Chapter 7

ties between causality and machine learning

Causal Representation Learning

In this final lecture, we discuss problems where we do not directly observe the variables which are related by a structural causal model. This situation occurs widely in practice:

Example 7.1. *An image classifier observes photos of cats and dogs. The variable* Animal *is a cause of variables such as* Wears Collar *and* Has Long Whiskers*. A classifier should understand that putting a collar on a cat does not make it a dag, and making a dog's whiskers longer does not make it a cat. However, the classifier does not have direct access to these variables, besides the label* Animal*. For both robustness and interpretability, we may want the classifier to learn the variables* Wears Collar *and* Has Long Whiskers *from the pixels that it observes.*

diff between causality and interpretibility?

Example 7.2. *An automated diagnostic tool observes a patient's vital signs and fMRI scans of a patient's brain after they are admitted to the emergency room. The variable* Has Stroke *is a cause of variables such as* Stress Response *and* Aspiration*, which are not directly observed, but can be inferred through vital signs such as blood pressure and blood oxygenation. To be transferable between hospitals, which may have different rates of stroke due to demographic differences, the tool should model these and other variables so that it can reason about how to change its disease classifications.*

7.1 Causal Disentanglement

Definition 7.1. *$X \in \mathbb{R}^p$ follows an* **(additive-noise) causal disentanglement model** *if there exists $Z \in \mathbb{R}^d$ such that Z follows a structural causal model M, and $X = g(Z) + \nu$ for some* **mixing function** *g and ν following a product distribution.*

See Figure 7.1.

Remark 7.1. *In this setup, we do not allow that the observed variables cause either the latent variables or each other. The literature calls this a* **measurement model**.

Given X generated from a causal disentanglement model, we have two goals:

- **Goal 1:** Recover the causal graph over Z.
- **Goal 2:** Recover the ~~mixing~~ function g.

As a special case, we may consider a setting where recovery of the mixing function gives us the ability to perfect *disentangle* our data.

Definition 7.2. *We call a causal disentanglement model* **deterministic** *if $X = g(Z)$, i.e., $\nu = 0$ almost surely.*

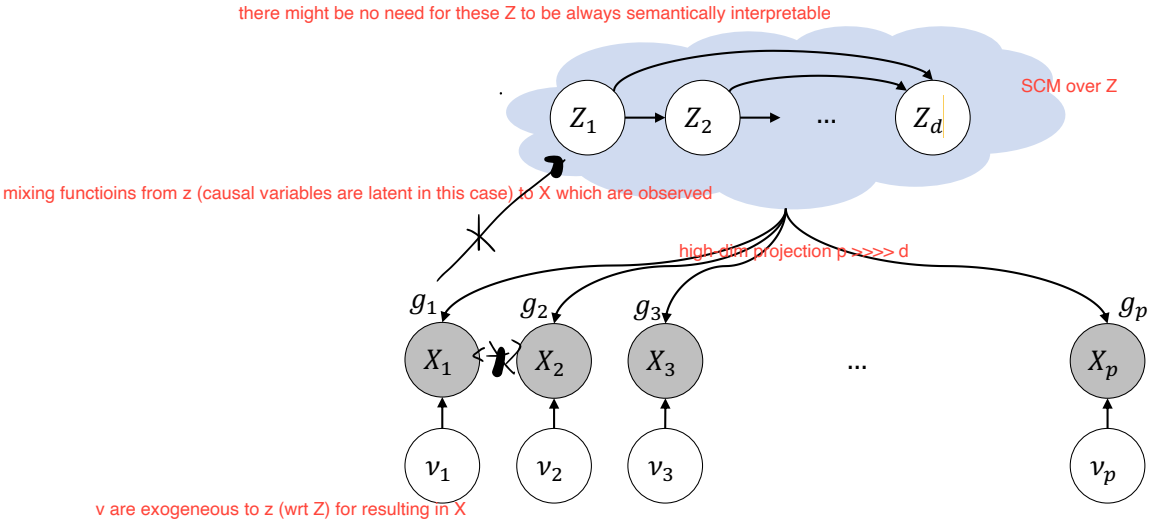

Figure 7.1: The causal disentanglement model.

In the deterministic setting, if g is invertible, then $h := g^{-1}$ disentangles a sample $X^{(m)}$ into its causal representation $Z^{(m)} = h(X^{(m)})$.

In this lecture, we will focus on the more well-studied case of linear models.

Definition 7.3. *We call a causal disentanglement model* **linear** *if (i) Z follows a linear structural causal model, and (ii) g is a linear function.*

identifiable means can be cauptured as a unique form not either one of it.

Even under both of these assumptions, the causal graph over the variables Z is not identifiable from observational data, even up to Markov equivalence, as shown by the following example.

Example 7.3. *Let $\varepsilon_1, \varepsilon_2$ be independent random variables, and let*

$$Z = \begin{bmatrix} 1 & 0 \\ A_{12} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \qquad X = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \tag{7.1}$$

Alternatively, let

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \qquad X = \begin{bmatrix} G_{11} + G_{12}A_{12} & G_{12} \\ G_{21} + G_{22}A_{12} & G_{22} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \tag{7.2}$$

In both models Equation (7.1) and Equation (7.2), we have

$$X = \begin{bmatrix} G_{11} + G_{12}A_{12} & G_{12} \\ G_{21} + G_{22}A_{12} & G_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix},$$

i.e., the distribution of X is the same in both models.

if it could be either two different types of form, we say it's not "identifiable"

Approaches to Identifiability of Latent DAG models

The previous example shows that some additional assumptions or data are required to learn a latent representation. At least three options are apparent:

collection of mixing functions

- **Restrict the form of the mixing function.** If we had $G = I$, then we could identify the latent DAG up to Markov equivalence using only observational data. We can hope to recover the DAG under less stringent conditions on the form of the mixing function. A common assumption is that each latent variable Z_i has a **pure child** (also called an **anchor**) - an observed variable that depends only on Z_i and no other latent variables. The pure child assumption essentially says that some *submatrix* of G is a scaled version of the identity.

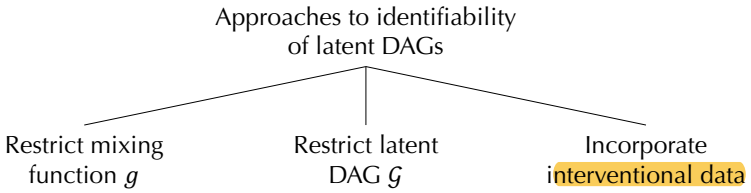

Figure 7.2: Approaches to identifiability of latent DAG models.

- **Restrict the form of the latent DAG.** A "dual" approach would restrict the latent DAG instead of the mixing function. For example, linear independent component analysis (ICA) assumes that the latent variables Z_i are all independent, so **Goal 1** is satisfied by assumption.
- **Incorporate interventional data.** Finally, we may incorporate data from interventions on the latent variables to find $\mathcal{G}$.

7.2 Covariance of Linear DAG models

We will consider identifiability in the linear setting, both by restriction of the mixing function and by incorporating interventional data. Both of these strategies will use some basic properties of covariance matrices for linear DAG models.

why you said Zj -> Zi? not Zi -> Zj?

Definition 7.4. *Let $\mathcal{G}$ be a DAG on nodes $Z_1, \ldots, Z_d$. Let $A \in \mathbb{R}^{d\times d}$ such that $A_{ij} = 0$ unless $X_i \to X_j$ in $\mathcal{G}$, and let $\Omega := \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, where $\sigma_i^2 > 0$ for $i \in [d]$. We say that M is a linear structural causal model with* **weight matrix** *A and* **exogenous variances** *$\sigma_1^2, \ldots, \sigma_p^2$ if*

Xis are typo - it's Z

$$Z_i = \sum_{X_j \in \mathrm{pa}_{\mathcal{G}}(X_i)} A_{ij} Z_j + \sigma_i \varepsilon_i \qquad \forall\ i = 1, \ldots, d$$

not X but Z

for ε such that $\mathrm{Cov}(\varepsilon) = I$.

Remark 7.2. *Given a linear structural causal model with weight matrix A, we can write the equations in matrix-vector form as*

$$Z = A^\top Z + \Omega^{1/2}\varepsilon$$

Since $\mathcal{G}$ is a DAG, A is, up to permutation, equal to an upper triangular matrix. It is well-known that such matrices are invertible, thus we also have that

$$Z = (I - A^\top)^{-1}\Omega^{1/2}\varepsilon$$

Proposition 7.1. *Given a linear structural causal matrix with weight matrix A, we have*

$$\mathrm{Cov}(Z) = (I - A)^{-\top}\Omega(I - A)^{-1}$$

Proof. By linearity of expectation,

$$\mathrm{Cov}(Z) = \mathbb{E}[ZZ^\top] = (I - A^\top)^{-1}\Omega^{1/2} \cdot \mathbb{E}[\varepsilon\varepsilon^\top] \cdot \Omega^{1/2}(I - A)^{-1}$$

□

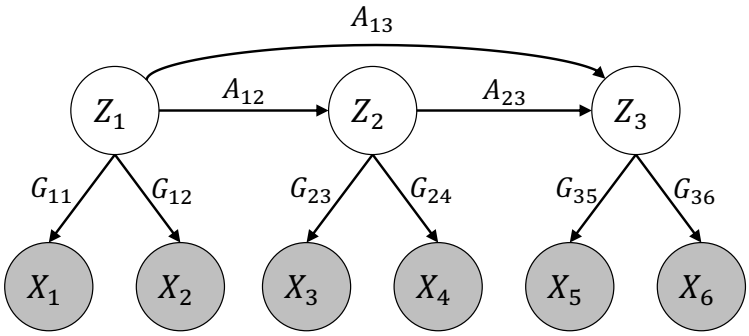

Figure 7.3: A pure measurement model.

7.3 The trek rule

We will now develop a useful characterization of the entries of the covariance matrix in terms of paths in a DAG.

Definition 7.5. *Let $\mathcal{G}$ be a DAG. A* **directed path** *$\gamma = \langle \gamma_1, \ldots, \gamma_M \rangle$ from X_i to X_j is a sequence of nodes such that $\gamma_1 = X_i$, $\gamma_M = X_j$, and $\gamma_m \to \gamma_{m+1}$ in $\mathcal{G}$ for all $m = 1, \ldots, M-1$. The set of all directed paths from X_i to X_j in $\mathcal{G}$ is denoted $\mathcal{P}_{\mathcal{G}}(X_i, X_j)$.*

Given a directed path $\gamma = \langle \gamma_1, \ldots, \gamma_M \rangle$ and a weight matrix A, the **weight** *of γ, denoted $w(\gamma)$, is*

$$w(\gamma) = \prod_{m=1}^{M-1} A_{\gamma_m, \gamma_{m+1}}$$

Proposition 7.2. *Let $\mathcal{G}$ be a weighted DAG with weight matrix A. Then $(A^k)_{ij}$ is the sum of all length-k paths from X_i to X_j.*

Proof. We have

$$(A^2)_{ij} = \sum_{k=1}^{p} A_{ik} A_{kj}$$

i.e., $(A^2)_{ij}$ is the sum of all length-2 paths from X_i to X_j. Generally, we have

$$(A^k)_{ij} = \sum_{k=1}^{p} (A^{k-1})_{ik} A_{kj}$$

and the result follows by induction. □

Proposition 7.3. *Let $\mathcal{G}$ be a weighted DAG with weight matrix A. Then*

$$[(I - A)^{-1}]_{ij} = \sum_{\gamma \in \mathcal{P}_{\mathcal{G}}(X_i, X_j)} w(\gamma)$$

Proof. We have

$$\begin{aligned} \left(\sum_{k=0}^{p} A^k \right) (I - A) &= \sum_{k=0}^{p} A^k - \sum_{k=1}^{p+1} A^k \\ &= A^0 = I \end{aligned}$$

Thus, $(I - A)^{-1} = \sum_{k=0}^{p} A^k$. The result follows from Proposition 7.2. □

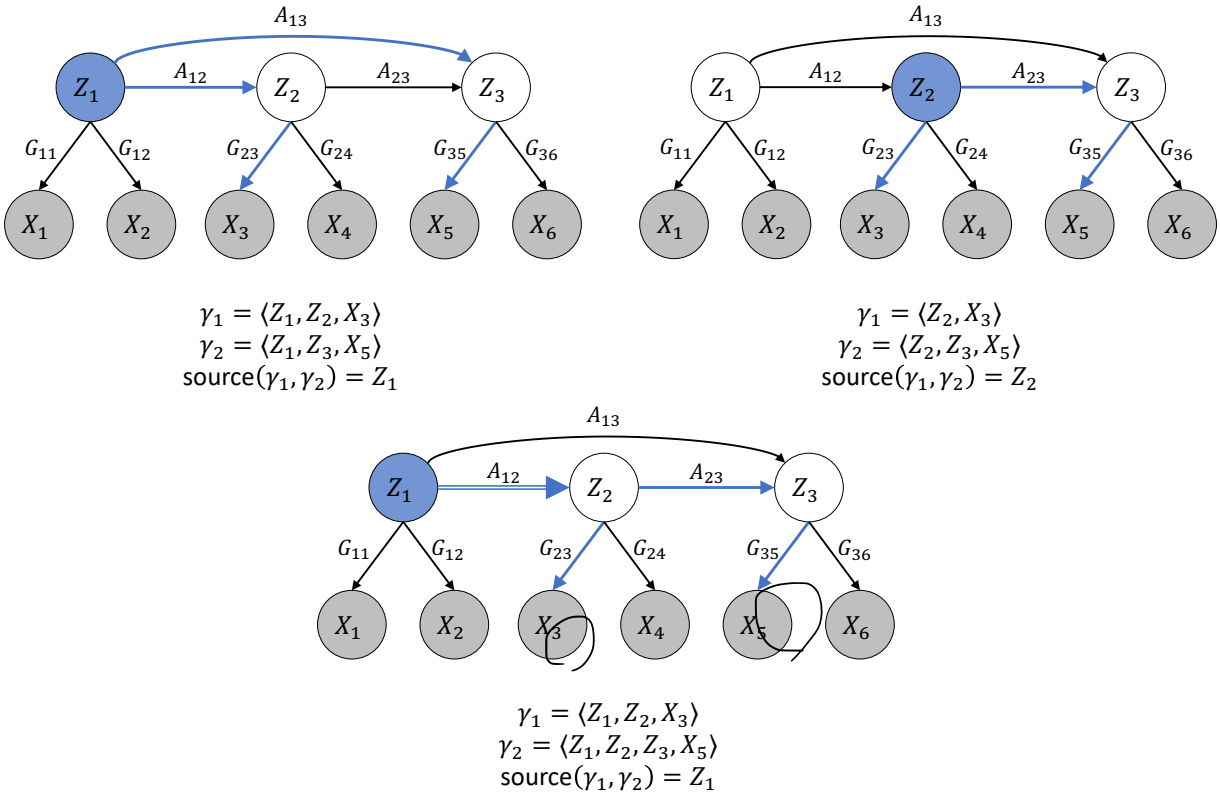

Figure 7.4: The treks between X_3 and X_5. Edges in the trek are in blue, with double-lines if the edge is used twice. The source node of each trek is also in blue.

Definition 7.6. *A* **trek** *from X_i to X_j in $\mathcal{G}$ is an ordered tuple of directed paths (γ_1, γ_2), where:*

- *the sink of γ_1 is X_i,*
- *the sink of γ_2 is X_j, and*
- *γ_1 and γ_2 have the same source, denoted $\text{source}(\gamma_1, \gamma_2)$.*

Denote the set of all treks from X_i to X_j by $\mathcal{T}_{\mathcal{G}}(X_i, X_j)$.

Example 7.4. *Figure 7.4 shows all treks between X_3 and X_5. Two treks have Z_1 as the source, since there is one path from Z_1 to X_3 and there are two paths from Z_1 to X_5. The third trek has Z_2 as a source. Notice that a trek can use an edge twice if it is in both paths γ_1 and γ_2.*

Theorem 7.1 (Trek Rule)**.** *Let $\mathcal{G}$ be a DAG. Let M be a linear structural causal model with causal graph $\mathcal{G}$, weight matrix A, and exogenous variance matrix Ω. Then*

ij component of Cov(Z)

$$\Sigma_{ij} = \sum_{(\gamma_1, \gamma_2) \in \mathcal{T}_{\mathcal{G}}(X_i, X_j)} w(\gamma_1) \cdot w(\gamma_2) \cdot \sigma^2_{\text{source}(\gamma_1, \gamma_2)}$$

+ additional q.
after getting cov(Z) how to determine orientation?

Proof. We have by Proposition 7.3 that

$$\begin{aligned}
\Sigma_{ij} &= \sum_{k=1}^{p} \sigma_k^2 \cdot [(I - A)^{-\top}]_{ik} \cdot [(I - A)^{-1}]_{kj} \\
&= \sum_{k=1}^{p} \sigma_k^2 \left(\sum_{\gamma_1 \in \mathcal{P}_{\mathcal{G}}(X_k, X_i)} w(\gamma_1) \right) \left(\sum_{\gamma_2 \in \mathcal{P}_{\mathcal{G}}(X_k, X_j)} w(\gamma_2) \right)
\end{aligned}$$

□

Example 7.5. *Following Example 7.4, we have that*

$$\mathrm{Cov}(X_3, X_5) = B_{12}G_{23}B_{13}G_{35}\sigma_1^2 + B_{12}G_{23}B_{12}B_{23}G_{35}\sigma_1^2 + G_{23}B_{23}G_{35}\sigma_2^2$$

7.4 Latent DAG recovery via mixing function sparsity

Let $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Then

what does this notation mean?

$$\Sigma_{[13],[45]} = \begin{bmatrix} G_{11}G_{24}A_{12}\sigma_1^2 & G_{11}G_{35}(A_{12}A_{23}+A_{13})\sigma_1^2 \\ G_{23}G_{24}A_{12}^2\sigma_1^2 + G_{23}G_{24}\sigma_2^2 & G_{23}G_{35}A_{12}(A_{12}A_{23}+A_{13}) + G_{23}G_{35}A_{23}\sigma_2^2 \end{bmatrix}$$

2x2 determinants?

In particular, if $A_{13} = 0$, we have

$$\begin{aligned}\Sigma_{[13],[45]} &= \begin{bmatrix} G_{11}G_{24}A_{12}\sigma_1^2 & G_{11}G_{35}A_{12}A_{23}\sigma_1^2 \\ G_{23}G_{24}A_{12}^2\sigma_1^2 + G_{23}G_{24}\sigma_2^2 & G_{23}G_{35}A_{12}^2A_{23} + G_{23}G_{35}A_{23}\sigma_2^2 \end{bmatrix} \\ &= \begin{bmatrix} A_{12}G_{11}\sigma_1^2 \\ A_{12}^2G_{23}\sigma_1^2 + G_{23}\sigma_2^2 \end{bmatrix} \begin{bmatrix} G_{24} & A_{23}G_{35} \end{bmatrix}\end{aligned}$$

i.e., $\Sigma_{[13],[45]}$ is rank-1. this means that whether we have an edge or not between two nodes

This does not hold generically if $A_{13} \neq 0$, and gives a way to distinguish whether the skeleton of $\mathcal{G}$ includes an edge between Z_1 and Z_3.

This is a special case of a more general result, which we will state but not prove.

Definition 7.7. *Let $\mathbf{A}, \mathbf{B}, \mathbf{S_A}, \mathbf{S_B}$ be subsets of nodes in $\mathcal{G}$, not necessarily disjoint. We say $(\mathbf{S_A}, \mathbf{S_B})$* **trek-separates** $\mathbf{A}$ *and* $\mathbf{B}$ *if for every trek γ_1, γ_2 from a node in $\mathbf{A}$ to a node in $\mathbf{B}$, either γ_1 intersects $\mathbf{S_A}$, or γ_2 intersects $\mathbf{S_B}$.*

Theorem 7.2 (Trek Separation, Sullivant et al. (2010))**.** *Given a linear structural equation model with weight matrix A, we have*

$$\mathrm{rank}(\Sigma_{A,B}) \leq \min\{|\mathbf{S_A}| + |\mathbf{S_B}| : (\mathbf{S_A}, \mathbf{S_B}) \text{ trek separates } \mathbf{A} \text{ from } \mathbf{B}\}$$

where equality holds generically over A.

Example 7.6. *If we do not have the edge $Z_1 \to Z_3$ in Figure 7.3, then we have that $A = \{X_1, X_3\}$ is trek-separated from $B = \{X_4, X_5\}$ by $(\mathbf{S_A}, \mathbf{S_B}) = (\{Z_2\}, \varnothing)$.*

This forms the basis for one of the first algorithms for learning a latent DAG, introduced by Silva et al. (2006). This algorithm assumes that every observed node is a pure child of some latent variable. Observed variables can then be clustered: if two observed variables X_i and X_j have the same latent parent, then that parent trek-separates them from the set of all other variables. This can be used in combination with conditional independence testing to cluster the observed variables, and trek-separation can be further used to discover the latent DAG up to Markov equivalence.

7.5 Latent DAG recovery via interventions

First, we take note of two "trivial" sources of non-identifiability that can be fixed by taking canonical choices.

- **Scaling.** Consider a fixed G and A. Let Λ be a diagonal matrix. Then $\widehat{G} = G\Lambda^{-1}$ and $\widehat{A} = \Lambda A$ satisfies $\widehat{G}\widehat{A} = GA$, with $\widehat{A}$ and A having the same sparsity pattern.
- **Permutation.** Consider a fixed G and A. Let P be a permutation matrix. Let $\widehat{G} = GP^\top$ and $\widehat{A} = PAP^\top$. Then $X = \widehat{G}\widehat{A}P\varepsilon$. We can get rid of this by fixing A to be lower triangular, so that B is upper triangular.

Identifiability from perfect interventions

Let $B = \Omega^{-1/2}(I - A^\top)$. Then we can re-write

$$\mathrm{Cov}(Z) = B^{-1}B^{-\top} \qquad \text{and } \mathrm{Cov}(Z)^{-1} = B^\top B$$

Then, if $X = GZ$ for some $G \in \mathbb{R}^{d\times d}$ invertible and $H := G^{-1}$, we have

add no noise , i.e. deterministic setting

$$\mathrm{Cov}(X) = GB^{-1}B^{-\top}G^\top \qquad \text{and } \mathrm{Cov}(X)^{-1} = H^\top B^\top BH$$

Let B denote the weight matrix of Z in the observational setting. Suppose B^a comes from a single-node intervention on Z_2, so that

how to perform intervention on latent node Z?

$$B^a = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B^a_{22} \end{bmatrix}$$

only changes the second row cuz it's Z_2

Similarly, suppose that B^b comes from a single-node intervention on Z_1, so that

$$B^b = \begin{bmatrix} B^b_{11} & 0 \\ 0 & B_{22} \end{bmatrix}$$

only changes the second row cuz it's Z_1

Thus, we have

$$\begin{aligned}
\mathrm{Cov}(Z)^{-1} &= \begin{bmatrix} B_{11}^2 & B_{11}B_{12} \\ B_{11}B_{12} & B_{12}^2 + B_{22}^2 \end{bmatrix} \\
\mathrm{Cov}_a(Z)^{-1} &= \begin{bmatrix} B_{11}^2 & B_{11}B_{12} \\ B_{11}B_{12} & B_{12}^2 + (B^a_{22})^2 \end{bmatrix} \\
\mathrm{Cov}_b(Z)^{-1} &= \begin{bmatrix} (B^b_{11})^2 & 0 \\ 0 & B_{22}^2 \end{bmatrix}
\end{aligned}$$

These induce the inverse covariance matrices

$$\begin{aligned}
\Theta &:= \mathrm{Cov}(X)^{-1} = H^\top B^\top BH, \\
\Theta_a &:= \mathrm{Cov}(X)_a^{-1} = H^\top {B^a}^\top B^a H, \text{ and} \\
\Theta_b &:= \mathrm{Cov}(X)_b^{-1} = H^\top {B^b}^\top B^b H,
\end{aligned}$$

respectively.

Recovering a basis for the rows of H.

We inspect the ranks of the differences $\Theta - \Theta_a$ and $\Theta - \Theta_b$. First,

$$\begin{aligned}
\Theta - \Theta_a &= H^\top (B^\top B - B_a^\top B_a) H \\
&= H^\top \begin{bmatrix} 0 & 0 \\ 0 & B_{22}^2 - (B^a_{22})^2 \end{bmatrix} H
\end{aligned}$$

So, $\mathrm{rank}(\Theta - \Theta_a) = 1$, and $\mathrm{rowspan}(\Theta - \Theta_a) = \mathbf{h}_2$, the second row of H. Pick a unit-length vector $\hat{\mathbf{q}}_2$ as a basis for this subspace.

Next,

$$\begin{aligned}
\Theta - \Theta_b &= H^\top (B^\top B - B_b^\top B_b) H \\
&= H^\top \begin{bmatrix} B_{11}^2 - (B^b_{11})^2 & B_{11}B_{12} \\ B_{11}B_{12} & B_{12}^2 \end{bmatrix} H
\end{aligned}$$

Note that this matrix is rank 2, so we can distinguish that Θ_a comes from an intervention on Z_2, while Θ_b comes from an intervention on Z_1. Now, given $\widehat{\mathbf{q}}_2$, we can project the row space and column space of this matrix onto the orthogonal complement of $\langle \widehat{\mathbf{q}}_2 \rangle$, giving

$$\begin{aligned}\text{proj}_{\widehat{\mathbf{q}}_2^\perp}(\Theta - \Theta_b) &= \begin{bmatrix} q_{11} & 0 \\ q_{12} & 0 \end{bmatrix} \begin{bmatrix} B_{11}^2 - (B_{11}^b)^2 & B_{11}B_{12} \\ B_{11}B_{12} & B_{12}^2 \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} \\ 0 & 0 \end{bmatrix} \\ &= \left(B_{11}^2 - (B_{11}^b)^2\right) \begin{bmatrix} q_{11} \\ q_{12} \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} \end{bmatrix}\end{aligned}$$

where $\mathbf{q}_1 = \text{proj}_{\widehat{\mathbf{q}}_2^\perp}(\mathbf{h}_1)$.

So, $\text{rank}(\text{proj}_{\widehat{\mathbf{q}}_2^\perp}(\Theta - \Theta_b)) = 1$, and $\text{rowspan}(\text{proj}_{\widehat{\mathbf{q}}_2^\perp}(\Theta - \Theta_b)) = \mathbf{q}_1$. Pick a unit-length vector $\widehat{\mathbf{q}}_1$ as a basis for this subspace, and let

$$\widehat{Q} = \begin{bmatrix} \widehat{\mathbf{q}}_1 \\ \widehat{\mathbf{q}}_2 \end{bmatrix}$$

Recovering H.

We may express H as

$$H = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \widehat{Q}$$

In particular,

$$\begin{aligned}\widehat{Q}^\top \Theta \widehat{Q} &= R^\top B^\top B R \\ \widehat{Q}^\top \Theta_a \widehat{Q} &= R^\top {B^a}^\top B^a R \\ \widehat{Q}^\top \Theta_b \widehat{Q} &= R^\top {B^b}^\top B^b R\end{aligned}$$

Since Θ, Θ_a and Θ_b are positive definite, and positive definiteness is preserved under conjugation, these matrices are also positive definite. Given a positive definite matrix M, there is a unique upper triangular matrix U, the *Cholesky factor*, with positive diagonal elements such that $M = U^\top U$. B, B^a, B^b and R are upper triangular. Further, B, B^a, and B^b have positive diagonal, and R may be induced to have positive diagonal by changing the signs of $\widehat{\mathbf{q}}_1$ and $\widehat{\mathbf{q}}_2$. Thus, by the Cholesky decomposition, we can recover the matrices $C = BR$, $C^a = B^a R$, and $C^b = B^b R$.

Finally, we have that

$$\begin{aligned}C^a &= \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22}^a \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \\ &= \begin{bmatrix} B_{11}R_{11} & B_{11}R_{12} + B_{12}R_{22} \\ 0 & B_{22}^b R_{22} \end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}C^b &= \begin{bmatrix} B_{11}^b & 0 \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \\ &= \begin{bmatrix} B_{11}^b R_{11} & B_{11}^b R_{12} \\ 0 & B_{22} R_{22} \end{bmatrix}\end{aligned}$$

Thus, we may recover the 2nd row of R up to scaling by extracting the second row of C^a, and the 1st row of R up to scaling by extracting the first row C^b. This will give H up to scaling, as desired.

7.6 Additional Remarks

- **Recovery from single-node perfect interventions.** The generalization of our final result is in Seigal et al. (2022). We showed:

  - Recovery of the latent graph and the mixing matrix for p observed variables and d latent variables, with $p \geq d$, under a single-node perfect intervention on each latent variable.
  - The necessity, in the worst case, of a single-node intervention on each latent variable.

- **Recovery under other interventions.** Other recent papers consider do-interventions Ahuja et al. (2022) and soft interventions Varici et al. (2023).

- **Causality-inspired representation learning.** In this lecture, we focused on learning a latent DAG model, a task which we called *causal disentanglement*. The term *causal representation learning* is used more broadly, and often refers to learning only a *subset* of the latent variables which are used in downstream tasks such as prediction. See Arjovsky et al. (2019) and Schölkopf et al. (2021).

Bibliography

Ahuja, K., Wang, Y., Mahajan, D., and Bengio, Y. (2022). Interventional causal representation learning. *arXiv preprint arXiv:2209.11924*.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards causal representation learning. *arXiv preprint arXiv:2102.11107*.

Seigal, A., Squires, C., and Uhler, C. (2022). Linear causal disentanglement via interventions. *arXiv preprint arXiv:2211.16467*.

Silva, R., Scheines, R., Glymour, C., Spirtes, P., and Chickering, D. M. (2006). Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2).

Sullivant, S., Talaska, K., and Draisma, J. (2010). Trek separation for gaussian graphical models.

Varici, B., Acarturk, E., Shanmugam, K., Kumar, A., and Tajer, A. (2023). Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*.