Università degli Studi di Milano - Bicocca

DIPARTIMENTO DI ECONOMIA, METODI QUANTITATIVI E STRATEGIE DI IMPRESA
Corso di Dottorato di Ricerca in Statistica XXVI ciclo

Tesi di dottorato di Ricerca in Statistica

A Causal Graphs-based approach for assessing Gender Gaps with an application to child health & nutrition in China

Candidata:
Caligaris Silvia

Coordinatore e Relatore:
Ch.ma Prof.ssa Fulvia Mecatti

Supervisori della ricerca:
Ch.ma Prof.ssa Franca Crippa
Ch.ma Prof.ssa Patrizia Farina

Academic Year 2012 – 2013

Contents

List of Figures

List of Tables

Introduction

Epidemiology typically deals with *causality*, namely learning what statistical relations imply, or do not imply, about *cause-effect* relations; causal claims like “smoking causes cancer” or “human papilloma virus causes cervical cancer” have long been a standard part of the epidemiology literature, bringing out concepts such as “exposure”, “outcome”, “confounder” etc.
Most of research purposes in health, social and behavioural sciences have, likewise, causal nature. Consider, for example, questions as: what is the efficacy of a treatment for a given illness? does a reduction of taxation lead to an increase in consumption? what are the main factors causing alcoholism? These are all causal matters, since they require some knowledge of the data-generating process and they cannot be computed from the data alone, nor from the distributions originating data.

This thesis aims to bring together causality and gender studies. With a particular focus on assessing *gender gap*, we would aim at developing sound causal-statistics tools able to answer to causal questions such as: does gender affect wages? can data prove an employer guilty of labour recruiting gendered discrimination?
Our purpose is on learning this kind of cause-effect relationships from observational data, and in addition compiling them into a consistent mesh that may be used to describe the mechanism originating gender gap, or at least an appropriate abstraction of it. We shall call such meshes or networks of cause-effect relationships *causal models.*
Anyway, in order to build up a bridge between causality and gender gap analysis, we need to formulate two *translation devices*: the first from the language of causality to standard statistical language of probability distribu-

tions, while the second from statistical to the sociological language of "gender studies".

Chapter 1 deals with the translation from causality to statistics; answering causal questions, indeed, systematically requires extensions in the standard mathematical language of statistics. In particular, recent developments in graphical models allowed understanding of the relationships between graphs and probabilities on one hand, and graphs and causal inference on the other, allowing to bring causality back into statistical modelling and analysis (Pearl, 1998).

Such tools, which we refer to as *causal graphs*, are a class of models that provide:

- a notational system for concepts and relationships that do not easily find an equivalent expression in the standard mathematical language of algebraic equations and probability calculus;
- a simple, flexible device to clearly display the causal structure relating variables even in case of complex systems of variables as well as big-data;
- an intuitive visual tool capable of representing direct cause-effect relationships as well as indirect causation by means of a graph;
- a powerful symbolic machinery for deriving the consequence of causal assumptions when such assumptions are combined with statistical data.

Such approach, thus, combines features of structural equation models (SEMs) used in economics and social science (Goldberger, 1973; Duncan, 1975), potential-outcome framework of Neyman (1923) and Rubin (1974), as well as graphical models developed for probabilistic reasoning and causal analysis (Pearl, 1988; Lauritzen, 1996; Spirtes et al., 2000; Pearl, 2000a).

The second translation occurring from statistics to gender issue, is deepen in Chapter 2 .

The main contribution of statistics in exploring gender gaps has been, traditionally, in terms of "measurements", meant as indexes, rankings, ratios and the like. In this thesis, conversely, we move the statistical perspective from the static "measuring" to the dynamic "understanding" of cause-effect relationships; an index, for instance, catches a certain phenomenon and measures its changing over time, but it is not able to explain how such phenomenon has been originated and how it is changing after intervention.
It should be noticed that, gender -as a social structure- is not a *cause* itself of gender disparities, nor is it the cause of differences in access to resources; gender is, rather, an individual characteristic having physical and socio-cultural attributes. It is neither a *manipulative* variable, in the following sense. In epidemiological studies, if we were interested, for example, in whether smoking affects cancer in a population, we could perform a randomized experiment where every member of the population is randomly assigned either to the group subject to manipulation, namely to "smoking treatment", or to the control group where no manipulation is performed. We could, then, observe the consequences, in terms of effects on cancer. Conversely, direct manipulation is not possible for gender, since a random assignment of the "gender treatment" can not be implemented.

The causal approach to gender issues requires thus, first, a re-definition in a gender perspective in order to include gendered-concepts such as "gender equality", "gender inequality" and "gender gap". Then, we explore the potential of graphical models as a language able to untangle the complex relationship among variables selected for statistically assessing gender disparities. We focus on causal graphs as tools for both representing the causal mechanism as well as estimating gender gap. More in detail, extending the use of Pearl's intervention calculus (Pearl, 2000), we proposed a new measure of gender gap, in terms of the causal effect of gender on a target outcome selected for assessing a certain field of interest.

Empirical evidence of the usefulness and effectiveness of causal graph approach in gender studies is then given, in Chapter 3 , by an application to real data; the application focuses in exploring the existence of a potential gender gap in child nutrition and health in China, with particular attention to children and adolescents among 0-17 years.

Conclusions and future methodological developments, finally, close this thesis.

Chapter 1

Causal Graphs: the Theoretical Framework

Causal relationships represent the fundamental building blocks of our physical reality and of human understanding about such reality. Every day we deal with causes-effect considerations, whether we are deciding to go to work by car or by bus, or evaluating the effects of taking an aspirin for a headache or predicting who will win the elections.
The need of exploring cause-effect relationships among variables or events is common to many sciences as physical, behavioural, social as well as biological and the appropriate methodology for unrevealing such relationships from data has been object of several debates.

In the last decades, thanks to advances in computer science and developments in graphical models, causality has been transformed from a concept mainly belonging to philosophy into a mathematical object with well-defined language and logic. Causality has been "mathematized" (Pearl, 2000).

Answering "causal questions" requires extensions in the standard mathematical language of statistics. Indeed the aim of standard statistical analysis, typified by regression, estimation, and hypothesis testing techniques, is to assess parameters of a distribution from samples drawn of that distribution. Causal analysis goes one step further, inferring not only beliefs or probabilities under static conditions, but also their dynamics, in terms of cause-effect.

For instance, changes induced by treatments or external interventions.

1.1 Translating causality logic to statistical language

As a student of statistics, many times I met up the statement "Correlation do not imply Causation"; anyway I had never fully understood and deepened its meaning and the arising implications before writing this thesis.

Shipley (2002) effectively exemplifies the concept of such statement describing causal processes as hidden three-dimensional objects whose all it is possible to see are shadows, just two-dimensional projections of their actual aspect.
Statisticians often deal with shadows, as they cannot directly observe the actual causal mechanism and all they can peek are the consequences of these processes in the form of complicated patterns of associations and independence in the data. But as with shadows, these relationships are incomplete and potentially ambiguous as they represents only projections of the original causal processes.
Our aim is thus to find a mathematical tool able, from "correlation shadows", to uncover and effectively represent the data generating process as well to deduce the dynamics of the events in terms of causes and effects.

In order to study causal processes using statistics, it is first necessary to translate from the language of causality to the base of statistical language: the probability theory.
Such a rigorous translation device did not exist until recently. The earliest attempt to formulate causal relationship mathematically was made by Sewall Wright in the early 20th century (Wright, 1934) with the "method of path coefficients", but only in the last decades with the larger attention in statistics (Lauritzen, 1996; Whittaker, 2009) and computer science (Pearl, 1988; Kalisch *et al.*, 2012) such translation has been completely developed.

When translating between Italian and English, something may be lost: a slight change in inflection or context of a word can change the meaning

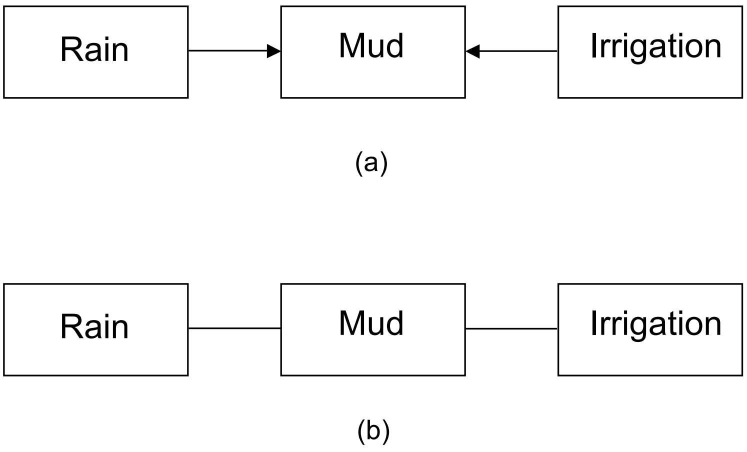


Figure 1.1: (a) Causal relationship between rain, mud and irrigation; (b) Observational relationship between rain, mud and irrigation.

in a disastrous way. For example the italian words “strani rumori” (strange noises) sound close to the english expression “strange rumors”; but if an italian boy said to an english friend: “I have heard strange rumors..” (meaning noises), the poor english guy would think about some strange gossip, with misunderstanding consequences! The same mistakes may occur in translating the language of causality into the language of probability distributions: there is a deep distinction between a causal model, an observational model and a statistical model; such differences will be now illustrated with a simple example.

The statement “rain causes mud” implies an asymmetric relationship that can be represented by the “→” symbol to refer to a causal relationship with the convention that, unless a causal relationship is explicitly indicated, it does not exist. Indeed in Figure 1.1(a) the missing arrow between “rain” and “irrigation” means that they have no causal relationship since the two are causally independent.

The observational model in Figure 1.1(b), related to the causal one in frame (a), represents the statement that “having observed rain will give us information about what we might observe concerning mud”. It deals with information, not causes and is not asymmetric, hence the linking symbol “—” is used.

Although rain and irrigation are causally independent, they are not observationally independent given the state of mud.

The statistical model differs from the observational one only in degree, not in kind. The statistical model could be expressed for instance by the mathematical relationship:

$$Mud(cm) = 0,1 * Rain(cm) + 0,2 * Irrigation(cm) \tag{1.1}$$

meaning that, to equal irrigated water, an increase in the quantity of fallen rain will result in an increase in cm of mud as there is a positive relation between them.

According to Pearl (1997) a great part of the actual confusion between correlation and causation is due to a mistranslation of the word "cause", a word having a connotation of asymmetry that cannot find a correct expression in the symbol "="; indeed the symbols "→" and "=" do not have an equivalent meaning as well as saying "mud does not cause rain" and "mud is independent of rain" . Equation 1.1 can correctly be rearranged to *predict* the amount of rain from the amount of mud recognising this as causally nonsensical:

$$Rain(cm) = 10 * Mud(cm) - 20 * Irrigation(cm) \tag{1.2}$$

In summary, the reasons for choosing tools as diagrams, able to encode causal relationships and represent the causal models, are that:

1. they are able of displaying relationships in an elementary and intuitively way;

2. they are "directional", thus being capable of representing cause-effect relationships;

3. they embed the theory of probability in order to handle with uncertainty;

4. they are visual tools for representing both direct and indirect causation.

1.2 The Mathematical Language of Graphs

In order to develop a translation device to move between causal models and observational (statistical) models, it is required the necessary and sufficient conditions needed to specify a joint probability distribution that must exist given a causal process. It is required the necessary and sufficient conditions to specify the "correlational shadow" that will be cast by a causal process. A feasible translation strategy would involve three points:

1. as the algebra cannot express causal relationships, we need a "new" mathematical language allowing it;

2. we need a tool that unambiguously will convert the statement expressed through directed graphs into statements involving conditional independences of random variables. This translation device will be called "d-separation", 1.2.2;

3. we need to define the assumptions and conditions connecting probabilities with causal graphs.

1.2.1 Notation and Terminology

Graphical models can be thought of as road maps: in order to use them one needs the physical map with symbol such as dots and lines (Kalisch *et al.*, 2012); secondly it is necessary a rule for interpreting the symbols. In causal graphs, the map consists of vertices and edges and the interpretation rule is called "d-separation". In this subsection we introduce some notational terminology of this causal map. In the next subsection we will deep the interpretation rules.

A *graph* $G = (V, E)$ consists of a set $V = \{1, ..., d\}$ of *vertices* (or nodes) and a set E of *edges* (or arcs) connecting some vertices pairwise (Lauritzen, 1996).

Graphical models have a "natural" causal semantics; they represent statistical models where vertices will correspond to random variables $X = (X_i | i \in V)$ and edges will denote the "relationship" between them. If two variables are

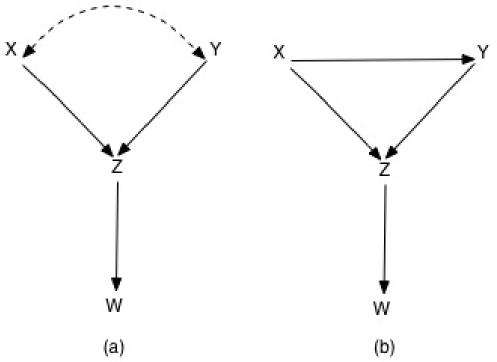


Figure 1.2: (a) A graph containing both directed and bidirected edges; (b) A directed acyclic graph (DAG) with the same skeleton of (a).

connected by an edge, are called *adjacent*. The adjacency set of a vertex X_i, is defined as the collection of all vertices that are adjacent to X_i in G and it is denoted by $adj_i(G)$. A *subgraph* of a graph G is a graph whose vertex set is a subset of that of G, and whose adjacency relation is a subset of that of G restricted to this subset.

In literature different classes of graphs can be distinguished for using different kinds of edges: *directed*, as marked by a single arrowhead on the edge, and *undirected* if unmarked links. In some applications we will also use *bidirected* edges, as marked with two arrowheads (see Figure 1.2 (a)) to denote the existence of unobserved common causes not showed in the graph, sometimes called *confounders*.

When all arcs are directed as in Figure 1.2 (b), we will have a *directed* graph; they may include directed *cycles*, (e.g. $X \to Y$, $Y \to X$) representing mutual causation or feedback processes, but not self-loops (e.g. $X \to X$). Conversely a graph not containing cycles is said *acyclic*.
Three basic classes of graphs can be found in the literature:

1. undirected graphs (UGs),
2. directed acyclic graphs (DAGs) or even Causal Bayesian Networks, a term coined by Pearl in 1985 to emphasize the subjective nature of the input information and the information' updating based on the reliance on Bayes's conditioning (Consonni and Leucari, 2001),

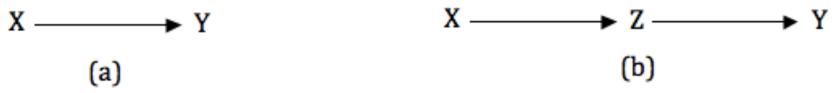


Figure 1.3: (a) X is a direct cause of Y; (b) X is an indirected cause of Y.

3. and chain graphs (Marchetti and Lupparelli, 2011) which are a generalization of the first two.

Our discussion mainly will involve DAGs also representing causal structure even called *Causal Graphs.*
We call a DAG a *causal graph* if, given a set of random variables V, for every pair $X, Y \in V$, we draw an edge from X to Y if and only if X is a direct *cause* of Y relative to V.
The notation $X \rightarrow Y$ means that X is a *direct cause* of Y, denoting that a causal relationship between the two vertices exists independently of any other vertex in the causal explanation, as depicted in figure 1.3(a); conversely an *indirected cause* is a causal relationship between two vertices that is conditional on the behaviour of other vertices, that is, if there is a sequence of directed arrows that can be followed from X to Y via one or more intermediate variables, as in figure 1.3(b).

A *skeleton* of a graph G is the resultant undirected graph obtained stripping away all arrowheads from the edges in G while a *path* is an unbroken sequence of edges, which may go either along - directed path - or against the arrows - undirected path -. For instance, in figure 1.2(a) the sequence $(X, Z), (Z, Y), (Y, X)$, moreover $(X, Z), (Z, W)$ defines a directed path. If no such directed path exists, then the two vertices are *causally independent.*
A non-endpoint vertex X_i in a path, as Z in Figure 1.2(a), is a *collider* if the path contains a pattern such as $\rightarrow X_i \leftarrow$. Three vertices $\langle X_i, X_j, X_k \rangle$ are called *unshield triple* if the couples of vertices (X_i, X_j) and (X_j, X_k) are adjacent but (X_i, X_k) are not adjacent. An unshield triple $\langle X_i, X_j, X_k \rangle$ is called a *v-structure* if X_j is a collider on the path $\langle X_i, X_j, X_k \rangle$, as for exam-

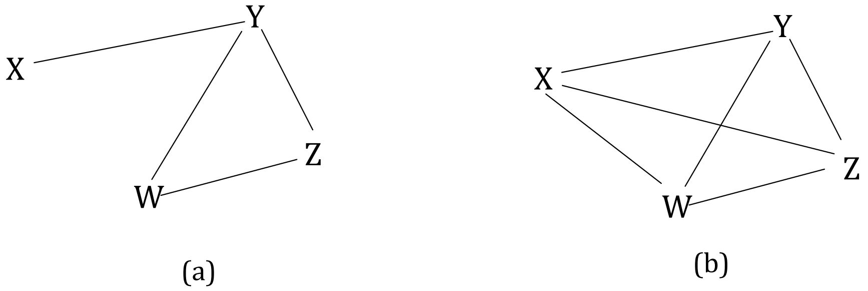


Figure 1.4: (a) A graph connected with 4 vertices and 4 edges; (b) A graph with 4 vertices and 4 edges complete.

ple the set (X, Z, Y) in figure 1.2(a).
Furthermore, a graph is *connected* if there is an undirected path between any two vertices as in 1.4(a) while it is said *complete* if every pair of its vertices are adjacent as in figure 1.4(b).

The relationships depicted in a graph make use of kinship terminology: as *parents, children, descendants, ancestors, spouses* etc. For example in Figure 1.2(a), both X and Y are parents of Z whilst Z is a descendant of X and Y. W has three ancestors, namely X, Z and Y as there are three paths respectively from X, Z and Y to W. X is a spouse of Y while, if they were connected by an undirected edge, X would be neighbour of Y. A *family* is a set of nodes, containing a node and all its parents. For instance, figure 1.2(a) for families are showed: $\{X\}$, $\{X, Z, Y\}$, $\{Y\}$ and $\{Z, W\}$).
Each child-parent family in a DAG represents a deterministic function:

$$x_i = f_i(pa_i, u_i) \quad i = 1, ..., n. \tag{1.3}$$

where pa_i denote the parents directly determining the value of the vertex X_i in G; u_i with $(1 \leq i \leq n)$ represents unobserved variables (including the errors ϵ_i) into a set U of background variables with distribution function $P(u)$. For

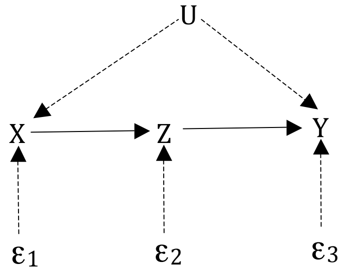


Figure 1.5: Path diagram corresponding to equations 1.4

example the set of equations:

$$\begin{aligned} x &= f_i(u, \epsilon_1), \\ z &= f_2(x, \epsilon_2), \\ y &= f_3(z, u, \epsilon_3), \end{aligned} \tag{1.4}$$

finds its graphical representation in figure 1.5, where X, Y, Z represent observed variables, f_1, f_2, f_3 are unknown arbitrary functions, and $U, \epsilon_1, \epsilon_2, \epsilon_3$ are unobservables that we can regard either as latent variables or as disturbances.

Notice that equation 1.3 is a non-linear, non-parametric generalization of the standard linear structural equation models (SEMs) (Goldberger, 1972; Wright, 1921)

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + \epsilon_i \quad i = 1, ..., n$$

with the only exception that the functional form of the equations, as well as the distribution of the disturbance terms, will remain unspecified.

1.2.2 Causal Markov Condition and d-separation

A causal model is defined *Markovian* if, when represented in a graph, it contains no directed cycles and if its ϵ_i's are mutually independent (no bi-directed arcs), while a model is said *Semi-Markovian* if its graph is acyclic and if it contains dependent errors (Lauritzen, 1996; Pearl, 2000).

Markovian models are equivalent to the SEM's literature *recursive models*

(Bollen, 1989). Indeed we can look at causal models in non-linear structural equation models as the equivalent of paths in linear ones: they differ from the latter in that their parents pa_i are defined, as shown in equation 1.3, as non-trivial argument of the function f_i, rather than variables obtaining non-zero coefficients. Moreover, bidirected arcs mean two-sided dependency instead of correlation (Pearl, 2000).

DAGs are a purely mathematical objects that may be interpreted both causally and probabilistically. Following the causal interpretation, a DAG G represents a causal structure such that the directed edge from X to Y means that X is a *direct cause* of Y; under the probability interpretation, a DAG G also referred to as a Bayesian Network (BN), represents a probability distribution P that satisfy the Markov Property, namely that each variable is independent of its non-descendant in the graph given the state of its parents. The intuitions, connecting causal graphs with the probability distributions generated, are generalized in two fundamental assumptions: Causal Markov Condition and Causal Faithfulness Condition. The latter will be described in the subsection 1.2.3. For what concern the first assumption, instead, let consider the following Theorem, known as:

Theorem 1.2.1 (**Causal Markov Condition (CMC)**). *Any distribution generated by a Markovian model can be factorized as:*

$$P(x_i, x_2, ..., x_n) = \prod_i P(x_i|pa_i) \tag{1.5}$$

where $X_1, X_2, ..., X_n$ are the endogenous variable in the model, and pa_i are the parents of X_i in the associated causal model.

Thus, given a set of variables whose causal structure can be represented by a DAG, the CMC, also known as Parental Screening (Pearl, 1988; Whittaker, 2009), represents one of the bridge principles linking the causal interpretation of a DAG to its probabilistic interpretation. Reformulating the CMC in terms of the causal DAG, it affirms that given a set of variables whose relationships structure can be represented by a DAG, every variable is probabilistically independent of its non-descendants conditional on its parents.

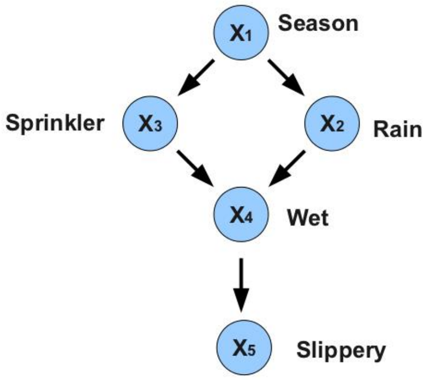


Figure 1.6: Graph illustrating causal relations among five variables.

For example, the figure 1.6 describes relationships among the seasons of the year (X_1), whether rain falls (X_2), whether a sprinkler is turned on (X_3), whether the pavement is wet (X_4) and whether it is slippery (X_5).
Thanks to the decomposition:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4)$$

if we would want, for instance, to compute the probability of a slippery pavement we would only need to know whether it is wet or not. In other words, it does not matter whether it rains or not if we already know that it is wet. Note that we have not yet assigned any causal meaning to the graph. This theorem supports the intuition that once that direct causes of X_i are known and controlled, the probability on X_i is in fact determined.

The conditional independences pattern for a given graph G hold by CMC, may not be obvious to identify. They can be read off using a purely graphical criterion proposed by Pearl (1988) that represents the translation device between the language of causality and the language of probability distributions: the *d-separation criterion*, where the d denotes *directional*.
D-separation gives the necessary and sufficient conditions for two vertices in a causal DAG to be observationally, probabilistically, independent upon conditioning on any other set of vertices. In other words, d-separation captures the conditional independence constraints entailed by the Markov property in

a DAG. For instance, in three disjoint sets of variables X, Y, Z, in order to verify if X is independent by Y given Z, we need to test if the nodes in Z "block" all paths from nodes in X to nodes in Y, in the sense that Z's nodes stop the flow of information from X to Y. To do this, we must check the direction of the arrowheads.
More formally:

Definition 1.2.1 (**The d-Separation Criterion**). *A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if:*

1. *p contains a **chain** $i \rightarrow m \rightarrow j$ or a **fork** $i \leftarrow m \rightarrow j$, such that the middle node m is in Z, or;*

2. *p contains a **collider** $i \rightarrow m \leftarrow j$, such that the middle node m is NOT included in Z and, at the same time, no descendant of m is included in Z.*

The Causal Markov Condition (CMC) can therefore be rephrased as to saying that for any three disjoint subsets of variables A, B and C, if A and B are d-separated by C in the causal DAG, then A and B are independent conditional on C.
In the example illustrated by figure 1.6, we can see that the set $\{X_2, X_1, X_3\}$ represents a fork such that X_2 and X_3 are marginally dependent; however they become independent (blocked) once we condition on the middle variable X_1. Notice that, X_2 and X_3 are not d-separated by $Z = \{X_4, X_5\}$ meaning that learning whether it is slippery outside somehow makes X_2 and X_3 dependent. Intuitively if we know that it is slippery outside and it is not raining, then the sprinkler must be turned on, which seems reasonable.
The d-rules fit with the intuition that two variables will be correlated if one causes the other or if there is an uncontrolled common prior cause of both variables. The rules also reflect the non-intuitive fact that a statistical association between two variables can be induced by conditioning on a common effect of both variables (Greenland *et al.*, 1999; Hernán *et al.*, 2004). Indeed if a collider on a path is in the covariate set, this collider does not block the path.

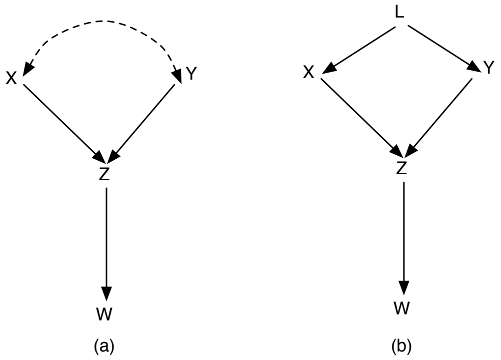


Figure 1.7: (a) A graph containing a bidirected edge; (b) Bidirected arc has been interpreted as a latent cause L affecting both X and Y.

The d-separation criterion is valid even in semi-Markovian models, when bidirected arcs are interpreted as emanating from latent common parents. In figure 1.7, for example, the bidirected arc (a) has been replaced in (b) with a fork such that the middle variable L is a latent common cause affecting both X and Y. This is also possible in linear semi-Markovian models where each latent variable is restricted to influence at most two observed variables (Spirtes, 1996).

Coming back on the analogy of a "correlational shadows" of the underlying causal process, the d-separation is the method by which one can predict these shadows. Although causal models and observational models are not the same thing, there is indeed a one to one correspondence between the set of conditional independences implied by the recursive decomposition 1.5, and the set of triples (X, Z, Y) that satisfy the d-separation criterion in the graph. Such connection is illustrated in the following Theorem (Geiger, Verma and Pearl 1990):

Theorem 1.2.2 (Probabilistic Implications of d-Separation). *If the sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every probability distribution such DAG G can represent. Conversely, if X and Y are not graphically d-separated by Z, then X and Y are dependent conditional on Z in at least one distribution compatible with G.*

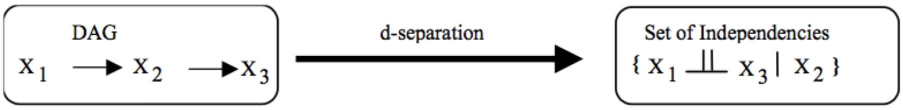


Figure 1.8: From d-separation relationships in a DAG to statistical independence relationships in the data.

In practice, as showed in figure 1.8, it means that once we have specified the acyclic causal graph, if it actually represents the unknown causal process, then every d-separation relationship existing in our graph should be mirrored in an equivalent statistical independence in the observational data. The converse part of the theorem says that the absence of d-separation, implies that there exists some distribution factorizing over the graph in which X and Y are dependent given Z.
Notice that the previous statement is very general as it does not depend on any distribution assumptions of the random variables or on the functional form of causal relationships.

1.2.3 Causal Faithfulness Condition

In this section we illustrate the second fundamental axiom that connects probability with causal graphs as anticipated in the previous subsection (Spirtes *et al.*, 2000).

Given a causal graph G and the Causal Markov assumption, we assume that any distribution compatible with G has the independence relations obtained by applying d-separation to it. However, this does not imply that the distribution has exactly these and no additional independences. Consider the following example (Scheines, 1997) and suppose that figure 1.9 is a graph describing the actual causal relationships between smoking, exercise and health. The + and - signs denote positive or negative effects. Applying d-separation to the graph, it would show no independences. However it might occur that in some probability distributions that this graph produces, smoking is independent of health “by chance”.

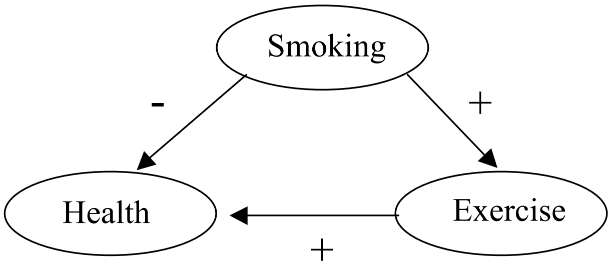


Figure 1.9: The “true ”graph describing the causal relationships between smoking, exercise and health where + and - signs denote positive or negative effects.

In figure 1.9, smoking has a negative direct effect on health but also a positive indirect effect on it, even though it might be absurd that smoking has positive effect on exercise. It could happen that we might have no association at all between smoking and health if the two effects would occur to cancel each other out. In this case we would say that the probability distribution is *unfaithful* to the causal graph that generated it, as there are any independence relations in the population that are not a consequence of the Causal Markov Condition (or d-separation).
In order to guarantee that such positive and negative effects never perfectly balance and thus cancel one another, we assume *faithfulness*, namely that whatever independences occurring in a population arise not from incredible coincidence but rather from structure.

Definition 1.2.2 (**Causal Faithfulness Condition (CFC)**)**.** *A probability distribution P is said to be faithful with respect to a graph G if conditional independences of the distribution are exactly the same as those encoded by G via d-separation, or equivalently, by the CMC.*

More precisely: consider a random vector X with probability distribution P (as denoted by $\sim P$); the CFC of P with respect to G means that for any i, $j \in V$ with $i \neq j$ and any set $\mathbf{s} \subseteq \mathrm{V}$, X_i and X_j are conditionally independent given $\{ X_r ; \mathrm{r} \in \mathbf{s}\} \iff$ node i and node j are d-separated by the set $\mathbf{s}$ (Kalisch and Bühlmann, 2008).

The CMC and CFC together set up a perfect correspondence between conditional independence constraints and d-separation features of the causal DAG. In other words, we can say that an oracle of conditional independence constraints over the given set of observed variables, translates into an oracle of d-separation features in the causal DAG (Zhang, 2006).

1.2.4 Markov Equivalence

After proposing a causal model and finding that the observational data do not contradict any of the d-separation statement of our causal model, it is possible to determine more causal models consistent with the same data. Indeed, under the shadow metaphor, it could happen that more than one three-dimensional object has the same contours behind a shadow, and consequently, we cannot detect the "true" object.

It is still true under the CMC and CFC assumptions that correlation does not imply causation. In general, the "true" causal graph is under-determined by a pattern of correlations, while, there must be multiple causal graphs that, given a pattern of conditional independence constraints, satisfy the CMC and CFC. In this case the two DAGs are said *observationally equivalent* or *Markov equivalent* (Andersson *et al.*, 1997). The possibility to have a whole class of equivalent models logically arises by the assumption that causal relations cannot be inferred by statistical data only and their distribution as Wright (1921) stated: "prior knowledge of the causal relations is assumed as prerequisite".

A criterion to determine when if observational equivalence occurs, follows directly from the d-separation criterion:

Theorem 1.2.3 (Observational Equivalence). *Two DAGs are observationally equivalent if and only if they have the same skeleton and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow (Verma and Pearl, 1990).*

For example in Figure 1.7 (b) if we reversed the arrow's direction from L to Y, we would obtain the same skeleton and v-structure, by gaining an equivalent DAG. Thus the directionality of the link $L \rightarrow Y$ cannot be determined on the basis of probabilistic information only. Conversely, the arrows $Y \rightarrow Z$ and $Z \rightarrow W$, if reversed, would create different v-structures and therefore not an equivalent DAG.

Notice that in standard SEM, models are observationally indistinguishable if they are *covariance equivalent*, that is, if every covariance matrix generated by one model can be generated also by another model (Bollen, 1989). It can be verified that Teorema 1.2.3 extends to covariance equivalence (Pearl and Verma, 1995).

Also notice the methodological importance of Theorem 1.2.3: it asserts that we are never testing *a* model, instead a whole *class* of observationally equivalent models are tested, providing a clear representation of competing alternatives for considerations. As a consequence, the space of DAGs can be partitioned into equivalent classes, where all members of an equivalent class encode the same conditional independence information.

A common tool for visualizing equivalence classes of DAGs is a representation through a *complete partially directed acyclic graph* (CPDAG) (Chickering, 2002). This is a graph with the same skeleton as the graphs in the equivalence class in which:

- the directed edges represent arrows that are common to all DAGs in the equivalence class, and;
- the undirected edges correspond to edges that are directed one way in some DAGs and the other way in another DAGs included in the equivalence class, as exemplified in Figure 1.10.

By using CPDAG the problems of having multiple representations of the same equivalent class is eliminated.

In semi-Markovian models the rules for generating equivalent models are more complicated. The basic principle is that if we interpret any bidirected arc $X <-- > Y$ as representing a common latent cause L, affecting both X and Y, that is $X \leftarrow L \rightarrow Y$, then the "if " part of Theorem 1.2.3 holds valid, allowing for any edge-replacement which do not either destroy or create new v-structures.

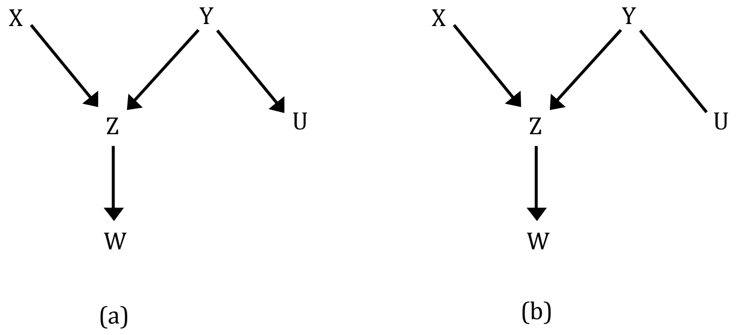


Figure 1.10: (a) a DAG G and (b) the CPDAG representation for *Class*(G)

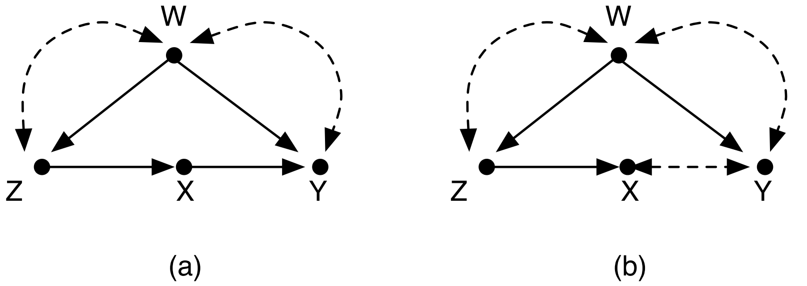


Figure 1.11: (a) Model with $X \longrightarrow Y$; (b) Model replaced with $X <-- > Y$

Generalizing, an edge $X \rightarrow Y$ can be replaced by either a bidirected arc $X <-- > Y$ or a directed arc $X \leftarrow Y$, according to the following rules:

Rule 1: An arrow $X \rightarrow Y$ can be replaced by $X <-- > Y$ only if every neighbour or parent of X cannot be d-separated from Y;

Rule 2: An arrow $X \rightarrow Y$ can be reversed in $X \leftarrow Y$ if, before reversal:
(i) every neighbour or parent of Y (excluding X) is inseparable from X and;
(ii) every neighbour or parent of X is inseparable from Y.

where for neighbour we mean a node connected through a bidirected arc. In the example depicted in Figure 1.11, we can reply $X \rightarrow Y$ with $X <-- > Y$: indeed Z, the X's parent, is inseparable from Y because even if we block the path through X, the path through W cannot be blocked.

1.3 Causal Structure Learning

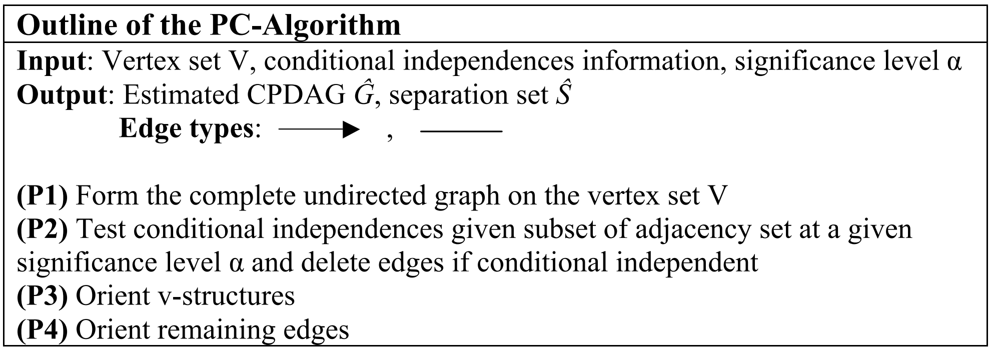
Causal structure learning represents the first of the two main functions of causal graphs.
We need a proper inference procedure for inferring features of the unknown causal structure from probabilistic independence and dependence relationships encoded in observational data. In literature, there are sound and complete algorithms for extracting causal information out of an oracle of probabilistic independence, With "sound", it means that anytime the algorithm returns an answer, that answer is correct;"complete" is intended in the sense that any feature of the causal structure left undecided by the inference procedure is indeed under-determined by facts of probabilistic independence and dependence. The output of such algorithms is a graphical object representing an equivalence class of causal structures, displaying all and only those common features shared by all causal structures that satisfy the Causal Markov Condition (CMC) and Causal Faithfulness Condition (CFC) with the oracle.

There are, in general, three different methods used in learning the structure of a causal graph from data: 1) via constraint-based, 2) score-based and 3) hybrid algorithms.
The constrain-based refers to conditional independence statements in the data and uses these conditional independences to reconstruct the structure (Spirtes *et al.*, 2000). The score-based method defines a search on the space of all causal graphs by using a goodness-of-fit score defined by the implementer, which says how good a graph is compared to the others. Hybrid algorithms combine aspects of both constraint-based and score-based algorithm methods, as they use conditional independence tests usually to reduce the search space, and at the same time goodness of fit scores to find the optimal graph in the reduced space. The methods will be detailed in the following subsections.

1.3.1 Constraint-based Methods: the PC and FCI algorithms

Constrain-based algorithms are all based on the inductive causation (IC) algorithm by Verma and Pearl (1990). They perform conditional independence statistical tests on the data set since conditional independence statements can reduce the variables under consideration and greatly aid in the task of untangling and understanding the interactions among the selected variables (Zhang, 2006).
Several conditional independence tests from information theory and traditional statistics are available for use in constraint-based learning algorithms. After selection of the conditional independence test to apply on the data set, some assumptions are required, such as Causal Markov, Faithfulness defined, respectively, in sections 1.2.2 and 1.2.3. In addition, we assume *Causal Sufficiency*, namely given a set of variables $\mathbf{V}$, $\mathbf{V}$ is said "causal sufficient" if for every pair of variables V_i,$V_j \in \mathbf{V}$, every common direct cause of V_i and V_j relative to $\mathbf{V}$ is also a member of $\mathbf{V}$. This means that there are both no unmeasured common causes and no unmeasured selection variables.

Several constrain-based algorithms have been proposed, among the most known, we cite the Spirtes, Glymour and Scheines's algorithm (SGS) (Spirtes 1993/2000), the Inferred Causation (IC) (Pearl and Verma, 1993), the Grow Shrink (GS) (Margaritis, 2003), Incremental Association (IAMB) (Tsamardinos et al, 2003).
Since DAGs encode conditional independences, information on the latter helps to infer aspects of the former. This concept is the basis of the PC algorithm, the most straightforward algorithm - where PC stands for the initials of its inventors Peter Scheines and Clark Glymour (Spirtes *et al.*, 2000). PC algorithm is able to reconstruct the (unknown) causal structure of the underlying DAG model given a conditional independence oracle up to its Markov equivalence class (Kalisch *et al.*, 2012).
According to the constraint-based approach, the PC algorithm is clearly divided into two parts, namely statistical inference from data and causal inference from probability. As outlined in table 1.1, it starts from a complete,



Outline of the PC-Algorithm

Input: Vertex set V, conditional independences information, significance level α

Output: Estimated CPDAG $\hat{G}$, separation set $\hat{S}$

Edge types: ⟶ , ―――

(P1) Form the complete undirected graph on the vertex set V
(P2) Test conditional independences given subset of adjacency set at a given significance level α and delete edges if conditional independent
(P3) Orient v-structures
(P4) Orient remaining edges

Table 1.1: Main steps of the PC Algorithm

undirected graph and recursively deletes edges according to a conditional independence rule. As mentioned before, the same list of conditional independences can be modelled by different DAGs; it has been shown that two DAGs represent the same conditional independences statements if and only if they have the same skeleton and the same v-structures, that is, they are Markov equivalent (Verma and Pearl, 1990). On the contrary, given a conditional independence oracle, one can only determine a DAG up to its equivalent class. Therefore the PC algorithm cannot uniquely determine the DAG so that the output of the algorithm will actually be the equivalent class (CPDAG) that describes the conditional independence information in the data.

The PC algorithm is sound and complete (i.e. maximally informative) under the assumptions of causal sufficiency and faithfulness (Spirtes *et al.*, 2000; Zhang, 2006). Moreover it is computationally feasible and consistent, even with high-dimensional sparse DAGs (Kalisch and Bühlmann, 2007) and it is efficiently implemented in the R-package pcalg (Kalisch *et al.*, 2012).

The main steps of PC Algorithm are summarized in Table 1.1; we now describe it in more detail since it will be mentioned in the application in Chapter 3.
The PC Algorithm starts with a complete undirected graph, G_0 (step P1 in table 1.1.
Subsequently in stage (P2) a series of conditional independences tests is per-

formed and edges are deleted following a two-phases rule:
1) all pairs of nodes are tested for marginal independences; if two nodes, say i and j, are tested marginally independent at level α, the edge between them is deleted and the empty set is saved as separation sets $\widehat{S}[i,j]$ and $\widehat{S}[j,i]$, as defined in 1.2.1. After all pairs have been tested for marginal independence and some edges might have been removed, phase 1) ends producing a graph results, denoted G_1.
2) all pairs of still adjacent nodes (i,j) in G_1 are tested for conditional independence given any single node in $\text{adj}(G_1,i) \setminus \{j\}$ or $\text{adj}(G_1,j) \setminus \{i\}$, where $\text{adj}(G,i)$ denotes the set of nodes in graph G adjacent to node i. If there is any node k such that V_i and V_j are conditionally independent given V_k, the edge between i and j is removed and node k is saved as separation sets $\widehat{S}[i,j]$ and $\widehat{S}[j,i]$. When all adjacent pairs have been tested given one adjacent node, a new graph G_2 results.
The PC algorithm proceeds increasing step by step the size of the conditioning set until all adjacency sets in the current graph are smaller then the the size of the conditioning set (Kalisch *et al.*, 2012). The resulting graph is a skeleton in which every edge is still undirected.
In step (P3) each triple of vertices (i,j,k) are considered, where the pairs (i,k) and (j,k) are each adjacent in the skeleton but (i,j) not. Each triple is oriented according to information saved in the conditioning sets $\widehat{S}[i,j]$ and $\widehat{S}[j,i]$ (Spirtes *et al.*, 2000); for instance, the triple $i - j - k$ is oriented as $i \rightarrow k \leftarrow j$ if k is not included in $\widehat{S}[i,j]$ nor in $\widehat{S}[j,i]$.
Finally in (P4) some of the remaining edges are tried to be oriented following two basic principles: not to create cycles and not to create new v-structures. The resulting output is the equivalence class (CPDAG) that describes the conditional independence information in the data, in which every edge is either undirected or directed. Notice that in order to improve the visual representation of the output, undirected edges are depicted as bidirected edges as long as at least one directed edge is present. Every DAG in this equivalence class can represent the true causal structure.
A typical output of the PC algorithm is shown in Figure 1.12: it is a graph containing relationships pattern both directed edges and undirected edges.

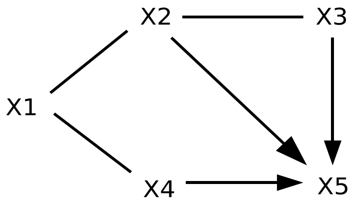


Figure 1.12: An example of PC algorithm output

Although the "true" causal graph is not fully known, this output summarizes valuable causal information, for example, that X_2,X_3,X_4 are drawn as *direct* causes of X_5 (Zhang, 2006).

It is important to notice that the assumption of causal sufficiency may not be satisfied.
It represents an open problem that has been discussed for over ten years (Spirtes et al. 1993/2000). Indeed in practice, it can happen that variables of interest suffer from confounding due to latent common causes or even that a unit is sampled in virtue of the value of certain variable(s), called selection variable(s), that are causally influenced by some other variables in the system. The point is that in these situations any probabilistic relationship inferrable from data is conditional upon (certain values of) the latent or the selection variable(s). In such cases, the set of observed variables may be *causally insufficient* so that DAGs do not provide a feasible representation. This is because CMC typically fails in the sense of not entailing the actual conditional independence constraints and hence not being causally *accurate*. The main problems concerning latent and selection variables are: 1) causal inference based on the PC algorithm may be incorrect; 2) the space of DAGs is not closed under marginalization and conditioning (Richardson and Spirtes, 2002). To illustrate problem 1) consider Figure 1.13(a) representing a DAG with observed variables $\mathbf{X}$=$\{X_1, X_2, X_3\}$ and latent variables $\mathbf{L}$=$\{L_1, L_2\}$ (Colombo *et al.*, 2012). In the system of the observed variables, the only conditional independence is the one between X_1 and X, even $X_1 \perp X_3$. The only DAG on X implying this single conditional independence relationship is $X_1 \rightarrow X_2 \leftarrow X_3$, and this will therefore be the output of the PC algorithm as

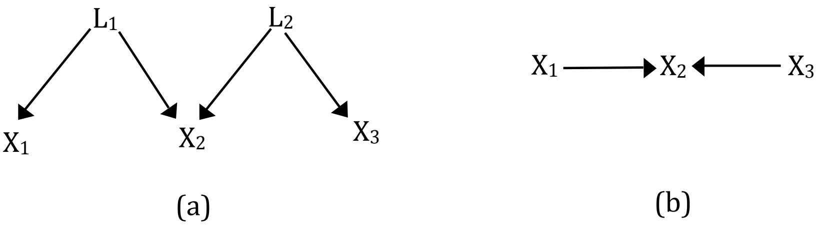


Figure 1.13: (a) DAG with latent variables; (b) CPDAG

showed in Figure 1.13(b). Anyway such output would lead us to incorrectly believe that both X_1 and X_3 are *causes* of X_2, while the underlying DAG including latent variables, shows neither a directed edge $X_1 \rightarrow X_2$ nor one $X_2 \leftarrow X_3$.

Regarding problem 2), consider, for example, the DAG $X_1 \rightarrow X_2 \leftarrow L_1 \rightarrow X_3 \leftarrow X_4$. This DAG implies the following set of conditional independences among the observed variables:

$$
\begin{aligned}
&X = \{X_1, ..., X_4\} : \\
&X_1 \perp X_3, X_1 \perp X_4, X_2 \perp X_4, X_1 \perp X_3 | X_4, X_1 \perp X_4 | X_2, X_1 \perp X_4 | X_3 \\
&and \quad X_2 \perp X_4 | X_1
\end{aligned}
$$

and others implied by these. No DAG on X would entail exactly this very set of conditional independences via d-separation. This means that in the related models, when some variables are ignored via marginalization or when other variables are fixed by selecting and fixing their value via conditioning, there is no graph of the same class capturing the modified independence statements. In this sense the space of DAGs is not closed under marginalization and conditioning, since the distribution obtained by either marginalizing or conditioning on some of the variables may not be faithful to any DAG on the observed variables.

We would need an alternative representation to depict the presence of latent common causes and selection variables in the causal process that generates the data. We also would need a proper generalization of DAGs called *maxi-*

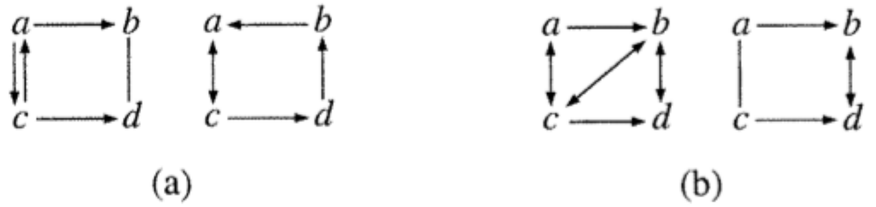


Figure 1.14: (a) Mixed graphs that are not ancestral; (b) Ancestral mixed graphs.

mal ancestral graphs (MAGs) (Richardson and Spirtes, 2002).

A MAG is a *mixed* graph in which every missing edge corresponds to a conditional independence relationship. Besides directed edges (→), it may also contain bi-directed edges (↔), associated with the presence of latent common causes, and undirected edges (-), associated with the presence of selection variables. In short, arrowheads in a MAG are interpreted as "non-cause", and tails are interpreted as "cause", of either an observed variable or a selection variable. A MAG is *ancestral* as it does not contain any directed or almost directed cycle and there is no edge into any vertex in the undirected component of an ancestral graph; examples of ancestral and non ancestral graphs are shown in figure 1.14. It is even *maximal* as every missing edge corresponds to at least one independence in the corresponding independence model (Richardson and Spirtes, 2002). Notice that maximality corresponds also to the property known as the pairwise Markov property.

Furthermore, syntactically, DAGs and UGs are special cases of MAG. In fact, every DAG with latent and selection variables can be transformed into a unique MAG over the observed variables. Several DAGs can indeed lead to the same MAG, whilst a MAG describes infinitely many DAGs since no restrictions are made on the number of latent and selection variables.

A distinctive property of MAGs is that they can represent such in-principle-testable constraints without explicitly introducing latent and selection variables. Given any DAG G over V=O∪L∪S, where O denotes a set of observed variables, L denotes a set of latent or unobserved variables, and S denotes a set of unobserved selection variables to be conditioned upon, then there

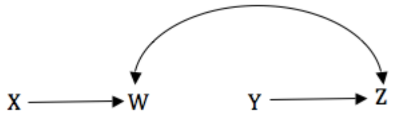


Figure 1.15: An ancestral graph and m-separation

exists a MAG over O alone.
According to Spirtes and Richardson (2002), MAGs encode conditional independence relationships among the observed variables via m-separation, an extension of the original definition of d-separation for DAGs, and defined as:

Definition 1.3.1 (**m-separation**). *In an ancestral graph G, a path π between vertices X and Y is said m-connecting given a set Z (possible empty) not included in X, Y, if:*

- *every non-collider on π is not a member of Z, or*
- *every collider on π is an ancestor of some member of Z*

For example, in the ancestral graph in figure 1.15, X and Y are m-separated given Z.
Definition 1.3.1 is thus, for MAGs, the equivalent of d-separation for DAGs in that the notions of "collider" and "non-collider" now allow for bi-directed and undirected edges. Furthermore, m-separation allows to detect the maximality in an alternative way: an ancestral graph is indeed said to be maximal if for any two non-adjacent vertices, there is a set of vertices that m-separates them.

Recalling that finding a unique DAG from an independence oracle is in general impossible, we can only detect the equivalence class of DAGs in which the true DAG must lie and visualize the equivalence class using a CPDAG. The same is true for MAGs: finding a unique MAG from an independence oracle is mostly not practicable. One may only report on the equivalence class in which the true MAG lies and represent it by a *partial ancestral graph* (PAG) (Zhang, 2008b). A PAG is an ancestral graph containing the following types of edges: ∘-∘, ∘-, ∘ →, →, ↔, -, and edges with the following interpretation:

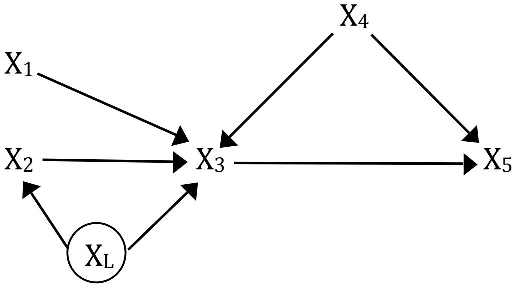


Figure 1.16: DAG with a latent variable X_L

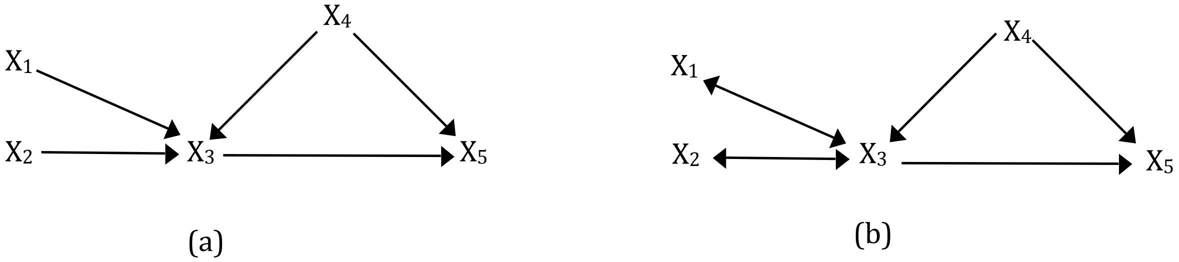


Figure 1.17: Two Markov equivalent MAGs

1. there is an edge between X and Y if and only if V_X and V_Y are conditionally dependent given V_S for all sets V_S consisting of all selection variables and subsets of the observed variables;

2. a tail on an edge implies that this tail is present in all MAGs in the equivalence class;

3. an arrowhead on an edge means that this arrowhead is present in all MAGs in the equivalence class;

4. a o- edgemark means that there is at least one MAG in the equivalence class where the edgemark is a tail, and at least one where the edgemark is an arrowhead.

Note that in the case where no selection effect is present (i.e., S = ∅), the causal MAG will not contain any undirected edges.

For example, suppose that Figure 1.16 shows the true, usually unknown,

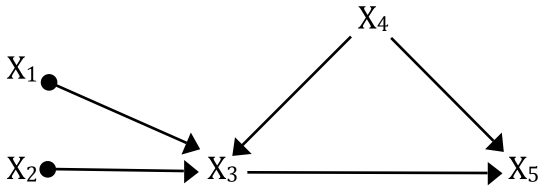


Figure 1.18: The PAG in our example

causal structure among variables, where X_L represents a latent common cause between X_2 and X_3. The causal MAGs corresponding to the causal DAG are depicted in Figure 1.17(a). This MAG might represent some other DAGs as well, for example, can also represent the DAG with an extra latent common cause of X_1 and X_3. The two MAGs would be Markov equivalent.
This motivates the representation of equivalence classes of MAGs through a PAG. For instance, the PAG for our example is drawn in Figure 1.18, which displays all the commonalities among MAGs that are Markov equivalent to the MAGs in Figure 1.17, both panels (a) and (b).

Given the exact correspondence between d-separation relations among the observed variables in the causal DAG and m-separation relations in the causal MAG, the causal Markov condition (CMC) 1.2.2, and its converse, the causal Faithfulness condition (CFC) 1.2.3, imply that conditional independence among the observed variables, correspond to m-separation in the causal MAG, which forms the basis of constraint-based learning algorithms (Zhang, 2008a).
A representative constraint-based causal discovery algorithm for causally insufficient systems is known as the fast causal inference (FCI) algorithm (Spirtes *et al.*, 2000, 1995) and its output can be interpreted as a PAG (Zhang, 2008a). The orientation rules of this algorithm were extended and proven to be complete in Zhang (2008b). The basic idea of the FCI algorithm is similar to the PC algorithm described in the previous subsession with the two stages, the adjacency stage and the orientation stage. However, it makes additional conditional independence tests and uses more orientation rules.

It represents a generalization of the PC algorithm allowing arbitrarily many latent and selection variables.

The rule to determine the adjacencies in a PAG within the FCI algorithm is: if X_i is not an ancestor of X_j, and X_i and X_j are conditionally independent given some set $\mathbf{Y} \cup \mathbf{S}$ where $\mathbf{Y} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$, then X_i and X_j are conditionally independent given $\mathbf{Y} \cup \mathbf{S}$ for some subset Y of a given set D-SEP(X_i, X_j) or of D-SEP(X_j, X_i) (see Spirtes *et al.* (2000) pag.134 for a definition). Therefore, in order to determine whether there is an edge between X_i and X_j in an FCI-PAG, it is necessary to test whether $X_i \perp X_j \mid (Y \cup S)$ restricted for all possible subsets $Y \subseteq$ D-SEP(X_i, X_j) and $Y \subseteq$ D-SEP(X_j, X_i). Anyway the sets D-SEP(X_i, X_j) cannot be inferred from the observed conditional independences, so Spirtes et al. defined a sort of superset, called *Possible D-SEP*, as follow:

Definition 1.3.2 (Possible D-SEP). *Let C be a graph with any of the following edge types: $\circ\!\!-\!\!\circ$, $\circ\!\rightarrow$, $\leftrightarrow$. Possible D-SEP(X_i, X_j) in C, denoted in shorthand by $pds(C, X_i, X_j)$, is defined as follows: $X_k \in pds(C, X_i, X_j)$ if and only if there is a path π between X_i and X_k in C such that for every subpath $\langle X_m, X_l, X_h \rangle$ of π, X_l is a collider on the subpath in C or $\langle X_m, X_l, X_h \rangle$ is a triangle in C.*

The definition of Possible D-SEP requires some information about both the skeleton and orientation of edges; therefore, Step 1 of the FCI algorithm finds an initial skeleton, C_1, as in the PC-algorithm, by starting from a complete graph with edges $\circ\!\!-\!\!\circ$ and performing conditional independence tests given subsets of increasing size of the adjacency sets of the vertices. An edge between X_i and X_j is deleted if a conditional independence is found, and the set responsible for this conditional independence is saved in the separation set, sepset(X_i, X_j) and sepset(X_j, X_i) . After completing Step 1, the skeleton is a superset of the final skeleton (Colombo *et al.*, 2012).

In Step 2 of the FCI algorithm, unshielded triples $X_i \ast\!\!-\!\!\circ\ X_j\ \circ\!\!-\!\!\ast\ X_k$ are orientated as v-structures $X_i \ast \rightarrow X_j \leftarrow \ast\ X_k$ if and only if X_j is not in sepset(X_i, X_k) and sepset(X_k, X_i).

The graph C_2 resulting from Step 2, contains sufficient knowledge to com-

pute the Possible D-SEP sets. Thus, in Step 3, the FCI algorithm computes $\text{pds}(C_2, X_i,)$ for every $X_i \in X$. Then for every element X_j in $\text{adj}(C_2, X_i)$,the algorithm tests whether $X_i \perp X_j \mid (Y \cup S)$ for every subset Y of $\text{pds}(C_2, X_i,)$ $\setminus \{X_i, X_j\}$ and of $\text{pds}(C_2, X_j,) \setminus \{X_j, X_i\}$. The tests are arranged in a hierarchically beginning with conditioning sets of small size. The edge between X_i and X_j is removed if there exists a set Y making X_i and X_j conditionally independent given $Y \cup S$, and the set Y is saved as the separation set in $\text{sepset}(X_i, X_j)$ and $\text{sepset}(X_j, X_i)$ (Colombo *et al.*, 2012).
In Step 4, the v-structures are thus oriented again based on the updated skeleton and the updated information in sepset. Finally, in Step 5 the algorithm replaces as many $\circ$-edges as possible by arrowheads and tails by using the orientation rules described by Zhang (2008b).

1.3.2 Score-based Method: the hill-climbing algorithm

Score-based algorithms assign a score to each candidate graph trying to maximize it with some heuristic search algorithm such as *hill-climbing*, *tabu search*, *simulated annealing*, *Tree Augmented Naive Bayes* as well as other genetic algorithms. According to this approach, learning a causal structure is considered an optimization problem where a quality measure (the score) of a causal structure, given the training data, needs to be maximized. In particular, assuming a causal structure G and a data set D, the score is given by the posterior probability of G given data:

$$Score(G, D) = Pr(G|D) = \frac{Pr(D|G)Pr(G)}{Pr(D)} \tag{1.6}$$

A score-based algorithm attempts to maximize this score returning the structure G that maximizes it. It constitutes a computational problem the fact that the space of all possible structures is at least exponential in the number of variables p; indeed, there are $p(p-1)/2$ possible undirected edges and $2^{(p-1)/2}$ possible structures for every subset of these edges, not mentioning that there may be more than one orientation of the edges for each choice. Thus heuristic search algorithms, which suggest some approximation, are employed to solve the optimization problem. The quality measure, i.e. the score, can be based on several criteria; for instance, on a Bayesian approach, minimum description length and information criteria. Most of scoring criteria derived in the literature are decomposable. It means that those metrics have the practical property that the score of the whole graph can be decomposed as either sum or products of the score of individual nodes. This allows for local scoring and thus local searching methods. Popular score functions are:

- The *likelihood* and the *log-likelihood* scores;
- The Akaike (AIC) and Bayesian (BIC) information criterion scores, as defined as:

  $$AIC = LogL(X_1, ..., X_v) - d \quad BIC = LogL(X_1, ..., X_v) - \frac{d}{2} log(n) \tag{1.7}$$

  Note that BIC is equivalent to the minimum description length (MDL) described by Rissanen (1978) and used as a Bayesian network score in Lam and Bacchus (1994).
- The logarithm of the *Bayesian Dirichlet equivalent* score, a score equivalent Dirichlet posterior density (Heckerman et al. 1995);
- The logarithm of the K2 score, which is in fact another Dirichlet posterior density (Cooper and Herskovits 1992) defined as:

  $$K2(X_i) = \prod_{j=1}^{L_i} \frac{(R_i - 1)!}{(\sum_{k=1}^{R_i} n_{ijk} + R_i - 1)!} \prod_{k=1}^{R_i} n_{ijk}! \tag{1.8}$$

  where R_i denotes the number of states of X_i, while L_i the number of possible configurations of the parent set $Pa_G(X_i)$ of X_i .
- The score equivalent Gaussian posterior density for continuous variables, which follows a Wishart distribution (Geiger and Heckerman 1994).

A popular score-based algorithm for causal structure learning is the *hill-climbing* greedy search. It is based on the principle of taking the (local) best

```
Procedure $B = BIChillclimb(D)$

1.  $E \leftarrow \emptyset$
2.  $T \leftarrow ProbabilityTables(E, D)$
3.  $B \leftarrow \langle U, E, T \rangle$
4.  $score \leftarrow -\infty$
5.  do:
    (a)  $maxscore \leftarrow score$
    (b)  for each attribute pair $(X, Y)$ do
    (c)      for each $E' \in \{E \cup \{X \rightarrow Y\},$
                               $E - \{X \rightarrow Y\},$
                               $E - \{X \rightarrow Y\} \cup \{Y \rightarrow X\}\}$
    (d)          $T' \leftarrow ProbabilityTables(E', D)$
    (e)          $B' \leftarrow \langle U, E', T' \rangle$
    (f)          $newscore \leftarrow BICscore(B', D)$
    (g)          if $newscore > score$ then
                     $B \leftarrow B'$
                     $score \leftarrow newscore$
6.  while $score > maxscore$
7.  Return $B$
```

Figure 1.19: Pseudo-code for the hill-climbing search algorithm for constructing a graph from a given data set D.

choice at each stage of the algorithm in order to find the global optimum (top of the hill) of specified objective function by essentially looking at the local gradient and following the curve in the direction of the steepest ascent. Hence, it consistently replaces the current solution with the best of its neighbours, as long as it scores better than the current. The search starts from either an empty, full, or possibly random graph.

The procedure *ProbabilityTables()* reported in figure 1.19, produces estimates of the parameters of the local joint probability distributions (compactly, pdfs), given a causal structure, typically through a maximum-likelihood estimation of the probability entries from the data set D. For multinomial local pdfs, it consists on counting the number of cases that fall into each table entry of each multinomial probability table in the graph. In fact, the main loop algorithm attempts every possible single-edge addition, removal, or reversal, updating the current graph with the one which increases the score.

This is iterated until no single-edge change increasing the score, is reachable (Margaritis, 2003).
Multiple restarts from random points (initial graphs) in the space allow for increasing the chances to reach a global maximum.
The hill-climbing algorithm is efficiently implemented in R-packages: *bnlearn* (Scutari, 2010) for categorical or continuous data and *deal* (Boettcher and Dethlefsen 2003), which implements a hill-climbing search for mixed data.
As for constrain-based methods, literature offers many other heuristic search algorithms. See, for instance, genetic algorithms (Larranaga et al., 1996) and Greedy Equivalent Search (GES) (Meek, 1997).

1.3.3 Pros and Cons of Constrain-based and Score-based Methods

As a causal graph is a structure encoding the joint distribution of the attributes, it may suggest that the method to be preferred is the one that better fits the data, leading to the scoring-based learning algorithms. Anyway, according to d-separation criterion, causal graphs also encode conditional independence relationships. As a consequence, using feasible statistical tests (such as Chi-squared test and mutual information test), we can find the conditional independence structure among the attributes and use these relationships as constraints to construct the causal graph's structure.
It is clear that a "best practice" does not exist as some pros and cons arise to both methods.
In this section we are going to explore differences and similarities of such methods, with the purpose of understanding, whatever method may be chosen, what is lost and what is gained in terms of causal structure learning.

In literature there are conflicting opinions: for example Heckerman *et al.* (1997) show how the score-based methods often have advantages over constrain-based ones, for allowing finer distinctions among model structures as well as better inference for combining information from different models. Conversely Friedman *et al.* (1997) prove that the general scoring-based methods may result in poor classifiers; score-based algorithms would, indeed, tend to favour complete graph structures in which every variable is connected to every other variable, leading to overfitting.

Both constrain-based and score-based approaches involve choices which could lead to different outputs: the conditional independence test as well as the alpha level in the first, and the score function to maximize in the second. In Zhang (2006) it is underlined how, from a practical point of view, the score-based approach is in general more stable with small to moderate sample sizes than the constraint-based approach. It also always returns a unique object unlike, for example, the PC algorithm that often gives as output a class of objects (the class of equivalent DAGs) visualized in a unique complete partially directed acyclic graph (CPDAG) with bidirected edges.

1.4 Estimating the Causal Effect from Observational Data: Pearl's do-calculus

Given a DAG, one could be also interested to estimate the size of the existing causal effects between pairs of variables or to predict effects of actions and interventions. For example, we may want to explore the probability distribution of some variables Y, possibly conditional on some other variables Z, as a variable X were manipulated to take some value in some way and whether there is a link between the pre-intervention probability and the post-intervention probability. Notice that this represents the primitive object of analysis in the potential-outcome framework or counterfactual analysis (Neyman-Rubin-Holland Model, 1986).

The Pearl's do-calculus, also called intervention calculus (Pearl, 2000), provides a useful language to specify how the pre-intervention distribution would change in response to external interventions, indicated by the operator *do*.

1.4.1 The intervention calculus's framework

When we talk about "intervention", we deal with a certain kind of interventions: first of all, for intervention on a variable X, we are meaning that the

direct target of the intervention is X. Then, intervention must be local, that is, it should affect only the target variable while local mechanism for other variables should remain unaffected by the intervention.
Manipulation (or intervention) can be thought as a local surgery with respect to causal mechanisms, as remote changes occurring to other variables after the intervention are due to propagation via the original causal mechanisms unaffected by the intervention (Zhang, 2006).

In order to represent the effect of an intervention on a set of variables we need a new notation. Assume, for instance, that we have variables $X_1, X_2, ..., X_n \subseteq V$ and we want to interpret the counterfactual phrase "had X_i taken the value x_i'" in terms of a hypothetical modification in the model. Such kind of sentences appears to be counterfactual, because they deal with unobserved quantities that differ from those actually observed (Pear, 2000). Formally, interventions and counterfactuals are defined by means of the mathematical operator called $do(X_i = x_i')$ or $do(x_i')$ for short, while an equivalent notation, using *set(x)* instead of *do(x)*, was used in Pearl (1995). It simulates physical intervention by removing certain functions from the model and replaces them with the constant $X_i = x_i$, while keeping any other variable unchanged. The new model thus created, when solved for the distribution of X_i, provides the "causal" effect defined as:

Definition 1.4.1 (**Causal Effect**). *Given two disjoint sets of variables, X_i and Y, the causal effect of X_i on $Y$is denoted as $P(y|do(X_i = x_i'))$ and gives the distribution of Y that would occur if treatment condition $X_i = x_i$ was enforced uniformly over the population via some intervention.*

The modification of an existing model entails the transformation between the pre-intervention and post-intervention distributions which can be expressed in the *truncated factorization* following directly by equation 1.5:

$$P(x_1, ..., x_n|do(X_i = x_i')) = \begin{cases} \prod_{j=1, j\neq i}^{n} P(x_j|pa_j)|_{x_i = x_i'} & \text{if } x_i = x_i' \\ 0 & \text{if } x_i \neq x_i' \end{cases} \tag{1.9}$$

where pa_j denote the parents of variable X_j. This formula uses the DAG structure to write the post-intervention distribution, in the left-hand side, in

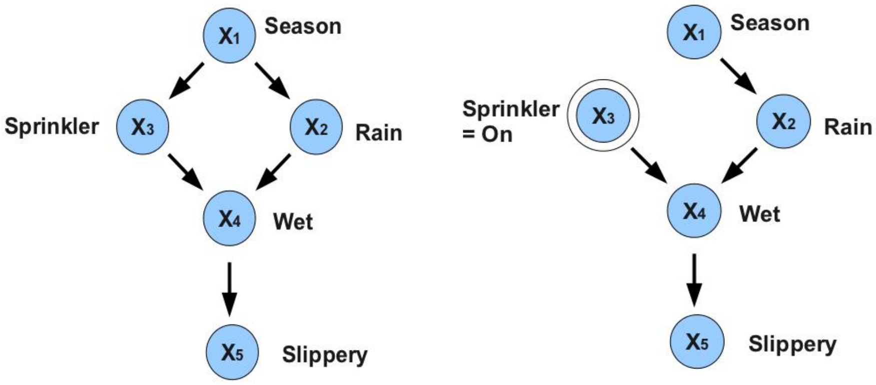


Figure 1.20: DAG representing the intervention of "turning the sprinkler on"

terms of the pre-intervention distribution $P(x_j|pa_j)$. Equation 1.9 reflects the removal of $P(x_i|pa_i)$'s term from the product; since the intervention "forces" X_i to be equal to x'_i, then pa_i no longer influences X_i.
Graphically, this is equivalent to removing all the links between PA_i and X_i while keeping intact the rest of the network.

For example, let consider again the rain's DAG shown in figure 1.6 and suppose we observe a particular spot on the street during some hour; moreover, let remind that the random variable X_3 denotes whether the sprinkler was on during that hour ($X_3 = 1$ if on, $X_3 = 0$ if off). If we wanted to represent the causal effect of turning the sprinkler on, we should remove the edge from X_1 to X_3 and assign the value $X_3 = 1$, even $X_3 = On$. The resulting graph can be seen in figure 1.20 : deleting the edge represents the understanding that, when we physically turn the sprinkler on, season has no longer any effect about the state of the sprinkler. The resulting joint distribution on the remaining variables will be:

$$P_{X_3=on}(x_1, x_2, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_4|x_2, X_3 = On)P(x_5|x_4)$$

. Notice, furthermore, the difference between the "intervention" $do(X = x')$ and the "observation" $X = x'$. We recall that the random variable X_4 denotes whether the street was wet at the end of that hour ($X_4 = 1$ if wet, $X_4 = 0$

otherwise). Assuming that $P(X_3 = 1) = 0.1$, $P(X_4 = 1|X_3 = 1) = 0.99$ (the street is almost always still wet at the end of the hour when the sprinkler is on that hour) and $P(X_4 = 1|X_3 = 0) = 0.02$ (if the sprinkler is off, the street is rarely wet on that hour). From the "observation" of $X_4 = 1$, we can calculate, with Bayesian conditioning, the probability $P(X_3 = 1|X_4 = 1) = 0.85$, namely, by observing the street to be wet, the probability that there was the sprinkler on in the last hour is 0.85. However, if we take a tank of water and force the street to be wet at a randomly chosen hour, we "intervene" as $do(X_4 = 1)$, then $P(X_3 = 1|do(X_4 = 1)) = P(X_3 = 1) = 0.1$. Thus, the distribution of the random variable describing sprinkler is quite different when making an observation versus when making an intervention.

If we are interested in the effect of one variable $X_i \subseteq V$ on another variable $Y = X_{n+1}$, we have to compute the distribution of Y after an intervention $do(X_i = x_i)$. Integrating out $w = \{x_1, ..., x_n\}$ in equation 1.9, it simplifies as:

$$P(Y = y|do(X_i = x_i')) = \begin{cases} P(x_i), & \text{if } Y \in pa_i \\ \int_w P(y|x_i', pa_i)P(pa_i)\partial pa_i, & \text{if } Y \notin pa_i \end{cases} \tag{1.10}$$

where $P(.)$ and $P(.|x_i', pa_i)$ are pre-intervention distributions.

According to Pearl (2000), it is equivalent to summarize the distribution generated by an intervention via expectation, thus equation 1.10 becomes:

$$E(Y|do(X_i = x_i')) = \begin{cases} E(Y), & \text{if } Y \in pa_i \\ \int_w E(Y|x_i', pa_i)P(pa_i)\partial pa_i, & \text{if } Y \notin pa_i \end{cases} \tag{1.11}$$

and the causal effect of $do(X_i = x_i')$ on Y is formulated as:

$$\frac{\partial}{\partial x} E(Y|do(X_i = x))|_{x=x_i'} \tag{1.12}$$

Following Rubin's definition of causal effect as $E(Y_{x'}) - E(Y_{x''})$, where x' and x'' are two levels of a treatment variable X_i, equation 1.12 may, equivalently, be rewritten in terms of the difference:

$$E[Y|do(X_i = x')] - E[Y|do(X_i = x'')] \tag{1.13}$$

Another measure of causal effect is then given by the ratio:

$$E[Y|do(X_i = x')]/E[Y|do(X_i = x''] \tag{1.14}$$

The causal effect we have analysed so far, $P(y|do(x'))$, measures the *total* effect of a variable (or a set of variables) X_i on a response variable Y . In many cases, anyway, the target of investigation and attention is focused instead on the *direct* effect of X_i on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X_i while all other factors in the analysis are held fixed.

When we want hence to evaluate the direct effect of a variable X_i on a variable Y, it may be necessary to *adjust* our measurements for possible covariates Z or "confounders". The word *confounding* refers in literature many different concepts (Greenland *et al.*, 1999): in the oldest usage, predominating in sociology and epidemiology is a *bias* described as a mixing of effects of extraneous factors, called confounders; in a more recent usage, it is synonym for *non-collapsibility*, namely that any statistical relationship between two variables may be reversed by including additional factors in the analysis. The idea behind adjustment consists in partitioning the population in groups homogeneous relative to Z, assessing the effect of X_i on Y in each homogeneous group and then averaging the result.

In Pearl (1993), the author shows two simple graphical (and hence visual) algorithms for checking whether a set of variables $Z \subseteq V$ would be sufficient for identifying $P(y|do(x_i')$. The first is:

Definition 1.4.2 (**The Back-Door Criterion**). *Given a DAG and a set Z of variables in the graph, it can be shown that Z is a sufficient set of covariates for "adjustment" for X_i, Y if, upon adjustment for Z:*

1. *No nodes in Z are descendant of X_i, and;*
2. *Z blocks every path between X_i and Y that contains an arrow in X_i (namely all paths that end with an arrow pointing to X_i).*

When the back-door criterion is fulfilled by a set of measured covariates, it is possible to estimate the total average causal effect of X_i on Y, as follow:

Theorem 1.4.1 (**Back-Door Adjustment**). *If a set of variables Z satisfies the back-door criterion relative to (X_i, Y), then the causal effect -as defined by 1.4.1- of X_i on Y is identifiable and is given by the formula:*

$$P(y|do(X_i = x')) = \int_z (y|x, z)P(z) \tag{1.15}$$

Note that the expression in 1.10 for $Y \notin pa_i$ is a special case of back-door adjustment where $Z = pa_i$ since pa_i satisfies the back-door criterion relative to (X_i, Y) if $Y \notin pa_i$.

The condition 1) of definition 1.4.2 reflects the prevailing practice that "the concomitant observations should be quite unaffected by the treatment" (Cox 1958, p.48). Anyway, confounders that are affected by the treatment can be used to facilitate causal inference as well. Indeed, the Pearl's *front-door criterion* (Pearl and Robins, 1995) constitutes the second building block to identifying causal effects in presence of confounders.

Definition 1.4.3 (**The Front-Door Criterion**). *A set Z of variables is said to satisfy the front-door criterion relative to an ordered pair of variables (X_i, Y) if:*

1. *Z intercepts all directed path from X_i to Y;*

2. *there is back-door path from X_i to Z;*

3. *all back-door paths from Z to Y are blocked by X_i.*

Theorem 1.4.2 (**Front-Door Adjustment**). *If Z satisfies the front-door criterion relative to (X_i, Y), then the causal effect of X_i on Y is identifiable and is given by the formula:*

$$P(y|do(X_i = x')) = \int_z (z|x) \int_{x'} (y|x', z)P(z) \tag{1.16}$$

Because, by clause 1) of 1.4.3, Z blocks all directed paths from X_i to Y, any causal dependence of Y on X_i must be mediated by a dependence of Y on Z:

$$P(y|do(X = x')) = \int_z P(y|do(Z = z))P(Z = z|do(X = x'))$$

Clause (2) says that we can estimate the effect of X_i on Z directly,

$$P(Z = z|do(X = x')) = P(Z = z|X = x')$$

Clause (3) say that X_i satisfies the back-door criterion for estimating the effect of Z on Y, so really we are using the back-door criterion.

Hence, both back-door and front-door criteria are *sufficient* for estimating causal effects from probabilistic distributions in presence of confounders. The criteria enable the analyst to search for an optimal set of covariates, namely a set Z that minimizes measurement cost or sampling variability (Tian *et al.*, 1998). Applications to epidemiological research are given in Greenland *et al.* (1999) and in Rothman *et al.* (2008), where the set Z is called "sufficient set"; *admissible* or *decounfounding* set are alternative terms (Pearl, 2000). Both criteria described above lead to the definition of *controlled direct effect* (CDE), where the term "controlled" just stays to indicate the fact that total effect is adjusted for the set of confounders Z. Focusing on differences of expectations, CDE is formulated as:

$$CDE \triangleq E[Y|do(X_i = x'), do(z)] - E[Y|do(X_i = x''), do(z)] \qquad (1.17)$$

where Z is any set of mediating variables that intercept all indirect paths between X_i and Y. Graphical identification conditions for expressions of the type $E(Y|do(x), do(z_1), do(z_2), ..., do(z_k))$ were derived by Pearl and Robins (1995) (see Pearl, 2000a, Chapter 4) using sequential application of the back-door condition 1.4.2.

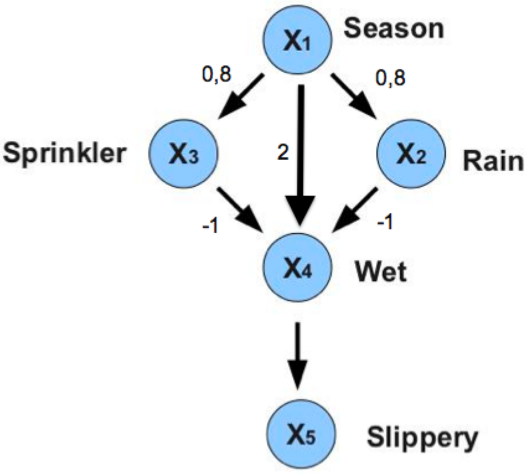


Figure 1.21: Graphical representation of the example

1.4.2 Intervention calculus vs regression: an example

Before proceeding in the intervention calculus's theory, a simple example can illustrate the difference between computing the causal effect and a multiple regression; indeed the first can be rather seen as a regression with selection of right covariates (Maathuis *et al.*, 2009). Now we call back the rain's DAG to illustrate such difference with a simple example; let us consider the following model (see also figure 1.21) where:

$$
\begin{aligned}
&X_1 = \varepsilon_3\\
&X_2 = 0.8X_1 + \varepsilon_2\\
&X_3 = 0.8X_1 + \varepsilon_3\\
&X_4 = -X_3 + 2X_1 - X_2 + \varepsilon_4
\end{aligned}
$$

where $\varepsilon_1 \sim N(0,1)$, $\varepsilon_2 \sim N(0,0.36)$, $\varepsilon_3 \sim N(0,0.36)$ and $\varepsilon_4 \sim N(0,1)$. Note that X_1, X_2 and X_3 all have variance 1, so that we are able to compare their regression coefficients or "causal effects".

By applying a multiple linear regression $X_4 = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_4$, we compute coefficients $\beta_1 = 2$, $\beta_2 = \beta_3 = -1$ thus X_1 would appear as the "most important" variable in the model.

Now let us apply intervention calculus. The parental sets are: $pa_1 = \varnothing$,

$pa_2 = X_1$ and $pa_3 = X_1$; θ_i with $i = 1, 2, 3$, represents the causal effect of X_1, X_2 and X_3 on X_4.
We now compute $\theta_1 = \beta_{1|\varnothing} = 0.4$, $\theta_2 = \beta_{2|X_1} = -1$ and $\theta_3 = \beta_{3|X_1} = -1$. Comparing the causal effect with the regression coefficients, we see that $\theta_2 = \beta_2 = -1$ as well as $\theta_3 = \beta_3 = -1$ but $\theta_1 \neq \beta_1$, in particular to note that in the intervention calculus X_1 appears the least important variable.

This example shows how computing regression and causal effects leads to different results as the set of controlled variables is different; thus a variable may show a strong association with a target variable while having a small causal effect.
Since X_4 is not a parent of any of the other variables considered, the distinction between intervention calculus and multiple regression can be interpreted as causal effect θ_i measures the *total* effect of explicative variables $X_i, i = 1, 2, 3$ on the response variable X_4, namely the sensitivity of X_4 to interventional changes. On the other hand, the regression parameter β_i measures the *direct* effect of X_i on X_4 in terms of sensitivity of the target output to interventional changes in X_i with $i = 1, 2, 3$ when all other variables in the model are held fixed (Maathuis *et al.*, 2009).

1.4.3 Some issues and the IDA algorithm

Although every term in the factorization 1.9 is assumed to be known as the pre-intervention probabilities can be consistently estimated from observational data, some complications may arise.

First, as discussed in the previous section, the set of observed variables may be causally insufficient (section 1.3.1), namely variables of interest may suffer from confounding due to latent common causes, or even that a unit is sampled in virtue of the value of selection variables. Furthermore, the observed variable may be unobservable. Even assuming a fully knowledge of the causal DAG with latent variables, the prediction of certain intervention effects may not be possible as our knowledge about the pre-intervention probability concerns merely the marginal probability of the observed variables instead of the joint probability of all variables in the DAG (Pearl in

1995, 1998, 2000, and more recently Tian and Pearl in 2004 dealt with this situation).
Secondly, at the beginning of this section, explaining calculus intervention's framework, we assumed to fully know the "true" DAG. This is seldom the case as mostly we try to infer the causal structure from observational data, and at most we can discover some features concerning the "true" causal graph. Thus obstacles in predicting certain intervention effects could arise by causal insufficiency due to the presence of latent variables as well as insufficiency of the causal information inferred from data or even, in the worst case, by a combination of both the two. (Zhang, 2006).

As the assumption for determining causal effects of knowing the true DAG is, in most of cases, unrealistic, Maathuis *et al.* (2009) has proposed a new methodology for determining causal effects in unknown DAG. The idea is to estimate the equivalence class of the true underlying DAG and then to apply the do-calculus on each DAG in the equivalence class. This procedure results in a *multiset* - namely a sort of set where the multiplicity of elements matters- of possible causal effects including the "true" causal effect. In this way a lower *bound* of the true causal effect can be computed.
These ideas are incorporated in the IDA algorithm (Intervention calculus when the DAG is Absent) implemented in the r package's pcalg (Kalisch *et al.*, 2012).

To estimate the bounds on causal effect, denoted as θ_i, of X_i on the response variable Y, IDA algorithm requires two main assumptions:

1. The distribution of $(X_1, ...X_i, ..., X_p, Y)$ is multivariate normal; furthermore, it is Markovian and faithful to the true (unknown) causal DAG.;

2. $X_1, ..., X_p$ have equal variance.

Assumption 2) is made for convenience, in order to easily compare the causal effects of different variables, while assumption 1) implies that the causal effect $E(Y|x'_i, pa_i)$, defined in equation 1.12, is linear in x'_i, pa_i and can be easily computed as:

$$\theta_i = E(Y|x_i, pa_i) = \beta_0 + \beta_i x'_i + \beta^T_{pa_i} pa_i \qquad (1.18)$$

for some values β_0, $\beta_i \in \mathbb{R}$ and $\beta_{pa_i} \in \mathbb{R}^{|pa_i|}$, where $|pa_i|$ is the cardinality of the set pa_i. Furthermore:

$$\theta_i = \int E(Y|x_i', pa_i)P(pa_i)\partial pa_i = \beta_i x_i' + \int \beta_{pa_i}^T pa_i P(pa_i)\partial pa_i \qquad (1.19)$$

According to Maathuis *et al.* (2009), the causal effect of X_i on Y when $Y \notin pa_i$ can now be computed by means of equation 1.12 yielding β_i which coincides with the regression coefficient of X_i in regressing Y on X_i and pa_i. Thus the causal effect of X_i on Y is given by:

$$\theta_i = \begin{cases} 0, & \text{if } Y \in pa_i \\ \beta_{i|pa_i}, & \text{if } Y \notin pa_i \end{cases} \qquad (1.20)$$

Note that the causal effect is zero when $Y \in pa_i$, since Y is then a direct cause of X_i.

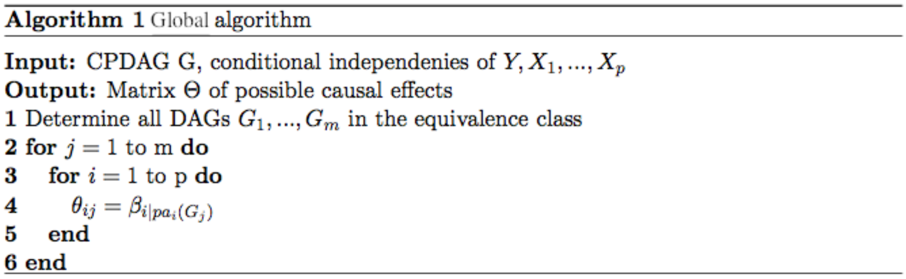
Since, under assumption 2), the causal effect does not depend on the value x_i, it can be interpreted, for any x_i, as:

$$\theta_i = E[Y|do(X_i = x_i + 1)] - E[Y|do(X_i = x_i)] \qquad (1.21)$$

There are two versions of IDA algorithm. A population version assumes that all conditional independences are known exactly and gives the correct complete partially directed acyclic graph (CPDAG) G. In the sample version, the conditional independences, estimated by data, are used as input to produce an estimated CPDAG $\widehat{G}$.

This second version of IDA appears as more useful when dealing with real data as in most cases we do not know the "true" causal graph. Let assume that we have a sample of n iid observations of $(X_1, ..., X_p, Y) = (X_1, ..., X_{p+1})$; first it is necessary to estimate the partial correlations $\hat{\rho}_{nij|S}$ between X_i and X_j given a set of other variables S, then we need to apply the PC algorithm to estimate the corresponding CPDAG G (Kalisch and Bühlmann, 2007). This requires multiple testing for Z-transformed partial correlations as given by:

$$\hat{Z}_{nij|S} = \frac{1}{2}\ln\frac{1 + \rho_{nij|S}}{1 - \rho_{nij|S}} \qquad (1.22)$$



Algorithm 1 Global algorithm

```
Input: CPDAG G, conditional independenies of $Y, X_1, ..., X_p$
Output: Matrix $\Theta$ of possible causal effects
Determine all DAGs $G_1, ..., G_m$ in the equivalence class
for $j = 1$ to m do
    for $i = 1$ to p do
        $\theta_{ij} = \beta_{i|pa_i(G_j)}$
    end
end
```

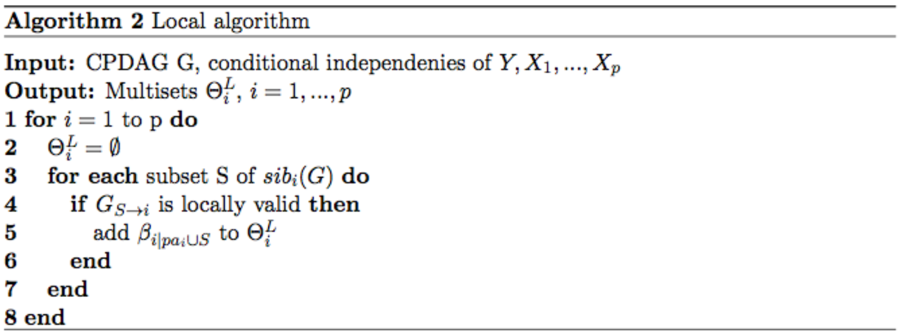
Figure 1.22: Global version of IDA algorithm

Under assumption 1) and for $\rho_{ij|S} = 0$, $\hat{Z}_{nij|S} \sim N(0, (n - |S| - 3) - 1)$ then $\rho_{ij|S} \neq 0$ occurs if:

$$|\hat{Z}_{nij|S}|\sqrt{n - |S| - 3} > \Phi^{-1}(1 - \frac{\alpha}{2}) \tag{1.23}$$

where Φ is the standard normal distribution function and $0 < \alpha < 1$ is a parameter representing the significance level of a single partial correlation test. The choice of an appropriate value for α is not straightforward at all. For instance, it can be chosen via a Bayesian Information Criterion (BIC) (see Maathuis *et al.* (2009) p.13).
Through the estimated CPDAG $\hat{G}(\alpha)$ is then possible to estimate the multisets $\hat{\Theta}_{ni}(\alpha) = \{\hat{\beta}_{i|pa_i(G_1)}, \hat{\beta}_{i|pa_i(G_2)}, ..., \hat{\beta}_{i|pa_i(G_m)}\}$ of possible causal effects for every G_j, $j = 1, ..., m$ DAG in the equivalent class, using the sample versions of equation 1.20.

Both a global and a local algorithm are implemented in pcalg r's package. In global one, according to the pseudocode described in figure 1.22, the set of possible causal effects for all DAGs in the equivalence class of the estimated CPDAG are computed. This method is suitable for small graphs (Maathius indicates up to 10 nodes) while as the number of covariates increases, it quickly becomes infeasible. Thus a faster "localized" algorithm has been developed. The local version of the sample algorithm is based on the following idea. Let assume for the computation of the causal effects $\hat{\Theta}_{n1}$ of X_1 on Y, that the key elements are the X_1's parents in the different DAGs in all the equivalence class. All possible parental sets of X_1 have to be determined by considering the CPDAG G, namely all sets $pa_1(G) \cup S$ where $S \subseteq sib_1(G)$



Algorithm 2 Local algorithm

```
Input: CPDAG G, conditional independenies of $Y, X_1, ..., X_p$
Output: Multisets $\Theta_i^L$, $i = 1, ..., p$
for $i = 1$ to p do
  $\Theta_i^L = \emptyset$
  for each subset S of $sib_i(G)$ do
    if $G_{S \to i}$ is locally valid then
      add $\beta_{i|pa_i \cup S}$ to $\Theta_i^L$
    end
  end
end
```

Figure 1.23: Local version of IDA algorithm

and sib_1 stays for X_1's siblings, the vertices linked with X_1 by an undirected edge. The sets S determine the direction of the edges between X_1 and the nodes in $sib_1(G)$. Particularly, all edges between X_1 and nodes in S must be directed towards X_1 and all edges between X_1 and nodes in $sib_1(G) \setminus S$ must be directed away from X_1. According to such a rule, $G_{S \to 1}$ denotes the graph obtained by changing all undirected edges $X_j - X_1$ where $X_j \in S$ into directed edges $X_j \to X_1$, and all undirected edges $X_j - X_1$ where $X_j \in sib_1(G) \setminus S$ into directed edges $X_j \leftarrow X_1$ (Maathuis *et al.*, 2009).

$G_{S \to 1}$ is said *locally valid* if, compared with G, it does not contain any other v-structure with X_1 as a collider; therefore we can check if $G_{S \to 1}$ is locally valid for each subset S. By excluding all locally valid sets, we gain all feasible parental sets of X_1. A new multiset $\hat{\Theta}_{n1}^L$ is then formed by taking all elements $\beta_{1|pa_1 \cup S}$ for which $G_{S \to 1}$ is locally valid. This procedure is summarized in figure 1.23. In Maathuis *et al.* (2009) it is shown that the global and the local algorithms drive to the same sets of distinct values, namely $\hat{\Theta}_{n1}^L = \hat{\Theta}_{n1}$, except for multiplicities.

It follows that the multisets of total causal effects of X_1 on Y from the global and the local method have the same unique values. Thus for computing the lower bound of the causal effects it does not matter which algorithm is used. Anyway, with the local algorithm we loose information about multiplicities of such values and therefore, for instance, their probability to occur.

Chapter 2

Causality and Gender Gap

The aim of this Chapter is to "drop" the causal approach of Chapter 1 in a gendered perspective, in particular for assessing potential gaps occurring among women and men.

After defining, briefly, the general framework and the terminology of gender studies, we develop a translation device that, from the language of causality, would allow to express gender inequality/equality occurrence in a statistical as well as graphical way.

On the basis of the Pearl's do-calculus, then, we propose a "causal" definition of Gender Gap, in terms of causal effect of variable "Gender" on a target outcome concerning a certain field of interest. It allows to give a synthetic measure of how the target outcome would change in response to external intervention (manipulation) on variable "Gender".

2.1 Gender Framework

In order to show the effectiveness of causal graphs in assessing gender gap, we need to introduce few key-concepts for framing the gender issues landscape and for contextualizing the research goals.

"Gender" is certainly a word in common use; anyway, a more formal and accurate definition is required. The importance that gender plays within the equalities recognised at international level, moreover, it is relatively recent.

For this reason, the demand of “ad hoc” statistical tools for assessing gender gap is pressing and represents an open field of research.
In this section we discuss more in details such topics.

2.1.1 Gender, Gender Gap and Development

With the term *Gender*, we refer to differences between men and women not determined biologically as a result of sexual characteristics but as cultural expectations and socially determined roles, behaviours, activities, and attributes that a given society considers appropriate for men and women (APA, 2012) and (WHO, 2013). It is a central organizing principle of societies, and often governs the processes of production and reproduction, consumption and distribution (FAO, 1997).
Globally, communities interpret biological differences between men and women to create “gender-normative” behaviours and to determine women’s and men’s different access to rights, resources, education, power in society and even health behaviours. Although the nature and degree of these differences vary from society to society, they typically favour men, creating an imbalance in power and gender inequalities in all countries.

Gender equality “between women and men exists when both genders are able to share equally in the distribution of power and knowledge and have equal opportunities, rights and obligations” (UNESCO, 2011).
However the definition of gender equality depends on the understanding of gender differences. Are all differences also inequalities? Or are some differences valued not a sign of inequality? Does reaching gender equality merely mean changing the position of women, or does it mean a much deeper transformation that includes changing the lives of men as well? (UNECE, Developing Gender Statistics, 2010).
According to a first interpretation, equality recalls a single standard of evaluation with the implication that unless there is sameness there is not equality as it happens in the cases of equal pay for work of equal value.
In a second approach, there is equal valuation of different contributions, thus there is not a simple single standard against which men’s and women’s posi-

tions are assessed. This is, for instance, what concerns the unpaid care work and whether and if so, how can be treated as equivalent to paid work.
In a third perspective, equality between the genders will be only achieved through the transformation of practices and standards of both men and women, for example as reconciling work and family life by making the workplace compatible with family care for both parents. This approach demands major structural changes throughout society, since it involves a transformation of the whole social environment.

As stated before, the concept of gender equality is strictly linked to the understanding of gender differences. Systematic differences in men and women' outcomes are called *gender gaps*. Well known examples are: the ratio between women's and men's earnings for the same amount of work is 0.77, according to the most recent statistics from the U.S. Census (source: National Committee on Pay Equity, 2012); in 2008, U.N. Secretary-General Ban Ki-moon reported that one in every three women is likely "to be beaten, coerced into sex or otherwise abused in her lifetime" (source: OneWorld); since 1980 "women live longer then men all over the world" (Kalben, 2000).

In recent times, a focus on gender-equality has been widely recognised as vital to international development, involving politicians, economists and human rights activists.

> "To ensure equal opportunities between different groups of the population is one of the issues underlying the economic development and one of the main tools for reducing poverty" (World Bank, 2005).

At an international level, the intrinsic and instrumental value of gender equality has been widely recognised in several intergovernmental resolutions as the Millennium Declaration (UN, 2000), adopted by all Member States of the United Nations in 2000. It provides the framework for measuring progress towards the eight Millennium Development Goals (MDGs) [1] to be achieve by 2015. Gender equality and women's empowerment are development objectives in their own right (MDG 3 and 5), as well as critical channels for

[1] http://www.beta.undp.org/undp/en/home/mdgoverview.html

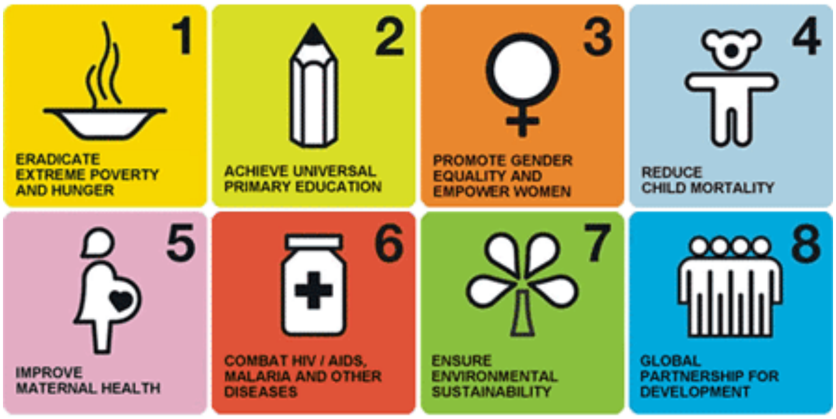

1
ERADICATE EXTREME POVERTY AND HUNGER
2
ACHIEVE UNIVERSAL PRIMARY EDUCATION
3
PROMOTE GENDER EQUALITY AND EMPOWER WOMEN
4
REDUCE CHILD MORTALITY
5
IMPROVE MATERNAL HEALTH
6
COMBAT HIV / AIDS, MALARIA AND OTHER DISEASES
7
ENSURE ENVIRONMENTAL SUSTAINABILITY
8
GLOBAL PARTNERSHIP FOR DEVELOPMENT

Figure 2.1: The eight UN Millenium Development Goals

achieving the other MDGs: they help to promote universal primary education (MDG 2), to reduce under-five mortality (MDG 4),to improve maternal health (MDG 5) and to reduce the risk of contracting HIV/AIDS (MDG 6).

According to 2013 World Development Report, the actions efforts should focus on four priority areas: 1) reducing excess female mortality and closing education gaps where they remain; 2) improving access to economic opportunities for women; 3) increasing women's voice and agency in the household and in society; and finally 4) limiting the reproduction of gender inequality across generations. Nevertheless, there is one cross-cutting priority: supporting evidence-based public action through improved quality of data, better knowledge generation and sharing, and better learning. In this, statistics may offer its contribution.

2.1.2 Gender Statistics: what, why, how

The expression *gender statistics* calls for a double interpretation (Mecatti *et al.*, 2012) as in the most widespread expression refers to the popular mix-up of statistical methodological tools and its typical products such as indexes, tables and graphs; whereas in a broader sense, it implies a forward-looking perspective which is inspired by the increasing demand of gender sensitive statistical information coming from society, official agencies, economy.

Gender statistics represents a proper independent field of statistics that cuts across traditional applications in social, economical, human and life science with the aims of:

- surfacing and quantitatively assessing gaps and issues based on either gender–as a social structure–or sex–as a biological factor–;
- providing information for formulating and monitoring data-driven policies;
- disseminating statistics and informing society.

Gender statistics is crucial for decision-making, “since without an understanding of the differences in the operation and effects of the policy on different population groups, such as on women and men, the full implications of the policy may not be understood and its objectives may not be fulfilled” (UNECE, 2010) [2].

Sex-disaggregated data are the foundation needed to show the existing gendered-based differences. However, this alone is, in general, not sufficient for producing sound results and reaching the objectives above mentioned.

The process of identifying gender relevance requires the identification of the areas that might contain significantly gendered dimensions, of current policy issues as well as a deep comprehension of the conceptual frameworks and methods used in official statistics. Indeed, several important frameworks and methods traditionally used in official statistics are biased against either women or men. It is a popular example the statistical concept of “economy”, which traditionally focuses on the monetized meaning represented in measures such as the Gross Domestic Product that omits unpaid household work.

Gender statistics may improve national statistical systems, through an unbiased review of definitions and concepts, an improvement of data collection detecting the development of new methods able to reflect diversities and inequalities between women and men in the entire society.

[2] $http : //live.unece.org/fileadmin/DAM/stats/publications/Developing_{G}ender_{S}tatistics.pdf$

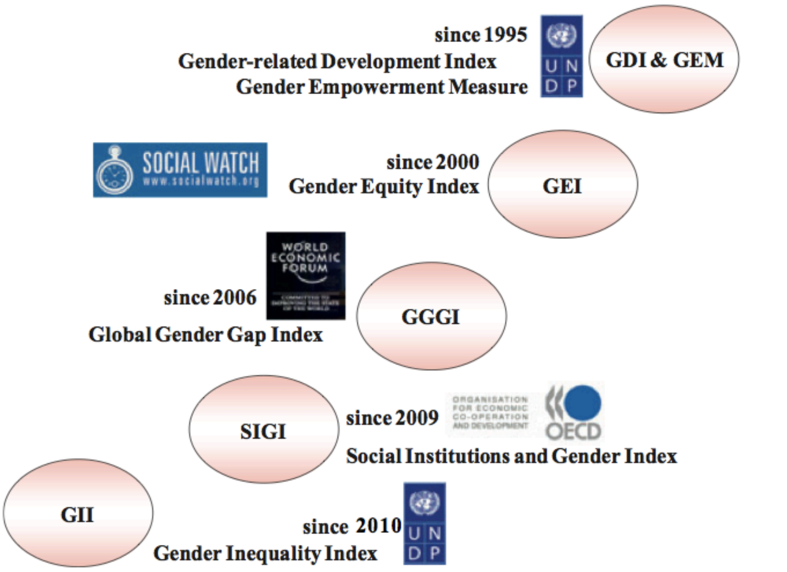


Figure 2.2: Main gender composite indicators released by International Agencies.

Several gender statistical measures have been released by international and supranational agencies since 1995, with the purpose to compare and possibly rank countries with respect to chosen gender-related macro-themes, such as economics, politics, education, health. Most of such measurements are composite indicators. A composite indicator is a number, usually in [0, 1] or [0, 100], provided by the aggregation of a set of simple indicators, each singled out to measure one particular component or aspect of the underlying latent dimension, this being the way to grasp the multi-dimensionality of the macro-subject under study (Mecatti *et al.*, 2012). These indicators are calculated as female/male ratio such that a value equal to 1 or 100 represents perfect equality while, when equal to 0, maximum inequality is shown. The main gender composite indicators are depicted in Figure 2.2. However, it is widely recognised that any composite indicator is limited by the subjectivity with which it is built up (OECD, 2008): it is enough to think to the many choices available in relation to weighting/aggregating system, standardization, to the variables/dimensions to be considered and excluded, to the link functions etc.

The arbitrariness in the many choices involved in the process risks to make opaque the index's construction with the consequence to produce unreliable indicators and objectionable rankings.

Another crucial aspect is the skepticism regarding whether a single indicator was able to grasp different gender related problems concerning both developed and in transition countries (Barnes and Bouchama 2011). In fact, once established that gender inequalities exist in most parts of the world, from Japan to Morocco, from Uzbekistan to the United States, yet the very definition of gender gaps varies, since inequalities between women and men are not the same everywhere and can take many different forms. Gender inequality is not a homogeneous phenomenon; it is, in fact, a collection of disparate and inter-linked issues such as natality and mortality inequality, basic human rights inequalities, special-opportunities inequality, professional inequalities, ownership inequality, household inequality and the like (Caligaris *et al.*, 2013). For instance, in developing-countries, gender disparities still affect human rights. Conversely, in OECD countries, equal access to education, health, survival ect., are values enshrined and guaranteed by national Constitutions and gender gap occur under different shapes, such as work-life balance, or political participation.
Gender statistics, thus requires the development of *ad hoc* methods and studies able to drop gender gaps in a well defined political, economic and social cultural framework.

2.2 Gender Gap Measurement: a Causal Inference Perspective

One of most quoted sentences of David Hume ties together causality and counterfactual dependences:

> we may define a *cause* to be an object followed by another.. or,
> in other word, where, if the first object had not been, the second
> never had existed.
> (Hume, 1748)

As discussed in the previous section, the term "gender gap" is usually used to indicate a differential in attitudes, behaviours, abilities, opportunities, access to resources among individuals based on either gender or sex. An action is said to be "discriminatory", with respect to gender, if the treatment of an individual would have been different had that person been of different gender. The challenge is, hence, to determine if a gender gap actually occurs but even to which extent a gender discrimination contributes to explain disparate outcomes between women and men. This is certainly a counterfactual matter, since, according to Hume's idea, if gender inequality had not been, differentials based on gender would have never existed.
In this section, we go back to causality, but in a new perspective: a gender perspective. We, indeed, show the translating bridge from the causal approach to gender framework, with the purpose to build up a new set of tools able to catch the "shady" causal mechanism originating gender gap.

2.2.1 Is Gender a Cause? A shift in emphasis

Due to its composite nature, as widely discussed in section 2.1.2, gender gaps cannot be *directly* observed. In order to identify the presence or absence of a gender discrimination, defined as a differential treatment of individuals based on gender, researchers typically observe an individual's gender (e.g., female) and a particular outcome (e.g., wages) and try to determine whether that outcome would have been different, had the individual been of a different gender (e.g.,male).
However, gender -as a social structure- or sex -as a biological factor- are not *causes* themselves of gender disparities in wages, nor are they the causes of differences in access to education, family wealth, or health outcomes and the like. Gender is an individual *characteristic* which may have physical and socio-cultural attributes so that, in this sense, it is not a manipulative variable.

In his article, Holland (1986) reported that he and Don Rubin had once made up the motto, "no causation without manipulation". Causal inference is indeed fundamentally related to experimentation, which is why the

randomized controlled trial (RCT) is widely considered to be the "gold standard" for the establishment of causality (Kaufman, 2008).
The idea of analysing, in observational studies, the causal effect of "things" that we, as human beings, could never influence, is incoherent because such things could never be the subject of a randomized experiment. The Rubin/Holland stance is clear: gender - as race - is an *attribute* that the person possesses and thus "it cannot be a *cause* in an experiment, because the notion of potentially exposability does not apply to it" (Holland, 1986); for the authors it makes no sense to talk about the causal effect of gender because, first, attributes are not subject to change by means of intervention, hence not to manipulation (a sex-transformation surgery?) and secondly, some *immutable characteristics*, such as gender, refer to both conception and identity of the person.

Meanwhile, other scholars have explored an alternative way with the idea that "perceptions" of immutable characteristics other than the "actual" traits are manipulable. Because we typically cannot observe the mechanisms through which the gender gap occurs, gender becomes a proxy for *gender perception*. This is intended as the way which a person is viewed or view itself as belonging to a gender. It is, thus, a social construct related to economical, educational, political, health or social decision, evaluation or procedure (APA, 2012).
By considering gender as a purely social construct, we can manipulate information about it without randomization, rather making such information unavailable; for example, Goldin and Rouse (1997) explore how information about the gender of applicants to symphony orchestras was removed through the creation of blind auditions. In such case we can assess the effect of a shift from having the information available to not having it available. If X is a binary variable representing the access to information on gender, the two possible values are x' =available and x'' =not available. According to intervention calculus notation, described in section 1.4.1 of Chapter 1, the manipulation is thus denoted with $do(X = x'')$ and disparities based on gender are not possible because the information is unavailable.

A shift in the emphasis on the gender perceptions requires some well-

defined causal questions to be posed.
Indeed, when studying the causal effect of gender, we do not contemplate an intervention in the form of manipulation of this attribute or even its perceptions(Greiner and Rubin, 2011); rather, the purpose is to establish whether a gender gap actually occurs and, if so, to grasp its nature, whether discriminatory or not, and to devise reducing policies. In this thesis, our aim is to draw inference on causal effects of gender perceptions in order not to study what would happen if we did intervene to alter such perceptions, rather to detect gender gap and decide whether to intervene in some remedial way.

2.2.2 Re-declining causal vocabulary in a gender perspective

Before delving to how to measure gender gap in a causal approach, we need to drop standard causal terminology in a gender perspective, allowing a new gendered interpretation.

The *unit* i of analysis is a person in some defined role, such as an applicant for a job, a student, a worker, a political candidate, a person in need of care etc. The *treatment* G_i is the unit's gender (gender perception), with $G_i \in \{0,1\}$. It is an immutable characteristic as perceived by the decider, namely the employer, the professor, the society, the family, etc; the unit i is *exposed* to treatment G_i when its gender is submitted to the decider's perception, for example, the employer seeing a particular gender listed on application; thus it is subject to manipulations as unavailability of information. The *timing of treatment* assignment is presumptively the moment the decider first perceives the unit's gender. Defining treatment as occurring at the moment of first perception captures the fact that variables, whose values are determined after that moment, may be affected by the perception itself. For example, in an applicant's job interview, an employer may evaluate a unit perceived to be male more favourably than an otherwise functionally identical unit perceived to be female (or vice versa), so the evaluation is considered as an intermediate outcome (Greiner and Rubin, 2011).

In this new gendered framework, we can define *gender discrimination* as the effect of the perceived gender G_i on the target outcome, indicated with Y (wage, job assumption, educational attainment, health status, etc.) controlling all the other variables. *Gender gap*, denoted in the following with $\Gamma(G)$, can be conceived as a difference in outcome between women and men; such gap however can arise either from gender discrimination (e.g. employment decision) or from biological differences (e.g. hormones) or socio-economical differences (e.g. educational achievement).

In observational studies, *confounders* are any measured or unmeasured quantity (Z) that are associated with (but not affected by) gender (G), causally preceding the target outcome (Y), and acting to *confound* the observed relationships. In figure 2.3, a straightforward example is given. The variables "Education" and "Productivity" can be viewed as confounders of the causal effect of "Gender" on "Wages"; they can considered as pre-treatment variables since they took place before the decider perceived the unit's gender, therefore suggesting to condition on them.

Causal Graphs, discussed in Chapter 1, are useful tools able to describe statistical models and to capture, by observational data, the causal relationships in a set of variables.
By including the variable "Gender" in such set, *Gender Equality* can be indeed visualized with the aid of Directed Acyclic Graphs (DAGs). Denoted, in the corresponding DAG, with G the node representing "Gender" and with Y the node referred to the target outcome, gender equality on the target outcome occurs in absence of a direct edge connecting G with Y. Statistically, it corresponds to conditional independence of the two variables, given all of the other variables in the graph (Pearl, 2000).
For example, in figure 2.3, if we consider the causal graph depicting the relationships among variables: Gender, Wage, Education and Productivity, the lack of a direct edge between the variable "Gender" and the variable "Wage" would mean that there is gender equality in wages, among people with same education and productivity. This suggests that gender has no direct causal effect on wages.

Conversely, the edge between the variable "Gender" and "Education" would

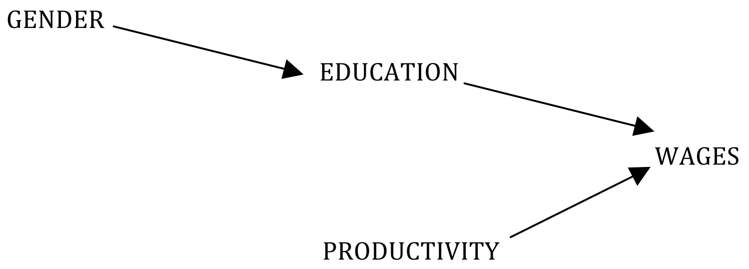


Figure 2.3: Causal relationship among the variables Gender, Education, Wages and Productivity

indicate a gender inequality in education, namely that gender affects education. According to what mentioned in section 1.2, the notation "Gender" → "Education" means, in fact, that "Gender" is a *direct cause* of "Education". Statistically, it implies no conditional independence between gender and education.

DAG allows even to represent graphically manipulation on variable gender that modify a select set of functions in the underlying model.
In section 1.4.1, we described how the truncated factorization in equation 1.9, due to intervention, graphically translates into removing all the links between the manipulated variable and its parents in the graph. In the example depicted in figure 2.4, the manipulation on gender implies inserting a new random variable $do(Gender = 1)$ to the graph that breaks the link between the variable "Gender" and its parent "Gender stereotypes", keeping intact the rest of the network. Once set $(Gender = 1)$, indeed, gender stereotypes don't affect gender any more.

2.2.3 Measuring Gender Gap: the causal effect

The underlying idea is that, if there were no gender gap, then we would expect people homogeneous in certain characteristics to have equal outcome (Fienberg and Haviland, 2003).
For instance, we are interested in causal relationships among variables: gender, gender gap in labour market and wages. In absence of a gender gap,

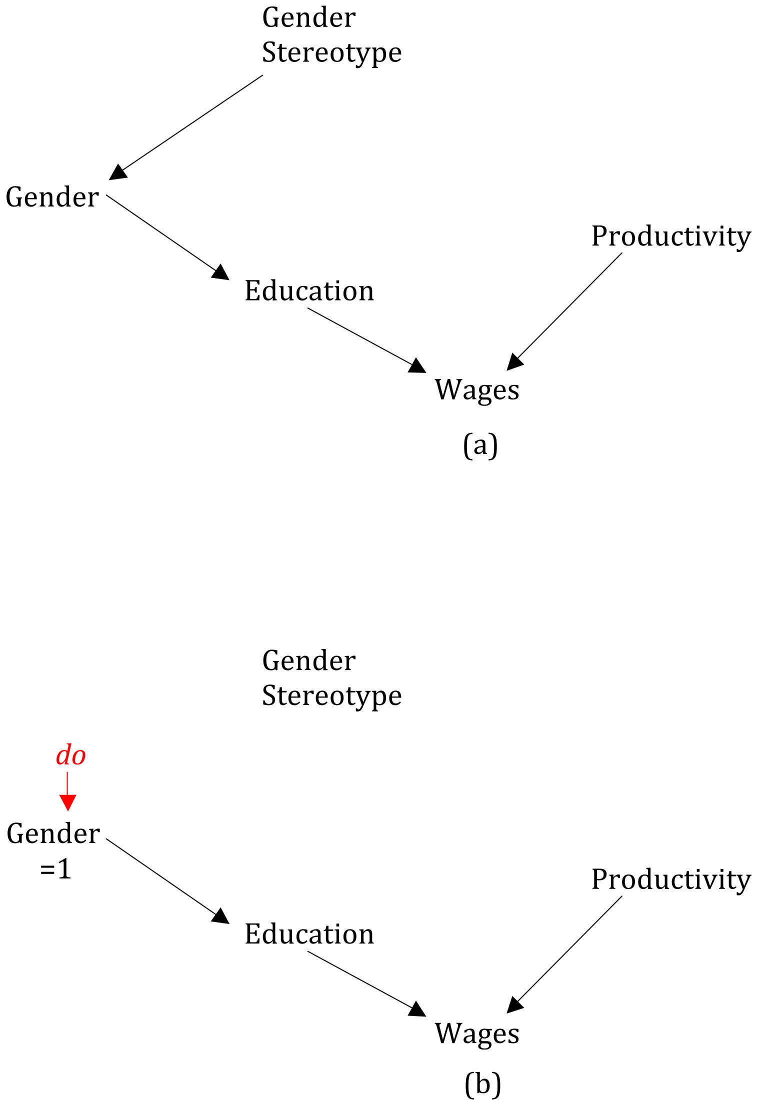
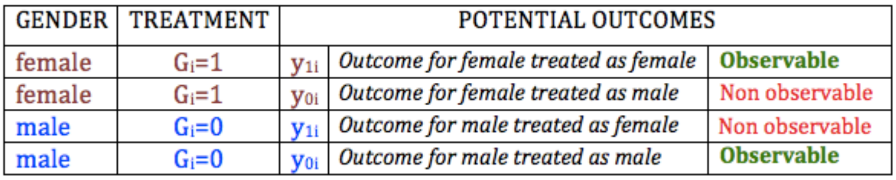


Figure 2.4: Causal relationships among the variables Gender stereotypes, Gender, Education, Wages and Productivity (a) pre-manipulation and (b) post-manipulation on Gender.

male and female workers with the same education and productivity would earn equal. The measurement of gender gap is thus clearly a counterfactual matter: “How much would a woman have gained if she had been a man?”. The answer to such kind of question is fundamental to infer whether a causal relationship between gender, as a proxy of gender perception, and wages occurs or does not occur, which, in turn, leads to determine whether a gender gap in labour market exists.
According to Rubin's experimental approach, the causal effect of gender would consist of the difference between two outcomes: the outcome in the case the individual were male and the outcome if the individual were female; in this sense, it is a counterfactual question. Although each individual i has an hypothetical potential outcome under either circumstances, only one of these outcomes is observed or, better, realized. Under this respect, it is a missing data problem as described by Rubin (1974).

Pearl (2009) suggests to use the notion of potential outcomes in the absence of a potential experiment, in order to attribute a cause to an effect. We need first to re-formulate the approach of literature in gender gap measurement context; the aim is therefore to understand how much of the observed gender gap (for example in terms of wages) is due to discrimination based on gender membership after adjusting for differences in the other observed factors.
The question can be formalized as following.
Let us consider a population of N individuals $(i = 1, ..., N)$ belonging to either of two mutually exclusive gendered groups indexed by $G_i \in \{0, 1\}$ such that the division of individuals into these two groups is based on whether an individual i has been exposed to treatment G_i or not. We set $G_i = 1$ if i has been treated, for instance, if female; otherwise, if male, we set $G_i = 0$, namely non-treated. Moreover, $\sum_{i=1}^{N} G_i = N_1$ and $N_0 + N_1 = N$ *i.e.* there are N_1 (N_0) treated or female (non-treated or male) individuals in the population. For each unit i, a vector of covariates, Z_i, is also observed while the realized (observable) target outcome is denoted by y_i for each unit i.
However, as summarized in table 2.1, of the four potential outcomes, only two could have been observed. For women, as $G_i = 1$, only the treated out-



GENDER	TREATMENT	POTENTIAL OUTCOMES		
female	G_i=1	y_{1i}	*Outcome for female treated as female*	**Observable**
female	G_i=1	y_{0i}	*Outcome for female treated as male*	Non observable
male	G_i=0	y_{1i}	*Outcome for male treated as female*	Non observable
male	G_i=0	y_{0i}	*Outcome for male treated as male*	**Observable**

Table 2.1: Table summarizing observable and non observable potential outcomes for females and males.

come y_{1i} is realized; conversely for men, for which $G_i = 0$, we can only observe the outcome y_{0i}. Consequently, for each individual i, we dispose of only one observable outcome and the other becomes counterfactual.

The gender effect (treatment effect) at individual level i, results as $\tau_i = y_{1i} - y_{0i}$, namely the difference between the target outcome if the individual i were treated as female and the outcome if it were treated as a male. The causal effect at population level, even called *average causal effect* (ACE), is instead given by $\tau_{ACE} = E(y_{1i} - y_{0i})$. However, as two potential outcome y_{1i}, y_{0i} from distinct intervention $G_i = 1$ and $G_i = 0$ cannot be observed for each unit i, potential outcome approach leads to what is said "black-box observation".

Our aim is now to illustrate the propose strategy for measuring gender gap. It is innovative as, unlike the existing gender gap indexes which provide a static snap of gender inequalities, it would give rather a dynamic sequence of snaps, showing how gender would affect the target outcome. It allows to catch the *causal nature* of gender gap in terms of causal effect of gender on the target variable Y.

We need, thus, to express queries about causal effect of gender as queries about the marginal distribution of the counterfactual variable of interest Y, written $P[Y(G_i = 1) = y]$.

According to subsection 1.4.1, counterfactual matters find expression by means of *do* operator. Indeed, the opaque English phrase "the value that Y would obtain in unit i, had G_i been $female$" and the physical processes that transfer changes in G_i into changes in Y, find their formal translation in the expression $E[Y|do(G_i = female)]$.

Resorting to causal effect definition in equation 1.13, we develop a new definition of gender gap as *gross* gender gap, denoted with $\Gamma_{gross}(G)$ and meant as the causal effect of gender on outcome Y. It has been denoted with "gross" to be distinguished by the "net" one, which will be described hereinafter. The gross gender gap is defined by the difference in conditional expected values:

$$\Gamma_{gross}(G) = E[y_1|do(G_i = 1)] - E[(y_0|do(G_i = 0)] \tag{2.1}$$

where y_1 and y_0 are the observed outcome as though one is treated as a member of the female or the male group; respectively, $G_i = 1$ indicates that respondents are women, $G_i = 0$ indicates that respondents are male. If we were interested, for example, in the wage gap, the notation y_i would represent the observed (log) wage and $\Gamma_{gross}(G)$ the gap between the expected women's and men's (log) wages.

Notice that the formulation in equation 2.1, although expressed by expected values, does not impute to gender gap a compensative connotation. Indeed, these expected values are conditioned on intervention on Gender variable, allowing to keep distinct the outcome of female group from the outcome of male one.

Following the alternative causal effect definition proposed in equation 1.14, another measure of gross gender gap in given by the ratio:

$$\Gamma_{gross}(G) = E[y_1|do(G_i = 1)]/E[y_0|do(G_i = 0)] \tag{2.2}$$

Notice that the gender gap in equations 2.1 or 2.2 is denoted as *gross*, in the following sense. Besides the gender membership, other characteristics, denoted with Z_i, might affect gender gap and whose distributions could differ between genders (for example the productivity and education in wage gap). In such cases, the quantity $\Gamma_{gross}(G)$ does not represent adequately the target of investigation as unadjusted for such characteristics.

Conversely, we are interested in the *net* gender gap, as the quantification of an effect adjusted by other variables in the model, i.e. the sensitivity of the target outcome to changes in the variable gender while all other factors in the analysis are held fixed. Therefore it is necessary to control such pre-treatment characteristics, defined in Chapter 1 as confounders, in the estimation process with the purpose to isolate and identify the actual gender gap due only to either gender discrimination or biological differences. The idea behind adjustment consists in partitioning the population in groups homogeneous relative to such confounding characteristics, assessing the effect of G_i on Y in each homogeneous group and then averaging the result.

Let note that the distinction occurring between "gross" and "net" gender gap finds its analogous in causal theory in Pearl's "total" and "direct" causal effect (Pearl, 2009), as illustrated in section 1.4.1. This represents a re-definition, in a gender perspective.

Recalling the example of wages, we would like to decompose the gross amount of wage gap in the amount due to differences in other characteristics, e.g. education, productivity etc., and the remaining variables in set considered. In order to gain this decomposition, we would need, ideally, full information on each individual i, with his/her characteristics besides group membership (Fienberg and Haviland, 2003); specifically we would want to know wages for each male (female) as he (she) would be paid were he (she) a member of female (male) group. Therefore, we need the estimation of such missing counterfactual matter (Fienberg and Haviland, 2003) in order to detect the "actual" gender gap, net of differences due to any other confounding variables.

The gender gap conditional on confounding characteristics $Z_i = z$, namely the net gender gap, then becomes (Słoczyński, 2013):

$$\Gamma_{net}(G) = E[(y_1|do(Z_i = z), do(G_i = 1)] - E[(y_0|do(Z_i = z), do(G_i = 0)] \tag{2.3}$$

or alternatively:

$$\Gamma_{net}(G) = \frac{E[(y_1|do(Z_i = z), do(G_i = 1)]}{E[(y_0|do(Z_i = z, do(G_i = 0)]} \tag{2.4}$$

In the wage example, the $\Gamma_{net}(G)$ allows for comparing men's wages with wages of women homogeneous for other characteristics Z_i, and for estimating the corresponding actual wage gender gap.

In equation 2.3, Z_i represents a sufficient or, according to section 1.4.1, "decounfounding" set of variables for estimating consistently these missing counterfactuals. Potential outcome literature refers to this as the *strong ignorability* (Greiner and Rubin, 2011), meaning that Z_i is an admissible set of covariates, if, given Z_i, the value that the outcome y_i would result had G_i been 1 (or 0) is independent of G_i, namely that $(y_0, y_1) \perp G_i|Z_i$.
Notice that, as mentioned in Chapter 1, ignorability fails to provide a workable criterion to guide the choice of such covariates (Pearl, 2000), since counterfactuals are unobservable. Pearl and Robins (1995)'s back-door criterion (1.4.2) and front-door criterion (1.4.3), as discussed in section 1.4.1, provide simple graphical solutions to assess the adequacy of controlling for a particular covariate set Z_i.
The net gender gap $\Gamma_{net}(G)$ in equation 2.3, thus, would be nothing that the Pearl (2009)'s *controlled direct effect* (CDE), defined in Chapter 1 in equation 1.17, of variable gender on the target outcome Y, defined as:

$$\Gamma_{net}(G) = CDE \triangleq E[Y|do(G_i = 1), do(z)] - E[Y|do(G_i = 0), do(z)] \quad (2.5)$$

where Z_i is a set of mediating variables satisfying the back-door or front-door criterion, hence intercepting all indirect paths between G and Y in the corresponding causal graph.

By assuming that Z_i satisfies the back-door or front-door criteria relative to (G_i, y_i), it follows that gender group membership is independent of the outcome y_0 (Heckman et al., 1998), then $E[y_0|do(G_i = 1), do(Z_i = z)] = E[y_0|do(G_i = 0), do(Z_i = z)]$. This can be interpreted as: for equal characteristics Z_i, the classification of an individual in the male or female group would not affect its expected wage. With these assumptions, we can now estimate the decomposition of the gross gender gap into the portion associated with confounders and another portion that is not, namely the controlled effect of gender on the outcome.
Let the average outcome, respectively, in female group $G_i = 1$ be expressed as:

$$E(y_1|G_i = 1) = \sum_{Z_i} p_{Wz} E(y_1|G_i = 1, Z_i = z) \quad (2.6)$$

where $E(y_1|G_i = 1, Z_i = z)$ is the expected outcome in the women' group with the observed characteristic $Z_i = z$, and p_{Wz} is the proportion of women with characteristic $Z_i = z$. Equation 2.6 provides an estimation of the "effect of treatment on the treated". In our interpretation this is the causal effect of gender on women's outcomes under the condition of gender discrimination. Similarly for the men's group $G_i = 0$ we set:

$$E(y_0|G_i = 0) = \sum_{Z_i} p_{Mz} E(y_0|G_i = 0, Z_i = z) \tag{2.7}$$

where, conversely, $E(y_0|G_i = 0, Z_i = z)$ would indicate the estimation of the "effect of treatment on the untreated". Substituting equations 2.6 and 2.7 in 2.1, we obtain the gross gender gap expressed as:

$$\begin{aligned}
&\Gamma_{gross}(G) = \\
&E(y_1|G_i = 1) - E(y_0|G_i = 0) \\
&= \sum_{Z_i} p_{Wz} E(y_1|G_i = 1, Z_i = z) - \sum_{Z_i} p_{Mz} E(y_0|G_i = 0, Z_i = z) \\
&= \sum_{Z_i} p_{Wz} [E(y_1|G_i = 1, Z_i = z) - E(y_0|G_i = 0, Z_i = z)] \\
&- \sum_{Z_i} [p_{Mz} - p_{Wz}] E(y_0|G_i = 0, Z_i = z) \qquad (2.8)
\end{aligned}$$

The assumption that Z_i were a deconfounding (or sufficient) set for estimating the controlled effect of gender on target outcome, allows to use equation 2.8, which may be observed, to estimate:

$$\begin{aligned}
&\Gamma_{gross}(G) = \\
&E(y_1|G_i = 1) - E(y_0|G_i = 0) \\
&= \sum_{Z_i} p_{Wz} [E(y_1|G_i = 1, Z_i = z) - E(y_0|G_i = 1, Z_i = z)] \qquad (2.9) \\
&- \sum_{Z_i} [p_{Mz} - p_{Wz}] E(y_0|G_i = 0, Z_i = z) \qquad (2.10)
\end{aligned}$$

Notice that the sum in equation 2.9 measures the net gender gap, namely the gross gap adjusted by other factors Z_i. In the wage example, it would express the actual wage gap as the difference between the expected female

wages as members of female group and the expected earning of men with similar characteristics but as they would be paid were they members of female group. Conversely, the sum in (2.10), would represents the remaining portion of gross gender gap (Fienberg and Haviland, 2003). In the wage case, it would represent the expected wage of men as belonging to male group. We have thus broken the male expectation in two parts adjusted for confounders. The first gives the expected outcome as they belonged to female group, while the second the expected outcome as members of the actual male group.
Such operation allow, thus, to decompose the gross gender gap in two parts, the first of which providing the net, actual gender gap as adjusted per counfounding characteristics.

In summary, in this chapter we have dealt with measuring gender gap under a causal perspective.
The main question has been how to detect and represent gender discrimination and non-gender discrimination factors affecting the gender gap in a causal graph. Pearl (2009) points out the difficulty arising from cases where an actual experiment is not the case. Gender gap measurement represents one of such cases as the observed variable, gender, is not manipulable, at least in the strict meaning.
The first purpose of this chapter lied in providing an effective graphical language for making concepts as "gender equality" and "gender gap" precise and explicit.
Then, extending the use of the intervention calculus, we proposed a new measure of gender gap under a causal approach. Gender gap, in equation 2.1, has been reinterpreted in terms of the causal effect of gender on a target outcome selected for assessing a certain field of interest; this re-definition allowed to grasp the causal side of the phenomenon.
Moreover, although the factors affecting gender gap and differing by gender but not affected by gender discrimination are difficult to disclose, the methodology proposed in this chapter, based on Pearl's back-door and front-door criteria, allows to define a sufficient set of covariates whose distributions differ by gender but not depend on gender discrimination. It enabled to parti-

tion any difference in outcome by gender. In this way, it has been possible to define a *net* measure of gender gap, in equation 2.5, which allows to "clean" the resulting gender gap from possible counfounders' effects.

The method developed in this chapter facilitates the drawing of quantitative causal inferences from a combination of qualitative assumptions encoded in the graph, and non-experimental observations. The performance of such methodology has been tested on a real data set, with both respects of practical applicability and interpretation capacity, as well as compared with standard methods such as ordinal logistic regression. This is discussed in Chapter 3.

Chapter 3

Causality in Gender Discrimination in China: an application

The aim of this last chapter is to empirically explore both the applicability and the informative potential of the method advanced in Chapter 2.

By means of an application to real data, we show how a causal graph-based approach would represent an effective and innovative statistical tool able to explore and catch gender gap. The application focuses in exploring the existence of a gender discrimination risk in child nutrition and health in China, with particular attention to children and adolescents of age 0-17. First, we briefly describe the Chinese framework, paying the attention on two "alarm bells" which make a deeper gendered analysis necessary:

1. an unbalanced sex ratio at birth; and
2. an excess of female child mortality.

Then, we go in details of the China Health and Nutrition Survey (CHNS) (section 3.2.1), describing the data and illustrating the used method. In order to show the advantages resulting by the causal network approach here proposed and illustrated in the previous chapters, we also conduct a comparison with standard statistical methods. Finally, we estimate the gender gap

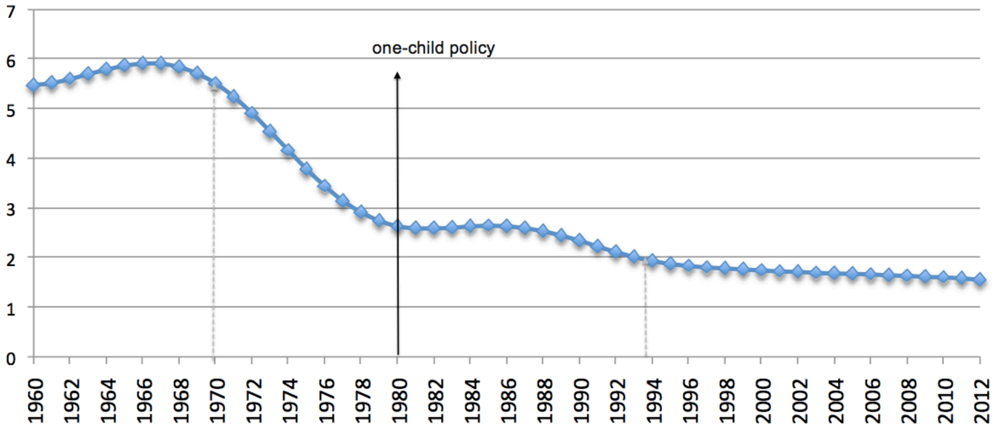


Figure 3.1: China Fertility rate from 1960 to 2012. Source: World Bank, 2012.

in child care in China by applying the methodology illustrated in Chapter 2 such as the measure proposed in section 2.2.3.

3.1 Why China

China's fertility rate, calculated as average births per woman in her lifetime, started its gradual decrease since end-1960s, falling, only over ten years between 1970s and 1980s, from 5 children from less than 3 births per woman. It dropped below replacement level, i.e. 2 births per woman, in the early 1990s and has continued its downward trend ever since, as showed in figure 3.1, recording 2013's total fertility estimate at 1.55 children per woman (Cai, 2008; Guo, 2009). Many factors contributed to such fertility transition, common to many other countries in the world, including Italy; among them: transforming economy, rising education, dropping mortality, changing gender roles including female participation in the labor force etc. But some other argued that China's family planning policies, aiming at controlling the size of population and carried out since 1970s, had played a major role (Wang, 2012).

Such intricate policies can be summarized as a combination of:

- propaganda, attempting to persuade people of the essentiality of family planning to both their own benefits and the national development;
- family planning services, covering free contraceptives and low-cost medical examinations or surgeries;
- birth quota, settings a ceiling on the number of children per married couple (Wang, 2012).

The one-child policy was launched by the Chinese Government in 1979-1980 and decreased the total population by 400 million people compared to the population that the country was predicted to reach without the policy (National Population and Family Planning Commission of China 2008); furthermore, local officers were eager to achieve specific goals in the short term and would even adopt some forceful measures directly acted on the women physically (Short et al., 2000; White 2006).

The dichotomic fertility policy is characterized by an urban-rural duality system, according to which the overwhelming majority of urban residents is subject to strict one-child policy and most people in rural areas is subject to one-and-half policy (Guo, 2002) admitting to have another child if the first is a daughter. In last years, softening of policy and relaxing of requirements for second birth has occurred, although details of regulations varied from province to province.

Anyway it is not possible to attribute China's fertility drop only to the one-child policy (Morgan *et al.*, 2009); indeed throughout the 1980s, when the birth control policy was most enforced, the fertility rate in China was above the replacement level reflecting the difficulties in its implementation, while only from 1990s the rate fell under the replacement level, where still now it is.

Beyond the suspicious of under-reported births, several studies (Chen et al. 2009, Poston 2000, Gu et al. 2007) have showed that fertility variation in China is closely linked to variations in economic and social development: with the rapid modernization and urbanization, the fertility desires are changing (Ma, 2007, 2008; Zheng, 2010). During the past two decades, indeed, the ideal children number of both rural and urban residents has decreased in China; at confirmation of such phenomenon, a recent meta-review of survey of mean ideal family size (MIFS) in China, conducted by Basten (2013), found a MIFS range in urban areas of 1.0-1.5 children per woman, and 1.2 to 1.8 in rural areas.

However, notice that fertility drop is a phenomenon affecting half of the world's population. What is recalling the attention on China are, in fact, two indicators, sentinels of both pre-natal and post-natal gender discrimination against girls: sex ratio at birth and infant mortality.

A skewed sex ratios at birth (SRB), above the "natural", worldwide empirically observed SRB of 103 to 107 male births for 100 females, have been a feature of numerous Asian countries in recent years including Korea, Taiwan, Hong Kong, India and Vietnam (Guilmoto, 2009). However, in the past decades, China has consistently shown the most skewed SRB in the world (Poston and Zhang, 2009), as showed in figure 3.2, reaching the maximum level of 120 male births per 100 female births in 2009. Then, each year since 2009, it has fallen, mainly because of measures adopted by the Chinese government, including the Care for Girls campaign, reaching 118.06 in 2010, 117.78 in 2011 and 117.70 in 2012. Anyway, wide regional differences still occur: from the lowest SRB values equal to 106 - and hence in the normal range - recorded in Tibet and Xinjiang, to the most extreme values of Jiangxi (123) and Anhui (129).

Such skewed SRB may have global consequences in terms of female deficit if we consider that China's population is 22 per cent of the planet total; furthermore it inevitably leads to a generation of exceeding males with unavoidable consequences in terms of lower fertility, more rapid ageing, difficulties in partnership formation and a squeeze on the marriage market (Guilmoto, 2009; Jiang *et al.*, 2013).

Such abnormal SRB in China is possibly due to a complex and interconnected set of drivers (Yi *et al.*, 1993).

Firstly, the patriarchal culture, according to which men take more responsibilities than women, as they have to afford economic and endowment support, carry on the family line, bring honour and authority, care parents in old age

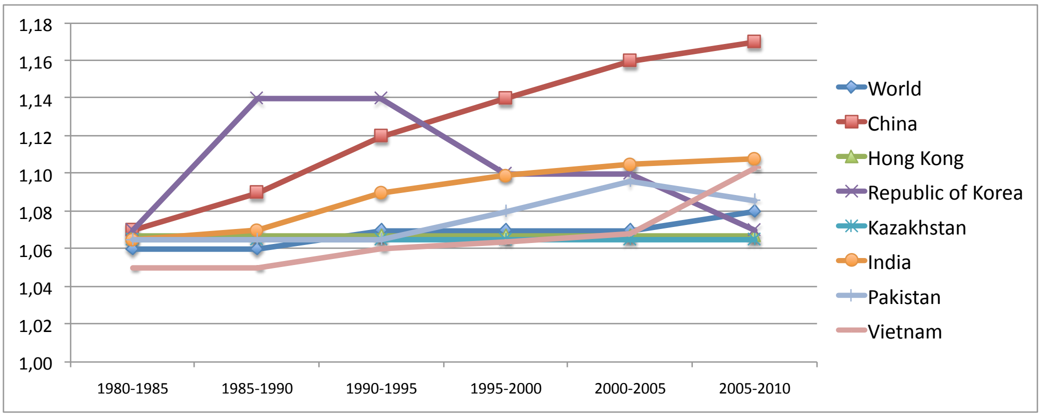
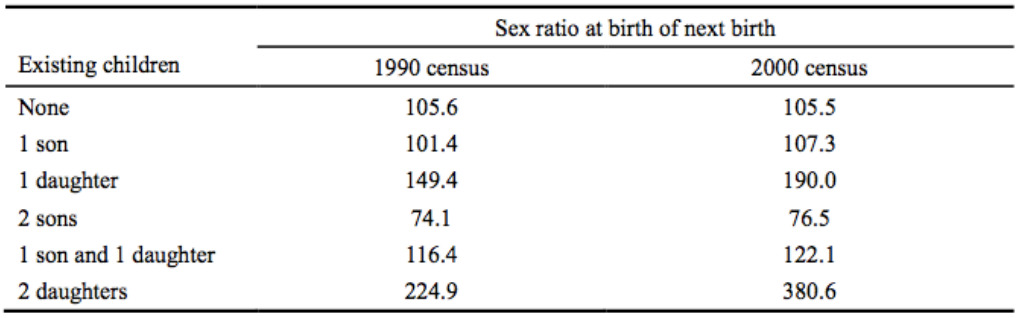


Figure 3.2: Sex Ratio at Birth trends in Asian countries. Source: UNDP, 2012.

and so on (Gupta et al., 2004; Attane, 2005). In this context, the strong traditional ideology of son preference is still prevalent in China and widely recognised in literature (Poston and Zhang, 2009); Yuan and Shi (2005) have, among others, pointed out how in Asian countries, no matter what the level of economic and social development is, the preference for boys over girls is widely accepted.

Even the rapid fertility decline is strictly linked to high SRB; in a low fertility context and in an ideal of small family sizes (≤ 2 children), some couples seek assistance from sex-selection technology to meet their target with fewer births. One-child policy and son preference as well, interfere in such relationship occurring between lower fertility and sex-selection use: indeed the son preference thought "if the child has to be one, at least that is son" leads families to resort to sex-selection to guarantee that this only child is male.

Another crucial elements, understudied within the SRB literature, is the influence of parity and sex composition of children already born (Poston and Zhang, 2009; Jiang *et al.*, 2013); the composition of children's gender is clearly linked to sex-selection technologies usage and, consequently, to SRB (Gupta, 2005) in the following way. Table 3.1 shows how sex composition of existing children would influence the gender of next birth: for example, the sex ratio at parity two, ranges from 107.3 in cases where the first child is a boy through to 190.0 when the first child is a girl, demonstrating that



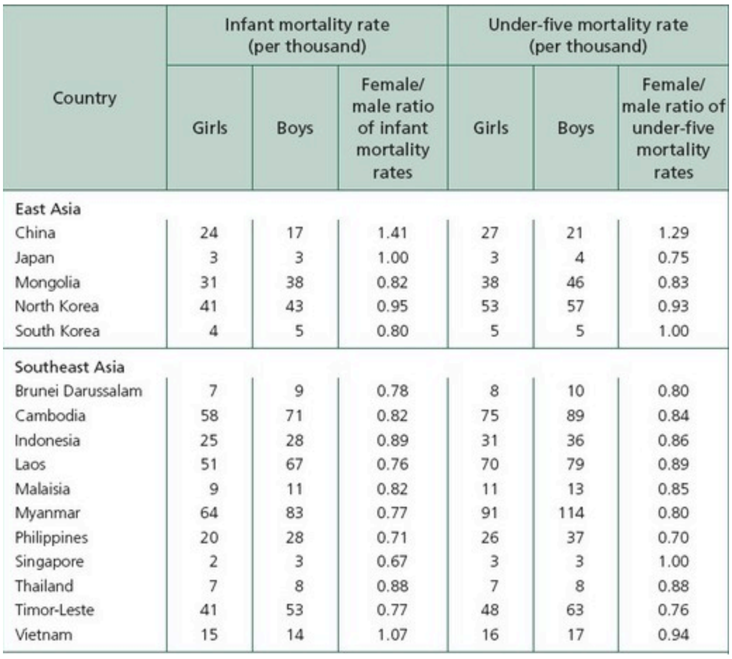
	Sex ratio at birth of next birth	
Existing children	1990 census	2000 census
None	105.6	105.5
1 son	101.4	107.3
1 daughter	149.4	190.0
2 sons	74.1	76.5
1 son and 1 daughter	116.4	122.1
2 daughters	224.9	380.6

Table 3.1: Sex ratios at birth by sex composition of existing children. Source: 1990 census data from Zeng et al. (1993), 2000 census data from Sun (2005).

in two-child ideal, if the first child is female, there is a strong incentive to ensure that the second child is male.

Recent studies (Guo, 2007; Yang, 2012) have confirmed how macro-level elements, such as accessibility of prenatal sex determination, sex-selection technology and economic development as well as individual children composition have an impact on the gender of next birth.

The second "alarm bell", involving, conversely, post-natal discrimination in China, is an observed excess of female infant (child) mortality, EFIM (EFCM) defined as the higher than "normal" ratio of female/male infant (child) mortality in a population (Li *et al.*, 2004).
Infant and child mortality is usually determined by a complex interplay of biomedical, socio-economic, demographic and environmental factors that impact mortality at different stages in the life course (Mosley and Chen, 1984).
In absence of sex discrimination, biomedical factors are the major determinants of gender differences, and both infant and child mortality is higher for males than for females (Coale, 1991); such difference would compensate the "natural" excess of males at birth and hence would ensure a ratio closer to 100 in the important years for reproduction (Banister, 2004).
Anyway in China, as showed in figure 3.3, a worrying observed excess of both infant and child female mortality is a signal of a girl child survival problem, reflection of an insufficient investment in girls and of unequal rights in the early stage of human life.



Country	Infant mortality rate (per thousand)			Under-five mortality rate (per thousand)		
	Girls	Boys	Female/ male ratio of infant mortality rates	Girls	Boys	Female/ male ratio of under-five mortality rates
East Asia						
China	24	17	1.41	27	21	1.29
Japan	3	3	1.00	3	4	0.75
Mongolia	31	38	0.82	38	46	0.83
North Korea	41	43	0.95	53	57	0.93
South Korea	4	5	0.80	5	5	1.00
Southeast Asia						
Brunei Darussalam	7	9	0.78	8	10	0.80
Cambodia	58	71	0.82	75	89	0.84
Indonesia	25	28	0.89	31	36	0.86
Laos	51	67	0.76	70	79	0.89
Malaisia	9	11	0.82	11	13	0.85
Myanmar	64	83	0.77	91	114	0.80
Philippines	20	28	0.71	26	37	0.70
Singapore	2	3	0.67	3	3	1.00
Thailand	7	8	0.88	7	8	0.88
Timor-Leste	41	53	0.77	48	63	0.76
Vietnam	15	14	1.07	16	17	0.94

Figure 3.3: Infant, under-five mortality rate and EFIM in Asia. Source: WHO, Statistical Information System, 2008

The main reasons for the excess of female mortality, among children and infants, is in fact the inequality in health care between sons and daughters (Li *et al.*, 2013); indeed the rooted son preference in China translates in discrimination against girls regarding nutrition, health care, and thus causing EFCM. Studies on child mortality (Li *et al.*, 2004) have confirmed that medical treatment for boys is significantly better than for girls , in addition female infanticide still exists (Li *et al.*, 2004).

These considerations, regarding differences in the scale of the skewed SRB, together with an evident excess of female child mortality in China, suggest that there is something special about China that requires to be highlighted and analysed in its own right.
As reported before, a number of studies have argued that this key differential is due to the family planning policy which, in the end, limits the number of children that couples are legally entitled to have. However, it is important to notice that such policy interacts with a wide array of other factors which

shapes childbearing and, ultimately, affects the SRB and EFCM: if couples have had sons, fertility policies might have little effect on their choices; conversely, to those who only have daughters, fertility policies affect them markedly.

We can say, hence, that in the context of low fertility levels and corresponding fertility intention, as well as sex-selection availability, son preference manifests itself more strongly.

It is, thus, necessary to develop feasible statistical methods as well as political strategies able to detect and fight gender discrimination, a phenomenon that, specially in China, is rooted and hidden in its history, culture and traditions.

3.2 Data and a Standard Statistical Approach

3.2.1 Description of the dataset

Data are collected in China Health and Nutrition Survey (CHNS) 2009 and they are available at $http://www.cpc.unc.edu/projects/china$; the project is the result of a collaborations between the Carolina Population Center at the University of North Carolina at Chapel Hill and the National Institute of Nutrition and Food Safety at the Chinese Center for Disease Control and Prevention. It represents a multi-purpose longitudinal survey (rounds in 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, 2011 and soon 2013) covering key public health risk factors, health outcomes, as well as demographic, social and economic factors in depth at the individual, household and community levels. Its aim is to examine the effects of health, nutrition, and family planning policies and to see how the social and economic transformation of Chinese society is affecting the health and nutritional status of its population.

The survey is conducted over a three-day period using a multi-stage, random cluster process in order to draw a sample of about 4,400 households with a total of 19,644 individuals covering 9 provinces with substantial variations in geography, economic development, public resources, and health indicators. In addition, detailed community data are collected in surveys of food markets, health facilities, family planning officials, and other social services and community leaders.

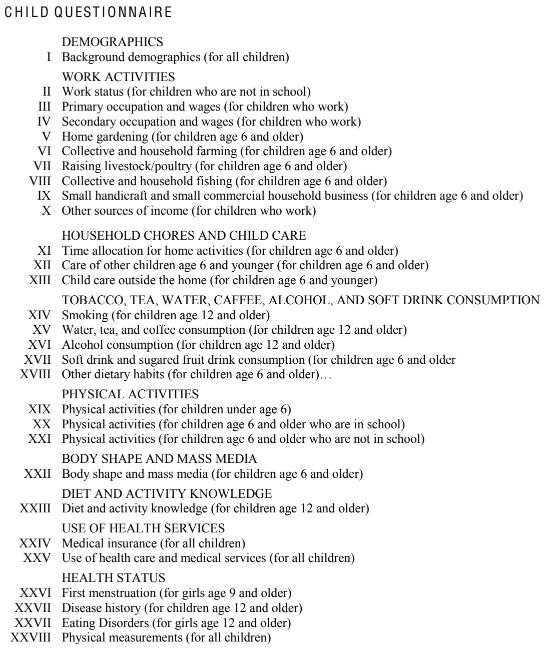
We focus on the 2009 household survey, the most recent available at the beginning of this thesis. Our analysis's unit is any child between 0-17 years collected in the survey, while our purpose is to infer if a gender discrimination in child health, nutrition and care occurs in China.
A preliminary stage of profound editing and cleaning of the data set was necessary in order to create a tailored dataset for the specific purpose of the present research. In particular, in order to collect in a single file all information needed for the analysis, the first step consisted in jointing 17 different datasets concerning child's demographic, social, economical, nutritional and health data resulting by the child's questionnaire (table 3.2). The second operation aimed to match, for every unit of analysis, key information about child's mother and father, data extracted by other 20 datasets collected through the adult's survey (table 3.3). Referring to child's household and individual codes, it has been possible to combine all these data; such procedures have been automatized with *if* and *for* loops in Matlab (Higham and Higham, 2005), in order to ensure fast and accurate matches.

In the dataset we collected information of different nature: demographic such as gender, age, residence province etc., have been included in the final dataset as they are; economic, educational and social variables have been aggregated in synthetic indicators while others have been built on purpose to catch specific phenomena. A summary prospect of the considered variables is showed in table 3.14.

In particular, the variables can be classified in the following macro-areas:

- *Demographic*: child's age, child's gender, urban/rural registration, parents' age etc.

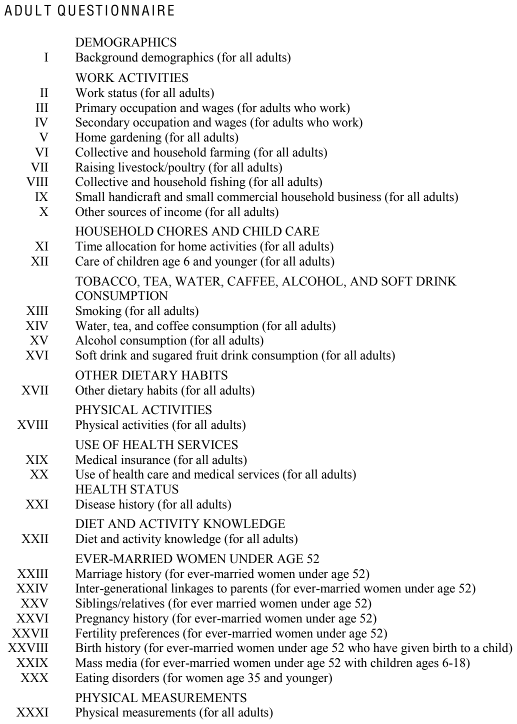
- *Time and Gender of Births related*: child's birth order, gender sequence of births and birth spacing.
  They represent crucial indicators to detect latent factors such as son preference and sex-selection; indeed a child with certain sex-birth-order characteristics will influence the next child's gender in order to attain



CHILD QUESTIONNAIRE

	DEMOGRAPHICS
I	Background demographics (for all children)
	WORK ACTIVITIES
II	Work status (for children who are not in school)
III	Primary occupation and wages (for children who work)
IV	Secondary occupation and wages (for children who work)
V	Home gardening (for children age 6 and older)
VI	Collective and household farming (for children age 6 and older)
VII	Raising livestock/poultry (for children age 6 and older)
VIII	Collective and household fishing (for children age 6 and older)
IX	Small handicraft and small commercial household business (for children age 6 and older)
X	Other sources of income (for children who work)
	HOUSEHOLD CHORES AND CHILD CARE
XI	Time allocation for home activities (for children age 6 and older)
XII	Care of other children age 6 and younger (for children age 6 and older)
XIII	Child care outside the home (for children age 6 and younger)
	TOBACCO, TEA, WATER, CAFFEE, ALCOHOL, AND SOFT DRINK CONSUMPTION
XIV	Smoking (for children age 12 and older)
XV	Water, tea, and coffee consumption (for children age 12 and older)
XVI	Alcohol consumption (for children age 12 and older)
XVII	Soft drink and sugared fruit drink consumption (for children age 6 and older
XVIII	Other dietary habits (for children age 6 and older)…
	PHYSICAL ACTIVITIES
XIX	Physical activities (for children under age 6)
XX	Physical activities (for children age 6 and older who are in school)
XXI	Physical activities (for children age 6 and older who are not in school)
	BODY SHAPE AND MASS MEDIA
XXII	Body shape and mass media (for children age 6 and older)
	DIET AND ACTIVITY KNOWLEDGE
XXIII	Diet and activity knowledge (for children age 12 and older)
	USE OF HEALTH SERVICES
XXIV	Medical insurance (for all children)
XXV	Use of health care and medical services (for all children)
	HEALTH STATUS
XXVI	First menstruation (for girls age 9 and older)
XXVII	Disease history (for children age 12 and older)
XXVII	Eating Disorders (for girls age 12 and older)
XXVIII	Physical measurements (for all children)

Table 3.2: Table of contents of child questionnaire



ADULT QUESTIONNAIRE

	DEMOGRAPHICS
I	Background demographics (for all adults)
	WORK ACTIVITIES
II	Work status (for all adults)
III	Primary occupation and wages (for adults who work)
IV	Secondary occupation and wages (for adults who work)
V	Home gardening (for all adults)
VI	Collective and household farming (for all adults)
VII	Raising livestock/poultry (for all adults)
VIII	Collective and household fishing (for all adults)
IX	Small handicraft and small commercial household business (for all adults)
X	Other sources of income (for all adults)
	HOUSEHOLD CHORES AND CHILD CARE
XI	Time allocation for home activities (for all adults)
XII	Care of children age 6 and younger (for all adults)
	TOBACCO, TEA, WATER, CAFFEE, ALCOHOL, AND SOFT DRINK CONSUMPTION
XIII	Smoking (for all adults)
XIV	Water, tea, and coffee consumption (for all adults)
XV	Alcohol consumption (for all adults)
XVI	Soft drink and sugared fruit drink consumption (for all adults)
	OTHER DIETARY HABITS
XVII	Other dietary habits (for all adults)
	PHYSICAL ACTIVITIES
XVIII	Physical activities (for all adults)
	USE OF HEALTH SERVICES
XIX	Medical insurance (for all adults)
XX	Use of health care and medical services (for all adults)
	HEALTH STATUS
XXI	Disease history (for all adults)
	DIET AND ACTIVITY KNOWLEDGE
XXII	Diet and activity knowledge (for all adults)
	EVER-MARRIED WOMEN UNDER AGE 52
XXIII	Marriage history (for ever-married women under age 52)
XXIV	Inter-generational linkages to parents (for ever-married women under age 52)
XXV	Siblings/relatives (for ever married women under age 52)
XXVI	Pregnancy history (for ever-married women under age 52)
XXVII	Fertility preferences (for ever-married women under age 52)
XXVIII	Birth history (for ever-married women under age 52 who have given birth to a child)
XXIX	Mass media (for ever-married women under age 52 with children ages 6-18)
XXX	Eating disorders (for women age 35 and younger)
	PHYSICAL MEASUREMENTS
XXXI	Physical measurements (for all adults)

Table 3.3: Table of contents of adult questionnaire

MACRO-AREA	NAME VARIABLE	DESCRIPTION	VALUES	MISSING
Demographic	URBAN/RURAL	*Type of household registartion*	1=URBAN SITE 2=RURAL SITE	0%
	GENDER	*Child's gender*	1=MALE; 2= FEMALE	0%
	AGE_C	*Child's age*		0%
	AGE_F	*Father's age*		3,50%
	AGE_M	*Mother's age*		0,60%
Time & Gender Births	BIRTH_ORDER	*Child's birth order; gained from ages of siblings (even >18)*	0= ONLY CHILD; 1=FIRST CHILD; 2=SECOND CHILD; 3= THIRD OR LOWER CHILD	0%
	GENDER_SEQUENCE	*Sequence of gender's siblings*	1=ONLY MALE; 2= ONLY GIRL; 11= ONLY MALES; 22= ONLY GIRLS; 12= AT LEAST 1 MALE BEFORE 1 OR MORE GIRLS; 21= AT LEAST 1 GIRL BEFORE 1 OR MORE MALES	0%
	TIME_DISTANCE	*Time in months before previous birth; gained from previous birth date*	0= ONLY OR FIRST CHILD	0,07%
	AVER_TIME	*Average time from previuos birth; average calculated among all observations*	1=MORE TIME THAN AVERAGE; 0= ON AVERAGE; -1= LESS TIME THAN AVERAGE	0,07%
Medical	MED.INS.CHILD	*If child has a medical insurance or not*	0=NO 1=YES	0%
	MED.INS_F	*If father has a medical insurance or not*	0=NO 1=YES	16,40%
	MED.INS_M	*If mother has a medical insurance or not*	0=NO 1=YES	9%
	BMI_C	*Child's BMI cut-offs; calculated as weight/(height)^2*	1=UNDERWEIGHT; 2=NORMAL; 3=OVERWEIGHT; 4=OBESE	0%
	BMI_F	*Father's BMI cut-offs; calculated as weight/(height)^2*	1=UNDERWEIGHT; 2=NORMAL; 3=OVERWEIGHT; 4=OBESE	24,10%
	BMI_M	*Mother's BMI cut-offs; calculated as weight/(height)^2*	1=UNDERWEIGHT; 2=NORMAL; 3=OVERWEIGHT; 4=OBESE	12,40%
Educational & Working	EDU_P	*Highest level of formal completed education between parents*	0=NO EDUCATION; 1= PRIMARY SCHOOL; 2=MIDDLE SCHOOL; 3=COLLEGE/UNIVERSITY; 9=UNKNOWN	0%
	JOB_POS_P	*Highest job position between parents*	1=OWNER MANAGER; 2=INDEPENDENT OPERATOR; 3=PERMANENT EMPLOYEE; 4=CONTRACTOR WITH OTHER PERSON OR ENTERPRISE; 5= TEMPORARY WORKER; 6=PAID FAMILY WORKER; 7=UNPAID FAMILY WORKER; 8=OTHER; 9=UNKNOWN; 20=NOT WORKING	0%
Child's grewing up context	DIETBEHAV_F	*If father has a good knowledge of healthy diet behaviours*	2=VERY GOOD KNOWLEDGE; 1=GOOD KNOWL; 0=NEUTRAL; -1=BAD KNOWL; -2=VERY BAD KNOWL	21,90%
	DIETBEHAV_M	*If mother has a good knowledge of healthy diet behaviours*	2=VERY GOOD KNOWLEDGE; 1=GOOD KNOWL; 0=NEUTRAL; -1=BAD KNOWL; -2=VERY BAD KNOWL	15,80%
	PRIOR_F	*Father's priority*	1=CHILD'S HEALTH; 0=SAME IMPORTANCE; -1=MONEY	16,60%
	PRIOR_M	*Mother's priority*	1=CHILD'S HEALTH; 0=SAME IMPORTANCE; -1=MONEY	9,10%
	POT_HELP	*If a potential help from grandparents*	0=NO 1=YES	12,50%
Parents' primary socialization	BS_F	*Father's siblings composition; to consider in which gender context the father has grown up*	0=ONLY CHILD; 1= MALE CULTURE; 2= FEMALE CULTURE; 3=MIX CULTURE	11,50%
	BS_M	*Mother's siblings composition; to consider in which gender context the mother has grown up*	0=ONLY CHILD; 1= MALE CULTURE; 2= FEMALE CULTURE; 3=MIX CULTURE	10%
Fertility intentions	ADD_CHILD	*If mother wants one or more additional child*	0=NO; 1=YES, ONE; 2=YES, MORE THAN ONE	12,30%

Table 3.4: Prospect of the variables considered in the analysis.

the desired sex composition. In general, more than one or two daughters are culturally not welcome, and girls who are born into a family that already has daughters, are the most likely to be least valued and thus discriminated by the household. A desire for sons, combined with a growing desire for a small family thus suggests that selective discrimination against higher-birth-order girls is likely to occur (Das Gupta 1989).

Even timing between children is an important signal: indeed each abortion increases the spacing between births as the uterus needs at least two menstrual cycles to recover before conception.

- *Medical*: if child and parents have any medical insurance, child's and parents' Body Mass Index (or BMI, calculated as $weight/height^2$) cut-offs.

  In particular, "child's BMI" constitutes a key variable for our research, representing the target outcome for measuring the effect of a potential gender gap on child's nutrition. Indeed both weight and height are sensitive measures of long-term health and nutrition in childhood, reflecting the intra-household resource allocation, the exposure to infectious diseases as well as the access to medical facilities. It is important to underline that, for our research purposes, both undernourishment than obesity constitute indicators of poor childcare and lack of attention to the child's well-being.

  According to WHO (2004) recommendations, it is necessary to consider in the analysis BMI cut-off points targeted for Chinese population, in order to avoid bias due to biological and anthropomorphic differences. For this reason, we referred to appropriate cut-off points proposed by the Working Group on Obesity in China (WGOC) (Zhou, 2002; Jiang *et al.*, 2006) and widely recognized as standards for Chinese BMI; anyway for children, since an international agreement has not been reached yet, we resorted to BMI cut-off points, as used in Jiang *et al.* (2006)'s and Shang *et al.* (2005)'s studies. In tables 3.5 the international, namely the WHO international standard, and chi-

international (WHO)		***Chinese (WGOC)***	
< 18.5	*underweight*	<18.5	*underweight*
18.5-24.99	*normal*	18.5-23.9	*normal*
25-29.9	*overweight*	24-27.9	*overweight*
>30	*obese*	>28	*obese*

Table 3.5: International and Chinese BMI cut-offs for adults. Source: WHO, 2012 and WGOC, 2002.

nese adults' cut-off points are compared. Child's chinese cut-off points, evaluated both per sex and age, are reported in table 3.6.

- *Educational and Working*: highest level of completed education and highest job position between mother and father.
  We are, indeed, interested in verifying if in more educated and well-off families a more gender balanced child's nutrition occurs or not.

- *Child's growing up environment*: parents' knowledge about good diet behaviours, parents' priorities and potential grandparents help in caring child.
  The variables concerning the parents' diet knowledge are composite indicators, resulting by the aggregation of several answers about healthy and unhealthy eating behaviours as well as food preferences (questions from U376 to U393 in the adult's questionnaire). The indicators have been built up by attributing a score to each question: +1 in case of correct eating behaviours, 0 to neutral answers, and -1 for unhealthy behaviours; then the resulting sum of each score has been rescaled in the range from -2 to +2 giving the final indicator.
  "Parents' priorities" is as well as a synthetic indicator, given by the aggregation of 5 questions regarding the importance attributed by child's parent respectively to good income and child's wealth (questions U405-U409). Such indicator is born with the aim of evaluating the collocation of child's wealth within the family values.

	normal		***overweight***		***obese***	
age	boys	girls	boys	girls	boys	girls
0	12,5	12,3	14,9	14,7	16,2	16,3
1	14,1	13,8	20,2	19,5	21,9	19,6
2	14	13,7	18,3	17,7	19,8	19,1
3	13,8	13,4	17,6	17,3	18,9	18,8
4	13,4	13	16,7	17,2	18,1	18,9
5	13	12,7	16,8	17,4	18,6	19,4
6	12,6	12,6	17,2	17,8	19,4	20,3
7	12,5	12,7	17,8	18,3	20,3	21,2
8	12,6	12,8	18,4	18,9	21,3	22,2
9	12,8	13	19	19,5	22,2	23,2
10	13,2	13,3	19,7	20,3	23,1	24,2
11	13,5	13,6	20,3	21,1	23,9	25,2
12	13,9	14,1	20,9	21,8	24,7	26,1
13	14,2	14,8	21,5	22,5	25,4	26,8
14	14,7	15,3	22	23	26	27,4
15	15,2	15,8	22,5	23,4	26,5	27,7
16	15,6	16,3	23,1	23,7	27,1	27,9
17	15,9	16,6	23,5	23,9	27,6	28

Table 3.6: Children' lower bound cut-offs. Source: WGOC, 2002 for overweight and obesity; Shang *et al.* (2005) for normal cut-off.

Finally, the "potential help" is a composite indicator involving the inter-generational linkages to grandparents. It results from the combination of two data about each grandparent: if he/she lives close or far from the family and if he/she needs help in daily life. From such information it is thus possible to deduce if they represent a potential help in childcare; indeed in the Chinese labour market, characterized by limited flexibility in work arrangements such as part-time, lack of daycare facilities such as parental leaves etc. and high migration to the workplace, the proximity and involvement of grandparents in child's assistance are essential factors for understanding child's wealth and care status (Chen *et al.*, 2000).

- *Parents' primary socialization*: father' and mother' siblings gender composition. These two variables have the purpose to explore the maternal/paternal primary socialization respecting the gender discrimination; in fact if the father/mother has grown up in a son-preference context, he/she would more probably tend to replicate or allow for gender discrimination. Conversely if parents are born in a free from gender discrimination context, they would more likely avoid differences between sons and daughters.

- *Fertility intentions*: number of eventual additional desired children. Such variable allowed us to insert in the research the fertility preferences, which, as we have already illustrated in section 3.1, have a strong impact on the risk of gender discrimination.

Some observations had to been deleted by the final file as the corresponding child's BMI was missing; it, indeed, embodies the essential information for our purposes of analysis since it represents the target outcome to establish if a gender discrimination in child's nutrition would exist. Hence the resulting dataset contains 1313 children from 0 to 17 years for which we have complete information about demographic, parents' educational and job variables; for what concerns, instead, the other variables, some missing entries may occur. The corresponding percentages of missing cases for each variable, are showed in the last column of table 3.14.

3.2.2 Descriptive analysis

In this subsection, we present some descriptive statistics in order to summarize the data.

Boys represent the 55% of the sample, while girls the 45%; 28% of children live in urban areas, whilst the remaining 72% in rural ones.
Only child represent the 46% of population while in 48,4% of cases we are in presence of families composed by 2 children and only in 5.6% by 3 or more, showing the effects of one-child policy on fertility choices.
For what concern gender siblings composition, excluding the only child, firstborn girl and younger brother's sequence represents the 24% of cases, twice compared to the opposite sequence (first son and second girl); a data which might hide the research of a son, after a daughter, even by sex-selection.
Exploring the BMIs, underweight children are the 4%, while the overweight the 10.7% and the obese the 6.8%; for what concerns parents, the undernourishment percentage is 3.8% for fathers and 6.2% for mothers while the firsts

are overweight or obese in 31.6% of cases and the seconds in 27%.

In 83% of cases, child has a medical insurance, while father in 77% and mother in 81%.

The 88% of parents have a primary or middle-school education and higher in only 8%; half are working as independent operators, others 20% as employees and 7% of them are unemployed.

Almost all parents have a good or very good knowledge of diet behaviours. For both fathers and mothers their priority is good incomes in about 20% of cases; conversely, child's health has precedence for 12% among fathers and 15% among mothers.

The 80% of households can rely on a potential help of grandparents for child-care.

Approximatively 50% of parents come from a family composed by both brothers and sisters while in about 20% by male sibling only; a household dominated by women has represented the growth environment for 20% of mothers and 12% of fathers.

Even the 78% of mothers do not want another child, the 7% just one more and only 2.5% more than one.

Exploring the distribution of child's BMI within gender groups, it clearly emerges how differences among girls and boys occur; indeed the number of boys belonging to normal, overweight and either obese categories overcomes the number of girls, as showed in picture 3.4, while within the category of underweights, girls represent the majority.

Furthermore, analysing the child's BMI in relation to the gender sequence of the child's siblings, such disparities result even more evident, as shown by the comparison in figure 3.5. Excluding only child for which significant gender differences do not occur, in general boys tend to be less underweight than girls; moreover they are overweight or even obese more frequently when they have an older sibling, regardless their gender.

Conversely, girls suffer of undernourishment specially when they have a younger brother and they tend to be overweight, in most of cases, when they are the oldest, independently by the gender of younger siblings. For

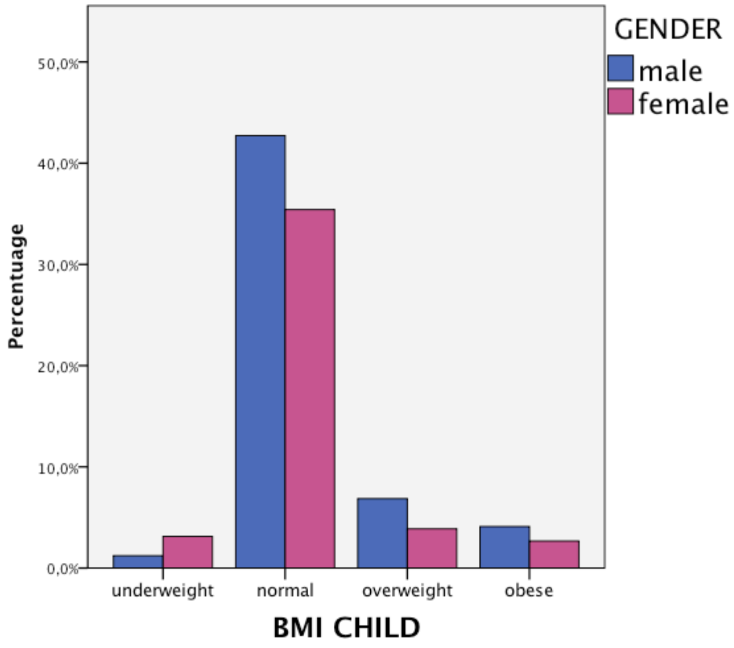


Figure 3.4: Child's BMI per gender.

what concerns, finally, girls' obesity, it would not seem significantly varying with siblings' gender composition.

Moreover, from table 3.7, we calculated the relative female/male ratio for each gender sibling composition, in order to evaluate the gap between boys and girls. For what concerns the undernourishment F/M ratio, it is close to 1 for only child and MF gender composition (namely when there is at least one boy before a girl), showing a gender balance. Conversely when a girl has a younger brother, she risks 6.3 times more than brother to be underweight;

		GENDER							
		M				F			
		SEQUENCE				SEQUENCE			
		M	MM	MF	FM	F	MF	FM	FF
CHILD BMI	underweight	9	1	3	3	12	3	19	7
	normal	261	104	59	137	188	69	124	84
	overweight	53	14	8	15	30	0	10	11
	obese	33	8	1	12	18	6	6	5

Table 3.7: Frequency of child's BMI per gender sequence of siblings.

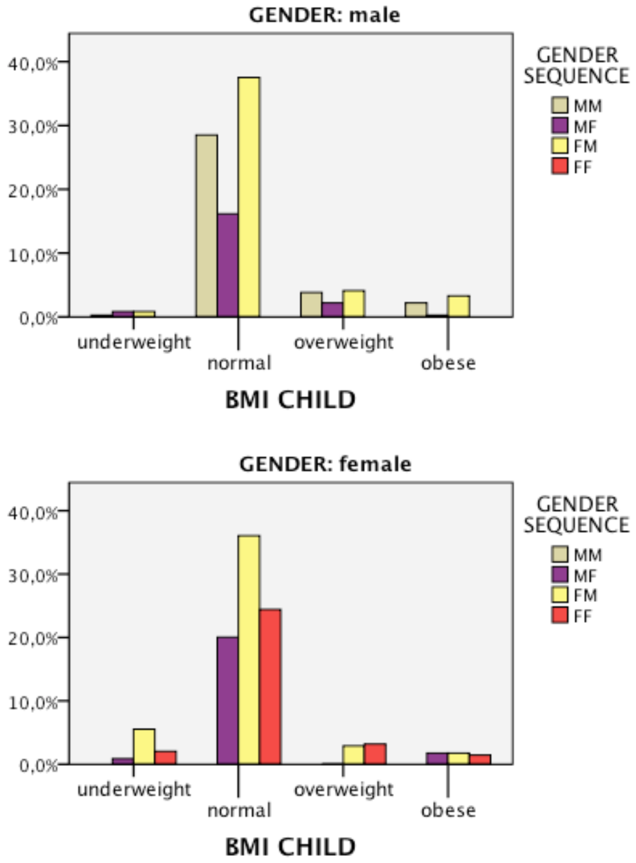


Figure 3.5: Boys' and girls' BMI per gender composition of siblings

		GENDER							
		M				**F**			
		BIRTH'S ORDER				*BIRTH'S ORDER*			
		0	***1***	***2***	***3 +***	***0***	***1***	***2***	***3 +***
	underweight	9	3	4	0	12	24	5	0
BMI	*normal*	261	103	161	36	188	136	119	22
CHILD	*overweight*	53	10	21	6	30	9	9	3
	obese	33	0	15	6	18	4	12	1

Table 3.8: Frequency of child's BMI per birth order.

the gap, then, arises in cases of same sex siblings, as the risk of being underweight for daughters is 7 times more than for sons. Looking at obesity, it affects, in general, more boys than girls, specially only boys and boys with elder sister; anyway, the gap is reversed when the family is composed by a son and a younger sister, as the girl has a risk of being obese 6 times more than her brother.
In general, there are no significant gender gaps existing among only child; things change when there are siblings. From these considerations, clearly emerges how sibling's gender is a crucial factor to understand nutrition pattern; having a brother implies for a daughter twice the risk of being malnourished, in sense of both lack than excess of nutrition, compared with sons.

Same considerations are valid for the interaction between child's BMI and birth order as reported in table 3.8; the risk of being undernourished, being the first-born child, is 8 times higher for girls than for boys while it decreases if the girl is a second or later child. Moreover, first-born girls have even a higher risk of being obese, confirming the previous results about gender sequence of siblings. For what concern the second order births, the gender disparities weaken, even if a higher risk of being overweight persists for second-born boys; also for higher birth's orders a tendency through excess of nutrition is addressed to sons.

In table 3.9 are reported some standard tests of association for categorical data with $\alpha = 0.05$; when at least one of two variables is nominal, we have calculated the Cramer's V, defined as:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where χ^2 is derived by Pearson's Chi-squared test, N is the number of observations and k the smallest number of rows or columns; it ranges between 0 and 1, where 0 represents no association while 1 complete association. For ordinal characters, filled in orange in table 3.9, we calculated the Gamma coefficient, which takes values in the range $[-1;+1]$ and provides a measure of both strength and direction of association. It finds its formulation as:

$$G = \frac{N_c - N_d}{N_c + N_d}$$

where N_c represents the number of concordant pairs, namely the number of pairs of cases ranked in the same order on both variables, while N_d the number of discordant pairs or even number of pairs of cases ranked differently on the variables.

In blue are highlighted the measures of associations between the pairs of variables "Gender" and "Child's BMI", and "Gender Sequence of siblings" and "Child's BMI"; they show a relationship, although weak nevertheless indicating, thus, that a gender discrimination risk occurs, since in presence of gender equality the two variables should be independent. Moreover, from such measures would result that child's BMI would be positively related, at an α level of 0.05, to parents' BMIs, parents' education and, to a lesser extent, to child's gender and child's medical insurance; conversely it would be negatively affected by parents' age.

Finally, notice that "Gender" is associated only with "Birth Order", "Gender Sequence" and "Child's BMI", confirming that understanding these factors' interactions, represents the challenge but even the crucial point of the research.

Cramer's V/ Gamma coeff.	URB/RUR	GENDER	BIRTH_ORD	GENDER_SEQ	AVER_TIME	MED.INS.C	BMI_C	EDU_P	ADD_CHILD
URB/RUR	-	-	-	0,18	0,2	0,07	-	-	-
GENDER	-	-	-	0,8	-	-	0,13	-	-
AGE_F	0,15	-	0,16	**0,32**	**0,33**	0,2	-0,2	-0,1	**-0,35**
AGE_M	0,11	-	**0,33**	**0,28**	**0,32**	0,18	-0,18	-	**-0,44**
BIRTH_ORDER	**0,18**	0,16	-	-	-	-		-	**-0,56**
GENDER_SEQUENCE	**0,19**	**0,8**	-	0,15	-	0,1	0,16	-	0,25
AVER_TIME	**0,2**	-	-	-	-	-	-	-	-
MED.INS.CHILD	0,07	-	-	0,1	-	-	0,11	-	-
MED.INS_F	0,09	-	0,09	-	-	0,3	-	-	-
MED.INS_M	0,09	-	0,1	0,12	-	0,27	-	-	-
BMI_C	0,13	***0,13***	-	***0,16***	-	0,11	-	-	-
BMI_F	-	-	-0,09	0,18	0,26	-	**0,35**	0,24	-
BMI_M	-	-	0,19	0,2	0,24	-	**0,27**	-0,27	-0,2
EDU_P	0,19	-	**-0,36**	0,3	**0,3**	0,09	0,26	-	0,28
JOB_POS_P	**0,29**	-	-0,17	**0,35**	-	0,16	-	-	-
DIETBEHAV_F	-	-	-	-	-	-	-	-	-
DIETBEHAV_M	-	-	-	0,17	-	-	-	-	-
PRIOR_F	-	-	-	0,18	-	-	-	0,2	-
PRIOR_M	-	-	-	0,13	-	-	-	-	-
POT_HELP	-	-	-	0,1	-	-	-	0,11	0,11
BS_F	-	-	0,13	0,25	-	0,14	0,09	-	0,12
BS_M	0,11	-	0,14	0,26	-	0,1	0,07	0,18	-
ADD_CHILD	-	-	**-0,56**	0,25	-	-	-	0,28	-

Table 3.9: Measures of associations among variables.

3.2.3 A standard model: ordinal logistic regression

In this subsection we are going to show the results provided by a standard statistical model: the ordinal regression, which is a generalized linear model specially tailored for the case of predicting ordinal variables.

The model is based on the assumption that there is a latent continuous outcome variable and that the observed ordinal outcome arises from discretizing the underlying continuum into j-ordered groups (Agresti, 2002).
We use the *logit* link function leading to the ordinal logistic regression; our ordinal response variable Y is child's BMI, having j=4 categories and hence values 1,...,4 while independent variables are both ordinal and categorical. Thus the resulting odds are defined as:

$$\theta_1 = \frac{P(Y \leq 1)}{1 - P(Y \leq 1)}; \qquad \theta_2 = \frac{P(Y \leq 2)}{1 - P(Y \leq 2)}; \qquad \theta_3 = \frac{P(Y \leq 3)}{1 - P(Y \leq 3)}$$

The fourth category $Y = 4$ i.e. obese, has score 1 for being the reference category, as usually assumed in literature.
The cumulative logistic model is given by:

$$logit(Y \leq j) = \ln(\theta_j) = \alpha_j + \boldsymbol{\beta}\mathbf{X} \qquad j = 1, 2, 3 \tag{3.1}$$

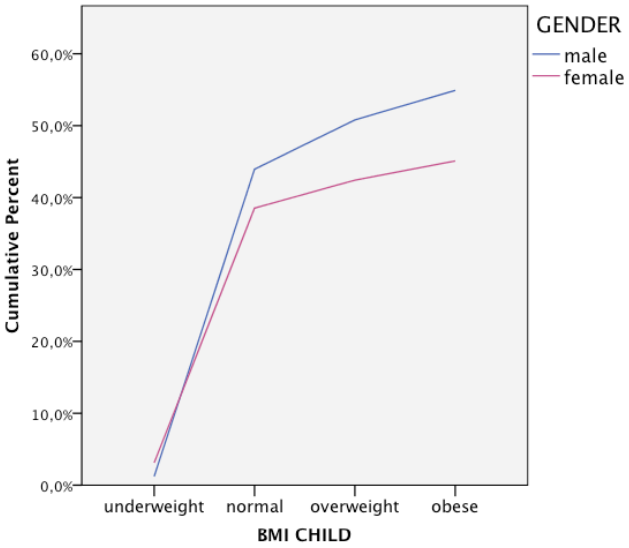


Figure 3.6: Plot of observed cumulative percentages for variable "Gender"

where α_j are the intercepts and $\boldsymbol{\beta}$ the vector of regression coefficients both depending on the category j. As we have only one common parameter β for each covariate, the cumulative odds result as:

$$\theta_j = e^{\alpha_j + \beta \mathbf{X}} \qquad j = 1, 2, 3 \tag{3.2}$$

meaning that the 3 odds for each cut-off category differ only with regard to the intercept α_j; that is why the model is called even *odds proportional* (Agresti, 2002; O'Connell, 2006).
Both the stringent proportional odds assumption, that the no multicollinearity assumption require to be tested.

In figure 3.6 we have examined the cumulative percentage plot of the child's BMI with separate curves for boys and girls. Consider category $Y = 1$, i.e."underweight": a larger percentage of girls than boys belongs to this category; as additional percentages are added (the cumulative percentage for "normal" and "overweight") the opposite occurs, and the cumulative percentages for boys become more and more large as BMI increases.

Then we have computed the same cumulative percentage plot even for

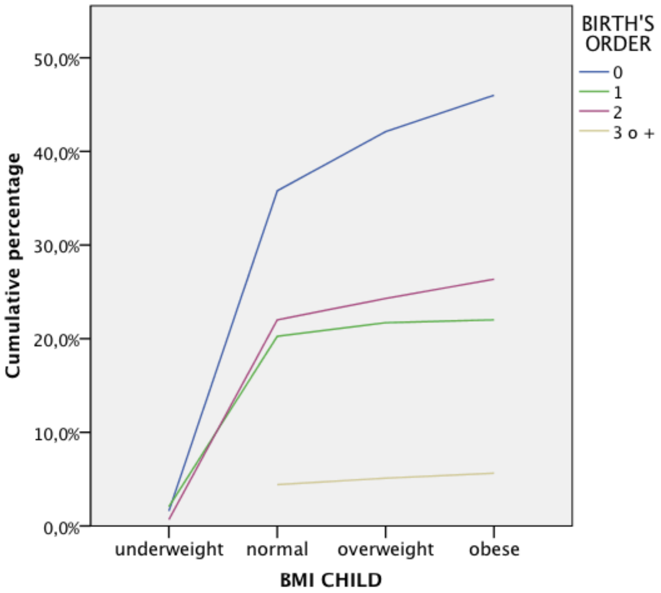


Figure 3.7: Plot of observed cumulative percentages for variable “Birth’s Order”

birth’s order, figure 3.7, and gender sequence, figure 3.8. For what concerns underweight children, no important differences exist depending on their birth’s order or their gender’ siblings; anyway to the growth of child’s BMI, differences appear, becoming larger and larger. In particular a significantly greater percentage of only child than children with siblings are overweight or obese; conversely the third-born or more children are the least frequent in these categories.
From figure 3.8 appears as disparities depending on gender composition of siblings occur: the cumulative percentage for only boys grows more and more with the increase of BMI, followed by sequence FM (first female, second male); conversely the lowest cumulative curve belongs to daughters.

A cumulative odds ordinal logistic regression with proportional odds was run to determine the effect of our variables on the child’s BMI. We have computed the ordinal logistic regression with SPSS PLUM procedure (O’Connell, 2006) reducing the number of predictors as we dealt with missing entries. Indeed PLUM procedure considers only observations having complete data for

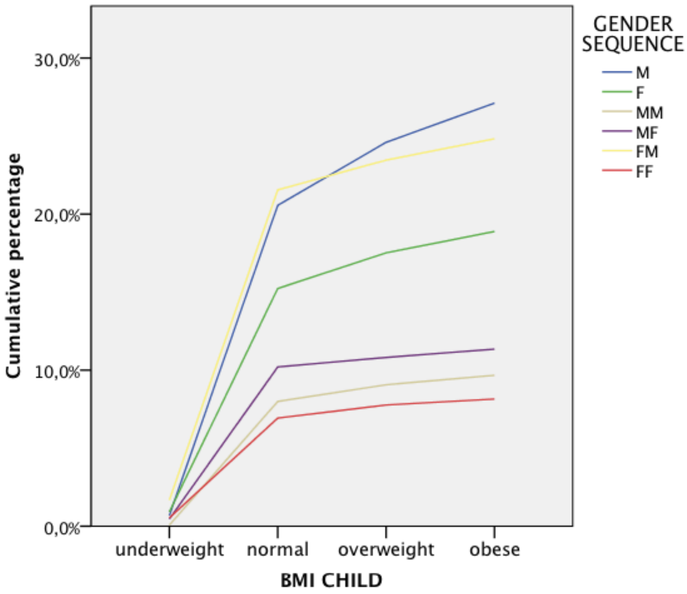
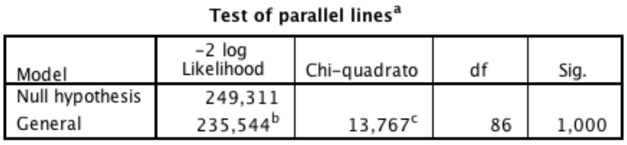


Figure 3.8: Plot of observed cumulative percentages for variable "Gender Sequence"

all variables; however, considering all independent variables, we would have had only 288 complete cases and the results of regression would be poor.
For this reason, we have selected the predictors corresponding to factors considered in literature, as the most affecting gender gaps in China, such as child's gender, child's birth order, urban/rural area, gender sequence, child's medical insurance, highest level of parents' education, parents' priorities, grandparents' help, parents' siblings and additional child. In this way we have included almost the 72% of cases.

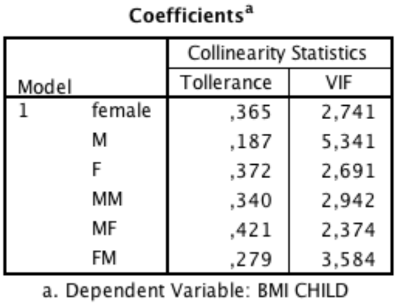
The assumption of odds proportional has been tested by a full likelihood ratio test, comparing the residual of the fitted location model to a model with varying location parameters, $\chi^2(86)$=13.767, p=1.000 as showed in the output in table 3.10.
To test the assumption of multicollinearity, before we had to create dummy variables for our categorical variables. For example, for variable "Gender Sequence", we created 5 (j-1) dummy variables, where each of these dummy variables will indicate the "membership" of a particular category j of the



Test of parallel lines[a]

Model	-2 log Likelihood	Chi-quadrato	df	Sig.
Null hypothesis	249,311			
General	235,544[b]	13,767[c]	86	1,000

Table 3.10: Test of Parallel Lines



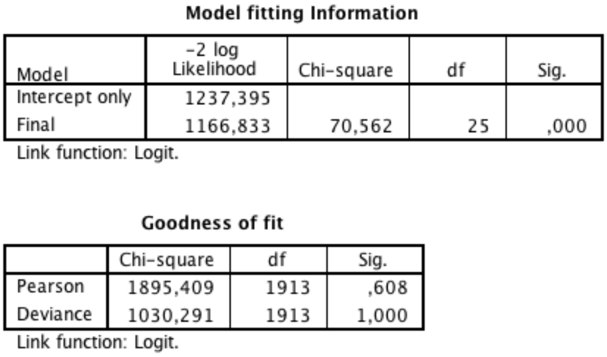
Coefficients[a]

Model		Collinearity Statistics	
		Tollerance	VIF
1	female	,365	2,741
	M	,187	5,341
	F	,372	2,691
	MM	,340	2,942
	MF	,421	2,374
	FM	,279	3,584

a. Dependent Variable: BMI CHILD

Table 3.11: Tests of collinearity for variables “Gender” and “Gender Sequence”

categorical variable; in such way the dummy “MF” would indicate if the child belongs to the sequence “first male, second female” or not and so on for the other categories. After transforming categorical variables in dummies, we have tested the hypothesis; in table 3.11 we have reported a short frame of the output, showing only the test of collinearity for variables “Gender” and “Gender Sequence”. From our results we can be fairly confident that we do not have a problem with collinearity in this particular data set, as all the Tolerance values are greater than 0.1.



Model fitting Information

Model	-2 log Likelihood	Chi-square	df	Sig.
Intercept only	1237,395			
Final	1166,833	70,562	25	,000

Link function: Logit.

Goodness of fit

	Chi-square	df	Sig.
Pearson	1895,409	1913	,608
Deviance	1030,291	1913	1,000

Link function: Logit.

Table 3.12: Model fitting information and Goodness of fit

SPSS generates two tests of the overall goodness-of-fit of the model, as showed in the bottom part of tables 3.12; the deviance goodness-of-fit test indicated that the model was a good fit to the observed data, $\chi^2(644)$=249.311 p=1.000, but most cells were sparse with zero frequencies in 71.8% of cells. However, the final model significantly predicted the dependent variable over and above the intercept-only model, $\chi^2(25)$=70.562, p<.001 as showed in the Model Fitting Table 3.12.

The estimates of the parameters of the model are presented in table 3.13. In our model, we have assumed the slope coefficients to be the same for all j-1=3 equations so it is just the thresholds differing between the three equations, as highlighted in the table. Below are reported the estimates of the slope coefficients which can be used to write the cumulative logit equations. This parameter estimate (slope coefficient) represents the change in the log odds of being in that specific category rather than the reference category; for example, let consider "UR-RU" variable, the reference category is UR-RU=2, that is rural, thus, there is an increase in the log odds of .131 of scoring higher on the dependent variable, namely having an higher BMI, for children living in urban areas compared to children living in rural ones. Anyway, as measuring changes in log odds does not have intuitive meaning, we wish to report the change in terms of the odds that is, the ratio of the odds between the two categories, which is called the odds ratio. For a specific comparison, the odds ratio is the exponential of the log odds of the slope coefficient, namely, the exponential of .131, which is $e^{0.131}$ = 1.14 (95% CI, .81 to 1.604) even reported in the "ExpB" column in table 3.13. It means that for a child living in urban area, the odds of scoring a higher BMI is just over 1.14 times that of a child living in rural one.

The values underlined in red, in table 3.13, represent the statistical significant (p>.05) coefficients. For example, analysing birth order, only the estimate relative to first-born is significant. It means that the odd of being first-born and having an higher BMI was almost 4 times less than of being third-born or more, Wald $\chi^2(1)$=11.327, p=.001. Conversely, "Gender" is not statistically significant. It means that there is not sufficient evidence to say that child's gender has an effect on child's BMI. The same occurs for "Gender Sequence";

since we found no statistically significant estimates, gender compositions of siblings have no significant differences of BMI.

The odds of children having parents with primary education (95% IC 0.169 to 0.732) to be undernourished was about 3 times more than odds of children having highly educated parents (odds of 1.000), a statistically significant effect, Wald $\chi^2(1)$=7.8, p=.005. Finally, the only other two statistically significant odds, concern fathers' siblings: indeed both only child fathers, than fathers having exclusively sisters, have children recording a higher BMI 2 times more, in the first case, and one and half more, in the latter case, than children whose fathers have both female and male siblings, (95% CI 1.030 to 3.901, $\chi^2(1)$=4.195, p=.04 for the firsts and 1.043 to 2.289, $\chi^2(1)$=4.705, p=.03 for the lasts).

3.3 The Causal Graphs-based approach

The proposal of this section is to evidence how an approach based on causal graphs is able to reveal causal relationships that standard statistical tools, as measures of associations and logistic ordinal regression, miss to uncover. In the first part of the section, we focus on learning the causal structure from the observational data. In particular, we resort to and discuss two different R-packages, *bnlearn* (Scutari, 2010) and *pcalg* (Kalisch *et al.*, 2012), supporting causal graphs and the algorithms described in Chapter 1. These packages are available from CRAN; the other suggested package *Rgraphviz* can be installed from Bioconductor and is loaded along with bnlearn if present.

In the second part of the section, we then use the resulting networks, produced by the algorithms, in order to provide a synthetic measure of gender gap, intended, according to Chapter 2, as the causal effect of child's gender on the child's BMI.

3.3.1 The Learning Phase

In learning phase, the aim is to detect a proper causal structure from observational data, able to graphically represent the causal relationships among

VARIABLE	B	LowBou	UppBou	StdErr	Wald	df	Sig	Exp_B	LowB	UppB
BMI_C = underweight	-4,045	-5,529	-2,561	0,757	28,55	1	**0**	**0,018**	0,004	0,077
BMI_C = normal	0,802	-0,639	2,242	0,735	1,19	1	0,275	2,229	0,528	9,416
BMI_C = overweight	2,057	0,604	3,511	0,742	7,695	1	**0,006**	**7,826**	1,829	33,483
UR_RU=urban	0,131	-0,211	0,472	0,174	0,561	1	0,454	1,14	0,81	1,604
UR_RU=rural	0					0		1		
GENDER=male	0,501	-0,056	1,058	0,284	3,103	1	0,078	1,65	0,945	2,882
GENDER=female	0					0		1		
BIRTH_ORDER=only child	-0,858	-1,843	0,127	0,503	2,913	1	0,088	0,424	0,158	1,136
BIRTH_ORDER=first-born child	-1,377	-2,178	-0,575	0,409	11,33	1	**0,001**	**0,252**	0,113	0,563
BIRTH_ORDER=second-born child	-0,376	-1,125	0,374	0,382	0,965	1	0,326	0,687	0,325	1,453
BIRTH_ORDER=third or more born child	0					0		1		
SEQUENCE=M	-0,119	-0,843	0,605	0,369	0,104	1	0,747	0,888	0,43	1,831
SEQUENCE=F	0					0		1		
SEQUENCE=MM	-0,154	-1,064	0,757	0,465	0,109	1	0,741	0,858	0,345	2,132
SEQUENCE=MF	-0,466	-1,25	0,317	0,4	1,363	1	0,243	0,627	0,287	1,373
SEQUENCE=FM	-0,627	-1,345	0,09	0,366	2,939	1	0,086	0,534	0,261	1,094
SEQUENCE=FF	0					0		1		
EDU_P=no education	-1,491	-6,816	3,834	2,717	0,301	1	0,583	0,225	0,001	46,241
EDU_P=primary school	-1,046	-1,78	-0,312	0,374	7,8	1	**0,005**	**0,351**	0,169	0,732
EDU_P=middle school	-0,429	-0,962	0,104	0,272	2,484	1	0,115	0,651	0,382	1,11
EDU_P=college/university	0					0		1		
PRIOR_F=child's health	0,224	-0,366	0,814	0,301	0,553	1	0,457	1,251	0,693	2,257
PRIOR_F=same child as income	0,381	-0,112	0,875	0,252	2,291	1	0,13	1,464	0,894	2,399
[PRIOR_F=good income	0					0		1		
PRIOR_M=child's health	-0,11	-0,679	0,46	0,291	0,142	1	0,706	0,896	0,507	1,584
PRIOR_M=same child as income	0,055	-0,394	0,504	0,229	0,058	1	0,81	1,057	0,674	1,656
[PRIOR_M=good income	0					0		1		
BS_M=only child	-0,071	-0,885	0,743	0,415	0,029	1	0,865	0,932	0,413	2,102
BS_M=male culture	-0,007	-0,395	0,381	0,198	0,001	1	0,972	0,993	0,674	1,464
BS_M=female culture	-0,093	-0,57	0,384	0,244	0,146	1	0,702	0,911	0,565	1,469
BS_M=mix culture	0					0		1		
BS_F=only child	0,696	0,03	1,361	0,34	4,195	1	**0,041**	**2,005**	1,03	3,901
BS_F=male culture	0,129	-0,295	0,553	0,216	0,354	1	0,552	1,137	0,744	1,738
BS_F=female culture	0,435	0,042	0,828	0,201	4,705	1	**0,03**	**1,545**	1,043	2,289
BS_F=mix culture	0					0		1		
POT_HELP=no	-0,203	-0,867	0,46	0,339	0,361	1	0,548	0,816	0,42	1,584
POT_HELP=yes	0					0		1		
ADD_CHILD=no	-0,091	-1,044	0,862	0,486	0,035	1	0,851	0,913	0,352	2,368
ADD_CHILD=yes, one	0,114	-0,938	1,167	0,537	0,045	1	0,832	1,121	0,391	3,212
ADD_CHILD=yes, more than one	0					0		1		

Legenda:
B=parameter estimate of the slope coefficient
LowBou=lower bound of 95% CI for B
UppBou=upper bound of 95% CI for B
Exp_B= exponential of B, or odds ratio
LowB= lower bound of 95% CI for Exp_B
UppB= upper bound of 95% CI for Exp_B

Table 3.13: Estimates of parameters of the ordinal logistic regression.

MACRO-AREA	NAME VARIABLE	DESCRIPTION	VALUES	MISSING
Demographic	URBAN/RURAL	*Type of household registartion*	1=URBAN SITE 2=RURAL SITE	0%
	GENDER	*Child's gender*	1=MALE; 2= FEMALE	0%
	AGE_C	*Child's age*		0%
	AGE_F	*Father's age*		3,50%
	AGE_M	*Mother's age*		0,60%
Time & Gender Births	BIRTH_ORDER	*Child's birth order; gained from ages of siblings (even >18)*	0= ONLY CHILD; 1=FIRST CHILD; 2=SECOND CHILD; 3= THIRD OR LOWER CHILD	0%
	GENDER_SEQUENCE	*Sequence of gender's siblings*	1=ONLY MALE; 2= ONLY GIRL; 11= ONLY MALES; 22= ONLY GIRLS; 12= AT LEAST 1 MALE BEFORE 1 OR MORE GIRLS; 21= AT LEAST 1 GIRL BEFORE 1 OR MORE MALES	0%
	TIME_DISTANCE	*Time in months before previous birth; gained from previous birth date*	0= ONLY OR FIRST CHILD	0,07%
	AVER_TIME	*Average time from previuos birth; average calculated among all observations*	1=MORE TIME THAN AVERAGE; 0= ON AVERAGE; -1= LESS TIME THAN AVERAGE	0,07%
Medical	MED.INS.CHILD	*If child has a medical insurance or not*	0=NO 1=YES	0%
	MED.INS_F	*If father has a medical insurance or not*	0=NO 1=YES	16,40%
	MED.INS_M	*If mother has a medical insurance or not*	0=NO 1=YES	9%
	BMI_C	*Child's BMI cut-offs; calculated as weight/(height)^2*	1=UNDERWEIGHT; 2=NORMAL; 3=OVERWEIGHT; 4=OBESE	0%
	BMI_F	*Father's BMI cut-offs; calculated as weight/(height)^2*	1=UNDERWEIGHT; 2=NORMAL; 3=OVERWEIGHT; 4=OBESE	24,10%
	BMI_M	*Mother's BMI cut-offs; calculated as weight/(height)^2*	1=UNDERWEIGHT; 2=NORMAL; 3=OVERWEIGHT; 4=OBESE	12,40%
Educational & Working	EDU_P	*Highest level of formal completed education between parents*	0=NO EDUCATION; 1= PRIMARY SCHOOL; 2=MIDDLE SCHOOL; 3=COLLEGE/UNIVERSITY; 9=UNKNOWN	0%
	JOB_POS_P	*Highest job position between parents*	1=OWNER MANAGER; 2=INDEPENDENT OPERATOR; 3=PERMANENT EMPLOYEE; 4=CONTRACTOR WITH OTHER PERSON OR ENTERPRISE; 5= TEMPORARY WORKER; 6=PAID FAMILY WORKER; 7=UNPAID FAMILY WORKER; 8=OTHER; 9=UNKNOWN; 20=NOT WORKING	0%
Child's grewing up context	DIETBEHAV_F	*If father has a good knowledge of healthy diet behaviours*	2=VERY GOOD KNOWLEDGE; 1=GOOD KNOWL; 0=NEUTRAL; -1=BAD KNOWL; -2=VERY BAD KNOWL	21,90%
	DIETBEHAV_M	*If mother has a good knowledge of healthy diet behaviours*	2=VERY GOOD KNOWLEDGE; 1=GOOD KNOWL; 0=NEUTRAL; -1=BAD KNOWL; -2=VERY BAD KNOWL	15,80%
	PRIOR_F	*Father's priority*	1=CHILD'S HEALTH; 0=SAME IMPORTANCE; -1=MONEY	16,60%
	PRIOR_M	*Mother's priority*	1=CHILD'S HEALTH; 0=SAME IMPORTANCE; -1=MONEY	9,10%
	POT_HELP	*If a potential help from grandparents*	0=NO 1=YES	12,50%
Parents' primary socialization	BS_F	*Father's siblings composition; to consider in which gender context the father has grown up*	0=ONLY CHILD; 1= MALE CULTURE; 2= FEMALE CULTURE; 3=MIX CULTURE	11,50%
	BS_M	*Mother's siblings composition; to consider in which gender context the mother has grown up*	0=ONLY CHILD; 1= MALE CULTURE; 2= FEMALE CULTURE; 3=MIX CULTURE	10%
Fertility intentions	ADD_CHILD	*If mother wants one or more additional child*	0=NO; 1=YES, ONE; 2=YES, MORE THAN ONE	12,30%

Table 3.14: Prospect of the variables considered in the analysis.

variables encoded in the data. As mentioned in section 1.3, two main approaches exist to learn the unknown causal structure: via constrain-based or via score-based algorithms. The firsts are implemented in both packages, while the only available score-based algorithm, the *Hill-Climbing*, is supported exclusively in bnlearn.

Irrespective of the approach used, the structure of a DAG can be recovered from observational data, up to d-separation equivalence (Pearl, 2000), only if the three assumptions described in section 1.2 are satisfied; such assumptions are in general untestable from observational data and must come from subject matter experts.
The Causal Markov Condition (CMC, section 1.2.2) asserts, briefly, that each variable in the model is independent of its nondescendants given its immediate causes, thus it requires to consider the factors affecting child's gender. Generally in statistics, gender is considered an independent variable which cannot be controlled or influenced by any other cause. Anyway in China, where son preference is still strong and sex-selective abortion is a widespread practice, child's gender might be the outcome of parents' fertility preferences; for these reasons, information regarding birth's order and sex composition of existing siblings must be considered as potential factors affecting the resulting child's gender. The Markov assumption is, in the end, treated more as a guiding principle (Karwa *et al.*, 2011) ensuring that all relevant causes are included in the analysis rather than an actual assumption.
The faithfulness assumption (CFC, section 1.2.3) ensures that the population generating the causal model has exactly those independence relations specified by the DAG structure, and no additional ones. By assuming faithfulness, we eliminate the cases where there are any independences in the population that are not consequence of the CMC.
Finally, the last assumption requires that there are no latent variables in the model violating the CMC; again, this assumption is strong, whose validity could be ensured by verification from experts opinion. For example, in our analysis, the introduction of information regarding child's birth order and gender composition of siblings has been led by literature and previous studies on factors affecting child's nutrition in Asia (Prashant Kumar Singh,

2013; Li *et al.*, 2004).

After such necessary digression about model's assumptions, the learning phase can begin; it started applying the score-based algorithm Hill-Climbing (hc) greedy search on the space of directed graphs 1.3.2 (for more details, see 1.3.2). The optimized implementation, used by default, uses score caching, score decomposability and score equivalence to reduce the number of duplicated tests; in particular, the learned procedure used 781 tests and resulted in a completely direct DAG with 24 direct arcs. The score function we choose to maximize is Bayesian Information Criteria (BIC) (Schwarz, 1978) rather than Akaikes Information Criteria (AIC) as it is derived within a Bayesian framework and reflects sample size. Furthermore, BIC generally penalizes free parameters and therefore it would pick the more parsimonious model than AIC might suggest. Anyway alternative scenario maximizing AIC led to similar outputs.

Here we report the learnt model, expressed in terms of conditional independences, in R code:

```
> bn.hc

  Bayesian network learned via Score-based methods

  model:
   [MED.INS_M][BMI_M|MED.INS_M][DIETBEHAV_M|MED.INS_M][BS_M|MED.INS_M][POT_
       HELP|BS_M][BS_F|BS_M][ADD_CHILD|BS_M][AGE_FA|BS_F]
   [EDU_P|POT_HELP][BIRTH_ORDER|EDU_P][AGE_MO|AGE_FA][JOBPOS_P|EDU_P][UR_RU|
       JOBPOS_P][GENDER|BIRTH_ORDER][AVER_TIME|BIRTH_ORDER]
   [MED.INS_F|JOBPOS_P][SEQUENCE|GENDER:BIRTH_ORDER][MED.INS_C|MED.INS_F][
       BMI_C|GENDER][BMI_F|MED.INS_F][DIETBEHAV_F|MED.INS_F]
   [PRIOR_F|MED.INS_F][PRIOR_M|MED.INS_M:PRIOR_F]
  nodes:                                  23
  arcs:                                   24
    undirected arcs:                      0
    directed arcs:                        24
  average markov blanket size:            2.17
  average neighbourhood size:             2.09
  average branching factor:               1.04

  learning algorithm:                     Hill-Climbing
  score:                                  Bayesian Information Criterion
  penalization coefficient:               3.590035
  tests used in the learning procedure:   781
  optimized:                              TRUE
```

```
> score(bn.hc,dat,type=''bic")
[1] -25727.07
```

For a clearer visualization and, consequently, interpretation, in figure 3.9, we report the output gained with Rgraphviz package. It represents the causal graph regarding the variables of table 3.14, with the exception, however, of “Child’s age” and “Time Distance”. The first has been eliminated as the “child’s BMI” is computed already per child’s age, while the second as the number of months occurring between births is already incorporated in the synthetic variable “Average Time”. Thus they would be redundant variables. In red the two variables of interest are underlined: the input “Child’s Gender” and the outcome “Child’s BMI”. As shown in the output, there is a direct edge between them, that is a causal relationship; we can conclude, thus, that a gender discrimination occurs as in absence of disparities the outcome should be independent of child’s gender (no direct edge).
Notice that the causal graph detects a relationship occurring between gender and child’s BMI that regression do not. But it gives even a further information: the direction of such relationships. Indeed from output in figure 3.9, we can conclude, not only that there is a relationship between the two target variables, as applying measure of associations, but even that child’s gender can be interpreted as a *cause* affecting child’s BMI, since the whole dependence/independence structure of the variable set is depicted in the graph, in a full multi-variate approach. Even the *only* direct cause. Notice that unlike the regression, where the symbol “=”, linking the dependent and independent variables, allows the symmetry $x = (y - \varepsilon)/\beta$, in causal graph the direction of the dependence, encoded in the arrow, cannot be in any way reversed.

Both in ordinal regression than in causal graphs, the variables “Birth’s Order” and “Parents’ Education” have a relationship with child’s BMI. Anyway, in PLUM procedure, the relation is expressed in terms of change in the log odds of being in that specific category rather than the reference category; for example, considering parents’ education, the odds to be undernourished

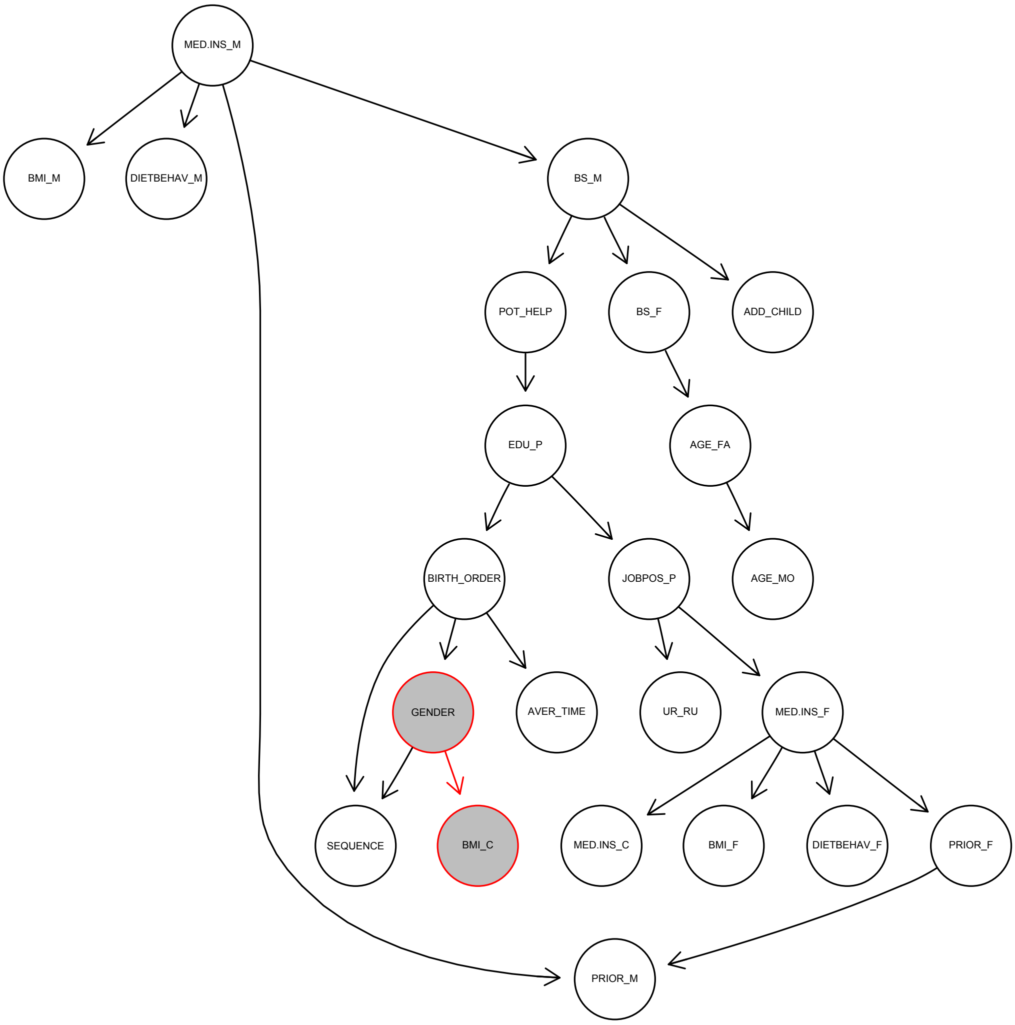


Figure 3.9: Causal Graph learned by Hill-Climbing, BIC score.

of children having parents with primary education, is about 3 times more than odds of children having highly educated parents. In causal graph we have, again, one additional information: parents' education has, of course, an impact on child's BMI but mediated (conditioned) by birth's order and gender's effect.
Causal graphs, thus, allow for detecting the conditioned effects of variables, or even called *indirect causes*, as once we intervene on birth's order, parents' education have no more effect on child's BMI.

Subsequently, we have deepen the causal relationship between child's gender and child's BMI conditioning on birth's order; namely we were interested in testing if gender differences in child nutrition and care occurred between only child and children having siblings.
In figures 3.10 and 3.11, we have reported the two causal graphs showing, respectively, the causal structure for only child and for children with siblings. In case of only child, there is gender equality as no arrow connects the variable "Gender" with "Child's BMI", thus they are independent. It means that in only-child households, no matters if child is a boy or a girl for nutritional and caring decisions. Conversely, in families with more than one child, child's gender influences the food's and care's allocation as child's gender has an indirect effect, mediated by birth's order, on child's BMI. The graph 3.11 is besides interesting for another aspect, since we can detect two clusters of variables independent each others: one involving child's gender, birth's order, gender sequence of siblings and child's BMI, and the other containing all the other variables of the analysis. It means that, for children having brothers or sisters, what it is crucial in determining their nutritional outcomes, are not the socio-economical conditions where they grew up, but instead their birth's and gender's placing in the household.
This highlights, even more, that a gender gap exists since the nutrition of a son/a daughter would not be to depend on parental economical situation or on if he/she lives in urban or rural areas as one might expect, but rather by the fact that just that boy/girl may have a younger/older brother/sister.

According to Pearl's d-separation criterion 1.2.1, the set {birth's order, gender sequence, average time, child's BMI} in figure 3.11 represents

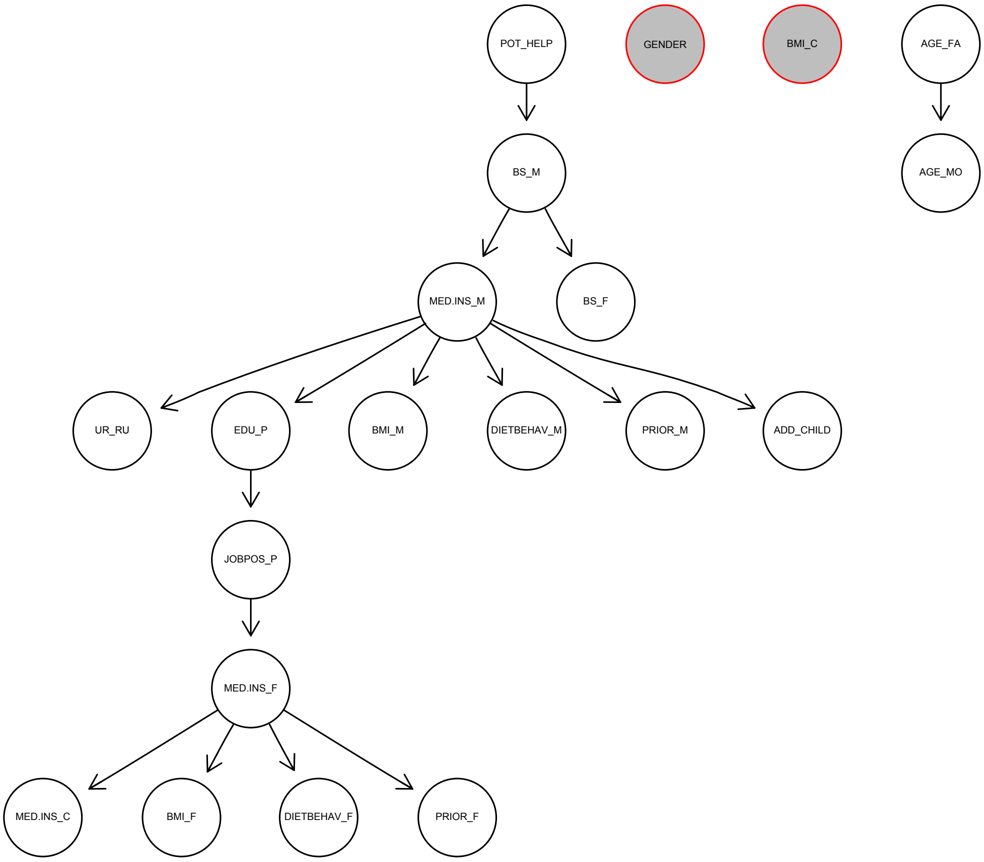


Figure 3.10: Causal Graph learned by Hill-Climbing, BIC score, for only child

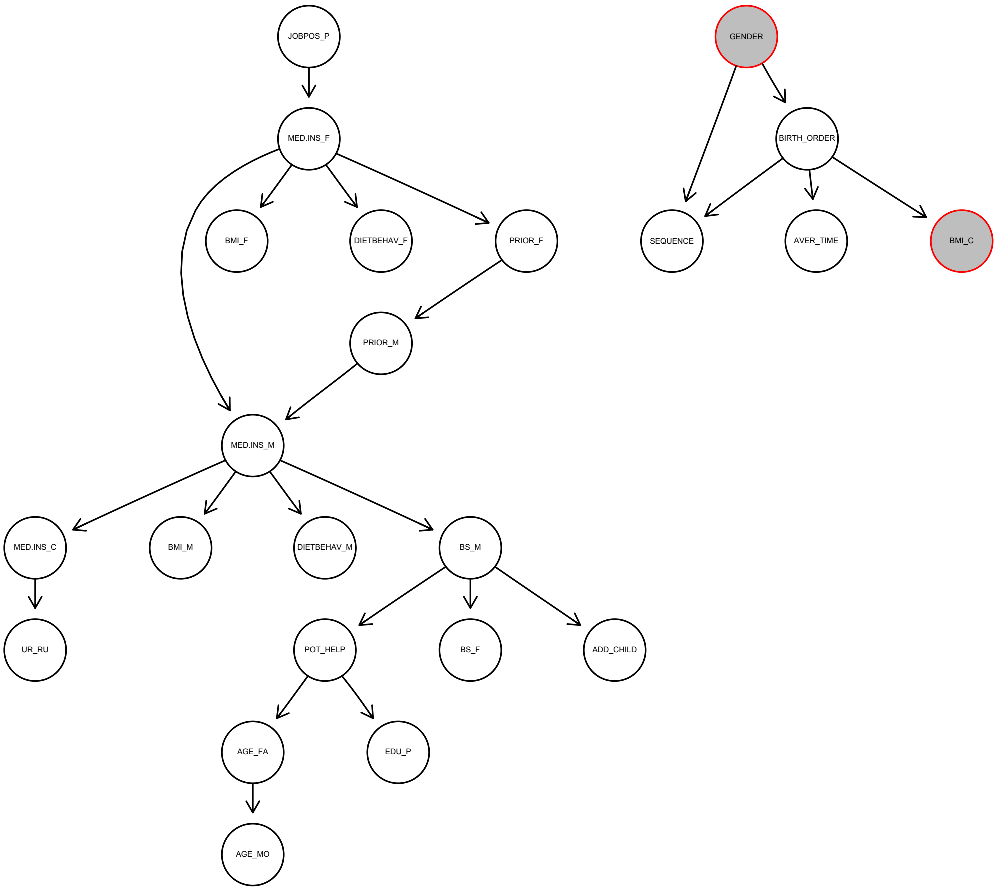


Figure 3.11: Causal Graph learned by Hill-Climbing, BIC score, for children with siblings

a fork, such that “gender sequence”, “average time” and “child’s BMI” are marginally dependent but become independent (blocked) once we condition on the middle variable birth’s order. Furthermore the set {gender, gender sequence, birth’s order} contains both a fork and a collider, implying that child’s gender acts on gender sequence both directly, and indirectly through birth’s order. Such a structure is consistent with the theoretical framework; indeed, in a country where sex-selective technology is available and where son preference is predominant, it seems reasonable that the previous child’s gender would influence the following child’s gender and hence the resulting gender sequence of siblings.

For completeness of the analysis, we attempted to run even constrain-based algorithms as GS, IAMB, and PC (see section 1.3.1), available on the considered R-packages, with the aim of a comparison with hill-climbing. They returned different outputs respect to the score-based one, thus a discussion is required. For example, the Incremental Association algorithm, implemented in bnlearn, produced a completely undirected causal network, as shown in the following R code:

```
> bn.iamb

  Bayesian network learned via Constraint-based methods

  model:
    [undirected graph]
  nodes:                                 23
  arcs:                                  20
    undirected arcs:                     20
    directed arcs:                       0
  average markov blanket size:           1.74
  average neighbourhood size:            1.74
  average branching factor:              0.00

  learning algorithm:                    Incremental Association
  conditional independence test:         Mutual Information
  alpha threshold:                       0.05
  tests used in the learning procedure:  644
  optimized:                             TRUE
```

The corresponding output is shown in figure 3.12. Notice that gender results independent by child’s BMI, anyway the algorithm detects a cluster of depen-

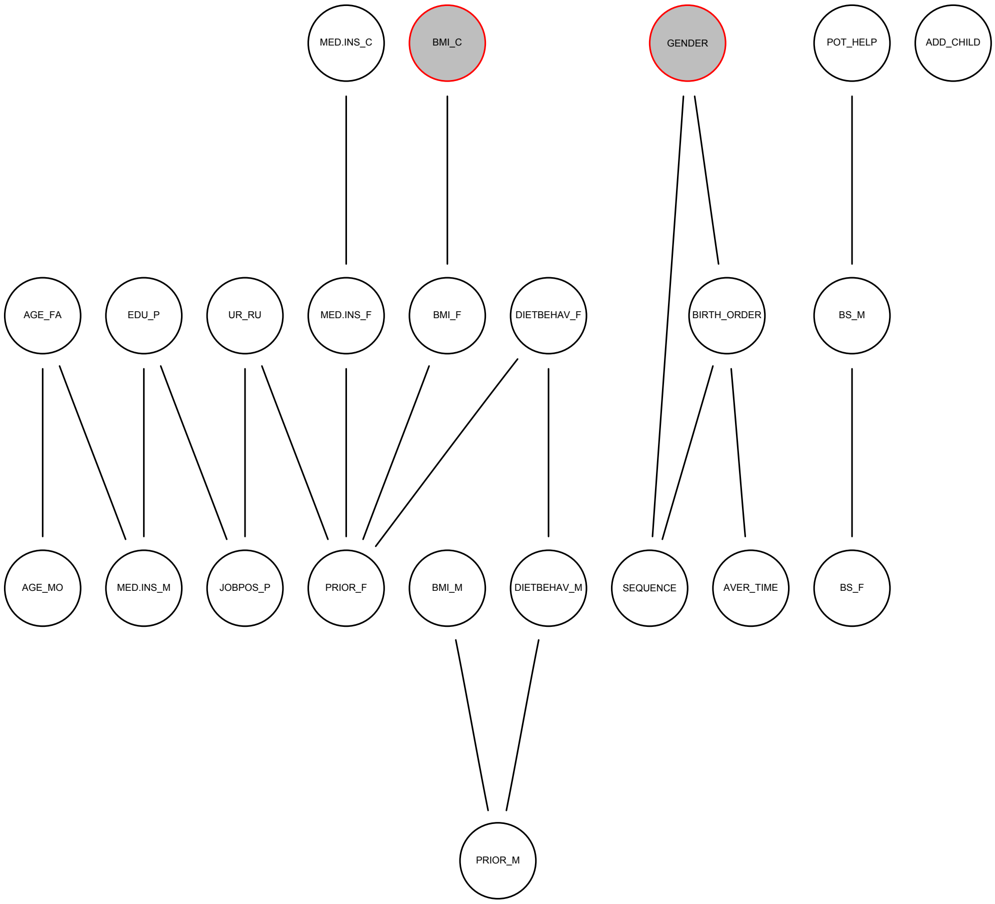


Figure 3.12: Causal Graph learned by constrain-based method Incremental Association (iamb).

dent variables: gender, birth's order, gender sequence and average time, that is similar to figure 3.11. Notice that likewise the measures of associations provided in subsection 3.2.2, the IAMB algorithm was able to detect only pairwise associations between couples of variables, represented by undirected edges. This constitutes a significant weakness; indeed the main advancement regarding causal graphs, with respects to standard statistical methods, is that they are able to reveal the direction of causal-effect relationships, depicted as directed edges. The same limit occurred with GS algorithm.

Even considering a different R-package, pcalg, the results do not improve. The output of PC algorithm, for instance, reported in figure 3.13, is even less

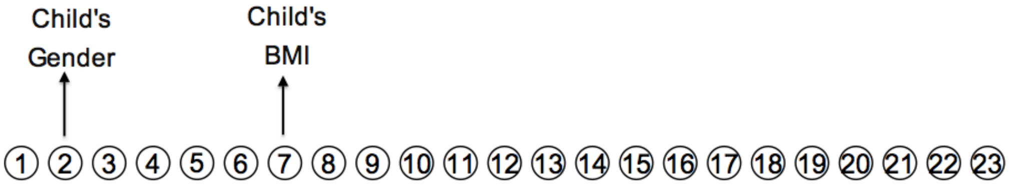


Figure 3.13: Causal Graph learned by constrain-based PC algorithm.

informative, showing an evident failure of the procedure. Indeed it reported an output in which all variables result independent (no edges connecting no couples of vertices). This is clearly unlikely, given the previous results as well as the a priori knowledge coming from reference literature existing on the phenomenon under analysis.

The outputs resulting by constrain-based algorithm application, require, thus, some considerations. The lack of effectiveness of the constrain-based algorithms supported by different R-packages, suggests how the issue would lie in the procedure rather than in their practical implementation in R software. Certainly, the number of independence tests in case, like our, of high dimensionality as well as high number of categories for each variable, increases exponentially. In literature it is recognised that when sample size or the conditional set is big, then the possibility of rejecting the null hypothesis is lower and independence will be assumed. Lack of support translates, thus, in independence. Notice that, in case of accepting independence, it does not mean that data support independence, but that there is no evidence in the data against it. In this sense the constrain-based algorithm are poorly informative, rather than wrong.

It is, otherwise true, that the complexity due to high dimensionality and high number of categories is the same also for score-based algorithm, in our particular case, for hill climbing. However, as we mentioned in section 1.3.2, most of scoring criteria derived in the literature are “decomposable”, in the sense that the score can be written as a sum of measures, each of which is a function only of one node and its parents. It allows to such methods to efficiently perform even when dealing with arising complexity.

Hence, comparing different algorithms as well as different computational approaches, the hill-climbing results, for our specific dataset, to be the most effective, since it provides additional causal information, such as directionality, that, conversely, other procedures are not able to give.

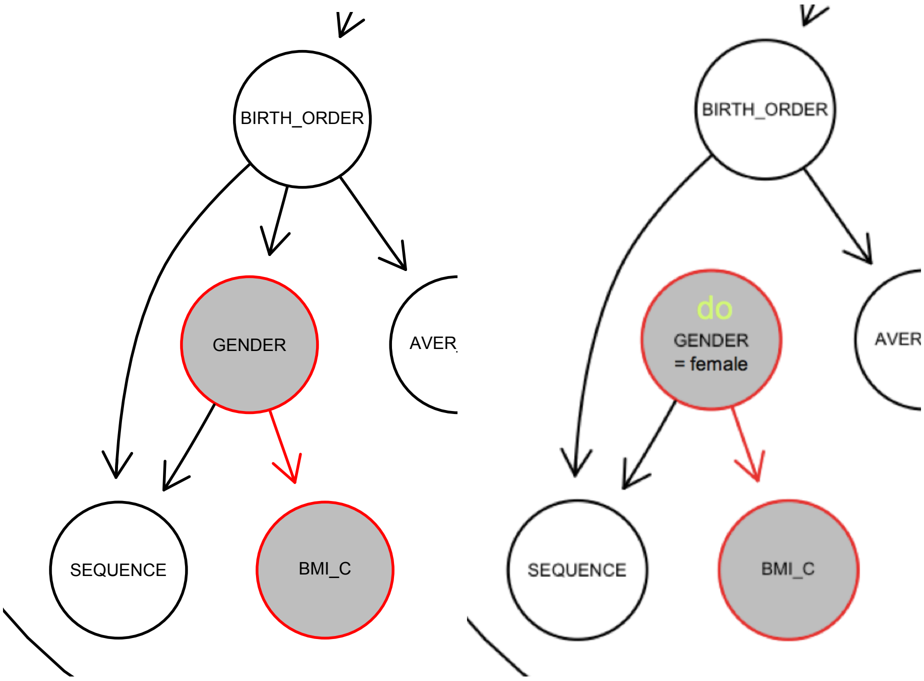
What emerges from learning phase, is that child's gender has a causal effect on child's BMI even if birth's order has a crucial role in such relationship, since differences among only child and children having siblings occur. Furthermore, a pattern of key-variables including gender, birth's order, sex composition of siblings and timing among birth is evident; it finds consistency in literature (Li *et al.*, 2004) and shows how, more cultural and gendered than economical factors have a causal effect in child's nutrition and health in China.

3.3.2 The Estimation Phase

In this second phase, our purpose is to provide conceivable estimates of the causal effect existing between child's gender and child's BMI. As described in section 1.4.3, some automatized solutions implemented in R software already exist; among them, the IDA algorithm in pcalg package, which is able to estimate bounds of causal effects when the DAG structure is unknown. Unluckily such procedure requires some strict assumptions that our data do not meet, for instance, the normality for the distribution of variables representing potential causes, as child's gender.

In order to provide a synthetic measure of gender gap in child's nutrition in China, we propose another approach. Resorting to the causal structure resulting from the learning phase, we apply the method illustrated in Chapter 2.

According to definition 1.4.1, the causal effect of gender on child's BMI, is denoted as $P[BMI\_C|do(Gender = g)]$ where $g \in \{male, female\}$ and it is a function from "Child's Gender" to the space of probability distribution on "BMI_ C". The atomic intervention of do-operator results in removing the links between gender variable and its set of parents from the graph resulting from learning phase and shown in 3.9, denoted as DAG D. It creates, thus,



(a) Pre-intervention

(b) Post-intervention

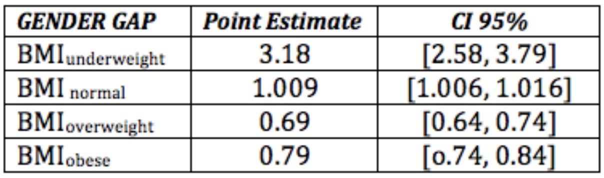
Figure 3.14: A frame of the original DAG D (figure 3.9) and the post-intervention DAG D_{Gender}. It results by forcing "gender" to take the particular value Gender=female and deleting the arcs between "gender" and its parents.

a new DAG D_{Gender}, depicted in figure 3.14.

According to equation 2.2, the gender gap, denoted as $\Gamma(G)$ is given by the causal effect of changing the treatment variable from female to male on the target level of outcome. The gender gap on undernourishment level, for example, is given by:

$$\Gamma(G)_{underweight} = \frac{E[(BMI\_C = underweight)|do(Gender = female)]}{E[(BMI\_C = underweight)|do(Gender = male)]}$$

where $E[BMI\_C = underweight|do(Gender = female)]$ is the conditional expectation for the child of being undernourished under the atomic interven-



GENDER GAP	Point Estimate	CI 95%
$\text{BMI}_{\text{underweight}}$	3.18	[2.58, 3.79]
$\text{BMI}_{\text{ normal}}$	1.009	[1.006, 1.016]
$\text{BMI}_{\text{overweight}}$	0.69	[0.64, 0.74]
$\text{BMI}_{\text{obese}}$	0.79	[o.74, 0.84]

Table 3.15: Point estimates and Confidence Intervals at 95% of Gender Gaps from Causal Model

tion $Gender = female$. Equivalently for overweight and obese BMI's levels:

$$\Gamma(G)_{overweight} = \frac{E[(BMI\_C = overweight)|do(Gender = female)]}{E[(BMI\_C = overweight)|do(Gender = male)]}$$

and

$$\Gamma(G)_{obese} = \frac{E[(BMI\_C = obese)|do(Gender = female)]}{E[(BMI\_C = obese)|do(Gender = male)]}$$

Recalling the analogy between "net" and "direct" as well as "gross" and "total" developed in Chapter 2, the gender gaps below are all net, since child's gender has a *direct* causal effect on outcome in graph D. According to equation 1.14, the gender gaps have been formulated in terms of ratio rather than difference of expected values as in Chapter 2. Anyway, the two expression are equivalent but since in literature gender gaps are generally expressed in terms of female/male ratios, we choose the first formulation to be compliance.

Table 3.15 shows the causal effects of child's gender on child's BMI, namely the gender gaps, derived by the causal graph in figure 3.9. On the basis of the results, a girl would be undertaken to a risk of low BMI; in particular the expected undernourishment increases 3.18 times for girls when compared with boys. Conversely boys record a potential gender gap in obesity direction as the expected obesity decreases 1.27 times when switching from female to male.

Figure 3.15 depicts the estimates of means and confidence intervals at 95% of gender gaps, showing the overlap amongst the estimates of causal gender effects. It can be seen that the causal effect of switching from female to male is very marked for undernourishment level while it tends to decrease

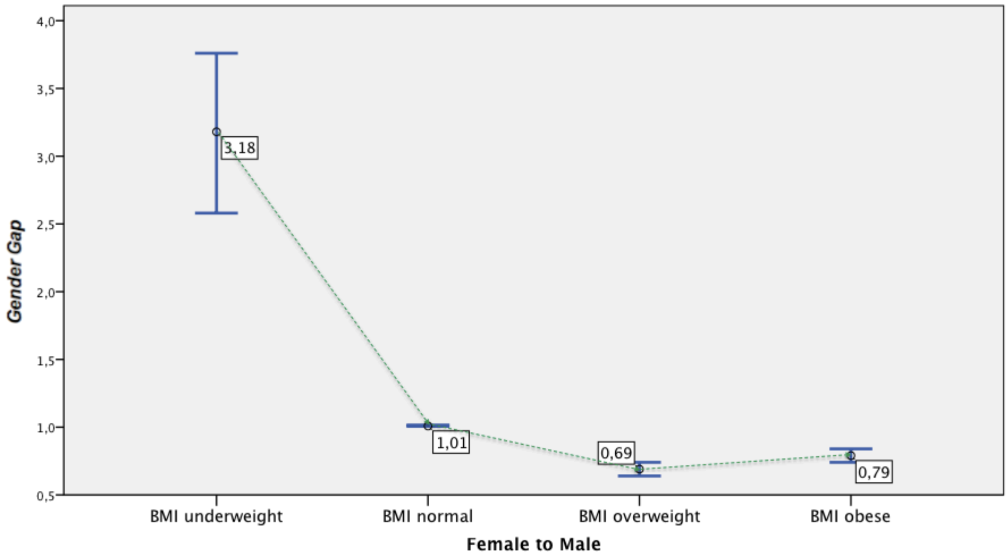


Figure 3.15: Overlap between the Confidence Intervals of Gender Gaps for each BMI's level.

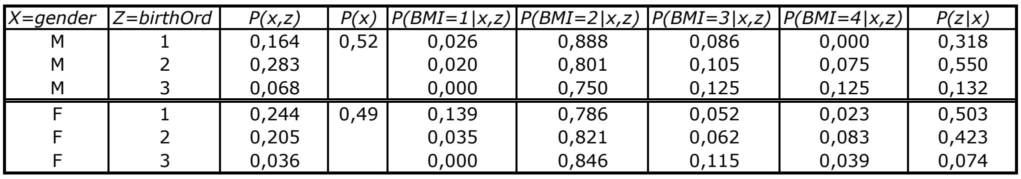
for higher levels of BMI; in particular for obese and overweight levels, it falls under the equality ratio of 1 meaning that the gender gap is reversed at boys detriment.

Since our aim is to deep the gender gaps on child's nutrition and care, it is important to take into account the gaps in both extremes of BMI's distribution: undernourishment and obesity. Our results show that child's gender has an evident effect on child's BMI; anyway, the phenomenon assumes different forms polarized for gender: for girls gender gap would translate in undernourishment while for boys in overweight.

We already demonstrated the importance of birth's order in assessing gender gaps in China.

In figure 3.11, in particular, we learnt the causal structure relative to children having brothers or sisters, underlining how the gender effect is, in such case, indirect as mediated by birth's order. The gender gap is in this case "gross", as it incorporates the birth's order effect; hence, an adjustment for such confounding characteristics is required, as indicated in equation 2.4.

The set $Z = \{birth's\ order\}$ represents a sufficient set of covariates for adjustment since it meets the conditions of Pearl's Front-Door Criterion 1.4.3,



X=gender	Z=birthOrd	P(x,z)	P(x)	P(BMI=1|x,z)	P(BMI=2|x,z)	P(BMI=3|x,z)	P(BMI=4|x,z)	P(z|x)
M	1	0,164	0,52	0,026	0,888	0,086	0,000	0,318
M	2	0,283		0,020	0,801	0,105	0,075	0,550
M	3	0,068		0,000	0,750	0,125	0,125	0,132
F	1	0,244	0,49	0,139	0,786	0,052	0,023	0,503
F	2	0,205		0,035	0,821	0,062	0,083	0,423
F	3	0,036		0,000	0,846	0,115	0,039	0,074

Table 3.16: Probabilities involved in front-door formula

in particular:

1. we assume, according to graph in figure 3.11, that child's gender has no effect on child's BMI except as mediated by birth's order effect;

2. we must even assume that, even if a latent factor is affecting the child's BMI, it nevertheless has no effect on the birth's order except indirectly, through child's gender;

3. likewise, we assume that no other factor that affects birth's order has any influence on child's gender.

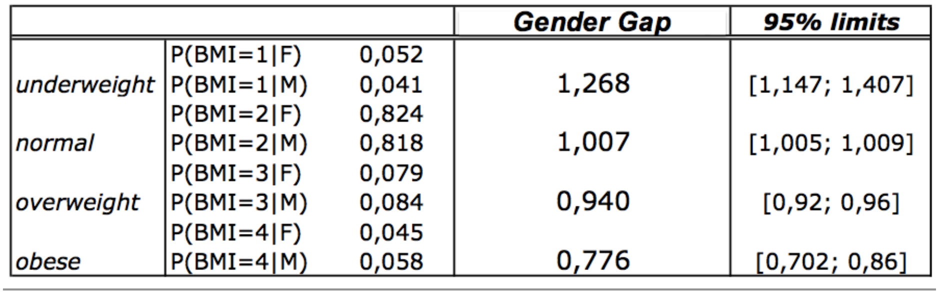
Applying the front-door formula in equation 1.4.2, the post-intervention distribution under the intervention level $Gender = female$, is given by:

$$E[BMI\_C = bmi|do(Gender = female)] \\ = \sum_{birOrd} P(birOrd|Gender = female) \sum_{Gender} P(bmi|Gender, birOrd)P(Gender) \tag{3.3}$$

while for $Gender = male$:

$$E[BMI\_C = bmi|do(Gender = male)] \\ = \sum_{birOrd} P(birOrd|Gender = male) \sum_{Gender} P(bmi|Gender, birOrd)P(Gender) \tag{3.4}$$

In table 3.16, we report the values of conditional and marginal probabilities involved in front-door adjustment for birth's order. For instance, the conditional expectation $E(BMI\_C = undeweight|Gender = female)$ results



			Gender Gap	95% limits
	P(BMI=1|F)	0,052		
underweight	P(BMI=1|M)	0,041	1,268	[1,147; 1,407]
	P(BMI=2|F)	0,824		
normal	P(BMI=2|M)	0,818	1,007	[1,005; 1,009]
	P(BMI=3|F)	0,079		
overweight	P(BMI=3|M)	0,084	0,940	[0,92; 0,96]
	P(BMI=4|F)	0,045		
obese	P(BMI=4|M)	0,058	0,776	[0,702; 0,86]

Table 3.17: Point estimates and Confidence Intervals at 95% of Gender Gap for children with siblings, adjusted per birth's order.

from:

$$
\begin{aligned}
&E[(BMI\_C = underweight|do(Gender = female)] = \\
&0.503(0.139 * 0.485 + 0.026 * 0.515) + 0.423(0.034 * 0.485 + 0.020 * 0.515) + \\
&+ 0.074(0 * 0.485 + 0 * 0.515) = 0.052
\end{aligned}
\tag{3.5}
$$

The gender gap for BMI's level= bmi, net of birth's order effect, is given by the ratio between equation 3.3 and 3.4:

$$
\Gamma(G)_{bmi} = \frac{\sum_{birOrd} P(birOrd|Gender = female) \sum_{Gender} P(bmi|Gender, birOrd)P(Gender)}{\sum_{birOrd} P(birOrd|Gender = male) \sum_{Gender} P(bmi|Gender, birOrd)P(Gender)}
\tag{3.6}
$$

Table 3.17 shows, thus, the resulting estimates and corresponding interval levels at 95%, of gender gap for children with siblings, adjusted for the confounder birth's order. Notice the most meaningful results: girls have an expectation 1.3 times more of being undernourished, conversely boys 1.3 times more of being obese. The gender gap, thus as highlighted before, from girls' detriment for low level of BMI, reverses to boys' disadvantage for high BMI's levels; on the contrary, for normal BMI's level, gender equality occurs.

In figure 3.16 are illustrated the corresponding estimates of means and confidence intervals of table 3.17. From such representation it is even more

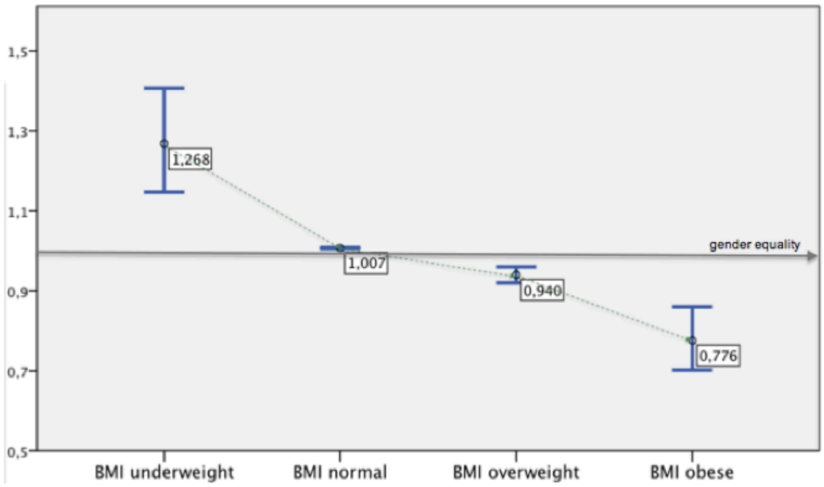


Figure 3.16: Overlap between the Confidence Intervals of Gender Gaps for each BMI's level for children with siblings, adjusted per "birth's order".

evident the reversing gender gap trend at increasing BMI and, in particular, the switching from female to male detriment between normal and overweight levels.

A graphical comparison about the gender gaps trends for all children and children with siblings results in figure 3.17, matching figures 3.15 and 3.16 in the same y scale. The whole sample size of $N = 1313$ is halved to $N_{sibl} = 709$ once we consider solely children with siblings. Notice that gender gap to girls detriment in underweight is dramatically more marked when we consider the whole sample. It results since the girls' expectation of being undernourished is higher (.07) when we include all children than for girls having siblings (.052); moreover, the whole sample male expectation of being underweight is lower (.022) than for boys with siblings (.041). It means that undernourishment consistently affects only girl as well.

Conversely the boys' gender gap in overweight including all children, fades once we focus on children with siblings reaching almost equality level. It is due to an increase of 4 percentage points of male overweight expectation once we consider only boys with siblings.

For normal and obese levels of BMI, no significant changes occur.

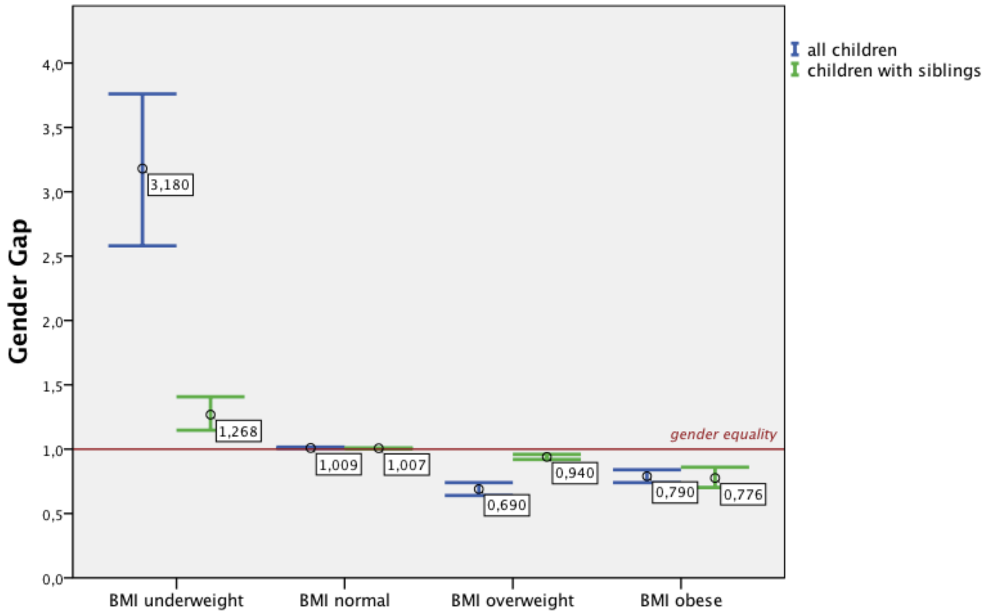


Figure 3.17: Comparison between gender gaps trends for all children and children with siblings

3.4 Discussion

The examination of causal graphs to child's wealth in China reveals that there is a considerable scope of their application to estimate nutritional gender gap in childhood. The purpose of this chapter was to examine the applicability and perform a comparison of different methods in the context of a relevant gender gap issue. A comparison has been conducted between standard statistical tools such as measure of associations and ordinal regression, and causal diagrams framework.

From this exploratory analysis emerges that a relationship between child's gender and child's BMI exists, as confirmed by measure of associations as well as causal graphs, not from ordinal regression. Anyway the nature of this relationships is different depending on the used method. Indeed Cramer's V, for instance, says us that there is a slight positive dependence between child's gender and its nutritional status, but we don't know if such relationship is due to an actual gender gap or to an effect of a common factor acting on both

of them. Conversely causal graph approach, not only informs us that there is a cause-effect relationship between gender and child's nutrition outcome, but even detects the structure of the whole causal mechanism, providing information about the action of confounders and the presence of colliders.
More in details, it is widely recognized in literature that in China birth's order is a crucial factor interacting with gender in the resources allocation within households. Anyway, standard statistical methods are not able to reveal the actual role of birth's order in the child's gender-nutrition relationship; measures of associations, for instance, involve only pairs of variables at a time, while ordinal regression provides the effect of each predictor, in terms of log odds, solely on the dependent variable. On the contrary, causal graphs are capable to detect and graphically represent how such factors, called confounders, act and interact with other variables in the causal process. This clearly emerged in our application, deepening the differences existing among only child and children with siblings. For the latter, indeed, we found that birth's order acts as a confounder in the nutritional effects of child's gender thus an adjustment on the estimate of gender gap has been required.
Causal graphs, in the specific context of gender issues, represent hence an effective tool able, not only to detect gender gaps, but also to uncover the latent causal process originating disparities among females and males.

Causal networks allowed even for providing a measure of the gender gap, through the interventional calculus. Pearl's do-operator does not find an equivalent in standard statistics as it simulates physical interventions by deleting certain functions from the model, replacing them with a constant. Anyway, to obtain a better validation of the methods, future research should aim at empirical evidence from simulation, allowing that the "true" causal structure as well as the "true" causal effect would be known a priori, and the quality and size of results controlled.

Conclusions and Research Perspectives

In this thesis we focused on causal graph-based approach to solve two distinct gendered issues under a “causal” perspective. We aimed:

1. to uncover the causal structure among variables selected for statistically assessing gender disparities;

2. to provide a new measurement of gender gap, in terms of causal effect of variable gender on a target outcome.

For what concerns the first aim, we developed a translation device between the language of gender studies and the language of causality. It consisted, firstly, in dropping standard causal terminology, as the concepts of “treatment”, “exposition”, “confounder” etc., under a gender perspective. Then, it involved the development of *ad hoc* graphical tool able to catch gender equality/inequality occurrences from observational data. These occurrences are detected in the causal graph as absence/presence of direct edges connecting the nodes corresponding to variable gender and to target outcome, given all of the other variables.
Compared to traditional statistical tools, such approach would allow to identify upstream, the causal relationships acting among variables, than, downstream, how such causal mechanisms would translate in gender gap. All these complex cause-effect relationships are, in fact, synthesized in a graph, depicting both direct than indirect relationships.

An avenue of further research in this direction, would certainly concern the development of a graphical tool allowing for variables selection. Indeed, in case of numerous variables, the corresponding causal relationships depicted in the graph may be very intricate and the identification of direct and indirect causes affecting the target outcome, could be difficult. An excess of variables is exactly what might be behind the failure of the constrain-based algorithms in our application, as discussed in subsection 3.3.1. A method taking into account the variable's count of descendants, as well as the number of direct edges between the variable and the target outcome, would thus allow for selecting factors originating gender gap.

Regarding the second purpose, we presented a "causal" interpretation of gender gap, in terms of *average causal effect* of changing (intervening on) the gender variable from female to male level on the target outcome. In order to specify how the resulting distribution of outcome would change in response to intervention on gender, we appealed to Pearl's do-calculus. Anyway, in the applicative context, an automated procedure for computing the estimate of the gender gap still does not exist; indeed the actual computation has been carried on by calculating manually the conditional expectation of female to male ratio.

Future research should involve the development of an algorithm for estimating the causal effects when we deal with categorical variables. This would be an important advancement, since it would allow the systematic estimate of gender gap without the use of a strict distributional assumption, as in IDA's case (e.g. assuming all variables are Gaussian).

A further interesting direction of research regards counterfactual reasonings as heart of real-time policy analysis; indeed, discovering how the current state of things deviates from the one expected i.e. gender equality, as well as determining what went wrong in a certain planned activity and how could be rectified, are exercises of counterfactual thinking.

In the gender issues' perspective, the evaluation of gendered policies' effect as well as the formulation of future actions for reducing gender gap in economics and society, represent thought of policy analysis to explore possible scenarios. In fact, in order to highlight the impact of a gendered policy involving decision variable X on outcome variable Y, namely the gender gap in a specific field, it is necessary to examine past data and estimate conditional expectation $E[Y|do(X = x)]$, where x is a particular instantiation of X under the policy studied.

However, although gender policies are endogenous in the analysis phase of past data, they become exogenous when we want to predict the potential causal effect of a certain decision; the question about how to manage such matters is still open and in need of future researches.

Bibliography

Agresti, A. (2002). *Categorical data analysis.* John Wiley & Sons.

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, **25**(2), 505–541.

APA (2012). Guidelines for psychological practice with lesbian, gay, and bisexual clients. *http://www.apa.org/pi/lgbt/resources/guidelines.aspx*, **67**(1), 10–42.

Basten, S. (2013). Re-examining the fertility assumptions for pacific asia in the un's 2010 world population prospects. *Barnett Papers in Social Research, University of Oxford Department of Social Policy and Intervention.*

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley New York.

Caligaris, S., Crippa, F., and Mecatti, F. (2013). A narrower perspective? from a global to a developed-countries gender gap index: a gender statistics exercise. *Statistica*, **73**(2), 2–18.

Chen, F., Short, S. E., and Entwisle, B. (2000). The impact of grandparental proximity on maternal childcare in china. *Population Research and Policy Review*, **19**(6), 571–590.

Chickering, d. M. (2002). Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, **2**, 445–498.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, **40**(1), 294–321.

Consonni, G. and Leucari, V. (2001). Model determination for directed acyclic graphs. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **50**(3), 243–256.

Fienberg, S. and Haviland, A. (2003). Discussion of statistics and causal inference: A review. *Test*, **12**, 319–327.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, **29**(2-3), 131–163.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*, **40**(6), 979–1001.

Goldin, C. and Rouse, C. (1997). Orchestrating impartiality: The impact of" blind" auditions on female musicians. Technical report, National Bureau of Economic Research.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, **10**(1), 37–48.

Greiner, D. J. and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, **93**(3), 775–785.

Guilmoto, C. Z. (2009). The sex ratio transition in asia. *Population and Development Review*, **35**(3), 519–549.

Gupta, M. D. (2005). Explaining asia's "missing women": A new look at the data. *Population and development review*, **31**(3), 529–535.

Heckerman, D., Meek, C., and Cooper, G. (1997). A bayesian approach to causal discovery. Technical report, Technical Report MSR-TR-97-05, Microsoft Research.

Hernán, M. A., Hernandez-Diaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, **15**(5), 615–625.

Higham, D. D. J. and Higham, N. J. (2005). *MATLAB guide*. Siam.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, **81**(396), 945–960.

Jiang, Q., Li, Y., Tai, X., and Basten, S. (2013). Effect of children composition on the sex of next birth in the context of low fertility in rural china. *http://www.iussp.org*.

Jiang, Y.-F., Cole, T., Pan, H.-Q., Ju, M.-F., Lin, Z.-F., Dong, X.-Y., and Zhang, L. (2006). Body mass index percentile curves and cut off points for assessment of overweight and obesity in shanghai children. *World Journal of Pediatrics*, **1**, 35–39.

Kalben, B. B. (2000). Why men die younger: causes of mortality differences by sex. *North American Actuarial Journal*, **4**(4), 83–111.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, **8**, 613–636.

Kalisch, M. and Bühlmann, P. (2008). Robustification of the pc-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, **17**(4), 773–789.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, **47**(11), 1–26.

Karwa, V., Slavković, A. B., and Donnell, E. T. (2011). Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams. *The Annals of Applied Statistics*, **5**(2B), 1428–1455.

Kaufman, J. S. (2008). Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example. *Social science & medicine*, **66**(8), 1659–1669.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Oxford University Press, USA.

Li, S., Zhu, C., and Feldman, M. W. (2004). Gender differences in child survival in contemporary rural china: a county study. *Journal of Biosocial Science*, **36**(1), 83–109.

Li, S., Jiang, Q., Wei, Y., Feldman, M., Attane, I., Carlton-Ford, S., Kent, M., Haub, C., Eberstadt, N., Garrett, L., *et al.* (2013). Female child survival in china: past present and prospects for the future. *Sociological Studies of Children and Youth*, **10**(2), 231–255.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, **37**(6A), 3133–3164.

Marchetti, G. M. and Lupparelli, M. (2011). Chain graph models of multivariate regression type for categorical data. *Bernoulli*, **17**(3), 827–844.

Margaritis, D. (2003). *Learning Bayesian network model structure from data*. Ph.D. thesis, University of Pittsburgh.

Mecatti, F., Crippa, F., and Farina, P. (2012). A special gen (d) re of statistics: Roots, development and methodological prospects of gender statistics. *International Statistical Review*, **80**(3), 452–467.

Morgan, S. P., Zhigang, G., and Hayford, S. R. (2009). China's below-replacement fertility: Recent trends and future prospects. *Population and Development Review*, **35**(3), 605–629.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Sage.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pearl, J. (1993). Bayesian analysis in expert systems. comment: Graphical models, causality and intervention. *Statistical Science*, **8**(3), 266–269.

Pearl, J. (1997). The new challenge: From a century of statistics to the age of causation. In *Computing Science and Statistics*, pages 415–423.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ Press.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, **3**, 96–146.

Pearl, J. and Robins, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in artificial intelligence*, volume 11, pages 444–453. Citeseer.

Pearl, J. and Verma, T. S. (1995). A theory of inferred causation. *Studies in Logic and the Foundations of Mathematics*, **134**, 789–811.

Poston, D. L. and Zhang, L. (2009). China's unbalanced sex ratio at birth: How many surplus boys have been born in china since the 1980s? In *Gender Policy and HIV in China*. Springer.

Prashant Kumar Singh, S. P. (2013). Sibling composition and child malnutrition in south asia, 1992-2007. In *International Union for the Scientific Study of Population*.

Richardson, T. and Spirtes, P. (2002). Ancestral graph markov models. *The Annals of Statistics*, **30**(4), 962–1030.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern epidemiology*. Wolters Kluwer Health.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, **66**(5), 688.

Scheines, R. (1997). An introduction to causal inference. *Department of Philosophy*, **430**.

Scutari, M. (2010). Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, **35**(i03).

Shang, L., Xu, Y.-y., Jiang, X., and Hou, R.-l. (2005). Body mass index reference curves for children aged 0-18 years in shaanxi, china. *International journal of biomedical science: IJBS*, **1**(1), 57.

Shipley, B. (2002). *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference*. Cambridge: Cambridge University Press.

Słoczyński, T. (2013). Population average gender effects. Technical report, Discussion Paper Series, Forschungsinstitut zur Zukunft der Arbeit.

Spirtes, P. (1996). Using d-separation to calculate zero partial correlations in linear models with correlated errors. Technical report, Carnegie Mellon.

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation Prediction & Search 2e*. MIT press.

Tian, J., Paz, A., and Pearl, J. (1998). *Finding minimal d-separators*. Computer Science Department, University of California.

UNESCO (2011). Priority gender equality guidelines. *UNESCO Publications*, page http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/BSP/GENDER/GE

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*.

Wang, F. (2012). Family planing policy in china: Measurement and impact on fertility. *Working paper*.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

WHO, E. C. (2004). Appropriate body-mass index for asian populations and its implications for policy and intervention strategies. *Lancet*, **363**(9403), 157.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, **20**(7), 557–585.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, **5**(3), 161–215.

Yi, Z., Ping, T., Baochang, G., Yi, X., Bohua, L., and Yongpiing, L. (1993). Causes and implications of the recent increase in the reported sex ratio at birth in china. *Population and Development Review*, **19**, 283–302.

Yuan, X. and Shi, H. (2005). Abnormal high sex ratio at birth and the family planning policy in china (in chinese). *Population Research*, **29**, 11–17.

Zhang, J. (2006). *Causal inference and reasoning in causally insufficient systems*. Ph.D. thesis, PhD thesis, Carnegie Mellon University.

Zhang, J. (2008a). Causal reasoning with ancestral graphs. *The Journal of Machine Learning Research*, **9**, 1437–1474.

Zhang, J. (2008b). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, **172**(16), 1873–1896.

Zhou, B. (2002). Effect of body mass index on all-cause mortality and incidence of cardiovascular diseases–report for meta-analysis of prospective studies open optimal cut-off points of body mass index in chinese adults. *Biomedical and environmental sciences: BES*, **15**(3), 245.