Homology, Homotopy and Applications, vol. 16(1), 2014, pp.245–264

DESCRIBING HIGH-ORDER STATISTICAL DEPENDENCE USING "CONCURRENCE TOPOLOGY," WITH APPLICATION TO FUNCTIONAL MRI BRAIN DATA

STEVEN P. ELLIS AND ARNO KLEIN

(*communicated by Gunnar Carlsson*)

Abstract

In multivariate data analysis dependence beyond pair-wise can be important. With many variables, however, the number of simple summaries of even third-order dependence can be unmanageably large.

"Concurrence topology" is an apparently new method for describing high-order dependence among up to dozens of dichotomous (i.e., binary) variables (e.g., seventh-order dependence in 32 variables). This method generally produces summaries of dependence of manageable size. (But computing time can be lengthy.) For time series, this method can be applied in both the time and Fourier domains.

Write each observation as a vector of 0's and 1's. A "concurrence" is a group of variables all labeled "1" in the same observation. The collection of concurrences can be represented as a filtered simplicial complex. Holes in the filtration indicate relatively weak or negative association among the variables. The pattern of the holes in the filtration can be analyzed using persistent homology.

We applied concurrence topology on binarized, resting-state, functional MRI data acquired from patients diagnosed with attention-deficit hyperactivity disorder and from healthy controls. An exploratory analysis finds a number of differences between patients and controls in the topologies of their filtrations, demonstrating that concurrence topology can find in data high-order structure of real-world relevance.

This research is supported in part by United States PHS grants MH62185 and MH084029.
Received May 22, 2013, revised October 25, 2013; published on May 27, 2014.
2010 Mathematics Subject Classification: 62H17, 92C55, 62M15.
Key words and phrases: dichotomous data, high-order dependence, Fourier analysis of time series, computational homology, persistent homology, fMRI, ADHD.
Article available at `http://dx.doi.org/10.4310/HHA.2014.v16.n1.a14`

1. Introduction

We propose an apparently new nonparametric method, "concurrence topology," for describing the high-order dependence structure of multivariate binary ("dichotomous") data. It does this by translating the data into a filtered simplicial complex and then analyzing the topology of the filtration. In this paper, we analyze the topology using persistent homology [**12**]. We call this approach to concurrence topology "concurrence homology." (Computer scientists also use topology to study "concurrences" in distributed systems [**16**]. Apparently, this problem inspired "directed algebraic topology" [**17**]. *Prima facie* our notion of "concurrence" is quite different from the usage in these subjects.)

In this paper, we first explain how concurrence topology works and then demonstrate it in analysis of "functional connectivity" in resting-state functional magnetic resonance imaging data (fMRI; [**20**], [**33**], and Section 8). This data set consists of multivariate time series of "blood oxygen level dependent (BOLD)" values for each of 25 patients diagnosed with attention deficit hyperactivity disorder (ADHD), and 41 healthy controls (Section 8). (Concurrence topology applies to binary data, so we first dichotomized the fMRI BOLD time-series values; see Appendix A.)

Functional connectivity appears to be particularly appropriate in understanding ADHD. To quote [**23**], "... a change in perspective in etiological models of ADHD has occurred. These models shift the focus of the assumed pathology from regional brain abnormalities to dysfunctions in distributed network organization. ... As a result, the analysis of brain connectivity has become more and more critical." Others have used fMRI to reveal abnormalities in functional connectivity in ADHD [**28**]. We find other abnormalities using concurrence topology.

Using fMRI to understand ADHD is an active area of research in psychiatry. The fact that we find differences between ADHD and control groups using concurrence topology demonstrates that concurrence topology can find in data structure of real-world relevance.

In this paper, we use concurrence topology to describe functional connectivity in each subject, then apply standard inferential statistical methods to the subject-wise descriptions (Section 9).

1.1. Order of dependence

A binary variable X can be thought of as taking values in the set $\{0, 1\}$. With a nod to fMRI terminology, say that X is "active" when it is "1." Informally, variables $X_1, \ldots, X_p$ are "positively associated" if, when some of the variables are active, all p variables tend to be active.

Concurrence homology is sensitive to relatively weak or negative (i.e., nonpositive) association. Thus, variables $X_1, \ldots, X_p$ are relatively weakly or negatively associated if, compared to the number of times (frequency) at which *some* of them are active, the frequency at which they are *all* active is low. The frequency at which some of them are active can be checked by looking at fewer than p variables at a time. Similarly, the definition of the interaction term $\lambda_{11\ldots1}^{X_1,\ldots,X_p}$ in a log-linear model [**1**, p. 143] involves not just the product $X_1 \ldots X_p$ but also lower-dimensional marginals.

If a feature of the joint distribution can be detected by looking at p variables at a time, but not by looking only at $p-1$ variables at a time, then we say that

feature pertains to "pth-order dependence" among the variables. For example, Pearson, Kendall, and Spearman correlations are measures of second-order dependence because a correlation matrix for a collection of variables can be computed by looking at the variables two at a time. The odds ratio [**1**, p. 15] is also second-order.

In this paper, we focus on "high-order" dependence, by which we mean dependence of order at least 3. Table 1 displays "toy" data sets that illustrate the need to look at orders of dependence higher than 2: The three data sets are identical in orders 1 and 2, but differ in order 3. For each data set, the rows represent individual cases or observations. Each variable is active, and each pair of variables is simultaneously active, the same number of times in all three data sets. But the triplet of variables X, Y, and Z is simultaneously active 0, 1, and 2 times in data sets I, II, and III, respectively. A real data example analogous to this is described in Section 10.

I					II					III				
V	W	X	Y	Z	V	W	X	Y	Z	V	W	X	Y	Z
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	1	0	0	0	0	1	0	0	0	1	0
0	0	1	0	1	0	0	0	1	0	0	0	1	0	0
0	0	1	1	0	0	0	1	0	0	0	0	1	1	1
0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
0	0	0	1	1	0	0	1	0	1	0	0	0	1	0
0	0	1	0	1	0	0	1	1	0	0	0	1	0	0
0	0	1	1	0	0	0	1	1	1	0	0	1	1	1
1	0	0	0	1	1	0	0	0	1	1	0	0	0	1
0	1	1	0	0	0	1	1	0	0	0	1	1	0	0
1	1	0	0	0	1	1	0	0	0	1	1	0	0	0

Table 1: Three data sets identical up to second-order, but not at third-order.

Typically, formulating a traditional statistical model involves stringent choices of which variables should be included in a model and in what way. If one can use prior knowledge as a guide in making these choices, a traditional statistical model can be a powerful way to learn from data. However, our interest is in a more data-driven approach to data analysis. Say a method is "agnostic" if *a priori*, for $k = 1, 2, \ldots$ all groups of k variables are treated identically and few assumptions are put on the joint distribution. (So the method is "nonparametric." The term "agnostic" has a different meaning in learning theory [**22**].) This rules out much *a priori* structural assumption. An example of an agnostic method is principal component analysis [**21**, Chapter 8], a second-order method.

There are apparently few nonparametric agnostic methods that can cope with the "combinatorial explosion" (Section 2) that is inherent in describing high-order dependence among more than a few variables. Other such methods include independent component analysis (ICA) [**19**] and latent variable methods [**3**]. (ICA is a popular method for analyzing fMRI data. For more discussion of analysis of fMRI data, see [**2, 25, 33, 30**].)

The aforementioned methods and ours capture very different aspects of high-order

dependence. Hence it is difficult to compare them to concurrence topology. For that reason, and in the interest of brevity, in this paper we do not compare concurrence topology to other methods.

1.2. Supplementary material

An updated version of the software we used for the computations described in this paper (Section 6), plus fairly extensive documentation, is posted on StatLib (`http://lib.stat.cmu.edu/`; search for "ConcurrenceTopology"). Two documents in that posting that may interest the reader are [**13**] ("SomeBackground") and [**14**] ("ConcurrenceTopol_Notes"). The data we used for this paper, with documentation, plus further results, as well as the software, can be found at `http://binarybottle.com/concurrencetopology/`; [**13**] and [**14**] are also posted there.

1.3. Preview

In Section 2, we explain how agnostic analysis of high-order dependence in many variables leads to a "combinatorial explosion." Concurrence topology makes some headway in overcoming this obstacle. In Sections 3 and 4, we explain how, in concurrence topology, binary data is translated into a filtered simplicial complex. In Section 5 the connection between persistent homology and statistical dependence is made. Section 6 briefly describes the algorithm and software we used in our work. With time-series data, like our fMRI data, concurrence topology can be used to analyze the data in either the "time" or "Fourier" domains. This is explained in Section 7. Section 8 discusses fMRI in general and our fMRI data set in particular.

Section 9 describes the formal statistical methods we applied to the descriptions generated by concurrence topology. Section 10 presents some of the findings we obtained. Section 11 discusses "localization," by which we mean identifying specific "short cycles" representing homology classes. We then go on to describe some of our findings based on localization. We sum up briefly in Section 12.

Two appendices provide further details of the methods and findings.

2. "Combinatorial explosion"

An agnostic analysis of order p means examining all combinations of p variables at a time. If $p > 2$ and there are many variables, the number of combinations can be unmanageably large.

As an example, in Section 10 we look at seventh-order dependence among the regions of the "default mode network (DMN)" [**32**] in each subject in our fMRI data set. In our interpretation the DMN consists of 40 regions (supplemental material; see Section 1.2). For each subject, we discarded eight regions (Appendix A). Thus, we examine seventh-order dependence in a 32-way table.

An agnostic seventh-order log-linear analysis would result in $\binom{32}{7} = 3,365,856$ distinct seventh-order interactions for each subject (compared to the 6144 fMRI BOLD values—192 time points in 32 regions—in each subject's data). The rapid growth in $\binom{V}{p}$ as p increases deserves to be called a "combinatorial explosion." (See [**1**, p. 150]; [**30**, Section 8.3.1] makes essentially the same point.)

By contrast, we found that the data summaries produced by concurrence homology included at most hundreds of numbers per subject, even when "localization"

(Section 11) was employed. Moreover, those numbers are structured in a way that aids interpretation. Thus, concurrence homology provides parsimonious descriptions of high-order dependence (the cost is in computation time).

3. Concurrences

We now describe the general process by which binary data is translated into a filtered simplicial complex. Concurrence topology is based on "concurrences." A binary variable can be coded "0" and "1." In concurrence topology, the data consist of binary variables observed on multiple "units." The data from a single unit is an "observation." Thus, an observation is a V-tuple of 0's and 1's, where V is the number of variables. Focus for now on the case in which the unit is a single time point in an fMRI run for a single person. (This is the "time domain" analysis. Another choice of unit is made in Section 7.) The variables correspond to brain regions. A "1" means the region is active; "0" indicates lack of activity. (To see how "active" is defined, see Appendix A.) But, in principle, concurrence topology can be applied to any collection of binary vectors.

A "concurrence" is a group of variables that are all "1" in the same observation. In effect, we throw away the 0's and just retain the 1's. (So if an observation consists entirely of 0's, it is dropped.) More precisely, we retain the names of the variables coded "1" in the observation. Call the number of variables in the concurrence its "length." The concurrences from a data set constitute a "concurrence list." In a concurrence list, the order of the concurrences is irrelevant, but multiplicity of concurrences is important.

For illustration, consider the data sets in Table 1. The concurrence list in data set I is YZ, XZ, XY, YZ, XZ, XY, VZ, WX, and VW. (The first and fifth rows are dropped. We ignore the order, but not the frequency of appearance, of concurrences.)

4. The filtered Curto–Itskov complex

What we call the "filtered Curto–Itskov (simplicial) complex" is constructed in two steps. (We explain the appellation "Curto–Itskov" presently.) Let $\mathcal{C}$ be a concurrence list.

1) *Filter the concurrence list:* For each concurrence, C, in the list, count how many times it appears as a subset (proper or not) of concurrences in the list. That count is the "frequency" of C. Thus, even if C only appears as such once in the list, it can have a frequency greater than 1. If $f = 1, 2, \ldots$, let $\mathcal{C}_f$ be the concurrence list consisting of concurrences that have frequency $\geqslant f$. We call f the (absolute) "frequency level" of $\mathcal{C}_f$. Thus, if $f_1 < f_2$, then any concurrence in $\mathcal{C}_{f_2}$ is a subset of a concurrence in $\mathcal{C}_{f_1}$. Call the collection $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \ldots$ the "filtered concurrence list" of the data.

2) *Construct the filtered Curto–Itskov complex from the filtered concurrence list:* The "frame," K_f, in the filtered Curto–Itskov complex at frequency level f is obtained by interpreting each concurrence $C \in \mathcal{C}_f$ as an abstract simplex, σ_C. K_f is the abstract simplicial closure of $\{\sigma_C : C \in \mathcal{C}_f\}$. Thus, $K_1 \supset K_2 \supset K_3 \cdots$. This descending filtered simplicial complex is the "filtered Curto–Itskov complex," $\mathcal{K}$, of the data. (In the fMRI

data, each time series has the same length, 192. This allows us to use absolute, i.e., integer frequencies, f. In general, relative (i.e., fractional) frequencies must be used. A population version of the filtered Curto–Itskov complex might be indexed by a continuum of frequencies; see [**13**].)

The filtered complex $\mathcal{K}$, together with the number of observations, is equivalent to the contingency table [**1**, Section 2.1.1] for the binary data [**13**].

We call investigation of the joint distribution of multivariate dichotomous data by analyzing the topology of the corresponding filtered Curto–Itskov complex "concurrence topology."

The use of descending rather than ascending filtrations is natural in this context because frequency level is a statistically meaningful index. But the fact that $\mathcal{K}$ is descending does not interfere with computing its persistent homology: A persistent classes is "born" at a higher frequency level than that at which it "dies."

In Figure 1, each row is the filtered Curto–Itskov complex for a data set in Table 1. We see that the filtered Curto–Itskov complexes—in particular their 1-dimensional persistent homology—do distinguish the three data sets.

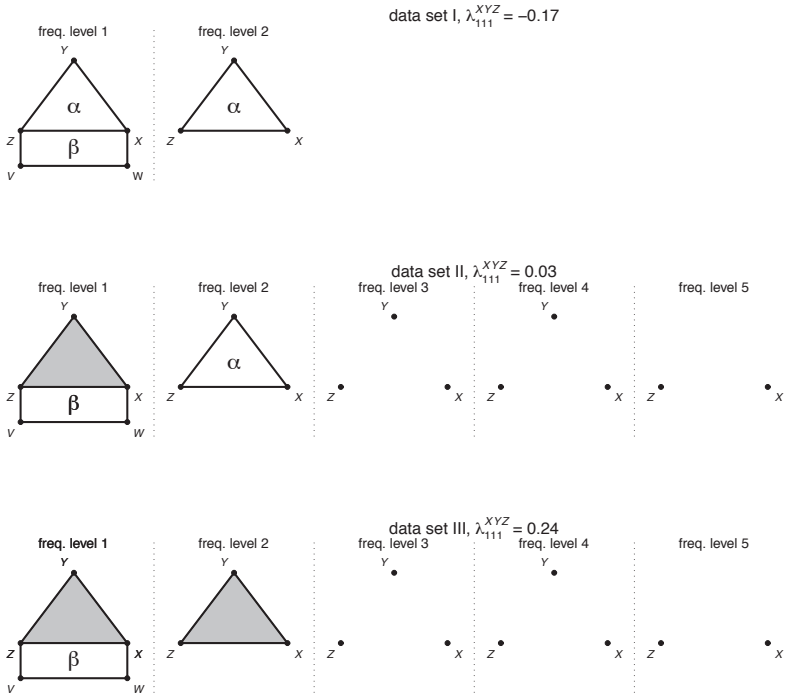

Figure 1: Rows are filtered Curto–Itskov complexes for data sets in Table 1. Columns, separated by dotted vertical lines, correspond to frequency levels. "λ_{111}^{XYZ}" is a third-order interaction in a log-linear model; "α" and "β" label persistent classes in dimension 1.

Figure 1 also displays the values of the third-order interaction term λ_{111}^{XYZ} in a log

linear model [**1**, Chapter 5] for each data set. This term pertains to the frequency of the event $\{X = 1, Y = 1, Z = 1\}$. Notice that there is a perfect negative association between the lifespans of the homology class α represented by $XY + XZ + YZ$ and the values of λ_{111}^{XYZ}. Thus, it seems that the lifespan of the homology class α does indicate how negative or weak is the third-order dependence among X, Y, and Z. We do not claim that there will always be such a neat pattern, but it does provide "experimental evidence" in favor of our contention that holes in a filtered Curto–Itskov complex indicate weak or negative association (Section 5).

Our work on concurrence topology is inspired by Curto and Itskov [**10**], who investigated a question in theoretical neuroscience by applying topological methods to simulated data. From each simulation, Curto and Itskov constructed the complex K_1 and studied its homology. For their purpose, it was not necessary to build a filtration. In essence, they only needed to know whether each cell in a contingency table was 0 or not. But typically for data analysis, one needs to know the actual values in the table. To represent those values geometrically a single Curto–Itskov complex is not sufficient.

Plotting *death* vs. *birth* yields a "persistence plot" for each dimension d. (Since we index the filtered Curto–Itskov complex by frequency level, our "persistence plot" is different from, but trivially equivalent to, the standard "persistence diagram" [**12**, p. 152].)

Figure 2 shows the persistence plot in dimension 1 (third- and higher-order dependence by equation (1), below) for the regions in the DMN for control subject "sub01912." Thus, e.g., the dot marked by an asterisk indicates a persistent 1-dimensional homology class, call it α, that is born in frequency level 13 and dies in frequency level 3. One expects that classes like this one, with a long lifespan, are less likely to appear by chance and are more likely to reflect negative, rather than merely weak, association among the variables (Section 9.1).

It turns out that, indeed, classes similar to α appear in most subjects' data in our fMRI data set. Investigating this led us to find one of several ways of using concurrence homology to discriminate ADHD subjects from controls (Section 11.1.1).

5. Homology and statistical dependence

In this paper, we analyze the persistent homology of the filtered Curto–Itskov complex belonging to each subject in the fMRI data set. (We used $\mathbb{Z}/2 = \{0, 1\}$ coefficients. In one analysis (Section 10) we made use of the Euler characteristic.) Recently, there has been much interest in using persistent homology for data analysis (e.g., [**15, 8**]). In particular, persistent homology has been applied to brain data [**24, 9**]. However, concurrence topology appears to be a new method.

One cannot detect a d-dimensional persistent homology class, η, in a Curto–Itskov complex by looking only at $d + 1$ variables at a time, but one can detect a d-dimensional homology class by looking at groups of $d + 2$ variables at a time. Detection of the class may require looking at multiple groups of $d + 2$ variables at the same time. Thus, η reflects dependence of order $d + 2$ or higher.

The persistent homology class η is represented by the sum of at least $d + 2$ d-simplices, each corresponding to $d + 1$ variables active at the same time. However, for η to exist also requires one or more groups of $d + 2$ of the same variables to *not*

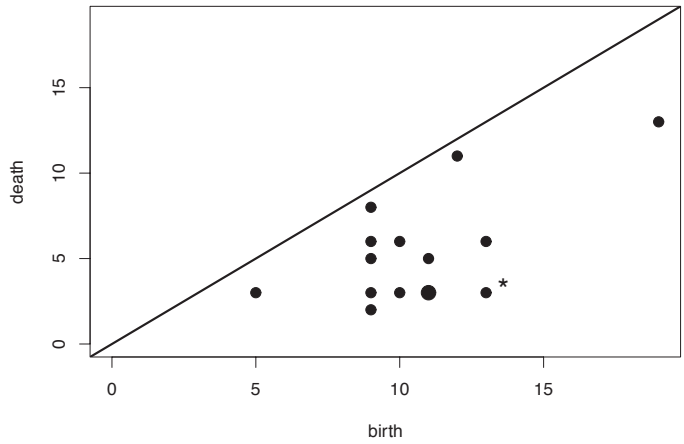

Figure 2: Dimension 1 persistence plot for the fMRI BOLD values in the time domain in the DMN for subject "sub01912." The larger dot indicates two coinciding points. The dot near the asterisk represents an interesting persistent class discussed in Section 11.1.1.

be active at the same time. Thus, existence of η reflects a shortage of groups of $d+2$ active variables compared to groups of $d+1$ active variables. Thus, at least in the data set at hand, the $d+2$ variables are relatively weakly or negatively associated. To sum up:

A d-dimensional persistent homology class of a filtered
Curto–Itskov complex indicates relatively weak or negative
statistical dependence of order $d+2$
or higher among the variables. (1)

(For more discussion of this issue, see Section 9.1.)

Remark 5.1. A d-dimensional persistent homology class of a filtered Curto–Itskov complex that is "born" at a high frequency level actually indicates strong *absolute* positive association because a d-dimensional hole is bounded by at least $d+2$ d-simplices, each corresponding to a concurrence of length $d+1$. Strong, but not perfect, positive association is needed to generate so many long concurrences, particularly if d is large.

Remark 5.2. In "classical" statistical methods, the sample size (e.g., number of time points in an fMRI run) has to exceed the number of variables (number of regions). This is the case in the data we describe in this paper. However, the filtered Curto–Itskov complex and its persistent homology are well defined and computable even if the number of variables equals or exceeds the sample size, an important case [**18**].

6. Algorithm and software

We wrote our own concurrence homology software in R [**31**]. Other software for computing homology include Dionysus (`http://mrzv.org/software/dionysus`), Perseus (`www.math.rutgers.edu/~vidit/perseus.html`), and CHomP (`http://chomp.rutgers.edu/`).

In our algorithm, we compute persistent homology relative to a subfiltration of acyclic subspaces [**27**]. Some of the theory underlying the software is discussed in [**14**].

The distribution of the time needed to compute the homology for each subject had a very long right-hand tail. Usually, a few hours sufficed to compute the persistent homology for an individual subject, but sometimes even a week did not.

7. Concurrence topology in the Fourier domain

High-order spectral analysis of multivariate time series is a well-studied subject [**5**]. There is a concurrence topology version of this. For each subject, the fMRI BOLD data consist of a multivariate time series with one component per region. Concurrence topology of fMRI in the "time domain" is carried out by constructing the filtered Curto–Itskov complex from direct dichotomization of BOLD values (Appendix A) and treating every time point as a separate observation. In the "Fourier domain," instead of dichotomizing the BOLD signal itself, one dichotomizes the periodograms (proportional to the squared moduli of the finite Fourier transform [**7**, p. 120]) of the component series.

Define concurrence in the Fourier domain just as in the time domain, but treat vectors of periodogram values at different Fourier angular frequencies [**4**, p. 42] as separate observations. This allows the study of high-order dependence while taking into account the time series nature of fMRI BOLD data.

Working in the Fourier domain gives information not provided by the time domain analysis. In the time domain analysis, concurrences are based on simultaneous behavior of the regional BOLD time series. The timing of brain activity is ignored. But functional connectivity among brain regions might not manifest itself in simultaneous activity. For example, suppose the BOLD time series in several regions are quite similar but shifted relative to each other in time. The time domain analysis may not detect the strong dependence among these regions. However, the periodograms for these series will be very similar, and one would expect them all to be included in concurrences at a number of Fourier frequencies.

8. fMRI data

An active brain region attracts oxygenated blood. This gives rise to a "blood-oxygen-level dependent (BOLD)" signal that can be detected by a magnetic resonance (MR) machine. This use of MR is called "functional Magnetic Resonance Imaging (fMRI)." A typical fMRI image of the brain is taken about once every 2 seconds with a spatial resolution of about $3 \times 3 \times 5\ mm^3$. So activity of different parts of the brain can be recorded over time. "This exciting technology has revolutionized the scientific study of the mind" [**30**, p. 1].

In some fMRI experiments, subjects in an MR machine are asked to perform specified tasks. But task-free or resting-state fMRI, in which subjects are asked to think about nothing in particular, has become a popular form of fMRI experiment to study baseline or spontaneous activity [**33**]. Ours are resting-state fMRI data.

The fMRI data set we use in this paper was produced at New York University and distributed as part of the "1000 Functional Connectomes" project (`http://fcon_1000.projects.nitrc.org/`). At the time we began our work, this was the largest publicly available resting-state fMRI data set containing clinical data. This data set includes 41 healthy controls ("NewYork_a_part1") and 25 adults diagnosed with ADHD ("NewYork_a_ADHD").

The samples were highly imbalanced with respect to age and gender. Only 20% of the ADHD group was female, while about half of the controls were. About 25% of the controls were children (younger than 20; median age = 12), while there were no children in the ADHD group. Among adults, ages ranged from about 21 to about 50 in each group. The median age in the ADHD group was 37, while in the control group the median adult age was 27.

We computed BOLD values for 92 regions, including 40 in the DMN (supplemental material) at 192 time points. Prior to applying concurrence topology we dropped some regions in a subject-wise fashion (Appendix A).

9. Analysis of fMRI data

9.1. Sampling within and between subjects

We again take up the issue discussed in Section 5, but we now allow randomness. Let T = sample size (= number of fMRI time points). Let σ be a simplex in the filtered Curto–Itskov complex of a subject's fMRI data. If $f = 1, 2, \ldots$ is the highest frequency level such that $\sigma \in K_f$, then $\hat{P}(\sigma) := f/T$ is an estimate of the probability, $P(\sigma)$, of occurrence of the concurrence corresponding to σ. If T is large relative to $V :=$ number of variables (= number of regions) and the observations are independent, then one expects the estimate, $\hat{P}$, of the joint distribution, P, to be good.

For a given subject, one does not expect the fMRI BOLD signal in different time points to be independent. However, it is reasonable to suppose that the multivariate BOLD time series is "mixing." (A time series is "mixing" if the statistical dependence between the observations taken before time t and after time $t + \Delta t$ drops off as $\Delta t > 0$ increases [**7**, Section 1.3]). Appropriate forms of mixing will ensure that $\hat{P}$ estimates P well for T large compared to V (e.g., [**6**, Definitions 1.2 and 1.6 and Theorem 1.14]).

These considerations are relevant to interpretation of the persistent homology of the filtered Curto–Itskov complex. To see this, consider a simple example in dimension 1. Hypothetically, suppose one analyzes BOLD values in only three regions, A, B, and C. If f_{ABC} is the largest frequency level in the corresponding filtered Curto–Itskov complex at which the 2-simplex ABC is present, then f_{ABC}/T is an estimate of the probability of observing the *concurrence* ABC. Similarly for the 1-simplices AB, AC, and BC. Suppose $m_{ABC} := \min\{f_{AB}, f_{AC}, f_{BC}\} > f_{ABC}$, then $z := AB + AC + BC$ represents a persistent homology class, α, whose lifespan is $m_{ABC} - f_{ABC}$.

The extent to which α is a reproducible feature of brain activity depends on the probability of observing $m_{ABC} > f_{ABC}$ in other resting-state fMRI runs. Now,

f_{ABC}/T, f_{AB}/T, etc., are estimates of probabilities. As such, they have sampling variation to which the first two paragraphs of this section apply. If the lifespan of α is sufficiently long, then in another sample, despite this sampling variation, it is probable that z would represent homology in the filtered Curto–Itskov complex for the new sample. (If there are more than three regions the picture is much more complex.)

While the ideas presented above in this section help one to interpret persistence plots, *they are largely irrelevant to the formal statistical inferences in this paper.* That is because our statistical analyses are based on sampling *subjects*. The filtered Curto–Itskov complex of a subject is treated as a descriptive statistic whose sampling properties within that subject therefore do not matter. We model the subjects in each group (ADHD or control) as independent draws from the relevant population. Our statistical analyses are based on that model. Thus, we performed inference *between* subjects, not *within* subjects.

9.2. Data analytic methods

For each subject, we computed summaries of the homology of the filtered Curto–Itskov complex based on his/her dichotomized fMRI BOLD data or periodograms thereof and compared the distribution of those summaries between groups (i.e., ADHD and control) or, in one instance, between genders.

Our purpose in this study is to develop methods for using concurrence topology. If a method revealed something of interest in the fMRI data (usually group differences), then we took that as an indication that the method might be a promising one for use elsewhere.

Thus, the analyses we undertook were exploratory. Operationally, to "reveal something of interest in the fMRI data" meant finding an effect that was nominally statistically significant at the $\alpha = 0.05$ level in an appropriate test. (We used Wilcoxon rank sum and chi-squared tests and generalized least squares (GLS) [**29**].) "Statistical significance" was used merely as a screening method and flag that indicated analytical methods that might be worthwhile for future use.

We *do not* claim that these findings are firm conclusions about ADHD, only that they are worth testing in independent samples. Unless stated otherwise, *all findings we mention concerning the fMRI data set are statistically significant in this operational, uncorrected sense.* Because our analyses are only exploratory, to save space we omit some details of the analyses performed.

For each subject, we computed persistent homology (in both the time and Fourier domains) in dimensions 0 through 5 (corresponding to dependence orders up to 7 or more, by (1)) in the DMN. We also computed persistent homology in both domains in dimensions 0 through 2 in the whole brain. In some cases, we also computed the corresponding localization (Section 11) and/or Euler characteristics for each subject.

Since the fMRI data set is quite imbalanced with respect to age and gender (Section 8), we sometimes analyzed only the data in adults and/or otherwise controlled for age and/or gender.

In some analyses, for each dimension d we summarized the main features of a persistence plot by nine "moments": The first moment was the number of persistent classes of dimension d. The other moments are, for $i, j = 0, 1, 2$ (not both 0), defined

to be the averages

$$moment_{ij} := M^{-1} \sum_{(birth,death,m)\in plot_d} m \times \left[birth^i \, (birth - death)^j\right]^{1/(i+j)},$$

where $plot_d$ is the collection of triples $(birth, death, m)$ in which $(birth, death)$ is a point, with mulitplicity m, in the persistence plot for dimension d. M is the sum of all the multiplicities. Thus, "$birth - death$" is the lifespan of the class(es) plotted at $(birth, death)$. Note that, in line with the reasoning in Section 9.1, the longer the lifespan a class has, the more weight it receives in $moment_{ij}$ $(j = 1, 2)$.

For the DMN, we computed persistent homology in dimensions $d = 0, \ldots, 5$. Hence for the DMN we obtained for each subject $6 \times 9 = 54$ moments in each domain (time and Fourier). For the whole brain, we computed persistent homology in dimensions 0 through 2, so each subject has a $3 \times 9 = 27$ moments in the whole brain in each domain. We analyzed these multivariate summaries using GLS with $moment^{1/3}$ as the response variable. (We took the cube root of *moment* in order to reduce the skewness of its distribution.)

10. Some findings

In the whole brain and Fourier domain, GLS analyses just described show a difference between the groups in the persistent homology in dimensions 1 and 2, particularly the former. Using the GLS analysis, we also picked up group differences in the DMN in the time domain in dimensions 4 and 5.

The group difference in the DMN and time domain in dimension 4 (representing 6th- and higher-order dependence by (1)) was a robust finding, in the sense that it manifested itself in a number of analyses. The essence of the difference is simply that a smaller proportion of ADHD subjects (64.0%) had any homology in the time domain in the DMN in dimension 4 than did controls (92.6%).

This finding is reminiscent of Table 1: In the DMN, in the time domain we find no statistically significant differences between the two groups in dimensions 0 through 3 (orders of dependence 2 through 5 and up). Only in dimension 4 and, perhaps, 5 do we see a difference. This is another example of why it is important to examine high-order dependence. (Of course, some other method of analysis might find differences in orders of dependence 2 through 5, so the parallel with Table 1 is not perfect.)

In the DMN, in the Fourier domain the Euler characteristic of the frame, K_1, in frequency level 1 is typically higher among the ADHD subjects (mean = 1.68, standard deviation (SD) = 2.53) than it is among the controls (mean = 0.415, SD = 1.12), another robust finding.

As an informal analysis, we observed in some experiments that the homology one gets from simulated data in which all the regions function independently of each other is far different from what one finds in the real fMRI data. Obviously, brain regions do not function independently of each other. It is reassuring that concurrence topology recognizes this in the data.

We describe further findings concerning the fMRI data set in Section 11.1.

11. Localization

"Localization" offers a higher-resolution description of the topology of the filtered Curto–Itskov complex. Having found a persistent homology class, it is natural to ask what variables (regions, in our case) are involved. Existence of a persistent homology class of the filtered complex requires the cooperation of all variables, but some variables are more directly involved than others.

In our persistent homology algorithm, we use relative homology (relative to acyclic subcomplexes; see Section 6), but in this section we only refer to *absolute* cycles. The fewest number of terms that a d-cycle can have is $d+2$. We say that a cycle is "short" if it includes exactly $d+2$ terms.

Short representatives of a homology class are the ones most directly involved in the hole corresponding to the class. Thus, to interpret a homology class, it makes sense to focus on its short representatives (if there are any). (However, since the homology classes depend on all variables, short cycles are only defined in the context of the entire set of variables.) We computed *all* short cycles of *all* homology classes (having short representatives), not just of classes in a basis. (Note that [**11**] discusses a different notion of localization.) Localization was carried out separately for each frame; i.e., persistence of homology classes was ignored in the localization.

11.1. Localization in the fMRI data

11.1.1. Dimension 1 in the DMN and time domain

In the DMN and time domain we found 7427 distinct short 1-cycles across all subjects. (There are 40 regions in the DMN. $\binom{40}{3} = 9880$ distinct short cycles are theoretically possible for a single subject; median number of distinct short 1-cycles per subject = 260.) One subject had a homology class in dimension 1 containing 164 short 1-cycles in a single frequency level (frame).

We selected the most important short cycles using two criteria. One is that the number of subjects having the cycle be large, and the other is that, in line with the reasoning in Section 9.1, the lifespan of the cycle be long. A cycle may represent homology across a range of frequency levels. The "lifespan" of the cycle is the number of frequency levels in which it does so. The lifespan of a cycle can never be longer than that of the persistent homology class to which it belongs.

A short cycle representing the persistent homology class plotted at the point marked by "*" in Figure 2 appears in 13 subjects and, for subject "sub01912," has cycle lifespan = 8 (supplementary material). Call this cycle "z." In subject "sub01912," this triplet of regions is well connected at second-order but, comparatively speaking, not even indirectly well connected at order 3.

To see if the appearance of a cycle in 13 subjects is remarkable, we performed an analysis under the null hypothesis that all possible 9880 triplets of DMN regions are equally likely to be short cycles in a given subject. We assumed that short cycles were selected from the 9880 independently *between* subjects, but not necessarily independently *within* subjects. Then, based on a simple model, an upper bound on the probability that *some* triplet will be a short cycle for 13 or more subjects is only 0.021 (Appendix B). Thus, z appears to be rather special.

Now, presence of z itself does not differentiate the ADHD and control groups, but the 29 short cycles that are homologous to z in subject "sub01912" *do* distinguish

the groups. (We include short cycles that are homologous to z in any frame where z exists and does not bound. This is not quite the same as taking all short cycles in the *persistent* class, call it α, to which z belongs, because z is not present in α in all frames where α is alive.)

We can refine this. Each of 16 of the 29 short cycles appears at least twice in each diagnostic group (supplementary material). Nineteen out of 25 ADHD subjects (76%) have at least one of the 16 short cycles, but only 18 out of 41 controls (44%) have any. This difference was another of our robust findings.

The 16-cycles contain cerebral regions implicated in differences found in the literature between resting-state fMRI of ADHD patients and healthy controls, such as the precuneus, the anterior and posterior cingulate, and the inferior and medial frontal lobe (pars orbitalis and medial orbital frontal regions [**23**]).

The frequencies of occurrence of each of the 13 regions involved in any of the 16 short cycles are very similar in the two groups. Neither do the groups differ in frequency of occurrence of any *particular* short cycle among the 16. It appears that there is a particular persistent class or family of related classes that occur in many of the subjects' filtered Curto–Itskov complexes. We are detecting a subtle feature in the data that might be common in people in the general healthy and ADHD populations, but more commonly in the latter. This conjecture, like all our findings, needs to be checked in an independent sample.

11.1.2. Dimension 4 in the DMN in the time domain

In dimension $d = 4$, a short cycle involves six regions. Out of $\binom{40}{6} = 91{,}390$ theoretically possible 4-dimensional short cycles in the DMN time domain, 1497 appear in the data. The median number of distinct short 4-cycles per subject is 12.5.

Call a class "narrow" if it has at least one short representative cycle. Say that two narrow classes are "adjacent" if their sum is also narrow. The presence of adjacent pairs of classes in dimension 4 does not discriminate the diagnostic groups, but it does discriminate genders: only 1 out of the 25 females have any adjacent class pairs, but 13 out of the 41 males do.

11.1.3. Dimension 2 in the whole brain in the Fourier domain

There are 92 regions in the "whole brain." Out of $\binom{92}{4} = 2,794,155$ distinct theoretically possible 2-dimensional short cycles in the whole brain and Fourier domain, 7933 appear in the data. The median number of distinct short 2-cycles per subject = 57.5.

The "corpus callosum" consists of white matter, and until recently only gray matter was believed to produce a BOLD signal [**26**]. However, *each* of the five corpus callosum regions in our data set appears in at least 909 short 2-cycles, which is more often than any noncorpus callosum region appears and much larger than the median number of times (249) that noncorpus callosum regions appear. Of the 2,794,155 possible quadruplets of whole brain regions, 20% include a region from the corpus callosum, but of the distinct short 2-cycles in the data, 65% include a corpus callosum region. Thus, the corpus callosum frequently takes part in quadruplets that are weakly connected at fourth-order. (We performed no formal tests here.)

12. Discussion and Conclusions

Concurrence topology is a general nonparametric strategy for describing high-order dependence in dichotomous data. In this paper, we focus on "concurrence homology," a particular approach to concurrence topology. Using a resting state fMRI data set as a test bed, we explored a number of different ways of deploying concurrence homology. These included persistence, Euler characteristics, and several different ways of mining localizations, and we found numerous interesting apparent structures in the data. These findings are only exploratory, but we intend to try to replicate our findings in an appropriate independent data set. Still, the fact that we find differences between the groups demonstrates that concurrence topology can find structure in data of real-world relevance.

Concurrence homology is computationally intensive, but, with that proviso, concurrence homology can be applied, not just to fMRI BOLD data, but to any multivariate binary data. Moreover, we are confident that improved software will greatly expand the range of data that can be analyzed using concurrence homology.

An important upshot of our work is evidence that, apparently, it *is* worthwhile to study high-order dependence in data.

Acknowledgement
Comments by an anonymous referee led to numerous improvements in this paper.

A. Dichotomization

Concurrence topology is designed for binary data. The fMRI BOLD signal is continuous. For each region in each subject, we determined at which time points the region is "active" and at which it is "inactive" by dichotomizing fMRI BOLD values. There is no single level of fMRI BOLD that demarcates activity from inactivity, because fMRI BOLD levels in different regions are incomparable. So a separate threshold is needed for each region (in each subject).

A potential complication is that in some cases dichotomizing can merely amplify noise. Brain functional connectivity means covariation. Without variation, there is no covariation. The little variation shown by a nearly constant activity level is liable to be noise. Dichotomizing such a slightly varying noise series will amplify it and introduce a noisy binary component in the multivariate series.

Therefore, in the fMRI data, for *each subject separately* we discarded the 20% least variable regions. So different subjects may have different regions dropped. (One subject had fMRI BOLD values of 0 for all time points in two regions, likely due to either missing or inaccurate automated labeling of the regions. For that subject, those two regions were also dropped.) This was done separately for the whole brain and DMN. We measured variability by a robust version of the coefficient of variation: interquartile range divided by median.

Whether or not our reasoning in favor of dropping the least variable regions is sound, it is expedient: if all regions are included, the computation of homology takes much longer than it does when low variability regions are dropped.

We stress that the analysis does not start *after* the 20% least variable regions are dropped. Dropping the least variable regions is the *first step* in the analysis.

So this step does not compromise the agnostic nature (Section 1) of our method. The distribution of regions that were dropped did not differ between the ADHD and control groups, but did depend on age and sex.

In the time domain, separately for each subject and region retained for that subject, we deemed as "active" the 20% (39) time points at which the fMRI BOLD value was highest. In the Fourier domain, for each subject and region we set the threshold at the 90th percentile of power, because the fMRI BOLD time series had low power in about the highest half of the Fourier frequencies and 20% of half the Fourier frequencies is the same as 10% of all of them.

The thresholds used in dichotomization are tuning constants of the method. Our choices of thresholds are based on informal experiments on a smaller data set independent of our fMRI data set. More experimentation with tuning constants is needed, but it is difficult because of the lengthy computing times.

B. Upper bound on probability that a short 1-cycle appears in 13 or more subjects in the DMN

Here we back up the claim made in Section 11.1.1 that, under a null hypothesis we make explicit, the probability that in the DMN and time domain *some* triplet will be a short 1-cycle for 13 or more subjects is only 0.021 or less.

The number of short time domain DMN 1-cycles among the 66 subjects has a mean of 260 with an SD of 81. The distribution is fairly normal looking (Figure 3) but slightly skewed. Table 2 shows group-wise summaries.

Group	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ADHD	108.0	209.0	273.0	263.4	310.0	499.0
Controls	102.0	194.0	258.0	258.2	299.0	426.0

Table 2: Counts of short 1-cycles in DMN in time domain by group.

Only about 3% of the subjects have more than $m = 400$ short cycles. The majority have fewer than 300. For simplicity, assume a conservative model in which *each* subject gets 400 short cycles independently across subjects.

Let S be the set of all $N := \binom{40}{3} = 9880$ theoretically possible short cycles. Let $p(z)$ be the probability that a specific, but arbitrary, short cycle, z, appears in a specific, but arbitrary, subject's list of m short cycles. Assume $p(z) = p$ is constant in z. That is, no short cycle is special. (This is the null hypothesis.)

If $Z \subset S$, let $|Z|$ denote the cardinality of Z. By assumption, the subject will get assigned to him/her some set $Z \subset S$ of m short cycles. The probability that the subject gets Z varies with Z, but the subject will get *some* Z. (We make no assumption about the probability that the subject gets a given Z beyond the null hypothesis $p(z) = p$.) Thus, for a specific, arbitrary subject,

$$\sum_{Z \subset S;\, |Z| = m} \text{Prob}\{\text{subject gets } Z\} = 1. \tag{2}$$

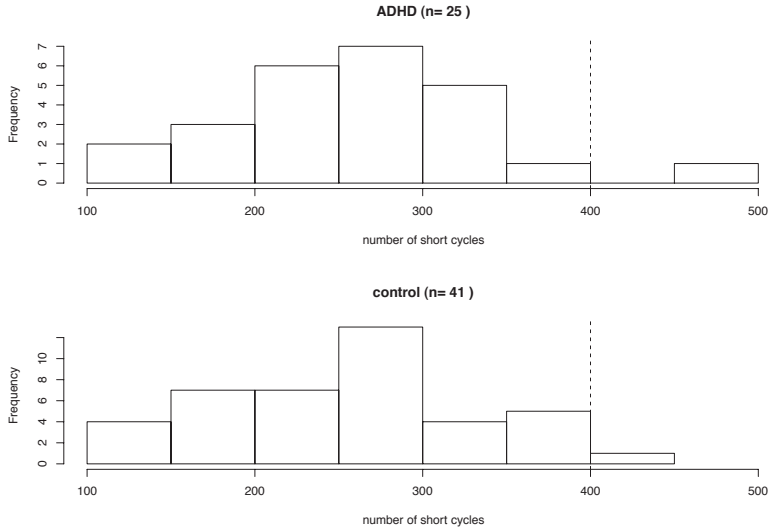

Figure 3: Subject-wise number of short time domain 1-cycles in the DMN in each diagnostic group. The vertical dashed line at abscissa = 400 shows the constant value used in the probability calculation.

Then, by (2),

$$\begin{aligned} Np &= \sum_{z\in S} \text{Prob}\{\text{subject gets cycle } z\} \\ &= \sum_{z\in S} \sum_{Z\subset S;\, z\in Z;\, |Z|=m} \text{Prob}\{\text{subject gets } Z\} \\ &= \sum_{Z\subset S;\, |Z|=m} \sum_{z\in Z} \text{Prob}\{\text{subject gets } Z\} \\ &= \sum_{Z\subset S;\, |Z|=m} m\,\text{Prob}\{\text{subject gets } Z\} \\ &= m \times 1. \end{aligned}$$

Thus, $p = m/N = 0.040$.

Hence, assuming independence across subjects and our simple model, the probability that a particular triplet pops up in 13 or more of the 66 subjects can be computed from the binomial distribution. The probability is 2.131×10^{-6}. (We also estimated this probability by simulation, using the actual numbers of short cycles per subject instead of the fixed number 400. The estimate computed by simulation is $< 5 \times 10^{-7}$.) Hence a Bonferroni upper bound on the probability that *some* triplet is found in 13 or more subjects is $N \times 2.131 \times 10^{-6} \approx 0.021$.

This bound is small. Hence one is inclined to reject the null hypothesis and to believe that the short cycle that did appear among the short time domain default mode 1-cycles of 13 subjects is special.

References

[1] A. Agresti, *Categorical Data Analysis*, Wiley, New York, 1990.

[2] F.G. Ashby, *Statistical Analysis of fMRI Data*, MIT Press, Cambridge, Mass., 2011.

[3] D. Bartholomew, M. Knott, and I. Moustaki, *Latent Variable Models and Factor Analysis: A Unified Approach*, Wiley, Chichester, 2011.

[4] P. Bloomfield, *Fourier Analysis of Time Series: An Introduction*, Wiley, New York, 1976.

[5] B. Boashash, E.J. Powers, and A.M. Zoubir, editors, *Higher-Order Statistical Signal Processing*, Longman, Melbourne, 1995.

[6] R.C. Bradley, *Introduction to Strong Mixing Conditions*, volume 1, Kendrick Press, Heber City, Utah, 2007.

[7] D.R. Brillinger, *Time Series: Data Analysis and Theory*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[8] G. Carlsson, Topology and data, *Bulletin of the American Math. Soc.* 46 (2009), 255–308.

[9] M.K. Chung, P. Bubenik, and P.T. Kim, Persistence diagrams of cortical surface data, in *IPMI '09*, 386–397, 2009.

[10] C. Curto and V. Itskov, Cell groups reveal structure of stimulus space, *PLoS Computational Biology* 4 (2008).

[11] T.K. Dey, A. Hirani, and B. Krishnamoorthy, Optimal homologous cycles, total unimodularity, and linear programming, *SIAM J. Computing* 40 (2008), 1026–1044.

[12] H. Edelsbrunner and J.L. Harer, *Computational Topology: An Introduction*, American Mathematical Society, Providence, 2010.

[13] S.P. Ellis, Background on the concurrence topology method and software, unpublished manuscript available at `http://binarybottle.com/concurrencetopology/`, 2012.

[14] S.P. Ellis, Notes on computational homology, unpublished manuscript available at `http://binarybottle.com/concurrencetopology/`, 2013.

[15] R. Ghrist, Barcodes: The persistent topology of data, *Bulletin of the American Math. Soc.* 45 (2008), 61–75.

[16] E. Goubault, Geometry and concurrency: a user's guide, *Mathematical Structures in Computer Science* 10 (2000), 411–425.

[17] M. Grandis, *Directed Algebraic Topology: Models of Non-Reversible Worlds*, Cambridge University Press, Cambridge, 2009.

[**18**] P. Hall, J.S. Marron, and A. Neeman, Geometric representation of high dimension, low sample size data, *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 67(3) (2005), 427–444.

[**19**] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, New York, 2001.

[**20**] P. Jezzard, P. Matthews, and S. Smith, *Functional MRI: An Introduction to Methods*, Oxford University Press, Oxford, 2002.

[**21**] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, NJ, third edition, 1992.

[**22**] M.J. Kearns, R.E. Schapire and L.M. Sellie, Toward efficient agnostic learning, *Machine Learning* 17 (1994), 115–141.

[**23**] K. Konrad and S.B. Eickhoff, Is the ADHD brain wired differently? A review on structural and functional connectivity in Attention Deficit Hyperactivity Disorder, *Human Brain Mapping* 31 (2010), 904–916.

[**24**] H. Lee, M. Chung, H. Kang, B.-N. Kim, and D.S. Lee, Discriminative persistent homology of brain networks, In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium*, 841–844, 2011.

[**25**] K. Li, L. Guo, J. Nie, G. Li, and T. Liu, Review of methods for functional brain connectivity detection using fMRI, *Computerized medical imaging and graphics: The official journal of the Computerized Medical Imaging Society* 33(2) (March 2009), 131–139.

[**26**] E. Mazerolle, S. Beyea, J. Gawryluk, K. Brewer, C. Bowen, and R. D'Arcy, Confirming white matter fMRI activation in the corpus callosum: Co-localization with DTI tractography, *Neuroimage* 50 (2010), 616–621.

[**27**] M. Mrozek, P. Pilarczyk, and N. Żelazna, Homology algorithm based on acyclic subspace, *Computers & Mathematics with Applications* 55(11) (2008), 2395–2412.

[**28**] Y. Paloyelis, M.A. Mehta, J. Kuntsi, and P. Asherson, Functional MRI in ADHD: a systematic literature review, *Expert Review of Neurotherapeutics* 7 (2007), 1337–1356.

[**29**] J.C. Pinheiro and D.M. Bates, *Mixed-Effects Models in S and S-PLUS*, Springer, New York, 2000.

[**30**] R.A. Poldrack, J.A. Mumford, and T.E. Nichols, *Handbook of Functional MRI Data Analysis*, Cambridge University Press, Cambridge, 2011.

[**31**] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. `http://www.R-project.org/`

[**32**] L.Q. Uddin, A.C. Kelly, B.B. Biswal, F.X. Castellanos, and M.P. Milham, Functional connectivity of default mode network components: Correlation, anticorrelation, and causality, *Human Brain Mapping* 30 (2009), 625–637.

[**33**] M.P. van den Heuvel and H.E. Hulshoff Pol, Exploring the brain network: A review on resting-state fMRI functional connectivity, *European Neuropsychopharmacology: the journal of the European College of Neuropsychopharmacology* 20(8) (May 2010), 519–534.

Steven P. Ellis ellisst@nyspi.columbia.edu

Unit 42, New York State Psychiatric Institute, Columbia University, 1051 Riverside Dr., New York, NY 10032, U.S.A.

Arno Klein arno@binarybottle.com

Sage Bionetworks, 1100 Fairview Avenue North, MS: MI-C815, Seattle, WA 98109-1024, U.S.A.