BACKGROUND ON THE CONCURRENCE TOPOLOGY METHOD AND SOFTWARE

STEVEN P. ELLIS

1. Introduction

This document is part of the supplementary material for the paper "Describing high-order statistical dependence using 'Concurrence Topology', with application to functional MRI brain data" by Steven P. Ellis and Arno Klein [EK13]. Here we give an overview of how the concurrence topology software works. A bit of theory is also included. For instructions on how to use the software see the file "SoftwareHowTo.txt". For some related mathematical theory see the file "ConcurrenceTopol_Notes.pdf". Both these files are found in the directory "ConcurrenceTopology_software" in the supplemental material.

2. The filtered Curto-Itskov complex

2.1. **Simplicial complexes.** Concurrence topology is designed for binary data[1]. 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 basic idea is to represent a list of concurrences by a "simplicial" complex (Munkres [Mun84]). Simplicial complexes can be defined abstractly, but as an aid to intuition we consider geometric complexes. A simplicial complex is a collection of simplices that fit together nicely. Simplices are shapes. There are simplices of any nonnegative dimension. A 0-dimensional simplex ("0-simplex") is just an individual point. A 1-dimensional simplex ("1-simplex") is a closed line segment. The endpoints of a line segment are the "vertices" of the 1-simplex. We regard vertices as 0-simplices, and we say that the 1-simplex is "spanned" by the two vertices. A 2-simplex is a filled-in triangle. It is "spanned" by its three vertices. A 3-simplex is a tetrahedron, i.e., a solid pyramid with a triangular base. It is spanned by its four vertices (corners). Etc. A simplex is specified by listing its vertices. The sides, edges, and vertices of a 3-simplex are generically referred to as the "faces" of the simplex. A simplex is considered a face of itself. If τ is a face of a simplex σ, but $\tau \neq \sigma$ then we say that τ is a "proper face" of σ.

Date: 3/25/'12.

[1]Binary variables can be coded "0"or "1". The process of translating the data into shapes is sensitive to the coding. This creates an issue for nominal data, e.g., "male-female". If there are only one or two nominal variables in the data set, it might be reasonable to analyze the data with all possible codings. Otherwise, a blanket rule for coding of nominal variables is needed. An example of such a rule is the following. Suppose, as is typical, that for no nominal variable is it the case that the numbers of observations in the two nominal categories are exactly equal. In that case, for each nominal variable code the smaller of the two categories as "1" and the other as "0".

This terminology extends to simplices of all dimensions. A d-dimensional simplex will have $d+1$ vertices. The number of p-dimensional faces of a d-simplex is $\binom{d+1}{p}$ (Thus, a d-simplex has $2^{d+1}-1$ faces altogether.) The number of faces can thus be quite large. This is a manifestation of a "combinatorial explosion". This fact has important consequences when it comes to computing the homology groups of simplicial complexes (sections 5 and 7).

A collection of vertices in "general position" span a unique simplex. $d+1$ points are in general position if they do *not* lie in a plane of dimension $d-1$ or less. Since we are using topological methods, we will not need the exact locations of the vertices. That they are in general position is sufficient.

A *simplicial complex*, K, is a collection of simplices with the following properties.

(1) Any two simplices in K either do not intersect at all or their intersection is a face of both of them.
(2) If a simplex $\sigma \in K$ then all faces of σ also belong to K.

The "dimension" of a complex is the largest dimension of any of the constituent simplices. The vertices of a complex are just all the vertices of all the simplices in the complex.

In concurrence topology all simplicial complexes will be finite, i.e., consist of only finitely many simplices. So a finite complex has finite dimension.

2.2. **From concurrences to simplicial complex.** Suppose one has V variables (in our case each variable corresponds to a brain region). To each variable associate a point in space. These points will be the vertices of a simplicial complex. In order to be free to have simplices spanned by any subset of these points, we need place them in general position, which means they must sit in a space of dimension at least $V-1$.

So for the fMRI data, after dropping the 20% least variable regions (section 12in Ellis and Klein [EK13]), in order to house a simplicial complex having one vertex per region, we potentially need 73 dimensional space for whole brain analyses. For the DMN ("Default Mode Network", Uddin *et al* [UKB+09]), 31 dimensional space is needed. Note that the locations of the points need not have anything to do with the physical locations of the corresponding regions in the brain. Our interest is functional, not structural (i.e., anatomical).

Suppose one has a list, $\mathcal{C}$, of concurrences among the V variables. (See section 3.1 for some discussion of alternative ways of generating concurrence lists.) The "Curto-Itskov complex" determined from $\mathcal{C}$ consists of all the simplices determined as follows. If variables $v_0, \ldots, v_d$ are concurrent then we join the vertices corresponding to those variables by the simplex spanned by the vertices corresponding to those variables. That simplex is included in the complex. (This idea is reminiscent of statistical graphical models, Lauritzen [Lau96]. Apparently, there is no connection.)

This is a well-defined procedure. If variables $v_0, \ldots, v_d$ are concurrent in an observation, then obviously so is any subset of $\{v_0, \ldots, v_d\}$. But the simplex corresponding to the subset is just a face of the simplex corresponding to variables $v_0, \ldots, v_d$ so it is automatically included in the complex, by property 2 of simplicial complexes.

In the time domain, an important property of the Curto-Itskov complex is that *all temporal information is lost* in its construction. concurrence topology represents each concurrence as a simplex. If the concurrence were to take place at a different time, it would still appear as the same simplex in the complex. In the Fourier domain time dependence is preserved, but angular frequency information is lost: A collection of regions is concurrent in the Fourier domain if their fMRI BOLD time series all have power at the same angular frequency. It does not matter what that angular frequency is.

3. Polytope

The Curto-Itskov complex is annotated with "points of interest" and boundaries between simplices. A simplicial complex is not a shape. It is a collection of shapes. However, a complex clearly determines a shape, which is created by assembling all the constituent simplices together in space and ignoring their identities as individual simplices. That shape is called the "polytope" of the simplex. This is illustrated by figure 1. A shape that is the polytope of some simplicial complex is called a "polyhedron". (A polyhedron will actually be the polytope of infinitely many different simplicial complexes.) In concurrence homology, the version of concurrence topology that is the focus of the paper, we analyze the holes in the polytopes.

Going from complex to polytope discards information. Replacement of a simplicial complex by its polytope loses information. E.g., from figure 1(2) we cannot tell how the shape is divided up into 2-simplices in figure 1(1). Moreover, we describe the polytope topologically, which also sheds information. E.g., in figure 1(2) one hole is triangular, the other rectangular. Topology cannot make that distinction. (However, in "localization", section 6, we resurrect that distinction.) One consequence of this approach is the following principle.

(1) An indirect connection is as good as a direct connection.

It is because of this principle that single linkage cluster analysis is the appropriate agglomerative cluster analysis analogue of concurrence homology in dimension 0. Concurrence topology is predicated on the hope that the information that remains in the topology of the polytope is still helpful.

(Actually, the simplicial complex structure is used in computing homology. But the results of the homology calculation are the same for all simplicial complexes having the same polytope. In that sense the homology depends only on the polytope.)

Principle (1) is a little disturbing. Certainly we wish to distinguish between a case in which regions A and D are frequently active at the same time from one in which, say, A and B are frequently active at the same time, B and C are frequently active at the same time, and C and D are frequently active at the same time. Homology erases that distinction.

However, the flip side of this actually furthers our aims: If a group of regions are not even indirectly connected, then unambiguously they are weakly connected. I.e., (1) serves as a noise reduction technique in finding weak connectivity. Assume that communication among brain regions causes simultaneously elevated fMRI BOLD levels in the regions. Then if a group of regions is not even indirectly functionally connected, then even indirect communication among the regions must be weak.

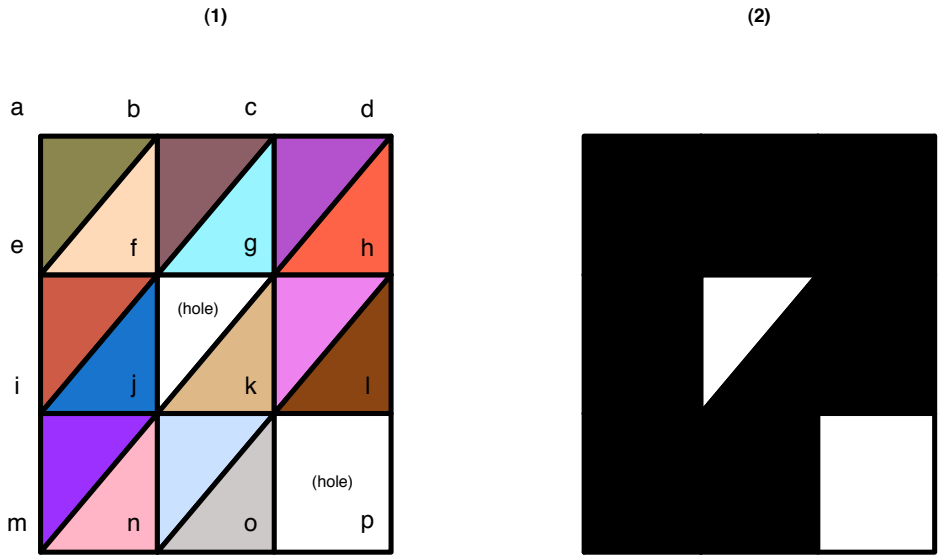

FIGURE 1. (1) A two-dimensional simplicial complex with two 1-dimensional holes (in white). (2) The polytope of the complex in (1).

3.1. **Filtering.** If a concurrence never appears then the corresponding simplex is not present in the Curto-Itskov complex. If a concurrence appears the simplex is present, but the simplex can only be present once in the complex whether the concurrence appears one time or 20 times.

For the purposes of capturing connectivity, this is a serious limitation of the Curto-Itskov complex. The Curto-Itskov complex is dichotomous in the sense that either a simplex is in the complex or it is not. Association among variables, on the other hand, is quantitative. The more frequently a concurrence appears in binary multivariate data, the stronger is the association or connectivity among the variables.

In order to capture frequency information geometrically, we "filter" the concurrences by frequency and construct a series of lists, each one containing the next. The result is a "filtered concurrence list". A concurrence is a collection of binary variables (regions, in dichotomized fMRI data). To filter, one begins with a list, $\mathcal{C}$, of concurrences. A single concurrence may be present multiple times in the list, but to translate $\mathcal{C}$ into a complex, K, as in section 2.2, one ignores multiplicity.

Define a "basic" concurrence in a single multivariate "0–1" observation to be the group of *all* variables that are "1" in that observation, if there are any. Thus, if R is a basic concurrence in an observation (so R consists of variables) then any variable not in R must be "0" in that observation. A concurrence that only appears as a proper subset of a basic concurrence is not basic.

Let R be a basic concurrence in $\mathcal{C}$. Define the number of "appearances" of R in $\mathcal{C}$ to be the number of concurrences R' in the list such that $R \subset R'$. Thus, in general R "appears" more often than it is actually listed in $\mathcal{C}$, but it has to be present at least once (as a basic concurrence).

The first step in constructing a filtration from a list $\mathcal{C}$ is to produce a "filtered concurrence list," $\mathcal{C}_1 \supseteq \mathcal{C}_2 \supseteq \cdots \supseteq \mathcal{C}_F$, where each $\mathcal{C}_i$ $(i = 1, \ldots, F)$ is a concurrence list. The first concurrence list in the series, $\mathcal{C}_1$, corresponding to frequency level 1, consists of all basic concurrences. Such concurrences appear at least once. Thus, $\mathcal{C}_1 = \mathcal{C}$. The list $\mathcal{C}_2$ consists of all the simplices that correspond to all basic concurrences that appear at least twice. Clearly, $\mathcal{C}_2 \subset \mathcal{C}_1$. Etc. $\mathcal{C}_k$ consists of all concurrences that appear at least k times. (For the data discussed in the paper every subject had the same number, *viz.* 192, time points. This makes permissible to use absolute frequencies, i.e., counts, as index. Otherwise, relative frequencies need to be used.)

Given a filtered concurrence list, $\mathcal{C}_1 \supseteq \mathcal{C}_2 \supseteq \cdots \supseteq \mathcal{C}_F$, to construct from it a filtered simplicial complex, one just translates each $\mathcal{C}_i$ into a simplicial complex, K_i, as in section 2.2 $(i = 1, \ldots, F)$. This results in a filtered complex $K_1 \supseteq K_2 \supseteq \cdots \supseteq K_F$. *NOTE: We index filtered complexes in the manner opposite to the conventional one (Edelsbrunner and Harer* [EH10, pp. 70 and 151]*) because that seems more natural in concurrence topology.*

By beginning with different lists $\mathcal{C}$ one gets different filtrations. Define a "degree 1" filtration to be one obtained by taking the initial concurrence list, $\mathcal{C}$, to consist of only basic concurrences. Suppose one includes in $\mathcal{C}$ not just basic concurrences, but also all pairwise intersections of basic concurrences. From this $\mathcal{C}$ one obtains a "degree 2" filtration. Including as well all intersections of triples of basic concurrences, leads to a "degree 3" filtration, etc. We just use degree 1 filtrations.

An attractive idea is to combine filtering by frequency with varying dichotomization cutoffs (section 12in Ellis and Klein [EK13]) to produce a doubly indexed filtration. We did not attempt this, which in any case appears to be hard to work with (Carlsson *et al* [CSZ09]).

3.2. **Relationship between filtered complex and contingency table.** Here we show that the filtered Curto-Itskov complex together with the number of observations (number of time points, in the fMRI example) contains all the information in the contingency table of the data. And *vice versa*.

Let V be the number of variables (regions). Concurrences correspond to cells of the multiway contingency table, T. T can be regarded as an nonnegative integer valued function on the set of 2^V strings of length V consisting of "0"'s and "1"'s. Call such a string a "cell". Denote by $\mathbf{0}$ the cell that consists of all "0"'s. Any cell, c, except $\mathbf{0}$ corresponds

to a concurrence (In practice, except when V is small, the table will be extremely sparse. I.e., $T(c) = 0$ for all but a tiny fraction of cells, c.)

Consider a filtered Curto-Itskov complex, $\mathcal{K}$, with frames

$$K_{f_1} \supseteq K_{f_2} \supseteq \ldots \supseteq K_{f_F},$$

where $0 < f_1 < f_2 < \cdots < f_F$. The f_i's might be absolute (i.e., counts) or relative frequencies. In this section frequencies will be absolute, but in the next section they will be relative. Suppose $\mathcal{K}$ was formed from a "degree 1" filtered concurrence list as in section 3.1.

We show how to compute the contingency table, T, from $\mathcal{K}$ and the number, N, of observations (time points, in fMRI case). Let σ be a simplex belonging to some K_{f_i}. Let R_σ be the corresponding concurrence and let c_σ be the corresponding cell. If R_σ is not basic (as defined in section 3.1; i.e., R_σ appears merely as a proper subset of another concurrence), then σ will only appear in $\mathcal{K}$ as a proper face of some other simplex. In that case let $f_\sigma := 0$ and let $T(c_\sigma) = 0$.

Suppose C_σ *is* basic. Then σ is a "facet" in some K_{f_i}, i.e., $\sigma \in K_{f_i}$, but σ is not a proper face of any simplex in K_{f_i}. Let f_σ be the largest frame index of $\mathcal{K}$ in which σ appears. The simplex σ will be a facet in K_{f_σ}. Define

$$\phi_\sigma := f_\sigma - \sum_{\sigma \text{ is a proper face of } \sigma'} f_{\sigma'}, \tag{2}$$

where the sum is taken over all simplices, σ', in any K_{f_i}, properly containing σ. (Such a σ' must belong to K_f with $f < f_\sigma$.) Note that, since $f_{\sigma'} = 0$ unless σ' is a facet in some K_{f_i}, only facets σ' will contribute nontrivially to the sum in (2). Let $T(c_\sigma) = \phi_\sigma$.

Define

$$\phi := \sum_\sigma \phi_\sigma \leq N.$$

Let $T(\mathbf{0}) = N - \phi$. The resulting table T will be the contingency table for the dataset from which $\mathcal{K}$ was constructed.

Conversely, we can derive the Curto-Itskov complex from the contingency table. Let $c \neq \mathbf{0}$ be a cell, let R^c be the corresponding concurrence, and let σ^c be the simplex corresponding to R^c. If $T(c) = 0$, then R^c will not appear in the data and σ^c will only be present in $\mathcal{K}$, if at all, as a proper face of some other simplex. Otherwise, let

$$f^c := \sum_{R^{c'} \supset R^c} T(c'),$$

where the sum is taken over all concurrences $R^{c'}$ corresponding to a cell c' and containing, not necessarily properly, R^c. Place the simplex σ^c (and all its faces) in frame f^c.

4. Population level

We have presented concurrence homology as a descriptive method. However, it is natural to ask, what population functional does concurrence homology is estimate? Let X be a vector of V "0–1" variables. In our fMRI data, each subject has the same number of time

points, *viz.* 192. That gives us the luxury of labeling the frames by *absolute* frequency of occurrence. In general, one would label the frames by the *relative* frequency, which in this instance means dividing by 192.

Similarly, one could describe the population distribution by a table, T, such that for every cell c, $T(c) = Prob\{X = c\}$. If one applies to T the recipe described in the preceding section, the result is a filtered complex, $\mathcal{K}$, indexed by probability. The empirical filtered Curto-Itskov complex, indexed by relative frequency, is an estimate of $\mathcal{K}$. We have made no attempt to assess how good this estimate is, its sampling variability, or how it might be improved upon.

5. Homology

Here we briefly and informally describe how homology (with $GF(2) = \mathbb{Z}/2\mathbb{Z} := \{0, 1\}$ coefficients) works (Munkres, [Mun84], Sato [Sat99], Edelsbrunner and Harer [EH10], Zomorodian [Zom05]). For simplicity, consider 1-dimensional homology of a simplicial complex K (section 2.1). The homology of K only depends on its polytope, $|K|$, but the complex K provides a "scaffolding" that we use to compute the homology of $|K|$.

Loosely speaking, a "1-cycle" is a union of one or more closed polygons made up of 1-simplices. Roughly speaking, two 1-cycles z and z' are "homologous" if the space between them is completely filled in by 2-simplices in K. The collection of cycles that are homologous to z, called a "homology class", is typically large. This homology class is denoted by $[z]$. We say that any 1-cycle in $[z]$, e.g., z itself, "represents" $[z]$.

We are interested in the cycles that go around one or more holes in $|K|$. If z goes around a hole, then so does any cycle homologous to z. We ignore homology classes that do not go around holes, so when we refer to a "homology class" we always assume that its representatives go around a hole. (So by "homology class" we always mean"nontrivial homology class".) The "one-dimensional homology group of K", $H_1\big(K; GF(2)\big)$, is just the collection of its one-dimensional homology classes.

Figure 1 illustrates. (This complex fits on a plane. In general, a complex with 16 vertices might not fit in Euclidean space of fewer than 15 dimensions.) We write cycles as formal sums of 1-simplices. So, e.g., $jf + fg + gj$ and $ae + ei + im + mj + jg + gd + dc + cb + ba$ are homologous cycles that go around the hole in the middle of the complex. The cycle $jn + nk + kh + hg + gj$ does not go around a hole, so we ignore it.

We described homology in dimension 1, but one can describe $H_d\big(K; GF(2)\big)$ in a similar fashion for any dimension $d = 1, 2, \ldots$. ($H_0\big(K; GF(2)\big)$ is a little different.) In our set up, $H_d\big(K; GF(2)\big)$ is just a vector space. The dimension of this vector space (not to be confused with the dimension d) is the Betti number β_d. Since we only consider finite complexes K all Betti numbers will be finite. In fact, only finitely many of them will be nonzero.

5.1. **Euler characteristic.** For $d = 0, 1, 2, \ldots$ the d^{th} Betti number, β_d, of a shape, X, describes the pattern of d-dimensional holes in X. For example a (hollow) 2-dimensional sphere with k holes in it has $\beta_1 = k - 1$. Another example, which is just the first example in disguise, is as follows. If X is a (2-dimensional) disk with k holes, then $\beta_1 = k$. If

$\dim X = D < \infty$ (this is the case for all shapes considered in concurrence topology), then $\beta_d = 0$ if $d > D$.

Let $d = 1, 2, \ldots$. A d-dimensional hole is enclosed by d-dimensional simplices. If β_{d-1} is large, i.e., if there are many holes in K of dimension $d-1$, that might cut into K's supply of d-dimensional simplices, making it harder for d-dimensional holes to form, thereby reducing β_d. Hence, as a measure of (weak or negative) order $d+2$ dependence, β_d is influenced by lower order dependence.

A similar phenomenon occurs in linear and log linear models and to get around this interactions are often defined as sums of marginal means that alternate in sign according to the dimension of the margin (Scheffé [Sch59, Table 4.6.1, p. 125], Agresti [Agr90, Section 5.3.1, p. 143]). This suggests that the alternating sum $\beta_0 - \beta_1 + \beta_2 - \beta_3 + \cdots$ might be a useful single number summary of dependence. In fact, using a different version of homology (integer coefficients; we use $GF(2)$ coefficients), this alternating sum is just the Euler characteristic, $\chi(K)$ (Munkres, [Mun84, p. 124], Richeson [Ric08]). There is an algorithm for computing $\chi(K)$ that is usually quite fast (section 7.2).

5.2. **Persistence.** Suppose $K_1 \supset K_2 \supset \cdots \supset K_N$ is a filtered simplicial complex. If z is a 1-cycle in K_j then z is also a 1-cycle in K_{j-1}. But if z surrounds a hole in K_j it might not do so in K_{j-1} because K_{j-1} might include one or more 2-simplicies, not already in K_j, that fill in the hole that z surrounds. Thus, z might not represent a homology class in K_{j-1}. In this case we say the homology class represented by z in K_j "dies" in K_{j-1}. If, on the other hand, z still surrounds a hole in K_{j-1} then the homology class $[z] \in H_d\big(K_j\big)$ "persists" in $H_d\big(K_{j-1}\big)$. The class may die in some lower level of the filtration or it may never die. I.e., z may still surround a hole in K_1.

Conversely, the appearance of some simplices in K_{j-1} not already present in K_j might create a hole in K_{j-1} that is not present in K_j. This will give rise to a new homology class. In that case, we say that a new homology class is "born" in K_{j-1}. By a "persistent homology class" we mean the collection of homology classes in various frames that are related to each other as described above. Identifying the births and deaths of the homology classes in the various K_j's is "persistent homology" (e.g., Edelsbrunner and Harer [EH10], Zomorodian [Zom05]). Plotting *death* vs. *birth* yields a "persistence plot" for each dimension d.

6. Localization

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. We saw that in the complex portrayed in figure 1(1), the cycles $z_1 := jf + fg + gj$ and $z_2 := ae + ei + im + mj + jg + gd + dc + cb + ba$ represent the same homology class in $H_1\big(K; GF(2)\big)$. Both cycles wrap around the triangular hole in the middle of the drawing. However, while z_2 only loosely wraps around the hole, the cycle z_1, hugs the hole tightly. It is natural to regard the cycle z_1 as the "location" of the hole.

Let K be a simplicial complex. Call a cycle representing a class in $H_d\big(K; GF(2)\big)$ $(d = 1, 2, \ldots)$ a "short cycle" if it includes only $d + 2$ d-dimensional simplices. Note that $d + 2$ is the smallest number of d-dimensional simplices that can form a d-cycle. Thus, $jf + fg + gj$ is a short cycle.

Not all homology classes are represented by short cycles. The hole in the lower right hand corner of figure 1(1) cannot be represented by the sum of three 1-simplices because it has four sides.

For a given dimension $d > 0$, we used an algorithm (section 7.6) that finds *all* short d-cycles that represent *any* homology class at any frequency level. Chen [CF08] and Dey *et al* [DHK08] concern themselves with a different problem, viz., finding *one* short, or otherwise optimal, representative cycle for one or, perhaps, each homology class in a *basis*.

7. Homology algorithms

Partly as an exercise to learn more about computational homology, we wrote our own computational homology software. Other software for computing persistent homology include the Dionysus (http://mrzv.org/software/dionysus), the Perseus Software Project (www.math.rutgers.edu/ vidit/perseus.html), and CHomP (http://chomp.rutgers.edu/). We implemented the algorithms described below in R (R Development Core Team [R D08]).

The computations described below can be rather expensive in terms of computing time. The distribution of running times required for the subjects in our fMRI dataset has a very long right tail. Using our software, the per subject running times varied from less than an hour to as long as 10 days! We expect that if at least some of our code were written in a compiled language the result would be a substantial increase in speed.

7.1. **Filtered complex.** References on homology include Munkres [Mun84], Edelsbrunner and Harer [EH10], and Kaczynski *et al* [KMM04]. Persistent homology describes the relationship among the homology of a filtered complex

$$K_1 \supseteq K_2 \supseteq \cdots \supseteq K_n = K. \tag{3}$$

Here K_1, K_n and K are simplicial complexes. Call the K_i's "frames". The sequence (3) gives rise to a corresponding sequence of homology homomorphisms. (In our software we use $GF(2)$ coefficients.) Persistence has to do with the images of homology classes under the homomorphisms induced by the inclusion maps.

"Boiling down:" The main obstacle to computing persistent homology is a "combinatorial explosion". This manifests itself in the fact that a high dimensional complex contains very many simplices. To shed some simplices we employ a step analogous to an "elementary collapse" (Kaczynski et al [KMM04, Definition 2.64, p. 71], Edelsbrunner and Harer [EH10, p. 72],). Care must be used in boiling down because we want the boiled down complexes to continue to be nested as in (3).

7.2. **Euler characteristic.** Suppose A and B are two subcomplexes of a complex K. Then

$$\chi(A \cup B) = \chi(A) + \chi(B) - \chi(A \cap B), \tag{4}$$

where $\chi(A)$ is the Euler characteristic of A, etc. (See section 5.1.) Equation (4) plus the fact that the Euler characteristic of a simplex is 1 form the basis for a recursive algorithm for computing Euler characteristics. We find that for the complexes we encountered in our analysis of the fMRI data this recursive algorithm was quite fast.

7.3. **Dimension 0.** We use non-reduced homology in dimension 0.

7.4. **"Excision trick".** The most important feature of our algorithm is what we call the "excision trick". This idea was also proposed in Mrozek *et al* [MPZ08] in the context of cubical homology (Kaczynski *et al* [KMM04]). While this trick can always be used, it works exceedingly well for the fMRI data. The standard method for computing homology overcomes the combinatorial explosion by brute force. The excision trick reduces the computational effort by a considerable degree.

The excision trick is based on the simple observation that given a complex K and an acyclic subcomplex L, the homology of K in positive dimension is isomorphic to that of the pair (K, L). But the d-dimensional relative chain group, $C_d(K, L)$, of (K, L) has as a basis all the d-simplices in $K \setminus L$. Thus, if most of the simplices in K lie in L, then the basis of $C_d(K, L)$ will contain fewer simplices (*many* fewer in our experience) than does that of the absolute chain group $C_d(K)$.

In order to implement this idea, one has to find an acyclic subcomplex, L. We employ a greedy algorithm that, using a highest dimensional simplex in K as a "seed", endeavors to sweep up additional simplices in decreasing order of dimension to form the acyclic subcomplex L. A collection of the largest simplices will contain the vast majority of the simplices in K. This allows one to largely bypass the combinatorial explosion.

A problem is that creating the acyclic subcomplex L can itself sometimes be very time consuming. However, this process is controlled by two parameters. One can adjust these parameters to make the process of assembling L less aggressive and therefore less time consuming. The price of doing this is that L is smaller and one ends up with more simplices to examine. For the fMRI data, however, the excision trick allows us to ignore the vast majority (typically over 90%, often close to 100%) of the simplices.

To use the excision trick for computing persistent homology for a filtered complex as in (3), one must generate a filtration of pairs:

$$(K_1, L_1) \supseteq (K_2, L_2) \supseteq \cdots \supseteq (K_n, L_n). \tag{5}$$

(Call the pairs (K_i, L_i) "frame pairs".) This means that the excision trick needs to be used in a coordinated way among the frames to preserve the filtration.

7.5. **Persistent homology.** Edelsbrunner and Harer [EH10, pp. 152–157] present an algorithm for computing persistent homology. It is unclear how to modify their algorithm to handle relative homology so that the "excision trick" can be used. In our method we compute persistent homology in two stages, but we expect that our method is reasonably efficient.

Let F be the number of the maximum frequency level. (For the fMRI data, in the time domain F was always 39.) Fix a dimension d. For $i = 1, \ldots, F$, let $\partial_{d+1\to d;i}$ be

the boundary matrix from $d+1$ to d in frequency level i. If it is important, we specify if $\partial_{d+1\to d;i}$ is reduced (as in "matrix reduction", not as in "reduced homology") or not. The first step is to compute (relative) homology for each frame pair (K_i, L_i) separately. We use a reduction algorithm (Munkres [Mun84, §11], Edelsbrunner and Harer [EH10, Section IV.2], Zomorodian [Zom05, Section 7.3.1]) to compute homology for each frame pair (K_i, L_i). Thus, we begin with matrices for the boundary operators. We then perform a reduction of the *columns*. We do *not* reduce the rows because, in the second stage, we will want, in each dimension, bases for the chain groups that are comparable across the frame pairs (K_i, L_i) in (5). As we perform the reduction we record all the column operations so we can express each column in the reduced matrix as a sum of simplices.

The second stage proceeds as follows.

```
Initialize α, ζ, and lifespan to be arbitrary lists of length F;
    For k = 1 to F;
        (*) Set ζ_k (k^th entry in ζ)
                = list of representative relative d-cycles for a basis
                of H_d(K_k, L_k; GF(2)) in frequency level k;
        Set lifespan_k to be a vector of 0's, one for each cycle in ζ_k;
        Set α_k to be ∂_{d+1→d;k} (reduced);
    End For;
    For i = F to 1 (reverse order);
        For j = i to 1 (reverse order);
            (**) Append columns corresponding to the relative cycles
                in ζ_i to the right side of α_j;
            Reduce α_j left to right. Replace α_j by the reduced matrix;
            For each column in α_j corresponding to a cycle in ζ_i
                that is not now a 0 column, increment the
                corresponding entry in lifespan_i by 1;
        End For;
    End For;
Return lifespan and ζ;
```

Step (*) is performed using matrix reduction. Notice that as the loop over i progresses the number of columns in any matrix in α is nondecreasing.

In dimension 0 we used non-reduced homology. The relation between absolute 0-dimensional homology 0-dimensional homology relative to an acyclic subcomplex is simple in a single complex, but is not so simple in persistent homology. Extra processing is needed to take into account the connected components containing the acyclic subcomplexes.

7.6. **Localization algorithm.** To perform localization at frequency level i and dimension d, we take each d-simplex, σ, in the complement $K_i \setminus L_i$ and, for each variable v not in σ, form the cycle $z := \partial v\sigma$ (where $v\sigma$ is the collection of variables including v and the variables in σ). Then we check to see if z is an absolute cycle *of the complex* K_i. This is done in two steps, first, we check that the d-simplicies in z not in L_i form a relative cycle.

This can be done by matrix reduction using $\partial_{d\to d-1;i}$, unreduced. Next, we check to see if the d-simplicies in z not in $K_i \setminus L_i$ all belong to L_i. If both these criteria are met then z is an absolute cycle of K_i.

However, it is also important to know which homology class z belongs to. This can be done by matrix reduction as follows. As in step (**) of the persistent homology algorithm, append columns corresponding to relative cycles repesenting a basis of $H_d(K_i, L_i)$ to the right side of $\partial_{d+1\to d;i}$ (reduced). Then on the right hand side of the resulting matrix, append a column representing z. Call the resulting matrix A. Next, reduce A left to right. Since z is a cycle, the last column of A, the one representing z, will be reduced to 0. By tracing the reduction process, one comes up with an expression for the class in $H_d(K_i, L_i)$ to which z belongs.

The fact we can begin with $\sigma \in K_i \setminus L_i$ greatly lessens the number of chains that need to be checked. Other simple steps also reduce the number of candidate chains. In our experience, we needed to check at most 10 or 20% of the $\binom{V}{d+2}$ possible chains (V = number of variables, i.e., regions), nearly always *much* fewer. At times we availed ourselves of the fact that this calculation can be run in parallel.

References

[Agr90] Alan Agresti, *Categorical data analysis*, Wiley, New York, 1990.

[CF08] Chao Chen and Daniel Freedman, *Quantifying homology classes*, Symposium on Theoretical Aspects of Computer Science (Bordeaux), 2008, pp. 169–180.

[CSZ09] Gunnar Carlsson, Gurjeet Singh, and Afra Zomorodian, *Computing multidimensional persistence*, Algorithms and Computation, 20th International Symposium, ISAAC 2009, Honolulu, Hawaii, USA, December 16-18, 2009. Proceedings (Yingfei Dong, Ding-Zhu Du, and Oscar H. Ibarra, eds.), Springer, 2009, pp. 730–739.

[DHK08] T. K. Dey, A. Hirani, and B. Krishnamoorthy, *Optimal homologous cycles, total unimodularity, and linear programming*, SIAM J. Computing **40** (2008), 1026–1044.

[EH10] Herbert Edelsbrunner and John L. Harer, *Computational Topology: An Introduction*, American Mathematical Society, Providence, 2010.

[EK13] Steven P. Ellis and Arno Klein, *Describing high-order statistical dependence using "concurrence topology", with application to functional MRI brain data*, (posted on arXiv, http://arxiv.org/abs/1212.1642), 2013.

[KMM04] Tomasz Kaczynski, Konstantin Mischaikow, and Marian Mrozek, *Computational Homology*, Springer, New York, 2004.

[Lau96] Steffen L. Lauritzen, *Graphical Models*, Oxford University Press, New York, 1996.

[MPZ08] Marian Mrozek, PawełPilarczyk, and Natalia Żelazna, *Homology algorithm based on acyclic subspace*, Computers & Mathematics with Applications **55** (2008), no. 11, 2395 – 2412.

[Mun84] J. R. Munkres, *Elements of Algebraic Topology*, Benjamin/Cummings, Menlo Park, CA, 1984, Reprinted by Perseus Publishing, Cambridge.

[R D08] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0.

[Ric08] David S. Richeson, *Euler's Gem: The Polyhedron Formula and the Birth of Topology*, Princeton University Press, Princeton, 2008.

[Sat99] H. Sato, *Algebraic Topology: An Intuitive Approach*, Translations of Mathematical Monographs, vol. 183, American Mathematical Society, Providence, RI, 1999.

[Sch59] Henry Scheffé, *The Analysis of Variance*, Wiley, New York, 1959.

[UKB+09] Lucina Q. Uddin, A.M. Clare Kelly, Bharat B. Biswal, F. Xavier Castellanos, and Michael P. Milham, *Functional connectivity of default mode network components: Correlation, anticorrelation, and causality*, Human Brain Mapping **30** (2009), 625 – 637.

[Zom05] Afra J. Zomorodian, *Topology for Computing*, Cambridge Monographs on Applied and Computational Mathematics, vol. 16, Cambridge, Cambridge, 2005.

Steven P. Ellis, Unit 42, New York State Psychiatric Institute at Columbia University, 1051 Riverside Dr., New York, NY 10032, U.S.A., e-mail: spe4@columbia.edu