ADA Project:
Measuring Functional Connectivity in fMRI data:
Statistical Methods and Comparison to Structure

Shashank Singh
Department of Statistics
Machine Learning Department
Carnegie Mellon University

Internal Advisor:
Barnabás Póczos
Machine Learning Department
Carnegie Mellon University

External Advisor:
Timothy Verstynen
Department of Psychology
Carnegie Mellon University

February 11, 2016

Abstract

Functional magnetic resonance imaging (fMRI) provides one of our richest sources of data about the functioning human brain. Functional connectivity is a common tool used in the analysis of fMRI data for understanding the relationships between different regions of interest (ROIs). However, functional connectivity can be measured via several different methods, and the choice of method can significantly affect the results of data analysis using functional connectivity. In this project, we study, through empirical comparison, the behaviors of several functional connectivity measures on fMRI data from a brain region (the basal ganglia) with well understood structural connectivity, with the goal of advancing our ability to interpret functional connectivity and our understanding of which questions different methods are suited to answer. We focus particularly on the role of nonlinear dependencies between ROIs and the abilities of different methods to recover the structural connectivity of the basal ganglia, using resting-state fMRI data from healthy adults.

Note: This is work in progress. The many TODOs throughout this work will be filled in soon and more empirical results (and perhaps methods) will be added. As such, notes regarding any errors, sent sss1@andrew.cmu.edu, are highly appreciated.

Contents

1 Project Introduction

Functional Magnetic Resonance Imaging (fMRI) is a functional, whole-brain neuroimaging modality that measures BOLD (blood-oxygen-level dependent) signal. BOLD signal is a measure of the oxygenated blood flow to a region, which increases as a response to increased metabolic demand in that region (hemodynamic response). Thus, BOLD signal is measured as a surrogate for brain activity. fMRI measures BOLD signal within each cell (known as a *voxel*) of a fine 3-dimensional grid in which the brain rests, roughly once per second. Thus, fMRI is used to measure activity in precise regions throughout the brain with high spatial resolution (on the order of millimeters), albeit with low temporal resolution (typically ≤ 1 Hz) relative to the dynamics of neurons, which can fire hundreds of times per second.

A common tool for analyzing of fMRI is *functional connectivity*, which refers to temporal coherence of activity between different brain regions (Hlinka et al. [2011]). Functional connectivity is a purely statistical construct, distinguished from *effective connectivity*, which refers to a causal, mechanistic dependence between brain regions (Stephan and Friston [2010]). [TODO: Better explain the distinction between effective and structural connectivities.] Functional connectivity can be measured in several ways; most methods we discuss use (linear or nonlinear) measures of statistical dependence between regions or predictive power of one or more regions on another region (either concurrently or later in time).

The goal of this project is to comparatively evaluate a variety of methods for measuring functional connectivity in fMRI data. In particular, we would like to understand the role of nonlinear dependencies, and to devise effective methods for inferring higher-order connectivity; that is, for distinguishing pairs of regions that are "directly" functionally connected from those for which functional connectivity can be explained by functional connectivity between intermediate regions). Finally, we would like to better understand the relationship between different measures of functional and structural connectivity, by studying the predictive power of the former on the latter.

2 Related Work

2.1 Overview of functional imaging

The dominant tool in the analysis of fMRI data over the past two decades has been the general linear model (GLM), [1] used primarily to detect average changes in local brain activity occurring in response to stimulus (Grinband et al. [2008]). Specifically, if Y is the fMRI data matrix (where $Y_{i,j}$ is the activation of the j^{th} ROI during the i^{th} trial of an experiment), X is the design matrix (where $X_{i,j}$ is value of the j^{th} stimulus covariate during the i^{th} trial of the experiment), the GLM fits a matrix β of regression coefficients (where $\beta_{i,j}$ is the regression coefficient of the i^{th} stimulus covariate of the j^{th} ROI) to the equation

$$Y = X\beta + U, \quad \text{where} \quad U_{i,j} \sim \mathcal{N}(0, \sigma^2),$$

using the ordinary least squares criterion.

[1] Widely known, in the context of neuroimaging, as *statistical parametric mapping* (SPM).

The GLM is a form of *mass-univariate* analysis in that it separately models the dependence of each ROI on the stimulus, while ignoring inter-voxel interactions. Over the past decade, however, it has become evident that rich functional information is encoded in the interactions between brain regions and how these change over time and in response to stimulus (Li et al. [2009]).

[TODO: Introduce basic functional connectivity analysis and its issues (outlining the following subsections).]

[TODO: Discuss grouping voxels as ROIs.]

2.1.1 Regions of Interest and Brain Parcellation

While the fMRI records data from over 10^5 voxels, before performing data analysis, we reduce the data by grouping voxels into *regions of interest* (ROIs). This process is referred to as *parcellation* of the brain. An ROI consists of a group of (spatially contiguous) voxels that has, been decided to form a single unit, based on prior knowledge either of the geometric or anatomic connectivity of the region, or of the functional role of the region. Within an ROI, the data from voxels are combined to form a single time series, typically by uniformly averaging the signals over voxels at each time point.

[TODO: Many possible parcellations, with different numbers and sizes of ROIs; choice depends on what is being studied. Occasionally (rarely), individual voxels are used as ROIs (i.e., the data are not changed at all; whole-brain study); cite FARM, and maybe others.]

We do this for at least 3 reasons:

1. The original data are quite noisy, in terms of both measurement and localization of the BOLD signal. ROIs are better localized, because our spatial scale (and hence error tolerance) increases from the order of millimeters to the order of centimeters, and suffer less measurement error, because we average signal over many (admittedly, highly dependent) voxels.

2. ROIs have biological interpretation as structural or functional regions of the brain, whereas voxels are essentially arbitrary partition of the brain.

3. ROI data is far more manageable, both computationally and statistically, because it not only reduces dimensionality but does so in a manner that drastically reduces collinearity in the data.

[TODO: Discuss resting-state fMRI (e.g., Biswal et al. [1995]).]

2.2 Simultaneous vs. Lagged Effects

[TODO: Add some discussion of time-scales (specifically, neural interactions are much faster than the BOLD signal), and how this affects analysis and interpretation.]

Cecchi et al. [2008] and Garg et al. [2011] develop a (very computationally expensive) [2] sparse full-brain autoregressive model to learn Granger causal relationships between individual voxels. They

[2] In fact, a key element of this work was access to IBM's Blue Gene/P supercomputer (Almasi et al. [2008]).

conjecture a "daring but falsifiable *information flow hypothesis* which states that the structure of the information flow in the fMRI data reflects, however partially, the flow of information in the underlying neural networks."

However, limitations of fMRI, and in particular of the BOLD signal, must be considered when measuring functional connectivity. For example, in a recent meta-analysis of various fMRI studies, Webb et al. [2013] studied the use of Granger causality to measure functional connectivity, and found that Granger Causality was strongly predicted by brain vasculature. Specifically, over a large collection of data sets, they showed that whether a voxel A Granger causes another voxel B is strongly influenced by ambient vascular anatomy, with A is systematically more likely than otherwise to Granger cause B if A is a vascular source, B is a vascular sink, or both. [3] This suggests that, default-state blood flow may sometimes have a stronger effect on Granger causality [4] of the BOLD signal than actual neural activity.

One response has been that functional connectivity should only be analyzed differentially; that is, functional connectivity should only be studied in terms of how it differs within a subject over time or between experimental conditions. [TODO: ADD CITATION; How is this consistent with using resting-state data to map functional networks?]

2.3 Linearity of functional connectivity

Hlinka et al. [2011] and Hartman et al. [2011] argue that linear correlation is sufficient to capture functional connectivity, citing an increase of only 5% in mutual information when accounting for nonlinear dependencies, and that differences in cluster and graph-theoretic properties of the data are unnoticeable.

2.4 Functional versus effective connectivity

Since we cannot typically perform causal interventions on the BOLD signal (let alone on brain activity), any attempts at causal understanding based on functional connectivity are subject to the confounding effects. The term *effective connectivity* refers to a more nuanced measure of connectivity incorporating a causal direction. [TODO: can only estimate effective connectivity in low-dimensions (i.e., a small number of ROIs).]

Li and Wang [2009] use a modification of the PC algorithm with bounded false discovery rate in order to learn a functional connectivity network via a series of Gaussian conditional independence tests.

2.5 Other related work

Klinkner et al. [2005] learn a state-space model over neurons from joint spike trains, and then estimate mutual information (normalized by minimum entropy) to fit a Chow-Liu tree over the neurons, pointing out that computing higher-order approximations to the true distribution is NP

[3] A *vascular source* or *sink* is a location where oxygenated blood enters or leaves (respectively) the brain.

[4] or other methods that measure directed functional connectivity by measuring the predictive power of a voxel on another voxel later in time.

hard. This simplifies the estimation and interpretation of mutual information in a time-series context, but requires fitting the data to a discrete state-space.

[TODO: Discuss Grosenick et al. [2013], Craddock et al. [2013], Carroll et al. [2009], Ryali et al. [2012], Greicius et al. [2003] .]

[Salvador et al. [2005] argue that conditional independence between multivariate time-series is better approached in the frequency domain than the time domain.]

3 Description of Data and Preprocessing

[TODO: Update this for new data sets] Each of 60 subjects (30 males, 30 females, ages 19-29) was scanned on a 3T Siemens Verio at the Scientific Imaging & Brain Research Center at CMU. Participants remained in a resting state throughout the experiment, remaining conscious and staring at a fixation point.

For each subject, the main object of the data is a length $T = 1{,}200$ sample time series of BOLD signal in each of $V = 160{,}990$ voxels. Specifically, for each subject, we have a $V \times T$ matrix. For each voxel, spatial (xyz) coordinates are also known.

The 160,990 voxels comprise 600 known regions of interest (ROIs), as identified by anatomical studies of the brain. One way to reduce the high dimensionality of the data is to aggregate data within these ROIs to produce a 600 dimensional time series. We do so simply by averaging the signal (at each time point) over all voxels within each ROI.

We first perform some basic denoising and normalization on the raw data. Denoising consists of removing large slow-wave oscillations via PCA by computing the first 3 eigenvectors of the signal, reconstructing the signal from those three eigenvectors, and then subtracting the result from the original signal. We then normalize the data by Z-scoring, so that each voxel has mean 0 and standard deviation 1 over time. An example of the resulting data is given in Figure 1.

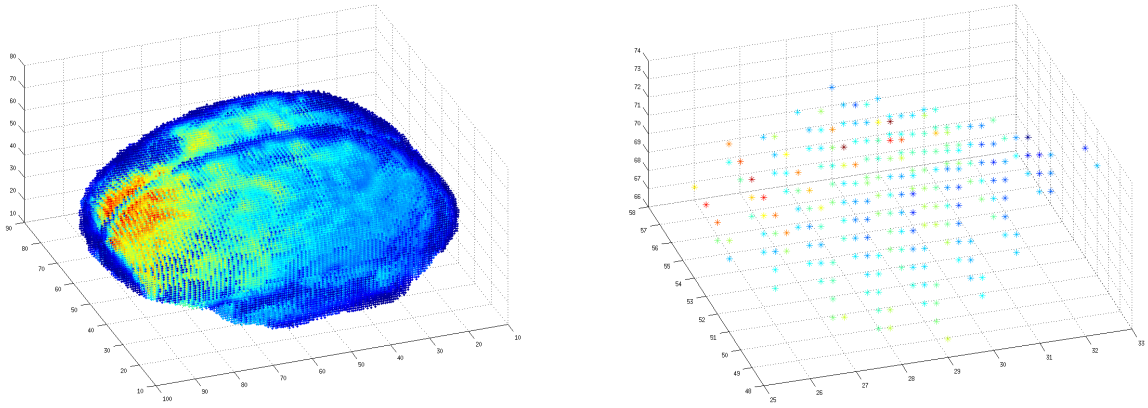
4 Statistical Formulation and Approaches

One statistical framework that approximates our problem is learning a minimal [5] graphical model whose variables are activity levels of ROIs. [TODO: discuss directedness of edges and conditionality of dependence within this graphical model.]

In addition to utilizing established methods such as variants of sparse linear regression and correlation, we would like to explore some more recently proposed methods, such a information-theoretic approaches, to see whether they can better measure functional connectivity (for example, by capturing nonlinear dependence or conditioning).

[TODO: Add table with detailed breakdown of all methods (including the following properties: pairwise/seedwise (regression)/full graph, IID/time series/in between (cross-correlation).]

[5] In the sense of having no unnecessary edges.



(a) 1a (b) 1b

Figure 1: BOLD signals for a single time frame from a single subject for all voxels (1a) and for all ROIs, averaged over voxels in each ROI (1b). Red indicates high activity in a voxel, while blue indicates lower activity. In this case, the highest activity is observed in prefrontal areas.

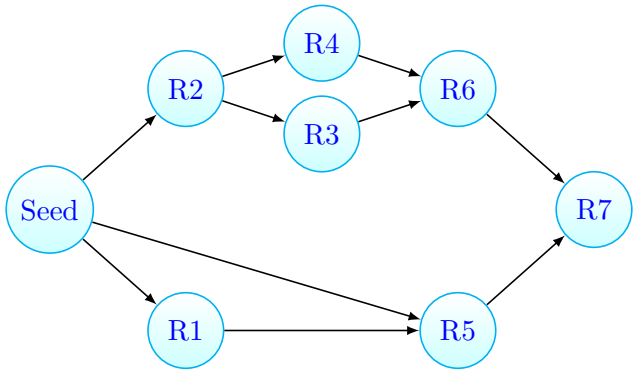


Figure 2: TODO: caption

5 Regression and Sparsity Based Methods

One measure of the relative dependence of a variable Y on other variables X is the coefficient of linearly regressing Y over X. In particular, assuming the linear model holds, if the coefficient is zero, then Y is conditionally independent of X given the other covariates. This leads to a framework for identifying functional connectivity that has recently received much attention (e.g., Grosenick et al. [2013], Craddock et al. [2013], Carroll et al. [2009], Ryali et al. [2012]): for each seed ROI Y, regress Y over all other ROIs with some manner of variable selection or sparsity constraint.

It is important to note that we are interested in using these methods *only* for variable selection (and not for parameter estimation). Hence, we avoid the problem of post-selection inference, a common source of methodological error in such methods.

	Parametric (Linear)		Semi/Nonparametric	
	Prediction-Based	Similarity-Based	Prediction-Based	Similarity-Based
Time Series	Granger Causality	Cross-Correlation	Transfer Entropy	Cross-Information
IID	Elastic Net	Correlation	FuSSO	Mutual Information

Table 1: Conceptual breakdown of several of the dependence measures we consider.

5.1 Lasso and Elastic Net

[TODO: Write down the objective functions for Lasso and Elastic Net.] The above framework makes the automatic neighborhood selection performed by L_1 regularization an obvious candidate. However, because we also expect multicollinearity amongst nearby voxels (and may also like to be able to select more than n voxels), L_2 regularization is also desirable. Hence, we propose to use elastic net regularization, a weighted combination of L_1 and L_2 regularization. Specifically, we use the elastic net regression implementation in the MATLAB package `glmnet`. [6]

Specifically, for each ROI, we used the elastic net to regress that ROI over all other ROIs, and consider ROIs i and j to be connected if and only if the coefficient of regressing i over j is nonzero, or vice versa. This has the additional benefit of being extremely computationally fast for a range of penalty weights (used as our threshold parameter), due to fast algorithms for computing the lasso path. One could also connect ROIs more conservatively by required both the coefficient of regressing i over j to be nonzero, *and* vice versa. [TODO: Mention that this idea is from Meinshausen and Bühlmann [2006].]

When using lasso or the elastic net on a very large number of ROIs, pre-screening the ROIs [TODO: cite Ryali et al., who used elastic net.]

5.2 Forward-Backward Stepwise Regression

A classical method for variable selection in linear regression is the forward-backward stepwise method introduced by Efroymson [1960]. This method iteratively adds and removes variables by performing, in each iteration, an F-test for whether each variable should be included in the linear model. Forward-backward stepwise regression has received much criticism from the statistical community as being a "data dredging" technique, and, to the best of our knowledge, forward-backward stepwise regression has not been previously used in the problem of functional connectivity detection (perhaps because much of this criticism came before the advent of fMRI).

Our motivation for using stepwise regression (rather than just) sparse regression when detecting functional connectivity is that stepwise regression gives relatively more information about *which* ROI(s) mediate(s) the relationship between two ROIs that are correlated but not partially correlated, given certain all other ROIs, by studying the full history of the stepwise regression (i.e., which and in what order variables were added/removed). Furthermore, since we use the technique only for variable selection and not inference, we avoid most of its controversial aspects.

[6] The MATLAB implementation of `glmnet` that we use is available at `http://web.stanford.edu/~hastie/glmnet_matlab/`.

6 Dependence Measure Based Methods

Let X and Y denote random variables corresponding to time series from two distinct ROIs. Suppose a statistic $D(X, Y)$ is a *dependence measure*, by which we mean that D satisfies

$$D(X, Y) \geq 0, \text{ with equality if and only if } X \text{ and } Y \text{ are } \textit{not} \text{ functionally connected.} \quad (1)$$

If we have a sufficiently accurate estimator $\hat{D}$ of $D(X, Y)$ and we assume that randomly permuting the time series Y destroys an functional connectivity with X, then we can perform a permutation test [7] to test whether $D(X, Y) = 0$, and hence whether X and Y are functionally (dis)connected, at a false positive rate α. [8] In this section, we propose a several dependence measures D which we use with permutation tests to detect functional connectivity.

6.1 Linear Methods

Correlation: Assuming that functional connectivity is captured by the joint stationary distribution of X and Y, and that this joint stationary distribution is Gaussian, the simplest dependence measure to use is the correlation,

$$\rho(X, Y) = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]\,\mathrm{Var}[Y]}.$$

If, furthermore, the joint time series (X, Y) mixes sufficiently quickly, we can estimate correlation well using the sample correlation.

6.1.1 Time Series Methods

Cross-Correlation: Using correlation makes extremely strong assumptions on the temporal dependence of the data; specifically, if assumes that only concurrent samples of X and Y are dependent. Cross-correlation, introduced in functional connectivity analysis by Cao et al. [1999], generalizes correlation to account for lags in the dependence of one ROI on another. Specifically, [TODO: elaborate.]

Coherence: The sample coherence (see, e.g., Müller et al. [2001]) roughly measures correlation in the frequency domain is another dependence measure widely used in functional connectivity analysis. [TODO: elaborate.]

Granger Causality: For each pair of ROIs i and j, perform an F-test to determine whether ROI i Granger causes ROI j, using BIC to determine the time lag and tuning the Type I error probability α through edge cross-validation. [9] [TODO: elaborate.]

[7] For fMRI data, the null hypothesis (that two ROIs are independent) does not yield *all* permutations as equally likely, due to the presence of autocorrelation. Hence, we actually perform a *block* permutation test, where we split the data into a small number (e.g., 8-10) of consecutive blocks, permute these blocks (retaining the order of data *within* each block), and then estimate the dependence between paired blocks of the two ROIs (averaged over blocks).

[8] Although, in practice, α would typically be fixed beforehand (e.g., $\alpha = 0.05$), in the sequel, to compare different methods, we will treat α as the free parameter underlying the ROC curves.

[9] See `http://www.mathworks.com/matlabcentral/fileexchange/25467-granger-causality-test` for a MATLAB implementation of the Granger Causality Test that we use.

Previous work (Webb et al. [2013]) has noted that Granger Causality is affected by the vascular structure of the brain, which is independent of the functional activity we are trying to measure. For example, brain regions serving as vascular sources tend to appear as Granger causal, while regions serving as vascular sinks tend to appear as Granger consequential.

Transfer Entropy: Given a joint time series $(\mathcal{X}, \mathcal{Y}) = \{(X_i, Y_i)\}_{i=1}^{\infty}$, the transfer entropy $T^n_{\mathcal{X}\to\mathcal{Y}}$ from X to Y at time n is

$$\begin{aligned} T^n_{\mathcal{X}\to\mathcal{Y}} &= H(Y_t|Y_1, ..., Y_{n-1}) - H(Y_t|X_1, ..., X_{n-1}, Y_1, ..., Y_{n-1}) \\ &= I(Y_t; X_1, ..., X_{n-1}|Y_1, ..., Y_{n-1}), \end{aligned}$$

(i.e., the shared information between the head of $\mathcal{Y}$ and the past of $\mathcal{X}$, given the past of $\mathcal{Y}$). Introduced by Schreiber [2000] and applied to neural data (magnetoencephalography (MEG) recordings) by Vicente et al. [2011], transfer entropy captures both (asymmetric) time series and non-linear aspects of dependencies between variables. Importantly, transfer entropy does not include the mutual information between X_n and Y_n, except through the past of $\mathcal{X}$. A closely related quantity is the *mutual information rate* (MIR) $I_R(\mathcal{X}; \mathcal{Y})$ defined by

$$I_R(\mathcal{X}; \mathcal{Y}) = \lim_{n\to\infty} \frac{I(X_1, ..., X_n; Y_1, ..., Y_n)}{n}.$$

Although this is not entirely obvious, [10] the MIR can be re-written as

$$I_R(\mathcal{X}; \mathcal{Y}) = \lim_{n\to\infty} T^n_{\mathcal{X}\to\mathcal{Y}} + T^n_{\mathcal{Y}\to\mathcal{X}} + I(X_n; Y_n|X_1, ..., X_{n-1}, Y_1, ..., Y_{n-1}).$$

From this, one can see that the MIR differs from the transfer entropy in three ways: (1) the MIR is a limit as $n \to \infty$, rather than an instantaneous measure, (2) the MIR is symmetric in $\mathcal{X}$ and $\mathcal{Y}$, and (3), the MIR includes concurrent information between X_n and Y_n (given their pasts). Of course, any combination of these three differences gives rise to a different quantity, each with its own potentially interesting properties and interpretation.

Unfortunately, capturing time-series dependencies requires [TODO: Explain dimensionality issues and Markov assumption.]

Barnett et al. [2009] note that, in the case of jointly Gaussian time series, non-zero transfer entropy is exactly equivalent to Granger causality. Since our fMRI data is typically approximately Gaussian, we may expect transfer entropy to behave very similarly to the Granger causality test. However, unless we explicitly make the assumption that the data are jointly Gaussian, (nonparametric) *estimation* transfer entropy in practice may be significantly more difficult than performing the Granger causality test.

6.2 Non-linear Methods

Mutual Information: For two random variables X and Y taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, the *Shannon mutual information* $I(X; Y)$ between X and Y is defined as

$$I(X; Y) := \int_{\mathcal{X}\times\mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = \mathbb{E}_{X,Y}\left[\log \frac{p(X, Y)}{p(X)p(Y)}\right],$$

[10] TODO: Cite 10-704 project report.

where p denotes the appropriate (marginal or joint) density function in each instance. Intuitively, $I(X;Y)$ is a measure of dependence between X and Y (for example, if X and Y are Gaussian, then is a (strictly) increasing function of the absolute value of the correlation $|\rho(X,Y)|$). Furthermore, classic results in information theory are that $I(X;Y) \geq 0$ and that $I(X;Y) = 0$ if and only if X and Y are independent. Recently, there has been much work on estimating mutual information using samples from a joint density (Singh and Póczos [2014], Krishnamurthy et al. [2014], Moon and Hero [2014]). We use an older estimator due to Kraskov et al. [2004] that is less well-understood theoretically, but is computationally quite efficient, and has been shown empirically to have good statistical error in many cases.

Distance Correlation: The distance correlation, introduced by Székely et al. [2007], is a modification of the correlation satisfying condition (1) for *any* joint distribution with finite second moments (whereas the correlation satisfies this only for linearly dependent variables). The *distance covariance* $\mathrm{dCov}[X,Y]$ between X and Y is defined as the grand mean of the element-wise product of the pairwise distance matrices of the two samples. The distance correlation $\mathrm{dCorr}[X,Y]$ between X and Y is then written in terms of the distance covariance by the same formula relating the usual covariance to the usual correlation:

$$\mathrm{dCorr}[X,Y] := \frac{\mathrm{dCov}[X,Y]}{\sqrt{\mathrm{dCov}[X,X]\,\mathrm{dCov}[Y,Y]}}.$$

Unfortunately, estimating the distance correlation takes time quadratic in the sample size, and so the permutation test is quite slow. [11] [TODO: See if Székely et al. [2007] has a faster test.]

7 Graphical Model Approaches

[TODO: Introduce these methods as those which estimate all edges (the whole graph) together.

7.1 Graphical Lasso

Under the assumption that the ROIs are jointly Gaussian with covariance matrix Σ, the precision matrix Σ^{-1} encodes the partial correlations of each pair of ROIs given all other ROIs. This somewhat captures the notion of functional connectivity, in the sense that a pair of ROIs is likely functionally connected if this partial correlation is nonzero. On the other hand, this partial correlation may be zero even if the pair is functionally connected (e.g., if three ROIs are identical, then the partial correlation between any two conditioned on the other will be 0).

The graphical lasso (also known as the *glasso*), introduced by Friedman et al. [2008], estimates a sparse approximation Θ of Σ^{-1} by solving the optimization problem

$$\operatorname*{argmax}_{\Theta \succeq 0} \log\det\Theta - \mathrm{tr}(\Sigma\Theta) - \rho\|\Theta\|_1,$$ [12]

This (extremely computationally efficient) algorithm gives a way of identifying (a subset of) pairs of ROIs that should be functionally connected. In particular, we run the graphical lasso on the

[11] See `http://www.mathworks.com/matlabcentral/fileexchange/49968-dcorr--x--y--` for a MATLAB implementation of the Distance Correlation measure that we use.

[12] $\{\Theta \succeq 0\}$ is the set of positive semi-definite matrices. This is in fact a Gaussian maximum likelihood estimate.

matrix $\hat{\Sigma}$ of empirical covariances of the ROIs (tuning the coefficient ρ of the L_1 penalty of Θ through edge cross-validation) to produce and empirical precision matrix $\hat{\Theta}$, and identify ROIs i and j as functionally connected if and only if $|\hat{\Theta}_{i,j}| > 0$.

7.2 Chow-Liu Approximation

The Chow-Liu tree, introduced by Chow and Liu [1968], is defined as the graphical model with distribution $\hat{P}$ that minimizes the KL divergence $D_{KL}\left(P \middle\| \hat{P}\right)$ between the true distribution and the approximated distribution $\hat{P}$, over all tree-structured distributions $\hat{P}$. It can be shown that

$$D_{KL}\left(P \middle\| \hat{P}\right) = -\sum_{i=1}^{n} I(X_i, X_{(i)}) - H(X) + \sum_{i=1}^{n} H(X_i),$$

where (i) denotes the parent of the i^{th} variable in the tree (for an arbitrary root node). The first term is clearly minimized (over (i)'s) by the maximum mutual information spanning tree (MMIST), while the latter terms do not depend on the (i)'s. This suggests the Chow-Liu algorithm, which estimates the Chow-Liu tree by the empirical MMIST. The empirical MMIST is itself computed using estimated pairwise mutual informations. The tree can be further sparsified to a forest by removing the edges with the least mutual information, so that the number of edges can be chosen via edge cross-validation.

7.3 PC Algorithm

[TODO: Outline this and provide citation.]

8 Stepwise Approaches

Here, we discuss a few methods that attempt to build a local dependence tree around a seed voxel by iteratively identifying the direct neighbors of the seed, then identifying their neighbors, then identifying those neighbors' neighbors, etc. We call these methods *stepwise* approaches because the build the dependence tree one *layer* at a time.

8.1 Naïve Stepwise Approach

Suppose we want to distinguish different orders of connectivity; that is, if a voxel C is connected to a voxel A through another voxel B, but C is not connected directly to A, we would like distinguish this from the case that C is connected directly to A.

Given a seed voxel s, one approach to determining the order of connectivity might be the following:

1. Fix seed ROI s.
2. Restrict to subset $C = \{c_1, \ldots, c_n\}$ of voxels correlated with seed s.

3. For each c_i in C, compute $p_i := \min_{j \in C} |\hat{\rho}(c_s, c_i | c_j)|$.

4. If $p_i > \rho_0$, put i in Layer 1

5. Else, make i child of $j = \text{argmin}_{j \in C} |\hat{\rho}(s, i|j)|$.

8.2 Stepwise Clustering Approach

One issue with the above stepwise procedure is that its results highly are unstable with respect to addition of new voxels. This is, in general, true of most methods which identifies the set of Layer 1 neurons before identifying the sets of children of each Layer 1 neuron.

Hence, it may be desirable to first identify sets of of voxels belonging to each subtree, and then identify the root of each subtree. The following is one such approach.

The symmetrized partial correlation $|\rho(s, v_1|v_2)| + |\rho(s, v_2|v_1)|$ can be thought of as a distance between v_1 and v_2, measuring the difference in the information they share with s.

1. Restrict to subset $C = \{c_1, \ldots, c_n\}$ of voxels correlated with seed s.

2. Compute a matrix $D \in \mathbb{R}^{n \times n}$ such that each

$$D_{i,j} = \frac{1}{2} \left(|\hat{\rho}(s, c_i | c_j)| + |\hat{\rho}(s, c_j | c_i)| \right).$$

3. Fix a number L_1 of voxels to be placed in the first layer.

4. Find L_1 clusters via any clustering algorithm that is a function of D (e.g., spectral clustering).

5. For each cluster $C_1, \ldots, C_{L_1}$, somehow identify a representative voxel r_i to connect to s.

6. Make all other voxels in C_i children of r_i.

9 Nonlinear Dependencies and Functional Connectivity

[TODO: Introduce this section, recalling motivation, prior work, and questions]

9.1 Detecting Nonlinearity

[TODO: Interpret these plots]

9.2 Testing for nonlinearity

In the previous section, we motivated the difference $I(X;Y) - I_{lin}(X;Y)$ as a measure of nonlinearity of a dependency, under the assumption that the marginals of X and Y are Gaussian. [TODO: Justify this assumption.] In particular, $I(X;Y) - I_{lin}(X;Y) \geq 0$, with equality if and only if the dependence between X and Y is entirely explained by their correlation. Hence, for a given

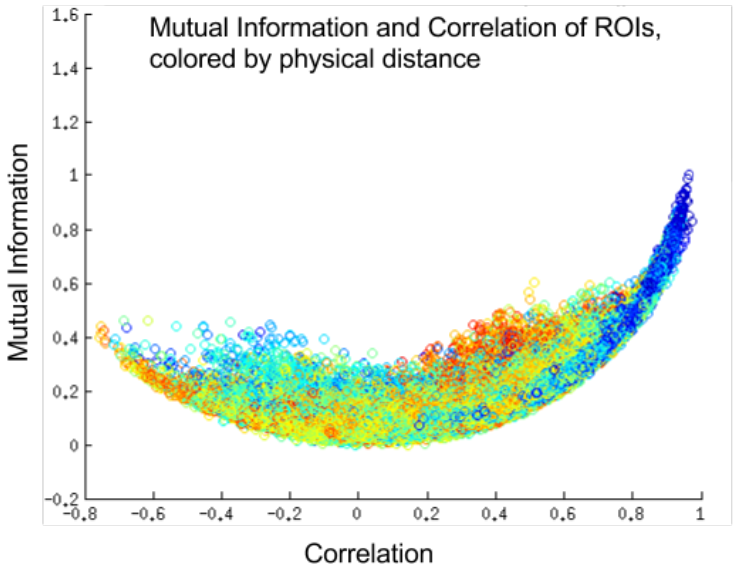


Figure 3: Plot includes one point for each pair of ROIs ($\approx$ 3600 pairs). Distance above curve indicates non-linearity of dependence.

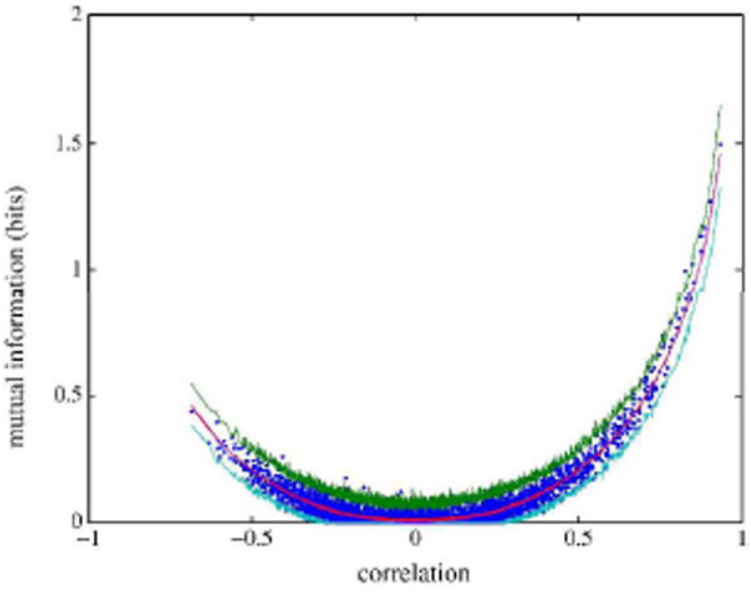


Figure 4: Same figure (colored by MI percentile), as presented in Hlinka et al. [2011].

$\alpha \in (0, 1)$, it is straightforward to perform an α-level permutation test of the null hypothesis H_0 that any dependence between X and Y is explained by correlation against the alternative H_1 that there is a nonlinear dependence between X and Y. Specifically, for a large number N of permutations and a type I error rate α, the test is conducted as follows:

1. Estimate $\hat{I}_0 := \hat{I}(X;Y) - \hat{I}_{lin}(X;Y)$.

2. For $i \in \{1, \ldots, N\}$, estimate $\hat{I}_i := \hat{I}(X; Y_i') - \hat{I}_{lin}(X; Y_i')$ for a random permutation Y_i' of Y.

3. Let $p := \frac{1}{N} \sum_{i=1}^{N} 1_{\{I_0 < I_i\}}$ and reject H_0 if $p < \alpha$.

9.3 Results, Further Investigations, and Ramifications

We performed the above test for a uniformly random sample of 1500 pairs of ROIs displaying a correlation strength of $\rho \geq 0.2$. At the $\alpha = 0.02$ confidence level, we rejected H_0 for 844 pairs (56% of pairs); i.e., 56% of pairs were identified as significantly nonlinearly related.

However, further investigation of these pairs revealed that the appearance of nonlinear dependencies may in fact be recording artifacts rather than manifestations of true functional connectivity.

Figure 5 shows the joint activities of some of these pairs over the course of the experiment. Here, the behaviors of many ROIs appear much more strongly dependent on time than on the paired ROI. In particular, most of the dependencies appear to consist of a ROI's activity first increasing and then decreasing (or vice versa), while another ROI's activity steadily increases or decreases, over the course of the experiment. On the other hand when looking at pairs of ROIs with significant linear relationships but *not* significant nonlinear relationships (see Figure 6), time of recording appears to have little or no effect on the ROI behaviors, while each paired ROI appears to exert significant influence. Hence, the changes underlying *nonlinear* dependence may be driven more by temporal drift (presumably an artifact of the signal) than by functional connectivity.

[TODO: discuss how we attempt to correct for this via linear and non-parametric regression.]

10 Comparison to Structural Data

[TODO: Describe basal ganglia structure (including diagram).]

[TODO: Justify structural connectivity as a baseline for functional connectivity, including the necessary caveats/assumptions (e.g., functional should average to structural over enough subjects and time).]

10.1 Graph Analysis

[TODO: Insert true and estimated circular graphs]

10.2 ROC Analysis

The methods discussed in the previous sections are designed to work with data from a single scan of a single subject. Hence, from each scan subject, we can estimate a functional connectome, and evaluate quality of the prediction against structure. Specifically, we can analyze the trade-off between detection ability (hit rate) and precision (false positive rate). Furthermore, since we have many predictions, assuming the data are IID across scans, we can estimate confidence for each proportion using a simple binomial model (treating each hit or false positive as a Bernoulli random variable). Figure 7 shows these ROC curves for several methods.

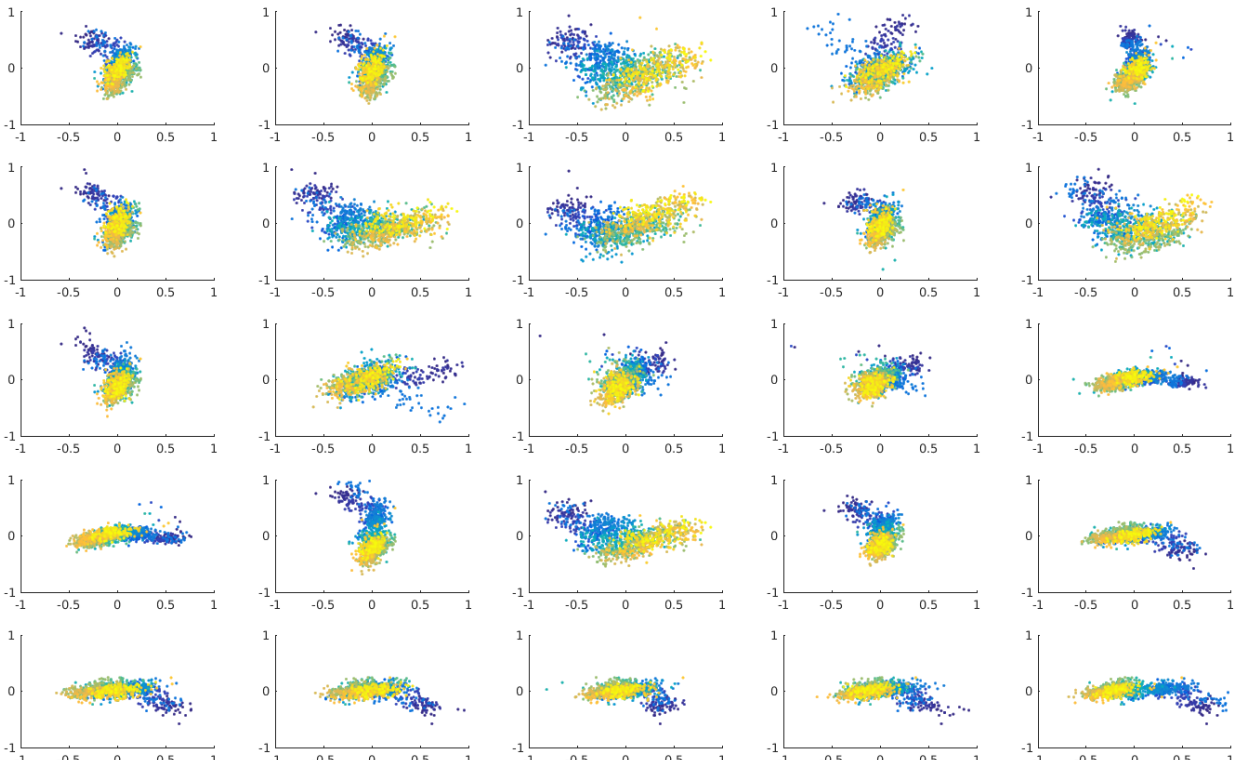


Figure 5: Scatter plots of the activities of each of 25 pairs of ROIs, colored by time of recording. Blue indicates time early in the recording, while yellow indicates time later in the recording. The 25 pairs were chosen uniformly at random from the 844 pairs identified as significantly nonlinearly dependent.

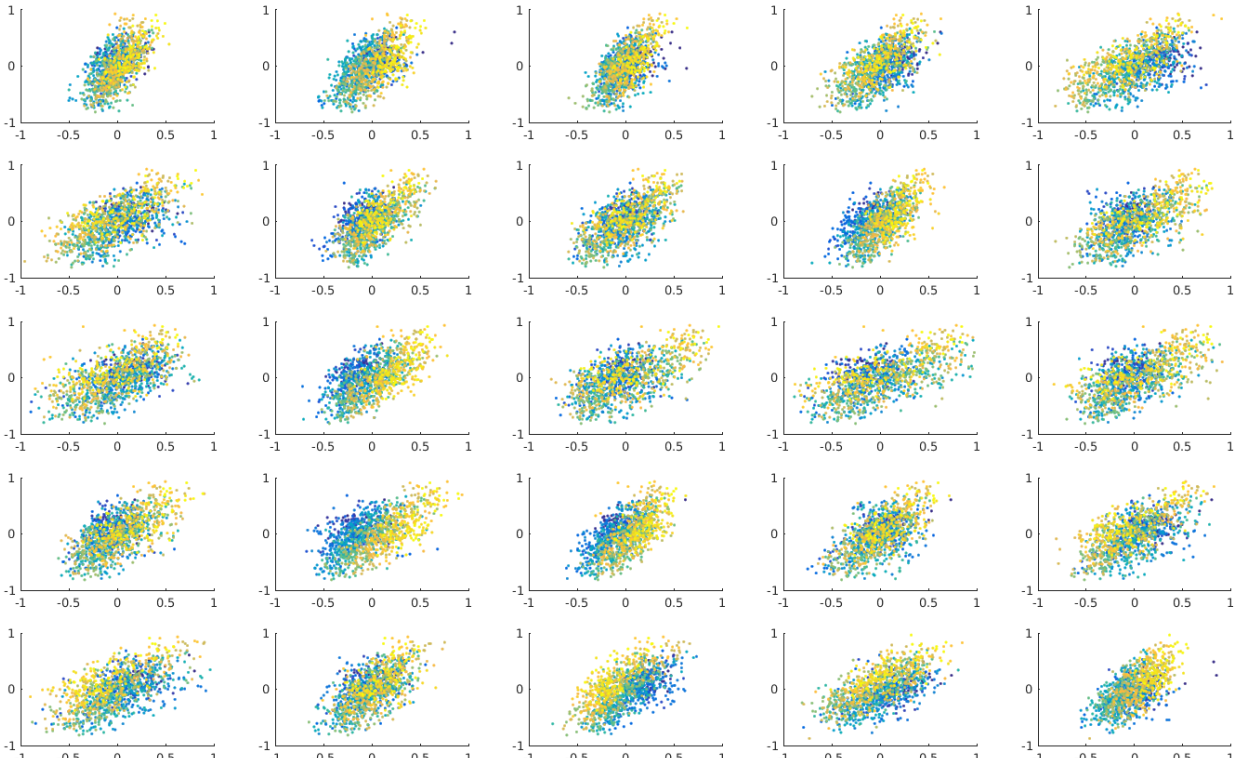


Figure 6: Similar plot as Figure 5 except using 25 ROI pairs randomly sampled from those identified as significantly linearly correlated but *not* significantly nonlinearly dependent.

On the other hand, the assumptions discussed in the beginning of this section suggest that functional connectivity should reflect structure only over sufficiently long scans and over sufficiently many subjects. Thus, our predictions may be more accurate (and more robust) if we pool the estimated functional connectome over subjects, producing a single connectome rather than subject-wise connectomes. Hence, we also present ROC curves for this approach, whose accuracy can be compared to that of subject-wise predictions. Since we have only one prediction for each edge, estimating confidence intervals for this ROC curve is more difficult, and we resort to resampling 60 scans (with replacement) 1000 times to generate bootstrapped versions of the pooled connectome, from which we estimate confidence intervals. Figure 8 shows these ROC curves for several methods.

[TODO: FIX SWAPPING OF HR AND FPR!]

10.3 Discussion of ROC Analysis Results

[TODO: Discuss how pairwise methods actually estimate unconditional graphical model, and why mutual information gives the closest (of these) to the conditional model (via data processing inequality).]

Varying the Pooling Threshold: Empirically, we see that many of the methods (especially pairwise methods, such as Granger causality and those based on permutation tests), tend to be too aggressive (selecting too many edges). This is not surprising, since almost all pairs of ROIs are dependent, even if their dependence is indirect. One way to counteract this is by pooling edges more conservatively. We define a *pooling threshold*, $[0, 1]$-valued parameter indicating the fraction of scans in which an edge must appear in order for it to be included in the pooled graph estimate. [TODO: verify that this makes sense and discuss our choice of pooling thresholds (50% and 95%).]

11 Data Analysis TODOs

1. Rerun methods without nonparametric regression (just linear regression).
2. Run ROC analysis on distance correlation.

12 Future Work

12.1 Other Methods

Post-Selection Inference: We used sparse regression methods (Lasso, elastic net, and forward-backward regression) purely for their variable selection ignoring the estimated values of the regression coefficients (and, it particular, whether there is significant evidence that they are nonzero). While this avoids the biases inherent in post-selection inference, it does leave open the possibility that we are including edges between variables having no statistically significant relationship. Very recently, statistical tools have been developed to test significance of regression coefficients after certain variable selection procedures such as Lasso and forward stepwise regression (Taylor et al. [2014]). Using these procedures might give a more precise estimate of the connecitivity graph by excluding some edges that are included by the variable selection simply due to random chance, as

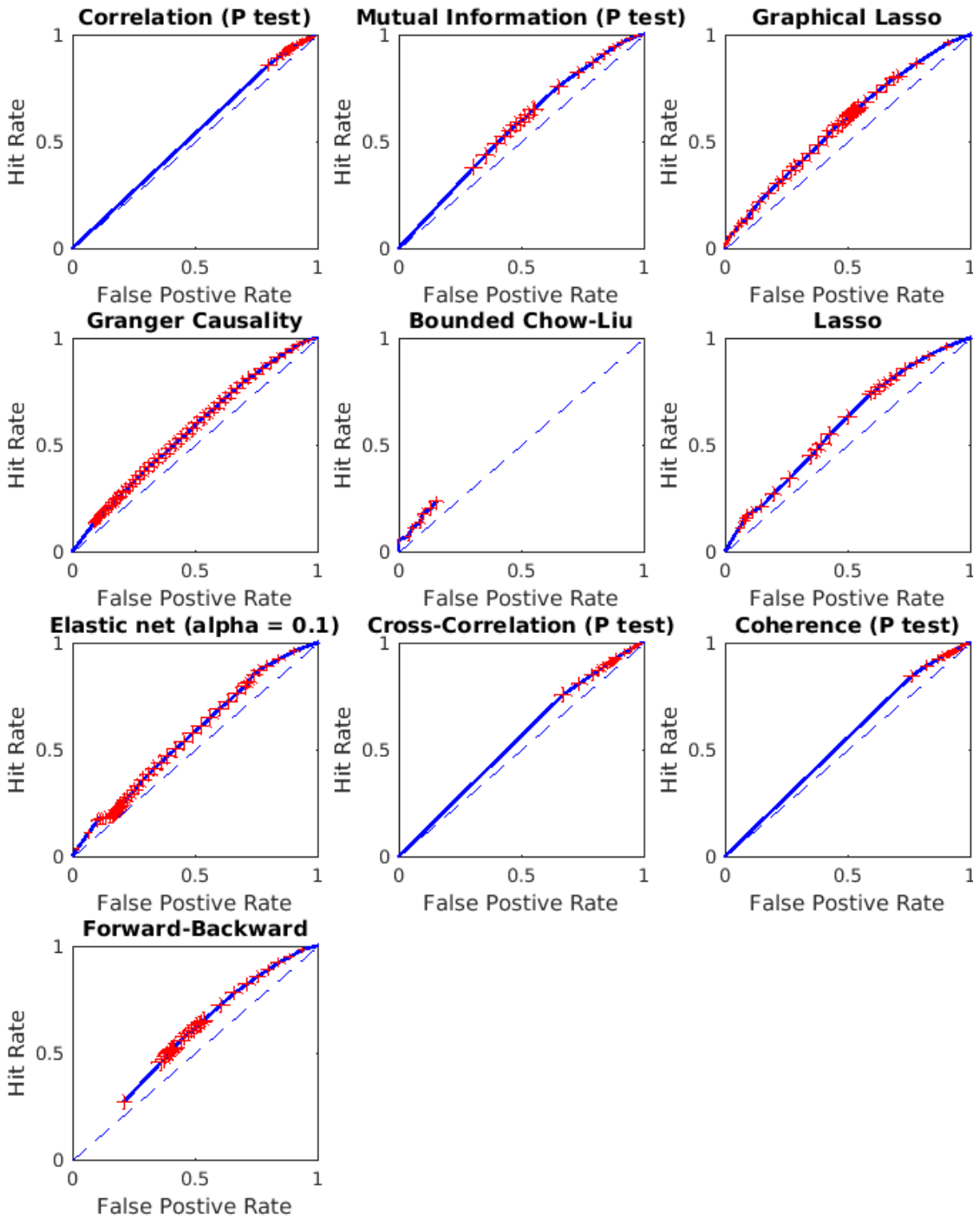


Figure 7: Mean scan ROC curves for each method. Hit and false positive rates were computed for the adjacency matrix estimated for each scan and parameter value, and averaged over scans. Blue lines are linear interpolations of these averaged ROC points, and crosses are 95%-confidence intervals for each point (estimated using a central limit approximation to a binomial model).

well as giving a standardized measure (a conditional p-value) of the confidence with which an edge is included.

Regularized Local Chow-Liu: The Bounded Chow-Liu algorithm we tested appears to perform well at identifying structural edges, but is overly constrained by fitting a tree to the *entire* set of ROIs (which necessarily has fewer edges than we expect to be present).

The Chow-Liu algorithm aims to find the most explanatory tree graphical model for the entire set of variables. We conjecture that there exists a simple modification of the Chow-Liu algorithm, a *local Chow-Liu algorithm* which aims instead to find the most explanatory tree graphical model for any particular seed variable. Such an algorithm might seek to minimize the KL-divergence of the estimated conditional distribution of the seed variable from the true conditional distribution; however, simply doing this provably results in a star graph, with every variable connected directly to the seed. [TODO: add a proof of this to the appendix.] A better approach might add a regularization term, increasing with the degree of the seed variable.

Such an approach would consistent with a prevalent scientific opinion that brain connectivity ought to be *locally tree-like* (though not globally a tree), in order to information to follow processing *streams* or pathways (e.g., the dorsal and ventral visual streams (Goodale and Milner [1992])). [13]

[13] A *directed acyclic graph* (DAG) may be a more accurate, but harder-to-estimate, structural model.

References

Gheorghe Almasi, Sameh Asaad, Ralph E Bellofatto, H Randall Bickford, Matthias A Blumrich, Bernard Brezzo, Arthur A Bright, Jose R Brunheroto, Jose G Castanos, Dong Chen, et al. Overview of the ibm blue gene/p project. *IBM Journal of Research and Development*, 52(1-2):199–220, 2008.

Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4): 537–541, 1995.

Jin Cao, Keith Worsley, et al. The geometry of correlation fields with an application to functional connectivity of the brain. *The Annals of Applied Probability*, 9(4):1021–1057, 1999.

Melissa K. Carroll, Guillermo A. Cecchi, Irina Rish, Rahul Garg, and A. Ravishankar Rao. Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1):112 – 122, 2009. ISSN 1053-8119.

Guillermo A Cecchi, Rahul Garg, and A Ravishankar Rao. Inferring brain dynamics using granger causality on fmri data. In *ISBI*, pages 604–607, 2008.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14:462–467, 1968.

R. Cameron Craddock, Michael P. Milham, and Stephen M. LaConte. Predicting intrinsic brain activity. *NeuroImage*, 82(0):127 – 136, 2013. ISSN 1053-8119.

MA Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, 1:191–203, 1960.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Rahul Garg, Guillermo A. Cecchi, and A. Ravishankar Rao. Full-brain auto-regressive modeling (FARM) using fMRI. *NeuroImage*, 58(2):416 – 441, 2011. ISSN 1053-8119.

Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

Michael D Greicius, Ben Krasnow, Allan L Reiss, and Vinod Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences*, 100(1):253–258, 2003.

Jack Grinband, Tor D Wager, Martin Lindquist, Vincent P Ferrera, and Joy Hirsch. Detection of time-varying signals in event-related fmri designs. *Neuroimage*, 43(3):509–520, 2008.

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E. Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72(0):304 – 321, 2013. ISSN 1053-8119.

D. Hartman, J. Hlinka, M. Paluš, Dante Mantini, and M. Corbetta. The role of nonlinearity in computing graph-theoretical properties of resting-state functional magnetic resonance imaging brain networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(1):013119, 2011.

Jaroslav Hlinka, Milan Paluš, Martin Vejmelka, Dante Mantini, and Maurizio Corbetta. Functional connectivity in resting-state fMRI: is linear correlation sufficient? *Neuroimage*, 54(3):2218–2225, 2011.

Kristina Lisa Klinkner, Cosma Rohilla Shalizi, and Marcelo F. Camperi. Measuring shared information and coordinated activity in neuronal networks. *arXiv preprint q-bio/0506009*, 2005.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman. Nonparametric estimation of Renyi divergence and friends. In *International Conference on Machine Learning (ICML)*, 2014.

Junning Li and Z. Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the PC algorithm. *J. Mach. Learn. Res.*, 10:475–514, June 2009. ISSN 1532-4435.

Kaiming Li, Lei Guo, Jingxin Nie, Gang Li, and Tianming Liu. Review of methods for functional brain connectivity detection using fmri. *Computerized Medical Imaging and Graphics*, 33(2):131–139, 2009.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

K.R. Moon and A.O. Hero. Ensemble estimation of multivariate f-divergence. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 356–360, June 2014.

Karsten Müller, Gabriele Lohmann, Volker Bosch, and D Yves Von Cramon. On multivariate spectral analysis of fmri time series. *NeuroImage*, 14(2):347–356, 2001.

Srikanth Ryali, Tianwen Chen, Kaustubh Supekar, and Vinod Menon. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852 – 3861, 2012. ISSN 1053-8119.

Raymond Salvador, John Suckling, Christian Schwarzbauer, and Ed Bullmore. Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):937–946, 2005.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Shashank Singh and Barnabás Póczos. Exponential concentration of a density functional estimator. In *Advances in Neural Information Processing Systems*, pages 3032–3040, 2014.

Klaas Enno Stephan and Karl J. Friston. Analyzing effective connectivity with functional magnetic resonance imaging. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(3):446–459, 2010.

Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889*, 2014.

Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropya model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

J. Taylor Webb, Michael A. Ferguson, Jared A. Nielsen, and Jeffrey S. Anderson. BOLD Granger causality reflects vascular anatomy. *PLoS ONE*, 8(12):e84279, 12 2013.

A Parameter Selection

A.1 Parameter of each Method

Dependence Methods (Correlation, Mutual Information, Cross-Correlation, Distance Correlation, Coherence): The primary parameter for most of these methods is the Type I error rate $\alpha \in [0, 1]$, which is swept from 0 to 1 to generate ROC curves. 200 permutations were used for permutation tests, largely due to computational constraints; the relatively small number of permutations likely accounts for the coarse parameter coverage of some of the ROC curves generated in this manner.

Graphical Lasso: The graphical lasso has a single regularization/penalty parameter, $\rho \in [0, \infty)$. ρ was swept uniformly from 0 to 1 to generate ROC curves..

Granger Causality: The Granger Causality test has two parameters: the Type I error rate $\alpha \in [0, 1]$ (which was swept from 0 to 1 to generate ROC curves) and the maximum lag T_{max}, which identifies the maximum that the "causal" time series should be lagged before the response time series. The actual lag is chosen from this range of lags using the Bayesian information criterion (BIC), and hence the exact value of T_{max} is unlikely to affect the choice, as long as it is sufficiently large. We used $T_{max} = 3$, since we expect functional connectivity.

Elastic Net/Lasso: The elastic net has two parameters: the regularization/penalty weight $\lambda \in [0, \infty)$ and the L_1-L_2 trade-off weight $\alpha \in [0, 1]$. To generate ROC curves, we swept λ over 50 (logarithmically distributed) values ranging from 10^{-14} to 10^{0} and simply use $\alpha = 1$ (Lasso) and $\alpha = 0.1$ to get a good contrast.

Forward-Backward Regression: Forward-Backward Regression has a single parameter: the maximum p-value $\alpha \in [0, 1]$ below which a variable is added to the model; α is swept from 0 to 1 to generate ROC curves. [14]

A.2 Edge Cross-Validation

Unlike standard prediction problems where a prediction is made for each sample, when learning a graphical model, we make a (relatively small) fixed number of predictions (one for each edge) using all samples of the covariates. This makes it difficult to perform cross-validation in order to select models or parameters, while still maintaining independent training and test datasets. We attempt to solve this by splitting our ground truth (which we define here as the structural connectivity), selecting models/parameters which minimize error on the first half (the *training set*), and then evaluating accuracy on the second half (the *test set*). Since we are optimizing over only a small number of parameters (at most 2, as in the elastic net and granger causality approaches), we are unlikely to over-fit. We call this procedure *edge cross-validation*, and use it throughout our evaluation. We average test errors over several training-test splits (using the same training-test splits across methods, for fairness).

[14] We use the same α as the *minimum* p-value below at which a variable is *removed* from the model.

B Visualization

[TODO: finish this section and move to appendix.]

The high dimensional nature of fMRI data makes visualizing results quite difficult. This even worse when we are studying connectivity, because there are $O(n^2)$ possible pairwise connections amongst n voxels. Here, we discuss some basic approaches to visualizing fMRI data.

B.1 Brain-Like Visualizations

For biological interpretation, it would be desirable to preserve the relative positions of voxels or ROIs within the brain. When plotting all voxels, one brain-like

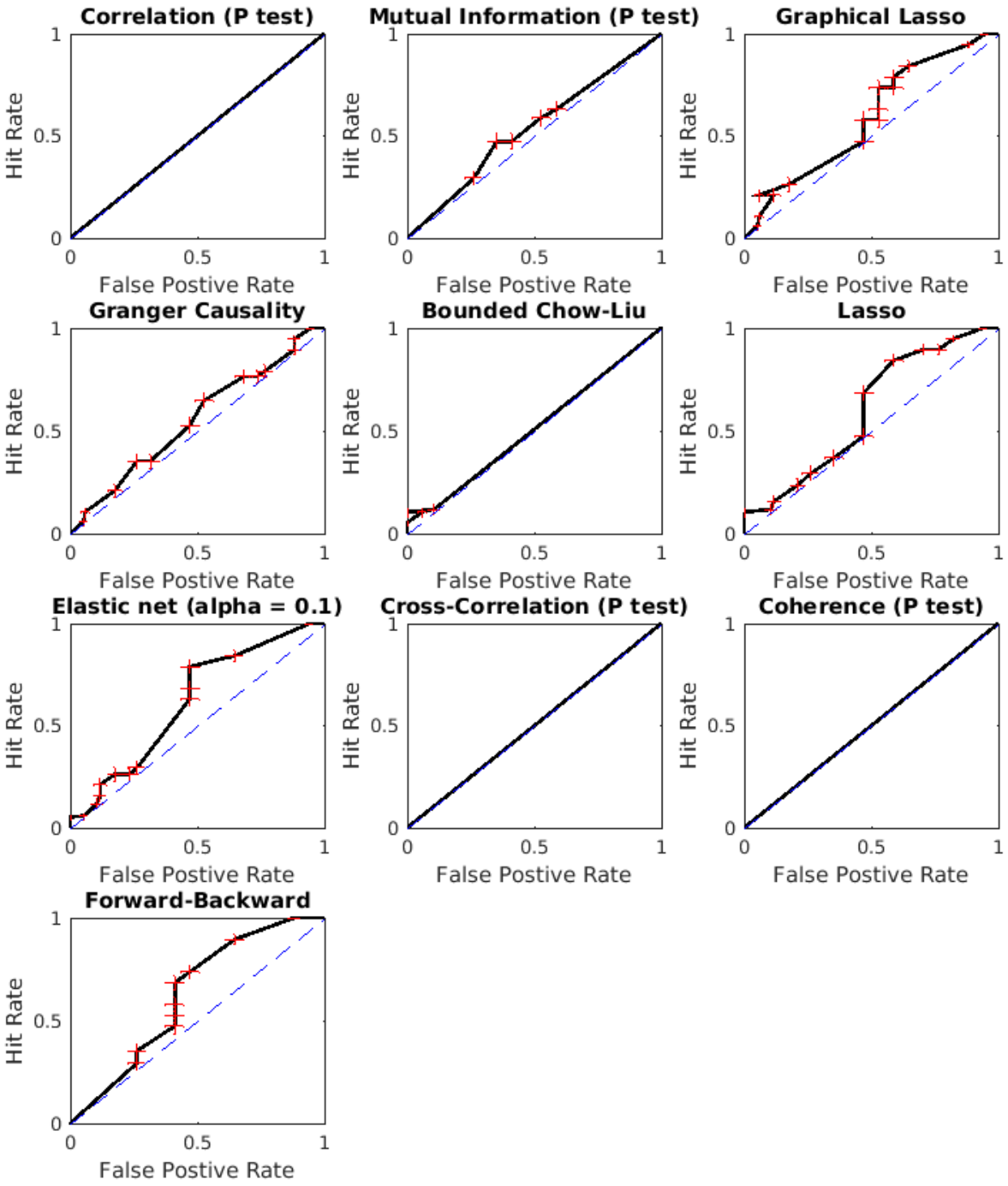


Figure 8: Pooled ROC curves for each method. At each parameter value, hit and false positive rates were computed for the adjacency matrix pooled across scans (at a pooling threshold of 0.5). Blue lines are linear interpolations of these averaged ROC points, and crosses are bootstrapped 95%-confidence intervals for each point (resampling scans whose results were pooled).

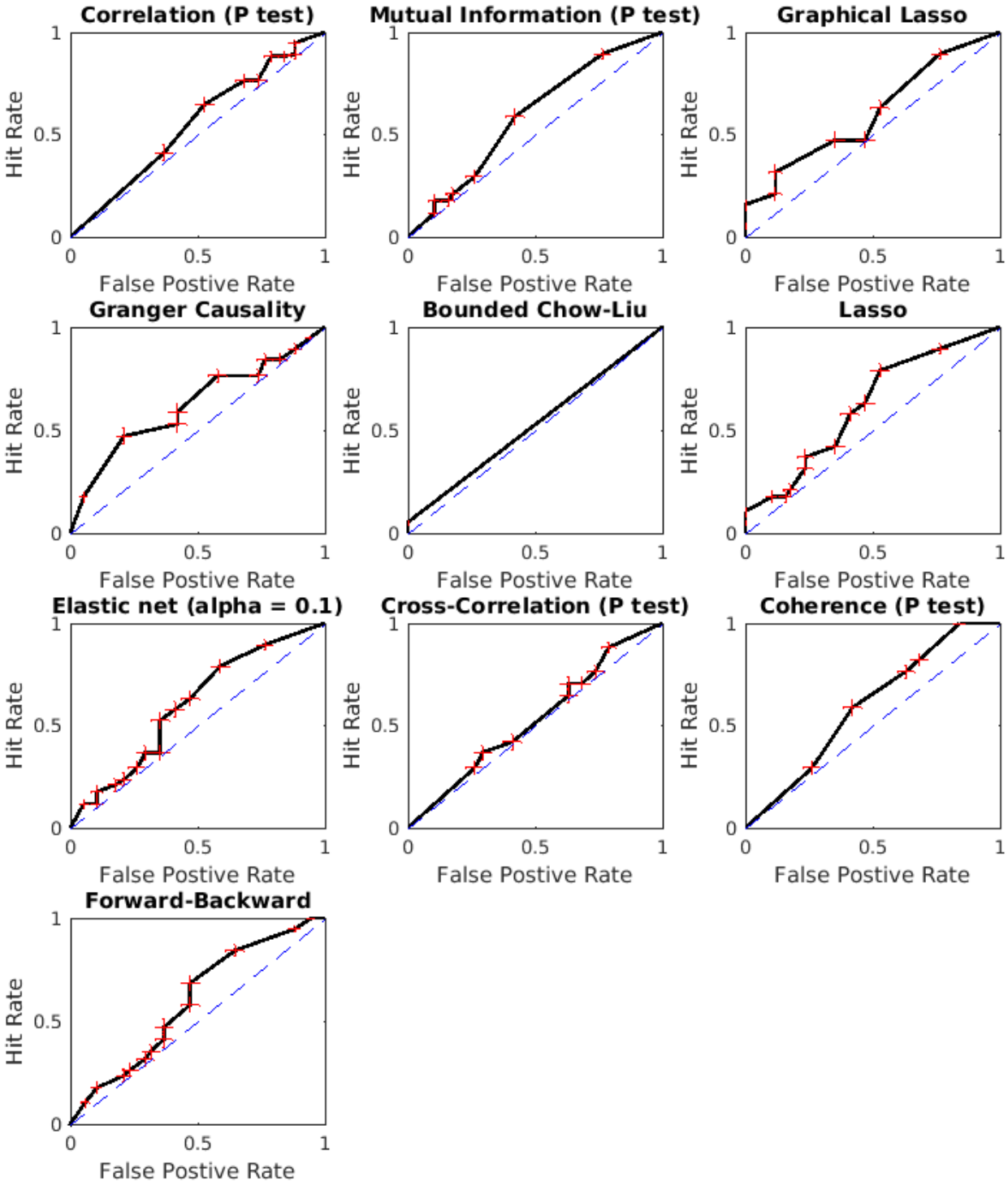


Figure 9: Same as Figure 8, but with a pooling threshold of 0.95.