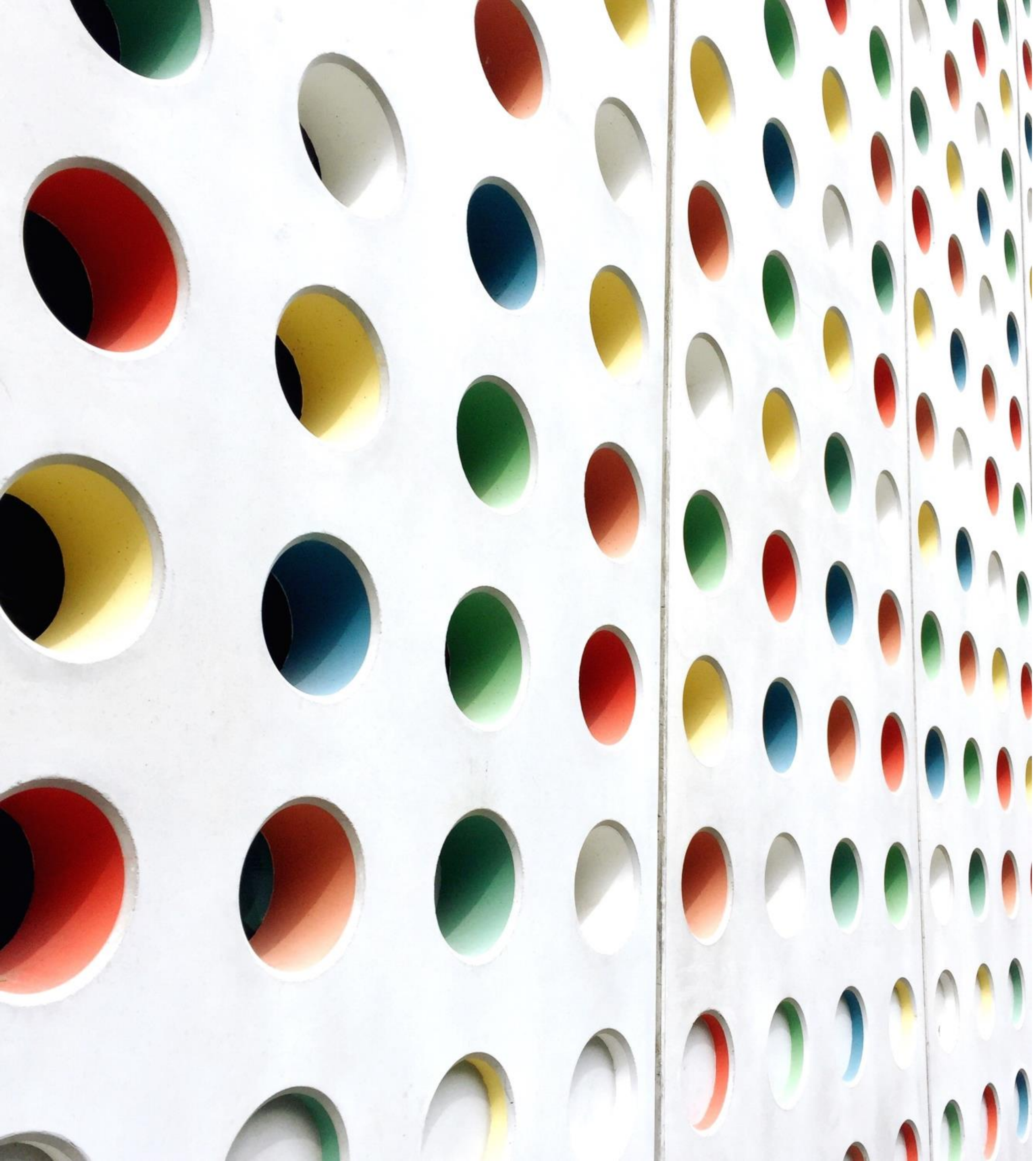
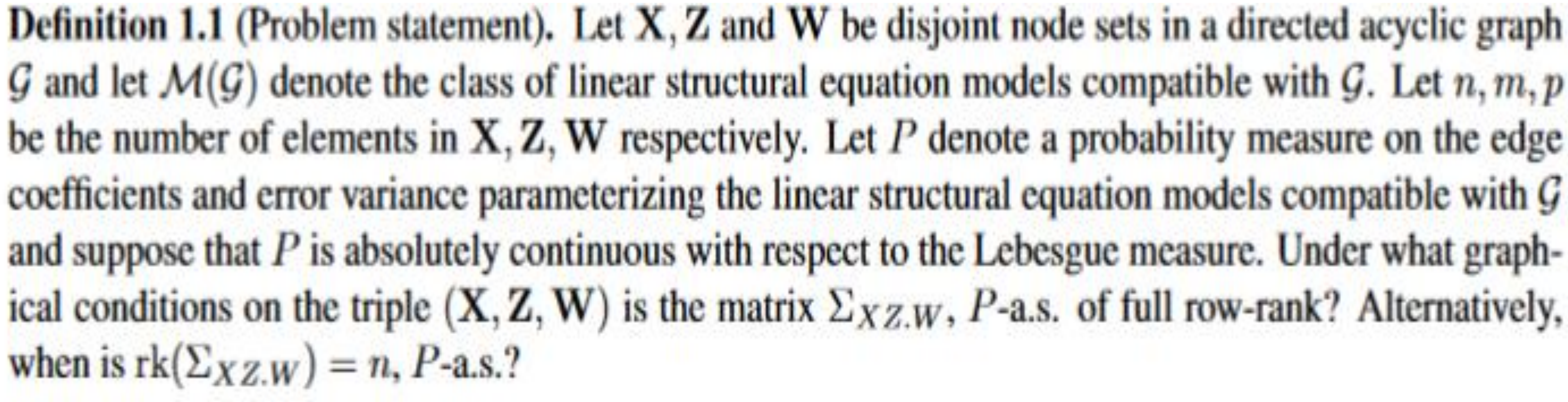


Graphical Conditions for Identifiability of Instrumental Variables in Linear SEMs

Dara Ó Callaráin

Problem Statement

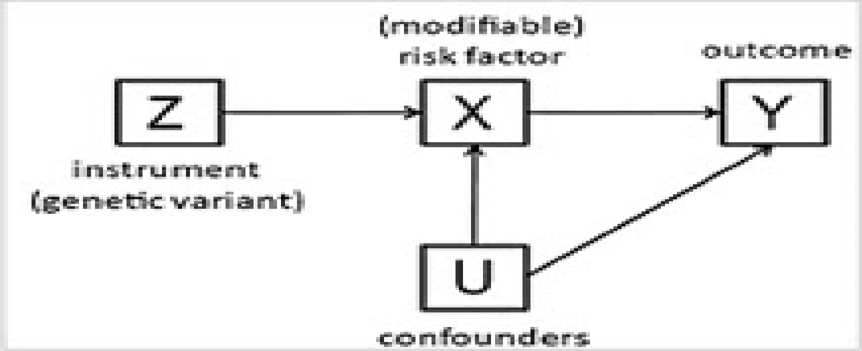


Definition 1.1 (Problem statement). Let $\mathbf{X}, \mathbf{Z}$ and $\mathbf{W}$ be disjoint node sets in a directed acyclic graph $\mathcal{G}$ and let $\mathcal{M}(\mathcal{G})$ denote the class of linear structural equation models compatible with $\mathcal{G}$. Let n, m, p be the number of elements in $\mathbf{X}, \mathbf{Z}, \mathbf{W}$ respectively. Let P denote a probability measure on the edge coefficients and error variance parameterizing the linear structural equation models compatible with $\mathcal{G}$ and suppose that P is absolutely continuous with respect to the Lebesgue measure. Under what graphical conditions on the triple $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ is the matrix $\Sigma_{XZ.W}$, P-a.s. of full row-rank? Alternatively, when is $\text{rk}(\Sigma_{XZ.W}) = n$, P-a.s.?

The goal of this presentation is to transmit an intuition about this problem and its solution.

From Random Variables to DAGs

Graph theory lends us an interesting mathematical setting to consider linear relationships between random variables. Directed Acyclic Graphs (DAGs) encode random variables as nodes. and statistical relationships as arrows, as below.



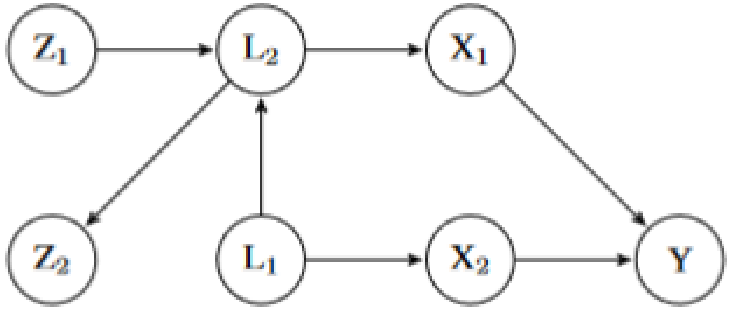
For the above graph for example, *D* is generated linearly from *Z* and *U*, along with an error term assumed to following a standardized normal distribution, N(0,1).

This graph also represents the concept of instrumental variables, the focus of this project. Suppose we were trying to determine the causal effect of *D* on *Y* . Due to the presence of an unmeasured confounding variable, *U*, it is not sufficient to have data on just *D* and *Y*. We say that the relationship between *D* and *Y* is confounded by *U*. To overcome this, we use an instrumental variable, in this case *Z*, which allows us to determine the linear effect of *D* on *Y*.

The goal of this project is to formulate graphical conditions which give that the so-called moment condition on IVs has a unique solution, which is crucial for identifiability.

General Setting for Problem Statement

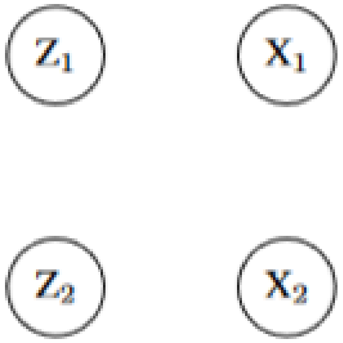
In the general case, which we explore in this project, we allow $\boldsymbol{Z}$ and $\boldsymbol{X}$ to be sets with multiple elements. We also introduce a conditioning set, $\boldsymbol{W}$. Y is still considered as a singleton. All other variables are denoted $\boldsymbol{L}$, giving DAGs like the following:



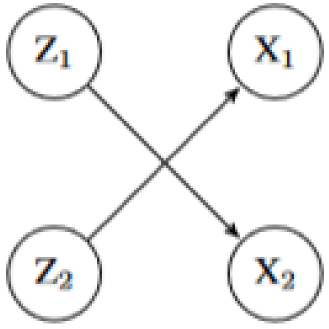
Formally, we ask whether there exists at least one SEM compatible with a given DAG, such that the corresponding covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is full row-rank. This can be thought of as the multi-element analogue of the condition for IVs that $\boldsymbol{Z}$ be strongly correlated with $\boldsymbol{X}$, essentially that $\boldsymbol{Z}$ is a faithful proxy variable for $\boldsymbol{X}$.

Intuitively, we ask whether the structure of the DAG “preserves” enough information about $\boldsymbol{X}$ to be able to deduce the linear relationships using instrumental variables. In the following slide we explore a few examples.

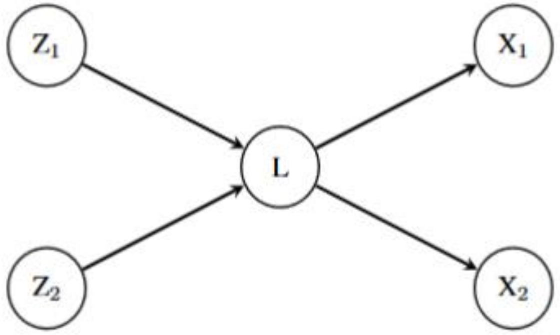
Examples



In this DAG, there is no relationship between Z and X. Z and X are hence independent. It follows that the covariance matrix of interest will not be of full rank.



On the other hand, this DAG will give full row-rank to our covariance matrix. Each Z "faithfully" generates the corresponding X.

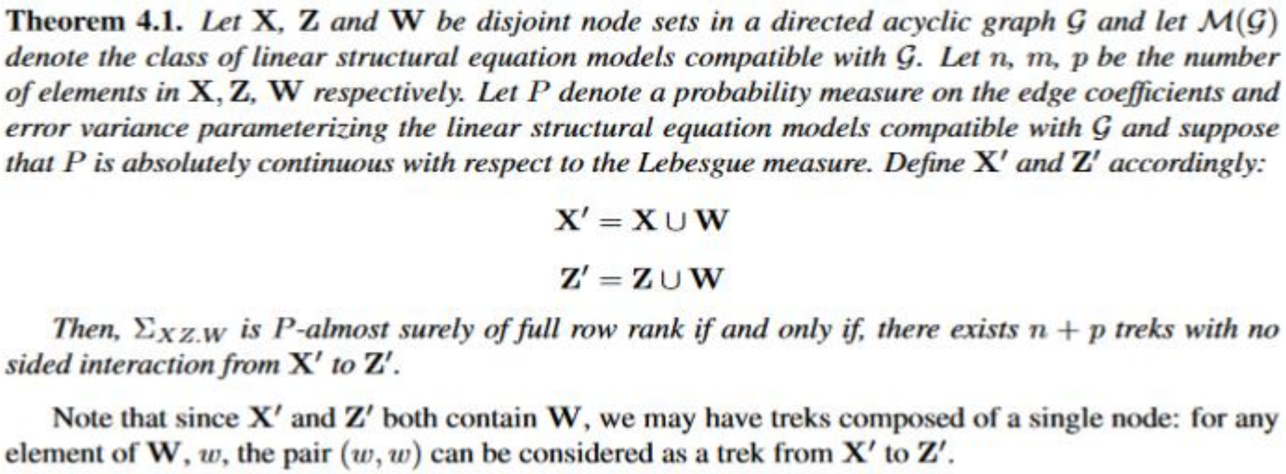


Although this DAG has two treks from Z to X, each trek system has a sided interaction. This gives that this is not full row-rank. Intuitively, the information being transmitted by the treks "interfere" and the rank drops.

It seems that the answer to our question lies in the existence of paths?

Main result

In these slides, we present in some detail the main result of our paper.



Theorem 4.1. *Let* $\mathbf{X}$, $\mathbf{Z}$ *and* $\mathbf{W}$ *be disjoint node sets in a directed acyclic graph* $\mathcal{G}$ *and let* $\mathcal{M}(\mathcal{G})$ *denote the class of linear structural equation models compatible with* $\mathcal{G}$. *Let* n, m, p *be the number of elements in* $\mathbf{X}, \mathbf{Z}, \mathbf{W}$ *respectively. Let* P *denote a probability measure on the edge coefficients and error variance parameterizing the linear structural equation models compatible with* $\mathcal{G}$ *and suppose that* P *is absolutely continuous with respect to the Lebesgue measure. Define* $\mathbf{X}'$ *and* $\mathbf{Z}'$ *accordingly:*

$$\mathbf{X}' = \mathbf{X} \cup \mathbf{W}$$

$$\mathbf{Z}' = \mathbf{Z} \cup \mathbf{W}$$

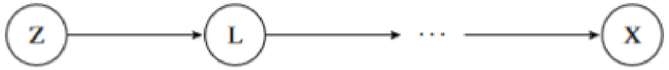
Then, $\Sigma_{XZ.W}$ *is* P*-almost surely of full row rank if and only if, there exists* $n + p$ *treks with no sided interaction from* $\mathbf{X}'$ *to* $\mathbf{Z}'$.

Note that since $\mathbf{X}'$ and $\mathbf{Z}'$ both contain $\mathbf{W}$, we may have treks composed of a single node: for any element of $\mathbf{W}$, w, the pair (w, w) can be considered as a trek from $\mathbf{X}'$ to $\mathbf{Z}'$.

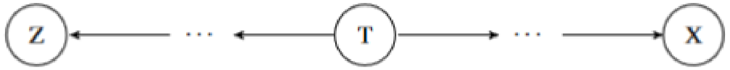
We consider the meaning of treks on the following slide, which form the basis for the above result.

Treks, not Paths!

Surprisingly perhaps, the main mechanism of interest in the solution is not path systems but trek systems. The difference between the two concepts is illustrated below.



Path



Trek

Treks essentially encode confounding relationships, so this reflects that instrument and treatment must be correlated but not necessarily have a direct causal relationship.

The existence of a system of treks from $\boldsymbol{Z}$ to $\boldsymbol{X}$ with the extra condition of no sided interaction will guarantee that our covariance matrix is full row-rank.

Thank you!
Any Questions?