
Singular Learning Theory for Causal Discovery under Unobserved Confounding

Mujin Zhou

University of Illinois Urbana-Champaign
mujinz2@illinois.edu

Ignavier Ng

Carnegie Mellon University
ignavierng@cmu.edu

Junzhe Zhang

Syracuse University
jzhan403@syr.edu

Abstract

Score-based causal discovery under unobserved confounding requires both a consistent score and an efficient graph search. Linear models with correlated errors are naturally encoded by acyclic directed mixed graphs (ADMGs), whose model families include singular cases where the Bayesian Information Criterion (BIC) approximation can break down. Drawing on singular learning theory, we develop a score-based framework for linear Gaussian ADMGs. For a finite candidate family with compact analytic priors, we prove that optimizing the Widely Applicable Bayesian Information Criterion (WBIC) identifies the data-generating graph’s algebraic equivalence class, defined by its polynomial covariance constraints, with probability tending to one for almost every generating model under the prior. We further approximate the WBIC at scale with Automatic Differentiation Variational Inference (ADVI) and embed it in a differentiable search that relaxes directed and bidirected adjacency through Gumbel-Softmax gates. On synthetic and real-world benchmarks, the method improves on a BIC-scored baseline for both singular and generic ADMGs.

1 INTRODUCTION

Learning the causal structure behind a phenomenon from data is a central problem across the data sciences. A common approach models the system by a directed acyclic causal graph whose nodes are random variables (measured or latent) and whose edges point from causes to effects (Pearl, 2000; Spirtes et al., 2000; Peters et al., 2017; Bareinboim et al., 2022), and then pieces together the constraints the system imprints on the observational distribution $P(\mathbf{V})$ to in-

fer that graph. In linear models, each d -separation in the graph entails a conditional independence, i.e., a vanishing partial correlation. The converse, faithfulness, turns these independences into testable constraints that narrow the compatible graphs (Pearl, 1988; Meek, 1995; Uhler et al., 2013; Zhang, 2008a) and underpins a wide range of structure learning algorithms (Verma and Pearl, 1988; Glymour et al., 1987; Spirtes and Glymour, 1991; Spirtes et al., 2000). With linear structural functions and independent noises, d -separation statements capture *all* testable constraints of the model (Verma and Pearl, 1988).

Unobserved confounding breaks this picture. Latent confounders correlate the exogenous noise terms, and the resulting model is naturally represented by an *Acyclic Directed Mixed Graph* (ADMG) (Richardson and Spirtes, 2002), whose bidirected edges mark latent common causes. Such models impose constraints beyond conditional independence, most notably *tetrad constraints* (Sullivant et al., 2010): polynomial equalities on the observed covariance that separate structures indistinguishable by independence tests. Because these constraints surface in the likelihood, they favor *score-based* discovery (Chickering, 2002), which ranks structures by a global fit criterion. Yet with latent variables the induced family of distributions is generally *not* a curved exponential family (Geiger et al., 2001), so the Bayesian Information Criterion (BIC), an asymptotic approximation to the model evidence for such families (Haughton, 1988; Schwarz, 1978), loses its footing: the Fisher information of linear Gaussian models with correlated errors can be singular (Drton, 2009), and the asymptotics behind the BIC then fail (Drton and Maathuis, 2017).

Watanabe’s singular learning theory (Watanabe, 2009) supplies a principled remedy. The Widely Applicable Bayesian Information Criterion (WBIC) (Watanabe, 2013) replaces the dimension penalty with a data-driven estimate of the *real log-canonical threshold* (RLCT), a measure of effective complexity that stays valid when the Fisher information is singular. This

makes the WBIC a natural score for linear Gaussian ADMGs with latent confounders, but two gaps remain. Its expansion alone does not yield structure identification, so we derive the score separation needed for a prior-almost-everywhere consistency result; and its exact computation requires posterior expectations that are intractable in practical models, calling for scalable approximate inference.

Additionally, search over candidate causal graphs poses its own difficulties. Early methods for latent-variable settings, such as FCI (Spirtes et al., 2000) and its variants (Colombo et al., 2012; Zhang, 2008b), test conditional independencies, while score-based methods (Chickering, 2002; Hauser and Bühlmann, 2012) have largely targeted the fully observed case, where the BIC is consistent and decomposable. Continuous relaxations such as NOTEARS (Zheng et al., 2018) recast DAG search as constrained optimization amenable to gradients, and Bhattacharya et al. (2021) carry this idea to bow-free ADMGs, in which no pair of variables shares both a directed and a bidirected edge. Such graphs can still be nonidentifiable on a measure-zero set of parameters, yet the BIC scores them, and its nominal-dimension approximation can fail at singular points; graphs with bows, which this search excludes, can be nonidentifiable at every parameter (Example 1). Score-based causal discovery over linear ADMGs thus faces two coupled challenges: (i) a consistent score for singular model classes, and (ii) an efficient search over ADMG equivalence classes.

This paper aims to address these challenges with a score-based approach for linear Gaussian ADMGs. (1) We prove prior-almost-every-target consistency of exact WBIC optimization over a finite family of linear Gaussian ADMGs with compact analytic priors. The identification target is the algebraic equivalence class (AEC), the graphs sharing the same polynomial covariance constraints. We verify Watanabe’s fundamental conditions for ADMGs (Watanabe, 2013, 2009) and derive a strict RLCT gap outside the true AEC from polynomial geometry, accommodating bows and nonidentifiable parametrizations without presuming an RLCT faithfulness. (2) We develop a scalable ADVI approximation of the WBIC (Kucukelbir et al., 2017) that replaces the intractable tempered posterior expectation with a mean-field Gaussian surrogate. (3) We propose a differentiable search over ADMG equivalence classes that relaxes directed and bidirected adjacency with Gumbel-Softmax gates and reports the recovered Markov equivalence class. Proofs and experimental details are provided in the appendix.

1.1 Preliminaries

Capital letters denote random variables (X), lowercase their values (x), and bold letters sets ($\mathbf{V}$) and their values ($\mathbf{v}$); $P(\mathbf{V})$ is the distribution over $\mathbf{V}$.

We adopt the language of *structural causal models* (SCMs) Pearl (2000). A linear semi-Markovian SCM M over endogenous variables $\mathbf{V}$ and exogenous variables $\mathbf{U}$ is a system $V = \Lambda V + U$, in which a coefficient $\lambda_{ij} \neq 0$ signals a direct effect of V_j on V_i . The disturbances $\mathbf{U}$ are centered Gaussian with *error covariance* $\mathcal{E}$: its diagonal holds the error variances, and an off-diagonal $\omega_{ij} \neq 0$ indicates an unobserved common cause of V_i and V_j . For acyclic models the endogenous covariance is $\Sigma = (I - \Lambda)^{-1} \mathcal{E} (I - \Lambda)^{-\top}$. The theoretical model retains free error variances; variables are standardized to unit variance for empirical evaluation. The causal graph of M is an *Acyclic Directed Mixed Graph* (ADMG) $\mathcal{G} = (\mathbf{V}, \mathbf{D}, \mathbf{B})$ Richardson and Spirtes (2002): directed edges $\mathbf{D}$ ($\rightarrow$) mirror the nonzero entries of Λ , bidirected edges $\mathbf{B}$ ($\leftrightarrow$) mirror the nonzero off-diagonal entries of $\mathcal{E}$, and we identify coefficients with their edges when no confusion arises.

Given n i.i.d. observational samples $\mathbf{D}_n = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$, Bayesian model selection ranks candidate classes $\mathcal{M}$ by the *Bayesian free energy*, the negative log evidence $F_n(\mathcal{M}) = -\log \int \prod_{i=1}^n p(\mathbf{v}_i | \theta, \mathcal{M}) \varphi(\theta | \mathcal{M}) d\theta$, where θ gathers the free entries of Λ and $\mathcal{E}$ and φ is a prior; smaller values are preferred. For *regular* models, curved exponential families with nonsingular Fisher information, a Laplace approximation yields

$$F_n(\mathcal{M}) = -\log p(\mathbf{D}_n | \hat{\theta}, \mathcal{M}) + \frac{d}{2} \log n + O_p(1), \quad (1)$$

whose leading terms, with MLE $\hat{\theta}$ and $d = \dim(\theta)$, constitute the Bayesian Information Criterion (BIC) Schwarz (1978); Haughton (1988). Linear Gaussian ADMGs with correlated errors can violate regularity, and Section 3 traces the consequences.

Score-based discovery then seeks the structure with the best score. In place of combinatorial search Chickering (2002), continuous relaxations such as NOTEARS Zheng et al. (2018) optimize a weighted adjacency matrix W by gradient methods, enforcing acyclicity through $h(W) = \text{tr}[\exp(W \circ W)] - p$, which vanishes exactly when W is acyclic ($\circ$ is the Hadamard product and p the number of variables).

2 CAUSAL DISCOVERY IN SINGULAR MODELS

Linear Gaussian SCMs over ADMGs are not, in general, regular statistical models, and at their singulari-

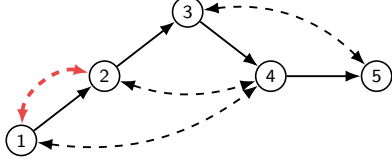


Figure 1: The bow ADMG $\mathcal{G}$ of Example 1, a variant of Drton et al.’s Example 1 (Drton et al., 2011) with the bidirected edge $1 \leftrightarrow 5$ moved to $1 \leftrightarrow 2$; solid edges are directed, dashed edges bidirected. The highlighted edge $1 \leftrightarrow 2$ (red) forms a bow with $1 \rightarrow 2$; removing it yields $\mathcal{G}'$, which induces the same model.

ties the BIC expansion of Eq. (1) no longer tracks the free energy (Drton, 2009; Drton and Maathuis, 2017). We first exhibit a case in which this failure changes the selected graph, then introduce the WBIC and its computation, and revisit the example.

Example 1. Write $\phi_{\mathcal{G}}$ for the parametrization that maps the structural parameters $(\Lambda, \mathcal{E})$ of a graph $\mathcal{G}$ to the observed covariance matrix Σ . Consider the five-node ADMG $\mathcal{G}$ of Fig. 1, with directed edges $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$ and bidirected edges $1 \leftrightarrow 2$, $1 \leftrightarrow 4$, $2 \leftrightarrow 4$, and $3 \leftrightarrow 5$. It modifies Example 1 of Drton et al. (2011) (Appendix C) by moving the bidirected edge $1 \leftrightarrow 5$ to $1 \leftrightarrow 2$, which forms a bow with $1 \rightarrow 2$. Because X_1 has no parents, no instrument identifies this bow: for every parameter, replacing U_2 by $U_2 - (\omega_{12}/\omega_{11})U_1$ maps $(\lambda_{12}, \omega_{12}, \omega_{22}, \omega_{24})$ to $(\lambda_{12} + \omega_{12}/\omega_{11}, 0, \omega_{22} - \omega_{12}^2/\omega_{11}, \omega_{24} - \omega_{12}\omega_{14}/\omega_{11})$ and leaves Σ unchanged. Each Σ is therefore generated by a curve of parameters, $\phi_{\mathcal{G}}$ is injective nowhere, and the model has $d = 13$ free parameters but only 12 effective degrees of freedom (Prop. 5 in Appendix B). We take $\lambda_{12} = 3$, $\lambda_{23} = -\frac{1}{2}$, $\lambda_{34} = \lambda_{45} = 1$, and let $\mathcal{E}$ have diagonal entries 2 and $\omega_{12} = \omega_{14} = \omega_{24} = \omega_{35} = 1$.

We simulate $n = 1000$ samples and use the BIC to compare $\mathcal{G}$ with the bow-free submodel $\mathcal{G}' = \mathcal{G} \setminus \{1 \leftrightarrow 2\}$ (Table 1, $R = 30$ replicates, lower is better). The reparametrization above maps every parameter of $\mathcal{G}$ to one of $\mathcal{G}'$ with the same covariance, so the two graphs induce the same distributions and their maximized likelihoods coincide in every replicate. The BIC nevertheless charges $\mathcal{G}$ for all 13 parameters and prefers $\mathcal{G}'$ by exactly $\frac{1}{2} \log n \approx 3.45$ (8358.4 versus 8361.9), rejecting the true graph in every replicate; the gap grows with n , although no amount of data can separate the two graphs. The example exposes the breakdown of the nominal-dimension approximation and motivates a penalty attuned to the likelihood geometry for models with unobserved confounding. ■

Example 1 shows the nominal $\frac{d}{2} \log n$ penalty mispricing complexity at a singularity. The WBIC Watanabe (2013) bypasses the Laplace expansion around

Table 1: Selection scores (mean $\pm$ s.d. over $R = 30$ replications, $n = 1000$) for the bow graph $\mathcal{G}$ and its bow-free submodel $\mathcal{G}'$, which induce the same model; lower is better, and the last column gives the paired gap. **Bold** marks the graphs each criterion selects: the BIC keeps only $\mathcal{G}'$, penalizing $\mathcal{G}$ by exactly $\frac{1}{2} \log n \approx 3.45$ in every replicate, whereas the WBIC and ADVI keep both graphs of the AEC.

	$\mathcal{G}$ (Fig. 1)	$\mathcal{G}'$ (Fig. 1 w/o $1 \leftrightarrow 2$)	Gap $\mathcal{G} - \mathcal{G}'$
BIC	8361.88 $\pm$ 55.66	8358.42 $\pm$ 55.66	+3.45 $\pm$ 0.00
WBIC	8360.60 $\pm$ 55.73	8360.15 $\pm$ 55.80	+0.45 $\pm$ 2.56
ADVI	8328.56 $\pm$ 55.67	8327.75 $\pm$ 55.77	+0.81 $\pm$ 0.57

the MLE and approximates the free energy $F_n(\mathcal{M})$ of Section 2 by a single tempered average of the loss. With the empirical negative log-likelihood $L_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log p(\mathbf{v}_i | \theta, \mathcal{M})$, it is defined by

$$\text{WBIC}(\mathcal{M}) = \mathbb{E}_{\theta}^{\beta}[nL_n(\theta)], \quad \beta = \frac{1}{\log n}, \quad (2)$$

where $\mathbb{E}_{\theta}^{\beta}[g] = \int g(\theta) \pi_n^{\beta}(\theta) d\theta$ averages under the tempered posterior $\pi_n^{\beta}(\theta) \propto e^{-n\beta L_n(\theta)} \varphi(\theta | \mathcal{M})$, obtained by raising the likelihood to the power β . The inverse temperature $\beta = 1/\log n$ flattens the likelihood just enough for this expectation to match $F_n(\mathcal{M})$ to leading order, for singular and regular models alike (Watanabe, 2013), so the WBIC needs neither the MLE nor the effective dimension.

We next apply the WBIC to causal discovery in linear causal models over ADMGs, showing that it can mitigate the BIC’s mispricing of complexity at singularities such as the bow in Example 1. More specifically, $\mathbb{G}$ is a finite family of ADMGs on the same p observed variables, $\mathcal{M}(\mathcal{G})$ is the linear Gaussian model induced by $\mathcal{G} \in \mathbb{G}$, and we write $\text{WBIC}(\mathcal{G})$, or $\text{WBIC}_n(\mathcal{G})$ to stress the sample size, for $\text{WBIC}(\mathcal{M}(\mathcal{G}))$. Each graph $\mathcal{G}$ has parameter space $\Theta_{\mathcal{G}}$ and a fixed prior $\Pi_{\mathcal{G}}$ with density $\varphi_{\mathcal{G}}$ and compact support $W_{\mathcal{G}} \subset \Theta_{\mathcal{G}}$, satisfying the fundamental conditions of Watanabe (2013, Section 3). The parameter $\theta = (\Lambda, \mathcal{E})$ ranges over the entries permitted by $\mathcal{G}$ (zeros off $\mathbf{D}$ in Λ and off $\mathbf{B}$ in $\mathcal{E}$) and determines the model-implied covariance $\Sigma(\theta) = (I - \Lambda)^{-1} \mathcal{E} (I - \Lambda)^{-\top}$ and the distribution $p_{\mathcal{G}, \theta} = \mathcal{N}(0, \Sigma(\theta))$. The data-generating distribution is q , and $\mathcal{G}$ realizes q within $W_{\mathcal{G}}$ if $q = p_{\mathcal{G}, \theta}$ for some $\theta \in W_{\mathcal{G}}$. Structure learning thereby becomes model selection over $\mathbb{G}$: each graph is scored by Eq. (2) under its own parameter space and prior, and we select any minimizer $\hat{\mathcal{G}}_n \in \arg \min_{\mathcal{G} \in \mathbb{G}} \text{WBIC}_n(\mathcal{G})$.

For a fixed graph, $\text{WBIC}(\mathcal{G})$ is a posterior mean, so Markov chain Monte Carlo can estimate it without the normalizing constant of π_n^{β} . We run random-

walk Metropolis–Hastings (MH) on the free coordinates of $\theta = (\Lambda, \mathcal{E})$, rejecting any proposal whose error covariance is not positive definite (Algorithm 2, Appendix E). Discarding T_0 warm-up iterations of a chain $\theta^{(1)}, \dots, \theta^{(T)}$, we average the loss,

$$\widehat{\text{WBIC}}_{\text{MH}}(\mathcal{G}) = \frac{1}{T - T_0} \sum_{t=T_0+1}^T nL_n(\theta^{(t)}), \quad (3)$$

which converges to $\text{WBIC}(\mathcal{G})$ as $T \rightarrow \infty$ for an ergodic chain. Each candidate graph needs its own chain, which is affordable when only a few graphs are compared, as in the running example.

Example 2 (Example 1, continued). We score $\mathcal{G}$ and $\mathcal{G}'$ on the same $R = 30$ datasets with Eq. (3) at $\beta = 1/\log 1000 \approx 0.145$, running one chain of $T = 60,000$ iterations ($T_0 = 15,000$ warm-up) per graph and replicate, long enough to keep the Monte Carlo error small relative to the BIC’s $\frac{1}{2} \log n$ gap. The WBIC shows no systematic preference: the paired difference $\mathcal{G} - \mathcal{G}'$ is 0.45 ± 2.56 across replicates, and $\mathcal{G}$ scores lower in 14/30. Unlike the BIC, it does not penalize the redundant bow parameter (Table 1). ■

2.1 From RLCT to Algebraic Equivalence

We now account for and characterize what exact WBIC minimization identifies, in two steps: the WBIC splits into a fit term shared by all graphs that realize q plus a complexity penalty given by the RLCT, and, for almost every generating parameter drawn from the model prior, this penalty is smallest exactly on the algebraic equivalence class of the generating graph.

Proposition 1 (WBIC for linear Gaussian ADMGs). *Under the standing setup, let $\mathcal{G} \in \mathbb{G}$ realize the centered Gaussian truth q within $W_{\mathcal{G}}$, and let S_n denote the sample covariance of n i.i.d. observations from q . The empirical negative log-likelihood of $\mathcal{M}(\mathcal{G})$ is*

$$nL_n(\theta) = \frac{n}{2} \left[p \log 2\pi + \log \det \Sigma(\theta) + \text{tr}(\Sigma(\theta)^{-1} S_n) \right], \quad (4)$$

and, writing $\theta_{0,\mathcal{G}} \in W_{\mathcal{G}}$ for any parameter realizing q , the WBIC of $\mathcal{G}$ satisfies

$$\text{WBIC}(\mathcal{G}) = nL_n(\theta_{0,\mathcal{G}}) + \lambda_{\mathcal{G}} \log n + O_p(\sqrt{\log n}), \quad (5)$$

where $\lambda_{\mathcal{G}} = \lambda_{\mathcal{G}}(q)$ is the real log-canonical threshold (RLCT) for the model and prior at q .

Prop. 1 splits the WBIC into a fit term and a complexity penalty. The fit term is evaluated at a parameter $\theta_{0,\mathcal{G}}$ realizing q ; at a singularity this parameter need not be unique, but any such parameter induces q , so the term is unaffected. The penalty coefficient $\lambda_{\mathcal{G}}$

is the *real log-canonical threshold* (RLCT), the coefficient of $\log n$ in the asymptotic expansion of the free energy Watanabe (2009). It quantifies how fast the prior mass of near-optimal parameters shrinks as the likelihood concentrates: writing $K_{\mathcal{G}}(\theta) = \text{KL}(q \parallel p_{\mathcal{G},\theta})$ for a graph realizing q within its prior support, the standing setup gives

$$\Pi_{\mathcal{G}}\{\theta \in W_{\mathcal{G}} : K_{\mathcal{G}}(\theta) \leq \varepsilon\} = \varepsilon^{\lambda_{\mathcal{G}}(q)+o(1)}, \quad \varepsilon \downarrow 0,$$

where the exponent notation suppresses logarithmic factors (Watanabe, 2009). This KL neighborhood surrounds the whole set of parameters realizing q , and a smaller RLCT means slower decay and hence a smaller leading WBIC penalty. In a regular model with an interior truth and positive prior density, the KL divergence is locally quadratic in the d identifiable directions, so $\lambda_{\mathcal{G}} = d/2$ and Prop. 1 recovers the leading BIC terms of Eq. (1). At singularities, where multiple parameters realize q and form a positive-dimensional set, the RLCT can fall strictly below $d/2$, and the WBIC then charges a smaller penalty that reflects the singular geometry. Decomposition (5) thus weighs fit against this decay rate. A graph whose support cannot realize q incurs a positive KL gap and hence an order- n disadvantage; among graphs that realize q , the fit terms coincide, and a strict RLCT difference separates their scores at order $\log n$.

Which graph realizing q has the smallest RLCT? The answer lies in the polynomial equalities of the full covariance models: for almost every generating parameter, any competing graph that realizes q within its prior support but obeys different polynomial equalities suffers a strict RLCT gap. The equalities a graph imposes are the polynomials that vanish on every covariance matrix it can generate, that is, on the image of its covariance map $\phi_{\mathcal{G}}(\theta) = (I - \Lambda)^{-1} \mathcal{E}(I - \Lambda)^{-\top}$, $\theta \in \Theta_{\mathcal{G}}$, from Example 1. Grouping graphs by these equalities yields the following equivalence classes.

Definition 2 (Algebraic equivalence class (van Ommen and Mooij, 2017)). For $\mathcal{G} \in \mathbb{G}$, defines

$$\mathcal{C}_{\mathcal{G}} = \phi_{\mathcal{G}}(\Theta_{\mathcal{G}}), \quad V_{\mathcal{G}} = \overline{\mathcal{C}_{\mathcal{G}}}^Z,$$

where the bar denotes real Zariski closure, i.e., the smallest real algebraic set (common zero set of polynomials) containing $\mathcal{C}_{\mathcal{G}}$. ADMGs are *algebraically equivalent* if their full covariance models satisfy the same polynomial equalities, equivalently $V_{\mathcal{G}} = V_{\mathcal{H}}$. The algebraic equivalence class is defined as:

$$\text{AEC}(\mathcal{G}) = \{\mathcal{H} \in \mathbb{G} : V_{\mathcal{H}} = V_{\mathcal{G}}\}.$$

Algebraic equivalence covers conditional independences and other polynomial covariance constraints, such as tetrads (Sullivant et al., 2010) and Verma

constraints (Verma and Pearl, 1990; Tian and Pearl, 2002b). It is weaker than full model equivalence, but stronger than sharing one distribution: two graphs generating the same distribution need not be algebraically equivalent. Conversely, graphs with different parameter counts can be algebraically equivalent.

Example 3 (Example 1, continued: Zariski closure and AEC). Write $\Sigma = (\sigma_{ij})$. The closure V_G comprises the symmetric matrices satisfying every polynomial equality valid throughout $\mathcal{C}_G$, for instance $\sigma_{13}\sigma_{22} - \sigma_{12}\sigma_{23} = 0$, which expresses $X_1 \perp X_3 \mid X_2$ for positive-definite Gaussian covariances. The closure can leave the positive-definite cone: scaling $\mathcal{E}$ by $t \downarrow 0$ shows $0 \in V_G$, although $0 \notin \mathcal{C}_G$. Since neither graph contains $1 \leftrightarrow 5$, both also satisfy the tetrad constraint $\sigma_{14}\sigma_{25} - \sigma_{15}\sigma_{24} = 0$ (Sullivant et al., 2010). Together with $X_1 \perp X_5 \mid \{X_2, X_3, X_4\}$, these two equalities cut out V_G near generic points (Prop. 6 in Appendix B). More strongly, the reparametrization of Example 1 maps $\mathcal{C}_G$ into $\mathcal{C}_{G'}$, while $\mathcal{C}_{G'} \subseteq \mathcal{C}_G$ because $\mathcal{G}' \subset \mathcal{G}$. Hence $\mathcal{C}_G = \mathcal{C}_{G'}$ and $V_G = V_{G'}$, so $\text{AEC}(\mathcal{G}) = \{\mathcal{G}, \mathcal{G}'\}$ within $\mathbb{G} = \{\mathcal{G}, \mathcal{G}'\}$. This is also an instance of van Ommen and Mooij (2017, Theorem 2): $\mathcal{G}$ is generically infinite-to-one, and deleting the single edge $1 \leftrightarrow 2$ yields the identifiable graph $\mathcal{G}'$. The bow adds a parameter but no equality constraint. ■

The RLCT ties WBIC scores to the varieties V_G of Def. 2: for almost every generating parameter, the generating graph’s RLCT is half the dimension of its variety, and every compatible graph with a different variety pays strictly more.

Lemma 3 (RLCT gap outside the AEC). *Under the standing setup, fix $\mathcal{G}^* \in \mathbb{G}$. For $\Pi_{\mathcal{G}^*}$ -almost every θ^* , the distribution $q = p_{\mathcal{G}^*, \theta^*}$ satisfies*

- (i) $\lambda_{\mathcal{G}^*}(q) = \frac{1}{2} \dim V_{\mathcal{G}^*}$, and
- (ii) $\lambda_{\mathcal{G}}(q) > \frac{1}{2} \dim V_{\mathcal{G}^*}$ for every $\mathcal{G} \in \mathbb{G}$ that realizes q within $W_{\mathcal{G}}$ and satisfies $V_{\mathcal{G}} \neq V_{\mathcal{G}^*}$.

Both parts hold except on a measure-zero set of generating parameters, whose covariances satisfy additional equality constraints. Part (i) says that the generating graph’s RLCT is set by the dimension of its covariance model rather than by its number of free parameters: when $\mathcal{G}^*$ is generically identifiable, $\dim V_{\mathcal{G}^*}$ equals its parameter count d , and part (i) recovers the BIC penalty $d/2$. Part (ii) concerns competitors that also realize q : outside the excluded set, such a graph outside the AEC has a strictly larger variety, $V_{\mathcal{G}^*} \subsetneq V_{\mathcal{G}}$, and thus imposes fewer polynomial equalities, and it pays a strictly larger RLCT. Among graphs fitting q , the WBIC therefore favors those imposing the most equality constraints. Combining Prop. 1 with Lem. 3 yields the main identification guarantee.

Theorem 4 (AEC consistency). *Fix $\mathcal{G}^* \in \mathbb{G}$. For*

$\Pi_{\mathcal{G}^*}$ -almost every θ^* , if $q = p_{\mathcal{G}^*, \theta^*}$ and the observations are i.i.d. from q , then any measurable choice $\hat{\mathcal{G}}_n \in \arg \min_{\mathcal{G} \in \mathbb{G}} \text{WBIC}_n(\mathcal{G})$ satisfies

$$P_q^{\otimes n} \left\{ \hat{\mathcal{G}}_n \in \text{AEC}(\mathcal{G}^*) \right\} \longrightarrow 1.$$

The two quantifiers must be kept apart: “almost every” concerns the generating parameter θ^* under its prior, whereas, for each such parameter, identification holds in probability over samples. The guarantee also degrades gracefully under prior misspecification: if θ^* is instead drawn from a different prior $\Pi_{\mathcal{H}}$, the asymptotic prediction error is at most the total variation distance $d_{\text{TV}}(\Pi_{\mathcal{H}}, \Pi_{\mathcal{G}^*})$ between priors. See Prop. 15 in Appendix D.4 for a detailed analysis.

Example 4 (Example 1, continued: AEC identification). Read against Lem. 3 and Thm. 4, Table 1 has a direct interpretation. The singularity of $\mathcal{G}$ is generic: every parameter lies on a curve of equivalent parameters, so the target is not one of the excluded measure-zero points, and Lem. 3(i) gives $\lambda_{\mathcal{G}} = \frac{1}{2} \dim V_{\mathcal{G}} = 6$ rather than the nominal $d/2 = 13/2$. The submodel $\mathcal{G}'$ is regular at its representation of q and has the same variety, so $\lambda_{\mathcal{G}'} = 6$ as well, and the two graphs share the fit term. Thm. 4 therefore guarantees that the WBIC minimizer lands in $\text{AEC}(\mathcal{G}) = \{\mathcal{G}, \mathcal{G}'\}$; it does not determine which member is selected, consistent with Example 2. The BIC’s choice $\mathcal{G}'$ also lies in this class, so its failure is not one of AEC identification: the $\frac{1}{2} \log n$ gap is an error in approximating the free energy, which separates two observationally equivalent graphs and discards the true one. ■

Thm. 4 leaves open what the WBIC identifies at targets in the excluded measure-zero set. There, graphs from different AECs can generate the same distribution with the same RLCT. A natural conjecture is a local AEC identification: the WBIC should not discriminate among graphs that realize the target with equal RLCT. Drton et al.’s original example, whose singularity occurs at such a point, illustrates this direction (Appendix C): the WBIC keeps both graphs, whereas the BIC selects the sparser one, which lies outside the generating graph’s AEC (Prop. 9).

3 DIFFERENTIABLE CAUSAL DISCOVERY

Thm. 4 guarantees that exact WBIC optimization identifies the true AEC for prior-almost every generating parameter, but it is not yet a practical discovery procedure: each candidate graph requires a costly tempered-posterior expectation, and the graph search space is discrete. We remove these obstacles in turn, first making the WBIC computation scalable

for a fixed ADMG through variational inference, then reusing the same approximation inside a continuous relaxation of the ADMG structure.

Fix an ADMG $\mathcal{G} = (\mathbf{V}, D, B)$ with directed and bidirected edge sets D and B . Its parameters $\theta = (\Lambda, \Omega)$ comprise the coefficients λ_{ij} of the directed edges $V_j \rightarrow V_i \in D$ and, in ω , the error variances together with the covariances of the bidirected edges $V_i \leftrightarrow V_j \in B$; every other coefficient and error covariance is held at zero. Rather than run a Metropolis–Hastings chain for each candidate graph, we approximate the tempered posterior $\pi_n^\beta(\theta \mid \mathbf{D}_n, \mathcal{G}) \propto p(\mathbf{D}_n \mid \theta, \mathcal{G})^\beta \varphi_{\mathcal{G}}(\theta)$ on $W_{\mathcal{G}}$, which defines the WBIC of Eq. (2), with Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2017).

ADVI operates in an unconstrained space, but the error covariance must be symmetric positive definite. We therefore reparameterize $\zeta = T_{\mathcal{G}}(\theta)$, with $\zeta = (\zeta_\lambda, \zeta_L, \zeta_{\text{diag}}) \in \mathbb{R}^{d_{\mathcal{G}}}$. The map is the identity on the directed coefficients, $\lambda_{ij} = \zeta_{\lambda, ij}$ for $V_j \rightarrow V_i \in D$, and builds the error covariance from a Cholesky factor, $\Omega_\omega = LL^\top$, whose free entries are its diagonal and its entries on bidirected edges; the remaining subdiagonal entries are completed so that Ω_ω inherits the zero pattern of B : $L_{ii} = \exp(\zeta_{\text{diag}, i})$, and for $i > j$,

$$L_{ij} = \begin{cases} \zeta_{L, ij}, & V_i \leftrightarrow V_j \in B, \\ -\frac{1}{L_{jj}} \sum_{k < j} L_{ik} L_{jk}, & \text{otherwise,} \end{cases} \quad (6)$$

filled row by row so that each completed entry depends only on entries already determined. Since $(LL^\top)_{ij} = \sum_{k < j} L_{ik} L_{jk} + L_{ij} L_{jj}$, the completion gives $(\Omega_\omega)_{ij} = 0$ for every pair without a bidirected edge, and the positive diagonal gives $\Omega_\omega \succ 0$. This is the free-entry construction for G-Wishart matrices (Atay-Kayis and Massam, 2005), applied to the covariance rather than the precision. Conversely, each positive-definite Ω with the zero pattern of B has a unique Cholesky factor with positive diagonal, whose entries off B obey the completion. Hence $\zeta \mapsto \Omega_\omega$ is a smooth bijection from $\mathbb{R}^{p+|B|}$ onto the admissible error covariances, and $T_{\mathcal{G}}$ is a diffeomorphism from $\Theta_{\mathcal{G}}$ onto $\mathbb{R}^{d_{\mathcal{G}}}$.

We approximate the tempered posterior by a mean-field Gaussian $q_\psi(\zeta) = \prod_{k \in \mathcal{I}} \mathcal{N}(\zeta_k; \mu_k, \sigma_k^2)$, where $\mathcal{I}$ indexes the free coordinates ($\zeta_{\lambda, ij}$ for $(i, j) \in D$, $\zeta_{L, ij}$ for $i > j$ with $(i, j) \in B$, and $\zeta_{\text{diag}, i}$ for $1 \leq i \leq p$) and $\psi = (\mu, \log \sigma)$. Mapping back through $\theta = T_{\mathcal{G}}^{-1}(\zeta)$ approximates the tempered posterior over θ , whose density in ζ coordinates is $\pi_n^\beta(\zeta \mid \mathbf{D}_n, \mathcal{G}) \propto p(\mathbf{D}_n \mid T_{\mathcal{G}}^{-1}(\zeta), \mathcal{G})^\beta \pi_\zeta(\zeta \mid \mathcal{G})$. We fit ψ by maximizing the tempered evidence lower bound (ELBO),

$$\mathcal{L}_\beta(\psi; \mathcal{G}) = \mathbb{E}_{q_\psi(\zeta)}[\beta \ell_n(\zeta; \mathbf{D}_n, \mathcal{G})] - \text{KL}(q_\psi(\zeta) \parallel \pi_\zeta(\zeta)), \quad (7)$$

using the reparameterization $\zeta = \mu + \sigma \odot \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$, for Monte Carlo gradients and stochastic gradient ascent, and write $\psi^* = \arg \max_\psi \mathcal{L}_\beta(\psi; \mathcal{G})$. The fit q_{ψ^*} then stands in for the tempered posterior, and we estimate the WBIC by averaging the loss over its samples,

$$\widehat{\text{WBIC}}_\beta(\mathcal{G}) = \frac{1}{S} \sum_{s=1}^S n L_n(\theta^{(s)}), \quad (8)$$

$$\theta^{(s)} = T_{\mathcal{G}}^{-1}(\mu^* + \sigma^* \odot \varepsilon_s), \quad \varepsilon_s \sim \mathcal{N}(0, I).$$

The subscript β records the temperature at which q_{ψ^*} is fitted. The single temperature $\beta = 1/\log n$ keeps this approximation compatible with singular learning theory: the effective sample size $n\beta = n/\log n$ lets the posterior concentrate on the optimal set, yet slowly enough that its leading behavior is governed by the same singular geometry as the WBIC.¹

Example 5 (Example 1, continued). On the $n = 1000$ data of Example 1, the ADVI surrogate (Eqs. (7)–(8)) avoids most of the BIC’s penalty on the redundant parameter (Table 1): its paired gap is 0.8 ± 0.6 in favor of $\mathcal{G}'$, under a quarter of the BIC’s $\frac{1}{2} \log n \approx 3.45$. The consistent sign of this small gap ($\mathcal{G}'$ scores lower in 28 of 30 replicates) suggests that the mean-field family slightly underfits the curve of equivalent parameters of $\mathcal{G}$. This finite-sample evidence motivates the variational score for differentiable search, without establishing its asymptotic consistency. ■

3.1 Differentiable Search over ADMGs

So far WBIC-ADVI assumes a fixed pair of masks (D, B) . To search over ADMGs, we relax them to continuous random gates whose logits we optimize jointly with the variational parameters, leaving the parameters, priors, variational family, and score untouched; Algorithm 1 summarizes the details of the proposed method. Specifically, let α_{ij}^D be the logit of the directed edge $V_j \rightarrow V_i$ ($i \neq j$) and α_{ij}^B that of the bidirected edge $V_i \leftrightarrow V_j$ ($i < j$), with edge probabilities $P_{ij}^D = \sigma(\alpha_{ij}^D) \mathbf{1}\{i \neq j\}$ and $P_{ij}^B = \sigma(\alpha_{ij}^B)$. During training, we draw binary concrete gates

$$\tilde{A}_{ij}^e = \sigma\left(\frac{\alpha_{ij}^e + \log u_{ij}^e - \log(1 - u_{ij}^e)}{\tau}\right), \quad e \in \{D, B\}, \quad (9)$$

with $u_{ij}^e \sim \text{Unif}(0, 1)$ and Gumbel temperature $\tau > 0$, setting $\tilde{A}_{ii}^D = 0$, $\tilde{A}_{ii}^B = 1$, and $\tilde{A}_{ji}^B = \tilde{A}_{ij}^B$.

Each sampled graph enters the construction through $D = \tilde{A}^D$ and $B = \tilde{A}^B$: the directed coefficients be-

¹In practice we evaluate the surrogate at a colder $\beta = c/\log n$ with $c = 4$ (any $c \in [2, 4]$ works well), a heuristic that offsets the mean-field family’s underestimation of posterior correlations along the singular set while keeping the over-parameterized graph distinguishable from its submodels; the theory uses $\beta = 1/\log n$.

Algorithm 1 Differentiable discovery over ADMGs

Require: data $\mathbf{D}_n$, β , τ , penalties $\tilde{\lambda}_D, \tilde{\lambda}_B, \tilde{\lambda}_{\text{acyc}}$

- 1: Initialize α^D, α^B and ψ .
- 2: **for** $t = 1, \dots, T$ **do**
- 3: Sample gates $\tilde{A}^D, \tilde{A}^B$ by Eq. (9).
- 4: Sample $\zeta^{(m)} = \mu + \sigma \odot \varepsilon^{(m)}$, $\varepsilon^{(m)} \sim \mathcal{N}(0, I)$, and build $\theta^{(m)} = (\Lambda, \Omega)$ by Eq. (10), $m \leq M$.
- 5: Ascend $\mathcal{L}_\beta(\psi; \tilde{A}^D, \tilde{A}^B)$ of Eq. (7) in ψ .
- 6: Resample $\theta^{(m)}$ from the updated q_ψ and estimate $\widehat{\text{WBIC}}_\beta = \frac{1}{M} \sum_m n L_n(\theta^{(m)})$.
- 7: Descend $\mathcal{J}_\beta$ of Eq. (11) in α^D, α^B ; update $\tilde{\lambda}_{\text{acyc}}$.
- 8: **end for**
- 9: Threshold P^{D*}, P^{B*} at κ_D, κ_B ; report the ADMG $(\hat{D}, \hat{B})$ and its PAG (Zhang, 2008b).

come $\Lambda = \tilde{A}^D \circ \zeta_\Lambda$, and each subdiagonal entry of L interpolates between its free and completed values,

$$L_{ij} = \tilde{A}_{ij}^B \zeta_{L,ij} + (1 - \tilde{A}_{ij}^B) \left(-\frac{1}{L_{jj}} \sum_{k < j} L_{ik} L_{jk} \right), \quad (10)$$

for $i > j$, with the diagonal of Eq. (6). At binary gates, Eq. (10) reduces exactly to Eq. (6), so every discrete ADMG reached by the relaxation is evaluated on its own model class. Let $\mathcal{L}_\beta(\psi; \tilde{A}^D, \tilde{A}^B)$ be the tempered ELBO of Eq. (7) with the masks replaced by these gates, its expected log-likelihood estimated on reparameterized samples $\zeta^{(1)}, \dots, \zeta^{(M)}$ and its KL term taken over all candidate coordinates. We learn a sparse acyclic ADMG with two coupled updates. The variational parameters ψ maximize $\mathcal{L}_\beta$, which is where the temperature enters: q_ψ approximates the tempered posterior of the relaxed model instead of collapsing onto a likelihood optimum. The graph logits minimize the penalized WBIC estimate

$$\begin{aligned} \mathcal{J}_\beta(\alpha^D, \alpha^B) = & \widehat{\text{WBIC}}_\beta + \tilde{\lambda}_D \sum_{i \neq j} P_{ij}^D \\ & + \tilde{\lambda}_B \sum_{i < j} P_{ij}^B + \tilde{\lambda}_{\text{acyc}} h(P^D), \end{aligned} \quad (11)$$

where $\widehat{\text{WBIC}}_\beta$ is the estimate of Eq. (8) for the relaxed graph under q_ψ , h is the NOTEARS acyclicity function (Zheng et al., 2018), and $\tilde{\lambda}_D, \tilde{\lambda}_B, \tilde{\lambda}_{\text{acyc}}$ govern sparsity and acyclicity. Algorithm 1 alternates single gradient steps on the two objectives, obtaining both gradients by reparameterizing ζ and the gates.

After optimization we threshold the edge probabilities, $\hat{D}_{ij} = \mathbf{1}\{P_{ij}^{D*} > \kappa_D\}$ and $\hat{B}_{ij} = \mathbf{1}\{P_{ij}^{B*} > \kappa_B\}$ for $i < j$. A pair lacking a reported bidirected edge has $\omega_{ij} = 0$ under Eq. (6), so $(\hat{D}, \hat{B})$ is an ADMG in the model class of the Preliminaries, which we report together with its PAG. Whereas exact WBIC

optimization targets the generating graph’s AEC for prior-almost every parameter (Thm. 4), we evaluate on the coarser conditional-independence information of a PAG (Zhang, 2008a), which is computable. PAG agreement does not imply agreement on the further polynomial constraints that define an AEC. Reporting the AEC itself would require a graphical characterization of algebraic equivalence for linear Gaussian ADMGs, which remains open; van Ommen and Mooij (2017) use the half-trek criterion to give sufficient graphical conditions for algebraic equivalence.

4 EXPERIMENTS

We evaluate the ADVI surrogate for WBIC (ADVI below) on ADMGs that are singular by construction, where regular asymptotic criteria are unreliable and the singular correction should matter most, on randomized ADMGs of increasing size to gauge behavior on generic (mostly regular) structures, and on the Sachs protein-signaling benchmark to check that the method recovers a meaningful PAG on real data.

In all experiments, we compare against ABIC Bhat-tacharya et al. (2021) as the baseline. For a common evaluation target, we convert both the ground-truth (D, B) and the estimated $(\hat{D}, \hat{B})$ to PAGs and compute all metrics at the PAG level: structural Hamming distance (SHD), Skeleton F1, Arrowhead F1. These metrics assess conditional-independence structure and are coarser than AEC identification; they do not require WBIC to be invariant within a Markov equivalence class. For PAGs, SHD adds 1 when the presence of an edge differs and, when an edge is present in both, adds 1 for each endpoint mark that differs. The appendix details the experimental setup.

Singular ADMGs. We first evaluate on three ADMGs that are singular by construction, denoted Singular I–III. Singular I (3 nodes) has directed $1 \rightarrow 3$, $2 \rightarrow 3$ and bidirected $1 \leftrightarrow 2$, $1 \leftrightarrow 3$; Singular II (4 nodes) has directed $1 \rightarrow 3$, $3 \rightarrow 4$, $2 \rightarrow 4$ and bidirected $1 \leftrightarrow 2$, $1 \leftrightarrow 4$, $2 \leftrightarrow 3$; Singular III (5 nodes) has directed $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$ and bidirected $1 \leftrightarrow 4$, $1 \leftrightarrow 5$, $2 \leftrightarrow 4$, $3 \leftrightarrow 5$. Data are generated from the linear Gaussian ADMG model $X = \varepsilon(I - \Lambda)^{-1}$, and $\varepsilon \sim \mathcal{N}(0, \Omega)$ with parameters chosen so that the weighted model lies on a singular locus. For Singular I every valid parameter is singular: we draw the nonzero directed coefficients from $\text{Unif}(-1, 1)$, the diagonal of Ω from $\text{Unif}(1, 3)$, and the nonzero bidirected entries from $\text{Unif}(-1, 1)$, rejecting until $\Omega \succ 0$. For Singular II and III most free parameters are sampled the same way, and specific entries are then solved so that the relevant rank condition fails— λ_{13} for Singular II, and ω_{24}, ω_{44} for Singular III (full formulas in

Table 2: PAG-level results for singular and randomized ADMGs, averaged over 10 runs per configuration with $n = 2000$. Entries are mean $\pm$ standard deviation.

Singular					General				
Graph	Method	SHD	Skel. F1	Arrow. F1	Graph	Method	SHD	Skel. F1	Arr. F1
I	ADVI	0.20 ± 0.42	0.96 ± 0.08	0.00 ± 0.00	4 nodes	ADVI	0.60 ± 1.07	0.96 ± 0.07	0.27 ± 0.44
	ABIC	1.40 ± 1.17	0.84 ± 0.08	0.00 ± 0.00		ABIC	1.50 ± 1.72	0.86 ± 0.20	0.10 ± 0.32
II	ADVI	0.40 ± 0.97	0.98 ± 0.04	0.00 ± 0.00	5 nodes	ADVI	1.60 ± 1.96	0.93 ± 0.08	0.40 ± 0.52
	ABIC	0.40 ± 0.52	0.96 ± 0.05	0.00 ± 0.00		ABIC	3.10 ± 2.38	0.83 ± 0.12	0.34 ± 0.46
III	ADVI	6.00 ± 1.56	0.88 ± 0.08	0.02 ± 0.06	8 nodes	ADVI	12.10 ± 6.12	0.83 ± 0.12	0.56 ± 0.30
	ABIC	7.00 ± 1.89	0.83 ± 0.07	0.07 ± 0.18		ABIC	14.50 ± 4.62	0.83 ± 0.08	0.50 ± 0.14

the appendix). Every sampled point is checked to satisfy $\Omega \succ 0$ with a rank-deficient condition. In this singular regime, a larger β amplifies the score gap between the singular model and its competitors, so we set $\beta = 4/\log n$. For each graph, we generate $R = 10$ independent datasets of size $n = 2000$.

Table 2 reports averages over the ten runs. On Singular I and II both methods recover very similar skeletons; on the harder five-node Singular III, ADVI attains lower SHD and higher skeleton F1.

Randomized ADMGs. To assess performance beyond hand-constructed singular instances, we evaluate on randomly generated ADMGs of increasing size ($p \in \{4, 5, 8\}$ nodes). For each size we sample random directed and bidirected structures and draw parameters from the same linear Gaussian model as above, generating $R = 10$ datasets of size $n = 2000$ per configuration. Because these graphs are generic, most parameter points are regular, so this setting probes whether the singular correction remains competitive when it is not strictly needed. Alongside skeleton F1 we report arrowhead F1 to capture orientation quality, together with SHD and its standard deviation.

Table 2 shows that ADVI consistently attains lower SHD than ABIC and generally achieves comparable or better F1 scores. For skeleton recovery, ADVI performs better at 4 and 5 nodes, while the two methods are nearly indistinguishable at 8 nodes, with ABIC having a slightly smaller variance. For arrowhead recovery, ADVI outperforms ABIC across all three graphs. The SHD advantage of ADVI increases with graph size. The larger SHD standard deviation of ADVI at $p = 8$ indicates increased variability as the graph becomes more complex.

Sachs Data. Finally, we evaluate on a log-transformed version of the Sachs et al. (2005) protein signaling dataset, following the preprocessing of Ramsey and Andrews (2018); the ground-truth graph is taken from Figure 4 of the same study. At the PAG level, ADVI recovers a biologically meaningful struc-

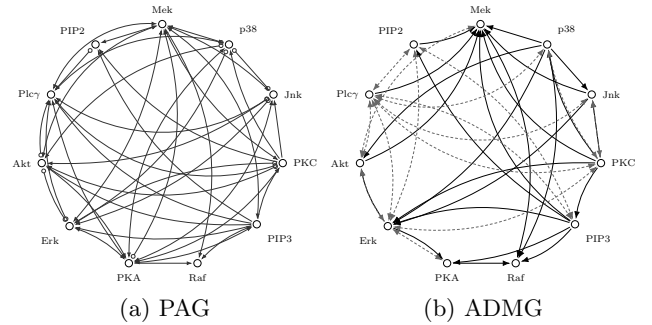


Figure 2: Causal structures learned by ADVI on the Sachs data. (a) Complete PAG. (b) Returned ADMG.

ture, achieving an SHD of 35 against 38 for ABIC and a higher skeleton F1 (0.79 vs. 0.73). The gain is driven by recall: ADVI recovers more of the true skeleton (0.91 vs. 0.72), while ABIC is marginally more precise (0.74 vs. 0.71), reflecting its sparser output. Fig. 2a shows the PAG estimated by ADVI, whose edges align with the benchmark’s canonical signaling pathways.

5 CONCLUSION

We presented a score-based framework for causal discovery over linear Gaussian ADMGs grounded in singular learning theory. For finite candidate families with compact analytic priors, exact WBIC optimization identifies the generating graph’s algebraic equivalence class with probability tending to one for prior-almost every generating parameter. The proof derives the required RLCT separation from polynomial geometry, allowing bows and nonidentifiable parametrizations; singular targets in the excluded measure-zero set need not be identified. We developed an ADVI approximation and a differentiable search over directed and bidirected adjacency, reporting PAGs as a coarser evaluation target. Establishing the variational expansion and score separation, characterizing behavior at such targets, and evaluating AEC identification beyond PAG-level constraints remain future directions.

References

- Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92(2):317–335.
- Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2022). On Pearl’s hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. ACM.
- Bellot, A., Zhang, J., and Bareinboim, E. (2024). Scores for learning discrete causal graphs with unobserved confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11043–11051.
- Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. (2021). Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 2314–2322.
- Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.
- Colombo, D., Maathuis, M., Kalisch, M., and Richardson, T. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321.
- Drton, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012.
- Drton, M., Eichler, M., and Richardson, T. S. (2009). Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(81):2329–2348.
- Drton, M., Foygel, R., and Sullivant, S. (2011). Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886.
- Drton, M. and Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393.
- Evans, R. J. (2016). Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648.
- Geiger, D., Heckerman, D., King, H., and Meek, C. (2001). Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505–529.
- Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Academic Press, Orlando, FL.
- Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355.
- Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464.
- Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR)*.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45.
- Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations (ICLR)*.
- Meek, C. (1995). Causal inference and causal explanation with background knowledge. In Besnard, P. and Hanks, S., editors, *Uncertainty in Artificial Intelligence 11*, pages 403–410. Morgan Kaufmann, San Francisco.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
- Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 435–443.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.
- Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.
- Ramsey, J. and Andrews, B. (2018). FASK with interventional knowledge recovers edges from the Sachs model. *arXiv preprint arXiv:1805.03108*.
- Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030.
- Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

- Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, 2nd edition.
- Sullivant, S., Talaska, K., and Draisma, J. (2010). Trek separation for Gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685.
- Tian, J. and Pearl, J. (2002a). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI)*, pages 567–573.
- Tian, J. and Pearl, J. (2002b). On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 519–527.
- Uhler, C., Raskutti, G., Bühlmann, P., and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463.
- van Ommen, T. and Mooij, J. M. (2017). Algebraic equivalence of linear structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, pages 352–359.
- Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270.
- Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, Cambridge, UK.
- Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897.
- Zhang, J. (2008a). Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(47):1437–1474.
- Zhang, J. (2008b). 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.
- Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. (2018). DAGs with NOTEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31.

Singular Learning Theory for Causal Discovery under Unobserved Confounding:

Supplementary Materials

A RELATED WORK

This appendix situates our contribution within five lines of research: constraint-based discovery under latent confounding, equality constraints beyond conditional independence, score-based discovery with the BIC, the singular statistics of linear structural equation models, and continuous (gradient-based) structure learning. We review each in turn and close by positioning our contribution against the closest prior work.

Constraint-based discovery and graphical models of confounding. A first family of methods recovers causal structure from the conditional independences in the data. The semantics linking d -separation in a graph to conditional independence in the distribution (Verma and Pearl, 1988; Pearl, 1988), together with the faithfulness assumption (Meek, 1995; Uhler et al., 2013), makes these independences testable constraints that prune the space of compatible graphs. When some variables are unobserved, the appropriate representation is no longer a DAG but a mixed graph: ancestral graphs and the closely related acyclic directed mixed graphs (ADMGs) encode both directed effects and bidirected edges for latent common causes while retaining a well-characterized Markov structure (Richardson and Spirtes, 2002; Zhang, 2008a). The constraint-based PC algorithm (Spirtes and Glymour, 1991) recovers structure from independence tests in the fully observed case, and the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2000) together with its refinements extends this to latent confounding, learning the corresponding partial ancestral graph (PAG); Zhang (2008b) established the completeness of its orientation rules under latent confounding and selection bias, and Colombo et al. (2012) extended the approach to the high-dimensional regime. These methods are consistent but rely solely on conditional-independence constraints and inherit the statistical fragility of repeated hypothesis tests.

Equality constraints beyond conditional independence. Linear models with correlated errors impose constraints on the observed covariance that go strictly beyond conditional independence. Trek separation characterizes the vanishing of subdeterminants (minors) of the covariance matrix in Gaussian graphical models, generalizing the classical tetrad conditions (Sullivant et al., 2010). Such constraints reflect the testable implications of latent-variable models more broadly (Pearl, 1995), and the algebraic and Markov structure of marginalized (latent-projected) models has been studied through the margins of Bayesian networks (Evans, 2016). Relatedly, identification theory characterizes when a causal effect can be recovered in the presence of latent confounders (Tian and Pearl, 2002a). These results motivate scoring whole model classes rather than testing individual independences, but they do not by themselves yield a tractable selection criterion.

Score-based discovery and the BIC. Score-based methods instead assign each candidate structure a goodness-of-fit-plus-complexity score and search for its optimum. The Bayesian Information Criterion (Schwarz, 1978; Haughton, 1988) is the canonical choice, and greedy search over Markov equivalence classes (GES) identifies the optimal structure for fully observed DAGs (Chickering, 2002), with extensions to interventional equivalence classes (Hauser and Bühlmann, 2012). For linear ADMGs with correlated errors, the maximum-likelihood fit underlying any such score can be computed by iterative conditional fitting (Drton et al., 2009). The asymptotic justification of the BIC, however, presumes a regular (curved exponential family) model with nonsingular Fisher information (Geiger et al., 2001); precisely this assumption fails for the latent-confounded linear Gaussian models we consider.

Singular models and identifiability of linear SEMs. Drton (2009) showed that linear Gaussian models with correlated errors can have a singular Fisher information at certain parameter values, so that likelihood-ratio statistics depart from their nominal χ^2 asymptotics. Drton et al. (2011) characterized when the parametrization of a linear structural equation model is globally (as opposed to merely generically) identifiable, exhibiting the singular instances that the BIC misscores; Drton and Maathuis (2017) survey the resulting challenges for structure

learning with latent variables. Watanabe’s singular learning theory (Watanabe, 2009) provides the geometry needed to handle such models, replacing the model dimension by the real log-canonical threshold, and the Widely Applicable Bayesian Information Criterion (WBIC) (Watanabe, 2013) turns this into a single tempered-posterior estimator of the leading free-energy terms. Bellot et al. (2024) bring the WBIC to bear on structure learning directly, scoring candidate graphs by this singular-aware criterion through a greedy search. Our score builds directly on this theory and verifies its hypotheses for linear ADMGs.

Continuous and differentiable structure learning. A more recent line recasts the combinatorial search over graphs as continuous optimization. NOTEARS (Zheng et al., 2018) replaces the discrete acyclicity constraint with a smooth algebraic penalty, enabling gradient-based DAG learning, but targets the fully observed setting. Gradients through discrete edge indicators are obtained with the Gumbel-Softmax / Concrete relaxation (Jang et al., 2017; Maddison et al., 2017), and otherwise intractable Bayesian objectives are made differentiable through automatic differentiation variational inference (Kucukelbir et al., 2017). Closest to our setting, Bhattacharya et al. (2021) develop a differentiable causal-discovery procedure for ADMGs under unmeasured confounding, relaxing the bidirected structure alongside the directed one.

Positioning. Our approach differs from this prior work in the score it optimizes. Continuous methods such as NOTEARS apply to fully observed DAGs, and existing differentiable ADMG discovery (Bhattacharya et al., 2021) optimizes standard, regular-model scores that — as the singular examples of Drton et al. (2011) show — can prefer the wrong structure when the latent-confounded model is singular. We instead optimize a singular-learning-theory score: we verify Watanabe’s fundamental conditions for realizable linear Gaussian ADMG targets, derive prior-almost-everywhere AEC consistency of exact WBIC optimization under compact analytic priors, develop a differentiable ADVI approximation, and relax *both* the directed and the bidirected adjacency within a single continuous program, reporting the recovered Markov equivalence class as a PAG. Foundational treatments of causal models and inference underlie all of the above (Pearl, 2000; Peters et al., 2017; Bareinboim et al., 2022).

B SINGULARITY OF THE BOW GRAPH

This appendix proves the claims of Examples 1 and 3. Throughout, $\mathcal{G}$ is the bow ADMG of Fig. 1, with directed edges $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$ and bidirected edges $1 \leftrightarrow 2$, $1 \leftrightarrow 4$, $2 \leftrightarrow 4$, $3 \leftrightarrow 5$, so that $d = 13$, and $\mathcal{G}' = \mathcal{G} \setminus \{1 \leftrightarrow 2\}$. Prop. 5 shows that the parametrization $\phi_{\mathcal{G}}$ is singular at every parameter and that deleting the bow edge leaves its image unchanged; Prop. 6 then describes the common variety $V_{\mathcal{G}} = V_{\mathcal{G}'}$ near generic points by three explicit equality constraints. Both proofs rest on a single reparametrization of the error vector: because vertex 1 has no parents, any multiple of U_1 can be moved from the error of X_2 into the coefficient λ_{12} without changing the joint distribution of X .

Proposition 5. *For $\theta = (\Lambda, \mathcal{E}) \in \Theta_{\mathcal{G}}$ and $s \in \mathbb{R}$, let $\theta(s)$ agree with θ except for*

$$\begin{aligned} \lambda_{12}(s) &= \lambda_{12} - s, & \omega_{12}(s) &= \omega_{12} + s\omega_{11}, \\ \omega_{22}(s) &= \omega_{22} + 2s\omega_{12} + s^2\omega_{11}, & \omega_{24}(s) &= \omega_{24} + s\omega_{14}. \end{aligned}$$

Then the following hold.

- (i) $\theta(s) \in \Theta_{\mathcal{G}}$ and $\phi_{\mathcal{G}}(\theta(s)) = \phi_{\mathcal{G}}(\theta)$ for every $s \in \mathbb{R}$. Every fiber of $\phi_{\mathcal{G}}$ therefore contains a curve, and $\phi_{\mathcal{G}}$ is injective nowhere.
- (ii) The Jacobian of $\phi_{\mathcal{G}}$ has rank at most $12 < d$ at every parameter, so the Fisher information of the parametrization is singular everywhere on $\Theta_{\mathcal{G}}$.
- (iii) $\mathcal{C}_{\mathcal{G}} = \mathcal{C}_{\mathcal{G}'}$. Hence $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}')$, $V_{\mathcal{G}} = V_{\mathcal{G}'}$, and $\dim V_{\mathcal{G}} = 12$.

Proof. The proof proceeds in three steps, one for each claim. We first realize the curve $s \mapsto \theta(s)$ through a linear change of the error vector and verify that it stays in $\Theta_{\mathcal{G}}$ and preserves the covariance. We then differentiate this curve to obtain a kernel vector of the Jacobian, and finally move along the curve until the bow coefficient ω_{12} vanishes.

We begin with (i). Since vertex 1 has no parents, $X_1 = U_1$. For $s \in \mathbb{R}$, let

$$A_s = I + s e_2 e_1^\top, \quad U(s) = A_s U,$$

so that

$$U_2(s) = U_2 + sU_1, \quad U_j(s) = U_j \quad (j \neq 2).$$

Let $\Lambda(s)$ agree with Λ except for $\lambda_{12}(s) = \lambda_{12} - s$. The structural equation of X_2 is preserved, since

$$\lambda_{12}(s)X_1 + U_2(s) = (\lambda_{12} - s)X_1 + U_2 + sX_1 = \lambda_{12}X_1 + U_2 = X_2,$$

and no other structural equation involves λ_{12} or U_2 . The pair $(\Lambda(s), U(s))$ therefore generates the same random vector X , and in particular the same covariance Σ . The error covariance of $U(s)$ is

$$\mathcal{E}(s) = \text{Cov}(A_s U) = A_s \mathcal{E} A_s^\top.$$

Since $\det A_s = 1$, the matrix A_s is invertible and $\mathcal{E}(s) \succ 0$. Expanding the product, the entries of $\mathcal{E}(s)$ coincide with those of $\mathcal{E}$ except for

$$\omega_{12}(s) = \omega_{12} + s\omega_{11}, \quad \omega_{22}(s) = \omega_{22} + 2s\omega_{12} + s^2\omega_{11}, \quad \omega_{2j}(s) = \omega_{2j} + s\omega_{1j} \quad (j \in \{3, 4, 5\}).$$

For $j = 4$ this is the displayed formula for $\omega_{24}(s)$. For $j \in \{3, 5\}$, the graph $\mathcal{G}$ contains none of $2 \leftrightarrow 3$, $2 \leftrightarrow 5$, $1 \leftrightarrow 3$, $1 \leftrightarrow 5$, so

$$\omega_{23}(s) = \omega_{23} + s\omega_{13} = 0, \quad \omega_{25}(s) = \omega_{25} + s\omega_{15} = 0.$$

Hence $\mathcal{E}(s)$ has the zero pattern of $\mathcal{G}$, $\theta(s) = (\Lambda(s), \mathcal{E}(s)) \in \Theta_{\mathcal{G}}$, and

$$\phi_{\mathcal{G}}(\theta(s)) = \phi_{\mathcal{G}}(\theta) \quad \text{for every } s \in \mathbb{R}.$$

The curve $s \mapsto \theta(s)$ is injective because $\lambda_{12}(s) = \lambda_{12} - s$, so every fiber of $\phi_{\mathcal{G}}$ contains a curve and $\phi_{\mathcal{G}}$ is injective nowhere. This proves (i).

We next prove (ii). Differentiating the identity $\phi_{\mathcal{G}}(\theta(s)) = \phi_{\mathcal{G}}(\theta)$ at $s = 0$ gives

$$D\phi_{\mathcal{G}}(\theta) \dot{\theta}(0) = \frac{d}{ds} \phi_{\mathcal{G}}(\theta(s)) \Big|_{s=0} = 0,$$

where the tangent vector $\dot{\theta}(0)$ has coordinates

$$(\dot{\lambda}_{12}, \dot{\omega}_{12}, \dot{\omega}_{22}, \dot{\omega}_{24}) = (-1, \omega_{11}, 2\omega_{12}, \omega_{14})$$

and all other coordinates zero. Since $\dot{\lambda}_{12} = -1$, this vector is nonzero at every θ , so

$$\text{rank } D\phi_{\mathcal{G}}(\theta) \leq d - 1 = 12 \quad \text{for every } \theta \in \Theta_{\mathcal{G}}.$$

The Fisher information of the Gaussian model in the coordinates θ is the pullback $D\phi_{\mathcal{G}}(\theta)^\top I(\Sigma) D\phi_{\mathcal{G}}(\theta)$ of the Fisher information $I(\Sigma)$ of the covariance parametrization, so its kernel contains $\dot{\theta}(0)$ as well. Hence the Fisher information is singular everywhere on $\Theta_{\mathcal{G}}$, which proves (ii).

It remains to prove (iii). Since $\omega_{11} > 0$, we may take $s = -\omega_{12}/\omega_{11}$ in (i), which gives

$$\omega_{12}(s) = \omega_{12} - \frac{\omega_{12}}{\omega_{11}} \omega_{11} = 0.$$

The parameter $\theta(s)$ therefore belongs to $\Theta_{\mathcal{G}'}$ and has the same covariance as θ ; this is the map used in Example 1. Since $\theta \in \Theta_{\mathcal{G}}$ was arbitrary,

$$\mathcal{C}_{\mathcal{G}} \subseteq \mathcal{C}_{\mathcal{G}'}$$

Conversely, $\mathcal{G}' \subset \mathcal{G}$, so every parameter of $\mathcal{G}'$ is a parameter of $\mathcal{G}$ with $\omega_{12} = 0$, and $\mathcal{C}_{\mathcal{G}'} \subseteq \mathcal{C}_{\mathcal{G}}$. Hence $\mathcal{C}_{\mathcal{G}} = \mathcal{C}_{\mathcal{G}'}$. Equal images give equal models, $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}')$, and equal real Zariski closures, $V_{\mathcal{G}} = V_{\mathcal{G}'}$.

Finally, we compute the dimension. Since $\mathcal{G}'$ has 12 free parameters, $\dim \mathcal{C}_{\mathcal{G}'} \leq 12$. At the $\mathcal{G}'$ -representation of the parameter point of Example 1, namely

$$\lambda_{12} = \frac{7}{2}, \quad \omega_{22} = \frac{3}{2}, \quad \omega_{24} = \frac{1}{2},$$

with all other entries unchanged, the Jacobian of $\phi_{\mathcal{G}'}$ has full rank 12 (verified symbolically over $\mathbb{Q}$). Full rank persists in a neighborhood of this point, so by the constant-rank theorem $\phi_{\mathcal{G}'}$ maps a neighborhood onto a 12-dimensional submanifold of $\mathcal{C}_{\mathcal{G}'}$, and $\dim \mathcal{C}_{\mathcal{G}'} = 12$. Since semialgebraic sets and their real Zariski closures have the same dimension,

$$\dim V_{\mathcal{G}} = \dim V_{\mathcal{G}'} = \dim \mathcal{C}_{\mathcal{G}'} = 12.$$

This completes the proof. $\square$

Part (iii) also follows from van Ommen and Mooij (2017, Theorem 2), which shows that a generically infinite-to-one graph is algebraically equivalent to any generically finite-to-one graph obtained from it by deleting the fewest possible edges; the reparametrization above shows, in addition, that the two images coincide exactly.

Prop. 5 already shows that $\mathcal{G}$ and $\mathcal{G}'$ are algebraically equivalent. The next result makes their common variety explicit near generic points. For index sets A, B , write $\Sigma_{A,B}$ for the submatrix with rows A and columns B .

Proposition 6. *Define the polynomials*

$$\begin{aligned} f_1(\Sigma) &= \sigma_{13}\sigma_{22} - \sigma_{12}\sigma_{23}, \\ f_2(\Sigma) &= \sigma_{14}\sigma_{25} - \sigma_{15}\sigma_{24}, \\ f_3(\Sigma) &= \det \Sigma_{\{1,2,3,4\},\{5,2,3,4\}}, \end{aligned}$$

which, for positive-definite Σ , express $X_1 \perp X_3 \mid X_2$, the tetrad constraint $\text{rank } \Sigma_{\{1,2\},\{4,5\}} \leq 1$, and $X_1 \perp X_5 \mid \{X_2, X_3, X_4\}$, respectively, and let $Z = \{\Sigma : f_1(\Sigma) = f_2(\Sigma) = f_3(\Sigma) = 0\}$. Then $V_{\mathcal{G}} = V_{\mathcal{G}'} \subseteq Z$. Moreover, let θ' be a parameter of $\mathcal{G}'$ at which $\phi_{\mathcal{G}'}$ has full rank 12 and the gradients of f_1, f_2, f_3 are linearly independent at $\Sigma' = \phi_{\mathcal{G}'}(\theta')$. Then there is a neighborhood N of Σ' with $V_{\mathcal{G}} \cap N = Z \cap N$. Both conditions hold for generic θ' .

Proof. The proof proceeds in two steps. We first show that f_1, f_2, f_3 vanish on the covariance model, which gives the inclusion $V_{\mathcal{G}} \subseteq Z$. We then show that, near a generic point, Z is a 12-dimensional manifold that the image of $\phi_{\mathcal{G}'}$ fills out, which gives the reverse inclusion locally.

We begin with the inclusion. Substituting $\Sigma = \phi_{\mathcal{G}}(\theta)$ into f_1, f_2, f_3 gives the zero polynomial in θ (verified symbolically), and the vanishing also has a direct explanation. Since $X_3 = \lambda_{23}X_2 + U_3$ and U_3 is uncorrelated with $X_1 = U_1$ and with $X_2 = \lambda_{12}U_1 + U_2$, because $\mathcal{G}$ contains neither $1 \leftrightarrow 3$ nor $2 \leftrightarrow 3$,

$$\sigma_{j3} = \lambda_{23} \sigma_{j2} \quad (j \in \{1, 2\}).$$

Similarly, $X_5 = \lambda_{45}X_4 + U_5$ and U_5 is uncorrelated with X_1 and X_2 , because $\mathcal{G}$ contains neither $1 \leftrightarrow 5$ nor $2 \leftrightarrow 5$, so

$$\sigma_{j5} = \lambda_{45} \sigma_{j4} \quad (j \in \{1, 2\}).$$

Substituting these identities gives

$$f_1(\Sigma) = \lambda_{23}\sigma_{12}\sigma_{22} - \sigma_{12}\lambda_{23}\sigma_{22} = 0, \quad f_2(\Sigma) = \lambda_{45}\sigma_{14}\sigma_{24} - \lambda_{45}\sigma_{14}\sigma_{24} = 0.$$

For f_3 , let $B = \{2, 3, 4\}$. Expanding the determinant through the Schur complement of $\Sigma_{B,B} \succ 0$ gives

$$f_3(\Sigma) = \det \begin{pmatrix} \sigma_{15} & \Sigma_{1,B} \\ \Sigma_{B,5} & \Sigma_{B,B} \end{pmatrix} = \det \Sigma_{B,B} (\sigma_{15} - \Sigma_{1,B} \Sigma_{B,B}^{-1} \Sigma_{B,5}),$$

where the factor in parentheses is the partial covariance of X_1 and X_5 given X_B . It vanishes because 1 and 5 are m-separated given $\{2, 3, 4\}$ in both graphs. Hence each f_k vanishes on $\mathcal{C}_{\mathcal{G}} = \mathcal{C}_{\mathcal{G}'}$ (Prop. 5(iii)). Since Z is a real algebraic set containing $\mathcal{C}_{\mathcal{G}}$, it contains the Zariski closure, and

$$V_{\mathcal{G}} = V_{\mathcal{G}'} \subseteq Z.$$

We next prove the local equality. Fix θ' as in the statement and let $\Sigma' = \phi_{\mathcal{G}'}(\theta')$. The space of symmetric 5×5 matrices has dimension 15, and the gradients of f_1, f_2, f_3 are linearly independent at Σ' . By the implicit function theorem, for a small enough neighborhood N of Σ' , the set $Z \cap N$ is a connected real-analytic manifold of dimension

$$15 - 3 = 12.$$

Since $\phi_{\mathcal{G}'}$ has full rank 12 at θ' , it has rank 12 near θ' , and by the constant-rank theorem it maps a small enough neighborhood of θ' onto a 12-dimensional submanifold $M \subseteq Z \cap N$ through Σ' . A submanifold of full dimension is open, so M is an open subset of $Z \cap N$. Every polynomial vanishing on $V_{\mathcal{G}}$ vanishes on $M \subseteq \mathcal{C}_{\mathcal{G}'}$ and hence, by analytic continuation on the connected manifold $Z \cap N$, on all of $Z \cap N$. Since $V_{\mathcal{G}}$ is the common zero set of the polynomials vanishing on it,

$$Z \cap N \subseteq V_{\mathcal{G}},$$

and together with the first step, $V_{\mathcal{G}} \cap N = Z \cap N$.

Finally, we show that both conditions hold for generic θ' . Each condition fails exactly on the zero set of a polynomial in θ' : for the first, the sum of the squared 12×12 minors of the Jacobian of $\phi_{\mathcal{G}'}$; for the second, the sum of the squared 3×3 minors of the gradient matrix of (f_1, f_2, f_3) at $\phi_{\mathcal{G}'}(\theta')$. Neither polynomial vanishes identically, since both conditions hold at a randomly drawn rational parameter of $\mathcal{G}'$ (verified symbolically). The union of their zero sets is therefore a proper algebraic subset, so both conditions hold for generic θ' . This completes the proof. $\square$

Prop. 6 is local: it does not assert that f_1, f_2, f_3 generate the vanishing ideal of $V_{\mathcal{G}}$, or that Z has no other irreducible components. In contrast, Drton et al.'s graph of Appendix C violates f_2 and f_3 through its edge $1 \leftrightarrow 5$, and its variety is 13-dimensional.

C DRTON ET AL.'S EXAMPLE 1

Example 1 modifies the five-node ADMG of Drton et al. (2011, Example 1). This appendix studies the original graph, whose singularity, unlike that of the bow, occurs only at an exceptional parameter point. We first fix the graph and its parameter points and report the simulated scores (Example 6 and Table 3). Prop. 7 then shows that the singular distribution is shared with a sparser submodel, while Example 7 and the derivation that follows exhibit an explicit polynomial that separates the two varieties. Prop. 9 combines the two facts: at the singular point, the BIC selects the sparser graph, and hence the wrong AEC, with probability tending to one. The example thus complements the bow of Example 1, where the BIC's choice stays inside the generating graph's AEC (Example 4). We close by reading the scores at the regular and singular points against these computations.

Throughout this appendix, $\mathcal{G}$ denotes Drton et al.'s graph, with directed path $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$ and bidirected edges $1 \leftrightarrow 4$, $1 \leftrightarrow 5$, $2 \leftrightarrow 4$, $3 \leftrightarrow 5$ (Fig. 3), and $\mathcal{G}' = \mathcal{G} \setminus \{1 \leftrightarrow 5\}$ is the same bow-free submodel as in the main text. We say two ADMGs are *distributionally equivalent* at a distribution q when both their model classes contain q . Since $\mathcal{G}'$ is obtained from $\mathcal{G}$ by setting a single free parameter to zero, $\mathcal{M}(\mathcal{G}') \subseteq \mathcal{M}(\mathcal{G})$ as model classes; the substantive claim of Prop. 7 below is therefore that the *particular* singular distribution generated by $\mathcal{G}$ also lies in the smaller class $\mathcal{M}(\mathcal{G}')$. Throughout we use the model-implied covariance $\Sigma(\theta) = (I - \Lambda)^{-1} \mathcal{E} (I - \Lambda)^{-\top}$ with $\theta = (\Lambda, \mathcal{E})$ on the graph-restricted support, as fixed in the Preliminaries.

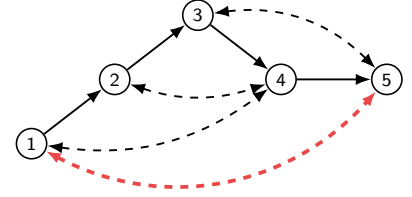


Figure 3: The ADMG $\mathcal{G}$ of Drton et al.'s Example 1 (Drton et al., 2011); solid edges are directed, dashed edges bidirected. At the parameter point of Example 6 the induced Gaussian model is singular, and the highlighted edge $1 \leftrightarrow 5$ (red) can be removed without changing the distribution, yielding $\mathcal{G}'$.

Example 6 (Drton et al.'s Example 1). Take $\lambda_{12} = 3$, $\lambda_{23} = -\frac{1}{2}$, $\lambda_{34} = \lambda_{45} = 1$, and let $\mathcal{E}$ have diagonal entries 2 and $\omega_{14} = \omega_{15} = \omega_{24} = \omega_{35} = 1$. Here $\mathcal{G}$ is generically but not globally identifiable: $\phi_{\mathcal{G}}$ fails to be injective at this point and its image is singular there, so a positive-dimensional set of parameters yields the same Σ and the model has fewer than $d = 13$ effective degrees of freedom, precisely the regime in which the BIC misjudges the effective dimension (Drton et al., 2011). Flipping the sign to $\lambda_{23} = +\frac{1}{2}$ removes the singularity: Σ then determines $(\Lambda, \mathcal{E})$ uniquely, and the model is regular.

We simulate $n = 1000$ samples and compare $\mathcal{G}$ with $\mathcal{G}'$, whose deleted edge $1 \leftrightarrow 5$ carries the coefficient that is unidentified along the singular direction (Table 3, $R = 30$ replicates, lower is better). At the regular point, the edge is identified; its deletion degrades the fit, and the BIC clearly prefers $\mathcal{G}$ (8014.8 versus 8279.0). At the singular point the edge is redundant, since $\mathcal{G}$ and $\mathcal{G}'$ induce the same distribution (Prop. 7); yet the BIC, charging the full nominal dimension through $\frac{d}{2} \log n$, picks $\mathcal{G}'$ (8011.8 versus 8014.5), a graph outside the generating graph's AEC (Prop. 9). The WBIC, estimated as in Example 2 but with chains of $T = 15,000$ iterations ($T_0 = 3,000$ warm-up; Appendix E), agrees with the BIC in the regular regime (8017.6 versus 8282.2) but leaves the two essentially tied at the singular point, with $\mathcal{G}$ marginally ahead (8012.9 versus 8013.9). The ADVI surrogate behaves likewise: it clearly favors $\mathcal{G}$ in the regular regime, and at the singular point its gap shrinks to 0.7 (7981.9 versus 7981.2), two orders of magnitude below the regular margin of 266.6. $\blacksquare$

Proposition 7. *At the parameter point of Example 6 ($\lambda_{12} = 3$, $\lambda_{23} = -\frac{1}{2}$, $\lambda_{34} = \lambda_{45} = 1$, $\mathcal{E}$ with diagonal 2 and $\omega_{14} = \omega_{15} = \omega_{24} = \omega_{35} = 1$), the centered Gaussian distribution $\mathcal{N}(0, \Sigma)$ generated by $\mathcal{G}$ is also generated*

Table 3: Selection scores (mean $\pm$ s.d. over $R = 30$ replications, $n = 1000$) for Drton et al.’s graph $\mathcal{G}$ and its sparser submodel $\mathcal{G}'$; lower is better and **bold** marks the regular-regime winner. Every criterion favors $\mathcal{G}$ by ≈ 264 – 267 at the regular point; at the singular point the BIC prefers $\mathcal{G}'$, which lies outside the generating graph’s AEC, by 2.7, whereas the WBIC and ADVI keep the two within 1.

Criterion	Regular ($\lambda_{23} = +1/2$)		Singular ($\lambda_{23} = -1/2$)	
	$\mathcal{G}$ (Fig. 3)	$\mathcal{G}'$ (Fig. 3 w/o $1 \leftrightarrow 5$)	$\mathcal{G}$ (Fig. 3)	$\mathcal{G}'$ (Fig. 3 w/o $1 \leftrightarrow 5$)
BIC	8014.82 $\pm$ 55.77	8279.01 $\pm$ 60.41	8014.46 $\pm$ 55.79	8011.79 $\pm$ 55.57
WBIC	8017.56 $\pm$ 55.28	8282.18 $\pm$ 61.42	8012.94 $\pm$ 55.94	8013.93 $\pm$ 57.28
ADVI	7981.82 $\pm$ 55.67	8248.41 $\pm$ 60.54	7981.88 $\pm$ 55.53	7981.19 $\pm$ 55.68

by $\mathcal{G}'$. Consequently the data-generating distribution q satisfies $q \in \mathcal{M}(\mathcal{G}) \cap \mathcal{M}(\mathcal{G}')$, and the two graphs are distributionally equivalent at q .

Proof. The proof proceeds in three steps. We first evaluate the covariance generated by $\mathcal{G}$ at the stated point, then exhibit a parameter of $\mathcal{G}'$, and finally check that the two parameters generate the same covariance.

Evaluating $\Sigma(\theta) = (I - \Lambda)^{-1} \mathcal{E} (I - \Lambda)^{-\top}$ at the stated point of $\mathcal{G}$ gives the exact covariance

$$\Sigma = \begin{pmatrix} 2 & 6 & -3 & -2 & -1 \\ 6 & 20 & -10 & -6 & -3 \\ -3 & -10 & 7 & 5 & \frac{9}{2} \\ -2 & -6 & 5 & 5 & \frac{9}{2} \\ -1 & -3 & \frac{9}{2} & \frac{9}{2} & 6 \end{pmatrix}. \quad (12)$$

The error covariance $\mathcal{E}$ of $\mathcal{G}$ is positive definite, since its eigenvalues are $2 - \sqrt{3} \approx 0.268$, 1, 2, 3, and $2 + \sqrt{3} \approx 3.732$. Hence the stated point is a valid parameter of $\mathcal{G}$.

We next consider $\mathcal{G}'$, where $\omega_{15} = 0$, at the parameter point θ' given by

$$\lambda_{12} = 3, \quad \lambda_{23} = -\frac{1}{2}, \quad \lambda_{34} = 1, \quad \lambda_{45} = \frac{1}{2},$$

with $\mathcal{E}'$ having diagonal $(\omega_{11}, \dots, \omega_{55}) = (2, 2, 2, 2, \frac{11}{4})$ and off-diagonal entries

$$\omega_{14} = \omega_{24} = 1, \quad \omega_{35} = 2,$$

all other entries, including ω_{15} , being zero. This $\mathcal{E}'$ is positive definite: its eigenvalues are 2, $2 \pm \sqrt{2}$, and $\frac{19 \pm \sqrt{265}}{8}$, the smallest being $\frac{19 - \sqrt{265}}{8} \approx 0.340 > 0$. Hence θ' is a valid parameter of $\mathcal{G}'$.

Finally, substituting θ' into $\Sigma(\theta') = (I - \Lambda')^{-1} \mathcal{E}' (I - \Lambda')^{-\top}$ yields, as an exact algebraic identity (verified symbolically over $\mathbb{Q}$),

$$\Sigma(\theta') = \Sigma,$$

the matrix of Eq. (12). Both graphs therefore induce the same centered Gaussian $\mathbb{N}(0, \Sigma)$, and $q = \mathbb{N}(0, \Sigma)$ lies in $\mathcal{M}(\mathcal{G}) \cap \mathcal{M}(\mathcal{G}')$. This is the asserted distributional equivalence. $\square$

This calculation concerns a single singular distribution shared by two graph models, not algebraic equivalence of their full covariance models. In fact, $\mathcal{G}$ has generic image dimension 13, whereas $\mathcal{G}'$ has only 12 free parameters, so their varieties differ. This target belongs to the measure-zero set excluded by Thm. 4. The scores in Table 3 illustrate behavior there, but neither their near tie nor their ranking establishes asymptotic identification of the generating graph’s AEC at this distribution.

Example 7 (Example 6, continued: Zariski closure and AEC). Recall $\mathcal{G}' = \mathcal{G} \setminus \{1 \leftrightarrow 5\}$ and write $\Sigma = (\sigma_{ij})$. The closure $V_{\mathcal{G}}$ comprises the symmetric matrices satisfying every polynomial equality valid throughout $\mathcal{C}_{\mathcal{G}}$, for instance $\sigma_{13}\sigma_{22} - \sigma_{12}\sigma_{23} = 0$, which expresses $X_1 \perp X_3 \mid X_2$ for positive-definite Gaussian covariances. The closure can leave the positive-definite cone: scaling $\mathcal{E}$ by $t \downarrow 0$ shows $0 \in V_{\mathcal{G}}$, although $0 \notin \mathcal{C}_{\mathcal{G}}$. Deleting $1 \leftrightarrow 5$ adds the tetrad constraint (Sullivant et al., 2010)

$$T(\Sigma) := \sigma_{14}\sigma_{25} - \sigma_{15}\sigma_{24} = 0, \quad (13)$$

whereas in $\mathcal{G}$ the parametrization gives $T(\Sigma) = -\omega_{15}(\lambda_{23}\lambda_{34}\omega_{22} + \omega_{24})$, which equals 0 at the singular point and -2 at the regular one. Since $\mathcal{C}_{\mathcal{G}'} \subseteq \mathcal{C}_{\mathcal{G}}$ and T vanishes throughout $V_{\mathcal{G}'}$ but not throughout $V_{\mathcal{G}}$, we obtain $V_{\mathcal{G}'} \subsetneq V_{\mathcal{G}}$ and $\mathcal{G}' \notin \text{AEC}(\mathcal{G})$. Within $\mathbb{G} = \{\mathcal{G}, \mathcal{G}'\}$ the AECs are therefore the singletons $\{\mathcal{G}\}$ and $\{\mathcal{G}'\}$, although the two graphs share the singular distribution. Both identities are derived below. ■

We now derive the two identities used in Example 7. Along the directed path, the structural equations read

$$X_1 = U_1, \quad X_2 = \lambda_{12}X_1 + U_2, \quad X_3 = \lambda_{23}X_2 + U_3, \quad X_4 = \lambda_{34}X_3 + U_4, \quad X_5 = \lambda_{45}X_4 + U_5.$$

We begin with the conditional independence. Since U_3 is uncorrelated with $X_1 = U_1$ and with $X_2 = \lambda_{12}U_1 + U_2$, we have

$$\sigma_{13} = \lambda_{23}\sigma_{12}, \quad \sigma_{23} = \lambda_{23}\sigma_{22},$$

and therefore

$$\sigma_{13}\sigma_{22} - \sigma_{12}\sigma_{23} = \lambda_{23}\sigma_{12}\sigma_{22} - \lambda_{23}\sigma_{12}\sigma_{22} = 0$$

in both graphs.

We next compute the tetrad. Besides the variances, the only nonzero error covariances of $\mathcal{G}$ are ω_{14} , ω_{15} , ω_{24} , and ω_{35} . Expanding $X_4 = \lambda_{23}\lambda_{34}(\lambda_{12}U_1 + U_2) + \lambda_{34}U_3 + U_4$ and $X_5 = \lambda_{45}X_4 + U_5$ gives

$$\begin{aligned} \sigma_{14} &= \lambda_{12}\lambda_{23}\lambda_{34}\omega_{11} + \omega_{14}, \\ \sigma_{24} &= \lambda_{23}\lambda_{34}(\lambda_{12}^2\omega_{11} + \omega_{22}) + \lambda_{12}\omega_{14} + \omega_{24}, \\ \sigma_{15} &= \lambda_{45}\sigma_{14} + \omega_{15}, \\ \sigma_{25} &= \lambda_{45}\sigma_{24} + \lambda_{12}\omega_{15}. \end{aligned}$$

Substituting the last two identities into $T(\Sigma)$, the terms proportional to λ_{45} cancel:

$$T(\Sigma) = \sigma_{14}(\lambda_{45}\sigma_{24} + \lambda_{12}\omega_{15}) - (\lambda_{45}\sigma_{14} + \omega_{15})\sigma_{24} = \omega_{15}(\lambda_{12}\sigma_{14} - \sigma_{24}).$$

Substituting the first two identities, the terms in ω_{11} and ω_{14} cancel as well, so

$$\lambda_{12}\sigma_{14} - \sigma_{24} = -(\lambda_{23}\lambda_{34}\omega_{22} + \omega_{24}) \quad \text{and} \quad T(\Sigma) = -\omega_{15}(\lambda_{23}\lambda_{34}\omega_{22} + \omega_{24}).$$

Setting $\omega_{15} = 0$ proves Eq. (13) throughout $\mathcal{C}_{\mathcal{G}'}$, and hence throughout its Zariski closure. For the two parameter choices in Example 6, $\lambda_{34} = \omega_{15} = \omega_{24} = 1$ and $\omega_{22} = 2$, so

$$T(\Sigma) = -(2\lambda_{23} + 1),$$

which is 0 at $\lambda_{23} = -\frac{1}{2}$ and -2 at $\lambda_{23} = \frac{1}{2}$. This supplies an explicit polynomial witness for $V_{\mathcal{G}'} \subsetneq V_{\mathcal{G}}$, independently of the dimension comparison above. It also shows why a shared singular distribution need not imply algebraic equivalence of the graphs.

Remark 8 (The underlying singularity). The equivalence reflects a genuine singularity of $\phi_{\mathcal{G}}$. Linearizing the map from the $d = 13$ free parameters of $\mathcal{G}$ to the 15 distinct entries of Σ at the point of Example 6, the Jacobian has rank 12, leaving a one-dimensional kernel spanned by

$$(\partial\lambda_{45}, \partial\omega_{55}, \partial\omega_{15}, \partial\omega_{35}) \propto (-\tfrac{1}{2}, -\tfrac{1}{2}, -1, +1),$$

with all other coordinates zero. The dominant component is $\partial\omega_{15}$, so ω_{15} is the coefficient that is unidentified along the flat direction, and the effective dimension is $12 < 13 = d$. The submodel $\mathcal{G}'$ lies on the constant- Σ fibre through this point, reached by driving ω_{15} to 0 (with the nonlinear fibre adjusting λ_{45} , ω_{55} , ω_{35} to the values above). This deficient effective dimension is exactly what defeats the nominal-dimension penalty of the BIC in Eq. (1) and what the RLCT-based WBIC of Prop. 1 corrects.

The two facts established so far, that $\mathcal{G}'$ realizes the singular distribution and that it lies outside $\text{AEC}(\mathcal{G})$, combine into a failure of the BIC to identify the generating graph's AEC. Write $\ell_n(\Sigma') = \sum_{i=1}^n \log \mathbb{N}(\mathbf{v}_i; 0, \Sigma')$ for the Gaussian log-likelihood, and for an ADMG $\mathcal{H}$ with $k_{\mathcal{H}}$ free parameters let

$$\hat{\ell}_{\mathcal{H}} = \sup_{\theta \in \Theta_{\mathcal{H}}} \ell_n(\phi_{\mathcal{H}}(\theta)), \quad \text{BIC}_n(\mathcal{H}) = -\hat{\ell}_{\mathcal{H}} + \frac{k_{\mathcal{H}}}{2} \log n,$$

which are the leading terms of Eq. (1); here $k_{\mathcal{G}} = 13$ and $k_{\mathcal{G}'} = 12$.

Proposition 9. *Let $q = \mathbb{N}(0, \Sigma)$ with Σ the covariance of Eq. (12), and let the observations be i.i.d. from q . Then*

$$\text{BIC}_n(\mathcal{G}) - \text{BIC}_n(\mathcal{G}') = \frac{1}{2} \log n - O_p(1) \xrightarrow{P_q} \infty,$$

so that $P_q^{\otimes n}\{\text{BIC}_n(\mathcal{G}') < \text{BIC}_n(\mathcal{G})\} \rightarrow 1$. Since $\mathcal{G}' \notin \text{AEC}(\mathcal{G})$, the BIC selects a graph outside the generating graph's AEC with probability tending to one.

Proof. The proof compares the maximized log-likelihoods of both graphs with that of the saturated Gaussian model. We first sandwich their difference between zero and a saturated likelihood-ratio statistic, and then show that this statistic is bounded in probability while the penalty gap grows like $\frac{1}{2} \log n$.

Since $\mathcal{G}' \subset \mathcal{G}$, we have $\mathcal{C}_{\mathcal{G}'} \subseteq \mathcal{C}_{\mathcal{G}}$, and therefore $\hat{\ell}_{\mathcal{G}'} \leq \hat{\ell}_{\mathcal{G}}$. Since $\Sigma \in \mathcal{C}_{\mathcal{G}'}$ by Prop. 7, we also have $\ell_n(\Sigma) \leq \hat{\ell}_{\mathcal{G}'}$. Finally, every covariance in $\mathcal{C}_{\mathcal{G}}$ is positive definite, so $\hat{\ell}_{\mathcal{G}} \leq \hat{\ell}_{\text{sat}} := \sup_{\Sigma' \succ 0} \ell_n(\Sigma')$. Combining the three inequalities,

$$0 \leq \hat{\ell}_{\mathcal{G}} - \hat{\ell}_{\mathcal{G}'} \leq \hat{\ell}_{\text{sat}} - \ell_n(\Sigma).$$

With probability one, $S_n = n^{-1} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top$ is positive definite for $n \geq p$ and maximizes the saturated likelihood, so

$$\hat{\ell}_{\text{sat}} - \ell_n(\Sigma) = \frac{n}{2} \left[\text{tr}(\Sigma^{-1} S_n) - \log \det(\Sigma^{-1} S_n) - p \right].$$

The saturated model is a regular exponential family with Σ in the interior of its parameter space, so by Wilks' theorem $2(\hat{\ell}_{\text{sat}} - \ell_n(\Sigma))$ converges in law to a χ^2 distribution with 15 degrees of freedom, one for each distinct entry of Σ . In particular,

$$\hat{\ell}_{\mathcal{G}} - \hat{\ell}_{\mathcal{G}'} = O_p(1).$$

Therefore

$$\text{BIC}_n(\mathcal{G}) - \text{BIC}_n(\mathcal{G}') = \frac{k_{\mathcal{G}} - k_{\mathcal{G}'}}{2} \log n - (\hat{\ell}_{\mathcal{G}} - \hat{\ell}_{\mathcal{G}'}) = \frac{1}{2} \log n - O_p(1) \xrightarrow{P_q} \infty,$$

and $P_q^{\otimes n}\{\text{BIC}_n(\mathcal{G}') < \text{BIC}_n(\mathcal{G})\} \rightarrow 1$. Finally, $\text{AEC}(\mathcal{G}) = \{\mathcal{G}\}$ within $\mathbb{G} = \{\mathcal{G}, \mathcal{G}'\}$ by Example 7, so $\mathcal{G}' \notin \text{AEC}(\mathcal{G})$. This completes the proof. $\square$

The BIC's preference for $\mathcal{G}'$ is not a finite-sample artifact: its penalty gap grows like $\frac{1}{2} \log n$, whereas the fit gap stays bounded, so the preference strengthens with n . Its error is one of AEC identification, in contrast with the bow of Example 1, where the BIC's choice remains inside the generating graph's AEC (Example 4). Since q also lies in $\mathcal{M}(\mathcal{G}')$, no procedure can tell $\mathcal{G}$ from $\mathcal{G}'$ at q alone. The BIC nevertheless commits to the class of the sparser graph by charging nominal dimension, whereas the WBIC, whose penalty is the RLCT, keeps the two graphs essentially tied.

We close by reading Table 3 against these computations. At the regular point, $T = -2 \neq 0$ places the target covariance outside $V_{\mathcal{G}'}$, so $\mathcal{G}'$ cannot realize q . It trails by an order- n KL gap, matching the large WBIC margin in Table 3, and the WBIC's preference for $\mathcal{G}$ selects the generating graph's singleton AEC, as Thm. 4 predicts.

At the singular point, by contrast, $\mathcal{G}'$ also realizes the target (Prop. 7), so the target covariance lies in $V_{\mathcal{G}'}$, a proper algebraic subset of $V_{\mathcal{G}}$ (consistently, $T = 0$), and belongs to the measure-zero set excluded by Lem. 3 and Thm. 4. Nor is it a generic point of $\mathcal{G}'$: rank computations place it where $V_{\mathcal{G}'}$ meets the image of the rank-deficient stratum $\{\lambda_{23}\lambda_{34}\omega_{22} + \omega_{24} = 0\}$ of $\mathcal{G}$, an 11-dimensional set, so it satisfies an equality constraint beyond those of $\mathcal{G}'$. Thm. 4, applied to either graph, therefore makes no claim here. The same computations suggest that both graphs realize q with RLCT 6: $\mathcal{G}'$ is regular at its representation of q , and the parameters of $\mathcal{G}$ realizing q form a smooth curve (Remark 8) on which the Jacobian of $\phi_{\mathcal{G}}$ has rank 12 at the points examined. The two graphs thus lie in different AECs, and $\mathcal{G}$ generates distributions with $T \neq 0$ arbitrarily close to q ; yet at q itself their fit terms coincide and, by these computations, so do their RLCTs, so they would agree at the WBIC's leading order. The near ties of the WBIC and ADVI in Table 3 are what a local analogue of AEC identification would predict: both graphs that are indistinguishable at this order are kept, whereas the BIC, charging nominal dimension, selects $\mathcal{G}'$ and hence the wrong AEC (Prop. 9). The example is an instance suggesting that the WBIC's behavior may extend beyond the generic targets covered by Thm. 4.

D PROOFS

This appendix collects the proofs of the theoretical results in the main text. Appendix D.1 fixes the standing setup shared by all results. Appendix D.2 verifies Watanabe’s fundamental conditions for realizable Gaussian targets (Lem. 10) and deduces the WBIC expansion of Prop. 1. Appendix D.3 proves the RLCT gap (Lem. 3) and AEC consistency (Thm. 4), and Appendix D.4 proves the prior-misspecification bound (Prop. 15). The consistency proof follows the standard singular-learning playbook: a generic-geometry argument removes a prior-null set of generating parameters, a local splitting of the Kullback–Leibler divergence computes the RLCT of every graph that realizes the target, and the WBIC expansion converts the resulting strict RLCT gap into a gap between scores.

D.1 Parameter supports and priors

The following conditions are part of the standing setup for all results in the main text and their proofs, and they underlie the setup preceding Prop. 1. Bows and nonidentifiable parametrizations are allowed. For each $\mathcal{G} \in \mathbb{G}$, let $\Theta_{\mathcal{G}} \subset \mathbb{R}^{k_{\mathcal{G}}}$ be its open space of free structural parameters, with $\mathcal{E} \succ 0$, and fix a compact, full-dimensional semialgebraic support $W_{\mathcal{G}} \Subset \Theta_{\mathcal{G}}$ satisfying $W_{\mathcal{G}} = \overline{\text{int } W_{\mathcal{G}}}$. Fix a prior $\Pi_{\mathcal{G}}$ supported on $W_{\mathcal{G}}$, with bounded, strictly positive density $\varphi_{\mathcal{G}}$ analytic on a neighborhood of $W_{\mathcal{G}}$ and normalized so that $\int_{W_{\mathcal{G}}} \varphi_{\mathcal{G}}(\theta) d\theta = 1$. These choices remain fixed throughout the results and their proofs. We suppress graph subscripts when discussing a single graph, writing W , Θ , Π , and φ .

Throughout, $\theta = (\Lambda, \mathcal{E})$ collects the $k_{\mathcal{G}}$ free parameters of an ADMG $\mathcal{G}$, and

$$\phi_{\mathcal{G}}(\theta) = (I - \Lambda)^{-1} \mathcal{E} (I - \Lambda)^{-\top}$$

is its covariance map. The sets $\mathcal{C}_{\mathcal{G}}$ and $V_{\mathcal{G}}$ are as in Def. 2, and we write $q = \mathcal{N}_p(0, S)$ for the target distribution. Covariance matrices are vectorized when taking Jacobians, and all fiber dimensions below are real local or semialgebraic dimensions.

D.2 Fundamental conditions and Gaussian admissibility

For a realizable truth q , the WBIC expansion in Prop. 1 follows from Watanabe’s theorem whenever the triple $(q, p(\cdot | \theta), \varphi)$ of true density, model, and prior satisfies Watanabe’s fundamental conditions (Watanabe, 2013). These are mild regularity requirements that make the free-energy geometry well defined: analyticity so the singularities can be resolved, compactness and a smooth prior so the defining integral is controlled, and a finite-variance bound tying the Kullback–Leibler divergence to the fluctuations of the log-likelihood. Writing $f(x, \theta) = \log\{q(x)/p(x | \theta)\}$ for the log-density ratio, $K(\theta) = \mathbb{E}_X[f(X, \theta)]$ for the Kullback–Leibler divergence (the expectation $\mathbb{E}_X$ taken under the true density q , i.e. $X \sim q$), $L^s(q)$ for the space of functions s -integrable under q , and W for the parameter set, they read:

Condition 1 (Compactness). $W \subset \mathbb{R}^d$ is compact with non-empty interior and has a boundary described by finitely many analytic equalities and inequalities.

Condition 2 (Prior factorization). $\varphi(\theta) = \varphi_1(\theta) \varphi_2(\theta)$ with $\varphi_1 \geq 0$ analytic and $\varphi_2 > 0$ of class C^∞ (i.e. smooth, with derivatives of all orders).

Condition 3 (Analytic integrability). There is an open $W' \supset W$ on which $\theta \mapsto f(x, \theta)$ is an $L^s(q)$ -valued analytic function for some $s \geq 6$.

Condition 4 (Relatively finite variance). There exist $\epsilon, c > 0$ such that $K(\theta) \geq c \mathbb{E}_X[f(X, \theta)^2]$ for all $\theta \in W_\epsilon = \{\theta \in W : K(\theta) \leq \epsilon\}$.

Lemma 10. *Under the standing setup, if the centered Gaussian truth q is realized by a parameter in $W_{\mathcal{G}}$, then Conditions 1–4 hold.*

The proof verifies the four conditions in turn. Conditions 1 and 2 are immediate from the standing setup. Condition 3 holds because the log-density ratio is a quadratic polynomial in x whose coefficients are analytic in θ . Condition 4 is proved in covariance coordinates, which avoids expanding the divergence in potentially singular structural parameters.

Proof of Lem. 10. Fix a graph and suppress its subscript. Since $W \in \Theta$ is compact, its error covariances have eigenvalues in some fixed interval $[c_0, C_0] \subset (0, \infty)$, and its structural coefficients are bounded. Assume $S = \phi(\theta_0)$ for some $\theta_0 \in W$; this representation need not be interior or identifiable.

We begin with Condition 1. The support W is compact with nonempty interior by hypothesis, and its semialgebraic description uses finitely many polynomial equalities and inequalities, which are analytic. This establishes the required boundary regularity. Condition 2 holds by taking $\varphi_1 = \varphi$ and $\varphi_2 = 1$ on a neighborhood of W , where the density is analytic and positive.

We next verify Condition 3. Acyclicity implies $\Lambda^p = 0$ after topological ordering, so

$$(I - \Lambda)^{-1} = \sum_{j=0}^{p-1} \Lambda^j,$$

and ϕ is a polynomial map. On Θ , $\phi(\theta) \succ 0$, and matrix inversion and the log-determinant are analytic on the positive-definite cone. Hence

$$f(x, \theta) = \log q(x) + \frac{1}{2} \left[p \log(2\pi) + \log \det \phi(\theta) + x^\top \phi(\theta)^{-1} x \right]$$

is a quadratic polynomial in x with coefficients analytic in θ . To see that it is an $L^s(q)$ -valued analytic function, fix $\theta_1 \in \Theta$ and expand the coefficients in convergent power series,

$$\frac{1}{2} \log \det \phi(\theta) = \sum_{\alpha} a_{\alpha} (\theta - \theta_1)^{\alpha}, \quad \frac{1}{2} \phi(\theta)^{-1} = \sum_{\alpha} B_{\alpha} (\theta - \theta_1)^{\alpha},$$

with $\sum_{\alpha} (|a_{\alpha}| + \|B_{\alpha}\|) r^{|\alpha|} < \infty$ for some $r > 0$. For $\|\theta - \theta_1\|_{\infty} < r$, this gives

$$f(x, \theta) = \log q(x) + \frac{p}{2} \log(2\pi) + \sum_{\alpha} (a_{\alpha} + x^\top B_{\alpha} x) (\theta - \theta_1)^{\alpha},$$

and each coefficient satisfies

$$\|a_{\alpha} + x^\top B_{\alpha} x\|_{L^s(q)} \leq |a_{\alpha}| + \|B_{\alpha}\| \|x\|^2 \|1\|_{L^s(q)}.$$

Gaussian moments of every order are finite, so $\log q$ and $\|x\|^2$ lie in $L^s(q)$ and the series converges absolutely in $L^s(q)$ for every finite s , including $s \geq 6$. Taking $W' = \Theta$ proves Condition 3.

It remains to verify Condition 4. We work in covariance coordinates rather than expanding in potentially singular structural parameters. Set

$$A(\theta) = S^{1/2} \phi(\theta)^{-1} S^{1/2}, \quad X = S^{1/2} Z,$$

where $Z \sim \mathcal{N}_p(0, I)$. Since $\log \det A = \log \det S - \log \det \phi(\theta)$, $X^\top \phi(\theta)^{-1} X = Z^\top A Z$, and $X^\top S^{-1} X = Z^\top Z$, the log-density ratio becomes

$$f(X, \theta) = \frac{1}{2} \left[\log \det \phi(\theta) - \log \det S + X^\top (\phi(\theta)^{-1} - S^{-1}) X \right] = \frac{1}{2} \left[Z^\top (A - I) Z - \log \det A \right].$$

Using $\mathbb{E}[Z^\top M Z] = \text{tr } M$ and $\text{Var}(Z^\top M Z) = 2 \text{tr } M^2$ for symmetric M , the Gaussian formulas give

$$K(\theta) = \frac{1}{2} \{ \text{tr } A - p - \log \det A \} = \frac{1}{2} \sum_{i=1}^p g(a_i),$$

$$\text{Var}_q f(X, \theta) = \frac{1}{2} \text{tr}(A - I)^2 = \frac{1}{2} \sum_{i=1}^p (a_i - 1)^2,$$

where $a_1, \dots, a_p > 0$ are the eigenvalues of $A = A(\theta)$ and $g(a) = a - 1 - \log a \geq 0$. Near $a = 1$, the function g has a positive quadratic term: since $g(1) = g'(1) = 0$ and $g''(a) = a^{-2}$, Taylor's theorem gives, for some ξ between 1 and a ,

$$g(a) = \frac{(a-1)^2}{2\xi^2} \geq \frac{2}{9}(a-1)^2 \quad \text{for } a \in (0, \frac{3}{2}].$$

Sufficiently small divergence forces A into this range. Indeed, g is increasing on $[1, \infty)$ and $K(\theta) \geq \frac{1}{2}g(a_i)$ for each i , so if $K(\theta) \leq \epsilon < \frac{1}{2}g(\frac{3}{2})$, then $a_i < \frac{3}{2}$ for every i . On W_ϵ we therefore obtain

$$K(\theta) \geq \frac{1}{2} \sum_{i=1}^p \frac{2}{9}(a_i - 1)^2 = \frac{2}{9} \text{Var}_q f(X, \theta),$$

and consequently

$$\mathbb{E}_q[f(X, \theta)^2] = \text{Var}_q f(X, \theta) + K(\theta)^2 \leq \left(\frac{9}{2} + \epsilon\right)K(\theta).$$

This argument is uniform over all representations of S and does not require a nonsingular structural Fisher information. Taking $c = (\frac{9}{2} + \epsilon)^{-1}$ proves Condition 4. This completes the proof. $\square$

With Lem. 10 in hand, Prop. 1 is a direct consequence of Watanabe's main theorem.

Proof of Prop. 1. For a centered Gaussian model, the log-density of a single observation is

$$\log p(\mathbf{v} \mid \theta, \mathcal{M}(\mathcal{G})) = -\frac{1}{2} \left[p \log 2\pi + \log \det \Sigma(\theta) + \mathbf{v}^\top \Sigma(\theta)^{-1} \mathbf{v} \right].$$

Summing over the sample and using

$$\sum_{i=1}^n \mathbf{v}_i^\top \Sigma(\theta)^{-1} \mathbf{v}_i = n \text{tr}(\Sigma(\theta)^{-1} S_n), \quad S_n = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i^\top,$$

gives Eq. (4). Since $\mathcal{G}$ realizes q within $W_{\mathcal{G}}$, Lem. 10 shows that the fundamental conditions hold. With $\beta = 1/\log n$, Watanabe (2013, Theorem 4) then gives

$$\text{WBIC}(\mathcal{G}) = nL_n(\theta_{0,\mathcal{G}}) + \lambda_{\mathcal{G}} \log n + U_n \sqrt{\lambda_{\mathcal{G}} \log n / 2} + O_p(1),$$

where U_n converges in law to a Gaussian random variable. In particular $U_n = O_p(1)$, so the last two terms are $O_p(\sqrt{\log n})$, which is Eq. (5). The fit term does not depend on the choice of $\theta_{0,\mathcal{G}}$, because every parameter realizing q induces the same density. This completes the proof. $\square$

D.3 Proof of Lem. 3 and Thm. 4

Fix the generating graph $\mathcal{G}^*$ and write

$$Y = V_{\mathcal{G}^*}, \quad d = \dim Y.$$

The proof proceeds in four steps. We first show that the varieties $V_{\mathcal{G}}$ are irreducible and use this to construct a proper algebraic subset $E \subsetneq Y$ outside which the comparison geometry of every candidate graph is generic (Lems. 11 and 12); the generating parameters mapped into E have prior measure zero (Lem. 13). We then compute local RLCTs through an analytic splitting of the Kullback–Leibler divergence (Lem. 14), which yields the RLCT gap of Lem. 3. Finally, we convert this gap into a separation of WBIC scores for competitors that realize q , treat competitors that cannot realize q within their prior support through an order- n fit gap, and conclude by a union bound over the finite family $\mathbb{G}$.

Lemma 11. *For every $\mathcal{G} \in \mathbb{G}$, the variety $V_{\mathcal{G}}$ is irreducible.*

Proof. It suffices to show that the vanishing ideal $I(V_{\mathcal{G}})$ of $V_{\mathcal{G}}$ is prime. By the proof of Lem. 10, $\phi_{\mathcal{G}}$ is a polynomial map. Let g, h be polynomials with $gh \in I(V_{\mathcal{G}})$. Then gh vanishes on $\mathcal{C}_{\mathcal{G}} = \phi_{\mathcal{G}}(\Theta_{\mathcal{G}})$, so

$$(g \circ \phi_{\mathcal{G}})(h \circ \phi_{\mathcal{G}}) = 0 \quad \text{on } \Theta_{\mathcal{G}}.$$

The left-hand side is a polynomial in θ that vanishes on the nonempty open set $\Theta_{\mathcal{G}}$, so it is the zero polynomial. Since the polynomial ring $\mathbb{R}[\theta]$ is an integral domain, $g \circ \phi_{\mathcal{G}} \equiv 0$ or $h \circ \phi_{\mathcal{G}} \equiv 0$. In the first case g vanishes on $\mathcal{C}_{\mathcal{G}}$ and hence on its Zariski closure $V_{\mathcal{G}}$, so $g \in I(V_{\mathcal{G}})$; likewise $h \in I(V_{\mathcal{G}})$ in the second case. Thus $I(V_{\mathcal{G}})$ is prime, and $V_{\mathcal{G}}$ is irreducible. $\square$

Lemma 12. *There is a proper algebraic subset $E \subsetneq Y$ such that, for every $S \in \mathcal{C}_{\mathcal{G}^*} \setminus E$ and every $\mathcal{G} \in \mathbb{G}$:*

- (i) $S \in V_{\mathcal{G}}$ implies $Y \subseteq V_{\mathcal{G}}$;
- (ii) every $\theta \in F_{\mathcal{G}}(S) := \{\theta \in \Theta_{\mathcal{G}} : \phi_{\mathcal{G}}(\theta) = S\}$ satisfies $\text{rank } D\phi_{\mathcal{G}}(\theta) \geq d$;
- (iii) if $Y \subsetneq V_{\mathcal{G}}$ and $F_{\mathcal{G}}(S) \neq \emptyset$, then $\dim F_{\mathcal{G}}(S) \leq k_{\mathcal{G}} - d - 1$.

Proof. We construct E as a finite union of proper algebraic subsets of Y , one for each property and each candidate graph, and use the irreducibility of Y (Lem. 11) to keep the union proper.

For (i), consider a graph $\mathcal{G}$ with $Y \not\subseteq V_{\mathcal{G}}$. Then $Y \cap V_{\mathcal{G}}$ is an algebraic subset of Y , and it is proper by assumption; we include it in E . For $S \in Y \setminus E$ the implication in (i) then holds, since $S \in V_{\mathcal{G}}$ together with $Y \not\subseteq V_{\mathcal{G}}$ would give $S \in Y \cap V_{\mathcal{G}} \subseteq E$.

For (ii), consider the semialgebraic set of rank-deficient parameters

$$C_{\mathcal{G},d} = \{\theta \in \Theta_{\mathcal{G}} : \text{rank } D\phi_{\mathcal{G}}(\theta) < d\}.$$

It admits a finite smooth semialgebraic stratification on which the restricted map has constant rank. On each stratum this rank is at most $\text{rank } D\phi_{\mathcal{G}} < d$, so every stratum image has dimension at most $d - 1$, and so does $\phi_{\mathcal{G}}(C_{\mathcal{G},d})$. Since semialgebraic sets and their real Zariski closures have the same dimension,

$$\dim\left(Y \cap \overline{\phi_{\mathcal{G}}(C_{\mathcal{G},d})}^Z\right) \leq d - 1 < \dim Y.$$

This algebraic subset therefore cannot contain Y , and we include it in E . For $S \in Y \setminus E$, no $\theta \in F_{\mathcal{G}}(S)$ lies in $C_{\mathcal{G},d}$, which is (ii).

For (iii), consider a graph $\mathcal{G}$ with $Y \subsetneq V_{\mathcal{G}}$, and let

$$Z_{\mathcal{G}} = \phi_{\mathcal{G}}^{-1}(Y) \subseteq \Theta_{\mathcal{G}}.$$

This is an algebraic subset relative to the open parameter domain. It is proper: if $Z_{\mathcal{G}} = \Theta_{\mathcal{G}}$, then $\mathcal{C}_{\mathcal{G}} \subseteq Y$ and hence $V_{\mathcal{G}} \subseteq Y$, contradicting $Y \subsetneq V_{\mathcal{G}}$. Some polynomial vanishing on Y therefore pulls back to a polynomial that is not identically zero on $\Theta_{\mathcal{G}}$, and $\dim Z_{\mathcal{G}} \leq k_{\mathcal{G}} - 1$. Stratify $Z_{\mathcal{G}}$ into finitely many smooth semialgebraic strata on which the restricted map has constant rank. For the strata of rank below d , the Zariski closures of their images have dimension below d , and we include their intersections with Y in E . Every remaining stratum M maps into Y , which has dimension d , so its constant rank is exactly d , and by the constant-rank theorem its fibers have dimension

$$\dim M - d \leq k_{\mathcal{G}} - 1 - d.$$

Now let $S \in Y \setminus E$. Then $F_{\mathcal{G}}(S) \subseteq Z_{\mathcal{G}}$ meets only strata of rank d , since S lies outside the image closures of the others, and $F_{\mathcal{G}}(S)$ is the finite union of its intersections with these strata. Hence $\dim F_{\mathcal{G}}(S) \leq k_{\mathcal{G}} - d - 1$, which is (iii).

There are finitely many candidate graphs and finitely many strata, so E is a finite union of proper algebraic subsets of Y . Since Y is irreducible, this union remains proper. This completes the proof. $\square$

The exceptional set E is small in the prior sense, which is what allows the conclusions to hold for prior-almost every generating parameter.

Lemma 13. *Let E be as in Lem. 12. Then*

$$\Pi_{\mathcal{G}^*} \{\theta^* : \phi_{\mathcal{G}^*}(\theta^*) \in E\} = 0.$$

Proof. Since $E \subsetneq Y$ is algebraic, its vanishing ideal strictly contains that of Y , so there is a polynomial h that vanishes on E but not identically on Y . The polynomial $h \circ \phi_{\mathcal{G}^*}$ in θ^* is not identically zero: otherwise h would vanish on $\mathcal{C}_{\mathcal{G}^*}$, and hence on its Zariski closure Y . Moreover,

$$\{\theta^* : \phi_{\mathcal{G}^*}(\theta^*) \in E\} \subseteq \{\theta^* : h(\phi_{\mathcal{G}^*}(\theta^*)) = 0\}.$$

The zero set of a nonzero polynomial has Lebesgue measure zero, and $\Pi_{\mathcal{G}^*}$ has a bounded density, so the left-hand set has prior measure zero. $\square$

The boundary of the full-dimensional semialgebraic support also has Lebesgue measure zero. For the remainder of the proof we therefore fix

$$\theta^* \in \text{int } W_{G^*} \quad \text{with} \quad S = \phi_{G^*}(\theta^*) \notin E,$$

and write $q = \mathcal{N}_p(0, S)$. For $\mathcal{G} \in \mathbb{G}$, let $K_{\mathcal{G}}(\theta) = \text{KL}(q \| p_{\mathcal{G}, \theta})$. The next lemma reduces the RLCT at a single representation of S to the rank of the Jacobian and an analytic residual.

Lemma 14. *Let $\theta_0 \in F_{\mathcal{G}}(S)$ and $r = \text{rank } D\phi_{\mathcal{G}}(\theta_0)$. There are local analytic coordinates $(u, v) \in \mathbb{R}^r \times \mathbb{R}^{k_{\mathcal{G}}-r}$ centered at θ_0 , with nonsingular change of variables, in which*

$$K_{\mathcal{G}}(u, v) = u_1^2 + \cdots + u_r^2 + R(v), \quad R \geq 0,$$

for an analytic function R with $R(0) = 0$. Consequently, the local RLCT at θ_0 , defined by integration over an ambient neighborhood with positive analytic density, is

$$\lambda_{\mathcal{G}, \theta_0}(q) = r/2 + \lambda(R), \quad (14)$$

where $\lambda(R) = 0$ if $R \equiv 0$ and $\lambda(R) > 0$ otherwise. Restricting the integration to $W_{\mathcal{G}}$ cannot decrease the exponent.

Proof. We first relate $K_{\mathcal{G}}$ to the covariance map. For centered Gaussians,

$$K_{\mathcal{G}}(\theta) = \frac{1}{2} \left[\text{tr}(\phi_{\mathcal{G}}(\theta)^{-1} S) - p + \log \det \phi_{\mathcal{G}}(\theta) - \log \det S \right].$$

Expanding in $\Delta = \phi_{\mathcal{G}}(\theta) - S$, the constant and linear terms cancel and

$$K_{\mathcal{G}}(\theta) = \frac{1}{4} \|S^{-1/2} \Delta S^{-1/2}\|_F^2 + O(\|\Delta\|^3).$$

The quadratic term is positive definite in covariance coordinates, so

$$K_{\mathcal{G}}(\theta) \asymp \|\phi_{\mathcal{G}}(\theta) - S\|_F^2$$

near θ_0 . In particular, $K_{\mathcal{G}}$ is analytic and nonnegative and vanishes at θ_0 together with its gradient, and its Hessian at θ_0 equals $\frac{1}{2} D\phi_{\mathcal{G}}(\theta_0)^{\top} \mathcal{I}_S D\phi_{\mathcal{G}}(\theta_0)$ for a positive-definite matrix $\mathcal{I}_S$. This Hessian has rank r . The analytic splitting lemma therefore provides local analytic coordinates (u, v) centered at θ_0 , with nonsingular change of variables, in which

$$K_{\mathcal{G}}(u, v) = u_1^2 + \cdots + u_r^2 + R(v),$$

where R is analytic with $R(0) = 0$; setting $u = 0$ shows that $R \geq 0$.

We next integrate out the quadratic coordinates. The positive analytic density and the Jacobian of the change of variables are bounded above and below near θ_0 , so the local Laplace integral factorizes up to constants:

$$\int e^{-nK_{\mathcal{G}}(\theta)} d\theta \asymp \int e^{-n(u_1^2 + \cdots + u_r^2)} du \int e^{-nR(v)} dv.$$

The first factor decays as $n^{-r/2}$. If $R \equiv 0$, the second factor is bounded above and below, and we set $\lambda(R) = 0$. Otherwise, R is a nonzero nonnegative analytic function vanishing at the origin, and resolution-based Laplace asymptotics show that the second factor decays as $n^{-\lambda(R)}$ up to logarithmic factors, with $\lambda(R) > 0$ (Watanabe, 2009). The exponents add, which gives Eq. (14). Finally, restricting the integration to $W_{\mathcal{G}}$ can only shrink the integral, so the exponent cannot decrease. $\square$

We can now compute the RLCTs that Lem. 3 compares.

Proof of Lem. 3. By Lem. 13, it suffices to prove both parts for the fixed θ^* chosen above. We first prove (i), which concerns the generating graph, and then (ii), which concerns compatible graphs outside the AEC.

For (i), we bound the rank of $D\phi_{G^*}$ from both sides. The image $\mathcal{C}_{G^*}$ has dimension d , so

$$\text{rank } D\phi_{G^*}(\theta) \leq d \quad \text{for every } \theta \in \Theta_{G^*};$$

otherwise the rank would exceed d on an open set, and the constant-rank theorem, applied at a point of locally maximal rank, would place a submanifold of dimension larger than d inside $\mathcal{C}_{\mathcal{G}^*}$. Conversely, Lem. 12(ii) gives $\text{rank } D\phi_{\mathcal{G}^*}(\theta_0) \geq d$ at every representation $\theta_0 \in F_{\mathcal{G}^*}(S)$. Since the rank is lower semicontinuous, it equals d on a neighborhood of each representation. By the constant-rank theorem, the fiber $F_{\mathcal{G}^*}(S)$, which is the zero set of $K_{\mathcal{G}^*}$, is near θ_0 a smooth manifold of dimension $k_{\mathcal{G}^*} - d$. In the coordinates of Lem. 14, with $r = d$, this zero set is

$$\{u = 0\} \times \{v : R(v) = 0\},$$

which has dimension $k_{\mathcal{G}^*} - d$, the dimension of the v -space, only if $R \equiv 0$. Hence the local exponent at every representation equals $d/2$. A finite cover of the compact prior-supported fiber $F_{\mathcal{G}^*}(S) \cap W_{\mathcal{G}^*}$ by such neighborhoods bounds the global exponent below by $d/2$, and the interior representation θ^* , where the prior density is positive, attains this exponent. Thus

$$\lambda_{\mathcal{G}^*}(q) = d/2. \quad (15)$$

For (ii), let $\mathcal{G} \in \mathbb{G}$ realize q within $W_{\mathcal{G}}$, so that $S \in \phi_{\mathcal{G}}(W_{\mathcal{G}})$, and suppose $V_{\mathcal{G}} \neq Y$. Since $S \in \mathcal{C}_{\mathcal{G}} \subseteq V_{\mathcal{G}}$, Lem. 12(i) gives $Y \subseteq V_{\mathcal{G}}$, and hence

$$Y \subsetneq V_{\mathcal{G}}.$$

Let $\theta_0 \in F_{\mathcal{G}}(S)$ and $r = \text{rank } D\phi_{\mathcal{G}}(\theta_0)$. By Lem. 12(ii) and (iii),

$$r \geq d \quad \text{and} \quad \dim F_{\mathcal{G}}(S) \leq k_{\mathcal{G}} - d - 1.$$

If $r > d$, Eq. (14) gives a local exponent of at least $r/2 > d/2$. If $r = d$, then R cannot vanish identically: otherwise the local zero set $\{u = 0\}$ of $K_{\mathcal{G}}$, which is contained in $F_{\mathcal{G}}(S)$, would have dimension $k_{\mathcal{G}} - d$, contradicting the fiber bound. Hence $\lambda(R) > 0$, and again the local exponent exceeds $d/2$.

It remains to turn these local bounds into a strict global gap. At each point of the compact set $F_{\mathcal{G}}(S) \cap W_{\mathcal{G}}$, choose a neighborhood O and an exponent $s_O > d/2$ such that $\int_O K_{\mathcal{G}}^{-s_O} d\theta < \infty$, which is possible because the local exponent exceeds $d/2$. Compactness gives a finite subcover $O_1, \dots, O_m$, and

$$s = \min_{1 \leq j \leq m} s_{O_j} > d/2.$$

Away from $O_1 \cup \dots \cup O_m$, the continuous function $K_{\mathcal{G}}$ is bounded away from zero on $W_{\mathcal{G}}$. Since the prior density is bounded, it follows that

$$\int_{W_{\mathcal{G}}} K_{\mathcal{G}}(\theta)^{-s} \varphi_{\mathcal{G}}(\theta) d\theta < \infty,$$

so the zeta function $z \mapsto \int_{W_{\mathcal{G}}} K_{\mathcal{G}}(\theta)^z \varphi_{\mathcal{G}}(\theta) d\theta$ is holomorphic on $\{\text{Re } z > -s\}$, and its largest pole lies at or below $-s$. Hence the prior-supported RLCT satisfies

$$\lambda_{\mathcal{G}}(q) \geq s > d/2 = \lambda_{\mathcal{G}^*}(q). \quad (16)$$

No full-rank assumption at competing representations is needed. Eqs. (15) and (16) prove parts (i) and (ii). This completes the proof. $\square$

Proof of Thm. 4. Fix θ^* as above, so that both parts of Lem. 3 hold, and let $\mathcal{G} \in \mathbb{G}$ with $\mathcal{G} \notin \text{AEC}(\mathcal{G}^*)$. We show that

$$P_q^{\otimes n} \{ \text{WBIC}_n(\mathcal{G}) \leq \text{WBIC}_n(\mathcal{G}^*) \} \longrightarrow 0,$$

separately for competitors that realize q within their prior support and for those that do not, and then combine the two cases by a union bound.

We first treat a graph $\mathcal{G}$ whose prior support realizes q , that is, $S \in \phi_{\mathcal{G}}(W_{\mathcal{G}})$. The generating graph realizes q as well, since $\theta^* \in W_{\mathcal{G}^*}$. For each such graph, Lem. 10 and Watanabe (2013, Theorem 4), applied as in the proof of Prop. 1, give

$$\text{WBIC}_n(\mathcal{G}) = - \sum_{i=1}^n \log q(v_i) + \lambda_{\mathcal{G}}(q) \log n + o_{P_q}(\log n), \quad (17)$$

where the fit term is common to all graphs realizing q because $nL_n(\theta_{0,\mathcal{G}}) = -\sum_{i=1}^n \log q(\mathbf{v}_i)$. Since $V_{\mathcal{G}} \neq Y$, subtracting Eq. (17) for $\mathcal{G}^*$ and using Eqs. (15) and (16) gives

$$\frac{\text{WBIC}_n(\mathcal{G}) - \text{WBIC}_n(\mathcal{G}^*)}{\log n} \xrightarrow{P_q} \lambda_{\mathcal{G}}(q) - d/2 > 0,$$

and hence $P_q^{\otimes n}\{\text{WBIC}_n(\mathcal{G}) \leq \text{WBIC}_n(\mathcal{G}^*)\} \rightarrow 0$.

We next treat a graph $\mathcal{G}$ with $S \notin \phi_{\mathcal{G}}(W_{\mathcal{G}})$. Since $K_{\mathcal{G}}$ is continuous and positive on the compact set $W_{\mathcal{G}}$,

$$\delta_{\mathcal{G}} := \min_{\theta \in W_{\mathcal{G}}} K_{\mathcal{G}}(\theta) > 0.$$

Write $H(q) = -\mathbb{E}_q \log q(X)$ and $L(\theta) = -\mathbb{E}_q \log p(X | \theta, \mathcal{G}) = H(q) + K_{\mathcal{G}}(\theta)$, and let $m = \min_{W_{\mathcal{G}}} L = H(q) + \delta_{\mathcal{G}}$. The empirical loss converges uniformly on $W_{\mathcal{G}}$: its parameter-dependent part is determined by S_n , and by Eq. (4),

$$\sup_{\theta \in W_{\mathcal{G}}} |L_n(\theta) - L(\theta)| = \frac{1}{2} \sup_{\theta \in W_{\mathcal{G}}} \left| \text{tr}(\phi_{\mathcal{G}}(\theta)^{-1}(S_n - S)) \right| \leq \frac{\sqrt{p}}{2} \sup_{\theta \in W_{\mathcal{G}}} \|\phi_{\mathcal{G}}(\theta)^{-1}\| \|S_n - S\|_F \xrightarrow{P_q} 0,$$

since the inverse covariances are uniformly bounded on the compact set $W_{\mathcal{G}}$ and $S_n \rightarrow S$ in probability. We next show that the tempered posterior concentrates near the minimizers of L . Write $\beta_n = 1/\log n$, fix $\eta > 0$, and work on the event $\sup_{W_{\mathcal{G}}} |L_n - L| < \eta/4$, whose probability tends to one. On this event, $L_n < m + \eta/2$ on $\{L < m + \eta/4\}$ and $L_n > m + 3\eta/4$ on $\{L \geq m + \eta\}$, so

$$\pi_n^{\beta_n}\{L \geq m + \eta\} \leq \frac{e^{-n\beta_n(m+3\eta/4)}}{e^{-n\beta_n(m+\eta/2)} \Pi_{\mathcal{G}}\{L < m + \eta/4\}} = \frac{e^{-n\beta_n\eta/4}}{\Pi_{\mathcal{G}}\{L < m + \eta/4\}}.$$

The denominator is positive: $\{L < m + \eta/4\}$ is a nonempty relatively open subset of $W_{\mathcal{G}}$, which meets $\text{int } W_{\mathcal{G}}$ because $W_{\mathcal{G}} = \text{int } \bar{W}_{\mathcal{G}}$, and the prior density is positive there. Since $n\beta_n = n/\log n \rightarrow \infty$, the right-hand side tends to zero. Splitting the posterior expectation of $L - m \geq 0$ over $\{L < m + \eta\}$ and its complement then gives

$$\left| \mathbb{E}_{\theta}^{\beta_n}[L_n] - m \right| \leq \sup_{W_{\mathcal{G}}} |L_n - L| + \eta + \left(\max_{W_{\mathcal{G}}} L - m \right) \pi_n^{\beta_n}\{L \geq m + \eta\},$$

whose right-hand side converges in probability to η . Since $\eta > 0$ is arbitrary, $\mathbb{E}_{\theta}^{\beta_n}[L_n] \rightarrow m$ in probability. Applying this argument to $\mathcal{G}$, with $m = H(q) + \delta_{\mathcal{G}}$, and to $\mathcal{G}^*$, with $m = H(q)$, we obtain

$$\frac{\text{WBIC}_n(\mathcal{G})}{n} \xrightarrow{P_q} H(q) + \delta_{\mathcal{G}}, \quad \frac{\text{WBIC}_n(\mathcal{G}^*)}{n} \xrightarrow{P_q} H(q),$$

so again $P_q^{\otimes n}\{\text{WBIC}_n(\mathcal{G}) \leq \text{WBIC}_n(\mathcal{G}^*)\} \rightarrow 0$. This case includes graphs that could realize q outside their chosen prior support, and it uses no misspecified singular expansion.

We finally combine the two cases. If a measurable minimizer $\hat{\mathcal{G}}_n$ lies outside $\text{AEC}(\mathcal{G}^*)$, then $\text{WBIC}_n(\hat{\mathcal{G}}_n) \leq \text{WBIC}_n(\mathcal{G}^*)$, so the finite union bound gives

$$P_q^{\otimes n}\{\hat{\mathcal{G}}_n \notin \text{AEC}(\mathcal{G}^*)\} \leq \sum_{\mathcal{G} \notin \text{AEC}(\mathcal{G}^*)} P_q^{\otimes n}\{\text{WBIC}_n(\mathcal{G}) \leq \text{WBIC}_n(\mathcal{G}^*)\} \rightarrow 0.$$

This holds for every $\theta^* \in \text{int } W_{\mathcal{G}^*}$ with $\phi_{\mathcal{G}^*}(\theta^*) \notin E$, a set of full prior measure by Lem. 13. Therefore

$$\Pi_{\mathcal{G}^*}\left(\left\{\theta^* : \lim_{n \rightarrow \infty} P_{\mathcal{G}^*, \theta^*}^{\otimes n}\{\hat{\mathcal{G}}_n \in \text{AEC}(\mathcal{G}^*)\} = 1\right\}\right) = 1.$$

The conclusion concerns the full models' polynomial equalities and does not require identical finite-sample scores within the class. This completes the proof. $\square$

D.4 Prior misspecification

Thm. 4 is stated for generating parameters drawn from the prior $\Pi_{\mathcal{G}^*}$ used to score $\mathcal{G}^*$. The next proposition lets the generating parameter follow a different prior $\Pi_{\mathcal{H}}$, while every graph is still scored under its own prior. Its proof integrates the pointwise conclusion of Thm. 4 against $\Pi_{\mathcal{H}}$ and charges the exceptional set to the total variation distance between the two priors.

Proposition 15. *Under the standing setup, fix $\mathcal{G}^* \in \mathbb{G}$ and let $\Pi_{\mathcal{H}}$ be a prior on $\Theta_{\mathcal{G}^*}$. Draw $\theta^* \sim \Pi_{\mathcal{H}}$ and, given θ^* , i.i.d. observations from $p_{\mathcal{G}^*, \theta^*}$, and write $P_{\mathcal{H}}$ for their joint law. Then any measurable choice $\widehat{\mathcal{G}}_n \in \arg \min_{\mathcal{G} \in \mathbb{G}} \text{WBIC}_n(\mathcal{G})$ satisfies*

$$\limsup_{n \rightarrow \infty} P_{\mathcal{H}}\{\widehat{\mathcal{G}}_n \notin \text{AEC}(\mathcal{G}^*)\} \leq d_{\text{TV}}(\Pi_{\mathcal{H}}, \Pi_{\mathcal{G}^*}),$$

where $d_{\text{TV}}(\Pi_{\mathcal{H}}, \Pi_{\mathcal{G}^*}) = \sup_A |\Pi_{\mathcal{H}}(A) - \Pi_{\mathcal{G}^*}(A)|$, the supremum running over Borel sets A . In particular, if $\Pi_{\mathcal{H}}$ is absolutely continuous with respect to $\Pi_{\mathcal{G}^*}$, then $P_{\mathcal{H}}\{\widehat{\mathcal{G}}_n \in \text{AEC}(\mathcal{G}^*)\} \rightarrow 1$.

Proof. The proof proceeds in two steps. We first express the error probability under $P_{\mathcal{H}}$ as the integral of a pointwise error probability against $\Pi_{\mathcal{H}}$, and then split this integral over the exceptional set of Thm. 4 and its complement.

For $\theta^* \in \Theta_{\mathcal{G}^*}$, let

$$f_n(\theta^*) = P_{\mathcal{G}^*, \theta^*}^{\otimes n}\{\widehat{\mathcal{G}}_n \notin \text{AEC}(\mathcal{G}^*)\} \in [0, 1].$$

The Gaussian density is jointly continuous in the observations and θ^* , and $\widehat{\mathcal{G}}_n$ is a measurable function of the observations, so f_n is measurable by Fubini's theorem. Conditioning on θ^* ,

$$P_{\mathcal{H}}\{\widehat{\mathcal{G}}_n \notin \text{AEC}(\mathcal{G}^*)\} = \int f_n d\Pi_{\mathcal{H}}.$$

We next isolate the exceptional set. Let

$$N = \{\theta^* : f_n(\theta^*) \not\rightarrow 0\} = \left\{ \theta^* : \limsup_{n \rightarrow \infty} f_n(\theta^*) > 0 \right\}.$$

This set is measurable, and $\Pi_{\mathcal{G}^*}(N) = 0$ by Thm. 4. Splitting the integral over N^c and N and using $f_n \leq 1$,

$$\int f_n d\Pi_{\mathcal{H}} = \int_{N^c} f_n d\Pi_{\mathcal{H}} + \int_N f_n d\Pi_{\mathcal{H}} \leq \int_{N^c} f_n d\Pi_{\mathcal{H}} + \Pi_{\mathcal{H}}(N).$$

On N^c , $f_n \rightarrow 0$ pointwise and $0 \leq f_n \leq 1$, so the first term tends to zero by dominated convergence. For the second term, since $\Pi_{\mathcal{G}^*}(N) = 0$,

$$\Pi_{\mathcal{H}}(N) = \Pi_{\mathcal{H}}(N) - \Pi_{\mathcal{G}^*}(N) \leq d_{\text{TV}}(\Pi_{\mathcal{H}}, \Pi_{\mathcal{G}^*}).$$

Taking the limit superior proves the bound. If $\Pi_{\mathcal{H}} \ll \Pi_{\mathcal{G}^*}$, then $\Pi_{\mathcal{H}}(N) = 0$, the limit superior is zero, and $P_{\mathcal{H}}\{\widehat{\mathcal{G}}_n \in \text{AEC}(\mathcal{G}^*)\} \rightarrow 1$. This completes the proof. $\square$

E METROPOLIS–HASTINGS ESTIMATION OF THE WBIC

This appendix details the Metropolis–Hastings (MH) estimator of Eq. (3), which produces the WBIC rows of Tables 1 and 3. Fix an ADMG $\mathcal{G} = (\mathbf{V}, \mathbf{D}, \mathbf{B})$. The chain runs on the free coordinates $\theta = (\Lambda, \mathcal{E})$ of $\mathcal{G}$: the coefficients λ_{ij} of the directed edges in $\mathbf{D}$, the error variances ω_{ii} , and the error covariances ω_{ij} of the bidirected edges in $\mathbf{B}$, with all other entries held at zero. Its target is the tempered posterior

$$\pi_n^\beta(\theta \mid \mathbf{D}_n, \mathcal{G}) \propto e^{-n\beta L_n(\theta)} \varphi_{\mathcal{G}}(\theta), \quad \beta = \frac{1}{\log n},$$

whose normalizing constant, the denominator of Eq. (2), is never needed because it cancels in the acceptance ratio. Algorithm 2 summarizes the procedure.

Each loss evaluation uses Eq. (4), so the data enter only through the sample covariance S_n . A proposal outside the positive-definite cone has zero prior density and is rejected without evaluating the likelihood. Every rejection, of either kind, repeats the current state, which is counted again in the final average. The Gaussian proposal is symmetric, so the proposal densities cancel from the acceptance ratio, and the chain is reversible with respect to π_n^β . Under standard ergodicity conditions, its average in Algorithm 2 therefore converges to the tempered expectation of Eq. (2) as $T \rightarrow \infty$. Scoring a family $\mathbb{G}$ requires one such chain per candidate graph.

Algorithm 2 WBIC estimation by random-walk Metropolis–Hastings

Require: data $\mathbf{D}_n$, graph $\mathcal{G}$, inverse temperature β , iterations T , warm-up length $T_0 < T$, step sizes $s_\Lambda, s_\mathcal{E}$

- 1: Initialize $\theta^{(0)} = (\Lambda^{(0)}, \mathcal{E}^{(0)})$ on the free coordinates of $\mathcal{G}$, with $\mathcal{E}^{(0)} \succ 0$.
 - 2: **for** $t = 1, \dots, T$ **do**
 - 3: Propose $\theta' = \theta^{(t-1)} + \varepsilon$, with independent increments $\varepsilon_k \sim \mathcal{N}(0, s_\Lambda^2)$ on the coordinates of Λ and $\varepsilon_k \sim \mathcal{N}(0, s_\mathcal{E}^2)$ on the coordinates of $\mathcal{E}$.
 - 4: **if** the error covariance $\mathcal{E}'$ of θ' is not positive definite **then**
 - 5: Reject: $\theta^{(t)} \leftarrow \theta^{(t-1)}$.
 - 6: **else**
 - 7: $\log \alpha \leftarrow n\beta [L_n(\theta^{(t-1)}) - L_n(\theta')] + \log \varphi_{\mathcal{G}}(\theta') - \log \varphi_{\mathcal{G}}(\theta^{(t-1)})$.
 - 8: Draw $u \sim \text{Unif}(0, 1)$. Set $\theta^{(t)} \leftarrow \theta'$ if $\log u < \log \alpha$, and $\theta^{(t)} \leftarrow \theta^{(t-1)}$ otherwise.
 - 9: **end if**
 - 10: **end for**
 - 11: **return** $\widehat{\text{WBIC}}_{\text{MH}}(\mathcal{G}) = \frac{1}{T-T_0} \sum_{t=T_0+1}^T nL_n(\theta^{(t)})$.
-

We now record the settings used for Tables 1 and 3. For Examples 1, 2, and 6, $n = 1000$, so $\beta = 1/\log 1000 \approx 0.145$. Unlike the ADVI surrogate of Section 4, the MH estimator is always run at this standard temperature. The prior places independent $\mathcal{N}(0, 2^2)$ densities on the directed coefficients and a flat density on the free entries of $\mathcal{E}$, restricted to positive-definite matrices with the zero pattern of $\mathcal{G}$. Each chain starts from $\lambda_{ij} \sim \mathcal{N}(0, 0.5^2)$ on the directed edges and from $\mathcal{E}^{(0)}$ with diagonal entries 1.5 and entries 0.3 on the bidirected edges. The step sizes are $s_\Lambda = 0.08$ and $s_\mathcal{E} = 0.06$, and every iteration perturbs all free coordinates jointly. For Table 1, each chain runs $T = 60,000$ iterations, of which the first $T_0 = 15,000$ are discarded; for Table 3, each chain runs $T = 15,000$ iterations with $T_0 = 3,000$. In each of the $R = 30$ replicates, we run one chain for $\mathcal{G}$ and one for $\mathcal{G}'$ on the same data, with a shared random seed. As a numerical safeguard, the implementation replaces any proposed diagonal entry $\omega_{ii} \leq 10^{-6}$ by $|\omega_{ii}| + 10^{-4}$ before the positive-definiteness check. This makes the proposal slightly asymmetric at the boundary $\omega_{ii} = 0$, and it is inactive elsewhere.

F EXPERIMENTAL SETUP

This appendix collects the full specification of the experiments summarized in the main text, so that each result can be reproduced from the settings given here. We describe, in turn, how the ground-truth ADMGs are generated for the randomized and singular benchmarks, the configuration of the RICF-based ABIC baseline, and the ADVI optimization procedure used by our method, including its schedules and hyperparameters. Throughout, the two methods are evaluated on the same observations in every repetition. Unless noted otherwise, we verify the graph support, the positive definiteness of Ω , and the rank deficiency of the relevant constraint matrix before generating any observations; the rank check combines the numerical matrix rank with the smallest singular value.

General random ADMGs. For the randomized experiments, the directed and bidirected skeletons are drawn from two independent Erdős–Rényi models, with edge probabilities $p_{\text{dir}} = 0.35$ and $p_{\text{bidir}} = 0.20$, respectively. Both skeletons are fixed before the nonzero SCM parameters are sampled.

Singular ADMGs. The construction of Singular I is described in the main text. For Singular II, we first sample the free diagonal and bidirected entries of Ω , keeping only positive-definite draws that satisfy $|\omega_{12}|, |\omega_{23}| \geq 0.2$. We then sample λ_{34} and λ_{24} freely and set

$$\lambda_{13} = \frac{\omega_{22}\omega_{33} - \omega_{23}^2}{\omega_{12}\omega_{23}}.$$

The lower bounds on $|\omega_{12}|$ and $|\omega_{23}|$ keep the denominator away from zero, so this assignment is well defined and enforces the rank-deficient constraint

$$\det(M_3) = \omega_{22}\omega_{33} - \omega_{23}^2 - \lambda_{13}\omega_{12}\omega_{23} = 0.$$

Table 4: Use of the bow-free constraint for ABIC.

Experiment	Bow-free constraint
Sachs	No
General randomized ADMGs	No
Singular I	No
Singular II	Yes
Singular III	Yes

For Singular III, we sample the free directed coefficients and the free entries of Ω , and then impose two zero constraints by setting

$$\begin{aligned}\omega_{24} &= -\lambda_{23}\lambda_{34}\omega_{22}, \\ \omega_{44} &= -\lambda_{12}\lambda_{23}\lambda_{34}\omega_{14} + (\lambda_{23}\lambda_{34})^2\omega_{22}.\end{aligned}$$

These choices are equivalent to enforcing

$$\begin{aligned}m_1 &= \lambda_{23}\lambda_{34}\omega_{22} + \omega_{24} = 0, \\ m_2 &= \lambda_{12}\lambda_{23}\lambda_{34}\omega_{14} + \lambda_{23}\lambda_{34}\omega_{24} + \omega_{44} = 0.\end{aligned}$$

A proposed Singular III point is rejected unless the resulting ω_{44} is positive and $\Omega \succ 0$.

ABIC baseline. For the baseline we use the RICF-based ABIC procedure of Bhattacharya et al. (2021), thresholding the final edge weights at 0.05: any directed or bidirected edge whose fitted weight is at most 0.05 in magnitude is dropped from the returned ADMG. The ABIC method performs five random restarts on each dataset; it then re-fits the BIC for the resulting candidates and selects the lowest one as the final result. While we conducted five random restarts for the singular and random experiments, we used a single initialization for the Sachs experiment to stay within our computational budget. The bow-free constraint is imposed only where indicated in Table 4: Singular II and Singular III use it because their ground-truth graphs are bow-free, while the remaining experiments use the unconstrained ADMG class.

ADVI optimization. For each data set, we run the continuous ADMG search from five independently initialized structure-logit configurations ranging from sparse to dense graph tendencies. The initialization configurations are

$$(-3, -3), \quad (-3, 0), \quad (0, 0), \quad (3, 0), \quad (3, 3),$$

where each pair denotes the initial means of the directed and bidirected structure logits, respectively. The continuous search alternates between variational and graph updates. Given graph gates A_D and A_B , the variational parameters are updated by maximizing the tempered ELBO

$$\mathcal{L}_\beta = \beta \mathbb{E}_{q_\psi}[\ell_n(\Lambda, L)] - \text{KL}(q_\psi \| p), \quad \beta = \frac{c}{\log n},$$

where ℓ_n is the Gaussian linear-SEM log likelihood and q_ψ is a mean-field Gaussian distribution over the coefficient and Cholesky coordinates.

The Gumbel soft gates temperature is annealed geometrically from $\tau_0 = 1.0$ to $\tau_{\text{final}} = 0.2$. The acyclicity penalty is also increased geometrically from $\lambda_{\text{acyc}}^{\text{init}}$ to $\lambda_{\text{acyc}}^{\text{final}}$ over the complete optimization run. In the synthetic experiments, the parameters are $T = 3000, \eta = 0.01, M_{\text{VI}} = M_{\text{graph}} = 8, \lambda_{\text{acyc}}^{\text{init}} = 1, \lambda_{\text{acyc}}^{\text{final}} = 100$. For all three singular graphs, the random graphs of 4 nodes and 5 nodes, the sparsity penalty parameters $\lambda_D = 1, \lambda_B = 1$. For 8 nodes random graph, $\lambda_D = 10, \lambda_B = 10$.

After continuous optimization, each relaxed solution is converted into a discrete ADMG using fixed probability thresholds

$$\hat{D}_{ij} = \mathbf{1}\{P_{ij}^D > 0.5\}, \quad \hat{B}_{ij} = \mathbf{1}\{P_{ij}^B > 0.5\}.$$

The directed part is subsequently projected to a DAG. Edges whose inclusion would create a directed cycle are considered in increasing order of importance, with higher-probability edges retained first and cycle-forming edges

removed. The resulting directed graph is therefore acyclic, while the bidirected matrix is symmetrized from its upper-triangular thresholded form.

Distinct thresholded ADMGs obtained from the five initializations are deduplicated. Each distinct discrete graph is then refitted independently using fixed-graph ADVI. During this refit, the graph is held fixed, and the variational parameters are optimized using

$$\mathcal{L}_\beta^{\text{fixed}} = \beta \mathbb{E}_{q_\psi}[\ell_n(\Lambda, L)] - \text{KL}(q_\psi \| p).$$

For the default experiments, fixed-graph refitting uses 2000 optimization steps, learning rate 0.01, and 8 Monte Carlo samples per update.

The final score for a refitted graph is estimated from $S = 1000$ samples from the fitted variational posterior,

$$\widehat{\text{WBIC}}_\beta(\mathcal{G}) = -\frac{1}{S} \sum_{s=1}^S \ell_n(\Lambda^{(s)}, L^{(s)}), \quad (\Lambda^{(s)}, L^{(s)}) \sim q_{\psi^*}.$$

Among the distinct refitted ADMGs, we select the graph with the smallest estimated score. For evaluation, the selected ADMG and the ground-truth ADMG are converted to PAGs, and structural metrics are computed at the PAG level.