A proposal for how to "prove" that a brain property may cause a mental property

Joshua T. Vogelstein, R. Jacob Vogelstein, Carey Priebe
Johns Hopkins University

January 9, 2010

Abstract

1 Introduction

Suppose that you wanted to make the following argument:

Hypothesis 1. *A particular property of a brain causes a particular mental property.*

The "brain property" could be nearly anything: the genetic expression profile of a subset of neurons in a particular developmental stage, or microtubules becoming entangled, or columnar organization, or the bumps in our skulls, etc. Similarly, the "mental property" could be almost anything: intelligence, capacity for love, knowing the square-root of pi, etc. How might you go about *proving* Hypothesis 1 (H1)?

Here's how we would do it. First, we would change H1 to be *testable*, and change our desiderata from *proving* its accuracy to *collecting evidence* to support its claim. More specifically, we would make the following argument:

Hypothesis 2. *A change in particular property of a brain can predict a corresponding change in a particular mental property*[1].

Now, to collect evidence in support of H2, we need the following:

1. A way to describe any brain, b. Our description should live in the space of all possible descriptions of brains (or, at least those under consideration), $\mathcal{B}$.
2. A way to describe a mental property, m, which lives in the space of mental properties, $\mathcal{M}$. For instance, if the mental property under investigation is IQ, then $\mathcal{M}$ may be all possible scores on a particular IQ test.
3. A mapping from brain space to mental space, which enables us to relate m and b. Call this mapping, g, so $g : \mathcal{B} \mapsto \mathcal{M}$.

If we choose to take a statistical perspective, then both brains and minds are random variables. More specifically, let B be a random variable corresponding to a brain, so that any particular brain, b, is a sample from the space of brains, $\mathcal{B}$. The probability of B taking any value $b \in \mathcal{B}$ is given by the *marginal* distribution $F_B[b = B]$. Similarly, let M be a random variable corresponding to the mental property under investigation, so m is a sample from the space of mental properties, $\mathcal{M}$. The probability of M taking any value $m \in \mathcal{M}$ is given by the *prior* distribution $F_M[m = M]$. The mapping, g, tells us the relationship between any m and any b. Or, more formally, the mapping tells us about the *posterior* distribution of minds given brains, $F_{M|B}[m = M|b = B]$.

If we knew the *joint* distribution of minds and brains, $F_{BM}[b = B, m = M]$, and the marginal distribution of brains, F_B, then finding the mapping from brains to minds would be trivial: we would simply use Bayes rule to obtain the posterior, $F_{M|B} = F_{BM}/F_B$. However, in practice these distributions are typically unknown. Therefore, we must *estimate* g from a corpus *data*. Assume we have collected n brain/mental pairs. Then, define the corpus of data as the

[1] Note the relationship between this and *statistical-supervenience*[**?**]

collection of all such pairs: $\mathcal{D}_n = \{(b_1, m_1), \ldots, (b_n, m_n)\}$. The estimated mapping, g_n, then takes a new brain and the old *training data*, and makes predictions about the mental property m. Formally, $g_n : \mathcal{B} \times (\mathcal{B}, \mathcal{M})^n \mapsto \mathcal{M}$, so $g_n(b; \mathcal{D}_n) = \widehat{m}$.

This work contributes a perspective on a possible space of brains, $\mathcal{B}$, and several principled ways of choosing g_n. In so doing, we also describe a set of desiderata for both $\mathcal{B}$ and g_n to satisfy. We make no claim that our definitions or desiderata are novel or unique. Rather, the novel contribution (if any), is the overarching statistical perspective that unifies previously (potentially) disparate ideas into a single coherent framework.

2 A possible description of brains with certain desirable properties

Many possible spaces for the descriptions of brains exist. For instance, phrenologists thought that the "sulci and gyri" of the *skull* were sufficient to explain various mental properties, including intelligence. The space of brains they considered then, were all possible skull shapes. Clearly, such a space is not sufficient to compare the evidence support phrenological theory with an alternate theory, such as it is the sulci and gyri of the *brain* that determine those mental properties. So, can we enumerate a set of desiderata which we would use to select our brain-space? Here is a possible set:

1. $\mathcal{B}$ should be sufficiently large to be able account for whatever properties of the brain are casually related to the properties of cognition under investigation.
2. $\mathcal{B}$ should be sufficiently large to be able span multiple *levels of explanation*. That is, we might want to be able to compare a hypothesis about the role of genetic expression profiles with another hypothesis about default mode networks.
3. $\mathcal{B}$ should be only as large as necessary, and no larger. In particular, if we do not believe that total brain volume can *cause* a particular mental property, that it need not be included.
4. The properties of any particular brain, $b \in \mathcal{B}$, should be either measurable or estimatable, such that experimental observation may be used to obtain them.
5. $\mathcal{B}$ should admit algorithms that (are guaranteed to be able to) capture the relationship of interest.
6. $\mathcal{B}$ should also admit *causal* studies, which entail modifying a particular $b \in \mathcal{B}$, to modify the corresponding mental property, $m \in \mathcal{M}$.

Note that the above desiderata are neither complete nor unique; rather, they provide (hopefully) a reasonable set of criteria for evaluating any proposed $\mathcal{B}$. With this in mind, we propose the notion of a *brain-graph*. Specifically, we say that the brain may be well characterized as a labeled, attributed multigraph (which is a generalized notion of a or network). Formally, we define a brain-graph, $b \in \mathcal{B}$ as a 4-tuple, $\mathcal{B} = (\mathcal{V}, \mathcal{E}, \mathcal{X}_V, \mathcal{X}_E)$, defined by the following:

- The set of vertices (nodes), $V = \{V_i\}_{i \in [n_v]}$. If we assume, for instance, that neurons are the fundamental (atomic) unit of computation, then each vertex could correspond to a neuron. Regardless of what vertices represent, formally, we let $V_i \in \mathcal{V}_i = \{0, 1\}$ for $i \in [n_v] = \{1, 2, \ldots, n_v\}$, $n_v \leq \infty$, and $\mathcal{V} = \mathcal{V}_i^{n_v}$.
- The set of edges, $E = \{E_{ij}\}_{i,j \in [n_v]}$. Again, if we assume that neurons are the fundamental unit of computation, then each edge could correspond to a synapse. To simplify matters we may only consider the presence or absence of a synapse, in which case $E_{ij} \in \{0, 1\}$. Or, we may consider the effective strength of a synapse, in which case $E_{ij} \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers, $\{0, 1, 2, \ldots\}$. Further, we could allow for the possibility of multiple "kinds" of synapses between any pair of neurons, such as chemical and electrical. In such a case, we have $E_{ijk} \in \mathcal{E}_{ijk} \subseteq [n_v]^2 \times [n_k]$, where $n_k \leq \infty$ is the maximum number of categorically different edges. In any case, we can define the space of edges as $\mathcal{E} = \mathcal{E}_{ijk}^{n_v^2 \times n_k}$.
- If vertices have features (or labels/attributes), then let X_i correspond to the feature vector for vertex i (features may correspond to *anything* about the vertex). Again, assuming vertices represent neurons, features may indicate neurotransmitter released, morphological properties, receptive fields, etc. Formally, $X_i \in \mathcal{X}_1 \subseteq \mathbb{R}^{d_V}$ for $i \in [n_v]$, and d_V is the dimensionality of the feature vectors, so $\mathcal{X}_V = \mathcal{X}_1^{n_v}$. It may be the case that certain features are measurable/observable, and others are hidden. If so, let $X_i = X_i^o \cup X_o^h$, and $X_i^o \cap X_i^h = \emptyset$.

- If edges also have features, then let X_{ij} correspond to the feature vector for edge (i, j). If edges are synapses, then edge features might include things like probability of release, post-synaptic potential shape, etc. Let $X_{ij} \in \mathcal{X}_2 \subseteq \mathbb{R}^{d_E}$ for $(i, j) \in [n_v] \times [n_v]$, and d_E is the dimensionality of the edge features, so $\mathcal{X}_E = \mathcal{X}_2^{n_v^2}$. In the scenario where categorically different edges exist, let an additional index k indicates edge features for each category k, so $\mathcal{X}_E = \mathcal{X}_2^{n_v^2 \times n_k}$.

Given such a brain-space $\mathcal{B}$, a natural question is: does this notion of brain-graphs satisfy the above desiderata? Let's see:

1. The space of all possible brain-graphs does appear to be quite large, potentially incorporating many small and large details of brains.
2. Because each vertex V_i could correspond to a neuron, a column, a neuroanatomical region, etc., indeed, brain-graphs can span multiple levels of explanation. Even for a given brain, it is possible to let some V_i's represent neurons, and other V_i's represent neuroanatomical regions, even if this is a redundant characterization of the brain, in that the neurons are *within* the neuroanatomical region.
3. While the brain-graph space is large, it is not "all-encompassing." For instance, the space of all possible functions is not within brain-graph space. Thus, the brain-graph space seems to exclude at least some possibilities.
4. For a space to admit algorithms guaranteed to capture the relationship of interest, one must prove limiting results. For instance, Stone proved in 1977 [**?**] that the k_n nearest neighbor algorithm is guaranteed to converge to the Bayes optimal solution. In 2010, Vogelstein et al. [**?**] proved a similar result holds for brain-graphs.
5. Brain-graphs clearly admit causal studies, as one could modify the number of nodes, or the value of edges or features, and (potentially) observe a corresponding change in a mental property.

Thus, it seems that brain-graphs indeed satisfy the above criteria. This is not to say that brain-graphs are the only space one could define to satisfy these criteria, merely that brain-graphs are sufficient. So, given such a space of brains, the next question is: "how can one choose a mapping between brains and minds?"

3 Possible approaches to choosing mappings with desirable properties

Given $\mathcal{B}$, what can we do with it? As stated above, our goal is to relate these models to properties of cognition. More specifically, let $\mathcal{M}$ characterize the space of the mental (cognitive) property, and $g \in \mathcal{G}$ be some mapping to learn. Then:

- If $\mathcal{M} = \{0, 1\}$, then g may be a two-way classifier: $g : \mathcal{B} \to \{0, 1\}$.
- If $\mathcal{M} = \{0, 1, \ldots, n_m\}$, then g may be an n_m-way classifier: $g : \mathcal{B} \to \{0, 1, \ldots, n_m\}$.
- If $\mathcal{M} = \mathbb{R}^a$, then g is a (multivariate-) regressor: $g : \mathcal{B} \to \mathbb{R}^a$.

In general, solving the above problems—which means finding g—will depend on the joint distribution of brains and minds, $F = F_{BM}$. In practice, however, F is typically unknown, and therefore g must be estimated from the data. Assume we have a corpus of training data, $\mathcal{D}_n = \{(B_1, M_1), \ldots, (B_n, M_n)\}$, where n is the number of training samples. Our goal then is to compute $g_n : \mathcal{B} \times (\mathcal{B}, \mathcal{M})^n \to \mathcal{M}$, which takes as input an observed brain-graph b and n training paris $\{(b_1, m_1), \ldots, (b_n, m_n)\}$ and produces a prediction $\widehat{m} = g_n(b; \mathcal{D}_n)$. The particular g_n should be the one that minimizes some loss function, $L_F(g_n)$, over the space of all possible g_n's, $\mathcal{G}$. For instance, when $|\mathcal{M}| = 2$, $\mathcal{G}$ is all possible two-way classifiers, and a potentially reasonable loss function is $L_F(g_n) = \mathbb{E}[P_F[g_n(B; \mathcal{D}_n) \neq M | \mathcal{D}_n]]$.

Importantly, in addition to finding a g_n that minimizes some loss-function, g_n should admit a way to *morph* any brain's mental property m, by modifying b. This will be discussed at greater length in the sequel.

Thus, given a mental property, a decision about how to represent it, $\mathcal{M}$, and a loss function, L, our task is to find a good algorithm, g_n. Two complementary strategies are possible: model-free and model-based. Model-free algorithms have the advantage that no model need be specified. Thus, in theory, model-free algorithms have the advantage of having little or no bias. Unfortunately, this freedom comes with the cost of relatively high variance. On the other hand, model-based algorithms can significantly reduce variance, but (almost) necessarily increase bias. Importantly, many standard algorithms, including linear, quadratic, and support vector based classifiers/regressors *implicitly* define a model, and are therefore not strictly "model-free".

3.1 Model-free algorithms

Model-free algorithms often operate on interpoint distance space, as opposed to the explicit data space [**?**]. More formally, given any two brain-graphs, b_1 and b_2, first define an interpoint (pseudo-) distance metric: $\rho : \mathcal{B} \times \mathcal{B} \mapsto [0, \infty)$. This reduces the problem from operating in $\mathcal{B}$ to operating in $\mathbb{R}$. Because the data collected is often corrupted by noise, it is typical to also introduce a smoothing function: $s : \mathcal{B} \mapsto \mathcal{B}$. For the brain-graph scenario, this may correspond to inferring unobserved edges. Thus, the smoothing-derived (pseudo-) distance metric, ρ' is defined as: $\rho' = \rho(s(b_1), s(b_2))$.

Perhaps the prototypical model-free algorithm is the k_n nearest neighbor (kNN) algorithm. Vogelstein et al. (2010) showed that a kNN classifier is a universally consistent classifier (meaning, achieves the Bayes optimal performance), for any F_{BM}, under a Frobenius norm distance metric. In other words, they let defined $\rho(\cdot) = \|\cdot\|_F$, and s was simply the identity (that is, no smoothing). Simulations showed that for only a few hundred simulated sample data points, this kNN achieved misclassification error rate below 10%. In practice, it is often the case that other model-free algorithms outperform (in accuracy) any particular kNN, including class cover catch digraphs, decision trees, and various ensemble approaches such as random forests []. Unfortunately, scant theoretical works is available to provide proofs of universal consistency for these alternate model-free algorithms.

3.2 Model-based algorithm

The other possible strategy is to propose a class of models, $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d\}$, that describe the data (note that the choice of how to determine the dimensionality d of the model specifies whether the model is parametric, semiparametric, or nonparametric). Ideally, the class of models is sufficiently large to include models very close to the "truth", and be able to find a Minimally Sufficient Model (MSM; by analogy with minimally sufficient statistics) within the class, that sufficiently explains the data, with the "smallest" $\boldsymbol{\Theta}$ (a potentially useful measure of size is the set cardinality). Classification/regression with g_n is then performed on the estimate of $\boldsymbol{\theta}$, which lives in some $\boldsymbol{\Theta}$ space smaller than $\mathcal{B}$, thereby reducing the variance, without increasing bias too much (hopefully). The model-based approach also (potentially) offers the advantage of *interpretability*, if the parameters, $\boldsymbol{\theta}$, correspond to interpretable features of the brain.

Let b_i correspond to brain-graph i. The model-based approach then typically assumes that each b_i is sampled identically and independently from some parametric distribution, $b_i \overset{iid}{\sim} P(b_i|\boldsymbol{\theta}) \, \forall i \in [n]$. Because $\boldsymbol{\theta}$ is typically unknown, an estimate, $\widehat{\boldsymbol{\theta}}$, must be found. Given such an estimate, g_n operates directly on the estimated parameters, as opposed to the brain-graphs [2]

The maximum likelihood approach then specifies to find The task is then to find an estimate $\widehat{\boldsymbol{\theta}}$, such that

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i \in [n]} P(b_i|\boldsymbol{\theta}) \tag{1}$$

If a prior over the parameters is specified

Generative Model Consider the following generative model for attributed multigraphs:

- Let c be a *class identity*, where $c \in \mathcal{C} = \{0, 1, \ldots, C\}$. The distribution of class identity, which is a random variable, is given by $f_c = MN(\boldsymbol{\pi})$, where MN indicates a multinomial, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_)$
- Let $\boldsymbol{\theta}_c$ be the *parameters* for class c, where $\boldsymbol{\theta}_c \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$, for some $d = d_V + d_E \in \mathbb{Z}$, where the dimensionality of the parameters is implicitly a function of the data, $\mathcal{D}_n$.
- Let b be a *brain-graph*, where $b(\boldsymbol{\theta}_c) \in \mathcal{B} = (\mathcal{V}, \mathcal{E}, \mathcal{X}_V, \mathcal{X}_E)$, where for clarity, we restrict edges to be integer weights (i.e., only include a single category of edge attributes, but this may straightforwardly generalized)

To sample graphs from this generative model, assuming that C and V are given (the number of classes and vertices per graph, respectively), one can use the following procedure (generalizing to the unknown V case is straightforward and therefore omitted):

- sample $c \sim f_c$

[2] Is it interesting to note that neither paradigm actually operates directly on the data? The model-free approach operates

- sample $\boldsymbol{\theta}_c \sim f_{\boldsymbol{\theta}}(\cdot|c)$
- for $i \in [n_v]$, sample $X_i \sim f_{X_V}(\cdot|\boldsymbol{\theta}_c^V)$
- for $(i,j) \in [n_v] \times [n_v]$
  - sample $X_{ij} \sim f_{X_E}(\cdot|\boldsymbol{\theta}_c^E)$
  - sample $E_{ij} \sim f_E(\cdot|X_i, X_j, X_{ij})$

Note that we have partitioned $\boldsymbol{\theta}_c$ into $\boldsymbol{\theta}_c^V$ and $\boldsymbol{\theta}_c^E$. The probability of obtaining any graph, when using this procedure, is therefore given by:

$$P(G|C,V) = \left(\prod_{i,j\in[n_v]} f_E(E_{ij}|X_i,X_j,X_{ij}) f_{X_E}(X_{ij}|\boldsymbol{\theta}_c^E)\right)\left(\prod_{i\in[n_v]} f_{X_V}(X_i|\boldsymbol{\theta}_c^V)\right) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_c|c) f_c(c) \tag{2}$$

Evaluating Eq. 2 requires defining $f = \{f_c, f_{\boldsymbol{\theta}}, f_{X_v}, f_{X_E}, f_E\}$ (note that $f_{\boldsymbol{\theta}}$ implicitly depends on defining $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$, which, in turn, requires a rule for deciding d based on the data). The most natural choice for f_c is a multinomial, that is, $f_c = MN(\vec{\pi})$, where $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_C\}$, $\pi_c > 0$, and $\sum_c \pi_c = 1$. Our task is then to choose the rest of f that yields: (i) models that display the properties of the data, and (ii) are statistically tractable, meaning that $\boldsymbol{\theta}$ may be estimated consistently from the data [3]. To proceed, we first write the posterior of interest. Importantly, it is often the case that some attributes and edges are hidden. Define $\boldsymbol{E} = \{E_{ij}\}_{i,j\in[n_v]}$, and let $\boldsymbol{E} = \boldsymbol{E}^h \cup \boldsymbol{E}^o$, where $\boldsymbol{E}^h$ corresponds to hidden edges, and $\boldsymbol{E}^o$ corresponds to observed edges. Similarly, let $X = X_V \cup X_E$, and then $X = X^h \cup X^o$. Finally, define $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c\}_{c\in[C]}$. Thus, we are interested in estimating/maximizing:

$$P(\boldsymbol{\theta}, X^h, \boldsymbol{E}^h | C, V, X^o, \boldsymbol{E}^o) \propto ... \tag{3}$$

Several strategies for maximizing Eq. 3 are possible. First, one could use a Gibbs sampling strategy, iteratively sampling $\boldsymbol{\theta}$, X^h, and $\boldsymbol{E}^h$. Second, one could use an expectation maximization algorithm, recursively finding the expected values for X^h and $\boldsymbol{E}^h$, and then use them to estimate $\boldsymbol{\theta}$. The precise definition of f might necessitate approximating both these strategies. Alternately, greedy or variational approaches might be more efficient.

3.3 Classification problem

Imagine we are in the classification setting. What does it mean that $P(B|M=0) \neq P(B|M=1)$? It must mean that, for some subset of edges, the distribution is different for brain-graphs in the two different classes. More formally, let $S = \{(i,j)|(i,j) \in S\}$. If class conditional posteriors are different, then it must be the case that $P(\{E_{ij}\}_{(i,j)\in S}|C=0) \neq P(\{E_{ij}\}_{(i,j)\in S}|C=1)$. Given the above class of models, we may write:

$$\begin{aligned} P(\{E_{ij}\}_{(i,j)\in S}|c) &= f(\cdot|X, \boldsymbol{\theta}_c) \\ &= f_E(\cdot|X) f_X(\cdot|\boldsymbol{\theta}_c) f_{\boldsymbol{\theta}}(\cdot|c) \\ &= f_E(\cdot|X) f_{X^h}(\cdot|X^o, \boldsymbol{\theta}_c) f_{\boldsymbol{\theta}}(\cdot|c) \end{aligned} \tag{4}$$

Thus, given a specification for f, the task is to find S, estimate the parameters $\boldsymbol{\theta}$, compute the likelihood under the two models, and compare. Finding S, however, is, in general, non-trivial.

Limitations/extensions attributed vertices, hyperedges

4 Simulated applications

5 Discussion

[3] The above generative framework generalizes and unifies previously proposed stochastic graph models, including Random Dot Product Graphs, stochastic block models, etc.