Abstract

For multivariate data dependence beyond pair-wise can be important. When one has more than a few variables, however, the number of simple summaries of even third-order dependence can be unmanageably large.

"Concurrence topology" is an apparently new nonparametric framework for describing high-order dependence among up to dozens of dichotomous ordinal variables (e.g., we compute descriptions of seventh-order dependence in 32 variables). This method generally produces summaries of p^{th}-order dependence of manageable size no matter how big p is. (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 "1" in the same observation. The collection of concurrences can be represented as a sequence of shapes ("filtration").

Holes in the filtration indicate weak or negative association among the variables. The pattern of the holes in the filtration can be analyzed using computational topology.

We applied this method to dichotomized functional MRI data. The dataset includes subjects diagnosed with ADHD and healthy controls.

In an exploratory analysis we found numerous differences between the diagnostic groups in the topology of the filtrations. (Keywords: dichotomous data, high-order dependence, Fourier analysis of time series, computational homology, persistent homology, fMRI, ADHD)

Describing High-Order Statistical Dependence Using "Concurrence Topology", with Application to Functional MRI Brain Data

Steven P. Ellis and Arno Klein

December 7, 2012

1 Introduction

We propose an apparently new nonparametric framework, "concurrence topology", for describing high-order dependence structure of multivariate ordinal, binary data. It does this by translating the data into a series of shapes and then analyzing the topology of those shapes. (The variables need to be ordinal because the translation process is sensitive to the coding.) In this paper our main focus is a specific approach to concurrence topology we call

"concurrence homology (CH)".

As a test bed for concurrence topology we analyze "functional connectivity" in resting-state functional magnetic resonance imaging data (fMRI, section 8 below, [Jezzard et al., 2002]). 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. (Concurrence topology applies to binary data so we first dichotomized the fMRI BOLD time series values, section 12.) Others have used fMRI to reveal abnormalities in functional connectivity in ADHD [Paloyelis et al., 2007]. We find other differences using concurrence topology.

An ordinal 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.

CH is sensitive to weak or negative, i.e., nonpositive, association. Thus, variables $X_1, \ldots, X_p$ are 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 term $\lambda_{11\ldots1}^{X_1,\ldots,X_p}$ in a log linear model [Agresti, 1990, Section 5.3.1, p. 143] involves not just the marginal distribution of $X_1, \ldots, X_p$ 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 has to do with the "p^{th}-order dependence" among the variables. For example, Pearson, Kendall, and Spearman correlation are measures of 2^{nd}-order dependence because a correlation matrix for a collection of variables can be computed by looking at the variables two at a time. In this paper we focus on "high-order" dependence, by which we mean dependence of order at least three.

Assuming *a priori* a structure for the dependence among the variables or designating *a priori* some variables as "predictors" and others as "responses" can be a powerful way to learn from data. However, our interest is in "agnostic" methods. A method is "agnostic" if *a priori*, for $k = 1, 2, \ldots$ all groups of k variables are treated identically. This rules out much *a priori* structural assumptions. An example of an agnostic method is principal component analysis, a second-order method.

There are apparently few nonparametric agnostic methods that can cope

with the "combinatorial explosion" (section 5) that is inherent in describing high-order dependence among more than a few variables. Other such methods include independent component analysis (ICA, [Hyvärinen et al., 2001]), latent variable methods [Bartholomew et al., 2011], and, perhaps, the method of [Dunson and Xing, 2009]. (There is a large literature on analysis of fMRI data, e.g., [Ashby, 2011, Li et al., 2009, van den Heuvel and Hulshoff Pol, 2010].)

The aforementioned methods and ours capture very different aspects of high-order dependence. Hence, *prima facie* these methods are not competitors. For that reason, and in the interest of brevity, in this paper we do not compare concurrence topology to other methods.

2 Toy examples

Concurrence topology can be explained via "toy examples". (Further explanation can be found in the supplementary material.) Consider the three multivariate datasets shown in table 1. Each has five variables. The univariate marginal distributions are the same across all three datasets. The bivariate marginals are also the same. But the datasets differ in third-order.

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 datasets identical up to second-order, but not at third-order.

For example, in dataset I X, Y, and Z are never all active in the same row, but in the other two datasets they are.

Our method is based on "concurrences". A concurrence is a group of variables that are all active 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.) Call the number of variables in the concurrence the "order" of the concurrence.

For example, the "concurrence list" in data set I in table 1 is YZ, XZ, XY, YZ, XZ, XY, VZ, WX, and VW. (We ignore the order, but not the frequency of appearance, of concurrences.)

We represent a concurrence list as a shape. In general, the shape will not fit on a plane, or in three-dimensional space, for that matter. But the shapes corresponding to data sets I, II, and III do fit on a plane. Choose points ("vertices") on the plane, each corresponding to a variable. If two variables form a concurrence in the list then connect the corresponding vertices by a line segment ("1-simplex"). If three variables are concurrent, connect them by a solid triangle ("2-simplex"). (Of course, if three variables are concurrent then each pair of them are. Implicit in connecting three variables by a 2-simplex is connecting each pair by a 1-simplex.) Vertices may need to be

rearranged so that none fall in the middle of a 2-simplex.

We call this shape the "Curto-Itskov complex" of the concurrence list. (Strictly speaking, a Curto-Itskov complex is more than just a shape. It is a "simplicial complex", a collection of simplices. The name "Curto-Itskov" is explained presently.) The Curto-Itskov complexes of the data in table 1 are shown in the first column of figure 1.

The shapes in the first column of figure 1 distinguish data set I from data sets II and III, but do not distinguish data sets II and III from each other. To do this we construct a decreasing series of shapes, indexed by "frequency level", which is how often a concurrence appears in a concurrence list, $\mathcal{C}$.

We say that a concurrence in $\mathcal{C}$ appears even if it appears as part of a larger concurrence. Thus, in data set II the concurrence XYZ appears once, XY appears twice, and X appears five times. The shape (Curto-Itskov complex), or "frame", in frequency level f is constructed in the manner we just described but from the concurrence list, $\mathcal{C}_f$ that consists of all concurrences in $\mathcal{C}$ that *appear at least f times.* (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 theoretical population version of the filtered Curto-Itskov complex might be indexed by

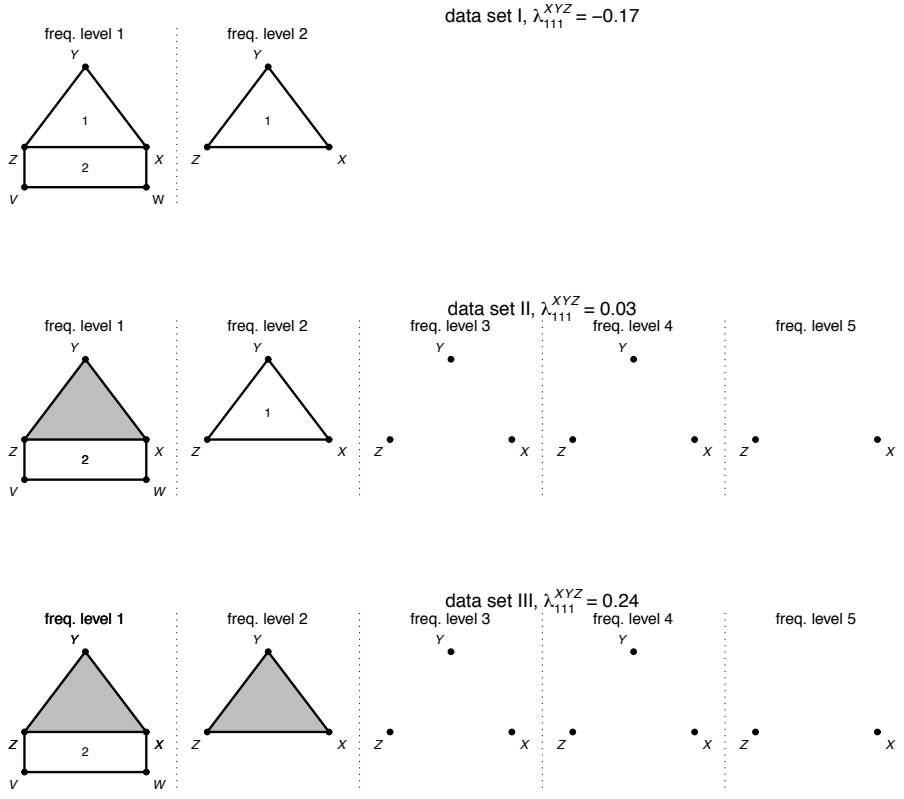

Figure 1: Rows are filtered Curto-Itskov complexes for data sets in table 1. Columns, separated by dotted vertical lines, correspond to frequency levels. "λ_{111}^{XYZ}" is a third-order interaction in a log linear model. "1" and "2" label holes.

a continuum of frequencies; supplemental material.) We call this series of shapes the "filtered Curto-Itskov complex" of the data. (Conventionally in topology a filtered simplicial complex would be indexed in the opposite order. Strictly speaking, we use *descending* filtrations.)

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 do distinguish the three data sets. (In fact, the filtered Curto-Itskov complex for a data set together with the number of observations is equivalent to the contingency table for the data set; supplementary material.)

Our work on concurrence topology is inspired by Curto and Itskov [Curto and Itskov, 2008], who investigated a question in theoretical neuroscience is by applying topological methods to simulated data. From each simulation Curto and Itskov constructed a single shape, like those in the first column of figure 1, and studied the holes in that shape. For their purpose it was not necessary to build a filtration, i.e., a series of shapes. In essence, they only needed to know whether each cell in a contingency table was 0 or not. But, in general, for data analysis one needs to know, not just which values in the table are nonzero, but the actual values in the table. To represent those values geometrically a single Curto-Itskov complex is not sufficient.

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

3 Holes

The main principle of our approach is that holes in filtered Curto-Itskov complexes represent negative or weak association among the variables. The representation takes into account lower-order dependence.

Figure 1 also displays the values of the third-order interaction term λ_{111}^{XYZ} in a log linear model for each data set. This term pertains to the frequency of the event $X = 1, Y = 1, Z = 1$. (The contingency tables for these data sets contain 0-cells. For that reason we added 1/2 to each cell in the tables before computing λ_{111}^{XYZ}.)

Notice that there is a perfect negative association between the number of empty triangles with vertices labeled X, Y, and Z and the values of λ_{111}^{XYZ}. Thus, it seems that the number of these empty triangles 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. We provide a general argument for that contention in section 4.

Now consider the rectangular holes in figure 1 with vertices V, W, X, and Z. If an empty triangle pertains to third-order dependence, one might guess that an empty rectangle pertains to fourth-order dependence. But the order of dependence corresponding to a hole depends, not on the number of its corners (vertices), but on its *dimension*.

A loop of wire represents a 1-dimensional hole because a length of wire is 1-dimensional. The void inside a basketball is 2-dimensional because the basketball consists of sheets of rubber glued together and sheets are 2-dimensional shapes. (Determining the dimension of a hole is not always as straightforward as these examples suggest.)

To illustrate, consider data set IV listed in table 2. Figure 2(a) shows the filtered Curto-Itskov complex for data set IV. In this case the filtered Curto-Itskov complex includes just one frame, which forms a hollow tetrahedron, i.e. a pyramid with a triangular base. To represent it as a plane figure it has been opened up. (To restore the Curto-Itskov complex, imagine folding the figure along each of the lines VW, VX, and WX to bring the three Z's into

V	W	X	Z
0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0

Table 2: Data set IV: A data set with weak fourth-order dependence.

coincidence.)

The sides of the shape are two-dimensional and enclose a two-dimensional hole. For this data set the fourth-order term λ_{1111}^{VWXZ} is -0.27, while the third-order terms λ_{111}^{VWX}, etc., are all -0.14 so, empirically, the shape seems to encode 4^{th}-order dependence.

A hole in a Curto-Itskov complex is a global feature of the joint distribution. I.e., it involves *all* the variables. For example, consider data set V, obtained by dropping the first observation (row) in data set IV. Its filtered Curto-Itskov complex contains only one frame (figure 2(b)). Data set V includes the concurrences WX, WZ, and XZ but not WXZ. Thus, the simplices involving only W, X, and Z form an empty triangle. That, however, is not a hole in the frame because the 2-simplices VWX, VWZ, and VXZ fill in the empty triangle. The fact that holes involve all the variables means that

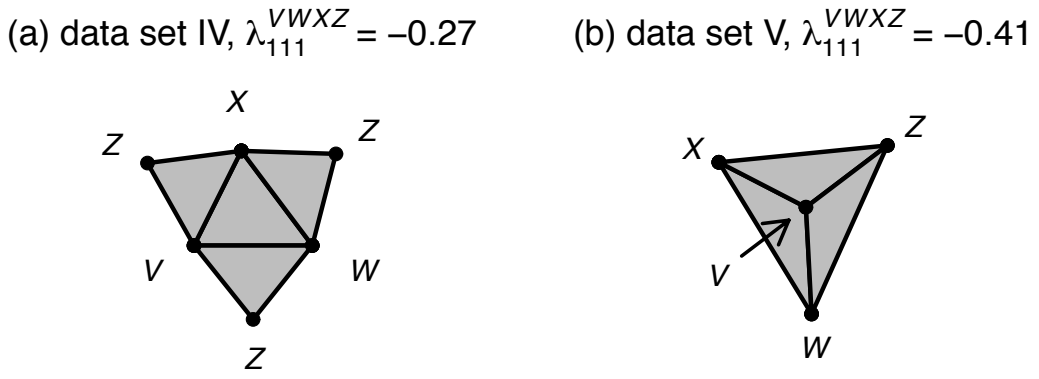

Figure 2: Curto-Itskov complexes for data sets IV and V.

there are usually few holes in filtered Curto-Itskov complexes. This furthers the goal of parsimony (section 5). Now, for data set V, $\lambda_{111}^{WXZ} = -0.41$, which is fairly far from 0. Nonetheless, concurrence topology finds no evidence of weak or negative dependence in this data set.

4 More cases and more variables

Below we consider real brain imaging data sets that have 32 or 74 variables (after some variables have been dropped, section 12). In the latter case it is common for 60 or more variables to be active in a single time point for a single subject. (We analyzed the concurrence topology for each subject separately, then studied the distributions of summaries thereof across subjects.)

We can still apply our approach to such data sets. This can be done abstractly, but as an aid to intuition we start by imagining that the variables correspond to points ("vertices") in general position (for any $k = 1, 2, \ldots$ no set of k points lie on a plane of dimension $k-2$) in a high-dimensional space.

There are simplices of any dimension [Munkres, 1984, §1]. We already observed that a line segment and a solid triangle are simplices of dimensions 1 and 2, respectively. A 3-simplex is a *solid* tetrahedron. (A 0-simplex is a single point.) A simplex is completely determined by its vertices ("corners").

For any collection of concurrent variables, even 60 or more, insert into the high-dimensional space the simplex whose vertices correspond to the concurrent variables. We call the resulting collection of simplices the "Curto-Itskov complex" of the data set. Considering multiplicity of concurrences as in section 2, one gets a series of shapes, the "filtered Curto-Itskov complex".

One cannot detect a d-dimensional hole, η, in a Curto-Itskov complex by only looking at $d+1$ variables at a time, but one can detect a d-dimensional hole by looking at groups of $d+2$ variables at a time. Thus, η pertains to order $d+2$ dependence.

Moreover, the hole η is bounded by at least $d+1$ 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* be active at the same time. Thus, existence of η reflects a shortage of active groups of $d+2$ variables compared to active groups of $d+1$ variables. To sum up:

A d-dimensional hole in a filtered Curto-Itskov complex indicates

weak or negative association of order $d+2$. (1)

"Homology theory" is a branch of algebraic topology that is concerned with tunnels, holes, voids, cavities, etc., in shapes. [Munkres, 1984, Edelsbrunner and Harer, 2010]. The homologically correct term for "hole" is "homology class". In this paper we use homology (with $\mathbb{Z}/2 = \{0, 1\}$ coefficients) to describe the patterns of holes in filtered Curto-Itskov complexes. We call this approach to concurrence topology "concurrence homology (CH)".

We wrote our own CH software in *R* [R Development Core Team, 2008] (supplemental material). 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/).

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 homology, but sometimes a week did not.

5 "Combinatorial explosion"

Later (section 10) we look at seventh-order dependence among the regions of the "default mode network (DMN)" ([Uddin et al., 2009]), in each subject in our fMRI dataset. In our interpretation the DMN consists of 40 regions (supplemental material). For each subject we discarded eight regions (section 12).

An agnostic seventh-order log linear analysis would result in $\binom{32}{7} = 3,365,856$ distinct λ's for each subject (compared to the 6,144 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 [Agresti, 1990, Section 5.4, p. 150].)

By contrast we found that the data summaries produced by CH included at most hundreds of numbers per subject, even if "localization" (section 11) was employed. Moreover, those numbers are structured in a way that aids

interpretation. Thus, CH provides parsimonious descriptions of high-order dependence (the cost is in computation time).

6 Persistence

In the filtered Curto-Itskov complex for data set I, shown in the top row of figure 1, two triangles appear. But the two triangles are related: Moving in decreasing order of frequency level, i.e., from right to left, a triangular hole (labeled "1") appears (is "born") at frequency level 2 and "persists" at frequency level 1. We say that the persistent triangle in data set I "dies in frequency level 0". In data set II again a triangular hole appears (is "born") at frequency level 2, but it "dies" in frequency level 1. By a "persistent (homology) class" we mean a collection of homology classes (holes) in various frequency levels that are related to each other in this way.

Hence, instead of saying that the filtered Curto-Itskov complex of data set I has three 1-dimensional holes, we say that it has two 1-dimensional persistent classes with lifespans 2 and 1. Identifying persistent classes and their the births and deaths is "persistent homology" [Edelsbrunner and Harer, 2010].

Recently there has been much interest in using persistent homology for

data analysis, e.g., [Ghrist, 2008] and [Carlsson, 2009]. In particular, persistent homology has been applied to brain data [Lee et al., 2011, Chung et al., 2009]. However, concurrence topology appears to be a new method.

Table 3 shows that births and deaths of all persistent classes in data sets I, II, and III. Note that the persistent homology of the Curto-Itskov complexes discriminates the three datasets. (CH in dimension 0 registers second-order dependence. It tracks, not holes, but clusters, or, more precisely, connected components, and is akin to single linkage cluster analysis, [Everitt et al., 2011]. The roles of "X" and "Z" can be reversed.)

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", [Edelsbrunner and Harer, 2010, p. 152].) Figure 3 shows the persistence plot in dimension 1 (third-order dependence by equation (1)) for the regions in the DMN for control subject "sub01912". Note that the format of the plot is the same no matter what order of dependence is portrayed. (Plots like this can be averaged over groups; supplementary material.)

	Data set						
	I		II		III		
dimension	birth	death	birth	death	birth	death	comp./hole
0	2	0	5	0	5	0	X
			5	2	5	2	Z
			4	2	4	2	Y
1	2	0	2	1	(not present)		1
1	1	0	1	0	1	0	2

Table 3: Persistent homology of the filtered complexes in figure 1 in dimensions 0 and 1. The column "comp./hole" identifies the component (dimension 0) and hole (dimension 1) whose births and deaths are listed.

Thus, e.g., the point marked by an "*" indicates a persistent 1-dimensional homology class that is born in frequency level 13 and dies in frequency level 3. So as one moves downward from frequency level 13 to 3, the frames include increasing numbers of simplices, but this hole is not filled in until 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.

It turns out that the class corresponding to the point indicated by the "*" is rather special and investigating it led us to find one of several ways of using CH to discriminate ADHD subjects from controls (section 11.1.1).

7 Concurrence topology in the Fourier domain

High-order spectral analysis of multivariate time series is a well-studied subject [Boashash et al., 1995]. 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 the direct dichotomization of BOLD values (section 12). In the "Fourier domain", instead of dichotomizing the BOLD signal itself, one dichotomizes the periodograms [Brillinger, 2001] of the component series.

Define concurrence in the Fourier domain just as in time domain, but instead of time points, use angular frequencies. This allows the study of high-order dependence while taking into account the time series nature of fMRI.

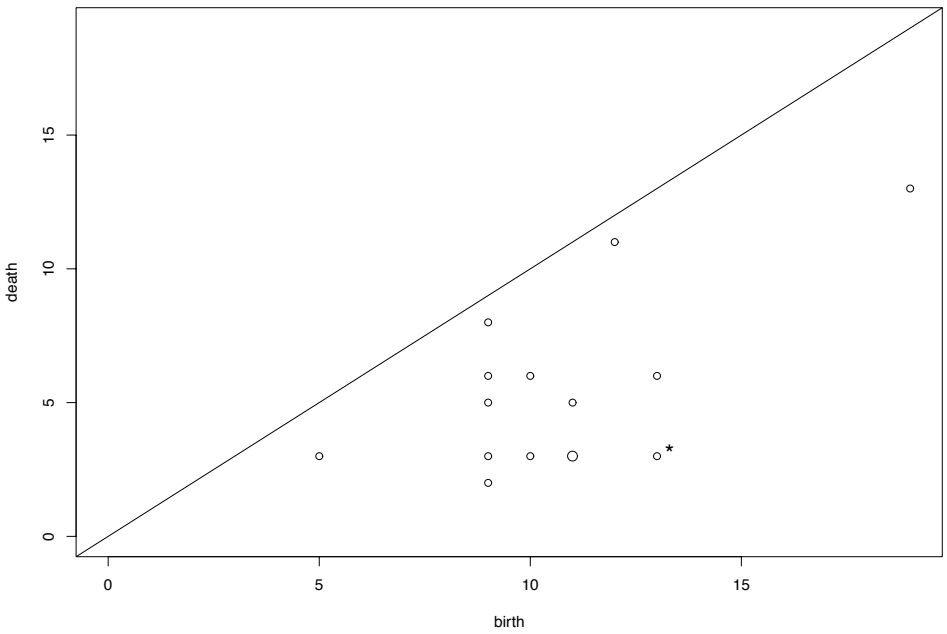

Figure 3: Dimension 1 persistence plot for the joint distribution of a the fMRI BOLD values in the time domain in the default mode network for subject "sub01912". The larger circle indicates two coinciding points. The asterisk indicates a special persistent class discussed in section 11.1.1.

8 fMRI data

The fMRI data set was generated 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 dataset containing clinical data available. 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 = 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). Prior to applying CH we dropped some regions in a subject-wise fashion (section 12).

9 Data analysis of fMRI data

For each subject we computed summaries of the homology of his/her fMRI BOLD data and compared the distribution of those summaries between groups (or, in one instance, between genders). Thus, we performed inference between subjects, not within subjects. Our purpose in this study is to develop methods for using CH in fMRI data. 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 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), [Pinheiro and Bates, 2000].) "Statistical significance" was merely a flag that indicated analytical methods that might be worthwhile for future use.

Unless stated otherwise, *all findings we mention concerning the fMRI dataset are statistically significant in this operational, uncorrected sense.* Because our analyses are only exploratory, to save space we omit many 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 2 through 7) in the DMN. We also computed persistent homology (time and Fourier domains) in dimensions 0 through 2 in the whole brain. In some cases we also computed the corresponding localization (section 11) and/or the "Euler characteristics" for each subject. (The "Euler characteristic" is a one number summary of the homology of a shape [Munkres, 1984, Richeson, 2008].)

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

In some analyses we summarized the main features of a persistence plot by nine "moments": The first "moment" was the number of points in the plot (counting multiplicity). The other "moments" are, for $i, j = 0, 1, 2$ (not both 0), $\left[\text{average of } (birth^i)(lifespan^j)\right]^{1/(i+j)}$. For the DMN we computed persistent homology in dimensions 0 through 5. Hence, for the DMN we obtained for each subject $6 \times 9 = 54$ moments. 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. We analyzed these multivariate

summaries using GLS with *moment* as the response variable.

10 Some findings

Using the GLS analysis just described we picked up group differences in the DMN in the time domain in dimensions 4 and 5 and in the whole brain in the Fourier domain in dimensions 1 and 2.

The group difference in the DMN in the time domain in dimensions 4 was a robust finding, in the sense that it manifested itself in a number of analyses. The essence of the difference is that fewer ADHD subjects (64.0%) had any homology in the time domain in the DMN in dimension 4 (i.e., only 64% had any 4-dimensional holes; this represents 6^{th}-order dependence by equation (1)) than did controls (92.6%).

In the DMN in the Fourier domain the Euler characteristic of the frame in frequency level 1 is typically higher among the ADHD subjects (mean = 1.68, 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 inde-

pendently of each other is far different from what one finds in the real fMRI 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 hole (i.e., homology class), it is natural to ask what variables (regions, in our case) are involved? Existence of a hole in the filtered complex requires the cooperation of all variables, but some variables are more directly involved than others.

For simplicity consider dimension $d = 1$ (third-order dependence, by equation (1)). In dimension 1, a "cycle" ("1-cycle") is a union of one or more empty polygons made up of 1-simplices. (In general, a d-cycle is a union of hollow polyhedra whose faces are d-simplices.) E.g., using an obvious notation, in figure 1, data set I, $XY + YZ + ZX$ is a 1-cycle. That cycle is "short" because it consists of just three 1-simplices. In general, a "short" d-cycle is one consisting of $d+2$ simplices, the smallest number that can form a d-cycle.

In homology we are interested in cycles that wrap around holes. All holes are surrounded by cycles. But not all homology classes have short cycles. E.g., in figure 1 again, there are no short cycles that surround the hole marked "2". Conversely, even if a homology class has a short cycle, in general it will also have cycles that are not short. By "localization" of a hole, we mean finding *all* short cycles, but only short cycles, that surround that hole, if there are any. ([Dey et al., 2008] discusses a different notion of localization.) We computed the localization of *all* holes in various dimensions in all subjects.

11.1 Localization in the fMRI data

11.1.1 Dimension 1 in the DMN in the time domain

In the DMN in the time domain in dimension 1 we found 7,427 distinct short cycles across all subjects. (There are 40 regions in the DMN. $\binom{40}{3} = 9,880$ distinct short cycles are theoretically possible; median number of distinct short 1-cycles per subject = 260.) One subject has a one-dimensional homology class (hole) containing 164 short 1-cycles in a single frequency level.

We select the most important short cycles using two criteria. The first is the number of subjects having the cycle and the second is the lifespan of the

cycle. 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.

The short cycle whose homology class persistence is plotted at the point marked by "*" in figure 3 appears in 13 subjects and, for subject "sub01912", has cycle lifespan = 8 (supplementary material). Call this cycle z. This triplet of regions tend to be well connected at second-order, but, comparatively speaking, not even indirectly well connected at order 3.

We performed an analysis under the null hypothesis that all possible 9,880 triplets of default mode regions are equally likely to be short cycles in a given subject. We assumed that short cycles were selected from the 9,880 independently *between* subjects, but not necessarily independently *within* subjects. Then, based on a simple model, a heavily Bonferronized upper bound on the probability of a triplet being a short cycle for 13 or more subjects is 0.021. Thus, z is rather special.

Now, z itself does not differentiate the ADHD and control groups, but the 29 short cycles that wrap tightly around the same hole that z does in subject "sub01912" *do* distinguish the groups.

We can refine this. 16 of the 29 short cycles appear at least twice each in each diagnostic group (supplementary material). 19 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 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 16. Something more subtle is at work. It appears that there is a particular hole or family of related holes that occur in many of the subjects' filtered Curto-Itskov complexes. This hole or family is more common, though, among the ADHD subjects than among the controls.

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 1,497 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. Thus, the holes marked "1" in figure 1 are narrow, while the holes marked

"2" are not. Homology classes can be added ([Munkres, 1984, Chapter 1]). Say that two narrow classes are "adjacent" if their sum is also narrow. (The holes corresponding to the adjacent classes do not actually have to be next to each other in space.)

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-Fourier domain 7,933 appear in the data. The median number of distinct short 2-cycles per subject = 57.5. (A short 2-cycle involves four regions.)

The "corpus callosum" consists of white matter and until recently only grey matter was believed to produce a BOLD signal [Mazerolle et al., 2010]. However, the median number of times that one of the five corpus callosum regions (supplemental material) appears in a short cycle is 1,305, while the median for non-corpus callosum regions is 249. Of the theoretically possible 2-dimensional short cycles in the whole brain, 20% include a region from the

corpus callosum, but, of the distinct short 2-cycles in the data, 65% include at least one corpus callosum region.

Thus, the corpus callosum frequently takes part in quadruplets that are weakly connected at fourth-order. However, *positive* dependence at third-order is needed in order to form the four 2-simplices in a short 2-cycle. So the five corpus callosum regions cannot be said to be weakly functionally connected to other regions in general.

12 Dichotomization

Concurrence topology is designed for binary data. The fMRI BOLD signal is continuous. For each region 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 regional fMRI BOLD levels 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 ("*rcv*"): 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 if low variability regions are dropped.

We stress 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 the 20% (39) time points at which the fMRI BOLD value was highest as "active". In the Fourier domain, for each subject and region we set the threshold at the 90^{th} 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.

13 Discussion and Conclusions

Concurrence topology is a nonparametric framework for describing high-order dependence in ordinal dichotomous data. We described a particular approach to concurrence topology, called "concurrence homology (CH)". We explored a number of different ways of deploying CH using an fMRI data set as a test bed. These included persistence, Euler characteristics, and several different ways of mining localizations and 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 .

CH is computationally intensive but, with that proviso, concurrence topol-

ogy can be applied, not just to fMRI BOLD data, but to any multivariate ordinal dichotomous data. Moreover, we are confident that improved software will greatly expand the range of data that can be analyzed using CH.

Supplementary material for this paper includes software (Concurrence-Topology, Copyright 2012, Steven P. Ellis, Apache v2.0 License) and data, with further documentation, as well as a few details of our findings concerning the fMRI data set.

References

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

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

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

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

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

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

[Chung et al., 2009] Chung, M. K., Bubenik, P., and Kim, P. T. (2009). Persistence diagrams of cortical surface data. In *IPMI'09*, pages 386–397.

[Curto and Itskov, 2008] Curto, C. and Itskov, V. (2008). Cell groups reveal structure of stimulus space. *PLoS Computational Biology*, 4:1000205.

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

[Dunson and Xing, 2009] Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.*, 104:1042–1051.

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

[Everitt et al., 2011] Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 5th edition.

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

[Hyvärinen et al., 2001] Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley, New York.

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

[Lee et al., 2011] Lee, H., Chung, M., Kang, H., Kim, B.-N., and Lee, D. S. (2011). Discriminative persistent homology of brain networks. In *Biomed-*

ical Imaging: From Nano to Macro, 2011 IEEE International Symposium, pages 841 – 844.

[Li et al., 2009] Li, K., Guo, L., Nie, J., Li, G., and Liu, T. (2009). 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):131–9.

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

[Munkres, 1984] Munkres, J. R. (1984). *Elements of Algebraic Topology*. Benjamin/Cummings, Menlo Park, CA.

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

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

[R Development Core Team, 2008] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Richeson, 2008] Richeson, D. S. (2008). *Euler's Gem: The Polyhedron Formula and the Birth of Topology*. Princeton University Press, Princeton.

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

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