Aprendizaje Automático para Datos en Grafos

Laboratorio 4

Graciana Castro
4.808.848-2
gcastro@fing.edu.uy

Julian O'Flaherty
6.285.986-9
julian.o.flaherty@fing.edu.uy

1. Introducción

2. Inferencia de topología en datos sintéticos

2.1. Generación del grafo sintético

Para generar un grafo sintético, utilizaremos el siguiente algoritmo:

1. Sorteamos N puntos de forma uniforme en el cuadrado $[0,1] \times [0,1]$. Esos serán los vértices de nuestro grafo.

2. Para cada par de puntos i y j que sorteamos antes, tomamos como peso de la arista

$$w_{ij} = \begin{cases} 0 & \text{si } i = j \\ e^{-\frac{d(i,j)}{2\sigma^2}} & \text{si } i \neq j \end{cases}$$

   donde $d(i,j)$ es la distancia euclidea en $\mathbb{R}^2$ y σ es un parámetro fijo.

3. Descartamos las aristas cuyo peso w_{ij} sea menor que un número fijo $r > 0$.

En la figura 1 se muestra un grafo sintético generado con el algoritmo anterior para $N = 50$, $\sigma = 0{,}5$ y $r = 0{,}6$.

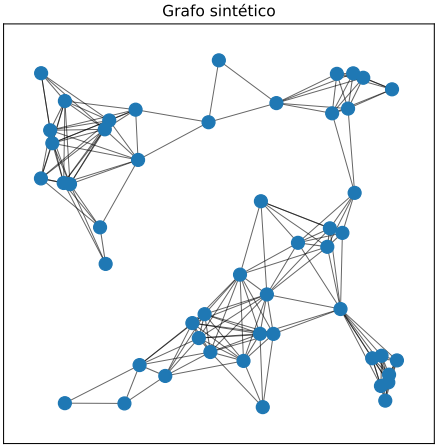


Figura 1: Grafo sintético generado con $N = 50$, $\sigma = 0{,}5$ y $r = 0{,}6$.

A cada nodo del grafo sintetico, le asignaremos una señal $x_i \sim \mathcal{N}(0, \mathbf{L}^\dagger)$, donde $\mathbf{L}^\dagger$ es la pseudoinversa de la matriz laplaciana $\mathbf{L}$ del grafo. En la figura 2 se muestra un *sample* de señales donde cada columna j es la señal asociada al nodo j del grafo. La cantidad de filas, o el tamaño de la muestra, es el parámetro n_{samples} del algoritmo, por lo que si n_{samples} es menor que N, la matriz de covarianza empírica no es invertible.

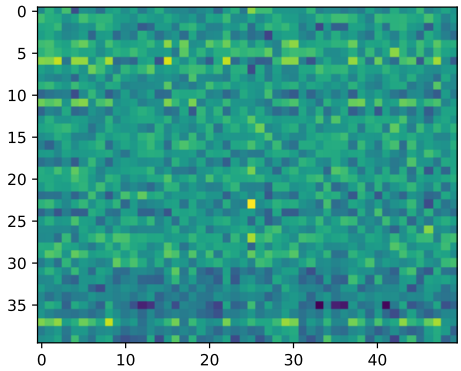


Figura 2: Sample de señales asignadas a los nodos del grafo sintético.

2.2. Estimación de estructura con Graphical Lasso

Estimaremos la estructura del grafo a partir de la matriz de datos utilizando el método *Graphical Lasso*. Este método busca al estimador de máxima verosimilitud de la matriz de precisión $\boldsymbol{\Theta}$ que cumple:

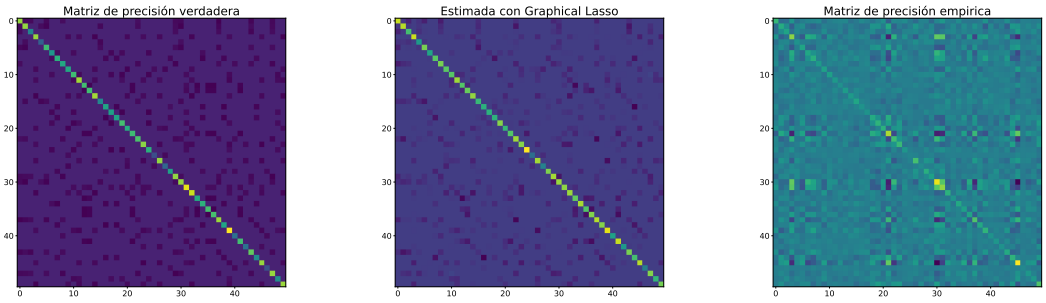
$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta} \succeq 0} \left\{ \log\det \boldsymbol{\Theta} - \operatorname{trace}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \lambda \|\boldsymbol{\Theta}\|_1 \right\} \tag{1}$$

Una de las propiedades importantes del Graphical Lasso, es que cuando $\lambda = 2\sqrt{\frac{\log N}{P}}$,

$$\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_2 \leq \sqrt{\frac{d_{\text{máx}}^2 \log N}{P}} \quad \text{w.h.p.}$$

con el parámetro λ siendo parámetro de regularización. En la figura 3 se puede observar la estimación obtenida con *Graphical Lasso* comparada con la estimación empírica cuando el número de samples utilizado n_{samples} es menor que la cantidad de nodos N. En este caso, la estimación empírica resulta mala, dado que la matriz de covarianza empírica no es invertible. El método de *Graphical Lasso* obtiene un resultado muy similar a la matriz de precisión verdadera, donde el valor de λ se calcula mediante validación cruzada.

En la figura 4 observamos el efecto del variar el valor de λ sobre la matriz de precisión estimada. A medida que aumenta la regularización, la matriz va perdiendo ceros tendiendo a una matriz diagonal, lo cual es esperable dado que el valor de λ está asociado a la norma 1 de la matriz de precisión.



(a) Precisión verdadera

(b) Estimación mediante Graphical Lasso

(c) Estimación empírica ($\Theta = \Sigma^{-1}$)

Figura 3: Comparación entre la matriz de precisión verdadera, la estimación mediante Graphical Lasso y la estimación obtenida con empíricamente ($\Theta = \Sigma^{-1}$), cuando $n_{\text{samples}} = 40$ y $N = 50$

2.3. Estimación de estructura con Meinshausen y Bühlmann

2.4. Estimación de estructura con Meinshausen y Bühlmann

El método de Meinshausen y Bühlmann propone una alternativa eficiente para la estimación de la estructura de grafos mediante selección de vecindarios con Lasso. Este método es particularmente útil en problemas de alta dimensionalidad donde $p \gg n$, ya que permite identificar los ceros estructurales en la matriz de precisión que representan restricciones de independencia condicional entre variables.

La idea central consiste en estimar el vecindario de cada nodo de forma individual mediante regresión. Para cada nodo a, se define su vecindario como el conjunto mínimo de variables que, al ser conocidas, hacen que X_a sea condicionalmente independiente del resto. Esto se formula como un problema de regresión Lasso donde X_a es la variable respuesta y las demás son predictoras:

$$\hat{\theta}_{a,\lambda} = \underset{\theta:\theta_a=0}{\operatorname{argmin}} \left(n^{-1} \|X_a - X\theta\|_2^2 + \lambda \|\theta\|_1 \right) \tag{2}$$

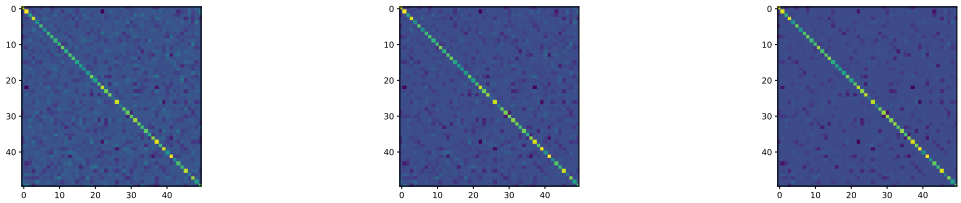
El vecindario estimado de a se determina por los coeficientes no nulos de $\hat{\theta}_{a,\lambda}$. Una vez obtenidos todos los vecindarios, el conjunto de aristas del grafo se construye mediante la regla AND:

$$\hat{E}_{\lambda,\wedge} = \{(a,b) : a \in \hat{ne}_{\lambda,b} \wedge b \in \hat{ne}_{\lambda,a}\} \tag{3}$$

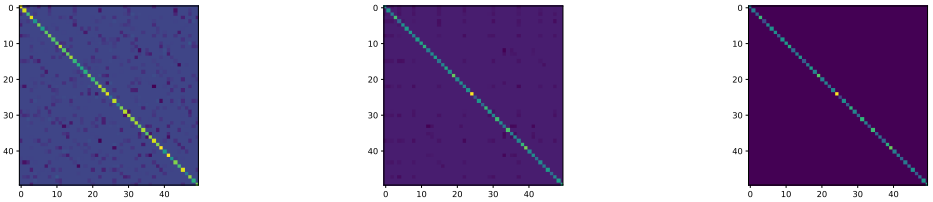
Una propiedad importante de este método es que mantiene consistencia incluso cuando $p \gg n$, siempre que el parámetro λ se elija adecuadamente.

Para implementar este método, seguimos el siguiente algoritmo:

1. Normalizamos la matriz de datos $\mathbf{X}$ por columnas.
2. Inicializamos una matriz $\mathbf{B}$ de ceros de tamaño $p \times p$.
3. Para cada nodo j del grafo:
   a) Extraemos la columna j de $\mathbf{X}$ como variable objetivo $\mathbf{y}_j$.



(a) $\lambda = 0{,}01$ (b) $\lambda = 0{,}02$ (c) $\lambda = 0{,}05$



(d) $\lambda = 0{,}1$ (e) $\lambda = 0{,}5$ (f) $\lambda = 1$

Figura 4: Estimación mediante Graphical Lasso para distintos valores de λ. La matriz de precisión verdadera se muestra en la figura 3a. No se muestran valores menores a 0,01 porque hay poca diferencia notable, y a partir de 0,005 el sistema queda mal condicionado y no puede ser resuelto.

b) Utilizamos las demás columnas como matriz de predictores $\mathbf{X}_j$.

c) Resolvemos el problema de regresión Lasso para obtener los coeficientes $\hat{\boldsymbol{\theta}}_j$.

d) Insertamos los coeficientes en la fila j de $\mathbf{B}$, colocando un cero en la posición j.

4. Devolvemos la matriz $\mathbf{B}$ que contiene los coeficientes de regresión para cada nodo.

En la figura 5 se observa que el método de Meinshausen y Bühlmann logra capturar la estructura principal de la matriz de coeficientes verdadera, identificando correctamente las conexiones más relevantes. Sin embargo, aparecen algunos falsos positivos y negativos, lo que refleja el compromiso entre detección y ruido inherente al uso de Lasso.

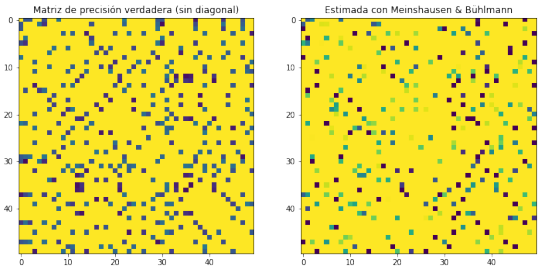


Figura 5: Comparación entre la matriz de coeficientes verdadera y la matriz de coeficientes estimada con el método de Meinshausen y Bühlmann.

2.5. Estimación de estructura con Kalofolias

El método propuesto por Kalofolias plantea aprender la estructura de un grafo a partir de señales que residen en sus nodos, asumiendo que estas señales cambian suavemente entre nodos conectados. La contribución clave consiste en reformular el problema tradicional de minimización del término de suavidad $\text{tr}(X^T LX)$ mostrando que es equivalente a una norma ℓ_1 ponderada aplicada a la matriz de adyacencia W:

$$\text{tr}(X^T LX) = \frac{1}{2}\|W \circ Z\|_{1,1} \tag{4}$$

donde $Z_{i,j} = \|x_i - x_j\|_2$ es la matriz de distancias por pares entre las señales. Esta equivalencia permite formular el problema de aprender el grafo como:

$$\min_{W \in \mathcal{W}_m} \|W \circ Z\|_{1,1} - \alpha \mathbf{1}^T \log(W\mathbf{1}) + \beta\|W\|_F^2 \tag{5}$$

El término logarítmico $-\alpha \mathbf{1}^T \log(W\mathbf{1})$ actúa sobre el vector de grados de los nodos, promoviendo la conectividad del grafo, mientras que el término $\beta\|W\|_F^2$ controla la dispersión de los pesos. Los parámetros $\alpha > 0$ y $\beta \geq 0$ permiten balancear estos objetivos.

En la figura 6 se presentan los resultados obtenidos con diferentes combinaciones de α y β, comparados con la matriz verdadera. Se aprecia cómo estos parámetros influyen en la estructura del grafo aprendido, permitiendo ajustar el balance entre conectividad y dispersión de pesos según las características deseadas.

2.6. Comparación y visualización de los grafos aprendidos

En la figura 7 se presenta una comparación visual entre el grafo verdadero y los grafos aprendidos mediante los métodos de Graphical Lasso, Meinshausen y Bühlmann, y Kalofolias. Cada subfigura muestra la estructura de conexiones entre nodos, permitiendo observar las similitudes y diferencias en la topología inferida por cada método en relación con el grafo original.

La comparación visual de la figura 7 muestra que Graphical Lasso recupera bien la estructura principal del grafo, aunque introduce algunas conexiones espurias; Meinshausen y Bühlmann produce un grafo más disperso, capturando solo las conexiones más fuertes y perdiendo muchas débiles; y Kalofolias logra un balance intermedio, con una topología similar pero no idéntica al grafo original.

3. Inferencia de topología en MNIST

Para esta parte trabajaremos el dataset MNIST, que contiene imágenes de dígitos escritos a mano. El objetivo es aprender la estructura del grafo que conecta las imágenes basándonos en sus características visuales, y luego utilizar esta estructura para realizar clustering espectral y evaluar su desempeño.

3.1. Descripción del dataset y preprocesamiento

Para trabajar, nos quedaremos con un subconjunto de 2000 imágenes (doscientas por cada dígito) de dimensión 28×28 píxeles. Cada imagen la convertimos en un vector de dimensión 784 (aplanando la matriz) y normalizamos los valores de píxeles al rango $[0, 1]$.

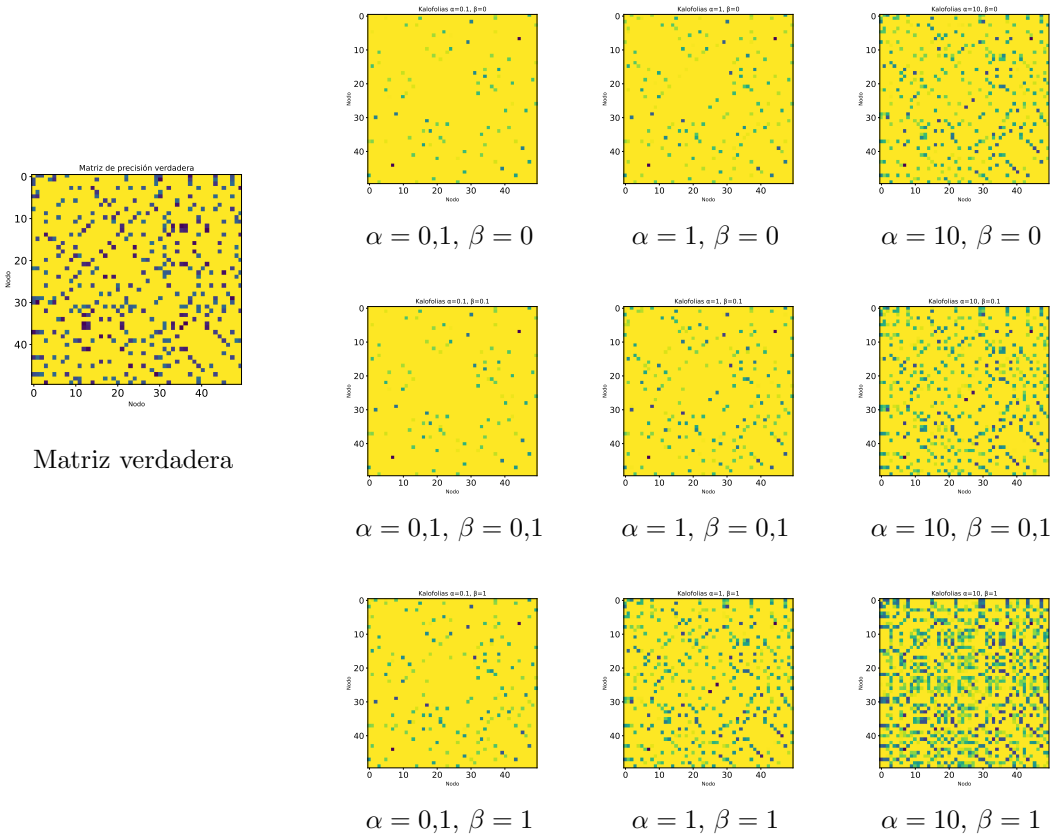


Figura 6: Resultados del método de Kalofolias para distintos valores de α y β, comparados con la matriz verdadera (izquierda) y una grilla 3x3 de variaciones (derecha).

3.2. Construcción de la matriz de features y t-SNE

3.3. Grafo aprendido con Meinshausen y Bühlmann

3.4. Grafo aprendido con Kalofolias y umbralización

3.5. Clustering espectral sobre grafos aprendidos

3.5.1. Métricas de evaluación

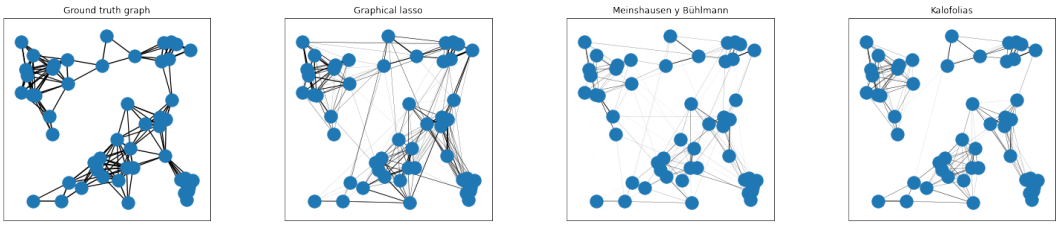
3.5.2. Resultados y comparación

3.6. Visualización en el plano t-SNE por método

3.7. Análisis de aciertos y errores por método

4. Conclusiones

5. Referencias



(a) Grafo verdadero (b) Graphical Lasso (c) Meinshausen y Bühlmann (d) Kalofolias

Figura 7: Comparación visual de los grafos: verdadero y aprendidos por distintos métodos.