Como base fundamental de este trabajo, a continuación, se expone a detalle la formulación matemática del modelo. Para su exposición, se encoje una presentación de *arriba hacia abajo*, es decir, de la parte más general del modelo (el modelo probit) a la parte más profunda (las funciones de proyección f_j); este enfoque facilita en gran medida su comprensión. En general, se trata de respetar la notación que usada en los libros de **hastie2008elements** y **james2013introduction**, asimismo, al comienzo de este trabajo se presenta un compendio de los símbolos y signos usados.

Se asume la siguiente estructura: $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ es el conjunto de datos observados independientes, con n el tamaño de la muestra donde, $y_i \in \{0, 1\} \quad \forall i = 1 \ldots, n$ son las variables de respuesta binarias o *output*, $\mathbf{x}_i \in \mathcal{X}^d \subseteq \mathbb{R}^d \quad \forall\, i = 1 \ldots, n$ las covariables, regresores o *input*[1] y $d \in \mathbb{N}$ la dimensionalidad de las covariables. Estos datos se organizan y se representan en una tabla (o matriz) como la presentada en la Tabla 1. En ella, cada fila $i = 1, \ldots, n$ representa una observación, la primer columna es el vector de respuestas y las columnas subsecuentes $j = 1 \ldots, d$ representa una variable o dimensión. Asimismo, se define el espacio de covariables como el producto cartesiano de los rangos de cada columna j, es decir:

1. Se utiliza la convención de usar negritas para distinguir las observaciones, o vistos de otra forma, los vectores fila $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n}) \quad i = 1, \ldots, n$

$$\mathcal{X}^d = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d] \subseteq \mathbb{R}^d \quad \text{con}$$

$$a_j = \text{mín}\left\{x_{1,j}, \ldots, x_{n,j}\right\} \quad \forall j = 1, \ldots, d$$

$$b_j = \text{máx}\left\{x_{1,j}, \ldots, x_{n,j}\right\} \quad \forall j = 1, \ldots, d.$$

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

Tabla 1: Estructura de los datos

El modelo, se define a continuación $\forall i = 1, \ldots, n$:

$$z_i \,|\, \mathbf{x}_i \sim \mathrm{N}(z_i \,|f(\mathbf{x}_i), 1) \tag{1}$$

$$y_i = \begin{cases} 1 & \text{si y solo si } \; z_i > 0 \\ \\ 0 & \text{si y solo si } \; z_i \leq 0 \end{cases} \tag{2}$$

$$f(\mathbf{x}_i) = f_0 + f_1(x_{i,1}) + f_2(x_{i,2}) + \ldots + f_d(x_{i,d}) \tag{3}$$

$$f_j(x_{i,j}) = \sum_{l=1}^{N^*} w_{j,l} \Psi_{j,l}(x_{i,j}, \mathcal{P}_j) \quad \forall j = 1, \ldots, d \tag{4}$$

La expresión (1) introduce n variables independientes z_i, llamadas variables latentes, con distribución normal condicionada a los datos observados $\mathbf{x}_i$. Las variables latentes, se asocian a cada respuesta y_i por medio de la expresión (2). Esta definición, es equivalente a la definición de un modelo probit tradicional. Los modelos GLM y la demostración de equivalencia entre definiciones, se detallan en la Sección 0.1 (**maccullagh1989generalized**; **sundberg2016exponential**). Se adopta este enfoque pues **albert1993bayesian** desarrollan un método de simulación, bajo el paradigma bayesiano, para el cómputo exacto de las distribuciones posteriores de los parámetros $w_{j,l}$ de los modelos probit.

Posteriormente, la ecuación (3) especifica la media de las variables latentes definidas en la expresión (1), es decir, se le da forma funcional a $\mathbb{E}[z_i|\mathbf{x}_i] = f(\mathbf{x}_i)$. A esta función $f(\cdot)$ se le conoce como función de predicción. La idea es suponer que la relación entre los componentes de las covariables $j = 1, \ldots, d$, es modelable como la

suma de funciones (usualmente suaves) f_j más un término independiente f_0. Esta definición corresponde a los GAM introducidos en **hastie1986generalized**; en la Sección 0.2, se estudia el detalle de este tipo de modelos.

Finalmente, la expresión (4) define a las funciones f_j, para $j = 1 \ldots, d$, en la parte más profunda del modelo. Estas funciones realizan una transformación no lineal de las covariables $x_{i,j}$. Este proceso se lleva a cabo mediante una expansión en bases funcionales revisada a detalle en la Sección 0.3. El objetivo de esta expansión es ponderar cada función base $\Psi_{j,l}(x_{i,j}, \mathcal{P}_j)$ por parámetros desconocidos $w_{j,l}$ los cuales se deben de estimar. Del mismo modo, las formas funcionales de las funciones base $\Psi_{j,l}$ dependen de tres componentes: las covariables $x_{i,j}$, una partición $\mathcal{P}_j$ para cada dimensión y el número total de funciones base $N^* \in \mathbb{N}$. Por el momento, se decide dejar a las funciones bases $\Psi_{j,l}$ no especificadas pues su construcción, y definición del número de ellas N^*, es compleja. Sin embargo, más adelante en las ecuaciones (27) y (28) se presentan a detalle. El modelo en si, es lineal en los parámetros más no en las covariables pues, se verá, que las funciones $\Psi_{j,l}(\cdot)$ introducen polinomios de orden mayor, particularmente, se escoge la expansión en bases de polinomios truncados presentada en **mallik1998automatic** pues resulta muy atractivo el nivel de flexibilidad logrado.

Antes de continuar y para esclarecer el trabajo un poco más, en la Figura 1 se presenta un diagrama del modelo y sus componentes. De izquierda a derecha y para toda $i = 1, \ldots, n$: se busca transformar de forma no lineal a cada una de las covariables $x_{i,j} \quad \forall j = 1, \ldots, d$ a través polinomios por partes condensados en las

funciones f_j. Estas transformaciones dependen de los parámetros por estimar $w_{j,l}$, donde $l = 1, \ldots, N^*$, y la partición de cada dimensión P_j. Una vez se tienen los datos transformados, se suman las funciones f_j con un intercepto f_0 para obtener una función de predicción f. Esta función actúa como la media de la variable latente z_i, a su vez, z_i conecta a la respuesta y_i con la información adicional contenida en las covariables $\mathbf{x}_i$, a través de un modelo probit para lograr la clasificación binaria en y_i.

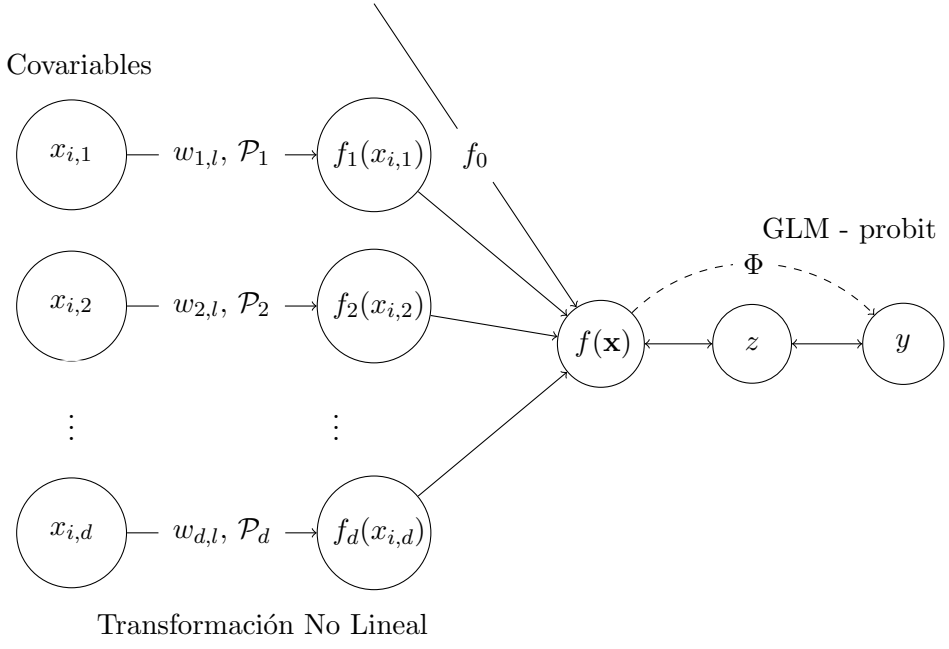

Figura 1: Diagrama del modelo

Las aparentemente complejas interacciones entre todos los componentes del modelo no son más que respuestas estructurales a un proceso de *síntesis* de la información. El modelo está buscando un patrón en las covariables $\mathbf{x}_i$ para la correcta clasificación de

su respuesta binaria asociada y_i. Para llevar a cabo este proceso y bajo este modelo, se realizan tres transformaciones:

$$f_j : \mathcal{X}_j = [a_j, b_j] \to \mathbb{R} \quad \forall j = 1, \ldots, d \tag{5}$$

$$f : \mathbb{R}^d \to \mathbb{R} \tag{6}$$

$$\Phi : \mathbb{R} \to (0, 1). \tag{7}$$

La primera (5), mapea cada dato $x_{i,j}$, dentro de su espacio correspondiente $\mathcal{X}_j$, a un número real. La segunda (6), colapsa las $j = 1, \ldots, d$ dimensiones o variables a un único número real. Se espera que f logre separe el espacio d-dimensional a regiones más identificables (para la clasificación) que las originales. La última (7), escala este real al intervalo unitario el cual, se interpreta como la probabilidad de que la respuesta asociada y_i sea uno o cero. Este proceso de síntesis de información es precisamente el objetivo del modelo, por lo tanto, es fundamental que cada componente esté bien ligada a los otros. De igual forma, es fundamental que la estimación de los parámetros del modelo $w_{i,j}$ sea correcta, de tal forma que cada transformación preserve la información y efectivamente se encuentren los patrones buscados. El Capítulo **??**, cuenta con visualizaciones que vuelven estos conceptos teóricos en algo más concreto.

El resto del Capítulo, se enfocará en estudiar a detalle cada uno de los componentes que constituyen al modelo. Sin embargo y antes de continuar, vale la pena aludir a

la celebre frase:

All models are wrong but some are useful[2]

Es la perspectiva del autor que, tratar de construir un modelo que explique perfectamente los datos sería una tarea inútil. Sin embargo, esto no significa que no se pueda intentar discernir patrones en ellos y esto, es justamente lo que se busca con la construcción de este modelo; además de profundizar sobre conceptos claves para el aprendizaje de máquina y entender las bases de modelos más avanzados.

0.1. Modelos lineales generalizados (GLM)

Los modelos lineales generalizados, surgen como una generalización del modelo lineal ordinario:

$$y_i = \beta_0 + \beta^t \mathbf{x}_i + \epsilon_i \qquad \forall i = 1, \ldots, n$$

donde $y_i \in \mathbb{R}$, $\beta \in \mathbb{R}^{d+1}$ es un vector de parámetros y ϵ_i es error estadistico (**sundberg2016exponential**). Los GLM, busca flexibilizar las regresiones ordinarias al darle diferentes rangos a la respuesta y_i, por ejemplo, los casos donde los datos están restringidos a un subconjunto de los reales como lo es el caso binario. Sin embargo, esta modificación vuelve al modelo más complejo y deriva en técnicas diversas para la estimación de β. Asimismo, la generalización del modelo lleva a que

2. **box1979robustnessinthe**

la intrepretación de los parámetros sea más compleja.[3]

Los GLM se especifican (de manera muy general) de la siguiente manera:

$$\begin{aligned} y &\sim F(\mu(\mathbf{x})) \\ \eta &= \beta^t \mathbf{x} \\ \mu &= g^{-1}(\eta) \end{aligned} \tag{8}$$

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. 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 (linealmente) la variabilidad sistemática de los datos.[4]

g: función liga que une la media μ de la distribución con el predictor lineal,[5] es decir: $\mu(x) = \mathbb{E}[y|x] = g^{-1}(\beta^t x)$. g puede ser cualquier función monótona que, idealmente, mapee de forma suave y biyectiva el dominio de la media μ

3. 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. $\ln(\pi_1/\pi_0) = \beta^t x$, donde π_k con $k = \{0, 1\}$, es la probabilidad de que la respuesta y sea 0 o 1 respectivamente.

4. Como restricción adicional, en el modelo clásico se pide que $\dim(\beta) = d < n$.

5. Al especificar un GLM, particularmente al trabajar con distribuciones exponenciales, es usual parametrizar la distribución no con la media μ sino con el parámetro canónico θ. Si la función g es tal que $\eta \equiv \theta$ entonces se dice que g es la función liga canónica.

con el rango del predictor lineal η.

Modelo probit

Dado que para este trabajo se busca construir un clasificador donde los datos son binarios, i.e. $y_i \in \{0, 1\}\ \forall i = 1, \ldots, n$; la discusión se centrará en la distribución Bernoulli pues, dado su dominio, resulta natural modelar con ella los datos. Esto es:

$$y_i \sim \text{Be}(y_i \,|\, p_i). \tag{9}$$

Esta restricción acota mucho las variedades de modelos que se pueden obtener bajo la especificación de (8). Sin embargo, en **maccullagh1989generalized** se presenta con todo formalismo y rigor la rica extensión de los GLM's además de ejemplos prácticos de su aplicación y de su desarrollo histórico.

La distribución Bernoulli (9), tiene una estructura sencilla que puede ser resumida en las siguientes expresiones $\forall i = 1, \ldots, n$:

$$f(y_i; p_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \quad \text{donde } \ y_i \in \{0, 1\} \tag{10}$$

$$= \exp\left\{ y_i \ln\left(\frac{p_i}{1 - p_i}\right) + \ln(1 - p_i) \right\} \tag{11}$$

$$\mathbb{E}[y_i] = \mu_i = P(y_i = 1) = p_i \tag{12}$$

$$\mathbb{V}[y_i] = p_i(1 - p_i). \tag{13}$$

En (10) se observa la función de densidad Bernoulli en su forma tradicional que al ser expresada como en (11) cumpla la definición de la familia exponencial.[6] Dada el soporte y la definición de la distribución Bernoulli, la media de la distribución $\mu = p$ coincide con la probabilidad de que la variable aleatoria tome el valor de uno como lo resume la ecuación (12). Asimismo, la varianza queda especificada por el mismo parámetro p, expresión (13).

Esta propiedad de la media (12) es de gran utilidad para un GLM por dos razones. Primero, además de modelar la media $\mu = p$, 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 mapeean de forma biyectiva $\mathbb{R}$, el dominio del predictor lineal η, al intervalo $(0, 1)$, el dominio de la media. Dadas las 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 rebase cierto umbral, crecen rápidamente y toman valores más cercanos a uno, *activando* así la probabilidad de que y sea un éxito.[7]

En particular, en este trabajo se escoge como función liga a la función probit, la

6. 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 $h(y), yA(\theta)$ funciones conocidas y θ el parámetro canónico, en este caso $\theta(p) = \ln p/(1-p)$.

7. En un contexto de redes neuronales, lo que se activa es la neurona y recientemente, se usa mucho la función $ReLu(x) := \text{máx}\{0, x\}$.

inversa de la función de acumulación normal estándar $\Phi(\cdot)$, i.e. $g(\mu) = g(p) = \text{probit}(p) = \Phi^{-1}(p)$, dándole nombre al modelo. Esta decisión, es consecuencia del trabajo de **albert1993bayesian**; en el, los autores describen un algoritmo bajo el paradigma bayesiano para la estimación de los parámetros del modelo que se estudiará más a detalles en la Sección 0.1.1 y **??**. Dado que la notación puede ser confusa, en la Figura 2 se presenta una representación gráfica de la función liga para un modelo probit.[8]

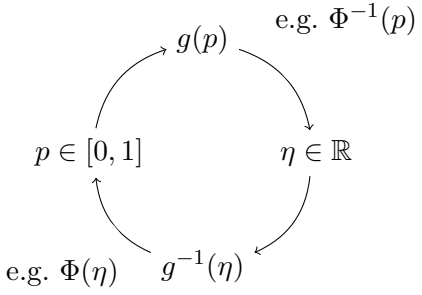

Figura 2: Esquema de función liga g para un modelo probit

Juntando todos los componentes, se está en posibilidades de detallar el modelo probit en su forma más rigurosa, rescatando la notación de un GLM con sus respectivas

8. Para no caer en redundancia 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

covariables $\mathbf{x}_i$:

$$y_i|\mathbf{x}_i \sim \text{Be}(y_i|p_i)\forall i = 1, \ldots, n \tag{14}$$

$$\eta = f(\mathbf{x}_i) \tag{15}$$

$$p_i = \Phi(\eta) = \Phi(f(\mathbf{x}_i)) \tag{16}$$

Equivalencia en las definiciones del modelo

El lector notará, sin embargo, que la especificación del modelo probit en las ecuaciones anteriores no corresponde a la definición mostrada al principio del Capítulo. Sin embargo, a continuación se prueba de forma muy sencilla la equivalencia entre ellas.

Teorema 1. *Un modelo probit especificado en* (14), (15) *y* (16)*, es equivalente a un modelo de variable latente como el presentado en* (1) *y* (2)*.*

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

$$\begin{aligned}
\mathbb{E}[y_i \,|\, x_i] &= p_i \\
&= P(y_i = 1 \,|\, \mathbf{x}_i) \qquad \text{por (14)} \\
&= \Phi(f(\mathbf{x}_i)) \qquad \text{por (16)}
\end{aligned}$$

Lo cual, es equivalente a introducir n variables aleatorias $\tilde{z}_i \sim N(\tilde{z}_i|0,1)$ tales que:

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

Donde $z_i = \tilde{z}_i + f(\mathbf{x}_i)$ es una transformación inyectiva de $\tilde{z}_i$ tal que $z_i|\mathbf{x}_i \sim N(z_i|f(\mathbf{x}_i), 1)$ idéntico a la expresión (1). Asimismo, al tener la igualdad $P(y_i = 1|\mathbf{x}_i) = P(z_i > 0|\mathbf{x}_i)$ y su probabilidad complementaria $P(y_i = 0|\mathbf{x}_i) = P(z_i \leq 0|\mathbf{x}_i)$, se da una correspondencia uno a uno entre espacios de probabilidad y se puede definir y_i en terminos de z_i y viceversa, dando lugar a la definición (2). Q.E.D.

La ecuación (15) realmente no influye en la prueba pues esta puede tener la forma funcional que se requiera para la aplicación especifica, ya sea lineal $\eta = \beta\mathbf{x}_i$ o algo diferente como se opta en este trabajo $\eta = f(\mathbf{x}_i)$.

0.1.1. La variable latente z y el predictor lineal f

Para entender como se conectan las n variables latente z_i con sus respectivos predictores lineales $f(\mathbf{x}_i)$, se necesita profundizar un poco más en el objetivo del modelo. En la Sección **??** se detalla el detalle algorítmico, derivado de la implicación bayesiana, de las variables latentes z y su conexión con los parámetros $\mathbf{w}$.

, se necesita entender que es posible estructurar estos modelos como *modelos de variable latente* (**albert1993bayesian**). Bajo esta formulación, se asume que la relación entre y y x no es directa, sin embargo, existe una variable no observada z estructural que ayuda a discernir un vínculo entre ellas. En la Figura 3 se tiene una representación gráfica de esto.

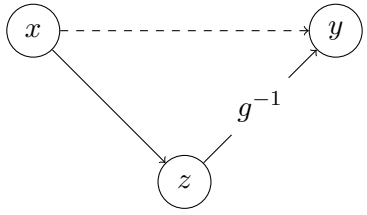

Figura 3: Modelo de variable latente

Tradicionalmente, la normalidad en z es derivada de suponer normalidad en los errores; es decir, dada la regresión lineal $z = \beta^t x + e$ se supone (y se debe verificar) que $e \sim N(0, \sigma^2)$. Lo cual lleva a $z \sim N(\beta^t x, \sigma^2)$. Además este supuesto facilita la estructura de los modelos y el algoritmo de ajuste. Bajo un paradigma frecuentista, la estimación de los parámetros β se reduce a encontrar los estimadores de mínimos cuadrados. Sin embargo, bajo el paradigma bayesiano, dentro de un modelo probit como el de este trabajo, se adopta la normalidad en z pues en **albert1993bayesian**, se sugiere un algoritmo *Gibbs sampler* con distribuciones truncadas de la normal para encontrar β.

La función liga probit se escoge como consecuencia de la normalidad en z, o viceversa, dependiendo de como se quiera ver. Esta función $\Phi^{-1}(p)$ es la inversa de la función

de acumulación normal estándar:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

la cual no tiene forma analítica cerrada, sin embargo, es claramente sigmoide. Esta función, cumple el propósito de modelar y cuantificar la incertidumbre pues está transformando la variable real z en una probabilidad p con su característica forma de "s". Se hace notar, que se podría haber usado una función más flexible o que, incluso, se podría dejar la función como una parte del modelo a estimar. Sin embargo, al adoptar el algoritmo antes citado, se requiere esta especificación.[9] La parte flexible de este modelo se encuentra en el proyector lineal.

Habiendo definido la función liga, distinguir entre si $y = 1$ ó 0, éxito o fracaso respectivamente, se reduce a distinguir en que área del espacio de covariables $\mathcal{X}^d$ se encuentra el dato. Esto se debe a que $y = 1$ cuando $\Phi(z) > 1/2$ que sucede *si y sólo si* $z > 0$ lo cual, depende en gran media de su media; en este caso la función de proyección $f(\mathbf{x})$. Si esta función es muy positiva en alguna región, implicará que el modelo tiene mucha evidencia para confiar que, al menos en esa área, la respuesta y es un éxito. El razonamiento, funciona de forma análoga para los casos donde $y = 0$, claramente, para esas regiones, se busca que $f(\mathbf{x})$ sea negativa. Por lo tanto, es fundamental para el modelo que se realice una correcta estimación de los parámetros de la función de proyección. Nótese además, que z le agrega cierta

9. En modelos multinomiales bayesianos, tomar esta decisión estructural lleva a que se puede asumir una estructura de interdependencia en los errores aleatorios. $\epsilon_k \sim N_k(0, \Sigma)$ con Σ una matriz de correlaciones

estocasticidad al modelo. Bajo la suposición que existe una pareja $(y_i, \mathbf{x}_i)$ tal que $f(\mathbf{x}_i) = 0$; alrededor de una vecindad de este punto, no se tendría evidencia para clasificar a y_i como un éxito o como un fracaso; sería mejor dejar la clasificación a la suerte.

Otro factor importante a considerar, es que el modelo supone que la varianza de z es constante, específicamente $\sigma^2 = 1$. Dado que la escala de z es completamente arbitraria pues es una variable auxiliar, se puede *restringir* z al rango que se desee. El método de simulación para z usando una distribución normal truncada se simplifica ligeramente usando esta varianza unitaria. Se verá en los resultados, sin embargo, que dada la naturaleza global de los polinomios que se usan, la escala de z, o al menos la estimación de su media $\hat{f}(\mathbf{x})$, puede variar mucho dependiendo de los datos, mas esto no representa un problema. Pues, en la practica, al usar el algoritmo de Albert y Chibb z sirve mucho más, para hacer la ligadura de y hacia f y no viceversa. En z se codifica, mediante una distribución normal truncada, los casos de éxito y de fracasos de y; posteriormente, se estima el vector β para la función f. Por ello, en la Figura 1, se representan las flechas de y a f por medio de z y solidas, en contraposición con la flecha punteada que va, directamente de f a y y pasa por la función Φ. Los detalles y su justificación probabilista, se tocan en detalle en el Capítulo **??**.

Es importante mencionar, que el *corte* que se hace en $z = 0$ para la clasificación, es resultado de hacer una clasificación binaria. En modelos multinomiales también se debe tomar en cuenta los intervalos en $\mathbb{R}$ para los que la observación se clasificaría en alguna de las posibles clases y por ende, estimar los umbrales o usar una función

diferente a las sigmoides. Este hecho, lleva a la realización de que z y su media f son *ajenas más no independientes*. Esto quiere decir que la parametrización de z como una normal $N(\mu, 1)$ es equivalente a la parametrización $N(0, 1)$. Este hecho se hará más claro cuando se hable del papel de β_0 en la función de proyección.

Finalmente, si se quisiera ver la relación de x en y directamente, se puede lograr usando el teorema de la probabilidad total. Se puede calcular (al menos de forma teórica), la distribución marginal de y dado $\mathbf{x}$ sumando sobre z:

$$\begin{aligned} P(y|x) &= \int_{-\infty}^{\infty} P(y|z)P(z|x)\,dz \\ &= \int_{-\infty}^{\infty} p(y;\Phi(z))p(z;f(\mathbf{x}),\sigma^2)\,dz \\ &= \int_{-\infty}^{\infty} \Phi(z)^y(1-\Phi(z))^{1-y}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(z-f(\mathbf{x}))^2}\,dz \end{aligned}$$

Sin embargo, está claro que esta derivación no lleva a ningún resultado analítico cerrado pues la relación es bastante más compleja como para resultar en una distribución tradicional; si lo hiciera, el propósito de la z se perdería.

Recapitulando, mediante la función liga Φ se une la media p, la probabilidad de éxito o fracaso de la respuesta y con los datos $\mathbf{x}$. Esto se logra, a través de una variable auxiliar z cuya media $f(\mathbf{x})$ es una función de proyección lineal.

$$P(y=1) = p(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = g^{-1}(f(\mathbf{x})) = \Phi(f(\mathbf{x})) \tag{17}$$

0.2. Función de predicción f

En la sección anteriór, se vió que tradicionalmente, se asumía z como una combinación lineal de los parámetros β y las covariables x.[10] Pero, como se explica 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 romper la linealidad. En 1986, Hastie y Tibshirani introducen los modelos aditivos generalizados (GAM), una clase de modelos donde se rompe la linealidad en las covariables, flexibilizando aún más el modelo.

Se hace notar que esta generalización es sutil pues el modelo aún conserva una parte lineal *a lo largo*.[11] Se ve en la ecuación (3) que el modelo aún es lineal en las β_j y en las f_j. Donde se pierde la linealidad es *hacia abajo*,[12] pues cada f_j en realidad es la transformación no lineal de la covariables x_j. Además, las dimensiones j's se asumen independientes entre si pues, se busca que corresponden a diferentes *características* o variables con las que se planea modelar la variable de respuesta.

En este modelo, el proyector f de la ecuación (3), no hace otra cosa más que ir colapsando dimensiones, en particular: $\mathcal{X}^d \to \mathbb{R}$ que posteriormente se colapsa por medio de Φ en $[0, 1]$. Es por esto que se le llama función de proyección, pues *proyecta*

10. Segunda ecuación de (8)

11. Con esta frase se hace alusión a que, en una tabla de datos de tamaño $n \times d$, siendo cada fila una observación y cada columna una variable, el *largo* se piensa como la segunda dimensión de tamaño d. Por lo tanto, al usar esta expresión se busca considerar todas las variables. En este trabajo y en su implementación, se usa el subíndice j para denotar esta idea.

12. De forma análoga, esta idea, hace alusión a las diferentes observaciones. Denotado por el subíndice i, el cual no se ha usado para simplificar la notación.

el espacio $\mathcal{X}^d$ en $\mathbb{R}$. Sin embargo, la forma en la que lo haga, debe de ser lo más precisa posible pues el modelo recae en que este colapso detecte los patrones correctos en las covariables que llevan a la correcta identificación de y. Por lo tanto, f como función de proyección es el corazón del modelo, por lo que su correcto entrenamiento es fundamental. La idea, recapitulando, es que f separe el espacio de covariables para que sea positiva en las regiones donde se tengan éxitos y negativa donde se tengan fracasos; para ello, es fundamental entender los GAM.

0.2.1. Modelos aditivos generalizados (GAM)

Un GAM como se introduce en **hastie1986generalized** remplaza la forma lineal $\sum_1^d \beta_j x_j = \beta^t x$ con una suma de funciones *suaves* $\sum_j^d f_j(x_j)$. Estas funciones no tienen una forma analiítica cerrada y son no especificadas, es decir, no hay tienen una forma funcional concreta y representable algebraicamente. Donde recae la fuerza de estos modelos, es que se estiman usando técnicas no paramétricas de suavizamiento[13] como lo sería un suavizamiento loess. En estos modelos, se supone que por más grande que sea $\mathcal{X}^d$, la relación que existe entre cada una de las dimensiones j, se puede explicar de manera aditiva, es por ello que cada función f_j tiene como argumento exclusivamente de x_j. Esta especificación fue revolucionaria pues no sólo regresa interpretabilidad al modelo, sino que simplifica la estimación usando técnicas prácticamente automáticas con el algoritmo de *backfitting*. La idea fundamental de

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

este algoritmo será de vital importancia para el ajuste de el modelo. Los principal ventaja de los GAM es que logran descubrir efectos no lineales en las covariables, justamente lo que se busca. La función f de este trabajo, es una versión versión modificada de un GAM con tres cambios fundamentales.

La primera modificación es que se está ponderando cada f_j por un parámetro β_j, esto, para suavizar aún más cada dimensión y captar el patrón general y no tanto los componentes individuales de cada x_j. 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 parámetro β_j como el efecto que tiene la dimensión i en particular para el modelo. Asimismo, se reincorpora un término independiente β_0 pues este ayuda a ajustar la escala en la estimación de los parámetros. La inclusión de este parámetro es fundamental para la correcta especificación del modelo pues ayuda a dar un *sesgo o nivel* base contra el cual comparar la suma y escalar la f para que sea compatible con el umbral de corte en 0 haciendo equivalente la parametrización de z con una distribución normal estándar. Por convención $f_0(\cdot) = 1$ por lo que se puede re-expresar la ecuación (3) como:

$$f(\mathbf{x}) \approx \beta_0 + \sum_{i=1}^{d} \beta_i f_i(x_i) = \beta^t \mathbf{f}(\mathbf{x}),$$

donde usando notación vectorial $\beta \in \mathbb{R}^{d+1}$ y $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{d+1}$:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_0 \\ f_1(x_1) \\ f_2(x_2) \\ \vdots \\ f_d(x_d) \end{bmatrix} = \begin{bmatrix} 1 \\ w_1^t \Psi_1(x_1, \mathcal{P}_1) \\ w_2^t \Psi_2(x_2, \mathcal{P}_2) \\ \vdots \\ w_d^t \Psi_d(x_d, \mathcal{P}_d) \end{bmatrix} \tag{18}$$

La segunda modificación, se ve en la expresión anterior (18). Aunque es muy práctico manejar las f_j's como indeterminadas y estimarlas con procedimientos de suavizamiento no paramétricos, también se puede optar por la vía en la que se especifica su forma funcional, en este caso $w_j^t \Psi_j(\cdot)$. No por ello, se le quita flexibilidad al procedimiento; esto se verá con todo detalle en la sección 0.3, donde se trata de adapar el procedimiento de **mallik1998automatic**. Esta modificación, obedece a que para ciertas aplicaciones, sirve hacer el modelo paramétrico en donde cada f_j se modela en su expansión de bases, vease Capitulo 9.1 y Ejemplo 5.2.2 de **hastie2008elements**.

La tercera modificación, es que en la practica, se tiene una aproximación en vez de una igualdad para f. El simple hecho de asumir que existe una f que puede separar el espacio en regiones positivas y negativas es uno de los supuestos más fuertes del modelo, esto responde a que el *error aleatorio*, sistemático de los datos, está siendo capturado por una aproximación a la f real, dentro de cada una de las f_j's. Bajo el paradigma frecuentista, se desarrolla toda una amplia teoría de convergencia y se explica a detalle el porqué de esta modificación usando técnicas de análisis más avanzadas (**bergstrom1985estimation**), (**stone1985additive**).

En la peculiaridad de que $d = 2$, se podrá visualizar $f(\mathbf{x})$ en $\mathbb{R}^3$ como una serie de picos y valles donde será positiva en caso de ser éxito y negativa en caso contrario.

Modelos en bases de funciones lineales

Finalmente, se aclara que este modelo podría ser confundido con un modelo en bases de funciones lineales[14] como los presentados en Capitulo 3 de **bishop2006pattern**:

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{d} \beta_j f_j(\mathbf{x}), \tag{19}$$

claramente con una forma funcional simular. La diferencia radica en que cada f_j es función de todas las covariables en lugar de sólo la que le corresponde en el índice. A estas funciones se les conocen como funciones base, sobre las que se hará una exposición más a detalle en la siguiente sección. La f una vez más es lineales en β, no-lineales en $\mathbf{x}$, pero la diferencia radica en que f_j es de naturaleza global. Algunos ejemplos son:

bases gaussianas,

$$f_j(\mathbf{x}) = \exp\left\{-\frac{(\mathbf{x} - \mu_j)^2}{2s^2}\right\}$$

14. *Linear basis function models* por falta de una mejor tradición.

funciones sigmoidales,

$$f_i(\mathbf{x}) = \sigma\left(\frac{\mathbf{x} - \mu_j}{s}\right).$$

Estos modelos son muy útiles en la estimación de mezclas de distribuciones y en la estimación de curvas. Sin embargo, estos son un grupo de modelos diferentes y su aplicación no es tanto inferencial como lo que se busca. En realidad, estos modelos son más una extensión al modelo que se construye en este trabajo.

0.3. Funciones f_j - polinomios por partes

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 medida f 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 en polinomios por partes. Toda la siguiente sección se concentra en darle formas funcionales a Ψ y a explicar el papel de los pesos w. Se usa como referencia en la exposición, las primeras dos secciones de el Capitulo 5 de **hastie2008elements**.

Expansión en bases funcionales

Saliendo por un momento del domino de la estadística, se definen las expansiones en bases de funciones. Sin entrar mucho en los detalles técnicos, dado un espacio funcional, se puede representar cualquiera de sus elementos, en este caso una funciones arbitrarias h, como la combinación lineal de los elementos de la base Ψ (también funciones) y constantes w. En particular (y dados los objetivos del trabajo) se considera el espacio funcional que mapea $\mathbb{R}^d$ a $\mathbb{R}$, quedando entonces la expansión:

$$h(\mathbf{x}) = \sum_{l=1}^{N^*} w_l \Psi_l(\mathbf{x}) = w^t \Psi(\mathbf{x}) \tag{20}$$

con el vector de funciones base $\Psi(\mathbf{x})^t = (\Psi_1(\mathbf{x}), \ldots, \Psi_{N^*}(\mathbf{x}))^t$, donde cada elemento Ψ_l es también una función con el mismo mapeado que h, $w^t = (w_1, \ldots, w_{N^*})^t$ un vector de coeficientes constantes y N^* un entero mayor o igual a la dimensión del espacio funcional que se maneja[15].

Regresando a las regresiones en el mundo de la estadística. Se busca representar la media condicional de la respuesta y por una función que depende de los datos: $h(\mathbf{x}) = \mathbb{E}[y\,|\mathbf{x}]$. Se puede pensar, que esta h también puede ser expresada como su expansión en bases funcionales.[16] La idea, es que se remplace (o se aumente) la cantidad de covariables $\mathbf{x}$ con transformaciones de estas, capturadas en el vector $\Psi(\mathbf{x})$. Por ejemplo:

15. Dependiendo de el espacio funcional, en ocasiones $N^* = \infty$
16. Un supuesto fuerte pero necesario en ocasiones.

$\Psi_l(\mathbf{x}) = x_j \quad \forall\, l = 1, \ldots, d$, donde se recupera un GLM tradicional.

$\Psi_l(\mathbf{x}) = \ln x_j$ ó $x_j^{1/2}$ donde se tienen transformaciones no lineales en cada una (o algunas) de las covariables.

$\Psi_l(\mathbf{x}) = ||\mathbf{x}||$ una transformación no lineal de todas las covariables.[17]

$\Psi_l(\mathbf{x}) = x_j^l \quad \forall l = 0, \ldots, N^*$ donde se tiene una expansión en bases polinómicas.

$\Psi_l(\mathbf{x}) = x_j x_k \quad \forall l,$ p.a. j, k donde se incluyen términos de interacción.

Esta representación engloba muchos de los modelos y transformaciones posibles en el mundo de las regresiones, uniendo temas de análisis funcional con estadística aplicada. Además de que en general han resultado ser de gran utilidad en la práctica. Se hace notar que el último ejemplo rompe con la aditividad inherente de las combinaciones lineales, demostrando que esta generalización no está restringida a ser completamente aditiva.

Dependiendo del tipo de datos y el propósito del modelo, puede ser conveniente usar algún tipo de funciones base sobre las otras. Sin embargo, sobre todo cuando se tiene poca o ninguna experiencia con los datos, se busca una representación más 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.

17. Como se vio en la ecuación (19)

Una de estas familias, es la de los polinomios por partes usadas en este modelo. Una desventaja de estos métodos, sin embargo, 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*. Algunos de los métodos más comunes para lograrlo son los siguientes:

Métodos de restricción: donde 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 perfecto de este tipo.

Métodos de selección de variables: como lo son los modelos CART y MARS, donde se explora de forma iterativa las funciones base y se incluyen aquellas que contribuyan a la regresión de forma significativa.

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*.

Consideraciones para la expansión en bases de este trabajo

Para simplificar un poco la exposición y reducir la notación, se supone por lo pronto que $d = 1$, por lo tanto $\mathbf{x} = x$ y se puede pensar únicamente en funciones que mapean reales a reales. Esto permite librar el subíndice j para indicar componente del vector $\mathbf{x}$ y usarlo para otros fines.

Para el modelo de este trabajo, se aplicaron las ideas de Denison, Mallick y Smith a los modelos GLM presentados con anterioridad. Los autores presentan un método revolucionario, automático y bayesiano, que permite estimar con un alto grado de precisión relaciones funcionales entre la variable de respuesta y y sus covariable $x \in \mathbb{R}$. En el trabajo original, se buscaba ajustar una curva tal que $y = h(x)$. El modelo en su forma estadística se plantea para un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{n}$:

$$y_i = h(x_i) + e_i \quad i = 1, \ldots, n \tag{21}$$

donde las e_i son variables aleatorios con media cero. Este método, combina los procedimientos paramétricos y no paramétricos desarrollados con anterioridad para hacer más robusto el algoritmo de **hastie1986generalized**. La idea, es ajustar un *polinomio por partes* muy flexible. Estos polinomios, se componen de partes de menor orden entre *nodos* adyacentes. Una de las muchas genialidad de su trabajo es que estos nodos, tradicionalmente fijos, se vuelven parámetros a estimar, usando un paradigma bayesiano. Y no sólo eso, sino que permiten *aumentar o disminuir el número de nodos* desarrollando un algoritmo Gibbs sampler trans-dimensional. Esta generalización, logra estimaciones tan robustas, que logran aproximar funciones continuas *casi en todas partes*, como lo son la función Doppler, funciones por bloques y funciones con picos pronunciados (**mallik1998automatic**).

0.3.1. *Splines*

Antes de exponer estos polinomios tan flexibles, se busca entender que son los polinomios por partes simplificando (bastante) el trabajo de **wahba1990splines**. Sea $x \in [a, b] \subseteq \mathbb{R}$, se busca separar $[a, b]$ en J intervalos. Por lo tanto, se construye una partición correspondiente $\mathcal{P} = \{\tau_1, \tau_2, \ldots, \tau_{J-1}\}$ tal que $a \leq \tau_1 < \ldots < \tau_{J-1} \leq b$. Estas τ's son llamadas *nodos*. Se hace notar, que se puede incluir o no la frontera dependiendo de la especificación.[18] Con estos nodos seleccionados, se puede hacer una representación de h en su expansión de bases como en la ecuación (20), donde cada Ψ_j será una función que depende, tanto de la partición como de la variable x. Por ejemplo, se puede pensar en un caso sencillo donde se tiene que $J = 3$ y a cada subintervalo les corresponde una función $\Psi_j \quad j = 1, \ldots, 3$. Simplificando aún más, se hacen que estas Ψ_j's sean funciones constantes en cada intervalo. Por lo tanto, las funciones base son:

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

con $I(\cdot)$ la función indicadora que vale 1 si x se encuentra en la región y 0 en otro

18. Dependiendo de si se busca hacer inferencia o no fuera del los intervalos.

caso. Por lo tanto, la expansión es:

$$\begin{aligned} h(x) &= \sum_{j=1}^{J} w_j \Psi_j(x) \\ &= w_1 I(x < \tau_1) + w_2 I(\tau_1 \leq x < \tau_2) + w_3 I(\tau_2 \leq x). \end{aligned}$$

Resultando en una función *escalonada*, en el sentido de que para cada región de x se tiene un nivel w_j.[19] Esta aproximación, podría servir para datos que estén agrupados por niveles, sin embargo, rara vez será este el caso.

Con este ejemplo sencillo, se ilustra a grandes rasgos como funcionan los polinomios por partes. Sin embargo, en cada intervalo se puede ajustar un polinomio de grado arbitrario, aumentando así, el número de funciones base. Adicionalmente, se pueden añadir restricciones de continuidad en los nodos, y no sólo continuidad entre los polinomios, sino continuidad en las derivadas, lo cual logra una estimación más robusta. Esta es la magia de los polinomios por partes, que se les puede pedir cuanta *suavidad* (o no) se requiera, entendido como la continuidad de la (K)-ésima derivada. Tradicionalmente, se construyen polinomios cúbicos con segunda derivada continua en los nodos. Esto, pues resultan en curvas suaves al ojo humano, además de que logran aproximar una gran cantidad de funciones.

19. Sin entrar en el detalle, usando una función de perdida cuadrática, es fácil demostrar que cada $\hat{w}_j = \bar{x}_j$ es decir, para cada región, el mejor estimador constante, es el promedio de los puntos de esa región.

Orígenes y justificación de su uso

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. Dependiendo de las particularidades se pueden denotar funciones diferentes. 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. Se hace notar, que existen muchos tipos de *splines*, además de que pueden ser, puede ser más flexibles o más rápidos en su implementación computacional como los B-Splines. En **deboor1978splines** y más recientemente **wahba1990splines** se hacen tratados extensivos sobre ellos.

Como breviario historico, los splines originales, los desarrolla el matemático I. J. Schoenberg como la solución al problema de encontrar la función h en el espacio de Sobolev W_M de funciones con $M-1$ derivadas continuas y M-ésima derivada integrable al cuadrado que minimice:

$$\int_a^b (h^{(M)}(x))^2 \, dx,$$

sujeta a que interpole los puntos $h(x_i) = h_i \quad i = 1, 2, \ldots, n$ (**schoenberg1964spline**). Posteriormente, la teoría sobre los splines se fue expandiendo y fueron adoptados por ramas de la matemática tan diversas como los gráficos por computadora y, como es el caso, la estadística computacional. Bajo este contexto, los splines también surgen de forma orgánica pues, la ecuación (21) se puede plantear como encontrar

la función h que minimice:

$$\sum_{i=1}^{n}(y_i - h(x_i))^2 + \lambda \int_a^b (h^{(M)}(x))^2 \, dx, \tag{22}$$

para alguna $\lambda > 0$. Donde, la solución se demuestra que son *splines cúbicos naturales* ($M = 4$). Cabe mencionar, que esta formulación del problema engloba muchas de técnicas estadísticas interesantes además de conceptos de optimización. El lector reconocerá que el primer término claramente es la *suma de residuales cuadrados (RSS)* y el segundo término del sumando es un caso particular de los métodos de regularización mencionados anteriormente. No es el enfoque del trabajo entrar en estos detalles pues, cambios menores en la formulación y diferentes elecciones de λ llevan a modelos que cada uno merece una tesis por si mismo. Sin embargo, es importante mencionar que la regularización y modelos de este tipo, son algunos de los más usados y útiles en ML, pues logran captar patrones muy complejos al incluir muchos términos de orden superior e interacciones sin sobreajustar en los datos. Como ejemplo, se puede encontrar fronteras de clasificación circulares usando un modelo logístico normal en $\mathbb{R}^2$ al incluir todos los términos polinomiales y las interacciones hasta orden 6. Por lo pronto, lo esencial en la expresión (22) es que al tratar de minimizar el RSS se puede caer en problemas de sobre-ajuste en donde los parámetros no estén capturando efectos y patrones subyacentes, sino sólo se trata de seguir los datos. Para compensar la complejidad, se penaliza la función a minimizar con segundo termino que controla el número de parámetros y la suavidad deseada mediante λ. A este segundo término, se le conoce como *penalización* y crece a medida

que h se vuelve más complicada.[20]

Posterior a estas formulaciones, los splines vuelven a ser relevantes con el modelo aditivo de Hastie y Tibshirani. Ellos extienden la formulación de un espacio de covariables en una sola dimensión, a muchas. La formulación del problema es prácticamente la misma que (22) pero ahora se busca estimar d funciones h, dando lugar a tener más parámetros λ:

$$y = \sum_{j=0}^{d} h_j(x_j) + \epsilon$$

$$\text{RSS}(h_0, h_1, \ldots, h_d) = \sum_{i=1}^{n} [y_i - \sum_{j=0}^{d} h_j(x_{ij})]^2 \; + \; \sum_{j=1}^{d} \lambda_j \int h_j^{''}(t_j)\, dt_j$$

con la convención de que h_0 es una constante. Ellos muestran que $h_j \quad j = 1, \ldots, d$ son splines cúbicos. Sin embargo, sin restricciones adicionales, el modelo no sería *identificable*, es decir, la h_0 podría ser cualquier cosa. Para asegurar la unicidad de la solución se añade la condición de que las funciones estimadas, promedien cero sobre los datos:

$$\sum_{i=1}^{n} h_j(x_{ij}) = 0 \quad \forall j \tag{23}$$

Esto lleva a la conclusión natural de que h_0 sea la media de las variables de respuesta, es decir: $h_0 = \bar{y}$. Por lo que si se ve cada dimensión j, se tiene que su función

20. Si el lector tiene una intuición de análisis, notará que integrar la función al cuadrado, corresponde con el producto interno de las funciones pertenecientes al espacio de Hilbert $\mathcal{L}_2([a, b])$.

correspondiente h_j está centrada alrededor de la media $\bar{y}$. Esta idea es fundamental para el modelo de este trabajo. En el, se permite que h_j sean *arbitrarias* para toda j, por lo que sólo se necesita que tenga la magnitud necesaria para ajustar los datos. Es decir, dada h_0, la estimación y entrenamiento de los parámetros que definen por completo a h_{j^*} (con j^* alguna $j = 1, \ldots, d$) deben ser tales para que esta ajuste los *residuales parciales*:

$$\hat{h}_{j^*} = y - h_0 - \sum_{\substack{j=1 \\ j \neq j^*}}^{d} h_j \tag{24}$$

y se vaya captando en esta h_{j^*} la información aún no captada por el modelo. Esta lógica, además de brillante, es la que le da fuerza a los GAM, pero sólo se puede entender de forma completa hasta que se estudie el algoritmo de *backfitting* en la Sección **??**.

Formalización matemática de *splines*

Retomando la discusión de la página 28, se está buscando definir un polinomio de grado $M - 1$ por partes en J intervalos. Tomando una expansión de bases para cada intervalo, como en el primer ejemplo que se dio, el número de funciones base aumenta en J por cada grado que se agregue, dando un total de $J \times M$ bases funcionales, y en consecuencia, el mismo número de parámetros por estimar. Esto ocurre porque se necesita definir una base de tamaño M para cada subintervalo

$j = 1, \ldots, J$. Es decir: $\mathcal{B}_j = \left\{1, x, x^2, \ldots, x^{M-1}\right\}$, para $j = 1, \ldots, J$. Sin embargo, esto lleva a 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 $J - 1$ parámetros que corresponden a los $J - 1$ nodos. De la misma forma, cada grado de continuidad nodal en las derivadas que se le pida al polinomio, lo restringe y por ende, devuelve el mismo número de funciones bases. Sea K este número, es decir, se construye un polinomio por partes con continuidad hasta la K-ésima derivada, que tiene un total de:

$$N^*(M, J, K) = M \times J - K(J - 1) \tag{25}$$

bases funcionales. Por ende, este polinomio tiene el mismo número de parámetros por estimar w.[21] Es claro que N^* es la *dimensión mínima* necesaria para construir polinomios por partes con estas características. Pues, el número de funciones N^* está, a su vez, en función de M definiendo el grado, el número de intervalos J (por ende el número de nodos) y el número de restricciones K.

Por lo pronto, y para continuar con una exposición constructiva, se centra la discusión cuando $K = M - 1$, devolviendo la definición de spline: polinomios de grado $M-1$ con continuidad hasta la $(M-2)$-derivada. Por ende, $N^* = M+J-1$. Ahora, se recuerda que el objetivo es darle forma funcional a Ψ. Para lograr esto, habiendo

21. En ocasiones es más fácil pensar en K como el número de restricciones que se imponen en los nodos. Así, $K = 0$ implica que los intervalos son independientes, $K = 1$, implica que los polinomios se conectan, $K = 2$ implica continuidad en la primera derivada y así sucesivamente. Naturalmente $K < M$

incorporado el número de bases, se define la función auxiliar *parte positiva*:

$$x_+ = \text{máx}\left\{0, x\right\}.$$

Esta función, ayuda a se pueda representar la expansión en bases de una forma relativamente sencilla. A esta expansión, se le conoce como *expansión en bases truncada*:

$$\begin{aligned} h(x) &= \sum_{i=1}^{M+J-1} w_i \Psi_i(x, \mathcal{P}) \\ &= \sum_{i=1}^{M} w_i \; x^{i-1} + \sum_{j=1}^{J-1} w_{M+i} \; (x - \tau_i)_+^{M-1}. \end{aligned} \tag{26}$$

El primer sumando de (26) representa el *polinomio base*[22] de grado $M-1$ que afecta a todo el rango. El segundo sumando, está compuesto únicamente de funciones parte positivas que se van activando a medida que x recorre el rango $[a, b]$ a la derecha y va pasando por los nodos. Estas funciones parte positiva, capturan el efecto de todos los intervalos anteriores que, al combinarlos con el primer sumando definen un polinomio de grado $M-1$ en todo el intervalo.[23] Esta derivación de las bases, surge cuando se integra un polinomio por partes constante $M-1$ veces. En cada iteración, las constantes se juntan y se integran por si solas, independientemente de los intervalos, lo cual deriva en este polinomio base. De forma explicita, se tiene que

22. *Baseline*, una vez más a falta de una mejor traducción
23. En realidad lo hace en todo $\mathbb{R}$

$\Psi(x, \mathcal{P})$ es:

$$\begin{aligned}
&\Psi_1(x, \mathcal{P}) = 1 \\
&\Psi_2(x, \mathcal{P}) = x \\
&\vdots \\
&\Psi_M(x, \mathcal{P}) = x^{M-1} \\
&\qquad \text{(el } \textit{polinomio base}\text{)} \\
&\Psi_{M+1}(x, \mathcal{P}) = (x - \tau_1)_+^{M-1} \\
&\vdots \\
&\Psi_{M+J-1}(x, \mathcal{P}) = (x - \tau_{J-1})_+^{M-1} \\
&\qquad \text{(la base truncada)},
\end{aligned}$$

las cuales forman un espacio lineal de funciones $(M + J - 1)$-dimensional. En la particularidad que $M = 4$, se les conoce como splines cúbicos y son los más usados cuando se buscan funciones suaves. En la práctica han resultado ser de gran utilidad pues el ojo humano no detecta la posición de los nodos.

0.3.2. Polinomios por parte flexibles

Independientemente de la elección de parametros en la construcción del polinomio, se tiene el problema de seleccionar la posición de los nodos. Existen procedimientos adaptativos, como los propuestos en **friedman1991multivariate**. No obstante, y

como ya se mencionó anteriormente, **mallik1998automatic**, proponen un método bayesiano más atractivo, que aunque no se implemente en este trabajo, se implementa su expansión en bases aún más general. Esta es una ligera modificación a la ecuación (26) la cual la convierte, de un spline, a un polinomio por partes más general, con grado arbitrario de continuidad en las derivadas. Dejando atrás el supuesto que $K = M - 1$ y devolviendole esa flexibilidad al modelo. Su expansión en bases resulta:

$$h(x) = \sum_{l=1}^{N^*} w_l \ \Psi_l(x, \mathcal{P}) = w^t \Psi(x, \mathcal{P}) \qquad \text{con } N^* = J \times M - K(J-1) \tag{27}$$

$$= \sum_{i=1}^{M} w_{i,0} \ x^{i-1} + \sum_{i=K}^{M-1} \sum_{j=1}^{J-1} w_{i,j} \ (x - \tau_j)_+^i \tag{28}$$

la cual es la expansión de bases implementada en el modelo final.

Dado que se tiene una doble suma, es necesario incluir un segundo índice, al menos temporalmente, a los pesos. El primer índice, denotado por i está asociado al grado de su función base; si $i = 2$ entonces, $w_{2,j}$ está asociado a una término de grado 1 cuando $j = 0$, pero a uno de grado 2 si $j > 0$.[24] El segundo índice $j = 1, \ldots, J-1$ denota el nodo al que está asociado el peso. Como convención, si $j = 0$, se hace referencia al polinomio base que siempre tiene efecto. En el segundo sumando de (28) la primera suma comienza en K. Recordando, K es el número de restricciones de continuidad que se imponen al polinomio en los nodos. Por ejemplo, $K = 0$ implicaría que cada polinomio es independiente; $K = 2$, se tiene continuidad en la función y en

24. Esta desgraciada disparidad en la notación surge para ser consistente con la anterior, y no se puede indexar directamente en el primer sumando.

la primera derivada, etc. En el caso que $K = M - 1$ se regresa a la ecuación (26) y se recuperan los splines que, por construcción, son suaves. La suavidad, aunque útil, no siempre es necesaria. Existen muchas funciones con primera y segunda derivada que varían rápidamente e incluso funciones discontinuas que no se podrían estimar usando splines, todo depende de los datos. Esta construcción, con su doble suma, permite tener $M-K$ términos por nodo, codificando así las continuidades arbitrarias en las derivadas.[25] La ecuación (27) es una vez más la expansión en bases arbitrarias, igual a (20) pero definiendo de forma completa a N^*. Además, si finalmente en esta ecuación se deja que $h(x)$ sea igual a $f_j(x_j)$ para toda $j = 1, \ldots, d$, se regresa a la ecuación canónica del modelo (4) presentada al principio de este trabajo. Este era el último componente que quedaba por definir, completando así la exposición matemática del modelo.

Para ayudar con la interpretación y lectura de la ecuación (28), la Tabla 2, de la página 39, hace un compendio de los polinomios por partes. Esto ayuda no sólo a esclarecer la notación, sino a formar una biyección entre w_l, $w_{i,j}$ y Ψ_l que posteriormente ayudará a expresar todo de forma matricial en su implementación en código.

Antes de cerrar la sección, se centra la atención en los nodos τ. A estos, se les ha dado poca importancia hasta el momento. Como ya se mencionó antes, en **mallik1998automatic** se desarrolla, además de la ecuación (28) un paradigma ba-

25. Esta codificación es sutil pues, al hacer los cálculos de continuidad, hay que considerar los límites izquierdos y derechos, los cuales existen siempre. Sin embargo, los términos $(x - \tau)_+^K$ se desvanecen únicamente hasta la K-esima derivada. Para la $(K + 1)$-derivada, el coeficiente correspondiente se suma a la función y rompe la continuidad pues no corresponde con el límite izquierdo

w_l	$w_{i,js}$	$\Psi_l(x,\mathcal{P})$		
Subíndice l	Subíndices i,j	Función Base		
1	$1,0$	1	M elementos	
2	$2,0$	x		
$\vdots$	$\vdots$	$\vdots$		
M	$M,0$	x^{M-1}		
$M+1$	$K,1$	$(x-\tau_1)_+^K$	$M-K$	$J-1$ veces
$M+2$	$K+1,1$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(M-K)$	$M-1,1$	$(x-\tau_1)_+^{M-1}$		
$M+(M-K)+1$	$K,2$	$(x-\tau_1)_+^K$	$M-K$	
$M+(M-K)+2$	$K+1,2$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+2(M-K)$	$M-1,2$	$(x-\tau_1)_+^{M-1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-2)(M-K)+1$	$K,J-1$	$(x-\tau_{J-1})_+^K$	$M-K$	
$M+(J-2)(M-K)+2$	$K+1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-1)(M-K)$	$M-1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		

Tabla 2: Biyección entre w_l, $w_{i,j}$ y sus correspondientes funciones base Ψ_l.

Se termina con $N^* = M+(J-1)(M-K) = J\times M - K*(J-1)$ términos, ecuación (25). Por construcción, se es consistente con la definición (26) si $K = M-1$.

yesiano en el que los nodos, son tratados como parámetros y por ende sus posiciones cambian. La ventaja de que estos estén indeterminados, es que se pueden concentrar en los lugares donde la función varia más. Y al contrario, si la función es relativamente suave para alguna sección, se usan pocos nodos. Aunque hubiera sido bueno implementar este proceso, el algoritmo que *mueve* los nodos va ligado directamente a un proceso de eliminación y nacimiento de estos, haciendo que la J sea variable. En el trabajo original, esto no era un problema pues sólo se hacían estimaciones para una dimension, $d = 1$. En el contexto de este modelo probit, implementar el algoritmo trans-dimensional que los autores proponen, hubiera implicado que N^* estrella, no fuera constante para toda variable, $j = 1, \ldots, d$. Sino que se tendría N_j^*, incorporado otra capa de complejidad innecesaria. Además, el algoritmo habría sido radicalmente diferente. 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 ve que para fines prácticos, el modelo funciona de maravilla y finalmente en el Capítulo **??** se discute que habría cambiado de haberse implementado.

0.3.3. Consideraciones matemáticas adicionales

A pesar de la utilidad de los splines (y los polinomios por parte), todos sufren de problemas más allá del rango de entrenamiento $[a, b]$. Pues, su naturaleza global hace que fuera de la región con nodos, los polinomios crezcan o decrezcan rápidamente. Por lo tanto, extrapolar con polinomios o splines es peligroso y podría llevar a esti-

maciones erróneas. Para corregir esto, en ocasiones, se puede imponer una restricción adicional para que el polinomio sea lineal en sus extremos. Se usa 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. Su expansión en bases, también se deriva de la ecuación (26). Es razonable que esta modificación mejore la fuerza predictiva fuera de el dominio de entrenamiento. Sin embargo, en general, en un contexto de regresión, se recomienda no hacer inferencia fuera de el espacio de covariables $\mathcal{X}$, pues en realidad, no se tiene evidencia para tomar conclusiones en esta región. Todo depende de los datos y el objetivo del modelo.

Se han usando los parámetros M, J y K para hablar del número de funciones base N^*, ecuación (25), pero se recuerda que también, dictan el número de *grados de libertad* del modelo. Es decir, el número de pesos o coeficientes w, los cuales son igual o más importantes que las bases, no sólo porque son parámetros a estimar, sino que son los que dictan el *ajuste* a los datos a diferencia de Ψ que únicamente los operan.

Al estar trabajando en espacios funcionales, la elección de base es relativamente arbitraria y se podría cambiar 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 fácilmente y tiene una forma funcional relativamente sencilla además, la interpretación de los coeficientes w es inmediata. Sin embargo, no es óptima computacionalmente cuando

J es grande. En la practica, usualmente se implementan B-Splines[26] que se derivan de lo vistos anteriormente. No obstante, para no complicar más la exposición (y el algoritmo en si) se implementó una versión optimizada de (28) con base en la Tabla (2) que funciona bastante rápido inclusive cuando J es grande.

En la practica, los parámetros M, J y K se calibran pues, como ya se mencionó anteriormente, hacer J variable y automático es muy complejo. Asimismo, la elección de M y K requeriría cierta exploración previa de los datos. No obstante, existen algoritmos que realizan esta tarea que, para los fines de este trabajo no aportaría mucho. Además de que los resultados que se obtuvieron por el método de calibración son bastante buenos.

Si se le da rigor al modelo, en realidad, hay dos expansiones en bases. La primera la primera *a lo largo* de la ecuación lineal (3) cuyos coeficientes son β y las funciones base $\mathbf{f}$. Posteriormente, se tiene la expansión de la ecuación de polinomios por partes de (4) cuyos coeficientes son w_j y las funciones base Ψ_j para toda j. Esto explica el salto conceptual y notacional que se da entre las representaciones de (19) y (20).

Al tener en mente que se tienen d covariables, y por ende d polinomios por partes, además de la estructura lineal de (27) se puede sustituir (4) dentro de (3) dando la

26. Vease el Capítulo 5.5 de **wasserman2007all** o el Apéndice del Capítulo 5 en **hastie2008elements**.

siguiente estructura con doble suma:

$$
\begin{aligned}
f(\mathbf{x}) &\approx \sum_{j=0}^{d} \beta_j f_j(x_j) \\
&\approx \beta_0 + \sum_{j=1}^{d} \beta_j \left[\sum_{l=1}^{N^*} w_{j,l} \Psi_{j,l}(x_j, \mathcal{P}_j) \right].
\end{aligned}
$$

Lo cual, es perfectamente lineal. Se tienen $1 + d \times N^*$ términos que se pueden acomodar en un solo vector. Sin embargo, se tiene un cruce de parámetros interesante: la multiplicación de $\beta_j \quad \forall j$ contra $w_{j,l} \quad \forall l$. Tradicionalmente, no se usan β y se deja que se capture ese efecto dentro de las f_j como en los GAM. Sin embargo, dado que el objetivo de este trabajo es la predicción, más que la estimación de funciones, se opta por dar una nueva capa de suavizamiento con β. No existe forma de garantizar ortogonalidad de β contra las todas las w, por lo tanto, se le da prioridad a la correcta estimación de w pues captura un mayor efecto además de que, por la construcción de los polinomios por partes, si está garantizada la ortogonalidad contra las funciones bases Ψ.