Conditional independence & statistics primer

Kaie Kubjas, 30.09.2020

- Homework deadline is on Friday at 23:59
- Exercise session this week: discussion of Homework 2 in breakout rooms
- Need to receive at the end of the course 60% of the homework points
- Optional extra homework: can be submitted any time during the course (50% of a regular homework)
- If you missed a reading task: contact me

Agenda

- Discrete conditional independence models
- Gaussian conditional independence models
- Primary decompositions of conditional independence ideals
- Statistics primer

Conditional independence

Conditional independence

<u>Def:</u> Let $A, B, C \subseteq [m]$ be pairwise disjoint subsets. We say that X_A is conditionally independent of X_B given X_C if and only if

$$f_{A \cup B|C}(x_A, x_B \,|\, x_C) = f_{A|C}(x_A \,|\, x_C) f_{B|C}(x_B \,|\, x_C)$$

for all x_A, x_B, x_C.

- Notation $X_A \perp\!\!\!\perp X_B \,|\, X_C$ or $A \perp\!\!\!\perp B \,|\, C$.

Discrete conditional independence models

Discrete random variables

- A vector of discrete random variables $X = (X_1, \dots, X_m)$
- X_j takes values in $[r_j]$
- X takes values in the Cartesian product $\mathscr{R} = \prod_{j=1}^{m} [r_j]$
- For $A \subseteq [m]$, let $X_A = (X_a)_{a \in A}$ and $\mathscr{R}_A = \prod_{a \in A} [r_a]$

Marginal distribution

Let $A, B, C \subseteq [m]$ be pairwise disjoint subsets. The notation $p_{i_A, i_B, i_C, +}$ denotes the marginal probability $P(X_A = i_A, X_B = i_B, X_C = i_C)$ which can be written as

$$p_{i_A, i_B, i_C, +} = \sum_{j_{[m]\setminus A \cup B \cup C} \in \mathscr{R}_{[m]\setminus A \cup B \cup C}} p_{i_A, i_B, i_C, j_{[m]\setminus A \cup B \cup C}}.$$

Discrete conditional independence

Prop: If X is a discrete random vector, then the conditional independence statement $X_A \perp\!\!\!\perp X_B \,|\, X_C$ holds if and only if

$$p_{i_A,i_B,i_C,+} \cdot p_{j_A,j_B,i_C,+} - p_{i_A,j_B,i_C,+} \cdot p_{j_A,i_B,i_C,+} = 0$$

for all $i_A, j_A \in \mathscr{R}_A$, $i_B, j_B \in \mathscr{R}_B$ and $i_C \in \mathscr{R}_C$.

Example: Let $m = 2$. Then $X_1 \perp\!\!\!\perp X_2$ holds if and only if

$$p_{i_1,j_1}p_{i_2,j_2} - p_{i_1,j_2}p_{i_2,j_1} = 0 \text{ for all } i_1, i_2 \in [r_1], j_1, j_2 \in [r_2].$$

$$X_A \perp\!\!\!\perp X_B \mid X_C \Leftrightarrow p_{i_A,i_B,i_C,+} \cdot p_{j_A,j_B,i_C,+} - p_{i_A,j_B,i_C,+} \cdot p_{j_A,i_B,i_C,+} = 0$$

Discrete conditional independence ideal

Def: The conditional independence ideal $I_{A \perp\!\!\!\perp B|C}$ is generated by the polynomials $p_{i_A,i_B,i_C,+} \cdot p_{j_A,j_B,i_C,+} - p_{i_A,j_B,i_C,+} \cdot p_{j_A,i_B,i_C,+}$ for all $i_A, j_A \in \mathscr{R}_A, i_B, j_B \in \mathscr{R}_B$ and $i_C \in \mathscr{R}_C$.

Example: Let $m = 2$ and consider the ordinary independence statement $X_1 \perp\!\!\!\perp X_2$. Then

$$I_{1 \perp\!\!\!\perp 2} = \langle p_{i_1,j_1} p_{i_2,j_2} - p_{i_1,j_2} p_{i_2,j_1} : i_1, i_2 \in [r_1], j_1, j_2 \in [r_2] \rangle.$$ [poll]

Conditional independence ideal

Def: If $\mathscr{C} = \{X_{A_1} \perp\!\!\!\perp X_{B_1} \,|\, X_{C_1}, X_{A_2} \perp\!\!\!\perp X_{B_2} \,|\, X_{C_2}, \ldots\}$ is a set of conditional independence statements, then the conditional independence ideal is defined as

$$I_{\mathscr{C}} = \sum_{A \perp\!\!\!\perp B | C \in \mathscr{C}} I_{A \perp\!\!\!\perp B | C}.$$

Discrete conditional independence model

Def: The probability simplex in $\mathbb{R}^{\mathscr{R}}$ is

$$\Delta_{\mathscr{R}} = \left\{ p \in \mathbb{R}^{\mathscr{R}} : \sum_{i \in \mathscr{R}} p_i = 1, p_i \geq 0 \text{ for all } i \right\}.$$

Def: The discrete conditional independence model $\mathscr{M}_{\mathscr{C}} := V(I_{\mathscr{C}}) \cap \Delta_{\mathscr{R}} \subseteq \Delta_{\mathscr{R}}$ consists of all probability distributions that satisfy all the conditional independence statements in $\mathscr{C}$. [poll]

Gaussian conditional independence models

Multivariate normal distribution

Let PD_m be the set of $m \times m$ symmetric positive definite matrices.

Def: Suppose $\mu \in \mathbb{R}^m$ and $\Sigma \in PD_m$. Then a random vector $X = (X_1, \ldots, X_m)$ is distributed according to the multivariate normal distribution $\mathcal{N}_m(\mu, \Sigma)$ if it has the density function

$$\phi_{\mu,\Sigma}(y) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu) \right\}.$$

Gaussian conditional independence models

<u>Prop:</u> The conditional independence statement $X_A \perp\!\!\!\perp X_B \mid X_C$ holds for a multivariate normal random vector $X \sim \mathcal{N}(\mu, \Sigma)$ if and only if the submatrix $\Sigma_{A\cup C, B\cup C}$ of the covariance matrix Σ has rank $\#C$. [poll]

- The set of symmetric matrices of rank at most k is an algebraic variety defined by $(k+1)\times(k+1)$ subdeterminants.
- The $(k+1)\times(k+1)$ subdeterminants are called the $(k+1)$-minors.

Gaussian conditional independence ideal

Def: The Gaussian conditional independence ideal $J_{A \perp\!\!\!\perp B|C}$ is the following ideal in $\mathbb{R}[\sigma_{ij}, 1 \leq i \leq j \leq m]$:

$$J_{A \perp\!\!\!\perp B|C} = \langle (\#C + 1) \text{ minors of } \Sigma_{A \cup C, B \cup C} \rangle.$$

Def: If $\mathscr{C}$ is a collection of conditional independence statements, then the conditional independence ideal is defined as

$$J_{\mathscr{C}} = \sum_{A \perp\!\!\!\perp B|C \in \mathscr{C}} J_{A \perp\!\!\!\perp B|C}$$

Gaussian conditional independence model

Def: The Gaussian conditional independence model is a subset of PD_m, the set of $m \times m$ symmetric positive definite matrices:

$$\mathscr{M}_{\mathscr{C}} := V(J_{\mathscr{C}}) \cap PD_m.$$

Gaussian conditional independence

- Let $m = 3$ and $\mathscr{C} = \{1 \perp\!\!\!\perp 3, 1 \perp\!\!\!\perp 3 \mid 2\}$.
- Then

$$J_{\mathscr{C}} = \langle \sigma_{13}, \det \Sigma_{\{1,2\},\{2,3\}} \rangle.$$

- The Gaussian conditional independence model consists of all covariance matrices $\Sigma \in PD_3$ satisfying $\sigma_{13} = 0$ and $\sigma_{12}\sigma_{23} - \sigma_{13}\sigma_{22} = 0$.
- Alternatively we can consider $\sigma_{13} = 0$ and $\sigma_{12}\sigma_{23} = 0$.

Gaussian conditional independence

- Alternatively we can consider $\sigma_{13} = 0$ and $\sigma_{12}\sigma_{23} = 0$.

- The solutions to these equations are given by the union of two linear spaces:

$$L_1 = \{\Sigma : \sigma_{13} = \sigma_{12} = 0\}, \quad L_2 = \{\Sigma : \sigma_{13} = \sigma_{23} = 0\}.$$

- These components correspond to $X_1 \perp\!\!\!\perp (X_2, X_3)$ and $X_3 \perp\!\!\!\perp (X_1, X_2)$.

- Hence $X_1 \perp\!\!\!\perp X_3$ and $X_1 \perp\!\!\!\perp X_3 \,|\, X_2 \implies X_1 \perp\!\!\!\perp (X_2, X_3)$ or $X_3 \perp\!\!\!\perp (X_1, X_2)$.

Primary decomposition

Primary decomposition

- An ideal Q is called primary if $f \cdot g \in Q$ implies that either $f \in Q$ or $g^k \in Q$ for some $k \in \mathbb{N}$.
- A primary decomposition of an ideal I is a representation $I = Q_1 \cap \cdots \cap Q_r$ where each Q_i is a primary ideal.
- Every ideal has a primary decomposition. A minimal primary decomposition can be computed in Macaulay2.

Irreducible decompositions

- A variety V is called reducible if there exist varieties $V_1, V_2 \subsetneq V$ such that $V = V_1 \cup V_2$. A variety that is not reducible, is called irreducible.

- A primary decomposition of an ideal I, gives a decomposition of $V(I)$:

$$V(I) = V(Q_1) \cup \cdots \cup V(Q_r).$$

Primary decomposition of CI ideals

Intersection axiom

Prop (Intersection axiom): Suppose that $f(x) > 0$ for all x. Then

$$X_A \perp\!\!\!\perp X_B \,|\, X_{C \cup D} \text{ and } X_A \perp\!\!\!\perp X_C \,|\, X_{B \cup D} \implies X_A \perp\!\!\!\perp X_{B \cup C} \,|\, X_D.$$

- The condition $f(x) > 0$ for all x is stronger than necessary.

Failure of the intersection axiom

- Let X_1, X_2, X_3 be binary random variables.
- Let $\mathscr{C} = \{1 \perp\!\!\!\perp 2 \,|\, 3, 1 \perp\!\!\!\perp 3 \,|\, 2\}$.
- Intersection axiom:

$$X_A \perp\!\!\!\perp X_B \,|\, X_{C \cup D} \text{ and } X_A \perp\!\!\!\perp X_C \,|\, X_{B \cup D} \implies X_A \perp\!\!\!\perp X_{B \cup C} \,|\, X_D \text{ [poll]}$$

- $A = \{1\}, B = \{2\}, C = \{3\}, D = \varnothing$
- Hence $X_A \perp\!\!\!\perp X_{B \cup C} \,|\, X_D$ is $X_1 \perp\!\!\!\perp (X_2, X_3)$

Failure of the intersection axiom

- The CI ideal is generated by four 2×2-minors of the matrix

$$\begin{pmatrix} p_{111} & p_{112} & p_{121} & p_{122} \\ p_{211} & p_{212} & p_{221} & p_{222} \end{pmatrix}.$$

- The CI ideal has the primary decomposition

$$\mathscr{C}_I = I_{1 \perp\!\!\!\perp \{2,3\}} \cap \langle p_{111}, p_{211}, p_{122}, p_{222} \rangle \cap \langle p_{112}, p_{212}, p_{121}, p_{221} \rangle.$$

- The first component corresponds to the conclusion of the intersection axiom.
- The other components correspond to families of probability distributions that might not satisfy the conclusion of the intersection axiom.

Failure of the intersection axiom

- For discrete random variables, precise conditions can be given which guarantee that the intersection axiom holds.
- The condition is given in terms of a certain graph having one connected component.
- See Chapter 4.3.1 in “Algebraic Statistics”

Conclusion

- CI ideal associated to a set of conditional independence statements
- Discrete case: The variety of the CI ideal intersected with the probability simplex consists of these joint probabilities that satisfy the CI statements
- Gaussian case: The variety of the CI ideal intersected with the positive definite cone gives these densities that satisfy the CI statements
- Primary decompositions of ideals are used to study CI implications
- We will return to conditional independence statements in the graphical models section

Statistics primer

1. What is the difference between probability and statistics?

- In probability, we assume the probability distributions are known. In statistics, we start from data, and infer certain properties of the underlying distribution (possibly with hypothesis testing).

- In the case of probability, we already know the distribution with which we are working and want to know more about its characteristics and how we can change some of the features. In statistics, we are presented with some sampled data and have to make an educated guess to which distribution the sample set could belong.

- Probability and statistics are two sides of the same coin.

Statistical models

- A statistical model is a collection of density functions or probability distributions.
- A parametric statistical model is the image of a mapping from a finite dimensional parameter space $\Theta \subseteq \mathbb{R}^d$ to a space of density functions or probability distributions, i.e. $p_\star : \Theta \to \mathscr{M}_\Theta, \theta \mapsto p_\theta$.
- An implicit statistical model is defined via constraints on densities or probability distributions. [poll]

2. Can a model be parametric and implicit?

- Yes, for example the model of independence (Example 5.1.4).
- Let X_1 and X_2 be two discrete random variables with state spaces $[r_1]$ and $[r_2]$. Let $\mathscr{R} = [r_1] \times [r_2]$.
- Implicit description: The model of independence consist of all distributions $p \in \Delta_{\mathscr{R}}$ such that $P(X_1 = i_1, X_2 = i_2) = P(X_1 = i_1)P(X_2 = i_2)$.
- Parametric description: Let $\Theta = \Delta_{r_1 - 1} \times \Delta_{r_2 - 1}$ and $\theta = (\alpha, \beta) \in \Theta$. Then $P_\theta(X_1 = i_1, X_2 = i_2) = \alpha_{i_1}\beta_{i_2}$.
- How would you get the implicit description from the parametric description?

3. The book uses $X_1, \ldots, X_m$ and $X^{(1)}, \ldots, X^{(n)}$. What is the difference between the two notations?

- $X_1, \ldots, X_m$ denote random variables with underlying distributions, whose values are generally assumed as unknown. $X^{(1)}, \ldots, X^{(n)}$ are data points, or specific instances / realizations of the random variables.

Data

- Independent and identically distributed data $D = \{X^{(1)}, X^{(2)}, \dots, X^{(n)}\}$ means that $X^{(i)}$ are realizations of random variables that have the same distribution and that are mutually independent

- Independent and identically distributed = i.i.d.

- Discrete case: The probability of observing the data D is

$$p_\theta(D) = \prod_{i=1}^{n} p_\theta(X^{(i)}).$$

Data

- Discrete case: If the random variable has the state space $[r]$, then we can define the vector of counts $u \in \mathbb{N}^r$ by

$$u_j = \#\{i : X^{(i)} = j\}.$$

- The probability of observing data D becomes

$$p_\theta(D) = \prod_{i=1}^{r} p_\theta(j)^{u_j}.$$

Next time

- Exponential families or likelihood inference
- Group work topics: the method of moments, the cone of sufficient statistics, exponential random graph models, phylogenetic models