Learning causal graphs via nonlinear sufficient dimension reduction

Eftychia Solea
Queen Mary University of London

Joint work with Bing Li (Penn State, USA) and
Kyongwon Kim (Yonsei University, South Korea)

2 May, 2026

Outline

Outline:

- Directed acyclic graphical (DAG) models
- Methods for estimating Gaussian and non-Gaussian DAGs
- Sufficient dimension reduction
- Estimating DAGs via nonlinear sufficient dimension reduction

Directed graphical models

- Inferring cause-effect relationships from observational data is fundamental in many scientific disciplines, such as epidemiology, neuroscience, genetics, sociology, and finance.

- The underlying causal relationships among the data are often encoded in a directed acyclic graph (DAG).

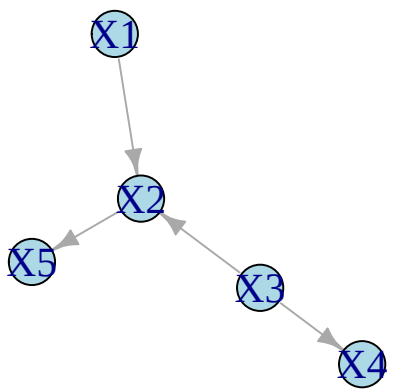
Directed acyclic graphical models

- Mathematically, let $X = (X^1, \ldots, X^p)$ be a random vector.
- Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be a graph, where $\mathsf{V} = \{1, \ldots, p\}$ is the set of nodes that corresponds to the components of X and $\mathsf{E} \subseteq \{(i, j) \in \mathsf{V} \times \mathsf{V}, i \neq j\}$ is the set of edges.
- A graph G is called a directed graph if for any $(i, j) \in \mathsf{V} \times \mathsf{V}$, with $i \neq j$, at most one of the ordered pairs (i, j) or (j, i) belongs to E.
- Graphs containing no cycles are called acyclic.

Toy example: a 5-node DAG

- To illustrate consider the DAG with $p = 5$ nodes and the directed edge set

$$\mathsf{E} = \{(1,2),\ (3,2),\ (2,5),\ (3,4)\}.$$



- **The statistical goal is to estimate the directed edges in** E **using a sample of observations from** X**.**

Directed acyclic graphical models

- When estimating a DAG (Directed Acyclic Graph) from data, one of the common assumptions is the faithfulness assumption.

- The joint distribution P_X of X is said to be faithful with respect to G if for any $i, j \in \mathrm{V}$ with $i \neq j$ and any set $S \subseteq \mathrm{V} \setminus \{i, j\}$,

  nodes i and j are d-separated by the set S in G $\iff$ $X^i \perp\!\!\!\perp X^j \,|\, X^S$,

  where $X^S = \{X^k : k \in S\}$.

Directed acyclic graphs (DAGs)

- Two DAGs that encode the same set of d-separation relations are indistinguishable based on conditional independence information alone.

- Consider the following two DAGs:

$$\text{(a)} \quad X_1 \to X_2 \to X_3 \qquad\qquad \text{(b)} \quad X_1 \leftarrow X_2 \to X_3$$

- Both DAGs encode the same conditional independence relation:

$$X_1 \perp\!\!\!\perp X_3 \mid X_2$$

- X_1 and X_3 become independent after conditioning on X_2
- Therefore, these two DAGs are indistinguishable using conditional independence information alone.

Directed acyclic graphs (DAGs)

- Therefore, from observational data, we typically cannot identify the exact DAG.
- We can only identify the Markov equivalence class of the true DAG G.
- The Markov equivalence class of G is the set of all DAGs that share the same d-separation structure (conditional independence relationships).

Example: Markov equivalence class

- Consider the DAG:

$$X_1 \to X_2 \to X_3$$

- Its Markov equivalence class consists of all DAGs that encode the same conditional independence relation:

$$X_1 \perp\!\!\!\perp X_3 \mid X_2$$

- In this case, the equivalence class is:

$$X_1 \to X_2 \to X_3,$$
$$X_1 \leftarrow X_2 \leftarrow X_3,$$
$$X_1 \leftarrow X_2 \to X_3.$$

- All these DAGs imply that X_1 and X_3 become independent after conditioning on X_2.

Characterization of Markov equivalence

- Two DAGs G_1 and G_2 are Markov equivalent if and only if
  - they have the same skeleton: the undirected graph obtained by ignoring edge directions
  - they have the same set of *v*-structures: A triple (i, k, j) forms a v-structure if

    $$i \rightarrow k \leftarrow j$$

    and i and j are not adjacent.

Example: Markov equivalence class

- Consider the true DAG, G:

$$X_1 \to X_2 \to X_3$$

- Skeleton:

$$X_1 - X_2 - X_3$$

- No v-structures.
- The Markov equivalence class of G consists of all DAGs with the same skeleton and no v-structures:

$$X_1 \to X_2 \to X_3,$$
$$X_1 \leftarrow X_2 \leftarrow X_3,$$
$$X_1 \leftarrow X_2 \to X_3.$$

- All these DAGs encode the same *d*-separation (conditional independence) relation:

$$X_1 \perp\!\!\!\perp X_3 \mid X_2.$$

Completed partially directed acyclic graphs (CPDAGs)

- A Markov equivalence class of DAGs can be uniquely represented by a **completed partially directed acyclic graph (CPDAG)**, which may contain both directed and undirected edges.
- In a CPDAG:
  - A directed edge $i \to j$ appears if and only if this orientation is the same in *every* DAG in the Markov equivalence class.
  - An undirected edge $i - j$ appears if both orientations $i \to j$ and $j \to i$ occur in different DAGs within the equivalence class.
- The CPDAG encodes exactly the set of conditional independence (d-separation) relations shared by all DAGs in the Markov equivalence class.

CPDAG representation

- **Example:** For the equivalence class

$$\{X_1 \to X_2 \to X_3,\ X_1 \leftarrow X_2 \leftarrow X_3,\ X_1 \leftarrow X_2 \to X_3\},$$

- All DAGs in the corresponding Markov equivalence class imply:

$$X_1 \perp\!\!\!\perp X_3 \mid X_2.$$

- The CPDAG is:

$$X_1 - X_2 - X_3$$

- The CPDAG encodes this conditional independence, but does not determine the direction of the edges.

Estimation of DAGs

- Under the **faithfulness assumption** with respect to the true DAG G, the goal is to recover the **Markov equivalence class** of G (or equivalently its CPDAG) from observational data.
- Constraint-based methods, such as the PC algorithm (Spirtes et al., 2000), estimate the CPDAG by performing a sequence of conditional independence tests.

Estimation of DAGs

The PC algorithm

The PC algorithm estimates the CPDAG in two steps

- **Step 1: Estimate skeleton**
  - Performs a sequence of conditional independence tests to estimate the skeleton of the true DAG. In particular,

$$i \text{ and } j \text{ are not connected in } \mathrm{ske}(\mathsf{G}) \iff X^i \perp\!\!\!\perp X^j \mid X^S$$
$$\text{for some } S \subset \mathsf{V} \setminus \{i, j\},$$

- **Step 2: Partial edge orientation**
  - Then, in the second step, the algorithm uses the results from the first step to determine a partial orientation of the edges. In particular, identify v-structures and apply orientation rules.

Estimation of Gaussian DAGs

- Many approaches have been proposed to estimate the CPDAG of the true DAG under the assumption that $\mathbf{X}$ is a Gaussian random vector.
- Kalisch and Bühlman (2007) studied the PC algorithm for estimating high-dimensional Gaussian DAGs.

Gaussian DAGs

- Advantage of GGM: The equivalence

$$X^i \perp\!\!\!\perp X^j \,|\, X^S \iff \rho_{ij|S} = 0,$$

where $\rho_{ij|S}$ is the partial correlation between X^i and X^j given X^S.

- Therefore, under a Gaussian model, edges in the skeleton satisfy

$$i \text{ and } j \text{ are not connected in } \mathrm{ske}(\mathsf{G}) \iff \rho_{ij|S} = 0$$

Gaussian DAGs

- Disadvantage of GGM: The Gaussian assumption is very restrictive. For example,
  - The data are skewed.
  - There are nonlinear or heteroscedastic relations among the data.

Estimation of DAGs

- Learning a DAG typically requires repeated tests of conditional independence:

$$X_i \perp\!\!\!\perp X_j \mid X_S.$$

- **Challenge:** the conditioning set S can be large, which makes estimation unstable and inefficient in high dimensions.

Key challenge: develop methods that relax Gaussian assumptions while avoiding high-dimensional conditioning sets, which can reduce estimation accuracy.

Non-gaussian DAGs

- Harris and Drton (2013) propose the Gaussian copula DAG model using rank correlations.
  $\Rightarrow$ Disadvantage. The Gaussian copula assumption can be violated under nonlinear interactions.

Non-gaussian DAGs

- To strike a balance between model flexibility and dimensionality, Lee et al. (2020) proposed a fully nonparametric estimation approach based on additive conditional independence (ACI) Li et al. (2014).

- ACI is a three-way statistical relation similar to conditional independence, but its implementation relies on one-dimensional kernels, thereby mitigating the curse of dimensionality.
  $\Rightarrow$ Disadvantages: ACI can be difficult to interpret and relies on additive structure assumptions, which may be restrictive in practice.

Our proposal and contributions

1. We introduce a new nonparametric method for estimating a DAG with the PC-algorithm.

2. To balance model flexibility and scalability, we apply nonlinear sufficient dimension reduction (SDR) to replace the conditioning set X_S with a lower-dimensional representation.

3. The reduced representation is then used within the conditional cross-covariance operator to evaluate conditional independence and estimate the skeleton of the graph.

4. We illustrate our methodology by simulation and a real data analysis.

General Framework of SDR

Linear Sufficient Dimension Reduction (SDR)

- Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$. Linear SDR seeks a matrix $\beta \in \mathbb{R}^{p \times d}$, $d \leq p$, such that

$$Y \perp\!\!\!\perp X \mid \beta_1^\top X, \ldots, \beta_d^\top X$$

- The column space $\mathcal{S}(\beta) = \mathrm{span}\{\beta_1, \ldots, \beta_d\}$ is called a **dimension reduction subspace**.
- The intersection of all such subspaces is the **central subspace**, denoted by $\mathcal{S}_{Y|X}$.
- The goal of linear SDR is to estimate $\mathcal{S}_{Y|X}$ or a basis thereof.

General Framework of SDR

Linear SDR: Methods

- Several classical estimators of $\mathcal{S}_{Y|X}$ include:
  - Sliced Inverse Regression (SIR) (Li, 1991)
  - Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991)
- These methods exploit low-dimensional structure in conditional moments of $X \mid Y$.

General Framework of SDR

Nonlinear sufficient dimension reduction

- Linear SDR extends to a nonlinear setting by replacing linear projections with measurable functions (Lee et al., 2013; Li et al., 2011).
- Let $\mathcal{H}_X$ be a Hilbert space of measurable functions on the support of X.
- Nonlinear SDR seeks functions $f_1, \ldots, f_d \in \mathcal{H}_X$ such that

$$Y \perp\!\!\!\perp X \mid f_1(X), \ldots, f_d(X).$$

General Framework of SDR

Nonlinear sufficient dimension reduction

- Define the central σ-field

$$\mathcal{G}_{Y|X} = \sigma\{f_1(X), \ldots, f_d(X)\}.$$

- Let $\mathcal{H}_X(\mathcal{G}_{Y|X})$ denote the subspace of $\mathcal{H}_X$ spanned by the functions in $\mathcal{H}_X$ that are $\mathcal{G}_{Y|X}$-measurable functions.

- This subspace is called the **central class**, denoted by

$$\mathfrak{S}_{Y|X}.$$

- The goal of nonlinear SDR is to estimate $\mathfrak{S}_{Y|X}$ or a basis for it.

General framework of SDR

Nonlinear SDR

- Commonly used nonlinear SDR estimation methods:
  - GSIR and GSAVE (Lee et al., 2013)
  - Kernel dimension reduction methods, such as the KCCA (Fukumizu et al., 2007; Wu, 2008) and the kernel sliced inverse regression.

- Our goal: Connect nonlinear SDR with DAG estimation, enabling more flexible and efficient inference of causal structures

Nonlinear SDR and DAGs

- **Assumption 1:** For each $i, j \in \mathrm{V}$ with $i \neq j$, and for any $S \subseteq \mathrm{V} \setminus \{i, j\}$, there exist measurable functions $f_1^{ij}, \ldots, f_d^{ij}$ such that

$$(X_i, X_j) \perp\!\!\!\perp X_S \mid f_1^{ij}(X_S), \ldots, f_d^{ij}(X_S). \tag{1}$$

- In other words, the conditional distribution of $(X_i, X_j) \mid X_S$ depends on X_S only through the lower-dimensional representation $U^{ij,S} = \left(f_1^{ij}(X_S), \ldots, f_d^{ij}(X_S)\right)$.

- Replace X^S with $U^{ij,S}$ without loss of information on conditional distribution of (X^i, X^j) given X^S.

Nonlinear SDR and DAGs

- If Assumption 1 holds, then

$$X_i \perp\!\!\!\perp X_j \mid X_S \iff X_i \perp\!\!\!\perp X_j \mid U^{ij,S},$$

  where $U^{ij,S} = \left(f_1^{ij}(X_S), \ldots, f_d^{ij}(X_S)\right)$.

- So we reduce a high-dimensional conditioning problem to a low-dimensional one.

Nonlinear SDR and DAG

- Under Assumption 1

$$X_i \perp\!\!\!\perp X_j \mid X_S \iff X_i \perp\!\!\!\perp X_j \mid U^{ij,S}, \tag{2}$$

- Under faithfulness,

$$i \text{ and } j \text{ are not connected in } \mathrm{ske}(\mathsf{G}) \iff X^i \perp\!\!\!\perp X^j \mid X^S \quad \text{for some } S \subset \mathsf{V} \setminus \{i,j\}, \tag{3}$$

Nonlinear SDR and DAG

- By (2) and (3), we can use the criterion

$$X_i \perp\!\!\!\perp X_j \mid U^{ij,S}$$

to infer the skeleton of the DAG G after applying nonlinear SDR to (X_i, X_j) given X_S

$$i \text{ and } j \text{ are not connected in } \mathrm{ske}(\mathsf{G}) \iff X^i \perp\!\!\!\perp X^j \mid U^{ij,S} \text{ for some } S \subset \mathsf{V} \setminus \{i,j\}\}.$$

- we recover the skeleton of the DAG using low-dimensional conditioning instead of high-dimensional conditioning.

Nonlinear SDR and DAGs

The result suggests a two-step strategy for DAG learning:

- **Step 1: Nonlinear SDR.** For each pair (i, j) and conditioning set $S \subseteq \mathrm{V} \setminus \{i, j\}$, apply nonlinear SDR to construct

$$U_{ij,S} = \left(f_1^{ij}(X_S), \ldots, f_d^{ij}(X_S)\right),$$

  which provides a low-dimensional representation of X_S preserving information relevant for (X_i, X_j).

- **Step 2: Conditional independence testing.** Second, we perform conditional independence testing to assess

$$X_i \perp\!\!\!\perp X_j \mid U_{ij,S}.$$

- Edges are removed from the skeleton whenever conditional independence is detected, yielding an estimate of the skeleton of the DAG.

Step 1: Nonlinear SDR

- We use GSIR for nonlinear sufficient dimension reduction.
- At Step 1, our goal is to estimate functions that span the central class of (X_i, X_j) given X_S:

$$\mathfrak{S}_{(X_i,X_j)|X_S}.$$

- This central class contains all square-integrable functions of X_S that are relevant for predicting (X_i, X_j).

Step 1: Nonlinear SDR

Reproducing Kernel Hilbert Space (RKHS)

- To handle nonlinear relationships, we embed random variables into a reproducing kernel Hilbert space (RKHS).
- For X_S, we define the RKHS

$$\mathcal{H}_{X_S} = \overline{\text{span}}\{\kappa_{X_S}(\cdot, x_S) : x_S \in \Omega_{X_S}\},$$

  where $\kappa_{X_S} : \Omega_{X_S} \times \Omega_{X_S} \to \mathbb{R}$ is a positive definite kernel.
- Here Ω_{X_S} denotes the support of X_S, typically a product space $\prod_{k \in S} \Omega_{X_k}$.

Step 1: Nonlinear SDR

GSIR and regression operator

- **Assumption 2:** For all $i \neq j$ and $S \subseteq \mathsf{V} \setminus \{i, j\}$,
  - $\mathbb{E}\big[\kappa_{X_S}(X_S, X_S)\big] < \infty, \quad \mathbb{E}\big[\kappa_{X_i X_j}((X_i, X_j), (X_i, X_j))\big] < \infty.$
  - $\operatorname{ran}\big(\Sigma_{X_S,(X_i X_j)}\big) \subseteq \operatorname{ran}\big(\Sigma_{X_S X_S}\big).$
- Then covariance operators

$$\Sigma_{X_S(X_i X_j)} \in \mathcal{B}(\mathcal{H}_{X_i X_j}, \mathcal{H}_{X_S}), \quad \Sigma_{X_S X_S} \in \mathcal{B}(\mathcal{H}_{X_S})$$

  are well-defined.

Step 1: Nonlinear SDR

GSIR and regression operator

- Define the regression operator

$$B_{X_S(X_iX_j)} = \Sigma^{-1}_{X_SX_S}\Sigma_{X_S(X_iX_j)}.$$

- This operator maps $\mathcal{H}_{X_iX_j} \to \mathcal{H}_{X_S}$ and is well-defined under Assumption 2.
- It is referred to as the **GSIR regression operator**.

Step 1: Nonlinear SDR

GSIR and regression operator

- Under some mild assumptions,

$$\overline{\mathrm{ran}}\left(B_{X^s(X^iX^j)}\right) = \mathfrak{S}_{(X^iX^j)|X^s}. \tag{4}$$

- This means that the central class is fully characterised by the range of the GSIR regression operator.

Step 1: Nonlinear SDR

GSIR and regression operator

- We assume $B_{X^S(X^iX^j)}$ is a finite-rank operator with rank d_S^{ij}.
- This assumption implies:
  - the central class is finite-dimensional,
  - estimation can be reduced to a finite-dimensional eigenvalue problem,
  - the range of $B_{X_S,(X_iX_j)}$ can be recovered via $d_{ij,S}$ leading eigenfunctions.

Step 1: Nonlinear SDR

Recovering the central class

- A basis $\{f_1^{ij}, \ldots, f_{d_S^{ij}}^{ij}\}$ of the central class, $\mathfrak{S}_{(X^iX^j)|X^S}$, can be found by solving the following iterative maximization problem for each $i, j \in \mathrm{V}$ and $S \subset \mathrm{V} \setminus \{i, j\}$.
- For each $k = 1, \ldots, d_S^{ij}$

$$\begin{aligned} \text{maximize} \quad & \langle f, \Sigma_{X^S(X^iX^j)} \Sigma^{-1}_{(X^iX^j)(X^iX^j)} \Sigma_{(X^iX^j)X^S} f \rangle \\ \text{subject to} \quad & f \in \mathcal{H}_{X^S}, \langle f, \Sigma_{X^SX^S} f \rangle = 1, \\ & \langle f, \Sigma_{X^SX^S} f_1 \rangle = \cdots = \langle f, \Sigma_{X^SX^S} f_{k-1} \rangle = 0, \end{aligned} \tag{5}$$

Step 1: Nonlinear SDR

To summarize, our goal consists of two steps:

- **Step 1:** For each $i, j \in \mathrm{V}$ and $S \subseteq \mathrm{V} \setminus \{i, j\}$, solve (5) to obtain the basic functions $f_1^{ij}(X_S), \ldots, f_{d_{ij,S}}^{ij}(X_S)$ of the central class $\mathfrak{S}_{(X_i, X_j)|X_S}$.
- Let
$$U_{ij,S} = (f_1^{ij}(X_S), \ldots, f_{d_{ij,S}}^{ij}(X_S))$$
- **Step 2** is to test the conditional independence relation
$$X_i \perp\!\!\!\perp X_j \mid U_{ij,S}.$$
and hence the skeleton of the graph.

Step 2: Conditional independence testing

Conjoined conditional cross-covariance operator and conditional independence

- The conjoined conditional cross-covariance operator (CCCO) is defined as:

$$\Sigma_{X_i X_j | U_{ij,S}} = \Sigma_{X_i X_j} - \Sigma_{X_i U_{ij,S}} \Sigma^{-1}_{U_{ij,S} U_{ij,S}} \Sigma_{U_{ij,S} X_j}.$$

- Under some assumptions, the CCCO satisfies the key property

$$\Sigma_{X_i X_j | U_{ij,S}} = 0 \iff X_i \perp\!\!\!\perp X_j \mid U_{ij,S}.$$

Step 2: Conditional independence testing

CCCO and skeleton recovery

- This results allow us to use CCCO to evaluate conditional independence and the skeleton of the true DAG, G

$$(i,j) \notin \mathrm{ske}(\mathsf{G}) \iff \exists S \subseteq \mathsf{V} \setminus \{i,j\} \text{ such that } \|\Sigma_{X_i X_j | U_{ij,S}}\|_{\mathrm{HS}} = 0.$$

Simulation studies

1. We evaluate the performance of our DAG estimator, referred to as the DAG-PC algorithm, through simulation comparisons with other methods and a data application.

2. We compare it with three existing PC algorithms:

Method A : Gaussian-PC algorithm

Method B : Rank-PC algorithm

Method C : Kernel-PC algorithm

Method D : DAG-PC algorithm.

Simulation studies

1. We use the structural Hamming distance (SHD; Tsamardinos et al., 2006) to measure the efficiency of estimating the CPDAG.

2. We use the area under the curve (AUC) of the receiver operating characteristic curve (ROC) to measure the estimation efficiency of the skeleton

Simulation studies

1. We generate a $p \times p$ dimensional adjacency matrix D as follows:

2. After deciding on the topological ordering among nodes, we fill the lower-diagonal elements of D with 0s or 1s according to the Bernoulli distribution with sparsity parameter $s \in (0, 1)$.

3. The expected number of neighbors of each node i, denoted by $\mathbb{E}[N_i]$, is $s(p-1)$.

Simulation studies

1. Given the adjacency matrix D, we generate a p-dimensional random vector $X = (X^1, \ldots, X^p)$ sequentially via

$$\epsilon^1, \ldots, \epsilon^p \overset{\text{i.i.d}}{\sim} N(0,1),$$

$$X^1 = \epsilon^1, \quad X^i = \sum_{j=1}^{i-1} d_{ij} \cos(X^j) + \epsilon^i,$$

   where d_{ij} is the (i,j)-th element of the adjacency matrix D.
2. We choose the sample size n to be 100, 150 and 200 and the network size p to be 5, 10 and 50.
3. We consider two scenarios with $E[N_i]$ are 2 and 4, respectively.

Simulation results: Estimation of CPDAG

A smaller SHD means more accurate estimation of the CPDAG

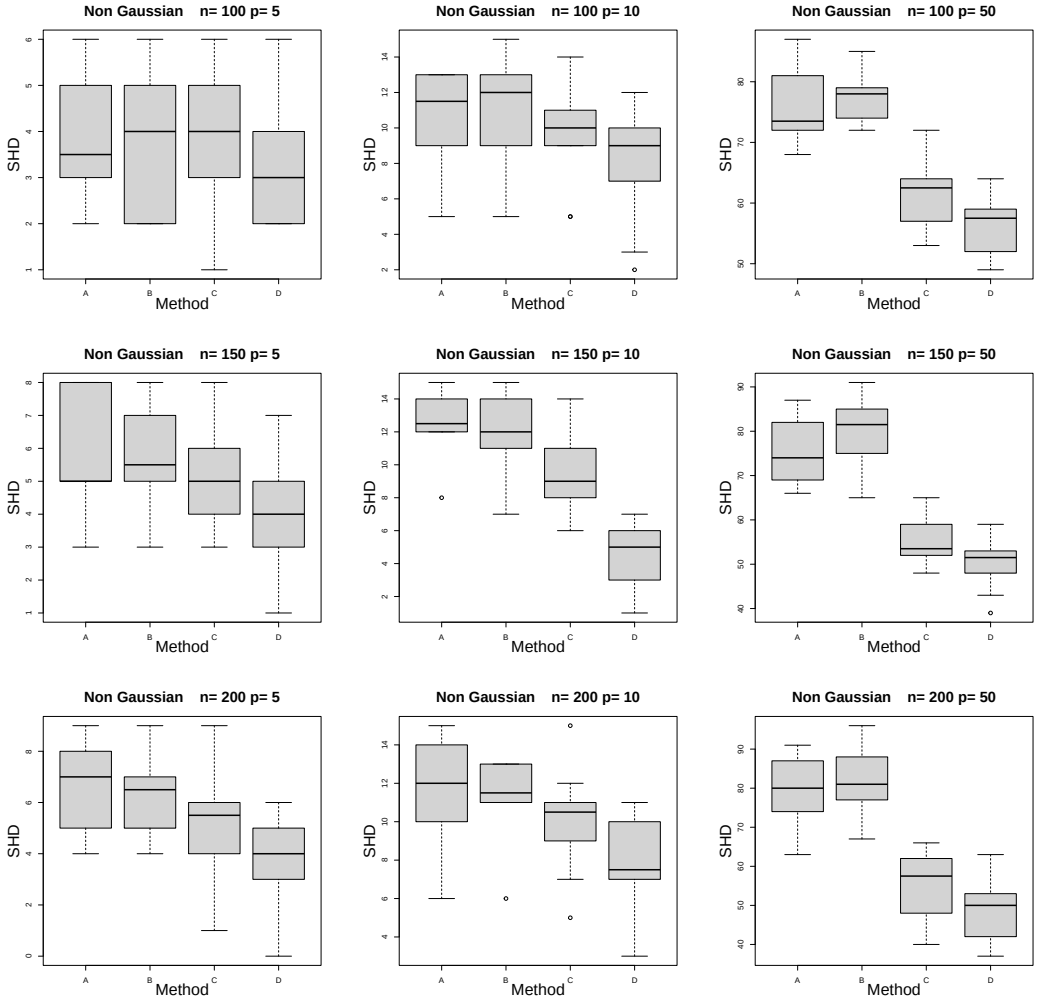


Figure: Comparisons of SHD values for non-Gaussian settings under $E[N_i] = 2$.

Simulation results: Estimation of the skeleton

A smaller AUS means more accurate estimation of the skeleton

Non-Gaussian		$E[N_i] = 2$		
n	Method	$p = 5$	$p = 10$	$p = 50$
100	A	0.57(0.09)	0.54(0.03)	0.53(0.01)
	B	0.55(0.09)	0.53(0.03)	0.53(0.01)
	C	0.59(0.09)	0.57(0.02)	0.53(0.01)
	D	0.62(0.08)	0.57(0.04)	0.56(0.01)
150	A	0.52(0.03)	0.52(0.03)	0.55(0.02)
	B	0.51(0.02)	0.53(0.05)	0.55(0.01)
	C	0.59(0.08)	0.61(0.05)	0.58(0.02)
	D	0.61(0.11)	0.61(0.07)	0.60(0.01)
200	A	0.54(0.06)	0.53(0.03)	0.54(0.01)
	B	0.54(0.06)	0.53(0.03)	0.54(0.01)
	C	0.66(0.13)	0.59(0.03)	0.61(0.02)
	D	0.70(0.11)	0.61(0.04)	0.63(0.02)

Table: Comparison of AUC values between the true and estimated graph for non-Gaussian settings with $E[N_i] = 2$.

Data application

1. We apply our method to the flow cytometry dataset (Sachs et al., 2005), which includes simultaneously measured $p = 11$ phosphorylated phosphoproteins and phospholipids in the single cell.

2. The goal of this application was to demonstrate that our method can accurately discover causal relationships in the latent signaling network.

Data application

1. To demonstrate the effectiveness of our approach, we applied Methods A, B, C, and D to $n = 90$ observations, and compared the estimated CPDAG with the true CPDAG.

2. The true DAG can be found in (Sachs et al., 2005).

Data application

1. We repeated this subsampling procedure 10 times and reported the mean and standard deviation of the SHD values in Table 2.

2. Our proposed method (Method D) has the lowest SHD values, indicating its competitiveness for investigating causal relations among cells.

Method	A	B	C	D
SHD	21.2(1.31)	23.6(1.07)	21.1(1.91))	20.6(0.79)

Table: Comparison of the mean and standard deviation of SHD values among four methods, with the standard deviation presented in parentheses.

References

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(14):361–383.

Harris, N. and Drton, M. (2013). Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11).

Kalisch, M. and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3).

Lee, K.-Y., Li, B., and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41(1):221–249.

Lee, K.-Y., Liu, T., Li, B., and Zhao, H. (2020). Learning causal networks via additive faithfulness. *Journal of Machine Learning Research*, 21(51):1–38.

Li, B., Artemiou, A., and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210.

Li, B., Chun, H., and Zhao, H. (2014). On an additive semigraphoid model for statistical networks with application to pathway analysis. *Journal of the American Statistical Association*, 109(507):1188–1204.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Spirtes, P., Glymour, C., and Scheines, R. (2000). Causation, prediction, and search. adaptive computation and machine learning series. *The MIT Press*, 49:77–78.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78.

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610.

Toy example: Markov equivalent DAGs

- Consider the following two DAGs:

$$\text{(a)} \quad X_1 \to X_2 \to X_3 \qquad\qquad \text{(b)} \quad X_1 \leftarrow X_2 \to X_3$$

- Both DAGs encode the same d-separation relation:

$$X_1 \perp\!\!\!\perp X_3 \mid X_2$$

- In both cases:
  - X_1 and X_3 are dependent marginally
  - X_1 and X_3 become independent after conditioning on X_2
- Therefore, these two DAGs are indistinguishable using conditional independence information alone.

EXTRA SLIDES: ASSUMPTIONS

1. For any $S \subseteq \mathsf{V}$, $\mathcal{H}_{X^S}$ is a dense subset of $L_2(P_{X^S})$ modulo constants; that is for any $f \in L_2(P_{X^S})$ and any $\epsilon > 0$, there exists a $g \in \mathcal{H}_{X^S}$ such that $\mathrm{var}[f(X^S) - g(X^S)] < \epsilon$.
2. Assumption ensures the kernel function, κ_{X^S}, to be sufficiently rich so that it is a characteristic kernel with respect to $L_2(P_{X^S})$. This means that projections onto $L_2(P_{X^S})$ can be well approximated by elements in the RKHS, $\mathcal{H}_{X^S}$. This assumption is satisfied by the Gaussian RBF kernel, but not by the polynomial kernel.

EXTRA SLIDES: ASSUMPTIONS

1. Let $L_2(P_{X^S})$ be the L_2-space with respect to the distribution P_{X^S} of X^S. Similar to Lee et al. (2013), we assume that the central class, $\mathfrak{S}_{(X^i X^j)|X^S}$, is a complete class which means that for any $\mathcal{G}_{(X^i X^j)|X^S}$-measurable $f \in L_2(P_{X^S})$ such that $\mathbb{E}(f(X^S)|(X^i, X^j)) = 0$ almost surely, $f(X^S) = 0$ almost surely.