Nonlinear Association Criterion, Nonlinear Granger Causality and Related Issues with Applications to Neuroimage Studies

Chenyang Tao[a,b], Jianfeng Feng[a,b,c,*]

[a] *Centre for Computational Systems Biology and School of Mathematical Sciences, Fudan University, Shanghai 200433, PR China*
[b] *Department of Computer Science, Warwick University, Coventry CV4 7AL, UK*
[c] *School of Life Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai 200433, PR China*

Abstract

Background

Quantifying associations in neuroscience (and many other scientific disciplines) is often challenged by high-dimensionality, nonlinearity and noisy observations. Many classic methods have either poor power or poor scalability on data sets of the same or different scales such as genetical, physiological and image data.

New Method

Based on the framework of reproducing kernel Hilbert spaces we proposed a new nonlinear association criteria (NAC) with an efficient numerical algorithm and p-value approximation scheme. We also presented mathematical justification that links the proposed method to related methods such as kernel generalized variance, kernel canonical correlation analysis and Hilbert-Schmidt independence criteria. NAC allows the detection of association between arbitrary input domain as long as a characteristic kernel is defined. A MATLAB package was provided to facilitate applications.

Results

Extensive simulation examples and four real world neuroscience examples

*Corresponding author
Email address: jianfeng64@gmail.com (Jianfeng Feng)

including functional MRI causality, Calcium imaging and imaging genetic studies on autism [Brain, 138(5):13821393 (2015)] and alcohol addiction [PNAS, 112(30):E4085–E4093 (2015)] are used to benchmark NAC. It demonstrates the superior performance over the existing procedures we tested and also yields biologically significant results for the real world examples.

Comparison with Existing Method(s)

NAC beats its linear counterparts when nonlinearity is presented in the data. It also shows more robustness against different experimental setups compared with its nonlinear counterparts.

Conclusions

In this work we presented a new and robust statistical approach NAC for measuring associations. It could serve as an interesting alternative to the existing methods for datasets where nonlinearity and other confounding factors are present.

Keywords: nonlinear association, canonical correlation analysis, reproducing kernel Hilbert space, Granger causality, regularization, variable selection, permutation, parametric approximation of p-value

---

1. Introduction

Searching for association and causality between systems has long been the primary objective for many research fields ranging from engineering to biology (Hlaváčková-Schindler et al., 2007; Buchanan, 2012; Sugihara et al., 2012; Reshef et al., 2011). Common choices for univariate association measures include classical univariate tests statistics such as Pearson's ρ, Spearman's ρ (rank correlation) and Kendall's τ (McDonald, 2009). In the multivariate case, maximum correlation techniques such as canonical correlation analysis (CCA) (Richard and Dean, 2002) and partial least square (McIntosh and Lobaugh, 2004) are the most common. But they all depend on strong assumptions such as linearity, and are only capable of capturing specific modes of dependency between two systems.

The human brain, regarded by many as the most complex organ, exhibits many nontrivial and significantly nonlinear relationships involving many levels. Three examples illustrate this.

1. A visual input on the retina involves many complex and possibly nonlinear transforms that lead to the corresponding output in the inferior temporal cortex (Kendrick et al., 2011). The result is a level of input recognition that outperforms vision recognition algorithms running on modern von Neumann computers.
2. A gene, such as DISC1, affects different pathways to cause schizophrenia, via complex, largely unknown nonlinear process Gong et al. (2014). A simple linear association might be able to catch *some* of the information obtained in a genome wide association study, but almost certainly not all of it.
3. At the neuronal network level, neurons receive and emit spikes as is characterized by classical Hodgkin-Huxley type models (Keener and Sneyd, 2010). These models are intrinsically nonlinear, reflecting the underlying dynamical behavior of excitable cells.

These examples motivate the need to quantify a nonlinear association between two systems (sets of random variables) in brain research and other subjects. As a matter of fact, many recent successes of applying nonlinear methods in neuroimaging studies (Ge et al., 2012, 2015a; Stephan et al., 2008; Viviani et al., 2005), especially for the support vector machines (Cuingnet et al., 2011), further justify such need. Other than this, the association measure should also be able to deal with high dimensional input and robust to low signal to noise ratio (SNR), which are the characterizing properties of the data in neuroimaging studies.

Following Shannon's seminal paper on information theory Shannon (1948), measures such as mutual information (MI), that quantify shared information, are the best known general purpose measure of associations between random variables. Although this has been found to be successful in low dimensional applications (Reshef et al., 2011), generalization to higher dimensions is non-

trivial; this normally requires estimation of a joint density or involves nearest neighbor search, which by themselves are challenging tasks (Li and Racine, 2007; Kraskov et al., 2004). Characteristic function based tests have also been proposed (Feuerverger, 1993; Kankainen, 1995), but they have only been used to test the association between univariate variables. In recent years, distance based correlation measures have emerged as an elegant solution for testing association between high dimensional variables Székely et al. (2009). These, however, turned out to be a special case of a kernel independence test, by choosing an appropriate kernel function Gretton et al. (2009); Sejdinovic et al. (2013). Yet another common testing procedure for detecting association in high dimensional datasets (especially GWAS studies) is called the least square kernel machine (LSKM) Liu et al. (2007). This is also known to be equivalent to a kernel independence test (Hua and Ghosh, 2014). Currently there are two kernel approaches for measuring independence, both are related to the covariance operator in Reproducing Kernel Hilbert Space but derived from different heuristics. One is known as the Hilbert-Schmidt independence criterion (HSIC) and the other is the kernel generalized variance (KGV). The asymptotic distribution of the HSIC has been derived Gretton et al. (2005, 2007) and its normalized variant, called the normalized cross-covariance operator (NOCCO), has been proposed in Fukumizu et al. (2007b) without knowing the asymptotic distribution. The KGV has found numerous successful applications, but mostly as a metric score that quantifies independence (Liu et al., 2013; Bach and Jordan, 2002). Other closely related work includes (Dauxois and Nkiet, 1998) and its kernel extension (Huang et al., 2006).

In this work we propose an approach based on a nonlinear association criterion (NAC) which is similar to the kernel generalized variance, with the key difference that our approach only needs a partial eigendecomposition and allows us to evaluate the significance of the NAC estimate, which the original kernel generalized variance approach lacks. We also endeavored to make the computational intensive nonlinear test affordable for large scale neuroimaging investigations by proposing a $O(N)$ time complexity algorithm, which has a

smaller constant factor and is much faster than other KCCA-based counterparts we are aware of. Some theoretical justifications are provided to show how an NAC is related to an HSIC.

As a specific application, we show how NAC can be used to investigate causal relationships. Causality, originally considered by economists, has been intensively studied in recent years in neuroimaging and other fields in neuroscience, despite applicability to fMRI data remaining controversial, due to the low temporal resolution that is inherent in the data acquisition. We note that incorporating NAC into the classical Granger causality gives rise to a new procedure that allows the detection of more general forms of information flow.

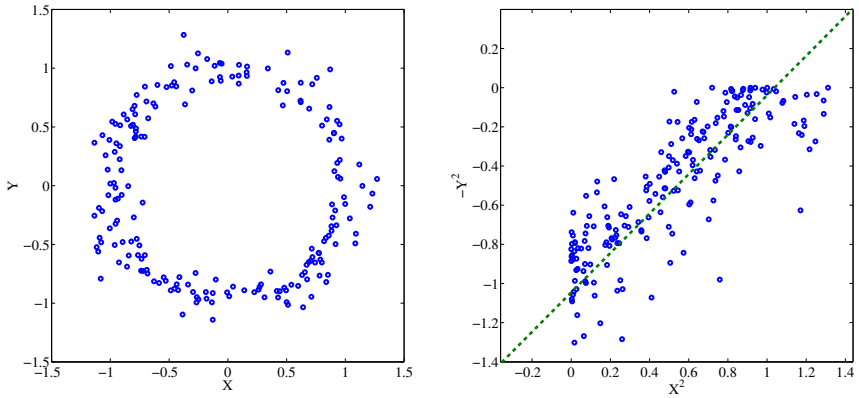
This paper is organized as follows: in Section 2 we begin with a short review of the concepts regarding kernel independence measures followed by an intuitive introduction of an NAC. We then go into the details of NAC, showing how to evaluate the statistical significance of an empirical estimate and maximize its power. In Section 3 the proposed NAC is compared with other association measures using extensive simulations. In Section 4 we then benchmark the performance of NAC using a few real-world examples, to illustrate its advantages over classical approaches. The examples are from resting state fMRI, synthetic calcium waves and imaging genetics. We conclude the paper with a discussion. A MATLAB toolbox of NAC and a few other nonlinear association measures is available online from `http://www.dcs.warwick.ac.uk/~feng/nac/`.

2. Methods

2.1. Basics of the proposed method

We begin with a motivating example where two variables $X = \{x_i\}_{i=1}^N$ and $Y = \{y_i\}_{i=1}^N$ are statistically related but the standard *linear* measure of correlations $\mathrm{corr}(X, Y)$[1] fails. This example also serves to explain the basic idea of

[1] $\mathrm{corr}(\{x_i\}, \{y_i\}) = \frac{1}{N}\sum_{i=1}^N \tilde{x}_i \tilde{y}_i$, where $\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$, $\tilde{y}_i = \frac{y_i - \bar{y}}{\sqrt{\sum_i (y_i - \bar{y})^2}}$, $\bar{x} = \frac{1}{N}\sum_i x_i$ and $\bar{y} = \frac{1}{N}\sum_i y_i$



(a) (b)

Figure 1: (a) Cloud points scattered along a unit circle. (b) Same points with proper nonlinear transformation applied.

the general method proposed in this work. Assume that we have collected N pairs of observations $\{(x_i, y_i)\}_{i=1}^N$ which are uniformly scattered along a unit circle, see Figure 1(a). The aim is to determine the extent to which x and y are correlated. It is obvious, from a two-dimensional plot of the (x_i, y_i) that their correlation cannot be detected via measures such as Pearson's ρ statistic (i.e., corr$(\{x_i\}, \{y_i\})$), since for each pair of observations, (x_i, y_i), there are other pairs of observations, in the vicinity of $(-x_i, y_i)$, $(x_i, -y_i)$, $(-x_i, -y_i)$, which lead to a near zero correlation. In order to resolve this issue, we shall adopt an alternative approach, which involves transforming the observations so the correlation can be properly detected. In this case the transformation on the observations is to work with new variables, for example $(x_i^2, -y_i^2)$, which leads to a correlation close to unity, see Figure 1(b). This is an explicitly nonlinear transformation. Of course, real data could be far more complicated than the simple circle used in this illustration. For example, in neuro-imaging, X could represent a set of DTI summary statistics({MD, FA, ...}), brain tissue makeup ({WM, GM, CSF, ...}), connectivity pattern or activation profile of a set of related tasks for a particular voxel, while Y could represent a set of behavior

evaluations or a set of candidate SNPs. Generally, a reasonable association measure should deal with high dimensional inputs and naturally transform the data points, to achieve a best linear association. We show below how to construct such an estimator.

The idea of transforming the data, to achieve maximal linear correlation, and use it to measure its nonlinear association, has existed for a long time. The first nonlinear association measure that adopts this idea is known as the Hirschfeld-Gebelein-Rényi maximal correlation Hirschfeld and Wishart (1935); Gebelein (1941); Rényi (1959). This is defined as

$$R = \max_{f,g} \operatorname{corr}(f(X), g(Y)), \tag{1}$$

where two unknown nonlinear functions $f(x)$ and $g(y)$ are introduced and the maximisation in Eq. (1) is over a space of functions known as *Borel measurable functions*. It turns out that the search for the functions $f(x)$ and $g(y)$ is computationally infeasible because the space of functions is too large. We thus restrict the space of the functions so the maximisation in Eq. (1) is computationally feasible. A Reproducing Kernel Hilbert Space (RKHS) serves for this purpose.

Before proceeding, let us first intuitively motivate the basic concepts of RKHS and introduce the notation used in our association measure (NAC); a full treatment and detailed discussion given in the Supplementary Material. Interested readers may refer to (Scholkopf and Smola, 2001) for more technical details.

A real-valued symmetric function κ is said to be a kernel function if, for any N distinctive points $\{\omega_i\}_{i=1}^{N}$, its kernel matrix K (also known as Gram matrix), defined by

$$(K)_{ij} := \kappa(\omega_i, \omega_j) \tag{2}$$

is positive definite. One such kernel function commonly used for variables in a d-dimensional space is the Gaussian kernel

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right), \tag{3}$$

where $\|\boldsymbol{x}\|^2 = \sum_{j=1}^{d} x_i^2$ and σ is a parameter. Given two kernel functions $\kappa_X(x, x')$ and $\kappa_Y(y, y')$ and N individual samples $\{(x_i, y_i)\}_{i=1}^{N}$, we confine the choice of functions we shall use in Eq. (1) to be the Hilbert space, $\mathcal{H}$, defined by functions having the forms

$$f(x) = \sum_{i=1}^{N} \alpha_i \kappa_X(x, x_i), \qquad g(y) = \sum_{i=1}^{N} \beta_i \kappa_Y(y, y_i). \tag{4}$$

Here the α_i and the β_i represent weights that we will choose to maximize the linear correlation in Eq. (1).

We make the assumption here and throughout the paper that the kernel functions have been centralized (the sum over each column of the kernel matrix is zero). We then have the estimates of covariances and variances as

$$\begin{aligned} \operatorname{cov}(f(X), g(Y)) &= \frac{1}{N}\boldsymbol{\alpha}^T K_{XY} \boldsymbol{\beta}, \\ \operatorname{var}(f(X), f(X)) &= \frac{1}{N}\boldsymbol{\alpha}^T K_{XX} \boldsymbol{\alpha}, \\ \operatorname{var}(g(Y), g(Y)) &= \frac{1}{N}\boldsymbol{\beta}^T K_{YY} \boldsymbol{\beta}, \end{aligned}$$

where a T superscript denotes a matrix transpose, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ...)^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, ...)^T$, $K_{XY} = K_X K_Y$, $K_{XX} = K_X K_X$, $K_{YY} = K_Y K_Y$ and K_X, K_Y are the kernel matrices, as defined in Eq. (2), for X and Y, respectively. Thus the empirical estimate of R (Eq. (1)) is

$$\hat{r}_{\mathcal{H}} = \max_{f,g \in \mathcal{H}} \operatorname{corr}(f(X), g(Y)) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^N} \frac{\operatorname{cov}(f, g)}{\sqrt{\operatorname{var}(f)\operatorname{var}(g)}}. \tag{5}$$

By taking the derivative with respect to $\boldsymbol{\alpha}, \boldsymbol{\beta}$, the above formulation results in the eigenvalue problem

$$\begin{pmatrix} 0 & K_{XY} \\ K_{YX} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \rho \begin{pmatrix} K_{XX} & 0 \\ 0 & K_{YY} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \tag{6}$$

where the largest eigenvalue $\hat{\rho}_1$ corresponds to $\hat{r}_{\mathcal{H}}$. We call the k-th largest eigenvalue $\hat{\rho}_k$ the k-th largest $\mathcal{F}$-correlation, and the k-th eigenvector correspond to functionals in H_X and H_Y achieves that correlation. And we denote the functionals defined by the k-th eigenvector $[\boldsymbol{\alpha}^{(k)}; \boldsymbol{\beta}^{(k)}]$ as $f^{(k)}$ and $g^{(k)}$

respectively. With some simple algebraic manipulations one could show that $\mathrm{corr}(f^{(i)}, g^{(j)}) = 0$ for $i \neq j$ and $\mathrm{corr}(f^{(i)}, g^{(i)}) = \hat{\rho}_i$ for $i = j$.

The standard kernel canonical correlation methods from the literature (Wu et al., 2011; Ashrafulla et al., 2013; Dong et al., 2015) essentially only uses $\hat{r}_{\mathcal{H}}$, the largest eigenvalue. However, such practice could sometimes lead to suboptimal solutions. To see this, let us look at another motivating example. In addition to the point cloud circle $(X, Y) = \{(x_i, y_i)\}_{i=1}^{N}$, we also have $(U, V) = \{(u_i, v_i)\}_{i=1}^{N}$ as independent copies of (X, Y) and $(S, T) = \{(s_i, t_i)\}_{i=1}^{N}$ as *i.i.d.* samples of $\mathcal{N}(0, 1)$. For the nonlinear association between (X, U) and (Y, V), we find $f_1(x) = x^2 - \mathbb{E}[x^2]$, $f_2(u) = u^2 - \mathbb{E}[u^2]$, $g_1(y) = y^2 - \mathbb{E}[y^2]$ and $g_2(v) = v^2 - \mathbb{E}[v^2]$ satisfies

$$\mathrm{corr}(f_1, g_1) = \mathrm{corr}(f_2, g_2) \approx 1,$$
$$\mathrm{corr}(f_1, f_2) = \mathrm{corr}(g_1, g_2) = 0,$$

while for any functions $f_3(\cdot) : S \to \mathbb{R}$ and $g_3(\cdot) : T \to \mathbb{R}$,

$$\mathrm{corr}(f_3, g_3) = 0.$$

(For clarity of discussion we only consider functions of one variable.) If only the largest correlation pair is considered, then one may erroneously reach the conclusion that the association between (X, S) and (Y, T) is as strong as that between (X, U) and (Y, V). Further evaluation of the second correlation pair makes it evident that (X, U) and (Y, V) are more closely related.

With the preceding examples in mind, we can now define the NAC measure that we shall introduce in this work. To account for the contribution from ensuing correlated function pairs, we propose using the leading k eigenvalues of Eq.(6) according to

$$\mathrm{NAC}(X; Y|k) = -\frac{1}{2} \sum_{j=1}^{k} \log(1 - \rho_j^2), \tag{7}$$

which pieces together as much evidence of independence as is possible and results in a more flexible measure of independence. This is the key distinction

between our method and (Gretton et al., 2005; Dauxois and Nkiet, 1998) where all $\{\rho_j\}_{j=1}^N$ are calculated. In what follows, we shall abbreviate $\mathrm{NAC}(X;Y|k)$ to $\mathrm{NAC}(k)$ when this does not cause any confusion. We note that in addition to NAC being a generalization of existing methods, we give, in the Appendix, an alternative motivation of NAC as being related to mutual information. We also give the relationship between NAC and another kernel independence measure (HSIC). We note proper regularization is essential for the estimation of NAC. We relegate those details and our computation recipe to the Appendix.

2.2. Permutation and parametric estimation based inference

To make statistical inference about the detected NAC, we need to construct the distribution under the *null* hypothesis, where X and Y are independent. Although the NAC statistic looks similar to the *Wilks' lambda* commonly used in the statistical inference of CCA, to our knowledge no known theoretical results has been obtained for the kernel case. Thus we resort to permutation based test (Westfall, 1993). Let $\boldsymbol{\pi}$ denote a permutation of index $\{1,\cdots,N\}$. We randomly generate B permutations and denote them as $\{\boldsymbol{\pi}^{(b)}\}_{b=1}^B$. We call $\{X_{\boldsymbol{\pi}},Y\} = \{X_{\boldsymbol{\pi}_i},Y_i\}_{i=1}^N$ a $\boldsymbol{\pi}$ permutation sample. By calculating $\mathrm{NAC}^{(b)} := \mathrm{NAC}(Y,X_{\boldsymbol{\pi}^{(b)}}|k)$ for $b = 1,\cdots,B$ we obtain the empirical *null* distribution $\{\mathrm{NAC}^{(b)}\}_{b=1}^B$ of $\mathrm{NAC}(Y,X|k)$ under the independent hypothesis.

However, it is the common situation that we need to carry out various multi-comparison corrections to control for the false positives, e.g. Bonferroni, Holmes, False Discovery Rate, etc. (Efron, 2010) To obtain a desired statistical significance, one needs to increase the permutation runs, which is very computationally intensive. For example, if one is interested in the pair-wise relationship between p variables, then approximately $p^2/2$ corrections should be made, for $p = 100$ and family-wise error rate (FWER) 0.05, one need to run at least $B = 10^5$ permutations to achieve the specified significance, and numerical recipes often suggest running 10 times more permutations than the minimum requirement to stabilize the result. The computational burden obstacles the practical application of computation-intensive nonlinear association procedures.

This motivates us to use parametric distributions to approximate the *null* distribution with reasonable permutation runs (Ge et al., 2012), thus relieving the computational burden evaluating the significance level of extreme events. The significance of the test statistics will be evaluated using the distribution parameters estimated from a small number of permutation runs, say B = 500. Then we can safely proceed to the statistical correction procedures using these estimated p-value. We carried out extensive numerical experiments and found that the *null* distributions of NAC are well approximated by the *Gamma* distribution, given that the input variables are appropriately 'normalized', in our case copula-transformed. This observation is also corroborated by the fact that HSIC's theoretically derived *null* distribution follows mixed chi-square and is also well-approximated by the *Gamma* distribution (Gretton et al., 2005). More specifically, we observe that the *Gamma* approximation of NAC gives an accurate or slightly conservative estimate of the p-value most of the time. Occasionally in some cases *Gamma* approximation of NAC gives inflated p-values, but it is still safe for us to report the empirical p-value instead. This offers us the possibility to significantly reduce the computational burden. We emphasize that if the approximated p-value p_{approx} is to be used, diagnostic plot should be reported along with it to reassure that the approximation is reliable. To ensure the stability of the parameter estimation of the *null* distribution, we found that it is necessary to cut off a small portion of the left side of the distribution to give a stable estimation. In our studies, we discard 2% of the left most permutation statistics when estimating the parameter of the *Gamma* distribution. We also note that one may also use *Extreme Value Theory* (Coles et al., 2001) to directly estimate the tail, rather than the entire distribution of the *null*. For large scale inference, the excess amount of computation could be avoided by stoping the permutations procedure when the majority of the first B' permutations, where $B' \ll B$, result in larger NAC statistics.

2.3. Conditional NAC

In clinical research the conditional association is crucial. The researchers need to remove the confounding factors such as age, gender, etc. Yet currently the recipes for assessing nonlinear conditional association are limited. Fukumizu et al. (2007b) proposed to use the in-slice permutation to obtain the *null* distribution of NOCCO estimator. But the applicability of in-slice permutation is only valid when covariates are low-dimensional and closely clustered. Other nonlinear conditional association measures include (Póczos and Schneider, 2012). We propose to use regularized kernel regression (Shawe-Taylor and Cristianini, 2004) on the original feature space to remove the covariate effects. The optimal regularization parameter is chosen via cross validation, thus ensuring the maximum variance related to covariates being removed, then the NAC is calculated from the residuals as described in earlier sections. We identify the NAC of the residuals that regress out Z as the conditional NAC, i.e.

$$\mathrm{NAC}(X, Y|Z, k) \coloneqq \mathrm{NAC}(\tilde{X}, \tilde{Y}|k),$$

where $\tilde{X}$ and $\tilde{Y}$ are the residuals regress out Z. We note that parametric models (e.g. linear regression) are preferred compared with kernel regression to obtain the residual provided the assumptions are valid because they are more statistically efficient than their kernel counterparts.

2.4. Signal Reconstruction and identifying the contributing components

In addition to being informed that the two systems are nonlinearly associated or causally linked, researchers would be more interested to know what is the original signal that associates the two systems. Fortunately it is possible to reverse the associated signals from the observed signals with the kernel method. We notice recall that while NAC use the leading eigenvalues to quantify the association, those associated eigenvectors give the empirical functional f and g that maximize the correlation between $\mathcal{H}_X$ and $\mathcal{H}_Y$, i.e.

$$f^{(k)}(\cdot) = \sum_{i=1}^{N} \alpha_i^{(k)} \kappa_X(\cdot, x_i), g^{(k)}(\cdot) = \sum_{i=1}^{N} \beta_i^{(k)} \kappa_Y(\cdot, y_i).$$

So by evaluating $f^{(k)}$ and $g^{(k)}$ we can estimate the signal with the k-th largest correlation. After the first k eigen-signals have been extracted, the users might want to take an extra step to use well established ICA procedures (Hyvärinen and Oja, 2000; Bach and Jordan, 2003) to further decompose the reconstructed signal space into independent ones.

Identifying the contributing components in the original space is crucial for the practitioners as it helps the practitioners to refine their model according the evidence from the observed data and may offer better interpretability. Modern high throughput technologies often provide researchers with excessive number of features. Those irrelevant features will create difficulties in model interpretation, extra computational cost and reduced generalization ability. We propose that the kernel nonlinear association procedure could be used to fulfill the job, thus obtaining a kernel version of the feature selection procedure. And our procedure is different from existing HSIC-based feature selection procedures Song et al. (2007, 2012). Basically, we identify the candidate variables via the discrepancies between marginal distribution of variables in original space belonging to different slices with respect to the kernel identified associated components, i.e. reconstructed associated signals. The elimination of non-contributing components may further boost the association signal we detected.

Using the signal reconstruction, we can obtain two associated components from K_Y and K_X respectively for the largest correlation coefficient, denoted as $c_Y^{(1)}$ and $c_X^{(1)}$ respectively. Then we slice $c_X^{(1)}$ and $c_Y^{(1)}$ into s parts and we divide the samples of X into subgroups according to slicing scheme for $c_Y^{(1)}$ and divide Y according slicing scheme for $c_X^{(1)}$.

Then we test whether the sliced samples come from the same marginal distribution using Kruskal-Wallis test for each component of original X and Y. For the purpose of screening candidates, one can also threshold the p-value using a prespecified threshold. Similar procedure could be repeated for other leading eigen-components. This allows us to identify the contributing components within only one kernel association step, rather than repetitively computing the kernel matrix and performing eigen-decomposition as in (Song et al., 2007).

2.5. Nonlinear Granger Causality

Originally proposed by N. Wiener (Wiener, 1956) and later formalized by E. Granger in the context of multivariate auto-regression (MVAR) (Granger, 1969), Granger causality has gained popularity among researchers over the past few decades. For Granger Causality, people test for whether the alternative model (i.e. Eq. (9)) better explains the observed series compared with the *null* model (i.e. Eq. (8)),

$$\boldsymbol{y}_t^{(p')} = \psi_1(\boldsymbol{y}_{t-1}^{(p)}, \boldsymbol{z}_t^{(r)}) + \boldsymbol{n}_{1,t}, \tag{8}$$

$$\boldsymbol{y}_t^{(p')} = \psi_2(\boldsymbol{y}_{t-1}^{(p)}, \boldsymbol{x}_{t-1}^{(q)}, \boldsymbol{z}_t^{(r)}) + \boldsymbol{n}_{2,t}, \tag{9}$$

where $\boldsymbol{z}_t$ is the confounding covariates. $\boldsymbol{y}_t^{(p')}, \boldsymbol{y}_t^{(p)}$, $\boldsymbol{x}_t^{(q)}$ and $\boldsymbol{z}_t^{(r)}$ are delayed embeddings of the respective series (i.e. Eq. (10~13)), ψ_1 and ψ_2 are the (nonlinear) mappings that the system evolves. Temporal overlapping of the lag embedding vectors is avoided, i.e.

$$\boldsymbol{y}_t^{(p')} = \left[\boldsymbol{y}_t, \boldsymbol{y}_{t+1}, \cdots, \boldsymbol{y}_{t+p'-1}\right], \tag{10}$$

$$\boldsymbol{y}_{t-1}^{(p)} = \left[\boldsymbol{y}_{t-1}, \cdots, \boldsymbol{y}_{t-p}\right], \tag{11}$$

$$\boldsymbol{z}_{t-1}^{(r)} = \left[\boldsymbol{z}_t, \cdots, \boldsymbol{z}_{t-r+1}\right], \tag{12}$$

$$\boldsymbol{x}_{t-1}^{(q)} = \left[\boldsymbol{x}_{t-1}, \cdots, \boldsymbol{x}_{t-q}\right]. \tag{13}$$

The magnitude of GC is defined as the log of the ratio of residual variance (in the case of univariate time series)

$$GC = \log \frac{var(\boldsymbol{n}_{1,t})}{var(\boldsymbol{n}_{2,t})}, \tag{14}$$

or more generally for the multivariate time series,

$$GC_\phi = \log \frac{\phi(cov(\boldsymbol{n}_{1,t}))}{\phi(cov(\boldsymbol{n}_{2,t}))}, \tag{15}$$

where ϕ could either be determinant or trace. The linear assumption in the original formulation of GC is rather restrictive. The growing recognition of the nonlinear phenomenon in the scientific community has inspired many nonlinear extensions of GC Chen et al. (2004); Marinazzo et al. (2008); Wu et al. (2011);

Sun (2008). Here we propose a new nonlinear extension of GC based on NAC. Slightly different from the GC that uses variance to characterize the association between X and Y, we define the nonlinear Granger Causality directly as the kernel association condition on external drives and history of Y:

$$\mathrm{NGC}(k) \coloneqq \mathrm{NAC}(\boldsymbol{y}_t^{(p')}, \boldsymbol{x}_{t-1}^{(q)} | \boldsymbol{y}_{t-1}^{(p)}, \boldsymbol{z}_t^{(r)}, k).$$

2.6. Choice of parameters

The sensitivity of NAC depends on the regularization parameter λ. When the λ is too small, it allows the NAC go underregularized and prone to overfitting the noise as signal, thus diminishing the significance of detected association. To the other extreme, when it is overly-regularized, the approximation error will dominate thus jeopardizing the algorithm's power to detect association. To find the optimal regularization parameter for empirical data, we propose using an n-fold cross-validation procedure. First the data is divided into n none-overlapping batches. Each time one batch, say the j^{th} batch is selected as test set and the rest $n-1$ batches are used to estimate the association functional f and g respectively. The estimated $\hat{f}_j^{(i)}$ and $\hat{g}_j^{(i)}$ is applied to the test batch and the correlation coefficient $\rho_j^{(i)}$ between $\hat{f}_j^{(i)}$ and $\hat{g_j}^{(i)}$ on the test batch is evaluated (we refer to this the predictive correlation). We denote the mean (or the median) of (absolute value of) out-of-sample correlation coefficient for the i-th functional pair as $\hat{\rho}_i$. This procedure is repeated for every candidate regularization λ parameter and the one with the out-of-sample correlation will be selected as the optimal regularization parameter λ_o to be used for the testing the association on the entire sample. The score metric we maximize is the same as kernel association statistics, i.e. $\lambda_{opt} = \arg\max_\lambda \sum_{i=1}^{k} \log(1 - \hat{\rho}_{\lambda,i}^2)$. By slight abuse of notation we write $\lambda = a$ for $\lambda_X = \lambda_Y = a$, i.e. fixing the regularization parameters to be the same, for most of the illustrative examples in the following sections unless explicitly stated. In principle λ_X and λ_Y should be optimized separately to maximize the sensitivity.

The number of eigenvalues also affects the sensitivity of the kernel association procedure. Retaining only the first correlation coefficient, like it is done

by most of CCA applications, renders NAC (i.e., $k = 1$ in Eq. (7)) blunt in situations where no single dominant but multiple relatively weak independent associations are functioning between two systems. If multiple (second, third, i.e. $k = 2, 3$, etc.) coefficients are retained, we might be able to accumulate the evidence to make a more informed inference. But keeping too many eigenvalues increases the variance of the estimator, thus diminishing the sensitivity. In this regard, to maximize the sensitivity of NAC, the number of eigenvalues should be optimized. A straight forward solution is first construct a *null* distribution of the out-of-sample predictive correlation with respect to each eigenvalue and then retain the leading eigenvalues that the true predictive correlation show significant deviation from the *null*. The pseudocode of the algorithm is detailed in Algorithm 1 in Appendix. However, this procedure is computationally intensive as it involves solving too many eigenvalue problems. A more efficient but inaccurate strategy is to skip the eigenvalue problem and directly estimate the correlations. For example, f and g estimates under the *null* could be be simulated by drawing random numbers from the standard normal distribution. This efficient heuristic approach works fine in practice (see Appendix for examples). Another subjective way to determine the number of eigenvalues is to plot the predictive correlation and choose the number where an abrupt fall of the predictive correlation is observed.

3. Simulation studies

In this section we demonstrate our proposed method using a few numerical examples. In all the following studies, Gaussian kernel in Eq. (3) is used as the kernel function and the bandwidth parameter σ^2 is set to the median of pair-wise ℓ^2 distance as suggested by Gretton et al. (2007). For the purpose of demonstration, we simply fix the regularization parameter to be equal (write $\lambda = \lambda_X = \lambda_Y$) while in practice we strongly recommend this to be optimized for each dataset separately. We will abbreviate the alternative case as *alt*.

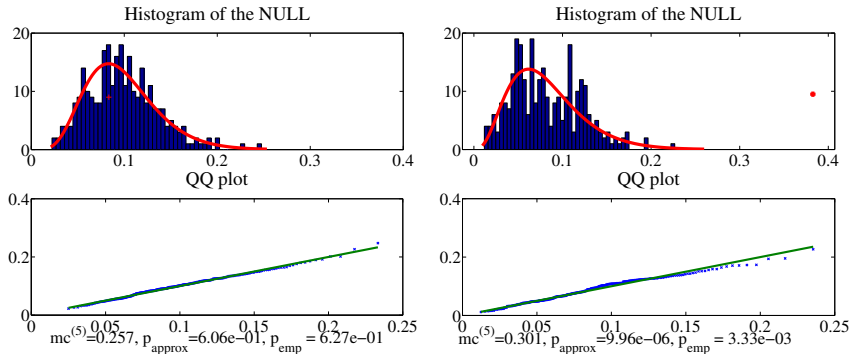
Example 1: Simple linear mixing example

To begin, we generate the data from the following linear model, and will demonstrate nonlinear examples later. We have

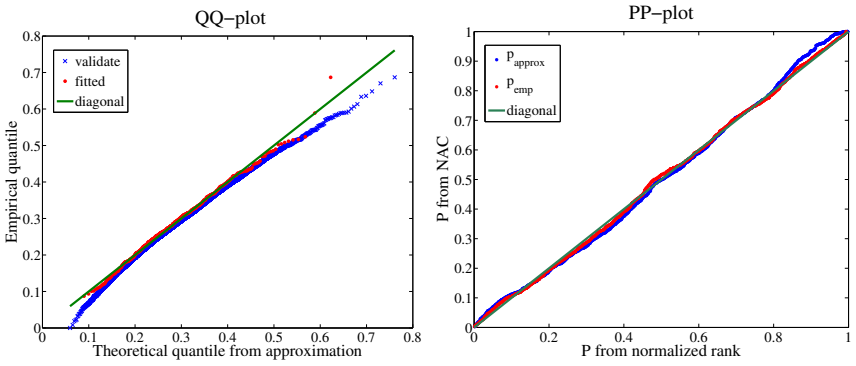
$$\begin{cases} \mathbf{x} &= A_1\mathbf{u} + B_1\mathbf{v} + \mathbf{n_x} \\ \mathbf{y} &= A_2\mathbf{u} + B_2\mathbf{v} + \mathbf{n_y}, \end{cases}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two random vectors, $\mathbf{u}$ is the signal that responsible for the association, and $\mathbf{v}$ are confounding factors that we intend to eliminate from the model, $\mathbf{n_x}$ and $\mathbf{n_y}$ are the standard independent Gaussian random perturbations. The $\{\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{n}_x, \boldsymbol{n}_y\}$ are all drawn from *i.i.d.* standard Gaussian. $A_i, B_i (i = 1, 2)$ are the mixing matrices with each entry drawn from *i.i.d.* $[0, 1]$ uniform distribution unless otherwise specified. In our experiment, we set the dimension of $\mathbf{x}$, $\mathbf{y}$, $\mathbf{u}$ and $\mathbf{v}$ to $1, 5, 2$ and 4 respectively. For each experiment, we drew $N = 300$ samples from the model and ran $B = 300$ permutations. We retained the first $k = 5$ eigencomponents and the regularization parameter was set to $\lambda = 0.1$. Entries of the A_i matrices are set to zero in the *null* case. Linear regression is used to remove the covariate effects of $\boldsymbol{v}$ from $\boldsymbol{x}$ and $\boldsymbol{y}$. In Figure 2(a) we present the diagnostic plot of NAC without covariate influence (set $B_i = 0$) for one *null* and one *alt* realization respectively. In both cases the empirical *null* distribution of NAC follow the fitted *Gamma* distribution very well and NAC identifies the *alt* case to be significant.

We further verified the validity of using *Gamma* approximation for the p-value. For the *null* case, $B = 10,000$ permutations were run while only the first 500 of them were used to estimate the parameters for the *Gamma* distribution. The QQ-plot of points both used (fitted) or not used (validate) in approximation is shown in Figure 2(b). The large permutation experiment showed that the *Gamma* approximation gives slightly conservative estimate for the significance level, validating its role as a surrogate for the empirical p-values obtained from large permutations. The discrepancy for the smaller quantiles is caused by the truncation used when fitting the *Gamma* distribution. Since this part of the quantile corresponds to very large p-values ($p > 0.9$) so it is of no practical interest in association studies. We then explored the specificity of the proposed



(a)



(b) (c)

Figure 2: (a) Diagnostic plot from the kernel association routine. Upper panel is the histogram of the permutation statistics and *Gamma* fit, the red plus sign indicates the calculated kernel association from the original data. Lower panel depict the QQ-plot. Left column: *null* model. Right column: *alt* model. For this realization *alt* model sampled $A_1 = [0.87, 0.17]$, $A_2 = [0.69, 0.74; 0.76, 0.64; 0.72, 0.40; 0.03, 0.60; 0.68, 0.21]$. $\text{mc}^{(k)}$ denote the mean of first k eigenvalues(i.e. mean $\mathcal{F}$-correlation). (b) QQ-plot for the *Gamma* approximation of NAC, red for fitted permutation runs and blue for validation runs (c) PP-plot for the rejection rate of *null* model using two-sides covariate removal, red using the approximated p-value and blue for empirical p-value.

NAC with covariate removal procedure. $M = 1,000$ simulations were carried out using the *null* model. We tested the specificity of the procedure with covariates removal on one side (X) and on both sides (X and Y). Both approaches showed expected rejection rate. More general kernel regression gives similar result. We show only the specificity result of two-sides covariate removal as P-P plot in Figure 2(c). Note that statistical model consistent with the data should be used in covariate removal to avoid inflation of p-value.

Example 2: Nonlinear Logistic mapping

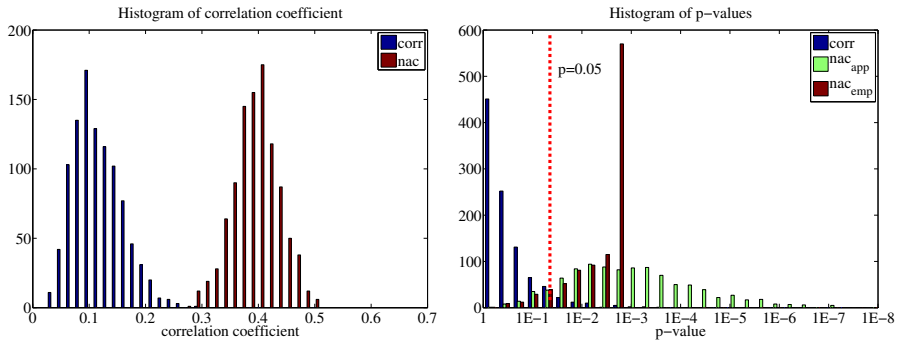
We now turn our attention to a nonlinear model, the logistic mapping, to show how NAC can pick up a nonlinear association, but a linear association simply fails. The model we used is

$$\begin{cases} \mathbf{d}_1 &= \mathbf{s} + \mathbf{n}_1 \\ \mathbf{d}_2 &= a\mathbf{s} \odot (\mathbf{1} - \mathbf{s}) + \mathbf{n}_2 \end{cases}$$

where $\mathbf{s}$ are i.i.d random variables uniformly distributed in $[0, 1]$ and $\odot$ is the element-wise multiplication. Various settings of the model with different parameters are extensively tested as described below.

We first compared the performance of linear correlation and NAC on this classic nonlinear model. The dimension of both $\mathbf{d}_1$ and $\mathbf{d}_2$ is set to 5. $\mathbf{n}_1, \mathbf{n}_2$ are i.i.d random variables uniformly distributed in $[0, 1]$. We set $a = 1$. The NAC parameters are set to $k = 5, \lambda = 1, B = 500$. $N = 200$ samples are generated for each experiment and repeated for $M = 1,000$ times. The distribution of correlation and p-values are shown in Figure 3. It is clear that the linear correlation cannot detect the correlation and NAC significantly outperforms the linear correlation in this case. Also note that the empirical p-value estimate is limited by the number of permutations runs, in this case 2×10^{-3}, while the approximated p-value gives a conservative estimate of the significance level under the *null* hypothesis with reasonable cost. For small p-values, *Gamma* approximation will spare the enormous computations required by empirical approach to reach the desired significance.

We show the sensitivity dependence on the parameters of NAC and auto-



(a) (b)

Figure 3: Performance of NAC on a nonlinear model. (a) Histogram of correlation strength. blue: histogram of the maximum absolute linear correlation between $d_{1,i}$ and $d_{2,i}$; red: histogram of the largest eigenvalue of NAC. (b) histogram of the p-values in $\log_{10}$ scale. Only the p-values for linear $corr(d_{1,1}, d_{2,1})$ are included to make the histogram visually comparable. The p-values of linear correlation are uncorrected for multiple comparisons. blue:corr;green:approximated p-value of NAC; red: empirical p-value of NAC.

matic parameter selection procedure using this example together with an illustrative example of NGC using the Logistic mapping in Appendix.

Example 3: Variable selection

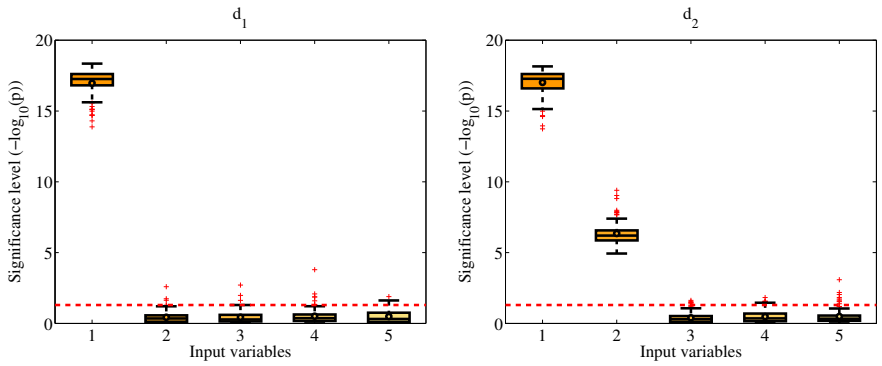
We used the following model

$$\begin{aligned} d_1 &= [x_t, \hat{x}_t^{(1)}, \cdots, \hat{x}_t^{(4)}]_{t=1}^N \\ d_2 &= [\sin(x_{t+1}), \cos(x_{t+1}), \hat{x}_{t+1}^{(5)}, \cdots, \hat{x}_{t+1}^{(7)}]_{t=1}^N, \end{aligned} \tag{16}$$

here x_t are generated by the Logistic model

$$x_{t+1} = \alpha\, x_t\,(1 - x_t) + n_t$$

with $\alpha = 3.6$ and $\{n_t\}$ *i.i.d.* draws from $\mathcal{N}(0, 4\times10^{-4})$. Sample paths of x_t starts from a uniformly distributed random initial point between $[0, 1]$ and $\{\hat{x}_t^{(i)}\}_{i=1}^4$ are the randomly permuted version of x_t, i.e. breaking the causal link but retaining the marginal distribution. Each $\hat{x}_t^{(i)}$ is permuted independently. The NAC parameters are set to be $k = 1, B = 500, \lambda = 0.1$ and the number of slices is set to $s = 3$. For each experiment we draw $N = 300$ samples and repeat



(a) (b)

Figure 4: Results of identifying contributing components. Box plot of the significance level (negative $\log_{10}$) of the Kruskal-Wallis test using $B = 100$ simulation runs ($N = 100$ sample points each) with 3-slices. (a) is the results for d_1 dataset, (b) is for d_2. Red dash line denotes thresholding $p = 0.05$.

the experiments for $M = 100$ times. We present the result in Figure 4. This procedure successfully identified the component 1 from d_1 and the component $1, 2$ from d_2 to be contributing components of the association detected while the others show no sign of significant association, which matches the ground truth structure.

Example 4: Comparison between NAC and other nonlinear measures

In this example we compared the performance of NAC using varying number of eigencomponents (k) with HSIC, NOCCO and MI. We note that NAC is equivalent to kernel canonical correlation analysis (KCCA) commonly used in the neuroimaging literature when only the first eigenvalue is used (Wu et al., 2011) so we use NAC(1) as a surrogate for KCCA. We use the following two toy models, which were specifically designed to have high dimensionality and low SNR commonly encountered in brain imaging researches, to benchmark the performance of NAC and its nonlinear counterparts:

Model A

$$\begin{aligned} \mathbf{d}_1 &= [s_1, \cdots, s_5, n_1^{(1)}, \cdots, n_5^{(1)}] \\ \mathbf{d}_2 &= [\alpha\,[s_1, \cdots, s_5] + \beta\,[n_1^{(2)}, \cdots, n_5^{(2)}], n_6, \cdots, n_{10}], \end{aligned} \tag{17}$$

where $s_i, i = 1, \cdots, 5$ and $n_i^{(j)}, i = 1, \cdots, 5, j = 1, 2, n_i, i = 6, \cdots, 10$ are *i.i.d.* standard Gaussian random variables and we set $\alpha = 0.5$ and $\beta = 2$. For the *null* case, we randomly shuffle the rows of $\boldsymbol{d}_1$.

Model B

$$\begin{aligned} \mathbf{d}_1 &= [s_1, \cdots, s_5, n_1, n_2] \\ \mathbf{d}_2 &= [h(s_1, \cdots, s_5), n_1', n_2'], \end{aligned} \tag{18}$$

where $h(s_1, \cdots, s_5) = [\cos(2\pi s_1), \cdots, \cos(2\pi s_5)] \times W + [r_1, \cdots, r_5]$. W is a linear mixing matrix to mask the signals with each entry drawn from standard Gaussian. $\{n_1, n_2, n_1', n_2'\}$ and $\{r_i\}_{i=1}^5$ are *i.i.d.* standard Gaussian variables. $\{s_i\}_{i=1}^5$ are independently sampled from $[0, 1]$ uniform distribution. For the *null* case, we also randomly shuffle the rows of $\boldsymbol{d}_1$.

$N = 100$ samples were drawn for each experiment, and the association was evaluated by the NAC using $k = 1, 3, 5$, denoting NAC(k) respectively. We used k-nearest neighbor MI estimation (Kraskov et al., 2004) implemented in (Stögbauer et al., 2004) to calculate MI with the number of nearest neighbors set to $k_{nei} = 5$. We adapted the HSIC code given in (Gretton et al., 2005) to calculate both HSIC and NOCCO statistics. The p-value for HSIC statistic is obtained using moment-matching *Gamma* approximation while the p-value for NOCCO and MI is obtained via permutation. The number of all permutation runs was set to $B = 500$. In all kernel based statistics (NAC, HSIC and NOCCO) Gaussian kernel was used and bandwidth parameter σ were fixed to be the median of pair-wise distance as suggested by (Gretton et al., 2005). For those methods need regularization (NAC and NOCCO) the regularization parameter was fixed to $\lambda = 0.1$. We repeated the experiments $M = 1,000$ times for all cases.

We use the $-\log_{10} p$ to draw the Receiver Operating Characteristic (ROC) curve and use the score known as Area Under ROC Curve (AUC) to evaluate the performance of all algorithms in discriminating *null* and *alt* cases. See Figure

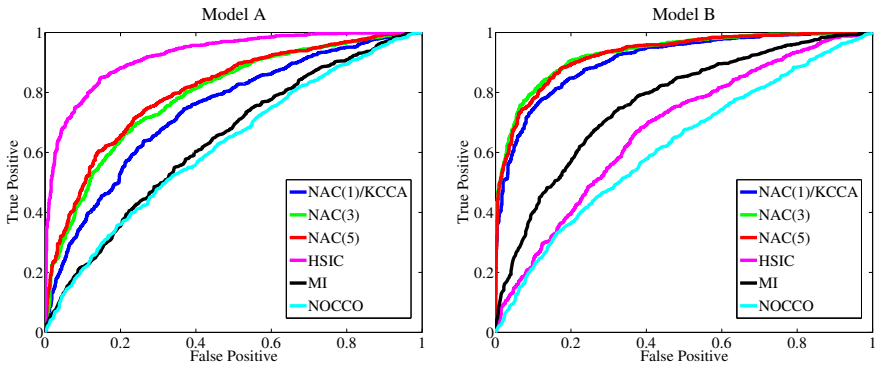


Figure 5: ROC curves of NAC and HSIC for linear case (model A, left) and nonlinear case (model B, right) (blue: NAC(1); green: NAC(3); red: NAC(5); purple: HSIC; black: MI; cyan: NOCCO)

Model	NAC(1)/KCCA	NAC(3)	NAC(5)	HSIC	MI	NOCCO
A	0.741	0.792	0.806	**0.924**	0.640	0.623
B	0.908	**0.931**	0.927	0.679	0.767	0.620

Table 1: AUC scores for NAC(k), $k = 1, 3, 5$, and HSIC,MI,NOCCO of two toy models. Maximum AUC score achieved is marked in bold.

5 for the ROC curves and Figure H.3 for the smoothed histogram of $-\log_{10} p$ in the *alt* case. The AUC scores are tabulated in Table. 1. One could see while HSIC is more sensitive than NAC in model A, NAC significantly outperforms HSIC in model B. This suggests there is no universal optimal choice for nonlinear association measure, at least between NAC and HSIC. And different choice of k gives different sensitivity in the experiments, confirming our hypothesis that in the case of multiple independent associated components, using $k > 1$ gives sensitivity gain over KCCA. We note NAC consistently outperformed MI and NOCCO for both models while HSIC failed to do so. This robustness against noise might make NAC more favorable to its competitors in low SNR brain imaging studies.

4. Real-world Datasets

After testing our NAC intensively in the previous subsection, now we apply it to four real world datasets.

4.1. High Temporal Resolution fMRI Causality

We use the enhanced Nathan Kline Institute-Rockland Sample (NKIRS) (Nooner et al., 2012) to compare the performance of traditional GC and NGC. This dataset achieved unprecedented sampling rates for full brain coverage compared with normal fMRI studies (TR: 0.645 sec vs. 2.0 ~ 2.5 sec) through the acquisition of multiple slices simultaneously in the same time. This allows more refined characterization of the brain dynamics. The entire dataset consists of resting-state fMRI scannings of 231 subjects aged from 6 to 85. Readers are referred to the project's website for more detailed description of the dataset. 900 volumes were collected for each subject during the 10-min acquisition period. First 50 volumes were discarded to allow for scanner stabilization. Data preprocessing was conducted using *DPARSF* (Chao-Gan and Yu-Feng, 2010) together with *SPM8* (Ashburner et al., 2012). Slice timing correction was omitted because of the short TR. All images were realigned and normalized to the standard MNI template with voxel size $3 \times 3 \times 3\,\mathrm{mm}^3$. Then the images were spatially smoothed with FWHM = 8 mm and band-pass filtered between 0.01 ~ 0.08 Hz to boost the signal-to-noise ratio. The brain volume is parcellated into 90 regions using the automated anatomical labeling atlas (AAL) (Tzourio-Mazoyer et al., 2002). Regional mean signal were extracted and z-transformed for subsequent analyses.

After preprocessing, the regional signals were delay-embedded to represent the present dynamical state of the brain. More specifically, denote x_t^i as the signal from brain region i at time point t, the current state of the brain is represented by the vector $\boldsymbol{x}_t^i = [x_t^i, \cdots, x_{t-(p-1)l}^i]$. The the lag size l and embedding dimension p could be empirically determined using delayed mutual information and false nearest neighbors respectively, which are common practice in the reconstruction of attractors of nonlinear dynamical systems (Kantz and Schreiber,

2004; Stark, 2001). In the current study, we fixed the parameters to be $l = 1$ and $p = 3$, matching the TR of normal fMRI studies. We are interested in the connectommic changes during the normal aging process. So we ran both GC and NGC on the NKIR sample and benchmarked NGC against GC in this real world fMRI study.

For the NGC between brain region i and j, the embedded dynamical state sequence of the two regions are denoted as $\boldsymbol{X}^i = \{\boldsymbol{x}_t^i\}_{t=1}^T$ and $\boldsymbol{X}^j = \{\boldsymbol{x}_t^j\}_{t=1}^T$ respectively. The directional NGC from i to j is defined as $\mathrm{NGC}(\boldsymbol{x}_{t-p}^i, \mathbf{x}_t^j | \boldsymbol{x}_{t-p}^j)$. Conditioning on $\boldsymbol{x}_{t-p}^j$ allows us to remove the auto-correlation for $\boldsymbol{X}^j$ thus making sure that we are evaluating the predictive gain from historical records of $\boldsymbol{X}^i$ for the future $\boldsymbol{X}^j$. Also the embedded vectors are temporally non-intersecting to make sure the conditioning operation will not completely eliminate the information at certain dimension. We used the kernel regression on both sides to carry out the conditioning operation. The kernel regression regularization parameter λ_{kr} was numerated from $\exp(-6)$ to $\exp(8)$ in a log-linear step-size of 1 and the regularization parameter with the minimal 10-fold cross-validation residual variance was chosen to carry out the kernel regression. The NAC parameters were set to $k = 3$ and $\lambda = 0.1$. For the GC study, the mean of the embedded vector $\boldsymbol{x}_t^j$ (denoted as $\bar{\mathbf{x}}_t^j$) is used as the response variable. Then it is regressed against $\{\boldsymbol{x}_{t-m}^j\}$ in the *null* model and $\{\boldsymbol{x}_{t-m}^i, \boldsymbol{x}_{t-m}^j\}$ in the alternative model. The log-ratio of the residual variances of $\boldsymbol{x}_t^j$ from the two models was used as the statistics for GC. To compare the performance of NGC and other related kernel association measures, we implemented a modified KCCA as reference. The modification we adopted significantly reduced the computational burden of the KCCA in (Wu et al., 2011), so it takes roughly the same amount of time to calculate compared with NGC. We observe that the PCA of the kernel matrix only has two dominant eigen-directions. (one account for 97% and the other 2% of the variance). Thus the kernel matrix essentially could be well approximated by a low-rank approximation. Instead of using all the columns of kernel matrix, we use only the first 50 columns to perform the PCA.

While the p-values for GC statistics could be easily obtained via F-test, the

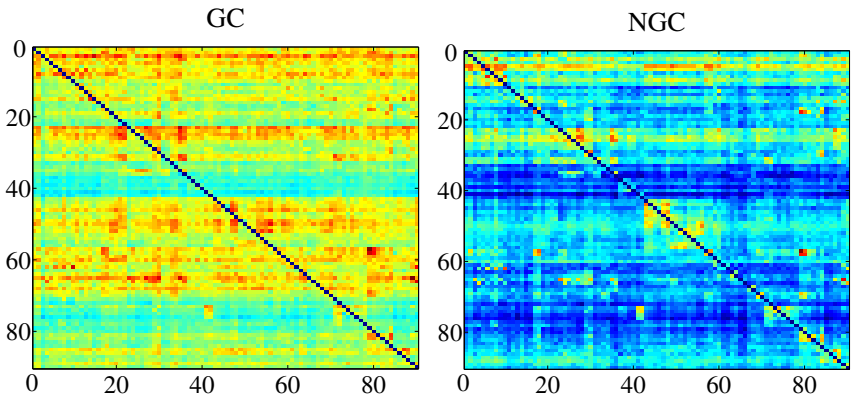
p-values for NGC and KCCA are too computationally intensive for the number of causal pairs (8010 pairs for each subject) we calculated, although the *Gamma* approximation still appear to be valid (see Figure H.4). In this regard we made no attempt to calculate the p-value for the NGC. Instead, we calculate only the causal statistics for each individual, and then perform group-level analysis.

We assign the subjects into two age groups, the young group and elderly group. The young group consists of subjects aged from 14 to 30 (76 subjects, 39 male, 4 left handed, 6 no preference for handedness) while subjects over 50 are assigned to the elderly group (80 subjects, 28 male, 3 left handed, 4 no preference for handedness). We ran two-sample T-test to compare the group difference of the causal statistics and regressed them against age using gender, handedness as covariates on all subjects older than 14. The population mean maps of GC/NGC are shown in Figure 6(a) and Figure 6(b) visualizes the $-\log$-significance map of the regression coefficients. The number of significant links adjusted for multiple comparison with both Bonferroni $p_{\text{FWE}} = 0.05$ and FDR $q = 0.05$ for all three methods are tabularized in Table 2. Kernel methods clearly show more sensitivity towards age related changes in brain connectomics compared with their linear counterparts. We observe in Figure 7(a) that the distribution of the regression coefficients on age are shifted towards the positive side for NGC, implicating a global trend that the strength of the connection intensifies with age.

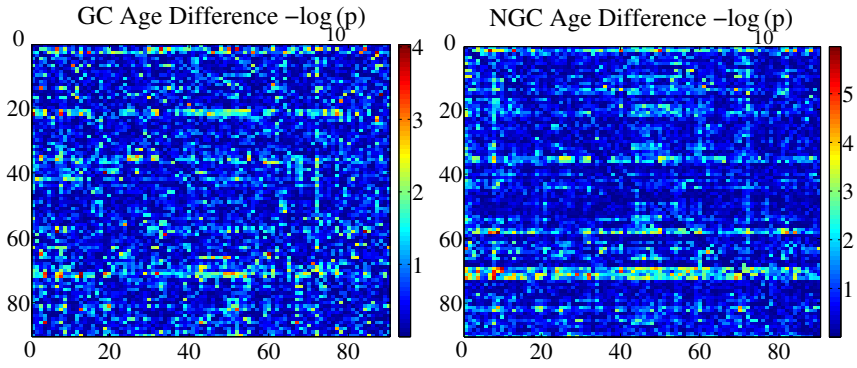
Correction	Model	NGC	GC	KCCA
Bonferroni	Regression	24	2	19
	T-test	5	0	9
FDR	Regression	736	10	682
	T-test	205	0	373

Table 2: Number of significant links corrected for multiple comparisons with significance level $p_{\text{FWE}} = 0.05$ (Bonferroni) and $q = 0.05$ (FDR).

We note that among the connections reported by NGC, most of the links

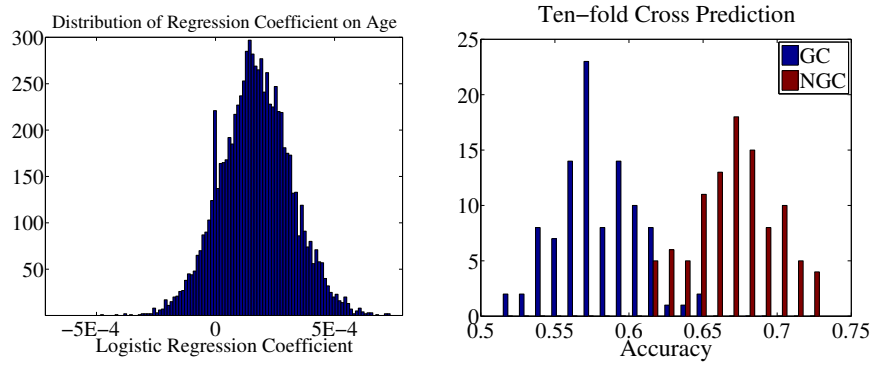


(a)



(b)

Figure 6: (a) population mean of GC and NGC maps (b) negative $\log_{10}$ p-value of the two-sample T-test with GC and NGC. The X-axis is the target AAL label and Y-axis is the source AAL-label.



(a) (b)

Figure 7: (a) distribution of regression coefficients on age using NGC (b) ten-fold cross-prediction using GC and NGC

originated from caudate, implicating a significant role it plays in the age related connectommic alternation. Previous studies reported age related reductions in caudate volume (Jerniganemail and Gamst, 2005; Abedelahi and Hasanzadeh, 2013), thus lending evidence that caudate may subject to severe influence of the aging process. Also the connectivity pattern of the motor areas and somatosensory cortex are also altered. The cingulum also exhibits some age related changes. And the frontal mid regions are subject to the changes too. We present in Figure 8 the mean negative $\log_{10}$ p-value of the connection from left and right caudate to all other AAL brain regions.

To further validate the effectiveness of NGC in extracting informative features, we compare the out-of-sample prediction performance of GC and NGC. We randomly split the original sample into ten partitions with similar size. Each time we pick one partition as the testing sample, and the rest as the training sample. Two sample T-test is performed on the training sample, and the top 20 functional connections were selected. We use MATLAB to fit a Logistic Lasso with 10-fold cross-validation using these 20 features. The parameters with minimal deviance on the training sample were used for label prediction on the testing sample. We picked each partition in turn and by comparing the

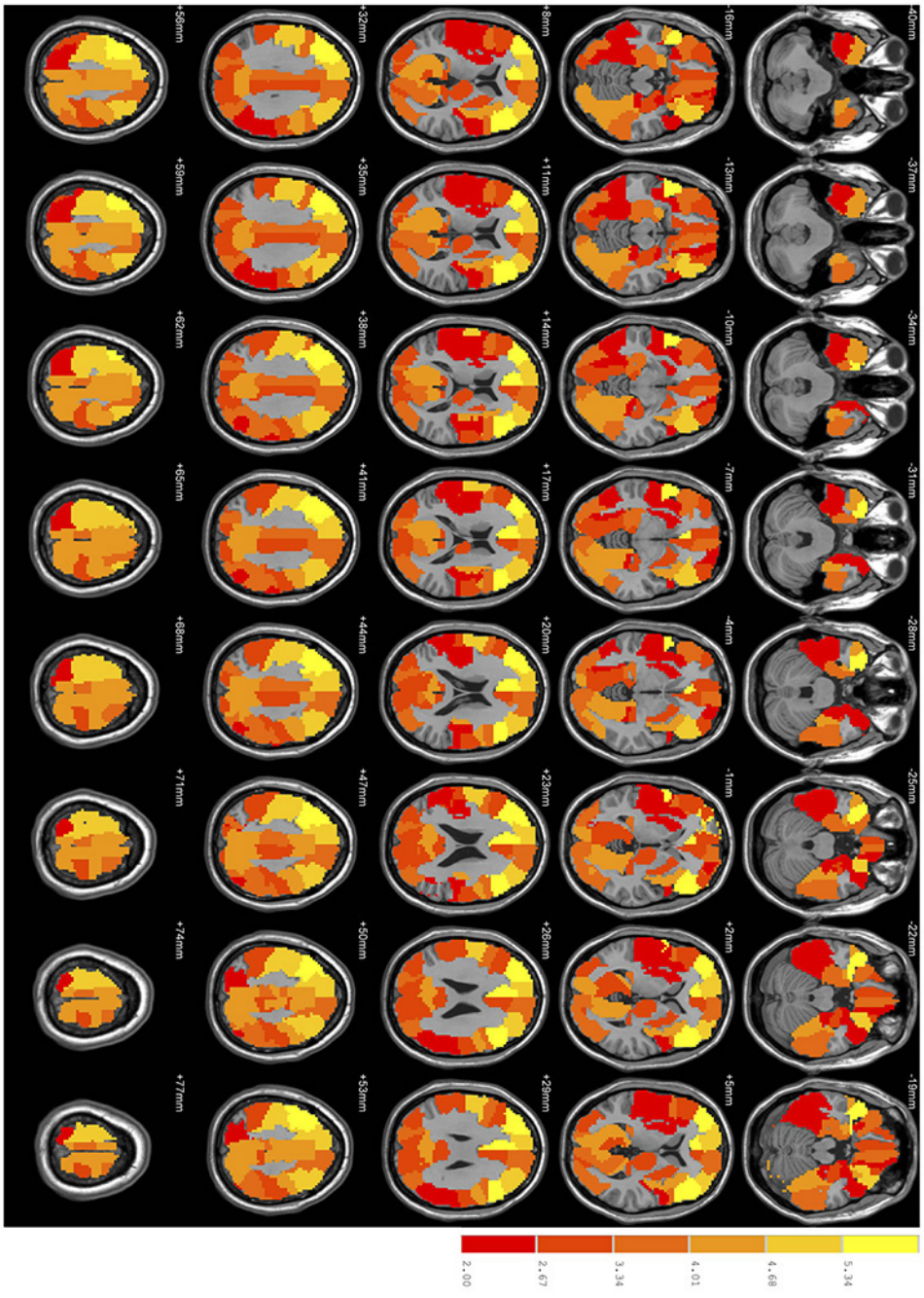


Figure 8: Caudate connection to the rest of the brain. The cerebrum is parcellated into 90 regions based on AAL template.

ground truth labels of testing sample we obtained an estimate of out-of-sample prediction performance. This procedure was repeated for 100 times to estimate the stability. The results are shown in Figure 7(b). NGC achieved more than 10% boost in prediction accuracy compared with GC.

4.2. Neuronal network reconstruction

We also use the synthetic Calcium imaging dataset from ChaLearn 2014 Connectomics Challenge (Orlandi et al., 2014) in this study. The data was generated from a realistic model that is extensively studied and validated (Stetter et al., 2012). The defects of the calcium fluorescence imaging technology is also simulated: limited time resolution and light scattering artifacts. Four training sets (normal-1 to normal-4) each containing 1 hour recording of $1,000$ neurons at the sampling rate of 59 Hz is provided by the challenge committee along with their respective ground truth synaptic connection. The reconstruction requires a ranking of all connections reflecting the confidence that there is a directed neuronal connection. The prediction results are evaluated with the area under the receiver operating characteristic curve (ROC), here after referred to as AUC score. The first principal component of the raw Calcium fluoresce signal is identified as synchronized dynamics of the network. The OOPSI filter (Vogelstein et al., 2010) is then applied to the synchronized dynamics to inference the spiking times. The differenced Calcium fluorescence signal is used as the proxy for firing rate within a time-bin. To reduce the dimensionality of the problem we only consider the firing rate at the spiking times (roughly 600 data points).

The synthetic Calcium imaging dataset is used to compare the performance of network reconstruction using cross correlation and NAC. The results are tabulated in Table. 3 for zero-lag and one-lag cases. While NAC provides similar, if not slightly better performance in zeros-lag, it significantly outperforms linear correlation in the one-lag scenario. All the best performance is achieved using multiple eigenvalues, which further validates the necessity of characterizing association using appropriate number of eigenvalues. The joint distribution of pre/post-synaptic neuron's differenced Calcium fluorescence signal (reflecting

firing rates within the sampling period) is given in Figure 9. X-axis is the signal intensity for the pre-synaptic neuron and Y-axis is the signal intensity for the post-synaptic neuron. Upper panel are for the neuron pairs with a direct connection and lower panel for the unconnected pairs. Left column is zero-lag ($x(t)$ vs $y(t)$) and right column is one-lag ($x(t)$ vs $y(t+1)$). As can be seen, the probability mass of zero-lag dynamics exhibits a highly linear trend (probability mass concentrates along the diagonal), which could be well captured both by linear correlation and NAC. This explains their similar performance in the zero-lag case. However, it exhibits highly nonlinear trend in the one-lag case (probability mass concentrates in a highly curved domain), where correlation fails. However such deviation from linearity has much less impact on the performance of NAC. This demonstrates NAC is more suitable for cases where the linearity assumption is not guaranteed. We note that the NAC result is by no means the best reconstruction result as we deliberately chose the dynamical regimes previously regarded as less informative when performing reconstructions (Stetter et al., 2012). And most of the best performing reconstruction methods (Orlandi et al., 2014) are hybrid methods that combine multi-modal information and including NAC might boost their performance further.

4.3. Other two applications

In addition to the two examples above, NAC has also found successful applications in two separate research projects. The results have been reported in separate articles so we only give a brief description here, interested readers may refer to the respective article for details. In (Cheng et al., 2015) we used NAC to quantify the dependency between altered functional connectivity links and the autism scores using data from the autism brain imaging data exchange repository (ABIDE; `http://fcon_1000.projects.nitrc.org/indi/abide/`) hosted by the 1000 Functional Connectome Project/International Neuroimaging Data-sharing Initiative (INDI), although NAC is termed KGV in that paper. In (Ojelade et al., 2015), NAC was used to carry out an imaging-genetic study. The association between a candidate gene and ethanol addiction from animal

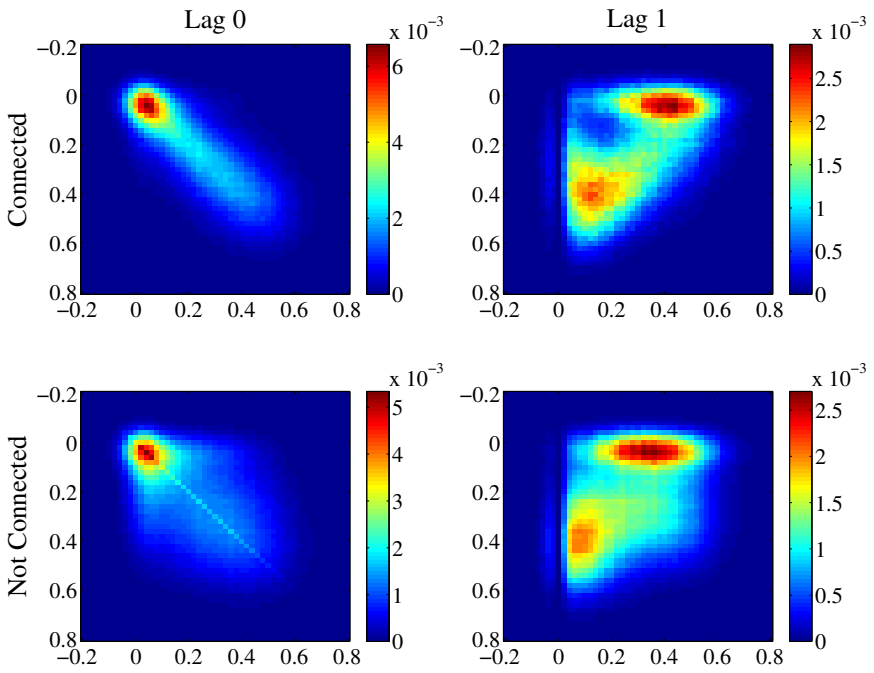


Figure 9: Heatmap of the presynaptic neuron vs postsynaptic neuron firing intensity density. x-axis presynaptic, y-axis postsynaptic. Upper row, neuron-pairs directly connected; lower row, neuron pairs not directly connected; left column, zero lag; right column, 20-ms lag (one sampling interval lag). Neuron pairs not directly connected show stronger nonlinearity.

lag 0	**corr**	**NAC(1)**	**NAC(2)**	**NAC(3)**
normal-1	0.89497	0.89530	**0.89627**	0.89613
normal-2	0.89016	0.89201	**0.89324**	0.89292
normal-3	0.89576	0.89602	0.89750	**0.89752**
normal-4	0.88841	0.88909	**0.88990**	0.88952
lag 1	**corr**	**NAC(1)**	**NAC(2)**	**NAC(3)**
normal-1	0.55005	0.70944	0.73249	**0.74996**
normal-2	0.53640	0.71083	0.73611	**0.75471**
normal-3	0.56430	0.70796	0.73901	**0.76270**
normal-4	0.53560	0.70606	0.72825	**0.74582**

Table 3: AUC scores of connectomics reconstruction with Pearson correlation(corr) and NAC(k), $k = 1, 2, 3$ for the four test dataset at lag 0 and 1. Maximum AUC scores are marked in bold.

study is verified by NAC across multiple human GWAS datasets. NAC also found association between that gene and the neuroimaging traits in the reward circuit thus shedding light on how it affects the addiction. In both studies the proposed NAC gives biologically meaningful results.

5. Conclusion

In this study we detailed a nonlinear kernel association measure and presented a practical recipe along with a MATLAB toolbox. Via a series of simulation studies we demonstrated its superior performance against its linear or nonlinear competitors. Finally, we benchmarked its performance over two real-world examples. In the high temporal resolution fMRI study the features selected by NAC turned out to be more informative than the ones selected by GC; this observation is further confirmed by the connectomics reconstruction study, as NAC achieves better reconstruction accuracy, especially in nonlinear regimes. The other two applications again validate the use of NAC in neuroimaging studies. All these provide evidences that nonlinear association measures should be

preferred in neuroimaging studies. And we note that with our computation recipe applied, NAC could be calculated in linear time, thus making it attractive even for large datasets. For example, in Ojelade et al. (2015) the largest sample we analyzed with NAC includes $4,604$ subjects.

The nonlinear association measure we adopted and the results obtained have implications more generally for the manner in which neuroscientists make inferences using their fMRI data. Currently most of analyses in the literature draw their conclusions solely based on only linear correlation, and they may have missed many significant associations. Also inconsistent even conflicting results reported using linear methods could, perhaps, be better explained by nonlinear interactions. In the case of strong nonlinear effects the validity of those conclusions obtained by linear models are questionable.

In future work, we will extend our framework in a number of directions. Firstly we will further extend this work to perform brain-wide kernel association studies. While our approach is more general than the existing kernel brain-wide association approach (Ge et al., 2012, 2015a,b), it comes with the price of a computational burden. A more computationally feasible strategy should be developed to address the challenge, possibly utilizing the low-dimensional structure of the brain rather than a brute force massive univariate approach (Zhu et al., 2014). Additionally, another pending issue is to determine the proper threshold for the resulting brain-wide statistic map; we are working with *Gamma* distributions and currently there is no existing result for a *Gamma* Random Field. Although we can approximate the *Gamma* distribution with a χ^2 distribution, the validity needs to be confirmed via further numerical investigations. Other than those aforementioned aspects, the result of smoothness regarding the random field no longer applies. As we are working in high dimensional domain, different features might have different smoothness. How to mitigate the effect of inconsistent smoothness of different random variables remains to be explored. Secondly, we plan to look into alternatives that extend single kernel association to multi-kernel association. Multi-kernel learning has found successful applications in computational vision and bioinformatics (Bucak

et al., 2014). Studies have found that a multi-kernel approach better integrates information especially for features extracted from different modalities compared with single kernel approach. Recently, Stein-type shrinkage estimators of the cross-covariance operator have been proposed to improve the finite-sample performance Muandet et al. (2013); Ramdas and Wehbe (2014). We note these estimators belong to the framework of spectrum regularization. These estimators performs a certain 'soft-thresholding' (i.e. shrinking all eigenvalues) to regularize the empirical operator while NAC takes the 'hard-thresholding' (i.e. cutoff at k) approach. Nevertheless, plug-in these regularized estimators may further boost the sensitivity on small samples.

Acknowledgements

All of the authors have no conflict of interest. CY Tao is supported by the China Scholarship Council (CSC). JF Feng is a Royal Society Wolfson Research Merit Award holder, partially supported by the Key Program of National Natural Science Foundation of China (NSFC No. 91230201), National Centre for Mathematics and Interdisciplinary Sciences (NCMIS) in the Chinese Academy of Sciences, and National High Technology Research and Development Program of China (863) under grant No. 2015AA020507. The authors would also like to thank Prof. D Waxman for proof reading; Dr. XX Ji, Dr. B Xu, Dr. TY Jia, Dr. W Cheng for their valuable feedback using NAC in research; Prof. W Lin, Prof. WL Lu for their advices; Prof. K Fukumizu, Dr. L Zhao, Dr. WB Sheng, Dr. XJ Gan, Dr. Y Guo and Dr. XN He for fruitful discussions.

References

Abedelahi, A., Hasanzadeh, H., 2013. Mri based morphometry of caudate nucleus in normal persons. Journal of Paramedical Sciences 4 (2).

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., Gitelman, D., Kiebel, S., Kilner, J., Litvak, V., et al., 2012. Spm8 manual. Functional Imaging Laboratory, Institute of Neurology.

Ashrafulla, S., Haldar, J. P., Joshi, A. A., Leahy, R. M., 2013. Canonical granger causality between regions of interest. NeuroImage 83, 189–199.

Bach, F. R., Jordan, M. I., 2002. Learning graphical models with mercer kernels. In: Advances in Neural Information Processing Systems. Vol. 15. MIT Press, pp. 1009–1016.

Bach, F. R., Jordan, M. I., 2003. Kernel independent component analysis. The Journal of Machine Learning Research 3, 1–48.

Barrett, A. B., Barnett, L., Seth, A. K., 2010. Multivariate granger causality and generalized variance. Physical Review E 81 (4), 041907.

Bucak, S. S., Jin, R., Jain, A. K., 2014. Multiple kernel learning for visual object recognition: A review. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36 (7), 1354–1369.

Buchanan, M., 2012. Cause and correlation. Nature Physics 8 (12), 852.

Chao-Gan, Y., Yu-Feng, Z., 2010. Dparsf: a matlab toolbox for pipeline data analysis of resting-state fmri. Frontiers in Systems Neuroscience 4, 13.

Chen, Y., Rangarajan, G., Feng, J., Ding, M., 2004. Analyzing multiple nonlinear time series with extended granger causality. Physics Letters A 324 (1), 26–35.

Cheng, W., Rolls, E. T., Gu, H., Zhang, J., Feng, J., 2015. Autism: reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self. Brain 138 (5), 1382–1393.

Coles, S., Bawa, J., Trenner, L., Dorazio, P., 2001. An Introduction to Statistical Modeling of Extreme Values. Vol. 208. Springer.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., Initiative, A. D. N., et al., 2011. Automatic classification of patients with alzheimer's disease from structural

mri: a comparison of ten methods using the adni database. NeuroImage 56 (2), 766–781.

Dauxois, J., Nkiet, G. M., 1998. Nonlinear canonical analysis and independence tests. The Annals of Statistics 26 (4), 1254–1278.

Dong, L., Zhang, Y., Zhang, R., Zhang, X., Gong, D., Valdes-Sosa, P. A., Xu, P., Luo, C., Yao, D., 2015. Characterising nonlinear relationships in functional imaging data using eigenspace maximal information canonical correlation analysis (emicca). NeuroImage 109, 388–401.

Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Vol. 1. Cambridge University Press.

Feuerverger, A., 1993. A consistent test for bivariate dependence. International Statistical Review/Revue Internationale de Statistique, 419–433.

Fukumizu, K., Bach, F. R., Gretton, A., 2007a. Statistical consistency of kernel canonical correlation analysis. The Journal of Machine Learning Research 8, 361–383.

Fukumizu, K., Bach, F. R., Jordan, M. I., 2004. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. The Journal of Machine Learning Research 5, 73–99.

Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B., 2007b. Kernel measures of conditional dependence. In: Advances in Neural Information Processing Systems. Vol. 20. MIT Press, pp. 489–496.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., Nichols, T. E., 2012. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. NeuroImage 63 (2), 858–873.

Ge, T., Nichols, T. E., Ghosh, D., Mormino, E. C., Smoller, J. W., Sabuncu, M. R., Initiative, A. D. N., et al., 2015a. A kernel machine method for detect-

ing effects of interaction between multidimensional variable sets: An imaging genetics application. NeuroImage 109, 505–514.

Ge, T., Nichols, T. E., Lee, P. H., Holmes, A. J., Roffman, J. L., Buckner, R. L., 2015b. Massively expedited genome-wide heritability analysis (megha). Proceedings of the National Academy of Sciences of the United States of America 112 (8), 2479–2484.

Gebelein, H., 1941. Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. Zamm-zeitschrift Fur Angewandte Mathematik Und Mechanik 21, 364–379.

Gong, X., Lu, W., Kendrick, K. M., Pu, W., Wang, C., Jin, L., Lu, G., Liu, Z., Liu, H., Feng, J., 2014. A brain-wide association study of disc1 genetic variants reveals a relationship with the structure and functional connectivity of the precuneus in schizophrenia. Human Brain Mapping 35 (11), 5414–5430.

Granger, C. W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, 424–438.

Gretton, A., Fukumizu, K., Sriperumbudur, B. K., 12 2009. Discussion of: Brownian distance covariance. The Annals of Applied Statistics 3 (4), 1285–1294.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., Smola, A. J., 2007. A kernel statistical test of independence. In: Advances in Neural Information Processing Systems. Vol. 20. MIT Press, pp. 585–592.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B., 2005. Kernel methods for measuring independence. The Journal of Machine Learning Research 6, 2075–2129.

Henderson, H. V., Searle, S. R., 1981. On deriving the inverse of a sum of matrices. SIAM Review 23 (1), 53–60.

Hirschfeld, H. O., Wishart, J., 1935. A Connection between Correlation and Contingency. Mathematical Proceedings of The Cambridge Philosophical Society 31.

Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., Bhattacharya, J., 2007. Causality detection based on information-theoretic approaches in time series analysis. Physics Reports 441 (1), 1–46.

Hua, W.-Y., Ghosh, D., 2014. Equivalence of kernel machine regression and kernel distance covariance for multidimensional trait association studies. arXiv preprint arXiv:1402.2679.

Huang, S.-Y., Lee, M.-H., Hsiao, C. K., 2006. Kernel canonical correlation analysis and its applications to nonlinear measures of association and test of independence. available online.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Networks 13 (4), 411–430.

J Reddi, S., Póczos, B., 2013. Scale invariant conditional dependence measures. In: Dasgupta, S., McAllester, D. (Eds.), Proceedings of the 30th International Conference on Machine Learning. pp. 1355–1363.

Jerniganemail, T. L., Gamst, A. C., 2005. Changes in volume with age-consistency and interpretation of observed effects. Neurobiology of Aging 26 (9), 1271–1274.

Kankainen, A., 1995. Consistent testing of total independence based on the empirical characteristic function. Vol. 29. University of Jyväskylä.

Kantz, H., Schreiber, T., 2004. Nonlinear Time Series Analysis. Vol. 7. Cambridge University Press.

Keener, J., Sneyd, J., 2010. Mathematical Physiology I: cellular physiology. Vol. 1. Springer Science & Business Media.

Kendrick, K. M., Zhan, Y., Fischer, H., Nicol, A. U., Zhang, X., Feng, J., 2011. Learning alters theta amplitude, theta-gamma coupling and neuronal synchronization in inferotemporal cortex. BMC neuroscience 12 (1), 55.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Physical review E 69 (6), 066138.

Li, Q., Racine, J. S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press.

Liu, D., Lin, X., Ghosh, D., 2007. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics 63 (4), 1079–1088.

Liu, Y., Qiao, N., Zhu, S., Su, M., Sun, N., Boyd-Kirkup, J., Han, J.-D. J., 2013. A novel bayesian network inference algorithm for integrative analysis of heterogeneous deep sequencing data. Cell Research 23 (3), 440.

Marinazzo, D., Pellicoro, M., Stramaglia, S., 2008. Kernel method for nonlinear granger causality. Physical Review Letters 100 (14), 144103.

McDonald, J. H., 2009. Handbook of Biological Statistics. Vol. 2. Sparky House Publishing, Baltimore, Maryland.

McIntosh, A. R., Lobaugh, N. J., 2004. Partial least squares analysis of neuroimaging data: applications and advances. NeuroImage 23, S250–S263.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Gretton, A., Schölkopf, B., 2013. Kernel mean estimation and stein's effect. arXiv preprint arXiv:1306.0842.

Nooner, K. B., Colcombe, S. J., Tobe, R. H., Mennes, M., Benedict, M. M., Moreno, A. L., Panek, L. J., Brown, S., Zavitz, S. T., Li, Q., et al., 2012. The nki-rockland sample: a model for accelerating the pace of discovery science in psychiatry. Frontiers in Neuroscience 6, 152.

Ojelade, S., Jia, T., et al., Gunter, S., 2015. Rsu1 regulates ethanol consumption in drosophila and humans. Proceedings of the National Academy of Sciences of the United States of America 112 (30), E4085E4093.

Orlandi, J. G., Ray, B., Battaglia, D., Guyon, I., Lemaire, V., Saeed, M., Soriano, J., Statnikov, A., Stetter, O., 2014. First connectomics challenge: From imaging to connectivity. available online.

Póczos, B., Schneider, J. G., 2012. Nonparametric estimation of conditional information and divergences. In: International Conference on Artificial Intelligence and Statistics. pp. 914–923.

Ramdas, A., Wehbe, L., 2014. Stein shrinkage for cross-covariance operators and kernel independence testing. arXiv preprint arXiv:1406.1922.

Rényi, A., 1959. On measures of dependence. Acta Mathematica Academiae Scientiarum Hungarica 10 (3-4), 441–451.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., Sabeti, P. C., 2011. Detecting novel associations in large data sets. Science 334 (6062), 1518–1524.

Richard, A. J., Dean, W. W., 2002. Applied Multivariate Statistical Analysis. Prentice Hall, New York.

Saad, Y., 1992. Numerical Methods for Large Eigenvalue Problems. Vol. 158. SIAM.

Scholkopf, B., Smola, A. J., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., et al., 2013. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. The Annals of Statistics 41 (5), 2263–2291.

Shannon, C. E., 1948. A mathematical theory of communication. Bell System Technical Journal 27 (3), 379–423.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel methods for pattern analysis. Cambridge University Press.

Song, L., Smola, A., Gretton, A., Bedo, J., Borgwardt, K., 2012. Feature selection via dependence maximization. The Journal of Machine Learning Research 98888 (1), 1393–1434.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M., Bedo, J., 2007. Supervised feature selection via dependence estimation. In: Ghahramani, Z. (Ed.), Proceedings of the 24th International Conference on Machine Learning. ACM, Omni Press, pp. 823–830.

Stark, J., 2001. Delay reconstruction: Dynamics versus statistics. In: Mees, A. I. (Ed.), Nonlinear Dynamics and Statistics. Springer, pp. 81–103.

Stephan, K. E., Kasper, L., Harrison, L. M., Daunizeau, J., den Ouden, H. E., Breakspear, M., Friston, K. J., 2008. Nonlinear dynamic causal models for fmri. NeuroImage 42 (2), 649–662.

Stetter, O., Battaglia, D., Soriano, J., Geisel, T., 2012. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. PLoS Computational Biology 8 (8), e1002653.

Stögbauer, H., Kraskov, A., Astakhov, S. A., Grassberger, P., 2004. Least-dependent-component analysis based on mutual information. Physical Review E 70 (6), 066123.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., Munch, S., 2012. Detecting causality in complex ecosystems. Science 338 (6106), 496–500.

Sun, X., 2008. Assessing nonlinear granger causality from multivariate time series. In: Machine Learning and Knowledge Discovery in Databases. Springer, pp. 440–455.

Székely, G. J., Rizzo, M. L., et al., 2009. Brownian distance covariance. The Annals of Applied Statistics 3 (4), 1236–1265.

Tikhonov, A. N., Leonov, A. S., Yagola, A. G., 1998. Nonlinear Ill-posed Problems. Chapman & Hall, London.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. NeuroImage 15 (1), 273–289.

Viviani, R., Grön, G., Spitzer, M., 2005. Functional principal component analysis of fmri data. Human Brain Mapping 24 (2), 109–129.

Vogelstein, J. T., Packer, A. M., Machado, T. A., Sippy, T., Babadi, B., Yuste, R., Paninski, L., 2010. Fast nonnegative deconvolution for spike train inference from population calcium imaging. Journal of Neurophysiology 104 (6), 3691–3704.

Westfall, P. H., 1993. Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment. Vol. 279. John Wiley & Sons.

Wiener, N., 1956. The theory of prediction. In: Beckenbach, E. F. (Ed.), Modern Mathematics for Engineers. 1. McGraw-Hill: New York, NY, USA, Ch. 8, pp. 165–183.

Wu, G., Duan, X., Liao, W., Gao, Q., Chen, H., 2011. Kernel canonical-correlation granger causality for multiple time series. Physical Review E 83 (4), 041921.

Zhu, H., Khondker, Z., Lu, Z., Ibrahim, J. G., 2014. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. Journal of the American Statistical Association 109 (507), 977–990.

Appendix A. Inner product on $\mathcal{H}_\Omega$

For $f, g \in \mathcal{H}_\Omega$, we have two sequences $\{f^{(m)}\}_{m=1}^{\infty}, \{g^{(m)}\}_{m=1}^{\infty}$, such that $f^{(m)}, g^{(m)} \in \mathcal{H}_{\Omega,\kappa}^{\mathcal{LS}}$ and $f = \lim_{m\to\infty} f^{(m)}$, $g = \lim_{m\to\infty} g^{(m)}$. First we define

the inner product on $\mathcal{H}^{\mathcal{LS}}_{\Omega,\kappa}$. For

$$f^{(m)}(\cdot) = \sum_{i=1}^{n_{f,m}} \alpha_i \kappa(\cdot, \omega_i^{f,m}), g^{(m)}(\cdot) = \sum_{i=1}^{n_{g,m}} \beta_i \kappa(\cdot, \omega_i^{g,m}), \tag{A.1}$$

we write matrix $K^{(m)} \in \mathbb{R}^{n_{f,m} \times n_{g,m}}$ as

$$(K^{(m)})_{ij} := \kappa(\omega_i^{f,m}, \omega_j^{g,m}), i \in [1, \cdots, n_{f,m}], j \in [1, \cdots, n_{g,m}]. \tag{A.2}$$

The inner product between $f^{(m)}$ and $g^{(m)}$ is given by

$$\langle f^{(m)}, g^{(m)} \rangle_{\mathcal{H}^{\mathcal{LS}}_{\Omega,\kappa}} := \boldsymbol{\alpha}^T K^{(m)} \boldsymbol{\beta}, \tag{A.3}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_{n_{f,m}}\}^T$ and $\boldsymbol{\beta} = \{\beta_1, \cdots, \beta_{n_{g,m}}\}^T$. Taking m to infinity, we have the inner product on $\mathcal{H}_\Omega$ as

$$\langle f, g \rangle_{\mathcal{H}_{\Omega,\kappa}} := \lim_{m \to \infty} \langle f^{(m)}, g^{(m)} \rangle_{\mathcal{H}^{\mathcal{LS}}_{\Omega,\kappa}}. \tag{A.4}$$

Appendix B. Alternative derivation of NAC

Suppose we have two sets of random vectors, $X \in \mathbb{R}^{d_1}$ and $Y \in \mathbb{R}^{d_2}$, where X can be further decomposed as $X = (U, V)$, where $U \in \mathbb{R}^{d_{11}}, V \in \mathbb{R}^{d_{12}}$ and $d_1 = d_{11} + d_{12}$. We want to determine whether Y and V are independent conditioned on U, i.e.

$$p_{Y|U,V}(y|u, v) = p_{Y|U}(y|u). \tag{B.1}$$

We will also interpret this from an information theoretical perspective. We denote $I(X;Y)$ as the mutual information shared by X and Y and $I(X;Y|U)$ the conditional mutual information. The decomposition of mutual information between X and Y could be written as

$$I(Y;X) = I(Y;U) + I(Y;V|U) \tag{B.2}$$

So if U contains all the relevant information shared with Y, e.g. $I(Y,X) = I(Y,U)$, then we have $I(Y;V|U = \cdot) = 0$ almost surely, i.e. Y and V are independent a.s. given U. However, to carry out entropy based statistics requires the estimation of the probability density function, which is prohibitively data

consuming in a high dimensional space, unless certain restrictive constraints are imposed on the distribution, e.g. the candidate distributions are in certain parametric family. Thus such information theoretic measures are only applicable to low dimensional and data intensive problems.

To overcome the difficulties encountered using the mutual information criteria to characterize independency, we must seek some new mathematical tools. As it turned out the cross-covariance operators on RKHS offer a nice solution. Let us denote $\mathcal{H}_X = (\Omega_X, \kappa_X)$ as a RKHS defined on a domain Ω_X with kernel function $\kappa_X(\cdot,\cdot)$ and similarly for $\mathcal{H}_Y$. The most important property of RKHS is the so-called reproducing kernel property:

$$\langle f, \kappa_X(\cdot, x)\rangle = f(x), \text{for } f \in \mathcal{H}_X \tag{B.3}$$

Now we define the cross-covariance operator between two RKHSs. For a random vector (X,Y) defined on $\Omega_X \times \Omega_Y$ and RKHS $(\mathcal{H}_X, \mathcal{H}_Y)$, the cross-covariance operator Σ_{XY} is then defined by

$$\langle g, \Sigma_{XY} f\rangle_{H_Y} := \text{cov}(f(X), g(Y)), \text{ for any } \ f \in \mathcal{H}_X \text{ and } g \in \mathcal{H}_Y. \tag{B.4}$$

Using Theorem 1 and Corollary 2 in Fukumizu et al. (2004), we know that

$$\mathbb{E}_{Y|X}[g(Y)|X = x] = \left(\Sigma_{XX}^{-1}\Sigma_{XY} g\right)(x), \text{ for all } g \in \mathcal{H}_Y \text{ and } x \in X \tag{B.5}$$

To better understand this, see the simplest case where (X,Y) are joint Gaussian random variables, $\alpha \in \mathcal{R}^{d_2}$ and choose κ_X, κ_Y to be Euclidian inner product. This gives us

$$\mathbb{E}_{Y|X}[\alpha^T Y|X = x] = x^T \Sigma_{XX}^{-1}\Sigma_{XY}\alpha \tag{B.6}$$

The cross covariance operator maps α to a functional $\Sigma_{XX}^{-1}\Sigma_{XY}\alpha$ on $\mathcal{R}^{d_1}$. Now we define the conditional covariance operator as

$$\Sigma_{YY|U} = \Sigma_{YY} - \Sigma_{YU}\Sigma_{UU}^{-1}\Sigma_{UY} \tag{B.7}$$

Theorem 3 in Fukumizu et al. (2004) gives us that under some mild conditions,

$$\langle g, \Sigma_{YY|U} f\rangle_{H_Y} = \mathbb{E}_U[\text{cov}_{Y|U}(f(Y), g(Y)|U)] \tag{B.8}$$

So take $f = g$, the above equation becomes

$$\langle f, \Sigma_{YY|U} f \rangle_{H_Y} = \mathbb{E}_U[\mathrm{var}_{Y|U}(f(Y)|U)] \tag{B.9}$$

Thus the operator $\Sigma_{YY|U}$ measures the residual variance when the information contained in U is removed. In analogy to the Gaussian variables, that gives

$$\langle \alpha^T Y, \Sigma_{YY|U} \alpha^T Y \rangle_{H_Y} = \mathbb{E}_U[\mathrm{var}_{Y|U}(\alpha^T Y|U)] = \alpha^T \Sigma_{YY|U} \alpha \tag{B.10}$$

Heuristically if all the information in Y is already contained in U, then we have $\langle f, \Sigma_{YY|U} f \rangle_{\mathcal{H}_Y}$=0. Theorem 4 and 5 in Fukumizu et al. (2004) give mathematically rigorous statement of this intuition. Based on Theorem 5 in Fukumizu et al. (2004) , for an arbitrary random vector U, if we can find a contrast function h

$$h(Y,U): \qquad (Y \times U, \mathbb{P}_{Y\times U}) \to \mathbb{R} \tag{B.11}$$

$$\Sigma_{YY|U} \mapsto h(\Sigma_{YY|U}) \tag{B.12}$$

such that h preserves the order of $\Sigma_{YY|U}$ in probability, i.e. for a finite sample $D(N)$ of random vectors $\{Y, U_1, U_2\}$ with sample size N, satisfying $\Sigma_{YY|U_1} \geq \Sigma_{YY|U_2}$, it holds

$$\lim_{N\to\infty} \mathbb{P}_{D(N)} \left[h(\hat{\Sigma}_{YY|U_1}) \geq h(\hat{\Sigma}_{YY|U_2}) \right] \to 1, \tag{B.13}$$

where $\hat{\Sigma}_{YY|U_i}, i = 1, 2$ are the empirical estimate of covariance operator, then we can quantify the independency between Y and W using h by comparing $h(\Sigma_{YY|U})$ and $h(\Sigma_{YY})$. The freedom to choose h provides a range of independency criteria, popular choices of h include the trace, the first eigenvalue and the determinant, etc. Barrett et al. (2010) In this paper we focus on the determinant. Using the equality

$$\det(A - BC^{-1}B^T) = \frac{\det\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}}{\det(C)}, \tag{B.14}$$

the contrast that we are evaluating is reduced to

$$\det(\Sigma_{YY|U}) = \frac{\det\begin{pmatrix} \Sigma_{YY} & \Sigma_{YU} \\ \Sigma_{UY} & \Sigma_{UU} \end{pmatrix}}{\det(\Sigma_{UU})}. \tag{B.15}$$

960 A symmetrical version of the objective function is obtained by augment it with $\det(\Sigma_{YY})$ term and taken the negative of log transform

$$h_d(Y,U) = -\log \frac{\det\begin{pmatrix} \Sigma_{YY} & \Sigma_{YU} \\ \Sigma_{UY} & \Sigma_{UU} \end{pmatrix}}{\det(\Sigma_{YY})\det(\Sigma_{UU})}. \tag{B.16}$$

Actually, the symmetrized objective function corresponds to the determinant of the normalized conditional cross-covariance operator $V_{YY|U}$ defined in Fukumizu et al. (2007b); Sun (2008) and $V_{YY|U} = O$ does correspond to independence under

965 certain mild conditions.

Before we proceed further, we revisit the connection between h_d and mutual information. First note that for multivariate Gaussian random vectors (X,Y), with covariance matrix

$$C = \begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix} \tag{B.17}$$

the mutual information between X and Y is

$$I(X;Y) = -\frac{1}{2}\log \frac{\det\begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix}}{\det(C_{XX})\det(C_{YY})}. \tag{B.18}$$

970 So $h_d(Y,U)$ is just the mutual information upto a constant factor in the case of multivariate Gaussian variables. As sketched in Bach and Jordan (2003), $h_d(Y,U)$ actually gives the second order approximation to the true mutual information near independence when Y,U are both univariate random variables, where the higher order terms vanishes in the Gaussian case.

975 Next we express $h_d(Y,U)$ as the generalized eigenvalue problem. Consider the following generalized eigenvalue problem

$$\begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1+\rho)\begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{B.19}$$

where (α, β) is eigenvector and ρ is the eigenvalue. The above generalized eigenvalue problem could be reformulated as

$$\begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix}^{-1} \begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1+\rho) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{B.20}$$

Using linear algebra the determinant of the left hand side of the matrix above can be rewritten as

$$\det\left(\begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix}^{-1} \begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix} \right) = \frac{\det \begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix}}{\det(C_{XX}) \det(C_{YY})} = \prod_{i=1}^{d_1+d_2} (1+\rho_i) \tag{B.21}$$

Suppose we have drawn N samples $\{(x_i, y_i)\}_{i=1}^N$ from (X, Y) and define K_X, K_Y as

$$(K_X)_{i,j} \coloneqq \kappa_X(x_i, x_j), \quad (K_Y)_{i,j} \coloneqq \kappa_Y(y_i, y_j), i, j \in [1, \cdots, N] \tag{B.22}$$

K is also known as the Gram matrix in the literature. Then the empirical estimate of the conditional covariance operator $\Sigma_{YY|X}$ is

$$\Sigma_{YY|X} = \frac{1}{N}\left(K_{YY} - K_{YX} K_{XX}^{-1} K_{XY}\right), \tag{B.23}$$

where $K_{XX} \coloneqq K_X K_X, K_{YY} \coloneqq K_Y K_Y, K_{XY} = (K_{YX})^T \coloneqq K_X K_Y$. In analogy to the Gaussian case, we could normalize by $\det(\Sigma_{YY})$ to symmetrize the statistic, which gives the following eigenvalue problem

$$\begin{pmatrix} K_{XX} & 0 \\ 0 & K_{YY} \end{pmatrix}^{-1} \begin{pmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1+\rho) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{B.24}$$

since $\det(\Sigma_{XY}) = \left(\frac{1}{N}\right)^N \det(K_{YY} - K_{YX} K_{XX}^{-1} K_{XY})$ and $\det(\Sigma_{YY}) = \left(\frac{1}{N}\right)^N \det(K_{YY})$. Using the symmetry of K_X and K_Y we know that the ρ_i in Eq. (B.24) comes in pairs, i.e. if $1+\rho_i$ is the eigenvalue of Eq. (B.24), then so is $1-\rho_i$. Ordering the eigenvalues in a descending order, $\{\rho_i\}_{i=1}^N$ where $1 \geq \rho_1 \geq \cdots \geq \rho_N \geq 0$, we approximate the mutual information between X and Y as

$$\mathrm{NAC}(X; Y|N) = -\frac{1}{2} \sum_{i=1}^{N} \log(1-\rho_i^2) \tag{B.25}$$

We note that ρ_i vanishes as i goes to N, thus we define our $\mathrm{NAC}(X;Y|N)$ by truncating Eq. (B.25) to the first k leading terms

$$\mathrm{NAC}(X;Y|k) \coloneqq -\frac{1}{2}\sum_{i=1}^{k}\log(1-\rho_i^2) \tag{B.26}$$

Appendix C. Connections to existing HSIC variants

In this section we establish the link between NAC and existing HSIC variants. The normalized cross-covariance operator is defined to be

$$V_{XY} \coloneqq \Sigma_{XX}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}^{-\frac{1}{2}} \tag{C.1}$$

and the conditional cross-covariance operator is defined as

$$V_{XY|Z} \coloneqq V_{XY} - V_{XZ}V_{ZY} \tag{C.2}$$

thus by plugin the definition of V_{XY} we have

$$\begin{aligned} V_{YY|X} &= V_{YY} - V_{YX}V_{XY} && \text{(C.3)}\\ &= \Sigma_{YY}^{-\frac{1}{2}}\left(\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right)\Sigma_{YY}^{-\frac{1}{2}} && \text{(C.4)} \end{aligned}$$

It is obvious that by replacing the cross-covariance operators $\Sigma_{XX}, \Sigma_{XY}, \Sigma_{YY}$ with their empirical estimate K_{XX}, K_{XY}, K_{YY}, we immediately have

$$-\frac{1}{2}\log\det(V_{YY|X}) = \mathrm{NAC}(X;Y|N) \tag{C.5}$$

The Hilber-Schmidt norm of operator $\mathcal{A}_{XY}$ is defined as

$$\|\mathcal{A}_{XY}\|_{HS}^2 \coloneqq \sum_i \langle\phi_i, \mathcal{A}_{XY}\psi_i\rangle_X^2, \tag{C.6}$$

where the above sum is assumed to converge and $\{\phi_i\}_{i=1}^{\infty}$, $\{\psi_i\}_{i=1}^{\infty}$ are the orthonormal bases of $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively, i.e.

$$\langle\phi_i,\phi_i\rangle_X = \langle\psi_i,\psi_i\rangle_Y = 1, \tag{C.7}$$

$$\langle\phi_i,\phi_j\rangle_X = \langle\psi_i,\psi_j\rangle_Y = 0, i \neq j. \tag{C.8}$$

The norm does not depend on the choice of orthonormal bases. A well-known application of using Hilber-Schmidt norm to assess the statistical dependency between variables is HSIC defined as

$$\text{HSIC} \coloneqq \|\Sigma_{XY}\|_{HS}.$$

Assume the base pairs $\{\tilde{\phi}_i, \tilde{\psi}_i\}$ are obtained by solving the NAC (the eigencomponents of the corresponding KCCA). In analogy to multivariate CCA we could show $\{\tilde{\phi}_i, \tilde{\psi}_i\}$ satisfy

$$\langle \tilde{\phi}_i, \Sigma_{XY} \tilde{\psi}_i \rangle_X = \rho_i, \langle \tilde{\phi}_i, \Sigma_{XX} \tilde{\phi}_j \rangle_X = \langle \tilde{\psi}_i, \Sigma_{YY} \tilde{\psi}_j \rangle_Y = \delta_{ij}, \tag{C.9}$$

where $\{\rho_i\}$ are the eigenvalues of the eigenvalue problem and δ_{ij} is the Kronecker delta. By define $\breve{\phi}_i \coloneqq \Sigma_{XX}^{1/2} \tilde{\phi}_i$ and $\breve{\psi}_i \coloneqq \Sigma_{YY}^{1/2} \tilde{\psi}_i$, we have

$$\langle \breve{\phi}_i, V_{XY} \breve{\psi}_i \rangle_X = \rho_i, \langle \breve{\phi}_i, \breve{\phi}_j \rangle_X = \langle \breve{\psi}_i, \breve{\psi}_j \rangle_Y = \delta_{ij}, \tag{C.10}$$

which suggest $\{\breve{\phi}_i\}$ and $\{\breve{\psi}_i\}$ are orthonormal basis in $\mathcal{H}_X$ and $\mathcal{H}_Y$. Then $\text{NOCCO} = \|V_{XY}\|_{HS}^2 = \sum_i \rho_i^2$ Fukumizu et al. (2007b,a) and $\text{NAC}(k) = -\frac{1}{2} \sum_{i=1}^{k} \log(1 - \rho_i^2)$. When the dependence is weak ($\rho_i \sim 0$ for $i = 1, \cdots, k$, $\rho_i \ll \rho_k$ for $i > k$), by Taylor expansion we have

$$\text{NOCCO} \approx 2\,\text{NAC}(k). \tag{C.11}$$

Thus the two measures are related.

We also remark the difference in the empirical estimation between NAC and HSIC/NOCCO. NAC explicitly attempts to find the optimal bases $\{(\phi_i, \psi_i)\}_i$ via maximizing the empirical association of selected bases. Thus if no regularization is adopted it will overfit the data, the degeneracy as mentioned in main text. On the other hand, HSIC does not suffer from this drawback since it explicitly estimate the Hilbert Schmidt norm (sum of squared spectrum of cross-covariance operator) without inferring any bases. This leads to a regularization-free and easy to calculate statistics but the price to pay is that it is less sensitive in certain cases compared with NAC.

Appendix D. Regularization of NAC

In this section, we address the regularization issue by penalizing the norm of functionals in RKHS. Recall that ρ_k is the k-th largest $\mathcal{F}$-correlation of functional in H_X and H_Y and NAC(k) is calculated from $\{\rho_j\}_{j=1}^k$. Some caution is required, when the $\{\rho_j\}_{j=1}^k$ are numerically estimated. Let us take Gaussian kernels as an example. If the Gram matrix is of full rank, for example if Gaussian kernel is used, and eigenvalues are estimated without regularization, the maximum correlation coefficient between $f^{(1)}$ and $g^{(1)}$ will always be 1. To resolve this issue, we penalize the norm of f and g using a Tikhonov-type regularization Tikhonov et al. (1998) on the variance term in Eq. (5)

$$\mathcal{F}^{\epsilon}(f,g) \coloneqq \frac{\mathrm{cov}_{XY}(f,g)}{\sqrt{\left(\mathrm{var}_X(f)+\epsilon_X\|f\|^2_{\mathcal{H}_X}\right)\left(\mathrm{var}_Y(g)+\epsilon_Y\|g\|^2_{\mathcal{H}_Y}\right)}},$$

where ϵ_X and ϵ_Y are regularization parameters and $\|f\|^2_{\mathcal{H}_X}, \|g\|^2_{\mathcal{H}_Y}$ are given by

$$\|f\|^2_{\mathcal{H}_X} = \boldsymbol{\alpha}^T K_X \boldsymbol{\alpha}, \|g\|^2_{\mathcal{H}_Y} = \boldsymbol{\beta}^T K_Y \boldsymbol{\beta}.$$

We could further write the regularized variance term as

$$\begin{aligned}
\mathrm{var}_X(f)+\epsilon_X\|f\|^2_{\mathcal{H}_X} &= \frac{1}{N}\boldsymbol{\alpha}^T K_X^2\boldsymbol{\alpha}+\epsilon_X\|f\|^2 \\
&= \frac{1}{N}\boldsymbol{\alpha}^T K_X^2\boldsymbol{\alpha}+\epsilon\boldsymbol{\alpha}^T K_X\boldsymbol{\alpha} \\
&= \frac{1}{N}\boldsymbol{\alpha}^T (K_X+\frac{N\epsilon_X}{2}I)^2\boldsymbol{\alpha}+\mathcal{O}(\epsilon_X^2),
\end{aligned}$$

and similarly for $\mathrm{var}_Y(g)+\epsilon_Y\|g\|^2_{\mathcal{H}_Y}$. By omitting the infinitesimal terms $\mathcal{O}\left(\epsilon_X^2\right)$, $\mathcal{O}\left(\epsilon_Y^2\right)$ and denoting $\lambda=(\lambda_X,\lambda_Y), \lambda_X \coloneqq \frac{N\epsilon_X}{2}$, $\lambda_Y \coloneqq \frac{N\epsilon_Y}{2}$, we define

$$\mathrm{var}^{\lambda}_X(f) \coloneqq \frac{1}{N}\boldsymbol{\alpha}^T(K_X+\lambda_X I)^2\boldsymbol{\alpha}, \mathrm{var}^{\lambda}_Y(g) \coloneqq \frac{1}{N}\boldsymbol{\beta}^T(K_Y+\lambda_Y I)^2\boldsymbol{\beta}. \tag{D.1}$$

By replacing the variance term with λ-regularized variance term, we define the regularized $\mathcal{F}$-correlation

$$\mathcal{F}^{\lambda}(f,g) \coloneqq \frac{\mathrm{cov}_{XY}(f,g)}{\sqrt{\mathrm{var}^{\lambda}_X(f)\,\mathrm{var}^{\lambda}_Y(g)}}. \tag{D.2}$$

Denoting $K_{XX}^{\lambda} \coloneqq (K_X + \lambda_X I)^2, K_{YY}^{\lambda} \coloneqq (K_Y + \lambda_Y I)^2$, we have the regularized NAC formulation

$$\begin{pmatrix} 0 & K_{XY} \\ K_{YX} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \rho^{\lambda} \begin{pmatrix} K_{XX}^{\lambda} & 0 \\ 0 & K_{YY}^{\lambda} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}$$

and

$$\mathrm{NAC}^{\lambda}(k) \coloneqq -\frac{1}{2} \sum_{j=1}^{k} \log\left(1 - \left(\rho_j^{\lambda}\right)^2\right). \tag{D.3}$$

We will only use λ-regularized NAC so all λ superscripts will be omitted.

Appendix E. Empirical Estimate of Cross Covariance Operator

In this section, we derive the empirical estimator of Σ_{XY}.

First we note that the definition of Σ_{XY} is given by

$$\Sigma_{XY} \coloneqq \mathrm{cov}\left(f(X), g(Y)\right), \text{ where } f \in \mathcal{H}_X, g \in \mathcal{H}_Y \tag{E.1}$$

for any empirical sample $\{(x_i, y_i)\}_{i=1}^N$, $f = \sum_{i=1}^N \alpha_i \kappa_X(\cdot, x_i), g = \sum_{i=1}^N \beta_i \kappa_Y(\cdot, Y)$. Then the empirical estimate of $\mathrm{cov}\left(f(X), g(Y)\right)$ is given by

$$\begin{aligned} \mathrm{cov}_{emp}\left(f(X), g(Y)\right) &= \frac{1}{N} \sum_{i=1}^{N} f(x_i) g(y_i) && \text{(E.2)} \\ &= \frac{1}{N} \left(K_X \boldsymbol{\alpha}\right)^T \left(K_Y \boldsymbol{\beta}\right) && \text{(E.3)} \\ &= \frac{1}{N} \left\langle K_X \boldsymbol{\alpha}, K_Y \boldsymbol{\beta} \right\rangle && \text{(E.4)} \\ &= \left\langle \boldsymbol{\alpha}, \left(\frac{1}{N} K_X K_Y\right) \boldsymbol{\beta} \right\rangle && \text{(E.5)} \end{aligned}$$

Thus the empirical estimate of Σ_{XY} is given by $\frac{1}{N} K_X K_Y$.

Appendix F. Copula transformation and invariance property

We note that a desirable property of the dependence measure is that it is invariant to certain transformations in the original space. While the invariance property with respect to general invertible transformation is difficult to achieve, we propose to use the copula transformation Póczos and Schneider (2012) to

make NAC invariant to the monotonic increasing transformation of marginal variables. Copular transformation maps the variables from $\mathbb{R}^d$ to compact set $[0,1]^d$, by mapping the original variable $(X_1, \cdots, X_d)$ to $(F_1(X_1), \cdots, F_d(X_d))$, where $\{F_i\}_{i=1}^d$ are the cumulative distribution function (CDF) of marginals. The empirical copula transformation could be implemented via plug-in in the empirical CDF(simply take the rank of the marginals). The invariance of NOCCO under invertible transformations and the consistency result of copula variant of NOCCO has been proved in J Reddi and Póczos (2013), and the results could be easily extended to NAC.

Appendix G. Computational recipe

We found that numerically solving a $2N$ dimensional symmetric eigenproblem is more computationally intensive than solving a N dimensional asymmetric eigenproblem. So we rewrite the original formulation as

$$\begin{cases} \rho K_{XX}^{\lambda} \alpha = K_{XY} \beta, \\ K_{YX} \alpha = \rho K_{YY}^{\lambda} \beta, \end{cases} \tag{G.1}$$

this leads us to a N-dimensional equivalent eigenproblem

$$(K_{YY}^{\lambda})^{-1} K_{YX} (K_{YY}^{\lambda})^{-1} K_{XY} \beta = \rho^2 \beta \tag{G.2}$$

that could be handled more efficiently. The kernel matrix K_X and K_Y need not to be explicitly calculated and stored in the memory via utilizing incomplete Cholesky decomposition (ICD). The Cholesky decomposition factorize the semi-positive definite matrix K as $K = LL^T$, where $L \in \mathcal{R}^{N \times N}$ is a lower triangular matrix. By properly pivoting the rows L could be truncated to $\tilde{L} \in \mathcal{R}^{N \times p}$, where $p \ll N$ and $K = \tilde{L}\tilde{L}^T + E$ with small $\|E\|$. The calculation of incomplete Cholesky decomposition of the kernel matrix only needs the knowledge of up to k-th column of the kernel matrix instead of the entire kernel matrix at step $\underline{k}$. So evaluating ICD directly from the data in original space significantly relieves the computational burden of kernel evaluation and subsequent matrices operations provided that the kernel could be well approximated by $\tilde{L}\tilde{L}^T$ with $\tilde{L}$ a 'high'

truncated lower triangular matrix. The Kailath identity Henderson and Searle (1981),

$$(A + BC)^{-1} = A^{-1} - A^{-1}B(I + CA^{-1}B)^{-1}CA^{-1}, \tag{G.3}$$

will speed up the matrix inversion operation of the form $(K + \lambda I)^{-1}$ in the calculation of NAC if one has already factorized the kernel matrix using incomplete Cholesky decomposition(ICD), by replacing B with $\tilde{L}$, C with $\tilde{L}^T$ and A with λI, where $\tilde{L}\tilde{L}^T$ being the ICD of K.

The calculation of kernel association and reconstruction of the signal depends on solving the leading k eigenpairs $\{(\rho_i^2, \xi_i)\}_{i=1}^k$, s.t.

$$(K_{YY}^{\lambda})^{-1}K_{YX}(K_{YY}^{\lambda})^{-1}K_{XY}\xi_i = \rho^2\xi_i.$$

For our application, we propose to use the Ritz value and Ritz vector to achieve efficient yet accurate approximation of the leading eigenpairs Saad (1992). For $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n\backslash\{0\}$, $K_k = span\{b, Ab, \cdots, A^{k-1}b\}$ is called the Krylov subspace of A and b. Denote $Q_k = \{q_1, \cdots, q_k\}$ as the orthonormal basis for the Krylov subspace K_k, which could be obtained by either QR decomposition, Lanczos (Hermitian) or Arnoldi (general) algorithm, $H_k = Q_k' A Q_k$ is the projection of operator A onto subspace K_k. The eigenpairs $\{(\lambda_i, v_i)\}$ of the H_k is called Ritz value and Ritz vector of A and provides $\{(\lambda_i, Q_k v_i)\}$ as an approximation of for eigenpairs of A. Thus the original n-dimensional eigenvalue problem could be efficiently approximated by solving a k-dimensional eigenvalue problem, in turn dramatically cuts the computational cost. To obtain the k leading eigenpairs with the desired accuracy, we need the Krylov subspace's dimension larger than k (popular choice is $d = \min(k + 10, 2k)$), and the program might need several restarts to avoid the degenerated arithmetic precision. The combination of incomplete Cholesky decomposition and Ritz approximation offers a 5 X ~ 15 X speedup over the standard implementation on the test fMRI studies described above. Also algorithm could be easily implemented on a modern GPU device for further acceleration.

Using more sophisticated eigenvalue approximation schemes could further improve the stability and efficiency. For details please refer to Saad (1992).

The numerical scheme for the eigenvalue problem could be further accelerated using singular value decomposition (SVD). Recall that the following eigenvalue problem is to be solved

$$\begin{pmatrix} K_X^\lambda & K_{XY} \\ K_{YX} & K_Y^\lambda \end{pmatrix}\begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\zeta} \end{pmatrix} = (1+\rho)\begin{pmatrix} K_X^\lambda & 0 \\ 0 & K_Y^\lambda \end{pmatrix}\begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\zeta} \end{pmatrix}. \tag{G.4}$$

By taking $\widetilde{\boldsymbol{\xi}} = K_X^\lambda \boldsymbol{\xi}$, $\widetilde{\boldsymbol{\zeta}} = K_Y^\lambda \boldsymbol{\zeta}$ and eliminating the block diagonal matrix in (G.4), we have

$$\begin{pmatrix} I & K_X^r K_Y^r \\ K_Y^r K_X^r & I \end{pmatrix}\begin{pmatrix} \widetilde{\boldsymbol{\xi}} \\ \widetilde{\boldsymbol{\zeta}} \end{pmatrix} = (1+\rho)\begin{pmatrix} \widetilde{\boldsymbol{\xi}} \\ \widetilde{\boldsymbol{\zeta}} \end{pmatrix}, \tag{G.5}$$

where $K_X^r \coloneqq \left(K_X^\lambda\right)^{-1} K_X$ and $K_Y^r \coloneqq \left(K_Y^\lambda\right)^{-1} K_Y$. Denote the low rank factorization $K_X \approx L_x L_x^\top$ and $K_Y \approx L_y L_y^\top$. We perform SVD on $L_x = U_x S_x V_x$ and $L_y = U_y S_y V_y$, so $K_X \approx U_x \Lambda_x U_x^\top$ and $K_Y = U_y \Lambda_y U_y^\top$. Then we have

$$K_X^r = U_x R_x U_x^\top,$$

where R_x is applying the function $x \mapsto \frac{x}{x+\lambda_x}$ to Λ_x's elements.

$$\begin{aligned} K_X + \lambda_x I &= [U_x, U_x^\perp](\widetilde{\Lambda}_x + \lambda_x I)[U_x, U_x^\perp]^\top \\ (K_X + \lambda_x I)^{-1} &= [U_x, U_x^\perp]\left(\widetilde{\Lambda}_x + \lambda_x\right)^{-1}[U_x, U_x^\perp]^\top \\ [U_x, U_x^\perp]^\top U_x &= [I_{r_x}; 0] \end{aligned}$$

where $\widetilde{\Lambda}_x = \mathrm{diag}(\Lambda_x, 0)$.

This gives the following factorization of l.h.s matrix in (G.5)

$$\begin{pmatrix} I & K_X^r K_Y^r \\ K_Y^r K_X^r & I \end{pmatrix} = [\mathbf{U}, \mathbf{V}]\mathrm{diag}(\mathbf{R}, I)[\mathbf{U}, \mathbf{V}]^\top$$

$$\mathbf{U} \coloneqq \mathrm{diag}(U_x, U_y), \mathbf{V} \coloneqq \mathrm{diag}(U_x^\perp, U_y^\perp)$$

and

$$\mathbf{R} \coloneqq \begin{pmatrix} I & R_x U_x^\top U_y R_y \\ R_y U_y^\top U_x R_x & I \end{pmatrix}.$$

This means it is equivalent to solve the eigenvalue problem of $\mathbf{R}$, which is a $(r_x + r_y) \times (r_x + r_y)$ matrix. Denote $(\check{\boldsymbol{\xi}}^\top, \check{\boldsymbol{\zeta}}^\top)^\top$ is the eigenvector estimated for $\mathbf{R}$, then the eigenvector for the original problem is $(\widetilde{\boldsymbol{\xi}}^\top, \widetilde{\boldsymbol{\zeta}}^\top)^\top = \mathbf{U}(\check{\boldsymbol{\xi}}^\top, \check{\boldsymbol{\zeta}}^\top)^\top$.

$$\begin{aligned}
\boldsymbol{\xi} &= \left(K_X^{\lambda}\right)^{-1} \widetilde{\boldsymbol{\xi}} \\
&= [U_x, U_x^{\perp}] \left(\widetilde{\Lambda}_x + \lambda_x\right)^{-1} [U_x, U_x^{\perp}]^\top U_x \check{\boldsymbol{\xi}} \\
&= U_x \left(\Lambda_x + \lambda_x\right)^{-1} \check{\boldsymbol{\xi}}
\end{aligned}$$

The complete algorithm for estimating the number of eigenvalues to keep is detailed as follows:

Algorithm 1: Estimate the number of eigenvalues to keep

Input: $X, Y, M, B, R_{\text{thres}}$

Split the samples into M batches with batch size $\{N_m\}_{m=1}^M$

Denote $\mathcal{I}_m$ the samples for m-th batch and $\mathcal{I}_m^C$ the rest of samples

for $\underline{m = 1 : M}$ **do**

  Estimate $\{f^{(i)}, g^{(i)}\}_{i=1}^l$ using $\mathcal{I}_m^C$

  Estimate $c(i, m) \coloneqq |\text{corr}(f^{(i)}, g^{(i)})|$ using $\mathcal{I}_m$

end for

Set $C(i) \coloneqq \text{mean}(c(i, :))$ for $i = 1 : l$

for $\underline{b = 1 : B}$ **do**

  Randomly shuffle Y to break the dependecy

  for $\underline{m = 1 : M}$ **do**

    Estimate $\{f^{(i)}, g^{(i)}\}_{i=1}^l$ using $\mathcal{I}_m^C$

    Estimate $c_{null}(i, m) \coloneqq |\text{corr}(f^{(i)}, g^{(i)})|$ using $\mathcal{I}_m$

  end for

  Set $C_{null}(b, i) \coloneqq \text{mean}(c_{null}(i, :))$

end for

Set $R(i) \coloneqq \#\{C_{null}(:, i) \le C(i)\}$

Set k_{opt} to largest k satisfying $R(i) \ge R_{\text{thres}}$ for all $i \le k$

Appendix H. Additional simulation examples

Example-2: Parameter dependence supplemental

We can show how the sensitivity of NAC depends on the choice of regularization parameter λ. Here the dimension of $\mathbf{d}_1$ and $\mathbf{d}_2$ are both equal to 3. $\boldsymbol{n}_1$ is set to 0 and $\boldsymbol{n}_2, \boldsymbol{s}$ are independently drawn from $[0, 1]$ uniform distribution. $a = 1$ as before. $N = 500$ samples are generated for each experiment and it is repeated for $M = 100$ times. The NAC parameters are set to $k = 5, B = 500$. We calculate NAC for the regularization parameter λ in $\{0.01, 0.1, 1, 10, 100\}$. We calculate both the 10-fold out-of-sample NAC and the approximated significance of each experiment. The results are shown in Figure 1(a). This confirms that the λ that maximizes out-of-sample NAC is a good estimate for optimal regularization parameter. Out-of-sample NAC is calculated via plug-in the mean (alternatively median) of out-of-sample correlation coefficients to the NAC formula.

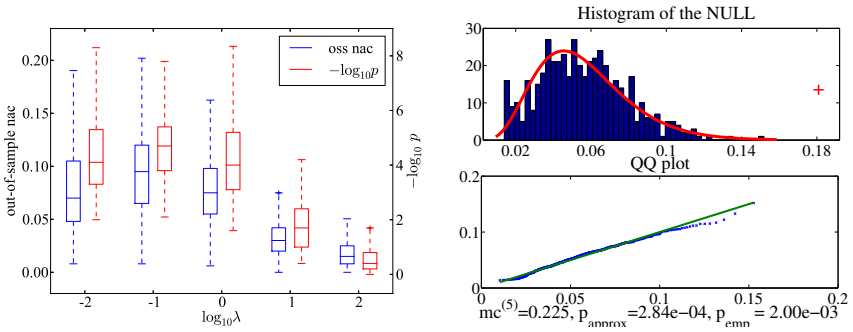
We further tested our procedure to automatically determine the number of eigencomponents k in NAC. The dimension of $\mathbf{d}_1$ and $\mathbf{d}_2$ are both equal to 3. The first free variables of $\boldsymbol{s}$ are *i.i.d.* draws from $[0, 1]$ uniformly while the rest two are set to zero. $\boldsymbol{n}_1$ and $\boldsymbol{n}_2, \boldsymbol{s}$ are *i.i.d.* draws from $[0, 0.7]$ uniform distribution. We set $a = 3.7$. $M = 100$ runs are executed. First 5 eigen-components are computed with varying sample size $N = \{200, 400, 600, 800, 1000\}$. The samples are equally divided into 2 batches and each time one batch is used as test set and the other is used as training set. We set the threshold to 0.95 and plot the probability that the k-th eigen-component being selected. The mean out-of-sample association exhibits a sharp transition between the third and fourth eigen-component, with sufficiently large sample size (Figure 1(b)), suggesting $k = 3$ is a good choice for the problem. And with great probability our procedure selects this k.

Example 2: NGC supplemental

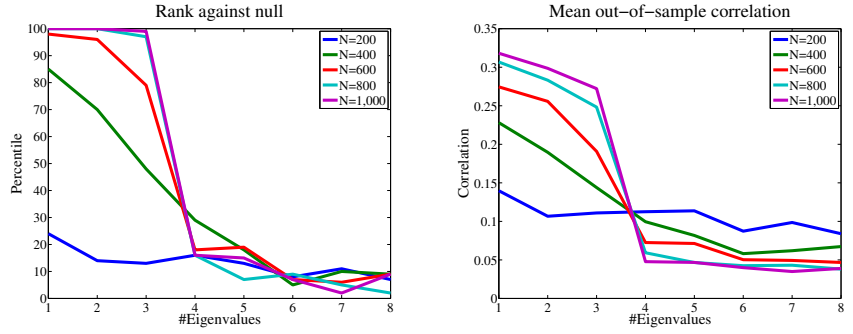
In this example, we demonstrate NGC as an application of conditional NAC. The following Logistic model with random perturbation is used in this example

$$x_{t+1} = \alpha\, x_t\,(1 - x_t) + n_t$$

where $\alpha = 3.6$ and $\{n_t\}$ are *i.i.d.* draws from $\mathcal{N}(0, 4 \times 10^{-4})$. We test the following three causal links:



(a)



(b)

Figure H.1: (a) (Left) Blue: 10-fold out-of-sample NAC statistics against λ. Red: the dependence of negative $\log_{10} p$ against λ. (Right) Diagnostic plot for one realization of the experiments at optimal regularization λ. The red-doted line corresponds to the $p = 0.05$ threshold. (b) 2-fold cross-validation procedure on the number of eigenvalues with different sample-size($N = \{200, 400, 600, 800, 1000\}$). Left: probability the k^{th} eigenvalue against the *null* distribution reached 95% percentile. Right: estimated mean out-of-sample association with the respective eigenvalues.

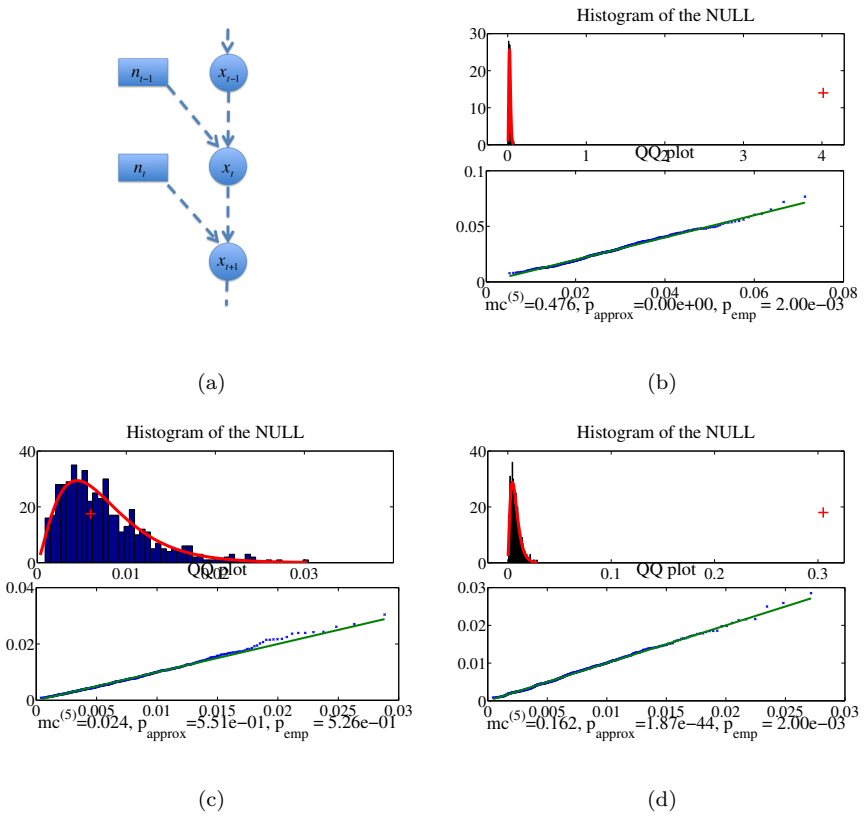


Figure H.2: Results of conditional NAC. (a) The randomly perturbed Logistic Model (b) Unconditional NAC between X_t and X_{t+2} (c) NAC between X_t and X_{t+2} condition on X_{t+1} (d) NAC between X_{t+1} and X_{t+2} condition on X_t

(i) $x_t \to x_{t+2}$,

(ii) $x_t \to x_{t+2}|x_{t+1}$,

(iii) $x_{t+1} \to x_{t+2}|x_t$.

$N = 500$ points were sampled and NAC parameters were set to $k = 5, B = 500$ and $\lambda = 0.1$. NAC successfully identifies the causal link (i) and (iii), while no causality is detected in link (ii), i.e. the false causal link is detected through conditioning. See Figure H.2 for the model schematic and diagnostic plots for one realization of each link.

Example 4: Comparison supplemental

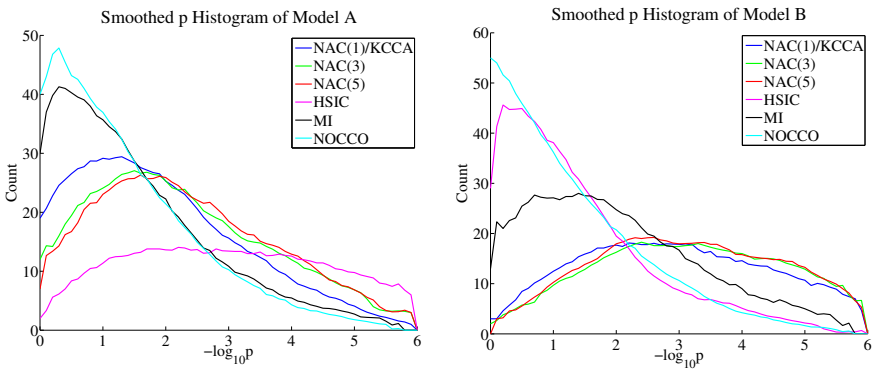


Figure H.3: Smoothed $\log_{10} p$ histograms of NAC, HSIC, MI and NOCCO for linear case (model A, left) and nonlinear case (model B, left) (blue NAC(1), green NAC(3), red NAC(5), purple HSIC, black MI, cyan NOCCO)

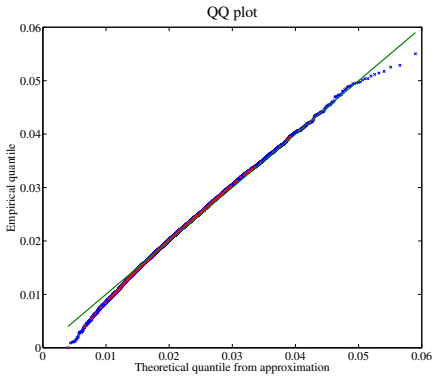


Figure H.4: QQ-plot for the Gamma approximation of NAC with real-world fMRI data, red for fitted permutation runs and blue for validation runs.

Appendix I. Remarks on NAC and related measures

We remark that NAC achieved significant computation reductions compared with other related kernel approaches. For example, in KGC Marinazzo et al. (2008) and KCCA-GC Wu et al. (2011) the full eigenvalue problem is been solved, but only the leading k principal components or the first eigenvalue is been actually used. This results in unnecessary computational overhead. As we show in this work an accurate approximation of the leading eigenvalue / vectors is suffice for the purpose of association detection while significantly reduce the computation burden. Utilizing our strategy computing NAC will undoubtedly boost the efficiency of those kernel methods.

Appendix J. Top 50 connections identified in high-resolution fMRI study

Table J.1: NGC Linear

rank	from	to	p-value	rank	from	to	p-value
1	Postcentral_R	Frontal_Mid_Orb_L	1.44E-07	26	Caudate_R	Cingulum_Ant_R	9.21E-06
2†,*	Caudate_L	Insula_L	2.96E-07	27	Caudate_R	Calcarine_R	1.05E-05
3†	Precentral_R	Frontal_Mid_L	4.66E-07	28*	Caudate_R	Precentral_L	1.06E-05
4†,*	Caudate_R	Frontal_Mid_R	4.78E-07	29*	Caudate_L	Postcentral_R	1.08E-05
5†,*	Caudate_R	Frontal_Sup_R	6.06E-07	30	Paracentral_Lobule_L	Temporal_Pole_Sup_R	1.15E-05
6*	Caudate_L	Precentral_R	1.03E-06	31†	Caudate_L	Frontal_Mid_L	1.18E-05
7	Paracentral_Lobule_L	Frontal_Mid_Orb_R	1.09E-06	32	Caudate_R	Precuneus_R	1.42E-05
8	SupraMarginal_L	Precentral_L	1.18E-06	33	Temporal_Inf_R	Rectus_L	1.42E-05
9†,*	Caudate_L	Cingulum_Ant_R	1.28E-06	34	Paracentral_Lobule_R	Frontal_Sup_Medial_R	1.50E-05
10	Caudate_R	Frontal_Inf_Tri_L	1.59E-06	35	Postcentral_R	Cingulum_Post_L	1.57E-05
11	Caudate_R	Temporal_Pole_Sup_R	1.63E-06	36	Postcentral_R	Parietal_Sup_R	1.58E-05
12*	Caudate_L	Frontal_Mid_R	1.96E-06	37	Caudate_R	Frontal_Sup_Medial_R	1.58E-05
13†,*	Caudate_L	Frontal_Inf_Tri_L	2.05E-06	38*	Cingulum_Post_L	Frontal_Med_Orb_R	1.61E-05
14*	Caudate_L	Rectus_L	3.10E-06	39	Temporal_Sup_L	Supp_Motor_Area_R	1.61E-05
15	Precentral_R	Frontal_Inf_Tri_L	3.10E-06	40	Caudate_L	Frontal_Sup_Medial_R	1.81E-05
16*	Paracentral_Lobule_L	Parietal_Sup_L	3.20E-06	41	Temporal_Sup_L	Paracentral_Lobule_R	1.86E-05
17	Caudate_R	Parietal_Sup_L	3.63E-06	42	Paracentral_Lobule_L	Frontal_Mid_Orb_L	2.01E-05
18*	Paracentral_Lobule_R	Frontal_Sup_R	4.17E-06	43	Precentral_R	Cingulum_Post_L	2.04E-05
19	Caudate_L	Occipital_Inf_L	4.90E-06	44	Caudate_L	Precuneus_L	2.05E-05
20	Caudate_R	Frontal_Inf_Orb_L	5.18E-06	45†	Caudate_R	Frontal_Mid_L	2.11E-05
21	Caudate_L	Fusiform_L	5.30E-06	46†,*	Caudate_L	Supp_Motor_Area_L	2.16E-05
22*	Caudate_L	Supp_Motor_Area_R	5.35E-06	47	Caudate_R	Paracentral_Lobule_L	2.17E-05
23†,*	Postcentral_R	Frontal_Mid_L	5.62E-06	48*	Caudate_L	Cingulum_Mid_R	2.21E-05
24†,*	Caudate_L	Temporal_Pole_Sup_R	5.63E-06	49	Precentral_R	Rolandic_Oper_R	2.24E-05
25	Paracentral_Lobule_R	Frontal_Mid_Orb_R	7.57E-06	50	Caudate_L	Lingual_L	2.35E-05

† common with GC (13 edges), * common with KCCA (18 edges)

Table J.2: GC Regression

rank	from	to	p-value	rank	from	to	p-value
1	Caudate_R	Frontal_Inf_Tri_L	1.06E-06	26	Olfactory_R	Temporal_Sup_R	1.81E-04
2	Temporal_Inf_R	Frontal_Inf_Tri_L	4.35E-06	27	Caudate_R	ParaHippocampal_R	1.82E-04
3	Precentral_R	Frontal_Mid_L	9.81E-06	28	Caudate_R	Frontal_Mid_R	2.05E-04
4	Caudate_L	Cingulum_Ant_R	1.04E-05	29	Amygdala_R	Temporal_Sup_L	2.09E-04
5	Postcentral_R	Frontal_Mid_L	1.22E-05	30	Caudate_L	Frontal_Mid_R	2.55E-04
6	Caudate_R	Frontal_Mid_L	1.37E-05	31	Caudate_L	Supp_Motor_Area_L	2.58E-04
7	Caudate_L	Occipital_Mid_L	1.52E-05	32	Olfactory_R	Thalamus_L	2.68E-04
8	Caudate_L	Frontal_Inf_Tri_L	2.21E-05	33	Caudate_R	Occipital_Mid_R	2.93E-04
9	Olfactory_R	Frontal_Inf_Tri_L	2.28E-05	34	Rolandic_Oper_R	Thalamus_R	2.96E-04
10	Amygdala_R	Frontal_Inf_Tri_L	6.11E-05	35	Caudate_L	Temporal_Pole_Sup_R	3.03E-04
11	Caudate_R	Frontal_Sup_R	7.40E-05	36	Precentral_R	Temporal_Pole_Sup_R	3.07E-04
12	Caudate_R	Occipital_Inf_R	7.89E-05	37	Temporal_Inf_R	Occipital_Inf_L	3.34E-04
13	Caudate_L	Frontal_Sup_R	9.07E-05	38	Hippocampus_L	Parietal_Sup_L	3.41E-04
14	Angular_L	Frontal_Sup_L	9.55E-05	39	Temporal_Mid_R	Frontal_Sup_R	3.58E-04
15	Caudate_L	Amygdala_R	1.14E-04	40	Cingulum_Post_R	Cingulum_Ant_L	3.59E-04
16	Caudate_L	Thalamus_L	1.16E-04	41	Caudate_L	Putamen_L	3.60E-04
17	Caudate_L	Frontal_Mid_L	1.18E-04	42	Precentral_R	Cingulum_Mid_L	3.64E-04
18	Frontal_Inf_Orb_R	Temporal_Pole_Mid_R	1.25E-04	43	ParaHippocampal_R	Frontal_Inf_Tri_L	3.89E-04
19	Temporal_Mid_R	Calcarine_L	1.29E-04	44	Caudate_L	Occipital_Mid_R	3.95E-04
20	Cingulum_Post_L	Frontal_Sup_L	1.34E-04	45	Caudate_L	Insula_L	3.97E-04
21	Caudate_R	Amygdala_R	1.36E-04	46	Cingulum_Post_R	Fusiform_L	4.27E-04
22	Caudate_L	Occipital_Sup_R	1.40E-04	47	Cingulum_Post_R	Frontal_Sup_L	4.28E-04
23	Caudate_L	Occipital_Sup_L	1.46E-04	48	Temporal_Inf_R	Caudate_R	4.35E-04
24	Frontal_Sup_L	Temporal_Pole_Mid_L	1.66E-04	49	Olfactory_R	Precentral_L	4.38E-04
25	Frontal_Inf_Tri_R	Frontal_Sup_L	1.76E-04	50	Caudate_R	Precuneus_L	4.51E-04

Table J.3: KCCA Regression

rank	from	to	p-value	rank	from	to	p-value
1	Cingulum_Post_L	Cingulum_Ant_L	7.69E-09	26	Thalamus_L	Frontal_Sup_R	1.37E-05
2	Caudate_L	Frontal_Mid_R	2.66E-07	27	Caudate_L	Amygdala_R	1.41E-05
3	Postcentral_R	Frontal_Mid_L	4.35E-07	28	Caudate_L	Supp_Motor_Area_L	1.43E-05
4	SupraMarginal_R	Frontal_Sup_L	4.86E-07	29	Frontal_Sup_R	Frontal_Mid_R	1.48E-05
5	Caudate_L	Cingulum_Mid_R	4.88E-07	30	Precentral_L	Lingual_R	1.50E-05
6	Caudate_L	Postcentral_R	6.48E-07	31	Paracentral_Lobule_R	Frontal_Sup_R	1.53E-05
7	Angular_L	Temporal_Sup_L	7.75E-07	32	Temporal_Inf_R	Frontal_Inf_Tri_L	1.56E-05
8	Caudate_R	Frontal_Sup_R	8.01E-07	33	Caudate_L	Temporal_Pole_Sup_R	1.67E-05
9	Caudate_L	Frontal_Inf_Orb_L	1.12E-06	34	Caudate_R	Occipital_Sup_L	1.82E-05
10	Cingulum_Post_L	Frontal_Sup_Medial_L	1.37E-06	35	Caudate_R	Supp_Motor_Area_L	1.89E-05
11	Caudate_L	Cingulum_Ant_R	1.83E-06	36	Cingulum_Post_L	Frontal_Sup_L	2.12E-05
12	Cingulum_Post_L	Frontal_Med_Orb_L	2.18E-06	37	Cingulum_Post_L	Temporal_Pole_Mid_R	2.40E-05
13	Caudate_L	Insula_L	3.67E-06	38	Caudate_R	Precuneus_L	2.50E-05
14	SupraMarginal_R	Frontal_Mid_R	3.69E-06	39	Paracentral_Lobule_L	Parietal_Sup_L	2.53E-05
15	Cingulum_Post_R	Cingulum_Ant_L	3.89E-06	40	Temporal_Pole_Mid_L	Precentral_R	2.53E-05
16	Caudate_L	Supp_Motor_Area_R	4.45E-06	41	Caudate_L	Rectus_L	2.64E-05
17	Caudate_R	Occipital_Sup_R	4.60E-06	42	Caudate_R	Precentral_L	2.69E-05
18	Caudate_R	Frontal_Mid_R	5.64E-06	43	Temporal_Sup_L	Temporal_Mid_L	2.70E-05
19	Temporal_Sup_L	Frontal_Mid_Orb_L	6.13E-06	44	Cingulum_Post_L	Cingulum_Ant_R	2.73E-05
20	Caudate_L	Precentral_R	7.11E-06	45	ParaHippocampal_R	Frontal_Mid_R	2.77E-05
21	Frontal_Med_Orb_L	Parietal_Inf_R	1.03E-05	46	Cingulum_Post_L	Frontal_Med_Orb_R	2.90E-05
22	Caudate_L	SupraMarginal_L	1.06E-05	47	Temporal_Sup_R	Frontal_Mid_Orb_L	2.92E-05
23	Caudate_L	Frontal_Sup_R	1.16E-05	48	Caudate_L	Frontal_Inf_Tri_L	3.01E-05
24	Angular_R	Cuneus_R	1.18E-05	49	Caudate_R	Temporal_Pole_Mid_R	3.15E-05
25	Frontal_Sup_L	Frontal_Mid_R	1.33E-05	50	Frontal_Mid_R	Angular_R	3.17E-05

Table J.4: NGC T-test

rank	from	to	p-value	rank	from	to	p-value
1	Postcentral_R	Frontal_Mid_Orb_L	1.08E-06	26	Temporal_Sup_L	Paracentral_Lobule_R	5.52E-05
2	Precentral_R	Frontal_Mid_L	2.54E-06	27	Paracentral_Lobule_L	Occipital_Sup_L	5.68E-05
3	Paracentral_Lobule_L	Frontal_Mid_Orb_R	3.12E-06	28	Paracentral_Lobule_R	Frontal_Sup_Medial_R	6.26E-05
4	Caudate_R	Frontal_Mid_R	4.82E-06	29	Precentral_R	Temporal_Pole_Sup_R	7.15E-05
5	Precentral_R	Frontal_Inf_Tri_L	4.89E-06	30	Temporal_Sup_L	Frontal_Mid_Orb_L	7.22E-05
6	Precentral_R	Occipital_Sup_R	8.00E-06	31	Frontal_Inf_Tri_L	Parietal_Inf_L	7.26E-05
7	Paracentral_Lobule_R	Frontal_Mid_R	8.23E-06	32	SupraMarginal_R	Parietal_Inf_R	7.59E-05
8	Caudate_L	Cingulum_Ant_R	1.05E-05	33	Cingulum_Post_L	Frontal_Mid_R	7.89E-05
9	Caudate_L	Frontal_Mid_R	1.10E-05	34	Postcentral_R	Parietal_Sup_R	7.97E-05
10	SupraMarginal_L	Precentral_L	1.23E-05	35	Caudate_L	Frontal_Inf_Tri_L	8.21E-05
11	Paracentral_Lobule_R	Frontal_Mid_Orb_R	1.25E-05	36	Paracentral_Lobule_R	Postcentral_L	8.24E-05
12	Paracentral_Lobule_R	ParaHippocampal_R	1.33E-05	37	Paracentral_Lobule_R	Precentral_L	8.48E-05
13	Caudate_L	Insula_L	1.37E-05	38	Precentral_R	Temporal_Pole_Mid_R	8.51E-05
14	Postcentral_R	Frontal_Mid_L	1.42E-05	39	Precentral_R	Temporal_Sup_R	8.86E-05
15	Paracentral_Lobule_R	Frontal_Sup_R	1.64E-05	40	Caudate_L	Occipital_Inf_L	9.26E-05
16	Precentral_R	ParaHippocampal_R	2.13E-05	41	Paracentral_Lobule_L	ParaHippocampal_R	1.02E-04
17	Caudate_R	Cingulum_Ant_R	2.74E-05	42	Caudate_R	Frontal_Inf_Oper_R	1.04E-04
18	Paracentral_Lobule_L	Frontal_Mid_Orb_L	2.79E-05	43	Frontal_Inf_Tri_R	Paracentral_Lobule_L	1.04E-04
19	Caudate_R	Precentral_L	2.82E-05	44	Postcentral_R	Occipital_Sup_R	1.07E-04
20	Postcentral_R	Temporal_Pole_Sup_R	3.75E-05	45	Frontal_Inf_Tri_L	Cingulum_Post_R	1.22E-04
21	Postcentral_R	Temporal_Sup_R	3.96E-05	46	Caudate_R	Frontal_Inf_Tri_L	1.26E-04
22	Caudate_R	Paracentral_Lobule_L	4.47E-05	47	Paracentral_Lobule_L	Putamen_L	1.32E-04
23	Postcentral_R	Frontal_Mid_R	4.54E-05	48	Caudate_R	Frontal_Inf_Orb_L	1.33E-04
24	Paracentral_Lobule_L	Parietal_Sup_L	4.56E-05	49	Temporal_Sup_R	Paracentral_Lobule_R	1.41E-04
25	Caudate_R	Frontal_Sup_R	4.91E-05	50	Caudate_L	Precentral_R	1.43E-04

Table J.5: GC T-test

rank	from	to	p-value	rank	from	to	p-value
1	Caudate_R	Frontal_Inf_Tri_L	9.31E-05	26	Frontal_Inf_Tri_R	Frontal_Sup_L	6.26E-04
2	Postcentral_R	Frontal_Mid_L	9.64E-05	27	Precentral_R	Occipital_Sup_R	6.84E-04
3	Frontal_Inf_Orb_R	Temporal_Pole_Mid_R	1.06E-04	28	Frontal_Sup_L	Fusiform_R	7.17E-04
4	ParaHippocampal_R	Supp_Motor_Area_L	1.18E-04	29	Caudate_L	Supp_Motor_Area_L	7.17E-04
5	Postcentral_R	Temporal_Pole_Sup_R	1.42E-04	30	Frontal_Mid_L	Frontal_Inf_Oper_L	7.28E-04
6	Olfactory_R	Precentral_L	1.54E-04	31	Frontal_Sup_Medial_R	Temporal_Pole_Mid_L	7.75E-04
7	Frontal_Sup_L	Temporal_Pole_Mid_L	1.58E-04	32	Precuneus_R	Caudate_L	7.89E-04
8	Caudate_L	Frontal_Inf_Tri_L	1.63E-04	33	Caudate_L	Occipital_Mid_L	8.39E-04
9	Angular_L	Parietal_Sup_R	1.72E-04	34	Frontal_Sup_L	Cuneus_R	9.25E-04
10	Precentral_R	Cingulum_Mid_L	2.36E-04	35	Angular_R	Occipital_Sup_L	9.35E-04
11	Temporal_Mid_R	Frontal_Sup_R	2.47E-04	36	Rolandic_Oper_R	Thalamus_R	9.59E-04
12	Precentral_R	Occipital_Mid_R	2.53E-04	37	Frontal_Sup_L	Temporal_Pole_Sup_L	9.90E-04
13	Parietal_Sup_L	Cingulum_Mid_R	2.69E-04	38	Temporal_Sup_L	Supp_Motor_Area_R	1.01E-03
14	Caudate_L	Cingulum_Ant_R	3.02E-04	39	Temporal_Mid_R	Calcarine_L	1.03E-03
15	Olfactory_R	Frontal_Mid_R	3.72E-04	40	Supp_Motor_Area_L	ParaHippocampal_R	1.06E-03
16	Caudate_R	Frontal_Mid_R	4.32E-04	41	Olfactory_R	Thalamus_L	1.07E-03
17	Frontal_Sup_L	Frontal_Inf_Orb_L	4.41E-04	42	Cingulum_Post_L	Frontal_Sup_L	1.14E-03
18	Caudate_L	Frontal_Sup_R	4.51E-04	43	Cingulum_Post_L	Rectus_L	1.19E-03
19	Precentral_R	Temporal_Pole_Sup_R	4.65E-04	44	Caudate_L	Occipital_Sup_L	1.20E-03
20	Temporal_Mid_R	Precuneus_L	4.68E-04	45	Frontal_Inf_Tri_L	Parietal_Inf_L	1.25E-03
21	Precentral_R	Frontal_Mid_L	5.37E-04	46	Cingulum_Post_R	Frontal_Med_Orb_L	1.28E-03
22	Temporal_Sup_R	Occipital_Mid_R	5.54E-04	47	Olfactory_L	Frontal_Mid_R	1.29E-03
23	Temporal_Sup_R	Frontal_Inf_Tri_L	5.73E-04	48	Caudate_L	Frontal_Mid_R	1.33E-03
24	Angular_L	Frontal_Sup_L	6.11E-04	49	Fusiform_R	Frontal_Sup_Medial_R	1.34E-03
25	Caudate_R	Frontal_Sup_R	6.24E-04	50	Caudate_R	Occipital_Inf_R	1.37E-03

Table J.6: KCCA T-test

rank	from	to	p-value	rank	from	to	p-value
1	Angular_L	Temporal_Sup_L	6.32E-08	26	Olfactory_R	Precentral_L	2.52E-05
2	Cingulum_Post_L	Cingulum_Ant_L	4.44E-07	27	Precentral_L	Lingual_R	2.59E-05
3	Cingulum_Post_L	Frontal_Med_Orb_L	2.00E-06	28	Frontal_Med_Orb_R	Parietal_Inf_R	2.64E-05
4	Caudate_L	Cingulum_Ant_R	2.59E-06	29	ParaHippocampal_R	Frontal_Mid_R	2.78E-05
5	Temporal_Sup_L	Frontal_Mid_Orb_L	2.89E-06	30	Frontal_Inf_Orb_R	Temporal_Pole_Mid_R	3.03E-05
6	Frontal_Med_Orb_L	Parietal_Inf_R	3.60E-06	31	Caudate_L	Postcentral_R	3.15E-05
7	Postcentral_R	Frontal_Mid_L	3.68E-06	32	Cingulum_Post_R	Frontal_Mid_R	3.35E-05
8	Caudate_L	Cingulum_Mid_R	4.25E-06	33	Postcentral_L	Temporal_Sup_L	5.11E-05
9	SupraMarginal_R	Frontal_Mid_R	5.17E-06	34	Postcentral_L	Postcentral_R	5.13E-05
10	Paracentral_Lobule_L	Frontal_Mid_Orb_L	6.83E-06	35	Postcentral_R	Temporal_Pole_Sup_R	5.15E-05
11	Postcentral_R	Temporal_Sup_L	7.28E-06	36	Temporal_Pole_Mid_L	Precentral_R	5.29E-05
12	Precentral_R	Occipital_Sup_R	1.01E-05	37	Temporal_Mid_R	Lingual_R	5.38E-05
13	Caudate_R	Supp_Motor_Area_L	1.13E-05	38	Cingulum_Post_L	Occipital_Sup_R	5.62E-05
14	SupraMarginal_R	Frontal_Sup_L	1.19E-05	39	Frontal_Inf_Orb_L	Cingulum_Mid_L	5.76E-05
15	Caudate_L	Frontal_Mid_R	1.22E-05	40	Angular_L	Cingulum_Mid_L	5.86E-05
16	Frontal_Sup_L	Frontal_Mid_R	1.28E-05	41	Postcentral_R	Pallidum_R	6.40E-05
17	Cingulum_Post_L	Temporal_Pole_Mid_R	1.58E-05	42	Olfactory_L	Frontal_Mid_R	6.44E-05
18	Postcentral_R	Frontal_Mid_R	1.61E-05	43	Paracentral_Lobule_L	Putamen_L	6.44E-05
19	Caudate_R	Precentral_L	1.72E-05	44	Caudate_L	Frontal_Inf_Orb_L	6.49E-05
20	Paracentral_Lobule_L	Fusiform_R	1.76E-05	45	Caudate_R	Frontal_Mid_R	6.59E-05
21	Paracentral_Lobule_R	Frontal_Mid_R	2.08E-05	46	Angular_R	Frontal_Mid_R	6.86E-05
22	Postcentral_R	Occipital_Sup_R	2.11E-05	47	Thalamus_L	Frontal_Sup_R	6.95E-05
23	Angular_R	Cuneus_R	2.13E-05	48	Caudate_R	Frontal_Sup_R	6.99E-05
24	Caudate_L	Supp_Motor_Area_L	2.17E-05	49	Postcentral_R	Rolandic_Oper_R	7.22E-05
25	Temporal_Sup_R	Lingual_R	2.29E-05	50	Postcentral_R	Occipital_Sup_L	7.33E-05