

# DETECCIÓN NO-PARAMÉTRICA DE ESTRUCTURAS GEOMÉTRICAS ANÓMALAS EN REDES

QUANTIL S.A.S.

Juan Pablo Lozano

Septiembre - 2016

**1** Descripción del Problema.

**2** Preliminares.

**3** Redes.

# Descripción y Motivación

Detectar cadenas de nodos anómalos en redes.

- Lineales.
- Anillos.
- Lattices 2D.
- Lattices Multidimensionales.

Se tiene una red de  $n$  nodos y se supone que cada nodo  $i$  en la estructura se relaciona con una muestra aleatoria de una variable aleatoria  $Y_i$ .

## Descripción y Motivación

Evaluar la siguiente hipótesis:

- $H_0 = Y_i \sim P \quad \forall i = 1, \dots, n.$
- $H_1 = Y_i \sim Q$  para  $i \in R.$

Donde  $R$  es una región de nodos contiguos de la red,  $P$  y  $Q$  son distribuciones tales que  $P \neq Q$ .

Es una prueba de hipótesis compuesta pues se evalúan todas las regiones posibles de la red.

$P$  y  $Q$  pueden no ser conocidas.

# Descripción y Motivación

Problemas:

- ¿Qué pasa cuando  $n \rightarrow \infty$ ? ¿Qué pasa con las pruebas?
- ¿Que pasa con el tamaño de las regiones anómalas?

# Descripción y Motivación

Aplicaciones:

- Secuencias anómalas de ADN.
- Virus en redes computacionales.
- Manchas en imágenes.

# Preliminares

Sea  $\chi$  un conjunto,  $\mathbb{F}$  su “feature space” con producto interno y  $\psi$  el mapa entre ellos tal que:

$$\psi : \chi \rightarrow \mathbb{F} \text{ donde } x \rightarrow \psi(x).$$

Como  $\mathbb{F}$  tiene producto interno, se puede evaluar como:  $k(x, x') = \langle \psi(x), \psi(x') \rangle$ . A  $k$  se le llama función kernel.

No se necesita saber la estructura de  $\psi$  para evaluar el producto interno. (Truco del kernel).

# Preliminares

Si  $k$  es semi-definido positivo, siempre existe un  $\psi : \mathcal{X} \rightarrow \mathbb{F}$  tal que  $k(x, x') = \langle \psi(x), \psi(x') \rangle$ .

## Teorema (Representación de Riesz)

Si  $A : \mathcal{H} \rightarrow \mathbb{R}$  es un operador lineal acotado en un espacio de Hilbert  $\mathcal{H}$ , existe un  $g_A \in \mathcal{H}$  tal que:

$$A(f) = \langle f, g_A \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$$

# Preliminares

## Definición

Para un conjunto arbitrario  $\chi$  y  $\mathcal{H}$  un espacio de Hilbert sobre las funciones  $f : \chi \rightarrow \mathbb{R}$ .

Se dice que  $\mathcal{H}$  es un espacio de Hilbert con kernel reproducible si:  
 $\forall x \in \chi, \exists M > 0 : L_x[f] := f(x) \leq M \|f\|_H, \forall f \in \mathcal{H}$ .

# Preliminares

## Proposición (Propiedad de Reproducción)

Para cada  $x \in \chi$  existe una función  $k_x \in \mathcal{H}$  tal que:

$$L_x[f] = f(x) = \langle f, k_x \rangle, \forall f \in \mathcal{H}$$

Se dice que  $k$  es el kernel reproducible del punto  $x$ .

El RKHS está completamente caracterizado por su kernel reproducible.

$$\mathcal{H} = \overline{\text{span}[k(x, \cdot) | x \in \chi]}$$

# Preliminares

Cualquier función medible  $f$  sobre un conjunto  $\chi$  es integrable sobre la medida de Dirac. Más aún:

$$\int f(t) d\delta_x(t) = f(x).$$

Cuando  $f$  pertenece al espacio de Hilbert con kernel reproducible de funciones sobre  $\chi$  se tiene:

$$\int f(t) d\delta_x(t) = \int \langle f, k(t, \cdot) \rangle_{\mathcal{H}} d\delta_x(t) \quad (1)$$

$$= \left\langle f, \int k(t, \cdot) d\delta_x(t) \right\rangle_{\mathcal{H}} \quad (2)$$

$$= \langle f, k(x, \cdot) \rangle_{\mathcal{H}} \quad (3)$$

# Preliminares

Se puede definir un feature map  $\psi$  que vaya del espacio de medidas de Dirac a  $\mathcal{H}$  como:

$$\delta_x \rightarrow \int_{\mathcal{X}} k(y, \cdot) d\delta_x(y).$$

La medida de Dirac es una medida de probabilidad sobre  $(\mathcal{X}, \mathcal{A})$ . Donde  $\mathcal{A}$  es una  $\sigma$ -álgebra sobre  $\mathcal{X}$ .

Se puede extender esto a medidas de probabilidad en general.

Si  $M_+^1(\mathcal{X})$  es el espacio de medidas de probabilidad sobre un espacio medible  $\mathcal{X}$ , se puede definir la representación en  $\mathcal{H}$  de cualquier medida de probabilidad  $\mathbb{P}$  como:

$$\mu : M_+^1(\mathcal{X}) \rightarrow \mathcal{H}, \mathbb{P} \rightarrow \int k(x, \cdot) d\mathbb{P}(x) = \mathbb{E}_{\mathbb{P}}[k(x, \cdot)]$$

## Preliminares

A estas funciones se les conoce como embebimientos en espacios de Hilbert. En particular  $\mu_{\mathbb{P}}$  es el embebimiento de media de la medida  $\mathbb{P}$ .

### Proposición

Si  $\mathbb{E}_{\chi \sim \mathbb{P}}[\sqrt{k(\chi, \chi)}] \leq \infty$  entonces  $\mu_{\mathbb{P}} \in \mathcal{H}$  y  $\mathbb{E}_{\mathbb{P}}[f(x)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$

En otras palabras esta es la propiedad de reproducción del operador de esperanza en  $\mathcal{H}$ .

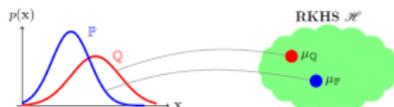


Figure: Ilustración.

# Preliminares

¿Para que se quiere hacer todo esto?

Por medio de funciones de kernel conocidas como kernels característicos, se puede capturar toda la información de una distribución  $\mathbb{P}$ . Lo que quiere decir que  $\mu_{\mathbb{P}}$  es inyectiva. ( $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ ).

Por esta vía se quiere llegar a hacer la prueba de hipótesis que dice que los nodos de una red vienen todos de una misma distribución.

# Preliminares

## Definición (MMD)

Para dos distribuciones de probabilidad  $\mathbb{P}$  y  $\mathbb{Q}$ , se define la medida de máxima discrepancia media (MMD) entre ellas como:

$$MMD[\mathbb{P}, \mathbb{Q}] = \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_{\mathbb{P}}[f(x)] - \mathbb{E}_{\mathbb{Q}}[f(x)].$$

A partir de esta definición se puede mostrar también que  $MMD[\mathbb{P}, \mathbb{Q}] = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$ . Se determina una métrica sobre distribuciones que permite saber si dos distribuciones son iguales o no.

## Preliminares

MMD tiene un estimador no paramétrico insesgado que se define por medio de:

$$MMD_u^2[X, Y] = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(y_i, y_j) \quad (4)$$

$$- \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, y_j) \quad (5)$$

Donde  $X = [x_1, \dots, x_n]$  y  $Y = [y_1, \dots, y_m]$  son muestras de dos distribuciones  $\mathbb{P}$  y  $\mathbb{Q}$ .

# Redes Lineales.

Comenzando con las redes más sencillas: redes lineales.

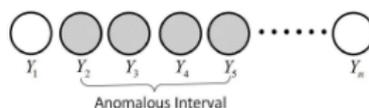


Figure: Red lineal.

## Redes Lineales.

Se denomina  $I$  un conjunto de índices de nodos consecutivos el cual se conoce como intervalo. Se quiere revisar si los nodos del intervalo están asociados a una distribución distinta al resto de los nodos.

Sobre una red de  $n$  nodos, se define el conjunto  $\mathcal{I}_n^{(a)} = \{I : I_{min} < |I| < I_{max}\}$ . Se supone que los intervalos de tamaño delimitado entre  $I_{min}$  y  $I_{max}$  son los posibles candidatos a ser anómalos.

La prueba de hipótesis es entonces:

$H_0$  :  $Y_i$  es i.i.d. para todo  $i = 1, \dots, n$  de una distribución  $\mathbb{P}$ .

$H_1$  : Existe un intervalo  $I \in \mathcal{I}_n^{(a)}$  tal que  $Y_{i \in I} \sim \mathbb{Q}$  y  $Y_{i \notin I} \sim \mathbb{P}$ .

Para  $\mathbb{P} \neq \mathbb{Q}$ .

## Redes Lineales.

La hipótesis alternativa es compuesta ya que hay muchos intervalos de distintos tamaños en el conjunto de posibles candidatos anómalos.

Las distribuciones son arbitrarias y no se conocen a priori.

Es de interés ver que pasa cuando  $n \rightarrow \infty$  y las repercusiones que esto tiene en los límites de los tamaños de intervalos anómalos. Es decir, se necesita cierto tamaño de intervalos para que la prueba funcione bien.

Para medir el desempeño del test se mira el riesgo minimax  $R_m^n = P(H_1|H_0) + \max_{i \in \mathcal{I}_n^{(a)}} P(H_0|H_{1,I})$ .

# Redes Lineales.

## Definición

Un test se dice consistente si el riesgo minimax  $R_m^n \rightarrow 0$  cuando  $n \rightarrow \infty$ .

## Redes Lineales.

Para cada intervalo  $I$  se toma  $Y_I$  como las muestras del intervalo  $I$  y  $Y_{\bar{I}}$  como las muestras fuera del intervalo  $I$ .

Se computa  $MMD_{u,I}^2(Y_I, Y_{\bar{I}})$  para cada intervalo  $I \in \mathcal{I}_n^{(a)}$ . Lo cual es una estimación de  $MMD_u(\mathbb{P}, \mathbb{Q})$ .

Bajo la hipótesis nula, todos los nodos son generados por la misma distribución lo que implica que para todos los intervalos evaluados  $MMD_{u,I}^2(Y_I, Y_{\bar{I}}) = 0$ .

Bajo la hipótesis alternativa, existe algún intervalo anómalo que hará que  $MMD_{u,I}^2(Y_I, Y_{\bar{I}})$  se aleje de 0 pues las distribuciones no son iguales.

## Redes Lineales.

Específicamente se evalúa el siguiente test para aceptar o rechazar la hipótesis nula:

$$\max_{I \in \mathcal{I}_n^{(a)}} MMD_{u,I}^2(Y_I, Y_{\bar{I}}) \begin{cases} \geq t, & \text{se rechaza } H_0 \\ \leq t, & \text{se acepta } H_0 \end{cases}$$

Donde  $t$  es un umbral.

# Redes Lineales.

## Teorema

Suponiendo el test anterior. Si el kernel está acotado  $0 < k(x, y) \leq K$  para todo  $x$  y  $y$ . Entonces los errores de tipo I y tipo II están acotados por:

$$P(H_0|H_1) \leq \sum_{i: I_{min} \leq |I| \leq I_{max}} e^{-\frac{t^2 |I| (n-|I|)}{8K^2 n}}$$

$$P(H_1|H_0) \leq e^{-\frac{(MMD^2[p, q] - t)^2 |I| (n-|I|)}{8K^2 n}}, \text{ para todo } I \in \mathcal{I}_n^{(a)}$$

Donde  $t$  es el umbral del test que satisface  $t < MMD^2[p, q]$ .

## Redes Lineales.

Además, el test es (universalmente) consistente si:

$$I_{min} \geq \frac{16K^2(1+\eta)}{t^2} \log(n) \rightarrow I_{min} = \omega(\log(n))$$

$$I_{max} \leq n - \frac{16K^2(1+\eta)}{t^2} \log \dots \log (n)$$

un numero arbitrario  $k$

$$\rightarrow n - I_{max} = \Omega(\log \dots \log (n))$$

un numero arbitrario  $k$

Para cualquier constante positiva  $\eta$ .

## Redes Lineales.

Si se conocen las distribuciones a priori, se puede definir un umbral  $t = (1 - \delta)MMD^2[\mathbb{P}, \mathbb{Q}]$  para  $\delta \in (0, 1)$ .

Si no se conocen las distribuciones entonces  $t$  se escala a medida que  $\lim_{n \rightarrow \infty} t_n = 0$ .

## Redes de Anillos.

La idea es similar a las redes lineales. El número de intervalos es diferente a las redes lineales.

El número de intervalos es mayor pues no se reduce a medida que los nodos del intervalo aumenten.

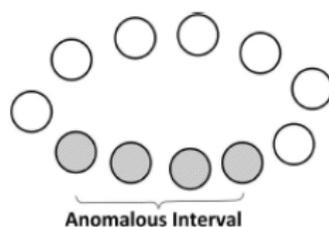


Figure: Red Anillo.

## Redes Anillos.

Específicamente se evalúa el siguiente test para aceptar o rechazar la hipótesis nula:

$$\max_{I \in \mathcal{I}_n^{(a)}} MMD_{u,I}^2(Y_I, Y_{\bar{I}}) \begin{cases} \geq t, & \text{se rechaza } H_0 \\ \leq t, & \text{se acepta } H_0 \end{cases}$$

Donde  $t$  es un umbral.

## Redes Anillos.

Al igual que antes los errores se pueden acotar:

$$P(H_0|H_1) \leq e^{2\log(n) - \frac{t^2 \min\{I_{\min}(n-I_{\min}), I_{\max}(n-I_{\max})\}}{8K^2n}}$$

$$P(H_1|H_0) \leq e^{-\frac{(MMD^2[p,q]-t)^2 |I|(n-|I|)}{8K^2n}}, \text{ para todo } I \in \mathcal{I}_n^{(a)}$$

Donde  $t$  es el umbral del test que satisface  $t < MMD^2[p, q]$ .

## Redes Anillos.

Además, el test es (universalmente) consistente si:

$$I_{min} \geq \frac{16K^2(1 + \eta)}{t^2} \log(n) \rightarrow I_{min} = \omega(\log(n))$$

$$I_{max} \leq n - \frac{16K^2(1 + \eta)}{t^2} \log(n) \\ \rightarrow n - I_{max} = \omega(\log(n))$$

Para cualquier constante positiva  $\eta$ .

## Lattice 2D.

Se pinta un lattice en 2D y se quiere encontrar un disco anómalo.

Ya no se observan intervalos sino discos. Se define entonces el conjunto de todos los posibles discos anómalos.

$\mathcal{D}_n^{(a)} = \{D : D_{min} \leq |D| \leq D_{max}\}$ . Donde  $|D|$  es el número de nodos que caen en el disco.

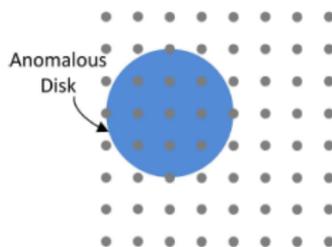


Figure: Disco anómalo en red lattice en 2D.

## Lattice 2D.

Específicamente se evalúa el siguiente test para aceptar o rechazar la hipótesis nula:

$$\max_{D \in \mathcal{D}_n^{(a)}} MMD_{u,D}^2(Y_D, Y_{\bar{D}}) \begin{cases} \geq t, & \text{se rechaza } H_0 \\ \leq t, & \text{se acepta } H_0 \end{cases}$$

Donde  $t$  es un umbral.

## Lattice 2D.

Al igual que antes los errores se pueden acotar:

$$P(H_0|H_1) \leq e^{3\log(n) - \frac{t^2 \min\{D_{\min}(n^2 - D_{\min}), D_{\max}(n^2 - D_{\max})\}}{8K^2 n^2}}$$

$$P(H_1|H_0) \leq e^{-\frac{(MMD^2[p,q] - t)^2 |D|(n^2 - |D|)}{8K^2 n}}, \text{ para todo } D \in \mathcal{D}_n^{(a)}$$

Donde  $t$  es el umbral del test que satisface  $t < MMD^2[p, q]$ .

## Lattice 2D.

Además, el test es (universalmente) consistente si:

$$I_{min} \geq \frac{24K^2(1 + \eta)}{t^2} \log(n) \rightarrow I_{min} = \omega(\log(n))$$

El  $I_{max}$  está acotado por  $cn^2$  donde  $c < 1$  por la geometría.

Para cualquier constante positiva  $\eta$ .

## Resultados.

Tomando una red lineal y de anillo de 200 nodos, se calculo el riesgo minimax para distintos límites del tamaño de los intervalos.

Utilizaron un kernel gaussiano y definieron  $\mathbb{P} \sim \mathcal{N}(0, 1)$  y  $\mathbb{Q} \sim \mathcal{N}(1, 1)$ .

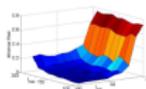


Figure 4: Minimax risk for a line network

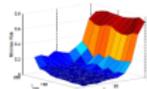


Figure 5: Minimax risk for a ring network

Riesgo.png

Figure: Graficas del desempeño del algoritmo.

# Resultados.

Table 1: Minimax risk for a line network

$I_{min} \setminus I_{max}$	100	110	130	160	190
1	0.73	0.73	0.72	0.76	0.76
11	0.58	0.61	0.55	0.59	0.59
31	0.09	0.11	0.10	0.09	0.27
61	0.03	0.03	0.05	0.06	0.21
91	0.02	0.02	0.04	0.05	0.24

Table 2: Minimax risk for a ring network

$I_{min} \setminus I_{max}$	100	110	130	160	190
1	0.73	0.78	0.74	0.74	0.71
11	0.58	0.61	0.53	0.60	0.63
31	0.09	0.10	0.12	0.12	0.27
61	0.03	0.03	0.03	0.05	0.27
91	0.02	0.03	0.03	0.05	0.24

Figure: Tablas del desempeño del algoritmo.

Se nota como a medida que se disminuye  $I_{max}$  y se aumenta  $I_{min}$  se mejora el desempeño. Esto es lógico pues el número de candidatos anómalos disminuye (reduce la dificultad de detección).

Además aumenta el número de muestras mínimas que deben haber dentro y fuera del intervalo, mejorando la información de las distribuciones.

## Resultados.

Se compara el método contra distintos tests para una red lineal de 100 nodos:

- t-test
- Smirnov (KS multidimensional)
- KDR (Kernel dimensionality reduction) (PCA con kernels)
- KFDA (Kernel fisher discriminant analysis)
- MMD.

Se tomó  $\mathbb{P} \sim \mathcal{N}(0, 2)$  y  $\mathbb{Q}$  como una mezcla de gaussianas  $\mathcal{N}(-1, 1)$  y  $\mathcal{N}(1, 1)$  con igual probabilidad. De esta manera  $\mathbb{P}$  y  $\mathbb{Q}$  tienen la misma media y varianza.

# Resultados.

Table 3: Comparison of nonparametric approaches over a line network

$(I_{\min}, I_{\max})$	t-test	Smirnov	KFDA	KDR	MMD
(10,95)	0.90	0.92	0.70	0.66	<b>0.66</b>
(10,50)	0.88	0.90	0.51	0.56	<b>0.55</b>
(45,95)	0.89	0.93	0.54	0.43	<b>0.43</b>
(45,50)	0.83	0.62	0.06	0.06	<b>0.05</b>

Figure: Tablas de comparacion de algoritmos.

Los mejores resultados se tienen para KDR y MMD mientras que el t-test y Smirnov no son muy buenos.

T-test compara medias y desviaciones estándar mientras que para Smirnov se necesitan muchos datos.

# Referencias

- Zou S., Yiang Y., Poor V., 2016. Nonparametric Detection of Geometric Structures over Networks.
- Muandet K. et al, 2016. Kernel Mean Embedding of Distributions: A Review and Beyond.
- Gretton A., 2012. A Kernel Two-Sample Test.