Pasar de un modelo tan estructurado a su implementación computacional no resulto fácil. Sin embargo, se logró desarrollar un algoritmo que estima todos los parámetros del modelo de una forma eficiente y que funciona en la práctica. En el fondo, el algoritmo recae en el método de Gibbs sampling propuesto en (**albert1993bayesian**), por lo que se hace una breve introducción a la escuela de inferencia bayesiana, y en el algortimo de backfitting descrito en (**hastie1986generalized**). Al algoritmo se le titula: *bayesian piece wise polinomial model (bpwpm)*. Para facilitar la utilización del modelo en diversas bases de datos, así como su validación y visualización, a la par del algoritmo se desarrolló un paquete de código abierto (con el mismo nombre) para el software estadístico `R`. Al darle un tratamiento bayesiano a los parámetros, más que estimarlos, se busca regresar una muestra de tamaño arbitrario de sus correspondientes distribuciones posteriores. La idea, es que estas distribuciones posteriores, se haya capturado toda la información de los datos de entrenamiento.

Se considera, que una buena forma de entender el algoritmo es *visualizando* tanto los datos como los objetos que componen el modelo, por lo tanto se hace un paréntesis notacional. De las ecuaciones del modelo: (**??**) a (**??**), se tienen dos grupos de parámetros por estimar, $\beta \in \mathbb{R}^{d+1}$ y $w_j \in \mathbb{R}^{N^*} \quad \forall j = 1, \ldots, d$. Donde:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} \quad \text{y} \quad w_1 = \begin{bmatrix} w_{1,1} \\ \vdots \\ w_{1,N^*} \end{bmatrix} \quad \ldots \quad w_d = \begin{bmatrix} w_{d,1} \\ \vdots \\ w_{d,N^*} \end{bmatrix}$$

Se hace énfasis en que existen d vectores w_j, cada uno de tamaño N^*. Por lo tanto, se tienen sun total de $1 + d + dN^*$ parámetros. Se usa el símbolo $\mathbf{w}$ para designar todos los vectores w_j, haciendo de este una matriz, es decir: $\mathbf{w} \in \mathbb{R}^{d \times N^*}$. Cuando se habla de datos: $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, estos se pueden representar en una tabla (o matriz):

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

Donde el vector de observaciones binarias $\mathbf{y} = (y_1, \ldots, y_n)^t$ es la primer columna de la tabla, y la matriz de covariables $\mathbf{X}$ es el resto.

Bajo esta representación, se da contexto cuando se habla de que la estimación debe reflejar los patrones *hacia abajo* y *hacia lo largo*. Hacia abajo, se está captando la información existente entre las observaciones; cada f_j, mediante su parámetro w_j, representa una transformación no lineal de la variable (o dimensión) j. Hacia lo largo, la función de proyección f suma cada f_j a través de β, ponderando los efectos individuales

de cada variable. Mantener el balance entre la estimación de β a lo largo y w_j hacia abajo, es fundamental para el algoritmo. Analizando este hecho, se concluye que la estimación de ambos grupos de parámetros, se puede ver como una regresión separada para cada uno, y por ende, estos pueden ser estimados por el mismo algoritmo. Esto responde a la dualidad que se exploró en el capítulo pasado de que ambas expresiones son expansiones en bases funcionales. El puente que conecta, y controla el balance entre ambas, son los residuales parciales. Los siguientes capítulos, se concentran en explicar e implementar este curioso patrón.

0.1. Fundamentos de la estadística bayesiana

Dado el problema de describir fenómenos bajo incertidumbre, existen dos escuelas dominantes de la estadística: la frecuentista y la bayesiana. La primera, aunque increíblemente útil, está hasta cierto punto limitada y en ocasiones termina derivando en colecciones de algoritmos. La teoría bayesiana, por el contrario, nombrada así en honor a Thomas Bayes (1702 - 1761), es una rama que enfatiza el componente *probabilista*, dando coherencia al proceso de inferencia (**mendoza2011estadistica**) y (**bernardo2001bayesian**). La estadistica bayesiana está axiomatizada bajo la *teoría de la decisión*. Esta teoría formaliza conceptos económicos como la *coherencia entre preferencias* y *utilidad*, sobre los que desarrolla un marco metodológico para la toma de decisiones.

Esta metodología, además de proveer técnicas concretas para resolver problemas, también formaliza en una forma de pensar sobre la probabilidad como una *medida racional para cuantificar la incertidumbre* condicionando sobre el conocimiento existente. Este paradigma es el que más corresponde con el sentido que usualmente se le da a la palabra. La inferencia sobre creencias (o parámetros), se realiza mediante una *actualización* de estas en luz de nueva evidencia, modificando su medida de incertidumbre. El mecanismo que permite realizar esto, es la aclamada formula de Bayes. De manera informal se puede describir como: dado un evento E bajo condiciones C, la probabilidad *posterior* del evento, es proporcional a la probabilidad *previa* que se tiene sobre este, ponderado por la probabilidad de ocurrencia de las condiciones presentes, es decir:

$$P(E|C)\ \alpha\ P(C|E)P(E) \tag{1}$$

El término central $P(C|E)$ es una medida descriptiva de las condiciones (usualmente datos) llamada *verosimilitud*. Se hace notar que para poder hacer cualquier intento de descripción, se debe especificar el *modelo probablistico* que se asume describe el estado por el que se dan las condiciones C.
En un contexto matemático más formal, la cuantificación de la incertidumbre se da a través de medidas de

probabilidad $\pi(\cdot)$, que describan el fenómeno observado. Estas medidas de probabilidad, usualmente son funciones que dependen de cantidades desconocidas llamados parámetros θ. Aunque desconocidas, se tienen ciertas creencias u conocimiento previo, *a priori*, sobre ellos, descritos por su correspondiente medida de probabilidad $\pi(\theta)$. Además, se tienen datos $\mathbf{X}$, interpretados como *evidencia*, a los cuales se les asigna un modelo de probabilidad dependiente de los parámetros, es decir, su verosimilitud: $\pi(\mathbf{X}|\theta)$. Usando la formula de Bayes, podemos actualizar el conocimiento que se tiene sobre los parámetros haciendo:

$$\pi(\theta|\mathbf{X}) \; \alpha \; \pi(\mathbf{X}|\theta)\pi(\theta)$$

La idea es que este proceso de actualización sea a la vez, un proceso de aprendizaje, en el cual los parámetros capturen la información contenida en los datos.

La teoría frecuentista, adopta un enfoque diferente para el aprendizaje. Se asume que no hay incertidumbre en los parámetros dado los datos y, por lo tanto, estos son tomados como fijos. El mecanismo que permite su estimación, usualmente consiste en plantear una función objetivo y optimizarla. Por ejemplo, si se escoge la verosimilitud $\pi(\mathbf{X}|\theta)$, se busca dar un estimador que la maximice, pues equivaldría a encontrar los parámetros que hagan más *posibles* los datos, bajo el modelo planteado. Si por el contrario, es escoge una función como la RSS de los modelos ANOVA (primer sumando de (**??**)), se busca la θ que minimice estos errores, así, el modelo logra capturar toda la variabilidad que puede sobre los datos. Independientemente del paradigma estadístico que se escoja, siempre es importante la validación del modelo y de sus supuestos. Además, tanto teoría bayesiana como frequentista han resultado de infinita utilidad en la practica y el avance de la estadística y ciencia en general.

Una de las dificultades que surgen en la estadística bayesiana, es que la obtención de resultados analíticos cerrados es difícil o muy tedioso una vez que los modelos se empiezan a complicar. Por ejemplo, en las ecuaciones anteriores, se ha usado el argumento de proporcionalidad α. Esto pues, para que se de la igualdad, el lado derecho de la ecuación (2) se debe de dividir entre $\pi(\mathbf{X}) = \int \pi(X|\tilde{\theta})\pi(\tilde{\theta})\, d\tilde{\theta}$, el cual usualmente es difícil, sino imposible, de calcular. A este término se le conoce como *constante de proporcionalidad* y su función es la de reescalar la expresión del lado derecho para que en realidad se tenga una distribución en el izquierdo. Usualmente, se escoger distribuciones *conjugadas*, para que tanto la distribución a priori como la posterior sean de la misma familia y por ende conocidas. Sin embargo, con los avances en el poder computacional disponible y técnicas numéricas para resolver integrales, se han desarrollado muchos métodos para aplicar el proceso de aprendizaje, independientemente de que tan complejo sea el modelo o las distribuciones, iniciales y resultantes. Muchos de estos métodos recaen en la

teoría de las *cadenas de Markov*, como lo es, el Gibbs sampler presentado en la sección. 0.2

Estimadores Bayesianos

Una vez realizado el proceso de actualización, el estadista se enfrenta con un problema. Se tiene una distribución posterior de probabilidad para los parámetros de interés, usualmente dada por una muestra y no por una distribución analítica. Sin embargo, por practicidad y utilidad, en ocaciones se busca dar un *estimador puntual*. Por ejemplo, si se necesita dar un estimador $\hat{\theta}$ para usarlo en otros cálculos, o si $\pi(\theta|\mathbf{X})$ es multidimensional. Para superar este problema teórico, se ha adoptado por usar *funciones de perdida* $L(\hat{\theta}, \theta)$[1]. Estas, miden las *consecuencias* que se dan, al tomar $\hat{\theta}$ como el verdadero valor del parámetro θ, es decir, las funciones de perdida evalúan que tan bien se está representando el valor de θ con un estimador puntual. Por ello, vale la pena usar funciones que penalicen la distancia entre θ y $\hat{\theta}$. Sin entrar mucho en los detalles técnicos, se tiene que calcular:

$$\hat{\theta} = \mathbb{E}[L(\hat{\theta}, \theta)] = \int_{\Theta} L(\hat{\theta}, \theta)\pi(\theta)\, d\theta \tag{2}$$

con Θ el espacio de todas las posibles valores de θ. Sin embargo, se demuestra que para funciones de perdida sencillas, pero intuitivas, se tiene que el estimador puntual posterior es alguna medida de centralidad de la distribución posterior. Por ejemplo:

- Función de pérdida cuadrática: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, deriva en la media posterior, es decir: $\hat{\theta} = \mathbb{E}[\theta|\mathbf{X}]$
- Función de perdida valor absoluto: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, deriva en la mediana de la distribución posterior.
- Función de pérdida 0-1: $L(\hat{\theta}, \theta) = I[\hat{\theta} \neq \theta]$, deriva en la moda de la distribución posterior.

En la práctica, estas cantidades son fáciles de calcular cuando se tiene una muestra simulada de θ proveniente de la distribución posterior. En el paquete, se implementa una forma sencilla de obtener estimadores puntuales con cualquiera de las 3 funciones de pérdida. Sin embargo, se verá que los resultados no varían mucho. Ver Apéndice **??**.

0.1.1. Funciones de probabilidad condicional completas

Retomando el modelo que concierne a este trabajo, se tienen dos grupos de parámetros, β y $\mathbf{w}$. Sin embargo, dados los supuestos del modelo, por el uso de la variable latente z, esta también se debe de incluir como parámetro pues es la liga entre la respuesta y y los datos $\mathbf{X}$, pero vista de forma bayesiana, también se debe de simular. Por lo tanto, quedan los parámetros: $\theta = (\mathbf{z}, \beta, \mathbf{w})$ con $\mathbf{z} = (z_1, \ldots, z_n)^t$.

1. Formalmente se tiene un problema de decisión.

Esta sección concierne desglosar el proceso de aprendizaje sobre ellos; esta derivación es importante en si pues es la que induce el algoritmo. Usando la notación presentada al inicio de esta sección, los supuestos propuestos en las ecuaciones del modelo (**??**) a (**??**) y sustituyendo en (2) se tiene:

$$\begin{aligned}
\pi(\mathbf{z}, \beta, \mathbf{w}|\mathbf{y}, \mathbf{X}) \quad &\alpha \quad \pi(\mathbf{y}|\mathbf{X}, \mathbf{z}, \beta, \mathbf{w})\ \pi(\mathbf{z}, \beta, \mathbf{w}) && (3)\\
&\alpha \quad \pi(\mathbf{y}|\mathbf{z})\ \pi(\mathbf{z}|\mathbf{X}, \beta, \mathbf{w})\ \pi(\beta, \mathbf{w}) && (4)\\
&\alpha \quad \pi(\mathbf{y}|\mathbf{z})\ \pi(\mathbf{z}|\mathbf{X}, \beta, \mathbf{w})\ \pi(\beta)\ \pi(\mathbf{w}) && (5)\\
&\alpha \quad \prod_{i=1}^{n} \mathrm{Be}[y_i|\Phi(z_i))]\ \phi[z_i|f(\mathbf{x}_i), 1] \times \pi(\beta)\ \pi(\mathbf{w}) && (6)
\end{aligned}$$

donde $\phi(\cdot|\mu, \sigma^2)$ es la función de densidad de una variables aleatoria normal con media μ y varianza σ^2, asimismo $\mathrm{Be}(\cdot|p)$ es función de densidad de una variable Bernoulli con probabilidad de éxito p. Esta factorización es válida dados los supuestos, donde se hace notar la forma que conecta $\mathbf{z}$ a las dos partes del modelo a través de la función de proyección f que contiene tanto a β como $\mathbf{w}$. Además, se presenta una forma extendida (aunque simplificada) de la verosimilitud para todas las observaciones $i = 1, \ldots, n$. Aunque aún no se han especificado las formas funcionales para las distribuciones a priori $\pi(\mathbf{w})$ y $\pi(\beta)$, estas se pueden separar ya que se asumen independientes. Esta propiedad, combinada con la forma funcional en expansiones de bases, lleva a que se piense en hacer una estimación *por bloques*, es decir, se estima primero β y posteriormente $\mathbf{w}$ en un bucle iterativo, pues esta es la idea de un Gibbs sampler.

0.2. Simulación bayesiana: el Gibbs sampler

- Explicar que el Gibbs sampler es un método MCMC. Directo del paper de Chibb

0.2.1. Algoritmo de Albert y Chibb

- Hacer derivación de la condicional para β paper de Albert + Chibb - Poner pseudocódigo - Argumentar por qué es igual para w pero en lugar de usar las z de regresores se usan los residuales parciales. - Mencionar que en el paper se hacen más algoritmos para diferentes cosas

0.2.2. Especificación probabilistica para el modelo

Para los parámetros, se usan las siguientes distribuciones *apriori*:

$$\beta = (\underline{e}ta_0, \beta_1, \ldots, \beta_d)^t \sim N_d(\mu_0, \Sigma_0) \tag{7}$$

$$w^{(i)} = (w_1^{(i)}, \ldots, w_J^{(i)})^t \sim N_J(\mu_0^{(i)}, \Sigma_0^{(i)}) \quad i = 1, \ldots, d \tag{8}$$

0.3. Algoritmo *bpwpm*

En forma de pseudocódigo el algoritmo tiene la siguiente forma:
A diferencia de la exposición del modelo, el algoritmo debe de construir de abajo hacia arriba, pues se necesita tener una estimación puntual de los parámetros para poder calcular las funciones intermedias y que todo quede definido de forma numérica.

```
Parametros inciales:
WHILE (...)
    Transformación de X -> Phi -> F (Función: estimate_PWP)
    Simulación de betas (Función simulate_beta)
```

- Hacer énfasis en el apéndice y el paquete.

0.3.1. Algoritmo de *backfitting* para ajuste de modelos GAM

- Justificación final para las w's
- Usamos los nodos iniciales en cuantiles determinados.

- taus: HMC
- β Estimar por máxima verosimilitud pero dentro del Gibbs con el método ABC
- $w's$ BAyesianas + importantes que las betas.

- Explicar Alortimo y hacer pseudocódigo de cada sección - Explicar lógica del algoritmo - Explicar desarrolo de paquetes en R - Explicar bien la parte de los residuales y el algorimto backfitting, por que las f_j son arbitrarias y pueden interpolar a los residuales para hacer el ajuste. Esto también explica las β pues si se pueden capturar chigón los residuales con una sola dimensión, te vale verga la siguiente :). Yei bitches

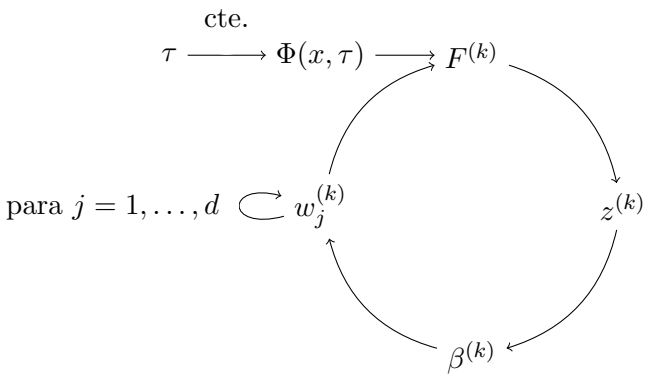

Figura 1: Esquema del algoritmo