El uso de modelos gráficos probabilísticos en predicción. Una aplicación con datos sobre la secuenciación genómica para detectar cáncer de cerebro.

Leonardo Daniel de la Cruz Cuaxiloa
Asesor: Dr. Gonzalo Pérez de la Cruz

Licenciatura en Actuaría
Facultad de Ciencias, UNAM

Junio 2025

- Los modelos gráficos gaussianos no dirigidos permiten representar dependencias condicionales entre variables cuantitativas.
- El algoritmo *Graphical Lasso* es sumamente eficiente para modelos de alta dimensionalidad.
- Existen diversos estudios sobre la vía de las kinureninas,en los cuales se ha estudiado si existe una asociación con la progresión del cáncer de cerebro [Cervenka et al., 2017; Pérez de la Cruz et al., 2023], o la manera en que el sistema inmunológico responde a la enfermedad [Vázquez Cervantes et al., 2022]
- Se propone usar datos genómicos de las principales enzimas de la vía de las kinureninas desde un enfoque de predicción.

- Los modelos gráficos gaussianos no dirigidos permiten representar dependencias condicionales entre variables cuantitativas.
- El algoritmo *Graphical Lasso* es sumamente eficiente para modelos de alta dimensionalidad.
- Existen diversos estudios sobre la vía de las kinureninas,en los cuales se ha estudiado si existe una asociación con la progresión del cáncer de cerebro [Cervenka et al., 2017; Pérez de la Cruz et al., 2023], o la manera en que el sistema inmunológico responde a la enfermedad [Vázquez Cervantes et al., 2022]
- Se propone usar datos genómicos de las principales enzimas de la vía de las kinureninas desde un enfoque de predicción.

- Los modelos gráficos gaussianos no dirigidos permiten representar dependencias condicionales entre variables cuantitativas.
- El algoritmo *Graphical Lasso* es sumamente eficiente para modelos de alta dimensionalidad.
- Existen diversos estudios sobre la vía de las kinureninas,en los cuales se ha estudiado si existe una asociación con la progresión del cáncer de cerebro [Cervenka et al., 2017; Pérez de la Cruz et al., 2023], o la manera en que el sistema inmunológico responde a la enfermedad [Vázquez Cervantes et al., 2022]
- Se propone usar datos genómicos de las principales enzimas de la vía de las kinureninas desde un enfoque de predicción.

- Los modelos gráficos gaussianos no dirigidos permiten representar dependencias condicionales entre variables cuantitativas.
- El algoritmo *Graphical Lasso* es sumamente eficiente para modelos de alta dimensionalidad.
- Existen diversos estudios sobre la vía de las kinureninas,en los cuales se ha estudiado si existe una asociación con la progresión del cáncer de cerebro [Cervenka et al., 2017; Pérez de la Cruz et al., 2023], o la manera en que el sistema inmunológico responde a la enfermedad [Vázquez Cervantes et al., 2022]
- Se propone usar datos genómicos de las principales enzimas de la vía de las kinureninas desde un enfoque de predicción.

Modelos Gráficos Probabilísticos No Dirigidos

- Un **modelo gráfico probabilístico no dirigido** es un modelo probabilístico multivariada basado en la estructura de una gráfica no dirigida $\mathrm{G} = \{\mathrm{V}, \mathrm{E}\}$. En estos modelos cada vértice v_i es asociado con una variable aleatoria x_{v_i}.
- En particular, diremos que el modelo cumple con la propiedad de **Markov por pares** si:

$$\{v_l, v_m\} \notin \mathrm{E} \iff x_{v_l} \perp\!\!\!\perp x_{v_m} | \boldsymbol{x}_{\mathrm{V}\setminus\{v_l, v_m\}}.$$

- En otras palabras, cuando no exista una arista entre dos vértices en una gráfica, las variables aleatorias asociadas a dichos vértices serán condicionalmente independientes, dado el resto de las variables.

Modelos Gráficos Probabilísticos No Dirigidos

- Un **modelo gráfico probabilístico no dirigido** es un modelo probabilístico multivariada basado en la estructura de una gráfica no dirigida $\mathrm{G} = \{\mathrm{V}, \mathrm{E}\}$. En estos modelos cada vértice v_i es asociado con una variable aleatoria x_{v_i}.
- En particular, diremos que el modelo cumple con la propiedad de **Markov por pares** si:

$$\{v_l, v_m\} \notin \mathrm{E} \iff x_{v_l} \perp\!\!\!\perp x_{v_m} | \boldsymbol{x}_{\mathrm{V}\setminus\{v_l, v_m\}}.$$

- En otras palabras, cuando no exista una arista entre dos vértices en una gráfica, las variables aleatorias asociadas a dichos vértices serán condicionalmente independientes, dado el resto de las variables.

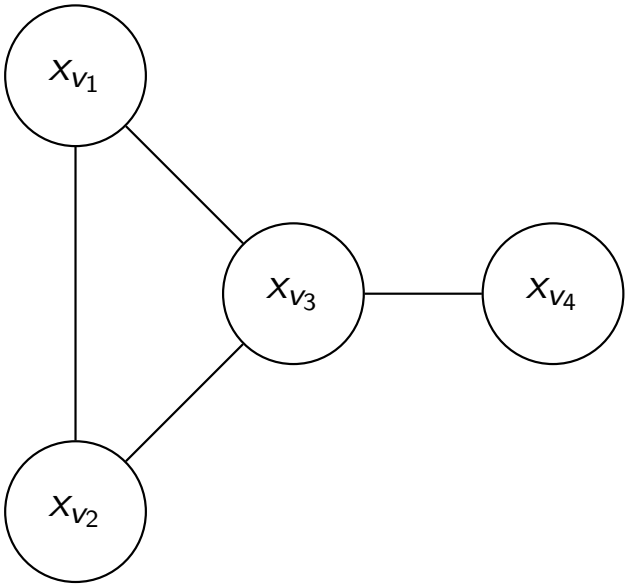
Modelos Gráficos Probabilísticos No Dirigidos

- Un **modelo gráfico probabilístico no dirigido** es un modelo probabilístico multivariada basado en la estructura de una gráfica no dirigida $\mathrm{G} = \{\mathrm{V}, \mathrm{E}\}$. En estos modelos cada vértice v_i es asociado con una variable aleatoria x_{v_i}.
- En particular, diremos que el modelo cumple con la propiedad de **Markov por pares** si:

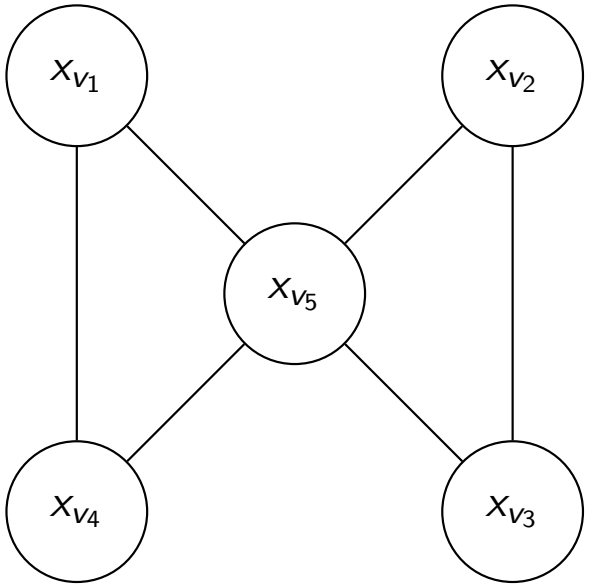
$$\{v_l, v_m\} \notin \mathrm{E} \iff x_{v_l} \perp\!\!\!\perp x_{v_m} | \boldsymbol{x}_{\mathrm{V}\setminus\{v_l, v_m\}}.$$

- En otras palabras, cuando no exista una arista entre dos vértices en una gráfica, las variables aleatorias asociadas a dichos vértices serán condicionalmente independientes, dado el resto de las variables.

Ejemplo



$$G_1 = (V_1, E_1)$$



$$G_2 = (V_2, E_2)$$

Ejemplo

A continuación se presentan dos ejemplos de modelos gráficos probabilísticos no dirigidos.

(a) $\mathrm{G}_1 = (\mathrm{V}_1, \mathrm{E}_1)$:

- $\mathrm{V}_1 = \{v_1, v_2, v_3, v_4\}$
- $\mathrm{E}_1 = \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_2, v_3\}, \{v_3, v_4\}\}$

(b) $\mathrm{G}_2 = (\mathrm{V}_2, \mathrm{E}_2)$:

- $\mathrm{V}_2 = \{v_1, v_2, v_3, v_4, v_5\}$
- $\mathrm{E}_2 = \{\{v_1, v_4\}, \{v_1, v_5\}, \{v_2, v_3\}, \{v_2, v_5\}, \{v_3, v_5\}, \{v_4, v_5\}\}$

Modelos Gráficos Gaussianos No Dirigidos

- Este trabajo se centra en los modelos gráficos gaussianos, un caso particular de los modelos gráficos de Markov. En estos modelos, se asume que las variables son cuantitativas y siguen una distribución normal o gaussiana multivariada $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}.$$

Modelos Gráficos Gaussianos No Dirigidos

- En particular, en estos modelos se puede verificar que

$$\theta_{lm} = 0 \iff \{v_l, v_m\} \notin \mathrm{E} \iff x_{v_l} \perp\!\!\!\perp x_{v_m} | \boldsymbol{x}_{\mathrm{V}\setminus\{v_l,v_m\}},$$

donde

$$\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & \ddots & \vdots \\ \theta_{p1} & \cdots & \theta_{pp} \end{pmatrix}.$$

- $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ es la inversa de la matriz de covarianza, también llamada matriz de precisión o concentración.

Modelos Gráficos Gaussianos No Dirigidos

- En particular, en estos modelos se puede verificar que

$$\theta_{lm} = 0 \iff \{v_l, v_m\} \notin \mathrm{E} \iff x_{v_l} \perp\!\!\!\perp x_{v_m} | \boldsymbol{x}_{\mathrm{V}\setminus\{v_l,v_m\}},$$

donde

$$\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & \ddots & \vdots \\ \theta_{p1} & \cdots & \theta_{pp} \end{pmatrix}.$$

- $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ es la inversa de la matriz de covarianza, también llamada matriz de precisión o concentración.

Clasificación Supervisada

- El objetivo es la construcción de una regla $\widehat{f}$ que permita asignar una y solo una clase $\mathrm{C}_j \in \{\mathrm{C}_1, \ldots, \mathrm{C}_k\}$ a una nueva observación, con base en las mediciones observadas de p variables $\boldsymbol{x} = (x_1, \ldots, x_p)$ asociadas a la nueva observación.
- La construcción de la regla se basa en un conjunto de de datos etiquetados:

$$\mathcal{D} = [(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)]^{\mathrm{T}} = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Clasificación Supervisada

- El objetivo es la construcción de una regla $\widehat{f}$ que permita asignar una y solo una clase $\mathrm{C}_j \in \{\mathrm{C}_1, \ldots, \mathrm{C}_k\}$ a una nueva observación, con base en las mediciones observadas de p variables $\boldsymbol{x} = (x_1, \ldots, x_p)$ asociadas a la nueva observación.
- La construcción de la regla se basa en un conjunto de de datos etiquetados:

$$\mathcal{D} = [(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)]^{\mathrm{T}} = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Planteamiento General

- El objetivo es estimar la siguiente probabilidad:

$$\mathrm{P}\left(y = \mathrm{C}_j | \boldsymbol{x}\right), \quad j \in \{1, ..., k\}.$$

- De forma equivalente:

$$\mathrm{P}\left(y = \mathrm{C}_j | \boldsymbol{x}\right) = \frac{\mathrm{P}(y = \mathrm{C}_j, \boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})}$$
$$\propto \mathrm{P}(\boldsymbol{x} | y = \mathrm{C}_j)\mathrm{P}(y = \mathrm{C}_j)$$

- Por lo tanto, la regla final $\widehat{f}$ será la etiqueta C_j para la cual la **probabilidad estimada** sea máxima. Es decir:

$$\widehat{f}(\boldsymbol{x}) = \arg\max_{\mathrm{C}_j} \left\{ \widehat{\mathrm{P}}\left(y = \mathrm{C}_j | \boldsymbol{x}\right) \right\}.$$

Planteamiento General

- El objetivo es estimar la siguiente probabilidad:

$$\mathrm{P}\left(y = \mathrm{C}_j | \boldsymbol{x}\right), \quad j \in \{1, ..., k\}.$$

- De forma equivalente:

$$\begin{aligned}\mathrm{P}\left(y = \mathrm{C}_j | \boldsymbol{x}\right) &= \frac{\mathrm{P}(y = \mathrm{C}_j, \boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \\ &\propto \mathrm{P}(\boldsymbol{x} | y = \mathrm{C}_j)\mathrm{P}(y = \mathrm{C}_j)\end{aligned}$$

- Por lo tanto, la regla final $\widehat{f}$ será la etiqueta C_j para la cual la **probabilidad estimada** sea máxima. Es decir:

$$\widehat{f}(\boldsymbol{x}) = \arg\max_{\mathrm{C}_j} \left\{ \widehat{\mathrm{P}}\left(y = \mathrm{C}_j | \boldsymbol{x}\right) \right\}.$$

Planteamiento General

- El objetivo es estimar la siguiente probabilidad:

$$\mathrm{P}\left(y = \mathrm{C}_j | \boldsymbol{x}\right), \quad j \in \{1, ..., k\}.$$

- De forma equivalente:

$$\begin{aligned}\mathrm{P}\left(y = \mathrm{C}_j | \boldsymbol{x}\right) &= \frac{\mathrm{P}(y = \mathrm{C}_j, \boldsymbol{x})}{\mathrm{P}(\boldsymbol{x})} \\ &\propto \mathrm{P}(\boldsymbol{x} | y = \mathrm{C}_j)\mathrm{P}(y = \mathrm{C}_j)\end{aligned}$$

- Por lo tanto, la regla final $\widehat{f}$ será la etiqueta C_j para la cual la **probabilidad estimada** sea máxima. Es decir:

$$\widehat{f}(\boldsymbol{x}) = \arg\max_{\mathrm{C}_j} \left\{ \widehat{\mathrm{P}}\left(y = \mathrm{C}_j | \boldsymbol{x}\right) \right\}.$$

Análisis de Discriminante

- Lineal, se asume el supuesto distribucional:

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}\right), \quad j \in \{1, ..., k\}.$$

- Cuadrático, se asume el supuesto distribucional:

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}_{\mathrm{C}_j}\right), \quad j \in \{1, \ldots, k\}.$$

- En ambos casos los parámetros utilizados con los estimadores máximo verosímiles.

Análisis de Discriminante

- Lineal, se asume el supuesto distribucional:

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}\right), \quad j \in \{1, ..., k\}.$$

- Cuadrático, se asume el supuesto distribucional:

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}_{\mathrm{C}_j}\right), \quad j \in \{1, \ldots, k\}.$$

- En ambos casos los parámetros utilizados con los estimadores máximo verosímiles.

Análisis de Discriminante

- Lineal, se asume el supuesto distribucional:

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}\right), \quad j \in \{1, ..., k\}.$$

- Cuadrático, se asume el supuesto distribucional:

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}_{\mathrm{C}_j}\right), \quad j \in \{1, \ldots, k\}.$$

- En ambos casos los parámetros utilizados con los estimadores máximo verosímiles.

Modelo UGGM-QDA

- Asumiremos el supuesto distribucional

$$\boldsymbol{x}|\left\{y=\mathrm{C}_j\right\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}_{\mathrm{C}_j}\right), \quad j \in \{1, \ldots, k\}.$$

- Adicionalmente, asumiremos una estructura de gráfica no dirigida asociada.
- Utilizaremos el algoritmo *glasso* [Hastie et al., 2015] para determinar la estructura de la red, así como para la estimación de los parámetros.
- Este modelo ha sido propuesto en [Chen, 2022].

Modelo UGGM-QDA

- Asumiremos el supuesto distribucional

$$\boldsymbol{x}|\left\{y=\mathrm{C}_j\right\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}_{\mathrm{C}_j}\right), \quad j \in \{1, \ldots, k\}.$$

- Adicionalmente, asumiremos una estructura de gráfica no dirigida asociada.
- Utilizaremos el algoritmo *glasso* [Hastie et al., 2015] para determinar la estructura de la red, así como para la estimación de los parámetros.
- Este modelo ha sido propuesto en [Chen, 2022].

Modelo UGGM-QDA

- Asumiremos el supuesto distribucional

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}_{\mathrm{C}_j}\right), \quad j \in \{1, \ldots, k\}.$$

- Adicionalmente, asumiremos una estructura de gráfica no dirigida asociada.
- Utilizaremos el algoritmo *glasso* [Hastie et al., 2015] para determinar la estructura de la red, así como para la estimación de los parámetros.
- Este modelo ha sido propuesto en [Chen, 2022].

Modelo UGGM-QDA

- Asumiremos el supuesto distribucional

$$\boldsymbol{x}|\{y = \mathrm{C}_j\} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\mathrm{C}_j}, \boldsymbol{\Sigma}_{\mathrm{C}_j}\right), \quad j \in \{1, \ldots, k\}.$$

- Adicionalmente, asumiremos una estructura de gráfica no dirigida asociada.
- Utilizaremos el algoritmo *glasso* [Hastie et al., 2015] para determinar la estructura de la red, así como para la estimación de los parámetros.
- Este modelo ha sido propuesto en [Chen, 2022].

glasso

- De esta forma, dado un valor de $\lambda \geq 0$,se plantea el siguiente problema de optimización:

$$\widehat{\boldsymbol{\Theta}}_{\text{glasso}} = \arg\max_{\boldsymbol{\Theta}} \left\{ \ell_{\boldsymbol{x}_{\mathrm{V}}}(\boldsymbol{\Theta}) - \lambda \sum_{l \neq m} |\theta_{lm}| \right\},$$

donde $\ell_{\boldsymbol{x}_{\mathrm{V}}}$ corresponde a la función de log verosimilitud asociada, es decir:

$$\begin{aligned} \ell_{\boldsymbol{x}_{\mathrm{V}}}(\boldsymbol{\Theta}) &= \frac{1}{n} \sum_{i=1}^{n} \ln\left[\mathrm{P}(\boldsymbol{x}_i)\right] \\ &= \ln(|\boldsymbol{\Theta}|) - \operatorname{tr}(\mathrm{S}\boldsymbol{\Theta}). \end{aligned}$$

Adicional al modelo propuesto UGGM-QDA, en la sección de aplicación se consideraron los siguientes modelos de clasificación supervisada.

- Análisis de discriminante lineal.
- Análisis de discriminante cuadrático.
- Regresión logística.
- Regresión logística con penalización *ElasticNet*.
- Máquina de soporte vectorial.
- Bosques aleatorios.

QDA v.s UGGM-QDA

- Para evaluar el desempeño del modelo UGGM-QDA en comparación con el modelo QDA en términos de poder predictivo, se llevaron a cabo simulaciones considerando distintos tamaños de muestra y número de variables.
- En total, se generaron 16 conjuntos de datos, cada uno compuesto por tres clases (Clase 1, Clase 2 y Clase 3).
- Cada simulación consideró tamaños de muestra de 50, 100, 500 y 1000 observaciones por clase, combinados con distintos números de variables: 20, 25, 30 y 35.
- Para cada uno de estos conjuntos, se generó un conjunto de evaluación $\mathcal{D}_{\text{Test}}$, compuesto por 10000 observaciones por clase.

QDA v.s UGGM-QDA

- Para evaluar el desempeño del modelo UGGM-QDA en comparación con el modelo QDA en términos de poder predictivo, se llevaron a cabo simulaciones considerando distintos tamaños de muestra y número de variables.
- En total, se generaron 16 conjuntos de datos, cada uno compuesto por tres clases (Clase 1, Clase 2 y Clase 3).
- Cada simulación consideró tamaños de muestra de 50, 100, 500 y 1000 observaciones por clase, combinados con distintos números de variables: 20, 25, 30 y 35.
- Para cada uno de estos conjuntos, se generó un conjunto de evaluación $\mathcal{D}_{\text{Test}}$, compuesto por 10000 observaciones por clase.

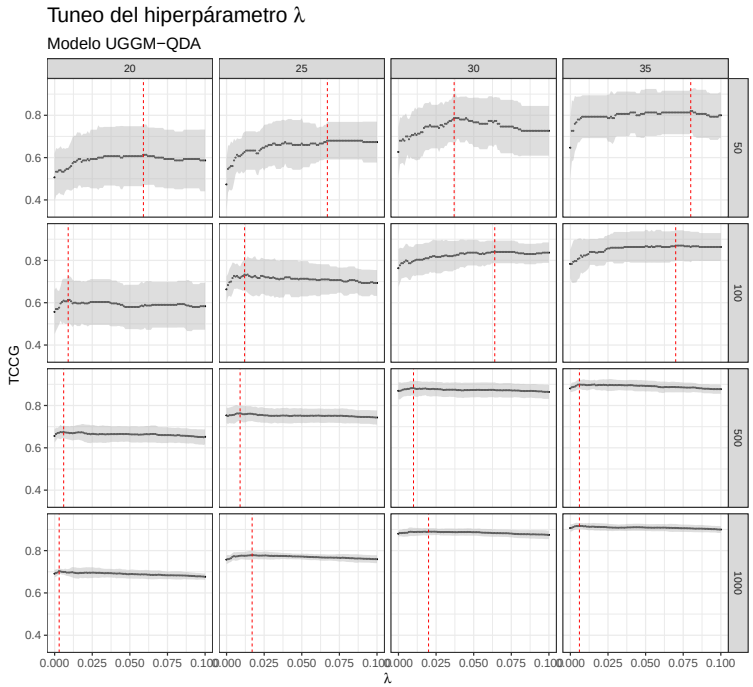
QDA v.s UGGM-QDA

- Para evaluar el desempeño del modelo UGGM-QDA en comparación con el modelo QDA en términos de poder predictivo, se llevaron a cabo simulaciones considerando distintos tamaños de muestra y número de variables.
- En total, se generaron 16 conjuntos de datos, cada uno compuesto por tres clases (Clase 1, Clase 2 y Clase 3).
- Cada simulación consideró tamaños de muestra de 50, 100, 500 y 1000 observaciones por clase, combinados con distintos números de variables: 20, 25, 30 y 35.
- Para cada uno de estos conjuntos, se generó un conjunto de evaluación $\mathcal{D}_{\text{Test}}$, compuesto por 10000 observaciones por clase.

QDA v.s UGGM-QDA

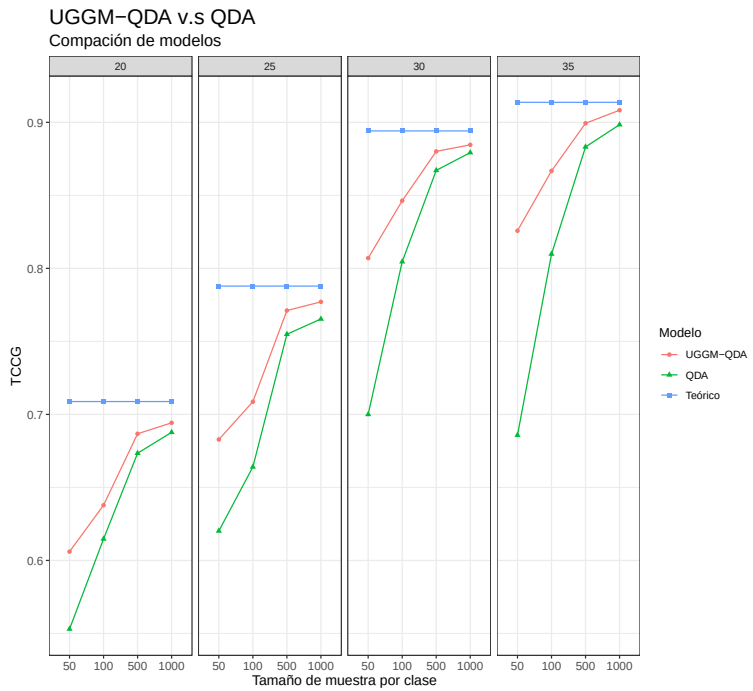
- Para evaluar el desempeño del modelo UGGM-QDA en comparación con el modelo QDA en términos de poder predictivo, se llevaron a cabo simulaciones considerando distintos tamaños de muestra y número de variables.
- En total, se generaron 16 conjuntos de datos, cada uno compuesto por tres clases (Clase 1, Clase 2 y Clase 3).
- Cada simulación consideró tamaños de muestra de 50, 100, 500 y 1000 observaciones por clase, combinados con distintos números de variables: 20, 25, 30 y 35.
- Para cada uno de estos conjuntos, se generó un conjunto de evaluación $\mathcal{D}_{\text{Test}}$, compuesto por 10000 observaciones por clase.

Tuneo del hiperparámetro λ



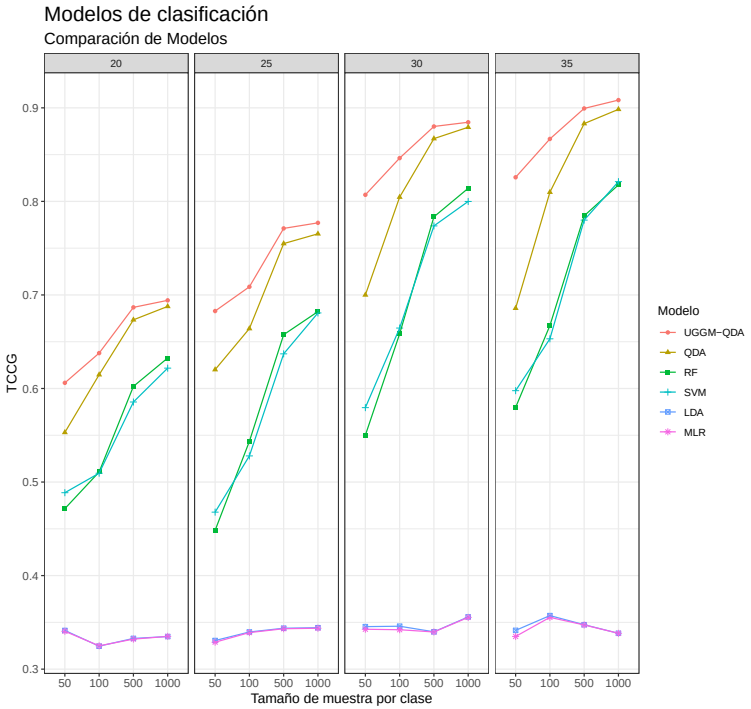
Haz clic en la figura para abrirla en una nueva ventana.

Comparativa de los modelos QDA, UGGM-QDA y valores teóricos



Haz clic en la figura para abrirla en una nueva ventana.

Comparativa con todos los modelos



Haz clic en la figura para abrirla en una nueva ventana.

La vía de las Kinureninas

- El triptófano es un aminoácido esencial que obtenemos a través de la alimentación y cumple funciones clave en nuestro organismo.
- La mayor parte del triptófano se procesa a través de una serie de reacciones químicas que conforman la llamada vía de las kinureninas.
- La vía de las kinureninas está controlada por dos enzimas principales: la indoleamina 2,3-dioxigenasa (IDO) y la triptófano 2,3-dioxigenasa (TDO)
- Diversas investigaciones han encontrado que la vía de las kinureninas está estrechamente relacionada con el desarrollo y la progresión del cáncer cerebral.

La vía de las Kinureninas

- El triptófano es un aminoácido esencial que obtenemos a través de la alimentación y cumple funciones clave en nuestro organismo.
- La mayor parte del triptófano se procesa a través de una serie de reacciones químicas que conforman la llamada vía de las kinureninas.
- La vía de las kinureninas está controlada por dos enzimas principales: la indoleamina 2,3-dioxigenasa (IDO) y la triptófano 2,3-dioxigenasa (TDO)
- Diversas investigaciones han encontrado que la vía de las kinureninas está estrechamente relacionada con el desarrollo y la progresión del cáncer cerebral.

La vía de las Kinureninas

- El triptófano es un aminoácido esencial que obtenemos a través de la alimentación y cumple funciones clave en nuestro organismo.
- La mayor parte del triptófano se procesa a través de una serie de reacciones químicas que conforman la llamada vía de las kinureninas.
- La vía de las kinureninas está controlada por dos enzimas principales: la indoleamina 2,3-dioxigenasa (IDO) y la triptófano 2,3-dioxigenasa (TDO)
- Diversas investigaciones han encontrado que la vía de las kinureninas está estrechamente relacionada con el desarrollo y la progresión del cáncer cerebral.

La vía de las Kinureninas

- El triptófano es un aminoácido esencial que obtenemos a través de la alimentación y cumple funciones clave en nuestro organismo.
- La mayor parte del triptófano se procesa a través de una serie de reacciones químicas que conforman la llamada vía de las kinureninas.
- La vía de las kinureninas está controlada por dos enzimas principales: la indoleamina 2,3-dioxigenasa (IDO) y la triptófano 2,3-dioxigenasa (TDO)
- Diversas investigaciones han encontrado que la vía de las kinureninas está estrechamente relacionada con el desarrollo y la progresión del cáncer cerebral.

Enfoque en Predicción

- Fundamento en investigaciones previas sobre la expresión de las enzimas de la vía de las kinureninas en diferentes tipos de gliomas y su relación con características tumorales clave [Pérez de la Cruz et al., 2023].
- Los datos utilizados en este análisis corresponden a valores de expresión genética normalizados (*RSEM norm count*) de muestras provenientes de los proyectos TCGA, TARGET y GTEx, accesibles a través de la plataforma Xena [Goldman et al., 2020]. Estas bases de datos incluyen información de pacientes con distintos tipos de cáncer y tejidos sanos, recopilada mayormente en instituciones de Estados Unidos.

Enfoque en Predicción

- Fundamento en investigaciones previas sobre la expresión de las enzimas de la vía de las kinureninas en diferentes tipos de gliomas y su relación con características tumorales clave [Pérez de la Cruz et al., 2023].
- Los datos utilizados en este análisis corresponden a valores de expresión genética normalizados (*RSEM norm count*) de muestras provenientes de los proyectos TCGA, TARGET y GTEx, accesibles a través de la plataforma Xena [Goldman et al., 2020]. Estas bases de datos incluyen información de pacientes con distintos tipos de cáncer y tejidos sanos, recopilada mayormente en instituciones de Estados Unidos.

Principales Enzimas involucradas en la vía de las kinureninas

Nombre	**Abreviación**	**Variable**
Tryptophan dioxygenase	TDO	x_1
Indoleamine dioxygenase 1	IDO1	x_2
Indoleamine dioxygenase 2	IDO2	x_3
Arylformamidase	AFMID	x_4
Glutamic-oxaloacetic transaminase	GOT2	x_5
Aminoadipate aminotransferase	AADAT	x_6
Kynureninase	KYNU	x_7
Kynurenine monooxygenase	KMO	x_8
Quinolinic acid phosphoribosyl transferase	QPRT	x_9
3-HANA dioxygenase	HAAO	x_{10}
Aminocarboxymuconate semialdehyde decarboxylase	ACMSD	x_{11}

Grupos

- De un total de 19,131 muestras disponibles, se seleccionaron aquellas correspondientes a gliomas de bajo grado (TCGA LGG, $n = 523$), glioblastoma multiforme (TCGA GM, $n = 166$) y muestras de tejido cerebral sano (GTEx brain cortex, $n = 105$; GTEx brain anterior cingulate cortex (Ba24), $n = 83$; y GTEx brain frontal cortex (Ba9), $n = 95$).

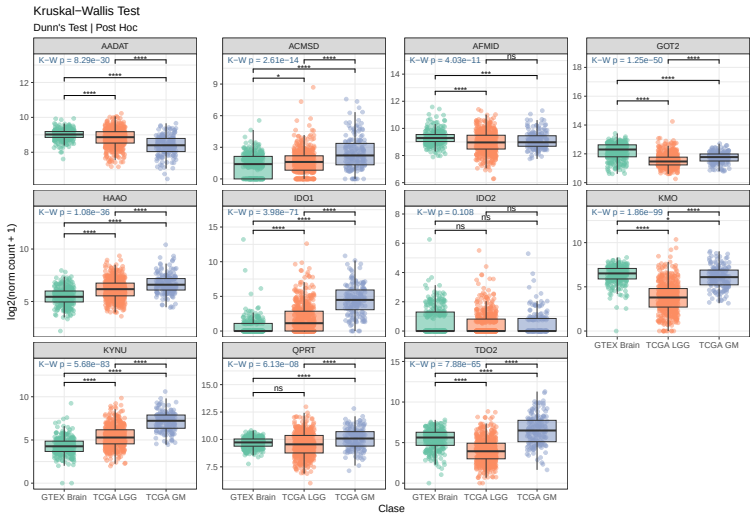
Grupo	**Abreviación**	**Tamaño de muestra**
GTEX Brain	GTEX Brain	283
TCGA Brain Lower Grade Glioma	TCGA LGG	523
TCGA Glioblastoma Multiforme	TCGA GM	166

- Evaluación diferencias estadísticamente significativas entre grupos, se aplicó la prueba de Kruskal-Wallis, seguida de pruebas Dunn para analizar diferencias entre pares de grupos.
- Matriz de diagramas de dispersión junto con las correlaciones de Pearson y la distribución de cada variable, diferenciando entre grupos.
- Análisis de componentes principales para reducir la dimensionalidad de los datos y explorar la variabilidad explicada por combinaciones lineales.

- Evaluación diferencias estadísticamente significativas entre grupos, se aplicó la prueba de Kruskal-Wallis, seguida de pruebas Dunn para analizar diferencias entre pares de grupos.
- Matriz de diagramas de dispersión junto con las correlaciones de Pearson y la distribución de cada variable, diferenciando entre grupos.
- Análisis de componentes principales para reducir la dimensionalidad de los datos y explorar la variabilidad explicada por combinaciones lineales.

- Evaluación diferencias estadísticamente significativas entre grupos, se aplicó la prueba de Kruskal-Wallis, seguida de pruebas Dunn para analizar diferencias entre pares de grupos.
- Matriz de diagramas de dispersión junto con las correlaciones de Pearson y la distribución de cada variable, diferenciando entre grupos.
- Análisis de componentes principales para reducir la dimensionalidad de los datos y explorar la variabilidad explicada por combinaciones lineales.

Pruebas no Paramétricas



Haz clic en la figura para abrirla en una nueva ventana.

Esquema de Estimación

- Para calcular el poder predictivo de los modelos, se empleó un esquema de Repeated u-v Fold Cross Validation, con $u = 100$ y $v = 5$, lo que significa que se realizaron 100 repeticiones de una validación cruzada de 5 pliegues, dando un total de 500 procesos diferentes. En los casos en donde se realiza calibración de hiperparámetros, esto se realiza en cada uno de los 500 procesos.

Modelos

Modelo	Descripción	Hiperparámetros	Malla de valores	Esquema de *tuneo*
MLR 1	Regresión Logística Multinomial con efectos principales.	*cut off*: α	$\alpha = 0.5$	-
MLR 2	Regresión Logística Multinomial con efectos principales e interacciones de segundo orden con penalización *ElasticNet*.	*ridge*: α *lasso*: λ	$\alpha \in \{0, 0.5, 1\}$. 100 valores por default de la función `cv.glmnet()` del paquete `glmnet` para λ.	Esquema de validación cruzada con 3 subconjuntos (*folds*), mediante de la función `cv.glmnet()`.
LDA 1	Análisis de Discriminante Lineal.	-	-	-
LDA 2	Análisis de Discriminante Lineal en su versión de UGGM.	*glasso*: λ	$\lambda = 0.01$.	-
QDA 1	Análisis de Discriminante Cuadrático.	-	-	-
QDA 2	Modelo UGGM-QDA.	*glasso*: λ	$\lambda \in [0, 1]$ con incrementos de 0.01.	Esquema de validación cruzada con 5 subconjuntos (*folds*), mediante la implementación del Algoritmo 6.
SVM 1	Máquina de Soporte Vectorial.	*kernel*: radial *cost*: 1 *gamma*: $\frac{1}{n}$	-	-
SVM 2	Máquina de Soporte Vectorial.	*kernel*: radial *cost*: c *gamma*: γ	$c \in \{1, 11, 21, \ldots, 91\}$ $\gamma \in \{0.1, 0.2, \ldots, 1\}$	Esquema de validación cruzada con 5 subconjuntos (*folds*), mediante la función `svm.tune()`.
RF 1	Bosques Aleatorios.	*num.trees*: 500 *mtry*: $\lfloor \sqrt{11} \rfloor$ *node.size*: 1	-	-
RF 2	Bosques Aleatorios.	*num.trees*: *nt* *mtry*: *m* *node.size*: *ns*	$nt \in \{50, 100, 500\}$ $m \in \{1, 2, \ldots, 11\}$ $ns \in \{1, 10, 15\}$	Evaluación exhaustiva de cada combinación de hiperparámetros.

Haz clic en la figura para abrirla en una nueva ventana.

Resultados Train

	$\mu(\mathcal{D}_{\text{Train}})$			
	TCCG	**TCCC$_1$**	**TCCC$_2$**	**TCC C$_3$**
Multinomial Logistic Regression (MLR 1)	86.1	89.2	90.3	67.3
Multinomial Logistic Regression (MLR 2)	88.0	91.0	91.5	71.8
Linear Discriminant Analysis (LDA 1)	84.6	88.6	86.4	71.9
Linear Discriminant Analysis (LDA 2)	83.3	82.4	91.5	58.8
Quadratic Discriminant Analysis (QDA 1)	87.8	91.6	88.6	78.9
Quadratic Discriminant Analysis (QDA 2)	81.7	89.8	74.5	90.4
Support Vector Machine (SVM 1)	91.4	93.7	94.7	76.6
Support Vector Machine (SVM 2)	90.6	94.0	93.3	76.3
Random Forest (RF 1)	100	100	100	100
Random Forest (RF 2)	97.6	98.5	97.0	98.2

Haz clic en la figura para abrirla en una nueva ventana.

Resultados Test

	$\mu(\mathcal{D}_{\text{Test}})$			
	TCC G	**TCC C1**	**TCC C2**	**TCC C3**
Multinomial Logistic Regression (MLR 1)	85.0	88.4	89.3	65.5
Multinomial Logistic Regression (MLR 2)	85.5	89.0	89.4	67.1
Linear Discriminant Analysis (LDA 1)	83.7	87.8	85.6	70.7
Linear Discriminant Analysis (LDA 2)	82.6	81.8	91.1	57.2
Quadratic Discriminant Analysis (QDA 1)	85.5	89.8	87.4	72.0
Quadratic Discriminant Analysis (QDA 2)	80.3	88.6	73.4	87.8
Support Vector Machine (SVM 1)	86.3	89.1	91.0	66.7
Support Vector Machine (SVM 2)	86.3	90.6	89.9	67.3
Random Forest (RF 1)	85.2	87.0	90.2	66.5
Random Forest (RF 2)	85.0	90.8	84.3	77.7

Haz clic en la figura para abrirla en una nueva ventana.

Modelo UGGM-QDA

- A partir de los resultados obtenidos tanto en los datos simulados como en el caso práctico, se concluye que el modelo UGGM-QDA puede ser útil en problemas de clasificación. Sin embargo, existen aspectos que podrían mejorar significativamente su desempeño.
  - Uso de valores λ diferenciados por clase.
  - Evaluación de diferentes métricas para la calibración de λ.
  - Incorporación de la calibración de las probabilidades a priori.
  - Definición de un modelo UGGM-LDA.

Vía de las kinureninas

- Los resultados obtenidos mostraron que los modelos de clasificación lograron Tasas de Clasificación Correcta Global superiores al 80 %, lo que sugiere que estas variables tienen el potencial de ser utilizadas en el diagnóstico de las siguientes condiciones:
  - Personas sanas.
  - Personas con gliomas de bajo grado.
  - Personas con glioblastoma multiforme, un tipo de tumor cerebral altamente agresivo.

Cervenka, I., Agudelo, L. Z., and Ruas, J. L. (2017). Kynurenines: Tryptophan's metabolites in exercise, inflammation, and mental health. *Science*, 357(6349).

Chen, L.-P. (2022). Netda: An r package for network-based discriminant analysis subject to multilabel classes. *Journal of Probability and Statistics*, 2022(1):1041752.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., and Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the xena platform. *Nature Biotechnology*, 38(6):675–678.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Pérez de la Cruz, G., Pérez de la Cruz, V., Navarro Cossio, J., Vázquez Cervantes, G. I., Salazar, A., Orozco Morales, M., and Pineda, B. (2023). Kynureninase promotes immunosuppression and predicts survival in glioma patients: In silico data analyses of the chinese glioma genome atlas (cgga) and of the cancer genome atlas (tcga). *Pharmaceuticals (Basel)*, 16(3).

Vázquez Cervantes, G. I., Navarro Cossio, J. Á., Pérez de la Cruz, G., Salazar, A., Pérez de la Cruz, V., and Pineda, B. (2022). Bioinformatic analysis of kynurenine pathway enzymes and their relationship with glioma hallmarks. *Metabolites*, 12(11).