Learning Local Neighborhoods of Non-Gaussian Graphical Models

Sarah Liaw[1], **Rebecca Morrison**[2], **Youssef Marzouk**[3], **Ricardo Baptista**[1]

[1]California Institute of Technology
[2]University of Colorado Boulder
[3]Massachusetts Institute of Technology
{sliaw, rsb}@caltech.edu, rebeccam@colorado.edu, ymarz@mit.edu

Abstract

Identifying the Markov properties or conditional independencies of a collection of random variables is a fundamental task in statistics for modeling and inference. Existing approaches often learn the structure of a probabilistic graph, which encodes these dependencies, by assuming that the variables follow a distribution with a simple parametric form. Moreover, the computational cost of many algorithms scales poorly for high-dimensional distributions, as they need to estimate all the edges in the graph simultaneously. In this work, we propose a scalable algorithm to infer the conditional independence relationships of each variable by exploiting the local Markov property. The proposed method, named Localized Sparsity Identification for Non-Gaussian Distributions (L-SING), estimates the graph by using flexible classes of transport maps to represent the conditional distribution for each variable. We show that L-SING includes existing approaches, such as neighborhood selection with Lasso, as a special case. We demonstrate the effectiveness of our algorithm in both Gaussian and non-Gaussian settings by comparing it to existing methods. Lastly, we show the scalability of the proposed approach by applying it to high-dimensional non-Gaussian examples, including a biological dataset with more than 150 variables.

Introduction

Given a collection of random variables $\mathbf{X} = (X_1, \ldots, X_d)$ with probability measure ν_π and Lebesgue density π, discovering the conditional independence relationships of $\mathbf{X}$ is an important task in statistics. These dependencies are represented in a graph as edges E between vertices V, which correspond to the variables. The resulting graph structure $\mathcal{G} = (V, E)$ is known as a probabilistic graphical model or a Markov random field.

Many real-world processes, such as gene expression levels, generate continuous and non-Gaussian data, requiring methods that can handle such distributions. Gene expression is regulated by complex networks involving transcription factors which exhibit nonlinear dynamics, leading to non-Gaussian distributions (Marko and Weil 2012). Other applications include data of financial market returns and climate variables. These datasets are high-dimensional, and existing algorithms that assume normality may fail to correctly characterize the relevant conditional dependencies.

Structure learning algorithms estimate graphs that capture and summarize the conditional dependencies within a dataset, thereby performing model selection. These algorithms are broadly categorized into global and local methods (Koller and Friedman 2009). Global methods reconstruct the entire graph based on global Markov properties, whereas local methods identify the neighborhood set $Nb(k)$ for each node k in the graph $\mathcal{G}$ based on local Markov properties. The neighborhood set defines variables $X_{Nb(k)}$ such that the conditional density of variable X_k satisfies $\pi(x_k|x_{-k}) = \pi(x_k|x_{Nb(k)})$. In other words, X_k is conditionally independent of variables outside the neighborhood $X_{-Nb(k)}$ given $X_{Nb(k)}$. For distributions with a positive probability density functions, global and local Markov properties are equivalent (Pearl and Paz 1986).

In this work we focus on the following graph recovery problem: given i.i.d. samples $\{\mathbf{x}^i\}_{i=1}^n$ from an (unspecified and possibly non-Gaussian) probability distribution, identify the local neighborhood structure, i.e., local Markov properties, of each variable in the graph.

Existing algorithms that do not assume Gaussianity, such as Sparsity Identification in Non-Gaussian Distributions (SING), rely on the iterative estimation of a transport map to represent the underlying data distribution (Morrison, Baptista, and Marzouk 2017; Baptista et al. 2024). To do so, SING requires estimating the joint density of $\mathbf{X}$, which is often computationally costly for high-dimensional distributions. To overcome this challenge, we propose Localized Sparsity Identification for Non-Gaussian Distributions (L-SING), which learns transport maps representing the conditional distributions of each node in parallel. Our approach exploits the local Markov property to construct a probabilistic graphical model that encodes the conditional dependencies of the nodes. We summarize the algorithm as follows: (1) learn a transport map for each node from samples; (2) use the estimated maps to define a matrix $\Omega \in \mathbb{R}^{d\times d}$, referred to as a *generalized precision*, which encodes conditional dependencies between non-Gaussian variables; (3) determine the edge set of the graph from the sparsity of Ω.

We show that L-SING generalizes existing methods, such as neighborhood selection with the Lasso and the nonparanormal approach. Finally, we show L-SING's scalability and effectiveness by empirically evaluating it on benchmark problems and a high-dimensional dataset of gene expression levels in ovarian cancer patients.

Related Work

Parametric methods for estimating Markov properties assume the observed data follows a probability distribution whose density and properties are defined by some parameters. For Gaussian random variables, conditional dependence is encoded by the sparsity of the precision (inverse covariance) matrix, where zero entries indicate conditional independence between the corresponding variables. Graph structure learning thus reduces to identifying the non-zero entries of the precision matrix in this setting. A widely used approach for this task is the graphical lasso (GLASSO), which solves an L_1-penalized maximum likelihood estimation problem for the precision matrix (Banerjee, Ghaoui, and d'Aspremont 2008). Friedman, Hastie, and Tibshirani (2007) further improved computational efficiency of GLASSO by introducing a coordinate descent algorithm. On the other hand, parametric methods have also been developed to identify local Markov properties by estimating each variable's neighborhood independently. These include greedy selection strategies (Bresler 2015) and penalized maximum likelihood estimators, such as the neighborhood selection method using Lasso regression (Meinshausen and Bühlmann 2006).

The relationship between the sparsity of the inverse covariance matrix and conditional independence, which is central to many algorithms, does not immediately generalize to non-Gaussian distributions. To address general distributions, semi-parametric methods, such as Gaussian copulas, have been proposed to model non-Gaussian data. For example, Liu, Lafferty, and Wasserman (2009) assumed observations are generated from marginal nonlinear transformations of a multivariate Gaussian random vector with known Markov properties. However, the class of distributions explicitly described by known copulas is relatively limited. While Gaussian copulas introduce non-Gaussianity, they may still preserve aspects of the underlying Gaussian structure (Morrison, Baptista, and Basor 2022). Therefore, algorithms that perform well on Gaussian copula-transformed data may struggle with more complex non-Gaussian dependencies.

Previously, Baptista et al. (2024); Morrison, Baptista, and Marzouk (2017) proposed SING, which learns the global Markov structure of continuous and non-Gaussian distributions. SING constructs a lower-triangular transport map to estimate the graph structure and iteratively refines the estimated map until the number of edges converges. To do so, SING estimates a multivariate transport map for the entire set of variables simultaneously, storing $\mathcal{O}(d^2)$ entries and thus becoming computationally and memory intensive in high dimensions. Alternatively, L-SING learns local structure by independently estimating each node's neighborhood. Thus, L-SING eliminates the need to maintain a global transport map and instead allowing each node's map to be computed in parallel. As a result, L-SING scales more efficiently to high-dimensional datasets and can use more expressive maps for each node without exceeding memory constraints.

Recently, Dong and Wang (2022) proposed a neighborhood selection method that approximates the conditional density of each variable based on a smoothing spline ANOVA decomposition. In contrast, L-SING constructs transport maps to represent arbitrary conditional distributions. For certain classes of distributions, constructing transport maps offers a more computationally efficient alternative to direct density estimation.

Transport Maps

A core step of L-SING is estimating a transport map using samples from π. Transport maps represent a target random variable as a transformation of a reference random variable, such as a standard normal. Given a target probability measure ν_π and a reference probability measure ν_η, both defined on $\mathbb{R}^d$, a transport map $S\colon \mathbb{R}^d \to \mathbb{R}^d$ is a measurable function that couples these two variables such that the pushforward measure of ν_π through S is ν_η. We denote the pushforward measure as $S_\sharp \nu_\pi$. This condition implies $S(\mathbf{X}) = \mathbf{Z}$ for $\mathbf{X} \sim \nu_\pi$ and $\mathbf{Z} \sim \nu_\eta$.

If the measures ν_π and ν_η have densities π and η, respectively, we denote the pushforward density as $S_\sharp \pi = \eta$. When S is invertible, the measure of $S^{-1}(\mathbf{Z})$ corresponds to the pullback density, denoted as $S^\sharp \eta = \pi$. For a diffeomorphism S, the pushforward and pullback densities can be expressed using the change-of-variables formula:

$$S^\sharp \eta(\mathbf{x}) = \eta \circ S(\mathbf{x}) \cdot |\det \nabla S(\mathbf{x})|$$
$$S_\sharp \pi(\mathbf{z}) = \pi \circ S^{-1}(\mathbf{z}) \cdot \left|\det \nabla S^{-1}(\mathbf{z})\right|,$$

where $\det \nabla S(\mathbf{x})$ is the determinant of the Jacobian of the map S evaluated at $\mathbf{x}$, and $\circ$ is the composition operator.

In this work, we choose a standard isotropic Gaussian reference, $\nu_\eta = \mathcal{N}(\mathbf{0}, \mathbf{I_d})$ and solve an optimization problem to learn the transport map S given only samples from the target density π. When π and η are strictly positive and smooth densities, Baptista, Marzouk, and Zahm (2023); Marzouk et al. (2016) showed how to learn a lower-triangular transport map S which has the form:

$$S(\mathbf{x}) = \begin{bmatrix} S^1(x_1) \\ S^2(x_1, x_2) \\ S^3(x_1, x_2, x_3) \\ \vdots \\ S^d(x_1, \ldots, x_d) \end{bmatrix}, \tag{1}$$

where S^k is a monotone increasing function of x_k for all $(x_1, \ldots, x_{k-1})$, i.e., $\partial_k S^k := \partial_{x_k} S^k > 0$. A feature of lower-triangular maps (1) is their suitability for causal dependencies. Each component S^k represents the conditional distribution of node X_k given the preceding variables in the specified ordering $(X_1, \ldots, X_{k-1})$. In particular, for a reference distribution with independent components whose density factorizes as $\eta(\mathbf{x}) = \prod_{k=1}^d \eta_k(x_k)$, component S^k pushes forward the marginal conditional distribution of X_k to the marginal distribution of Z_k. That is

$$S^k(x_1, \ldots, x_{k-1}, \cdot)^\sharp \eta_k(x_k) = \pi(x_k | x_1, \ldots, x_{k-1}), \tag{2}$$

for all values of the conditioning variables $(x_1, \ldots, x_{k-1})$. Moreover, if X_k is conditionally independent of X_j given $X_{(1:k-1)\setminus j}$, then both the density $\pi(x_k | x_1, \ldots, x_{k-1})$ and the map S^k do not depend on x_j; see Spantini, Bigoni, and Marzouk (2018) for more details on the relationship between Markov properties of π and the sparsity of triangular maps.

In L-SING, we exploit the relationship between the sparsity of the map component and conditional dependence. In particular, we seek map components S^k that represent the conditional distribution of each variable X_k given all other variables $X_{-k} \in \mathbb{R}^{d-1}$. Given that the last variable of a triangular transport map depends on all variables, this can be seen as learning a component $S^k \colon \mathbb{R}^d \to \mathbb{R}$ to describe the conditional distribution for X_k as

$$S^k(x_{-k}, \cdot)^\sharp \eta_k(x_k) = \pi(x_k | x_{-k}). \tag{3}$$

Our goal is to extract the neighborhood set $Nb(k) \subset \{1, \dots, d\} \setminus k$ for X_k by learning a map S^k that (sparsely) depends on a subset $Nb(k)$ of its inputs.

Parameterization of Transport Map

To parameterize the transport map component S^k in (3), we use Unconstrained Monotonic Neural Networks (UMNNs) to construct invertible transformations (Wehenkel and Louppe 2019). UMNNs define a strictly monotonic function U as

$$U(x; \psi) = \int_0^x f(t; \psi)\, dt + \beta,$$

where $f(t; \psi) \colon \mathbb{R} \to \mathbb{R}^+$ is a strictly positive parametric function implemented via an unconstrained neural network with parameters ψ, and $\beta \in \mathbb{R}$ is a learnable bias term. The positivity of f is enforced using activation functions like $\exp$ and softplus applied to the network's output.

Since the derivative $U'(x; \psi) = f(x; \psi) > 0$ everywhere, $U(x, ; \psi)$ is strictly monotonic and therefore invertible. This property makes UMNNs suitable for parameterizing transport maps. Additionally, UMNNs allow f to depend on auxiliary input variables, enabling flexible modeling of monotonic functions without restrictive architectural constraints.

In L-SING, an UMNN is used to parameterize each component of the transport map. In particular, we seek $S^k \in \mathcal{S}_k$ where $\mathcal{S}_k$ is a class of monotonic functions implemented as a UMNN with d input features $\forall k \in [1, d]$ with $\partial_k S^k > 0$. We note that the partial derivative of S^k only requires a single forward pass through the neural network representing the function f. This is useful in L-SING to evaluate the conditional density $(S^k)^\sharp \eta_k$, which requires the derivative of the map component.

Learning the Transport Map

To learn the parameters of S^k, we formulate the optimization as the solution to a single convex problem. The objective is to minimize

$$S^k \mapsto \mathbb{E}_{\pi(x_{-k})}[D_{\mathrm{KL}}(S^k(x_{-k}, \cdot)_\sharp \pi(\cdot | x_{-k}) || \eta_k)],$$

which is equivalent to maximizing the log-likelihood of $\pi(x_k | x_{-k})$ under the model given by the transport map (Marzouk et al. 2016). Simply, the approach aims to find the transport map S^k that transforms the conditional distribution $\pi(x_k | x_{-k})$ into a simpler reference distribution η_k.

Given M samples from π, a regularized maximum likelihood estimation problem for S^k is given by

$$\min_{S^k} \frac{1}{M} \sum_{i=1}^{M} \left[\frac{1}{2} (S^k)^2(\mathbf{x}^i) - \log \partial_k S^k(\mathbf{x}^i) \right] + \lambda \Phi(S^k)$$
$$\text{s.t. } S^k \in \mathcal{S}_k; \;\; \partial_k S^k > 0, \qquad (\pi - \text{a.e.}) \tag{4}$$

where $\lambda > 0$ is a regularization parameter and Φ is the regularization penalty term from Rosasco et al. (2012) that is used to promote sparse functional dependence:

$$\Phi(S^k) := \sum_{j=1}^{d} \sqrt{\frac{1}{M} \sum_{i=1}^{M} \left(\frac{\partial S^k(\mathbf{x}^i)}{\partial x_j} \right)^2}. \tag{5}$$

The optimal parameter λ is determined in our experiments by minimizing the validation loss for the objective in (4).

Connections to Existing Methods

To show the generality of L-SING, we demonstrate the connection of the learning problem above to existing methods for Gaussian and nonparanormal distributions.

(Gaussian Case) For a Gaussian vector $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^d$ and non-singular covariance $\Sigma \in \mathbb{R}^{d \times d}$, Meinshausen and Bühlmann (2006) proposed a neighborhood selection method for estimating the neighborhood sets $\{Nb(k) : 1 \le k \le d\}$ through successive linear regressions. The neighborhood selection with Lasso estimates the regression coefficients $\theta^k \in \mathbb{R}^{d-1}$ of X_k given co-variates X_{-k} by solving the regularized optimization problem

$$\hat{\theta}^k = \arg\min_{\theta} \|\mathbf{X}_k - \mathbf{X}_{-k}\theta\|_2^2 + \lambda \|\theta\|_1, \tag{6}$$

where the rows of $\mathbf{X}_k \in \mathbb{R}^M, \mathbf{X}_{-k} \in \mathbb{R}^{M \times (d-1)}$ contain M data samples, and $\|\theta\|_1$ is the L_1-norm of the coefficient vector, which is used to promote sparsity in the solution. The non-zero entries of the resulting vector $\hat{\theta}^k$ defines the neighborhood set $Nb(k)$.

Proposition 1. *For a linear transport map component S^k, the optimization problem in* (4) *reduces to the optimization problem in* (6) *for neighborhood selection with the Lasso.*

Proof. For a linear map component $S^k(\mathbf{x}_{1:d}) = \sum_{l=1}^{d} a_l x_l$ where $(a_l) \in \mathbb{R}$ are scalar parameters, the objective in (4) is

$$\min_{S^k} \frac{1}{M} \sum_{i=1}^{M} \left[\frac{1}{2} (S^k)^2(\mathbf{x}^i) - \log \partial_k S^k(\mathbf{x}^i) \right] + \lambda \Phi(S^k)$$
$$= \min_{a_{1:d}} \frac{1}{M} \sum_{i=1}^{M} \left[\frac{1}{2} \left(\sum_{l=1}^{d} a_l x_l^i \right)^2 - \log a_k \right] + \lambda \sum_{j=1}^{d} |a_j|$$

Consider the change of variables $b_l = \frac{a_l}{a_k}$ for all $l \neq k$. Then,

we can write the optimization problem as

$$\min_{\mathbf{b}_{-k},a_k} \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{2}(a_k)^2\left(\sum_{l\neq k} b_l x_l^i + x_k^i\right)^2 - \log a_k\right] + \lambda\left(\sum_{j\neq k}|b_j a_k| + |a_k|\right),$$

where $\mathbf{b}_{-k} = (b_1, \ldots, b_{k-1}, b_k, \ldots, b_{d-1})$. We identify the re-regression coefficients $\mathbf{b}_{-k}$ by solving the problem

$$\min_{\mathbf{b}_{-k}} \|\mathbf{b}_{-k}\mathbf{X}_{-k} + \mathbf{X}_k\|^2 + \tilde{\lambda}\|\mathbf{b}_{-k}\|_1,$$

where $\tilde{\lambda} = \lambda(2M)/|a_k|$. This objective has the form as the problem in (6) up to a sign in the coefficients $\mathbf{b}_{-k}$. □

While the relationship between L-SING and neighborhood selection with Lasso is only shown for Gaussian distributions, L-SING extends beyond Gaussians by using nonlinear functions (e.g., higher-order polynomials) to parameterize the transport map.

(Nonparanormal Case) A random vector $\mathbf{X} = (X_1, \ldots, X_d)$ has a nonparanormal distribution if there exists a set of functions $\{f_j\}_{j=1}^d$ such that $\mathbf{Z} = \mathbf{f}(\mathbf{X}) \sim \mathcal{N}(\mu, \Sigma)$, where $\mathbf{f}(\mathbf{X}) = (f_1(X_1), \ldots, f_d(X_d))$. Liu, Lafferty, and Wasserman (2009) proposed applying GLASSO to the transformed data to estimate the undirected graph from the sparsity pattern of the estimated precision matrix:

$$\hat{\Theta} = \arg\min_{\Theta\geq 0}\left(\mathrm{Tr}(S(\tilde{\mathbf{f}})\Theta) - \log\det(\Theta) + \lambda\sum_{j\neq k}|\Theta_{jk}|\right),$$

where $S(\tilde{\mathbf{f}})$ is a sample covariance estimator of $\tilde{\mathbf{f}}(\mathbf{X})$ based on an estimator $\tilde{\mathbf{f}}$ of $\mathbf{f}$.

Proposition 2. *Let $\mathbf{X}$ be a random vector following a nonparanormal distribution, i.e., there exists monotonic functions $\mathbf{f}$ such that $\mathbf{f}(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ for some strictly positive definite Σ. Then, the transport map $S(\mathbf{x}) = \Sigma^{-1/2}\mathbf{f}(\mathbf{x})$ pushes forward $\mathbf{X}$ to a standard normal random variable.*

Proof. To show that this is a valid transport map, we show that $S(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. First, we recall that the covariance matrix Σ is symmetric and positive definite, and hence invertible. For a strictly positive definite Σ, we can use the Cholesky decomposition (or another matrix square root) to obtain $\Sigma^{-1/2}$. Next, we verify that $S(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Given that $f(\mathbf{X})$ is Gaussian, we just need to verify the first two moments of the linear transformation of $\mathbf{f}(\mathbf{X})$:

$$\mathbb{E}[S(\mathbf{X})] = \mathbb{E}[\Sigma^{-1/2}\mathbf{f}(\mathbf{X})] = \Sigma^{-1/2}\mathbb{E}[\mathbf{f}(\mathbf{X})] = \mathbf{0}$$
$$\mathrm{Cov}[S(\mathbf{X})] = \Sigma^{-1/2}\mathbb{E}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^T](\Sigma^{-1/2})^T = \mathbf{I}.$$

□

The result above shows that nonparanormal distributions can be characterized using transport maps and their components define its conditionals. Thus, L-SING includes nonparanormal methods as a special case.

Computing the Generalized Precision

Finally, we show how to compute a matrix encoding the conditional independence structure using the estimated transport map. We recall that in the Gaussian setting, the inverse covariance matrix Σ^{-1} is also the precision matrix, where $\Sigma^{-1}_{kj} = 0 \leftrightarrow X_k \perp\!\!\!\perp X_j \,|\, X_{-kj}$ (Loh and Wainwright 2012). For non-Gaussian distributions, we extend this concept using the transport maps that represent the conditional distributions of $\mathbf{X}$. Following Morrison, Baptista, and Marzouk (2017), we consider the generalized precision $\Omega \in \mathbb{R}^{d\times d}$ with entries

$$\Omega_{jk} = \mathbb{E}|\partial_j\partial_k \log\pi(\mathbf{x})| = \mathbb{E}|\partial_j\partial_k \log\pi(x_k|x_{-k})|.$$

Spantini, Bigoni, and Marzouk (2018) showed that $\partial_j\partial_k \log\pi(\mathbf{x}) = 0$ for all $\mathbf{x} \iff X_j \perp\!\!\!\perp X_k \,|\, X_{-kj}$, so the sparsity of the generalized precision matrix encodes pairwise conditional independence properties (similarly to the precision matrix in the Gaussian setting) for distributions whose log-density is twice differentiable.

Given that the map component S^k characterizes the conditional distribution of the target density, we can express the entries of Ω in terms of the map as:

$$\begin{aligned}\Omega_{jk} &= \mathbb{E}|\partial_j\partial_k[\log\eta_k(S^k(\mathbf{x})) + \log\partial_k S^k(\mathbf{x})]| \\ &= \mathbb{E}\left|\partial_j\partial_k\left[-\frac{1}{2}(S^k)^2(\mathbf{x}) + \log\partial_k S^k(\mathbf{x})\right]\right|.\end{aligned}$$

An estimator of this matrix entry based on N i.i.d. samples from π is given by

$$\hat{\Omega}_{jk} := \frac{1}{N}\sum_{i=1}^{N}\left|\partial_j\partial_k\left[-\frac{1}{2}(S^k)^2(\mathbf{x}^i) + \log\partial_k S^k(\mathbf{x}^i)\right]\right|$$

For each variable k, we use the computed transport map S^k to determine the neighborhood set $\mathrm{Nb}(k)$ based on the nonzero entries $\hat{\Omega}_{jk}$ for $j \in \{1, \ldots, d\}\setminus k$. These entries identify edges in the graphical model. In particular, we say that there exists an edge between variables X_k and X_j if Ω_{jk} meets the following conditional independence criteria:

$$\begin{aligned}&X_k \perp\!\!\!\perp X_j \,|\, X_{V\setminus\{kj\}} \Leftrightarrow X_k \perp\!\!\!\perp X_{j\in V\setminus\mathrm{Nb}(k)} \,|\, X_{\mathrm{Nb}(k)} \\ &\Leftrightarrow \partial_k\partial_j \log\pi(\mathbf{x}) = 0 \;\forall\, \mathbf{x} \in \mathbb{R}^d\end{aligned}$$

Thus, the estimation of the generalized precision matrix allows us to compute an edge set that reflects the conditional independence structure of the data. We note that for each pair of edges (j, k), we obtain two estimators for the conditional independence based on the dependence of map component S^j on x_k and S^k on x_j. To reconcile these two estimates, which are theoretically equal when the estimator for the conditional distributions is correct, we compute a symmetrized version of the generalized precision matrix.

L-SING Algorithm

In this section, we present L-SING for learning the Markov structure of a continuous and (possibly) non-Gaussian distributions. The complete procedure is outlined in Algorithm 1.

In practice, we split the provided samples from the target distribution π into training, validation, and estimation

Algorithm 1: L-SING Algorithm

Input i.i.d. samples $\{\mathbf{x}_i\}_{i=1}^{M+N}$, transport class $\mathcal{S}_k$.
Output Generalized precision matrix $\hat{\Omega}$.

for each map component $S^k \in \mathcal{S}_k$ **do**
  for fixed number of epochs **do**
    Compute the regularized loss using M samples
    Back-propagate loss and update S^k parameters
  end for
  Compute entries of the generalized precision matrix $\hat{\Omega}_{jk}\ \forall j = 1, \ldots, d$ using N samples
  Set $\hat{\Omega}_{kk} = 1$
end for
Return $\hat{\Omega}^{\text{L-SING}} := \frac{1}{2}(\hat{\Omega} + \hat{\Omega}^T)$.

sets. The regularized objective in (4) is optimized using the training set, while the unregularized negative log-likelihood is evaluated on the validation set to select the optimal regularization parameter λ. To ensure model generalization, we implement early stopping during the estimation of S^k. That is, training stops if the validation loss fails to improve for 10 consecutive epochs. Finally, we evaluate $\hat{\Omega}$ using the estimation set to avoid biases arising from reusing samples for both learning the map and computing the generalized precision.

Edge Set Generation

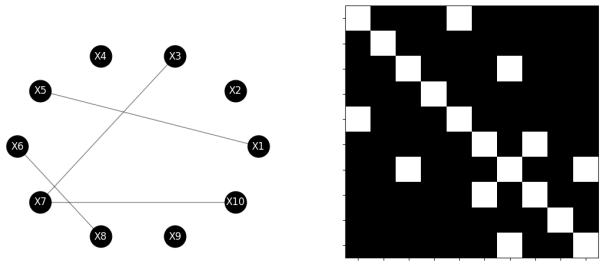
After computing and normalizing the generalized precision matrix $\hat{\Omega}$ (scaled to have maximum value 1), we generate a sparse edge set for the graphical model by thresholding:

1. Choose a threshold value $\tau > 0$.
2. For each pair of variables (j, k) for all $j, k \in [1, d]$:
   - If $|\hat{\Omega}_{jk}| > \tau$, add an edge between variables j and k.
   - Otherwise, no edge is added.

Given that the choice of threshold τ affects the sparsity of the graph, we demonstrate the sensitivity of the graph sparsity and false positive rates (FPR) to variations in τ in our numerical experiments. In practice, τ can be selected based on prior knowledge about the expected graph density.

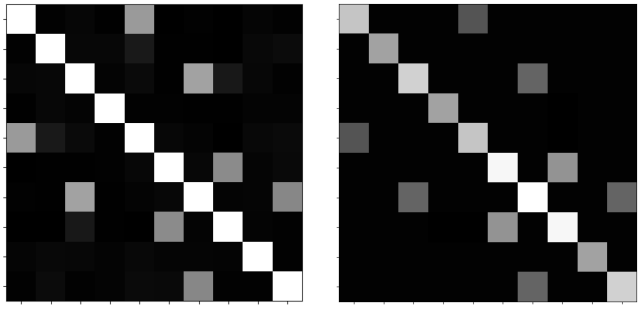
Numerical Results

We now aim to answer the following questions: (1) Can L-SING accurately quantify the conditional dependencies of $\mathbf{X}$ without relying on assumptions about the distribution of $\mathbf{X}$? (2) Is L-SING computationally tractable for high-dimensional problems? The first and second experiments address question (1), while the second and third experiments address question (2). Additionally, we compare the performance of $\hat{\Omega}^{\text{L-SING}}$ to existing methods on the same test dataset (see the arXiv version for detailed experimental setups). The code to reproduce the numerical experiments is available at: https://github.com/SarahLiaw/L-SING.



(a) (b)

Figure 1: (a) The undirected graphical model; (b) Adjacency matrix of true graph (white corresponds to an edge, black to no edge) for the 10-dimensional Gaussian distribution.



(a) $\hat{\Omega}^{\text{L-SING}}$ (b) $\hat{\Omega}^{\text{GLASSO}}$

Figure 2: Generalized precision matrix for L-SING and the estimated precision matrix for GLASSO, which are computed using $N = 10,000$ Gaussian evaluation samples.

Gaussian Distribution

To validate L-SING against existing parametric methods, we evaluated it on data sampled from a Gaussian distribution. We first generated a symmetric, positive definite matrix $\Sigma^{-1} = \Omega$, where Ω is the true precision matrix. Using this distribution, we drew $M = 5,000$ training samples from a d-dimensional multivariate normal distribution $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, with $d = 10$. After training L-SING, we computed $\hat{\Omega}$ using an evaluation set of $N = 10,000$ samples drawn from the same distribution.

Evaluation. Figure 1 visualizes the true underlying graph and its adjacency matrix. Figure 2 compares the estimated precision matrices $\hat{\Omega}$ computed using L-SING and GLASSO.

The heatmap colors in Figures 2a and 2b represent the magnitude of the entries in $\hat{\Omega}$. Figure 2a shows that L-SING accurately recovers the overall sparsity of Ω. For $\tau = 0.2$, the estimated adjacency matrix matches the true adjacency matrix in Figure 1b, with all edges correctly identified. The non-zero off-diagonal entries are within ± 0.1 of the corresponding values in the true normalized precision matrix (Figure 1b).

Figure 2b presents results for the Gaussian simulation using GLASSO. Comparing the off-diagonal entries of $\hat{\Omega}^{\text{GLASSO}}$ to Ω, edges $(1, 5)$ and $(3, 7)$ are within a ± 0.03 range of their true values, while all other edges fall within ± 0.1. Note that GLASSO operates directly on the test sam-

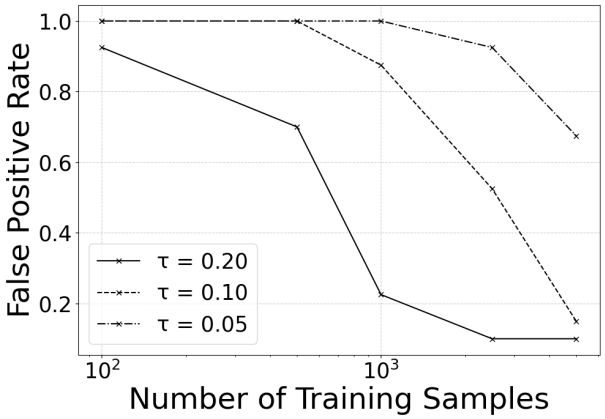
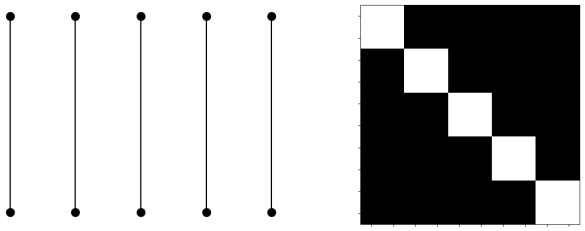


Figure 3: Sensitivity of false positives with L-SING for different thresholds and sample sizes on the Gaussian data.



(a) (b)

Figure 4: (a) The undirected graphical model; (b) Adjacency matrix of true graph for the butterfly distribution ($d = 10$).

ples without requiring training. Figure 3 plots the FPR against the number of training samples for $\tau = 0.20, 0.10, 0.05$. L-SING's FPR decreases with increasing M, which demonstrates its consistency. Higher τ (e.g., 0.20) results in lower FPR across all sample sizes, indicating more conservative edge detection. The choice of τ results in a trade-off between sensitivity and specificity in edge detection.

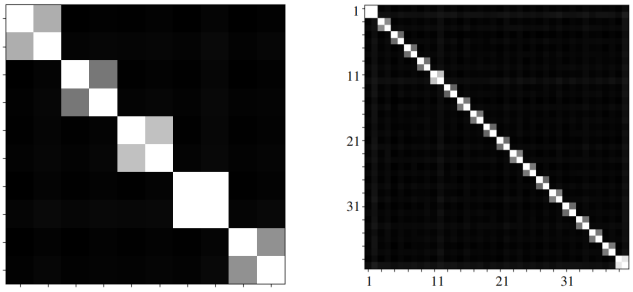
Butterfly Distribution

Next, we evaluate L-SING on a non-Gaussian butterfly distribution, which exhibits nonlinear dependencies. Consider r pairs of random variables (X, Y), where:

$$X \sim \mathcal{N}(0,1) \qquad Y = WX, \text{ with } W \sim \mathcal{N}(0,1).$$

Figure 4 displays the probabilistic graph and the corresponding support of the generalized precision matrix for $r = 5$ pairs ($d = 10$). The variables are ordered $X_1, Y_1, \ldots, X_r, Y_r$, where X_i corresponds to odd-numbered columns/rows in the heatmap and Y_i to even-numbered ones for all $i \in [1, r]$. This ordering is consistent across all plots of the identified graph. While each one-dimensional marginal of the butterfly distribution is symmetric and unimodal, the two-dimensional marginals exhibit strong non-Gaussianity.

Evaluation. Figure 5a shows the estimated generalized precision matrix $\hat{\Omega}$ for $r = 5$ pairs ($d = 10$), as computed using UMNN map components with $[64, 64, 64]$ hidden layers and $M = 5,000$ training samples. L-SING successfully recovers



(a) $\hat{\Omega}^{\text{L-SING}}$ for $d = 10$ (b) $\hat{\Omega}^{\text{L-SING}}$ for $d = 40$

Figure 5: Estimated generalized precision matrix for the Butterfly distribution with $d = 10$ and $d = 40$ variables using L-SING with $N = 10,000$ evaluation samples.

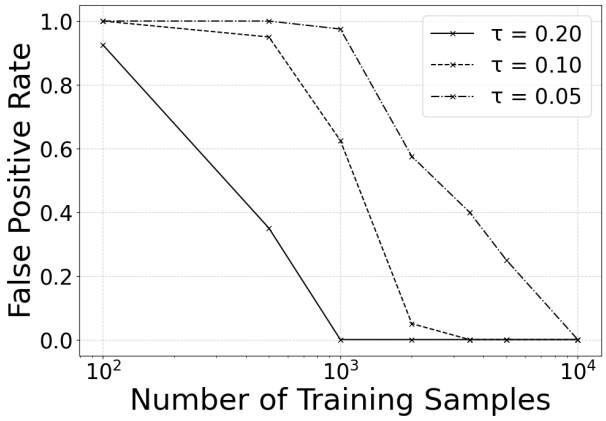
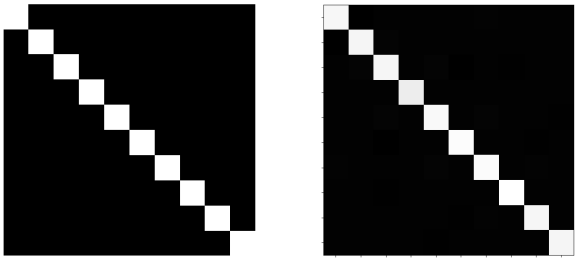


Figure 6: Sensitivity of false positives with L-SING for different thresholds on the 10-dimensional butterfly distribution.

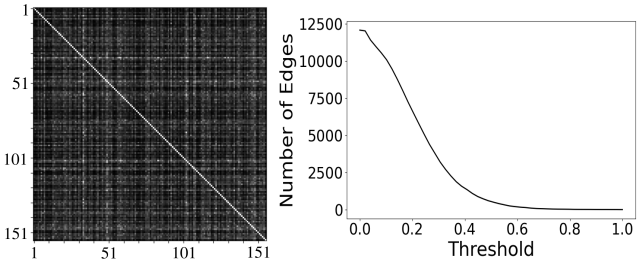
the true sparse structure of the graphical model for $\tau = 0.2$ and 0.1, as shown in Figure 6. To show the scalability of L-SING, Figure 5b presents $\hat{\Omega}$ for $r = 20$ pairs ($d = 40$), using $M = 5,000$ training samples and the same UMNN architecture. With $\tau = 0.1$, L-SING achieves an F_1 score of ≈ 0.941, indicating high precision and recall in identifying the correct edges, even in higher dimensions. An FPR of $\approx 6.58 \times 10^{-3}$ indicates that L-SING introduces minimal spurious edges, demonstrating its ability to accurately recover the structure of the butterfly distribution even as the dimensionality increases.

Figure 7 shows that both GLASSO and the nonparanormal (npn) method yield incorrect graphs for this dataset. These methods only identify the diagonal entries in Figure 4b, corresponding to self-dependencies of each variable, and fail to recover the dependencies between X_i and Y_i for all $i \in [1, r]$. GLASSO's failure arises from its reliance on a Gaussian assumption, which does not hold for the butterfly distribution. Similarly, the nonparanormal method, which applies a truncated ECDF transformation to each variable's marginal distribution before running GLASSO, fails because the butterfly distribution lies outside the class of distributions assumed by Liu, Lafferty, and Wasserman (2009). Consequently, it cannot recover the true edges between variable pairs.



(a) $\hat{\Omega}^{\text{npn}}$ (b) $\hat{\Omega}^{\text{GLASSO}}$

Figure 7: Conditional independencies with (a) the nonparanormal and (b) GLASSO for the butterfly distribution.



(a) $\hat{\Omega}^{\text{L-SING}}$ (b)

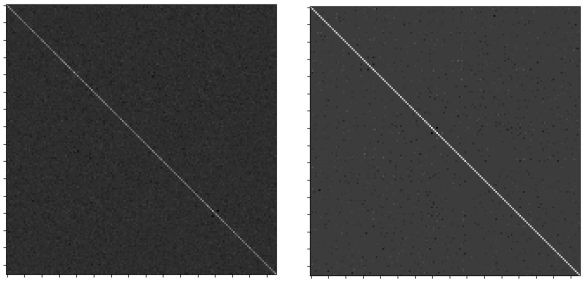
Figure 8: (a) Generalized precision for the Ovarian Cancer dataset; (b) Sensitivity of graph sparsity/recovered edge count to changes in the threshold τ.

Ovarian Cancer Dataset

Finally, we address question (2) by demonstrating the scalability of L-SING on the high-dimensional curated Ovarian Data (Ganzfried et al. 2013), comprising gene expression profiles from 578 ovarian cancer patients sourced from The Cancer Genome Atlas (TCGA). Following the data processing procedure in Shutta et al. (2022), we selected biologically relevant genes. Specifically, we identified two gene sets from the Molecular Signatures Database: genes down- and up-regulated in mucinous ovarian tumors compared to normal ovarian epithelial cells. After intersecting these genes with those available in TCGA, the final dataset included 156 genes (variables) and 578 samples, split into 346 training, 117 evaluation, and 115 validation samples. For L-SING, estimating S^k and computing the corresponding entries of $\hat{\Omega}_k$ took 42.3 seconds, while GLASSO took 13.1 seconds.

Evaluation. Figure 8a presents $\hat{\Omega}$, as computed using UMNN map components with $[64, 128, 128]$ hidden layers. Figure 8b shows the effect of τ on graph sparsity, showing that larger τ leads to sparser graphs, with the most significant changes in edge count and sparsity occurring for $\tau \in [0, 0.4]$. This range is critical for estimating the graph. Rather than relying on manual graph inspection, we follow the threshold selection procedure in Shutta et al. (2022) and set $\tau = 0.2$ based on sparsity and edge count. Notably, 58% of the entries in $\hat{\Omega}^{\text{L-SING}}$ are less than 0.2, resulting in a sparse graph as weak connections are effectively pruned.

To avoid biases in model interpretation due to qualitative inspection, we adopt a quantitative approach using centrality measures to compare against existing methods, given the absence of a ground truth graph for this problem. Specifically, we compute an average centrality rank based on betweenness, degree, hubscore, and closeness centralities. Using the support of the generalized precision matrix from L-SING (Figure 8a), we observe that the gene CTSE exhibits the highest mean centrality rank. This aligns with prior findings by Marquez et al. (2005), which identify CTSE as an up-regulated and specific marker for mucinous ovarian cancers.



(a) $\hat{\Omega}^{\text{GLASSO}}$ (b) $\hat{\Omega}^{\text{npn}}$

Figure 9: Estimated pairwise conditional independencies on evaluation samples from the Ovarian Cancer problem.

We compare these results to those of Shutta et al. (2022), who use GLASSO (Figure 9a). They highlight EPCAM as the second most important gene in ovarian cancer, supported by Spizzo et al. (2006), but did not identify CTSE as significant. In contrast, L-SING ranks EPCAM 14th out of 156 genes based on mean centrality. A direct comparison between L-SING and GLASSO is challenging due to the lack of a ground truth, as L-SING captures additional non-Gaussian dependencies that GLASSO, by design, cannot detect. Finally, we compare $\hat{\Omega}^{\text{GLASSO}}$ and $\hat{\Omega}^{\text{npn}}$. The difference, measured by the Frobenius norm, is $\|\hat{\Omega}^{\text{GLASSO}} - \hat{\Omega}^{\text{npn}}\|_F = 6.43$, indicating that relying solely on Gaussian-based methods provides an incomplete representation of the network structure. We demonstrate that parametric methods like GLASSO may fail to detect non-Gaussian relationships, underscoring the advantage of using L-SING to recover nonlinear dependencies.

Conclusion

We have proposed L-SING, a method for learning the structure of high-dimensional graphical models underlying non-Gaussian distributions using transportation of measures. Unlike previous methods that estimate the joint distribution all at once, L-SING constructs a generalized precision matrix by learning each variable's neighborhood independently. This local approach allows for parallelization, making L-SING computationally tractable (e.g., more memory efficient than global methods) in high-dimensional settings. We have also shown the broad applicability of L-SING by establishing theoretical connections to existing methods and through empirical comparisons. Future work involves extending L-SING to handle mixed continuous and discrete variables and developing thresholding and scoring strategies to reduce sensitivity to tuning parameters, as done in Zhao et al. (2024) for the Gaussian setting.

Acknowledgments

SL acknowledges support from the Citadel Global Fixed Income SURF Endowment. YM acknowledges support from DOE ASCR award DE-SC0023187 and from the Office of Naval Research under award N00014-20-1-259. RB is grateful for support from the von Kármán instructorship at Caltech, the Air Force Office of Scientific Research MURI on "Machine Learning and Physics-Based Modeling and Simulation" (award FA9550-20-1-0358) and a Department of Defense (DoD) Vannevar Bush Faculty Fellowship (award N00014-22-1-2790) held by Andrew M. Stuart.

References

Banerjee, O.; Ghaoui, L. E.; and d'Aspremont, A. 2008. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, 9(15): 485–516.

Baptista, R.; Marzouk, Y.; and Zahm, O. 2023. On the Representation and Learning of Monotone Triangular Transport Maps. *Foundations of Computational Mathematics*.

Baptista, R.; Morrison, R.; Zahm, O.; and Marzouk, Y. 2024. Learning Non-Gaussian Graphical Models via Hessian Scores and Triangular Transport. *Journal of Machine Learning Research*, 25(85): 1–46.

Bresler, G. 2015. Efficiently Learning Ising Models on Arbitrary Graphs. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, 771–782. New York, NY, USA: Association for Computing Machinery. ISBN 9781450335362.

Dong, H.; and Wang, Y. 2022. Nonparametric Neighborhood Selection in Graphical Models. *Journal of Machine Learning Research*, 23: 1–36. Submitted 2/22; Revised 7/22; Published 10/22.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2007. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3): 432–441.

Ganzfried, B. F.; Riester, M.; Haibe-Kains, B.; Risch, T.; Tyekucheva, S.; Jazic, I.; Wang, X. V.; Ahmadifar, M.; Birrer, M. J.; Parmigiani, G.; Huttenhower, C.; and Waldron, L. 2013. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database (Oxford)*, 2013: bat013.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press. ISBN 9780262013192.

Liu, H.; Lafferty, J.; and Wasserman, L. 2009. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 10(80): 2295–2328.

Loh, P.-l.; and Wainwright, M. J. 2012. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In Pereira, F.; Burges, C.; Bottou, L.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Marko, N. F.; and Weil, R. J. 2012. Non-Gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. *PLoS One*, 7(10): e46935. Epub 2012 Oct 31.

Marquez, R. T.; Baggerly, K. A.; Patterson, A. P.; Liu, J.; Broaddus, R.; Frumovitz, M.; Atkinson, E. N.; Smith, D. I.; Hartmann, L.; Fishman, D.; Berchuck, A.; Whitaker, R.; Gershenson, D. M.; Mills, G. B.; Jr, R. C. B.; and Lu, K. H. 2005. Patterns of gene expression in different histotypes of epithelial ovarian cancer correlate with those in normal fallopian tube, endometrium, and colon. *Clinical Cancer Research*, 11(17): 6116–6126.

Marzouk, Y.; Moselhy, T.; Parno, M.; and Spantini, A. 2016. *Sampling via Measure Transport: An Introduction*, 1–41. Springer International Publishing. ISBN 9783319112596.

Meinshausen, N.; and Bühlmann, P. 2006. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3).

Morrison, R.; Baptista, R.; and Basor, E. 2022. Diagonal nonlinear transformations preserve structure in covariance and precision matrices. *Journal of Multivariate Analysis*, 190: 104983.

Morrison, R.; Baptista, R.; and Marzouk, Y. 2017. Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pearl, J.; and Paz, A. 1986. GRAPHOIDS: Graph-Based Logic for Reasoning about Relevance Relations. *Probabilistic and Causal Inference*.

Rosasco, L.; Villa, S.; Mosci, S.; Santoro, M.; and verri, A. 2012. Nonparametric sparsity and regularization. arXiv:1208.2572.

Shutta, K. H.; Vito, R. D.; Scholtens, D. M.; and Balasubramanian, R. 2022. Gaussian graphical models with applications to omics analyses. *Statistics in Medicine*, 41(25): 5150–5187.

Spantini, A.; Bigoni, D.; and Marzouk, Y. 2018. Inference via Low-Dimensional Couplings. *Journal of Machine Learning Research*, 19(66): 1–71.

Spizzo, G.; Went, P.; Dirnhofer, S.; Obrist, P.; Moch, H.; Baeuerle, P. A.; Mueller-Holzner, E.; Marth, C.; Gastl, G.; and Zeimet, A. G. 2006. Overexpression of epithelial cell adhesion molecule (Ep-CAM) is an independent prognostic marker for reduced survival of patients with epithelial ovarian cancer. *Gynecologic Oncology*, 103(2): 483–488.

Wehenkel, A.; and Louppe, G. 2019. Unconstrained Monotonic Neural Networks. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhao, B.; Zhai, P. S.; Wang, Y. S.; and Kolar, M. 2024. High-dimensional Functional Graphical Model Structure Learning via Neighborhood Selection Approach. arXiv:2105.02487.