Como base fundamental de este trabajo, a continuación, se expondrá a detalle el modelo. El objetivo, es construir un clasificador binario flexible con buena fuerza predictiva. La notación se irá explicando conforme aparece pero existe un compendio en el Apéndice **??**. En general, se trata de respetar la notación que usan en los libros (**hastie2008elements**) y (**james2013introduction**)

Se supone la siguiente estructura en los datos:

- $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ con n el tamaño de la muestra.
- $y_i \in \{0, 1\} \quad \forall i = 1 \ldots, n$ variables de respuesta binarias o *output*.
- $\mathbf{x}_i \in \mathcal{X}^d \subseteq \mathbb{R}^d \quad \forall\, i = 1 \ldots, n$ covariables, regresores o *input*.
- $d \in \mathbb{N}$ dimensionalidad de mis covariables.

El modelo en si, se presenta a continuación de forma general para cualquier pareja de datos $(y, \mathbf{x})$:

$$y \,|\, z \sim \mathrm{Be}(y \,|\Phi(z)) \tag{1}$$

$$z \,|\, x \sim \mathrm{N}(z \,|f(\mathbf{x}), 1) \tag{2}$$

$$f(\mathbf{x}) \approx \sum_{j=0}^{d} \beta_j f_j(x_j) \tag{3}$$

$$f_j(x_j) \approx \sum_{l=1}^{N^*} w_{j,l} \Psi_{j,l}(x_j, \mathcal{P}_j) \quad \forall j = 0, 1, \ldots, d \tag{4}$$

En las expresiones (1) y (3), dejando de un lado ecuación (2), se tiene una versión ligeramente modificada de un GLM (Sec. 0.1). Esto, pues la variable de respuesta y es binaria modelada con una distribución Bernoulli. Además, (3) es una función de proyección lineal como las que se usan el los modelos tradicionales. En el contexto de un modelo probit, esta función f, busca separar el espacio d-dimensional de covariables $\mathcal{X}^d$ en regiones identificables en una sola dimensión $\mathbb{R}$. Esta función de proyección, asume que la dependencia entre mis covariables se puede modelar como la suma ponderada de los componentes f_j (Sec. 0.2). Para poder hacer la liga entre ambas ecuaciones, se requiere de la incorporación de una variable latente z, vista en la ecuación (2), esta variable es meramente estructural y será modelada a través de una distribución normal, lo cual lleva a tener un modelo probit. Finalmente (4) hace una transformación no lineal de cada dimensión j y trata de encontrar las tendencias individuales de cada una de las covariables. Esto se logra, haciendo un suavizamiento por medio de polinomios por partes que dependen de 3 objetos: una partición del intervalo $\mathcal{P}_j$, un vector de pesos w_j y parámetros que captura la N^* especificando la

forma y grado de los polinomios. La forma funcional de Ψ es compleja y relativamente arbitraria dependiendo de la selección de la base, por lo tanto, no se especifican aún y se deja para la Sección 0.3. Se hace notar que el componente bayesiano se explora hasta el Capítulo **??** pues va estrechamente ligado con su implementación. En la Figura 1 se hace una representación visual del modelo para su mejor comprensión.

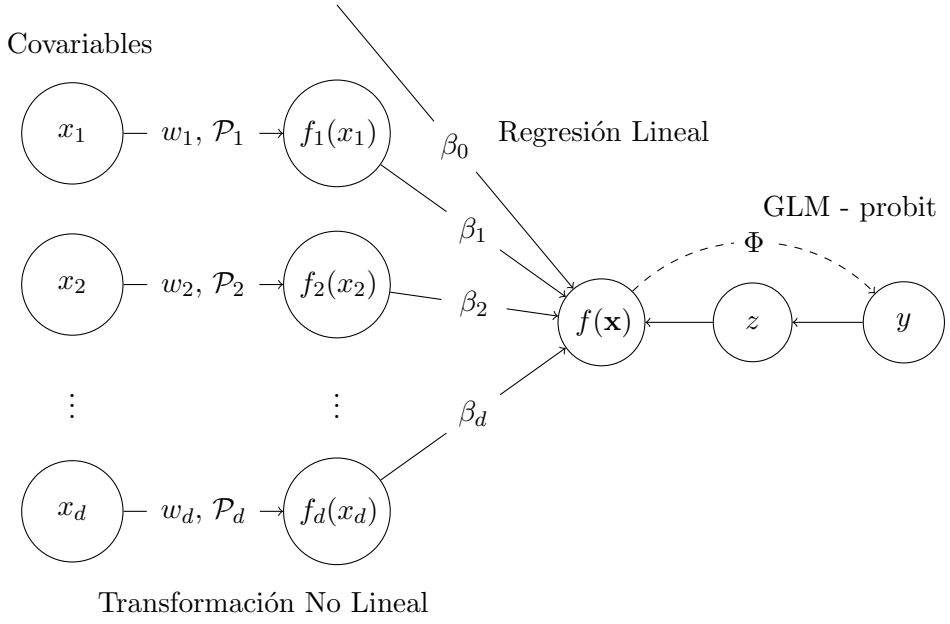

Figura 1: **Diagrama del modelo.** Se hace una transformación no lineal de las covariables x_j a través de los parámetros w_j y P_j. Con los datos transformados f_j, se lleva a cabo un modelo probit con función liga Ψ para lograr la clasificación binaria en y.

Antes de continuar, vale la pena recordar que:

> *All models are wrong but some are useful*[1]

Escoger un modelo que explique perfectamente los datos o que logre predecir todo sería una tarea inútil. Sin embargo, no significa que no se pueda intentar discernir un patrón y es justamente lo que se busca con la construcción de este modelo. Además de entender a profundidad un modelo que sirve como base para modelos que se están usando en el mundo de la inteligencia artificial. En particular, este modelo tiene la ventaja que es flexible y, al menos en teoría, debería de servir para representar una gran cantidad de datos.

0.1. Modelos Lineales Generalizados (GLM)

Los modelos lineales generalizados, (**sundberg2016exponential**) y (**maccullagh1989generalized**), surgen como una generalización del modelo lineal ordinario $y = \beta^t x + \epsilon$ donde $y \in \mathbb{R}$. En esta generalización, se busca darle otros rangos a y pues tenemos casos donde está restringida a un subconjunto de

1. (**box1979robustnessinthe**)

los reales como lo es el caso binario. Sin embargo, este cambio vuelve el modelo más complejo y lleva a técnicas diferentes en la estimación de los parámetros β. Además, se pierde algo de la interpretabiliad del modelo[2]. Sin embargo, han resultado ser realmente útiles.

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

$$
\begin{aligned}
y &\sim F(\theta(x)) \\
z &= \beta^t x \\
\theta &= g^{-1}(z)
\end{aligned}
$$

con los siguientes tres elementos:

1. F: **Tipo de distribución** de la familia exponencial que describa el dominio de las respuestas y. Por ejemplo: Bernoulli si y es binaria, Poisson si $y \in \mathbb{Z}^+$ o una distribución Gamma si $y \in \mathbb{R}^+$

2. z: **Proyector lineal** que explique (linealmente) la variabilidad sistemática de tus datos. En el modelo tradicional $\dim(\beta) = d < n$.

3. g: **Función liga** que una la media (o los parámetros canónicos) θ de mi distribución con el proyector lineal. Es decir: $\theta(x) = \mathbb{E}[y|x] = g^{-1}(\beta^t x)$.

   Como ejemplos clásicos se tiene la función $\text{logit}(p) = \ln(p/(1-p))$ o la $\text{probit(p)} = \Phi^{-1}(p)$, donde $p = \mathbb{E}[y|x]$ y $\Phi(\cdot)$ es la función de acumulación de una distribución normal estándar. En la Figura 2 podemos ver una representación gráfica para su mejor comprensión.

Para este trabajo, se busca construir un clasificador binario por lo que $y \in \{0, 1\}$, por lo cual, es natural modelar y como una distribución Bernoulli. Notese que si $Y \sim \text{Be}(y|p)$ se tienen las siguientes propiedades:

$$
\begin{aligned}
\mathbb{E}[Y] &= p = P(y = 1) \\
\mathbb{V}[Y] &= p(1-p)
\end{aligned}
$$

Por lo que solo se tiene un parámetro p y la varianza queda determinada automáticamente. Además, esta especificación deja como opción para las función liga, a las inversas de las funciones *sigmoidales* $s(x)$.

2. Dependiendo de la especificación, su interpretación puede ser complicada. Por ejemplo, cuando se tiene un modelo logit tradicional, se logra expresar el logaritmo de la proporción de probabilidades (*Log-Odds-Ratio*) como una combinación lineal de las covariables. $\ln(\pi_i/\pi_0) = \beta^t x$

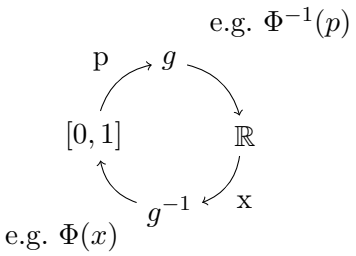

Figura 2: **Esquema de función liga** g

Las funciones sigmoidales, son funciones $s : \mathbb{R} \to (0, 1)$, estrictamente monótonas y por ende, biyectivas. Algunos ejemplos son las ya mencionadas logit, probit y la curva de Gompertz[3]. Estas funciones cumplen un papel de activación, es decir, una vez que se rebase cierto umbral, crecen rápidamente y toman valores más cercanos a uno, *activando* asi la probabilidad de que y sea un éxito. [4]. Esto las hace perfectas herramientas para ligar el proyector lineal $z \in \mathbb{R}$ con una probabilidad $p \in [0, 1]$.

0.1.1. Uso de la Variable Latente

Ahora, para entender el papel que juega z, 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 nos ayuda a discernir un vínculo entre ellas. En la Figura 3 tenemos esa representación del modelo.

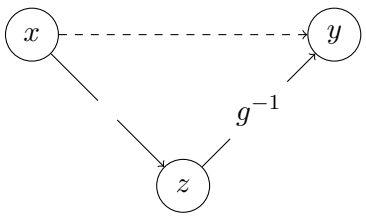

Figura 3: **Modelo de variable latente**

Tradicionalmente, la normalidad en z es derivada de asumir normalidad en los errores; es decir, dada la regresión lineal $z = \beta^t x + e$ se asume (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

3. Para no caer en redundancia de notación para este trabajo se tiene a partir de ahora: $s(x) = g^{-1}(x) = \Phi(x)$

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

frequentista, 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 cerrada. Sin embargo, tiene justamente las propiedades que se necesitan pues 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.[5]. 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 $\mathcal{X}^d$ nos encontramos. Esto se debe a que $y = 1$ cuando $\Phi(z) > 1/2$ que sucede *si y solo 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, buscamos 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 *estocasticidad* al modelo. Supongamos 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 un volado.

Otro factor importante a considerar, es que el modelo asume 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

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

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 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 hablemos del papel de la 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.

$$P(y=1) = p(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = g^{-1}(f(\mathbf{x})) = \Phi(f(\mathbf{x})) \tag{5}$$

0.2. Función de proyección f

Tradicionalmente, se asumía que z es una combinación lineal de los parámetros β y las covariables x, pero, como se explica en la pagina 6 de (**james2013introduction**), conforme avanzaron los métodos y el poder computacional disponible se fueron desarrollando técnicas cada vez más poderosas que permitían romper la linealidad. En 1986, Hastie y Tibshirani introducen los Modelos Aditivos Generalizados (GAM), una clase de modelos donde se lleva a romper la linealidad en las covariables, esto permite flexibilizar aún el modelo. Como se vio anteriormente, los GLM siguen la forma especificada en la Sección: 0.1, sin embargo, al tratar el modelo con la variable latente $z|x \sim N(f(\mathbf{x}), \sigma^2)$ con función liga probit, se tiene la ecuación (5).

Esta nueva *capa* no hace otra cosa más que ir colapsando dimensiones, ie: $\mathcal{X}^d \to \mathbb{R} \to [0,1]$. Es por esto que se llama función de proyección, pues *proyecta* el espacio $\mathcal{X}^d$ en $\mathbb{R}$. Sin embargo, la forma en la que lo hace debe de ser muy sutil 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 tengamos éxitos y negativa donde tengamos fracasos; para ello, es fundamental entender los GAM.

0.2.1. Modelos Aditivos Generalizados (GAM)

Un GAM como lo introducen Hastie y Tibshirani en (**hastie1986generalized**) remplaza la forma lineal $\sum_1^d \beta_i x_i = \beta^t x$ con una suma de funcione suaves $\sum_i^d f_i(x_i)$. Estas funciones no tienen una forma cerrada y son no especificadas, es decir, no hay tienen una forma funcional concreta y representable algebraicamente. Donde recae la fuerza del modelo, es que se estiman usando técnicas de suavizamiento no paramétricas[6] como lo sería un Suavizamiento Loess. En estos modelos, se asume que por más grande que sea $\mathcal{X}^d$, la relación que existe entre cada una de las dimensiones se puede explicar de manera aditiva. Esta especificación fue revolucionaria pues no solo regresa interpretabilidad al modelo, sino que simplifica

6. 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 *All of Non-parametric statistics*(**wasserman2007all**).

la estimación usando técnicas prácticamente automáticas con el algoritmo de *backfitting*. Además, Los GAM, demuestran que son efectivos en descubrir efectos no lineales en las covariables.

Para este trabajo, se escoge la ecuación (3):

$$f(\mathbf{x}) \approx \sum_{i=0}^{d} \beta_i f_i(x_i)$$

Esta función se escoge respondiendo a que este modelo, se planea usar para aplicaciones en econometría, donde las dimensiones i's son independientes entre si y corresponden a diferentes *características* o variables con las que se planea modelar la variable de respuesta. Por ejemplo, se puede pensar que cada $x_i \quad \forall i = 1, \ldots, d$ como d series de tiempo con las que se planea predecir cuando vender o comprar cierta acción.

La f es una versión versión modificada de un GAM tradicional con tres cambios fundamentales. La primera modificación es que estamos ponderando cada f_i por un parámetro β_i, esto es, para suavizar aún más cada dimensión y captar el patrón general y no tanto los componentes individuales de cada x_i. Se puede pensar en cada f_i como una transformación no-lineal de x_i (como lo sería una transformación logarítmica o una transfomación Box-Cox) por lo que se puede dar una interpretación al parámetro β_i como el efecto que tiene la dimensión i en particular para el modelo. Se hace notar que se deja espacio para un término independiente β_0. Por convención $f_0(\cdot) = 1$ por lo que se puede re-expresar la ecuación anterior (3) como:

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

Donde usando notación vectorial $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ y $F \in \mathbb{R}^{d+1}$. 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 ecuivalente la parametrización de z con una normal estándar.

La segunda modificación, es una más sutil, aunque es muy práctico manejar las f_i's como indeterminadas y estimarlas con procedimientos de suavizamiento no parametricos, también se puede optar por la vía en la que se especifica su forma funcional, no por ello quitandoles flexibilidad. Los detalles de esto lo dejaremos para la Sección 0.3 donde se trata de adapart el procedimiento de (**mallik1998automatic**). Esta modificación, obedece a que para ciertas aplicaciones, sirve hacer el modelo paramétrico en donde

cada f_i se modela en su expansión de bases y se puede hacer el ajuste con un algoritmo de minimos cuadrados. (Vease Capitulo 9.1 y Ejemplo 5.2.2 de (**hastie2008elements**)).

La tercera modificación, es que estamos trabajando con una aproximación en vez de una igualdad para f. Se hace notar que esta 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_i's. En la siguiente sección se explora el porqué de esta aproximación usando técnicas de análisis más avanzadas.

Imaginar o peor aún, visualizar f es complicado, pero el objetivo es que (si se tienen datos continuos) f agrega los efectos de cada una de las componentes y separa el espacio en regiones de éxitos y fracasos. Es decir, si tenemos puntos en $\mathbb{R}^2$, f se podrá visualizar en $\mathbb{R}^3$ como una serie de picos y valles donde será positiva en caso de ser éxito y negativa en caso contrario.

Finalmente, notemos que este modelo se podría confundir con un Modelos en Bases de Funciones Lineales como lo presentado en Capitulo 3 en (**bishop2006pattern**) donde se tiene:

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{d} \beta_i f_i(\mathbf{x}) \tag{6}$$

La diferencia radica en que cada f_i es función de todas mis covariables en lugar de solo la dimensión i. A estas funciones se les conocen como *funciones base* y son lineales en β, pero no-lineales para $\mathbf{x}$. Se hará una exposición más a detalle en la Sección 0.3 pero, las posibilidades son ilimitadas para estas funciones, algunos ejemplos son:

- **Bases Gaussianas:**
$$f_i(\mathbf{x}) = \exp\left\{-\frac{(\mathbf{x}-\mu_i)^2}{2s^2}\right\}$$
- **Funciones sigmoidales:**
$$f_i(\mathbf{x}) = \sigma\left(\frac{\mathbf{x}-\mu_i}{s}\right)$$

Sin embargo, estos son un grupo de modelos completamente diferente cuyas aplicaciones usualmente son en estimación de curvas y no tanto inferencia como lo busca este trabajo.

0.3. Funciones f_i

Finalmente se trata la parte más profunda del modelo, las funciones f_i que, como se mencionó anteriormente, son transformaciones no lineales de cada componente x_i que buscan suavizar la nube de datos, para posteriormente sumarlas entre si y dar una medida f que resuma la información. 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* como lo visto en la ecuación (6). Toda la siguiente sección se concentra en darle formas funcionales a las Ψ's. Se usa como referencia el captiulo 5 de (**hastie2008elements**).

Una expansión en bases de una función $h : \mathbb{R}^d \to \mathbb{R}$ es:

$$h(\mathbf{x}) = \sum_{j=1}^{J} w_j \Psi_j(\mathbf{x}) \tag{7}$$

Donde, $\Psi_j(\mathbf{x})$ es la j-ésima transformación no lineal de $\mathbf{x}$ y una vez especificadas y estimadas, el procedimiento (hacia arriba en el modelo) se hace de forma tradicional pues recobra su estructura lineal. Algunos ejemplos son:

- $\Psi_j(\mathbf{x}) = x_j$ donde $j = 1, \ldots, d$ y se tiene el modelo lineal más sencillo.
- $\Psi_j(\mathbf{x}) = \ln x_j$ ó $x_j^{1/2}$ donde se tienen transformaciones no lineales en cada una de las covariables.
- $\Psi_j(\mathbf{x}) = ||\mathbf{x}||$ una transformación lineal de todas las covariables.[7]

Dependiendo del tipo de datos y de aproximación que se busque, puede ser conveniente usar forma sobre la otra; existen muchas más posibles expansiones de bases. 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 estos. El método más común, es tomar una familia de funciones como los son los polinomios por partes o familias de funciones flexibles que logren representar una gran variedad de patrones. En estos métodos, se cuenta con una gran cantidad de funciones base por lo que se requiere controlar la complejidad; las formas más comunes de lograrlo son:

- **Métodos de Restricción**: como los son los métodos aditivos usados en este trabajo.
- **Métodos de Selección**: como lo son los modelos CART y MARS.

7. Como se vio en la ecuación (6)

- **Métodos de Regularización**: donde se busca controlar los coeficientes, como los son los modelos *Ridge* y *LASSO*.

Simplificando un poco la exposición, por lo pronto, se puede pensar únicamente en funciones reales, por lo que se deja de usar el subíndice i para indicar el componente del vector $\mathbf{x}$.

Para este trabajo, se aplica el procedimiento de (**mallik1998automatic**). Los autores presentan un método revolucionario, que permite estimar con un alto grado de precisión relaciones funcionales entre la variable de respuesta y y el regresor $x \in \mathbb{R}$. Se puede pensar que se busca ajustar una curva tradicional. Esto es, para un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$ se plantea el modelo:

$$y_i = h(x_i) + e_i \quad i = 1, \ldots, n \tag{8}$$

con las e_i errores aleatorios de media cero. Este método, combina los procedimientos paramétricos y no paramétricos desarrollados antes 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. La genialidad del su trabajo es que estos nodos, tradicionalmente fijos, se vuelven parámetros a estimar, usando un paradigma bayesiano. Y no solo 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.

0.3.1. Polinomios por partes y splines

Antes de llegar a estos polinomios tan flexibles, se busca entender que son los polinomios por partes simplificando la exposición 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 y que a cada intervalo le corresponde una función Ψ_j. Con estos nodos, se puede representar a la función global h en su expansión de bases como en la ecuación (7), donde cada Ψ_j es una función que depende de la partición y de x. Por ejemplo, se puede pensar en un caso sencillo donde se tiene que $J = 3$ y se quiere ajustar funciones constantes en cada intervalo. Entonces, las funciones base correspondientes serían:

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

Con $I(\cdot)$ la función indicadora que vale 1 si x se encuentra en la región y 0 en otro caso. Por lo tanto,

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

Lo cual es una función *escalonada*, en el sentido de que para cada región de x tenemos un nivel w_j.[8] Esta aproximación a mis datos podría servir para datos que estén agrupados por niveles, sin embargo, rara vez será ese el caso.

Entre cada pareja de nodos, se puede buscar ajustar un polinomio de grado arbitrario. Adicionalmente, se pueden construir polinomios con restricciones como 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* queramos, entendido como la continuidad de la K-ésima derivada. Tradicionalmente, se construyen polinomios cúbicos con segunda derivada continua en los nodos. Esto, pues resulta en funciones suaves al ojo humano que logran aproximar una gran cantidad de funciones.

Orígenes y justificación de su uso

La palabra *spline*, [9] se usa para designar a este grupo de polinomios por parte. Sin embargom dependiendo de como se definan pueden denotar funciones muy diversas; hasta ahora no hay consenso en la literatura. Para este trabajo se denota a un *spline de grado M* como un polinomio por partes de grado $M - 1$ con continuidad hasta la $M - 2$ derivada. (**wasserman2007all**). Se hace notar, que se tienen definiciónes de splines muy diferentes a las presentadas aquí, todo está en la definición de la partición que además, puede ser tan flexible como se requiera, por ejemplo los B-Splines.

8. 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 constatne, es el promedio de los puntos de esa región.

9. A diferencia de el texto tradicional de (**deboor1978splines**)

Los splines, surgen en (**schoenberg1964spline**) donde se plantea el problema: encontrar h en el espacio de Sovolev W_{M-1} de funciones con $M-2$ 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, ie: $h(x_i) = h_i \quad i = 1, 2, \ldots, n$. Sin embargo, se hace notar que sea como sea la especificación, se tiene un problema con la la naturaleza global de los polinomios, es decir, se necesita controlar lo que pasa más allá de los nodos de la frontera. Por lo que usualmente se escogen condiciones adicionales o linealidad pasando los nodos. En un contexto estadístico, el problema (8) se puede plantear como encontrar la función h que minimice:

$$\sum_{i=1}^{n}(y_i - h(x_i))^2 + \lambda \int_a^b (h^{(m)}(x))^2\,dx \tag{9}$$

para alguna $\lambda > 0$, donde la solución se demuestra que son *splines naturales* que se estudian más adelante, específicamente *splines cúbicos naturales* si $m = 2 \quad (M = 4)$. 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 procedimiento conocido como *regularización*. No es el enfoque entrar a detalle en cada uno de estos pues merecen una tesis por si mismas, sin embargo, se definen en el Apéndice **??**. Por lo pronto, lo esencial, es que al tratar de minimizar el RSS se puede caer en problemas de sobreajuste en donde los parámetros no estén capturando efectos y patrones subyacentes, sino solo estén tratando de seguir los datos. Para compensar, se penaliza el modelo con segundo termino que controla la complejidad del modelo y la suavidad deseada mediante la λ. Esto se logra, incorporando un termino de penalización el cual crece a medida que h se vuelve más complicada.[10]

Una vez más, Hastie y Tibshirani, con su modelo aditivo y función de perdida cuadrática con penalización en la segunda derivada:

10. 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 $\mathcal{L}_2([a, b])$. Más detalles de esto en la Sección **??**

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

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

más la convención de que h_0 es una constant y las λ_j los parámetros de suavizamiento, muestran que $h_j \quad j = 1, \ldots, d$ son splines cúbicos. Sin embargo, el modelo no es identificable pues h_0 puede ser arbitraria. Por lo que se necesita una restricción adicional para que el mínimo sea único, esta es:

$$\sum_{i=1}^{n} h_j(x_{ij}) = 0 \quad \forall j \tag{10}$$

es decir, las funciones promedian en cero sobre los datos. Lo que nos lleva a que: $h_0 = \bar{y}$. Por lo que si viéramos cada dimensión j, tendríamos que las h_j estarían centradas alrededor de la media $\bar{y}$. Además, la implementación de esta formulación, es conocida como *Algoritmo Backfitting* que será revisado en el Capítulo **??**.

Formalización matemática de splines

Retomando la discusión de la página 11, se busca definir un polinomio de grado $M-1$ por partes en J intervalos. Tomando una expansión de bases para cada intervalo, como en el ejemplo anterior, el número de funciones base aumenta en J por cada grado que se agregue, dando un total de $J*M$ bases funcionales, y en consecuencia, el mismo número de parámetros por estimar. Esto ocurre porque necesitamos definir $\mathcal{B}_j = \left\{1, x, x^2, \ldots, x^{M-1}\right\}$ para cada j. Sin embargo, esto llevaría a polinomios que se comportan de forma independiente en cada intervalo y no se conectan. La primera condición que se le impone es continuidad en los nodos, lo cual devuelve $J-1$ parámetros que corresponden a los $J-1$ nodos donde se da la continuidad. Cada grado de continuidad en las derivadas que se le pida al polinomio, se restringe el modelo y por ende, el número de funciones bases necesarias a un total de:

$$N^*(M, J, K) = M * J - K * (J - 1) \tag{11}$$

donde K es el número de restricciones para cada nodo ($K \leq M - 1$), definiendo hasta que número de derivada es continua. Independientemente de la base que escojamos, $N*$ será la *dimensión mínima de*

la base es decir, el número de funciones necesarias para representar un polinomio en función de M que define su grado, el número de intervalos J, por ende el número de nodos y K. Por lo pronto, se centra la discusión cuando $K = M - 1$ regresando a la definición de spline: polinomios de grado $M - 1$ con continuidad hasta la $M - 2$ derivada. Por ende, la dimensión queda: $N^* = M + J - 1$

Ahora, para definir la expansión de bases, se define la función auxiliar *parte positiva*:

$$x_+ = \text{máx}\{0, x\}$$

quedando una expansión en bases truncada:

$$\begin{aligned} h(x) &= \sum_{i=1}^{M+J-1} w_i \Psi_i(x, \mathcal{P}) \\ &= \sum_{i=1}^{M} w_i \, x^{i-1} + \sum_{i=1}^{J-1} w_{M+i} \, (x - \tau_i)_+^{M-1} \end{aligned} \tag{12}$$

El primer sumando de (12), representa el *polinomio base* 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 se mueve 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 cúbico en todo $[a, b]$. Se hace notar, que esta derivación de las bases, surge cuando se integra un polinomio por partes constante $M - 1$ veces. En cada iteración, se juntan las constantes y se integran por si solas, independientemente de los intervalos, lo cual lleva a este *polinomio base*. De forma más explicita, tenemos las bases:

$$\Psi_1(x, \mathcal{P}) = 1$$
$$\Psi_2(x, \mathcal{P}) = x$$
$$\vdots$$
$$\Psi_M(x, \mathcal{P}) = x^{M-1}$$
el *polinomio base*
$$\Psi_{M+1}(x, \mathcal{P}) = (x - \tau_1)_+^{M-1}$$
$$\vdots$$
$$\Psi_{M+J-1}(x, \mathcal{P}) = (x - \tau_{J-1})_+^{M-1}$$
la base truncada

las cuales forman un espacio lineal de funciones $(M + J - 1)$-dimensional.

A estos splines, se les conoce como splines cúbicos y son los más usados cuando se buscan funciones suaves.

A pesar de la utilidad de los splines por su suavidad, todos sufren de problemas más allá del rango de entrenamiento. Su naturaleza global hace que, fuera de la región con nodos, los polinomios crecen o decrecen rápidamente. Por lo tanto, extrapolar con polinomios o splines es peligroso y podría llevar a estimaciones erróneas. Para corregir esto, en ocasiones, se puede imponer la restricción de que el polinomio deba ser lineal más allá de los nodos frontera. Para designarlos, se les agrega el adjetivo de *natural*. Esta modificación, libera $2 * (M - 2)$ funciones bases, pues quita todas las bases de orden mayor a 1 en los dos nodos frontera. Adicionalmente, es razonable que se mejore la fuerza predictiva fuera de el dominio de entrenamiento. Todo depende de los datos y el tipo de funciones que se esté tratando de estimar. Su expansión en bases, también se deriva de la ecuación (12).

Hasta ahora, se han usando los parámetros M, J y K para definir el número de funciones base N^*, ecuación (11), pero también, sirven para definir los *grados de libertad* que se tienen. Esto se debe a que no solo nos dicen el número de funciones bases y la dimensión del espacio lineal, sino que nos indican el número de parámetros w_j's a estimar.

Otra consideración, es que, al estar trabajando en espacios funcionales, la elección de base es relativamente arbitraria y podemos intercambiarla como lo haríamos con una transformación de coordenadas en un

espacio euclidiano. Cada base tiene sus beneficios y simplicidad. Aquí se escoge una expansión de bases truncadas pues es explicada fácilmente y tiene una forma funcional relativamente sencilla. Sin embargo, no es óptima computacionalmente cuando J es grande. En la practica, usualmente se implementan B-Splines[11] que se derivan de los vistos anteriormente.

0.3.2. Polinomios por parte flexibles

Independientemente de la selección de parametros en la construcción del polinomio, se tiene el problema de la selección de los nodos. Existen procedimientos adaptativos, como los propuestos en (**friedman1991multivariate**). Sin embargo, y como ya se mencionó más atrás, en 1998, Denison, Mallik y Smith proponen un método bayesiano más atractivo.

Para poder explicar su método, se tiene que hacerle una modificación a la ecuación (12) para convertirla, de un spline, a un polinomio por partes más general con grado arbitrario de continuidad en las derivadas. En su expansión de bases se tiene:

$$h(x) = \sum_{i=1}^{N^*} w_i^* \ \Psi_i(x, \mathcal{P}) \qquad \text{con } N^* = J * M - K * (J-1) \tag{13}$$

$$= \sum_{i=1}^{M} w_{i,0} \ x^{i-1} + \sum_{i=K}^{M-1} \sum_{j=1}^{J-1} w_{i,j} \ (x - \tau_j)_+^i \tag{14}$$

Dado que se tiene una doble suma, es necesario incluir un segundo índice, al menos temporalmente, a los pesos. De, modo que el primer índice, denotado por i refleja el grado asociado a su término[12]. Por lo tanto, si $i = 2$ entonces, $w_{2,j}$ está asociado a una término de grado 1. El segundo índice j denota al nodo al que está asociado el peso. Como convención, si $j = 0$, se hace referencia al *polinomio base* que siempre tendrá efecto. En el segundo sumando de (14) 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 la primera derivada, etc. En el caso de que $K = M - 1$ regresamos a la ecuación (12) y tenemos una vez más splines que, por construcción, son suaves. La suavidad, aunque importante, no siempre es requerida. 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 constucción, con su doble suma,

11. Vease el Capítulo 5.5 de (**wasserman2007all**) o el Apéndice del Capítulo 5 en (**hastie2008elements**)

12. Desgraciadamente y para ser consistentes con la notación anterior, no se puede indexar directamente, es decir, se le tiene que restar 1 para obtener el grado en los primeros terminos, pero en los posteriores si es directo.

w_j^*	$w_{n,m}$	$\Psi_j(x,\mathcal{P})$		
Subíndice j	Subíndices n,m	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
$M+2$	$K+1,1$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(M-K)$	$M-1,1$	$(x-\tau_1)_+^{M-1}$		
$M+(M-K)+1$	$K,2$	$(x-\tau_1)_+^K$	$M-K$	
$M+(M-K)+2$	$K+1,2$	$(x-\tau_1)_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+2(M-K)$	$M-1,2$	$(x-\tau_1)_+^{M-1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-2)(M-K)+1$	$K,J-1$	$(x-\tau_{J-1})_+^K$	$M-K$	
$M+(J-2)(M-K)+2$	$K+1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		
$\vdots$	$\vdots$	$\vdots$		
$M+(J-1)(M-K)$	$M-1,J-1$	$(x-\tau_{J-1})_+^{K+1}$		

Figura 4: Biyección entre w_j^*, $w_{n,m}$ y sus correspondientes funciones base Ψ_j

permite tener $M-K$ términos por nodo, codificando así las continuidades en las derivadas[13]. En la ecuación (13) se usa w_i^* solamente para denotar que se puede seguir expresando como una combinación lineal. Finalmente, hágase que $h(x)$ sea igual a $f_i(x_i)$. Con este cambio de notración,(13) es equivalente a (4). Este era el último componente fundamental por definir del modelo, completando así su exposición.

En la Figura 4 de la página 18, se hace un compendio de los polinomios por partes. Esto ayuda no solo a esclarecer las cosas, sino a formar una biyección entre w_i^*, $w_{n,m}$ y Ψ_i que posteriormente ayudará a expresar todo de forma matricial en su implementación en código.

Por lo tanto, se termina teniendo $N^* = M + (J-1)(M-K) = JM - K(J-1)$ términos una vez más. Y por construcción, como se vio anteriormente, la biyección, es consistente con la definición en (12) para el caso especifico que $K = M-1$.

13. Esta codificación es sutil pues, al hacer los cálculos de continuidad, tenemos 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

Antes de cerrar el capítulo, se centra la atención en los nodos τ. A estos, se les ha dado poca importancia hasta el momento, pues se han considerado como fijos. Como ya se mencionó antes, en (**mallik1998automatic**) se desarrolla, además de la ecuación (14) un paradigma bayesiano para que los nodos, sean tratados como parámetros y por ende sus posiciones son variables. 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 plana para alguna sección, se necesitan pocos nodos. En el Capitulo **??**, se continua con esta exposición y se detalla el proceso para la selección de la posición de los nodos. Sin embargo, cabe recalcar que a diferencia del trabajo original donde el número de nodos era variable, en este trabajo se usa J fija. Esto corresponde a que se busca simplificar el algoritmo sin tener que usar métodos que aumenten o disminuyan el número de dimensiones. En la practica, la J se tiene que calibrar, sin embargo, no ha resultado ser un problema adicional pues normalmente, se busca suavizar más que estimar estimar funciones especificas complejas como era el objetivo del trabajo original.

Consideraciones finales

Al tener en mente que se tienen d covariables, y por ende d polinomios por partes, además de la estructura lineal de (13) podemos sustituir (4) dentro de (3) y se tiene la siguiente estructura con doble suma:

$$\begin{aligned} f(\mathbf{x}) &\approx \sum_{i=0}^{d} \beta_i f_i(x_i) \\ &\approx \beta_0 + \sum_{i=1}^{d} \beta_i \left[\sum_{j=1}^{N^*} w_{i,j} \Psi_{i,j}(x_i, \mathcal{P}_i) \right] \end{aligned}$$

Lo cual, es perfectamente lineal. Se tienen $1 + d * N^*$ términos y se pueden acomodar en un solo vector. Sin embargo, se tiene un cruce de parámetros interesante, la multiplicación de la $\beta_i \quad \forall i$ contra $w_{i,j} \quad \forall j$. Tradicionalmente, no se usan β's y se deja que se capture ese efecto dentro de las f_i como en los modelos aditivos normales. 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 las β's. No existe forma de garantizar ortogonalidad de las β's contra las w's, por lo tanto, se le da prioridad a la correcta estimación de w pues captura un mayor efecto además de que, por construcción de los polinomios por partes, si está garantizada la ortogonalidad contra las funciones bases Ψ'.