Matroids and statistical dependency

Art Duval, Amy Wagler

University of Texas at El Paso

Joint Mathematics Meeting
AMS Contributed Paper Session on
Matrices and Matroids
San Diego
January 12, 2018

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

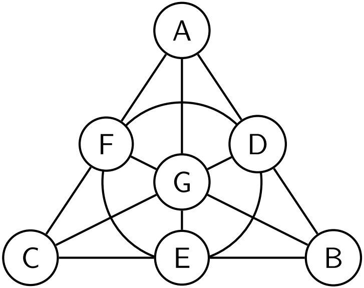
- Can three variables be somehow (statistically) dependent, even when no two of them are?
- Yes. For instance, $Z = 1 + XY + \epsilon$.

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

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 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.

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 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.

So when does this happen?

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

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, 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.

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

- $A \subseteq \text{cl}(A)$ (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 $z \in \text{cl}\{x, y\}$ 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

- $A \subseteq \text{cl}(A)$ (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 $z \in \text{cl}\{x, y\}$ 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).