Discovering Relationships and their Structures Across Disparate Data Modalities

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

[1]Johns Hopkins University, [2]Temple University, [3]Child Mind Institute

Determining how certain properties are related to other properties is fundamental to scientific discovery. As data collection rates accelerate, it is becoming increasingly difficult, yet ever more important, to determine whether one property of data (e.g., cloud density) is related to another (e.g., grass wetness). Only if two properties are related are further investigations into the geometry of the relationship warranted. While existing approaches can test whether two properties are related, they may require unfeasibly large sample sizes in real data scenarios, and do not address how they are related. Our key insight is that one can adaptively restrict the analysis to the "jointly local" observations—that is, one can estimate the scales with the most informative neighbors for determining the existence and geometry of a relationship. "Multiscale Graph Correlation" (MGC) is a framework that extends global procedures to be multiscale; consequently, MGC tests typically require far fewer samples than existing methods for a wide variety of dependence structures and dimensionalities, while maintaining computational efficiency. Moreover, MGC provides a simple and elegant multiscale characterization of the potentially complex latent geometry underlying the relationship. In several real data applications, MGC uniquely detects the presence and reveals the geometry of the relationships.

Identifying the existence of a relationship is the critical initial step in the investigation of any property within a dataset. Only if there is a statistically significant relationship does it make sense to further investigate; such questions arise in high-throughput screening [1], developing imaging biomarkers for diseases [2], causal analyses [3], and machine learning tasks [4]. 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 [5]). This seminal paper prompted the development of entirely new ways of thinking about and quantifying relationships (see [6, 7] 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), and nonlinear relationships (such as nonlinear oscillators), even with very small sample sizes as is common in modern science. Second, we desire methods that provide insight into the geometry of the underlying relationship—rather than merely its existence—to help guide further experimentation and analysis.

While many statistical and machine learning approaches have been developed over the last 120 years to combat the first issue—detecting dependence for any kind of data and relationship—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 [8, 9]. In the 1970s and 1980s, nearest neighbor style approaches were popularized [10, 11], but they were sensitive to algorithm parameters resulting in poor empirical performance. The distance correlation test (DCORR) was recently shown to be able to detect any dependency with sufficient observations [12], at arbitrary dimensions [13], and structured data [14]. Empirically, with a relatively small sample size, DCORR performs well on high-dimensional linear data, whereas two other tests (Heller, Heller, and Gorfine's test (HHG) [15] and a kernel test (HSIC) [16]) perform well on low-dimensional nonlinear data, but no test performs particularly well on high-dimensional nonlinear data, which characterizes a large fraction of real data challenges in the current big data era.

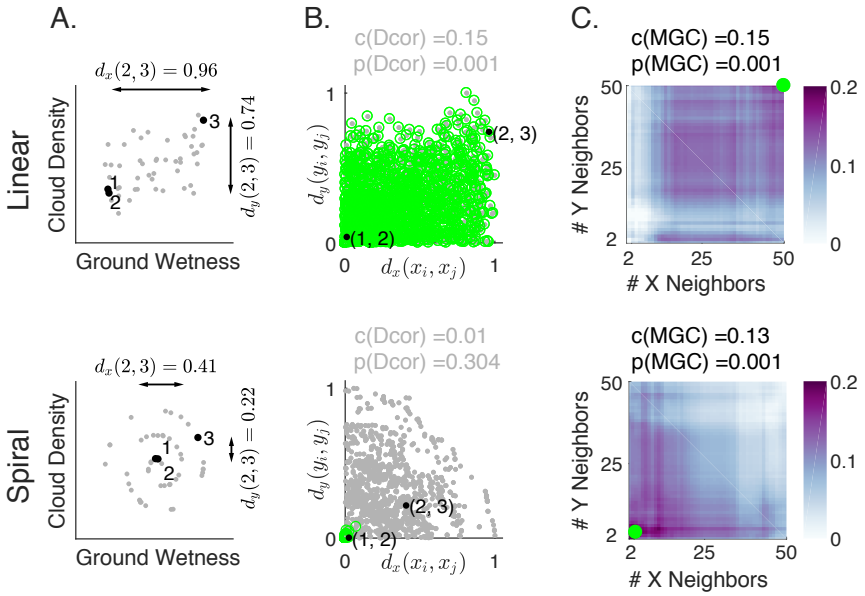

Figure 1: Conceptualization of Multiscale Graph Correlation (MGC) simulating cloud density (x_i) and grass wetness (y_i). We present two different relationships: linear (top) and nonlinear spiral (bottom; see Appendix C for simulation details). (**A**) Scatterplots of the raw data using 50 pairs of samples for each scenario. Samples 1, 2, and 3 (black) are highlighted; arrows show x and y distances between these pairs of points. (**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, and is the scale that MGC selects, resulting 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.

Topological and geometric data analysis has recently emerged as a novel approach to combat the second issue: characterizing the topology and geometry of the relationship [17]. Such methods build multiscale characterization of point cloud data, much like recent developments in harmonic analysis [18]. However, those tools typically operate in unsupervised settings, lack statistical guarantees, and are often quite computationally burdensome.

We surmised that both (i) empirical performance in high-dimensional low-sample size settings, and (ii) quantitative characterization of the geometry of the relationship, could be satisfactorily addressed via extending existing dependence tests to be adaptive to the data [19]. Specifically, existing tests rely on a fixed a priori selection of an algorithmic parameter, such as the kernel bandwidth [20], intrinsic dimension [18], and/or local scale [10, 11]. Indeed, the Achilles Heel of manifold learning has been the requirement to manually choose these parameters [21]. Making these methods adaptive, for example, through post-hoc cross-validation, often adds an unacceptable computational burden, and may weaken or destroy any statistical guarantee. There is, therefore, a need for a statistically valid and computationally efficient adaptive method.

To illustrate the importance of adapting to different kinds of relationships, imagine investigating 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 are prevalent in nature and mathematics, and are canonical in evaluations of manifold learning techniques [22], thereby motivating its use here purely for illustrative purposes.

Under the linear relationship, 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, 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 highlighted in Figure 1A and B). On the other hand, consider the nonlinear (spiral) relationship (bottom plots). 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, every pair of distances is informative with respect to the relationship, while under the nonlinear relationship, only a subset of the distances are—in particular, the "jointly local" distances. By characterizing the strength of dependence at all scales (Figure 1C), one can obtain both an understanding of the geometry underlying the relationship, and determine which distances are sufficiently close to warrant inclusion for assessing overall dependence, thereby improving sensitivity and specificity of the test.

The key, therefore, to successfully determining the presence and geometry of a relationship is to adaptively estimate the number of neighbors that are particularly informative. This is especially important in high-dimensional data, where simple visualizations do not reveal relationships to the unaided human eye. Our methodology—called "Multiscale Graph Correlation" (MGC)—extends essentially all previously proposed pairwise comparison-based approaches to enable estimation of the optimal scales. Crucially, MGC adaptively estimates the informative scales for any relationship—linear or nonlinear, low-dimensional or high-dimensional, unstructured or structured—in a computationally efficient and statistically consistent fashion, therefore effectively guaranteeing equally good or better statistical performance compared to existing global methods in any setting. Moreover, the estimated scales are informative about the geometry of the dependence structure, therefore providing further guidance for subsequent experimental or analytical steps. MGC is thus a hypothesis-testing and geometry-characterizing methodology that builds on recent developments in manifold learning (operating on pairwise comparisons) by combining them with complementary developments in multiscale (topological and/or geometric) analyses. It is this union of these disparate disciplines spanning data science that enables improved theoretical and empirical performance.

The Multiscale Graph Correlation Procedure

MGC is a six step procedure to determine the presence and geometry of dependencies, as follows.

1. Compute two Euclidean distance matrices, one consisting of distances between all pairs of one property (e.g., cloud densities) and the other consisting of distances between all pairs of the other property (e.g., grass wetnesses).

2. Compute the "joint distance matrix" by taking the element-wise product of the centered distance matrices. Centering is required to avoid bias, and is the only difference between MANTEL and DCORR, two previously proposed global methods.

3. Compute all the nearest neighbor graphs for each property. Specifically, for each property, for each point, MGC finds the closest neighbor, and then the second closest neighbor, etc., and similarly for the other property.

4. Estimate the set of all local generalized correlations, that is, the generalized correlation that only includes the k smallest distances for each point in one property, and the l smallest distances for each point in the other property (the "MGC Image" is the matrix of all local generalized correlations).

5. Estimate the optimal local correlation (the MGC test statistic) by finding the smoothed maximum of all local correlations. Smoothing avoids biases and provides MGC with better finite-sample performance and theoretical guarantees (see Appendix A and [23] for details).

6. Determine whether the relationship is significantly dependent via a permutation test. Specifically, MGC permutes the labels of either property many times (typically 1000), and computes a permuted MGC statistic for each, thereby estimating the null distribution of the test statistic, which it then uses to compute the p-value.

We note several clarification points here. The first step of MGC is the same as many non-parametric methods. However, global methods then compute the "generalized correlation", which is simply the correlation between all distances (see Appendix A for details on the global methods). In contrast, MGC computes MGC-Image, which characterizes the geometry of the relationship (Figure 1C). Note how different the MGC Images look for linear versus spiral relationships. Similarly, the optimal scales are quite different for the linear versus spiral relationship: for the linear relationship the optimal scale is global, but for the spiral it is quite local. The green dots in Figure 1B show the set of distances amongst the (k, l) nearest neighbors that MGC selected for these particular simulations. And the green dot in Figure 1C shows MGC's estimated optimal scale. The permutation procedure sidesteps the multiple hypothesis testing problem by only computing one p-value for the MGC test statistic, ensuring that it is a valid test (meaning that the false positive rate is properly controlled at the specified type I error rate; see Appendix B for details). Finally, running MGC is straightforward—it requires inputting n paired samples of two measured properties, or two dissimilarity matrices of size $n \times n$. Our open source implementation[1] requires $O(n^2 \log n)$ time to compute the test statistic, p-value, and MGC-Image, which is about the same running -time complexity as other methods and situates it to be useful in a wide variety of contexts. The following sections document MGC's empirical, computational, and theoretical properties; MGC pseudocodes are provided in Appendix B.

MGC Requires Substantially Fewer Samples to Achieve the Same Power Across Essentially All Dependencies and Dimensions

When, and to what extent, does MGC outperform other approaches, and when does it not? To address these questions, we formally pose the following hypothesis test (see Appendix A for details):

$$H_0 \colon X \text{ and } Y \text{ are independent}$$
$$H_A \colon X \text{ and } Y \text{ are not independent.}$$

The standard criterion for evaluating statistical tests is to compute the probability that it correctly rejects a false null hypothesis, i.e. the testing power at a given type 1 error level. In a complementary

[1] In both MATLAB and R from our website, https://neurodata.io/.

theoretical manuscript [23], we established the theoretical properties of MGC on sample and population level, proved its validity and universal consistency for dependence testing against all distributions of finite second moments, and demonstrated its finite-sample advantage over distance correlation.

Here, we address the empirical performance of MGC as compared with multiple popular tests: (i) DCORR, as discussed above, (ii) MCORR, a modified version of DCORR designed to be unbiased for sample data [13], (iii) HHG, a distance-based test that is very powerful for detecting low-dimensional nonlinear relationships [15]. (iv) HSIC, a kernel-based (and therefore local rather than global) test [16]. (v) MANTEL, which is historically widely used in biology and ecology [24]. (vi) RV coefficient [5, 7], 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 [25–28]. Note that while we focus on high-dimensional settings, Appendix D 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 [29], (x) KENDALL's tau correlation [30], and (xi) MIC [31]. Under the regularity condition that the data distribution has finite second moment, the first four tests are universally consistent (meaning they are theoretically guaranteed to be able to detect any dependency with enough samples), whereas the other tests are consistent only for linear or monotone relationships.

We consider 20 different noisy dependence relationships, mostly taken from the existing literature, including "monotonic" $(1-5)$, "non-monotonic" $(6-19)$, and independent (20) relationships [13, 15, 32–34]. Function details are in Appendix C. The visualization of one-dimensional noise-free (black) and noisy (gray) samples is shown in Supplementary Figure E1; note that the "monotonic" relationships are approximately linear, while "non-monotonic" are strongly nonlinear. For each relationship, we compute the power of each method relative to MGC for a range of dimensions 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 is either the best or close to the best in testing power, for all the simulations and dimensions. Supplementary Figure E2 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 and non-monotone relationships in Table 1. As expected, MGC requires substantially fewer samples than competing methods to achieve the same or higher power. In general, traditional linear correlations (PEARSON, RV, CCA, SPEARMAN, KENDALL) always perform the best in monotonic simulations, the distance-based methods like MCORR, DCORR, MGC; HHG and HSIC are slightly worse, while MIC and MANTEL are the worst. For non-monotone dependencies, traditional correlations fail to detect the existence of dependencies, while MGC is the best approach, followed by HHG and HSIC. In high-dimensional non-monotonic relationships, which motivated the development of MGC, the second best method is MANTEL, and it requires $1.6\times$ as many samples as MGC to achieve the same power. In prior work we proved that MANTEL is not a universally consistent test [23]. The second best test that is universally consistent requires $1.9\times$ as many samples as MGC, demonstrating that MGC could reduce the time and cost of experiments to achieve sufficient power by a factor of two.

As mentioned above, 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 [13]. Supplementary Figure E3 directly compares multiscale generalizations of MANTEL and

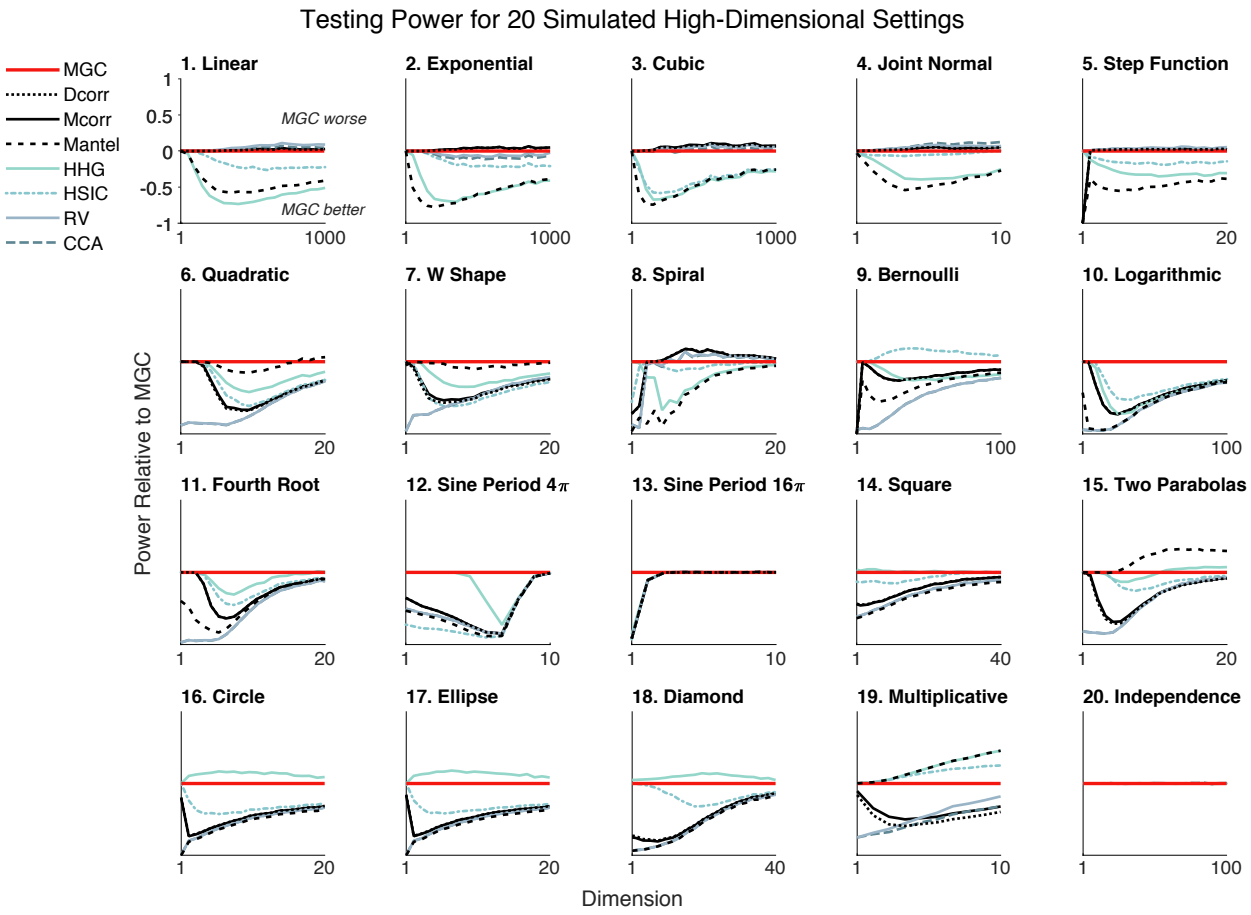

Figure 2: Power of different methods for 20 different dependencies (see Algorithm C2 for details). It includes eight different tests: MGC (solid red), DCORR, MCORR, and MANTEL (black dotted, solid, and dashed lines, respectively), HHG and HSIC (solid and dash-dotted light green, respectively), RV and CCA (solid and dashed gray, respectively). Each panel shows the testing power relative to the power of MGC (e.g., power of MCORR minus the power of MGC) at significance level $\alpha = 0.05$ versus the dimensionality of $\boldsymbol{x}$'s for $n = 100$. Any line below zero at any point indicates that method's power is less than MGC's power for the specified setting and dimensionality. MGC empirically better (or similar) power than all other methods in almost all relationships and all dimensions. For the the independent relationship (#20), all methods yield power 0.05 as they should. Note that MGC is always plotted "on top" of the global variants if there is overlap, therefore, some of the global variants are not always visible from the display.

MCORR as dimension increases, demonstrating that empirically, MGC nearly dominates its global variant for essentially all dimensions and simulation settings considered here. Supplementary Figure E4 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.

Table 1: The median sample size for each method to achieve power 85% at type 1 error level 0.05, grouped into monotone (type 1-5) and non-monotone simulations (type 6-19) for both one- and ten-dimensional settings, normalized by the number of samples required by MGC. In other words, a 2 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 or almost the fewest samples in all settings. Specifically, in the non-monotone and high-dimensional settings, the second best method requires $1.6\times$ the number of samples as MGC.

Dependency Type	Monotone		Non-Monotone	
Dimensionality	1D	10D	1D	10D
MGC	**1**	1	**1**	**1**
DCORR	**1**	1	2.6	3.2
MCORR	**1**	1	2.8	3.1
HHG	1.4	1.7	**1**	1.9
HSIC	1.4	1.7	1.1	2.4
MANTEL	1.4	3	1.8	1.6
PEARSON / RV / CCA	**1**	**0.8**	>10	>6
SPEARMAN / KENDALL	**1**	n/a	>10	n/a
MIC	2.4	n/a	2	n/a

MGC is Sufficiently Computationally Efficient

MGC is able to extend global methods without incurring large costs in computational time. Though a naïve implementation of MGC requires $\mathcal{O}(n^4)$ operations, we have devised a nested parallel implementation that requires only $\mathcal{O}(n^2 \log n/T)$ operations where T is the number of parallel threads (see Algorithm C6 for details). Since T is often larger than $\log n$, in practice, MGC is actually $\mathcal{O}(n^2)$, and a constant factor slower than its global counterpart. For example, at size $n = 5000$ and dimension $p = 1$, MCORR requires around 0.5 seconds to compute the test statistic, whereas MGC requires 5 seconds. The cost and time to obtain $2.5\times$ more data typically far exceeds a few seconds. In 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 Appendix E.V.

MGC Characterizes the Geometry of Dependence

Beyond simply determining whether a relationship exists, the next question is often about the nature or structure of that relationship, to provide insight or guide further experimentation. A single scalar quantity (such as effect size) is inadequate given the vastness and complexities of possible relationships. Other tests would require a secondary procedure to characterize the relationship, which introduces complicated "post selection" statistical quandaries that remain mostly unsolved [35]. Instead, MGC provides a simple, intuitive, and nonparametric (and therefore infinitely flexible) description of the geometry of any relationship.

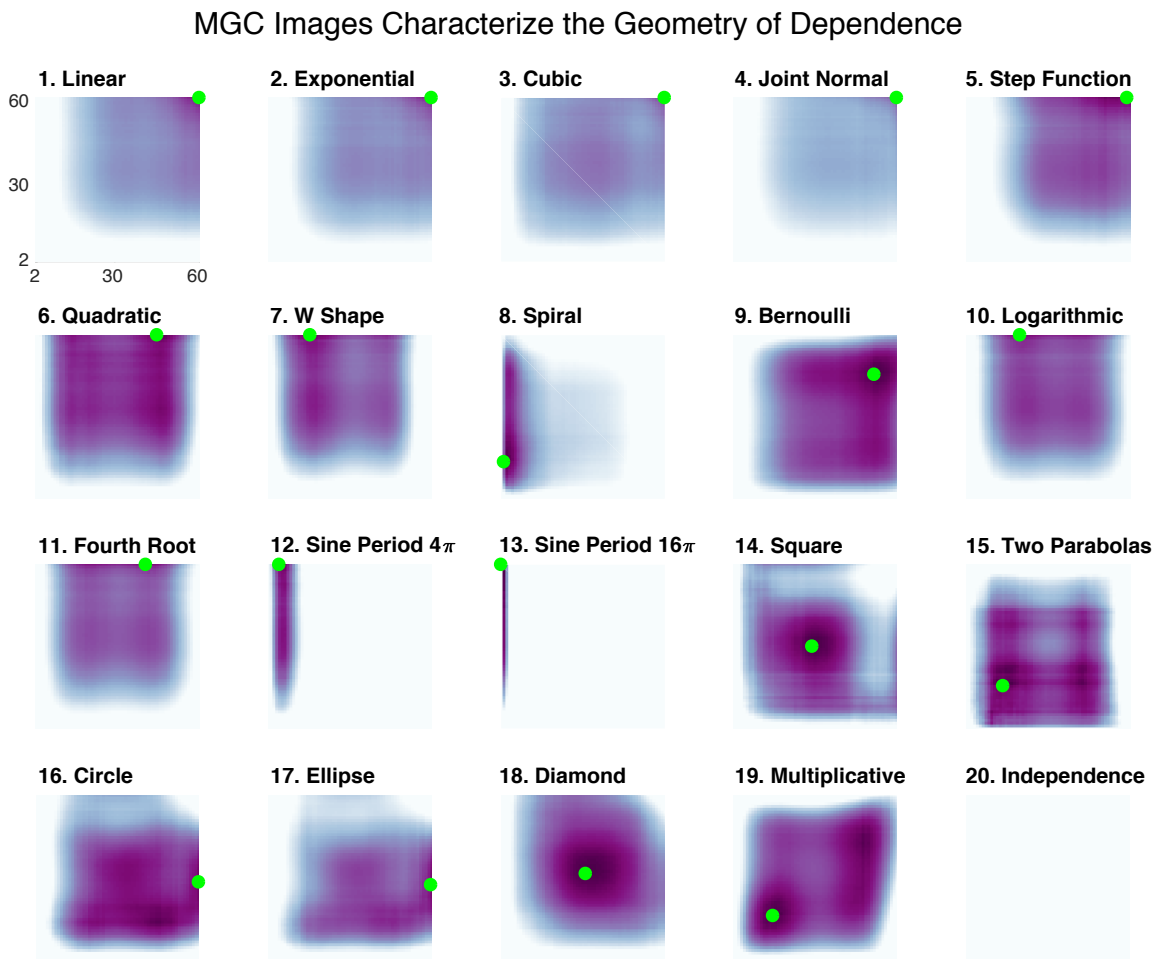

Figure 3: The MGC Image characterizes the geometry of the dependence function. For each of the 20 panels, the abscissa and ordinate denote the number of neighbors for X and Y, respectively, and the color denotes the magnitude of each local correlation. For each simulation, the sample size is 60, and both X and Y are one-dimensional. Each dependency has a different MGC Image 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, i.e., 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 Images 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. A visualization of each dependency is provided in Appendix Figure E1, and the MGC Images for HD simulations are provided in Figure E5.

The MGC Image is an image that shows, for a given dependence relationship, the local correlation as a function of the scales of the two properties. Figure 3 provides the MGC Image for all 20 different one-dimensional relationships; the optimal scales are shown with green dots. For the monotonic dependencies (1-5), the optimal scale is always the largest scale, i.e., 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 Appendix A.VI) :

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

Thus, one can formally use MGC not just to determine whether two properties are related, but also to determine aspects of the geometry of that 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. We know of no other testing procedure that has this property.

Furthermore, similar dependencies have similar MGC Images and often similar optimal scales. For example, logarithmic (10) and fourth root (11), though very different functions analytically, are geometrically similar, and yield very similar MGC Images. Similarly, (12) and (13) are trigonometric functions, and they share a narrow range of significant local generalized correlations. Both circle (16) and ellipse (17), as well as square (14) and diamond (18), are closely related geometrically and also in MGC Images. This indicates that the MGC Image characterizes the geometry of these relationships, differentiating different dependence structures and assisting subsequent analysis steps. Moreover, in [23] we proved that the sample MGC-Image (which MGC estimates) converges to the true MGC-Image provided by the underlying joint distribution of the data. In other words, each relationship has a specific image 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 multi-scale geometric or topological characterizations of data, such as persistence diagrams [17].

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

Here we investigate two particularly interesting properties of the human psyche: personality and creativity. Both have been extensively studied, yielding quantitative metrics for evaluating them using structured interviews [36, 37]. We utilized two previously published datasets to determine whether MGC could yield insight into the relationship between our brains and these mental properties.

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

First, we investigated the relationship between resting-state functional magnetic resonance (rs-fMRI) activity and personality [38] (see Appendix E.I for details). The first row of Table 2 compares the p-value of different methods, and Figure 4A shows the MGC Image 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 Image provides a characterization of the dependence, for which the optimal scale indicates that the dependency is strongly nonlinear. Interestingly, the MGC Image 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 [37] (see Appendix E.II for details). 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 Image in Figure 4B closely resembles a linear relationship where the optimal scale is global. The close-to-linear relationship is also evident 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. 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 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. This experiment demonstrates that for high-dimensional structured data, MGC is able to reveal dependency with relatively small sample size while parametric techniques or estimating a regression function can often be ineffective.

The performance in the real data closely matches the simulations in terms of the superiority of MGC: the first dataset is a strongly nonlinear relationship, for which MGC has the lowest p-value, followed by HHG and HSIC and then all other methods; the second data-set is a close-to-linear relationship, for which global methods often 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 relationship for brain network versus creativity, the only method able to make either of those claims. Finally, we also assessed the frequency with which MGC obtained false positive results using brain activity data, based on experiments from [39, 40]. Supplementary Figure E6 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 for subsequent analysis. We sought to discover new useful protein biomarkers from a quantitative proteomics technique that measures protein and peptide abundance called Selected Reaction Monitoring (SRM) [41] (see Appendix E.III for details). 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.

We obtained a dataset consisting 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 [42].

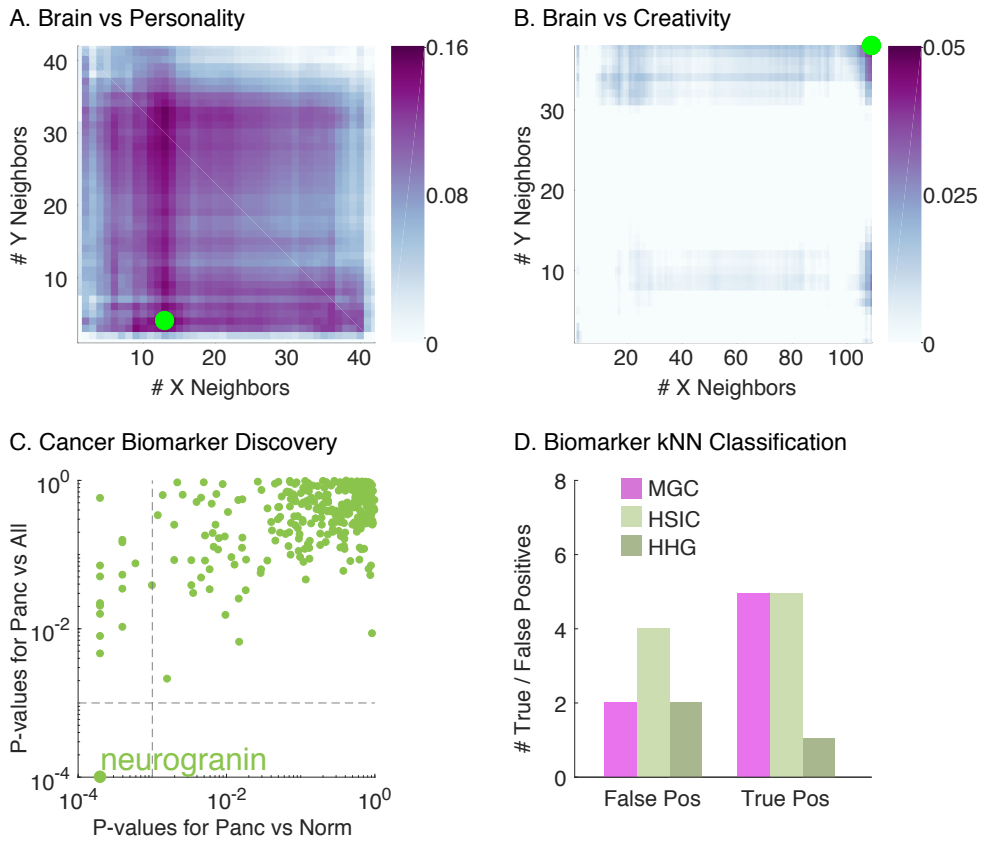

Figure 4: Demonstration that MGC successfully detects dependency, distinguishes linearity from non-linearity, and identifies the most informative feature in a variety of real data experiments. **(A)** The MGC Image for brain activity versus personality. MGC has a large test statistic and a significant 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 Image 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 features identified by each testing method on the peptide data. The peptide identified by MGC achieves the best true and false positive rates, as compared to the peptides identified by HSIC or HHG.

Therefore, our first step was to check whether MGC could correctly identify ovarian biomarkers. Indeed, the pepetides that have been validated previously are also identified by MGC (see Appendix E.III). 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 all other subjects. Figure 4C shows the p-value of each peptide achieved by MGC, which uniquely revealed one particular protein, neurogranin, that exhibited a strong dependency 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. 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. The rest of the global methods did not identify any markers.

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 subset of the peptides selected by HSIC or HHG, in terms of both few 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 generalized correlation coefficients. Regularization is central to high-dimensional and ill-posed problems, where dimensionality is larger than sample size. The connection made here between regularization and dependence testing opens the door towards considering other regularization techniques for correlation-based dependence testing, including HHG. Second, MGC can be thought of informally as learning a metric because it chooses amongst a set of n^2 truncated distances. We could therefore capitalize on recent advances in metric learning [43]. In particular, deep learning can be thought of as metric learning [44], and generative adversarial networks [45] are implicitly testing for equality, which is closely related to dependence. 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 [46], analysis of variance [47], conditional dependence [48, 49], and feature selection [50, 51], so MGC can be adapted for them as well. In fact, MGC can also implement a two-sample (or generally the K-sample) test [52, 53], so further comparisons of MGC to standard methods for two-sample testing will be interesting.

Finally, although energy statistics have not yet been 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 an undesirably strong effect on the performance of different machine learning algorithms [21]. 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 [54] and local lin-

ear embedding [55] for example, must also choose a scale, but have no valid criteria for doing so. MGC could therefore be used to provide insight into multimodal dimensionality reduction as well.

The fact that MGC provides an estimate of the informative scales suggests several theoretical steps to extend this work. First, we could provide further theoretical guidance for choosing the optimal scale in finite samples, which could possibly further improve performance. Second, because the multiscale significance maps provide insight into the geometry of dependence, we could theoretically determine a mapping from these maps to the set of all nonlinear functions to provide a formal characterization of the geometry of the dependency.

MGC also addresses a particularly vexing statistical problem that arises from the fact that methods for two statistical tasks are dissociated from one another: methods for determining whether two properties are related, and methods for determining how they are related. The reason this dissociation creates a problem is that the statistical assumptions underlying the "how related" methods become compromised in the process of determining "whether related": this is the so-called "post-selection inference" problem [35]. 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 determining 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 suggest that we could reduce computational time to close to linear [56], although with some weakening of the theoretical guarantees [57]. Alternately, semi-external memory implementations would allow the running of MGC on any data as long as the interpoint comparison matrix fits on disk rather than main memory [58–61]. Another approach would be to derive an approximation to the asymptotic null distribution for MGC, obviating the need for the permutation test, but at the cost of potential finite-sample bias.

Finally, MGC is easy to use: it merely requires pairs of samples to run, and all the code is available in MATLAB (from our website https://neurodata.io/) and as a R package on Comprehensive R Archive Network (CRAN); code to fully reproduce all the figures in this manuscript is also available from our website. Because MGC is open source and reproducible, and obtains state of the art empirical and theoretical results, enables MGC to be useful in a wide range of applications. We showed its value in diverse applications spanning neuroscience , which motivated this work, and an 'omics example. Other domains, extending beyond science even, to include finance, pharmaceuticals, commerce, and security, face similar questions of dependence and thus could likewise benefit from the methodology proposed here.

Bibliography

[1] Zhang, J. H., Chung, T. D. & Oldenburg, K. R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J. Biomol. Screen. **4**, 67–73 (1999).

[2] Prescott, J. W. Quantitative imaging biomarkers: the application of advanced image processing and analysis to clinical and preclinical decision making. J. Digit. Imaging **26**, 97–108 (2013).

[3] Pearl, J. Causality: Models, Reasoning, and Inference (Cambridge University Press, 2000), first edition edn.

[4] Hastie, T., Tibshirani, R. & Friedman, J. H. Elements of Statistical Learning (Springer, New York, 2001).

[5] Pearson, K. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London **58**, 240–242 (1895).

[6] Reimherr, M. & Nicolae, D. On Quantifying Dependence: A Framework for Developing Interpretable Measures. Statistical Science **28**, 116–130 (2013).

[7] Josse, J. & Holmes, S. Measures of dependence between random vectors and tests of independence. arXiv (2013). URL http://arxiv.org/abs/1307.7383.

[8] Hoeffding, W. A Non-Parametric Test of Independence. Annals of Mathematical Statistics **19**, 546–557 (1948).

[9] Renyi, A. On measures of dependence. Acta Mathematica Academiae Scientiarum Hungarica **10**, 441–451 (1959).

[10] Friedman, J. & Rafsky, L. Graph-Theoretic Measures of Multivariate Association and Prediction. Annals of Statistics **11**, 377–391 (1983).

[11] Schilling, M. Multivariate Two-Sample Tests Based on Nearest Neighbors. Journal of the American Statistical Association **81**, 799–806 (1986).

[12] Szekely, G. & Rizzo, M. Brownian Distance Covariance. Annals of Applied Statistics **3**, 1233–1303 (2009).

[13] Szekely, G. & Rizzo, M. The distance correlation t-test of independence in high dimension. Journal of Multivariate Analysis **117**, 193–213 (2013).

[14] Lyons, R. Distance covariance in metric spaces. Annals of Probability **41**, 3284–3305 (2013).

[15] Heller, R., Heller, Y. & Gorfine, M. A consistent multivariate test of association based on ranks of distances. Biometrika **100**, 503–510 (2013).

[16] Gretton, A. & Gyorfi, L. Consistent Nonparametric Tests of Independence. Journal of Machine Learning Research **11**, 1391–1423 (2010).

[17] Edelsbrunner, H. & Harer, J. L. Computational Topology: An Introduction (American Mathematical Society, 2009), new ed. edition edn.

[18] Allard, W. K., Chen, G. & Maggioni, M. Multi-scale geometric methods for data sets II: Geometric Multi-Resolution Analysis. Applied and Computational Harmonic Analysis **32**, 435–462 (2012). URL http://linkinghub.elsevier.com/retrieve/pii/S1063520311000868.

[19] Zhang, Z., Wang, J. & Zha, H. Adaptive manifold learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**, 253–265 (2012).

[20] Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B. & Smola, A. J. A kernel method for the two-sample-problem. In Advances in neural information processing systems, 513–520 (2006).

[21] Levina, E. & Bickel, P. J. Maximum likelihood estimation of intrinsic dimension. In Advances in Neural Information Processing Systems (2004).

[22] Lee, J. A. & Verleysen, M. Nonlinear dimensionality reduction (Springer Science & Business Media, 2007).

[23] Shen, C., Priebe, C. E. & Vogelstein, J. T. From Distance Correlation to Multiscale Generalized Correlation. submitted (2018).

[24] Mantel, N. The detection of disease clustering and a generalized regression approach. Cancer Research **27**, 209–220 (1967).

[25] Hotelling, H. Relations Between Two Sets of Variates. Biometrika **28**, 321–377 (1936).

[26] Witten, D. M., Tibshirani, R. & Hastie, T. A Penalized Matrix Decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**, 515–534 (2009).

[27] Witten, D. M. & Tibshirani, R. Penalized Classification using Fisher's Linear Discriminant. Journal of the Royal Statistical Society, Series B **73**, 753–772 (2011).

[28] Tenenhaus, A. & Tenenhaus, M. Regularized Generalized Canonical Correlation Analysis. Psychometrika **76**, 257–284 (2011).

[29] Spearman, C. The Proof and Measurement of Association between Two Things. The American Journal of Psychology **15**, 72 (1904). URL http://www.jstor.org/stable/1412159?origin=crossref.

[30] Kendall, M. G. Rank Correlation Methods (London: Griffin, 1970).

[31] Reshef, D. et al. Detecting Novel Associations in Large Data Sets. Science **334**, 1518–1524 (2011).

[32] Szekely, G., Rizzo, M. & Bakirov, N. Measuring and Testing Independence by Correlation of Distances. Annals of Statistics **35**, 2769–2794 (2007).

[33] Simon, N. & Tibshirani, R. COMMENT ON "DETECTING NOVEL ASSOCIATIONS IN LARGE DATA SETS". arXiv (2012). URL http://arxiv.org/abs/1401.7645.

[34] Gorfine, M., Heller, R. & Heller, Y. COMMENT ON "DETECTING NOVEL ASSOCIATIONS IN LARGE DATA SETS". Technical Report (2012). URL http://ie.technion.ac.il/~gorfinm/files/science6.pdf.

[35] Berk, R. et al. Valid post-selection inference. The Annals of Statistics **41**, 802–837 (2013).

[36] Costa, & McCrae, R. R. Neo PI-R professional manual, vol. 396 (1992).

[37] Jung, R. E. et al. Neuroanatomy of creativity. Human Brain Mapping **31**, NA–NA (2009). URL http://doi.wiley.com/10.1002/hbm.20874.

[38] Adelstein, J. et al. Personality Is Reflected in the Brain's Intrinsic Functional Architecture. PLoS ONE **6**, e27633 (2011).

[39] Eklund, A., Andersson, M., Josephson, C., Johannesson, M. & Knutsson, H. Does parametric fMRI analysis with SPM yield valid results?—An empirical study of 1484 rest datasets. NeuroImage **61**, 565–578 (2012).

[40] Eklund, A., Nichols, T. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences **113**, 7900–7905 (2016).

[41] Wang, Q. et al. Mutant proteins as cancer-specific biomarkers. Proceedings of the National Academy of Sciences of the United States of America 2444–9 (2011).

[42] Wang, Q. et al. A Selected Reaction Monitoring Approach for Validating Candidate Biomarkers. In preparation (2017).

[43] Xing, E. P., Ng, A. Y., Jordan, M. I. & Russell, S. Distance metric learning with application to clustering with side-information. Advances in neural information processing systems **15**, 505–512 (2003).

[44] Giryes, R., Sapiro, G. & Bronstein, A. M. Deep neural networks with random gaussian weights: A universal classification strategy. CoRR, abs/1504.08291 (2015).

[45] Goodfellow, I. et al. Generative adversarial nets. In Advances in neural information processing systems, 2672–2680 (2014).

[46] Szekely, G. J. & Rizzo, M. L. A new test for multivariate normality. Journal of Multivariate Analysis **93**, 58–80 (2005).

[47] Rizzo, M. L. & Szekely, G. J. DISCO analysis: A nonparametric extension of analysis of variance. Annals of Applied Statistics **4**, 1034–1055 (2010).

[48] Székely, G. J. & Rizzo, M. L. Partial distance correlation with methods for dissimilarities. The Annals of Statistics **42**, 2382–2412 (2014). URL http://projecteuclid.org/euclid.aos/1413810731.

[49] Wang, X., Pan, W., Hu, W., Tian, Y. & Zhang, H. Conditional Distance Correlation. Journal of the American Statistical Association **110**, 1726–1734 (2015). URL http://www.tandfonline.com/doi/full/10.1080/01621459.2014.993081.

[50] Li, R., Zhong, W. & Zhu, L. Feature Screening via Distance Correlation Learning. Journal of American Statistical Association **107**, 1129–1139 (2012).

[51] Zhong, W. & Zhu, L. An iterative approach to distance correlation-based sure independence screening. Journal of Statistical Computation and Simulation **85**, 2331–2345 (2015). URL http://www.tandfonline.com/doi/full/10.1080/00949655.2014.928820.

[52] Szekely, G. J. & Rizzo, M. L. Testing for Equal Distributions in High Dimension. InterStat **10** (2004).

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

[54] Tenenbaum, J. B., de Silva, V. & Langford, J. C. A Global Geometric Framework for Nonlinear Dimension Reduction. Science **290**, 2319–2323 (2000).

[55] Saul, L. K. & Roweis, S. T. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science **290**, 2323–2326 (2000).

[56] Huo, X. & Szekely, G. Fast Computing for Distance Covariance. Technometrics **58**, 435–447 (2016).

[57] Zhang, Q., Filippi, S., Gretton, A. & Sejdinovic, D. Large-scale kernel methods for independence testing. Statistics and Computing 1–18 (2017).

[58] Zheng, D. et al. FlashGraph: Processing Billion-Node Graphs on an Array of Commodity SSDs. In USENIX Conference on File and Storage Technologies (2015).

[59] Zheng, D., Mhembere, D., Vogelstein, J. T., Priebe, C. E. & Burns, R. FlashMatrix: Parallel, Scalable Data Analysis with Generalized Matrix Operations using Commodity SSDs. arXiv **1604.06414** (2016). URL http://arxiv.org/abs/1604.06414v1.

[60] Zheng, D., Burns, R., Vogelstein, J. T., Priebe, C. E. & Szalay, A. S. An SSD-based eigensolver for spectral analysis on billion-node graphs. arXiv:1602.01421 (2016).

[61] Zheng, D. et al. Semi-External Memory Sparse Matrix Multiplication on Billion-node Graphs in a Multicore Architecture. arXiv (2016). URL http://arxiv.org/abs/1602.02864.

[62] Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. EQUIVALENCE OF DISTANCE-BASED AND RKHS-BASED STATISTICS IN HYPOTHESIS TESTING. Annals of Statistics **41**, 2263–2291 (2013).

[63] Good, P. Permutation, Parametric, and Bootstrap Tests of Hypotheses (Springer, 2005).

[64] Guillot, G. & Rousset, F. Dismantling the Mantel tests. Methods in Ecology and Evolution **4**, 336–344 (2013).

[65] Szekely, G. & Rizzo, M. Partial distance correlation with methods for dissimilarities. Annals of Statistics **42**, 2382–2412 (2014).

[66] Rizzo, M. & Szekely, G. Energy distance. Wiley Interdisciplinary Reviews: Computational Statistics **8**, 27–38 (2016).

[67] Craddock, C. et al. Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC). Frontiers in Neuroinformatics **42** (2015). URL http://www.frontiersin.org/neuroinformatics/10.3389/conf.fninf.2013.09.00042/full.

[68] Koutra, D., Shah, N., Vogelstein, J. T., Gallagher, B. J. & Faloutsos, C. DeltaCon: A Principled Massive-Graph Similarity Function. ACM Transactions on Knowledge Discovery from Data (2015).

[69] Gray Roncal, W. et al. MIGRAINE: MRI Graph Reliability Analysis and Inference for Connectomics. Global Conference on Signal and Information Processing (2013). URL http://arxiv.org/abs/1312.4875.

[70] Sussman, D. L., Tang, M., Fishkind, D. E. & Priebe, C. E. A consistent dot product embedding for stochastic blockmodel graphs. Journal of the American Statistical Association **107**, 1119–1128 (2013). URL http://arxiv.org/abs/1108.2228.

[71] Shen, C., Vogelstein, J. T. & Priebe, C. Manifold Matching using Shortest-Path Distance and Joint Neighborhood Selection. arXiv (2016). URL http://arxiv.org/abs/1412.4098.

[72] Tang, M. et al. A semiparametric two-sample hypothesis testing problem for random graphs. Journal of Computational & Graphical Statistics (2016). URL http://www.tandfonline.com/doi/abs/10.1080/10618600.2016.1193505.

[73] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological) 289–300 (1995).

[74] Karsani, S., Saihen, N., Zain, R., Cheong, S. & Rahman, M. Comparative proteomics analysis of oral cancer cell lines: identification of cancer associated proteins. Proteome Science 3 (2014).

[75] Sun, Y. et al. Facile preparation of salivary extracellular vesicles for cancer proteomics. Scientific Reports 24669 (2016).

[76] Lee, H., Lim, C., Cheong, Y., Singh, M. & Gam, L. Comparison of Protein Expression Profiles of Different Stages of Lymph Nodes Metastasis in Breast Cancer. International Journal of Biological Sciences 353–362 (2012).

[77] Lam, C. Y. et al. Identification and Characterization of Tropomyosin 3 Associated with Granulin-Epithelin Precursor in Human Hepatocellular Carcinoma. PLoS ONE e40324 (2012).

[78] Biswal, B. B. et al. Toward discovery science of human brain function. Proceedings of the National Academy of Sciences **107**, 4734–4739 (2010).

Acknowledgment

This work was partially supported by the Child Mind Institute Endeavor Scientist Program, the National Science Foundation Division of Mathematical Sciences 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, 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, and Dr. Yakir Reshef for insightful suggestions.

A 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 more statistical treatment on MGC is in [23], which introduces the population version of Mgc and various theoretical properties.

A.I 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, i.e., $F_{XY} = F_X F_Y$. 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, i.e., 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, i.e., 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 [31, 53, 56], which are not generalizable to multidimensional data.

A.II Generalized Correlation

Instead of computing on the sample observations directly, most state-of-the-art dependence tests 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. Note that we will use slightly different notations in the appendix: in the main paper a_{ij} and b_{ij} denote the Euclidean distance, while in the appendix they denote the centered distance with $\tilde{a}_{ij}$ and $\tilde{b}_{ij}$ denoting the Euclidean distance.

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" [29, 30] can be written as:

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

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, rather than the individual

data samples. A generalized correlation always has the range $[-1, 1]$, 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 as generalized correlation coefficients, where A and B are derived from sample observations rather than distances. Distance-based methods like DCORR and MANTEL operate on the distance metric, which may be chosen on the basis of domain knowledge, or by default they use the Euclidean distance; 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 [62]), 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, e.g., a generalized correlation, the permutation test is often an effective choice [63], 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 [64].

A.III 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 named the distance variances. The centering scheme is important to guarantee the universal consistency of DCORR, whereas MANTEL uses a simple centering scheme and thus not universal consistent.

Let $c(X, Y)$ be the population distance correlation, that is, the distance correlation between the underlying random variables X and Y. Szekely 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 [14].

It turns out that the sample distance correlation has a finite-sample bias, especially as the dimension p or q increases [13]. 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 (2013) [13, 65, 66] 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, which is the default choice MGC is based on in this paper.

A.IV Local Generalized Correlations

Local generalized correlations can be thought of as further generalizations of generalized correlation coefficients. In particular, given any matrices A and B, we can define a set of local variants of them 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:

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

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 variant of any global generalized correlation coefficient is defined to effectively excludes large distances:

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

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. Thus, c^{kl} is the local sample generalized correlation at a given scale. The MGC Image can be constructed by computing all local generalized correlations, which allows the discovery of the optimal correlation. For any aforementioned generalized correlation (DCORR, MCORR, HSIC, MANTEL, PEARSON), its local generalized correlations can be directly defined by Equation 3, by plugging in the respective a_{ij} and b_{ij} from Equation 1.

A.V 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, i.e., 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 C3). Thus, we can write:

$$c^*(\mathcal{X}_n, \mathcal{Y}_n) = \max_{(k,l) \in R} c^{kl}(\mathcal{X}_n, \mathcal{Y}_n) \tag{4}$$

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

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 [23].

A.VI 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 relation-

ship.

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, i.e., $\|y_i - y_j\| = \|Wx_i - Wx_j\| = \|x_i - x_j\|$. 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 $\texttt{Mgc}(\mathcal{X}_n, \mathcal{Y}_n) = \texttt{Mcorr}(\mathcal{X}_n, \mathcal{Y}_n) = 1$. □

A.VII Computational Complexity

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 C5). 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 C6). 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, e.g., 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 $\mathcal{O}(n^2 \max\{\log n, p, q, r\}/T)$.

B MGC Algorithms and Testing Procedures

Six algorithms are presented in order:

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

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 MGC((x_i, y_i), for $i \in [n]$)
(1) Calculate all pairwise distances:
for $i, j := 1, \ldots, n$ **do**
$a_{ij} = \delta_x(x_i, x_j)$ ▷ δ_x is the distance between pairs of x samples
$b_{ij} = \delta_y(y_i, y_j)$ ▷ δ_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 C3
(3) Calculate the p-value
$pval(c^*) = \text{PermutationTest}(A, B, r, c^*)$ ▷ Algorithm C4
end function

Pseudocode C2 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 C3. 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, α)
  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
    $c_0^*[t] =$ MGCSampleStat(A^0, B^0) ▷ MGC statistic under the null
    $c_1^*[t] =$ MGCSampleStat(A^1, B^1) ▷ MGC statistic under the alternative
  end for
  $\omega_\alpha \leftarrow$ 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

Pseudocode C3 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, i.e., 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 $(k, l) \in \mathbb{N} \times \mathbb{N}$.

function MGCSampleStat(A, B)
  $\mathcal{C} = \text{MGCAllLocal}(A, B)$ ▷ All local correlations
  $\tau = \text{Thresholding}(\mathcal{C})$ ▷ 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
  $\mathcal{R} = \text{Connected}(\mathcal{R})$ ▷ 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 C4 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**
    $\pi = \text{RandPerm}(n)$ ▷ generate a random permutation of size n
    $c_0^*[t] = \text{MGCSampleStat}(A, B(\pi, \pi))$ ▷ 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 C5 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** $\mathcal{E}^Z = \text{Sort}(Z)$ **end for** ▷ parallelized sorting
  for $Z := A, B$ **do** $Z = \text{Center}(Z)$ **end for** ▷ center distance matrices
  $\tilde{c}^{kl} \leftarrow tr((A \circ \mathcal{E}^A)^\intercal \times (B \circ (\mathcal{E}^B)^\intercal))$ ▷ un-normalized local distance covariance
  $v^A \leftarrow tr((A \circ \mathcal{E}^A)^\intercal \times (A \circ (\mathcal{E}^A)^\intercal))$ ▷ local distance variances
  $v^B \leftarrow tr((B \circ \mathcal{E}^B)^\intercal \times (B \circ (\mathcal{E}^B)^\intercal))$
  $e^A \leftarrow \sum_{i,j=1}^n (A \circ \mathcal{E}^A)_{ij}$ ▷ sample means
  $e^B \leftarrow \sum_{i,j=1}^n (B \circ \mathcal{E}^B)_{ij}$
  $c^{kl} \leftarrow \left(\tilde{c}^{kl} - e^A e^B / n^2\right) / \sqrt{(v^A - (e^A/n)^2)\,(v^B - (e^B/n)^2)}$ ▷ center and normalize
end function

Pseudocode C6 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 $k \leq R(A_{\cdot j}, i)$ and $l \leq R(B_{\cdot j}, i)$, 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** $\mathcal{E}^Z = \text{Sort}(Z)$ **end for**
  for $Z := A, B$ **do** $Z = \text{Center}(Z)$ **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**
    $c^{kl} \leftarrow \left(\tilde{c}^{kl} - e^A_k e^B_l/n^2\right) / \sqrt{\left(v^A_k - e^{A2}_k/n^2\right)\left(v^B_l - e^{B2}_l/n^2\right)}$
  end for
end function

C 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 [32–34]. 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 $w^\mathsf{T}\boldsymbol{x}$ 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 of 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 $F_{Y|\boldsymbol{X}}$ and F_X, as well as any additional auxiliary distributions.

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

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

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

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

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

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

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

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

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

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

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

7. W Shape $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: $U \sim \mathcal{U}(-1, 1)^p$,
$$\begin{aligned} X &\sim \mathcal{U}(-1, 1)^p, \\ Y &= 4\left[\left((w^\mathsf{T} X)^2 - \tfrac{1}{2}\right)^2 + w^\mathsf{T} U/500\right] + 0.5\kappa\epsilon. \end{aligned}$$
8. Spiral $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: $U \sim \mathcal{U}(0, 5)$, $\epsilon \sim \mathcal{N}(0, 1)$,
$$\begin{aligned} X_{[d]} &= U \sin(\pi U) \cos^d(\pi U) \text{ for } d = 1, \ldots, p-1, \\ X_{[p]} &= U \cos^p(\pi U), \\ Y &= U \sin(\pi U) + 0.4p\epsilon. \end{aligned}$$
9. Uncorrelated Bernoulli $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: $U \sim \mathcal{B}(0.5)$, $\epsilon_1 \sim \mathcal{N}(0, I_p)$, $\epsilon_2 \sim \mathcal{N}(0, 1)$,
$$\begin{aligned} X &\sim \mathcal{B}(0.5)^p + 0.5\epsilon_1, \\ Y &= (2U - 1)w^\mathsf{T} X + 0.5\epsilon_2. \end{aligned}$$
10. Logarithmic $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: $\epsilon \sim \mathcal{N}(0, I_p)$
$$\begin{aligned} X &\sim \mathcal{N}(0, I_p), \\ Y_{[d]} &= 2\log_2(|X_{[d]}|) + 3\kappa\epsilon_{[d]} \text{ for } d = 1, \ldots, p. \end{aligned}$$
11. Fourth Root $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:
$$\begin{aligned} X &\sim \mathcal{U}(-1, 1)^p, \\ Y &= |w^\mathsf{T} X|^{\frac{1}{4}} + \frac{\kappa}{4}\epsilon. \end{aligned}$$
12. Sine Period 4π $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: $U \sim \mathcal{U}(-1, 1)$, $V \sim \mathcal{N}(0, 1)^p$, $\theta = 4\pi$,
$$\begin{aligned} X_{[d]} &= U + 0.02pV_{[d]} \text{ for } d = 1, \ldots, p, \\ Y &= \sin(\theta X) + \kappa\epsilon. \end{aligned}$$
13. Sine Period 16π $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: 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
$$\begin{aligned} X_{[d]} &= U \cos\theta + V \sin\theta + 0.05p\epsilon_{[d]}, \\ Y_{[d]} &= -U \sin\theta + V \cos\theta, \end{aligned}$$
for $d = 1, \ldots, p$.
15. Two Parabolas $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: $\epsilon \sim \mathcal{U}(0, 1)$, $U \sim \mathcal{B}(0.5)$,
$$\begin{aligned} X &\sim \mathcal{U}(-1, 1)^p, \\ Y &= \left((w^\mathsf{T} X)^2 + 2\kappa\epsilon\right) \cdot (U - \tfrac{1}{2}). \end{aligned}$$
16. Circle $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: $U \sim \mathcal{U}(-1, 1)^p$, $\epsilon \sim \mathcal{N}(0, I_p)$, $r = 1$,
$$\begin{aligned} X_{[d]} &= r\left(\sin(\pi U_{[d+1]}) \prod_{j=1}^{d} \cos(\pi U_{[j]}) + 0.4\epsilon_{[d]}\right) \text{ for } d = 1, \ldots, p-1, \\ X_{[p]} &= r\left(\prod_{j=1}^{p} \cos(\pi U_{[j]}) + 0.4\epsilon_{[p]}\right), \\ Y &= \sin(\pi U_{[1]}). \end{aligned}$$

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 $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^p \times \mathbb{R}^p$: $u \sim \mathcal{N}(0, I_p)$,

$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}(0, I_p), \\ \boldsymbol{y}_{[d]} &= u_{[d]} \boldsymbol{x}_{[d]} \text{ for } d = 1, \ldots, p. \end{aligned}$$

20. Multimodal Independence $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: 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

$$\begin{aligned} X &= U/3 + 2U' - 1, \\ Y &= V/3 + 2V' - 1. \end{aligned}$$

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 E1.

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 $n = 100$.

D Supplementary Figures

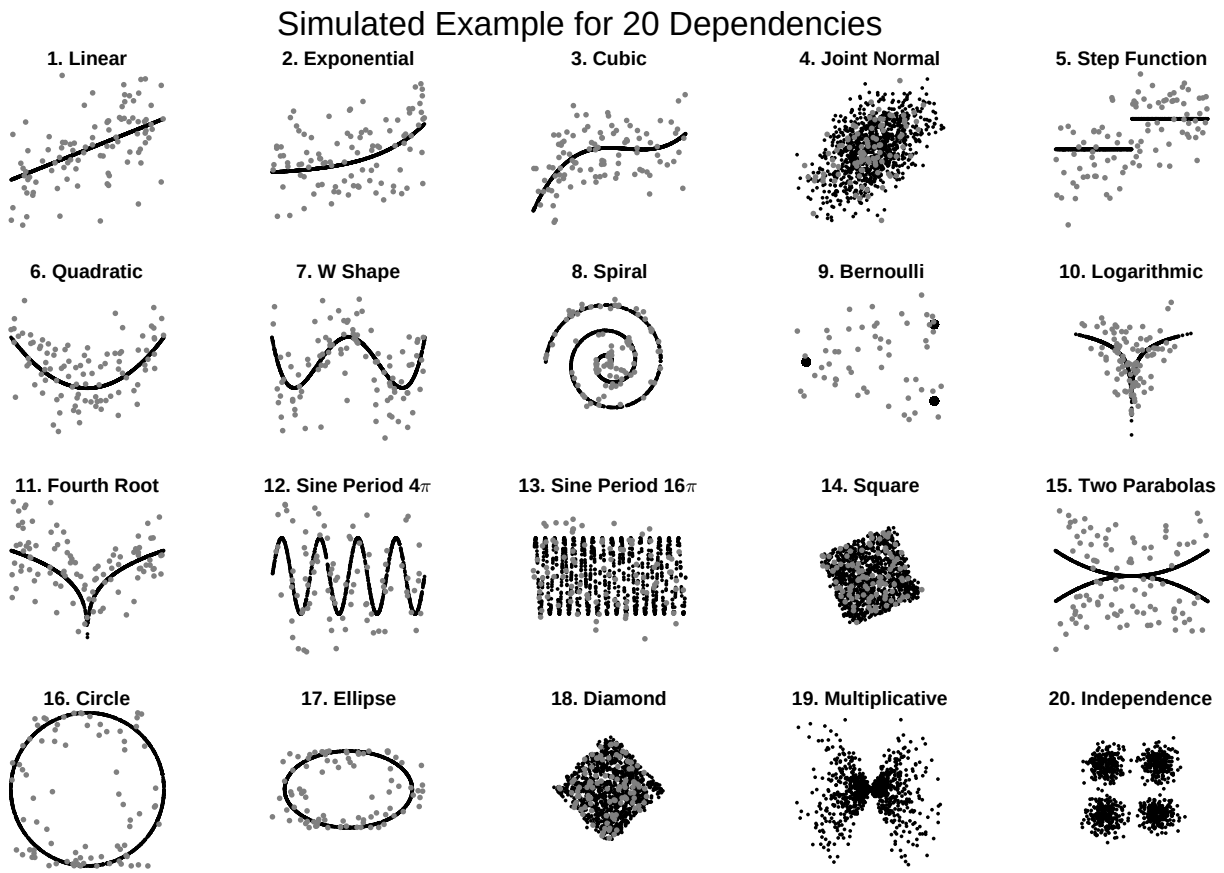

Figure E1: Visualization of the 20 dependencies at $p = q = 1$. For each, $n = 100$ points are sampled with noise ($\kappa = 1$) to show the actual sample data used for one-dimensional relationships (gray dots). For comparison purposes, $n = 1000$ points are sampled without noise ($\kappa = 0$) to highlight each underlying dependency (black dots). Note that only black points are plotted for type 19 and 20, as they do not have the noise parameter κ.

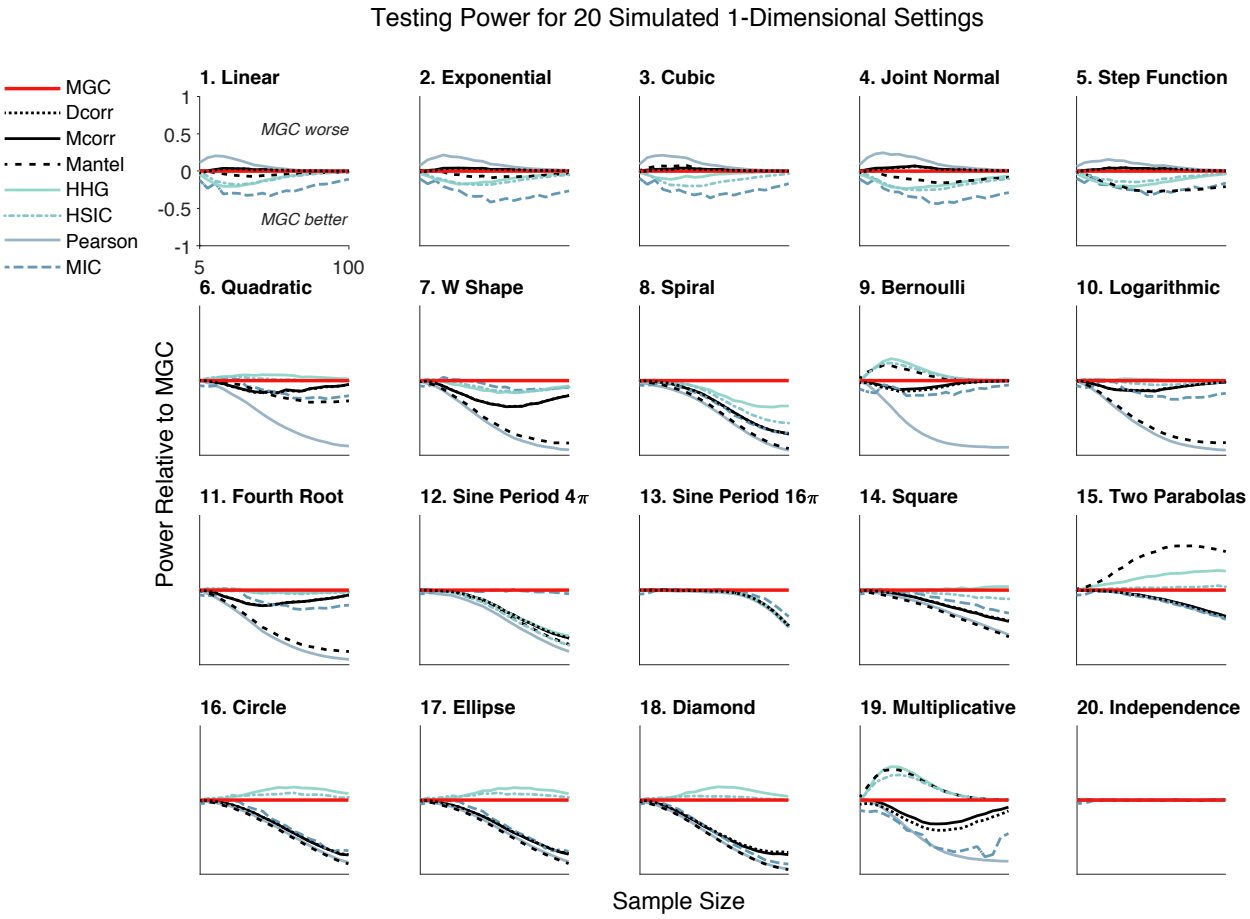

Figure E2: The same power plots as in Figure 2, except the 20 dependencies are one-dimensional with noise, and the x-axis shows sample size increasing from 5 to 100. Again, `Mgc` empirically achieves similar or better power than the previous state-of-the-art approaches on most problems. Note that `Mic` is included in 1D case; `RV` and `Cca` both equal `Pearson` in 1D; `Kendall` and `Spearman` are too similar to `Pearson` in power and thus omitted in plotting.

Figure E3: The same set-ups as in Figure 2, comparing different MGC implementations versus its global counterparts. The default MGC builds upon MCORR throughout the paper, and we further consider MGC on MANTEL to illustrate the generalization. The magenta line shows the power difference between MCORR and MGC , and the cyan line shows the power difference between MANTEL and the MGC version of MANTEL. Indeed, MGC is able to improve the global counterpart in testing power under nonlinear dependencies, and maintains similar power under linear and independent dependencies.

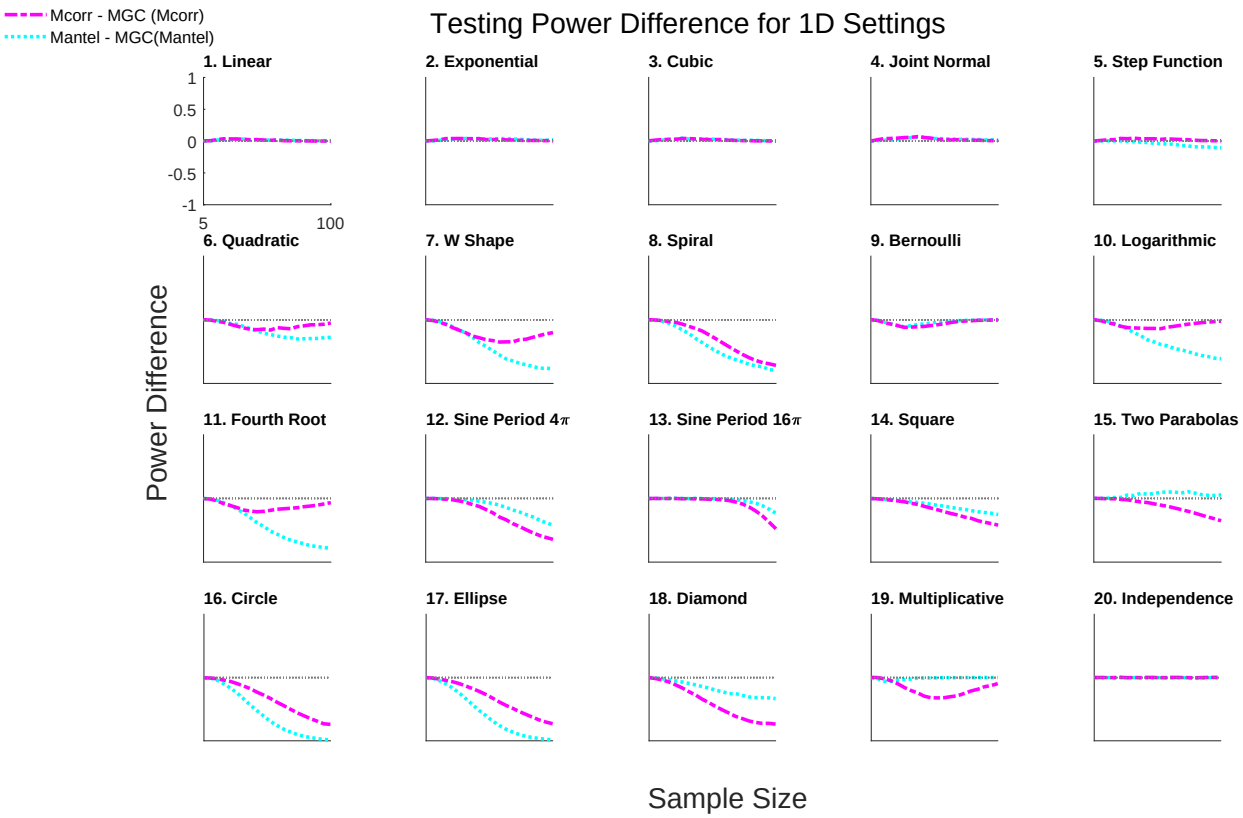

Figure E4: The same power plots as in Figure E3, except the 20 dependencies are one-dimensional with noise, and the x-axis shows sample size increasing from 5 to 100.

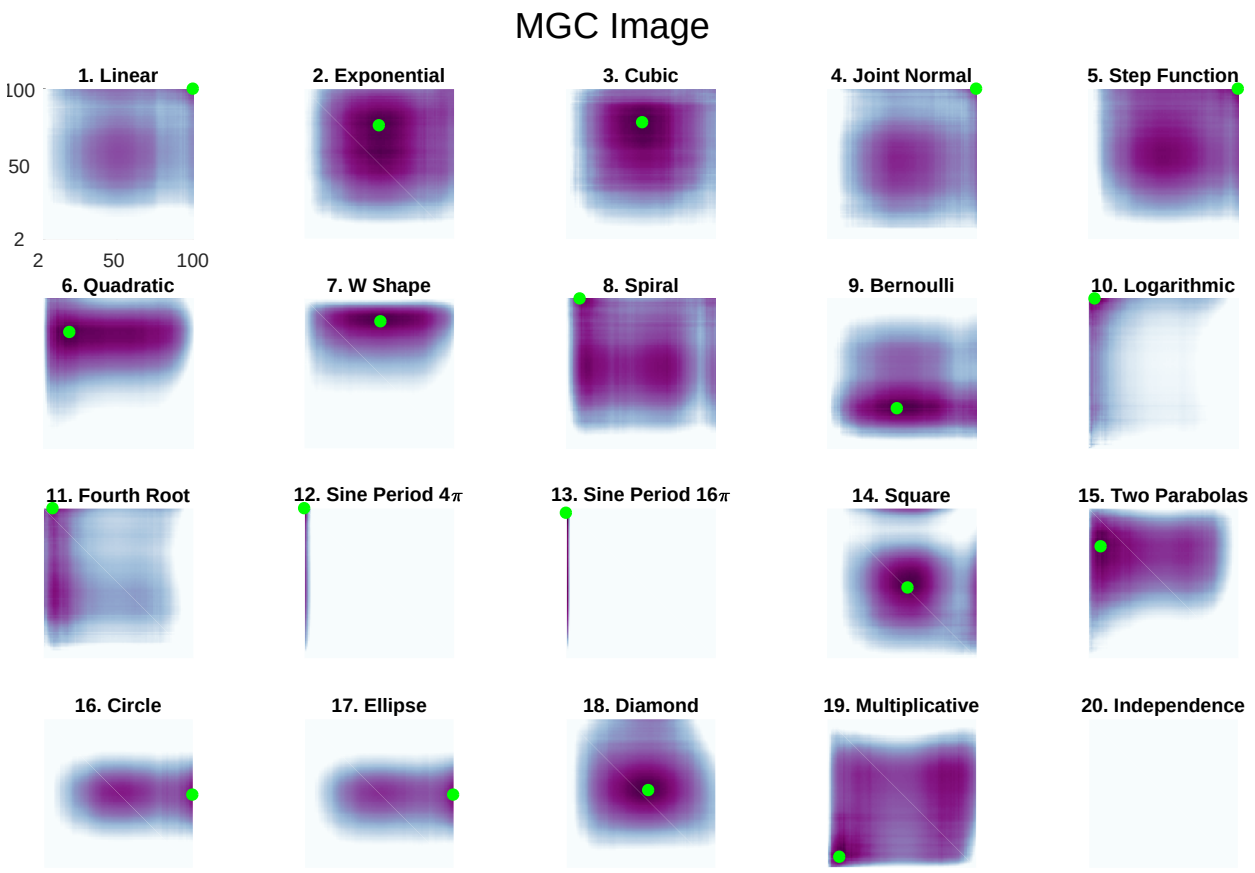

Figure E5: The MGC Image for the 20 panels for high-dimensional dependencies. For each simulation, the sample size is 100, and the dimension is selected as the dimension such that MGC has a testing power above 0.5. It has similar behavior and interpretation as the 1-dimensional power maps in Figure 3, i.e., the linear relationships optimal scales are global, and similar dependencies share similar MGC Images.

E Real Data Processing

E.I 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 [36]. 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. [38] 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 [67] 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.

E.II 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 [37]. 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 [37]. Previously published results explored the relationship between graphs and CCI [68], 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 [69]. We compute the distance between the graphs using the semi-parametric graph test statistic [70–72], embedding each graph into two dimensions and aligning the embeddings via a Procrustes analysis.

E.III 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 [41]. 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 [42]. 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 [73] to control the false discovery rate. A summary of the results are reported in Table 3.

Table 3: 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	Sample 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

All methods are able to successfully detect a dependence between peptide abundances in ovarian cancer samples versus normal samples (Table 3, line 1). This is likely because there are so many individual peptides that have different abundance distributions between ovarian and normal samples (Table 3, 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 ≈ 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 of the five methods found a dependence at the 0.05 level, with Sample MGC obtaining the lowest p-value (Table 3, 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 (Table 3, line 4). Because we were interested in identifying peptides that were uniquely useful for pancreatic cancer, we then compared pancreatic samples to all others. Only MGC, HSIC, and HHG identified peptides that expressed different abundances in this more challenging case (Table 3, line 5). To identify peptides that are unique to pancreatic cancer, we looked at the set of peptides that were both different from normals and different from all non-pancreatic cancer samples (Table 3, line 6). All

three method 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 [74–77]. Therefore, initial literature search suggests that tropomyosin is likely falsely identified by HHG and less useful as a pancreatic cancer marker, meaning that only MGC identified putative pancreatic cancer biomarkers without also identifying likely false positives.

Furthermore, although neurogranin is not identified by other methods, it is always the most dependent peptide in all methods except MIC. Namely, all of PEARSON, DCORR, MCORR, MANTEL, HHG, HSIC, and MGC rank neurogranin as the most significant protein by p-value; the only difference is that the p-values are not significant enough for other methods after multiple testing adjustment. Also, the three peptides identified by HHG are also the top three in MGC; and if we further investigate the top three peptides in all methods, they always come from these three peptides, and another peptide (mitogen-activated protein kinase); the only exception is MIC, whose top three peptides do not coincide with all other correlation measures, which suggests it may detect too many false positives. Along with the classification result showing that neurogranin along has the best classification error, this experiment strongly indicate that MGC, HSIC, HHG are the top methods in dependency testing, able to amplify the signal, and do not detect false signals.

E.IV 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. [39, 40], 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 $1{,}000$ Functional Connectomes Project consisting of a total of $1{,}583$ subjects [78]. Figure E6 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 [39, 40]. Moreover, even the proposed solutions to those issues make linearity assumptions, thereby limiting detection to only a small subset of possible dependence functions.

E.V Running Time Report

Here we list 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 32GB 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, i.e., compute the test statistic only. PEARSON runs the fastest, trailed by

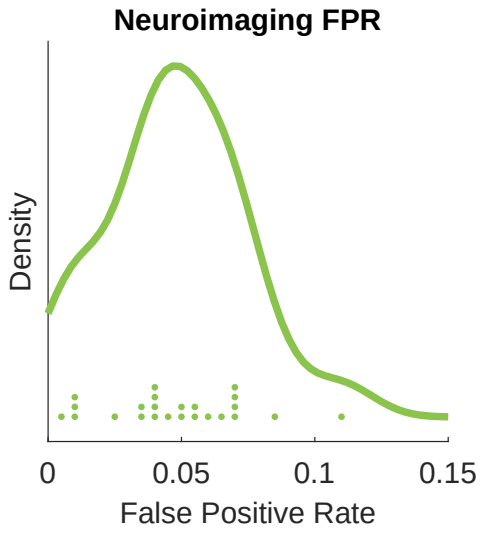

Figure E6: We demonstrate that MGC is a valid test that does not inflate the false positives in screening and variable selection. This figure shows the density estimate for the false positive rates of applying MGC to select the "falsely significant" brain regions versus independent noise experiments; dots indicate the false positive rate of each experiment. The mean $\pm$ standard deviation is 0.0538 ± 0.0394.

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.

Table 4: 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