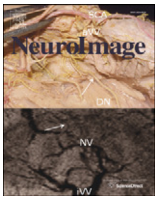
NeuroImage 57 (2011) 323–330

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Comments and Controversies

On meta-analyses of imaging data and the mixture of records[☆]

J.D. Ramsey [a,*], P. Spirtes [a], C. Glymour [a,b]

[a] Department of Philosophy, Carnegie Mellon University, United States
[b] Florida Institute for Human and Machine Cognition, United States

ARTICLE INFO

Article history:
Received 2 April 2010
Revised 24 July 2010
Accepted 27 July 2010
Available online 12 August 2010

ABSTRACT

Neumann et al. (2010) aim to find directed graphical representations of the independence and dependence relations among activities in brain regions by applying a search procedure to merged fMRI activity records from a large number of contrasts obtained under a variety of conditions. To that end, Neumann et al., obtain three graphical models, justifying their search procedure with simulations that find that merging the data sampled from probability distributions characterized by two distinct Bayes net graphs results in a graphical object that combines the edges in the individual graphs. We argue that the graphical objects they obtain cannot be interpreted as representations of conditional independence and dependence relations among localized neural activities; specifically, directed edges and directed pathways in their graphical results may be artifacts of the manner in which separate studies are combined in the meta-analytic procedure. With a larger simulation study, we argue that their simulation results with combined data sets are an artifact of their choice of examples. We provide sufficient conditions and necessary conditions for the merger of two or more probability distributions, each characterized by the Markov equivalence class of a directed acyclic graph, to be describable by a Markov equivalence class whose edges are a union of those for the individual distributions. Contrary to Neumann et al., we argue that the scientific value of searches for network representations from imaging data lies in attempting to characterize large scaled neural mechanisms, and we suggest several alternative strategies for combining data from multiple experiments.

Introduction

As functional magnetic resonance imaging time series from psychological experiments and resting states have become available in increasing number, search procedures adapted from economics and machine learning have been applied to attempt to extract from multiple data sets processing information represented by graphical causal models (e.g., Chen and Herskovitz (2010), Chen et al. (2009), Friston et al. (2003), Gates et al. (2010), Roebroeck et al. (2005), Laird et al. (2008), Marreiros et al. (2009), Ramsey et al. (2010), Rajapakse and Zhou (2007)) Recently, Neumann et al. (2010) have applied a heuristic Bayes net search algorithm to combined samples from a large, diverse collection of fMRI contrasts. The stated aim of the Neumann study is not to identify neural processing mechanisms but simply to characterize the dependence and conditional independence structures of the multiple fMRI studies by a single graphical representation. To that end, Neumann et al., produce three graphical models using an automated search applied to reductions of fMRI data assembled from multiple, diverse studies in BrainMap. They validate their search procedure with simulations that find that merging the data from two distinct Bayes nets results in a graph that combines the edges in the individual Bayes nets.

In what follows, we define relevant notions and describe and analyze the Neumann et al. results. We argue that their procedure yields graphical objects that in general do not result, even in the large sample limit, in a combination of edges from the original directed acyclic graphs (DAGs). We argue that their graphical results cannot be used to characterize statistical dependencies in the individual studies of their meta-analyses. We describe sufficient conditions and necessary conditions for that aim to be possible. We argue that their simulation results are an artifact of their examples and do not generalize. Finally, we question the scientific value of network representations of imaging data that are separated from causal hypotheses, and consider several possible alternative strategies for meta-analysis.

Graphical models and Markov equivalence classes

Directed graphs can be used to represent a family of joint probability distributions on the variables, or hypothetical causal relations among the variables, or both simultaneously. Causal representations are common in fMRI applications, but Neumann et al., explicitly reject such an interpretation, and seek to represent only distributional features.

[☆] This research was supported by the James S. McDonnell Foundation.

[*] Corresponding author. Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, United States.

E-mail address: jdramsey@andrew.cmu.edu (J.D. Ramsey).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.07.065

In the simplest case, such graphs have no directed cycles (they are directed acyclic graphs, or DAGs). Two nodes, X, Y, in a DAG are said to be *adjacent* if $X \rightarrow Y$ *or* $X \leftarrow Y$ occurs in the graph. The vertices of a DAG represent random variables, and depending on the range of values of the variables, a DAG may be associated with a family of probability distributions, for example multinomial or Gaussian distributions, and appropriate parameters (e.g., linear coefficients and disturbance variances). Specification of the parameter values (e.g., linear coefficient values, or conditional probability values) then determines a particular joint distribution on the possible assignments of values to all variables. The topology of the DAG represents a set of restrictions on the joint probability distribution on those variables characterized (in kinship terminology) by a Markov property: *Let X be any variable in DAG G. For every allowed assignment of values to the variables of G, conditional on the values of its parents, X is independent in probability of all variables that are not descendants of X in G.*[1]

In causal terminology: *Conditional on its direct causes as represented in G, X is independent of all variables in G that are not effects of X.*

Equivalently, the joint probability of an allowable assignment of values to the variables is always equal to the product, over all variables, of the probability of the value of each variable conditional on the values assigned to its parents. The graphical topology of a DAG thus encodes the set of conditional independence relations shared by all probability distributions appropriate for the DAG. The conditional independence relations encoded by a DAG can be calculated by purely graphical algorithms; for example, Pearl's (1988) d-separation algorithm. When a DAG captures all and only the conditional independence relations in a distribution, we say the DAG *represents* the distribution, or, equivalently, that the DAG and distribution are *faithful* to one another.

Two directed acyclic graphs (DAGs) are said to be *Markov equivalent* if, by the Markov property, they imply the same set of conditional independence relations among their variables. For example, the graphs $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$ and $X \leftarrow Y \leftarrow Z$ constitute the Markov equivalence class encoding that X is independent of Z conditional on Y, commonly abbreviated: $X \perp\!\!\!\perp Z|Y$. A Markov equivalence class of DAGs is commonly represented by a mixed graph (called a CPDAG by Neumann et al., and often called a *pattern* in the computer science literature) with a directed edge $X \rightarrow Y$ provided $X \rightarrow Y$ occurs in every DAG in the class, and an undirected edge $X–Y$ if $X \rightarrow Y$ occurs in some DAG in the class and $X \leftarrow Y$ also occurs in some DAG in the class. All Markov equivalent graphs have the same adjacencies, but as the example indicates, not necessarily the same directions for all edges. If in all graphs in a Markov equivalence class $X \rightarrow Y$ occurs, and Z is adjacent to Y but not adjacent to X, then the edge joining Y and Z has the same direction in all DAGs in the Markov equivalence class.

Directed graphs faithful to a probability distribution also represent probabilistic dependence relations for some kinds of random variables and parameterizations of the distributions. For example, if there is a directed path from X to Y in a DAG, and, as in the Neumann et al., paper, all variables are binary (in which case the DAG and a specification of the probability distribution has come to be known as a Bayes net), then X and Y are dependent unconditionally. The same is true for linear models with independent noises, but is not true in general for models with categorical variables taking 3 or more values.

The Neumann et al. meta-analyses

Neumann et al. "propose a new exploratory method for the discovery of partially directed functional networks from fMRI meta-analysis data. The method performs structure learning of Bayesian networks in search of directed probabilistic dependencies between brain regions…we infer with our method possible functional interdependencies between brain regions from observational data alone" (p. 1372). A more precise goal can be surmised from their effort to validate their method: Assuming that the functional dependencies of each individual imaging study could be represented by an (unknown) Bayes net, from the combined data infer a graphical object each of whose edges occurs in at least one of the individual Bayes nets, and, so far as possible, includes all such edges. Further, we assume that Neumann et al. intended that the resulting graphical objects would at least preserve some of the statistical interpretability of Bayes nets, for example, that directed paths between variables indicate statistical association of the terminal nodes. In what follows we show that their method cannot reliably achieve these goals, and we describe the special conditions under which the goals are feasible. In particular, we show that their method cannot distinguish, on the one hand, between edges and pathways that are due to functional associations of brain areas and, on the other hand, edges and pathways that are artifacts of pooling studies from different subjects under different experimental conditions.

Neumann et al. carry out a search for graphical representations separately on four groups of fMRI data with variables from several brain regions. The number of regions in a group varies from 5 to 10. The data are extracted from BrainMap and represent a sample from 2050 fMRI contrasts. Each case specifies a value (active/not active) for the regions. 100 samples are drawn from the data for each group, with sample sizes varying from 196 to 569 for the respective groups, and their search procedure is applied. An edge is postulated if it occurs in more than 50% of the trials.

Neumann et al., use a heuristic iterative search strategy for graphs, starting with an empty graph and adding the edge that most increases the posterior probability of the model, estimated via a Monte Carlo Markov Chain procedure. Their resulting graphs are shown in Figs. 1–3 (their Figs. 10–12).

Dotted edges represent associations that do not quite meet the 50% criterion. About the undirected edges, for example in Fig. 1, Neumann et al. write that: "the directionality between rdPMC and cerebellum bilaterally could not be determined due to graph equivalence" (Neumann et al. (2010), p. 1381). On each sampled data set, their search algorithm always produces a DAG, which they convert

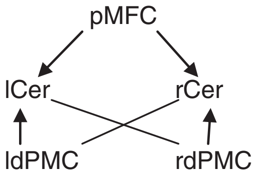


Fig. 1. Neumann et al.'s Fig. 10. See text for description.

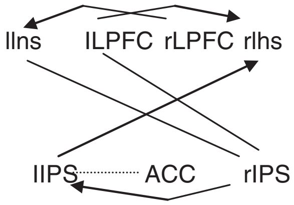


Fig. 2. Neumann et al.'s Fig. 11. See text for description.

[1] Independence of random variables is defined in terms of independence of measurable subsets of their values. In the Neumann case, the variables range over two values, {0,1}, and all subsets of values are measurable. Independence of variables X and Y thus *means* $pr(X=0, Y=0)=pr(X=0)pr(Y=0)$, and $pr(X=1, Y=0)=pr(X=1) pr(Y=0)$; etc. Conditional independence of variables similarly means, by definition, conditional independence for all measurable sets of values of the variables. Thus, for binary variables, X and Y are independent conditional on Z if and only if $pr(X=0, Y=0|Z=0)=pr(X=0| Z=0)pr(Y=0|Z=0)$, and $pr(X=0, Y=0|Z=1)=pr(X=0| Z=1)pr(Y=0|Z=1)$, etc.

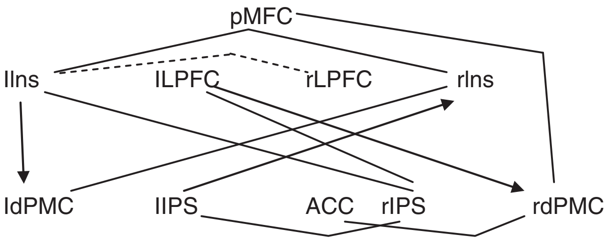


Fig. 3. Neumann et al.'s Fig. 12. See text for description.

to a description—a pattern or a CPDAG—of a Markov equivalence class, which may contain an undirected edge. A directed or undirected edge is postulated if it occurs in at least 50% of the patterns thus obtained.

Neumann et al. warrant their procedure by a simulation study using data from the graphs in Fig. 4 (their Fig. 7) and by a simulation study with data from the two graphs shown in Fig. 5 (their Fig. 6). All variables are binary.

They consider samples of 1000 from a single random parameterization (that is, a specification of the probability tables giving the probability of each value of a variable conditional on each assignment of values to its parents) of each graph, and combine the samples in various proportions. They repeat each experiment with a given proportion 100 times, keeping the parameterization fixed. They infer an edge if it occurs in at least 50% of the trials for a given proportionate combination. They recover either the edges in the graph most frequently represented in the combination, or, when the proportions are equal or nearly equal, they recover the union of the edges of the two graphs. All of the edges of Fig. 5 are recovered by their procedure when the source graphs are equinumerous, and all but one of the edges of Fig. 4.

Neumann et al., do not claim that the graphical objects they obtain with their procedure are always DAGs, or descriptions of Markov equivalence classes of DAGs. That leaves open the question of how to interpret the graphical objects they obtain: what dependence and independence relations are characterized by a graphical objects found by their procedure? We argue that the directed edges and directed paths in the Newmann graphs cannot be interpreted as meaning there are associations of the variables connected by that edge or path in *any* of the individual studies they aggregate, nor, when two variables measured only in different studies are related by a directed edge or directed path in the Newmann graph, does it mean that the activity of the brain ever produces associations of those variables. While Neumann et al. do not give a causal interpretation of their graphs, they surely intended to capture associations produced by brain activity, not artifacts of their aggregation.

Interpreting the graphical objects from the Neumann meta-analyses

None of their empirical results shown in Figs. 1–3 are graphs of Bayes nets, nor do any of them characterize a unique Markov equivalence class. In Markov equivalence classes, if there is an edge $X \rightarrow Y$, and Z is adjacent to Y but not to X, then the Y–Z adjacency must be oriented. Thus, in Fig. 1, both of the undirected edges must be oriented in any Markov equivalence class that agrees with the directed edges and adjacencies. Similarly, in Fig. 2, the rIPS and Iins adjacency must be oriented. In Fig. 3, the IdPMC–rIns adjacency must be oriented; the ACC–rdPMC adjacency must be oriented; and the IIns–rIns adjacency must also be oriented. These facts suggest that their empirical graphs represent, not Markov equivalence classes, but *sets* of Markov equivalence classes of the original DAGs; each such Markov equivalence class can be characterized by giving alternative directions to undirected edges—in other words, alternative sets of conditional independence and conditional dependence relations. That interpretation is not tenable because of a further problem: for binary variables, the existence of a directed path between two variables in a member of a Markov Equivalence class indicates that the variables are associated in every distribution faithful to the DAGs in the Markov equivalence class. In the graphs obtained by the Neumann et al. procedure, directed paths from one variable to another cannot always be understood to imply that the variables are associated. The data Neumann et al. use are assembled from a variety of studies with different stimulus conditions and different subject groups. If in one experiment X and Y are dependent but both are independent of Z, and in another experiment X is independent of Y and Z but Y and Z are dependent, with DAG representations for the respective empirical cases by $X \rightarrow Y$ Z, and X $Y \rightarrow Z$, X and Z will be independent in each experimental case. But their search, if it succeeds as they intend, will produce a graphical object with a directed path $X \rightarrow Y \rightarrow Z$. In other words, directed paths in their graphical objects cannot generally be interpreted to mean that the variables connected by such paths are actually dependent in any brain activity.

Is it at least the case that if the Neumann et al. procedure produces a graph with a directed or undirected edge between two variables, then in principle—say in the large sample limit—the adjacent variables are associated in one of the individual experiments? It is not. With elementary algebra, Yule (1903) showed that, except in special circumstances, combining two or more distinct distributions in each of which two binary variables are independent yields a distribution in which the two variables are dependent.[2] Sufficient conditions for independence in a distribution $p = kp_1 + (1 - k)\ p_2\ (0<k<1)$ formed from joint distributions p_1 and p_2 both defined on random variables X, Y are that $p_1(X=1) = p_2(X=1)$ or that $p_1(Y=1) = p_2(Y=1)$.

Yule's point applies as well to *conditional* independence relations. Let $p(X,Y,Z) = kp_1 + (1 - k)\ p_2$, and let X,Y be independent conditional on Z in both p1 and p2. Then for k not equal to 0 or 1, X, Y are independent conditional on Z in p if and only if $p_1(X|Z)p_1(Y|Z) + p_2(X|Z)p_2(Y|Z) = p_1(X|Z)p_2(Y|Z) + p_2(X|Z)p_1(Y|Z)$. Related considerations hold for categorical variables with more than two values. Yule's point suggests that, except in special cases, meta-analysis searches of the kind Neumann et al. conduct using combined data sets from two or more fMRI time series cannot be correct in the large sample limit.

For Gaussian distributions, vanishing covariances and vanishing correlations are preserved under what Yule called the "mixture of records" when the means of corresponding variables are the same in both distributions, but vanishing partial covariances and partial correlations, which mark conditional independence in Gaussian

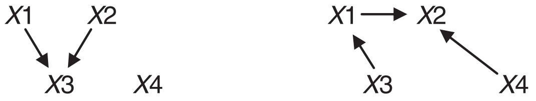


Fig. 4. Unmixed graphs from Neumann et al.'s Fig. 6.

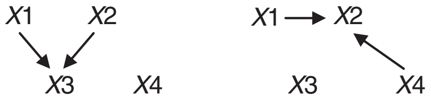


Fig. 5. Unmixed graphs from Neumann et al.'s Fig. 7.

[2] Yule's point is the converse of Simpson's (1951) later but better known "paradox." In Yule's case two variables that are independent conditional on each value of a third variable are nonetheless jointly dependent; in Simpson's two variables that are dependent conditional on each value of a third variable are nonetheless jointly independent. Yule's phenomenon holds for most combinations of probability distributions; Simpson's depends on combining separate distributions in special proportions.

distributions, are not preserved under mixing except under specific conditions.

Yule's argument shows that mixing samples can result in novel dependencies if the sample size is sufficiently large. Search procedures will in such circumstances asymptotically return a superset of edges of the union of the sets of edges in the Bayes nets describing the separate distributions, with consequent loss of information about the orientations of edges. A mixed distribution may not be representable by any DAG, may be representable by a DAG that is not the union of the graphs of the component distributions, and may even be representable by a DAG that reverses edges in a component graph. We give some examples and then consider the case more generally.

Consider the two graphs in Fig. 5, used in the Neumann et al. simulations. The Neumann et al. search returns the union of the edges in the two graphs, but Yule's argument shows that is an artifact of the sample size, the example, and the search procedure. Mixing two distinct distributions for binary variables each represented by one of the graphs in Fig. 5 results in a distribution represented only by a complete graph, in which every pair of edges is adjacent. Such graphs are members of a single Markov equivalence class, and that is what a correct search procedure would find, probability 1, in the large sample limit. For a simpler example with the same point, consider mixing two distinct distributions over binary variables represented by the same graph, $X \rightarrow Y \leftarrow Z$. The result of a correct (for single data sets) search over the combined data is the Markov equivalence class representing all acyclic complete directed graphs on X, Y, and Z. The results of course vary with sample size and the parameter values for the joint distributions. In contrast, if the distributions are Gaussian with zero means, mixing two data sets faithful to $X \rightarrow Y \leftarrow Z$ yields that very graph. But combinations of Gaussian distributions have their own problems. For example, consider two Gaussian distributions with zero means, each respectively represented by one of the graphs in Fig. 6.

In a mixed distribution, X and Z will be independent, but will be dependent conditional on Y, and the mixed distribution will be represented by the graph in Fig. 7.

In sum, the procedure Neumann et al. use does not preserve Markov properties, and the directed edges and directed paths in their graphical output cannot be generally interpreted as marking real associations. We have no principle for reading the statistical content of their output.

When Markov properties are preserved in combined distributions

Consider more generally whether the Markov property holds for the combined distribution of a collection of units if it holds for distinct sub-populations of the collection. The sub-populations can vary in several different ways. They might share the same graph, but have different parameter values and hence different distributions. They might have different graphs, but all be subgraphs of one acyclic supergraph. Or they might have different graphs, and not be subgraphs of one acyclic supergraph. To describe the interactions more generally, represent the parameter for each variable V by an exogenous variable directed into V. Consider the case of a linear model:

$$Y = aX + e_Y.$$

$$X \rightarrow Y \quad Z \qquad X \quad Y \rightarrow Z$$

Fig. 6. Unmixed graphs, in example with Gaussian distributions with zero means.

$$X \rightarrow Y \leftarrow Z$$

Fig. 7. The mixed distribution for the distributions in Fig. 6.

Suppose that for each unit in the population, the value of the coefficient is the value of the random variable $coeff_{X \rightarrow Y}$. Treat the parameter as a regular random variable, and treat the parameters as if they were ordinary causes which, when included in a graphical representation, satisfy the Markov condition. If every system in the population has the same value for the coefficient, then $coeff_{X \rightarrow Y}$ is independent of everything and can be left out of the graph representing the population. On the other hand, if $coeff_{X \rightarrow Y}$ is different for different members of the population, then the graph is: $X \rightarrow Y \leftarrow coeff_{X \rightarrow Y}$. Note that the model as a whole is no longer linear, because $coeff_Y$ multiplies X rather than being added to X. If the value of $coeff_{X \rightarrow Y}$ varies among units but is independent of X and e_Y, then it can also be marginalized out of the graph without affecting the Markov Condition for the non-parameter variables.

Suppose we have a binary model, $X \rightarrow Y$, where $\theta_{Y=0|X=0} = P(Y=0|X=0) = a$ and $\theta_{Y=0|X=1} = P(Y=0|X=1) = b$. Again, if a and b are different for different units in the population, then the graph can be represented as in Fig. 8, assuming that $\theta_{Y=0|X=0}$ and $\theta_{Y=0|X=1}$ are independent of each other and of all of the other non-descendants of Y. Note also that the model as a whole is not binary any more, as $\theta_{Y=0|X=0}$ and $\theta_{Y=0|X=1}$ are continuous.

Suppose now we mix together different populations $\mathrm{Pop}_1, \ldots, \mathrm{Pop}_n$. Suppose in each population that there is a DAG, $G_1, \ldots, G_n$, respectively, and that the union of the DAGs is also a DAG G. G now contains all of $G_1, \ldots, G_n$ — each of these can be considered a specialization of G in which the parameters take on a certain value. For example, if it is a linear model, and G contains $X \rightarrow Y$, but G_1 is a subDAG of G in which the edge $X \rightarrow Y$ is missing, then $coeff_{X \rightarrow Y} = 0$ in G_1. Suppose we mix together populations with DAGs G_1 and G_2, and that G_1 is $X\ Y \rightarrow Z$, but G_2 is $X \rightarrow Y \rightarrow Z$. Suppose that $coeff_{Y \rightarrow Z}$ is the same in Pop_1 and Pop_2. Then $coeff_{X \rightarrow Y} = 0$ in Pop_1 but not in Pop_2. However, $coeff_{X \rightarrow Y}$ and $coeff_{Y \rightarrow Z}$ are not correlated because $coeff_{Y \rightarrow Z}$ is the same for both populations. So if these two populations are mixed, the Markov Assumption is true for the combined population and the combined non-parameter causal graph $X \rightarrow Y \rightarrow Z$. The graph with the parameters included is shown in Fig. 9.

In this graph, the independence of X and Z conditional on Y is preserved. The same reasoning holds when all units in the population have a different coefficient, as long as the coefficients are independent of each other.

If, by contrast, $coeff_{Y \rightarrow Z}$ is different in Pop1 and Pop2, then $coeff_{X \rightarrow Y}$ and $coeff_{Y \rightarrow Z}$ are correlated. So if these two populations are mixed, the Markov Conditions does not hold for the distribution of the combined population and the combined causal graph $X \rightarrow Y \rightarrow Z$. If we include the parameter variables, and assume that the Markov Assumption applies to $coeff_{X \rightarrow Y}$ and $coeff_{Y \rightarrow Z}$, from the fact that they are correlated we can write $coeff_{X \rightarrow Y} \leftrightarrow coeff_{Y \rightarrow Z}$. (The analysis is essentially the same if $coeff_{X \rightarrow Y} \rightarrow coeff_{Y \rightarrow Z}$ or $coeff_{X \rightarrow Y} \leftarrow coeff_{Y \rightarrow Z}$.) It follows that X and Z are dependent conditional on Y, and X is

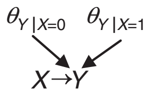


Fig. 8. $X \rightarrow Y$, where X and Y are binary, treating parameters as random variables.

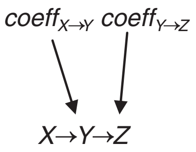


Fig. 9. $X \rightarrow Y \rightarrow Z$, for the linear case, treating parameters as random variables, where the coefficients are independent.

not independent of Z conditional on Y. This arrangement is shown in Fig. 10.

The same analysis applies when we mix together populations with DAGs G_1 and G_2, when that G_1 is $X\ Y \rightarrow Z$, but G_2 is $X \rightarrow Y\ Z$. Then $coeff_{X\rightarrow Y} = 0$ and $coeff_{Y\rightarrow Z} \neq 0$ in Pop_1 and $coeff_{X\rightarrow Y} \neq 0$ and $coeff_{Y\rightarrow Z} = 0$ in Pop_2. In that case $coeff_{X\rightarrow Y}$ and $coeff_{Y\rightarrow Z}$ have to be correlated, and X is not independent of Y conditional on Z, no matter what the proportions of Pop1 and Pop2 in the mixture are (other than the extremes of $k = 0$ or $k = 1$).

Now suppose we mix together four populations with DAGs G_1 through G_4, and that G_1 is $X\ Y \rightarrow Z$, G_2 is $X \rightarrow Y\ Z$, G_3 is $X \rightarrow Y \rightarrow Z$, and G_4 is $X\ Y\ Z$. Then it is possible to find mixing proportions to make $coeff_{X\rightarrow Y}$ and $coeff_{Y\rightarrow Z}$ independent, in which case X is independent of Z conditional on Y in the mixed population, and the Markov Condition applies to the combined graph $X \rightarrow Y \rightarrow Z$. For example, mixing the four sub-populations in equal proportions is a sufficient but unnecessary condition for making $coeff_{X\rightarrow Y}$ and $coeff_{Y\rightarrow Z}$ independent. Arbitrary proportions of each population in the mixture, however, would lead to $coeff_{X\rightarrow Y}$ and $coeff_{Y\rightarrow Z}$ being dependent. The same analysis applies to cases where the combined graph is cyclic.

Suppose we mix just two sub-populations with graphs G_1 and G_2 respectively, $X \rightarrow Y$ and $Y \rightarrow X$, then $coeff_{X\rightarrow Y} \neq 0$ and $coeff_{Y\rightarrow X} = 0$ in Pop1, and $coeff_{X\rightarrow Y} = 0$ and $coeff_{Y\rightarrow X} \neq 0$ in Pop2. Hence if these two populations are mixed together, $coeff_{X\rightarrow Y}$ and $coeff_{Y\rightarrow X}$ are dependent no matter what the non-trivial proportions are. On the other hand, if four populations are mixed together with graphs $X\ Y$, $X \rightarrow Y$, $X \leftarrow Y$, and $X\ Y$, then if the right mixing proportions are chosen, $coeff_{X\rightarrow Y}$ and $coeff_{Y\rightarrow X}$ are independent. If this is the case, the Markov Condition fails for the population distribution because the population as a whole looks like a cyclic system.

Neumann et al., aim to recover the union of subsets of edges of graphical models of binary variables for several data sets, provided the individual graphs do not conflict in the directions of edges. Whether a search on a finite sample from such a mixture will return the union of edges depends on the individual graphs, the parameter values the sample size, and the search procedure. In fMRI meta-analysis, the true graphs are unknown and the parameter values are unknown.

Simulation studies

Neumann et al. warrant their method with simulations of two mixed cases, each involving four variables, as illustrated in Figs. 4 and 5 above. In the first of these two simulation studies, the union of the edges of the two graphs generating the combined samples contains four of six possible adjacencies. Only two false positive adjacency errors are logically possible in the mixed data. To see how their procedure would do when we mix data from structures when there is more possibility for error, and to compare the results on mixed and unmixed data, we conducted our own simulation study.

We generated at random two sets of 10 directed acyclic graphs on 8 vertices with 8 edges. Each graph was parameterized by drawing from a uniform distribution the conditional probability of each variable given a value assignment of its parents. A data set of 1000 cases was then obtained by simulation from each of the 20 parameterized graphs. Over 100 repetitions for each pair of DAGs, a sample with replacement of 500 cases was drawn from the 1000 case sample from each DAG, and the samples combined (Table 1). For each combined sample for each graph pair, a Markov equivalence class search was conducted, and the edges obtained that occurred in 50% or more of the resulting Markov equivalence classes. The average, over all graph pairs of the error rates for adjacencies in each graph pair were calculated, counting an adjacency in the output of the procedure as a false positive if it occurred in the Markov equivalence class of neither of the paired graphs. Likewise, averages of edge error rates for directed and undirected edges were calculated, counting a directed edge in the output as a false positive if it occurred in the Markov equivalence class of neither of the paired graphs, and an undirected edge as a false positive if it occurred in the Markov equivalence class of neither of the paired graphs. The false positive rate for each graph pair was calculated as the percentage of edges in the output that are false positives. False negatives were counted as edges in one or the other paired Markov equivalence classes of DAGs that were not in the output. An edge directed as A → B in the output but as A ← B in both source graphs was counted as a false positive and a false negative. The Markov equivalence class search procedures were also run on 100 samples with replacement from the same samples of 1000 and of 500 for each of the 20 individual DAGs, and >50% rule applied to obtain graphical objects (Table 2). Errors were counted in the same way.

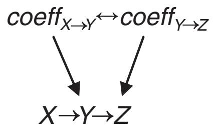

$coeff_{X\rightarrow Y} \leftrightarrow coeff_{Y\rightarrow Z}$
$X \rightarrow Y \rightarrow Z$

Fig. 10. $X \rightarrow Y \rightarrow Z$, for the linear case, treating parameters as random variables, where the coefficients are dependent.

Table 1
$N = 500$, uncombined. Averages of true positive errors, false positive errors, false negative errors, and false negative rate over simulation runs of 500 cases, for the case where datasets are not combined pairwise, as described in the text. 'TP' records the average number of edges in the output graph also contained in the graph generating the data; 'FP' records the average number of edges in the output not contained in the input graph, 'FN' records the average number of edges in the input graph that are not contained in the output graph, and 'FPR' records the average of the number of false positives to number of edges returned. For edge counts, an edge X↔Y is counted as a prediction of both X → Y and Y → X.

	TP	FP	FN	FPR
CPC, directed edge errors	3.35	1.40	4.65	0.31
GES, directed edge errors	3.40	1.20	4.60	0.29
CPC, adjacency errors	4.65	0.00	3.35	0.00
GES, adjacency errors	4.80	0.00	3.20	0.00

The Metropolis–Hasting (MH) search procedure for DAGs and CPDAGs that Neumann et al. used is heuristic and comparatively slow and the scoring and stopping procedure chosen is not specified in their paper. We used instead two correct search procedures, the Conservative PC algorithm (CPC, Ramsey et al., 2006) and the Greedy Equivalence Search (GES, Meek, 1997), implemented in the TETRAD suite of search algorithms (Tetrad IV, 2010). Both procedures produce patterns and neither requires the user to specify a stopping criterion. CPC is provably asymptotically uniformly consistent under i.i.d. sampling when there are no latent confounders of recorded variables. CPC requires the user to specify an alpha value for tests of conditional independence, set at .01 in our simulations. GES is asymptotically consistent.[3] Either procedure tends to give better results than MH with i.i.d. samples on reasonable finite samples, and the searches are

[3] GES is consistent in the sense that its output converges probability 1 to the true Markov equivalence class from i.i.d. samples from a distribution faithful to a DAG. CPC is uniformly consistent in the sense that given i.i.d. samples from a distribution faithful to a DAG, for all positive ε, there exists an N such that the probability of the algorithm producing the true Markov equivalence class on samples of size > N is greater than 1 − ε. On small i.i.d. samples from such a distribution, CPC produces graphs that represent disjunctions of Markov equivalence classes, by noting explicitly where identifications of collider paths (of the form X → Y ← Z, for variables X, Y, and Z) are ambiguous because of conditional independence information that conflicts with Markov properties. The convergence results do not depend on whether the graphs are understood causally or merely as a summary of distribution facts. For any DAG with binary variables and a multinomial distribution, a result of Meek's (Spirtes, et al., 2000) shows the set of probability distributions faithful to the DAG has probability 1 for any smooth measure on the possible distributions.

Table 2
$N = 1000$, uncombined. Averages of true positive errors, false positive errors, false negative errors, and false negative rate over simulation runs, for the case where datasets are not combined pairwise, as described in the text. 'TP' records the average number of edges in the output graph also contained in the graph generating the data; 'FP' records the average number of edges in the output not contained in the input graph, 'FN' records the average number of edges in the input graph that are not contained in the output graph, and 'FPR' records the average of the number of false positives to number of edges returned. For edge counts, an edge X ↔ Y is counted as a prediction of both X → Y and Y → X.

	TP	FP	FN	FPR
CPC, directed edge errors	4.10	1.30	3.90	0.24
GES, directed edge errors	3.50	1.40	4.50	0.28
CPC, adjacency errors	5.00	0.05	3.00	0.01
GES, adjacency errors	5.00	0.15	3.00	0.03

Table 3
$N = 500 + 500$, combined: Averages of true positive errors, false positive errors, false negative errors, and false negative rate over simulation runs, for the case where datasets are combined in pairs, as described in the text. 'TP' records the average number of edges in the output graph also contained in the graph generating the data; 'FP' records the average number of edges in the output not contained in the input graph, 'FN' records the average number of edges in the input graph that are not contained in the output graph, and 'FPR' records the average of the number of false positives to number of edges returned. For edge counts, an edge X↔Y is counted as a prediction of both X → Y and Y → X.

	TP	FP	FN	FPR
CPC, directed edge errors	2.46	2.85	12.86	0.52
GES, directed edge errors	2.14	2.52	13.18	0.52
CPC, adjacency errors	4.34	1.01	10.98	0.16
GES, adjacency errors	3.98	0.71	11.34	0.15

roughly 20 times faster, reducing to an hour the 24 h that the entire process would require with MH. The results of the simulation studies are shown in the following tables. Data may be found at http://www.phil.cmu.edu/projects/tetrad/on.metaanalysis/data.zip.

Neumann et al. report that in models with six edges their procedure produces few false positives but fails to capture all edges—they do not detail how many are missing. We find for the simpler problem–identifying variable pairs that are adjacent in one or the other source graphs–with eight edges in each source graph, on average the searches on the uncombined data for each graph are very accurate, producing almost no false positives and omitting about three variable pairs. With the combined data the accuracy decreases, producing more but still few false positives, but a much larger number of false negatives (Table 3). When directions of edges are considered, for the uncombined data the number of false positives remains small, less than two on average, and about half of the true directed edges are not found. When the data are combined, however, the number of false positives more than doubles, and the number of true edges not found more than triples. Note that, a priori, the chance of false positives in combined data is less than in the uncombined data, and the chance of false negatives is greater. Our simulation results are not any estimate of the actual error rates in the Neumann et al., analyses of real data. The various error measures are generally unknown in real cases, and will vary with the unknown true parameterizations, the unknown total number of edges in the various sources, the sample sizes etc.

Discussion

Although Neumann et al., reject a causal interpretation of their search results on the grounds that "causal relationships in general cannot be inferred from observational data alone,"[4] proposals for search procedures for neural processing mechanisms represented by directed graphs are increasingly common. The graphical structure of a DAG and a probability distribution over its variables can be unambiguously associated with a family of distributions that would result from various (usually hypothetical) exogenous interventions on the represented variables (Spirtes et al., 1993; Pearl, 2000), yielding a causal statistical model. Graphical causal models characterize abstract features of almost all forms of statistical models that have been proposed to represent causal cascades among locally aggregated neural activities in psychological tasks, including structural equation models, dynamic causal models, Granger causal models and others. Contrary to Neumann et al., Spirtes et al., (1993, 2000), and a large related literature in machine learning, prove that various search procedures are consistent estimators of various causal relations from observational data subject, as with all statistical estimators, to sampling and distribution assumptions. We nonetheless agree with Neumann et al. that their meta-analysis method does not discover causal relations representable by a DAG or Markov equivalence class, because it does not reliably discover the dependence and conditional independence structures of component systems.

The problems we have advanced with regard to the Neumann et al. study apply only to methods, like theirs, that attempt to infer associations true of individuals by pooling data across individuals, when those data are samples from different probability distributions. Our particular objections to the Neumann et al. method do not apply when, as is more common in imaging studies, the inferences are to means of distributions, or when distribution parameters are estimated separately on individual data and group averages of these estimates are reported. We emphasize that whether a data analysis procedure runs afoul of Yule's problem, or related issues, may require careful thinking about the method and its assumptions. For example, "random effects" models assume the sample is drawn from a probability distribution that includes an independent distribution over parameters; in that case, as shown above, pooling data is unexceptionable if the assumption is correct. Again, independent components methods, such as the LiNGAM family of algorithms (Shimizu et al., 2005) require sample sizes larger than those typically provided in fMRI studies and it is therefore tempting to pool data from separate experimental runs. There seem to be no analyses of how robust such algorithms are when data are pooled. The effects of mixing of records on clustering algorithms is difficult to assess, in part because there is rarely an objective standard, even in simulation, against which to judge the accuracy of a clustering.

When spurious edges or edge reversals will be obtained by mixing data sets each of which is faithful to a graphical causal model—although not necessarily the same one—is a complex, and usually unknown, function of the actual causal structures in the respective data sets, the ranges of the variables, the distributions in the several data sets, the distribution of parameters, the sample sizes and the search methods. If, for example, two binary variables are independent in each of two data sets and the probabilities are not too dissimilar, standard search procedures combining 1000 data cases from each will preserve independence; if, in contrast, the probabilities for a value of 1 for each of the two variables in the two data sets are respectively (.9, .2) and (.2, .9) a spurious association is found from mixing. Different results may obtain with different sample sizes and proportions. For Gaussian systems which consist of dependent variables with distributions faithful to a common DAG, the independence of the distribution of linear coefficients across data sets is sufficient to obtain correct results with centered variables. It is, we

[4] Neumann et al. write: (p. 1382), "Further, it is important to note that directionality in a Bayesian network does not imply causal relationships. In fact, causal relationships in general cannot be inferred from observational data alone. This requires the application of external intervention (Pearl, 2000), a fact that holds true for all directed network models." Pearl's book describes the experimental implications of causally interpreted DAG models, however discovered, but does not make the claim Neumann et al. attribute.

think, sometimes reasonable to assume that coefficients—whether linear coefficients or conditional probabilities in categorical systems—are invariant within subjects on repeated trials and independent between subjects under similar experimental conditions, but selected subgroups may have parameters that are dependent. We do not know conditions for safely mixing samples from different, continuous, non-Gaussian distributions. When combining samples from two distributions faithful to graphs containing edges directly oppositely between the same two variables, the result will often be a spurious edge between one of or both of the variables and other variables. Such associations may result in misdirections of other edges.

It should be noted that the problem addressed in this paper—finding causal relations or independence and dependence relations using multiple samples from different distributions with different graphical structures—is different from the problem of finding the possible causal or independence structures when the data are measures of different (but intersecting) sets of variables sampled from the same distribution. The latter problem is solved in Tillman et al. 2009. A related problem arises when the samples are from different distributions but share a common directed graphical structure. Ramsey et al. (2010) offers a method for that problem.

A natural idea for searching for graphical models of mixed data is to use a Bayesian hypermodel with a scoring search like that of GES. Suppose there are N data sets. Let R be the union of the variables in the data sets, and let $\Theta = <\Theta_1, \ldots \Theta_N>$ be a configuration, for each of the N data sets, of all possible parameters relating the variables. For example, in linear models, Θ will contain for each data set, k, and for each member of R in that data set, a parameter ranging over variance values, and for each ordered pair, $<X, Y>$ of variables in k, a parameter for the linear dependence of Y on X. Put a suitable prior distribution over Θ, compute posterior distributions conditional on Θ, and search iteratively for the best graphical model in each data set. Theissen et al. (1997) have implemented such a procedure for Gaussian models. They show that on simulated data from mixtures of three graphical models (two of which are identical, and without any reversed edges between the models) with varying sample sizes, from 93 to 3000, the procedure tends to misidentify the number of components, but is quite accurate in the total set of directed edges it finds. Unfortunately, their simulated example is too specialized: there are only three data generating models, two of the data generating graphs are identical, there are no differences in edge directions between data generating graphs, and all linear coefficients in the data generating models were fixed at 1. Their procedure deserves further research for the linear case; for models with binary or categorical valued variables, however, the procedure seems infeasible for fMRI meta-analysis of the kind Neumann et al. attempt: the parameter space Θ would be too large.

Another strategy is to unmix the sample distribution—decompose it, for example, into a collection of covariance matrices—and then apply a search procedure separately to each component (Figuieredo and Jain, 2002; Verbeek et al., 2003). These procedures start either with a one component Gaussian model and add components as necessary, or with a large multi-component model and eliminate components as needed, in both cases estimating parameters by variants of expectation maximization. Simulation results are given for small (3) component cases. It is not known how the procedures would scale up to much larger numbers of data sets, or how well they would perform with non-Gaussian distributions. These procedures deserve further investigation.

Meta-analysis might also be approached by searching for graphical causal structure on various individual time series and clustering the results. For sparse graphs, with variable sets larger than the usual sets of regions of interest identified in fMRI studies, search with CPC, GES or related algorithms requires only a fraction of a second, and analysis of thousands of series is quite feasible. Clustering can be done directly on numerical representation of graphical structure using, for example, hamming distances.

Conclusion

We believe, contrary to Neumann et al., that the chief point of network representations obtained by search methods applied to imaging data ought to be to guide inference to processing mechanisms, and that producing networks merely as a summary of joint distributions is of little scientific value. To that end, the mathematical facts of mixtures, together with individual variation and simulation results of the kind we have illustrated, provide reasons to be skeptical of meta-analyses that use conventional machine learning techniques on combined samples from multiple fMRI studies with different experimental paradigms, or with combinations of neurotypical and neuroatypical groups. In such cases, we believe there is often insufficient reason to believe that the causal pathways and distribution properties and sufficiently similar, or vary with appropriate randomness, so as to preserve Markov properties when data are combined. With multiple neurotypical subjects in a common experimental paradigm it may be reasonable, although not necessarily correct, to suppose that there is a common causal structure or superstructure, and that parameters are independent between subjects. It is more risky, we believe, to combine data from neuroatypicals sharing a common clinical diagnostic syndrome–autistics or schizophrenics for example–which may be the result of (or result in) varying neurological processing anomalies. Neural processing may also differ by age, sex and other biological conditions. In any case, the application of search methods to multiple data sets requires careful attention to the sometimes subtle conditions under which a procedure is reliable.

Acknowledgment

We thank Jane Neumann for very helpful comments on a draft of this note.

References

Chen, R., Herskovitz, E.H., 2010. Machine-learning techniques for building a diagnostic model for very mild dementia. NeuroImage 52, 234–244.

Chen, H., Yang, Q., Liao, W., Gong, Q., Shan, S., 2009. Evaluation of the effective connectivity of supplementary motor areas during motor imagery using Granger causality mapping. NeuroImage 47, 1844–1853.

Figuieredo, M., Jain, A., 2002. Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24, 381–396.

Friston, K., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.

Gates, K., Molenaar, P.C.M., Hillary, F.G., Ram, N., Rovine, M., 2010. Automatic search for fMRI connectivity mapping: an alternative to Granger causality testing using formal equivalences among SEM path modeling, VAR, and unified SEM. NeuroImage 50, 1118–1125.

Laird, A., Robbins, J.M., Li, K., Price, L.R., Cykowski, M.D., Narayana, S., Laird, R., Franklin, C., Fox, P., 2008. Modeling motor connectivity using TMS/PET and structural equation modeling. NeuroImage 41, 424–436.

Marreiros, A., Kiebel, S.J., Friston, K.J., 2009. A dynamic causal model study of neural population dynamics. NeuroImage 51, 91–101.

Meek, C., 1997. Graphical Models: Selecting Causal and Statistical Models, Ph.D Thesis, Department of Philosophy, Carnegie Mellon University.

Neumann, J., Fox, P., Turner, R., Lohmann, G., 2010. Learning partially directed functional networks from meta-analysis imaging data. NeuroImage 49 (2), 1372–1384.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann.

Pearl, J., 2000. Causality. The Cambridge University Press, Cambridge.

Rajapakse, J.C., Zhou, Juan, 2007. Learning effective brain connectivityity with dynamic Bayesian networks. NeuroImage 37, 749–760.

Ramsey, J., Spirtes, P., Zhang, J., 2006. Adjacency-Faithfulness and Conservative Causal Inference. Proceedings of the 22nd Convergence on Uncertainty in Artificial Intelligence. AUAI Press, Oregon, pp. 401–408.

Ramsey, J., Hanson, C., Hanson, S., Halchenko, Y., Poldrack, R., Glymour, C., 2010. Six problems for causal inference from fMRI. NeuroImage 49 (2), 1545–1558.

Roebroeck, A., Formisano, E., Gabriel, R., 2005. Mapping directed influence over the brain using Granger causality and fMRI. NeuroImage 25, 230–242.

Shimizu, S., Hyvärinen, A., Kano, Y., Hoyer, P.O., 2005. Discovery of non-gaussian linear causal models using ICA. Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005), pp. 526–533.

Simpson, E.H., 1951. The interpretation of interaction in contingency tables. J. R. Stat. Soc. B 13, 238–241.

Spirtes, P., Glymour, C., Scheines, R., 1993, 2000. Causation, Prediction and Search. Springer.2nd ed. MIT Press.

TETRAD IV, 2010. http://www.phil.cmu.edu/tetrad/.

Theissen, B., Meek, Heckerman, D., Chickering, D.M., 1997. Learning mixtures of DAG models. Proc. Of the 13th Conference on Uncertainty in AI, pp. 504–513.

Tillman, R.E., Danks, D., Glymour, C., 2009. Integrating locally learned causal structures with overlapping variables. Advances in Neural Information Processing Systems, 21, pp. 1665–1672.

Verbeek, J., Vlassis, N., Krobe, B., 2003. Efficient greedy learning of Gaussian mixture models. Neural Comput. 15, 469–485.

Yule, G., 1903. Notes on the theory of the distribution of attributes in statistics. Biometrika 2, 121–134.