Matroids and statistical dependency

Art Duval, Amy Wagler

University of Texas at El Paso

AMS Central Sectional Meeting
Special Session on Geometric Combinatorics and Combinatorial Commutative Algebra

University of North Texas, Denton
September 9, 2017

AD supported by Simons Foundation Grant 516801

Set dependence

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

Set dependence

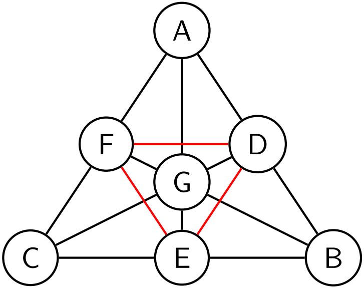
- Can three variables be somehow (statistically) dependent, even when no two of them are?
- Yes. For instance, $Z = 1 + XY + \epsilon$.

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

We even show that, under not uncommon assumptions, set dependence gives us a matroid. Useful to statisticians in at least two ways:

Matroids

We even show that, under not uncommon assumptions, set dependence gives us a matroid. 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 matroid circuits would also inform which interactions ($x_1 x_2$ products) should be investigated for inclusion to the model.

Matroids

We even show that, under not uncommon assumptions, set dependence gives us a matroid. 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 matroid 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

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

How to measure dependence

Note in previous example:

- 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

Note in previous example:

- 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.

Question

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

How to measure dependence

Note in previous example:

- 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.

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 - Ψ, 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.

When does our notion of statistical independence and dependence of sets of variables also lead to a matroid?

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 $x \in \text{cl}(A \cup y) - \text{cl}(A)$, then $y \in \text{cl}(A \cup x)$

- $x \in \text{cl}(A \cup y) - \text{cl}(A)$ means that in using $A \cup y$ to determine x, we must use (can't ignore) y. ("model parsimony")
- $y \in \text{cl}(A \cup x)$ 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, a common statistical assumption:

Definition (PRDS)

(Positive regression dependency on each one from a subset.) For any increasing set D, for for each $i \in I_0$, $P(\mathbf{X} \in D | X_i = x)$ is nondecreasing in x.

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 $z \in \mathrm{cl}\{x, y\}$, but $z \neq x, y$. 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

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

Example

When $A = x$ is a single element and $\text{cl}(x) = \{x, y\}$. We need to avoid $z \in \text{cl}\{x, y\}$, but $z \neq x, y$. 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

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

Dependence axioms

How we actually show, via cumulants, 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 $z \in I$.