El modelo general presentado en este trabajo, aunque pesado en notación y relativamente abstracto, resultó ser muy efectivo al llevarlo a la práctica. A lo largo de este capítulo, se hará una exploración intuitiva y visual de sus capacidades. Se remarca que todas las gráficas presentadas, se generaron con el mismo paquete `bpwpm` que realiza la estimación de los parámetros $\boldsymbol{\beta}$; pues, los mismos objetos que las funciones devuelven, pueden ser utilizados para hacer gráficas que evalúan el modelo y reflejen la intuición subyacente.

Para mostrar los resultados y las capacidades del modelo se presentan seis ejemplos breves. Los primeros cinco, corresponden a bases de datos simuladas en dos dimensiones, es decir, se tienen dos covariables ($\mathbf{x}_i \in \mathbb{R}^2 \ \forall i$), con diferentes patrones para las respuestas $\mathbf{y}$ con fronteras tanto lineales como no lineales. El objetivo, es poder visualizar lo flexibles que son las fronteras de clasificación: la parte no lineal del modelo. Asimismo, al trabajar con bases de datos donde $\mathcal{X}^2 \subseteq \mathbb{R}^2$, se puede visualizar la función $\eta(\mathbf{x})$ en tres dimensiones. El último ejemplo corresponde a una base de datos médicos reales donde cada observación representa un tumor que puede o no ser cancerígeno, las covariables representan ciertas características sobre éste. Al aumentar la dimensionalidad, el modelo ya no es representable visualmente pero se siguen obteniendo buenos resultados.

A todos los modelos presentados a lo largo de este capítulo se les realizó un análisis de convergencia revisando las medias ergódicas de las cadenas. Sin embargo, únicamente se estudia a detalle para el ejemplo 2 de forma que no se saturara más la presentación.

## 0.1. Evaluación del modelo

Antes de poder presentar los ejemplos, se definen las dos métricas que se usarán para probar la efectividad de los modelos.[1] Al trabajar con modelos de clasificación binaria, una forma intuitiva de medir su efectividad es a través de un simple conteo de *errores y aciertos*. Este conteo, se presenta en una *matriz de confusión* que desglosa la clasificación en sus respectivas categorías binarias. Asimismo, se presenta la función *log-loss* ($ll$) que no solo pondera la clasificación sino la *precisión* de ésta, medida a través de las probabilidades ajustadas $\hat{p}_i$ que se le asigna a cada observación $i$.

Las matrices de confusión (tabla 1), son un método descriptivo con base en las tablas de contingencia que calcula la frecuencia de los aciertos y errores separando por grupos. Donde $\hat{y}$ es la variable predicha de la respuesta $y$ y # el símbolo que denota *número de*. Asimismo, se define la precisión del modelo como:

$$\text{precisión} = \frac{\text{Número de clasificaciones correctas}}{\text{Número total de observaciones}}$$

| | $\hat{y} = 0$ | $\hat{y} = 1$ | |
|---|---|---|---|
| $y = 0$ | #0's ✓ | #0's clasificados como 1 | # de observaciones 0 |
| $y = 1$ | #1's clasificados como 0 | #1's ✓ | # de observaciones 1 |
| | # de 0's estimados | # de 1's estimados | Total de obs. $= n$ |

Tabla 1: La matriz de confusión

1. Técnicamente, es diferente la selección de la evaluación de los modelo. La primera se refiere a *qué modelo se debe de escoger* entre varias alternativas, controlando la interacción entre sesgo, varianza y complejidad, es decir, previniendo el sobre-ajuste. Mientras que la evaluación, hace referencia a *qué tan bien generaliza* el modelo ante la presencia de nuevos datos, capítulo 7 de **hastie2008elements**. Dados los objetivos del trabajo, se hace énfasis en la evaluación.

Sin embargo, la matriz de confusión resulta deficiente para comparar modelos completamente diferentes que resulten en exactamente la misma clasificación, por ello, se define una métrica más formal. Sea $\mathbf{y} = (y_1, \ldots, y_n)^t$ el vector de respuestas observadas y $\hat{\mathbf{p}} = (\hat{p_1}, \ldots, \hat{p_n})^t$ el vector de probabilidades ajustadas, donde:

$$\hat{p}_i = \hat{P}_{\text{modelo}}(y_i = 1 \,|\, \mathbf{x}_i)$$

es la probabilidad estimada por el modelo de que la observación $y_i$ sea igual a uno. Con lo anterior, se define un vector de respuestas ajustadas $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^t$, haciendo la predicción en el corte $\hat{y}_i = 1 \iff \hat{p}_i > 0.5$ .[2]

**Definición 0.1.** La función *log-loss* $ll : \{0,1\}^n \times [0,1]^n \to \mathbb{R}^+$ tiene la forma:

$$ll(\mathbf{y}, \hat{\mathbf{p}}) = -\sum_{i=1}^{n} [y_i \ln(\hat{p_i}) + (1 - y_i) \ln(1 - \hat{p_i})]. \tag{1}$$

La función $ll$ logra medir la *bondad de ajuste* del modelo, tomando en cuenta tanto la clasificación en si (a través de los valores binarios de $y_i$) como la precisión de esta (a través de las probabilidades ajustadas $\hat{p}_i$). Idealmente $ll = 0$ si se da una clasificación perfecta y conforme tome valores más positivos, el modelo realiza un peor trabajo. Esto se debe a que la función es convexa y se penaliza cuando las probabilidades ajustadas están muy lejos de la real. Asimismo, si la clasificación fue incorrecta pero la probabilidad fue cercana a 0.5 no se penaliza tanto. En la práctica y bajo un enfoque frecuentista, la función $ll$ puede ser vista como una función objetivo por optimizar y más recientemente se ha utilizado para para entrenar y comparar modelos de clasificación como lo son las redes neuronales (**nielsen2015neural**).[3]

2. Este corte, es resultado de la simetría en cero de la función de acumulación normal estándar $\Phi$, derivado de la ecuación (**??**).

3. Bajo un contexto de selección de modelos, la función $ll$ se relaciona con el criterio de Akaike y la

## 0.2. Ejemplo 1 - las capacidades del modelo `bpwpm`

El primer ejemplo que se analiza, busca ejemplificar los componentes del modelo en general y sus capacidades. Para ello, se simularon un total de $n = 350$ observaciones separadas en dos grupos, cada uno con tamaños $n_0 = 200$ y $n_1 = 150$ respectivamente ($n = n_0 + n_1$). Los datos se simularon de dos distribuciones normales bivariadas:

$$\underset{i=1,\ldots,200}{\text{Grupo 0:}} \quad \mathbf{x}_i \sim \mathcal{N}_2 \left( \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \middle| \, \boldsymbol{\mu}_0 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \, \Sigma_0 = \begin{pmatrix} 0.25 & 0.35 \\ 0.35 & 1 \end{pmatrix} \right)$$

$$\underset{i=201,\ldots,350}{\text{Grupo 1:}} \quad \mathbf{x}_i \sim \mathcal{N}_2 \left( \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \middle| \, \boldsymbol{\mu}_1 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \, \Sigma_1 = \begin{pmatrix} 1 & -0.24 \\ -0.24 & 0.64 \end{pmatrix} \right)$$

Las medias $\boldsymbol{\mu}_j \; j = \{0, 1\}$ se toman relativamente alejadas y las covarianzas corresponden a las correlaciones $\rho_0 = 0.7$ y $\rho_1 = 0.3$ respectivamente. Estos parámetros de simulación se escogen a través de un proceso empírico resultando en una estructura simple donde los grupos están claramente separados y hay poco traslape. Asimismo, el espacio de covariables queda definido: $\mathcal{X}^2 \approx [0.3, 7.5] \times [-0.5, 5.9]$. Se codifica el grupo 0 ($y = 0$) de color rojo y el grupo 1 ($y = 1$) de color azul.[4] La base de datos final se presenta en la figura 1.

métrica AIC contra el que se puede ponderar la complejidad del modelo: $\text{AIC}(\lambda) = 2(\lambda - ll)$ penalizando los modelos que sobre-ajustan los datos. Bajo un contexto de estadística bayesiana es usual utilizar la métrica BIC: *bayesian information criterion* la cual es sencilla de calcular para este modelo resultando en $\text{BIC}(\lambda) = \lambda \ln(n) - 2 \times ll$.

4. Se recomienda visualizar la versión digital de este trabajo donde se aprecian con más claridad los colores. Disponible en `https://github.com/PaoloLuciano/Tesis-Latex/`

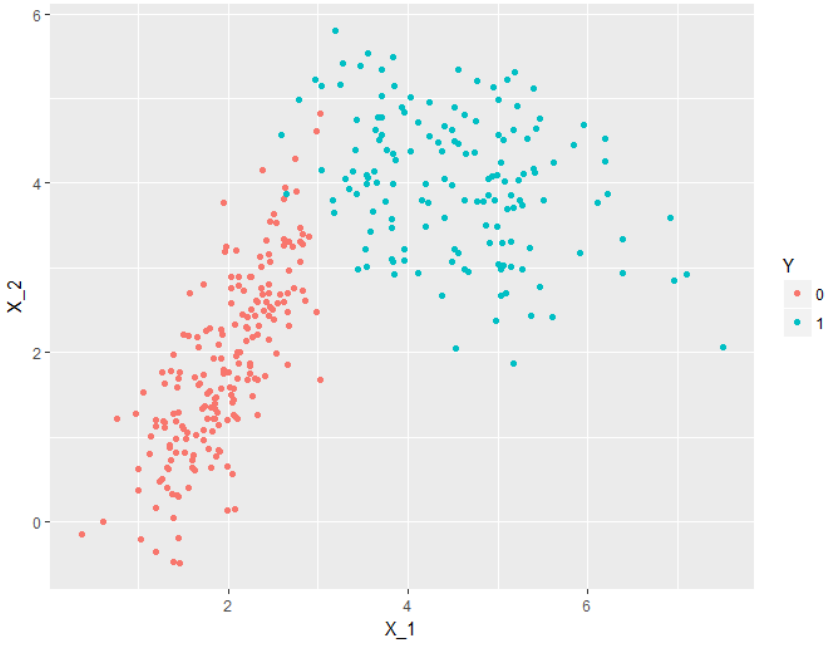



Figura 1: Ejemplo 1 - datos normales bivariados

**Tres realizaciones del modelo**

El objetivo principal de esta simple base de datos es ejemplificar el tipo de fronteras alcanzables por $\eta$, mostrando una clara separación entre los dos grupos sin sobre-ajustar, para ello, se estiman los parámetros para tres realizaciones del modelo. Para la primera se escoge el modelo más sencillo, una frontera lineal con un solo nodo (M = 2, J = 2 y K = 1). La segunda realización, consta de parábolas continuas mas no suaves sobre cuatro nodos (M = 3, J = 5 y K = 1). Finalmente la tercera realización consta de splines cúbicos en 3 nodos (M = 4, J = 3 y K = 3),[5] en la tabla 2 se resume lo anterior.[6]

5. Polinomios por partes cúbicos suaves hasta la segunda derivada.

6. Se recuerda que $M-1$ corresponde al grado de los polinomios, $J-1$ es el número de nodos, $K$ el parámetro que controla la suavidad, $N^* = JM - K(J-1) - 1$ el número de funciones base (por expansión

| Parámetro | Realización 1 | Realización 2 | Realización 3 |
|---|---|---|---|
| $M$ | 2 | 3 | 4 |
| $J$ | 2 | 5 | 3 |
| $K$ | 1 | 1 | 3 |
| $N^*$ | 2 | 10 | 5 |
| $\lambda$ | 5 | 21 | 11 |

Tabla 2: Ejemplo 1 - tres realizaciones del modelo

Para las tres realizaciones se simularon $N_{\text{sim}} = 15{,}000$ valores de $\boldsymbol{\beta}$ y se opta por no usar periodo de *burn-in* ni suavizamiento para las cadenas, es decir: $k^* = 0$ y $k_{\text{thin}} = 0$, esto para hacer a las cadenas más comparables entre sí. La única modificación que se realiza entre las tres realizaciones es que, para la tercera, se estandarizan las covariables.[7] Al usar polinomios de orden mayor, en este caso polinomios de tercer grado, el algoritmo puede caer en problemas numéricos pues $\hat{\eta}$ puede crecer muy rápido fuera de $\mathcal{X}^d$; se expande sobre este tema en el capítulo **??**.

En las figuras 2, 3 y 4 se presentan imágenes que ejemplifican las tres realizaciones del modelo respectivamente. En las imágenes 2a, 3a y 4a se visualizan las diferentes tipos de fronteras que el modelo logra estimar. Con estas fronteras, se nota claramente como es determinante la elección de $M$, $J$ y $K$ en sus formas. El modelo logra estimar tanto fronteras relativamente rígidas (imagen 2a) como fronteras más suaves en las imágenes subsecuentes 3a y 4a. Asimismo, para cada realización, se tiene la representación en 3D de cada función $\hat{\eta}$ que preserva la suavidad (o rugosidad) de sus componentes. Asimismo, rescatando las ideas de los GAM, se puede colapsar cada expansión de polinomios por partes en sus correspondientes

de cada covariable) y $\lambda = 1 + d * N^*$ el número total de parámetros.

7. Se resta la media y se divide entre la desviación estándar muestral de cada covariable.

funciones $\hat{f}_j$ y visualizarla como la transformación no lineal de cada covariable. Por ejemplo, en la imagen 2c se observa que $\hat{f}_1(x_1)$ está compuesta por rectas que se conectan en el nodo, mientras que 4d representa $\hat{f}_2(x_2)$, un polinomio cúbico suave hasta la segunda derivada.

| | $\hat{y}=0$ | $\hat{y}=1$ | |
|---|---|---|---|
| $y=0$ | 198 | 2 | 200 |
| $y=1$ | 2 | 148 | 150 |
| | 200 | 150 | 350 |

(a) Matriz de confusión para todas las realizaciones

| Realización | *ll* |
|---|---|
| 1 | 0.04088 |
| 2 | 0.03464 |
| 3 | 0.03498 |

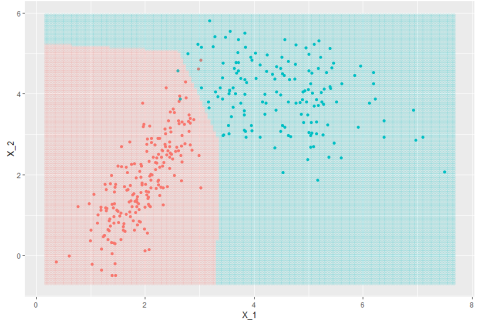
(b) *log-loss*

Tabla 3: Ejemplo 1 - resultados

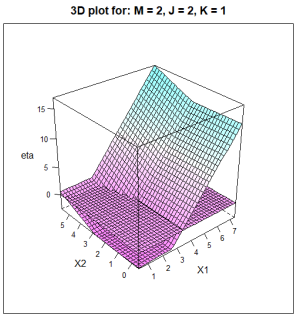
Al estar tratando con una base de datos tan sencilla, no es el enfoque comparar los modelos resultantes entre si pues no se está tomando en cuenta la complejidad a través del número de parámetros. Más bien, se está tratando de ejemplificar las varias fronteras que el modelo logra para la misma clasificación desglosada en la tabla 3a. Al compartir la matriz de confusión, por ende, las realizaciones también comparten una precisión de 98.9 %. De la matriz y las imágenes, se observa que se clasifican de forma incorrecta solo cuatro observaciones. Sería inverosímil tratar de alcanzar una precisión del 100 % pues implicaría sobre-ajustar el modelo. Para estas cuatro observaciones incorrectamente clasificadas, no se tiene la suficiente evidencia como para clasificarlas en su categoría contraria. Sin embargo, los modelos se pueden comparar más a fondo por medio de la métrica *ll* presentada en la tabla 3b. Aunque muy similares, la definición de la métrica indica que la realización dos es la mejor por un pequeño margen pues es la más cercana a cero.

Dado que éste es un ejemplo introductorio la estimación de los parámetros se realizó *dentro de la muestra* (*in-sample*), esto quiere decir que el modelo se entrena con las mismas observaciones contra las que se busca predecir.[8] Cabe mencionar que para esta sencilla base de

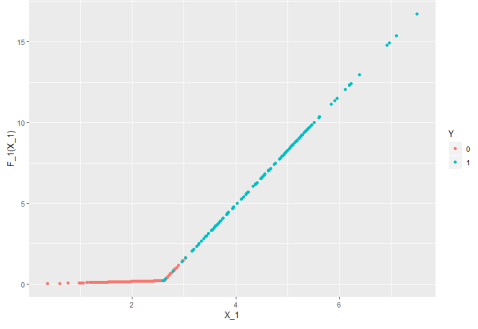
8. El efecto que ésto puede tener es que se sobre-ajuste a la muestra de entrenamiento donde, si se introdujeran nuevos datos el modelo probablemente no haría clasificaciones tan acertadas. En aprendizaje



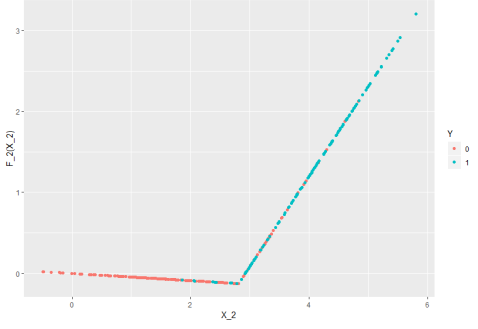

(a) Frontera de predicción




(b) Representación 3D de $\hat{\eta}$

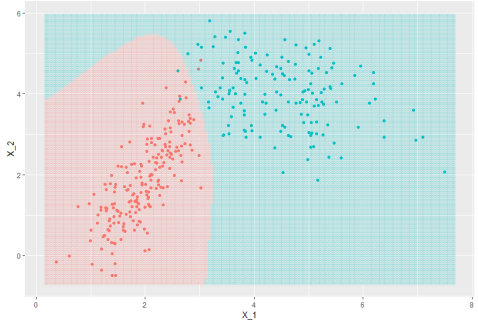



(c) $\hat{f}_1(x_1)$

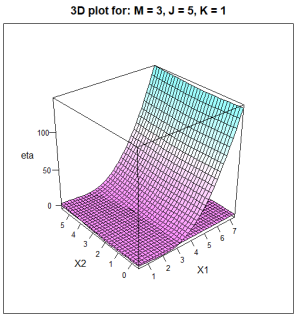



(d) $\hat{f}_2(x_2)$

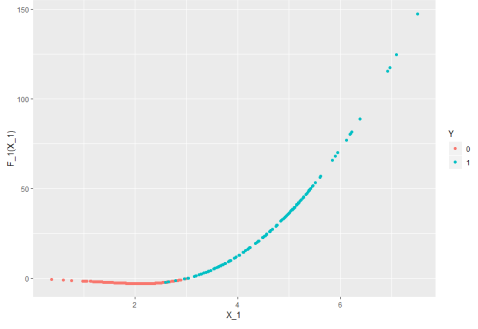
Figura 2: Realización 1 - fronteras lineales con un nodo (M = 2, J = 2 y K = 1)



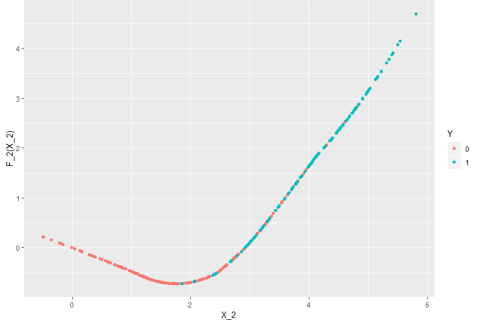

(a) Frontera de predicción




(b) Representación 3D de $\hat{\eta}$

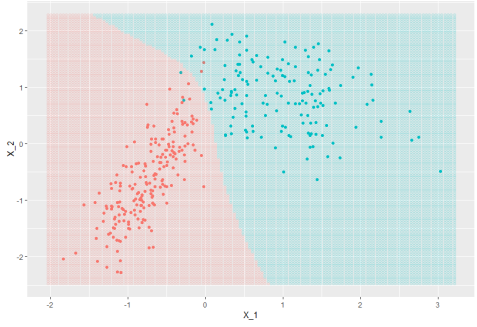



(c) $\hat{f}_1(x_1)$

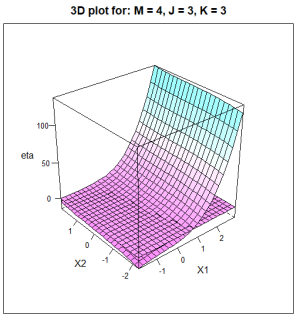



(d) $\hat{f}_2(x_2)$

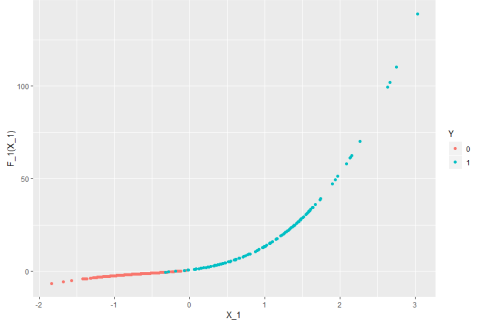
Figura 3: Realización 2 - parábolas continuas mas no suaves (M = 3, J = 5 y K = 1)



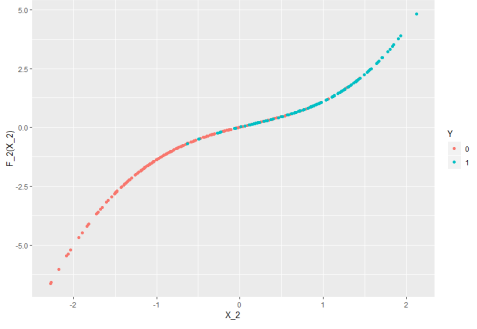

(a) Frontera de predicción




(b) Representación 3D de $\hat{\eta}$




(c) $\hat{f}_1(x_1)$




(d) $\hat{f}_2(x_2)$

Figura 4: Realización 3 - *splines* cúbicos (M = 4, J = 3 y K = 3)

datos en particular, usar un modelo complejo como el *bpwpm* no es del todo necesario pues la base podría ser clasificada con la misma precisión por un modelo que use un predictor lineal en covariables. Sin embargo, se usa la base de datos para ejemplificar los tipos de fronteras flexibles. Asimismo, presentar las formas funcionales que toman las funciones $f_j$ tampoco aportaría mucho pues están compuestas de muchos términos aditivos que no vale la pena desglosar.

## 0.3. Ejemplo 2 - comparación contra un probit lineal

Aprovechando la familiaridad de la base de datos anterior, se decide modificarala para que existan dos regiones de clasificación separadas. Se tomaron trece puntos, más allá de $x_1 \approx 5.5$ y se voltea su clasificación. En la imagen 6a se presenta esta base de datos modificada.

Con afán de comparar las predicciones del modelo *bpwpm* presentado en este trabajo contra uno más convencional, primeramente utiliza un modelo probit lineal en covariables. Es decir, se estiman los parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^t$ del modelo:[9]

$$\begin{aligned} p_i = P(y_i = 1) = \mathbb{E}[y|\mathbf{x}_i] = \Phi(\eta(\mathbf{x}_i)) \quad \Rightarrow \\ \eta(\mathbf{x}_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} \qquad \forall i = 1, \ldots, n \end{aligned} \tag{2}$$

De donde se obtienen los resultados presentados en la tabla 4 y la figura 5. De la imagen anterior, todo lo que quede por arriba de recta será clasificado como uno y todo lo que quede por debajo como cero.

de máquina, debido al tamaño de la $n$, es usual separar la base de datos en dos, una para entrenamiento y otra para probar la efectividad; a este proceso se le conoce como *validación cruzada* y es análogo a la selección de modelos a través de las métricas AIC y BIC.

9. La estimación se realiza bajo el paradigma frecuentista usando el método de mínimos cuadrados a través de la función `glm(..., family = binomial(link = 'probit'))` en `R`.

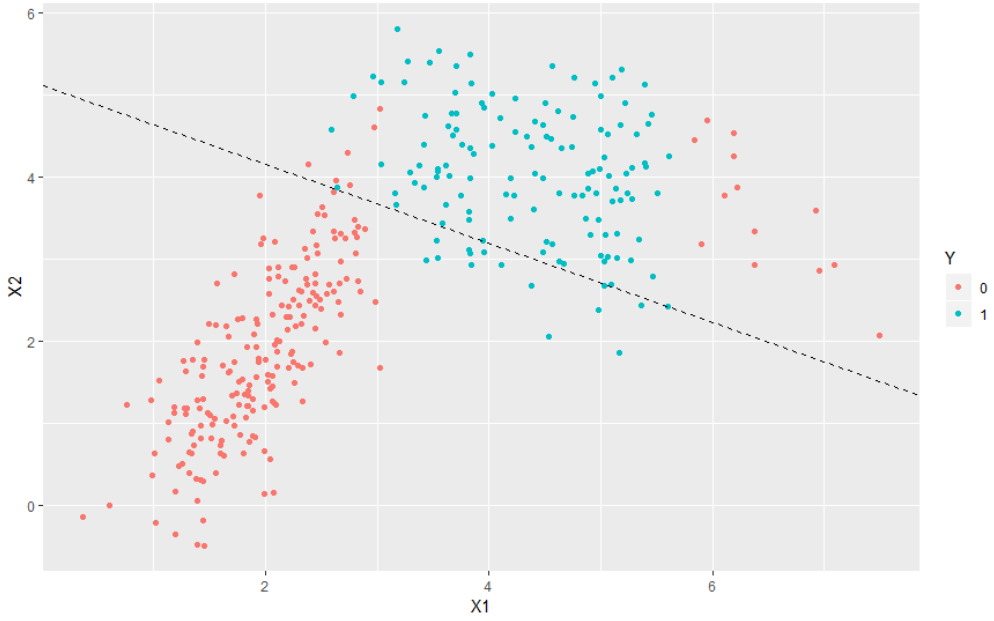



Figura 5: Frontera de predicción para modelo probit lineal en covariables

Por ende, el modelo resultante es:

$$\Phi^{-1}(\hat{p}_i) = \hat{\eta}(\mathbf{x}) = -4.67 + 0.45x_{i,1} + 0.91x_{i,2}, \tag{3}$$

de donde se puede obtener explicitamente la ecuación de la frontera de predicción igualando (3) a 0.5, es decir:

$$\begin{aligned} \Phi(\hat{\eta}(\mathbf{x}_i)) \equiv 0.5 \quad &\iff \\ \hat{\eta}(\mathbf{x}_i) = 0 \quad &\iff \\ 0.45x_{i,1} + 0.91x_{i,2} = 4.67 \quad & \end{aligned} \tag{4}$$

| Parámetro | Estimado |
|---|---|
| $\hat{\beta}_0$ | $-4.67$ |
| $\hat{\beta}_1$ | 0.45 |
| $\hat{\beta}_2$ | 0.91 |

| Info. predicción | |
|---|---|
| Est. Puntual | No aplica |
| Precisión | 90 % |
| *log-loss* | 0.28072 |

| | $\hat{y}=0$ | $\hat{y}=1$ | |
|---|---|---|---|
| $y=0$ | 194 | 19 | 213 |
| $y=1$ | 16 | 121 | 137 |
| | 210 | 140 | 350 |

Tabla 4: Ejemplo 2 - resultados para modelo probit lineal

Para contrastar, ahora se estiman los parámetros del modelo *bpwpm* especificando $M = 3, J = 3$ y $K = 2$ (resumen en la tabla 5). Para este ejemplo, se opta por analizar a fondo todos los componentes y hacer un análisis más detallado de su convergencia, por lo tanto, se presentan los resultados de los estimadores en la tabla 6 e imágenes generadas en la figura 6.

| Parámetros | | Parámetro Sim. |
|---|---|---|
| $M = 3$ | $N^* = 4$ | $N_{\text{sim}} = 10{,}000$ |
| $J = 3$ | $\lambda = 9$ | $k^* = 7{,}500$ |
| $K = 2$ | $n = 350$ | $k_{\text{thin}} = 0$ |

Tabla 5: Ejemplo 2 - regiones disjuntas de clasifcación

Juntando todo, el modelo tiene como predictor lineal a la siguiente forma funcional:

$$\begin{aligned}\Phi^{-1}(\hat{p}_i) = \hat{\eta}(\mathbf{x}) &= \hat{f}_0 + \hat{f}_1(x_{i,1}) + \hat{f}_2(x_{i,2}) \\ &= \overbrace{\hat{\beta}_0}^{\hat{f}_0} \\ &+ \overbrace{\hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,1}^2 + \hat{\beta}_3 (x_{i,1} - \hat{\tau}_{1,1})_+^2 + \hat{\beta}_4 (x_{i,1} - \hat{\tau}_{1,2})_+^2}^{\hat{f}_1(x_{i,1})} \\ &+ \overbrace{\hat{\beta}_5 x_{i,2} + \hat{\beta}_6 x_{i,2}^2 + \hat{\beta}_7 (x_{i,2} - \hat{\tau}_{2,1})_+^2 + \hat{\beta}_8 (x_{i,2} - \hat{\tau}_{2,2})_+^2}^{\hat{f}_2(x_{i,2})},\end{aligned} \tag{5}$$

la cual queda perfectamente definida si se sustituyen los valores de $\boldsymbol{\beta}$ y nodos contenidos en $\mathcal{P}$ presentados en la tabla 6. La ecuación (5) permite observar la expansión en bases final de $\eta(\cdot)$ para esta realización y elección de parámetros. Asimismo, en esta expansión se observa su forma aditiva y el desglose de las funciones $\hat{f}_j$. Es necesario remarcar que la transformación que realizan las funciones no lineales $f_j$, se observa contrastando las imágenes 6a y 6b. Es decir, el espacio inicial de covariables $\mathcal{X}^2$ (imagen 6a) tiene una forma que no puede ser separable por una frontera lineal. Sin embargo al llevar a cabo la transformación no lineal de esta base (imagen 6a), se deriva en un espacio $\widetilde{\Psi}(\mathcal{X})^9 \subseteq \mathbb{R}^9$ que si puede ser separable por una recta. En consecuencia, la frontera de clasificación es disjunta, pues el modelo identifica dos regiones donde los datos deben ser clasificados como cero (imágenes 6c y 6d). Una vez más, se tienen esos pocos puntos que no quedan bien clasificados, incluyendo uno nuevo cerca de las coordenadas cartesianas $(5.8, 2.3)$. Para esta base de datos en particular, se debe usar un nodo adicional cerca de la segunda región, ya que la curvatura, deriva de él. Asimismo, la suavidad de las funciones $\hat{f}$ deriva de la elección de $K$.

Contrastando los resultados del modelo probit lineal (tabla 4) contra el modelo *bpwpm* (tabla 6), se observa que se tiene una mejora en precisión sustancial pues se enfatiza que la

flexibilidad en la frontera viene derivada del número relativamente grande de parámetros. Uno de los beneficios es que para el modelo probit lineal, esta frontera se puede derivar de forma explícita en la ecuación (4), mientras que para el modelo *bpwpm* implicaría resolver numéricamente la ecuación derivada de la expresión no lineal (5). Asimismo, al comparar la métrica *log-loss*, se observa que se tiene una mejora importante. En cuanto a tiempo de estimación computacional, no se tiene una diferencia significativa entre los dos modelos.

| Info. predicción | |
|---|---|
| Est. Puntual | Media posterior |
| Precisión | 98.6 % |
| *log-loss* | 0.04505 |

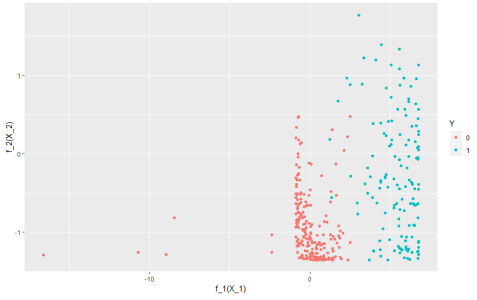
| | $\hat{y}=0$ | $\hat{y}=1$ | |
|---|---|---|---|
| $y=0$ | 210 | 2 | 200 |
| $y=1$ | 2 | 135 | 137 |
| | 212 | 138 | 350 |

| $\boldsymbol{\beta}$ | Valor | |
|---|---|---|
| $\hat{\beta}_0$ | $-2.03$ | $\hat{f}_0$ |
| $\hat{\beta}_1$ | $-1.74$ | $\hat{f}_1(x_{i,1})$ |
| $\hat{\beta}_2$ | 0.90 | |
| $\hat{\beta}_3$ | $-0.07$ | |
| $\hat{\beta}_4$ | $-3.68$ | |
| $\hat{\beta}_5$ | $-1.01$ | $\hat{f}_2(x_{i,2})$ |
| $\hat{\beta}_6$ | 0.13 | |
| $\hat{\beta}_7$ | 0.31 | |
| $\hat{\beta}_8$ | $-0.25$ | |

| $\mathcal{P}$ | Valor | |
|---|---|---|
| $\hat{\tau}_{1,1}$ | 2.07 | Nodos |
| $\hat{\tau}_{1,2}$ | 3.69 | |
| $\hat{\tau}_{2,1}$ | 2.00 | |
| $\hat{\tau}_{2,2}$ | 3.52 | |

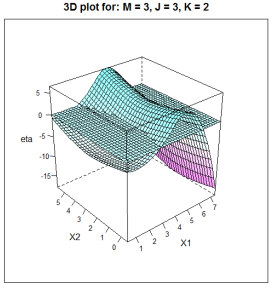
Tabla 6: Ejemplo 2 - resultados


(a) Base del ejemplo 1 modificada

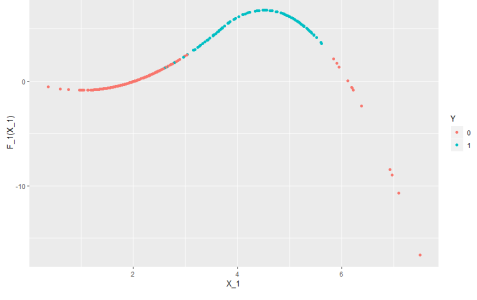



(b) Transformación no lineal

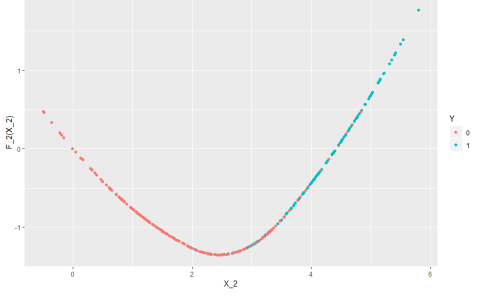

(c) Frontera de predicción




(d) Representación 3D de $\hat{\eta}$




(e) $\hat{f}_1(x_1)$




(f) $\hat{f}_2(x_2)$

Figura 6: Ejemplo 2 - regiones disjuntas de clasificación ($M = 3$, $J = 3$ y $K = 2$)

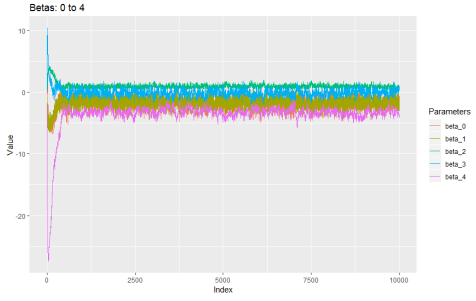
### 0.3.1. Análisis de convergencia

Para finalizar este ejemplo, se busca realizar un análisis detallado de la convergencia de las cadenas pues es parte fundamental del estudio de un modelo bayesiano. Por lo tanto en la tabla 7 se presentan resúmenes numéricos de los parámetros $\boldsymbol{\beta}$ y las cadenas completas en la figura 7.[10]

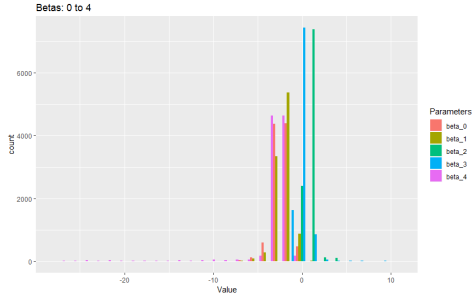
Al analizar las gráficas y los resúmenes, se nota cómo ciertos parámetros como $\hat{\beta}_4$ fluctúan mucho en su estimación en un comienzo, sin embargo de 7e se observa cómo el modelo converge a la larga. Asimismo, de los histogramas y trazas de las cadenas, se observa que éstas tienden a estar bien formadas y presentan vagamente una distribución normal multivariada, estabilizándose conforme el algoritmo de muestreo Gibbs converge al espacio de probabilidad buscado. El periodo de burn-in, se escoge en $k* = 7,500$ pues se busca dar estimaciones puntuales de la media posterior lo más exactas posibles y pareciera que a partir de $k*$ se logra esto pues tanto las cadenas como la media ergódica no fluctúan demasiado. Si únicamente se mostraran los datos de las cadenas recortadas, los histogramas estarían perfectamente formados y tendrían una desviación estándar de aproximadamente uno por construcción.

Para este modelo flexible, aunque los parámetros no estén confundidos, existe la posibilidad de que algunos de ellos converjan a cero pues son innecesarios para la estimación, por ejemplo $\hat{\beta}_3$ y $\hat{\beta}_6$. Para identificar estos parámetros, se pueden aplicar pruebas de hipótesis o procedimientos de selección de variables ya que se cuenta con toda la información que se tendría en un modelo tradicional; sin embargo, se enfatizan los resultados de predicción del modelo completo y no la interpretabilidad de los parámetros individuales.

10. Tanto las imágenes como los resúmenes, aún no han sido ajustados por el periodo de burn-in, de ahí la disparidad contra las estimaciones puntuales de 6. Asimismo, para este ejemplo se presenta una visualización animada en formato `.gif` de como el algoritmo converge conforme avanza el número de iteraciones, consultando `https://github.com/PaoloLuciano/Tesis-Latex`

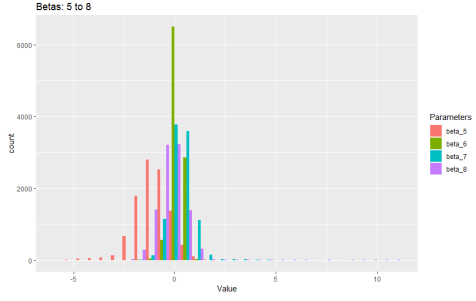



(a) Trazas de $\hat{\beta}_0$ a $\hat{\beta}_4$

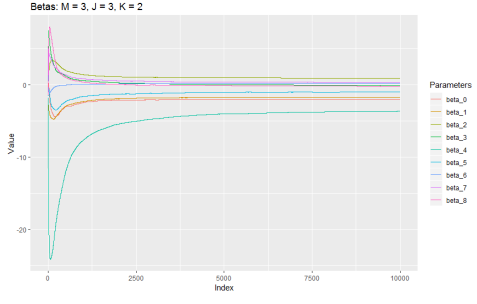



(b) Histogramas de $\hat{\beta}_0$ a $\hat{\beta}_4$


(c) Trazas de $\hat{\beta}_5$ a $\hat{\beta}_8$




(d) Histogramas de $\hat{\beta}_5$ a $\hat{\beta}_8$




(e) Media ergódica

Figura 7: Ejemplo 2 - análisis de convergencia

| Métrica | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ |
|---|---|---|---|---|---|
| Mínimo | −6.79 | −6.63 | −0.39 | −2.11 | −27.40 |
| Primer Cuartíl | −2.54 | −2.22 | 0.66 | −0.48 | −3.72 |
| Media | −2.03 | −1.73 | 0.90 | −0.07 | −3.68 |
| Mediana | −1.98 | −1.69 | 0.85 | −0.09 | −3.28 |
| Tercer Cuartíl | −1.44 | −1.17 | 1.05 | 0.27 | −2.86 |
| Máximo | 0.85 | 1.56 | 4.21 | 10.38 | −1.07 |
| Desviación Estandar | 0.89 | 0.87 | 0.45 | 0.72 | 2.61 |

| Métrica | $\hat{\beta}_5$ | $\hat{\beta}_6$ | $\hat{\beta}_7$ | $\hat{\beta}_8$ |
|---|---|---|---|---|
| Mínimo | −5.59 | −1.64 | −1.89 | −2.47 |
| Primer Cuartíl | −1.49 | −0.04 | −0.05 | −0.69 |
| Media | −1.00 | 0.13 | 0.31 | −0.21 |
| Mediana | −0.97 | 0.13 | 0.27 | −0.27 |
| Tercer Cuartíl | −0.44 | 0.31 | 0.63 | 0.13 |
| Máximo | 2.44 | 1.47 | 6.67 | 11.0 |
| Desviación Estandar | 0.87 | 0.28 | 0.60 | 0.94 |

Tabla 7: Ejemplo 2 - resúmenes numéricos para las cadenas de $\boldsymbol{\beta}$

## 0.4. Ejemplos 3 a 5 - otros resultados interesantes

Los ejemplos presentados a continuación, son más expositivos que analíticos, es decir, se enfatizan los resultados más que los detalles matemáticos como se hizo en la sección anterior. Estos ejemplos y bases de datos simuladas, buscan sobre todo poner a prueba las capacidades no lineales del modelo y estresar las interacciones entre las dimensiones. Al estar tratando con regiones de clasificación más complejas, la predicción correcta sería imposible para un modelo lineal en covariables.

## Ejemplo 3 - región parabólica

Para este ejemplo, se generaron $n = 400$ datos en $\mathbb{R}^2$ usando coordenadas polares al tomar ángulos con un rango entre $[-1, 1]$. Posteriormente se tomaron diferentes radios para diferenciar cada grupo y finalmente se les sumó ruido blanco a los puntos para que existiera una región de confusión. Las simulación derivó en un patrón de datos cuya frontera es curva, casi parabólica. Dadas las características de los datos, se piensa que usar polinomios por partes parabólicos y suaves ($M = 3$ y $K = 2$) es una buena opción para modelarlos. Los parámetros escogidos para la realización final del modelo, se presentan en la tabla 8. Asimismo, los resultados e imágenes se presentan en la tabla 9 y la figura 8 respectivamente.

| Parámetros | | Parámetro Sim. |
|---|---|---|
| $M = 3$ | $N^* = 5$ | $N_{\text{sim}} = 10{,}000$ |
| $J = 4$ | $\lambda = 11$ | $k^* = 2{,}500$ |
| $K = 2$ | $n = 400$ | $k_{\text{thin}} = 0$ |

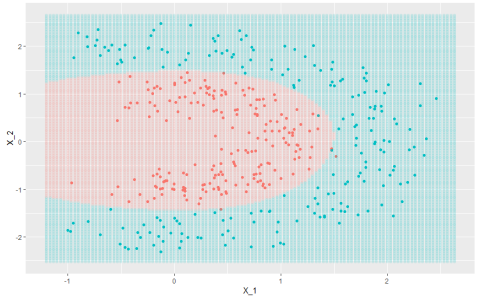
Tabla 8: Ejemplo 3 - región parabólica

| Info. predicción | |
|---|---|
| Est. Puntual | Media posterior |
| Precisión | 99.2 % |
| *log-loss* | 0.04352 |

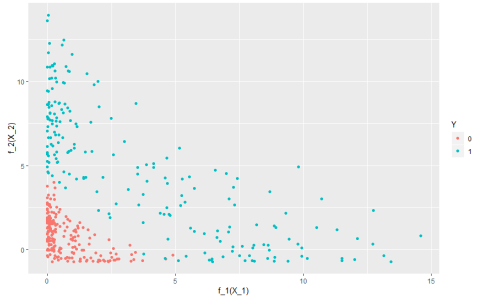
| | $\hat{y} = 0$ | $\hat{y} = 1$ | |
|---|---|---|---|
| $y = 0$ | 198 | 2 | 200 |
| $y = 1$ | 1 | 199 | 200 |
| | 199 | 201 | 400 |

Tabla 9: Ejemplo 3 - resultados

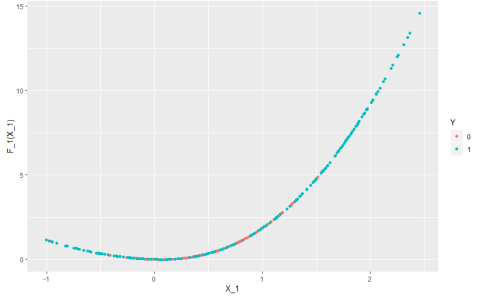
Esta es una realización particularmente interesante pues con un total $\lambda = 11$ parámetros se logra una precisión alta además de obtener convergencia relativamente rápido ($N_{\text{sim}} = 10{,}000$ y $k^* = 2{,}500$). Analizando el modelo de forma gráfica, se observa claramente que la segunda transformación $\hat{f}_2(x_2)$ (imagen 8d) captura la parte parabólica. A la vez, la primera transformación $\hat{f}_1(x_1)$ (imagen 8c) le da poco peso a la región donde hay confusión entre los los grupos pero posteriormente crece en donde hay certidumbre. Asimismo, se presenta



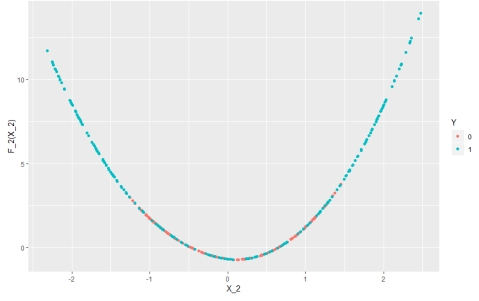

(a) Frontera



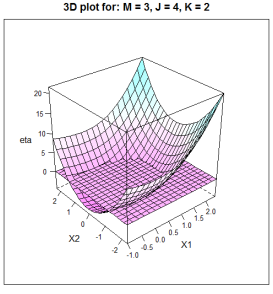

(b) Transformación no lineal



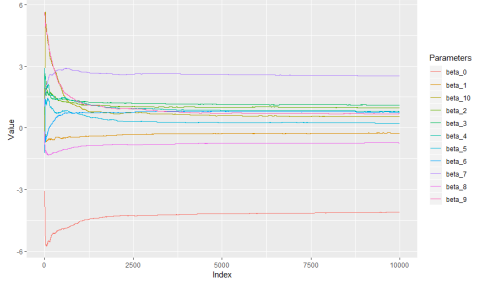

(c) $\hat{f}_1(x_1)$




(d) $\hat{f}_2(x_2)$




(e) Representación 3D de $\hat{\eta}$




(f) Medias ergódicas

Figura 8: Ejemplo 3 - parábolas suaves ($M = 3$, $J = 4$ y $K = 2$)

el espacio de la transformación no lineal en la imagen 8b en donde se observa que el grupo rojo cero, se concentra en la esquina inferior izquierda, representando la posible separación lineal en este espacio transformado.

## Ejemplo 4 - región ovalada

Para esta base de datos en particular se busca replicar algo similar a la imagen del capítulo introductorio **??** de la página **??**. Se obtuvo una base de datos pequeña del curso en linea de aprendizaje de máquina de **andrew2018ml** que presenta una frontera de clasificación ovalada.[11] Esta base de datos se usa para entrenar modelos saturados logit con regularización, **hastie2008elements**, logrando predecir fronteras curvas con modelos tradicionalmente lineales al incluir interacciones de orden mayor entre covariables. Por lo tanto, se decidió probarlo también con el modelo presentado para contrastar.

El modelo una vez más, fue ajustado usando parábolas suaves las cuales resultaron ser excelentes herramientas. Los parámetros escogidos para la realización final del modelo, se presentan en la tabla 10 con resultados e imágenes en la tabla 11 y figura 9 respectivamente.

| Parámetros | | Parámetro Sim. |
|---|---|---|
| $M = 3$ | $N^* = 3$ | $N_{\text{sim}} = 2{,}000$ |
| $J = 2$ | $\lambda = 7$ | $k^* = 500$ |
| $K = 2$ | $n = 118$ | $k_{\text{thin}} = 0$ |

Tabla 10: Ejemplo 4 - región ovalada

Para esta realización del modelo, se buscó estresar su flexibilidad al incluir el menor número

11. Este curso, se ofrece de forma gratuita en la plataforma de MOOC's Coursera.

| Info. predicción | |
|---|---|
| Est. Puntual | Media posterior |
| Precisión | 78.8 % |
| *log-loss* | 0.4714 |

| | $\hat{y}=0$ | $\hat{y}=1$ | |
|---|---|---|---|
| $y=0$ | 48 | 12 | 60 |
| $y=1$ | 13 | 45 | 58 |
| | 61 | 57 | 118 |

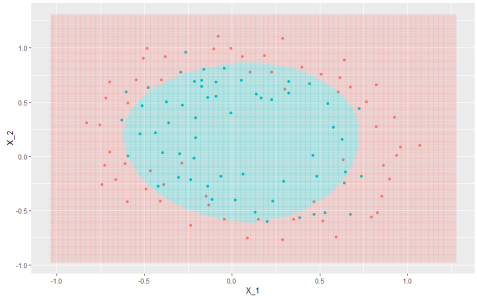
Tabla 11: Ejemplo 4 - resultados

de términos posibles usando un solo nodo ($\lambda = 7$ y $J = 2$) y cadenas cortas ($N_{\text{sim}} = 2{,}000$). Aunque una precisión de 78.8 % no resulte tan atractiva, es la precisión que se presenta en el curso en línea y permanece constante aún si se aumenta $\lambda$. La métrica *ll* mejora (marginalmente) sobre la presentada en el curso ($\approx 0.5$), sin embargo, se logra una reducción significativa en el número de parámetros pues el modelo saturado de **andrew2018ml** inicia con 28 parámetros.[12] Asimismo, como se observa en la figura 9f las cadenas convergen rápidamente. Todo el poder del modelo, recae en la forma funcional de las funciones $\hat{f}_j$ al poder estimar regiones irregularmente curvas, con pocas observaciones y parámetros.

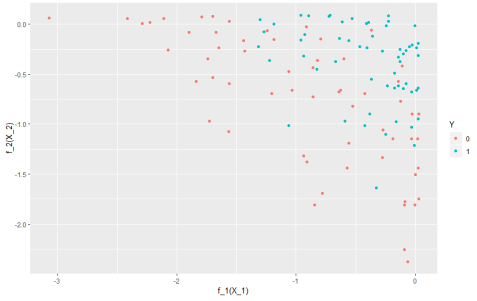
**Ejemplo 5 - *yin-yang*, limitaciones del modelo**

Para finalizar con las bases de datos simulados, el modelo *bpwpm* se llevó al límite de sus capacidades sobre un patrón de puntos, intuitivo al ojo humano, pero difícil de identificar por un modelo de este tipo. Los datos tratan de simular un *yin-yang* que se puede observar en la figura 10a. La simple simulación de la base de datos representó un reto donde se conjuntaron varias áreas de la matemática aplicada. En el software `GeoGebra`, se generó el

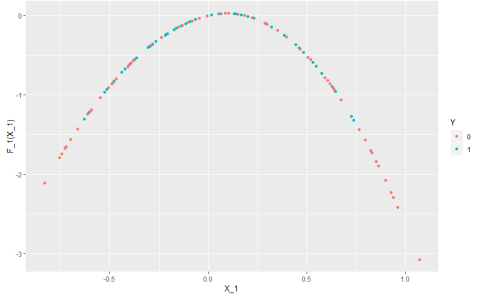
12. Dada la regularización, muchos de estos parámetros se desvanecían.



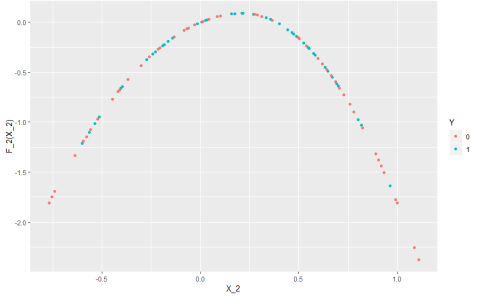

(a) Frontera



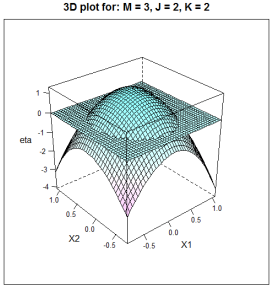

(b) Transformación no lineal



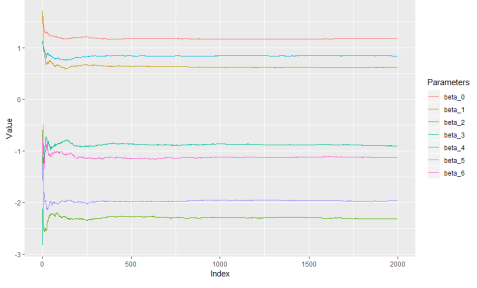

(c) $\hat{f}_1(x_1)$




(d) $\hat{f}_2(x_2)$




(e) Representación 3D de $\hat{\eta}$

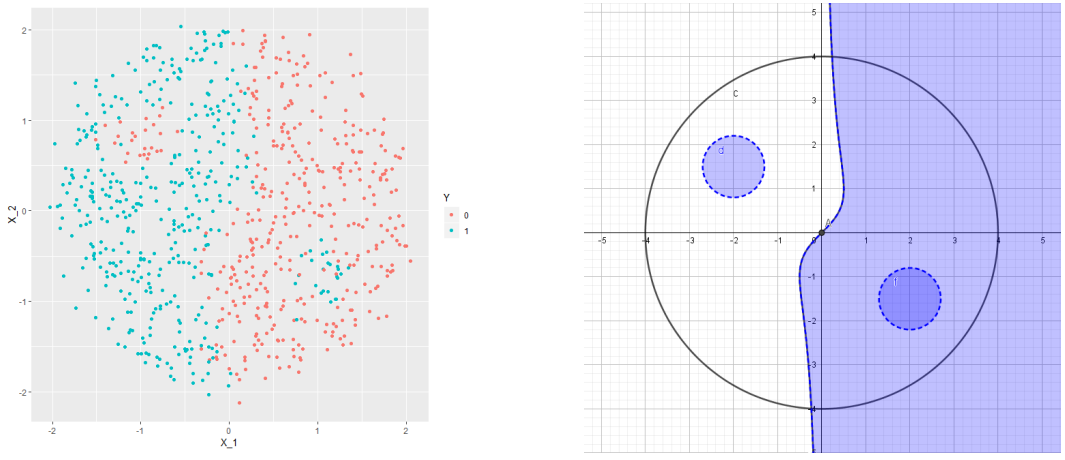



(f) Medias ergódicas

Figura 9: Ejemplo 4 - parábolas suaves en un nodos ($M = 3$, $J = 2$ y $K = 2$)

diagrama presentado en la figura 10b que consiste de las siguientes desigualdades cartesianas:

$$x^2 + y^2 < 16,$$
$$(x+2)^2 + (y-1.5)^2 < 0.49,$$
$$(x-1.5)^2 + (y+2)^2 < 0.49,$$
$$x < \frac{y}{(1+y^2)}.$$



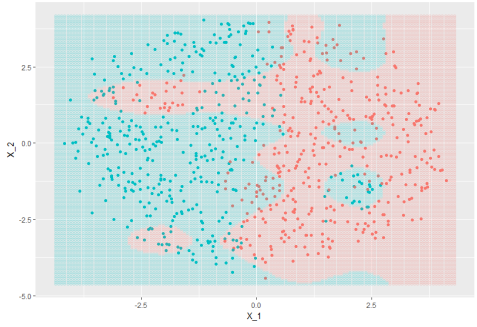

(a) Datos simulados

(b) Salida del software `GeoGebra`

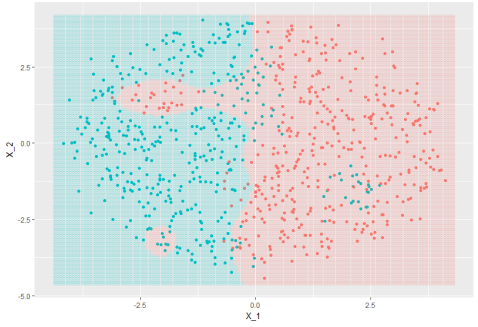
Figura 10: Ejemplo 5 - patrón yin-yang

Una vez dibujadas las ecuaciones, se generó una base de datos de aproximadamente $n \approx 800$ observaciones con una distribución uniforme dentro del círculo usando coordenadas polares. A estos puntos se les asignó la categoría cero, posteriormente se cambió la categoría a los puntos que cumplieran con las desigualdades. Después, se le añadió algo de ruido normal a cada punto para darle aleatoriedad a la base de datos pero manteniendo el patrón y finalmente, se escala la base para centrarla en cero.

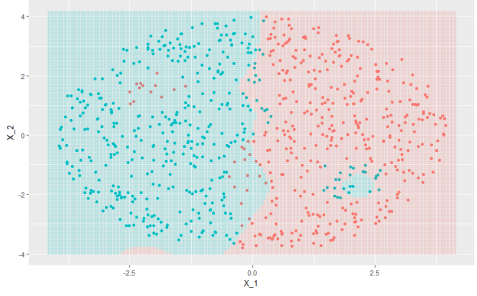
Se corrieron un sinfín de realizaciones del modelo, tratando de calibrar los parámetros $M$, $J$ y $K$ para captar de la mejor manera posible el patrón. Sin embargo y aunque el modelo casi siempre lograba una precisión de cerca de 85 %, no se logra la clasificación esperada identificando los puntos de color dentro de las áreas opuestas. De cualquier forma se observa cómo el algoritmo está tratando de encontrar este patrón. En la figura 11 se pueden ver fronteras de algunos de los mejores modelos.[13] Para las imágenes 11c y 11b, se observa



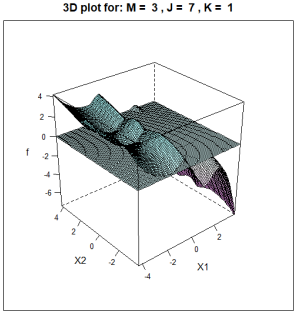

(a) Sobre-ajuste




(b) Mejor modelo




(c) Falta de precisión




(d) Gráfico 3D para uno de los modelos

Figura 11: Fronteras de varios modelos para datos yin-yang

cómo el modelo está tratando de encontrar las regiones anidadas, sin embargo, nunca se

13. En la imagen 11a, el modelo detecta relativamente bien la curva que separa las regiones y detecta de forma aislada el circulo azul de la esquina inferior derecha.

logra de forma precisa. Finalmente la imagen 11d, muestra una de las muchas representaciones 3D que se hicieron al tratar de ajustar esta base de datos. Precisamente en esta última imagen esconde el porqué no se logró hacer la estimación correcta: la dependencia implícita entre los nodos. Estos nodos, en realidad están dividiendo el espacio bi-dimensional en una malla cuadriculada donde las interacciones son difíciles de discernir. Conforme aumenta el número de nodos, más complejo se vuelve el modelo. Es por ello, que los picos y valles se repiten en un patrón uniforme. Asimismo, dada la naturaleza global de los polinomios y esta interacción, el modelo tiene esta estructura decreciente siempre, derivando que los picos y los valles nunca alcancen las regiones extremas en polos opuestos. De igual forma, la uniformidad y simetría impar, inherente a esta base de datos, llevó a que la estimación de los parámetros fuera óptima dentro de las capacidades del modelo. Otra desventaja de esta base, es que estos modelos se tuvieron que correr con un número grande de nodos $J \approx 20$, derivando en un número de parámetros aún mayor.

## 0.5. Ejemplo 6 - el modelo en la práctica

Hasta ahora, todos los resultados de este trabajo han sido sobre bases de datos simuladas. Claramente se forman imágenes atractivas por construcción, sin embargo, no se está prediciendo nada en realidad pues se utiliza una metodología *dentro de muestra* para enfatizar las posibles fronteras del modelo. Por lo tanto y como último ejemplo, se presenta la base de datos de cáncer de mama de la Universidad de Wisconsin. Esta base de datos es citada en varios trabajos de los años noventa, donde se tratan de hacer clasificaciones binarias usando una serie de procedimientos más robustos que los tradicionales GLM, **mangasarian1990pattern** y **bennett1992robust**.

De manera general y sin entrar en el detalle biológico de las variables como tal, se presenta un

análisis exploratorio preliminar que se lleva a cabo para seleccionar, de forma completamente subjetiva, las que se consideren relevantes. La base de datos cuenta con $n = 699$ observaciones de las cuales el 34.5 % representan pacientes infectados con tumores malignos representados por el color rojo (etiqueta cero). Se cuenta con diez covariables (dimensiones) médicas sobre las características de los tumores como lo son: el tamaño, la uniformidad de la pared celular y la cromatina.[14] En la figura 12, se muestran los gráficos de puntos pareados para todas las posibles combinaciones de covariables además de cierta información adicional. Se hace notar

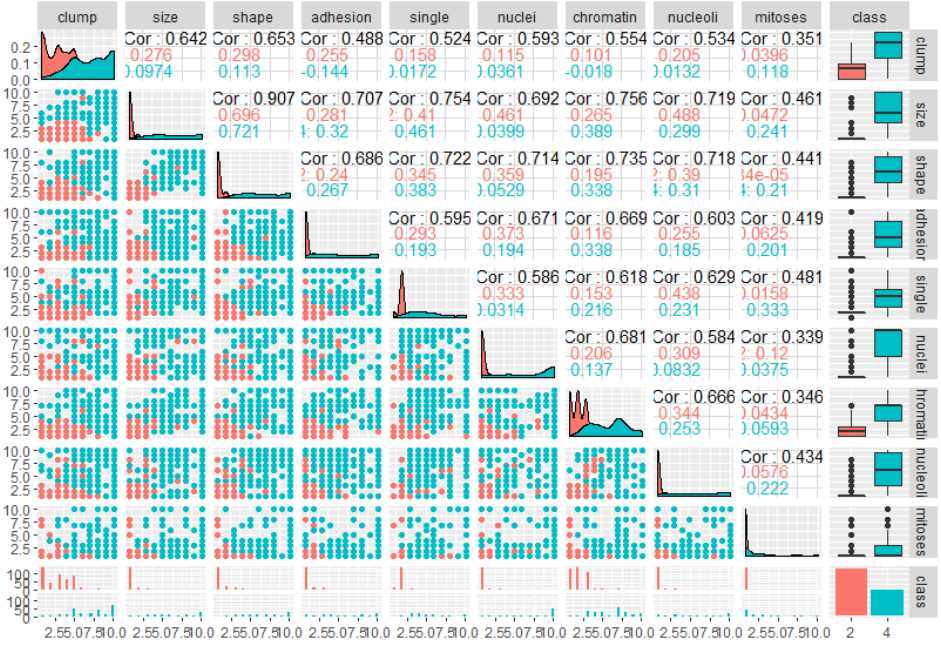
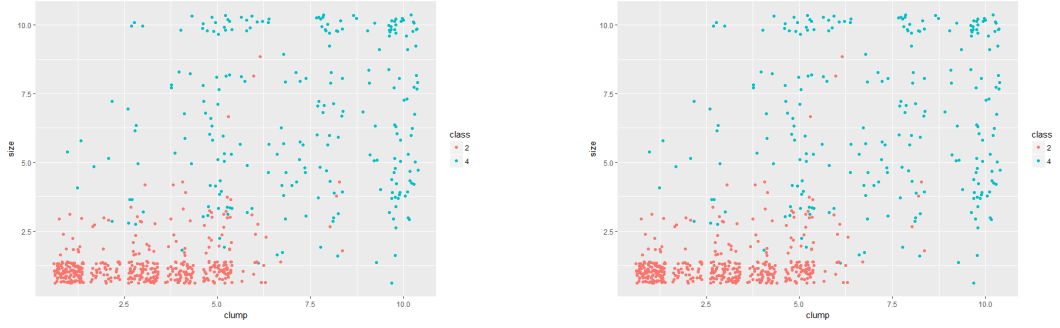



Figura 12: Análisis exploratorio para selección de variables

que las covariables están codificadas en una escala a 10 puntos, por lo tanto, la representación gráfica de los datos se ve más como una cuadrícula que como un espacio real de variables. Derivado de esta exploración previa, se seleccionan las covariables *clump, size* y *chromatin*

14. Forma en la que se presenta la cadena de ADN en el núcleo celular.

debido a que parecieran ser las que mejor separan el espacio.[15] En la figura 13 se presentan dos gráficos de puntos con algo de ruido para hacer notar que las regiones son un poco más complejas de lo que podría parecer en una primera exploración, además se tienen puntos idénticos con clasificaciones contrarias. Sin embargo, a simple vista se detecta cierto patrón en los datos.




(a) Variables *clump* y *chromatin*

(b) Variables *clump* y *size*

Figura 13: Gráficos de puntos con ruido para separar las observaciones

Para poder hablar de predicción como tal, tiene que existir una base de datos contra la cual probar las estimaciones del modelo. Por lo tanto, la base original se parte en dos: un conjunto de entrenamiento con el 60 % de las observaciones ($n_{\text{train}} = 409$) y un conjunto de prueba con las observaciones restantes ($n_{\text{test}} = 274$) sobre las que se evaluará el modelo.[16] La realización final de entrenamiento del modelo se resume en la tabla 12, se escogen segmentos de recta continuos sobre tres nodos. Los resultados numéricos sobre la base de datos de prueba se presentan en la tabla 13 y el análisis de convergencia a través de las medias ergódicas en la figura 14a. Asimismo, se presenta cada $\hat{f}_j \quad j = 1, 2, 3$ en las figuras 14b, 14c y 14d respectivamente.

15. Estas covariables corresponden a el espesor de los tumores, su tamaño y la textura de la cromatina en las células respectivamente.

16. La diferencia de 16 observaciones entre la suma de entrenamiento y prueba, contra las 699 originales, se debe a que estas estaban incompletas y por lo tanto se descartan.

Haciendo una predicción fuera de muestra los resultados son buenos logrando una precisión del 95.6 %. Asimismo, se resalta que inclusive en dimensiones ($d = 3$) más altas si se escogen los parámetros adecuados $M$, $J$ y $K$, el número total de parámetros ($\lambda = 13$) no necesita aumentar demasiado para lograr una buena separación del espacio.[17] Sin embargo, derivado también del número de covariables es que no se pueden hacer una visualización en el plano cartesiano $\mathbb{R}^2$ como en los ejemplos anteriores. No obstante, la convergencia es clara y los resultados buenos, incluso usando segmentos de recta y un número de nodos pequeño. Se hace notar que la codificación de las covariables usando una escala de diez puntos no es óptima para un modelo que se construye pensando en un espacio real de covaraibles $\mathcal{X}^d$, sin embargo, no parece afectar la estimación de los parámetros.

| Parámetros | | Parámetro Sim. |
|---|---|---|
| $M = 2$ | $N^* = 4$ | $N_{\text{sim}} = 50{,}000$ |
| $J = 4$ | $\lambda = 13$ | $k^* = 10{,}000$ |
| $K = 1$ | $n = 409$ | $k_{\text{thin}} = 0$ |

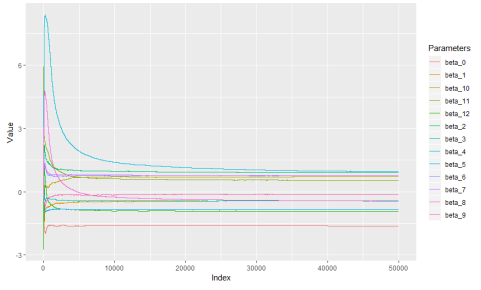
Tabla 12: Ejemplo 6 - datos médicos de cáncer

| Info. predicción | |
|---|---|
| Est. Puntual | Media posterior |
| Precisión | 95.6 % |
| *log-loss* | 0.1561 |

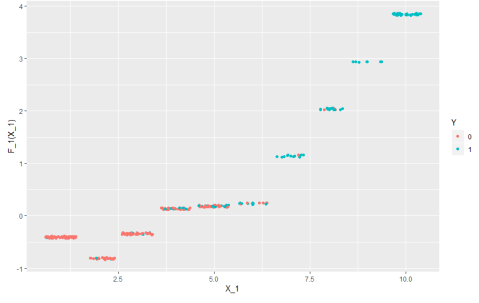
| | $\hat{y} = 0$ | $\hat{y} = 1$ | |
|---|---|---|---|
| $y = 0$ | 169 | 9 | 178 |
| $y = 1$ | 3 | 93 | 96 |
| | 172 | 102 | 274 |

Tabla 13: Ejemplo 6 - resultados

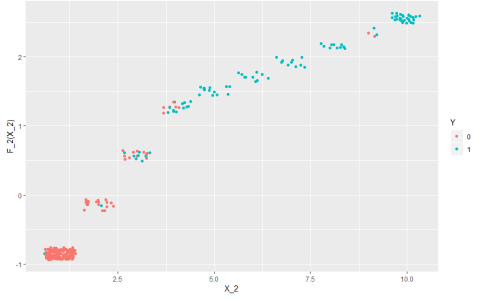
17. Se corrieron otras realizaciones del modelo aumentando el número de covariables $d$ y ajustando los parámetros $M$, $J$ y $K$. Sin embargo, no logró aumentar significativamente la precisión del modelo.




(a) Medias ergódica




(b) $\hat{f}_1(x_1)$ con ruido




(c) $\hat{f}_2(x_2)$ con ruido


(d) $\hat{f}_3(x_3)$ con ruido

Figura 14: Media ergódica y funciones $\hat{f}_j(x_j) \quad j = 1, 2, 3$