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

The"mind-brain supervenience" conjecture suggests that all mental properties are derived from the physical properties of the brain. To address the question of whether the mind supervenes on the brain, we frame a supervenience hypothesis in rigorous statistical terms. Specifically, we propose a modified version of supervenience (called ε-supervenience) that is amenable to experimental investigation and statistical analysis. To illustrate this approach, we perform 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). ε-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given misclassification rate > 0, regardless of the relationship between the two. This may provide motivation for cross-disciplinary research between neuroscientists and statisticians.

The mind-brain supervenience notion was canonized in 1970 with the following quote from Donald Davidson: [1]

> [mind-brain] 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.

We consider a special case of supervenience of considerable interest. Specifically, this work addresses a novel version of a local supervenience between mental properties and brain properties, with particular emphasize on brain-graphs (i.e., connectivity structure). The determination of supervenience (or lack thereof) of a mental property on a brain property has potentially important implications in a number of fields of inquiry. Neural network theory and artificial intelligence often implicitly take a generalized notion of local mind-brain supervenience as an assumption; which if falsified, might change modern approaches to learning [2, 3]. Cognitive neuroscience similarly seems to operate under such assumptions, which if falsified, might result in novel perspectives and theories [4, 5]. And the question of mind-brain supervenience continues to be debated amongst philosophers [6].

This work does not attempt to resolve any particular mind-brain supervenience debates. Rather, we propose a statistical approach for framing mind-brain supervenience questions. This approach depends on defining the space of mental and brain properties under investigation and a statistical model characterizing the possible distributions governing their relationship. Such definitions transform supervenience from a conjecture or an assumption, into a hypothesis which can be tested.

Results

Statistical supervenience; a definition

Let $\mathcal{M} = \{m_1, m_2, \ldots\}$ be a set of possible mental properties. For example, m might indicate a person's intelligence, psychological state, current thought, gender identity, etc. Similarly, let $\mathcal{B} = \{b_1, b_2, \ldots\}$ be a set of possible brain properties. For example, b might denote the number of cells in a person's brain at some time t, or the collection of spike trains of all neurons in the brain during some time period t to t'. Given these definitions, Davidson's conjecture may be concisely and formally stated thusly: $m \neq m' \implies b \neq b'$, 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 1 for more details and some examples). To facilitate both statistical analysis and empirical investigation, we convert this supervenience relation from a logical to a probabilistic relation.

Let F_{MB} indicate a joint distribution of minds and brains. Statistical supervenience can thusly be defined:

Definition 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, 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 (which could be considered a generalization of correlation; see Appendix 1 for details).

Statistical supervenience is equivalent to perfect classification accuracy

If minds statistically supervene on brains, $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, then if two minds differ, there must be some brain-based difference to account for the mental difference. This means that there exists a deterministic function mapping each brain to its supervening mind, $g : \mathcal{B} \mapsto \mathcal{M}$. One could therefore, in principle, construct this function. When the space of all possible minds is finite; that is, $|\mathcal{M}| < \infty$, g is called a *classifier*. Define misclassification rate, the probability that g misclassifies b under distribution $F = F_{MB}$, as

$$L_F(g) = \mathbb{P}[g(B) \neq M] = \sum_{(m,b) \in \mathcal{M} \times \mathcal{B}} \mathbb{I}\{g(b) \neq m\} \mathbb{P}[B = b, M = m] \tag{1}$$

where $\mathbb{I}$ denotes the indicator function taking value unity whenever its argument is true, and zero otherwise. The Bayes optimal classifier g^* minimizes $L_F(g)$ over all classifiers, that is: $g^* = \operatorname{argmin}_g L_F(g)$. The *Bayes error*, or Bayes risk, $L^* = L_F(g^*)$, is the minimum possible misclassification rate.

The primary result of casting supervenience in a statistical framework is the following theorem:

Theorem 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $L^* = 0$. Formally, $\mathcal{M} \overset{S}{\sim}_F \mathcal{B} \Leftrightarrow L^* = 0$.*

Proof. Let $\mathcal{M}_b = \{m : \mathbb{P}[B = b|M = m] > 0\}$.

$$L^* = 0 \Leftrightarrow |\mathcal{M}_b| = 1 \,\forall\, b \in \mathcal{B}.$$

In words, for L^* to equal 0, it must be the case that there exists only a single m for each b. □

The above argument shows (for the first time to our knowledge) that statistical supervenience and zero Bayes error are equivalent. Statistical supervenience can therefore be thought of as a constraint on the possible distributions on minds and brains. Specifically, let $\mathcal{F}$ indicate the set of all possible joint distributions on minds and brains, and let $\mathcal{F}_s$ be subset of distributions for which supervenience holds. Theorem 1 implies that $\mathcal{F}_s = \{F_{MB} : L^* = 0\} \subseteq \mathcal{F}$.

Testing for statistical supervenience

The above theorem implies that if we desire to know whether a particular mental property is supervenient to a particular brain property, we must check whether $L^* = 0$. Unfortunately, L^* is typically unknown. Fortunately, we can approximate L^* using training data.

Let g_n be a classifier induced by the data, $g_n : \mathcal{B} \times (\mathcal{M} \times \mathcal{B})^n \mapsto \mathcal{M}$. Assume that training data $\mathcal{T}_n = \{(M_1, B_1), \ldots, (M_n, B_n)\}$ are each sampled identically and independently (iid) from the true (but unknown) joint distribution, F_{MB}. The misclassification rate of such a classifier is given by:

$$L_F(g_n) = \sum_{(m,b) \in \mathcal{M} \times \mathcal{B}} \mathbb{I}\{g_n(b; \mathcal{T}_n) \neq m\} \mathbb{P}[B = b, M = m], \tag{2}$$

Calculating the misclassification rate is often intractable in practice because it requires a sum over all possible mind/brain property pairs. Instead, misclassification rate is approximated by "hold-out" error. Let $\mathcal{H}_{n'} = \{(M_{n+1}, B_{n+1}), \ldots, (M_{n+n'}, B_{n+n'})\}$ be a set of n' hold-out samples, each sampled iid from F_{MB}. The hold-out approximation to misclassification rate is given by

$$\widehat{L}_F^{n'}(g_n) = \sum_{(m,b) \in \mathcal{H}_{n'}} \mathbb{I}\{g_n(b; \mathcal{T}_n) \neq m\}, \tag{3}$$

which can be used as a surrogate for L^*. By definition of g^*, the expectation of $\widehat{L}_F^{n'}(g_n)$ is greater than or equal to g^* for any g_n and all n. Thus, we can construct a hypothesis test for L^* using the surrogate $\widehat{L}_F^{n'}(g_n)$.

A statistical test proceeds by calculating a test statistic and a critical value c_α. We reject if the test statistic is more extreme than the critical value, or equivalently, if the p-value is less than α. The p-value, the probability of rejecting a true null hypothesis, is a functional of the distribution of the test statistic.

Ideally, we might consider the following hypothesis test: $H_0 : L^* > 0$ and $H_A : L^* = 0$. Unfortunately, because the null hypothesis is not closed, the alternate hypothesis lies on the boundary, so the p-value is always equal to unity [7]. To proceed, therefore, we introduce a relaxed notion of supervenience:

Definition 2. *Given $\varepsilon > 0$, $\mathcal{M}$ is said to ε-supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$, if and only if $L_F(g^*) < \varepsilon$.*

Given this relaxation, consider the problem of testing for ε-supervenience:

$$H_0^\varepsilon : L^* \geq \varepsilon$$
$$H_A^\varepsilon : L^* < \varepsilon$$

Let $\widehat{n} = n' \widehat{L}_F^{n'}(g_n)$ be the test statistic. The distribution of $\widehat{n}$ is available under the least favorable distribution of the alternate hypothesis. For the above hypothesis test, the p-value is given by the binomial cumulative distribution function, $\mathbb{B}(\widehat{n}; n', p_{H_A}) = \sum_{k \in [\widehat{n}]_0} \text{Binomial}(k; n'; p_{H_A})$, where p_{H_A} is Bernoulli probability under least favorable distribution of the alternate hypothesis and $[\widehat{n}]_0 = \{0, 1, \ldots, \widehat{n}\}$. In this composite alternate hypothesis, $p_{H_A} = \varepsilon$. We therefore reject whenever this p-value is less than α; rejection implies that we are $100(1-\alpha)$% confident that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$. The definition of ε-supervenience therefore admits, for the first time to our knowledge, a viable statistical test of supervenience, given a specified ε and α.

The power and limits of an ideal test of supervenience

Ideally, as n increases, the power of a test (the probability of a rejecting the null when it is false) goes to unity. For any statistical test, if the p-value converges in distribution to δ_0 (a point mass at zero), then whenever $\alpha > 0$, power goes to unity. For the above statistical test, if $g_n \to g^*$ as $n \to \infty$, then $L_F^{n'}(g_n) \to L^*$ as $n' \to \infty$. Thus, if $L^* < \varepsilon$, the p-value converges; that is, $\mathbb{B}(\widehat{n}; n', \widehat{L}_F^{n'}(g_n)) \to \delta_0$.

Unfortunately, the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ depends on the (unknown) distribution $F = F_{MB}$ [8]. Furthermore, arbitrarily slow convergence theorems regarding the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ demonstrate that there is no universal n, n' which will guarantee that the test has power greater than any specified target $\beta > \alpha$ [9]. 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 $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ holds, but we can never be confident in its negation; rather, it may be the case that the evidence in favor of $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ is insufficient for any number of reasons, including that we simply have not yet collected enough data. Thus, without restrictions on F_{MB}, arbitrarily slow convergence theorems imply that our theorem of ε-supervenience does not strictly satisfy Popper's *falsifiability* requirement [10].

The above considerations indicate that to obtain a statistical test with unity power, a consistent classifier (one for which $g_n \to g^*$ as $n \to \infty$) is required. Below, we show that under a very general mind-brain model, a consistent classifier is readily available.

***Gedankenexperiment* 1.** *Let the physical property under consideration be brain connectivity structure, so b is a brain-graph ("connectome") with vertices representing neurons (or collections thereof) and edges representing synapses (or collections thereof). Further let $\mathcal{B}$, the brain observation space, be the collection of all graphs on a finite number of vertices, and $\mathcal{M}$, the mental property observation space, be finite. Now, imagine collecting very large amounts of very accurate exchangeable 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 Methods for details). Therefore, the existence of a universally consistent classifier guarantees that eventually (in n, n') we will be able to conclude $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ 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, the proof holds for other matrix norms (which might speed up convergence), and the regression scenario, where $|\mathcal{M}|$ is infinite (again, see Methods for details).*

Discussion

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. This definition admits, for the first time to our knowledge, a viable statistical test for supervenience. Furthermore, when we restrict the space of minds and brains to the setting of *Gedankenexperiment* 1, we have shown that k_n-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, a failure to reject the null has many possible explanations. First, the amount of data might be insufficient given the particular distance implemented; collecting more data or utilizing a more informative distance might resolve this difficulty. Second, ε-supervenience is defined between a mental property and a brain property. Thus, a failure to reject supervenience on a particular brain property does not entail non-supervenience on any other brain property. Third, in general, neither mental properties nor brain properties are directly measurable; rather, one typically measures a function of such properties. For instance, instead of measuring intelligence, one considers the results of an IQ test as a proxy for intelligence. Similarly, instead of measuring neural spike trains, one estimates spike trains from some measurable neural signal, like voltage. While one could conceivably ask questions such as: "is IQ score at some time supervenient on the the output of a voltage measuring device?", these are less elegant and general than their non-observable analogs, such as: "is intelligence supervenient on spike trains?"

Thus, given measurement of mental and brain properties that we believe reflect the properties of interest, and a sufficient amount data satisfying the exchangeability assumption, a rejection of H_0^ε entails that we are $100(1-a)\%$ confident that the mental property under investigation does not ε-supervene on the brain property under investigation. Unfortunately, failure to reject is more ambiguous. $\varepsilon-$supervenience tests can therefore be thought of as constraining the space of possible brain properties upon which a mental property supervenes.

Determining that a mental property does not supervene on *any* brain property is beyond the capacities of this formalism.

Interestingly, much of contemporary research in neuroscience and cognitive science could be cast as mind-brain supervenience investigations. Specifically, any investigation that aims to "explain" some behavioral or psychological phenomenon by reference to neural (or other brain) activity or structure, could be thought of as mind-brain supervenience investigations. Thus, the proposed ε-supervenience framework is perhaps a useful unifying perspective for these fields. Perhaps this is especially true in light of the advent of "connectomics" [11, 12], a field devoted to estimating whole organism brain-graphs, and relating them to function. Testing supervenience of various mental properties on these brain-graphs will likely therefore become increasingly compelling; so the framework developed herein could be fundamental to these investigations. For example, questions about whether connectivity structure alone is sufficient to explain a particular mental property is one possible mind/brain ε-supervenience investigation. Appendix 2 presents an illustrative simulated example using the only currently known connectome—that of a Caenorhabditis elegans—demonstrating that ε-supervenience can be rejected at some critical value c_α with only a reasonably small number of samples. Similar supervenience tests on larger animals will require either higher-throughput imaging modalities [13, 14] or more coarse brain-graphs [15, 16], or both.

Methods

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, $\widehat{m} = \operatorname{argmax}_{m'} \mathbb{I}\{\sum_{i=1}^{k_n} m_{(i)} = m'\}$. 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.

Let g_n be the k_n nearest neighbor (k_nNN) classifier when there are n training samples. A collection of such classifiers $\{g_n\}$, with k_n increasing with n, is called a classifier sequence. A universally consistent classifier sequence is any classifier sequence that is guaranteed to converge to the Bayes optimal classifier regardless of the true distribution from which the data were sampled; that is, a universally consistent classifier sequence satisfies $g_n \to g^*$ as $n \to \infty$ for all F_{MB}.

The k_nNN classifier is consistent if (i) $k_n \to \infty$ as $n \to \infty$ and (ii) $k_n/n \to 0$ as $n \to \infty$ [17]. In Stone's original proof [17], 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 [8]. When brain-graphs are represented by their adjacency matrices, one can stack the columns of the adjacency matrices, effectively embedding graphs into a vector space, in which case Stone's theorem applies. Stone's original proof also applied to the scenario when $|\mathcal{M}|$ was infinite, resulting in a universally consistent regression algorithm as well.

Note that the above extension of Stone's original theorem to the graph domain implicitly assumed that vertices were labeled, such that elements of the adjacency matrices could easily be compared across graphs. In theory, when vertices are unlabeled, one could first map each graph to a quotient space invariant to isomorphisms, and then proceed as before. Unfortunately, there is no known polynomial time complexity algorithm for graph isomorphism [18], so in practice, dealing with unlabeled vertices will likely be computationally challenging.

References

[1] Davidson, D. *Experience and Theory*, chapter Mental Eve. Duckworth (1970).

[2] Haykin, S. *Neural Networks and Learning Machines*. Prentice Hall, 3rd edition, (2008).

[3] Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, (2008).

[4] Fodor, J. A. *Concepts: Where Cognitive Science Went Wrong (Oxford Cognitive Science)*. Oxford University Press, USA, (1998).

[5] Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. *Cognitive Neuroscience: The Biology of the Mind (Third Edition)*. W. W. Norton & Company, (2008).

[6] Kim, J. *Physicalism, or Something Near Enough (Princeton Monographs in Philosophy)*. Princeton University Press, (2007).

[7] Bickel, P. J. and Doksum, K. A. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I (2nd Edition)*. Prentice Hall, (2000).

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

[9] Devroye, L. *Utilitas Mathematica* **483**, 475–483 (1983).

[10] Popper, K. R. (1959).

[11] Sporns, O., Tononi, G., and Kotter, R. *PLoS Computational Biology* **1**(4), e42 (2005).

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

[13] Hayworth, K. J., Kasthuri, N., Schalek, R., Lichtman, J. W., Program, N., Angeles, L., and Biology, C. *World* **12**(Supp 2), 86–87 (2006).

[14] Bock, D. D., Lee, W.-C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzel, A. W., Yurgenson, S., Soucy, E. R., Kim, H. S., and Reid, R. C. *Nature* **471**(7337), 177–182 March (2011).

[15] Palm, C., Axer, M., Gräß el, D., Dammers, J., Lindemeyer, J., Zilles, K., Pietrzyk, U., and Amunts, K. *Frontiers in Human Neuroscience* **4** (2010).

[16] Johansen-Berg, H. and Behrens, T. E. *Diffusion MRI: From quantitative measurement to in-vivo neuroanatomy*. Academic Press, (2009).

[17] Stone, C. J. *The Annals of Statistics* **5**(4), 595–620 July (1977).

[18] Garey, M. R. and Johnson, D. S. *Computers and intractability. A guide to the theory of NP-completeness. A Series of Books in the Mathematical Sciences*. WH Freeman and Company, San Francisco, Calif, (1979).

Acknowledgments

The authors would like to acknowledge helpful discussions with J. Lande, B. Vogelstein, and S. Seung.

Author Contributions

JTV, RJV, and CEP conceived of the manuscript. JTV and CEP wrote it. CEP ran the experiment.

Additional Information

The authors have no competing financial interests to declare.