"Concurrence Topology:" A New Method for Describing High-Order Statistical Dependence in Data

Steven P. Ellis (largely joint work with Arno Klein)

SoCG 7/21/'15

Abstract

Data analytic methods possessing the following three features are desirable: (1) The method describes "high-order dependence" among variables. (2) It does so with few preconceptions. And (3) it can handle at least dozens, maybe hundreds of variables. However, if approached in a naive fashion, data analysis having these three features triggers a "combinatorial explosion": The output from the analysis can include thousands, maybe millions of numbers. Few methods exist possessing all three features yet which avoid the combinatorial explosion. An apparently new methd, "Concurrence Topology (CT)", often succeeds in doing this.
CT takes the usual approach in "topological data analysis", viz., it translates the data into a series of shapes, a "filtration". Except CT does this, not for point clouds, but for binary data. The vertices correspond to variables in the data.
A persistent homology class of the filtration indicates a mutually inhibitory negative relationship among the variables. Normally, there are no more than a few dozen homology classes, so CT avoids the combinatorial explosion. Often one can identify small

- Free R code exists that implements (most of) the procedures described in this talk.
- Reference: S.P. Ellis, A. Klein (2014) "Describing high-order statistical dependence using 'concurrence topology', with application to functional MRI brain data," *Homology, Homotopy, and Applications*, 16, 245–264.

CONCERNED WITH DATA ANALYSIS CHARACTERIZED BY THREE FEATURES

FEATURE 1: HIGH-ORDER DEPENDENCE

- A feature of a data set that is visible when looking at k variables at a time but which is invisible when looking at $k-1$ variables at a time reflects "k^{th}-order dependence" in the data.
  - Example: (Pearson, Spearman, Kendall) correlation is a second order feature of data.
- "High-order dependence" means dependence of order at least 3.

EXAMPLE:

I			II			III		
x	y	z	x	y	z	x	y	z
0	0	0	0	0	0	0	0	1
0	1	1	0	0	1	0	1	0
1	0	1	0	1	0	1	0	0
1	1	0	1	0	0	1	1	1
0	0	0	0	1	1	0	0	1
0	1	1	1	0	1	0	1	0
1	0	1	1	1	0	1	0	0
1	1	0	1	1	1	1	1	1

- Three data sets identical up to second order, but different at third order.

FEATURE 2: "AGNOSTIC" DATA ANALYSIS

- Typically, formulating a conventional statistical model involves choices.
  - Which variable should be the response (dependent) variable?
  - Which variables should be the predictors (independent variables)?
  - Which variables should be included in interactions?
- If you have prior knowledge to guide you, conventional statistical modeling is a powerful way to learn from data.
- A more data-driven approach is *"agnostic data analysis":*
  - For every $k = 1, 2, \ldots$, treat all groups of k variables the same *a priori.*

FEATURE 3: "LARGE" NUMBER OF VARIABLES

- In this talk "large number" means "dozens", maybe a hundred or so.

"COMBINATORIAL EXPLOSION"

- The three features constitute an "explosive mixture".
- Agnostically describing k^{th}-order dependence in a data set means examining *all* combinations of k variables at a time.
- If there are many variables and $k > 2$, the number of combinations can be huge.

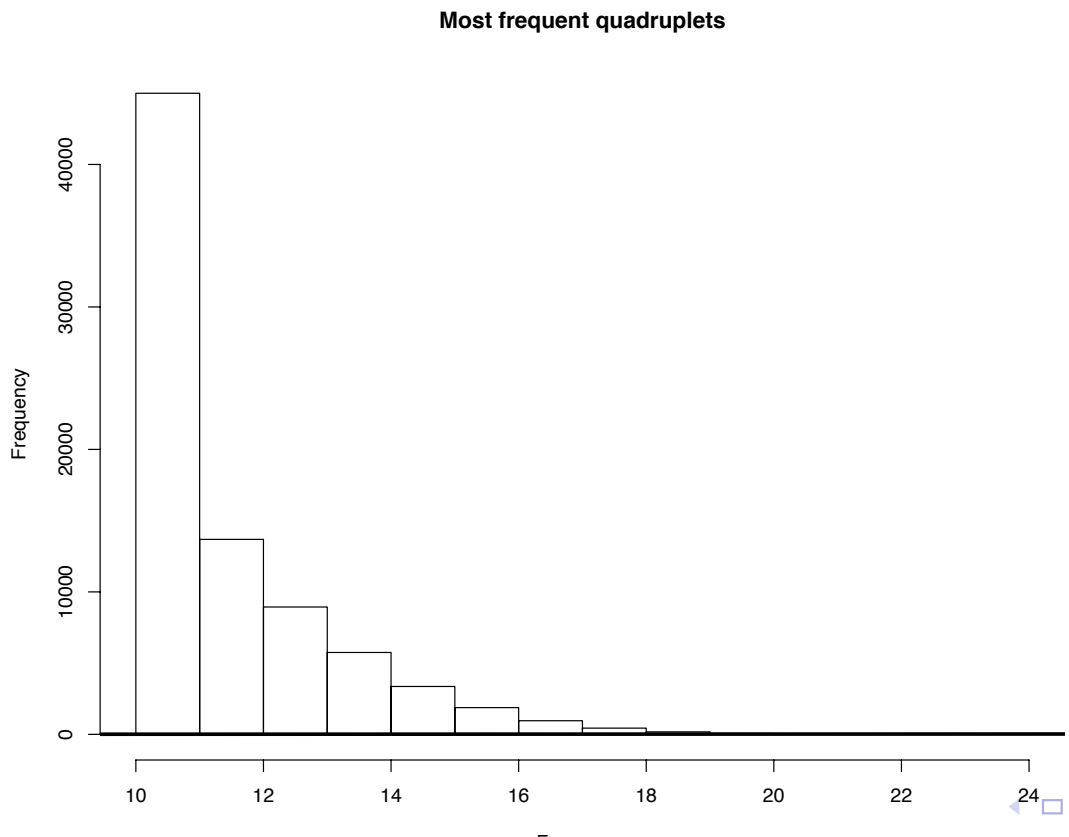
"COMBINATORIAL EXPLOSION:" EXAMPLE

Analysis of seventh-order dependence among the regions of the brain "default mode network" in an fMRI data set.

- 32 variables.
- Naive agnostic seventh order analysis of 32 variables means looking at $\binom{32}{7} = 3,365,856$ combinations of 7 variables.
- Data contained only 6,144 numbers.
- It is difficult to make sense of so many numbers.

"COMBINATORIAL EXPLOSION," EXAMPLE

- Look at fourth order dependence in one subject's dichotomized BOLD in 74 ROI's.
- $\binom{74}{4} = 1,150,626$.
- Of these $1,013,129$ appear in data.
- Histogram of frequencies of the $80,271$ that appear most often.



SOME METHODS THAT AGNOSTICALLY CAPTURE HIGH ORDER DEPENDENCE IN MANY VARIABLES

"Unsupervised" methods:

- There seem to be few established unsupervised methods that capture high order dependence.
- Independent Components Analysis
- Tensor based methods:
  - "Parallel factor analysis"
  - "Tucker 3"
  - Only go up to third order dependence?

“Supervised” methods

- Many machine learning methods tap into high order dependence.

Experimental methods.

CONCURRENCE TOPOLOGY (CT)

- Apparently new "unsupervised" method for high-order agnostic analysis of dependence among dozens (a hundred?) of variables.
- CT is radically different from methods mentioned above.
- I will not compare CT to those other methods.
- Since there are few methods there's no need to choose among them: "Use all of them."
- So comparing methods to see which is best is not urgent.

“CURTO-ITSKOV”

- The germ of the idea for CT came from a theoretical neuroscience talk I heard by the mathematician Carina Curto.

CONCURRENCE TOPOLOGY (CT), continued

- (So far!) CT only handles dichotomous, i.e., binary, variables.
  - Quantitative or ordinal data can be binarized by thresholding.
  - Nominal data can be binarized by collapsing categories.
- CT is often able to extract a moderate number of high order features from a combinatorial explosion.
- CT detects certain forms of *negative or weak association* among the variables.
- CT uses topology.

"TOPOLOGICAL DATA ANALYSIS"

- Given: Noisy high-dimensional data clustered near a manifold.
  - (How common are such data sets?)
- Objective: Use data to learn about the topology of the manifold.
- Or in TDA one examines the topology of super/sub level sets of a function.
- CT is not topological data analysis?

CONCURRENCE TOPOLOGY, continued

- In CT the joint distribution of a group of *binary*, i.e. "0-1", variables is represented as a filtered simplicial complex ("filtration").
  - Filtered simplicial complex is a monotonic series of simplicial complexes ("frames").
- *In CT one studies the joint distribution of variables via the persistent homology of the filtration.*
  - Describe how holes penetrate multiple frames (or not).

FIRST STEP: COMPUTE "CONCURRENCES"

- "Observation:" A vector of 0's and 1's.
  - The components of the vector are the values of the variables.
- The "concurrence" for an observation is the list of the variables that are "1" in that observation.
- Result is "concurrence complex".

EXAMPLE: TOY DATA SETS

I			II			III		
1	1	0	1	1	1	1	1	1
x	y	z	x	y	z	x	y	z
0	0	0	0	0	0	0	0	1
0	1	1	0	0	1	0	1	0
1	0	1	0	1	0	1	0	0
1	1	0	1	0	0	1	1	1
0	0	0	0	1	1	0	0	1
0	1	1	1	0	1	0	1	0
1	0	1	1	1	0	1	0	0
1	1	0	1	1	1	1	1	1

- Concurrences for data set I: (ignore row of all 0's), yz, xz, xy, yz, xz, xy.
- Concurrences for data set II: z, y, x, yz, xz, xy, xyz.
- Concurrences for data set III: $z, y, x, xyz, z, y, x, xyz$.

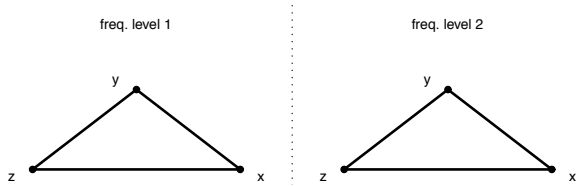
STEP 2: FILTER CONCURRENCE LIST

- For each concurrence, C, in the list, count how many times it appears as a subset (proper or not) of concurrences in the list.
- The "frequency" of C.
  - 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 $\geq f$.
  - We call f the (absolute) "frequency level" of $\mathcal{C}_f$.
- Call the collection $\mathcal{C}_1 \supset \mathcal{C}_2 \supset \mathcal{C}_3 \supset \cdots$ the "filtered concurrence list" of the data.

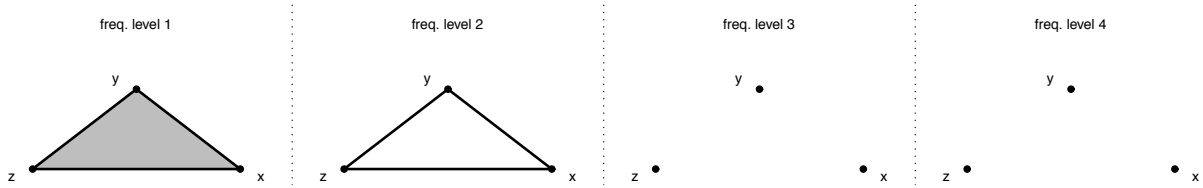
STEP 3: CONSTRUCT FILTERED "CURTO-ITSKOV COMPLEX"

- Interpret each concurrence $C \in \mathcal{C}_f$ as an abstract simplex, σ_C.
- Complex K_f is abstract simplicial closure of $\{\sigma_C : C \in \mathcal{C}_f\}$.
- $K_1 \supset K_2 \supset K_3 \supset \cdots$ is "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 other cases fractional frequencies are needed.
- Note that this is a *descending* filtration.
  - This is consequence of interpretation of the index f as frequency.
  - This reversed indexing only changes persistence analysis in trivial ways.

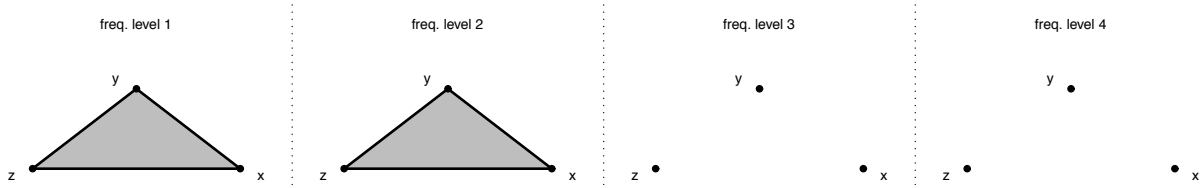
data set I, λ_{111}^{xyz}= –0.80



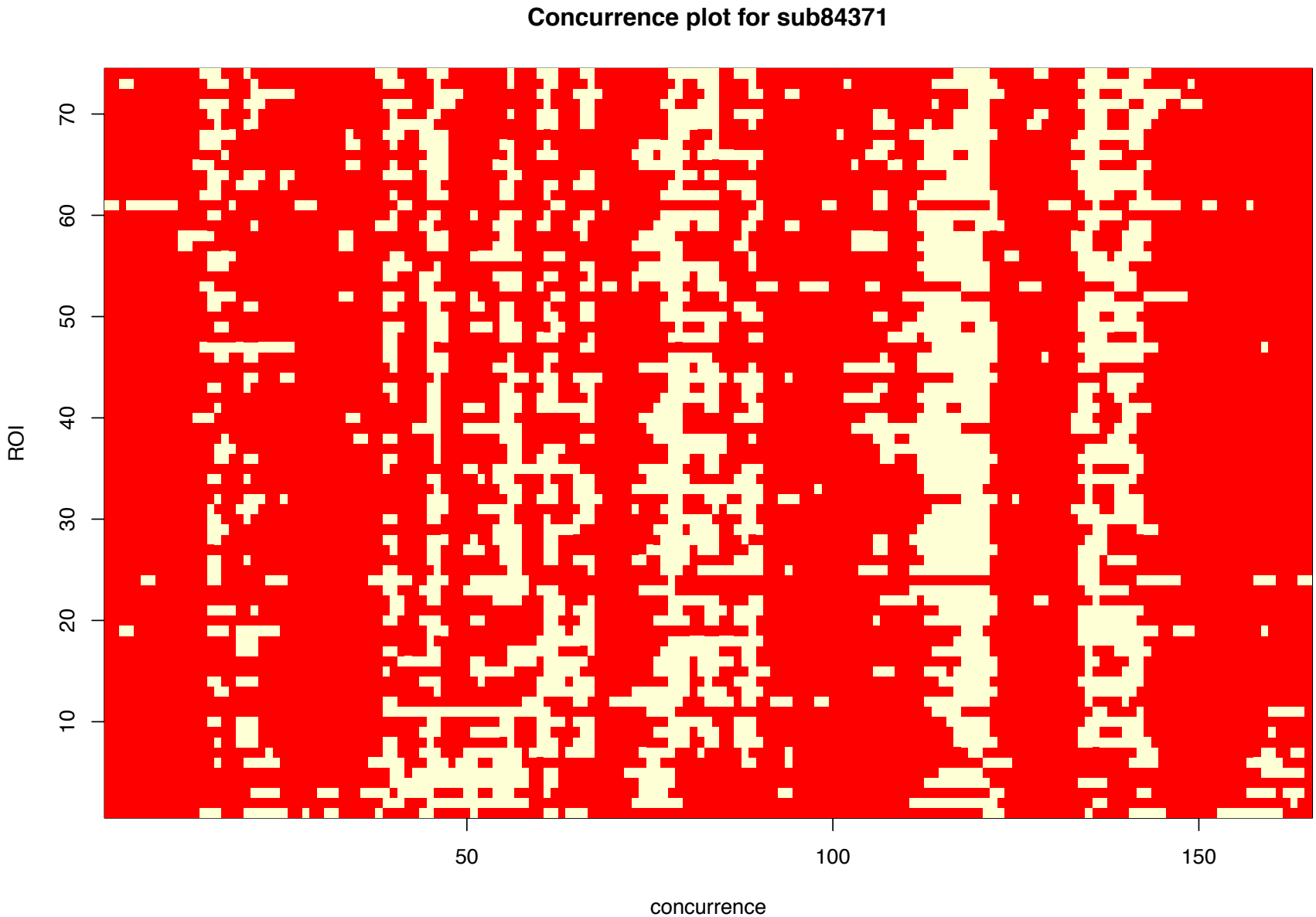
data set II, λ_{111}^{xyz}= 0.0



data set III, λ_{111}^{xyz}= 0.80



CONCURRENCE: EXAMPLE



HOLES

- Holes (homology classes) come in different "dimensions" : 0, 1, 2, ...
- Moving in *decreasing* order of frequency, we may find that a persistent homology class of a given dimension appears in a frame.
- We say the homology class is "born" at that frequency level.
- The same homology class may also be present in frames of lower frequency level.
- We say the homology class "dies" at the first frequency level (moving downward) at which the homology class is absent.
- The difference *birth* − *death* is the "lifespan" or "persistence" of the homology class.

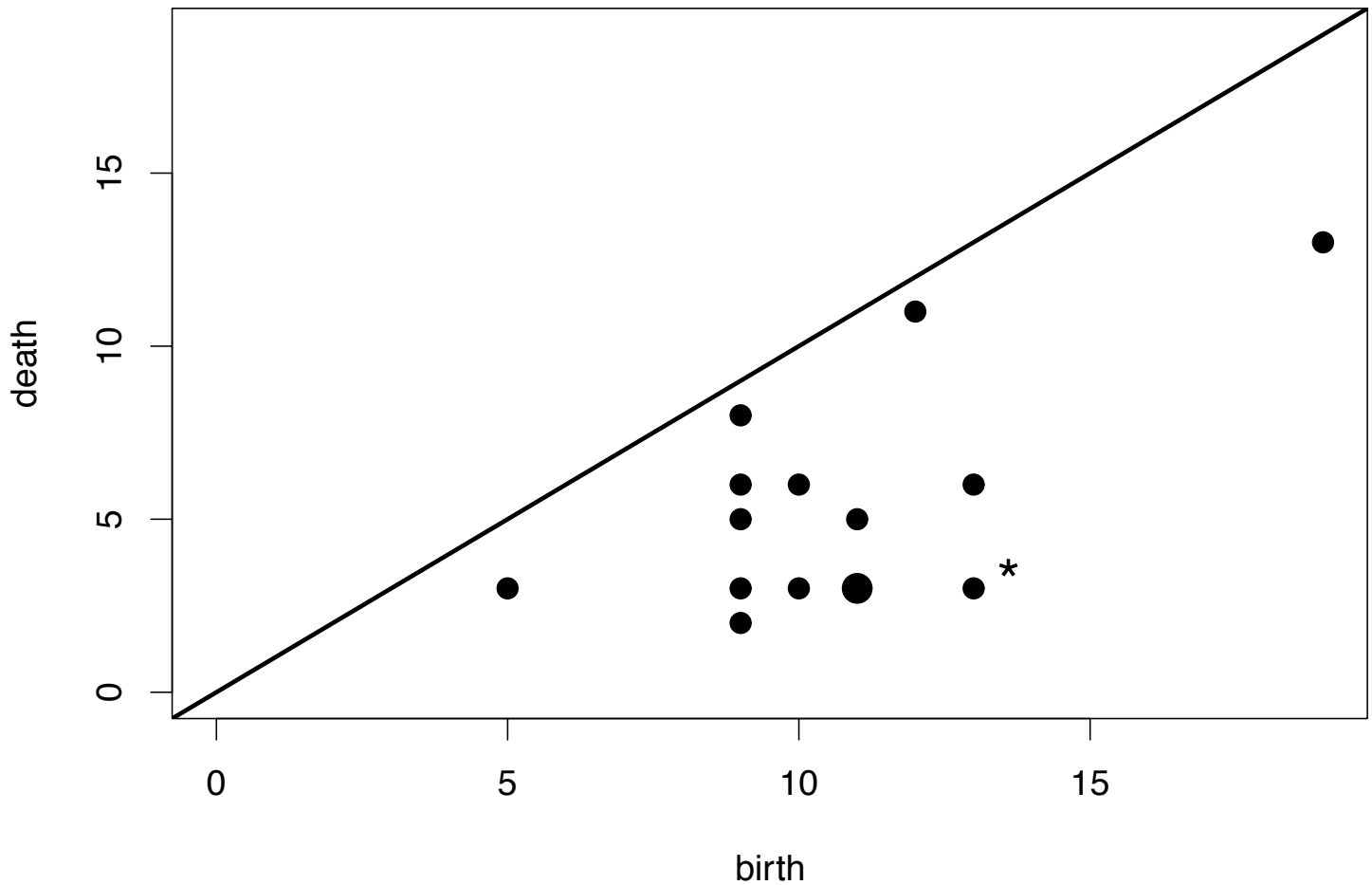
MEANING OF HOMOLOGY CLASSES IN CT

- In CT topology one learns about data by studying the homology classes in the data's filtration and their lifespans.
- Homology classes in the filtration indicate relative *weakness or negativity in joint distribution of variables.*
- The longer the lifespan of a homology class in a data set's filtration, the more likely it is that the homology class represents signal, not noise.

"PERSISTENCE PLOTS"

- Plotting death vs. birth yields a "persistence plot".
- The lifespan of a homology class equals the distance it's point in the persistence plot is below the $birth = death$ diagonal.

PERSISTENCE PLOT FOR ONE SUBJECT IN DIMENSION 1



SPARSITY

- The number of persistent homology classes in a filtered Curto-Itskov complex tends to be small, a few dozen at the most.
- CT therefore avoids the combinatorial explosion.

fMRI

- “fMRI” stands for “functional Magnetic Resonance Imaging”.
  - Numerical values generated by fMRI is called "BOLD".
- It images the *functioning* of the brain of a living person over time.

DATA SET

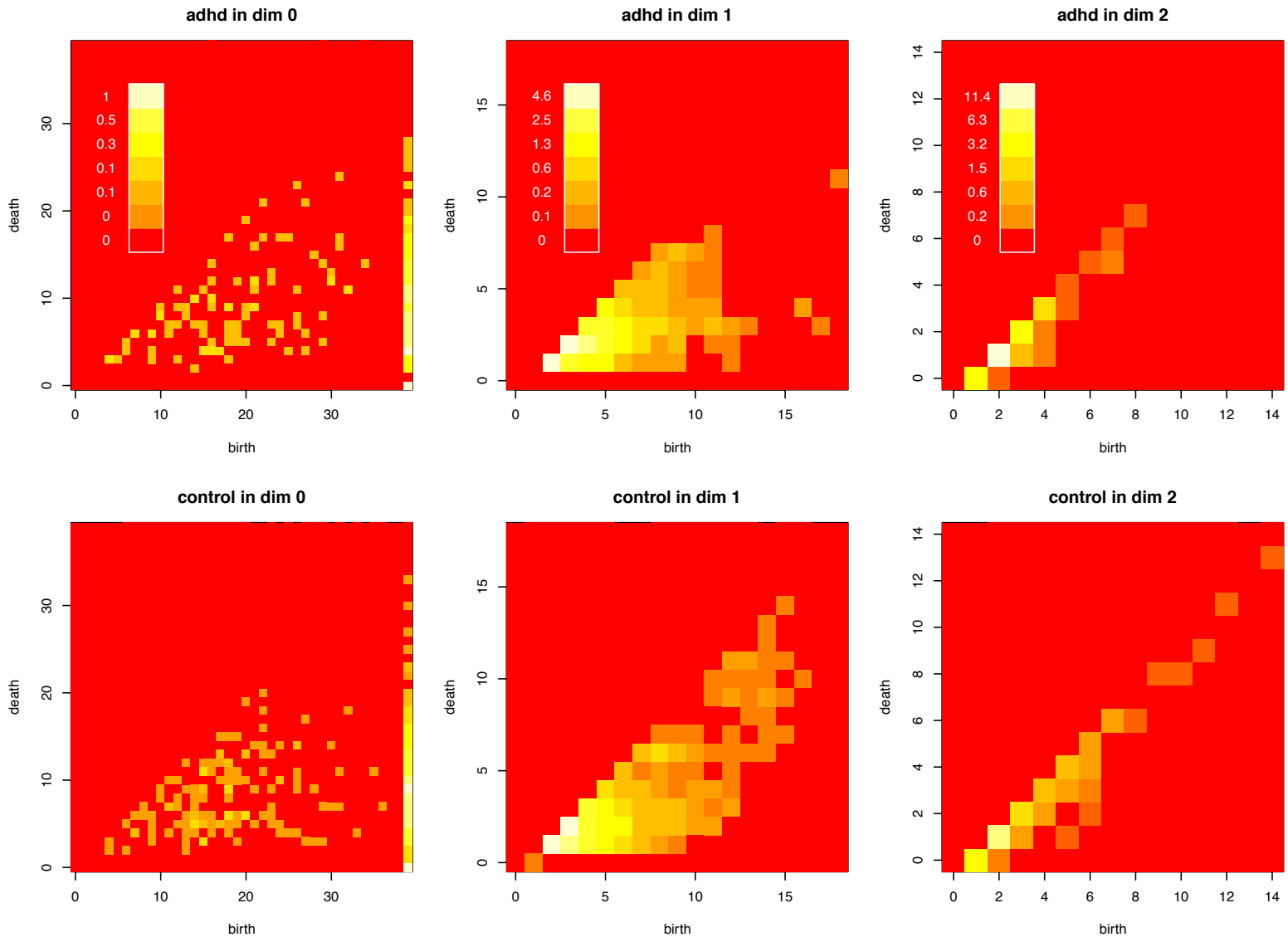
- Publicly available fMRI data.
  - 25 subjects with Attention Deficit Hyperactivity Disorder (ADHD).
  - 41 healthy controls.
  - Data include BOLD values for each subject in 92 brain regions at 192 time points.

WE APPLIED CT TO EACH SUBJECT'S fMRI DATA *SEPARATELY*.

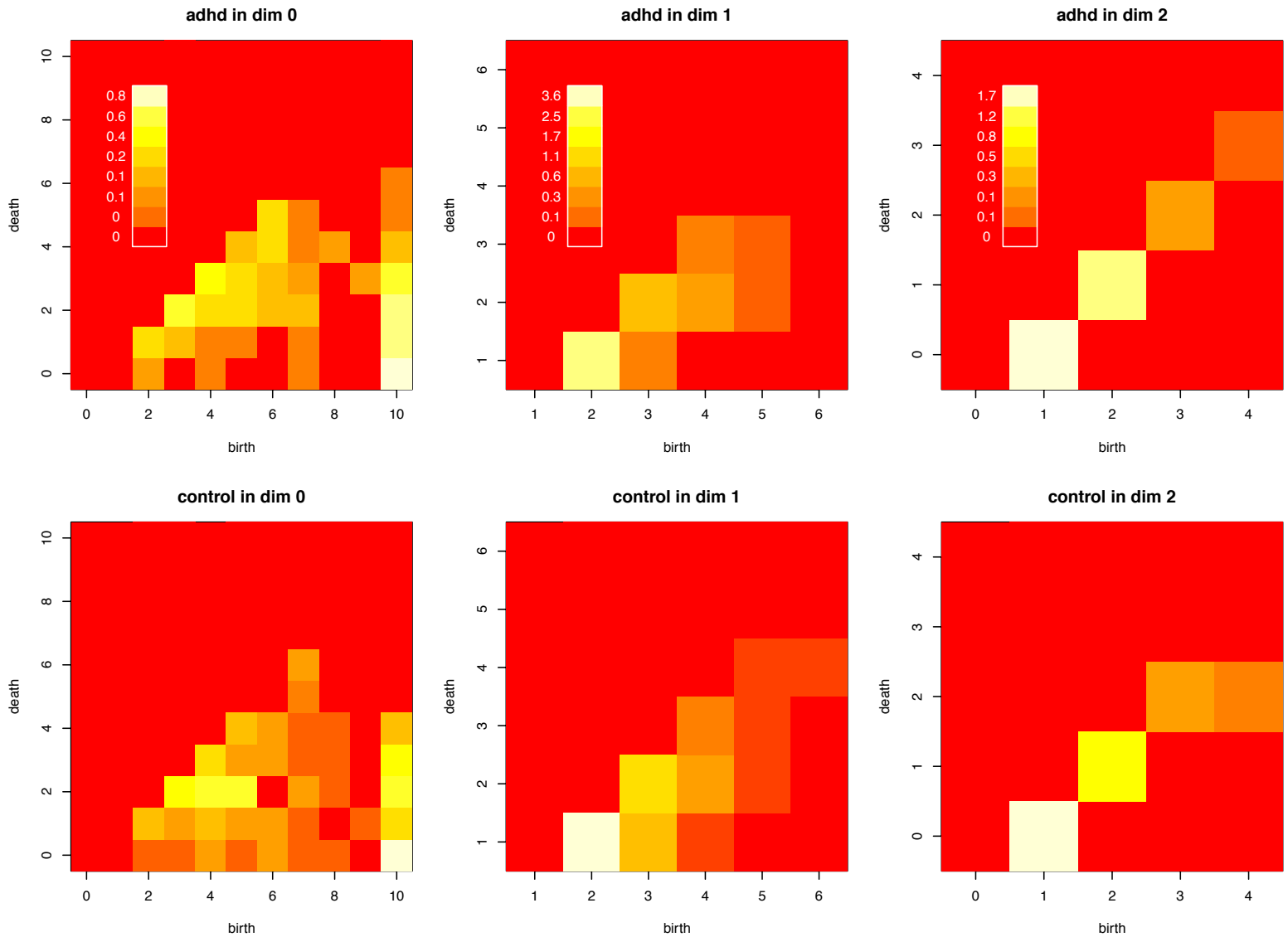
- Looked at persistent homology up to dimension 5.
  - Pertains to dependence (connectivity) of orders of *seven* or more.
  - Like fitting a LS regression model with one or more sixth-order interactions.

AVERAGE PERSISTENCE PLOTS

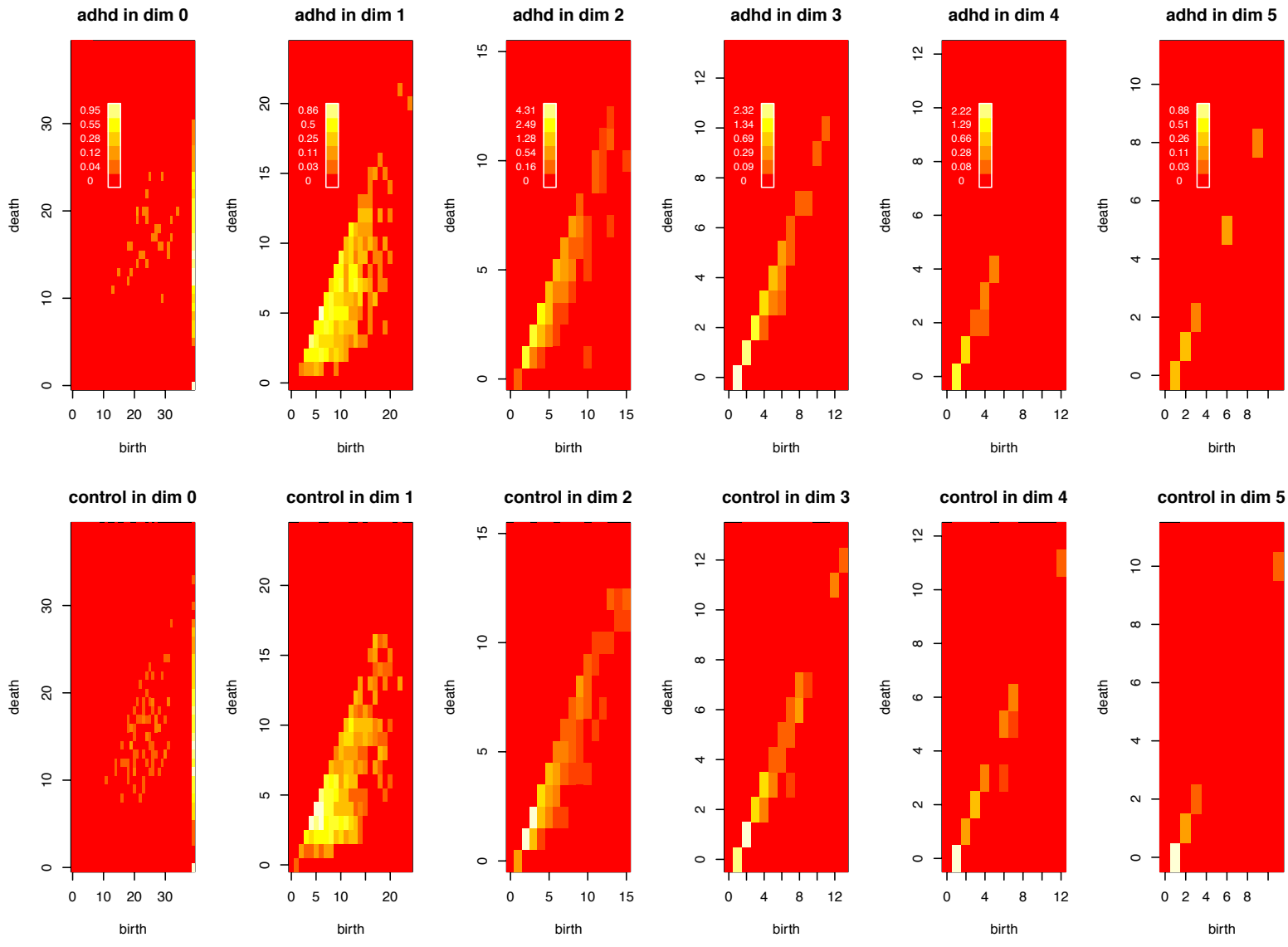
Whole Brain, Time Domain



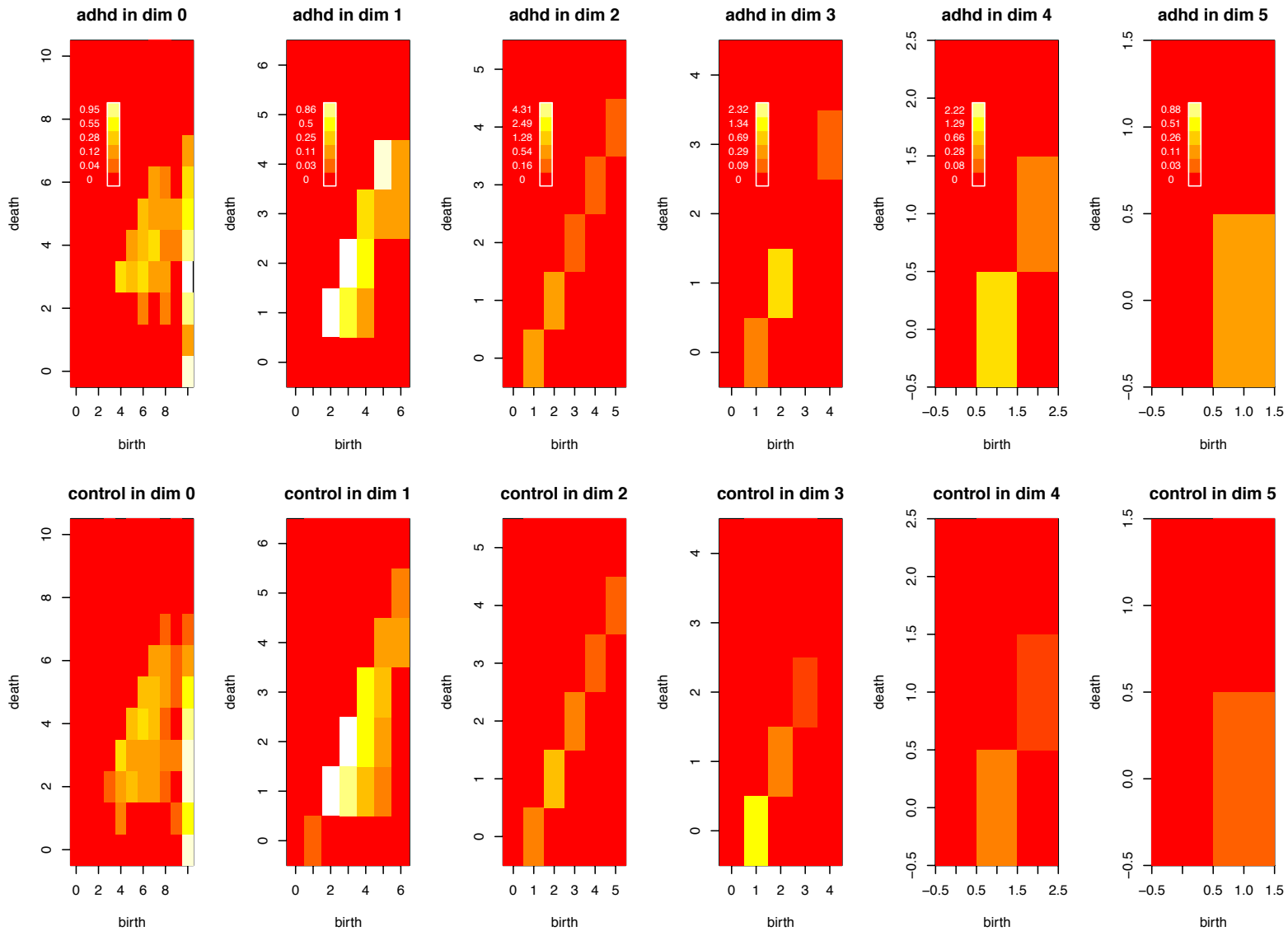
Whole Brain, Fourier Domain



Default Mode Network, Time Domain



Default Mode Network, Fourier Domain



- We found a number of differences between the ADHD and control groups based on persistence plots.
- Including a difference in dimension 4, which corresponds to dependence of order 6.

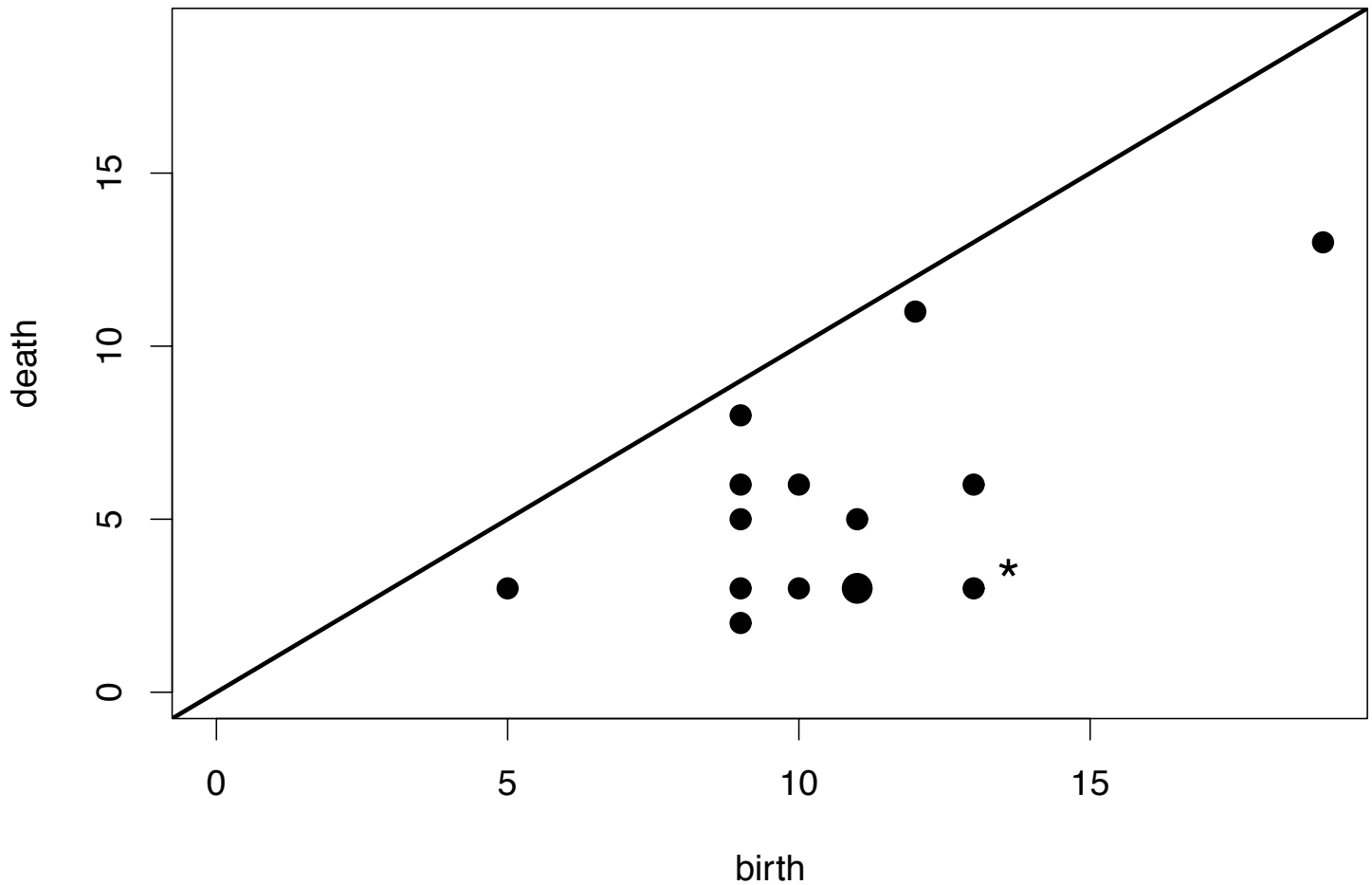
BEYOND PERSISTENCE PLOTS: "LOCALIZATION"

- Homology classes involve all the variables.
- Homology classes can often be "localized" by identifying groups of variables ("short cycles") most closely associated with them.
- Short cycles can be examined to see if they're interesting.

LOCALIZATION: EXAMPLE

- In fMRI data we found interesting short cycles in dimension 1.
  - 3rd-order dependence.
  - "Short cycle" consists of a triple of regions.
  - Each subject has a few hundred short cycles.
  - One subject had 164 short 1-cycles in a single homology class in a single frame.
- Combinatorial explosion: 9,880 different possible triplets of regions.

PERSISTENCE PLOT FOR ONE SUBJECT IN DIMENSION 1, TIME DOMAIN, DMN



SPECIAL SHORT CYCLES

- One short cycle in dimension 1 is found in 13 subjects.
  - This is statistically significant.
- 16 short cycles associated with the same class distinguish ADHD from controls.
  - Most subjects have at least one of these 16 short cycles.
  - More common in ADHD.

(CURRENT WORK) BEYOND PERSISTENCE PLOTS: PRODUCTS

- Suppose one has two groups of variables, $\ldots, U, V, W, \ldots$ and $\ldots, u, v, w, \ldots$.
- Suppose CT analysis of $\ldots, U, V, W, \ldots$ turns up a non-trivial homology class α and CT analysis of $\ldots, u, v, w, \ldots$ turns up a non-trivial homology class β.
  - So there is negative dependence among $\ldots, U, V, W, \ldots$ and among $\ldots, u, v, w, \ldots$.
- Suppose $\ldots, U, V, W, \ldots$ and $\ldots, u, v, w, \ldots$ are *independent of each other as groups* of variables.

HOMOLOGY CROSS PRODUCT

- *Then* in fairly large data sets the homology cross product $\alpha \times \beta$ will be tend to be non-trivial in the product space.
- This provides a way of looking for independence between groups of variables.
- Using this method I have found interesting things in real data.