On the Lasso for Graphical Continuous Lyapunov Models

Philipp Dettling   PHILIPP.DETTLING@TUM.DE
Technical University of Munich, Department of Mathematics

Mathias Drton   MATHIAS.DRTON@TUM.DE
Technical University of Munich, Department of Mathematics and Munich Data Science Institute
Munich Center for Machine Learning (MCML)

Mladen Kolar   MKOLAR@USC.EDU
University of Southern California, Marshall School of Business

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Graphical continuous Lyapunov models offer a new perspective on modeling causally interpretable dependence structure in multivariate data by treating each independent observation as a one-time cross-sectional snapshot of a temporal process. Specifically, the models assume that the observations are cross-sections of independent multivariate Ornstein-Uhlenbeck processes in equilibrium. The Gaussian equilibrium exists under a stability assumption on the drift matrix, and the equilibrium covariance matrix is determined by the continuous Lyapunov equation. Each graphical continuous Lyapunov model assumes the drift matrix to be sparse, with a support determined by a directed graph. A natural approach to model selection in this setting is to use an ℓ_1-regularization technique that, based on a given sample covariance matrix, seeks to find a sparse approximate solution to the Lyapunov equation. We study the model selection properties of the resulting lasso technique to arrive at a consistency result. Our detailed analysis reveals that the involved irrepresentability condition is surprisingly difficult to satisfy. While this may prevent asymptotic consistency in model selection, our numerical experiments indicate that even if the theoretical requirements for consistency are not met, the lasso approach is able to recover relevant structure of the drift matrix and is robust to aspects of model misspecification.

Keywords: Graphical models, ℓ_1-regularization, Lyapunov equation, support recovery

1. Introduction

Directed graphical models are powerful tools for exploring cause-effect relationships in multivariate data (Pearl, 2009; Spirtes et al., 2000; Peters et al., 2017). The causal aspect of the models is built on the assumption that each variable is a function of parent variables and independent noise. This approach is also known as structural causal modeling. For directed acyclic graphs (DAGs), the resulting models have simple interpretations and statistically favorable density factorization properties that facilitate large-scale analyses (Maathuis et al., 2019). The situation is more complicated when the graph is allowed to contain directed cycles, which represent feedback loops (Bongers et al., 2021). Although a model can still be defined by solving structural equations, directed cycles prevent density factorization, making it more difficult to perform tasks such as computation of maximum likelihood estimates (Drton et al., 2019) or model selection (S. Richardson, 1996; Améndola et al., 2020), even in the case of linear models. Importantly, the interpretation of the models also becomes more involved and typically appeals to dynamic processes in a post-hoc way. For example,

Fisher (1970) provided an interpretation based on data averaged over time, while alternative interpretations in terms of differential equations were suggested by Mooij et al. (2013) and Bongers and Mooij (2018).

Fitch (2019) and Varando and Hansen (2020) proposed a different perspective, in which the distribution of an observed sample $X_1, \dots, X_n \in \mathbb{R}^p$, independent and identically distributed, is modeled through a temporal process in equilibrium. Specifically, each random vector X_i is assumed to be a single cross-sectional observation of a multivariate Ornstein-Uhlenbeck process (a multivariate continuous-time autoregressive process), which leads to X_i being multivariate normal. We emphasize that the n observations are obtained from n independent processes. This setup is suitable, in particular, for applications in biology where each of the n independent organisms may live up to the time of measurement (e.g., for gene expression) but is sacrificed for the measurement.

The p-dimensional Ornstein-Uhlenbeck process is the solution to the stochastic differential equation

$$\mathrm{d}X(t) = M(X(t) - a)\,\mathrm{d}t + D\,\mathrm{d}W(t), \tag{1.1}$$

where $W(t)$ is a Wiener process, and $a \in \mathbb{R}^p$ and $M, D \in \mathbb{R}^{p \times p}$ are non-singular parameter matrices. The *drift matrix* M is the key object of interest in the work of Fitch (2019) and Varando and Hansen (2020) as it determines the causal relations between the coordinates of the Ornstein-Uhlenbeck process $X(t)$; see also Mogensen et al. (2018). Provided M is stable (i.e., all eigenvalues have a strictly negative real part), $X(t)$ admits an equilibrium distribution that is multivariate normal with a positive definite covariance matrix. This covariance matrix, denoted by Σ, is determined as the unique matrix that solves the Lyapunov equation

$$M\Sigma + \Sigma M^\top + C = 0, \tag{1.2}$$

where $C = DD^\top$. The vector a is the mean vector of the equilibrium distribution. Subsequently, we will assume without loss of generality that $a = 0$, i.e., our observations are centered. Moreover, we will focus on the case where the positive definite volatility matrix C is known up to a scalar multiple. Since scaling (M, C) does not change the solution Σ in (1.2), this case can be studied by reducing to the setting where C is fully known; see Remark B.1 for a further detailed discussion.

Our interest is now in the selection of models that postulate that the drift matrix $M = (M_{ij})$ is sparse. In other words, we consider the estimation of the sparsity pattern (or support) of the drift matrix M. This support is naturally represented by a directed graph $G = (V, E)$ with a vertex set $V = \{1, \dots, p\}$ and an edge set E that includes the edge $i \to j$ precisely when $M_{ji} \neq 0$. A stable matrix M will have negative diagonal entries, and therefore the edge set E will always contain all self-loops $i \to i$. However, we will not draw the self-loops in figures showing graphs.

Example 1 *The graph $G = (\{1, 2, 3\}, \{1 \to 2, 2 \to 3, 1 \to 1, 2 \to 2, 3 \to 3\})$, shown in Figure 1, corresponds to the support of the matrix*

$$M = \begin{pmatrix} m_{11} & 0 & 0 \\ m_{21} & m_{22} & 0 \\ 0 & m_{32} & m_{33} \end{pmatrix}.$$

We remark that Young et al. (2019) considered a related setup with discrete-time vector autoregressive (VAR) processes, which leads to the discrete Lyapunov equation.

Figure 1: Directed graph on 3 nodes.

1.1. Support Recovery with the Direct Lyapunov Lasso

In this paper, we will study an ℓ_1-regularization method to estimate the support of the drift matrix M from an i.i.d. sample consisting of centered observations $X_1, \ldots, X_n \in \mathbb{R}^p$. Let

$$\hat{\Sigma} = \hat{\Sigma}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \tag{1.3}$$

be the sample covariance matrix. The *Direct Lyapunov Lasso* finds a sparse estimate of M as a solution of the convex optimization problem

$$\min_{M \in \mathbb{R}^{p \times p}} \frac{1}{2} \| M\hat{\Sigma} + \hat{\Sigma}M^\top + C \|_F^2 + \lambda \|M\|_1 \tag{1.4}$$

with tuning parameter $\lambda > 0$. This method is considered in numerical experiments by Fitch (2019) as well as by Varando and Hansen (2020) who additionally explore non-convex methods based on regularizing Gaussian likelihood or a Frobenius loss. The Direct Lyapunov Lasso yields matrices in $\mathbb{R}^{p \times p}$ that can be non-stable. If a stable estimate is required in such a case, one can appeal to projection onto the set of stable matrices; e.g., using techniques by Noferini and Poloni (2021).

1.2. Organization of the Paper

We connect the Direct Lyapunov Lasso to more standard lasso problems by vectorizing the Lyapunov equation and describing the structure of the Hessian matrix for the smooth part of the Direct Lyapunov Lasso objective (Section 2). In Section 3, we present a result of statistical consistency (Theorem 2), where the solution $\hat{M}$ of (1.4) is shown to converge in the max norm at a rate $\|\hat{M} - M^*\|_\infty = O(\sqrt{(dp)/n})$ with d being the number of nonzero entries in the true drift matrix M^*. Theorem 2 only holds under an irrepresentability condition, which turns out to be more subtle than in the classical lasso regression. As we explore in Section 4, the condition is highly dependent on the structure of the graph associated with the true signal. In Section 5, we present large-scale simulations (up to $p = 50$) where the performance of the Direct Lyapunov Lasso on synthetic data is measured. Despite the theoretical requirements for consistency not being met, we observe performance at useful levels across various metrics. Finally, in Section 6, we apply the Direct Lyapunov Lasso to obtain an estimate of a protein signaling network that recovers important connections in a network that was previously reported as a gold standard.

1.3. Motivating Example

Before developing a detailed analysis of the Direct Lyapunov Lasso, we present an example that illustrates the behavior of estimates for growing sample size and highlights the impact of the irrepresentability condition. We defer some of the details of how the example is designed to Appendix A.
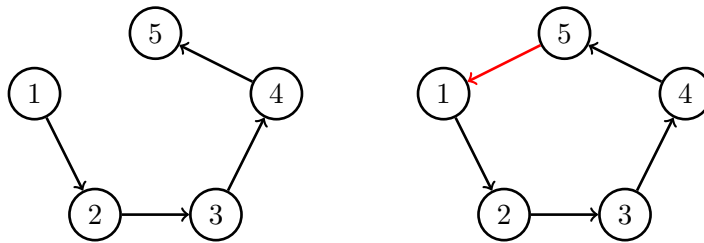
Figure 2: Left: The graph G_1, a path 1 to 5. Right: The graph G_2, the 5-cycle.

Example 2 *Let G_1 be the directed path from 1 to 5, and let G_2 be the 5-cycle obtained by adding the edge $5 \to 1$; see Figure 2. For G_1 we define a (well-conditioned) stable matrix M_1^* by setting the diagonal to $(-2, -3, -4, -5, -6)$ and the four nonzero subdiagonal entries to 0.65. For G_2, we consider two cases. First, we add the fixed entry $m_{15} = 0.65$ to M_1^* to obtain the matrix M_2^*. Second, we similarly include m_{15} but select it randomly (uniform on $[0.5, 1]$) in 100 instances. Using the Lyapunov equation (1.2) with $C = 2I_p$, for each setting and each one of 6 different sample sizes n we simulate 100 Gaussian datasets. We also consider $n = \infty$, i.e., taking the population covariance matrices as input to the method. We calculate solutions to the Direct Lyapunov Lasso (1.4) for 100 choices of the regularization parameter λ. From these solutions we compute the maximum accuracy, the maximum F_1-score, and the area under the ROC curve. Figure 3 plots the performance measures, averaged over the 100 datasets in each pairing of setup and sample size. There the blue curves refer to G_1, the red curves to G_2 with $m_{15} = 0.65$ fixed, and the green curves to G_2 with m_{15} chosen randomly. We observe that for every sample size and performance measure, the Direct Lyapunov Lasso performs better for the path G_1 than for the cycle G_2. When the sample size is $n = 10^4$, we observe an almost perfect recovery of G_1. However, increasing the sample size when recovering G_2 does not result in perfect recovery. The choice of $m_{15} = 0.65$ is not particularly unfortunate—averaging over various completions does not improve the metrics. We conclude that while learning useful structure in either case, the Direct Lyapunov Lasso is consistent only for the considered path. Our subsequent analysis explains this behavior, which is a consequence of the failure of the irrepresentability condition in (3.1).*

1.4. Notation

Let $b \in [1, \infty]$. The ℓ_b-norm of $v \in \mathbb{R}^p$ is $\|v\|_b = (\sum_{i=1}^p |v_i|^b)^{1/b}$, with $\|v\|_\infty = \max_{1 \le i \le n} |v_i|$. We may apply this vector norm to a matrix $A = (a_{ij}) \in \mathbb{R}^{p \times p}$ and obtain the norm $\|A\|_b = (\sum_{i=1}^p \sum_{j=1}^n |a_{ij}|^b)^{1/b}$. In particular, $\|A\|_F := \|A\|_2$ is the Frobenius norm. We denote the associated operator norm by $\|A\|_b = \max\{\|Ax\|_b : \|x\|_b = 1\}$. Specifically, we use $\|A\|_2$ to denote the spectral norm, given by the maximal singular value of A, and $\|A\|_\infty = \max_{1 \le i \le p} \sum_{j=1}^p |a_{ij}|$ to denote the maximum absolute row sum.

For an index set S, we write $A_{.S}$ for the submatrix of $A = (a_{ij})$ that is obtained by selecting the columns indexed by S. The matrices $A_{S.}$ and A_{SS} are defined analogously by selection of rows or both rows and columns, respectively. The vec-operator stacks the columns of A, giving the vector $\mathrm{vec}(A) = (a_{11}, a_{21}, \ldots, a_{p1}, \ldots, a_{1p}, \ldots, a_{pp})^\top \in \mathbb{R}^{p^2}$. The diag-operator turns a vector $v \in \mathbb{R}^p$ into the diagonal matrix $\mathrm{diag}(v) \in \mathbb{R}^{p \times p}$ that has v_i as its i-th diagonal entry. The Kronecker product of A and another matrix $B = (b_{uv}) \in \mathbb{R}^{p \times p}$ is denoted by $A \otimes B$. It is a matrix in $\mathbb{R}^{p^2 \times p^2}$
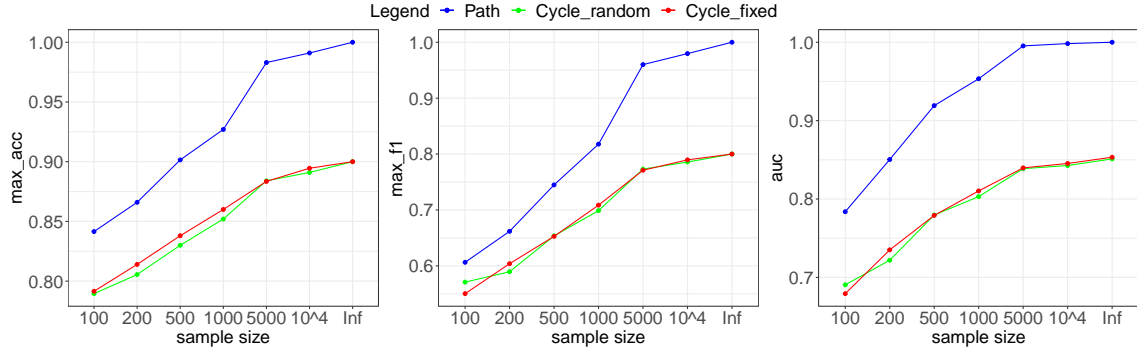
Figure 3: Performance measures for sample sizes $n = 10^1, \ldots, 10^4, \infty$ and models given by the graphs G_1 and G_2 from Figure 2, one choice of edgeweights for the path, one choice of edgeweights for the cycle (Cycle fixed) and 100 random completions to the 5-cycle (Cycle random). Left: maximal accuracy, Middle: maximal F_1-score, Right: area under the ROC curve.

with the entries $[A \otimes B]_{(i-1)p+j,(k-1)p+l} = A_{ik}B_{jl}$. The commutation matrix in $\mathbb{R}^{p^2 \times p^2}$ is the permutation matrix $K^{(p,p)}$ such that $K^{(p,p)}\mathrm{vec}(A) = \mathrm{vec}(A^\top)$. See, e.g., Bernstein (2009).

Finally, we write Sym_p for the space of symmetric matrices in $\mathbb{R}^{p \times p}$. We use PD_p to denote the cone of $p \times p$ positive definite matrices. The set of stable matrices in $\mathbb{R}^{p \times p}$ is denoted Stab_p.

2. Gram Matrix of the Direct Lyapunov Lasso

In this section, we rewrite the smooth part of the objective of the Direct Lyapunov Lasso from (1.4) in terms of the vectorized drift matrix and present the resulting Hessian matrix.

The Lyapunov equation from (1.2) is a linear matrix equation and may be rewritten as

$$A(\Sigma)\,\mathrm{vec}(M) + \mathrm{vec}(C) = 0, \tag{2.1}$$

where the $p^2 \times p^2$ matrix $A(\Sigma)$ has its rows and columns indexed by pairs $(i,j) \in \{1, \ldots, p\}^2$ and takes the form

$$A(\Sigma) = (\Sigma \otimes I_p) + (I_p \otimes \Sigma)K^{(p,p)}. \tag{2.2}$$

We have $\mathrm{vec}(M\Sigma) = (\Sigma \otimes I_p)\mathrm{vec}(M)$ and $\mathrm{vec}(\Sigma M^\top) = (I_p \otimes \Sigma)K^{(p,p)}\mathrm{vec}(M)$. By the symmetry of the Lyapunov equation, $A(\Sigma)$ has two copies of each row corresponding to an off-diagonal entry in the Lyapunov equation. Retaining this redundancy will be helpful for later arguments, as it preserves the Kronecker product structure in (2.2). A display of the matrix $A(\Sigma)$ is provided in Example C.1 in the appendix. Define the *Gram matrix*

$$\Gamma(\Sigma) := A(\Sigma)^\top A(\Sigma) \ \in \mathbb{R}^{p^2 \times p^2} \tag{2.3}$$

and the vector

$$g(\Sigma) := -A(\Sigma)\mathrm{vec}(C) \ \in \mathbb{R}^{p^2}. \tag{2.4}$$

Omitting a constant from the objective function, the Direct Lyapunov Lasso problem from (1.4) may be reformulated as

$$\min_{M \in \mathbb{R}^{p \times p}} \frac{1}{2}\,\mathrm{vec}(M)^\top \Gamma(\hat{\Sigma})\mathrm{vec}(M) - g(\hat{\Sigma})^\top \mathrm{vec}(M) + \lambda \|\mathrm{vec}(M)\|_1. \tag{2.5}$$

As noted in the introduction, one difficulty that arises in the analysis of the solution of (2.5) is the fact that the Gram matrix has entries that are quadratic polynomials in Σ with p terms (i.e., the number of terms scales with the size of the problem). This fact can be seen in the appearance of Σ^2 in the following formula for the Gram matrix.

Lemma 1 *The Gram matrix for a given covariance matrix Σ is equal to*

$$\Gamma(\Sigma) = A(\Sigma)^\top A(\Sigma) = 2(\Sigma^2 \otimes I_p) + (\Sigma \otimes \Sigma)K^{(p,p)} + K^{(p,p)}(\Sigma \otimes \Sigma).$$

Proof Apply the rules $(A \otimes B)^\top = (A^\top \otimes B^\top)$, $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and $K^{(p,p)}(A \otimes B)K^{(p,p)} = B \otimes A$ to deduce that

$$\begin{aligned} A(\Sigma)^\top A(\Sigma) &= [(\Sigma \otimes I_p) + K^{(p,p)}(I_p \otimes \Sigma)][(\Sigma \otimes I_p) + (I_p \otimes \Sigma)K^{(p,p)}] \\ &= 2(\Sigma^2 \otimes I_p) + (\Sigma \otimes \Sigma)K^{(p,p)} + K^{(p,p)}(\Sigma \otimes \Sigma). \end{aligned}$$

■

3. Consistent Support Recovery with the Direct Lyapunov Lasso

We now provide a probabilistic guarantee that the Direct Lyapunov Lasso is able to recover the support of the true population drift matrix that defines the data-generating distribution. Let M^* denote the true value of the drift matrix in (1.1), and let Σ^* be the associated true covariance matrix of the observations. We write $\hat{M}$ for the solution of the Direct Lyapunov Lasso problem in (1.4). The support of M^* is the set of all indices of nonzero elements and is denoted by

$$S \equiv S(M^*) = \{(j,k) : M^*_{jk} \neq 0\}.$$

We write $d = |S|$ for the size of the support of M^*. The support set of the estimate $\hat{M}$ is

$$\hat{S} \equiv S(\hat{M}) = \{(j,k) : \hat{M}_{jk} \neq 0\}.$$

Let $\hat{\Gamma} = \Gamma(\hat{\Sigma})$, $\Gamma^* = \Gamma(\Sigma^*)$, $\hat{g} = g(\hat{\Sigma})$, $g^* = g(\Sigma^*)$. Furthermore, let $\Delta_\Gamma = \hat{\Gamma} - \Gamma^*$ and $\Delta_g = \hat{g} - g^*$, and define the quantities

$$c_{\Gamma^*} = \|(\Gamma^*_{SS})^{-1}\|_\infty \quad \text{and} \quad c_{M^*} = \|\text{vec}(M^*)\|_\infty.$$

The definition of c_{Γ^*} requires Γ^*_{SS} to be invertible, which is an implicit assumption on the identifiability of the parameters; see Remark B.2 in the Appendix. By suitably adapting work on structure learning for undirected graphical models (Lin et al., 2016), one can derive a deterministic guarantee for success of the Direct Lyapunov Lasso provided the estimation errors Δ_Γ and Δ_g are sufficiently small (Theorem D.1 in the Appendix). This result leads to the following probabilistic result.

Theorem 2 *Suppose that the data are generated as n i.i.d. draws from the Gaussian equilibrium distribution of a p-dimensional Ornstein-Uhlenbeck process defined by a drift matrix $M^* \in Stab_p$ and a matrix $C \in \text{PD}_p$. Let S be the support of M^*. Assume that Γ^*_{SS} is invertible and that the irrepresentability condition*

$$\|\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}\|_\infty < 1 - \alpha \tag{3.1}$$

holds for $\alpha \in (0, 1]$. Let $c_{\Sigma^*} = \|\|\Sigma^*\|\|_2$, $c_C = \|vec(C)\|_2$,

$$\tilde{c} = \max \left\{ \frac{4 \max\{1, c_{\Sigma^*}^2\}(4 + 8c_{\Sigma^*})^2}{c_3}, 16c_1^2 c_{\Sigma^*}^2 (4 + 8c_{\Sigma^*})^2, \frac{16 \max\{1, c_{\Sigma^*}^2\}c_C^2}{c_3}, 64c_1^2 c_{\Sigma^*}^2 c_C^2 \right\},$$

$$c_* = \frac{6}{\alpha} c_{\Gamma^*},$$

where $\{c_i\}_{j=1}^3$ are universal constants (from Theorem E.4 below) with $c_1 > \max\{1, \|\|\Sigma^*\|\|_2\}$. Suppose the sample size satisfies $n > \tau_1 \tilde{c} dp \max\{c_*^2, 1/4\}$ for $\tau_1 > 1$, and the regularization parameter is chosen as

$$\lambda > \frac{3c_{M^*}(2 - \alpha)}{\alpha} \sqrt{\frac{\tau_1 \tilde{c} dp}{n}}.$$

Then the following statements hold with probability at least $1 - c_2 \exp\left(-\tau_1 p\right)$:

a) *The minimizer $\hat{M}$ is unique, has its support included in the true support ($\hat{S} \subset S$), and satisfies*

$$\|\hat{M} - M^*\|_\infty < \frac{2c_{\Gamma^*}}{2 - \alpha} \lambda.$$

b) *Furthermore, if*

$$\min_{\substack{1 \le j < k \le m \\ (j,k) \in S}} |M_{jk}^*| > \frac{2c_{\Gamma^*}}{2 - \alpha} \lambda,$$

then $\hat{S} = S$ and $\text{sign}(\hat{M}_{jk}) = \text{sign}(M_{jk}^*)$ for all $(j, k) \in S$.

The reader may be surprised by the sample size requirement of $n = \Omega(dp)$; recall that $d = |S|$ is the size of the support of M^*. Since S includes the diagonal of M^*, we have $d \ge p$. Under sparsity, however, dp is not much larger than the number of unknown parameters p^2.

This said, $\Omega(dp)$ is far larger than the sample size requirement a reader may be familiar with from the glasso for learning undirected conditional independence, which is on the order of $d^2 \log p$ but with d being the maximum number of nonzero entries in any row of a true precision matrix (Ravikumar et al., 2011). This allows for far higher-dimensional settings, but crucially relies on the glasso having a Hessian that concentrates well entry-wise and a simple connection between the covariance matrix and the sparse precision matrix. In contrast, the Lyapunov Lasso has a denser Hessian/Gram matrix that includes entries that become heavier-tailed as the dimension p grows.

In order to prove Theorem 2, we combine the aforementioned deterministic analysis, stated in Theorem D.1, with the concentration results we obtain in Section E. We defer the detailed proof to Appendix F. Here, we present only a short sketch.

Proof [Sketch] Applying Theorem D.1 to obtain the probabilistic result in Theorem 2 requires showing that a probability of the form $\mathbb{P}(\|\|(\Gamma(\hat{\Sigma}) - \Gamma(\Sigma^*))_{.S}\|\|_\infty \ge \epsilon_1)$ is small when the sample size is sufficiently large. Representing Hessian Γ as Kronecker products of the covariance matrix Σ, Lemma 1 permits to trace back the problem of deriving a concentration inequality for Γ to finding concentration inequalities for Σ. Ultimately, this connection is made in Lemma E.6 securing that if

$$\|\Delta_\Sigma\|_2 = \|\|\hat{\Sigma} - \Sigma^*\|\|_2 < \min \left\{ \frac{\epsilon_1}{\sqrt{d}(4 + 8c_{\Sigma^*})}, \frac{\epsilon_2}{2c_C} \right\},$$

then it holds that

$$\|(\Delta_\Gamma)._S\|_\infty < \epsilon_1 \quad \text{and} \quad \|\Delta_g\|_\infty < \epsilon_2.$$

A known concentration result for Σ is given in Theorem E.4. Combining this with Lemma E.6 and using the concentration result for Γ in Theorem D.1 yields Theorem 2. ∎

Sample size aside, the crucial assumption for support recovery is the irrepresentability condition. A detailed analysis of this condition is the subject of the next section.

4. Irrepresentability Condition

The irrepresentability condition is vital for Theorem 2. The condition is well known from the standard lasso regression, but turns out to be much more subtle for the Direct Lyapunov Lasso. In regression and in the Lyapunov model, the irrepresentability condition makes an assumption about the Gram matrix in light of the signal. However, the Gram matrix in regression depends solely on the predictors, whereas the Gram matrix for the Lyapunov model is obtained from the matrix $A(\Sigma)$ which depends on the signal itself (Example C.1).

We present an analysis of irrepresentability for the Direct Lyapunov Lasso under weak dependence and find that the condition is very restrictive. Indeed, it leads to a restrictive ordering condition on the diagonal of the drift matrix; in particular, the support must define a DAG. For cyclic graphs, it seems difficult to construct general examples of irrepresentability. Simulations suggest that such examples do exist, but are rare. We refer to Appendix G for details on cyclic graphs as well as a discussion of a weaker notion of irrepresentability that is necessary for consistency.

In our study of the irrepresentability condition, we will consider the case where the volatility matrix C is a multiple of identity; specifically, we assume $C = 2I_p$ *throughout this section*. Other diagonal matrices C would also be tractable for analysis and would yield analogous conclusions. Before proceeding, we recall that a matrix $M^* \in \text{Stab}_p$ with support S satisfies the irrepresentability condition if

$$\rho(M^*) := \|\Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}\|_\infty \tag{4.1}$$

is strictly smaller than 1; the condition in (3.1) stated an explicit gap $\alpha > 0$. In the following, we will refer to the number $\rho(M^*)$ as the *irrepresentability constant* of M^*. When drawing an analogy to regression problems, Γ^* corresponds to a covariance matrix of predictors and the irrepresentability constant captures how well non-signals (here data-derived features that correspond to non-edges) can be predicted from signals (here data-derived features that correspond to edges). In standard lasso regression, the irrepresentability condition is fulfilled when each irrelevant predictor exhibits little correlation with the active predictors. In particular, the condition would hold in a neighborhood of a diagonal Gram matrix.

Example 3 *Consider the graph $G = (V, E)$ in Figure 1, a path on 3 nodes. For small $e \in \mathbb{R}$, we define two stable matrices $M_1(e)$ and $M_2(e)$ with support given by G. We set their diagonals to $\text{diag}(M_1(e)) = (-1/2, -1, -3/2)$ and $\text{diag}(M_2(e)) = (-3/2, -1, -1/2)$, respectively, and set all nonzero off-diagonal entries equal to e. Note that the diagonal of $M_2(e)$ is the reverse of the diagonal of $M_1(e)$. In Figure 4, we plot the two irrepresentability constants $\rho(M_1(e))$ and $\rho(M_2(e))$ as functions of the off-diagonal value e. We observe that irrepresentability holds in a neighborhood of $M_1(0)$, but not around $M_2(0)$.*
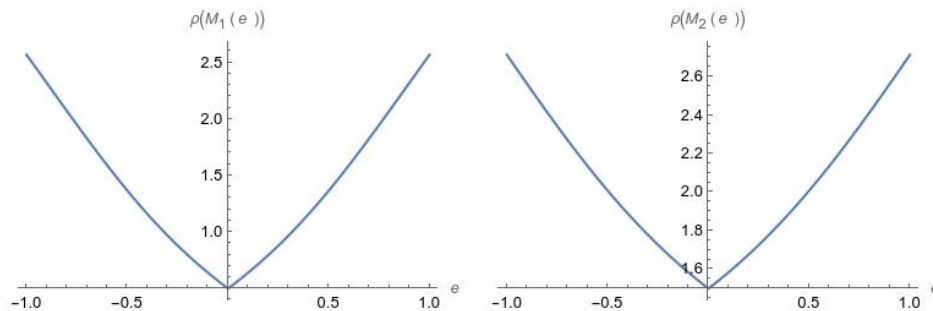
Figure 4: Values of the irrepresentability constants $\rho(M_1(e))$ and $\rho(M_2(e))$ for the two matrices from Figure 1 plotted against the size of the off-diagonal entries e. Left: $\rho(M_1(e))$ where $\mathrm{diag}(M_1(e)) = (-1/2, -1, -3/2)$. Right: $\rho(M_2(e))$ where $\mathrm{diag}(M_2(e)) = (-3/2, -1, -1/2)$.

In the example, the order of diagonal entries is seen to impact whether irrepresentability holds near a diagonal matrix. As we prove in the theorem below, this fact is not a coincidence, but rather a general phenomenon. Let $S \subseteq \{(i,j) : 1 \le i, j \le p\}$ be a given support set. We say that the irrepresentability condition for the support S holds uniformly over a set $U \subset \mathrm{Stab}_p$ if there exists $\alpha > 0$ such that $\rho(M^*) \le 1 - \alpha$ for all $M^* \in U$ with support $S(M^*) = S$. By our convention, the edge set of a directed graph $G = (V, E)$ determines the support set $S_G = \{(j, i) : i \to j \in E\}$.

Theorem 3 *Let $G = (V, E)$ be a graph with p nodes. Let $M^0 = \mathrm{diag}(-d_1, \ldots, -d_p)$ be a stable diagonal matrix. Then, the irrepresentability condition for support S_G holds uniformly over a neighborhood of M^0 if and only if*

$$d_i < d_j \ \text{for every edge}\ i \to j \in E.$$

In particular, it is necessary that the graph G is a DAG.

We give a proof and an illustrating example in Appendix G.1.

5. Simulation Studies

In this section, we present simulation studies that provide insight into the performance of the Direct Lyapunov Lasso in seemingly unfavorable settings. First, most drift matrices do not satisfy the irrepresentability condition; compare Section G.3 in the Appendix. Second, while our assumption that C is fixed up to a scalar multiple is made similarly in the related case where actual time series data is considered (Gaïffas and Matulewicz, 2019), it is an assumption that may be overly simple for many applications. Nevertheless, our simulations suggest robustness of the Direct Lyapunov Lasso to the irrepresentability condition not being fulfilled and to mild misspecification of the volatility matrix C, where by robustness we mean that a part of signal is being learned correctly.

For the simulations in this section, we use a similar setting as in Varando and Hansen (2020). Each stable matrix M was generated with $M_{ij} = \omega_{ij}\epsilon_{ij}$ for $i \ne j$ and $M_{ii} = -\sum_{j\ne i}|M_{ij}| - |\epsilon_{ii}|$ where $\omega_{ij} \sim \mathrm{Bernoulli}(d)$ and $\epsilon_{ij} \sim N(0, 1)$. Unlike in Varando and Hansen (2020), we consider four different choices for C. The label in brackets corresponds to the one used in Figure 5.

1) We choose $C = 2I_p$ (C_ID).

2) We choose C diagonal with $C_{ii} \sim \text{Unif}[0.5, 4]$ (C_Random_Diag).

3) We choose C diagonal with $C_{ii} \sim \text{Unif}[2, 4]$ (C_Random_Min_Diag).

4) We choose C symmetric but non-diagonal. Let $\tilde{\omega}_{ij} \sim \text{Bernoulli}(2/p)$ and $\tilde{\epsilon}_{ij} \sim N(0, 1)$ be independent random variables, $i, j = 1, \ldots, p$. Then the off-diagonal entries of C are set to $C_{ij} = \tilde{\omega}_{ij}\tilde{\epsilon}_{ij} + \tilde{\omega}_{ji}\tilde{\epsilon}_{ji}$ and the diagonal entries to $C_{ii} = \sum_{j \neq i} |C_{ij}| + |\tilde{\epsilon}_{ii}| + 0.5$ (C_Random_Full).

For each $k \in \{1, 2, 3, 4\}$ and $p = \{10, 15, 20, 25, 30, 40, 50\}$, the edge probability is set as $d = k/p$. For each choice of C, we generate 100 pairs of signals (M, C). We generate $N = 1000$ observations from a multivariate Gaussian distribution with covariance matrix solving the Lyapunov equation for (M, C). Note that $p^2 > N$ for $p = \{40, 50\}$ which corresponds to the high-dimensional setting. Then we apply the Direct Lyapunov Lasso with $C = 2I_p$ for model selection. The results are calculated along the λ-grid:

$$0 < \frac{\lambda_{\max}}{10^4} = \lambda_1 < \cdots < \lambda_{100} = \lambda_{\max},$$

with $\lambda_{\max}$ being the minimal λ-value such that M is diagonal. We compute the maximum accuracy, the maximum F_1-score, the area under the ROC curve and the area under the precision curve; more details on the metrics are given in Definition G.5. The metrics are averaged over the 4 different sparsity levels k and the 100 randomly selected drift matrices M. The results are shown in Figure 5.

Choice 1) for C is used when applying Direct Lyapunov Lasso for model selection. Thus, it is natural to expect the best results for this choice. Choices 2) and 3) allow for variability on the diagonal. The second choice allows for larger differences $(\text{Unif}[0.5, 4])$ in the size of the diagonal entries, while the third choice is more conservative $(\text{Unif}[2, 4])$. Choices 1) and 3) perform best in our simulations. We observe that there are few differences in all metrics among the choices between choice 1) and choice 3), indicating that the Direct Lyapunov Lasso with $C = 2I_p$ possesses a certain robustness to the exact diagonal matrix C of the data generating model. This is true for all $p \in \{10, 15, 20, 25, 30, 40, 50\}$. For choice 2), we observe that the results in all metrics except maximum accuracy fall with increasing p. For $p = 40$ and especially for $p = 50$, the results for all metrics are similar to choice 4). Choice 4) allows for data generating models for which C is no longer diagonal. For this choice, the worst results are to be expected as the matrix C used for data generation is much different from the one used for estimation. Another interesting point revealed by the simulations is that although the irrepresentability condition is not satisfied in almost any of the signals, it is still possible to get estimates that recover much of the support of the drift matrix.
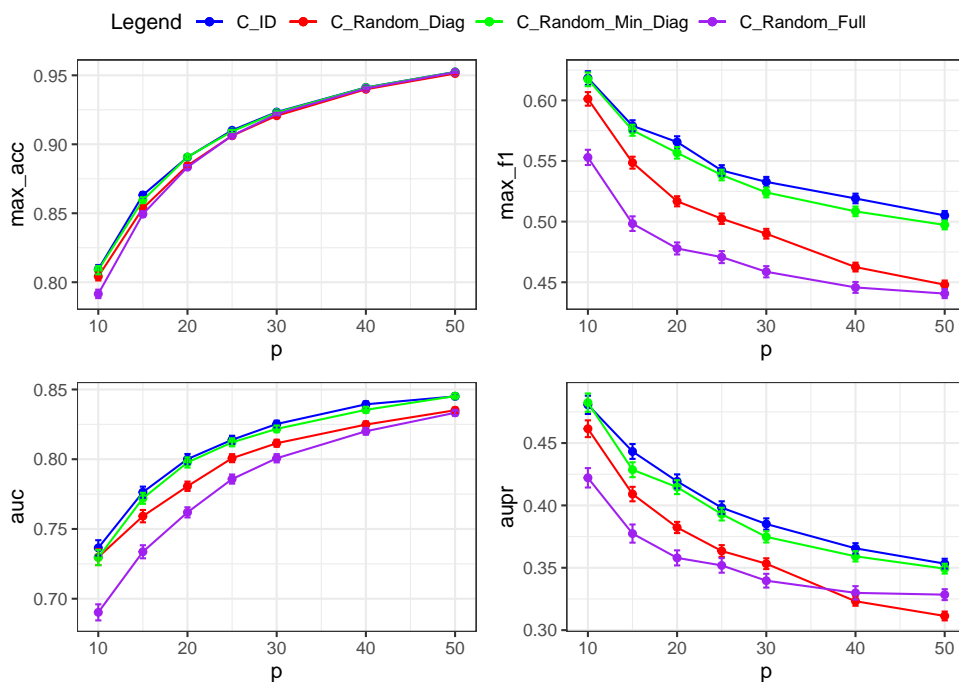
Figure 5: The maximum accuracy (top left), maximum F_1-score (top right), area under the ROC curve (bottom left) and area under the precision curve (bottom right) in support recovery with the Direct Lyapunov Lasso using parameter $C = 2I_p$. The data has been generated using the choices 1) C_ID, 2) C_Random_Diag, 3) C_Random_Min_Diag, and 4) C_Random_Full. The error bars are the estimated standard errors of the average of a metric for a specific problem size over the 400 randomly generated drift matrices.

6. Real World Data Example

The data set collected and analyzed first by Sachs et al. (2005) has become a test bed for graphical model selection algorithms. Some recent examples are the review paper on causal discovery methods based on graphical models by Glymour et al. (2019), the application of classical structure equation models allowing for cycles by Améndola et al. (2020) or the application of Lyapunov models and a specific model selection technique by Varando and Hansen (2020). The data set consists of flow cytometry measurements of 11 phosphorylated proteins and phospholipids in human T-cells captured under different experimental conditions, resulting in 14 independent data sets of varying sizes ($n = 707$ to $n = 927$). When flow cytometry is applied, cells are destroyed during the measurement process, and hence the measurements are collected at one point in time. Each sample consists of quantitative and simultaneous measurements of the 11 phosphorylated proteins and phospholipids of single cells. The way in which the data set was collected matches the motivation underlying the Lyapunov models. In presenting the experiments below, our goal is to exemplify the fact that Lyapunov models concern cross-sectional data and not time series and to showcase the use of the models in a widely studied data example. However, the Gaussianity of Lyapunov models (as well as other aforementioned models) is clearly an oversimplification for the considered flow cytometry data.

We apply the Direct Lyapunov Lasso (1.4) with $C = 2I_p$ and with an adaptation of the extended Bayesian Information Criterion (BIC) criterion for tuning parameter selection (Chen and Chen, 2008; Foygel and Drton, 2010). First, we standardize every column of the data by calculating $X_{.,i}^{std} = (X_{.,i} - \mu)/\sigma$ where μ is the mean and σ the standard deviation of $X_{.,i}$. We apply the Direct Lyapunov Lasso to obtain estimates along the regularization path that is the logarithmic sequence

$$0 < \frac{\lambda_{\max}}{10^4} = \lambda_1 < \cdots < \lambda_{100} = \lambda_{\max},$$

where $\lambda_{\max}$ is the minimal λ-value such that the estimate is diagonal. Extracting the non-zero structure, each estimate $\hat{M}_j$ defines a directed graph G_j for $j = 1, \ldots, 100$. To decide which graph to select, we use the extended BIC. We first minimize the two times negative Gaussian log-likelihood

$$L(M) = n\big[\log \det\big(\Sigma(M, 2I_p)\big) + \operatorname{tr}\big(\hat{\Sigma}(\Sigma(M, 2I_p))^{-1}\big)\big] \tag{6.1}$$

for all 100 models obtained by restricting the support of M to G_j, $j = 1, \ldots, 100$. We denote the minima by $\hat{L}_1, \ldots, \hat{L}_{100}$ and substitute these values into

$$EBIC_\gamma(G_j) = (|E_j| + p) \log n + 4\gamma |E_j| \log p + \hat{L}_j, \tag{6.2}$$

where E_j is the edge set of G_j. Selecting the graph with the lowest score yields the Lyapunov model selected by the extended BIC criterion.

In Figure 6a) we present the graph estimate for the protein signaling network using data set 7. A graph that shows conventionally accepted "ground truth" molecule interactions is given in Figure 6b) and it includes ambiguities for some connections (Ramsey and Andrews, 2018, Section 2). The estimate has 17 edges and 11 connections when counting the 2-cycles only once, while the ground truth has 20 edges with no 2-cycles. The red edges are the edges that are correctly recovered by the estimate. The orange edges are the edges where the reversed edges are present in the ground truth. Black edges are additional edges that are not present in the ground truth. Among the correctly estimated connections are the direct enzyme-substrate relationships Raf $\rightarrow$ Mek and PIP3 $\rightarrow$ Pcl$-\gamma$. The connections Pcl$-\gamma \rightarrow$ PIP2 and PKA $\rightarrow$ Raf are missing in Figure 6, but the Pcl$-\gamma \rightarrow$ PIP3 $\rightarrow$ PIP2 and Pcl$-\gamma \rightarrow$ PIP3 $\rightarrow$ PIP2 pathways suggest the presence of these interactions. Only the connection Mek $\rightarrow$ Erk is not present at all. In general, Direct Lyapunov Lasso with extended BIC is an intuitive and easy-to-implement method that produces a sparse estimate with most edges (or their reverse) present in the ground truth and even some additional edges such as Akt $\rightarrow$ Raf can be interpreted as connecting pieces of meaningful pathways.

7. Conclusion

We investigated the model selection properties of the Direct Lyapunov Lasso when applied to data distributed according to the graphical continuous Lyapunov model. Although the optimization problem that the Direct Lyapunov Lasso solves is similar to the lasso-penalized linear regression objective, there are several surprising differences. We established a reasonable bound on the sample complexity by carefully investigating the Hessian matrix whose elements are sums of p products of covariances. The irrepresentability condition is more subtle under the Lyapunov model than it is in the linear regression setting. We formulated conditions under which the irrepresentability condition is guaranteed to hold for DAGs based on the topological ordering of the nodes. Despite
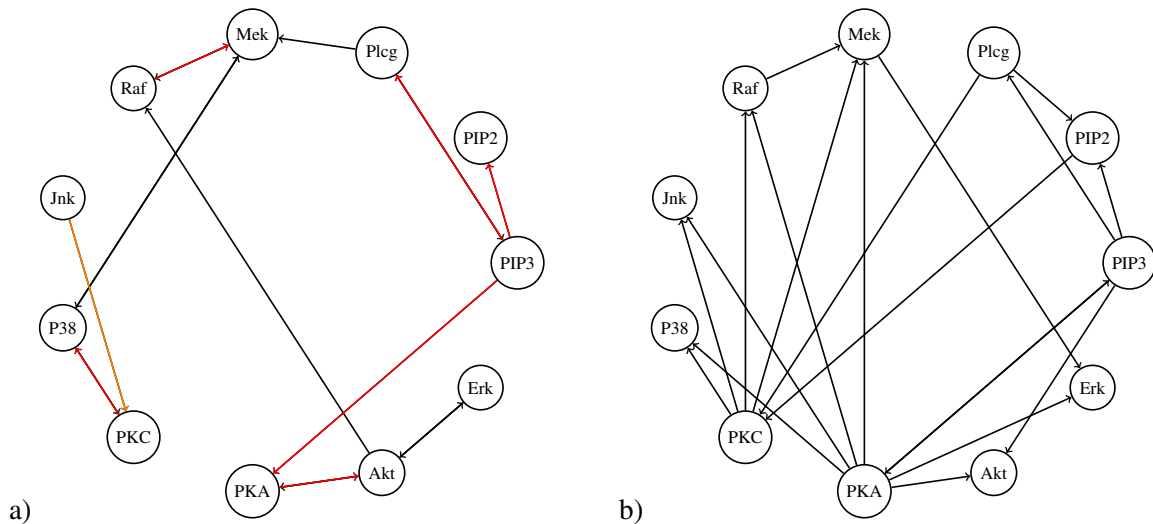
Figure 6: a) Estimated Sachs Network using the Direct Lyapunov Lasso and the EBIC criterion with $\gamma = 1$ for scoring (Dataset 7). b) Ground truth (consensus) network of Sachs et al. (2005).

the irrepresentability condition rarely being fulfilled for randomly selected drift matrices and the problem of misspecification of the volatility matrix when applying the Direct Lyapunov Lasso, we showed that the method is rather robust and is able to detect key features of sparse structures also in seemingly unfavorable settings. Similarly, the combination of Direct Lyapunov Lasso and extended BIC is quite intuitive and easy-to-implement, but still manages to recover important structures of a protein-signaling network purely based on observational data.

Acknowledgments

References

Carlos Améndola, Philipp Dettling, Mathias Drton, Federica Onori, and Jun Wu. Structure learning for cyclic linear causal models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 999–1008, 2020.

Dennis S. Bernstein. *Matrix mathematics*. Princeton University Press, Princeton, NJ, second edition, 2009. Theory, facts, and formulas.

Stephan Bongers and Joris M. Mooij. From random differential equations to structural causal models: the stochastic case. *arXiv.org preprint*, 1803.08784, 2018.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *Ann. Statist.*, 49(5):2885–2915, 2021.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Philipp Dettling, Roser Homs, Carlos Améndola, Mathias Drton, and Niels R. Hansen. Identifiability in continuous Lyapunov models. *SIAM J. Matrix Anal. Appl.*, 44(4):1799–1821, 2023.

Mathias Drton, Christopher Fox, and Y. Samuel Wang. Computation of maximum likelihood estimates in cyclic structural equation models. *Ann. Statist.*, 47(2):663–690, 2019.

Franklin M. Fisher. A correspondence principle for simultaneous equation models. *Econometrica*, 38(1):73–92, 1970.

Katherine E. Fitch. Learning directed graphical models from Gaussian data. *arXiv.org preprint*, 1906.08050, 2019.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

Stéphane Gaïffas and Gustaw Matulewicz. Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process. *J. Multivariate Anal.*, 169:1–20, 2019.

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

Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854, 2016.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th conference on Uncertainty in Artificial Intelligence (UAI)*, pages 350–360, 2018.

Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: The deterministic case. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 440–448, 2013.

Vanni Noferini and Federico Poloni. Nearest Ω-stable matrix via Riemannian optimization. *Numer. Math.*, 148(4):817–851, 2021.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017.

Joseph Ramsey and Bryan Andrews. Fask with interventional knowledge recovers edges from the sachs model. *arXiv.org preprint*, 1805.03108, 2018.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional co-variance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5: 935–980, 2011.

Thomas S. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 454–461. 1996.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000.

Gherardo Varando and Niels R. Hansen. Graphical continuous Lyapunov models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 989–998, 2020.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5): 2183–2202, 2009.

Martin J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.

Wolfram Research, Inc. Mathematica version 13.2, 2022. Champaign, IL, 2022.

William Chad Young, Ka Yee Yeung, and Adrian E. Raftery. Identifying dynamical time series model parameters from equilibrium samples, with application to gene regulatory networks. *Statistical Modelling*, 19(4):444–465, 2019.

Appendix A. Supplementary Information for Example 2

Here, we give the exact choices of stable drift matrices and the estimation procedure for Example 2.

Let G_1 be the path from 1 to 5, and let G_2 be the 5-cycle obtained by adding the edge $5 \to 1$; see Figure 2. For G_1 we define a (well-conditioned) stable matrix M_1^* by setting the diagonal to $(-2, -3, -4, -5, -6)$ and the four nonzero subdiagonal entries to 0.65. For G_2, we consider two cases. In the first case, we add the entry $m_{15} = 0.65$ to M_1^* to obtain the matrix M_2^*. We then draw 100 samples of size $n = 100, 200, 500, 1000, 5000, 10^4, 10^5, \infty$ from $N(0, \Sigma_j^*)$ for $j = 1, 2$, where Σ_j^* is the covariance matrix obtained from M_j^*. When $n = \infty$, the population covariance matrices are taken as input to the method. In the second case, we generate 100 stable matrices $M_{2,1}^*, \dots, M_{2,100}^*$ from M_1^* by selecting 100 entries m_{15} according to a uniform distribution on $[0.5, 1]$. Let $\Sigma_{2,1}^*, \dots, \Sigma_{2,100}^*$ be the corresponding equilibrium covariance matrices. In the second case, we generate one sample from $N(0, \Sigma_j^*)$, $j = 1, (2, 1), \dots, (2, 100)$ for each of the sample sizes given above. Direct Lyapunov Lasso is used for support recovery, with the penalty parameter λ chosen on a grid $\lambda_1 = \lambda_{\max}/10^4, \dots, \lambda_{100} = \lambda_{\max}$ that is equidistant on the log-scale. The value $\lambda_{\max}$ is the minimal λ-value such that the estimate is diagonal. To implement the Direct Lyapunov Lasso, we use the R package `glmnet`, which runs a coordinate descent algorithm for fitting the Lasso, see Friedman et al. (2010). For each data set, we calculate the maximum accuracy, the maximum F_1-score and the area under the ROC curve. We give the details for the metrics in Definition G.5.

Appendix B. Volatility Matrix and Identifiability

In this section, we provide more insight into the assumption that the volatility matrix C is known. Based on this assumption, the question of parameter identifiability is solved for most graphs. This property of a model is necessary to derive consistency results as we do in this work. We give the basic idea of parameter identifiability and provide a reference.

Remark B.1 *Purely from the covariance matrix Σ, it is not possible to determine whether the true parameter pair is (M, C) or $(\gamma M, \gamma C)$ with $\gamma > 0$, since the Lyapunov equation is scaling invariant. Assuming that C is known is the best we can do when C is unknown up to a multiplicative scalar $\gamma > 0$. This accommodates, in particular, the homoscedastic case with $C = \gamma I_p$, where I_p denotes the identity matrix. From a practical perspective, the case $C = 2I_p$ is the most useful, as it mirrors the equal variance assumption for structural equation models, see Peters and Bühlmann (2014). The assumption that C is known might seem to be restrictive, but it is also made in related work on estimation of the drift matrix for data collected from a single time series Gaïffas and Matulewicz (2019). Moreover, the above-mentioned identifiability theory needs to be further developed to adequately address the case where C is unknown. However, we do think that this should be the subject of further research.*

Remark B.2 *Even after reducing to the case of a known volatility matrix C, the drift matrix M is not identifiable without exploiting further structure, such as sparsity. Indeed, the Lyapunov equation is a symmetric matrix equation with $(p + 1)p/2$ individual equations, whereas M contains p^2 unknown parameters. However, M becomes identifiable when it is known to be suitably sparse. For example, it can be shown that the Lyapunov equation for a given C never has two different lower-triangular solutions. In particular, the matrix M from Example 1 can always be uniquely*

recovered from the Lyapunov equation when its graph/support is known. The work by Dettling et al. (2023) proves that unique recovery of M is always possible for graphs that do not contain cycles of length two. Many graphs with two-cycles permit almost sure unique recovery when the sparse entries of the drift matrix are randomly selected according to a continuous distribution, although here a concise sufficient condition has not been found.

Appendix C. Display of the Matrix $A(\Sigma)$

In this section, we present a display of the matrix $A(\Sigma)$.

Example C.1 *When $p = 3$, the matrix $A(\Sigma)$ is a 9×9 matrix and has the form*

	$(1,1)$	$(2,1)$	$(3,1)$	$(1,2)$	$(2,2)$	$(3,2)$	$(1,3)$	$(2,3)$	$(3,3)$
$(1,1)$	$2\Sigma_{11}$	0	0	$2\Sigma_{12}$	0	0	$2\Sigma_{13}$	0	0
$(1,2)$	Σ_{21}	Σ_{11}	0	Σ_{22}	Σ_{12}	0	Σ_{23}	Σ_{13}	0
$(1,3)$	Σ_{31}	0	Σ_{11}	Σ_{23}	0	Σ_{12}	Σ_{33}	0	Σ_{13}
(2,1)	Σ_{21}	Σ_{11}	0	Σ_{22}	Σ_{12}	0	Σ_{23}	Σ_{13}	0
$(2,2)$	0	$2\Sigma_{21}$	0	0	$2\Sigma_{22}$	0	0	$2\Sigma_{23}$	0
$(2,3)$	0	Σ_{31}	Σ_{21}	0	Σ_{23}	Σ_{22}	0	Σ_{33}	Σ_{23}
(3,1)	Σ_{31}	0	Σ_{11}	Σ_{23}	0	Σ_{12}	Σ_{33}	0	Σ_{13}
(3,2)	0	Σ_{31}	Σ_{21}	0	Σ_{23}	Σ_{22}	0	Σ_{33}	Σ_{23}
$(3,3)$	0	0	$2\Sigma_{31}$	0	0	$2\Sigma_{23}$	0	0	$2\Sigma_{33}$

Rows with an italicized index correspond to strictly upper triangular entries in the Lyapunov equation from (1.2).

Appendix D. Deterministic Result on Support Recovery

In this section, we provide the deterministic result that Theorem 2 is based on. We adapt Theorem 1 by Lin et al. (2016) to arrive at our deterministic result. This requires resolving only a few differences, as we describe in Remark D.2. The underlying construction for the proof is the Primal-Dual-Witness (PDW) method (Wainwright, 2009).

Theorem D.1 *Let $M^* \in Stab_p$ be the true drift matrix, and let S be its support. Assume that Γ^*_{SS} is invertible and that the irrepresentability condition*

$$\|\Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}\|_\infty < 1 - \alpha \tag{D.1}$$

holds with parameter $\alpha \in (0, 1]$. Furthermore, assume that $\hat{\Gamma}$ is a matrix such that

$$\|(\Delta_\Gamma)_{\cdot S}\|_\infty < \epsilon_1, \qquad \|\Delta_g\|_\infty < \epsilon_2,$$

with $\epsilon_1 \leq \alpha/(6c_{\Gamma^*})$. If

$$\lambda > \frac{3(2-\alpha)}{\alpha} \max\{c_{M^*}, \epsilon_1, \epsilon_2\},$$

then the following statements hold:

a) *The LSGE $\hat{M}$ is unique, has its support included in the true support ($\hat{S} \subseteq S$), and satisfies*

$$||\hat{M} - M^*||_\infty < \frac{2c_{\Gamma^*}}{2-\alpha}\lambda.$$

b) *If*

$$\min_{\substack{1 \le j < k \le m \\ (j,k) \in S}} |M_{jk}^*| > \frac{2c_{\Gamma^*}}{2-\alpha}\lambda,$$

then $\hat{S} = S$ and $sign(\hat{M}_{jk}) = sign(M_{jk}^*)$ for all $(j,k) \in S$.

Proof The proof is very similar to the proof of Theorem 1 in Lin et al. (2016). However, there are a few subtle differences and missing explanations that we add in this proof. For all the calculations that are already carried out in Lin et al. (2016), we refer to the original manuscript for these passages.

We use the PDW technique to prove the result. The estimate $\hat{M}$ satisfies the KKT conditions

$$\hat{\Gamma}\text{vec}(\hat{M}) - \hat{g} + \lambda\hat{z} = 0, \tag{D.2}$$

where $\hat{z} \in \partial\|\text{vec}(\hat{M})\|_1$ is an element of the subdifferential of the ℓ_1-norm, that is, the elements of the vector $\hat{z} \in \mathbb{R}^{p^2}$ satisfy that elements

$$\hat{z}_{(i,j)} = \begin{cases} \text{sign}(\text{vec}(\hat{M})_{(i,j)}) & \text{if } \text{vec}(\hat{M})_{(i,j)} \neq 0, \\ \in [-1,1] & \text{if } \text{vec}(\hat{M})_{(i,j)} = 0. \end{cases}$$

Here, we index $\hat{z}$ by pairs (i,j) with $1 \le i, j \le p$. The optimization problem in (2.5) is convex as Γ is positive semidefinite by construction, and the KKT conditions are necessary and sufficient for a solution to be optimal for the problem. The PDW technique constructs, in three steps, a primal-dual pair $(\hat{M}, \hat{z})$ that satisfies (D.2) and has the support of $\hat{M}$ contained in S.

Since the true signal $M^* \in \text{Stab}_p$ and $C \in PD_p$, there exists a unique positive definite Σ^* determined by the continuous Lyapunov equation in (1.2). As a result

$$\Gamma^*\text{vec}(M^*) - g^* = 0,$$

and we can rewrite the KKT conditions in (D.2) in the following block form

$$\begin{bmatrix} \Gamma_{SS}^* & \Gamma_{SS^c}^* \\ \Gamma_{S^cS}^* & \Gamma_{S^cS^c}^* \end{bmatrix} \begin{bmatrix} (\Delta_M)_S \\ (\Delta_M)_{S^c} \end{bmatrix}$$
$$+ \begin{bmatrix} (\Delta_\Gamma)_{SS} & (\Delta_\Gamma)_{SS^c} \\ (\Delta_\Gamma)_{S^cS} & (\Delta_\Gamma)_{S^cS^c} \end{bmatrix} \begin{bmatrix} \text{vec}(\hat{M})_S \\ \text{vec}(\hat{M})_{S^c} \end{bmatrix} + \begin{bmatrix} (\Delta_g)_S \\ (\Delta_g)_{S^c} \end{bmatrix} + \lambda \begin{bmatrix} \hat{z}_S \\ \hat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where $\Delta_M = \text{vec}(\hat{M}) - \text{vec}(M^*)$. We now construct a pair $(\hat{M}, \hat{z})$ that satisfies the equation.

Step 1. We solve the restricted optimization problem

$$\text{vec}(\tilde{M}) = \arg\min_{\text{vec}(M)_{S^c}=0} \frac{1}{2}\text{vec}(M)^\top \hat{\Gamma}\text{vec}(M) - \hat{g}^\top \text{vec}(M) + \lambda\|\text{vec}(M)\|_1. \tag{D.3}$$

Since Γ_{SS}^* is invertible, under our assumptions, $\hat{\Gamma}_{SS}$ is also invertible. The matrix $\hat{\Gamma}_{SS}$ can be expressed as

$$\hat{\Gamma}_{SS} = \Gamma_{SS}^* + (\hat{\Gamma}_{SS} - \Gamma_{SS}^*) = \Gamma_{SS}^* + (\Delta_\Gamma)_{SS}$$

Factoring out Γ_{SS}^*, we obtain

$$\Gamma_{SS}^* + (\Delta_\Gamma)_{SS} = \Gamma_{SS}^*(I_{|S|} + (\Gamma_{SS}^*)^{-1}(\Delta_\Gamma)_{SS})$$

where $I_{|S|}$ denotes the identity matrix of size $|S| \times |S|$. Then, the matrix $\hat{\Gamma}_{SS}$ is invertible if

$$\rho((\Gamma_{SS}^*)^{-1}(\Delta_\Gamma)_{SS}) < 1.$$

This is true as the spectral norm is bounded by the maximum absolute row sum norm and

$$\|(\Gamma_{SS}^*)^{-1}(\Delta_\Gamma)_{SS}\|_\infty \leq \|(\Gamma_{SS}^*)^{-1}\|_\infty\|(\Delta_\Gamma)_{SS}\|_\infty < 1$$

with the second inequality being true because of $\|(\Delta_\Gamma)_{SS}\|_\infty < \epsilon_1 < \alpha/6c_{\Gamma^*} < 1/c_{\Gamma^*}$ and $c_{\Gamma^*} = \|(\Gamma_{SS}^*)^{-1}\|_\infty$.

Therefore, the solution $\text{vec}(\tilde{M})$ is unique. Furthermore, we have

$$(\text{vec}(\tilde{M}))_S = (\hat{\Gamma}_{SS})^{-1}(\hat{g}_S - \lambda\text{sign}((\text{vec}(\tilde{M}))_S)).$$

Let $\tilde{\Delta}_M = \text{vec}(\tilde{M}) - \text{vec}(M^*)$. Following the proof of Theorem 1 in Lin et al. (2016), we have

$$\|\tilde{\Delta}_M\|_\infty \leq \frac{c_{\Gamma^*}}{1 - \alpha/6} \cdot \frac{6 - \alpha}{3(2 - \alpha)}\lambda = \frac{2c_{\Gamma^*}}{2 - \alpha}\lambda. \tag{D.4}$$

Step 2. Let $\tilde{z}_S = \text{sign}(\text{vec}(\tilde{M})_S)$. Then $\tilde{z}_S \in \partial\|\text{vec}(\tilde{M})\|_1$.

Step 3. Let

$$\tilde{z}_{S^c} = \frac{1}{\lambda}\left[-\Gamma_{S^cS}^*(\Gamma_{SS}^*)^{-1}((\Delta_\Gamma)_{SS}\text{vec}(\tilde{M})_S + (\Delta_g)_S) + (\Delta_\Gamma)_{S^cS}\text{vec}(\tilde{M})_S\right.$$
$$\left. + (\Delta_g)_{S^c} + \lambda\Gamma_{S^cS}^*(\Gamma_{SS}^*)^{-1}\text{sign}(\text{vec}(\tilde{M})_S)\right]. \tag{D.5}$$

We show that $\|\tilde{z}_{S^c}\|_1 < 1$, which is a dual feasibility condition. Once this is shown, we have that the pair $(\text{vec}(\tilde{M}), \tilde{z})$ satisfies (D.2) by construction, and $(\text{vec}(\hat{M}), \hat{z}) = (\text{vec}(\tilde{M}), \tilde{z})$ is the solution to the optimization problem in (2.5). Furthermore, Lemma 1 of Wainwright (2009) implies that the strict dual feasibility implies that $\hat{S} \subseteq S$. Following Theorem 1 in Lin et al. (2016), we have

$$\|\tilde{z}_{S^c}\|_\infty \leq \underbrace{\frac{2 - \alpha}{\lambda}\|(\Delta_\Gamma)_{\cdot S}\text{vec}(M^*)_S\|_\infty}_{G_1} + \underbrace{\frac{2 - \alpha}{\lambda}\|(\Delta_\Gamma)_{\cdot S}\|_\infty\|\Delta_S\|_\infty}_{G_2}$$
$$+ \underbrace{\frac{2 - \alpha}{\lambda}\|\Delta_g\|_\infty}_{G_3} + (1 - \alpha).$$

For G_1, we have that

$$G_1 \leq \frac{2 - \alpha}{\lambda}\|(\Delta_\Gamma)_{\cdot S}\|_\infty\|\text{vec}(M^*)\|_\infty = \frac{2 - \alpha}{\lambda}c_{M^*}\epsilon_1 \leq \frac{\alpha}{3}.$$

For G_3, we have that

$$G_3 = \frac{2-\alpha}{\lambda}\|\Delta_g\|_\infty < \frac{2-\alpha}{\lambda}\epsilon_2 \leq \frac{\alpha}{3}.$$

Finally, for G_2, we have that

$$G_2 < \frac{2-\alpha}{\lambda} \cdot \frac{\alpha}{6c_{\Gamma*}} \cdot \epsilon_1 \cdot \frac{c_{\Gamma^*}}{1-\alpha/6} \cdot \frac{6-\alpha}{3(2-\alpha)} < \frac{\alpha}{3}.$$

Combining these bounds, we have that $\|\tilde{z}_{S^c}\|_\infty < 1$, which establishes the strict dual feasibility.

Finally, for any $(j,k) \in S$, we have that

$$|\hat{M}_{jk}| \geq |M_{jk}^*| - |\hat{M}_{jk} - M_{jk}^*| > \min_{\substack{1 \leq j < k \leq p \\ (j,k) \in S}} |M_{jk}^*| - \|\text{vec}(\hat{M}) - \text{vec}(M^*)\|_\infty > 0,$$

which shows that $\hat{S} = S$. ∎

Remark D.2 *The distinction between Theorem D.1 and Theorem 1 of Lin et al. (2016) lies in the steps of our analysis that involve the maximal absolute row sum norm in the bound for the difference between the estimated Hessian $\hat{\Gamma}$ and the true Hessian $\Gamma^\star$. In Lin et al. (2016), the bound was based on the maximal entry. This difference requires adjustments in certain steps of the proof. Consequently, our above proof also provides a more detailed explanation of certain arguments that were omitted by Lin et al. (2016), but are of greater significance in our work. For example, we address the issue of invertibility of $\hat{\Gamma}$. We also indicate when parts of the proof by Lin et al. (2016) are unaffected to ensure clarity and consistency.*

Appendix E. Probabilistic Analysis

The Direct Lyapunov Lasso depends on the loss being sufficiently close to its population version in the sense of $\Delta_\Gamma = \hat{\Gamma} - \Gamma^*$ and $\Delta_g = \hat{g} - g^*$ being sufficiently small. In this section, we bound Δ_Γ and Δ_g in terms of $\Delta_\Sigma = \hat{\Sigma} - \Sigma^*$ and, subsequently, use a concentration inequality for $\|\Delta_\Sigma\|_2$ to probabilistically bound Δ_Γ and Δ_g.

Deriving an inequality for $\hat{\Gamma}$ is most critical as the matrix contains sums of products of covariances and a careful analysis is required to obtain a non-trivial requirement on the sample size. Let $\Gamma(\Sigma) = \Gamma_1(\Sigma) + \Gamma_2(\Sigma)$, where

$$\Gamma_1(\Sigma) = 2(\Sigma^2 \otimes I_p) \quad \text{and} \quad \Gamma_2(\Sigma) = (\Sigma \otimes \Sigma)K^{(p,p)} + K^{(p,p)}(\Sigma \otimes \Sigma).$$

Lemma E.1 *Let $c_{\Sigma^*} = \|\Sigma^*\|_2$. Then*

$$\|\Gamma_1(\hat{\Sigma}) - \Gamma_1(\Sigma^*)\|_2 \leq 2\|\Delta_\Sigma\|_2^2 + 4c_{\Sigma^*}\|\Delta_\Sigma\|_2.$$

Proof Using that $\|A \otimes B\|_2 = \|A\|_2\|B\|_2$, we obtain that

$$\begin{aligned}
\|\Gamma_1(\hat{\Sigma}) - \Gamma_1(\Sigma^*)\|_2 &= 2\|(\hat{\Sigma}^2 - (\Sigma^*)^2) \otimes I_p)\|_2 \\
&= 2\|\hat{\Sigma}^2 - (\Sigma^*)^2)\|_2 \\
&\leq 2\|\Delta_\Sigma\|_2^2 + 2\|\Delta_\Sigma\Sigma^*\|_2 + 2\|\Sigma^*\Delta_\Sigma\|_2.
\end{aligned}$$

Since the spectral norm of a symmetric matrix is the absolute maximal eigenvalue, and the eigenvalues of a squared matrix are the squared eigenvalues of the original matrix, we find as claimed that

$$\||\Gamma_1(\hat{\Sigma}) - \Gamma_1(\Sigma^*)\||_2 \leq 2\||\Delta_\Sigma\||_2^2 + 4\||\Sigma^*\||_2\||\Delta_\Sigma\||_2.$$

∎

Lemma E.2 *Let $c_{\Sigma^*} = \||\Sigma^*\||_2$. Then*

$$\||\Gamma_2(\hat{\Sigma}) - \Gamma_2(\Sigma^*)\||_2 \leq 2\||\Delta_\Sigma\||_2^2 + 4c_{\Sigma^*}\||\Delta_\Sigma\||_2.$$

Proof The commutation matrix $K^{(p,p)}$ is an orthonormal matrix. Therefore, $\||K^{(p,p)}\||_2 = 1$ and

$$\||K^{(p,p)}(\hat{\Sigma} \otimes \hat{\Sigma} - \Sigma^* \otimes \Sigma^*)\||_2 = \||(\hat{\Sigma} \otimes \hat{\Sigma} - \Sigma^* \otimes \Sigma^*)K^{(p,p)}\||_2 = \||\hat{\Sigma} \otimes \hat{\Sigma} - \Sigma^* \otimes \Sigma^*\||_2.$$

We obtain that

$$\begin{aligned}
\||\Gamma_2(\hat{\Sigma}) - \Gamma_2(\Sigma^*)\||_2 &\leq 2\||\hat{\Sigma} \otimes \hat{\Sigma} - \Sigma^* \otimes \Sigma^*\||_2 \\
&= 2\||\Delta_\Sigma \otimes \Delta_\Sigma + \Delta_\Sigma \otimes \Sigma^* + \Sigma^* \otimes \Delta_\Sigma + \Sigma^* \otimes \Sigma^* - \Sigma^* \otimes \Sigma^*\||_2 \\
&\leq 2\||\Delta_\Sigma \otimes \Delta_\Sigma\||_2 + 2\||\Delta_\Sigma \otimes \Sigma^*\||_2 + 2\||\Sigma^* \otimes \Delta_\Sigma\||_2 \\
&\leq 2\||\Delta_\Sigma\||_2^2 + 4\||\Sigma^*\||_2\||\Delta_\Sigma\||_2,
\end{aligned}$$

which was the claim. ∎

For a matrix $A \in \mathbb{R}^{p \times d}$, it holds that $\|A\|_\infty \leq \sqrt{d}\|A\|_2$. Then it follows from Lemma E.1 and Lemma E.2 that

$$\|(\Delta_\Gamma)_{\cdot S}\|_\infty \leq \sqrt{d}\left(4\||\Delta_\Sigma\||_2^2 + 8c_{\Sigma^*}\||\Delta_\Sigma\||_2\right). \tag{E.1}$$

We note that bounding $\|(\Delta_\Gamma)_{\cdot S}\|_\infty$ using $\|(\Delta_\Gamma)_{\cdot S}\|_\infty$, as was done in Lin et al. (2016), leads to a worse bound. While such an approach might seem simpler, it does not exploit the structure of the Hessian Γ in Lemma 1.

We now provide a bound on $\|\Delta_g\|_\infty$.

Lemma E.3 *We have $\|\Delta_g\|_\infty \leq 2c_C\||\Delta_\Sigma\||_2$, where $c_C = \|vec(C)\|_2$.*

Proof Similar to the proof of Lemma E.1 and Lemma E.2, we have

$$\begin{aligned}
\|\Delta_g\|_\infty &\leq \|\Delta_g\|_2 \\
&\leq c_C\||\Sigma^* \otimes I_p - (I_p \otimes \Sigma^*)K^{(p,p)} - \hat{\Sigma} \otimes I_p + (I_p \otimes \hat{\Sigma})K^{(p,p)}\||_2 \\
&\leq c_C(\||I_p \otimes (\hat{\Sigma} - \Sigma^*)\||_2 + \||(\hat{\Sigma} - \Sigma^*) \otimes I_p\||_2) \quad \text{(since } \||K^{(p,p)}\||_2 = 1\text{)} \\
&= 2c_C\||\Delta_\Sigma\||_2.
\end{aligned}$$

∎

The bounds in (E.1) and Lemma E.3 depend on the spectral norm of Δ_Σ. We adapt Theorem 6.5 in Wainwright (2019) to our setting to upper bound $\||\Delta_\Sigma\||_2$ under the assumption that $(x_i)_{i=1}^n$ are sub-Gaussian.

Theorem E.4 (Theorem 6.5. in Wainwright (2019)) *Suppose that $(X_i)_{i=1}^n$ are σ sub-Gaussian random variables. Then the sample covariance matrix $\hat{\Sigma}$ in (1.3) satisfies*

$$\mathbb{P}\left(\frac{\|\hat{\Sigma}-\Sigma^*\|_2}{\sigma^2} \geq c_1\left\{\sqrt{\frac{p}{n}}+\frac{p}{n}\right\}+\delta\right) \leq c_2\exp(-c_3 n\min\{\delta,\delta^2\}) \quad \forall\delta\geq 0,$$

where $\{c_j\}_{j=0}^3$ are universal constants.

Corollary E.5 *Let $\{c_j\}_{j=1}^3$ be the universal constants from Theorem E.4, but ensuring that $c_1 > \max\{1, 1/\|\Sigma^*\|_2\}$. Let $(X_i)_{i=1}^n$ be Gaussian random variables. For any $\epsilon \in (4c_1\|\Sigma^*\|_2\sqrt{p/n}, 2)$, we have*

$$\mathbb{P}\left(\|\hat{\Sigma}-\Sigma^*\|_2 \geq \epsilon\right) \leq c_2\exp\left(-\frac{c_3}{4\max(1, \|\Sigma^*\|_2^2)}n\epsilon^2\right).$$

Proof A Gaussian random vector is sub-Gaussian with parameter $\sigma = \|\Sigma^*\|_2$.
Set $\delta = \min\left(\frac{\epsilon}{2\|\Sigma^*\|_2}, \frac{\epsilon}{2}\right)$. Since $\frac{p}{n} < \frac{\epsilon^2}{16c_1^2\|\Sigma^*\|_2^2}$, we have

$$\|\Sigma^*\|_2\left(c_1\left\{\sqrt{\frac{p}{n}}+\frac{p}{n}\right\}+\delta\right) < c_1\|\Sigma^*\|_2\left\{\frac{\epsilon}{4c_1\|\Sigma^*\|_2}+\frac{\epsilon^2}{16c_1^2\|\Sigma^*\|_2^2}\right\}+\frac{\epsilon}{2}$$

$$=\frac{\epsilon}{4}+\frac{\epsilon^2}{16c_1\|\Sigma^*\|_2}+\frac{\epsilon}{2} < \frac{\epsilon}{4}+\frac{\epsilon}{4}+\frac{\epsilon}{2} = \epsilon.$$

Since $\delta < 1$, it holds that $\delta^2 < \delta$. Then

$$\mathbb{P}\left(\|\hat{\Sigma}-\Sigma^*\|_2 \geq \epsilon\right) \leq \mathbb{P}\left(\|\hat{\Sigma}-\Sigma^*\|_2 \geq \|\Sigma^*\|_2\left(c_1\left\{\sqrt{\frac{p}{n}}+\frac{p}{n}\right\}+\delta\right)\right)$$

$$\leq c_2\exp(-c_3 n\delta^2) = c_2\exp\left(-\frac{c_3}{4\max(1, \|\Sigma^*\|_2^2)}n\epsilon^2\right).$$

$\blacksquare$

We finally have the following result.

Lemma E.6 *In the event that*

$$\|\Delta_\Sigma\|_2 = \|\hat{\Sigma}-\Sigma^*\|_2 < \min\left\{\frac{\epsilon_1}{\sqrt{d}(4+8c_{\Sigma^*})}, \frac{\epsilon_2}{2c_C}\right\}$$

it holds that

$$\|(\Delta_\Gamma)_{\cdot S}\|_\infty < \epsilon_1 \quad \text{and} \quad \|\Delta_g\|_\infty < \epsilon_2.$$

Proof The result follows directly from (E.1), where $\|\Delta_\Sigma\|_2^2 \leq \|\Delta_\Sigma\|_2$, and Lemma E.3. $\blacksquare$

Appendix F. Proof Probabilistic Guarantee on Support Recovery

Using the preparation in Appendix E, we prove the main result.

Proof [Proof of Theorem 2] We prove the result in three steps.

1) It has to hold that

$$\frac{\epsilon}{\sqrt{d}(4 + 8c_{\Sigma^*})}, \frac{\epsilon}{2c_C} \in \left(4c_1 \|\Sigma^*\|_2 \sqrt{p/n}, 2\right).$$

2) Then Corollary E.5 gives us that

$$\|\Delta_\Sigma\|_2 < \min \left\{ \frac{\epsilon}{\sqrt{d}(4 + 8c_{\Sigma^*})}, \frac{\epsilon}{2c_C} \right\}$$

with probability at least $1 - c_2 \exp(-\tau_1 p)$. Then $\|(\Delta_\Gamma)._S\|_\infty < \epsilon$ and $\|\Delta_g\|_\infty < \epsilon$, using Lemma E.6.

3) We verify that $\epsilon \leq \frac{\alpha}{6c_{\Gamma^*}}$ under the assumption on the sample size. Then, the result follows from Theorem D.1.

In the following, we go through the steps in detail.

1) Using the lower bound on the sample size, it holds that

$$\frac{\epsilon}{\sqrt{d}(4 + 8c_{\Sigma^*})} = \frac{\sqrt{\tau_1 \tilde{c} dp/n}}{\sqrt{d}(4 + 8c_{\Sigma^*})}$$

$$< \frac{\sqrt{\tau_1 \tilde{c} dp/\tau_1 \tilde{c} dp \max\{c_*^2, 1/4\}}}{\sqrt{d}(4 + 8c_{\Sigma^*})}$$

$$= \frac{\sqrt{1/\max\{c_*^2, 1/4\}}}{\sqrt{d}(4 + 8c_{\Sigma^*})}$$

$$\leq \sqrt{1/\max\{c_*^2, 1/4\}}$$

$$\leq \sqrt{4} = 2.$$

Using $\tau_1 \geq 1$, we obtain

$$\frac{\epsilon}{\sqrt{d}(4 + 8c_{\Sigma^*})} = \frac{\sqrt{\tau_1 \tilde{c} dp/n}}{\sqrt{d}(4 + 8c_{\Sigma^*})}$$

$$> \frac{\sqrt{\tilde{c}}\sqrt{p/n}}{(4 + 8c_{\Sigma^*})}$$

$$\geq \frac{\sqrt{(4 + 8c_{\Sigma^*})^2 16 c_1^2 c_{\Sigma^*}^2} \sqrt{p/n}}{(4 + 8c_{\Sigma^*})}$$

$$= 4c_1 c_{\Sigma^*} \sqrt{p/n}.$$

536

2) Using Corollary E.5 we obtain

$$\mathbb{P}\left(\|\Delta_\Sigma\|_2 \geq \frac{\epsilon}{\sqrt{d}(4+8c_{\Sigma^*})}\right)$$

$$\leq c_2 \exp\left(-\frac{c_3}{4\max(1,c_{\Sigma^*}^2)}n\frac{\tau_1\tilde{c}dp/n}{d(4+8c_{\Sigma^*})^2}\right)$$

$$\leq c_2 \exp\left(-\frac{c_3\tilde{c}}{4\max(1,c_{\Sigma^*}^2)(4+8c_{\Sigma^*})^2}\tau_1 p\right)$$

$$\leq c_2 \exp\left(-\tau_1 p\right)$$

3) We verify that $\epsilon \leq \frac{\alpha}{6c_{\Gamma^*}}$ under the assumption on the sample size.

$$\epsilon = \sqrt{\tau_1\tilde{c}dp/n}$$
$$\leq \sqrt{\tau_1\tilde{c}dp/\tau_1\tilde{c}dp\max\{c_*^2,1/4\}}$$
$$= \sqrt{1/\max\{c_*^2,1/4\}}$$
$$\leq \frac{\alpha}{6c_{\Gamma^*}}$$

For the same choice of ϵ and $\epsilon/2c_C$ steps 1) - 3) can be carried out analogously and we obtain

$$\mathbb{P}\left(\|\Delta_\Sigma\|_2 \geq \frac{\epsilon}{2c_C}\right) \leq c_2 \exp\left(-\tau_1 dp\right).$$

The result follows by applying Theorem D.1. ∎

Appendix G. Irrepresentability Condition

This section is divided into four parts. First, we give the proof of Theorem 3 and provide an illustration. Second, we discuss a weaker notion of the irrepresentability condition (4.1) that is necessary for the recovery of the support and is more often fulfilled. Third, we provide a detailed simulation study comparing the fulfillment of the irrepresentability condition and its weaker notion. Finally, we show that the impact of the weak irrepresentability condition is already dramatically increasing the performance of the Direct Lyapunov Lasso.

G.1. Irrepresentability Proof and Example

Proof [Theorem 3] Let $\Sigma^0 = \Sigma(M^0, C)$ be the covariance matrix associated to the drift matrix M^0. As we assume that $C = 2I_p$, we have

$$\Sigma^0 = -(M^0)^{-1} = \text{diag}(1/d_1,\dots,1/d_p).$$

Writing $\Gamma^0 = \Gamma(\Sigma^0)$ for the resulting Gram matrix, we define the *local* irrepresentability constant

$$\tilde{\rho}_G(M^0) = \|\Gamma^0_{S_G^c S_G}(\Gamma^0_{S_G S_G})^{-1}\|_\infty.$$

If a small open ball around M^0 contains a matrix M, then the ball also contains all matrices that are obtained from M by negating one or more of the off-diagonal entries. Hence, by continuity, the irrepresentability condition for support S_G holds uniformly over a neighborhood of M^0 if and only if (i) the submatrix $\Gamma^0_{S_G S_G} = (\Gamma^0)_{S_G S_G}$ is invertible and (ii) $\tilde{\rho}_G(M^0) < 1$.

Since Σ^0 is diagonal, plugging it into the coefficient matrix from (2.2) gives a symmetric matrix with entries

$$
A(\Sigma^0)_{(i,j),(k,l)} = \begin{cases}
2/d_l & \text{if } i = j = k = l, \\
1/d_l & \text{if } i = k,\ j = l \text{ and } k \neq l, \\
1/d_l & \text{if } i = l,\ j = k \text{ and } k \neq l, \\
0 & \text{otherwise.}
\end{cases}
$$

The entries of the Gram matrix $\Gamma^0 = \Gamma(\Sigma^0)$ are the inner products of the columns of $A(\Sigma^0)$. That is,

$$
\Gamma^0_{(i,j),(k,l)} = \begin{cases}
4/d_l^2 & \text{if } i = j = k = l, \\
2/d_l^2 & \text{if } i = k,\ j = l \text{ and } k \neq l, \\
2/(d_k d_l) & \text{if } i = l,\ j = k \text{ and } k \neq l, \\
0 & \text{otherwise.}
\end{cases}
$$

Note that the only off-diagonal entries in Γ^0 occur when the row index is (i,j) and the column index is (j,i) with $i \neq j$. We display the matrices $A(\Sigma^0)$ and Γ^0 for a graph with $p = 3$ nodes in Example G.1.

Case I: Graph contains a two-cycle. Suppose G contains a two-cycle, say $k \to l \to k$ with $k \neq l$. The two edges on the cycle index two columns of $A(\Sigma^0)$ that are linearly dependent. Indeed, the column indexed by (k,l) has only two nonzero entries in rows (k,l) and (l,k), both of which are equal to d_l, and the same holds for the column indexed (l,k) except that the common value of its two nonzero entries is d_k. The columns (k,l) and (l,k) of Γ^0 are similarly linearly dependent. Therefore, the submatrix $\Gamma^0_{S_G S_G}$ fails to be invertible, if the graph G contains a two-cycle. Consequently, the irrepresentability condition holds uniformly over a neighborhood of M^0 only if G is free of two-cycles, in which case we call G *simple*.

Case II. Graph is simple. In the rest of the proof suppose that G is simple. In this case, the submatrix $\Gamma^0_{S_G S_G}$ is diagonal with entries

$$
\Gamma^0_{(k,l),(k,l)} = \begin{cases}
4/d_l^2 & \text{if } k = l, \\
2/d_l^2 & \text{if } k \neq l,
\end{cases}
$$

where $l \to k$ is an edge of G. The second submatrix of interest, $\Gamma^0_{S_G^c S_G}$, also has only one nonzero entry in each column. If $l \to k$ is an edge, indexing column (k,l), then the entry is

$$
(\Gamma^0_{S_G^c S_G})_{(l,k),(k,l)} = 2/(d_k d_l).
$$

Note that G being simple implies that $k \to l$ is not an edge of G. Multiplying the second submatrix to the inverse of the first, we obtain that

$$
\begin{aligned}
&(\Gamma^0_{S^c_G S_G}(\Gamma^0_{S_G S_G})^{-1})_{(i,j),(l,k)} \\
&= \begin{cases} d_k/d_l & \text{if } (i,j) = (k,l) \text{ and } (l,k) \in S_G, \ (k,l) \in S^c_G, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Since $\tilde{\rho}_G(M^0)$ is obtained via the maximum absolute row sum, we have $\tilde{\rho}_G(M^0) < 1$ if and only if $d_i/d_j < 1$ for all pairs $(j,i) \in S_G$, or equivalently, all edges $i \to j \in E$, as the theorem claims. If G contains a cycle of at least length 3, there exists a sequence of edges in E such that $i_1 \to i_2 \to i_3 \to i_m \to i_1$ with $i_1, \ldots, i_m \in V$. Then, we have $\tilde{\rho}_G(M^0) < 1$ if and only if

$$
d_{i_1}/d_{i_2} < 1, \quad d_{i_2}/d_{i_3} < 1, \quad \ldots \quad d_{i_{m-1}}/d_{i_m} < 1, \quad d_{i_m}/d_{i_1} < 1.
$$

Multiplying yields

$$
d_{i_1}/d_{i_2} \cdot d_{i_2}/d_{i_3} \cdot \ldots \cdot d_{i_{m-1}}/d_{i_m} \cdot d_{i_m}/d_{i_1} = 1
$$

which contradicts that all individual quotients are smaller than one. ∎

We illustrate the matrix calculations in the proof of Theorem 3 for a graph on $p = 3$ nodes.

Example G.1 *We consider the 3-chain $G = (V, E)$ displayed in Figure 1, and the matrices*

$$
M^0 = diag(-d_1, -d_2, -d_3) \quad and \quad \Sigma^0 = diag(1/d_1, 1/d_2, 1/d_3).
$$

Ordering rows as
$(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)$ *and columns as*
$(1,1), (2,1), (3,1), (1,2), (2,2), (3,2), (1,3), (2,3), (3,3)$, *we find*

$$
A(\Sigma^0) = \begin{pmatrix}
2/d_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/d_1 & 0 & 1/d_2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/d_1 & 0 & 0 & 0 & 1/d_3 & 0 & 0 \\
0 & 1/d_1 & 0 & 1/d_2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2/d_2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/d_2 & 0 & 1/d_3 & 0 \\
0 & 0 & 1/d_1 & 0 & 0 & 0 & 1/d_3 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/d_2 & 0 & 1/d_3 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2/d_3
\end{pmatrix}
$$

and for Γ^0 using the labelling $(1,1),(2,1),(3,1),(1,2),(2,2),(3,2),(1,3),(2,3),(3,3)$ both for rows and columns we obtain

$$
\begin{pmatrix}
4/d_1^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2/d_1^2 & 0 & 2/d_1d_2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2/d_1^2 & 0 & 0 & 0 & 2/d_1d_3 & 0 & 0 \\
0 & 2/d_1d_2 & 0 & 2/d_2^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 4/d_2^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2/d_2^2 & 0 & 2/d_2d_3 & 0 \\
0 & 0 & 2/d_1d_3 & 0 & 0 & 0 & 2/d_3^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2/d_2d_3 & 0 & 2/d_3^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4/d_3^2
\end{pmatrix}.
$$

Since

$$
\begin{aligned}
S_G &= \{(1,1),(2,1),(2,2),(3,2),(3,3)\} \quad \text{and} \\
S_G^c &= \{(3,1),(1,2),(1,3),(2,3)\}
\end{aligned}
$$

we obtain

$$
(\Gamma_{S_G S_G}^0)^{-1} = \mathrm{diag}(d_1^2/4, d_1^2/2, d_2^2/4, d_2^2/2, d_3^2/4),
$$

$$
\Gamma_{S_G^c S_G}^0 = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & 2/d_1d_2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2/d_2d_3 & 0
\end{pmatrix},
$$

and

$$
\Gamma_{S_G^c S_G}^0 (\Gamma_{S_G S_G}^0)^{-1} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & d_1/d_2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & d_2/d_3 & 0
\end{pmatrix}.
$$

To have $\|\Gamma_{S_G^c S_G}^0 (\Gamma_{S_G S_G}^0)^{-1}\|_\infty < 1$, we need $d_1/d_2 < 1$ and $d_2/d_3 < 1$. With the edges $1 \to 2$ and $2 \to 3$ present in G, this requirement coincides with the statement of Theorem 3.

G.2. Necessity of the Weak Irrepresentability Condition

In Theorem 2 we show that the irrepresentability condition

$$
\|\Gamma_{S^c S}^*(\Gamma_{SS}^*)^{-1}\|_\infty \leq (1 - \alpha), \quad \alpha \in (0,1)
$$

is sufficient for model selection consistency. As we show in the subsequent Proposition, a weaker version of the condition is indeed necessary for model selection consistency.

Definition G.1 *Let $M^* \in \mathrm{Stab}_p$ and $S = S(M)$ its corresponding support set. Then, the weak irrepresentability condition is fulfilled if*

$$
\|\Gamma_{S^c S}^*(\Gamma_{SS}^*)^{-1}\mathrm{sign}(\mathrm{vec}(M^*))_S\|_\infty \leq 1. \tag{G.1}
$$

If the slightly weaker condition (G.1) is violated and the entries in the drift matrix fulfill a minimal signal strength condition, we cannot recover the correct support asymptotically.

Proposition G.2 *Consider the setting of Corollary 2. Let $M^* \in \mathrm{Stab}_p$ with $S = S(M^*)$ such that*

$$\min_{\substack{1 \leq j < k \leq p \\ (j,k) \in S}} |M^*_{jk}| > \frac{2c_{\Gamma^*}}{2 - \alpha} \lambda$$

holds and that the weak irrepresentability condition (G.1) is violated. For a fixed positive definite matrix C, the equilibrium distribution for M^* is given by $\mathcal{N}(0, \Sigma^*)$. Let $X_1, \ldots, X_p \in \mathbb{R}^p$ be an i.i.d sample of centered observations and let

$$\hat{\Sigma}^n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

be the sample covariance. We denote the estimate obtained by the Direct Lyapunov Lasso (1.4) using $\hat{\Sigma}^n$ by $\hat{M}^n$. Then it holds that

$$\mathbb{P}(S(\hat{M}^n) = S(M^*)) \longrightarrow 0 \qquad \text{for } n \to \infty.$$

Proof The proof is based on the proof of Theorem D.1. Since the optimization problem (1.4) is convex, the KKT - conditions

$$\hat{\Gamma}^n \mathrm{vec}(\hat{M}^n) - \hat{g}^n + \lambda \hat{z}^n = 0, \tag{G.2}$$

with

$$\hat{z}^n_{(i,j)} = \begin{cases} \mathrm{sign}(\mathrm{vec}(\hat{M}^n)_{(i,j)}) & \text{if } \mathrm{vec}(\hat{M}^n)_{(i,j)} \neq 0, \\ \in [-1, 1] & \text{if } \mathrm{vec}(\hat{M}^n)_{(i,j)} = 0, \end{cases}$$

are necessary and sufficient for optimality of $\hat{M}^n$. Assume that $S(\hat{M}^n) = S(M^*)$. Then, $\hat{M}^n$ is the unique solution of the support restricted problem (D.3) and following the calculations in the proof of Theorem D.1, the subgradient $\hat{z}^n_{S^c}$ is given by

$$\hat{z}^n_{S^c} = \frac{1}{\lambda} \Big[-\Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}((\Delta^n_\Gamma)_{SS}\mathrm{vec}(\hat{M}^n)_S + (\Delta^n_g)_S) + (\Delta^n_\Gamma)_{S^c S}\mathrm{vec}(\hat{M}^n)_S$$

$$+ (\Delta^n_g)_{S^c} + \lambda\Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}\mathrm{sign}(\mathrm{vec}(\hat{M}^n)_S) \Big]. \tag{G.3}$$

We need $\|\tilde{z}^n_{S^c}\|_\infty \leq 1$ for $\hat{M}^n$ to satisfy the KKT-condtions (G.2). Using Lemma E.6 together with Corollary E.5, we obtain that $\Delta^n_g \xrightarrow{P} 0$ and that $\Delta^n_\Gamma \xrightarrow{P} 0$. Moreover, the inequality (D.4) holds for n large enough for $\hat{M}^n$ resulting in

$$\|\mathrm{vec}(\hat{M}^n)_S - \mathrm{vec}(M^*)_S\|_\infty \leq \frac{2c_{\Gamma^*}}{2 - \alpha} \lambda.$$

Then, we obtain for the weak irrepresentability condition that

$$\Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}\mathrm{sign}(\mathrm{vec}(\hat{M}^n)_S) = \Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}\mathrm{sign}(\mathrm{vec}(M^*)_S).$$

Therefore, we obtain

$$\|\tilde{z}_{S^c}^n\|_\infty \xrightarrow{P} \|\Gamma^*_{S^c S}(\Gamma^*_{SS})^{-1}\text{sign}(\text{vec}(M^*))_S)\|_\infty > 1.$$

Asymptotically, the subgradient condition is violated for $\hat{M}^n$ with $S(\hat{M}^n) = S(M^*)$ and probability 1. Hence,

$$\mathbb{P}(S(\hat{M}^n) = S(M^*)) \longrightarrow 0$$

as the sample size $n \to \infty$. ∎

Remark G.3 *It is also easily possible to construct drift matrices fulfilling the weak irrepresentability condition* (G.1). *The same construction as in Theorem 3 can be used.*

G.3. Simulation Studies: Irrepresentability Condition vs. Weak Irrepresentability Condition

In this section, we want to answer two urgent questions. We have shown that for every DAG there exist non-trivial stable drift matrices such that the irrepresentability condition (3.1) holds. The same is possible for the weak irrepresentability condition (G.1). These signals were constructed to be in a neighborhood of diagonal matrices whose diagonal entries are ordered in accordance with the topological ordering of the DAG. As the size of the graphs increases, this diagonal ordering becomes more restrictive. Moreover, there might be signals that have a different diagonal ordering, but still fulfill the irrepresentability condition. Therefore, the first question is how often the conditions are fulfilled when selecting random drift matrices according to a predetermined distribution.

Given a graph $G = (V, E)$, we generate signals $M^* \in \text{Stab}_p(E)$ by drawing from the uniform distribution on the subset of matrices in $\text{Stab}_p(E)$ that have all entries in $[-1, 1]$. The sampling is carried out by rejection sampling, with rejection of matrices that are not stable.

We consider connected graphs with $p = 2, 3, 4$ nodes and at most $p(p + 1)/2$ edges. This includes all DAGs but also many cyclic graphs. Furthermore, we only consider one labeling of vertices for every graph. For every graph, we check for one million simulated signals M^* if $\rho(M^*) < 1$ and store the signals that meet the irrepresentability condition (3.1). The frequency of signals that meet the irrepresentability condition is shown in Figure G.4.

The frequency with which the irrepresentability condition is fulfilled decreases with increasing number of edges. The decrease is not monotonic in the number of edges, since the restrictiveness is tied to whether an edge adds a new condition on the quotient of the diagonal elements as presented in Theorem 3. An investigation of the drift matrices in Figure G.1 shows that those who fulfill the irrepresentability condition (3.1) all have a diagonal ordering according to our theoretical result.

Example G.2 *Consider the graph shown in Figure G.2. The drift matrices supported on this graph have the highest frequency of irrepresentability among the graphs with three edges in Figure G.1. Since there is no edge between the nodes $\{1, 2, 3\}$, the only conditions on the diagonal are $d_1/d_4 < 1, d_2/d_4 < 1$ and $d_3/d_4 < 1$. Translated, this means that d_4 has to be bigger than d_1, d_2, d_3.*

Following Theorem D.1, the conditions on the diagonal elements for the drift matrices supported in Figure G.3 are $d_1/d_2 < 1, d_2/d_3 < 1$ and $d_3/d_4 < 1$. In particular, these conditions also contain the requirement that d_4 has to be bigger than d_1, d_2, d_3. In addition, they contain the requirement that d_3 has to be bigger than d_1, d_2 and that d_2 has to be bigger than d_1.
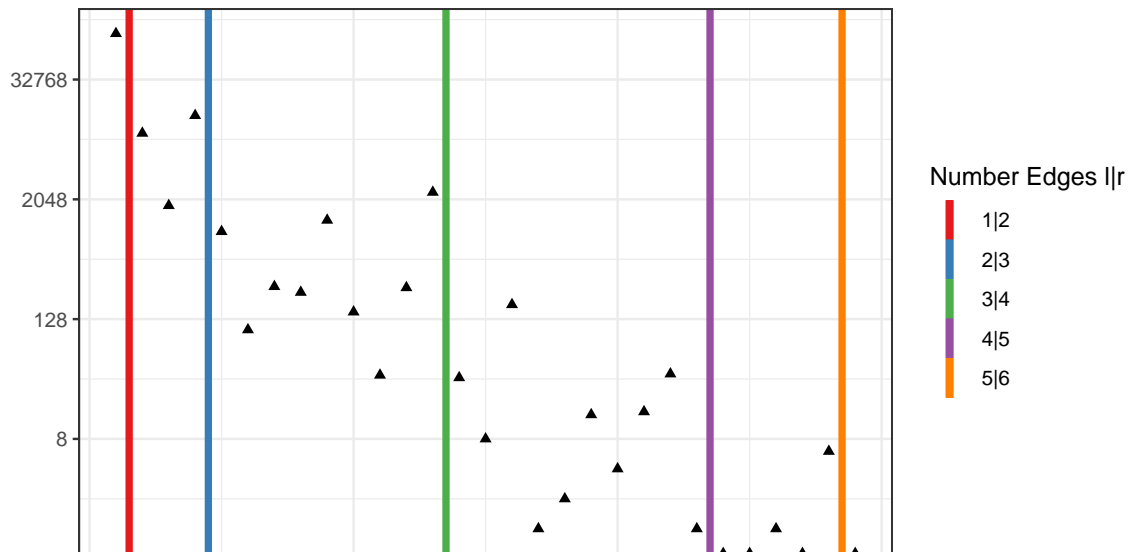
Figure G.1: Frequency of the irrepresentability condition (3.1) being met for one million simulated stable matrices M^* for DAGs up to 4 nodes. The number of edges is given by the coloring.
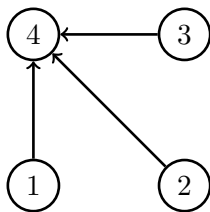

Figure G.2: The graph on three nodes with highest frequency of simulated signals satisfying irrepresentability.

Another important observation is that the condition is extremely restrictive when selecting stable drift matrices according to a uniform distribution. In Figure G.1 we observe that already if a graph on 4 nodes has 3 or more edges, the irrepresentability condition is only fulfilled in less than 1 % of the cases. There even exist some graphs for which the irrepresentability condition is never met. These graphs are displayed in Table 1. We tried to find stable drift matrices by applying the above mentioned selection procedure ten million times to these critical graphs. For only two of the graphs we were able to select drift matrices fulfilling the irrepresentability condition. Theorem 3 guarantees that there must exist stable drift matrices supported over the two remaining graphs. Using Theorem 3, we put one choice for each of the two graphs in red in Table 1.
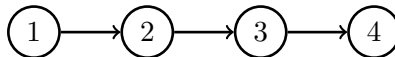
Figure G.3: The path from 1 to 4.
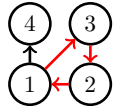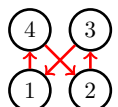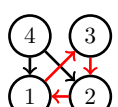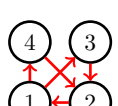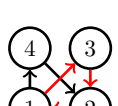
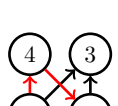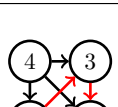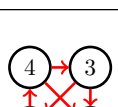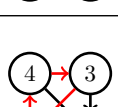**Table 1: Left:** The four graphs where none of the one million randomly selected drift matrices M fulfilled the irrepresentability condition in Figure G.1. **Right:** Drawing another ten million drift matrices, we obtain for the second and third graph drift matrices that fulfill the irrepresentability condition (black). For the first and fourth graph, we use Theorem 3 to construct drift matrices that fulfill the irrepresentability condition (red).

$$\begin{pmatrix} -0.5 & 0 & 0 & 0.05 \\ 0.05 & -1 & 0.05 & 0.05 \\ 0.05 & 0 & -0.75 & 0 \\ 0 & 0 & 0 & -0.25 \end{pmatrix}$$

$$\begin{pmatrix} -0.584860503 & 0.03949857 & 0.0000000 & -0.05605342 \\ 0.000000000 & -0.35729470 & 0.0000000 & -0.00303305 \\ 0.005031837 & -0.08209815 & -0.7782385 & 0.00000000 \\ 0.000000000 & 0.00000000 & 0.0000000 & -0.22854795 \end{pmatrix}$$

$$\begin{pmatrix} -0.7388917 & 0.0000000 & -0.1277403 & 0.01491351 \\ -0.1184546 & -0.9615896 & 0.0000000 & -0.09631827 \\ 0.0000000 & 0.0000000 & -0.4652617 & 0.04858871 \\ 0.0000000 & 0.0000000 & 0.0000000 & -0.23807701 \end{pmatrix}$$

$$\begin{pmatrix} -1 & 0.05 & 0.05 & 0.05 \\ 0 & -0.75 & 0.05 & 0.05 \\ 0 & 0 & -0.5 & 0.05 \\ 0 & 0 & 0 & -0.25 \end{pmatrix}$$

We also carried out the simulation study for simple cyclic graphs. None of the cyclic graphs on 4 nodes fulfilled the irrepresentability condition for ten million randomly selected drift matrices for each graph structure. This is not a proof that the irrepresentability condition (3.1) is never met for a cyclic graph, but at least a strong computational evidence. In a next step we carry out the same sampling procedure for graphs on 4 nodes for the weak irrepresentability condition (G.1) than we did previously for the irrepresentability condition (3.1). The results are displayed in Figure G.4.

Table 2: Left: All simple cyclic graphs with 4 nodes, up to relabelling of the nodes. Edges on cycles are highlighted in red. **Right:** Specific choice of matrices M matching the graph on the left and fulfilling the weak irrepresentability condition (G.1), all entries are rounded to 10 digits.

$$\begin{pmatrix} -0.0444620792 & -0.5733500496 & 0.0000000000 & 0.0000000000 \\ 0.0000000000 & -0.0153532191 & 0.0054622865 & 0.0000000000 \\ 0.8317033453 & 0.0000000000 & -0.8824298000 & 0.0000000000 \\ 0.0000000000 & 0.0000000000 & 0.0000000000 & -0.3405775614 \end{pmatrix}$$

$$\begin{pmatrix} -0.9780979650 & 0.1042322782 & 0.0000000000 & 0.3752107187 \\ 0.0000000000 & -0.7998522464 & -0.4260628200 & 0.0000000000 \\ 0.2079165080 & 0.0000000000 & -0.6517819995 & 0.0000000000 \end{pmatrix}$$

544

$$\begin{pmatrix} -0.6792729949 & -0.6022921619 & 0.0000000000 & 0.0000000000 \\ 0.0000000000 & -0.1733464822 & 0.5762203289 & 0.0000000000 \\ 0.0383909321 & 0.0000000000 & -0.1785332798 & 0.0000000000 \\ 0.2089620568 & 0.0000000000 & 0.0000000000 & -0.6556593408 \end{pmatrix}$$

$$\begin{pmatrix} -0.5008390141 & 0.0000000000 & -0.3301411900 & 0.0000000000 \\ 0.0000000000 & -0.0754047022 & 0.0000000000 & -0.2224099669 \\ 0.0000000000 & 0.9894780936 & -0.8953534714 & 0.0000000000 \\ -0.4568265276 & 0.0000000000 & 0.0000000000 & -0.6545859827 \end{pmatrix}$$

$$\begin{pmatrix} -0.9852473154 & 0.0237436080 & 0.0000000000 & -0.1801203806 \\ 0.0000000000 & -0.9146776730 & -0.6301784553 & -0.3625553502 \\ 0.0314035588 & 0.0000000000 & -0.7371845325 & 0.0000000000 \\ 0.0000000000 & 0.0000000000 & 0.0000000000 & -0.2936787312 \end{pmatrix}$$

$$\begin{pmatrix} -0.6168078599 & -0.4643970933 & 0.0000000000 & 0.0000000000 \\ 0.0000000000 & -0.8265482867 & 0.0118716909 & 0.4726413568 \\ 0.3998511671 & 0.0000000000 & -0.8792877044 & 0.0000000000 \\ -0.5496377517 & 0.0000000000 & 0.0000000000 & -0.7865214688 \end{pmatrix}$$

$$\begin{pmatrix} -0.2066421132 & -0.0034684981 & 0.1383411973 & 0.0000000000 \\ 0.0000000000 & -0.9617960961 & 0.0000000000 & -0.7641737331 \\ 0.0000000000 & -0.3169060163 & -0.7561623598 & 0.0000000000 \\ -0.7012514030 & 0.0000000000 & 0.0000000000 & -0.2419070452 \end{pmatrix}$$

$$\begin{pmatrix} -0.8234110032 & -0.6069790549 & 0.0000000000 & 0.0000000000 \\ 0.0000000000 & -0.4768311884 & 0.0000000000 & -0.5430481988 \\ -0.1151224086 & 0.5541216009 & -0.8947804412 & 0.0000000000 \\ -0.1818817416 & 0.0000000000 & 0.0000000000 & -0.6244826200 \end{pmatrix}$$

$$\begin{pmatrix} -0.7566250684 & 0.1517044385 & 0.0000000000 & 0.0068894741 \\ 0.0000000000 & -0.9917302341 & 0.5077337530 & 0.3153799707 \\ 0.0895817326 & 0.0000000000 & -0.7472212519 & -0.1730670566 \\ 0.0000000000 & 0.0000000000 & 0.0000000000 & -0.3600410065 \end{pmatrix}$$

$$\begin{pmatrix} -0.8680259003 & 0.4557597358 & -0.0925138230 & 0.0000000000 \\ 0.0000000000 & -0.9139470784 & -0.1607573517 & 0.3138186112 \\ 0.0000000000 & 0.0000000000 & -0.9212171654 & -0.9521876550 \\ -0.5101859323 & 0.0000000000 & 0.0000000000 & -0.2475099666 \end{pmatrix}$$

$$\begin{pmatrix} -0.6688544271 & 0.0000000000 & -0.7215559445 & 0.0000000000 \\ -0.4272868899 & -0.9967063963 & 0.0374428187 & -0.8531300114 \\ 0.0000000000 & 0.0000000000 & -0.6779836947 & -0.5781906121 \\ -0.6749138949 & 0.0000000000 & 0.0000000000 & -0.5980373188 \end{pmatrix}$$

Comparing the results in Figure G.4 with those in Figure G.1, we observe that the weak irrepresentability condition is fulfilled much more often than the irrepresentability condition. The reason is that the sign vector in (G.1) enables fortunate cancellation. Moreover, this allows us to find a suitable drift matrix for every simple cyclic graph on 4 nodes. In Table 2, we list all cyclic graphs on 4 nodes together with examples of drift matrices that satisfy the irrepresentability condition. The
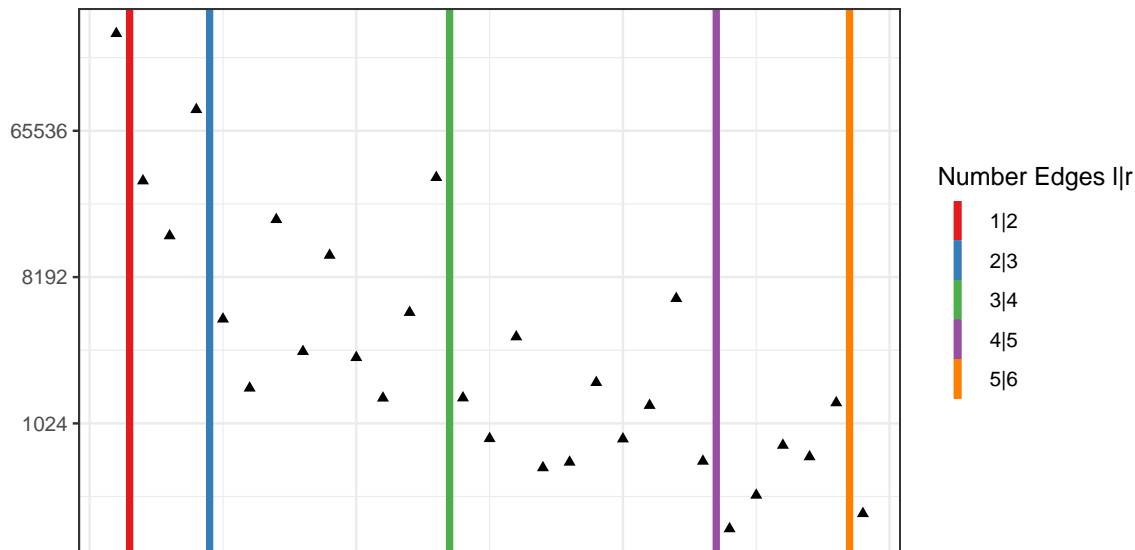
Figure G.4: Frequency of the weak irrepresentability condition (G.1) being met for one million simulated stable matrices M^* for DAGs up to 4 nodes. The number of edges is given by the coloring.
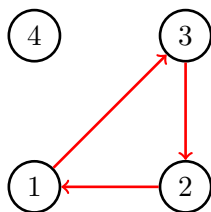

Figure G.5: 3-cycle in a four node setting.

selection of graphs includes all graphs that contain at least one directed cycle and are simple (i.e., do not contain a two-cycle).

Calculations are carried out with the statistical software R. A natural suspicion is that these very few matrices were only selected due to numerical imprecision. In addition, one might wonder if the 10 digits are really necessary. Example G.3 provides more insight using a representative from Table 2.

Example G.3 *For the graph in Figure G.5 (or first row of Table 2) the matrix*

$$M = \begin{pmatrix} -0.0444620792 & -0.5733500496 & 0.0000000000 & 0.0000000000 \\ 0.0000000000 & -0.0153532191 & 0.0054622865 & 0.0000000000 \\ 0.8317033453 & 0.0000000000 & -0.8824298000 & 0.0000000000 \\ 0.0000000000 & 0.0000000000 & 0.0000000000 & -0.3405775614 \end{pmatrix}$$

fulfills the weak irrepresentability condition. The margins to satisfy the weak irrepresentability condition are thin. Rounding the entries of M potentially yields matrices M that do not satisfy the weak irrepresentability condition. The matrix M displayed in this example results in a value for the

left side of (G.1) of 0.9960339 while the 2 - digit version yields a value of 1.011801, i.e. the longer version fulfills the weak irrepresentability condition while the shorter version does not. This is the reason for the long displays in Table 2. However, for the matrix M in this Example, we are able to rationalize the entries with a tolerance of 0.0001 to obtain

$$M_R = \begin{pmatrix} -2/45 & -43/75 & 0 & 0 \\ 0 & -1/65 & 1/183 & 0 \\ 84/101 & 0 & -15/17 & 0 \\ 0 & 0 & 0 & -31/91 \end{pmatrix}$$

fulfilling the weak irrepresentability condition with all calculations being carried out rationally in Mathematica *(Wolfram Research, Inc., 2022). This allays the concern that these matrices only exist due to numerical imprecision in the calculations.*

Summing up the situation for simple cyclic graphs, extensive computation was necessary to present an example for every simple cyclic graph up to four nodes. We were unable to discern the structure that would suggest how to construct such examples in general.

The last class of graphs that misses are the non-simple graphs. We omit discussing them in detail, but the work by Dettling et al. (2023) suggests that there exist drift matrices M^* supported over non-simple graphs such that Γ^*_{SS} is invertible. This leaves the possibility for drift matrices fulfilling the irrepresentability condition. We observed that the drift matrices that satisfy the weak irrepresentability condition fulfill the diagonal ordering of Theorem D.1 for the "DAG part" of the graph over which the drift matrix is supported.

G.4. Simulation Studies: Impact of the Weak Irrepresentability Condition

Corollary 2 ensures that if the irrepresentability condition (3.1) is fulfilled and some assumptions about minimal signal strength and sample size hold, we are able to recover the support of a drift matrix correctly when applying the Direct Lyapunov Lasso (1.4). We were not able to prove this for the weak irrepresentability condition (G.1), only its necessity in Proposition G.2 in case a minimal signal requirement is fulfilled. Nevertheless, the condition is quite close to the sufficient condition and is fulfilled much more often, as we show in Section G.3. Therefore, we want to investigate the impact of the fulfillment of the weak irrepresentability condition on support recovery. The positive computational results in this section also translate to the irrepresentability condition as every drift matrix fulfilling the irrepresentability condition also fulfills the weak irrepresentability condition.

For every DAG on 4 nodes, we select 10 drift matrices fulfilling the weak irrepresentability condition. The selection procedure is the same that we use to obtain Figure G.4 (uniform distribution of stable matrices with entries between -1 and 1). Furthermore, we select 100 stable drift matrices supported over the DAGs that do no necessarily fulfill the irrepresentability condition. Based on the drift matrices M^* and the Lyapunov equation (1.2) with $C = 2I_p$, we calculate the equilibrium covariance matrices Σ^*. We then sampled the data with $n = 100$ from the normal distributions $\mathcal{N}(0, \Sigma^*)$. Then, we apply the Direct Lyapunov Lasso (1.4) along a regularization path

$$\lambda_1 = \lambda_{\max}, \ldots, \lambda_{100} = \frac{\lambda_{\max}}{10^4}$$

where $\lambda_{\max}$ is chosen on an initial grid such that $\hat{M}$ is diagonal. For the estimates $\hat{M}_1, \ldots, \hat{M}_{100}$ obtained along the regularization path, we calculate some basic metrics regarding support recovery of the data generating M^*.

Definition G.4 *Let $\hat{M} \in \mathbb{R}^{p \times p}$ be an estimate and let M^* be the estimation target. Then, we define*

$$tp = |\{\hat{M}_{ij} : \hat{M}_{ij} \neq 0 \, and \, M_{ij}^* \neq 0\}|,$$
$$fp = |\{\hat{M}_{ij} : \hat{M}_{ij} \neq 0 \, and \, M_{ij}^* = 0\}|,$$
$$tn = |\{\hat{M}_{ij} : \hat{M}_{ij} = 0 \, and \, M_{ij}^* = 0\}|,$$
$$fn = |\{\hat{M}_{ij} : \hat{M}_{ij} = 0 \, and \, M_{ij}^* \neq 0\}|.$$

While these metrics already provide some insights, there exist more refined metrics to evaluate the performance of a structure learning algorithm.

Definition G.5 *Let $\hat{M} \in \mathbb{R}^{p \times p}$ be an estimate and let M^* be the estimation target and let tp, fp, tn, fn be defined as in Definition G.4. Then, we define*

$$\textbf{\textit{tpr}} \text{ (true positive rate)} = \frac{tp}{tp + fn},$$
$$\textbf{\textit{fpr}} \text{ (false positive rate)} = \frac{fp}{fp + tn},$$
$$\textbf{\textit{acc}} \text{ (accuracy)} = \frac{tp + tn}{tp + tn + fp + fn},$$
$$\textbf{f}_1\textbf{\textit{-score}} = \frac{2tp}{2tp + fp + fn},$$
$$\textbf{\textit{pr}} \text{ (precision)} = \frac{tp}{tp + fp}.$$

*Calculating tpr and fpr for all regularization parameters, we define the roc curve as plotting tpr vs. fpr with fpr ranging from 0 to 1 using interpolation and extrapolation if necessary. The **auc roc** or just auc is then defined as the area under the roc curve. Calculating pr and tpr for all regularization parameters, we define the pr curve as plotting pr vs. tpr with tpr ranging from 0 to 1 using interpolation and extrapolation if necessary. The **aupr** curve is then defined as the area under the precision curve.*

For the estimates $\hat{M}_1, \ldots, \hat{M}_{100}$ obtained for each DAG and for each initial drift matrix M^*, we calculate the metrics mean tpr, mean fpr and max acc, max f_1-score. All metrics are then averaged over the 10 drift matrices that satisfy the weak irrepresentability condition per DAG or over the 100 randomly selected drift matrices, respectively. The results are displayed in Figure G.6. The empty triangles correspond to the average over the randomly selected drift matrices while the full triangles correspond to the average over the drift matrices fulfilling the weak irrepresentability condition.

Generally, there are many subtleties to be discovered in the plots. For conciseness, we limit our discussion to the key observation that across all metrics, the results for the signals that fulfill the weak irrepresentability condition are almost perfect and much better than for randomly selected ones. Of course, for graphs with fewer edges, more randomly selected drift matrices already fulfill the weak irrepresentability condition, which explains why the difference is not severe.

Lastly, we present the results for the area under the roc curve (auc) using the exact same simulation setup as for Figure G.6. The auc is particularly insightful as the roc curve is obtained by plotting the trade-off of tpr vs. fpr. An auc value of 0.5 means that the method applied performs
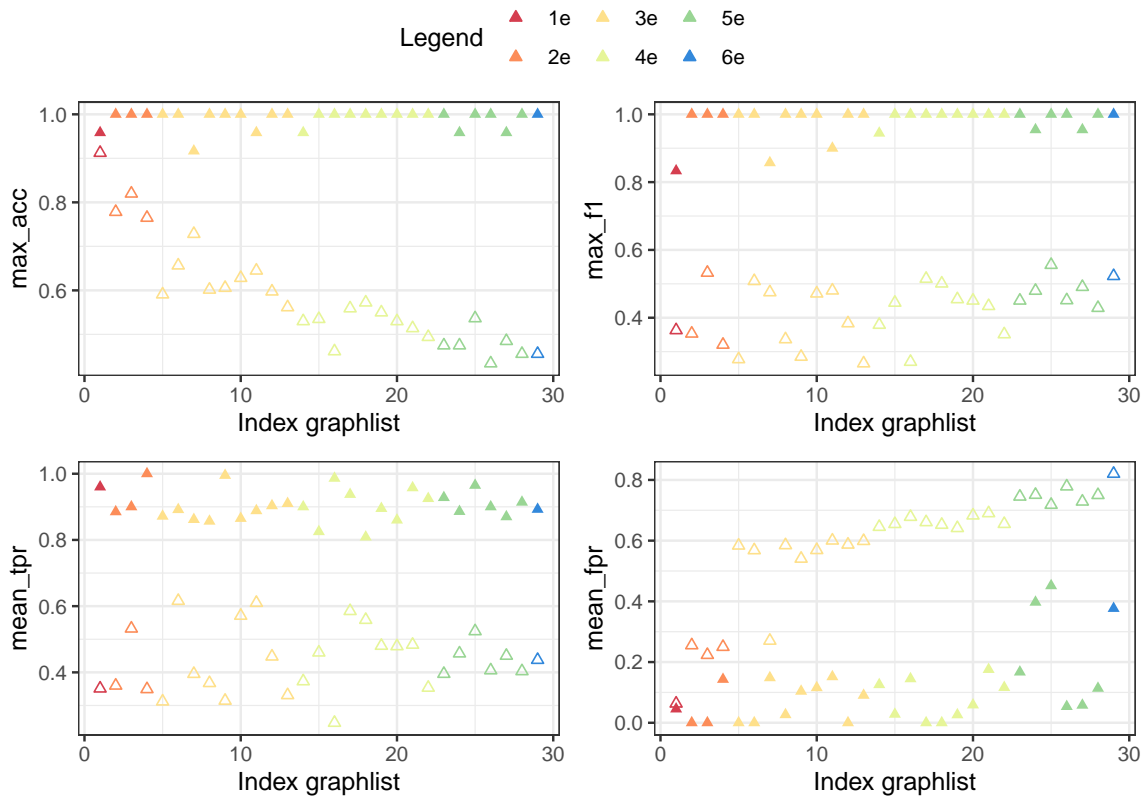
Figure G.6: Four metrics measuring the quality of the estimate for DAGs with up to 4 nodes. The number of edges is given by the coloring. **Empty:** irrepresentability condition in general not fulfilled, **Full:** weak irrepresentability condition fulfilled.

badly (random guessing), while a value of 1 is optimal. For drift matrices fulfilling the weak irrepresentability condition, we observe that the auc is above 0.9 for almost all graphs that fulfill the weak irrepresentability condition while the performance is very poor for randomly selected ones.

We do not include further simulations for cyclic graphs in the above setting, which is mainly because we already struggle to find 10 drift matrices supported over cyclic graphs fulfilling the weak irrepresentability condition. In particular, we struggle to find 10 "really different" drift matrices that do not only differ by a small margin in the individual entries.
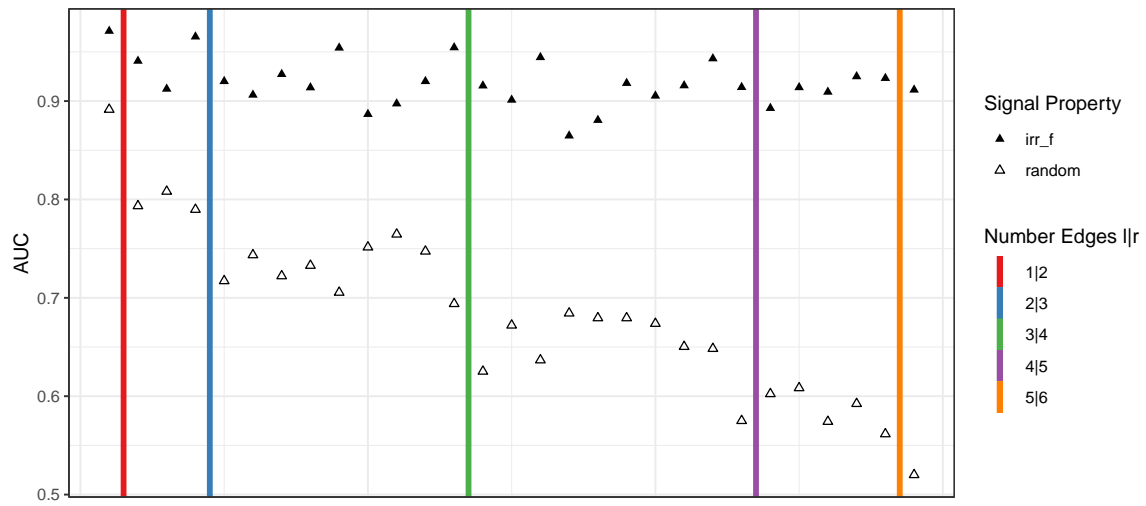
Figure G.7: The auc values for DAGs with up to 4 nodes. The number of edges is given by the coloring. **Empty:** irrepresentability condition in general not fulfilled, **Full:** weak irrepresentability condition fulfilled.