Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

---

box1976science

Como base fundamental de este trabajo, a continuación, se inicia la construcción de un modelo de clasificación binaria flexible. En la perspectiva del autor, no se está tratando de construir un modelo que replique el proceso generador de los datos. Más bien, se está tratando de construir una útil abstracción de la realidad a través de un modelo estadístico. Vale la pena tener en mente que escoger cualquier enfoque de modelado, es un proceso reduccionista y por ende, falible. No obstante, no significa que no se puedan discernir patrones en los datos y aprender de ellos: extraer la señal implícita del ruido.

Al modelo desarrollado se le titula *bayesian piecewise polynomial model (bpwpm)* por sus siglas en inglés, nombre que engloba los componentes fundamentales de este. El modelo en si, es un modelo estructurado notacionalmente pesado por lo que se opta por presentación constructiva. Además, esta presentación tiene la peculiaridad que sigue de cerca el desarrollo histórico del aprendizaje estadístico. No obstante, en la sección 0.4 se puede encontrar el modelo de forma preliminar, mientras que la versión más completa se presenta en la sección **??**.[1]

1. Aunque al comienzo de este trabajo se presenta un glosario de los símbolos y signos usados, se busca respetar la notación usada en los libros **hastie2008elements** y **james2013introduction**.

0.1. Modelos lineales generalizados (GLM)

A medida que avanzó la disciplina de la estadística durante el siglo veinte, se desarrollaron de forma independiente muchos modelos que permitieron estudiar una mayor variedad de datos: binarios, proporciones y datos continuos cuyo error se no se distribuye normal. Sin embargo, todos ellos surgen como una extensión del modelo de regresión lineal y son agrupados en **maccullagh1989generalized** nombrádolos modelos lineales generalizados o GLM por sus siglas en inglés.

De forma análoga, el modelo pertinente se construirá a partir del modelo de regresión lineal definido de una forma peculiar.[2]

$$
\begin{aligned}
y_i|\mathbf{x}_i &\sim \mathcal{N}(y_i \,|\, \mu(\mathbf{x}_i), \sigma^2) \quad \forall i = 1, \ldots, n \\
\mu(\mathbf{x}_i) &= \beta_0 + \tilde{\boldsymbol{\beta}}^t \mathbf{x}_i,
\end{aligned}
$$

donde $y_i \in \mathbb{R}$, $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^d$ es un vector de parámetros y $\mu(\mathbf{x}_i) = \mathbb{E}[y_i|\mathbf{x}_i]$ la media de la regresión.[3] Las regresiones lineales, están acotadas a casos donde la variable de respuesta y_i tenga soporte real, lo cual reduce mucho los datos a los que este modelo es aplicable. Por lo tanto, la principal modificación en los GLM es que busca flexibilizar este soporte a diferentes variables de respuesta. Esta modificación vuelve al modelo más complejo y deriva en diversas métodos para la estimación de $\boldsymbol{\beta}$

2. El escoger esta definición particular, se irá esclareciendo conforme se construya el modelo *bpwpm*.

3. Se respeta la convención de usar negritas para distinguir vectores $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})^t$. Asimismo, se utiliza $\tilde{\boldsymbol{\beta}}$ para distinguir al vector de dimensionalidad d y a $\boldsymbol{\beta}$ para distinguir al vector que contiene el término independiente, es decir, $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$.

diferentes de la tradicional minimización de residuales cuadrados (RSS). Asimismo, la generalización del modelo lleva a que la intrepretación de los parámetros no sea trivial generalmente.[4]

4. Por ejemplo, en un modelo logit que busca la predicción de variables binarias, se logra expresar el logaritmo de la proporción de probabilidades (*Log-Odds-Ratio*) como una combinación lineal de las covariables. $\ln(p_1/p_0) = \boldsymbol{\beta}^t x$, donde p_k con $k = \{0, 1\}$, es la probabilidad de que la respuesta y sea 0 o 1 respectivamente.

Definición 0.1. La familia de modelos lineales generalizados (GLM), **sundberg2016exponent** tienen la forma:

$$\begin{aligned} y &\sim F(y \mid \mu(\mathbf{x})) \\ \eta(\mathbf{x}) &= \beta_0 + \tilde{\boldsymbol{\beta}}^t \mathbf{x} \\ g(\mu) &= \eta(\mathbf{x}), \end{aligned} \tag{1}$$

cuentan con los siguientes tres elementos:

F: distribución de la familia exponencial que describe el dominio de las respuestas y, cuya media $\mu(\cdot)$ es dependiente de las covariables.[5] Por ejemplo: Bernoulli si y es binaria, Poisson si $y \in \mathbb{Z}^+$ o una distribución Gamma si $y \in \mathbb{R}^+$.

η: predictor lineal que explique la variabilidad sistemática de los datos.

g: función liga que une la media μ de la distribución con el predictor lineal,[6] es decir: $g(\mu(x)) = g(\mathbb{E}[y|x]) = \eta(\mathbf{x}) = \boldsymbol{\beta}^t \mathbf{x}$, donde g puede ser cualquier función monótona que idealmente mapee de forma suave y biyectiva el dominio de la media μ con el rango del predictor lineal η (**hardle2004semiparametric**).

5. Al trabajar con distribuciones de la familia exponencial es usual parametrizar la distribución no con la media μ sino con el parámetro canónico θ.

6. Si la función g es tal que $\eta \equiv \theta$ entonces se dice que g es la función liga canónica.

0.1.1. El modelo binario

Dado que se busca construir un clasificador supervisado donde las respuestas observadas sean binarias, es decir: $y_i \in \{0, 1\}\ \forall i = 1, \ldots, n$, se enfoca la discusión en la familia de modelos binarios, caraterizados por la distribución Bernoulli la cual surge de manera natural. Es decir, se permite que $F = \text{Be}$ quedando entonces:

$$y_i \sim \text{Be}(y_i \,|\, p_i) \tag{2}$$

con $\mu = p$. La distribución Bernoulli (2) tiene una estructura sencilla que puede ser resumida en la siguiente función de masa de probabilidad: $\forall i = 1, \ldots, n$,

$$f(y_i|p_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \tag{3}$$

donde,

$$\begin{aligned} y_i &\in \{0, 1\}, \\ \mathbb{E}[y_i] &= \mu_i = P(y_i = 1) = p_i \\ \mathbb{V}[y_i] &= p_i(1 - p_i). \end{aligned}$$

En (3) se observa la función de masa de probabilidad Bernoulli en su forma tradicional que puede ser reexpresada para que cumpla la definición de la familia exponencial.[7] Dado el soporte y la definición de la distribución Bernoulli, la media de la

7. Una distribución (de un solo parámetro) se dice que pertenece a la familia exponencial si se puede expresar de la forma: $f(y; \theta) = h(y) \exp\{y \cdot \theta - A(\theta)\}$ con $h(y)$, $A(\theta)$ funciones conocidas y θ el parámetro canónico, en el caso Bernoulli: $\theta(p) = \ln p/(1 - p)$.

distribución $\mu = p$ coincide con la probabilidad de que la variable aleatoria tome el valor de uno. Asimismo, la varianza queda especificada por el mismo parámetro p.

El que la media conocida con la probabilidad de éxito en una distribución Bernoulli ($\mu = p$) es de gran utilidad en un contexto de clasificación por varias razones. Primero, al modelar la media, se está caracterizando por completo la distribución y la predicción de la variable y. Segundo, se restringen las posibles funciones liga a las funciones que mapean de forma biyectiva $\mathbb{R}$, el dominio del predictor lineal η, al rango de la media, el intervalo $(0, 1)$ que se interpreta una probabilidad. Dadas estas propiedades buscadas, es usual usar como función liga a las inversas de funciones *sigmoidales*. Las funciones sigmoidales, son funciones $s : \mathbb{R} \to (0, 1)$ estrictamente monótonas y por ende, biyectivas. Algunos ejemplos son la ya mencionada logit, la función probit que concierne a este trabajo o la curva de Gompertz. Estas funciones cumplen un papel de activación, es decir, una vez que el predictor lineal cruce cierto umbral, crecen rápidamente y toman valores más cercanos a uno, *activando* así la probabilidad de que y sea uno.[8]

El modelo probit

En particular, para este trabajo se escoge como función liga a la función probit dándole nombre al modelo. Este es un supuesto fuerte que responde a la forma en la

8. En un contexto de aprendizaje de máquina, se les conoce como funciones de activación a las inversas de la funciones liga g^{-1} que no necesariamente tienen que ser biyectivas (**bishop2006pattern**). Por ejemplo, en redes neuronales es común utilizar la función $ReLu(x) := \text{máx}\{0, x\}$ la cual no es suave (**3blue1brown2017**).

cual se define el modelo que, como se verá en el teorema 0.4 y capítulos subsecuentes, permite desarrollar un algoritmo para la estimación bayesiana de los parámetros, **albert1993bayesian**.

La función probit es la función inversa de la función de acumulación normal estándar $\Phi(\cdot) : \mathbb{R} \to (0, 1)$, por lo tanto $g(\mu) = g(p) = \text{probit}(p) = \Phi^{-1}(p)$. Dado que la notación puede ser confusa, en la figura 1 se presenta una representación gráfica de la función liga para un modelo probit.[9]

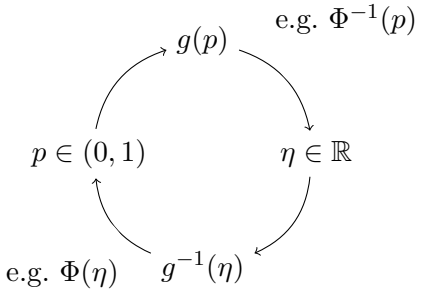

Figura 1: Esquema de función liga g para un modelo probit

Juntando todos los componentes, se está en posibilidades de definir el modelo probit en su forma más rigurosa, el cual es la base para la construcción del clasificador binario que se busca.

9. Para no caer en redundancia de notación se tiene a partir de ahora: $s(x) = g^{-1}(x) = \Phi(x)$ la función de acumulación normal estándar. Asimismo, se deja de usar μ para referirse a la media y se utiliza únicamente p.

Definición 0.2. Rescatando la notación de un GLM (1) con sus respectivas covariables $\mathbf{x}_i$ se tiene el modelo probit (versión 1):

$$y_i \,|\, \mathbf{x}_i \sim \text{Be}(y_i \,|\, p_i) \qquad \forall i = 1, \ldots, n \tag{4}$$

$$\eta(\mathbf{x}_i) = \beta_0 + \tilde{\boldsymbol{\beta}}^t \mathbf{x}_i \tag{5}$$

$$p_i = \Phi(\eta_i) = \Phi(\eta(\mathbf{x}_i)) \tag{6}$$

La definición 0.2 aunque sencilla, no corresponde a la que se usará para definir el modelo *bpwpm* final pues, como busca utilizar un paradigma bayesiano de aprendizaje resultará más sencillo definir al modelo probit introduciendo una estructura adicional conocida como *variable latente*, dando lugar a toda una clase de modelos por si mismos. Por lo tanto este se redefine, declarando así las dos primeras líneas del modelo *bpwpm*.

Definición 0.3. El modelo probit (versión 2):

$$y_i = \begin{cases} 1 & \iff \quad z_i > 0 \\ 0 & \iff \quad z_i \leq 0 \end{cases} \tag{7}$$

$$z_i \,|\, \mathbf{x}_i \sim \mathcal{N}(z_i \,|\, \eta(\mathbf{x}_i), 1) \tag{8}$$

Esta definición introduce n variables latentes z_i, independientes entre si que se distribuyen de forman normal $\mathcal{N}$. Por ser latentes, se relacionan de forma unívoca con

las respuestas y_i formando una clase de equivalencia entre la probabilidad de dos eventos, permitiendo que se asocie el soporte binario de y_i con el soporte real de z_i. Es decir, (7) y (8) implican:

$$P(y_i = 1|\mathbf{x}_i) = P(z_i > 0 \,|\, \mathbf{x}_i) = \Phi(\eta(\mathbf{x}_i)). \tag{9}$$

A continuación, se prueba la equivalencia entre las dos definiciones y la identidad fundamental (9)

Teorema 0.4. *Un modelo probit definido en 0.2 es equivalente a un modelo de variable latente como el presentado en 0.3.*

Demostración. Dado un modelo probit 0.2 se tiene sin perdida de generalidad $\forall i = 1, \ldots, n$:

$$\begin{aligned} \mathbb{E}[y_i \,|\, \mathbf{x}_i] &= p_i \\ &= P(y_i = 1 \,|\, \mathbf{x}_i) \\ &= \Phi(\eta(\mathbf{x}_i)) \end{aligned}$$

Por las ecuaciones por (4) y (6). Lo anterior es equivalente a introducir n variables

aleatorias $\tilde{z}_i \sim \mathcal{N}(\tilde{z}_i \,|\, 0, 1)$ tales que:

$$\begin{aligned}
\Phi(\eta(\mathbf{x}_i)) &= P(\tilde{z}_i \leq \eta(\mathbf{x}_i) \,|\, \mathbf{x}_i) && \text{por definición de la función de acumulación} \\
&= P(\tilde{z}_i > -\eta(\mathbf{x}_i) \,|\, \mathbf{x}_i) && \text{por simetría de la distribución normal} \\
&= P\left(\frac{\tilde{z}_i + \eta(\mathbf{x}_i)}{1} > 0 \,\middle|\, \mathbf{x}_i \right) \\
&= P(z_i > 0 \,|\, \mathbf{x}_i).
\end{aligned}$$

Donde $z_i = \tilde{z}_i + \eta(\mathbf{x}_i)$ es una transformación biyectiva de $\tilde{z}_i$ tal que,

$$z_i \,|\, \mathbf{x}_i \sim \mathcal{N}(z_i \,|\, \eta(\mathbf{x}_i), 1),$$

lo cual es idéntico a la expresión (8). Asimismo, al tener la igualdad

$$P(y_i = 1 \,|\, \mathbf{x}_i) = P(z_i > 0 \,|\, \mathbf{x}_i)$$

y por ende su probabilidad complementaria $P(y_i = 0 \,|\, \mathbf{x}_i) = P(z_i \leq 0 \,|\, \mathbf{x}_i)$, se define una clase de equivalencia entre probabilidades. Es decir, se puede definir y_i en términos de z_i y viceversa, dando lugar a la ecuación (7) y la identidad (9).

El argumento es casi idéntico si la demostración se comienza suponiendo la definición 0.3 con variable latente y se construye hasta llegar a un GLM como el definido en 0.2. Sin embargo, se tiene la peculiaridad de que la varianza debe de ser igual a uno para ser que la correspondencia entre definiciones sea univoca.[10] Q.E.D.

10. Comenzar con $z_i \,|\, \mathbf{x}_i \sim \mathcal{N}(z_i|\eta(\mathbf{x}_i), \sigma^2)$ con $\sigma^2 \neq 1$ deriva en que $p_i = \Phi(\eta(\mathbf{x}_i)/\sigma)$ lo cual es diferente a lo que se tiene en (6).

La clave en la prueba recae en que en la definición 0.2 del modelo probit se especifica a la función liga a Φ mientras que en la definición 0.3 se deriva de la normalidad de las variables latentes z_i. La razón principal para adoptar este enfoque es que en **albert1993bayesian** se desarrolla un método numérico vía simulación, bajo el paradigma bayesiano, para el cómputo exacto de las distribuciones posteriores del vector completo de de coeficientes de regresión $\boldsymbol{\beta}$ el cual resultaba atractivo para los objetivos del trabajo. Esta idea se refinará en el capítulo **??**; asimismo, es la razón para haber definido el modelo de regresión como se hizo.

Se hace notar que la ecuación (5) del GLM no influye en la prueba pues esta puede tener otras formas funcionales tan complejas como se requiera para la su aplicación especifica, ya sea lineal $\eta_i = \beta_0 + \boldsymbol{\beta}^t \mathbf{x}_i$ como en (1) o algo no lineal como se opta en este trabajo.

Liga entre la variable latente z y η

Para entender como se conectan las n variables latentes z_i con sus respectivos predictores lineales $\eta(\mathbf{x}_i)$, se necesita profundizar un poco más en el objetivo del modelo. Recapitulando, mediante la función liga Φ se une la media p_i, la probabilidad de que la respuesta y_i sea uno con los datos $\mathbf{x}_i$. Esto se logra, a través de una variable latente z_i definida con distribución normal cuya media $\eta(\mathbf{x}_i)$, la función de predicción,

es una transformación de las covariables $\mathbf{x}_i$. Es decir,

$$P(z_i > 0|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i) = p_i(\mathbf{x}_i) = \mathbb{E}[y_i|\mathbf{x}_i] = g^{-1}(\eta(\mathbf{x}_i)) = \Phi(\eta(\mathbf{x}_i)). \quad (10)$$

Este enfoque funciona, además de por el componente algorítmico, por la siguiente idea. Si se quiere crear una regla de decisión que clasifique observaciones en categorías binarias con base en cierta información, parecería intuitivo condensar esa información de forma que proporcione suficiente evidencia para inducir la clasificación. Traduciendo en términos matemáticos, la información $\mathbf{x}_i$ se condensa en la función $\eta(\mathbf{x}_i)$ la cual induce la clasificación de y_i a través de la cadena de identidades (10). Por ejemplo, si se tiene una $\eta(\mathbf{x}_i) >> 0$ para alguna observación i, implicaría que $P(z_i > 0|\mathbf{x}_i)$ es cercano a uno (por el dominio de Φ) y por lo tanto, es muy probable que y_i sea un uno. El argumento es idéntico para la probabilidad complementaria.

Al final, el modelo está resumiendo información al ir transformando espacios una función a la vez. El siguiente paso en el modelo consiste en detallar la transformación que debe realizar el predictor lineal $\eta(\mathbf{x}_i)$.[11] Tradicionalmente como se mencionó en (1) esta transformación era lineal tanto en parámetros como en covariables, dando lugar a fronteras de decisión lineales. Sin embargo, el siguiente paso lógico es modificar estos modelos para que las fronteras puedan ser más flexibles, rompiendo la linealidad en las covariables para lograr encontrar patrones más complejos.

11. En la literatura más enfocada en aprendizaje estadístico, es habitual de nombrar a la función $\eta(\mathbf{x}_i)$ como *proyector lineal* pues cumple la función de proyectar el espacio de las covariables a otro espacio donde sean linealmente separables, donde no necesariamente son de la misma dimensión, **bishop2006pattern**.

0.2. La función de predicción η

0.2.1. Una breve introducción a los GAM

Como se detalla en la página 6 de **james2013introduction**, conforme avanzaron los métodos y el poder computacional disponible se fueron desarrollando técnicas cada vez más poderosas que permitieron flexibilizar el supuesto de linealidad en covariables. En particular, **hastie1990generalized** se agrupan una clase de modelos a los que se les da el nombre de modelos aditivos generalizados (GAM). Estos modelos logran identificar relaciones no lineales utilizando, usualmente, métodos no paraméticos de suavizamiento en los datos adoptando así, un enfoque de *dejar que los datos hablen por si mismos.*[12] Fundamentalmente, lo que se realiza es transformar el espacio de covariables $\mathcal{X}$ por otro donde se puedan identificar patrones ocultos.

Definición 0.5. Un GAM tiene la forma $\forall i = 1, \ldots, n$:

$$\mathbb{E}[y_i|\mathbf{x}_i] = g^{-1}\left[f_0 + f_1(x_{i,1}) + \ldots + f_d(x_{i,d})\right], \tag{11}$$

con g^{-1} la inversa de la función liga definida en (0.1) y el predictor lineal $\eta(\mathbf{x}_i) = f_0 + f_1(x_{i,1}) + \ldots + f_d(x_{i,d})$.

La idea fundamental de los GAM, es asumir que los efectos en las covariables se pueden modelar como una suma de funciones por componentes, es decir, cada cova-

12. Página 1 de **hastie1990generalized**

riable $x_j \quad \forall j = 1, \ldots, d$ está siendo transformada de forma no lineal e independiente por una función asociada $f_j, \quad \forall j$. De esta forma, se preserva la interpretabilidad del modelo lineal pero se flexibiliza η. La funciones f_j que ahora componen el predictor lineal η se busca que sean tan flexibles como sea necesario, permitiendo que el estadístico pueda hacer menos suposiciones rígidas sobre los datos. Estas funciones f_j son suaves y no especificadas (*no paramétricas*), es decir, no tienen una forma funcional concreta y representable algebraicamente. Sin embargo, es justamente ahí donde recae la fuerza de los GAM: al dejar a las funciones f_j ser no especificadas, se permite que estas capturen los efectos necesarios en los datos para hacer la mejor asociación posible con y_i, a este proceso se le llama suavizamiento.

Un suavizador, se puede definir de forma general, como una herramienta que resume la tendencia de la respuesta y como función de las covariables $\mathbf{x}$ y produce un estimador f que es menos variable (ruidoso) que la respuesta en si. Como se mencionó con anterioridad, estos suavizadores son de naturaleza no paramétrica pues no se asume una dependencia rígida de y en $\mathbf{x}$.[13] Como su nombre lo indica, usualmente usualmente se busca que estos sean monótonas y *suficientemente suaves*, entendido como k veces diferenciables, sin embargo, no siempre es el caso. Como ejemplos prácticos de métodos no paramétricos, se encuentran los ajustes de medias móviles y el suavizamiento LOESS, (**cleveland1988locally**).[14]

13. Las técnicas no paraméticas están fuera del alcance de este trabajo. Sin embargo, vale la pena una mención especial por su funcionalidad, practicalidad y forma intuitiva, además del sinfín de aplicaciones que tienen. Una guía comprensiva de estas se encuentra en el libro **wasserman2007all**.

14. El suavizamiento LOESS, *locally estimated scatterplot smoothing*, es un tipo de regresión local que ajusta modelos más simples a subconjuntos de los datos para construir una función global que describa de forma no lineal la variabilidad intrínseca presentada.

En un GAM la estimación de las funciones f_j, se lleva a cabo tradicionalmente empleando el algoritmo de ajuste hacia atrás (*backfitting algorithm*), **hastie1986generalized**. Este procedimiento, busca dar estimadores de cada f_j de forma iterativa por componentes, utilizando como regresores los residuales parciales. Por ejemplo, sea $d = 2$ y $g^{-1}(w) = w$ la función identidad, quedando así el modelo:

$$\mathbb{E}[y_i|\mathbf{x}_i] = f_0 + f_1(x_1) + f_2(x_2).$$

Dados estimadores preliminares $\hat{f}_0$ y $\hat{f}_1$ de las respectivas funciones f_0 y f_1, se definen los residuales parciales: $\mathbb{E}[y_i|\mathbf{x}_i] - (\hat{f}_0 + \hat{f}_1(x_1))$ sobre los cuales se busca suavizar f_2. Este proceso resulta en una mejor estimación de la función f_2, con la cual, se puede mejorar el estimador de f_1. Ese proceso se lleva a cabo de forma iterativa, hasta que el cambio en las funciones funciones f_j sea menor que un umbral especificado.[15] Este algoritmo, se puede extender para d y g arbitrarias y es bastante flexible a modificaciones. En un GAM, las curvas resultantes de las funciones f_j son suaves y lejos de ser lineales. Asimismo, sus formas, pueden ayudar a entender el fenómeno subyacente.

Los GAM en el contexto de este trabajo

Sin dudar la elegancia y practicidad de los métodos no paramétricos, para este trabajo, se opta modificar el enfoque original de los GAM y darles una forma rígida a las funciones f_j, regresando a los dominios de la estadística semiparamétrica.

15. La demostración de convergencia de un GAM se encuentra en **stone1985additive**.

Esta decisión, pues se busca profundizar en los polinomios por partes estudiados en la siguiente sección 0.3 que componen a las funciones f_j. Aunque pareciera una desviación considerable del trabajo original de **hastie1990generalized**, en realidad en el apéndice **??** se detalla como los polinomios por partes son el resultado de plantear la idea de suavizamiento como un problema de optimización. Asimismo, los GAM son tan flexibles en su definición (y concepción) que es usual restringir las funciones f_j con formas funcionales concretas.[16]

Bajo esta óptica, para este trabajo se retienen dos de las ideas fundamentales de los GAM: aditividad y las transformaciones por componentes de las covariables. Es decir, la definición de un GAM (11) sustituye el predictor lineal tradicional de los GLM (1), $\eta(\mathbf{x}_i) = \beta_0 + \tilde{\boldsymbol{\beta}}^t \mathbf{x}_i$, por una suma de funciones $\sum_j^d f_j(x_j)$ más un intercepto constante f_0 (que juega el papel de β_0) dando lugar a la ecuación (12):

$$\eta(\mathbf{x}_i) = f_0 + f_1(x_{i,1}) + f_2(x_{i,2}) + \ldots + f_d(x_{i,d}). \tag{12}$$

Esta ecuación, será el predictor lineal usando en la implementación final del *bpwpm*.

Se hace notar, que a diferencia de los modelos lineales donde se tiene a los parámetros $\boldsymbol{\beta}$ incluidos en la expansión de η, en los GAM los parámetros se incluyen dentro de cada una de las f_j pues, los efectos de cada covariable son resumidos dentro de las mismas transformaciones. Aunque se pueden agregar parámetros que ponderen cada f_j sobre-parametrizar puede llevar a la incorrecta especificación del modelo y caer en problemas de identificabilidad de los parámetros.

16. Capitulo 9.1 y Ejemplo 5.2.2 de **hastie2008elements**

Al entender que cada f_j es una transformación no-lineal de x_j (como lo sería una transformación logarítmica o una transformación Box-Cox) se le regresa cierta interpretabilidad al modelo. Es decir, cada $f_j(x_{i,j})$ es el efecto que tiene la covariable j, para una observación i, en la clasificación. Por lo tanto y heredado de la identidad (10) si f_j es más positiva para esta observación i, se tiene mayor evidencia (en el componente j) de que la respuesta binaria asociada y_i sea uno. En la peculiaridad de que $d = 2$, se podrá visualizar, no solo las funciones f_j de manera independiente, sino toda $\eta(\mathbf{x}_i)$ en $\mathbb{R}^3$ como una serie de picos y valles donde será positiva en caso de que y_i sea clasificada como uno y negativa en caso de que sea cero. Esta propiedad se puede visualizar **??** en la imagen de la página **??** entre otras gráficas del trabajo.

La inclusión de un término independiente f_0 es importante en los GAM pues es uno de los resultados de la derivación mencionada en el apéndice **??**. Asimismo, se debe considerar el caso tal que $f_j(x_j) = 0 \quad \forall j$ de donde se necesita un término independiente f_0. Para este trabajo al término f_0 se le da el mismo tratamiento que el de un parámetro independiente convencional, por lo tanto, se estima usando el mismo procedimiento que todos los demás parámetros. Este hecho se esclarecerá en las secciones subsecuentes. Las imágenes **??** y **??** de la página **??**, son solo algunos ejemplos de las posibles formas finales que pueden adoptan las funciones f_j. Para esa realización particular del modelo, están compuestas por segmentos de recta que no son suaves.

0.3. Funciones f_j

Finalmente se trata la parte más profunda del modelo, las funciones f_j que, como se mencionó anteriormente, son transformaciones no lineales de cada componente x_j. Lo que buscan es suavizar la nube de datos, para posteriormente sumarlas entre si y dar una media η que resuma toda la información en un número real. Como se menciona en la introducción de **hardle2004semiparametric**, el suavizamiento de los datos es central en la estadística inferencial. La idea es extraer la señal entre el ruido y para ello, se intenta estimar y modelar la estructura subyacente. Este suavizamiento, se llevará a cabo usando una expansión en bases funcionales, particularmente el tipo de polinomios por partes presentados en **mallik1998automatic**. Toda la siguiente sección se concentra en describir las formas funcionales de las sub-funciones que componen a las funciones f_j y por ende a η.

0.3.1. Expansión en bases funcionales

Saliendo por un momento del domino de la estadística, se definen las expansiones en bases funcionales. Sin entrar mucho en los detalles técnicos, dado un espacio funcional[17] se puede representar cualquiera de sus elementos, en este caso una función arbitraria h, como la combinación lineal de los elementos de la base $\Psi_l : \mathbb{R}^d \to \mathbb{R}$ y constantes $\beta_l \in \mathbb{R}$, análogo al espacio vectorial canónico $\mathbb{R}^d$. En particular (y dados los objetivos del trabajo) se considera el espacio funcional que mapea $\mathbb{R}^d$ a $\mathbb{R}$,

17. Espacio vectorial cuyos elementos son funciones con una topología dada.

definiendo entonces la expansión en bases funcionales:

$$h(\mathbf{x}) = \sum_{l=1}^{N} \beta_l \Psi_l(\mathbf{x}) = \tilde{\boldsymbol{\beta}}^t \Psi(\mathbf{x}). \tag{13}$$

Bajo esta definición, $\boldsymbol{\Psi}(\mathbf{x}) = (\Psi_1(\mathbf{x}), \ldots, \Psi_N(\mathbf{x}))^t$ es un vector cuyos elementos $\Psi_l(\mathbf{x})$ son llamados funciones base y tienen el mismo mapeado que h. De la misma forma $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_N)^t$ es un vector de coeficientes constantes. Finamente, $N \in \mathbb{N}$ es un entero mayor o igual a la dimensión del espacio funcional que se maneja.[18]

En un contexto estadístico de regresión, se definen los modelos lineales de bases funcionales,[19] capitulo 3 de **bishop2006pattern**, como:

$$h(\mathbf{x}) = \beta_0 + \sum_{l=1}^{N} \beta_l \Psi_l(\mathbf{x}) = \beta_0 + \tilde{\boldsymbol{\beta}}^t \boldsymbol{\Psi}(\mathbf{x}), \tag{14}$$

lo cual es idéntico a (13) con la adición del término independiente β_0. Para los objetivos del modelo *bpwpm*, se busca representar una transformación de la media condicional de la respuesta por una función dependiente de los datos, es decir: $h(\mathbf{x})$ es equivalente a $g(\mathbb{E}[y\,|\mathbf{x}]) = \eta(\mathbf{x})$. Por lo tanto se puede pensar que esta función h es análoga a la función de predicción η, que también puede ser expresada como su expansión en bases funcionales.[20]

18. Dependiendo de el espacio funcional y la complejidad de la función real por estimar h, en ocasiones se requiere que $N = \infty$ para que se de la igualdad estricta (**bergstrom1985estimation**).

19. *Linear basis function models.*

20. Un supuesto fuerte pero útil pues la verdadera η puede no ser expresada como una suma de bases Ψ.

La idea, es que se remplace (o se aumente) la cantidad de covariables $\mathbf{x}$ con transformaciones de estas, capturadas en el vector $\Psi(\mathbf{x})$ de igual o mayor dimensión. Como ejemplos ilustrativos de ψ se tiene en la literatura:

$\Psi_l(\mathbf{x}) = x_l \quad \forall\, l = 1, \ldots, N = d$, recuperan de un GLM tradicional.

$\Psi_l(\mathbf{x}) = \ln x_l$ ó $x_l^{1/2}$ para alguna $l = 1, \ldots, N = d$, donde se tienen transformaciones no lineales en cada una (o algunas) de las covariables.

$\Psi_l(\mathbf{x}) = \exp\left\{-\dfrac{(x_l - \mu_l)^2}{2s^2}\right\} \quad l = 1, \ldots, d$ una expansión en bases gaussianas con μ_j el parámetro que gobierna la ubicación y s la escala de las funciones bases.

$\Psi_l(\mathbf{x}) = x_j^a I(\tau_b \leq x_j < \tau_c)$ para alguna j y $\forall l = 1, \ldots, N$ con $a \in \mathbb{N}$ y τ_b, τ_c nodos fijos. Dando lugar a una expansión en bases polinómicas como la que se usa en este trabajo (sección 0.3.2).

$\Psi_l(\mathbf{x}) = x_j x_k \quad \forall l = 1, \ldots, N$, para alguna j, k. Dando lugar a un modelo con interacciones.

Como se ve, esta representación es tan flexible que engloba muchos de los modelos y transformaciones posibles en el mundo de las regresiones, uniendo temas de análisis funcional con estadística aplicada. Asimismo esta representación ha resultado ser de gran utilidad en casos donde los patrones entre las covariables son no lineales. Se hace notar que el último ejemplo rompe con la aditividad inherente de las covariables

en los modelos que se han estudiado hasta ahora, mostrando que esta generalización no está restringida a ser completamente aditiva en covariables. Sin embargo h, por su construcción, siempre es lineal en los parámetros $\boldsymbol{\beta}$ pero, usualmente, no lineal en las covariables, dependiendo de la forma de $\Psi(\mathbf{x})$.

Dependiendo del tipo de datos y el propósito del modelo, puede ser conveniente usar algún tipo de funciones base sobre otras. Sin embargo, sobre todo cuando se tiene poca o ninguna experiencia con los datos, se busca una representación flexible (por no decirla ingenua) de éstos. El método más común es tomar una familia grande de funciones que logre representar una gran variedad de patrones. No obstante, una desventaja de estos métodos es que al contar con una cantidad muy grande de funciones base y por ende parámetros, se requiere controlar la complejidad del modelo para evitar el *sobre-ajuste*.[21] Algunos de los métodos más comunes para lograrlo son los siguientes, **hastie2008elements**:

> Métodos de restricción: se selecciona un conjunto finito de funciones base y su tipo, limitando así las posibles expansiones. Los modelos aditivos como los usados en este trabajo, son un ejemplo de esto.
>
> Métodos de selección de variables: como lo son los modelos CART y MARS,[22] donde se explora de forma iterativa las funciones base y se incluyen aquellas que contribuyan a la regresión de forma significativa.

21. Seguir los datos tan de cerca que se pierda la señal entre el ruido.
22. *Classification & regression tree* (**breiman1984classification**) y *multivariate adaptive regression splines* (**friedman1991multivariate**) respectivamente.

Métodos de regularización: donde se busca controlar la magnitud los coeficientes, buscando que la mayoría de ellos sean cero, como los son los modelos *Ridge* y *LASSO* entre otros.[23]

Para los objetivos de este trabajo, lo que se busca expresar en su expansión de bases funcionales no es la función de predicción η completa, sino cada uno de sus componentes aditivos f_j. Al aislar cada función f_j que dependen únicamente de una variable real $x_j \quad \forall j$, es decir: $f_j : \mathbb{R} \to \mathbb{R} \quad \forall j$, se puede simplificar la exposición y reducir el número de índices pues sus expresiones algebraicas son idénticas.

0.3.2. Polinomios por partes y *splines*

Los polinomios por partes, por su flexibilidad, ha resultado ser de gran utilidad en diversas ramas de las matemáticas. En particular, el mundo de la estadística surgen de forma natural como solución a varios problemas de modelado (ver apéndice **??**). No obstante, antes de exponer la representación final de las funciones f_j, se da una exposición constructiva de los polinomios por partes. Se usa como referencia las primeras dos secciones de el Capitulo 5 de **hastie2008elements** y **wahba1990splines**.

Sea $x \in [a, b] \subseteq \mathbb{R}$, se busca separar $[a, b]$ en J intervalos. Por lo tanto, se define una partición correspondiente $\mathcal{P} = \{\tau_1, \ldots, \tau_{J-1}\}$ tal que $a \leq \tau_1 < \ldots < \tau_{J-1} \leq b$. Las constantes τ son llamadas *nodos*.[24] Con los nodos seleccionados, se puede representar

23. *Least absolute shrinkage and selection operator* (**hoerl1970ridge**; **tibshirani1996regression**)

24. En la definición, se puede incluir o no la frontera dependiendo de si se busca hacer inferencia

a diferentes niveles de precisión una función arbitraria h, a través de su expansión análoga a la ecuación (13), donde cada Ψ_j será una función que depende, tanto de la partición $\mathcal{P}$ como de la variable real x.

Para ilustrar se presenta un ejemplo sencillo. Primero, se parte el intervalo en tres pedazos ($J = 3$) definiendo una partición con dos nodos, es decir: $\mathcal{P} = \{\tau_1, \tau_2\}$. Posteriormente, a cada subintervalo se le asocia una función Ψ_j tales que:

$$\begin{aligned}
\Psi_1(x, \mathcal{P}) &= I(x < \tau_1) \\
\Psi_2(x, \mathcal{P}) &= I(\tau_1 \leq x < \tau_2) \\
\Psi_3(x, \mathcal{P}) &= I(\tau_2 \leq x),
\end{aligned}$$

con $I(\cdot)$ la función indicadora que vale uno si x se encuentra en la región y cero en otro caso. Con esta definición, se construye una función por partes h:

$$\begin{aligned}
h(x) &= \sum_{l=1}^{J} \beta_l \Psi_l(x) \\
&= \beta_1 I(x < \tau_1) + \beta_2 I(\tau_1 \leq x < \tau_2) + \beta_3 I(\tau_2 \leq x).
\end{aligned}$$

Esta función h es una función escalonada, en el sentido de que cada región de x tiene un nivel β_j, Se hace notar que esta definición transforma el espacio original $x \in [a, b] \subseteq \mathbb{R}$ de una dimensión a uno de tres dimensiones. Esta idea de aumento artificial en la dimensionalidad a lo largo del rango del intervalo es una idea fundamental para

fuera del intervalo acotado de los datos.

este trabajo[25]

Con este ejemplo, al partir el intervalo (aumentando la dimensionalidad) y construir funciones más sencillas sobre ellos, se ilustra a grandes rasgos como funcionan los polinomios por partes. Sin embargo, estos pueden ser mucho más flexibles pues a cada intervalo se puede ajustar un polinomio de grado arbitrario $(M-1)$.[26] Adicionalmente, se puede añadir restricciones de continuidad en los nodos y no sólo continuidad entre los polinomios, sino continuidad en las derivadas. Esta es la flexibilidad de los polinomios por partes, que se les puede pedir cuanta *suavidad*, o no, se requiera, entendida como la continuidad de la $(\tilde{K})$-ésima derivada.

Número total de funciones bases N

Para formalizar la idea anterior, al tomar una expansión de bases para cada intervalo, el número de funciones base aumenta en J por cada grado que se agregue, dando un total de $J \times M$ bases funcionales. Esto ocurre porque se necesita definir una base de tamaño M para cada subintervalo $j = 1, \ldots, J$, es decir, $\mathcal{B} = \left\{1, x, x^2, \ldots, x^{M-1}\right\}$ con $\mathcal{B}$ la base. Esta definición, deriva en polinomios que se comportan de forma independiente en cada intervalo y no se conecten. Naturalmente, la primera condición en la que se piensa, es imponer continuidad en los nodos lo cual devuelve $(J-1)$

25. Bajo un contexto de regresión, dado un conjunto de observaciones $\{(y_i, x_i)\}_{i=1}^n$, si se buscara estimar los parámetros $\boldsymbol{\beta}$ usando una función de perdida cuadrática, se puede demostrar que cada $\hat{\beta}_j = \bar{y}_j$, es decir, para cada región el mejor estimador constante, es el promedio de los puntos de esa región.

26. Se usa esta convención pues, para representar un polinomio de grado $M-1$ se necesitan M elementos en la base.

parámetros que corresponden a los $(J-1)$ nodos. De la misma forma, cada grado de continuidad en las derivadas que se le pida al polinomio, lo restringe y por ende, devuelve el mismo número de funciones bases, se denota por $\tilde{K}$ este número. Sin embargo, es más intuitivo pensar en un parámetro $K = \tilde{K} + 1$ como el número de restricciones que se imponen en los nodos. Es decir, $K = 0$ implica intervalos independientes, $K = 1$, implica que los polinomios se conectan, $K = 2$ implica continuidad en la primera derivada ($\tilde{K} = 1$) y así sucesivamente. Bajo esta definición los polinomios por partes tienen un total de:

$$N(M, J, K) = JM - K(J-1) \tag{15}$$

bases funcionales y por ende, el mismo número de parámetros $\boldsymbol{\beta}$ por estimar. Dada la construcción y las características de M, J y K se derivan de forma trivial las restricciones para estos parámetros: $M > K \geq 0$ y $J > 1$.

La palabra *spline* usualmente se usa para designar a un grupo particular de polinomios por parte. Sin embargo, no hay consenso en la literatura de su definición exacta. Para este trabajo se usa la definición de **wasserman2007all** y **hastie2008elements**. Un *spline de grado M* es un polinomio por partes de grado $M-1$ y continuidad hasta la $(M-2)$-derivada, es decir, se impone la restricción adicional $K = M - 1$. Se hace notar, que existen muchos tipos de *splines*, como lo son los B-Splines. Dependiendo de la aplicación, se pueden construir más o menos flexibles o más rápidos en su implementación computacional. En **deboor1978splines** y más recientemente **wahba1990splines** se hacen tratados extensivos sobre ellos y sus generalizaciones. Los *splines* cúbicos se han popularizado en la literatura, pues resultan en curvas

suaves al ojo humano, reteniendo suficiente flexibilidad para aproximar una gran cantidad de funciones.

Polinomios por parte flexibles

Habiendo definido M, J y K y por ende el número de funciones bases N, finalmente se le puede dar forma funcional a la familia de funciones base Ψ que se usan en este trabajo, legado del trabajo **mallik1998automatic**. Se define primero, la función auxiliar *parte positiva* para poder escribir polinomios por partes en una sola línea. Sea $a \in \mathbb{R}$ entonces:

$$a_+ = \text{máx}\left\{0, a\right\},$$

simplificando la notación.

Definición 0.6. Expansión en bases truncada, **mallik1998automatic**:

$$h(x) = \sum_{l=1}^{N} \beta_l \ \Psi_l(x, \mathcal{P}) = \tilde{\boldsymbol{\beta}}^t \boldsymbol{\Psi}(x, \mathcal{P}) \tag{16}$$

$$\text{donde, } \ N = JM - K(J-1)$$

$$= \underbrace{\sum_{\hat{\imath}=0}^{M-1} \beta_{\hat{\imath},0} \ x^{\hat{\imath}}}_{\text{polinomio base}} + \underbrace{\sum_{\hat{\imath}=K}^{M-1} \sum_{\hat{\jmath}=1}^{J-1} \beta_{\hat{\imath},\hat{\jmath}} \ (x - \tau_{\hat{\jmath}})_+^{\hat{\imath}}}_{\text{parte truncada}} \tag{17}$$

Esta expansión en bases, que representa un polinomio por partes, es prácticamente la

expansión de bases implementada en el modelo final. Se hace notar, que la flexibilidad viene derivada de las múltiples posibles elecciones para M, J y K.

Al primer sumando de (17) se le conoce como polinomio base (*baseline polynomial*), pues afecta a todo el intervalo de definición $[a, b]$ al no estar afectado por los nodos. El segundo sumando, conocido como la parte truncada, controla la suavidad entre los nodos. Es decir, por cada nodo $\hat{\jmath} = 1, \ldots, J-1$ se tienen $M - K$ funciones parte positivas $(\cdot)_+$ que se activan (se vuelven positivas) a medida que x recorre el su dominio $[a, b]$ hacia la derecha y va pasando por los nodos τ_j. Estas funciones parte positiva, van capturando los efectos de los intervalos anteriores que, al combinarlos con el primer sumando definen un polinomio de grado $M - 1$ en todo el intervalo.[27] La principal utilidad de esta expansión, es que engloba todas las ideas antes mencionadas en tres parámetros: M, J y K, al escogerlos, se pueden representar un gran número polinomios por partes. Por ejemplo, si $M = 3$, $J = 5$ y $K = 0$ se tiene un polinomio por partes en 5 subintervalos (4 nodos) donde cada subintervalo es una parábola independiente de la anterior, es decir, las parábolas no son continuas entre si. Por el contrario, si $K = M - 1$ se devuelve a la definición de *splines*, o por último, si $M = 1$, $J = 3$ y $K = 0$, se tienen constantes por segmentos como en el ejemplo introductorio de los polinomios por parte.

Para la facilitar la interpretación de los parámetros y la expansión de (17), los parámetros β cuentan dos índices: $\hat{\imath}$ y $\hat{\jmath}$. El índice $\hat{\imath}$ siempre estará asociado al grado de su función base asociada, es decir, si $\hat{\imath} = 2$ se está hablando de un término de

27. Esta expansión, surge de integrar un polinomio por partes, constante en cada subintervalo, $M - 1$ veces, pues las constantes de integración se pueden agrupar en el polinomio base.

grado 2. En el segundo sumando (la parte truncada) el índice $\hat{\imath}$ comienza en K para codificar las restricciones de continuidad.[28] El segundo índice $\hat{\jmath} = 1, \ldots, J-1$ describe el nodo al que está asociado el parámetro. Como convención, si $\hat{\jmath} = 0$, se hace referencia al primer sumando (el polinomio base) que siempre está activo sobre el intervalo.

La ecuación (16) es una expansión en bases arbitrarias análoga a definición de (13). Sin embargo, que en (16) se hace referencia a β con un solo índice $l = 1, \ldots, N^*$ mientras que en (17) con dos. Esta disparidad surge de la necesidad de una doble interpretación de la expresión; como una expansión de bases arbitrarias Ψ_l y su correspondiente expansión en bases truncadas. Sin embargo, existe una biyección notacional univoca entre los elementos β_l, $\beta_{i,j}$ y Ψ_l presentada en la tabla 1 de la página 29. Esta tabla ayuda no sólo a esclarecer la notación, sino a expresar los polinomios de forma matricial que posteriormente se implementará en el algoritmo.

Los nodos τ: el trabajo de mallik1998automatic

Las ideas de **mallik1998automatic**, van más allá de la ecuación (17). En su trabajo, los autores presentan un método automático bayesiano para estimar relaciones funcionales complejas. En su trabajo original plantean el problema para un conjunto

28. Esta codificación es sutil pues, al hacer la demostración de continuidad, hay que considerar los límites izquierdos y derechos. Los límites izquierdos siempre coinciden con la función en el nodo. Sin embargo, los términos $(x-\tau)_+^K$ se desvanecen únicamente hasta la (K)-ésima derivada. Para la $(K+1)$-derivada, el coeficiente correspondiente se suma a la función y rompe la continuidad pues no corresponde el límite derecho.

β_l	$\beta_{\hat{\imath},\hat{\jmath}}$	$\Psi_l(x,\mathcal{P})$		
Subíndice l	Subíndices $\hat{\imath},\hat{\jmath}$	Función Base		
1	$0,0$	1	M elementos	
2	$1,0$	x		
$\vdots$	$\vdots$	$\vdots$		
M	$M-1,0$	x^{M-1}		
$M+1$	$K,1$	$(x-\tau_1)_+^K$	$M-K$	$J-1$ veces
$M+2$	$K+1,1$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(M-K)$	$M-1,1$	$(x-\tau_2)_+^{M-1}$		
$M+(M-K)+1$	$K,2$	$(x-\tau_2)_+^K$	$M-K$	
$M+(M-K)+2$	$K+1,2$	$(x-\tau_2)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+2(M-K)$	$M-1,2$	$(x-\tau_2)_+^{M-1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-2)(M-K)+1$	$K,J-1$	$(x-\tau_{J-1})_+^K$	$M-K$	
$M+(J-2)(M-K)+2$	$K+1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-1)(M-K)$	$M-1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		

Tabla 1: Biyección notacional entre β_l, $\beta_{i,j}$ y sus correspondientes funciones base Ψ_l.

Se tiene un total de $N = M + (J-1)(M-K) = JM - K(J-1)$ términos, ecuación (15). Por construcción, se es consistente con la definición de *spline* si $K = M-1$.

de datos $\{(y_i, x_i)\}_{i=1}^{n}$ donde buscaban ajustar una curva tal que $\mathbb{E}[y \mid x] = h(x)$, de forma análoga:

$$y_i = h(x_i) + e_i \quad i = 1, \ldots, n \tag{18}$$

donde e_i son variables aleatorias con media cero ($\mathbb{E}[e_i] = 0 \ \forall i$). Se observa como bajo este contexto, h es análoga a la η definida con anterioridad.

Para lograrlo, utilizan el polinomio definido en (17) y desarrollan un procedimiento bayesiano para la estimación de los nodos τ que son tradicionalmente fijos. Este procedimiento permite modelar a la vez J aumentando o disminuyendo la cantidad de nodos, desarrollando un algoritmo de muestreo Gibbs trans-dimensional, es decir, el algoritmo cambia el número de parámetros en cada iteración. Esta generalización, logra estimaciones robustas que logran aproximar funciones continuas *casi en todas partes* como lo son la función Doppler, funciones por bloques y funciones con picos pronunciados. Con lo anterior, los autores ilustran que el supuesto de suavidad en h, aunque útil, no siempre es necesaria. Muchas funciones discontinuas no se podrían estimar del todo usando polinomios continuos como los *splines*. Al final, todo depende de la *rugosidad* de los datos y el propósito del modelo.

La ventaja de que nodos sean parámetros por estimar, es que se estos pueden concentrar en los lugares donde la función varia más. Al contrario, si la función es relativamente suave para algún intervalo se utilizan pocos nodos. Sin embargo y para propósitos de este trabajo, los nodos se toman determinados desde el principio. Su número $(J - 1)$ es definido por el estadístico y su localización se escoge en los

cuantiles del rango de las covariables.[29] En el capítulo **??** se detalla como la simplificación de no incorporar los nodos como parámetros ayuda bastante a la velocidad del algoritmo. Posteriormente en el capítulo **??**, se observará que para fines prácticos, el modelo funciona muy bien y finalmente en el capítulo **??** se discute que habría cambiado de haberse implementado.

0.4. Primer vistazo al modelo *bpwpm*

Después de esta extensa discusión teórica, finalmente se está en posición de sintetizar muchas de las ideas construidas con anterioridad y dar de forma preliminar una visión general del modelo.

Se supone lo siguiente: $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ es el conjunto de datos observados independientes, con n el tamaño de la muestra, donde $y_i \in \{0, 1\}$ son las variables de respuesta binarias, $\mathbf{x}_i \in \mathcal{X}^d \subseteq \mathbb{R}^d$ las covariables o regresores y $d \in \mathbb{N}$ la dimensionalidad de las covariables.[30] Estos datos se organizan y se representan en una tabla (o matriz) como la presentada en la tabla 2. En ella, cada fila $i = 1, \ldots, n$ representa una observación. La primer columna corresponde al vector de respuestas y las columnas subsecuentes $j = 1, \ldots, d$ representan una covariable. Es útil pensar en estas columnas como d *dimensiones* que contienen información que induce la clasificación binaria en y_i.

29. Es decir, si se tiene J intervalos, se toman los nodos como los cuantiles que acumulan probabilidad $1/J$ en el rango $[a, b]$.

30. En el lenguaje de aprendizaje de máquina, es usual hablar de *outputs* e *inputs o features* para referirse a y_i y $\mathbf{x}_i$ respectivamente (**alpaydin2014introduction**).

$$\left[\begin{array}{c|c} y_1 & \mathbf{x}_1 \\ \vdots & \vdots \\ y_n & \mathbf{x}_n \end{array}\right] = \left[\begin{array}{c|ccc} y_1 & x_{1,1} & \dots & x_{1,d} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n,1} & \dots & x_{n,d} \end{array}\right]$$

Tabla 2: Estructura asumida en los datos

Asimismo, se define el espacio de covariables $\mathcal{X}^d$ como el producto cartesiano de los rangos de cada covariable j. Esta definición, está relacionada con los polinomios por partes f_j estudiados en la sección 0.3. Para el modelo, se supone que $\mathcal{X}^d$ es cerrado y acotado de la forma:

$$\begin{aligned} \mathcal{X}^d &= \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \\ &= [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d] \subseteq \mathbb{R}^d \\ \text{con} \qquad a_j &= \text{mín}\left\{x_{1,j}, \ldots, x_{n,j}\right\} \\ b_j &= \text{máx}\left\{x_{1,j}, \ldots, x_{n,j}\right\} \end{aligned}$$

para todo $j = 1, \ldots, d$.

Definición 0.7. El modelo *bpwpm* (preliminar). Para cada observación $i = 1, \ldots, n$:

$$y_i = \begin{cases} 1 & \iff \quad z_i > 0 \\ 0 & \iff \quad z_i \leq 0 \end{cases} \tag{7}$$

$$z_i \,|\, \mathbf{x}_i \sim \mathcal{N}(z_i \,|\, \eta(\mathbf{x}_i), 1) \tag{8}$$

$$\eta(\mathbf{x}_i) = f_0 + f_1(x_{i,1}) + f_2(x_{i,2}) + \ldots + f_d(x_{i,d}) \tag{12}$$

$$f_j(x_{i,j}) = \sum_{l=1}^{N^*} \beta_{j,l} \Psi_l(x_{i,j}, \mathcal{P}_j) \quad \forall j = 1, \ldots, d \tag{19}$$

donde, $\eta(\mathbf{x}_i)$ es un predictor no lineal que mapea $\mathcal{X}^d$ a $\mathbb{R}$, N^* es el número total de funciones base $\Psi(\cdot)_l$, $\beta_{j,l}$ los parámetros de modelo y $\mathcal{P}_j$ una partición en nodos del espacio de covariables.

Esta definición, rescata el las identidades (7) y (8) modelo probit (definición 0.3), haciendo la liga de las covariables reales $\mathbf{x}_i$ con la variable de respuesta binaria y_i a través de las variables latentes z_i. Asimismo, recupera las ideas de los GAM sintetizadas la ecuación (12), especificando la media de las variables latentes z_i, al darle forma funcional aditiva a $\mathbb{E}[z_i|\mathbf{x}_i] = \eta(\mathbf{x}_i)$.

Por último, esta definición introduce la identidad (19) definiendo a cada una de las funciones f_j $\forall j$ en la parte más profunda del modelo. Estas funciones $f_j : \mathcal{X}_j = [a_j, b_j] \to \mathbb{R}$, como se estudió en la sección 0.3, realizan una transformación no lineal de las covariables $x_{i,j}$ mediante una expansión en bases funcionales. El objetivo de

esta expansión es expresar cada f_j de una forma flexible, a través de la suma ponderada de funciones bases $\Psi_{j,l}(x_{i,j}, \mathcal{P}_j)$ y parámetros desconocidos $\beta_{j,l}$ los cuales se deben de estimar. Asimismo, las funciones bases dependen de tres componentes: las covariables $x_{i,j}$, una partición $\mathcal{P}_j$ para cada intervalo $\mathcal{X}_j = [a_j, b_j] \quad \forall j = 1, \ldots, d$ y el número total de funciones base $N^* \in \mathbb{N}$. Sus formas funcionales, no son más que truncamientos de orden mayor en las covariables, por ejemplo: $(x_{i,j} - a)_+^b$ con a, b constantes definidas por N^* y $(\cdot)_+$ la función parte positiva, dando lugar a la expansión en polinomios por partes similar a la presentada en **mallik1998automatic**.

No obstante, bajo las definiciones anteriores aún se tiene un problema de confusión en los parámetros. Al ya tener un término independiente $f_0 = \beta_0$ en (21), para preservar la identificabilidad de los parámetros se deben realizar unas pequeñas modificaciones a (17). Los parámetros confundidos pueden tener dos orígenes. Primero, si se permite que $K = 0$ (polinomios discontinuos) el segundo sumando tendría términos independientes no deseados. Esto se arregla fácilmente imponiendo la restricción de continuidad en los polinomios, es decir, $K > 0$.[31] Segundo, se debe retirar el término independiente inherente en el polinomio base, es decir, comenzar el primer sumando de (17) en uno en vez de cero. Esta modificación retira una función base modificando N y convirtiendola en la N^* que se observa en (19).[32] Juntando estas

31. De manera preeliminar, se implementó una versión del algoritmo que permitía esta confusión. El ajuste no mejoraba cuando $K = 0$ y solamente causaba que las cadenas simuladas de los parámetros no convergieran debidamente. Sin embargo, en los polinomios resultantes si se observaba la discontinuidad.

32. **mallik1998automatic** resuelven este problema de identificabilidad al solamente usar una covariable para la estimación de las curvas, permitiendo retirar uno de los parámetros independientes sin penalización. Asimismo, su algoritmo automático trans-dimensional les permitía tener polinomios por partes discontinuos.

cambios (17) se redefine como:

$$
\begin{aligned}
h(x) &= \sum_{l=1}^{N^*} \beta_l \; \Psi_l(x, \mathcal{P}) \\
\text{con } \; N^* &= J \times M - K(J-1) - 1 \\
\text{donde: } \; M &> K > 0 \text{ y } J > 1 \\
&= \sum_{\hat{\imath}=1}^{M-1} \beta_{\hat{\imath},0} \; x^{\hat{\imath}} + \sum_{\hat{\imath}=K}^{M-1} \sum_{\hat{\jmath}=1}^{J-1} \beta_{\hat{\imath},\hat{\jmath}} \; (x - \tau_{\hat{\jmath}})_+^{\hat{\imath}}. \qquad (20)
\end{aligned}
$$

Lo cual, es finalmente la expansión que se implementa en el modelo *bpwpm*. Solamente basta igualar $h(x)$ a $f_j(x_j)$ para toda $j = 1, \ldots, d$ y se termina por definir a la ecuación canónica (19).

Para esclarecer un poco más el trabajo, en la figura 2 se presenta un diagrama del modelo y sus componentes. De izquierda a derecha y para toda $i = 1, \ldots, n$: se busca transformar de manera no lineal a cada una de las covariables observadas $x_{i,j} \quad \forall j = 1, \ldots, d$ a través polinomios por partes condensados en las funciones f_j. Estas transformaciones dependen de parámetros desconocidos $\beta_{j,l}$ con $l = 1, \ldots, N^*$ y la partición de cada dimensión $\mathcal{P}_j$. Una vez se tienen las covariables transformados, se suman las funciones f_j con un intercepto local f_0 para obtener una función de predicción η. Esta función actúa como la media de la variable latente z_i que relaciona a la respuesta y_i con $\mathbf{x}_i$. La relación se realiza a través de la función Φ para lograr la clasificación binaria en y_i.

Las aparentemente complejas interacciones entre todos los componentes del modelo

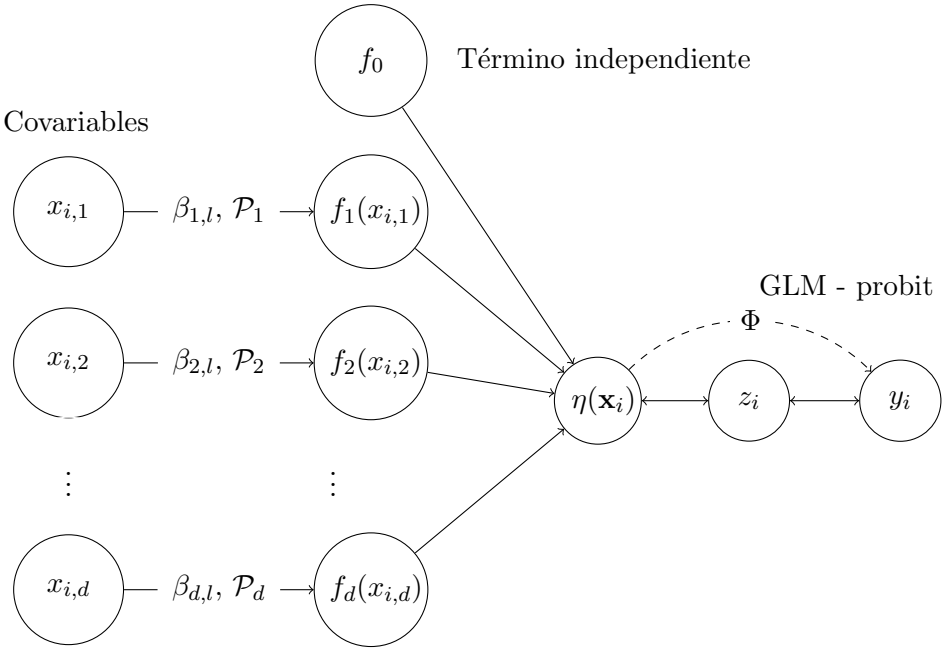

Figura 2: Diagrama del modelo *bpwpm*

no son más que respuestas estructurales a un proceso de *síntesis* de la información. El modelo está buscando identificar patrones en las covariables $\mathbf{x}_i$ para la clasificación de su respuesta binaria asociada y_i. Este proceso, se lleva a cabo mediante tres transformaciones $f_j(x_{i,j})\ \forall j$, $\eta(\mathbf{x}_i)$ y finalmente $\Phi(\eta(\mathbf{x}_i))$ las cuales cumplen el propósito de ir colapsando dimensiones. Se espera que este proceso, logre separar de forma flexible el espacio d-dimensional $\mathcal{X}^d$ a regiones más identificables (para la clasificación) que las regiones originales; donde finalmente, se le asigne una probabilidad a cada región de clasificación mediante Φ. El Capítulo **??** cuenta con visualizaciones que esperan aterrizar estos conceptos teóricos en algo más concreto. No sin antes

especificar por completo el resto del modelo en los siguientes capítulos.

0.4.1. Consideraciones matemáticas adicionales

Bajo la óptica de la implementación del modelo, se hace énfasis en la linealidad de los parámetros más no de las covariables. Al sustituir (19) en (12) este hecho se hace aún más evidente:

$$\begin{aligned}
\eta(\mathbf{x}_i) &= f_0 + \sum_{j=1}^{d} f_j(x_{i,j}) \\
&= f_0 + \sum_{j=1}^{d} \boldsymbol{\beta}_j^t \boldsymbol{\Psi}(\mathbf{x}_i, \mathcal{P}_j) \\
&= f_0 + \sum_{j=1}^{d} \left[\sum_{l=1}^{N^*} \beta_{j,l} \Psi_l(x_{i,j}, \mathcal{P}_j) \right] \qquad (21) \\
&= \boldsymbol{\beta}^t \widetilde{\boldsymbol{\psi}}_i(\mathbf{x}_i). \qquad (22)
\end{aligned}$$

En donde cada sumando interior de (21) tiene una expansión de bases funcionales definida por la ecuación (20).[33] Esta representación, permite visualizar la linealidad en parámetros del modelo, asimismo, la función f_0, al ser constante puede ser interpretada como otro parámetro adicional, es decir: $f_0 \equiv \beta_0$. La linealidad en los parámetros de la función de predicción η, derivan en que (21) pueda ser re-expresada simplemente como el producto punto de un largo vector de parámetros $\boldsymbol{\beta} \in \mathbb{R}^\lambda$ y un vector nombrado $\widetilde{\boldsymbol{\psi}}_i(\mathbf{x}_i)$, (ecuación (22)), que representa un renglón de la ma-

33. No se hace la sustitución pues la notación resultante es innecesariamente compleja.

triz de diseño $\widetilde{\Psi}$ de mayor dimensión ($d < N^*$) que incorpora la doble suma de las correspondientes expansiones en bases y todas las observaciones $i = 1, \ldots, n$.

Bajo estas observaciones, el predictor lineal η se puede re-expresar en su forma vectorial compacta $\boldsymbol{\eta}$, como:

$$\boldsymbol{\eta}(\mathbf{X}) = \widetilde{\Psi}(\mathbf{X})\boldsymbol{\beta}, \tag{23}$$

donde $\mathbf{X} \in \mathbb{R}^{n \times d}$ es la matriz de covariables, $\boldsymbol{\beta}$ el vector de parámetros con un total de $\lambda = 1 + d \times N^*$ elementos y $\widetilde{\Psi}(\mathbf{X}) \in \mathbb{R}^{n \times \lambda}$ la transformación no lineal definida con anterioridad. Vistos en sus correspondientes formas matriciales, las estructuras tienen las siguientes formas:

$$\boldsymbol{\eta}(\mathbf{X}) = \begin{bmatrix} \eta(\mathbf{x}_1) \\ \eta(\mathbf{x}_2) \\ \vdots \\ \eta(\mathbf{x}_n) \end{bmatrix} \qquad \boldsymbol{\beta} = \left\{ \begin{array}{l} \left[\begin{array}{c} \beta_0 \\ \hdashline \beta_1 \\ \vdots \\ \beta_{N^*} \end{array} \right] \begin{array}{l} \text{término independiente} \\ \\ \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \boldsymbol{\beta}_1 : \ N^* \text{ términos} \end{array} \\ \left[\begin{array}{c} \beta_{N^*+1} \\ \vdots \\ \beta_{2N^*} \end{array} \right] \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \boldsymbol{\beta}_2 : \ N^* \text{ términos} \\ \hdashline \vdots \\ \beta_{d \times N^*} \end{array} \right\} d \text{ veces}$$

$$\widetilde{\Psi}(\mathbf{X}) = \left[\begin{array}{c:c:c:c} 1 & f_1(x_{1,1}) & \dots & f_d(x_{1,d}) \\ \vdots & \vdots & & \vdots \\ 1 & f_1(x_{n,1}) & \dots & f_d(x_{n,d}) \end{array}\right] \tag{24}$$

$$= \left[\begin{array}{c:c:c:c:c:c:c} 1 & \Psi_1(x_{1,1},\mathcal{P}_1) & \dots & \Psi_{N^*}(x_{1,1},\mathcal{P}_1) & \dots & \Psi_1(x_{1,d},\mathcal{P}_d) & \dots & \Psi_{N^*}(x_{1,d},\mathcal{P}_d) \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 1 & \Psi_1(x_{n,1},\mathcal{P}_1) & \dots & \Psi_{N^*}(x_{n,1},\mathcal{P}_1) & \dots & \Psi_1(x_{n,d},\mathcal{P}_d) & \dots & \Psi_{N^*}(x_{n,d},\mathcal{P}_d) \end{array}\right]$$

Bajo esta definición, el modelo se simplifica y se observa que en realidad cada f_j es una expansión de cada covariable x_j en más términos que se le añaden al predictor lineal, idea fundamental del modelo. Es decir, el espacio original de covariables $\mathcal{X}^d$ de dimensionalidad d, $\mathcal{X}^d$ se transforma de forma no lineal en otro espacio $\widetilde{\Psi}(\mathcal{X})^\lambda$ de dimensionalidad λ donde se pueden manifestar patrones ocultos que en este nuevo espacio son separables linealmente. Asimismo y desde la perspectiva de un GAM, el espacio de covariables original se puede transformar en uno de igual dimensión representado por las *nuevas covaraibles* $f_j(x_j)$ el cual es más sencillo de separar. En la particularidad en que $d = 2$, esta perspectiva tiene la ventaja que se puede seguir visualizando, para ilustrar, se sugiere contrastar las imágenes **??** con **??** y **??** con **??** en el capítulo **??**.

A pesar de la utilidad de estos polinomios por parte, todos sufren de problemas más allá del rango de definición $\mathcal{X}_j = [a_j, b_j] \quad \forall j = 1, \dots, d$. Pues, su naturaleza global hace que fuera de la región con nodos los polinomios crezcan o decrezcan rápidamente. Por lo tanto, extrapolar con polinomios es peligroso y podría llevar a predicciones erróneas. Para corregir esto, en ocasiones, se puede imponer una

restricción adicional para que el polinomio sea lineal en los extremos de su dominio, añadiendo el adjetivo de *natural* para designarlos. Esta modificación, libera $2(M-2)$ funciones bases, pues quita todas las bases de orden mayor a 1 en los dos nodos frontera. Es razonable que esta modificación mejore la fuerza predictiva fuera de el dominio de entrenamiento. Sin embargo, en un contexto de regresión (o clasificación) general, se recomienda no hacer inferencia fuera de el espacio de covariables $\mathcal{X}^d$, pues en realidad, no se tiene evidencia para tomar conclusiones en esta región.

Al estar trabajando con espacios funcionales en la definición de $\widetilde{\Psi}$, la elección del tipo de base funcional es relativamente arbitraria y se podría modificar como lo hace una transformación de coordenadas en un espacio euclidiano; cada base tiene sus beneficios y desventajas. Para esta exposición, se escoge la expansión en bases truncadas pues es explicada con facilidad y tiene una forma funcional relativamente sencilla. Además, la interpretación de los coeficientes $\boldsymbol{\beta}$ es inmediata. Sin embargo, no es buena computacionalmente pues el algoritmo recae en el cálculo de matrices inversas que aumenta proporcionalmente con n. En la práctica, usualmente se implementan los *B-Splines* o bases ortogonales que se derivan de lo estudiado.[34] No obstante, para no complicar más la exposición (y el algoritmo en si) se implementó una versión vectorizada de la ecuación (20) con base en la tabla 1 y la estructura (24) que se ejecuta bastante rápido inclusive cuando n y J son grandes.

En la práctica, los parámetros M, J y K se calibran comparando diferentes alternativas de modelos pues, como ya se mencionó anteriormente, hacer J variable y esti-

34. Vease el capítulo 5.5 de **wasserman2007all** o el apéndice del capítulo 5 en **hastie2008elements**.

marlo automáticamente hubiera escapado de los objetivos del trabajo.[35] Finalmente, cabe mencionar que aunque el modelo no sufra de problemas de identificabilidad en los parámetros, no se puede asegurar la no-colinealidad entre las columnas de $\widetilde{\Psi}$ por construcción, por lo que se podrían dar problemas en la estimación.[36]

35. En el capítulo **??** y **??** se discute la selección de covariables como extensiones al modelo.

36. Bajo el paradigma frecuentista y esta forma funcional, los parámetros también se podrían estimar por un procedimiento de mínimos cuadrados, en donde de haberlos, serían evidentes los problemas de colinealidad en la matriz de covarianzas $\widetilde{\Psi}^t\widetilde{\Psi}$.