

# Tesis De Maestría

## Test de Hipótesis Sobre Homología Persistente Utilizando la Distancia de Fermat



Universidad de Buenos Aires - Facultad De Ciencias Exactas y Naturales

Diego Javier Battocchio

Director: Dr. Pablo Groisman

## Qué buscamos y qué aprendimos

# Qué buscamos y qué aprendimos

## **Buscamos**

- Aplicar herramientas de pruebas de hipótesis para homología persistente.
- Validar resultados de la bibliografía.
- Comparar esos resultados con los obtenidos usando la distancia de Fermat.

## **Encontramos**

- Fermat es más potente en topologías complejas.
- Los métodos de la bibliografía fallan en ciertas topologías y para  $D > 2$ .
- Fermat es consistente con la inspección visual en datos reales.

## **Mientras tanto aprendimos**

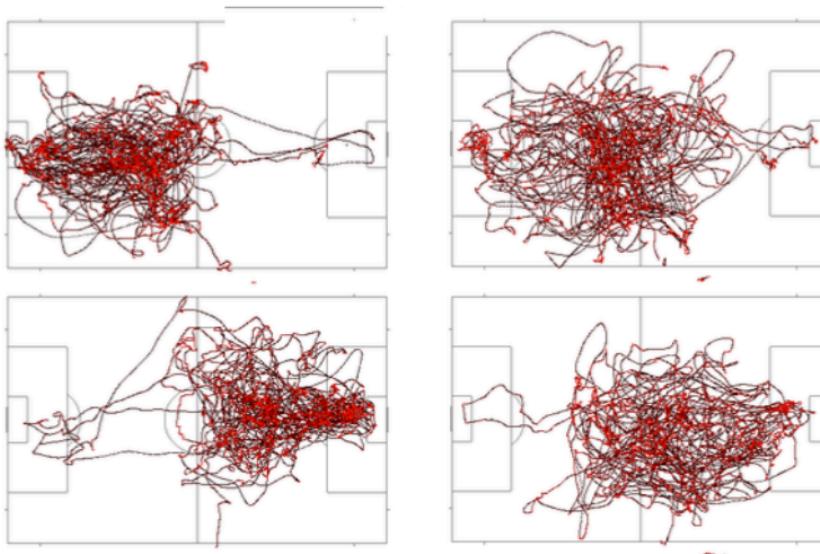
- Topología aplicada a análisis de datos.
- Geometría computacional.
- Publicar un paquete de R en CRAN.

## **Problema y Motivación**

# Problema y Motivación

## Homología Persistente (Análisis Topológico de Datos)

Los “agujeros” en los datos revelan estructura: la homología persistente permite detectarlos y analizarlos de forma confiable.



## Problema y Motivación (II)

---

La distancia de Fermat puede ayudarnos a mejorar los resultados en estas tareas!

### Bot Detection on Social Networks Using Persistent Homology

Minh Nguyen<sup>1</sup>, Mehmet Aktas<sup>1,\*</sup> and Esra Akbas<sup>2</sup>

<sup>1</sup> Department of Mathematics and Statistics, University of Central Oklahoma, Edmond, OK 73034, USA;

mnguyen3@uco.edu

<sup>2</sup> Department of Computer Science, Oklahoma State University, Stillwater, OK 74078, USA;

eakbas@okstate.edu

\* Correspondence: maktas@uco.edu

### Coverage in sensor networks via persistent homology

VIN DE SILVA

ROBERT GHIRST

### Persistent Homology of Delay Embeddings and its Application to Wheeze Detection

Saba Emrani, Thanos Gentimis and Hamid Krim

February 21, 2014

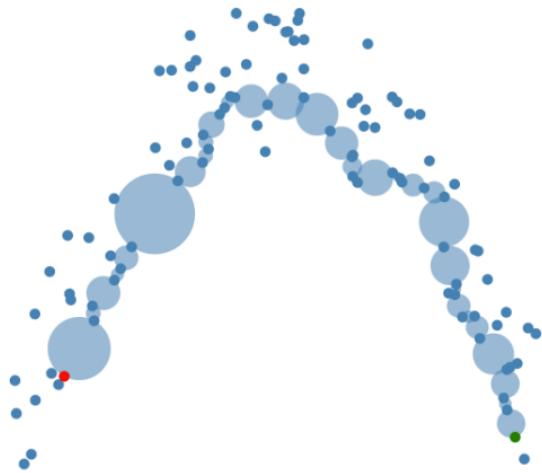
## **Conceptos Utilizados**

## Distancia de Fermat

---

$$d_F(p, q) = \min_{\mathcal{S}_N^K} \sum_{i=1}^{K-1} \|x^i - x^{i+1}\|^\lambda$$

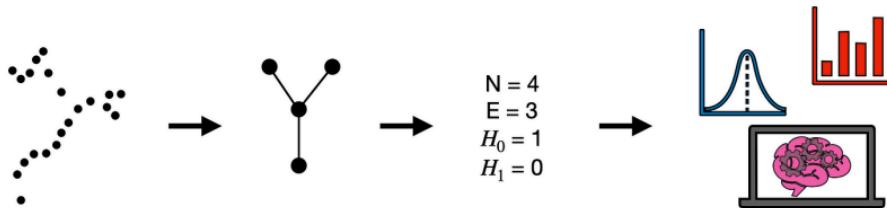
$$x^1 = p, \quad x^K = q$$



Incorpora información topológica en la definición de distancia entre puntos

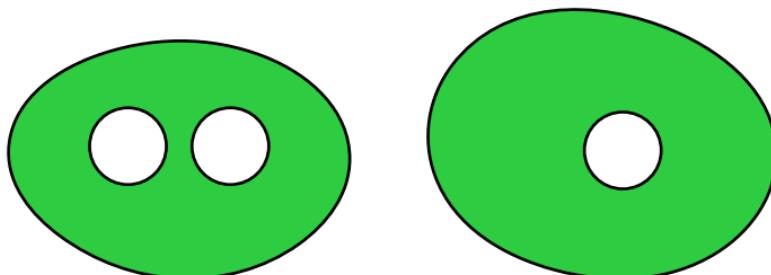
# Análisis topológico de datos

---



## Homología

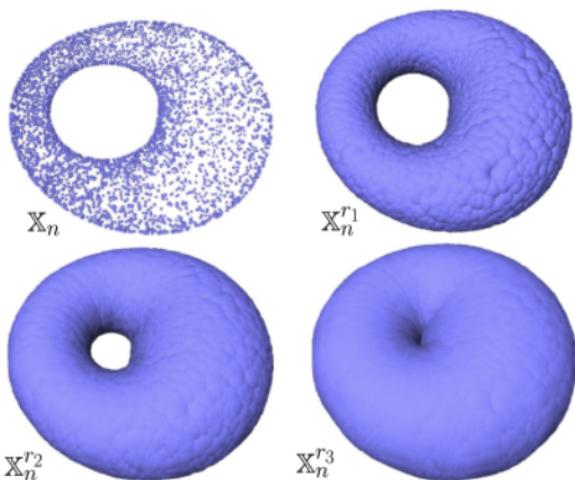
Área de estudio que clasifica espacios topológicos en términos de agujeros de diferentes dimensiones.



¿Cómo inferir la homología de una variedad  $\mathcal{M}$  a partir de una muestra  $\mathcal{S}_N$ ?

$$B(\mathbf{x}_i, \varepsilon) = \{\mathbf{x} \mid d(\mathbf{x}_i, \mathbf{x}) < \varepsilon\}$$

$$\mathcal{S}_N^\varepsilon = \cup_{\mathbf{x} \in \mathcal{S}_N} B(\mathbf{x}, \varepsilon)$$



Calcular la homología  $H(\mathcal{M})$  directamente a partir de  $\mathcal{S}_N^\varepsilon$  es muy difícil.

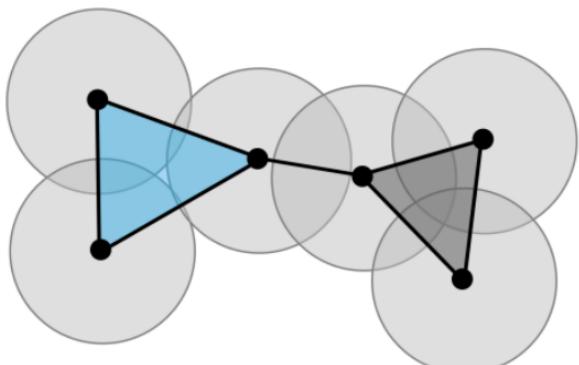
### Complejo de Čech

$\check{\text{C}}\text{ech}(\mathcal{S}_N, \varepsilon)$ : Conjunto de simplices  $\sigma$  con vértices  $v_1, \dots, v_k \in \mathcal{S}_N$  tales que

$$\bigcap_{i=1}^k B(v_i, \varepsilon) \neq \emptyset$$

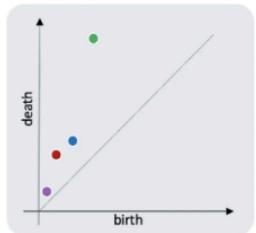
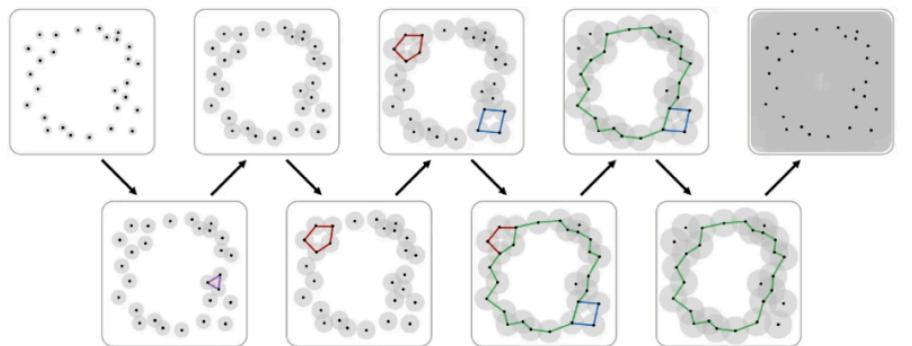
### Complejo de Vietoris-Rips

$V(\mathcal{S}_N, \varepsilon)$ : Simplices con vértices en  $\mathcal{S}_N$  de diámetro máximo  $2\varepsilon$ .



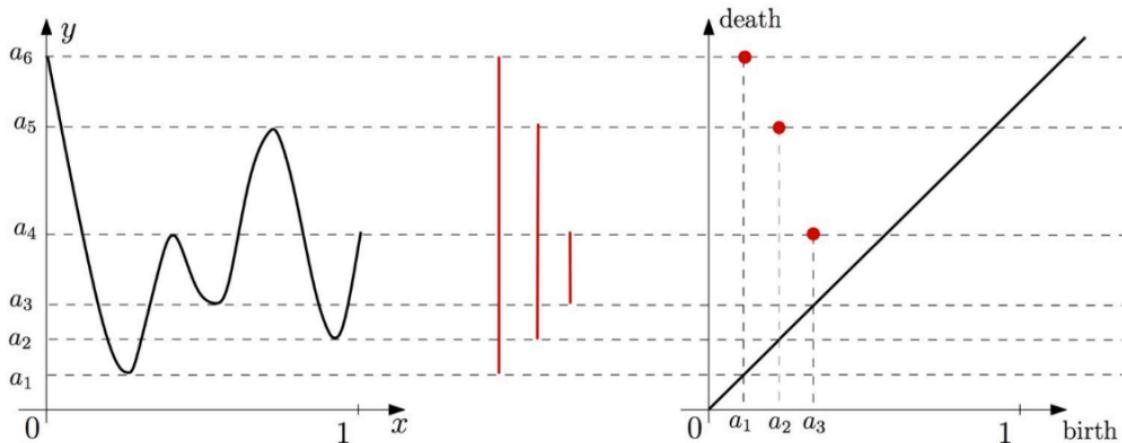
$$\check{\text{C}}\text{ech}(\mathcal{S}_N, \varepsilon) \subset V(\mathcal{S}_N, \varepsilon) \subset \check{\text{C}}\text{ech}(\mathcal{S}_N, \sqrt{2}\varepsilon)$$

Se obtiene aumentando progresivamente el  $\varepsilon$  y evaluando las cualidades topológicas de  $\mathcal{S}_N^\varepsilon$  para cada valor



## Diagramas de persistencia de conjuntos de nivel

$$\mathcal{S}_N^\varepsilon = \cup_{\mathbf{x} \in \mathcal{S}_N} B(\mathbf{x}, \varepsilon) = L_\varepsilon = \left\{ \mathbf{x} \mid \min_{\mathbf{x}_i \in \mathcal{S}_N} d(\mathbf{x}_i, \mathbf{x}) < \varepsilon \right\}$$

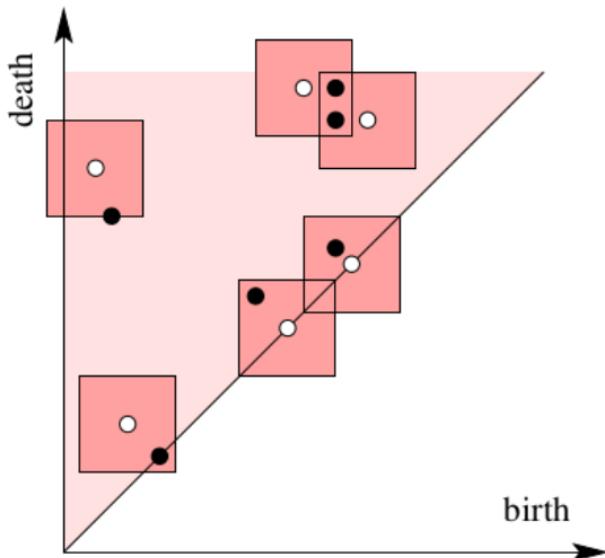


## Distancia entre diagramas de persistencia

$\mathcal{X}, \mathcal{Y}$  : Diagramas de persistencia

$$\eta : \mathcal{X} \rightarrow \mathcal{Y}$$

$$W_\infty(\mathcal{X}, \mathcal{Y}) = \inf_{\eta: X \rightarrow Y} \sup_{x \in \mathcal{X}} \|x - \eta(x)\|_\infty$$



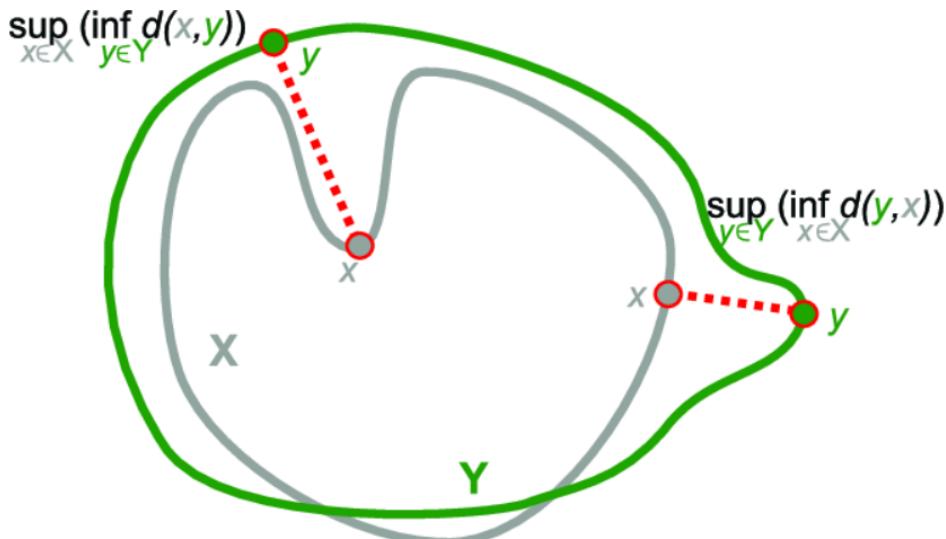
## Distancia de Hausdorff

---

$$d(x, Y) = \inf_{y \in Y} d(x, y)$$

$$d_H(X, Y) = \max \left\{ \sup_{y \in Y} d(y, X), \sup_{x \in X} d(x, Y) \right\}$$

$$W_\infty(\mathcal{P}(X), \mathcal{P}(Y)) \leq d_H(X, Y)$$



# Regiones de Confianza

---

## Intervalos de confianza

$$\mathbb{P}\{\theta_L(X) \leq \theta \leq \theta_U(X)\} = 1 - \alpha$$

## Regiones de confianza

$$\mathbb{P}\{\theta \in A(X)\} = 1 - \alpha$$

## Cálculo de regiones de confianza

El cálculo analítico de las regiones de confianza suele ser imposible.

## Técnicas de Muestreo

Analizar la distribución empírica de  $\hat{\theta}^j = T(\mathcal{S}_N^j)$

### Bootstrap

$\mathcal{S}_N^j$ : muestras de tamaño  $N$  **con reposición**

### Sub-muestreo

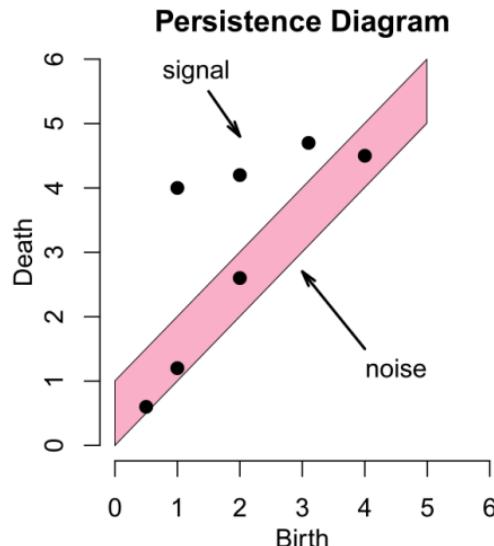
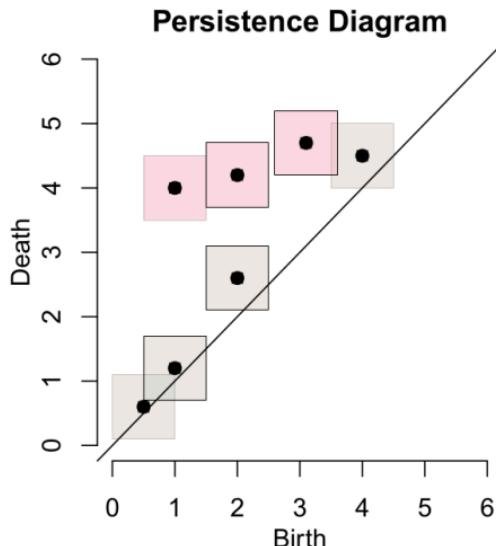
$\mathcal{S}_N^j, b$ : muestras de tamaño  $b$  **sin reposición**

## Regiones de Confianza

## Regiones de confianza para diagramas de persistencia

$$\liminf_{n \rightarrow \infty} \mathbb{P}\left(0 \leq W_\infty(\mathcal{P}, \hat{\mathcal{P}}) \leq \theta_n\right) \geq 1 - \alpha$$

$$\mathcal{C}_n = \left\{ \tilde{\mathcal{P}} \mid W_\infty(\tilde{\mathcal{P}}, \hat{\mathcal{P}}) < \theta_n \right\}$$



## **Desarrollo y Métodos Utilizados**

$$\theta_j = d_H(\mathcal{S}_N, \mathcal{S}_N^j)$$

$$L_{b(t)} = \frac{1}{M} \sum_{j=1}^M I(\theta_j > t)$$

$$c_b = 2L_b^{-1}(\alpha)$$

Se demuestra

$$\mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > c_b) \leq \mathbb{P}(d_H(\mathcal{S}_N, \mathcal{M}) > c_b) = \alpha + \mathcal{O}\left(\frac{b}{N}\right)^{\frac{1}{4}}$$

## Algoritmo

INTERVALO SUB-MUESTREO( $\mathcal{S}_N, b, \alpha, M$ ):

```

1    $\theta \leftarrow \text{array}(M)$ 
2   for  $j \leftarrow 0$  to  $M$ :
3        $\mathcal{S}_N^j \leftarrow \text{submuestra}(\mathcal{S}_N, b)$ 
4        $\theta[j] \leftarrow d_H(\mathcal{S}_N, \mathcal{S}_N^j)$ 
5   return 2 quantile( $\theta, 1 - \alpha$ )

```

La distancia utilizada para calcular  $d_H(\mathcal{S}_N, \mathcal{S}_N^j)$  puede ser Fermat o euclídea

Eficiente: No es necesario computar el diagrama de persistencia

## Distancia de Hausdorff para nubes de puntos

$$d_H(\mathcal{A}_N, \mathcal{B}_N) = \max \left\{ \max_{\mathbf{a} \in \mathcal{A}_N} \min_{\mathbf{b} \in \mathcal{B}_N} d(\mathbf{a}, \mathbf{b}), \max_{\mathbf{b} \in \mathcal{B}_N} \min_{\mathbf{a} \in \mathcal{A}_N} d(\mathbf{a}, \mathbf{b}) \right\}$$

Nuestro caso particular  $\mathcal{A}_N = \mathcal{S}_N$ ,  $\mathcal{B}_N = \mathcal{S}_N^j$

$$d_H(\mathcal{S}_N, \mathcal{S}_N^j) = \max_{\mathbf{a} \in \mathcal{S}_N} \min_{\mathbf{b} \in \mathcal{S}_N^j} d(\mathbf{a}, \mathbf{b})$$

## Estimación robusta de distancia de Hausdorff

$$\hat{d}_H(\mathcal{S}_N, \mathcal{S}_N^j) = \text{percentil}(\gamma) \min_{\mathbf{b} \in \mathcal{S}_N^j} d(\mathbf{a}, \mathbf{b})$$

$$f_h(x) = \sum_{i=1}^N \frac{1}{h^D} K\left(\frac{\|x - x_i\|_2}{h}\right)$$

evaluado en una grilla de puntos para construir el diagrama de persistencia.

Computacionalmente muy costoso

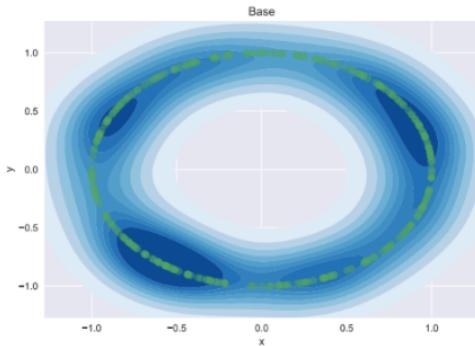
Puede omitir detalles sutiles de la topología del espacio original

Resulta más estable al ruido y datos atípicos

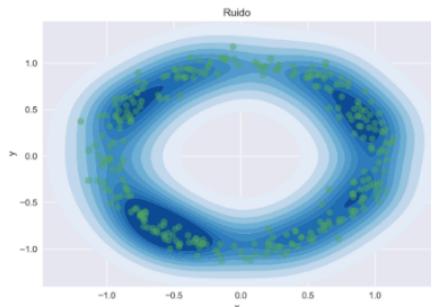
BOOTSTRAP DENSIDAD( $\mathcal{S}_N, h, \alpha, M$ ):

- 1  $\theta \leftarrow \text{array}(M)$
- 2  $\hat{f}_h \leftarrow \text{estimador de densidad de } f$   
basado en  $\mathcal{S}_N$
- 3  $\hat{\mathcal{P}} \leftarrow \text{diagrama de persistencia de } \hat{f}_h$
- 4 **for**  $j \leftarrow 0$  **to**  $M$ :
- 5      $\mathcal{S}_N^j \leftarrow \text{muestra bootstrap de } \mathcal{S}_N$
- 6      $\hat{f}_h^j \leftarrow \text{estimador de densidad de } f$   
basado en  $\mathcal{S}_N^j$
- 7      $\hat{\mathcal{P}}^j \leftarrow \text{diagrama de persistencia de } \hat{f}_h^j$
- 8      $\theta[j] \leftarrow W_\infty(\hat{\mathcal{P}}, \hat{\mathcal{P}}^j)$
- 9 **return**  $\text{quantile}(\theta, 1 - \alpha)$

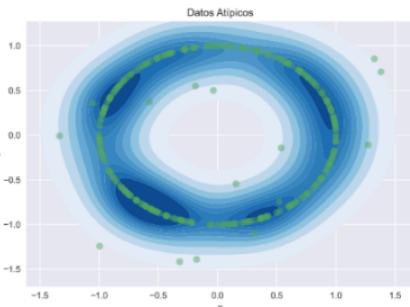
# **Conjuntos de datos**



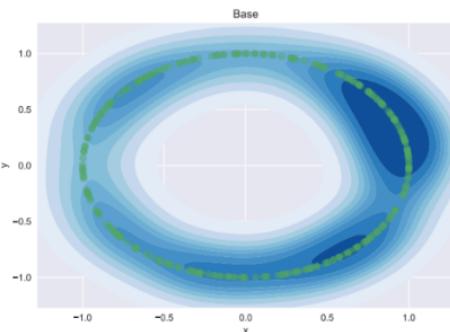
(a) Muestreo uniforme



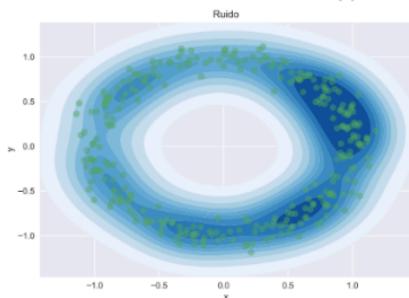
(b) Ruido agregado



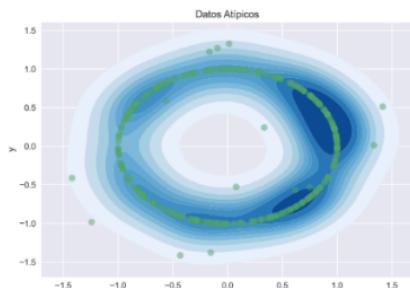
(c) Datos atípicos



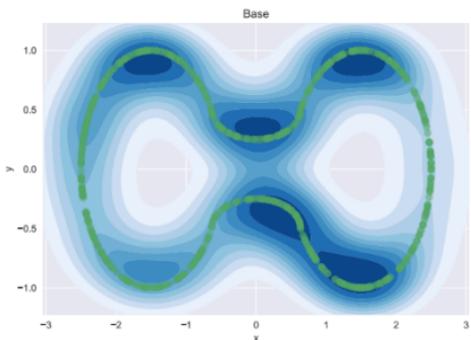
(a) Muestre uniforme



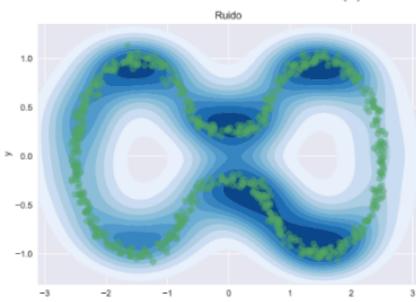
(b) Ruido agregado



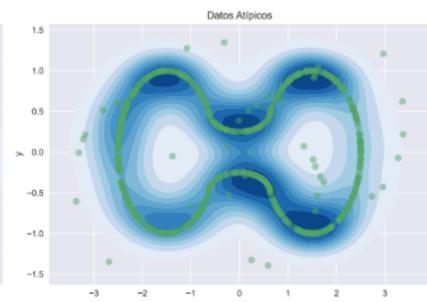
(c) Datos atípicos



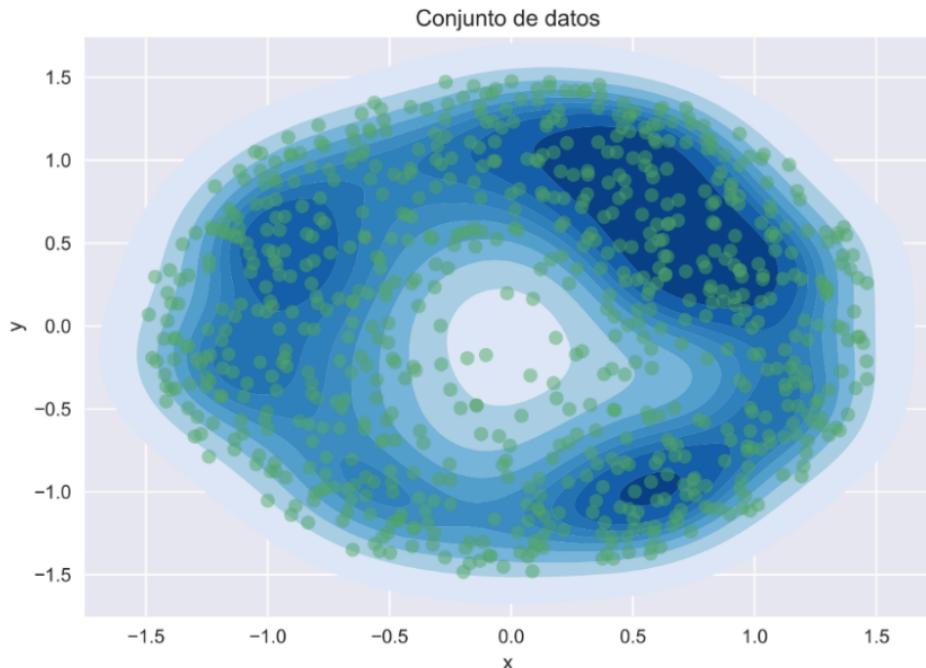
(a) Muestre uniforme



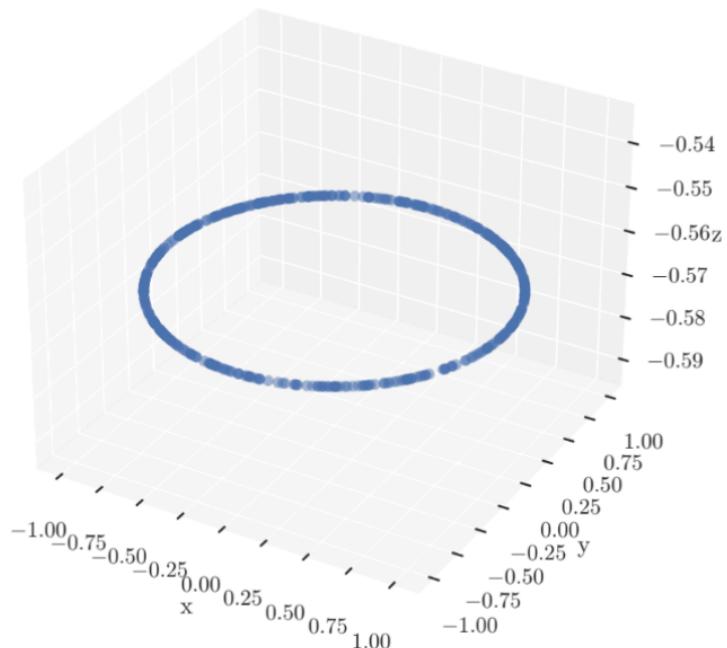
(b) Ruido agregado



(c) Datos atípicos



Circunferencia con  $D = 3$



## Soccer Video and Player Position Dataset

Svein Arne Pettersen<sup>1</sup>, Dag Johansen<sup>1</sup>, Håvard Johansen<sup>1</sup>, Vegard Berg-Johansen<sup>4</sup>,  
Vamsidhar Reddy Gaddam<sup>2,3</sup>, Asgeir Mortensen<sup>2,3</sup>, Ragnar Langseth<sup>2,3</sup>, Carsten Griwodz<sup>2,3</sup>,  
Håkon Kvale Stensland<sup>2,3</sup>, Pål Halvorsen<sup>2,3</sup>

<sup>1</sup>University of Tromsø, Norway

<sup>3</sup>University of Oslo, Norway

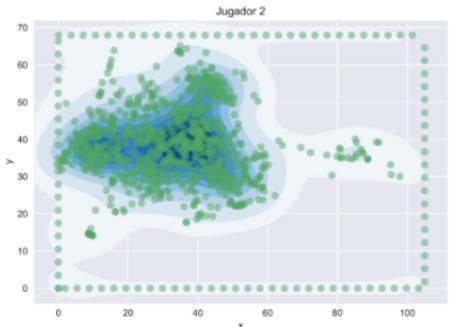
<sup>2</sup>Simula Research Laboratory, Norway

<sup>4</sup>Tromsø Idrettslag, Norway

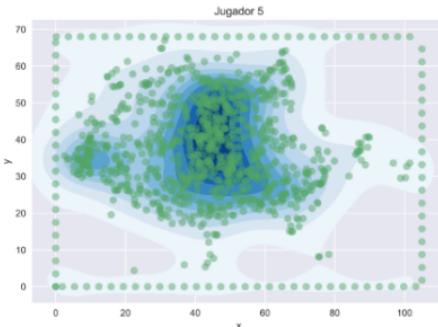
{svein.arne.pettersen, dag, haavardj}@uit.no      vegard.berg-johansen@til.no  
{vamsidhg, asgeirom, ragnarla, griff, haakonks, paalh}@ifi.uio.no

## Jugadores analizados

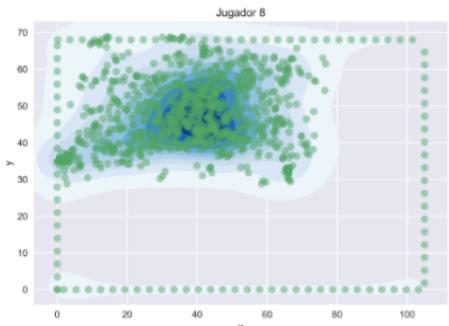
- Defensor Central
- Mediocampista
- Lateral Izquierdo
- Mediocampista



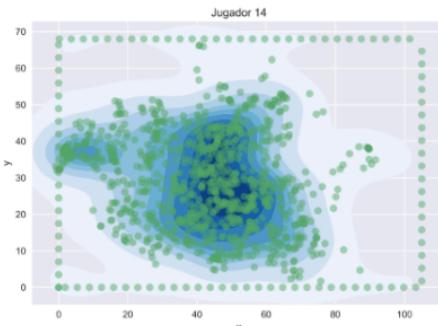
(a) Jugador 2: Defensor Central



(b) Jugador 5: Mediocampista



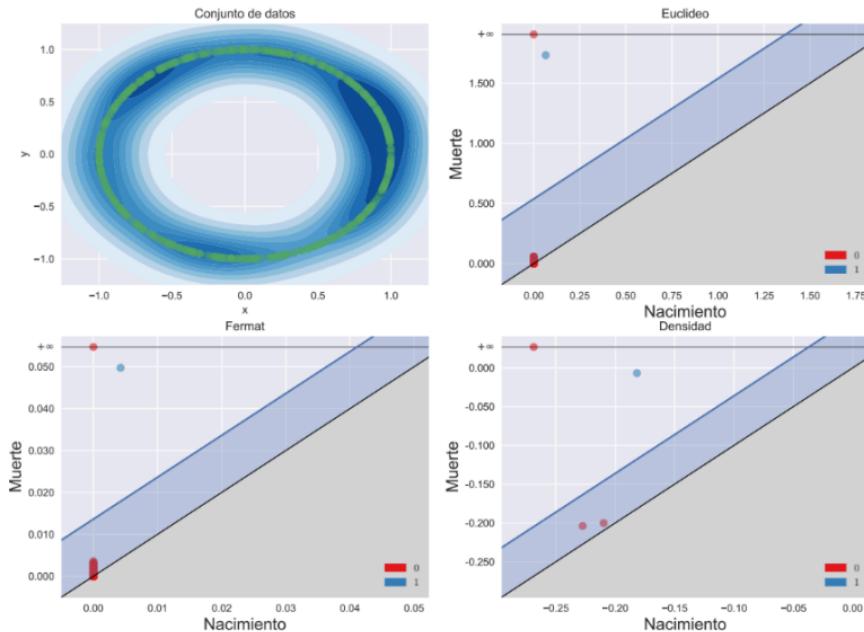
(c) Jugador 8: Lateral Izquierdo



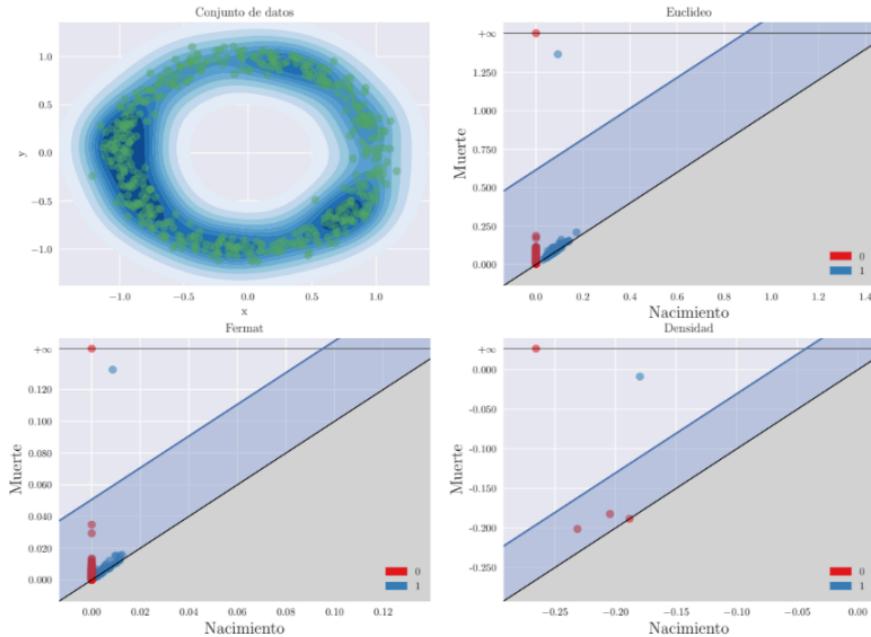
(d) Jugador 14: Mediocampista

# **Resultados**

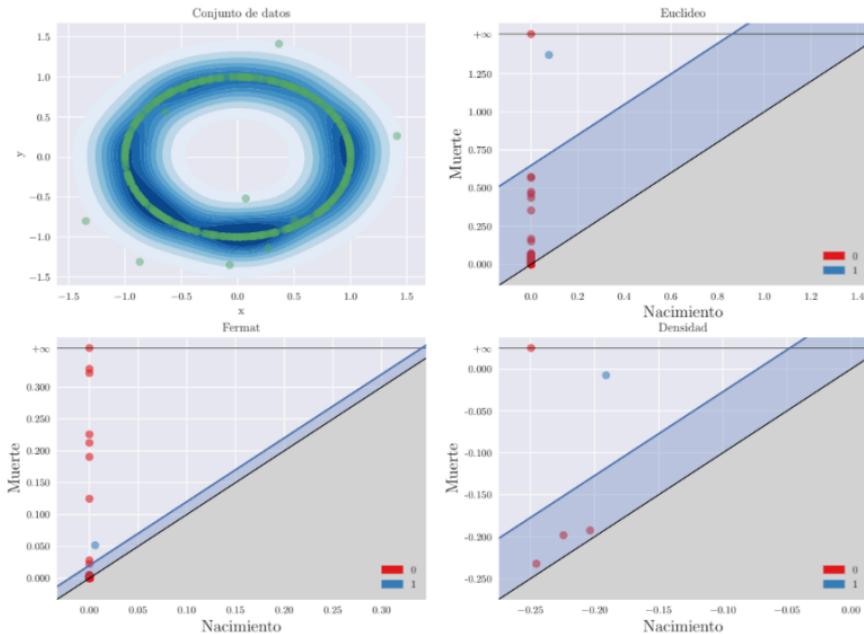
### Original



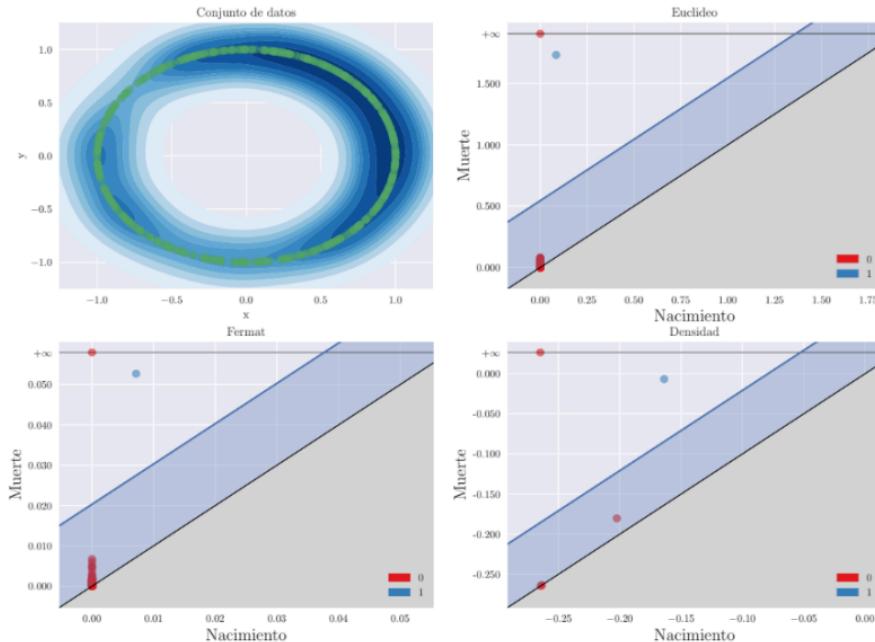
### Ruido



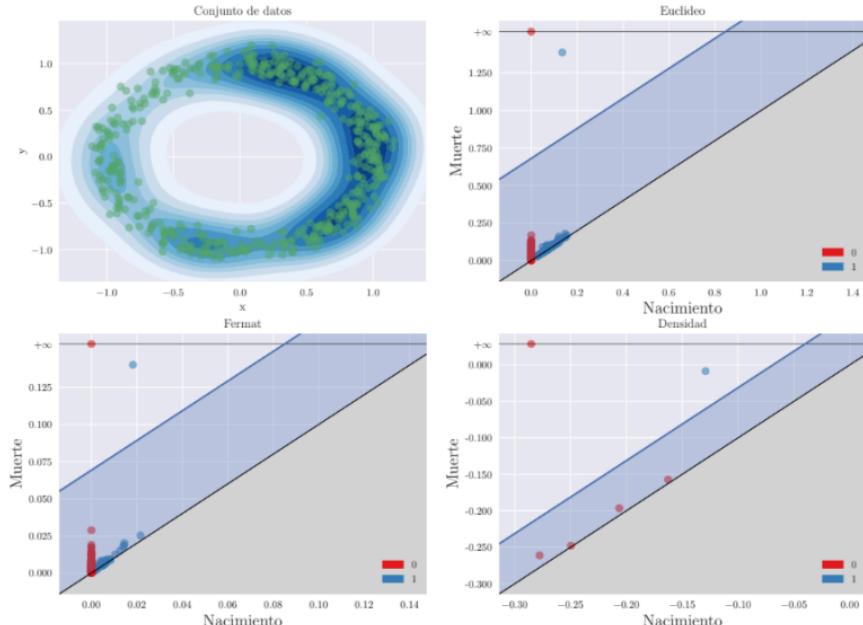
### Datos Atípicos



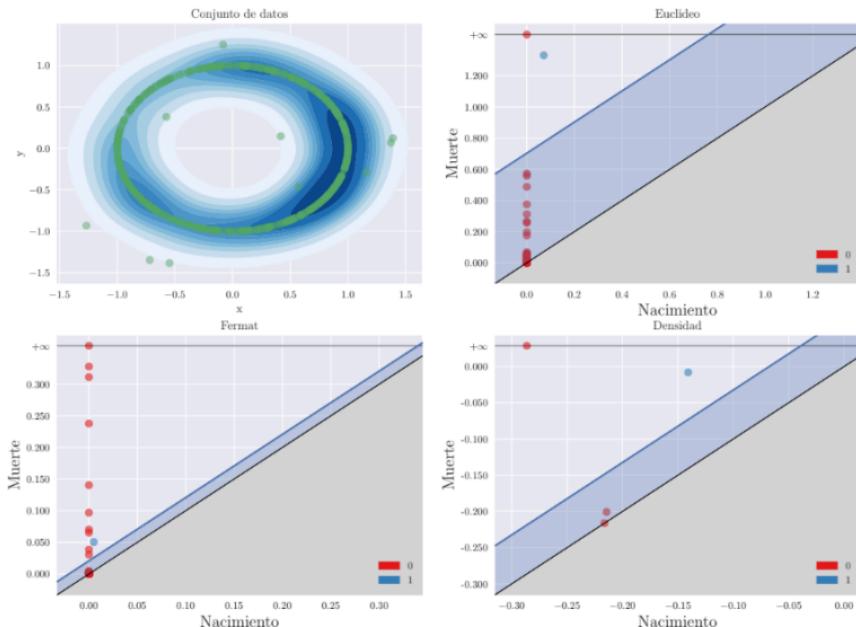
## Original



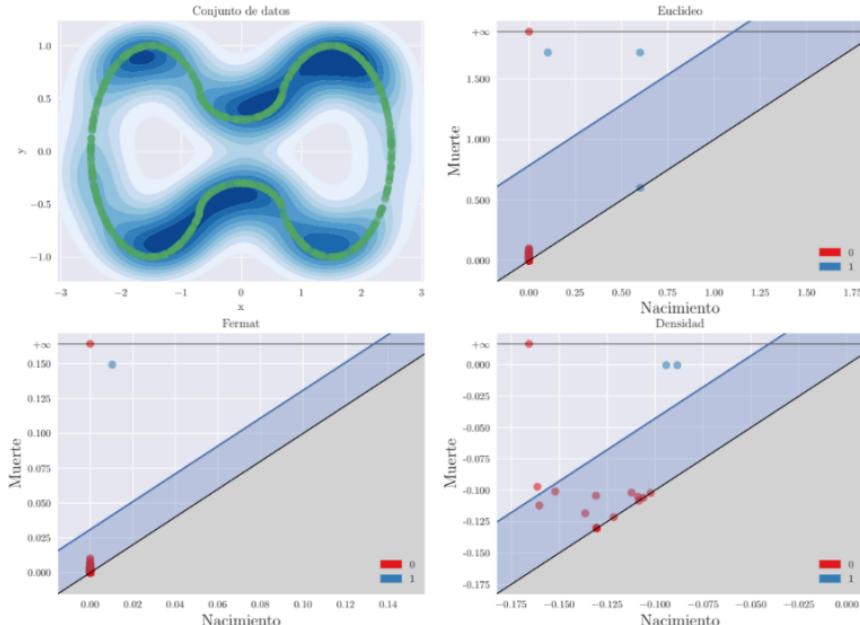
### Ruido



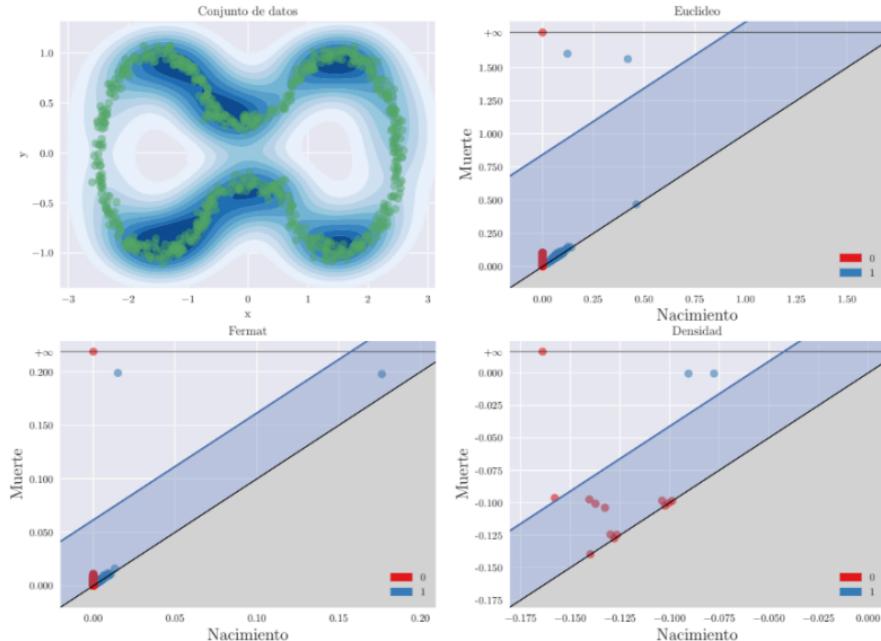
## Datos Atípicos



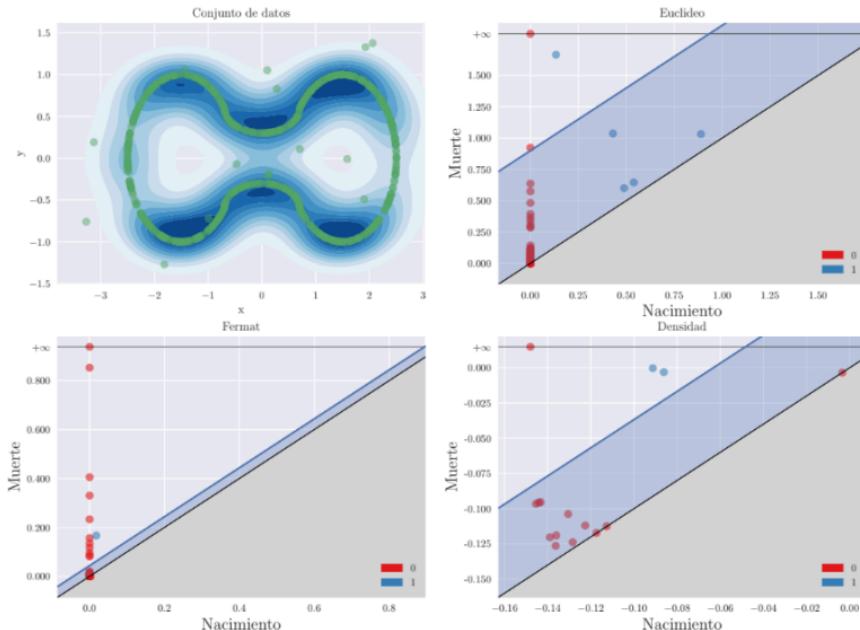
## Original



### Ruido

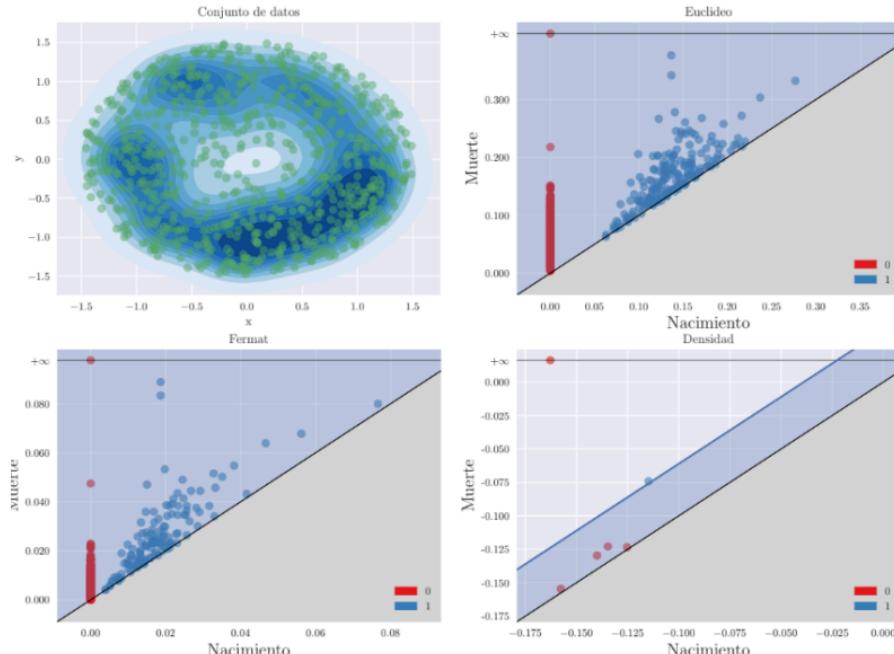


### Datos Atípicos

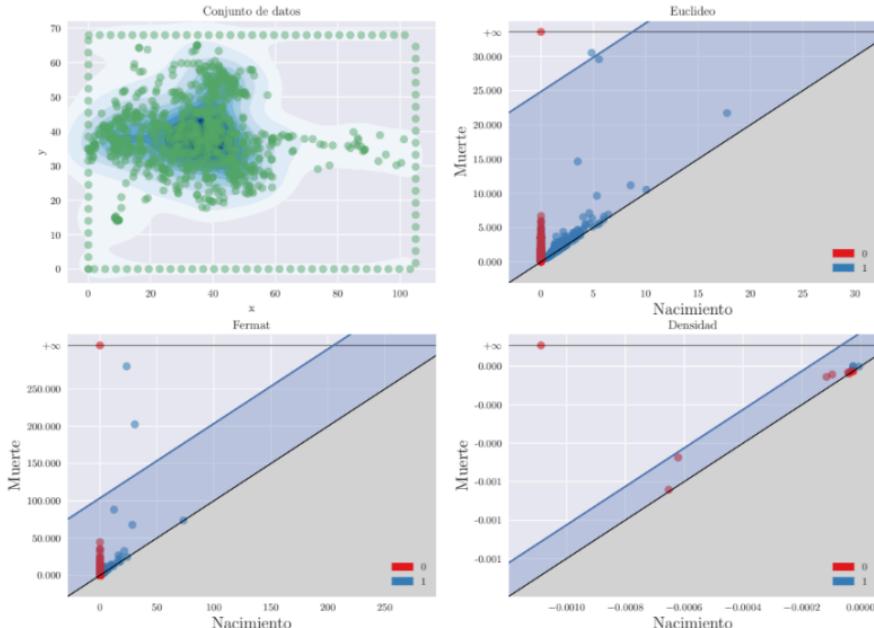


# Diagramas de Persistencia

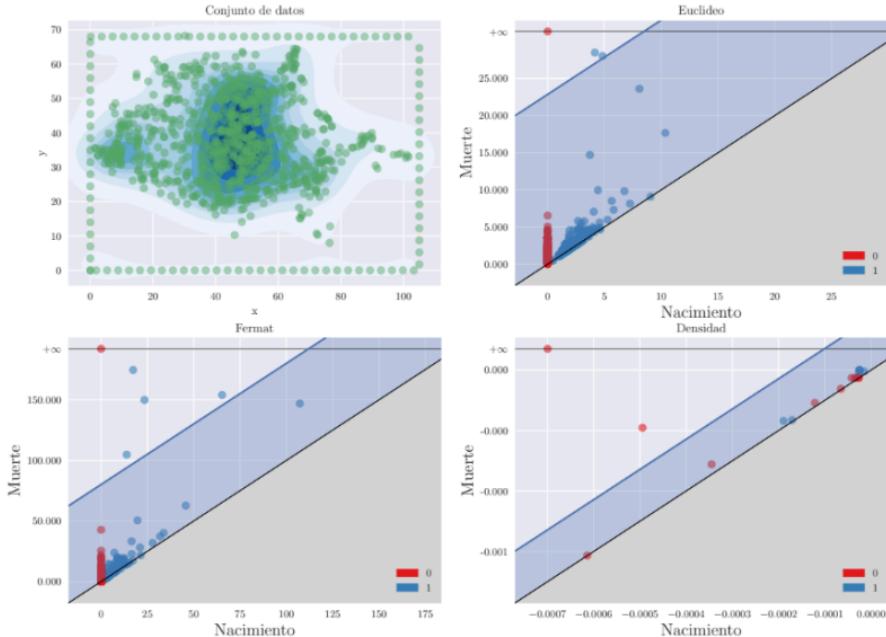
## Círculo Relleno



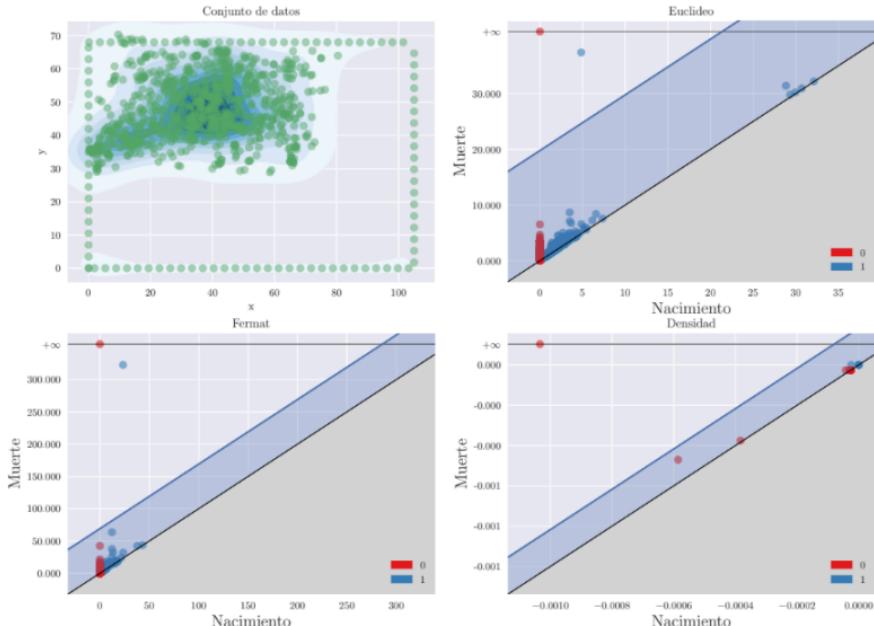
## Defensor Central (2)



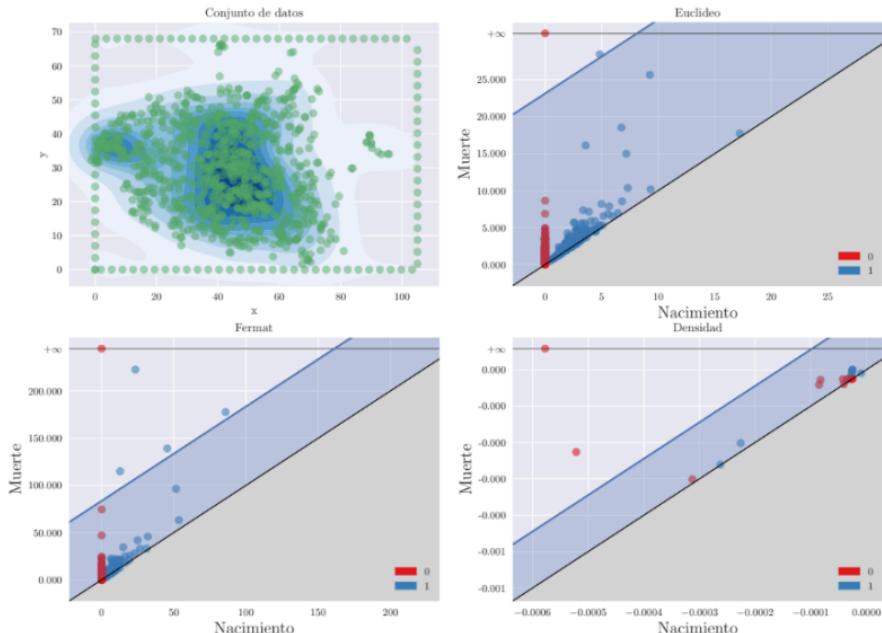
## Mediocampista (5)



## Lateral Izquierdo (8)



## Mediocampista (14)



## Dimensiones Superiores

---

	Método		
Dimensiones $D$	Euclídeo	Fermat	KDE
2	8.99	8.07	2.28
3	8.43	8.00	31.65
4	9.06	8.37	1501.22

KDE demora exponencialmente más tiempo para dimensiones superiores a  $D = 2$

		Circunferencia Uniforme	Circunferencia Gaussiana	Anteojos
Métodos	Agujeros			
Euclídeo	1	100%	100%	0
	2	0	0	100%
Fermat	1	100%	100%	100%
	2	0	0	0
Kde	1	100%	100%	0
	2	0	0	100%

		Circunferencia Uniforme	Circunferencia Gaussiana	Anteojos
Métodos	Agujeros			
Euclídeo	1	100%	100%	0
	2	0	0	100%
Fermat	1	100%	100%	98%
	2	0	0	2%
Kde	1	100%	100%	0
	2	0	0	100%

		Circunferencia Uniforme	Circunferencia Gaussiana	Anteojos
Métodos	Agujeros			
Euclídeo	0	0	0	8%
	1	100%	100%	80%
	2	0	0	12%
	3	0	0	0
Fermat	0	0	0	0
	1	98%	100%	84%
	2	2%	0	14%
	3	0	0	2%
Kde	0	0	0	0
	1	100%	100%	0
	2	0	0	100%
	3	0	0	0

Métodos	Agujeros	Detecciones	
		0	1
Euclídeo	0	100%	0
	1	0	0
Fermat	0	100%	0
	1	0	0
Kde	0	50%	50%
	1	50%	50%

El método KDE falla en detectar la ausencia de agujeros para el 50% de las corridas.

## **Conclusiones y Proximos Pasos**

# Conclusiones y Próximos Pasos

---

## Conclusiones

- La distancia de Fermat supera a los métodos propuestos en la bibliografía en los conjuntos de datos sintéticos.
- Fermat detecta componentes conexas adicionales en presencia de datos atípicos, potencial uso como herramienta de detección de outliers.
- Fermat logra resultados consistentes con la inspección visual en conjuntos de datos reales.

## Próximos pasos

- Explorar el comportamiento de los hiperparámetros en las simulaciones, en particular para  $\lambda$  en la distancia de Fermat.
- Extender el estudio a conjuntos de datos reales más diversos.
- Profundizar en el uso de Fermat como herramienta de detección de datos atípicos.

**¡Muchas Gracias!**

**¿Preguntas?**