TOOLS AND RESOURCES

Discovering and deciphering relationships across disparate data modalities

Joshua T Vogelstein[1,2]*, Eric W Bridgeford[1], Qing Wang[1], Carey E Priebe[1], Mauro Maggioni[1], Cencheng Shen[3]

[1]Johns Hopkins University, Baltimore, United States; [2]Child Mind Institute, New York, United States; [3]University of Delaware, Delaware, United States

Abstract Understanding the relationships between different properties of data, such as whether a genome or connectome has information about disease status, is increasingly important. While existing approaches can test whether two properties are related, they may require unfeasibly large sample sizes and often are not interpretable. Our approach, 'Multiscale Graph Correlation' (MGC), is a dependence test that juxtaposes disparate data science techniques, including k-nearest neighbors, kernel methods, and multiscale analysis. Other methods may require double or triple the number of samples to achieve the same statistical power as MGC in a benchmark suite including high-dimensional and nonlinear relationships, with dimensionality ranging from 1 to 1000. Moreover, MGC uniquely characterizes the latent geometry underlying the relationship, while maintaining computational efficiency. In real data, including brain imaging and cancer genetics, MGC detects the presence of a dependency and provides guidance for the next experiments to conduct.

DOI: https://doi.org/10.7554/eLife.41690.001

Introduction

Identifying the existence of a relationship between a pair of properties or modalities is the critical initial step in data science investigations. Only if there is a statistically significant relationship does it make sense to try to decipher the nature of the relationship. Discovering and deciphering relationships is fundamental, for example, in high-throughput screening (***Zhang et al., 1999***), precision medicine (***Prescott, 2013***), machine learning (***Hastie et al., 2001***), and causal analyses (***Pearl, 2000***). One of the first approaches for determining whether two properties are related to—or statistically dependent on—each other is Pearson's Product-Moment Correlation (published in 1895; ***Pearson, 1895***). This seminal paper prompted the development of entirely new ways of thinking about and quantifying relationships (see ***Reimherr and Nicolae, 2013*** and ***Josse and Holmes, 2013*** for recent reviews and discussion). Modern datasets, however, present challenges for dependence-testing that were not addressed in Pearson's era. First, we now desire methods that can correctly detect any kind of dependence between all kinds of data, including high-dimensional data (such as 'omics), structured data (such as images or networks), with nonlinear relationships (such as oscillators), even with very small sample sizes as is common in modern biomedical science. Second, we desire methods that are interpretable by providing insight into how or why they discovered the presence of a statistically significant relationship. Such insight can be a crucial component of designing the next computational or physical experiment.

While many statistical and machine learning approaches have been developed over the last 120 years to combat aspects of the first issue—detecting dependencies—no approach satisfactorily addressed the challenges across all data types, relationships, and dimensionalities. Hoeffding and Renyi proposed non-parametric tests to address nonlinear but univariate relationships (***Hoeffding, 1948***; ***Rényi, 1959***). In the 1970s and 1980s, nearest neighbor style approaches were

***For correspondence:** jovo@jhu.edu

Competing interests: The authors declare that no competing interests exist.

Funding: See page 23

Received: 03 September 2018
Accepted: 14 January 2019
Published: 15 January 2019

Reviewing editor: Dane Taylor, University of Buffalo, United States

eLife digest If you want to estimate whether height is related to weight in humans, what would you do? You could measure the height and weight of a large number of people, and then run a statistical test. Such 'independence tests' can be thought of as a screening procedure: if the two properties (height and weight) are not related, then there is no point in proceeding with further analyses.

In the last 100 years different independence tests have been developed. However, classical approaches often fail to accurately discern relationships in the large, complex datasets typical of modern biomedical research. For example, connectomics datasets include tens or hundreds of thousands of connections between neurons that collectively underlie how the brain performs certain tasks. Discovering and deciphering relationships from these data is currently the largest barrier to progress in these fields. Another drawback to currently used methods of independence testing is that they act as a 'black box', giving an answer without making it clear how it was calculated. This can make it difficult for researchers to reproduce their findings – a key part of confirming a scientific discovery. Vogelstein et al. therefore sought to develop a method of performing independence tests on large datasets that can easily be both applied and interpreted by practicing scientists.

The method developed by Vogelstein et al., called Multiscale Graph Correlation (MGC, pronounced 'magic'), combines recent developments in hypothesis testing, machine learning, and data science. The result is that MGC typically requires between one half to one third as big a sample size as previously proposed methods for analyzing large, complex datasets. Moreover, MGC also indicates the nature of the relationship between different properties; for example, whether it is a linear relationship or not.

Testing MGC on real biological data, including a cancer dataset and a human brain imaging dataset, revealed that it is more effective at finding possible relationships than other commonly used independence methods. MGC was also the only method that explained how it found those relationships.

MGC will enable relationships to be found in data across many fields of inquiry – and not only in biology. Scientists, policy analysts, data journalists, and corporate data scientists could all use MGC to learn about the relationships present in their data. To that extent, Vogelstein et al. have made the code open source in MATLAB, R, and Python.

DOI: https://doi.org/10.7554/eLife.41690.002

popularized (**Friedman and Rafsky, 1983**; **Schilling, 1986**), but they were sensitive to algorithm parameters resulting in poor empirical performance. 'Energy statistics', and in particular the distance correlation test (Dcorr), was recently shown to be able to detect any dependency with sufficient observations, at arbitrary dimensions, and structured data under a proper distance metric (**Székely et al., 2007**; **Székely and Rizzo, 2009**; **Szekely and Rizzo, 2013**; **Lyons, 2013**). Another set of methods, referred to a 'kernel mean embedding' approaches, including the Hilbert Schmidt Independence Criterion (Hsic) (**Gretton and Gyorfi, 2010**; **Muandet et al., 2017**), have the same theoretical guarantees, which is shown to be a kernel version of the energy statistics (**Sejdinovic et al., 2013**; **Shen and Vogelstein, 2018**). The energy statistics can perform very well with a relatively small sample size on high-dimensional linear data, whereas the kernel methods and another test (Heller, Heller, and Gorfine's test, Hhg) (**Heller et al., 2013**) perform well on low-dimensional nonlinear data. But no test performs particularly well on high-dimensional nonlinear data with typical sample sizes, which characterizes a large fraction of real data challenges in the current big data era.

Moreover, to our knowledge, existing dependency tests do not attempt to further characterize the dependency structure. On the other hand, much effort has been devoted to characterizing 'point cloud data', that is, summarizing certain global properties in unsupervised settings (for example, having genomics data, but no disease data). Classic examples of such approaches include Fourier (**Bracewell and Bracewell, 1986**) and wavelet analysis (**Daubechies, 1992**). More recently, topological and geometric data analysis compute properties of graphs, or even higher order simplices (**Edelsbrunner and Harer, 2009**). Such methods build multiscale characterization of the samples, much like recent developments in harmonic analysis (***Coifman and Maggioni, 2006***; ***Allard et al., 2012***). However, these tools typically lack statistical guarantees under noisy observations and are often computationally burdensome.

We surmised that both (i) empirical performance in all dependency structures, in particular high-dimensional, nonlinear, low-sample size settings, and (ii) providing insight into the discovery process, can be addressed via extending existing dependence tests to be *adaptive* to the data (***Zhang et al., 2012***). Existing tests rely on a fixed *a priori* selection of an algorithmic parameter, such as the kernel bandwidth (***Gretton et al., 2006***), intrinsic dimension (***Allard et al., 2012***), and/or local scale (***Friedman and Rafsky, 1983***; ***Schilling, 1986***). Indeed, the Achilles Heel of manifold learning has been the requirement to manually choose these parameters (***Levina and Bickel, 2004***). Post-hoc cross-validation is often used to make these methods effectively adaptive, but doing so adds an undesirable computational burden and may weaken or destroy any statistical guarantees. There is therefore a need for statistically valid and computationally efficient adaptive methods.

To illustrate the importance of adapting to different kinds of relationships, consider a simple illustrative example: investigate the relationship between cloud density and grass wetness. If this relationship were approximately linear, the data might look like those in ***Figure 1A*** (top). On the other hand, if the relationship were nonlinear—such as a spiral—it might look like those in ***Figure 1A*** (bottom). Although the relationship between clouds and grass is unlikely to be spiral, spiral relationships

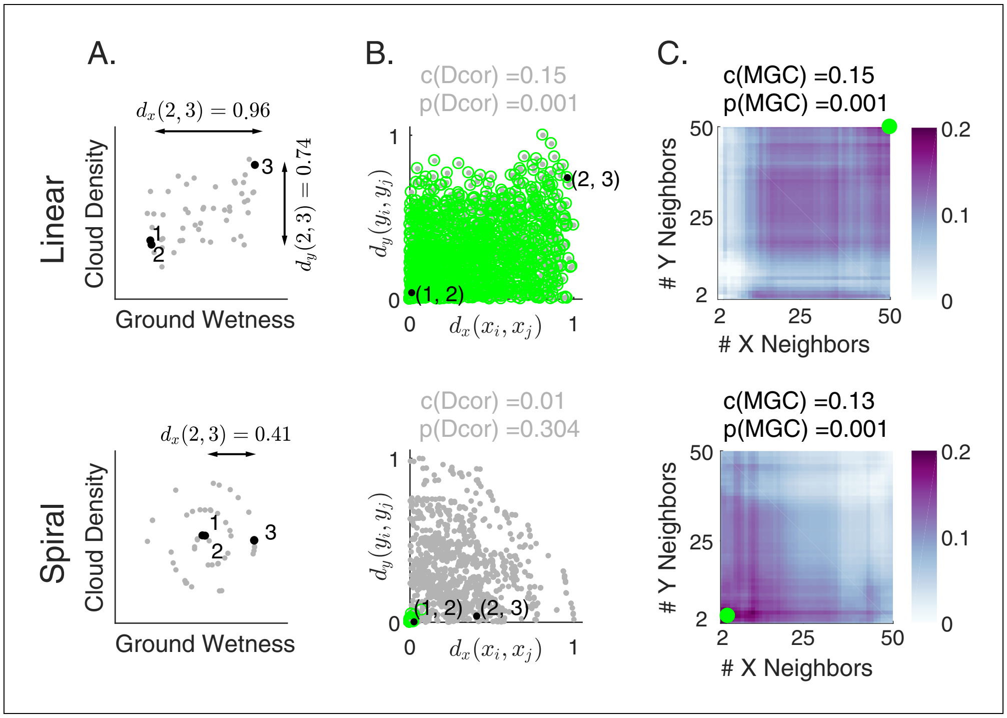


Figure 1. Illustration of Multiscale Graph Correlation (Mgc) on simulated cloud density (x_i) and grass wetness (y_i). We present two different relationships: linear (top) and nonlinear spiral (bottom; see Materials and methods for simulation details). (**A**) Scatterplots of the raw data using 50 pairs of samples for each scenario. Samples 1, 2, and 3 (black) are highlighted; arrows show x distances between these pairs of points while their y distances are almost 0. (**B**) Scatterplots of all pairs of distances comparing x and y distances. Distances are linearly correlated in the linear relationship, whereas they are not in the spiral relationship. Dcorr uses all distances (gray dots) to compute its test statistic and p-value, whereas Mgc chooses the local scale and then uses only the local distances (green dots). (**C**) Heatmaps characterizing the strength of the generalized correlation at all possible scales (ranging from 2 to n for both x and y). For the linear relationship, the global scale is optimal, which is the scale that Mgc selects and results in a p-value identical to Dcorr. For the nonlinear relationship, the optimal scale is local in both x and y, so Mgc achieves a far larger test statistic, and a correspondingly smaller and significant p-value. Thus, Mgc uniquely detects dependence and characterizes the geometry in both relationships.

DOI: https://doi.org/10.7554/eLife.41690.003

are prevalent in nature and mathematics (for example, shells, hurricanes, and galaxies), and are canonical in evaluations of manifold learning techniques (**Lee and Verleysen, 2007**), thereby motivating its use here.

Under the linear relationship (top panels), when a pair of observations are close to each other in cloud density, they also tend to be close to each other in grass wetness (for example, observations 1 and 2 highlighted in black in ***Figure 1A***, and distances between them in ***Figure 1B***). Similarly, when a pair of observations are far from each other in cloud density, they also tend to be far from each other in grass wetness (see for example, distances between observations 2 and 3). On the other hand, consider the nonlinear (spiral) relationship (bottom panels). Here, when a pair of observations are close to each other in cloud density, they also tend to be close to each other in grass wetness (see points 1 and 2 again). However, the same is not true for large distances (see points 2 and 3). Thus, in the linear relationship, the distance between every pair of points is informative with respect to the relationship, while under the nonlinear relationship, only a subset of the distances are.

For this reason, we juxtapose nearest neighbor mechanism with distance methods. Specifically, for each point, we find its k-nearest neighbors for one property (e.g. cloud density), and its l-nearest neighbors for the other property (e.g. grass wetness); we call the pair (k, l) the 'scale'. *A priori*, however, we do not know which scales will be most informative. We compute all distance pairs, then efficiently compute the distance correlations for all scales. The local correlations (***Figure 1C***, described in detail below) illustrate which scales are relatively informative about the relationship. The key, therefore, to successfully discover and decipher relationships between disparate data modalities is to adaptively determine which scales are the most informative, and the geometric implication for the most informative scales. Doing so not only provides an estimate of whether the modalities are related, but also provides insight into how the determination was made. This is especially important in high-dimensional data, where simple visualizations do not reveal relationships to the unaided human eye.

Our method, 'Multiscale Graph Correlation' (MGC, pronounced 'magic'), generalized and extends previously proposed pairwise comparison-based approaches by adaptively estimating the informative scales for any relationship — linear or nonlinear, low-dimensional or high-dimensional, unstructured or structured—in a computationally efficient and statistically valid and consistent fashion. This adaptive nature of MGC effectively guarantees an improved statistical performance. Moreover, the dependency strength across all scales is informative about the structure of a statistical relationship, therefore providing further guidance for subsequent experimental or analytical steps. MGC is thus a hypothesis-testing and insight-providing approach that builds on recent developments in manifold and kernel learning, with complementary developments in nearest-neighbor search, and multiscale analyses.

The multiscale graph correlation procedure

MGC is a multi-step procedure to discover and decipher dependencies across disparate data modalities or properties. Given n samples of two different properties, proceed as follows (see Materials and methods and (**Shen et al., 2018**) for details):

1. Compute two distance matrices, one consisting of distances between all pairs of one property (e.g. cloud densities, entire genomes or connectomes) and the other consisting of distances between all pairs of the other property (e.g. grass wetnesses or disease status). Then center each distance matrix (by subtracting its overall mean, the column-wise mean from each column, and the row-wise mean from each row), and denote the resulting n-by-n matrices A and B.
2. For all possible values of k and l from 1 to n:
    a. Compute the k-nearest neighbor graphs for one property, and the l-nearest neighbor graphs for the other property. Let G_k and H_l be the adjacency matrices for the nearest neighbor graphs, so that $G_k(i,j) = 1$ indicates that $A(i,j)$ is within the k smallest values of the i^{th} row of A, and similarly for H_l.
    b. Estimate the local correlations—the correlation between distances restricted to only the (k, l) neighbors—by summing the products of the above matrices, $c^{kl} = \sum_{ij} A(i,j)G_k(i,j)B(i,j)H_l(i,j)$.
    c. Normalize c^{kl} such that the result is always between -1 and $+1$ by dividing by $\sqrt{\sum_{ij} A^2(i,j)G_k(i,j) \times \sum_{ij} B^2(i,j)H_l(i,j)}$.

3. Estimate the optimal local correlation c^* by finding the smoothed maximum of all local correlations $\{c^{kl}\}$. Smoothing mitigates biases and provides MGC with theoretical guarantee and better finite-sample performance.
4. Determine whether the relationship is significantly dependent—that is, whether c^* is more extreme than expected under the null—via a permutation test. The permutation procedure repeats steps 1–4 on each permutation, thereby eliminating the multiple hypothesis testing problem by only computing one overall p-value, rather than one p-value per scale, ensuring that it is a valid test (meaning that the false positive rate is properly controlled at the specified type I error rate).

Computing all local correlations, the test statistic, and the p-value requires $O(n^2 \log n)$ time, which is about the same running time complexity as other methods (***Shen et al., 2018***).

Results

MGC typically requires substantially fewer samples to achieve the same power across all dependencies and dimensions

When, and to what extent, does MGC outperform other approaches, and when does it not? To address this question, we formally pose the following hypothesis test (see Materials and methods for details):

$$H_0 : X \, \mathrm{and} \, Y \, \mathrm{are\ independent}$$
$$H_A : X \, \mathrm{and} \, Y \, \mathrm{arenotindependent}.$$

The standard criterion for evaluating statistical tests is the testing power, which equals the probability that a test correctly rejects the null hypothesis at a given type one error level, that is power = Prob(H_0 is rejected $|H_0$ is false). The higher the testing power, the better the test procedure. A consistent test has power converging to 1 under dependence, and a valid test controls the type one error level under independence. In a complementary manuscript (***Shen et al., 2018***), we established the theoretical properties of MGC, proving its validity and universal consistency for dependence testing against all distributions of finite second moments.

Here, we address the empirical performance of MGC as compared with multiple popular tests: (i) DCORR, a popular approach from the statistics community (***Székely et al., 2007***; ***Székely and Rizzo, 2009***), (ii) MCORR, a modified version of DCORR designed to be unbiased for sample data (***Szekely and Rizzo, 2013***), (iii) HHG, a distance-based test that is very powerful for detecting low-dimensional nonlinear relationships (***Heller et al., 2013***). (iv) HSIC, a kernel dependency measure (***Gretton and Gyorfi, 2010***) formulated in the same way as DCORR except operating on kernels, (v) MANTEL, which is historically widely used in biology and ecology (***Mantel, 1967***). (vi) RV coefficient (***Pearson, 1895***; ***Josse and Holmes, 2013***), which is a multivariate generalization of PEARSON's product moment correlation whose test statistic is the sum of the trace-norm of the cross-covariance matrix, and (vii) the CCA method, which is the largest (in magnitude) singular value of the cross-covariance matrix, and can be viewed as a different generalization of PEARSON in high-dimensions that is more appropriate for sparse settings (***Hotelling, 1936***; ***Witten et al., 2009***; ***Witten and Tibshirani, 2011***). Note that while we focus on high-dimensional settings, Appendix 1 shows further results in one-dimensional settings, also comparing to a number of tests that are limited to one dimension, including: (viii) PEARSON's product moment correlation, (ix) SPEARMAN's rank correlation (***Spearman, 1904***), (x) KENDALL's tau correlation (***Kendall, 1970***), and (xi) MIC (***Reshef et al., 2011***). Under the regularity condition that the data distribution has finite second moment, the first four tests are universally consistent, whereas the other tests are not.

We generate an extensive benchmark suite of 20 relationships, including different polynomial (linear, quadratic, cubic), trigonometric (sinusoidal, circular, ellipsoidal, spiral), geometric (square, diamond, W-shape), and other functions. This suite includes and extends the simulated settings from previous dependence testing work (***Székely et al., 2007***; ***Simon and Tibshirani, 2012***; ***Gorfine et al., 2012***; ***Heller et al., 2013***; ***Szekely and Rizzo, 2013***). For many of them, we introduce high-dimensional variants, to more extensively evaluate the methods; function details are in Materials and methods. The visualization of one-dimensional noise-free (black) and noisy (gray) samples is shown in ***Figure 2—figure supplement 1***. For each relationship, we compute the power of each method relative to MGC for ~20 different dimensionalities, ranging from 1 up to 10, 20, 40, 100, or 1000. The high-dimensional relationships are more challenging because (1) they cannot be easily visualized and (2) each dimension is designed to have less and less signal, so there are many noisy dimensions. ***Figure 2*** shows that MGC achieves the highest (or close to the highest) power given 100 samples for each relationship and dimensionality. ***Figure 2—figure supplement 2*** shows the same advantage in one-dimension with increasing sample size.

Moreover, for each relationship and each method we compute the required sample size to achieve power 85% at error level 0.05, and summarize the median size for monotone relationships (type 1–5) and non-monotone relationships (type 6–19) in ***Table 1***. Other methods typically require double or triple the number of samples as MGC to achieve the same power. More specifically, traditional correlation methods (PEARSON, RV, CCA, SPEARMAN, KENDALL) always perform the best in monotonic simulations, distance-based methods including MCORR, DCORR, MGC, HHG and HSIC are slightly worse, while MIC and MANTEL are the worst. MGC's performance is equal to linear methods on monotonic relationships. For non-monotonic relationships, traditional correlations fail to detect the existence of dependencies, DCORR, MCORR, and MIC, do reasonably well, but HHG and MGC require the fewest samples. In the high-dimensional non-monotonic relationships that motivated this work, and are common in biomedicine, MGC significantly outperforms other methods. The second best test that is universally consistent (HHG) requires nearly double as many samples as MGC, demonstrating that MGC could half the time and cost of experiments designed to discover relationships at a given effect size.

MGC extends previously proposed global methods, such as MANTEL and DCORR . The above experiments extended MCORR , because MCORR is universally consistent and an unbiased version of DCORR (***Szekely and Rizzo, 2013***). ***Figure 2—figure supplement 3*** directly compares multiscale generalizations of MANTEL and MCORR as dimension increases, demonstrating that empirically, MGC nearly dominates its global variant for essen- tially all dimensions and simulation settings considered here. ***Figure 2—figure supplement 4*** shows a similar result for one-dimensional settings while varying sample size. Thus, not only does MGC empirically nearly dominate existing tests, it is a framework that one can apply to future tests to further improve their performance.

MGC deciphers latent dependence structure

Beyond simply testing the existence of a relationship, the next goal is often to decipher the nature or structure of the relationship, thereby providing insight and guiding future experiments. A single scalar quantity (such as effect size) is inadequate given the vastness and complexities of possible relationships. Existing methods would require a secondary procedure to characterize the relationship, which introduces complicated 'post selection' statistical quandaries that remain mostly unresolved (***Berk et al., 2013***). Instead, MGC provides a simple, intuitive, and nonparametric (and therefore infinitely flexible) 'map' of how it discovered the relationship. As described below, this map not only provides interpretability for how MGC detected a dependence, it also partially characterize the geometry of the investigated relationship.

The MGC-*Map* shows local correlation as a function of the scales of the two properties. More concretely, it is the matrix of c^{kl}'s, as defined above. Thus, the MGC-Map is an n-by-n matrix which encodes the strength of dependence for each possible scale. ***Figure 3*** provides the MGC-Map for all 20 different one-dimensional relationships; the optimal scale to achieve $\hat{t}_*$ is marked with a green dot. For the monotonic dependencies (1-5), the optimal scale is always the largest scale, that is the global one. For all non-monotonic dependencies (6-19), MGC chooses smaller scales. Thus, a global optimal scale implies a close-to-linear dependency, otherwise the dependency is strongly nonlinear. In fact, this empirical observation led to the following theorem (which is proved in Materials and methods):

Theorem 1. *When (X, Y) are linearly related (meaning that Y can be constructed from X by rotation, scaling, translation, and/or reflection), the optimal scale of MGC equals the global scale. Conversely, a local optimal scale implies a nonlinear relationship.*

Thus, the MGC-Map explains how MGC discovers relationships, specifically, which scale has the most informative pairwise comparisons, and how that relates to the geometry of the relationship. Note that MGC provides the geometric characterization 'for free', meaning that no separate procedure is required; therefore, MGC provides both a valid test and information about the geometric relationship.

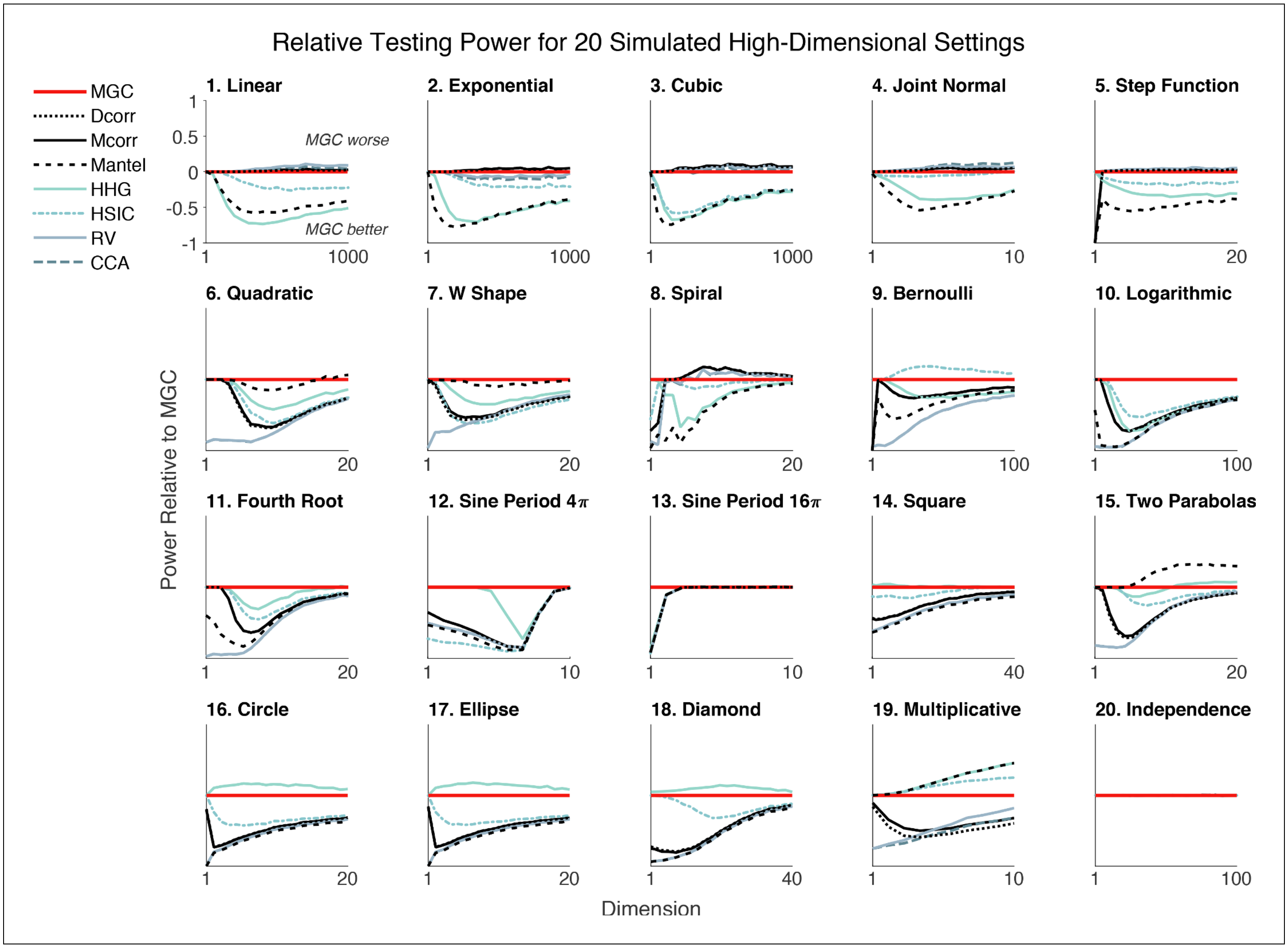


Figure 2. An extensive benchmark suite of 20 different relationships spanning polynomial, trigonometric, geometric, and other relationships demonstrates that Mgc empirically nearly dominates eight other methods across dependencies and dimensionalities ranging from 1 to 1000 (see Materials and methods and ***Figure 2—figure supplement 1*** for details). Each panel shows the testing power of other methods relative to the power of Mgc (e.g. power of Mcorr minus the power of Mgc) at significance level $\alpha = 0.05$ versus dimensionality for $n = 100$. Any line below zero at any point indicates that that method's power is less than Mgc's power for the specified setting and dimensionality. Mgc achieves empirically better (or similar) power than all other methods in almost all relationships and all dimensions. For the independent relationship (#20), all methods yield power 0.05 as they should. Note that Mgc is always plotted 'on top' of the other methods, therefore, some lines are obscured.

DOI: https://doi.org/10.7554/eLife.41690.004

The following figure supplements are available for figure 2:

Figure supplement 1. Visualization of the 20 dependencies at $p = q = 1$.

DOI: https://doi.org/10.7554/eLife.41690.005

Figure supplement 2. The same power plots as in ***Figure 2***, except the 20 dependencies are one-dimensional with noise, and the x-axis shows sample size increasing from 5 to 100.

DOI: https://doi.org/10.7554/eLife.41690.006

Figure supplement 3. The same set-ups as in ***Figure 2***, comparing different Mgc implementations versus its global counterparts.

DOI: https://doi.org/10.7554/eLife.41690.007

Figure supplement 4. The same power plots as in ***Figure 3***, except the 20 dependencies are one-dimensional with noise, and the x-axis shows sample size increasing from 5 to 100.

DOI: https://doi.org/10.7554/eLife.41690.008

Table 1. The median sample size for each method to achieve power 85% at type one error level 0.05, grouped into monotone (type 1–5) and non-monotone relationships (type 6–19) for both one- and ten-dimensional settings, normalized by the number of samples required by MGC.

In other words, a 2.0 indicates that the method requires double the sample size to achieve 85% power relative to MGC. PEARSON, RV, and CCA all achieve the same performance, as do SPEARMAN and KENDALL. MGC requires the fewest number of samples in all settings, and for high-dimensional non-monotonic relationships, all other methods require about double or triple the number of samples MGC requires.

Dimensionality	One-Dimensional			Ten-Dimensional		
Dependency type	Monotone	Non-Mono	Average	Monotone	Non-Mono	Average
MGC	**1**	**1**	**1**	**1**	**1**	**1**
DCORR	**1**	2.6	2.2	**1**	3.2	2.6
MCORR	**1**	2.8	2.4	**1**	3.1	2.6
HHG	1.4	**1**	1.1	1.7	1.9	1.8
HSIC	1.4	1.1	1.2	1.7	2.4	2.2
MANTEL	1.4	1.8	1.7	3	1.6	1.9
PEARSON / RV / CCA	**1**	>10	>10	**0.8**	>10	>10
SPEARMAN / KENDALL	**1**	>10	>10	n/a	n/a	n/a
MIC	2.4	2	2.1	n/a	n/a	n/a

DOI: https://doi.org/10.7554/eLife.41690.009

The following source data is available for Table 1:

Source data 1. Testing power sample size data in one dimension.

DOI: https://doi.org/10.7554/eLife.41690.010

Source data 2. Testing power sample size data in high-dimensions.

DOI: https://doi.org/10.7554/eLife.41690.011

Moreover, similar dependencies have similar MGC-Maps and often similar optimal scales. For example, logarithmic (10) and fourth root (11), although very different functions analytically, are geometrically similar, and yield very similar MGC-Maps. Similarly, (12) and (13) are trigonometric functions, and they share a narrow range of significant local scales. Both circle (16) and ellipse (17), as well as square (14) and diamond (18), are closely related geometrically and also have similar MGC-Maps. This indicates that the MGC-Map partially characterizes the geometry of these relationships, differentiating different dependence structures and assisting subsequent analysis steps. Moreover, in ***Shen and Vogelstein, 2018***, we proved that the sample MGC-Map (which MGC estimates) converges to the true MGC-Map provided by the underlying joint distribution of the data. In other words, each relationship has a specific map that characterizes it based on its joint distribution, and MGC is able to accurately estimate it via sample observations. The existence of a population level characterization of the joint distribution strongly differentiates MGC from previously proposed multiscale geometric or topological characterizations of data, such as persistence diagrams (***Edelsbrunner and Harer, 2009***).

MGC is computationally efficient

MGC does not incur large computational costs and has a similar complexity as existing methods. Though a naïve implementation of MGC requires $\mathcal{O}(n^4)$ operations, we devised a nested implementation that requires only $\mathcal{O}(n^2 \log n)$ operations. Moreover, obtaining the MGC-Map costs no additional computation, whereas other methods would require running a secondary computational step to decipher geometric properties of the relationship. MGC can also trivially be parallelized, reducing computation to $\mathcal{O}(n^2 \log n / T)$, where T is the number of cores (see Algorithm C1 for details). Since T is often larger than $\log n$, in practice, MGC can be $\mathcal{O}(n^2)$, meaning only a constant factor slower than DCORR and HSIC, which is illustrated in Figure 6 of ***Shen and Vogelstein, 2018***. For example, at sample size $n = 5000$ and dimension $p = 1$, on a typical laptop computer, DCORR requires around 0.5 s to compute the test statistic, whereas MGC requires no more than 5 s. But the cost and time to obtain 2.5× more data (so DCORR has same average power as MGC) typically far exceeds a few seconds. In

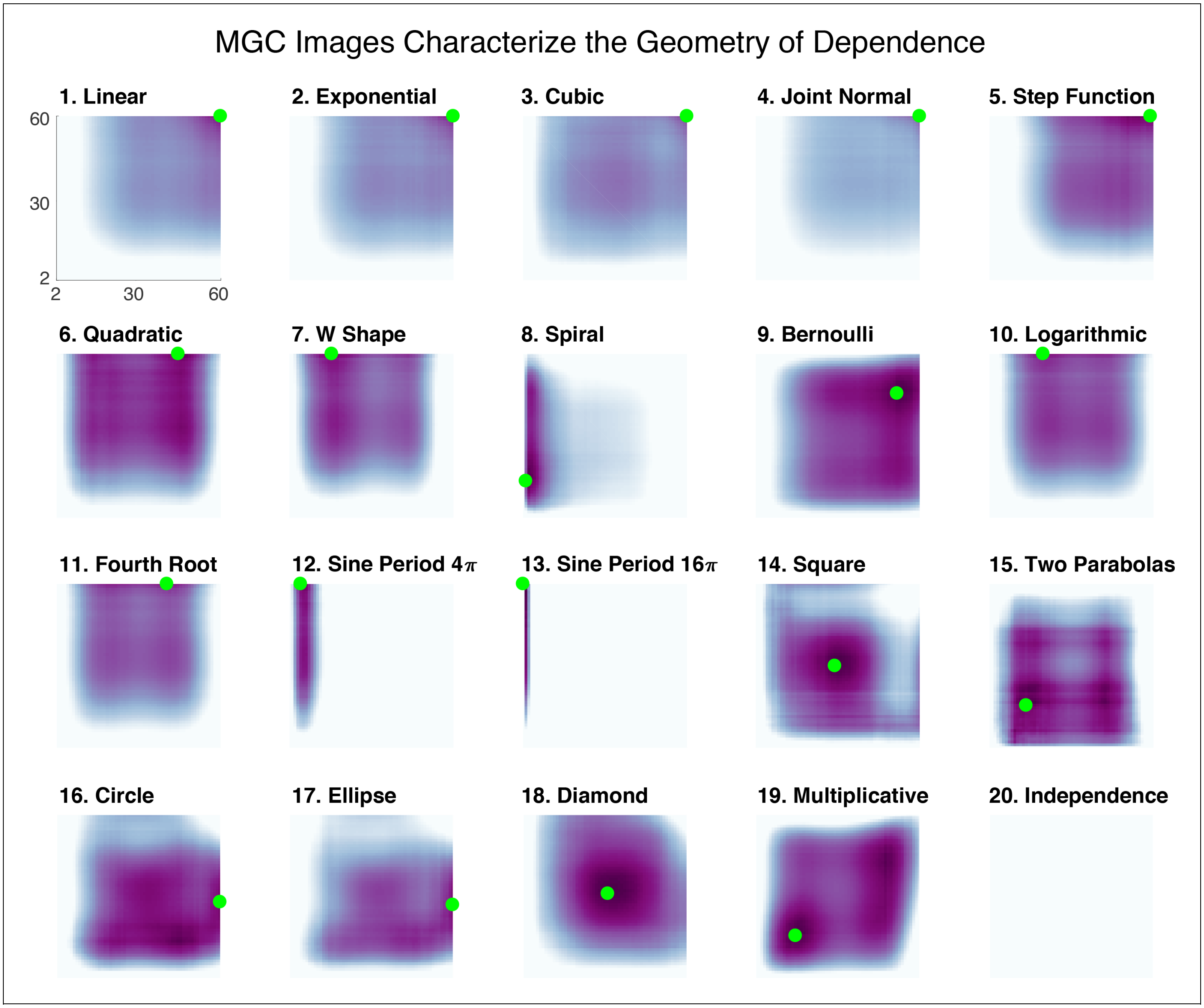


Figure 3. The MGC-Map characterizes the geometry of the dependence function. For each of the 20 panels, the abscissa and ordinate denote the number of neighbors for X and Y, respectively, and the color denotes the magnitude of each local correlation. For each simulation, the sample size is 60, and both X and Y are one-dimensional. Each dependency has a different MGC-Map characterizing the geometry of dependence, and the optimal scale is shown in green. In linear or close-to-linear relationships (first row), the optimal scale is global, that is the green dot is in the top right corner. Otherwise the optimal scale is non-global, which holds for the remaining dependencies. Moreover, similar dependencies often share similar MGC-Maps and similar optimal scales, such as (10) logarithmic and (11) fourth root, the trigonometric functions in (12) and (13 , 16) circle and (17) ellipse, and (14) square and (18) diamond. The MGC-Maps for high-dimensional simulations are provided in ***Figure 3—figure supplement 1***.

DOI: https://doi.org/10.7554/eLife.41690.012

The following figure supplement is available for figure 3:

Figure supplement 1. The MGC-Map for the 20 panels for high-dimensional dependencies.

DOI: https://doi.org/10.7554/eLife.41690.013

comparison, the cost to compute a persistence diagram is typically $\mathcal{O}(n^3)$, which is orders of magnitude slower when $n > 10$. The running time of each method on the real data experiments are reported in Materials and methods.

Mgc uniquely reveals relationships in real data

Geometric intuition, numerical simulations, and theory all provide evidence that Mgc will be useful for real data discoveries. Nonetheless, real data applications provide another necessary ingredient to justify its use in practice. Below, we describe several real data applications where we have used Mgc to understand relationships in data that other methods were unable to provide.

Mgc discovers the relationships between brain and mental properties

The human psyche is of course dependent on brain activity and structure. Previous work has studied two particular aspects of our psyche: personality and creativity, developing quantitative metrics for evaluating them using structured interviews (***Costa and McCrae, 1992***; ***Jung et al., 2009***). However, the relationship between brain activity and structure, and these aspects of our psyche, remains unclear (***DeYoung et al., 2010***; ***Xu and Potenza, 2012***; ***Bjørnebekk et al., 2013***; ***Sampaio et al., 2014***). For example, prior work did not evaluate the relationship between entire brain connectivity and all five factors of the standard personality model (***Costa and McCrae, 1992***). We therefore utilized Mgc to investigate published open access data (see Materials and methods for details).

First, we analyzed the relationship between resting-state functional magnetic resonance (rs-fMRI) activity and personality (***Adelstein et al., 2011***). The first row of ***Table 2*** compares the p-value of different methods, and ***Figure 4A*** shows the Mgc-Map for the sample data. Mgc is able to yield a significant p-value (< 0.05), whereas all previously proposed global dependence tests under consideration (Mantel, Dcorr, Mcorr, or Hhg) fail to detect dependence at a significance level of 0.05. Moreover, the Mgc-Map provides a characterization of the dependence, for which the optimal scale indicates that the dependency is strongly nonlinear. Interestingly, the Mgc-Map does not look like any of the 20 images from the simulated data, suggesting that the nonlinearity characterizing this dependency is more complex or otherwise different from those we have considered so far.

Second, we investigated the relationship between diffusion MRI derived connectivity and creativity (***Jung et al., 2009***). The second row of ***Table 2*** shows that Mgc is able to ascertain a dependency between the whole brain network and the subject's creativity. The Mgc-Map in ***Figure 4B*** closely resembles a linear relationship where the optimal scale is global. The close-to-linear relationship is also supported from the p-value table as all methods except Hsic are able to detect significant dependency, which suggests that there is relatively little to gain by pursuing nonlinear regression techniques, potentially saving valuable research time by avoiding tackling an unnecessary problem. The test statistic for both Mgc and Mcorr equal 0.04, which is quite close to zero despite a significant p-value, implying a relatively weak and noisy relationship. A prediction of creativity via linear regression turns out to be non-significant, which implies that the sample size is too low to obtain useful predictive accuracy (not shown), indicating that more data are required for single subject predictions. If one had first directly estimated the regression function, obtaining a null result, it would remain unclear whether a relationship existed. This experiment demonstrates that for high-dimensional and potentially structured data, Mgc is able to reveal dependency with relatively small sample size while parametric techniques and directly estimating regression functions can often be ineffective.

Table 2. The p-values for brain imaging vs mental properties.

Mgc *always* uncovers the existence of significant relationships and discovers the underlying optimal scales. Bold indicates significant p-value per dataset.

Testing Pairs/Methods	Mgc	Dcorr	Mcorr	Hhg	Hsic
Activity vs Personality	**0.043**	0.667	0.441	0.059	0.124
Connectivity vs Creativity	**0.011**	**0.010**	**0.011**	**0.031**	0.092

DOI: https://doi.org/10.7554/eLife.41690.015

The following source data is available for Table 2:

Source data 1. p-value data for activity vs personality.

DOI: https://doi.org/10.7554/eLife.41690.016

Source data 2. p-value data for connetivity vs creativity.

DOI: https://doi.org/10.7554/eLife.41690.017

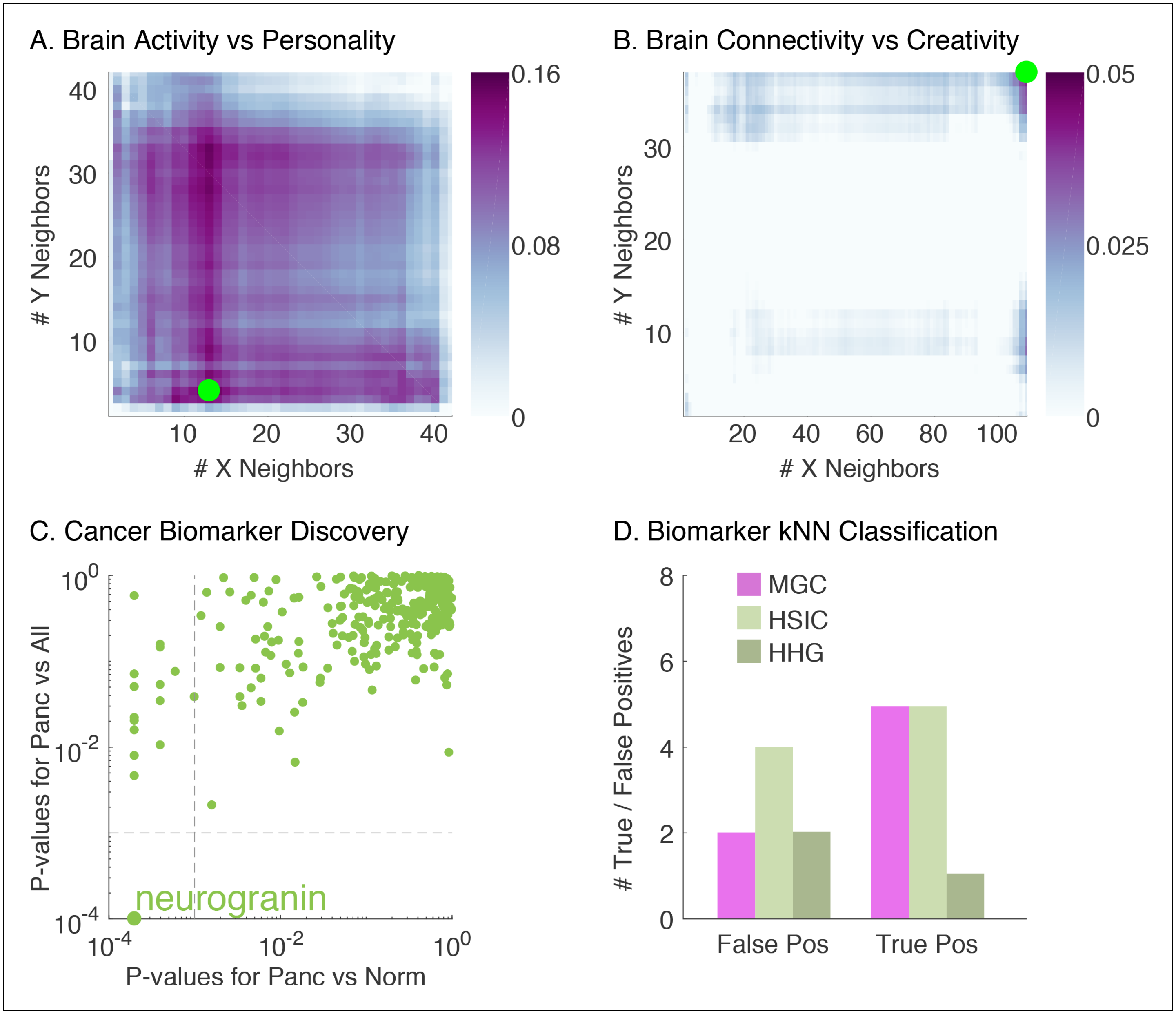


Figure 4. Demonstration that MGC successfully detects dependency, distinguishes linearity from nonlinearity, and identifies the most informative feature in a variety of real data experiments. (**A**) The MGC-Map for brain activity versus personality. MGC has a large test statistic and a significant p-value at the optimal scale (13, 4), while the global counterpart is non-significant. That the optimal scale is non-global implies a strongly nonlinear relationship. (**B**) The MGC-Map for brain connectivity versus creativity. The image is similar to that of a linear relationship, and the optimal scale equals the global scale, thus both MGC and MCORR are significant in this case. (**C**) For each peptide, the x-axis shows the p-value for testing dependence between pancreatic and healthy subjects by MGC, and the y-axis shows the p-value for testing dependence between pancreatic and all other subjects by MGC. At critical level 0.05, MGC identifies a unique protein after multiple testing adjustment. (**D**) The true and false positive counts using a k-nearest neighbor (choosing the best $k \in [1, 10]$) leave-one-out classification using only the significant peptides identified by each testing method. The peptide identified by MGC achieves the best true and false positive rates, as compared to the peptides identified by HSIC or HHG.

DOI: https://doi.org/10.7554/eLife.41690.014

The performance in the real data closely matches the simulations results: the first dataset exhibits a strongly nonlinear relationship, for which MGC has the lowest p-value, followed by HHG and HSIC and then all other methods; the second dataset exhibits a close-to-linear relationship, for which global methods perform the best while HHG and HSIC are trailing. Moreover, MGC detected a complex nonlinear relationship for brain activity versus personality, and a nearly linear but noisy relationship for brain network versus creativity, the only method able to make either of those claims. In a separate experiment, we assessed the frequency with which MGC obtained false positive results using brain activity data, based on experiments from **Eklund et al. (2012)**; **Eklund et al. (2016)**. **Appendix 1—figure 1** shows that MGC achieves a false positive rate of 5% when using a significance level of 0.05, implying that it correctly controls for false positives, unlike typical parametric methods on these data.

MGC identifies potential cancer proteomics biomarkers

MGC can also be useful for a completely complementary set of scientific questions: screening proteomics data for biomarkers, often involving the analysis of tens of thousands of proteins, peptides, or transcripts in multiple samples representing a variety of disease types. Determining whether there is a relationship between one or more of these markers and a particular disease state can be challenging but is a necessary first step (**Frantzi et al., 2014**). We sought to discover new useful protein biomarkers from a quantitative proteomics technique that measures protein and peptide abundance called Selected Reaction Monitoring (SRM) (**Wang et al., 2011**). Specifically, we were interested in finding biomarkers that were unique to pancreatic cancer, because it is lethal and no clinically useful biomarkers are currently available (**Bhat et al., 2012**).

The data consist of proteolytic peptides derived from the blood samples of 95 individuals harboring pancreatic ($n = 10$), ovarian ($n = 24$), colorectal cancer ($n = 28$), and healthy controls ($n = 33$). The processed data included 318 peptides derived from 121 proteins. Previously, we used these data and other techniques to find ovarian cancer biomarkers (a much easier task because the dataset has twice as many ovarian patients) and validated them with subsequent experiments (**Wang et al., 2017**). Therefore, our first step was to check whether MGC could correctly identify ovarian biomarkers. Indeed, the peptides that have been validated previously are also identified by MGC. Emboldened, using the same dataset, we applied MGC to screen for biomarkers unique to pancreatic cancer. To do so, we first screened for a difference between pancreatic cancer and healthy controls, identifying several potential biomarkers. Then, we screened for a difference between pancreatic cancer and all other conditions, to find peptides that differentiate pancreatic cancer from other cancers. **Figure 4C** shows the p-value of each peptide assigned by MGC, which reveals one particular protein, neurogranin, that exhibits a strong dependency specifically with pancreatic cancer. Subsequent literature searches reveal that neurogranin is a potentially valuable biomarker for pancreatic cancer because it is exclusively expressed in brain tissue among normal tissues and has not been linked with any other cancer type (**Yang et al., 2015**; **Willemse et al., 2018**). In comparison, HSIC identified neurogranin as well, but it also identified another peptide; HHG identified the same two by HSIC, and a third peptide. A literature evaluation of these additional peptides shows that they are upregulated in other cancers as well and are unlikely to be useful as a pancreatic biomarker (**Helfman et al., 2018**; **Lam et al., 2012**). The rest of the global methods did not identify any markers at significance level 0.05, see Materials and methods for more details and **Appendix 1—table 2** for identified peptide information using each method.

Since there is no ground truth yet in this experiment, we further carried out a classification task using the biomarkers identified by the various algorithms, using a k-nearest-neighbor classifier to predict pancreatic cancer, and a leave-one-subject-out validation. **Figure 4D** shows that the peptide selected by MGC (neurogranin) works better than any other subset of the peptides selected by HSIC or HHG, in terms of both fewer false positives and negatives. This analysis suggests MGC can effectively be used for screening and subsequent classification.

Discussion

There are a number of connections between MGC and other prominent statistical procedures that may be worth further exploration. First, MGC can be thought of as a regularized or sparsified variant of distance or kernel methods. Regularization is central to high-dimensional and ill-posed problems, where dimensionality is larger than sample size. Second, MGC can also be thought of as learning a metric because it chooses the optimal scale amongst a set of n^2 truncated distances, motivating studying the relationship between MGC and recent advances in metric learning (**Xing et al., 2003**). In particular, deep learning can be thought of as metric learning (**Giryes et al., 2015**), and generative adversarial networks (**Goodfellow et al., 2014**) are implicitly testing for equality, which is closely related to dependence (***Sutherland et al., 2016***). While MGC searches over a two-dimensional parameter space to optimize the metric, deep learning searches over a much larger parameter space, sometimes including millions of dimensions. Probably neither is optimal, and somewhere between the two would be useful in many tasks. Third, energy statistics provide state of the art approaches to other problems, including goodness-of-fit (***Székely and Rizzo, 2005***), analysis of variance (***Rizzo and Székely, 2010***), conditional dependence (***Székely and Rizzo, 2014***; ***Wang et al., 2015***), and feature selection (***Li et al., 2012***; ***Zhong and Zhu, 2015***), so MGC can be adapted for them as well. Indeed, MGC can also implement a two-sample (or generally the K-sample) test (***Szekely and Rizzo, 2004***; ***Heller et al., 2016***; ***Shen and Vogelstein, 2018***). Specifically, for more than two modalities, one may use summation of pairwise MGC test statistics, similar to how energy statistic is generalized to K-sample testing from two-sample testing (***Rizzo and Székely, 2010***; ***Rizzo and Székely, 2016***; ***Shen and Vogelstein, 2018***), or how canonical correlation analysis is generalized into more than two modalities (***Kettenring, 1971***; ***Tenenhaus and Tenenhaus, 2011***; ***Shen et al., 2014***). Finally, although energy statistics have not yet been explicitly used for classification, regression, or dimensionality reduction, MGC opens the door to these applications by providing guidance as to how to proceed. Specifically, it is well documented in machine learning literature that the choice of kernel, metric, or scale often has a strong effect on the performance of different machine learning algorithms (***Levina and Bickel, 2004***). MGC provides a mechanism to estimate scale that is both theoretically justified and computationally efficient, by optimizing a metric for a task wherein the previous methods lacked a notion of optimization. Nonlinear dimensionality reduction procedures, such as Isomap (***Tenenbaum et al., 2000***) and local linear embedding (***Roweis and Saul, 2000***) for example, must also choose a scale, but have no principled criteria for doing so. MGC could be used to provide insight into multimodal dimensionality reduction as well.

The default metric choice of MGC in this paper is always the Euclidean distance, but other metric choices may be more appropriate in different fields, and using the strong negative type metric as specified in ***Lyons (2013)*** guarantees consistency. However, if multiple metric choices are experimented to yield multiple MGC p-values, then the optimal p-value should be properly corrected for multiple testing. Alternatively, one may use the maximum MGC statistic among multiple metric choices, apply the same procedure in each permutation (i.e. in each permutation, use the same number of metric choices and take the maximum MGC as the permuted statistic), then derive a single p-value. Such a testing procedure properly controls the type one error level without the need for additional correction.

MGC also addresses a particularly vexing statistical problem that arises from the fact that methods methods for discovering dependencies are typically dissociated from methods for deciphering them. This dissociation creates a problem because the statistical assumptions underlying the 'deciphering' methods become compromised in the process of 'discoverying'; this is called the 'post-selection inference' problem (***Berk et al., 2013***). The most straightforward way to address this issue is to collect new data, which is costly and time-consuming. Therefore, researchers often ignore this fact and make statistically invalid claims. MGC circumvents this dilemma by carefully constructing its permutation test to estimate the scale in the process of estimating a p-value, rather than after. To our knowledge, MGC is the first dependence test to take a step towards valid post-selection inference.

As a separate next theoretical extension, we could reduce the computational space and time required by MGC. MGC currently requires space and time that are quadratic with respect to the number of samples, which can be costly for very large data. Recent advances in related work demonstrated that one could reduce computational time of distance-based tests to close to linear via faster implementation, subsampling, random projection, and null distribution approximation (***Huo and Székely, 2016***; ***Huang and Huo, 2017***; ***Zhang et al., 2018***; ***Chaudhuri and Hu, 2018***), making it feasible for large amount of data. Alternately, semi-external memory implementations would allow running MGC even as the interpoint comparison matrix exceeds the size of main memory (***Da Zheng et al., 2015***; ***Da Zheng et al., 2016a***; ***Da Zheng et al., 2016b***; ***Da Zheng et al., 2016c***).

Finally, MGC is easy to use. Source code is available in MATLAB, R, and Python from https://mgc.neurodata.io/ (***Bridgeford et al., 2018***; experiments archived at https://github.com/elifesciences-publications/MGC-paper). Code for reproducing all the figures in this manuscript is also available from the above websites. We showed MGC's value in diverse applications spanning neuroscience (which motivated this work) and an 'omics example. Applications in other domains facing similar questions of dependence, such as finance, pharmaceuticals, commerce, and security, could likewise benefit from Mgc.

Materials and methods

Mathematical details

This section contains essential mathematical details on independence testing, the notion of the generalized correlation coefficient and the distance-based correlation measure, how to compute the local correlations, and the smoothing technique. A statistical treatment on MGC is in ***Shen and Vogelstein, 2018***, which introduces the population version of Mgc and various theoretical properties.

Testing independence

Given pairs of observations $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{R}^p \times \mathbb{R}^q$ for $i = 1, \ldots, n$, assume they are independently identically distributed as $(X, Y) \overset{iid}{\sim} F_{XY}$. If the two random variables X and Y are independent, the joint distribution equals the product of the marginals, that is $F_{XY} = F_X F_Y$. The statistical hypotheses for testing independence is as follows:

$$H_0 : F_{XY} = F_X F_Y,$$

$$H_A : F_{XY} \neq F_X F_Y.$$

Given a test statistic, the testing power equals the probability of rejecting the independence hypothesis (i.e. the null hypothesis) when it is false. A test statistic is consistent if and only if the testing power increases to 1 as sample size increases to infinity. We would like a test to be universally consistent, that is consistent against all joint distributions. Dcorr, Mcorr, Hsic, and Hhg are all consistent against any joint distribution of finite second moments and finite dimension.

Note that p is the dimension for $\boldsymbol{x}$'s, q is the dimension for $\boldsymbol{y}$'s. For Mgc and all benchmark methods, there is no restriction on the dimensions, that is the dimensions can be arbitrarily large, and p is not required to equal q. The ability to handle data of arbitrary dimension is crucial for modern big data. There also exist some special methods that only operate on one-dimensional data, such as (***Reshef et al., 2011***; ***Heller et al., 2016***; ***Huo and Székely, 2016***), which are not directly applicable to multidimensional data.

Correlation measures

To achieve consistent testing, most state-of-the-art dependence measures operate on pairwise comparisons, either similarities (such as kernels) or dissimilarities (such as distances).

Let $\mathcal{X}_n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \in \mathbb{R}^{p \times n}$ and $\mathcal{Y}_n = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\} \in \mathbb{R}^{q \times n}$ denote the matrices of sample observations, and δ_x be the distance function for $\boldsymbol{x}$'s and δ_y for $\boldsymbol{y}$'s. One can then compute two $n \times n$ distance matrices $\tilde{A} = \{\tilde{a}_{ij}\}$ and $\tilde{B} = \{\tilde{b}_{ij}\}$, where $\tilde{a}_{ij} = \delta_x(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\tilde{b}_{ij} = \delta_y(\boldsymbol{y}_i, \boldsymbol{y}_j)$. A common example of the distance function is the Euclidean metric (L^2 norm), which serves as the starting point for all methods in this manuscript.

Let A and B be the transformed (e.g., centered) versions of the distance matrices $\tilde{A}$ and $\tilde{B}$, respectively. Any 'generalized correlation coefficient' (***Spearman, 1904***; ***Kendall, 1970***) can be written as:

$$c(\mathcal{X}_n, \mathcal{Y}_n) = \frac{1}{z} \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ij}, \tag{1}$$

where z is proportional to the standard deviations of A and B, that is $z = n^2 \sigma_a \sigma_b$. In words, c is the global sample correlation across *pairwise comparison matrices* A and B, and is normalized into the range $[-1, 1]$, which usually has expectation 0 under independence and implies a stronger dependency when the correlation is further away from 0.

Traditional correlations such as the Pearson's correlation and the rank correlation can be written via the above correlation formulation, by using A and B directly from sample observations rather than distances. Distance-based methods like Dcorr and Mantel operate on the Euclidean distance by default, or other metric choices on the basis of domain knowledge; then transform the resulting distance matrices $\tilde{A}$ and $\tilde{B}$ by certain centering schemes into A and B. Hsic chooses the Gaussian kernel and computes two kernel matrices, then transform the kernel matrices $\tilde{A}$ and $\tilde{B}$ by the same centering scheme as Dcorr. For Mgc, A and B are always distance matrices (or can be transformed to distances from kernels by ***Sejdinovic et al. (2013)***), and we shall apply a slightly different centering scheme that turns out to equal Dcorr.

To carry out the hypothesis testing on sample data via a nonparametric test statistic, for example a generalized correlation, the permutation test is often an effective choice (***Good, 2005***), because a p-value can be computed by comparing the correlation of the sample data to the correlation of the permuted sample data. The independence hypothesis is rejected if the p-value is lower than a pre-determined type 1 error level, say 0.05. Then the power of the test statistic equals the probability of a correct rejection at a specific type 1 error level. Note that Hhg is the only exception that cannot be cast as a generalized correlation coefficient, but the permutation testing is similarly effective for the Hhg test statistic; also note that the *iid* assumption is critical for permutation test to be valid, which may not be applicable in special cases like auto-correlated time series (***Guillot and Rousset, 2013***).

Distance correlation (Dcorr) and the Unbiased Version (Mcorr)

Define the row and column means of $\tilde{A}$ by $\bar{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} \tilde{a}_{ij}$ and $\bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} \tilde{a}_{ij}$. Dcorr defines

$$a_{ij} = \begin{cases} \tilde{a}_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}, & \text{if } i \neq j, \\ 0, & \text{if } i = j, \end{cases}$$

and similarly for b_{ij}. For distance correlation, the numerator of ***Equation 1*** is named the distance covariance (Dcov), while σ_a and σ_b in the denominator are the square root of each distance variance. The centering scheme is important to guarantee the universal consistency of Dcorr, whereas Mantel uses a simple centering scheme and thus not universally consistent.

Let $c(X, Y)$ be the population distance correlation, that is, the distance correlation between the underlying random variables X and Y. ***Székely et al. (2007)*** define the population distance correlation via the characteristic functions of F_X and F_Y, and show that the population distance correlation equals zero if and only if X and Y are independent, for any joint distribution F_{XY} of finite second moments and finite dimensionality. They also show that as $n \to \infty$, the sample distance correlation converges to the population distance correlation, that is, $c(\mathcal{X}_n, \mathcal{Y}_n) \to c(X, Y)$. Thus the sample distance correlation is consistent against any dependency of finite second moments and dimensionality. Of note, the distance covariance, distance variance, and distance correlation are always non-negative. Moreover, the consistency result holds for a much larger family of metrics, those of strong negative type (***Lyons, 2013***).

It turns out that the sample distance correlation has a finite-sample bias, especially as the dimension p or q increases (***Szekely and Rizzo, 2013***). For example, for independent Gaussian distributions, the sample distance correlation converges to 1 as $p, q \to \infty$. By excluding the diagonal entries and slightly modifies the off-diagonal entries of $\mathcal{A}$ and $\mathcal{B}$, Szekely and Rizzo (***Szekely and Rizzo, 2013***; ***Székely and Rizzo, 2014***) show that Mcorr is an unbiased estimator of the population distance correlation $c(\boldsymbol{x}, \boldsymbol{y})$ for all p, q, n, which is approximately normal even if $p, q \to \infty$. Thus it enjoys the same theoretical consistency as Dcorr and always has zero mean under independence.

Local correlations

Given any matrices A and B, we can define a set of local correlations as follows. Let $R(A_{\cdot j}, i)$ be the 'rank' of $\boldsymbol{x}_i$ relative to $\boldsymbol{x}_j$, that is, $R(A_{\cdot j}, i) = k$ if $\boldsymbol{x}_i$ is the k^{th} closest point (or 'neighbor') to $\boldsymbol{x}_j$, as determined by ranking the $n-1$ distances to x_j. Define $R(B_{i\cdot}, j)$ equivalently for the Y's, but ranking relative to the rows rather than the columns (see below for explanation). For any neighborhood size k around each $\boldsymbol{x}_i$ and any neighborhood size l around each $\boldsymbol{y}_j$, we define the local pairwise comparisons:

$$\widetilde{a}_{ij}^{k} = \begin{cases} a_{ij}, & \text{if } R(A_{\cdot j}, i) \leq k, \\ 0, & \text{otherwise;} \end{cases} \qquad \widetilde{b}_{ij}^{l} = \begin{cases} b_{ij}, & \textit{if } R(B_{i\cdot}, j) \leq l, \\ 0, & \text{otherwise;} \end{cases} \tag{2}$$

and then let $a_{ij}^{k} = \widetilde{a}_{ij}^{k} - \bar{a}^{k}$, where $\bar{a}^{k}$ is the mean of $\{\widetilde{a}_{ij}^{k}\}$, and similarly for b_{ij}^{l}.

The *local* correlation coefficient at a given scale is defined to effectively exclude large distances:

$$c^{kl}(\mathcal{X}_n,\mathcal{Y}_n) = \frac{1}{z_{kl}} \sum\nolimits_{i,j=1}^{n} a_{ij}^{k} b_{ij}^{l}, \tag{3}$$

where $z_{kl} = n^2 \sigma_a^k \sigma_b^l$, with σ_a^k and σ_b^l is the standard deviations for the truncated pairwise comparisons. The Mgc-Map can be constructed by computing all local correlations, which allows the discovery of the optimal correlation. For any aforementioned correlation (Dcorr, Mcorr, Hsic, Mantel, Pearson), one can define its local correlations by using ***Equation 3*** and plugging in the respective a_{ij} and b_{ij} from ***Equation 1***.

As most nonlinear relationships intrinsically exhibit a local linear structure, considering the nearest-neighbors is able to amplify the dependency signal over the global correlation. There could be two other scenarios: when the small distances in one modality mostly correspond to large distances in another modality, or when the large distances in one modality correspond to large distance in another modality. For the first scenario, the small distances become negative terms after centering while the large distances become positive terms after centering, so adding their product to c^{kl} will cause the test statistic to be smaller — in fact, as distance correlation is shown to be > 0 under dependence (***Székely et al., 2007***), the first scenario cannot happen for all distances pairs. For the second scenario, one can experiment using the large distances (or the furthest neighbors) only by reversing the ranking scheme in local correlation to descending order. However, whenever the large distances are highly correlated, the small distances must also be highly correlated after centering by the mean distances, so global correlation coefficient like Dcorr already handles this scenario. Therefore considering the nearest-neighbor may significantly improve the performance over global correlation, while considering the other scenarios does not.

Mgc as the optimal local correlation

We define the multiscale graph correlation statistic as the optimal local correlation, for which the family of local correlation is computed based on Euclidean distance and Mcorr transformation.

Instead of taking a direct maximum, Mgc takes a smoothed maximum, that is the maximum local correlation of the largest connected component R such that all local correlations within R are significant. If no such region exists, Mgc defaults the test statistic to the global correlation (details in Algorithm C2). Thus, we can write:

$$c^{*}(\mathcal{X}_n,\mathcal{Y}_n) = \max_{(k,l)\in R} c^{kl}(\mathcal{X}_n,\mathcal{Y}_n)$$

$$R = \text{Largest Connected Component of } \{(k,l) \text{ such that } c^{kl} > \max(\tau, c^{nn})\}.$$

Then the optimal scale equals all scales within R whose local correlations are as large as c^{*}. The choice of τ is made explicit in the pseudo-code, with further discussion and justification offered in ***Shen and Vogelstein, 2018***.

Proof for theorem 1

Theorem 1. When (X,Y) are linearly related (rotation, scaling, translation, reflection), the optimal scale of Mgc equals the global scale. Conversely, that. the optimal scale is local implies a nonlinear relationship.

Proof. It suffices to prove the first statement, then the second statement follows by contrapositive. When (X,Y) are linearly related, $Y = WX + b$ for a unitary matrix W and a constant b up-to possible scaling, in which case the distances are preserved, that is $\|y_i - y_j\| = \|Wx_i - Wx_j\| = \|x_i - x_j\|$. It follows that $\texttt{Mcorr}(\mathcal{X}_n,\mathcal{Y}_n) = 1$, so the global scale achieves the maximum possible correlation, and the largest connected region R is empty. Thus the optimal scale is global and $\texttt{Mgc}(\mathcal{X}_n,\mathcal{Y}_n) = \texttt{Mcorr}(\mathcal{X}_n,\mathcal{Y}_n) = 1$.

Computational complexity of each step

The distance computation takes $\mathcal{O}(n^2 \max\{p,q\})$, and the ranking process takes $\mathcal{O}(n^2 \log n)$. Once the distance and ranking are completed, computing one local generalized correlation requires $\mathcal{O}(n^2)$

(see Algorithm C4). Thus, a naive approach to compute all local generalized correlations requires at least $\mathcal{O}(n^2 \max\{n^2, p, q\})$ by going through all possible scales, meaning possibly $\mathcal{O}(n^4)$ which would be computationally prohibitive. However, given the distance and ranking information, we devised an algorithm that iteratively computes all local correlations in $\mathcal{O}(n^2)$ by re-using adjacent smaller local generalized correlations (see Algorithm C5). Therefore, when including the distance computation and ranking overheads, the MGC statistic is computed in $\mathcal{O}(n^2 \max\{\log n, p, q\})$, which has the same running time as the HHG statistic, and the same running time up to a factor of $\log n$ as global correlations like DCORR and MCORR, which require $\mathcal{O}(n^2 \max\{p, q\})$ time. By utilizing a multi-core architecture, MGC can be computed in $\mathcal{O}(n^2 \max\{\log n, p, q\}/T)$ instead. As $T = \log(n)$ is often a small number, for example T is no more than 30 at 1 billion samples, thus MGC can be effectively computed in the same complexity as DCORR. Note that the permutation test adds another r random permutations to the n^2 term, so computing the p-value requires $\mathcal{O}(n^2 \max\{\log n, p, q, r\}/T)$.

MGC algorithms and testing procedures

Six algorithms are presented in order:

- Algorithm C1 describes MGC in its entirety (which calls most of the other algorithms as functions).
- Algorithm C2 computes the MGC test statistic.
- Algorithm C3 computes the p-value of MGC by the permutation test.
- Algorithm C4 computes the local generalized correlation coefficient at a given scale (k, l), for a given choice of the global correlation coefficient.
- Algorithm C5 efficiently computes all local generalized correlations, in nearly the same running time complexity as computing one local generalized correlation.
- Algorithm C6 evaluates the testing power of MGC by a given distribution.

For ease of presentation, we assume there are no repeating observations of X or Y, and note that MCORR is the global correlation choice that MGC builds on.

Pseudocode C1 Multiscale Graph Correlation (MGC); requires $\mathcal{O}(n^2 \max(\log n, p, q, r)/T)$ time, where r is the number of permutations and T is the number of cores available for parallelization.

Input: n samples of (x_i, y_i) pairs, an integer r for the number of random permutations.

Output: (i) MGC statistic c^*, (ii) the optimal scale (k, l), (iii) the p-value $p(c^*)$,

```
function MG($(x_i, y_i)$), for $i \in [n]$)
(1) Calculate all pairwise distances:
    for $i, j := 1, \ldots, n$ do
        $a_{ij} = \delta_x(x_i, x_j)$                    $\delta_x$ is the distance between pairs of $x$ samples
        $b_{ij} = \delta_y(y_i, y_j)$                    $\delta_y$ is the distance between pairs of $y$ samples
    end for
    Let $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$.
(2) Calculate Multiscale Correlation Map $\mathcal{C}$ & MGC Test Statistic:
    $[c^*, \mathcal{C}, k, l] = \text{MGCSampleStat}(A, B)$                    Algorithm C2
(3) Calculate the p-value
    $pval(c^*) = \text{PermutationTest}(A, B, r, c^*)$                    Algorithm C3
end Function
```

Pseudocode C2 MGC test statistic. This algorithm computes all local correlations, take the smoothed maximum, and reports the (k, l) pair that achieves it. For the smoothing step, it: (i) finds the largest connected region in the correlation map, such that each correlation is significant, that is larger than a certain threshold to avoid correlation inflation by sample noise, (ii) take the largest correlation in the region, (iii) if the region area is too small, or the smoothed maximum is no larger than the global correlation, the global correlation is used instead. The running time is $\mathcal{O}(n^2)$.

Input: A pair of distance matrices $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$.

Output: The MGC statistic $c^* \in \mathbb{R}$, all local statistics $\mathcal{C} \in \mathbb{R}^{n \times n}$, and the corresponding local scale $(k, l) \in \mathbb{N} \times \mathbb{N}$.

Continued on next page

```
function MGCSampleStat($(A, B)$)
    $\mathcal{C} = \mathrm{MGCAllLocal}(A, B)$                    All local correlations
    $\tau = \mathrm{Thresholding}(\mathcal{C})$                   find a threshold to determine large local correlations
    for $i,j := 1,\ldots,n$do $r_{ij} \leftarrow \mathbb{I}(c^{ij}>\tau)$end for   identify all scales with large correlation
    $\mathcal{R} \leftarrow \{r_{ij} : i,j = 1,\ldots,n\}$        binary map encoding scales with large correlation
    $\mathcal{R} = \mathrm{Connected}(\mathcal{R})$               largest connected component of the binary matrix
    $c^* \leftarrow \mathcal{C}(n,n)$                             use the global correlation by default
    $k \leftarrow n, l \leftarrow n$
    if $\left(\sum_{i,j} r_{ij}\right) \geq 2n$ then              proceed when the significant region is sufficiently large
        $[c^*, k, l] \leftarrow \max(\mathcal{C} \circ \mathcal{R})$   find the smoothed maximum and the respective scale
    end if
end Function
```

Input: $\mathcal{C} \in \mathbb{R}^{n \times n}$.

Output: A threshold τ to identify large correlations.

```
function Thresholding $\mathcal{C}$
    $\tau \leftarrow \sum_{c^{ij}<0}(c^{ij})^2 / \sum_{c^{ij}<0} 1$   variance of all negative local generalized correlations
    $\tau \leftarrow \max\{0.01, \sqrt{\tau}\} \times 3.5$            threshold based on negative correlations
    $\tau \leftarrow \max\{\tau, 2/n, c^{nn}\}$
end Function
```

Pseudocode C3 Permutation Test. This algorithm uses the random permutation test with r random permutations for the p-value, requiring $\mathcal{O}(rn^2 \log n)$ for Mgc. In the real-data experiment, we always set $r = 10{,}000$. Note that the p-value computation for any other global generalized correlation coefficient follows from the same algorithm by replacing Mgc with the respective test statistic.

Input: A pair of distance matrices $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$, the number of permutations r, and Mgc statistic c^* for the observed data.

Output: The p-value $pval \in [0, 1]$.

```
function PermutationTest($A$, $B$, $r$, $c^*$)
    for $t := 1,\ldots,r$ do
        $\pi = \mathrm{RandPerm}(n)$                              generate a random permutation of size $n$
        $c_0^*[t] = \mathrm{MGCSampleStat}(A, B(\pi, \pi))$       calculate the permuted Mgc statistic
    end for
    $pval(c^*) \leftarrow \frac{1}{t}\sum_{t=1}^{r} \boldsymbol{I}(c^* \leq c_0^*[t])$   compute p-value of Mgc
end function
```

Pseudocode C4 Compute local test statistic at a given scale. This algorithm runs in $\mathcal{O}(n^2)$ once the rank information is provided, which is suitable for Mgc computation if an optimal scale is already estimated. But it would take $\mathcal{O}(n^4)$ if used to compute all local generalized correlations. Note that for the default Mgc implementation uses single centering, the centering function centers A by column and B by row, and the sorting function sorts A within column and B within row. By utilizing $T = \log(n)$ cores, the sorting function can be easily parallelized to take $\mathcal{O}(n^2 \log(n)/T) = \mathcal{O}(n^2)$.

Input: A pair of distance matrices $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$, and a local scale $(k, l) \in \mathbb{N} \times \mathbb{N}$.

Output: The local generalized correlation coefficient $c^{kl} \in [-1, 1]$.

```
function LocalGenCorr($A$, $B$, $k$, $l$)
    for $Z := A, B$ do $\mathcal{E}^Z = \mathrm{Sort}(Z)$ end for    parallelized sorting
    for $Z := A, B$ do $Z = \mathrm{Center}(Z)$ end for              center distance matrices
    $\tilde{c}^{kl} \leftarrow tr((A \circ \mathcal{E}^A)^{\mathrm{T}} \times (B \circ (\mathcal{E}^B)^{\mathrm{T}}))$   un-normalized local distance covariance
```

Continued on next page

```
        $v^A \leftarrow tr((A \circ \mathcal{E}^A)^{\mathrm{T}} \times (A \circ (\mathcal{E}^A)^{\mathrm{T}}))$        local distance variances
        $v^B \leftarrow tr((B \circ \mathcal{E}^B)^{\mathrm{T}} \times (B \circ (\mathcal{E}^B)^{\mathrm{T}}))$
        $e^A \leftarrow \sum_{i,j=1}^n (A \circ \mathcal{E}^A)_{ij}$        sample means
        $e^B \leftarrow \sum_{i,j=1}^n (B \circ \mathcal{E}^B)_{ij}$
        $c^{kl} \leftarrow (\tilde{c}^{kl} - e^A e^B/n^2)/\sqrt{\left(v^A - (e^A/n)^2\right)\left(v^B - (e^B/n)^2\right)}$        center and normalize
    end function
```

Pseudocode C5 Compute the multiscale correlation map (i.e., all local generalized correlations) in $\mathcal{O}(n^2 \log n/T)$. Once the distances are sorted, the remaining algorithm runs in $\mathcal{O}(n^2)$. An important observation is that each product $a_{ij}b_{ij}$ is included in c^{kl} if and only if (k,l) satisfies $k \le R(A_{\cdot j}, i)$ and $l \le R(B_{\cdot j}, i)$, so it suffices to iterate through $a_{ij}b_{ij}$ for $i,j := 1, \ldots, n$, and add the product simultaneously to all c^{kl} whose scales are no more than $(R(A_{\cdot j}, i), R(B_{\cdot j}, i))$. To achieve the above, we iterate through each product, add it to c^{kl} at $(kl) = (R(A_{\cdot j}, i), R(B_{\cdot j}, i))$ only (so only one local scale is accessed for each operation); then add up adjacent c^{kl} for $k,l = 1, \ldots, n$. The same applies to all local covariances, variances, and expectations.

Input: A pair of distance matrices $(A,B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$.

Output: The multiscale correlation map $\mathcal{C} \in [-1,1]^{n \times n}$for $k,l = 1, \ldots, n$.

```
function MGCAllLocal($A$, $B$)
    for $Z := A, B$ do $\mathcal{E}^Z = \textsc{Sort}(Z)$ end for
    for $Z := A, B$ do $Z = \textsc{Center}(Z)$end for
    for $i,j := 1, \ldots, n$ do        iterate through all local scales to calculate each term
        $k \leftarrow \mathcal{E}^Z_{ij}$
        $l \leftarrow \mathcal{E}^Z_{ij}$
        $\tilde{c}^{kl} \leftarrow \tilde{c}^{kl} + a_{ij}b_{ij}$
        $v^A_k \leftarrow v^A_k + a^2_{ij}$
        $v^B_l \leftarrow v^B_l + b^2_{ij}$
        $e^A_k \leftarrow e^A_k + a_{ij}$
        $e^B_l \leftarrow e^B_l + b_{ij}$
    end for
    for $k := 1, \ldots, n-1$ do        iterate through each scale again and add up adjacent terms
        $\tilde{c}^{1,k+1} \leftarrow \tilde{c}^{1,k} + \tilde{c}^{1,k+1}$
        $\tilde{c}^{k+1,1} \leftarrow \tilde{c}^{k+1,1} + \tilde{c}^{k+1,1}$
        for $Z := A, B$ do $v^Z_{k+1} \leftarrow v^Z_k + v^Z_{k+1}$ end for
        for $Z := A, B$ do $e^Z_{k+1} \leftarrow e^Z_k + e^Z_{k+1}$ end for
    end for
    for $k,l := 1, \ldots, n-1$ do
        $\tilde{c}^{k+1,l+1} \leftarrow \tilde{c}^{k+1,l} + \tilde{c}^{k,l+1} + \tilde{c}^{k+1,l+1} - \tilde{c}^{k,l}$
    end for
    for $k,l := 1, \ldots, n$ do
        $c^{kl} \leftarrow (\tilde{c}^{kl} - e^A_k e^B_l/n^2)/\sqrt{\left(v^A_k - {e^A_k}^2/n^2\right)\left(v^B_l - {e^B_l}^2/n^2\right)}$
    end for
```

Continued on next page

```
end function
```

Pseudocode C6 Power computation of Mgc against a given distribution. By repeatedly sampling from the joint distribution F_{XY}, sample data of size n under the null and the alternative are generated for r Monte-Carlo replicates. The power of Mgc follows by computing the test statistic under the null and the alternative using Algorithm C2. In the simulations we use $r = 10{,}000$ MC replicates. Note that power computation for other benchmarks follows from the same algorithm by plugging in the respective test statistic.

Input: A joint distribution F_{XY}, the sample size n, the number of MC replicates r, and the type 1 error level α.

Output: The power β of Mgc.

```
function MGCPower($F_{XY}$, $n$, $r$, $\alpha$)
    for $t := 1, \ldots, r$ do
        for $i := [n]$ do
            $x_i^0 \overset{iid}{\sim} F_X$, $y_i^0 \overset{iid}{\sim} F_Y$                    sample from null
            $(x_i^1, y_i^1) \overset{iid}{\sim} F_{XY}$,                             sample from alternative
        end for
        for $i,j := 1, \ldots, n$ do
            $a_{ij}^0 = \delta_x(x_i^0, x_j^0)$, $b_{ij}^0 = \delta_y(y_i^0, y_j^0)$         pairwise distances under the null
            $a_{ij}^1 = \delta_x(x_i^1, x_j^1)$, $b_{ij}^1 = \delta_y(y_i^1, y_j^1)$         pairwise distances under the alternative
        end for
        $c_0^*[t] = \text{MGCSampleStat}(A^0, B^0)$                           Mgc statistic under the null
        $c_1^*[t] = \text{MGCSampleStat}(A^1, B^1)$                           Mgc statistic under the alternative
    end for
    $\omega_\alpha \leftarrow \text{Cdf}_{1-\alpha}(c_0^*[t], t \in [r])$                    the critical value of Mgc under the null
    $\beta \leftarrow \sum_{t=1}^{r} (c_1^*[t] > \omega_\alpha)/r$                        compute power by the alternative distribution
end function
```

Simulation dependence functions

This section provides the 20 different dependency functions used in the simulations. We used essentially the exact same relationships as previous publications to ensure a fair comparison (***Székely et al., 2007***; ***Simon and Tibshirani, 2012***; ***Gorfine et al., 2012***). We only made changes to add white noise and a weight vector for higher dimensions, thereby making them more difficult, to better compare all methods throughout different dimensions and sample sizes. A few additional relationships are also included.

For each sample $\boldsymbol{x} \in \mathbb{R}^p$, we denote $\boldsymbol{x}_{[d]}, d = 1, \ldots, p$ as the d^{th} dimension of the vector $\boldsymbol{x}$. For the purpose of high-dimensional simulations, $w \in \mathbb{R}^p$ is a decaying vector with $w_{[d]} = 1/d$ for each d, such that $w^T\boldsymbol{x}$ is a weighted summation of all dimensions of $\boldsymbol{x}$. Furthermore, $\mathcal{U}(a,b)$ denotes the uniform distribution on the interval (a,b), $\mathcal{B}(p)$ denotes the Bernoulli distribution with probability p, $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution with mean μ and covariance Σ, U and V represent some auxiliary random variables, κ is a scalar constant to control the noise level (which equals 1 for one-dimensional simulations and 0 otherwise), and ϵ is a white noise from independent standard normal distribution unless mentioned otherwise.

For all the below equations, $(X,Y) \overset{iid}{\sim} F_{XY} = F_{Y|X}F_X$. For each relationship, we provide the space of (X,Y), and define $F_{Y|X}$ and F_X, as well as any additional auxiliary distributions.

1. Linear $(X,Y) \in \mathbb{R}^p \times \mathbb{R}$,

$$X \sim \mathcal{U}(-1,1)^p,$$
$$Y = w^T X + \kappa\epsilon.$$

2. Exponential $(X,Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$\begin{aligned}&X\sim\mathcal{U}(0,3)^p,\\&Y=exp(w^TX)+10\kappa\epsilon.\end{aligned}$$

3. Cubic $(X,Y)\in\mathbb{R}^p\times\mathbb{R}$:

$$\begin{aligned}&X\sim\mathcal{U}(-1,\,1)^p,\\&Y=128(w^TX-\frac{1}{3})^3+48(w^TX-\frac{1}{3})^2-12(w^TX-\frac{1}{3})+80\kappa\epsilon.\end{aligned}$$

4. Joint normal $(X,Y)\in\mathbb{R}^p\times\mathbb{R}^p$: Let $\rho=1/2p$, I_p be the identity matrix of size $p\times p$, J_p be the matrix of ones of size $p\times p$, and $\Sigma=\begin{bmatrix}I_p & \rho J_p\\ \rho J_p & (1+0.5\kappa)I_p\end{bmatrix}$. Then

$$(X,Y)\sim\mathcal{N}(0,\Sigma).$$

5. Step Function $(X,Y)\in\mathbb{R}^p\times\mathbb{R}$

$$\begin{aligned}&X\sim\mathcal{U}(-1,1)^p,\\&Y=\boldsymbol{I}(w^TX>0)+\epsilon,\end{aligned}$$

where $\boldsymbol{I}$ is the indicator function, that is $\boldsymbol{I}(z)$ is unity whenever z true, and zero otherwise.

6. Quadratic $(X,Y)\in\mathbb{R}^p\times\mathbb{R}$:

$$\begin{aligned}&X\sim\mathcal{U}(-1,1)^p,\\&Y=(w^TX)^2+0.5\kappa\epsilon.\end{aligned}$$

7. W Shape $(X,Y)\in\mathbb{R}^p\times\mathbb{R}:\ U\sim\mathcal{U}(-1,1)^p$,

$$\begin{aligned}&X\sim\mathcal{U}(-1,1)^p,\\&Y=4\left[\left((w^TX)^2-\frac{1}{2}\right)^2+w^TU/500\right]+0.5\kappa\epsilon.\end{aligned}$$

8. Spiral $(X,Y)\in\mathbb{R}^p\times\mathbb{R}:\ U\sim\mathcal{U}(0,5),\ \epsilon\sim\mathcal{N}(0,1)$

$$\begin{aligned}&X_{[d]}=U\sin(\pi U)\cos^d(\pi U)\,\textbf{for}\,d=1,\ldots,p-1,\\&X_{[d]}=U\cos^p(\pi U),\\&Y=U\sin(\pi U)+0.4p\epsilon.\end{aligned}$$

9. Uncorrelated Bernoulli $(X,Y)\in\mathbb{R}^p\times\mathbb{R}:\ U\sim\mathcal{B}(0.5)\,\epsilon_1\sim\mathcal{N}(0,I_p),\ \epsilon_2\sim\mathcal{N}(0,1)$,

$$\begin{aligned}&X\sim\mathcal{B}(0.5)^p+0.5\epsilon_1,\\&Y=(2U-1)w^TX+0.5\epsilon_2.\end{aligned}$$

10. Logarithmic $(X,Y)\in\mathbb{R}^p\times\mathbb{R}^p:\ \epsilon\sim\mathcal{N}(0,I_p)$

$$\begin{aligned}&X\sim\mathcal{N}(0,I_p),\\&Y_{[d]}=2\log_2(|X_{[d]}|)+3\kappa\epsilon_{[d]}\,\textbf{for}\,d=1,\ldots,p.\end{aligned}$$

11. Fourth Root $(X,Y)\in\mathbb{R}^p\times\mathbb{R}^p:$

$$\begin{aligned}&X\sim\mathcal{U}(-1,1)^p,\\&Y=|w^TX|^{\frac{1}{4}}+\frac{\kappa}{4}\epsilon.\end{aligned}$$

12. Sine Period $4\pi(X,Y)\in\mathbb{R}^p\times\mathbb{R}^p:U\sim\mathcal{U}(-1,1),V\sim\mathcal{N}(0,1)^p,\theta=4\pi$,

$$\begin{aligned}&X_{[d]}=U+0.02pV_{[d]}\,\textbf{for}\,d=1,\ldots,p,\\&Y=\sin(\theta X)+\kappa\epsilon.\end{aligned}$$

13. Sine Period $16\pi\,(X,Y)\in\mathbb{R}^p\times\mathbb{R}^p$: Same as above except $\theta=16\pi$ and the noise on Y is changed to $0.5\kappa\epsilon$.

14. Square $(X,Y)\in\mathbb{R}^p\times\mathbb{R}^p$: Let $U\sim\mathcal{U}(-1,1)$, $V\sim\mathcal{U}(-1,1)$, $\epsilon\sim\mathcal{N}(0,1)^p$, $\theta=-\frac{\pi}{8}$. Then

$$X_{[d]} = U\cos\theta + V\sin\theta + 0.05p\epsilon_{[d]},$$
$$Y_{[d]} = -U\sin\theta + V\cos\theta,$$

for $d = 1, \ldots, p$.

15. Two Parabolas $(X,Y) \in \mathbb{R}^p \times \mathbb{R}$: $\epsilon \sim \mathcal{U}(0,1)$, $U \sim \mathcal{B}(0.5)$,

$$X \sim \mathcal{U}(-1,1)^p,$$
$$Y = \left((w^{\mathrm{T}}X)^2 + 2\kappa\epsilon\right) \cdot (U - \frac{1}{2}).$$

16. Circle $(X,Y) \in \mathbb{R}^p \times \mathbb{R} : U \sim \mathcal{U}(-1,1)^p, \epsilon \sim \mathcal{N}(0, I_p), r = 1,$

$$X_{[d]} = r\left(\sin(\pi U_{[d+1]})\prod_{j=1}^{d}\cos(\pi U_{[j]}) + 0.4\epsilon_{[d]}\right) \textbf{for } d = 1, \ldots, p-1,$$
$$X_{[p]} = r\left(\prod_{j=1}^{p}\cos(\pi U_{[j]}) + 0.4\epsilon_{[p]}\right),$$
$$Y = \sin(\pi U_{[1]}).$$

17. Ellipse $(X,Y) \in \mathbb{R}^p \times \mathbb{R}$: Same as above except $r = 5$.
18. Diamond $(X,Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Same as 'Square' except $\theta = -\frac{\pi}{4}$.
19. Multiplicative Noise $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^p \times \mathbb{R}^p : u \sim \mathcal{N}(0, I_p)$,

$$x \sim \mathcal{N}(0, I_p),$$
$$y_{[d]} = u_{[d]}\boldsymbol{x}_{[d]} \textbf{ for } d = 1, \ldots, p.$$

20. Multimodal Independence $(X,Y) \in \mathbb{R}^p \times \mathbb{R}^p : \text{Let } U \sim \mathcal{N}(0, I_p), V \sim \mathcal{N}(0, I_p), U' \sim \mathcal{B}(0.5)^p, V' \sim \mathcal{B}(0.5)^p$. Then

$$X = U/3 + 2U' - 1,$$
$$Y = V/3 + 2V' - 1.$$

For each distribution, X and Y are dependent except (20); for some relationships (8,14,16-18) they are independent upon conditioning on the respective auxiliary variables, while for others they are 'directly' dependent. A visualization of each dependency with $D = D_y = 1$ is shown in ***Figure 2—figure supplement 1***.

For the increasing dimension simulation in the main paper, we always set $\kappa = 0$ and $n = 100$, with p increasing. Note that $q = p$ for types 4, 10, 12, 13, 14, 18, 19, 20,, otherwise $q = 1$. The decaying vector w is utilized for $p > 1$ to make the high-dimensional relationships more difficult (otherwise, additional dimensions only add more signal). For the one-dimensional simulations, we always set $p = q = 1$, $\kappa = 1$ and $n = 100$.

Acknowledgements

This work was partially supported by the Child Mind Institute Endeavor Scientist Program, the National Science Foundation award DMS-1712947, the National Security Science and Engineering Faculty Fellowship (NSSEFF), the Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), the Defense Advanced Research Projects Agency's (DARPA) SIMPLEX program through SPAWAR contract N66001-15-C-4041, the XDATA program of DARPA administered through Air Force Research Laboratory contract FA8750-12-2-0303, DARPA Lifelong Learning Machines program through contract FA8650-18-2-7834, the Office of Naval Research contract N00014-12-1-0601, the Air Force Office of Scientific Research contract FA9550-14-1-0033. The authors thank Dr. Brett Mensh of Optimize Science for acting as our intellectual consigliere, Julia Kuhl for help with figures, and Dr. Ruth Heller, Dr. Bert Vogelstein, Dr. Don Geman, and Dr. Yakir Reshef for insightful suggestions. And we'd like to thank Satish Palaniappan, Sambit Panda, and Junhao (Bear) Xiong for porting the code to Python.

Additional information

Funding

Funder	Grant reference number	Author
Child Mind Institute	Endeavor Scientist Program	Joshua T Vogelstein
National Science Foundation		Joshua T Vogelstein
Defense Advanced Research Projects Agency		Joshua T Vogelstein
Office of Naval Research		Joshua T Vogelstein
Air Force Office of Scientific Research		Joshua T Vogelstein

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Joshua T Vogelstein, Conceptualization, Formal analysis, Supervision, Funding acquisition, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing; Eric W Bridgeford, Software; Qing Wang, Data curation; Carey E Priebe, Supervision; Mauro Maggioni, Conceptualization; Cencheng Shen, Conceptualization, Formal analysis, Investigation, Visualization, Methodology, Writing—review and editing

Author ORCIDs

Joshua T Vogelstein http://orcid.org/0000-0003-2487-6237
Cencheng Shen http://orcid.org/0000-0003-1030-1432

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.41690.029
Author response https://doi.org/10.7554/eLife.41690.030

Additional files

Supplementary files

- Transparent reporting form

DOI: https://doi.org/10.7554/eLife.41690.018

Data availability

To facilitate reproducibility, we make all datasets available from: https://github.com/neurodata/MGC-paper/tree/master/Data/Preprocessed (copy archived at https://github.com/elifesciences-publications/MGC-paper).

References

Adelstein JS, Shehzad Z, Mennes M, Deyoung CG, Zuo XN, Kelly C, Margulies DS, Bloomfield A, Gray JR, Castellanos FX, Milham MP. 2011. Personality is reflected in the brain's intrinsic functional architecture. *PLOS ONE* **6**:e27633. DOI: https://doi.org/10.1371/journal.pone.0027633, PMID: 22140453

Allard WK, Chen G, Maggioni M. 2012. Multi-scale geometric methods for data sets II: geometric Multi-Resolution analysis. *Applied and Computational Harmonic Analysis* **32**:435–462. DOI: https://doi.org/10.1016/j.acha.2011.08.001

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B* **57**:289–300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berk R, Brown L, Buja A, Zhang K, Zhao L. 2013. Valid post-selection inference. *The Annals of Statistics* **41**:802–837. DOI: https://doi.org/10.1214/12-AOS1077

Bhat K, Wang F, Ma Q, Li Q, Mallik S, Hsieh TC, Wu E. 2012. Advances in biomarker research for pancreatic cancer. *Current Pharmaceutical Design* **18**:2439–2451. DOI: https://doi.org/10.2174/138161282112092439, PMID: 22372502

Biswal BB, Mennes M, Zuo XN, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL, Colcombe S, Dogonowski AM, Ernst M, Fair D, Hampson M, Hoptman MJ, Hyde JS, Kiviniemi VJ, Kötter R, Li SJ, Lin CP, et al. 2010. Toward discovery science of human brain function. *PNAS* **107**:4734–4739. DOI: https://doi.org/10.1073/pnas.0911855107, PMID: 20176931

Bjørnebekk A, Fjell AM, Walhovd KB, Grydeland H, Torgersen S, Westlye LT. 2013. Neuronal correlates of the five factor model (FFM) of human personality: multimodal imaging in a large healthy sample. *NeuroImage* **65**:194–208. DOI: https://doi.org/10.1016/j.neuroimage.2012.10.009, PMID: 23063449

Bracewell RN, Bracewell RN. 1986. *The Fourier Transform and Its Applications*. New York: McGraw-Hill.

Bridgeford E, Shen C, Vogelstein J. 2018. R package for MGC code. 034795d. GitHub. https://github.com/neurodata/mgc

Chaudhuri A, Hu W. 2018. A fast algorithm for computing distance correlation. *Arxiv*. https://arxiv.org/abs/1810.11332.

Coifman RR, Maggioni M. 2006. Diffusion wavelets. *Applied and Computational Harmonic Analysis* **21**:53–94. DOI: https://doi.org/10.1016/j.acha.2006.04.004

Costa PT, McCrae RR. 1992. *Neo PI-R Professional Manual*. **396** PAR Psychological Assessment R.

Craddock C, Sikka S, Cheung B, Khanuja R, Ghosh SS, Yan C, Li Q, Lurie D, Vogelstein J, Burns R, Colcombe S, Mennes M, Kelly C, Martino AD, Castellanos FX, Milham M. 2013. Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (C-PAC). *Frontiers in Neuroinformatics* **7**. DOI: https://doi.org/10.3389/conf.fninf.2014.08.00117

Da Zheng DM, Burns R, Vogelstein JT, Priebe CE, Szalay AS. 2015. FlashGraph: processing Billion-Node graphs on an array of commodity SSDs. USENIX Conference on File and Storage Technologies.

Da Zheng DM, Vogelstein JT, Priebe CE, Burns R. 2016a. FlashMatrix: parallel, scalable data analysis with generalized matrix operations using commodity SSDs. *arXiv* . http://arxiv.org/abs/1604.06414v1.

Da Zheng RB, Vogelstein JT, Priebe CE, Szalay AS. 2016b. An SSD-based eigensolver for spectral analysis on billion-node graphs. *arXiv*. https://arxiv.org/abs/1602.01421.

Da Zheng DM, Lyzinski V, Vogelstein JT, Priebe CE, Burns R. 2016c. Semi-External Memory Sparse Matrix Multiplication on Billion-node Graphs in a Multicore Architecture. *arXiv*. http://arxiv.org/abs/1602.02864.

Daubechies I. 1992. *Ten Lectures on Wavelets*. SIAM.

DeYoung CG, Hirsh JB, Shane MS, Papademetris X, Rajeevan N, Gray JR. 2010. Testing predictions from personality neuroscience. brain structure and the big five. *Psychological Science* **21**:820–828. DOI: https://doi.org/10.1177/0956797610370159, PMID: 20435951

Edelsbrunner H, Harer JL. 2009. *Computational Topology: An Introduction*. American Mathematical Society.

Eklund A, Andersson M, Josephson C, Johannesson M, Knutsson H. 2012. Does Parametric fMRI analysis with SPM yield valid results? an empirical study of 1484 rest datasets. *NeuroImage* **61**:565–578. DOI: https://doi.org/10.1016/j.neuroimage.2012.03.093, PMID: 22507229

Eklund A, Nichols TE, Knutsson H. 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *PNAS* **113**:7900–7905. DOI: https://doi.org/10.1073/pnas.1602413113, PMID: 27357684

Frantzi M, Bhat A, Latosinska A. 2014. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clinical and Translational Medicine* **3**:7. DOI: https://doi.org/10.1186/2001-1326-3-7, PMID: 24679154

Friedman JH, Rafsky LC. 1983. Graph-Theoretic measures of multivariate association and prediction. *The Annals of Statistics* **11**:377–391. DOI: https://doi.org/10.1214/aos/1176346148

Giryes R, Sapiro G, Bronstein AM. 2015. Deep neural networks with random gaussian weights: a universal classification strategy. *Arxiv*. https://arxiv.org/abs/1504.08291.

Good P. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. 2014. Generative adversarial nets. In: *Advances in Neural Information Processing System*. MIT Press. p. 2672–2680.

Gorfine M, Heller R, Heller Y. 2012. *Comment on Detecting Novel Associations in Large Data Sets*: Israel Institute of Technology.

Gretton A, Borgwardt KM, Rasch M, Schölkopf B, Smola AJ. 2006. A kernel method for the two-sample-problem. In: *Advances in Neural Information Processing Systems*. MIT Press. p. 513–520.

Gretton A, Gyorfi L. 2010. Consistent nonparametric tests of independence. *Journal of Machine Learning Research* **11**:1391–1423.

Guillot G, Rousset F. 2013. Dismantling the mantel tests. *Methods in Ecology and Evolution* **4**:336–344. DOI: https://doi.org/10.1111/2041-210x.12018

Hastie T, Tibshirani R, Friedman JH. 2001. *Elements of Statistical Learning*. New York: Springer.

Helfman D, Flynn P, Khan P, Saeed A. 2018. Tropomyosin as a regulator of cancer cell transformation. *Advances in Experimental Medicine and Biology* **644**:124–131. DOI: https://doi.org/10.1007/978-0-387-85766-4_10

Heller R, Heller Y, Gorfine M. 2013. A consistent multivariate test of association based on ranks of distances. *Biometrika* **100**:503–510. DOI: https://doi.org/10.1093/biomet/ass070

Heller R, Heller Y, Kaufman S, Brill B, Gorfine M. 2016. Consistent distribution-free -sample and independence tests for univariate random variables. *Journal of Machine Learning Research* **17**:1–54.

Hoeffding W. 1948. A Non-Parametric test of independence. *The Annals of Mathematical Statistics* **19**:546–557. DOI: https://doi.org/10.1214/aoms/1177730150

Hotelling H. 1936. Relations between two sets of variates. *Biometrika* **28**:321–377. DOI: https://doi.org/10.1093/biomet/28.3-4.321

Huang C, Huo X. 2017. A statistically and numerically efficient independence test based on random projections and distance covariance. *arXiv*. https://arxiv.org/abs/1701.06054.

Huo X, Székely GJ. 2016. Fast computing for distance covariance. *Technometrics* **58**:435–447. DOI: https://doi.org/10.1080/00401706.2015.1054435

Josse J, Holmes S. 2013. Measures of dependence between random vectors and tests of independence. *arXiv*. http://arxiv.org/abs/1307.7383.

Jung RE, Segall JM, Jeremy Bockholt H, Flores RA, Smith SM, Chavez RS, Haier RJ. 2009. Neuroanatomy of creativity. *Human Brain Mapping* **43**:NA. DOI: https://doi.org/10.1002/hbm.20874

Karsani SA, Saihen NA, Zain RB, Cheong SC, Abdul Rahman M. 2014. Comparative proteomics analysis of oral cancer cell lines: identification of cancer associated proteins. *Proteome Science* **12**:3. DOI: https://doi.org/10.1186/1477-5956-12-3, PMID: 24422745

Kendall MG. 1970. *Rank Correlation Methods*. London: Griffin.

Kettenring JR. 1971. Canonical analysis of several sets of variables. *Biometrika* **58**:433–451. DOI: https://doi.org/10.1093/biomet/58.3.433

Koutra D, Shah N, Vogelstein JT, Gallagher BJ, Faloutsos C. 2015. DeltaCon: a principled massive-graph similarity function. *ACM Transactions on Knowledge Discovery From Data* **10**. DOI: https://doi.org/10.1145/2824443

Lam CY, Yip CW, Poon TC, Cheng CK, Ng EW, Wong NC, Cheung PF, Lai PB, Ng IO, Fan ST, Cheung ST. 2012. Identification and characterization of tropomyosin 3 associated with granulin-epithelin precursor in human hepatocellular carcinoma. *PLOS ONE* **7**:e40324. DOI: https://doi.org/10.1371/journal.pone.0040324, PMID: 22792281

Lee HH, Lim CA, Cheong YT, Singh M, Gam LH. 2012. Comparison of protein expression profiles of different stages of lymph nodes metastasis in breast cancer. *International Journal of Biological Sciences* **8**:353–362. DOI: https://doi.org/10.7150/ijbs.3157, PMID: 22393307

Lee JA, Verleysen M. 2007. *Nonlinear Dimensionality Reduction*. Springer Science & Business Media.

Levina E, Bickel PJ. 2004. Maximum likelihood estimation of intrinsic dimension. In: *Advances in Neural Information Processing Systems*. MIT Press.

Li R, Zhong W, Zhu L. 2012. Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**:1129–1139. DOI: https://doi.org/10.1080/01621459.2012.695654, PMID: 25249709

Lyons R. 2013. Distance covariance in metric spaces. *The Annals of Probability* **41**:3284–3305. DOI: https://doi.org/10.1214/12-AOP803

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**:209–220. PMID: 6018555

Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning* **10**:1–141. DOI: https://doi.org/10.1561/2200000060

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

Pearson K. 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58**:240–242. DOI: https://doi.org/10.1098/rspl.1895.0041

Prescott JW. 2013. Quantitative imaging biomarkers: the application of advanced image processing and analysis to clinical and preclinical decision making. *Journal of Digital Imaging* **26**:97–108. DOI: https://doi.org/10.1007/s10278-012-9465-7, PMID: 22415112

Reimherr M, Nicolae DL. 2013. On quantifying dependence: a framework for developing interpretable measures. *Statistical Science* **28**:116–130. DOI: https://doi.org/10.1214/12-STS405

Rényi A. 1959. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae* **10**:441–451. DOI: https://doi.org/10.1007/BF02024507

Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. 2011. Detecting novel associations in large data sets. *Science* **334**:1518–1524. DOI: https://doi.org/10.1126/science.1205438, PMID: 22174245

Rizzo ML, Székely GJ. 2010. DISCO analysis: a nonparametric extension of analysis of variance. *The Annals of Applied Statistics* **4**:1034–1055. DOI: https://doi.org/10.1214/09-AOAS245

Rizzo ML, Székely GJ. 2016. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics* **8**:27–38. DOI: https://doi.org/10.1002/wics.1375

Roncal WG, Koterba ZH, Mhembere D, Kleissas DM, Vogelstein JT, Burns R, Bowles AR, Donavos DK, Ryman S, Jung RE, Wu L, Calhoun VD, Jacob Vogelstein R. 2013. MIGRAINE: mri graph reliability analysis and inference for connectomics. Global Conference on Signal and Information Processing.

Roweis ST, Saul LK. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**:2323–2326. DOI: https://doi.org/10.1126/science.290.5500.2323, PMID: 11125150

Sampaio A, Soares JM, Coutinho J, Sousa N, Gonçalves ÓF. 2014. The big five default brain: functional evidence. *Brain Structure and Function* **219**:1913–1922. DOI: https://doi.org/10.1007/s00429-013-0610-y, PMID: 23881294

Schilling MF. 1986. Multivariate Two-Sample tests based on nearest neighbors. *Journal of the American Statistical Association* **81**:799–806. DOI: https://doi.org/10.1080/01621459.1986.10478337

Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* **41**:2263–2291. DOI: https://doi.org/10.1214/13-AOS1140

Shen C, Sun M, Tang M, Priebe CE. 2014. Generalized canonical correlation analysis for classification. *Journal of Multivariate Analysis* **130**:310–322. DOI: https://doi.org/10.1016/j.jmva.2014.05.011

Shen C, Vogelstein JT, Priebe CE. 2017. Manifold matching using shortest-path distance and joint neighborhood selection. *Pattern Recognition Letters* **92**:41–48. DOI: https://doi.org/10.1016/j.patrec.2017.04.005

Shen C, Priebe CE, Vogelstein JT. 2018. From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association*:1–39. DOI: https://doi.org/10.1080/01621459.2018.1543125

Shen C, Vogelstein JT. 2018. The exact equivalence of distance and kernel methods for hypothesis testing. *arXiv*. https://arxiv.org/abs/1806.05514.

Simon N, Tibshirani R. 2012. Comment on "Detecting Novel Associations In Large Data Sets" by Reshef Et Al, Science Dec 16, 2011. *arXiv*. http://arxiv.org/abs/1401.7645.

Spearman C. 1904. The proof and measurement of association between two things. *The American Journal of Psychology* **15**:72. DOI: https://doi.org/10.2307/1412159

Sun Y, Xia Z, Shang Z, Sun K, Niu X, Qian L, Fan L-Y, Cao C-X, Xiao H. 2016. Facile preparation of salivary extracellular vesicles for cancer proteomics. *Scientific Reports* **6**. DOI: https://doi.org/10.1038/srep24669

Sussman DL, Tang M, Fishkind DE, Priebe CE. 2012. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* **107**:1119–1128. DOI: https://doi.org/10.1080/01621459.2012.699795

Sutherland DJ, Tung H-Y, Strathmann H, De S, Ramdas A, Smola A, Gretton A. 2016. Generative models and model criticism via optimized maximum mean discrepancy. International Conference on Learning Representations.

Székely GJ, Rizzo ML, Bakirov NK. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**:2769–2794. DOI: https://doi.org/10.1214/009053607000000505

Szekely GJ, Rizzo ML. 2004. Testing for equal distributions in high dimension. *InterStat* **10**.

Székely GJ, Rizzo ML. 2005. A new test for multivariate normality. *Journal of Multivariate Analysis* **93**:58–80. DOI: https://doi.org/10.1016/j.jmva.2003.12.002

Székely GJ, Rizzo ML. 2009. Brownian distance covariance. *The Annals of Applied Statistics* **3**:1236–1265. DOI: https://doi.org/10.1214/09-AOAS312

Szekely GJ, Rizzo M. 2013. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis* **117**:193–213. DOI: https://doi.org/10.1016/j.jmva.2013.02.012

Székely GJ, Rizzo ML. 2014. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics* **42**:2382–2412. DOI: https://doi.org/10.1214/14-AOS1255

Tang M, Athreya A, Sussman DL, Lyzinski V, Park Y, Priebe CE. 2017. A semiparametric Two-Sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics* **26**:344–354. DOI: https://doi.org/10.1080/10618600.2016.1193505

Tenenbaum JB, de Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**:2319–2323. DOI: https://doi.org/10.1126/science.290.5500.2319, PMID: 11125149

Tenenhaus A, Tenenhaus M. 2011. Regularized generalized canonical correlation analysis. *Psychometrika* **76**:257–284. DOI: https://doi.org/10.1007/s11336-011-9206-8

Wang Q, Chaerkady R, Wu J, Hwang HJ, Papadopoulos N, Kopelovich L, Maitra A, Matthaei H, Eshleman JR, Hruban RH, Kinzler KW, Pandey A, Vogelstein B. 2011. Mutant proteins as cancer-specific biomarkers. *PNAS* **108**:2444–2449. DOI: https://doi.org/10.1073/pnas.1019203108, PMID: 21248225

Wang X, Pan W, Hu W, Tian Y, Zhang H. 2015. Conditional distance correlation. *Journal of the American Statistical Association* **110**:1726–1734. DOI: https://doi.org/10.1080/01621459.2014.993081, PMID: 26877569

Wang Q, Zhang M, Tomita T, Vogelstein JT, Zhou S, Papadopoulos N, Kinzler KW, Vogelstein B. 2017. Selected reaction monitoring approach for validating peptide biomarkers. *PNAS* **114**:13519–13524. DOI: https://doi.org/10.1073/pnas.1712731114

Willemse EAJ, De Vos A, Herries EM, Andreasson U, Engelborghs S, van der Flier WM, Scheltens P, Crimmins D, Ladenson JH, Vanmechelen E, Zetterberg H, Fagan AM, Blennow K, Bjerke M, Teunissen CE. 2018. Neurogranin as cerebrospinal fluid biomarker for alzheimer disease: an assay comparison study. *Clinical Chemistry* **64**:927–937. DOI: https://doi.org/10.1373/clinchem.2017.283028, PMID: 29523639

Witten DM, Tibshirani R, Hastie T. 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**:515–534. DOI: https://doi.org/10.1093/biostatistics/kxp008, PMID: 19377034

Witten DM, Tibshirani R. 2011. Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B* **73**:753–772. DOI: https://doi.org/10.1111/j.1467-9868.2011.00783.x

Xing EP, Ng AY, Jordan MI, Russell S. 2003. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems* **15**:505–512.

Xu J, Potenza MN. 2012. White matter integrity and five-factor personality measures in healthy adults. *NeuroImage* **59**:800–807. DOI: https://doi.org/10.1016/j.neuroimage.2011.07.040, PMID: 21840401

Yang J, Korley FK, Dai M, Everett AD. 2015. Serum neurogranin measurement as a biomarker of acute traumatic brain injury. *Clinical Biochemistry* **48**:843–848. DOI: https://doi.org/10.1016/j.clinbiochem.2015.05.015, PMID: 26025774

Zhang JH, Chung TD, Oldenburg KR. 1999. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of Biomolecular Screening* **4**:67–73. DOI: https://doi.org/10.1177/108705719900400206, PMID: 10838414

Zhang Z, Wang J, Zha H. 2012. Adaptive manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**:253–265. DOI: https://doi.org/10.1109/TPAMI.2011.115, PMID: 21670485

Zhang Q, Filippi S, Gretton A, Sejdinovic D. 2018. Large-scale kernel methods for independence testing. *Statistics and Computing* **28**:113–130. DOI: https://doi.org/10.1007/s11222-016-9721-7

Zhong W, Zhu L. 2015. An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation* **85**:2331–2345. DOI: https://doi.org/10.1080/00949655.2014.928820

Appendix 1

DOI: https://doi.org/10.7554/eLife.41690.019

Real data processing

Brain activity vs personality

This experiment investigates whether there is any dependency between resting brain activity and personality. Human personality has been intensively studied for many decades; the most widely used and studied approach is the NEO Personality Inventory-Revised the characterized personality along five dimensions (***Costa and McCrae, 1992***).

This dataset consists of 42 subjects, each with 197 time-steps of resting-state functional magnetic resonance activity (rs-fMRI) activity, as well as the subject's five-dimensional 'personality'. Adelstein et al. (***Adelstein et al., 2011***) were able to detect dependence between the activity of certain brain regions and dimensions of personality, but lacked the tools to test for dependence of whole brain activity against all five dimensions of personality.

For the five-factor personality modality, we used the Euclidean distance. For the brain activity modality, we derived the following comparison function. For each scan, (i) run Configurable Pipeline for the Analysis of Connectomes pipeline (***Craddock et al., 2013***) to process the raw brain images yielding a parcellation into 197 regions of interest, (ii) run a spectral analysis on each region and keep the power of band, (iii) bandpass and normalize it to sum to one, (iv) calculate the Kullback-Leibler divergence across regions to obtain a similarity matrix across comparing all regions. Then, use the normalized Hellinger distance to compute distances between each subject.

Brain connectivity vs creativity

This experiment investigates whether there is any dependency between brain structural networks and creativity. Creativity has been extensively studied in psychology; the 'creativity composite index' (CCI) is an index similar to an 'intelligence quotient' but for creativity rather than intelligence (***Jung et al., 2009***).

This dataset consists of 109 subjects, each with diffusion weighted MRI data as well as the subject's CCI. Neural correlates of CCI have previously been investigated, though largely using structural MRI and cortical thickness (***Jung et al., 2009***). Previously published results explored the relationship between graphs and CCI (***Koutra et al., 2015***), but did not provide a valid test.

We used Euclidean distance to compare CCI values. For the raw brain imaging data, we derived the following comparison function. For each scan we estimated brain networks from diffusion and structural MRI data via Migraine, a pipeline for estimating brain networks from diffusion data (***Roncal et al., 2013***). We compute the distance between the graphs using the semi-parametric graph test statistic (***Sussman et al., 2012***; ***Shen et al., 2017***; ***Tang et al., 2017***), embedding each graph into two dimensions and aligning the embeddings via a Procrustes analysis.

Proteins vs cancer

This experiment investigated whether there is any dependency between abundance levels of peptides in human plasma and the presence of cancers. Selected Reaction Monitoring (SRM) is a targeted quantitative proteomics technique for measuring protein and peptide abundance in complicated biological samples (***Wang et al., 2011***). In a previous study, we used SRM to identify 318 peptides from 33 normal, 10 pancreatic cancer, 28 colorectal cancer, and 24 ovarian cancer samples (***Wang et al., 2017***). Then, using other methods, we identifed three peptides that were implicated in ovarian cancer, and validated them as legitimate biomarkers with a follow-up experiment.

In this study, we performed the following five sets of tests on those data:

1. Ovarian vs. normal for all proteins,
2. Ovarian vs. normal for each individual protein,
3. Pancreas vs. normal for all proteins,
4. Pancreas vs. all others for each individual protein,
5. Pancreas vs. normal for each individual protein.

These tests are designed to first validate the Mgc method from ovarian cancer, then identify biomarkers unique to pancreatic cancer, that is, find a protein that is able to tell the difference between pancreas and normals, as well as pancreas vs all other cancers. For each of the five tests, we create a binary label vector, with 1 indicating the cancer type of interest for the corresponding subject, and 0 otherwise. Then each algorithm is applied to each task. For all tests we used Euclidean distances and the type 1 error level is set to $\alpha = 0.05$. The three test sets assessing individual proteins provide 318 p-values; we used the Benjamini-Hochberg procedure (***Benjamini and Hochberg, 1995***) to control the false discovery rate. A summary of the results are reported in ***Appendix 1—table 1***.

Appendix 1—table 1. Results for cancer peptide screening. The first two rows report the p-values for the tests of interest based on all peptides. The next four rows report the number of significant proteins from individual peptide tests; the Benjamini-Hochberg procedure is used to locate the significant peptides by controlling the false discovery rate at 0.05.

	Testing pairs / Methods	**Mgc**	**Mantel**	**Dcorr**	**Mcorr**	**Hhg**
1	Ovar vs. Norm: p-value	**0.0001**	**0.0001**	**0.0001**	**0.0001**	**0.0001**
2	Ovar vs. Norm: # peptides	218	190	186	178	225
3	Pancr vs. Norm: p-value	**0.0082**	0.0685	0.0669	0.0192	0.0328
4	Panc vs. Norm: # peptides	9	7	6	7	11
5	Panc vs. All: # peptides	1	0	0	0	3
6	# peptides unique to Panc	1	0	0	0	2
7	# false positives for Panc	**0**	n/a	n/a	n/a	1

DOI: https://doi.org/10.7554/eLife.41690.020

The following source data is available for Appendix 1—table 2:

Appendix 1—table 1—Source data 1. Ovarian testing results.

DOI: https://doi.org/10.7554/eLife.41690.021

Appendix 1—table 1—Source data 2 Pancreatic testing results.

DOI: https://doi.org/10.7554/eLife.41690.022

Appendix 1—table 1—Source data 3. Peptide screening results for pancreatic.

DOI: https://doi.org/10.7554/eLife.41690.023

All methods are able to successfully detect a dependence between peptide abundances in ovarian cancer samples versus normal samples (***Appendix 1—table 1***, line 1). This is likely because there are so many individual peptides that have different abundance distributions between ovarian and normal samples (***Appendix 1—table 1***, line 2). Nonetheless, Mgc identified more putative biomarkers than any of the other methods. While we have not checked all of them with subsequent experiments to identify potential false positives, we do know from previous experiments that three peptides in particular are effective biomarkers.

All three peptides have p-value $\approx$ 0 for all methods including Mgc, that is, they are all correctly identified as significant. However, by ranking the peptides based on the actual test statistic of each peptide, Mgc is the method that ranks the three known biomarkers the lowest, suggesting that it is the least likely to falsely identify peptides.

We then investigated the pancreatic samples in an effort to identify biomarkers that are unique to pancreas. We first checked whether the methods could identify a difference using all the peptides. Indeed, three methods found a dependence at the 0.05 level, with Mgc obtaining the lowest p-value (***Appendix 1—table 1***, line 3). We then investigated how many individual peptides the methods identified; all of them found 6 to 11 peptides with a significant difference between pancreatic and normal samples (***Appendix 1—table 1***, line 4). Because we were interested in identifying peptides that were uniquely useful for pancreatic cancer, we then compared pancreatic samples to all others. At significance level 0.05, only Mgc, Hsic, and Hhg identified peptides that expressed different abundances in this more challenging case, and we list the top four peptides in ***Appendix 1—table 2*** along with the corrected p-value for each peptide.

Appendix 1—table 2. For each of Mgc, Dcorr, Mcorr, Hhg, Hsic, Mantel, Pearson, and Mic, list the top four peptides identified for Panc vs All and the respective corrected p-value using Benjamini-Hochberg. Bold indicates a significant peptide at type 1 error level 0.05. The top candidates are very much alike except Mic. In particular, neurogranin is consistently among the top candidates for all methods, but is only significant while using Mgc, Hsic, and Hhg; there are two other significant proteins from Hsic and Hhg, but they do not further improve the classification performance comparing to just using neurogranin. Note that the p-values from Mantel and Pearson are always 1 after Benjamini-Hochberg correction, so their respective top peptides are identified using raw p-values without correction.

method	**Top four identified peptides**			
Mgc	**neurogranin**	fibrinogen protein 1	tropomyosin alpha-3	ras suppressor protein 1
p-value	**0.03**	0.33	0.49	0.52
Dcorr	neurogranin	fibrinogen protein 1	kinase 6	twinfilin-2
p-value	0.41	0.60	0.60	0.93
Mcorr	neurogranin	fibrinogen protein 1	kinase 6	tropomyosin alpha-3
p-value	0.45	0.80	0.80	0.83
Hsic	**neurogranin**	**tropomyosin alpha-3**	kinase 6	tripeptidyl-peptidase 2
p-value	**0.01**	**0.01**	0.09	0.09
Hhg	**neurogranin**	**fibrinogen protein 1**	**tropomyosin alpha-3**	platelet basic protein
p-value	**0.03**	**0.03**	**0.03**	0.11
Mantel	neurogranin	adenylyl cyclase	tropomyosin alpha-3	alpha-actinin-1
p-value	1	1	1	1
Pearson	neurogranin	adenylyl cyclase	tropomyosin alpha-3	alpha-actinin-1
p-value	1	1	1	1
Mic	kinase B	S100-A9	ERF3A	thymidine
p-value	0.15	0.15	0.15	0.15

DOI: https://doi.org/10.7554/eLife.41690.024

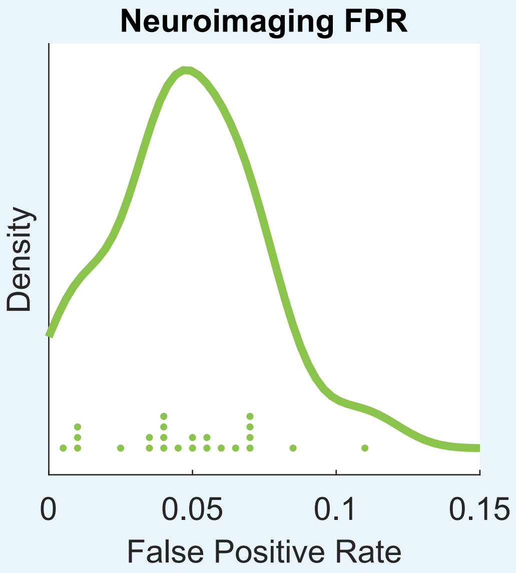
All three methods reveal the same unique protein for pancreas: neurogranin. Hsic identifies another peptide (tropomyosin alpha-3 chain isoform 4), and Hhg identifies a third peptide (fibrinogen-like protein 1 precursor). However, fibrinogen-like protein 1 precursor is not significant for p-value testing between pancreatic and normal subjects. On the other hand, tropomyosin is a ubiquitously expressed protein, since normal tissues and other cancers will also express tropomyosin and leak it into blood, whereas neurogranin is exclusively expressed only in brain tissues. Moreover, there exists strong evidence of tropomyosin 3 upregulated in other cancers (***Karsani et al., 2014***; ***Sun et al., 2016***; ***Lee et al., 2012***; ***Lam et al., 2012***). Therefore, it suggests that the other two peptides identified by Hhg and Hsic are likely false positives.

In fact, neurogranin is always one of the top 4 candidates in all methods except Mic; the only difference is that the corrected p-values are not significant enough for other methods. Along with the classification result in ***Figure 4D*** showing that neurogranin alone has the best classification error, Mgc discovers an ideal candidate for potential biomarker. Moreover, the fact that Mgc, Hhg and Hsic discover the dependency while others cannot implies a nonlinear relationship.

MGC does not inflate false positive rates in screening

In this final experiment, we empirically determine that MGC does not inflate false positive rates via a neuroimaging screening. To do so, we extend the work of Eklund et al. (***Eklund et al., 2012***; ***Eklund et al., 2016***), where a number of parametric methods are shown to largely inflate the false positives. Specifically, we applied MGC to test whether there is any dependency between brain voxel activities and random numbers. For each brain region, MGC attempts to test the following hypothesis: Is activity of a brain region independent of the time-varying stimuli? Any region that is selected as significant is a false positive by construction. By testing each brain region separately, MGC provides a distribution of false positive rates. If MGC is valid, the resulting distribution should be centered around the significance level, which is set at 0.05 for these experiments.

We considered 25 resting state fMRI experiments from the 1000 Functional Connectomes Project consisting of a total of 1583 subjects (***Biswal et al., 2010***). ***Appendix 1—figure 1*** shows the false positive rates of MGC for each dataset, which are centered around the critical level 0.05, as it should be. In contrast, many standard parametric methods for fMRI analysis, such as generalized linear models, can significantly increase the false positive rates, depending on the data and pre-processing details (***Eklund et al., 2012***; ***Eklund et al., 2016***). Moreover, even the proposed solutions to those issues make linearity assumptions, thereby limiting detection to only a small subset of possible dependence functions.



Appendix 1—figure 1. We demonstrate that MGC is a valid test that does not inflate the false positives in screening and variable selection. This figure shows the density estimate for the false positive rates of applying MGC to select the 'falsely significant' brain regions versus independent noise experiments; dots indicate the false positive rate of each experiment. The mean ± standard deviation is 0.0538 ± 0.0394.

DOI: https://doi.org/10.7554/eLife.41690.025

Running time report in experiments

Appendix 1—table 3 lists the actual running time of MGC versus other methods for testing on the real data, based on a modern desktop with a six core I7-6850K CPU and 32 GB memory on MATLAB 2017a on Windows 10. The first two experiments are timed based on 1000 permutations, while the screening experiment is timed without permutation, that is compute the test statistic only. Pearson runs the fastest, trailed by MIC and then DCORR. PEARSON and MIC are only possible to run in the screening experiment, as the other two experiments are multivariate. The running time of MGC is a constant times (about 10) higher than that of Dcorr, and Hhg is implemented in a running time of $O(n^3)$ and thus significantly slower.

Appendix 1—table 3. The actual testing time (in seconds) on real data.

Data	Personality	Creativity	Screening
Mgc	2.5	7.5	1.9
Dcorr	0.2	0.4	0.18
Hsic	0.5	1.7	0.23
Hhg	6.3	53.4	12.3
Pearson	NA	NA	0.03
Mic	NA	NA	0.1
Mrule			

DOI: https://doi.org/10.7554/eLife.41690.026