Trabajo Fin de Máster

Modelos de predicción y clasificación con alta dimensión en el número de covariables

Laura Freijeiro González

Máster en Técnicas Estadísticas

Curso 2017-2018

Propuesta de Trabajo Fin de Máster

Título en galego: Modelos de Predicción e Clasificación con alta dimensión no número de covariables

Título en español: Modelos de Predicción y Clasificación con alta dimensión en el número de covariables

English title: Prediction and Clasification models with high dimensional data in the covariates number

Modalidad: Modalidad A

Autora: Laura Freijeiro González, Universidad de Santiago de Compostela

Directores: Wenceslao González Manteiga, Universidad de Santiago de Compostela; Manuel Febrero Bande, Universidad de Santiago de Compostela

Breve resumen del trabajo:

El ámbito de los modelos de regresión y de clasificación ha sido muy estudiado desde el principio de los desarrollos metodológicos en la Estadística. Desde sus inicios, con los clásicos modelos de regresión lineal múltiple y la regla lineal discriminante de Fisher, hasta los años más recientes, con los estudios que cubren los modelos no lineal, de tipo no paramétrico o semiparamétrico para ambos contextos. Más recientemente el fenómeno de la cantidad masiva de datos o los llamados datos de alta dimensión, generó adaptaciones o modificaciones de los modelos al contexto en el que el número de covariables sea muy amplio, incluso en proporción al tamaño muestral de la muestra de partida. El desarrollo a llevar a cabo en esta propuesta de trabajo fin de máster es revisar las diversas adaptaciones más importantes de los distintos modelos de regresión y clasificación al contexto en el que el número de covariables sea comparable al número de datos.

Recomendaciones:

Se recomienda haber cursado los tópicos “Modelos de Regresión”, “Modelos de Datos Funcionales”, “Estadística Matemática” y el “Análisis Multivariante” del máster de Técnicas Estadísticas.

Don Wenceslao González Manteiga, Catedrático de Universidad de la Universidad de Santiago de Compostela , y don Manuel Febrero Bande, Catedrático de Universidad de la Universidad de Santiago de Compostela , informan que el Trabajo Fin de Máster titulado

Modelos de predicción y clasificación con alta dimensión en el número de covariables

fue realizado bajo su dirección por doña Laura Freijeiro González para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En santiago, a 03 de julio de 2018.

El director:

Don Wenceslao González Manteiga

El director:

Don Manuel Febrero Bande

La autora:

Doña Laura Freijeiro González

Índice general

Resumen IX

Prefacio XI

1. Regresión en alta dimensión. 1
1.1. Modelo lineal múltiple. Problemas en alta dimensión . . . 1
1.1.1. Regresión Ridge . . . 10
1.1.2. Regresión LASSO . . . 15
1.1.3. Regresión con penalización Elastic Net . . . 17
1.1.4. Regresión LAR . . . 18
1.2. Modelos aditivos. Problemas en alta dimensión. . . . 20
1.2.1. Estimación de las funciones $\boldsymbol{f_j}$ . . . 22
1.2.2. Regularización L_2 . . . 27
1.2.3. Regularización L_1: modelo SpAM . . . 28
1.3. Regresión lineal generalizada. Problemas en alta dimensión . . . 30
1.3.1. Regularización L_2 . . . 40
1.3.2. Regularización L_1 . . . 41
1.3.3. Regularización Elastic Net . . . 42
1.3.4. Regresión LAR . . . 42
1.4. Regresión logística. Problemas en alta dimensión . . . 42
1.4.1. Regularización L_2 . . . 48
1.4.2. Regularización L_1 . . . 49
1.4.3. Regularización Elastic Net . . . 49
1.4.4. Regresión LAR . . . 50
1.5. Modelos aditivos generalizados. Problemas en alta dimensión . . . 50
1.5.1. Estimación de las funciones $\boldsymbol{f_j}$ . . . 50
1.5.2. Regularización L_2 . . . 56
1.5.3. Regularización L_1: modelo SpAM generalizado . . . 56

2. Clasificación en alta dimensión 59
2.1. Análisis Discriminante . . . 59
2.1.1. Reglas discriminantes . . . 60
2.1.2. Criterios de elección de la regla discriminante . . . 61
2.1.3. Análisis Lineal Discriminante (LDA): Regla lineal de Fisher. Problemas en alta dimensión . . . 61
2.1.4. Regla discriminante cuadrática (QDA). Problemas en alta dimensión . . . 64
2.1.5. Estimación de las reglas discriminantes . . . 64
2.2. Análisis discriminante regularizado . . . 67
2.3. La regresión logística como regla discriminante . . . 68
2.4. Reglas de clasificación con matrices de covarianzas diagonales . . . 69
2.4.1. Regla de clasificación estimada de matriz de covarianzas diagonal . . . 69
2.4.2. Encogimiento por centroides cercanos . . . 71
2.5. Métodos de clasificación no paramétricos . . . 72
2.5.1. Regla de clasificación de los K-vecinos más cercanos . . . 72
2.5.2. Regla de clasificación de las K-medias . . . 74

2.6. Support Vector Machine (SVM) 75
2.6.1. Clasificador SVM para dos poblaciones. Extensión al caso de L poblaciones 76
2.6.2. Cálculo de los SVM generalizados para la clasificación en dos poblaciones. Extensión al caso de L poblaciones 80

Bibliografía **83**

Resumen

Resumen en español

En este trabajo se lleva a cabo un análisis teórico tanto de los Modelos de Regresión usuales, como de las reglas de clasificación clásicas del Análisis Discriminante en el contexto de la alta dimensión, especialmente cuando se cuenta con un número de covariables comparable al número de datos ($p > n$). En primer lugar se muestra el desarrollo habitual y las características de los algoritmos, así como su ajuste e implementación, para posteriormente introducir los problemas que surgen en este ámbito con el fin de estudiar su repercusión y proponer soluciones o alternativas.

Referente a los Modelos de Regresión se analizan tanto las versiones lineales (*modelo lineal múltiple* y *modelo lineal generalizado*) como aquellas con estructura no lineal (*modelos aditivos* y *modelos aditivos generalizados*). En todos los casos se exponen los inconvenientes que surgen en el contexto de interés, tanto debido al mal comportamiento de los estimadores como a la dificultad de selección de variables, proponiendo alternativas que solucionan estos inconvenientes.

Por otra parte, en el caso del Análisis Discriminante, se muestran las reglas usuales de discriminación (*regla discriminante lineal* y *regla discriminante cuadrática*), explicando las trabas que afrontan estas cuando se estiman en este nuevo marco, debidas principalmente al mal condicionamiento de las matrices de covarianzas. Con el fin de solucionar este problema se propondrán enfoques de carácter tanto paramétrico, es decir manteniendo sus hipótesis iniciales, como no paramétricos, los cuales serán válidos en multitud de contextos.

English abstract

This master thesis carries out a theoretical analysis of the usual Regression Models as well as the classic classification rules of the Discriminant Analysis in a high dimensional or *Big Data* context. This takes into account the special case when the covariates number is high in comparison with the amount of available data ($p > n$). Firstly, their algorithm developments as well as their features and characteristics are shown, so as their estimation and implementation. Then, emerging problems in this new context are introduced, studying their impact and proposing solutions or alternatives.

In reference to Regression Models, this will analyze the lineal versions (*multiple lineal models* and *generalized lineal models*), so as those with non-lineal structure (*additive models* and *generalized additive models*), exposing disadvantages which come out in the recent context. Next, it shows how these are mainly due to the bad behavior of common estimators or because of the large volume of covariates. Eventually, penalized versions which are able to settle these inconveniences are proposed.

On the other hand, in the case of Discriminant Analysis, the usual rules are presented (*lineal discriminant rule* and *quadratic discriminant rule*), explaining what problems these methods have to deal with when these are estimated in a high dimensional framework, like the ill-conditioned covariance matrices. In order to solve this problem several parametric approaches will be introduced, keeping the initial hypothesis or assuming non parametrical assumptions.

Prefacio

El avance de diversos campos y medios tecnológicos hace que cada vez sea mayor el volumen de datos con el que nos encontramos diariamente. Algo tan usual como internet produce cada segundo una inmensa cantidad de información a la que cada usuario contribuye. Otros campos como los datos económicos en bolsa, el estudio genómico en sectores biológicos o sanitarios, la información suministrada por ordenadores en un ámbito informático o los datos proporcionados por los clientes de una empresa no son más que ejemplos comunes de esta situación.

Proporcionalmente al aumento de información surge la necesidad de conocer cómo emplearla adecuadamente y conseguir extraer conclusiones. La estadística busca proporcionar medios y respuestas a diversos problemas reales, con el fin de averiguar qué se puede esperar que ocurra y qué toma de decisiones podría ser la adecuada, entre otros fines. La necesidad de saber manejar y aprovechar la información se va haciendo más necesaria progresivamente, de forma que nace el término *Big Data* abriendo un nuevo campo de estudio estadístico.

El *Big Data* se define comúnmente a través de las denominadas tres llaves: volumen, velocidad y variedad. Se considera un problema de *Big Data* cuando cumple alguna o varias de dichas facetas, donde el volumen, característica que impulsa el nombre, designa que el conjunto de información a tratar es de grandes dimensiones informativas. Uno de los principales problemas surge a la hora de trabajar con esta gran cantidad de datos o cuando las características de cada vector de datos individuales, son muy próximas o superiores en número al conjunto de muestras recogidas. La velocidad, por su parte, caracteriza la rapidez con la que se generan y tratan nuevos datos a tiempo real, surgiendo así el problema de incorporar las nuevas observaciones instantáneamente, ya que estas podrían generarse en minutos o segundos. Por consiguiente, agregar la nueva información de una forma rauda al proceso de estimación del modelo conseguirá que este sea más preciso y útil haciéndolo más competitivo. Por último, la variedad expresa la diversidad de formatos en los que puede obtenerse la información, ya sea a través de imágenes, mensajes, sensores, señales de GPS u otros modos. Muchos intelectuales han denominado este nuevo fenómeno como el microscopio del siglo XXI, considerando el *Big Data* la herramienta que permite extraer de una enorme nube de información las características claves para conseguir entenderla, analizarla y por consiguiente aprovecharla.

Este trabajo se va a centrar en dos tipos de modelos importantes que pueden surgir en un análisis de datos dentro de un entorno *Big Data*: los modelos de regresión y las reglas de clasificación del análisis discriminante en el caso de la Alta Dimensión, esto es, cuando el número de variables disponibles es más grande que el número de casos. El impacto de la alta dimensión en los modelos tradicionales de la regresión o del análisis discriminante puede acarrear complicaciones, ya que además de la dificultad computacional añadida debida a la gran cantidad de datos, estos pueden dejan de funcionar de forma adecuada, surgiendo la necesidad de disponer de modificaciones o alternativas. Por tanto, en este documento se va a ahondar en la situación de alta dimensión donde las características de cada elemento o covariables, p, son en cantidad superiores en número al conjunto de datos recogidos, n, siendo el contexto en el cual estos modelos presentan problemas, el cual se denotará por $p > n$.

En primer lugar se revisan los modelos más tradicionales de regresión y discriminación, desarrollando y explicando la teoría matemática que permite obtenerlos y justifica su correcto funciona-

miento. A continuación se explica cual es la causa de que dejen de funcionar estos procedimientos en una situación de alta dimensión, especialmente en el contexto de interés donde $p > n$. Seguidamente se muestran soluciones, métodos nuevos que permiten realizar la regresión y el análisis discriminante, explicando en que consisten y analizando sus ventajas.

En el caso de los modelos de regresión se estudian tanto las versiones lineales como las no lineales, permitiendo cualquier tipo de estructura entre las variables explicativas y la variable respuesta que se desea explicar y de la cual se quieren extraer conclusiones. En ambos contextos se mostrarán los problemas que son necesarios afrontar cuando el número de covariables es similar o más grande al número de muestras, tales como las malas propiedades de los estimadores usuales o la falta de interpretabilidad del modelo ajustado, debido al gran número de covariables que entran en juego, haciendo latente la necesidad de contar con algoritmos que permitan seleccionar únicamente las variables importantes y deshacerse de las restantes. Ante estos problemas se mostrarán algunas soluciones que mitigan ambos problemas, mediante la inclusión de versiones regularizadas.

Por otro lado, en el caso del análisis discriminante se expondrán las reglas de clasificación clásicas, tanto en el caso de igualdad de covarianzas entre clases como cuando esta hipótesis no se cumple, mostrando los problemas existentes en el contexto de alta dimensión al querer estimar dichas reglas. Para aliviar estos inconvenientes se propondrán diversas alternativas, tanto manteniendo las hipótesis iniciales de las que parten las reglas discriminantes lineal o cuadrática, como proporcionando otros métodos de carácter no paramétrico que permiten realizar clasificación en cualquier contexto. Se obtiene de esta forma otros algoritmos que permiten clasificar una nueva muestra en uno de los grupos determinados de antemano en el caso de que el número de muestras sea menor que el de variables explicativas, $p > n$.

Capítulo 1

Regresión en alta dimensión.

Los tradicionalmente conocidos como modelos de regresión permiten explicar una variable de interés, como puede ser el beneficio o rendimiento de una empresa o el nivel de glucosa de un paciente diabético, en base a un determinado conjunto de variables que guardan relación con esta, como serían, por ejemplo, cierto tipo de características relacionadas con la actividad de la empresa o determinadas medidas sanitarias del paciente, respectivamente. De esta forma se construye un modelo que satisface dos propósitos. El primero es realizar predicciones de la variable de interés en base a las explicativas, lo que puede interpretarse como, una vez conocido el valor de las características de la empresa o las medidas del paciente, o ante diferentes valores posibles de estas, se es capaz de obtener una predicción del beneficio o rendimiento medio así como del nivel de glucosa que es de esperar en cada situación. El segundo propósito por su parte, muestra como cada una de las características de la empresa influye en el beneficio, al igual que cada variable medida en el paciente influye en su nivel de glucosa, haciendo latente cuáles se consideran más o menos importantes para obtener un buen resultado o para detectar una subida o bajada importante de glucosa.

En un contexto de alta dimensión, los modelos de regresión usuales para datos vectoriales dejan de funcionar de forma adecuada y por tanto la implementación de su construcción a través de los métodos tradicionales deja de ser eficiente. Así mismo ante bases de datos con un gran número de covariables aparece otro problema adicional que es determinar qué subconjunto de datos es el más relevante a la hora de explicar una determinada variable, ya que la consideración de todas las covariables haría que se perdiera totalmente la interpretabilidad del modelo de regresión ajustado.

En este capítulo se presentarán diversos modelos de regresión, recogiendo estructuras tanto lineales como no lineales para contextos multidimensionales, donde el número de covariables puede ser tan o más grande que el número de muestras del que se dispone. Se empezará mostrando las características de cada uno de estos modelos, así como su ajuste e implementación además de características y propiedades varias. A continuación se explicará qué ocurre en el contexto de alta dimensión, donde estos dejan de funcionar de forma adecuada o surgen problemas a la hora de dictaminar qué variables son las más relevantes, ligados estos últimos con la interpretabilidad eficiente del modelo. Este escenario se corresponde con el caso donde el número de covariables es mayor que el de muestras ($p > n$) y se propondrán alternativas para conseguir solucionar estos problemas.

1.1. Modelo lineal múltiple. Problemas en alta dimensión

El **modelo de regresión lineal múltiple** es una extensión del modelo lineal simple, el cual sirve para expresar la dependencia de una variable continua Y con respecto a un conjunto de variables $(X_1, \ldots, X_p) = X$ de forma lineal como bien indica su nombre. Dicha variable Y es denominada *variable respuesta* mientras que las $\{X_1, \ldots, X_p\}$ son las *variables explicativas*, las cuales aportan información del comportamiento de la variable inicial. Estos modelos buscan conocer la dependencia de la variable respuesta respecto a cada una de las explicativas y además, pretenden realizar

predicciones del valor de la variable Y en base a unos valores prefijados de las variables explicativas.

Se distinguen dos tipos de diseño experimental del modelo según las variables explicativas: *diseño fijo* o *diseño aleatorio*. El *diseño aleatorio* se fundamenta en que tanto las variables explicativas como la variable respuesta son aleatorias, mientras que en el *diseño fijo* los valores de las variables explicativas están fijados antes de realizar el experimento. Consideraremos de ahora en adelante un diseño fijo.

La regresión se formaliza como la media condicionada de la variable respuesta en función del valor que tomen las variables explicativas:

$$m(x) = \mathbb{E}(\, Y \mid X = x\,) \quad x \in \mathbb{R}^p.$$

De esta forma se descompone la variable respuesta en función de su media condicionada más un error no observable ϵ

$$Y = m(X) + \epsilon.$$

Para poder expresar y construir este modelo se asumen tradicionalmente cuatro hipótesis: *linealidad*, *homocedasticidad*, *normalidad* e *independencia*.

La suposición de *linealidad* de la variable respuesta en torno a las variables explicativas está fundamentada en que se busca construir como función de regresión una función lineal, con lo que el modelo será de la forma

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

Al aceptar la *homocedasticidad* estamos suponiendo que la varianza del error será la misma para cualesquiera valores de las variables explicativas, es decir $\mathbb{V}(\epsilon|X = x) = \sigma^2$ para todo $x \in \mathbb{R}^p$. Por su parte, la *normalidad* se basa en que el error posee distribución normal $\epsilon \in N(0, \sigma^2)$ y por tanto la distribución de la variable Y será normal. Finalmente, la *independencia* nos garantiza que si tomamos una muestra de tamaño n

$$\left\{ \left(y_i, X_i = (x_{i1}, \ldots, x_{ip})^t\right) \right\}_{i=1}^n,$$

las variables que representan los errores $\epsilon_1, \ldots, \epsilon_n$ son mutuamente independientes y por tanto, nuestras variables $y_1, \ldots, y_n$ también serán mutuamente independientes.

De esta forma para el individuo i-ésimo el modelo será

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad \forall i \in \{1, \ldots, n\}$$

y lo podemos expresar de forma matricial por $y = \mathbf{X}\beta + \epsilon$:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \tag{1.1}$$

con $y, \epsilon \in \mathcal{M}_{n\times 1}$, $\mathbf{X} \in \mathcal{M}_{n\times(p+1)}$ y $\beta \in \mathcal{M}_{(p+1)\times 1}$. Donde ahora y y $\mathbf{X}$ serán respectivamente un vector y una matriz de valores conocidos, dados por los datos muestrales.

Estimación de β: método de mínimos cuadrados

A la hora de estimar los parámetros del modelo, es decir, el vector β y la varianza del error σ^2, el método más empleado es el de **mínimos cuadrados**, Figura 1.1. La filosofía de este método se basa en minimizar los errores cometidos al aproximar el valor de una variable por el obtenido con el modelo. Para este fin se busca minimizar las distancias entre la estimación dada por el modelo y los datos conocidos. De esta forma se toma como estimador $\hat{\beta}$ el valor que

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2 = \min_{\beta} (y - \mathbf{X}\beta)^t (y - \mathbf{X}\beta) = \min_{\beta} \|y - \mathbf{X}\beta\|^2 = \min_{\beta} \phi(\beta).$$

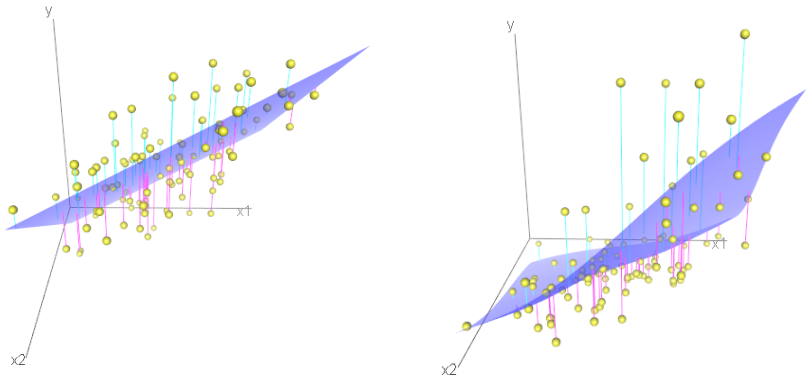

Figura 1.1: Ilustración del método de mínimos cuadrados en tres dimensiones. Este método se basa en ajustar una superficie (azul oscuro) que minimice la suma de los residuos al cuadrado tanto positivos (azul claro) como negativos (rosa). La superficie ajustada podrá tener carácter lineal (izquierda), como ocurre en el caso de la regresión lineal, o no lineal (derecha), al considerar otros contextos.

Denotando

$$\begin{aligned}\phi(\beta) :=& \|y - \mathbf{X}\beta\|^2 = (y - \mathbf{X}\beta)^t (y - \mathbf{X}\beta) = (y^t - (\mathbf{X}\beta)^t)(y - \mathbf{X}\beta) \\ =& (y^t - \beta^t \mathbf{X}^t)(y - \mathbf{X}\beta) = yty - \beta^t \mathbf{X}^t y - y^t \mathbf{X}\beta + \beta^t \mathbf{X}^t \mathbf{X}\beta,\end{aligned}$$

derivando e igualando a cero esta función para hallar sus mínimos se llega a

$\frac{\partial \phi(\beta)}{\partial \beta} = -y^t \mathbf{X} + \beta^t \mathbf{X}^t \mathbf{X} = 0 \Rightarrow \beta^t \mathbf{X}^t \mathbf{X} = y^t \mathbf{X} \Rightarrow \mathbf{X}^t \mathbf{X}\beta = \mathbf{X}^t y$

y se sabe que es un mínimo dado que

$\frac{\partial \phi^2(\beta)}{\partial \beta^2} = \mathbf{X}^t \mathbf{X}$ y $\det(\mathbf{X}^t \mathbf{X}) \geq 0$ por ser $\mathbf{X}^t \mathbf{X}$ semidefinida positiva.

De esta forma se obtiene la expresión que da las **ecuaciones normales de regresión**

$$\mathbf{X}^t \mathbf{X}\beta = \mathbf{X}^t y.$$

Por tanto habrá solución para las ecuaciones normales cuando exista $(\mathbf{X}^t \mathbf{X})^{-1}$

$$\hat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t y. \tag{1.2}$$

Además, dicho estimador es insesgado.:

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &\overset{(a)}{=} \mathbb{E}((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y) \overset{(b)}{=} (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{E}(y) \\ &\overset{(c)}{=} (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{E}(\mathbf{X}\beta+\epsilon) \overset{(d)}{=} (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbb{E}(\mathbf{X}\beta)+\mathbb{E}(\epsilon)) \\ &\overset{(e)}{=} (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\beta = \beta\end{aligned} \tag{1.3}$$

donde en (a) se usa (1.2), (b) se cumple porque $\mathbf{X}$ es una matriz de diseño fijo, en (c) se está aplicando el resultado visto en (1.1), en (d) las propiedades elementales de la media y en (e) que tanto $\mathbf{X}$ como β son una matriz y un vector con valores fijos respectivamente, además de que el vector de errores tiene como media el vector nulo.

El Teorema 1.1 garantiza que el estimador de mínimos cuadrados $\hat{\beta}$ (1.2) es el mejor estimador insesgado que se puede obtener en términos de minimización de la varianza.

Teorema 1.1 (Gauss Markov)**.** *Suponiendo que $m(X) \equiv \mathbb{E}[Y] = X\beta$ y $\mathbb{V}(X) = \sigma^2\mathbf{I}$ y sea $\tilde{\phi} = c^tY$ cualquier estimador lineal insesgado de $\phi = z^t\beta$, donde z es un vector arbitrario. Entonces:*

$$\mathbb{V}(\tilde{\phi}) \geq \mathbb{V}(\hat{\phi})$$

donde $\hat{\phi} = z^t\hat{\beta}$, siendo $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tY$ el estimador de mínimos cuadrados de β. Notar que, dado que z es arbitrario, este teorema implica que cada elemento de $\hat{\beta}$ es el mínimo estimador insesgado de la varianza.

Demostración. *Dado que $\tilde{\phi}$ es una transformación lineal de Y, $\mathbb{V}(\tilde{\phi}) = c^tc\sigma^2$. Para comparar las varianzas de $\hat{\phi}$ y $\tilde{\phi}$ es útil expresar $\mathbb{V}(\hat{\phi})$ en términos de c. Para hacer esto, basta tener en cuenta que $\tilde{\phi}$ es insesgado lo que implica que*

$$\mathbb{E}[c^tY] = z^t\beta \Rightarrow c^t\mathbb{E}[Y] = z^t\beta \Rightarrow c^tX\beta = z^t\beta \Rightarrow c^tX = z^t.$$

Por lo tanto la varianza de $\hat{\phi}$ puede ser escrita como

$$\mathbb{V}(\hat{\phi}) = \mathbb{V}(z^t\hat{\beta}) = \mathbb{V}(c^t\mathbf{X}\hat{\beta}).$$

Esto es la varianza de una transformación lineal de $\hat{\beta}$, y la matriz de covarianzas de $\hat{\beta}$ es $(\mathbf{X}^t\mathbf{X})^{-1}\sigma^2$, así que

$$\mathbb{V}(\hat{\phi}) = \mathbb{V}(c^t\mathbf{X}\hat{\beta}) = c^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tc\sigma^2 = c^t\mathbf{H}c\sigma^2$$

(donde $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ (1.6) es la matriz de influencia o matriz hat). Ahora las varianzas de ambos estimadores pueden ser directamente comparadas, y puede verse que

$$\mathbb{V}(\tilde{\phi}) \geq \mathbb{V}(\hat{\phi})$$

si

$$c^t(\mathbf{I}-\mathbf{H})c \geq 0$$

por la idempotencia de $(\mathbf{I}-\mathbf{H})$, pero esta última condición está diciendo que una suma de cuadrados no puede ser menor que 0, lo cual es claramente cierto.

Es necesario darse cuenta que este teorema usa la independencia y la suposición de igualdad de varianzas, pero en cambio no asume normalidad. Una interpretación de este resultado es que el estimador de mínimos cuadrados visto en (1.2) siempre tiene esta estructura, independientemente de la distribución que sigan los errores del modelo de regresión y además este siempre es el mejor estimador insesgado en términos de minimización de la varianza. De esta forma puede concluirse que aunque no se pueda asumir la hipótesis de normalidad siempre se puede obtener un estimador

insesgado con mínima varianza mediante el método de mínimos cuadrados.

También puede verse que el estimador $\hat{\beta}$ (1.2) obtenido por el método de mínimos cuadrados coincide con el que se obtendría a través de la **maximización de la verosimilitud**, siempre que se suponga normalidad en los errores. Es decir, este valor coincide con el más verosímil a la hora de estimar el parámetro β del modelo. Puede verse este resultado de forma sencilla, si se tiene en cuenta la suposición de normalidad del modelo por la cual la función de densidad del residuo vendría dada por

$$f(Y) = (2\pi\sigma^2)^{-n/2} e^{-\|Y - X\beta\|/(2\sigma^2)}.$$

De esta forma la función de verosimilitud tendría la expresión

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\|y - \mathbf{X}\beta\|/(2\sigma^2)} \tag{1.4}$$

donde y y $\mathbf{X}$ son respectivamente el vector y matriz dados por los datos muestrales. Si se maximiza la expresión de $L(\beta, \sigma^2)$ respecto de β, derivando e igualando a cero, se llega a que maximizar esta expresión respecto del parámetro β es equivalente a minimizar

$$S = \|y - \mathbf{X}\beta\|^2,$$

sin importar el valor de σ^2. De este modo se ve que se llega al mismo estimador para β que se obtiene mediante mínimos cuadrados.

Dado que los estimadores de máxima verosimilitud son eficientes y consistentes [1], se tiene que en el caso de asumir que los errores siguen una distribución normal, el estimador de mínimos cuadrados también posee dichas características, pues se ha demostrado que en este contexto ambos coinciden. Se ve así que este estimador cuenta con buenas propiedades.

A continuación se presentará el problema de estimación del vector de parámetros, β, cuando se trabaja en un contexto donde los datos vienen dados mediante un diseño aleatorio. Además, se verán escenarios donde no es posible asumir las hipótesis de homocedasticidad o independencia, estudiando cómo es la estimación del modelo en estos contextos.

Estimación bajo diseño aleatorio

Puede verse que si en vez de considerar diseño fijo se tuviese un diseño aleatorio, con las mismas hipótesis de partida, el estimador de β (1.2) seguiría siendo el mismo. Para demostrar este resultado basta comprobar que mediante el método de maximización de la verosimilitud bajo diseño aleatorio es necesario optimizar la misma función que en el caso de diseño fijo, lo cual hace que este estimador vuelva a coincidir con el que se obtiene por mínimos cuadrados. Por tanto, para probar este resultado se tiene en cuenta que ahora la función de densidad conjunta para (X, Y) viene dada por

$$f(X, Y) = f_X(X) f(\, Y \mid X \,),$$

[1] La eficiencia de un estimador garantiza que este posee la mínima varianza posible mientras que la consistencia asegura convergencia del estimador a su valor real cuando el tamaño muestral tiende a infinito.

obteniéndose que la función a maximizar es

$$L(\beta,\sigma^2)=\underbrace{\left[\prod_{i=1}^{n} f_x(x_i)\right]}_{f_x(x)}\underbrace{(2\pi\sigma^2)^{-n/2}e^{-\|y-\mathbf{X}\beta\|/(2\sigma^2)}}_{f(\,y|\mathbf{X}\,)},$$

donde $f_x(x)$ es una función constante en términos de β,σ^2. Por tanto se ve que la función a maximizar es $f(\,y\mid\mathbf{X}\,)$ la cual se corresponde con (1.4).

En consecuencia, el estimador del modelo GLM obtenido por mínimos cuadrados es el mismo tanto para diseño fijo como para diseño aleatorio, teniendo en cuenta que se están asumiendo las hipótesis de normalidad e independencia. La diferencia entre ambos estimadores $\hat{\beta}$ radica en la amplitud de sus intervalos de confianza, los cuales serán más grandes en el caso de diseño aleatorio.

Escenario heterocedástico

En el caso donde no se pueda asumir la hipótesis de igualdad de varianzas, puede verse que el estimador obtenido por máxima verosimilitud, o lo que es equivalente mediante mínimos cuadrados, se corresponde igualmente con $\hat{\beta}$ (1.2). Para esto basta tener ahora en cuenta que en este caso se tiene que

$$m(Y)=X\beta,\quad Y\sim N(X\beta,\mathbf{V}\sigma^2) \tag{1.5}$$

donde $\mathbf{V}$ es cualquiera matriz diagonal definida positiva, la cual se correspondería con $\mathbf{I}_p$ en el contexto homocedástico suponiendo independencia.

En este caso la ecuación de verosimilitud para β es

$$L(\beta,\sigma^2)=\frac{1}{\sqrt{(2\pi\sigma^2)^n|\mathbf{V}|}}e^{-(y-\mathbf{X}\beta)^t\mathbf{V}^{-1}(y-\mathbf{X}\beta)/(2\sigma^2)}$$

y si $\mathbf{V}$ es conocida puede verse que el estimador de máxima verosimilitud de β se alcanza minimizando la función

$$S_V=(y-\mathbf{X}\beta)^tV^{-1}(y-\mathbf{X}\beta),$$

la cual se corresponde con la respectiva expresión de mínimos cuadrados en este nuevo contexto. Además, puesto que $\mathbf{V}$ depende de parámetros desconocidos estos también pueden ser estimados mediante máxima verosimilitud y una vez obtenidos ser usados para calcular $\hat{\beta}$.

Datos dependientes

Otro contexto de interés es aquel en el cual los datos presentan algún tipo de dependencia entre sí y por tanto no es factible aceptar la hipótesis de independencia. Esto se traduce en que la matriz de covarianzas, tanto bajo homocedasticidad como heterocedasticidad, deja de ser diagonal.

De esta forma, la matriz $\mathbf{V}$ (1.5) será simétrica pero no diagonal en el caso de que las varianzas no sean iguales, mientras que tendrá estructura simétrica con unos en la diagonal en el caso de estar bajo homocedasticidad. Cuando no se cumpla la hipótesis de independencia, puede verse que se puede adaptar este contexto a uno similar al caso anterior, en el cual se suponía que los datos seguían una distribución normal multivariante con media desconocida y matriz de covarianzas conocida salvo una constante de proporcionalidad.

Si se denota $\mathbf{L}$ cualquiera matriz tal que $\mathbf{L}^t\mathbf{L} = \mathbf{V}$, por ejemplo obtenida a través de una descomposición de Choleski, se puede ver que

$$\begin{aligned} S_V &= (y - \mathbf{X}\beta)^t \mathbf{V}^{-1}(y - \mathbf{X}\beta) \\ &= (y - \mathbf{X}\beta)^t (\mathbf{L}^t\mathbf{L})^{-1}(y - \mathbf{X}\beta) \\ &= \|(\mathbf{L}^t)^{-1}y - (\mathbf{L}^t)^{-1}\mathbf{X}\beta\|^2, \end{aligned}$$

es la función objetivo de mínimos cuadrados la cual puede ser minimizada por los métodos clásicos de optimización, como por ejemplo a través de una descomposición QR de $(\mathbf{L}^t)^{-1}\mathbf{X}$.

Puesto que $(\mathbf{L}^t)^{-1}Y$ no es más que una transformación del vector normal aleatorio Y, este término sigue una distribución normal multivariante con esperanza $\mathbb{E}[(\mathbf{L}^t)^{-1}Y] = (\mathbf{L}^t)^{-1}X\beta$ y matriz de covarianzas

$$V_{(\mathbf{L}^t)^{-1}Y} = (\mathbf{L}^t)^{-1}\mathbf{V}\mathbf{L}^{-1}\sigma^2 = (\mathbf{L}^t)^{-1}\mathbf{L}^t\mathbf{L}\mathbf{L}^{-1}\sigma^2 = \mathbf{I}\sigma^2.$$

Por lo tanto $(\mathbf{L}^t)^{-1}Y \sim N((\mathbf{L}^t)^{-1}X\beta, \mathbf{I}\sigma^2)$. En otras palabras, la transformación ha resultado en un nuevo problema de modelado lineal en el cual la variable respuesta son datos aleatorios independientes que siguen una distribución normal, con varianza constante. De esta forma se cumplen todas las hipótesis que se habían supuesto inicialmente y se puede obtener la estimación del parámetro β para la construcción del modelo de regresión de la forma vista inicialmente.

Estimación σ^2

A la hora de conocer el valor de la varianza, σ^2, necesitaríamos conocer los errores del modelo, pero estos no están disponibles ya que son desconocidos. Para solventar este problema se aproximan los errores por los residuos y se calcula un estimador de la varianza en base a ellos.

Los errores de predicción, denominados **residuos** son aquellos que miden la discrepancia entre una predicción $\hat{Y}_i$ y el dato observado de la forma:

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - x_i\hat{\beta} \quad \forall i \in \{1, \ldots, n\}.$$

Se puede definir un vector de residuos

$$\hat{\epsilon} = y - \hat{y} = (\mathbf{I}_n - \mathbf{H})y = \mathbf{M}y \quad \text{siendo}$$

$$\begin{aligned} &\mathbf{M} = \mathbf{I}_n - \mathbf{H} \text{ la } \textbf{matriz generadora de residuos} \text{ y} \\ &\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \text{ la conocida como } \textbf{matriz de influencia} \text{ o } \textbf{matriz hat.} \end{aligned} \tag{1.6}$$

La matriz de influencia, $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, recibe este nombre dado que muestra la influencia que tiene cada valor de respuesta en cada valor ajustado. Los elementos de su diagonal son conocidos como los **leverage**, los cuales se encargan de recoger la influencia que tiene cada valor de la respuesta en cada uno de los valores ajustados para la misma observación. Estos efectos permitirán detectar qué datos del modelo se pueden considerar **influyentes**, entendiendo por estos aquellos datos que modifican en gran medida la forma del modelo ajustado, es decir obviando estos se tendría un modelo bastante distinto al que se obtiene ajustando este al incluirlos.

De esta forma, al no ser observables los errores, se usan los residuos para estimar la varianza σ^2. En consecuencia, el estimador de la varianza del error es

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}\hat{\epsilon_i}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 = \frac{RSS}{n-(p+1)} = \frac{\hat{\epsilon}^t\hat{\epsilon}}{n-(p+1)} \overset{(a)}{=} \frac{y^t\mathbf{M}y}{n-(p+1)} \tag{1.7}$$

donde (a) se deduce de que $\mathbf{M}$ es una matriz simétrica e idempotente.

De igual forma que ocurría con el vector de parámetros β, el estimador de la varianza existirá cuando exista $(\mathbf{X}^t\mathbf{X})^{-1}$.

Problemas en alta dimensión ($p > n$)

En el contexto de alta dimensión considerado, es decir, cuando el número de variables explicativas es similar o más grande que el número de muestras de las que disponemos ($p > n$), dado que $\mathbf{X} \in \mathcal{M}_{n\times(p+1)}$ y $\mathbf{X}^t\mathbf{X} \in \mathcal{M}_{(p+1)\times(p+1)}$; la matriz $\mathbf{X}^t\mathbf{X}$ es singular y por tanto no vamos a poder definir su inversa de forma única.

Corolario 1.2. *Sean $\mathbf{A} \in \mathcal{M}_{(p+1)\times n}$ y $\mathbf{B} \in \mathcal{M}_{n\times(p+1)}$ matrices con $p > n$, se tiene por tanto que*

$$\left.\begin{array}{c} rango(\mathbf{AB}) \leq rango(\mathbf{A}) \\ y \\ rango(\mathbf{AB}) \leq rango(\mathbf{B}) \end{array}\right\} \Rightarrow rango(\mathbf{AB}) \leq n$$

dado que $rango(\mathbf{A}) \leq n$ y $rango(\mathbf{B}) \leq n$ por ser $p > n$.

Usando el Corolario 1.2 se puede demostrar que $\mathbf{X}^t\mathbf{X}$ es singular cuando $p > n$:

Demostración. *Existirá la inversa de la matriz $\mathbf{X}^t\mathbf{X} \in \mathcal{M}_{(p+1)\times(p+1)}$, de forma única, cuando su determinante sea distinto de cero, lo cual equivale a que*

$$\exists(\mathbf{X}^t\mathbf{X})^{-1} \Leftrightarrow |\mathbf{X}^t\mathbf{X}| \neq 0 \Leftrightarrow rango(\mathbf{X}^t\mathbf{X}) = p+1.$$

Por el Corolario 1.2 se tiene garantizado que $rango(\mathbf{X}^t\mathbf{X}) \leq n < p$ de modo que no existe la inversa de la matriz $\mathbf{X}^t\mathbf{X}$.

Por tanto, surge el problema de que tal y como están formulados los estimadores por mínimos cuadrados en alta dimensión no existe unicidad de los mismos. A continuación se verá como modificar este método para poder obtener dichos estimadores y que consecuencias conlleva.

Dado que la matriz $\mathbf{X}^t\mathbf{X}$ es cuadrada, simétrica y semidefinida positiva por construcción, el *Teorema espectral* garantiza que existe una matriz ortogonal $\mathbf{U}$ y una matriz diagonal $\mathbf{D}$ tal que:

$$\mathbf{X}^t\mathbf{X} = \mathbf{UDU}^t = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} u_1^t \\ \vdots \\ u_n^t \end{pmatrix} \tag{1.8}$$

donde u_i es el autovector propio asociado al correspondiente autovalor $\lambda_i \geq 0$ de la matriz $\mathbf{X}^t\mathbf{X}$, $\forall i = 1, \ldots, n$. De (1.8) se sigue que

$$|\mathbf{X}^t\mathbf{X}| = |\mathbf{U}\mathbf{D}\mathbf{U}^t| \overset{(a)}{=} |\mathbf{U}||\mathbf{D}||\mathbf{U}^t| \overset{(b)}{=} |\mathbf{D}| = \prod_{i=1}^{n} \lambda_i$$
$$\text{por tanto} \quad \exists (\mathbf{X}^t\mathbf{X})^{-1} \text{ si } \lambda_i \neq 0 \quad \forall i \in \{1, \ldots, n\}$$

la igualdad (a) se deduce de las propiedades elementales de los determinantes, mientras que la (b) de que la matriz $\mathbf{U}$ es ortogonal ($|\mathbf{U}| = |\mathbf{U}^t| = 1$).

Cuando un autovalor de la matriz $\mathbf{X}^t\mathbf{X}$ es cero o muy próximo a cero ocasiona problemas, en el primer caso se deja de tener invertibilidad en la matriz (como caso particular esto sucede cuando $p > n$, pues no disponemos de información suficiente para que el rango de la matriz sea el total); mientras que en el segundo caso hace obtener un estimador $\hat{\beta}$ bastante inexacto. Para ver la interpretación de esto último, recordando por (1.3) que la esperanza del estimador es β, basta probar que la covarianza de $\hat{\beta}$ viene dada por $\mathbb{C}(\hat{\beta}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$:

$$\begin{aligned}\mathbb{C}(\hat{\beta}) =& \mathbb{C}(\hat{\beta}, \hat{\beta}) = \mathbb{C}((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y, (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y) \\ =& (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{C}(y, y)\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \overset{(a)}{=} (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\ =& \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\end{aligned}$$

en (a) se está empleando la hipótesis de homocedasticidad.

Como se vio en (1.8) la matriz $\mathbf{X}^t\mathbf{X}$ se podía descomponer en producto de matrices, dos de ellas ortogonales y otra diagonal. Es fácil ver que los autovalores de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ son los inversos de los obtenidos en dicha descomposición.

$$(\mathbf{X}^t\mathbf{X})^{-1} = (\mathbf{U}\mathbf{D}\mathbf{U}^t)^{-1} = (\mathbf{U}^t)^{-1}\mathbf{D}^{-1}\mathbf{U}^{-1} \overset{(a)}{=} \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^t$$

en (a) se tiene en cuenta que la matriz $\mathbf{U}$ es ortogonal y por tanto $\mathbf{U}^{-1} = \mathbf{U}^t$.

Se ve que la descomposición espectral de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ es $\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^t$ y por tanto que los autovalores de esta matriz son los inversos de los elementos que conformaban la diagonal de la matriz $\mathbf{D}$: $\frac{1}{\lambda_i}$, $i = 1, \ldots, n$. Consecuentemente, dado que $\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ sigue siendo una matriz cuadrada, simétrica, semidefinida positiva y que σ^2 es una constante; los autovalores de esta matriz de covarianzas vienen dados por: $\frac{\sigma^2}{\lambda_i}$, $i = 1, \ldots, n$. Cuando existe al menos un i tal que $\lambda_i = 0$ se vio que se pierde la invertibilidad de la matriz $\mathbf{X}^t\mathbf{X}$, pues esta está mal condicionada. Además, si existe algún i tal que el autovalor correspondiente λ_i es muy pequeño puede llevar a que la matriz de covarianzas del estimador $\hat{\beta}$ posea una dirección de variación que está ejerciendo una gran influencia a la hora de estimar β y esto puede ocasionar grandes errores en el proceso.

De esta forma se hace latente que el método de mínimos cuadrados, pese a proporcionar un estimador insesgado de β, tiene el inconveniente de que la varianza de este puede llegar a ser muy grande y por tanto no ser una elección óptima. A continuación se proponen soluciones para este problema como son la *regresión Ridge*, la *regresión LASSO*, la regresión con penalización *Elastic Net* o la *regresión LAR*, las cuales proporcionan estimadores sesgados pero a cambio reducen su variabilidad en el caso de las tres primeras y proporcionan un enfoque de mínimos cuadrado modificado en el caso de la última.

Una forma de corregir el problema de que la matriz $\mathbf{X}^t\mathbf{X}$ sea singular es sumarle la misma cantidad constante y positiva a todos los autovalores de dicha matriz. Basta tener en cuenta que la matriz $\mathbf{X}^t\mathbf{X}$ es semidefinida positiva, de modo que aplicando esta modificación se podría conseguir que ninguno de sus autovalores se anulase y por tanto que se pudiese definir un estimador de forma similar a la vista en (1.2) pero con una matriz de covarianzas bien condicionada. También se puede considerar la opción de aplicar modificaciones al método de estimación por mínimos cuadrados que

permita obtener el vector $\hat{\beta}$ sin importar la no invertibilidad de $\mathbf{X}^t\mathbf{X}$.

Otro de los problemas de la estimación por mínimos cuadrados es que muchas veces existe una gran cantidad de variables explicativas para estimar la variable respuesta, lo que hace que sea difícil interpretar el modelo. Para solucionar este problema se busca reducir el número de variables explicativas "eficaces". Hay dos formas de hacerlo, la primera se centra en la significación que tiene cada coeficiente que acompaña a cada variable explicativa, eliminando del modelo aquellas variables que son muy poco o nada significativas hasta que sea posible explicar este con el menor número. Esto es útil cuando se busca una buena interpretación de las variables y como influye cada una sobre la variable respuesta y entre ellas. Por otra parte, la segunda forma se centra en mejorar las aproximaciones del modelo, es decir, se basa en quedarse con aquellas variables que hacen más preciso el modelo, entendiendo como tal que sus residuos son menores.

Vamos a estudiar métodos que van a permitir hallar los estimadores de los parámetros del modelo lineal por mínimos cuadrados por una parte y añadir una penalización proporcional al número de parámetros por otra en el caso de la regresión Ridge, LASSO o de Elastic Net. Esto último ayudará a reducir el número de variables explicativas que formarán parte del modelo y a hacer este más fácil de interpretar. Finalmente, se mostrará un planteamiento modificado en la estimación de mínimos cuadrados que permitirá solucionar los problemas de alta dimensión y determinar las variables más eficientes a través de la regresión LAR, mediante un procedimiento iterativo que se base en recoger adecuadamente las relaciones entre variables a través de las correlaciones con los residuos del modelo.

1.1.1. Regresión Ridge

En 1970 nace la **regresión Ridge** a manos de Arthur E. Hoerl y Robert W. Kennard en el artículo *"Ridge regression: Biased estimation for nonorthogonal problems"* (véase [7]). La finalidad principal de esta no es otra que solventar el problema de la no invertibilidad de la matriz $\mathbf{X}^t\mathbf{X}$.

En múltiples ocasiones, los parámetros de regresión estimados por mínimos cuadrados no son adecuados. Es decir, cuando la matriz $\mathbf{X}^t\mathbf{X}$ no es invertible el estimador de mínimos cuadrados $\hat{\beta}$ visto en (1.2) se aleja de dar una buena estimación del vector β ya que presenta una variabilidad mucho mayor. Si denotamos por D_1 la distancia de $\hat{\beta}$ a β entonces su cuadrado es:

$$D_1^2 = (\hat{\beta} - \beta)^t(\hat{\beta} - \beta) \tag{1.9}$$

donde

$$\begin{aligned} \mathbb{E}(D_1^2) = \mathbb{E}(\|\hat{\beta} - \beta\|^2) = \mathbb{E}\left(\sum_{i=1}^{p}(\hat{\beta}_i - \beta_i)^2\right) = \sum_{i=1}^{p}\mathbb{E}[(\hat{\beta}_i - \beta_i)^2] \\ \overset{(a)}{=} \sum_{i=1}^{p}\mathbb{V}(\hat{\beta}_i) \overset{(b)}{=} \sigma^2 \mathrm{tr}(\mathbf{X}^t\mathbf{X})^{-1} \end{aligned} \tag{1.10}$$

deduciéndose (a) de que $\hat{\beta}$ es insesgado y (b) de que $\mathbb{C}(\hat{\beta}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.

Cuando el error ϵ sigue una distribución normal como se supuso, utilizando propiedades de operadores cuadráticos (véase [16]), se obtiene:

$$\mathbb{V}(D_1^2) = 2\sigma^4 \mathrm{tr}(\mathbf{X}^t\mathbf{X})^{-2}. \tag{1.11}$$

Denotando los autovalores de la matriz $\mathbf{X}^t\mathbf{X}$ por

$$\alpha_{\text{máx}} = \alpha_1 \geq \cdots \geq \alpha_{p+1} = \alpha_{\text{mín}} \geq 0,$$

se observa que se puede expresar la media de la distancia al cuadrado (1.10) por

$$\mathbb{E}(D_1^2) = \sigma^2 \sum_{i=1}^{p+1} \frac{1}{\alpha_i} \tag{1.12}$$

y la varianza (1.11) por

$$\mathbb{V}(D_1^2) = 2\sigma^4 \sum_{i=1}^{p+1} \left(\frac{1}{\alpha_i}\right)^2. \tag{1.13}$$

Fijándose en (1.12) y (1.13) se concluye que cuando la matriz $\mathbf{X}^t\mathbf{X}$ es singular, es decir, cuando tiene autovalores nulos, la distancia de $\hat{\beta}$ a β tiende a ser muy grande y por tanto el estimador del método de mínimos cuadrados aporta una aproximación muy mala.

A continuación se expone un método que se basa en sumar una cantidad positiva a los autovalores para lograr que la matriz $\mathbf{X}^t\mathbf{X}$ sea siempre invertible.

La **regresión Ridge** impone una penalización del tamaño de los coeficientes de la regresión a la hora de minimizar la suma residual de cuadrados, esta penalización es de tipo L_2. Es decir, ahora se buscaría un vector de coeficientes $\hat{\beta}^{\mathrm{RR}}$ que cumpla:

$$\hat{\beta}^{\mathrm{RR}} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \right\} \tag{1.14}$$

donde $\lambda \geq 0$ es el parámetro que controla la cantidad de reducción. Cuanto más grande es el valor de λ, mayor es la reducción y por tanto el modelo se queda con menos variables explicativas ya que los coeficientes son aproximados hacia cero. Una forma equivalente de escribir este problema es:

$$\begin{aligned} &\hat{\beta}^{\mathrm{RR}} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ &\text{sujeto a } \sum_{j=1}^{p} {\beta_j}^2 \leq t, \end{aligned} \tag{1.15}$$

en el cual se hace explícita la restricción del tamaño de los parámetros. Hay una clara correspondencia entre el parámetro λ en (1.14) y t en (1.15). Se observa que la relación entre ambos parámetros es proporcionalmente inversa.

Al término independiente β_0 no se le impone ninguna penalización, esto se debe a que una penalización de este parámetro haría que el proceso dependiese del origen escogido para Y, pero dado que la solución para la regresión Ridge no es equivariante ante cambios de escala un cambio sumando una constante c a cada variable y_i no se solucionaría con un cambio en la predicción por la misma cantidad c. Para estimar los parámetros de $\hat{\beta}^{\mathrm{RR}}$ se usan las variables centradas $x_{ij} - \bar{x}_j$ en vez de las x_{ij} iniciales y se estima β_0 por $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ ya que este parámetro no tiene restricciones. Para el resto de elementos a estimar se lleva a cabo una estimación por mínimos cuadrados con las variables centradas sin tener en cuenta el intercepto.

De ahora en adelante se denotará por $\beta = (\beta_1, \ldots, \beta_p)^t$ y la matriz de diseño $\mathbf{X}$ tendrá la forma

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Esto lleva a que

$$\begin{aligned}\hat{\beta}^{\mathrm{RR}} &= \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}{\beta_j}^2\right\}\\ &= \min_{\beta}\left\{\left((y-\beta_0\mathbf{I}_n)-\mathbf{X}\beta\right)^t\left((y-\beta_0\mathbf{I}_n)-\mathbf{X}\beta\right)+\lambda\beta^t\beta\right\}\end{aligned}$$

donde

$$\phi(\beta) := (y-\mathbf{X}\beta)^t(y-\mathbf{X}\beta)+\lambda\beta^t\beta,$$

desarrollando la expresión se llega a que

$$\begin{aligned}\phi(\beta) &= (y^t-(\mathbf{X}\beta)^t)(y-\mathbf{X}\beta)+\lambda\beta^t\beta\\ &= (y^t-\beta^t\mathbf{X}^t)(y-\mathbf{X}\beta)+\lambda\beta^t\beta\\ &= yty-y^t\mathbf{X}\beta-\beta^t\mathbf{X}^ty+\beta^t\mathbf{X}^t\mathbf{X}\beta+\lambda\beta^t\beta.\end{aligned}$$

Ahora derivando e igualando a cero para obtener el mínimo

$$\begin{aligned}\frac{\partial\phi(\beta)}{\partial\beta} &= -y^t\mathbf{X}+\beta^t\mathbf{X}^t\mathbf{X}+\lambda\beta^t=0\Rightarrow\beta^t(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)=y^t\mathbf{X}\\ &\Rightarrow\beta(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)=\mathbf{X}^ty\Rightarrow\beta=(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^ty\end{aligned}$$

de esta forma, los parámetros se estiman por

$$\begin{aligned}\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n}y_i\\ \hat{\beta}^{\mathrm{RR}} &= (\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^ty\end{aligned}\tag{1.16}$$

Al añadir una constante positiva a la diagonal de $\mathbf{X}^t\mathbf{X}$ antes de la inversión se consigue que el problema sea no singular y que se pueda obtener la matriz inversa, incluso si $\mathbf{X}^t\mathbf{X}$ era no inversible. Con esta penalización L_2 se garantiza que existen los estimadores por mínimos cuadrados como se ve en (1.16).

Ahora, el estimador obtenido en (1.16) deja de ser insesgado, pues:

$$\begin{aligned}\mathbb{E}(\hat{\beta}^{\mathrm{RR}}) &\overset{(a)}{=} \mathbb{E}((\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^ty)\overset{(b)}{=}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^t\mathbb{E}(y)\\ &\overset{(c)}{=}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^t\mathbb{E}(\mathbf{X}\beta)\overset{(d)}{=}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^t\mathbf{X}\beta\neq\beta\end{aligned}\tag{1.17}$$

donde en (a) se emplea (1.16), en (b) que $(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I}_p)^{-1}$ es una matriz de diseño fijo y finalmente en (c) y en (d) propiedades vistas del modelo lineal general.

En cambio, observando la media del error al cuadrado visto en (1.10) y comparándola con la obtenida para el nuevo estimador que es de la forma:

$$\begin{aligned}
\mathbb{E}((D_1^2)^{RR}) =& \mathbb{E}\left((\hat{\beta}^{\mathrm{RR}}-\beta)^t(\hat{\beta}^{\mathrm{RR}}-\beta)\right)\\
\overset{(a)}{=}& \mathbb{E}\left((\hat{\beta}-\beta)^t\mathbf{Z}^t\mathbf{Z}(\hat{\beta}-\beta)\right)+(\mathbf{Z}\beta-\beta)^t(\mathbf{Z}\beta-\beta)\\
\overset{(b)}{=}& \sigma^2\mathrm{tr}((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}^t\mathbf{Z})+\beta^t(\mathbf{Z}-\mathbf{I})^t(\mathbf{Z}-\mathbf{I})\beta\\
\overset{(c)}{=}& \sigma^2\mathrm{tr}\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-1}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}\right]\\
&+\beta^t\left[-\lambda(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-1}\right]^t\left[-\lambda(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-1}\right]\beta\\
=& \sigma^2\mathrm{tr}\left[(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}\mathbf{X}^t\mathbf{X}\right]+\beta^t\lambda^2(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}\beta\\
=& \sigma^2\mathrm{tr}\left[(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})-\lambda(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}\right]\\
&+\lambda^2\beta^t(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}\beta\\
=& \sigma^2\left(\mathrm{tr}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-1}-\lambda\cdot\mathrm{tr}(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}\right)\\
&+\lambda^2\beta^t(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}\beta\\
\overset{(d)}{=}& \sigma^2\sum_{i=1}^{p}\frac{\alpha_i}{(\alpha_i+\lambda)^2}+\lambda^2\beta^t(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-2}\beta
\end{aligned} \tag{1.18}$$

donde en (a) se está usando que $\hat{\beta}^{\mathrm{RR}}=\mathbf{Z}\hat{\beta}$ con $\mathbf{Z}=((\mathbf{X}^t\mathbf{X})+\lambda\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}$, propiedades de la media y se tiene en cuenta (1.3). En (b) se siguen aplicando propiedades de la media y (1.10). En (c) se emplea que $\mathbf{Z}=((\mathbf{X}^t\mathbf{X})+\lambda\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}=\mathbf{I}-\lambda(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-1}$ (véase [7]). Finalmente en (d) se usa que la traza de una matriz es la suma de sus autovalores.

Se va a tener garantizado que siempre va a existir un valor λ tal que el estimador $\hat{\beta}^{\mathrm{RR}}$ va a tener una media del error al cuadrado (1.18) menor que la que se obtiene para el estimador de mínimos cuadrados:

Teorema 1.3 (Teorema de existencia)**.** *Siempre va a existir un* $\lambda>0$ *tal que* $\mathbb{E}((D_1^2)^{RR})<\mathbb{E}(D_1^2)$.

En [7] se puede ver la demostración del Teorema 1.3 el cual garantiza que escogiendo un λ adecuado el estimador de la regresión Ridge será mejor que el obtenido por mínimos cuadrados. De esta forma no sólo se soluciona el problema principal sino que además se obtiene un estimador mejor en términos del error cuadrático medio.

Este nuevo estimador tiene matriz de covarianzas

$$\begin{aligned}
\mathbb{C}(\hat{\beta}^{\mathrm{RR}}) &= \mathbf{Z}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{C}(y,y)\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}^t\\
&= \sigma^2\mathbf{Z}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}^t.
\end{aligned} \tag{1.19}$$

Cuanto más grande es el valor del parámetro λ menor es la varianza (1.19) para $\hat{\beta}^{\mathrm{RR}}$ pero mayor es el sesgo (1.17).

A la hora de estimar el parámetro λ se busca aquel que minimice la suma residual de cuadrados (RSS). Si B es un estimador cualquiera del vector β la suma residual de cuadrados puede ser escrita como

$$\begin{aligned}
\phi &= (y-\mathbf{X}B)^t(y-\mathbf{X}B)=((y-\mathbf{X}\hat{\beta})+(\mathbf{X}\hat{\beta}-\mathbf{X}B))^t((y-\mathbf{X}\hat{\beta})+(\mathbf{X}\hat{\beta}-\mathbf{X}B))\\
&= (y-\mathbf{X}\hat{\beta})^t(y-\mathbf{X}\hat{\beta})+2(y-\mathbf{X}\hat{\beta})^t(y-\mathbf{X}\hat{\beta})+(\mathbf{X}\hat{\beta}-\mathbf{X}B)^t(\mathbf{X}\hat{\beta}-\mathbf{X}B)\\
&\overset{(a)}{=} \underbrace{(y-\mathbf{X}\hat{\beta})^t(y-\mathbf{X}\hat{\beta})}_{\phi_{\mathrm{mín}}}+\underbrace{(B-\hat{\beta})^t\mathbf{X}^t\mathbf{X}(B-\hat{\beta})}_{\phi(B)}
\end{aligned} \tag{1.20}$$

usando en (a) que $(y-\mathbf{X}\hat{\beta})^t(\mathbf{X}\hat{\beta}-\mathbf{X}B)=(\mathbf{X}\hat{\beta}-\mathbf{X}B)^t(y-\mathbf{X}\hat{\beta})=(\hat{\beta}-B)^t\mathbf{X}^t(y-\mathbf{X}\hat{\beta})=0$ pues $\mathbf{X}^ty-\mathbf{X}^t\mathbf{X}\hat{\beta}=\mathbf{X}^ty-\mathbf{X}^ty=0$ ya que $\hat{\beta}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^ty$.

Los contornos de la constante ϕ son las superficies de los hiperelipsoides centrados en $\hat{\beta}$ (el estimador de mínimos cuadrados). El valor de ϕ es el valor mínimo $\phi_{\text{mín}}$ más un valor en forma cuadrática $(B-\hat{\beta})$. Fijado un incremento $\phi_0 > 0$ van a existir un conjunto de B_0 que satisfacen la relación $\phi = \phi_{\text{mín}} + \phi_0$. Fijado un valor ϕ, se escoge un único estimador B con mínima longitud de la forma

$$\begin{aligned} &\text{minimizar } B^tB \\ &\text{sujeto a } (B-\hat{\beta})^t\mathbf{X}^t\mathbf{X}(B-\hat{\beta}) = \phi_0, \end{aligned} \tag{1.21}$$

se emplean multiplicadores de Lagrange para resolver el problema

$$\text{minimizar } F = B^tB + (1/\lambda)[(B-\hat{\beta})^t\mathbf{X}^t\mathbf{X}(B-\hat{\beta})^t - \phi_0] \tag{1.22}$$

con $(1/\lambda)$ el multiplicador de Lagrange. Entonces

$$\frac{\partial F}{\partial B} = 2B + (1/\lambda)[2(\mathbf{X}^t\mathbf{X})B - 2(\mathbf{X}^t\mathbf{X})\hat{\beta}] = 0. \tag{1.23}$$

Esto se reduce a

$$B = \hat{\beta}^{\text{RR}} = [\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^t y \tag{1.24}$$

donde λ es escogido para verificar la restricción (1.21).

Ahora, de forma análoga a la vista en (1.7) se ve que se puede estimar σ^2 por:

$$\hat{\sigma}^2 = \frac{y^t\mathbf{M}y}{n-p}$$

donde $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$ y ahora $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^t$ es la matriz hat ajustada de la regresión Ridge.

Por tanto la regresión Ridge soluciona el problema de existencia y unicidad de la matriz inversa de $\mathbf{X}^t\mathbf{X}$ en alta dimensión. Además la imposición de la penalización L_2 en el tamaño de los coeficientes alivia los problemas que surgen cuando las variables están muy correlacionadas, ya que cuando esto sucede un coeficiente muy grande y positivo puede ser anulado por otro de magnitud similar y negativo a causa de que ambas variables a las que acompañan están muy correlacionadas entre si.

Por otra parte ahora el estimador de β es sesgado como probamos en (1.17) frente al insesgado obtenido al principio por mínimos cuadrados en (1.3).

Estimador	**Fórmula**
Ridge	$\hat{\beta}_j/(1+\lambda)$
LASSO	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

Tabla 1.1: Comparación entre los estimadores de β obtenidos mediante mínimos cuadrados, $\hat{\beta}_j$, frente a su equivalencia por el método Ridge o LASSO.

Se puede ver que ahora la equivalencia entre este nuevo estimador y el de mínimos cuadrados viene dada por la expresión $\hat{\beta}_j^{\text{RR}} = \hat{\beta}_j/(1+\lambda)$, siendo $\hat{\beta}_j$ el parámetro obtenido por mínimos cuadrados, Tabla 1.1. Puesto que $\lambda > 0$ se ve que ahora se reduce el valor de mínimos cuadrados

para todos los coeficientes $\hat{\beta}_j$, haciendo que estos sean más próximos a cero. De esta forma, para valores de λ suficientemente grandes se conseguirá que muchos de los coeficientes tengan un valor muy cercano a cero, lo cual era el objetivo deseado.

A la hora de hallar el valor del parámetro λ se pueden aplicar diversos métodos de la literatura como la **validación cruzada** o la **validación cruzada generalizada**. Los algoritmos de estos métodos así como el desarrollo de los mismos para este problema se pueden ver en [17] o [3].

1.1.2. Regresión LASSO

La **regresión LASSO**, *"Least Absolute Shrinkage and Selection Operator"*, fue introducida por Robert Tibshirani en 1996. En el artículo *" Regression shrinkage and selection via the LASSO"* [17] explica este nuevo método y expone las ventajas que presenta frente a la regresión Ridge.

Al igual que la Ridge, minimiza la suma residual de cuadrados imponiendo ahora una restricción en la suma de los valores absolutos de los coeficientes. Esta restricción lleva a que algunos coeficientes sean nulos y que se consiga un modelo más fácil de interpretar. Al igual que pasaba con la regresión Ridge se busca un estimador que deje de ser insesgado pero que a cambio reduzca la varianza de los valores estimados y por tanto que mejore en conjunto la precisión de la predicción.

Ahora se impone una penalización L_1, este método busca hallar el vector de coeficientes que cumpla:

$$\hat{\beta}^{\mathrm{RL}} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{1.25}$$

Al igual que ocurría en la regresión Ridge en (1.15) este problema se puede expresar como

$$\begin{aligned} &\hat{\beta}^{\mathrm{RL}} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ &\text{sujeto a } \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{1.26}$$

De igual forma que antes se establece una relación inversa entre el coeficiente λ de (1.25) y t de (1.26).

Ahora la penalización $L_2 : \sum_{j=1}^{p} {\beta_j}^2$ de (1.15) se remplaza por una penalización $L_1 : \sum_{j=1}^{p} |\beta_j|$ de (1.26). Esta nueva penalización hace que la solución no vaya a ser lineal en el término y_i y no se puede expresar el estimador de forma explícita matricialmente como ocurría en la regresión Ridge. Se debe a que el cálculo del estimador LASSO es un problema de programación cuadrático. En el Algoritmo 1.4 se muestra un *método de descenso de coordenadas* para resolver este problema. Otros métodos de resolución son tratados en [17].

De nuevo, para hallar el valor de λ se puede recurrir a métodos de validación cruzada al igual que en el caso de la penalización Ridge.

De esta forma, para cada situación se tendría que resolver un problema de programación cuadrática donde se obtendría un estimador $\hat{\beta}^{\mathrm{RL}}$. Este esquema soluciona el problema de la no invertibilidad y permite obtener el modelo lineal general estimado. La filosofía de la regresión LASSO es muy similar a la que se exponía para la regresión Ridge pero con la ventaja de que esta nueva permite reducir el número de parámetros no nulos de forma más eficiente. Se puede ver esta idea

en un caso bidimensional, como se muestra en la Figura 1.2 donde se ve que los contornos de la función de mínimos cuadrados en el caso de la regresión LASSO corta la región azul en un vértice situado en un eje cardinal, lo que hace que la otra componente sea nula (en este ejemplo $\beta_1 = 0$). En cambio no pasa lo mismo para la regresión Ridge. Esta situación se extiende a espacios de dimensión mayor ($p > 2$) haciendo que el modelo obtenido a través de la regresión LASSO sea más fácil de interpretar que el aportado por la regresión Ridge.

Algoritmo 1.4 (Coordinate descent LASSO)**.**

1. *Se inicializa* $\hat{\beta}_j = 0$ *para todo* $j = 1, \ldots, p$.
2. *Para cada* $j = 1, \ldots, p$
   - 2.1. *Se calculan los residuos* $R_j = y - \sum_{k \neq j} \hat{\beta}_k X_k$.
   - 2.2. *Se proyecta cada residuo sobre* X_j *obteniendo* $P_j = X_j^t R_j$.
   - 2.3. *Se obtiene el umbral suave* $\hat{\beta}_j = [1 - \lambda/|P_j|]_+ P_j$.
3. *Se itera el paso* (2) *hasta convergencia.*

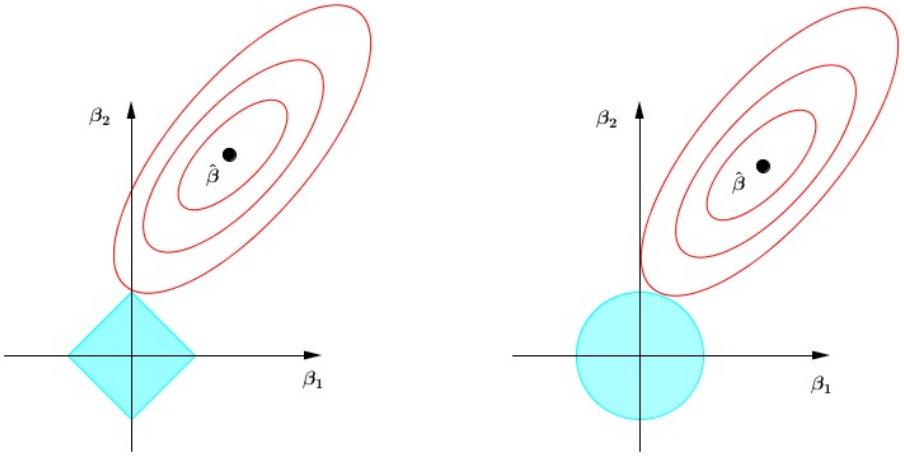

Figura 1.2: Gráfica de estimación de la regresión LASSO (izquierda) y la regresión Ridge (derecha). Las regiones azules son las áreas donde $|\beta_1| + |\beta_2| \leq t$ y ${\beta_1}^2 + {\beta_2}^2 \leq t^2$ respectivamente, mientras que las elipses rojas son los contornos de la función de mínimos cuadrados.

Ahora, en el caso de la regresión LASSO, la correspondencia entre el estimador obtenido por este método y el de mínimos cuadrados viene dada por la expresión recogida en la Tabla 1.1, $\hat{\beta}_j^{\mathrm{RL}} = \mathrm{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$, donde sign(.) denota el signo y $(.)_+$ iguala a cero todas aquellas cantidades que no sean positivas. La interpretación de este coeficiente es que todo elemento del estimador de mínimos cuadrados cuyo valor absoluto sea menor o igual que λ, $|\hat{\beta}_j| \leq \lambda$, es ajustado a cero, mientras que en caso contrario se encoge este valor una cantidad λ hacia el origen. En conclusión se extrae que determinando un valor de $\lambda > 0$ lo suficientemente grande no sólo se garantiza que se encogen más los coeficientes que se obtendrían por mínimos cuadrados hacia cero sino que se puede conseguir que un gran número de coeficientes se tomen como nulos.

En vez de la penalización Ridge o la LASSO se podrían haber impuesto otras restricciones de la forma $\sum_j |\beta_j|^q \leq t$ para un valor de q determinado y un valor de t fijado, solucionando el problema y dando lugar a otros problemas de optimización que actuarían de formas diferentes. En la Figura 1.3 se muestran algunas de las fronteras que se obtendrían con estas penalizaciones en el

caso bidimensional.

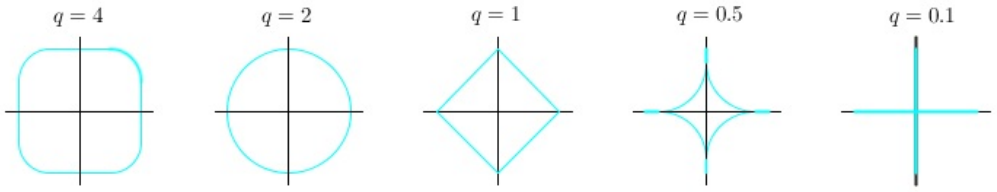

Figura 1.3: Contornos del valor constante de la restricción $\sum_j |\beta_j|^q$ para valores dados de q.

1.1.3. Regresión con penalización Elastic Net

Cuando surge la situación donde las variables están altamente correladas la penalización LASSO es indiferente ante la elección de un conjunto de variables correladas, mientras que la penalización Ridge tiende a encoger los coeficientes de las variables correladas hacia el de la otra. Ante esta situación se puede recurrir a la penalización **Elastic Net**, la cual es un compromiso entre ambas y tiene la forma

$$\sum_{j=1}^{p}(\alpha|\beta_j| + (1-\alpha){\beta_j}^2). \tag{1.27}$$

El segundo término refuerza el promedio de las variables altamente correladas mientras que el primero impulsa una separación en los coeficientes de estas variables promediadas enfrentándose al problema de la alta correlación.

Para hallar el nuevo estimador el problema consiste en

$$\begin{aligned}\hat{\beta}^{\mathrm{EN}} = \min_{\beta} \Bigg[& \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \\ & + \lambda\sum_{j=1}^{p}(\alpha|\beta_j| + (1-\alpha)\beta_j^2)\Bigg]\end{aligned} \tag{1.28}$$

donde el parámetro α determina la mezcla de la penalización. Los parámetros α y λ pueden ser escogidos empleando de nuevo métodos como el de la validación cruzada. Igual que antes, el estimador $\hat{\beta}^{\mathrm{EN}}$ obtenido de resolver (1.28), soluciona el problema de obtención de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ de una forma análoga a la vista para la regresión Ridge o LASSO.

El problema 1.28 también se puede formular como

$$\hat{\beta}^{\text{EN}} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 ,$$
$$\text{sujeto a } \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2) \leq t.$$

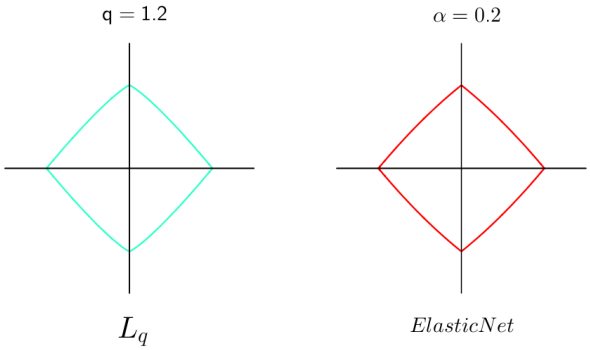

Figura 1.4: A la izquierda se muestra el valor constante de $\sum_j |\beta_j|^q$ para $q = 1.2$ y a la derecha la penalización Elastic Net $\sum_j (\alpha {\beta_j}^2 + (1-\alpha)|\beta_j|)$ para $\alpha = 0.2$. Aunque se parecen gráficamente, la Elastic Net no es diferenciable mientras que la penalización $q = 1.2$, sí.

En la Figura 1.4 se hace una comparación en dimensión dos de la penalización de red elástica con una restricción muy cercana a la LASSO: $\sum_j |\beta_j|^{1.2}$. Se obtienen áreas dadas por la correspondiente restricción muy similares salvo que la penalización $q = 1.2$ es diferenciable en todos los puntos de su frontera mientras que la penalización Elastic Net o de red elástica no.

1.1.4. Regresión LAR

La **regresión LAR**, cuyas siglas se corresponden en inglés con *Least Angle Regression*, fue introducida por Bradley Efron et. al. (véase [2]). Esta metodología sigue un procedimiento similar al *criterio forward* empleado en la selección de parámetros que entran en juego en un modelo de regresión. Cabe recordar que este último se basa en ir añadiendo variables al modelo e ir comprobando si su inclusión mejora o no los resultados obtenidos a la hora de explicar la variable Y, penalizando el número de variables debido al aumento de dificultad que supone su posterior interpretación.

De esta forma, siguiendo la filosofía del criterio forward, se empieza con todos los coeficientes que acompañan a las variables del modelo igualados a cero. El primer paso que sigue el método LAR es identificar la variable más correlada x_j, en términos de valor absoluto, con la respuesta. Una vez determinada, se ajusta una regresión linear simple de esta variable con la respuesta, obteniéndose su correspondiente coeficiente de mínimos cuadrados y un vector de residuos, el cual es ortogonal a esta variable. A continuación se mueve el valor del coeficiente β_j de cero hacia su valor de mínimos cuadrados lo máximo posible hasta que otra de las variables explicativas, x_k, tenga tanta correlación con el residuo como tiene la variable inicial x_j. En este punto el criterio LAR se aleja de la selección forward, ya que en vez de seguir a lo largo de x_j, se desplaza en una dirección

equiangular al grupo de variables activas $\{x_j, x_k\}$, es decir en la dirección que forma el menor ángulo con ambas variables (*"least angle direction"*), hasta que otra de las explicativas restantes alcanza el nivel de correlación de estas dos y se añade por tanto al conjunto de variables activas. En la Figura 1.5 se ilustra el funcionamiento del algoritmo para $p = 2$, se ve como en cada paso la estimación LAR $\hat{\mu}_k$ se aproxima a la correspondiente estimación por mínimos cuadrados $\bar{y}_k$, pero sin llegar a alcanzarla. Después de mín$(n-1, p)$ pasos se obtiene un conjunto de variables explicativas que serán empleadas de la forma usual para construir un modelo lineal de mín$(n-1, p)$ parámetros.

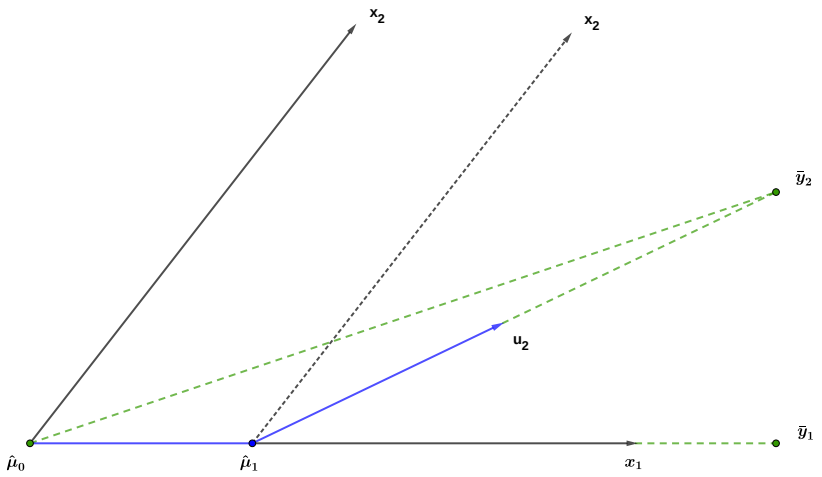

Figura 1.5: El algoritmo LAR en el caso de tenerse $p = 2$ covariables, $\bar{y}_2$ es la proyección de y en $\mathcal{L}(x_1, x_2)$ (espacio lineal que se extiende sobre x_1 y x_2). Empezando en $\hat{\mu}_0 = 0$, el vector residual $\bar{y}_2 - \hat{\mu}_0$ tiene mayor correlación con x_1 que con x_2; la siguiente estimación LAR es $\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 x_1$, donde $\hat{\gamma}_1$ es tomado de forma que $\bar{y}_2 - \hat{\gamma}_1$ biseca el ángulo entre x_1 y x_2; luego se obtiene $\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_2 u_2$, donde u_2 es el bisector unitario; $\hat{\mu}_2 = \bar{y}_2$ en el caso de $p = 2$, aunque esto no es cierto si $p > 2$. La ruta azul indica como sería un procedimiento *Stagewise* típico.

Algoritmo 1.5 (Least Angle Regression)**.**

1. *Se empieza tomando los residuos $r = y - \bar{y}$ y los coeficientes $\beta_1, \ldots, \beta_p = 0$.*
2. *Se busca la variable x_j más correlada con r.*
3. *Se lleva el valor de β_j de 0 hacia su valor mediante la estimación por mínimos cuadrados $\langle x_j, r \rangle$, hasta que otra de las variables explicativas x_k tenga tanta correlación con el residuo actual como x_j.*
4. *Mover β_j y β_k en la dirección definida por los valores obtenidos para estos parámetros mediante el método de mínimos cuadrados del actual residuo en (x_j, x_k), hasta que otra de las variables explicativas x_l tenga tanta correlación con el residuo actual como estas dos.*
5. *Continuar de esta forma hasta que todas las p variables explicativas hayan sido añadidas. Después de* mín$(n-1, p)$ *pasos, nos encontraremos con la solución completa de mínimos cuadrados.*

La condición de finalización del algoritmo, es decir el paso 5, en el caso de alta dimensión estudiado donde $p > n - 1$, fuerza a que el algoritmo LAR alcance la solución residual cero tras

$n-1$ pasos (cabe recordar que las variables son centradas, de ahí el -1). Este algoritmo soluciona tanto el problema de no invertibilidad de la matriz $\mathbf{X}^t\mathbf{X}$ como de la selección de variables, ya que considerada a lo sumo $n-1$ covariables cuando $p > n$, consiguiendo un modelo eficiente.

Denotando por $\mathcal{A}_k$ el conjunto de variables activas al principio del paso k-ésimo, y $\beta_{\mathcal{A}_k}$ el vector de coeficientes de esas variables, el cual tiene $k-1$ valores no nulos mas el que acaba de añadirse que es cero. Si $r_k = y - \mathbf{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$ es el residuo actual, entonces la dirección para este paso es

$$\delta_k = (\mathbf{X}^t_{\mathcal{A}_k}\mathbf{X}_{\mathcal{A}_k})^{-1}\mathbf{X}^t_{\mathcal{A}_k}r_k.$$

Por tanto, el perfil del coeficiente evoluciona como $\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha \cdot \delta_k$. Se puede ver que por construcción dicha dirección verifica que las correlaciones entre las $k-1$ variables con coeficientes no nulos y los correspondientes residuos son iguales y además estas van decreciendo al añadirse variables al conjunto activo. Si el vector ajustado al principio del paso es $\hat{f}_k$, entonces esto implica que $\hat{f}_k(\alpha) = f_k + \alpha \cdot u_k$, donde $u_k = \mathbf{X}_{\mathcal{A}_k}\delta_k$ es la nueva dirección del ajuste. En conclusión, puede verse como que el nombre de *"least angle"* (menor ángulo) viene dado mediante una interpretación geométrica de este proceso debido a que u_k genera el menor (e igual) ángulo con cada una de las variables explicativas que conforman $\mathcal{A}_k$.

Por construcción, los coeficientes en LAR cambian de manera lineal por partes. Teniendo en cuenta que no necesitamos tomar pequeños pasos y volver a verificar las correlaciones en el paso 3, usando la información proporcionada por la covarianza de los predictores y la linealidad del algoritmo, podemos calcular la longitud exacta del paso al comienzo del mismo.

Se puede aplicar una simple modificación del algoritmo 1.5 la cual sigue un enfoque LASSO y también es linear por trozos.

Algoritmo 1.6 (Least Angle Regression: LASSO Modification)**.**

4. *Si un coeficiente no nulo llega a cero, se elimina la variable del conjunto activo y se recalcula la dirección actual de mínimos cuadrados del conjunto.*

El algoritmo LAR con la modificación LASSO es extremadamente eficiente, ya que requiere el mismo orden computacional que un ajuste simple por mínimos cuadrados usando p variables. El método LAR converge a un estimador de mínimos cuadrados completo al cabo de p pasos, mientras que el ajuste mediante un método de tipo LASSO puede llevar más pasos, pese a que ambos son bastante similares. De esto se concluye que el Algoritmo 1.6 es una forma eficiente de calcular una solución para cualquier problema LASSO, especialmente en el caso de alta dimensión donde $p > n$.

1.2. Modelos aditivos. Problemas en alta dimensión.

Hasta ahora se ha mostrado como expresar o explicar la dependencia lineal entre dos variables, tanto en el caso unidimensional como multivariante, a través de los modelos lineales de regresión. Pero, ¿qué ocurre si la relación entre dos variables tiene una naturaleza distinta? En este contexto los modelos de regresión presentados no proporcionarán resultados satisfactorios y es necesario recurrir a otras opciones.

Como puede verse en Hastie et. al. [6], se podría pensar en generalizar el modelo de regresión lineal múltiple (1.1) a través de la utilización de superficies suavizadoras que recojan los efectos no lineales proporcionados por las variables. De esta forma se buscaría hallar una estimación no paramétrica o tipo kernel del modelo de regresión

$$Y = f(X_1, \ldots, X_p) + \epsilon.$$

Este enfoque ocasiona problemas en el caso de alta dimensión donde p es grande, ya que se tiene un gran número de variables explicativas, lo cual hace que el espacio de covariables con el que

se trabaja tenga gran dimensión. Esto es debido a que en esta situación se ocasionan problemas referentes al carácter local de los entornos, ya que los entornos con un número fijo de puntos son cada vez menos locales según aumenta la dimensión del espacio. Esta situación es común cuando $p > n$, ya que se suele tener un gran número de covariables, haciendo que tenga especial interés comprender los problemas de este planteamiento en dicha situación.

Para entender este problema de forma sencilla basta tener en cuenta qué ocurre en el caso de tres dimensiones ante entornos cúbicos y pensar en una extensión de esta idea al caso de espacios de dimensión mayor. En la Figura 1.6 se ilustra este ejemplo. Basta suponer que los puntos están distribuidos de forma uniforme en un cubo unitario p-dimensional y que interesa construir un entorno cúbico en el origen para capturar, en promedio, un porcentaje $100 \times lapso\,\%$ de los datos. Es fácil ver que este subcubo debe tener un largo de lado $lapso^{1/p}$. De esta forma se tiene que para $p = 1$ y un $lapso = 0.1$ el lado del entorno cúbico mide 0.1, mientras que si la dimensión aumenta a $p = 10$ el lado pasaría a medir 0.8, perdiéndose el carácter local. Por tanto se concluye que el concepto de local, en términos de porcentaje de datos, falla en alta dimensión. Este inconveniente se conoce como *la maldición de la dimensionalidad*.

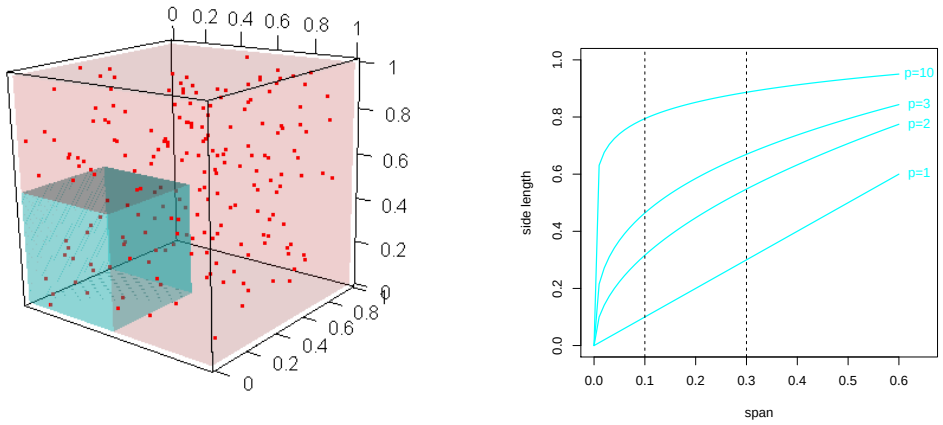

Figura 1.6: El cubo más grande (rojo) representa una distribución uniforme en p dimensiones, mientras el subcubo azul representa el $\%lapso$ del volumen. El gráfico de la derecha muestra la longitud de lado necesaria para alcanzar esta proporción a lo largo de cualquier coordenada, para diferentes dimensiones p.

Este método es útil cuando la dimensión de p es pequeña, pero según adquiere valores más grandes, por ejemplo $p \geq 4$, se pierde la interpretabilidad de dichas superficies, haciendo que estos modelos sean difíciles de entender en un caso de alto dimensión. Debido a esta dificultad será necesario buscar otros enfoques que permitan dotar de mayor interpretabilidad al ajuste.

Con el fin de conseguir explicar la dependencia no necesariamente lineal existente entre las variables mediante una estructura más fácil de analizar surgen los **modelos aditivos**. Estos consideran que cada una de las variables explicativas x_{ij} con $i = 1, \ldots, n$ y $j = 1, \ldots, p$ se relaciona con la variable respuesta a través de una función f_j no necesariamente lineal, de la forma $y_i = f_j(x_{ij})$, y además que estos efectos se suman a la hora de explicar esta última, $y_i = \sum_{j=1}^{p} f_j(x_{ij})$.

En consecuencia el modelo considerado tendría una estructura

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \tag{1.29}$$

donde y_i es la variable respuesta, x_{ij} las $j = 1, \ldots, p$ covariables o variables explicativas, f_j las funciones suaves y ϵ_i variables aleatorias i.i.d. $N(0, \sigma^2)$ que recogen el error del modelo.

Ahora, la formulación del modelo (1.29), mediante la utilización de las funciones f_j, permite recoger estructuras de carácter no lineal entre las variables explicativas con la respuesta, dotando de mayor flexibilidad al ajuste. De esta forma se proporciona una clara ventaja frente al modelo lineal. A la hora de estimar los parámetros de (1.29) se ve que ahora no se tendrá que hallar un vector de parámetros sino la respectiva función f_j de cada una de las $j = 1, \ldots, p$ variables explicativas. Para esta finalidad se emplean principalmente dos técnicas

- **Suavización tipo núcleo** o *Kernel smoothing*: emplea funciones núcleo o *Kernel* para conseguir aproximar la estructura de las funciones f_j utilizando la información proporcionada por los datos muestrales.
- **Representación en una base funcional:** cada una de las funciones f_j se aproxima en términos de una combinación lineal de q_j elementos de una base funcional debidamente escogida, como podría ser la *polinómica*, la de *splines* (ver Figura 1.7) o la de *wavelets* entre otras, de la cual se desconoce el valor de sus escalares. De este modo se pasa del problema de estimar la estructura de la función a tener que hallar los valores de los escalares que conforman la combinación lineal.

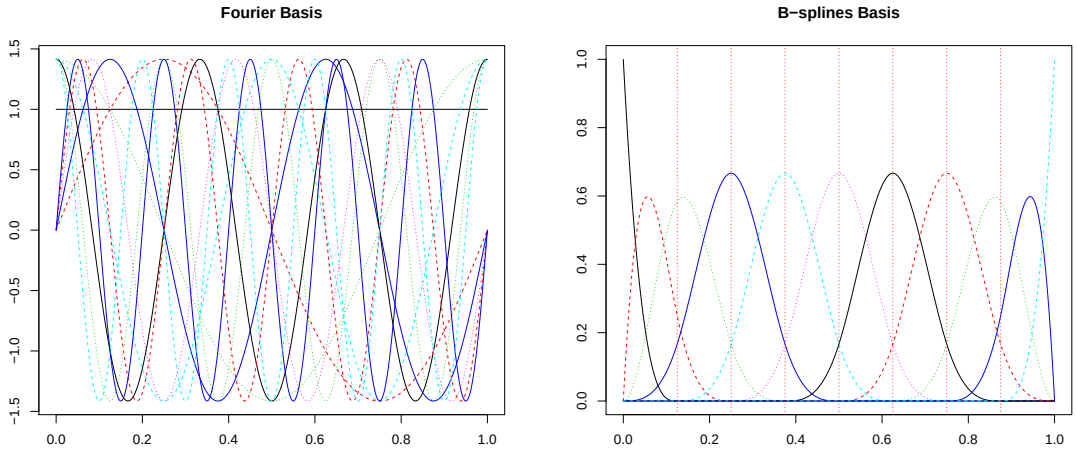

Figura 1.7: Ejemplos de bases funcionales donde se consideran 11 elementos de una base de Fourier (izquierda) y otros 11 de una base de b-splines (derecha).

Implícito en (1.29) está impuesta la condición de que $\mathbb{E}[f_j(x_j)] = 0$, ya que de otra forma habrá constantes libres en cada una de las funciones y no se tendría una única solución. Es decir, no se diferenciaría entre f_1 y $f_1 + c$ ya que el efecto $c \in \mathbb{R}$ podría ser eliminado por las constantes de las restantes $j = 2, \ldots, p$ funciones.

1.2.1. Estimación de las funciones f_j

Tanto en el caso de emplear la **suavización tipo núcleo** como la **representación en una base funcional** es necesario recurrir a algoritmos iterativos para poder proporcionar estimadores de las funciones f_j, precio que hay que pagar por añadir generalidad al modelo. En ambos casos el algoritmo iterativo más utilizado para ajustar un modelo aditivo es el denominado *backfitting algorithm*. Este procedimiento se apoya en el supuesto de que si la estructura aditiva del modelo (1.29) es correcta, entonces para cada k se cumple que $\mathbb{E}\left(Y - \alpha - \sum_{j \neq k} f_j(X_j) | X_k\right) = f_k(X_k)$.

Por tanto, este resultado proporciona las bases para un algoritmo iterativo que calcule todas las f_j, las cuales serán dadas en base a los datos y a los suavizadores S_j utilizados.

Algoritmo 1.7 (Backfitting)**.**

1. *Se inicializan los parámetros mediante* $\alpha = \bar{y}$, *con* $y = (y_1, \ldots, y_n)^t$ *y* $f_j = f_j^0$ *para* $j = 1, \ldots, p$.
2. *Para cada una de las funciones* f_j, $j = 1, \ldots, p$, *se recalcula iterativamente su valor a través de la expresión*
   $f_j = S_j \left(y - \alpha - \sum_{k \neq j} f_k | x_j \right)$.
3. *Se continua con el paso anterior hasta que no se aprecien cambios notables en todas las funciones individuales y se pueda concluir que el algoritmo ha alcanzado la convergencia.*

En este algoritmo hay que tener en cuenta que $S_j(Y \mid X_j)$ denota un suavizador de la respuesta Y en base al predictor X_j, produciendo una función. Lo que se busca es ajustar todas las funciones simultáneamente de forma que los pasos de suavización individual tengan sentido. Cuando se reajusta f_j, lo que se hace es eliminar todos los efectos que tienen las restantes variables sobre la respuesta Y antes de suavizar este residuo parcial en términos de X_j. Claramente se concluye que esto es únicamente apropiado si todas las funciones eliminadas también son correctas y por lo tanto la iteración.

Uno de los requisitos del algoritmo 1.7 es proporcionar iterantes iniciales para las funciones, f_j^0. Puesto que en muchas circunstancias no se cuenta con información acerca de la forma o estructura que estas pueden tener, una buena alternativa es tomar como iterantes iniciales los coeficientes de las regresiones lineales de la variable Y con cada una de las explicativas.

Una de las características notable del backfitting algorithm es que evita la maldición de la dimensionalidad debido a que se está ajustando un modelo que considera la suma de los efectos individuales de las variables. De esta forma los modelos aditivos son una propuesta totalmente válida a la hora de recoger efectos no lineales en un modelo de regresión cuando el número de covariables es mayor que el tamaño muestral, es decir en la situación de $p > n$.

Estimación mediante representación en bases funcionales: método de mínimos cuadrados penalizado

Como se explicaba anteriormente, a la hora de estimar las funciones f_j del modelo (1.29) se puede recurrir a la **representación en bases funcionales**. Este método se basa en expresar cada una de las funciones f_j en términos de una base funcional, consiguiendo que el modelo (1.29) tenga una estructura similar a la del caso lineal teniendo que estimarse ahora únicamente un vector de parámetros que puede ser calculado a través de procedimientos conocidos. De esta forma, una vez que se selecciona una base funcional adecuada con q_j elementos, $\{b_{jq}\}_{q=1}^{q_j}$, se representa cada una de las funciones $f_j(x)$, $j = 1, \ldots, p$, como

$$f_j(x) = \sum_{q=1}^{q_j} b_{jq}(x)\beta_{jq}, \tag{1.30}$$

siendo β_{jq} coeficientes desconocidos. Si se sustituye (1.30) en (1.29) se obtiene el modelo lineal (1.31).

$$y_i = \sum_{j=1}^{p} \sum_{q=1}^{q_j} b_{jq}(x)\beta_{jq} + \epsilon_i \tag{1.31}$$

De esta forma se puede ver que ahora se puede reescribir (1.29) como una estructura lineal respecto a los elementos de las bases funcionales escogidos, pues teniendo en cuenta (1.31) se puede ver que es posible expresar el modelo como

$$y = \mathbf{X}\beta + \epsilon, \tag{1.32}$$

donde ahora $y = (y_1, \ldots, y_n)$, $\mathbf{X} = [\{b_{1q}\}_{q=1}^{q_1}, \ldots, \{b_{pq}\}_{q=1}^{q_p}] \in \mathcal{M}_{n\times\sum_{j=1}^{p} q_j}$ y $\beta = [\{\beta_{1q}\}_{q=1}^{q_1}, \ldots, \{\beta_{pq}\}_{q=1}^{q_p}]^t$ es un vector de dimensión $\sum_{j=1}^{p} q_j$.

Ante la consideración de (1.31) surge el problema de la elección adecuada de los grados de suavización o número de elementos de la base funcional q_j, involucrados en la representación de cada una de las funciones $f_j \quad \forall j = 1, \ldots, p$. Se podría pensar en determinar cada valor q_j a través de criterios backwards, empezando por un valor grande para q_j e ir reduciéndolo siempre que el modelo no se vea demasiado afectado, pero este método es problemático dado que los modelos de $k-1$ y k nodos no son necesariamente anidados y por lo tanto no es adecuado compararlos. También se podría considerar el aplicar este algoritmo partiendo de una malla densa de nodos e ir eliminando y comparando los resultados secuencialmente, pero las distancias desiguales entre puntos podría llevar a un mal funcionamiento del modelo. Además, es necesario tener en cuenta que el ajuste del modelo depende en fuerte medida de las localizaciones escogidas para los nodos.

Una forma de controlar el suavizado sin necesidad de alterar las dimensiones de las bases es considerar una base de dimensión mayor a la que se cree que podría ser razonable y controlar el suavizado excesivo a través de la agregación de una penalización. Esta penalización se añadirá en el proceso de estimación de los parámetros del modelo (1.32) a través de mínimos cuadrados y se encargará de sancionar la excesiva curvatura de cada estimación, medida por $\int_{\mathcal{D}_j} [f_j^{''}(x)]^2 dx$, a través de un parámetro $\lambda_j \geq 0$ que será necesario determinar. Cuanto más grande sea el valor de λ_j, es decir cuando $\lambda_j \to \infty$, mayor será la penalización y por tanto la estimación de f_j tenderá a ser una línea recta, mientras que si $\lambda_j = 0$ se obtiene el estimador del caso no penalizado para f_j. Así, para estimar los parámetros de (1.32) será necesario, una vez conocidos λ_j's, resolver el problema

$$\min_{\beta} \quad \|y - \mathbf{X}\beta\|^2 + \lambda_1 \int_{\mathcal{D}_1} [f_1^{''}(x)]^2 dx + \cdots + \lambda_p \int_{\mathcal{D}_p} [f_p^{''}(x)]^2 dx \tag{1.33}$$

donde $\mathcal{D}_j$ con $j = 1, \ldots, p$ son los dominios de las variables x_j.

Puesto que cada f_j es lineal en base a los parámetros β_j, la penalización se puede escribir como una forma cuadrática en β_j obteniendo que

$$\int_{\mathcal{D}_j} [f_j^{''}(x)]^2 dx = \beta^t \mathbf{S}_j \beta, \tag{1.34}$$

donde $\mathbf{S}_j = \int_{\mathcal{D}_j} (B_j^{''}(x))^t (B_j^{''}(x)) dx \in \mathcal{M}_{\sum_{j=1}^{p} q_j \times \sum_{j=1}^{p} q_j}$ es una matriz conocida, ya que $B_j^{''}(x) = [b_{j1}^{''}(x), \ldots, b_{jq_j}^{''}(x)] \in \mathcal{M}_{n\times q_j}$ y $\mathbf{S}_j$ es una matriz de ceros salvo en $S_{j_{(k+q_{j-1}, l+q_{j-1})}} = (b_{jk}^{''}(x) \cdot b_{jl}^{''}(x))$ con $k, l = 1, \ldots, q_j$ y tomando $q_0 = 0$. En consecuencia sus valores dependen de los elementos de la base funcional empleada para representar las funciones f_j. Esta expresión se encarga de imponer la penalización para la variable j-ésima.

Teniendo este último resultado en cuenta se puede ver que el problema a resolver para obtener los parámetros del modelo (1.33) puede reescribirse como

$$\min_{\beta} \quad \|y - \mathbf{X}\beta\|^2 + \beta^t \mathbf{S} \beta, \tag{1.35}$$

siendo $\mathbf{S} = \lambda_1 \mathbf{S}_1 + \cdots + \lambda_p \mathbf{S}_p \in \mathcal{M}_{\sum_{j=1}^{p} q_j \times \sum_{j=1}^{p} q_j}$ tal que

$$\mathbf{S} = \begin{pmatrix} \overbrace{\lambda_1\tilde{\mathbf{S}}_1}^{q_1} & \cdots & \overbrace{0}^{q_p} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p\tilde{\mathbf{S}}_p \end{pmatrix} \begin{matrix} \} \ q_1 \\ \\ \} \ q_p \end{matrix} \tag{1.36}$$

donde $\tilde{\mathbf{S}}_j$ representan las submatrices $q_j \times q_j$ de $\mathbf{S}_j$ cuyos elementos no son nulos.

Se puede ver que el problema (1.35) es un **problema de mínimos cuadrados penalizado** con una penalización de tipo Ridge. De esta forma el problema de estimar los grados de suavización del modelo pasa a ser estimar los parámetros de suavizado λ_j.

Al igual que ocurre en el caso de la Regresión Ridge en el contexto del modelo lineal se pude ver, siguiendo un procedimiento equivalente, que el estimador para β viene dado por la expresión

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^t y, \tag{1.37}$$

donde la no invertibilidad de $\mathbf{X}^t\mathbf{X}$, en el caso de alta dimensión donde $p > n$ se subsana al sumarle la matriz $\mathbf{S} = \lambda_1\mathbf{S}_1 + \cdots + \lambda_p\mathbf{S}_p$. Esto se debe a que los autovalores de $\mathbf{X}^t\mathbf{X} + \mathbf{S}$ vienen dados por $\gamma_1 + \omega_1, \ldots, \gamma_p + \omega_p$, siendo γ_j los correspondientes a $\mathbf{X}^t\mathbf{X}$ y ω_j los obtenidos de $\mathbf{S}$, de forma que aunque $\gamma_j = 0$ para algún j, bastaría que $\omega_j \neq 0$ para poder garantizar la invertibilidad de la matriz. Puesto que ω_j son los autovalores de $\lambda_1\mathbf{S}_1 + \cdots + \lambda_p\mathbf{S}_p$ se garantiza que si los autovalores de cada $\mathbf{S}_j$ son mayores o iguales que cero, por definición de los λ_j estos también lo serán, pues $\omega_j = \lambda_1\omega_{j1} + \cdots + \lambda_p\omega_{jp}$.

Ahora, teniendo en cuenta que $\mathbf{X} = [\{b_{1q}\}_{q=1}^{q_1}, \ldots, \{b_{pq}\}_{q=1}^{q_p}] \in \mathcal{M}_{n\times\sum_{j=1}^p q_j}$, para ver que $\mathbf{X}^t\mathbf{X} \in \mathcal{M}_{\sum_{j=1}^p q_j \times \sum_{j=1}^p q_j}$ es una matriz singular en el caso de que $p > n$, basta tener en cuenta que $\sum_{j=1}^p q_j \geq p$, puesto que al menos debe haber un elemento en cada una de las bases funcionales consideradas. De esta forma, aplicando el corolario 1.2, se tiene que $\text{rango}(\mathbf{X}^t\mathbf{X}) \leq \text{rango}(\mathbf{X}) \leq n < p \leq \sum_{j=1}^p q_j \Rightarrow \text{rango}(\mathbf{X}^t\mathbf{X}) < \sum_{j=1}^p q_j$, lo cual concluye que el determinante de la matriz es cero y por tanto no existe su inversa.

Se tiene la expresión (1.37) puesto que (1.35) se puede reescribir como

$$\begin{aligned} \min_\beta \ \|y - \mathbf{X}\beta\|^2 + \beta^t\mathbf{S}\beta &= \min_\beta \ (y - \mathbf{X}\beta)^t(y - \mathbf{X}\beta) + \beta^t\mathbf{S}\beta \\ &= \min_\beta \ y^t y - y^t\mathbf{X}\beta - (\mathbf{X}\beta)^t y + (\mathbf{X}\beta)^t(\mathbf{X}\beta) + \beta^t\mathbf{S}\beta \end{aligned}$$

así que basta denotar $\phi(\beta) := y^t y - y^t\mathbf{X}\beta - (\mathbf{X}\beta)^t y + (\mathbf{X}\beta)^t(\mathbf{X}\beta)$ y despejar $\frac{\partial\phi(\beta)}{\partial\beta} = 0$ en función de β para obtener el estimador:

$$\begin{aligned} \frac{\partial\phi(\beta)}{\partial\beta} = 0 &\Rightarrow -y^t\mathbf{X} + (\mathbf{X}\beta)^t\mathbf{X} + \beta^t\mathbf{S} = 0 \\ &\Rightarrow -y^t\mathbf{X} + \beta^t\mathbf{X}^t\mathbf{X} + \beta^t\mathbf{S} = 0 \\ &\Rightarrow -y^t\mathbf{X} + \beta^t(\mathbf{X}^t\mathbf{X} + \mathbf{S}) = 0 \\ &\Rightarrow \beta^t(\mathbf{X}^t\mathbf{X} + \mathbf{S}) = y^t\mathbf{X} \\ &\Rightarrow \beta^t = y^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \mathbf{S})^{-1} \\ &\Rightarrow \beta = (\mathbf{X}^t\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^t y. \end{aligned}$$

Obtención de los coeficientes λ de las penalizaciones

Dado que se conoce una expresión explícita para la estimación de β (1.37), una vez determinadas las penalizaciones λ_j, únicamente es necesario estimar el valor de estos parámetros para ajustar el modelo. El criterio que se emplea para este fin es el de **validación cruzada**.

En la elección de los valores λ_j es necesario tener en cuenta que si estos son muy grandes los datos serán sobresuavizados y si por el contrario sus valores son demasiado pequeños se producirá un fenómeno de infrasuavizado. Así, se hace latente la importancia de determinar el valor adecuado para cada una de las penalizaciones a la hora de obtener buenos estimadores para el modelo. La filosofía que sigue la validación cruzada es escoger el estimador $\hat{f}_j$ que sea lo más parecido posible a la función real desconocida f_j. Un criterio adecuado siguiendo esta idea sería tomar los valores λ_j que minimizaran

$$M_j = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_{ji} - f_{ji})^2,$$

donde se ha denotado $\hat{f}_{ji} = \hat{f}_{ji}(x_i)$ y $f_{ji} = f_{ji}(x_i)$, $j = 1, \ldots, p$.

Puesto que las funciones f_j son desconocidas no se puede obtener el valor de M_j directamente, pero sí es posible calcular una estimación de $\mathbb{E}(M) + \sigma^2$ que se corresponde con el error cuadrático esperado al predecir una nueva variable. Si se denota $\hat{f}^{[-i]}$ como el modelo ajustado con todos los datos salvo el i-ésimo, y_i, se definen las puntuaciones o *scores* de la *validación cruzada ordinaria* por

$$\mathcal{V}_{jo} = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_{ji}^{[-i]} - y_i)^2.$$

Substituyendo $y_i = \sum_{j=1}^{p} f_{ji} + \epsilon_i$ se obtiene que

$$\begin{aligned}\mathcal{V}_{jo} &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{ji}^{[-i]} - \sum_{j=1}^{p} f_{ji} - \epsilon_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{ji}^{[-i]} - \sum_{j=1}^{p} f_{ji}\right)^2 - 2\left(\hat{f}_{ji}^{[-i]} - \sum_{j=1}^{p} f_{ji}\right)\epsilon_i + \epsilon_i^2.\end{aligned} \tag{1.38}$$

Dado que $\mathbb{E}(\epsilon_i) = 0$, y ϵ_i y $\hat{f}_{ji}^{[-i]}$ son independientes para todo $j = 1, \ldots, p$, entonces el segundo término de la suma de (1.38) desaparece al tomar la esperanza y por tanto se llega a que

$$\mathbb{E}(\mathcal{V}_{jo}) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(\hat{f}_{ji}^{[-i]} - f_{ji})^2\right) + \sigma^2. \tag{1.39}$$

Teniendo en cuenta que $\hat{f}_j^{[-i]} \approx \hat{f}_j$, con igualdad en el límite al aumentar el tamaño muestral, se puede ver que $\mathbb{E}(\mathcal{V}_o) \approx \mathbb{E}(M) + \sigma^2$, también consiguiendo igualdad en el límite al aumentar el tamaño muestral. Por tanto se justifica que escoger cada valor λ_j con la finalidad de minimizar $\mathcal{V}_{jo}$ es un resultado razonable si lo ideal es minimizar M_j. Determinar el valor de λ_j mediante el criterio de minimizar $\mathcal{V}_{jo}$ se conoce como **validación cruzada ordinaria**.

Pese a que la validación cruzada ordinaria parece un resultado razonable, si el modelo es juzgado únicamente por su habilidad de ajustar los datos con los cuales se están prediciendo los

parámetros o funciones, entonces siempre se escogerán modelos complicados sobre otros más simples. En cambio, optar por un modelo que maximiza la habilidad de predecir datos con los cuales sus componentes no han sido estimadas, no sufre de este problema.

Puesto que es ineficiente calcular $\mathcal{V}_{jo}$ con $j = 1, \ldots, p$ dejando cada vez un dato fuera y ajustando el modelo para cada uno de los n resultantes grupos, se puede probar que $\mathcal{V}_{jo}$ puede ser obtenido teniendo todos los datos muestrales en cuenta a través de la expresión

$$\mathcal{V}_{jo} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{ji})^2/(1 - A_{ii})^2,$$

donde $\mathbf{A} = \mathbf{X}(\mathbf{X}^t\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^t$ es la correspondiente matriz de influencia. En la práctica los pesos $(1 - A_{ii})$ suelen ser reemplazados por el peso medio $\mathrm{tr}(\mathbf{I} - \mathbf{A})/n$, obteniendo las puntuaciones o *scores* de *validación cruzada generalizada* dados por

$$\mathcal{V}_{jg} = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_{ji})^2}{[tr(\mathbf{I} - A)/n]^2}. \tag{1.40}$$

La **validación cruzada generalizada**, a la hora de estimar los parámetros del modelo, tiene ventajas sobre el método ordinario, además de tener ventajas en términos de la invariancia. Se puede ver que cuando el tamaño muestral tiende a infinito minimizar $\mathcal{V}_{jg}$ (1.40) es equivalente a minimizar $\mathbb{E}(M)$.

Problemas en alta dimensión ($p > n$)

A pesar de que los modelos aditivos, a diferencia de los modelos lineales estudiados anteriormente, son totalmente estimables en el contexto de alta dimensión considerado ($p > n$), aunque más costosos computacionalmente, estos afrontan otras problemas no menos importantes que es necesario solucionar.

Mientras que en los modelos lineales, tanto en el caso del modelo lineal múltiple como en el del modelo lineal generalizado, es necesario afrontar el problema de la posible **colinealidad** entre sus covariables, ahora los modelos aditivos hacen frente a un fenómeno similar denominado **concurvidad** [2], el cual es mucho más complicado de detectar debido a su naturaleza funcional. Una solución para evitar estos problemas es conseguir reducir el número de variables explicativas que entran en juego, determinando cuáles son las más eficaces a la hora de predecir la variable respuesta y de cuales se puede prescindir, haciendo el modelo más sencillo, más fácil de interpretar y con menos probabilidad de sufrir el fenómeno de concurvidad.

Puesto que en el caso de alta dimensión donde $p > n$ es muy difícil determinar qué variables son relevantes y de cuales se puede prescindir para ajustar el modelo aditivo, se propone la incorporación de una penalización relativa al número de covariables que entran en juego.

1.2.2. Regularización L_2

Una forma de implantar este criterio de selección es, al igual que en el caso de la regresión Ridge, añadir una penalización de tipo L_2.

De nuevo, empleando la representación sobre bases funcionales, esta vendrá dada a través de un parámetro $\tilde{\lambda} > 0$ y el nuevo estimador del vector de parámetros $\hat{\beta}$ será aquel que minimice el problema equivalente a (1.35) ahora dado por la estructura

[2] La **colinealidad** se cumple cuando una o más variables son una combinación lineal de otra, mientras que se afirma que existe **concurvidad** cuando la función relativa a una variable del modelo aditivo (f_j) es una combinación lineal de cierto número de funciones asociadas a otras variables.

$$\min_{\beta} \quad \|y - \mathbf{X}\beta\|^2 + \tilde{\lambda}\mathbf{I} + \beta^t \mathbf{S}\beta.$$

Con un razonamiento análogo al que se aplicaba para obtener la estructura explícita de este estimador en (1.37) se ve que ahora este viene dado a través de la estructura

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X} + \tilde{\lambda}\mathbf{I} + \mathbf{S})^{-1}\mathbf{X}^t y. \tag{1.41}$$

El método genérico empleando otro tipo de suavizador se ilustra en el caso de los modelos GAM en la subsección 1.5.2, teniendo en cuenta que los modelos aditivos son un caso particular de estos tomando como función link la identidad se tiene un procedimiento apto en ambos contextos.

1.2.3. Regularización L_1: modelo SpAM

Dado que los modelos aditivos únicamente tienen un buen comportamiento estadístico y computacional cuando el número de covariables p no es relativamente grande en comparación con el tamaño muestral n, su utilidad está limitada en el contexto de alta dimensión donde $p > n$.

Para solucionar estos problemas Ravikumar et. al. [14] introducen los modelos no paramétricos de regresión conocidos como **SpAM (*Sparse Additive Models*)**, los cuales se encargarán de imponer una penalización tipo L_1 o *LASSO* que permitirá ajustar el modelo aditivo cuando haya más covariables que datos. Estos métodos extienden las ventajas de los modelos lineales dispersos o *sparse* al contexto no paramétrico.

Por simplicidad, para exponer el algoritmo, se va a suponer $\mathbb{E}[Y_i] = 0$ para todo $i = 1, \ldots, n$, condición fácil de verificar sin más que trabajar con las respuestas centradas en caso de que no se cumpla. Ahora el problema a minimizar sería

$$\min_{f_j \in \mathcal{H}_j,\ 1 \leq j \leq p} \mathbb{E}\left(y - \sum_{j=1}^{p} f_j(x_j)\right)^2, \tag{1.42}$$

con $y = (y_1, \ldots, y_n)^t$, donde la esperanza se toma con respecto a X y al ruido ϵ.

En (1.42) $\mathcal{H}_j$ denota el subespacio de Hilbert de $L_2(P)$ de funciones P-medibles $f_j(x_j)$, asociadas a la variable X_j, que tienen media nula, $\mathbb{E}[f_j(X_j)] = 0$. Teniendo en cuenta que P es la distribución conjunta de (X_i, Y_i) y $L_2(P)$ es al norma dada por

$$\|f\| = \sqrt{\int_0^1 f^2(x) dP(x)} = \sqrt{\mathbb{E}[f]^2}.$$

Ahora, si se introduce un parámetro escalar para cada función β_j en (1.42) y se impone la restricción de que a lo sumo el valor absoluto de los parámetro sumen una cantidad $L > 0$ prefijada de antemano, se obtiene un problema donde la solución será de tipo *sparse*. Este problema vendría dado por la estructura

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^p,\ g_j \in \mathcal{H}_j} \mathbb{E}\left(y - \sum_{j=1}^{p} \beta_j g_j(x_j)\right)^2 \\ &s.a.: \\ &\sum_{j=1}^{p} |\beta_j| \leq L, \\ &\mathbb{E}[g_j^2] = 1, \quad j = 1, \ldots, p \end{aligned} \tag{1.43}$$

donde cada g_j es una función mientras que $\beta = (\beta_1, \ldots, \beta_p)^t$ es un vector de parámetros. La restricción impuesta de que $\sum_{j=1}^{p} |\beta_j| \leq L$ se corresponde con la interpretación de que β se encuentre dentro de la bola de tipo L_1 dada por $B_1 = \{\beta : \|\beta\|_1 \leq L\}$. Esta condición proporciona una solución del problema (1.43) de tipo *sparse*, ya que la solución óptima para β se alcanza en la frontera del conjunto B_1, lo cual fuerza a que algunos coeficientes de dicho vector sean nulos, al igual que ocurría en el caso de la regresión LASSO.

Se puede reescribir el problema (1.43) de forma más compacta obteniendo que sería necesario resolver

$$\begin{aligned} & \min_{f_j \in \mathcal{H}_j} \mathbb{E}\left(y - \sum_{j=1}^{p} f_j(x_j)\right)^2 \\ s.a.: & \\ & \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(x_j))} \leq L. \end{aligned} \tag{1.44}$$

De esta forma, viendo la estructura de (1.44) es fácil ver que para obtener la solución de este problema se pueden emplear métodos de penalización como el **método del lagrangiano aumentado** que proporciona el nuevo problema de minimizar la función

$$\mathcal{L}(f, \lambda) = \frac{1}{2}\mathbb{E}\left[y - \sum_{j=1}^{p} f_j(x_j)\right]^2 + \lambda \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(x_j))}. \tag{1.45}$$

Una vez que se tiene una muestra de datos, la solución correspondiente para cada f_j de (1.45) puede ser calculada mediante un *procedimiento de descenso por coordenadas*, el cual fija f_k para todo $k \neq j$ y obtiene f_j a través de la expresión (1.46) proporcionada por el Teorema 1.8 (cuya demostración se muestra en [14]), finalmente se itera sobre j.

Teorema 1.8. *Los $f_j \in \mathcal{H}_j$ que minimizan el problema (1.45) cumplen que*

$$f_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}[P_j^2]}}\right]_+ P_j, \tag{1.46}$$

donde $[.]_+$ denota la parte positiva, y $P_j = \mathbb{E}[\, R_j \mid X_j \,]$ denota la proyección del residuo $R_j = y - \sum_{k \neq j} f_k(x_k)$ sobre $\mathcal{H}_j$.

Para obtener una versión muestral del resultado expuesto basta con insertar estimadores poblacionales en lugar de los parámetros desconocidos, al igual que se hacía en el caso del algoritmo backfitting. De esta forma se empieza estimando la proyección $P_j = \mathbb{E}[\, R_j \mid X_j \,]$ por una versión suavizada del residuo dada por

$$\hat{P}_j = \mathbf{S}_j R_j,$$

donde S_j es un suavizador lineal. Por otro lado el valor de $\mathbb{E}[P_j^2]$ se puede estimar por

$$\hat{s}_j = \frac{1}{\sqrt{n}}\|\hat{P}_j\| = \sqrt{\bar{Q}_j} \quad \text{con} \quad Q_j = \hat{P}_j^2.$$

Usando estas estimaciones en el procedimiento de descenso por coordenadas se está en condiciones de proporcionar un algoritmo para ajustar el modelo SpAM.

Algoritmo 1.9 (SpAM Backfitting).

1. *Se inicializan las funciones* $\hat{f}_j = 0$, *para todo* $j = 1, \ldots, p$.
2. *Para cada* $j = 1, \ldots, p$:
   - 2.1. *Se calcula el residuo* $R_j = y - \sum_{k \neq j} \hat{f}_k(x_k)$.
   - 2.2. *Se estima* P_j *a través de* $\hat{P}_j = \mathbf{S}_j R_j$.
   - 2.3. *Estimación de la norma por* $\hat{s}_j^2 = \frac{1}{n} \sum_{i=1}^n \hat{P}_j^2(i)$.
   - 2.4. *Se obtiene el umbral suave* $\hat{f}_j = [1 - \lambda/\hat{s}_j]_+ \hat{P}_j$.
   - 2.5. *Se centran las funciones estimadas:* $\hat{f}_j = \hat{f}_j - \bar{\hat{f}}_j$.
3. *Se repite el paso* (2) *hasta alcanzar convergencia.*

Este algoritmo no es más que una versión funcional del *algoritmo por descenso de coordenadas* que puede ser empleado para resolver el problema LASSO, Algoritmo 1.4.

En el caso particular de emplear la representación de las funciones f_j sobre bases funcionales (véase (1.30)) se veía que era necesario seleccionar un número adecuado de componentes q_j. Bajo la suposición simple de que las funciones $f_j \in \mathcal{T}_j$, siendo $\mathcal{T}_j$ un espacio de Sobolev de orden dos definido por

$$\mathcal{T}_j = \left\{ f_j \in \mathcal{H}_j : \ f_j(x_j) = \sum_{q=1}^{\infty} b_{jq}(x_j)\beta_{jq}, \quad \sum_{q=1}^{\infty} b_{jq}^2 q^{2\nu_j} \leq C^2 \right\}$$

para algún $0 < C < \infty$ y tomando normalmente $\nu_j = 2$, se tiene que $\|\tilde{f}_j - f_j\| = \mathcal{O}(1/d^4)$, denotando $\tilde{f}$ la representación de $f_j(x)$ sobre todos los elementos de la base seleccionada y f_j la representación sobre el subconjunto de cardinal q_j. Sea $S = \{j : \ \tilde{f}_j \neq 0\}$. Si se asume la condición de *sparsity* se tiene que $|S| = \mathcal{O}(1)$ y se sigue que $\|\tilde{m} - m\| = \mathcal{O}(1/d^4)$, donde $m = \sum_j f_j$. De esta forma una elección adecuada para q_j es tomar $q_j \backsimeq n^{1/5}$ para todo $j = 1, \ldots, p$ para garantizar que el sesgo producido por el truncamiento es de orden $\|\tilde{m} - m\| = \mathcal{O}(n^{-4/5})$.

En este caso el Algoritmo (1.9) es exactamente el algoritmo de descenso de coordenadas que minimiza

$$\frac{1}{2n} \left\| y - \sum_{j=1}^{p} \mathbf{B}_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{p} \sqrt{\frac{1}{n} \beta_j^t \mathbf{B}_j^t \mathbf{B}_j \beta_j},$$

tomando ahora como matriz de proyectado $\mathbf{S}_j = \mathbf{B}_j (\mathbf{B}_j^t \mathbf{B}_j)^{-1} \mathbf{B}_j^t$, la cual se encarga del suavizado.

1.3. Regresión lineal generalizada. Problemas en alta dimensión

Los **modelos lineales generalizados** fueron formulados por John Nelder y Robert Wedderburn en 1972 (véase [11] y [12]). Estos modelos buscaban unificar diferentes modelos estadísticos de regresión dentro de un marco global que incluyese las distintas posibilidades de incorporar variables tanto discretas como continuas de distribución no gaussiana en la respuesta. Dentro de este grupo se encuentran modelos como el modelo de regresión logística o el modelo lineal general.

Los modelos lineales generalizados o GLM (*Generalized Linear Models*) hacen uso de una *función link* $g(.)$ para transformar la variable respuesta y conseguir de esta forma expresar los datos

a través de una estructura lineal como la vista en (1.1). Así, se tendría que un GLM viene dado a través de la estructura

$$g(\mu_i) = X_i\beta \quad \text{con} \quad i = 1, \ldots, n\,, \tag{1.47}$$

donde $\mu_i \equiv \mathbb{E}[Y_i]$, con Y_i una variable i.i.d. Y. X_i es la i-ésima fila de la matriz del modelo, X, conocida una vez que se obtiene una muestra de las covariables (la cual pasa a denotarse por $\mathbf{X}$), mientras que β vuelve a ser un vector de parámetros desconocidos que será necesario estimar.

Normalmente, en el caso de un modelo GLM se suelen asumir las hipótesis de que las variables respuesta Y_i son *independientes entre sí* y de que estas *pertenecen a alguna distribución de la familia exponencial*, como puede ser la Poisson, Binomial, Gamma o Normal entre otras. Cabe notar que en el caso de la distribución normal se tomaría como función link, $g(.)$, la identidad y se estaría en un contexto de regresión lineal múltiple, por tanto se procedería de la forma ya explicada en la sección anterior.

Estimación de β: método de máxima verosimilitud

Dado que los modelos lineales generalizados se especifican en términos del "predictor lineal" $X\beta$, a la hora de estimar $\hat{\beta}$ muchas de las principales ideas y conceptos del modelo lineal estudiado anteriormente son llevadas a cabo con ligeras modificaciones para conseguir modelizar este caso más general. Esto se debe a que las ideas son las mismas que en el caso lineal expuesto anteriormente, salvo que ahora es necesario determinar una distribución que no tiene que ser necesariamente la gaussiana y una función link adecuada.

Sin embargo, esta generalización implica algún coste. El ajuste del modelo tendrá que hacerse ahora de forma iterativa, y los resultados distribucionales, usados para realizar inferencia, son ahora aproximados y justificados a través de resultados asintóticos, en vez de ser exactos.

Se puede ver que es posible obtener expresiones para la media y varianza de cada una de las distribuciones pertenecientes a la familia exponencial en base a los parámetros a, b y ϕ (Tabla 1.2), lo cual permite relacionar el método de máxima verosimilitud con el de mínimos cuadrados. Para ver esto se empieza considerando la log-verosimilitud de dicha distribución en base al parámetro θ, dado un vector $y = (y_1, \ldots, y_n)^t$, la cual se corresponde con $\log[f_\theta(y)]$. Viendo el Corolario 1.10 se puede ver que esta se denota por

$$l(\theta) = [y\theta - b(\theta)]/a(\phi) + c(y, \phi)$$

y por tanto

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= [y - b'(\theta)]/a(\phi) \\ \frac{\partial^2 l}{\partial \theta^2} &= -b''(\theta)/a(\phi). \end{aligned} \tag{1.48}$$

Corolario 1.10 (familia exponencial)**.** *La función de densidad o función de masa de probabilidad de una distribución perteneciente a la familia exponencial siempre se puede factorizar como*

$$f_\theta(y) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)],$$

donde b, a y c son funciones arbitrarias que pueden variar con i, ϕ es un "factor de escala" arbitrario y θ es el denominado "parámetro canónico". Todos estos parámetros dependen de la distribución considerada.

Si se sustituye el vector de observaciones y por la variable aleatoria Y se tiene que $l(\theta)$ se convierte en una variable aleatoria y por tanto tiene sentido calcular la esperanza de $\partial l(\theta)/\partial\theta$. Ahora utilizando propiedades de la función de densidad o probabilidad, teniendo en cuenta que en el caso de que la variable fuese discreta bastaría cambiar la integral por un sumatorio, se llega a que

$$\begin{aligned}\int f_\theta(Y)dY = 1 \Rightarrow \log\left(\int f_\theta(Y)dY\right) = \log(1) \overset{(a)}{\Rightarrow} \int \log[f_\theta(Y)]dY = 0 \\ \Rightarrow \int l(\theta)dY = 0,\end{aligned} \tag{1.49}$$

donde en (a) se utiliza la continuidad de la función logarítmica sobre $f_\theta(Y) \geq 0$.

En base a lo obtenido en (1.49) se puede ver fácilmente que se verifica

$$\frac{\partial}{\partial\theta}\int l(\theta)dY = 0 \Rightarrow \int \frac{\partial l(\theta)}{\partial\theta}dY = 0 \Rightarrow \mathbb{E}\left[\frac{\partial l(\theta)}{\partial\theta}\right] = 0. \tag{1.50}$$

Teniendo en cuenta (1.48) y (1.50) se concluye que

$$\mathbb{E}\left[\frac{\partial l(\theta)}{\partial\theta}\right] = [\mathbb{E}(Y) - b'(\theta)]/a(\phi) = 0 \Rightarrow \mathbb{E}(Y) = b'(\theta), \tag{1.51}$$

lo cual muestra que la media de cualquier variable aleatoria perteneciente a una distribución de la familia exponencial viene dada por la primera derivada del parámetro específico b, en términos de θ. Esta expresión es la clave que permite relacionar los parámetros β a ajustar en el modelo GLM con los parámetros canónicos de la familia exponencial.

Si se diferencia la expresión (1.49) una vez más se tiene

$$\frac{\partial^2}{\partial\theta^2}\int l(\theta)dY = 0 \Rightarrow \int \frac{\partial^2}{\partial\theta^2}l(\theta)dY = 0 \Rightarrow \mathbb{E}\left[\frac{\partial^2 l(\theta)}{\partial\theta^2}\right] = 0. \tag{1.52}$$

Juntando las expresiones (1.48), (1.51) y (1.52) se llega a que

$$\begin{aligned}&\mathbb{E}\left[\frac{\partial^2 l(\theta)}{\partial\theta^2}\right] + \left(\mathbb{E}\left[\frac{\partial l(\theta)}{\partial\theta}\right]\right)^2 = \frac{-b''(\theta)}{a(\phi)} + \left(\mathbb{E}\left[\frac{Y - b'(\theta)}{a(\phi)}\right]\right)^2 = 0 \\ &\Rightarrow \frac{-b''(\theta)}{a(\phi)} + \frac{(\mathbb{E}[Y - b'(\theta)])^2}{a(\phi)^2} = 0 \Rightarrow \frac{(\mathbb{E}[Y - b'(\theta)])^2}{a(\phi)^2} = \frac{b''(\theta)}{a(\phi)} \\ &\Rightarrow \frac{\mathbb{V}(Y)}{a(\phi)^2} = \frac{b''(\theta)}{a(\phi)} \Rightarrow \mathbb{V}(Y) = b''(\theta)a(\phi).\end{aligned}$$

En un principio a puede ser cualquier función de ϕ, lo cual no da problemas cuando dicho parámetro es conocido. Sin embargo, si este dato es desconocido el desarrollo se vuelve mucho más complejo a menos que se pueda escribir $a(\phi) = \phi/\omega$, donde ω será una constante conocida que puede variar de observación a observación y $\phi = \sigma^2$ es un valor constante que se conoce como *parámetro de dispersión*. Esta última suposición recoge las distribuciones tradicionales pertenecientes a la familia exponencial y además permite heterocedasticidad en la varianza, ya que esta se reescribiría ahora en términos de los parámetros distribucionales como

$$\mathbb{V}(Y) = b''(\theta)\phi/\omega = b''(\theta)\sigma^2/\omega. \tag{1.53}$$

De esta forma, si se supone que la variable Y pertenece a una distribución de la familia exponencial cuya densidad se puede denotar por $f_{\theta i}(y_i)$, es decir $Y \sim f_{\theta i}(y_i)$, con parámetro canónico θ_i este está totalmente determinado por μ_i (véase Tabla 1.2) y por lo tanto por β a través de la relación (1.47). Dado un vector de observaciones de Y, $y = (y_1, \ldots, y_n)^t$, es posible estimar el vector de parámetros β a través del **método de máxima verosimilitud**. Puesto que se ha asumido la hipótesis de que los Y_i son mutuamente independientes se tiene que la verosimilitud de β viene dada por

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i),$$

y por tanto la log-verosimilitud tiene la estructura

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \log[f_{\theta_i}(y_i)] \\ &\overset{(a)}{=} \sum_{i=1}^{n} [y_i\theta_i - b_i(\theta_i)]/a_i(\phi) + c_i(\phi, y_i), \end{aligned}$$

donde la dependencia respecto de β en la última expresión se refleja a través del término θ_i. El paso (a) se deduce a través de las propiedades de la familia exponencial, Corolario 1.10.

En la práctica se van a considerar los casos donde $a_i(\phi) = \phi/\omega_i$, donde ω_i es una constante conocida, en cuyo caso

$$l(\beta) = \sum_{i=1}^{n} \omega_i [y_i\theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i). \tag{1.54}$$

Para maximizar esta última expresión en términos de β basta diferenciar respecto a cada uno de sus componentes e igualar a cero, de modo que se obtendría la expresión

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i) \frac{\partial \theta_i}{\partial \beta_j} \right),$$

y dado que por la regla de la cadena se tiene que

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j},$$

junto con que por (1.51) se sabe que $\mu \equiv \mathbb{E}[Y] = b'(\theta) \Rightarrow \frac{\partial \mu_i}{\partial \theta_i} = b_i''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b_i''(\theta_i)}$, se llega a que

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{[y_i - b_i'(\theta_i)]}{b_i''(\theta_i)/\omega_i} \frac{\partial \mu_i}{\partial \beta_j}.$$

Teniendo en cuenta que $\mathbb{E}[Y] = b'(\theta)$ (1.51) y que $\mathbb{V}(Y) = b''(\theta)\phi/\omega = V(\mu)\phi$ (1.53), donde $V(\mu) = b''(\theta)/\omega$, basta aplicar estos resultados en la última ecuación para llegar a que las ecuaciones que hay que resolver para obtener β son

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall j. \tag{1.55}$$

Sin embargo, estas ecuaciones son las mismas que habría que resolver para hallar β mediante **mínimos cuadrados no lineales ponderados** si los pesos $V(\mu_i)$ fuesen conocidos por adelantado y fuesen independientes de β. En este caso la función objetivo a minimizar, obtenida mediante mínimos cuadrados ponderados, sería

$$S = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}, \tag{1.56}$$

donde los términos μ_i dependen de forma no lineal de β, pero los pesos $V(\mu_i)$ son tratados como fijos. De esta forma se tendría que

$$\frac{\partial S}{\partial \beta_j} = \sum_{i=1}^{n} \frac{-2(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}. \tag{1.57}$$

Encontrar el estimador de mínimos cuadrados involucra resolver $\partial S/\partial \beta_j = 0 \quad \forall j$, pero este sistema de ecuaciones (1.57) es igual que (1.55), una vez que se han determinado los valores de $V(\mu_i)$. Por tanto se hace latente la relación existente entre el estimador obtenido por el método de máxima verosimilitud y el de mínimos cuadrados ponderados, una vez que se conocen los valores $V(\mu_i)$. Esto permite desarrollar un algoritmo que obtiene el estimador de máxima verosimilitud resolviendo iterativamente problemas de mínimos cuadrados (**IRLS**) o empleando métodos numéricos como **el método de Newton-Raphson** para resolver el problema (1.55). A continuación se expone la implementación de ambos así como sus características.

	Normal	**Poisson**	**Binomial**	**Gamma**
$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$	$\frac{\mu^y \exp(-\mu)}{y!}$	$\binom{n}{y} \left(\frac{\mu}{n}\right)^y \left(1 - \frac{\mu}{n}\right)^{n-y}$	$\frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} \exp\left(-\frac{\nu y}{\mu}\right)$
Range	$-\infty < y < \infty$	$y = 0, 1, 2, \ldots$	$y = 0, 1, \ldots, n$	$y > 0$
θ	μ	$\log(\mu)$	$\log\left(\frac{\mu}{n-\mu}\right)$	$-\frac{1}{\mu}$
ϕ	σ^2	1	1	$\frac{1}{\nu}$
$a(\phi)$	$\phi(=\sigma^2)$	$\phi(=1)$	$\phi(=1)$	$\phi\left(=\frac{1}{\nu}\right)$
$b(\theta)$	$\frac{\theta^2}{2}$	$\exp(\theta)$	$n\log(1+e^\theta)$	$-\log(-\theta)$
$c(y,\phi)$	$-\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$	$-\log(y!)$	$\log\binom{n}{y}$	$\nu\log(\nu y) - \log(y\Gamma(\nu))$
$V(\mu)$	1	μ	$\mu(1-\mu/n)$	μ^2
$g_c(\mu)$	μ	$\log(\mu)$	$\frac{\mu}{n-\mu}$	$\frac{1}{\mu}$
$D(y,\hat{\mu})$	$(y-\hat{\mu})^2$	$2y\log\left(\frac{y}{\hat{\mu}}\right) - 2(y-\hat{\mu})$	$2\left[y\log\left(\frac{y}{\hat{\mu}}\right) + (n-y)\log\left(\frac{n-y}{n-\hat{\mu}}\right)\right]$	$2\left[\frac{y-\hat{\mu}}{\hat{\mu}} - \log\left(\frac{y}{\hat{\mu}}\right)\right]$

Tabla 1.2: Algunas distribuciones pertenecientes a la familia exponencial. Hay que tener en cuenta que cuando $x = 0$, entonces $x\log(x/\hat{\mu})$ es remplazado por 0 (su límite cuando $x \to 0$).

Algoritmo iterativo de mínimos cuadrados ponderados (IRLS)

Esta correspondencia sugiere un método iterativo para resolver (1.55). El método es conocido como **método IRLS** (*Iteratively Re-weighted Least Squares*), el cual fue introducido por John Nelder y Robert Wedderburn en 1972 (véase [11] y [12]). Para presentar este algoritmo se denotará $\hat{\beta}^{[k]}$ el estimador del vector de parámetros en la k-ésima iteración, mientras que $\eta^{[k]}$ y $\mu^{[k]}$ serán los vectores con elementos $\eta_i^{[k]} = \mathbf{X}_i\hat{\beta}^{[k]}$ y $\mu_i^{[k]} = g^{-1}(\eta_i^{[k]})$ respectivamente, donde $g^{-1}(.)$ es la inversa de la función link con la que se trabaja. Empezando con una estimación inicial, $\hat{\beta}^{[0]}$, los siguientes pasos son iterar hasta que se obtenga convergencia en la secuencia de $\hat{\beta}^{[k]}$:

Algoritmo 1.11 (IRLS)**.**

1. *Calcular los términos $V(\mu_i^{[k]})$ usando el $\hat{\beta}^{[k]}$ actual.*

   Para esto se tiene en cuenta que $\mu_i^{[K]} = g^{-1}(\mathbf{X}_i\hat{\beta}^{[k]})$ y que $V(\mu_i^{[K]})$ es la función de varianza evaluada en $\mu_i^{[K]}$.

2. *Dadas estas estimaciones se empleará el método de aproximación lineal del modelo para minimizar (1.56) con respecto a β, con la finalidad de obtener $\hat{\beta}^{[k+1]}$ (el término $V(\mu_i^{[k]})$ en este paso es tratado como un valor fijo y no como una función de β).*

3. *Fijar k como $k+1$.*

En el paso 2 del algoritmo se habla del *método de aproximación lineal del modelo*, el cual se basa en estimar el parámetro β del modelo de mínimos cuadrados no lineales a través de ir aproximando este sucesivamente por una versión lineal y ajustando esta última. De esta forma en cada iteración del algoritmo el ajuste mediante el modelo lineal aproximado proporciona una mejora en la dificultad de estimación de los parámetros y al alcanzar convergencia se obtienen las estimaciones correspondientes al caso de mínimos cuadrados no lineales del modelo original.

Formalmente, si se quiere ajustar un modelo con estructura

$$\mathbb{E}[Y] \equiv \mu = f(\beta),$$

donde Y es la variable respuesta y f es ahora una función no lineal de valor vectorial en base a β, como ocurre en el caso de considerar una distribución perteneciente a la familia exponencial distinta de la normal, se tiene intuitivamente que una función objetivo a considerar para ajustar este modelo es

$$S = \sum_{i=1}^{n}\{y_i - f_i(\beta)\}^2 = \|y - f(\beta)\|^2,$$

donde y denota al vector de las n observaciones muestrales.

Si las funciones f_i son lo suficientemente suaves entonces este problema no lineal de mínimos cuadrados puede ser resuelto a través de la iteración de mínimos cuadrados lineales. Para hacer esto, es necesario empezar con una estimación inicial $\hat{\beta}^{[k]}$ y después desarrollar la expansión de Taylor de primer orden de cada f_i alrededor de $\hat{\beta}^{[k]}$. Así, la función objetivo pasaría a ser

$$S \approx S^{[k]} = \|y - f(\hat{\beta}^{[k]}) + \mathbf{J}^{[k]}\hat{\beta}^{[k]} - \mathbf{J}^{[k]}\beta\|^2,$$

donde $\mathbf{J}^{[k]}$ es la matriz jacobiana la cual satisface $J_{ij}^{[k]} = \partial f_i/\partial\beta_j$, estando las derivadas evaluadas en $\hat{\beta}^{[k]}$. Si se define el vector de *pseudodatos*

$$z^{[k]} = y - f(\hat{\beta}^{[k]}) + \mathbf{J}^{[k]}\hat{\beta}^{[k]},$$

se puede ver que la función objetivo puede ser reescrita como

$$S^{[k]} = \|z^{[k]} - \mathbf{J}^{[k]}\beta\|^2 = (z^{[k]} - \mathbf{J}^{[k]}\beta)^t(z^{[k]} - \mathbf{J}^{[k]}\beta),$$

teniéndose ahora un problema de mínimos cuadrados lineales en términos de β, el cual puede ser fácilmente minimizado respecto a β (al igual que se hacía en el modelo lineal múltiple), para obtener un mejor estimador de este parámetro, $\hat{\beta}^{[k+1]}$,

$$\hat{\beta}^{[k+1]} = \left((\mathbf{J}^{[k]})^t\mathbf{J}^{[k]}\right)^{-1}(\mathbf{J}^{[k]})^t z^{[k]}.$$

Si la función $f(\beta)$ no difiere mucho de una función lineal, entonces este procedimiento puede ser iterado hasta que la secuencia de $\hat{\beta}^{[k]}$ converja al estimador final de mínimos cuadrados $\hat{\beta}$.

En cambio, si la función f se aleja bastante de la linealidad el algoritmo puede no alcanzar la convergencia. Ante este problema basta implementar una "reducción de pasos" en cada iteración, donde se emplea el vector $\Delta = \hat{\beta}^{[k+1]} - \hat{\beta}^{[k]}$ como dirección de búsqueda de un nuevo estimador $\hat{\beta}^{[k+1]}$ óptimo. Esta *búsqueda lineal* se lleva a cabo cuando $\hat{\beta}^{[k+1]}$ no consigue reducir el valor de S evaluada en $\hat{\beta}^{[k]}$, realizando una nueva búsqueda en la dirección $\hat{\beta}^{[k]} + \alpha\Delta$, tomando cada vez valores más pequeños de $\alpha \in (0,1)$ hasta que S decrezca. Para el cálculo del valor adecuado de α siempre se puede recurrir a métodos como el *método de búsqueda uniforme*, el *método de dicotomía* o el *método de la sección aúrea*. Si además la función S cuenta con buenas propiedades como son garantizar diferenciabilidad y al menos pseudo-convexidad se pueden emplear algoritmos como el *método de bisección* o el *método de Newton*. Geométricamente, esto es equivalente a desarrollar una actualización de los pasos a través del ajuste del promedio $y\alpha + (1-\alpha)f(\hat{\beta}^{[k]})$, en vez de los datos originales y: visto de esta forma es claro que para valores de α suficientemente pequeños, cada iteración debe hacer decrecer el valor de S, hasta que se alcance un mínimo.

Volviendo al esquema de estimación de los parámetros del modelo de regresión lineal generalizado, puede verse que este método es más lento de lo que necesita serlo. El paso 2, por si mismo, involucra iteración, pero parece que no tiene mucho sentido iterar el método de mínimos cuadrados no lineales hasta convergencia antes de que la sucesión de valores $V(\mu_i^{[k]})$ haya alcanzado la convergencia. Debido a esto se suele reemplazar el paso 2 por

2. Usar $\hat{\beta}^{[k]}$ como el valor inicial y realizar únicamente una iteración del *método de aproximación lineal del modelo* para obtener $\hat{\beta}^{[k+1]}$.

La aplicación de este enfoque da como resultado un esquema más compacto y claro. Para ver esto se va a escribir el problema de mínimos cuadrados no lineales ponderados (1.56) en forma matricial. Por tanto, se define la matriz diagonal $\mathbf{V}_{[k]}$ donde $V_{[k]ii} = V(\mu_i^{[k]})$ de modo que el criterio a minimizar se puede expresar a través de la estructura

$$S = \left\| \sqrt{\mathbf{V}_{[k]}^{-1}}[y - \mu(\beta)] \right\|^2$$

y, si se reemplaza μ por su correspondiente desarrollo de Taylor de primer orden centrado en $\hat{\beta}^{[k]}$ se tiene que

$$S \approx \left\| \sqrt{\mathbf{V}_{[k]}^{-1}} \left[y - \mu^{[k]} - \mathbf{J}(\beta - \hat{\beta}^{[k]}) \right] \right\|^2,$$

donde $\mathbf{J}$ es la matriz jacobiana, con elementos $J_{ij} = \partial\mu_i/\partial\beta_j|_{\hat{\beta}^{[k]}}$. Ahora

$$g(\mu_i) = X_i\beta \Rightarrow g'(\mu_i)\frac{\partial\mu_i}{\partial\beta_j} = X_{ij}$$

y por consiguiente

$$J_{ij} = \left.\frac{\partial\mu_i}{\partial\beta_j}\right|_{\hat{\beta}^{[k]}} = X_{ij}/g'(\mu_i^{[k]}).$$

Definiendo $\mathbf{G}$ como la matriz diagonal con elementos $G_{ii} = g'(\mu_i^{[k]})$, $\mathbf{J} = \mathbf{G}^{-1}\mathbf{X}$, se puede expresar S como

$$\begin{aligned} S &\approx \left\| \sqrt{\mathbf{V}_{[k]}^{-1}}\mathbf{G}^{-1}\left[\mathbf{G}(y-\mu^{[k]}) + \eta^{[k]} - \mathbf{X}\beta\right] \right\|^2 \\ &= \left\| \sqrt{\mathbf{W}_{[k]}}(z^{[k]} - \mathbf{X}\beta) \right\|^2, \end{aligned} \tag{1.58}$$

donde, por definición de *pseudodatos*,

$$z_i^{[k]} = g'(\mu^{[k]})(y_i - \mu_i^{[k]}) + \eta_i^{[k]},$$

z es ahora una versión linealizada de la función link aplicada a los datos: $g(y) \backsimeq g(\mu) + (y-\mu)g'(\mu) = \eta + (y-\mu)g'(\mu)$, mientras que $\mathbf{W}$ es una matriz diagonal de pesos cuyos elementos vienen dados por

$$W_{ii}^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}. \tag{1.59}$$

Teniendo estas expresiones en cuenta se tiene el siguiente esquema de iteración

Algoritmo 1.12 (IRLS: approximated linear version)**.**

1. *Usando los valores actuales de $\mu^{[k]}$ y $\eta^{[k]}$ se calculan los pseudodatos $z^{[k]}$ y los pesos iterativos $W^{[k]}$.*
2. *Minimizar la suma de cuadrados $\left\| \sqrt{\mathbf{W}_{[k]}}(z^{[k]} - \mathbf{X}\beta) \right\|^2$ con respecto a β para obtener $\hat{\beta}^{[k+1]}$ y a continuación $\eta^{[k+1]} = \mathbf{X}\hat{\beta}^{[k+1]}$ y $\mu^{[k+1]}$. Incrementar k en una unidad.*

El valor final de convergencia, $\hat{\beta}$, resuelve (1.56) y es por tanto el estimador de máxima verosimilitud de β. Este algoritmo tiene la buena propiedad de que converge en casi todas las circunstancias, aunque hay algunas excepciones como por ejemplo, modelos pobres o excesivamente flexibles de datos binomiales.

Se puede ver que para empezar la iteración se necesita únicamente los valores $\mu^{[0]}$ y $\eta^{[0]}$, pero no $\hat{\beta}^{[0]}$. Debido a esto la iteración se suele inicializar fijando $\mu_i^{[0]} = y_i$ y $\eta_i^{[0]} = g(\mu_i^{[0]})$, con un ligero ajuste de $\mu_i^{[0]}$, según sea necesario, para evitar valores no finitos de $\eta_i^{[0]}$.

En el caso de alta dimensión donde el número de covariables es mayor que el tamaño muestral, $p > n$, este método presenta problemas a la hora de estimar β. Para entender las dificultades que surgen basta ver que minimizar la suma de cuadrados (1.58) del paso 2 del algoritmo es equivalente a resolver el problema

$$\frac{\partial}{\partial\beta}\left\| \sqrt{\mathbf{W}^{[k]}}(z^{[k]} - \mathbf{X}\beta) \right\|^2 = 0 \Rightarrow \frac{\partial}{\partial\beta}\left(\sqrt{\mathbf{W}^{[k]}}(z^{[k]} - \mathbf{X}\beta)\right)^t \left(\sqrt{\mathbf{W}^{[k]}}(z^{[k]} - \mathbf{X}\beta)\right) = 0. \tag{1.60}$$

Teniendo en cuenta que

$$\begin{aligned} \left(\sqrt{\mathbf{W}^{[k]}}(z^{[k]} - \mathbf{X}\beta)\right)^t \left(\sqrt{\mathbf{W}^{[k]}}(z^{[k]} - \mathbf{X}\beta)\right) &= (z^{[k]} - \mathbf{X}\beta)^t \left(\sqrt{\mathbf{W}^{[k]}}\right)^t \sqrt{\mathbf{W}^{[k]}}(z^{[k]} - \mathbf{X}\beta) \\ &\overset{(a)}{=} (z^{[k]} - \mathbf{X}\beta)^t\mathbf{W}^{[k]}(z^{[k]} - \mathbf{X}\beta) = ((z^{[k]})^t - \beta^t\mathbf{X}^t)\mathbf{W}^{[k]}(z^{[k]} - \mathbf{X}\beta) \\ &= \left(z^{[k]}\right)^t \mathbf{W}^{[k]}z^{[k]} - \left(z^{[k]}\right)^t \mathbf{W}^{[k]}\mathbf{X}\beta - \beta^t\mathbf{X}^t\mathbf{W}^{[k]}z^{[k]} + \beta^t\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X}\beta, \end{aligned}$$

donde en (a) se tiene en cuenta que $\mathbf{W}^{[k]}$ es una matriz diagonal, resolver (1.60) equivale a que

$$\begin{aligned} -\left(z^{[k]}\right)^t \mathbf{W}^{[k]}\mathbf{X} + \left(\beta^{[k+1]}\right)^t \mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X} = 0 &\Rightarrow \left(\beta^{[k+1]}\right)^t \mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X} = \left(z^{[k]}\right)^t \mathbf{W}^{[k]}\mathbf{X} \\ &\Rightarrow \mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X}\beta^{[k+1]} = \mathbf{X}^t\mathbf{W}^{[k]}z^{[k]} \\ &\Rightarrow \beta^{[k+1]} = (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}^{[k]}z^{[k]}. \end{aligned} \tag{1.61}$$

De esta forma se puede ver que en cada iteración del algoritmo IRLS se obtiene $\beta^{[k+1]}$ a través de la expresión (1.61), haciéndose latente la necesidad de que exista la inversa de $\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X}$. Puesto que dicha matriz es semidefinida positiva, pues puede expresarse como $\left(\sqrt{\mathbf{W}^{[k]}}\mathbf{X}\right)^t\left(\sqrt{\mathbf{W}^{[k]}}\mathbf{X}\right)$, se tiene que todos sus autovalores serán mayores o iguales que cero. Por tanto, el problema surge en el caso de que $p > n$, ya que la matriz heredaría la propiedad de no invertibilidad de $\mathbf{X}^t\mathbf{X}$ debida a su mal condicionamiento, haciendo que no sea posible obtener $(\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}$ y en consecuencia que no se pueda despejar $\beta^{[k+1]}$.

Además, dado que no se ha supuesto la hipótesis de normalidad en las variables no se tendrá una distribución exacta para $\hat{\beta}$, como ocurría en el caso del modelo lineal simple, por el contrario será necesario recurrir a resultados aproximados. Como el método IRLS se basa en la estimación de máxima verosimilitud de β, mediante un algoritmo iterativo, nos podemos apoyar en las propiedades de los estimadores de máxima verosimilitud y afirmar que este estimador tiene una distribución asintótica normal de la forma

$$\hat{\beta} \sim N(\beta, \boldsymbol{\mathcal{I}}^{-1}) \equiv N(\beta, (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}\phi) \equiv N(\beta, (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}\sigma^2), \tag{1.62}$$

donde $\boldsymbol{\mathcal{I}}^{-1}$ es la matriz inversa de la información de Fisher, la cual en este caso se corresponde con $\boldsymbol{\mathcal{I}}^{-1} = (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}\phi = (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}\sigma^2$.

En consecuencia se ve que este estimador es asintóticamente insesgado.

Como la varianza de $\hat{\beta}$ viene expresada a través de la inversa del producto matricial $\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X}$, como se vio en (1.62), se tiene que en el caso de alta dimensión considerado esta matriz tendrá autovalores nulos, lo cual hace que el estimador tenga mucha variabilidad en ciertas direcciones y por lo tanto que sea muy poco preciso. De esta forma, pese a contar con la propiedad de insesgadez asintótica, se tiene un estimador muy inexacto y que por lo tanto es incapaz de ajustarse al verdadero valor de β.

Algoritmo de Newton-Raphson

Otro modo de hallar el estimador β del modelo es resolver el sistema (1.55) a través de métodos numéricos como pueden ser el **método de Newton-Raphson**, el cual se apoya en la información proporcionada por las derivadas para llegar al valor óptimo.

Puesto que cada μ_i no es necesariamente una función lineal respecto de β_j se tiene que el problema (1.55) no posee una solución explícita. Una forma de conseguir obtener un óptimo del problema es emplear técnicas numéricas que utilicen la información proporcionada por la derivada para hallar el valor de β. El método de Newton-Raphson se basa en obtener los nuevos estimadores a través del esquema

$$\beta_{k+1} = \beta_k - \boldsymbol{\mathcal{H}}(\beta_k)^{-1}\nabla f(\beta_k) \tag{1.63}$$

hasta alcanzar convergencia, donde β_k es el iterante anterior, $\boldsymbol{\mathcal{H}}$ es la derivada de (1.55), la cual se corresponde con la matriz hessiana de (1.54) y $\nabla f(\beta_k)$ representa la ecuación (1.55) evaluada

en β_k, siendo el gradiente de la expresión (1.54).

Para desarrollar el esquema anterior en términos de los parámetros del modelo basta obtener las expresiones de la matriz hessiana y del gradiente en base a estos. Para esta finalidad se puede ver que (1.55) se puede reescribir como

$$\frac{\partial l}{\partial \beta_j} = 0 \overset{(a)}{\Rightarrow} \frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{b_i''(\theta_i)/\omega_i}\frac{d\mu_i}{d\eta}x_j = 0 \overset{(b)}{\Rightarrow} \sum_{i=1}^{n}(y_i-\mu_i)W_{ii}\frac{d\eta}{d\mu_i}x_j = 0 \tag{1.64}$$

donde en (a) se está teniendo en cuenta que $\mu_i = b_i'(\theta_i)$ y se está aplicando la regla de la cadena $\frac{\partial \mu_i}{\partial \beta_j} = \frac{d\mu_i}{d\eta}\frac{\partial \eta}{\partial \beta_j}$ con $\eta = \sum_{j=1}^{p}\beta_j x_j \Rightarrow \frac{\partial \eta}{\partial \beta_j} = x_j$. En (b) se está multiplicando y dividiendo por $\left(\frac{d\eta}{d\mu_i}\right)^2$ y se está definiendo $W_{ii}^{-1} = \left(\frac{d\eta}{d\mu_i}\right)^2 \frac{b_i''(\theta_i)}{\omega_i} \equiv \frac{1}{V(\mu_i)g'(\mu_i)^2}$.

Ahora, teniendo esto en cuenta es fácil obtener los elementos de la matriz hessiana, pues vienen dados por

$$\frac{\partial^2 l_r}{\partial \beta_j^2} = \sum_{i=1}^{n}\left[(y_i-\mu_i)\frac{\partial}{\partial \beta_j}\left\{W_{ii}\frac{d\eta}{d\mu_i}\right\} + W_{ii}\frac{d\eta}{d\mu_i}x_r\frac{\partial}{\partial \beta_j}(y_i-\mu_i)\right].$$

Como esta expresión no es fácil de obtener, muchas veces se aproxima la matriz hessiana por la esperanza de esta dando lugar al algoritmo denominado **fisher scoring**, una variante del método de Newton-Raphson para estimación por máxima verosimilitud. En el caso de estar empleando las funciones link canónicas, como las vistas en la Tabla 1.2, se tiene que ambas expresiones coinciden. Esto es debido a que en este contexto la función lineal ponderada $W\frac{d\eta}{d\mu}$ en la ecuación de máxima verosimilitud es una constante. De esta forma se tendría que se puede substituir la matriz hessiana por la matriz cuyos términos se definen a través de

$$\mathbb{E}\left[\frac{\partial^2 l_r}{\partial \beta_j^2}\right] \overset{(a)}{=} -\sum_{i=1}^{n}W_{ii}\frac{d\eta}{d\mu_i}x_r\frac{\partial \mu}{\partial \beta_j} \overset{(b)}{=} -\sum_{i=1}^{n}W_{ii}x_r x_j = -\mathbf{X}^t\mathbf{W}\mathbf{X} \tag{1.65}$$

donde (a) se debe a que $\mathbb{E}\left[\sum_{i=1}^{n}(y_i-\mu_i)\right] = 0$ y en (b) se tiene que $\mu = \sum_{j=1}^{p}x_j\beta_j$.

De esta forma se ve que la esperanza de la matriz hessiana, $\boldsymbol{\mathcal{H}}$, viene dada por $\boldsymbol{\mathcal{H}} = -\mathbf{X}^t\mathbf{W}\mathbf{X}$, siendo la opuesta de la matriz de la información de Fisher salvo una constante. Si se define $\mathbf{A} = \mathbf{X}^t\mathbf{W}\mathbf{X}$ y se reemplaza este valor en (1.63) se tiene que

$$\beta_{k+1} = \beta_k + \mathbf{A}^{-1}\nabla f(\beta_k) \Rightarrow \mathbf{A}\beta_{k+1} = \mathbf{A}\beta_k + \nabla f(\beta_k). \tag{1.66}$$

Ahora, dado que

$$(A\beta_k)_r = \sum_{s=1}^{p}A_{rs}\beta_s \overset{(a)}{=} \sum_{i=1}^{n}W_{ii}x_r\eta,$$

teniendo (a) por

$$A_{rs} = \sum_{i=1}^{n}W_{ii}x_r x_s \Rightarrow A_{rs}\beta_s = \sum_{i=1}^{n}W_{ii}x_r x_s\beta_s \Rightarrow \sum_{s=1}^{p}A_{rs}\beta_s = \sum_{s=1}^{p}\sum_{i=1}^{n}W_{ii}x_r x_s\beta_s$$
$$\Rightarrow \sum_{s=1}^{p}A_{rs}\beta_s = \sum_{i=1}^{n}W_{ii}x_r\underbrace{\sum_{s=1}^{p}x_s\beta_s}_{\eta},$$

se llega a que

$$(\mathbf{A}\beta_{k+1})_r \overset{(a)}{=} \sum_{i=1}^{n} W_{ii}x_r \underbrace{\eta + (y_i - \mu_i \frac{d\eta}{d\mu_i})}_{z} = \sum_{i=1}^{n} W_{ii}x_r z = \mathbf{X}^t\mathbf{W}z,$$

aplicando en (a) las expresiones obtenidas en (1.66) y (1.64).

Por tanto se llega a que el esquema de iteración sigue la forma

$$\beta^{[k+1]} = \mathbf{A}^{-1}\mathbf{X}^t\mathbf{W}^{[k]}z^{[k]} = (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}^{[k]}z^{[k]},$$

la cual coincide con el esquema obtenido para el método IRLS visto en (1.61). De esta forma se hace latente que es equivalente emplear uno u otro método para obtener el parámetro β.

Dado que la matriz hessiana $\boldsymbol{\mathcal{H}}(\beta_k) = -\mathbf{X}^t\mathbf{W}_k\mathbf{X}$ es semidefinida negativa por construcción se podrá asegurar que el método de Newton-Raphson conduce a un mínimo del problema (1.55) cuando se pueda garantizar que todos sus autovalores son no nulos, sin más que partir de un buen iterante inicial. En el caso de alta dimensión considerado, $p > n$, pueden aparecer problemas, ya que se tendrán autovalores de la matriz hessiana con valor cero, haciendo que esta no sea invertible y que por lo tanto no se pueda implementar este método.

En este contexto será necesario recurrir a regularizaciones sobre la función de verosimilitud que se quiere optimizar para poder obtener el vector β. A continuación se proponen algunas penalizaciones que permiten obtener un estimar $\hat{\beta}$ en el modelo GLM cuando el número de covariables es mayor que el de muestras. Finalmente, en la siguiente sección se mostrará este inconveniente en el caso particular de la regresión logística, la cual será útil como regla de clasificación en el contexto de $p > n$, en esta se plantearán soluciones que permitan desarrollar el procedimiento.

1.3.1. Regularización L_2

Una forma de solucionar el problema de la no invertibilidad de la matriz $\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X}$, debido a su mal condicionamiento cuando $p > n$, es obtener el estimador β de la máxima verosimilitud penalizada. Se puede imponer una penalización cuadrática en (1.54) al igual que se hace en el caso de la regresión Ridge en el modelo lineal múltiple (Sección 1.1.1) llegando a que el problema a resolver sería

$$\max_{\beta} \left\{ \sum_{i=1}^{n} (\omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)) - \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{1.67}$$

con $\lambda > 0$.

También se puede ver que el problema (1.67) se puede formular como

$$\hat{\beta}^{\mathrm{RR}} = \max_{\beta} \sum_{i=1}^{n} (\omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)),$$
$$\text{sujeto a } \sum_{j=1}^{p} {\beta_j}^2 \leq t,$$

donde al igual que ocurría en la regresión Ridge no se penaliza el intercepto y se estandarizan los predictores para que la penalización sea significativa.

Este es un criterio cóncavo, por tanto se puede hallar una solución usando métodos de programación no lineal, sin más que trabajar con el problema de minimización convexo equivalente como

se expone en el corolario 1.13.

Corolario 1.13 (Equivalencia concavidad-convexidad)**.** *Dada una función cóncava f y el problema*

$$\max_x f(x)$$

este es equivalente a

$$\min_x -f(x).$$

Teniendo en cuenta todos los desarrollos presentados tanto en el caso del *método IRLS* como en el *algoritmo de Newton-Raphson* es fácil ver que para este problema se tendría el esquema

$$\beta^{[k+1]} = (\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X} - \lambda\mathbf{I}_p)^{-1}\mathbf{X}^t\mathbf{W}^{[k]}z^{[k]},$$

donde ahora la correspondiente matriz hessiana, $\mathbf{X}^t\mathbf{W}^{[k]}\mathbf{X} - \lambda\mathbf{I}_p$, se puede asegurar que es definida negativa para un valor de λ adecuado, consiguiendo que se pueda definir su inversa de forma única y solucionando el problema de no invertibilidad.

De esta forma la implementación de una penalización de tipo L_2 mantiene las buenas propiedades de la regresión Ridge y además permite ajustar un modelo lineal generalizado cuando el número de covariables es mucho mayor que el de muestras.

1.3.2. Regularización L_1

Otra forma de proceder es implementar una penalización de tipo L_1, análoga a la que se imponía en la regresión LASSO (Sección 1.1.2) y que mantendrá sus buenas propiedades. Llevando a cabo esta idea el nuevo problema a resolver sería

$$\max_\beta \left\{ \sum_{i=1}^{n} (\omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)) - \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{1.68}$$

con $\lambda > 0$ el factor que controla la penalización.

De nuevo, es fácil ver que el problema (1.68) es equivalente a la formulación

$$\hat{\beta}^{\mathrm{RL}} = \max_\beta \sum_{i=1}^{n} (\omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)),$$
$$\text{sujeto a } \sum_{j=1}^{p} |\beta_j| \leq t,$$

donde se vuelve a no penalizar el intercepto y se estandarizan los predictores para que la penalización sea significativa.

Al igual que ocurría aplicando una penalización de tipo L_2 se está ante un problema de maximización cóncavo, siendo posible hallar su solución sin más que resolver el problema equivalente a (1.68) de minimización convexo, como se vio a través del corolario 1.13. Para este fin bastaría con aplicar el algoritmo de Newton u otras técnicas numéricas. Cabe notar que debido a la no diferenciabilidad de la restricción L_1 en este caso no es posible obtener una expresión explícita para $\hat{\beta}$ como pasaba al imponer una penalización de tipo L_2.

De nuevo, la utilización de una penalización de valor absoluto o L_1 hará que varios coeficientes sean nulos, lo cual permitirá discernir cuáles son las variables más importantes en el modelo y por tanto seleccionar un conjunto eficiente de ellas para explicar el comportamiento de la variable respuesta.

1.3.3. Regularización Elastic Net

Finalmente, siguiendo el esquema del modelo lineal múltiple, se presentará una penalización de tipo Elastic Net para el modelo lineal generalizado (GLM). Al igual que se presentó en la Sección 1.1.3 esta será un compromiso entre las penalizaciones L_2 y L_1, permitiendo obtener un estimador $\hat{\beta}$ y reduciendo el número de variables explicativas que entran en juego en dicho modelo.

En este contexto aparece un parámetro $\alpha \in (0,1)$ en el problema a resolver, el cual tendrá que ser determinado de antemano y es el que controla el aporte de cada una de las penalizaciones. De esta forma para valores de α cercanos a cero tendrá mayor peso la penalización L_2 mientras que si su valor es próximo a la unidad primará la penalización L_1. Teniendo esto en cuenta el problema que será necesario resolver es

$$\max_{\beta} \left\{ \sum_{i=1}^{n} (\omega_i [y_i \theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)) - \lambda \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2) \right\}, \tag{1.69}$$

donde $\lambda > 0$.

Se tiene un problema equivalente a (1.69) dado por

$$\begin{aligned} \hat{\beta}^{\text{EN}} &= \max_{\beta} \sum_{i=1}^{n} (\omega_i [y_i \theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)), \\ \text{sujeto a } & \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2) \leq t, \end{aligned}$$

y si se tiene en cuenta que la nueva restricción es un promedio de las penalizaciones L_1 y L_2 se puede garantizar que (1.69) es de nuevo un problema de maximización cóncavo que puede ser resuelto de las formas ya comentadas en los casos de las penalizaciones de tipo Ridge o LASSO.

1.3.4. Regresión LAR

Al igual que en el caso del modelo lineal múltiple se pueden extender las ideas de la **regresión LAR** al modelo de regresión lineal generalizado.

Para este fin basta tener en cuenta que a la hora de estimar los parámetros del modelo se aplica el Algoritmo 1.12, basado en resolver de forma iterativa un problema de mínimos cuadrados ponderados, hasta obtener convergencia. Dado que en cada iteración este procedimiento busca optimizar un problema de mínimos cuadrados, se puede proceder a su resolución mediante el algoritmo empleado en el ajuste de la regresión LAR, resolviendo cada uno de estos problemas mediante el Algoritmo 1.5. De esta forma se ve que en cada una de las k iteraciones, en el segundo paso del Algoritmo 1.12, en vez de minimizar la suma de cuadrados $\left\| \sqrt{\mathbf{W}_{[k]}} (z^{[k]} - \mathbf{X}\beta) \right\|^2$ se lleva a cabo el Algoritmo 1.5 tomando $y = \sqrt{\mathbf{W}_{[k]}} z^{[k]}$ y las variables explicativas con valores correspondientes a los de la matriz $\mathbf{X} = \sqrt{\mathbf{W}_{[k]}}\mathbf{X}$.

Se obtiene así para cada iteración k-ésima un vector $\hat{\beta}^{[k+1]}$ que cuenta con únicamente mín$(n-1, p)$ valores no nulos.

1.4. Regresión logística. Problemas en alta dimensión

Como caso particular de los modelos lineal generalizados se va a mostrar el desarrollo de la **regresión logística**, la cual cuenta con gran importancia dentro del contexto de clasificación (Sec-

ción 2) y por tanto tiene gran interés conocer sus características y ventajas.

La **regresión logística** estudia problemas de regresión donde la variable respuesta es discreta a diferencia de la regresión lineal donde se consideraba que esta era continua. Este modelo se crea con el objetivo de conseguir afrontar una toma de decisiones ante dos posibles sucesos. Un ejemplo de esta finalidad sería emplearlo para dictaminar si un paciente padece o no una enfermedad en base a los valores de ciertos parámetros especiales relacionados con esta.

De este modo se considera el caso donde la variable respuesta Y es una **variable binaria (o dicotómica)**, es decir, sólo puede tomar dos valores. Se representarán estos valores por 0 y 1 respectivamente.

Por ser la variable respuesta binaria su distribución es una Bernoulli cuya media será una probabilidad de éxito, obteniéndose

$$\mathbb{E}(\, Y_i \mid X = x_i \,) = \mathbb{P}(\, Y_i = 1 \mid X = x_i \,), \tag{1.70}$$
$$\mathbb{V}(\, Y_i \mid X = x_i \,) = \mathbb{P}(\, Y_i = 1 \mid X = x_i \,)[1 - \mathbb{P}(\, Y_i = 1 \mid X = x_i \,)], \tag{1.71}$$

siendo ahora Y_i el valor de la variable respuesta para el individuo $i-$ésimo y $x_i \in \mathbb{R}^p \quad \forall p \in \mathbb{N}$ un vector que contiene los valores de las variables explicativas para $i \ \in \{1, \ldots, n\}$.

Un modelo lineal representaría la respuesta de la siguiente manera:

$$Y_i = x_i^t \beta + \epsilon_i \quad \text{para } i \ \in \{1, \ldots, n\}$$

donde β es el vector de coeficientes y ϵ_i el error.

Este modelo es inadecuado ya que no se cumplen las suposiciones básicas que se tenían en cuenta para construir el modelo lineal múltiple. No se tiene linealidad, pues si se intenta expresar la media de Y como una función lineal, es decir $\mathbb{E}(\, Y_i \mid X = x_i \,) = x_i^t \beta$, provoca que en muchos casos las predicciones no estén en el soporte $[0, 1]$ en el cual tienen que estar siempre, ya que (1.70) es una probabilidad de éxito. Por otra parte, ahora el modelo es heterocedástico ya que fijándonos en (1.71) se observa que la varianza no es constante, ya que esta depende de cada observación. Finalmente se ve que al estar ante una distribución Bernoulli no se cumple la normalidad y que la hipótesis de independencia sería la única que podría cumplirse y será necesario que lo haga para el modelo logístico. Bajo estas condiciones se hace obvia la necesidad de considerar otro modelo.

Por tanto, un enfoque diferente que consiga solucionar estos problemas se basa en buscar construir un modelo para la probabilidad de éxito condicionada a cada valor de la variable explicativa, la cual coincide con la función

$$\pi(x) = \mathbb{P}(\, Y = 1 \mid X = x \,).$$

Si ahora se desea considerar un modelo lineal habría que aplicar a $\pi(x)$ una función que transforme el intervalo $[0, 1]$ de definición en toda la recta real, de forma que se trabajaría con

$$g(\pi(x, \beta)) = x^t \beta,$$

donde g es la **función link** de los modelos lineales generalizados. En esta situación, teniendo una variable respuesta dicotómica es común considerar la **función logit** que tiene la forma

$$g(p) = \log\left(\frac{p}{1-p}\right) \quad \forall p \in [0, 1],$$

donde sustituyendo p por la probabilidad de éxito, aplicar la función logit se basa en aplicar un logaritmo (en base e) al cociente entre la probabilidad de éxito y la probabilidad de fracaso. Este cociente se conoce como la **odds** (disparidad o ventaja)

$$Odds(Y) = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}.$$

Mientras que la probabilidad de éxito toma valores en $[0,1]$ la odds los toma en $[0,+\infty)$. Ahora al aplicar el logaritmo a la odds se tiene como soporte $(-\infty,+\infty)$ siendo, por consiguiente, susceptible de ser explicada mediante un modelo lineal. De esta forma, el modelo consistirá en expresar el logaritmo de la odds de la variable respuesta como función lineal de la variable explicativa

$$\log\left(\frac{\pi(x,\beta)}{1-\pi(x,\beta)}\right) = x^t\beta.$$

Si se piensa ahora el modelo como la representación de la probabilidad de éxito se ve que es necesario invertir la función logit. Esto es posible ya que produce una correspondencia biunívoca y creciente entre $[0,1]$ y el intervalo $(-\infty,+\infty)$, siendo tanto ella como su inversa funciones suaves (derivables con derivadas continuas). La inversa de la función logit es

$$g^{-1}(x) = \frac{e^x}{1+e^x}.$$

Por tanto, el modelo logístico va a expresar la probabilidad de éxito de la siguiente manera

$$\pi(x,\beta) = g^{-1}(x,\beta) = \frac{e^{x^t\beta}}{1+e^{x^t\beta}} > 0, \tag{1.72}$$

la cual va a tomar siempre un valor positivo y mayor que cero, puesto que $e^{x^t\beta} > 0$.

Se observa en (1.72) que si $x^t\beta = 0$ la probabilidad será de 0.5; si $x^t\beta > 0$ esta será mayor que 0.5 teniendo como asíntota el valor 1 y recíprocamente si es menor será más pequeña de 0.5 teniendo como asíntota el valor 0. Por lo tanto el modelo logístico parece adecuado para representar el comportamiento de la función de regresión como probabilidad de éxito de la variable respuesta dicotómica.

A la hora de explicar y dotar de interpretación los coeficientes β de la regresión logística se va a empezar considerando una situación sencilla en la que la variable explicativa X es discreta y sólo toma dos valores. Esto se puede hacer separando los datos en dos grupos ya que basta codificar por $X=0$ el primer grupo y $X=1$ el segundo, de forma que el modelo de regresión adopta la estructura

$$\pi(0,\beta) = \mathbb{P}(\,Y=1 \mid X=0\,) = \frac{e^{\beta_0}}{1+e^{\beta_0}},$$
$$\pi(1,\beta) = \mathbb{P}(\,Y=1 \mid X=1\,) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}},$$

siendo β_0 el intercepto asociado al grupo de referencia y β_1 el parámetro que representa el incremento al pasar del primer grupo al segundo.

Si se quiere estimar la probabilidad de éxito para el grupo de referencia, $X=0$, se puede ver que esta es de la forma

$$\hat{p}_0 = \frac{\sum_{i=1}^n I_{\{x_i=0\}}Y_i}{\sum_{i=1}^n I_{\{x_i=0\}}}.$$

Pensando que β será tal que $\hat{p}_0 = \pi(0, \hat{\beta})$, como cabe esperar, para obtener el valor de $\hat{\beta}_0$ basta tener en cuenta que

$$e^{\hat{\beta}_0} = \frac{\pi(0,\hat{\beta})}{1-\pi(0,\hat{\beta})} = \frac{\hat{p}_0}{1-\hat{p}_0} \Rightarrow \hat{\beta}_0 = \log\left(\frac{\hat{p}_0}{1-\hat{p}_0}\right).$$

De esta forma el intercepto β_0 resulta ser el logaritmo de la odds de la variable respuesta condicionada al grupo de referencia. Si β_0 es negativo, la probabilidad de éxito es menor de 0.5; mientras que si es positivo es mayor de 0.5 y si fuese nulo sería exactamente 0.5.

Actuando de manera similar con el otro grupo se tiene

$$e^{\hat{\beta}_0+\hat{\beta}_1} = \frac{\pi(1,\hat{\beta})}{1-\pi(1,\hat{\beta})} = \frac{\hat{p}_1}{1-\hat{p}_1}$$

de forma que $e^{\hat{\beta}_0+\hat{\beta}_1}$ es la odds correspondiente al segundo grupo. Ahora para extraer el parámetro $\hat{\beta}_1$ es necesario efectuar el cociente entre ambas odds

$$e^{\hat{\beta}_1} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{e^{\hat{\beta}_0}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}$$

este cociente se conoce como la **odds ratio** que se define para dos poblaciones respecto de una variable binaria Y como el cociente de la odds en una y otra población, de la forma

$$OddsRatio = \frac{Odds(\,Y \mid \text{Prob.2}\,)}{Odds(\,Y \mid \text{Prob.1}\,)} = \frac{\mathbb{P}(\,Y=1 \mid \text{Prob.2}\,)/\mathbb{P}(\,Y=0 \mid \text{Prob.2}\,)}{\mathbb{P}(\,Y=1 \mid \text{Prob.1}\,)/\mathbb{P}(\,Y=0 \mid \text{Prob.1}\,)}$$

y se estima por

$$\widehat{Odds\ Ratio} = e^{\hat{\beta}_1} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}.$$

Esta indica el incremento de la odds del primer grupo al pasar al segundo grupo. Los coeficientes de regresión pueden presentar signo negativo o positivo, mientras que la odds y la odds-ratio son siempre cantidades positivas, y se valora si son mayores o menores que uno.

Bajo este modelo de regresión logística simple, en el caso particular de una variable explicativa continua, a diferencia del caso discreto binario considerado, no hay expresiones explícitas para los estimadores de los parámetros β_0 y β_1, pero estos se pueden interpretar a través de la odds y de la odds-ratio. Así, bajo este modelo

$$Odds(X=x_0) = \frac{e^{\beta_0+\beta_1 x_0}/(1+e^{\beta_0+\beta_1 x_0})}{1-e^{\beta_0+\beta_1 x_0}/(1+e^{\beta_0+\beta_1 x_0})} = e^{\beta_0+\beta_1 x_0},$$

$$Odds\ Ratio(X=x_0+1 \text{ frente a } X=x_0) = \frac{e^{\beta_0+\beta_1(x_0+1)}}{e^{\beta_0+\beta_1 x_0}} = e^{\beta_1},$$

de modo que e^{β_1} representa ahora la odds debida a haber incrementado la variable explicativa en una unidad, pero ahora a diferencia de la regresión lineal hay que tener en cuenta que se incrementa de manera multiplicativa.

Estimación de los parámetros del modelo

Como ya se explicó en el caso del modelo de regresión lineal generalizado y puesto que el modelo de regresión logística es un caso particular de este conjunto, se emplea el **método de máxima verosimilitud** para estimar los parámetros del modelo.

Se parte de una muestra aleatoria simple

$$(x_1, y_1), \ldots, (x_n, y_n)$$

donde cada $y_i \in \text{Bernoulli}(\pi(x_i, \beta))$ de forma que su función de densidad es

$$f(y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \text{ con } y_i \in \{0, 1\},$$

siendo $p_i = \pi(x_i, \beta)$ la probabilidad de éxito. De esta manera la función de verosimilitud es:

$$L(\beta) = \prod_{i=1}^{n} \left[\pi(x_i, \beta)^{y_i}(1 - \pi(x_i, \beta))^{1-y_i}\right]$$

y su logaritmo

$$\log(L(\beta)) = \sum_{i=1}^{n} \left[y_i \log(\pi(x_i, \beta)) + (1 - y_i)\log(1 - \pi(x_i, \beta))\right]. \tag{1.73}$$

Teniendo en cuenta que

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\left[y_i \log(\pi(x_i, \beta)) + (1 - y_i)\log(1 - \pi(x_i, \beta))\right] \\
&= y_i \frac{1}{\pi(x_i, \beta)} \frac{\partial \pi(x_i, \beta)}{\partial \beta} + (1 - y_i)\frac{1}{1 - \pi(x_i, \beta)} \frac{-\partial \pi(x_i, \beta)}{\partial \beta} \\
&= \frac{\partial \pi(x_i, \beta)}{\partial \beta}\left(\frac{y_i}{\pi(x_i, \beta)} - \frac{(1 - y_i)}{1 - \pi(x_i, \beta)}\right) \\
&= \frac{\partial \pi(x_i, \beta)}{\partial \beta}\left(\frac{y_i - y_i\pi(x_i, \beta) - \pi(x_i, \beta) + y_i\pi(x_i, \beta)}{\pi(x_i, \beta)(1 - \pi(x_i, \beta))}\right) \\
&= \frac{\partial \pi(x_i, \beta)}{\partial \beta}\left(\frac{y_i - \pi(x_i, \beta)}{\pi(x_i, \beta)(1 - \pi(x_i, \beta))}\right),
\end{aligned}$$

y si se deriva (1.73) respecto de β se obtiene

$$\frac{\partial \log(L(\beta))}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \pi(x_i, \beta)}{\partial \beta} \frac{1}{\pi(x_i, \beta)(1 - \pi(x_i, \beta))}[y_i - \pi(x_i, \beta)]. \tag{1.74}$$

Si ahora se supone que la función de regresión paramétrica adopta la forma $\pi(x, \beta) = \frac{e^{x^t\beta}}{1+e^{x^t\beta}}$ entonces

$$\begin{cases}
\dfrac{\partial \pi(x, \beta)}{\partial \beta} = \dfrac{x^t e^{x^t\beta}(1 + e^{x^t\beta}) - e^{x^t\beta}x^t e^{x^t\beta}}{(1 + e^{x^t\beta})^2} = \dfrac{x^t e^{x^t\beta}}{(1 + e^{x^{\text{`}}\beta})^2} \\
x^t\pi(x, \beta)(1 - \pi(x, \beta)) = x^t \dfrac{e^{x^t\beta}}{1 + e^{x^t\beta}}\left(1 - \dfrac{e^{x^t\beta}}{1 + e^{x^t\beta}}\right) = x^t \dfrac{e^{x^t\beta}}{1 + e^{x^t\beta}} \dfrac{1 + e^{x^t\beta} - e^{x^{\text{`}}\beta}}{(1 + e^{x^t\beta})} \\
\qquad\qquad = \dfrac{x^t e^{x^t\beta}}{(1 + e^{x^t\beta})^2}
\end{cases}$$

lleva a que

$$\frac{\partial \pi(x, \beta)}{\partial \beta} = x^t\pi(x, \beta)(1 - \pi(x, \beta)). \tag{1.75}$$

Ahora si se sustituye (1.75) en (1.74) y se iguala a cero para obtener el máximo se obtiene la expresión

$$\frac{\partial \log(L(\beta))}{\partial \beta} = \sum_{i=1}^{n} x_i^t [y_i - \pi(x_i, \beta)] = 0. \tag{1.76}$$

Para que β sea el estimador se ha de cumplir que esta derivada esté igualada a cero, por lo que a (1.76) se las conoce como las **ecuaciones de verosimilitud**.

En el caso particular de un modelo logístico simple (variable explicativa unidimensional) para permitir una constante como parámetro del modelo se toma $x = (1, x_1)$, de forma que $x^t\beta = \beta_0 + \beta_1 x_1$. En esta situación se obtendrían las ecuaciones de verosimilitud:

$$\sum_{i=1}^{n} [y_i - \pi(x_i, \beta)] = 0,$$
$$\sum_{i=1}^{n} x_{i,1} [y_i - \pi(x_i, \beta)] = 0,$$

donde la primera ecuación dice que los errores suman cero, lo que se interpreta como que el número de éxitos ha de coincidir con la suma de proporciones estimadas.

Se ve que las ecuaciones de verosimilitud no tienen solución explícita pues $\pi(x, \beta)$ no es una función lineal respecto de β, de modo que para resolver estas ecuaciones va a ser necesario recurrir a la información proporcionada por la matriz hessiana. Esta tiene la expresión

$$\frac{\partial^2 \log L(\beta)}{\partial \beta^2} = -\sum_{i=1}^{n} x_i x_i^t \pi(x_i, \beta)(1 - \pi(x_i, \beta)),$$

donde $\partial\beta^2$ hace referencia a las derivadas parciales del tipo $\partial\beta_i \partial\beta_j$. Se ve que $x_i x_i^t$ es una matriz simétrica, semidefinida positiva y de rango uno, lo que se sigue conservando al multiplicar por $\pi(x_i, \beta)(1 - \pi(x_i, \beta))$, puesto que esta última expresión es mayor que cero teniendo en cuenta la estructura de $\pi(x_i, \beta)$ vista en (1.72). De esta forma la suma será definida positiva cuando los x_i no estén contenidos en un espacio lineal de dimensión inferior, cuya interpretación se reduce a que los vectores x_i sean independientes entre sí. En consecuencia, la matriz hessiana será semidefinida o definida negativa respectivamente, donde en este último caso la raíz de las ecuaciones de verosimilitud será un máximo de la función de verosimilitud, obteniendo así un estimador de máxima verosimilitud.

Denotando por

$$\mathbf{X} = \begin{pmatrix} x_1^t \\ \vdots \\ x_n^t \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \pi(x_1, \beta)(1 - \pi(x_1, \beta)) & & 0 \\ & \ddots & \\ 0 & & \pi(x_n, \beta)(1 - \pi(x_n, \beta)) \end{pmatrix},$$

entonces la matriz hessiana se puede escribir como

$$\frac{\partial^2 \log L(\beta)}{\partial \beta^2} = -\sum_{i=1}^{n} x_i x_i^t \pi(x_i, \beta)(1 - \pi(x_i, \beta)) = -\mathbf{X}^t \mathbf{V} \mathbf{X}. \tag{1.77}$$

Ahora recordando las ecuaciones de verosimilitud (1.76) pero expresándolas en forma matricial se tiene

$$\frac{\partial \log L(\beta)}{\partial \beta} = \mathbf{X}^t(y - \pi(\mathbf{X}, \beta)) = 0,$$

donde estas ecuaciones no tienen una solución explícita. Por tanto, para hallar los estimadores del modelo logístico simple hay que recurrir a procedimientos iterativos como el **método de Newton-Raphson** o el **método IRLS**.

El método de Newton-Raphson se basa en a partir de un iterante inicial (β_0), adecuadamente escogido (pues este método posee convergencia de tipo local), ir calculando los sucesivos iterantes. Un buen iterante inicial podría ser un estimador $\hat{\beta}$ obtenido empleando el método de mínimos cuadrados en este problema. En nuestro caso cada iterante se calcula de la forma:

$$\beta_{k+1} = \beta_k + (\mathbf{X}^t\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^t(y - \pi(\mathbf{X}, \beta)) \quad k \in \{0, 1, 2, \dots\}, \tag{1.78}$$

hasta que se obtiene convergencia.

En alta dimensión surge el problema que al igual que le pasaba a la matriz $\mathbf{X}^t\mathbf{X}$ ahora la matriz $\mathbf{X}^t\mathbf{V}\mathbf{X}$ va a estar mal condicionada y esto puede impedir la convergencia del método de Newton-Raphson así como hacer que la matriz hessiana no sea definida negativa.

1.4.1. Regularización L_2

Al igual que ocurría en el modelo lineal múltiple se va a ilustrar como solucionar el problema del calculo del estimador β imponiendo una penalización L_2 en (1.73) como hicimos en la regresión Ridge (Sección 1.1.1) de forma que se buscaría hallar

$$\max_{\beta} \left\{ \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta^t x_i) - \log(1 + e^{\beta_0 + \beta^t x_i}) \right] - \lambda \sum_{j=1}^{p} \left(\|\beta_j\|_2 \right)^2 \right\} \tag{1.79}$$

o lo que es análogo resolver el problema equivalente

$$\hat{\beta}^{\mathrm{RR}} = \max_{\beta} \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta^t x_i) - \log(1 + e^{\beta_0 + \beta^t x_i}) \right],$$
$$\text{sujeto a } \sum_{j=1}^{p} {\beta_j}^2 \leq t.$$

Como ocurría en la regresión Ridge no se penaliza el intercepto y se estandarizan los predictores para que la penalización sea significativa.

Este es un criterio cóncavo, por tanto se puede hallar una solución usando métodos de programación no lineal, sin más que trabajar con el problema de minimización convexo equivalente como se expone en el corolario 1.13.

De forma alternativa, usando las mismas aproximaciones cuadráticas empleadas en el algoritmo de Newton de la sección anterior se puede resolver (1.80) por repetida aplicación del algoritmo de penalización Ridge. Se ve que la matriz equivalente a la de (1.77) en el caso de maximizar (1.80) deja de ser singular, ya que pasa lo mismo que ocurría en la regresión Ridge con la matriz $\mathbf{X}^t\mathbf{X}$ y por tanto garantiza que siempre se puede estimar β por el método de Newton-Raphson por la forma vista en la sección anterior en (1.78) pero tomando como $\mathbf{X}^t\mathbf{V}\mathbf{X}$ la nueva matriz de este problema .

1.4.2. Regularización L_1

Ahora se va a ilustrar como solucionar el problema de cálculo del estimador β imponiendo una penalización L_1 en (1.73) al igual que se hizo en el caso de la regresión LASSO (1.1.2) de forma que se buscaría hallar

$$\begin{aligned} &\max_{\beta} \left\{ \sum_{i=1}^{n} [y_i \log(\pi(x_i,\beta)) - (1-y_i)\log(1-\pi(x_i,\beta))] - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\ &= \max_{\beta} \left\{ \sum_{i=1}^{n} \left[y_i \beta^t x_i - \log(1+e^{\beta^t x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\ &= \max_{\beta} \left\{ \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta^t x_i) - \log(1+e^{\beta_0+\beta^t x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \end{aligned} \tag{1.80}$$

o lo que es análogo resolver el problema

$$\begin{aligned} &\hat{\beta}^{\mathrm{RL}} = \max_{\beta} \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta^t x_i) - \log(1+e^{\beta_0+\beta^t x_i}) \right], \\ &\text{sujeto a } \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned}$$

El problema resultante de optimización (1.80) es un problema de maximización cóncavo, pero de nuevo puede ser resuelto a través de pasar a resolver el equivalente de minimización convexo. Este puede ser resuelto por el algoritmo de Newton así como por otras técnicas numéricas. Ahora, al igual que se comentaba para la regularización L_1 la matriz correspondiente a la $\mathbf{X}^t\mathbf{V}\mathbf{X}$ de (1.77) obtenida en este problema se va a poder invertir, por lo que se va a poder implementar el algoritmo de Newton-Raphson y obtener la estimación del parámetro β.

1.4.3. Regularización Elastic Net

Al igual que se hizo en la Sección 1.1.3 en el caso del modelo lineal general se puede imponer una penalización elástica (*"Elastic Net"*) para estimar β en el modelo logístico, la cual recoja un compromiso entre la penalización de tipo L_1 y la L_2. Así, la estimación de β imponiendo esta penalización en (1.73) resulta en el problema

$$\max_{\beta} \left\{ \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta^t x_i) - \log(1+e^{\beta_0+\beta^t x_i}) \right] - \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + (1-\alpha) \|\beta_j\|_2^2 \right) \right\}, \tag{1.81}$$

con $\alpha \in (0,1)$.

Es fácil ver que es equivalente hallar la solución de (1.81) a resolver

$$\begin{aligned} &\hat{\beta}^{\mathrm{EN}} = \max_{\beta} \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta^t x_i) - \log(1+e^{\beta_0+\beta^t x_i}) \right], \\ &\text{sujeto a } \sum_{j=1}^{p} (\alpha |\beta_j| + (1-\alpha) {\beta_j}^2) \leq t. \end{aligned}$$

Al igual que ocurría en el caso de las penalizaciones L_1 y L_2, teniendo en cuenta que la nueva restricción impuesta no es más que un promedio de estas, se tiene que (1.81) es de nuevo un problema de maximización cóncavo, debido a que la suma de funciones cóncavas mantiene la concavidad.

De esta forma se puede ver que se puede hallar la solución mediante métodos numéricos como Newton-Raphson y que además esta modificación también resuelve los problemas de no invertibilidad de la matriz $\mathbf{X}^t\mathbf{V}\mathbf{X}$ de (1.77) en el caso donde $p > n$, consiguiendo hallar un estimador de β.

1.4.4. Regresión LAR

De nuevo, al igual que se veía en el caso general de los GLM (Sección 1.3.4), es posible aplicar un enfoque tipo LAR al caso particular de la regresión logística.

En este contexto basta tener en cuenta que se está bajo la situación particular donde la función link g es de tipo *logit*, por tanto es suficiente introducir esta información en la implementación del Algoritmo 1.12 utilizando la regresión LAR (Algoritmo 1.5), como se ha visto en la Sección 1.3.4.

De esta forma se tiene un nuevo enfoque que soluciona los problemas de tener que hallar la inversa de la matriz $\mathbf{X}^t\mathbf{V}\mathbf{X}$, la cual está mal condicionada en el caso de alta dimensión donde se tienen más covariables que datos, $p > n$, como se extrapola de la explicación dada en la Sección 1.1.4.

1.5. Modelos aditivos generalizados. Problemas en alta dimensión

Al igual que ocurre con los modelos lineales generalizados en el caso de la regresión lineal, se puede definir una clase más general de los modelos aditivos.

Un **modelo aditivo generalizado (GAM)** es aquel que tras aplicar una función link g en la variable respuesta se puede expresar esta como un modelo aditivo, es decir como una suma de funciones suaves en las covariables.

De esta forma un modelo aditivo generalizado tiene una estructura

$$g(\mu_i) = \mathbf{X}_i^*\theta + f_1(x_{i1}) + f_2(x_{i2}) + \dots \tag{1.82}$$

donde $\mu_i = \mathbb{E}(y_i)$, siendo y_i la variable respuesta la cual pertenece a una distribución de la familia exponencial. Aquí $\mathbf{X}_i^*$ es la i-ésima fila de la matriz del modelo que recoge cada una de las componentes estrictamente paramétricas, cuyo vector de parámetros es θ, mientras que f_j son las funciones suaves relacionadas con las covariables x_j que presentan una correspondencia no lineal.

1.5.1. Estimación de las funciones f_j

De nuevo será necesario estimar la forma de las funciones f_j, para ello se ilustrará ahora como proceder en el caso generalizado empleando la representación sobre bases funcionales por simplicidad.

Estimación mediante representación en bases funcionales: método de máxima verosimilitud

A la hora de estimar los parámetros de (1.82) basta tener en cuenta la representación de las funciones f_j a través de bases funcionales de la forma $f_j(x_j) = \sum_{q=1}^{q_j} \beta_{jq} b_{jq}(x_j)$ como se ha visto en el caso del modelo aditivo (1.30), lo cual se puede expresar como

$$f_j = \tilde{\mathbf{X}}_j \tilde{\beta}_j,$$

siendo $\tilde{\mathbf{X}}_j = \{b_{jq}\}_{q=1}^{q_j} = [b_{j1}, \ldots, b_{jq_j}]$, con $b_{ji} = (b_{j1}(x_1), \ldots, b_{jn}(x_n))^t$, la matriz de elementos de la base y $\tilde{\beta}_j = [\beta_{j1}, \ldots, \beta_{jq_j}]^t$ el vector de coeficientes de la representación.

Puesto que el modelo no está identificado a menos que las funciones suaves estén sujetas a la restricción de estar centradas, sino estas estarían identificadas salvo la suma de una constante, es decir no se puede diferenciar entre $f_1(x_1)$ y $f_1(x_1) - c$, se impone la condición de que

$$1^t \tilde{\mathbf{X}}_j \tilde{\beta}_j = 0 \tag{1.83}$$

para garantizar que la media se anule. La condición (1.83) puede ser reescrita a través de una reparametrización que garantiza que existe una matriz $\mathbf{Z}$, con $q_j - 1$ columnas ortogonales, que satisface

$$1^t \tilde{\mathbf{X}}_j \mathbf{Z} = 0. \tag{1.84}$$

Ahora, reparametrizando los términos de suavizado en $q_j - 1$ parámetros nuevos, β_j, tal que $\tilde{\beta}_j = \mathbf{Z}\beta_j$, se obtiene una nueva matriz modelo para el término j^{th}, $\mathbf{X}_j = \tilde{\mathbf{X}}_j \mathbf{Z}$, tales que $f_j = \mathbf{X}_j \beta_j$ definidas a través de esta nueva matriz automáticamente cumplen la condición de estar centradas.

Considerando estos resultados se puede expresar el modelo inicial (1.82) como un modelo lineal general con estructura

$$g(\mu_i) = \mathbf{X}_i \beta \tag{1.85}$$

siendo $\mathbf{X} = [\mathbf{X}^*, \mathbf{X}_1, \mathbf{X}_2, \ldots] \in \mathcal{M}_{n \times \sum_{j=0}^{p} q_j}$ y $\beta^t = [\theta^t, \beta_1^t, \beta_2^t, \ldots] \in \mathcal{M}_{1 \times \sum_{j=0}^{p} q_j}$ tales que

$$S = \left(\overbrace{\mathbf{X}^*}^{q_0} \quad \overbrace{\mathbf{X}_1}^{q_1} \quad \overbrace{\mathbf{X}_2}^{q_2} \quad \cdots \quad \overbrace{\mathbf{X}_p}^{q_p} \right) \quad \text{y} \quad \beta = \left(\overbrace{\theta^t}^{q_0} \quad \overbrace{\beta_1^t}^{q_1} \quad \overbrace{\beta_2^t}^{q_2} \quad \cdots \quad \overbrace{\beta_p^t}^{q_p} \right)^t.$$

Teniendo en cuenta la expresión obtenida en (1.85) se ve que ahora los parámetros a predecir son los de un modelo lineal general GLM y por tanto estos se pueden obtener a través de la maximización de la verosimilitud. Si q_j es suficientemente grande y por tanto se tiene la posibilidad razonable de representar de forma precisa las funciones f_j's desconocidas y β es estimado por la maximización de la verosimilitud ordinaria, entonces hay una alta probabilidad de que se produzca sobreajuste. Para solucionar este problema a la hora de obtener las estimaciones de los parámetros del modelo GAM se recurre a la **maximización de la verosimilitud penalizada**, la cual se encarga de penalizar la curvatura excesiva de las estimaciones de los términos f_j.

Empleando la forma cuadrática $\tilde{\beta}^t \tilde{\mathbf{S}}_j \tilde{\beta}$ para penalizar la excesiva curvatura en la estimación de cada una de las $j = 1, \ldots, p$ funciones suaves, obtenida esta de igual modo que en el caso aditivo no generalizado (1.34). De nuevo $\tilde{\mathbf{S}}_j$ es una matriz de coeficientes conocidos expresada en términos de una base funcional. La reparametrización de centrado aplicada anteriormente sobre las funciones convierte esta penalización en $\beta_j^t \bar{\mathbf{S}}_j \beta_j$, donde ahora $\bar{\mathbf{S}}_j = \mathbf{Z}^t \tilde{\mathbf{S}}_j \mathbf{Z}$ siendo $\mathbf{Z}$ la matriz de (1.84). Además, es conveniente reescribir esta expresión en función de todos los términos de β, la cual pasaría a ser $\beta^t \mathbf{S}_j \beta$, donde $\mathbf{S}_j \in \mathcal{M}_{\sum_{j=0}^{p} q_j \times \sum_{j=0}^{p} q_j}$ es simplemente $\bar{\mathbf{S}}_j \in \mathcal{M}_{q_j \times q_j}$ rellenada con ceros cumpliendo que $\beta^t \mathbf{S}_j \beta \equiv \beta_j^t \bar{\mathbf{S}}_j \beta_j$. Teniendo esto en cuenta se obtiene que la verosimilitud penalizada para este modelo, conocidas las penalizaciones λ_j, es

$$l_p(\beta) = l(\beta) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \beta^t \mathbf{S}_j \beta,$$

la cual puede reescribirse de forma vectorial como

$$l_p(\beta) = l(\beta) - \frac{1}{2}\beta^t \mathbf{S}\beta \tag{1.86}$$

con $\beta^t = [\theta^t, \beta_1^t, \ldots, \beta_p^t] \in \mathcal{M}_{1\times\sum_{j=0}^p q_j}$ y $\mathbf{S} = \sum_{j=1}^p \lambda_j \mathbf{S}_j \in \mathcal{M}_{\sum_{j=0}^p q_j \times \sum_{j=0}^p q_j}$.

Derivando la expresión (1.86) respecto de β e igualando a cero para obtener el valor que maximiza la verosimilitud se obtiene que

$$\frac{\partial l_p}{\partial \beta_j} = \frac{\partial l}{\partial \beta_j} - [\mathbf{S}\beta]_j = \frac{1}{\phi}\sum_{i=1}^n \frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} - [\mathbf{S}\beta]_j = 0,$$

donde $\mathbb{V}(y_i) = \phi V(\mu_i)$ y $[.]_j$ denota la j-ésima fila de un vector.

Puede verse que esta última expresión es equivalente a resolver el problema mediante mínimos cuadrados penalizados no lineales, siempre que los pesos $V(\mu_i)$ sean conocidos de antemano y fuesen independientes de β. Por tanto, teniendo esto en cuenta, se llega a que maximizar la última ecuación es igual a minimizar la expresión

$$P = \frac{1}{\phi}\sum_{i=1}^n \frac{(y_i - \mu_i)^2}{V(\mu_i)} + \beta^t \mathbf{S}\beta,$$

donde se asume $V(\mu_i)$ conocida.

Además puede verse que en el entorno de algún estimador del parámetro $\hat{\beta}^{[k]}$ (véase la sección section 2.1.2 de [20]) se tiene que

$$P \approx \|\sqrt{\mathbf{W}^{[k]}}(z^{[k]} - \mathbf{X}\beta)\|^2 + \beta^t \mathbf{S}\beta,$$

con $\mathbf{W}$ cierta matriz de pesos y z un vector basado en una transformación de la variable respuesta. Permitiendo desarrollar un algoritmo recursivo para hallar la solución de este problema de optimización.

Al igual que en el caso de los GLM's lo que se hace es aproximar la función de máxima verosimilitud a optimizar por un problema de mínimos cuadrados y se halla la solución de este último de modo iterativo. De esta forma para llevar a cabo el ajuste del modelo se emplea iterativamente el algoritmo **penalizado de mínimos cuadrados ponderados** (*P-IRLS: penalized iteratively re-weighted least squares*) hasta obtener convergencia.

Algoritmo 1.14 (P-IRLS)**.**

1. *Dada una estimación del parámetro $\beta^{[k]}$, y el vector correspondiente a la estimación de la media $\mu^{[k]}$, se calcula*

$$\omega_i = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2} \quad y \quad z_i = g(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + \mathbf{X}_i\hat{\beta}^{[k]},$$

donde $\mathbb{V}(Y_i) = V(\mu^{[k]})\phi$ y X_i es la i-ésima fila de la matriz $\mathbf{X}$.

2. *Obtener $\beta^{[k+1]}$ como el valor β tal que es solución de*

$$\min_\beta \|\sqrt{\mathbf{W}}(z - \mathbf{X}\beta)\|^2 + \sum_{j=1}^p \lambda_j \beta^t \mathbf{S}_j \beta, \tag{1.87}$$

donde $\mathbf{W}$ es la matriz diagonal con $W_{ii} = w_i$ y z el vector de la respuesta transformada obtenido en el paso anterior.

De esta forma, teniendo en cuenta la expresión a minimizar en (1.87), puede verse que el valor de $\beta^{[k+1]}$ viene dado por

$$\beta^{(k+1)} = (\mathbf{X}^t\mathbf{W}^{(k)}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^t\mathbf{W}^{(k)}z^{(k)}. \tag{1.88}$$

Al igual que ocurría en el caso de los modelos aditivos, en un contexto de alta dimensión donde se supone que el número de covariables es mayor que el tamaño muestral, $p > n$, se tiene que no existirá la inversa de $(\mathbf{X}^t\mathbf{W}^{(k)}\mathbf{X})$, puesto que esta matriz tiene autovalores nulos. En cambio, como los autovalores de la matriz $\mathbf{S}$, por definición, cumplen que son todos positivos se puede garantizar que la matriz $\mathbf{X}^t\mathbf{W}^{(k)}\mathbf{X} + \mathbf{S}$ sí es invertible y se tiene por tanto que la expresión (1.88) está bien definida a la hora de estimar el valor de $\beta^{(k+1)}$.

Hasta ahora, al igual que se hacía en el caso de los modelos aditivos, se ha mostrado un algoritmo recursivo (P-IRLS) que permite obtener una estimación del vector β del modelo una vez conocidos los valores λ_j de penalización. De este modo se hace necesario recurrir a procedimientos que permitan conocer los valores óptimos de las penalizaciones para poder conseguir un buen estimador de β. En este contexto se diferenciarán los casos donde el parámetro de escala ϕ sea conocido, **criterio UBRE**, y donde no se puede utilizar esta información, **validación cruzada generalizada**.

Criterios de estimación de los parámetros de penalización λ_j

Al igual que ocurre con los modelos aditivos, el problema de ajustar el modelo (1.82) pasa a ser el de estimar las penalizaciones λ_j. Así, en el caso de un modelo aditivo generalizado se distinguen principalmente dos algoritmos para estimar el vector de penalizaciones $\lambda = (\lambda_1, \ldots, \lambda_p)^t$ que depende de si el parámetro de escala de la familia exponencial, ϕ, es conocido, **criterio UBRE** (*UBRE: Un-Biased Risk Estimator*), o si en cambio no se dispone de esta información, **método de validación cruzada generalizada** (*GCV: generalized cross validation*).

De forma similar al caso de los modelos aditivos se puede ver que la función objetivo a minimizar de un modelo GAM para la obtención del parámetro β del modelo puede ser escrita en términos de la deviance del modelo como

$$\min_{\beta} \quad D(\beta) + \sum_{j=1}^{m} \lambda_j \beta^t \mathbf{S}_j \beta.$$

Así, una vez conocido λ esta función objetivo puede ser aproximada cuadráticamente por la expresión

$$\min_{\beta} \quad \|\sqrt{\mathbf{W}}(z - \mathbf{X}\beta)\|^2 + \beta^t\mathbf{S}\beta, \tag{1.89}$$

la cual captura de forma razonable la dependencia de la deviance penalizada en λ y β, en el entorno del λ actual, y la correspondiente minimización de los valores de β.

Usando los mismos argumentos que se emplean en el caso de los modelos aditivos ahora sobre la función objetivo (1.89) se obtienen los correspondientes criterios para la selección de los parámetros λ_j. A continuación se explican los desarrollos imprescindibles para obtener dichas medidas.

Criterio UBRE

En el caso de los modelos aditivos, si el parámetro de escala de la familia exponencial, ϕ, es conocido, el valor de $\lambda = (\lambda_1, \ldots, \lambda_p)^t$ puede ser tomado de forma que $\hat{\mu}$ sea lo más cercano posible a la verdadera media $\mu \equiv \mathbb{E}(Y)$. Una medida útil de esta proximidad es el *Error Cuadrático Medio Esperado* (*MSE: expected mean square error*) que se mide a través de

$$\mathbb{E}(M) = \mathbb{E}\left(\|\mu - \mathbf{X}\hat{\beta}\|^2/n\right) = \mathbb{E}(\|y - \mathbf{A}y\|^2)/n - \sigma^2 + 2 \cdot \mathrm{tr}(\mathbf{A})\sigma^2/n. \tag{1.90}$$

Por tanto parece razonable escoger como parámetros de suavización aquellos que minimicen una estimación del valor esperado del MSE, esto es minimizar el **Estimador de Riesgo Insesgado (*UBRE: Un-Biased Risk Estimator*)**,

$$V_u(\lambda) = \|y - \mathbf{A}y\|^2/n - \sigma^2 + 2 \cdot \mathrm{tr}(\mathbf{A})\sigma^2/n, \tag{1.91}$$

el cual depende del parámetro de suavizado a través del término $\mathbf{A}$, donde $\mathbf{A} = (\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{X}$.

Si el valor σ^2 es conocido entonces el estimar λ a través de V_u (1.91) funciona adecuadamente, en cambio si no se conoce el valor del parámetro y es necesario estimar este aparecen problemas. Por ejemplo, bajo el supuesto de gaussianidad, teniendo en cuenta que

$$\mathbb{E}(\|y - \mathbf{A}y\|^2) = \sigma^2(n - \mathrm{tr}(\mathbf{A}))$$

se tiene que el estimador del error del modelo es $\hat{\sigma}^2 = \frac{\|y - \mathbf{A}y\|^2}{n - \mathrm{tr}(\mathbf{A})}$. Si se añade esta información a (1.90) se tiene

$$M = \mathbb{E}\left(\|\mu - \mathbf{X}\hat{\beta}\|^2/n\right) = \frac{\mathrm{tr}(\mathbf{A})}{n}\sigma^2$$

y por consiguiente el estimador MSE será $\tilde{M} = \mathrm{tr}(\mathbf{A})\hat{\sigma}^2/n$. El problema es que este estimador $\tilde{M}$ no es un fundamento adecuado para la selección de los parámetros del modelo ya que depende del tamaño muestral. Esto repercute en que al aumentar el número de parámetros incluidos en el modelo, al estar dividiendo el tamaño muestral la expresión a minimizar, será necesario que se reduzca proporcionalmente $\hat{\sigma}^2$ para obtener que un modelo con menos parámetros es igual de bueno que el que considera más. Esto hará que el valor de $\tilde{M}$ sea siempre menor que si se considerasen más parámetros, de modo que un mayor número de elementos siempre producirá una mejora en el ajuste del modelo.

Realizando un razonamiento similar pero en vez de tener (1.90) como criterio a minimizar se tiene en cuenta que en el caso de un modelo aditivo generalizado (GAM) la expresión a optimizar viene dada en términos de la deviance, pudiéndose expresar esta por la forma cuadrática vista en (1.89), se obtiene que el estimador UBRE en el caso de un modelo GAM viene dado por la expresión

$$V_u^w = \frac{1}{n}\|\sqrt{\mathbf{W}}(z - \mathbf{X}\beta)\|^2 - \sigma^2 + \frac{2}{n}\mathrm{tr}(\mathbf{A})\sigma^2, \tag{1.92}$$

siendo un estimador local válido únicamente para el λ empleado en el cálculo de z y $\mathbf{W}$. Puede obtenerse un criterio global que vendría dado por la expresión

$$V_u^w = \frac{1}{n}D(\hat{\beta}) - \sigma^2 + \frac{2}{n}\mathrm{tr}(\mathbf{A})\sigma^2, \tag{1.93}$$

el cual es efectivamente una transformación lineal del criterio AIC (para más información ver sección 2.1.3 de [20]).

Criterio de validación cruzada generalizada (GCV)

Otra forma de obtener los valores óptimos del vector λ, cuando no se conoce el parámetro de escala, es recurrir a procedimientos como los de validación cruzada. En particular a la **validación cruzada generalizada (*GCV: generalized cross validation*)** puesto que la validación cruzada ordinaria es muy costosa computacionalmente además de sufrir de falta de invariancia.

De nuevo, en el caso de los modelos aditivos, se obtiene que el criterio de validación cruzada generalizada viene dado por

$$V_g = \frac{n\|y - \hat{\mu}\|^2}{[n - \mathrm{tr}(\mathbf{A})]^2}, \tag{1.94}$$

que no es más que una modificación del criterio de validación cruzada ordinaria a través de una rotación debidamente escogida aplicada a la matriz de influencia $\mathbf{A}$, la cual consigue que sus efectos A_{ii} sean lo más igualados posibles.

Desarrollando las mismas ideas pero aplicadas el caso de los modelos aditivos generalizados donde, como se recordó anteriormente, ahora la función a minimizar se aproxima por (1.89), se llega a que para este tipo de construcciones el criterio de validación cruzada generalizada se basa en

$$V_g^w = \frac{n\|\sqrt{\mathbf{W}}(z - \mathbf{X}\beta)\|^2}{[n - \mathrm{tr}(\mathbf{A})]^2}, \tag{1.95}$$

el cual es un criterio de índole local ya que se ha obtenido mediante una estimación local de la función objetivo. En el caso de querer un criterio global puede verse que en el caso de la validación cruzada generalizada este tiene la forma

$$V_g^w = \frac{nD(\hat{\beta})}{[n - \mathrm{tr}(\mathbf{A})]^2}. \tag{1.96}$$

De este modo, teniendo en cuenta los criterios (1.92), (1.93), (1.95) y (1.96) obtenidos se puede ver que en el caso de los modelos GAM hay dos estrategias numéricas posibles para la estimación de los parámetros de suavizado usando la minimización de V_u o V_g:

- $V_{u/g}$ puede ser minimizado directamente, lo que significa que el esquema P-IRLS debe ser iterado hasta convergencia para cada conjunto de prueba de los parámetros de suavizado. Este procedimiento es conocido como *outer iteration*.

- $V_{u/g}^w$ puede ser minimizado y los parámetros de suavizado seleccionados para cada modelo linear penalizado de trabajo de cada iteración P-IRLS. Este método se conoce como *performance iteration*.

Pese a que el *performance algorithm* es computacionalmente eficiente puede presentar problemas de convergencia, mientras que el *outer algorithm* es computacionalmente más costoso pero a cambio se tiene mayor estabilidad.

Al igual que en el caso de los modelos lineales, cuando $p > n$ surge el problema de ser capaces de determinar qué variables del conjunto son las más importantes a la hora de explicar la variable respuesta. Ante este problema se propondrán a continuación varias penalizaciones y métodos que solucionan este problema en el caso de los modelos GAM.

1.5.2. Regularización L_2

Recordando la expresión que se tenía en (1.87) se puede ver que es fácil imponer una penalización de tipo Ridge al igual que se explica en [19]. Una penalización de este estilo consigue que los coeficientes obtenidos tengan un valor más próximo a cero, facilitando la determinación de cuáles podrían ser aquellos más propensos a tener un coeficiente nulo y por lo tato a no formar parte del modelo.

Recordando, una vez obtenida la aproximación lineal del problema a resolver iterativamente, que la expresión a optimizar es (1.89) puede verse que es fácil imponer una estimación de tipo Ridge sin más que tomar la matriz fija semidefinida positiva de penalización $\mathbf{P} = \tilde{\lambda}\mathbf{I}$, $\tilde{\lambda} > 0$. De esta forma se tiene que la expresión genérica a optimizar de [19] en el caso de imponer una penalización $\mathbf{P}$ con estas características,

$$\min_{\beta} \quad \|\sqrt{\mathbf{W}}(z - \mathbf{X}\beta)\|^2 + \beta^t \mathbf{P}\beta + \beta^t \mathbf{S}\beta,$$

pasa a ser resolver el problema

$$\min_{\beta} \quad \|\sqrt{\mathbf{W}}(z - \mathbf{X}\beta)\|^2 + \tilde{\lambda}\beta^t\beta + \beta^t \mathbf{S}\beta. \tag{1.97}$$

El problema (1.97) puede ser resuelto de modo iterativo a través del algoritmo IRLS una vez que se han determinado los parámetros de suavizado recogidos en $\mathbf{S}$, $\lambda_j \quad j = 1, \ldots, p$ y el parámetro de penalización $\tilde{\lambda}$. Para la estimación del valor adecuado de estos últimos, al igual que se vio en el modelo GAM no penalizado, se pueden emplear criterios como el UBRE o el de validación cruzada, sin más que tener en cuenta que ahora en la matriz de influencia $\mathbf{A}$, la cual entra en juego, hay que incluir la penalización tipo Ridge impuesta. De esta forma la nueva matriz de influencia, denotada por $\tilde{\mathbf{A}}$ viene dada por la expresión

$$\tilde{\mathbf{A}} = (\mathbf{X}^t\mathbf{W}\mathbf{X} + \tilde{\lambda}\mathbf{I} + \mathbf{S})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{X}.$$

Basta substituir esta en los criterios citadas para obtener las estimaciones deseadas.

1.5.3. Regularización L_1: modelo SpAM generalizado

Al igual que en el caso de los modelos aditivos se puede imponer una penalización de tipo L_1 para conseguir determinar que variables son las más importantes a la hora de explicar e interpretar un modelo aditivo generalizado.

Al igual que se veía en la Sección 1.2.3 ahora se recurrirá a resolver un problema de lagrangiano aumentado, el cual vendrá dado a través de la expresión

$$\mathcal{L}(f, \lambda) = -l(f) + \lambda \left(\sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(x_j))} - L \right), \tag{1.98}$$

siendo $l(f)$ la correspondiente función de log-verosimilitud.

Basta imponer las condiciones de *Karush-Kuhn-Tucker* sobre (1.98) que vienen dadas por

$$\frac{\partial}{\partial f}\left[-l(f_j)\right] + \lambda v_j = 0 \quad \text{con } v_j \text{ un elemento del subgradiente de } \partial\sqrt{\mathbb{E}(f_j^2)}$$

y despejar en función de cada f_j para obtener una expresión que permita llegar, mediante iteración, a una solución de (1.98). En el proceso hay que tener en cuenta que puesto que esta condición no es lineal sobre las f_j es necesario linealizar el gradiente de la log-verosimilitud sobre la estimación actual.

A continuación se va a ejemplificar este procedimiento en el caso del **modelo logístico aditivo**. En este caso se tiene que la función que se desea modelar en torno a una estructura aditiva es

$$P(\,Y=1\mid X\,)\equiv p(X,f)=\frac{\exp\left(\sum_{j=1}^{p}f_j(X_j)\right)}{1+\exp\left(\sum_{j=1}^{p}f_j(X_j)\right)}$$

donde $Y\in\{0,1\}$ y la log-verosimilitud poblacional se corresponde ahora con la expresión

$$l(f)=\mathbb{E}[yf(\mathbf{X})-\log(1+\exp f(\mathbf{X}))]. \tag{1.99}$$

Ahora, recordando el algoritmo P-IRLS que se desenvolvía en el caso del modelo aditivo generalizado (algoritmo 1.14) se calculaba una aproximación lineal de la respuesta, la cual ahora se correspondería con

$$Z_i=f_0(X_i)+\frac{y_i-p(X_i,f_0)}{p(X_i,f_0)(1-p(X_i,f_0))} \tag{1.100}$$

donde f_0 representa el estimador actual y $p(X_i,f_0)(1-p(X_i,f_0))=w(X_i)$ son los pesos y se llevaba a cabo un procedimiento de backfitting ponderado de $(\mathbf{X},\mathbf{Z})$ con pesos $w(X_i)$. El suavizado de la ponderación viene ahora dado por

$$\hat{P}_j=\frac{S_j(wR_j)}{S_jw}.$$

Para incorporar la penalización de tipo *sparse* se llega a que la expresión para el lagrangiano (1.98) en el caso del modelo logístico es

$$\mathcal{L}(f,\lambda)=\mathbb{E}\left[\log(1+e^{f(\mathbf{X})}-yf(\mathbf{X}))\right]+\lambda\left(\sum_{j=1}^{p}\sqrt{\mathbb{E}(f_j^2(x_j))}-L\right),$$

y la condición estacionaria de KKT para cada f_j es que se cumpla que $\mathbb{E}(\,p-y\mid X_j\,)+\lambda v_j=0$, con v_j una componente del subgradiente de $\partial\sqrt{\mathbb{E}(f_j^2)}$. Dado que, como se explicó anteriormente, esta imposición no es lineal en base al término f_j se aproxima el gradiente de la log-verosimilitud por una aproximación lineal en torno al estimador actual f_0. De esta forma, teniendo en cuenta la expresión (1.100) obtenida, se llega a que la condición a verificar es $\mathbb{E}[w(\mathbf{X})(f(\mathbf{X})-\mathbf{Z})|X_j]+\lambda v_j=0$. En conclusión, se tiene que cuando $\mathbb{E}(f_j^2)\neq 0$, esta condición implica que

$$\left(\mathbb{E}(w|X_j)+\frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right)f_j(X_j)=\mathbb{E}(wR_j|X_j).$$

En el caso de ajustar el modelo para una muestra finita, en base a los términos suavizadores escogidos S_j, se tiene la expresión

$$f_j=\frac{S_j(wR_j)}{S_jw+\lambda\sqrt{\mathbb{E}(f_j^2)}}. \tag{1.101}$$

Si $\|S_j(wR_j)\|<\lambda$, entonces $f_j=0$. En caso contrario será necesario calcular el valor de dicha función. Puesto que la relación proporcionada en (1.101) no es lineal en términos de f_j no se podrá despejar este factor y será necesario recurrir a procedimientos iterativos para obtener su valor. Se propone el esquema iterativo basado en calcular, hasta obtener convergencia, las funciones f_j mediante

$$f_j\leftarrow\frac{S_j(wR_j)}{S_jw+\lambda/\sqrt{\mathbb{E}(f_j^2)}}.$$

Capítulo 2

Clasificación en alta dimensión

Las reglas de clasificación proporcionan un algoritmo que permite determinar en qué grupo clasificar una nueva observación dentro de las posibles L clases consideradas, teniendo únicamente en cuenta sus características o el valor de sus parámetros. De esta forma se tiene un procedimiento que permite extraer conclusiones en base a la discriminación realizada, como podría ser conocer si un paciente está sano o enfermo en base a su condición fisiológica, determinar qué tipo de prevalencia tiene una persona a sufrir una dolencia o enfermedad según su genética y antecedentes, o establecer en qué grupo de riesgo se encuentra un futuro cliente que quiere contratar un seguro. Para lograr este fin el análisis discriminante, suponiendo que los grupos de estudios están totalmente determinados de antemano, se encarga de buscar las coincidencias y discrepancias entre dichas clases para conseguir dictaminar criterios que permitan entender qué hace único cada grupo y qué tipos de datos se asocian con estos. De igual importancia es determinar y conocer como será la estructura y forma de las fronteras de decisión, lo cual permitirá entender cómo de lejos está un dato de una clase particular de los restantes.

Ante un escenario de alta dimensión donde se tiene un número de covariables muy próximo o mayor al tamaño muestral, $p > n$, las reglas estimadas de clasificación no poseen un buen comportamiento. Esto es debido a que tienen que afrontar problemas como el mal condicionamiento o el desastre de la dimensionalidad, distorsionando los resultados obtenidos y haciendo dudosa su eficiencia. De nuevo será necesario recurrir a regularizaciones o modificaciones de estas para poder llevar a cabo la clasificación.

En lo que respecta a este capítulo se introducirán diversos métodos de clasificación que dan lugar a distintas reglas que pueden ser empleadas en la práctica ante la necesidad de clasificar nuevos datos. Se expondrán algoritmos que hacen suposiciones bien distribucionales o de la matriz de covarianzas así como otros que no necesitan de estas y proporcionan otros enfoques diferentes. En estos se tendrá en cuenta la posibilidad de que las fronteras de decisión tenga forma lineal u otras estructuras no necesariamente lineales, proporcionando flexibilidad en los ajustes. Se mostrará la filosofía de cada una de estas reglas, explicando como realizar su estimación y comentando las ventajas e inconvenientes que poseen comúnmente, haciendo especial hincapié en el caso de alta dimensión considerado ($p > n$). Finalmente se propondrán soluciones o alternativas para la estimación de las reglas de clasificación en este contexto, permitiendo obtener clasificadores que se puedan utilizar cuando el número de covariables sea mayor que el de muestras.

2.1. Análisis Discriminante

El científico *Ronald Aylmer Fisher* fue el inventor del **análisis discriminante lineal** en 1936. Este campo ayuda a identificar las características que diferencian (discriminan) a dos o más grupos. Es una técnica estadística capaz de decidir qué variables permiten diferenciar a los grupos y cuántas de estas son necesarias para alcanzar la mejor clasificación posible. La finalidad es crear una función discriminante capaz de decidir a que grupo pertenece un individuo tras analizar las

determinadas variables que este proporciona. Para crear dicha función se recogen muestras de cada conjunto y se analiza como son sus características, lo que tienen en común dentro de un mismo grupo y en que difieren entre los restantes. Con esta información se extraen patrones que permiten catalogar.

De esta forma el análisis discriminante surge ante la necesidad de clasificar a un nuevo individuo en base a p características observadas de este en un grupo, teniendo que escoger entre L grupos distintos. Con este fin se crean las **reglas discriminantes** las cuales son funciones que indican a qué clase o grupo pertenece la nueva observación teniendo en cuenta sus características.

2.1.1. Reglas discriminantes

Vamos a situarnos en el contexto donde partimos de L grupos $G_1, \ldots, G_L$. Se tiene por tanto que ahora los datos muestrales serán de la forma $\{(X_i, y_i)\}_{i=1}^n \in \mathbb{R}^p \times \{1, \ldots, L\}$, siendo y_i un valor discreto que indica la pertenencia de cada dato X_i a uno de los G_l grupos, $l = 1, \ldots, L$.

Una **regla discriminante aleatorizada** es una aplicación:

$$\begin{aligned} &\varphi : R \longrightarrow [0,1] \\ &\quad x \to \varphi_l(x) = \mathbb{P}(\, y \in G_l \mid X = x\,) \quad l \in \{1, \ldots, L\} \end{aligned} \tag{2.1}$$

siendo $R \subset \mathbb{R}^p$ el soporte donde toma valores el vector de variables x.

Para cada individuo del grupo l-ésimo se pueden cometer $L-1$ errores al clasificarlo en un grupo de los restantes. Las probabilidades condicionadas de cada uno de estos errores para la regla (2.1) son

$$\mathbb{P}(\, j \mid l\,) = \int \varphi_j f_l(x) dx \quad \forall j \in \{1, \ldots, L\} \text{ con } j \neq l,$$

siendo f_l la función de densidad para un determinado individuo si procede del grupo l-ésimo.

Una **regla discriminante no aleatorizada** es de la forma

$$\varphi_l(x) = \begin{cases} 1 & \text{si } x \in R_l, \\ 0 & \text{en otro caso.} \end{cases} \tag{2.2}$$

Donde en (2.2) R_l es la región del soporte donde se encuentran los individuos del grupo l y por tanto $\varphi_l(x) = 1$ cuando x pertenezca a dicho grupo. Esta regla es determinista.

En este caso, las probabilidades condicionadas de los errores serían

$$\mathbb{P}(\, j \mid l\,) = \int_{R_j} f_l(x) dx.$$

Se dice que una regla discriminante φ es *preferible* a otra φ^t si se cumple que para todo $l \in \{1, \ldots, L\}$

$$\mathbb{P}_{\varphi}(\, j \mid l\,) \leq \mathbb{P}_{\varphi^t}(\, j \mid l\,) \quad \forall j \in \{1, \ldots, L\} \text{ con } j \neq l$$

y se dirá *estrictamente preferible* si es preferible y además existe un $l \in \{1, \ldots, L\}$ tal que

$$\mathbb{P}_{\varphi}(\, j \mid l\,) < \mathbb{P}_{\varphi^t}(\, j \mid l\,) \quad \text{para algún } j \in \{1, \ldots, L\} \text{ con } j \neq l.$$

Una regla discriminante se dirá *admisible* cuando no haya otra regla que sea preferible estrictamente a ella.

2.1.2. Criterios de elección de la regla discriminante

A la hora de establecer el criterio de elección de una regla discriminante puede obtarse por varios criterios como la **razón de verosimilitudes**, que asigna el individuo al grupo más verosímil, o la **minimización del coste total de clasificación incorrecta**, donde clasificar mal un dato supone un coste a tener en cuenta y se busca que el coste total esperado sea lo menor posible. También hay otros criterios conocidos como la **minimización de la probabilidad total de clasificación incorrecta**, donde se quiere disminuir la probabilidad de que un individuo sea mal clasificado, o el **criterio minimax** que se sitúa en el peor caso pretendiendo minimizar la probabilidad total de clasificación incorrecta.

En este trabajo se van a desarrollar las reglas de clasificación en base a **maximizar la probabilidad a posteriori**, donde se asignará la observación futura al grupo con mayor probabilidad. Suponiendo que $f_l(x)$ es la densidad o masa de probabilidad condicionada de X a la clase $G = l$ (se denotará así ocasionalmente la clase l-ésima: G_l) y π_l la probabilidad de que X pertenezca a dicha clase ($\sum_{j=1}^{L} \pi_j = 1$), por el Teorema de Bayes se sabe que

$$\mathbb{P}(\, G = l \mid X = x\,) = \frac{f_l(x)\pi_l}{\sum_{j=1}^{L} f_j(x)\pi_j}, \tag{2.3}$$

de esta forma se clasificará x en la clase l si se cumple que

$$\mathbb{P}(\, G = l \mid X = x\,) = \max_{l} \mathbb{P}(\, G = l \mid X = x\,) = \max_{l} \frac{f_l(x)\pi_l}{\sum_{j=1}^{L} f_j(x)\pi_j} \quad \forall l \in \{1, \ldots, L\}.$$

2.1.3. Análisis Lineal Discriminante (LDA): Regla lineal de Fisher. Problemas en alta dimensión

El **análisis lineal discriminante** o *LDA (Linear Discriminant Analysis)* se basa en que la variable poblacional sigue una distribución normal o Gaussiana. Tanto en el caso simple como en el multivariante distintas distribuciones se pueden expresar a través de una distribución normal aplicando una transformación, pudiendo emplear el análisis lineal discriminante con estos datos transformados. Una forma de conseguir este objetivo es a través de las transformaciones de la familia Box-Cox como pueden ser la transformación logarítmica o la raíz cuadrada (véase capítulo 10, página 305 de [13]), las cuales permiten conseguir normalidad en variables escalares. Por tanto este método permite barrer un amplio campo de contextos.

Bajo este supuesto la densidad en cada clase seguirá una normal multivariante ($N_p(\mu_l, \mathbf{\Sigma}_l)$) de forma que la función de densidad de la clase l-ésima es

$$f_l(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_l|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_l)^t {\mathbf{\Sigma}_l}^{-1}(x-\mu_l)\right\}.$$

El análisis lineal discriminante típicamente supone que la estructura de covarianzas es la misma para todas las clases, es decir $\mathbf{\Sigma}_l = \mathbf{\Sigma} \quad \forall l \in \{1,\ldots,L\}$, denotando por $\mathbf{\Sigma}_l$ la matriz de covarianzas de la clase $l \in \{1,\ldots,L\}$ y siendo $\mathbf{\Sigma}$ una matriz simétrica y semidefinida positiva. Esta hipótesis puede comprobarse en la práctica mediante un contraste de hipótesis sobre la igualdad de las matrices de covarianzas entre clases, usando el test de razón de verosimilitudes como puede verse en el Capítulo 9 de [16], siendo aceptable cuando no haya evidencias suficientes para rechazar la hipótesis nula. A la hora de comparar dos clases l y m se toman logaritmos para que quede una expresión más sencilla (esto es posible dado que la transformación logarítmica es creciente), de esta forma la comparación se basaría en

$$\begin{aligned} \log\frac{\mathbb{P}(\,G=l \mid X=x\,)}{\mathbb{P}(\,G=m \mid X=x\,)} &= \log\frac{f_l(x)}{f_m(x)} + \log\frac{\pi_l}{\pi_m} \\ &\overset{(a)}{=} \log\frac{\pi_l}{\pi_m} - \frac{1}{2}(\mu_l+\mu_m)^t\mathbf{\Sigma}^{-1}(\mu_l-\mu_m) + x^t\mathbf{\Sigma}^{-1}(\mu_l-\mu_m). \end{aligned} \tag{2.4}$$

La ecuación (2.4) es lineal en x, lo que da el nombre al método. La suposición de igualdad de matriz de covarianzas en las distintas clases hace que en (a) se cancelen los factores normalizados además de las partes cuadráticas de los exponentes de la función de densidad.

La frontera entre dos clases viene dada por $\mathbb{P}(\,G=l \mid X=x\,) = \mathbb{P}(\,G=m \mid X=x\,)$. En esta situación las fronteras entre las distintas clases serán hiperplanos en el espacio p-dimensional de los parámetros, de forma que el espacio $\mathbb{R}^p$ quedará dividido en L regiones bien diferenciadas y separadas por hiperplanos.

En el caso particular donde sólo existen dos posibles clases, sobrescribiendo (2.4) se obtiene

$$\begin{aligned} \log\frac{\mathbb{P}(\,G=1 \mid X=x\,)}{\mathbb{P}(\,G=2 \mid X=x\,)} &= \log\frac{f_1(x)}{f_2(x)} + \log\frac{\pi_1}{\pi_2} \\ &= \log\frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1+\mu_2)^t\mathbf{\Sigma}^{-1}(\mu_1-\mu_2) + x^t\mathbf{\Sigma}^{-1}(\mu_1-\mu_2), \end{aligned} \tag{2.5}$$

de forma que x será clasificado en el primer grupo cuando su probabilidad sea mayor, es decir, cuando $\log\frac{\mathbb{P}(\,G=1|X=x\,)}{\mathbb{P}(\,G=2|X=x\,)} > 0$ y esta condición es equivalente a que

$$\lambda^t\left[x - \frac{1}{2}(\mu_1+\mu_2)\right] > \log\left(\frac{\pi_2}{\pi_1}\right) \quad \text{siendo } \lambda = \mathbf{\Sigma}^{-1}(\mu_1-\mu_2), \tag{2.6}$$

esta expresión (2.6), es conocida como la **regla lineal discriminante de Fisher**.

Si además las probabilidades de pertenecer a cada clase son iguales ($\pi_1 = \pi_2$), se tendría la condición:

$$\lambda^t\left[x - \frac{1}{2}(\mu_1+\mu_2)\right] > 0 \Rightarrow \lambda^t x > \frac{1}{2}(\lambda^t\mu_1 + \lambda^t\mu_2) \Rightarrow x - \frac{1}{2}\mu_1 > \frac{1}{2}\mu_2.$$

Es fácil ver que si $\mathbf{\Sigma}_l = \mathbf{\Sigma} \quad \forall l \in \{1,\ldots,L\}$ y $\pi_l = \pi_2$, teniendo en cuenta (2.5), la condición $\log\frac{\mathbb{P}(G=1|X=x)}{\mathbb{P}(G=2|X=x)} > 0$ equivale a que

$$\frac{1}{2}\underbrace{(x-\mu_2)^t\mathbf{\Sigma}^{-1}(x-\mu_2)}_{D_2^2} > \frac{1}{2}\underbrace{(x-\mu_1)^t\mathbf{\Sigma}^{-1}(x-\mu_1)}_{D_1^2}, \tag{2.7}$$

donde D_1^2 y D_2^2 son las distancias de Mahalanobis del vector x a los centroides de cada clase, μ_1 y μ_2 respectivamente. De esta forma el criterio (2.7) dice que cuando se puede suponer que ambas clases comparten la misma estructura de covarianzas y tienen la misma probabilidad a priori, se clasificará una nueva muestra en el primer grupo si $D_2^2 > D_1^2$ lo cual se corresponde con que el nuevo dato esté más próximo del centroide del primer grupo que del segundo en términos de la distancia de Mahalanobis.

Se define la **función de discriminación lineal** para cada clase como

$$\delta_l(x) = x^t \mathbf{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^t \mathbf{\Sigma}^{-1} \mu_l + \log \pi_l. \tag{2.8}$$

De esta forma se clasificará x en aquella clase con mayor probabilidad, es decir, en la clase $G(x) = \max_l \delta_l(x)$. En la práctica no se van a conocer los parámetros de la distribución normal (vector de medias, matriz de covarianzas y probabilidades de cada clase) con lo que es necesario estimarlos mediante los datos muestrales

$$\begin{aligned}
&\hat{\pi}_l = \frac{n_l}{n} \quad \text{donde } n_l \text{ es el número de observaciones de la clase } l = 1, \dots, L, \\
&\hat{\mu}_l = \sum_{g_j = l} \frac{x_j}{n_l} \quad \text{siendo } g_j \text{ una variable dummy que indica la clase a la que pertenece cada dato,} \\
&\hat{\mathbf{\Sigma}} = \sum_{l=1}^{L} \sum_{g_j = l} \frac{(x_j - \hat{\mu}_l)(x_j - \hat{\mu}_l)^t}{(n - L)}.
\end{aligned} \tag{2.9}$$

En la situación de alta dimensión de interés, es decir, cuando $p > n$ va a surgir el problema de que no es posible construir la regla discriminante, dado que al estimar la matriz de covarianzas $\mathbf{\Sigma}$ esta no va a ser invertible y por tanto no se puede definir la función de discriminación lineal (2.8).

Para entender qué problemas surgen en este contexto referentes a la invertibilidad de $\hat{\mathbf{\Sigma}}$ se puede ver que es posible descomponer esta matriz en un producto matricial de la forma $\tilde{\mathbf{X}}^t \tilde{\mathbf{X}}/c$, donde $\tilde{\mathbf{X}} \in \mathcal{M}_{n \times p}$ y $c \in \mathbb{R}$, apareciendo de nuevo el problema de que rango$(\tilde{\mathbf{X}}) \leq n < p$ que fuerza a que $\hat{\mathbf{\Sigma}} \in \mathcal{M}_{p \times p}$ tenga determinante nulo. Teniendo en cuenta la estructura de covarianzas dada en (2.9) se puede ver que tomando

$$\tilde{\mathbf{X}} = \left(x_{11} - \hat{\mu}_1 \,; \quad \cdots \quad ; x_{1n_1} - \hat{\mu}_1 \,; \quad \cdots \quad ; x_{L1} - \hat{\mu}_L \,; \quad \cdots \quad ; x_{Ln_k} - \hat{\mu}_L \right)^t$$

se cumple la igualdad $\tilde{\mathbf{X}}^t \tilde{\mathbf{X}}/c$. Donde $\hat{\mu}_l$, $l = 1, \dots, L$, es la estimación del centroide de cada clase y x_{ln_l} es cada uno de los vectores muestrales referentes a la clase l-ésima.

Además, puesto que toda matriz de covarianzas es simétrica y semidefinida positiva trivialmente por construcción, puede verse que $\mathbf{\Sigma}$ admite una descomposición espectral de la forma

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^t \Rightarrow \mathbf{\Sigma}^{-1} = (\mathbf{U}\mathbf{D}\mathbf{U}^t)^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^t,$$

con $\mathbf{U}$ una matriz ortogonal y $\mathbf{D}$ una matriz diagonal, siendo ambas matrices cuadradas de dimensión $p \times p$. Observando la matriz $\mathbf{D}$ se ve que es necesario estimar p elementos, pero al disponer de una muestra de tamaño n y darse que $p > n$, al igual que ocurría en regresión, no se dispone de suficiente información para estimar dichos valores y hacer que la matriz tenga rango máximo. Por tanto no se podrá invertir $\mathbf{D}$ de forma única, ya que posee autovalores nulos, con lo que $\mathbf{\Sigma}$ será singular a efectos prácticos, habiendo direcciones en las cuales la covarianza se dispara. Estas direcciones se corresponden con los autovectores asociados a los autovalores nulos de $\mathbf{D}$.

2.1.4. Regla discriminante cuadrática (QDA). Problemas en alta dimensión

Si volviendo al punto de partida, bajo las mismas hipótesis, ahora no se tiene garantizado que $\mathbf{\Sigma}_l = \mathbf{\Sigma} \quad \forall l \in \{1, \ldots, L\}$, entonces la expresión en (2.4) cambia pues ya que no se producen las oportunas cancelaciones en los términos, obteniendo por consiguiente

$$\begin{aligned} \log \frac{\mathbb{P}(\, G = l \mid X = x\,)}{\mathbb{P}(\, G = m \mid X = x\,)} &= \log \frac{f_l(x)}{f_m(x)} + \log \frac{\pi_l}{\pi_m} \\ = \log \frac{\pi_l}{\pi_l} + \frac{1}{2} \log \left(\frac{|\mathbf{\Sigma}_m|}{|\mathbf{\Sigma}_l|} \right) &- \frac{1}{2} \left[(x - \mu_l)^t \mathbf{\Sigma}_l^{-1} (x - \mu_l) - (x - \mu_l)^t \mathbf{\Sigma}_m^{-1} (x - \mu_l) \right]. \end{aligned} \tag{2.10}$$

Ahora esta ecuación es cuadrática respecto a la variable x. La **función de discriminación cuadrática** para cada clase adoptaría la forma

$$\delta_l(x) = -\frac{1}{2} \log(|\mathbf{\Sigma}_l|) - \frac{1}{2} (x - \mu_l)^t \mathbf{\Sigma}_l^{-1} (x - \mu_l) + \log(\pi_l), \tag{2.11}$$

donde x se clasificaría en el grupo $G(x) = \text{máx}_l \, \delta_l(x)$.

Interpretando dicha función se ve que la frontera entre dos clases deja de ser lineal y pasa a ser una hipersuperficie, la cual está dada por la ecuación cuadrática que satisface el conjunto de puntos $\{x : \delta_l(x) = \delta_m(x)\} \quad \forall l, m \in \{1, \ldots, L\}$ con $l \neq l$.

Los estimadores para la regla de discriminación cuadrática son los mismos que los vistos en (2.9), excepto que ahora la matriz de covarianzas se estimada por separado para cada clase. Al igual que ocurría antes, cuando hay mayor número de variables que de muestras ($p > n$), las matrices de covarianzas no van a ser invertibles y por lo tanto no será posible estimar la regla de clasificación de nuevo.

En la Figura 2.1 se ilustra en dos dimensiones las diferentes fronteras de decisión en el caso de dos clases correspondientes a la Regla Lineal de Fisher y a la Regla Discriminante Cuadrática. En la Regla de Fisher se observa que al suponerse $\mathbf{\Sigma}_l = \mathbf{\Sigma}$ la frontera de decisión es lineal, mientras que en la regla cuadrática al no poder hacerse esta suposición la frontera de decisión pasa a ser la solución de una función cuadrática.

2.1.5. Estimación de las reglas discriminantes

En la práctica normalmente se van a desconocer las distribuciones de X condicionales a cada uno de los grupos. En su lugar se va a disponer de muestras procedentes de cada grupo de la forma:

$$\begin{gathered} X_{11}, \ldots, X_{1n_1} \text{ del grupo } G_1 \\ \vdots \\ X_{L1}, \ldots, X_{Ln_L} \text{ del grupo } G_L \end{gathered}$$

conocidas como *muestras de entrenamiento*, las cuales sirven para estimar la regla discriminante a través de estimaciones de sus respectivas distribuciones.

Siendo

$$\hat{R}_l = \left\{ x : \frac{\hat{f}_l(x)}{\sum_{\substack{m=1 \\ m \neq l}}^{L} \hat{f}_m(x)} > c \right\} \quad \forall l \in \{1, \ldots, L\},$$

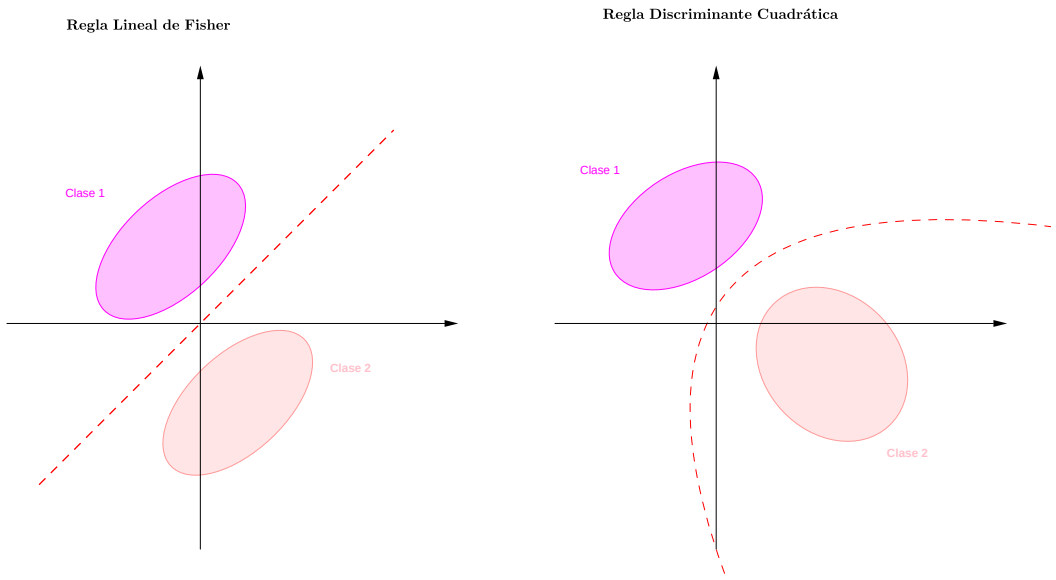

Figura 2.1: Gráfica de las fronteras de clasificación de las distintas reglas discriminantes. Las regiones rosas marcan la zona donde se encuentran los datos de cada grupo, son delimitadas por la matriz de covarianzas de cada clase. Las líneas discontinuas rojas son las respectivas fronteras de decisión de cada regla.

donde $\hat{f}_l(x) \quad l \in \{1, \ldots, L\}$ es una estimación oportuna de la función de densidad del grupo l, dependiendo de si esta sigue un modelo paramétrico o si en cambio es no paramétrico.

La regla discriminante estimada con $\hat{R} = (\hat{R}_1, \ldots, \hat{R}_L)$ se aplicaría más adelante para clasificar a un nuevo individuo en uno de los L-grupos en base a su resultado en el vector X.

En la sección anterior 2.1.1 se vio que la regla discriminante óptima se obtenía conociendo $R = (R_1, \ldots, R_L)$, donde las regiones de clasificación $R_l \quad l \in \{1, \ldots, L\}$ dependían de las verdaderas funciones de distribución $f_l \quad l \in \{1, \ldots, L\}$. Como las funciones de densidad son desconocidas hay que estimarlas, de forma que la regla deja de ser óptima. Para comprobar si la regla estimada sigue siendo similar a la óptima se definen las **tasas de error en reglas estimadas**.

Tasas de error en reglas estimadas

Es interesante conocer cómo afecta la estimación de la regla discriminante a los errores de clasificación, con esta finalidad se distingue entre **Tasas de error óptimas** y **Tasas de error efectivas**.

Las **Tasas de error óptimas** son las probabilidades de error de la regla discriminante óptima. Se denotan de la forma

$$e_{m,opt} = \int_{\bigcup_{\substack{l=1 \\ l \neq m}}^{L} R_l} f_m(x)dx \quad m \in \{1, \ldots, L\},$$

siendo

$$e_{opt} = \sum_{l=1}^{L} \pi_l e_{l,opt}.$$

En cambio, las **Tasas de error efectivas** son las probabilidades de error de la regla discriminante estimada:

$$e_{m,efe} = \int_{\bigcup_{\substack{l=1 \\ l\neq m}}^{L} \hat{R}_l} f_m(x)dx \quad m \in \{1,\ldots,L\},$$

con

$$e_{efe} = \sum_{l=1}^{L} \pi_l e_{l,efe}.$$

Nótese que las regiones $\hat{R}_l$ se obtienen de las muestras de entrenamiento. En la medida en que estas muestras son aleatorias, las regiones de clasificación serán aleatorias, y también lo serán las tasas de error efectivas. Por tanto, tiene sentido calcular su esperanza. La **esperanza de las tasas de error efectivas** no es más que la esperanza de las tasas de error anteriores de la forma

$$\mathbb{E}(e_{efe}) = \sum_{l=1}^{L} \pi_l \mathbb{E}(e_{l,efe}).$$

Es inmediato ver que

$$e_{l,opt} \leq e_{l,efe} \quad \text{y} \quad e_{opt} \leq e_{efe}.$$

Las tasas de error óptimas se refieren a la situación ideal donde las distribuciones de cada grupo son conocidas, como esto en general no ocurre, lo que interesa es calcular las tasas de error efectivas ya que estas miden el error al clasificar un nuevo individuo mediante la regla estimada.

Las tasas de error efectivas son desconocidas ya que dependen de las funciones de densidad f_l, de forma que es necesario plantear estimadores de las mismas.

Estimación de las tasas de error efectivas

Para conseguir estimar las tasas de error efectivas se van a tratar varios métodos:

- **Estimadores plug-in** Se obtienen sustituyendo f_l por $\hat{f}_l$

$$e_{m,pl} = \int_{\bigcup_{\substack{l=1 \\ l\neq m}}^{L} \hat{R}_l} \hat{f}_m(x)dx \quad m \in \{1,\ldots,L\}.$$

- **Tasas de error aparentes** Las tasas de error aparentes son las frecuencias relativas de observaciones mal clasificadas

$$e_{m,apa} = \frac{m_l}{n_l} \quad l \in \{1,\ldots,L\},$$

siendo m_l el número de individuos de la muestra de entrenamiento procedente del grupo G_l que son mal clasificados por la regla estimada.

Las tasas de error aparentes tienden a subestimar el error real, es decir infraestiman las tasas de error efectivas, ya que se aplica la regla estimada a los mismos datos con los que ha sido construida. El mismo fenómeno ocurre con los estimadores plug-in, ya que entre la regla estimada y $\hat{f}_m$ hay una relación de optimalidad que no se cumple con las tasas de error efectivas. Por este motivo se proponen correcciones de las tasas de error aparentes como la

corrección por validación cruzada. Consiste en construir la regla discriminante con todas las observaciones menos una, y aplicar después esa regla para clasificar al individuo omitido. Las tasas de error se calcularían después como las frecuencias relativas de individuos mal clasificados.

2.2. Análisis discriminante regularizado

En un contexto de alta dimensión donde se tienen más covariables que datos, $p > n$, independientemente del número de clases que entren en juego, $L \geq 2$, se ha visto que no es posible estimar las reglas de clasificación tanto en el caso lineal (LDA) como en el caso cuadrático (QDA), Secciones 2.1.3 y 2.1.4.

Este problema surge ante la falta de información necesaria para poder calcular la inversa de las matrices de covarianzas, teniendo que estimarse únicamente una en el caso lineal, donde se asume que esta estructura es idéntica para cada grupo, y L en el caso cuadrático donde no se considera esta hipótesis. Por lo tanto una forma de solucionar este problema es reemplazar las correspondientes matrices de covarianzas por una aproximación de estas que mantenga dentro de lo posible sus características.

Ante este enfoque se proponen métodos de regularización que permitan invertir la matriz y por lo tanto estimar las reglas discriminantes. Varios métodos son propuestos por Guo et. al. en [4] o en Friedman et. al. [3]. De nuevo se supondrá que las variables proceden de una distribución normal y se verá que este método puede ser implementado tanto en el caso de poder asumir que las clases tienen la misma matriz de covarianzas o cuando esta hipótesis no se verifique y sea necesario estimar dicha matriz para cada grupo.

Si se denota por $\hat{\mathbf{\Sigma}}$ el correspondiente estimador de máxima verosimilitud de la matriz de covarianzas, puede verse que en vez de utilizar este estimador para obtener las funciones discriminantes (2.8) y (2.11), es posible emplear una penalización de esta cantidad que solucione el problema de su no invertibilidad. Dentro de las posibles regularizaciones que se pueden implementar algunas de ellas vienen dadas por $\tilde{\mathbf{\Sigma}}$ mediante

$$\tilde{\mathbf{\Sigma}} = \alpha\hat{\mathbf{\Sigma}} + (1-\alpha)\mathbf{I}, \tag{2.12}$$

$$\tilde{\mathbf{\Sigma}} = \alpha\hat{\mathbf{\Sigma}} + (1-\alpha)\mathrm{diag}(\hat{\mathbf{\Sigma}}), \tag{2.13}$$

$$\tilde{\mathbf{\Sigma}} = \alpha\hat{\mathbf{\Sigma}} + (1-\alpha)\hat{\sigma}^2\mathbf{I}, \tag{2.14}$$

$$\tilde{\mathbf{\Sigma}} = \lambda\hat{\mathbf{\Sigma}} + \mathbf{I}, \tag{2.15}$$

$$\tilde{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}} + \lambda\mathbf{I}. \tag{2.16}$$

con $\alpha \in [0,1]$ ó $\lambda \in [0,\infty)$.

Se puede ver que si se ignoran los términos constantes las tres formas de regularización presentadas son equivalentes en términos de la función discriminante, ya que lo que hacen es distorsionar el valor de la diagonal de la matriz de covarianzas sumando una cantidad positiva, lo cual se traduce en que ninguno de sus autovalores llegue a ser nulo.

Ante la utilización de estas versiones penalizadas se consigue estabilizar la estructura de covarianza muestral estimada, ya que no habrá más autovalores nulos y por lo tanto se resolverá el problema referente a la no invertibilidad. A pesar de estas ventajas se tiene ahora un estimador sesgado, frente al resultado insesgado que proporcionaba el estimador de máxima verosimilitud.

Esta penalización también cuenta con la ventaja de que consigue estabilizar las covarianzas individuales puesto que no habrá autovalores nulos en las versiones regularizadas. Esto hará que al calcular las reglas discriminantes no se tengan al menos $p - n$ direcciones con un gran peso, como puede verse interpretando (2.8) o (2.11), lo cual distorsionaba los grupos y las fronteras de decisión, por lo tanto los resultados de la clasificación.

2.3. La regresión logística como regla discriminante

Otra forma de afrontar el problema de discriminar en alta dimensión es recurrir a la **regresión logística** empleando esta como una regla de clasificación.

Para ello se ajusta un modelo de regresión en base a las valores conocidos de la muestra de entrenamiento y se estiman sus parámetros. De este modo cuando se quiera decidir en qué grupo clasificar una nueva observación basta obtener el valor correspondiente de su variable respuesta en base al modelo estimado y decidir en torno a este. En las Secciones 1.4 y 1.5 se mostró el desarrollo de este método y se vieron formas de solucionar los problemas que surgían cuando $p > n$ (Secciones 1.4.1, 1.4.2, 1.5.2 y 1.5.3).

Dado que el modelo visto consigue discriminar entre únicamente dos clases ($L = 2$), codificando los grupos por $y = 0$ e $y = 1$ respectivamente, se pueden emplear métodos como el *OVO (one versus one)*, también conocido como *majority voting*, o el *OVA (one versus all)* para conseguir clasificar en un contexto donde se tenga un número de $L > 2$ grupos.

Este primer procedimiento se basa en que para saber dónde clasificar un vector de tamaño p: $(x_1, \ldots, x_p)$, entre $L > 2$ clases, se procede a enfrentar las clases dos a dos para comprobar a cual de todos los grupos pertenece el nuevo dato con mayor probabilidad. De modo que se llevan a cabo $\binom{L}{2}$ contrastes acerca de la probabilidad que tiene la muestra de pertenecer a cada clase y finalmente se clasifica en aquella que obtenga el mejor resultado en términos de mayor número de victorias. Es decir, se acaba catalogando la muestra en el grupo que después de los $\binom{L}{2}$ contrastes ha sido escogido más veces como aquel que tiene la probabilidad más grande de que pertenezca la muestra. En caso de empate final se puede repetir el procedimiento entre los grupos igualados para conseguir un desempate, en caso de no ser capaz de alcanzar este se puede proceder a clasificar en cualquiera de los grupos seleccionados o sortear el resultado.

La segunda opción, *OVA*, realiza sólo L comparaciones o contrastes. En cada una selecciona uno de los $L > 2$ grupos y agrupa el resto de datos en otro grupo artificial, calcula la regla de clasificación y dictamina qué probabilidad tiene el nuevo dato de pertenecer al grupo l-ésimo, $l = 1, \ldots, L$. Tras haber calculado las correspondientes probabilidades se decide clasificar la nueva observación en el grupo para el cual se ha obtenido la mayor probabilidad tras los L pasos.

La forma de proceder es la siguiente, dentro de cada uno de estos contrastes se etiqueta un grupo con $y = 0$ y el otro con $y = 1$. De forma que si se quiere contrastar un grupo $l \in \{1, \ldots, L\}$ frente a un grupo $m \in \{1, \ldots, L\}$ con $m \neq l$ se tendrían los datos para cada uno de estos grupos respectivamente:

$$(0, x_{l1}), \ldots, (0, x_{ln_l}) \text{ con } n_l = \text{número de elementos del grupo l,}$$
$$(1, x_{m1}), \ldots, (1, x_{mn_m}) \text{ con } n_m = \text{número de elementos del grupo m,}$$

con los que se emplea la regresión logística para estimar que valor toma $\pi(x) = \mathbb{P}(Y = 1 \mid X = x)$ con el procedimiento ya visto. Si este es mayor de 0.5 la muestra tiene más posibilidades de pertenecer al grupo correspondiente a $Y = 1$ en este contraste, si es menor a $Y = 0$ y si es exactamente 0.5 tiene la misma probabilidad de pertenecer a ambos grupos, con lo que se podría dictaminar un empate o sortear el resultado. De esta forma, estudiando los $\binom{L}{2}$ o L resultados

respectivamente se dirá que la muestra pertenece al grupo que resulte el más probable.

2.4. Reglas de clasificación con matrices de covarianzas diagonales

En el análisis discriminante tanto lineal como cuadrático, a la hora de construir la regla de clasificación surgía el problema de la no invertibilidad de la matriz $\mathbf{\Sigma}$. En lo que sigue se propone una regla de clasificación, bajo las hipótesis consideradas, para solucionar este problema.

Esta regla se construye bajo la suposición de que la matriz de covarianzas de cada clase es diagonal, o lo que es equivalente que las variables son independientes entre sí. De esta forma se obtiene la **regla de clasificación estimada de matriz de covarianzas diagonal**, la cual proporciona un nuevo algoritmo que permite clasificar nuevos datos sin tener que enfrentarse al problema del mal condicionamiento de las matrices de covarianzas.

Pese a que esta nueva regla soluciona los problemas de estimación de las reglas discriminantes vistas cuando $p > n$, tiene el inconveniente de que entran en juego las $j = 1, \ldots, p$ covariables en su estimación, haciendo que el modelo resultante sea difícil de interpretar. Ante esta situación se propone una regularización que permite discernir qué variables son las más importantes cuando se quiere clasificar y cuales, por el contrario, se pueden omitir en el proceso. De esta forma se presenta el procedimiento de **encogimiento por centroides cercanos**.

2.4.1. Regla de clasificación estimada de matriz de covarianzas diagonal

Se ha explicado que cuando $p > n$ no se puede ajustar una regla de discriminación lineal ni cuadrática, por tanto, es necesario regularizar el método. En este nuevo procedimiento se va a asumir que las características de cada variable son independientes dentro de cada clase, esto es, que la matriz de covarianzas para cada grupo es diagonal. Para comprobar si se va a poder asumir este supuesto se pueden realizar test como el presentado en el Capítulo 3 de Seber [16].

La **regla de clasificación estimada de matriz de covarianzas diagonal** (también conocida como "regla de la independencia", véase [1]) para cada clase l va a depender de si la matriz de covarianzas es la misma o diferente para cada grupo. Estos casos se corresponden con la regla lineal o cuadrática respectivamente.

Distinta matriz de covarianzas entre grupos

En este caso se define la siguiente regla de clasificación para la clase l:

$$\delta_l(x^*) = -\sum_{j=1}^{p} \frac{(x_j^* - \mu_{lj})^2}{s_j^2} + 2\log \pi_l, \tag{2.17}$$

donde $x^* = (x_1^*, \ldots, x_p^*)^t$ es una muestra tomada, s_j es la desviación estándar de la característica j en la clase l-ésima y $\mu_{lj} = \sum_{l \in G_l} x_{lj}/n_l$ es la media de los n_l valores para la característica j en la clase l.

Se observa que en (2.17) la primera parte es la distancia estandarizada de x^* al centroide de la clase l-ésima al cuadrado, con signo negativo, mientras que la segunda es una corrección basada en la probabilidad a priori π_l de que una muestra aleatoria pertenezca al grupo l. De esta forma cuando la distancia sea idéntica entre dos grupos la probabilidad a priori será la que decida en que grupo clasificar. Por tanto, se clasificará en el grupo m cuando:

$$C(x^*) = m \quad \text{si } \delta_m(x^*) = \max_l \delta_l(x^*). \tag{2.18}$$

En consecuencia se va a clasificar en aquel grupo donde (2.17) sea menos negativa, es decir, se va a clasificar en el grupo m cuando su distancia estandarizada al cuadrado a su centroide sea la menor de todas, teniendo en cuenta la probabilidad que existe de que una muestra pertenezca al grupo m. Este método es equivalente al método de *clasificación de los centroides cercanos* o *algoritmo de K-medias*, Sección 2.5.2, después de una estandarización (véase sección 12.3, capítulo 12, del libro [3]), ya que se basa en clasificar en el grupo donde la distancia dada entre la muestra y su centroide sea la menor, usando la norma euclídea.

Se observa que ahora sería necesario estimar los centroides de cada clase, las probabilidades de pertenecer a una cierta clase y la desviación típica de cada característica, lo que siempre se puede llevar a cabo y soluciona el problema de la regla de discriminación cuadrática. Para una clase l se estimarían de la forma

$$\begin{aligned}
&\hat{\mu}_{lj} = \bar{x}_{lj} = \sum_{l \in G_l} \frac{x_{lj}}{n_l} \quad \forall j \in \{1, \ldots, p\} \Rightarrow \hat{\mu}_l = (\hat{\mu}_{l1}, \ldots, \hat{\mu}_{lp})^t = (\bar{x}_{l1}, \ldots, \bar{x}_{lp})^t, \\
&\hat{\pi}_l = \frac{n_l}{n} \quad \text{con } n_l : \text{número de muestras del grupo } l \quad \text{y} \quad n = \sum_{l=1}^{L} n_l, \\
&\hat{s}_j^2 = \sigma_j^2 = \sum_{l \in G_l} \frac{(x_{lj} - \hat{\mu}_{lj})^2}{n_l} = (\mathbf{X}_l - \hat{\boldsymbol{\mu}}_l)^t \mathbf{N}_l (\mathbf{X}_l - \hat{\boldsymbol{\mu}}_l) \quad \forall j \in \{1, \ldots, p\},
\end{aligned} \tag{2.19}$$

siendo $\mathbf{N}_l$ una matriz diagonal dada por $\mathbf{N}_l = \text{diag}(1/n_l, \ldots, 1/n_l) \in \mathcal{M}_{n_l \times n_l}$ y

$$\mathbf{X}_l = \begin{pmatrix} x_{l11} & \cdots & x_{l1p} \\ \vdots & \ddots & \vdots \\ x_{ln_l1} & \cdots & x_{ln_lp} \end{pmatrix} \in \mathcal{M}_{n_l \times p}, \quad \hat{\boldsymbol{\mu}}_\mathbf{l} = (\hat{\mu}_l^t, \ldots, \hat{\mu}_l^t) \in \mathcal{M}_{n_l \times p}.$$

Igual matriz de covarianzas entre grupos

Cuando la matriz de covarianzas es la misma entre las distintas clases, como ocurre en el análisis discriminante lineal, basta obtener las desviaciones típicas s_j de las características para una clase, ya que las demás serán idénticas. Por lo tanto se estimarán estas teniendo ahora en cuenta todos los datos muestrales y será necesario estimar únicamente p valores.

La regla de clasificación y el criterio para escoger el grupo a donde pertenece una nueva muestra siguen siendo los mismos que en (2.17) y (2.18) salvo que ahora s_j tendrá el mismo valor para todos los grupos, con lo que llega estimarla una sola vez.

En [1] se ha probado teóricamente que este método suele dar mejor resultado en alta dimensión que el análisis discriminante lineal o cuadrático estandarizado. En un contexto de alta dimensión es común que sea más óptimo trabajar con casos más sencillos, haciendo suposiciones como la considerada en este apartado, frente a regularizar métodos que dejan de funcionar. Esto es debido a que esta última opción precisa de estimaciones que aumentan en gran medida el error al clasificar. De esta forma aunque aceptar que las matrices de covarianzas son diagonales parece una suposición

muy fuerte no lo es tanto en comparación con realizar estimaciones en el caso de la alta dimensión cuando $p > n$.

2.4.2. Encogimiento por centroides cercanos

El **encogimiento por centroides cercanos** es una regularización del método de clasificación de matriz de covarianzas diagonal.

Este método persigue la filosofía de desprenderse de aquellas variables que no aportan información relevante a la hora de decidir en qué grupo clasificar una nueva muestra. Para este fin se busca contraer la media de cada clase hacia la media global, para cada característica por separado. De esta forma se definen las distancias d_{lj}, empleando los estimadores de (2.19), por

$$d_{lj} = \frac{\bar{x}_{lj} - \bar{x}_j}{m_l(\sigma_j + s_0)}, \tag{2.20}$$

donde $\bar{x}_j$ es la media total estimada de la característica j, $m_l^2 = 1/n_l - 1/n$ y s_0 es una constante positiva pequeña, normalmente se toma la mediana de los valores de los σ_j.

Al seguir suponiendo que la matriz de covarianzas para cada clase es diagonal, la varianza del numerador $\bar{x}_{lj} - \bar{x}_j$ es $m_l^2\sigma_j^2$ y por lo tanto el denominador estandariza la expresión. Por su parte, la constante s_0 hace que los valores de los d_{lj} no sean desorbitados cuando σ_j sea muy próximo o tome el valor cero. Se distinguen de nuevo dos casos dependiendo de si la matriz de covarianzas se supone igual para cada clase o distinta, afectando únicamente en la estimación de los valores σ_j.

Se encoge el valor de d_{lj} hacia cero usando un *criterio de umbralización suave* similar al visto en el caso de la regresión LASSO (Sección 1.1.2):

$$d'_{lj} = \text{sign}(d_{lj})(|d_{lj}| - \Delta)_+;$$

cada d_{lj} es reducido, en valor absoluto, por una cantidad pequeña $\Delta > 0$ y es ajustado a cero cuando el resultado es negativo. Aplicando esta expresión a (2.20) se obtiene

$$\tilde{\bar{x}}_{lj} = \bar{x}_j + m_l(\sigma_j + s_0)d'_{lj}.$$

Si se emplea $\tilde{\bar{x}}_{lj}$ en lugar del original $\bar{x}_{lj}$ en (2.17) se consigue la nueva regla de clasificación

$$\delta_l(x^*) = -\sum_{j=1}^{p} \frac{(x_j^* - \tilde{\bar{x}}_{lj})^2}{\sigma_j^2} + 2\log\pi_l, \tag{2.21}$$

donde al igual que anteriormente, se clasificará una muestra en un grupo m cuando $C(x^*) = m$ si $\delta_m(x^*) = \text{máx}_l\, \delta_l(x^*)$.

De esta forma se ve que únicamente las características que tienen un d'_{lj} no nulo para alguna de las respectivas clases van a jugar un papel relevante en la regla de clasificación. Esto, al igual que ocurría en el caso de los coeficientes de la regresión LASSO, permite desprenderse de aquellas variables que no aportan información provechosa a la hora de discriminar, reduciendo de esta forma la dimensión del problema y dotando de mayor interpretabilidad al algoritmo, especialmente en contextos donde se cuenta con un gran número de covariables como por ejemplo en algunas situaciones donde $p > n$.

El valor d'_{lj} será cero cuando se tenga que $(|d_{lj}| - \Delta)_+ \leq 0$, lo cual es equivalente a $|d_{lj}| < \Delta$ para cierto valor $\Delta > 0$ pequeño fijado de antemano. La interpretación de este resultado es la

siguiente, se tomará $d'_{lj} = 0$ cuando $|d_{lj}|$ sea próximo a cero, es decir cuando el centroide $\bar{x}_{lj}$ tenga un valor similar a $\bar{x}_j$ (véase 2.20). De esta forma, si no hay diferencias significativas entre el centroide de la característica j-ésima, $\bar{x}_j$, y su valor correspondiente en la muestra l-ésima, $\bar{x}_{lj}$, entonces se trabajará con $\tilde{\bar{x}}_{lj} = \bar{x}_j$ en vez de con $\bar{x}_{lj}$. La construcción de estos nuevos centroides de esta manera permite prescindir de aquella información que no es relevante y tener en cuenta sólo aquellas muestras que tienen mayor poder discriminatorio, en las cuales se aprecian diferencias con el centroide global, permitiendo la construcción de reglas de clasificación más óptimas.

Además, se puede ver que la regla (2.21) puede ser usada para construir estimadores de las probabilidades de pertenecer a cada una de las clases de la forma

$$\hat{p}_l(x^*) = \frac{e^{\frac{1}{2}\delta_l(x^*)}}{\sum_{m=1}^{L} e^{\frac{1}{2}\delta_m(x^*)}}.$$

Estas probabilidades pueden ser útiles para valorar las clasificaciones o decidir no clasificar una muestra determinada.

2.5. Métodos de clasificación no paramétricos

Hasta ahora se han expuesto métodos de carácter paramétrico, los cuales requieren de hipótesis distribucionales para poder estimar una regla de clasificación. Estas condiciones pueden ser problemáticas cuando no se pueden aceptar las hipótesis requeridas o cuando la estimación de ciertos parámetros distribucionales es muy inexacta o ineficiente.

En estos contextos se puede recurrir a reglas de carácter no paramétrico como la **regla de clasificación de los K-vecinos más cercanos** o la **regla de las K-medias**, las cuales requieren únicamente de la información muestral para ser capaces de implementar un análisis discriminante.

2.5.1. Regla de clasificación de los K-vecinos más cercanos

La **regla de clasificación de los K-vecinos más cercanos**, a diferencia de los algoritmos anteriores, permite realizar una discriminación entre L clases sin necesidad de suponer ninguna hipótesis distribucional ni de ajustar un modelo determinado. Estas características permiten barrer un campo más amplio de situaciones de las que se tenían en cuenta dentro del análisis discriminante lineal o cuadrático.

De esta forma, dado un punto x_0, para clasificarlo entre uno de los L grupos determinados, se encuentran los K puntos muestrales más cercanos a este, denotados por $x_{(r)} \quad r = 1, \ldots, K$, y después se clasifica el nuevo dato empleando la técnica de *majority voting* entre los K vecinos seleccionados. Esta técnica se basa en que, una vez determinados los K puntos más cercanos, se cuenta cuántos datos de cada clase hay dentro de dicho subgrupo y a continuación se clasifica la nueva observación en el grupo que tenga un mayor número. Ante los empates se puede proceder aumentando el número de datos seleccionados en una unidad hasta que se consiga determinar el grupo que se repite más veces dentro del entorno.

En consecuencia, únicamente se necesita una distancia que caracterice el espacio de covariables para poder implementar este algoritmo. Normalmente, antes de proceder se estandarizan las variables explicativas de forma que tengan media cero y varianza unidad, dado que es posible que estén medidas en unidades distintas y esto pueda ocasionar problemas.

Por simplicidad se puede considerar la distancia euclídea correspondiente a dicho espacio para los datos estandarizados. En caso de preferir no estandarizar los datos y tener en cuenta las estructuras de covarianzas existentes entre estos, para definir las distancias se pueden emplear métricas

tales como la de Mahalanobis. De esta forma, para poder emplear este procedimiento únicamente será necesario conocer

$$d_{(r)} = \|x_{(r)} - x_0\|, \quad r = 1, \ldots, K.$$

Además, puede verse que este algoritmo es capaz de adaptarse al caso de variables cualitativas u ordinales, sin más que modificar la elección de las distancias consideradas (estos cambios se recogen en el capítulo 14 de [3]). Proporcionando una metodología válida para diversos tipos de datos.

En conclusión, se tiene un método sencillo que permite resolver gran cantidad de problemas de clasificación. A menudo este método es una buena elección cuando las fronteras de decisión son muy irregulares, caso que no es posible recoger de forma adecuada empleando técnicas como el LDA o el QDA, donde se suponen fronteras regulares.

En el caso de alta dimensión en el número de covariables aparece el problema de la *maldición de la dimensionalidad*, introducido anteriormente en el Sección 1.2, donde los entornos pierden su carácter local según aumenta la dimensión de p. En consecuencia, la consideración de los K vecinos más cercanos puede seleccionar puntos que realmente están bastante alejados del nuevo dato, produciendo sesgo y degradando el rendimiento de la regla.

Para solucionar este problema Hastie et. al. proponen la **regla discriminante adaptativa de los vecinos cercanos** o *DANN (Discriminant Adaptative Nearest-Neighbour)* [5]. Este método opta por adaptar la métrica usada en la clasificación de forma que los entornos resultantes se extiendan en las direcciones en las cuales las probabilidades de pertenecer a cada clase no cambien demasiado.

Esto se debe a que en un contexto de alta dimensión en el espacio de las covariables, las probabilidades de cada clase quizás cambien únicamente en un subespacio de menor dimensión, haciendo que sea ventajoso adaptar la métrica. De este modo, en un primer momento se forma un entorno de M puntos y a continuación se emplea la distribución de los puntos para decidir como deformar el entorno, esto es, adaptar la métrica. Posteriormente, la métrica adaptada es después empleada por la regla de vecinos cercanos sobre el punto a clasificar. Puesto que este proceso se lleva a cabo sobre cada punto que se quiere clasificar se obtiene una métrica distinta para cada caso.

Se define así la *métrica discriminante adaptativa de los vecinos cercanos* para un punto x_0 como

$$D(x, x_0) = (x - x_0)^t \bar{\mathbf{\Sigma}} (x - x_0), \tag{2.22}$$

donde

$$\bar{\mathbf{\Sigma}} = \mathbf{W}^{-1/2}[\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} + \epsilon \mathbf{I}]\mathbf{W}^{-1/2} = \mathbf{W}^{-1/2}[\mathbf{B}^* + \epsilon \mathbf{I}]\mathbf{W}^{-1/2}. \tag{2.23}$$

Aquí $\mathbf{W} = \sum_{r=1}^{K} \pi_r \mathbf{W}_r$ designa a la matriz de covarianzas de las clases agrupadas, siendo π_r la probabilidad a priori del dato r-ésimo de pertenecer a la clase l, $l = 1, \ldots, L$. Por su parte $\mathbf{B} = \sum_{r=1}^{K} \pi_r (\bar{x}_r - \bar{x})(\bar{x}_r - \bar{x})^t$ es la matriz de covarianzas entre clases, donde $\mathbf{W}$ y $\mathbf{B}$ se calculan teniendo en cuenta únicamente los K datos seleccionados y ϵ es un pequeño parámetro de ajuste a determinar.

Teniendo en cuenta la forma de $\bar{\mathbf{\Sigma}}$ en (2.23) se ve que es necesario calcular la inversa de $\mathbf{W} \in \mathcal{M}_{p \times p}$, la cual no es invertible al igual que ocurría en el caso del análisis discriminante tanto lineal como cuadrático con las matrices de covarianzas. Para solucionar este problema se puede pensar en reemplazar $\mathbf{W}$ por una versión penalizada de esta que sí sea invertible, $\tilde{\mathbf{W}}$, pudiendo emplear técnicas de regularización como las presentadas en (2.12)-(2.16), la cuales introducirán

sesgo en el estimador pero permitirán obtener la métrica adaptada.

La interpretación de la métrica (2.22) es que esta empieza ajustando entornos esféricos respecto de $\mathbf{W}$ y después los alarga en la dirección de los autovalores nulos de $\mathbf{B}^*$, direcciones en las cuales se tiene localmente que las medias muestrales no son distintas. Finalmente el parámetro ϵ redondea el entorno, pasando de una banda infinita a un elipsoide. En la Figura 2.2 se ilustra un ejemplo de este procedimiento.

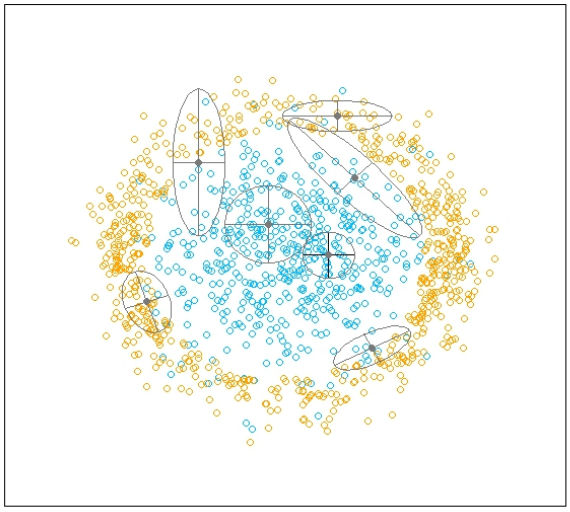

Figura 2.2: Ejemplo de aplicación del algoritmo de K-vecinos cercanos *DANN* entre dos grupos (azul y naranja) para varios puntos. Los elipsoides muestran los entornos adaptados a cada dato.

2.5.2. Regla de clasificación de las K-medias

De modo similar al método de los K-vecinos más cercanos, Sección 2.5.1, la **regla de clasificación de las K-medias** o de *K-means* puede ser empleada en cualquier circunstancia donde se cuente con L grupos totalmente determinados de antemano, sin necesidad de la suposición de ninguna hipótesis distribucional. Pese a que este es un método de agrupamiento, el cual se encarga de determinar K grupos de un conjunto de datos, puede ser empleado como regla de clasificación para determinar a que grupo pertenece un nuevo dato cuando las clases están totalmente determinadas de antemano.

De nuevo, lo único necesario para poder desempeñar esta clasificación será la consideración de una métrica adecuada que permita medir las distancias entre los elementos existentes en el espacio formado por las covariables.

Una vez seleccionada una distancia, como podría ser la euclídea o la Mahalanobis en p dimensiones, este algoritmo se basa en clasificar una nueva observación en el grupo con el centroide más cercano. Para este fin se estimará cada centroide de cada grupo mediante lo datos muestrales y cuando se incluya una nueva muestra en una clase se recalcularán estas medidas centrales.

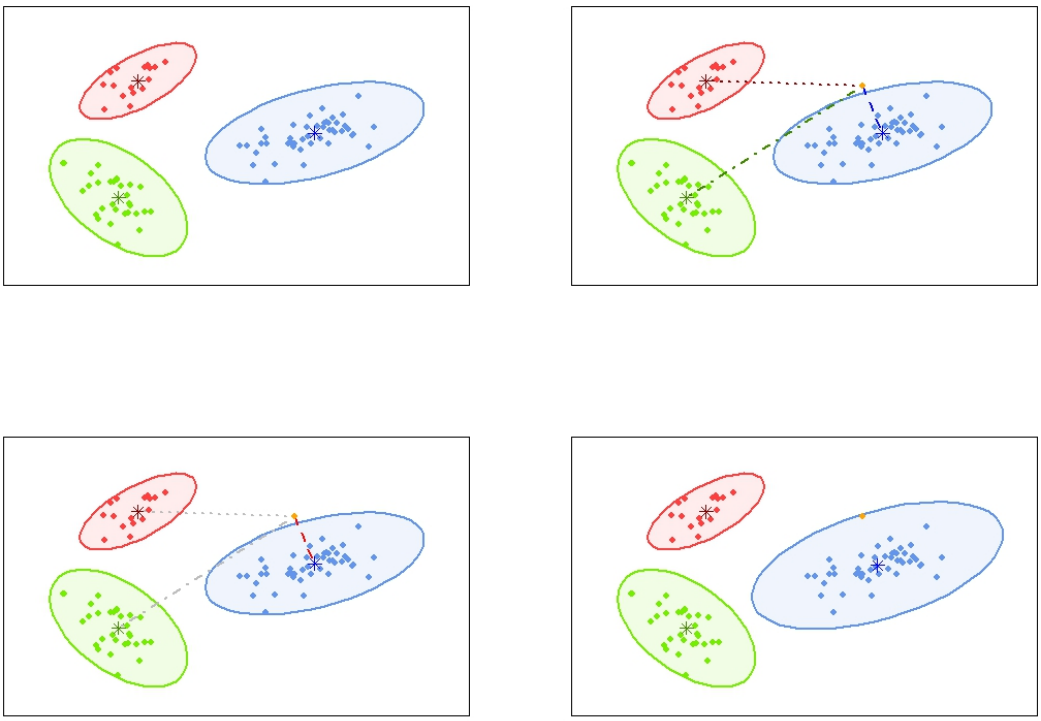

Figura 2.3: Ejemplo de aplicación del algoritmo de K-medias considerando $L = 3$ grupos en dos dimensiones. Una vez determinado o estimado el centroide de cada grupo se clasifica un nuevo dato (naranja) mediante la comparación de las distancias a cada media (líneas granate, verde oscuro y azul oscuro). En este caso es fácil ver que el grupo más próximo en términos de centro de gravedad es el grupo azul. De esta forma se añade la nueva observación a la clase y se recalcula el vector de medias para volver a implementar el algoritmo cuando se necesite clasificar un nuevo dato.

La Figura 2.3 muestra un ejemplo de como procede este algoritmo para clasificar una nueva observación.

2.6. Support Vector Machine (SVM)

En el caso de un problema de clasificación entre dos clases distintas se han visto técnicas que se basan en la obtención de un hiperplano o hipersuperficie que permita la mayor separación entre ellas. Esto permite clasificar a un individuo o elemento viendo en que región cae de las separadas por dicho hiperplano o hipersuperficie. En esta sección se van a describir generalizaciones de fronteras de decisión lineales y no lineales para la clasificación. Además se van a tener en cuenta los casos no separables, es decir, aquellos en que las clases se solapen.

En 1963 los matemáticos Vapnik y Nelder introducen unos nuevos métodos de clasificación que resuelven este problema en su artículo *"Recognition of Patterns with help of Generalized Portraits"* (véase [18]). Estas técnicas son conocidas en inglés como *Support Vector Machine (SVM)*, que traduciremos como **maquinaria de soporte de vectores**.

Estos procedimientos tienen el beneficio de que no será necesario asumir ninguna hipótesis sobre los datos, esto se traduce en que no importará la distribución que puedan seguir o la ca-

racterización de sus parámetros poblacionales como la estructura de la matriz de covarianzas. De esta forma se está ante otro método de clasificación no paramétrico. En conclusión se obtiene un método mucho más genérico que permitirá discriminar entre múltiples escenarios y que además se verá que funciona de forma eficiente en el contexto de alta dimensión $p > n$.

2.6.1. Clasificador SVM para dos poblaciones. Extensión al caso de L poblaciones

El objetivo de este nuevo procedimiento es construir de forma óptima un hiperplano que separe perfectamente dos clases.

Si se supone que se tiene un conjunto de n datos de la forma $(x_1, y_1), \ldots, (x_n, y_n)$ con $x_i \in \mathbb{R}^p$. Ahora, siguiendo una codificación binaria de forma similar a la regresión logística, se va a tomar $y_i \in \{-1, 1\}$ para designar que un dato pertenece a uno u otro grupo. Se puede ver que es posible definir un hiperplano de la forma

$$\{x : f(x) = x^t\beta + \beta_0\} \tag{2.24}$$

donde β es un vector unidad ($\|\beta\| = 1$). De esta forma una regla de clasificación inducida por $f(x)$ es

$$G(x) = \text{sign}[x^t\beta + \beta_0],$$

la cual es posible gracias a la particular codificación que toman los y_i.

En (2.24) $f(x)$ da la distancia de un punto x (en signo positivo o negativo según la clase en la que se encuentre) al hiperplano $f(x) = x^t\beta + \beta_0 = 0$, puesto que $f(x) - 0 = x^t\beta + \beta_0$.

Clases sin solapamiento

En el caso donde las clases son totalmente separables se va a poder encontrar una función $f(x) = x^t\beta + \beta_0$ que cumpla $y_i f(x_i) > 0 \quad \forall i$. Basta tener en cuenta que en el caso separable siempre va a existir un hiperplano $x^t\beta + \beta_0 = 0$ tal que los puntos de una clase verificarán $x^t\beta + \beta_0 > 0$ mientras que los del segundo grupo la condición contraria, $x^t\beta + \beta_0 < 0$. De esta forma basta codificar los y_i relativos a la primera clase por 1 y los restantes por -1 para garantizar que $y_i(x^t\beta + \beta_0) > 0$, lo cual equivale a que $y_i f(x_i) > 0$ para todo i. Puesto que se tiene una desigualdad estricta se puede garantizar que existe una cantidad $M > 0$ desconocida tal que $y_i(x^t\beta + \beta_0) \geq M \quad \forall i$.

Por tanto, se puede encontrar el hiperplano que crea el mayor margen entre los datos muestrales de las clases -1 y 1 resolviendo el problema de optimización

$$\begin{cases} \max\limits_{\beta, \beta_0, \|\beta\|=1} M \\ \text{sujeto a } y_i(x_i^t\beta + \beta_0) \geq M, \quad i = 1, \ldots, n. \end{cases} \tag{2.25}$$

El margen va a tener de ancho una medida de $2M$, donde $M > 0$ es la máxima distancia que se puede considerar desde el hiperplano a cada clase verificando que ningún dato va a quedar dentro del margen. Es decir, los datos de la primera clase verificarán ahora que $x^t\beta + \beta_0 \geq M$ mientras que los de la segunda serán aquellos que cumplan que $x^t\beta + \beta_0 \leq -M$. En la Gráfica 2.4 se muestra un ejemplo bidimensional de esta situación. Para medir esta distancia se puede usar una norma como la euclídea del correspondiente espacio $\mathbb{R}^p$.

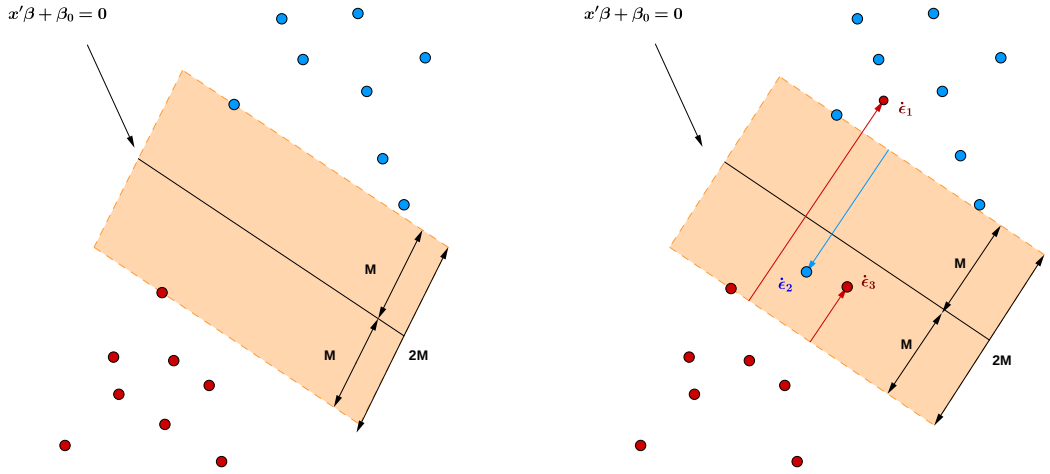

Figura 2.4: Soporte clasificador de vectores. El panel de la izquierda muestra el caso separable. La frontera de decisión es la línea gruesa, mientras que las líneas discontinuas designan la frontera del margen maximal de anchura $2M = 2/\|\beta\|$. La gráfica de la derecha muestra el caso con solapamiento. Los puntos $\dot{\epsilon}_i$ están en el lado equivocado de su margen, distando una cantidad $\dot{\epsilon}_i = M\epsilon_i$; los puntos que están en el lado correcto distan $\dot{\epsilon}_i = 0$. El margen es maximizado sujeto a una restricción $\sum \epsilon_i \leq cte$. Por tanto $\sum \dot{\epsilon}_i$ es la distancia total de los puntos que se encuentran en el lado equivocado de su margen.

Puede verse que el margen maximal se corresponde con $M = 1/\|\beta\|$. Para esto basta con tener en cuenta que los hiperplanos que determinan la posición de los datos de cada clase pueden ser reescritos como $H_1 = x^t\beta+\beta_0 = 1$ para la clase correspondiente a $y = 1$ y $H_{-1} = x^t\beta+\beta_0 = -1$ para $y = -1$ considerando valores adecuados de β y β_0, en consecuencia dándose que $y_i(x_i^t\beta + \beta_0) \geq 1$. Así, si se toma un punto $x_0 \in H_{-1}$, cumpliendo por tanto que $x_0^t\beta + \beta_0 = -1$, para calcular la distancia entre H_1 y H_{-1} sólo es necesario calcular la distancia perpendicular entre x_0 y H_1, la cual denotaremos por $2M$. Como $\frac{\beta}{\|\beta\|}$ es el vector normal unitario del hiperplano H_1, entonces se tiene que

$$\left(x_0 + 2M\frac{\beta}{\|\beta\|}\right)^t \beta + \beta_0 = 1, \tag{2.26}$$

dado que $x_0 + 2M\frac{\beta}{\|\beta\|}$ es un punto de H_1 por definición de $2M$. Expandiendo la ecuación (2.26) se llega a que

$$x_0^t\beta + 2M\frac{\beta^t}{\|\beta\|}\beta + \beta_0 = 1 \Rightarrow x_0^t\beta + 2M\frac{\|\beta\|^2}{\|\beta\|} + \beta_0 = 1 \Rightarrow x_0^t\beta + 2M\|\beta\| + \beta_0 = 1$$
$$\Rightarrow \underbrace{x_o^t\beta + \beta_0}_{-1} = 1 - 2M\|\beta\| \Rightarrow -1 = 1 - 2M\|\beta\| \Rightarrow 2M = \frac{2}{\|\beta\|}.$$

En conclusión, el problema (2.25) puede reescribirse como

$$\begin{cases} \min\limits_{\beta,\beta_0} \|\beta\| \\ \text{sujeto a } y_i(x_i^t\beta + \beta_0) \geq 1, \quad i = 1, \ldots, n \end{cases} \tag{2.27}$$

donde se está teniendo en cuenta la restricción de la norma de β y que $M = \frac{1}{\|\beta\|}$. El problema (2.27) es un problema convexo de optimización y representa la manera común de escribir el criterio para el soporte de vectores en el caso de dos clases completamente separadas.

Clases con solapamiento

Si ahora se supone que hay solapamiento entre ambas clases, la forma de enfrentar este problema para obtener el margen es maximizar M pero permitiendo a algunos datos estar dentro del "lado equivocado del margen". Para este fin se define el vector de pesos $\xi = (\xi_1, \ldots, \xi_n)^t$, el cual mide las distancias entre cada dato y su clase en base a la separación dada por el margen, es decir, si un dato se encuentra dentro de su clase su distancia será cero mientras que si está en el lado incorrecto del margen esta será positiva. De esta forma la restricción impuesta en (2.25) cambia, para la cual se pueden ver dos nuevas formas de imponerla:

$$y_i(x_i^t\beta + \beta_0) \geq M - \xi_i, \tag{2.28}$$

ó

$$y_i(x_i^t\beta + \beta_0) \geq M(1 - \xi_i), \tag{2.29}$$

$\forall i\ \xi_i \geq 0, \quad \sum_{i=1}^n \xi_i \leq cte.$

Las opciones (2.28) y (2.29) dan diferentes soluciones. La restricción (2.28) mide el solapamiento en *distancia absoluta* al margen mientras que la segunda (2.29) lo mide en *distancia relativa*, la cual cambia según la anchura M del margen. La primera de las nuevas restricciones proporciona un problema de optimización no convexo, mientras que la segunda da a resolver un problema de optimización convexo y por lo tanto facilita la obtención de una solución. En consecuencia se emplea la segunda opción (2.29) para generalizar el criterio de clasificación del soporte de vectores.

Veamos la idea de esta formulación. El valor ξ_i en la restricción $y_i(x_i^t\beta + \beta_0) \geq M(1 - \xi_i)$ se interpreta como la cantidad proporcional por la cual la predicción $f(x_i) = x_i^t\beta + \beta_0$ está en el lado erróneo del margen. Por tanto, acotando la suma $\sum \xi_i$ se está acotando la proporción total de predicciones que caen en el lado equivocado. Una mala clasificación para un nuevo dato x_i ocurre cuando $\xi_i > 1$, así que acotando $\sum \xi_i$ por un valor C positivo determinado de antemano, se consigue que el número máximo de muestras mal clasificadas sea a lo sumo C. Pese a poder determinar el máximo número de muestral mal clasificadas se desconocerán los datos que se han clasificado de forma errónea siempre que no se tenga información a priori del grupo al que pertenecen.

De esta forma se transforma el problema anterior de optimización para dos clases bien separadas (2.27) en uno que tiene en cuenta el solapamiento entre ellas de la forma

$$\text{mín}\, \|\beta\| \quad \text{sujeto a} \begin{cases} y_i(x_i^t\beta + \beta_0) \geq 1 - \xi_i \quad \forall i, \\ \xi_i \geq 0, \quad \sum \xi_i \leqslant cte. \end{cases} \tag{2.30}$$

El problema (2.30) es la forma usual en la que se define el clasificador para el método de soporte de vectores en el caso de dos clases con solapamiento. Este es un problema de optimización cuadrático convexo. Por la naturaleza de este criterio se ve que los puntos que están bien situados dentro de la frontera de su clase no juegan un gran papel a la hora de decidir la forma de esta. Es una buena propiedad que diferencia este método del Análisis Discriminante Lineal que determinaba la frontera de decisión a través de las covarianzas y los centroides de las clases.

Calculo de la regla discriminante

Seguidamente se expondrá como calcular el clasificador para el soporte de vectores para ambos casos: clases separadas y clases con solapamiento. Para este fin se resuelve el problema (2.30) usando multiplicadores de Lagrange. Se expresa (2.30) en la forma equivalente:

$$\begin{aligned} &\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{sujeto a} \quad \xi_i \geqslant 0, \quad y_i(x_i^t\beta + \beta_0) \geqslant 1 - \xi_i \quad \forall i \end{aligned} \tag{2.31}$$

se observa que en (2.31) el "parámetro de coste" C remplaza a la constante de (2.30); el caso de clases separadas se corresponde con $C = \infty$.

La función de Lagrange para (2.31) es

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(x_i^t\beta + \beta_0) - (1 - \xi_i)] - \sum_{l=1}^{n}\mu_i\xi_i \tag{2.32}$$

con lo que minimizamos β, β_0 y ξ_i derivando en (2.32) respecto a cada variable e igualando a cero obteniendo

$$\beta = \sum_{i=1}^{n}\alpha_i y_i x_i, \tag{2.33}$$

$$0 = \sum_{i=1}^{n}\alpha_i y_i, \tag{2.34}$$

$$\alpha_i = C - \mu_i, \quad \forall i. \tag{2.35}$$

Sustituyendo (2.33)-(2.35) en (2.32) se obtiene el Lagrangiano de la función objetivo del problema dual

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\alpha_i\alpha_{i'}y_i y_{i'}x_i^t x_{i'}, \tag{2.36}$$

el cual da una cota inferior de la función objetivo (2.31) para cada punto factible. Se maximiza L_D sujeto a $0 \leq \alpha_i \leq C$ y $\sum_{i=1}^{n}\alpha_i y_i = 0$. Además de (2.33)-(2.35) las condiciones de Karush-Kuhn-Tucker incluyen las restricciones

$$\alpha_i[y_i(x_i^t\beta + \beta_0) - (1 - \xi_i)] = 0, \tag{2.37}$$

$$\mu_i\xi_i = 0, \tag{2.38}$$

$$y_i(x_i^t\beta + \beta_0) - (1 - \xi_i) \leq 0, \tag{2.39}$$

para $i = 1, \ldots, n$.

Juntando estas ecuaciones (2.33)-(2.39) queda caracterizada de forma única la solución para el problema primal y para el dual.

De (2.33) se extrae que la solución para β tiene la forma

$$\hat{\beta} = \sum_{i=1}^{n}\hat{\alpha}_i y_i x_i, \tag{2.40}$$

con coeficientes no nulos $\hat{\alpha}_i$, unicamente para aquellas observaciones i para las cuales la restricción en (2.39) son exactamente conocidas (de acuerdo con (2.37)). Estas observaciones son llamadas los *vectores del soporte* dado que $\hat{\beta}$ es representado en términos de ellas únicamente. Sobre estos puntos del soporte, algunos se quedarán justo en el borde del margen ($\hat{\xi}_i = 0$) y por tanto (2.38) y (2.35) serán caracterizadas por $0 < \hat{\alpha}_i < C$; el residuo ($\hat{\xi}_i > 0$) tiene $\hat{\alpha}_i = C$. De (2.37) se puede ver que todos esos puntos del margen ($0 < \hat{\alpha}_i$, $\hat{\xi}_i = 0$) pueden ser usados para hallar β_0, y por tanto que es posible usar una media de todas las soluciones para conseguir estabilidad numérica.

Maximizar el dual (2.36) es un problema de programación cuadrática convexo más sencillo que el primal (2.32), y puede ser resuelto con técnicas estándar.

Dadas las soluciones $\hat{\beta}_0$ y $\hat{\beta}$, la función de decisión se puede escribir como:

$$\hat{G}(x) = \text{sign}[\hat{f}(x)] = \text{sign}[x^t\hat{\beta} + \hat{\beta}_0] \tag{2.41}$$

El parámetro de ajuste de este procedimiento es el parámetro de coste C. El valor óptimo de C puede ser estimado por validación cruzada.

Para generalizar este método en la clasificación entre $l > 2$ grupos distintos pueden llevarse a cabo procesos como el *one versus one (OVO)* o el *one versus all (OVA)* que se explicaban en el caso de emplear la regresión logística como una regla de clasificación, Sección 2.3. La diferencia es que ahora, para determinar en qué clase o grupo se va a clasificar la nueva muestra se calcula la denominada *confianza* (la distancia al hiperplano con su correspondiente signo) para cada uno de los l clasificadores, de modo que la clase elegida será aquella que obtenga la mayor confianza en valor absoluto.

Bajo una situación de alta dimensión, es decir, cuando $p > n$ la falta de regularidad del método de clasificación del soporte de vectores a menudo trabaja tan bien como la mejor versión regularizada, lo que hace que este método siga siendo útil. Además, el sobreajuste no suele ser un problema, en parte a causa de la insensibilidad de pérdida de clasificación errónea. Esta opción es atractiva porque normalmente los grupos pueden ser perfectamente separados por un hiperplano, al menos que existan vectores iguales entre las clases, lo cual permite afrontar un problema de grandes dimensiones como en el contexto donde $p > n$ obteniendo una frontera de decisión lineal en la mayoría de los casos.

2.6.2. Cálculo de los SVM generalizados para la clasificación en dos poblaciones. Extensión al caso de L poblaciones

Existen otros procedimientos que permiten hacer la clasificación más flexible aumentando el espacio de parámetros y permitiendo una mejora en la discriminación. Para aumentar este espacio se usan expansiones en bases funcionales como pueden ser las polinómicas o de splines.

Las fronteras lineales en este espacio ampliado se transforman en fronteras no lineales en el espacio inicial de parámetros y normalmente permiten mejorar la separación de las clases. De esta forma se consiguen fronteras no lineales que se adaptan mejor para conseguir una discriminación más refinada sin más que trabajar de la forma vista anteriormente con las fronteras lineales del espacio ampliado.

Una vez seleccionadas las funciones de base que se denotarán ahora por $\{h_m(x)\}_{m=1}^M$, el proceso es el mismo que el visto anteriormente salvo que ahora se ajusta el clasificador del soporte de vectores usando los nuevos parámetros $h(x_i) = (h_1(x_i), \ldots, h_M(x_i))^t \quad i = 1, \ldots, n$ y se tendrá una función no lineal $\hat{f}(x) = h(x)^t\hat{\beta} + \hat{\beta}_0$.

La regla de clasificación estimada será $\hat{G}(x) = \text{sign}(\hat{f}(x))$, donde se clasifica un nuevo dato en uno u otro grupo dependiendo del signo de $\hat{f}(x)$.

La maquinaria clasificadora de soporte de vectores, como se ha traducido la expresión inglesa *"support vector machine classifier"*, es una extensión de esta idea, donde la dimensión del espacio aumentado puede llegar a ser muy grande, infinita incluso.

A la hora de calcular el clasificador del método SVM se puede ver el problema de optimización (2.32) y su solución de una forma diferente que involucraría sólo a las variables a través de productos interiores. Se lleva a cabo este proceso con los parámetros del espacio transformado $h(x_i)$. De esta forma la función Lagrangiana del problema dual (2.36) tendría la estructura

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} \langle h(x_i), h(x_{i'}) \rangle = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_{i'} y_i y_{i'} K(x_i, x_{i'}). \quad (2.42)$$

Se observa que (2.42) involucra a $h(x)$ a través del producto interior, en consecuencia basta conocer el valor de una función kernel

$$K(x, x') = \langle h(x), h(x') \rangle,$$

la cual calcule el producto interior en el espacio transformado. K deberá ser una función simétrica y semidefinida positiva.

De (2.33) se ve que la función solución $f(x)$ puede ser escrita de la forma

$$\begin{aligned} f(x) &= h(x)^t \beta + \beta_0 \\ &= \sum_{i=1}^{n} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0, \end{aligned} \quad (2.43)$$

como antes, dados α_i, β_0 puede ser determinado resolviendo $y_i f(x_i) = 1$ en (2.43) para los x_i para lo cuales $0 < \alpha_i < C$.

De este modo, de (2.43) se extrae que la solución puede escribirse como

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_l y_i K(x, x_i) + \hat{\beta}_0. \quad (2.44)$$

El papel del parámetro C es evitar que el espacio ampliado de parámetros sea demasiado grande, ya que la separación perfecta se consigue a menudo en dicho espacio.

De esta forma se obtiene un método de clasificación que proporciona fronteras de separación entre clases no lineales que no son necesariamente regulares. Al igual que en el clasificador de soporte de vectores (sección 2.6.1), cuando nos encontramos en alta dimensión con $l > 2$ grupos se pueden aplicar métodos vistos como el *OVO* o el *OVA*.

Bibliografía

[1] BICKEL, Peter J.; LEVINA, Elizaveta. Covariance regularization by thresholding. The Annals of Statistics, 2008, p. 2577-2604.

[2] EFRON, Bradley, et al. Least angle regression. The Annals of statistics, 2004, vol. 32, no 2, p. 407-499.

[3] FRIEDMAN, Jerome; HASTIE Trevor; TIBSHIRANI Robert. The elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer, 2009.

[4] GUO, Yaqian; HASTIE, Trevor; TIBSHIRANI, Robert. Regularized linear discriminant analysis and its application in microarrays. Biostatistics, 2006, vol. 8, no 1, p. 86-100.

[5] HASTIE, Trevor; TIBSHIRANI, Robert. Discriminant adaptive nearest neighbor classification and regression. En Advances in Neural Information Processing Systems. 1996. p. 409-415.

[6] HASTIE Trevor; TIBSHIRANI Robert. Generalized Additive Models. First Edition. Chapman and Hall, 1990.

[7] HOERL, Arthur E.; KENNARD, Robert W. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 1970, vol. 12, no 1, p. 55-67.

[8] HONG, Zi-Quan; YANG, Jing-Yu. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern recognition, 1991, vol. 24, no 4, p. 317-324.

[9] LOHR, Steve. The age of big data. New York Times, 2012, vol. 11.

[10] MCAFEE, Andrew, et al. Big data. The management revolution. Harvard Bus Rev, 2012, vol. 90, no 10, p. 61-67.

[11] McCULLAGH, P.; NELDER, J. Generalized Linear Models. Second Edition. Chapman and Hall, 1989.

[12] NELDER, John A.; BAKER, R. Jacob. Generalized linear models, 1972.

[13] PEÑA, Daniel. Análisis de datos multivariantes. McGraw Hill, 2002.

[14] RAVIKUMAR, Pradeep, et al. Spam: Sparse additive models. En Proceedings of the 20th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2007. p. 1201-1208.

[15] R CORE TEAM (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[16] SEBER, G.A.F. Multivariate observations. Wiley, 1984.

[17] TIBSHIRANI, Robert. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society. Series B (Methodological), 1996, p. 267-288.

[18] VAPNIK, V. N.; LERNER, A. Ya. Recognition of Patterns with help of Generalized Portraits. Avtomat. i Telemekh, 1963, vol. 24, no 6, p. 774-780.

[19] WOOD, Simon N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association, 2004, vol. 99, no 467, p. 673-686.

[20] WOOD, Simon N. Generalized additive models: an introduction with R. Chapman and Hall, 2006.