Are mental properties supervenient on brain properties?

Joshua T. Vogelstein[1], R. Jacob Vogelstein[2], Carey E. Priebe[1]
[1]Department of Applied Mathematics & Statistics,
Johns Hopkins University, Baltimore, MD, 21218,
[2]National Security Technology Department,
Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723

December 31, 2010

Abstract

The "mind-brain supervenience" conjecture suggests that all mental properties (e.g. consciousness, intelligence, personality, etc.) are derived from the physical properties of the brain. The validity of this conjecture has been argued in philosophical terms for over 2,500 years. Alternative conjectures, including various non-physical causes of mental properties, seem rather implausible to many, but proving or disproving these alternatives has remained elusive.

To enable addressing the question, through empirical means, of whether the mind supervenes on the brain, here we frame a supervenience hypothesis in rigorous mathematical terms and propose a modified version of supervenience (called ε-supervenience) that is amenable to empirical investigations and statistical analysis. To elucidate this approach, we posit a thought experiment that illustrates how the probabilistic theory of pattern recognition can be used to make a one-sided determination of ε-supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the brain-graph or *connectome*), and ε-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given misclassification rate $\varepsilon > 0$, regardless of the relationship between the two.

The potential significance of this work can be divided into distinct disciplines. To the philosopher, this work demonstrates an example of morphing philosophical conjectures into statistical hypotheses that are amenable to experimental investigations, allowing the philosopher to add empirical support to her rational arguments. To the statistician, this work points out the limitations of hypothesis testing in a novel domain, and is suggestive of more general limitations previously (perhaps) under-appreciated. To the neuroscientist, cognitive scientist and psychologist, this work indicates that much of contemporary research can be considered supervenience investigations, further motivating brain-graph related research paradigms.

1 Introduction

Questioning the relationship between the mind (our thoughts, beliefs, preferences, emotions, intelligences, etc.) and the brain (the physical structure inside our skulls) dates back at least as far as 400 BCE, when Plato wrote the dialogues, in which he posited immateriality of the soul [1]. Approximately two millennia passed before these ideas reached their canonical form through Descartes's discussion of mind-body dualism [2]. Then, in the 20th century, Donald Davidson stated and popularized the mind-brain supervenience conjecture, which claims that an agent cannot alter in some mental property without altering in some physical property [3]. Contemporary fields of neural network theory and neuroscientific inquiry often assume mind-brain supervenience, or an even stronger assumption about mind-brain causality, but no previously proposed notion of supervenience seems amenable to empirical investigation. This work is an attempt to bridge the gap between these philosophical conjectures and experimentally testable hypotheses.

The primary contributions of this work are as follows. A notion of supervenience amenable to empirical investigation is formally introduced. This renders the mind-brain dualism debate a hypothesis, rather than an assumption, both expanding the space of questions amenable to hypothesis testing, and placing limits on this space. Because hypothesis tests (implicitly sometimes) depend on a model, a very general model of brains and their associated mental properties is proposed. Fortunately, this formulation admits universally consistent classifiers, that is, classifiers guaranteed to find the relationship between minds and brains, if one exists, given sufficient data. Many previous investigations relating brains and mental properties can therefore be considered ε-supervenience hypothesis tests. This paradigm, therefore,

generalizes previous approaches, embedding them in a rigorous statistical framework, suggesting avenues for future research.

2   Statistical supervenience

Donald Davidson canonized the mind-brain supervenience relation in 1970 with the following quote: [3]

> supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect.

This conjecture may be concisely and formally stated. Let b correspond to an agent's brain, which is a particular element from the set of all possible brains, $\mathcal{B}$. Similarly, let m correspond to an agent's mind, which is a particular element from the set of all possible minds, $\mathcal{M}$. Thus, the above supervenience conjecture implies that: $m \neq m' \implies b \neq b'$, where $(m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$ are mind-brain pairs. This mind-brain supervenience relation does not imply an injective relation, a causal relation, or an identity relation (see Appendix A for more details).

To facilitate both statistical analysis and empirical investigation, we convert this supervenience relation from a logical to a probabilistic relation. Let $\mathbb{P}[M, B]$ indicate a joint distribution of minds and brains.

Definition 1. $\mathcal{M}$ *is said to statistically supervene on* $\mathcal{B}$ *for distribution* $\mathbb{P} = \mathbb{P}[M, B]$, *denoted* $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B}$, *if and only if* $\mathbb{P}[m \neq m' | b = b'] = 0$, *or equivalently* $\mathbb{P}[m = m' | b = b'] = 1$.

Statistical supervenience is therefore a probabilistic relation on sets (related to, but distinct from correlation; see Appendix A for details).

3   Statistical supervenience is equivalent to perfect classification accuracy

If minds statistically supervene on brains, $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B}$, then two different minds must supervene on two different brains. This means that there exists a unique mapping from each brain to a single mind. In other words, one can in principle construct a function $g(b) : \mathcal{B} \mapsto \mathcal{M}$, that is a deterministic mapping from brains to minds. It may be the case that subsets of brains form equivalence classes, such that any brain in that subset is mapped to the same mind (see Appendix A for examples). Assuming for the moment that the space of all possible minds is finite, that is $|\mathcal{M}| < \infty$, then we call any such function a *classifier* (this assumption will later be relaxed). Let $\widehat{m}$ denote the output of a classifier, $g(b) = \widehat{m}$. Define misclassification rate as $L_{\mathbb{P}}(g) = \mathbb{P}[g(B) \neq M]$, which denotes the probability that g misclassifies b. The Bayes optimal classifier g^* minimizes $L_{\mathbb{P}}(g)$ over all classifiers, that is: $g^* = \operatorname{argmin}_g L_{\mathbb{P}}(g)$. Thus, the *Bayes error*, or Bayes risk, $L_{\mathbb{P}}(g^*)$ is the minimum possible misclassification rate. The primary result of casting supervenience as a statistical framework is the following theorem:

Theorem 1. $\mathcal{M}$ *is said to statistically supervene on* $\mathcal{B}$ *for distribution* $\mathbb{P} = \mathbb{P}[M, B]$, *denoted* $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B}$, *if and only if* $L_{\mathbb{P}}(g^*) = 0$. *Formally,* $\mathcal{M} \overset{S}{\sim}_{\mathbb{P}} \mathcal{B} \Leftrightarrow L_{\mathbb{P}}(g^*) = 0$.

If minds supervene on brains, then, by the definition of supervenience, there exists a function that maps each brain deterministically to a particular mind. This means that one could draw a decision boundary between all equivalence classes of brains, each class corresponding to a different mind, and no mind will reside within two different equivalence classes. Thus, the optimal classifier would correctly find these decision boundaries, and therefore have no opportunity to err. $\square$

Thus, the above argument shows (for the first time to our knowledge) that statistical supervenience and Bayes error are equivalent. Further, statistical supervenience can be thought of as a constraint on the possible models. Specifically, let $\mathcal{P}$ indicate the set of all possible joint distributions on minds and brains, and let $\mathcal{P}_s$ be subset of models for which supervenience holds. Then, $\mathcal{P}_s = \{\mathbb{P}[M, B] : L_{\mathbb{P}}(g^*) = 0\} \subseteq \mathcal{P}$.

4   A hypothesis test for supervenience

While the above theorem is of potential theoretical interest, because the arguments rest on knowing $\mathbb{P}[M, B]$ and g^*, which are typically unknown, they are pragmatically useless. However, both $\mathbb{P}[M, B]$ and g^* could be estimated from data. Let $\mathcal{T}_n = \{(m_1, b_1), (m_2, b_2), \ldots, (m_n, b_n)\}$ be a set of random samples taking their values in $\mathcal{M} \times \mathcal{B}$, each independently and identically distributed according to model $\mathbb{P}[M, B]$. Generalizing the concept of a classifier g to allow incorporation of training data, consider $g_n : \mathcal{B} \times (\mathcal{M} \times \mathcal{B})^n \mapsto \mathcal{M}$ which takes as input an observed brain b and training data $\mathcal{T}_n$, and produces a classification $g_n(b; \mathcal{T}_n) = \widehat{m}$. Misclassification rate for this classifier will be a random variable, because the training data $\mathcal{T}_n$ are random samples. The expected misclassification rate for this classifier is therefore approximated by "hold-out" error: $\widehat{L}_{\mathbb{P}}^{n'}(g_{\widetilde{n}}) = \mathbb{P}[g_{\widetilde{n}}(B) = M | \mathcal{T}_{\widetilde{n}}]$, where $\widetilde{n} = n - n'$, and $n' < n$ is the number of held-out training samples (samples not used to obtain $g_{\widetilde{n}}(\cdot)$). The approximate number of misclassified minds therefore has a binomial distribution: $n' \widehat{L}_{\mathbb{P}}^{n'}(g_{\widetilde{n}}) \sim \text{Binomial}(n', L_{\mathbb{P}}(g_{\widetilde{n}}))$.

Before explicitly considering the problem of testing for statistical supervenience, we define a relaxed notion of supervenience:

Definition 2. *Given $\varepsilon > 0$, $\mathcal{M}$ is said to ε-supervene on $\mathcal{B}$ for distribution $\mathbb{P} = \mathbb{P}[M, B]$, denoted $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$, if and only if $L_{\mathbb{P}}(g^*) < \varepsilon$.*

Given this relaxation, consider the problem of testing for $\varepsilon-$supervenience. Let the null hypothesis be H_0: $L_{\mathbb{P}}(g_n) \geq \varepsilon$, and the alternative hypothesis be H_A: $L_{\mathbb{P}}(g_n) < \varepsilon$. We reject for values of the test-statistic lower than the critical value, that is, we reject if and only if $n' \widehat{L}_{\mathbb{P}}^{n'}(g_{\widetilde{n}}) < c_\alpha(n', \varepsilon)$. The critical value is available under the least favorable distribution $\text{Binomial}(n', \varepsilon)$. Thus, rejection implies that we are $100(1 - \alpha)\%$ confident that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$. The definition of ε-supervenience therefore admits, for the first time to our knowledge, a statistical test of supervenience, given a specified ε and α.

Importantly, the utility of any statistical test depends both on the p-value, the probability of obtaining a test statistic at least as extreme as the observed value (under the assumed model), and its power, the probability that the test will reject a false null hypothesis. Ideally, the power of this test would go to unity, as $n, n' \to \infty$. A sufficient condition for power to approach unity is that g_n is a *consistent* classifier. A classifier is consistent if and only if its expected misclassification rate converges to the Bayes optimal limit with sufficient data, that is $\mathbb{E}[L_{\mathbb{P}}(g_n)] \to L_{\mathbb{P}}(g^*)$ as $n \to \infty$.

Unfortunately, the rate of convergence of $L_{\mathbb{P}}(g_n)$ to $L_{\mathbb{P}}(g^*)$ depends on the (unknown) distribution $\mathbb{P} = \mathbb{P}[M, B]$ [4]. Furthermore, arbitrarily slow convergence theorems regarding the rate of convergence of $L_{\mathbb{P}}(g_n)$ to $L_{\mathbb{P}}(g^*)$ demonstrate that there is no universal n, n' which will guarantee that the test has power greater than any specified target $\beta > \alpha$ [5]. For this reason, the test outlined above can provide only a one-sided conclusion: if we reject we can be $100(1 - \alpha)\%$ confident that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ holds, but we can never be confident in its negation; rather, it may be the case that the evidence in favor of $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ is insufficient for any number of reasons, including that we simply have not yet collected enough data. Thus, without restrictions on $\mathbb{P}[M, B]$, arbitrarily slow convergence theorems imply that our theorem of ε-supervenience does not strictly satisfy Popper's *falsifiability* requirement [6].

5   A thought experiment demonstrating unity power

To ensure consistency and therefore unity power, the classifier $g_n(\cdot)$ must be able to converge to the truth, regardless of the true model, $\mathbb{P}$. We therefore make explicit a model for brain, and show that under this very general model, universally consistent classifiers are available.

Thought experiment 1. *Let the physical property under consideration be brain connectivity structure ("connectome"), so b is a brain-graph (or, network) with vertices representing neurons (or neuroanatomical regions) and edges representing synapses (or white matter tracts). Further let $\mathcal{B}$, the observation space, be the collection of all graphs on a finite number of vertices, and let $|\mathcal{B}|$ be countable. Now, imagine collecting very large amounts of very accurate independent and identically distributed brain-graph data and the associated mental property indicators. A k_n-nearest neighbor classifier using an isomorphism-matching Frobenius norm is universally consistent (see Appendix B for proof). Therefore, the existence of a universally consistent classifier guarantees that eventually (in n, n') we will be able to conclude $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ for this mind/brain property pair, if indeed ε-supervenience holds. This logic holds for directed graphs or multigraphs or hypergraphs with discrete edge weights and vertex attributes. Furthermore, Appendix B also extends the proof to deal with other matrix norms (which might speed up convergence), and the regression scenario, where $|\mathcal{M}|$ is infinite.*

6   Discussion

6.1   Summary

We have introduced the notion of ε-supervenience, which states that the Bayes optimal misclassification rate for any mind/brain property pair is less than ε. Furthermore, when we restrict to the setting of Thought Experiment 1, we have shown that k-NN classifiers are universally consistent, such that one can derive a hypothesis test, with confidence level α, that is guaranteed to converge to the Bayes optimal misclassification rate, given sufficient data, no matter the true (but unknown) distribution of mind/brain pair properties. Alas, this is a one-sided test, so although power converges to unity, one can never determine whether more data is necessary to get a lower p-value, or particular ε-supervenience does not hold.

6.2   Practical issues

Importantly, we are *not* claiming that actually determining ε-supervenience in humans when the vertices represent individual neurons is practically possible for any particular mental property at this time. Rather, the claim is that one had sufficient data (a very large number of exchangeable mind/brain property pairs), then *in theory*, some level ε-supervenience hypothesis test could be performed. Even in such a scenario, a universally consistent classifier is just one of many possible kinds of classifiers, and not necessarily the best one (in terms of $\widehat{L}$), for any particular dataset. Further, even if a universally consistent classifier is used, a more informative and tractable distance on $\mathcal{B}$ may be desired, as the k_n-nearest neighbor classifier under a Frobenius norm may have a rate of convergence so slow and a computational demand so high as to be impractical (though, see Appendix C for a simulated example in which convergence is relatively fast). Whichever classifier is used, it is likely to benefit from a large amount of domain specific knowledge, which the proposed classifier completely neglects.

6.3   A unified quest

A central (perhaps *the* central) quest in much of neuroscience, psychology, and cognitive science is to discover the brain properties that subvene under various mental properties, although questions are rarely cast within a supervenience formalism. Moreover, the particular brain properties that are often believed to subvene under these mental properties are neural circuits, or brain-subgraphs. To this end, many investigations in these fields include schematic diagrams showing a particular brain-subgraph subvening under a particular mental phenotype. This practice transcends the evolutionary hierarchy of neuroscientific research. For instance, in the invertebrate literature, vertices correspond to particular labeled neurons, and edges correspond to synapses [7]. In the vertebrate literature, vertices often correspond to types of neurons in particular regions, and edges correspond to tendencies of connections [8]. For primates [9] and humans [10] vertices frequently represent functionally distinct neuroanatomical regions, and edges represent regional interconnectivity. Furthermore, this practice also transcends analytical background, including anatomists [11], philosophers [12], statisticians [13], and physicists [14]. The near ubiquity of this practice suggests that a fundamental quest is to determine which brain-subgraphs subvene under which mental properties (although perhaps causality, not supervenience, is the true desideratum). Perhaps supervenience is therefore a framework that can fruitfully be applied to myriad and varied neuropsychological investigations.

6.4   Human applications

In recent years, with the advent of the field of "connectomics" [15, 16], neuroimaging has seen an explosion of studies investigating the human connectome, and relating connectomes to cognitive properties. These studies can all be thought of as testing various ε-supervenience hypotheses. For instance, a recent study showed that using data from diffusion tensor imaging [17], one can nearly perfectly differentiate between schizophrenics and controls [18]. As the resolution and signal-to-noise ratio of magnetic resonance imaging continue to improve, especially with more advanced techniques such as High Angular Diffusion Imaging [19], Q-Ball Imaging [20], and diffusion spectrum imaging [21], similar results could be obtained with other, more subtle cognitive properties. Furthermore, the utilization of other imaging technologies, such as polarized light imaging [22] and high-throughput electron microscopy [23, 24], will continue to improve the effective resolution of these inferred connectomes from human brains. While

determining from a brain scan whether a particular individual knows calculus might be quite distant, many other cognitive and psychological supervenience hypotheses have already been tested, and the gap between testing for calculus and testing for schizophrenia seems to be diminishing.

6.5   Quantum networks

Recently, several authors have suggested the possibility that brains are better characterized as quantum computers, instead of classical computers [25, 26]. Assuming that the computer can be represented as a network, the above results hold regardless of whether computations in the brain are quantum or classical. This follows because quantum networks merely speed up computation for certain classes of problems, they cannot, however, solve problems that classical computers cannot [27]. This means, if the above analysis failed to reject the null at level α, the interpretation does not change if one assumes quantum versus classical computations.

6.6   Stochastic Supervenience

Possible explanations of how it might be the case that $\varepsilon > 0$ include stochastic supervenience [28], and supernatural causal effects. Thus, the above analysis could be considered an empirical test for whether we have souls, or, perhaps whether souls play a causal role in our mental properties over and above the physical role played by the brain, or whether the data we have suggests that the probability that our souls play a measurable causal role over and above the physical is less than ε.

6.7   Dynamics vs. statics

The above *gedanken* experiment did not require simulating any dynamics; rather, the dynamics are necessarily a function of the model parameters (statics). Similarly, for the question of mind-brain supervenience in humans, one need not ever observe any activity of the brain, one must merely observe the model, which determines the activity (in a potentially stochastic process). Thus, this approach to understanding the relationship between mind and brain is distinct from the standard systems neuroscience paradigm, in which the goal is typically to understanding the neural activity "code." In contrast, if mind-brain supervenience is true, this motivates a search for the neural connectivity "code," an *engram* for memories [29, 30, 31, 32, 33], or more generally a *mengram*, the neural signature of any mental property, be it cognitive, psychological, or otherwise.

6.8   Concluding thoughts

This thought experiment, together with (i) the formal definition of ε-supervenience, (ii) the brain-graph model, and (iii) the universal consistency proof on graphs, is the first demonstration (to our knowledge) that empirically investigating supervenience is at least theoretically possible. The above discussion suggests that many previously conducted investigations either assume supervenience, or test it. Further, new technologies facilitate testing supervenience of mental properties on brain-graphs more easily.

A   Relations between sets

In this appendix we aim to provide more intuition regarding supervenience, by discussing the limitations and extent of its implications.

First, a supervenient relation does not imply an injective relation. An injective relation is any relation that preserves distinctness. Thus if minds are injective on brains, then $b \neq b' \implies m \neq m'$ (note that the directionality of the implication has been switched relative to supervenience). For instance, it might be the case that a brain could change without the mind changing. Consider the case that a single subatomic particle shifts its position by a Plank length, changing brain state from b to b'. It is possible (likely?) that the mental state supervening on brain state b remains m, even after b changes to b'. In such a scenario, the mind might still supervene on the brain, but the relation from brains to minds is not injective. This argument also shows that supervenience is not necessarily a *symmetric* relation. Minds supervening on brains does not imply that brains supervene on minds.

Second, supervenience does not imply causality. For instance, consider an analogy where M and B correspond to two coins being flipped, each possibly landing on heads or tails. Further assume that every time one lands on heads so does the other, and every time one lands on tails, so do the other. This implies that M supervenes on B, but assumes nothing about whether M causes B, or B causes M, or some exogenous force causes both.

Third, supervenience does not imply identity. Consider, for example, acceleration and velocity. Clearly, acceleration supervenes on velocity, as acceleration cannot change without velocity changing (assuming one does not consider gravity as acceleration). Similarly, velocity supervenes on position, as velocity cannot change without position changing. Therefore, acceleration supervenes on position, by the transitive property of supervenience, but it is not the case that a change in acceleration is equal to a change in position. Rather, position can change with constant velocity, meaning without acceleration changing.
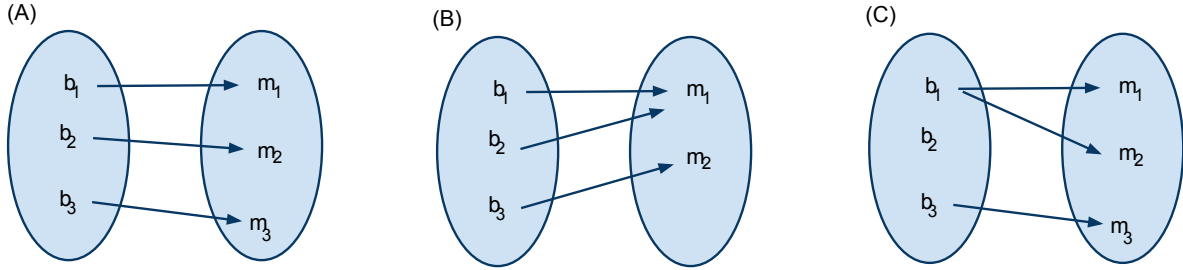

Figure 1: Possible relations between minds and brains. (A) Minds supervene on brains, and it so happens that there is a bijective relation from brains to minds. (B) Minds supervene on brains, and it so happens that there is a surjective (a.k.a., onto) relation from brains to minds. (C) Minds are *not* supervenient on brains, because two different minds supervene on the same brain.

What supervenience does imply, however, is the following. Imagine finding two different minds. If $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$, then the brains subvening under those two minds must be different. In other words, there cannot be two different minds, either of which could supervene on a single brain. Figure 1 shows several possible relations between the sets of minds and brains.

Note that statistical supervenience is distinct from statistical correlation. *Statistical correlation* between brain states and mental states is defined as $\rho_{MB} = \mathbb{E}[(B - \mu_B)(M - \mu_M)]/(\sigma_B \sigma_M)$, where μ_X and σ_X are the mean and variance of X, and $\mathbb{E}[X]$ is the expected value of X. If $\rho_{MB} = 1$, then both $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ and $\mathcal{B} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{M}$. Thus, perfect correlation implies supervenience, but supervenience does not imply correlation.

B k_n nearest neighbor algorithm

Consider the following problem setup. We have a collection of training data, $\mathcal{T}_n = \{(m_i, b_i)\}_{i=1}^n$, each sampled exchangably from some unknown joint distribution, $(m_i, b_i) \overset{iid}{\sim} \mathbb{P}[M, B]$, where m_i and b_i are the observed mental and brain properties of experiment i, respectively. A new brain, b, called the "test brain", is then observed, and one

desires to find the most likely class of the new brain, m. It is further assumed that the test mind/brain pair is sampled from the same distribution as the training data, $(m, b) \sim \mathbb{P}[M, B]$, and m is unobserved. Further assume that m can take one of a finite number of possible values, that is, $|\mathcal{M}| < \infty$.

The 1-nearest neighbor (1-NN) classifier works as follows. Compute the distance between the test brain and all the training brains, $d_i = d(b, b_i)$ for all $i \in [n]$, where $[n] = 1, 2, \ldots, n$. Then, sort them, $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$, and their corresponding mental properties, , $m_{(1)}, m_{(2)}, \ldots, m_{(n)}$, where parenthetical indices indicate rank order. The 1-NN algorithm predicts that the unobserved mind is of the same class as the closest brain's class: $\widehat{m} = m_{(1)}$. The k_n nearest neighbor is a straightforward generalization of this approach. It says that the test mind is in the same class as which ever class is the plurality class of the k_n nearest neighbors, $\widehat{m} = \text{argmax}_{m'} \mathbb{I}\{\sum_{i=1}^{k_n} m_{(i)} = m'\}$. Given a particular choice of k_n (the number of nearest neighbors to consider), and a choice of $d(\cdot, \cdot)$ (the distance metric used to compare the test datum and training data), one then has a relatively simple and intuitive algorithm.

Unfortunately, no such algorithm is universally consistent. Let g_n be the k_n nearest neighbor classifier when there are n training points. Then, a collection of such algorithms, $\{g_n\}$, with k_n increasing with n, can be universally consistent under certain constraints. In particular, as n increases, k_n must also increase, but not quite as quickly. Formally, k_n must satisfy: (i) $k_n \to \infty$ as $n \to \infty$ and (ii) $k_n/n \to 0$ as $n \to \infty$. In Stone' original proof, b was assumed to be a d-dimensional vector, and the L_2 norm ($d(b, b') = \sum_{j=1}^{d} (b_j - b'_j)^2$, where j indexes elements of the d-dimensional vector) was shown to satisfy the constraints on a distance metric for this collection of classifiers to be universally consistent. Later, others extended these results to apply to any L_p norm [4]. When brain-graphs are represented by their adjacency matrices, one can stack the columns of the adjacency matrices, effectively embedding graphs into finite Euclidean space, in which case Stone's theorem applies. Stone's original proof applied to the scenario when $|\mathcal{M}|$ was infinite, resulting in a regression algorithm as well.

C Simulation

As an example of a feasible experiment, one may consider a species whose nervous system consists of the same (small) number of labeled neurons for each organism. *Caenorhabditis elegans* is believed to be such a species [34]. The hermaphroditic C. elegans' somatic nervous system consists of 279 interconnected neurons. While the graph with these neurons as vertices and edges defined by chemical synapses between neurons is not identical across individuals, it is reasonably consistent [34]. Furthermore, these animals exhibit a rich behavioral repertoire that depends on circuit properties [35]. Thus, one may design an experiment by describing the joint distribution $\mathbb{P}[M, B]$ via class-conditional distributions $\mathbb{P}[B|M = m_j]$ for the C. elegans brain-graph for two mental properties of interest, m_0 and m_1, along with the prior probability of class membership $\mathbb{P}[M = m_1]$. Here the mental property corresponds to the C. elegans exhibiting (or not exhibiting) a particular behavior (e.g., response to an odor).

Simulations suggest that one may build a classifier, practically and with a manageable training sample size n, that demonstrates ε-supervenience with reasonable choices for ε and α, and a plausible joint distribution $\mathbb{P}[M, B]$ (Figure 2). To generate the data, let E_{ij} be an integer-valued random variable whose value indicates the number of synapses (edges) between neurons (vertices) i and j. Let the class-conditional random variable $E_{ij}|M = m_0$ be distributed Poisson($A_{ij} + \eta$), where A_{ij} is the number of chemical synapses between neuron i and neuron j according to [36], with noise parameter $\eta = 0.05$. Let $\mathcal{E}$ be the set of edges deemed responsible for odor-evoked behavior according to [37]. Therefore, the distribution of these edges must differ between the two classes. The class-conditional random variable $E_{ij}|M = m_1$ is distributed Poisson($A_{ij} + z_{ij}$), where the signal parameter $z_{ij} = \eta$ for all edges not in $\mathcal{E}$, and z_{ij} is uniformly sampled from ▤▤] for all edges within $\mathcal{E}$.

We consider k_n-nearest neighbor classification of labeled multigraphs (directed, with loops) on 279 vertices, under Frobenius norm. The k_n-nearest neighbor classifier used here satisfies $k_n \to \infty$ as $n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, ensuring universal consistency. (Better classifiers can be constructed for the joint distribution $\mathbb{P}[M, B]$ used here; however, we demand universal consistency.)

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D superresolution imaging [38] combined with neurite tracing algorithms [39, 40, 41] allow the collection of a C. elegans brain-graph within a day. Genetic manipulations, laser ablations, and training paradigms can each be used to obtain a non-wild type population for use as $M = m_1$ [35], and the class of each organism (m_0 vs. m_1) can also be determined automatically [42].
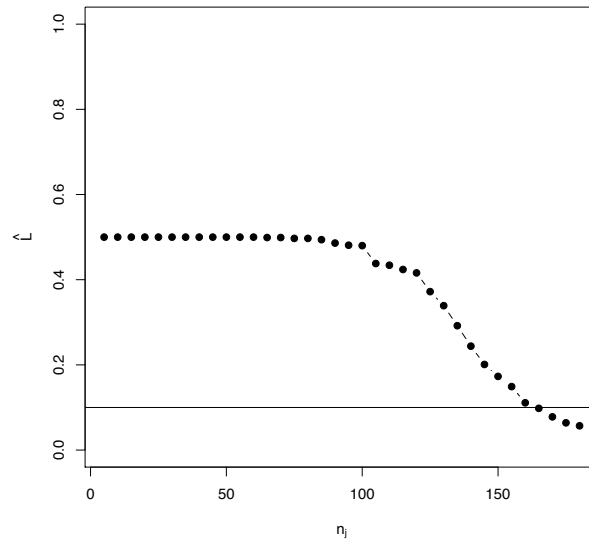
Figure 2: C. elegans graph classification simulation results. $\widehat{L}$ (the misclassification rate estimated upon with 1000 testing samples) is plotted as a function of class-conditional training sample size $n_j = \widetilde{n}/2$, suggesting that for $\varepsilon = 0.1$ we can determine that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ holds with 99% confidence with just a few hundred training samples generated from $\mathbb{P}[M, B]$. Each dot depicts an estimate for $L_{\mathbb{P}}(g_{\widetilde{n}})$; standard errors are $(L_{\mathbb{P}}(g_{\widetilde{n}})(1 - L_{\mathbb{P}}(g_{\widetilde{n}}))/1000)^{1/2}$; e.g., $n_j = 180$; $k_n = 53$; $\widehat{L}_F^{1000}(g_{\widetilde{n}}) = 0.057$; standard error less than 0.01. We reject $H_0 : L_{\mathbb{P}}(g^*) \geq 0.10$ at $\alpha = 0.01$. $L_{\mathbb{P}}(g^*) \approx 0$ for this simulation.

Acknowledgments

References

[1] Plato, *Plato: complete works*. Hackett Pub Co, 1997.

[2] R. Descartes, *Meditationes de prima philosophia*. 1641.

[3] D. Davidson, *Experience and Theory*, ch. Mental Events. Duckworth, 1970.

[4] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[5] L. Devroye, "On arbitrarily slow rates of global convergence in density estimation," *Probability Theory and Related Fields*, vol. 62, no. 4, pp. 475–483, 1983.

[6] K. Popper, "The logic of scientific discovery," 1959.

[7] R. J. G. Geoffrey North, ed., *Invertebrate neurobiology*. CSHL Press, 2007.

[8] G. Shepherd, *The synaptic organization of the brain*. Oxford University Press New York, 2004.

[9] D. Felleman and D. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral cortex*, vol. 1, no. 1, p. 1, 1991.

[10] P. v. Z. L. N.-P. S. Mori, S. Wakana, *MRI Atlas of Human White Matter*. Elsevier Science, 2005.

[11] M. Abeles, *Corticonics*. Cambridge University Press, 1991.

[12] C. Koch and J. Davis, *Large-scale neuronal theories of the brain*. The MIT Press, 1994.

[13] R. Rao, B. Olshausen, and M. Lewicki, *Probabilistic models of the brain: Perception and neural function*. The MIT Press, 2002.

[14] C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, eds., *Methods and Models in Neurophysics*. Elsevier, 2003.

[15] O. Sporns, G. Tononi, and R. Kotter, "The human connectome: A structural description of the human brain," *PLoS Computational Biology*, vol. 1, no. 4, p. e42, 2005.

[16] P. Hagmann, *From diffusion MRI to brain connectomics*. PhD thesis, Institut de traitement des signaux, 2005.

[17] P. J. Basser, J. Mattiello, and D. LeBihan, "Mr diffusion tensor spectroscopy and imaging.," *Biophys J*, vol. 66, pp. 259–267, Jan 1994.

[18] B. A. Ardekani, A. Tabesh, S. Sevy, D. G. Robinson, R. M. Bilder, and P. R. Szeszko, "Diffusion tensor imaging reliably differentiates patients with schizophrenia from healthy volunteers.," *Hum Brain Mapp*, Mar 2010.

[19] D. Tuch, T. Reese, M. Wiegell, N. Makris, J. Belliveau, and V. Wedeen, "High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity," *Magnetic Resonance in Medicine*, vol. 48, no. 4, pp. 577–582, 2002.

[20] D. Tuch, "Q-ball imaging," *Magnetic Resonance in Medicine*, vol. 52, no. 6, pp. 1358–1372, 2004.

[21] V. Wedeen, P. Hagmann, W. Tseng, T. Reese, and R. Weisskoff, "Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging," *Magnetic Resonance in Medicine*, vol. 54, no. 6, pp. 1377–1386, 2005.

[22] C. Palm, M. Axer, D. Gräßel, J. Dammers, J. Lindemeyer, K. Zilles, U. Pietrzyk, and K. Amunts, "Towards ultra-high resolution fibre tract mapping of the human brain–registration of polarised light images and reorientation of fibre vectors," *Frontiers in Human Neuroscience*, vol. 4, 2010.

[23] W. W. Denk and H. Horstmann, "Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure," *PLOS Biol.*, vol. 2, p. e329, 2004.

[24] K. Hayworth, N. Kasthuri, R. Schalek, and J. Lichtman, "Automating the collection of ultrathin serial sections for large volume TEM reconstructions," *Microscopy and Microanalysis*, vol. 12, no. S02, pp. 86–87, 2006.

[25] R. Penrose and M. Gardner, *The emperor's new mind: concerning computers, minds, and the laws of physics.* Oxford University Press, USA, 1999.

[26] J. Satinover, *The quantum brain: the search for freedom and the next generation of man.* Wiley, 2002.

[27] M. A. Nielson and I. L. Chuang, *Quantum Computation and Quantum Information.* Cambridge University Press, 2000.

[28] C. F. Craver, "Stochastic supervenience," 2009.

[29] R. W. Semon, *The Mneme.* G. Allen & Unwin ltd., 1921.

[30] K. Lashley, "In search of the engram," *Symposia of the society for experimental biology*, vol. 4, no. 454-482, p. 30, 1950.

[31] W. Zhang and D. Linden, "The other side of the engram: experience-driven changes in neuronal intrinsic excitability," *Nature Reviews Neuroscience*, vol. 4, no. 11, pp. 885–900, 2003.

[32] R. Shema, T. Sacktor, and Y. Dudai, "Rapid erasure of long-term memory associations in the cortex by an inhibitor of PKM {zeta}," *Science*, vol. 317, no. 5840, p. 951, 2007.

[33] J. Berry, W. Krause, and R. Davis, "Olfactory memory traces in Drosophila," *Progress in brain research*, vol. 169, pp. 293–304, 2008.

[34] R. M. Durbin, *Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans.* PhD thesis, University of Cambridge, 1987.

[35] M. de Bono and A. V. Maricq, "Neuronal substrates of complex behaviors in c. elegans.," *Annu Rev Neurosci*, vol. 28, pp. 451–501, 2005.

[36] L. Varshney, B. Chen, E. Paniagua, D. Hall, and D. Chklovskii, "Structural Properties of the Caenorhabditis elegans Neuronal Network," *ArXiv*, 2009.

[37] S. H. Chalasani, N. Chronis, M. Tsunozaki, J. M. Gray, D. Ramot, M. B. Goodman, and C. I. Bargmann, "Dissecting a circuit for olfactory behaviour in caenorhabditis elegans.," *Nature*, vol. 450, pp. 63–70, Nov 2007.

[38] A. Vaziri, J. Tang, H. Shroff, and C. V. Shank, "Multilayer three-dimensional super resolution imaging of thick biological samples.," *Proc Natl Acad Sci U S A*, vol. 105, pp. 20221–20226, Dec 2008.

[39] M. Helmstaedter, K. L. Briggman, and W. Denk, "3d structural imaging of the brain with photons and electrons.," *Curr Opin Neurobiol*, vol. 18, pp. 633–641, Dec 2008.

[40] Y. Mishchenko, "Automation of 3d reconstruction of neural tissue from large volume of conventional serial section transmission electron micrographs.," *J Neurosci Methods*, vol. 176, pp. 276–289, Jan 2009.

[41] J. Lu, J. C. Fiala, and J. W. Lichtman, "Semi-automated reconstruction of neural processes from large numbers of fluorescence images," *PLoS ONE*, vol. 4, p. e5655, 05 2009.

[42] S. D. Buckingham and D. B. Sattelle, "Strategies for automated analysis of c. elegans locomotion.," *Invert Neurosci*, vol. 8, pp. 121–131, Sep 2008.