Exchangeability and Invariance: A Causal Theory

Jiji Zhang

(Very Preliminary Draft)

1. Motivation: Lindley-Novick's Puzzle

In their seminal paper on the role of exchangeability in statistical inference, D. Lindley and M. Novick (henceforth L-N 1981) presented an interesting problem based on Simpson's paradox. Consider the following simple three-way contingency table:

	Y=1	Y=0		Y=1	Y=0
X=1	18	12		2	8
X=0	7	3		9	21
	Z=1			Z=0	

Despite the small numbers that appear in the table, L-N invited us to expand the sample size by whatever proportion to make it a large-sample scenario – accidental correlation due to sample variation is not an issue here. Calculate the relative frequencies, and we find a standard case of Simpson's reversal: $P(Y=1 \mid X=1) > P(Y=1 \mid X=0)$, but $P(Y=1 \mid X=1, Z=0) < P(Y=1 \mid X=0, Z=0)$ and $P(Y=1 \mid X=1, Z=1) < P(Y=1 \mid X=0, Z=1)$. In other words, X and Y are positively correlated in the whole population, but, when the population is partitioned by Z, are negatively correlated in each of the two subpopulations. Now consider a new, similar unit whose Z attribute is unknown. Suppose we prefer $Y=1$ to $Y=0$ for the new unit, and we can choose to make its X attribute 1 or 0. What is more effective: to choose $X = 1$ or to choose $X = 0$?

The data set alone does not determine an answer to the question. L-N nicely illustrated this point by two interpretations of the events. The first scenario is a medical experiment. X represents the treatment status of a patient with two values: treated ($X=1$) or not treated ($X=0$); Y represents the outcome with two values: recover ($Y=1$) or fail to recover ($Y=0$); Z represents the sex of a patient: male ($Z=1$) or female ($Z=0$). In this scenario, given the data set, it is quite intuitive that one should not give the treatment to a new patient.

The second scenario is an agricultural experiment, in which X represents the variety of a plant, with two values: white ($X=1$) or black ($X=0$), Y represents the yield, either high ($Y=1$) or low ($Y=0$); and Z represents the height of a plant, either tall ($Z=1$) or short ($Z=0$). Contrary to the first one, our intuition in this second scenario seems to favor planting the white variety, the analogue of giving the treatment in the first scenario. So the intuitively correct answer differs in the two scenarios even though the two data sets are completely analogous.

L-N appealed to de Finetti's notion of exchangeability to explain the difference: in the second, agricultural case, the new unit is (or should be judged to be) exchangeable with the sampled units in variable *Y* given variable *X*, but it is not so in the first, medical case. This difference in exchangeability judgment matters because it leads to different assessments of the probabilities of possible values of *Y* in the new unit conditional on different values of *X*: $p(Y_{new}=1 \mid X_{new}=1)$ is less than $p(Y_{new}=1 \mid X_{new}=0)$ in the first scenario, but is greater than $p(Y_{new}=1 \mid X_{new}=0)$ in the second. (We will see more details in Section 2.) Accordingly, $X_{new}=0$ is the right choice in the first case, but $X_{new}=1$ the right one in the second.

L-N's discussion contained only a few remarks on causation. They deliberately avoided the language of cause and effect in their official story. To people who are not as averse to causal talk, however, the difference in causal structure between the two scenarios seems more fundamental. As commented in C. Meek and C. Glymour (1994), and concurred by J. Pearl (2000), "the terminology [of exchangeability] seems to us only to suggest that judgments of exchangeability are related, in a way that remains to be clarified, to judgments about uniformity of causal structure, and that an explicit account of the interaction of causal beliefs and probabilities is necessary to understand when exchangeability should and should not be assumed." (Meek and Glymour 1994, pp. 1013, footnote 19)

The present paper seeks to provide such an explicit account that links exchangeability judgments to causal beliefs, an account Meek and Glymour probably would endorse. For the account will be developed by way of analogy to a causal theory of "invariance under intervention" developed by P. Spirtes, C. Glymour and R. Scheines (henceforth S-G-S, 1993), which Meek and Glymour drew upon in their diagnosis of L-N's two scenarios, and which also formed a basis of Pearl (1995)'s celebrated intervention calculus. The approach is motivated by the observation that L-N's procedure for identifying relevant probabilities via exchangeability judgments is in close parallel to S-G-S' method for calculating what they call *post-intervention* probabilities. Let us start with this parallel.

2. Two Resolutions of the L-N Puzzle

For L-N, statistical inference is essentially "a procedure whereby one passes from data on a set of units to statements about a further unit". The output statements of the procedure are of course probabilistic in character, and the most relevant concept of probability here, from L-N's perspective, is personal probability. For instance, in the above scenario of medical experiment, one is interested, among other things, in whether a new patient will recover if he or she receives the treatment, based on observations on a number of patients. The relevant probability is *p*(*new patient will recover* | *treatment is given; data on other patients*), representing the degree of credence a person would (or should) have about the recovery of the new patient given the information that the treatment is given, and the information contained in the data about other units. Following L-N, we may suppress the reference to data (and for that matter, reference to other background information), and simply write *p*(*new patient will recover* | *treatment is given*), or $p(Y_{new}=1 \mid X_{new}=1)$ in the notation from Section 1.

In this framework, exchangeability judgments play an important role because they establish a close tie between personal probabilities and (large-sample) relative frequencies, thanks to de Finetti's profound representation theorem and extensions thereof. Let <u_1, …, u_m> denote a sequence of units, e.g., patients in the medical scenario and plants in the agricultural scenario. For a random variable X (or a set of variables $\mathbf{X}$), let X_i ($\mathbf{X}_i$) denote the realization of X ($\mathbf{X}$) on unit u_i. Here are definitions of exchangeability adapted from L-N's article:

Definition 1 (Exchangeability) The *m* units <u_1, …, u_m> are (judged to be) exchangeable in **X** if for every permutation π of the units,

$$p(\mathbf{X}_1, \ldots, \mathbf{X}_m) = p(\mathbf{X}_{\pi(1)}, \ldots, \mathbf{X}_{\pi(m)})$$

Definition 2 (Conditional Exchangeability) The *m* units <u_1, …, u_m> are (judged to be) exchangeable in **X** conditional on **Y** if for every value setting **y** of **Y** and every permutation π of the units,

$$p(\mathbf{X}_1, \ldots, \mathbf{X}_m \mid \mathbf{Y}_i = \mathbf{y}, \text{ for all } i) = p(\mathbf{X}_{\pi(1)}, \ldots, \mathbf{X}_{\pi(m)} \mid \mathbf{Y}_i = \mathbf{y}, \text{ for all } i)$$

Definition 3 (Strong Conditional Exchangeability) The *m* units <u_1, …, u_m> are (judged to be) strongly exchangeable in **X** conditional on **Y** if for every permutation π of the units,

$$p(\mathbf{X}_1=\mathbf{x}_1,\ldots, \mathbf{X}_m=\mathbf{x}_m \mid \mathbf{Y}_1=\mathbf{y}_1,\ldots, \mathbf{Y}_m=\mathbf{y}_m) = p(\mathbf{X}_{\pi(1)}=\mathbf{x}_1,\ldots, \mathbf{X}_{\pi(m)}=\mathbf{x}_m \mid \mathbf{Y}_{\pi(1)}=\mathbf{y}_1,\ldots, \mathbf{Y}_{\pi(m)}=\mathbf{y}_m)$$

The notion of (what is here called) strong conditional exchangeability is discussed only in the appendix of L-N's article, but will play an important role in the present paper. Obviously strong conditional exchangeability implies conditional exchangeability --- the latter considers only situations where the conditioning variables take the same value in all units. An important consequence of this difference is that exchangeability in **Y** together with conditional exchangeability in **X** given **Y** do not imply (joint) exchangeability in **<X, Y>**, but exchangeability in **Y** plus strong conditional exchangeability in **X** given **Y** do imply --- and are actually equivalent to --- (joint) exchangeability in <**X**, **Y**>.

The important role of exchangeability judgments is this: if *n* units in the sample and one further unit are judged to be exchangeable in X, then when *n* is sufficiently large, the personal probability of a possible value of X in the new unit (given the data) can, for practical purposes, be equated with the relevant frequency of that value of X in the *n* units. Following L-N, we write this as $p(X = x) = P(X = x)$, where capitalized P is used to denote relative frequency in the data (or propensity as L-N called it). This link between personal probability and relative frequency extends to conditional exchangeability as well: when the units in the data and one further unit are exchangeable in X conditional on Y, we have $p(X=x \mid Y=y) = P(X = x \mid Y=y)$, provided that the number of units with $Y=y$ in the sample is sufficiently large. Like L-N, we shall not concern ourselves here with the formal demonstration of the link between personal probability and relative frequency,[1] or issues that arise in small sample.

Back to the two hypothetical experiments. L-N's analysis of the medical experiment goes as follows. A plausible judgment is that the new unit is exchangeable with the sampled units in outcome (Y) conditional on treatment (X) and sex (Z). Hence $p(Y \mid X, Z) = P(Y \mid X, Z)$. Moreover, since the sex of the new patient is unknown (and can't possibly be affected by the treatment), it is independent of the treatment: $p(Z \mid X) = p(Z)$. Finally, the new patient is exchangeable with the sampled units in sex, which implies that $p(Z) = P(Z)$. It follows that:

$$\begin{aligned} p(Y=1|X=1) &= p(Y=1 \mid X=1, Z=1)\, p(Z=1|X=1) + p(Y=1 \mid X=1, Z=0)\, p(Z=0|X=1) \\ &= p(Y=1 \mid X=1, Z=1)\, p(Z=1) + p(Y=1 \mid X=1, Z=0)\, p(Z=0) \\ &= P(Y=1 \mid X=1, Z=1)\, P(Z=1) + P(Y=1 \mid X=1, Z=0)\, P(Z=0) \\ &= 0.4 \end{aligned}$$

$$\begin{aligned} p(Y=1|X=0) &= p(Y=1 \mid X=0, Z=1)\, p(Z=1|X=0) + p(Y=1 \mid X=0, Z=0)\, p(Z=0|X=0) \\ &= p(Y=1 \mid X=0, Z=1)\, p(Z=1) + p(Y=1 \mid X=0, Z=0)\, p(Z=0) \\ &= P(Y=1 \mid X=0, Z=1)\, P(Z=1) + P(Y=1 \mid X=0, Z=0)\, P(Z=0) \\ &= 0.5 \end{aligned}$$

So, for the new patient with unknown sex, the probability of recovery given treatment is less than the probability of recovery without treatment, and treatment should be withheld.

By contrast, if the scenario is the agricultural experiment, a plausible judgment is that the new unit is exchangeable with the sampled units in yield (Y) conditional on variety (X). So immediately we get:

$$\begin{aligned} p(Y = 1 \mid X = 1) = P(Y=1 \mid X=1) = 0.5 \\ p(Y = 1 \mid X = 1) = P(Y=1 \mid X=1) = 0.4 \end{aligned}$$

Thus $X = 1$ is preferred to $X = 0$ for the new unit in this case --- a different conclusion than the one reached in the medical case. Exchangeability judgments underlie this difference, from L-N's point of view.

However, from the perspective of authors like Meek and Glymour (1994) and Pearl (2000), exchangeability is a rather elusive and unnecessary notion to invoke here. For them, a crucially relevant distinction is that between conditioning and intervening, between standard conditional probability (propensity in L-N's term) and probability (propensity) resulting from an intervention. In the cases under discussion, one can estimate conditional probabilities from the data, but what we need is the probability of Y=1 given that X is *intervened* to take a value. The latter in general is not determinable from the data alone, but depends on the underlying structure of the data generating process. Identification of such post-intervention probabilities from pre-intervention

[1] Regazzini (1996) contains an explicit statement of the relationship between probability and relative frequency when infinite exchangeability can be assumed. Diaconis (1978) gave a nice account of the failure of de Finetti's theorem when only finite exchangeability can be assumed, as well as the finite forms of the representation theorem. For simplicity we will only consider categorical variables in this paper.

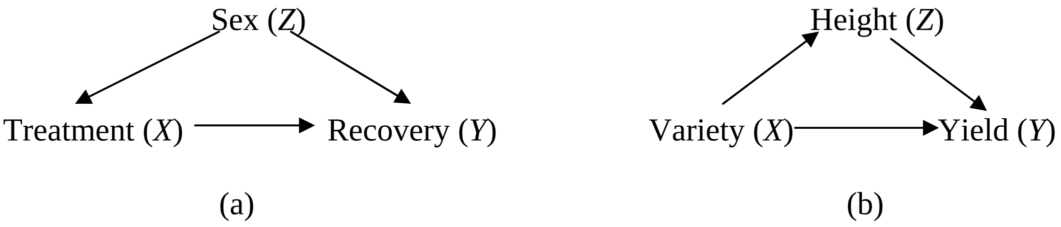
probabilities together with information about causal structure has been studied by quite a few authors in the last two decades (Robins 1988, Spirtes et al. 1993, Pearl 2000, Dawid 2002). We will focus on S-G-S' work in this paper, which was what Meek and Glymour relied upon.

An important notion in S-G-S' work is what they call *invariance* --- invariance under interventions. We will see more details in the next section, but roughly the probability of an event (or conditional event) is termed invariant under an intervention if it has the same value before and after the intervention. Invariant probabilities play a pivotal role in S-G-S' method for identifying post-intervention probabilities, in which the general strategy is to search for an expression of a post-intervention probability in question in terms of invariant probabilities, because invariant post-intervention probabilities are easily identified as the corresponding pre-intervention probability.

For example, as Pearl (2000) forcefully argued, a plausible causal structure for the medical case is depicted in Figure 1a, and a plausible causal structure for the agricultural case is depicted in Figure 1b. Using S-G-S' criterion introduced below, the causal structure in Figure 1b, but not that in Figure 1a, implies that $P(Y|X)$ is invariant under an intervention of X, and hence

$$P_{X:=1}(Y=1) = P_{X:=1}(Y=1|X=1) = P(Y=1|X=1) = 0.5$$
$$P_{X:=0}(Y=1) = P_{X:=0}(Y=1|X=0) = P(Y=1|X=0) = 0.4$$

where $P_{X:=x}(\bullet)$ denotes the post-intervention probability function given that X is intervened to have value x. This is why, according to the causal theorists, choosing X=1 (or better X:=1) for the new unit is preferred to choosing X=0 in the agricultural case.



(a)

(b)

Figure 1

On the other hand, the causal structure in Figure 1a implies that $P(Y \mid X, Z)$ and $P(Z)$ are invariant under intervention of X (and also that X and Z become independent upon an intervention of X). It is then easy to derive the following:

$$\begin{aligned} & P_{X:=1}(Y=1) = P_{X:=1}(Y=1|X=1) \\ &= P_{X:=1}(Y=1 \mid X=1, Z=1)\, P_{X:=1}(Z=1 \mid X=1) + P_{X:=1}(Y=1 \mid X=1, Z=0)\, P_{X:=1}(Z=0 \mid X=1) \\ &= P_{X:=1}(Y=1 \mid X=1, Z=1)\, P_{X:=1}(Z=1) + P_{X:=1}(Y=1 \mid X=1, Z=0)\, P_{X:=1}(Z=0) \\ &= P(Y=1 \mid X=1, Z=1)\, P(Z=1) + P(Y=1 \mid X=1, Z=0)\, P(Z=0) \\ &= 0.4 \end{aligned}$$

$$\begin{aligned}
&P_{X:=0}(Y=1) = P_{X:=0}(Y=1|X=0)\\
&= P_{X:=0}(Y=1 \mid X=0, Z=1)\, P_{X:=0}(Z=1 \mid X=0) + P_{X:=0}(Y=1 \mid X=0, Z=0)\, P_{X:=0}(Z=0 \mid X=0)\\
&= P_{X:=0}(Y=1 \mid X=0, Z=1)\, P_{X:=0}(Z=1) + P_{X:=0}(Y=1 \mid X=0, Z=0)\, P_{X:=0}(Z=0)\\
&= P(Y=1 \mid X=0, Z=1)\, P(Z=1) + P(Y=1 \mid X=0, Z=0)\, P(Z=0)\\
&= 0.5
\end{aligned}$$

Hence the preference for choosing X=0 (or better X:=0). The similarity of the causal reasoning to the previous reasoning by way of exchangeability is obvious. Both try to find an expression of the probability of interest in terms of "manageable" pieces. In L-N's setting, a "manageable" personal probability is one that can be equated to the relative frequency in the data set due to the (conditional) exchangeability of the new unit with the already sampled units. In S-G-S' setting, a "manageable" post-intervention probability is one that is *invariant* in the sense that it can be equated to the corresponding pre-intervention probability.

This parallel suggests that we may develop an explicit causal account of exchangeability analogous to S-G-S' general theory of invariance. Let us pay a closer look at the latter.

3. S-G-S' Graphical Theory of Invariance

The theory is based on a graphical representation of causal structure, as employed in Figure 1. Here are some basic graph theoretical notions. A *directed graph* is a pair <**V, E**>, where **V** is a set of vertices and **E** is a set of arrows. An arrow is an ordered pair of vertices, <X, Y>, represented by $X \rightarrow Y$. Given a graph G(**V, E)**, if <X, Y>$\in$ **E**, then X and Y are said to be *adjacent*, and X is called a *parent* of Y, and Y a *child* of X. The set of X's parents in G is usually denoted by $\mathbf{PA}_G(X)$, or simply $\mathbf{PA}(X)$ when the reference graph is clear. A *path* in G is a sequence of distinct vertices <V_1, ..., V_n>, such that for all $0 \leq i \leq n-1$, V_i and V_{i+1} are adjacent in G. A *directed path* in G from X to Y is a sequence of distinct vertices <V_1,...,V_n>, such that $V_1=X$, $V_n=Y$ and for all $0 \leq i \leq n-1$, V_i is a parent of V_{i+1} in G, i.e., all arrows on the path point in the same direction. X is called an *ancestor* of Y, and Y a *descendant* of X if X=Y or there is a directed path from X to Y. Finally, *directed acyclic graphs* (DAGs) are directed graphs in which there are no directed cycles, or in other words, there are no two distinct vertices in the graph that are ancestors of each other.

DAGs are widely used in statistics and artificial intelligence for multivariate analysis, in which the vertices in the graph represent random variables, and (absence of) arrows in the graph encode statistical independence by the so-called Markov property. On the other hand, it seems natural to interpret DAGs causally: an arrow from X to Y means that X is a direct cause of Y relative to the set of variables in the graph (Spirtes et al. 1993). S-G-S did not attempt to further analyze the notion of "direct cause", but postulated that a causal DAG so interpreted constraints the joint probability in the following way:

Causal Markov Condition: For a set of variables **V** whose causal structure is properly represented by a DAG G[2], the joint probability distribution of **V** factorizes as follows:

$$P(\mathbf{V}) = \prod_{X\in\mathbf{V}} P(X \mid \mathbf{PA}_G (X))$$

The factorization condition is a familiar one in the statistics literature, known there as the Markov property that gives the probabilistic interpretation of DAGs. The causal Markov condition basically says that the causal interpretation and the probabilistic interpretation of DAGs are consistent. The formulation given here is equivalent to the better known version in the philosophical literature: every variable is independent of its non-effects (non-descendants in the causal DAG) conditional on its direct causes (parents in the causal DAG).

S-G-S frequently referred to population distributions in their work, indicating a frequency view of probability. The causal Markov condition, however, can also be cast into a subjectivist form involving personal probabilities, as Meek and Glymour discussed in their article. The account of exchangeability developed in the next section will build on the subjectivist version.

As already intimated, information about causal structure is useful for estimating consequences of interventions. To facilitate formal treatment, interventions are taken to be ideal in the sense that they are effective and local (with no side effects). When we talk about an intervention of a variable *X*, we mean, among other things, that the *direct* target of the intervention is *X*. *Effectiveness* means that the value of *X* or the probability distribution of *X* --- and if the intervention is supposed to depend on some other variables, what variables to depend upon --- are completely fixed by the intervention. Since the intended effect of an intervention on its direct target is usually known, the assumption of effectiveness immediately gives us the post-intervention (conditional) probability of *X*. *Locality*, on the other hand, requires that the intervention should not *directly* affect any variable other than the direct target, and more importantly, local mechanisms for other variables should remain the same as before the intervention.

A formal implementation of these two requirements is given by econometricians, most notably Strotz and Wold (1960) and Fisher (1970), and is nicely recounted in Pearl (2000). A causal system represented by a DAG can also be represented by a set of (recursive) structural equations, in which each variable is equated with a function of its parents in the DAG and an error term. The equations are ``structural" in that they represent mechanisms with causal direction that do not admit ordinary algebraic transformations. Then an effective and local intervention on *X* can be simply implemented by replacing the original equation that defines the mechanism for *X* with a new equation introduced by the intervention and -- this is the important part -- leaving all the other equations unchanged. For example, in the simplest case where *X* is intervened to

[2] At least two things are meant by this relatively vague expression: (1) there is no feedback mechanism; and (2) every direct common cause of any two variables in **V** is also included in **V**, in which case **V** is termed *causally sufficient* by S-G-S. We will assume throughout that we are working with a set of variables whose causal structure can be properly represented by a DAG.

take a fixed value, the equation for *X* is simply ``wiped out" (or replaced by an uninteresting equation *X*=*x*), but all other equations remain the same.

Graphically, the above operation amounts to erasing all arrows into *X* in the original causal DAG, and possibly putting some of them back --- arrows from those of *X*'s parents that are conditioned on in the intervention. That is to say, an intervention may depend on some direct causes of *X* in the original causal system, but nullify other ones.[3] The resulting graph is referred to as the *manipulated* causal DAG. So the major difference between intervening and ordinary conditioning is that the former usually modifies the causal structure (and when it does not in special cases, it at least modifies the causal parameter, i.e., the local probability of the variable conditional on its causal parents). However, all variables that are not directly intervened upon remain governed by their local mechanisms, which imply that the probabilities of these variables conditional on their respective causal parents remain invariant. This forms the basis of S-G-S' graphical theory of invariance.

One of S-G-S' main insights is that one can model an effective and local intervention on *X* in a causal DAG as an extra direct cause of *X* (with no other connection to the graph). That is, given a causal DAG G over a set of variables **V**, we can introduce an extra variable called policy variable for *X* to represent a (local) intervention of *X*. The policy variable for *X* is simply an (extra) parent of *X* but otherwise not adjacent to any other variables in G. For example, Figure 2a augments the causal structure in Figure 1a with a policy variable for treatment. In general, several variables may be intervened simultaneously, and we will need to include one policy variable for each variable.[4] For a DAG G over **V**, and $\mathbf{X} \subseteq \mathbf{V}$, we refer to the graph with a policy variable added for each variable in **X** the ***X**-Policy-Augmented DAG* of G. For example, Figure 2a depicts the *X*-Policy-Augmented DAG of the DAG in Figure 1a, and Figure 2b depicts the one for the DAG in Figure 1b.

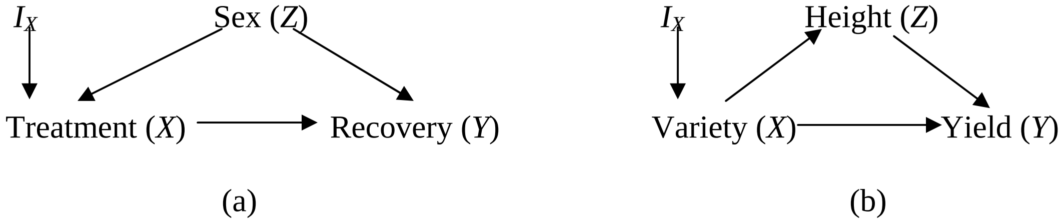


Figure 2

[3] It is S-G-S' restriction that an intervention of *X* can depend, if at all, only on *X*'s causal parents in the original causal DAG, but their theory can in fact carry over to deal with interventions that depend on any number of non-descendants of *X*, or in graphical terms, that add extra arrows into *X* as long as no directed cycles are created. Indeed our analogous account of exchangeability will opt for the more general setting.

[4] This implies that the multiple interventions are independent. We will follow S-G-S to make this simplifying assumption, but it is certainly possible to model correlated interventions in this framework as well.

In general a policy variable can take multiple values, representing different intervention schemes, with one of the values representing null intervention or absence of intervention. The benefit of introducing policy variables is to be able to simulate intervening with conditioning --- conditioning on policy variables. Let G be the causal DAG of a set of variables **V**, and **I** be the set of policy variables added to G to model interventions on $\mathbf{U} \subseteq \mathbf{V}$. We use $\mathbf{i_0}$ to denote the null value of **I** that represents absence of intervention, and so the pre-intervention probability can be written as:

$$\mathrm{P}_{\mathrm{pre}}(\mathbf{V}) = \mathrm{P}\ (\mathbf{V} \mid \mathbf{I} = \mathbf{i_0}) = \prod_{X\in\mathbf{V}} \mathrm{P}_{\mathrm{pre}}(X \mid \mathbf{PA}_{\mathrm{G}}\ (\mathrm{X}))$$

When **I** take a non-null value, say, $\mathbf{i_1}$, some intervention is at work, and the post-intervention probability, according to S-G-S, is the following:

$$\mathrm{P}_{\mathrm{post}}(\mathbf{V}) = \mathrm{P}\ (\mathbf{V} \mid \mathbf{I} = \mathbf{i_1}) = \prod_{X\in\mathbf{U}} \mathrm{P}_{\mathrm{post}}(X \mid \mathbf{PA}_{\mathrm{G}^*}(\mathrm{X}))\prod_{X\in\mathbf{V}\backslash\mathbf{U}} \mathrm{P}_{\mathrm{pre}}(X \mid \mathbf{PA}_{\mathrm{G}}(\mathrm{X}))$$

where G* refers to the manipulated causal DAG under the intervention coded by $\mathbf{i_1}$. Three things are worth noting. First, for $X\in\mathbf{U}$, $\mathrm{P}_{\mathrm{post}}(X \mid \mathbf{PA}_{\mathrm{G}^*}(\mathrm{X}))$ represents the conditional probability of X imposed by the intervention. This reflects the idea that the intervention is effective. Second, for all $X\in\mathbf{V}\backslash\mathbf{U}$, pre-intervention conditional probabilities are used, which reflects the idea of locality. Third, the form of factorization reflects the assumption that the post-intervention probability still satisfies the causal Markov condition, i.e., it factorizes according to the manipulated causal DAG.

S-G-S dubbed the above factorization of post-intervention probability *Manipulation Theorem*[5] (equivalent formulas can be found in Robins 1986 and Pearl 2000). Obviously the causal Markov condition can be regarded as a special case of the manipulation theorem, which obtains in the case of null intervention. We now define S-G-S' notion of invariance:

Definition 4 (Invariance) Let G be the causal DAG for **V**, and consider an intervention of $\mathbf{U} \subseteq \mathbf{V}$. For any two disjoint subsets $\mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, The probability of **Y** conditional on **Z** is said to be *invariant* under the said intervention if $\mathrm{P}_{\mathrm{pre}}(\mathbf{Y} \mid \mathbf{Z}) = \mathrm{P}_{\mathrm{post}}(\mathbf{Y} \mid \mathbf{Z})$.

It is clear from the manipulation theorem (or from the stipulation of the locality of interventions) that for any variable not directly intervened upon, the probability of the variable conditional on its causal parents --- which remain the same after the intervention --- is invariant. This, as already mentioned, is the basis of S-G-S' theory of invariance, whose philosophical significance is also explored in more detail in Woodward (2003). More generally, the device of policy variables gives a simple clue of invariance. Let $\mathbf{I} = \mathbf{i_k}$ represent the intervention in question, we have

[5] I prefer to call it manipulation principle or manipulation postulate, as it is really a statement of two restrictions on interventions: effectiveness and locality plus a postulate of causal Markov condition in the post-intervention population. And as I said in the text, the causal Markov condition can be regarded as a special case of the principle. It seems a little odd to call the latter a theorem while the former is regarded as a postulate. Anyhow I will respect S-G-S' choice of labels in this article.

$$\mathrm{P}_{\mathrm{pre}}(\mathbf{Y} \mid \mathbf{Z}) = \mathrm{P}\,(\mathbf{Y} \mid \mathbf{Z}, \mathbf{I} = \mathbf{i_0})$$
$$\mathrm{P}_{\mathrm{post}}(\mathbf{Y} \mid \mathbf{Z}) = \mathrm{P}\,(\mathbf{Y} \mid \mathbf{Z}, \mathbf{I} = \mathbf{i_k})$$

So if **Y** and **Z** are independent conditional on **I**, then $\mathrm{P}_{\mathrm{pre}}(\mathbf{Y} \mid \mathbf{Z}) = \mathrm{P}_{\mathrm{post}}(\mathbf{Y} \mid \mathbf{Z})$. We can now appeal to an elegant graphical criterion for conditional independence, known as d-separation. Given a path *p* in a DAG, a non-endpoint vertex *V* on *p* is called a ***collider*** if the two edges incident to *V* on *p* are both into *V* (i.e., $\rightarrow V \leftarrow$), otherwise *V* is called a ***non-collider***.

Definition 5 (d-separation) Given a DAG G, a path *p* in G between vertices *A* and *B* is *active* (or *d-connecting*) relative to a set of vertices **Z** ($A,B \notin \mathbf{Z}$) if
(i) every non-collider on *p* is not a member of **Z**; and
(ii) every collider on *p* is an ancestor of some member of **Z**.
A and *B* are said to be ***d-separated*** by **Z** if there is no active path between *A* and *B* relative to **Z**. Two disjoint sets of variables **A** and **B** are d-separated by **Z** if every vertex in **A** and every vertex in **B** are d-separated by **Z.**

Here is a well-known important result (Pearl 1988):

Lemma 1 (Pearl) Let G be a DAG over **V**, and P be the joint probability distribution of **V**. Suppose P satisfies the Markov property of G in the sense that it factorizes according to G. Then for any three disjoint sets of variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, if **Y** and **Z** are d-separated by **X**, then **Y** and **Z** are independent conditional on **X**.

So d-separation in a DAG is a sufficient condition for conditional independence when the probability distribution satisfies the Markov property of the DAG.[6] S-G-S' main result about invariance should appear very natural at this point.[7]

Proposition (Spirtes, Glymour, Scheines) Let G be the causal DAG for **V**, and consider an intervention of $\mathbf{U} \subseteq \mathbf{V}$. For any two disjoint subsets $\mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, the probability of **Y** conditional on **Z** is invariant under the intervention if in the **U**-Policy-Augmented DAG of G, the policy variables are d-separated from **Y** by **Z**.

It is easy to check that this proposition accommodates the basic facts of invariance: invariance of the probability of a variable not intervened upon conditional on its causal parents, because such a variable would be d-separated from the policy variables by its causal parents in the augmented graph. Back to the motivating example, we see that in Figure 2b the policy variable for *X* (variety) is d-separated from *Y* (yield) by *X*. That is why, as we said in the last section, P(*Y*|*X*) is invariant under an intervention of *X* in the agricultural case. By contrast, in Figure 2a, the policy variable for *X* (treatment) is not d-

[6] It is sufficient and necessary for conditional independence under *all* probability measures that satisfy the Markov property.

[7] This version of the theorem makes use of policy variables and augmented graphs. Due to the special topology of policy variables, the proposition can be equivalently formulated without mentioning policy variables or augmented graphs. However, the formulation in terms of policy variables is more intuitive, and is suitable for our present purpose.

separated from Y (recovery) by X (due to the active path $I \rightarrow X \leftarrow Z \rightarrow Y$), and we have to rely on other invariant probabilities to figure out the post-intervention probability of Y.

4. A Parallel Theory of Exchangeability

We are set to articulate Meek and Glymour's hint --- "judgments of exchangeability are related, in a way that remains to be clarified, to judgments about uniformity of causal structure" --- by developing an account of exchangeability parallel to S-G-S' theory of invariance. As we have seen, the basic setting of the problem of intervention is that a causal structure is given, and some local mechanisms of it are known to be disturbed in some way. By analogy, a basic question to be addressed here is the following: suppose we know that the set of sampled units are governed by a common causal structure over a set of attributes **V**. A similar, new unit comes in, whose symmetry with the previous units is known to break down (locally) in some ways --- henceforth we will refer to the content of such knowledge or belief as *symmetry-breaking information*. What are plausible exchangeability judgments given the information?

By analogy to null interventions, consider first the absence of symmetry breaking. That is to say, for all we know, the data generating process for the new unit is the same as the data generating process that governs the previous units. This is a default knowledge situation, deriving from the initial judgment that the new unit is "similar" to the previous units. In such a knowledge situation, where the causal structure for the new unit is judged to be no different than the previous units, it is natural and plausible to judge that the new unit and the sampled units are exchangeable in **V**.

Suppose the common causal structure is (known or believed to be) properly represented by a DAG G. The judgment of exchangeability of all the units in **V** is equivalent to a bunch of judgments of exchangeability and strong conditional exchangeability in a variable given its causal parents, given some postulate like the causal Markov condition. As we have seen, the causal Markov condition is a crucial component in S-G-S' theory of invariance, but in that framework this connection between probability and causal structure is postulated at the population level, and the concept of probability employed is more naturally interpreted along the frequentist lines. Meek and Glymour (1994) present a subjective version of the causal Markov condition, which they argue is also respected by Bayesian theoreticians and practitioners. This version can certainly be applied to a unit: conditional on the proposition that the causal structure over **V** of a unit u can be properly represented by a DAG G, the personal joint probability of **V** should factorize according to G: $p(\mathbf{V}) = \prod_{X \in \mathbf{V}} p(X \mid \mathbf{PA}_{\mathrm{G}}(\mathrm{X}))$.

But this single-unit Markov condition is of no relevance to exchangeability judgments, which concern multiple units. We need to generalize the condition to apply to multiple units, naturally, as follows:

Causal Markov Condition (for multiple units): Let <u_1, u_2, … > be a sequence of units that share a common causal structure among a set of attributes **V**. For any $n \geq 1$, conditional on the proposition that the common causal structure over **V** is properly represented by a DAG G, the personal probability should satisfy the following (again, suppressing the conditioning proposition):

$$p(\mathbf{V}_1, \ldots, \mathbf{V}_n) = \prod_{X\in\mathbf{V}} p(X_1, \ldots, X_n \mid \mathbf{PA}_G(X_1), \ldots, \mathbf{PA}_G(X_n))$$

It is not the right place to delve into a detailed critique of this condition. Suffice it to say that if the single-unit version of the causal Markov condition has force in constraining personal probability, it is *prima facie* plausible that this multiple-unit generalization has force as well. We will need it for the parallel account of exchangeability.

Given this causal Markov condition, the earlier judgment that the new unit and the previous units are exchangeable in **V** is equivalent to the judgment that for every $X\in\mathbf{V}$, the new unit and the previous units are strongly exchangeable in X conditional on X's causal parents $\mathbf{PA}_G(X)$.[8] (When $\mathbf{PA}_G(X) = \varnothing$, it is simply a judgment of exchangeability in X.)

This alternative reading of the exchangeability judgment is of course quite unnecessary in the situation we have being considering, in which no relevant information that breaks the symmetry between the previous units and the new unit is available. It becomes useful when there is symmetry-breaking information concerning specific attributes in **V**, as is the case in L-N's examples. For example, treatment is known to be assigned in a different way for the new patient than for the previous units; or the new patient is known to come from a city with different sex ratio than the city from which the previous patients are drawn; or one deliberately chose a new plant from the white variety rather than flipped a coin to select a variety first, etc. In such situations, it is not reasonable to assume exchangeability in **V** between the new unit and the previous units, but there is no reason to give up the (conditional) exchangeability judgments concerning attributes for which no symmetry-breaking information is available. So in L-N's medical scenario, it is still reasonable to judge that the new patient is exchangeable with the previous patients in *recovery* conditional on *treatment* and *sex*, and also exchangeable with them in *sex*, because the available symmetry-breaking information applies only to *treatment*.

The basic rationale behind all this seems to be precisely what Meek and Glymour hinted: belief of the uniformity in causal structure warrants judgment of exchangeability. When the new unit is believed to share the same causal structure over **V** as the previous units, it is natural and reasonable to believe that the new unit is exchangeable with the previous units in **V**. Likewise, when the new unit is believed to share the same local causal structure that governs the realization of an attribute X as the previous units, it is natural and reasonable to believe that the new unit is (strongly) exchangeable with the previous units in X conditional on its causal parents.

Recall that S-G-S' theory of invariance is built upon the stipulation that under an intervention, the probability of a variable not directly intervened upon conditional on its causal parents is invariant. These basic invariance statements plus the causal Markov condition may imply others, and S-G-S gave a sufficient graphical condition for the

[8] Strong conditional exchangeability is needed here to go the other direction, i.e., to derive the joint exchangeability in **V** from local exchangeability.

implied invariance. The stipulation of the basic invariance statements is of course analogous to our stipulation here that pending symmetry-breaking information concerning an attribute *X*, the new unit is to be judged strongly exchangeable with the previous units in *X* conditional on its causal parents. To complete the analogy, we need a parallel graphical condition.

S-G-S used the device of policy variables to, among other things, signal what variables are directly intervened upon and what are not. And the basic invariance statements apply to variables not directly intervened upon. We can use the same device to signal where the available symmetry-breaking information applies. Call the device *information variables*. More specifically, suppose the common causal structure over **V** for the previous units is properly represented by a DAG G. An information variable for $X \in \mathbf{V}$, I_x, is simply a dummy variable added to G that is a parent of *X* but is not adjacent to any other variable in G. It indicates that *X* is subject to a different local mechanism in the new unit than it is in the previous units. In general we may have such symmetry-breaking information for multiple variables in **V**, and we will add an information variable for each of them to G. Call the resulting graph the *Information-Augmented DAG* of G. We have already stipulated the following basic judgments of exchangeability.

Local Exchangeability Judgments: Given the Information-Augmented DAG of G, for every variable in the graph that is not a child of an information variable, the new unit is strongly exchangeable with the previous units in that variable conditional on its parents.

The (multiple-unit version of the) causal Markov condition formulated above applies to the "null" situation where no symmetry-breaking information is available. That is, if we use **0** to represent the null value setting of the set of information variables **I**, the causal Markov condition applies to $p(\mathbf{V}_1, \ldots, \mathbf{V}_m \mid \mathbf{I} = \mathbf{0})$. This is analogous to saying, in S-G-S' setting, that the causal Markov condition holds of the pre-intervention population. But more is needed to establish their theorem. They need also to assume that the causal Markov condition holds of the post-intervention population, for which the causal structure is represented by a sub-DAG of the original causal DAG. (Recall that they only consider interventions that amount to deleting some arrows in the original causal DAG, but not those that may create new arrows.) Same deal here. We need to assume that the causal Markov condition still holds when symmetry-breaking information is present.

Generalized Causal Markov Condition: for every setting **i** of the information variables,

$$p(\mathbf{V}_1, \ldots, \mathbf{V}_m \mid \mathbf{I} = \mathbf{i}) = \prod_{X \in \mathbf{V}} p(X_1, \ldots, X_m \mid \mathbf{PA}_{\mathrm{G}}(X_1), \ldots, \mathbf{PA}_{\mathrm{G}}(X_m), \mathbf{I} = \mathbf{i})$$

This condition is not reasonable if the symmetry-breaking information gives reason to believe that the causal DAG for the new unit is not a sub-DAG of G, the causal DAG for the previous units. In that case, we need to take the minimal DAG that is a supergraph of both the causal DAG for the new unit and the causal DAG for the previous units. Similar remarks apply to S-G-S' theory of invariance when an intervention does not result in a causal DAG that is a subgraph of the original causal DAG.

From the local exchangeability judgments and the generalized causal Markov condition, we can derive the following proposition, entirely analogous to S-G-S' theorem: *for any two disjoint sets of variables* **X, Y** $\subseteq$ **V**, *if* **X** *is d-separated from the information variables by* **Y** *in the Information-Augmented DAG of G, then the new unit and the previous units should be judged (strongly) exchangeable in* **X** *conditional on* **Y**.

Figure 2a and 2b can be reinterpreted as Information-Augmented DAG of Figure 1a and Figure 1b, respectively. We have already mentioned the exchangeability judgments warranted by Figure 2a. In Figure 2b, {*height*, *yield*} are d-separated from the information variable by *variety,* which agrees well with L-N's judgment that the new unit and the sampled units are exchangeable in *height* and *yield* conditional on *variety*.

It is worth noting that if we just postulate conditional exchangeability instead of strong conditional exchangeability in the local exchangeability judgments, they will not imply more exchangeability judgments, but they may imply more (conditional) probabilities that can be equated with the corresponding relative frequencies. In that case, our graphical condition is sufficient for equating the conditional probability of **X** given **Y** with the frequency distribution of **X** given **Y**.

The situation we have treated so far is relatively simple in that all sampled units are believed to conform to the same causal structure, against which a new unit is considered. More generally, we may consider a bunch of units, whose causal structures are believed to differ from one another in various ways. Suppose each causal structure is properly represented by a DAG, and the union of all the DAGs --- i.e., a graph in which an arrow is present if and only if it is present in one of the DAGs --- is still a DAG. Then one can take the super-DAG, insert information variables where appropriate, and read off exchangeability judgments according to the criterion given above. One important guide for inserting information variables is when a variable does not have a uniform local causal structure across all units, or in other words, when the variable has different causal parents in different units.

5. Conclusion

I have outlined a causal theory of exchangeability, parallel to S-G-S' theory of invariance under intervention. The basic idea stems from Meek and Glymour's hint that judgments about exchangeability are warranted by beliefs of the uniformity of causal structure. Supplemented with a generalized causal Markov condition, we can make use of causal graphs to compactly represent the set of exchangeability judgments implied by the uniformity principle.

The theory outlined above by no means implies that exchangeability judgments can be warranted only by knowledge or belief about full details of causal structures. Indeed, the guiding principle makes it clear that we do not need to know the details of causal structures to judge exchangeability as long as we believe that the units share the same causal structure. More specific beliefs are needed when symmetry-breaking information is present. But even then one need not know the full-blown causal structures to justify some exchangeability judgments. For example, if the only available symmetry-breaking information applies to an attribute *Z* which is believed to have no causal influence on another attribute *W* in any of the units, then the units can be judged to be exchangeable in *W* without any more specific belief or knowledge about causal structures. For *W* must be d-separated from the information variable in the relevant Information-Augmented DAG, whatever it is.

Our use of causal graphs in developing the theory suggests a possible interpretation of the causal graph for a population of units in terms of exchangeability. Roughly, the causal DAG represents relatively stable exchangeability judgments --- stable in the sense that it will not be disturbed by symmetry-breaking information about other variables in the graph. It seems to me a worthwhile project to articulate this interpretation, as well as its connection with the currently standard interventionist interpretation of causal graphs.

References

Dawid, P. (2002) Influence Diagrams for Causal Modelling and Inference. *International Statistical Review* 70: 161-189.

Diaconis, P. (1978) Finite Forms of de Finetti's Theorem on Exchangeability. *Synthèse*, Vol. 36, pp. 271- 281.

Fisher, F.M. (1970) A Correspondence Principle for Simultaneous Equation Models. *Econometrica* 38: 73-92.

Lindley, D. V. and M. Novick (1981) The Role of Exchangeability in Inference. *The Annals of Statistics*, 9 (1), pp. 45-58.

Meek, C., and C. Glymour (1994) Conditioning and Intervening. *British Journal for the Philosophy of Science* 45, pp. 1001-21.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligence Systems.* San Mateo, California: Morgan Kaufmann.

Pearl, J. (2000) Causality: *Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Regazzini, E. (1996) De Finetti’s Reconstruction of the Bayes-Laplace Paradigm, *Erkenntnis*, 45, pp. 159-176.

Robins, J. (1986) A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods -Applications to Control of the Healthy Worker Survivor Effect. *Mathematical Modeling* 7: 1393-1512.

Spirtes, P., C. Glymour, and R. Scheines (1993) *Causation,Prediction and Search.* New York: Springer-Verlag. (2000, 2nd ed.) Cambridge, MA: MIT Press.

Strotz, R.H., and H.A. Wold (1960) Recursive versus Nonrecursive Systems: An Attempt at Synthesis. *Econometrica* 28: 417-427