that HRC provides a better criterion for identifying pure measurement clusters under nonlinear mixing procedures.

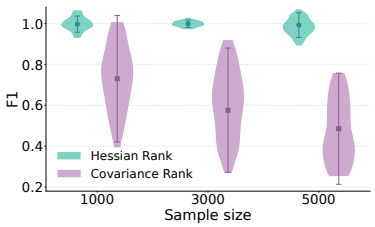


Figure 4: Clustering results between Hessian Rank constraint and covariance rank constraint.

Tables 1 and 2 report the results for causal discovery among latent variables. HRC consistently outperforms all baselines across different nonlinear mechanisms and latent dimensions. In Table 1, HRC achieves the best results under every nonlinear measurement function, with an average F1 of 0.974 and an average SHD of 0.156, while the strongest baseline only reaches

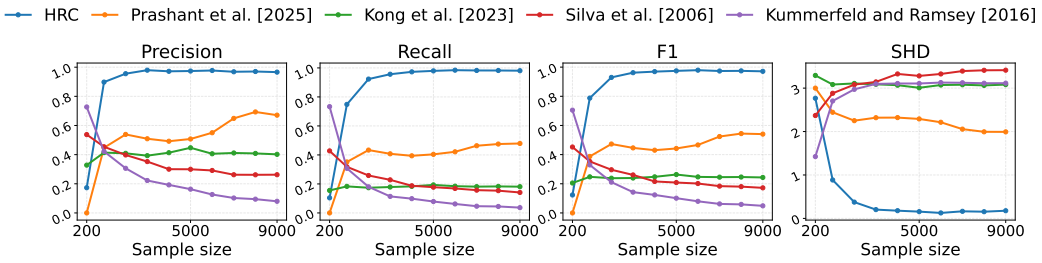


Figure 5: Causal discovery performance under different sample sizes.

0.442 F1 on average and has much larger SHD. This advantage remains stable across different nonlinearities, indicating that HRC can capture latent structural information beyond linear covariance dependence. Figure 4 further evaluates the effect of sample size. As the number of samples increases, HRC maintains high precision, recall, and F1. In contrast, the baselines remain substantially worse across the whole sample-size range, showing that HRC benefits more effectively from additional observations in nonlinear latent causal discovery. Please refer to Appendix G.3 for experiments regarding varying noise levels and sampling regions of latent variables.

5.2 Real-world Experiments

We validate HRC on the FRED-MD monthly macroeconomic database from the Federal Reserve Bank of St. Louis.[3] Following the economic grouping of the selected variables, we construct four latent factors: an interest-rate factor measured by FEDFUNDS, TB3MS, GS10, GS5, and GS10TB3Mx; a credit factor measured by BAA, BAA10YM, and COMPAPFF; an exchange-rate factor measured by TWEXMMTH, EXUSEU, and EXJPUSx; and an equity/volatility factor measured by S&P 500, NASDAQCOM, and VXOCLSX. When an exact series is unavailable in the FRED-MD vintage, we use the closest available proxy or construct the spread variable directly from the available series.

The four groups are specified from economic domain knowledge rather than learned by HRC-based FOFC. The four groups are specified from economic domain knowledge rather than learned by HRC-based FOFC. Figure 6 reports the structure returned by HRC-PC on these groups. This result is intended as a qualitative illustration. Some included spread variables are algebraically constructed from other macroeconomic series, and the monthly observations may remain temporally dependent after transformation. Consequently, this experiment is not covered by our i.i.d. density-estimation theory and should not be interpreted as evidence for the correctness of individual causal edges.

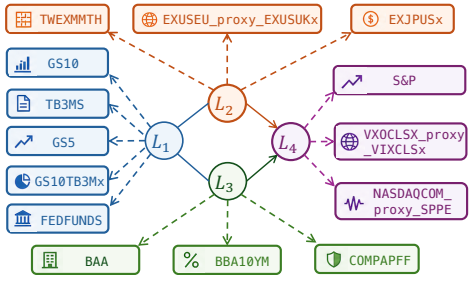


Figure 6: Illustrative HRC-PC output on four domain-specified groups from the FRED-MD dataset.

6 Conclusion

We proposed the Hessian Rank Constraint (HRC), a rank-based criterion for nonlinear latent causal discovery. HRC uses the approximately low-rank structure of the observed log-density cross-Hessian to characterize latent separation, with a controlled approximation error and an exact zero-remainder special case. Based on this criterion, we use HRC as the rank criterion within FOFC for latent-variable location and within a PC-style procedure for learning causal relations among latent variables up to Markov equivalence. Experiments on synthetic data show that HRC consistently outperforms covariance-based methods and recent nonlinear baselines under different nonlinear functions, latent dimensions, and sample sizes. The FRED-MD experiment further recovers an economically interpretable latent structure.

[3] https://www.stlouisfed.org/research/economists/mccracken/fred-databases

AI USE STATEMENT

We did not use generative AI tools for any task requiring disclosure under the ICLR 2027 AI policy; we used Claude and ChatGPT only to polish the writing, assist with auxiliary code (e.g., plotting and debugging), and edit figures, all of which were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-Gaussian and heterogeneous cases. Advances in Neural Information Processing Systems, 34:22822–22833, 2021.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In International conference on machine learning, pp. 372–407. PMLR, 2023.

Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, and Negar Kiyavash. Recursive causal structure learning in the presence of latent variables and selection bias. Advances in Neural Information Processing Systems, 34:10119–10130, 2021.

Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham Kakade. Learning linear Bayesian networks with latent variables. In International Conference on Machine Learning, pp. 249–257. PMLR, 2013.

Ricardo Baptista, Rebecca Morrison, Olivier Zahm, and Youssef Marzouk. Learning non-gaussian graphical models via hessian scores and triangular transport. Journal of Machine Learning Research, 25(85):1–46, 2024.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. Advances in Neural Information Processing Systems, 35:38319–38331, 2022.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. Advances in neural information processing systems, 32, 2019.

Ruichu Cai, Zhiyi Huang, Wei Chen, Zhifeng Hao, and Kun Zhang. Causal discovery with latent confounders based on higher-order cumulants. In International conference on machine learning, pp. 3380–3407. PMLR, 2023.

Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. The Annals of Statistics, 40(4):1935–1967, 2012. doi: 10.1214/11-AOS949.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. Biostatistics & Epidemiology, 1(1):161–187, 2017.

Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, Kun Zhang, and Ruichu Cai. Identification of linear latent variable model with arbitrary distribution. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pp. 6350–6357, 2022.

Zhengming Chen, Yewei Xia, Feng Xie, Jie Qiao, Zhifeng Hao, Ruichu Cai, and Kun Zhang. Identification of latent confounders via investigating the tensor ranks of the nonlinear observations. In Forty-second International Conference on Machine Learning, 2025.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics, pp. 294–321, 2012.

Ruifei Cui, Perry Groot, Moritz Schauer, and Tom Heskes. Learning the causal structure of copula models with latent variables. In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence, pp. 188–197, 2018.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=FhQSGhBlqv.

Patrick Forré and Joris M Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence, pp. 269–278, 2018.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. Frontiers in genetics, 10:524, 2019.

Daniel M Hausman and James Woodward. Independence, invariance and the causal Markov condition. The British journal for the philosophy of science, 50(4):521–583, 1999.

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.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. Journal of Machine Learning Research, 21(89):1–53, 2020.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. Advances in neural information processing systems, 35:5549–5561, 2022.

Aapo Hyvärinen. Independent component analysis: recent advances. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 2013.

Aapo Hyvärinen, Jarmo Hurri, and Patrik O Hoyer. Independent component analysis. In Natural image statistics: A probabilistic approach to early computational vision, pp. 151–175. Springer, 2009.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In The 22nd international conference on artificial intelligence and statistics, pp. 859–868. PMLR, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In International conference on artificial intelligence and statistics, pp. 2207–2217. PMLR, 2020.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. Advances in Neural Information Processing Systems, 34:18087–18101, 2021.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. Advances in Neural Information Processing Systems, 35:15687–15701, 2022.

Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. Advances in Neural Information Processing Systems, 36: 2010–2032, 2023.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1655–1664, 2016.

Sébastien Lachapelle and Simon Lacoste-Julien. Partial disentanglement via mechanism sparsity. arXiv preprint arXiv:2207.07732, 2022.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. Journal of Machine Learning Research, 27(71):1–90, 2026.

Te-Won Lee. Independent component analysis. In Independent component analysis: Theory and applications, pp. 27–66. Springer, 1998.

Adam Li, Amin Jaber, and Elias Bareinboim. Causal discovery from observational and interventional data across multiple environments. Advances in Neural Information Processing Systems, 36: 16942–16956, 2023.

Mingjia Li, Hong Qian, Tian-Zuo Wang, Shujun Li, Min Zhang, and Aimin Zhou. Strong and weak identifiability of optimization-based causal discovery in non-linear additive noise models. In Forty-second International Conference on Machine Learning, pp. 35753–35768, 2025.

Zhaolong Ling, Jiale Yu, Yiwen Zhang, Debo Cheng, Peng Zhou, Xingyu Wu, Bingbing Jiang, and Kui Yu. Local causal discovery without causal sufficiency. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 18737–18745, 2025.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research, 10(80):2295–2328, 2009.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. Journal of Machine Learning Research, 21(209):1–62, 2020.

Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 44(2):226–233, 1982.

Alex Markham and Moritz Grosse-Wentrup. Measurement dependence inducing latent causal models. In Conference on Uncertainty in Artificial Intelligence, pp. 590–599. PMLR, 2020.

Michael W. McCracken and Serena Ng. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics, 34(4):574–589, 2016. doi: 10.1080/07350015.2015.1086655.

Ehsan Mokhtarian, Sepehr Elahi, Sina Akbari, and Negar Kiyavash. Recursive causal discovery. Journal of Machine Learning Research, 26(61):1–65, 2025.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In Conference on Causal Learning and Reasoning, pp. 752–771. PMLR, 2023.

Ignavier Ng, Xinshuai Dong, Haoyue Dai, Biwei Huang, Peter Spirtes, and Kun Zhang. Score-based causal discovery of latent variable causal models. In Forty-first International Conference on Machine Learning, 2024.

Judea Pearl. Causality. Cambridge University Press, 2009.

Parjanya Prajakta Prashant, Ignavier Ng, Kun Zhang, and Biwei Huang. Differentiable causal discovery for latent hierarchical causal models. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=Bp0HBaMNRl.

Abbavaram Gowtham Reddy, Benin Godfrey L, and Vineeth N Balasubramanian. On causally disentangled representations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pp. 8089–8097, 2022.

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 Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 18741–18753, 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634, 2021.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. Journal of Machine Learning Research, 23 (241):1–55, 2022.

Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-Gaussian acyclic models for latent factors. Neurocomputing, 72(7-9):2024–2027, 2009.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. Journal of Machine Learning Research, 7(Feb):191–246, 2006.

Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. Journal of Machine Learning Research, 19(66):1–71, 2018.

Charles Spearman. Pearson's contribution to the theory of two factors. British Journal of Psychology. General Section, 19(1):95–101, 1928.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In Applied informatics, volume 3, pp. 3. Springer, 2016.

Peter Spirtes, Clark Glymour, and Richard Scheines. Discovery algorithms without causal sufficiency. In Causation, Prediction, and Search, pp. 163–200. Springer, 1993.

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

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. Advances in Neural Information Processing Systems, 36:34465–34492, 2023.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. The Annals of Statistics, 38(3):1665–1685, 2010.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. arXiv preprint arXiv:2301.08230, 2023.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. Advances in neural information processing systems, 33:14891–14902, 2020.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-Gaussian latent hierarchical structure. In International Conference on Machine Learning, pp. 24370–24387. PMLR, 2022.

Feng Xie, Yan Zeng, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Causal discovery of 1-factor measurement models in linear latent variable models with arbitrary noise distributions. Neurocomputing, 526:48–61, 2023.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius Von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In Forty-first International Conference on Machine Learning, 2024.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. Advances in Neural Information Processing Systems, 35:26492–26503, 2022.

Mingzhang Yin, Yixin Wang, and David M Blei. Optimization-based causal estimation from heterogeneous environments. Journal of Machine Learning Research, 25(168):1–44, 2024.

Naiwen Ying, Ping Zhang, Shanshan Luo, and Wang Miao. A generalized tetrad constraint for testing conditional independence given a latent variable. arXiv preprint arXiv:2504.14173, 2025.

Yan Zeng, Shohei Shimizu, Ruichu Cai, Feng Xie, Michio Yamamoto, and Zhifeng Hao. Causal discovery with multi-domain LiNGAM for latent factors. In Causal Analysis Workshop Series, pp. 1–4. PMLR, 2021.

Zhalama, Jiji Zhang, and Wolfgang Mayer. Weakening faithfulness: some heuristic causal discovery algorithms. International journal of data science and analytics, 3(2):93–104, 2017.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artificial Intelligence, 172(16-17):1873–1896, 2008.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, pp. 804–813, 2011.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In Forty-first International Conference on Machine Learning, 2024.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In The Eighth International Conference on Learning Representations, 2020.

Supplement to

"Learning the Structure of Nonlinear Latent Variable Models via Hessian Rank Constraint"

Appendix organization:

A Related Works **16**

A.1 Causal discovery . . . 16

A.2 Causal discovery among latent variables . . . 16

A.3 Causal representation learning . . . 16

B Notation Description **17**

C Definitions **18**

D Proof **18**

D.1 Proof of Theorem 1 . . . 18

D.2 Proof of Corollary 2.1 . . . 19

D.3 Proof of Theorem 2 . . . 19

D.4 Derivation of the KDE Cross-Hessian Formula . . . 20

D.5 Proof of Theorem 3 . . . 20

D.6 KDE Cross-Hessian Estimation and Rank Recovery . . . 23

D.7 Evaluation Points and Rank Aggregation . . . 24

D.8 Proof of Proposition 1 . . . 25

D.9 Proof of Theorem 4 . . . 25

D.10 Proof of Theorem 5 . . . 25

E Further Discussion on Assumptions **26**

E.1 Discussion on Pointwise Non-degeneracy . . . 26

E.1.1 Full-Dimensional Conditional Variation . . . 26

E.1.2 Zero Measurements and Informative Observations . . . 27

E.2 Exact Affinity of Conditional Derivatives . . . 29

F Details of FOFC Framework **29**

G Experiment **30**

G.1 Synthetic Dataset . . . 30

G.2 Real-world Dataset . . . 31

G.3 More Experiment Results . . . 31

H Broader Impact **31**

I Limitations **32**

A Related Works

A.1 Causal discovery

Causal discovery studies how causal relations can be inferred from observational or interventional data, and has been widely investigated in machine learning, statistics, and scientific discovery (Glymour et al., 2019; Huang et al., 2020; Zhu et al., 2020; Pearl, 2009). Most classical methods are built on graphical assumptions such as the causal Markov condition (Hausman & Woodward, 1999) and faithfulness (Zhalama et al., 2017). Under these assumptions, constraint-based methods use conditional independence tests to recover a Markov equivalence class of DAGs, with PC-style algorithms being representative examples (Glymour et al., 2019; Spirtes & Zhang, 2016; Zhang et al., 2011). Other methods formulate causal discovery as score-based search or continuous optimization over DAGs (Huang et al., 2018; Ng et al., 2024; Montagna et al., 2023; Li et al., 2025; Yin et al., 2024). A common limitation of many such methods is the causal sufficiency assumption, which excludes unobserved common causes among measured variables (Spirtes et al., 1993; Ling et al., 2025; Cai et al., 2023; Forré & Mooij, 2018). In practice, this assumption is often too strong, since latent variables can induce dependencies among observed variables and make the observed causal structure ambiguous. FCI and its variants address this issue by allowing latent confounders and representing the remaining uncertainty over causal structures (Spirtes et al., 2000; Pearl, 2009; Zhang, 2008; Colombo et al., 2012; Akbari et al., 2021; Mokhtarian et al., 2025). However, these methods primarily describe causal relations among observed variables. They do not, in general, identify the latent variables themselves or recover causal relations among them, which is the focus of latent-variable causal discovery.

A.2 Causal discovery among latent variables

A central goal in latent-variable causal discovery is to recover causal relations among latent variables (Huang et al., 2022; Dong et al., 2024). This problem is often studied in measurement-model settings (Xie et al., 2023; Markham & Grosse-Wentrup, 2020), where observed variables serve as indicators or descendants of latent causes. Beyond detecting latent confounding, one aims to locate the latent variables, determine their dimensions, and learn their causal structure. Since this structure is not identifiable without additional assumptions, existing methods exploit various statistical signatures. Rank-based methods use tetrad constraints or low-rank cross-covariance matrices to identify pure measurement clusters or infer the number of latent variables that d-separate two measured sets (Silva et al., 2006; Kummerfeld & Ramsey, 2016; Huang et al., 2022; Xie et al., 2022; Dong et al., 2024). Other methods rely on higher-order moments (Shimizu et al., 2009; Cai et al., 2019; Xie et al., 2020; Adams et al., 2021; Chen et al., 2022), matrix decomposition (Anandkumar et al., 2013; Chen et al., 2025), copula formulations (Cui et al., 2018), multi-domain information (Zeng et al., 2021; Li et al., 2023; Sturma et al., 2023; Zhang et al., 2024), or mixture-based oracles (Kivva et al., 2021; 2022). Despite their identifiability guarantees, many of these methods assume linear measurements, discrete latent variables, or other restrictive structures, which may not hold when observations are generated through nonlinear mechanisms.

Recent work has extended latent causal discovery to nonlinear measurement models. Kong et al. (2023) study nonlinear latent hierarchical structures but require latent variables to be deterministic functions of measured variables. Prashant et al. (2025) use the Jacobian rank of conditional expectation mappings to identify separating latent sets and learn the structure through differentiable optimization. Recent work has also generalized tetrad constraints to nonlinear models for testing conditional independence given a single latent variable (Ying et al., 2025), whereas our goal is to identify latent dimensions, locate latent variables, and recover their causal relations. Our work instead uses the cross-Hessian rank of the observed log-density, providing a nonlinear rank criterion that can be incorporated into existing clustering and PC-style latent discovery procedures.

Several adjacent lines of work also connect density derivatives with graphical structure. In latent Gaussian graphical models, marginalizing latent variables yields a sparse-minus-low-rank precision matrix, with the low-rank component induced by latent variables (Chandrasekaran et al., 2012); in the Gaussian special case, our cross-Hessian is a block of the marginal precision matrix. For continuous non-Gaussian graphical models, Spantini et al. (2018) connect Markov structure with sparse and low-dimensional transport couplings, while Hessian scores combined with transport maps have been used for graph learning (Baptista et al., 2024). SCORE exploits score and Hessian variation for causal ordering in nonlinear additive-noise models (Rolland et al., 2022), while the nonparanormal extends Gaussian graph estimation through componentwise monotone transformations (Liu et al., 2009). HRC differs by using the rank of a cross-Hessian block to characterize the dimension of latent variables that d-separate observed groups in a directed latent-variable model.

A.3 Causal representation learning

Our work is also related to causal representation learning, which aims to recover latent variables from observed data (Schölkopf et al., 2021; Ahuja et al., 2023; Brehmer et al., 2022). Many causal representation learning methods build on ideas from nonlinear independent component analysis (ICA) (Khemakhem et al., 2020; Hyvarinen et al., 2019), where identifiability of latent factors is obtained under additional assumptions such

as auxiliary variables (Khemakhem et al., 2020; Zhang et al., 2024), temporal dependence (Hyvarinen et al., 2019), sparsity (Xu et al., 2024; Lachapelle et al., 2026; Lachapelle & Lacoste-Julien, 2022), or nonstationarity (Yao et al., 2022). Other CRL approaches use interventions (Ahuja et al., 2023; Varici et al., 2023), multiple environments (Zhang et al., 2024), or weak supervision (Shen et al., 2022) to learn causally meaningful representations. These works share our interest in latent variables behind observations, but they mainly focus on representation recovery or disentanglement (Locatello et al., 2020; Reddy et al., 2022). In contrast, our goal is to locate latent variables and recover causal relations among them from observational data. Moreover, while ICA-based approaches often assume independent latent sources (Lee, 1998; Hyvärinen, 2013; Hyvärinen et al., 2009), our setting allows latent variables to be causally dependent.

B Notation Description

This section collects the main notations used in the theorem proofs for clarity and consistency.

Table A3: List of main notations, explanations, and corresponding values.

Notation	Explanation	Support / Value
Notation Style		
Uppercase Roman letters	random variables or random vectors	—
Lowercase Roman letters	realizations of the corresponding random variables or random vectors	—
Calligraphic letters	sets of nodes in the graph	—
Bold uppercase letters	matrices	—
P	probability distributions or probability measures	—
p	probability density functions	—
Graph and Variables		
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	directed acyclic graph with node set $\mathcal{V}$ and edge set $\mathcal{E}$	$\mathcal{V} = \mathcal{L} \cup \mathcal{X}$
$\mathcal{L} = \{L_i\}_{i=1}^{n}$	set of all latent nodes	$\lvert\mathcal{L}\rvert = n$
$\mathcal{X} = \{X_j\}_{j=1}^{m}$	set of all observed nodes	$\lvert\mathcal{X}\rvert = m$
n, m	numbers of latent and observed variables, respectively	$n, m \in \mathbb{N}^+$
A, B	two disjoint groups of observed variables	$A, B \subseteq \mathcal{X}$, $A \cap B = \emptyset$
$\mathcal{Z}$; Z, z	specified latent d-separator; associated random vector and realization	$\mathcal{Z} \subseteq \mathcal{L}$, $\lvert\mathcal{Z}\rvert = d$; $Z, z \in \mathbb{R}^d$
d^*	cardinality of a minimum set of latent variables that d-separates A and B	$d^* = \lvert\mathcal{Z}^*\rvert$
Q	conditioning set of latent nodes	$Q \subseteq \mathcal{L} \setminus \{L_i, L_j\}$, $\lvert Q\rvert = q$
C_k	pure cluster corresponding to latent variable L_k	$C_k \subseteq \mathcal{X}$
Structural Model		
$L_i = f_{L_i}(\mathrm{Pa}(L_i), \epsilon_{L_i})$	structural equation for latent variable L_i	$\mathrm{Pa}(L_i) \subseteq \mathcal{L}$
$X_j = f_{X_j}(\mathrm{Pa}(X_j), \epsilon_{X_j})$	structural equation for observed variable X_j	$\mathrm{Pa}(X_j) \subseteq \mathcal{V}$
$\mathrm{Pa}(\cdot)$	parent set of a variable in $\mathcal{G}$	/
$\epsilon_{L_i}, \epsilon_{X_j}$	independent noise terms of L_i and X_j	/
$p_{A,B}(a, b)$, $p_{A,B\mid Z}(a, b \mid z)$, $p_{Z\mid A,B}(z \mid a, b)$	marginal, conditional, and separator-posterior densities	/
Scores		
$\mathbf{s}_A(a, z)$, $\mathbf{s}_B(b, z)$	conditional score functions of A and B	$\in \mathbb{R}^{\lvert A\rvert}$ and $\mathbb{R}^{\lvert B\rvert}$
$\mathbf{u}_A(a), \mathbf{u}_B(b)$	intercept terms in Assumption 2	$\mathbf{u}_A \in \mathbb{R}^{\lvert A\rvert}$, $\mathbf{u}_B \in \mathbb{R}^{\lvert B\rvert}$
$\mathbf{V}_A(a), \mathbf{V}_B(b)$	slope matrices in the approximate affine decomposition	$\mathbf{V}_A \in \mathbb{R}^{\lvert A\rvert \times d}$, $\mathbf{V}_B \in \mathbb{R}^{\lvert B\rvert \times d}$
$\mathbf{r}_A(a, z), \mathbf{r}_B(b, z)$	conditional-score remainder terms	controlled by $\varepsilon_A(a, b), \varepsilon_B(a, b)$
$\mathrm{Cov}(Z \mid A = a, B = b)$	posterior covariance of a specified separator	$\in \mathbb{R}^{d \times d}$
$\mathbb{E}_{Z\mid A=a,B=b}[\cdot]$	posterior expectation with respect to $P_{Z\mid A=a,B=b}$	/
HRC		
$\mathbf{H}_{A,B}$	cross-Hessian function of the marginal log-density	Equation (2)
$\mathbf{H}^{\mathrm{aff}}_{A,B}$	affine benchmark induced by a latent d-separating set	Equation (5)
$\mathbf{E}_{A,B}$	cross-Hessian approximation error	$\mathbf{H}_{A,B} - \mathbf{H}^{\mathrm{aff}}_{A,B}$
$\mathrm{rank}_\tau(\mathbf{H}_{A,B})$	pointwise thresholded-rank function of the cross-Hessian	number of valid singular values exceeding τ
$\mathrm{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) = d^*$	robust HRC equality under Theorem 2	thresholded rank equals the minimum separator cardinality
$\widehat{\mathbf{H}}_{A,B}(y)$	empirical marginal cross-Hessian estimated from Y-samples	Equation (11)
$E(y)$	cross-Hessian estimation error	$\widehat{\mathbf{H}}_{A,B}(y) - \mathbf{H}_{A,B}(y)$
ε	high-probability error threshold	$\lVert E(y)\rVert_2 \le \varepsilon$
$\sigma_j(\cdot), \mathrm{SVD}_\varepsilon(\cdot)$	singular values and hard-thresholded SVD operator	retain singular values $> \alpha$

C DEFINITIONS

Definition 2 (Measurement Model)**.** *Given a latent variable model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \mathcal{L} \cup \mathcal{X}$, the subgraph containing all nodes in $\mathcal{V}$ and exactly those edges directed into the observed nodes in $\mathcal{X}$ is called the* ***measurement model*** *of $\mathcal{G}$ (Silva et al., 2006).*

D PROOF

D.1 PROOF OF THEOREM 1

Theorem 1 (Approximate Cross-Hessian Rank under d-Separation). Let A, B be disjoint observed groups and let $Z \in \mathbb{R}^d$ be associated with latent variables that d-separate them. Under Assumptions 1 and 2, define $\mathbf{H}^{\text{aff}}_{A,B} := \mathbf{V}_A(A)\,\mathrm{Cov}(Z \mid A, B)\mathbf{V}_B(B)^\top$ and let $\varepsilon_{A,B}$ be the tolerance defined in Theorem 1. Then

$$\begin{aligned} \mathbf{H}_{A,B} &= \mathbf{H}^{\text{aff}}_{A,B} + \mathbf{E}_{A,B}, \\ \|\mathbf{E}_{A,B}\|_2 &\le \varepsilon_{A,B}, \\ \sigma_j(\mathbf{H}_{A,B}) &\le \varepsilon_{A,B} \quad (j > d), \\ \mathrm{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) &\le d. \end{aligned}$$

Proof. Fix $y = (y_A^\top, y_B^\top)^\top$ with $p_{A,B}(y) > 0$ and write $q_y = P_{Z \mid A=y_A, B=y_B}$. Assumption 1 justifies all differentiations below. Because the model is Markov with respect to $\mathcal{G}$ and the chosen latent variables d-separate A and B, we have

$$p_{A,B|Z}(y_A, y_B \mid z) = p_{A|Z}(y_A \mid z) p_{B|Z}(y_B \mid z). \tag{A1}$$

Let $\mathbf{s}_A(y_A, z) = \nabla_{y_A} \log p_{A|Z}(y_A \mid z)$ and define $\mathbf{s}_B(y_B, z)$ analogously. Differentiating the latent marginalization once in each observed group and using Equation (A1) gives the score-covariance identity

$$\mathbf{H}_{A,B}(y) = \mathrm{Cov}_{q_y}(\mathbf{s}_A(y_A, Z), \mathbf{s}_B(y_B, Z))\,. \tag{A2}$$

For completeness, the mixed conditional derivative vanishes because $\log p_{A,B|Z} = \log p_{A|Z} + \log p_{B|Z}$; the remaining two terms obtained by differentiating $\log p_{A,B}$ are respectively the posterior expectation of the score outer product and the product of the posterior score expectations, whose difference is the covariance above.

Under Assumption 2, at the fixed point y,

$$\mathbf{s}_A = \mathbf{u}_A + \mathbf{V}_A Z + \mathbf{r}_A, \qquad \mathbf{s}_B = \mathbf{u}_B + \mathbf{V}_B Z + \mathbf{r}_B.$$

Let $Z_c := Z - \mathbb{E}_{q_y} Z$ and $\bar{\mathbf{r}}_G := \mathbf{r}_G - \mathbb{E}_{q_y}\mathbf{r}_G$. Centering the scores removes their intercepts and yields

$$\mathbf{s}_A - \mathbb{E}_{q_y}\mathbf{s}_A = \mathbf{V}_A Z_c + \bar{\mathbf{r}}_A, \qquad \mathbf{s}_B - \mathbb{E}_{q_y}\mathbf{s}_B = \mathbf{V}_B Z_c + \bar{\mathbf{r}}_B.$$

Substitution into Equation (A2) gives

$$\begin{aligned} \mathbf{H}_{A,B}(y) = \mathbf{V}_A\,\mathrm{Cov}_{q_y}(Z)\mathbf{V}_B^\top + \mathbf{V}_A\,\mathrm{Cov}_{q_y}(Z, \mathbf{r}_B) \\ + \mathrm{Cov}_{q_y}(\mathbf{r}_A, Z)\mathbf{V}_B^\top + \mathrm{Cov}_{q_y}(\mathbf{r}_A, \mathbf{r}_B). \end{aligned} \tag{A3}$$

The first term is $\mathbf{H}^{\text{aff}}_{A,B}(y)$ and the last three terms form $\mathbf{E}_{A,B}(y)$.

For square-integrable random vectors U, V, Cauchy–Schwarz implies

$$\|\,\mathrm{Cov}(U, V)\|_2 \le \left(\mathbb{E}\|U - \mathbb{E}U\|^2\right)^{1/2} \left(\mathbb{E}\|V - \mathbb{E}V\|^2\right)^{1/2}.$$

Using Equation (4) and the definition of ν_Z therefore gives

$$\|\,\mathrm{Cov}_{q_y}(Z, \mathbf{r}_B)\|_2 \le \nu_Z \varepsilon_B, \quad \|\,\mathrm{Cov}_{q_y}(\mathbf{r}_A, Z)\|_2 \le \nu_Z \varepsilon_A, \quad \|\,\mathrm{Cov}_{q_y}(\mathbf{r}_A, \mathbf{r}_B)\|_2 \le \varepsilon_A \varepsilon_B.$$

Applying submultiplicativity to Equation (A3) proves $\|\mathbf{E}_{A,B}(y)\|_2 \le \varepsilon_{A,B}(y)$. Finally, $\mathrm{rank}(\mathbf{H}^{\text{aff}}_{A,B}(y)) \le d$. Hence, for every valid singular-value index $j > d$, $\sigma_j(\mathbf{H}^{\text{aff}}_{A,B}(y)) = 0$, and Weyl's inequality gives

$$\sigma_j(\mathbf{H}_{A,B}(y)) \le \|\mathbf{E}_{A,B}(y)\|_2 \le \varepsilon_{A,B}(y).$$

Equivalently, $\mathrm{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) \le d$. The claim holds wherever the stated regularity and remainder bounds hold. When both remainders vanish, $\varepsilon_{A,B} = 0$ and the original exact rank upper bound is recovered. □

Relation to the linear Gaussian case. For non-degenerate jointly Gaussian distributions, the block-inverse calculation in the proof of Corollary 2.1 shows that cross-Hessian rank equals cross-covariance rank. Classical algebraic results then imply that, in linear Gaussian structural equation models, additional rank loss occurs only on exceptional measure-zero parameter subsets (Sullivant et al., 2010). This comparison motivates the explicit non-degeneracy condition used for exact rank recovery; no genericity claim is made for unrestricted nonlinear models.

D.2 Proof of Corollary 2.1

Corollary 2.1 (Linear Gaussian Special Case). Let A and B be disjoint observed groups. Let $\mathcal{Z}^*$ be a minimum latent d-separator between them, write $d^* = |\mathcal{Z}^*|$, and let $Z^* \in \mathbb{R}^{d^*}$ be its associated random vector. Suppose the full structural causal model is linear Gaussian, the joint distribution of (A, B) is non-degenerate, and Assumptions 1 and 3 hold. Then

$$\operatorname{rank}(\mathbf{H}_{A,B}) = d^*.$$

Proof. Under the stated assumptions, $Y = (A^\top, B^\top)^\top$ is a non-degenerate jointly Gaussian vector. Let Σ denote its covariance matrix and let $\Theta = \Sigma^{-1}$ be its precision matrix. Since the Gaussian log-density has Hessian $-\Theta$, its cross-Hessian block is $\mathbf{H}_{A,B} = -\Theta_{A,B}$. The block-inverse identity gives

$$\mathbf{H}_{A,B} = \Sigma_{A,A}^{-1}\Sigma_{A,B}\big(\Sigma_{B,B} - \Sigma_{B,A}\Sigma_{A,A}^{-1}\Sigma_{A,B}\big)^{-1}.$$

Because Σ is positive definite, the two matrices multiplying $\Sigma_{A,B}$ are invertible, and hence

$$\operatorname{rank}(\mathbf{H}_{A,B}) = \operatorname{rank}(\Sigma_{A,B}).$$

This establishes the equivalence between covariance-rank and cross-Hessian-rank constraints in the non-degenerate Gaussian case.

To verify the affine-derivative condition, write the conditional Gaussian regressions induced by the linear Gaussian model as

$$A \mid Z^* = z \sim \mathcal{N}(\mu_A + \Lambda_A z, \Omega_A), \qquad B \mid Z^* = z \sim \mathcal{N}(\mu_B + \Lambda_B z, \Omega_B),$$

where Ω_A and Ω_B are the positive-definite conditional residual covariance matrices for groups A and B, respectively. Their conditional scores satisfy

$$\nabla_a \log p_{A|Z^*}(a \mid z) = -\Omega_A^{-1}(a-\mu_A) + \Omega_A^{-1}\Lambda_A z, \qquad \nabla_b \log p_{B|Z^*}(b \mid z) = -\Omega_B^{-1}(b-\mu_B) + \Omega_B^{-1}\Lambda_B z.$$

Thus Assumption 2 holds with $\mathbf{V}_A = \Omega_A^{-1}\Lambda_A$, $\mathbf{V}_B = \Omega_B^{-1}\Lambda_B$, and $\mathbf{r}_A = \mathbf{r}_B = 0$. Hence $\varepsilon_{A,B} = 0$, and Theorem 1 gives

$$\mathbf{H}_{A,B} = \mathbf{V}_A \operatorname{Cov}(Z^* \mid A, B)\, \mathbf{V}_B^\top.$$

Assumption 3 makes the middle factor positive definite almost surely and the two slope matrices full column rank d^*. Therefore the cross-Hessian function has rank d^*. □

D.3 Proof of Theorem 2

Theorem 2 (Robust Hessian Rank Constraint). Let $\mathcal{Z}^*$ be a minimum-cardinality set of latent variables that d-separates A and B, with $d^* = |\mathcal{Z}^*|$. Under Assumptions 1–3, suppose that $\sigma_{d^*}(\mathbf{H}^{\text{aff}}_{A,B}) > 2\varepsilon_{A,B}$ when $d^* \geq 1$; no signal-gap condition is required when $d^* = 0$. Then

$$\operatorname{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) = d^*.$$

Proof. If $d^* = 0$, the empty set d-separates A and B, so $p_{A,B} = p_A p_B$ and $\mathbf{H}_{A,B} = 0$. Since $\varepsilon_{A,B} \geq 0$, no singular value exceeds the threshold and $\operatorname{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) = 0$.

Now suppose $d^* \geq 1$. Apply Theorem 1 to the minimum d-separating set $\mathcal{Z}^*$. It gives

$$\mathbf{H}_{A,B} = \mathbf{H}^{\text{aff}}_{A,B} + \mathbf{E}_{A,B}, \qquad \|\mathbf{E}_{A,B}\|_2 \leq \varepsilon_{A,B}.$$

Assumption 3 makes $\operatorname{Cov}(Z^* \mid A, B)$ positive definite and both slope matrices full column rank d^*. Hence $\operatorname{rank}(\mathbf{H}^{\text{aff}}_{A,B}) = d^*$. For every valid singular-value index $j > d^*$, Weyl's inequality yields

$$\sigma_j(\mathbf{H}_{A,B}) \leq \|\mathbf{E}_{A,B}\|_2 \leq \varepsilon_{A,B}.$$

For the smallest nonzero benchmark singular value, the same inequality and the signal-gap condition give

$$\sigma_{d^*}(\mathbf{H}_{A,B}) \geq \sigma_{d^*}(\mathbf{H}^{\text{aff}}_{A,B}) - \varepsilon_{A,B} > \varepsilon_{A,B}.$$

Thus exactly the first d^* singular values exceed the threshold $\varepsilon_{A,B}$, proving the stated thresholded-rank equality. In the zero-remainder case, $\varepsilon_{A,B} = 0$ and this reduces to the exact algebraic-rank result. □

D.4 Derivation of the KDE Cross-Hessian Formula

Write $y = (y_A^\top, y_B^\top)^\top$ and let $q(y) := \widehat{p}_h(y)$. Fix an evaluation point at which $q(y) > 0$. For coordinates $y_{A,i}$ and $y_{B,j}$,

$$\frac{\partial \log q(y)}{\partial y_{B,j}} = \frac{1}{q(y)} \frac{\partial q(y)}{\partial y_{B,j}}.$$

Differentiating this identity with respect to $y_{A,i}$ and applying the product rule gives

$$\begin{aligned}\frac{\partial^2 \log q(y)}{\partial y_{A,i}\, \partial y_{B,j}} &= \frac{\partial}{\partial y_{A,i}} \left[q(y)^{-1} \frac{\partial q(y)}{\partial y_{B,j}} \right] \\ &= q(y)^{-1} \frac{\partial^2 q(y)}{\partial y_{A,i}\, \partial y_{B,j}} - q(y)^{-2} \frac{\partial q(y)}{\partial y_{A,i}} \frac{\partial q(y)}{\partial y_{B,j}}.\end{aligned}$$

The (i, j) entry of $\nabla_{y_A} \nabla_{y_B}^\top \log q(y)$ is the left-hand side. Stacking the identities over all i and j therefore yields

$$\nabla_{y_A} \nabla_{y_B}^\top \log q(y) = \frac{\nabla^2_{y_A, y_B} q(y)}{q(y)} - \frac{\nabla_{y_A} q(y)\, \nabla_{y_B} q(y)^\top}{q(y)^2}.$$

To make the dependence on the KDE explicit, define $u^{(n)} = (y - Y^{(n)})/h$. If the kernel is twice differentiable, differentiation of the finite sum gives

$$\begin{aligned}\nabla_{y_A} \widehat{p}_h(y) &= \frac{1}{N h^{s+1}} \sum_{n=1}^{N} \nabla_{y_A} \mathcal{K}(u^{(n)}), \\ \nabla_{y_B} \widehat{p}_h(y) &= \frac{1}{N h^{s+1}} \sum_{n=1}^{N} \nabla_{y_B} \mathcal{K}(u^{(n)}), \\ \nabla^2_{y_A, y_B} \widehat{p}_h(y) &= \frac{1}{N h^{s+2}} \sum_{n=1}^{N} \nabla^2_{y_A, y_B} \mathcal{K}(u^{(n)}),\end{aligned}$$

where the derivatives on the right-hand side are taken with respect to the corresponding coordinate blocks of the kernel argument. Substituting these derivatives into the matrix identity above with $q = \widehat{p}_h$ gives Equation (11).

D.5 Proof of Theorem 3

Lemma A1 (Taylor remainder under Hessian Lipschitz continuity)**.** *Under the conditions of Theorem 3, let* $\mathbf{v}_{G,i}^\top$ *be the* i*th row of* $\mathbf{V}_G$*, define*

$$r_{G,i}(z) := s_{G,i}(y_G, z) - s_{G,i}(y_G, z_0) - \mathbf{v}_{G,i}^\top (z - z_0), \tag{A4}$$

and write $\widetilde{C}_{G,i} := \kappa_{G,i} + \delta L_s/3$*. Then, for every* $z \in \mathcal{B}(z_0, \delta)$*,*

$$|r_{G,i}(z)| \le \frac{1}{2} \widetilde{C}_{G,i} \|z - z_0\|_2^2. \tag{A5}$$

Proof. Fix G and i, and write $h = z - z_0$. Taylor's theorem with integral remainder gives

$$r_{G,i}(z) = \int_0^1 (1 - t) h^\top \nabla_z^2 s_{G,i}(y_G, z_0 + th) h \, \mathrm{d}t. \tag{A6}$$

Because $\mathcal{B}(z_0, \delta)$ is convex, the entire segment $z_0 + th$, $t \in [0, 1]$, remains in the region covered by the Hessian Lipschitz condition. Hence

$$\begin{aligned}\|\nabla_z^2 s_{G,i}(y_G, z_0 + th)\|_2 &\le \|\nabla_z^2 s_{G,i}(y_G, z_0)\|_2 \\ &\quad + \|\nabla_z^2 s_{G,i}(y_G, z_0 + th) - \nabla_z^2 s_{G,i}(y_G, z_0)\|_2 \\ &\le \kappa_{G,i} + t L_s \|h\|_2.\end{aligned} \tag{A7}$$

Substitution into Equation (A6) gives

$$\begin{aligned}|r_{G,i}(z)| &\le \|h\|_2^2 \int_0^1 (1 - t)\big(\kappa_{G,i} + t L_s \|h\|_2\big) \, \mathrm{d}t \\ &= \frac{1}{2} \kappa_{G,i} \|h\|_2^2 + \frac{1}{6} L_s \|h\|_2^3 \\ &\le \frac{1}{2} \left(\kappa_{G,i} + \frac{\delta L_s}{3} \right) \|h\|_2^2 = \frac{1}{2} \widetilde{C}_{G,i} \|h\|_2^2,\end{aligned} \tag{A8}$$

where the last inequality uses $\|h\|_2 \le \delta$. □

Lemma A2 (Covariance bound on a posterior ball)**.** *If a random vector Z is supported on $\mathcal{B}(z_0, \delta)$, then*

$$\|\operatorname{Cov}(Z)\|_2 \le \delta^2. \tag{A9}$$

Equivalently, $\operatorname{Var}(d^\top Z) \le \delta^2$ *for every unit vector d.*

Proof. For any unit vector d, the scalar projection $d^\top Z$ lies in an interval of length at most 2δ. Indeed, for any $z_1, z_2 \in \mathcal{B}(z_0, \delta)$,

$$|d^\top z_1 - d^\top z_2| \le \|d\|_2 \|z_1 - z_2\|_2 \le 2\delta. \tag{A10}$$

For a scalar random variable W supported on an interval $[u, v]$, Popoviciu's inequality gives $\operatorname{Var}(W) \le (v-u)^2/4$. Hence $\operatorname{Var}(d^\top Z) \le \delta^2$. Since $\operatorname{Cov}(Z)$ is positive semidefinite, the Rayleigh–Ritz characterization gives

$$\|\operatorname{Cov}(Z)\|_2 = \sup_{\|d\|_2=1} d^\top \operatorname{Cov}(Z) d = \sup_{\|d\|_2=1} \operatorname{Var}(d^\top Z) \le \delta^2. \tag{A11}$$

□

Lemma A3 (Spectral and Frobenius norms)**.** *For every matrix D, $\|D\|_2 \le \|D\|_F$.*

Proof. Let $\sigma_1(D) \ge \sigma_2(D) \ge \cdots \ge 0$ be the singular values of D. Then

$$\|D\|_2 = \sigma_1(D) \le \left(\sum_k \sigma_k(D)^2 \right)^{1/2} = \|D\|_F. \tag{A12}$$

□

Proof of Theorem 3. The proof consists of three steps.

Step 1: Decompose the cross-Hessian approximation error. Let $\ell(y, z) := \log p_{A,B|Z}(y \mid z)$ and $Q = P_{Z|A=y_A, B=y_B}$. Assumption 1 permits differentiation through the latent marginalization. First,

$$\begin{aligned} \nabla_{y_A} \log p_{A,B}(y) &= \frac{1}{p_{A,B}(y)} \int \nabla_{y_A} p_{A,B|Z}(y \mid z) p_Z(z) \, \mathrm{d}z \\ &= \int \nabla_{y_A} \ell(y, z) \frac{p_{A,B|Z}(y \mid z) p_Z(z)}{p_{A,B}(y)} \, \mathrm{d}z \\ &= \mathbb{E}_Q[\nabla_{y_A} \ell(y, Z)]. \end{aligned} \tag{A13}$$

Differentiating this identity with respect to y_B gives the second-order Fisher identity

$$\begin{aligned} \mathbf{H}_{A,B}(y) = \mathbb{E}_Q[\nabla_{y_A} \nabla_{y_B}^\top \ell(y, Z)] \\ + \operatorname{Cov}_Q(\nabla_{y_A} \ell(y, Z), \nabla_{y_B} \ell(y, Z)) . \end{aligned} \tag{A14}$$

Because Z d-separates A and B, $\ell(y, z) = \log p_{A|Z}(y_A \mid z) + \log p_{B|Z}(y_B \mid z)$. The mixed derivative in Equation (A14) is therefore zero, while the two gradients are the conditional scores. Thus

$$\mathbf{H}_{A,B}(y) = \operatorname{Cov}_Q(\mathbf{s}_A(y_A, Z), \mathbf{s}_B(y_B, Z)) . \tag{A15}$$

To expand this covariance componentwise, let $\mathbf{v}_{G,i}^\top$ denote the ith row of $\mathbf{V}_G$. Lemma A1 gives, for $i \in G$,

$$s_{G,i}(y_G, Z) = s_{G,i}(y_G, z_0) + \mathbf{v}_{G,i}^\top (Z - z_0) + r_{G,i}(Z). \tag{A16}$$

For $i \in A$ and $j \in B$, constants vanish under covariance, so Equations (A15) and (A16) imply

$$\begin{aligned} H_{ij}(y) = \mathbf{v}_{A,i}^\top \operatorname{Cov}_Q(Z) \mathbf{v}_{B,j} + \mathbf{v}_{A,i}^\top \operatorname{Cov}_Q(Z, r_{B,j}) \\ + \operatorname{Cov}_Q(r_{A,i}, Z) \mathbf{v}_{B,j} + \operatorname{Cov}_Q(r_{A,i}, r_{B,j}). \end{aligned} \tag{A17}$$

Stacking Equation (A17) over all $i \in A$ and $j \in B$ yields

$$\begin{aligned} \mathbf{H}_{A,B}(y) = \mathbf{V}_A \operatorname{Cov}_Q(Z) \mathbf{V}_B^\top + \mathbf{V}_A \operatorname{Cov}_Q(Z, \mathbf{r}_B) \\ + \operatorname{Cov}_Q(\mathbf{r}_A, Z) \mathbf{V}_B^\top + \operatorname{Cov}_Q(\mathbf{r}_A, \mathbf{r}_B). \end{aligned} \tag{A18}$$

The first term defines the affine Taylor benchmark $\mathbf{H}_{A,B}^{\text{aff}}(y) := \mathbf{V}_A \operatorname{Cov}_Q(Z) \mathbf{V}_B^\top$. Therefore the approximation error $\mathbf{E} := \mathbf{H}_{A,B}(y) - \mathbf{H}_{A,B}^{\text{aff}}(y)$ is exactly

$$\mathbf{E} = \mathbf{V}_A \operatorname{Cov}_Q(Z, \mathbf{r}_B) + \operatorname{Cov}_Q(\mathbf{r}_A, Z) \mathbf{V}_B^\top + \operatorname{Cov}_Q(\mathbf{r}_A, \mathbf{r}_B). \tag{A19}$$

Step 2: Bound the nonlinear Taylor terms. Let $\widetilde{C}_{G,i}$ be the curvature constants in Theorem 3. By Lemma A1,

$$|r_{G,i}(Z)| \le \frac{1}{2}\widetilde{C}_{G,i}\delta^2 \qquad Q\text{-almost surely.} \tag{A20}$$

The variance of each remainder component therefore satisfies

$$\begin{aligned}\mathrm{Var}_Q(r_{G,i}) &= \mathbb{E}_Q[(r_{G,i} - \mathbb{E}_Q r_{G,i})^2] \\ &\le \mathbb{E}_Q[r_{G,i}^2] \le \frac{1}{4}\widetilde{C}_{G,i}^2\delta^4.\end{aligned} \tag{A21}$$

For $i \in A$ and $j \in B$, scalar Cauchy–Schwarz gives

$$\begin{aligned}|\operatorname{Cov}_Q(r_{A,i}, r_{B,j})| &\le \sqrt{\mathrm{Var}_Q(r_{A,i})\,\mathrm{Var}_Q(r_{B,j})} \\ &\le \frac{1}{4}\widetilde{C}_{A,i}\widetilde{C}_{B,j}\delta^4.\end{aligned} \tag{A22}$$

Using Lemma A3 and summing the squared entrywise bounds gives

$$\begin{aligned}\|\operatorname{Cov}_Q(\mathbf{r}_A, \mathbf{r}_B)\|_2 &\le \|\operatorname{Cov}_Q(\mathbf{r}_A, \mathbf{r}_B)\|_F \\ &\le \frac{\delta^4}{4}\left(\sum_{i\in A}\widetilde{C}_{A,i}^2\right)^{1/2}\left(\sum_{j\in B}\widetilde{C}_{B,j}^2\right)^{1/2} \\ &= \frac{1}{4}\|\widetilde{\mathbf{C}}_A\|_2\|\widetilde{\mathbf{C}}_B\|_2\delta^4.\end{aligned} \tag{A23}$$

It remains to control the two latent–remainder covariance matrices. For a unit vector $d \in \mathbb{R}^d$, scalar Cauchy–Schwarz, Lemma A2, and Equation (A21) give

$$\begin{aligned}|d^\top \operatorname{Cov}_Q(Z, r_{G,i})| &= |\operatorname{Cov}_Q(d^\top Z, r_{G,i})| \\ &\le \sqrt{\mathrm{Var}_Q(d^\top Z)\,\mathrm{Var}_Q(r_{G,i})} \\ &\le \delta\left(\frac{1}{2}\widetilde{C}_{G,i}\delta^2\right) = \frac{1}{2}\widetilde{C}_{G,i}\delta^3.\end{aligned} \tag{A24}$$

Taking the supremum over unit d yields

$$\|\operatorname{Cov}_Q(Z, r_{G,i})\|_2 \le \frac{1}{2}\widetilde{C}_{G,i}\delta^3. \tag{A25}$$

The columns of $\operatorname{Cov}_Q(Z, \mathbf{r}_B)$ are $\operatorname{Cov}_Q(Z, r_{B,j})$. Hence, again using Lemma A3,

$$\begin{aligned}\|\operatorname{Cov}_Q(Z, \mathbf{r}_B)\|_2 &\le \|\operatorname{Cov}_Q(Z, \mathbf{r}_B)\|_F \\ &= \left(\sum_{j\in B}\|\operatorname{Cov}_Q(Z, r_{B,j})\|_2^2\right)^{1/2} \\ &\le \frac{1}{2}\delta^3\left(\sum_{j\in B}\widetilde{C}_{B,j}^2\right)^{1/2} = \frac{1}{2}\|\widetilde{\mathbf{C}}_B\|_2\delta^3.\end{aligned} \tag{A26}$$

The same argument applied to the rows of $\operatorname{Cov}_Q(\mathbf{r}_A, Z)$ gives

$$\|\operatorname{Cov}_Q(\mathbf{r}_A, Z)\|_2 \le \frac{1}{2}\|\widetilde{\mathbf{C}}_A\|_2\delta^3. \tag{A27}$$

Step 3: Bound the total approximation error. Applying the triangle inequality and submultiplicativity of the spectral norm to Equation (A19) gives

$$\begin{aligned}\|\mathbf{E}\|_2 \le{}& \|\mathbf{V}_A\|_2\|\operatorname{Cov}_Q(Z, \mathbf{r}_B)\|_2 \\ &+ \|\operatorname{Cov}_Q(\mathbf{r}_A, Z)\|_2\|\mathbf{V}_B\|_2 + \|\operatorname{Cov}_Q(\mathbf{r}_A, \mathbf{r}_B)\|_2.\end{aligned} \tag{A28}$$

Substituting Equations (A23), (A26), and (A27) into Equation (A28) yields

$$\begin{aligned}\|\mathbf{H}_{A,B}(y) - \mathbf{H}_{A,B}^{\mathrm{aff}}(y)\|_2 \le{}& \frac{1}{2}\delta^3\left(\|\mathbf{V}_A\|_2\|\widetilde{\mathbf{C}}_B\|_2 + \|\mathbf{V}_B\|_2\|\widetilde{\mathbf{C}}_A\|_2\right) \\ &+ \frac{1}{4}\delta^4\|\widetilde{\mathbf{C}}_A\|_2\|\widetilde{\mathbf{C}}_B\|_2,\end{aligned} \tag{A29}$$

which is Equation (10). □

Implication for rank. Let $\varepsilon_{\mathrm{nl}}$ denote the right-hand side of Equation (10). Since $\mathbf{H}^{\mathrm{aff}}_{A,B}(y)$ factors through $\mathbb{R}^d$, $\mathrm{rank}(\mathbf{H}^{\mathrm{aff}}_{A,B}(y)) \le d$. Weyl's inequality therefore gives, for every valid singular-value index $j > d$,

$$\sigma_j(\mathbf{H}_{A,B}(y)) \le \varepsilon_{\mathrm{nl}}.$$

Thus, Theorem 3 establishes approximate rank at most d rather than preservation of the algebraic rank. If $d = 0$, hard thresholding at $\varepsilon_{\mathrm{nl}}$ gives rank zero. If $d \ge 1$, $\mathrm{rank}(\mathbf{H}^{\mathrm{aff}}_{A,B}(y)) = d$, and

$$\sigma_d(\mathbf{H}^{\mathrm{aff}}_{A,B}(y)) > 2\varepsilon_{\mathrm{nl}},$$

then hard thresholding at $\varepsilon_{\mathrm{nl}}$ recovers rank d. Combining this approximation bound with a KDE estimation bound $\|\widehat{\mathbf{H}}_{A,B}(y) - \mathbf{H}_{A,B}(y)\|_2 \le \varepsilon_{\mathrm{est}}$ gives

$$\|\widehat{\mathbf{H}}_{A,B}(y) - \mathbf{H}^{\mathrm{aff}}_{A,B}(y)\|_2 \le \varepsilon_{\mathrm{nl}} + \varepsilon_{\mathrm{est}}.$$

Consequently, the same thresholded-rank conclusion holds with threshold $\varepsilon_{\mathrm{nl}} + \varepsilon_{\mathrm{est}}$ whenever the smallest nonzero benchmark singular value exceeds twice this total error.

D.6 KDE Cross-Hessian Estimation and Rank Recovery

Theorem A1 (Pointwise KDE Cross-Hessian Estimation Error). *Fix an interior point $y = (y_A^\top, y_B^\top)^\top \in \mathbb{R}^s$. Let U be a neighborhood of y on which $p_{A,B} \ge p_{\min} > 0$ and $p_{A,B}$ has bounded continuous partial derivatives up to order four. Let $\mathcal{K}$ be a twice continuously differentiable, compactly supported second-order kernel satisfying*

$$\int \mathcal{K}(u)\,du = 1, \qquad \int u\,\mathcal{K}(u)\,du = 0, \qquad \int \|u\|_2^2 |\mathcal{K}(u)|\,du < \infty, \tag{A30}$$

and suppose its partial derivatives up to order two are bounded and square-integrable. For $0 < \eta < 1$, choose a deterministic bandwidth $0 < h \le h_0 \le 1$ such that $y - h\ \mathrm{supp}(\mathcal{K}) \subset U$, and define

$$\Delta_{N,h,\eta} := C_1 \left\{ h^2 + \sqrt{\frac{\log(C_0/\eta)}{Nh^{s+4}}} \right\}. \tag{A31}$$

If $Nh^s \ge \log(C_0/\eta)$ and $\Delta_{N,h,\eta} \le p_{\min}/2$, then, with probability at least $1 - \eta$,

$$\left\| \widehat{\mathbf{H}}_{A,B}(y) - \mathbf{H}_{A,B}(y) \right\|_2 \le C \left\{ h^2 + \sqrt{\frac{\log(C_0/\eta)}{Nh^{s+4}}} \right\}. \tag{A32}$$

Here C, C_0, and C_1 are positive constants that depend only on the density and derivative bounds on U, the kernel, and the fixed dimensions.

Proof. Let $p = p_{A,B}$, and let γ be a multi-index of order $r = |\gamma| \le 2$. Differentiating the finite KDE sum gives

$$D^\gamma \widehat{p}_h(y) = \frac{1}{Nh^{s+r}} \sum_{i=1}^{N} D^\gamma \mathcal{K}\left(\frac{y - Y^{(i)}}{h} \right). \tag{A33}$$

Because $\mathcal{K}$ is twice continuously differentiable and compactly supported, integration by parts gives

$$\mathbb{E}[D^\gamma \widehat{p}_h(y)] = \int \mathcal{K}(u) D^\gamma p(y - hu)\,du. \tag{A34}$$

The bandwidth restriction ensures that every $y - hu$ in this integral lies in U. Taylor expansion of $D^\gamma p(y - hu)$ around y, together with the normalization, zero first moment, and finite second absolute moment in Equation (A30), therefore yields

$$|\mathbb{E}[D^\gamma \widehat{p}_h(y)] - D^\gamma p(y)| \le B_\gamma h^2. \tag{A35}$$

For the stochastic term, define $W_{i,\gamma} = h^{-(s+r)} D^\gamma \mathcal{K}((y - Y^{(i)})/h)$. Compact support, the upper density bound on U, and square integrability of $D^\gamma \mathcal{K}$ imply $\mathrm{Var}(W_{i,\gamma}) \le V_\gamma h^{-(s+2r)}$, while boundedness gives $|W_{i,\gamma}| \le L_\gamma h^{-(s+r)}$. Bernstein's inequality consequently gives, with the relevant failure probability,

$$|D^\gamma \widehat{p}_h(y) - \mathbb{E}[D^\gamma \widehat{p}_h(y)]| \le C_\gamma \left\{ \sqrt{\frac{\log(C_0/\eta)}{Nh^{s+2r}}} + \frac{\log(C_0/\eta)}{Nh^{s+r}} \right\}. \tag{A36}$$

A union bound over the finitely many partial derivatives of orders zero, one, and two makes these inequalities simultaneous. The condition $Nh^s \ge \log(C_0/\eta)$ makes the linear Bernstein term no larger than the square-root term. Since $r \le 2$ and $h \le 1$, combining the bias and stochastic bounds gives

$$\max_{|\gamma| \le 2} |D^\gamma \widehat{p}_h(y) - D^\gamma p(y)| \le \Delta_{N,h,\eta}. \tag{A37}$$

It remains to pass from density derivatives to the cross-Hessian of the log-density. Write $p_0 = p(y)$, $\mathbf{d}_A = \nabla_{y_A} p(y)$, $\mathbf{d}_B = \nabla_{y_B} p(y)$, and $M_{A,B} = \nabla^2_{y_A, y_B} p(y)$, and define

$$\Phi(p_0, \mathbf{d}_A, \mathbf{d}_B, M_{A,B}) := \frac{M_{A,B}}{p_0} - \frac{\mathbf{d}_A \mathbf{d}_B^\top}{p_0^2}. \tag{A38}$$

Then $\mathbf{H}_{A,B}(y) = \Phi(p_0, \mathbf{d}_A, \mathbf{d}_B, M_{A,B})$, with the analogous identity for the KDE quantities. On the event in Equation (A37),

$$\widehat{p}_h(y) \geq p(y) - \Delta_{N,h,\eta} \geq p_{\min}/2. \tag{A39}$$

Thus both denominators in Φ are uniformly bounded away from zero. Using Equation (A37) to bound the density, gradient, and mixed-Hessian perturbations, and expanding the two fractions in Equation (A38), gives $\|\widehat{\mathbf{H}}_{A,B}(y) - \mathbf{H}_{A,B}(y)\|_2 \leq C_2 \Delta_{N,h,\eta}$. Absorbing $C_1 C_2$ into C proves Equation (A32). □

Theorem A2 (**Rank Recovery under a Valid Error-Bound Threshold**). *Fix an evaluation point y at which the affine benchmark $\mathbf{H}^{\text{aff}}_{A,B}(y)$ has rank d^*. Suppose*

$$\|\mathbf{H}_{A,B}(y) - \mathbf{H}^{\text{aff}}_{A,B}(y)\|_2 \leq \varepsilon_{A,B}(y)$$

and, with probability at least $1 - \eta$,

$$\|\widehat{\mathbf{H}}_{A,B}(y) - \mathbf{H}_{A,B}(y)\|_2 \leq \alpha.$$

Let $\beta := \varepsilon_{A,B}(y) + \alpha$. If $d^* \geq 1$, suppose $\sigma_{d^*}(\mathbf{H}^{\text{aff}}_{A,B}(y)) > 2\beta$; no signal-gap condition is required when $d^* = 0$. Then, with probability at least $1 - \eta$,

$$\text{rank}_\beta\Big(\widehat{\mathbf{H}}_{A,B}(y)\Big) = d^*, \tag{A40}$$

where rank_β counts the singular values strictly larger than β.

Proof. By the triangle inequality, on the stated high-probability event,

$$\|\widehat{\mathbf{H}}_{A,B}(y) - \mathbf{H}^{\text{aff}}_{A,B}(y)\|_2 \leq \varepsilon_{A,B}(y) + \alpha = \beta.$$

Weyl's inequality therefore bounds every singular-value perturbation by β. Since the benchmark has rank d^*, all estimated singular values with index $j > d^*$ are at most β. If $d^* = 0$, this already shows that no singular value exceeds β. If $d^* \geq 1$, the signal-gap condition also gives

$$\sigma_{d^*}(\widehat{\mathbf{H}}_{A,B}(y)) \geq \sigma_{d^*}(\mathbf{H}^{\text{aff}}_{A,B}(y)) - \beta > \beta.$$

Thus exactly the first d^* singular values remain after hard thresholding at β. □

Scope. Theorem A2 concerns an ideal threshold that bounds the total deviation from the affine benchmark. The bootstrap and relative thresholds in Section 3.3 provide a data-adaptive finite-sample implementation and are not required by the theorem.

D.7 Evaluation Points and Rank Aggregation

For each queried pair (A, B), we select the KDE bandwidth h once by leave-one-out likelihood cross-validation. We begin with a prespecified finite candidate set of interior points, discard candidates whose fitted density is below a prespecified lower quantile, and denote the retained set by $\mathcal{Y}_{\text{eval}} = \{y_1, \ldots, y_{K_{\text{eval}}}\}$. After this selection, both h and $\mathcal{Y}_{\text{eval}}$ are held fixed when the original and bootstrap cross-Hessians are computed.

At each y_k, let $\tau_{\text{boot},k}$ be the bootstrap spectral-norm threshold and set $\tau_k = \max\{\tau_{\text{boot},k}, \rho\, \sigma_1(\widehat{\mathbf{H}}_{A,B}(y_k))\}$. The pointwise estimate $\widehat{r}_k$ is the number of singular values of $\widehat{\mathbf{H}}_{A,B}(y_k)$ that exceed τ_k. We report the modal value among $\widehat{r}_1, \ldots, \widehat{r}_{K_{\text{eval}}}$, resolving a tie in favor of the larger rank. All empirical HRC decisions use this aggregated rank.

This procedure is an operational aggregation rule. Theorem A2 applies when a valid pointwise estimator-error bound is available and does not itself establish bootstrap validity for data-dependent bandwidths or evaluation points. Selecting h and $\mathcal{Y}_{\text{eval}}$ on an independent pilot sample avoids reusing the estimation sample for these choices. For multiple fixed evaluation points, valid pointwise bounds at level η / K_{eval} can be combined by a union bound. Correspondingly, $q_{\text{boot}} = 1 - \eta / K_{\text{eval}}$ is used as a practical Bonferroni calibration rather than as a separate bootstrap-consistency guarantee.

Implementation details. The standalone rank calibration uses the hybrid cutoff with $\rho = 0.07$ and bootstrap quantile 0.95. In the end-to-end experiments, the cutoff uses 40 bootstrap replicates: clustering uses $\rho = 0.0554$ and bootstrap quantile 0.8914, while HRC-PC uses $\rho = 0.0736$ and bootstrap quantile 0.9906. Because τ_k is the maximum of its bootstrap and relative components, the active component is determined separately for each query; the code does not assume or record a globally dominant component. With capped KDE support and local-neighborhood sizes, the cost of one rank query grows approximately linearly with N for support selection, linearly with the number of bootstrap replicates and retained local points, and quadratically with the active coordinate dimension. The total cost is driven by the number of queries: for m observed variables, FOFC has at most $O(m^4)$ quartet queries, whereas PC-style search over n_L located latent variables and maximum conditioning-set size $q_{\max}$ has at most $O\Big(n_L^2 \sum_{q=0}^{q_{\max}} \binom{n_L-2}{q}\Big)$ queries.

D.8 Proof of Proposition 1

Proof. For a pure one-factor quartet, the minimum set of latent variables that d-separates the two groups in each pairwise partition has cardinality one. Theorem 2 therefore gives thresholded cross-Hessian rank one for all three partitions. For a mixed quartet, Assumption 4 ensures that at least one partition has minimum separating cardinality different from one, and Theorem 2 gives a thresholded rank different from one for that partition. Hence the HRC quartet decisions coincide with the population FOFC decisions. The cluster-growing and cluster-selection steps are unchanged, so the conclusion follows from Theorem 6 of Kummerfeld & Ramsey (2016). □

D.9 Proof of Theorem 4

Theorem 4 (Latent Conditional Independence via HRC). Assume a correctly identified pure nonlinear one-factor measurement model whose latent distribution is Markov and faithful to its latent DAG. Let L_i, L_j and Q be as in Theorem 4, with $|Q| = q$, and construct A, B from pure indicators as specified there. Suppose Assumptions 1–3 hold for every HRC query involved and, whenever $q \geq 1$, the signal-gap condition in Equation (8) also holds. Then

$$L_i \perp\!\!\!\perp L_j \mid Q \quad \Longleftrightarrow \quad \operatorname{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) = q.$$

Proof. For $q = 0$, use $Q = \varnothing$ and the convention that the affine benchmark associated with the empty conditioning set has rank zero. The argument below then reduces to marginal independence. We therefore write the proof for $q \geq 1$.

Suppose first that $L_i \perp\!\!\!\perp L_j \mid Q$. Faithfulness of the latent distribution implies that Q d-separates L_i and L_j. Purity of the measurement model then implies that Q d-separates the constructed observed groups A and B. Moreover, Q is a minimum latent d-separating set for these groups. Indeed, for each $L_k \in Q$, one pure indicator lies in A and another in B, and the path

$$X_k^A \leftarrow L_k \rightarrow X_k^B$$

remains active unless L_k is included. Hence every latent d-separating set contains all q members of Q. Theorem 2 therefore yields

$$\operatorname{rank}_{\varepsilon_{A,B}}(\mathbf{H}_{A,B}) = q.$$

Conversely, suppose the thresholded cross-Hessian rank equals q. Under the approximation and signal-gap conditions, Theorem 2 identifies this thresholded rank with the minimum latent d-separating cardinality. The pure-indicator paths above force every such set to contain all members of Q; a set of cardinality q must therefore be exactly Q. Thus Q d-separates A and B. Any active path between L_i and L_j given Q, together with the pure-indicator edges at its endpoints, would induce an active path between A and B, a contradiction. Hence Q d-separates L_i and L_j, and the causal Markov property gives $L_i \perp\!\!\!\perp L_j \mid Q$. □

D.10 Proof of Theorem 5

Theorem 5 (Correctness of HRC-PC). Suppose HRC-based FOFC correctly recovers the target latent variables and their pure measurement clusters. Assume that these latent variables are causally sufficient and that their distribution satisfies the causal Markov and faithfulness conditions with respect to the target latent DAG. If every HRC query made by HRC-PC agrees with the corresponding latent conditional-independence relation, then HRC-PC returns the Markov equivalence class of the target latent DAG.

Proof. Under the stated conditions, HRC-PC makes the same conditional-independence decisions as oracle PC applied to the recovered latent variables. Causal sufficiency, the causal Markov condition, and faithfulness therefore allow the standard correctness result for PC to be applied (Spirtes et al., 2000). Hence HRC-PC returns the Markov equivalence class of the target latent DAG. □

E Further Discussion on Assumptions

E.1 Discussion on Pointwise Non-degeneracy

E.1.1 Full-Dimensional Conditional Variation

Fix a minimum-cardinality latent separator $\mathcal{Z}^* \subseteq \mathcal{L}$ and, for brevity, write $d = |\mathcal{Z}^*|$ and $Z \in \mathbb{R}^d$ for its associated random vector. Assumption 3-(i) is a pointwise condition on the conditional distribution of this minimum separator. For $P_{A,B}$-almost every (a, b), it requires finite conditional second moments and

$$\mathrm{Cov}(Z \mid A = a, B = b) \succ 0. \tag{A41}$$

For every $v \in \mathbb{R}^d$,

$$v^\top \mathrm{Cov}(Z \mid A = a, B = b)v = \mathrm{Var}(v^\top Z \mid A = a, B = b). \tag{A42}$$

Consequently, the covariance is singular if and only if there exist a nonzero v and a scalar $c(a, b)$ such that $v^\top Z = c(a, b)$ holds $P_{Z|A=a,B=b}$-almost surely. Equivalently, the conditional distribution is supported on a proper affine subspace of $\mathbb{R}^d$. This is the precise failure excluded by Assumption 3-(i).

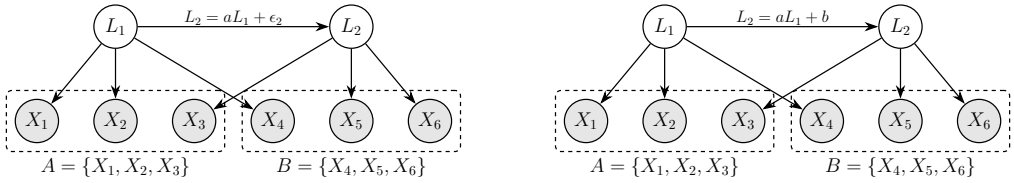
To illustrate the condition, consider the two-dimensional separator

$$Z = (L_1, L_2)^\top. \tag{A43}$$

Under conditional independence, the affine benchmark is, for $P_{A,B}$-almost every (a, b),

$$\mathbf{H}^{\mathrm{aff}}_{A,B} = \mathbf{V}_A(a)\ \mathrm{Cov}(Z \mid A = a, B = b)\, \mathbf{V}_B(b)^\top. \tag{A44}$$

Thus positive definiteness makes the middle matrix rank two, while Assumption 3-(ii) prevents the slope matrices from reducing the benchmark rank. The actual cross-Hessian differs from this benchmark by the remainder matrix in Theorem 1.



(a) Stochastic linear relation.

(b) Deterministic linear relation.

Figure A7: Two latent-variable graphs in which L_1 and L_2 each have children in both observed groups, so the minimum latent d-separator between A and B is $\{L_1, L_2\}$. In (a), $L_2 = aL_1 + \epsilon_2$ contains a non-degenerate exogenous noise term; Assumption 3-(i) additionally requires $\mathrm{Cov}((L_1, L_2)^\top \mid A = a, B = b) \succ 0$ for $P_{A,B}$-almost every (a, b). In (b), $L_2 = aL_1 + b$, so the conditional distribution of $(L_1, L_2)^\top$ is supported on the affine line $\{(l_1, l_2) : l_2 = al_1 + b\}$ and its conditional covariance is singular.

In both graphs of Figure A7, the fork paths through L_1 and L_2 show that neither latent variable alone d-separates A and B; hence $\{L_1, L_2\}$ is a minimum latent d-separator. We compare two cases: a stochastic linear relation, where L_2 receives its own exogenous noise, and a deterministic linear relation, where L_2 is completely determined by L_1.

First, consider the stochastic linear relation

$$L_2 = aL_1 + \epsilon_2, \qquad \epsilon_2 \sim \mathcal{N}(0, \sigma_2^2), \qquad \sigma_2^2 > 0. \tag{A45}$$

Here ϵ_2 provides an additional exogenous source of variation for L_2. Its presence alone is not used as a substitute for Assumption 3-(i). The required condition is stated directly: for $P_{A,B}$-almost every (a, b),

$$\mathrm{Cov}\Big((L_1, L_2)^\top \mid A = a, B = b\Big) \succ 0, \qquad \mathrm{rank}\Big(\mathrm{Cov}\Big((L_1, L_2)^\top \mid A = a, B = b\Big)\Big) = 2. \tag{A46}$$

If the observation matrices $\mathbf{V}_A(A)$ and $\mathbf{V}_B(B)$ both have full column rank two, then multiplication by these matrices does not reduce the rank. Hence

$$\mathrm{rank}(\mathbf{H}_{A,B}) = \mathrm{rank}\Big(\mathbf{V}_A(A)\ \mathrm{Cov}(Z \mid A, B)\, \mathbf{V}_B(B)^\top\Big) = 2. \tag{A47}$$

Therefore, in the stochastic linear case, the cross-Hessian rank correctly recovers the two-dimensional separating the latent set.

In contrast, consider the deterministic linear relation

$$L_2 = aL_1 + b. \tag{A48}$$

Let $v = (-a, 1)^\top$. The structural equality implies

$$v^\top Z = L_2 - aL_1 = b \tag{A49}$$

under every conditional distribution $P_{Z|A=a,B=b}$. Therefore,

$$\mathrm{Var}(v^\top Z \mid A = a, B = b) = 0, \tag{A50}$$

so $\mathrm{Cov}(Z \mid A = a, B = b)$ is singular and the conditional distribution is supported on the proper affine line $\{(l_1, l_2) : l_2 = al_1 + b\}$. Expanding this covariance gives

$$\mathrm{Cov}(Z \mid A = a, B = b) = \begin{pmatrix} \mathrm{Var}(L_1 \mid A = a, B = b) & a\ \mathrm{Var}(L_1 \mid A = a, B = b) \\ a\ \mathrm{Var}(L_1 \mid A = a, B = b) & a^2\ \mathrm{Var}(L_1 \mid A = a, B = b) \end{pmatrix}. \tag{A51}$$

Equivalently,

$$\mathrm{Cov}(Z \mid A = a, B = b) = \mathrm{Var}(L_1 \mid A = a, B = b) \begin{pmatrix} 1 \\ a \end{pmatrix} \begin{pmatrix} 1 & a \end{pmatrix}. \tag{A52}$$

This is an outer-product matrix, so whenever $\mathrm{Var}(L_1 \mid A = a, B = b) > 0$, we have

$$\mathrm{rank}(\mathrm{Cov}(Z \mid A = a, B = b)) = 1. \tag{A53}$$

Thus, the posterior covariance is not positive definite, and Assumption 3-(i) is violated. Substituting this degenerate posterior covariance into the affine benchmark yields

$$\mathrm{rank}(\mathbf{H}^{\mathrm{aff}}_{A,B}) \le \mathrm{rank}(\mathrm{Cov}(Z \mid A = a, B = b)) = 1. \tag{A54}$$

If both slope matrices are nonzero on the rank-one image of the conditional covariance, then the benchmark rank is exactly one. Nonzero score remainders may introduce additional small singular values into the actual cross-Hessian; these are controlled by the approximation tolerance in Theorem 1.

This example makes the role of Assumption 3-(i) explicit. In the stochastic model, the assumption requires the conditional covariance to have rank two for $P_{A,B}$-almost every (a, b). In the deterministic model, the identity $(-a, 1)^\top Z = b$ forces the conditional distribution onto a proper affine line, so the conditional covariance has rank one and the cross-Hessian rank is at most one. Thus, the relevant condition is full-dimensional conditional support, stated equivalently through positive-definite conditional covariance.

E.1.2 Zero Measurements and Informative Observations

For a given evaluation point, a nonzero vector $c \in \mathbb{R}^d$ is a zero-measurement direction for group A if $\mathbf{V}_A(a)c = 0$, and analogously for group B. In nonlinear models, this direction may vary with the observed value. Assumption 3-(ii) excludes such directions for $P_{A,B}$-almost every (a, b). In a linear Gaussian measurement model, the slope matrices are constant and have the same column ranks as the loading matrices. A zero loading column is one example, while linearly dependent nonzero columns can produce the same failure. Conversely, an individual zero loading coefficient does not violate the assumption if the complete loading matrix remains full column rank. When full column rank is feasible, rank-deficient loading matrices form a proper algebraic, and hence Lebesgue-measure-zero, subset of the unrestricted loading-parameter space. This parameter-space statement is distinct from the $P_{A,B}$-almost-everywhere qualification over evaluation points; under structural restrictions that force rank deficiency, the failure need not be measure zero within the restricted model class.

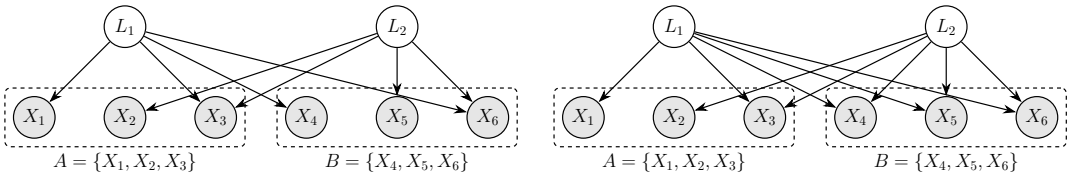
Under conditional independence, the affine benchmark is

$$\mathbf{H}^{\mathrm{aff}}_{A,B} = \mathbf{V}_A(A)\ \mathrm{Cov}(Z \mid A, B)\, \mathbf{V}_B(B)^\top. \tag{A55}$$

Assumption 3-(ii) requires both $\mathbf{V}_A(a)$ and $\mathbf{V}_B(b)$ to have full column rank d. Thus, both observed groups must collectively contain information about every direction of $Z \in \mathbb{R}^d$. If one group has a zero-measurement direction, the affine benchmark loses rank even when the posterior covariance is positive definite. The robust HRC theorem transfers this benchmark rank to the thresholded rank of the actual cross-Hessian only when its signal-gap condition is satisfied.

The requirement is imposed on each observed group as a whole; it does not require every observed variable to depend on every latent variable. For example, when $Z = (L_1, L_2)^\top$, one variable in group A may primarily measure L_1 and another may primarily measure L_2, provided that the group jointly spans both latent directions. What is excluded is a group-level loss of a latent direction.

Figure A8 illustrates the distinction. In both graphs, the separating latent set is $\mathcal{Z} = \{L_1, L_2\}$. In Figure A8a, both observed groups contain information about two independent latent directions. In Figure A8b, both latents



(a) Both groups are informative.

(b) Group B measures only one latent direction.

Figure A8: Illustration of the informative-observation condition. In both graphs, the separating latent set is $\mathcal{Z} = \{L_1, L_2\}$. In (a), both observed groups collectively span two latent directions. In (b), both latents have nonzero effects on group B, but their loading columns are proportional. Thus B measures only one linear combination of the latents, $\mathbf{V}_B(b)$ has rank one, and the cross-Hessian loses one graphically present latent direction.

can have nonzero effects on group B, but the two loading columns are proportional. Hence B responds only to one linear combination of (L_1, L_2) and has a zero-measurement direction, so its cross-Hessian cannot recover both graphically present latent directions.

To make this point explicit, consider the special case of a linear Gaussian measurement model:

$$A = \Lambda_A Z + \epsilon_A, \qquad B = \Lambda_B Z + \epsilon_B, \tag{A56}$$

where $Z = (L_1, L_2)^\top$, $\epsilon_A \sim \mathcal{N}(0, \Omega_A)$, and $\epsilon_B \sim \mathcal{N}(0, \Omega_B)$. Then

$$\frac{\partial \ln p_{A|Z}(A \mid Z)}{\partial A} = -\Omega_A^{-1} A + \Omega_A^{-1} \Lambda_A Z, \qquad \frac{\partial \ln p_{B|Z}(B \mid Z)}{\partial B} = -\Omega_B^{-1} B + \Omega_B^{-1} \Lambda_B Z. \tag{A57}$$

Thus, in this special case,

$$\mathbf{V}_A(A) = \Omega_A^{-1} \Lambda_A, \qquad \mathbf{V}_B(B) = \Omega_B^{-1} \Lambda_B. \tag{A58}$$

Since Ω_A and Ω_B are positive definite noise covariance matrices, multiplication by Ω_A^{-1} or Ω_B^{-1} does not change column rank. Therefore,

$$\operatorname{rank}(\mathbf{V}_A(A)) = \operatorname{rank}(\Lambda_A), \qquad \operatorname{rank}(\mathbf{V}_B(B)) = \operatorname{rank}(\Lambda_B). \tag{A59}$$

In the informative case, suppose, for example, that

$$\Lambda_A = \begin{pmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \\ \varepsilon_3 & \varepsilon_4 \end{pmatrix}, \qquad \Lambda_B = \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \\ \beta_3 & \beta_4 \end{pmatrix}, \tag{A60}$$

where the columns of both matrices are linearly independent. Then

$$\operatorname{rank}(\mathbf{V}_A) = \operatorname{rank}(\mathbf{V}_B) = 2. \tag{A61}$$

If Assumption 3-(i) also holds, so that

$$\operatorname{rank}(\operatorname{Cov}(Z \mid A, B)) = 2 \quad P_{A,B}\text{-almost surely}, \tag{A62}$$

then

$$\operatorname{rank}(\mathbf{H}_{A,B}) = \operatorname{rank}\Big(\mathbf{V}_A \operatorname{Cov}(Z \mid A, B)\, \mathbf{V}_B^\top\Big) = 2. \tag{A63}$$

Hence, the cross-Hessian rank correctly recovers the two-dimensional separating latent set.

In contrast, suppose the two loading columns for group B are proportional:

$$\Lambda_B = \begin{pmatrix} \beta_1 & c\beta_1 \\ \beta_2 & c\beta_2 \\ \beta_3 & c\beta_3 \end{pmatrix}, \qquad c \neq 0. \tag{A64}$$

Both latent variables may therefore have nonzero arrows into every variable in B, while B depends only on the combination $L_1 + cL_2$. Indeed, $\Lambda_B(-c, 1)^\top = 0$, and hence

$$\operatorname{rank}(\mathbf{V}_B) = \operatorname{rank}(\Lambda_B) = 1. \tag{A65}$$

Even if group A is fully informative and the posterior covariance of Z is positive definite,

$$\operatorname{rank}(\mathbf{H}_{A,B}) \leq \operatorname{rank}(\mathbf{V}_B) = 1. \tag{A66}$$

Thus, although the graph contains two latent paths and its minimum latent d-separating set has cardinality two, the observed cross-Hessian sees only one direction. This is precisely the rank loss excluded by the full-column-rank clause of Assumption 3-(ii). A literal zero loading column is the boundary case $c = 0$; the proportional-column construction shows that zero measurement refers more generally to an unobserved latent direction, not necessarily to an individually absent latent variable.

E.2 Exact Affinity of Conditional Derivatives

Assumption 2 is imposed relative to a specified set $\mathcal{Z} \subseteq \mathcal{L}$ of latent variables. Writing $d = |\mathcal{Z}|$ and $Z \in \mathbb{R}^d$ for its associated random vector, the assumption decomposes each conditional score into an affine term in z and a posterior-small remainder. It is not a global linearity condition on the structural mechanisms, nor is it required to hold simultaneously for every subset of $\mathcal{L}$. Exact affine constructions are included as the zero-remainder special case.

The condition permits genuinely nonlinear measurement functions for arbitrary separator dimension. Let $X \in (0,\infty)^m$ satisfy

$$X = \exp(\Lambda Z + \epsilon), \qquad \epsilon \sim \mathcal{N}(0, \Omega),$$

where the exponential is applied componentwise, $\Lambda \in \mathbb{R}^{m \times d}$, and $\Omega \succ 0$. For a realization x, let $D(x) := \mathrm{diag}(x)$. The conditional log-density is

$$\log p_{X|Z}(x \mid z) = -\frac{1}{2}(\log x - \Lambda z)^\top \Omega^{-1}(\log x - \Lambda z) - \sum_{i=1}^{m} \log x_i + C,$$

where $\log x$ is componentwise. Differentiation gives

$$\nabla_x \log p_{X|Z}(x \mid z) = \underbrace{-D(x)^{-1}\Omega^{-1}\log x - D(x)^{-1}\mathbf{1}}_{\mathbf{u}_X(x)} + \underbrace{D(x)^{-1}\Omega^{-1}\Lambda}_{\mathbf{V}_X(x)} z.$$

Thus the derivative is exactly affine in the full d-dimensional separator realization, even though the measurement function is nonlinear. Applying this construction separately to A and B yields Assumption 2. Figure 2 is the scalar special case $m = d = 1$, $\Lambda = 1$, and $\Omega = 1$. This example gives a concrete sufficient nonlinear class; it does not claim that every post-nonlinear model satisfies the assumption.

An iterated post-nonlinear special case. The same affine-derivative property can hold after several nonlinear structural layers. Set $W_0 = Z$ and g_0 to the identity map, and consider

$$W_k = g_k\left(M_k g_{k-1}^{-1}(W_{k-1}) + \epsilon_k\right), \qquad k = 1, \ldots, K,$$

where each g_k is a twice continuously differentiable diffeomorphism and the noises $\epsilon_k \sim \mathcal{N}(0, \Omega_k)$ are mutually independent and independent of Z. The intermediate variables $W_1, \ldots, W_{K-1}$ may be latent, while $X := W_K$ is observed. Defining $U_k := g_k^{-1}(W_k)$ gives the linear Gaussian recursion $U_k = M_k U_{k-1} + \epsilon_k$. Consequently, after marginalizing the intermediate variables,

$$U_K \mid Z = z \sim \mathcal{N}(\Lambda z, \Omega), \qquad \Lambda = M_K \cdots M_1,$$

where $\Omega \succ 0$ is the covariance obtained by propagating the layer noises. Let $h := g_K^{-1}$. A change of variables yields

$$\nabla_x \log p_{X|Z}(x \mid z) = \underbrace{\nabla_x \log|\det J_h(x)| - J_h(x)^\top \Omega^{-1} h(x)}_{\mathbf{u}_X(x)} + \underbrace{J_h(x)^\top \Omega^{-1}\Lambda}_{\mathbf{V}_X(x)} z.$$

Thus the conditional log-density derivative remains exactly affine in the full separator realization z, although the structural relations are iterated and nonlinear and the intermediate variables have been marginalized. Applying two such branches with conditionally independent noises to the groups A and B gives Assumption 2.

F Details of FOFC Framework

Find One Factor Clusters (FOFC) is a clustering framework for learning pure one-factor measurement clusters from observed indicators. The goal is to group observed variables that are likely to share the same latent parent. FOFC does not assume that the number of latent variables or the cluster assignments are known in advance. Instead, it searches for sets of observed variables whose covariance structure is consistent with a one-factor measurement model.

The algorithm is built on quartet tests. For a set of four observed variables, FOFC checks whether the quartet satisfies the rank pattern expected from a one-factor structure, which is usually implemented through vanishing tetrad constraints. A triple of variables is retained as a candidate pure triple if adding any other observed variable forms a valid one-factor quartet. These candidate triples are then merged into larger clusters, and the final output is obtained by selecting maximal non-overlapping clusters.

Here, QuartetTest(Q) denotes the statistical test used to decide whether a quartet Q is consistent with a one-factor measurement structure. In the original FOFC algorithm, this is based on vanishing tetrad constraints. The first stage finds candidate pure triples, the second stage grows these triples into larger candidate clusters, and the final stage selects maximal non-overlapping clusters as the recovered measurement model.

Algorithm A1 Find One Factor Clusters (FOFC)

Require: Observed variables $\mathcal{X}$, quartet test
Ensure: Recovered measurement clusters $\mathcal{C}$

$\mathcal{T} \leftarrow \emptyset$
for each triple $S \subset \mathcal{X}$ with $|S| = 3$ **do**
  $\mathrm{Pure} \leftarrow \mathrm{TRUE}$
  for each variable $X_i \in \mathcal{X} \setminus S$ **do**
    if QUARTETTEST$(S \cup \{X_i\}) = \mathrm{FALSE}$ **then**
      $\mathrm{Pure} \leftarrow \mathrm{FALSE}$
      break
    end if
  end for
  if $\mathrm{Pure} = \mathrm{TRUE}$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{S\}$
  end if
end for
$\mathcal{C}_{\mathrm{cand}} \leftarrow \mathcal{T}$
for each candidate cluster $C \in \mathcal{C}_{\mathrm{cand}}$ **do**
  for each variable $X_i \in \mathcal{X} \setminus C$ **do**
    $\mathrm{Expandable} \leftarrow \mathrm{TRUE}$
    for each pair $P \subset C$ with $|P| = 2$ **do**
      if $P \cup \{X_i\} \notin \mathcal{T}$ **then**
        $\mathrm{Expandable} \leftarrow \mathrm{FALSE}$
        break
      end if
    end for
    if $\mathrm{Expandable} = \mathrm{TRUE}$ **then**
      $C \leftarrow C \cup \{X_i\}$
    end if
  end for
end for
$\mathcal{C} \leftarrow \emptyset$
while $\mathcal{C}_{\mathrm{cand}} \neq \emptyset$ **do**
  Select the largest cluster $C^\star$ in $\mathcal{C}_{\mathrm{cand}}$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{C^\star\}$
  Remove from $\mathcal{C}_{\mathrm{cand}}$ all clusters that intersect $C^\star$
end while
return $\mathcal{C}$

G EXPERIMENT

G.1 SYNTHETIC DATASET

The synthetic data were generated from a nonlinear structural equation model with latent measurement structure. For each setting, we fixed the number of latent variables and the number of observed measurements per latent variable M. We first generated a directed acyclic graph over the latent variables by sampling a random topological ordering, constructing a connected directed tree, and then optionally adding extra directed edges. Each latent edge coefficient was sampled independently from $\mathrm{Unif}(0.7, 1.3)$. Root latent variables were sampled from $\mathrm{Unif}(a, b)$, while non-root latent variables were generated in topological order as

$$L_j = \sum_{i \in \mathrm{Pa}(j)} \beta_{ij} L_i + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, 0.2^2). \tag{A67}$$

Each latent variable L_j then generated r_{ind} pure observed measurements:

$$X_{j\ell} = b_{j\ell} f_{j\ell}\big((\ell + 1)L_j\big) + \eta_{j\ell}, \qquad \ell = 1, \ldots, r_{\mathrm{ind}}, \quad \eta_{j\ell} \sim \mathcal{N}(0, \sigma_x^2). \tag{A68}$$

The loading coefficient $b_{j\ell}$ had a random sign and magnitude sampled from $\mathrm{Unif}(2, 5)$. The nonlinear measurement function $f_{j\ell}$ was chosen from $\{\sin, \sinh, \tanh, \mathrm{softplus}\}$. In the single-function setting, all measurements in a dataset used the same nonlinear function; in the mixed setting, each measurement was assigned a function independently. The observation noise level σ_x was selected from a predefined grid. All datasets were generated

with fixed random seeds for reproducibility, and each observed variable was standardized before evaluation. All experiments are run on nodes with 128 logical CPU cores.

We consider four single-function settings, using $\sin$, $\sinh$, $\tanh$, or softplus for all measurements, and a mixed setting in which each measurement independently selects one of these functions. The number of latent variables is varied over $\{4, 6, 8, 10\}$. For each dimension, we generate 10 random latent graphs and run 10 random seeds per graph. We use sample sizes $N \in \{200, 1000, 2000, \ldots, 9000\}$. Latent-variable location is evaluated by F1, while latent causal discovery is evaluated by F1 and structural Hamming distance (SHD) for the latent Markov equivalence class. We report the mean and standard deviation over all graph–seed combinations.

G.2 Real-world Dataset

The experiment uses the August 2025 vintage of FRED-MD, a monthly U.S. macroeconomic database covering January 1959 through July 2025 (McCracken & Ng, 2016). We apply the series-specific transformation codes supplied by FRED-MD, retain the common non-missing period, and standardize each transformed series. Following the economic grouping of the selected variables, we construct four latent factors: an interest-rate factor measured by FEDFUNDS, TB3MS, GS10, GS5, and GS10TB3Mx; a credit factor measured by BAA, BAA10YM, and COMPAPFF; an exchange-rate factor measured by TWEXMMTH, EXUSEU, and EXJPUSx; and an equity/volatility factor measured by S&P 500, NASDAQCOM, and VXOCLSX. When an exact series is unavailable in the FRED-MD vintage, we use the closest available proxy or construct the spread variable directly from available series.

We apply HRC-PC using the KDE-based cross-Hessian estimator and bootstrap thresholding rule described in Section 3.3. Since no ground-truth latent graph is available, we report the recovered structure and examine whether its relations admit an economically meaningful interpretation.

G.3 More Experiment Results

Table A4: Sensitivity to noise level on L4M3 with $n = 3000$.

Noise	Cluster F1	Structure F1	Structure SHD
1.00	0.7143 ± 0.2771	0.5430 ± 0.3031	2.0000 ± 1.1466
0.75	0.8663 ± 0.2160	0.7277 ± 0.2291	1.3960 ± 1.0341
0.50	0.9387 ± 0.1475	0.8469 ± 0.1764	0.9540 ± 1.0593
0.25	0.9910 ± 0.0556	0.9502 ± 0.1088	0.2820 ± 0.5789
0.15	0.9993 ± 0.0149	0.9643 ± 0.0991	0.1900 ± 0.4923

Table A5: Sensitivity to latent variable range with $n = 3000$ and noise level fixed at 0.15.

Range	Cluster F1	Structure F1	Structure SHD
$[-0.5, 0.5]$	0.9993 ± 0.0149	0.9643 ± 0.0991	0.1900 ± 0.4923
$[-1, 1]$	0.9452 ± 0.1320	0.8268 ± 0.2060	0.8780 ± 0.9492
$[-1.5, 1.5]$	0.7543 ± 0.2451	0.6064 ± 0.2661	1.8280 ± 1.0792
$[-2, 2]$	0.6025 ± 0.2382	0.5133 ± 0.2901	2.1960 ± 1.2201
$[-2.5, 2.5]$	0.4990 ± 0.2203	0.4348 ± 0.3102	2.3980 ± 1.1757

Assumption 2 permits nonzero conditional-score remainders. Theorem 3 gives explicit sufficient conditions controlling these remainders through posterior concentration and conditional-score curvature. The sensitivity study probes related regimes by varying the observation-noise level and latent range. Tables A4 and A5 show improved empirical performance at lower noise levels and over narrower latent ranges. These results are consistent with the approximate-affine analysis, but they do not directly estimate its numerical error bound.

H Broader Impact

This work may help improve the interpretability of latent-variable models in domains where observed variables are noisy nonlinear measurements of hidden factors, such as economics, finance, health care, and scientific discovery. By using cross-Hessian rank constraints to identify latent dimensions and latent causal structure, the proposed method can provide more structured explanations of high-dimensional observations. However, in sensitive applications, the recovered graph should be viewed as exploratory evidence rather than definitive causal knowledge, and should be interpreted together with domain expertise and data-quality considerations.

I LIMITATIONS

The current application to latent-variable location and causal discovery follows the pure measurement model setting, where observed indicators are assumed to be associated with latent variables in a structured way. More general measurement structures, such as mixed indicators, correlated measurement errors, or complex observed causal relations, may require additional algorithmic development. The approximate HRC characterization relies on smoothness, controlled conditional-score remainders, non-degeneracy, and a singular-value gap; these are explicit sufficient conditions and are not claimed to hold for arbitrary nonlinear structural causal models. Proposition 1 and Theorem 5 are oracle-transfer results, while the bootstrap and relative-threshold calibration is a finite-sample implementation whose rank consistency is not established here. The FRED-MD case study additionally uses prespecified groups and temporally dependent observations, so it is an exploratory illustration outside the i.i.d. estimation theory rather than a benchmark with ground-truth structure.