Selección de modelos gráficos no-dirigidos en el contexto de alta dimensión

Violeta Roizman

Directora: Dra. Florencia Leonardi
Codirectora: Dra. Mariela Sued

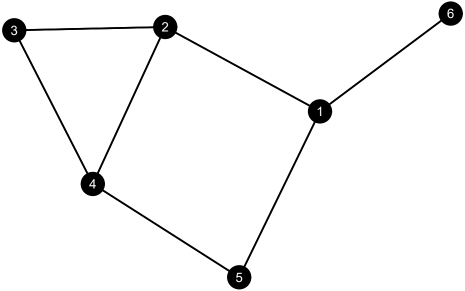
2017-05-28

Selección de modelos gráficos no-dirigidos en el contexto de alta dimensión

- modelos gráficos (?)
- no-dirigidos (?)
- contexto de alta dimensión (?)

Modelos gráficos

Grafo $\mathcal{G} = (V, E)$



$\mathcal{G}$ representa $\vec{X} = (X_1, ..., X_p)$ con $|V| = p$

Las aristas codifican dependencias condicionales entre variables

Modelos gráficos no-dirigidos

También llamados campos aleatorios de Markov

Propiedades:

- aristas sin orientación
- relaciones simétricas
- cada vértice es una variable
- **Un grafo $\mathcal{G}$ representa a un vector $\vec{X} \sim D$ si**:

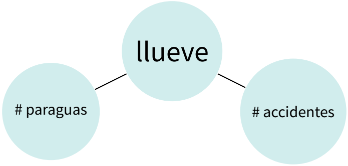
$$(i,j) \notin E \Longrightarrow X_i \perp X_j | X_{V \setminus \{i,j\}}$$

Independencia condicional

Definición

X, Y, Z variables aleatorias

$$X \perp Y | Z \iff p(x, y|z) = p(x|z)p(y|z)$$



Intuitivamente

Suponiendo que dado el valor de Z queremos adivinar el valor de X, ¿nos ayudaría conocer el valor de Y? En caso negativo tendremos que $X \perp Y|Z$.

Propiedades de Markov

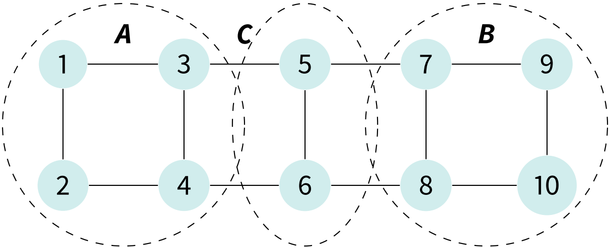
Dado un modelo gráfico decimos que se cumple:

- (**P**) Propiedad de Markov de a pares si $\forall (i,j) \notin E$:

$$X_i \perp X_j | X_{V \setminus \{i,j\}}.$$

- (**G**) Propiedad de Markov global si para $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C} \subset V$ disjuntos tales que $\boldsymbol{C}$ separa a $\boldsymbol{A}$ de $\boldsymbol{B}$ en $\mathcal{G}$ entonces

$$X_A \perp X_B | X_C.$$



Objetivo

Dada una muestra finita $\{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$ de $\vec{X} \sim D$

$$\boldsymbol{X} = \begin{bmatrix} \text{———} & \boldsymbol{x}^{(1)} & \text{———} \\ \text{———} & \boldsymbol{x}^{(2)} & \text{———} \\ & \vdots & \\ \text{———} & \boldsymbol{x}^{(n)} & \text{———} \end{bmatrix}$$

Objetivo: estimar $\mathcal{G}$ que represente a $\vec{X}$ a partir de $\boldsymbol{X}$

Teorema Hammersley-Clifford

Si D es estrictamente positiva entonces (**G**) $\iff$ (**P**)

Nos permite estimar el grafo con la propiedad (**P**) y deducir con (**G**)

Cuando ya estimamos el grafo

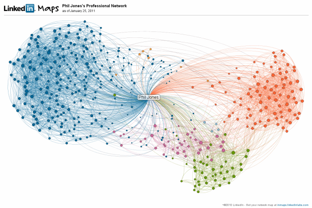
¿Para qué me sirve?

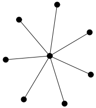
- **Análisis cualitativo:**

  Leer independencias condicionales

  ¿Las variables están agrupadas?

  ¿Existen variables importantes?





- **Inferencia:** a partir de el valor de algunas de las variables del grafo predecir el valor de otras

Alta dimensión

Base de datos:

- n = número de observaciones
- p = número de variables

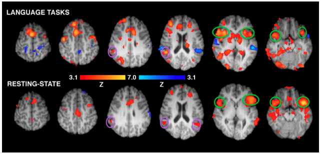
Contexto de alta dimensión:

$$n < p$$

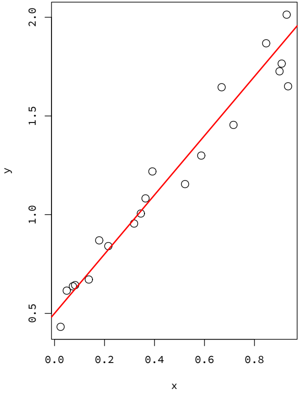
Algunas disciplinas:

genética

neurociencia



Ejemplo: problema de regresión



Cuadrados mínimos:

$$\underset{\boldsymbol{\beta}}{\text{minimizar}} \frac{1}{2n} \sum_{i=1}^{n} ||\mathbf{y} - \beta \mathbf{X}||_2^2$$

Si $\boldsymbol{p} > \boldsymbol{n}$:

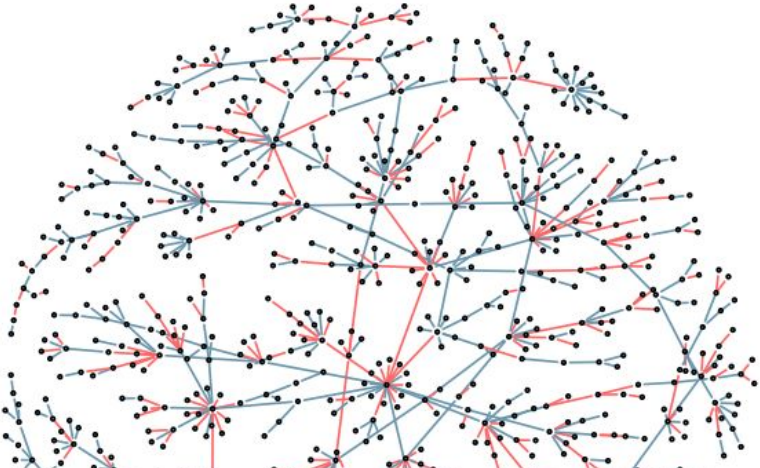
- No puede resolverse con las ecuaciones normales
- No tiene un único mínimo

Contexto de alta dimensión en modelos gráficos

muchas variables = muchos vértices

de parámetros a estimar = $\frac{p(p-1)}{2}$

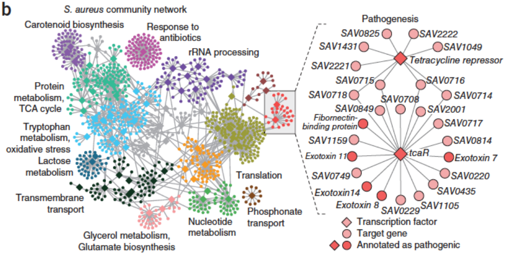
esperamos: pocas aristas



Contenido

- Dos ejemplos
- (Regresión lineal en alta dimensión)
- Modelo gaussiano
- (Regresión logística en alta dimensión)
- Modelos discretos
- Conclusiones y trabajo a futuro

Ejemplo 1: red de genes

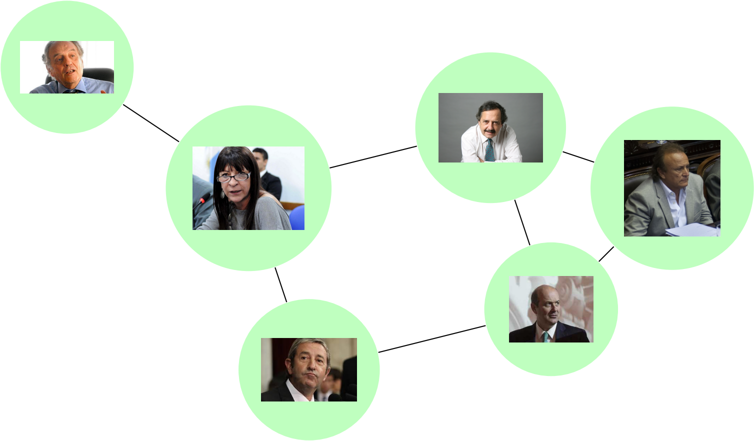


Ejemplo 2: diputados

Quiero saber como se relacionan los diputados al votar

Ley	**del Sel**	**Conti**	**Heller**	**Cobos**	**Sturtzenegger**	.
1	0	1	1	0	0	.
2	0	1	0	0	0	.
3	1	1	1	1	0	.
4	1	1	1	1	1	.
5	1	1	1	1	1	.
6	0	0	0	1	0	.
.	.	.	.	.	.	.

Buscamos algo así



(Regresión lineal en alta dimensión)

Penalización norma l_1: método *lasso*

Volvemos al problema de regresión:

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimizar}} \frac{1}{2n}\sum_{i=1}^{n} ||\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$$

Propiedades:

- convexo (único mínimo, fácil de optimizar)
- $\lambda \to 0$ **CUADRADOS MÍNIMOS** y $\lambda \to \infty$ $\hat{\boldsymbol{\beta}} = \mathbf{0}$
- introduce sesgo - menos varianza
- tiende a anular coeficientes (carácter selector)
- selecciona como mucho n variables

¿Por qué norma 1?: *ridge* usa $||\beta||_2$
estima a todos los coeficientes distintos $\neq 0$

Modelo Gaussiano

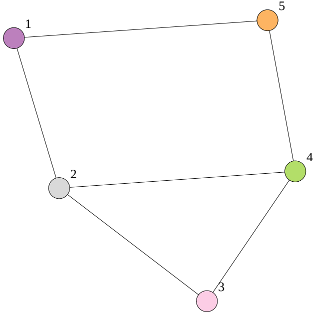
Normal multivariada

Si $\vec{X} \sim N(\boldsymbol{\mu}, \Sigma)$:

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Nos basamos en esta propiedad:

$$\Theta_{ij} := \Sigma^{-1}_{ij} = 0 \iff X_i \perp X_j | X_{V\setminus\{i,j\}} \iff (i,j) \notin E$$



$$\boldsymbol{\Theta} = \begin{pmatrix} 1 & 0.5 & \mathbf{0} & \mathbf{0} & 0.5 \\ 0.5 & 1 & 0.5 & 0.3 & \mathbf{0} \\ \mathbf{0} & 0.5 & 1 & 0.5 & \mathbf{0} \\ \mathbf{0} & 0.3 & 0.5 & 1 & 0.6 \\ 0.5 & \mathbf{0} & \mathbf{0} & 0.6 & 1 \end{pmatrix}$$

Graphical Lasso

Consiste en maximizar la función de verosimilitud penalizada:

$$l_\lambda(\Theta; \boldsymbol{x}^{(1)}, .., \boldsymbol{x}^{(n)}) = \log(\det(\Theta)) - \mathrm{tr}(\mathbf{S}\Theta) - \lambda\|\Theta\|_1$$

con

$$\mathbf{S} = n^{-1}\sum_{i=1}^{n}(\boldsymbol{x}^{(i)} - \widehat{\boldsymbol{\mu}}_{MV})(\boldsymbol{x}^{(i)} - \widehat{\boldsymbol{\mu}}_{MV})^T$$

y

$$\|\Theta\|_1 = \sum_{j\neq k}|\Theta_{jk}|$$

Una opción: Elección de λ con validación cruzada con verosilmilitud gaussiana negativa como función de pérdida

Graphical Lasso

Ya estimada Σ^{-1}, estimamos E_*:

$$\widehat{E} = \{(j,k) \in V \times V; \widehat{\Sigma_{jk}^{-1}} \neq 0\}.$$

Bajo ciertas condiciones de regularidad con alta probabilidad:

$$E_* \subseteq \widehat{E}$$

Las condiciones para la consistencia son bastante restrictivas

Surgen variaciones de este método (*Adaptive GLasso*)

Nodewise regression

Basándonos en esto:

$$X_j = -\sum_{k \in V \setminus \{j\}} \frac{\Theta_{jk}}{\Theta_{jj}} X_k + \varepsilon_j$$

Para cada $j \in V$ estimo $\widehat{\beta_k^{(j)}}$ haciendo la regresión lineal de X_j vs $X_{V \setminus j}$

$$\widehat{\beta_k^{(j)}} \neq 0 \iff \Theta_{jk} \neq 0$$

Necesito que Θ sea simétrica $\Longrightarrow$ Regla **OR** o **AND**

Stability Selection

Método basado en re-muestreo combinado con *GLasso*

Consiste en:

1. Elegir Λ una secuencia de posibles valores λ
2. Elegir un valor 50 $< \pi <$ 100 de corte
3. Estimar la estructura del grafo *nrep* veces para cada λ
4. Poner cada arista en el grafo final si para alguno de los $\lambda \in \Lambda$ aparece en más del π % de los grafos estimados

La estimación del grafo no depende fuertemente de la elección de Λ ni de la elección π

Estimando el grafo de genes ($p = 160, n = 115$)

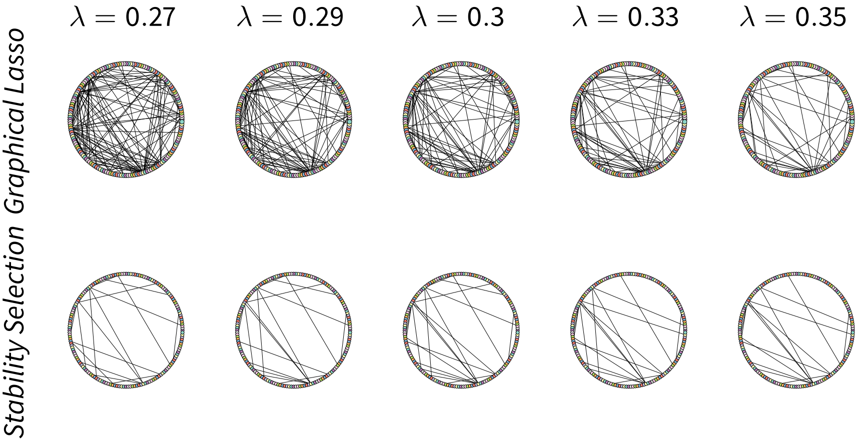


Figura: Comparación de la estimación del grafo para el conjunto de datos `riboflavin` ***GLasso*** vs. ***Stability Selection*** con distintos valores del parámetro λ

Estimando el grafo vacío

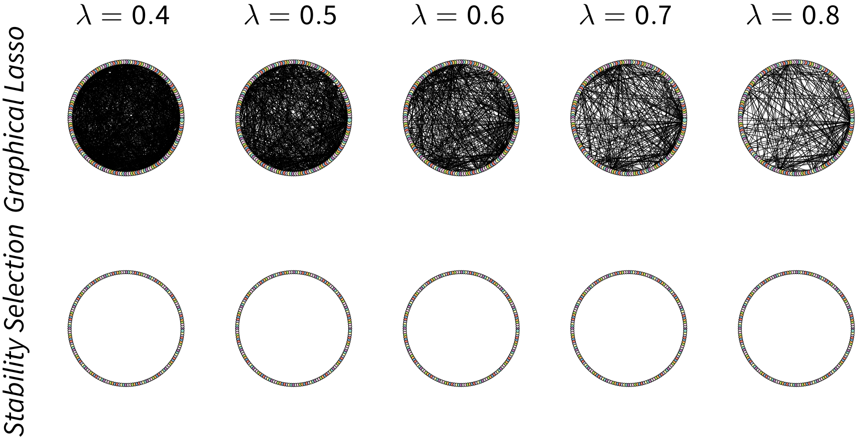


Figura: Comparación de la estimación del grafo vacío con ***GLasso*** vs. ***Stability Selection*** con distintos valores del parámetro λ

(Regresión logística en alta dimensión)

Regresión logística

Clásico método de **clasificación**

- $Y = 0$ o 1 es la variable respuesta
- $\vec{X}$ variables explicativas

Modelo la probabilidad $p(\vec{X})$ de que la variable respuesta valga uno:

$$logit(p(\vec{X})) := \log\left(\frac{p(\vec{X})}{1-p(\vec{X})}\right) = \beta_0 + \boldsymbol{\beta}_1 X_1 + \ldots + \boldsymbol{\beta}_p X_p, \qquad (\star)$$

Función de verosimilitud:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y^{(i)} \log(p(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - p(\mathbf{x}^{(i)})).$$

Agrego un término de penalización de la norma de coeficientes β

Regresión logística penalizada

Analizamos los datos `prostate`: expresión de 6033 genes obtenidas de 102 muestras de pacientes operados

Clasificar a estas muestras como tumores malignos o benignos

Los datos fueron divididos al azar en:
entrenamiento (75 %) y validación (25 %).

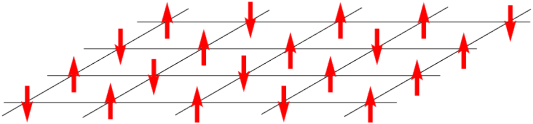
	Porcentaje de acierto	**Variables seleccionadas**
penalización l_1	84 %	23
penalización l_2	92 %	todas

Modelos discretos

Modelo Ising

Probabilidad puntual conjunta para $\mathbf{x} \in \{0,1\}^p$:

$$p(\mathbf{x}, \Theta, \alpha) = \exp\left[\sum_{j \in V} \alpha_j \mathbf{x}_j + \sum_{(j,k) \in E} \Theta_{jk} \mathbf{x}_j \mathbf{x}_k - \Phi(\Theta, \alpha)\right]$$



2–D Ising Model

De acá se deduce:

$$X_j \perp X_k | X_{V \setminus \{j,k\}} \iff \Theta_{jk} = 0.$$

Nodewise logistic regression

Se basa en:

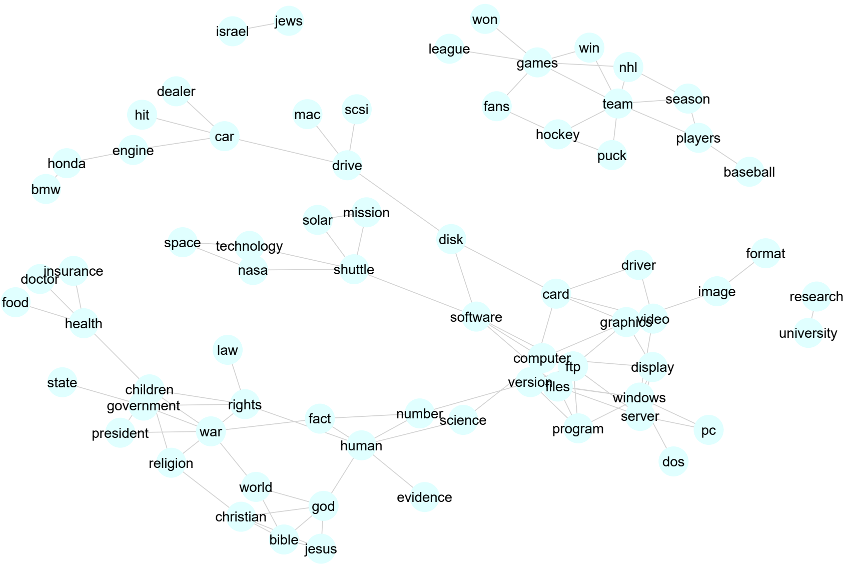
$$logit(P[X_j = 1 | X_{V \setminus j}]) = \alpha_j + \sum_{k \neq j} \Theta_{jk} X_k.$$

Para cada $j \in V$:
Hacemos una regresión logística de la variable X_j en función de $X_{V \setminus j}$

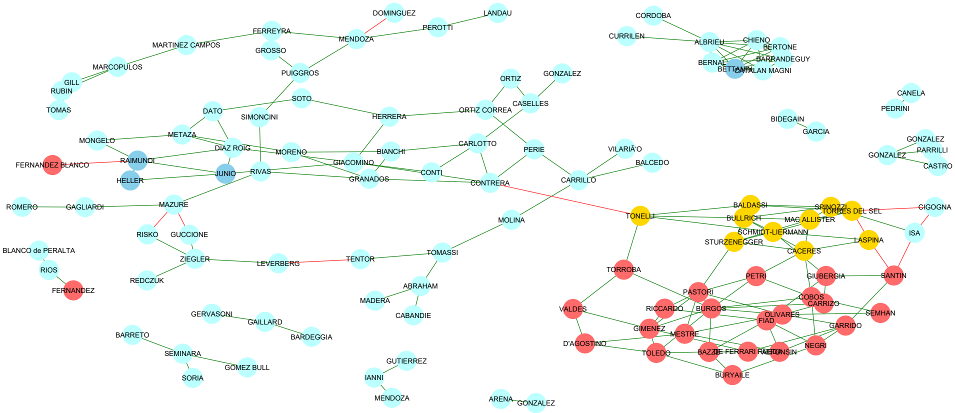
$$\beta_j^{(k)} = \Theta_{kj} \text{ y entonces } \beta_j^{(k)} \neq 0 \iff \Theta_{kj} \neq 0.$$

Por cada parámetro 2 estimaciones $\Longrightarrow$ usamos reglas **AND** u **OR**

Estimando el grafo para `palabras` ($n = 1000, p = 100$)



Estimando el grafo para `diputados` ($n = 183, p = 133$)



Estimando el grafo para `diputados`

El diputado radical por Corrientes Rodolfo Fernández dio quórum al oficialismo y renunció a su bloque

El diputado nacional Rodolfo Fernández (UCR-Corrientes) **desató la polémica en el radicalismo de la Cámara baja al prestar quórum al oficialismo** en el debate por las leyes económicas y difundir luego su renuncia al bloque radical.

Foto: Fernando Sturla / Télam

UBICACIÓN

ÚLTIMOS VIDEOS

> VER MÁS

La Policía británica ide...

Conclusiones y trabajo a futuro

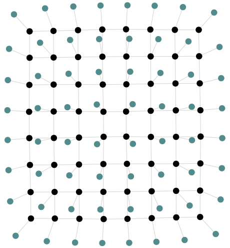
Conclusiones:

- Efecto selector de la penalización con norma 1
- *Stability Selection* como una forma simple de reducir la dependencia de λ
- Marcada separación en grupos de variables

Trabajo a futuro:

Estudiar algunas variantes de estos modelos

- continuos no normales
- modelos mixtos
- con datos faltantes
- con variables ocultas



¡Gracias!