Seminar in Directed Graphical Models and Causality

2. Equivalence of Factorization and Markov Properties and Independence Testing

Grigor Keropyan

Technical University of Munich

August 17, 2021

Outline

Equivalence of Markov Properties and Factorization

Theorem

Let $\mathcal{G}$ be a DAG. Suppose the joint distribution P^X has density with respect to a product measure λ. Then, the following conditions are equivalent

1. *The joint distribution P^X factorizes according to graph $\mathcal{G}$.*
2. *The joint distribution P^X is Global Markov w.r.t. $\mathcal{G}$.*
3. *The joint distribution P^X is Local Markov w.r.t. $\mathcal{G}$.*

Proof of Theorem

$2 \implies 3$: Assume Global Markov Property holds. For all $v \in \mathbf{V}, w \in \mathbf{V} \setminus \{\{v\} \cup \mathbf{PA}_v^{\mathcal{G}} \cup \mathbf{DE}_v^{\mathcal{G}}\}$ and every path $\pi = v, j_1, \ldots, j_n, w$ from v to w we have either $j_1 \in \mathbf{PA}_v^{\mathcal{G}}$ or $j_1 \in \mathbf{CH}_v^{\mathcal{G}}$.

1. $j_1 \in \mathbf{PA}_v^{\mathcal{G}}$: Since we have $v \leftarrow j_1$ in the path, j_1 cannot be a collider and it is in the conditioning set. So

   $$\{v\} \text{ and } \mathbf{V} \setminus \{\{v\} \cup \mathbf{PA}_v^{\mathcal{G}} \cup \mathbf{DE}_v^{\mathcal{G}}\} \text{ are d-separated by } \mathbf{PA}_v^{\mathcal{G}}.$$

   Global Markov Property gives

   $$v \perp\!\!\!\perp \mathbf{V} \setminus \{\{v\} \cup \mathbf{PA}_v^{\mathcal{G}} \cup \mathbf{DE}_v^{\mathcal{G}}\} | \mathbf{PA}_v^{\mathcal{G}}.$$

2. $j_1 \in \mathbf{CH}_v^{\mathcal{G}}$: Let's take minimal index k such that $j_k \leftarrow j_{k+1}$, which exists as w is not in descendants of v. This implies that j_k is a collider in the path and j_k and also it's descendants are not in the conditioning set as the graph is acyclic. So,

   $$v \perp\!\!\!\perp \mathbf{V} \setminus \{\{v\} \cup \mathbf{PA}_v^{\mathcal{G}} \cup \mathbf{DE}_v^{\mathcal{G}}\} | \mathbf{PA}_v^{\mathcal{G}}.$$

Proof of Theorem

$3 \implies 1$: Proof is by induction by number of nodes $m = \#\mathbf{V}$ in the graph.

1. $m = 1$ case there is nothing to prove.
2. Assume it is true for m. Let's prove for $m + 1$ nodes:
   - Since the graph is acyclic there is always a vertex v which is terminal (does not have any children).
   - Local Markov Property gives $f(x_v | x_{\mathbf{V}\setminus\{v\}}) = f(x_v | x_{\mathbf{PA}_v^{\mathcal{G}}})$.
   - $f(x) = f(x_v | x_{\mathbf{V}\setminus\{v\}}) f(x_{\mathbf{V}\setminus\{v\}}) = f(x_v | x_{\mathbf{PA}_v^{\mathcal{G}}}) f(x_{\mathbf{V}\setminus\{v\}})$
   - Clearly $f(x_{\mathbf{V}\setminus\{v\}})$ is Local Markov w.r.t. subgraph consists of $\mathbf{V} \setminus \{v\}$ nodes and by induction step we have it factorizes according to the subgraph.
   - Putting last two steps together we obtain

$$f(x) = f(x_v | x_{\mathbf{PA}_v^{\mathcal{G}}}) f(x_{\mathbf{V}\setminus\{v\}}) = \prod_{v \in \mathbf{V}} f(x_v | x_{\mathbf{PA}_v^{\mathcal{G}}}).$$

$1 \implies 2$: Missing! ONB if time allows. $\square$

Markov Equivalence class and Causal Minimality

- A distribution satisfies **causal minimality** with respect to graph $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$, but not to any proper subgraph of $\mathcal{G}$.
- Let's denote $\mathcal{M}(\mathcal{G}) := \{P^{\mathbf{X}} : P^{\mathbf{X}} \text{ is Markov w.r.t. } \mathcal{G}\}$ all the distributions which are Markov with respect to $\mathcal{G}$.
- Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are called **Markov equivalent** if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.
- The above holds if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy same set of d-separations.
- The set of all DAGs that are Markov equivalent to some DAG is called Markov equivalence class.

Independence Testing Preliminaries

- Let random vectors X and Y are defined on the $(\mathcal{X}, \mathcal{A}, \lambda_X)$ and $(\mathcal{Y}, \mathcal{B}, \lambda_Y)$ measure spaces, respectively.
- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed (i.i.d) copies of random vector (X, Y).
- Our aim is to test the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ of independence vs $H_1 : X \not\perp\!\!\!\perp Y$.
- A (randomized) independence test is a measurable function $\psi : (\mathcal{X} \times \mathcal{Y})^n \to [0, 1]$.
- Observing a sample $(X_1, Y_1, \ldots, X_n, Y_n) = (x_1, y_1, \ldots, x_n, y_n)$, we reject H_0 with probability $\psi(x_1, y_1, \ldots, x_n, y_n)$.

Error Types and Power Function

Type I error: Rejecting H_0 when it is true (i.e. saying X and Y are not independent).

Type II error: Accepting H_0 when it is false (i.e. saying X and Y are independent).

Power function: $\beta = E[\psi(X_1, Y_1, \ldots, X_n, Y_n)]$. In particular, for the non-randomized case $\beta = P(\psi(X_1, Y_1, \ldots, X_n, Y_n) = 1)$.

Neyman-Pearson Approach of testing

Neyman-Pearson approach: Requires test ψ to have (significance) level α, that is fixing $\alpha \in (0, 1)$

$$E[\psi(X_1, Y_1, \ldots, X_n, Y_n)] \leq \alpha \quad \text{if } X \text{ and } Y \text{ are independent,}$$

subject to power is maximized

$$E[\psi(X_1, Y_1, \ldots, X_n, Y_n)] \quad \text{if } X \text{ and } Y \text{ are not independent.}$$

Definition

The size of test ψ is $supE[\psi(X_1, Y_1, \ldots, X_n, Y_n)]$, when X and Y are independent. A test ψ has level α if its size is at most α.

Independence tests for Discrete Data

- Consider i.i.d.copies of (X, Y), where X taking the value x_i with probability q_i, for $i = 1, \ldots, I$ and Y taking the value y_j with probability r_j, for $j = 1, \ldots, J$.
- From the random sample of size n we can summarize the data into contingency table with I rows and J columns, where (i, j)th entry o_{ij} denotes the number of data points equal to (x_i, y_j).
- Denote $p_{ij} := P(X = x_i, Y = y_j)$ is the probability that observation will appear in (i, j)th cell.
- Testing H_0 is the same as testing $p_{ij} = q_i r_j$ for all i, j.
- Let $o_{i+} := \sum_{j=1}^{J} o_{ij}$ sum of the entries in ith column and $o_{+j} := \sum_{i=1}^{I} o_{ij}$ sum of the entries in jth row.
- Let $e_{ij} = o_{i+} o_{+j} / n$ be the "expected" number of observations in the (i, j)th cell under null hypothesis.

Example 1

- X represents a martial status (values: "Never married", "Married", "Divorced", "Widowed") and Y represents level of education (values: "Middle school or lower", "High school", "Bachelor's", "Master's", "PhD or higher")

	Middle school or lower	High school	Bachelor's	Master's	PhD or higher
Never married	18	36	21	9	6
Married	12	36	45	36	21
Divorced	6	9	9	3	3
Widowed	3	9	9	6	3

Table 1: Contingency table summarising the marital status and education level of 300 survey respondents. Source: https://www.spss-tutorials.com/chi-square-independence-test/. [1]

[1] Berrett and Samworth "USP: an independence test that improves on Pearson's chi-squared and the G-test"

Pearson's chi-squared test

- Fix significance level α.
- Test Statistic: $\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$.
- χ^2 test statistic is compared to the $(1-\alpha)$-level quantile of chi-squared distribution with $(I-1)(J-1)$ degrees of freedom.
- Reject the null hypothesis if the test statistic is greater than $(1-\alpha)$-level quantile, otherwise not.

Note: Distribution of random variable Y is chi-squared with dgrees of freedom k if

$$Y = \sum_{j=1}^{k} Z_j^2,$$

where $Z_1, \ldots, Z_k$ are i.i.d standard normal random variables.

Testing via Pearson's chi-squared test

- From the data in Table 1 we have $\chi^2 = 23.6$
- Corresponding degrees of freedom is $(4-1)(5-1) = 12$
- 0.95-level and 0.99-level quantile of chi-squared distribution with 12 degrees of freedom are 21.0 and 26.2
- So, we reject the null hypothesis at the 5% significance level, but not 1%
- p-value corresponding to the $\chi^2 = 23.6$ is 0.0235

G-test

- Fix significance level α.
- Test Statistic: $G = 2\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij} \log \frac{o_{ij}}{e_{ij}}$.
- G-test statistic is compared to the $(1-\alpha)$-level quantile of chi-squared distribution with $(I-1)(J-1)$ degrees of freedom.
- Reject the null hypothesis if the test statistic is greater than $(1-\alpha)$-level quantile, otherwise not.

Drawbacks of Pearson's chi-squared and G tests[1]

- There are cases when the tests do not control the probability of Type I error at the desired level.
- In case there is a row or a column without any observation the test statistics are undefined.
- The power properties of both Pearson's chi-squared test and the G-test are poorly understood.

[1] Berrett and Samworth "USP: an independence test that improves on Pearson's chi-squared and the G-test" 2021.

USP test of Independence[1]

- Define a measure of dependence in a contingency table by
$$D = \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - q_i r_j)^2.$$
- Clearly, $D = 0$ if and only if X and Y are independent.
- However, we do not have population values p_{ij}, q_i and r_j, that is why we will try to estimate D from the observational data.
- Define
$$h((x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)) =$$
$$\sum_{i=1}^{I} \sum_{j=1}^{J} (\mathbb{1}_{\{x_1=i, y_1=j\}} \mathbb{1}_{\{x_2=i, y_2=j\}} - 2 \cdot \mathbb{1}_{\{x_1=i, y_1=j\}} \mathbb{1}_{\{x_2=i\}} \mathbb{1}_{\{y_3=j\}} + \mathbb{1}_{\{x_1=i\}} \mathbb{1}_{\{y_2=j\}} \mathbb{1}_{\{x_3=i\}} \mathbb{1}_{\{y_4=j\}}).$$

[1] Berrett and Samworth "USP: an independence test that improves on Pearson's chi-squared and the G-test" 2021.

USP test of Independence

- $h((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))$ is an unbiased estimator of D, that is

$$E[h((X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4))] =$$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (P(X_1 = i, Y_1 = j)P(X_2 = i, Y_2 = j) - 2P(X_1 = i, Y_1 = j)P(X_2 = i)P(Y_3 = j) +$$

$$P(X_1 = i)P(Y_2 = j)P(X_3 = i)P(Y_4 = j)) =$$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij}^2 - 2p_{ij}q_i r_j + q_i^2 r_j^2) = \sum_{i=1}^{I} \sum_{j=1}^{J} (p_{ij} - q_i r_j)^2 = D.$$

- So, it is natural to define $\hat{D} = \sum_{\text{distinct } i_1, i_2, i_3, i_4 = 1}^{n} h((x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), (x_{i_3}, y_{i_3}), (x_{i_4}, y_{i_4}))$.

Theorem

The test statistic $\hat{D}$ is the unique minimum variance unbiased estimator (UMVUE) of D.

USP test of Independence

- Since in the permutation test we are only interested the rank of the test statistic we can disregard all the items that are the same in all permutation cases.
- after canceling out the same items we obtain $\hat{U}$ as a test statistic

$$\hat{U} = \frac{1}{n(n-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} (o_{ij} - e_{ij})^2 - \frac{4}{n(n-2)(n-3)} \sum_{i=1}^{I} \sum_{j=1}^{J} o_{ij} e_{ij}.$$

- Given a data $T = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
- Compute $\hat{U} = \hat{U}(T)$ on the original data.
- Choose B large enough (usually $B = 999$).
- For each $b = 1, \ldots, B$ generate an independent permutation $\sigma^{(b)}$ of $\{1, \ldots, n\}$ uniformly at random in all $n!$ possible choices.

USP test of Independence

- Construct permuted test sets $T^{(b)} = \{(X_1, Y_{\sigma^{(b)}(1)}), \ldots, (X_n, Y_{\sigma^{(b)}(n)})\}$.
- Compute permuted test statistics $\hat{U}^{(b)} = \hat{U}(T^{(b)})$.
- Under null hypothesis $\hat{U}^{(1)}, \ldots, \hat{U}^{(B)}$ have the same distribution as $\hat{U}$.
- In order to understand, under null hypothesis if out test statistic $\hat{U}$ is extreme among all $B+1$ test statistics $\hat{U}, \hat{U}^{(1)}, \ldots, \hat{U}^{(B)}$ we compute the rank of it and break the ties at random.
- For the case of α Type I we reject the null hypothesis if $\hat{U}$ is at least the $\alpha(B+1)$th largest element among the $B+1$ test statistics.

Conditional Independence Testing

- Let random vectors X, Y, Z are defined on the $(\mathcal{X}, \mathcal{A}, \lambda_X)$, $(\mathcal{Y}, \mathcal{B}, \lambda_Y)$ and $(\mathcal{Z}, \mathcal{C}, \lambda_Z)$ measure spaces, respectively.
- Let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ be independent and identically distributed (i.i.d) copies of random vector (X, Y, Z).
- Our aim is to test the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ of independence vs $H_1 : X \not\perp\!\!\!\perp Y|Z$.
- A (randomized) independence test is a measurable function $\psi_n : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^n \times [0, 1] \to \{0, 1\}$.
- Observing a sample $(X_1, Y_1, Z_1, \ldots, X_n, Y_n, Z_n) = (x_1, y_1, z_1 \ldots, x_n, y_n, z_n)$, we reject H_0 if $\psi(x_1, y_1, z_1 \ldots, x_n, y_n, z_n, \alpha) = 1$.

Hardness of Conditional Independence testing[1]

- Let $\mathcal{P}_0 \subset \mathcal{E}_0$ be subset of all distributions for which $X \perp\!\!\!\perp Y|Z$.
- Let $\mathcal{E}_{0,M} \subseteq \mathcal{E}_0$ be the subset of all distributions with support contained strictly in ℓ_∞ ball of radius M for all $M \in (0, \infty)$. (Note, $\mathcal{E}_{0,\infty} = \mathcal{E}_0$)
- Define $\mathcal{Q}_0 = \mathcal{E}_0 \setminus \mathcal{P}_0, \mathcal{P}_{0,M} = \mathcal{E}_{0,M} \cap \mathcal{P}_0$ and $\mathcal{Q}_{0,M} = \mathcal{E}_{0,M} \cap \mathcal{Q}_0$.

Definition

Given a $\alpha \in (0, 1)$ and null hypothesis $\mathcal{P}$ (i.e., all the distributions that null hypothesis holds). Suppose ϕ_n is a randomized independence test as described above. We say that ψ_n has valid level at sample size n if $sup_{P\in\mathcal{P}}\mathbb{P}_P(\psi_n = 1) \leq \alpha$.

[1] Shah and Peters "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure" 2021.

No-free-lunch[1]

Theorem

Given any $n \in \mathrm{N}, \alpha \in (0,1)$ *and* $M \in (0,\infty)$. *Suppose* ψ_n *is any randomized independence test that has valid level* α *for the null hypothesis* $\mathcal{P}_{0,M}$. *Then*

$$\mathbb{P}_Q(\psi_n = 1) \leq \alpha \quad \forall Q \in \mathcal{Q}_{0,M}.$$

Thus, ψ_n *does not have power against any alternative.*

[1] Shah and Peters "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure" 2021.

Distance Covariance[1]

- Let $\|\cdot\|$ denote the Euclidean norm of corresponding dimension.
- For every n, let $c_n = \pi^{\frac{1+n}{2}} / \Gamma((\frac{1+n}{2}))$, where Γ is the Gamma function (i.e., $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, dx$).
- Distance Covariance (dCov) of random vectors X and Y is defined as follows

$$dCov^2(X, Y) = \frac{1}{c_n c_m} \int_{\mathbf{R}^{m+n}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{\|t\|^{n+1} \|s\|^{m+1}} dtds$$

  where f_X, f_Y and $f_{X,Y}$ are the marginal and joint density functions respectively.
- Clearly, $dCov(X, Y) = 0 \iff X \perp\!\!\!\perp Y$.
- Let X_1 and X are i.i.d and Y_1, Y_2 and Y are i.i.d.. Then, dCov can be expressed as

$$dCov^2(X, Y) = E[\|X - X_1\| \, \|Y - Y_1\|] + E[\|X - X_1\|] E[\|Y - Y_1\|] - 2E[\|X - X_1\| \, \|Y - Y_2\|].$$

[1] Székely, Rizzo and Bakirov "Measuring and testing dependence by correlation of distances" 2007.

Conditional Distance Covariance[1]

- Let Z be another random vector.
- Conditional Distance Covariance (CdCov) of random vectors X and Y is defined as follows

$$CdCov^2(X,Y|Z=z) = \frac{1}{c_n c_m}\int_{\mathbf{R}^{m+n}} \frac{|f_{X,Y|Z}(t,s|z) - f_{X|Z}(t|z)f_{Y|Z}(s|z)|^2}{\|t\|^{n+1}\|s\|^{m+1}}\, dtds$$

- Clearly, $CdCov(X,Y|Z=z) = 0 \iff X \perp\!\!\!\perp Y|Z$.
- Let we observed i.i.d. values $(X_1,Y_1,Z_1),\ldots,(X_n,Y_n,Z_n)$.
- Define distance matrices $d^X = (d^X_{ij})^n_{i,j=1}$ and $d^Y = (d^Y_{ij})^n_{i,j=1}$, where $d^X_{ij} := \|X_i - X_j\|$ and $d^Y_{ij} := \|Y_i - Y_j\|$
- Define $d_{ijkl} := (d^X_{ij} + d^X_{kl} - d^X_{ik} - d^X_{jl})(d^Y_{ij} + d^Y_{kl} - d^Y_{ik} - d^Y_{jl})$, which is not symmetric.
- Define a symmetric form as $d^S_{ijkl} := d_{ijkl} + d_{ijlk} + d_{ilkj}$

[1] Wang, Pan, Hu, Tian and Zhang "Conditional Distance Correlation" 2015.

Conditional Distance Covariance[1]

- CdCov can be expressed as

$$CdCov^2(X,Y|Z=z) = \frac{1}{12}E[d^S_{1234}|Z_1=z, Z_2=z, Z_3=z, Z_4=z].$$

- For a vector ω, let $K_H(\omega) := |det(H)|^{-1}K(H\omega)$ be a kernel function, where H is a diagonal matrix $diag(h,\ldots,h)$ with a bandwidth parameter h.
- In practice K_H is usually the Gaussian kernel: $K_H(\omega) = (2\pi)^{-\frac{r}{2}}|det(H)|^{-1}exp(-\frac{1}{2}\omega^T H^{-2}\omega)$, where $\omega \in \mathbf{R}^r$.
- Let $K_{iu} := K_H(Z_i - Z_u)$ and $K_i(Z) := K_H(Z - Z_i)$ for all $i,u \in [1,n]$.
- Estimator for $CdCov^2(X,Y|Z)$ can be constructed in the following way which converges with probability to the population value as sample size grows

$$CdCov_n^2(X,Y|Z=z) := \sum_{i,j,k,l} \frac{K_i(Z)K_j(Z)K_k(Z)K_l(Z)}{12(\sum_{i=1}^n K_i(Z))^4} d^S_{ijkl}.$$

[1] Wang, Pan, Hu, Tian and Zhang "Conditional Distance Correlation" 2015.

CdCov Testing[1]

- Let $\rho_0^*(X, Y|Z) := E[CdCov_n^2(X, Y|Z)]$, which does not depend on Z.
- Since $CdCov^2(X, Y|Z)$ is non-negative, $\rho_0^*(X, Y|Z) = 0$ if and only if $X \perp\!\!\!\perp Y|Z$.
- Consider a plug-in estimate of $\rho_0^*(X, Y|Z)$ as

$$\hat{\rho}^*(X, Y|Z) := \frac{1}{n}\sum_{u=1}^{n} CdCov_n^2(X, Y|Z_u) = \frac{1}{n}\sum_{i,j,k,l} \frac{K_{iu}K_{ju}K_{ku}K_{lu}}{12(\sum_{i=1}^{n} K_{iu})^4} d_{ijkl}^S.$$

- We reject $H_0 : X \perp\!\!\!\perp Y|Z$ vs $H_1 : X \not\!\perp\!\!\!\perp Y|Z$ at level $\alpha \in (0, 1)$ if $\hat{\rho}^*(X, Y|Z) > \xi_{n,\alpha}$, where the threshold $\xi_{n,\alpha}$ is obtained by a local bootstrap procedure.

[1] Chakraborty and Shojaie "Nonparametric causal structure learning in high dimensions" 2021.

Exercises

1. If $P^{\mathbf{X}}$ is Markov and faithful with respect to graph $\mathcal{G}$, then $P^{\mathbf{X}}$ satisfies causal minimality with respect to $\mathcal{G}$. (Hint: use exercise 4 from previous lecture)
2. In Table 1 calculate G-test statistic and compare it with the Pearson's chi-squared test statistic (which is bigger and how we can interpret it?). Test the null hypothesis at the 5% and 1% significance levels. What is the p-value of test statistic?
3. Show that d^S_{ijkl} is indeed symmetric.
4. Show that for $n \geq 4$ we have

$$\hat{D} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(o_{ij} - e_{ij})^2}{n(n-3)} - \frac{4\sum_{i=1}^{I}\sum_{j=1}^{J} o_{ij}e_{ij}}{n(n-2)(n-3)}$$
$$+ \frac{\sum_{i=1}^{I} o_{i+}^2 + \sum_{j=1}^{J} o_{+j}^2}{n(n-1)(n-3)} + \frac{(3n-2)(\sum_{i=1}^{I} o_{i+}^2)(\sum_{j=1}^{J} o_{+j}^2)}{n^3(n-1)(n-2)(n-3)} - \frac{n}{(n-1)(n-3)}$$

and explain why in the setting of USP test it is enough to consider only the value $\hat{U}$.