Matroids and statistical dependency

Art Duval, Amy Wagler

University of Texas at El Paso

Discrete Math Seminar
Texas State University
March 2, 2018

Set dependence

- Can three variables be somehow (statistically) dependent, even when no two of them are?

Set dependence

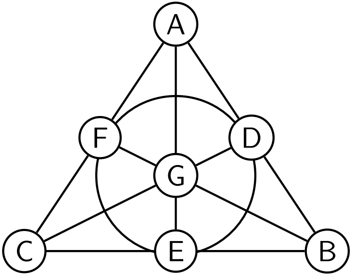
- Can three variables be somehow (statistically) dependent, even when no two of them are?
- Yes. For instance, $Z = 1 + XY + \epsilon$.

Set dependence

- Can three variables be somehow (statistically) dependent, even when no two of them are?
- Yes. For instance, $Z = 1 + XY + \epsilon$.
- We might expect to get any sort of simplicial complex (subsets of independent sets are independent).

Set dependence

- Can three variables be somehow (statistically) dependent, even when no two of them are?
- Yes. For instance, $Z = 1 + XY + \epsilon$.
- We might expect to get any sort of simplicial complex (subsets of independent sets are independent).
- We can even get the Fano plane: A, B, C independent, $D = AB, E = BC, F = CA, G = DEF$.



Matroids

If we are in a situation where set dependence gives us a matroid, this would be useful to statisticians in at least two ways:

Matroids

If we are in a situation where set dependence gives us a matroid, this would be useful to statisticians in at least two ways:

- In regression modeling, matroid structures could be used as a variable selection procedure to find the most parsimonious set of X's to predict a Y. The results of the minimally dependent sets [circuits] would also inform which interactions ($x_1 x_2$ products) should be investigated for inclusion to the model.
- In big data settings, a matroid would identify maximally independent sets [bases] so that multiplicity can be corrected at the circuit level rather than the full data set.

How to picture data

Each variable is a vector, whose components are measurements of this variable.

- m different variables
- n different trials
- m vectors in $\mathbb{R}^n$

Example

Three variables, four trials

$$\begin{aligned} X &= \begin{pmatrix} 3.1 & 1 & 4 & 2 \end{pmatrix} \\ Y &= \begin{pmatrix} 2 & 1 & 6.9 & 8 \end{pmatrix} \\ Z &= \begin{pmatrix} 5 & 2.1 & 11 & 9.9 \end{pmatrix} \end{aligned}$$

Dependence

Example

$$\begin{aligned} X &= (3.1 \quad 1 \quad 4 \quad 2\ ) \\ Y &= (\ 2 \quad 1 \quad 6.9 \quad 8\ ) \\ Z &= (\ 5 \quad 2.1 \quad 11 \quad 9.9) \end{aligned}$$

- Knowing the value of any two of X, Y, Z tells you approximately the value of the third;

Dependence

Example

$$
\begin{aligned}
X &= (3.1 \quad 1 \quad 4 \quad 2\ ) \\
Y &= (\ 2 \quad 1 \quad 6.9 \quad 8\ ) \\
Z &= (\ 5 \quad 2.1 \quad 11 \quad 9.9)
\end{aligned}
$$

- Knowing the value of any two of X, Y, Z tells you approximately the value of the third;
- but knowing only one variable tells you nothing about either of the others.

Dependence

Example

$$
\begin{aligned}
X &= (3.1 \quad 1 \quad 4 \quad 2\ ) \\
Y &= (\ 2 \quad 1 \quad 6.9 \quad 8\ ) \\
Z &= (\ 5 \quad 2.1 \quad 11 \quad 9.9)
\end{aligned}
$$

- Knowing the value of any two of X, Y, Z tells you approximately the value of the third;
- but knowing only one variable tells you nothing about either of the others.

Dependence

Example

$$\begin{aligned} X &= (3.1 \quad 1 \quad 4 \quad 2\ ) \\ Y &= (\ 2 \quad 1 \quad 6.9 \quad 8\ ) \\ Z &= (\ 5 \quad 2.1 \quad 11 \quad 9.9) \end{aligned}$$

- Knowing the value of any two of X, Y, Z tells you approximately the value of the third;
- but knowing only one variable tells you nothing about either of the others.

So this set is (minimally) dependent.

How to measure dependence

Example

$$\begin{aligned} X &= (3.1 \quad 1 \quad 4 \quad 2) \\ Y &= (2 \quad 1 \quad 6.9 \quad 8) \\ Z &= (5 \quad 2.1 \quad 11 \quad 9.9) \end{aligned}$$

Question

How can we identify statistically independent sets in general? And capture non-linear dependence? What is "close enough"?

How to measure dependence

Example

$$\begin{aligned} X &= (3.1 \quad 1 \quad 4 \quad 2\ ) \\ Y &= (\ 2 \quad 1 \quad 6.9 \quad 8\ ) \\ Z &= (\ 5 \quad 2.1 \quad 11 \quad 9.9) \end{aligned}$$

Question

How can we identify statistically independent sets in general? And capture non-linear dependence? What is "close enough"?

We will use

- Effective dependence
- Joint cumulants

These appear to be consistent measures of dependence.

Effective dependence

Effective dependence $= 1 - \Psi$, where

$$\Psi = \frac{|\det \Sigma|^{1/m}}{(\sum \lambda_i)/m} = \frac{\text{geometric mean}}{\text{arithmetic mean}}$$

is sphericity;

- Σ is covariance matrix (pairwise covariance of variables);
- λ_i are eigenvalues of Σ.

Joint cumulants

Definition

$$\prod_{a=1}^{b(\tau)} E(\prod_{i \in \tau_a} X_i) = \sum_{\sigma \leq \tau} \kappa_\sigma$$

By Möbius inversion, we can solve for κ's.

Example

$$E(X_1)E(X_2)E(X_3)E(X_4) = \kappa_{1|2|3|4}$$
$$E(X_1X_2)E(X_3)E(X_4) = \kappa_{1|2|3|4} + \kappa_{12|3|4}$$

So $\kappa_{12|3|4} = (E(X_1X_2) - E(X_1)E(X_2))E(X_3)E(X_4)$

Joint cumulants

Definition

$$\prod_{a=1}^{b(\tau)} E\left(\prod_{i\in\tau_a} X_i\right) = \sum_{\sigma\leq\tau} \kappa_\sigma$$

By Möbius inversion, we can solve for κ's.

Example

$$E(X_1)E(X_2)E(X_3)E(X_4) = \kappa_{1|2|3|4}$$
$$E(X_1X_2)E(X_3)E(X_4) = \kappa_{1|2|3|4} + \kappa_{12|3|4}$$

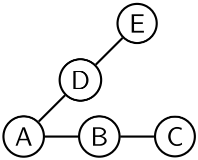
So $\kappa_{12|3|4} = (E(X_1X_2) - E(X_1)E(X_2))E(X_3)E(X_4)$

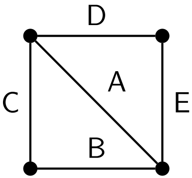
Our test of set dependence: If there is a partition of a set into two parts such that there is a cumulant dependence $\kappa_{\alpha|\beta} \neq 0$.

Matroids

Matroids make abstract ideas of independence, and model

- linear independence and dependence of sets of vectors in linear algebra;
- independent (cycle-free) sets of edges in graphs;
- etc.

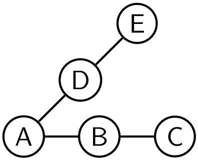


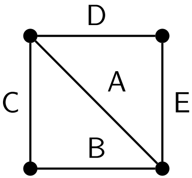


Matroids

Matroids make abstract ideas of independence, and model

- linear independence and dependence of sets of vectors in linear algebra;
- independent (cycle-free) sets of edges in graphs;
- etc.



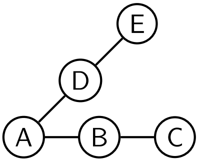


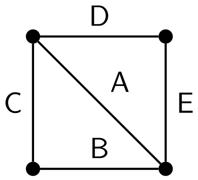
Remark

Not all matroids can be represented by vectors or graphs

Independent sets

- $\emptyset$ is independent.
- Any subset of an independent set is also independent.
- If I_1, I_2 independent, and $|I_2| = |I_1| + 1$, then $\exists x \in I_2 - I_1$ such that $I_1 \cup \{x\}$ is independent.

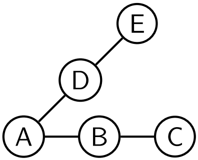


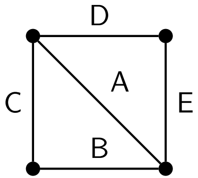


Bases

Maximally independent sets

- $\emptyset$ is not a basis.
- One basis cannot be a proper subset of another basis.
- If B_1, B_2 are bases and $x \in B$, then $\exists y \in B_2$ such that $(B_1 - \{x\}) \cup \{y\}$ is a basis.

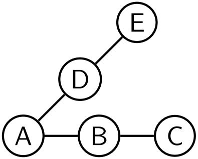


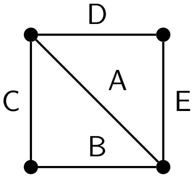


Circuits

Minimally dependent sets

- $\emptyset$ is not a circuit.
- One circuit cannot be a proper subset of another circuit.
- $(C_1 \cup C_2) - \{x\}$ contains a circuit for distinct circuits C_1, C_2.

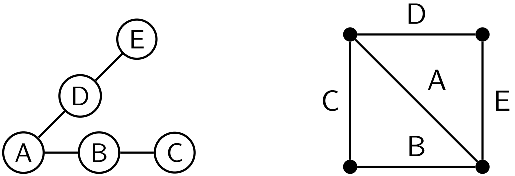




Rank function

Size of maximal independent subset of a set

- $r(\emptyset) = 0$.
- $r(A \cup \{x\}) = r(A)$ or $r(A) + 1$.
- If $r(A) = r(A \cup \{x\}) = r(A \cup \{y\})$, then $r(A \cup \{x, y\}) = r(A)$.



Closure axioms

A matroid on ground set E may be defined by closure axioms:

$$\mathrm{cl}\colon 2^E \to 2^E$$

- Closure axioms:
  - $A \subseteq \mathrm{cl}(A)$
  - If $A \subseteq B$, then $\mathrm{cl}(A) \subseteq \mathrm{cl}(B)$
  - $\mathrm{cl}(\mathrm{cl}(A)) = \mathrm{cl}(A)$
- Exchange axiom: If $x \in \mathrm{cl}(A \cup y) - \mathrm{cl}(A)$, then $y \in \mathrm{cl}(A \cup x)$

For us, $x \in \mathrm{cl}(A)$ means that knowing the values of all the variables in A implies knowing something about the value of x. (Sort of: x is a function of A, with statistical noise and fuzziness.)

Invertibility

Exchange axiom: If $x \in \mathrm{cl}(A \cup y) - \mathrm{cl}(A)$, then $y \in \mathrm{cl}(A \cup x)$

- $x \in \mathrm{cl}(A \cup y) - \mathrm{cl}(A)$ means that in using $A \cup y$ to determine x, we must use (can't ignore) y. ("model parsimony")
- $y \in \mathrm{cl}(A \cup x)$ means we can "solve" for y in terms of x and A. (This is sort of invertibility.)

Invertibility

Exchange axiom: If $x \in \mathrm{cl}(A \cup y) - \mathrm{cl}(A)$, then $y \in \mathrm{cl}(A \cup x)$

- $x \in \mathrm{cl}(A \cup y) - \mathrm{cl}(A)$ means that in using $A \cup y$ to determine x, we must use (can't ignore) y. ("model parsimony")
- $y \in \mathrm{cl}(A \cup x)$ means we can "solve" for y in terms of x and A. (This is sort of invertibility.)

Easiest way for a function (only way for continuous function) to be invertible is to be monotone in each variable. Fortunately, implied by a common statistical assumption:

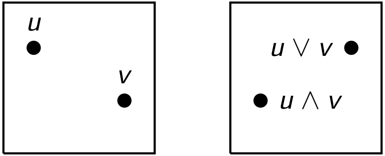
Definition (MTP_2)

(Multivariate Totally Positive of order 2.)
$f(u)f(v) \leq f(u \wedge v)f(u \vee v)$, where f is probability distribution, u and v are vectors of variable values, and $\wedge$ and $\vee$ denote element-wise minimum and maximum.

Multivariate Totally Positive of order 2

Definition (MTP_2)

$f(u)f(v) \leq f(u \wedge v)f(u \vee v)$, where f is probability distribution, u and v are vectors of variable values, and $\wedge$ and $\vee$ denote element-wise minimum and maximum.



Composition

Closure axioms

- $A \subseteq \mathrm{cl}(A)$ (easy)
- If $A \subseteq B$, then $\mathrm{cl}(A) \subseteq \mathrm{cl}(B)$ (easy)
- $\mathrm{cl}(\mathrm{cl}(A)) = \mathrm{cl}(A)$ (not so easy)

Composition

Closure axioms

- $A \subseteq \mathrm{cl}(A)$ (easy)
- If $A \subseteq B$, then $\mathrm{cl}(A) \subseteq \mathrm{cl}(B)$ (easy)
- $\mathrm{cl}(\mathrm{cl}(A)) = \mathrm{cl}(A)$ (not so easy)

Example

When $A = x$ is a single element and $\mathrm{cl}(x) = \{x, y\}$. We need to avoid $z \in \mathrm{cl}\{x, y\}$ for $z \neq x, y$. In other words, z depends on y, and y depends on x should mean that z depends on x directly. This is a kind of transitivity.

Composition

Closure axioms

- $A \subseteq \mathrm{cl}(A)$ (easy)
- If $A \subseteq B$, then $\mathrm{cl}(A) \subseteq \mathrm{cl}(B)$ (easy)
- $\mathrm{cl}(\mathrm{cl}(A)) = \mathrm{cl}(A)$ (not so easy)

Example

When $A = x$ is a single element and $\mathrm{cl}(x) = \{x, y\}$. We need to avoid $z \in \mathrm{cl}\{x, y\}$ for $z \neq x, y$. In other words, z depends on y, and y depends on x should mean that z depends on x directly. This is a kind of transitivity.

More generally, if Z is determined by $Y_1, \ldots, Y_p$, and each Y_i is determined by $X_1, \ldots, X_q$, then Z should be determined directly by $X_1, \ldots, X_q$. This is a kind of composition.

Remark

MTP_2 means the dependence will be strong enough to guarantee transitivity, and more generally composition.

Dependence axioms

How we actually show that we have a matroid. The dependent sets $\mathcal{D}$ in a matroid satisfy:

- $\emptyset \notin \mathcal{D}$
- If $D \in \mathcal{D}$ and $D' \supseteq D$, then $D' \in \mathcal{D}$
- If $I \notin \mathcal{D}$ but $I \cup x, I \cup y \in \mathcal{D}$, then $(I - z) \cup \{x, y\} \in \mathcal{D}$ for all $z \in I$.

We can prove that MTP_2 distributions satisfy this, using results of Fallat et al. (using that MTP_2 is an upward-stable singleton-transitive compositional semigraphoid).

Example: Cancer genes

Non-matroid analysis: Clusters

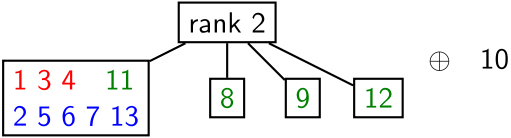
$$\{1,3,4\},\ \{2,5,6,7,13\},\ \{8,9,11,12\},\ \{10\}.$$

Example: Cancer genes

Non-matroid analysis: Clusters

$$\{1, 3, 4\}, \ \{2, 5, 6, 7, 13\}, \ \{8, 9, 11, 12\}, \ \{10\}.$$

Matroid analysis:

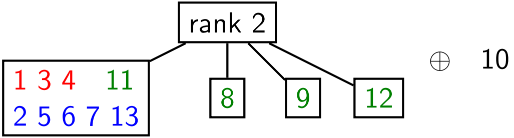


Example: Cancer genes

Non-matroid analysis: Clusters

$$\{1,3,4\},\ \{2,5,6,7,13\},\ \{8,9,11,12\},\ \{10\}.$$

Matroid analysis:

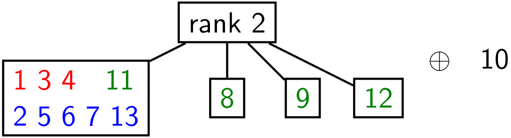


Example: Cancer genes

Non-matroid analysis: Clusters

$$\{1, 3, 4\},\ \{2, 5, 6, 7, 13\},\ \{8, 9, 11, 12\},\ \{10\}.$$

Matroid analysis:



Remark

This suggests two independent, possibly latent, variables explaining the left side of the diagram.