Causal Discovery from Multiple Data Sets with Non-Identical Variable Sets

Biwei Huang,[1] **Kun Zhang,**[1] **Mingming Gong,**[2] **Clark Glymour**[1]
[1]Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA
[2]School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia
{biweih, cg09}@andrew.cmu.edu, kunz1@cmu.edu, mingming.gong@unimelb.edu.au

Abstract

A number of approaches to causal discovery assume that there are no hidden confounders and are designed to learn a fixed causal model from a single data set. Over the last decade, with closer cooperation across laboratories, we are able to accumulate more variables and data for analysis, while each lab may only measure a subset of them, due to technical constraints or to save time and cost. This raises a question of how to handle causal discovery from multiple data sets with non-identical variable sets, and at the same time, it would be interesting to see how more recorded variables can help to mitigate the confounding problem. In this paper, we propose a principled method to *uniquely* identify causal relationships over the integrated set of variables from multiple data sets, in linear, non-Gaussian cases. The proposed method also allows distribution shifts across data sets. Theoretically, we show that the causal structure over the integrated set of variables is identifiable under testable conditions. Furthermore, we present two types of approaches to parameter estimation: one is based on maximum likelihood, and the other is likelihood free and leverages generative adversarial nets to improve scalability of the estimation procedure. Experimental results on various synthetic and real-world data sets are presented to demonstrate the efficacy of our methods.

1 Introduction

Learning causal relationships is one of the fundamental tasks in scientific developments. As it is often difficult to carry out randomized experiments, inferring causal relations from purely observational data, known as causal discovery, has drawn much attention. Most approaches in causal discovery are designed to learn causal relationships from a single data set generated by a fixed causal model, and they often assume that there are no hidden confounders (Spirtes, Glymour, and Scheines 1993; Heckerman, Geiger, and Chickering 1995; Shimizu et al. 2006; Zhang and Hyvärinen 2009), with several exceptions. Dealing with the confounding problem is challenging–much effort has been made, but existing methods usually resort to rather strong assumptions (e.g., graph structure constraints over observed variables or latent variables), and the estimated graphs usually have large indeterminacies (Spirtes, Glymour, and Scheines 1993; R. Silva and Spirtes 2006; Hoyer et al. 2008; Anandkumar et al. 2013).

Over the last decade, thanks to the improved ability to collect and store "big data" and closer cooperation across laboratories, we are able to accumulate more variables and data for analysis, while each lab may only measure a subset of them. For instance, in biology, to understand the gene expression process, each researcher or each lab is responsible for measuring a part of the genes or proteins. In electrophysiology, in each experiment, an electrode might only record activities generated by one or several nearby neurons, and it impedes recording and analyzing information flows between distinct areas. A solution is to record a subset of target areas in each experiment and repeat the experiments several times.

Thus, it is desirable to perform causal discovery from multiple data sets with non-identical sets of variables. A typical proposal to solve the problem is to analyze each data set separately and then integrate the results. However, this procedure may again suffer the confounding problem due to missing variables in each data set. In addition, it ignores the information shared by different data sets and may suffer low statistical reliability, especially when the sample size from each data set is small while the data dimension is high.

Several approaches have been proposed to dealing with the above scenario, such as the Integration of Overlapping Networks (ION) algorithm (Danks, Glymour, and Tillman 2009) and Causal discovery from Overlapping INtErventions (COmbINE) (Triantafillou and Tsamardinos 2015). They use conditional independence information to find a partial ancestral graph (PAG) on each data set independently and then search for PAGs over the integrated set of variables that are consistent with the input PAGs from all data sets. With this line of approaches, the final causal graph over the integrated set of variables is generally not unique and usually has large indeterminacies. Moreover, statistical inefficiency may affect the estimation of individual PAGs, especially in the case when the sample sizes of some data sets are small, and further affect the accuracy of final results.

In this paper, we focus on the linear, non-Gaussian case, where the causal relations are linear and noise variables are non-Gaussian. Interestingly, in this case, we are able to develop a principled method to *uniquely* identify causal relationships over the integrated set of variables from multiple data sets, called Causal Discovery from Multiple data sets with Non-identical variables (CD-MiNi). It also allows distribution shifts across data sets (some approaches can handle

distribution shifts, but they assume the same variable sets across data sets (Zhang et al. 2017; Ghassami et al. 2018; Huang et al. 2019a; 2019b)). Our main contributions are as follows:

- To the best of our knowledge, the proposed CD-MiNi is the first that can identify the entire directed acyclic causal structure from non-identical sets of variables.
- We show the theoretical identifiability of the causal structure under testable conditions.
- We present two practical estimation approaches. One exploits the Expectation-Maximization (EM) algorithm to maximize the likelihood. The other adopts generative adversarial nets, which is likelihood free and improves scalability of the algorithm.
- We further extend our approach to confounding cases, where the integrated set of variables is not causally sufficient (i.e., some direct common causes of two variables in this set are not in the set), and to cyclic cases.

2 Motivation and Related Work

Identification of causal relationships from observational data is attractive for the reason that traditional randomized experiments may be hard or even impossible to do. Over the past decades, prominent progress has been made in this area. Constraint-based and score-based methods make use of conditional independence constraints to find causal skeleton and determine orientations up to the Markov equivalence class (Spirtes, Glymour, and Scheines 1993; Heckerman, Geiger, and Chickering 1995; Meek 1997; Chickering 2003; Huang et al. 2018). It was later shown that with functional causal model-based approaches, it is possible to recover the whole causal graph under certain constraints on the functional class of causal mechanisms, by making use of asymmetries between causal and anti-causal directions. For example, in the case of linear causal relationships, the non-Gaussianity of noise terms helps to identify the causal direction; in the causal direction, the noise term is independent of hypothetical causes, while independence does not hold in the anti-causal direction. The linear non-Gaussian acyclic model (LiNGAM) (Shimizu et al. 2006) uses this property for causal discovery.

The above approaches are designed to do causal discovery from a single data set generated by a fixed causal model. In many cases, we aim to learn the causal model over the complete set of variables from multiple data sets, each of which records only a subset of variables. To handle this case, several approaches have been proposed. ION (Danks 2005; Danks, Glymour, and Tillman 2009) is the first algorithm proposed to do causal discovery in such cases. It first learns a PAG from each data set independently, by using constraint-based FCI (Spirtes, Glymour, and Scheines 1993) or score-based greedy equivalence search (Chickering 2003) with latent variable post-processing steps. Then the learned set of PAGs are input to an integration procedure, to learn PAGs over the integrated set of variables, which are consistent (i.e., with the same d-separation and d-connection relations) with the input graphs from individual data sets. The output of ION may contain multiple PAGs. Moreover, in practice with finite samples, the structures integrated may entail contradictory conditional independences and dependences.

COmbINE (Triantafillou and Tsamardinos 2015) and SAT-based causal discovery approaches (Hyttinen et al. 2013; Hyttinen, Eberhardt, and Hoyer 2012; Hyttinen, Eberhardt, and Järvisalo 2014; Tillman and Eberhardt 2014; Rantanen, Hyttinen, and Järvisalo 2018; Zhang et al. 2019) use a similar idea to ION, with the difference that they convert all constraints to SAT solvers, which are usually more efficient. The Multiple model Causal Inference (MCI) algorithm (Claassen and Heskes 2010) takes into account the variability of causal structures across experiments. The Integration of Overlapping Datasets (IOD) algorithm (Tillman and Spirtes 2011) learns equivalence classes directly from multiple datasets to avoid potential conflicts. Wang and Glymour (2019) further leveraged trek rules for better integration. Unfortunately, generally, this line of approaches may result in large indeterminacies in the estimated graph over the integrated set of variables, as noticed in Spirtes et al. (2010).

Figure 1 gives an illustration of such indeterminacies in the estimation results. Suppose there are two data sets. On data set 1, we observed x_2, x_3, and x_4, and on data set 2, we observed x_1, x_3, and x_4. Figure 1(a) gives the true causal structure over the four integrated variables. Figure 1(b1) and (b2) show the learned PAGs with FCI in the ideal case (with no statistical error) from data sets 1 and 2, respectively. Figure 1(c1)-(c8) are the output PAGs with ION, using Figure 1(b1) and (b2) as input. ION gives us 8 candidate PAGs. Even if excluding bidirected edges, there are still 5 possibilities. Hence, from the results, we can only determine that there are causal edges between x_1 and x_3, and between x_1 and x_4. We cannot be certain of the existence of other edges, let alone causal directions.

In contrast, we develop an approach which avoids such indeterminacies in the estimation results. The proposed CD-MiNi can *uniquely* identify the entire causal graph over the integrated set of variables.

3 Causal Model for Non-Identical Sets of Variables

Suppose we have data collected from M data sets. Let X^m ($m = 1, \cdots, M$) be the set of variables measured in the m-th data set, with the number of variables $|X^m| = d_m$. Each data set may contain different sets of variables but partially overlaps with each other. Let X be the union of variables from all M data sets: $X = \bigcup_{m=1}^{M} X^m$, with d variables in total.

We assume that the causal graph over X is a directed acyclic graph (DAG), and that each variable $x_i \in X$ ($i = 1, \cdots, d$) satisfies the following data generating process

$$x_i = \sum_{j \in \mathcal{P}_i} b_{ij} x_j + e_i, \tag{1}$$

where variable $x_j \in X$ is the direct cause of x_i, b_{ij} represents causal coefficient from x_j to x_i, and $\mathcal{P}_i$ is the index set of the direct causes of x_i. The noise term e_i is non-Gaussian distributed and is independent of the parents of x_i. We allow

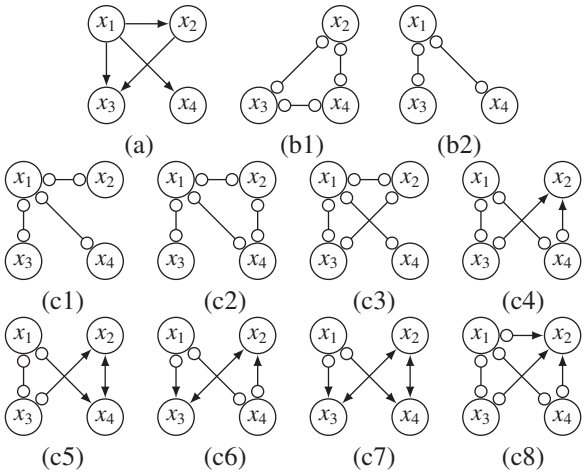


Figure 1: An example to illustrate that current approaches (e.g., ION) may introduce large indeterminacies in the causal graph over the integrated set of variables. (a) Ground truth. (b1)-(b2) Oracle PAGs from two data sets. (c1)-(c8) Output PAGs over the integrated set of variables.

that the noise distribution $p(e_i)$ varies across data sets, but the causal coefficient b_{ij}, for any i and j, is fixed. Note that in the individual data set, some x_i's are not observable, because the variable sets are not identical across data sets.

Denote by B the $d\times d$ causal adjacency matrix with entries b_{ij}. We can reorganize Eq. (1) into the matrix form:

$$X = (I - B)^{-1}E, \tag{2}$$

where $X = (x_1, \cdots, x_d)^{\mathrm{T}}$, $E = (e_1, \cdots, e_d)^{\mathrm{T}}$, and I is a $d \times d$ identity matrix. Let $A = (I - B)^{-1}$, and then $X = AE$.

Let $I^m = (I_1^m, \cdots, I_d^m)$ be a 0-1 binary vector indicating the measured variables in the m-th data set; i.e., $I_i^m = 1$ if and only if variable i is measured in the m-th data set. Furthermore, let $\mathcal{I}^m = \mathtt{diag}(I^m)$, i.e., a diagonal matrix with I^m on its diagonal. Then the m-th data set satisfies

$$X^m = A^m E^m = g(\mathcal{I}^m A)E^m, \tag{3}$$

where $A^m = g(\mathcal{I}^m A)$ is a $d_m\times d$ submatrix of A, with g removing all zero rows from matrix $\mathcal{I}^m A$, and $E^m = (e_1^m, \cdots, e_d^m)^{\mathrm{T}}$.

To remove g in the formula, equivalently, we fill in zero values to those missing variables in the m-th data set. Denote the m-th variable set after filling in zero values by $\tilde{X}^m$, and then it becomes:

$$\tilde{X}^m = \mathcal{I}^m A E^m, \qquad \text{for } m = 1, \cdots, M. \tag{4}$$

Remark: Note that with the representation in Eq. (4), we can leverage all data sets to estimate A in one step, instead of separately estimating A^m from the m-th data set, for $m = 1, \cdots, M$, and then concatenating A^m's to derive A. The advantages that our representation enjoys are mainly twofold:

1. Directly estimating A from all data sets improves computational efficiency. If we estimate each A^m separately, we need to run the estimation procedure M times (although they can be done in parallel), and furthermore, we need to find appropriate ordering and scaling of A^m's to derive A. Matching the rows in two A^m's is in principle a combinatorial problem.
2. Directly estimating the whole matrix of A also improves statistical efficiency. In the case where the sample size of individual data set is small, if we consider each data set separately, the estimation error of A^m may be large. Moreover, due to estimation errors, it may be hard to find appropriate ordering and scaling of A^m's to derive A. Instead, if we estimate A from all M data sets directly, we make use of the information shared across data sets to improve statistical efficiency. Furthermore, by estimating A in one step, it avoids possible conflicts during integration.

The detailed estimation procedure will be introduced in Section 5. In the next section, we show that the causal structure over X is identifiable under verifiable conditions.

4 Identifiability Conditions

Suppose the underlying causal graph over X is acyclic and that the variable set X is causally sufficient. In this section, we show that the causal model for non-identical variable sets, given in Eq. (3) or (4), is identifiable under verifiable conditions. Before moving forward, we first give the identifiability condition of independent component analysis (ICA).

Theorem 1 (Identifiability of ICA (Comon 1994; Eriksson and Koivunen 2004))**.** *Let X be observed variables, satisfying $X = AS$, with sources in S being independent of each other. Then the mixing matrix A can be identified up to scaling and column permutation, if one of the following two conditions is satisfied:*

1. There is no Gaussian source in S.
2. A is full column rank and at most one source is Gaussian.

Because the noise terms in Eq. (3) are independent of each other, and $|E^m| > |X^m|$, Eq. (3) satisfies the over-complete ICA model. For over-complete ICA, where the number of observed variables is smaller than that of sources, condition 2 is not satisfied, and thus to achieve identifiability, a natural way is to assume that there is no Gaussian source in S.

Let m_1 and m_2 be two data sets, and let $X_{\mathbf{i}} = X^{m_1} \cap X^{m_2}$, where $\mathbf{i}$ is the index of overlapping variables of m_1 and m_2. We make the following assumption on the mixing matrix A.

Assumption 1. *For any two overlapping data sets m_1 and m_2 with the number of overlapping variables $|\mathbf{i}| \geq 2$, there exist two different $i_1, i_2 \in \mathbf{i}$, so that $a_{i_1 j}a_{i_2 k} \neq a_{i_1 k}a_{i_2 j}$, for any $j \neq k$, where $a_{i_1 j}$, $a_{i_2 j}$, $a_{i_1 k}$, and $a_{i_2 k}$ are entries in A.*

This assumption removes ambiguity when integrating different data sets. The validity of the assumption can be checked after identifying A. Now we are ready to give identifiability conditions of causal relations, which are stated in Theorem 2.

Theorem 2. *Suppose X satisfies the generating process in Eq. (1). Under Assumption 1, the causal graph over X is identifiable, if the collected data sets satisfy the following conditions:*

1. for any data set i, there exists another data set j which has at least two overlapping variables with i;
2. all noises are non-Gaussian.

Below, we give a sketch of the proof. For complete proofs of theoretical results in the paper, please refer to the supplementary material.

Proof Sketch. According to the identifiability of overcomplete ICA, if Condition 2 holds, A^i is identifiable up to column permutation and scaling. Furthermore, if Condition 1 holds, A^i and A^j have two overlapping rows, and by considering all data sets together and estimating A directly, we guarantee that the shared rows of A^i and A^j are the same. Then A^i and A^j have the same column sequence and scale, by matching the ratio of the two overlapping rows, under Assumption 1. It follows that, for any i, A^i has the same column sequence and scale with others. Thus, the estimated A is the same as estimating the mixing matrix over X directly (up to column permutations and rescaling of the whole matrix), if we had such a data set including all variables. Therefore, A is identifiable up to column permutation and scaling.

We can further determine A by using the property that $B = I - A^{-1}$ can be permuted to a strict lower triangular matrix, where the permutation is unique. Therefore, A and the causal adjacency matrix B are uniquely identifiable. □

Based on the example in Figure 1, below we give an intuitive explanation of the identifiability conditions and how Assumption 1 removes ambiguity.

- Suppose there was only one overlapping variable, say x_4, between data sets 1 and 2. Denote the 4-th row of A from data set 1 by $A^1_{4:}$ and that from data set 2 by $A^2_{4:}$. Because the column permutations and scales of A^1 and A^2 are not determined and noise distributions may vary across data set, matching only $A^1_{4:}$ and $A^2_{4:}$ does not give a unique integration result. For example, let $A^1_{4:} = (1, 2, 3, 4)$ and $A^2_{4:} = (2, 1, 3, 4)$. Then $A^2_{4:}$ can be matched to $A^1_{4:}$ either by permuting its first two columns, or by multiplying the first column by $1/2$ and the second column by 2. Thus, only one overlapping variable across the two data sets is not enough.
- Now suppose there are two overlapping variables x_3 and x_4. For example, let $A^1_{\binom{3}{4}:} = \binom{1,2,3,4}{5,6,7,8}$ and $A^2_{\binom{3}{4}:} = \binom{1,2,3,4}{3,10,7,8}$, where Assumption 1 holds. To match $A^2_{\binom{3}{4}:}$ to $A^1_{\binom{3}{4}:}$, its column permutation and scaling can be uniquely determined by matching the ratio of the first row to the second row, more specifically, by permuting the first two columns and then multiplying the first column by $1/2$ and the second column by 2. However, if Assumption 1 does not hold (which is not the case here), the permutation and scaling may not be unique by matching the ratio of the first two columns.

Remark: Note that the conditions in Theorem 2, including linearity and non-Gaussianity, are verifiable. Specifically, for linearity, one may check for the linear relationships by scatter plot, or more rigorously, by linearity tests. It is non-trivial to extend our approach to nonlinear cases, as nonlinear ICA is a nontrivial task, and it is left as future work. For non-Gaussianity, it has been argued that in theory non-Gaussianity is ubiquitous in the linear case (Spirtes and Zhang 2016), and if needed, it can be checked by normality tests (see e.g., (Székely and Rizzo 2005)). In the supplementary material, we also discuss the case when noises are Gaussian, in which not identifiable without further assumption such as faithfulness.

Furthermore, we realize that in a special case when the noise distributions in different data sets are the same, Condition 1 in Theorem 2 can be relaxed to one-variable overlapping, which is stated in Corollary 1, which holds under Assumption 2. The intuition is as follows. Suppose only x_3 is shared in the example given in Figure 1. If the noise distribution does not change, then after enforcing the same variance of all noise terms across data sets, any entry in the third row of A^1 (the mixing matrix on data set 1) will be identical to some entry in the third row of A^2. Hence, we can permute and integrate A^1 and A^2.

Assumption 2. *For any two overlapping data sets m_1 and m_2, there exist $i \in \mathbf{i}$, so that $|a_{ij}| \neq |a_{ik}|$, for any $j \neq k$, where a_{ij} and a_{ik} are entries in A.*

Corollary 1. *Suppose X satisfies the generating process in Eq. (1) and that the noise distributions are fixed across data sets. Under Assumption 2, the causal graph over X is identifiable, if the collected data sets satisfy the following conditions:*

1. *for any data set i, there exists another data set j which has at least one overlapping variables with i;*
2. *all noises are non-Gaussian.*

Next, we show how to learn causal relations from measured data.

5 Model Estimation

We propose two estimation approaches to learning causal relationships over non-identical variable sets. One maximizes the log-likelihood, and the other adopts a likelihood-free framework, adversarial learning-based estimation, to improve scalability.

In estimation, instead of estimating A, we estimate the causal adjacency matrix B directly, where $B = I - A^{-1}$, which enjoys the following advantages:

- It is easy to add prior knowledge of causal connections. In practice, experts may have domain knowledge about the presence/absence or the strength of some causal edges.
- One can directly enforce sparsity constraints on the causal adjacencies. This cannot be easily achieved if the procedure estimates A directly, because even if B is sparse, $A = (I - B)^{-1}$ may not be sparse.
- The estimation procedure directly outputs the causal adjacency matrix, without additional steps of permutation and rescaling, as required in LiNGAM (Shimizu et al. 2006) analysis, which are usually expensive. In practical implementations, one may fix the diagonal entries of B to zero, with no update in the procedure.

Likelihood-Based Estimation

We first introduce a likelihood-based estimation approach, by maximizing the following log-likelihood over all M data

sets:

$$\ell(\boldsymbol{\theta}) = \sum_{m=1}^{M} \log P(\tilde{X}^m; \theta^m), \tag{5}$$

where θ^m denotes involved parameters in the model of m-th data set, and $\boldsymbol{\theta} = \{\theta^m\}_{m=1}^M$. Note that the parameter B is shared in all θ^m, for $m = 1, \cdots, M$.

We allow the noise distributions to vary across data sets. Denote the noise term in the i-th variable in the m-th data set by e_i^m. Specifically, to handle general non-Gaussian distributions, we model e_i^m by a mixture of Gaussian (MoG):

$$p(e_i^m) = \sum_{k=1}^{K} p(z_{i,k}^m = 1)\mathcal{N}(e_i^m; \mu_{i,k}^m, \sigma_{i,k}^m), \tag{6}$$

where $p(z_{i,k}^m = 1) = \pi_{i,k}^m$, with $\sum_{k=1}^K \pi_{i,k}^m = 1$, is the mixture proportion, $Z_i^m = (z_{i,1}^m, \cdots, z_{i,K}^m)$ is a K-dimensional binary vector in which a particular element $z_{i,k}^m$ is equal to 1 and all other elements are zero, and $\mathcal{N}(\cdot)$ denotes the density of a Gaussian distribution.

The causal model defined above can be seen as a latent variable model, with $U^m = \{\{e_i^m\}_{i=1}^d, \{Z_i^m\}_{i=1}^d\}$ as latent variables in the m-th data set, and $\theta^m = \{B, \{\pi_{i,k}^m\}_{i,k}, \{\mu_{i,k}^m\}_{i,k}, \{\sigma_{i,k}^m\}_{i,k}\}$ as free parameters in the m-th data set that are to be estimated. Thus, we use the Expectation-Maximization (EM) algorithm for maximum likelihood estimation of the model parameters, by iterating between two steps, expectation (E) and maximization (M), until convergence:

(E) Compute $p_{\theta^m}(U^m|X^m)$, for $m = 1, \cdots, M$, and the lower bound of the log-likelihood, $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$, with

$$Q(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_{m=1}^{M} \int p_{\theta^m}(U^m|X^m) \log p_{\theta'^m}(X, U^m)\, dU^m.$$

(M) Given $Q(\boldsymbol{\theta}', \boldsymbol{\theta})$, the lower bound is maximized with respect to the parameters, by computing: $\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}', \boldsymbol{\theta})$.

Note that because the EM lower bound Q can be decomposed into several terms which only depend on subsets of the parameters, the parameters can be updated independently.

With the EM algorithm, the computational complexity in each iteration is linear in the number of data sets but is cubic in the number of integrated variables. Hence, it is computationally demanding for large-scale problems. To circumvent this issue, below we propose adversarial learning-based estimation, which is likelihood free.

Adversarial Learning-Based Estimation

To improve scalability, we adopt a likelihood-free framework for parameter estimation, based on a recently proposed likelihood-free overcomplete ICA algorithm (Ding et al. 2019). It utilizes adversarial learning (Goodfellow et al. 2014) for parameter estimation, by minimizing appropriate distributional distances between the generated data and the observed data.

The adversarial learning-based estimation approach estimates the causal adjacency matrix B directly by back-propagation without explicit assumptions on the density function of noise terms. Specifically, each noise term e_i^m is modeled with a function $f_{\omega_i^m}$ that transforms a Gaussian variable u_i^m to the non-Gaussian e_i^m, with involved parameters ω_i^m; that is, $e_i^m = f_{\omega_i^m}(u_i^m)$, where $u_i^m \sim \mathcal{N}(0, 1)$, and the function $f_{\omega_i^m}$ is generated by a Multi-Layer Perceptron (MLP). Note that different noise terms are generated separately, so they are independent of each other and their distributions may be different. Thus, the generative model of data in the m-th data set is as follows:

$$\begin{aligned} \hat{\tilde{X}}^m &= \mathcal{I}^m A E^m = \mathcal{I}^m (I - B)^{-1} [f_{\omega_1^m}(u_1^m), \cdots, f_{\omega_d^m}(u_d^m)]^{\mathrm{T}} \\ &= G_{B,\omega^m}^m(\mathbf{u}^m), \end{aligned} \tag{7}$$

where $\boldsymbol{\omega}^m = [\omega_1^m, \cdots, \omega_d^m]^{\mathrm{T}}$, $\boldsymbol{u}^m = [u_1^m, \cdots, u_d^m]^{\mathrm{T}}$, and we use $G_{B,\omega^m}^m(\cdot)$ to denote the generating function. Figure 2 shows the graphical structure of the data generation process.

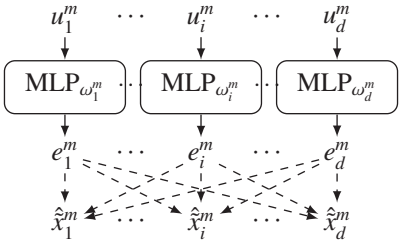


Figure 2: A graph representation of the generating process of the m-th dataset. The dashed lines represents the mixing matrix $I^m A$.

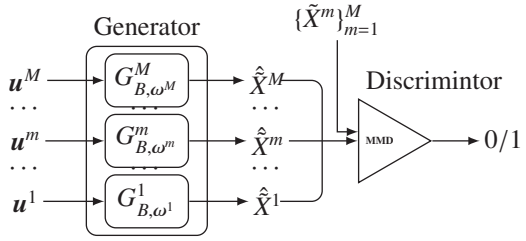


Figure 3: A graph illustration of the whole adversarial-learning process over the M data sets.

To learn the parameters B and $\boldsymbol{\Omega} = [\boldsymbol{\omega}^1, \cdots, \boldsymbol{\omega}^M]$, we minimize the Maximum Mean Discrepancy (MMD (Gretton et al. 2012)) between the joint distributions of true and generated data over all M data sets:

$$\begin{aligned} \hat{B}, \hat{\boldsymbol{\Omega}} &= \arg\min_{B,\boldsymbol{\Omega}} \sum_{m=1}^{M} \mathrm{MMD}(p(X^m), p(G_{B,\omega^m}^m(\boldsymbol{u}^m))) \\ &= \arg\min_{B,\boldsymbol{\Omega}} \sum_{m=1}^{M} \left\| \mathbb{E}_{X^m \sim P(X^m)}[\phi(X^m)] - \mathbb{E}_{\boldsymbol{u}^m \sim P(\boldsymbol{u}^m)}[\phi(G_{B,\omega^m}^m(\boldsymbol{u}^m))] \right\|^2, \end{aligned} \tag{8}$$

where ϕ is the feature map corresponding to a kernel function. All parameters are estimated efficiently by stochastic gradient descent. A graph illustration of the whole adversarial-learning process over the M data sets is shown in Figure 3, and the procedure is given in Algorithm 1.

6 Extensions to More General Cases

In this section, we show that our approach CD-MiNi can be easily extended to confounding cases and cyclic cases.

Algorithm 1 Adversarial Learning-Based Estimation of the Causal Adjacency Matrix

1. Get a minibatch of i.i.d. samples $\{\boldsymbol{u}^m\}_{m=1}^M$ from the standard Gaussian distribution.
2. Generate mixtures $\hat{\tilde{X}}^m$, for $m = 1, \cdots, M$, using Eq. (7).
3. Get a minibatch of samples from the distribution of observed mixtures $p(X^1), \cdots, p(X^M)$.
4. Update B and $\{\boldsymbol{\omega}^m\}_{m=1}^M$ by minimizing the empirical estimate of Eq. (8) on the minibatch.
5. Repeat step 1 to step 4 until max iterations reached.

Confounding Cases

In previous sections, we assume that there are no confounders relative to the integrated set of variables. In this section, we show that CD-MiNi can be extended to the case where there are hidden confounders.

Suppose we did not observe variables $Z = (z_1, \cdots, z_{d_z})$ in any data sets, while they influence two or more variables in X. The generating processes of X and Z are as follows:

$$\begin{cases} X &= B_1 X + B_2 Z + E_X, \\ Z &= B_3 Z + E_Z, \end{cases} \tag{9}$$

where B_1 and B_3 are the causal adjacency matrices over X and Z, respectively, and B_2 is the causal influence from Z to X. After reorganization, Eq. (9) becomes

$$\begin{aligned} X &= \begin{bmatrix} (I - B_1)^{-1} & (I - B_1)^{-1} B_2 (I - B_3)^{-1} \end{bmatrix} \begin{bmatrix} E_X \\ E_Z \end{bmatrix} \\ &= \begin{bmatrix} A_1 & A_1 B_2 A_3 \end{bmatrix} \begin{bmatrix} E_X \\ E_Z \end{bmatrix}, \end{aligned} \tag{10}$$

where $A_1 = (I - B_1)^{-1}$ and $A_3 = (I - B_3)^{-1}$. Specifically, for the m-th data set, we have

$$X^m = \begin{bmatrix} A_1^m & A_1^m B_2 A_3 \end{bmatrix} \begin{bmatrix} E_X^m \\ E_Z^m \end{bmatrix}, \tag{11}$$

for $m = 1, \cdots, M$, where A_1^m is a $d_m \times d$ submatrix of A_1. With a similar trick as in Eq. (4), we can re-formalize the above equation to:

$$\tilde{X}^m = \begin{bmatrix} \mathcal{I}^m A_1 & \mathcal{I}^m A_1 B_2 A_3 \end{bmatrix} \begin{bmatrix} E_X^m \\ E_Z^m \end{bmatrix}. \tag{12}$$

Denote $\begin{bmatrix} A_1 & A_1 B_2 A_3 \end{bmatrix}$ by H for future use.

Since all noises in E_X and E_Z are independent of each other, Eq. (12) satisfies the over-complete ICA model. Hence, similar to the proof of Theorem 2, H is identifiable up to column permutation and scaling, under conditions in Theorem 2 and Assumption 1.

However, since H is only identified up to column permutation and scaling, we cannot determine which columns belong to A_1 without further constraints, and thus, in the confounding case, the causal adjacency matrix B_1 is not uniquely identifiable. In general, the upper bound for the number of causal adjacency matrices is $\frac{(d+d_z)!}{d!d_z!}$, where d is the number of integrated set of observed variables, and d_z is the number of hidden confounders.

Cyclic Cases

Now we consider the case where there are cycles in the graph. In cyclic cases, without further constraints, there may be multiple graphs that are distribution equivalent (Lacerda et al. 2008).

The following theorem shows that with further constraints, including stability of the causal system and the cycles being disjoint, the causal graph over the integrated set of variables X is uniquely identifiable.

Theorem 3. *Suppose the causal graph over X is cyclic. Under Assumption 1, the causal graph over X is identifiable, if the collected data sets satisfy the following conditions:*
1. *for any data set i, there exists another data set j which has at least two overlapping variables with i;*
2. *all noises are non-Gaussian;*
3. *the cycles are disjoint, and the causal system is stable.*

7 Experimental Results

To show the efficacy of the proposed approach for causal discovery from non-identical data sets, we apply it to both synthetic and real-world data.

Synthetic Data

We generated synthetic data according to the functional causal model in Eq. (1). In particular, to generate data from the m-th data set, we first generated complete data of all d variables and then randomly removed some variables. We guaranteed that the generated data satisfy Assumption 1 and the two identifiability conditions in Theorem 2. Each noise term e_i is modeled with a mixture of two Gaussian components, with mean $\mu_{i,k} \sim \mathcal{U}(-0.6, -0.3) \cup \mathcal{U}(0.3, 0.6)$, variance $\sigma_{i,k}^2 \sim \mathcal{U}(0.1, 0.5)$, and the mixture proportion $\pi_{i,k} \sim \mathcal{U}(0.3, 0.6)$ with $\sum_{k=1}^2 \pi_{i,k} = 1$, where $\mathcal{U}(l, u)$ denotes a uniform distribution between l and u. The acyclic causal structure G was generated according to the Erdos-Renyi model. The non-zero entries of the causal adjacency matrix B was generated according to $b_{ij} \sim \mathcal{U}(-0.8, -0.3) \cup \mathcal{U}(0.3, 0.8)$. Note that the noise terms over different variables and different data sets are generated independently, while B is shared across data sets.

To show the finite sample size effect of the proposed method, we varied the sample size per data set $T = 200, 500, 1000, 2000$, the number of integrated variables $d = 4, 6, 8, 10, 20$, and the number of variables per data set $d_s = 3, 4, 5, 6$. For each setting, we generated 30 realizations.

We used the proposed likelihood-based (CD-MiNi-EM) and adversarial learning-based (CD-MiNi-AL) approaches. We compared them with other well-known approaches that are designed to handle such cases, including ION (Danks, Glymour, and Tillman 2009) and COmbINE (Triantafillou and Tsamardinos 2015). We took COmbINE as a representative of SAT-based methods for causal discovery (e.g., Triantafillou, Tsamardinos, and Tollis; Hyttinen et al.; Hyttinen, Eberhardt, and Hoyer; Hyttinen, Eberhardt, and Järvisalo; Rantanen, Hyttinen, and Järvisalo; Zhang et al. (2010; 2013; 2012; 2014; 2018; 2019)), because they are similar.

Since our proposed approaches can recover a unique DAG, while ION and COmbINE may result in a list of PAGs, it is hard to make a fair comparison directly. Thus, for the results from ION and COmbINE, we adopted the following two strategies to analyze the output PAGs. (1) We first only considered those edges which appear in all output PAGs, i.e., the intersection of output PAGs, which are denoted as *solid edges* in Triantafillou and Tsamardinos (2015). Moreover, since in PAGs, some orientations are not determined, we did the following steps: for $\circ\!\!-\!\!\circ$ or $\leftrightarrow$ in a PAG, we transfered it to $\rightarrow$ according to *the ground truth*; for $\circ\!\!\rightarrow$, we transfered it to $\rightarrow$. (2) We next considered all edges that appear in the output PAGs, i.e., a union of all PAGs, and for conflicted ones, we chose the correct one according to *the ground truth*. The undetermined directions were set in the same way. Below, we used *ION-solid* and *COmbINE-solid* to denote the results with Strategy (1) from ION and COmbIME, respectively, and *ION-union* and *COmbINE-union* the results with Strategy (2).

For the results from the proposed methods, the final graph is determined by setting a threshold on the estimated causal adjacency matrix $\hat{B}$; we used 0.1 as the threshold, that is, the estimated graph $\hat{G}_{ij} = 1$ if $|\hat{b}_{ij}| > 0.1$, and $\hat{G}_{ij} = 0$ if otherwise. Alternatively, one may use statistical test to determine the edges (Hoyer et al. 2008). We used the F_1 score to measure the accuracy of the estimated structure, which is the geometric mean of precision and recall.

In Figure 4, we reported the F_1 score (both mean and one standard deviation over 30 realizations) to measure the accuracy of learned causal graphs. Specifically, Figure 4(a) shows the F_1 score along with the sample size $T = 200, 500, 1000, 2000$, when the total number of variables $d = 6$, the number of variables per data set $d_s = 4$, and the number of data sets $M = 3$. Figure 4(b) shows the F_1 score along with the number of variables per data set $d_s = 3, 4, 5, 6$, when $d = 6$, $M = 3$, and $T = 1000$. Figure 4(c) shows the F_1 score along with the number of integrated variables $d = 4, 6, 8, 10, 20$. In this case, $d_s = d/2 + 1 = 3, 4, 5, 6, 11$, $T = 1000$, and $M = 3$. We found that CD-MiNi-EM, ION, and COmbINE are computationally demanding when $d = 20$, so we only reported their results when $d \leq 10$.

From the results, we can see that the proposed methods, CD-MiNi-EM and CD-MiNi-AL, have the best performance (the highest F_1 score) in all settings, and between them, CD-MiNi-AL is slightly better in most cases. More specifically, the accuracy tends to increase along with the sample size or the number of variables per data set with a fixed total number of variables. However, with the increase of total number of variables, they tend to perform less well; we suspect that it is because both EM and adversarial training may suffer more statistical errors and be more prone to local optimal, with the increase of the number of free parameters. With ION or COmbINE with Strategy (1), which only accounts for solid edges (edges appear in all output PAGs), only a very small percentage of edges is identified. With Strategy (2), which considers the union of edges from the output PAGs, we found that their accuracy has improved.

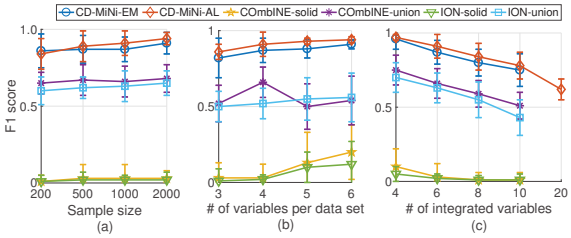


Figure 4: F_1 score of the learned causal structure.

Cellular Signaling Networks

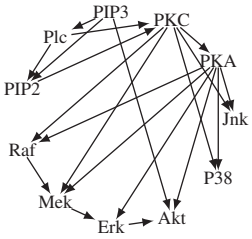
We applied CD-MiNi to multivariate flow cytometry data, which were measured from 11 phosphorylated proteins and phospholipids (Sachs et al. 2005). The 11 variables are Raf, Mek, Plc, PIP2, PIP3, Erk, Akt, Pka, Pkc, P38, and Jnk. We used the data set with general perturbation (anti-CD3/CD28). To test and verify our method, we split the 11 variables into 2 data sets and 3 data sets, respectively. The splits are according to different types of antibodies that were used to assay the target residues.

Figure 5 shows the estimated cellular signaling networks from the 2-dataset case, with CD-MiNi-EM (Figure 5(c)) and CD-MiNi-AL (Figure 5(d)). We compared the estimations with the ground-truth graph given in Figure 5(a); note that edges in the ground-truth graph are those with high confidence in biology, but other edges may also exist in the system. Solid lines denote found edges that are consistent with the ground-truth graph, and dashed edges are those that are not in the ground-truth graph. We found the following true edges with both estimations: Raf $\rightarrow$ Mek, Mek $\rightarrow$ Erk, Plc $\rightarrow$ PIP2, Plc $\rightarrow$ PKC, PIP3 $\rightarrow$ Plc, Erk $\rightarrow$ Akt, PKA $\rightarrow$ Raf, PKA $\rightarrow$ P38, PKA $\rightarrow$ Erk, PKC $\rightarrow$ PKA, PKC $\rightarrow$ P38. Figure 5(b) shows the F_1 scores from 1 data set ($M = 1$, without split), 2 data sets ($M = 2$), and 3 data sets ($M = 3$). Interestingly, the F_1 score seems to be approximately the same across different values of M, indicating that with our approach, we do not sacrifice the estimating accuracy because of the lack of observations for all variables together.

For more experimental details and the results from the 1-dataset case and the 3-dataset case, please refer to the supplementary material.

8 Conclusions

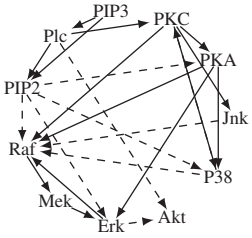
This paper proposed a principled approach CD-MiNi to identify causal relationships from multiple data sets with non-identical variable sets, under the linearity and non-Gaussianity assumption. We showed that the causal structure over the integrated set of variables is uniquely identifiable under certain technical conditions, which are testable given observed data. To the best of our knowledge, the proposed method is the first that can identify the entire DAG from non-identical sets of variables. We presented two types of estimation approaches: a likelihood-based approach and a likelihood-free approach based on adversarial training to improve scalability. The proposed methods showed promising results on flow cytometry data. We further showed that



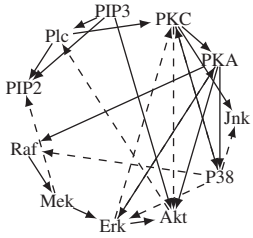
(a)

F_1	EM	AL
$M = 1$	0.65	0.68
$M = 2$	0.63	0.67
$M = 3$	0.66	0.65

(b)



(c)



(d)

Figure 5: Estimated cellular signaling networks. (a) Ground-truth graph (edges with high confidence in biology). (b) F_1 score according to the ground truth in (a). (c) Estimated graph with CD-MiNi-EM from 2 data sets. (d) Estimated graph with CD-MiNi-AL from 2 data sets.

our framework can be easily extended to more general cases, such as the confounding case and the cyclic case. A future line of research is to extend our approach to nonlinear cases.

Acknowledgements

We would like to acknowledge the support by National Institutes of Health under Contract No. NIH-1R01EB022858-01, FAIN-R01EB022858, NIH-1R01LM012087, NIH-5U54HG008540-02, and FAIN-U54HG008540, by the United States Air Force under Contract No. FA8650-17-C-7715, and by National Science Foundation EAGER Grant No. IIS-1829681. The National Institutes of Health, the U.S. Air Force, and the National Science Foundation are not responsible for the views reported in this article. KZ also benefited from funding from Living Analytics Research Center and Singapore Management University.

References

Anandkumar, A.; Hsu, D.; Javanmard, A.; and Kakade, S. 2013. Learning linear bayesian networks with latent variables. In *International Conference on Machine Learning (ICML)*, 249–257.

Chickering, D. M. 2003. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3:507–554.

Claassen, T., and Heskes, T. 2010. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems (NIPS)*, 415–423.

Comon, P. 1994. Independent component analysis, a new concept? *Signal Processing - Special issue on higher order statistics archive* 36(3):287–314.

Danks, D.; Glymour, C.; and Tillman, R. E. 2009. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems (NIPS)*, 1665–1672.

Danks, D. 2005. Scientific coherence and the fusion of experimental results. *The British Journal for the Philosophy of Science* 56(4):791–807.

Ding, C.; Gong, M.; Zhang, K.; and Tao, D. 2019. Likelihood-free overcomplete ICA and applications in causal discovery. *arXiv preprint arXiv:1909.01525*.

Eriksson, V., and Koivunen, J. 2004. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters* 11(7).

Ghassami, A. E.; Kiyavash, N.; Huang, B.; and Zhang, K. 2018. Multi-domain causal structure learning in linear systems. In *Advances in Neural Information Processing Systems (NIPS)*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, 2672–2680.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13:723–773.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20:197–243.

Hoyer, P.; Shimizu, S.; Kerminen, A.; and Palviainen, M. 2008. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning* 49:362–378.

Huang, B.; Zhang, K.; Lin, Y.; B., S.; and Glymour, C. 2018. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 1551–1560.

Huang, B.; Zhang, K.; Xie, P.; Gong, M.; Xing, E. P.; and Glymour, C. 2019a. Specific and shared causal relation modeling and mechanism-based clustering. *In Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, B.; Zhang, K.; Zhang, J.; Ramsey, J.; Sanchez-Romero, R.; Glymour, C.; and Schölkopf, B. 2019b. Causal discovery from heterogeneous/nonstationary data. In *arXiv preprint arXiv:1902.10073*.

Hyttinen, A.; Hoyer, P.; Eberhardt, F.; and Järvisalo, M. 2013. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Uncertainty in Artificial Intelligence (UAI)*.

Hyttinen, A.; Eberhardt, F.; and Hoyer, P. O. 2012. Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. In *arXiv preprint arXiv:1210.4879*.

Hyttinen, A.; Eberhardt, F.; and Järvisalo, M. 2014. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Uncertainty in Artificial Intelligence (UAI)*.

Lacerda, G.; Spirtes, P.; Ramsey, J.; and Hoyer, P. O. 2008. Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Meek, C. 1997. Graphical models: selecting causal and statistical models. *Ph.D Thesis, Carnegie Mellon University*.

R. Silva, R. Scheine, C. G., and Spirtes, P. 2006. Learning the structure of linear latent variable models. *Journal of Machine Learning Research* 7:191–246.

Rantanen, K. M. J.; Hyttinen, A. J.; and Järvisalo, M. J. 2018. Learning optimal causal graphs with exact search. *In Proceedings of the Ninth International Conference on Probabilistic Graphical Models*.

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.

Shimizu, S.; Hoyer, P.; Hyvärinen, A.; and Kerminen, A. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7:2003–2030.

Spirtes, P., and Zhang, K. 2016. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics* 3:1–28.

Spirtes, P.; Glymour, C.; Scheines, R.; and Tillman, R. 2010. Automated search for causal relations: Theory and practice. *Heuristics, Probability, and Causality: A Tribute to Judea Pearl* 467–506.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. Spring-Verlag Lectures in Statistics.

Székely, G. J., and Rizzo, M. L. 2005. A new test for multivariate normality. *Journal of Multivariate Analysis* 93:58–80.

Tillman, R. E., and Eberhardt, F. 2014. Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika* 41(1):41–64.

Tillman, R., and Spirtes, P. 2011. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 3–15.

Triantafillou, S., and Tsamardinos, I. 2015. Constraint-based causal discovery from multiple interventions over overlapping variable sets. In *Journal of Machine Learning Research*, volume 16, 2147–2205.

Triantafillou, S.; Tsamardinos, I.; and Tollis, I. 2010. Learning causal structure from overlapping variable sets. In *In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 860–867.

Wang, S., and Glymour, C. 2019. Unifying causal models with trek rules. *arXiv preprint arXiv:1909.01789*.

Zhang, K., and Hyvärinen, A. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence (UAI)*, 647–655.

Zhang, K.; Huang, B.; Zhang, J.; Glymour, C.; and Schölkopf, B. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*.

Zhang, J.; Eberhardt, F.; Mayer, W.; and Li, M. J. 2019. ASP-based discovery of semi-markovian causal models under weaker assumptions. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*.