Universidad Industrial de Santander

Asignatura I- Modelado Matemático I

Maestría en Matemática Aplicada

La topología en la ciencia de datos

ÁLVARO JOSÉ CHAPARRO VILLAMIZAR YASMÍN RUBIELA COTE CONTRERAS

PROFESOR:

Dr. Fabio Duvan Lora Clavijo

Universidad Industrial de Santander Facultad de Ciencias Escuela de Física Bucaramanga 2022

1 Introducción

En el mundo en que nos encontramos manejamos constantemente datos, los datos nos han brindado soluciones a problemas que antes no los tenían, la ciencia que se encarga de los datos es conocida como la ciencia de datos, ésta nos ha permitido avanzar en aspectos de la vida cotidiana, tales como inferir fenómenos meteorológicos y en la toma de decisiones.

Las matemáticas se ven involucradas en la ciencia de datos, pues esta permite moldearla y ayuda a sus predicciones. En las matemáticas se encuentra una rama que se llama topología, ésta se encarga del estudio de las propiedades de los espacios topológicos y todo lo relacionado con ellos, también estudia las propiedades de los cuerpos geométricos. El objetivo de este trabajo es presentar la relación que existe entre la ciencia de datos y la topología.

En la topología se puede encontrar la topología digital y la topología algebraica, se inicia este documento presentando el procesamiento de imágenes y los temas más recientes de trabajo en esta área, seguidamente presentamos como se involucrada la topología digital en el procesamiento de imágenes, realizando unos comentarios de cómo se relaciona ésta en dicho proceso, para encontrar más referencias ver [3] y [5].

Finalmente se estudiará la relación entre la topología algebraica y la ciencia de datos por medio de un experimento realizado en la universidad de California, San Francisco, en el año 2015; Donde se estudió la relación entre lesiones neurológicas y sus implicaciones locomotrices, en comportamientos, biocomportamientos y resistencia a medicamentos, por medio de un análisis de datos topológico.

Además se resaltan las ventajas del análisis de datos topológico frente a los análisis de datos tradicionales y las limitaciones al tener escasos datos, para encontrar más referencias ver [6] y [7].

2 El procesamiento de imágenes

El mundo está rodeado de imágenes, es por eso que el desarrollo de algoritmos en el procesamiento de imágenes ha aumentado, sobre todo enfocado en la inteligencia artificial (Taquia, 2017 [2]), donde sin lugar a dudas también se ha involucrado la ciencias de datos para el desarrollo y son las imágenes los datos que procesan dichos algoritmos.

Para hablar del procesamiento de imágenes debemos entender que es una imagen y sus características, pues bien una imagen desde el punto de vista computacional es una matriz de píxeles de colores verde, rojo y azul o por sus siglas RGB, por lo que parte del procesamiento se enfoca principalmente en el color, los pixels y la continuidad.

Para analizar la continuidad en una imagen es indispensable quitar el ruido de la imagen para ello se pueden aplicar diferentes técnicas una de las más conocidas es la difuminación y binarización, esta última "es una técnica que consiste en la realización de un barrido en la matriz de la imagen digital, por medio de bucles o recursividad, con el fin de que el proceso produzca la reducción de la escala de grises a dos únicos valores" (Magro 2013, p. 445) y difuminación como su nombre lo indica consiste en difuminar.

Hoy en día se está trabajando también en el aumento de resolución, lo que se conoce como súper resolución, ésta parte del hecho de que al aumentar el tamaño de una imagen, se pierde calidad en la misma, como sabemos, una imagen está compuesta de píxeles, pues bien cuando se hace un aumento de tamaño hay información que se pierde, quedan espacios donde no se conoce el valor del píxel.

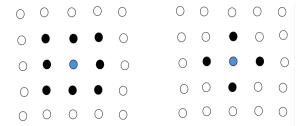


Figure 1: En esta imagen se muestra la relaciones de 8-adyacencia y 4-adyacencia, respectivamente donde los puntos negros son adyacentes al punto azul.

Para ello se han desarrollados varios métodos que permiten hacer la predicción de lo píxeles que no se conocen, los métodos que se suelen usar son métodos de interpolación como lo es rescalado bilineal y el vecino más cercano. El primero consiste en combinar proporcionalmente los colores asignando dicho valor al píxel desconocido y el segundo consiste en tomar el valor del píxel más cercano conocido.

Estos métodos hace que se pierda información, la alta resolución busca que cuando se cambie la resolución no se pierda información y ayude a generar más detalles en la imagen, para ello se deben utilizar redes convolucionales que permiten el estudio de patrones y así identificar la imagen y establecer las características faltantes a partir del entrenamiento del modelo a través de imágenes ya procesadas.

Otro idea que se está manejando predecir o mejorar la calidad de un vídeo, para ello es necesario conocer más imágenes donde nuevamente es necesario trabajar con los píxeles de las imágenes e interpolar las imágenes ya conocidas.

3 Topología digital

La topología digital está enfocada en el conjunto de los enteros y se define como el estudio de las relaciones topológicas en una imagen digital (castillo, 2012) y en lo que respecta al procesamiento de imágenes "se ocupa de las propiedades topológicas de la imagen digital y proporciona una base matemática sólida para operaciones de procesamiento de imágenes, como adelgazamiento de imágenes, seguimiento de bordes y etiquetado de componentes conectados" (shao and Li, 1992).

Para el aprovechamiento de las propiedades topológicas que podemos encontrar en las imágenes, se han desarrollado algoritmos topológicos en el adelgazamiento de imágenes, anteriormente se había definido qué es una imagen de forma computacional, pero ahora definiremos que se entiende como imagen en relación a la topología:

Definición 3.1. Una imagen digital de 2-dimensiones es una cuaterna $\rho = (\mathbb{Z}^2, m, n, N)$ donde (m, n) = (4, 8) o (8, 4)

Estos últimos relacionan las relaciones de adyacencia en un punto, se dice que un punto es adyacente a otro si difiere del otro en una unidad, la relación de adyacencia se puede evidenciar en la figura 1.

Ahora el adelgazamiento de imágenes también suele llamarse esqueletización o reducción de imagen, formalmente se puede definir de la siguiente manera

Definición 3.2. Sea $\rho = (\mathbb{Z}^2, m, n, N)$ y $\dot{p} = (\mathbb{Z}^2, m, n, \dot{N})$ dos imágenes digitales donde $\dot{N} \subset N$ donde \dot{p} conserva la topología de ρ





Figure 2: Imagen tomada de (Camara 2016, p. 32) [5].

Podemos encontrar esencialmente dos tipos de algoritmos, el primero busca que la imagen inicial (que podemos llamar ρ) sea topológicamente equivalente a la imagen final es decir \dot{p} . El segundo busca que la imagen inicial se transforme en una imagen con las mismas componentes conexas, conexas en el sentido topológico, estos algoritmos parte de una imagen en blanco y negro.

Para el primer caso, la idea es que el algoritmo determine los puntos que no afectan a la topología, y crear una imagen topológicamente equivalente. Un resultado de estos algoritmos se pueden evidenciar en la figura 2.

La figura 2 muestra el algoritmo de alternancia en campos, realizado en (camara, 2016 [5]). Este algoritmo se basa en la eliminación de puntos, en función de si el número de iteración es par o impar y si la suma de sus coordenadas es par o impar, para posteriormente establecer si el punto negro se elimina o no; para ello se establecen una serie condiciones.

Para cada iteración impar, un punto p negro es borrado si y sólo si:

- p es punto simple (punto simple es entendido como el punto que no altera la topología).
- \bullet p no es punto extremo.
- La suma de las coordenadas de p es impar.

Para cada iteración par, un punto p
 negro es borrado si y sólo si: 1. p es punto simple

- p es punto simple
- p no es punto extremo
- \bullet La suma de las coordenadas de p es par.

Este es uno de los algoritmos que permite que las imágenes resultantes sean topológicamente equivalentes, pero ¿Qué se entiende como imágenes topológicamente equivalentes?. Se puede decir que dos conjuntos son topológicamente equivalentes si existe un homeomorfismo (ver definición 4.3), pero bien un homeomorfismo exige que exista una función biyectiva, ahora cuando existe una función biyectiva en dos conjuntos se dice que tiene la misma cardinalidad, pero en el caso de las imágenes topológicamente equivalentes no es el caso, pues no se cuenta con el mismo numero de elementos y además el conjunto de las imágenes tiene finitos elementos, por tanto no es posible establecer una función biyectiva en el conjunto de las imágenes.

Cuando se dice que se busca que las imágenes sean topológicamente equivalentes se refiere a que las imágenes son homotópicas, la homotopía en matemáticas es una definición en la topología algebraica que busca establecer espacios que tenga la misma forma (un ejemplo clásico son circunferencias con bordes de diferente grosor). En las imágenes matemáticamente se establece una relación de homotopía a través de la topología digital, al establecer esta relación se dicen que las imágenes son topológicamente equivalentes, formalmente podemos definir homotopía de la siguiente manera.

Definición 3.3. Dos aplicaciones continuas $f, g :\to \dot{p}$ se dicen homótopas si existe otra aplicación (continua también) $H : \rho \times [0,1] \to \dot{p}$ tal que:

$$H(x,0) = f(x)$$

$$H(x,1) = g(x)$$

Definición 3.4. ρ es equivalente homeopáticamente equivalentes a \dot{p} si existen aplicaciones f y g tales que

$$f \circ g \simeq I_{\rho}$$

$$g \circ f \simeq 1_{\dot{p}}$$

Ahora bien, esto que tiene que ver con la ciencia de datos, pues bien para el desarrollo de algoritmos es necesario tener las imágenes para procesarlas, por tanto en estas se ve involucrado el manejo de datos, que en este caso seria los píxeles, también podemos destacar que cuando se hace un adelgazamiento de imágenes se disminuye los datos, los cuales denominamos píxeles; en el mundo que vivimos se necesitan grandes cantidades de datos a la hora de dar una mejor predicción, este adelgazamiento de imágenes o algoritmos podría ayudar a requerir menos datos a la hora de resolver un problema?.

En los documentos estudiados no se ha planteado respuesta a dicho interrogante, sin embargo el procesamiento de imágenes se ve relacionado con la ciencia de datos en la ejecución de algoritmos que conserven la topología donde es necesario los datos de las imágenes.

4 Topología algebraica

En la topología algebraica se encuentra que a la nube de puntos se le puede asociar un espacio topológico euclidiano conocido como complejo simplicial, este se vuelve de interés a la hora de extraer características topológicas tales como componentes conexos, agujeros y su estructura como gráfica, entre otras, ayudando al procesamiento y a su vez al análisis de información.

Para poder estudiar y entender el impacto del análisis de datos topológico como herramienta de estudio en el big data, se debe conocer las bases teóricas en la que se fundamenta, para ello a continuación se enuncian algunos conceptos claves que intervienen en el desarrollo del análisis de datos topológico.

Definición 4.1. Sea X un conjunto no vació. Una familia $\tau \subset \mathcal{P}(X)$ es una topología en X, si τ satisface las siguientes propiedades::

- \emptyset , $X \in \tau$;
- La unión de cualquier familia de elementos en τ esta en τ ;
- La intersección finita de elementos de τ , esta en τ .

Los elementos de τ son llamados abiertos en X. Un espacio topológico es la pareja (X,τ) , donde X es un conjunto y τ es una topología en X.

Definición 4.2. Sea $f: X \to Y$ una función entre espacios topológicos. Diremos que f es continua, si para cada abierto V en Y, $f^{-1}(V)$ es abierto enX.

Definición 4.3. Sea $f: X \to Y$ una función continua. Diremos que: f es un homeomorfismo si f es biyectiva y f^{-1} es continua. En este caso se dice que X y Y son homeomorfos y escribiremos $X \cong Y$.

Definición 4.4. Se dice que una aplicación $\varphi: A \to B$ $((A, *) y (B, *') estructuras algebraicas) es un morfismo, si para cada par de elemento <math>x, y \in A$, se tiene que:

$$\varphi(x * y) = \varphi(x) *' \varphi(y).$$

Definición 4.5. Sea X un conjunto finito, $(X = \{x_0, x_1, \ldots, x_k\} \subset \mathbb{R}^n)$ de puntos independientes. El simplex k-dimensional $\sigma = [x_0, x_1, \ldots, x_k]$ generado por X coincide con el menor simplex conexo que contiene a X. Los puntos de X son denominados vértices y el simplex generado por un subconjunto de X, se denominan caras de σ .

Definición 4.6. Un complejo simplicial K de \mathbb{R}^n es una colección de simplices (Simplex, en inglés) tales que:

- Cualquier cara de K, es un simplex de K;
- La intersección de cualquier par de simpleces de K, es vacío o una cara en común de ambos.

El complejo simplicial puede ser interpretado como generalizaciones de grafos en grandes dimensiones. Éste es el principal objeto de estudio del Análisis de datos topológicos(TDA, por sus siglas en ingles), pues es sobre el complejo que se dota una estructura para su análisis.

Por otro lado, la unión de simpleces de K es un subespacio de \mathbb{R}^n , que se denomina Polytope (Politopos o grafos) o underlying space (Espacio de interliniado), el cual adquiere la estructura de subespacio topológico inducido por \mathbb{R}^n , este espacio permite dotar a K de un espacio topológico completamente determinado por su estructura combinatoria.

Definición 4.7. El complejo simplicial abstracto finito es una familia K no vacía de subconjuntos de un conjunto de vértices $V = \{v_i\}_{i=1}^m$ tal que:

- $Si \ \alpha \in K, \ \beta \subset \alpha, \ entonces \ \beta \in K;$
- $V \subset K$, es decir cada $v \in V$, $\{v\} \in K$.

A los elemento de K se le denomina caras.

Estas últimas definiciones son comunmente usadas para construir complejos simpliciales por medio de una colección de datos, la siguiente definición ilustra cómo.

Definición 4.8. Dado un conjunto de datos X, inducidos en un espacio métrico (\mathcal{M},d) y un número real $\alpha \geq 0$. El complejo de Vietoris-Rips, $\operatorname{Rips}_{\alpha}(X)$ es el conjunto de simpleces $[x_0, x_1, \ldots, x_k]$ tales que $d_X(x_i, x_j) \leq \alpha$ para cada $i, j = 0, 1, \ldots, k$, este conjunto representa un complejo simplicial.

La importancia en la escogencia de dicho parámetro α , se ilustra en la figura 3, donde se comparan los espacios simpliciales obtenidos para distintos valores de α . Se puede observar que a medida que α se acerca a cero, el grafo obtenido se asemeja a un espacio discreto, mientras que a medida que α crece, el espacio asociado se vuelve demasiado interconectado.

En el centro de Información Sindicial visualizada y resultados para traumas neurológicos-SCI se llevó a cabo un experimento estadístico que buscaba observar las relaciones entre las lesiones cerebrales, de la médula espinal, biocomportamientos, fisiología y la respuesta de medicamentos en el tratamiento de esta patología. El objeto de estudio fue una población de roedores, los cuales presentaban daños en su medula espinal y lesiones cerebrales, a los cuales se les realizaron estudios histopatológicos (estudio de tejidos afectados), fisiológicos y pruebas locomotoras.

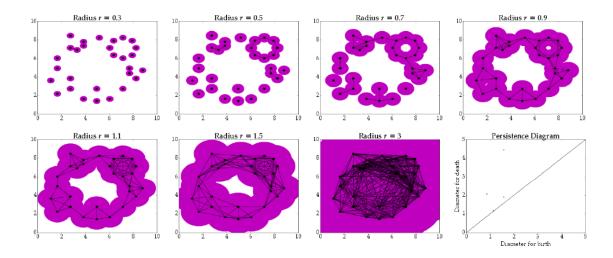


Figure 3: Ejemplos de espacios simpliciales generados para distintos valores de α (0.3, 0.5, 0.7, 0.9, 1.1, 1.5 y 3).

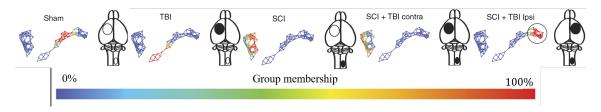


Figure 4: Se organizó una escala de colores que describe la concentración de la población sobre todo el complejo simplicial

Durante el presente estudio fue necesario utilizar técnicas tradicionales para la organización de datos, como es el análisis de componentes principales; por medio del cual, los datos se distribuyeron en el espacio tridimensional teniendo en cuenta la relación entre los individuos y sus resultados en los exámenes fisiológicos, biocomportamientos y resistencia de medicamentos.

Posteriormente se determina un coeficiente óptimo, denominado radio topológico, por medio del cual, una vez organizados los datos éstos se toman como centro y se traza una esfera de dicho radio. Cuando la intersección entre las esferas es distinto de vacío, sus respectivos centros se dice que están relacionados y se expresa dicha relación por medio de una arista, el conjunto de datos y aristas se conoce como espacio simplicial.

En la figura 4 se puede observar la distribución de la población, por medio de una escala de azul a rojo, en el que se refleja la densidad de la población, siendo azul una zona de baja concentración y rojo de alta concentración.

Esta organización ya sugiere relaciones sobre las variables, pues como vemos esta estructura es multiplemente conexa, ya que está dividida en dos regiones disjuntas. En una de ellas se encuentra la población de roedores con lesiones de la médula espinal y ambas lesiones en hemisferios opuestos, mientras que en la otra región encontramos la población de roedores sanos, con lesiones cerebrales y ambas lesiones en hemisferios opuestos. Esto nos sugiere que las lesiones de médula espinal son más letales que las cerebrales, pues está última población se presenta resultados fisiológicos similares a la población de roedores sanos.

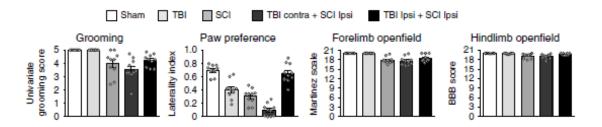


Figure 5: Comportamientos obtenidos tras realizar experimentos motrices.

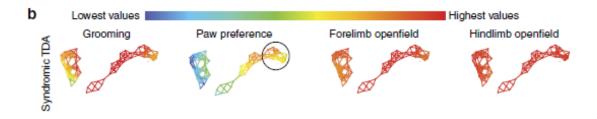
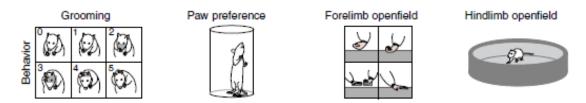


Figure 6: En este análisis las histopatologías fueron mapeadas del espacio topológico en sí mismo por medio de un análisis topológico de datos.

Para estudiar la relación entre las variables, los investigadores plantearon una serie de experimentos; primero se encuentra el Grooming, este examen consiste en la observación de las actividad que realizan los roedores para limpiarse del mugre o insectos de su piel, seguidamente el Paw preference es un reconocimiento espacial donde el individuo explora su entorno por medio de sus extremidades y por último las Open field consisten en actividades que miden niveles locomotores de reconocimiento espacial las cuales se dividen en dos; el Forelimb estudia las respuestas cerebrales para poder efectuar un desplazamiento y el Hindlimb el comportamiento que se realiza para reconocer el entorno en el que se encuentra.

Tras realizar los experimentos antes mencionados, cada dato o individuo obtuvo una valoración de acuerdo a sus observaciones, esta información se clasificó por medio de un diagrama de barras (ver figura 5), donde se puede observar que en la mayoría de exámenes los resultados obtenidos en cada población fueron bastante similares, sin embargo cuando se utilizan técnicas de análisis de datos obtenidas a través del estudio de las propiedades topológicas del espacio simplicial, se observa que esto no parece ser como los diagramas de barras indicaban (ver figura 4).



El análisis de datos topológico permite asociar cada dato del espacio simplicial con un color, el cual está relacionado a una escala de colores que hace referencia a la intensidad de las observaciones por medio de un homeomorfismo del espacio topológico simplicial en sí mismo, que colorea cada vértice (ver figura 6). Esta construcción permite evidenciar que:

• Existen patrones que relacionan las lesiones de la médula espinal y lesiones cerebrales con el

cambio locomotor de los individuos

• El incremento de la presión sanguínea después de una intervención quirúrgica sugiere la existencia de una baja tasa de recuperación.

5 Consideraciones finales

La topología y la ciencias datos se encuentran relacionadas no solo por la forma de dar estructura en el manejo de datos en el espacio simplicial, a la hora de dar un análisis de los datos obtenidos, si no que además realizar un análisis dejando de lado las formas tradicionales de realizarlo, permite observar comportamientos en los datos que por medio de formas tradicionales no se podrían establecer.

En lo que respecta al procesamiento de imágenes y topología digital, la ciencia de datos juega un papel importante, pues se necesitan datos para realizar el procesamiento de imágenes a través de los algoritmos; por ejemplo para establecer si una imagen es topológicamente equivalente a otra o realizar algoritmos para la mejora de la resolución de una imagen.

La principal herramienta del análisis de datos topológico son el estudio de persistencias homologías, por medio de las cuales se estudian las propiedades del espacio simplicial, estas persistencias han encontrado muchas aplicaciones en las investigación de datos en diversas ramas de estudio; estas incluyen procesamiento de imágenes (Carlsson, Ishkhanov, de Silva, Zomorodian, 2008 [8]; Perea Carlsson, 2014 [9]; Adcock, Carlsson, Carlsson, 2016 [10]), Análisis de series de tiempo (Perea, Deckard, Haase, Harer, 2015 [11]; Khasawneh Munch, 2016; Emrani, Gentimis, Krim, 2014 [12]), Filogenia (Chan, Carlsson, Rabadan, 2013 [14]), Neurociencia (Giusti, Pastalkova, Curto, Itskov, 2015 [15]; Dabaghian, Mémoli, Frank, Carlsson, 2012 [16]), y redes sensoriales (de Silva Ghrist, 2007 [17]; Adams Carlsson, 2015 [18]; Munch, Shapiro, Harer, 2012 [19]). ver (Referencia)

References

- [1] Chao, D; Change,C; Kaiming H & Xiaoou T (2015, 3 julio). Image Super-Resolution Using Deep Convolutional Networks. Disponible en https://arxiv.org/pdf/1501.00092.pdf.
- [2] Taquia, J. (2017). El procesamiento de imágenes y su potencial aplicación en empresas con estrategia digital. INTERFASES, 10, 11–29.
- [3] Shao, J & LI, D. (Ed.). (1992). DIGITAL TOPOLOGICAL AND MATRIX STRUCTURED IMAGE PROCESSING (Vol. 17). ISPRS Congress, Comm. ill, Washington D. C.
- [4] Magro, R. (2013, 31 diciembre). BINARIZACIÓN DE IMÁGENES DIGITALES Y SU ALGORITMIA COMO HERRAMIENTA APLICADA A LA ILUSTRACIÓN ENTOMOLÓGICA. Boletín de la Sociedad Entomológica Aragonesa. http://sea-entomologia.org/PDF/Boletin53/443464BSEA53BinarizacionRMagro.pdf.
- [5] Camara, A. (2016, junio). Algoritmos topológicos de adelgazamiento de imagenes digitales. Tesis para optar el título de matemático e informático.
- [6] Jiménez-Martínez, K. S., & Trejo-Medina, D. (2016, 26 mayo). Topología Algebraica para el Análisis de Datos. DSA Investigación y Desarrollo Aplicado. DOI: 10.13140/RG.2.1.4610.3922.
- [7] Nielson, J., & Paquette, J. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. nature communications. https://doi.org/10.1038/ncomms9581. (2015).
- [8] Carlsson, G., Ishkhanov, T., de Silva, V., Zomorodian, A. (2008). On the local behavior of spaces of natural images. International Journal of Computer Vision, 76, 1–12. http://dx.doi.org/10.1007/s11263-007-0056-x
- [9] Perea, J. A., Carlsson, G. (2014). A Klein-bottle-based dictionary for texture representation. International Journal of Computer Vision, 107(1), 75–97. http://dx.doi.org/10.1007/s11263-013-0676-2
- [10] Adcock, A., Carlsson, E., Carlsson, G. (2016). The ring of algebraic functions on persistence bar codes. Homology, Homotopy and Applications, 18(1), 381–402.
- [11] Perea, J. A., Deckard, A., Haase, S. B., Harer, J. (2015). SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. BMC Bioinformatics, 16(1).
- [12] Khasawneh, F. A., Munch, E. (2016). Chatter detection in turning using persistent homology. Mechanical Systems and Signal Processing, 70–71, 527–541. https://dx.doi.org/10.1016/j.ymssp.2015.09.046
- [13] Emrani, S., Gentimis, T., Krim, H. (2014). Persistent homology of delay embeddings and its application to wheeze detection. Signal Processing Letters, IEEE, 21(4), 459–463. http://dx.doi.org/10.1109/LSP.2014.2305700
- [14] Chan, J. M., Carlsson, G., Rabadan, R. (2013). Topology of viral evolution. Proceedings of the National Academy of Sciences, 110(46), 18566–18571.
- [15] Giusti, C., Pastalkova, E., Curto, C., Itskov, V. (2015). Clique topology reveals intrinsic geometric structure in neural correlations. Proceedings of the National Academy of Sciences, 112(44), 13455–13460. http://dx.doi.org/10.1073/pnas.1506407112
- [16] Dabaghian, Y., Mémoli, F., Frank, L., Carlsson, G. (2012). A topological paradigm for hip-pocampal spatial map formation using persistent homology. PLoS Computational Biology, 8(8), e1002581.

- [17] de Silva, V., Ghrist, R. (2007). Coverage in sensor networks via persistent homology. Algebraic Geometric Topology, 7, 339–358. http://dx.doi.org/10.2140/agt.2007.7.339
- [18] Adams, H., Carlsson, G. (2015). Evasion paths in mobile sensor networks. International Journal of Robotics Research, 34(1), 90–104. https://dx.doi.org/10.1177/0278364914548051
- [19] Munch, E., Shapiro, M., Harer, J. (2012). Failure filtrations for fenced sensor networks. The International Journal of Robotics Research, 31(9), 1044-1056. https://dx.doi.org/10.1177/0278364912451671