Trabajo Final de Máster - Curso 2023/2024

Selección de variables agrupadas en regresión logística desde una perspectiva bayesiana

Autora: **Carolina Mulet Rojas**

Tutora: ANABEL FORTE DELTELL

Máster en Bioestadística

Agradecimientos

A Mario, Anabel y Gonzalo.

Índice general

Índice de tablas

Índice de figuras

1. Diagnóstico temprano de la enfermedad de Alzheimer

En términos globales, la enfermedad de Alzheimer es la mayor causa de demencia[1] (representa entre un 60 y un 80 % de los casos de demencia), siendo los problemas de memoria, pensamiento y comportamiento la sintomatología más característica. Estos presentan un desarrollo lento, pero llegan a condicionar gravemente el día a día de las personas que la padecen.

La importancia de un diagnóstico temprano radica en aspectos puramente paliativos puesto que la enfermedad es irreversible y no puede detenerse. Un diagnóstico temprano puede mejorar el posible beneficio de los tratamientos administrados junto a la calidad de vida de las personas. Asimismo, permite aumentar la participación en ensayos clínicos para mejorar la investigación en la materia.

Sin embargo, el diagnóstico del Alzheimer requiere de exhaustivas evaluaciones médicas puesto que la causa precisa de una demencia es difícil de determinar. Además, los criterios de diagnóstico suelen basarse en datos clínicos pese a estar ampliamente aceptado que el desarrollo de la enfermedad tiene lugar décadas antes de la aparición de los primeros síntomas.

Actualmente se conoce la existencia de biomarcadores del líquido cefalorraquídeo, pero la obtención de dichas mediciones es costosa, invasiva y con posibles efectos secundarios en el paciente. Por ello, a lo largo de las últimas décadas se ha perseguido una definición biológica basada en biomarcadores de fácil obtención, como en muestras de plasma sanguíneo, que determine la respuesta del cuerpo a la neuropatología subyacente.

En Janeiro *et al.* (2020), se realizó una revisión de resultados obtenidos en la medición de biomarcadores de fluidos biológicos, con el objetivo de facilitar un diagnóstico temprano. En este trabajo se perseguirá el mismo objetivo con mediciones obtenidas del plasma de pacientes diagnosticados con y sin demencia en el hospital La Fe de Valencia.

[1] Término genérico que engloba la pérdida de memoria y de otras habilidades cognitivas que comprometan la vida cotidiana de las personas.

1.1. Complejidad del problema

El número de variables recogidas en las muestras de plasma sanguíneo es elevado, por lo que una selección de las variables es necesaria. Esto ya se realizó en Forte *et al.* (2024) mediante una aproximación bayesiana. Sin embargo, no se tuvo en cuenta que, dada la naturaleza de las variables, estas pueden ser clasificadas en base a criterios expertos en una serie de grupos donde presentan características similares: compuestos de peroxidación lipídicos, lípidos y microRNAs.

Es evidente que las variables de dichos grupos están relacionadas puesto que presentan una composición química o función biológica similar dependiendo de su naturaleza. Además, las unidades de medida en las que están recogidas las variables de cada tipo son las mismas. Por ello es razonable pensar que cada variable competirá con aquellas cuyas características sean similares a la hora de explicar la variable respuesta. Para conocer la importancia de cada variable se puede considerar una estructura de grupo que recoja la importancia de aquellas variables que están relacionadas. En este caso dicha caracterización viene dada por el experto y las variables se agruparán en los grupos especificados previamente para considerar dichas relaciones y similitudes en la selección de variables.

La selección de variables en presencia de grupos ha sido ampliamente estudiada en la aproximación clásica (Breheny, 2009, Yian y Lin, 2006, **?**). Por otro lado, en el contexto bayesiano se ha trabajado mucho la metodología de selección de variables en modelos de regresión lineal (Bayarri *et al.*, 2012), pero no tanto en modelos lineales generalizados. En este caso, la complejidad del estudio del Alzheimer también reside en el uso de modelos de regresión logística. Junto al desarrollo teórico de la selección de variables bayesiana tradicional, en este trabajo se desarrollará una propuesta para considerar agrupaciones de variables para, posteriormente, compararla con otras metodologías ya estudiadas.

1.2. Avance del trabajo

El trabajo se divide en 6 capítulos. El primero es la introducción al problema que motiva el presente trabajo: el estudio de biomarcadores en sangre para diagnosticar tempranamente la enfermedad de Alzheimer y las problemáticas que pueden comprometer la consecución de dicho objetivo.

En el segundo y tercer capítulos se desarrollará la base teórica de la metodología desarrollada. En el segundo se describirá el modelo genérico del que se parte y se desarrollará la base teórica bayesiana de los problemas selección de variables, sin considerar la estructura de grupos de las variables independientes, que será necesaria para el desarrollo de la metodología posterior. En el

tercero se introducirá el concepto de grupo de variables, junto a los beneficios de dicha consideración a la hora de resolver un problema de selección de variables. Además, se desarrollará tanto la base teórica necesaria para dicha consideración como la propuesta finalmente implementada.

En el cuarto capítulo se desarrollarán los resultados simulados obtenidos para contrastar la eficacia del método propuesto a la hora de seleccionar variables agrupadas en un modelo de regresión logística con datos simulados. Además, se comparará dicha metodología con otras ya existentes, como el método Lasso agrupado.

En el quinto capítulo se trabajará con los datos de pacientes de Alzheimer del hospital La Fe, con el objetivo de seleccionar posibles biomarcadores sanguíneos mediante el método propuesto y desarrollado en capítulos anteriores. Por último, en el sexto capítulo se expondrán las conclusiones del trabajo, realizando una discusión respecto a la metodología estudiada.

Para la realización de este trabajo se ha utilizado el paquete BAS (versión 1.7.1) con el programa estadístico R (versión 4.4.0). El código empleado está en https://github.com/camulro/Seleccion-de-variables-agrupadas.

2. Introducción a la selección de variables desde el punto de vista bayesiano

En este capítulo veremos los fundamentos de la selección de variables desde el punto de vista bayesiano, sin considerar la presencia de grupos. En primer lugar, se describirá el problema del que parte el presente trabajo y se introducirá el concepto del factor Bayes para el cálculo de las distribuciones a posteriori en un problema de regresión lineal. En segundo lugar, se discutirá la selección de las distribuciones previas y se introducirá esta para el caso de los modelos lineales generalizados. Por último, se desarrollarán las distribuciones previas elegidas tanto para el espacio de modelos como el de parámetros.

2.1. Descripción del problema

La selección de variables en el análisis estadístico es la búsqueda del mejor subconjunto de variables explicativas de entre un conjunto de variables potenciales. Cuando se trata de un modelo de regresión donde se tiene un predictor lineal en el que cada variable viene acompañada por un coeficiente, este problema de selección de variables se puede ver como la resolución de un problema de múltiples pruebas donde cada hipótesis contrasta la nulidad de los coeficientes asociados al subconjunto de covariables considerado.

Se han desarrollado distintos métodos tanto en el marco frecuentista como en el bayesiano. En el primero, destacamos el método Lasso (Tibshirani, 1996), el cual minimiza la suma de cuadrados residual sujeto a que la suma del valor absoluto de los coeficientes sea menor que una constante. Otras extensiones con diferentes funciones de penalti son: el método SCAD (Fan y Li, 2001), MCP (Zhang, 2010) y Elastic Net (Zou y Hastie, 2005). En el ámbito bayesiano destacamos la selección implementada en la librería BayesVarSel (Garcia-Donato y Forte, 2018), la cual implementa metodología bayesiana objetiva en modelos de regresión lineal; y la de BAS (Clyde *et al.*, 2011), siendo

esta última una parte esencial de este trabajo y a que retomaremos más adelante.

De forma genérica, al aumentar el número de variables en modelos de regresión mejora el ajuste de los datos, pero puede conllevar a un posible sobreajuste. Sin embargo, al reducir el número de variablesse obtiene un peor ajuste. Además, la introducción de una variable explicativa puede empeorar la confiabilidad del modelo, como sucede cuando las variables están correlacionadas. Por tanto, los métodos de selección de variables persiguen el equilibrio entre la simplicidad del modelo y la bondad del ajuste para los datos observados. El objetivo es conocer la importancia que tiene cada variable a la hora de explicar la variable respuesta.

Nótese que, en este contexto, hablar de selección de variables puede ser equivalente a hablar de selección de modelos, puesto que la distribución de la variable respuesta está fija y los diferentes modelos considerados se diferencian únicamente en el subconjunto de variables explicativas.

Sea $\boldsymbol{Y}$ la variable continua de interés, es decir, la variable respuesta, y $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ las variables explicativas consideradas. Bajo esta premisa, un modelo formado por cualquier subconjunto de dichas variables compite con el resto por explicar la variable respuesta a partir de la evidencia aportada por los datos. Denotaremos el espacio de modelos posibles por $\mathcal{M}$.

Sea $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ el vector de observaciones de la variable respuesta y x_{ij} la i-ésima observación de la variable $\boldsymbol{X}_j$, para $i = 1, \ldots, n$; $j = 1, \ldots, m$. El modelo completo en el problema de regresión logística descrito viene dado por:

$$y_i \sim Ber(\pi_i) \quad \text{con} \quad logit(\pi_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \quad i = 1, \ldots, n.$$

Sea $M \in \mathcal{M}$ un modelo cualquiera del espacio de modelos posibles para el problema anterior. Definimos el vector indicador $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^T$ tal que $\gamma_j = 1$ si la variable $\boldsymbol{X}_j$ se incluye en el modelo y $\gamma_j = 0$ en caso contrario, con $j = 1, \ldots, m$. Por tanto, un modelo cualquiera $M \in \mathcal{M}$ se puede escribir como

$$y_i \sim Ber(\pi_i) \quad \text{con} \quad logit(\pi_i) = \beta_0 + \sum_{j=1}^{m} \gamma_j(\beta_j x_{ij}) \quad i = 1, \ldots, n. \tag{2.1}$$

Los modelos a comparar tendrán siempre el intercepto, por lo que el modelo nulo está siempre anidado en el resto.

Se puede definir una biyección entre el espacio de modelos posibles y el conjunto m-dimensional de ceros y unos, $\{0, 1\}^m$, siendo m el número de variables independientes consideradas en el pro-

blema de selección. Formalmente, definimos dicha función biyectiva como sigue:

$$\begin{array}{rccl} f: & \mathcal{M} & \to & \{0,1\}^m \\ & M & \mapsto & f(M) = \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^T \quad \text{tal que} \quad \gamma_j = \begin{cases} 1 & \text{si } \boldsymbol{X}_j \text{ se incluye en } M, \\ 0 & \text{en caso contrario.} \end{cases} \end{array}$$

De esta manera, es equivalente hablar de un modelo cualquiera $M \in \mathcal{M}$ y de la m-tupla asociada de forma unívoca a dicho modelo a través de f. Además, se tiene que el número de elementos del conjunto $\mathcal{M}$ es el mismo que el de $\{0,1\}^m$, es decir, hay 2^m modelos posibles en el problema de selección considerado. En adelante se denotará por $M_{\boldsymbol{\gamma}}$ al modelo asociado a la m-tupla $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^T$ definido en la Ecuación 2.1.

Se asume que la matriz de diseño del modelo, que denotaremos por $\boldsymbol{X}$, tiene rango completo m y que el espacio de columnas asociado, $C(X)$, no contiene el vector $\mathbf{1}_n$. Además, se asume que el verdadero modelo , M_T, es parte de los 2^m modelos competitivos.

2.2. Factor Bayes y probabilidades a posteriori

Nuestro objetivo es calcular las probabilidades a posteriori para cada modelo competitivo del conjunto de modelos considerados, $\mathcal{M}$. Sin embargo, a menudo el cálculo directo es complejo.Jeffreys (1935) desarrolló una metodología para cuantificar la evidencia aportada por los datos, siendo el factor Bayes la pieza central de dicha teoría y la medida de evidencia en la que se basan las probabilidades a posteriori. Este se define como el cociente de dos verosimilitudes marginales o, más formalmente, como sigue:

Definición 2.2.1. *Dados dos modelos competitivos $M_{\boldsymbol{\gamma}_i}$ y $M_{\boldsymbol{\gamma}_j}$ se define el **factor Bayes**, a favor de $M_{\boldsymbol{\gamma}_i}$ y en contra de $M_{\boldsymbol{\gamma}_j}$, dados los datos $\boldsymbol{y}$ como*

$$B_{\boldsymbol{\gamma}_i \boldsymbol{\gamma}_j}(\boldsymbol{y}) = \frac{m_{\boldsymbol{\gamma}_i}(\boldsymbol{y})}{m_{\boldsymbol{\gamma}_j}(\boldsymbol{y})}, \tag{2.2}$$

siendo

$$m_{\boldsymbol{\gamma}_i}(\boldsymbol{y}) := p(\boldsymbol{y}|M_{\boldsymbol{\gamma}_i}) = \iint p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}_i}, M_{\boldsymbol{\gamma}_i}) \cdot p(\beta_0) \cdot p(\boldsymbol{\beta}_{\boldsymbol{\gamma}_i}|M_{\boldsymbol{\gamma}_i}) d\beta_0 d\boldsymbol{\beta}_{\boldsymbol{\gamma}_i}$$

la verosimilitud marginal del modelo $M_{\boldsymbol{\gamma}_i}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}_i}$ los parámetros específicos del modelo $M_{\boldsymbol{\gamma}_i}$ y β_0 el

coeficiente correspondente al intercepto, común a todos los modelos. En adelante y por simplicidad, se sustituye el subíndice γ_i por su propio subíndice i.

Se deduce de su propia definición (Ecuación 2.2) que el factor Bayes es transitivo, es decir, $B_{ij}(\boldsymbol{y}) = B_{il}(\boldsymbol{y}) \cdot B_{lj}(\boldsymbol{y})$ para cualquier modelo $M_l \in \mathcal{M}$. Por tanto, en lugar de comparar todos los modelos con todos, se comparan todos los modelos con uno fijo, con lo que se reduce el número de cálculos. En nuestro caso, el modelo fijado será el modelo nulo, M_0, puesto que está anidado en todos los modelos considerados. Denotaremos por $B_{i0}(\boldsymbol{y}) = \frac{m_i(\boldsymbol{y})}{m_0(\boldsymbol{y})}$ el factor Bayes del modelo M_i respecto al modelo nulo.

Además, por la forma del Teorema de Bayes, el factor Bayes es el valor que multiplica los odds a priori para convertirlo en odds a posteriori. Más formalmente,

$$\frac{\pi(M_i|\boldsymbol{y})}{\pi(M_0|\boldsymbol{y})} = \frac{m_i(\boldsymbol{y})}{m_0(\boldsymbol{y})} \cdot \frac{\pi(M_i)}{\pi(M_0)}, \tag{2.3}$$

donde $\pi(M_i)$ es la probabilidad previa y $\pi(M_i|\boldsymbol{y})$, la probabilidad a posteriori del modelo M_i en el espacio de probabilidad asociado al conjunto de posibles modelos.

En otras palabras, el factor Bayes cuantifica la evidencia que hay en los datos a favor del modelo M_i y en contra del nulo. Se deduce que, si $B_{i0}(\boldsymbol{y}) > 1$, entonces hay evidencia a favor del modelo M_i y en contra del nulo. La guía para interpretar el factor Bayes y las probabilidades a posteriori de Kass y Raftery (1995) se muestra en la Tabla 2.1. Cabe destacar que la utilidad de la guía disminuye a medida que el número de modelos a comparar aumenta.

Tabla 2.1: Interpretación del factor Bayes de Kass y Raftery (1995).

$B_{i0}(\boldsymbol{y})$	Probabilidad[1]	Evidencia en contra de M_0
De 1 a 3	De 0.5 a 0.75	No merece más que una mención
De 3 a 20	De 0.75 a 0.95	Sustancial
De 20 a 150	De 0.95 a 0.99	Fuerte
Mayor que 150	Mayor que 0.99	Decisiva

Por otro lado, haciendo uso del Teorema de Bayes la probabilidad a posteriori de un modelo $M_j \in \mathcal{M}$, $\pi(M_j|\boldsymbol{y})$, se puede expresar como

$$\pi(M_j|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M_j) \cdot \pi(M_j)}{\sum_{i:M_i \in \mathcal{M}} p(\boldsymbol{y}|M_i) \cdot \pi(M_i)}.$$

[1]La columna de probabilidad se obtiene asumiendo que, a priori, ambas hipótesis son equiprobables.

Dividiendo numerador y denominador en la ecuación anterior por la verosimilitud marginal del modelo nulo, obtenemos la probabilidad a posteriori del modelo M_j expresada en términos del factor Bayes:

$$\pi(M_j|\boldsymbol{y}) = \frac{B_{j0}(\boldsymbol{y}) \cdot \pi(M_j)}{\sum_{i:M_i \in \mathcal{M}} B_{i0}(\boldsymbol{y}) \cdot \pi(M_i)}. \tag{2.4}$$

Teniendo en cuenta que, para calcular de la distribución a posteriori con la Ecuación 2.4, es necesario tanto el cálculo del factor Bayes como el de la probabilidad sobre el espacio de modelos, se deben definir las probabilidades a priori sobre el espacio de modelos y las probabilidades a priori necesarias para calcular la distribución marginal, y con ella los factores Bayes.

2.3. Elección de las distribuciones previas

Una particularidad de la selección de modelos bayesiana es que los resultados son altamente sensibles a la elección de las distribuciones previas. Por otro lado, en nuestro problema se tienen muchos parámetros, con muchas combinaciones posibles, y un número muy elevado de modelos sobre los que establecer distribuciones a priori. Por tanto, es prácticamente imposible conseguir información experta para todas estas consideraciones. Para simplificar el análisis se utilizarán las llamadas distribuciones **previas por defecto**, las cuales pueden establecerse sin emplear información a priori y, además, pretenden ser objetivas para no introducir ninguna información extra al estudio que se está realizando (Berger, 2006). La elección de dichas previas no es única y se realizará una discusión al respecto.

2.3.1. Previas sobre el espacio de modelos

Una aproximación inicial podría ser la consideración de que todos los modelos sean equiprobables a priori, es decir, la asignación de una distribución previa constante a todos los modelos como hicieron Fernández *et al.* (2002):

$$\pi(M_i) = \frac{1}{2^m}.$$

Esto es equivalente a asignar a cada variable una probabilidad a priori de 0.5 de incluirse en el modelo verdadero. Sin embargo, un aspecto importante a tener en cuenta es que la selección de modelos se ve afectada por problemas de multiplicidad. Esto es debido a que la probabilidad de que un modelo muestre evidencia falsa se incrementa considerablemente al aumentar el número de comparaciones, puesto que aumenta el número de hipótesis consideradas a la vez, lo que puede derivar en efectos

espurios. En el caso del modelo de regresión lineal, Scott y Berger (2010) comprobaron que la multiplicidad debe ser controlada a través de las probabilidades previas de los modelos y que la previa constante no proporciona dicho control.

Una forma estándar es tratar la inclusión de variables como una sucesión de intentos intercambiables de una distribución Bernoulli con probabilidad de éxito común ρ, es decir:

$$\pi(M_i|\rho) = \rho^{m_i} \cdot (1-\rho)^{m-m_i},$$

donde $m_i = ||\boldsymbol{\gamma}_i||_2^2$ (número de variables incluidas en el modelo M_i), m el número de variables explicativas consideradas en el problema de selección y $\rho \in (0,1)$. En Scott y Berger (2010) se propone tomar previas inversamente proporcionales al número de modelos de una dimensión dada. Este tipo de distribuciones previas son las que se desarrollarán posteriormente para el caso de selección de variables en presencia de grupos.

2.3.2. Previas sobre el espacio de parámetros

Dadas las características de nuestro problema, tenemos un parámetro común, β_0, y una serie de parámetros específicos para cada modelo, β_i. La distribución previa para los parámetros de un modelo M_i se puede escribir como

$$p(\beta_0, \boldsymbol{\beta}_i) = p(\boldsymbol{\beta}_i|\beta_0) \cdot p(\beta_0). \tag{2.5}$$

Las distribuciones previas por defecto impropias, como la previa mínimo informativa $p(\beta_0) \propto 1$, pueden ser utilizadas para el parámetro común debido a que $p(\beta_0)$ aparece tanto en el numerador como en el denominador del factor Bayes. Sin embargo, para los parámetros específicos de cada modelo no se pueden asignar previas impropias ni vagas, puesto que el cálculo del factor Bayes resultante es inviable. Bayarri *et al.* (2012) desarrollaron una serie de criterios que toda distribución previa objetiva debe cumplir para obtener factores Bayes de forma cerrada[2], al realizar una selección de variables bayesiana. El cumplimiento de estos criterios se garantiza gracias a una serie de propiedades que cumplen las ***g-priors*** (Zellner, 1986), ampliamente estudiadas en la literatura.

[2]Una expresión se dice que tiene **forma cerrada** si está formada únicamente por constantes, variables y un conjunto finito de funciones básicas conectadas mediante operaciones aritméticas y composiciones de funciones. Esta propiedad es de interés en el cálculo del factor Bayes puesto que reduce significativamente la complejidad del calculo y el tiempo de computación del mismo.

Tabla 2.2: Propuestas de $g\text{-}priors$ más relevantes.

g fijo	
g-Prior	Propuesta
Previa de la información de la unidad	$g = n$
Previa del criterio del riesgo de inflación	$g = m^2$
Previa de Hannan-Quinn	$g = log(n)$
g como hiperparámetro	
Hyper-g-Prior	Propuesta
Previa de Cauchy	$g \sim IGa(\frac{1}{2}, \frac{n}{2})$
Hyper-g previa	$g\|a \sim p(g) \propto (1+g)^{\frac{-a}{2}}$
Hyper-$\frac{g}{n}$ previa	$g\|a \sim p(g) \propto (1+\frac{g}{n})^{\frac{-a}{2}}$
Previa robusta	$g\|a \sim p(g) \propto (1+g)^{\frac{-3}{2}}$

g-Priors en modelos de regresión lineal

Sea $\boldsymbol{X}_i$ la matriz de diseño del modelo M_i centrada respecto de la media. La distribución de las *g-priors* en modelos de regresión lineal se define como

$$\boldsymbol{\beta}_i|\boldsymbol{\gamma}_i, \sigma, g \sim N(\mathbf{0}, g \cdot \sigma^2(\boldsymbol{X}_i^T \boldsymbol{X}_i)^{-1}),$$

donde el parámetro g modifica la matriz de varianzas-covarianzas de los coeficientes $\boldsymbol{\beta}_i$.

La elección de g es clave (Liang y Berger, 2008) y las propuestas son mucha, algunas de las cuales se discutieron en Liang y Berger (2008). Se puede mantener g constante o considerarlo un hiperparámetro. Esta segunda opción conduce a una mezcla de $g\text{-}priors$ para los coeficientes $\boldsymbol{\beta}_i$, lo que produce una inferencia más robusta. Además, permite verificar los criterios ya mencionados de Bayarri *et al.* (2012) y admite la consideración de aquellas variables cuyo impacto en el modelo es menor. Las $g\text{-}priors$ más estudiadas en modelos de regresión lineal quedan recogidas en la Tabla 2.2.

En modelos de regresión lineal se sabe que las *g-priors* dan lugar a verosimilitudes marginales de forma cerrada. Como ya se mencionó, dicha propiedad es de interés puesto que permite que el tiempo de computación del cálculo de las probabilidades a posteriori sea menor y la búsqueda del modelo sea eficiente. Para modelos lineales generalizados las previas normales no dan lugar a distribuciones conjugadas, es decir, las distribuciones a posteriori resultantes no pertenecen a la misma familia que las distribuciones previas; y no pueden, por tanto, obtenerse en forma cerrada. Se ha visto que

pueden utilizarse aproximaciones de Laplace de la verosimilitud y, asignando una previa uniforme al parámetro común y previas normales (como las *g-priors*) a los parámetros específicos del modelo, se consigue eficiencia computacional (Li y Clyde, 2018). Esto está programado en la librería BAS, pero no se entrará en más detalle puesto que lo único que necesitamos es calcular las distribuciones previas de los parámetros.

g-Priors en modelos lineales generalizados

En Li y Clyde (2018) se analizan diferentes *g-priors* propuestas para modelos lineales generalizados y, para un modelo M_i, se propone

$$\boldsymbol{\beta}_i|\boldsymbol{\gamma}_i, g \sim N(\mathbf{0}, g \cdot \mathcal{J}_n(\hat{\boldsymbol{\beta}}_i)^{-1}), \tag{2.6}$$

donde $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_i)$ es la matriz Hessiana negativa de la log-verosimilitud y $\hat{\boldsymbol{\beta}}_i$ es el estimador máximo verosímil de los coeficientes $\boldsymbol{\beta}_i$.

Dicha formulación presenta beneficios en la aproximación del factor Bayes. Para ello es necesario reparametrizar la matriz de diseño del modelo, $\boldsymbol{X}_i$, a $(\boldsymbol{I}_n - \mathcal{P}_{\boldsymbol{I}_n})\boldsymbol{X}_i$ y, con una serie de consideraciones adicionales, se tiene que

$$\begin{aligned} \boldsymbol{\beta}_i|\boldsymbol{y}, i, g \quad &\xrightarrow{D} \quad N\Big(\frac{g}{1+g} \cdot \hat{\boldsymbol{\beta}}_i, \frac{g}{1+g} \cdot \mathcal{J}_n(\hat{\boldsymbol{\beta}}_i)^{-1}\Big), \\ \beta_0|\boldsymbol{y}, i \quad &\xrightarrow{D} \quad N\Big(\hat{\beta}_0, \mathcal{J}_n(\hat{\beta}_0)^{-1}\Big); \end{aligned}$$

donde $\xrightarrow{D}$ denota convergencia en la función de distribución y $\hat{\beta}_0$ y $\hat{\boldsymbol{\beta}}_i$ son los estimadores máximo verosímiles de β_0 y $\boldsymbol{\beta}_i$, respectivamente. Bajo esta premisa, se tiene la siguiente distribución marginal del modelo M_i:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\gamma}_i, g) &= \int p(\boldsymbol{y}|\boldsymbol{\beta}_i, \boldsymbol{\gamma}_i) \cdot p(\boldsymbol{\beta}_i|\boldsymbol{\gamma}_i, g) \partial \boldsymbol{\beta}_i \\ &\propto p(\boldsymbol{y}|\hat{\beta}_0, \hat{\boldsymbol{\beta}}_i, \boldsymbol{\gamma}_i) \cdot \mathcal{J}_n(\hat{\beta}_0)^{-\frac{1}{2}} \cdot (1+g)^{\frac{-p_i}{2}} \cdot e^{-\frac{Q_i}{2(1+g)}}, \end{aligned}$$

siendo p_i el rango de las columnas de la matriz de diseño del modelo, $\boldsymbol{X}_i$, y $Q_i = \hat{\boldsymbol{\beta}}_i^T \cdot \mathcal{J}_n(\hat{\boldsymbol{\beta}}_i) \cdot \hat{\boldsymbol{\beta}}_i$ el estadístico de Wald bajo la información observada.

Además, podemos escribir el factor Bayes correspondiente como sigue:

$$B_{i0} = \frac{m_i(\boldsymbol{y})}{m_0(\boldsymbol{y})} = e^{\frac{z_i}{2}} \cdot \sqrt{\frac{\mathcal{J}_n(\hat{\beta}_0)}{\mathcal{J}_n(\hat{\beta}_i)}} \cdot (1+g)^{\frac{-p_i}{2}} \cdot e^{-\frac{Q_i}{2(1+g)}} \tag{2.7}$$

siendo

$$z_i = 2log\left(\frac{p(\boldsymbol{y}|\hat{\beta}_0, \hat{\boldsymbol{\beta}}_i, \boldsymbol{\gamma}_i)}{p(\boldsymbol{y}|\hat{\beta}_0, \boldsymbol{\gamma}_i = \mathbf{0})}\right)$$

el estadístico del cociente de verosimilitudes para los modelos M_i y M_0, también llamado **estadístico *deviance***. Así, el factor Bayes resultante proporciona un ajuste del test del cociente de verosimilitudes con una función de penalti que depende de g y del estadístico de Wald.

Cabe plantearse la posibilidad de que los estimadores máximo verosímiles de los parámetros del modelo, $\hat{\beta}_0$ y $\hat{\boldsymbol{\beta}}_i$, no existan. Esto podría deberse a una separación de los datos en regresión binaria o a que las matrices de diseño de los modelos sean singulares, es decir, no tengan rango completo. Sin embargo, esta no es una característica de nuestro problema, por lo que no se profundizará en ello. Para más información, consultar Li y Clyde (2018).

La elección para la consecución del estudio fue de una $hyper\text{-}g\text{-}prior$, la **previa robusta**, introducida en Bayarri *et al.* (2012) y cuya expresión en modelos lineales generalizados se desarrollará en el siguiente capítulo.

2.4. Probabilidades de inclusión a posteriori

El objetivo final de la selección de variables bayesiana es calcular las **probabilidades de inclusión a posteriori** de las variables consideradas para conocer la importancia de cada una a la hora de explicar la variable respuesta Podemos expresar la probabilidades de inclusión de una variable $\boldsymbol{X}_i$ como sigue:

$$p(\gamma_i = 1|\boldsymbol{y}) = \sum_{\boldsymbol{\gamma}:\, \gamma_i = 1} \pi(M_{\boldsymbol{\gamma}}|\boldsymbol{y}). \tag{2.8}$$

Para ello, sería necesario calcular las distribuciones a posteriori de cada modelo $M_i \in \mathcal{M}$. No obstante, esto no siempre será posible cuando el número de variables consideradas sea elevado. Esta situación será estudiada en el capítulo siguiente.

Por último, una vez se tienen las probabilidades de inclusión de cada variable, queda establecer qué criterio seguir a la hora de elegir una variable en base a dicha probabilidad.

2.5. *Median posterior probability model* y *High posterior probability model*

Una vez elegidas las distribuciones a priori del espacio de modelos y del espacio de parámetros, calculadas las distribuciones a posteriori de cada modelo y obtenidas las probabilidades de inclusión de las variables mediante la Ecuación 2.8, queda establecer una regla de decisión para seleccionar las variables explicativas en función de las probabilidades de inclusión a posteriori obtenidas. En el marco bayesiano, se tienen dos criterios de inclusión de variables:

High posterior probability model (HPM) o modelo de la mayor probabilidad a posteriori: como su nombre indica, el modelo elegido es aquel que se obtiene con la mayor probabilidad a posteriori.

Median posterior probability model (MPM) o modelo de probabilidad mediana a posteriori: Una variable singular se incluye en el modelo si su probabilidad de inclusión a posteriori es mayor que 0.5 (Barbieri y Berger, 2004).

En un principio, se vio que el MPM coincide con el HPM en diseños ortogonales, correlacionados y anidados. Sin embargo, posteriormente se vio en Barbieri *et al.* (2021) que en determinadas situaciones bajo diseños correlacionados también tiene un buen comportamiento. Por tanto, será el MPM el principal criterio de selección que se seguirá en adelante.

Para considerar la presencia de grupos en el problema de selección de variables se utilizará lo visto a lo largo de este capítulo junto a la metodología que se desarrollará a continuación.

3. Selección de variables agrupadas

En este capítulo veremos cómo considerar la estructura de grupos en la selección de variables desde el punto de vista bayesiano. En primer lugar, se definirá el concepto de grupo de variables y se analizarán los posibles beneficios de dicha consideración en el análisis estadístico. En segundo lugar, se verán distintas metodologías para realizar dicho acercamiento y se desarrollará nuestra propuesta. Por último, se establecerá el algoritmo de *Gibbs Sampling* para realizar las simulaciones del espacio de modelos.

Cabe destacar que este enfoque es un área aún en estudio y, a diferencia de los resultados vistos anteriormente, la metodología no está propiamente establecida, sino que se basa en el trabajo preliminar de García-Donato y Paulo (2021) para el estudio de factores (variables categóricas).

3.1. ¿Qué es un grupo de variables?

La definición del concepto de grupo de variables es intrínsecamente amplia y, en la práctica, está sujeta a incertidumbre.

Definición 3.1.1. *De forma genérica, se dice que un conjunto de variables* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_l$ *con* $l \geq 2$ *es un* ***grupo de variables****, y se denotará por* $\mathcal{G}$*, si dichas variables están interconectadas o relacionadas en un contexto determinado.*

La consideración del grupo como conjunto de variables relacionadas aporta un conocimiento mayor de los datos mediante la identificación de asociaciones o patrones. En el caso concreto del alzheimer que motiva este trabajo, la agrupación de las variables viene dada por el juicio experto del personal sanitario. Por ejemplo, se espera que los niveles de los diferentes compuestos lipídicos obtenidos en un análisis sanguíneo estén relacionados.

La identificación de un grupo de variables puede no siempre provenir del conocimiento propio o de un juicio experto y las fuentes pueden ser muy diversas. De entre las que definen una metodología

más sistemática cabe destacar:

- Correlación entre variables: cuantifica la fuerza y dirección de una posible relación lineal entre dos variables. Hay cierto grado de arbitrariedad en la selección del umbral a partir del cual se define un grupo.
- PCA: realiza combinaciones lineales de las variables originales que explican una parte de la variabilidad original. Permite identificar grupos de variables relacionadas.
- Análisis de agrupamiento: *clusters* de individuos o variables construidos en función de su similaridad. Permite identificar grupos de variables que comparten características.

Se ha visto cómo de vasta puede ser la identificación de variables agrupadas. Sin embargo, en muchas ocasiones se suele asociar simplemente los conceptos de grupo de variables y correlación entre variables debido a la carencia de otras fuentes de información o de un sentido crítico experto sobre el tema. Cabe destacar que dos variables que hacen referencia al mismo concepto estarán muy probablemente relacionadas, pero dos variables relacionadas pueden no hacer referencia al mismo concepto.

3.1.1. Marco teórico de la selección de variables en presencia de grupos

Cuando hay muchas variables que intentan explicar la misma información sobre el proceso de interés, se pierde parte de dicha información porque compiten entre ellas por entrar en el modelo. Identificar una serie de variables relacionadas permite mejorar la eficiencia del modelo, lidiando con posible colinealidad entre variables y reduciendo dimensionalidad.

En Huang y Zhang (2010) se estudió el funcionamiento del método de selección de variables Lasso tradicional (Tibshirani, 1996) cuando hay grupos de coeficientes que tienden a 0 al mismo tiempo. Se vio que el método Lasso no funciona correctamente en esta situación, la cual se da con frecuencia cuando hay variables que presentan estructura de grupo. Ellos proponen una extensión llamada **Lasso agrupado**, que emplea el penalti L2 (Yian y Lin, 2006).

Otros método derivados del Lasso que se crearon para seleccionar variables agrupadas son el **MCP agrupado**, que emplea el penalti mínimo cóncavo (Breheny, 2009); y el **SCAD agrupado**, que emplea el penalti de desviación absoluta recortada sin problemas (Fan y Li, 2001). Estos serán los que se utilizarán más adelante para comparar con la metodología propuesta. Otra extensión que funciona mejor cuando hay variables altamente correlacionadas es la de Hastie *et al.* (2015), llamado ***Elastic Net*.**

Un caso particular de grupo de variables es el que se produce al introducir variables tipo factor a

un modelo de regresión, donde destacamos el trabajo de García-Donato y Paulo (2021) para modelos lineales. En este se desarrolló un método de selección de variables en presencia de factores, basada en la selección de variables bayesiana vista en el Capítulo 2, que asigna distribuciones previas sobre el espacio de modelos de forma jerárquica, en dos niveles. Esto permite, entre otras cosas, controlar los problemas de multiplicidad a causa del número de predictores.

Existen otras metodologías desarrolladas para lidiar con estos problemas pero nuestra propuesta está basada en la selección de variables bayesiana, tal y como la hemos estudiado en este trabajo. En el ámbito bayesiano, cuando hay un grupo de variables que compiten por entrar en el modelo porque explican la misma información de la variable respuesta, las probabilidades de inclusión respectivas se verán muy reducidas por esto, hecho que puede derivar en no introducir ninguna en el modelo. El objetivo de considerar la estructura de grupo reside en tratar dichas variables relacionadas en conjunto, de forma que ya no compitan entre ellas sino que se refuercen unas a otras para hacer su influencia más clara, reflejada en la probabilidad a posteriori del grupo.

3.2. Propuesta

Como se mencionó en el Capítulo 2, nuestro problema es uno de regresión logística en el que la variable respuesta, $\boldsymbol{Y}$, es binaria y el conjunto de variables explicativas está formado por k variables numéricas[1], $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$; y p grupos de variables, $\mathcal{G}_1, \ldots, \mathcal{G}_p$. Cada grupo $\mathcal{G}_j$ tiene $l_j \geq 2$ variables numéricas, para $j = 1, \ldots, p$. Sea $L = \sum_{j=1}^{p} l_j$, el número total de variables agrupadas.

Sea $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ la variable respuesta, con $Y_i \in \{0, 1\}$. Así $\boldsymbol{Y} \sim Ber(\boldsymbol{\pi})$ con $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^T$, probabilidad de éxito de la variable $\boldsymbol{Y}$. El modelo de regresión logística completo es

$$logit(\boldsymbol{\pi}) = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\beta}, \tag{3.1}$$

siendo $\boldsymbol{X}$ la matriz de n filas y k columnas, donde la fila i-ésima está formada por los valores de las variables singulares para la observación i-ésima, con $i = 1, \ldots, k$; y $\boldsymbol{Z} = [\boldsymbol{Z}^1 | \ldots | \boldsymbol{Z}^j | \ldots | \boldsymbol{Z}^p]$ es la matriz n filas y L columnas, donde $\boldsymbol{Z}^j = (z_{ir}^{(j)})$ es la submatriz de n filas y l_j columnas tal que $z_{ir}^{(j)}$ es la observación i-ésima de la variable r-ésima del grupo $\mathcal{G}_j$, para $r = 1, \ldots, l_j$ y $j = 1, \ldots, p$. Denotamos por $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)$ con $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jl_j})^T$, $j = 1, \ldots, p$; y $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ los vectores paramétricos asociados a las variables agrupadas y a las singulares respectivamente.

[1]En este trabajo nos centraremos en una selección considerando únicamente variables numéricas. Sin embargo, la extensión con variables categóricas sería sencilla siguiendo García-Donato y Paulo (2021).

A diferencia de la selección de variables en presencia de factores, la matriz de diseño cuando se consideran grupos de variables, $[\mathbf{1}_n|\boldsymbol{X}|\boldsymbol{Z}]$, sí es de rango completo. Esto es debido a que, tanto las variables singulares como las agrupadas, son numéricas y el espacio de columnas $C([\boldsymbol{X}|\boldsymbol{Z}])$ no contiene la columna $\mathbf{1}_n$. De esta manera, tenemos un total de 2^{k+L} modelos competitivos, cuyo conjunto denotaremos por $\mathcal{M}$, y asumimos que el verdadero modelo que genera los datos, M_T, forma parte de los modelos considerados.

Para continuar con la selección de variables es necesario definir cuándo se considera que una variable y un grupo de variables está activo en un modelo concreto.

Definición 3.2.1. *Se considera que una **variable** X_i está **activa** en un modelo $M \in \mathcal{M}$ si dicha variable se incluye en el modelo. Se considera que un **grupo** de variables $\mathcal{G} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_l\}$ está **activo** en un modelo $M \in \mathcal{M}$ si al menos una variable $\boldsymbol{Z}_r$ se incluye en el modelo, con $r \in \{1, \ldots, l\}$.*

Es decir, una variable $\boldsymbol{X}_i$ está activa en un modelo $M \in \mathcal{M}$ si $\alpha_i \neq 0$, para $i = 1, \ldots, k$. Un grupo $\mathcal{G}_j$ está activo en un modelo $M \in \mathcal{M}$ si existe $\beta_{jr} \neq 0$, para $r = 1, \ldots, l_j$, $j = 1, \ldots, p$.

Nótese que, a diferencia del caso de factores donde un grupo formado por los niveles del factor es o bien elegido por completo o bien no elegido, en este caso se tiene una selección jerárquica de dos niveles. Así, se quiere que las variables de un mismo grupo ya no compitan por explicar la variable respuesta, sino que sumen unas a otras para conocer la importancia del grupo en sí mismo. No obstante, con datos reales hay cierta incertidumbre sobre la propia definición de los grupos y estos también pueden verse penalizados al introducir variables poco relevantes a la hora de explicar la variable respuesta.

3.2.1. Metodología

El espacio de modelos posibles, $\mathcal{M}$, está formado por todas las posibles permutaciones de las columnas de la matriz de diseño $[\boldsymbol{X}|\boldsymbol{Z}]$ del modelo de la Ecuación 5.1.

Sea $\boldsymbol{\gamma}^T = (\gamma_1, \ldots, \gamma_k)$, $\boldsymbol{\delta}^T = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_p^T)$ y $\boldsymbol{\delta}_j^T = (\delta_{j1}, \ldots, \delta_{jl_j})$ tales que $\gamma_i = 1$ si la variable $\boldsymbol{X}_i$ está en el modelo, $\gamma_i = 0$ en caso contrario; y $\delta_{jr} = 1$ si la variable $z_r^{(j)}$ del grupo $\mathcal{G}_j$ está en el modelo, $\delta_{jr} = 0$ en caso contrario.

En adelante, identificaremos por $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \{0, 1\}^{k+L}$ al modelo $M_{(\boldsymbol{\gamma},\boldsymbol{\delta})} \in \mathcal{M}$, aquel cuya matriz de diseño sea $[\boldsymbol{X}_{\boldsymbol{\gamma}}|\boldsymbol{Z}_{\boldsymbol{\delta}}]$, matriz resultante de seleccionar las columnas de $\boldsymbol{X}$ y $\boldsymbol{Z}$ asociadas a las variables cuyos coeficientes γ_i y δ_{jr} sean distintos de 0 respectivamente, para $i = 1, \ldots, k$, $r = 1, \ldots, l_j$

y $j = 1, \ldots, p$. Denotamos por $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ y $\boldsymbol{\beta}_{\boldsymbol{\delta}}$ los coeficientes del predictor lineal correspondientes a las matrices $\boldsymbol{X}_{\boldsymbol{\gamma}}$ y $\boldsymbol{Z}_{\boldsymbol{\delta}}$, respectivamente.

Se tienen variables indicadoras para las variables singulares y agrupadas, $\boldsymbol{\gamma}$ y $\boldsymbol{\delta}$ respectivamente. Para construir la estructura de grupos deseada es necesario introducir una función que identifique qué grupos están activos.

Definición 3.2.2. *Sea $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \{0,1\}^{k+L}$ un modelo concreto con k variables singulares y L variables agrupadas. Definimos $\boldsymbol{\tau} = \boldsymbol{\tau}(\boldsymbol{\delta}) = (\tau_1, \ldots, \tau_p)$ con $\tau_j = \tau_j(\boldsymbol{\delta}_j) = 1$ si $\mathbf{1}^T\boldsymbol{\delta}_j \geq 1$ y $\tau_j = \tau_j(\boldsymbol{\delta}_j) = 0$ en caso contrario.*

Siguiendo con lo visto en la Sección 2.4, el objetivo final es calcular las probabilidades de inclusión para cada variable singular, grupo y variable agrupada, para conocer la importancia de cada uno a la hora de explicar la variable respuesta. Podemos expresar dicha probabilidad de inclusión, cuando consideramos grupos de variables, como sigue:

$$\begin{aligned} p(\gamma_i = 1|\boldsymbol{y}) &= \sum_{(\boldsymbol{\gamma},\boldsymbol{\delta}):\ \gamma_i=1} \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y}), \\ p(\tau_j = 1|\boldsymbol{y}) &= \sum_{(\boldsymbol{\gamma},\boldsymbol{\delta}):\ \tau_j=1} \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y}), \\ p(\delta_{jr} = 1|\boldsymbol{y}, \tau_j = 1) &= \sum_{(\boldsymbol{\gamma},\boldsymbol{\delta}):\ \delta_{jr}=1} \frac{\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y})}{p(\tau_j = 1|\boldsymbol{y})}, \end{aligned} \tag{3.2}$$

para una variable singular $\boldsymbol{X}_i$, grupo $\mathcal{G}_j$ y variable agrupada δ_{jr}.

Para ello, será necesario calcular las distribuciones a posteriori de cada modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \{0,1\}^{k+L}$. Esto no siempre será cierto, como ya se mencionó, cuando el número de variables consideradas sea elevado; y será estudiado más adelante en la Sección 3.2.2.

Distribuciones a posteriori

Como se hizo en el Capítulo 2, pero considerando ahora la presencia de grupos, se pueden expresar las probabilidades a posteriori del modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ como

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y}) \propto m_{(\boldsymbol{\gamma},\boldsymbol{\delta})}(\boldsymbol{y}) \cdot \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}), \tag{3.3}$$

siendo

$$m_{(\boldsymbol{\gamma},\boldsymbol{\delta})}(\boldsymbol{y}) = \iiint p(\boldsymbol{y}|\beta_0, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\delta}}) \cdot p(\beta_0, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\delta}}) d\beta_0 d\boldsymbol{\alpha}_{\boldsymbol{\gamma}} d\boldsymbol{\beta}_{\boldsymbol{\delta}}$$

y

$$m_{(\mathbf{0},\mathbf{0})}(\boldsymbol{y}) = \iiint p(\boldsymbol{y}|\beta_0) \cdot p(\beta_0) d\beta_0$$

las verosimilitudes marginales del modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ y el modelo nulo, respectivamente.

Se puede reescribir la Ecuación 3.3 en términos del factor Bayes del modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ respecto al modelo nulo, $B_{(\boldsymbol{\gamma},\boldsymbol{\delta})\mathbf{0}}$, como sigue:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y}) \propto B_{(\boldsymbol{\gamma},\boldsymbol{\delta})\mathbf{0}} \cdot \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) \quad \text{siendo} \quad B_{(\boldsymbol{\gamma},\boldsymbol{\delta})0} = \frac{m_{(\boldsymbol{\gamma},\boldsymbol{\delta})}(\boldsymbol{y})}{m_{(\mathbf{0},\mathbf{0})}(\boldsymbol{y})}. \tag{3.4}$$

El siguiente paso es elegir la distribución previa de los parámetros para calcular el factor Bayes correspondiente y obtener la distribución a posteriori de un modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \{0,1\}^{k+L}$.

Distribuciones previas de los parámetros: $p(\beta_0, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\delta}})$

La distribución previa de los parámetros específicos del modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ se puede escribir como $p(\beta_0, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\delta}}) = p(\boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\delta}}|\beta_0) \cdot p(\beta_0)$, con lo que el término $p(\beta_0)$ aparece tanto en el numerador como en el denominador del factor Bayes. Este hecho permite asignar distribuciones previas por defecto impropias al parámetro común. Sin embargo, para los parámetros específicos no se pueden asignar distribuciones vagas ni impropias y la elección más extendida es la de las *g-priors*, como ya se mencionó en el capítulo anterior.

Para el parámetro común a todos los modelos, β_0, se asignará una distribución previa uniforme. Para los parámetros específicos de cada modelo, $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ y $\boldsymbol{\beta}_{\boldsymbol{\delta}}$, se utilizará como ya se mencionó, la **previa robusta**, puesto que no requiere un excesivo tiempo de computación, presenta garantía teóricas y el factor Bayes correspondiente se puede obtener en forma cerrada.

En el caso de las $hyper\text{-}g\text{-}priors$, el parámetro g entra en la distribución a posteriori de $\boldsymbol{\beta}_i$ y la verosimilitud marginal a través del factor reductor, $\frac{g}{1+g}$, llamado así por su papel en la Ecuación 2.3.2; o del complementario, $u = \frac{1}{1+g}$, cuya expresión para la previa robusta con la notación de grupos de variables es:

$$p_r(u) = a_r \cdot \big(\rho_r(b_r + n)\big)^{a_r} \cdot \frac{u^{a_r - 1}}{\big(1 + (b_r - 1)u\big)^{a_r + 1}} \cdot \mathbf{1}_{\{0 < u < (\rho_r(b_r+n)+(1-b_r))^{-1}\}},$$

donde $a_r > 0$, $b_r > 0$ y $\rho_r \geq \frac{b_r}{b_r+n}$. Se recomienda tomar $a_r = 0.5$, $b_r = 1$ y $\rho_r = \frac{1}{1+p_{\boldsymbol{\gamma}}}$ para cumplir los criterios de selección de previas de Bayarri *et al.* (2012).

Bajo estas presunciones, el factor reductor complementario se distribuye como una Gamma truncada,

$$u \sim TG_{\left(0,\frac{p_{(\boldsymbol{\gamma},\boldsymbol{\delta})}+1}{n+1}\right)}\left(\frac{1}{2},0\right) \quad \Rightarrow \quad u|\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{\delta} \quad \xrightarrow{D} \quad TG_{\left(0,\frac{p_{(\boldsymbol{\gamma},\boldsymbol{\delta})}+1}{n+1}\right)}\left(\frac{p_{\boldsymbol{\gamma}}+1}{2},\frac{Q_{(\boldsymbol{\gamma},\boldsymbol{\delta})}}{2}\right),$$

y el factor Bayes correspondiente es

$$B_{(\boldsymbol{\gamma},\boldsymbol{\delta})0} = \left[\frac{\mathcal{J}_n(\hat{\beta}_0)}{\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{(\boldsymbol{\gamma},\boldsymbol{\delta})})}\right] \cdot v^{-\frac{p_{(\boldsymbol{\gamma},\boldsymbol{\delta})}}{2}} \cdot e^{\frac{z_{(\boldsymbol{\gamma},\boldsymbol{\delta})}}{2}-\frac{Q_{(\boldsymbol{\gamma},\boldsymbol{\delta})}}{2v}} \cdot \frac{Beta\left(\frac{a+p_{(\boldsymbol{\gamma},\boldsymbol{\delta})}}{2},\frac{b}{2}\right)\cdot\Phi_1\left(\frac{b}{2},r,\frac{a+b+p_{(\boldsymbol{\gamma},\boldsymbol{\delta})}}{2},\frac{s+Q_{(\boldsymbol{\gamma},\boldsymbol{\delta})}}{2v},1-\kappa\right)}{Beta\left(\frac{a}{2},\frac{b}{2}\right)\cdot\Phi_1\left(\frac{b}{2},r,\frac{a+b}{2},\frac{s}{2v},1-\kappa\right)},$$

donde $v \geq 1$, $r, s \in \mathbb{R}$, $\kappa > 0$, $\Phi_1(\alpha,\beta,\gamma,x,y)$ es la función hipergeométrica cofluente de dos variables y $(\alpha)_n = 1$, si $n = 0$, y $(\alpha)_n = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}$, si $n \in \mathbb{N}$, es el coeficiente de Pochammer.

Esta metodología ha sido implementada en la librería BAS mediante la función $bas.glm$, la cual realiza una reparametrización de la matriz de diseño y la aproximación de Laplace, para que las *g-prior* den lugar a eficiencia computacional en el cálculo de las distribuciones a posteriori en modelos lineales generalizados (Li y Clyde, 2018). Nuestro interés en ella reside únicamente en el cálculo de las marginales del factor Bayes, por lo que no se profundizará más en la metodología.

Distribuciones previas de los modelos: $\pi(\boldsymbol{\gamma},\boldsymbol{\delta})$

La parte fundamental de este trabajo es la selección de las previas sobre el espacio de modelos, puesto que la propuesta reside en asignar dichas previas de forma jerárquica, en dos niveles, de forma similar a como hicieron García-Donato y Paulo (2021) con factores. Veamos en primer lugar como sería una propuesta más tradicional, con la que se comparará más adelante.

Sin considerar la presencia de grupos, y siguiendo con la notación empleada hasta ahora, tendríamos $\boldsymbol{\gamma}$ y $\boldsymbol{\delta}$ como indicadores de las posibles variables en el modelo, es decir, sin considerar $\boldsymbol{\tau}$, variable indicadora de los grupos. Como se hizo en el Capítulo 2, habría dos posibilidades para la distribución previa de los modelos. Una primera aproximación sería la **previa constante**:

$$\pi(\boldsymbol{\gamma},\boldsymbol{\delta}) = \frac{1}{2^{k+L}}, \qquad (Const)$$

siendo 2^{k+L} el número de modelos compitiendo para el problema de selección de variables del modelo completo de la Ecuación 5.1. Sin embargo, la previa constante no controla la multiplicidad, en el sentido en el que produce probabilidades de inclusión marginales que dependen fuertemente del número de variables consideradas en cada grupo. En consecuencia, Scott y Berger (2010) proponen la asignación de una distribución previa inversamente proporcional al número de modelos para una dimensión determinada $r = \kappa(\boldsymbol{\gamma},\boldsymbol{\delta})$. En

este contexto, una implementación directa de dicha previa sería la que sigue, la cual llamaremos **previa SB**:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{\mathcal{F}_k^{l_1,\ldots,l_p}(\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta})) \cdot (L + k + 1)}, \qquad (SB)$$

donde $\mathcal{F}_k^{l_1,\ldots,l_p}(r)$ es el número de modelos del espacio de modelos posibles, $\mathcal{M}$, de dimensión r con k variables singulares y p grupos de $l_1, \ldots, l_p$ variables, respectivamente. Para el cálculo de esta última previa es necesario presentar el siguiente resultado, cuya demostración (ver Anexo) es análoga a la de García-Donato y Paulo (2021) en el caso de factores.

Proposición 3.2.1. *Sea $[\mathbf{1}|\boldsymbol{X}|\boldsymbol{Z}]$ la matriz de diseño del modelo completo de la Ecuación 5.1, de rango $1 + k + L$, con $n > 1 + k + L$ y $L = \sum_{j=1}^{p} l_j$ el número de variables agrupadas. Sea el modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}$, entonces:*

1. *$\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \mathbf{1}^T\boldsymbol{\gamma} + \sum_{j=1}^{p} \mathbf{1}^T\boldsymbol{\delta}_j \in [0,\ k + L]$, siendo 0 la dimensión del modelo nulo y $L + k$ la del modelo completo; es el rango de la matriz de diseño del modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ menos uno.*
2. *$\mathcal{F}_k^{l_1,\ldots,l_p}(r) = \sum_{0\leq i\leq k,\ 0\leq j_q\leq l_q,\ 1\leq q\leq p,\ i+\sum_{q=1}^{p} j_q = r} \binom{k}{i} \cdot \prod_{q=1}^{p} \binom{l_q}{j_q}$, con $r \in [0,\ k + L]$, es el número de modelos de $\mathcal{M}$ de dimensión r con k variables singulares y p grupos de $l_1, \ldots, l_p$ variables cada uno.*

En la mayoría de distribuciones previas por defecto del espacio de modelos, se tiene una probabilidad a priori de inclusión de cada variable de 0.5. Esto es independientemente del número de variables consideradas, propiedad ampliamente aceptada en la literatura científica. No obstante, esto no se verifica en ninguna de las dos propuestas anteriores, por lo que en adelante se seguirá una aproximación similar a la realizada por García-Donato y Paulo (2021) para encontrar alternativas que respeten dicha presunción.

Para ello se construirá la distribución previa de forma jerárquica, en dos niveles, introduciendo la información sobre los grupos a través de la variable indicadora $\boldsymbol{\tau}$. Como esta es una función determinista de $\boldsymbol{\delta}$, $\boldsymbol{\tau} = \boldsymbol{\tau}(\boldsymbol{\delta})$, es decir, se calcula a partir de las variables que consideramos agrupadas, se tiene:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\tau}) = \pi(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau}) \cdot \pi(\boldsymbol{\gamma}, \boldsymbol{\tau}).$$

Previa de $(\boldsymbol{\gamma}, \boldsymbol{\tau})$: $\pi(\boldsymbol{\gamma}, \boldsymbol{\tau})$

La distribución previa $\pi(\boldsymbol{\gamma}, \boldsymbol{\tau})$ es la marginal de aquellos parámetros que indican qué variables singulares y grupos se encuentran activos en el modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$. Se busca una previa que verifique que

$$p(\tau_j = 1) = p(\gamma_i = 1) = \frac{1}{2} \quad \text{para} \quad j = 1, \ldots, p; \quad i = 1, \ldots, k;$$

para tener la propiedad deseada. Cuando consideramos la presencia de grupos también tenemos la posibilidad de elegir la distribución previa constante o la de tipo Scott y Berger[2].

La distribución **previa constante** para la marginal $(\boldsymbol{\gamma}, \boldsymbol{\tau})$ es

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{2^{k+p}}, \qquad (Const_1)$$

mientras que la **previa de tipo Scott y Berger** es

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{(k+p+1) \cdot \binom{k+p}{\mathbf{1}^T\boldsymbol{\gamma}+\mathbf{1}^T\boldsymbol{\tau}}}, \qquad (SB_1)$$

siendo k el número de variables singulares, p el número de grupos, $\mathbf{1}^T\boldsymbol{\gamma}$ el número de variables singulares activas y $\mathbf{1}^T\boldsymbol{\tau}$ el número de grupos activos en el modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$.

Previa de $(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau})$: $\pi(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau})$

La distribución previa $\pi(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau})$ es la condicional que reparte la probabilidad entre los posibles modelos cuyas variables singulares y grupos sean los dados por las variables indicadoras $\boldsymbol{\gamma}$ y $\boldsymbol{\tau}$, respectivamente; es decir, en el conjunto de modelos:

$$\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau}) = \{\boldsymbol{\gamma}, \boldsymbol{\tau} \in \mathcal{M} : \ \boldsymbol{\gamma} = \boldsymbol{\gamma}, \ min\{\mathbf{1}^T\boldsymbol{\delta}_j, 1\} = \tau_j, \ j = 1, \ldots, p\},$$

que particiona el espacio de modelos posibles original, $\mathcal{M}$, en 2^{k+p} conjuntos de modelos.

En este caso, la distribución **previa constante** para la condicional $\pi(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau})$ viene dada por la cardinalidad del conjunto $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$:

$$\pi(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{|\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})|}, \qquad (Const_2)$$

siendo $|\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})|$ el número de modelos cuyas variables singulares y agrupadas tengan una determinada combinación de las variables singulares y grupos dados por $\boldsymbol{\gamma}$ y $\boldsymbol{\tau}$. La distribución **previa de tipo Scott y Berger**, es decir, la que reparte la probabilidad de forma inversamente proporcional al número de modelos de igual dimensión en $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$, es

$$\pi(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau}) = \frac{1}{\left(\mathcal{G}^{l_{j_1},\ldots,l_{j_{m_2}}}(\kappa(\boldsymbol{\gamma}, \boldsymbol{\tau}) - \mathbf{1}^T\boldsymbol{\gamma})\right) \cdot \left(\sum_{q=1}^{m_2} l_{j_q} - m_2 + 1\right)}, \qquad (SB_2)$$

siendo $j_1, \ldots, j_{m_2}$ los índices de $\boldsymbol{\delta}$ que indican las variables agrupadas activas, $m_2 = \mathbf{1}^T\boldsymbol{\tau}$ el número de grupos activos y $\mathcal{G}^{l_{j_1},\ldots,l_{j_{m_2}}}(r)$ el número de modelos de $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$ de dimensión r. Como antes, para el

[2]Asigna la probabilidad de forma inversamente proporcional al número de modelos de igual dimensión en $\mathcal{M}$.

cálculo de esta última previa es necesario presentar el siguiente resultado, cuya demostración es, una vez más, similar a la de García-Donato y Paulo (2021) en el caso de factores.

Proposición 3.2.2. *Sea* $[\mathbf{1}|\boldsymbol{X}|\boldsymbol{Z}]$ *la matriz de diseño del modelo completo de la Ecuación 5.1, de rango* $1+k+L$*, con* $n > 1+k+L$ *y* $L = \sum_{j=1}^{p} l_j$ *el número de variables agrupadas. Sea el modelo* $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$*, entonces:*

1. $|\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})| = \prod_{j=1}^{p}(2^{\tau_j l_j} - \tau_j)$ *es el número de elementos de* $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$*.*
2. $\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta})$ *es el número de variables activas, tanto singulares como agrupadas, presentes en el modelo* $(\boldsymbol{\gamma}, \boldsymbol{\delta})$*. Se tiene que* $\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in [\mathbf{1}^T\boldsymbol{\gamma} + \mathbf{1}^T\boldsymbol{\tau},\ \mathbf{1}^T\boldsymbol{\gamma} + \sum_{j=1}^{p} \tau_j \cdot l_j]$*.*
3. $\mathcal{G}^{l_{j_1},\dots,l_{j_{m_2}}}(r) = \sum_{1\le j_1\le l_1,\ \dots,\ 1\le j_{m_2}\le l_{j_{m_2}},\ \sum_{q=1}^{m_2} j_q = r} \prod_{q=1}^{m_2} \binom{l_q}{j_q}$ *es el número de modelos de* $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$ *de dimensión* $r + \mathbf{1}^T\boldsymbol{\gamma}$*, con* $r \in [m_2,\ l_{j_1} + \cdots + l_{j_{m_2}}]$*.*

En conclusión, si no se considera la presencia de grupos se tiene dos posibilidades para la asignación de la distribución previa del modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$:

(Const) $$\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{2^{k+L}},$$

(SB) $$\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{\mathcal{F}_k^{l_1,\dots,l_p}(\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta})) \cdot (L+k+1)}.$$

Y, si se considera la presencia de grupos de variables, en la selección de la distribución previa del modelo $(\boldsymbol{\gamma}, \boldsymbol{\delta})$, $\pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \pi(\boldsymbol{\delta}|\boldsymbol{\gamma}, \boldsymbol{\tau}) \cdot \pi(\boldsymbol{\gamma}, \boldsymbol{\tau})$, se tiene un total de cuatro posibilidades diferentes:

$$(Const - Const) \quad \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{2^{k+p}} \cdot \frac{1}{\prod_{j=1}^{p}(2^{\tau_j l_j} - \tau_j)},$$

$$(Const - SB) \quad \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{2^{k+p}} \cdot \frac{1}{\left(\mathcal{G}^{l_{j_1},\dots,l_{j_{m_2}}}(\kappa(\boldsymbol{\gamma}, \boldsymbol{\tau}) - \mathbf{1}^T\boldsymbol{\gamma})\right) \cdot \left(\sum_{q=1}^{m_2} l_{j_q} - m_2 + 1\right)},$$

$$(SB - Const) \quad \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{(k+p+1) \cdot \binom{k+p}{\mathbf{1}^T\boldsymbol{\gamma} + \mathbf{1}^T\boldsymbol{\tau}}} \cdot \frac{1}{\prod_{j=1}^{p}(2^{\tau_j l_j} - \tau_j)},$$

$$(SB - SB) \quad \pi(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{(k+p+1) \cdot \binom{k+p}{\mathbf{1}^T\boldsymbol{\gamma} + \mathbf{1}^T\boldsymbol{\tau}}} \cdot \frac{1}{\left(\mathcal{G}^{l_{j_1},\dots,l_{j_{m_2}}}(\kappa(\boldsymbol{\gamma}, \boldsymbol{\tau}) - \mathbf{1}^T\boldsymbol{\gamma})\right) \cdot \left(\sum_{q=1}^{m_2} l_{j_q} - m_2 + 1\right)};$$

de las cuales se estudiará el funcionamiento aplicando en ejemplos concretos en el siguiente capítulo.

3.2.2. Esquema de *Gibbs Sampling*

Se calculan las probabilidades a posteriori de los modelos mediante:

$$\pi((\boldsymbol{\gamma},\boldsymbol{\delta})^*|\boldsymbol{y}) = \frac{B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^*}(\boldsymbol{y})\cdot\pi((\boldsymbol{\gamma},\boldsymbol{\delta})^*)}{\sum_{(\boldsymbol{\gamma},\boldsymbol{\delta})} B_{(\boldsymbol{\gamma},\boldsymbol{\delta})}(\boldsymbol{y})\cdot\pi((\boldsymbol{\gamma},\boldsymbol{\delta}))},$$

con la problemática de que el denominador contiene más sumandos a medida que m aumenta.

Una vez calculadas dichas probabilidades, tendremos el HPM y MPM correspondientes, explicados en la Sección 2.5. Este último, cuando se consideran grupos, está formado por aquellas variables singulares con probabilidad mayor que 0.5 y aquellas variables agrupadas cuyo grupo y ellas mismas tengan una probabilidad a posteriori de 0.5.

Una enumeración exhaustiva de todos los modelos sería más o menos viable si $m \leq 25$, siendo $m = L + k$ el número de variables consideradas. Si $m > 25$, debido al gran número de modelos a comparar (2^m), la práctica usual es simular del espacio de modelos para obtener una muestra de las distribuciones a posteriori. Para ello se empleará el esquema de *Gibss sampling* de George y McCulloch (1997), donde se demuestra su funcionamiento óptimo, en comparación con otros métodos similares, debido a que aporta menos incertidumbre en la estimación de los parámetros. Dicho esquema tiene la siguiente estructura:

(1) Se inicializa el algoritmo con un modelo inicial $(\boldsymbol{\gamma},\boldsymbol{\delta})^{(0)} = ((\gamma,\delta)_1^{(0)},\ldots,(\gamma,\delta)_m^{(0)})^T$.

(2) A continuación se calcula el factor Bayes asociado al modelo inicial, $B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^{(0)}}$.

(3) Para $i = 1,\ldots,N$ se repiten los siguientes $m+1$ pasos:

$(i.j)$ Para $j \in \{1,\ldots,m\}$:

$(i.j.1)$ Se define el modelo $(\boldsymbol{\gamma},\boldsymbol{\delta})^* = ((\gamma,\delta)_1^{(i-1)},\ldots,1-(\gamma,\delta)_j^{(i-1)},\ldots,(\gamma,\delta)_m^{(i-1)})^T$.

$(i.j.2)$ Se calcula $B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^*0}$, factor Bayes asociado al modelo $M_{(\boldsymbol{\gamma},\boldsymbol{\delta})^*}$ con respecto al modelo nulo.

$(i.j.3)$ Se calcula:

$$r = \frac{B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^*0}\cdot\pi((\boldsymbol{\gamma},\boldsymbol{\delta})^*)}{B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^*0}\cdot\pi((\boldsymbol{\gamma},\boldsymbol{\delta})^{(i-1)}) + B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^*0}\cdot\pi((\boldsymbol{\gamma},\boldsymbol{\delta})^{(i-1)})}.$$

Se simula una observación de una distribución Bernuilli con probabilidad de éxito igual a r. Si se obtiene éxito, entonces se actualiza $(\boldsymbol{\gamma},\boldsymbol{\delta})^{(i-1)} = (\boldsymbol{\gamma},\boldsymbol{\delta})^*$ y $B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^{(i-1)}0} = B_{(\boldsymbol{\gamma},\boldsymbol{\delta})^*0}$. En caso de fracaso, no se actualiza y se continua con el paso siguiente.

$(i.m+1)$ Se define y se guarda $(\boldsymbol{\gamma},\boldsymbol{\delta})^{(i)} = (\boldsymbol{\gamma},\boldsymbol{\delta})^{(i-1)}$.

El factor Bayes de cada modelo se calculará asignando una distribución previa uniforme al coeficiente del intercepto y la *g-prior* robusta a los parámetros específicos de cada modelo. La distribución previa de los modelos se asignará de forma jerárquica, siendo la de tipo de Scott y Berger en ambos niveles (SBSB) la que se espera que obtenga mejores resultados.

Para calcular las probabilidades a posteriori del espacio de modelos, y con ellas las probabilidades de

inclusión de cada variable singular, grupo y variable agrupada, se utilizará la muestra generada mediante dicho algoritmo, $(\boldsymbol{\gamma}, \boldsymbol{\delta})^{(1)}, \ldots, (\boldsymbol{\gamma}, \boldsymbol{\delta})^{(N)}$.

4. Resultados en datos simulados

En este capítulo vamos a analizar la metodología desarrollada en el Capítulo 3 mediante una serie de experimentos basados en simulaciones, siguiendo lo que se hizo en García-Donato y Paulo (2021) para el caso de factores. Para ello, se analizarán cómo, diferentes aspectos clave centrados en el concepto de grupo de variables, modifican los resultados obtenidos mediante la metodología propuesta y sus implicaciones.

Para realizar la selección de variables de cada base de datos simulada, se muestrearon 1100 modelos con reemplazamiento, de los cuales 1000 se guardaron para calcular las probabilidades de inclusión de los grupos y variables. El código de la función donde se implementa la metodología propuesta, junto con el que se utilizó para simular los datos de las comparaciones que siguen, se encuentra en https://github.com/camulro/Seleccion-de-variables-agrupadas.

4.1. Comparación: Estructura de grupos vs No estructura de grupos

El objetivo es comparar el funcionamiento de la metodología propuesta, es decir, considerando la estructura de grupos a través de la previa sobre el espacio de modelos, con la selección bayesiana tradicional empleando la previa *SB*. Se empezará investigando en el control de la multiplicidad en lo que se refiere a la inclusión de los grupos en el MPM. Para poder realizar dicha comparación cuando no se introduce la estructura de grupos (pevia *SB*), se calculó la probabilidad de cada grupo como la suma de las probabilidades a posteriori de aquellos modelos con al menos una variable activa de dicho grupo. Este experimento se basa en el de Scott y Berger (2010), reproducido por García-Donato y Paulo (2021) para factores y adaptándolo ahora para grupos en modelos de regresión logística.

Se simularon un total de 10 bases de datos de un modelo de regresión logística, con $n = 600$ observaciones y p grupos con $l_j = 4$ variables cada uno. De estos, se eligió que los dos primeros fueran activos, con

parámetros de regresión:

$$\beta_1 = \begin{bmatrix} -0.6 \\ -0.56 \\ -0.53 \\ 0.49 \end{bmatrix}, \ \beta_2 = \begin{bmatrix} -0.45 \\ 0.4 \\ 0.35 \\ 0.3 \end{bmatrix}, \ \beta_j = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \text{ para } j \in \{3, \ldots, p\}.$$

El intercepto fue fijado, eligiendo $\beta_0 = 1$ y la matriz de varianzas-covarianzas de cada grupo fue simulada.

Para evaluar el control de la multiplicidad en lo que se refiere a los grupos, se aumentó el número de grupos espurios ($p - 2$) considerado en la selección de variables. En particular, se realizó la selección de cada banco de datos tres veces, con $p \in \{3, 5, 7\}$ grupos, es decir, $L \in \{12, 20, 28\}$ variables agrupadas.

Los resultados obtenidos se representan en las Figuras 4.1, 4.2 y 4.3 en la forma de las probabilidades de inclusión a posteriori de los grupos para los 10 bancos de datos simulados con $p = 3$, $p = 5$ y $p = 7$ grupos, respectivamente. Asimismo, en la Figura 4.4 se representan las probabilidades de inclusión a posteriori de las variables para los 30 bancos de datos a los que se les ha hecho la selección, en función del número de grupos introducido en cada uno.

Al no considerar la presencia de grupos (*SB*), el porcentaje de grupos espurios incluidos en el MPM (falsos positivos) es mayor, llegando a incluir un $40\,\%$ de los grupos falsos en el primer caso. Si bien con las previas que sí tienen la estructura de grupo, se observa un control óptimo de las señales falsas. Con todas las previas se ve reducida la probabilidad del segundo grupo al aumentar el número de grupos falsos. Sin embargo, con *SB* se observa además una ligera reducción de las probabilidades del primero, el que se incluye con mayor seguridad, hecho que no sucede con el resto de previas. Por tanto, se puede afirmar que cuando no se considera la estructura de grupos, no se obtiene un buen control de la multiplicidad a nivel del grupo.

Por otro lado, las probabilidades de inclusión a posteriori de las variables verdaderas obtenidas mediante *SB* tienden a ser menores que las del resto, viéndose considerablemente reducidas al aumentar el número de grupos falsos. En contraposición, aumentar el número de grupos falsos en el resto de previas apenas tiene ningún impacto en los resultados. No se comentarán las variables de los grupos espurios puesto que estas se rechazan, en la mayoría de casos, a través de la probabilidad del grupo.

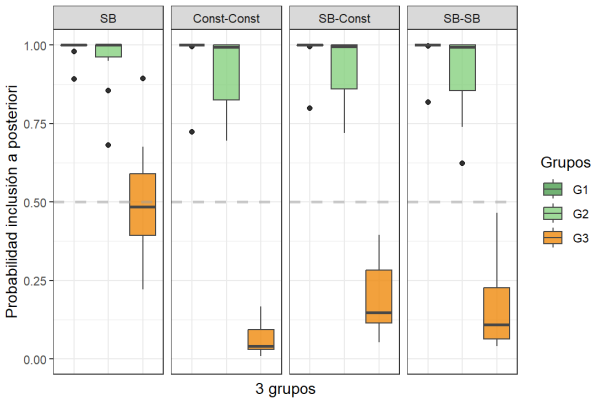


Figura 4.1: Probabilidades de inclusión a posteriori de los grupos del primer experimento de las 10 bases con 3 grupos.

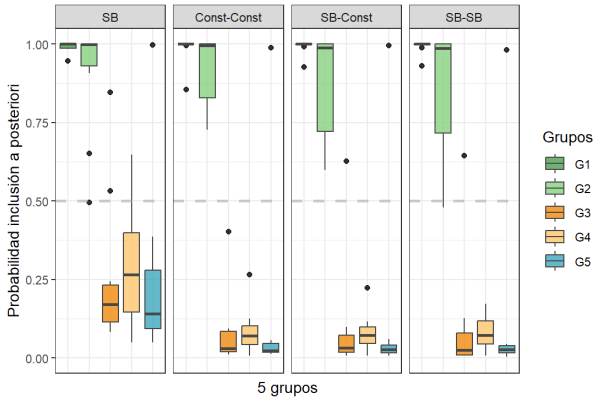


Figura 4.3: Probabilidades de inclusión a posteriori de los grupos del primer experimento de las 10 bases con 5 grupos.

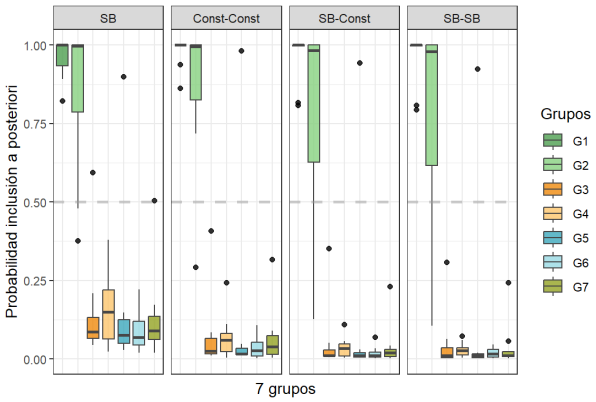


Figura 4.2: Probabilidades de inclusión a posteriori de los grupos del primer experimento de las 10 bases con 5 grupos.

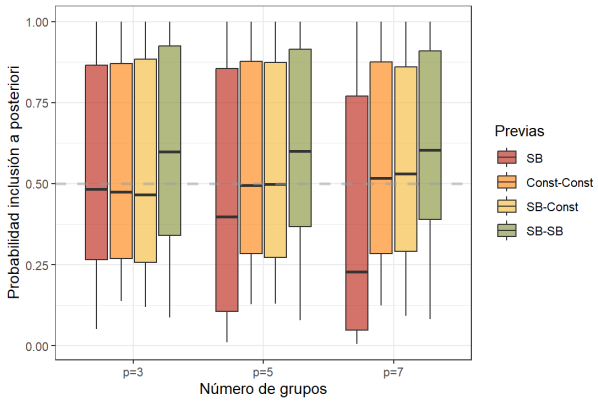


Figura 4.4: Probabilidades de inclusión a posteriori de las variables verdaderas de las 30 bases, en función del número de grupos.

En esta comparación no se han visto diferencias notables entre las distintas previas que sí consideran la estructura de grupos. Es razonable pensar, en base a los resultados de Scott y Berger (2010), que dichas previas tendrán funcionamientos diferentes en cuanto al control de la multiplicidad a nivel de las variables dentro de los grupos.

4.2. Comparación: Previas con estructura de grupos

El objetivo es comparar el funcionamiento de las distintas previas posibles en la metodología propuesta, es decir, considerando la estructura de grupos en la previa sobre el espacio de modelos. Se continuará investigando en el control de la multiplicidad, esta vez en lo que se refiere a la inclusión de las variables agrupadas (nivel intra-grupo) en el MPM. Este experimento fue propuesto en García-Donato y Paulo (2021) para el caso de

factores en modelos lineales, esperando obtener que la probabilidad de declarar erróneamente falsos positivos aumentara con la previa constante, mientras que dicha probabilidad se mantuviera casi constante con la previa de tipo Scott y Berger. Esto no fue confirmado.

4.2.1. Aumentando el número de variables agrupadas falsas

En primer lugar, si se ven afectadas las previas consideradas al aumentar el número de variables agrupadas espurias introducidas en la selección de variables.

Se simularon 3 bases de datos de un modelo de regresión logística, con $n = 600$ observaciones y $p = 4$ grupos, todos con el mismo número de variables agrupadas $l_j = l$. De estos, se eligió que únicamente el primero fuera activo con parámetros de regresión:

$$\beta_1 = \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \beta_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \text{ para } j \in \{2, 3, 4\}.$$

El intercepto fue fijado a $\beta_0 = 1$ y la matriz de varianzas-covarianzas de cada grupo fue simulada.

Para analizar el efecto de incrementar el número de variables agrupadas espurias, cada base de datos simulada se generó con un número diferente de variables agrupadas: $l \in \{5, 10, 15\}$. Este experimento se realizó dos veces, primero tomando $\beta_{11} = 0.56$ y $\beta_{12} = 0.49$; y luego $\beta_{11} = -0.43$ y $\beta_{12} = -0.37$; obteniendo así un total de 6 bases de datos simuladas con 4 grupos con distinto número de variables agrupadas.

Las probabilidades de inclusión a posteriori de cada grupo obtenidas para las primeras 3 bases simuladas, en función del número de variables considerado, quedan recogidas en la Tabla 4.1, y para las siguientes 3 bases en la Tabla 4.2. Se volvió a incluir la previa *SB* para contrastar con las probabilidades obtenidas sin considerar la estructura de grupos en un problema de estas características.

En el primer caso, no se observan diferencias entre las probabilidades a posteriori de los grupos obtenidas mediante las previas que tienen en cuenta la presencia de grupos. Además, estas no parecen verse afectadas por el aumento del número de variables espurias en cada grupo. Sin embargo, en los resultados cuando los coeficientes son menores sí se observan diferencias de interés. La probabilidad obtenida de los grupos verdaderos se ve reducida en todas las previas al aumentar el número de variables. Esto es especialmente notorio en las previas que son combinación de la constante, puesto que no incluyen erróneamente en el MPM el primer grupo en el caso con 15 variables por grupo. Por tanto, parece que dichas previas no proporcionan un buen control de la multiplicidad de las variables dentro de los grupos.

Además, se observa que, sin considerar la estructura de grupo, el control de la multiplicidad a nivel intra-grupo es deficiente en ambos casos.

Tabla 4.1: Probabilidades de inclusión de los grupos obtenidas para las bases simuladas con $\beta_{11} = 0.56$ y $\beta_{12} = 0.49$.

$\boldsymbol{\beta_{11} = 0.56}$, $\boldsymbol{\beta_{12} = 0.49}$												
	SB			*Const-Const*			*SB-Const*			*SB-SB*		
	$l=5$	$l=10$	$l=15$	$l=5$	$l=10$	$l=15$	$l=5$	$l=10$	$l=15$	$l=5$	$l=10$	$l=15$
$p(\tau_1 = 1\|\boldsymbol{y})$	1	1	1	1	1	1	1	1	1	1	1	1
$p(\tau_2 = 1\|\boldsymbol{y})$	0.31	0.37	0.24	0.1	0	0	0.07	0	0	0.07	0.10	0.05
$p(\tau_3 = 1\|\boldsymbol{y})$	0.20	0.11	0.09	0.07	0	0	0.05	0	0	0.05	0.03	0.01
$p(\tau_4 = 1\|\boldsymbol{y})$	0.22	0.62	0.50	0.04	0.01	0	0.03	0.01	0	0.04	0.23	0.16

Tabla 4.2: Probabilidades de inclusión de los grupos obtenidas para las bases simuladas con $\beta_{11} = -0.43$ y $\beta_{12} = -0.37$.

$\boldsymbol{\beta_{11} = -0.43}$, $\boldsymbol{\beta_{12} = -0.37}$												
	SB			*Const-Const*			*SB-Const*			*SB-SB*		
	$l=5$	$l=10$	$l=15$	$l=5$	$l=10$	$l=15$	$l=5$	$l=10$	$l=15$	$l=5$	$l=10$	$l=15$
$p(\tau_1 = 1\|\boldsymbol{y})$	0.89	0.64	0.43	0.99	0.89	0.23	0.98	0.67	0.05	0.98	0.90	0.78
$p(\tau_2 = 1\|\boldsymbol{y})$	0.08	0.04	0.02	0.03	0	0	0.02	0	0	0.05	0.03	0.03
$p(\tau_3 = 1\|\boldsymbol{y})$	0.09	0.03	0.02	0.03	0	0	0.02	0	0	0.06	0.03	0.02
$p(\tau_4 = 1\|\boldsymbol{y})$	0.03	0.02	0.01	0.01	0	0	0	0	0	0.02	0.01	0.01

Se han visto diferencias entre las previas consideradas en cuanto al control de la multiplicidad cuando aumenta el número de variables espurias. Sin embargo, cabe plantearse si esto también sucede cuando el número de variables que generan los datos se incrementa. Como la previa constante favorece elecciones con la mitad de variables, puesto que el número combinatorio correspondiente es mayor, se quiere ver si esto afecta también a la probabilidad de inclusión de las variables cuando aumenta el número de variables que generan los datos. Se prevé que las previas *Const-Const* y *SB-Const* muestren más dificultades para elegir todas las variables verdaderas, mientras que *SB-SB* no debería tener ese problema.

4.2.2. Aumentar el número de variables agrupadas verdaderas

Veamos si se ven afectadas las previas consideradas al aumentar el número de variables agrupadas que generan los datos. El experimento se llevó a cabo como el anterior, pero considerando todas las variables del primer grupo verdaderas.

Se simularon 3 bases de datos de un modelo de regresión logística, con $n = 600$ observaciones y $p = 4$ grupos, todos con el mismo número de variables agrupadas $l_j = l$. De estos, se eligió que únicamente el

primer grupo fuera activo. Para analizar el efecto de incrementar el número de variables agrupadas verdaderas, cada base de datos simulada se generó con un número diferente de variables agrupadas: $l \in \{5, 10, 15\}$. Además, este experimento se realizó dos veces como antes, modificando los coeficientes que generan los datos, obteniendo una vez más un total de 6 bases de datos simuladas. Los parámetros de regresión para las bases con $l = 5$ variables por grupo fueron:

$$\beta_1 = \begin{bmatrix} 0.63 \\ 0.58 \\ -0.53 \\ 0.5 \\ -0.46 \end{bmatrix}, \ \beta_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \quad \text{y} \quad \beta_1 = \begin{bmatrix} -0.41 \\ 0.39 \\ -0.35 \\ 0.3 \\ 0.27 \end{bmatrix}, \ \beta_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \text{con } j \in \{2, 3, 4\},$$

para el primer y segundo caso, respectivamente.

Para aumentar el número de variables a $l = 10$ y $l = 15$, se repitieron los mismos coeficientes de $l = 5$ dos y tres veces, respectivamente. El intercepto fue fijado una vez más a $\beta_0 = 1$ y la matriz de varianzas-covarianzas fue simulada.

En todos los casos, la probabilidad de inclusión del primer grupo fue 1 y la de los grupos espurios fue muy próxima a 0 para las previas que consideran la estructura de grupos.

Las probabilidades de inclusión a posteriori de cada grupo obtenidas para los 6 bancos de datos simulados, en función del número de variables considerado, se muestran en la Figura 4.5. Se volvió a incluir la previa *SB* para contrastar con las probabilidades obtenidas sin considerar la estructura de grupos.

Con $l = 5$ y $l = 10$ variables por grupo, las tres previas que consideran la presencia de grupos parecen obtener probabilidades similares, siendo *SB-SB* la que asigna mayores valores y obtiene resultados más precisos. Sin embargo, para los datos simulados con $l = 15$ variables por grupo se observan diferencias sustanciales. Como se esperaba, las previas que son combinación de la constante tienden a incluir en el MPM la mitad de las variables, a diferencia de *SB-SB* que incluye el $93\,\%$ y el $73\,\%$ en cada caso. Por tanto, parece ser *SB-SB* la que mayor control de la multiplicidad proporciona cuando se aumenta el número de variables que generan los datos.

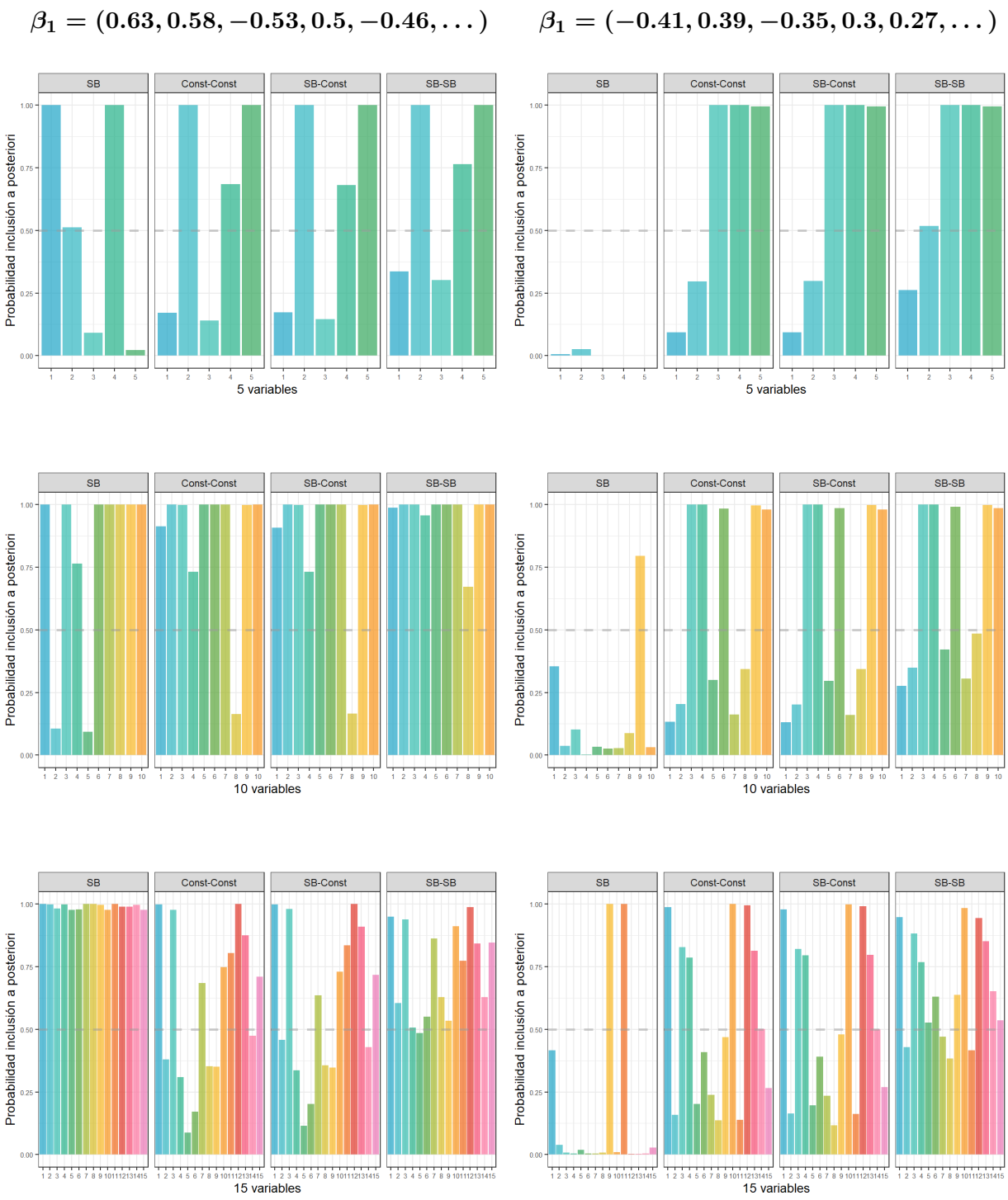


Figura 4.5: Probabilidades de inclusión a posteriori de las variables del primer grupo para las tres bases simuladas con $\beta_1 = (0.63, 0.58, -0.53, 0.5, -0.46, \dots)$ (izquierda) y para las tres simuladas con $\beta_1 = (-0.41, 0.39, -0.35, 0.3, 0.27, \dots)$ (derecha), en función del número de variables consideradas.

Hasta ahora hemos podido discernir el comportamiento de las diferentes previas consideradas y también se ha vislumbrado el papel que juega considerar la estructura de grupos en las distribuciones previas del espacio

de modelos. La diferencia entre las tres previas no parece importante cuando el número de variables de cada grupo no es elevado. Sin embargo, *SB-SB* parece ser más robusta, asignando un peso igual a todos los modelos con los mismo grupos activos y con el mismo número de variables cada uno.

El método propuesto parece discriminar correctamente los grupos y variables verdaderos de los falsos en la mayoría de los casos. Cabe plantearse, si su funcionamiento es también notable en comparación con otras metodologías ya existentes.

4.3. Comparación con otras metodologías

El objetivo es estudiar el funcionamiento de la metodología propuesta con la previa *SB-SB*, en comparación con otras ya ampliamente estudiadas, a través de la proporción de veces que cada método incluye erróneamente las variables explicativas falsas (falsos positivos) y que rechaza las verdaderas (falsos negativos). En el método propuesto se incluirá tanto el MPM como el HPM obtenido. Para ello se seguirá, una vez más, el experimento realizado por García-Donato y Paulo (2021) con factores. Los otros métodos de selección de variables en presencia de grupos considerados son: grupo Lasso (Yian y Lin, 2006), grupo MCP (Breheny, 2009) y grupo SCAD (**?**); explicados en el Capítulo 3.

Se simularon 20 bases de datos de un modelo de regresión logística, con $n = 300$ observaciones las 10 primeras, y $n = 600$ observaciones las 10 siguientes. Todas estaban formadas por $p = 4$ grupos, con $l_1 = l_3 = 8$ y $l_2 = l_4 = 4$ variables. Además, este experimento se realizó dos veces, modificando ls coeficientes que generan los datos, obteniendo un total de 40 bases de datos simuladas. Se eligió que los dos primeros grupos fueran activos con parámetros de regresión:

$$\beta_1 = \begin{bmatrix} 0 \\ 0 \\ 0.55 \\ 0.55 \\ 0.55 \\ 0.55 \\ -0.49 \\ -0.49 \end{bmatrix}, \ \beta_2 = \begin{bmatrix} 0 \\ 0 \\ 0.46 \\ 0.46 \end{bmatrix}, \ \beta_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \quad \text{y} \quad \beta_1 = \begin{bmatrix} 0.42 \\ 0 \\ \vdots \\ 0 \\ -0.4 \end{bmatrix}, \ \beta_2 = \begin{bmatrix} 0.35 \\ 0 \\ 0 \\ -0.3 \end{bmatrix}, \ \beta_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ j \in \{3, 4\},$$

para el primer y segundo caso, respectivamente. El intercepto fue fijado, eligiendo $\beta_0 = 1$ y la matriz de varianzas-covarianzas fue simulada una vez más.

La proporción de grupos incorrectamente seleccionados e incorrectamente no seleccionados de las 20

bases de datos simuladas con $\beta_1 = [0, 0, 0.55, 0.55, 0.55, 0.55, -0.49, -0.49]$ y $\beta_2 = [0, 0, 0.46, 0.46]$ queda recogida en la Tabla 4.3; y las 20 simuladas con $\beta_1 = [0.42, 0, 0, 0, 0, 0, 0, -0.4]$ y $\beta_2 = [0.35, 0, 0, -0.3]$ en la Tabla 4.4.

Tabla 4.3: Proporción de grupos incorrectamente no seleccionados ($\mathcal{G}_1$ y $\mathcal{G}_2$) e incorrectamente seleccionados ($\mathcal{G}_3$ y $\mathcal{G}_4$) en el primer caso.

$\boldsymbol{\beta_1 = [0, 0, 0.55, 0.55, 0.55, 0.55, -0.49, -0.49]}$, $\boldsymbol{\beta_2 = [0, 0, 0.46, 0.46]}$								
	$n = 300$				$n = 600$			
	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$
HPM	0	0.3	0	0	0	0.1	0	0
MPM	0	0.1	0	0	0	0.1	0	0
Grupo Lasso	0.4	0.3	0	0	0.1	0.1	0.2	0.1
Grupo MCP	0.3	0.5	0	0	0.1	0.3	0	0
Grupo SCAD	0.3	0.3	0	0	0.1	0.3	0.1	0

Tabla 4.4: Proporción de grupos incorrectamente no seleccionados ($\mathcal{G}_1$ y $\mathcal{G}_2$) e incorrectamente seleccionados ($\mathcal{G}_3$ y $\mathcal{G}_4$) en el segundo caso.

$\boldsymbol{\beta_1 = [0.42, 0, 0, 0, 0, 0, 0, -0.4]}$, $\boldsymbol{\beta_2 = [0.35, 0, 0, -0.3]}$								
	$n = 300$				$n = 600$			
	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$
HPM	0.7	0.6	0	0	0	0.3	0	0
MPM	0.4	0.5	0	0	0	0.3	0	0
Grupo Lasso	0.9	0.8	0	0	0.7	0.6	0	0
Grupo MCP	1	0.8	0	0	0.7	0.6	0	0
Grupo SCAD	0.9	0.8	0	0	0.7	0.6	0.1	0

En el primer caso, se observa que el porcentaje de falsos negativos es mucho mayor con la aproximación frecuentista, siendo especialmente notable con un menor número de observaciones. Además, con $n = 600$ observaciones se observa un ligero aumento de los falsos positivos con la aproximación frecuentista, mientras que el método propuesto nunca falla en la detección de los grupos falsos. En el segundo caso, donde hay menos variables verdaderas y los coeficientes son menores, se observan resultados mucho peores. En comparación con el MPM, los métodos clásicos llegan a fallar hasta un $70\,\%$ más en la detección de los grupos verdaderos. Por tanto, el método propuesto parece ser más preciso que otros métodos de selección de variables tradicionales.

Cabe destacar que, en ambos casos, los modelos MPM y HPM solo coinciden en los datos generados con 600 observaciones. Con menos observaciones, es el MPM el que mejores resultados presenta y, además, a mayor número de observaciones, la eficacia del método de selección es superior.

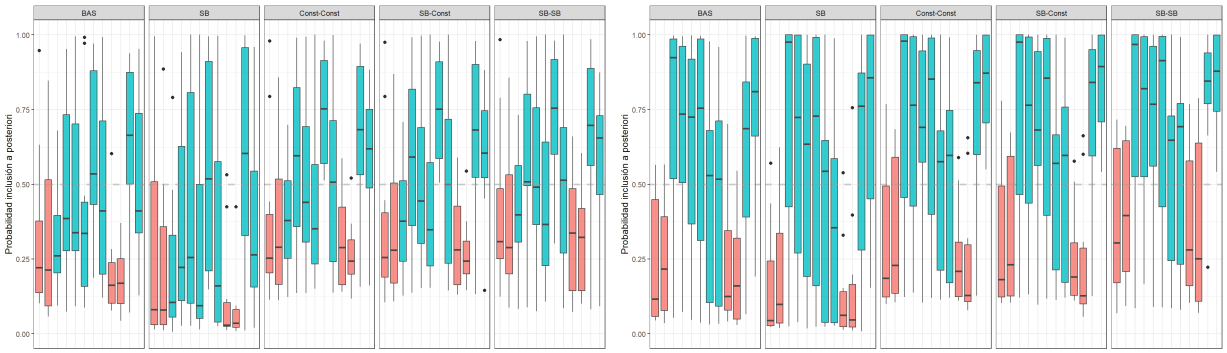
Por otro lado, se han comparado las probabilidades de inclusión a posteriori obtenidas con el método

propuesto y el implementado en la librería BAS (Clyde *et al.*, 2011). Este último no selecciona variables en presencia de grupos, pero era de interés analizar la diferencia entre estas dos aproximaciones.

Respecto a los grupos, considerando la estructura de grupos no se incluye ninguna variable de $\mathcal{G}_3$ y $\mathcal{G}_4$, mientras que con *BAS* sí sucede. Las probabilidades de inclusión a posteriori de las variables de los grupos verdaderos ($\mathcal{G}_1$ y $\mathcal{G}_2$) de los 40 datos simulados se muestran en la Figura 4.6, donde las primeras 8 variables se corresponden al primer grupo y las últimas cuatro, al segundo.

No se observan diferencias muy llamativas entre los métodos. Parece que considerando la estructura de grupos se obtiene una probabilidad a posteriori mayor para las variables verdaderas, con una proporción de variables incorrectamente seleccionadas un poco menor. Sin embargo, también aumenta ligeramente la proporción de variables incorrectamente seleccionadas.

$$\boldsymbol{\beta_1} = [\mathbf{0}, \mathbf{0}, \mathbf{0.55}, \mathbf{0.55}, \mathbf{0.55}, \mathbf{0.55}, -\mathbf{0.49}, -\mathbf{0.49}], \;\; \boldsymbol{\beta_2} = [\mathbf{0}, \mathbf{0}, \mathbf{0.46}, \mathbf{0.46}]$$



$$\boldsymbol{\beta_1} = [\mathbf{0.42}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, -\mathbf{0.4}], \;\; \boldsymbol{\beta_2} = [\mathbf{0.35}, \mathbf{0}, \mathbf{0}, -\mathbf{0.3}]$$

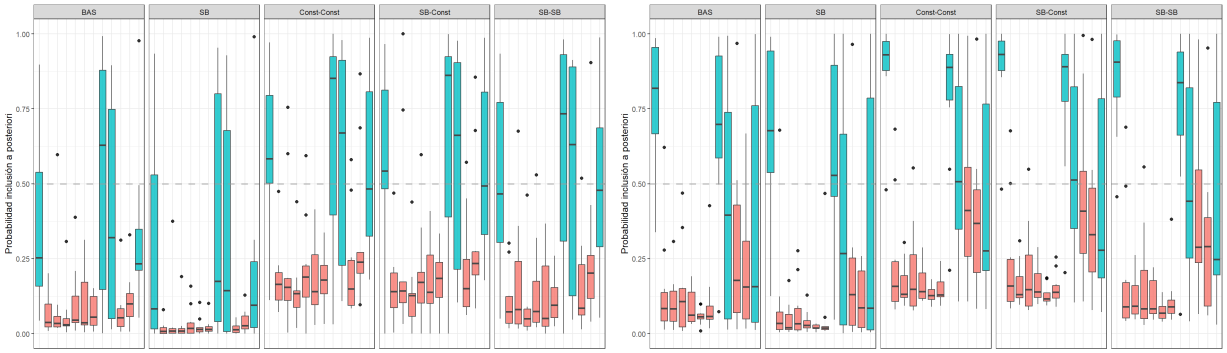


Figura 4.6: Probabilidades de inclusión a posteriori de las variables de los dos primeros grupos. Arriba los resultados obtenidos para el primer banco de datos, abajo para el segundo; siendo los de la izquierda considerando 300 observaciones y los de la derecha 600. El color azul hace referencia a las variables que generan los datos, mientras que el rojo refiere a las falsas.

Los resultados obtenidos son bastante concluyentes en cuanto a la eficacia del método propuesto en comparación con la alternativa frecuentista. Sin embargo, no ha sido posible diferenciar con la metodología de *BAS*, puesto que esta parece ser, en este caso, eficaz para seleccionar variables relacionadas sin considerar la presencia de grupos. Volveremos con esto más adelante, con unas consideraciones extra que aportará más información del funcionamiento de dichas metodologías.

4.4. Consideraciones adicionales

En los experimentos anteriores se han visto diferencias en cuanto al control de la multiplicidad entre la selección de variables tradicional con la previa *SB*, con la metodología propuesta, es decir, considerando la estructura de grupo a través de las previas del espacio de modelos y, a su vez, entre las distintas previas de esta última. Sin embargo, las situaciones estudiadas hasta ahora son muy concretas, con muchas suposiciones que no reflejan situaciones con datos reales y que pueden afectar los resultados obtenidos y comprometer las conclusiones. El objetivo ahora es profundizar en el funcionamiento y la eficacia de estas previas a la hora de seleccionar variables agrupadas, modificando algunas especificaciones intrínsecamente ligadas al concepto de grupo de variables que no se han tenido en cuenta hasta ahora. Para ello, se partirá de un modelo y se modificarán diferentes aspectos para analizar las posibles diferencias en los resultados de la selección en cada caso. También se incluirán los resultados obtenidos con *SB* para comparar con la aproximación sin tener en cuenta los grupos.

En adelante, se asociará el concepto de grupo con el de variables correlacionadas, puesto que suele suceder en situaciones reales donde las variables compiten por explicar el mismo concepto relativo a la variable respuesta. Además, es razonable considerar las variables altamente correlacionadas en conjunto, en lugar de individualmente, sin mayor contexto. Por otro lado, no es realista asumir que se tienen todas las variables que generan los datos, por lo que en la selección se considerarán solo las primeras variables de cada grupo y se omitirán el resto. El experimento se llevó a cabo extendiendo los realizados por Barbieri y Berger (2004), para contrastar el efecto de la correlación en la selección de variables.

En primer lugar, se quiere ver cómo afecta el grado de correlación de las variables a los resultados obtenidos con la metodología propuesta. Nuestro interés radica en el hecho de que la selección de variables clásica pierde su optimalidad en presencia de variables altamente correlacionadas, situación que se da a menudo con datos reales. Asimismo, también es de interés comparar con el desempeño de *BAS* en situaciones de estas características, puesto que parece razonable cuestionarlo y no se llegaron a ver diferencias en la sección anterior.

Se simularon 10 bases de datos de un modelo de regresión logística, con $n = 600$ observaciones y $p = 3$ grupos. Los grupos están formados por $l_1 = l_2 = l_3 = 20$ variables, con un coeficiente de correlación

$\rho = 0.9$. Se eligió que todas generaran los datos, con parámetros de regresión $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ y $\boldsymbol{\beta}_3$ simulados con una normal de media $\mu = 0.5$ y desviación típica $\sigma = 0.1$, 0.15 y 0.2, respectivamente. El intercepto fue fijado a $\beta_0 = 1$ y el coeficiente de correlación a $\rho = 0.9$ a través de la matriz de varianzas-covarianzas, imponiendo $Var(X_i) = Var(X_j) = 1$ y $Cov(X_i, X_j) = 0.9$ para cualquier par de variables X_i, X_j, $i \neq j$, y para cada grupo $\mathcal{G}_l$, con $i, j \in 1, \ldots, 20$ y $l \in \{1, 2, 3\}$. A partir de cada una de las 10 bases simuladas, se formaron 3 bases eligiendo únicamente las primeras 5, 7 y 9 variables de cada grupo respectivamente, haciendo un total de 30 bases a las que realizar la selección.

Los resultados obtenidos con *BAS*, en la Figura 4.7 se representan las probabilidades de inclusión a posteriori obtenidas para las variables de los 30 datos simulados, en función del número de variables considerado. En cuanto a los resultados obtenidos con el método propuesto, a nivel del grupo se obtiene una probabilidad de 1 para los tres grupos, con todas las previas y en todas las simulaciones realizadas. Respecto a las variables, se representan en la Figura 4.8 las probabilidades de inclusión a posteriori de las variables de cada grupo para los 30 bancos de datos considerados en la selección.

En todos los casos, las previas que son combinación la constante asignan probabilidades menores a las variables de los grupos, llegando a producir más de un $50\,\%$ de falsos negativos en las bases con 9 variables por grupo. En cambio, *SB* y *SB-SB* son más efectivas con un número de falsos negativos menor del $15\,\%$, mostrando un mayor control de la multiplicidad.

A pesar de que *BAS* parece tener un mejor desempeño que *Const-Const* y *SB-Const*, la proporción de variables incorrectamente no seleccionadas es, en todos los casos, mucho mayor que cuando se considera la estructura de grupo. Además, las probabilidades obtenidas disminuyen considerablemente a medida que aumenta el número de variables. Por tanto, *BAS* tampoco proporciona un óptimo control de la multiplicidad y, en situaciones donde las variables están altamente correlacionadas, la selección de variables con estructura de grupos a través de la previa *SB-SB* parece ser una mejor opción.

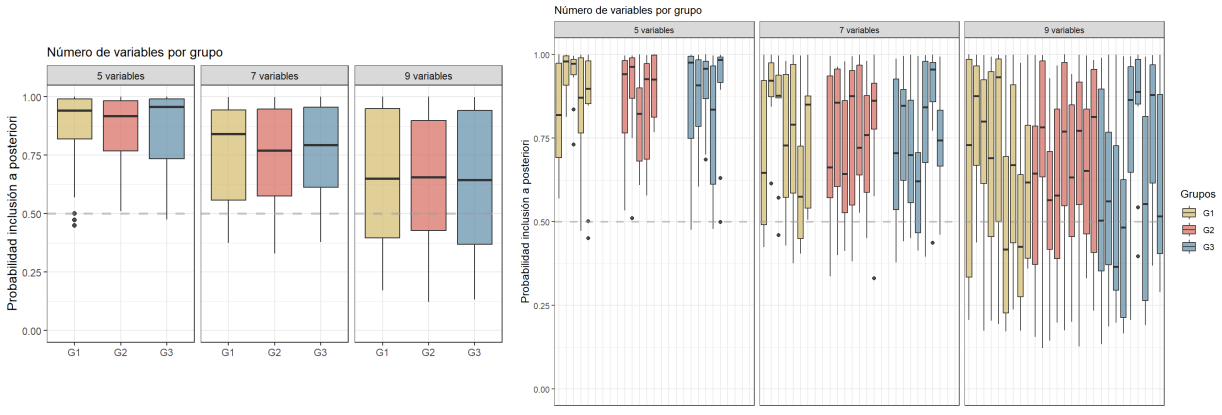


Figura 4.7: Probabilidad de inclusión a posteriori de las variables de cada grupo obtenidas con *BAS*, en función del número de variables por grupo. A la izquierda por grupo y a la derecha por variable, para cada número de variables considerado.

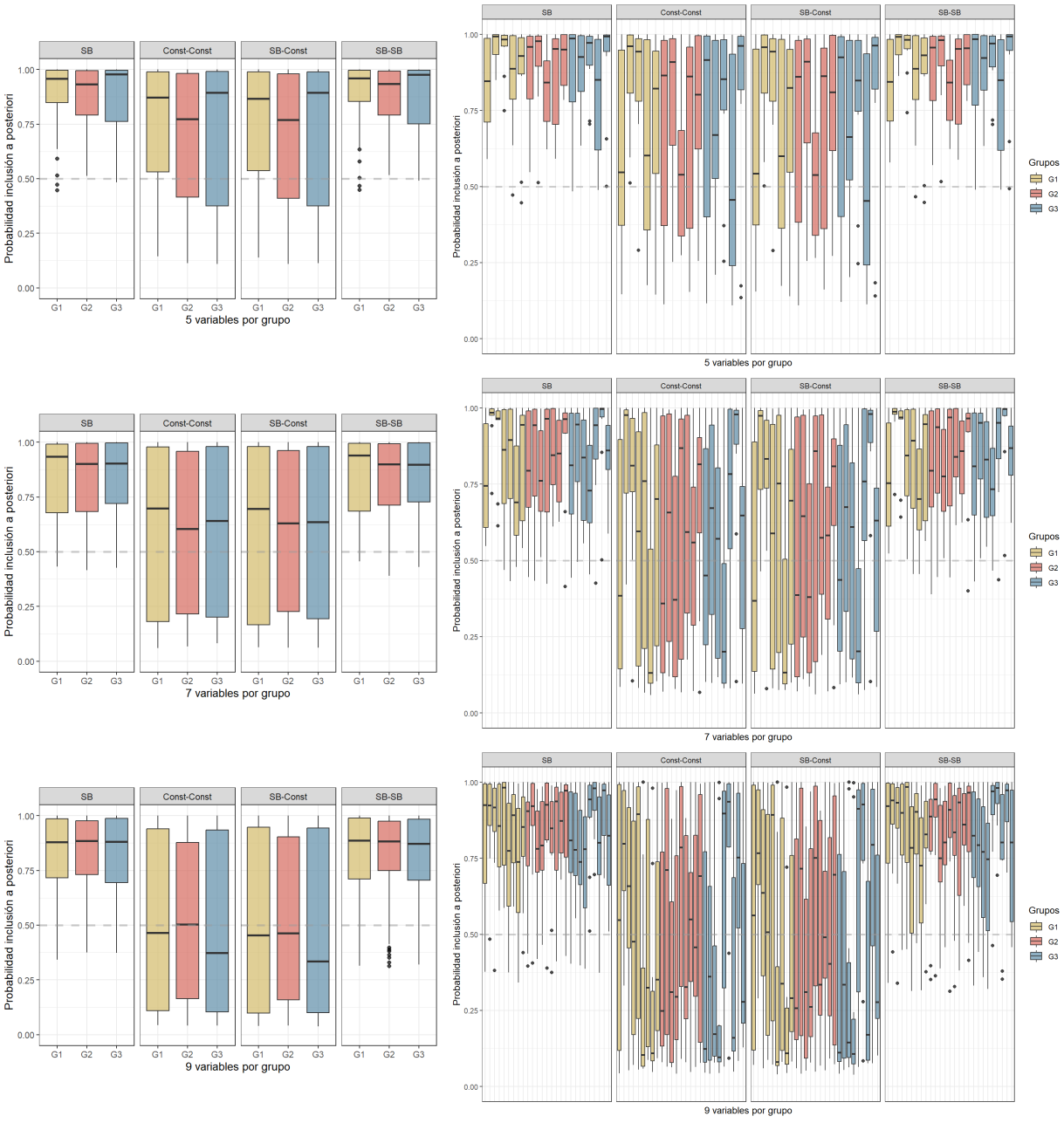


Figura 4.8: Probabilidad de inclusión a posteriori de las variables de los 30 bancos de datos seleccionados, en función del número de variables considerado. A la izquierda por grupo y a la derecha por variable.

Sin embargo, cabe preguntarse si los resultados obtenidos son consecuencia de la elevada correlación o de las datos en sí mismos.

4.4.1. Efecto de la correlación

El objetivo es contrastar si el grado de correlación entre las variables afecta a los resultados obtenidos con el método propuesto y si las diferencias observadas entre las diferentes previas son debidas a dicho efecto.

Los datos se simularon de un modelo de regresión logística con $n = 600$ observaciones y 3 grupos, con 20 variables cada uno y un coeficiente de correlación fijo. Los parámetros de regresión fueron simulados con una normal de media $\mu = 0.5$ y desviación típica $\sigma = 0.1$, 0.15 y 0.2, respectivamente. El intercepto fue nuevamente $\beta_0 = 1$.

Para evaluar el efecto de la correlación en los resultados de la selección, se simularon 5 bancos de datos con cada una de las correlaciones elegidas. A partir de cada uno de estos, se formaron 3 bases eligiendo únicamente las primeras 5, 7 y 9 variables de cada grupo respectivamente. Haciendo un total de 60 bancos de datos a los que realizar la selección, con distinta correlación y número de variables. Se eligió $\rho \in \{0.5, 0.6, 0.7, 0.8\}$ puesto que si ρ es menor, carece de sentido considerar grupos de variables en base a dicho criterio.

La probabilidad de inclusión a posteriori de los grupos fue de 1 para todas las previas y coeficiente de correlación considerados de los 60 bancos de datos. En la Figura 4.9 se representa la probabilidad de inclusión a posteriori de las variables de cada grupo de los 60 bancos de datos en función de la correlación elegida.

En todas las variables, con independencia del grupo, la probabilidad disminuye a medida que aumenta la correlación con las previas que son combinación de la constante. Además, estas tienen un desempeño peor en todos los casos debido al gran número de variables considerado. El efecto de la correlación con *SB* y *SB-SB* es muy leve. En todos los casos, estas son las que mejores resultados obtienen, incluyendo en el MPM todas las variables siempre, y son las que casi no se ven afectadas por la multiplicidad.

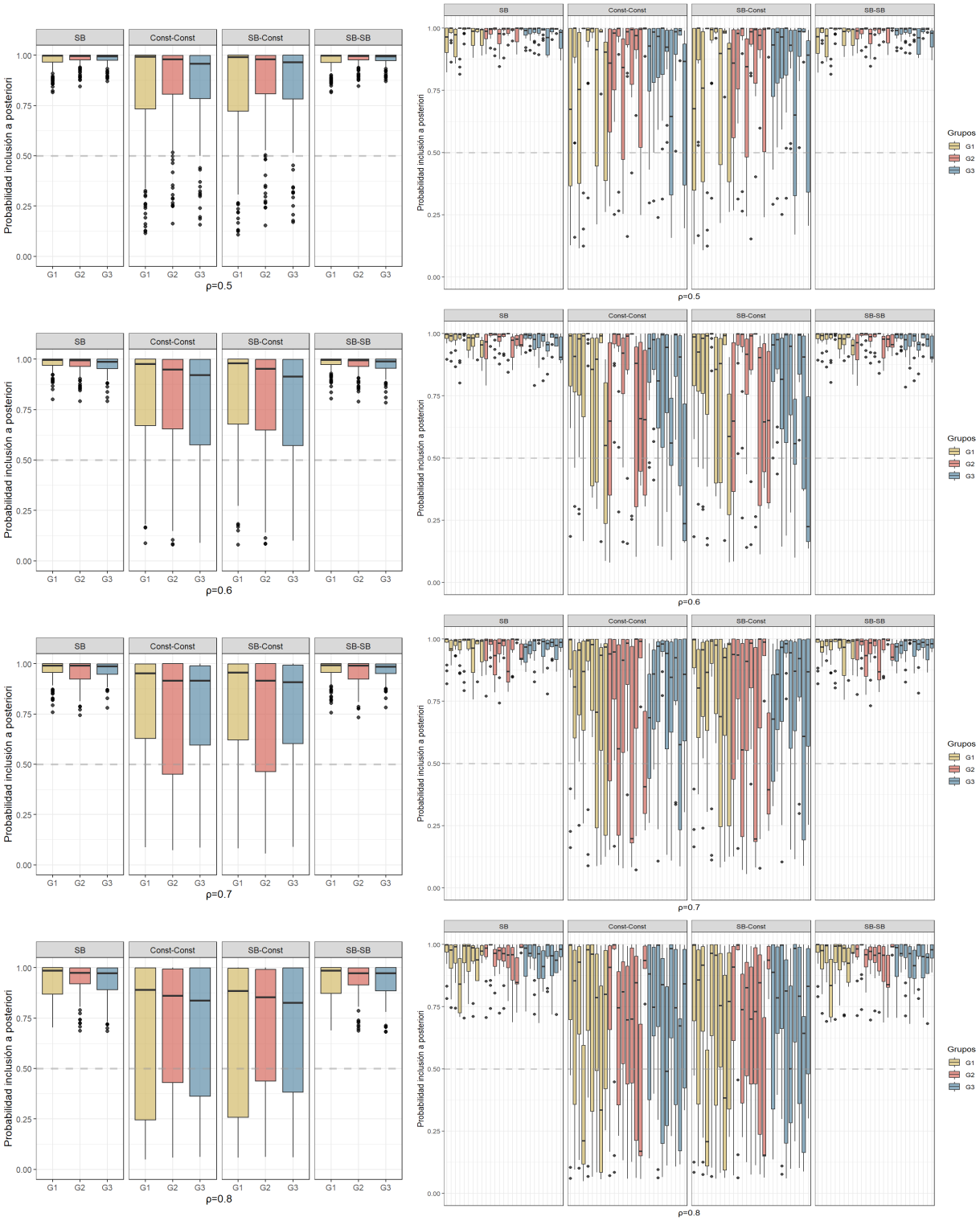


Figura 4.9: Probabilidad de inclusión a posteriori de las variables de los grupos de los 60 bancos de datos, en función de la correlación impuesta para las variables de cada grupo. A la izquierda por grupo y a la derecha por variable.

Ante variables correlacionadas, son las previas *SB* y *SB-SB* las que mejor control de la multiplicidad presenta y mejor desempeño parecen presentar. Cabe plantearse si los resultados obtenidos se mantienen al disminuir el valor de los coeficientes de las variables que generan los datos.

4.4.2. Correlación modificando los coeficientes

El objetivo es contrastar si el grado de correlación entre las variables también afecta a los resultados obtenidos con el método propuesto cuando los parámetros de regresión toman valores más pequeños.

Los datos se simularon como en el experimento anterior, cambiando únicamente los coeficientes de las variables que generan los datos. Estos se simularon esta vez con una normal de media $\mu = 0.2$ y desviación típica $\sigma = 0.1$, 0.15 y 0.2, valores menores que los anteriores.

La probabilidad de inclusión a posteriori de los grupos fue, una vez más, de 1 para todas las previas y coeficiente de correlación considerados de los 60 bancos de datos. En la Figura 4.10 se representan las probabilidades de inclusión a posteriori de las variables de cada grupo de los 60 datos, en función de la correlación elegida.

Para las previas que son combinación de la constante se observa el mismo comportamiento que en el experimento anterior, con unas probabilidades obtenidas menores en general. También se observan resultados peores para *SB* y *SB-SB*, siendo especialmente notorio el efecto de la correlación en la primera. Se aprecia que, no considerando la estructura de grupos en este caso, aumentar el grado de correlación modifica radicalmente las probabilidades obtenidas, empeorando los resultados en gran medida. Sin embargo, con *SB-SB* el efecto de la correlación es menor, con lo que resulta ser la mejor opción una vez más.

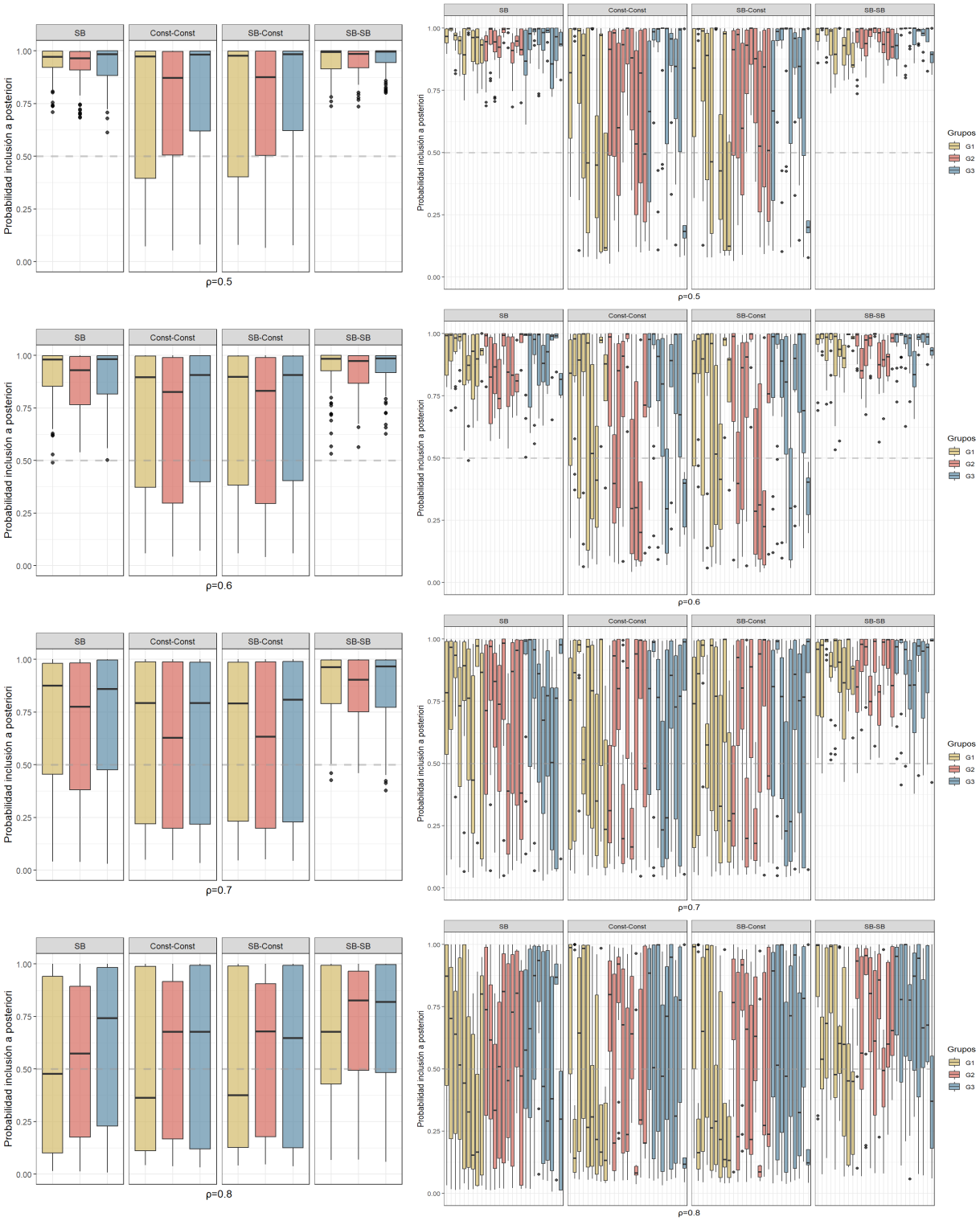


Figura 4.10: Probabilidad de inclusión a posteriori de las variables de los grupos de los 60 bancos de datos, en función de la correlación impuesta para las variables de cada grupo. A la izquierda por grupo y a la derecha por variable.

Hasta ahora, las variables de los grupos tenían la misma correlación y esta tomaba valores lo suficientemente grandes para que tuviera sentido considerar grupos de variables en base a la correlación de las mismas. Sin embargo, cabe plantearse cómo se desenvuelve la metodología propuesta cuando se generan las variables con una correlación entre ellas considerablemente menor.

4.4.3. Correlaciones pequeñas

El objetivo es contrastar si el el método propuesto es eficaz a la hora de seleccionar variables agrupadas con correlaciones pequeñas.

Se simularon 10 bases de datos de un modelo de regresión logística, con $n = 600$ observaciones y $p = 3$ grupos. Cada grupo tiene 20 variables, todas activas, y un coeficiente de correlación fijo. Los parámetros de regresión se simularon con una normal de media $\mu = 0.5$ y desviación típica $\sigma = 0.1$, 0.15 y 0.2 respectivamente, y el intercepto fue fijado a $\beta_0 = 1$. Para evaluar el efecto de una correlación menor en los resultados, se generaron los datos con un coeficiente de correlación $\rho = 0.2$ para las variables de $\mathcal{G}_1$ y $\rho = 0.8$ para las variables de $\mathcal{G}_2$ y $\mathcal{G}_3$. A partir de cada una de las 10 bases simuladas, se formaron 3 bases eligiendo únicamente las primeras 5, 7 y 9 variables de cada grupo respectivamente, haciendo un total de 30 bases a las que realizar la selección.

La probabilidad de inclusión a posteriori de los grupos fue de 1 para todas las previas. En la Figura 4.11 se muestran las probabilidades de inclusión a posteriori obtenidas para las variables de cada grupo de los 30 bancos de datos, en función del número de variables.

Se observan ligeras diferencias con los resultados obtenidos en el experimento 4.0 (Figura 4.8). Las probabilidades obtenidas para las variables del primer grupo son considerablemente mayores y no se ven afectadas por el aumento del número de variables con las previas *SB* y *SB-SB*. Para los grupos con correlaciones altas se observa lo visto en experimentos anteriores.

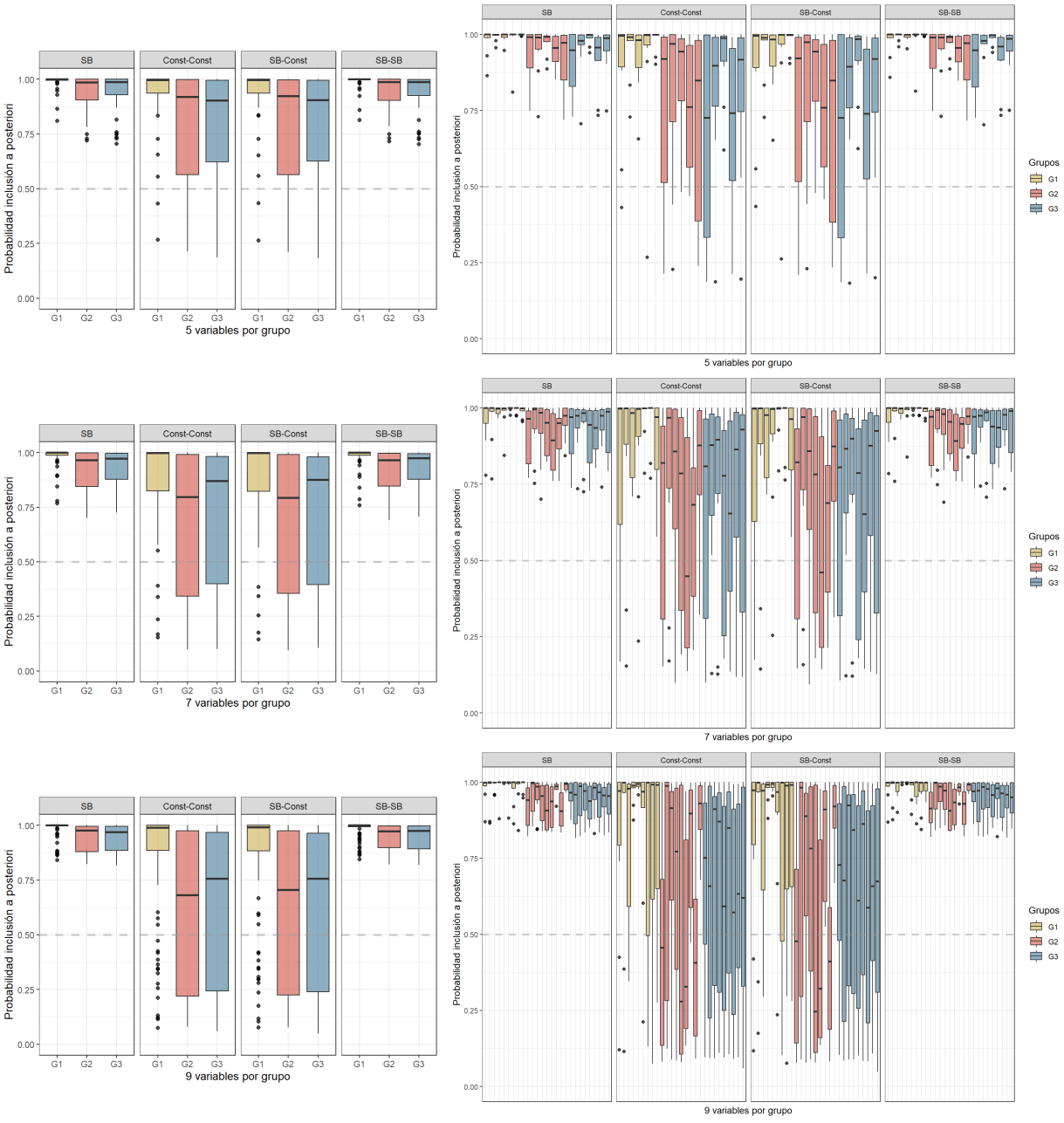


Figura 4.11: Probabilidad de inclusión a posteriori de las variables de los grupos de los 60 bancos de datos, en función del número de variables de cada grupo. A la izquierda por grupo y a la derecha por variable.

Se ha visto cómo afecta la correlación entre las variables de los grupos a la hora de seleccionar variables con la metodología propuesta. Sin embargo, hasta ahora se ha asumido que la determinación de los grupos es correcta y no se ha comprobado la influencia de las variables espurias en situaciones similares. Por tanto, en los dos siguientes experimentos se abordarán dichas cuestiones.

4.4.4. Grupos incorrectamente especificados

El objetivo es contrastar si el el método propuesto es eficaz a la hora de seleccionar variables agrupadas cuando los grupos se han especificado incorrectamente, situación que puede darse fácilmente con datos reales.

Como antes, se simularon 10 bases de datos de un modelo de regresión logística, con $n = 600$ observaciones y $p = 3$ grupos, cada uno con 20 variables activas con coeficiente de correlación $\rho = 0.75$. Los parámetros de regresión se simularon con una normal de media $\mu = 0.5$ y desviación típica $\sigma = 0.1$, 0.15 y 0.2 para cada grupo, y el intercepto fue fijado a $\beta_0 = 1$. Para evaluar el efecto cuando la especificación de los grupos de variables no es correcta se asignaron incorrectamente las dos primeras variables de $\mathcal{G}_2$ y las tres primeras de $\mathcal{G}_3$ a $\mathcal{G}_1$ en la selección de variables. A partir de cada una de las 10 bases simuladas, se formaron 3 bases eligiendo únicamente las primeras 5, 7 y 9 variables de cada grupo respectivamente, haciendo un total de 30 bases a las que realizar la selección.

En la Figura 4.12 se representan las probabilidades de inclusión a posteriori obtenidas para las grupos de los 30 bancos de datos, en función del número de variables. Estas fueron 1 en la mayoría de grupos. Sin embargo, cabe destacar que en el caso de 5 variables por grupo, las probabilidades asociadas al segundo grupo se ven ligeramente reducidas en todas las previas que consideran la estructura de grupo. Esto tiene sentido puesto que al haber menos variables, los aspectos relacionados que refuerzan conjuntamente tienen menos fuerza.

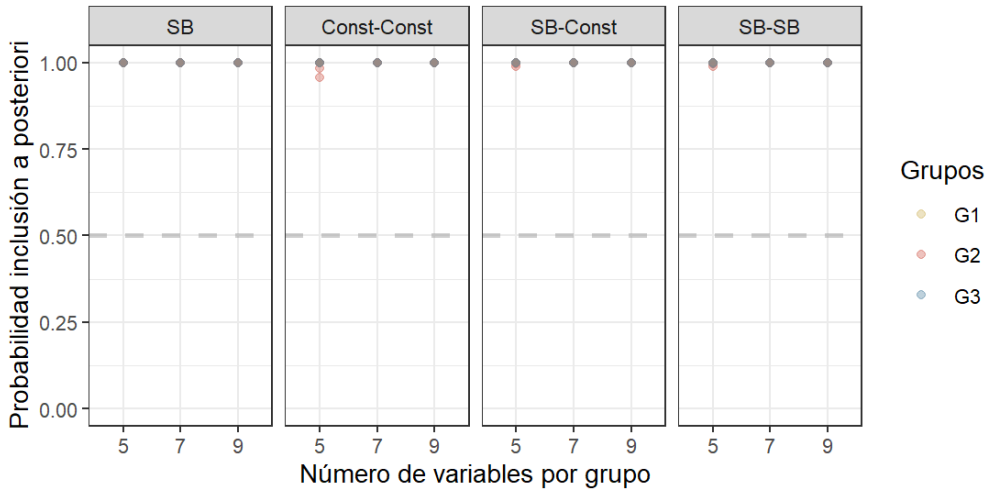


Figura 4.12: Probabilidades de inclusión a posteriori de los grupos de las 30 bases de datos, en función del número de variables.

En la Figura 4.13 se representan las probabilidades de inclusión a posteriori obtenidas para las variables de cada grupo de los 30 bancos de datos, en función del número de variables. En este caso, observamos que la asignación errónea no afecta a los resultados obtenidos, donde se observa el mismo patrón que en experimentos anteriores.

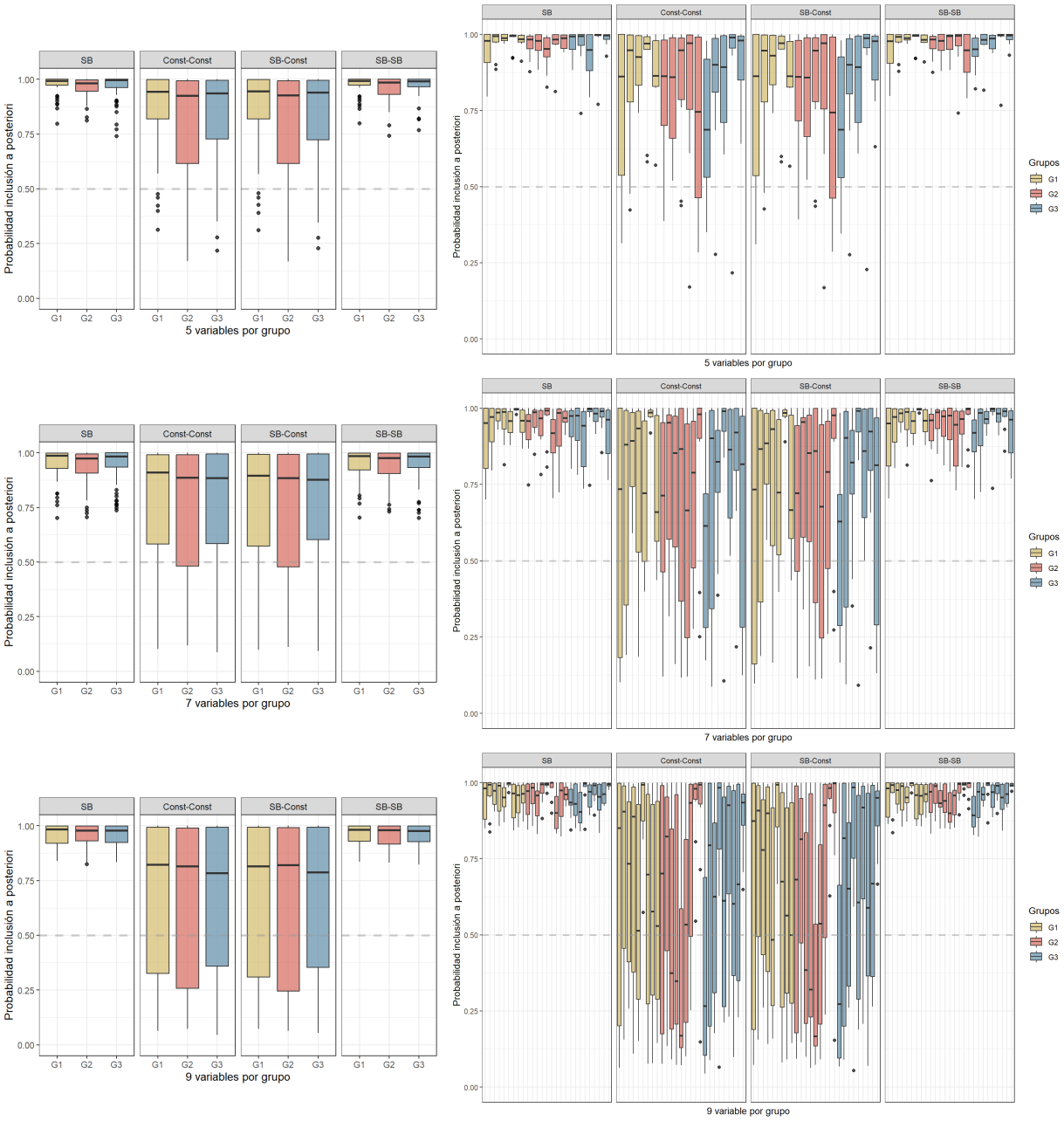


Figura 4.13: Probabilidad de inclusión a posteriori de las variables de los grupos de los 60 bancos de datos, en función del número de variables de cada grupo. A la izquierda por grupo y a la derecha por variable.

4.4.5. Variables espurias

El objetivo es contrastar si el el método propuesto es eficaz a la hora de seleccionar variables agrupadas en presencia de variables que no generan los datos.

Se simularon 5 bases de datos de un modelo de regresión logística, con $n = 600$ observaciones y $p = 3$ grupos, cada uno con 20 variables activas con coeficiente de correlación $\rho = 0.9$. Los parámetros de regresión se simularon con una normal de media $\mu = 0.5$ y desviación típica $\sigma = 0.1, 0.15$ y 0.2 para cada grupo, y el intercepto fue fijado a $\beta_0 = 1$. Para evaluar el efecto de la presencia de variables espurias en este contexto, se introdujo en la selección de variables un cuarto grupo, $\mathcal{G}_4$, que no genera los datos. A partir de cada una de las 5 bases simuladas, se formaron 3 bases eligiendo únicamente las primeras 5, 7 y 9 variables de cada grupo respectivamente, haciendo un total de 30 bases a las que realizar la selección.

En la Figura 4.14 se representan las probabilidades de inclusión a posteriori obtenidas para las grupos de los 15 bancos de datos, en función del número de variables. En todos los casos, se obtiene una probabilidad de 1 de inclusión de los grupos que generan los datos. Sin embargo, se incluye incorrectamente el cuarto grupo en el MPM con las previas que son combinación de la de Scott y Berger. Con *SB-Const* se incluyó erróneamente un $6.67\,\%$ de los grupos espurios, y con *SB-SB* se incluyó erróneamente un $20\,\%$. Cuando no se considera la estructura de grupos, el porcentaje de aquellos incorrectamente seleccionados aumenta hasta el $93.33\,\%$.

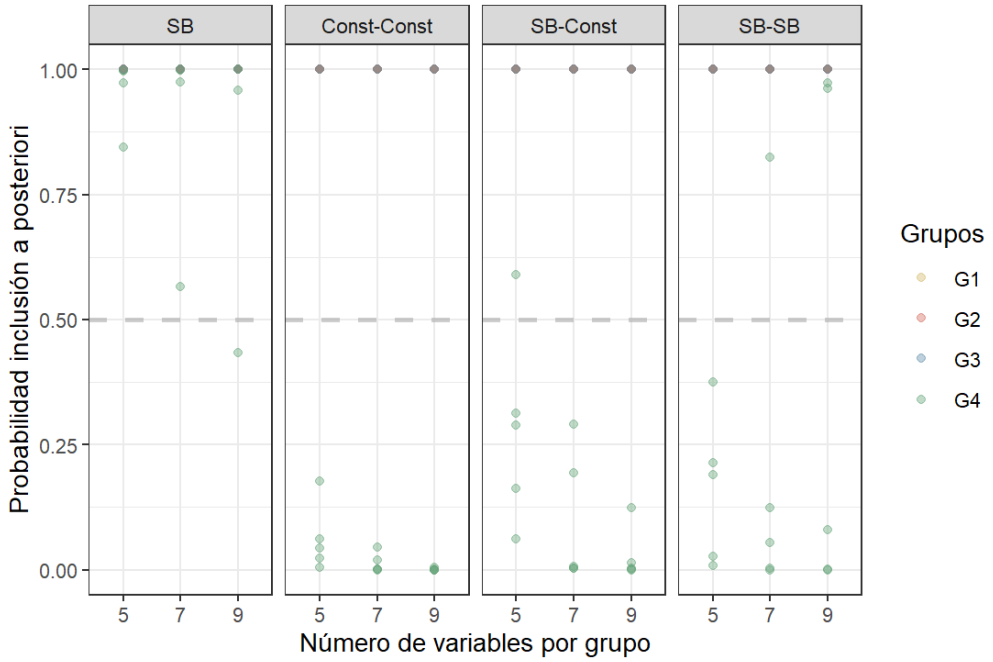


Figura 4.14: Probabilidades de inclusión a posteriori de los grupos de las 15 bases de datos, en función del número de variables.

La proporción de variables agrupadas de las 15 bases de datos incorrectamente no incluidas (falsos negativos) en el MPM se recoge en la Tabla 4.5, y de aquellas incorrectamente incluidas (falsos positivos), en la Tabla 4.6, en función del número de variables.

Una vez más se hace patente que las previas que son combinación de la constate no controlan la multiplici-

dad. Sin embargo, estas producen un menor número de falsos positivos que *SB-SB*. Por otro lado, no considerar la estructura de grupos en la selección de variables produce un alarmante número de falsos positivos, llegando al $80\,\%$. Esto es especialmente interesante puesto que, hasta ahora, se había visto un comportamiento muy similar con *SB-SB* en una situación de estas características.

Tabla 4.5: Proporción de variables incorrectamente no seleccionadas ($\mathcal{G}_1$, $\mathcal{G}_2$ y $\mathcal{G}_3$) de los 15 bancos de datos, en función del número de variables.

Variables incorrectamente no seleccionadas												
	SB			*Const-Const*			*SB-Const*			*SB-SB*		
	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$
5 variables	0	0.04	0.04	0.28	0.24	0.28	0.28	0.24	0.24	0	0	0
7 variables	0.09	0.09	0.06	0.43	0.31	0.46	0.43	0.29	0.46	0	0	0
9 variables	0.11	0.11	0.11	0.51	0.53	0.53	0.51	0.53	0.53	0	0	0

Tabla 4.6: Proporción de variables incorrectamente seleccionadas ($\mathcal{G}_4$) de los 15 bancos de datos, en función del número de variables.

Variables incorrectamente seleccionadas				
	SB	*Const-Const*	*SB-Const*	*SB-SB*
	$\mathcal{G}_4$	$\mathcal{G}_4$	$\mathcal{G}_4$	$\mathcal{G}_4$
5 variables	0.76	0.20	0.04	0
7 variables	0.80	0.23	0	0.2
9 variables	0.80	0.18	0	0.4

Se ha visto la necesidad de considerar la estructura de grupos a través de las previas sobre el espacio de modelos cuando se realiza una selección de variables en presencia de grupos. Además, con dicha estructura, la previa *SB-SB* parece ser la más robusta y eficaz, pese a que en algunas ocasiones también falla.

5. Alzheimer

Hasta ahora, la metodología propuesta se ha trabajado con datos simulados, donde se tenía un control de la situación, para estudiar su funcionamiento. En este capítulo se estudiará su desempeño con datos reales mediante el análisis de los datos de pacientes de Alzheimer del hospital La Fe de Valencia, motivación del trabajo. Como ya se mencionó, el objetivo es encontrar biomarcadores en plasma que diagnostiquen la enfermedad de forma temprana.

5.1. Formulación del problema

Sea $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ la variable respuesta, con

$$Y_i = \begin{cases} 1 & \text{si el individuo i-ésimo padece la enfermedad de Alzheimer (AD),} \\ 0 & \text{en caso contrario (no-AD);} \end{cases}.$$

Así $\boldsymbol{Y} \sim Ber(\boldsymbol{\pi})$ siendo $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^T$ probabilidad de éxito. El modelo de regresión logística completo es

$$logit(\boldsymbol{\pi}) = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\beta}, \tag{5.1}$$

siendo $\boldsymbol{X}$ la matriz de n filas y k columnas, donde la fila i-ésima está formada por los valores de las variables singulares para la observación i-ésima, con $i = 1, \ldots, k$; y $\boldsymbol{Z} = [\boldsymbol{Z}^1 | \ldots | \boldsymbol{Z}^j | \ldots | \boldsymbol{Z}^p]$ es la matriz n filas y L columnas, donde $\boldsymbol{Z}^j = (z_{ir}^{(j)})$ es la submatriz de n filas y l_j columnas tal que $z_{ir}^{(j)}$ es la observación i-ésima de la variable r-ésima del grupo $\mathcal{G}_j$, para $r = 1, \ldots, l_j$ y $j = 1, \ldots, p$. $\boldsymbol{\beta}$ y $\boldsymbol{\alpha}$ son los vectores paramétricos asociados a las variables agrupadas y singulares, respectivamente.

En la base de datos estudiada se tienen $n = 133$ individuos y $m = 38$ variables. Como ya se introdujo en 1, los grupos considerados son $p = 3$: compuestos de peroxidación lipídicos, lípidos y microRNAs; con $l_1 = 22$, $l_2 = 6$ y $l_3 = 8$ variables, respectivamente. De esta manera, se tienen un total de $k = 2$ variables singulares, $L = 36$ variables agrupadas y $2^m = 2^{38} = 274877906944$ modelos competitivos.

5.2. Análisis descriptivo del banco de datos

Entre las variables explicativas se tienen variables clínicas y variables agrupadas según el criterio clínico en: compuestos de peroxidación lipídicos, lípidos y microRNAs. El análisis descriptivo numérico queda recogido en las Tablas 5.1, 5.2, 5.3 y 5.4, respectivamente. Para describir el comportamiento de las variables agrupadas se han dibujado las Figuras 5.1, 5.4 y 5.6 para cada grupo de variables respectivamente, en función de si hay o no AD.

Tabla 5.1: Variables clínicas y demográficas recogidas en los pacientes.

Variable	Grupo AD (n=64)	Grupo no-AD (n=69)
Género (Femenino, n(%))	$34\ (53.1\,\%)$	$41\ (59.4\,\%)$
Edad (Años, mediana (IQR))	$70\ (66-73.3)$	$64\ (61-68)$

Tabla 5.2: Niveles en plasma de los compuestos de peroxidación lipídicos recogidos en los pacientes.

Biomarcador (nmol/L)	Grupo AD (n=64)		Grupo no-AD (n=69)	
	Mediana	IQR	Mediana	IQR
IsoF tot	0.16	(0.10, 0.27)	0.42	(0.31, 0.57)
NeuroF tot	0.23	(0.14, 0.43)	0.26	(0.13, 0.37)
ADT-420	0.01	(0, 0.21)	0	(0, 0.22)
ADT-207	0	(0, 0)	0	(0, 0)
CO5-776	0	(0, 0)	0	(0, 0)
CO5-778	0	(0, 0)	0	(0, 0)
CO5-769	0.05	(0, 0.15)	0	(0, 0.2)
IsoP tot	0.35	(0.27, 0.52)	1.23	(0.73, 1.55)
CO1-31	0.89	(0, 1.77)	1.3	(0.57, 1.90)
AG495m4R	0.16	(0.02, 0.27)	0.17	(0.10, 0.27)
NeuroP tot	0.00	(0, 0.12)	0.39	(0, 0.65)
1a1bdihomo	0	(0, 0)	2.73	(0, 4.08)
VB559m1	1.18	(0.263, 1.51)	3	(1.30, 4.05)
PGF2a	0.44	(0.21, 0.69)	0.50	(0.27, 0.82)
8-iso-PGF2a	0.01	(0, 0.05)	0.05	(0.02, 0.07)
5-iPF2a-VI	1.05	(0.39, 1.43)	2.13	(1.20, 2.98)
8-iso-PGE2	0.69	(0.24, 1.98)	1	(0.60, 1.58)
8-iso-15-keto-PGF2a	0.27	(0.15, 0.38)	0.5	(0.30, 0.82)
8-iso-15-keto-PGE2	0	(0, 0.37)	0.72	(0, 1.15)
2,3-dinor-iPF2a	0	(0, 0.02)	0	(0, 0.02)
PGE2	0.07	(0, 0.41)	0.32	(0.25, 0.45)
8-iso-15R-PGF2a	0.36	(0.26, 0.53)	0.50	(0.25, 0.67)

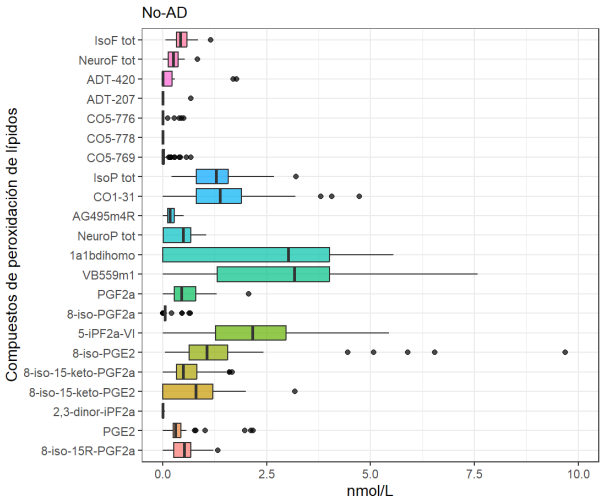


Figura 5.1: Diagrama de cajas para los compuestos de peroxidación lipídicos medidos en los pacientes sin AD.

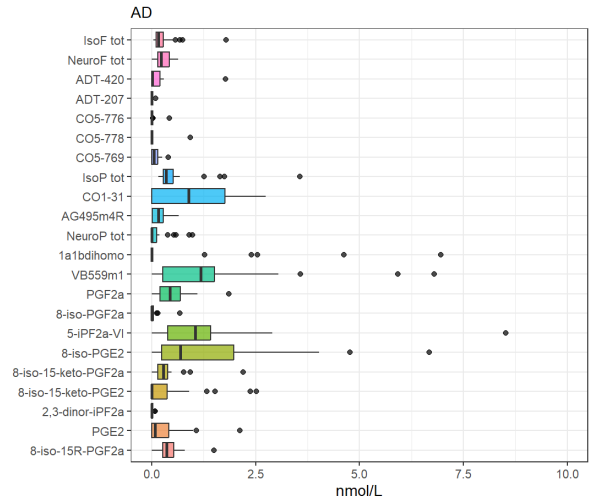


Figura 5.2: Diagrama de cajas para los compuestos de peroxidación lipídicos medidos en los pacientes con AD.

Tabla 5.3: Niveles en plasma de los lípidos recogidos en los pacientes.

Biomarcador (μ**g m/L**)	Grupo AD (n=64)		Grupo no-AD (n=69)	
	Mediana	IQR	Mediana	IQR
18:0 SM	74.6	(48.7, 105)	69.4	(50.9, 101)
DOPE	3.87	(1.44, 11.5)	7.33	(3.89, 11.8)
16:0 SM	159	(108, 206)	149	(116, 196)
16:1 SM	21.2	(15.1, 30.1)	20.9	(13.2, 28.5)
18:0 LPC	34.9	(22.8, 65.8)	48.8	(23.7, 75.1)
18:1 LPE	1.78	(1.14, 2.64)	1.45	(0.77, 1.85)

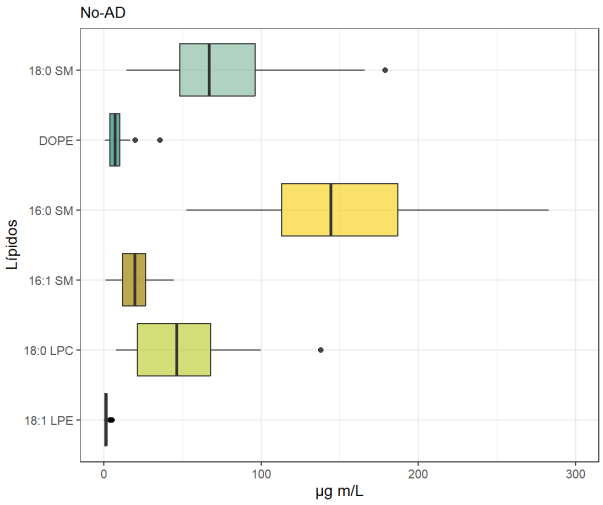


Figura 5.3: Diagrama de cajas para los lípidos medidos en los pacientes sin AD.

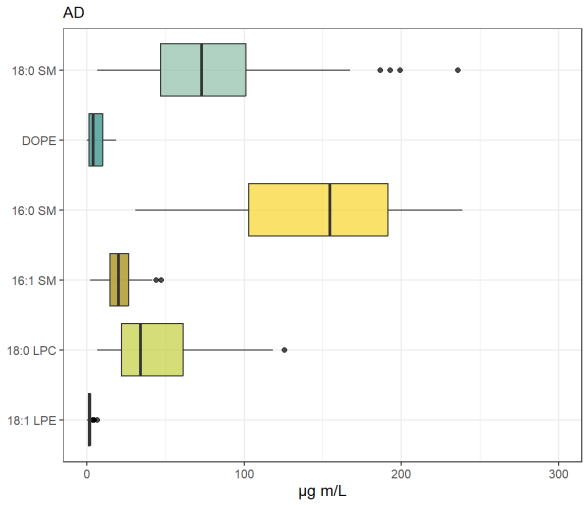


Figura 5.4: Diagrama de cajas para los lípidos medidos en los pacientes con AD.

Tabla 5.4: Niveles en plasma de los MicroRNAs recogidos en los pacientes.

Biomarcador (conteos normalizados)	Grupo AD (n=64)		Grupo no-AD (n=69)	
	Mediana	IQR	Mediana	IQR
miR-342-3p	1.21	(0.55, 2.39)	0.92	(0.65, 1.13)
miR-483-3p	1.06	(0.62, 1.59)	0.92	(0.62, 1.36)
miR-320b	1.14	(0.83, 1.66)	0.91	(0.57, 1.65)
miR-150-5p	1.28	(0.62, 1.82)	0.99	(0.60, 1.26)
miR-486-5p	0.92	(0.70, 1.82)	1.06	(0.66, 1.49)
miR-29a-3p	0.88	(0.55, 1.25)	1.07	(0.81, 2.12)
miR-486-3p	1.33	(0.54, 2.12)	0.99	(0.71, 1.92)
miR-92a-3p	1.28	(0.67, 1.62)	0.83	(0.63, 1.92)

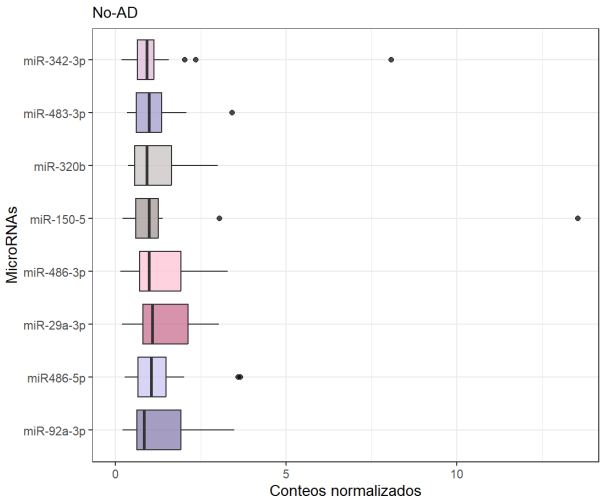


Figura 5.5: Diagrama de cajas para los MicroRNAs medidos en los pacientes sin AD.

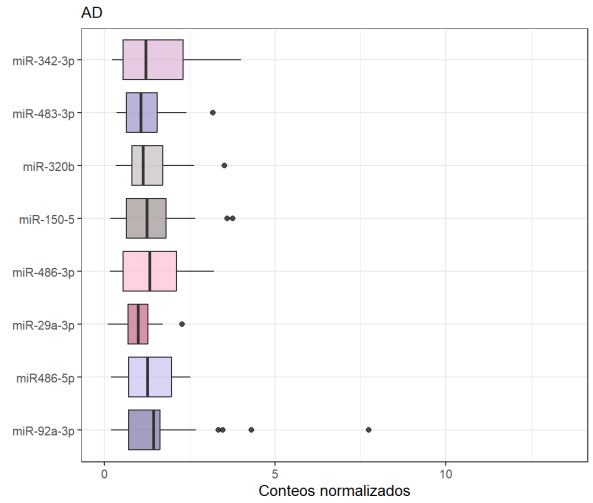


Figura 5.6: Diagrama de cajas para los MicroRNAs medidos en los pacientes con AD.

5.2.1. Estudio correlaciones

Las elipses tienen un parámetro de excentricidad escalado por el coeficiente de correlación correspondiente.

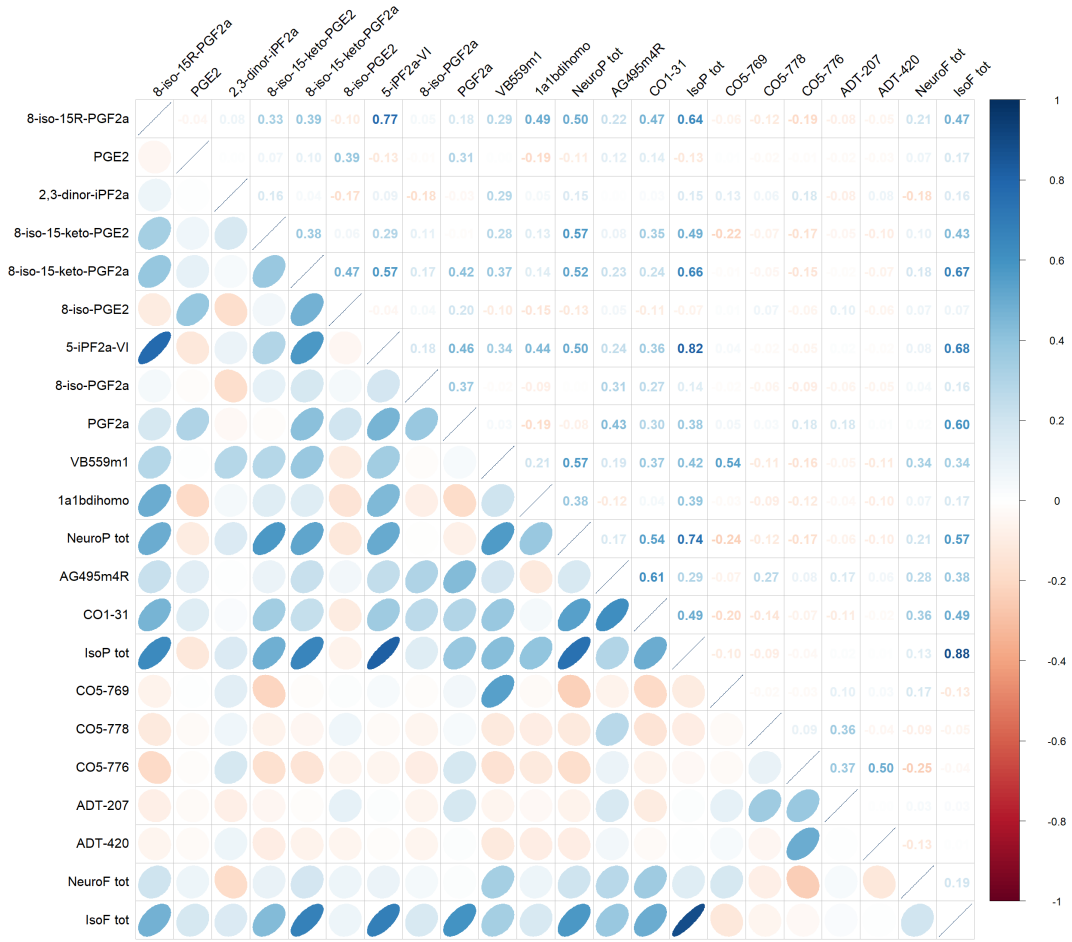


Figura 5.7: Gráfico de correlaciones para los compuestos de peroxidación lipídicos.

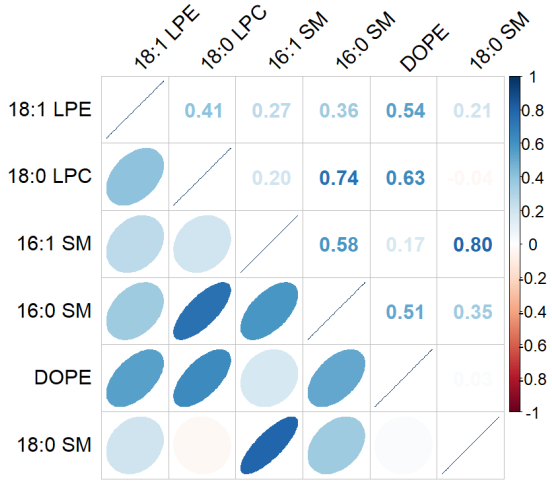


Figura 5.8: Gráfico de correlaciones para los lípidos.

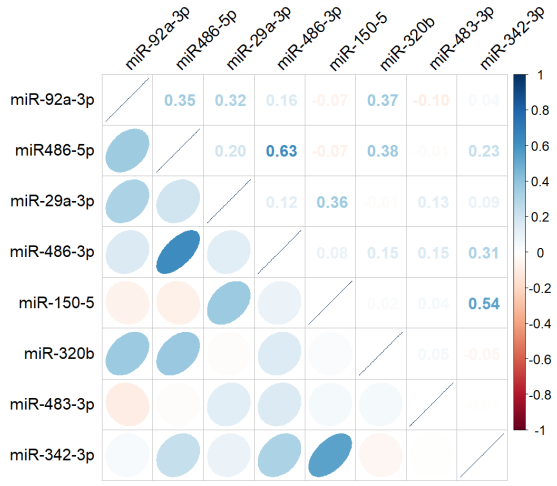


Figura 5.9: Gráfico de correlaciones para los microRNAs.

Las correlaciones obtenidas son todas mayores que -0.25, mayoritariamente positivas. Se observan muchas correlaciones altas, hecho que corrobora la presencia de relaciones entre las variables. Estas agrupaciones vienen dadas por el juicio médico.

5.3. Selección de variables en presencia de grupos

Antes de realizar la selección de variables, se imputó un banco de datos mediante la librería *mice* para paliar la gran cantidad de datos faltantes de la base de datos (el $34.40\,\%$).

Las probabilidades de inclusión a posteriori para las variables singulares y agrupadas obtenidas considerando la estructura de grupos con la previa *SB-SB* se muestran en las Tablas 5.5, 5.6, 5.8 y 5.7.

Tabla 5.5: Probabilidades de inclusión a posteriori de los CPL.

Compuestos de peroxidación	
Probabilidad del grupo	0.98
IsoF tot	0.99
NeuroF tot	0.98
ADT-420	0.97
ADT-207	0.98
CO5-776	0.97
CO5-778	0.98
CO5-769	0.99
IsoP tot	0.99
CO1-31	0.99
AG495m4R	0.99
NeuroP tot	0.98
1a1bdihomo	0.97
VB559m1	0.97
PGF2a	0.99
8-iso-PGF2a	0.98
5-iPF2a-VI	0.98
8-iso-PGE2	0.97
8-iso-15-keto-PGF2a	0.98
8-iso-15-keto-PGE2	0.98
2,3-dinor-iPF2a	0.98
PGE2	0.99
8-iso-15R-PGF2a	0.98

Tabla 5.6: Probabilidades de inclusión a posteriori de los lípidos.

Lípidos	
Probabilidad del grupo	0.98
18:0 SM	0.98
DOPE	0.99
16:0 SM	0.97
16:1 SM	0.98
18:0 LPC	0.99
18:1 LPE	0.99

Tabla 5.7: Probabilidades de inclusión a posteriori de las variables clínicas.

Variables singulares	
Edad	1
Genero	0.90

Tabla 5.8: Probabilidades de inclusión a posteriori de los MicroRNAs.

MicroRNAs	
Probabilidad del grupo	0.98
miR-342-3p	0.99
miR-483-3p	0.99
miR-320b	0.97
miR-150-5p	0.98
miR-486-5p	0.97
miR-29a-3p	0.99
miR-486-3p	0.97
miR-92a-3p	0.99

Realizando la selección de variables con la estructura de grupo a través de la previa *SB-SB*, obtenemos que

el MPM y el HPM coinciden, siendo este último el modelo con todas las variables singulares y agrupadas. Este resultado difiere en gran medida del obtenido por Forte *et al.* (2024), donde con la librería BAS obtuvo únicamente una variable con una probabilidad de inclusión a posteriori mayor que 0.5. Sin embargo, ellos no introdujeron los MicroRNAs en la selección de variables bayesiana ni imputaron los valores faltantes, por lo que es difícil realizar una comparación.

Los resultados obtenidos con la previa *SB-Const* son muy diferentes, cuyo MPM está formado por la mitad de las variables, como era de esperar.

6. Conclusiones

A diferencia de la selección de variables tradicional, donde tener un gran número de variables relacionadas se ve como una complicación añadida a la selección; considerar la estructura de grupos refuerza la contribución de dichas variables. Esto se debe a que, especificando una estructura de grupos, las variables relacionadas no compiten entre ellas, sino que se fortalecen unas a otras para hacer su influencia más clara. El hecho de tener variables relacionadas ya no implica que únicamente una sea necesaria. Considerar dicha información antes de que el proceso de selección empiece marca la diferencia, captando aspectos fuertemente relacionados de las mismas.

De entre las previas sobre el espacio de modelos que introducen la estructura de grupos, resulta más robusta y fiable *SB-SB*, en base a las simulaciones realizadas. Además, esta proporciona resultados más precisos cuando se compara con otras metodologías más tradicionales. No obstante, debido al elevado tiempo de computación de la función desarrollada, el número de simulaciones no fue alto y sería necesario un estudio más extenso para conseguir resultados más concluyentes.

En el análisis de los datos de Alzheimer se obtuvieron unas probabilidades muy próximas a 1 para todos los grupos y variables considerados. Sin embargo, estos resultados pueden no ser muy precisos puesto que se imputaron más de un $30\,\%$ de los datos, debido al gran número de valores faltantes. Como trabajo a futuro sería necesario realizar una imputación mejor que nos permita trabajar con dicha base de datos de forma más fiable.

En cuanto al código de la metodología desarrollada, sería necesario optimizarlo para realizar una investigación más precisa. Además, sería posible extenderlo a variables categóricas siguiendo el trabajo de García-Donato y Paulo (2021).

Bibliografía

Barbieri, M. M. y Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics, Ann. Statist*, 32(3):870–897.

Barbieri, M. M., Berger, J. O., George, E. I., y Rocková, V. (2021). The Median Probability Model and Correlated Variables. *Bayesian Analysis*, 16(4):1085–1112.

Bayarri, M. J., Berger, J. O., Forte, A., y García-Donato, G. (2012). Criteria for bayesian model choice with application to variable selection. *Annals of Statistics*, 40(3):1550–1577.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1:385–402.

Breheny, P., a. H. J. (2009). Penalized Methods for Bi-Level Variable Selection. *Statistics and Its Interface*, 2:369–380.

Clyde, M. A., Ghosh, J., y Littman, M. L. (2011). Bayesian Adaptive Sampling for Variable Selection and Model Averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.

Fan, J. y Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, (96):1348–1360.

Fernández, C., Ley, E., y Steel, M. (2002). Bayesian Modelling of Catch in a North-West Atlantic Fishery. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 51(3):257–280.

Forte, A., Lara, S., Peña-Bautista, C., Baquero, M., y Cháfer-Pericás, C. (2024). New approach for early and specific Alzheimer disease diagnosis from different plasma biomarkers. *Clin Chim Acta*.

Garcia-Donato, G. y Forte, A. (2018). Bayesian testing, variable selection and model averaging in linear models using R with BayesVarSel. *The R Journal*, 10(1):329.

García-Donato, G. y Paulo, R. (2021). Variable Selection in the Presence of Factors: A Model Selection Perspective. *Journal of the American Statistical Association*, 0(0):1–11.

George, E. y McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Hastie, T., Tibshirani, R., y Wainwright, M. (2015). *Statistical Learning with Sparsity The Lasso and Generalizations*. Monographs on Statistics and Applied Probability, 143. Chapman and Hall.

Huang, J. y Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics, Ann. Statist*, 38(4):1978–2004.

Janeiro, M. H., Ardanaz, C. G., Sola-Sevilla, N., Dong, J., C.-E. M., Solas, M., Puerta, E., y Ramírez, M. J. (2020). Biomarkers in Alzheimer's disease. *Advances in Laboratory Medicine*, 2(1):27–50.

Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222.

Kass, R. E. y Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Li, Y. y Clyde, M. A. (2018). Mixtures of g-Priors in Generalized Linear Models. *Journal of the American Statistical Association*, 113(524):1828–1845.

Liang, F., P.-R. M. G. C. M. y Berger, J. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Scott, J. G. y Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, (58):267–288.

Yian, M. y Lin, Y. (2006). Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society, Series B*, (68):49–67.

Zellner, A. (1986). *On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions. In Goel, P.; Zellner, A. (eds.). Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics and Statistics.* Vol. 6. New York: Elsevier.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, (38):894–942.

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. *The Annals of Statistics*, (67):301–320.

Anexo

A. Resultados de interés

A.1 Prueba de la proposición 4.2.1

Proposición 4.2.1. *Sea* $[\mathbf{1}|\boldsymbol{X}|\boldsymbol{Z}]$ *la matriz de diseño del modelo completo de la Ecuación 5.1, de rango* $1+k+L$*, con* $n > 1+k+L$ *y* $L = \sum_{j=1}^{p} l_j$ *el número de variables agrupadas. Sea el modelo* $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}$*, entonces:*

1. $\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta}) = \mathbf{1}^T\boldsymbol{\gamma} + \sum_{j=1}^{p} \mathbf{1}^T\boldsymbol{\delta}_j \in [0,\ k+L]$*, siendo 0 la dimensión del modelo nulo y* $L+k$ *la del modelo completo; es el rango de la matriz de diseño del modelo* $(\boldsymbol{\gamma}, \boldsymbol{\delta})$ *menos uno.*
2. $\mathcal{F}_k^{l_1,\ldots,l_p}(r) = \sum_{0\leq i\leq k,\ 0\leq j_q\leq l_q,\ 1\leq q\leq p,\ i+\sum_{q=1}^{p} j_q = r} \binom{k}{i} \cdot \prod_{q=1}^{p} \binom{l_q}{j_q}$*, con* $r \in [0,\ k+L]$*, es el número de modelos de* $\mathcal{M}$ *de dimensión* r *con* k *variables singulares y* p *grupos de* $l_1, \ldots, l_p$ *variables cada uno.*

Demostración: Los resultados 1 y 2 siguen de que $\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta}) + 1$ es el número de vectores linealmente independientes en:

$$\{\mathbf{1}|\gamma_1\boldsymbol{x}_1|\ldots|\gamma_k\boldsymbol{x}_k|\delta_{11}\boldsymbol{z}_{11}|\ldots|\delta_{1l_1}\boldsymbol{z}_{1l_1}|\ldots|\delta_{p1}\boldsymbol{z}_{p1}|\ldots|\delta_{pl_p}\boldsymbol{z}_{pl_p}\}.$$

El resultado 3 es una expresión multi-binomial estándar que define el número de maneras diferentes que hay de elegir i elementos de un conjunto de k y j_q elementos de un conjunto de l_q, para $j \in \{1, \ldots, p\}$.

□

A.2 Prueba de la proposición 4.2.2

Proposición 4.2.2. *Sea* $[\mathbf{1}|\boldsymbol{X}|\boldsymbol{Z}]$ *la matriz de diseño del modelo completo de la Ecuación 5.1, de rango* $1+k+L$*, con* $n > 1+k+L$ *y* $L = \sum_{j=1}^{p} l_j$ *el número de variables agrupadas. Sea el modelo* $(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in \mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$*, entonces:*

1. $|\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})| = \prod_{j=1}^{p} (2^{\tau_j l_j} - \tau_j)$ *es el número de elementos de* $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$.
2. $\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta})$ *es el número de variables activas, tanto singulares como agrupadas, presentes en el modelo* $(\boldsymbol{\gamma}, \boldsymbol{\delta})$*. Se tiene que* $\kappa(\boldsymbol{\gamma}, \boldsymbol{\delta}) \in [\mathbf{1}^T\boldsymbol{\gamma} + \mathbf{1}^T\boldsymbol{\tau},\ \mathbf{1}^T\boldsymbol{\gamma} + \sum_{j=1}^{p} \tau_j \cdot l_j]$.
3. $\mathcal{G}^{l_{j_1},\ldots,l_{j_{m_2}}}(r) = \sum_{1\leq j_1\leq l_1,\ \ldots,\ 1\leq j_{m_2}\leq l_{j_{m_2}},\ \sum_{q=1}^{m_2} j_q = r} \prod_{q=1}^{m_2} \binom{l_q}{j_q}$ *es el número de modelos de* $\mathcal{M}(\boldsymbol{\gamma}, \boldsymbol{\tau})$ *de dimensión* $r + \mathbf{1}^T\boldsymbol{\gamma}$*, con* $r \in [m_2,\ l_{j_1} + \cdots + l_{j_{m_2}}]$.

Demostración: Análoga a la de la Proposición 4.2.1.

□