Algorithmic Homology

Graham Ellis

Preface

This book is an introduction to the homology theory of topological spaces and discrete groups, focusing on algorithmic aspects of the subject. It is divided into two parts. Part I consists of the first three chapters and is a short introduction to simplicial, cubical, cellular and singular homology. It includes computer applications to data analysis, and finishes with a computer application of the Lefschetz fixed point theorem to dynamical systems. Part II is concerned with the homology of classifying spaces of discrete groups. Some of its chapters deal with low-dimensional homology, and introduce computational machinary based on presentations of groups by generators and relators. Other chapters deal with higher dimensional group homology and derive computations from explicit regular CW-structures on classifying spaces.

Part I is based on a masters level course of 20 lectures given at NUI Galway in 2010. Part II aims to describe some of the mathematical techniques implemented as part of the homological algebra package HAP [5] for the computer algebra system GAP [6].

Implementations of the algorithms described in this book can be found in GAP. These are used in examples, and links are provided to GAP readable versions of the examples.

Graham Ellis
March 2010

Part I

Contents

1 Betti numbers and applications **5**

1.1 Introduction 6

1.2 Euler-Poincaré characteristic 8

1.3 Digital image analysis 25

1.4 Statistical data analysis 42

1.5 Betti numbers 58

1.6 Persistence of betti numbers 74

1.7 Natural image statistics 88

2 Categories, functors and homology **97**

2.1 Introduction 98

2.2 Definitions of categories and functors 101

3 Homology computation **104**

Chapter 1

Betti numbers and applications

1.1 Introduction

This first chapter is an informal introduction to some of the basic notions of topology, to algorithms for computing them, and to potential applications. The reader will occasionally need to rely on intuition for understanding notions (such as "boundary", "orientation", ...) whose formal definitions are deferred to Chapter 2.

Topology is the study of those properties of an object that remain unchanged throughout continuous deformations of the object. Any user of the London Underground is familiar with this idea. The standard map of the Underground

is a continuous deformation of a more precise geographical map. It retains topological properties necessary for route planning, but discards geometric properties involving distance and angle.

One of the earliest contributors to the literature on topologiy was Leonhard Euler who, in 1736,

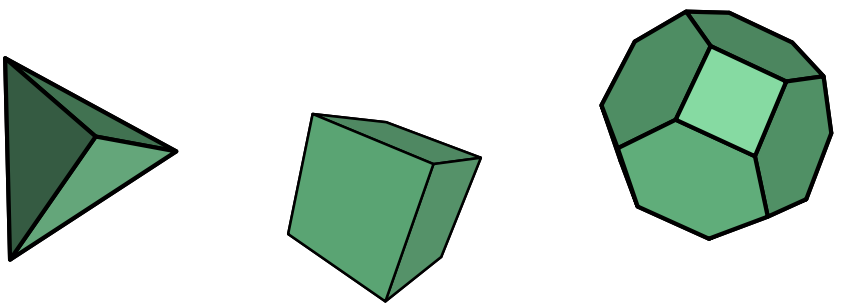
published a paper entitled *Solutio problematis ad geometriam situs pertinentis*. It uses graph theoretic invariants for studying the now famous problem of planning a route through Königsberg that crosses each of the city's bridges precisely once. The use of algebraic invariants in topology was pioneered by Henri Poincaré in a paper of 1885 entitled *Analysis Situs*, and in his five *Compléments à l'Analysis Situs* published between 1889 and 1905. The following extract from Poincaré's 1885 paper explains his motivation.

> Toutes les voies diverses où je m'étais engagé successivement me conduisaient à *l'Analysis Situs*. J'avais besoin des données de cette Science pour poursuivre mes études sur les courbes définis par les équations differentielles et pour les étendre aux equations differentielles d'ordre supérieur, et, en particulier, à celles du problème des trois corps. J'en avais besoin pour l'étude des fonctions non uniformes de deux variables. J'en avais besoin pour l'étude des intégrales multiples et pour l'application de cette étude au développment de la fonction perturbatrice. Enfin, j'entrevoyais dans *l'Analysis Situs* un moyen d'aborder un probléme important de théorie des groupes, la recherche des groups discrets ou des groupes finis contenus dans un groupe continu donné.

A natural starting point for any account of homology is the relationship between the number vertices, edges and faces of a convex solid which was spotted by Euler and generalized by Poincaré in *l'Analysis Situs*.

1.2 Euler-Poincaré characteristic

For any 3-dimensional convex solids such as the simplex, the cube and the permutahedron

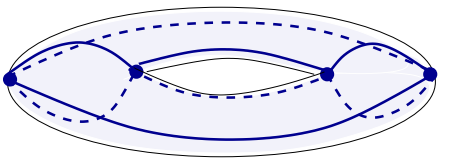


we can count the number V of vertices, the number E of edges and the number F of faces. For the three examples we find:

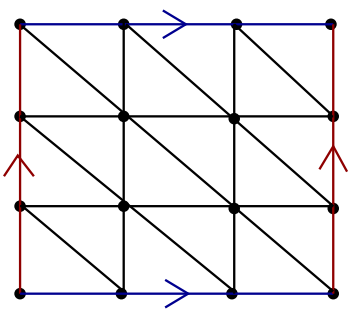
	V	E	F	V-E+F
simplex	4	6	4	2
cube	8	12	6	2
permutahedron	24	36	14	2

Euler mentioned the formula $V - E + F = 2$ in a letter to Christian Goldbach (1750). He published

the fomrula in two articles in 1752 and claimed that $V - E + F$ is equal to 2 when the surface of any 3-dimensional body is decomposed into a finite number of regions or *faces* by a graph involving a finite number of *vertices* and *edges*. Such a decomposition is called a *finite cell structure*. Some careful hypothesis is needed for Euler's claim since, for instance, the surface of a donought can be given a cell structure involving four curvilinear rectangles



yielding $V - E + F = 4 - 8 + 4 = 0$. The surface of a donought can also be given a cell structure involving 18 curvilinear triangles



again yielding $V - E + F = 9 - 27 + 18 = 0$. In our illustration of this triangulation, the two blue lines need to be identified, as do the two brown lines.

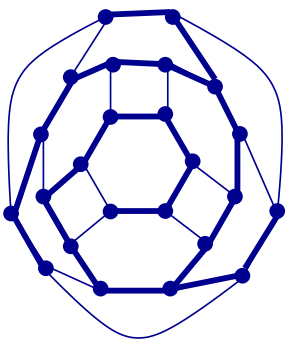
In any finite cell structure we assume that: the interiors of edges are homeomorphic to the open unit interval $U^1 = \{x \in \mathbb{R} : 0 < x < 1\}$; the interiors of faces are homeomorphic to the open unit 2-disc $U^2 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$; edges intersect only at vertices; and the boundary of any face is a union of vertices and edges.

Theorem 1 *The formula*

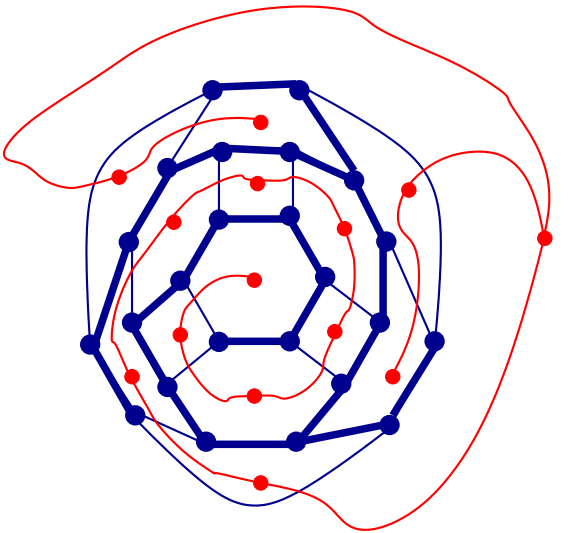
$$V - E + F = 2$$

holds for any finite cell structure on the 2*-sphere* $S^2 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$.

Proof. The vertices and edges form a connected planar graph Γ in which we can choose some maximal subtree T. It is possible that T contains no edges, in which case $V-E+F = 1-0+1 = 2$. So suppose that T contains some edge. The graph and a maximal subtree are illustrated for the permutahedron, with edges in the tree being shown thicker.



The dual graph Γ^* has one vertex for each face of the cellular decomposition. Two vertices of Γ^* are connected by an edge if the corresponding faces of Γ have a common edge. The edges in $\Gamma \setminus T$ correspond to a maximal tree T^* in Γ^*. This tree T^* is shown in red for the permutahedron.



The number F of faces in the cellular decomposition equals the number of vertices in the tree T^*. Thus F is one more than the number of edges in T^*. Therefore $F = 1 + E - E_T$ where E_T denotes the number of edges in T. Consequently $V - E + F = V - (E - E_T) - E_T + F = V - (F - 1) - E_T + F = V - E_T + 1 = 2$. □

We now introduce the language of convex polytopes and discuss a generaliation of Theorem 1 to this setting.

Recall that a subset $X \subset \mathbb{R}^n$ is *convex* if, for any pair of points x, y in X, the line segment joining x to y lies entirely in X. The *convex hull* of a finite set of points $S = \{v_1, v_2, \ldots v_p\}$ in $\mathbb{R}^n$,

$$\text{Conv}(S) = \{ \sum_{1 \le i \le p} t_i v_i \in \mathbb{R}^n : t_i \ge 0 \ \sum_{1 \le i \le p} t_i = 1\},$$

is the smallest convex subset of $\mathbb{R}^n$ containing all points of S. For any vector $w \in \mathbb{R}^n$ and any real number c the *half space*

$$H(w, c) = \{x \in \mathbb{R}^n : \langle x, w \rangle \ge c\}$$

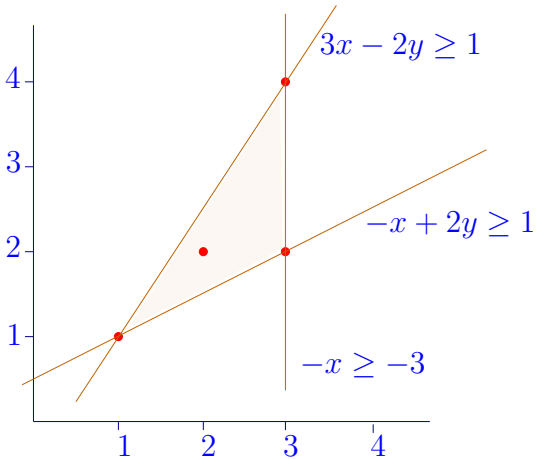
is a convex set (where $\langle x, w \rangle$ denotes the standard inner product). The term *convex polytope* is a synonym for the convex hull of a finite set of points. A basic theorem of convex polytopes is that for any finite set $S \subset \mathbb{R}^n$ we can find a finite collection of half spaces $H(w_1, c_1)$, $H(w_2, c_2)$, $\ldots$, $H(w_F, c_F)$ such that

$$\text{Conv}(S) = \bigcap_{1 \le i \le F} H(w_i, c_i) \ . \qquad (*)$$

To illustrate (*) consider the set of four points $S = \{v_1 = (1, 1), v_2 = (2, 2), v_3 = (3, 2), v_4 = (3, 4)\}$ in $\mathbb{R}^2$. Consider the three half spaces $H_1 = H((-1, 2), 1)$, $H_2 = H((-1, 0), -3)$, $H_3 = H((3, -2), 1)$. The equality

$$\text{Conv}(S) = H_1 \cap H_2 \cap H_3$$

is illustrated in the following picture.



We can also use this example to illustrate a proof of (*). In order to compute the half spaces whose intersection equals $P = \mathrm{Conv}(S)$, first observe that the convex hull P is clearly a projection onto the first two coordinates of the subspace $K \subset \mathbb{R}^6$ consisting of those vectors $(x, y, t_1, t_2, t_3.t_4)$

satisfying the following ten inequalities.

$$\begin{pmatrix} -1 & 0 & 1 & 2 & 3 & 3 \\ 1 & 0 & -1 & -2 & -3 & -3 \\ 0 & -1 & 1 & 2 & 2 & 4 \\ 0 & 1 & -1 & -2 & -2 & -4 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad (\dagger)$$

We need to explain how the variables t_1, t_2, t_3, t_4 can be "eliminated" from these ten inequalities to yield a description of $P \subset \mathbb{R}^2$ as the subset of those pairs $(x, y) \in \mathbb{R}^2$ satisfying

$$\begin{pmatrix} -1 & 2 \\ -1 & 0 \\ 3 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \geq \begin{pmatrix} 1 \\ -3 \\ 1 \end{pmatrix}. \qquad (\dagger\dagger)$$

Let us begin with the elimination of t_4. Note that, of the ten inequalities (†), four are of the form

$$t_4 \geq L(x, y, t_1, t_2, t_3)$$

where the L is a linear expression. Three are of the form

$$t_4 \leq U(x, y, t_1, t_2, t_3)$$

where U is a linear expression. The remaining three inequalities don't involve t_4. The projection $\pi\colon \mathbb{R}^6 \to \mathbb{R}^5$ onto the first five coordinates thus sends K to the subspace $\pi(K) \subset \mathbb{R}^5$ consisting of those vectors (x, y, t_1, t_2, t_3) satisfying the three inequalities $t_1 \geq 0$, $t_2 \geq 0$, $t_3 \geq 0$ and $4 \times 3 = 12$ inequalities of the form

$$L(x, y, t_1, t_2, t_3) \leq U(x, y, t_1, t_2, t_3).$$

A number of these 15 inequalities defining $\pi(K)$ may be redundant and we can remove them (see Exercise 4) to yield an irredundant set of inequalities defining $\pi(K)$. The same elimination technique (due to Fourier and rediscovered by Motzkin) can then be applied in turn to t_3, t_2 and t_1 to yield the inequalities (††) describing P as an intersection of half spaces.

The computation of half space inequalities for the convex hull of a finite set of points can be performed symbolically on a computer. The naive algorithm based on the above discussion suffices for computing the convex hull of a small number of points in low dimensional euclidean space. There are many possible refinements to this basic algorithm. The POLYMAKE [9] computer package for convex polyhedra incorporates a range of convex hull algorithms, each optimized for specific situations. In order to use POLYMAKE to compute the half space description of the above example the user first needs to create a file containing the following information.

```
POINTS
1 1 1
1 2 2
1 3 2
1 3 4
```

If the file is saved as `file1.txt` the user can then run the following command at the prompt.

```
$ polymake file1.txt FACETS
FACETS
-1 -1  2
 3 -1  0
-1  3 -2
```

Each row of output (a_0, a_1, a_2) corresponds to an inequality $a_0 + a_1 x + a_2 y \geq 0$ that defines a half space.

Returning to the theme of Theorem 1, let $P = \mathrm{Conv}(S)$ be the convex hull of a finite set $S \subset \mathbb{R}^n$, and let $H(w_1, c_1), \ldots, H(w_F, c_F)$ denote the minimal set of half spaces whose intersection is P. Each half space defines a *hyperplane*

$$\overline{H}(w_i, c_i) = \{x \in \mathbb{R}^n : \langle x, w_i \rangle = c_i\} \ .$$

A *face* of P is any non-empty intersection

$$P \cap \bigcap_{i \in T} \overline{H}(w_i, c_i)$$

with $T \subset \{1, \ldots, F\}$. The polytope itself is deemed to be a face corresponding to $T = \emptyset$. We say that a face has *dimension* d if $d+1$ is the maximal size of a linearly independent subset of the face.

The *Euler-Poincaré* characteristic of an n-dimensional polytope P is the sum

$$\chi(P) = \alpha_0 - \alpha_1 + \alpha_2 \cdots + (-1)^{n+1}\alpha_n$$

where α_d is the number of d-dimensional faces of P.

The numbers $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ for a convex hull $P = \text{Conv}(S)$ can be computed using POLYMAKE by storing S in a file, say `file2.txt`, and performing the following command at the prompt.

```
$ polymake file2.txt F_VECTOR
F_VECTOR
24 36 14
```

The Euler-Poincaré characteristic in this example is $\chi(P) = 24-36+14-1 = 1$ (and the convex hull is the permutahedron). A little experimentation with POLYMAKE would suggest that $\chi(P) = 1$ for any convex hull. Theorem 1 proves this for 3-dimensional polytopes. Poincaré proved the general result in *l'Analysis Situs*. In fact he proved a stronger result about "cellular spaces" which need not be convex.

A modern notion of cellular space is that of a *CW-space* introduced by J.H.C. Whitehead in the 1940s. Set

$$E^n = \{x \in \mathbb{R}^n : ||x|| \leq 1\},$$
$$U^n = \{x \in \mathbb{R}^n : ||x|| < 1\},$$
$$S^{n-1} = \{x \in \mathbb{R}^n : ||x|| = 1\}$$

where $||(x_1, \ldots, x_n)|| = \sqrt{x_1^2 + \ldots + x_n^2}$. Let X be a Hausdorff space with subspace $A \subset X$. A space Y is said to be obtained from X by *attaching an n-cell* to A if $X \subset Y$ and there exists a continuous map $\phi\colon E^n \to Y$ that restricts to a continuous map $\phi'\colon S^{n-1} \to A$ and that maps U^n homeomorphically into Y. We usually denote the open subset $\phi(U^n)$ by e^n, say that e^n is an *n-n-cell*, and that ϕ is the *attaching map*.

Definition A finite *CW-space* is a Hausdorff space X posessing a chain of subspaces

$$X^0 \subset X^1 \subset X^2 \subset \cdots \subset X^n = X$$

where X_0 is a discrete finite set and where, for each $d \geq 1$, there is a chain of subspaces

$$X^{d-1} = X_0^d \subset X_1^d \subset \cdots \subset X_{\alpha_d}^d = X^d$$

with X_i^d obtained from X_{i-1}^d by attaching a d-cell to X^{d-1}. We call X^d the *d-skeleton*, and say that X is of *dimension n*.

For an example of a CW-space consider the *torus* $T = S^1 \times S^1$. This can be viewed as the surface of a donought. The cellular decomposition of T described above, involving four 2-cells, is a CW-space. So too is the above cellular decomposition of T involving eighteen curvilinear triangles. Any polytope is a CW-space where the d-faces are the d-cells.

An extra technical axiom is required when CW-spaces with infinitely many cells are to be considered. One requires that the space X and the subspaces X^n have the weak topology: a subset $A \subset X$

is closed if and only if $A \cap X$ and $\phi^{-1}(A)$ are closed for all attaching maps ϕ. The prefix "CW" stands for "closure finite" and "weak topology".

Definition For a finite CW-space X of dimension n we define the Euler-Poincaré characteristic

$$\chi(X) = \alpha_0 - \alpha_1 + \alpha_2 \cdots + (-1)^{n+1}\alpha_n \ .$$

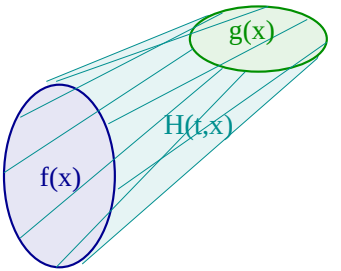
We usually abbreviate this terminology and refer to $\chi(P)$ as the *Euler characteristic*.

A given compact space can usually be endowed with many finite CW-space structures, and it can be proved that any two CW-structures yield the same Euler characteristic. In fact, a much stronger statemnet is true: two finite CW-spaces have the same Euler characteristic if they are *homotopy equivalent*. We finish this section with an introduction to the notion of *homotopy*, a notion on which algebraic topology is built.

Definition Two continuous functions $f, g\colon X \to Y$ between topological spaces X, Y are *homotopic* if there exists a continuous function

$$H\colon [0,1] \times X \to Y$$

such that $H(0,x) = f(x)$ and $H(1,x) = g(x)$ for all $x \in X$. We call H a *homotopy* and write $f \simeq g$.



We sometimes write $h_t(x) = H(t,x)$ and think of a homotopy as a continuous family $\{h_t(x)\}_{t\in[0,1]}$ of continuous maps. Exercise 5 is to show that $\simeq$ is an equivalence relation on continuous functions $f\colon X \to Y$.

Definition Spaces X and Y are said to have the same *homotopy type*, or to be *homotopy equivalent*, if there exist maps $f\colon X \to Y$, $g\colon Y \to X$ such that $fg \simeq 1_Y$ and $gf \simeq 1_X$. In this situation we write $X \simeq Y$.

Proposition 2 *Any non-empty convex subset $X \subset \mathbb{R}^n$ is homotopy equivalent to a point.*

Proof. Let v be some point in X, let $V = \{v\}$ be the one-point space, and consider the inclusion $\iota\colon V \to X$ and projection $\rho\colon X \to V$. The formula $H(t,x) = v + t(x - v)$ defines a homotopy

$H\colon [0,1] \times X \to X$ between the identity map $1_X\colon X \to X$ and the composite $\iota\rho\colon X \to X$. The composite $\rho\iota\colon V \to V$ is the identity (and thus homotopic to the identity). □

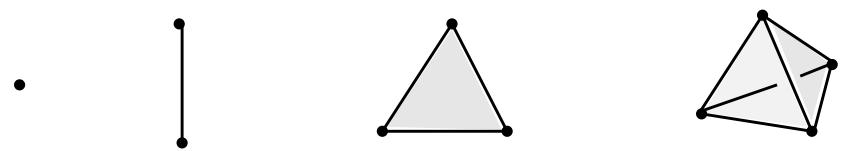
We are now in a position to state a generalisation of Theorem 1.

Theorem 3 *Two finite CW-spaces have the same Euler-Poincaré characteristic if they have the same homotopy type.*

An approach to a proof of Theorem 3 is briefly outlined in Chapter 2.

Exercises

1. An *n-simplex* Δ^n is the convex hull of a set $V = \{v_1, \ldots, v_{n+1}\}$ of $n+1$ linearly independent points in $\mathbb{R}^{n+1}$.



For $k \geq 0$ the k-dimensional faces of Δ^n are in bijection with the subsets of V of size k.

Exercise: Prove directly that $\chi(\Delta^n) = 1$.

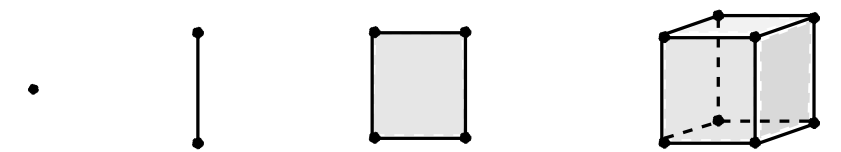
2. (a) Let α_k^X denote the number of k-cells in a CW-space X. For any two CW-spaces X, Y there is a cellular structure on the direct product $X \times Y$ satisfying

$$\alpha_k^{X\times Y} = \sum_{s,t\geq 0, s+t=k} \alpha_s^X \times \alpha_t^Y \ .$$

Exercise: Derive a formula for $\chi(X \times Y)$ in terms of $\chi(X)$ and $\chi(Y)$.

(b) Calculate $\chi(S^1 \times S^1 \times S^1)$ where S^1 is the unit circle in $\mathbb{R}^2$.

(c) Use part (a) to calculate directly the Euler characteristic $\chi(I^n)$ of the unit *n-cube*, where $I = [0, 1]$.



3. (a) Suppose that a CW=space $W = X \cup Y$ is a union of two CW-spaces X and Y whose CW-structures agree on the intersection $X \cap Y$.

Exercise: Find a formula for $\chi(W)$ in terms of $\chi(X)$, $\chi(Y)$ and $\chi(X \cap Y)$.

(b) Derive the formula $\chi(X) = 2 - 2g$ for the Euler charactersitic of a surface X of *genus* g (i.e. the surface of a doughnut with g holes).

4. Implement a GAP function for computing a minimal set of inequalities describing the convex hull of a set $S \subset \mathbb{R}^n$ as an intersection of half spaces. The implementation could involve the following basic functions.

   - `VectorListToInequalitiesMat(S)` inputs a $d \times n$ matrix S whose rows are the d vectors whose convex hull we are computing. It outputs an $m \times (n+d+1)$ matrix A whose rows $[a_0, a_1, \ldots, a_{n+d}]$ represent inequalities $a_0 + a_1 x_1 + \cdots a_n x_n + a_{n+1} t_1 + \cdots + a_{n+d} t_d \geq 0$ describing the convex hull.
   - `FourierMotzkinEliminate(A,i)` eliminates the ith column of the $m \times (n+d+1)$ matrix A, and inserts new rows corresponding to Fourier-Motzkin elimination of the ith variable.

   Fourier-Motzkin elimination can be used to test if a given system of inequalities has any solution: eliminate all but the first variable; the resulting system has a solution if and only if the original system has a solution.

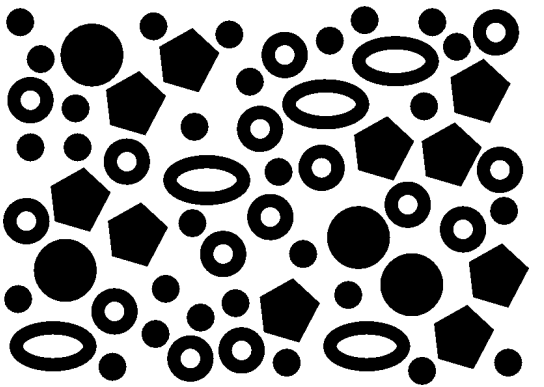
   To test if a given inequality E is implied by a system $Ax \geq b$ we can add the negation of E to the system; the resuling system *not* E *and* $Ax \geq b$ has no solution if and only if E is implied by $Ax \geq b$.

5. (a) Show that homotopy $\simeq$ is an equivalence relation on the set of continuous functions $f\colon X \to Y$ between two given spaces X, Y.

   (b) Show that homotopy type is an equivalence relation on spaces.

1.3 Digital image analysis

Given a black and white digital image of a collection of circular coins of various sizes, pentagonal coins, and washers of various shapes and sizes



we consider the problem of automatically counting: (i) the number of coins, (ii) the number of pentagonal coins, and (iii) the number of washers. A computer implementation of the Euler characteristic can be used for this problem.

We regard the black region of the image as a subspace $Y \subset \mathbb{R}^2$. This space Y has many path components (where, by definition, two points $y, y' \in Y$ belong to the same *path component* if there is a continuous function $p\colon [0,1] \to Y$ with $p(0) = y, p(1) = y'$). The number β_0 of path components

equals the total number of coins and washers. We can compute β_0 in GAP using the following two commands.

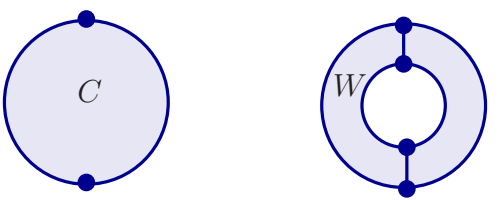
```
gap> Y:=ReadImageAsPureCubicalComplex("coinsandwashers.png",500);;
gap> Beta_0:=Bettinumbers(Y,0);
62
```

The first command reads a digital image into GAP and represents it as a CW-space. We postpone for a moment the precise details of this representation. But we should mention that any digital image is a collection of coloured pixels. The colour of a pixel is stored as a vector (r, g, b) with $0 \leq r, g, b \leq 255$. In converting an image to a CW-space we deem pixels to contribute to the space when $r + g + b \leq T$ where T is some threshold. In the example $T = 500$.

The Euler characteristics

$$\chi(\text{coin}) = 1, \qquad \chi(\text{washer}) = 0$$

can be computed from the following CW-spaces.



$$\chi(C) = V - E + F = 1 - 1 + 1 = 1, \qquad \chi(W) = V - E + F = 4 - 6 + 2 = 0.$$

Letting $\pi_0(Y)$ denote the collection of path components of Y, we have

$$\chi(Y) = \sum_{P \in \pi_0(Y)} \chi(P) \ .$$

and thus $\chi(Y)$ is equal to the total number of coins in the image. We can compute $\chi(Y)$ using the following GAP commands.

```
gap> Y:=ReadImageAsPureCubicalComplex("coinsandwashers.png",500);;
gap> chi:=EulerCharacteristic(Y);
42
```

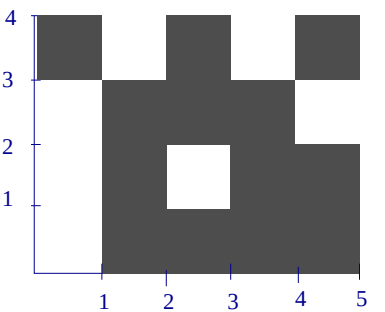
A pentagonal coin has a boundary curve with precisely five non-differentiable (or *singular*) points. The following GAP commands determine the number of pentagons.

```
gap> Y:=ReadImageAsPureCubicalComplex("coinsandwashers.png",500);;
gap> NumberOfPentagons:=0;;
gap> for i in [1..Bettinumbers(Y,0)] do
> P:=PathComponentOfPureCubicalComplex(Y,i);
> S:=SingularitiesOfPureCubicalComplex(P,3,30);
> if Bettinumbers(S,0)=5 then
>     NumberOfPentagons:=NumberOfPentagons+1;
> fi;
> od;
gap> Print(NumberOfPentagons,"\n");
10
```

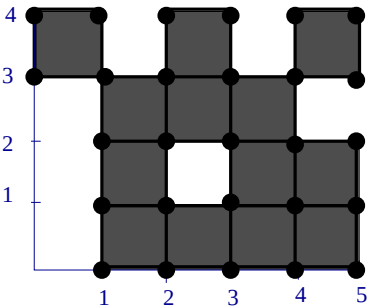
In order to explain these commands we must first address the question of combinatorially representating an image as a CW-space. A black and white image on a screen of $m \times n$ pixels can be represented as an $m \times n$ matrix of 0s and 1s, where 1 represents a black pixel and 0 a white pixel. We'll use the same convention for indexing screen pixels and matrix entries. (So the (i, j) pixel is i rows down from the top and j columns to the right.) Topologically we let a black pixel in the (i, j) position correspond to the unit square $[j-1, j] \times [m-i, m-i+1]$ in $\mathbb{R}^2$. The entire image then corresponds to a subspace Y of the rectangle $[0, n] \times [0, m]$ in $\mathbb{R}^2$. The subspace Y is a union of unit squares whose corners have integer coordinates. For instance, the matrix

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

corresponds to the following subspace Y in $\mathbb{R}^2$.



A cellular structure is needed for the computation of Euler characteristics. We endow Y with the structure of a CW-space in which each unit square has four vertices, four edges and one face, the vertices having integer coordinates.



In this example Y has Euler characteristic

$$\chi(Y) = 27 - 40 + 13 = 0$$

and the number of path components is

$$\beta_0 = 1.$$

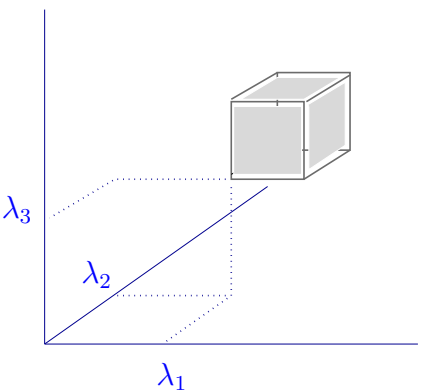
We formalize the above representation of digital images by introducting the notion of a *pure cubical complex*. We present definitions in sufficient generality for handling digital images of any integral dimension $d \geq 1$.

Definition A *pure cubical complex* of dimension d is an array $K = (k_\lambda)_{\lambda\in\Lambda}$ of 0s and 1s where the index λ ranges over an index set $\Lambda = \{1,\dots,n_1\}\times\{1,\dots,n_2\}\times\cdots\times\{1,\dots,n_d\} \subset \mathbb{N}^d$. The integer vector $(n_1,\dots,n_d)$ is called the *dimension vector* for K.

To each $\lambda = (\lambda_1,\dots,\lambda_d) \in \Lambda$ we associate a unit d-cube

$$[[\lambda]] = [[\lambda_1,\dots,\lambda_d]] = [\lambda_1, 1+\lambda_1]\times[\lambda_2, 1+\lambda_2]\times\cdots[\lambda_d, 1+\lambda_d] \subset \mathbb{R}^d$$

and endow this cube with the canonical CW-structure having 2^d 0-cells and $2d$ cells of dimension $d-1$. For $d=3$ we picture $[[\lambda]] = [[\lambda_1,\lambda_2,\lambda_3]]$ as follows.



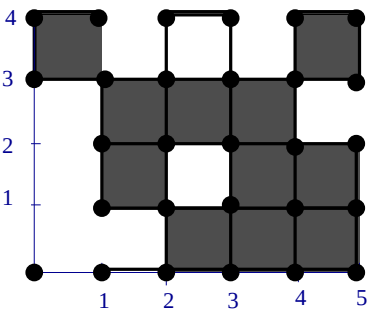
Definition The *geometric realization* of a pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$ is the subspace

$$|K| = \bigcup_{\lambda\in\Lambda, k_\lambda=1} [[\lambda]]$$

of $\mathbb{R}^d$ arising as the union of the associated d-cubes. We endow $|K|$ with the CW-structure inherited from the cubes $[[\lambda]]$.

This notion of geometric realization is consistent, up to an isometry, with the representation of 2-dimensional digital images described above.

For computational reasons it is useful to have combinatrial representations of CW-spaces arising as unions of unit cubes of various dimensions. For example, the CW-space



is a union of 0-dimensional, 1-dimensional and 2-dimensional cubes and can be represented by the

matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

of 0s and 1s where the 1s correspond to the (centres of the) cells in the CW-structure. As such a space is not the geometric realization of any pure cubical complex we introduce the following definition.

Definition A *cubical complex* of dimension d is an array $M = (m_\theta)_{\theta \in \Theta}$ of 0s and 1s where the index θ ranges over an index set $\Theta = \{1, \ldots, 2n_1+1\} \times \{1, \ldots, 2n_2+1\} \times \cdots \times \{1, \ldots, 2n_d+1\} \subset \mathbb{N}^d$. There is one axiom: if $m_\theta = 1$ then $m_{\theta'} = 1$ for any $\theta' \in \Theta$ obtained by adding 1 to, or substracting 1 from, an even entry in the index θ. The integer vector $(n_1, \ldots, n_d)$ is called the *dimension vector* for M.

It is important to note that pure cubical complexes are not just special cases of cubical complexes. The two structures are quite different.

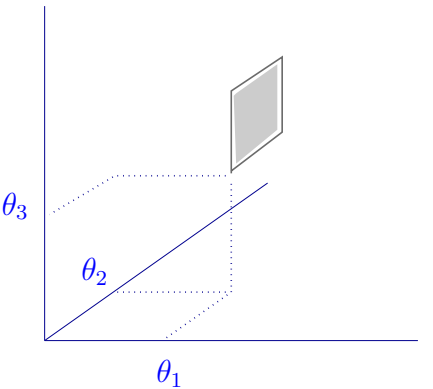
For $\theta_i \in \mathbb{Z}$ we define

$$\langle\langle \theta_i \rangle\rangle = \begin{cases} [\theta_i, 1+\theta_i], & \theta_i \text{ odd} \\ \{\theta_i\}, & \theta_i \text{ even} \end{cases}$$

and to each $\theta = (\theta_1, \ldots, \theta_d) \in \Theta$ we associate a unit cube

$$\langle\langle \theta \rangle\rangle = \langle\langle \theta_1 \rangle\rangle \times \langle\langle \theta_2 \rangle\rangle \times \cdots \times \langle\langle \theta_d \rangle\rangle$$

of dimension equal to the number of even entries in θ. For $d = 3$ and $\theta = (\theta_1, \theta_2, \theta_3)$ with odd θ_1 and even θ_2, θ_3 we picture $\langle\langle \theta \rangle\rangle$ as follows.



Definition The *geometric realization* of a cubical complex $M = (m_\theta)_{\theta \in \Theta}$ is the subspace

$$|M| = \bigcup_{\theta \in \Theta, m_\theta = 1} \langle\langle \theta \rangle\rangle$$

of $\mathbb{R}^d$ arising as the union of the associated unit cubes. We endow $|M|$ with the CW-structure inherited from the cubes $\langle\langle\theta\rangle\rangle$.

To each pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$ of dimension d with dimension vector $(n_1, \ldots, n_d)$ we associate a cubical complex $K^{cw} = (m_\theta)_{\theta\in\Theta}$ of the same dimension, and having the same dimension vector, by setting $m_\theta = 1$ for precisely those indexes θ that lie in the set

$$\{(2\lambda_1 + \epsilon_1, \ldots, 2\lambda_d + \epsilon_1) : \epsilon_i \in \{-1, 0, 1\}, k_{(\lambda_1,\ldots,\lambda_d)} = 1\}.$$

Definition The *Euler characteristic* of a cubical complex $M = (m_\theta)_{\theta\in\Theta}$ is the integer

$$\chi(M) = \sum_{\theta=(\theta_1,\ldots,\theta_d)\in\Theta} (-1)^{\theta_1+\cdots+\theta_d} m_\theta \ .$$

The *Euler characteristic* of a pure cubical complex K is the integer $\chi(K^{cw})$.

These definitions are designed to immediately yield the following.

Lemma 4 *For any pure cubical complex K and cubical complex M we have*

$$\chi(|K|) = \chi(K), \qquad \chi(|M|) = \chi(M).$$

The Euler characteristic for pure cubical complexes is implemented in GAP and illustrated in the above computer examples. We now turn to cubical notions of *path component* and *boundary singularity* which are also used in the above computer examples.

Definition A *path* in a d-dimensional pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$ is a sequence $s_1, s_2, \ldots, s_n$ of indices in Λ for which each $k_{s_i} = 1$ and $s_i - s_{i+1} \in \{-1, 0, 1\}^d$. We call s_1 the *source*, s_n the *target*, and write $s_1 \equiv s_n$. This defines an equivalence relation on those indices $\lambda \in \Lambda$ for which $k_\lambda = 1$. The number of equivalence classes is denoted by $\beta_0(K)$ and we clearly have the following.

Lemma 5 *$\beta_0(K)$ is the number of path components in the geometric realization $|K|$ of a pure cubical complex K.*

The following simple algorithm computes $\beta_0(K)$. It is presented in a form that destroys the pure cubical complex K.

Path Components Algorithm for Pure Cubical Complexes

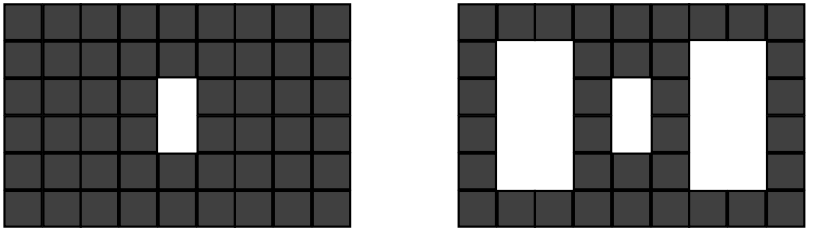
Input: A pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$.
Output: The number $N = \beta_0(K)$.
Procedure:
initialize $N := 1$;
while some $k_\lambda = 1$
    set N:=N+1;
    choose $\lambda \in \Lambda$ such that $k_\lambda = 1$ and set $k_\lambda := N$;
    while some $k_\lambda = 1$, $k_{\lambda'} = N$, $\lambda - \lambda' \in \{-1, 0, 1\}^d$
        set $k_\lambda := 1$;
set $N := N - 1$;
return N

We now work towards the definition of a singular boundary point in a pure cubical complex.

Definition The *boundary* of a d-dimensional pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$ is a pure cubical complex $\partial B = (b_\lambda)_{\lambda\in\Lambda}$ with the same index set. The equality $b_\lambda = 1$ holds if and only if $k_\lambda = 1$ and $k_{\lambda'} = 0$ for some index λ' with $\lambda' - \lambda \in \{-1, 0, 1\}^d$.

The following are examples of the geometric realization of a 2-dimensional pure cubical complex

and its boundary.



Definition Let $K = (k_\lambda)_{\lambda\in\Lambda}$ and $K' = (k'_\lambda)_{\lambda\in\Lambda}$ be pure cubical complexes with the same index set. Their *intersection*, *union*, *complement*, *difference* and *size* are defined as

$$K \cap K' = (\min(k_\lambda, k'_\lambda))_{\lambda\in\Lambda} \ ,$$

$$K \cup K' = (\max(k_\lambda, k'_\lambda))_{\lambda\in\Lambda} \ ,$$

$$\mathrm{Comp}(K) = (\overline{k}_\lambda)_{\lambda\in\Lambda} \ ,$$

$$K \setminus K' = K \cap \mathrm{Comp}(K'),$$

$$\mathrm{Size}(K) = \sum_{\lambda\in\Lambda} k_\lambda$$

where $\overline{1} = 0$ and $\overline{0} = 1$.

Definition Given an index set $\Lambda \subset \mathbb{N}^d$, an index $c \in \Lambda$ and a positive integer r we define the pure cubical complex $S(\Lambda, r, c) = (k_\lambda)_{\lambda \in \Lambda}$ by setting

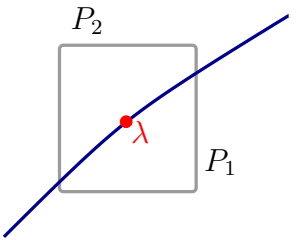
$$k_\lambda = \begin{cases} 1, & \text{if } \lambda - c \in \{-1, 0, 1\}^d \\ 0, & \text{otherwise} \end{cases} .$$

We refer to $S(\Lambda, r, c)$ as the *sphere* of radius r centred at c.

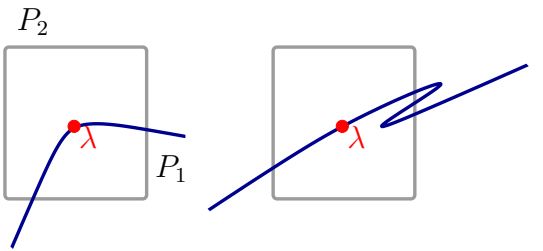
Definition Let K be a pure cubical complex with boundary $\partial K = (b_\lambda)_{\lambda \in \Lambda}$. Let r be a positive integer and τ a percentage. An index $\lambda \in \Lambda$ is said to be (r, τ)-*smooth* if

1. $b_\lambda = 1$,
2. the difference $D = S(\Lambda, r, \lambda) \backslash \partial K$ is a union $D = P_1 \cup P_2$ with $\beta_0(D) = 2$, $\beta_0(P_1) = \beta_0(P_2) = 1$,
3. and
$$\frac{|\text{Size}(P_1) - \text{Size}(P_2)|}{\text{Size}(D)} \leq \tau\%.$$

The following pictures a smooth index for $d = 2$



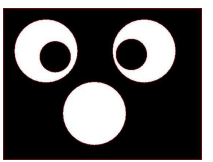
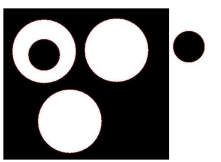
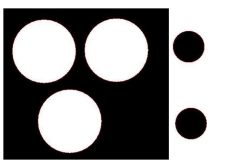
and the following pictures non-smooth indices for $d = 2$.

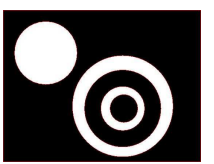
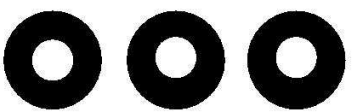


Definition Given a pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$, a positive integer r and a percentage τ we define the pure cubical complex Singularities$(K, r, \tau) = (b_\lambda)_{\lambda\in\Lambda}$ by setting $b_\lambda = 1$ precisely when λ is not (r, τ)-smooth. We call Singularities(K, r, τ) the *complex of boundary singularities*.

Exercises

1. The following five pure cubical complexes K_i each have $\beta_0(K_i) = 3$ and $\chi(K_i) = 0$.

$|K_1| =$  $|K_2| =$  $|K_3| =$ 

$|K_4| =$  $|K_5| =$ 

   Which pairs of complexes can be distinguished using $\beta_0(X)$, $\beta_0(\mathrm{Comp}(X))$, $\chi(X)$, $\chi(\mathrm{Comp}(X))$ where X is a path component, or union of path components, of K_i?

2. Twenty-five slices of a 3-dimensional pure cubical complex M are stored as files file1.png, ..., file25.png in a directory called "pictures". What information about the shape of M can be

deduced from the following GAP session.

```
gap> M:=ReadImageSequenceAsPureCubicalComplex("pictures",400);
Pure cubical complex of dimension 3.

gap> Bettinumbers(M);
[ 1, 0, 0, 0 ]

gap> B:=BoundaryOfPureCubicalComplex(M);;
gap> D:=PureCubicalComplexDifference(B,S);;
gap> Bettinumbers(D,0));
8

gap> P:=List([1..8],n->PathComponentOfPureCubicalComplex(D,n));;
gap> for n in [2..8] do
> U:=ThickenedPureCubicalComplex(P[1]);
> U:=ThickenedPureCubicalComplex(U);
> V:=ThickenedPureCubicalComplex(P[n]);
> V:=ThickenedPureCubicalComplex(V);
> W:=PureCubicalComplexIntersection(U,V);
> Print(Size(W),"  ");
> od;
128 95 95 50 0 0 0
```

1.4 Statistical data analysis

The following table contains simulated data. This kind of data could arise, for example, as the expression of three genes in 400 DNA specimens where the entry in the jth column of the ith row is the amount of gene j in specimen i.

	gene 1	gene 2	gene 3
specimen 1	8	35	7
specimen 2	3	38	65
specimen 3	7	9	17
specimen 4	1	56	57
specimen 5	1	44	5
⋮	⋮	⋮	⋮
specimen 400	2	94	79

It is common practice to search for natural clusters in such data, where a *cluster* is a collection of nearby data points that are distant from the remaining data points.

Clusters will depend on a choice of metric on the data points. We could consider each specimen as a point in $\mathbb{R}^3$ and use the *Manhattan metric*

$$\text{dist}(\ (x,y,z)\ ,\ (x',y',z')\ ) = |x-x'| + |y-y'| + |z-z'|\ .$$

Alternatively we could use the *euclidean metric*

$$\text{dist}(\ (x,y,z)\ ,\ (x',y',z')\ ) = \sqrt{(x-x')^2+(y-y')^2+(z-z')^2}$$

or the *Chebychev metric*

$$\text{dist}(\ (x,y,z)\ ,\ (x',y',z')\ ) = \text{Max}\{|x-x'|,|y-y'|,|z-z'|\}$$

or the *Pearson correlation metric*

$$\text{dist}(\ (x,y,z)\ ,\ (x',y',z')\ ) = 1-r$$

with

$$r = (x-\mu, y-\mu, z-\mu)\cdot(x'-\mu', y'-\mu', z'-\mu')/3\sigma\sigma'$$

and $\mu, \mu', \sigma, \sigma'$ the respective means and standard deviations.

Clusters also depend on how we measure distance between two sets A, B of data points. We could use *average linkage* where the distance between A and B is the average of $\text{dist}(u,v)$ for $u \in A, v \in B$. We could use *single linkage* where the distance between A and B is the minimum of $\text{dist}(u,v)$ for $u \in A, v \in B$. We could use *complete linkage* where the distance between A and B is the maximum of $\text{dist}(u,v)$ for $u \in A, v \in B$.

Not too much should be inferred from the presence of clusters in a data set since they depend on the choice of metric and linkage. But their existence might suggest directions for further statistical analysis.

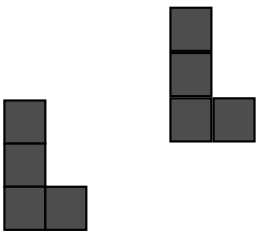
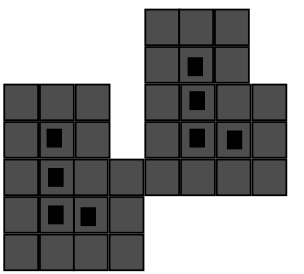
We can perform a single linkage, Chebychev metric, cluster analysis using the language of pure cubical complexes. For this we associate to each data point (x, y, z) the unit cube

$$[[x, y, z]] = [x, x + 1] \times [y, y + 1] \times [z, z + 1] \ .$$

Since the data points in our example have integer coordinates, the union of all 400 cubes can be viewed as the geometric realization $|K|$ of a 3-dimensional pure cubical complex.

Definition Given a pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$ of dimension d, we define the *thickened* pure cubcical complex $TK = (t_\lambda)_{\lambda\in\Lambda}$ by setting $t_\lambda = 1$ precisely when $k_{\lambda'} = 1$ for some index λ' with $\lambda - \lambda' \in \{-1, 0, 1\}^d$.

A 2-dimensional example is:

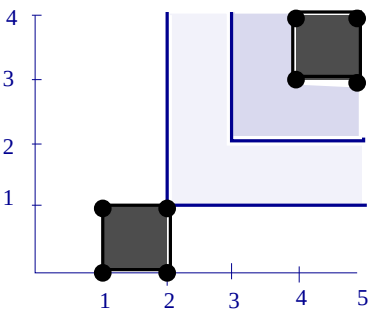
$|K| =$  $|TK| =$ 

The Chebychev distance between two distinct vectors $a, b \in \Lambda \subset \mathbb{N}^d$ can be described in terms of thickenings. Let A and B be pure cubical complexes with index set Λ and with geometric

realizations $|A| = [[a]]$ and $|B| = [[b]]$. The Chebychev distance between a and b is equal to $1 + n$ where n is the smallest integer such that

$$\beta_0(A \cup T^n B) = 1.$$

For example, the Chebychev distance between $a = (1, 0)$ and $b = (4, 3)$ is 3.



Single linkage cluster analysis with the Chebychev metric can be performed using the following

algorithm.

Chebychev Cluster Analysis Algorithm

1. Input a set D of data points in $\mathbb{N}^d$.
2. Convert the data points to a pure cubical complex $K = (k_\lambda)_{\lambda\in\Lambda}$ by setting $k_\lambda = 1$ precisely when $\lambda \in D$.
3. Compute the sequence of thickened pure cubical complexes

$$K, TK, T^2K, T^3K, \ldots$$

(where $T^nK = T(T^{n-1}K), T^0K = K$) .

4. Compute the integer sequence

$$\beta_0(K), \beta_0(TK), \beta_0(T^2K), \beta_0(T^3K), \ldots .$$

5. Search for long constant portions in the sequence $\beta_0(T^nK)$ $(n \geq 0)$ as these should represent data clusters.

The following GAP session applies an implementation of the cluster algorithm to the above simulated

data.

```
gap> K := ReadMatrixAsPureCubicalComplex("simulatedData.txt");;
gap> ListOfBettiZeros := [];;
gap> for i in [1..25] do
> BettiZero:=Bettinumbers(K,0);
> K:=ThickenedPureCubicalComplex(K);
> Add(ListOfBettiZeros,BettiZero);
> od;
gap> Print(ListOfBettiZeros,"\n");
[ 330, 38, 6, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 1 ]
```

The constant sequence $\beta_0(T^n K) = 2$ for $4 \leq n \leq 12$ suggests the presence of two natural clusters in the sample data.

As topologists, we might also consider the sequence of Euler characteristics

$$\chi(K),\ \chi(TK),\ \chi(T^2K),\ \chi(T^3K),\ \ldots$$

which is computed in the following GAP session.

```
gap> K := ReadMatrixAsPureCubicalComplex("simulatedData.txt");;
gap> EulerChars := [];;
gap> for i in [1..25] do
> EulerChar:=EulerCharacteristic(K);
```

```
> K:=ThickenedPureCubicalComplex(K);
> Add(EulerChars,EulerChar);
> od;
gap> Print(EulerChars,"\n");
[ 330, 6, -6, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, -3, -1, -1, -1,
  -1, -1, -1, -1, 1, 1, 1, 1 ]
```

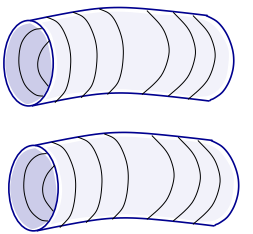
The sequences $\beta_0(T^nK)$, $\chi(T^nK)$ are presented in the following table.

n	$\beta_0(T^nY)$	$\chi(T^nY)$	n	$\beta_0(T^nY)$	$\chi(T^nY)$	n	$\beta_0(T^nY)$	$\chi(T^nY)$
0	330	330	9	2	0	18	1	-1
1	30	6	10	2	0	19	1	-1
2	6	-6	11	2	0	20	1	-1
3	4	4	12	2	0	21	1	1
4	2	0	13	1	-3	22	1	1
5	2	0	14	1	-1	23	1	1
6	2	0	15	1	-1	24	1	1
7	2	0	16	1	-1	25	1	1
8	2	0	17	1	-1			

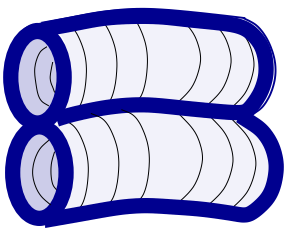
To explain these computational results topologically we should look for some "natural" subspace $X \subset \mathbb{R}^3$ with the properties that:

1. $[\beta_0(X), \chi(X)] = [2, 0]$,
2. X is contained in some neighbourhood $X \subset TX$ with $[\beta_0(TX), \chi(TX)] = [1, -1]$.

If some obvious candidate for X exists, we might consider further investigating the possibility that the data points are actually sampled from X. For instance, a space X consisting of two hollow tubes



has $\beta_0(X) = 2$ and Euler characteristic $\chi(X) = 0$. Furthermore, a certain thickening TX



has $\beta_0(TX) = 1$ and Euler characteristic $\chi(TX) = -1$.

Let us return to our table of simulated data and consider how we might perform a cluster analysis of β_0 and χ using some metric other than the Chebychev metric? The metric may bear little relation to any metric on $\mathbb{R}^3$. In fact, suppose that instead of being presented with a vector in $\mathbb{R}^3$ for each specimen we are simply presented with a table

	specimen 1	specimen 2	$\cdots$	specimen 400
specimen 1	0	31	$\cdots$	202
specimen 2	31	0	$\cdots$	195
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
specimen 400	202	195	$\cdots$	0

where the entry in the ith row and jth column is some measure of the distance between specimen i and specimen j. We won't assume any knowledge of the measure used. The table is stored as a symmetric matrix $S = (s_{ij})$. We need a method for investigating topology underlying the data in S.

There is no obvious method of associating a pure cubical complex to S. Rather, we take a 400-dimensional simplex Δ^{400} whose set of vertices we order and denote by $V = \{v_1, ..., v_{400}\}$. For any subset $A \subset V$ let $|A|$ denote the convex hull of the vertices in A. For any real number $t \geq 0$ we say that A is a *t-cluster* if $s_{ij} \leq t$ for all $v_i, v_j \in A$ and we write $||A|| \leq t$. We can then consider the space

$$X_S(t) = \bigcup_{||A|| \leq t} |A|$$

for any desired value of $t \geq 0$. This space inherits the structure of a CW-space from Δ^{400}. For any increasing sequence of real numbers $0 = t_0 < t_1 < t_2 < \ldots$ we can consider the chain of CW-space inclusions

$$X_S(t_0) \subseteq X_S(t_1) \subseteq X_S(t_2) \subseteq \ldots \subseteq \Delta^{400}.$$

Applying β_0 to this chain should highlight the presence of any (single linkage) clusters. To formalise the method we introduce terminology for regarding $X_S(t)$ as the *geometric realization* of a certain *simplicial complex*.

Definition A *simplicial complex* consists of an ordered set V of *abstract vertices* and a set K of finite nonempty subsets of V called *abstract simplices* such that:

1. $\{v\} \in K$ for any $v \in V$.
2. Any nonempty subset of an abstract simplex is an abstract simplex.

An abstract simplex σ containing exactly $k+1$ abstract vertices is called an *abstract n-simplex*. If $\sigma' \subseteq \sigma$ then σ' is called a *face* of σ.

Definition Let K be a simplicial complex with abstract vertex set V of finite size $|V|$. For the ith abstract vertex $v_i \in V$ we define $|v_i|$ to be the ith standard basis vector for $\mathbb{R}^{|V|}$. For an abstract

simplex $\sigma \in K$ we define $|\sigma|$ to be the convex hull of the set $\{|v| : v \in \sigma\}$. The *geometric realization* of K is the CW-space

$$|K| = \bigcup_{\sigma \in K} |\sigma| \ .$$

Definition For an $n \times n$ symmetric matrix $S = (s_{ij})$ and real number t the *Vietoris-Rips* simplicial complex $K_S(t)$ has abstract vertex set $V = \{1, \ldots, n\}$ and has abstract simplices those subsets $\sigma \subset V$ with the property that $s_{ij} \leq t$ for all $i, j \in \sigma$.

With these definitions we see that the CW-space $X_S(t)$ described above is the geometric realization $|K_S(t)|$ of the Vietoris-Rips simplicial complex. To illustrate the Vietoris-Rips complex on a small example we consider the symmetric group S_n of degree n. The *inversion number* $Inv(\pi)$ of a permutation $\pi \in S_n$ is the size of the set

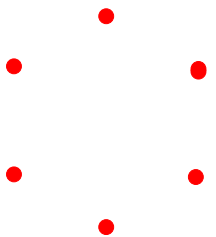
$$\bigcup_{1 \leq i \leq n} \{j < i \ : \ \pi(j) > \pi(i)\} \ .$$

A metric on S_n is given by

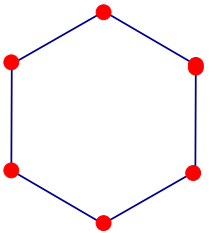
$$dist(\pi, \tau) = Inv(\pi\tau^{-1}).$$

Now take $n = 3$, order the six permutation in S_3 in some fashion, and form the 6×6 matrix $P = (p_{ij})$ with p_{ij} the distance between the ith and jth permutations in S_3. The geometric realization of the Vietoris-Rips complex $K_P(t)$ is: a discrete set of six points for $t = 0$; a hexagonal graph for $t = 1$;

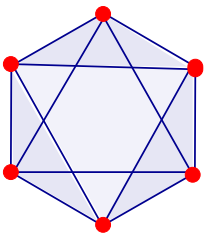
a CW-space homeomorphic to the sphere S^2 for $t = 2$; the 6-dimensional simplex Δ^6 for $t \geq 3$.



$t = 0$



$t = 1$



$t = 2$

Definition For any simplicial complex K the abstract vertices and abstract 1-simplices constitute a graph which we denote by $\Gamma(K)$. We define $\beta_0(K)$ to be the number of path components in this graph. The following is immediate.

Lemma 6 *$\beta_0(|K|) = \beta_0(K)$ for any simplicial complex K.*

We can represent a finite simplicial complex K on a computer simply by creating a list whose terms are lists of integers representing the abstract simplices. Recall that the incidence matrix M for the graph $\Gamma(K)$ is a symmetric matrix with 1 in the ith row and jth column if there is an edge between

the ith and jth verrtices; otherwise there is a 0 in the ith row and jth column of M. The incidence matrix can be used in the computation of $\beta_0(K)$.

Path Components Algorithm for Simplicial Complexes

Input: A finite simplicial complex K.
Output: The number $N = \beta_0(K)$.
Procedure:
create the incidence matrix $M = (m_{ij})$ for the graph $\Gamma(K)$;
initialize $N := 1$;
while some $m_{ij} = 1$
    set $N := N + 1$;
    choose some i, j with $m_{ij} = 1$;
    for each j'
        if $m_{ij'} = 1$ then set $m_{ij'} = N$;
set $N := N - 1$;
return N;

For our 400×400 symmetric matrix S the sequence

$$\beta_0(X_S(0)) \subseteq \beta_0(X_S(2)) \subseteq \beta_0(X_S(4)) \subseteq \beta_0(X_S(6)) \subseteq \ldots \subseteq \beta_0(X_S(50))$$

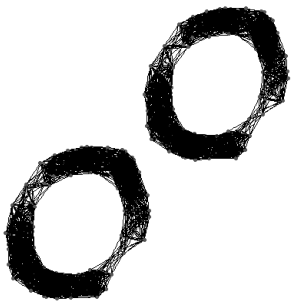
is computed in the following GAP session.

```
gap> Read("symmetricMatrix.txt"); # This creats the matrix S
gap> ListOfBettiZeros:=[];;
gap> for n in [0..25] do
> G:=SymmetricMatrixToGraph(S,2*n);
> Add(ListOfBettiZeros,PathComponentsOfGraph(G,0));
> od;
gap> Print(ListOfBettiZeros,"\n");
[ 390, 308, 148, 38, 14, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
  2, 2, 2, 1 ]
```

The constant sequence $\beta_0(X_S(2n)) = 2$ for $6 \leq n \leq 24$ suggests the presence of two principal clusters in the data set from which the symmetric matrix S was constructed. We in fact constructed the matrix $S = (s_{ij})$ by setting s_{ij} equal to the Manhattan distance between the ith and jth data points in our original table of 400 simulated data points. Thus a single linkage, Manhattan metric, cluster analysis suggests the same number of clusters in the data as did the previous single linkage, Chebychev metric, cluster analysis.

It is not practical to directly compute the Euler characteristic $\chi(X_S(t))$ for a high-dimensional CW-space $X_S(t) = |K_S(t)|$ with a vast number of cells. To overcome this practical difficulty we introduce more topological invaraints, namely *Betti numbers*, in the next chapter. However, it is possible to visualize the abstract graph $\Gamma(K_S(t))$ using the GAP command `GraphDisplay`. This function calls GRAPHVIZ software [10] for visualizing abstract graphs. It produces the following

picture of $\Gamma(K_S(t))$ for our 400×400 symmetric matrix S and $t = 20$.



This computer generated picture is certainly consistent with the Euler characteristic values obtained in the previous Chebychev metric analysis.

Exercises

1. For any set A and any collection $V = \{v_\lambda \subset A\}$ of subsets of A, the set $N(V)$ of all finite subsets $\sigma = \{v_{\lambda_1}, v_{\lambda_2}, \ldots\} \subset V$ with non-empty intersection $\cap_{v_\lambda \in \sigma} v_\lambda$ is a simplicial complex called the *nerve* of V.

   A d-dimensional pure cubical complex $K = (k_\lambda)_{\lambda \in \Lambda}$ gives rise to a finite collection $V = \{[[\lambda]] : \lambda \in \Lambda, k_\lambda = 1\}$ of unit cubes whose union is the geometric realisation $|K|$. Convince yourself that $\chi(K) = \chi(N(V))$ for the case $d = 2$. (The equality actually holds for all d.)

2. For any set A and any collection $S = \{s_\lambda \subseteq A\}$ of subsets of A, we can consider the set $V = \{v_{\lambda\mu} : s_\lambda \subsetneq s_\mu\}$ of proper inclusions with $s_\lambda, s_\mu \in S$. A *chain* in V is a finite subset

$$\{s_{\lambda_1} \subset s_{\lambda_2}, s_{\lambda_2} \subset s_{\lambda_3} \ldots, s_{\lambda_n} \subset s_{\lambda_{n+1}}\}$$

of V consisting of composable inclusions. The collection of all chains in V is a simplicial complex $Ord(S)$ called the *order complex* of S.

This order complex is particularly interesting when $A = G$ is a finite group and $S = ElAb_p(G)$ is the collection of elementary abelian p-subgroups for some prime p. The following GAP commands compute the euler characteristic of $Ord(ElAb_p(G))$ for G the nth (small) group of order g.

```
gap> G:=SmallGroup(g,n);;
gap> Ord:=QuillenComplex(G,p);;
gap> EulerCharacteristic(Ord);
```

Compute this euler characteristic for a range of small prime-power groups and formulate a conjecture.

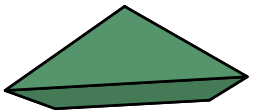
1.5 Betti numbers

For any finite CW-space X of dimension n one can compute a sequence of topological invariants $\beta_0(X), \beta_1(X), \ldots, \beta_n(X)$ known as *betti numbers*. They are non-negative integers which relate to the euler characteristic via the formula

$$\chi(X) = \beta_0(X) - \beta_1(X) + \beta_2(X) - \ldots (-1)^n \beta_n(X).$$

These invariants were first introduced by Poincaré and named after Enricho Betti. This section is an informal, examples based introduction to betti numbers. The reader will need to rely on intuition in understanding terms such as "boundary homomorphism" and "orientation". Precise definitions are postponed until Chapter 2.

Let us start with the CW-space X arising as the convex hull of the five points $(0,0,0)$, $(2,0,0)$, $(0,2,0)$, $(2,2,0)$, $(1,1,1)$ in $\mathbb{R}^3$.



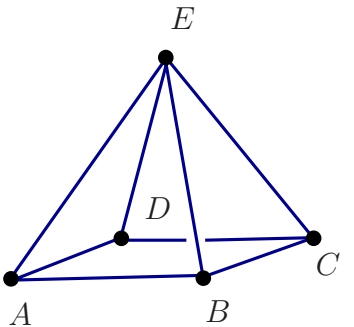
For each integer $n \geq 0$ we consider the vector space

$$C_n(X, \mathbb{Q}) = \mathbb{Q}^{\alpha_n}$$

over the rationals of dimension α_n equal to the number of n-dimensional cells in X. Thus

$$\begin{aligned} C_0(X, \mathbb{Q}) &= \mathbb{Q}^5 \ , \\ C_1(X, \mathbb{Q}) &= \mathbb{Q}^8 \ , \\ C_2(X, \mathbb{Q}) &= \mathbb{Q}^5 \ , \\ C_3(X, \mathbb{Q}) &= \mathbb{Q} \ , \\ C_n(X, \mathbb{Q}) &= 0 \quad (n \geq 4). \end{aligned}$$

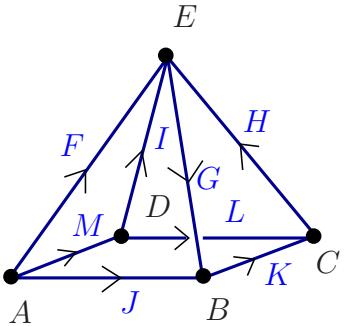
We label the n-cells of X and associate them to basis elements of $C_n(X, \mathbb{R})$. For instance, we can label the vertices by A, B, C, D, E,



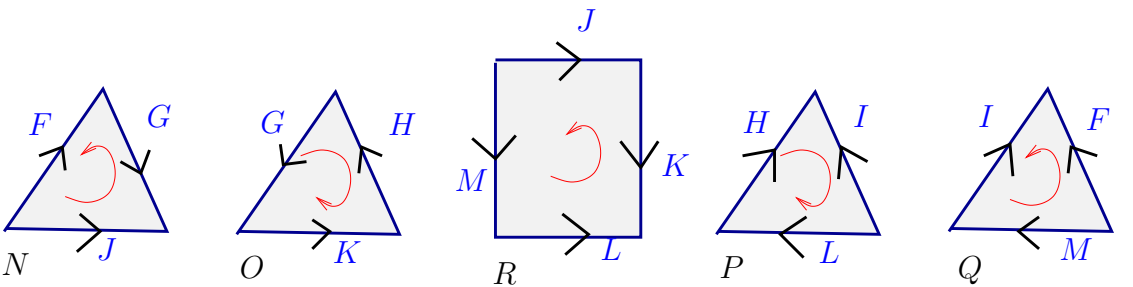
the edges by $F = \{A, E\}, G = \{B, E\}, H = \{C, E\}, I = \{D, E\}, J = \{A, B\}, K = \{B, C\}, L = \{C, D\}, M = \{A, D\}$, and the 2-faces by $N = \{F, G, J\}, O = \{G, H, K\}, P = \{H, I, L\}, Q = \{F, I, M\}, R = \{J, K, L, M\}$.

We denote a basis of $C_0(X,\mathbb{Q})$ by $\{e^0_A, e^0_B, e^0_C, e^0_D, e^0_E\}$, a basis of $C_1(X,\mathbb{Q})$ by $\{e^1_F, e^1_G, e^1_H, e^1_I, e^1_J, e^1_K, e^1_L, e^1_M\}$, a basis of $C_2(X,\mathbb{Q})$ by $\{e^2_N, e^2_O, e^2_P, e^2_Q, e^2_Q\}$, and a basis of $C_3(X,\mathbb{Q})$ by $\{e^3\}$..

We assign "orientations" to the edges of X by placing arrow heads on the them in some arbitrary fashion.



We assign "orientations" to the faces of X by placing circular arrows on the them in some arbitrary direction.



We assigne an "orientation" to the 3-cell by deeming all rays from its centre through the centres of its faces to have a common (let's say outward) direction. Our labelling and orientations can be used to derive formulae corresponding to the boundaries: $\partial(e^n)$ of the cells e^n.

$$\begin{aligned}
\partial(e^1_F) &= e^0_E - e^0_A & \qquad \partial(e^2_N) &= e^1_J - e^1_F - e^1_G \\
\partial(e^1_G) &= e^0_B - e^0_E & \qquad \partial(e^2_O) &= -e^1_G - e^1_H - e^1_K \\
\partial(e^1_H) &= e^0_E - e^0_C & \qquad \partial(e^2_P) &= e^1_H + e^1_L + e^1_I \\
\partial(e^1_I) &= e^0_E - e^0_D & \qquad \partial(e^2_Q) &= e^1_I - e^1_F - e^1_M \\
\partial(e^1_J) &= e^0_B - e^0_A & \qquad \partial(e^2_R) &= e^1_M + e^1_L - e^1_J - e^1_K \\
\partial(e^1_K) &= e^0_C - e^0_B & & \\
\partial(e^1_L) &= e^0_C - e^0_D & & \\
\partial(e^1_M) &= e^0_D - e^0_A & \qquad \partial(e^3) &= -e^2_N + e^2_O + e^2_P - e^2_Q - e^2_R
\end{aligned}$$

These boundary formulae determine a sequence of linear homomorphisms.

$$\cdots \to 0 \xrightarrow{\partial_4} C_3(X,\mathbb{Q}) \xrightarrow{\partial_3} C_2(X,\mathbb{Q}) \xrightarrow{\partial_2} C_1(X,\mathbb{Q}) \xrightarrow{\partial_1} C_0(X,\mathbb{Q}) \ .$$

An important feature is that

$$\text{image}(\partial_{n+1}) \subseteq \ker \partial_n \qquad \text{for } n \geq 1.$$

The inclusion holds trivially for $n \geq 3$ and is readily verified for $n = 1, 2$. This example motivates the following definitions and lemma.

Definition A *chain complex* over a field (resp. ring) $\mathbb{K}$ is an infinite sequence of linear homomorphisms of vector spaces (resp. modules) over $\mathbb{K}$

$$\cdots \xrightarrow{\partial_{n+1}} C_{n+1} \xrightarrow{\partial_n} C_n \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} C_0$$

satisfying $\partial_n \partial_{n+1} = 0$ for all $n \geq 1$. We denote such a chain complex by (C_*, ∂_*) or simply by C_*.

Definition Let C_* be a chain complex over a field in which all C_n are finite dimensional, and all but finitely many of the C_n have dimension equal to 0. The *euler characteristic* of C_* is the number

$$\chi(C_*) = \sum_{n \geq 0} (-1)^n \dim(C_n) \quad .$$

Definition Let C_* be a chain complex over a field. Its *betti numbers* are

$$\beta_n(C_*) = \dim\left(\frac{\ker(\partial_n)}{\text{image}(\partial_{n+1})}\right)$$

where we take ∂_0 to be the zero homomorphism $\partial_0 \colon C_0 \to 0$. Thus, if all C_n are finite dimensional, we have

$$\beta_n(C_*) = \dim(\ker(\partial_n)) - \dim(\text{image}(\partial_{n+1})) \qquad (n \geq 0).$$

Lemma 7 *Let C_* be a chain complex over a field in which all C_n are finite dimensional, and all but finitely many of the C_n have dimension equal to 0. Then*

$$\chi(C_*) = \sum_{n\geq 0} (-1)^n \beta_n(C_*) \ .$$

Proof. Let α_n denote the dimension of the vector space C_n. We have

$$\begin{aligned}
& \beta_0(C_*) - \beta_1(C_*) + \beta_2(C_*) \ldots \\
= \quad & (\dim(\ker(\partial_0)) - \dim(\text{image}(\partial_1))) \\
& -(\dim(\ker(\partial_1)) - \dim(\text{image}(\partial_2))) \\
& +(\dim(\ker(\partial_2)) - \dim(\text{image}(\partial_3))) \quad \ldots \\
= \quad & \alpha_0 - \dim(\text{image}(\partial_1) \\
& -(\alpha_1 - \dim(\text{image}(\partial_1)) - \dim(\text{image}(\partial_2))) \\
& +(\alpha_2 - \dim(\text{image}(\partial_2)) - \dim(\text{image}(\partial_3))) \quad \ldots \\
= \quad & \alpha_0 + \alpha_1 - \alpha_2 \quad \ldots \\
= \quad & \chi(C_*) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square
\end{aligned}$$

Our convex polytope example above can be extended to produce a chain complex $C_*(X, \mathbb{K})$ for any CW-space X and any field (or ring) $\mathbb{K}$. Recall that a CW-space X is built by attaching maps

$\phi\colon E^n \to X$ and that we refer to each $e^n = \phi(U^n)$ as an *n-cell*. We set $\overline{e}^n = \phi(E^n)$ and say that e^{n-1} is *in the boundary* of e^n if $e^{n-1} \subset \overline{e}^n$.

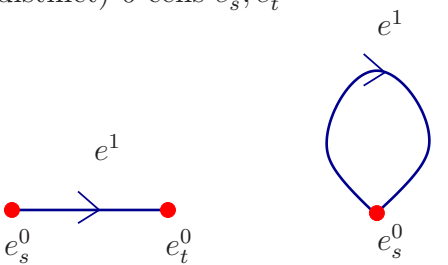
We take $C_n(X, \mathbb{K})$ to be the vector space (or free module) over $\mathbb{K}$ of dimension α_n equal to the number of n-cells in X. We allow the possibility of infinite α_n and, abusing notation slightly, identify the n-cells e_i^n with elements of some basis for this vector space (or module). The definition of the linear *boundary* homomorphisms

$$\partial_n\colon C_n(X, \mathbb{K}) \to C_{n-1}(X, \mathbb{K}), e_i^n \mapsto \sum_{1 \le j \le \alpha_{n-1}} b_{ij}^n e_j^{n-1}$$

is somewhat involved. An approach to a general and rigorous description of the homomorphism is outlined briefly in Chapter 2. Below we provide a description for certain (computationally important) special cases. With suitably defined boundary homomorphisms $(C_*(X, \mathbb{K}), \partial_*)$ is a chain complex which we refer to as the *cellular chain complex* of X over $\mathbb{K}$.

To define the first homomorphism $\partial_1\colon C_1(X, \mathbb{K}) \to C_0(X, \mathbb{K})$ it suffices to give each 1-cell an orientation by placing an arrow head on it in some (arbitrarily chosen) direction. Any 1-cell e^1 is

attached to two (not necessarily distinct) 0-cells e^0_s, e^0_t



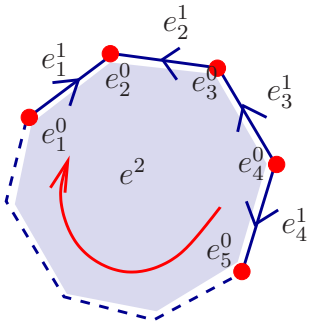
and we set

$$\partial_1(e^1) = e^0_t - e^0_s$$

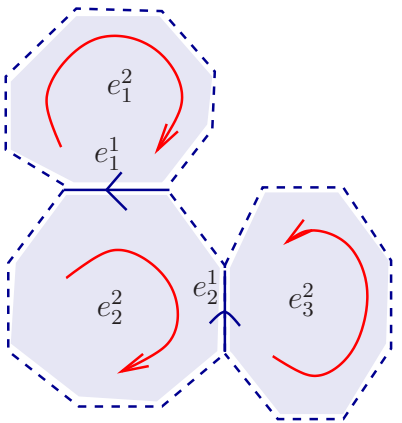
where the orientation on e^1 is from e^0_s to e^0_t.

A simplified context for the description of higher boundary homomorphisms is that of *regular* CW-spaces. These are defined as having the property that each attaching map restricts to a homeomorphism $\phi\colon E^n \to \overline{e}^n$. An important source of regular CW-spaces is those spaces arising from convex polytopes by removing inetriors of k-faces. In particular, the geometric realizations of finite simplicial complexes and cubical complexes are regular.

For a regular CW-space $b^n_{ij} = 0$ unless e^{n-1}_j lies in the boundary of e^n_i, in which case $b_{ij} = 1$ or $b_{ij} = -1$ depending on an issue of "orientation". In dimensions $n = 1, 2, 3$ the signs can be assigned by placing orientation arrows as in the above convex hull example. Then $\partial_{n-1}\partial_n(e^n) = 0$ for any basis elements in dimension $n = 2, 3$ because we can see that in the expression for $\partial_{n-1}\partial_n(e^n)$ each $(n-2)$-cell occurs precisely twice and with opposite signs.



$$\begin{aligned}\partial_1(\partial_2(e^2)) &= \partial_1(\ldots + e_1^1 - e_2^1 - e_3^1 + e_4^1 + \ldots)\\ &= \ldots - e_1^0 + e_2^0 - e_2^0 + e_3^0 - e_3^0\\ &\quad + e_4^0 - e_4^0 + e_5^0 \ldots\end{aligned}$$



$$\begin{aligned}\partial_2(\partial_3(e^3)) &= \partial_2(\ldots + e_1^2 \ldots + e_2^2 \ldots - e_3^2 \ldots)\\ &= \ldots + e_1^1 \ldots - e_1^1 \ldots - e_2^1 \ldots + e_2^1 \ldots\end{aligned}$$

The signs in ∂_n can of course be ignored if we work over the field $\mathbb{K} = \mathbb{F}_2$ of two elements. (Orientation is treated in more detail in Chapter 2.)

An approach to a proof of the following result is outlined in Chapter 2. Together with Lemma 7 this result implies Theorem 3.

Theorem 8 *If X, Y are homotopy equivalent CW-spaces then*

$$\beta_n(C_*(X,\mathbb{K})) = \beta_n(C_*(Y,\mathbb{K}))$$

for each $n \geq 0$ and any field $\mathbb{K}$.

Definition The *betti numbers* of a CW-space X can can be defined as

$$\beta_n(X) = \beta_n(C_*(X,\mathbb{Q}))$$

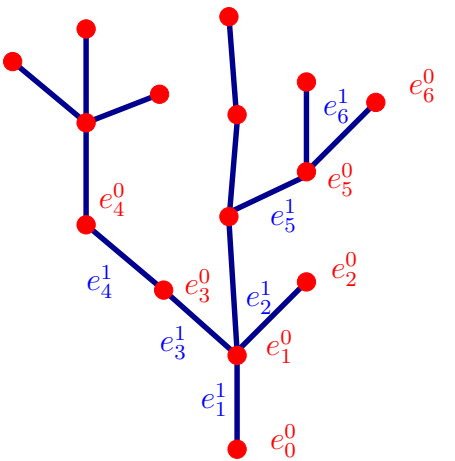
for $n \geq 0$. More generally we define

$$\beta_n(X,\mathbb{K}) = \beta_n(C_*(X,\mathbb{K})).$$

This definition is consistent with our previous definition of $\beta_0(X)$ thanks to the following.

Proposition 9 *A CW-space X has $\beta_0(C_*(X,\mathbb{K}))$ path components.*

Proof. It is immediate from the definition of the boundary homomorphism ∂_1 that $\beta_0(C_*(X,\mathbb{K})) = \beta_0(C_*(X'',\mathbb{K})) + \beta_0(C_*(X'',\mathbb{K}))$ whenever $X = X' \cup X''$ is the disjoint union of two CW-spaces. It thus suffices to show that $\beta_0(C_*(X,\mathbb{K})) = 1$ for a path-connected CW-space. So let us suppose that the CW-space X is path-connected. The 0-cells and 1-cells of X form a connected graph. We can choose a maximal tree in this graph, and label the tree's vertices $e^0_0, e^0_1, e^0_2, \ldots$ and edges $e^1_1, e^1_2, e^1_3, \ldots$ such that edge e^1_k has vertices e^0_k and $e^0_{k'}$ where $k' < k$.



For any $v > 0$ the vectors $\partial_1(e^1_1), \ldots, \partial_1(e_v)$ clearly span a v-dimensional subspace of the $(v+1)$-dimensional span of $e^0_0, \ldots, e^0_v$. It follows that $\beta_0(C_*(X,\mathbb{K})) = 1$ for any connected space X with a finite number $v > 0$ of vertices. If there are infinitely many vertices in the tree then, for any non-zero vector $z \in C_0(X,\mathbb{K})$ there is a v such that z lies in the span of $e^0_0, \ldots, e^0_v$. So it is just an

exercise to see that $\beta_0(C_*(X, \mathbb{K})) = 1$ even when the tree has infinitely many vertices. □

Note that path-connectedness is fairly clearly a homotopy invariant. Thus Proposition 9 proves Theorem 8 in the case $n = 0$.

A chain complex C_* of finite dimensional vector spaces can be represented on a computer as a sequence of matrices $\Delta_1, \Delta_2, \ldots$ representing the boundary homomorphisms $\partial_1, \partial_2, \ldots$. Computation of the betti numbers $\beta_n(C_*)$ is an easy computation

$$\beta_n(C_*) = (\text{number of columns in } \Delta_n) - \text{rank}(\Delta_n) - \text{rank}(\Delta_{n+1})$$

involving the ranks of matrices. The rank of a matrix can be computed as the number of vectors in some basis for the row span of the transposed matrix. For completeness we recall from elementary linear algebra the classical method for finding a basis of the row span of a matrix A.

Definition The *head index* of the ith row (a_{i*}) of a matrix $A = (a_{ij})$ is the least j for which $a_{ij} \neq 0$. A zero row is deemed to have no head index. The matrix is said to be in *semi-echelon form* if no two rows have the same head index.

To find a basis B for the row span of a matrix A over a field we simply use row operations to

convert A to a *semi-echelon matrix* B. A more algorithmic description follows.

Semi-Echelon Matrix Algorithm

Input: a matrix $A = (a_{ij})$ (which will be destroyed).
Output: a matrix B in semi-echelon form with the same row span as A.
Procedure:
initialise $B := A$;
while $B = (b_{ij})$ is not in semi-echelon form
    find a row $(b_{i_0 *})$ with least head index j for which other rows also have head index j;
    subtract suitable multiples of row $(b_{i_0 *})$ from the other rows with head index j to ensure that row $(b_{i_0 *})$ becomes the only row with head index j;
remove any zero rows from B;
return B;

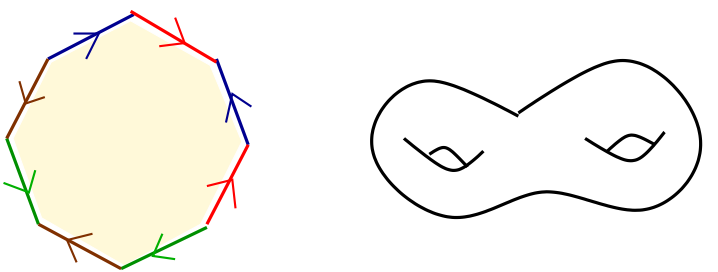
We are now in a position to revisit the analysis of the symmetric matrix S of distances between 400 sample points considered in Section 1.4. At the end of that section we produced a picture of the graph of a Vietoris-Rips complex $X_S(t)$ that suggests the presence of two "1-dimensional holes" in the data. The following computation of the betti numbers $\beta_1(X_S(t))$ for a range of values of t also

suggests the presence of two such holes.

```
gap> Read("symmetricMatrix.txt"); # This creats the matrix S
gap> ListOfBettiOnes:=[];;
gap> for t in 2*[0..22] do
> C:=RipsChainComplex(S,t);
> Add(ListOfBettiOnes,Bettinumbers(C,1));
> od;
gap> Print(ListOfBettiOnes,"\n");
[ 0, 0, 4, 36, 20, 12, 6, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2 ]
```

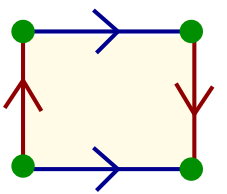
Exercises

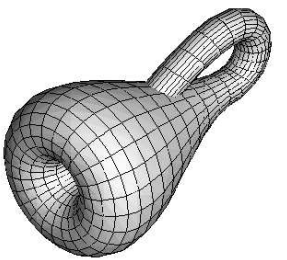
1. Identifying all pairs of similarly coloured directed edges of the following octagon is one method for constructing the surface T_2 of genus 2 (i.e. the surface of a donought with two holes).



Use the cell structure on T_2 arising from the above construction to calculate the chain complex $C_*(T_2, \mathbb{Q})$ and the betti numbers $\beta_n(T_2)$ for $n \geq 0$.

2. Identifying all pairs of simlarly coloured directed edges of the following rectangle is one method for contructing the surface K known as the *Klein bottle*.



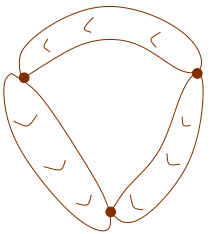


Use the cell structure on K arising from the above construction to calculate the chain complex $C_*(K, \mathbb{Q})$ and also the chain complex $C_*(K, F_2)$ over the field of two elements. Calculate $\beta_n(K, \mathbb{Q})$ and $\beta_n(K, \mathbb{F}_2)$ for $n \geq 0$.

3. Let X be the surface of the 3-simplex endowed with a CW-structure involving four triangles, six edges and four vertices. Determine the cellular chain complex $C_*(X, \mathbb{Q})$. Use the boundary matrices in this chain complex to determine $\beta_n(X)$ for $n > 0$.

4. Determine the Euler characteristic $\chi(X)$ for three sausage shaped (hollow) 2-spheres con-

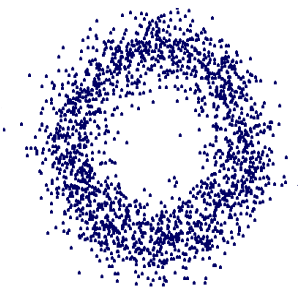
nected at three points as shown.



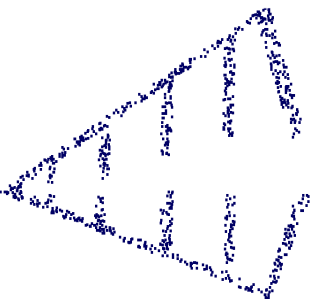
Also determine the betti numbers $\beta_n(X)$ for $n > 0$.

1.6 Persistence of betti numbers

There is a difficulty in interpreting a sequence of betti numbers $\beta_n(X_1), \beta_n(X_2), \beta_n(X_3), \ldots$ arising from a chain of subspaces $X_1 \subset X_2 \subset X_3 \subset \ldots$. To illustrate the difficulty we consider two sets of data points in the plane represented by the following images.



Data set A



Data set B

We can represent these data sets as two pure cubical complexes A, B and then form the chains $A \subset T(A) \subset T^2(A) \subset \ldots$ and $B \subset T(B) \subset T^2(B) \subset \ldots$ where $T^n(X)$ denotes an n-fold thickened

space. The following GAP session computes the first betti numbers of $T^n(A)$ for $0 \leq n \leq 70$.

```
gap> A:=ReadImageAsPureCubicalComplex("datasetA.eps",500);;
gap> Seq:=[A];;
gap> for i in [1..70] do
> Add(Seq,ThickenedPureCubicalComplex(Seq[Length(Seq)]));
> od;
gap> FirstBettis:=[];;
gap>  for T in Seq do
> Add(FirstBettis,Bettinumbers(T,1));
> od;
FirstBettis;
[ 10, 23, 26, 29, 10, 7, 2, 0, 0, 1, 1, 1, 2, 2, 3, 2, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1,
  1, 2, 2, 3, 3, 4, 2, 2, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0 ]
```

Note that $\beta_1(T^n(A)) = 1$ for many values of n, and that the circle also has $\beta_1(S^1) = 1$. This is certainly consistent with the view that data set A has been sampled from a space of the homotopy type of S^1. However, an analogous computation can be performed for data set B yielding the following sequence of first betti numbers.

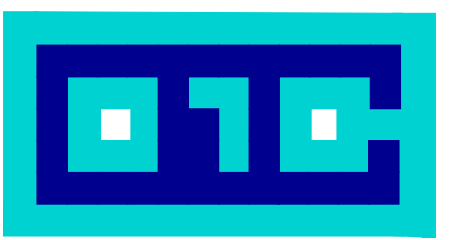
```
[ 11, 17, 24, 12, 7, 2, 0, 0, 1, 1, 3, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 3, 1, 1, 0, 1, 2, 3, 1, 1,
  1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ]
```

Again $\beta_1(T^n(B)) = 1$ for many values of n. But it is hard to imagine that data set B was sampled from any space X with $\beta_1(X) = 1$.

The difficulty in interpreting a sequence of betti numbers is that the process of "thickening" a space X to an inclusion $X \subset TX$ could result in all of the following (imprecisely stated) phenomena:

1. A hole in X might get filled in and not correspond to any hole in TX.
2. The thickening proces might create a hole in TX that was not present in X.
3. A hole in X might persist to TX.

X



TX

The illustration shows a cubical complex X and its cubical thickening TX with both $\beta_1(X) = 2$ and $\beta_1(TX) = 2$. One hole present in X gets filled in, one new hole is created in TX, and one hole persists from X to TX.

We need to explain how the intuitive notion of persistence of holes can be formalized into a notion of "persistence of betti numbers". The "persistent betti numbers" arising from a chain of subspaces $X \subset TX \subset T^2X \subset \dots$ will contain more robust information about the topology underlying the sequence.

Definition By the nth-*homology* of a chain complex C_* we mean the quotient module (or vector space)

$$H_n(C_*) = \frac{\ker(\partial_n \colon C_n \to C_{n-1})}{\text{image}(\partial_{n+1} \colon C_{n+1} \to C_n)}.$$

An advantage of dealing with homology of chain complexes rather than betti numbers is that we can consider linear homomorphisms $H_n(C_*) \to H_n(C'_*)$.

Let X be a CW-space and Y a CW-subspace of X. (Thus every cell of Y is a cell of X.) For each $n \geq 0$ there is a canonical (injective) linear homomorphism $\iota_n \colon C_n(Y, \mathbb{K}) \to C_n(X, \mathbb{K})$, and the following diagrams commute:

$$\begin{array}{ccc} C_{n+1}(Y,\mathbb{K}) & \xrightarrow{\iota_{n+1}} & C_{n+1}(X,\mathbb{K}) \\ \downarrow{\scriptstyle \partial_{n+1}} & & \downarrow{\scriptstyle \partial_{n+1}} \\ C_n(Y,\mathbb{K}) & \xrightarrow{\iota_n} & C_n(X,\mathbb{K}) \end{array}$$

This motivates the following.

Definition Let (C_*, ∂_*) and (C'_*, ∂'_*) be two chain complexes. A *chain map* $\phi_*: C_* \to C'_*$ is a collection of linear homomorphisms $\phi_n: C_n \to C'_n$ such that the diagram

$$\begin{array}{ccc} C_{n+1} & \xrightarrow{\phi_{n+1}} & C'_{n+1} \\ \downarrow{\scriptstyle \partial_{n+1}} & & \downarrow{\scriptstyle \partial'_{n+1}} \\ C_n & \xrightarrow{\phi_n} & C'_n \end{array}$$

commutes for all $n \geq 0$.

Proposition 10 *A chain map $\phi_*: C_* \to C'_*$ induces homomorphisms*

$$H_n(\phi_*): H_n(C_*) \to H_n(C'_*) \ ,$$

$$v + \text{image}(\partial_{n+1}) \mapsto \phi_n(v) + \text{image}(\partial'_{n+1})$$

for all $n \geq 0$.

Proof. The commutativity equations $\phi_n \partial_{n+1} = \partial'_{n+1} \phi_{n+1}$ imply that

$$\phi_n(\ker(\partial_n)) \subseteq \ker(\partial'_n) \ ,$$

$$\phi_{n+1}(\text{image}(\partial_{n+1})) \subseteq \text{image}(\partial'_{n+1}).$$

□

Any sequence of inclusions of CW-subspaces

$$X_1 \subset X_2 \subset X_3 \subset \ldots \subset X_t \qquad (*)$$

induces a sequence of chain maps

$$C_*(X_1, \mathbb{K}) \hookrightarrow C_*(X_2, \mathbb{K}) \hookrightarrow C_*(X_3, \mathbb{K}) \hookrightarrow \ldots \hookrightarrow C_*(X_t, \mathbb{K})$$

which in turn induce a sequence of homology homomorphism

$$H_n(X_1, \mathbb{K}) \to H_n(X_2, \mathbb{K}) \to H_n(X_3, \mathbb{K}) \to \ldots \to H_n(X_t, \mathbb{K})$$

for each $n \geq 0$.

Definition For a sequence of inclusions $(*)$, a field $\mathbb{K}$ and each $n \geq 0$ we define the nth *persistence matrix* over $\mathbb{K}$ to be the $t \times t$ integer matrix $P_n = (p_{ij})$ where

$$p_{ij} = \dim(\ \text{image}(H_n(X_i, \mathbb{K}) \longrightarrow H_n(X_j, \mathbb{K}))\ )$$

for $i \leq j$, and

$$p_{ij} = 0$$

for $i > j$.

The following GAP session computes the first persistence matrix P_1^A over the field $\mathbb{F}_2$ of two elements for the sequence $T^{12}(A) \subset T^{16}(A) \subset T^{20}(A) \subset \ldots \subset T^{40}(A)$ of forty thickenings of our data set A.

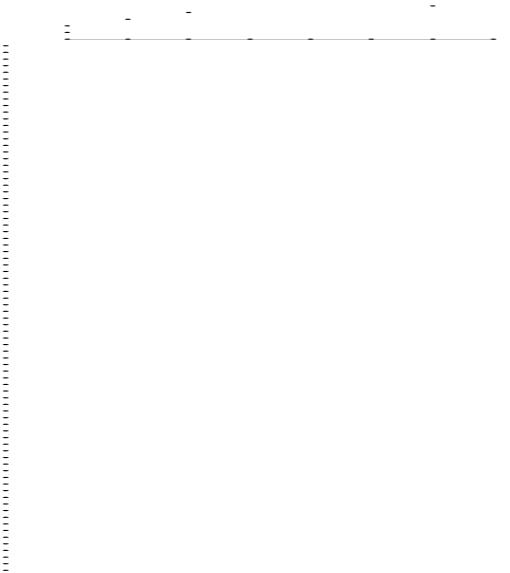
```
gap> A:=ReadImageAsPureCubicalComplex("dataset.eps",500);;
gap> Seq:=[A];;
gap> for i in [1..40] do
> A:=ThickenedPureCubicalComplex(A);
> if i in 4*[3..10] then Add(Seq,A); fi;
> od;
gap> P:=PersistentHomologyOfPureCubicalComplex(Seq,1,2);;
```

Similar commands compute the analogous persistence matrix P_1^B for our data set B.

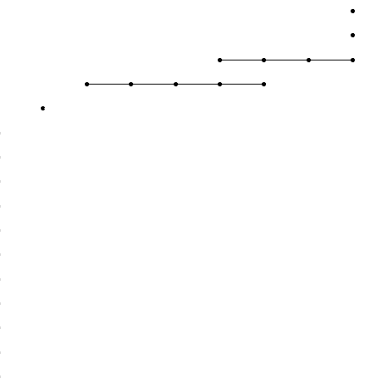
$$P_1^A = \begin{pmatrix} 80 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 2 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad P_1^B = \begin{pmatrix} 11 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

The information in a $t \times t$ persistence matrix $P_n = (p_{ij})$ can be conveniently displayed as a β_n-*bar code* involving t columns of dots, and horizontal lines joining certain dots in columns j and $j+1$.

We place precisely p_{jj} dots in column j, and join precisely $p_{j,j+1}$ of these dots by a horizontal line to dots in the next column. The horizontal lines between successive columns are positioned so as to form p_{ij} continuous lines between the ith and jth columns.



Mod 2 bar code for P_1^A



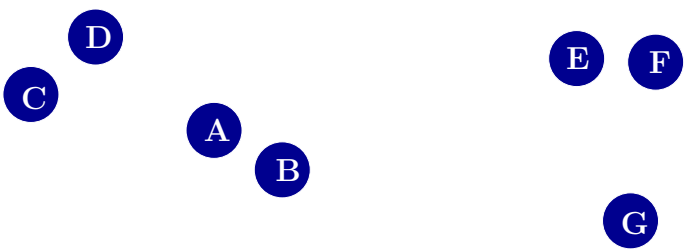
Mod 2 bar code for P_1^B

The above β_1-bar codes were computed using GAP's `BarCodeDisplay` command. The bar code for data set A suggests a single persistent "hole" during 40 thickenings of the data. In contrast, the

bar code for data set B suggests two non-simultaneous and slightly less persistent "holes" during the same range of thickenings.

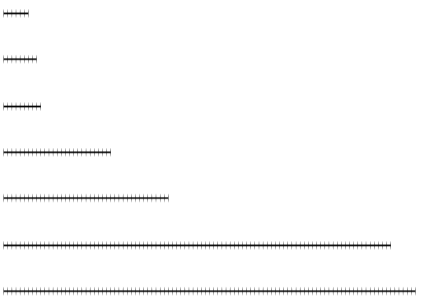
Persistence matrices P_n and β_n-bar codes provide a convenient record of the birth, death and persistence of certain topological features during a sequence of inclusions of CW-spaces. For the special case of $n = 0$ in the special situation where the sequence of inclusions corresponds to a sequence of thickenings of some data set, there exists a more traditional visual record known as a *phylogenetic tree* or *dendrogram*. We assume that whatever process is used for thickening a CW-space X to an inclusion $X \subset TX$, each path component of the thickened space TX contains at least one path component of X. A careful analysis of the homomorphisms $H_0(X, \mathbb{K}) \to H_)(TX, \mathbb{K})$ then shows that each row of the resulting persitence matrix P_0 decreases (non-strictly) from left to right. Equivalently, each bar in the β_0-bar code starts in the first column.

An example of a β_0-bar code can be computed from the following data set

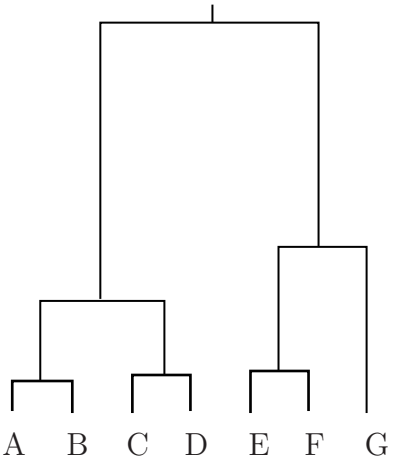


using single linkage, Chebychev metric clustering on the pure cubical complex determined by our

digital image of the data. GAP produces the following β_0-bar code.



The following is a manually produced *dendrogram* for the same data.



In both the bar code and the dendrogram the persistence of a cluster is represented by its length. The dendrogram contains extra information about the manner in which clusters form.

Dendrograms are much used in areas such as bioinformatics; an early example appears in Charles Darwin's notebooks. The notion of bar code was introduced recently in a paper [12] of A. Zomorodian and G. Garlsson. The notion of persistent homology was developed around the turn of the millenium by a number of independent groups. One of the developers, H. Edelsbrunner, describes the origins of the notion in [3].

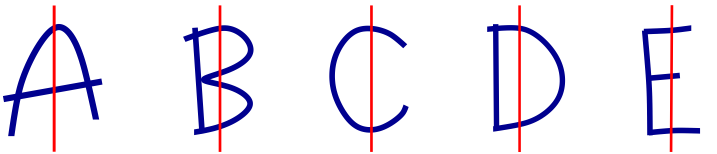
Exercises

1. Sketch the β_0 and β_1 bar codes (over $\mathbb{Q}$) that would corresponding to a single linkage, Chebychev metric cluster analysis of the following data points (when represented as a pure cubical complex).

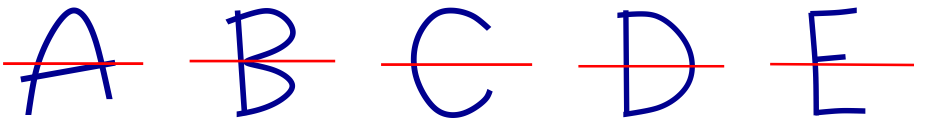
.

2. Consider the following five digital images.

For each image X we can construct an image X^v by adding a vertical line through the centre of gravity.



We can also construct an image X^h by adding a horizontal line through the centre of gravity.



The five images were read into GAP in some random order and converted to pure cubical complexes X_i $(1 \leq i \leq 5)$. For eaxh $Y = X_i, X_i^v, X_i^h$ a sequence of thickenings $Y \subset TY \subset T^2Y \subset \ldots \subset T^{19}Y$ was constructed and the sequences of betti numbers $\beta_1(T^nY)$ are recorded below. Determine the image to which each X_i corresponds.

$\beta_1(T^n(X_1))$	$[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]$
$\beta_1(T^n(X_1^v))$	$[2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,0,0,0]$
$\beta_1(T^n(X_1^h))$	$[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]$

$\beta_1(T^n X_2)$	$[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2]$
$\beta_1(T^n(X_2^v))$	$[2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,0,0,0]$
$\beta_1(T^n(X_2^h))$	$[2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]$
$\beta_1(T^n X_3)$	$[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]$
$\beta_1(T^n(X_3^v))$	$[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]$
$\beta_1(T^n(X_3^h))$	$[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1]$
$\beta_1(T^n X_4)$	$[2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2]$
$\beta_1(T^n(X_4^v))$	$[5,5,4,5,4,4,4,4,4,4,4,4,3,3,3,1,0,0,0,0]$
$\beta_1(T^n(X_4^h))$	$[2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2]$
$\beta_1(T^n X_5)$	$[1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1]$
$\beta_1(T^n(X_5^v))$	$[2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,1]$
$\beta_1(T^n(X_5^h))$	$[2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2]$

3. The following three images

were realised as pure cubical complexes X_1, X_2, X_3 in some random order. These were then thickened 15 times so that $\beta_1(T^{15}X_i) = 1$. The following complements $\overline{T^{15}X_i}$ were then

computed.

The table gives the size (i.e. number of blue pixels) in the path components of $\overline{T^{15}X_i}$. Determine which image corresponds to each X_i.

i	sizes of path components of X_i
1	$[79, 659, 12242, 2388, 31086]$
2	$[1733, 3589, 41788, 4938, 3101]$
3	$[1678, 2969, 46925, 1387, 3576, 339, 16]$

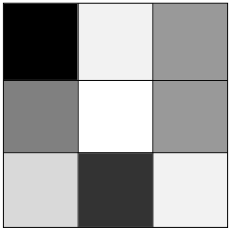
(The complement of X was taken as the difference between X and the smallest upright solid rectangle containing X. Feature recognition could be improved by taking instead the difference between X and its convex hull.)

1.7 Natural image statistics

Several recent applications of bar codes are described by Gunnar Carlsson in [1]. One of these, which we recall here, is based on the paper [2] and addresses the question: what distinguishes a typical digital photograph from a random collection of pixels?

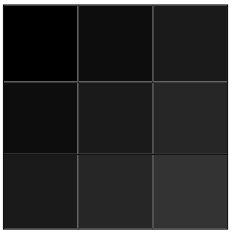
A digital image is an $m \times n$ array of pixels. Each pixel is coloured, and the spectrum of colours can be represented by an interval of real numbers. A digital image can thus be represented as a point in $\mathbb{R}^{m\times n}$. However, a random vector in $\mathbb{R}^{m\times n}$ is highly unlikely to correspond to any natural image.

Do those points of $\mathbb{R}^{m\times n}$ representing natural images form a submanifold with describable topological features? This question is rather ambitious. D. Mumford, A. Lee and K.Pedersen [8] have suggested investigating the more tractable space X consisting of all 3×3 patches taken from all possible natural images.



Such patches are represented by vectors in $\mathbb{R}^{3\times 3} = \mathbb{R}^9$.

A random sample of 3×3 patches from natural images will contain a high proportion of *low contrast* patches.



As such patches don't contain much structure, it is desirable to focus on samples of *high contrast* patches. For $v = (x_1, \ldots, x_9) \in \mathbb{R}^9$ define $\log(v) = (\log|x_1|, \ldots, \log|x_9|)$. For a suitable 9×9 symmetric matrix A, the *D-norm*

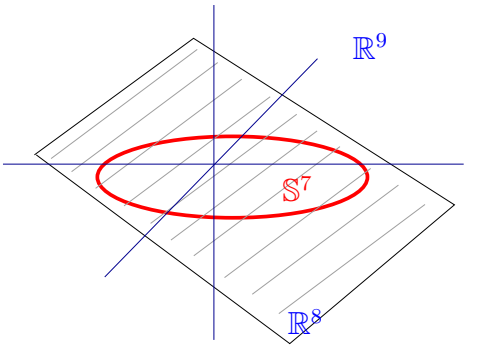
$$||v||_D = \sqrt{\log(v) A \log(v)^t}$$

was defined in [8] as a measure of the contrast of a patch. (Details of the matrix A are not so important for our discussion.)

A finite database $\mathcal{M}$ of high-contrast patches in $\mathbb{R}^9$ was compiled in [8] as follows. A large number of natural digital images were chosen and, for each image, 5000 random patches were selected. Of these 5000 patches per image, only the 20% with the highest D-norm were retained. As brightness of an image should not be too important for the topology of X, each database patch $v = (x_1, \ldots, x_9) \in \mathbb{R}^9$ was replaced by

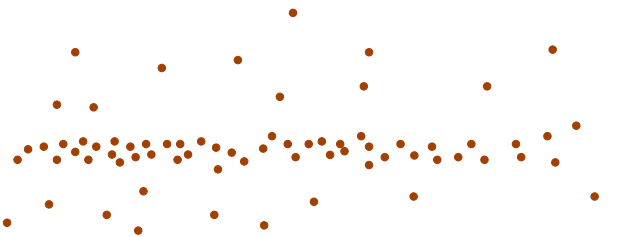
$$\overline{v} = (x_1 - \overline{x}, \ldots, x_9 - \overline{x})$$

with $\overline{x}$ the average value of the x_i. The database $\mathcal{M}$ thus lies in an 8-dimensional subspace of $\mathbb{R}^9$. As the contrast of an image should not be too important, each database patch $v \in \mathbb{R}^8$ was replaced by $(1/||v||_D)\, v$. So the database $\mathcal{M}$ actually lies in a 7-dimensional ellipsoid.



The goal is to understand how the database points lie in this ellipsoid.

Experimentation suggested that the density of the patches varies significantly over the ellipsoid.



In order to capture the essential topology underlying $\mathcal{M}$ some method is needed for discarding patches from low density regions. For each integer $k \geq 0$ and patch v in the database $\mathcal{M}$, the *k-codensity* of v was defined in [2] as

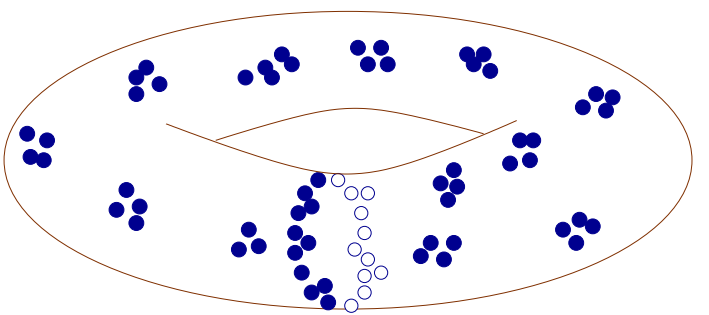
$$\delta_k(v) = d(v, u_k)$$

where u_k is the k-th nearest patch to v in $\mathcal{M}$, and d is our metric on $\mathcal{M}$. The idea is: points of low k-codensity come from dense regions of the sample.

For any choice of $k \geq 0$ and $0 < T < 100$ one can consider the following subsamples of patches:

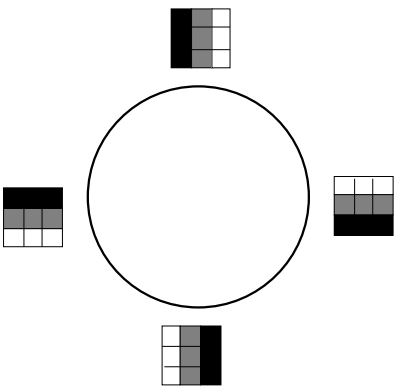
$$\mathcal{M}[k, T] = \{v \in \mathcal{M} \mid \delta_k(v) \text{ lies in the } T\% \text{ lowest values of } \delta_k \text{ in } \mathcal{M}\}$$

(To get a feel for these subspamples it might help to suppose, very hypothetically, that the points in $\mathcal{M}$ were distributed on a torus with the following kind of variable density.



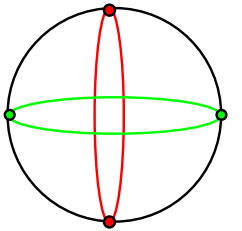
For very low values of k bar codes for the subsample $\mathcal{M}[k, T]$ would likely suggest it is taken from a space with $\beta_1 = 2$. For higher values of k the bar codes might suggest the subsample comes from a space with $\beta_1 = 1$.)

A β_1 bar code was computed for $\mathcal{M}[300, 30]$. It suggested that this particular subsample lies in a space X with $\beta_1(X) = 1$. One simple explanation is that X might have the homotopy type of a circle. This explanation, and the patches found in $\mathcal{M}[300, 30]$, are consistent with the following picture.

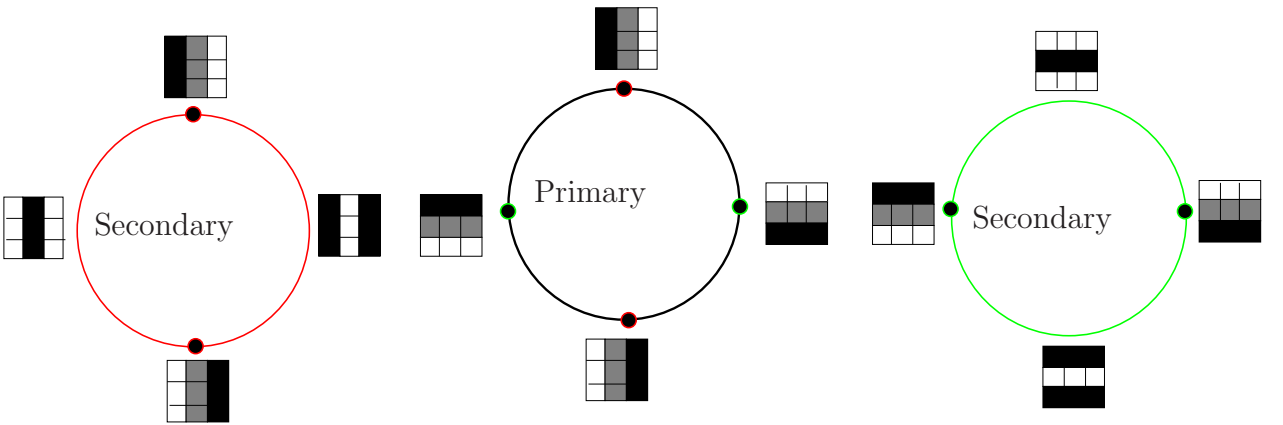


A β_1 bar code suggested that $\mathcal{M}[15, 30]$ lies in a space X with $\beta_1(X) = 5$. A simple explanation proposed in [2] is that X consists of three circles, coloured say black, red and green, where the black and red circles have two common points, the black and green circles have two common points, and

the red and green circles have no common point.

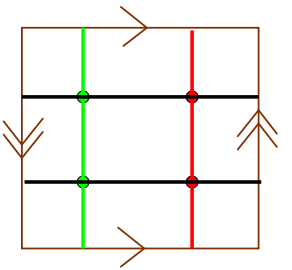


This three circle model fits well with the data.



A bar code analysis suggests that $\mathcal{M}[100, 10]$ is sampled from a space X with $\beta_0(X) = 0, \beta_1(X) =$

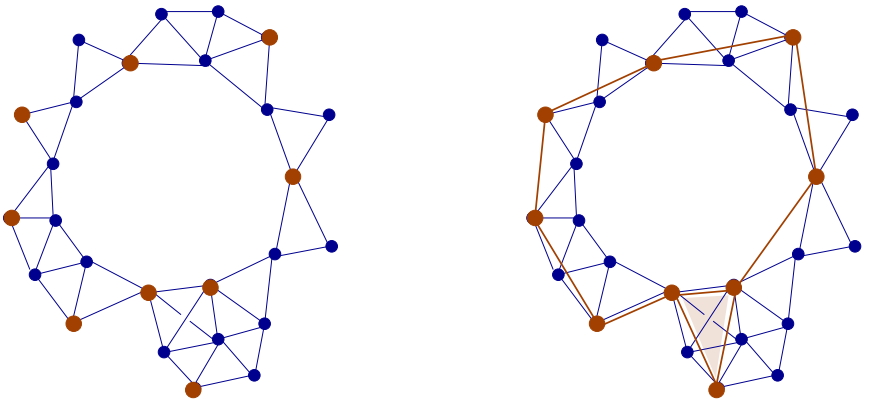
$1, \beta_2(X) = 1$. The Klein bottle has such Betti numbers. A possible explanation proposed in [2] is that the three circle model should be viewed as being embedded in the Klein bottle as follows.



The above summary of the analysis of natural images has avoided an important technical difficulty. How does one compute a bar code for a large sample of data points in $\mathcal{M}$? Even to store the symmetric matrix $S = (s_{ij})$ of distances between data points is problematic for large samples. The Vietoris-Rips complex is totally impractical for large samples.

One practical computational tool employed in [2] is the *witness complex* . Suppose we have a sample M of data points and a suitably chosen subset L whose elements are called *landmark points*. We store the $|M| \times |L|$ matrix S_L of distances from each point in M to each landmark point. For $t \geq 0$ the *witness complex* $W(S_L, t)$ is a simplicial complex whose vertices are the landmark points L. There is one k-simplex for each set Δ of $k+1$ landmark points with the following property: given any pair of vertices $u, v \in \Delta$ there is a point $c \in M$ which is within a distance t of both u and v.

As an illustration let us consider the following graph with 26 vertices. Suppose that its vertices correspond to data points for which we have the symmetric matrix S of pairwise distances, and that for some fixed t a blue edge has been inserted between a pair of vertices if the corresponding data points lie within a distance t of one another.



graph witness complex

Nine of the vertices have been coloured brown, and we regard these as forming a set L of landmark vertices. The witness complex $W(S_L, t)$ is illustrated on the right in brown. It has just 9 vertices, 10 edges and one 2-simplex. By contrast, the Vietoris-Rips complex $R(S, t)$ has 26 vertices, 46 edges, 22 2-simplices and one 3-simplex.

The above example illustrates that the witness complex can be much smaller than the Vietoris-Rips

complex. It also illustrates that the geometric realizations of the two simplicial complexes can be homotopy equivalent for suitably chosen landmark vertices.

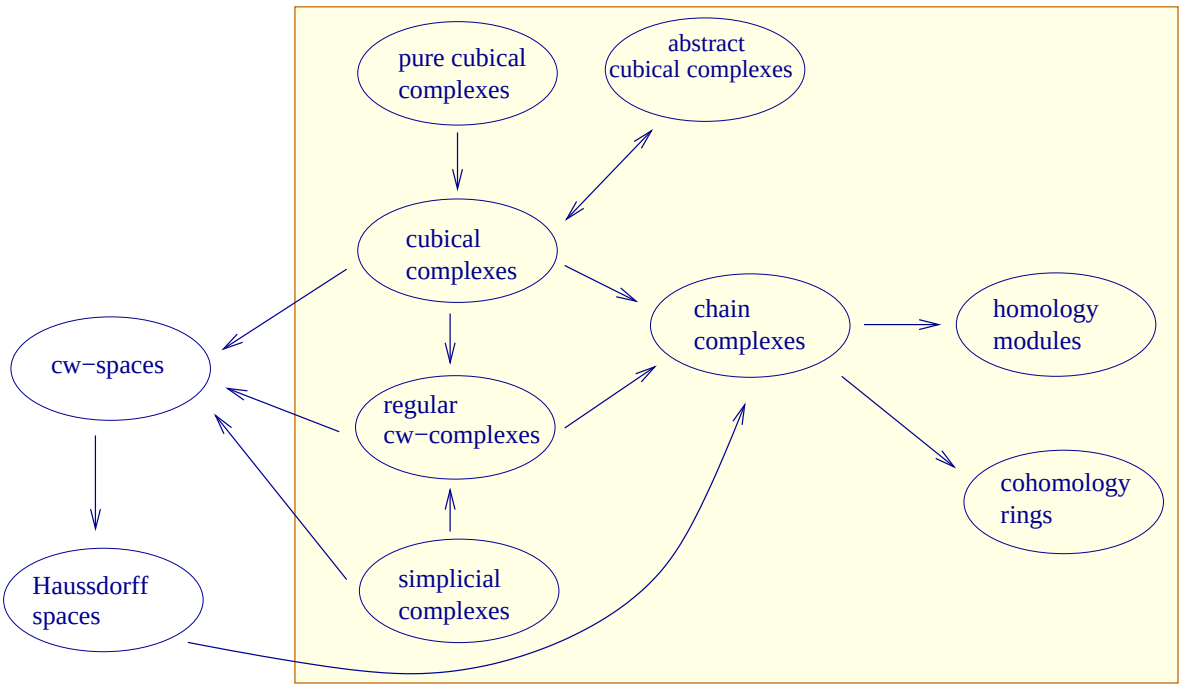
For $t \leq t'$ we have inclusions $R(S,t) \subset R(S,t')$, $W(S_L,t) \subset W(S_L,t')$ of simplicial complexes (meaning that any simplex of $R(S,t)$ is also a simplex of $R(S,t')$, and any simplex of $W(S_L,t)$ is also a simplex of $W(S_L,t')$). These inclusions induce obvious chain maps on the corresponding cellular chain complexes. Thus both the Vietoris-Rips complex and the witness complex can in principle be used to compute persistence matrices and bar codes. The witness complex is more often used in practice.

Chapter 2

Categories, functors and homology

2.1 Introduction

Algebraic topology involves a range of topological stuctures, such as those named in the following diagram, together with procedures or constructions for passing from one structure to another. Some



of these, such as those lying within the box, are fairly readily implemented on a computer. Others are more of theoretical importance. Much of the above diagram was encountered in Chapter 1. The present chapter provides fuller details, with particular emphasis on that portion of the diagram within the box. The language of *categories* and *functors* is useful for this task.

Categories and functors were introduced in a paper of 1945 by S. Eilenberg and S. MacLane as a language for expreseeing topological calculations. Some decades later MacLane [11] wrote:

> Given the conceptual background which I have been describing, we took the next following step of defining category and functor in our next joint paper [4] which we entitled simply "General theory of natural equivalences" - although it really began with categories and functors. It was perhaps a rash step to introduce so quickly such a sweeping generality - an evident piece of what was soon to be called "general abstract nonsense." One of our good friends (an admirer of Eilenberg) read the paper and told us privately that he thought the paper was without any content. Eilenberg took care to see to it that the editor of the *Transactions* sent the manuscript to a young referee (perhaps one who might be gently bullied). The paper was accepted by *Transactions*. I have sometimes wondered what could have happened had the same paper been submitted by a couple of wholly unknown authors. At any rate, we did think it was good, and that it provided a handy language to be used by topologists and others, and that it offered a conceptual view of parts of mathematics, in some way analogous to Felix Klein's "Erlanger programme." We did not then regard it as a field for further research efforts, but just as a language and an orientation - a limitation which we followed for a dozen years

> or so, till the advent of adjoint funtors.
>
> ...
>
> The use of categories as a language is well illustrated by the development of axiomatic homology theory. About 1940, the multiplicity of homology theories (simplicial, singular, Čech, Vietoris, Alexander, ...) seemed confusing. Then Eilenberg and Steenrod introduced their axioms, including the central one asserting that homology is a functor on (a category of) topological spaces to abelian groups. This could have been stated without the words or language "functor" and "category," but Steenrod in conversation emphasized the importance of these concepts. He said that the Eilenberg-MacLane paper on categories and functors had a more significant impact on him than any other research paper; other papers contributed results, while this paper changed his way of thinking.

In subsequent chapters we shall see that the conceptual view offered by the categorical language is not only applicable to parts of mathematics. It is of practical importance when implementing topological and homological algorithms (and no doubt other kinds of algorithms) on a computer. It provides a modularity that seems to guarantee readable and extensibile code without any significant cost to computational efficiency.

2.2 Definitions of categories and functors

A *category* $\mathcal{C}$ consists of:

- a collection $Ob(\mathcal{C})$ of things called *objects*,
- for each pair of objects $A, B \in Ob(\mathcal{C})$ a (possibly empty) collection $Mor_{\mathcal{C}}(A, B)$ of things called *morphisms* or *arrows* from A to B,
- a function
$$Mor_{\mathcal{C}}(B, C) \times Mor_{\mathcal{C}}(A, B) \longrightarrow Mor_{\mathcal{C}}(A, C),$$
$$(f, g) \mapsto f \circ g$$
called *composition* which is defined for each triple of objects $A, B, C \in Ob(\mathcal{C})$ admitting at least one arrow from A to B and at least one arrow from B to C.

The axioms for a category are:

1. (*Associativity*) The equation
$$(f \circ g) \circ h = f \circ (g \circ h)$$
holds for all morphisms f, g, h for which $(f \circ g) \circ h$ or $f \circ (g \circ h)$ is defined.

2. (*Identity*) For every object $B \in Ob(\mathcal{C})$ there is a morphism $1_B \in Mor_{\mathcal{C}}(B, B)$ such that $1_B \circ f = f$ and $g \circ 1_B = g$ for any morphisms $f \in Mor_{\mathcal{C}}(A, B)$, $g \in Mor_{\mathcal{C}}(B, C)$.

Some examples of categories are:

1. The category $\mathbb{K}$-*modules* whose objects are modules over a common ring $\mathbb{K}$ and whose morphisms are $\mathbb{K}$-linear homomorphisms of modules. Instances of this category are the categores *Abelian groups* when $\mathbb{K} = \mathbb{Z}$ and $\mathbb{K}$-*vector spaces* when $\mathbb{K}$ is a field.

2. The category *chain complexes over* $\mathbb{K}$ whose onjects are chain complexes over a common ring or field $\mathbb{K}$ and whose morphisms are chain maps.

3. The category *simplicial complexes* whose objects are simplicial complexes. A morphism between in this category between two simplicial complexes K, K' is a function $f: V \to V'$ from the vertex set of K to the vertex set of K' for which $\{f(v) : v \in \Delta\} \in K'$ whenever $\Delta \in K$. The function f need not be injective, so the image of a simplex could be a simplex of lower dimension.

4. The category *ordered simplicial complexes* whose objects are again simplicial complexes. Recall that we defined a simplicial complex to have an ordered set of vertices. The morphisms in this category are functions $f: V \to V'$ of vertex sets which, in addition to preserving simplices, satisfy $f(u) \leq f(v)$ whenever $u < v \in V$. We will need to consider this category for technical resaons.

5. The category *cubical complexes* whose objects are cubical complexes. A morphism in this category between two cubical complexes $K = (k_\theta)_{\theta\in\Theta}$ and $K' = (k'_\theta)_{\theta\in\theta'}$ is a function $f\colon \Theta \to \Theta'$ on index sets satisfying certain conditions involving the *degree* of an index $\theta = (\theta_1, \ldots, \theta_d) \in \Theta$; we take $\deg(\theta)$ to be the number of even valued coordinates θ_i in θ. A morphism must satisfy:

   (a) $k'_{f(\theta)} = 1$ if $k_\theta = 1$;

   (b) $\deg(f(\theta)) \leq \deg(\theta)$;

   (c) $f(\theta + \phi) = f(\theta) + f(\phi)$.

Chapter 3

Homology computation

Index

average linkage, 43

bar code, 80
betti numbers, 67
betti numbers of a chain complex, 62
boundary of pure cubical complex, 36

category, 100
cell structure, 10
cellular chain complex, 64
chain complex, 62
chain map, 78
Chebychev metric, 43
complement of pure cubical complex, 37
complete linkage, 43
complex of singularities, 39
cube, 22
cubical complex, 32
CW-space, 18

dendrogram, 83
difference of pure cubical complexes, 37

euclidean metric, 43
euler characteristic of chain complex, 62
euler characteristic of cubical complex, 34
Euler-Poincaré characteristic, 19

Fourier-Motzkin elimination, 15

genus, 23
geometric realization of cubical complex, 34
geometric realization of pure cubical complex, 31

geometric realization of simplicial complex, 51

homology of chain complex, 77
homotopy, 19
homotopy type, 20

intersection of pure cubical complexes, 37

klein bottle, 72

Manhattan metric, 42

nerve of a collection of sets, 56

path component, 25
path in pure cubical complex, 35
Pearson correlation metric, 43
persistence matrix, 79
phylogenetic tree, 83
pure cubical complex, 30

regular CW-space, 65

simlicial complex, 51
simplex, 21
simplicial complex, 51
single linkage, 43
singular point, 27
smooth point, 38

thickened pure cubical complex, 44

union of pure cubical complexes, 37

witness complex, 93

Bibliography

[1] G. Carlsson

[2] G. Carlsson, T. Ishkhanov, V. de Silva and A. Zomorodian

[3] H. Edelsbrunner et al.

[4] S. Eilenberg and S. MacLane, "General theory of natural equivalences", Trans. American Math. Soc. 58 (1945) 231-294.

[5] G. Ellis, HAP, a GAP package for homological algebra programming. http://www.gap-system.org/packages/Hap

[6] The GAP group, *Groups Algorithms Programming*, a computer system for discrete mathematics and algebra. http://www.gap-system.org

[7] R. Grhist

[8] D. Mumford, A. Lee and K. Pedersen

[9] E. Gawrilow and M. Joswig, Polymake software, http://www.opt.tu-darmstadt.de/polymake

[10] GraphViz software for displaying abstract graphs.

[11] S. MacLane, "Concepts and categories in perspective", in *A century of mathematics in America, Part I*, American Math. Soc., Providence R.I.

[12] A. Zomorodian and G. Carlsson