UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

$M-$estimadores penalizados para regresión logística

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Ciencias Matemáticas

Gonzalo Chebi

Director de tesis: Ana Bianco

Director adjunto: Graciela Boente

Lugar de trabajo: Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Fecha de defensa: 20 de diciembre de 2019

$M-$estimadores penalizados para regresión logística

Resumen

El modelo de regresión logística es ampliamente utilizado en problemas de clasificación cuando se tienen covariables que permiten explicar la pertenencia a alguno de los dos grupos en consideración. En estos modelos asegurar una buena clasificación e identificar variables con capacidad predictora es de suma importancia. En particular, el problema de selección de variables es relevante cuando el vector de coeficientes de regresión es ralo, es decir, cuando en el modelo verdadero unas pocas covariables son suficientes para poder predecir la variable respuesta. En el modelo de regresión lineal, un método efectivo para estimar modelos ralos consiste en agregar un término de penalización a la suma de cuadrados de los residuos a minimizar.

En esta tesis, se aborda el problema de estimación y selección de variables en el modelo de regresión logística ralo mediante métodos robustos que resisten la presencia de datos atípicos. Más precisamente, consideramos un modelo de regresión logística en el cual se observan p covariables, pero sólo hay un número k (desconocido) de variables explicativas activas que se desean identificar. Además de seleccionar variables, nuestro interés consiste en proveer inferencias estables cuando existe en la muestra un pequeño porcentaje de observaciones mal clasificadas que, si además corresponden a puntos de alta palanca, pueden tener una gran influencia. Para resolver estos problemas, se consideran versiones penalizadas y pesadas de los estimadores propuestos por Bianco y Yohai (1996). Por un lado, se muestra que la familia de pérdidas introducidas en dicho trabajo incluye otros estimadores considerados en la literatura. Por otra parte, se considera una amplia variedad de funciones de penalización y se propone la llamada penalidad Signo, que mejora sustancialmente el sesgo introducido por penalizaciones como Ridge o LASSO.

Bajo condiciones de regularidad, se obtienen resultados de consistencia y expresiones para la distribución asintótica de los estimadores propuestos. Se deducen además resultados que aseguran que los estimadores seleccionan variables de manera consistente. Se analiza por separado el caso en que la cantidad de covariables p es fija y cuando p diverge a infinito junto con el tamaño de la muestra n. Específicamente, en este último escenario mostramos que, bajo ciertas condiciones sobre la distribución de covariables y la penalidad utilizada, los estimadores propuestos son consistentes si $p/n \to 0$ y tienen la llamada propiedad oráculo si $pk/n \to 0$, donde k es la cantidad de covariables activas en el modelo de regresión logística.

Se propone un algoritmo que permite encontrar una solución aproximada de los problemas de minimización para las funciones de pérdida y penalización consideradas en la tesis. Se define además un procedimiento de convalidación cruzada robusto para elegir el parámetro de regularidad. Un extenso estudio de simulación permite investigar, para muestras finitas, el desempeño de los estimadores propuestos para distintas elecciones tanto de la función de pérdida como de la penalidad para conjuntos de observaciones con datos atípicos y sin ellos. En particular, los $M-$estimadores pesados con penalizaciones acotadas muestran sus ventajas bajo los diferentes esquemas de contaminación considerados. Finalmente, se aplican los métodos propuestos en esta tesis a conjuntos de datos reales.

Palabras clave: Clasificación, $M-$estimadores, Penalización, Regresión Logística, Robustez.

Penalized $M-$estimators for logistic regression

Abstract

The logistic regression model is widely used in classification problems where explanatory covariates with capability to explain the group membership are available. For these models, ensuring good classification properties and selecting a subset of variables with high prediction ability is a fundamental task. In particular, variable selection is specially important when the true underlying model has a sparse representation, i.e., when only a few explanatory variables are enough to predict the response variable. In the linear regression model, an effective method to estimate sparse models is to add a suitable penalization term to the residuals sum of squares that is minimized.

In this thesis, we address the problem of estimating and selecting variables under a sparse logistic regression model through methods that are robust against the presence of outliers. To be more precise, we consider logistic regression models in which p covariates are observed, but only k of them are active. Both the quantity k and the subset of active covariates are unknown and need to be estimated. Besides selecting variables, we aim to provide stable procedures against a small proportion of observations wrongly classified. In particular, these observations may be extremely harmful when they correspond to high leverage points. To solve these problems, we consider penalized and weighted versions of the estimators proposed by Bianco and Yohai (1996). On the one hand, we show that the family of loss functions introduced in that paper includes other estimators in the literature. On the other one, we consider a wide range of penalization functions and we propose the so called Sign Penalty, which substantially improves the bias introduced by popular penalizations such as Ridge or LASSO.

Under regularity conditions, we obtain consistency results and arrive to expressions for the asymptotic distribution of the proposed estimators. Moreover, we derive results ensuring that these estimators perform variable selection consistently. We separately analyse the case where the number of covariates p is fixed and the situation where p diverges to infinity with the sample size n. More precisely, in the latter scenario, we show that, under mild assumptions for the covariate distribution and the penalization function, the proposed estimators are consistent if $p/n \to 0$ and have the oracle property if $pk/n \to 0$, where k is the number of active covariates in the true logistic regression model.

We propose an algorithm that allows to find an approximate solution of the minimization problem, for the loss and penalty functions considered here. Moreover, we define a robust cross-validation procedure to select the tuning parameter. An extensive numerical study allows to investigate the performance of the proposed estimators for different loss and penalty choices. We consider the case of clean samples following a logistic regression model and also that the situation where misclassified data are added according to different contamination scenarios. In particular, the obtained results show the advantages of using weighted $M-$estimators combined with bounded penalty functions, under the considered outlier schemes. Finally, the proposed methods are illustrated on some real data.

Keywords: Classification, $M-$estimators, Penalization, Logistic Regression, Robustness.

Índice general

Capítulo 1

Introducción

En las últimas décadas, los avances tecnológicos en diversas áreas de la ciencia permitieron obtener, almacenar y procesar una gran cantidad de datos. Junto con estos avances, surgieron muchos problemas estadísticos de alto impacto. Por ejemplo, en medicina se realizan estudios genéticos para determinar el tratamiento óptimo para una determinada enfermedad y las campañas políticas desarrollan modelos estadísticos para determinar las mejores decisiones con el fin de ganar una elección. Además, las plataformas de transmisión online de películas y series desarrollan sistemas de recomendación en base a las opiniones de los consumidores, mientras que las redes sociales mejoran la experiencia de sus usuarios en base a información sobre los miembros y sus contactos. La Figura 1.1 muestra la increíble cantidad de datos que Internet genera cada minuto. Más aún, se puede ver que estos números están creciendo considerablemente, motivando nuevos problemas estadísticos altamente desafiantes. Sin embargo, muchos de estos problemas no pueden ser abordados con métodos estadísticos clásicos, por lo que es necesario desarrollar nuevas técnicas. Estos nuevos contextos que involucran una gran cantidad de datos suelen ser encasillados como problemas de *Big Data*.

Dentro de este nuevo paradigma, son muy frecuentes las situaciones en donde se cuenta con un conjunto de observaciones cuyo tamaño muestral y dimensión crecen, pero que frecuentemente incluyen variables o información superflua. Un ejemplo típico corresponde a datos del genoma humano, donde es posible que cada observación tenga miles de elementos genéticos y la cantidad de individuos de la muestra sea de tan sólo unos cientos, sin embargo, solo unos pocos elementos son relevantes a la hora de predecir una determinada enfermedad. En este contexto, los métodos estadísticos clásicos no son aplicables en general y necesitan ser adaptados para este nuevo escenario. Efron y Hastie (2016) hacen un excelente resumen sobre algunos métodos de inferencia adaptados al contexto de alta dimensión.

En particular, en los modelos de regresión, los datos de alta dimensión pueden provocar nuevos problemas. Esto se debe, en muchos casos, a que las estimaciones tradicionales tienden a producir un sobreajuste de los datos e incluso pueden no estar bien definidas cuando el tamaño de la muestra es menor a la dimensión de cada uno de sus elementos. Más aún, los algoritmos de minimización se pueden volver inestables. Por otro lado, muchas veces, además de contar con una buena predicción, es de interés que el modelo resutante sea *interpretable*, es decir, encontrar un subconjunto de covaria-

bles que permita realizar una buena estimación de la variable respuesta. Los métodos tradicionales que fueron desarrollados bajo el paradigma de que el número de observaciones, n, es mucho mayor que la cantidad de covariables, p, necesitan ser adaptados a este contexto ya que el vector de coeficientes estimados resultante típicamente tiene todas sus coordenadas distintas de cero. Al no haber una forma sencilla de seleccionar un subconjunto pequeño de variables explicativas para predecir la variable respuesta, la interpretabilidad es difícil en estos casos. La selección efectiva y automática de variables en regresión es, por lo tanto, necesaria para mejorar la interpretación y la precisión de las técnicas utilizadas, ya que elegir subconjuntos de variables muy pequeños puede llevar a una mala especificación del modelo, mientras que demasiadas variables agrava el problema conocido como *maldición de la dimensión*. Por esta razón, especialmente en modelos de regresión es de suma importancia seleccionar el correcto subconjunto de variables excluyendo aquellas innecesarias.

Dada la importancia y popularidad de los modelos de regresión, los métodos tradicionales necesitan ser adaptados. Para superar las nuevas dificultades provocadas por los datos de alta dimensión, es usual asumir que el vector verdadero de coeficientes de la regresión es ralo. Esto significa asumir que en el modelo verdadero unas pocas covariables son suficientes para poder predecir la variable respuesta, ver por ejemplo Tibshirani *et al.* (2015) para más detalles. Cabe señalar que, más alla de que matemáticamente éste parezca ser un supuesto fuerte, este tipo de enfoque ha tenido muy buenos resultados en el análisis de datos en las últimas décadas. Esto se debe, posiblemente, a que estos nuevos métodos, al asumir este supuesto, producen buenas aproximaciones ralas de los verdaderos modelos subyacentes para cada caso. El estimador de máxima verosimilitud (EMV) o el de mínimos cuadrados son los métodos clásicos para estimar el vector de coeficientes en modelos de regresión y regresión generalizada. Esta técnica es altamente eficiente cuando la dimensión es pequeña en comparación con el número de observaciones. Sin embargo, no produce estimadores ralos, es decir, los estimadores tienen todas sus componentes no nulas no permitiendo hacer selección de variables. En el modelo de regresión lineal, un método efectivo para estimar modelos ralos consiste en agregar un término de penalización a la suma de cuadrados de los residuos a minimizar.

El modelo de regresión logística es ampliamente utilizado en problemas de clasificación cuando se tienen covariables que permiten explicar la pertenencia a alguno de los dos grupos en consideración. En estos modelos asegurar una buena clasificación e identificar variables con capacidad predictora es de suma importancia. En particular, como en el modelo lineal, el problema de selección de variables es relevante cuando el vector de coeficientes de regresión es ralo, es decir, cuando en el modelo verdadero unas pocas covariables son suficientes para poder predecir la pertenencia al grupo. Por estas razones, uno de nuestros objetivos es proveer estimaciones confiables y precisas cuando el modelo de regresión logística es ralo, definiendo métodos de estimación que seleccionen automáticamente variables.

Cabe mencionar que, aún en situaciones donde la dimensión de las variables explicativa es baja, es muy habitual que una proporción pequeña de los datos sean atípicos, es decir, que tienen un comportamiento distinto al de la mayoría de los datos. En el caso particular del modelo de regresión logística, los datos atípicos están asociados a datos mal clasificados que, si además corresponden a puntos de alta palanca, pueden tener una gran influencia. Cuando este es el caso, el EMV puede verse severamente afectado tanto en sesgo como en eficiencia. La Estadística Robusta se encarga de

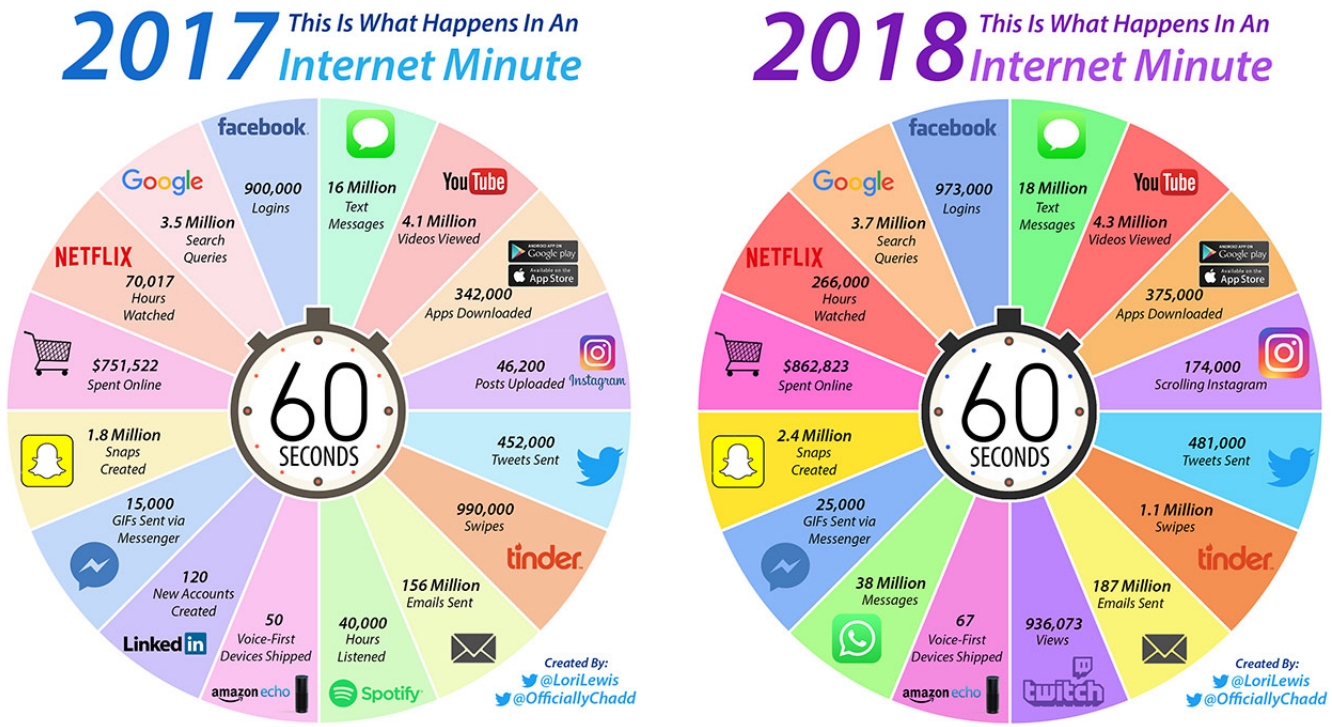


Figura 1.1: Comparación entre un minuto de Internet en 2017 y 2018. Fuente: `http://www.visualcapitalist.com/internet-minute-2018/`

este tipo de problemas y desarrolla estimadores que son menos sensibles cuando hay datos atípicos y casi tan eficientes como el EMV cuando no los hay, ver Maronna *et al.* (2019) para un análisis detallado de métodos robustos para distintos modelos.

Existen varias propuestas robustas para estimar el vector de coeficientes del modelo de regresión logística cuando la dimensión es baja y no se considera el problema de selección de variables. Pregibon (1982), Bianco y Yohai (1996), Cantoni y Ronchetti (2001), Croux y Haesbroeck (2003), Feng *et al.* (2014) y Basu *et al.* (2017) son algunos ejemplos de estas propuestas. Sin embargo, estos métodos sólamente son eficientes cuando el número de covariables es mucho menor a la cantidad de individuos de la muestra. Además, como el EMV, en el caso de modelos ralos, no permiten hacer una selección automática de las variables explicativas que realmente intervienen en el modelo. Por ello, en esta tesis nos focalizamos en la búsqueda de procedimientos de estimación bajo un modelo de regresión logística que seleccione variables y provea inferencias estables cuando existe en la muestra un pequeño porcentaje de observaciones mal clasificadas.

Como mencionamos antes, una estrategia muy popular al lidiar con situaciones de alta dimensión es sumar un término de penalización en el correspondiente problema de minimización. Hoerl y Kennard (1970) propusieron la penalización Ridge para resolver el problema de matrices de diseño mal condicionadas. Este término reduce el efecto del sobreajuste y previene problemas de multicolinearidad, mejorando así el comportamiento de los estimadores. Sin embargo, cuando se usa esta penalidad, el estimador resultante no selecciona variables. La penalización LASSO propuesta por Tibshirani (1996), simultáneamente selecciona variables y reduce el efecto del sobreajuste. Por otra parte, la penalidad Elastic Net considerada por Zou y Hastie (2005) combina estas dos estrategias y mejora a LASSO cuando hay alta correlación entre las covariables. Todas estas funciones de penalización

achican los coeficientes estimados hacia cero, generando así un sesgo no despreciable en la estimación. Algunas otras funciones de penalización, como la función SCAD, introducida por Fan y Li (2001) y la MCP propuesta por Zhang (2010), solucionan este problema.

Cabe observar que las penalizaciones Ridge, LASSO y Elastic Net son funciones convexas, lo cual permite encontrar el mínimo absoluto del problema de minimización correspondiente en el caso en que la función de pérdida también sea convexa. Lamentablemente, las penalizaciones SCAD y MCP no son funciones convexas, lo cual hace más difícil encontrar el mínimo absoluto del problema de minimización correspondiente. Sin embargo, en muchos casos, los mínimos locales de estos problemas también producen buenas estimaciones para los vectores de coeficientes de la regresión. Elsener y van de Geer (2018) hacen un profundo análisis de esta situación. En esta tesis, proponemos la llamada penalización **Signo** que, si bien no es convexa, reduce el sesgo introducido por LASSO, tiene una forma mucho más simple que SCAD y MCP y depende de un solo parámetro.

En el contexto de alta dimensión o de modelos ralos, algunos autores han considerado propuestas robustas para el modelo de regresión logística. Entre otros, podemos mencionar a Chi y Scott (2014), quienes consideran un estimador de mínimos cuadrados con una penalización Elastic Net y Kurnaz *et al.* (2018) quienes proponen un estimador basado en la suma podada de las desviaciones con una función de penalización Elastic Net. Vale la pena destacar que el estimador de mínimos cuadrados que utilizan Chi y Scott (2014) corresponde a un caso particular de los $M-$estimadores penalizados considerados en esta tesis. Por otro lado, los estimadores definidos en Kurnaz *et al.* (2018) no resultan Fisher–consistentes, ya que se considera simplemente la suma podada de los residuos *deviance*, pero no se agrega un término de corrección que asegure la consistencia de los estimadores al verdadero parámetro de regresión. Tibshirani y Manning (2013) introducen términos de corrimiento en la función de pérdida para prevenir el posible efecto de datos atípicos. Finalmente, Park y Konishi (2016) consideran un enfoque de desviaciones pesadas con penalización Elastic Net. Los pesos utilizados se basan en distancias de Mahalanobis computadas en un espacio de menor dimensión generado por las primeras componentes principales. Sin embargo, ninguno de estos autores estudia propiedades de consistencia, tasa de convergencia o de selección de variables de las propuestas dadas.

Las propiedades de los estimadores clásicos penalizados en el modelo lineal y lineal generalizado fueron ampliamente estudiadas. Una propiedad deseable de este tipo de estimadores es la llamada propiedad **oráculo**. Dado un modelo que depende de un parámetro $\boldsymbol{\beta}$, decimos que el estimador $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ tiene la propiedad oráculo si se cumplen las siguientes dos condiciones:

(a) Con probabilidad tendiendo a uno, $\widehat{\boldsymbol{\beta}}$ identifica todas las covariables no activas (es decir, todos los coeficientes que corresponden a covariables no activas son estimados como cero).

(b) La distribución asintótica del sub–vector que corresponde a las covariables activas es la misma que la que se obtendría si conociéramos de antemano cuál es el subconjunto de covariables activas y aplicáramos el método no penalizado sobre este modelo restringido.

Para el modelo de regresión lineal, Bühlmann y van de Geer (2011) dan un análisis completo sobre las propiedades de estos estimadores, no sólo para el caso de mínimos cuadrados sino que incluyen

además el caso de pérdidas convexas combinadas con una penalización LASSO o LASSO Adaptiva. Los resultados de consistencia y de selección de variables dependen de las llamadas hipótesis de compatibilidad e irrepresentabilidad. Una observación importante es que, para el modelo lineal, el estimador de mínimos cuadrados con penalización LASSO no tiene la propiedad oráculo (ver Zou, 2006). Para el modelo lineal generalizado, bajo condiciones análogas a las dadas en el modelo lineal, van de Geer y Müller (2012) prueban resultados de consistencia y selección de variables usando una función de pérdida convexa.

Para resolver los problemas de la penalización LASSO, Fan y Li (2001) definen la penalización SCAD y prueban resultados asintóticos para el caso en que la dimensión es fija. Entre otros resultados, muestran que, para modelos lineales generalizados, los estimadores de máxima verosimilitud con penalización SCAD tienen un máximo local con la propiedad oráculo. Posteriormente, Fan y Peng (2004) extendieron este resultado probando la propiedad oráculo cuando la dimensión p de las variables explicativas tiende a infinito junto con el tamaño de muestra n, bajo el régimen $p^5/n \to 0$. Más recientemente, Fan y Lv (2011) estudian el caso de dimensión no polinomial, es decir, cuando p puede tender a infinito exponencialmente en n. Más aún, dan condiciones bajo las cuales un máximo local es también global. Por otro lado, Huang y Xie (2007) obtienen resultados sobre las propiedades asintóticas de estimadores de mínimos cuadrados con penalizaciones cóncavas (como SCAD y MCP), para el modelo lineal. Para este mismo problema, Kim *et al.* (2008) también prueban que, con probabilidad tendiendo a uno, el estimador oráculo es en realidad una solución local del problema de minimización, mientras que Kim y Kwon (2012) dan condiciones bajo las cuales existe un único mínimo local para este problema. Algunos de estos resultados fueron extendidos por Xie y Huang (2009) a modelos parcialmente lineales.

Para el modelo de regresión lineal, Li *et al.* (2011) consideran $M-$estimadores con penalización cóncava. Cabe mencionar que los estimadores de Li *et al.* (2011) no son equivariantes y más aún, no permiten determinar qué puntos son atípicos de acuerdo al tamaño de sus residuos ya que no incorporan ningún estimador preliminar de escala. Por otra parte, estos estimadores utilizan una pérdida convexa que, como es conocido, no protege contra puntos de alta palanca. Para resolver estos problemas, Smucler (2016) y Smucler y Yohai (2017) adaptaron los $MM-$estimadores de regresión al caso ralo agregando una penalización Bridge o Bridge Adaptiva y estudiaron sus propiedades. Tanto estas últimas penalizaciones como las consideradas en Li *et al.* (2011) dan lugar a estimadores con la propiedad oráculo. Cabe mencionar, que las hipótesis utilizadas para probar resultados asintóticos cuando $p \to \infty$ en los trabajos antes mencionados sobre estimadores robustos, son diferentes a las hipótesis que consideramos en esta tesis. La principal diferencia radica en que en dichos trabajos se supone que las variables explicativas son fijas, mientras que nosotros las consideramos aleatorias. Por otra parte, al considerar un modelo de regresión logística, las observaciones son naturalmente heteroscedásticas, aunque no tenemos que dar un estimador preliminar de otro parámetro auxiliar como el de escala en el caso del modelo de regresión lineal.

Para modelos lineales generalizados, Avella–Medina (2016) introduce una familia de estimadores penalizados robustos que corresponde a una versión penalizada de los estimadores de Cantoni y Ronchetti (2001), es decir, estimadores que acotan la cuasi–verosimilitud. Como es bien sabido, este tipo de estimadores, no resultan adecuados bajo un modelo de regresión logística. Entre otros

resultados, Avella–Medina (2016) muestra que sus estimadores tienen la propiedad oráculo cuando la dimensión p es fija. Por otra parte, Loh (2017) y Loh y Wainwright (2017) presentan resultados asintóticos para $M-$ estimadores penalizados, en un marco muy general. Dicho marco admite que tanto la función de pérdida como la penalización sea no convexa, sin embargo en el caso de regresión lineal suponen que la escala es conocida y no permiten funciones de pérdida acotadas, de allí la relevancia de los resultados de Smucler (2016) y Smucler y Yohai (2017). A pesar de ello, es de destacar que el contexto general considerado en Loh (2017) y Loh y Wainwright (2017), que incluye por ejemplo el *Graphical LASSO*, permite obtener resultados de consistencia, selección consistente de variables y distribución asintótica bajo la condición llamada "Restricted Strong Convexity". En general, verificar esta hipótesis no es una tarea sencilla y por lo tanto, no asumiremos como válido este supuesto en esta tesis. Otro inconveniente de los resultados presentados en Loh (2017) y Loh y Wainwright (2017) es que los estimadores allí considerados se definen a partir de un problema de minimización restringido, considerando como espacio de parámetros una bola de ℓ_1. Esta restricción del espacio paramétrico facilita la obtención de resultados de consistencia.

Más recientemente, Elsener y van de Geer (2018) prueban las llamadas desigualdades oráculo "sharp" para $M-$ estimadores generales con pérdidas no convexas, utilizando una norma como función de penalización. La hipótesis principal de su trabajo es la condición que llaman "Two Point Margin", que puede ser vista como la versión poblacional de la condición "Restricted Strong Convexity". En particular, Elsener y van de Geer (2018) aplican sus resultados al caso de regresión logística utilizando el estimador de mínimos cuadrados, como en Chi y Scott (2014), pero con penalización LASSO en lugar de Elastic Net. Sin embargo, debido a la condición "Two Point Margin" las condiciones utilizadas resultan muy restrictivas, entre otras razones, porque requieren acotación de las variables explicativas respecto de la norma de Orlicz.

Los métodos presentados y estudiados en esta tesis son $M-$ estimadores de regresión penalizados para el modelo de regresión logística, es decir, son versiones penalizadas y pesadas de los estimadores propuestos por Bianco y Yohai (1996). Estos estimadores incluyen un término de penalización para producir estimaciones ralas del vector de coeficientes de la regresión. La función de pérdida que consideramos es una función acotada que controla los residuos deviance y por lo tanto, resulta no convexa. Esta familia de funciones de pérdida permite incluir muchos de los estimadores ya considerados en la literatura para el caso de modelos de regresión logística que asumen parámetros ralos, así como los estimadores propuestos por Chi y Scott (2014). Como en Croux y Haesbroeck (2003), permitimos además la inclusión de pesos para controlar el efecto de los puntos de alta palanca.

Esta tesis está organizada de la siguiente forma. El Capítulo 2 presenta algunos resultados preliminares que incluyen: la definición y propiedades asintóticas de los estimadores de máxima verosimilitud en el modelo de regresión logística, un estudio de su comportamiento cuando la relación entre el número de variables explicativas y la cantidad de observaciones es alta y otro sobre su sensibilidad a datos atípicos. Finalmente, concluimos dicho capítulo con una descripción de varias funciones de penalización consideradas en la literatura. El Capítulo 3 describe los estimadores robustos penalizados propuestos objeto de nuestro estudio e introduce una nueva penalidad en este contexto, la penalización Signo. En particular, como nuestro enfoque está basado en $M-$estimadores

penalizados, describimos un método robusto para elegir el parámetro de penalización. Muchos de nuestros resultados se basan en definiciones y resultados sobre procesos empíricos. Por esta razón y a los fines de completitud de esta tesis, en el Capítulo 4 presentamos un resumen de esas nociones, en su mayoría, extraídos de van der Vaart y Wellner (1996), van der Geer (2000) y Bühlmann y van de Geer (2011). Resultados de consistencia, tasas de convergencia, selección de variables y sobre la distribución asintótica de los estimadores propuestos para el caso en que la dimensión de la covariables es fija, pero el modelo es ralo, se detallan en el Capítulo 5; mientras que el estudio de dichas propiedades cuando el número de variables explicativas crece con el tamaño de muestra se obtiene en el Capítulo 6. En el Capítulo 7, se describe el algoritmo utilizado para calcular los estimadores propuestos y se presentan los resultados de un extenso estudio de simulación diseñado para evaluar el desempeño de los estimadores propuestos para distintas elecciones tanto de la función de pérdida como de la penalidad para conjuntos de muestras con datos atípicos y sin ellos. En particular, los $M-$estimadores pesados con penalizaciones acotadas muestran sus ventajas bajo los diferentes esquemas de contaminación considerados. Se presenta además en la Sección 7.3 el análisis de algunos conjuntos de datos. Finalmente, en el Capítulo 8 hacemos algunos comentarios finales y presentamos las conclusiones. Para facilitar la lectura, las demostraciones fueron relegadas a los apéndices de cada capítulo.

Capítulo 2

El modelo de regresión logística

2.1. El estimador de máxima verosimilitud

2.1.1. Un ejemplo simple y descripción del modelo

Consideremos el siguiente conjunto de datos disponible en `https://stats.idre.ucla.edu/stat/data/binary.csv` que corresponde a mediciones hechas sobre 400 aspirantes a ingresar a la universidad. Para cada aspirante, consideramos la variable respuesta `admit`, que vale 1 si fue admitido y 0 si no, y la variable explicativa `gpa` que es el promedio de dicho alumno en la secundaria. Nuestro interés radica en saber si la variable `gpa` ayuda a predecir la admisión de los alumnos. La Figura 2.1 muestra el boxplot de la variable `gpa` para cada unos de los dos grupos, es decir, para los alumnos admitidos y los no admitidos.

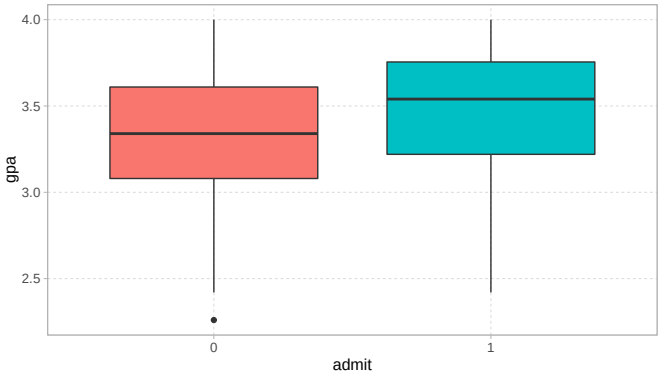


Figura 2.1: Boxplot de la variable `gpa` para los alumnos admitidos y no admitidos.

Los boxplots permiten observar que los alumnos admitidos tienen un `gpa` ligeramente mayor que el de los no admitidos. Nos interesa averiguar en qué medida la variable `gpa` explica la admisión de un alumno. Para ello, planteamos un modelo de regresión logística que supone que, a través de una función de vínculo, dicha dependecia es una función lineal de `gpa`, es decir, la probabilidad de que el aspirante i sea admitido dependerá únicamente de la cantidad $\gamma_0 + \beta_1\ \texttt{gpa}_i$, donde γ_0 y β_1 son coeficientes a estimar. Sin embargo, esta última cantidad varía en todo $\mathbb{R}$, mientras que las probabilidades varían en el intervalo $[0, 1]$. Para que dichos rangos sean compatibles, consideramos

la función **logística** $F : \mathbb{R} \to (0, 1)$, definida como

$$F(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}} \, . \tag{2.1}$$

Finalmente, el modelo de regresión logística asume que $\mathbb{P}(\texttt{admit}_i = 1) = F(\gamma_0 + \beta_1 \, \texttt{gpa}_i)$, para todo $i = 1, \ldots, 400$. Observemos que si β_1 es positivo, entonces, según este modelo, un alumno con mayor `gpa` tendrá una mayor probabilidad de ingresar a la universidad.

A diferencia de este ejemplo, en el que existe sólo una variable explicativa, en esta tesis consideramos conjuntos de datos con varias covariables, es decir, suponemos tener una muestra de observaciones i.i.d. $(Y_i, \mathbf{X}_i^{\mathrm{T}})$, $1 \leq i \leq n$ tales que $\mathbf{X}_i \in \mathbb{R}^p$ y $Y_i \in \{0, 1\}$. Diremos que $(Y_i, \mathbf{X}_i^{\mathrm{T}})$ satisfacen el modelo de regresión logística si la distribución condicional de $Y_i | \mathbf{X}_i \sim Bi(1, \pi_{0,i})$, donde

$$\pi_{0,i} = \mathbb{P}(Y_i = 1 | \mathbf{X}_i) = F(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0) = \frac{\exp\left(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right)}{1 + \exp\left(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0\right)} \, , \tag{2.2}$$

siendo $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ el vector de coeficientes de la regresión a estimar. Cabe observar que este modelo permite una ordenada al origen, como en el ejemplo anterior, tomando $\mathbf{X} = \begin{pmatrix} 1 \\ \mathbf{X}^\star \end{pmatrix}$ donde $\mathbf{X}^\star$ corresponden a las variables explicativas medidas (en el caso del ejemplo de admisión $\mathbf{X}^\star =$`gpa`).

2.1.2. Estimación de los coeficientes por máxima verosimilitud

Dada una muestra de observaciones i.i.d. $(Y_i, \mathbf{X}_i^{\mathrm{T}})$, $1 \leq i \leq n$, que satisface el modelo (2.2), el estimador clásico de los coeficientes $\boldsymbol{\beta}_0$ es el de máxima verosimilitud, $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$, que se define como

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MV}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmax}} \prod_{i=1}^{n} \left[F(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})\right]^{Y_i} \left[1 - F(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})\right]^{1 - Y_i} \, .$$

Cálculos sencillos muestran que una forma equivalente de definir este estimador es a través del problema de minimización siguiente

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MV}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} d(Y_i, \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}), \tag{2.3}$$

donde

$$d(y, t) = -\log(F(t)) y - \log(1 - F(t))(1 - y) \tag{2.4}$$

es llamada **devianza** o *deviance*, por su nombre en inglés. Si $p > n$, puede probarse que el mínimo del problema (2.3) se alcanza en el infinito, por lo que en este caso este estimador no estaría bien definido. Por esta razón, supondremos que $n \geq p$ siempre que nos refiramos a este estimador.

A diferencia de lo que ocurre con el estimador de mínimos cuadrados en el modelo lineal, los estimadores de máxima verosimilitud en regresión logística no tienen una forma analítica explícita, sino que se obtienen a partir de métodos numéricos de optimización. Cabe señalar que la función a minimizar en (2.3) es suave y convexa, lo cual facilita hallar numéricamente su mínimo global.

En el ejemplo de los aspirantes a la universidad, podemos estimar por máxima verosimilitud los coeficientes β_0 y β_1 usando la función `glm` en el lenguaje `R`, obteniendo $\widehat{\beta}_0 = -4.3576$ y $\widehat{\beta}_1 = 1.0511$. Observemos que los resultados obtenidos confirman nuestra sospecha inicial: según este modelo, tener un mayor `gpa` implica tener una mayor probabilidad de ser admitido en la universidad.

2.1.3. Propiedades asintóticas del estimador de máxima verosimilitud

En esta sección, recordamos algunos resultados de consistencia y distribución asintótica que conciernen al estimador de máxima verosimilitud definido en la Sección 2.1.2.

De ahora en más, dado un vector $\mathbf{b} = (b_1, \ldots, b_p)^{\mathrm{T}} \in \mathbb{R}^p$ indicaremos por $\|\mathbf{b}\|_\infty = \max_{1\leq j\leq p} |b_j|$, $\|\mathbf{b}\|_2^2 = \sum_{j=1}^p b_j^2$ y $\|\mathbf{b}\|_1 = \sum_{j=1}^p |b_j|$. Más generalmente, $\|\mathbf{b}\|_q^q = \sum_{j=1}^p |b_j|^q$ para todo $q > 0$.

Los resultados que mencionaremos, junto con sus demostraciones, pueden verse en Gourieroux y Monfort (1981). En dicho trabajo, a diferencia de esta tesis, se asume que las covariables son fijas, por este motivo las denotamos en minúscula, es decir, consideramos una muestra de observaciones i.i.d., $\left(Y_{n,i}, \mathbf{x}_{n,i}^{\mathrm{T}}\right)$, $1 \leq i \leq n$, con $n > p$ donde $\mathbf{x}_{n,i} \in \mathbb{R}^p$ son covariables fijas e $Y_{n,i} \sim Bi(1, F(\mathbf{x}_{n,i}^{\mathrm{T}}\, \boldsymbol{\beta}_0))$ para algún vector de coeficientes $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, donde eventualmente la primer coordenada de $\boldsymbol{\beta}_0$ puede corresponder a la ordenada al origen en cuyo caso, la primer componente de $\mathbf{x}_{n,i}$ es 1. Definamos

$$\boldsymbol{\Upsilon}_n = \sum_{i=1}^n F(\mathbf{x}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}_0)\,(1 - F(\mathbf{x}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}_0))\,\mathbf{x}_{n,i}\mathbf{x}_{n,i}^{\mathrm{T}}\,.$$

Supondremos que la matriz $\boldsymbol{\Upsilon}_n$ es inversible para todo n (esto no podría ocurrir si el tamaño de muestra n fuera menor a la cantidad de covariables p) y consideramos las siguientes hipótesis:

A1 Existe una constante M_0 tal que $\|\mathbf{x}_{n,i}\|_\infty < M_0$ para todo $i = 1, \ldots, n$ y para todo $n \in \mathbb{N}$.

A2 Sean $\iota_{1,n}$ y $\iota_{p,n}$ respectivamente el mínimo y máximo autovalor de la matriz $\boldsymbol{\Upsilon}_n$. Existe una constante M_1 tal que $\iota_{p,n}/\iota_{1,n} < M_1$.

El siguiente teorema establece que el estimador de máxima verosimilitud $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$ de $\boldsymbol{\beta}_0$ es fuertemente consistente.

Teorema 2.1. *Supongamos que se cumplen las condiciones **A1** y **A2**, entonces, $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}} \xrightarrow{c.s.} \boldsymbol{\beta}_0$ si y solo si $\lim_{n\to\infty} \iota_{1,n} = +\infty$.*

Finalmente, el Teorema 2.2 asegura que estos estimadores son asintóticamente normales, lo que permite obtener estadísticos para testear hipótesis sobre los parámetros y construir intervalos de confianza para cada uno de los coeficientes.

Teorema 2.2. *Supongamos que se cumplen las condiciones **A1** y **A2** y que $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}} \xrightarrow{p} \boldsymbol{\beta}_0$. Entonces, $\boldsymbol{\Upsilon}_n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{MV}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_p)$.*

Veremos más adelante que el Teorema 2.2 es compatible con los resultados de distribución asintótica de los estimadores presentados en esta tesis.

2.2. Algunos problemas del estimador de máxima verosimilitud

El objetivo de esta sección es ilustrar algunos problemas que puede tener el estimador de máxima verosimilitud ya sea cuando la dimensión de las covariables es alta en relación al tamaño muestral o cuando existen datos mal clasificados en la muestra. Para cada uno de estos casos, realizamos un pequeño estudio de simulación que ejemplifica el efecto correspondiente.

2.2.1. El problema de sobreajuste

Un primer problema del estimador de máxima verosimilitud surge cuando la cantidad p/n, que suponemos menor o igual a 1, es relativamente grande.

Para ello, consideremos una muestra de observaciones $\mathcal{M} = \{(Y_i, \mathbf{X}_i), i = 1, \ldots, n\}$ del modelo de regresión logística (2.2) sin ordenada al origen. Sea $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ el estimador de máxima verosimilitud de $\boldsymbol{\beta}_0$ basado en $\mathcal{M}$. La muestra $\mathcal{M}$ es usualmente llamada *muestra de entrenamiento*. Se define la **proporción de clasificaciones correctas dentro de la muestra** como $\#\{i : \widehat{Y}_i = Y_i\}/n$, donde $\widehat{Y}_i = 1$ si $F(\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\text{MV}}) > 1/2$ y $\widehat{Y}_i = 0$ en caso contrario. Vale la pena observar que la condición $F(\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\text{MV}}) > 1/2$ es equivalente a $\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\text{MV}} > 0$.

Supongamos ahora que, una vez obtenido el estimador $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ con la muestra $\mathcal{M}$, queremos ahora obtener predicciones para una nueva muestra $\mathcal{T} = \{(Y_{i,\mathcal{T}}, \mathbf{X}_{i,\mathcal{T}}), i = 1, \ldots, \widetilde{n}\}$ independiente de $\mathcal{M}$ y tal que $(Y_{i,\mathcal{T}}, \mathbf{X}_{i,\mathcal{T}}) \sim (Y_i, \mathbf{X}_i)$. La muestra $\mathcal{T}$ se denomina *muestra de testeo*. Definimos la **proporción de clasificaciones correctas fuera de la muestra** como $\#\{i : \widehat{Y}_{i,\mathcal{T}} = Y_{i,\mathcal{T}}\}/\widetilde{n}$, donde $\widehat{Y}_{i,\mathcal{T}} = 1$ si $F(\mathbf{X}_{i,\mathcal{T}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\text{MV}}) > 1/2$ y $\widehat{Y}_{i,\mathcal{T}} = 0$ en caso contrario.

Consideremos ahora el siguiente estudio de simulación donde elegimos como tamaños para las muestras de entrenamiento y de testeo $n = \widetilde{n} = 150$ y la dimensión de las covariables p en el conjunto $\{5, 10, 15, \ldots, 115, 120\}$. El vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ elegido es tal que $(\boldsymbol{\beta}_0)_j = \mathbb{I}_{j \leq 5}$, es decir, sus primeras 5 coordenadas son 1 y el resto son 0. Las covariables $\mathbf{X}_i$ se generan a partir de una distribución $N(\mathbf{0}, \mathbf{I}_p)$ y las variables Y_i se obtienen de forma independiente con distribución condicional $Y_i|\mathbf{X}_i \sim Bi(1, F(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0))$. Para cada uno de los valores de p, realizamos 400 replicaciones en donde obtenemos el estimador $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ y calculamos la proporción de clasificaciones correctas dentro y fuera de la muestra. Finalmente, promediamos los 400 resultados, obteniendo el gráfico que se muestra en la Figura 2.2. Los círculos rojos indican la proporción de clasificaciones correctas dentro de la muestra, mientras que los triangulos celestes la proporción fuera de la muestra.

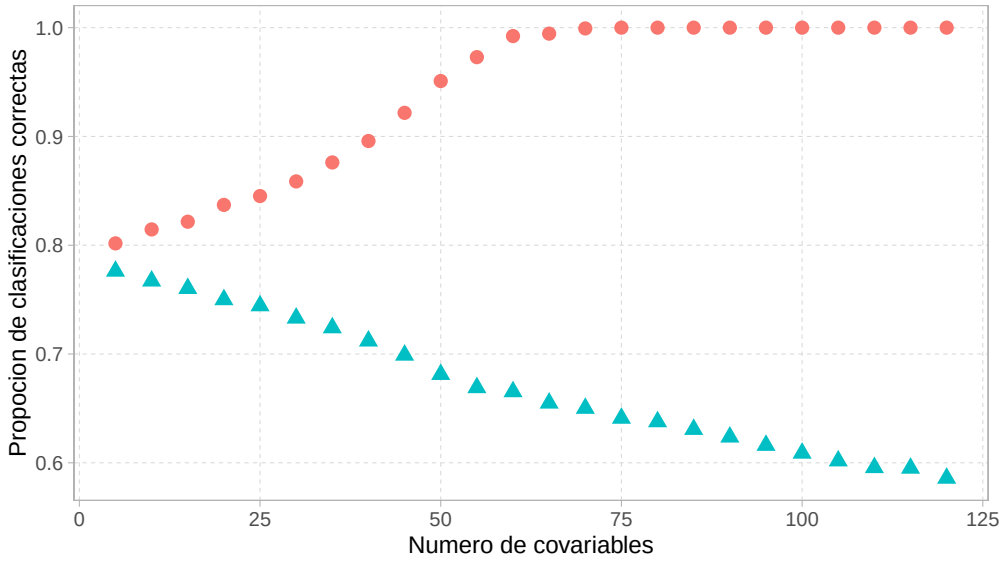


Figura 2.2: Proporción de clasificaciones correctas dentro y fuera de la muestra con $n = 150$ y $(\boldsymbol{\beta}_0)_j = \mathbb{I}_{j \leq 5}$. Cada uno de los puntos corresponde al promedio sobre 400 replicaciones. Los círculos rojos y los triangulos celestes indican la proporción de clasificaciones correctas dentro y fuera de la muestra, respectivamente.

La Figura 2.2 permite ver que a medida que aumenta la cantidad de covariables p, la proporción de clasificaciones correctas dentro de la muestra aumenta, mientas que fuera de la muestra disminuye. Este efecto se llama en la literatura *efecto de sobreajuste* y puede describirse como el efecto que se produce cuando el resultado de un análisis corresponde en forma demasiado exacta a un conjunto particular de observaciones y, por lo tanto, falla en ajustar o predecir futuras observaciones. Esto suele suceder cuando las mismas observaciones son utilizadas para estimar y ajustar un modelo y es mayor a medida que aumenta el número de covariables. En el caso que consideramos, cuando p/n es cercano a 1, el estimador $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ logra obtener excelentes resultados sobre la muestra de entrenamiento, pero falla al intentar predecir la muestra de testeo generada bajo el mismo modelo. En base a este comportamiento, podemos consluir que este estimador no es recomendable cuando la relación p/n es grande.

2.2.2. El efecto de datos atípicos

Un segundo problema del estimador $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ es su sensibilidad cuando la muestra de entrenamiento contiene datos atípicos, es decir, observaciones que provienen de una distribución distinta a la que corresponde a la mayoría de los datos. En este caso, el estimador de máxima verosimilitud produce estimaciones poco fiables.

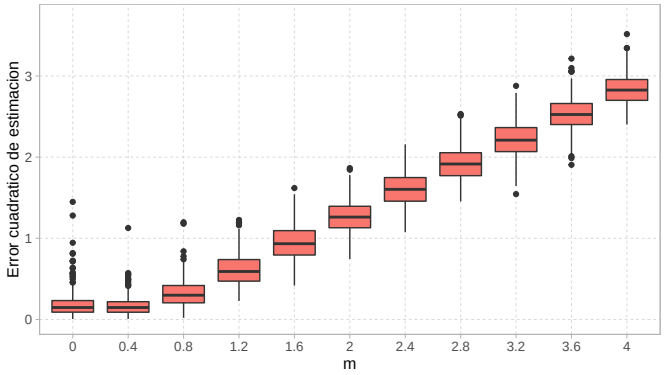
Para ejemplificar este problema, realizamos 400 replicaciones donde generamos muestras $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, del modelo de regresión logística (2.2) sin intercept, con $n = 300$, $p = 5$ y $\boldsymbol{\beta}_0 = \mathbf{1}_5^{\text{T}}$, donde $\mathbf{1}_p$ es el vector en $\mathbb{R}^p$ con todas sus coordenadas iguales a 1. Las covariables $\mathbf{X}_i$ y las variables respuesta Y_i se generan del mismo modo que en la Sección 2.2.1. Para introducir datos atípicos, consideramos un porcentaje $\varepsilon \in [0, 1]$ de datos mal clasificados. Más precisamente, dado $m \in \mathbb{R}_{\geq 0}$, se agregan a la muestra original $n_0 = [n\varepsilon]$ datos atípicos $(\widetilde{Y}_j, \widetilde{\mathbf{X}}_j)$, $1 \leq j \leq n_0$, donde $\widetilde{\mathbf{X}}_j = m\boldsymbol{\beta}_0/\sqrt{5} + \boldsymbol{\gamma}_j$, $\widetilde{Y}_j = 0$ y $\boldsymbol{\gamma}_j$ es una pequeña perturbación aleatoria de modo que los datos atípicos generados no sean todos idénticos. Cuanto mayor sea el valor de m, mayor influencia tendrán los datos atípicos en la estimación de $\boldsymbol{\beta}_0$ ya que corresponderán a datos mal clasificados de alta palanca.

Para cada muestra obtenida, según variamos m y ε, calculamos el estimador de máxima verosimilitud $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ y para resumir su comportamiento evaluamos el error cuadrático de estimación

$$\|\widehat{\boldsymbol{\beta}}_{\text{MV}} - \boldsymbol{\beta}_0\|_2^2 = (\widehat{\boldsymbol{\beta}}_{\text{MV}} - \boldsymbol{\beta}_0)^{\text{T}}(\widehat{\boldsymbol{\beta}}_{\text{MV}} - \boldsymbol{\beta}_0)\,.$$

La Figura 2.3 presenta los boxplots de los 400 errores cuadráticos obtenidos bajo los distintos escenarios de contaminación. El caso $\varepsilon = 0$ corresponde a los datos originales. En la Figura 2.3(a) se muestran los resultados cuando $\varepsilon = 0.05$ y varía m, mientras que en (b) varía el porcentaje de contaminación para la elección de $m = 2$. Como se observa en el gráfico, el estimador $\widehat{\boldsymbol{\beta}}_{\text{MV}}$ se ve fuertemente afectado tanto cuando aumenta el desplazamiento de la covariable como cuando se agrega un mayor porcentaje de datos atípicos. Por este motivo, en el Capítulo 3 describiremos estimadores resistentes, en particular, ante este tipo de datos atípicos.

(a) $\varepsilon = 0.05$



(b) $m = 2$

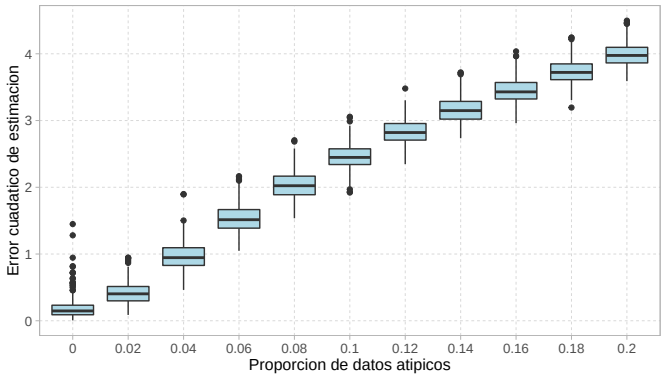


Figura 2.3: Boxplots de los errores cuadráticos de estimación cuando $n = 300$, $p = 5$, $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1)^{\mathrm{T}}$ y se agrega un $100\varepsilon\,\%$ de datos atípicos generados con desplazamiento m. En (a) se fijó $\varepsilon = 0.05$ y se muestran los resultados en función de m, mientras que en (b) se fijó $m = 2$ y se muestran los resultados en función de ε.

2.3. Funciones de penalización

En la bibliografía, la manera habitual de tratar modelos ralos es introduciendo un término de penalización que, por lo general, no incluye la ordenada al origen. Por esta razón y para simplificar la presentación, de ahora en más, consideraremos un modelo de regresión logística sin ordenada al origen.

Cuando el modelo es ralo, para reducir el efecto del sobreajuste del estimador $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$ se consideran estimadores de la forma

$$\widehat{\boldsymbol{\beta}}^{\mathrm{I}}_{\mathrm{MV}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} d(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) + I_{\boldsymbol{\lambda}}(\boldsymbol{\beta}), \tag{2.5}$$

donde $I_{\boldsymbol{\lambda}}$ es una función no negativa que depende de un vector de parámetros de ajuste $\boldsymbol{\lambda}$.

En muchas situaciones, cuando la dimensión p es mediana o grande, los modelos que se plantean son ralos. Por esta razón, es de interés realizar selección de variables, es decir, seleccionar el subconjunto de covariables asociadas a valores no nulos del coeficiente de regresión ya que este subconjunto es suficiente para predecir la respuesta en base a una nueva observación de las variables explicativas, lo cual facilita la interpretabilidad del modelo. Como todas las coordenadas del vector $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$ son distintas de cero con probabilidad uno, este estimador no arroja modelos intepretables. Tal como describiremos a continuación, ciertas funciones de penalización cumplen que su correspondiente

estimador tiene muchas de sus coordenadas iguales a cero con alta probabilidad, es decir, estiman los coeficientes y simultáneamente, realizan selección de variables.

En esta sección, explicamos brevemente las características de algunas funciones de penalización introducidas en la literatura.

2.3.1. Penalizaciones Ridge, Bridge, LASSO y Elastic Net

Para reducir el problema de colinealidad de las variables explicativas, Hoerl y Kennard (1970) introdujeron la penalización Ridge definida como $I_\lambda(\boldsymbol{\beta}) = (\lambda/2)\,\|\boldsymbol{\beta}\|_2^2$ para $\lambda \geq 0$. Esta penalización, además de dar buenos resultados cuando hay un alto grado de colinealidad, reduce el efecto del sobreajuste y mejora el condicionamiento del problema de minimización (2.3).

Sin embargo, el estimador resultante al usar la penalización Ridge no selecciona variables. Por el contrario, la penalización LASSO, introducida en Tibshirani (1996), sí selecciona variables con alta probabilidad. Esta penalidad se define como

$$I_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 \,. \tag{2.6}$$

Esta diferencia sustancial entre Ridge y LASSO se debe esencialmente a que las bolas de norma 1 y de norma 2 poseen una distinta estructura geométrica (ver Bühlmann y van de Geer, 2011).

Por último, Hastie y Zou (2005) proponen tomar una combinación convexa de estas dos penalizaciones, es decir, consideran la penalización $I_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \lambda\{\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)/2\|\boldsymbol{\beta}\|_2^2\}$ donde $\boldsymbol{\lambda} = (\lambda, \alpha) \in \mathbb{R}_{\geq 0} \times [0,1]$. Esta penalización, llamada Elastic Net, permite realizar selección de variables si $\alpha > 0$ y arroja mejores resultados que la penalización LASSO cuando hay un alto grado de colinealidad entre las variables. Otras penalizaciones utilizadas en el ámbito de los modelos de regresión lineal son las penalizaciones Bridge introducidas en Frank y Friedman (1993) que corresponden a tomar $I_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_q^q$ y aseguran, en el caso del modelo lineal, modelos ralos si $q < 1$.

A medida que el valor de λ aumenta, todas estas penalizaciones achican los coeficientes estimados reduciendo así la variabilidad del estimador resultante. Sin embargo, este achicamiento también introduce un sesgo no despreciable en la estimación. En el Capítulo 3 proponemos una forma sencilla de modificar la penalización LASSO que logra reducir el sesgo introducido en la estimación.

Es importante observar que todas estas funciones de penalidad son convexas, lo cual permite encontrar el mínmo global de (2.5) numéricamente.

2.3.2. Penalizaciones SCAD y MCP

Fan y Li (2001) estudiaron una clase de métodos de penalización que incluyen el LASSO y conjeturaron que los estimadores obtenidos con LASSO no tienen la propiedad oráculo. Por esa razón, introdujeron la penalización SCAD (Smoothly Clipped Absolute Deviation) que, a diferencia de las funciones ya descriptas en esta sección, es cóncava. Esta función de penalización da lugar a estimadores que poseen propiedades asintóticas mejores que las obtenidas usando penalizaciones del tipo Elastic Net.

Para explicar la idea principal de la penalización SCAD, consideremos el modelo lineal con matriz de diseño ortogonal, es decir, supongamos que tenemos observaciones $(Y_i, \mathbf{X}_i)$, $1 \leq i \leq n$, tales que

$$Y_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i, \qquad i = 1, \ldots, n$$

donde $\|\mathbf{X}_i\|_2 = 1$ y $\mathbf{X}_i^{\mathrm{T}} \mathbf{X}_j = 0$, si $i \neq j$, $\mathbb{E}(\varepsilon_i) = 0$ y $\mathrm{VAR}(\varepsilon_i) = \sigma^2$. Sea $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ el estimador de mínimos cuadrados ordinario y definamos también el estimador penalizado

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})^2 + I_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \,. \tag{2.7}$$

Sea S_λ el operador "soft–thresholding", definido como $S_\lambda(t) = \operatorname{sign}(t)(|t| - \lambda)_+$ con $(x)_+ = \max(x, 0)$. Más aún, sea $H_\lambda(t) = t\, \mathbb{I}_{|t| \geq \lambda}$ el operador "hard–thresholding". Tanto S_λ como H_λ se conocen como reglas de truncado.

No es difícil mostrar que, cuando $I_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$ es la penalidad LASSO, $\widehat{\beta}_j = S_\lambda(\widehat{\beta}_{\mathrm{LS},j})$, mientras que si $I_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_0 = \lambda \, \#\{j : \beta_j \neq 0, 1 \leq j \leq p\}$, se tiene que $\widehat{\beta}_j = H_\lambda(\widehat{\beta}_{\mathrm{LS},j})$.

Tal como se observa en la Figura 2.4, S_λ tiene un corrimiento respecto de la función identidad (incluso para valores grandes de $|t|$), lo cual puede introducir un sesgo importante en la estimación de $\boldsymbol{\beta}_0$. Por otra parte, H_λ es idéntica a la función identidad cuando $|t|$ es grande pero es una función discontinua, lo que produce problemas de inestabilidad numérica en el problema (2.7). Como se muestra en la Figura 2.4, la regla de truncado SCAD es un compromiso entre S_λ y H_λ: es una función continua y es igual a la identidad cuando $|t|$ es grande. La función de penalización definida como

$$I_\lambda^{\mathrm{SCAD}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \lambda |\beta_j| \, \mathbb{I}_{|\beta_j| \leq \lambda} + \sum_{j=1}^{p} \frac{a\lambda|\beta_j| - 0.5(\beta_j^2 + \lambda^2)}{a - 1} \, \mathbb{I}_{\lambda < |\beta_j| \leq a\lambda} + \sum_{j=1}^{p} \frac{\lambda^2 (a^2 - 1)}{2(a - 1)} \, \mathbb{I}_{|\beta_j| > a\lambda} \tag{2.8}$$

para algún $a > 2$ tiene a SCAD como regla de truncado. Haciendo abuso de nomenclatura, llamamos también SCAD a la función de penalización dada en (2.8).

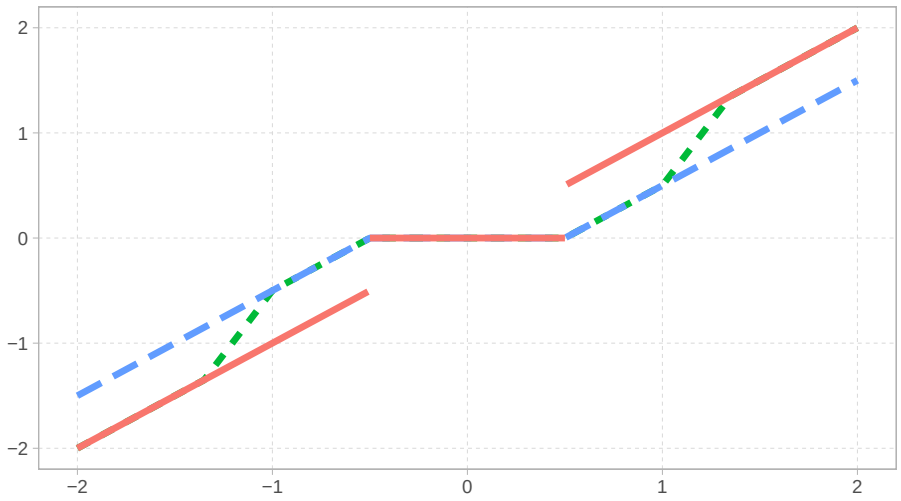


Figura 2.4: Gráfico de tres reglas de truncado distintas con $\lambda = 0.5$. La línea solida roja corresponde a $H_\lambda(t) = t\, \mathbb{I}_{|t| \geq \lambda}$, la línea a trazos largos celeste a $S_\lambda(t) = \operatorname{sign}(t)(|t| - \lambda)_+$, mientras que la línea verde a trazos cortos a la regla de truncado SCAD.

Otra penalidad muy utilizada es la penalización MCP definida por Zhang (2010) y que se basa en la siguiente idea. Definamos la máxima concavidad de una función τ como

$$\kappa(\tau) = \sup_{0<t_1<t_2} \frac{\tau'(t_1) - \tau'(t_2)}{t_2 - t_1}.$$

Dado $\alpha > 1$ y $\lambda > 0$, Zhang (2010) considera el siguiente problema de minimización:

$$\begin{cases} \text{minimizar} & \kappa(\tau) \\ \text{sujeto a} & \tau'(t) = 0 \quad \forall t \geq \alpha\lambda \quad \text{y} \quad \lim_{t\to 0^+} \tau'(0) = \lambda\,, \end{cases}$$

y prueba que la solución a este problema se alcanza cuando $\tau = J^{\text{MCP}}_{\lambda,\,a}$ donde

$$J^{\text{MCP}}_{\lambda,a}(t) = \left(\lambda|t| - \frac{t^2}{2\,a}\right) \mathbb{I}\{|t| \leq\, a\lambda\} + \frac{1}{2}\,a\lambda^2\,\mathbb{I}\{|t| >\, a\lambda\}\,. \tag{2.9}$$

Finalmente, define la penalización MCP como la función $I^{\text{MCP}}_{\lambda}(\boldsymbol{\beta}) = \sum_{j=1}^{p} J^{\text{MCP}}_{\lambda,\,a}(\beta_j)$.

Fan and Li (2001) sugieren que una buena función de penalización debería dar lugar a un estimador con las siguientes tres propiedades:

1. **Ausencia de sesgo:** es decir, los estimadores de los coeficientes de $\boldsymbol{\beta}_0$ que son grandes en módulo tienen poco sesgo.

2. **Estimación rala:** o sea, automáticamente estima como cero a los coeficientes pequeños o nulos para reducir la complejidad del modelo y mejorar su interpretabilidad.

3. **Estabilidad en las predicciones:** es decir, el estimador es estable con respecto a las componentes de $\mathbf{z} = \sum_{i=1}^{n} \mathbf{X}_i\, Y_i$ para obtener predicciones confiables.

Vale la pena observar que los estimadores que se obtienen a partir de las penalizaciones SCAD y MCP cumplen estas tres propiedades. Sin embargo, como estas penalizaciones son cóncavas, el correspondiente problema de minimización (2.5) puede no ser convexo y los algoritmos numéricos para obtener los estimadores son más complicados. Además estas dos penalidades, involucran la elección de dos parámetros de ajuste (a, λ), mientras que las penalizaciones LASSO o Ridge solamente involucran uno.

Capítulo 3

$M-$estimadores pesados penalizados

En este capítulo, introduciremos los $M-$estimadores penalizados para regresión logística que constituyen la propuesta de esta tesis y cuyo comportamiento asintótico estudiaremos en los Capítulos 5 y 6. De ahora en más, asumiremos que tenemos una muestra aleatoria $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$, $Y_i \in \{0, 1\}$, $\mathbf{X}_i \in \mathbb{R}^p$ del modelo de regresión logística (2.2). Antes de definir nuestros estimadores, recordaremos algunos estimadores robustos propuestos previamente para dicho modelo.

3.1. Estimadores robustos de Bianco y Yohai

Consideremos el estimador de máxima verosimilitud definido en (2.3). Tal como mostramos en la Sección 2.2.2, este estimador es muy sensible a la presencia de datos atípicos. A continuación, describiremos una familia de $M-$estimadores resistentes ante observaciones atípicas como las consideradas en la Sección 2.2.2.

Una primera propuesta para estimar en forma robusta el parámetro $\boldsymbol{\beta}_0$ fue dada por Pregibon (1982), quien consideró los estimadores definidos como

$$\widehat{\boldsymbol{\beta}}_{\text{PREG}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(d(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))\,,$$

donde $\rho : \mathbb{R} \to \mathbb{R}$ es una función monótona creciente pero que crece más lento que la función identidad, para acotar el efecto de residuos *deviance* $d(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$ grandes, donde $d(y, t)$ está definida en (2.4). Sin embargo, este estimador es asintóticamente sesgado.

Para resolver este problema, Bianco y Yohai (1996) agregan un término de corrección a la función objetivo considerada por Pregibon (1982). Más precisamente, sea $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ una función acotada, diferenciable y no decreciente con derivada $\psi = \rho'$. Los $M-$estimadores dados en Bianco y Yohai (1996) están definidos como

$$\widehat{\boldsymbol{\beta}}_{\text{M}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(d(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})) + G(F(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})) + G(1 - F(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))\,, \tag{3.1}$$

donde $G(t) = \int_0^t \psi(-\log u)\, du$ es el término de correción necesario para garantizar la Fisher–consistencia del estimador.

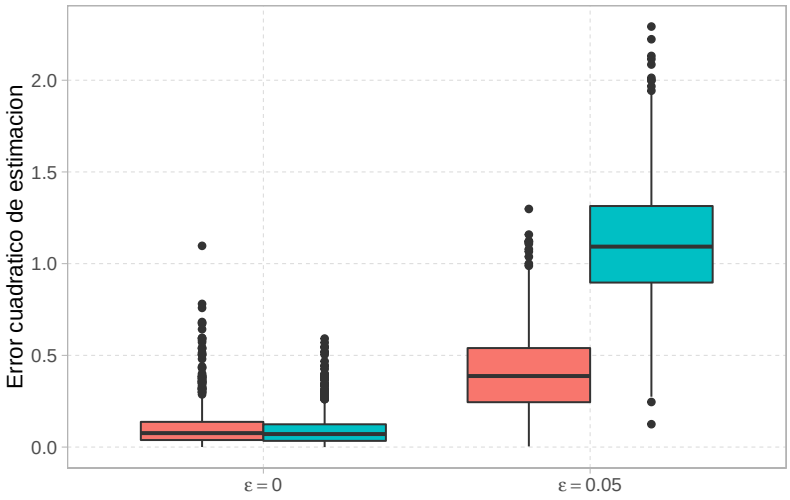


Figura 3.1: Boxplots de los errores cuadráticos de estimación cuando $n = 300$ y $\boldsymbol{\beta}_0 = (1, -0.5, 1.5)$. Las covariables para los datos no atípicos tienen distribución $N_3(\mathbf{0}_3, \mathbf{I}_3)$ y los datos atípicos fueron generados como en (3.2). Los boxplots en celeste y rosa corresponden al estimador de máxima verosimilitud y al $M-$estimador, respectivamente.

Para analizar el comportamiento del $M-$estimador $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$ ante observaciones mal clasificadas, consideramos el modelo de regresión logística (2.2) sin ordenada al origen y realizamos 1000 replicaciones generando muestras $(Y_i, \mathbf{X}_i)$, $1 \leq i \leq n$, con $n = 300$, $p = 3$ y $\boldsymbol{\beta}_0 = (1, -0.5, 1.5)$. Para las muestras sin contaminar, la distribución de las variables explicativas es $N_3(\mathbf{0}_3, \mathbf{I}_3)$. Para obtener muestras contaminadas, a cada una de esas muestras se le agregó una proporción ε de datos atípicos, donde cada dato atípico $(\widetilde{Y}, \widetilde{\mathbf{X}})$ fue elegido del siguiente modo: primero $\widetilde{\mathbf{X}}$ fue generado con distribución $N_3(\mathbf{0}_3, 10\,\mathbf{I}_3)$ y luego se tomó $\widetilde{Y}$ de modo a obtener datos mal clasificados, es decir,

$$\widetilde{Y} = \begin{cases} 1 & \text{si } \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}_0 < 0 \\ 0 & \text{si } \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}_0 \geq 0\,. \end{cases} \tag{3.2}$$

En la Figura 3.1, se muestran los boxplots del error cuadrático de estimación correspondientes al estimador de máxima verosimilitud (en celeste) y al $M-$estimador $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$ (en rosa) tomando como función ρ la función dada en (3.5). Puede observarse que el $M-$estimador da mejores resultados que el de máxima verosimilitud al agregar datos atípicos ya que los errores cuadráticos son menores. Por otro lado, cuando la muestra no contiene este tipo de datos, ambos métodos arrojan resultados similares.

3.1.1. Sobre la elección de la función ρ

Esta familia de estimadores tiene como casos particulares algunas funciones de pérdida ya consideradas en la literatura. Un primer ejemplo trivial se obtiene tomando $\rho(t) = t$, lo cual da lugar al estimador $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$. Por otra parte, en Bianco (1990) se prueba que el estimador de mínimos cuadrados $\widehat{\boldsymbol{\beta}}_{\mathrm{MC}}$ también pertenece a la familia de $M-$estimadores dada en (3.1) para una selección adecuada de la pérdida ρ. Recordemos que el estimador de mínimos cuadrados se define como

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MC}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - F(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}))^2 \tag{3.3}$$

y corresponde a la elección $\rho(t) = 1 - \exp(-t)$ en (3.1). Cabe mencionar que Chi y Schott (2014) agregan una penalización Elastic Net en el problema de minimización considerado en (3.3).

Otro ejemplo de estimadores incluidos entre los definidos por (3.1) corresponde a los estimadores de mínima divergencia introducidos en Basu *et al.* (2017). Estos autores observan que el estimador de máxima verosimilitud minimiza la medida de divergencia de Kullback–Leibler. Basándose en la llamada "power divergence" introducida en Basu *et al.* (1998), proponen estimar $\boldsymbol{\beta}_0$ mediante el estimador $\widehat{\boldsymbol{\beta}}_{\text{DIV}}$ definido como

$$\widehat{\boldsymbol{\beta}}_{\text{DIV}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \Delta_c(\widehat{\boldsymbol{\pi}}, \boldsymbol{\pi}(\boldsymbol{\beta}))\,, \tag{3.4}$$

donde $\widehat{\boldsymbol{\pi}} = (y_1, \ldots, y_n, 1-y_1, \ldots, 1-y_n)^{\mathrm{T}}/n$,

$$\begin{aligned}
\boldsymbol{\pi}(\boldsymbol{\beta}) &= \frac{1}{n}\left(F(\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}), \ldots, F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}), 1-F(\mathbf{x}_1^{\mathrm{T}}\boldsymbol{\beta}), \ldots, 1-F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})\right)^{\mathrm{T}}\,, \\
\Delta_c(\widehat{\boldsymbol{\pi}}, \boldsymbol{\pi}(\boldsymbol{\beta})) &= \frac{1}{n^{c+1}}\left\{\frac{n}{c} + \sum_{i=1}^{n} F^{c+1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \left(1-F(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^{c+1}\right. \\
&\quad \left. - \left(1+\frac{1}{c}\right)\left[y_i F^c(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + (1-y_i)\left(1-F(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^c\right]\right\}\,.
\end{aligned}$$

La constante c juega el papel de constante de calibración ya que, como se observa en Basu *et al.* (2017), el estimador de máxima verosimilitud se obtiene considerando la distancia $\Delta_0(\widehat{\boldsymbol{\pi}}, \boldsymbol{\pi}(\boldsymbol{\beta})) = \lim_{c\to 0} \Delta_c(\widehat{\boldsymbol{\pi}}, \boldsymbol{\pi}(\boldsymbol{\beta}))$. Es fácil ver que los estimadores de mínima divergencia $\widehat{\boldsymbol{\beta}}_{\text{DIV}}$ también pertenecen a la familia de estimadores (3.1) eligiendo como función ρ la función $\rho_{\text{DIV}} = (1+1/c)\{1-\exp(-ct)\}$. Observemos que si $c = 1$ obtenemos el estimador de mínimos cuadrados $\widehat{\boldsymbol{\beta}}_{\text{MC}}$.

Otra función ρ que será considerada en esta tesis es la definida por Croux y Haesbroeck (2003). Para definirla, recordemos que decimos que un conjunto de observaciones $(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)$ está *completamente separado* si existe un vector $\boldsymbol{\xi} \in \mathbb{R}^p$ tal que $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\xi} > 0$ si $Y_i = 1$ y $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\xi} < 0$ cuando $Y_i = 0$. Un conjunto de datos que no está completamente separado se dice *casi completamente separado* si existe un vector $\boldsymbol{\xi} \in \mathbb{R}^p$ tal que $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\xi} \geq 0$ si $Y_i = 1$, $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\xi} \leq 0$ cuando $Y_i = 0$ y existe $j \in \{1, \ldots, n\}$ tal que $\mathbf{X}_j^{\mathrm{T}}\boldsymbol{\xi} = 0$. Se dice que un conjunto de observaciones tiene superposición si no hay *separación completa* ni *separación casi completa*. Es sabido que los estimadores de máxima verosimilitud están bien definidos cuando la muestra tiene superposición. Para garantizar la existencia de los $M-$estimadores dados en (3.1) bajo esta misma condición, Croux y Haesbroeck (2003) sugieren tomar $\rho = \rho_c$ donde

$$\rho_c(t) = \begin{cases} te^{-\sqrt{c}} & \text{if } \ t \leq c \\ -2e^{-\sqrt{t}}\left(1+\sqrt{t}\right) + e^{-\sqrt{c}}\left(2\left(1+\sqrt{c}\right)+c\right) & \text{if } \ t > c\,, \end{cases} \tag{3.5}$$

donde c es la constante de calibración que permite determinar la eficiencia del $M-$estimador respecto del estimador de máxima verosimilitud. En este caso, tenemos que

$$\psi_c(t) = \rho_c'(t) = \begin{cases} e^{-\sqrt{c}} & \text{if } \ t \leq c \\ e^{-\sqrt{t}} & \text{if } \ t > c\,. \end{cases} \tag{3.6}$$

Dada un pérdida ρ, sea

$$\begin{aligned}
\phi(y,t) &= \rho(d(y,t)) + G(F(t)) + G(1-F(t)) \\
&= y\rho\left(-\log\left[F(t)\right]\right) + (1-y)\rho\left(-\log\left[1-F(t)\right]\right) + G(F(t)) + G(1-F(t)) \qquad (3.7)
\end{aligned}$$

donde $G(t) = \int_0^t \psi(-\log u)\, du$ y definamos $\Psi(0,t) = \partial\phi(0,t)/\partial t$. Supongamos que $\phi(0,t)$ es una función no decreciente, tal que $\lim_{t\to-\infty} \phi(0,t) = 0$. Croux y Haesbroeck (2003) prueban que las siguientes tres condiciones son suficientes para la existencia del estimador:

(a) la muestra tiene superposición,

(b) existe una constante L_0 tal que $\Psi(0,\cdot)$ es creciente en el intervalo $(-\infty, L_0]$ y monótona en el intervalo $[L_0, +\infty)$.

(c) para todo $t > 0$ se tiene que

$$\lim_{s\to+\infty} \frac{\Psi(0,ts)}{\Psi(0,-s)} = +\infty.$$

La Figura 3.2 presenta los gráficos de las funciones $\Psi(0,\cdot)$ para cuatro elecciones distintas de la función ρ. Las líneas celeste, lila , rosa y verde corresponden a $\rho(t) = t$, $\rho(t) = 1 - \exp(-t)$ que da origen a los estimadores de mínimos cuadrados, $\rho = \rho_c$ la función dada por Croux y Haesbroeck (2003) y $\rho = \rho_{\text{DIV}}$ que genera los estimadores de mínima divergencia, respectivamente. En estos últimos dos casos se tomó $c = 0.5$. Observemos que la cola derecha de la función ρ_c definida en (3.5) es más pesada que la cola izquierda, lo que sugiere que la condición (c) vale para esta elección de ρ.

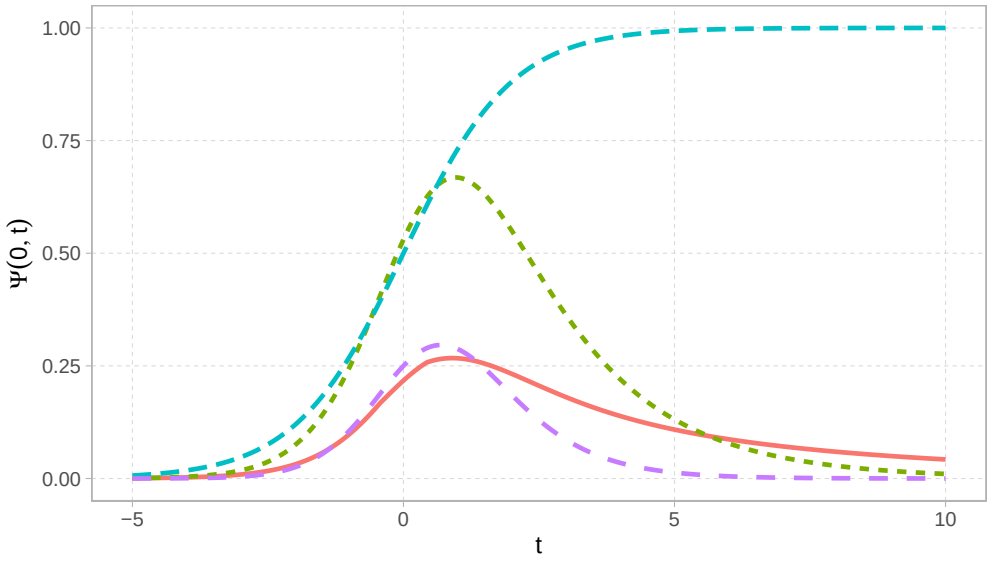


Figura 3.2: Gráfico de la función $\Psi(0,\cdot)$ para diferentes funciones ρ. Las líneas celeste, rosa, verde y lila corresponden a $\rho(t) = t$, $\rho = \rho_c$ definida en (3.5) con $c = 0.5$, $\rho = \rho_{\text{DIV}}$ y $\rho(t) = 1 - \exp(-t)$, respectivamente.

3.1.2. El $M-$estimador pesado

Para controlar el efecto de los puntos de alta palanca y obtener estimadores con función de influencia acotada, Croux y Haesbroeck (2003) proponen una versión pesada del estimador dado en (3.1). Para definirlos sea $w : \mathbb{R}^p \to \mathbb{R}$ una función de peso, es decir, una función tal que $0 \le w(\mathbf{x}) \le 1$. El $M-$estimador pesado se define como

$$\widehat{\boldsymbol{\beta}}_{\text{WM}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} w(\mathbf{X}_i) \left\{ \rho(d(Y_i, \mathbf{X}_i^{\text{T}}\boldsymbol{\beta})) + G(F(\mathbf{X}_i^{\text{T}}\boldsymbol{\beta})) + G(1 - F(\mathbf{X}_i^{\text{T}}\boldsymbol{\beta})) \right\}. \tag{3.8}$$

Usualmente, los pesos $w(\mathbf{X}_i)$ se basan en una distancia de Mahalanobis robusta de las variables explicativas, es decir, dependen de la distancia entre $\mathbf{X}_i^\star$ y un estimador de posición robusto de los datos, donde $\mathbf{X} = \begin{pmatrix} 1 \\ \mathbf{X}^\star \end{pmatrix}$ cuando el modelo incluye ordenada al origen y $\mathbf{X} = \mathbf{X}^\star$, en caso contrario. Más precisamente, sea $D^2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$, el cuadrado de la distancia de Mahalanobis donde suponemos que $\mathbf{X} = \mathbf{X}^\star$, por simplicidad. Dada una función $W : \mathbb{R} \to \mathbb{R}$ tal que $0 \leq W \leq 1$, podemos definir $w(\mathbf{x}) = W(D^2(\mathbf{x}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}^{-1}))$, donde para obtener una medida robusta de la palanca de las covariables, $\widehat{\boldsymbol{\mu}}$ es un estimador robusto de posición como la mediana espacial, y $\widehat{\boldsymbol{\Sigma}}^{-1}$ es un estimador robusto de la inversa $\boldsymbol{\Sigma}^{-1}$ de la matriz de dispersión.

3.1.3. Propiedades del estimador $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$

Notemos que los $M-$estimadores pesados definidos en (3.8) pueden ser escribirse como

$$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} L_n(\boldsymbol{\beta}) \,, \tag{3.9}$$

donde

$$L_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) w(\mathbf{X}_i) \,, \tag{3.10}$$

$\phi : \mathbb{R}^2 \to \mathbb{R}$ está dada por (3.7) y eventualmente $w(\mathbf{x}) \equiv 1$ para la versión no pesada. Si ρ tiene derivada continua, la función ϕ es contínuamente diferenciable con respecto a su segundo argumento y $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$ es una solución de las ecuaciones en derivadas

$$\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\, w(\mathbf{X}_i)\, \mathbf{X}_i = \mathbf{0} \,.$$

Por otra parte, cabe observar que

$$\Psi(y,t) = \frac{\partial}{\partial t}\phi(y,t) = -\,[y - F(t)]\,\nu(t) \,, \tag{3.11}$$

donde

$$\nu(t) = \psi\left(-\log F(t)\right)[1 - F(t)] + \psi\left(-\log\left[1 - F(t)\right]\right) F(t) \,. \tag{3.12}$$

Observemos también que las funciones ϕ y Ψ satisfacen $\phi(0,s) = \phi(1,-s)$ y $\Psi(0,s) = -\Psi(1,-s)$. Más aún, en virtud de (3.11) tenemos que

$$\mathbb{E}[\Psi(Y_1, \mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta})|\mathbf{X}_1] = \mathbf{0} \,, \tag{3.13}$$

que es usualmente conocida como la propiedad de Fisher–consistencia condicional.

El siguiente resultado muestra que los $M-$estimadores son efectivamente Fisher–consistentes, lo cual asegura que el procedimiento de estimación es asintóticamente insesgado y estima la cantidad de interés. Aunque, cuando $w(\mathbf{x}) \equiv 1$, una demostración de este resultado se puede encontrar en el Teorema 2.2 de Bianco y Yohai (1996), por completitud presentaremos una con un enfoque levemente distinto.

Por simplicidad, llamaremos $(Y, \mathbf{X})$ a un vector aleatorio con la misma distribución que las observaciones $(Y_i, \mathbf{X}_i)$, es decir, tal que $Y|\mathbf{X} \sim Bi(1, F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0))$.

Teorema 3.1. *Sea $\phi : \mathbb{R}^2 \to \mathbb{R}$ la función dada por (3.7) donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ cumple que $\rho(0) = 0$. Supongamos que $w(\mathbf{X})$ es una función de peso acotada y no negativa y que*

a) $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ es acotada con derivada ψ continua y acotada.

b) $\psi(t) \geq 0$ y existe un $c \geq \log 2$ tal que $\psi(t) > 0$ para todo $0 < t < c$.

Supongamos también que

$$\mathbb{P}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \cup w(\mathbf{X}) = 0) < 1 \text{ para todo } \boldsymbol{\alpha} \in \mathbb{R}^p \, , \boldsymbol{\alpha} \neq \mathbf{0} \, . \tag{3.14}$$

Luego, para todo $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, se tiene que $\mathbb{L}(\boldsymbol{\beta}_0) < \mathbb{L}(\boldsymbol{\beta})$ donde

$$\mathbb{L}(\boldsymbol{\beta}) = \mathbb{E}[\phi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X})] \, . \tag{3.15}$$

Demostración. Al igual que en el Teorema 2.2 de Bianco y Yohai (1996), tomando esperanza condicional tenemos que

$$\mathbb{L}(\boldsymbol{\beta}) = \mathbb{E}[\phi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X})] = \mathbb{E}\left[\mathbb{E}\phi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X})|\mathbf{X}\right] = \mathbb{E}[\phi(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X})] \, .$$

Dado un valor fijo de $\mathbf{x}$, denotamos $t = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ y $t_0 = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$. Supongamos que $t \neq 0$ y $t_0 \neq 0$. Mostraremos que la función $\phi(F(t_0), t)$ alcanza su único mínimo cuando $t = t_0$. Por simplicidad llamamos $\Phi(t) = \phi(F(t_0), t)$. Cálculos sencillos muestran que

$$\Phi'(t) = -(F(t_0) - F(t))\nu(t) \qquad \Phi''(t) = F(t)(1 - F(t))\nu(t) \, - \, (F(t_0) - F(t))\nu'(t) \, ,$$

donde $\nu(t)$ está definida en (3.12). Por lo tanto, $\Phi'(t_0) = 0$ y $\Phi''(t_0) = F(t_0)(1-F(t_0))\nu(t_0) > 0$ pues $\nu(t_0) > 0$. Más aún, como $t \neq 0$ y $t_0 \neq 0$, usando la condición b), tenemos que $\Phi'(t) > 0$ si $t > t_0$ y $\Phi'(t) < 0$ si $t < t_0$, lo que implica que Φ tiene un único mínimo en t_0. Luego, $\phi(F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) > \phi(F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$ para todo $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ tal que $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \neq 0$ y $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \neq 0$ lo que junto con la condición (3.14), concluye la demostración. ■

Denotaremos

$$\chi(y, t) = \partial\Psi(y, t)/\partial t = F(t)(1 - F(t))\nu(t) - (y - F(t))\nu'(t) \tag{3.16}$$

que existe siempre y cuando ρ tenga una segunda derivada continua. Si esto ocurre, es fácil ver que $\chi(0, s) = \chi(1, -s)$.

El comportamiento asintótico de $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$ y $\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$ se deduce del Teorema 3.3 de Bianco y Martinez (2009) y el Teorema 2.2 de Basu *et al.* (2017), respectivamente, en el caso en el que la dimensión p es fija y no se hacen supuestos acerca de cuán ralo es el vector $\boldsymbol{\beta}_0$. Más precisamente, estos autores muestran que $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(\mathbf{0}_p, \boldsymbol{\Sigma})$ donde $\boldsymbol{\Sigma} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ y las matrices $\mathbf{A}$ y $\mathbf{B}$ están definidas como

$$\mathbf{A} = \mathbb{E}\left(\chi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)w(\mathbf{X})\,\mathbf{X}\mathbf{X}^{\mathrm{T}}\right) \qquad \mathbf{B} = \mathbb{E}\left(\Psi^2(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)w^2(\mathbf{X})\,\mathbf{X}\mathbf{X}^{\mathrm{T}}\right) \, .$$

Notemos que este resultado es compatible con el Teorema 2.2, ya que cuando $\rho(t) = t$ se tiene que $\nu(t) = 1$ y esto implica que $\mathbf{A} = \mathbf{B}$.

El siguiente Lema muestra una forma equivalente de definir las matrices $\mathbf{A}$ y $\mathbf{B}$. Se omite su demostración, ya que es una consecuencia directa de (3.11) y de la expresión (3.16).

Lema 3.2. *Sea $\phi : \mathbb{R}^2 \to \mathbb{R}$ la función definida en (3.7), donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ tiene segunda derivada continua. Luego,*

$$\mathbf{A} = \mathbb{E}\left(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\left[1 - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\nu(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)w(\mathbf{X})\,\mathbf{X}\mathbf{X}^{\mathrm{T}}\right) \tag{3.17}$$

$$\mathbf{B} = \mathbb{E}\left(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\left[1 - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\nu^2(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)w^2(\mathbf{X})\,\mathbf{X}\mathbf{X}^{\mathrm{T}}\right). \tag{3.18}$$

Cabe mencionar que al considerar estimadores que acotan la devianza, la esperanza condicional $\mathbb{E}\left(\chi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\,|\mathbf{X}\right) = F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\left[1 - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\nu(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)$ no depende de ρ''.

3.2. $M-$Estimadores ralos

Los métodos definidos en la Sección 3.1 no dan estimaciones ralas de $\boldsymbol{\beta}_0$. Este hecho tiene dos consecuencias. Por un lado, el procedimiento no permite seleccionar variables automáticamente. Por el otro lado, las propiedades de robustez y eficiencia del estimador pueden tener un mal comportamiento cuando p/n es grande. Por este motivo, tiene sentido agregar a la función objetivo una función de penalización, tanto en el caso en que el vector $\boldsymbol{\beta}_0$ es ralo como cuando p tiende a infinito junto con n y se supone que sólo un pequeño número de componentes están activas. Como en la Sección 2.3, en la presentación de los $M-$estimadores pesados penalizados y en el estudio de sus propiedades asintóticos supondremos que el modelo no contiene ordenada al origen. De esta forma, los $M-$estimadores pesados penalizados se definen como

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X}_i) + I_\lambda(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, L_n(\boldsymbol{\beta}) + I_\lambda(\boldsymbol{\beta}), \tag{3.19}$$

donde $L_n(\boldsymbol{\beta})$ está definida como en (3.10) y la penalización $I_{\lambda_n}(\boldsymbol{\beta})$ es una función fija positiva que depende de un parámetro de ajuste λ que mide la complejidad del modelo de regresión logística.

Además de las penalizaciones descriptas en la Sección 2.3, el usuario puede elegir la penalización que introduciremos en la próxima sección.

3.2.1. Penalización Signo

Recordemos que la variable respuesta en el modelo de regresión logística es binaria y por lo tanto, acotada. Esto implica que al considerar, por ejemplo, el estimador de mínimos cuadrados definido en (3.3), el primer término en (3.19) será siempre menor o igual a 1. Por lo tanto, si la función de penalización no es acotada, puede ocurrir que este término domine el comportamiento de la función objetivo para ciertos valores del parámetro de ajuste λ.

En esta tesis, como alternativa a las penalizaciones SCAD y MCP ya mencionadas, introducimos la penalización **Signo**, definida como

$$I_\lambda(\boldsymbol{\beta}) = \lambda\frac{\|\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}\|_2}\mathbb{I}_{\boldsymbol{\beta}\neq\mathbf{0}} = \lambda\left\|\frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2}\right\|_1 \mathbb{I}_{\boldsymbol{\beta}\neq\mathbf{0}}\,, \tag{3.20}$$

que es una nueva propuesta en este contexto y corresponde a aplicar la penalidad LASSO a las direcciones en la bola unidad en ℓ_2. Esta función de penalización tiene las siguientes propiedades:

- al usarla en el problema (3.19), produce una regla de truncado, es decir, produce estimaciones ralas de $\boldsymbol{\beta}_0$,
- alcanza su mínimo cuando únicamente una coordenada de $\boldsymbol{\beta}$ es distinta de cero,
- alcanza su máximo cuando todas las coordenadas de $\boldsymbol{\beta}$ son iguales y distintas de cero,
- es invariante por escala, es decir, si $\boldsymbol{\beta} \neq \mathbf{0}$, entonces $I_\lambda(\boldsymbol{\beta}) = I_\lambda(c\boldsymbol{\beta})$ para todo $c > 0$.

Para ilustrar cómo esta penalización reduce el sesgo introducido por LASSO, generamos un conjunto de datos de tamaño $n = 300$ que cumple el modelo de regresión logística (2.2) cuando las covariables tienen dimensión $p = 20$, $(\boldsymbol{\beta}_0)_j = \mathbb{I}_{j\leq 5}$ y $\mathbf{X}_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$. Para esa muestra, estudiamos el comportamiento de $f(c)$ siendo $f(c) = L_n(c\boldsymbol{\beta}_0) + I_\lambda(c\boldsymbol{\beta}_0)$ cuando elegimos como $I_\lambda(\cdot)$ la penalidad LASSO y la penalización Signo con $\lambda = 0.04$. La Figura 3.3 da un gráfico de la función f, las líneas roja y celeste corresponden a la penalización LASSO y Signo, respectivamente. La línea punteada vertical indica el punto $c = 1$ que es el objetivo a estimar. En ambos casos, se eligió $\lambda = 0.04$. Se puede ver que la curva correspondiente a la penalización Signo alcanza su mínimo cerca de $c = 1$, mientras que la que corresponde a la penalización LASSO se minimiza cerca de $c = 0.5$. El comportamiento observado sugiere que, incluso en escenarios sin contaminación, LASSO achica en módulo a los coeficientes correspondientes a componentes no nulas.

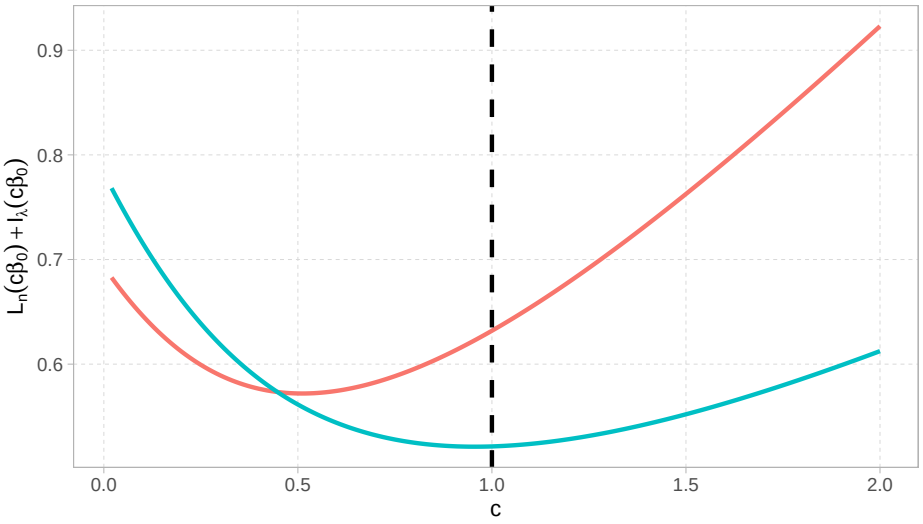


Figura 3.3: Grafico de la función $f(c) = L_n(c\boldsymbol{\beta}_0) + I_\lambda(c\boldsymbol{\beta}_0)$ con $\lambda = 0.04$. Las líneas roja y celeste corresponden a la penalización LASSO y Signo, respectivamente. La línea punteada vertical indica el punto $c = 1$ que es el objetivo a estimar.

3.2.2. Elección del parámetro de regularización λ

Como se discute, por ejemplo, en Chi y Scott (2014), Efron *et al.* (2004) y Meinshausen (2007), la elección del parámetro de ajuste o regularización λ juega un papel importante al estimar modelos ralos. Por otra parte, utilizar un método no robusto de selección de dicho parámetro puede llevar a estimadores que no resulten resistentes a datos atípicos. Este hecho ha sido ampliamente estudiado en regresión no paramétrica, donde el método de convalidación cruzada basado en mínimos

cuadrados resulta prácticamente constante en su dominio cuando hay outliers. Una consecuencia de este hecho es que ante la presencia de datos atípicos, las ventanas obtenidas pueden producir sobresuavizado o subsuavizado y el estimador obtenido no resulta robusto aún cuando se utilicen $M-$estimadores locales (ver Wang y Scott, 1994). Como se mostrará en la Sección 7.2.4, esta falta de robustez del procedimiento clásico de selección de λ motiva la necesidad de elegir el parámetro de penalización en forma robusta. En esta sección, introduciremos un procedimiento basado en $K-$fold robusto que permite dar elecciones más estables de λ.

Como es usual, primero dividimos aleatoriamente la muestra en K subconjuntos disjuntos que tengan aproximadamente la misma cantidad de observaciones. Sea $\mathcal{C}_j$, $1 \leq j \leq K$, el conjunto de índices del $j-$ésimo subconjunto de dicha partición y $n_j \geq 2$ su cardinal. De este modo, $\bigcup_{j=1}^{K} \mathcal{C}_j = \{1, \ldots, n\}$ y $\sum_{j=1}^{K} n_j = n$. Sea $\widetilde{\Lambda} \subset \mathbb{R}$ el conjunto de posibles valores para λ que se considerará y llamemos $\widehat{\boldsymbol{\beta}}_{\lambda}^{(-j)}$ al estimador robusto penalizado de $\boldsymbol{\beta}_0$, calculado con el parámetro de regularización $\lambda \in \widetilde{\Lambda}$ y sin usar las observaciones con índices en $\mathcal{C}_j$.

Fijado λ, para cada $i = 1, \ldots, n$ tal que $i \in \mathcal{C}_j$ para algún $j = 1, \ldots, K$, los residuos de predicción $\widehat{d}_{i,\lambda}$ se definen como

$$\widehat{d}_{i,\lambda} \,=\, d(Y_i, \mathbf{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\lambda}^{(-j)}) \,.$$

El método de convalidación cruzada clásico construye estimadores adaptivos minimizando la función

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \widehat{d}_{i,\lambda} \,, \tag{3.21}$$

que es la función objetivo utilizada usualmente para los estimadores penalizados clásicos definidos en (2.5). Sin embargo, este criterio puede verse afectado por observaciones mal clasificadas, aún cuando $\boldsymbol{\beta}_0$ se estime mediante el $M-$estimador pesado dado en (3.19), ya que los datos atípicos pueden tener residuos de predicción grandes con mucha influencia sobre $CV(\lambda)$. Para resolver este problema, es natural utilizar la misma función de pérdida ϕ y los mismos pesos w que se usan en el cómputo del $M-$estimador pesado. De esta forma, el criterio de convalidación cruzada robusto elige el parámetro de penalización que minimiza

$$RCV(\lambda) = \frac{1}{n} \sum_{1 \leq j \leq K} \sum_{i \in \mathcal{C}_j} \phi(Y_i, \mathbf{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\lambda}^{(j)}) \, w(\mathbf{X}_i) \,. \tag{3.22}$$

En particular, eligiendo $K = n$ obtenemos el criterio de convalidación cruzada que elimina solo una observación a la vez, que es un criterio popular aunque con mayor costo computacional.

Capítulo 4

Algunos resultados de procesos empíricos

En los Capítulos 5 y 6 se presentarán resultados relativos a las propiedades asintóticas de los estimadores propuestos en la Sección 3.2, tanto cuando p es fijo como cuando $\lim_{n\to\infty} p = \infty$. Muchas de las demostraciones en dichos capítulos se basan en resultados de procesos empíricos. Es por eso que, en este capítulo, damos una brevísima introducción a esta teoría y enunciamos algunos de los resultados que usaremos. Se puede ver una referencia completa de estos temas en Van der Vaart y Wellner (1996), van de Geer (2000) y Kosorok (2008).

Definición 4.1. Sea $\mathcal{H}$ un subconjunto de un espacio de Banach con norma $\|\cdot\|$.

(a) El **número de cubrimiento** $N(\varepsilon, \mathcal{H}, \|\cdot\|)$ es el mínimo número de bolas abiertas de radio ε (con norma $\|\cdot\|$) necesarias para cubrir al conjunto $\mathcal{H}$. Análogamente, si $\mathcal{T}$ es un subconjunto de un espacio métrico con distancia d, $N(\varepsilon, \mathcal{T}, d)$ es el mínimo número de bolas abiertas de radio ϵ (con respecto a la distancia d) necesarias para cubrir $\mathcal{T}$.

(b) Supongamos que $\mathcal{H}$ es una clase de funciones de $\mathbb{R}^m$ a $\mathbb{R}$. Dada una norma $\|\cdot\|$ sobre $\mathcal{H}$, un $\varepsilon-$**corchete** $[\ell, u]$ es un par de funciones $\ell, u : \mathbb{R}^m \to \mathbb{R}$ con $\ell(\mathbf{x}) \leq u(\mathbf{x})$, para todo $\mathbf{x} \in \mathbb{R}^m$, y $\|\ell - u\| < \epsilon$.

Diremos que un $\varepsilon-$corchete **cubre** un elemento g si $\ell(\mathbf{x}) \leq g(\mathbf{x}) \leq u(\mathbf{x})$ para todo $\mathbf{x} \in \mathbb{R}^m$.

El **número de cubrimiento corchete** $N_{[\,]}(\varepsilon, \mathcal{H}, \|\cdot\|)$ es el mínimo número de $\varepsilon-$corchetes necesario para cubrir todas las funciones del conjunto $\mathcal{H}$.

(c) Cuando $\mathcal{H}$ es un conjunto de funciones definidas en $\mathbb{R}^m$ para algún $m \in \mathbb{N}$ y tomando valores en $\mathbb{R}$, decimos que una función $H : \mathbb{R}^m \to \mathbb{R}$ es una **envolvente** de $\mathcal{H}$ si $|h(\mathbf{x})| \leq H(\mathbf{x})$ para todo $\mathbf{x} \in \mathbb{R}^m$ y $h \in \mathcal{H}$.

El siguiente Lema da la relación entre el número de cubrimiento y el cubrimiento corchete.

Lema 4.1. *Sea $\mathcal{H}$ una clase de funciones de $\mathbb{R}^m$ a $\mathbb{R}o$ con norma $\|\cdot\|$ y sea $\varepsilon > 0$, entonces, $N(\varepsilon/2, \mathcal{H}, \|\cdot\|) \leq N_{[\,]}(\varepsilon, \mathcal{H}, \|\cdot\|)$.*

El siguiente Lema, que corresponde al Lema 2.5 de van de Geer (2000), da una cota para el número de cubrimiento de una bola de $\mathbb{R}^p$:

Lema 4.2. *Sea $\mathcal{B}_m(K)$ una bola en $\mathbb{R}^m$ de radio K y sea $\varepsilon > 0$. Entonces,*

$$N(\varepsilon, \mathcal{B}_m(K), \|\cdot\|_2) \leq \left(\frac{4\,K + \varepsilon}{\varepsilon}\right)^m .$$

Una de las herramientas esenciales que provee esta teoría es la llamada **Ley de los Grandes Números Uniforme** (LGNU), que definimos a continuación.

Definición 4.2. Sean $X_1, \ldots, X_n$ elementos aleatorios i.i.d. que toman valores en un espacio de Banach $\mathcal{W}$ y sea $\mathcal{G}$ una familia de funciones definidas sobre $\mathcal{W}$ a valores en $\mathbb{R}$. Decimos que la familia $\mathcal{G}$ cumple la LGNU si

$$\mathbb{P}^* \left(\lim_{n\to\infty} \sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X_1) \right| = 0 \right) = 1 .$$

donde $\mathbb{P}^*$ indica la probabilidad exterior.

Condiciones sobre los números de cubrimiento y números de cubrimiento corchete de una familia de funciones para garantizar que la clase de funciones cumple la LGNU pueden verse en van de Geer (2000). Un caso particular de familias de funciones que cumplen esta ley son las llamadas **clases** $VC-$**subgrafo**, que definimos a continuación.

Definición 4.3. Sea $\mathcal{W}$ un espacio de Banach y sea $\mathcal{D}$ una familia de subconjuntos de $\mathcal{W}$.

(a) Dados elementos $\xi_1, \ldots, \xi_n \in \mathcal{W}$, denotamos $\Delta^{\mathcal{D}}(\xi_1, \ldots, \xi_n) = \#\{D \cap \{\xi_1, \ldots, \xi_n\}\}$. Cuando $\Delta^{\mathcal{D}}(\xi_1, \ldots, \xi_n) = 2^n$, decimos que $\mathcal{D}$ **desmenuza** al conjunto $\{\xi_1, \ldots, \xi_n\}$.

(b) Se define $m^{\mathcal{D}}(n) = \sup\{\Delta^{\mathcal{D}}(\xi_1, \ldots, \xi_n) \ : \ \xi_1, \ldots, \xi_n \in \mathcal{W}\}$.

(c) Llamamos índice de Vapnik-Chervonenkis o **índice VC** de la clase $\mathcal{D}$ a

$$V(\mathcal{D}) = \text{ínf}\{n \geq 1 : m^{\mathcal{D}}(n) < 2^n\} .$$

Decimos que $\mathcal{D}$ es de **clase** VC si su índice es finito, o sea, $V(\mathcal{D}) < \infty$.

(d) Sea $\mathcal{G}$ es una familia de funciones definidas sobre $\mathcal{W}$ que toman valores en $\mathbb{R}$. Dada $g : \mathcal{W} \to \mathbb{R}$, $g \in \mathcal{G}$, el subgrafo de g es el subconjunto de $\mathcal{W} \times \mathbb{R}$ definido por

$$\{(w, t) \in \mathcal{W} \times \mathbb{R} : t < g(w)\} .$$

Decimos que $\mathcal{G}$ es una clase $VC-$**subgrafo** si la familia de subgrafos de $\mathcal{G}$ es de clase VC. El índice $V(\mathcal{G})$ de la clase de funciones es el índice de la familia de subgrafos de $\mathcal{G}$.

Intuitivamente, una familia $\mathcal{D}$ de subconjuntos de $\mathcal{W}$ es de clase VC si existe un n tal que ningún conjunto de n puntos de $\mathcal{W}$ puede ser desmenuzado por $\mathcal{D}$. El siguiente teorema, que corresponde al Corolario 3.12 de van de Geer (2000), muestra una utilidad importante de las clases de funciones $VC-$subgrafo.

Teorema 4.3. *Sean $X_1, \ldots, X_n$ elementos aleatorios i.i.d en $\mathcal{W}$ y sea $\mathcal{G} = \{g : \mathcal{W} \to \mathbb{R}\}$ una familia de funciones $VC-$subgrafo. Supongamos que $\mathcal{G}$ tiene una función envolvente G tal que $\mathbb{E}|G(X_1)| < \infty$, entonces $\mathcal{G}$ cumple la LGNU.*

El Teorema 4.4 corresponde al Teorema 2.6.7 de van der Vaart y Wellner (1996) y da una cota del número de cubrimiento de una familia $VC-$subgrafo. Denotamos como $\|\cdot\|_{r,\mathbb{Q}}$ a la norma $L_r(\mathbb{Q})$ para $1 \leq r \leq \infty$.

Teorema 4.4. *Sea $\mathcal{G}$ una familia $VC-$subgrafo de índice $V(\mathcal{G})$ y sea G una envolvente de $\mathcal{G}$. Sea $r \geq 1$ y $\mathbb{Q}$ una probabilidad tal que $\|G\|_{r,\mathbb{Q}} > 0$. Luego, existe una constante universal K tal que*

$$N(\varepsilon\|G\|_{r,\mathbb{Q}}, \mathcal{G}, \|\cdot\|_{r,\mathbb{Q}}) \leq K\, V(\mathcal{G})\, (16e)^{V(\mathcal{G})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{G})-1)}$$

para $0 < \varepsilon < 1$.

El proceso empírico evaluado en g se define como $\mathbb{G}_n g = \sqrt{n}(P_n g - \mathbb{P}g)$, donde

$$P_n g = \frac{1}{n}\sum_{i=1}^{n} g(X_i), \qquad \mathbb{P}g = \int g d\mathbb{P} = \mathbb{E}g(X_1).$$

En algunas situaciones serán necesarias cotas explícitas sobre el proceso empírico

$$\|\mathbb{G}_n\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |\mathbb{G}_n g| = \sup_{g\in\mathcal{G}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X_1) \right|,$$

para lo cual usaremos el siguiente lema que corresponde al Teorema 2.14.1 de Van der Vaart y Wellner (1996).

Lema 4.5. *Sean $\mathbf{V}_1, \ldots, \mathbf{V}_n$ vectores aleatorios i.i.d. en $\mathbb{R}^m$. Sea $\mathcal{G}$ una familia de funciones $g : \mathbb{R}^m \to \mathbb{R}$ con envolvente G tal que $\|G\|_{2,\mathbb{P}}^2 = \mathbb{E}G^2(\mathbf{V}_1) < \infty$. Entonces, existe una constante $M > 0$ que no depende de n tal que*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \left| \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{V}_i) - \mathbb{E}g(\mathbf{V}_1)\right)\right|\right] \leq M\, \|G\|_{2,\mathbb{P}}\, J(1, \mathcal{G})$$

con

$$J(\delta, \mathcal{G}) = \sup_{\mathbb{Q}} \int_0^{\delta} \sqrt{1 + \log N(\epsilon\|G\|_{2,\mathbb{Q}}, \mathcal{G}, \|\cdot\|_{2,\mathbb{Q}})} d\epsilon\,,$$

donde el supremo se toma sobre todas las medidas discretas $\mathbb{Q}$ con $\|G\|_{2,\mathbb{Q}} > 0$.

En esta tesis, también serán necesarias cotas para los incrementos de un proceso empírico basado en funciones de pérdida Lipschitz. Los resultados que enunciamos a continuación corresponden a los Lemas 14.19 y 14.20 de Bühlmann y van de Geer (2011).

Consideremos vectores aleatorios independientes e idénticamente distribuidos $(Y_i, \mathbf{X}_1), \ldots, (Y_i, \mathbf{X}_n)$ donde $\mathbf{X}_i \in \mathbb{R}^p$, $Y_i \in \mathbb{Y}$. Sea $\gamma : \mathbb{R}^2 \to \mathbb{R}$ una función tal que para todo $y \in \mathbb{Y}$

$$|\gamma(y, s) - \gamma(y, \widetilde{s})| \leq C_\gamma |s - \widetilde{s}| \quad \forall s, \widetilde{s} \in \mathbb{R} \tag{4.1}$$

donde C_γ es una constante que no depende de y. Consideremos el proceso empírico

$$v_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\gamma(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) - \mathbb{E}\gamma(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right], \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

El siguiente Lema cuya demostración es análoga a la de los Lemas 14.19 y 14.20 de Bühlmann y van de Geer (2011), permite acotar los incrementos de este proceso empírico.

Lema 4.6. *Sea $\boldsymbol{\beta}^* \in \mathbb{R}^p$ fijo. Supongamos que vale (4.1), $\mathbb{E}\|\mathbf{X}_1\|_2^2 < \infty$ y definamos la distancia $D^2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbb{E}\left[\mathbf{X}_1^{\mathrm{T}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right]^2$.*

(a) Para todo $\delta > 0$ se tiene que

$$\mathbb{E}\left[\sup_{D(\boldsymbol{\beta},\boldsymbol{\beta}^*)\leq\delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}^*)|\right] \leq 4\ \delta\ C_\gamma\ \sqrt{\frac{p}{n}}.$$

(b) Para todo $M > 0$ se tiene que

$$\mathbb{E}\left[\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_1\leq M} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}^*)|\right] \leq 4M\ C_\gamma\ \sqrt{\frac{2\log(2p)}{n}}\ \mathbb{E}\left(\max_{1\leq j\leq p}\ \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right)$$

donde X_{ij} es la $j-$ésima coordenada de $\mathbf{X}_i$.

Por último, enunciamos un resultado que acota el número de cubrimiento corchete de una familia de funciones $\mathcal{F}$ por el número de cubrimiento de la familia que indexa a $\mathcal{F}$. Este lema corresponde al Teorema 2.7.11 de Van der Vaart y Wellner (1996).

Lema 4.7. *Sea $\mathcal{T}$ un subconjunto de un espacio métrico dotado con una distancia d. Sea $\mathcal{F} = \{f_t : \mathcal{W} \to \mathbb{R} : t \in \mathcal{T}\}$ una familia de funciones definidas sobre $\mathcal{W}$ e indexada en $\mathcal{T}$. Supongamos que existe una función $F : \mathcal{W} \to \mathbb{R}$ tal que $|f_s(w) - f_t(w)| \leq d(s,t)\ F(x)$ para todo $s, t \in \mathcal{T}$ y todo $w \in \mathcal{W}$. Luego, para cualquier norma $\|\cdot\|$ definida sobre $\mathcal{F}$, se tiene que*

$$N_{[]}(2\ \varepsilon\ \|F\|, \mathcal{F}, \|\cdot\|) \leq N(\varepsilon, \mathcal{T}, d).$$

Capítulo 5

Resultados asintóticos para p fijo

En este Capítulo, presentamos resultados asintóticos para los estimadores propuestos en la Sección 3.2 cuando la dimensión p es fija, pero el modelo es ralo. Bajo condiciones de regularidad, se prueba que los $M-$estimadores pesados penalizados son consistentes y seleccionan variables correctamente. Por otra parte, obtenemos expresiones para su distribución asintótica. En particular, mostramos que los $M-$estimadores obtenidos con la penalidad SCAD y MCP tienen la propiedad oráculo. Aunque nuestro interés se focaliza en penalizaciones acotadas, los resultados que presentamos son generales y pueden aplicarse, por ejemplo, a la penalización Ridge, Bridge o Elastic Net.

Como se menciona en la Introducción, las demostraciones de este capítulo fueron relegadas al final del mismo, a los Apéndices **A** a **D**. A lo largo de este Capítulo, supondremos que $(Y_i, \mathbf{X}_i) \in \{0,1\} \times \mathbb{R}^p$, $1 \leq i \leq n$, son vectores aleatorios independientes e idénticamente distribuidos que cumplen el modelo de regresión logística (2.2) y están definidos sobre el espacio de probabilidad $(\Omega, \mathcal{B}, \mathbb{P})$.

5.1. Consistencia

Consideraremos el siguiente conjunto de hipótesis respecto de la función ρ usada en (3.7):

R1 $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ es acotada, cumple $\rho(0) = 0$ y tiene derivada ψ acotada y continua.

R2 $\psi(t) \geq 0$ y existe un $c \geq \log 2$ tal que $\psi(t) > 0$ para todo $0 < t < c$.

R3 ρ es dos veces diferenciable. Además, ψ y $\psi' = \rho''$ son funciones acotadas.

Observación 5.1. Recordemos que para los $M-$estimadores que acotan la deviance

$$\phi(y,t) = \rho(d(y,t)) + G(F(t)) + G(1 - F(t))$$

y $\Psi(y,t) = \partial\phi(y,t)/\partial t = -[y - F(t)]\nu(t)$ donde $\nu(t)$ está definida en (3.12). Por otra parte, bajo **R1** y **R2**, la función $\Psi(y,\cdot)$ es continua y estrictamente positiva, mientras que

$$\chi(y,t) = F(t)(1 - F(t))\nu(t) - (y - F(t))\nu'(t)$$

está bien definida si además vale la hipótesis **R3**.

Vale la pena mencionar que cuando $\psi(t) > 0$ para todo $t > 0$, la constante c en **R2** puede ser tomada como ∞. Por ejemplo, esto ocurre al elegir la función ρ_c propuesta por Croux y Haesbroeck (2003) y definida en (3.5), pues su derivada ψ_c' es siempre positiva. ♣

Para los resultados de esta sección, también serán necesarias las siguientes hipótesis acerca de la distribución de $\mathbf{X}$.

X1 Para todo $\boldsymbol{\alpha} \in \mathbb{R}^p$, $\boldsymbol{\alpha} \neq \mathbf{0}$, se tiene que $\mathbb{P}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha} = 0) = 0$.

X2 w es una función acotada y no negativa con soporte $\mathcal{C}_w$ tal que $\mathbb{P}(\mathbf{X} \in \mathcal{C}_w) > 0$. Sin pérdida de generalidad, asumimos que $\|w\|_\infty = 1$.

X3 $\mathbb{E}w(\mathbf{X})\|\mathbf{X}\|^2 < \infty$.

X4 La matriz $\mathbf{A} = \mathbb{E}\left\{F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\left[1 - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\nu(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)w(\mathbf{X})\,\mathbf{X}\mathbf{X}^{\mathrm{T}}\right\}$ definida en (3.17) es no singular.

Observación 5.2. Las hipótesis **X1** y **X2** garantizan que los estimadores definidos en (3.9) son Fisher–consistentes. Como se verá en esta sección, este hecho permitirá deducir la consistencia de los estimadores definidos en (3.19). En realidad, para probar la Fisher–consistencia, alcanza con pedir que $\mathbb{P}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \cup w(\mathbf{X}) = 0) < 1$ para todo $\boldsymbol{\alpha} \neq \mathbf{0}$. No obstante, **X1** es necesaria para deducir que el ínfimo de la versión poblacional de la función objetivo no se alcanza en el infinito y de este modo poder concluir la consistencia de nuestros estimadores.

Observemos que **X1** y **X2** implican que la matriz $\mathbb{E}\left[w(\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}\right]$ es definida positiva. Por otro lado, si se cumple **X2** y consideramos la función de pérdida ρ_{DIV} considerada en Basu *et al.* (2017), la matriz $\mathbf{A}$ es no singular, ya que en ese caso $\mathbb{P}(\nu(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) > 0) = 1$. Del mismo modo, **X4** se verifica si $\mathbb{P}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha} = 0) < 1$, para todo $\boldsymbol{\alpha} \neq \mathbf{0}$, ϕ está dada por (3.7) y valen las hipótesis **R2** y **X2**. Por ejemplo, la función de pérdida introducida por Croux and Haesbroeck (2003) verifica **R2**. Más aún, si definimos $\Upsilon(t) = F(t)(1 - F(t))\nu(t)$, es fácil ver que $\mathbf{A}$ es no singular cuando $\mathbb{P}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha} = 0) < 1$ para cualquier $\boldsymbol{\alpha} \neq \mathbf{0}$ y se cumple alguna de las siguientes dos condiciones: (a) La función $\eta \mapsto \mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}w(\mathbf{X})\mathbb{I}_{\Upsilon(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\geq\eta}]$ es continua en η o (b) Existe algún $A > 0$ tal que $\mathbb{P}(\Upsilon(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) > A) = 1$. ♣

El siguiente resultado da un resultado de consistencia para la familia de estimadores definida en (3.19) cuando se considera una función ϕ general.

Teorema 5.1. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19). Supongamos que la función $\mathbb{L}(\boldsymbol{\beta}) = \mathbb{E}[\phi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X})]$ definida en* (3.15) *alcanza su único mínimo en $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ y que $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{c.s.} 0$ cuando $n \to \infty$. Si además*

$$\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2>\epsilon} \mathbb{L}(\boldsymbol{\beta}) > \mathbb{L}(\boldsymbol{\beta}_0)\,, \tag{5.1}$$

para todo $\epsilon > 0$ y se cumple la siguiente Ley Uniforme de los Grandes Números

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \left|\frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X}_i) - \mathbb{E}[\phi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X}_i)]\right| = 0\right) = 1\,, \tag{5.2}$$

entonces, $\widehat{\boldsymbol{\beta}}_n$ es fuertemente consistente para $\boldsymbol{\beta}_0$.

Vale la pena mencionar que, en los Teoremas 5.1 y 5.3, el parámetro λ_n puede ser fijo o aleatorio y en este último caso, el único requerimiento es que $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{c.s.} 0$. En particular, para las penalizaciones LASSO, Signo, Ridge, Bridge, SCAD y MCP descriptas en la Sección 2.3 la condición $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{c.s.} 0$ se verifica si $\lambda_n \xrightarrow{c.s.} 0$.

El siguiente lema da una cota para la dimensión de Vapnik-Chervonenkis (VC) para la familia de funciones $\phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ cuando el vector $\boldsymbol{\beta}$ varía en $\mathbb{R}^p$ y ϕ viene dada por (3.7). Este resultado será usado para probar una Ley Uniforme de los Grandes Números que finalmente garantizará la consistencia de nuestros estimadores.

Lema 5.2. *Sea $\phi : \mathbb{R}^2 \to \mathbb{R}$ la función definida en (3.7) donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R1** y **R2**. Luego, la familia de funciones*

$$\mathcal{F} = \{f_{\boldsymbol{\beta}}(y, \mathbf{x}) = \phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{x}) : \boldsymbol{\beta} \in \mathbb{R}^p\} \tag{5.3}$$

es VC-subgrafo con índice $V(\mathcal{F}) \leq 2p + 4$.

El siguiente teorema se deduce del Teorema 5.1 y del Lema 5.2 y permite obtener la consistencia fuerte de los estimadores definidos en (3.19) cuando se considera como función ϕ la función que controla valores grandes de los residuos *deviance*.

Teorema 5.3. *Sea $\phi : \mathbb{R}^2 \to \mathbb{R}$ definida como en (3.7) donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R1** y **R2**. Si además $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{c.s.} 0$ cuando $n \to \infty$ y se cumplen **X1** y **X2**, entonces el estimador $\widehat{\boldsymbol{\beta}}$ definido en (3.19) es fuertemente consistente para $\boldsymbol{\beta}_0$.*

5.2. Tasa de convergencia

Para obtener tasas de convergencia de los $M-$estimadores pesados penalizados, necesitaremos la siguente hipótesis sobre la función de penalización I_λ. De ahora en adelante, llamaremos $\mathcal{B}(\boldsymbol{\beta}, \epsilon)$ a la bola centrada $\boldsymbol{\beta}$ con radio ϵ respecto de la norma $\|\cdot\|_2$, es decir, $\mathcal{B}(\boldsymbol{\beta}, \epsilon) = \{\mathbf{b} \in \mathbb{R}^p : \|\mathbf{b} - \boldsymbol{\beta}\|_2 \leq \epsilon\}$.

P1 La penalización $I_\lambda(\boldsymbol{\beta})/\lambda$ es Lipschitz respecto de la norma $\|\cdot\|_1$ en un vecindario de $\boldsymbol{\beta}_0$, es decir, existe un $\epsilon > 0$ y una constante K que no depende de λ tal que si $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}(\boldsymbol{\beta}_0, \epsilon)$ entonces

$$|I_\lambda(\boldsymbol{\beta}_1) - I_\lambda(\boldsymbol{\beta}_2)| \leq \lambda K \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1.$$

Observación 5.3. Vale la pena mencionar que las penalizaciones Elastic Net, SCAD y MCP satisfacen **P1**, ya que $\|\boldsymbol{\beta}\|_2 \leq \|\boldsymbol{\beta}\|_1 \leq \sqrt{p}\,\|\boldsymbol{\beta}\|_2$. Por otra parte, la penalización Signo también satisface **P1** si $\|\boldsymbol{\beta}_0\|_2 \neq 0$. Notemos que si $I_\lambda(\boldsymbol{\beta}) = \lambda \sum_{\ell=1}^p J_\ell(|\beta_\ell|)$ donde $J_\ell(\cdot)$ es una función continuamente diferenciable, entonces I_λ satisface **P1**, en particular, este hecho implica que la penalización Bridge satisface **P1** para $q \geq 1$. ♣

Teorema 5.4. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) donde $\phi(y, t)$ viene dada por (3.7) y la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3**. Más aún, supongamos que $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$ y valen las hipótesis **X2**, **X3** y **X4**. Luego,*

(a) Si vale la hipótesis ***P1****, se tiene que* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\lambda_n + 1/\sqrt{n})$*. Por lo tanto si* $\lambda_n = O_{\mathbb{P}}(1/\sqrt{n})$*,* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1/\sqrt{n})$*, mientras que si* $\lambda_n\sqrt{n} \to \infty$*,* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\lambda_n)$*.*

(b) Supongamos que $I_{\lambda_n}(\boldsymbol{\beta}) = \sum_{\ell=1}^{p} J_{\lambda_n}(|\beta_\ell|)$ *donde* $J_{\lambda_n}(\cdot)$ *es no negativa, dos veces diferenciable en* $(0,\infty)$*,* $J'_{\lambda_n}(|\beta_{0,\ell}|) \geq 0$ *y* $J_{\lambda_n}(0) = 0$*. Sea*

$$a_n = \text{máx}\left\{J'_{\lambda_n}(|\beta_{0,\ell}|) : 1 \leq \ell \leq p \;\; y \;\; \beta_{0,\ell} \neq 0\right\} \quad y \quad \alpha_n = \frac{1}{\sqrt{n}} + a_n\,.$$

Por último, supongamos que existe un $\delta > 0$ *tal que*

$$\sup\{|J''_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta)| : \tau \in [-1,1]\,,\; 1 \leq \ell \leq p \;\; y \;\; \beta_{0,\ell} \neq 0\} \xrightarrow{p} 0\,.$$

Luego, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\alpha_n)$*.*

Observación 5.4. El item (a) del Teorema 5.4 muestra que, si la penalidad satisface la hipótesis **P1**, la tasa de convergencia del estimador depende del comportamiento de la velocidad de convergencia de λ_n a 0. En particular, si $\lambda_n\sqrt{n} \to \infty$, el estimador tiene una tasa de convergencia menor que $\sqrt{n}$. Este resultado es análogo al obtenido por Zou (2006) en el caso de regresión lineal para el estimador de mínimos cuadrados con penalidad LASSO. Para esta última penalidad, las tasas dadas en (a) y (b) coinciden pues $J_{\lambda_n}(v) = \lambda_n\, v$ de modo que $a_n = \lambda_n$ y para todo $\beta_{0,\ell} \neq 0$, $\tau \in [-1,1]$, $J''_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta) = 0$ para $\delta > 0$ suficientemente chico.

Por otro lado, las penalizaciones SCAD y MCP cumplen las condiciones del item (b) si $\lambda_n \to 0$. Más aún, cuando esto sucede, existe algún n_0 tal que para $n \geq n_0$, $a_n = 0$ y $\alpha_n = 1/\sqrt{n}$. Por lo tanto, para estas penalidades, la tasa $\sqrt{n}$ puede ser alcanzada asumiendo únicamente $\lambda_n \to 0$. Esta diferencia juega un rol importante con respecto a las propiedades de selección de variables, tal como se verá en la siguiente sección. ♣

5.3. Propiedades de selección de variables

En esta sección, obtendremos resultados acerca de las propiedades de selección de variables de los $M-$estimadores pesados penalizados. Dichas propiedades dependen de la función de penalidad considerada.

Sin pérdida de generalidad, supongamos que $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,I}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ donde $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$, $k \geq 1$, es el subvector de coordenadas activas de $\boldsymbol{\beta}_0$, es decir, el subvector de elementos de $\boldsymbol{\beta}_0$ distintos de cero. De ahora en más, usaremos la notación $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^{\mathrm{T}}, \boldsymbol{\beta}_{II}^{\mathrm{T}})^{\mathrm{T}}$, donde $\boldsymbol{\beta}_I \in \mathbb{R}^k$ y $\boldsymbol{\beta}_{II} \in \mathbb{R}^{p-k}$.

Teorema 5.5. *Sea* $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,I}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n,II}^{\mathrm{T}})^{\mathrm{T}}$ *el estimador definido en (3.19), donde* $\phi(y,t)$ *viene dada por (3.7) y la función* $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *satisface* ***R3****. Supongamos que* $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ *y se cumplen* ***X2*** *y* ***X3****. Supongamos además que, para todo* $C > 0$ *y* $\ell \in \{k+1, \ldots, p\}$*, existe una constante* $K_{C,\ell} \in \mathbb{R}$ *y* $N_{C,\ell} \in \mathbb{N}$ *tal que si* $\|\mathbf{u}\|_2 \leq C$ *y* $n \geq N_{C,\ell}$*, entonces*

$$I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}^{(-\ell)}}{\sqrt{n}}\right) \geq K_{C,\ell}\,\frac{\lambda_n}{\sqrt{n}}\,|u_\ell|\,, \tag{5.4}$$

donde $\mathbf{u}^{(-\ell)}$ *se obtiene reemplazando la* $\ell-$*ésima coordenada de* $\mathbf{u}$ *con cero y* u_ℓ *es la* $\ell-$*ésima coordenada de* $\mathbf{u}$*. Luego,*

(a) Dado $0 < \tau < 1$*, existe* $b > 0$ *y* $n_0 \in \mathbb{N}$ *tal que si* $\lambda_n = b/\sqrt{n}$ *y* $n \geq n_0$*, entonces*

$$\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \geq 1 - \tau\,.$$

(b) Si $\lambda_n \sqrt{n} \to \infty$*, entonces*

$$\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \to 1\,.$$

Para probar propiedades de selección de variables para $M-$estimadores pesados penalizados propuestos, debemos mostrar que la condición (5.4) vale para ciertas penalidades. Observemos en primer lugar, que (5.4) se cumple trivialmente para la penalidad LASSO. En la demostración del Corolario 5.6, probamos que SCAD, MCP y la penalidad Signo también verifican (5.4), lo que permite obtener el siguiente resultado.

Corolario 5.6. *Sea* $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,I}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n,II}^{\mathrm{T}})^{\mathrm{T}}$ *el estimador definido en (3.19), donde* $\phi(y,t)$ *viene dada por (3.7) y la función* $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *satisface* ***R3****. Supongamos que* $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ *y se cumplen* ***X2*** *y* ***X3****.*

(a) Si $I_{\lambda_n}(\boldsymbol{\beta})$ *es la penalización Signo, entonces, para todo* $\tau > 0$ *existe* $b > 0$ *y* $n_0 \in \mathbb{N}$ *tal que si* $\lambda_n = b/\sqrt{n}$ *y* $n \geq n_0$*, se cumple que*

$$\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \geq 1 - \tau\,.$$

(b) Si $I_{\lambda_n}(\boldsymbol{\beta})$ *es la penalización SCAD o MCP y* $\sqrt{n}\lambda_n \to \infty$*, entonces*

$$\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \to 1.$$

Observación 5.5. Una consecuencia del Corolario 5.6 es que las penalizaciones SCAD y MCP tienen la propiedad automática de selección de variables cuando $\sqrt{n}\lambda_n \to \infty$. Por otra parte, al utilizar las penalidades LASSO y Signo, no podemos asegurar selección de variables cuando el estimador tiene tasa $\sqrt{n}$. Recordemos que, para estas últimas dos penalidades, por el Teorema 5.4 el estimador tiene una tasa de convergencia menor que $\sqrt{n}$ cuando $\lambda_n\sqrt{n} \to \infty$. Por esta razón, solo podemos asegurar que prefijado $0 < \tau < 1$, podemos elegir una sucesión de parámetros λ_n de la forma $b/\sqrt{n}$ (de modo a asegurar que el estimador tenga tasa $\sqrt{n}$) y tales que el $M-$estimador penalizado selecciona variables con probabilidad mayor que $1 - \tau$.

Los resultados de la Sección 5.4 nos permitirán concluir que, para las penalidades LASSO y Signo, si el estimador tiene tasa $\sqrt{n}$, entonces $\limsup_n \mathbb{P}(\mathcal{A}_n = \mathcal{A}) < 1$, donde $\mathcal{A} = \{j : \beta_{0,j} \neq 0\} = \{1, \ldots, k\}$ es el conjunto de índices correspondientes a las coordenadas activas de $\boldsymbol{\beta}_0$ y $\mathcal{A}_n = \{j : \widehat{\beta}_{n,j} \neq 0\}$ es el conjunto de índices asociado a las coordenadas no nulas del estimador $\widehat{\boldsymbol{\beta}}_n$. Este resultado es análogo a la Proposición 1 de Zou (2006), que muestra que en esta situación el estimador LASSO puede ser inconsistente para selección de variables en el modelo lineal.

Vale la pena observar que $\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}$ si y solo si $\mathcal{A}_n \subset \mathcal{A}$, por lo tanto, si $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \to 1$ tenemos que $\mathbb{P}(\mathcal{A}_n \subset \mathcal{A}) \to 1$. Observemos que cuando $\mathcal{A}_n \subsetneq \mathcal{A}$ el $M-$estimador penalizado selecciona un submodelo con menos variables explicativas que el original comprimiendo la estimación de algunas componentes activas a 0; sin embargo, la propiedad oráculo de los estimadores basados en las penalizaciones SCAD o MCP dada en el Teorema 5.11 permite deducir que $\mathbb{P}(\mathcal{A}_n = \mathcal{A}) \to 1$. ♣

5.4. Distribución asintótica

En esta sección, obtenemos expresiones para la distribución asintótica de nuestros estimadores. Como la velocidad de convergencia a 0 del parámetro de penalización λ_n requerida para obtener estimadores con tasa $\sqrt{n}$ no es la misma para la penalización Signo que para las penalizaciones SCAD o MCP, consideraremos estos casos separadamente. Si bien la mayoría de los resultados sobre estimadores penalizados suponen que la sucesión de parámetros de penalización es determinística, en esta sección, como en el Teorema 5.4, admitimos que λ_n sea aleatorio, teniendo de esta manera un enfoque más realista.

Observemos que, bajo **X4**, la matriz $\mathbf{A}$ dada en (3.17) es definida positiva, por lo tanto el sub–bloque correspondiente a las coordenadas activas de $\boldsymbol{\beta}_0$ también lo es.

De ahora en más, $\mathbf{e}_\ell$ indicará el $\ell-$ésimo vector canónico y $\text{sign}(z)$ la función signo univariada, es decir, $\text{sign}(z) = z/|z|$ para $z \neq 0$ y $\text{sign}(0) = 0$.

Teorema 5.7. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) donde $\phi(y,t)$ viene dada por (3.7) y la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3**. Supongamos que se cumplen **X2**, **X3**, **X4**, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = O_{\mathbb{P}}(1)$, $\sqrt{n}\lambda_n \xrightarrow{p} b$. Sean $\mathbf{A}$ y $\mathbf{B}$ las matrices definidas en (3.17) y (3.18), respectivamente y consideremos la penalidad Signo dada por*

$$I_\lambda(\boldsymbol{\beta}) = \lambda \, \frac{\|\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}\|_2} \, .$$

Luego, si $\|\boldsymbol{\beta}_0\|_2 \neq 0$, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \text{argmin}_{\mathbf{z}} \, R(\mathbf{z})$, donde el proceso $R : \mathbb{R}^p \to \mathbb{R}$ está dado por

$$R(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + b \, \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z}) \, ,$$

con $\mathbf{w} \sim N_p(\mathbf{0}, \mathbf{B})$, $\mathbf{q}(\mathbf{z}) = \sum_{\ell=1}^{p} \nabla_\ell(\boldsymbol{\beta}_0)\mathbb{I}_{\{\beta_{0,\ell} \neq 0\}} + (\text{sign}(z_\ell)/\|\boldsymbol{\beta}_0\|_2) \; \mathbb{I}_{\{\beta_{0,\ell}=0\}} \, \mathbf{e}_\ell$ y

$$\begin{aligned} \nabla_\ell(\boldsymbol{\beta}) &= \left(-\frac{|\beta_\ell|\beta_1}{\|\boldsymbol{\beta}\|_2^3}, -\frac{|\beta_\ell|\beta_2}{\|\boldsymbol{\beta}\|_2^3}, \ldots, \text{sign}(\beta_\ell)\frac{\|\boldsymbol{\beta}\|_2^2 - \beta_\ell^2}{\|\boldsymbol{\beta}\|_2^3}, \ldots, -\frac{|\beta_\ell|\beta_p}{\|\boldsymbol{\beta}\|_2^3} \right) \\ &= -\frac{|\beta_\ell|}{\|\boldsymbol{\beta}\|_2^3}\,\boldsymbol{\beta} + \frac{\text{sign}(\beta_\ell)}{\|\boldsymbol{\beta}\|_2}\,\mathbf{e}_\ell \, . \end{aligned}$$

El siguiente resultado es análogo al Teorema 5.7 y es aplicable a penalizaciones del tipo Ridge, LASSO y cualquier combinación convexa de ellas, en particular Elastic Net y Bridge con $q > 1$.

Teorema 5.8. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) con $\phi(y,t)$ dada por (3.7) donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3** y sean $\mathbf{A}$ y $\mathbf{B}$ las matrices definidas en (3.17) y (3.18), respectivamente. Supongamos que, para algún $0 \leq \alpha \leq 1$, la función de penalización puede escribirse como*

$$I_\lambda(\boldsymbol{\beta}) = \lambda \left\{ (1-\alpha) \sum_{\ell=1}^{p} J_\ell(|\beta_\ell|) + \alpha \sum_{\ell=1}^{p} |\beta_\ell| \right\} , \tag{5.5}$$

*donde $J_\ell(\cdot)$ es una función continuamente diferenciable tal que $J'_\ell(0) = 0$. Supongamos además que $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_{\mathbb{P}}(1)$, $\sqrt{n}\lambda_n \xrightarrow{p} b$ y valen la hipótesis **X2**, **X3**, **X4**. Luego, si $\|\boldsymbol{\beta}_0\|_2 \neq 0$, entonces $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \text{argmin}_{\mathbf{z}} \, R(\mathbf{z})$, donde $R : \mathbb{R}^p \to \mathbb{R}$ está definido por*

$$R(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\,\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + b \, \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z}) \, ,$$

con $\mathbf{w} \sim N_p(\mathbf{0}, \mathbf{B})$, $\mathbf{q}(\mathbf{z}) = (q_1(\mathbf{z}), \ldots, q_p(\mathbf{z}))^{\mathrm{T}}$ *siendo*

$$q_\ell(\mathbf{z}) \;=\; (1-\alpha) J'_\ell(|\beta_{0,\ell}|)\,\mathrm{sign}(\beta_{0,\ell}) + \alpha\left\{\mathrm{sign}(\beta_{0,\ell})\mathbb{I}_{\{\beta_{0,\ell}\neq 0\}} + \mathrm{sign}(z_\ell)\mathbb{I}_{\{\beta_{0,\ell}=0\}}\right\}\,.$$

Observación 5.6. Notemos que cuando $\sqrt{n}\lambda_n \xrightarrow{p} 0$ $(b = 0)$, los $M-$estimadores penalizados basados en la penalidad Signo o en la penalidad dada en (5.5) tienen la misma distribución asintótica que los $M-$estimadores definidos en (3.9). Por otra parte, si $b > 0$ y $\alpha > 0$ en (5.5), argumentos análogos a los considerados para el modelo lineal en Knight y Fu (2000) permiten probar que la distribución asintótica de las coordenadas de $\widehat{\boldsymbol{\beta}}_n$ que corresponden a covariables no activas, es decir, la distribución asintótica de $\sqrt{n}\,\widehat{\boldsymbol{\beta}}_{n,II}$, le asigna probabilidad positiva al cero. Finalmente, si $\alpha = 0$ y $b > 0$, el achicamiento en la estimación de los coeficientes de la regresión aumenta con la magnitud de $\boldsymbol{\beta}_0$. Luego, para parámetros "grandes", el sesgo introducido por la penalidad diferenciable $J_\ell(\cdot)$ puede ser grande. ♣

Argumentos análogos a los utilizados en la demostración de la Proposición 1 de Zou (2006), junto con los Teoremas 5.7 y 5.8 permiten mostrar que, para las penalidades LASSO o Signo, si el $M-$estimador penalizado tiene tasa $\sqrt{n}$, entonces resulta inconsistente para selección de variables, como establece el siguiente resultado. Más aún, de la demostración se deduce que si $\sqrt{n}\lambda_n \xrightarrow{p} 0$, entonces $\mathbb{P}(\mathcal{A}_n = \mathcal{A}) \to 0$, es decir, necesitamos parámetros de regularización que converjan a 0 pero no demasiado rápido, para poder seleccionar variables con probabilidad no nula.

Corolario 5.9. *Sea* $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,I}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n,II}^{\mathrm{T}})^{\mathrm{T}}$ *el estimador definido en (3.19), donde* $\phi(y,t)$ *viene dada por (3.7) y la función* $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *satisface* ***R3****. Supongamos que* $\|\boldsymbol{\beta}_0\| \neq 0$, $\sqrt{n}\lambda_n \xrightarrow{p} b$, $\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ *y se cumplen* ***X2****,* ***X3****,* ***X4****. Entonces, para las penalidades Signo o LASSO, existe* $c < 1$ *tal que* $\limsup_n \mathbb{P}(\mathcal{A}_n = \mathcal{A}) \leq c < 1$*, donde* $\mathcal{A} = \{j : \beta_{0,j} \neq 0\}$ *es el conjunto de índices correspondientes a las coordenadas activas de* $\boldsymbol{\beta}_0$ *y* $\mathcal{A}_n = \{j : \widehat{\beta}_{n,j} \neq 0\}$.

Argumentos análogos a los utilizados en la demostración del Teorema 5.7, permiten obtener el comportamiento asintótica del $M-$estimador con penalización Signo cuando $\sqrt{n}\lambda_n \to \infty$. Un resultado análogo vale para penalizaciones que cumplan (5.5), como la penalización LASSO.

Teorema 5.10. *Sea* $\widehat{\boldsymbol{\beta}}_n$ *el estimador definido en (3.19) donde* $\phi(y,t)$ *viene dada por (3.7) y la función* $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *satisface* ***R3****. Supongamos que se cumplen* ***X2****,* ***X3****,* ***X4****,* $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = O_{\mathbb{P}}(\lambda_n)$, $\sqrt{n}\lambda_n \to \infty$*. Sea* $\mathbf{A}$ *la matriz definida en (3.17) y consideremos la penalidad Signo dada por*

$$I_\lambda(\boldsymbol{\beta}) = \lambda\,\frac{\|\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}\|_2}\,.$$

Luego, si $\|\boldsymbol{\beta}_0\|_2 \neq 0$, $(1/\lambda_n)\,(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{p} \mathrm{argmin}_{\mathbf{z}}\, R(\mathbf{z})$*, donde la función* $R : \mathbb{R}^p \to \mathbb{R}$ *está definida por*

$$R(\mathbf{z}) = \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})\,,$$

siendo $\mathbf{q}(\mathbf{z})$ *la función definida en el Teorema 5.7.*

Observación 5.7. El Lema 3 de Zou (2006) da un resultado análogo al Teorema 5.10 para el estimador de mínimos cuadrados con penalización LASSO, para el caso del modelo de regresión

lineal. Como en ese resultado, la tasa de convergencia de $\widehat{\boldsymbol{\beta}}_n$ es menor que $\sqrt{n}$ y el límite es una cantidad no aleatoria. Como en Zou (2006), la tasa óptima para $\widehat{\boldsymbol{\beta}}_n$ se logra cuando $\lambda_n = O_{\mathbb{P}}(1/\sqrt{n})$ pero a costa de no seleccionar variables. ♣

Finalmente, el siguiente teorema muestra la normalidad asintótica de $\widehat{\boldsymbol{\beta}}_{n,I}$ cuando la penalidad es consistente para seleccionar variables, es decir, cuando $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \to 1$. Para ello, recordemos que $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,I}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ donde $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$ es el subvector de coordenadas activas de $\boldsymbol{\beta}_0$ y definamos para $\mathbf{b} \in \mathbb{R}^k$,

$$\nabla I_\lambda(\mathbf{b}) = \frac{\partial I_\lambda\left((\mathbf{b}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}\right)}{\partial \mathbf{b}}\,.$$

Teorema 5.11. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) con $\phi(y,t)$ dada por (3.7) donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3**. Supongamos que valen las hipótesis **X2**, **X3** y que existe $\delta > 0$ tal que*

$$\sup_{\mathbf{b} \in \mathbb{R}^k : \|\mathbf{b} - \boldsymbol{\beta}_{0,I}\|_2 \leq \delta} \|\nabla I_{\lambda_n}(\mathbf{b})\|_2 = o\left(\frac{1}{\sqrt{n}}\right), \tag{5.6}$$

$\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \to 1$ y $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$. Sean $\mathbf{A}$ y $\mathbf{B}$ las matrices definidas en (3.17) y (3.18), respectivamente. Indiquemos por $\widetilde{\mathbf{A}} \in \mathbb{R}^{k\times k}$ y $\widetilde{\mathbf{B}} \in \mathbb{R}^{k\times k}$ a las submatrices de $\mathbf{A}$ y $\mathbf{B}$, respectivamente, asociadas a las primeras k covariables. Luego, si $\widetilde{\mathbf{A}}$ es inversible,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{D} N(\mathbf{0}_k, \widetilde{\mathbf{A}}^{-1}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{-1}).$$

Observación 5.8. Es fácil ver que tanto SCAD como MCP satisfacen (5.6) cuando $\lambda_n \to 0$. Este hecho, junto con el Corolario 5.6, implican que el $M-$estimador pesado penalizado definido en (3.19) con penalidades SCAD o MCP tienen la propiedad oráculo cuando $\lambda_n \to 0$ y $\sqrt{n}\,\lambda_n \to \infty$. ♣

5.5. Apéndice A: Demostraciones de la Sección 5.1

Demostración del Teorema 5.1. Como $\widehat{\boldsymbol{\beta}}_n$ minimiza $L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta})$, tenemos que

$$L_n(\widehat{\boldsymbol{\beta}}_n) \leq L_n(\widehat{\boldsymbol{\beta}}_n) + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) \leq L_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}_0)\,.$$

Por lo tanto,

$$\limsup_{n\to\infty} L_n(\widehat{\boldsymbol{\beta}}_n) \leq \limsup_{n\to\infty} L_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}_0).$$

Usando la Ley de los Grandes Números y que $I_{\lambda_n}(\boldsymbol{\beta}_0) \xrightarrow{c.s.} 0$ cuando $n \to \infty$, tenemos que, con probabilidad uno,

$$\limsup_{n\to\infty} L_n(\widehat{\boldsymbol{\beta}}_n) \leq \mathbb{L}(\boldsymbol{\beta}_0)\,. \tag{5.7}$$

Sea $\epsilon > 0$. Usando (5.2) tenemos que, con probabilidad uno,

$$\lim_{n\to\infty} \sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > \epsilon} |L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta})| = 0\,. \tag{5.8}$$

Notemos que $L_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}) + \mathbb{L}(\boldsymbol{\beta}) \geq -|L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta})| + \mathbb{L}(\boldsymbol{\beta})$. Luego,

$$\inf_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > \epsilon} L_n(\boldsymbol{\beta}) \geq -\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > \epsilon} |L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta})| + \inf_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > \epsilon} \mathbb{L}(\boldsymbol{\beta})\,.$$

Por lo tanto, con probabilidad uno,

$$\liminf_{n\to\infty} \inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2>\epsilon} L_n(\boldsymbol{\beta}) \geq \inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2>\epsilon} \mathbb{L}(\boldsymbol{\beta}) > \mathbb{L}(\boldsymbol{\beta}_0)\,, \tag{5.9}$$

donde la primera desigualdad es consecuencia de (5.8) y la segunda sigue de (5.1). Entonces, utilizando (5.9) y (5.7) obtenemos que con probabilidad uno existe $n_0 \in \mathbb{N}$ tal que $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq \epsilon$ para todo $n \geq n_0$, lo que concluye la demostración. ■

Demostración del Lema 5.2. Teniendo en cuenta que multiplicar por una función fija preserva el índice de una clase, basta probar el resultado cuando $w(\mathbf{x}) \equiv 1$. Supongamos que esta clase de funciones no es VC-subgrafo o que el índice VC de los subgrafos de las funciones de $\mathcal{F}$ es mayor a $2p+4$. Luego, existe un conjunto $\mathcal{C}_0 = \{(y_i, \mathbf{x}_i, r_i)\,,\ i = 1, \ldots, 2p+5\}$, $r_i \in \mathbb{R}$ que puede ser "desmenuzado" por los subgrafos de las funciones en $\mathcal{F}$. Como hay solamente dos valores posibles para y, es posible tomar un subconjunto $\mathcal{C} \subset \mathcal{C}_0$ tal que $|\mathcal{C}| = \ell = p+3$ y el valor correspondiente de y para todos los elementos de $\mathcal{C}$ es el mismo.

Sin pérdida de generalidad, supongamos que dicho valor común de y es cero y que los índices del conjunto $\mathcal{C}$ son $i = 1, \ldots, \ell$. Sea $\phi_0(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \phi(0, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ y $\phi_1(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \phi(1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. Luego, cuando $\phi(y,t)$ viene dada por (3.7) y valen las condiciones **R1** y **R2**, $\phi_0(s)$ es estrictamente creciente, mientras que $\phi_1(s)$ es estrictamente decreciente.

Supongamos que $\{(\mathbf{x}_1, r_1), \ldots, (\mathbf{x}_\ell, r_\ell)\}$ son la segunda y tercera coordenada de los ℓ elementos de $\mathcal{C}$. Entonces, para cada subconjunto $\mathcal{I} \subset \{1, \ldots, \ell\}$ existe un $\boldsymbol{\beta}_{\mathcal{I}}$ tal que $\phi_0(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{I}}) \geq r_i$ si y solo si $i \in \mathcal{I}$. Esto implica que $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{I}} \geq \phi_0^{-1}(r_i)$ si y solo si $i \in \mathcal{I}$.

Sea $\widetilde{\mathbf{x}}_i = (\mathbf{x}_i, \phi_0^{-1}(r_i))$, $1 \leq i \leq \ell$, y $\widetilde{\boldsymbol{\beta}}_{\mathcal{I}} = (\boldsymbol{\beta}_{\mathcal{I}}, -1)$. Luego, $\widetilde{\mathbf{x}}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_{\mathcal{I}} \geq 0$ si y solo si $i \in \mathcal{I}$. Observemos que esta última equivalencia implica que la familia de semiespacios de dimensión $p+1$ puede desmenuzar a un conjunto de $\ell = (p+3)$ elementos, lo cual es un absurdo (ver Ejemplo 3.7.4.c de van de Geer, 2000 o Pollard, 1984). Por lo tanto, concluimos que $\mathcal{F}$ es una familia VC-subgrafo y su índice VC cumple $V(\mathcal{F}) \leq 2p+4$. ■

El siguiente resultado corresponde al Lema 6.3 de Bianco y Yohai (1996) y juega un papel fundamental en la prueba de la consistencia de nuestro estimador. En dicho trabajo puede verse su demostración para el caso no pesado. La prueba para el caso general es completamente análoga usando la condición **X2**.

Lema 5.12. *Sea $\phi : \mathbb{R}^2 \to \mathbb{R}$ dada por (3.7), donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ cumple las condiciones **R1** y **R2**. Luego, si valen **X1** y **X2**, para todo $\|\mathbf{u}\|_2 = 1$ existe un $\epsilon_{\mathbf{u}}$ tal que*

$$\mathbb{E}\left(\liminf_{a\to\infty} \inf_{\mathbf{v}\in\mathcal{V}(\mathbf{u},\epsilon_{\mathbf{u}})} \phi(Y, a\,\mathbf{X}^{\mathrm{T}}\mathbf{v})\, w(\mathbf{X})\right) > \mathbb{L}(\boldsymbol{\beta}_0)\,,$$

donde $\mathcal{V}(\mathbf{u}, \epsilon) = \{\mathbf{v} : \|\mathbf{v} - \mathbf{u}\|_2 < \epsilon\}$.

Demostración del Teorema 5.3. Bastará ver que se satisfacen las condiciones del Teorema 5.1. El Teorema 3.1 implica que $\mathbb{L}(\boldsymbol{\beta}) = \mathbb{E}[\phi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{X})]$ tiene un único mínimo en $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Por otra parte, la convergencia uniforme requerida en (5.2) es consecuencia del Teorema 4.3, el Lema 5.2 y del hecho de que $|\phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})w(\mathbf{x})|$ es uniformemente acotada.

Resta probar la condición (5.1). Supongamos que esto es falso, es decir, supongamos que

$$\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2>\epsilon} \mathbb{L}(\boldsymbol{\beta}) \leq \mathbb{L}(\boldsymbol{\beta}_0)\,. \tag{5.10}$$

Sea $(\boldsymbol{\beta}_n)_{n\geq 1}$ una sucesión tal que $\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\|_2 > \epsilon$ y

$$\lim_{n\to\infty} \mathbb{L}(\boldsymbol{\beta}_n) = \inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2>\epsilon} \mathbb{L}(\boldsymbol{\beta})\,.$$

Supongamos primero que $(\boldsymbol{\beta}_n)_{n\geq 1}$ es acotada. Luego, existe una subsucesión $(\boldsymbol{\beta}_{n_j})_{j\geq 1}$ de $(\boldsymbol{\beta}_n)_{n\geq 1}$ que converge a un valor $\boldsymbol{\beta}^\star$. La continuidad de $\mathbb{L}(\boldsymbol{\beta})$ permite concluir que

$$\lim_{j\to\infty} \mathbb{L}(\boldsymbol{\beta}_{n_j}) = \mathbb{L}(\boldsymbol{\beta}^\star) > \mathbb{L}(\boldsymbol{\beta}_0)\,,$$

donde la última desigualdad vale por el Teorema 3.1, contradiciendo (5.10).

Por lo tanto, $\limsup_{n\to\infty} \|\boldsymbol{\beta}_n\|_2 = \infty$ y $\lim_{n\to\infty} \mathbb{L}(\boldsymbol{\beta}_n) = \inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2>\epsilon} \mathbb{L}(\boldsymbol{\beta}) \leq \mathbb{L}(\boldsymbol{\beta}_0)$.

Definamos $\boldsymbol{\beta}_n^\star = \boldsymbol{\beta}_n/\|\boldsymbol{\beta}_n\|_2$ y supongamos, eventualmente tomando una subsucesión, que $\lim_{n\to\infty} \boldsymbol{\beta}_n^\star = \boldsymbol{\beta}^\star$ donde $\|\boldsymbol{\beta}^\star\|_2 = 1$.

Recordemos que hemos denotado $\mathcal{V}(\mathbf{u},\epsilon) = \{\mathbf{v} : \|\mathbf{v}-\mathbf{u}\|_2 < \epsilon\}$. El Lema 5.12 implica que existe un $\epsilon^\star = \epsilon_{\boldsymbol{\beta}^\star}$ tal que

$$\mathbb{E} \liminf_{a\to\infty} \inf_{\mathbf{v}\in\mathcal{V}(\boldsymbol{\beta}^\star,\epsilon^\star)} \phi(Y, a\mathbf{X}^{\mathrm{T}}\mathbf{v})\, w(\mathbf{X}) > \mathbb{L}(\boldsymbol{\beta}_0)\,. \tag{5.11}$$

Sea $n_0 \in \mathbb{N}$ tal que $\boldsymbol{\beta}_n^\star \in \mathcal{V}(\boldsymbol{\beta}^\star,\epsilon)$ y $\|\boldsymbol{\beta}_n\|_2 > M$ para $n \geq n_0$. Luego,

$$\phi(Y, \|\boldsymbol{\beta}_n\|_2\, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_n^\star)\, w(\mathbf{X}) \geq \inf_{a>M} \inf_{\mathbf{v}\in\mathcal{V}(\boldsymbol{\beta}^\star,\epsilon)} \phi(Y, a\, \mathbf{X}^{\mathrm{T}}\mathbf{v})\, w(\mathbf{X})$$

de donde, usando el Lema de Fatou, obtenemos que

$$\lim_{n\to\infty} \mathbb{L}(\boldsymbol{\beta}_n) \;=\; \lim_{n\to\infty} \mathbb{L}(\|\boldsymbol{\beta}_n\|_2\, \boldsymbol{\beta}_n^\star) \geq \mathbb{E} \liminf_{a\to+\infty} \inf_{\mathbf{v}\in\mathcal{V}(\boldsymbol{\beta}^\star,\epsilon)} \phi(Y, a\mathbf{X}^{\mathrm{T}}\mathbf{v})\, w(\mathbf{X})\,.$$

Entonces, usando (5.11), deducimos que $\lim_{n\to\infty} \mathbb{L}(\boldsymbol{\beta}_n) > \mathbb{L}(\boldsymbol{\beta}_0)$ llegando nuevamente a un absurdo, lo que implica que se cumple (5.1). ■

5.6. Apéndice B: Demostraciones de la Sección 5.2

Para probar el Teorema 5.4, será necesario el siguiente Lema.

Lema 5.13. *Sea $\widetilde{\boldsymbol{\beta}}_n$ tal que $\widetilde{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_0$ y $\phi(y,t)$ es la función dada por (3.7) donde $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface la condición **R3**. Luego, si se cumple **X3**, se tiene que $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n) \xrightarrow{a.s.} \mathbf{A}$ donde $\mathbf{A}$ es la matriz dada en (3.17) y*

$$\mathbf{A}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\, w(\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\,. \tag{5.12}$$

Demostración. Alcanza con probar que $\|\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n) - \mathbf{A}_n(\boldsymbol{\beta}_0)\| \xrightarrow{a.s.} 0$, ya que $\mathbf{A}_n(\boldsymbol{\beta}_0) \xrightarrow{a.s.} \mathbf{A}$. Más aún, es suficiente demostrar que para todo $1 \leq k, j \leq p$,

$$A_{n,kj} = \frac{1}{n}\sum_{i=1}^{n} |\chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_n) - \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)|\, w(\mathbf{X}_i)\, |X_{ik}|\, |X_{ij}| \xrightarrow{a.s.} 0\,. \tag{5.13}$$

Sea $\epsilon > 0$. Como ϕ está dada por (3.7), la condición **R3** implica que la función χ está acotada. Tomemos $M > 0$ tal que

$$\mathbb{E}[\mathbb{I}_{\{\|\mathbf{X}\|>M\}} w(\mathbf{X})\|\mathbf{X}\|^2] < \frac{\epsilon}{4\|\chi\|_\infty} \tag{5.14}$$

y observemos que

$$\begin{aligned} A_{n,kj} &\leq \frac{1}{n}\sum_{i=1}^{n} |\chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_n) - \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)|\, w(\mathbf{X}_i)\, \|\mathbf{X}_i\|^2 \left(\mathbb{I}_{\{\|\mathbf{X}_i\|>M\}} + \mathbb{I}_{\{\|\mathbf{X}_i\|\leq M\}}\right) \\ &\leq 2\,\frac{1}{n}\sum_{i=1}^{n} \|\chi\|_\infty\, w(\mathbf{X}_i)\|\mathbf{X}_i\|^2 \mathbb{I}_{\{\|\mathbf{X}_i\|>M\}} + \frac{1}{n}\sum_{i=1}^{n} |\chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_n) - \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)|\, w(\mathbf{X}_i)\, \|\mathbf{X}_i\|^2 \mathbb{I}_{\{\|\mathbf{X}_i\|\leq M\}}\,. \end{aligned}$$

La función $\chi(y,t)$ es uniformemente continua en t si restringimos esta variable a un compacto. Como solamente hay dos valores posibles para y, es posible elegir $\delta > 0$ tal que si $|s| \leq M(\|\boldsymbol{\beta}_0\| + 1)$ y $|s'| \leq M(\|\boldsymbol{\beta}_0\| + 1)$ cumplen $|s - s'| < \delta$, entonces $|\chi(y,s) - \chi(y,s')| < \epsilon/(2M^2)$.

Usando que $\widetilde{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_0$ y la Ley Fuerte de los Grandes Números, tenemos que existe un conjunto $\mathcal{N} \subset \Omega$ de probabilidad cero tal que para todo $\omega \notin \mathcal{N}$, $\widetilde{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ y

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i\|^2 w(\mathbf{X}_i)\mathbb{I}_{\{\|\mathbf{X}_i\|>M\}} \to \mathbb{E}[\mathbb{I}_{\{\|\mathbf{X}\|>M\}} w(\mathbf{X})\|\mathbf{X}\|^2].$$

Sea $n_1 = n_1(\omega)$ tal que, para $n \geq n_1$, $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \leq \text{mín}\{1, \delta/M\}$ y

$$\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{X}_i)\|\mathbf{X}_i\|^2 \mathbb{I}_{\{\|\mathbf{X}_i\|>M\}} < \frac{\epsilon}{4\|\chi\|_\infty}\,.$$

Supongamos que $\|\mathbf{X}_i\| \leq M$ y $n \geq n_1$. Luego, usando que $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \leq 1$ es fácil ver que tanto $\mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_n$ como $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ tienen un valor absoluto no mayor a $M(\|\boldsymbol{\beta}_0\| + 1)$. Más aún, como $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \leq \delta/M$, también tenemos que $|\mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_n - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0| \leq \delta$, por lo tanto

$$|\chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_n) - \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)| \leq \frac{\epsilon}{2M^2}\,.$$

Finalmente, tomando en cuenta que

$$\frac{1}{n}\sum_{i=1}^{n} |\chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_n) - \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)|\, w(\mathbf{X}_i)|X_{ik}\|X_{ij}| \leq 2\|\chi\|_\infty \frac{\epsilon}{4\|\chi\|_\infty} + \frac{1}{n}\sum_{i=1}^{n} \frac{\epsilon}{2M^2} M^2 = \epsilon,$$

concluimos la demostración de (5.13). ■

El siguiente resultado es una extensión del Lema 5.13 y puede ser probado usando argumentos similares a los usados en el Lema 1 de Bianco y Boente (2002). Notemos que una consecuencia directa del Lema 5.14 es que $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n) \xrightarrow{p} \mathbf{A}$ siempre que $\widetilde{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$.

Lema 5.14. *Sea $\phi(y,t)$ la función dada por (3.7) donde $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3**. Luego, si se cumple la condición **X3**, para todo $\delta > 0$ se tiene que*

a) $\lim_{\boldsymbol{\beta}\to\boldsymbol{\beta}_0} \mathbb{E}\chi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\, w(\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{A}$,

b) $\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|<\delta} |\mathbf{A}_n(\boldsymbol{\beta}) - \mathbb{E}\chi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\, w(\mathbf{X})\mathbf{X}\mathbf{X}^{\mathrm{T}}| \xrightarrow{p} 0$

donde $\mathbf{A}$ *y* $\mathbf{A}_n(\boldsymbol{\beta})$ *están definidas en (3.17) y (5.12) respectivamente.*

Demostración del Teorema 5.4. Sea $W_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta})$ donde L_n está definida en (3.10). Usando un desarrollo de Taylor de orden 2 en $L_n(\widehat{\boldsymbol{\beta}}_n)$ alrededor de $\boldsymbol{\beta}_0$, tenemos que

$$W_n(\widehat{\boldsymbol{\beta}}_n) = L_n(\boldsymbol{\beta}_0) + (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}} \nabla L_n(\boldsymbol{\beta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}} \mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n),$$

donde $\widetilde{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \tau_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ es un punto intermedio y con $\tau_n \in [0,1]$, $\nabla L_n(\boldsymbol{\beta})$ es el gradiente de $L_n(\boldsymbol{\beta})$ dado por

$$\nabla L_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\, w(\mathbf{X}_i)\mathbf{X}_i$$

y $\mathbf{A}_n(\boldsymbol{\beta})$ es el Hessiano de la función $L_n(\boldsymbol{\beta})$, es decir,

$$\mathbf{A}_n(\boldsymbol{\beta}) = \frac{\partial^2}{(\partial\boldsymbol{\beta})^2} L_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\, w(\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\,.$$

Sea ε una constante positiva y sea ζ_1 el menor autovalor de la matriz $\mathbf{A}$. En virtud de la hipótesis **X4**, sabemos que $\zeta_1 > 0$. Como $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, del Lema 5.14 deducimos que $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n) \xrightarrow{p} \mathbf{A}$, por lo tanto existe n_1 tal que si $n \geq n_1$, $\mathbb{P}(\mathcal{B}_n) > 1 - \varepsilon/4$, donde $\mathcal{B}_n = \Big\{\|\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n) - \mathbf{A}\| < \zeta_1/2\Big\}$.

Por otro lado, el Teorema Central del Límite junto con (3.13) implica que

$$\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0) = O_{\mathbb{P}}(1)\,. \tag{5.15}$$

Luego, existe una constante M_1 para la cual $\mathbb{P}(\mathcal{C}_n) > 1 - \varepsilon/4$, donde $\mathcal{C}_n = \{\|\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)\|_2 < M_1\}$. Usando la definición de $\widehat{\boldsymbol{\beta}}_n$ se tiene que en $\mathcal{B}_n \cap \mathcal{C}_n$,

$$\begin{aligned}
0 &\geq W_n(\widehat{\boldsymbol{\beta}}_n) - W_n(\boldsymbol{\beta}_0) \\
&= (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}} \nabla L_n(\boldsymbol{\beta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}} \mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) - I_{\lambda_n}(\boldsymbol{\beta}_0) \\
&\geq -\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \frac{1}{\sqrt{n}} \|\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)\|_2 - \|\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_n) - \mathbf{A}\| \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 + \zeta_1 \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) - I_{\lambda_n}(\boldsymbol{\beta}_0) \\
&\geq -\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \frac{1}{\sqrt{n}} M_1 + \frac{\zeta_1}{2} \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) - I_{\lambda_n}(\boldsymbol{\beta}_0)\,.
\end{aligned} \tag{5.16}$$

Para probar (a), definimos el evento $\mathcal{D}_n = \{I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) - I_{\lambda_n}(\boldsymbol{\beta}_0) \leq K\,\lambda_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2\}$. Observemos que **P1** y el hecho de que $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$ implican que existe una constante K y $n_2 \in \mathbb{N}$ tal que si $n \geq n_2$, $\mathbb{P}(\mathcal{D}_n) \geq 1 - \varepsilon/2$. Luego, si $n \geq \max\{n_1, n_2\}$ tenemos que $\mathbb{P}(\mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n) > 1 - \epsilon$. Además, en $\mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{D}_n$ se cumple que

$$0 \geq -\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \frac{1}{\sqrt{n}} M_1 + \frac{\zeta_1}{2} \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 - K\,\lambda_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2,$$

lo que implica que

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq 2\left(\lambda_n + \frac{1}{\sqrt{n}}\right)\frac{M_1 + K}{\zeta_1}\,.$$

Por lo tanto, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\lambda_n + 1/\sqrt{n})$, completando la demostración.

Probaremos ahora el item (b). Supongamos, sin pérdida de generalidad, que $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,I}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ donde $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$ es el subvector de coordenadas activas de $\boldsymbol{\beta}_0$ (en particular, esto quiere decir que

las primeras k coordenadas de $\boldsymbol{\beta}_0$ son distintas de cero). Como $J_{\lambda_n}(0) = 0$ y $J_{\lambda_n}(s) \geq 0$, tenemos que

$$I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) = \sum_{\ell=1}^{k} J_{\lambda_n}(|\beta_\ell|) - J_{\lambda_n}(|\beta_{0,\ell}|) + \sum_{\ell=l+1}^{p} J_{\lambda_n}(|\beta_\ell|) \geq \sum_{\ell=1}^{k} J_{\lambda_n}(|\beta_\ell|) - J_{\lambda_n}(|\beta_{0,\ell}|)\,.$$

Por lo tanto, (5.16) implica

$$0 \geq -\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \frac{1}{\sqrt{n}} M_1 + \frac{\zeta_1}{2}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 + \sum_{\ell=1}^{k} J_{\lambda_n}(|\widehat{\beta}_{n,\ell}|) - J_{\lambda_n}(|\beta_{0,\ell}|)\,.$$

Sea $\delta > 0$ tal que $\sup\{|J'_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta)| : \tau \in [-1,1]\ ,\ 1 \leq \ell \leq p \ \text{ y } \ \beta_{0,\ell} \neq 0\} \xrightarrow{p} 0$. Luego, existe $n_3 \in \mathbb{N}$ tal que para todo $n \geq n_3$, $\mathbb{P}(\mathcal{E}_{n,1}) > 1 - \varepsilon/4$ donde

$$\mathcal{E}_{n,1} = \left\{\sup\{|J'_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta)| : \tau \in [-1,1]\ ,\ 1 \leq \ell \leq p \ \text{ y } \ \beta_{0,\ell} \neq 0\} \leq \frac{\zeta_1}{2}\right\}$$

y $\mathbb{P}(\mathcal{E}_{n,2}) \geq 1 - \varepsilon/4$ donde $\mathcal{E}_{n,2} = \{\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq \delta\}$. Sea $\mathcal{E}_n = \mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$. Haciendo un desarrollo de Taylor de primer orden, tenemos que, para $\ell = 1, \ldots, k$,

$$J_{\lambda_n}(|\widehat{\beta}_{n,\ell}|) - J_{\lambda_n}(|\beta_{0,\ell}|) = J'_{\lambda_n}(|\beta_{0,\ell}|)(|\widehat{\beta}_{n,\ell}| - |\beta_{0,\ell}|) + \frac{1}{2}J''_{\lambda_n}(\xi_{n,\ell})(|\widehat{\beta}_{n,\ell}| - |\beta_{0,\ell}|)^2\,,$$

donde $\xi_{n,\ell}$ pertenece al intervalo comprendido entre $|\widehat{\beta}_{n,\ell}|$ y $|\beta_{0,\ell}|$. Usando que $|\,|a| - |b|\,| \leq |a-b|$, $J'_{\lambda_n}(|\beta_{0,\ell}|) \geq 0$ y que en el evento $\mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{E}_n$, $|J''_{\lambda_n}(\xi_{n,\ell})| < \zeta_1/2$, pues $\text{máx}(0, |\beta_{0,\ell}| - \delta) < \xi_{n,\ell} \leq |\beta_{0,\ell}| + \delta$, tenemos que

$$\begin{aligned}
I_{\lambda_n}(\widehat{\beta}_n) - I_{\lambda_n}(\beta_0) &\geq \sum_{\ell=1}^{k} J_{\lambda_n}(|\widehat{\beta}_{n,\ell}|) - J_{\lambda_n}(|\beta_{0,\ell}|) \\
&\geq -\sum_{\ell=1}^{k} J'_{\lambda_n}(|\beta_{0,\ell}|)|\widehat{\beta}_{n,\ell} - \beta_{0,\ell}| - \frac{1}{2}\sum_{\ell=1}^{k} |J''_{\lambda_n}(\xi_{n,\ell})|(\widehat{\beta}_{n,\ell} - \beta_{0,\ell})^2 \\
&\geq -a_n \sum_{\ell=1}^{k} |\widehat{\beta}_{n,\ell} - \beta_{0,\ell}| - \frac{\zeta_1}{4}\sum_{\ell=1}^{k}(\widehat{\beta}_{n,\ell} - \beta_{0,\ell})^2 \\
&\geq -a_n\sqrt{k}\ \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 - \frac{\zeta_1}{4}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2\,.
\end{aligned}$$

Por lo tanto,

$$0 \geq -\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \frac{1}{\sqrt{n}} M_1 + \frac{\zeta_1}{2}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 - a_n\sqrt{k}\ \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 - \frac{\zeta_1}{4}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2\,,$$

lo que implica que $4\alpha_n(M_1 + \sqrt{k})/\zeta_1 \geq \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2$. Además, $\mathbb{P}(\mathcal{B}_n \cap \mathcal{C}_n \cap \mathcal{E}_n) \geq 1 - \varepsilon$, para $n \geq \text{máx}_{1 \leq i \leq 3}\, n_i$, concluyendo la demostración. ■

5.7. Apéndice C: Demostraciones de la Sección 5.3

Demostración del Teorema 5.5. Consideramos la descomposición $\boldsymbol{\beta}_0 = (\widetilde{\boldsymbol{\beta}}_0^{\mathrm{T}}, 0)^{\mathrm{T}}$ donde $\widetilde{\boldsymbol{\beta}}_0 \in \mathbb{R}^{p-1}$ y definimos

$$V_n(\mathbf{u}_1, u_2) = L_n\left(\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}, \frac{u_2}{\sqrt{n}}\right) + I_{\lambda_n}\left(\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}, \frac{u_2}{\sqrt{n}}\right),$$

con $L_n(\boldsymbol{\beta})$ definido como en (3.10).

Fijemos $\tau > 0$ y definamos $\tau^* = \tau/(2(p-k))$. Sea $C > 0$ tal que $\mathbb{P}(\mathcal{B}_n) \geq 1 - \tau^*$ donde $\mathcal{B}_n = \{\sqrt{n}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq C\}$. Luego, para todo $\omega \in \mathcal{B}_n$,

$$\widehat{\boldsymbol{\beta}}_n = \left(\widetilde{\boldsymbol{\beta}}_0^{\mathrm{T}} + \frac{\mathbf{U}_{1,n}^{\mathrm{T}}}{\sqrt{n}}, \frac{U_{2,n}}{\sqrt{n}}\right)^{\mathrm{T}}$$

donde $\mathbf{U}_{1,n} \in \mathbb{R}^{p-1}$, $U_{2,n} \in \mathbb{R}$, $\|\mathbf{U}_n\|_2 \leq C$ y $\mathbf{U}_n = (\mathbf{U}_{1,n}^{\mathrm{T}}, U_{2,n})^{\mathrm{T}}$ es tal que

$$\mathbf{U}_n = (\mathbf{U}_{1,n}^{\mathrm{T}}, U_{2,n})^{\mathrm{T}} = \underset{\|(\mathbf{u}_1,u_2)\|_2 \leq C}{\operatorname{argmin}} V_n(\mathbf{u}_1, u_2)\,.$$

Nuestro objetivo es probar que si $\|\mathbf{u}_1\|^2 + u_2^2 \leq C^2$ y $u_2 \neq 0$, $V_n(\mathbf{u}_1, u_2) - V_n(\mathbf{u}_1, 0) > 0$ con alta probabilidad. Consideramos $\mathbf{u}_1 \in \mathbb{R}^{p-1}$ y $u_2 \neq 0$ tal que $\|\mathbf{u}_1\|^2 + u_2^2 \leq C^2$. Observemos que $V_n(\mathbf{u}_1, u_2) - V_n(\mathbf{u}_1, 0) = S_{1,n} + S_{2,n}$ donde

$$\begin{aligned} S_{1,n} &= L_n\left(\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}, \frac{u_2}{\sqrt{n}}\right) - L_n\left(\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}, 0\right), \\ S_{2,n} &= I_{\lambda_n}\left(\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}, \frac{u_2}{\sqrt{n}}\right) - I_{\lambda_n}\left(\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}, 0\right). \end{aligned}$$

Primero, vamos a acotar $S_{1,n}$. Denotamos $\mathbf{u}_n^{(0)} = \left(\mathbf{0}_{p-1}^{\mathrm{T}}, u_2/\sqrt{n}\right)^{\mathrm{T}}$. Luego, el Teorema de Valor Medio implica que

$$S_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^*)\, w(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)}\,,$$

donde

$$\boldsymbol{\beta}_n^* = \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_0 + \dfrac{\mathbf{u}_1}{\sqrt{n}} \\ \\ \alpha_{n,1}\dfrac{u_2}{\sqrt{n}} \end{pmatrix},$$

y $\alpha_{n,1} \in [0,1]$. Además, usando nuevamente el Teorema de Valor Medio, se tiene que

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}[\Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^*) - \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)]\, w(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)} &= \frac{1}{n}\sum_{i=1}^{n} \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^{**})\, w(\mathbf{X}_i)\,(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)} \\ &= (\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_n(\boldsymbol{\beta}_n^{**})\mathbf{u}_n^{(0)}\,, \end{aligned}$$

donde $\mathbf{A}_n(\boldsymbol{\beta})$ es la matriz dada en (5.12) y

$$\boldsymbol{\beta}_n^{**} = \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_0 + \alpha_{n,2}\dfrac{\mathbf{u}_1}{\sqrt{n}} \\ \\ \alpha_{n,2}\,\alpha_{n,1}\dfrac{u_2}{\sqrt{n}} \end{pmatrix}$$

con $\alpha_{n,2} \in [0,1]$. Por lo tanto, usando que

$$S_{1,n} = \left\{\frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\, w(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}} + \frac{1}{n}\sum_{i=1}^{n}[\Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^*) - \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)]\, w(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}}\right\}\mathbf{u}_n^{(0)}\,,$$

obtenemos $S_{1,n} = S_{11,n} + S_{12,n}$ donde

$$\begin{aligned} S_{11,n} &= \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\, w(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)} = \frac{1}{n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\, w(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}}(\mathbf{0}_{p-1}^{\mathrm{T}}, u_2)^{\mathrm{T}}\,, \\ S_{12,n} &= (\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\, \mathbf{A}_n(\boldsymbol{\beta}_n^{**})\mathbf{u}_n^{(0)} = \frac{1}{n}(\mathbf{u}_1^{\mathrm{T}}, \alpha_{n,1}\, u_2)\, \mathbf{A}_n(\boldsymbol{\beta}_n^{**})(\mathbf{0}_{p-1}^{\mathrm{T}}, u_2)^{\mathrm{T}}\,. \end{aligned}$$

Usando (3.13) y el Teorema Central del Límite Multivariado, tenemos que $n\,|S_{11,n}| = O_{\mathbb{P}}(1)\,|u_2|$. Por otra parte,

$$\begin{aligned} |S_{12,n}| &\leq \frac{1}{n}\left|(\mathbf{u}_1, \alpha_{n,1}\, u_2)^{\mathrm{T}}\, \frac{1}{n}\sum_{i=1}^{n} \chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^{**})\, w(\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}(\mathbf{0}_{p-1}, u_2)\right| \\ &\leq \frac{1}{n}\|\chi\|_\infty \|(\mathbf{u}_1, \alpha_{n,1}\, u_2)\|_2 \left(\frac{1}{n}\sum_{i=1} w(\mathbf{X}_i)\|\mathbf{X}_i\|^2\right)|u_2| \leq \frac{1}{n}\|\chi\|_\infty\, C\, \left(\frac{1}{n}\sum_{i=1} w(\mathbf{X}_i)\|\mathbf{X}_i\|^2\right)|u_2|\,. \end{aligned}$$

Entonces, como **X3** y la Ley de los Grandes Números implican que $(1/n)\sum_{i=1} w(\mathbf{X}_i)\|\mathbf{X}_i\|^2 \xrightarrow{p} \mathbb{E}w(\mathbf{X}_i)\|\mathbf{X}\|^2$, se tiene

$$C\,\|\chi\|_\infty \left(\frac{1}{n}\sum_{i=1} w(\mathbf{X}_i)\|\mathbf{X}_i\|^2\right) = O_{\mathbb{P}}(1).$$

Como consecuencia, $n|S_{12,n}| = O_{\mathbb{P}}(1)\,|u_2|$ lo cual, junto con el hecho de que $n|S_{11,n}| = O_{\mathbb{P}}(1)\,|u_2|$, implica que $n|S_{1,n}| = A_n\,|u_2|$ con $A_n \geq 0$ y $A_n = O_{\mathbb{P}}(1)$.

Sea M_p tal que $\mathbb{P}(0 \leq A_n < M_p) \geq 1 - \tau^*$ para todo n. Vale la pena mencionar que M_p depende de C y por lo tanto, de τ. Luego, si $\mathcal{D}_n = \{n\, S_{1,n} > -M_p\,|u_2|\}$,

$$\mathbb{P}(\mathcal{D}_n) \geq \mathbb{P}(0 \leq A_n < M_p) \geq 1 - \tau^*\,. \tag{5.17}$$

Tomemos N_p y K_p (ambas cantidades dependientes de C) tales que si $n \geq N_p$ y $\|\mathbf{u}\|_2 \leq C$,

$$I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}^{(-p)}}{\sqrt{n}}\right) \geq K_p\,\frac{\lambda_n}{\sqrt{n}}\,|u_p|\,,$$

lo que implica que $n\,S_{2,n} \geq K_p\,\lambda_n\,\sqrt{n}\,|u_p|$. Entonces, en el evento $\mathcal{B}_n \cap \mathcal{D}_n$ y para $n \geq N_p$, tenemos que

$$V(\mathbf{U}_{1,n}, U_{2,n}) - V(\mathbf{U}_{1,n}, 0) = S_{1,n} + S_{2,n} \geq \frac{1}{n}|U_{2,n}|(K_p\,\lambda_n\,\sqrt{n} - M_p)\,. \tag{5.18}$$

Luego, si $b \geq M_p/K_p$,

$$\mathbb{P}(\widehat{\beta}_{n,p} = 0) \geq 1 - 2\tau^* = 1 - \tau/(p-k)\,.$$

Podemos ahora proceder secuencialmente con el mismo razonamiento para todas las coordenadas no activas, obteniendo así valores $N_s \in \mathbb{N}$, $K_s \in \mathbb{R}$, $M_s \in \mathbb{R}$ para $s = p, p-1, \ldots, k+1$.

El item (a) se obtiene tomando $n_0 = \max(N_p, N_{p-1}, \ldots, N_{k+1})$ y $b > \max(A_p, A_{p-1}, \ldots, A_{k+1})$ donde $A_j = M_j/K_j$.

Por otra parte, para (b) si $\sqrt{n}\lambda_n \to \infty$, entonces existe $\widetilde{n}_0$ tal que para $n \geq \widetilde{n}_0$, $\sqrt{n}\lambda_n > \max\{M_p/K_p, \ldots, M_{k+1}/K_{k+1}\}$. Por lo tanto, si $n_0 = \max\{\widetilde{n}_0, N_p, \ldots, N_{k+1}\}$, para todo $n \geq n_0$ tenemos que

$$\mathbb{P}(\widehat{\beta}_{n,p} = 0 \cap \widehat{\beta}_{n,p-1} = 0 \cap \cdots \cap \widehat{\beta}_{n,k+1} = 0) \geq 1 - \tau,$$

lo que concluye la demostración. ■

Demostración del Corolario 5.6. Para poder aplicar el Teorema 5.5, sólo resta probar que se cumple la condición (5.4). Sin perder generalidad, solamente probaremos que dicha condición se cumple para la última coordenada, es decir, para $\ell = p$. Fijamos $C > 0$ y tomamos $\mathbf{u} = (\mathbf{u}_1, u_2)$ con $\mathbf{u}_1 \in \mathbb{R}^{p-1}$, $u_2 \in \mathbb{R}$ y $\|\mathbf{u}_1\|_2^2 + u_2^2 \leq C^2$.

En primer lugar probaremos el item (a). Dado un vector $\widetilde{\mathbf{u}} \in \mathbb{R}^{p-1} - \{\mathbf{0}_{p-1}\}$ consideremos la función

$$h(u) = h_{\widetilde{\mathbf{u}}}(u) = J(\widetilde{\mathbf{u}}, u)\,,$$

donde $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2$. Para tener una idea más esclarecedora del comportamiento de esta función, en la Figura 5.1 se grafica $h_{\widetilde{\mathbf{u}}}$ cuando $\widetilde{\mathbf{u}} = \mathbf{1}_{15}$.

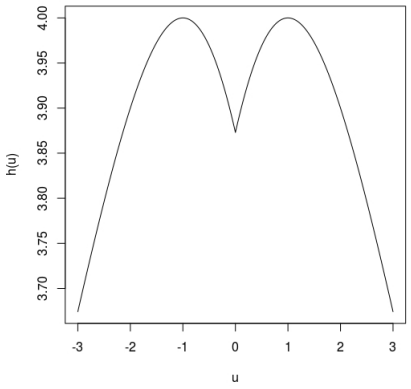


Figura 5.1: Gráfico de la función $h_{\widetilde{\mathbf{u}}}(u) = J(\widetilde{\mathbf{u}}, u)$ cuando $\widetilde{\mathbf{u}} = \mathbf{1}_{15}$.

La función $h(u)$ claramente no es derivable cuando $u = 0$, pero para $u \neq 0$ su derivada viene dada por

$$h'(u) = \frac{\operatorname{sign}(u)\left(\sum_{j=1}^{p-1} \widetilde{u}_\ell^2 + u^2\right) - u\left(\sum_{j=1}^{p-1} |\widetilde{u}_\ell| + |u|\right)}{\left(\sum_{j=1}^{p-1} \widetilde{u}_\ell^2 + u^2\right)^{\frac{3}{2}}}\,.$$

Cálculos sencillos muestran que los puntos críticos de h son $\pm\|\widetilde{\mathbf{u}}\|_2^2/\|\widetilde{\mathbf{u}}\|_1$, ambos máximos locales. Entonces, $h_{\widetilde{\mathbf{u}}}$ es una función creciente en $|u|$ cuando $|u| \leq \|\widetilde{\mathbf{u}}\|_2^2/\|\widetilde{\mathbf{u}}\|_1$. Más aún,

$$\lim_{u\to 0^+} h'(u) = \frac{1}{\|\widetilde{\mathbf{u}}\|_2} \qquad \text{y} \qquad \lim_{u\to 0^-} h'(u) = -\frac{1}{\|\widetilde{\mathbf{u}}\|_2}.$$

Dado un vector $\mathbf{u}_1$, indiquemos por

$$h_{n,\mathbf{u}_1}(u) = h_{\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}}(u) = J\left(\widetilde{\boldsymbol{\beta}}_0 + \frac{\mathbf{u}_1}{\sqrt{n}}, u\right)\,.$$

Usando que los puntos críticos de $h_{\widetilde{\mathbf{u}}}$ son $\pm\|\widetilde{\mathbf{u}}\|_2^2/\|\widetilde{\mathbf{u}}\|_1$, obtenemos que los puntos críticos de $h_{n,\mathbf{u}_1}$ son $c_n^+(\mathbf{u}_1) = \|\widetilde{\boldsymbol{\beta}}_0 + \mathbf{u}_1/\sqrt{n}\|_2^2/\|\widetilde{\boldsymbol{\beta}}_0 + \mathbf{u}_1/\sqrt{n}\|_1$ y $c_n^-(\mathbf{u}_1) = -c_n^+(\mathbf{u}_1)$ que convergen, respectivamente, a c^+ y $c^- = -c^+$, uniformemente sobre conjuntos compactos, donde

$$c^+ = \frac{\|\widetilde{\boldsymbol{\beta}}_0\|_2^2}{\|\widetilde{\boldsymbol{\beta}}_0\|_1} = \frac{\|\boldsymbol{\beta}_{0,I}\|_2^2}{\|\boldsymbol{\beta}_{0,I}\|_1}\,.$$

Esto implica que $\lim_{n\to\infty} \sup_{\|\mathbf{u}_1\|\leq C} |c_n^+(\mathbf{u}_1) - c^+| = 0$. Más aún,

$$\lim_{n\to\infty} \lim_{u\to 0^+} h'_{n,\mathbf{u}_1}(u) = \frac{1}{\|\widetilde{\boldsymbol{\beta}}_0\|_2} \qquad \text{y} \qquad \lim_{n\to\infty} \lim_{u\to 0^-} h'_{n,\mathbf{u}_1}(u) = -\frac{1}{\|\widetilde{\boldsymbol{\beta}}_0\|_2},$$

donde, nuevamente, la convergencia es uniforme sobre conjuntos compactos ya que $\|\widetilde{\boldsymbol{\beta}}_0\|_2 = \|\boldsymbol{\beta}_{0,I}\|_2 \neq 0$.

Sea $n_1 \in \mathbb{N}$ y $\delta > 0$ tal que para $n \geq n_1$ y $0 < |u| < \delta$ se tiene que

$$\sup_{\|\mathbf{u}_1\|\leq C} \left| h'_{n,\mathbf{u}_1}(u) - \frac{\operatorname{sign}(u)}{\|\widetilde{\boldsymbol{\beta}}_0\|_2} \right| < \frac{1}{2\|\widetilde{\boldsymbol{\beta}}_0\|_2}.$$

Tomamos ahora n_2 tal que $Cn^{-1/2} \leq \min(c^+/2, \delta)$ y $\sup_{\|\mathbf{u}_1\|\leq C} |c_n^+(\mathbf{u}_1) - c^+| \leq c^+/2$ para $n \geq n_2$. Entonces, si $n \geq N_p = \max(n_1, n_2)$ obtenemos que para todo $\mathbf{u}_1 \in \{\mathbf{v} \in \mathbb{R}^{p-1} : \|\mathbf{v}\| \leq C\}$,

$$Cn^{-1/2} \leq \min\left(\frac{c^+}{2}, \delta\right), c_n^+(\mathbf{u}_1) > \frac{c^+}{2} \quad \text{y} \quad \left|h'_{n,\mathbf{u}_1}(u)\right| \mathbb{I}_{\{0<|u|<\delta\}} < \frac{1}{2\|\widetilde{\boldsymbol{\beta}}_0\|_2}. \tag{5.19}$$

En particular, las funciones $h_{n,\mathbf{u}_1}(u)$ son crecientes en $|u|$ al restringirlas al intervalo $[-c^+/2, c^+/2]$.

Usando que $\|\mathbf{u}\|_2 \leq C$ (lo cual implica que $\|\mathbf{u}_1\| \leq C$ y $0 < |u_2| \leq C$) y (5.19), se tiene que

$$h_{n,\mathbf{u}_1}\left(\frac{u_2}{\sqrt{n}}\right) > h_{n,\mathbf{u}_1}(0).$$

Por otra parte, si ξ_n es un punto intermedio entre 0 y $u_2/\sqrt{n}$, obtenemos que $0 < |\xi_n| \leq Cn^{-1/2} < \delta$ y entonces

$$h_{n,\mathbf{u}_1}\left(\frac{u_2}{\sqrt{n}}\right) - h_{n,\mathbf{u}_1}(0) = |h'_{n,\mathbf{u}_1}(\xi_n)|\frac{|u_2|}{\sqrt{n}} > \frac{1}{2\|\widetilde{\boldsymbol{\beta}}_0\|_2}\,\frac{|u_2|}{\sqrt{n}}.$$

Finalmente, tenemos que

$$I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}^{(-p)}}{\sqrt{n}}\right) = \lambda_n\left\{h_{n,\mathbf{u}_1}\left(\frac{u_2}{\sqrt{n}}\right) - h_{n,\mathbf{u}_1}(0)\right\},$$

por lo que (5.4) se cumple tomando $K_{C,p} = 1/(2\|\widetilde{\boldsymbol{\beta}}_0\|_2) = 1/(2\|\boldsymbol{\beta}_0\|_2)$ y $N_{C,p} = N_p$. El resultado buscado se obtiene usando el item (a) del Teorema 5.5.

Probamos ahora el item (b). Para la penalización SCAD, es fácil ver que

$$I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}^{(-p)}}{\sqrt{n}}\right) = \mathrm{SCAD}_{\lambda_n,a}\left(\frac{u_2}{\sqrt{n}}\right),$$

donde

$$\mathrm{SCAD}_{\lambda,a}(\beta) = \begin{cases} \lambda|\beta| & \text{si} \quad |\beta| \leq \lambda \\ \dfrac{1}{a-1}\left(a\lambda|\beta| - \dfrac{\beta^2+\lambda^2}{2}\right) & \text{si} \quad \lambda < |\beta| \leq a\lambda \\ \dfrac{\lambda^2(a^2-1)}{2(a-1)} & \text{si} \quad |\beta| > a\lambda. \end{cases} \tag{5.20}$$

Tomamos n_0 tal que para $n \geq n_0$, $\sqrt{n}\lambda_n \geq C$. Si esto sucede, entonces $\mathrm{SCAD}_{\lambda_n,a}(u_2/\sqrt{n}) = \lambda_n|u_2|/\sqrt{n}$, por lo que cumple la condición (5.4) para $K_{C,p} = 1$ y $N_{C,p} = n_0$.

Para la penalización MCP, la demostración es muy similar. En este caso,

$$I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}^{(-p)}}{\sqrt{n}}\right) = \mathrm{MCP}_{\lambda_n,a}\left(\frac{u_2}{\sqrt{n}}\right),$$

donde

$$\text{MCP}_{\lambda,a}(\beta) = \begin{cases} \lambda|\beta| - \dfrac{\beta^2}{2a} & \text{si} \quad |\beta| \leq a\lambda \\ \dfrac{a\lambda^2}{2} & \text{si} \quad |\beta| > a\lambda. \end{cases} \tag{5.21}$$

Al igual que antes, tomando n_0 tal que si $n \geq n_0$, $\sqrt{n}\lambda_n \geq C/a$ deducimos que

$$\text{MCP}_{\lambda_n,a}\left(\frac{u_2}{\sqrt{n}}\right) = \lambda_n \frac{|u_2|}{\sqrt{n}} - \frac{u_2^2}{2na} = \lambda_n \frac{|u_2|}{\sqrt{n}}\left(1 - \frac{u_2}{2\sqrt{n}\lambda_n a}\right) \geq \frac{1}{2}\lambda_n \frac{|u_2|}{\sqrt{n}},$$

por lo que se cumple la condición (5.4) para $K_{C,p} = 1/2$ y $N_{C,p} = n_0$.

Tanto para SCAD como para MCP, el resultado buscado es consecuencia del item (b) del Teorema 5.5 . ■

5.8. Apéndice D: Demostraciones de la Sección 5.4

Para probar el Teorema 5.7 necesitaremos los dos siguientes lemas.

Lema 5.15. *Sea $\phi(y,t)$ dada por (3.7) donde la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3**. Asumamos que la matriz $\mathbf{A}$ definida en (3.17) es no singular y definamos el proceso $R_1 : \mathbb{R}^p \to \mathbb{R}$ como*

$$R_1(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}\,,$$

donde $\mathbf{w} \sim N_p(\mathbf{0}, \mathbf{B})$ y $\mathbf{B}$ es la matriz definida en (3.18). Más aún, sea $R_{n,1}(\mathbf{z})$ el proceso definido como

$$R_{n,1}(\mathbf{z}) \;=\; \sum_{i=1}^{n} \left\{ \phi\left(Y_i, \mathbf{X}_i^{\mathrm{T}}\left[\boldsymbol{\beta}_0 + \frac{\mathbf{z}}{\sqrt{n}}\right]\right) - \phi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0) \right\} w(\mathbf{X}_i)\,. \tag{5.22}$$

Luego, el proceso $R_{n,1}$ converge en distribución a R_1.

Demostración. Por el Teorema 2.3 de Kim y Pollard (1990), es suficiente probar la convergencia de las distribuciones finito–dimensionales y la equicontinuidad estocástica, es decir, basta mostrar que

(a) Dados $\mathbf{z}_1, \ldots, \mathbf{z}_s \in \mathbb{R}^p$, $(R_{n,1}(\mathbf{z}_1), \ldots, R_{n,1}(\mathbf{z}_s))^{\mathrm{T}} \xrightarrow{D} (R_1(\mathbf{z}_1), \ldots, R_1(\mathbf{z}_s))^{\mathrm{T}}$.

(b) Dados $\epsilon > 0$, $\eta > 0$ y $M < \infty$, existe un $\delta > 0$ tal que

$$\limsup_{n\to\infty} \mathbb{P}^* \left(\sup_{\substack{\|\mathbf{u}\|_2 \leq M, \|\mathbf{v}\|_2 \leq M \\ \|\mathbf{u}-\mathbf{v}\|_2 < \delta}} |R_{n,1}(\mathbf{u}) - R_{n,1}(\mathbf{v})| > \epsilon \right) < \eta,$$

donde $\mathbb{P}^*$ indica probabilidad exterior.

Probaremos primero (a). Sin pérdida de generalidad podemos suponer que $s = 1$, ya que para $s > 1$ la demostración se extiende trivialmente usando el dispositivo de Cramer–Wald, es decir, tomando proyecciones para cualquier $\mathbf{a} \in \mathbb{R}^s$.

Fijemos $\mathbf{z} \in \mathbb{R}^p$. Usando un desarrollo de Taylor de primer orden, tenemos que

$$R_{n,1}(\mathbf{z}) = \sqrt{n}\,\mathbf{z}^{\mathrm{T}}\nabla L_n(\boldsymbol{\beta}_0) + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{z}})\mathbf{z}\,,$$

donde $L_n(\boldsymbol{\beta})$ y $\mathbf{A}_n(\boldsymbol{\beta})$ están definidos en (3.10) y (5.12) respectivamente y

$$\widetilde{\boldsymbol{\beta}}_{\mathbf{z}} = \boldsymbol{\beta}_0 + \frac{\tau_n \mathbf{z}}{\sqrt{n}}\,,$$

es un punto intermedio con $\tau_n \in [0,1]$. La Fisher–consistencia dada en (3.13) y el Teorema Central del Límite Multivariado implican que

$$\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0) \xrightarrow{D} N_p(\mathbf{0}, \mathbf{B})\,,$$

ya que $\textsc{Var}[\Psi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\, w(\mathbf{X})\mathbf{X}]$ es igual a la matriz $\mathbf{B}$ definida en (3.18). Por otra parte, como consecuencia del Lema 5.14, se tiene que

$$\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{z}}) \xrightarrow{p} \mathbf{A}.$$

Por lo tanto, usando el Teorema de Slutsky, obtenemos que $R_{n,1} \xrightarrow{D} \mathbf{z}^{\mathrm{T}}\mathbf{w} + (1/2)\ \mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}$, lo que prueba (a).

Para obtener (b), realizamos un desarrollo de Taylor de primer orden de $R_{n,1}(\mathbf{u})$ y $R_{n,1}(\mathbf{v})$ alrededor de $\boldsymbol{\beta}_0$:

$$R_{n,1}(\mathbf{u}) - R_{n,1}(\mathbf{v}) = \sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)^{\mathrm{T}}(\mathbf{u}-\mathbf{v}) + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}})\mathbf{u} - \frac{1}{2}\mathbf{v}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v},$$

donde $\widetilde{\boldsymbol{\beta}}_{\mathbf{u}}$ y $\widetilde{\boldsymbol{\beta}}_{\mathbf{v}}$ están definidas como

$$\widetilde{\boldsymbol{\beta}}_{\mathbf{v}} = \boldsymbol{\beta}_0 + \frac{\tau_{\mathbf{v},n}\mathbf{v}}{\sqrt{n}} \qquad \widetilde{\boldsymbol{\beta}}_{\mathbf{u}} = \boldsymbol{\beta}_0 + \frac{\tau_{\mathbf{u},n}\mathbf{u}}{\sqrt{n}}\,,$$

con $\tau_{\mathbf{v},n}$, $\tau_{\mathbf{u},n} \in [0,1]$. Notemos que $\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)^{\mathrm{T}}(\mathbf{u}-\mathbf{v}) \le O_{\mathbb{P}}(1)\|\mathbf{u}-\mathbf{v}\|_2$ y

$$\begin{aligned}\mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}})\mathbf{u} - \mathbf{v}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v} &= \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}})\mathbf{u} - \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{u} + \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{u} \\ &- \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v} + \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v}.\end{aligned}$$

Por lo tanto, si $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 \le M$ e indicamos por $\|\mathbf{C}\|$ la norma de Frobenius de la matriz $\mathbf{C}$, obtenemos que

$$\begin{aligned}|R_{n,1}(\mathbf{u}) - R_{n,1}(\mathbf{v})| &\le O_p(1)\|\mathbf{u}-\mathbf{v}\|_2 + M^2\left\|\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}}) - \mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\right\| \\ &\quad +2\|\mathbf{u}-\mathbf{v}\|_2 M\left\|\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\right\|\,.\end{aligned}$$

El Lema 5.14 implica que $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}}) - \mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}}) \xrightarrow{p} 0$ y $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}}) \xrightarrow{p} \mathbf{A}$ uniformemente sobre $\{\mathbf{u}, \mathbf{v} \in \mathbb{R}^p : \text{máx}\{\|\mathbf{u}\|_2, \|\mathbf{v}\|_2\} \le M\}$, de donde fácilmente se concluye el resultado del item (b). ■

Cabe observar que el Teorema 2.3 de Kim y Pollard (1990) afirma que si se cumplen las condiciones (a) y (b) en la demostración del Lema 5.15, entonces el proceso estocástico límite existe y sus proyecciones finito–dimensionales son las mismas que las del proceso $R_1(\mathbf{z})$. Sin embargo, como los procesos estocásticos que concentran sus "caminos" en $\mathcal{C}_b(\mathbb{R}^p)$ están determinados por sus proyecciones finito–dimensionales, podemos concluir que R_1 debe ser dicho límite.

En el siguiente Lema, al igual que en el Teorema 5.7, permitimos que el parámetro de penalización λ_n sea aleatorio.

Lema 5.16. *Sea $I_{\lambda_n}(\boldsymbol{\beta})$ una penalización que satisface **P1** y tal que $\sqrt{n}\lambda_n = O_{\mathbb{P}}(1)$. Definamos*

$$R_{n,2}(\mathbf{z}) = n\left\{I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{z}}{\sqrt{n}}\right) - I_{\lambda_n}(\boldsymbol{\beta}_0)\right\}. \tag{5.23}$$

Luego, el proceso $R_{n,2}(\mathbf{z})$ es equicontinuo, es decir, dados $\epsilon > 0$, $\eta > 0$ y $M < \infty$ existe $\delta > 0$ tal que

$$\limsup_{n\to\infty} \mathbb{P}^*\left(\sup_{\substack{\|\mathbf{u}\|_2 \le M, \|\mathbf{v}\|_2 \le M \\ \|\mathbf{u}-\mathbf{v}\|_2 < \delta}} |R_{n,2}(\mathbf{u}) - R_{n,2}(\mathbf{v})| > \epsilon\right) < \eta,$$

Demostración. Notemos que **P1** implica que

$$\begin{aligned} |R_{n,2}(\mathbf{u}) - R_{n,2}(\mathbf{v})| &= n\left|I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{u}}{\sqrt{n}}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \frac{\mathbf{v}}{\sqrt{n}}\right)\right| \le n\,\lambda_n K\,\frac{\|\mathbf{u}-\mathbf{v}\|_1}{\sqrt{n}} \\ &\le \sqrt{n}\,\lambda_n\,K\,\sqrt{p}\,\|\mathbf{u}-\mathbf{v}\|_2\,, \end{aligned}$$

de donde se obtiene el resultado pues $\sqrt{n}\lambda_n = O_{\mathbb{P}}(1)$. ■

Demostración del Teorema 5.7. Consideremos el proceso estocástico indexado en $\mathbf{z}$ definido por $R_n(\mathbf{z}) = R_{n,1}(\mathbf{z}) + R_{n,2}(\mathbf{z})$, donde $R_{n,1}(\mathbf{z})$ y $R_{n,2}(\mathbf{z})$ están dados en (5.22) y (5.23), respectivamente, con $I_\lambda(\boldsymbol{\beta}) = \lambda J(\boldsymbol{\beta})$ y $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2$. Observemos que $\operatorname{argmin}_{\mathbf{z}} R_n(\mathbf{z}) = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$.

Para probar que $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = \operatorname{argmin}_{\mathbf{z}} R_n(\mathbf{z}) \xrightarrow{D} \operatorname{argmin}_{\mathbf{z}} R(\mathbf{z})$, usaremos el Teorema 2.7 de Kim y Pollard (1990). La condición (iii) de dicho resultado se verifica trivialmente y la condición (ii) es una consecuencia directa del Teorema 5.4. Luego, basta con probar la condición (i) del mencionado Teorema, es decir, debemos probar que el proceso $R_n(\mathbf{z})$ converge en distribución al proceso $R(\mathbf{z})$. Con este propósito, como en la demostración del Lema 5.15, basta con probar la convergencia de las distribuciones finito–dimensionales y la equicontinuidad estocástica, es decir, hay que probar que se cumplen las siguientes dos condiciones:

(a) Dados $\mathbf{z}_1, \ldots, \mathbf{z}_s \in \mathbb{R}^p$ $(R_n(\mathbf{z}_1), \ldots, R_n(\mathbf{z}_s))^{\mathrm{T}} \xrightarrow{D} (R(\mathbf{z}_1), \ldots, R(\mathbf{z}_s))^{\mathrm{T}}$.

(b) Dados $\epsilon > 0$, $\eta > 0$ y $M < \infty$, existe un $\delta > 0$ tal que

$$\limsup_{n\to\infty} \mathbb{P}^*\left(\sup_{\substack{\|\mathbf{u}\|_2 \le M, \|\mathbf{v}\|_2 \le M \\ \|\mathbf{u}-\mathbf{v}\|_2 < \delta}} |R_n(\mathbf{u}) - R_n(\mathbf{v})| > \epsilon\right) < \eta,$$

donde $\mathbb{P}^*$ indica probabilidad exterior.

Usando que $\boldsymbol{\beta}_0 \neq \mathbf{0}$, es fácil ver que la penalización Signo satisface **P1** y por lo tanto, la equicontinuidad descripta en (b) es consecuencia de los Lemas 5.15 y 5.16.

Solamente resta mostrar que se cumple (a). Tal como se observó en la prueba del Lema 5.15, es suficiente considerar el caso $s = 1$. Con este fin, fijemos $\mathbf{z} \in \mathbb{R}^p$.

Usando el Lema 5.15 tenemos que $R_{n,1}(\mathbf{z}) \xrightarrow{D} \mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}$. Por lo tanto, bastará probar que $R_{n,2}(\mathbf{z}) \xrightarrow{p} b\,\mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$. Notemos que, para $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2$, se tiene que

$$J(\boldsymbol{\beta}) = \sum_{\ell=1}^{p} \frac{|\beta_\ell|}{\|\boldsymbol{\beta}\|_2} = \sum_{\ell=1}^{p} J_\ell(\boldsymbol{\beta})$$

donde J_ℓ es diferenciable en todo punto salvo en el hiperplano $\{\boldsymbol{\beta} : \beta_\ell = 0\}$. Supongamos que $\beta_{0,\ell} \neq 0$. Luego, para n suficientemente grande, $\beta_{0,\ell} + z_\ell/\sqrt{n}$ tiene el mismo signo que $\beta_{0,\ell}$. Por lo tanto, usando el Teorema de Valor Medio, obtenemos que

$$J_\ell\left(\boldsymbol{\beta}_0 + \frac{\mathbf{z}}{\sqrt{n}}\right) - J_\ell(\boldsymbol{\beta}_0) = \left[\nabla J_\ell\left(\boldsymbol{\beta}_0 + \alpha_{n,\ell}\frac{\mathbf{z}}{\sqrt{n}}\right)\right]^{\mathrm{T}} \frac{\mathbf{z}}{\sqrt{n}}\,,$$

con $\alpha_{n,\ell} \in [0,1]$. Cuando $\beta_{0,\ell} = 0$, $J_\ell(\boldsymbol{\beta}_0) = 0$, entonces

$$J_\ell\left(\boldsymbol{\beta}_0 + \frac{\mathbf{z}}{\sqrt{n}}\right) - J_\ell(\boldsymbol{\beta}_0) = \frac{|z_\ell|}{\sqrt{n}} \frac{1}{\left\|\boldsymbol{\beta}_0 + \frac{\mathbf{z}}{\sqrt{n}}\right\|_2}\,.$$

Luego, $R_{n,2}(\mathbf{z})$ puede ser escrito como

$$R_{n,2}(\mathbf{z}) \;=\; \sqrt{n}\,\lambda_n \left\{ \sum_{\ell=1}^{p} \left[\nabla J_\ell\left(\boldsymbol{\beta}_0 + \alpha_{n,\ell}\frac{\mathbf{z}}{\sqrt{n}}\right)\right]^{\mathrm{T}} \mathbf{z}\, \mathbb{I}_{\{\beta_{0.\ell}\neq 0\}} + \sum_{\ell=1}^{p} |z_\ell| \frac{1}{\left\|\boldsymbol{\beta}_0 + \frac{\mathbf{z}}{\sqrt{n}}\right\|_2} \mathbb{I}_{\{\beta_{0.\ell}=0\}} \right\}$$

lo cual implica que $R_{n,2}(\mathbf{z}) \xrightarrow{p} b\,\mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$. Concluimos entonces, usando el Teorema de Slutzky, que se cumple la condición (a). ■

Demostración del Teorema 5.8. La prueba esencialmente utiliza la misma estrategia que la del Teorema 5.7. Consideramos el proceso estocástico indexado en $\mathbf{z}$ definido como $R_n(\mathbf{z}) = R_{n,1}(\mathbf{z}) + R_{n,2}(\mathbf{z})$, donde $R_{n,1}(\mathbf{z})$ y $R_{n,2}(\mathbf{z})$ están dados en (5.22) y (5.23), respectivamente, con

$$I_{\lambda_n}(\boldsymbol{\beta}) = \lambda_n \left\{ (1-\alpha)\sum_{\ell=1}^{p} J_\ell(|\beta_\ell|) + \alpha \sum_{\ell=1}^{p} |\beta_\ell| \right\}.$$

Para probar que $R_n(\mathbf{z})$ converge en distribución a $R(\mathbf{z})$, debemos probar que se cumplen las condiciones (a) y (b) descriptas en la demostración del Teorema 5.7.

Usando que $J_\ell(\cdot)$ es una función continuamente diferenciable, es fácil ver que $I_\lambda(\boldsymbol{\beta})$ satisface **P1**. Luego, la equicontinuidad del item (b) es consecuencia directa de los Lemas 5.15 y 5.16.

Para probar (a), nuevamente alcanza con considerar el caso $s = 1$. Con ese motivo, fijamos $\mathbf{z} \in \mathbb{R}^p$.

Por el Lema 5.15, tenemos que $R_{n,1}(\mathbf{z}) \xrightarrow{D} \mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}$, luego basta con estudiar la convergencia en probabilidad de $R_{n,2}(\mathbf{z})$. Notemos que $R_{n,2}(\mathbf{z}) = R_{n,2,1}(\mathbf{z}) + R_{n,2,1}(\mathbf{z})$ donde

$$\begin{aligned} R_{n,2,1}(\mathbf{z}) &= n\,\lambda_n(1-\alpha)\left\{ \sum_{\ell=1}^{p} J_\ell\left(\left|\beta_{0,\ell} + \frac{z_\ell}{\sqrt{n}}\right|\right) - J_\ell(|\beta_{0,\ell}|)\right\}, \\ R_{n,2,2}(\mathbf{z}) &= n\,\lambda_n\alpha\left\{ \sum_{\ell=1}^{p} \left|\beta_{0,\ell} + \frac{z_\ell}{\sqrt{n}}\right| - |\beta_{0,\ell}|\right\}. \end{aligned}$$

Usando argumentos estándar, es fácil ver que

$$\begin{aligned} R_{n,2,1}(\mathbf{z}) &\xrightarrow{p} b\,(1-\alpha)\sum_{\ell=1}^{p} J'_\ell\left(|\beta_{0,\ell}|\right)\,\mathrm{sign}(\beta_{0,\ell})\,z_\ell\,, \\ R_{n,2,2}(\mathbf{z}) &\xrightarrow{p} b\,\alpha\sum_{\ell=1}^{p}\left\{ z_\ell\,\mathrm{sign}(\beta_{0,\ell})\mathbb{I}_{\{\beta_{0,\ell}\neq 0\}} + |z_\ell|\mathbb{I}_{\{\beta_{0,\ell}=0\}}\right\}, \end{aligned}$$

uniformemente sobre conjuntos compactos. Por lo tanto, $R_{n,2}(\mathbf{z}) \xrightarrow{p} b\,\mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$ y el resultado buscado es consecuencia del Teorema de Slutsky. ■

Demostración del Corolario 5.9. Probaremos solamente el caso en que la penalización es la penalidad Signo, el caso de LASSO se obtiene en forma análoga.

Sean

$$\alpha_\ell = \frac{\operatorname{sign}(\beta_{0,\ell})}{\|\boldsymbol{\beta}_0\|_2} + \beta_{0,\ell} \sum_{j=1}^{k} \frac{|\beta_{0,j}|}{\|\boldsymbol{\beta}_0\|_2^3} \qquad \mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \quad \mathbf{A}_{11} \in \mathbb{R}^{k\times k}, \mathbf{A}_{22} \in \mathbb{R}^{(p-k)\times(p-k)}\,.$$

Como $\beta_{0,\ell} = 0$ si $\ell = k+1, \ldots, p$, el Teorema 5.7 implica que $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \mathbf{z}^\star = \operatorname{argmin}_{\mathbf{z}} R(\mathbf{z})$ donde

$$R(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + \frac{b}{\|\boldsymbol{\beta}_0\|_2} \sum_{\ell=k+1}^{p} |z_\ell| + b \sum_{\ell=1}^{k} z_\ell\, \alpha_\ell$$

con $\mathbf{w} \sim N_p(\mathbf{0}, \mathbf{B})$.

Observemos que si $\mathcal{A} = \mathcal{A}_n$ entonces $\widehat{\boldsymbol{\beta}}_{n,j} = 0$ para $j \notin \mathcal{A}$ de donde

$$\mathbb{P}(\mathcal{A} = \mathcal{A}_n) \le \mathbb{P}(\sqrt{n}\,\widehat{\boldsymbol{\beta}}_{n,j} = 0; \forall j \notin \mathcal{A})\,.$$

Por otra parte, como $\mathcal{A} = \{1, \ldots, k\}$, tenemos que $\sqrt{n}\widehat{\boldsymbol{\beta}}_{n,II} \xrightarrow{D} \mathbf{z}^\star_{II}$. De las propiedades de convergencia débil y como $\{\mathbf{0}_{p-k}\}$ es un cerrado, deducimos que

$$\limsup_{n\to\infty} \mathbb{P}(\mathcal{A} = \mathcal{A}_n) \le \limsup_{n\to\infty} \mathbb{P}(\sqrt{n}\,\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}) \le \mathbb{P}(\mathbf{z}^\star_{II} = \mathbf{0}_{p-k})\,,$$

con lo cual, bastará mostrar que $\mathbb{P}(\mathbf{z}^\star_{II} = \mathbf{0}_{p-k}) < 1$. Estudiaremos los casos $b = 0$ y $b > 0$.

Si $b = 0$, $R(\mathbf{z}) = \mathbf{z}^{\mathrm{T}}\mathbf{w} + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}$ de donde $\mathbf{z}^\star = -\mathbf{A}^{-1}\mathbf{w} \sim N(\mathbf{0}, \mathbf{A}^{-1}\,\mathbf{B}\,\mathbf{A}^{-1})$, con lo cual $\mathbb{P}(\mathbf{z}^\star_{II} = \mathbf{0}_{p-k}) = 0$.

Si $b > 0$, $R(\mathbf{z})$ no es derivable respecto de z_ℓ si $z_\ell = 0$, $\ell = k+1, \ldots, p$. Por las KKT condiciones de optimalidad (ver Bülmann y van de Geer, 2011) deducimos que

$$\begin{aligned} w_\ell + (\mathbf{A}\mathbf{z}^\star)_\ell + b\,\alpha_\ell = 0 \qquad & \ell = 1, \ldots k \\ |w_\ell + (\mathbf{A}\mathbf{z}^\star)_\ell| \le \frac{b}{\|\boldsymbol{\beta}_0\|_2} \qquad & \ell \notin \mathcal{A}\,, \end{aligned}$$

es decir, si $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}}$,

$$\mathbf{w}_I + \mathbf{A}_{11}\mathbf{z}^\star_I + \mathbf{A}_{12}\mathbf{z}^\star_{II} + b\,\boldsymbol{\alpha}_I = 0 \qquad \text{y} \qquad |\mathbf{w}_{II} + \mathbf{A}_{21}\mathbf{z}^\star_I + \mathbf{A}_{22}\mathbf{z}^\star_{II}| \le \frac{b}{\|\boldsymbol{\beta}_0\|_2}\,,$$

donde la desigualdad se entiende coordenada a coordenada. Por lo tanto, si $\mathbf{z}^\star_{II} = \mathbf{0}_{p-k}$ obtenemos que $\mathbf{w}_I + \mathbf{A}_{11}\mathbf{z}^\star_I + b\,\boldsymbol{\alpha}_I = 0$ y $|\mathbf{w}_{II} + \mathbf{A}_{21}\mathbf{z}^\star_I| \le b/\|\boldsymbol{\beta}_0\|_2$. Como $\mathbf{A}$ es definida positiva, deducimos que $\mathbf{z}^\star_I = -\mathbf{A}_{11}^{-1}(\mathbf{w}_I + b\,\boldsymbol{\alpha}_I)$. Por lo tanto, si $\mathbf{v} = \mathbf{w}_{II} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}(\mathbf{w}_I + b\,\boldsymbol{\alpha}_I) = (v_1, \ldots, v_{p-k})^{\mathrm{T}}$, tenemos que $\mathbf{v}$ tiene distribución normal multivariada y

$$\mathbb{P}(\mathbf{z}^\star_{II} = \mathbf{0}_{p-k}) \le \mathbb{P}\left(|v_\ell| \le \frac{b}{\|\boldsymbol{\beta}_0\|_2}, \quad \forall \ell = 1, \ldots, p-k\right) < 1\,,$$

lo que concluye la demostración. ■

Demostración del Teorema 5.10. Como en la demostración del Teorema 5.7 definamos $R_n(\mathbf{z}) = R_{n,1}(\mathbf{z}) + R_{n,2}(\mathbf{z})$, donde ahora

$$R_{n,1}(\mathbf{z}) = \frac{1}{n\ \lambda_n^2} \sum_{i=1}^{n} \left\{ \phi\left(Y_i, \mathbf{X}_i^{\mathrm{T}}\left[\boldsymbol{\beta}_0 + \lambda_n\ \mathbf{z}\right]\right) - \phi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\right\}\ w(\mathbf{X}_i)$$
$$R_{n,2}(\mathbf{z}) = \frac{1}{\lambda_n^2}\left\{I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \lambda_n\ \mathbf{z}\right) - I_{\lambda_n}(\boldsymbol{\beta}_0)\right\}\ ,$$

con $I_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2$. Observemos que $\operatorname{argmin}_{\mathbf{z}} R_n(\mathbf{z}) = (1/\lambda_n)\ (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$. Mostraremos primero que $R_{n,\ell}(\mathbf{z})$, $\ell = 1, 2$ es equicontinuo, es decir, dados $\epsilon > 0$, $\eta > 0$ y $M < \infty$ veremos que existe $\delta > 0$ tal que

$$\limsup_{n\to\infty} \mathbb{P}^* \left(\sup_{\substack{\max(\|\mathbf{u}\|_2, \|\mathbf{v}\|_2) \le M \\ \|\mathbf{u}-\mathbf{v}\|_2 < \delta}} |R_{n,\ell}(\mathbf{u}) - R_{n,\ell}(\mathbf{v})| > \epsilon \right) < \eta\ . \tag{5.24}$$

Como en el Lema 5.16, tenemos que **P1** implica que

$$I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \lambda_n\ \mathbf{u}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_0 + \lambda_n\ \mathbf{v}\right) \le \lambda_n K\ \left(\lambda_n\ \|\mathbf{u} - \mathbf{v}\|_1\right)$$

de donde $|R_{n,2}(\mathbf{u}) - R_{n,2}(\mathbf{v})| \le K\|\mathbf{u} - \mathbf{v}\|_1 \le\ K\ \sqrt{p}\,\|\mathbf{u} - \mathbf{v}\|_2$, lo que prueba (5.24) para $\ell = 2$.

Mostraremos ahora que $R_{n,1}(\mathbf{z})$ es equicontinuo. Como en la demostración del Lema 5.15, mediante un desarrollo de Taylor de primer orden de $R_{n,1}(\mathbf{u})$ y $R_{n,1}(\mathbf{v})$ alrededor de $\boldsymbol{\beta}_0$ obtenemos

$$R_{n,1}(\mathbf{u}) - R_{n,1}(\mathbf{v}) = \frac{1}{\lambda_n\ \sqrt{n}}\ \sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)^{\mathrm{T}}(\mathbf{u} - \mathbf{v}) + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}})\mathbf{u} - \frac{1}{2}\mathbf{v}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v},$$

donde $\widetilde{\boldsymbol{\beta}}_{\mathbf{u}}$ y $\widetilde{\boldsymbol{\beta}}_{\mathbf{v}}$ están definidas como

$$\widetilde{\boldsymbol{\beta}}_{\mathbf{v}} = \boldsymbol{\beta}_0 + \lambda_n\ \tau_{\mathbf{v},n}\mathbf{v} \qquad\qquad \widetilde{\boldsymbol{\beta}}_{\mathbf{u}} = \boldsymbol{\beta}_0 + \lambda_n\ \tau_{\mathbf{u},n}\mathbf{u}\,,$$

con $\tau_{\mathbf{v},n},\ \tau_{\mathbf{u},n} \in [0, 1]$. Como en la demostración del Lema 5.15, usando que $\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0) \xrightarrow{D} N_p(\mathbf{0}, \mathbf{B})$, deducimos que $\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)^{\mathrm{T}}(\mathbf{u} - \mathbf{v}) \le O_{\mathbb{P}}(1)\|\mathbf{u} - \mathbf{v}\|_2$. Por otra parte,

$$\begin{aligned}\mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}})\mathbf{u} - \mathbf{v}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v} &= \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}})\mathbf{u} - \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{u} + \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{u} \\ &- \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v} + \mathbf{u}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\mathbf{v}\,,\end{aligned}$$

luego, si $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 \le M$,

$$\begin{aligned}|R_{n,1}(\mathbf{u}) - R_{n,2}(\mathbf{v})| \ \le\ & \frac{1}{\lambda_n\ \sqrt{n}}\ O_p(1)\|\mathbf{u} - \mathbf{v}\|_2 + M^2\left\|\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}}) - \mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\right\| \\ & +2\|\mathbf{u} - \mathbf{v}\|_2 M\left\|\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}})\right\|\ .\end{aligned}$$

Usando que $\lambda_n\ \sqrt{n} \to \infty$ y que el Lema 5.14 implica que $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{u}}) - \mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}}) \xrightarrow{p} 0$ y $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{v}}) \xrightarrow{p} \mathbf{A}$ uniformemente sobre $\{\mathbf{u}, \mathbf{v} \in \mathbb{R}^p : \max\{\|\mathbf{u}\|, \|\mathbf{v}\|\} \le M\}$, se obtiene que (5.24) vale para $\ell = 1$.

Mostraremos ahora que dados $\mathbf{z}_1, \ldots, \mathbf{z}_s \in \mathbb{R}^p$, $(R_n(\mathbf{z}_1), \ldots, R_n(\mathbf{z}_s))^{\mathrm{T}} \xrightarrow{D} (R(\mathbf{z}_1), \ldots, R(\mathbf{z}_s))^{\mathrm{T}}$, donde $R(\mathbf{z}) = (1/2)\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$. Como en la demostración del Teorema 5.7, basta probarlo para $s = 1$. Fijemos $\mathbf{z} \in \mathbb{R}^p$. Nuevamente un desarrollo de Taylor de primer orden, tenemos que

$$R_{n,1}(\mathbf{z}) = \frac{1}{\lambda_n\ \sqrt{n}}\ \sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)^{\mathrm{T}}(\mathbf{u} - \mathbf{v}) + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{z}})\mathbf{z}\,,$$

donde $L_n(\boldsymbol{\beta})$ y $\mathbf{A}_n(\boldsymbol{\beta})$ están definidos en (3.10) y (5.12) respectivamente y $\widetilde{\boldsymbol{\beta}}_{\mathbf{z}} = \boldsymbol{\beta}_0 + \lambda_n\, \tau_n \mathbf{z}$, con $\tau_n \in [0,1]$ es un punto intermedio. La Fisher–consistencia dada en (3.13) y el Teorema Central del Límite Multivariado implican que

$$\sqrt{n}\nabla L_n(\boldsymbol{\beta}_0) \xrightarrow{D} N_p(\mathbf{0}, \mathbf{B})\,,$$

donde la matriz $\mathbf{B}$ definida en (3.18). Por otra parte, $\lambda_n\, \sqrt{n} \to \infty$, por lo tanto,

$$\frac{1}{\lambda_n\, \sqrt{n}}\, \sqrt{n}\nabla L_n(\boldsymbol{\beta}_0)^{\mathrm{T}}(\mathbf{u} - \mathbf{v}) \xrightarrow{p} 0\,.$$

Finalmente, por el Lema 5.14, se tiene que $\mathbf{A}_n(\widetilde{\boldsymbol{\beta}}_{\mathbf{z}}) \xrightarrow{p} \mathbf{A}$, luego $R_{n,1} \xrightarrow{p} (1/2)\mathbf{z}^{\mathrm{T}}\mathbf{A}\mathbf{z}$.

Por otra parte, $R_{n,2}(\mathbf{z}) \xrightarrow{p} \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$. Como en la demostración del Teorema 5.7 si $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2$, se tiene que $I_\lambda(\boldsymbol{\beta}) = \lambda\, J(\boldsymbol{\beta})$

$$J(\boldsymbol{\beta}) = \sum_{\ell=1}^{p} \frac{|\beta_\ell|}{\|\boldsymbol{\beta}\|_2} = \sum_{\ell=1}^{p} J_\ell(\boldsymbol{\beta})$$

donde J_ℓ es diferenciable en todo punto salvo en el hiperplano $\{\boldsymbol{\beta} : \beta_\ell = 0\}$. Recordemos que

$$R_{n,2}(\mathbf{z}) = \frac{1}{\lambda_n}\left(J\left(\boldsymbol{\beta}_0 + \lambda_n\, \mathbf{u}\right) - J(\boldsymbol{\beta}_0)\right)\,.$$

Supongamos que $\beta_{0,\ell} \neq 0$. Como en la demostración del Teorema 5.7, para n suficientemente grande, $\beta_{0,\ell} + \lambda_n\, z_\ell$ tiene el mismo signo que $\beta_{0,\ell}$. Por lo tanto, usando el Teorema de Valor Medio, obtenemos que

$$J_\ell\left(\boldsymbol{\beta}_0 + \lambda_n\, \mathbf{z}\right) - J_\ell(\boldsymbol{\beta}_0) = \left[\nabla J_\ell\left(\boldsymbol{\beta}_0 + \alpha_{n,\ell}\, \lambda_n\, \mathbf{z}\right)\right]^{\mathrm{T}} \lambda_n\, \mathbf{z}\,,$$

con $\alpha_{n,\ell} \in [0,1]$. Cuando $\beta_{0,\ell} = 0$, $J_\ell(\boldsymbol{\beta}_0) = 0$, entonces

$$J_\ell\left(\boldsymbol{\beta}_0 + \lambda_n\, \mathbf{z}\right) - J_\ell(\boldsymbol{\beta}_0) = \frac{\lambda_n\, |z_\ell|}{\left\|\boldsymbol{\beta}_0 + \lambda_n\, \mathbf{z}\right\|_2}\,.$$

Luego, $R_{n,2}(\mathbf{z})$ puede ser escrito como

$$R_{n,2}(\mathbf{z}) \;=\; \left\{\sum_{\ell=1}^{p} \left[\nabla J_\ell\left(\boldsymbol{\beta}_0 + \alpha_{n,\ell}\lambda_n\, \mathbf{z}\right)\right]^{\mathrm{T}} \mathbf{z}\, \mathbb{I}_{\{\beta_{0.\ell}\neq 0\}} + \sum_{\ell=1}^{p} |z_\ell| \frac{1}{\left\|\boldsymbol{\beta}_0 + \lambda_n\, \mathbf{z}\right\|_2}\, \mathbb{I}_{\{\beta_{0.\ell}=0\}}\right\}$$

lo cual implica que $R_{n,2}(\mathbf{z}) \xrightarrow{p} \mathbf{z}^{\mathrm{T}}\mathbf{q}(\mathbf{z})$. Por lo tanto, el proceso $R_n \xrightarrow{p} R$ y usando el Teorema 2.7 de Kim y Pollard (1990), deducimos que $(1/\lambda_n)\,(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = \operatorname{argmin}_{\mathbf{z}} R_n(\mathbf{z}) \xrightarrow{p} \operatorname{argmin}_{\mathbf{z}} R(\mathbf{z})$.

Demostración del Teorema 5.11. Sea $\gamma = \min\{|\beta_{0,j}| : 1 \leq j \leq k\}/2$. Consideremos el evento

$$\mathcal{E}_n = \{\widehat{\boldsymbol{\beta}}_{n,II} = \mathbf{0}_{p-k}\ \wedge \|\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I}\|_2 \leq \gamma\}.$$

Usando que $\mathbb{P}(\mathcal{E}_n) \to 1$ y la definición de $\widehat{\boldsymbol{\beta}}_n$, tenemos que

$$\mathbf{0}_k = \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{n,I})\, w(\mathbf{X}_i)\mathbf{X}_{i,I} + \nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,I}) + \mathbf{r}_n\,,$$

donde $\mathbf{X}_{i,I} \in \mathbb{R}^k$ es el subvector de $\mathbf{X}_i$ correspondiente a las k coordenadas activas y $\mathbb{P}(\mathbf{r}_n = \mathbf{0}_k) \to 1$.

Sea $\mathbf{a} \in \mathbb{R}^k$ fijo. El Teorema de Valor Medio implica que

$$\mathbf{a}^{\mathrm{T}}\mathbf{0}_k = \mathbf{a}^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_{0,I})\, w(\mathbf{X}_i)\mathbf{X}_{i,I} + \mathbf{a}^{\mathrm{T}}\widetilde{\mathbf{A}}_n(\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I}) + \mathbf{a}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,I}) + \mathbf{a}^{\mathrm{T}}\mathbf{r}_n\,,$$

donde $\widetilde{\mathbf{A}}_n = (1/n)\sum_{i=1}^n \chi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}_{n,I})\, w(\mathbf{X}_i)\mathbf{X}_{i,I}\mathbf{X}_{i,I}^{\mathrm{T}}$ con $\widetilde{\boldsymbol{\beta}}_{n,I} = \boldsymbol{\beta}_{0,I} + \tau_n(\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I})$ y $0 \leq \tau_n \leq 1$. Luego, tenemos que

$$\sqrt{n}\mathbf{a}^{\mathrm{T}}\widetilde{\mathbf{A}}_n(\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I}) = -\,\mathbf{a}^{\mathrm{T}}\frac{1}{\sqrt{n}}\sum_{i=1}^n \Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_{0,I})\, w(\mathbf{X}_i)\mathbf{X}_{i,I} - \mathbf{a}^{\mathrm{T}}\sqrt{n}\,\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,I}) - \mathbf{a}^{\mathrm{T}}\sqrt{n}\mathbf{r}_n\,.$$

La condición (5.6) implica que $\sqrt{n}\,\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,I}) \xrightarrow{p} 0$. Usando que $\mathbb{P}(\mathbf{r}_n = \mathbf{0}_k) \to 1$ y el Teorema Central del Límite, deducimos que $\sqrt{n}\,\mathbf{a}^{\mathrm{T}}\widetilde{\mathbf{A}}_n\,(\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{D} N(0, \mathbf{a}^{\mathrm{T}}\widetilde{\mathbf{B}}\mathbf{a})$, de donde se deduce que $\mathbf{c}_n = \sqrt{n}\widetilde{\mathbf{A}}_n(\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{D} N(\mathbf{0}, \widetilde{\mathbf{B}})$.

El Lema 5.14 y el hecho de que $\widehat{\boldsymbol{\beta}}_{n,I} \xrightarrow{p} \boldsymbol{\beta}_{0,I}$, implican que $\widetilde{\mathbf{A}}_n^{-1} \xrightarrow{p} \widetilde{\mathbf{A}}^{-1}$. Por lo tanto, usando que $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{n,I} - \boldsymbol{\beta}_{0,I}) = \widetilde{\mathbf{A}}_n^{-1}\mathbf{c}_n$, el resultado buscado es consecuencia del Teorema de Slutsky. ■

Capítulo 6

Resultados asintóticos para $p \to \infty$

El objetivo de este capítulo es probar resultados similares a los del Capítulo 5 para el caso en que p tienda a infinito junto con el tamaño de muestra n. Más específicamente, bajo este nuevo contexto, se prueban resultados de consistencia, tasa de convergencia y distribución asintótica del $M-$estimador penalizado, para ciertas elecciones de la función de penalización. También se presenta un resultado sobre sus propiedades de selección de variables. Por simplicidad, supondremos que $w(\mathbf{x}) \equiv 1$, para todo $\mathbf{x}$ y que $\{\lambda_n\}_{n\in\mathbb{N}}$ es una sucesión determinística. Resultados análogos podrían obtenerse cuando se considera el estimador pesado, adaptando adecuadamente los supuestos **Z1** a **Z7**.

Al estar estudiando la situación en la que la dimensión p de la covariables tiende a infinito junto con n, debemos considerar una sucesión de modelos. Más precisamente, consideraremos un arreglo triangular de variables Bernoulli independientes, $\{Y_{n,i} : 1 \leq i \leq n,\, n \geq 1\}$ y uno de variables aleatorias explicativas $\{\mathbf{X}_{n,i} : 1 \leq i \leq n,\, n \geq 1\}$ con $\mathbf{X}_{n,i} \in \mathbb{R}^{p_n}$, tales que:

$$\pi_{0,n,i} := \mathbb{P}(Y_{n,i} = 1|\mathbf{X}_{n,i}) = F(\mathbf{X}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}_{0,n}) = \frac{\exp(\mathbf{X}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}_{0,n})}{1+\exp(\mathbf{X}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}_{0,n})},$$

donde $\{\boldsymbol{\beta}_{0,n} : n \geq 1\}$ es la sucesión de vectores de coeficientes de regresión verdaderos. Supondremos que para cada n fijo, $(Y_{n,i}, \mathbf{X}_{n,i})$, $1 \leq i \leq n$, son independientes e igualmente distribuidas.

Para alivianar la notación, omitiremos el índice n en las cantidades ya descriptas arriba, por lo que $(Y_i, \mathbf{X}_i)$ con $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ indicará la $i-$ésima observacion de ese arreglo triangular.

Al trabajar con pérdidas y penalizaciones no convexas, consideraremos también el siguiente estimador restringido

$$\widehat{\boldsymbol{\beta}}_{n,\,R} = \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}) + I_\lambda(\boldsymbol{\beta}), \tag{6.1}$$

donde $R > 0$ es una constante fija y ϕ es la función dada en (3.7). Este tipo de restricciones es común al lidiar con funciones no convexas, ver por ejemplo Loh (2017) y el supuesto 3 (ii) en Elsener y van de Geer (2018). La principal razón de considerar un estimador restringido, es que, al restringir el problema de minimización a un compacto, las propiedades de consistencia son más fáciles de obtener. A pesar de ello, en esta tesis también daremos un resultado de consistencia para el estimador sin restringir definido en (3.19).

Consideraremos las siguientes hipótesis adicionales respecto de la elección de la función de pérdida

R4 $\psi(t) \geq 0$ y existe un $c \geq \log 2$ y $\tau > 0$ tal que $\psi(t) > \tau$ para todo $0 < t < c$.

R5 ρ es tres veces diferenciable, con derivadas ψ, ψ' y ψ'' acotadas.

Observación 6.1. Observemos que **R5** implica que la función $\phi(y,t)$ definida en (3.7) es tres veces derivable en su segunda coordenada y tanto ϕ como sus derivadas, son acotadas para $y \in \{0,1\}$. Por otra parte, si $\psi(0) \neq 0$ y se cumplen **R1** y **R2** para una constante $c > \log(2)$, como es el caso de la función introducida por Croux y Haesbroeck (2003), entonces vale la condición **R4**. ♣

Al igual que en el capítulo anterior, supondremos sin perder generalidad que $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,I}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ donde $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$ tiene todas sus coordenadas distintas de cero. Vale la pena destacar que la cantidad k de componentes no nulas puede depender de n, creciendo eventualmente con el tamaño de muestra. Sin embargo, para obtener resultados de distribución asintótica deberemos hacer algunos supuestos sobre los coeficientes $\boldsymbol{\beta}_{0,I}$, más precisamente sobre el comportamiento de

$$m_{0,n} = \min\{|\beta_{0,j}| : \beta_{0,j} \neq 0\}. \tag{6.2}$$

Como se menciona en Bühlmann y van de Geer (2011), cuanto mayor sea $m_{0,n}$, más fácil será seleccionar variables. Consideraremos también las siguientes hipótesis respecto de la distribución de las covariables $\mathbf{X}$. Dado un vector $\mathbf{v} \in \mathbb{R}^p$, usaremos la notación $\mathbf{v} = (\mathbf{v}_I, \mathbf{v}_{II})$ con $\mathbf{v}_I \in \mathbb{R}^k$ y $\mathbf{v}_{II} \in \mathbb{R}^{p-k}$.

Sean $\mathbf{X}_1, \ldots, \mathbf{X}_n$ las variables aleatorias explicativas correspondientes a la $n-$ésima fila del arreglo triangular y sea $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \sim \mathbf{X}_i$. Dada una matriz simétrica y semidefinida positiva $\mathbf{C} \in \mathbb{R}^{p \times p}$, definimos como $\iota_1(\mathbf{C})$ al mínimo autovalor de $\mathbf{C}$ y $\iota_p(\mathbf{C})$ al máximo autovalor de $\mathbf{C}$.

Z1 Sea X_{ij} es la $j-$ésima coordenada de $\mathbf{X}_i$, entonces

$$\mathbb{E}\left(\max_{1 \leq j \leq p} \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2\right) = O(1).$$

Z2 Existe una constante $K_1 > 0$ independiente de n tal que $\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]) \leq K_1$.

Z3 Existe una constante $\tau_1 > 0$ independiente de n tal que $\iota_1(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]) \geq \tau_1$.

Z4 Existe una constante $K_2 > 0$ independiente de n tal que $\boldsymbol{\beta}_0^{\mathrm{T}} \mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}] \boldsymbol{\beta}_0 \leq K_2^2$.

Z5 $\mathbf{X}$ tiene una distribución elíptica centrada con función característica

$$\phi_{\mathbf{X}}(\mathbf{t}) = \xi(\mathbf{t}^{\mathrm{T}} \boldsymbol{\Gamma} \mathbf{t})$$

para alguna función ξ que no depende de n y algún $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times p}$.

Z6 Existe una constante $K_3 > 0$ independiente de n tal que $\mathbb{E} \|\mathbf{X}_I\|_2^6 \leq K_3$.

Z7 Existe una constante $\tau_2 > 0$ tal que $\iota_1(\mathbf{B}_I) \geq \tau_2$ donde

$$\mathbf{B}_I = \mathbb{E}[\Psi^2(Y, \mathbf{X}_I^{\mathrm{T}} \boldsymbol{\beta}_{0,I}) \mathbf{X}_I \mathbf{X}_I^{\mathrm{T}}].$$

Observación 6.2. La hipótesis **Z1** es necesaria para obtener tasas de convergencia con orden $(p \log p/n)^{1/2}$ sin pedir condiciones de acotación superior sobre los autovalores de $\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]$. Dicho supuesto se cumple si $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_p)$ y $a_n = \log p\,/n \to 0$. Efectivamente, sea $V_n \sim \chi^2_n$, luego

$$\mathbb{E}\left(\max_{1\le j\le p} \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right) = \frac{1}{n}\,\mathbb{E}\left(\max_{1\le j\le p} V_n\right).$$

Por lo tanto, la desigualdad dada en Dasarathy (2011) permite obtener la cota

$$\mathbb{E}\left(\max_{1\le j\le p} \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right) \le \frac{a_n}{1-\exp(-2\,a_n)}\,.$$

Usando que $1-x \ge \exp(-2\,x)$, para $0 < x \le 1/2$, deducimos que **Z1** se verifica si $a_n = \log p\,/n \to 0$.

Por otra parte, **Z3**, **Z4** y **Z5** son necesarias para probar la consitencia del estimador sin restricciones definido en (3.19). En particular, **Z5** se cumple si $\mathbf{X}_{n,1} \sim S\mathbf{Z}$ donde S y $\mathbf{Z}$ son independientes, $\mathbb{P}(S > 0) = 1$ y $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{\Gamma})$. Esta clase de mezcla de normales incluye la normal contaminada, la distribución $\mathcal{T}_m$ de Student multivariada con m grados de libertad, además de la distribución normal.

Vale la pena mencionar que **Z3** y **Z4** implican que $\tau_1\|\boldsymbol{\beta}_0\|^2 \le \textsc{Var}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) \le K_2^2$. Por lo tanto, como $\|\boldsymbol{\beta}_0\|^2 = \|\boldsymbol{\beta}_{0,I}\|^2$ obtenemos que $\sum_{j=1}^{k}\beta_{0,j}^2 \le K_2^2/\tau_1$ para todo n aunque k crezca con el tamaño de muestra. En particular, $\max\{|\beta_{0,j}| : \beta_{0,j} \ne 0\}$ está acotado y $m_{0,n} = O(1/\sqrt{k})$, donde $m_{0,n} = \min\{|\beta_{0,j}| : \beta_{0,j} \ne 0\}$ fue definido en (6.2). Por lo tanto, si $k = k_n \to \infty$, el cumplimiento simultáneo de **Z3** y **Z4** implica que $m_{0,n} \to 0$.

La hipótesis **Z2** se utiliza para probar la tasa de convergencia $\sqrt{n/p}$. Finalmente, las hipótesis **Z6** y **Z7** son necesarias para probar la normalidad asintótica de nuestro estimador al usar las penalizaciones SCAD o MCP. ♣

Sea $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^{\mathrm{T}}$. Para obtener la distribución asintótica de los estimadores de $\boldsymbol{\beta}_{0,I}$ y la propiedad oráculo de los estimadores penalizados, consideraremos las siguientes hipótesis respecto al crecimiento de n, k, λ_n y los coeficientes de $\boldsymbol{\beta}_{0,I}$.

N1 $m_{0,n}\,\sqrt{n/k} \to \infty$, donde $m_{0,n} = \min\{|\beta_{0,j}| : \beta_{0,j} \ne 0\}$.

N2 $m_{0,n}/\lambda_n \to \infty$.

N3 $k/n = O(\lambda_n^2)$.

Observación 6.3. Vale la pena destacar que, si k y y $\boldsymbol{\beta}_{0,I}$ son fijos, **N1** se cumple, mientras que **N2** es equivalente a $\lambda_n \to 0$. Por otra parte, si existe $m_0 > 0$ independiente de n tal que $m_{0,n} > m_0$, entonces $k/n \to 0$ y $\lambda_n \to 0$ implican **N1** y **N2**, respectivamente. Si además $m_{0,n}$ está acotado superiormente, las condiciones anteriores son equivalentes. Finalmente, si $m_{0,n} = O(1/\sqrt{k})$, como cuando se cumplen **Z3** y **Z4**, entonces **N1** y **N2** implican $k^2/n \to 0$ y $k\,\lambda_n^2 \to 0$. Estas dos últimas condiciones coinciden si $\lambda_n = O(1/\sqrt{n})$, de lo contrario da una relación entre el parámetro de penalización y el crecimiento de la cantidad de coordenadas no nulas. ♣

Como en el Capítulo 5, relegamos a los apéndices de este capítulo todas las demostraciones de los resultados que se presentan.

6.1. Consistencia

Como en los capítulos anteriores, definamos $\mathbb{L}(\boldsymbol{\beta}) = \mathbb{E}\phi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$ y $L_n(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n}\phi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$. Cabe mencionar que a diferencia de lo que ocurre en el Capítulo 5, la función $\mathbb{L}(\cdot)$ depende de n ya que tanto la distribución de $\mathbf{X}$ como su dimensión y la de $\boldsymbol{\beta}$ dependen de $p = p_n$. Sin embargo, para evitar notación engorrosa no reforzaremos la dependencia en n salvo que sea necesario. Por otra parte, para dar una medida de cercanía basada en predicciones definamos

$$d_n^2(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = \mathbb{E}[F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)]^2\,, \tag{6.3}$$

donde $\mathbf{X} \sim \mathbf{X}_{n,1}$ y para reforzar la dependencia en n utilizamos el subíndice n en la distancia. Observemos que $d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0)$ corresponde a evaluar (6.3) en $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$, por lo tanto la muestra es fija y podemos escribir

$$d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) = \mathbb{E}\left\{[F(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)]^2 \middle| (Y_1, \mathbf{X}_1), \dots (Y_n, \mathbf{X}_n)\right\}\,.$$

El siguiente resultado muestra que el estimador propuesto en (3.19) da origen a predicciones consistentes. Más aún, el estimador resulta consistente cuando las covariables satisfacen **Z5**, en particular, si tienen distribución normal.

Teorema 6.1. *Sea $\mathbf{X} \sim \mathbf{X}_{n,1}$ y supongamos que se cumplen las hipótesis **R1** y **R4**. Entonces,*

(a) Si $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19), tenemos que

$$d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}} + I_{\lambda_n}(\boldsymbol{\beta}_0)\right)\,.$$

(b) Supongamos además que $\|\boldsymbol{\beta}_0\|_1 \leq R$ y que existe una constante M tal que $P(\|\mathbf{X}\|_\infty \leq M) = 1$, para todo $n \geq 1$. Si $\iota_1(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]) > 0$, entonces

$$\|\widehat{\boldsymbol{\beta}}_{n,\,R} - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\left(\sqrt{\frac{p}{n}} + I_{\lambda_n}(\boldsymbol{\beta}_0)\right)^{1/2}\left(\frac{1}{\iota_1(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}])}\right)^{1/2}\right).$$

*donde $\widehat{\boldsymbol{\beta}}_{n,\,R}$ es el estimador restringido definido en (6.1). Si además se cumple **Z3** entonces*

$$\|\widehat{\boldsymbol{\beta}}_{n,\,R} - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\left(\sqrt{\frac{p}{n}} + I_{\lambda_n}(\boldsymbol{\beta}_0)\right)^{1/2}\right).$$

*(c) Si además se cumplen las hipótesis **Z3**, **Z4** y **Z5**, $p/n \to 0$ y $I_{\lambda_n}(\boldsymbol{\beta}_0) \to 0$ entonces, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow{p} 0$, donde $\widehat{\boldsymbol{\beta}}_n$ es el estimador definido en (3.19).*

Observación 6.4. En el item (a), probamos que $F(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n)$ es consistente en la norma L_2 si $p/n \to 0$ e $I_{\lambda_n}(\boldsymbol{\beta}_0) \to 0$. La prueba se basa en el hecho de que la familia de funciones $\mathcal{F}$ definida en (5.3) es VC-subgrafo y el Teorema 4.5. A diferencia de lo que ocurre en el modelo de regresión lineal, este hecho no implica necesariamente que $\widehat{\boldsymbol{\beta}}_n$ sea consistente, pues la función de vínculo F es acotada. Por ello, es necesario asumir hipótesis adicionales, como el supuesto **Z3** que también es necesario en el caso de regresión lineal. El resultado del item (a) también vale para el estimador restringido

definido en (6.1) y por lo tanto, en el item (b), probamos que la implicación antes mencionada es cierta si consideramos el estimador restringido (6.1), las covariables son acotadas en norma $\|\cdot\|_\infty$ y $\iota_1(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}])$ se mantiene separado del 0, obteniendo un resultado preliminar de tasa que será mejorado en la Sección 6.2. Finalmente, en el item (c) mostramos que el resultado del item (a) garantiza la consistencia del estimador no restringido definido en (3.19) cuando la distribución de $\mathbf{X}$ tiene una distribución elíptica centrada cuya generadora de la característica no depende de n, lo que incluye el caso particular de mezcla de normales como se describió en la Observación 6.2. ♣

6.2. Tasa de convergencia

El siguiente teorema da tasas de convergencia para nuestros estimadores. Su prueba usa cotas para los incrementos de procesos empíricos dados en Bühlmann y van de Geer (2011) y el Teorema 3.2.5 de Van der Vaart y Wellner (1996), que utiliza una técnica conocida como "dispositivo de pelado" (peeling device).

Para obtener una tasa para ciertas penalidades como la SCAD y MCP bajo condiciones más débiles sobre λ necesitaremos el siguiente supuesto

P2 Existe un $\widetilde{\delta} > 0$ y sucesiones $\{a_n\}_{n\in\mathbb{N}}$ y $\{b_n\}_{n\in\mathbb{N}}$, $a_n \geq 0$, $b_n \geq 0$, tales que si $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \widetilde{\delta}$, entonces

$$I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) \geq -a_n\sqrt{k}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 - b_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 . \tag{6.4}$$

Observación 6.5. Cabe mencionar que las penalizaciones LASSO, SCAD o MCP verifican **P2**. En efecto, para la penalidad LASSO $a_n = \lambda_n$ y $b_n = 0$. Para las penalidades SCAD o MCP, podemos escribir $I_{\lambda_n}(\boldsymbol{\beta}) = \sum_{j=1}^p J_{\lambda_n}(|\beta_j|)$ donde $J_{\lambda_n}(\cdot)$ es no negativa, dos veces diferenciable en $(0,\infty)$, $J'_{\lambda_n}(|\beta_{0,\ell}|) \geq 0$ y $J_{\lambda_n}(0) = 0$. Dado $\delta_0 > 0$, definamos

$$\begin{aligned} a_n &= \text{máx}\left\{J'_{\lambda_n}(|\beta_{0,\ell}|) : 1 \leq \ell \leq p \text{ y } \beta_{0,\ell} \neq 0\right\} = \text{máx}\left\{J'_{\lambda_n}(|\beta_{0,\ell}|) : 1 \leq \ell \leq k\right\} \\ b_n &= b_n(\delta_0) = \sup\{|J''_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta_0)| : \tau \in [-1,1] \,,\ 1 \leq \ell \leq p \text{ y } \beta_{0,\ell} \neq 0\} \\ &= \sup\{|J''_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta_0)| : \tau \in [-1,1] \,,\ 1 \leq \ell \leq k\} . \end{aligned}$$

A partir de los argumentos utilizados la demostración del Teorema 5.4(b), obtenemos que

$$I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) \;\geq\; -a_n\sum_{\ell=1}^k |\beta_\ell - \beta_{0,\ell}| - \frac{1}{2}\sum_{\ell=1}^k |J''_{\lambda_n}(\xi_\ell)|(\beta_\ell - \beta_{0,\ell})^2 , \tag{6.5}$$

donde ξ_ℓ pertenece al intervalo comprendido entre $|\beta_\ell|$ y $|\beta_{0,\ell}|$. Por lo tanto, si $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \widetilde{\delta} < \delta_0$, $\text{máx}(0, |\beta_{0,\ell}| - \widetilde{\delta}) < \xi_{n,\ell} \leq |\beta_{0,\ell}| + \widetilde{\delta}$, de donde usando (6.5) se deduce que

$$\begin{aligned} I_{\lambda_n}(\beta) - I_{\lambda_n}(\beta_0) &\geq -a_n\sum_{\ell=1}^k |\beta_{n,\ell} - \beta_{0,\ell}| - b_n\sum_{\ell=1}^k (\beta_{n,\ell} - \beta_{0,\ell})^2 \\ &\geq -a_n\sqrt{k}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 - b_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 , \end{aligned}$$

es decir, que vale la desigualdad (6.4).

Si consideramos las penalizaciones SCAD o MCP, $J'_{\lambda_n}(t)$ y $J''_{\lambda_n}(t)$ son nulas si $t > a\,\lambda_n$ donde a es el segundo parámetro de ajuste, que suponemos fijo, en la definición de estas penalizaciones. Por lo tanto, si $m_{0,n} > a\lambda_n$ para $n \geq n_0$ donde $m_{0,n}$ está definido en (6.2), tenemos que $a_n = 0$ y $b_n = 0$ para n suficientemente grande. En particular esto se cumple si existe $m_0 > 0$ independiente de n tal que $m_{0,n} > m_0$ y $\lambda_n \to 0$ o si vale **N2**. Observemos que si valen **Z3** y **Z4** como $m_{0,n} = O(1/\sqrt{k})$, existe M tal que $\sqrt{k}\; m_{0,n} \leq M$ para todo n, luego la condición $m_{0,n} > a\lambda_n$ para $n \geq n_0$ implica que $\lambda_n = O(1/\sqrt{k})$. ♣

Teorema 6.2. *Supongamos que se cumple la hipótesis **R1** y que existen constantes $\eta > 0$ y $\tau > 0$ tal que si $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \eta$, entonces $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$, para todo $n \geq 1$. Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) o en (6.1) y supongamos que $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow{p} 0$. Luego,*

*(a) Si se cumplen las hipótesis **P1** y **Z1** y $\lambda_n = O(\sqrt{\log p/n})$, entonces*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p \log p}{n}}\right). \tag{6.6}$$

*(b) Si se cumplen las hipótesis **P1** y **Z2** y $\lambda_n = O(\sqrt{1/n})$, luego*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}}\right). \tag{6.7}$$

*(c) Supongamos que se cumple **P2**, entonces*

*(i) bajo **Z1**, si $b_n \to 0$ y $a_n\sqrt{k} = O(\sqrt{p \log p/n})$, se obtiene la tasa de convergencia (6.6).*

*(ii) si **Z2** vale, $b_n \to 0$ y $a_n\sqrt{k} = O(\sqrt{p/n})$, la tasa de convergencia está dada por (6.7).*

Una condición importante para la validez del teorema anterior estipula que existen $\eta > 0$ y $\tau > 0$ tal que $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$ siempre que $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \eta$. El siguiente resultado da condiciones bajo las cuales dicho supuesto se verifica.

Lema 6.3. *Supongamos que se cumplen **Z2**, **Z3**, **Z4** y **Z5** y que la función ρ satisface **R1** y **R4**, entonces existen $\eta > 0$ y $\tau > 0$ tal que $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$ siempre que $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \eta$.*

Observación 6.6. Cabe mencionar que los items (a) y (b) del Teorema 6.2 pueden ser aplicados a las penalizaciones LASSO, Elastic Net, SCAD, MCP y Signo, ya que en estos casos se cumple la hipótesis **P1**. Sin embargo, como se comentó en la Observación 6.5 las penalizaciones LASSO, SCAD o MCP verifican **P2**. Por lo tanto, el item (c) del Teorema 6.2 permite obtener las mismas tasas que las dadas en los items (a) y (b), pero con hipótesis menos restrictivas para λ_n. En particular, para la penalidad LASSO para obtener la misma tasa de convergencia del estimador, el parámetro de regularidad debe cumplir $\lambda_n\sqrt{n\; k/p} = O(1)$ en lugar de $\lambda_n\sqrt{n} = O(1)$. Al igual que en el caso en el que p es fijo, esta diferencia juega un rol importante en las propiedades de selección de variables. ♣

6.3. Propiedades de selección de variables

El objetivo de esta sección es probar que, con probabilidad tendiendo a uno, nuestro estimador selecciona variables correctamente al utilizar penalizaciones que verifican una condición un poco

más restrictiva que la condición (5.4) del Teorema 5.5. Dicha condición se verifica trivialmente para la penalización LASSO y como veremos en el Corolario 6.5 también se cumple para las penalizaciones SCAD y MCP.

Teorema 6.4. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) o (6.1), donde $\phi(y,t)$ viene dada por (3.7) y la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3**. Sea $\{\ell_n\}_{n\in\mathbb{N}}$ tal que $\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ y definamos*

$$c_n = \frac{\sqrt{\iota_p(\mathbf{B})}}{\sqrt{n}} + \frac{\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}])}{\ell_n}, \tag{6.8}$$

donde $\mathbf{B}$ está definida en (3.18). Supongamos que $\lambda_n c_n^{-1} \to \infty$ y que para todo $C > 0$ existe una constante $K_C > 0$ y $N_C \in \mathbb{N}$ tal que si $n \geq N_C$,

$$I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \frac{\mathbf{u}_2}{\ell_n}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \mathbf{0}_{p-k}\right) \geq K_C \frac{\lambda_n}{\ell_n} \|\mathbf{u}_2\|_2, \tag{6.9}$$

para todos los vectores $\mathbf{u}_1 \in \mathbb{R}^k$ y $\mathbf{u}_2 \in \mathbb{R}^{p-k}$ que satisfacen $\|\mathbf{u}_1\|_2^2 + \|\mathbf{u}_2\|_2^2 \leq C^2$. Entonces, se tiene que $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \to 1$.

Observación 6.7. Vale la pena mencionar que si $\ell_n = \sqrt{n/p}$ y se cumple **Z2** y una condición análoga para $\iota_p(\mathbf{B})$, es decir, si existe una constante $K > 0$ independiente de n tal que $\max\{\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]), \iota_p(\mathbf{B})\} \leq K$, entonces $c_n = O(\sqrt{p/n})$. Por lo tanto, bajo estos supuestos, la condición $\lambda_n c_n^{-1} \to \infty$ se cumple si $\lambda_n \sqrt{n/p} \to \infty$. Recordemos que la tasa requerida a la penalidad en el Teorema 6.2(b) para tener estimadores con tasa de convergencia $\sqrt{p/n}$ es $\lambda_n = O(\sqrt{1/n})$, con lo cual $\lambda_n \sqrt{n/p} \to 0$. Asimismo, la tasa requerida a la penalidad LASSO en el Teorema 6.2(c) también es incompatible con $\lambda_n \sqrt{n/p} \to \infty$, lo que es consistente con los resultados obtenidos en el Capítulo 5. Por otra parte, si existe $K^\star > 0$ tal que $\min\{\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]), \iota_p(\mathbf{B})\} \geq K^\star$, las dos condiciones $\lambda_n c_n^{-1} \to \infty$ y $\lambda_n \sqrt{n/p} \to \infty$, son equivalentes, por lo que en ese caso, la tasa de convergencia del parámetro de penalización requerida en el Teorema 6.4 es análoga a la del Teorema 5.5. ♣

Como en el caso en que la dimensión era fija, el siguiente corolario muestra que el Teorema 6.4 puede ser aplicado a las penalizaciones SCAD y MCP.

Corolario 6.5. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) o (6.1), donde $\phi(y,t)$ viene dada por (3.7) y la función $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisface **R3**. Sea $\{\ell_n\}_{n\in\mathbb{N}}$ tal que $\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ y definamos c_n como en (6.8). Supongamos que $\lambda_n c_n^{-1} \to \infty$ y $\lambda_n \ell_n \to \infty$. Si $I_{\lambda_n}(\boldsymbol{\beta})$ es la penalización SCAD o MCP, entonces $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \to 1$.*

Observación 6.8. Cuando solamente se asume la hipótesis **P1**, las condiciones pedidas para λ_n en el Teorema 6.2(a) y (b) (necesarias para obtener tasas de convergencia) son incompatibles con la condición $\lambda_n c_n^{-1} \to \infty$. Sin embargo, al utilizar la penalización SCAD o MCP, las tasas de convergencia dadas en el Teorema 6.2 se obtienen pidiendo solamente $\lambda_n \to 0$ cuando $m_{0,n} = \min\{|\beta_{0,j}| : \beta_{0,j} \neq 0\} > m_0$, para todo n. Por otra parte, bajo **Z3** y **Z4** como $m_{0,n} = O(1/\sqrt{k})$, $\lambda_n = O(1/\sqrt{k})$ lo cual no contradice las tasas de convergencia de λ_n pedidas en el Corolario 6.5. En particular, si $\ell_n = \sqrt{n/p}$ y existen $K > 0$ y $K^\star > 0$ tales que $K^\star \leq \min\{\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]), \iota_p(\mathbf{B})\} \leq \max\{\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]), \iota_p(\mathbf{B})\} \leq K$, la condición $\lambda_n c_n^{-1} \to \infty$ es equivalente a $n/(k\,p) \to \infty$, cuando se cumplen **Z3** y **Z4**, mientras que si vale **N2**, la condición $\lambda_n c_n^{-1} \to \infty$ implica $m_{0,n}\sqrt{n/p} \to \infty$.

Vale la pena mencionar que si se cumple **Z3**, entonces la tasa $\lambda_n \ell_n \to \infty$ pedida en el Corolario 6.5 es una consecuencia de $\lambda_n c_n^{-1} \to \infty$. ♣

A partir de los resultados obtenidos en los Teoremas 6.2 y 6.4, podemos obtener el siguiente corolario que permite mejorar la tasa de convergencia de los estimadores definidos en (3.19) o en (6.1). En primer lugar, observemos que $I_\lambda(\cdot) : \mathbb{R}^p \to \mathbb{R}$ y por lo tanto, en todos los resultados anteriores las penalizaciones constituyen una sucesión de funciones de penalización, no sólo por su dependencia en λ_n sino también por su dominio. Sin embargo, teniendo esto en claro para evitar notación engorrosa, entenderemos que cuando escribimos $I_\lambda(\boldsymbol{\beta})$ para $\boldsymbol{\beta} \in \mathbb{R}^p$ o $I_\lambda(\mathbf{b})$ con $\mathbf{b} \in \mathbb{R}^k$ consideramos penalizaciones con distinto dominio. Para mayor claridad, utilizaremos el subíndice k para indicar vectores en $\mathbb{R}^k$. Para probar el Corolario 6.6, será necesario el siguiente supuesto para la función de penalización.

P3 Dado $\mathbf{b}_k \in \mathbb{R}^k$ tal que $\mathbf{b}_k \neq \mathbf{0}$, se cumple $I_\lambda(\mathbf{b}_k) = I_\lambda((\mathbf{b}_k^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}})$.

Corolario 6.6. *Sea $\widehat{\boldsymbol{\beta}}_n$ el estimador definido en (3.19) o en (6.1). Supongamos que $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \to 1$ cuando $n \to \infty$. Definamos*

$$\widehat{\mathbf{b}}_k = \underset{\mathbf{b}_k \in \mathbb{R}^k}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}} \mathbf{b}_k) + I_\lambda(\mathbf{b}_k)\,.$$

*Si se cumple la hipótesis **P3** y $\|\widehat{\mathbf{b}}_k - \boldsymbol{\beta}_{0,I}\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$ entonces, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$.*

Observación 6.9. En primer lugar, observemos que la hipótesis **P3** se cumple para las penalizaciones LASSO, Signo, SCAD y MCP. Más generalmente, se cumple para toda penalización que puede escribirse como $I_{\lambda_n}(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_{\lambda_n}(|\beta_j|)$ donde $J_{\lambda_n}(0) = 0$.

Por otra parte, el Teorema 6.2 da condiciones para que $\|\widehat{\mathbf{b}}_k - \boldsymbol{\beta}_{0,I}\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$. De hecho, para deducir $\|\widehat{\mathbf{b}}_k - \boldsymbol{\beta}_{0,I}\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$ alcanza que las hipótesis **Z1**-**Z5** se reemplacen por versiones análogas en donde solamente consideramos las primeras k coordenadas de $\mathbf{X}$ y $\boldsymbol{\beta}_0$. Cabe notar que, en particular, la hipótesis **Z4** ya da una condición sobre las primeras k componentes de $\mathbf{X}$ y $\boldsymbol{\beta}_0$ pues es equivalente a que $\boldsymbol{\beta}_{0,I}^{\mathrm{T}} \mathbb{E}[\mathbf{X}_I \mathbf{X}_I^{\mathrm{T}}] \boldsymbol{\beta}_{0,I} \leq K_2^2$. ♣

6.4. Distribución asintótica

En esta sección, estudiamos la distribución asintótica del estimador $\widehat{\boldsymbol{\beta}}_n$ definido en (3.19). Mostramos que, para algunas penalidades, el estimador tiene la propiedad oráculo, es decir, que el estimador de $\widehat{\boldsymbol{\beta}}_{n,\,I}$ tiene la misma distribución asintótica que la del estimador obtenido sabiendo de antemano que las últimas $p-k$ coordenadas son 0 y utilizando esa restricción en el modelo de regresión logística.

Para dejar clara la dependencia en el tamaño de muestra, cuando k crece con n o cuando la distribución de $\mathbf{X}_I^{\mathrm{T}} \boldsymbol{\beta}_{0,I}$ depende de n, dado un vector $\mathbf{b}_k \in \mathbb{R}^k$, indicaremos por $\mathbf{A}_I^{(k)}(\mathbf{b}_k) \in \mathbb{R}^{k \times k}$ y $\mathbf{B}_I^{(k)}(\mathbf{b}_k) \in \mathbb{R}^{k \times k}$ a las matrices

$$\mathbf{A}_I^{(k)}(\mathbf{b}_k) = \mathbb{E}\left[\chi(Y, \mathbf{X}_I^{\mathrm{T}} \mathbf{b}_k) \mathbf{X}_I \mathbf{X}_I^{\mathrm{T}}\right] \qquad \text{y} \qquad \mathbf{B}_I^{(k)}(\mathbf{b}_k) = \mathbb{E}\left[\Psi^2(Y, \mathbf{X}_I^{\mathrm{T}} \mathbf{b}_k) \mathbf{X}_I \mathbf{X}_I^{\mathrm{T}}\right]\,.$$

Definamos $\mathbf{A}_I^{(k)} = \mathbf{A}_I^{(k)}(\boldsymbol{\beta}_{0,I})$ y $\mathbf{B}_I^{(k)} = \mathbf{B}_I^{(k)}(\boldsymbol{\beta}_{0,I})$. Observemos que, en este caso, dado $\mathbf{v}_k \in \mathbb{R}^k$ con $\|\mathbf{v}_k\|_2 = 1$ la cantidad $t^2 = \mathbf{v}_k^{\mathrm{T}}\mathbf{B}_I^{(k)}\mathbf{v}_k$ también depende de n. Sin embargo, por simplicidad de notación en los enunciados del Teorema 6.7 y del Corolario 6.8, escribiremos t en lugar de t_n. Definamos además para $\mathbf{b}_k = (b_1, \ldots, b_k)^{\mathrm{T}} \in \mathbb{R}^k$ con $b_j \neq 0$, $1 \leq j \leq k$,

$$\nabla I_\lambda(\mathbf{b}_k) = \frac{\partial I_\lambda\left((\mathbf{b}_k^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}\right)}{\partial \mathbf{b}_k}\,.$$

Teorema 6.7. *Sea $\mathbf{v}_k \in \mathbb{R}^k$ tal que $\|\mathbf{v}_k\|_2 = 1$ y sea $t^2 = \mathbf{v}_k^{\mathrm{T}}\mathbf{B}_I^{(k)}\mathbf{v}_k$. Supongamos que $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \to 1$, cuando $n \to \infty$, y que $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$. Más aún, supongamos que se cumplen las hipótesis **N1**, **R5**, **Z6** y **Z7**. Luego, si $k^2/n \to 0$ y*

$$\sqrt{n}\|\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,I})\|_2 \xrightarrow{p} 0, \tag{6.10}$$

tenemos que $\sqrt{n}\; t^{-1}\mathbf{v}_k^{\mathrm{T}}\mathbf{A}_I^{(k)}(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{D} N(0,1)$.

Finalmente, el siguiente colorario muestra que el Teorema 6.7 puede ser aplicado a las penalizaciones SCAD y MCP.

Corolario 6.8. *Sea $\mathbf{v}_k \in \mathbb{R}^k$ donde $\|\mathbf{v}_k\|_2 = 1$ y $t^2 = \mathbf{v}_k^{\mathrm{T}}\mathbf{B}_I^{(k)}\mathbf{v}_k$. Supongamos que $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \to 1$ cuando $n \to \infty$ y que $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$. Más aún, supongamos que se cumplen las hipótesis **N1**, **N2**, **N3**, **R5**, **Z6** y **Z7**. Si $k^2/n \to 0$ y I_{λ_n} es la penalización SCAD o MCP, entonces*

$$\sqrt{n}\; t^{-1}\mathbf{v}_k^{\mathrm{T}}\mathbf{A}_I^{(k)}(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{D} N(0,1)\,.$$

Observación 6.10. Vale la pena mencionar que si en el Teorema 6.7 y el Corolario 6.8, pedimos la condición de tasa sobre los estimadores obtenida en la Sección 6.2, es decir, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{p/n})$, la conclusión de dichos resultados sigue valiendo si $p^2/n \to 0$ en lugar de $k^2/n \to 0$ y si reemplazamos las hipótesis **N1** y **N3** por $m_{0,n}\sqrt{n/p} \to \infty$ y $p/n = O(\lambda_n^2)$, respectivamente. ♣

6.5. Apéndice A: Demostraciones de la Sección 6.1

Para probar la consistencia de los estimadores propuestos cuando $p \to \infty$, usaremos el Lema 4.5 que corresponde al Teorema 2.14.1 de Van der Vaart y Wellner (1996). La importancia de ese lema radica en que, al trabajar en un contexto en donde la dimensión p diverge a infinito, no se pueden usar resultados "límite"como la Ley de los Grandes Números. En cambio, en este contexto son útiles las cotas para el proceso empírico con un n fijo, lo cual es justamente lo que provee el Lema 6.9. Aplicaremos ese lema a la familia de funciones definida en (5.3), es decir, a la clase de funciones $\mathcal{F} = \{f(y, \mathbf{x}) = \phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Recordemos que para una función $f : \mathbb{R}^{p+1} \to \mathbb{R}$, $P_n f = (1/n)\sum_{i=1}^n f(Y_i, \mathbf{X}_i)$ y $Pf = \mathbb{E}[f(Y, \mathbf{X})]$.

Lema 6.9. *Sea ϕ definida como en (3.7). Si se cumplen las condiciones **R1** y **R2**, entonces existe una constante C_1 independiente de n y de p que satisface*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |(P_n - P)(f)|\right] \leq C_1\sqrt{\frac{p}{n}}.$$

Demostración. Como ρ es acotada, existe una constante $C = \|\phi\|_\infty > 0$ tal que $|\phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})| \leq \|\phi\|_\infty$ para todo $y \in \{0, 1\}$, $\mathbf{x} \in \mathbb{R}^p$ y $\boldsymbol{\beta} \in \mathbb{R}^p$, es decir que $F = \|\phi\|_\infty$ es una envolvente de la clase de funciones $\mathcal{F} = \{f(y, \mathbf{x}) = \phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^p\}$. En el Lema 5.2 probamos que $\mathcal{F}$ es VC-subgrafo con índice $V(\mathcal{F}) \leq 2p + 4$.

Usando el Lema 4.4, deducimos que, para alguna constante universal K_1,

$$\begin{aligned} N(\varepsilon\|\phi\|_\infty, \mathcal{F}, \|\cdot\|_{2,\mathbb{Q}}) &\leq K_1\, V(\mathcal{F})(16e)^{V(\mathcal{F})}\left(\frac{1}{\varepsilon}\right)^{2(V(\mathcal{F})-1)} \\ &\leq K_1(2p+4)(16e)^{2p+4}\left(\frac{2}{\varepsilon}\right)^{4p+6}. \end{aligned} \tag{6.11}$$

Por lo tanto, usando el Lema 4.5 obtenemos que para alguna constante universal $M > 0$

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sqrt{n}(P_n - P)(f)\right|\right] \leq M\, \phi_\infty \sup_{\mathbb{Q}} \int_0^1 \sqrt{1 + \log N\left(\varepsilon\, \|\phi\|_\infty, \mathcal{F}, \|\cdot\|_{2,\mathbb{Q}}\right)} d\varepsilon\,,$$

donde el supremo se toma sobre todas las medidas de probabilidad discretas $\mathbb{Q}$. Utilizando (6.11), deducimos que

$$\sqrt{1 + \log N\left(\varepsilon\, \|\phi\|_\infty, \mathcal{F}, \|\cdot\|_{2,\mathbb{Q}}\right)} \;\leq\; \sqrt{1 + \log\left(K_1(2p+4)\right) + (2p+4)(\log(16)+1) + (4p+6)\log\left(\frac{2}{\varepsilon}\right)}.$$

Usando que $\log p \leq p$ para $p \geq 1$ obtenemos que para alguna constante $C > 0$, independiente de n y de p, $1 + \log\left(K_1(2p+4)\right) + (2p+4)(\log(16)+1) \leq C(2p+4) \leq C\,(4p+6) \leq 16\, C\, p$. Por lo tanto, como $4p + 6 \leq 16p$, si $C_1 = \max(C + \log(2), 1)$ y $C_2 = 4\, M\, \|\phi\|_\infty\, C_1$, tenemos que

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sqrt{n}(P_n - P)(f)\right|\right] \;\leq\; \sqrt{p}\, C_2 \int_0^1 \sqrt{1 + \log\left(\frac{1}{\varepsilon}\right)} d\varepsilon\,,$$

de donde concluimos el resultado usando que $\int_0^1 \sqrt{1 - \log(\varepsilon)}\, d\varepsilon < \infty$. ■

Lema 6.10. *Sea* $M(\pi, \pi_0)$ *definida como*

$$M(\pi, \pi_0) = \pi_0 \rho(-\log \pi) + (1 - \pi_0)\rho(-\log(1 - \pi)) + G(\pi) + G(1 - \pi)\,, \tag{6.12}$$

para $(\pi, \pi_0) \in (0, 1) \times [0, 1]$.

(a) Si se cumplen las condiciones ***R1*** *y* ***R2****, entonces la función* $M(\pi, \pi_0)$ *puede ser extendida a una función continua en el conjunto* $[0, 1] \times [0, 1]$.

(b) Si se cumplen las condiciones ***R1*** *y* ***R4****, entonces existe una constante* $\tau > 0$ *tal que, para todo* $0 < \pi < 1$*, se cumple que*

$$M(\pi, \pi_0) - M(\pi_0, \pi_0) \geq \tau(\pi - \pi_0)^2.$$

Demostración. (a) Observemos que las hipótesis **R1** y **R2** implican que existe una constante finita ρ_∞ tal que

$$\lim_{t \to \infty} \rho(t) = \rho_\infty.$$

Por lo tanto, si definimos $M(0, \pi_0) = \pi_0 \rho_\infty + G(1)$ y $M(1, \pi_0) = (1 - \pi_0)\rho_\infty + G(1)$, es fácil verificar que la función $M(\pi, \pi_0)$ así extendida, es continua en $[0, 1] \times [0, 1]$, lo que concluye la demostración de (a).

(b) Supongamos $0 < \pi < 1$. Como en la demostración del Lema 2.1 de Bianco y Yohai (1996), tenemos que

$$\frac{\partial}{\partial \pi} M(\pi, \pi_0) = (\pi - \pi_0) g(\pi)$$

donde $g(\pi)$ está definida por

$$g(\pi) = \left(\frac{\psi(-\log \pi)}{\pi} + \frac{\psi(-\log(1 - \pi))}{1 - \pi} \right).$$

Notemos que la función g es una función simétrica alrededor de $\pi = 1/2$ y tal que, si $\pi \in [1/2, 1]$,

$$g(\pi) \geq \frac{\psi(-\log \pi)}{\pi} \geq \psi(-\log \pi) \geq \tau, \tag{6.13}$$

donde τ es la constante de la hipótesis **R4**.

El resultado buscado es entonces una consecuencia de (6.13) y del hecho de que

$$M(\pi, \pi_0) - M(\pi_0, \pi_0) = \int_{\pi_0}^{\pi} \frac{\partial M(\pi, \pi_0)}{\partial \pi}\Big|_{\pi = u} du. \quad \square$$

Lema 6.11. *Sea* $\mathbf{Z} = (Z_1, Z_2)^{\mathrm{T}}$ *un vector aleatorio con distribución elíptica centrada y función característica* $\phi_{\mathbf{Z}}(\mathbf{z}) = \xi(\mathbf{z}^{\mathrm{T}} \boldsymbol{\Upsilon} \mathbf{z})$*. Supongamos que* $\mathbb{E}(Z_j^2) < \infty$*, para* $j = 1, 2$*, de forma tal que si* $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{Z})$ *tenemos que*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = -2\xi'(0)\boldsymbol{\Upsilon}\,.$$

Supongamos además que $\sigma_2 > 0$ *y que existe* $K_2 > 0$ *tal que* $\sigma_2 \leq K_2$ *y la distribución de* $\mathbf{Z}$ *verifica*

$$\mathbb{E}\left\{[F(Z_1) - F(Z_2)]^2\right\} < \frac{[F(4\,K_2) - F(2\,K_2)]^2}{4}\,, \tag{6.14}$$

donde $F(t) = \exp(t)/(1 + \exp(t))$*. Entonces, tenemos que*

(a) existe una constante C_0 que solamente depende de ξ tal que $\sigma_1 \leq C_0\, K_2$.

(b) existe una constante C_2 que solamente depende de K_2 y ξ y cumple

$$\mathbb{E}[(Z_1 - Z_2)^2] \leq C_2\, \mathbb{E}[(F(Z_1) - F(Z_2))^2].$$

DEMOSTRACIÓN. Para probar el item (a), observemos que si $\sigma_1 = 0$, la desigualdad vale con $C_0 = 1$. Supongamos por lo tanto que $\sigma_1 \neq 0$. La desigualdad de Tchebychev implica que

$$\mathbb{P}(|Z_2| \leq 2K_2) \geq 3/4\,. \tag{6.15}$$

Sea c tal que

$$\mathbb{E}\left\{[F(Z_1) - F(Z_2)]^2\right\} < c < \frac{[F(4\,K_2) - F(2\,K_2)]^2}{4}\,, \tag{6.16}$$

tenemos entonces que

$$\begin{aligned} c &> \mathbb{E}\left\{[F(Z_1) - F(Z_2)]^2\right\} \geq \mathbb{E}\left\{[F(Z_1) - F(Z_2)]^2\ \mathbb{I}_{|Z_2|\leq 2K_2}\ \mathbb{I}_{|Z_1|>4K_2}\right\} \\ &\geq [F(4K_2) - F(2K_2)]^2\ \mathbb{P}(|Z_2| \leq 2K_2 \cap |Z_1| > 4K_2) \\ &\geq [F(4K_2) - F(2K_2)]^2\ \{\mathbb{P}(|Z_2| \leq 2K_2) - \mathbb{P}(|Z_1| \leq 4K_2)\}. \end{aligned}$$

En virtud de esta desigualdad y (6.16) deducimos que

$$\mathbb{P}(|Z_2| \leq 2K_2) - \mathbb{P}(|Z_1| \leq 4K_2) < \frac{1}{4}\,,$$

con lo cual usando (6.15), concluimos que

$$\mathbb{P}(|Z_1| \leq 4K_2) \geq 1/2. \tag{6.17}$$

Sea $V_j = Z_j/\sigma_j$, $j = 1, 2$. La función característica de $V_1 = Z_1/\sigma_1$ satisface

$$\phi_{V_1}(s) = \phi_{\mathbf{Z}}\left(\left(\frac{s}{\sigma_1}, 0\right)^{\mathrm{T}}\right) = \xi\left(\frac{s^2}{-2\xi'(0)}\right), \tag{6.18}$$

de modo que su distribución solo depende de ξ. Sea m la mediana de $|V_1|$. Observemos que (6.17) implica que $\mathbb{P}(|V_1| \leq 4K_2/\sigma_1) \geq 1/2$, entonces $4K_2/\sigma_1 \geq m$, de donde se deduce el resultado buscado.

Probaremos ahora el item (b). Supongamos primero que $|\rho| \neq 1$ y $\sigma_1 \neq 0$, con lo cual la matriz $\mathbf{\Sigma}$ es inversible. Sea $\mathbf{W} = \mathbf{\Sigma}^{-1/2}\mathbf{Z}$, luego $\mathbf{W}$ tiene distribución esférica y

$$\phi_{\mathbf{W}}(\mathbf{t}) = \xi\left(\frac{\mathbf{t}^{\mathrm{T}}\mathbf{t}}{-2\xi'(0)}\right).$$

Observemos que los autovalores de $\mathbf{\Sigma}$ están acotados por $\mathrm{TR}(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2$, con lo cual por el item (a), sus autovalores están acotados por $(C_0^2 + 1)K_2^2$.

Por otra parte, notemos que

$$\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Z} \geq \iota_1(\mathbf{\Sigma}^{-1})\|\mathbf{Z}\|_2^2 \geq \frac{1}{(C_0^2 + 1)K_2^2}\|\mathbf{Z}\|_2^2\,.$$

Sea m_1 la mediana de $\|\mathbf{W}\|_2$ que sólo depende de ξ. Luego, si $\|\mathbf{W}\|_2^2 \leq m_1^2$ tenemos que $\max\{|Z_1|, |Z_2|\} \leq C_1\, K_2$ donde $C_1 = m_1(C_0^2 + 1)^{1/2}$ sólo depende de ξ. Por lo tanto, el Teorema de

Valor Medio implica que existe $-C_1\,K_2 \le \theta \le C_1\,K_2$ tal que $[F(Z_1)-F(Z_2)]^2 = [F'(\theta)]^2\,(Z_1-Z_2)^2$. Usando que $F'(t) = F(t)(1-F(t))$ es par, creciente en $(-\infty,0]$ y decreciente en $[0,\infty)$ deducimos que $[F(Z_1)-F(Z_2)]^2 \ge [F'(C_1\,K_2)]^2\,(Z_1-Z_2)^2$ con lo cual

$$\mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\right\} \ge \mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\,\mathbb{I}_{\|\mathbf{W}\|_2\le m_1}\right\} \ge \mathbb{E}\left\{\left[F'(C_1\,K_2)\right]^2\,(Z_1-Z_2)^2\mathbb{I}_{\|\mathbf{W}\|_2\le m_1}\right\}.$$

Sea $\mathbf{v} = (1,-1)^{\mathrm{T}}$, entonces, como $\mathbf{W} = \mathbf{\Sigma}^{-1/2}\mathbf{Z}$ y $\mathbf{W}/\|\mathbf{W}\|_2$ es independiente de $\|\mathbf{W}\|_2$ obtenemos que

$$\begin{aligned}\mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\right\} &\ge \left[F'(C_1\,K_2)\right]^2\mathbb{E}\left(\mathbf{v}^{\mathrm{T}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{v}\,\mathbb{I}_{\|\mathbf{W}\|_2\le m_1}\right)\\ &= \left[F'(C_1\,K_2)\right]^2\mathbb{E}\left(\mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}^{1/2}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{\Sigma}^{1/2}\mathbf{v}\,\mathbb{I}_{\|\mathbf{W}\|_2\le m_1}\right)\\ &= \left[F'(C_1\,K_2)\right]^2\mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}^{1/2}\mathbb{E}\left(\frac{\mathbf{W}\mathbf{W}^{\mathrm{T}}}{\|\mathbf{W}\|_2^2}\right)\mathbb{E}\left(\|\mathbf{W}\|_2^2\mathbb{I}_{\|\mathbf{W}\|_2\le m_1}\right)\mathbf{\Sigma}^{1/2}\mathbf{v}\,.\end{aligned}$$

Más aún, como $\mathbb{E}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}}/\|\mathbf{W}\|_2^2\right) = (1/2)\mathbf{I}_2$ pues $\mathbf{W}/\|\mathbf{W}\|_2$ tiene distribución uniforme en la esfera unidad y $\mathbf{v}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v} = \mathbb{E}[(Z_1-Z_2)^2]$, obtenemos que

$$\mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\right\} \ge \frac{\left[F'(C_1\,K_2)\right]^2}{2}\mathbb{E}\left(\|\mathbf{W}\|_2^2\mathbb{I}_{\|\mathbf{W}\|_2\le m_1}\right)\mathbb{E}\left[(Z_1-Z_2)^2\right] = C\,\mathbb{E}\left[(Z_1-Z_2)^2\right].$$

Como m_1 es la mediana de $\|\mathbf{W}\|_2$, tenemos que $\mathbb{P}(\|\mathbf{W}\|_2 \le m_1) \ge 1/2$, de donde deducimos que $\mathbb{E}\left(\|\mathbf{W}\|_2^2\mathbb{I}_{\|\mathbf{W}\|_2\le m_1}\right) \ne 0$ y por lo tanto, $C>0$ concluyendo la demostración de (b) si $\rho \ne 1$.

Consideremos ahora el caso en que $|\rho| = 1$ y $\sigma_1 \ne 0$. Cuando esto sucede, existe una constante $a \ne 0$ tal que $Z_1 = aZ_2$, de donde $\sigma_1^2 = a^2\,\sigma_2^2$ y $V_2 = Z_2/\sigma_2 = Z_1/\sigma_1 = V_1$. Recordemos que llamamos m a la mediana de $|V_1| = |V_2|$, de donde $\mathbb{P}(|V_2| \le m) \ge 1/2$. Luego, si $|V_2| \le m$ entonces $|Z_1| \le m\sigma_1 \le m\,C_0K_2$ por (a) y $|Z_2| \le m\sigma_2 \le mK_2$, es decir, $\max\{|Z_1|,|Z_2|\} \le C_2\,K_2$ donde $C_2 = m\max\{C_0,1\}$. Por lo tanto, usando que $\mathbb{E}\left[(Z_1-Z_2)^2\right] = (a-1)^2\sigma_2^2$ y el Teorema del Valor Medio obtenemos que

$$\begin{aligned}\mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\right\} &\ge \mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\,\mathbb{I}_{|V_2|\le m}\right\} \ge \left[F'(C_2\,K_2)\right]^2\mathbb{E}\left[(Z_1-Z_2)^2\mathbb{I}_{|V_2|\le m}\right]\\ &\ge \left[F'(C_2\,K_2)\right]^2(a-1)^2\sigma_2^2\mathbb{E}\left[\frac{Z_2^2}{\sigma_2^2}\mathbb{I}_{|V_2|\le m}\right]\\ &= \left[F'(C_2\,K_2)\right]^2\mathbb{E}\left[(Z_1-Z_2)^2\right]\mathbb{E}\left(V_2^2\mathbb{I}_{|V_2|\le m}\right).\end{aligned}$$

Como la distribución de V_2 sólo depende de ξ, $\mathbb{E}\left[V_2^2\mathbb{I}_{|V_2|\le m}\right]$ solamente depende de ξ y es no nula pues $\mathbb{P}(|V_2| \le m) \ge 1/2$, concluyendo la demostración.

Finalmente, si $\sigma_1 = 0$, $\mathbb{P}(Z_1 = 0) = 1$, por lo tanto, debemos ver que $\mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\right\} \ge C\sigma_2^2$ para alguna constante C que sólo depende de ξ y K_2. En este caso, sea m_2 la mediana de V_2. Como en el caso $|\rho| = 1$ y $\sigma_1 \ne 0$, tenemos que el Teorema del Valor Medio implica que si $|Z_2| \le m\sigma_2 \le mK_2$ entonces $[F(0)-F(Z_2)]^2 \ge [F'(m\,K_2)]^2\,Z_2^2$, de donde

$$\mathbb{E}\left\{[F(Z_1)-F(Z_2)]^2\right\} \ge \mathbb{E}\left\{[F(0)-F(Z_2)]^2\,\mathbb{I}_{|V_2|\le m}\right\} \ge \left[F'(m\,K_2)\right]^2\sigma_2^2\mathbb{E}\left(V_2^2\mathbb{I}_{|V_2|\le m}\right),$$

lo que concluye la demostración. ■

Demostración del Teorema 6.1. Probaremos primero el item (a). Usando la definición de $\widehat{\boldsymbol{\beta}}_n$, tenemos que

$$L_n(\widehat{\boldsymbol{\beta}}_n) \le L_n(\widehat{\boldsymbol{\beta}}_n) + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) \le L_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}_0)\,,$$

lo cual implica que

$$\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) \leq [L_n(\boldsymbol{\beta}_0) - \mathbb{L}(\boldsymbol{\beta}_0)] - \left[L_n(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\widehat{\boldsymbol{\beta}}_n)\right] + I_{\lambda_n}(\boldsymbol{\beta}_0).$$

Sea C_1 la constante del Lema 6.9, donde supondremos sin pérdida de generalidad que $C_1 > 1$. Consideramos el evento

$$\mathcal{A}_{n,T} = \left\{ \sup_{\boldsymbol{\beta}} |L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta})| \leq C_1\, T \sqrt{\frac{p}{n}} \right\}.$$

En virtud del Lema 6.9 y de la desigualdad de Markov, se tiene que $\mathbb{P}(\mathcal{A}_{n,T}) \geq 1 - 1/T$, para $T > 1$. Entonces, en el evento $\mathcal{A}_{n,T}$, se verifica que

$$\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) \leq 2\, C_1\, T \sqrt{\frac{p}{n}} + I_\lambda(\boldsymbol{\beta}_0) \leq 2\, C_1\, T \left\{ \sqrt{\frac{p}{n}} + I_\lambda(\boldsymbol{\beta}_0) \right\}. \tag{6.19}$$

Un cálculo sencillo muestra que para M definida en (6.12)

$$\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) = \mathbb{E}\left\{ M(F(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n), F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)) - M(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0), F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)) \middle| (Y_1, \mathbf{X}_1), \dots (Y_n, \mathbf{X}_n) \right\},$$

por lo tanto, por el Lema 6.10 tenemos que existe una constante $\tau > 0$, independiente de n, tal que

$$\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\, \mathbb{E}\left\{ \left[F(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]^2 \middle| (Y_1, \mathbf{X}_1), \dots (Y_n, \mathbf{X}_n) \right\} = \tau\, d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0),$$

lo que junto con (6.19) concluye la demostración de (a).

Para probar el item (b), observemos que si $\max\{\|\widehat{\boldsymbol{\beta}}_n\|_1, \|\boldsymbol{\beta}_0\|_1\} \leq R$ y $\|\mathbf{X}\|_\infty \leq M$, entonces por la desigualdad de Hölder $\max\{|\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n|, |\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0|\} \leq M\, R$. Usando que $F'(t)$ es par, creciente en $(-\infty, 0]$ y decreciente en $[0, \infty)$, como en la demostración del Lema 6.11, obtenemos que

$$\begin{aligned} \mathbb{E}\left[(F(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0))^2\right] &\geq (F'(MR))^2\, \mathbb{E}[(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)] \\ &\geq (F'(MR))^2\, \iota_1(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}])\, \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2. \end{aligned}$$

En particular, si se cumple (**Z3**)

$$\mathbb{E}\left[(F(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0))^2\right] \geq \tau_1\, (F'(MR))^2\, \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2.$$

Por lo tanto, el resultado buscado es consecuencia del item (a).

Para probar el item (c), basta mostrar que dados $\varepsilon > 0$ y $\delta > 0$, existe n_0 tal que si $n \geq n_0$, se verifica que

$$\mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \leq \varepsilon) > 1 - \delta. \tag{6.20}$$

La hipótesis **Z5** implica que, para cualquier $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{Z}_{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^{\mathrm{T}}$ tiene distribución elíptica con segundo momento finito y función generadora ξ. De la hipótesis **Z3** deducimos que $\textsc{Var}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) \neq 0$. Por otra parte, $\textsc{Var}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) \leq K_2$ por **Z4**, donde K_2 no depende de n.

Como $p/n \to 0$ y $I_{\lambda_n}(\boldsymbol{\beta}_0) \to 0$, del item (a) tenemos que $d_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \xrightarrow{p} 0$. Definamos el evento

$$\mathcal{B}_n = \left\{ d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq \frac{[F(4K_2) - F(2K_2)]^2}{8} \right\}.$$

Si $\omega \in \mathcal{B}_n$, $\mathbf{Z}_{\widehat{\boldsymbol{\beta}}_n(\omega)}$ cumple las condiciones del Lema 6.11(b) con lo cual, existe C_2 que sólo depende de ξ y K_2, pero no depende ni de ω ni de n, tal que

$$D_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq C_2 d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0)\,,$$

donde $D_n^2(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = \mathbb{E}[(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2]$. Definamos además el conjunto

$$\mathcal{A}_{\varepsilon,n} = \left\{ d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq \frac{\varepsilon}{C_2}\tau_1 \right\}\,,$$

donde τ_1 está dada en **Z3**.

Como $d_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \xrightarrow{p} 0$, tenemos que $\lim_n \mathbb{P}(\mathcal{A}_{\varepsilon,n}) = \lim_n \mathbb{P}(\mathcal{B}_n) = 1$, de donde

$$\lim_n \mathbb{P}(\mathcal{A}_{\varepsilon,n} \cap \mathcal{B}_n) = 1\,,$$

y existe n_0 tal que si $n \geq n_0$, se cumple que $\mathbb{P}(\mathcal{A}_{\varepsilon,n} \cap \mathcal{B}_n) > 1 - \delta$.

Observemos que si $\omega \in \mathcal{A}_{\varepsilon,n} \cap \mathcal{B}_n$, tenemos que

$$D_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq C_2\; d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq \varepsilon\tau_1\;.$$

Por otra parte,

$$D_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \geq \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \iota_1(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]) \geq \tau_1 \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2\,,$$

de donde $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 < \varepsilon$. Por lo tanto, obtenemos que (6.20) vale, concluyendo la demostración. ■

6.6. Apéndice B: Demostraciones de la Sección 6.2

Demostración del Teorema 6.2. Probaremos el resultado para el estimador $\widehat{\boldsymbol{\beta}}_n$ definido en (3.19), la demostración para el estimador dado en (6.1) es análoga. Vale la pena mencionar que **R1** implica que $\mathbb{L}(\boldsymbol{\beta}) < \infty$ para todo $\boldsymbol{\beta}$.

Empecemos probando el item (a).

Sean $v_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta})$ y $\ell_n = \sqrt{n/(p\log p)}$. Para acotar el incremento del proceso empírico v_n, usaremos el Lema 4.6(a) definiendo $\gamma(y, s) = \phi(y, s)$, $y \in \{0, 1\}$. Observemos que $\gamma(y, s)$ es derivable respecto de su segundo argumento con derivada $\gamma'(y, s) = \Psi(y, s)$ donde Ψ está definida en (3.11). Por lo tanto, por **R1** $\|\gamma'\|_\infty \leq 4\|\psi\|_\infty < \infty$, luego por el Teorema de Valor Medio, tenemos que γ cumple (4.1) con $C_\gamma = 4\|\psi\|_\infty$. Por lo tanto, el Lema 4.6 implica que para todo $M > 0$

$$\mathbb{E}\left(\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \leq M} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \right) \leq 4MC_\gamma \sqrt{\frac{2\log(2p)}{n}}\, \mathbb{E}\left(\max_{1\leq j\leq p} \frac{1}{n}\sum_{i=1}^n X_{ij}^2 \right) \leq MC_1 \sqrt{\frac{\log p}{n}}\,,$$

donde la última desigualdad se deduce de **Z1** y la constante C_1 no depende de n ni de p.

Como $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \delta$ implica que $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \leq \sqrt{p}\,\delta$, obtenemos que

$$\mathbb{E}\left(\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \right) \leq \mathbb{E}\left(\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \leq \sqrt{p}\,\delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \right) \leq \frac{C_1\,\delta}{\ell_n}\,,$$

de donde, en virtud de la desigualdad de Markov, concluimos que para todo $C > 0$

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\leq\delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| > C\right) \leq \frac{C_1\,\delta}{\ell_n\,C}\,. \tag{6.21}$$

El resto de la demostración sigue las mismas ideas que las de la prueba del Teorema 3.2.5 de Van der Vaart y Wellner (1996) y se basa en el denominado "dispositivo de pelado"(peeling device).

Sea $c_n = \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq \eta)$, donde $\eta > 0$ es tal que $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$, para todo $n \geq 1$ y $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \eta$. Como $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow{p} 0$, tenemos que $c_n \to 0$. Para cada $j \in \mathbb{N}$, definamos los conjuntos

$$A_{n,j} = \{\boldsymbol{\beta} \in \mathbb{R}^p : 2^{j-1} \leq \ell_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq 2^j\}\,.$$

Sea M un número natural y observemos que como $\widehat{\boldsymbol{\beta}}_n$ minimiza $L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta})$, se tiene que $L_n(\widehat{\boldsymbol{\beta}}_n) + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) \leq L_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}_0)$, por lo tanto

$$\begin{aligned}
\mathbb{P}(\ell_n\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq 2^M) &\leq c_n + \sum_{\substack{j\geq M+1\\ 2^j\leq\ell_n\eta}} \mathbb{P}\left(\widehat{\boldsymbol{\beta}}_n \in A_{n,j}\right)\\
&\leq c_n + \sum_{\substack{j\geq M+1\\ 2^j\leq\ell_n\eta}} \mathbb{P}\left(\inf_{\boldsymbol{\beta}\in A_{n,j}} L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}_0) - I_{\lambda_n}(\boldsymbol{\beta}_0) \leq 0\right)\\
&= c_n + \sum_{\substack{j\geq M+1\\ 2^j\leq\ell_n\eta}} \mathbb{P}\left(\inf_{\boldsymbol{\beta}\in A_{n,j}} v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) + \mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \leq 0\right)\,.
\end{aligned}$$

Observemos que **P1** implica que $I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) \geq -|I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0)| \geq -\lambda_n K\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1$. Además, dado $\boldsymbol{\beta} \in A_{n,j}$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \eta$ si $2^j \leq \ell_n\eta$, con lo cual $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$. Por lo tanto, si $\boldsymbol{\beta} \in A_{n,j}$

$$v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) + \mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq -|v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| - \lambda_n K\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2\,,$$

lo que nos permite concluir que $\mathbb{P}(\ell_n\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq 2^M) \leq c_n + d_n$ donde $d_n = \sum_{\substack{j\geq M+1\\ 2^j\leq\ell_n\eta}} d_{n,j}$ con

$$d_{n,j} = \mathbb{P}\left(-\sup_{\boldsymbol{\beta}\in A_{n,j}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| - K\lambda_n \sup_{\boldsymbol{\beta}\in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \tau \inf_{\boldsymbol{\beta}\in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \leq 0\right)\,.$$

Observemos que si $\boldsymbol{\beta} \in A_{n,j}$,

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \geq \frac{2^{2j-2}}{\ell_n^2} \qquad \text{y} \qquad \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \leq \sqrt{p}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \frac{\sqrt{p}\,2^j}{\ell_n}\,,$$

por lo tanto,

$$-K\lambda_n \sup_{\boldsymbol{\beta}\in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \tau \inf_{\boldsymbol{\beta}\in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \geq -K\lambda_n\,\frac{\sqrt{p}\,2^j}{\ell_n} + \frac{\tau\,2^{2j-2}}{\ell_n^2} = \alpha_n\,,$$

de donde

$$d_{n,j} \leq \mathbb{P}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\leq\frac{2^j}{\ell_n}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \geq \alpha_n\right)\,.$$

Ahora bien, $\lambda_n = O(\sqrt{\log p/n})$, luego existe una constante $D > 0$ tal que $\lambda_n \leq D\sqrt{\log p/n}$ para todo n, con lo cual si

$$M \geq \frac{\log\left(\dfrac{8\,K\,D}{\tau}\right)}{\log 2} + 1 = M_0\,,$$

tenemos que $\alpha_n > 0$, de donde usando (6.21) obtenemos que, para todo $j \geq M+1$,

$$d_{n,j} \leq C_1 \frac{2^j}{\ell_n^2 \alpha_n}\,.$$

Ahora bien, como $\lambda_n \leq D\sqrt{\log p/n}$ para todo n, tenemos que $\lambda_n \sqrt{p} \leq D/\ell_n$ de donde $\ell_n^2\,\alpha_n \geq 2^j\left(\tau 2^{j-2} - KD\right) > \tau\, 2^{2\,j}/8$, si $j \geq M+1$ con lo cual $d_{n,j} \leq 2^{-j}\ (8\,C_1)/\tau$. Dado $\varepsilon > 0$,sea $N_\varepsilon \in \mathbb{N}$ tal que si $n \geq N_\varepsilon$, $c_n \leq \varepsilon/2$. Por otra parte, sea $M_\varepsilon \in \mathbb{N}$, $M_\varepsilon \geq M_0$ tal que $\sum_{j \geq M_\varepsilon} 2^{-j} < \tau\varepsilon/(16 C_1)$. Luego, si $n \geq N_\varepsilon$, tenemos que $\mathbb{P}\left(\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq 2^{M_\varepsilon}\right) \leq \varepsilon$, concluyendo la demostración de (a).

Probaremos ahora el item (b). En este caso, definamos $\ell_n = \sqrt{n/p}$.

Notemos que si $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \delta$, entonces $\mathbb{E}[(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2] \leq \delta^2 \iota_p$, donde por simplicidad indicamos por ι_p al máximo autovalor de $\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]$. Usando el Lema 4.6(a) obtenemos

$$\mathbb{E}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 \leq \delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)|\right) \leq \mathbb{E}\left(\sup_{\mathbb{E}[(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^2] \leq \delta^2\,\iota_p} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)|\right) \leq 4 C_\gamma \delta \sqrt{\iota_p}\sqrt{\frac{p}{n}}\,.$$

La condición **Z2** asegura que $\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}]) \leq K_1$, para todo n, de donde

$$\mathbb{E}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 \leq \delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)|\right) \leq 4 C_\gamma \sqrt{K_1}\delta\,\sqrt{\frac{p}{n}} = \frac{4C\sqrt{K_1}\delta}{\ell_n}\,.$$

El resto de la demostración es completamente análoga a la demostración del item (a).

A continuación, probaremos el item (c)(i).

Al igual que en la prueba del item (a), tomamos $\ell_n = \sqrt{n/(p\log p)}$. Siguiendo la misma cadena de desigualdades, llegamos a

$$\begin{aligned} V_n(\boldsymbol{\beta}) = v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) + \mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq -|v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \\ - a_n\sqrt{k}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 - b_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 + \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2\,. \end{aligned}$$

Con lo cual, si $\boldsymbol{\beta} \in A_{n,j}$ tenemos que

$$V_n(\boldsymbol{\beta}) \geq -\sup_{\boldsymbol{\beta} \in A_{n,j}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| - a_n\sqrt{k}\frac{2^j}{\ell_n} - b_n\frac{2^{2j}}{\ell_n^2} + \tau\frac{2^{2j-2}}{\ell_n^2}\,,$$

con lo cual $\mathbb{P}(\ell_n\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq 2^M) \leq c_n + d_n$ donde ahora $c_n = \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq \widetilde{\eta})$, con $\widetilde{\eta} = \text{mín}(\eta, \widetilde{\delta})$ y $d_n = \displaystyle\sum_{\substack{j \geq M+1 \\ 2^j \leq \ell_n \widetilde{\eta}}} d_{n,j}$ con

$$d_{n,j} = \mathbb{P}(\inf_{\boldsymbol{\beta} \in A_{n,j}} V_n(\boldsymbol{\beta}) \leq 0) \leq \mathbb{P}\left(\sup_{\boldsymbol{\beta} \in A_{n,j}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \geq \frac{\tau\, 2^{2j-2}}{\ell_n^2} - a_n\sqrt{k}\frac{2^j}{\ell_n} - b_n\frac{2^{2j}}{\ell_n^2}\right)\,.$$

Como $a_n\sqrt{k} = O(1/\ell_n)$, existe $D > 0$ tal que $a_n\sqrt{k} \le D/\ell_n$, para todo n. Sea n_0 tal que si $n \ge n_0$, $b_n \le \tau/8$ y sea

$$M \ge \frac{\log\left(\dfrac{16\,D}{\tau}\right)}{\log 2} + 1 = M_0\,.$$

Por lo tanto, si $n \ge n_0$ y $M \ge M_0$

$$\alpha_n = \frac{\tau\,2^{2j-2}}{\ell_n^2} - a_n\sqrt{k}\frac{2^j}{\ell_n} - b_n\frac{2^{2j}}{\ell_n^2} \ge \frac{2^{2j}}{\ell_n^2}\left(\frac{\tau}{4} - \frac{D}{2^j} - b_n\right) \ge \frac{2^{2j}}{\ell_n^2}\frac{\tau}{16}\,,$$

de donde

$$d_{n,j} \le \mathbb{P}\left(\sup_{\boldsymbol{\beta}\in A_{n,j}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \ge \alpha_n\right) \le \mathbb{P}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 \le \frac{2^j}{\ell_n}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \ge \frac{2^{2j}}{\ell_n^2}\frac{\tau}{16}\right)\,.$$

Esta desigualdad y (6.21) permiten concluir que $d_{n,j} \le (16\,C_1)/(\tau\,2^j)$ y la demostración sigue como en (a).

La demostración del item (c)(ii) es completamente análoga, tomando $\ell_n = \sqrt{n/p}$. ■

Demostración del Lema 6.3. En primer lugar, observemos que como $\mathbb{L}(\boldsymbol{\beta}) = \mathbb{E}\phi\,(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$, se cumple que

$$\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) = \mathbb{E}\left[\phi(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) - \phi(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\,.$$

Teniendo en cuenta que ρ cumple **R1** y **R4**, como veremos en el Lema 6.10 existe una constante $\tau_0 > 0$, independiente de n, tal que

$$\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) = \mathbb{E}\left[\phi(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) - \phi(F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0), \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right] \ge \tau_0\mathbb{E}\left[F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]^2\,.$$

Por otra parte, tenemos que

$$\mathbb{E}\left[F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]^2 \le \iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}])\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2\,.$$

La hipótesis **Z5** implica que $\mathbf{Z} = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)^{\mathrm{T}}$ tiene distribución elíptica con segundo momento finito, mientras que de **Z3** deducimos que $\textsc{Var}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) \ne 0$. Más aún, $\textsc{Var}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0) \le K_2^2$ por **Z4**, donde K_2 no depende de n. Por lo tanto, podemos elegir $\eta > 0$ suficientemente chico de manera tal que se cumpla la condición (6.14) del Lema 6.11 con $\sigma_2^2 = \textsc{Var}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)$ y K_2 la constante dada en la hipótesis **Z4**. Utilizando el Lema 6.11(b), podemos concluir que existe una constante C_1, independiente de n, tal que

$$\mathbb{E}\left[\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0\right]^2 \le C_1\mathbb{E}\left[F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]^2\,.$$

Las cotas anteriores implican que

$$\mathbb{L}(\boldsymbol{\beta})-\mathbb{L}(\boldsymbol{\beta}_0) \ge \tau_0\mathbb{E}\left[F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) - F(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\right]^2 \ge \tau_0 C_1^{-1}\mathbb{E}\left[\mathbf{X}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]^2 \ge \tau_0 C_1^{-1}\iota_1(\mathbb{E}\left[\mathbf{X}\mathbf{X}^{\mathrm{T}}\right])\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2^2\,,$$

lo que nos permite concluir que la condición del Teorema 6.2 sobre la función $\mathbb{L}$ efectivamente se verifica tomando $\tau = \tau_0\,\tau_1\,C_1^{-1}$ donde τ_1 está dada en **Z3**. ■

6.7. Apéndice C: Demostraciones de la Sección 6.3

Demostración del Teorema 6.4. Esta demostración es muy similar a la hecha para el Teorema 5.5. Dado $\tau > 0$, mostraremos que $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) > 1-\tau$ para n suficientemente grande. Definimos $V_n : \mathbb{R}^k \times \mathbb{R}^{p-k} \to \mathbb{R}$ como

$$V_n(\mathbf{u}_1, \mathbf{u}_2) = L_n\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \frac{\mathbf{u}_2}{\ell_n}\right) + I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \frac{\mathbf{u}_2}{\ell_n}\right),$$

donde $L_n(\boldsymbol{\beta})$ está dado en (3.10). Sea $C > 0$ tal que $\mathbb{P}(\mathcal{B}_n) \geq 1 - \tau/2$, donde

$$\mathcal{B}_n = \{\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq C\}\,.$$

Luego, para todo $\omega \in \mathcal{B}_n$, si escribimos

$$\widehat{\boldsymbol{\beta}}_n = \left(\boldsymbol{\beta}_{0,I}^{\mathrm{T}} + \frac{\mathbf{U}_{1,n}^{\mathrm{T}}}{\ell_n}, \frac{\mathbf{U}_{2,n}^{\mathrm{T}}}{\ell_n}\right)^{\mathrm{T}},$$

tenemos que $\|\mathbf{U}_n\|_2 \leq C$, donde $\mathbf{U}_n = (\mathbf{U}_{1,n}^{\mathrm{T}}, \mathbf{U}_{2,n}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{U}_{1,n} \in \mathbb{R}^k$, $\mathbf{U}_{2,n} \in \mathbb{R}^{p-k}$. Observemos que

$$(\mathbf{U}_{1,n}^{\mathrm{T}}, \mathbf{U}_{2,n}^{\mathrm{T}})^{\mathrm{T}} = \underset{\|\mathbf{u}_1\|_2^2 + \|\mathbf{u}_2\|_2^2 \leq C^2}{\operatorname{argmin}} V_n(\mathbf{u}_1, \mathbf{u}_2)\,. \tag{6.22}$$

Nuestro objetivo es probar que, con alta probabilidad, $V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}_{p-k}) > 0$ para todo $\|\mathbf{u}_1\|_2^2 + \|\mathbf{u}_2\|_2^2 \leq C^2$ tal que $\mathbf{u}_2 \neq \mathbf{0}_{p-k}$.

Consideremos $\mathbf{u}_1 \in \mathbb{R}^k$ y $\mathbf{u}_2 \neq \mathbf{0}_{p-k}$ tales que $\|\mathbf{u}_1\|_2^2 + \|\mathbf{u}_2\|_2^2 \leq C^2$. Observemos que $V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}_{p-k}) = S_{1,n} + S_{2,n}$ donde

$$\begin{aligned} S_{1,n} &= L_n\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \frac{\mathbf{u}_2}{\ell_n}\right) - L_n\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \mathbf{0}_{p-k}\right), \\ S_{2,n} &= I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \frac{\mathbf{u}_2}{\ell_n}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I} + \frac{\mathbf{u}_1}{\ell_n}, \mathbf{0}_{p-k}\right). \end{aligned}$$

En primer lugar, acotaremos $S_{1,n}$. El Teorema de Valor Medio implica que

$$S_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^*)\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)}\,,$$

donde $\mathbf{u}_n^{(0)} = (\mathbf{0}_k^{\mathrm{T}}, \mathbf{u}_2^{\mathrm{T}}/\ell_n)^{\mathrm{T}}$ y

$$\boldsymbol{\beta}_n^* = \begin{pmatrix} \boldsymbol{\beta}_{0,I} + \dfrac{\mathbf{u}_1}{\ell_n} \\ \\ \alpha_{n,1}\dfrac{\mathbf{u}_2}{\ell_n} \end{pmatrix},$$

para algún $\alpha_{n,1} \in [0,1]$. Por otra parte, usando nuevamente el Teorema de Valor Medio, tenemos que

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}[\Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^*) - \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)]\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)} &= \frac{1}{n}\sum_{i=1}^{n}\chi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^{**})(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)} \\ &= (\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{A}_n(\boldsymbol{\beta}_n^{**})\mathbf{u}_n^{(0)}\,, \end{aligned}$$

donde $\mathbf{A}_n(\boldsymbol{\beta})$ es la matriz dada en (5.12) y

$$\boldsymbol{\beta}_n^{**} = \begin{pmatrix} \boldsymbol{\beta}_{0,I} + \alpha_{n,2}\dfrac{\mathbf{u}_1}{\ell_n} \\ \\ \alpha_{n,2}\,\alpha_{n,1}\dfrac{\mathbf{u}_2}{\ell_n} \end{pmatrix},$$

con $\alpha_{n,2} \in [0,1]$. Luego, observando que

$$S_{1,n} = \left\{ \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}_i^{\mathrm{T}} + \frac{1}{n}\sum_{i=1}^{n}[\Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n^*) - \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)]\mathbf{X}_i^{\mathrm{T}} \right\} \mathbf{u}_n^{(0)}.$$

obtenemos que $S_{1,n} = S_{11,n} + S_{12,n}$ donde

$$S_{11,n} = \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}_i^{\mathrm{T}}\mathbf{u}_n^{(0)} = \frac{1}{n}\frac{1}{\ell_n}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}_i^{\mathrm{T}}(\mathbf{0}_k^{\mathrm{T}}, \mathbf{u}_2^{\mathrm{T}})^{\mathrm{T}}$$

$$S_{12,n} = (\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_0)^{\mathrm{T}}\,\mathbf{A}_n(\boldsymbol{\beta}_n^{**})\mathbf{u}_n^{(0)} = \frac{1}{\ell_n^2}(\mathbf{u}_1^{\mathrm{T}}, \alpha_{n,1}\mathbf{u}_2^{\mathrm{T}})\,\mathbf{A}_n(\boldsymbol{\beta}_n^{**})(\mathbf{0}_k^{\mathrm{T}}, \mathbf{u}_2^{\mathrm{T}})^{\mathrm{T}}\,.$$

Primero, acotaremos $S_{11,n}$. Usando el hecho de que $\mathbb{E}[\Psi(Y, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0)X] = 0$ y la desigualdad de Tchebychev, obtenemos que

$$\sqrt{\frac{1}{n\,\iota_p(\mathbf{B})}}\sum_{i=1}^{n} \Psi(Y_i, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{X}_i^{\mathrm{T}}\left(\mathbf{0}_k^{\mathrm{T}}, \frac{\mathbf{u}_2^{\mathrm{T}}}{\|\mathbf{u}_2\|_2}\right)^{\mathrm{T}} = O_{\mathbb{P}}(1),$$

donde la matriz $\mathbf{B}$ está definida en (3.18). Esto implica que

$$S_{11,n} = \frac{\sqrt{\iota_p(\mathbf{B})}}{\ell_n\sqrt{n}}\,\|\mathbf{u}_2\|_2\,O_{\mathbb{P}}(1)\,. \tag{6.23}$$

Por otra parte, usando que la función χ es acotada y aplicando la desigualdad de Markov, tenemos que

$$S_{12,n} = \frac{C\,\iota_p(\mathbb{E}[\mathbf{X}\mathbf{X}^{\mathrm{T}}])}{\ell_n^2}\,\|\mathbf{u}_2\|_2 O_{\mathbb{P}}(1)\,. \tag{6.24}$$

Por lo tanto, de (6.23) y (6.24) deducimos que $S_{1,n} = C(c_n/\ell_n)\|\mathbf{u}_2\|_2 O_{\mathbb{P}}(1)$.

Sea $M_C > 0$ tal que $\mathbb{P}(|S_{1,n}| > M_C(c_n/\ell_n)\|\mathbf{u}_2\|_2) < \tau/2$. Luego, tenemos que

$$\mathbb{P}(S_{1,n} > -M_C(c_n/\ell_n)\|\mathbf{u}_2\|_2) \geq 1 - \tau/2\,.$$

Usando (6.9), tenemos que existen constantes N_C y K_C tal que si $n \geq N_C$, $S_{2,n} \geq K_C(\lambda_n/\ell_n)\|\mathbf{u}_2\|_2$. Entonces, si $n \geq N_C$,

$$\mathbb{P}\left(S_{1,n} + S_{2,n} \geq \|\mathbf{u}_2\|_2\frac{c_n}{\ell_n}(K_C\lambda_n\,c_n^{-1} - M_C)\right) \geq 1 - \tau/2\,.$$

Como $\lambda_n c_n^{-1} \to \infty$, existe $n_1 > N_C$ tal que $K_C\lambda_n\,c_n^{-1} - M_C > 1$, para todo $n \geq n_1$. Por lo tanto, si $n \geq n_1$, $\mathbb{P}(S_{1,n} + S_{2,n} > 0) > 1 - \tau/2$, es decir,

$$\mathbb{P}\left(V_n(\mathbf{u}_1, \mathbf{u}_2) > V_n(\mathbf{u}_1, \mathbf{0}_{p-k})\right) > 1 - \tau/2\,,$$

para todo $\|\mathbf{u}_1\|_2^2 + \|\mathbf{u}_2\|_2^2 \leq C^2$ tal que $\mathbf{u}_2 \neq \mathbf{0}_{p-k}$, lo que junto con el hecho que $\mathbb{P}(\mathcal{B}_n) \geq 1 - \tau/2$, muestra que $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \geq 1 - \tau$ para un $n \geq n_1$ concluyendo la demostración. ■

Demostración del Corolario 6.5. Solamente debemos verificar que se cumple la condición (6.9). Observemos que para la penalización SCAD,

$$I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I}+\frac{\mathbf{u}_1}{\ell_n},\frac{\mathbf{u}_2}{\ell_n}\right)-I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I}+\frac{\mathbf{u}_1}{\ell_n},\mathbf{0}_{p-k}\right)=\sum_{j=1}^{p-k}\mathrm{SCAD}_{\lambda_n,a}\left(\frac{u_{2,j}}{\ell_n}\right),$$

donde $u_{2,j}$ es la $j-$ésima coordenada de $\mathbf{u}_2$ y $\mathrm{SCAD}_{\lambda_n,a}$ está definida en (5.20). Por otra parte, como $\lambda_n\ell_n\to\infty$, existe un $N_C\in\mathbb{N}$ tal que, para todo $n\geq N_C$, $\lambda_n\ell_n>C\geq u_{2,j}$ para $j=1,\ldots,p-k$. Entonces, para $n\geq N_C$,

$$I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I}+\frac{\mathbf{u}_1}{\ell_n},\frac{\mathbf{u}_2}{\ell_n}\right)-I_{\lambda_n}\left(\boldsymbol{\beta}_{0,I}+\frac{\mathbf{u}_1}{\ell_n},\mathbf{0}_{p-k}\right)=\sum_{j=1}^{p-k}\frac{\lambda_n}{\ell_n}|u_{2,j}|\geq\frac{\lambda_n}{\ell_n}\frac{1}{C}\|\mathbf{u}_2\|_2,$$

por lo que se cumple la condición (6.9) tomando $K_C=1/C$. La demostración para la penalización MCP es completamente análoga, usando la función $\mathrm{MCP}_{\lambda_n,a}$ definida en (5.21). ■

Demostración del Corolario 6.6. Debemos ver que dado $\delta>0$ existe un $C>0$ y un n_0 tal que $\mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2<C\sqrt{k/n})>1-\delta$ si $n\geq n_0$. Sea $\mathcal{A}_n=\{\widehat{\boldsymbol{\beta}}_{n,\,II}=\mathbf{0}_{p-k}\}$. Como $\mathbb{P}(\mathcal{A}_n)\to1$, existe $n_0\in\mathbb{N}$ tal que $\mathbb{P}(\mathcal{A}_n)>1-\delta/2$ para todo $n\geq n_0$. Por otra parte, como $\|\widehat{\mathbf{b}}_k-\boldsymbol{\beta}_{0,I}\|_2=O_{\mathbb{P}}\left(\sqrt{k/n}\right)$, existe $C>0$ tal que $\mathbb{P}(\mathcal{C}_n)>1-\delta/2$ donde $\mathcal{C}_n=\{\|\widehat{\mathbf{b}}_k-\boldsymbol{\beta}_{0,I}\|_2<C\sqrt{k/n}\}$. Luego, si $n\geq n_0$, $\mathbb{P}(\mathcal{A}_n\cap\mathcal{C}_n)>1-\delta$.

Usando la condición **P3**, es fácil ver que en $\mathcal{A}_n$ se tiene que $\widehat{\boldsymbol{\beta}}_{n,\,I}=\widehat{\mathbf{b}}_k$, por lo tanto, en $\mathcal{A}_n\cap\mathcal{C}_n$,

$$\|\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|_2=\|\widehat{\boldsymbol{\beta}}_{n,\,I}-\boldsymbol{\beta}_{0,I}\|_2=\|\widehat{\mathbf{b}}_k-\boldsymbol{\beta}_{0,I}\|\leq C\sqrt{\frac{k}{n}},$$

lo que concluye la demostración. ■

6.8. Apéndice D: Demostraciones de la Sección 6.4

Para probar el Teorema 6.7 necesitaremos los siguientes Lemas. El primero es una extensión directa de la desigualdad de Hölder al caso del producto de tres variables aleatorias, que incluimos a los fines de completitud, mientras que el segundo es un resultado análogo al Lema 6.9 sobre una familia indexada por parámetros en compactos de $\mathbb{R}^k$ para lo que usaremos el Lema 4.5.

Lema 6.12. *Sean p, q y r reales positivos tales que $(1/p)+(1/q)+(1/r)=1$. Sean U, V y W variables aleatorias tal que $\mathbb{E}|U|^p<\infty$, $\mathbb{E}|V|^q<\infty$ y $\mathbb{E}|W|^r<\infty$. Luego,*

$$\mathbb{E}|U\,V\,W|\leq\left(\mathbb{E}|U|^p\right)^{1/p}\,\left(\mathbb{E}|V|^q\right)^{1/q}\,\left(\mathbb{E}|W|^r\right)^{1/r}.$$

Demostración. Definamos $1/p^\star=(1/p)+(1/q)$. Luego, $(1/p^\star)+(1/r)=1$. Usando la desigualdad de Hölder, tenemos que

$$\mathbb{E}|U\,V\,W|\leq\left(\mathbb{E}|U\,V|^{p^\star}\right)^{1/p^\star}\left(\mathbb{E}|W|^r\right)^{1/r}. \tag{6.25}$$

Sean $p_1=p/p^\star$ y $q_1=q/p^\star$, luego $(1/p_1)+(1/q_1)=1$. Por lo tanto, aplicando la desigualdad de Hölder a las variables $U_1=|U|^{p^\star}$ y $V_1=|V|^{p^\star}$ obtenemos que

$$\mathbb{E}|U\,V|^{p^\star}=\mathbb{E}|U_1\,V_1|\leq\left(\mathbb{E}|U_1|^{p_1}\right)^{1/p_1}\left(\mathbb{E}|V_1|^{q_1}\right)^{1/q_1}=\left(\mathbb{E}|U|^p\right)^{\frac{p^\star}{p}}\left(\mathbb{E}|V|^q\right)^{\frac{p^\star}{q}}. \tag{6.26}$$

El resultado deseado se obtiene de (6.25) y (6.26). ■

De ahora en más indicaremos por $\mathcal{B}(\boldsymbol{\beta}_{0,I}) = \mathcal{B}_k(\boldsymbol{\beta}_{0,I}, 1) = \{\mathbf{b} \in \mathbb{R}^k : \|\mathbf{b} - \boldsymbol{\beta}_{0,I}\|_2 \leq 1\}$ la bola de radio 1 centrada en $\boldsymbol{\beta}_{0,I}$ y por $\mathcal{S}^{k-1} = \{\mathbf{b} \in \mathbb{R}^k : \|\mathbf{b}\|_2 = 1\}$ la esfera unidad en $\mathbb{R}^k$. Como en la Sección 6.5, para una función $h : \mathbb{R}^{k+1} \to \mathbb{R}$ indicaremos por $P_n h = (1/n)\sum_{i=1}^n h(Y_i, \mathbf{X}_{i,I})$ y $Ph = \mathbb{E}[h(Y, \mathbf{X}_I)]$. Para evitar notación engorrosa, indicaremos por $\mathbf{b}$, $\mathbf{v}$ y $\mathbf{w}$ a los vectores $\mathbf{b}_k, \mathbf{v}_k, \mathbf{w}_k \in \mathbb{R}^k$, respectivamente, cuando no exista confusión.

Lema 6.13. *Consideremos la familia de funciones*

$$\mathcal{H} = \{h_{\mathbf{w},\mathbf{b},\mathbf{u}} : \{0,1\}\times\mathbb{R}^k \to \mathbb{R} \text{ tales que } h_{\mathbf{w},\mathbf{b},\mathbf{u}}(y,\mathbf{z}) = \chi(y, \mathbf{z}^{\mathrm{T}}\mathbf{b})\mathbf{w}^{\mathrm{T}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{u} : \mathbf{b} \in \mathcal{B}(\boldsymbol{\beta}_{0,I}), \mathbf{w}, \mathbf{u} \in \mathcal{S}^{k-1}\}\,. \tag{6.27}$$

*Entonces, si se cumplen las hipótesis **R5** y **Z6**, existe una constante C que no depende de n ni de p tal que*

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} |(P_n - P)(h)|\right] \leq C\sqrt{\frac{k}{n}}.$$

Demostración. Observemos que la clase $\mathcal{H}$ tiene como envolvente natural $H(\mathbf{z}) = \|\chi\|_\infty \|\mathbf{z}\|_2^2$ y $\|H\|_{2,P} = \left[\mathbb{E}H^2(\mathbf{X}_I)\right]^{1/2} \leq \|\chi\|_\infty \left[\mathbb{E}\|\mathbf{X}_I\|_2^6\right]^{1/3} \leq \|\chi\|_\infty K_3^{1/3} < \infty$ donde K_3 es la constante independiente de n dada en **Z6**.

Por el Lema 4.1, $N(\varepsilon/2, \mathcal{H}, \|\cdot\|_{r,\mathbb{Q}}) \leq N_{[\,]}(\varepsilon, \mathcal{H}, \|\cdot\|_{r,\mathbb{Q}})$, para cualquier medida de probabilidad $\mathbb{Q}$ y $r \geq 1$. Luego, para poder aplicar el Lema 4.5 bastará con acotar adecuadamente $N_{[\,]}(\varepsilon, \mathcal{H}, \|\cdot\|_{r,\mathbb{Q}})$, para ello usaremos el Lema 4.7.

Mostraremos, por lo tanto, que la familia $\mathcal{H}$ es Lipschitz respecto de los elementos que la indexan. Sean $\mathbf{w}_1, \mathbf{w}_2, \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{S}^{k-1}$ y $\mathbf{b}_1, \mathbf{b}_2 \in \mathcal{B}(\boldsymbol{\beta}_{0,I})$. Por simplicidad de notación, llamemos $\boldsymbol{\theta}_j = (\mathbf{w}_j^{\mathrm{T}}, \mathbf{b}_j^{\mathrm{T}}, \mathbf{u}_j^{\mathrm{T}})$, $j = 1, 2$, y $h_{\boldsymbol{\theta}_j} = h_{\mathbf{w}_j,\mathbf{b}_j,\mathbf{u}_j}$. Indiquemos por $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ la distancia euclídea usual en $\mathbb{R}^{3k}$, es decir, $d^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2 + \|\mathbf{b}_1 - \mathbf{b}_2\|_2^2 + \|\mathbf{u}_1 - \mathbf{u}_2\|_2^2$.

Indiquemos por $\chi_1(y,s) = (\partial/\partial s)\chi(y,s)$. La hipótesis **R5** asegura que χ_1 es acotada. Por lo tanto, usando

$$\begin{aligned}|h_{\boldsymbol{\theta}_1}(y,\mathbf{z}) - h_{\boldsymbol{\theta}_2}(y,\mathbf{z})| \leq\ & |h_{\boldsymbol{\theta}_1}(y,\mathbf{z}) - h_{\mathbf{w}_1,\mathbf{b}_1,\mathbf{u}_2}(y,\mathbf{z})| \\ & + |h_{\mathbf{w}_1,\mathbf{b}_1,\mathbf{u}_2}(y,\mathbf{z}) - h_{\mathbf{w}_1,\mathbf{b}_2,\mathbf{u}_2}(y,\mathbf{z})| \\ & + |h_{\mathbf{w}_1,\mathbf{b}_2,\mathbf{u}_2}(y,\mathbf{z}) - h_{\boldsymbol{\theta}_2}(y,\mathbf{z})|\,,\end{aligned}$$

el Teorema de Valor Medio y la desigualdad de Cauchy-Schwartz obtenemos que

$$\begin{aligned}|h_{\boldsymbol{\theta}_1}(y,\mathbf{z}) - h_{\boldsymbol{\theta}_2}(y,\mathbf{z})| &\leq \|\chi\|_\infty \|\mathbf{z}\|_2^2 \left(\|\mathbf{u}_1 - \mathbf{u}_2\|_2 + \|\mathbf{w}_1 - \mathbf{w}_2\|_2\right) + \|\chi_1\|_\infty \|\mathbf{z}\|_2^3 \|\mathbf{b}_1 - \mathbf{b}_2\|_2 \\ &\leq \left(2\,\|\chi\|_\infty \|\mathbf{z}\|_2^2 + \|\chi_1\|_\infty \|\mathbf{z}\|_2^3\right)\ \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2\,,\end{aligned}$$

de donde deducimos que $|h_{\boldsymbol{\theta}_1}(y,\mathbf{z}) - h_{\boldsymbol{\theta}_2}(y,\mathbf{z})| \leq H_1(\mathbf{z})\, d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, con

$$H_1(\mathbf{z}) = \left(2\,\|\chi\|_\infty + \|\chi_1\|_\infty\right)\left(\|\mathbf{z}\|_2^2 + \|\mathbf{z}\|_2^3\right)\,.$$

Los Lemas 4.1 y 4.7 permiten concluir que, para cualquier medida de probabilidad $\mathbb{Q}$ y $r \geq 1$ tal que $\|H_1\|_{r,\mathbb{Q}} < \infty$,

$$N\left(\varepsilon\,\|H_1\|_{r,\mathbb{Q}}, \mathcal{H}, \|\cdot\|_{r,\mathbb{Q}}\right) \leq N_{[\,]}\left(2\varepsilon\|H_1\|_{r,\mathbb{Q}}, \mathcal{H}, \|\cdot\|_{r,\mathbb{Q}}\right) \leq N\left(\varepsilon, \mathcal{V}, d\right)\,,$$

con $\mathcal{V} = \mathcal{S}^{k-1} \times \mathcal{B}(\boldsymbol{\beta}_{0,I}) \times \mathcal{S}^{k-1}$.

Observemos que $\mathcal{V} = \mathcal{S}^{k-1} \times \mathcal{B}(\boldsymbol{\beta}_{0,I}) \times \mathcal{S}^{k-1} \subset \mathcal{B}_{3k}(\boldsymbol{\theta}_0, 3)$ donde $\boldsymbol{\theta}_0 = (\mathbf{0}_k^{\mathrm{T}}, \boldsymbol{\beta}_{0,I}^{\mathrm{T}}, \mathbf{0}_k^{\mathrm{T}})^{\mathrm{T}}$ y $\mathcal{B}_{3k}(\boldsymbol{\theta}, \delta)$ es la bola de centro $\boldsymbol{\theta}$ y radio δ en $\mathbb{R}^{3k}$. El Lema 4.2 implica que

$$N(\varepsilon, \mathcal{V}, d) \leq N(\varepsilon, \mathcal{B}_{3k}(\boldsymbol{\theta}_0, 3), d) \leq \left(\frac{12+\varepsilon}{\varepsilon}\right)^{3k}.$$

Por otra parte, como $H(\mathbf{z}) \leq H_1(\mathbf{z})$, H_1 también es una envolvente de la clase de funciones $\mathcal{H}$ y es la envolvente que utilizaremos en lo que sigue. Más aún,

$$\|H_1\|_{2,P} = \left[\mathbb{E}H_1^2(\mathbf{X}_I)\right]^{1/2} \leq (2\,\|\chi\|_\infty + \|\chi_1\|_\infty)\ \|\left(\|\mathbf{X}_I\|_2^2 + \|\mathbf{X}_I\|_2^3\right)\|_{2,P} \leq A\,,$$

con $A = (2\,\|\chi\|_\infty + \|\chi_1\|_\infty)\left(K_3^{1/3} + K_3^{1/2}\right)$. Luego, por el Lema 4.5 tenemos que existe una constante $M > 0$ que no depende de n tal que

$$\mathbb{E}\left[\sqrt{n}\sup_{h\in\mathcal{H}}|(P_n - P)(h)|\right] \leq M\|H_1\|_{2,P}\ J(1,\mathcal{H}) \leq MA\ J(1,\mathcal{H})\,, \tag{6.28}$$

donde $J(\delta,\mathcal{H}) = \sup_{\mathbb{Q}}\int_0^\delta \sqrt{1+\log N(\epsilon\|H_1\|_{2,\mathbb{Q}}, \mathcal{H}, \|\cdot\|_{2,\mathbb{Q}})}d\epsilon$. Sea $B = \sqrt{3}\,(1+\log(13))^{1/2}$, como

$$J(\delta,\mathcal{H}) \leq \int_0^\delta \sqrt{1+3k\log\left(\frac{12+\varepsilon}{\varepsilon}\right)}d\epsilon \leq B\,\sqrt{k}\ \int_0^1 \sqrt{1-\log(\varepsilon)}\,d\varepsilon\,,$$

y $\int_0^1 \sqrt{1-\log(\varepsilon)}\,d\varepsilon < \infty$, el resultado deseado se deduce de (6.28). ■

Demostración del Teorema 6.7. Recordemos que $m_{0,n} = \min\{|\beta_{0,j}| : \beta_{0,j} \neq 0\}$ y sea $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{n,1}, \ldots, \widehat{\beta}_{n,p})$. En primer lugar, notemos que

$$\begin{aligned}\mathbb{P}(\widehat{\beta}_{n,j} = 0 \text{ para algún } 1 \leq j \leq k) &\leq \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 > m_{0,n}) \\ &= \mathbb{P}\left(\sqrt{\frac{n}{k}}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 > \sqrt{\frac{n}{k}}m_{0,n}\right).\end{aligned}$$

Usando el hecho de que $\sqrt{n/k}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ y la hipótesis **N1**, tenemos que

$$\mathbb{P}(\mathcal{A}_n^c) = \mathbb{P}(\widehat{\beta}_{n,j} = 0 \text{ para algún } 1 \leq j \leq k) \to 0\,.$$

Por otra parte, $\mathbb{P}(\mathcal{B}_n) = \mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \to 1$. Por lo tanto, en $\mathcal{B}_n \cap \mathcal{A}_n$, todas las componentes de $\widehat{\boldsymbol{\beta}}_{n,\,I}$ son distintas de cero. Luego, como $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,\,I}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ minimiza $L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta})$, también minimiza $L_n\left((\mathbf{b}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}\right) + I_{\lambda_n}\left((\mathbf{b}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}\right)$ para $\mathbf{b} \in \mathbb{R}^k$. Por lo tanto, tenemos que

$$\mathbf{0}_k = \nabla\left(\frac{1}{n}\sum_{i=1}^n \phi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{n,\,I})\right) + \nabla\left(I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,I})\right) + \mathbf{r}_n\,,$$

donde $\mathbb{P}(\mathbf{r}_n = 0) \to 1$, lo cual implica que si $\mathbf{v} \in \mathbb{R}^k$, $\|\mathbf{v}\|_2 = 1$,

$$0 = \frac{1}{n}\sum_{i=1}^n \mathbf{v}^{\mathrm{T}}\Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{n,\,I})\mathbf{X}_{i,I} + \mathbf{v}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,I}) + \mathbf{v}^{\mathrm{T}}\mathbf{r}_n\,.$$

Dado $\mathbf{b} \in \mathbb{R}^k$, definamos

$$\mathbf{A}_{n,I}(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^n \chi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\mathbf{b})\mathbf{X}_{i,I}\mathbf{X}_{i,I}^{\mathrm{T}}\,.$$

Definamos

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \Psi\left(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\left[\theta\,\widehat{\boldsymbol{\beta}}_{n,\,I} + (1-\theta)\,\boldsymbol{\beta}_{0,I}\right]\right)\mathbf{v}^{\mathrm{T}}\mathbf{X}_{i,I}\,.$$

Luego, usando el Teorema de Valor Medio tenemos que $M_n(1) = M_n(0) + M_n'(\alpha)$, para algún $\alpha \in [0,1]$, es decir,

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_{0,I})\mathbf{v}^{\mathrm{T}}\mathbf{X}_{i,I} + \frac{1}{n}\sum_{i=1}^{n}\chi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_I^*)\mathbf{v}^{\mathrm{T}}\mathbf{X}_{i,I}\mathbf{X}_{i,I}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}) + \mathbf{v}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,I}) + \mathbf{v}^{\mathrm{T}}\mathbf{r}_n \\ &= \frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_{0,I})\mathbf{v}^{\mathrm{T}}\mathbf{X}_{i,I} + \mathbf{v}^{\mathrm{T}}\mathbf{A}_{n,I}(\boldsymbol{\beta}_I^*)(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}) + \mathbf{v}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,I}) + \mathbf{v}^{\mathrm{T}}\mathbf{r}_n\,, \end{aligned} \tag{6.29}$$

donde $\boldsymbol{\beta}_I^* = \alpha\widehat{\boldsymbol{\beta}}_{n,\,I} + (1-\alpha)\boldsymbol{\beta}_{0,I}$ para algún $\alpha \in [0,1]$.

Observemos que $\sqrt{n}\;t_n^{-1}\mathbf{v}^{\mathrm{T}}\mathbf{A}_I(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}) = S_{1,n} + S_{2,n} + S_{3,n}$ donde para reforzar la dependencia en n hemos escrito t_n en lugar de $t = (\mathbf{v}^{\mathrm{T}}\mathbf{B}_I^{(k)}\mathbf{v})^{1/2}$.

$$\begin{aligned} S_{1,n} &= \sqrt{n}\;t_n^{-1}\mathbf{v}^{\mathrm{T}}(\mathbf{A}_I - \mathbf{A}_I(\boldsymbol{\beta}_I^*))(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})\,, \\ S_{2,n} &= \sqrt{n}\;t_n^{-1}\mathbf{v}^{\mathrm{T}}(\mathbf{A}_I(\boldsymbol{\beta}_I^*) - \mathbf{A}_{n,I}(\boldsymbol{\beta}_I^*))(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})\,, \\ S_{3,n} &= \sqrt{n}\;t_n^{-1}\mathbf{v}^{\mathrm{T}}\mathbf{A}_{n,I}(\boldsymbol{\beta}_I^*)(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})\,. \end{aligned}$$

Probaremos que

$$S_{1,n} \xrightarrow{\;p\;} 0\,, \tag{6.30}$$

$$S_{2,n} \xrightarrow{\;p\;} 0\,, \tag{6.31}$$

$$S_{3,n} \xrightarrow{\;D\;} N(0,1)\,. \tag{6.32}$$

Empecemos probando (6.30). Dados $\varepsilon, \delta > 0$ debemos probar que $\mathbb{P}(|S_{1,n}| < \varepsilon) > 1 - \delta$ para n suficientemente grande. Como $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\sqrt{k/n}\right)$, dado $\delta > 0$ existe $C_1 > 0$ tal que $\mathbb{P}(\mathcal{D}_n) > 1 - \delta/4$, donde $\mathcal{D}_n = \{\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \le C_1\sqrt{k/n}\}$.

Observemos que $\chi_1(y,s) = (\partial/\partial s)\chi(y,s)$ es acotada, por **R5**, luego

$$\begin{aligned} |S_{1,n}| &= \left|\sqrt{n}\;t_n^{-1}\mathbf{v}^{\mathrm{T}}\,\mathbb{E}\left[(\chi(Y, \mathbf{X}_I^{\mathrm{T}}\boldsymbol{\beta}_{0,I}) - \chi(Y, \mathbf{X}_I^{\mathrm{T}}\boldsymbol{\beta}_I^*))\mathbf{X}_I\mathbf{X}_I^{\mathrm{T}}\right](\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})\right| \\ &\le \sqrt{n}\;t_n^{-1}\mathbb{E}\left|\chi_1(Y, \mathbf{X}_I^{\mathrm{T}}\boldsymbol{\beta}_I^{**})\mathbf{X}_I^{\mathrm{T}}(\boldsymbol{\beta}_{0,I} - \boldsymbol{\beta}_I^*)\mathbf{v}^{\mathrm{T}}\mathbf{X}_I\mathbf{X}_I^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})\right|\,, \end{aligned}$$

donde $\boldsymbol{\beta}_I^{**} = \alpha_1\boldsymbol{\beta}_I^* + (1-\alpha_1)\boldsymbol{\beta}_{0,I}$ para algún $\alpha_1 \in [0,1]$. Queremos destacar que la esperanza en la última igualdad está tomada solamente con respecto a Y y $\mathbf{X}_I$. Luego, usando que χ_1 es acotada y aplicando el Lema 6.12 a las variables aleatorias $U = (\boldsymbol{\beta}_{0,I} - \boldsymbol{\beta}_I^*)^{\mathrm{T}}\mathbf{X}_I$, $V = \mathbf{v}^{\mathrm{T}}\mathbf{X}_I$ y $W = (\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})^{\mathrm{T}}\mathbf{X}_I$ (tomando $p = q = r = 3$), obtenemos que

$$\begin{aligned} |S_{1,n}| &\le \|\chi_1\|_\infty\,\sqrt{n}\;t_n^{-1}\mathbb{E}\left|(\boldsymbol{\beta}_{0,I} - \boldsymbol{\beta}_I^*)^{\mathrm{T}}\mathbf{X}_I\;\mathbf{v}^{\mathrm{T}}\mathbf{X}_I\;(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})^{\mathrm{T}}\mathbf{X}_I\right| \\ &\le \|\chi_1\|_\infty\,\sqrt{n}\;t_n^{-1}\mathbb{E}\left[|(\boldsymbol{\beta}_{0,I} - \boldsymbol{\beta}_I^*)^{\mathrm{T}}\mathbf{X}_I|^3\right]^{1/3}\mathbb{E}\left[|\mathbf{v}^{\mathrm{T}}\mathbf{X}_I|^3\right]^{1/3}\mathbb{E}\left[|(\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})^{\mathrm{T}}\mathbf{X}_I|^3\right]^{1/3} \\ &\le \|\chi_1\|_\infty\,\sqrt{n}\;t_n^{-1}\mathbb{E}\left[\|\mathbf{X}_I\|_2^3\right]\|\boldsymbol{\beta}_{0,I} - \boldsymbol{\beta}_I^*\|_2\|\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}\|_2\,, \end{aligned}$$

donde, en la última desigualdad, usamos la desigualdad de Cauchy-Schwartz y el hecho que $\|\mathbf{v}\|_2 = 1$.

Luego, en $\mathcal{D}_n$ tenemos que

$$|S_{1,n}| \leq t_n^{-1}\|\chi_1\|_\infty C_1^2 \mathbb{E}\left[\|\mathbf{X}_I\|_2^3\right] \frac{k}{\sqrt{n}} \leq \|\chi_1\|_\infty C_1^2 K_3^{1/2}\, t_n^{-1} \frac{k}{\sqrt{n}}\,,$$

donde K_3 es la constante de la hipótesis **Z6**. Como **Z7** implica que $t_n = \mathbf{v}^{\mathrm{T}}\mathbf{B}_I^{(k)}\mathbf{v} \geq \iota_1\left(\mathbf{B}_I^{(k)}\right) \geq \tau_2$, deducimos que $|S_{1,n}| \leq \|\chi_1\|_\infty C_1^2 K_3^{1/2}\, \tau_2^{-1}\, k/\sqrt{n}$. Luego, como $k^2/n \to 0$ existe n_0 tal que si $n \geq n_0$, $\mathcal{D}_n \subset \{\|S_{1,n}\| \leq \varepsilon\}$ lo que concluye la demostración de (6.30).

A continuación probaremos que $S_{2,n} \xrightarrow{p} 0$. Sea $\mathbf{u}_n = (\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I})/\|\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}\|_2$. Observemos que

$$\begin{aligned} S_{2,n} &= \sqrt{n}\, t_n^{-1}\|\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}\|_2 \left\{ \mathbb{E}\left[\chi(Y, \mathbf{X}_I^{\mathrm{T}}\boldsymbol{\beta}_I^*)\mathbf{v}^{\mathrm{T}}\mathbf{X}_I\mathbf{X}_I^{\mathrm{T}}\mathbf{u}_n\right] - \frac{1}{n}\sum_{i=1}^n \chi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_I^*)\mathbf{v}^{\mathrm{T}}\mathbf{X}_{i,I}\mathbf{X}_{i,I}^{\mathrm{T}}\mathbf{u}_n \right\} \\ &= \sqrt{n}\, t_n^{-1}\|\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}\|_2 (P - P_n)(h_{\mathbf{v},\boldsymbol{\beta}_I^*,\mathbf{u}_n})\,, \end{aligned}$$

donde la función $h_{\mathbf{v}_n,\boldsymbol{\beta}_I^*,\mathbf{u}_n}$ está definida en (6.27). Sean $\varepsilon > 0$ y $\delta > 0$ positivos. Como $\mathbb{P}(\mathcal{B}_n) = \mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,II} = \mathbf{0}_{p-k}) \to 1$, existe n_0 tal que si $n \geq n_0$, $\mathbb{P}(\mathcal{B}_n) > 1 - \delta/4$. Por otra parte, recordemos que $\mathbb{P}(\mathcal{D}_n) > 1 - \delta/4$, donde $\mathcal{D}_n = \{\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq C_1\sqrt{k/n}\}$, con lo cual, si definimos el evento $\widetilde{\mathcal{D}_n} = \{\|\widehat{\boldsymbol{\beta}}_{n,\,I} - \boldsymbol{\beta}_{0,I}\|_2 \leq C_1\sqrt{k/n}\}$ tenemos que $\mathcal{B}_n \cap \mathcal{D}_n \subset \widetilde{\mathcal{D}_n}$ de donde $\mathbb{P}(\widetilde{\mathcal{D}_n}) > 1 - \delta/2$. Definamos además el evento

$$\mathcal{C}_n = \left\{ \sup_{h \in \mathcal{H}} |(P_n - P)(h)| < \frac{2C}{\delta}\sqrt{\frac{k}{n}} \right\}\,,$$

donde $\mathcal{H}$ está definido en (6.27) y C es la constance del Lema 6.13. Aplicando la desigualdad de Markov, tenemos que $\mathbb{P}(\mathcal{C}_n) > 1 - \delta/2$, por lo que $\mathbb{P}(\widetilde{\mathcal{D}_n} \cap \mathcal{C}_n) > 1 - \delta$.

Sea $n_0 \in \mathbb{N}$ tal que, para $n \geq n_0$, $C_1\sqrt{k/n} < 1$. Luego, en el evento $\widetilde{\mathcal{D}_n} \cap \mathcal{C}_n$, obtenemos que

$$|S_{2,n}| \leq \sqrt{n}\, t_n^{-1} C_1 \sqrt{\frac{k}{n}}\, \frac{2C}{\delta}\sqrt{\frac{k}{n}} \leq \frac{C_3}{\delta}\frac{k}{\sqrt{n}}\,,$$

donde $C_3 = 2CC_1/\tau_2$ y usamos nuevamente que $t_n \geq \tau_2$. Por último, el hecho de que $k^2/n \to 0$ implica que, existe $n_1 \geq n_0$ tal que, $\widetilde{\mathcal{D}_n} \cap \mathcal{C}_n \subset \{|S_{2,n}| \leq \varepsilon\}$ para $n \geq n_1$ lo que muestra que (6.30) vale.

Finalmente, debemos probar que $S_{3,n} \xrightarrow{D} N(0,1)$. Usando (6.29) tenemos que

$$\begin{aligned} S_{3,n} &= -\sqrt{n}\, t_n^{-1}\frac{1}{n}\sum_{i=1}^n \mathbf{v}^{\mathrm{T}}\Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_{0,I})\mathbf{X}_{i,I} - \sqrt{n}\, t_n^{-1}\mathbf{v}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,I}) - \sqrt{n}\, t_n^{-1}\mathbf{v}_n^{\mathrm{T}}\mathbf{r}_n \\ &= S_{31,n} + S_{32,n} + S_{33,n}\,. \end{aligned}$$

En virtud de (6.10), **Z7** y el hecho que $\mathbb{P}(\mathbf{r}_n = 0) \to 1$, es fácil ver que $S_{32,n} \xrightarrow{p} 0$ y $S_{33,n} \xrightarrow{p} 0$.

Probaremos ahora que $S_{31,n} \xrightarrow{D} N(0,1)$. Definamos

$$W_{n,i} = -t_n^{-1}\frac{\Psi(Y_i, \mathbf{X}_{i,I}^{\mathrm{T}}\boldsymbol{\beta}_{0,I})}{\sqrt{n}}\mathbf{v}^{\mathrm{T}}\mathbf{X}_{i,I}\,.$$

Luego, $S_{31,n} = \sum_{i=1}^n W_{n,i}$. Notemos que $\mathbb{E}W_{n,i} = 0$ para todo $n \in \mathbb{N}$ y $1 \leq i \leq n$, mientras que

$$\mathbb{E}W_{n,i}^2 = \textsc{Var}(W_{n,i}) = \frac{1}{n}\, t_n^{-2}\, \mathbf{v}^{\mathrm{T}}\mathbb{E}\left[\Psi^2(Y, \mathbf{X}_I^{\mathrm{T}}\boldsymbol{\beta}_{0,I})\mathbf{X}_I\mathbf{X}_I^{\mathrm{T}}\right]\mathbf{v} = \frac{1}{n}\, t_n^{-2}\, \mathbf{v}^{\mathrm{T}}\mathbf{B}_I^{(k)}\mathbf{v} = \frac{1}{n}\,,$$

lo cual implica que $\sum_{i=1}^n \mathbb{E}W_{n,i}^2 = 1$.

Para aplicar el Teorema Central del Límite para arreglos triangulares, se puede verificar la condición de Lindeberg, es decir, que para todo $\varepsilon > 0$, $\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}\left[W_{n,i}^2 \mathbb{I}_{|W_{n,i}|>\varepsilon}\right] = 0$ o la condición de Lyapunov, o sea, que para algún $\delta > 0$

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}\left[|W_{n,i}|^{2+\delta}\right] = 0\,,$$

que será la que probaremos. Observemos que

$$\mathbb{E}|W_{n,i}|^{2+\delta} = \frac{1}{t_n^{2+\delta}} \frac{1}{n^{1+\frac{\delta}{2}}} \mathbb{E}\left[|\Psi(Y, \mathbf{X}_I^{\mathrm{T}}\boldsymbol{\beta}_{0,I})|^{2+\delta} |\mathbf{v}^{\mathrm{T}}\mathbf{X}_I|^{2+\delta}\right]\,,$$

de donde, usando que $\|\mathbf{v}\|_2 = 1$, Ψ es acotada y la desigualdad de Cauchy-Schwartz, deducimos que

$$\sum_{i=1}^n \mathbb{E}[|W_{n,i}|^{2+\delta}] \leq \frac{1}{t_n^{2+\delta}} \frac{1}{n^{\frac{\delta}{2}}} \|\Psi\|_\infty^{2+\delta} \mathbb{E}\|\mathbf{X}_I\|^{2+\delta} \leq \frac{1}{n^{\frac{\delta}{2}}} \frac{1}{\tau_2^{2+\delta}} \|\Psi\|_\infty^{2+\delta} K_3^{\frac{2+\delta}{6}}\,,$$

donde la última desigualdad se deduce del hecho que t_n^{-1} es acotado y de la hipótesis **Z6**. Es decir, que la condición de Lyapunov se cumple, por lo tanto, usando el Teorema Central del Límte de Lindeberg–Feller para arreglos triangulares, concluimos que $S_{31,n} \xrightarrow{D} N(0,1)$. Finalmente, usando el Teorema de Slutsky, obtenemos el resultado buscado. ■

Demostración del Corolario 6.8. Bastará ver que las penalizaciones SCAD y MCP cumplen (6.10) bajo las condiciones **N2** y **N3**. Sea $a > 0$ el segundo parámetro de ajuste de las penalizaciones SCAD y MCP, que suponemos fijo y recordemos que, para cualquiera de estas dos penalidades podemos escribir $I_{\lambda_n}(\boldsymbol{\beta}) = \sum_{j=1}^p J_{\lambda_n}(|\beta_j|)$ donde $J_{\lambda_n}(t)$ es constante en $[a\lambda_n, \infty)$ y $J_{\lambda_n}(0) = 0$. Por lo tanto, para cualquier $\mathbf{b} \in \mathbb{R}^k$, $I_{\lambda_n}\left((\mathbf{b}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}\right) = \sum_{j=1}^k J_{\lambda_n}(|b_j|)$ y $\nabla I_{\lambda_n}(\mathbf{b}) = \sum_{j=1}^k J'_{\lambda_n}(|b_j|)$.

Como $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\sqrt{k/n}\right)$, dado $\delta > 0$ existe $C_1 > 0$ tal que $\mathbb{P}(\mathcal{D}_n) > 1 - \delta$, donde $\mathcal{D}_n = \{\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq C_1\sqrt{k/n}\}$.

Observemos que para todo $\omega \in \mathcal{D}_n$ y $1 \leq j \leq k$,

$$|\widehat{\beta}_{n,j}| \geq |\beta_{0,j}| - |\widehat{\beta}_{n,j} - \beta_{0,j}| \geq m_{0,n} - C_1\sqrt{\frac{k}{n}} = \lambda_n \left(\frac{m_{0,n}}{\lambda_n} - C_1\sqrt{\frac{k}{n}}\frac{1}{\lambda_n}\right)\,.$$

La hipótesis **N3** implica que existe $C_2 > 0$ tal que $k/n \leq C_2\lambda_n^2$, mientras que por **N2** existe $n_0 \in \mathbb{N}$ tal que $(m_{0,n}/\lambda_n) > a + 1 + C_1\sqrt{C_2}$, si $n \geq n_0$. Luego, si $n \geq n_0$, se cumple que, para todo $j = 1, \ldots, k$,

$$|\widehat{\beta}_{n,j}| \geq \lambda_n \left(\frac{m_{0,n}}{\lambda_n} - C_1\sqrt{C_2}\right) > a\lambda_n\,.$$

Finalmente, como $J_{\lambda_n}(t)$ es constante en $[a\lambda_n, \infty)$, se tiene que $\mathcal{D}_n \subset \{\|\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,I})\|_2 = 0\}$, lo que muestra que la condición (6.10) se verifica para estas penalizaciones. ■

Capítulo 7

Algoritmo y resultados computacionales

En este Capíıtulo, presentamos los resultados de un estudio de Monte Carlo diseñado para comparar el comportamiento de los estimadores clásicos y robustos penalizados para muestras finitas bajo distintos tipos de contaminación. La Sección 7.1 describe el algoritmo usado para calcular los $M-$estimadores penalizados, mientras que los modelos utilizados, los distintos esquemas de contaminación así como los resultados obtenidos se presentan en la Sección 7.2. Finalmente, en la Sección 7.3 se ilustra la utilidad de usar procedimientos robustos para proveer métodos de inferencia confiables.

7.1. El algoritmo

El procedimiento general para computar los estimadores propuestos es una implementación del algoritmo de descenso cíclico. Supongamos que se quieren obtener soluciones del problema de minimización (3.19) para una grilla de parámetros de ajuste $\widetilde{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ ordenados de mayor a menor, es decir, $\lambda_s > \lambda_{s+1}$ para $s = 1, \ldots, K-1$. Sobre esta grilla se elegirá el parámetro de regularización usando convalidación cruzada como se describió en la Sección 3.2.2. Dado $\lambda_s \in \widetilde{\Lambda}$ y un conjunto $\mathcal{I}_s \subset \{1, \ldots, p\}$, el algoritmo consiste en los siguientes pasos:

(a) Obtenga un estimador inicial $\widehat{\boldsymbol{\beta}}_{\text{INI}}$ y defina $M^{(0)} = M_n(\widehat{\boldsymbol{\beta}}_{\text{INI}})$, donde $M_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) + I_\lambda(\boldsymbol{\beta})$ está dado en (3.19). Fije inicialmente $\ell = 0$.

(b) $\ell \leftarrow \ell + 1$

Paso 1 Elija una permutación aleatoria del conjunto $\mathcal{I}_s$. Siguiendo el orden definido por los índices de esa permutación, minimice la función $M_n(\boldsymbol{\beta})$ respecto de una única coordenada de $\boldsymbol{\beta}$ por vez, dejando fijas las restantes. Esto involucra $\#(\mathcal{I}_s)$ problemas de minimización univariados. Denotamos como $\widetilde{\boldsymbol{\beta}}$ el vector estimado luego de pasar por todas las coordenadas de $\mathcal{I}_s$.

Paso 2 Obtenga el valor $c > 0$ que minimiza la función $M_n(c\widetilde{\boldsymbol{\beta}})$. Denotemos por $\widetilde{c}$ a dicho valor, $\boldsymbol{\beta}^{(\ell)} = \widetilde{c}\widetilde{\boldsymbol{\beta}}$ y $M^{(\ell)} = M_n(\boldsymbol{\beta}^{(\ell)})$.

Paso 3 Calcule $R^{(\ell)} = |M^{(\ell-1)} - M^{(\ell)}|/M^{(\ell)}$.

(c) Si el cociente $R^{(\ell)}$ es menor que un parámetro de tolerancia (prefijado), defina $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(\ell)}$. Si esto no ocurre, vuelva al paso (b).

Cuando el modelo incluye una ordenada el origen γ, se considera un paso intermedio entre los Pasos 2 y 3, en el cual se minimiza sobre $\mathbb{R}$ la función objetivo dejando fijo el valor de $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(\ell)}$ obtenido en el Paso 2.

En los resultados numéricos presentados en las Secciones 7.2 y 7.3, las optimizaciones univariadas de los Pasos 1 y 2 (y eventualmente también la del paso que permite calcular el estimador de la ordenada al origen) se llevaron a cabo usando la rutina `optim` en `R`, la cual implementa el método de optimización de Nelder-Mead. Varias partes del código fueron implementadas en `C++` e integradas con el paquete `Rcpp` en `R`.

7.1.1. Obtención de los conjuntos $\mathcal{I}_s$

Los conjuntos $\mathcal{I}_1, \ldots, \mathcal{I}_K$ son obtenidos secuencialmente como se describe a continuación. En primer lugar, definimos $\mathcal{I}_1 = \{1, \ldots, p\}$. Para $s = 2, \ldots, K$, definimos el conjunto $\mathcal{I}_s = \mathcal{I}_{1,s} \cup \mathcal{I}_{2,s}$ donde $\mathcal{I}_{1,s} = \{j : \widehat{\boldsymbol{\beta}}_j^{(s-1)} \neq 0\}$, siendo $\widehat{\boldsymbol{\beta}}_j^{(s-1)}$ la $j-$ésima coordenada del estimador computado con el parámetro de penalización λ_{s-1}. Por otra parte, cuando la penalización I_λ es Lipschitz, $\mathcal{I}_{2,s}$ se define como

$$\mathcal{I}_{2,s} = \left\{ j : \widehat{\boldsymbol{\beta}}_j^{(s-1)} = 0 \quad \text{y} \quad 0 \in \nabla L_n\left(\widehat{\boldsymbol{\beta}}^{(s-1)}\right)_j + tol. \left[\overline{\nabla} I_{\lambda_s}\left(\widehat{\boldsymbol{\beta}}^{(s-1)}\right)_j\right] \right\},$$

donde $\overline{\nabla}$ es el gradiente generalizado definido en Clarke (1975) y *tol* es un parámetro de tolerancia prefijado. Por ejemplo, al usar la penalización Signo, este conjunto se define como

$$\mathcal{I}_{2,s} = \left\{ j : \widehat{\boldsymbol{\beta}}_j^{(s-1)} = 0 \quad \text{y} \quad \left\|\widehat{\boldsymbol{\beta}}^{(s-1)}\right\|_2 \left|\nabla L_n\left(\widehat{\boldsymbol{\beta}}^{(s-1)}\right)_j\right| \geq tol.\, \lambda \right\}.$$

Este procedimiento es una versión heurística de las llamadas *Strong Safe Rules* definidas en Tibshirani *et al.* (2012). Vale la pena mencionar que restringir los problemas de minimización univariada al conjunto $\mathcal{I}_s$ es importante para mejorar el tiempo de cómputo de los estimadores.

7.1.2. Obtención del estimador inicial $\widehat{\boldsymbol{\beta}}_{\text{INI}}$

Un punto importante del algoritmo es la elección del estimador inicial $\widehat{\boldsymbol{\beta}}_{\text{INI}}$. Chi y Scott (2014) basan su estimador inicial en las condiciones de Karush-Kuhn-Tucker (KKT) para el siguiente problema de minimización:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - F(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\right)^2 + \lambda\|\boldsymbol{\beta}\|_1 .$$

Para $j = 1, \ldots, p$, estos autores calculan los *scores* $z_j = \left|\overline{Y}(1-\overline{Y})\mathbf{X}_{(j)}^{\mathrm{T}}(\mathbf{Y} - \overline{Y}\mathbf{1}_n)\right|$, donde $\mathbf{1}_n \in \mathbb{R}^n$ es vector con todas sus coordenadas iguales a 1, $\overline{Y} = \sum_{i=1}^n Y_i/n$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ y $\mathbf{X}_{(j)}$ es

la $j-$ésima columna de la matriz de diseño, es decir, $\mathbf{X}_{(j)} = (X_{1j}, \ldots, X_{nj})^{\mathrm{T}}$. Chi y Scott (2014) obtienen el estimador inicial fijando como 1 las coordenadas correspondientes a las variables con mayores *scores* (en valor absoluto) y como 0 a las restantes. Cabe señalar que este estimador inicial únicamente tiene sentido cuando las covariables están estandarizadas.

Este estimador inicial funciona relativamente bien cuando la muestra no tiene datos atípicos. Sin embargo, puede dar origen a un mal estimador inicial si hay datos mal clasificados. Para solucionar este problema, en nuestro algoritmo primero calculamos las cantidades $\kappa_{ij} = X_{ij}(Y_i - \overline{Y})$ y, para todo $j = 1, \ldots, p$, el *score* $\widetilde{z}_j$ se computa como la media $\alpha-$podada de $\{\kappa_{1j}, \kappa_{2j}, \ldots, \kappa_{nj}\}$. En las simulaciones y aplicaciones de este capítulo, se fijó $\alpha = 0.15$. Finalmente, dado $0 < \widetilde{\alpha} < 1$, seleccionamos la proporción $\widetilde{\alpha}$ de variables que tienen mayor valor abosluto del *score* y aplicamos el $M-$estimador pesado definido en (3.8) sobre dichas variables. Para los resultados numéricos realizados, se usó la proporción $\widetilde{\alpha} = 0.1$.

7.1.3. Obtención de la grilla de parámetros de regularización $\widetilde{\Lambda}$

El conjunto $\widetilde{\Lambda}$ de posibles candidatos de parámetros de ajuste fue obtenida como se describe a continuación. En primer lugar, definimos

$$\lambda^* = \frac{2}{n}\overline{\mathbf{Y}}(1 - \overline{\mathbf{Y}}) \max_{j \in \{1,\ldots,p\}} |\mathbf{X}_{(j)}^{\mathrm{T}}\mathbf{Y}|$$

donde, como en la sección anterior, $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\overline{\mathbf{Y}} = (\sum_{i=1}^n Y_i)/n$ y $\mathbf{X}_{(j)} = (X_{1j}, \ldots, X_{nj})^{\mathrm{T}}$. λ^* es el mínimo valor del parámetro de penalización para el cual $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ al usar la pérdida de mínimos cuadrados definida en (3.3) y la penalización LASSO. En los estudios numéricos que reportamos, tomamos $\lambda_{\text{máx}} = 2\lambda^*$ y $\lambda_{\text{mín}} = \tau\lambda_{\text{máx}}$, donde τ es una constante prefijada que se eligió igual a 0.2. La grilla $\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$ se obtiene de forma tal que sus logaritmos estén equiespaciados entre $\log(\lambda_{\text{mín}})$ y $\log(\lambda_{\text{máx}})$. En el estudio de simulación reportado en la Sección 7.2, se fijó $K = 12$.

De todos modos, dado un problema particular, es aconsejable verificar que la grilla obtenida sea adecuada. Por ejemplo, si el parámetro de ajuste finalmente elegido es uno de los extremos $\lambda_{\text{mín}}$ o $\lambda_{\text{máx}}$, es aconsejable ampliar la grilla en el extremo correspondiente.

Como mencionamos anteriormente, la elección del parámetro de penalización es un punto clave para el cómputo efectivo del estimador. Este hecho y la necesidad de utilizar un criterio de convalidación cruzada robusta se describieron en la Sección 3.2.2. Por esta razón, en este Capítulo, el parámetro λ se eligió mediante el criterio de convalidación cruzada tradicional dado en (3.21) al usar los estimadores clásicos, mientras que al considerar los $M-$estimadores robustos utilizamos la versión robusta de dicho criterio dada en (3.22).

7.2. Estudio de Monte Carlo

7.2.1. Los modelos estudiados

En esta sección, presentamos los resultados de un estudio de simulación para comparar el rendimiento empírico de las distintas funciones de pérdida y penalizaciones consideradas en esta tesis. Para

ello, en cada replicación y para diferentes posibles valores del tamaño muestral n, la dimensión de las covariables p y del parámetro de regresión $\boldsymbol{\beta}_0$, generamos observaciones i.i.d. $(\mathbf{X}_i, Y_i)$, $1 \leq i \leq n$, $\mathbf{X}_i \in \mathbb{R}^p$, $Y_i \sim Bi(1, F(\gamma_0 + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0))$ donde F es la función logística definida en (2.1) y la ordenada al origen $\gamma_0 = 0$. La distribución de las variables explicativas es $\mathbf{X}_i \sim N_p(0, \mathbf{I})$ para muestras sin contaminar. De ahora en más esa situación se indicará como **C0**.

Exploramos el comportamiento para distintos tamaños de muestra n y dimensiones p. Más precisamente, consideramos pares (n, p), donde $n \in \{150, 300\}$ y $p \in \{40, 80, 120\}$. En particular, el caso $p = 120$ y $n = 150$ plantea un gran desafío. Para generar un escenario ralo tomamos el parámetro de regresión verdadero igual a $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, 0, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$, es decir, las primeras cinco componentes son iguales a uno y las restantes son cero, obteniéndose valores de $\mathbb{E}(Y_i)$ iguales 0.50. En todos los casos, el número de replicaciones se tomó igual a $NR = 500$.

Para estudiar el efecto de datos atípicos sobre el comportamiento de los estimadores, consideramos dos esquemas de contaminación donde agregamos una proporción ε de datos atípicos. En el primer esquema, que corresponde a outliers de clase **A**, generamos puntos mal clasificados $(\widetilde{Y}, \widetilde{\mathbf{X}})$, con $\widetilde{\mathbf{X}} \sim N_p(0, 20\,\mathbf{I})$ y

$$\widetilde{Y} = \begin{cases} 1 & \text{si } \gamma_0 + \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}_0 < 0 \\ 0 & \text{si } \gamma_0 + \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}_0 \geq 0. \end{cases} \tag{7.1}$$

En el segundo escenario de contaminación, obtuvimos outliers de clase **B** como en Croux y Haesbroeck (2003), es decir, dado $m > 0$, fijamos $\widetilde{\mathbf{W}} = m\sqrt{p}\,\boldsymbol{\beta}_0/5$ y tomamos $\widetilde{\mathbf{X}} = \widetilde{\mathbf{W}} + \widetilde{\mathbf{U}}$, donde $\widetilde{\mathbf{U}} \sim N_p(\mathbf{0}, \mathbf{I}/100)$ es un ruido blanco introducido de modo a obtener valores distintos de las variables explicativas. La respuesta $\widetilde{Y}$, asociada a $\widetilde{\mathbf{X}}$, se tomó siempre igual a 0. Vale la pena observar que $\widetilde{\mathbf{W}}^{\mathrm{T}}\boldsymbol{\beta}_0 \approx m\sqrt{p}$, por lo que la palanca de los puntos crece con m. En nuestro estudio numérico, evaluamos la performance de los estimadores para $m = 0.5, 1, 2, 3, 4$ y 5.

Consideramos los siguientes escenarios de contaminación

- **CA1:** se agregó una proporción $\varepsilon = 0.05$ de datos atípicos de clase **A**.
- **CA2:** se agregó una proporción $\varepsilon = 0.10$ de datos atípicos de clase **A**.
- **CB1:** se agregó una proporción $\varepsilon = 0.05$ de datos atípicos de clase **B**.
- **CB2:** se agregó una proporción $\varepsilon = 0.10$ de datos atípicos de clase **B**.

7.2.2. Los estimadores considerados

Comparamos el comportamiento de los estimadores de máxima verosimilitud, es decir, cuando $\rho(t) = t$, que se indican con el subíndice MV en todas las tablas y figuras, con aquellos que acotan la *deviance* y también con sus versiones robustas pesadas. Las tres funciones de pérdida consideradas son $\rho(t) = 1 - \exp(-t)$ que da origen a los estimadores de mínimos cuadrados, la función ρ_c introducida por Croux and Haesbroeck (2003), dada en (3.5), y $\rho(t) = \rho_{\mathrm{DIV}}(t) = (c+1)(1+\exp(-ct))$ que corresponde a la pérdida de los estimadores de divergencia. Para estas dos últimas perdidas, la constante de calibración es $c = 0.5$. Estos estimadores se indican con el subíndice MC, M y DIV, respectivamente.

Consideramos además una versión pesada de los estimadores anteriores, que se denota WMV, WMC, WM o WDIV, de acuerdo a la función ρ considerada. Para obtener los pesos, definimos como en la Sección 3.1.2, el cuadrado de la distancia de Mahalanobis $D^2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$. Tomamos los pesos $w(\mathbf{x}) = W(D^2(\mathbf{x}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}^{-1}))$, donde para obtener medidas de palanca robustas $\widehat{\boldsymbol{\mu}}$ es la mediana espacial de $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $\widehat{\boldsymbol{\Sigma}}^{-1}$ es un estimador de la matriz de precisión $\boldsymbol{\Sigma}^{-1}$ calculado usando Graphical LASSO, como se describe a continuación. La función W es una función de peso de tipo *hard rejection* igual a $W(t) = \mathbb{I}_{[0,c_w]}(t)$. La constante c_w es adaptiva y se basa en los cuantiles de $d_i^2 = D^2(\mathbf{X}_i, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}^{-1})$. Más precisamente, calculamos los cuantiles 0.25 y 0.75 de $d_1^2, \ldots, d_n^2$ denotados q_1 y q_3, respectivamente y definimos el valor de corte $c_w = q_3 + 1.5(q_3 - q_1)$.

Este procedimiento depende de la elección del estimador de $\boldsymbol{\Sigma}^{-1}$ que puede ser más complejo a medida que la dimensión aumenta. Por ello, utilizamos el procedimiento propuesto en Öllerer y Croux (2015) y Tarr *et al.* (2016). Sea $\boldsymbol{\Sigma}_{ij} = \sigma_i\sigma_j\rho_{ij}$, donde $\rho_{ii} = 1$. Para estimar σ_j que representa una medida de dispersión de la $j-$ésima variable explicativa, usamos la mediana de las desviaciones absolutas (MAD) de la $j-$ésima coordenada de las observaciones, es decir, la MAD de $\{X_{1j}, \ldots, X_{nj}\}$, siendo $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$. Para estimar ρ_{ij} podemos utilizar el estimador τ de Kendall o el estimador ρ de Spearman. Luego, se define la matriz $\widehat{\boldsymbol{\Sigma}}$ elemento a elemento como $\widehat{\boldsymbol{\Sigma}}_{ij} = \widehat{\sigma}_i\widehat{\sigma}_j\widehat{\rho}_{ij}$. Finalmente, aplicamos el estimador Graphical LASSO definido en Friedman *et al.* (2008) a la matriz $\widehat{\boldsymbol{\Sigma}}$ para obtener $\widehat{\boldsymbol{\Sigma}}^{-1}$.

Para cada pérdida ρ, consideramos distintas funciones de penalización: las penalidades LASSO, Signo y MCP, indicadas con el superíndice L, S y MCP, respectivamente. Los estimadores tradicionales no ralos se calculan sin penalizar los coeficientes y se indican sin superíndice . Todos los estimadores se calcularon usando el algoritmo descripto en la Sección 7.1.

Bajo **C0** y los escenarios **CA1** y **CA2** comparamos todos los estimadores descriptos. Sin embargo, en vista de los resultados obtenidos para estas tres situaciones y para simplificar la presentación, bajo **CB1** y **CB2**, solo reportamos los resultados obtenidos con los estimadores basados en la deviance (MV) y a los basados en $\rho = \rho_c$ con ρ_c dada en (3.5). Para esta última, se consideraron el $M-$estimador con pesos iguales a 1 y el estimador pesado usando la distancia de Mahalanobis tal como es describió más arriba y solo se consideraron las penalidades Signo y MCP. Por otra parte, los valores de (n, p) se tomaron en el conjunto $\{(150, 40), (150, 80), (300, 80), (300, 120)\}$.

7.2.3. Las medidas resumen

Para evaluar el desempeño de cada estimador consideramos cinco medidas resumen. En lo que sigue, $\mathcal{T} = \{(Y_{i,\mathcal{T}}, \mathbf{X}_{i,\mathcal{T}}), i = 1, \ldots, m\}$, $m = 100$, es una nueva muestra generada en forma independiente de la muestra de entrenamiento $\mathcal{M} = \{(Y_i, \mathbf{X}_i), i = 1, \ldots, n\}$. La distribución de $(Y_{i,\mathcal{T}}, \mathbf{X}_{i,\mathcal{T}})$ es la dada por **C0**. Dados estimadores $\widehat{\boldsymbol{\beta}}$ del parámetro de regresion y $\widehat{\gamma}$ de la ordenada al origen calculados con la muestra de entrenamiento $\mathcal{M}$, llamaremos $\widehat{Y}_{i,\mathcal{T}} = \mathbb{I}_{\{\mathbf{X}_{i,\mathcal{T}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{\gamma} > 0\}}$ y $\Pi = \mathbb{P}(Y_{i,\mathcal{T}} = \mathbb{I}_{\{\mathbf{X}_{i,\mathcal{T}}^{\mathrm{T}}\boldsymbol{\beta}_0 + \gamma_0 > 0\}})$. Definimos entonces las medidas resumen

- **Error Cuadrático Medio de las Probabilidades**

$$\mathrm{PECM} = \frac{1}{m}\sum_{i=1}^{m}(F(\mathbf{X}_{i,\mathcal{T}}^{\mathrm{T}}\boldsymbol{\beta}_0 + \gamma_0) - F(\mathbf{X}_{i,\mathcal{T}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \widehat{\gamma}))^2\,,$$

- **Proporción de Clasificaciones Correctas**

$$\text{PCC} = \frac{1}{\Pi} \frac{\#\{i : \widehat{Y}_{i,\mathcal{T}} = Y_{i,\mathcal{T}}\}}{m},$$

- **Error Cuadrático de Beta**

$$\text{ECM} = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2,$$

- **Proporción de Verdaderos Positivos**

$$\text{PVP} = \frac{\#\{j : 1 \leq j \leq p,\ \beta_{0,j} \neq 0 \text{ y } \widehat{\beta}_j \neq 0\}}{\#\{j : 1 \leq j \leq p,\ \beta_{0,j} \neq 0\}},$$

- **Proporción de Verdaderos Nulos**

$$\text{PVN} = \frac{\#\{j : 1 \leq j \leq p,\ \beta_{0,j} = 0 \text{ y } \widehat{\beta}_j = 0\}}{\#\{j : 1 \leq j \leq p,\ \beta_{0,j} = 0\}}.$$

7.2.4. Estudio del comportamiento de la convalidación cruzada

En esta sección, reportamos los resultados de un estudio de simulación realizado para ilustrar la importancia de elegir el parámetro λ mediante un procedimiento robusto cuando se utilizan estimadores robustos, para asegurar la estabilidad de la estimación obtenida.

Para estudiar la performance de los estimadores consideraremos muestras sin contaminar y el escenario **CB1** con $m = 4$. Se evaluó el comportamiento de los estimadores $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{WM}}$ y $\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{WM}}$ cuando el parámetro de regularización se elige minimizando el método clásico de convalidación cruzada $CV(\lambda)$ o el método de convalidación cruzada robusta $RCV(\lambda)$ definidos en la Sección 3.2.2. Los tamaños de muestra y la dimensión de las covariables se eligieron tomando los valores $(n, p) = (150, 40), (150, 80), (300, 80)$ y $(300, 120)$. La Tabla 7.1 muestra las medias 10 % podadas de las medidas PECM, ECM, PVP y PVN para cada método sobre el escenario **C0** y cada uno de los pares (n, p) considerados, mientras que en la Tabla 7.2 se muestran los resutados para el escenario **CB1**.

En la Tabla 7.1 se puede observar que cuando no hay contaminación, los estimadores obtenidos utilizando CV o RCV dan resultados muy similares, incluso en algunos casos se observa una leve ventaja del comportamiento de los estimadores que utilizan el parámetro de regularización adaptivo elegido por el procedimiento robusto. Por ejemplo, los estimadores $\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{WM}}$ muestran leves mejoras en la medidas PECM y PVP cuando $n = 150$ al usar RCV.

La ventaja sustancial de usar el procedimiento de convalidación cruzada robusta por sobre el clásico puede verse en la Tabla 7.2. La proporción de verdaderos positivos (PVP) se ve fuertemente afectada cuando se utiliza convalidación cruzada clásica aún cuando se utilicen estimadores robustos, lo que muestra el importante rol que juega el procedimiento de selección de λ para asegurar la resistencia final del estimador ante la presencia de datos atípicos. Un ejemplo del efecto de las observaciones mal clasificadas introducidas artificalmente sobre el método CV puede observarse al utilizar el estimador $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{WM}}$ donde la proporción de verdaderos positivos es menor a 0.2 para varios de los pares (n, p) considerados, mientras que utilizando RCV, los valores de PVP son siempre mayores a 0.85. Las mismas conclusiones pueden aplicarse a los otros $M-$estimadores pesados. En general,

	n	150					300			
		CV		*RCV*			*CV*		*RCV*	
	p	40	80	40	80		80	120	80	120
PECM	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.033	0.046	0.029	0.038	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.010	0.011	0.010	0.011
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.022	0.028	0.022	0.027	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.006	0.008	0.007	0.008
ECM	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	1.708	2.239	1.642	2.014	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.501	0.546	0.507	0.539
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	1.144	1.377	1.150	1.414	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.318	0.348	0.327	0.359
PVP	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.926	0.906	0.966	0.944	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	1.000	1.000	1.000	1.000
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.932	0.924	0.942	0.936	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	1.000	1.000	1.000	1.000
PVN	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.968	0.965	0.949	0.963	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.961	0.955	0.960	0.953
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.988	0.971	0.986	0.970	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.977	0.972	0.977	0.971

Tabla 7.1: Promedio podado al 10 % de las medidas PECM, ECM, PVP y PVN en el escenario **C0**

	n	150					300			
		CV		*RCV*			*CV*		*RCV*	
	p	40	80	40	80		80	120	80	120
PECM	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.097	0.095	0.035	0.051	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.090	0.090	0.010	0.012
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.109	0.107	0.023	0.039	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.105	0.109	0.009	0.011
ECM	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	4.191	4.160	2.023	2.542	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	4.089	4.097	0.519	0.558
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	4.969	4.922	1.300	2.157	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	4.804	5.000	0.435	0.509
PVP	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.260	0.272	0.948	0.831	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.264	0.252	1.000	1.000
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.084	0.140	0.956	0.876	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.107	0.606	1.000	1.000
PVN	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.972	0.969	0.926	0.957	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.972	0.972	0.954	0.950
	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.954	0.960	0.971	0.960	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.967	0.947	0.972	0.968

Tabla 7.2: Promedio podado al 10 % de las medidas PECM, ECM, PVP y PVN en el escenario **CB1** con $m = 4$.

los $M-$estimadores pesados obtenidos utilizando el procedimiento clásico de convalidación cruzada resultan excesivamente ralos. Como es de suponer, este hecho afecta severamente a las medidas

PECM y ECM. Para todos los casos, estas medidas son menores cuando se utiliza RCV. Estas diferencias son mucho más significativas cuando se utilizan métodos pesados. Para estos estimadores, cuando $n = 150$, las medidas PECM y ECM son por lo menos el doble cuando se usa CV en vez de la alternativa robusta. Por otro lado, cuando $n = 300$, esta relación es mucho mayor: dichas medidas son aproximadamente 10 veces más grandes al utilizar CV en lugar de RCV, siendo asimismo casi 10 veces superiores a las obtenidas bajo **C0**.

Los resultados obtenidos muestran que la presencia de datos atípicos afecta la elección del parámetro de ajuste λ si se utiliza el criterio clásico CV. Por esta razón, para estudiar el efecto producido en dicho parámetro, la Figura 7.1 muestra superpuestos los estimadores de la densidad de los valores λ elegidos para cada estimador y para cada método de convalidación cruzada.

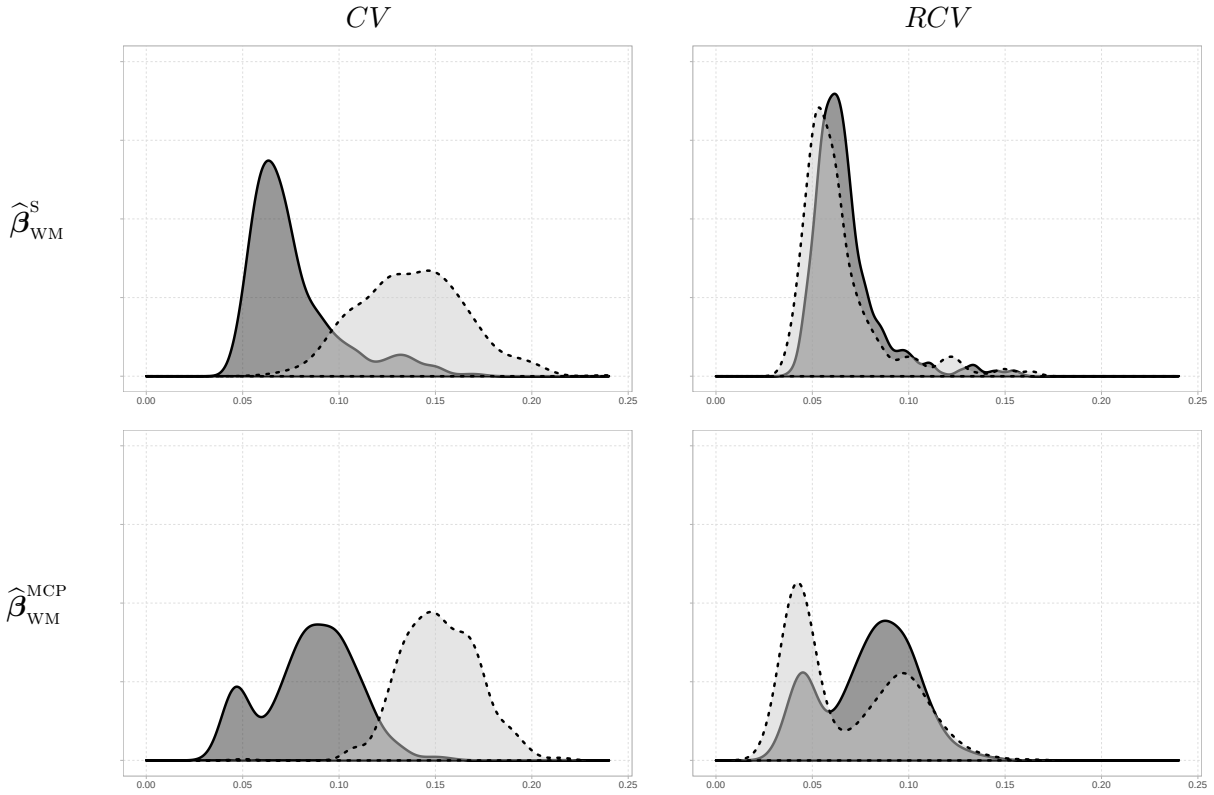


Figura 7.1: Estimadores de la densidad de los valores seleccionados del parámetro de ajuste λ para cada estimador cuando $(n, p) = (150, 80)$. A la izquierda figuran las densidades utilizando convalidación cruzada clásica, mientras que a la derecha se muestran los resultados con RCV. Las curvas sólidas rellenas con gris oscuro corresponden a muestras sin contaminar, mientras que las curvas punteadas rellenas con gris claro corresponden al escenario **CB1** con $m = 4$.

Como era de esperar, al combinar un procedimiento de estimación robusto con el método de convalidación cruzada robusto asociado, los valores de λ obtenidos minimizando RCV se mantienen estables dando origen a densidades semejantes tanto para muestras sin datos atípicos y con ellos (ver Figura 7.1). En cambio, al elegir el valor de λ que minimiza el procedimiento clásico CV, la contaminación afecta severamente a la selección del parámetro de ajuste cuya densidad se corre hacia la derecha, especialmente para los $M-$estimadores pesados, dando origen a valores mayores del parámetro de regularización. Este fenómeno se condice con los resultados de las Tablas 7.1 y 7.2 ya que los estimadores resultan más ralos dismimuyendo los valores de PVP.

Para evaluar la velocidad de convergencia a 0 del parámetro de penalización elegido por convalidación cruzada, se realizó un estudio numérico para distintos tamaños de muestra $n = 150, 200, 250, 300, 500$ y 1000 bajo **C0** cuando $p = 40$. Por simplicidad, se consideraron solamente dos funciones de pérdida, la pérdida $\rho(t) = t$ que da origen al estimador de máxima verosimilitud y la introducida por Croux y Haesbroeck (2003), cada una de ella combinadas con las penalidades Signo y MCP. Para el estimador clásico, se utilizó el procedimiento clásico de convalidación cruzada $CV(\lambda)$, mientras que para el robusto se empleó la versión robusta $RCV(\lambda)$ definida en la Sección 3.2.2. La Figura 7.2 presenta las medias sobre 500 replicaciones de λ_n en el gráfico (a) y de $\sqrt{n}\,\lambda_n$ en el (b), versus el tamaño muestral n. La Figura 7.2 (b) muestra que para el estimador clásico con penalización Signo, los valores de $\sqrt{n}\,\lambda_n$ rápidamente se estabilizan alrededor de 1.80, lo cual sugiere que $\sqrt{n}\,\lambda_n$ está acotado dando lugar a estimadores con tasa $\sqrt{n}$. Por otra parte, al utilizar la penalidad MCP, tanto para el estimador clásico como para el robusto, se observa que $\sqrt{n}\,\lambda_n$ crece con el tamaño muestral mientras que λ_n decrece a 0 (más lentamente que con la penalidad Signo, como era de esperar), sugiriendo que en este caso obtenemos un estimador con tasa $\sqrt{n}$ (ver Observación 5.4) que además selecciona variables consistentemente, por el Corolario 5.6.

7.2.5. Sobre los resultados obtenidos

Como es usual cuando se consideran métodos robustos, reportamos el promedio podado al 10 % de los resultados sobre 500 replicaciones. Las Tablas 7.3 y 7.4 reportan los resultados correspondientes a las medidas ECM, PVP y PVN bajo **C0**, las Tablas 7.5 a 7.7 los de las contaminaciones **CA1** y **CA2**, mientras que las Tablas 7.8 a 7.13 presentan los resultados obtenidos en los escenarios **CB1** y **CB2**. Por otro lado, los resultados correspondientes a las medidas PECM y PCC se presentan en las Tablas 7.15 a 7.20 del Apéndice **A**.

Las Figuras 7.3 a 7.8 presentan en forma gráfica los promedios con poda 10 % de las medidas PECM, PVP y PVN, bajo **C0**, **CA1** y **CA2**. En dichas figuras, la línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CA1** y **CA2**, respectivamente. Por otra parte, las líneas celeste, violeta, roja y verde corresponden a los estimadores que minimizan la *deviance* ($\rho(t) = t$), a los de mínimos cuadrados ($\rho(t) = 1 - \exp(-t)$), a los $M-$estimadores obtenidos usando la función de pérdida $\rho = \rho_c$ introducida en Croux y Haesbroeck (2003) dada en (3.5) y a los basados en $\rho = \rho_{\text{DIV}}$ que da origen a los estimadores de divergencia, respectivamente. Los gráficos superiores muestran los resultados cuando $w \equiv 1$ y los inferiores cuando $w(\mathbf{x}) = W(D^2(\mathbf{x}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}^{-1}))$. Finalmente, los gráficos del lado izquierdo de cada figura corresponden a los estimadores con penalización LASSO, los del centro están asociados a la penalidad Signo y los de la derecha a la MCP.

Por otra parte, las Figuras 7.9 a 7.14 presentan los valores de PECM, PVP y PVN, cuando se consideran muestras sin contaminar y contaminadas según los esquemas **CB1** y **CB2**. En todos los casos, la línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CB1** y **CB2**, respectivamente. En dichas Figuras, las líneas celeste, roja y verde corresponden a $\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{MV}}$, $\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{M}}$ y $\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{WM}}$ en el caso de las Figuras 7.9, 7.11 y 7.13 y a $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{MV}}$, $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{M}}$, $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{WM}}$ en las restantes.

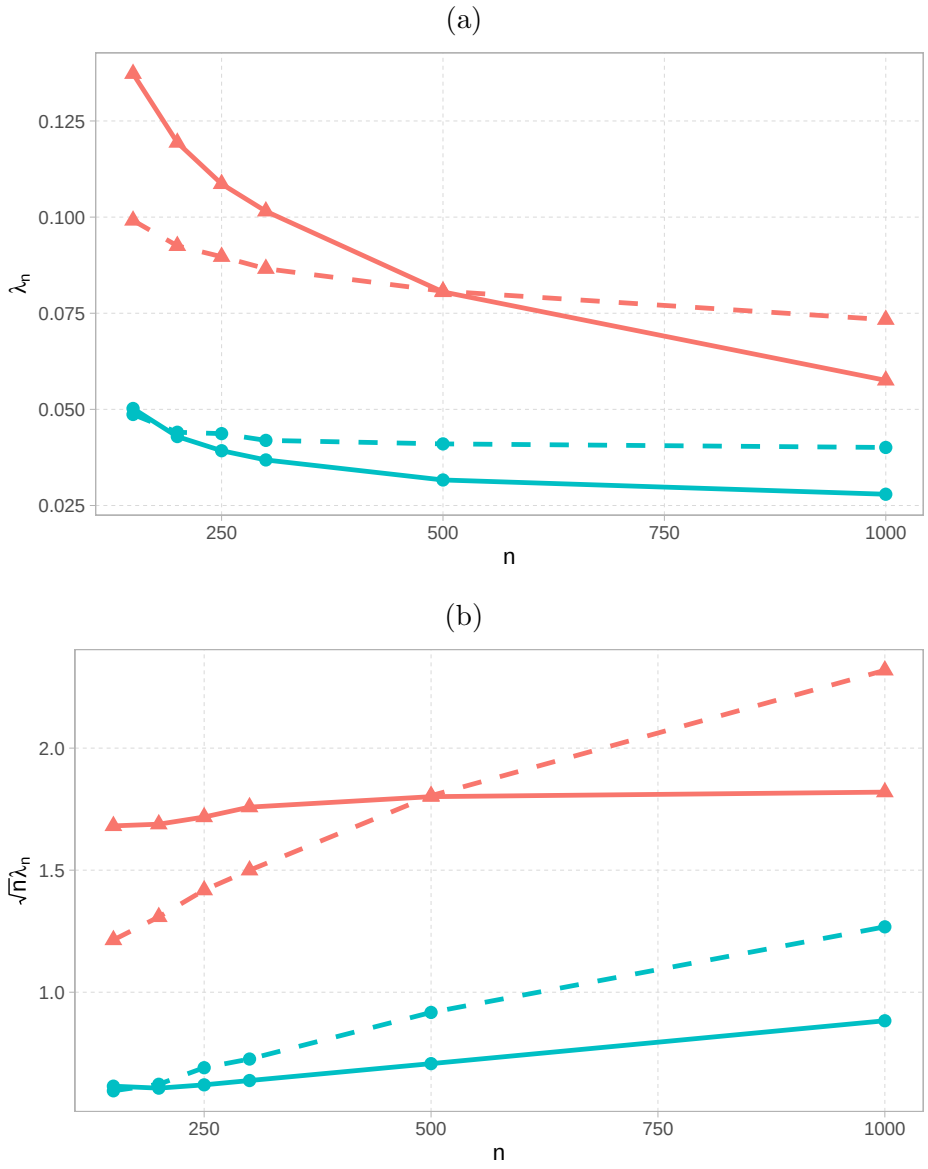


Figura 7.2: Medias sobre 500 replicaciones de los valores del parámetro de penalización, λ_n obtenidos por convalidación cruzada y de $\sqrt{n}\lambda_n$. Los triángulos rojos corresponden a $\rho(t) = t$ y los círculos azules a $\rho = \rho_c$. La línea continua corresponde a la penalización Signo, mientras que la entrecortada a la penalizalidad MCP. Los ejes del plot corresponden al tamaño muestral en el eje horizontal, mientras que en el eje horizontal se presenta λ_n (a) y $\sqrt{n}\lambda_n$ (b).

La Tabla 7.15 y la Figura 7.3 muestran que, para muestras sin contaminación, los estimadores penalizados con MCP muestran menores valores de PECM que las otras penalidades. En particular, para muestras de tamaño $n = 300$, los estimadores de máxima verosimilitud utilizando la penalidad MCP llegan a tener valores de PECM que son la tercera parte respecto de los obtenidos con la penalidad LASSO. Esa diferencia es aún mayor para los estimadores de mínimos cuadrados y los $M-$estimadores calculados con la función $\rho = \rho_c$ dada en (3.5). Los estimadores pesados bajo **C0** dan resultados semejantes a los no pesados, no solamente respecto del error cuadrático medio de las probabilidades sino respecto de todas las medidas consideradas (ver Tablas 7.15 a 7.4). Como se observa en las Tablas 7.15 y 7.3, el $M-$estimador penalizado con LASSO muestra mayor pérdida de eficiencia que con las demás penalidades, llegando a tener valores de PECM y ECM que al

menos duplican los de $\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MV}}$. La Figura 7.4 muestra el fenómeno antes mencionado. Efectivamente, cuando $n = 300$ y no hay contaminación en la muestra, las penalizaciones Signo y MCP dan menores valores de PECM que la penalización LASSO. Este fenómeno puede explicarse por el sesgo no despreciable, ya discutido en esta tesis, que introduce la penalidad LASSO incluso cuando n/p es grande. Para ambas penalizaciones acotadas, todas las funciones de pérdida dan resultados muy similares. La penalidad LASSO da resultados semejantes o mejores que la penalización Signo solamente al considerar los estimadores basados en $\rho(t) = t$ y $\rho = \rho_{\mathrm{DIV}}$.

Como es de esperar, los estimadores sin penalizar dan peores resultados que los obtenidos penalizando. Además, los errores PECM y ECM crecen cuando aumenta la dimensión. En particular, ese incremento es mayor al utilizar la penalidad Signo cuando $n = 150$ y $p = 120$, donde los valores de PECM casi duplican los obtenidos con $n = 150$ y $p = 40$ para la mayoría de los estimadores. Como hemos mencionado, el caso $(n, p) = (150, 120)$ es una situación que plantea un gran desafío en cuanto a la estimación de $\boldsymbol{\beta}_0$ y a la selección de variables.

Cabe mencionar que el comportamiento del PECM no siempre se condice con el del ECM. Para algunos casos, se obtiene un muy alto error de estimación de $\boldsymbol{\beta}_0$ pero un bajo error de predicción de las probabilidades. Esto sucede, por ejemplo, con las pérdidas $\rho = \rho_{\mathrm{DIV}}$ y con aquella que da origen a los estimadores mínimos cuadrados. Efectivamente, para algunas dimensiones el ECM de estos estimadores toma valores tan grandes que se reportan como ★. Una posible explicación de este hecho podría ser que estas pérdidas, a diferencia de lo que ocurre con la dada en (3.5), no cumplen las condiciones que garantizan la existencia del estimador no penalizado enunciadas en la Sección 3.1.1. Los resultados obtenidos sugieren que introducir una penalidad acotada no resuelve dicho problema de existencia, por lo que en algunas muestras el estimador puede explotar. Por lo antes mencionado, los estimadores calculados con la pérdida introducida por Croux y Haesbroeck (2003), tanto pesados como con $w \equiv 1$, producen mejores errores cuadráticos medios, ECM($\widehat{\boldsymbol{\beta}}$), que otras otras pérdidas acotadas, en particular, al usar la penalización Signo.

Respecto de la proporción de clasificaciones correctas y de las proporciones de verdaderos positivos y nulos, todos los estimadores penalizados dan resultados parecidos. Cabe mencionar que, cuando se usa la penalidad LASSO, se obtienen valores de PVN más bajos que con otras penalidades, dando origen a estimadores menos ralos. Este procedimiento muestra menor capacidad que MCP para identificar como 0 los coeficientes asociados a variables explicativas que no intervienen en el modelo. Ese problema también se observa, aunque en menor medida, cuando se consideran los estimadores de divergencia sin pesos penalizados con el Signo o los estimadores de máxima verosimilitud con la misma penalidad (ver Tabla 7.4).

La sensibilidad a datos anómalos de los estimadores basados en $\rho(t) = t$ y $w \equiv 1$, aún cuando se combinen con las penalidades consideradas, puede observarse en las distintas tablas. En particular, las Tablas 7.17 y 7.5 muestran que, al introducir datos atípicos según los esquemas **CA1** y **CA2**, los valores de PECM y ECM por lo menos triplican los obtenidos para muestras sin contaminar. Por otra parte, los valores de ECM, bajo **CB1** y **CB2**, pueden llegar a ser 5 veces más grandes que los obtenidos bajo **C0** (ver Tablas 7.8 y 7.9).

El mejor comportamiento de los $M-$estimadores pesados penalizados, bajo los esquemas de conta-

minación **CA1** y **CA2**, es evidente en las Figuras 7.3 y 7.4 donde se observa que el error cuadrático medio de las probabiliades (PECM) son cercanos a los obtenidos para las muestras limpias para las penalidades acotadas Signo y MCP. La ventaja de los estimadores pesados también se refleja en las proporciones de verdaderos positivos y nulos, como ilustran las Figuras 7.5 a 7.8. En el caso de estas medidas, la penalización LASSO da valores más altos de la probabilidad de verdaderos positivos, en detrimento de los valores de PVN ya que como mencionamos esta penalización tiene más problemas para identificar las variables explicativas no activas.

Cabe mencionar que, bajo **CA1** y **CA2**, los estimadores no pesados tienen valores de PECM mucho más altos que los pesados especialmente cuando $n = 150$. Estos valores pueden llegar a ser el doble de los obtenidos con los estimadores que controlan la palanca de las covariables, bajo **CA2**. Entre los estimadores con $w \equiv 1$, aquellos que dan menores valores de PECM son los estimadores correspondientes a $\rho = \rho_{\text{DIV}}$ y los basados en el procedimiento de mínimos cuadrados cuando se combinan con las penalizaciones Signo y MCP, en particular si $n = 300$.

En el escenario **CA1**, los estimadores basados en funciones de pérdidas acotadas dan lugar a resultados más estables. Por ejemplo, la Figura 7.6 muestra que la pérdida $\rho(t) = t$ es la única que tiene problemas con este nivel de contaminación. Por otro lado, la pérdida introducida por de Croux y Haesbroeck (2003) produce estimadores más ralos que las pérdidas $\rho = \rho_{\text{DIV}}$ y $\rho(t) = 1 - \exp(-t)$ que dan origen a los estimadores de mínima divergencia y a los de mínimos cuadrados.

Como puede observarse en la Tabla 7.7 al aumentar el nivel de contaminación (esquema **CA2**), todos los estimadores parecen volverse demasiado ralos, impactando en PVP que decrece casi a la mitad en los estimadores no pesados. Como es de esperar, esto se intensifica al utilizar las penalizaciones Signo y MCP combinados con $\rho(t) = t$. El efecto de la contaminación al 10 % también se observa, aunque en menor medida, en los $M-$estimadores con $\rho = \rho_c$ dada en (3.5). Los estimadores pesados tienen resultados similares a los obtenidos en el escenario **C0** respecto de su capacidad para detectar variables.

Al considerar los esquemas **CB1** y **CB2**, las Tablas 7.10 y 7.11 muestran el decrecimento existente en la probabilidad de verdaderos positivos, en particular, para valores pequeños de m ($m = 0.5, 1, 2$) que corresponden a los outliers intermedios que son los más difíciles de detectar. Los valores de PVP de los $M-$estimadores pesados, basados en la función introducida por Croux y Haesbroeck (2003), se recuperan al aumentar m, lo cual también se observa en los resultados obtenidos para la medida $PEMC$ que aumenta para valores bajos de m y decrece a medida que aumenta la palanca de los outliers (ver Figuras 7.9 y 7.10). Cabe mencionar que los valores de PVN obtenidos bajo **CB1** y **CB2** son semejantes a los obtenidos para muestras sin contaminar, salvo para los estimadores $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{MV}}$ que se ven más afectados por este tipo de contaminación. Estos resultados muestran que las penalidades Signo y MCP consiguen identificar las variables no activas (ver Tablas 7.12 y 7.13) aunque producen modelos más ralos que el verdadero.

De las Figuras 7.9 y 7.10 se desprende que, para las penalizaciones Signo y MCP, bajo **CB1** los estimadores basados en $\rho = \rho_c$ dada en (3.5) tienen valores PECM mucho menores que los obtenidos cuando $\rho(t) = t$, en particular cuando se consideran los $M-$estimadores pesados. Ese efecto es más claro cuando m es mayor igual a 3 ya que los estimadores $\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{MV}}$ y $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{MV}}$ tienen valores de PECM superiores a 0.10, es decir, 5 veces más grandes que los obtenidos bajo **C0**. El mejor comportamiento

de los estimadores pesados se debe a que este método detecta la mayoría de los datos atípicos, para valores grandes de m. En algunos casos, las ventajas de los $M-$estimadores robustos pesados se intensifican al utilizar la penalidad MCP. Por ejemplo, cuando $(n,p)=(300,80)$, el PECM de $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$ y $\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$ es muy similar al obtenido para datos sin contaminar. En el escenario de contaminación **CB2**, los estimadores basados la función definida en Croux y Haesbroeck (2003) empeoran su comportamiento y dan resultados muy similares a los obtenidos con el estimador clásico penalizado. La versión pesada de dichos estimadores muestra resultados aceptables de esta medida para valores grandes de m.

Resumiendo, para los esquemas de contaminación considerados, los $M-$estimadores pesados basados en la función $\rho=\rho_c$ dada en (3.5) combinados con las penalizaciones MCP y Signo, resultan ser los más estables y fiables.

n	150			300				150			300		
p	40	80	120	40	80	120		40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$	49.54	★	★	2.28	26.32	★	$\widehat{\boldsymbol{\beta}}_{\mathrm{WMV}}$	60.68	★	★	2.30	27.24	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MV}}$	1.23	1.50	1.75	0.90	0.95	1.03	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMV}}$	1.23	1.50	1.76	0.90	0.96	1.03
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	1.33	★	★	0.35	0.43	0.45	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMV}}$	1.38	★	★	0.35	0.44	0.45
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	1.02	1.18	1.52	0.20	0.25	0.24	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	1.03	1.16	1.54	0.21	0.27	0.24
$\widehat{\boldsymbol{\beta}}_{\mathrm{MC}}$	★	★	★	★	★	★	$\widehat{\boldsymbol{\beta}}_{\mathrm{WMC}}$	★	★	★	★	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MC}}$	2.78	2.81	2.93	2.50	2.52	2.54	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMC}}$	2.78	2.81	2.94	2.50	2.52	2.54
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MC}}$	2.80	25.44	★	0.49	0.63	0.69	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMC}}$	2.53	★	★	0.50	0.65	0.69
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	1.62	1.87	2.62	0.30	0.40	0.44	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	1.60	1.86	2.71	0.30	0.39	0.43
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	★	★	★	★	★	★	$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	★	★	★	★	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	1.50	1.75	1.98	1.18	1.23	1.32	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.50	1.75	1.99	1.18	1.24	1.32
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	1.72	★	★	0.40	0.52	0.54	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	1.76	★	★	0.41	0.53	0.54
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	1.24	1.43	1.86	0.25	0.31	0.28	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	1.23	1.43	1.88	0.24	0.30	0.29
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	★	★	★	3.69	★	★	$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	★	★	★	3.76	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	2.98	2.99	3.11	2.69	2.72	2.74	$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	2.98	3.00	3.11	2.69	2.72	2.74
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	1.60	2.00	2.78	0.38	0.50	0.55	$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	1.64	2.01	2.76	0.38	0.51	0.54
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	1.15	1.43	1.86	0.23	0.34	0.35	$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	1.15	1.41	1.87	0.23	0.33	0.36

Tabla 7.3: Promedio podado al 10 % del error cuadrático (ECM), bajo **C0**.

	$n = 150$			$n = 300$		
p	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MV}}$	1.00 / 0.61	1.00 / 0.68	1.00 / 0.72	1.00 / 0.69	1.00 / 0.70	1.00 / 0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.98 / 0.87	0.98 / 0.82	0.97 / 0.76	1.00 / 0.93	1.00 / 0.92	1.00 / 0.90
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.96 / 0.95	0.95 / 0.93	0.90 / 0.92	1.00 / 0.98	1.00 / 0.92	1.00 / 0.92
$\widehat{\boldsymbol{\beta}}_{\mathrm{MC}}$	1.00 / 0.09	1.00 / 0.18	1.00 / 0.24	1.00 / 0.03	1.00 / 0.09	1.00 / 0.13
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MC}}$	1.00 / 0.84	1.00 / 0.85	0.98 / 0.85	1.00 / 0.92	1.00 / 0.92	1.00 / 0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MC}}$	0.89 / 0.97	0.90 / 0.96	0.85 / 0.95	1.00 / 0.96	1.00 / 0.96	1.00 / 0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	0.91 / 0.98	0.91 / 0.97	0.86 / 0.97	1.00 / 0.99	1.00 / 0.97	1.00 / 0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	1.00 / 0.03	1.00 / 0.03	1.00 / 0.01	1.00 / 0.00	1.00 / 0.05	1.00 / 0.04
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	1.00 / 0.60	1.00 / 0.67	1.00 / 0.71	1.00 / 0.68	1.00 / 0.71	1.00 / 0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.95 / 0.91	0.95 / 0.87	0.94 / 0.84	1.00 / 0.94	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.95 / 0.96	0.93 / 0.93	0.88 / 0.92	1.00 / 0.98	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	1.00 / 0.01	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.02	1.00 / 0.01
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.98 / 0.88	0.99 / 0.88	0.97 / 0.88	1.00 / 0.94	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.96 / 0.95	0.94 / 0.96	0.86 / 0.95	1.00 / 0.96	1.00 / 0.95	1.00 / 0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.94 / 0.98	0.93 / 0.96	0.88 / 0.96	1.00 / 0.99	1.00 / 0.97	1.00 / 0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMV}}$	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMV}}$	1.00 / 0.61	1.00 / 0.68	1.00 / 0.72	1.00 / 0.69	1.00 / 0.71	1.00 / 0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMV}}$	0.98 / 0.87	0.98 / 0.81	0.97 / 0.76	1.00 / 0.93	1.00 / 0.92	1.00 / 0.89
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	0.96 / 0.95	0.95 / 0.93	0.90 / 0.92	1.00 / 0.98	1.00 / 0.93	1.00 / 0.92
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMC}}$	1.00 / 0.09	1.00 / 0.19	1.00 / 0.25	1.00 / 0.04	1.00 / 0.09	1.00 / 0.14
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMC}}$	1.00 / 0.84	1.00 / 0.85	0.99 / 0.85	1.00 / 0.92	1.00 / 0.92	1.00 / 0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMC}}$	0.88 / 0.97	0.89 / 0.96	0.83 / 0.95	1.00 / 0.95	1.00 / 0.96	1.00 / 0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	0.91 / 0.98	0.91 / 0.97	0.86 / 0.97	1.00 / 0.99	1.00 / 0.97	1.00 / 0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	1.00 / 0.03	1.00 / 0.03	1.00 / 0.01	1.00 / 0.00	1.00 / 0.05	1.00 / 0.04
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.00 / 0.60	1.00 / 0.67	1.00 / 0.71	1.00 / 0.68	1.00 / 0.71	1.00 / 0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.95 / 0.91	0.96 / 0.87	0.93 / 0.84	1.00 / 0.94	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.95 / 0.96	0.93 / 0.93	0.89 / 0.93	1.00 / 0.98	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	1.00 / 0.01	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.02	1.00 / 0.01
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.99 / 0.87	0.99 / 0.88	0.97 / 0.88	1.00 / 0.94	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.96 / 0.94	0.94 / 0.96	0.87 / 0.95	1.00 / 0.95	1.00 / 0.95	1.00 / 0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.94 / 0.98	0.93 / 0.96	0.88 / 0.96	1.00 / 0.99	1.00 / 0.97	1.00 / 0.97

Tabla 7.4: Promedio podado al 10 % del la proporción de verdaderos positivos/ proporción de verdaderos nulos (PVP/PVN) bajo **C0**.

	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
	$n = 150$			$n = 300$			$n = 150$			$n = 300$		
p	40	80	120	40	80	120	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$	7.43	★	★	2.37	6.50	★	4.93	★	★	3.33	4.67	33.30
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MV}}$	3.78	3.88	3.93	3.61	3.67	3.66	4.86	4.86	4.92	4.83	4.85	4.89
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	3.76	★	★	3.82	4.00	3.77	4.71	4.76	★	4.62	4.78	4.83
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	3.22	3.30	3.66	3.03	3.20	3.21	4.84	4.86	4.92	4.89	4.98	4.99
$\widehat{\boldsymbol{\beta}}_{\mathrm{MC}}$	★	★	★	★	★	★	★	★	★	★	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MC}}$	3.36	3.45	3.58	2.93	2.97	3.06	4.55	4.67	4.74	4.12	4.22	4.34
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MC}}$	3.09	3.74	4.27	1.05	1.21	1.44	4.05	4.41	4.54	2.57	2.86	3.49
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	2.28	2.84	3.71	0.73	0.92	1.12	3.64	4.31	4.84	1.68	2.09	2.88
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	★	★	★	4.47	★	★	★	★	★	3.06	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	2.39	2.58	2.79	1.93	2.01	2.14	4.27	4.33	4.42	3.83	3.86	3.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	2.80	★	★	1.08	1.17	1.43	4.05	4.31	★	3.13	3.13	3.48
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	2.08	2.49	3.09	0.86	0.90	1.10	3.58	3.89	4.29	2.19	2.31	2.70
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	★	★	★	2.33	★	★	★	★	★	2.59	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	3.86	3.94	4.04	3.60	3.63	3.64	4.80	4.84	4.91	4.75	4.76	4.78
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	3.36	3.68	4.12	1.41	1.49	2.02	4.44	4.59	4.70	4.32	4.37	4.61
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	2.30	2.84	3.82	1.11	1.08	1.41	4.41	4.86	5.00	3.77	4.14	4.63
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMV}}$	55.90	★	★	2.30	26.77	★	50.81	★	★	2.29	26.19	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMV}}$	1.24	1.52	1.77	0.90	0.97	1.05	1.24	1.53	1.78	0.90	0.97	1.07
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMV}}$	1.44	★	★	0.34	0.44	0.44	1.59	3.15	★	0.41	0.45	0.45
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	1.05	1.37	1.71	0.21	0.31	0.28	1.11	1.59	2.05	0.22	0.35	0.37
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMC}}$	★	★	★	★	★	★	★	★	★	★	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMC}}$	2.77	2.81	2.95	2.50	2.53	2.56	2.77	2.80	2.95	2.50	2.54	2.58
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMC}}$	3.43	★	★	0.50	0.63	0.70	343.84	★	★	0.55	0.68	0.73
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	1.60	2.27	2.68	0.30	0.48	0.48	1.73	2.59	2.99	0.31	0.55	0.68
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	★	★	★	★	★	★	★	★	★	★	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.50	1.77	2.02	1.18	1.25	1.33	1.50	1.78	2.04	1.18	1.26	1.36
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	1.80	★	★	0.41	0.53	0.53	1.94	3.16	★	0.47	0.54	0.54
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	1.24	1.72	2.04	0.25	0.36	0.33	1.34	1.92	2.34	0.25	0.41	0.45
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	★	★	★	3.72	★	★	★	★	★	3.68	★	★
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	2.98	3.00	3.10	2.69	2.73	2.77	2.98	3.00	3.12	2.69	2.74	2.79
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	1.89	2.30	2.85	0.39	0.48	0.51	2.27	2.44	2.99	0.65	0.55	0.56
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	1.24	1.78	2.27	0.23	0.42	0.47	1.24	2.04	2.62	0.24	0.46	0.57

Tabla 7.5: Promedio podado al 10 % del error cuarático (ECM), para los escenarios de contaminación **CA1** y **CA2**.

	$n = 150$			$n = 300$		
p	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MV}}$	0.81 / 0.69	0.79 / 0.75	0.76 / 0.79	0.94 / 0.65	0.95 / 0.71	0.95 / 0.73
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.54 / 0.85	0.75 / 0.73	0.91 / 0.67	0.60 / 0.94	0.61 / 0.88	0.61 / 0.81
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.70 / 0.79	0.71 / 0.82	0.68 / 0.84	0.81 / 0.78	0.78 / 0.81	0.77 / 0.82
$\widehat{\boldsymbol{\beta}}_{\mathrm{MC}}$	1.00 / 0.07	1.00 / 0.16	1.00 / 0.24	1.00 / 0.02	1.00 / 0.07	1.00 / 0.11
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MC}}$	0.93 / 0.70	0.93 / 0.72	0.92 / 0.75	0.99 / 0.75	1.00 / 0.75	1.00 / 0.75
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MC}}$	0.72 / 0.97	0.70 / 0.96	0.63 / 0.95	1.00 / 0.87	1.00 / 0.94	0.99 / 0.94
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	0.79 / 0.97	0.77 / 0.96	0.67 / 0.96	0.97 / 0.98	0.99 / 0.97	0.96 / 0.96
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	1.00 / 0.04	1.00 / 0.04	1.00 / 0.03	1.00 / 0.00	1.00 / 0.04	1.00 / 0.06
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.99 / 0.48	0.98 / 0.56	0.97 / 0.61	1.00 / 0.47	1.00 / 0.51	1.00 / 0.53
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.82 / 0.87	0.85 / 0.82	0.86 / 0.80	0.99 / 0.80	1.00 / 0.86	0.99 / 0.84
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.84 / 0.89	0.85 / 0.89	0.80 / 0.89	0.97 / 0.91	0.98 / 0.89	0.97 / 0.87
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	1.00 / 0.01	1.00 / 0.02	1.00 / 0.02	1.00 / 0.00	1.00 / 0.01	1.00 / 0.02
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.87 / 0.73	0.87 / 0.76	0.86 / 0.78	0.96 / 0.75	0.97 / 0.76	0.98 / 0.76
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.59 / 0.97	0.61 / 0.96	0.55 / 0.95	0.93 / 0.91	0.97 / 0.94	0.94 / 0.94
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.72 / 0.96	0.71 / 0.95	0.59 / 0.95	0.91 / 0.97	0.96 / 0.96	0.96 / 0.95
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMV}}$	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMV}}$	1.00 / 0.61	1.00 / 0.67	1.00 / 0.71	1.00 / 0.69	1.00 / 0.71	1.00 / 0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMV}}$	0.98 / 0.88	0.96 / 0.83	0.96 / 0.78	1.00 / 0.92	1.00 / 0.92	1.00 / 0.90
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	0.97 / 0.95	0.93 / 0.92	0.87 / 0.92	1.00 / 0.98	1.00 / 0.93	1.00 / 0.92
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMC}}$	1.00 / 0.09	1.00 / 0.18	1.00 / 0.21	1.00 / 0.04	1.00 / 0.09	1.00 / 0.12
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMC}}$	1.00 / 0.85	1.00 / 0.85	0.98 / 0.85	1.00 / 0.92	1.00 / 0.92	1.00 / 0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMC}}$	0.90 / 0.96	0.85 / 0.95	0.79 / 0.95	1.00 / 0.95	1.00 / 0.96	1.00 / 0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	0.91 / 0.98	0.86 / 0.97	0.79 / 0.97	1.00 / 0.99	1.00 / 0.97	1.00 / 0.97
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	1.00 / 0.03	1.00 / 0.02	1.00 / 0.01	1.00 / 0.00	1.00 / 0.05	1.00 / 0.03
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.00 / 0.60	1.00 / 0.66	1.00 / 0.70	1.00 / 0.69	1.00 / 0.71	1.00 / 0.71
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.96 / 0.91	0.93 / 0.88	0.91 / 0.86	1.00 / 0.93	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.96 / 0.95	0.92 / 0.93	0.85 / 0.92	1.00 / 0.98	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	1.00 / 0.01	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.02	1.00 / 0.01
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.99 / 0.88	0.99 / 0.88	0.97 / 0.87	1.00 / 0.94	1.00 / 0.94	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.89 / 0.90	0.86 / 0.94	0.80 / 0.95	1.00 / 0.90	1.00 / 0.93	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.94 / 0.98	0.90 / 0.96	0.82 / 0.96	1.00 / 0.99	1.00 / 0.97	1.00 / 0.96

Tabla 7.6: Promedio podado al 10 % del la proporción de verdaderos positivos/ proporción de verdaderos nulos (PVP/PVN) para el escenario **CA1**.

	$n = 150$			$n = 300$		
p	40	80	120	40	80	120
$\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}$	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MV}}$	0.45 / 0.78	0.44 / 0.84	0.41 / 0.87	0.57 / 0.77	0.60 / 0.83	0.56 / 0.86
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.33 / 0.94	0.38 / 0.88	0.60 / 0.78	0.46 / 0.98	0.42 / 0.94	0.33 / 0.92
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.34 / 0.81	0.33 / 0.87	0.33 / 0.89	0.40 / 0.82	0.31 / 0.90	0.34 / 0.91
$\widehat{\boldsymbol{\beta}}_{\mathrm{MC}}$	1.00 / 0.05	1.00 / 0.13	1.00 / 0.19	1.00 / 0.00	1.00 / 0.05	1.00 / 0.08
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MC}}$	0.60 / 0.77	0.59 / 0.82	0.58 / 0.84	0.81 / 0.69	0.84 / 0.73	0.83 / 0.75
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MC}}$	0.47 / 0.97	0.40 / 0.96	0.42 / 0.96	0.83 / 0.88	0.83 / 0.94	0.77 / 0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	0.53 / 0.95	0.43 / 0.95	0.38 / 0.95	0.88 / 0.94	0.90 / 0.95	0.79 / 0.95
$\widehat{\boldsymbol{\beta}}_{\mathrm{DIV}}$	1.00 / 0.02	1.00 / 0.05	1.00 / 0.05	1.00 / 0.00	1.00 / 0.03	1.00 / 0.06
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.67 / 0.63	0.68 / 0.71	0.67 / 0.74	0.87 / 0.53	0.87 / 0.60	0.85 / 0.65
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.48 / 0.90	0.51 / 0.88	0.57 / 0.84	0.73 / 0.82	0.75 / 0.87	0.69 / 0.87
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.65 / 0.81	0.60 / 0.86	0.56 / 0.88	0.88 / 0.78	0.90 / 0.83	0.86 / 0.85
$\widehat{\boldsymbol{\beta}}_{\mathrm{M}}$	1.00 / 0.00	1.00 / 0.02	1.00 / 0.02	1.00 / 0.00	1.00 / 0.00	1.00 / 0.01
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{M}}$	0.50 / 0.83	0.54 / 0.86	0.49 / 0.89	0.62 / 0.81	0.69 / 0.84	0.69 / 0.85
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.36 / 0.98	0.36 / 0.97	0.27 / 0.97	0.48 / 0.99	0.61 / 0.97	0.52 / 0.96
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.34 / 0.94	0.31 / 0.95	0.30 / 0.95	0.51 / 0.95	0.47 / 0.95	0.47 / 0.95
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMV}}$	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMV}}$	1.00 / 0.61	1.00 / 0.67	1.00 / 0.71	1.00 / 0.68	1.00 / 0.70	1.00 / 0.72
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMV}}$	0.96 / 0.88	0.93 / 0.85	0.94 / 0.79	1.00 / 0.90	1.00 / 0.92	1.00 / 0.90
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	0.96 / 0.94	0.91 / 0.93	0.83 / 0.91	1.00 / 0.97	1.00 / 0.95	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}_{\mathrm{WMC}}$	1.00 / 0.09	1.00 / 0.17	1.00 / 0.18	1.00 / 0.03	1.00 / 0.09	1.00 / 0.11
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WMC}}$	1.00 / 0.85	1.00 / 0.84	0.99 / 0.85	1.00 / 0.93	1.00 / 0.91	1.00 / 0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WMC}}$	0.89 / 0.94	0.82 / 0.94	0.73 / 0.95	1.00 / 0.94	1.00 / 0.95	1.00 / 0.95
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	0.90 / 0.97	0.82 / 0.96	0.71 / 0.96	1.00 / 0.99	1.00 / 0.97	1.00 / 0.96
$\widehat{\boldsymbol{\beta}}_{\mathrm{WDIV}}$	1.00 / 0.04	1.00 / 0.02	1.00 / 0.01	1.00 / 0.00	1.00 / 0.05	1.00 / 0.04
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	1.00 / 0.59	1.00 / 0.65	1.00 / 0.70	1.00 / 0.68	1.00 / 0.70	1.00 / 0.71
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.95 / 0.92	0.91 / 0.89	0.89 / 0.87	1.00 / 0.91	1.00 / 0.93	1.00 / 0.93
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.96 / 0.95	0.89 / 0.93	0.81 / 0.92	1.00 / 0.98	1.00 / 0.95	1.00 / 0.94
$\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}$	1.00 / 0.02	1.00 / 0.00	1.00 / 0.00	1.00 / 0.00	1.00 / 0.02	1.00 / 0.01
$\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.99 / 0.88	0.99 / 0.87	0.97 / 0.87	1.00 / 0.95	1.00 / 0.93	1.00 / 0.92
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.77 / 0.88	0.76 / 0.93	0.71 / 0.95	0.94 / 0.87	0.99 / 0.89	1.00 / 0.91
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.94 / 0.97	0.87 / 0.96	0.75 / 0.96	1.00 / 0.99	1.00 / 0.97	1.00 / 0.96

Tabla 7.7: Promedio podado al 10 % del la proporción de verdaderos positivos/ proporción de verdaderos nulos (PVP/PVN) para el escenario **CA2**.

$n = 150$	$p = 40$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	1.843	3.638	4.382	4.736	4.970	5.070	3.304	4.484	5.092	5.138	5.160	5.215
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	1.632	3.079	4.540	5.114	5.321	5.349	2.829	4.310	5.433	5.499	5.165	5.181
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	1.925	3.316	4.123	4.352	4.397	4.243	3.598	4.531	5.050	5.152	5.153	5.185
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	1.515	2.244	3.119	3.236	2.938	2.661	2.987	4.548	5.110	5.110	5.066	5.137
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	1.999	3.348	4.167	1.871	2.023	2.016	3.597	4.563	5.061	3.597	3.880	3.967
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	1.546	2.252	3.185	1.289	1.300	1.304	2.963	4.560	5.104	1.410	1.284	1.282
	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	3.335	4.365	4.951	5.221	5.666	5.194	4.327	5.162	5.221	5.152	5.210	5.258
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	2.657	4.238	5.360	5.554	5.526	5.276	3.972	5.365	5.990	5.107	5.141	5.196
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	3.378	4.139	4.564	4.510	4.457	4.562	4.451	5.141	5.189	5.172	5.204	5.209
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	2.413	3.556	3.580	3.132	2.849	3.064	4.532	5.065	5.125	5.000	5.027	5.081
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	3.391	4.176	4.594	2.597	2.542	2.516	4.477	5.155	5.150	4.072	4.340	4.273
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	2.469	3.544	3.571	2.204	2.157	2.107	4.526	5.070	5.068	2.418	1.890	1.904

Tabla 7.8: Promedio podado al 10 % del error cuadrático (ECM), para los escenarios de contaminación **CB1** y **CB2** con $n = 150$.

$n = 300$	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	1.571	3.589	4.621	5.013	5.079	5.123	3.520	4.974	5.071	5.216	5.284	5.296
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	1.537	3.819	4.966	5.216	5.194	5.060	3.135	4.766	5.376	5.134	5.237	5.250
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	1.046	1.607	1.736	1.286	1.277	1.932	3.732	4.963	5.080	5.175	5.215	5.221
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.981	1.061	0.618	0.465	0.426	0.677	3.045	4.972	5.028	5.102	5.109	5.082
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	1.061	1.612	1.749	0.509	0.519	0.507	3.768	4.976	5.074	2.681	2.354	1.719
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.996	1.081	0.653	0.425	0.435	0.433	3.030	4.968	4.946	0.363	0.426	0.446
	$p = 120$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	2.381	3.936	4.947	5.069	5.117	5.147	4.309	5.096	5.095	5.248	5.299	5.297
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	2.257	4.514	5.195	5.254	5.066	5.052	3.944	5.340	5.171	5.177	5.232	5.229
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	1.161	2.217	1.951	1.578	2.425	2.959	4.462	5.094	5.106	5.195	5.223	5.228
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	1.221	1.286	0.537	0.451	0.868	2.010	4.280	5.035	5.000	5.080	5.096	5.101
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	1.155	2.247	1.997	0.551	0.558	0.550	4.438	5.095	3.835	2.525	2.159	1.572
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	1.237	1.352	0.545	0.511	0.509	0.532	4.286	5.036	3.188	0.499	0.586	0.657

Tabla 7.9: Promedio podado al 10 % del error cuadrático (ECM), para los escenarios de contaminación **CB1** y **CB2** con $n = 300$.

$n = 150$	$p = 40$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.865	0.535	0.403	0.356	0.363	0.394	0.597	0.327	0.272	0.249	0.282	0.307
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.874	0.666	0.497	0.469	0.453	0.456	0.704	0.459	0.467	0.482	0.460	0.552
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.820	0.566	0.422	0.389	0.415	0.458	0.476	0.259	0.217	0.211	0.224	0.256
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.846	0.723	0.588	0.588	0.640	0.727	0.591	0.306	0.275	0.305	0.216	0.425
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.805	0.551	0.418	0.965	0.948	0.933	0.469	0.254	0.224	0.619	0.546	0.535
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.845	0.723	0.582	0.952	0.957	0.951	0.599	0.291	0.285	0.960	0.958	0.956
	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.632	0.443	0.365	0.454	0.542	0.509	0.375	0.257	0.314	0.229	0.259	0.308
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.715	0.546	0.460	0.398	0.445	0.469	0.518	0.420	0.474	0.446	0.554	0.602
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.648	0.493	0.388	0.418	0.419	0.420	0.354	0.199	0.185	0.216	0.252	0.261
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.726	0.506	0.519	0.603	0.682	0.643	0.283	0.213	0.263	0.224	0.438	0.472
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.647	0.483	0.380	0.826	0.831	0.837	0.341	0.191	0.219	0.484	0.434	0.431
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.720	0.511	0.512	0.881	0.876	0.865	0.275	0.224	0.315	0.901	0.923	0.914

Tabla 7.10: Promedio podado al 10 % del la proporción de verdaderos positivos (PVP) para los escenarios **CB1** y **CB2** con $n = 150$.

$n = 300$	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.991	0.744	0.571	0.605	0.674	0.626	0.715	0.372	0.311	0.454	0.476	0.459
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.961	0.726	0.474	0.455	0.526	0.492	0.744	0.593	0.521	0.533	0.611	0.619
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	1.000	0.952	0.887	0.912	0.908	0.837	0.673	0.338	0.289	0.391	0.356	0.359
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.987	0.970	0.981	0.996	1.000	0.978	0.736	0.312	0.401	0.491	0.530	0.532
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.999	0.945	0.868	1.000	1.000	1.000	0.672	0.329	0.324	0.753	0.786	0.869
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.985	0.968	0.978	1.000	1.000	1.000	0.725	0.301	0.444	1.000	1.000	1.000
	$p = 120$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.931	0.667	0.474	0.642	0.624	0.652	0.536	0.334	0.304	0.422	0.438	0.488
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.897	0.641	0.404	0.474	0.440	0.505	0.596	0.562	0.491	0.572	0.598	0.603
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.982	0.889	0.858	0.885	0.801	0.805	0.524	0.325	0.307	0.366	0.376	0.370
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.967	0.956	0.983	0.993	0.964	0.881	0.463	0.284	0.338	0.514	0.546	0.623
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.986	0.889	0.870	1.000	1.000	1.000	0.521	0.331	0.581	0.791	0.836	0.902
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.968	0.951	0.983	1.000	1.000	0.997	0.463	0.285	0.635	1.000	1.000	1.000

Tabla 7.11: Promedio podado al 10 % del la proporción de verdaderos positivos (PVP) para los escenarios **CB1** y **CB2** con $n = 300$.

$n = 150$	$p = 40$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.931	0.955	0.962	0.963	0.950	0.947	0.937	0.922	0.908	0.924	0.946	0.962
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.931	0.903	0.842	0.784	0.743	0.732	0.897	0.831	0.697	0.739	0.905	0.945
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.949	0.970	0.979	0.970	0.967	0.963	0.971	0.967	0.956	0.956	0.959	0.968
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.974	0.971	0.965	0.961	0.959	0.958	0.961	0.945	0.911	0.928	0.976	0.969
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.951	0.973	0.979	0.918	0.926	0.931	0.971	0.966	0.956	0.938	0.935	0.935
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.974	0.969	0.966	0.972	0.971	0.972	0.961	0.947	0.912	0.944	0.949	0.951
	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.934	0.930	0.916	0.887	0.869	0.902	0.913	0.887	0.884	0.953	0.968	0.966
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.915	0.878	0.819	0.811	0.813	0.869	0.865	0.803	0.770	0.923	0.955	0.958
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.959	0.962	0.962	0.964	0.963	0.971	0.958	0.955	0.956	0.966	0.972	0.972
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.965	0.965	0.953	0.951	0.950	0.958	0.956	0.942	0.933	0.966	0.963	0.962
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.960	0.962	0.963	0.956	0.957	0.956	0.959	0.955	0.954	0.947	0.947	0.951
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.965	0.963	0.953	0.960	0.960	0.960	0.958	0.937	0.934	0.941	0.944	0.943

Tabla 7.12: Promedio podado al 10 % del la proporción de verdaderos nulos (PVN) para los escenarios **CB1** y **CB2** con $n = 150$.

$n = 300$	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.950	0.957	0.964	0.957	0.960	0.969	0.953	0.926	0.941	0.965	0.968	0.968
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.953	0.917	0.840	0.767	0.801	0.932	0.910	0.763	0.753	0.957	0.964	0.964
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.957	0.953	0.955	0.953	0.956	0.964	0.964	0.963	0.963	0.969	0.974	0.975
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.971	0.971	0.952	0.938	0.944	0.965	0.963	0.942	0.944	0.967	0.964	0.964
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.957	0.953	0.956	0.951	0.954	0.953	0.963	0.964	0.960	0.944	0.943	0.940
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.971	0.971	0.947	0.972	0.972	0.972	0.964	0.940	0.949	0.961	0.958	0.959
	$p = 120$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.950	0.944	0.933	0.938	0.956	0.960	0.939	0.895	0.946	0.968	0.966	0.959
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.939	0.896	0.829	0.815	0.934	0.960	0.871	0.732	0.883	0.960	0.959	0.958
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$	0.947	0.949	0.953	0.953	0.959	0.958	0.957	0.955	0.965	0.972	0.970	0.972
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.964	0.965	0.930	0.931	0.958	0.968	0.962	0.934	0.960	0.959	0.952	0.951
$\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.947	0.949	0.953	0.950	0.950	0.951	0.959	0.955	0.950	0.940	0.941	0.940
$\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.964	0.964	0.929	0.968	0.968	0.967	0.961	0.933	0.956	0.953	0.952	0.952

Tabla 7.13: Promedio podado al 10 % del la proporción de verdaderos nulos (PVN) para los escenarios **CB1** y **CB2** con $n = 300$.

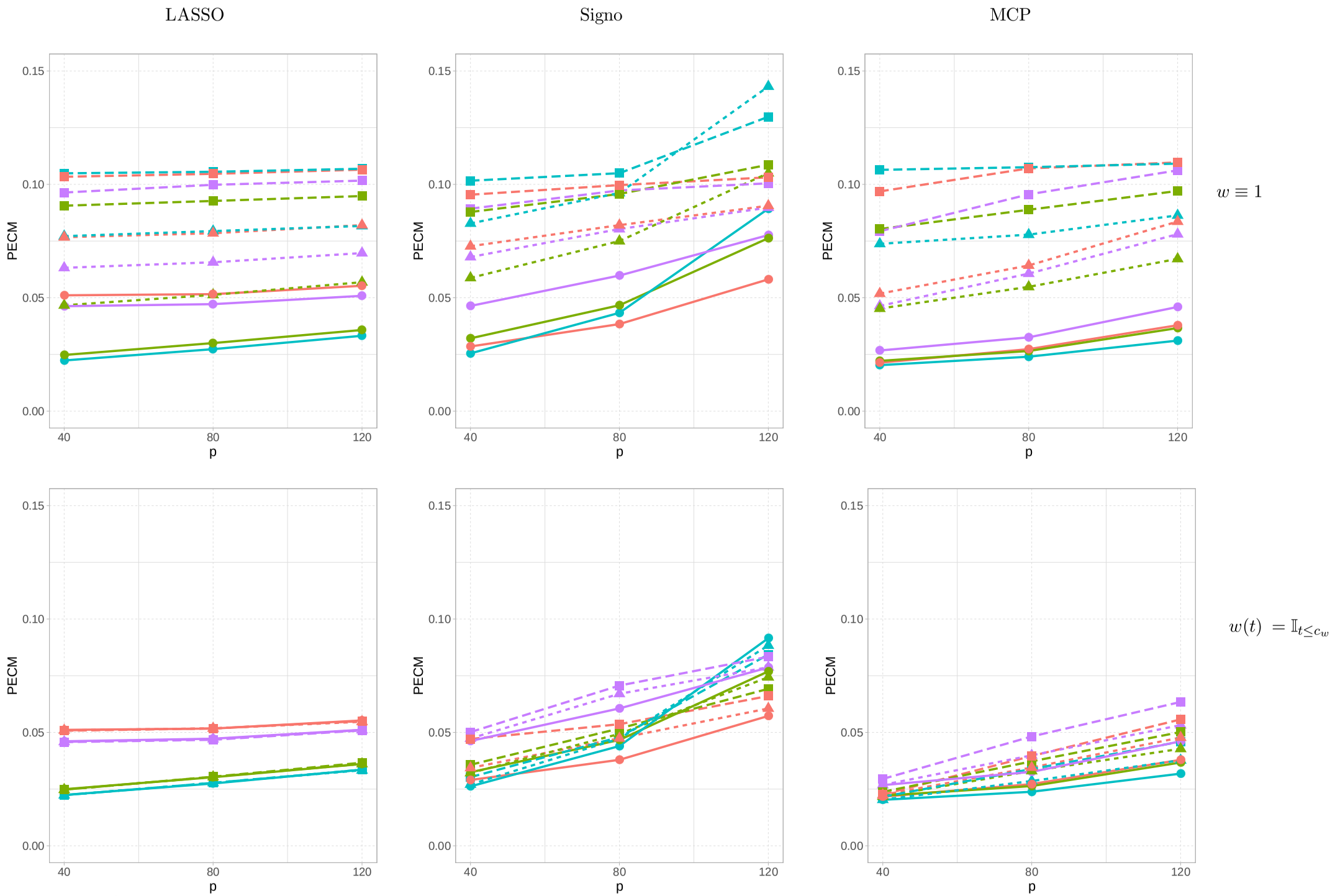


Figura 7.3: Error Cuadrático Medio de las Probabilidades para $n = 150$ en los escenarios **C0**, **CA1** y **CA2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CA1** y **CA2**, respectivamente. Por otra parte, las líneas roja, verde, celeste y violeta corresponden a $\rho = \rho_c$ dada en (3.5), $\rho = \rho_{\text{DIV}}$, $\rho(t) = t$ y $\rho(t) = 1 - \exp(-t)$.

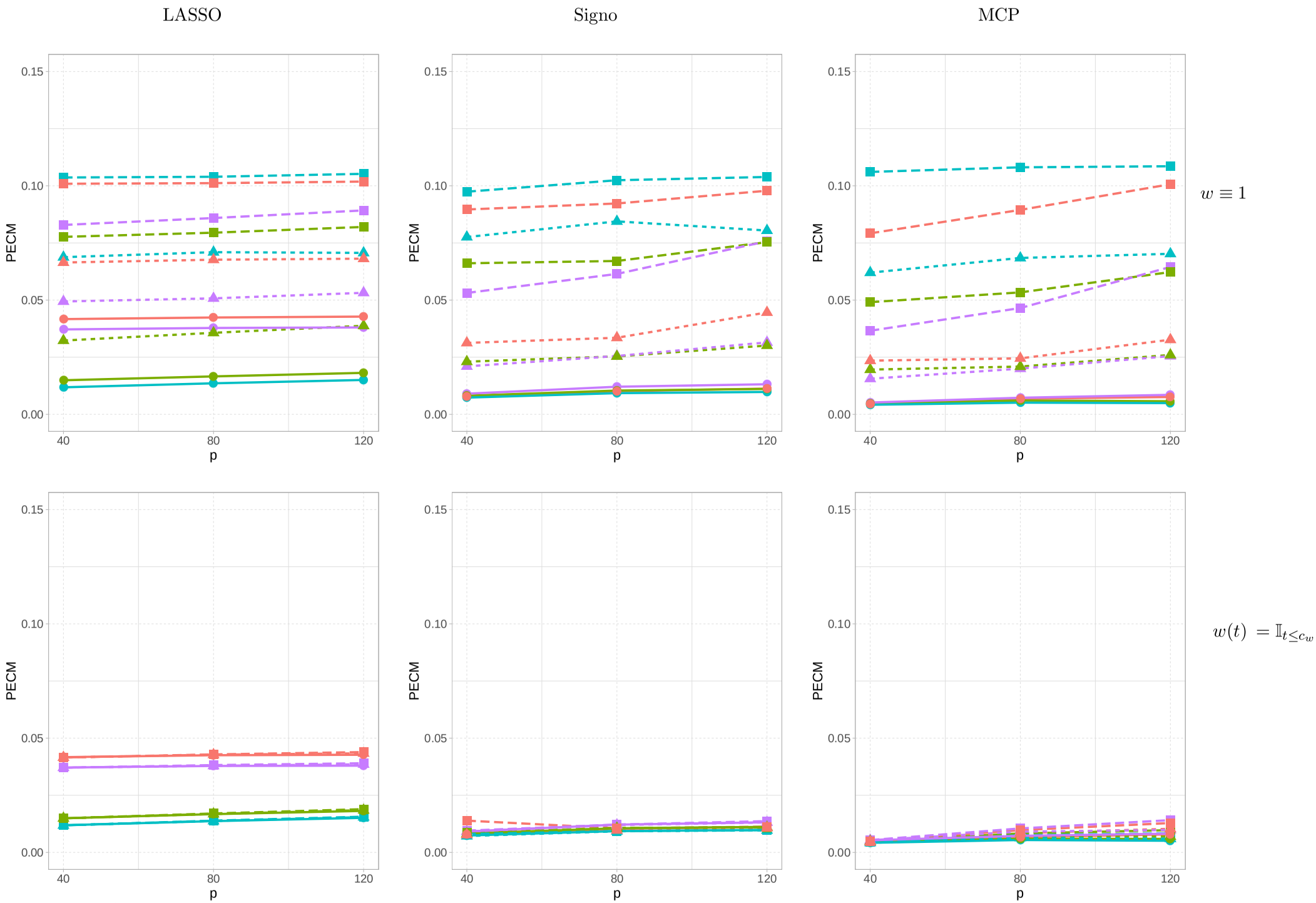


Figura 7.4: Error Cuadrático Medio de las Probabilidades para $n = 300$ en los escenarios **C0**, **CA1** y **CA2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CA1** y **CA2**, respectivamente. Por otra parte, las líneas roja, verde, celeste y violeta corresponden a $\rho = \rho_c$ dada en (3.5), $\rho = \rho_{\text{DIV}}$, $\rho(t) = t$ y $\rho(t) = 1 - \exp(-t)$.

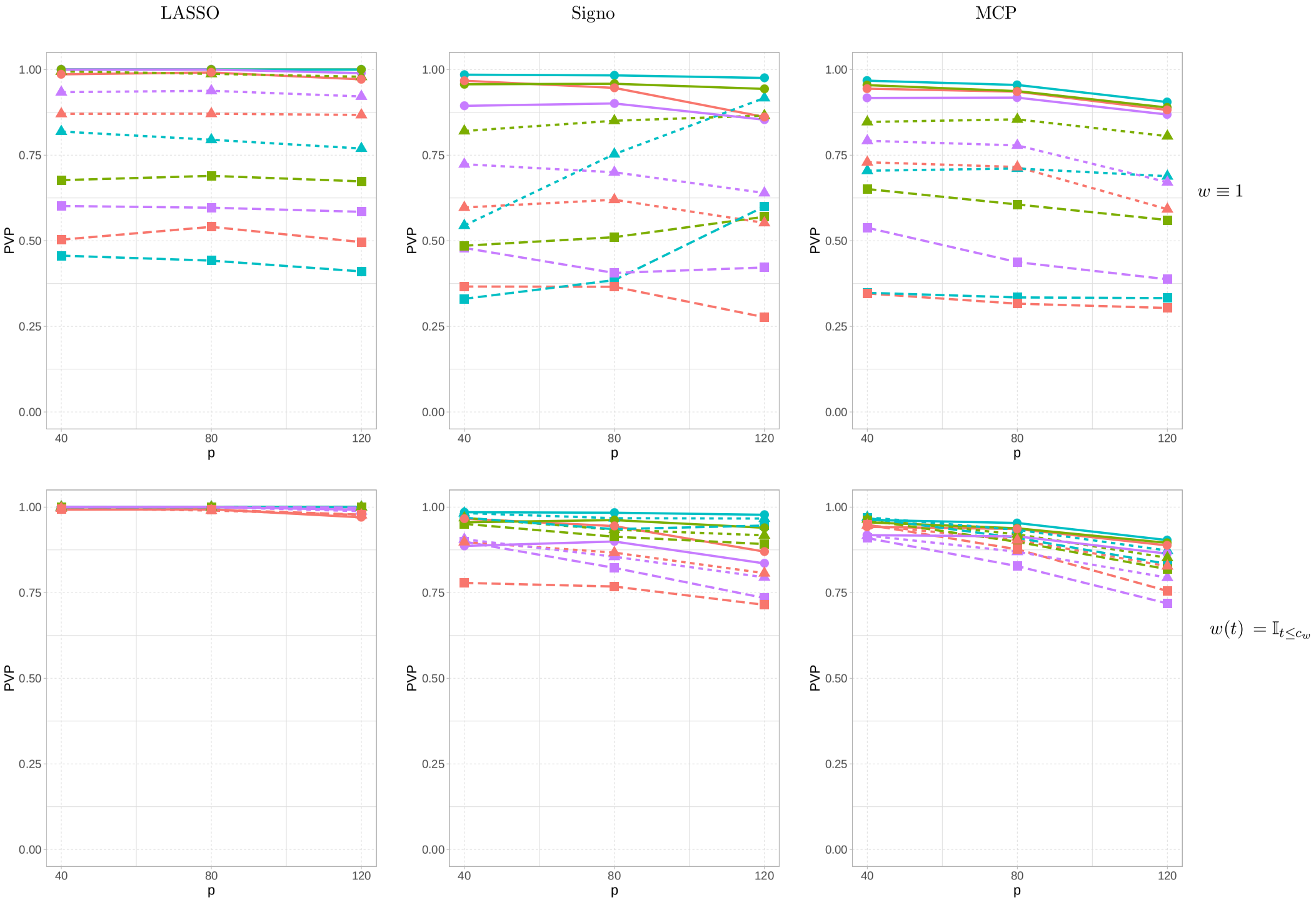


Figura 7.5: Proporción de Verdaderos Positivos para $n = 150$ en los escenarios **C0**, **CA1** y **CA2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CA1** y **CA2**, respectivamente. Por otra parte, las líneas roja, verde, celeste y violeta corresponden a $\rho = \rho_c$ dada en (3.5), $\rho = \rho_{\text{DIV}}$, $\rho(t) = t$ y $\rho(t) = 1 - \exp(-t)$.

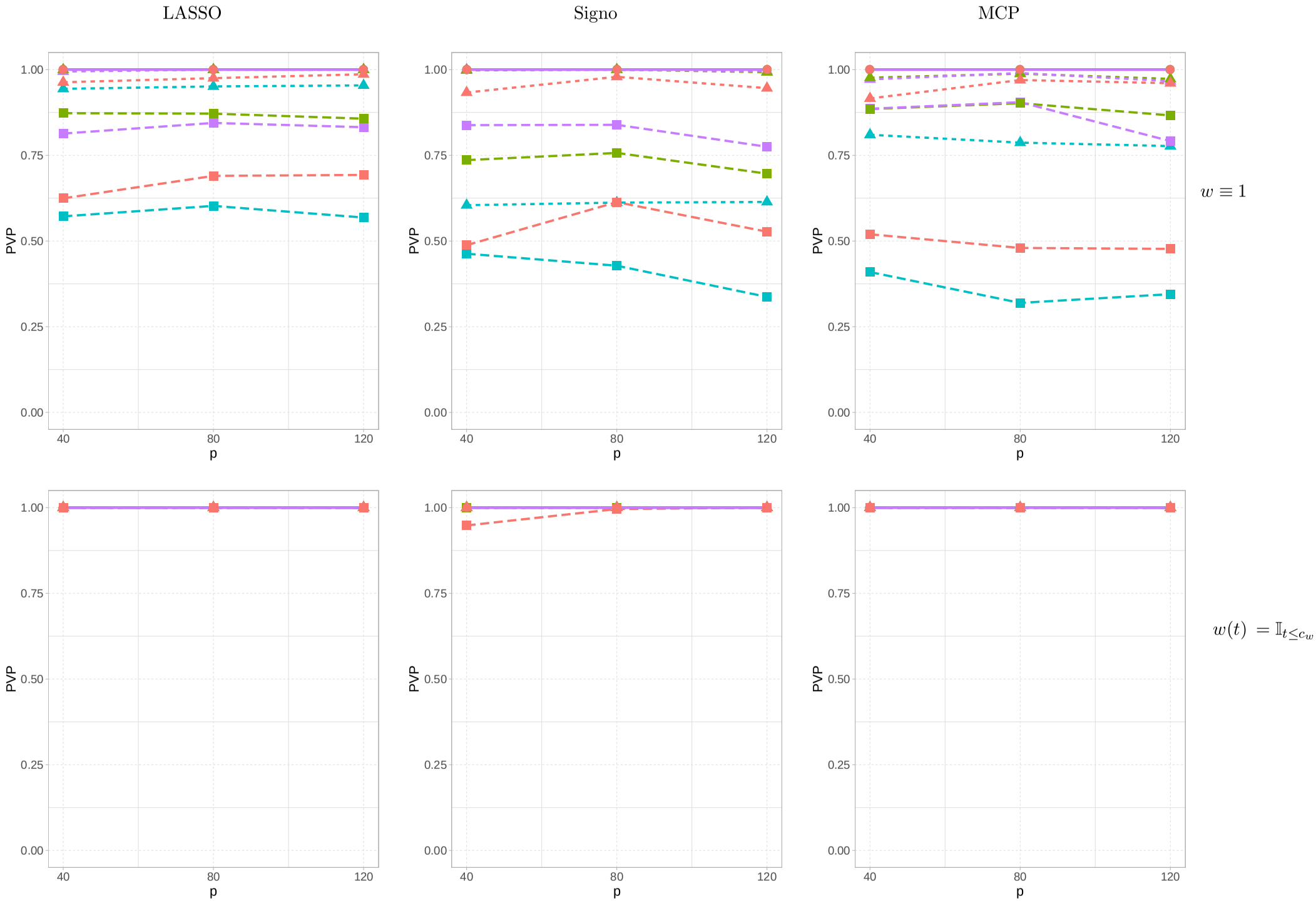


Figura 7.6: Proporción de Verdaderos Positivos para $n = 300$ en los escenarios **C0**, **CA1** y **CA2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CA1** y **CA2**, respectivamente. Por otra parte, las líneas roja, verde, celeste y violeta corresponden a $\rho = \rho_c$ dada en (3.5), $\rho = \rho_{\text{DIV}}$, $\rho(t) = t$ y $\rho(t) = 1 - \exp(-t)$.

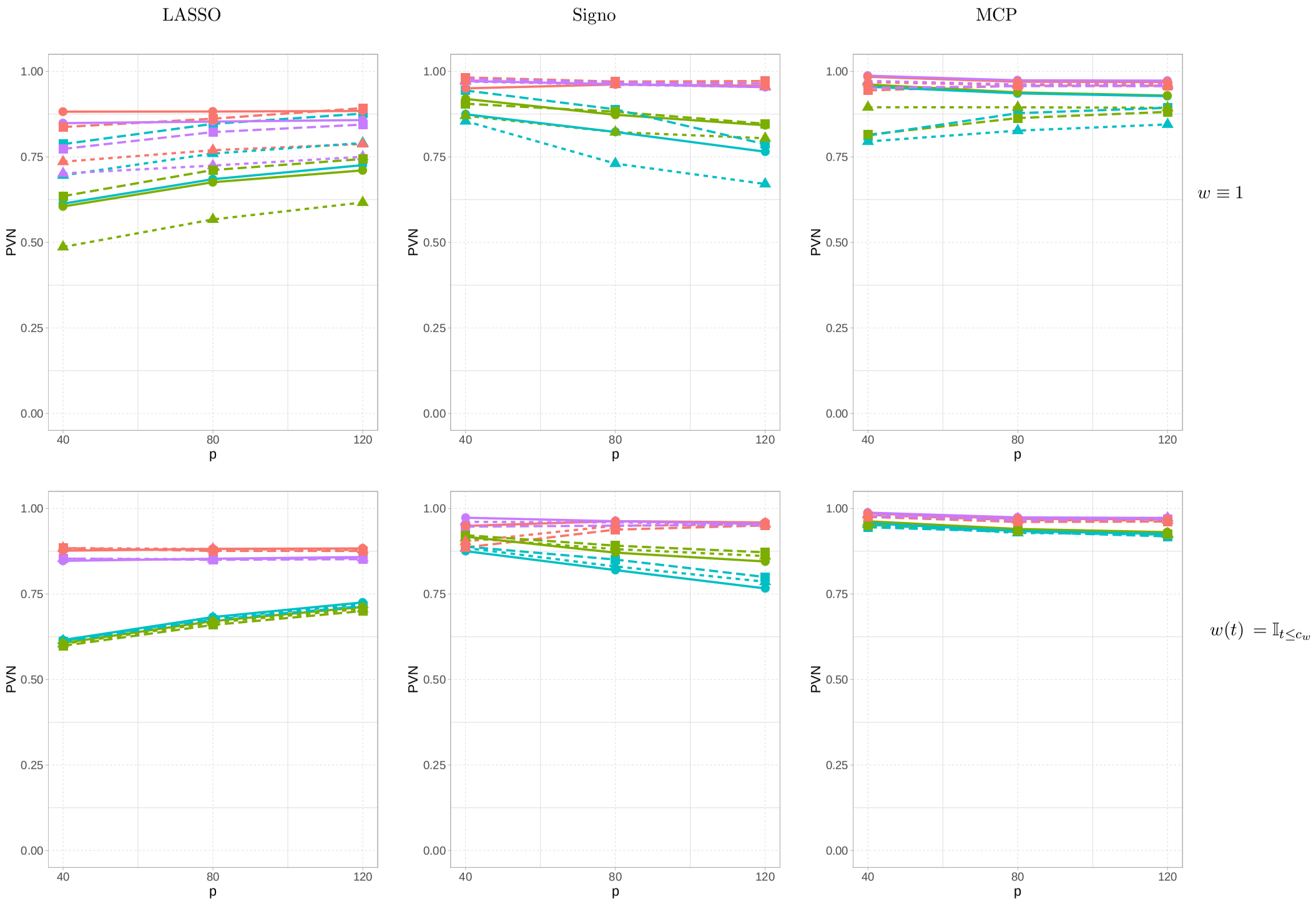


Figura 7.7: Proporción de Verdaderos Negativos para $n = 150$ en los escenarios **C0**, **CA1** y **CA2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CA1** y **CA2**, respectivamente. Por otra parte, las líneas roja, verde, celeste y violeta corresponden a $\rho = \rho_c$ dada en (3.5), $\rho = \rho_{\text{DIV}}$, $\rho(t) = t$ y $\rho(t) = 1 - \exp(-t)$.

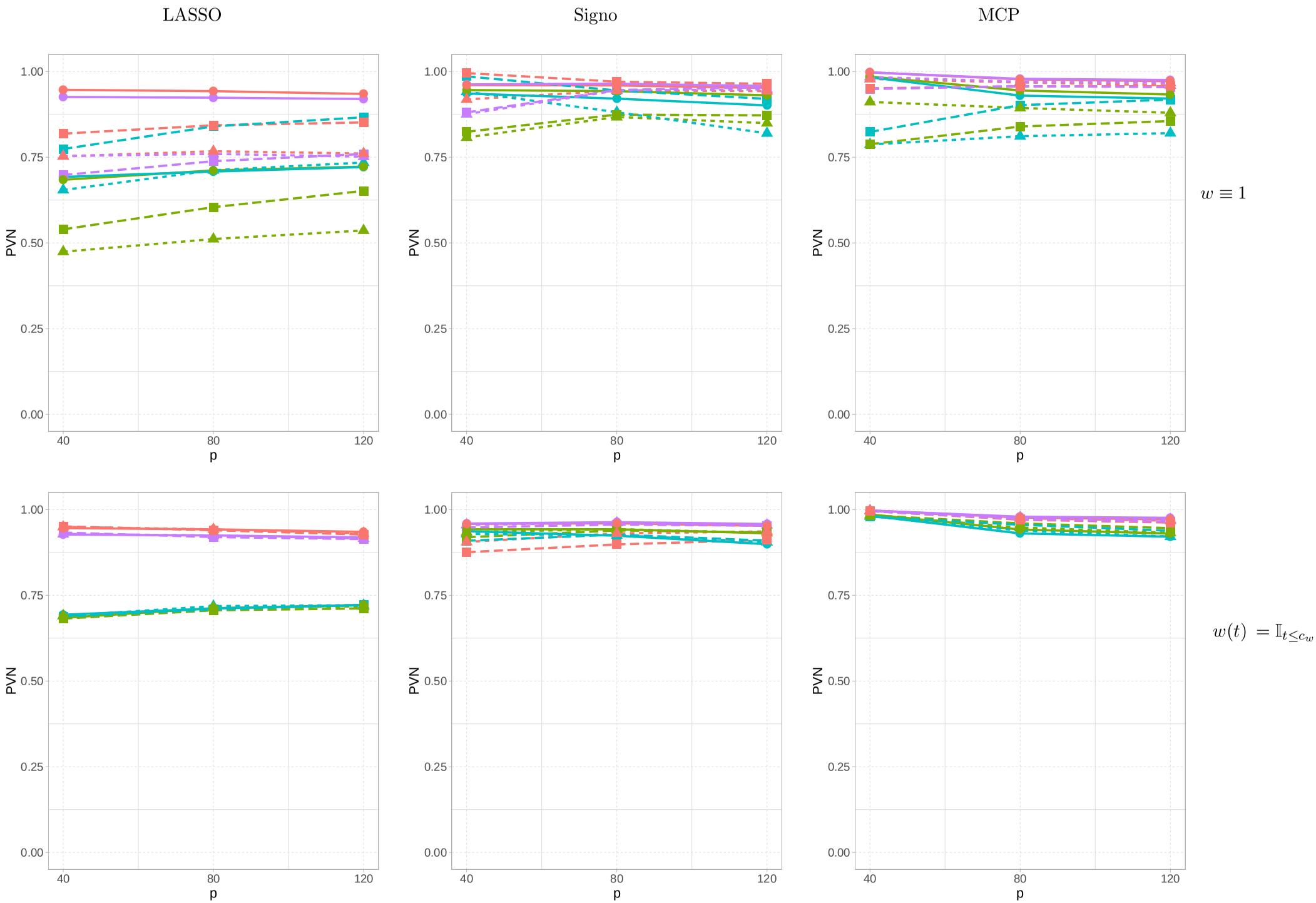


Figura 7.8: Proporción de Verdaderos Negativos para $n = 300$ en los escenarios **C0**, **CA1** y **CA2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CA1** y **CA2**, respectivamente. Por otra parte, las líneas roja, verde, celeste y violeta corresponden a $\rho = \rho_c$ dada en (3.5), $\rho = \rho_{\text{DIV}}$, $\rho(t) = t$ y $\rho(t) = 1 - \exp(-t)$.

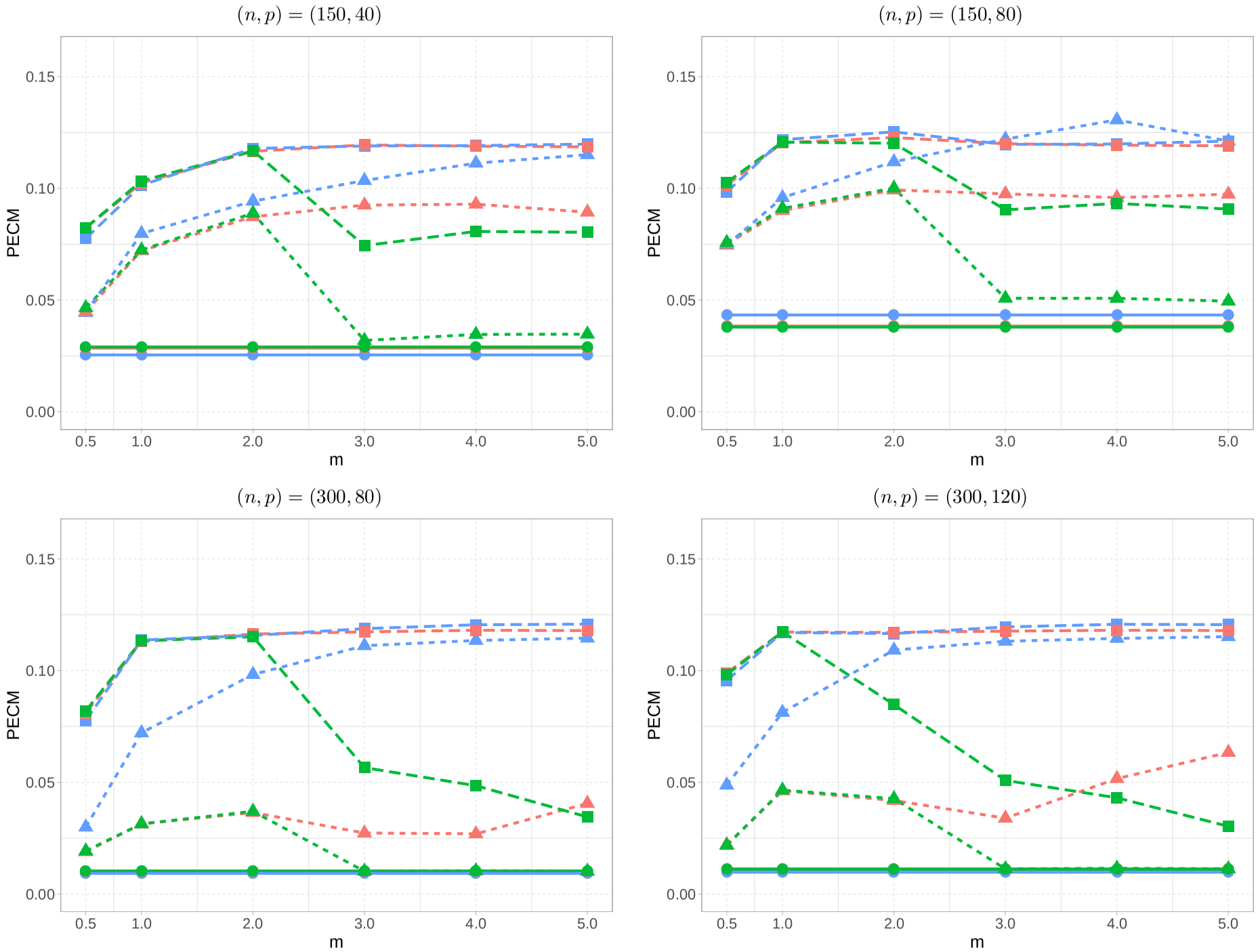


Figura 7.9: Error Cuadrático Medio de las Probabilidades para la penalización Signo en los escenarios **C0**, **CB1** y **CB2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CB1** y **CB2**, respectivamente. Por otra parte, las líneas roja, verde y celeste corresponden a $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$ y $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$.

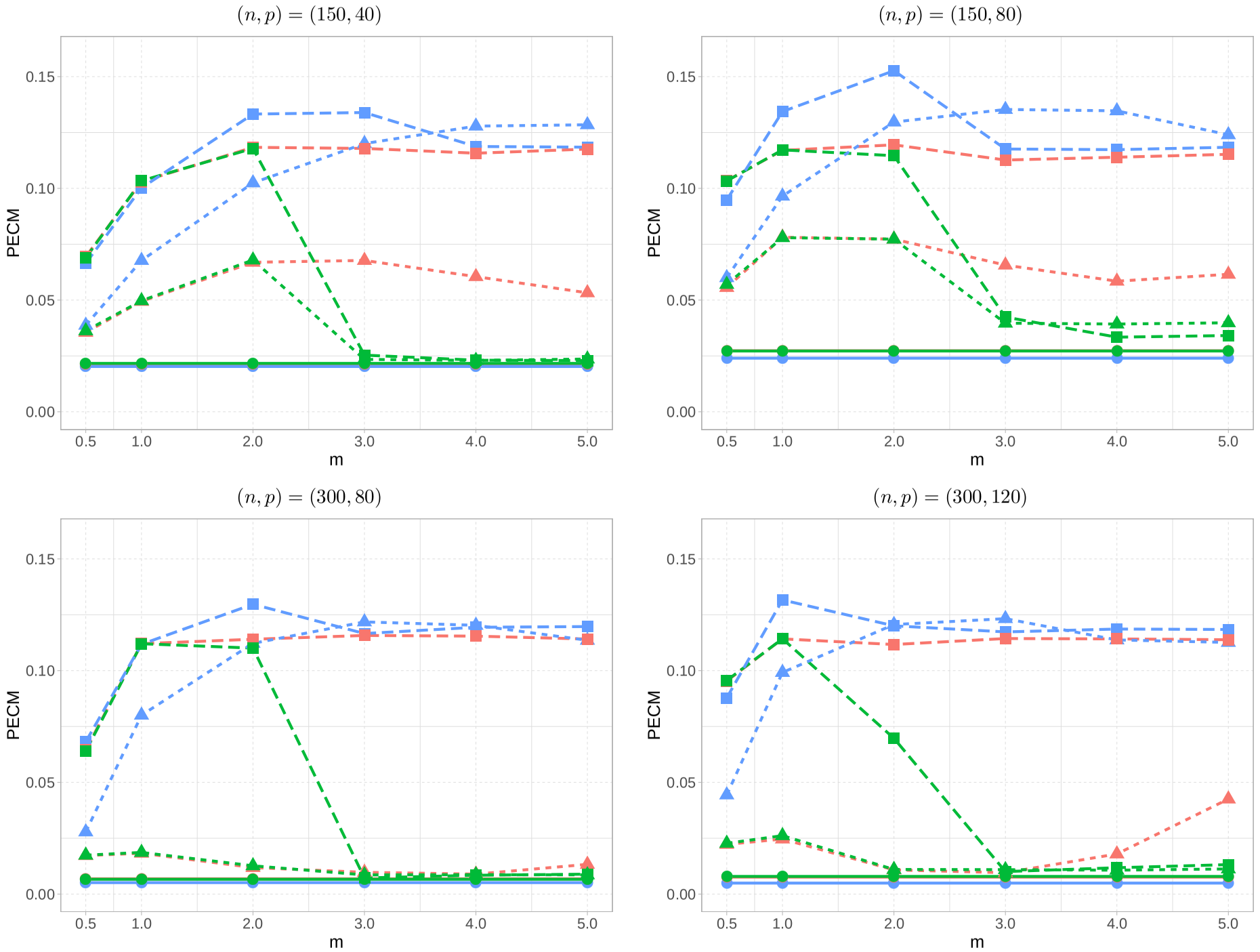


Figura 7.10: Error Cuadrático Medio de las Probabilidades para la penalización MCP en los escenarios **C0**, **CB1** y **CB2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CB1** y **CB2**, respectivamente. Por otra parte, las líneas roja, verde y celeste corresponden a $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{MCP}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{MCP}}$ y $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}^{\mathrm{MCP}}$.

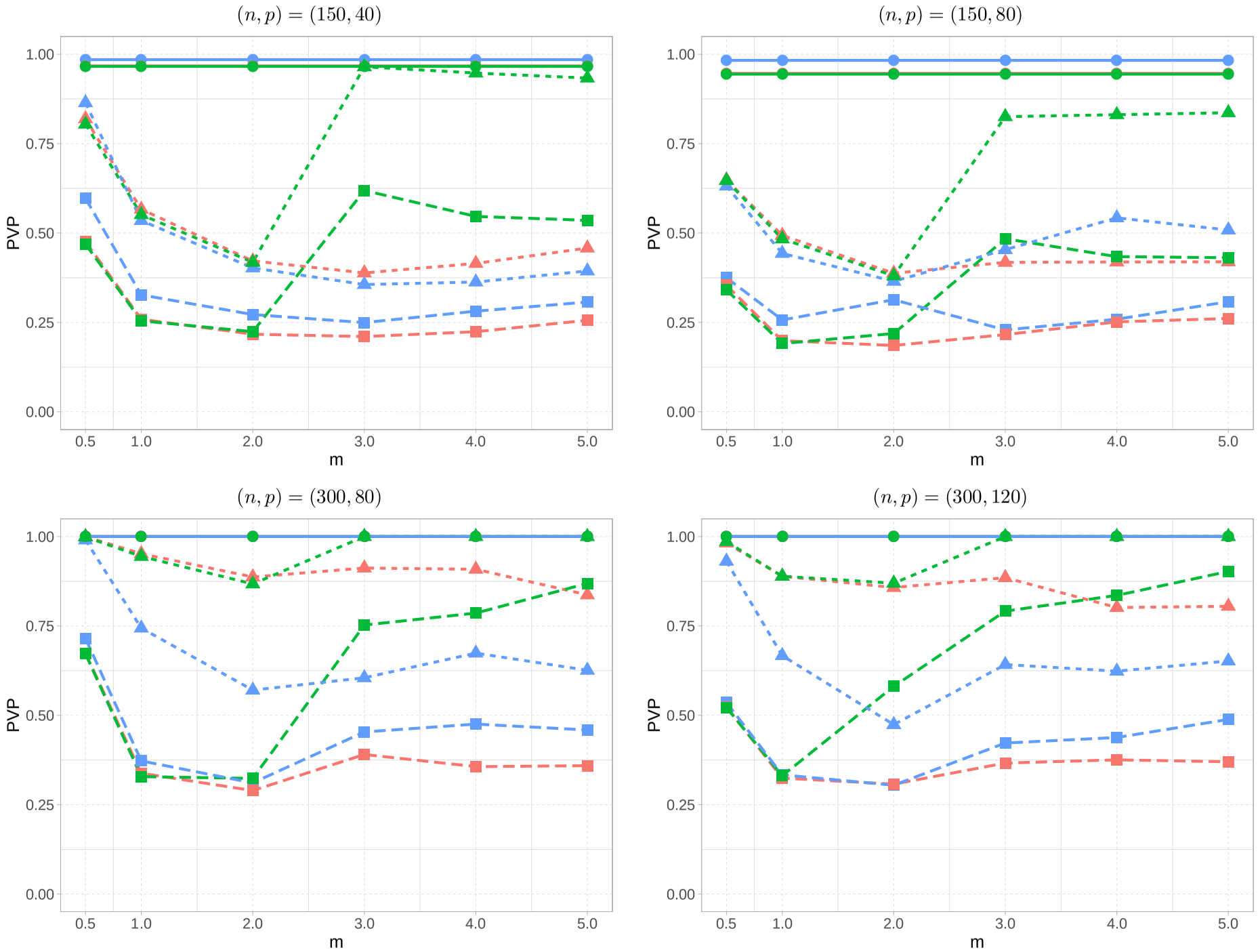


Figura 7.11: Proporción de Verdaderos Positivos para la penalización Signo en los escenarios **C0**, **CB1** y **CB2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CB1** y **CB2**, respectivamente. Por otra parte, las líneas roja, verde y celeste corresponden a $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$ y $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$.

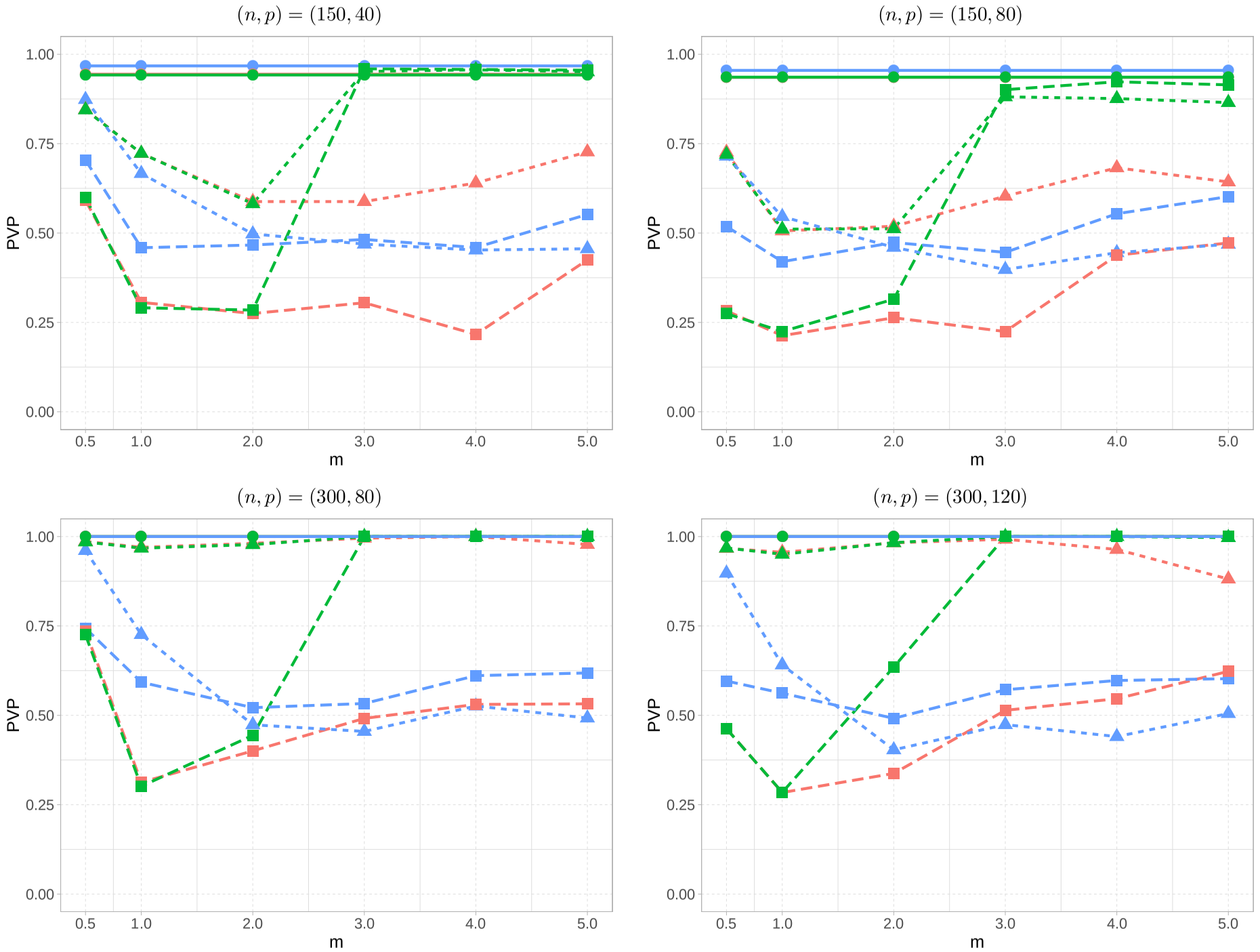


Figura 7.12: Proporción de Verdaderos Positivos para la penalización MCP en los escenarios **C0**, **CB1** y **CB2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CB1** y **CB2**, respectivamente. Por otra parte, las líneas roja, verde y celeste corresponden a $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{MCP}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{MCP}}$ y $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}^{\mathrm{MCP}}$.

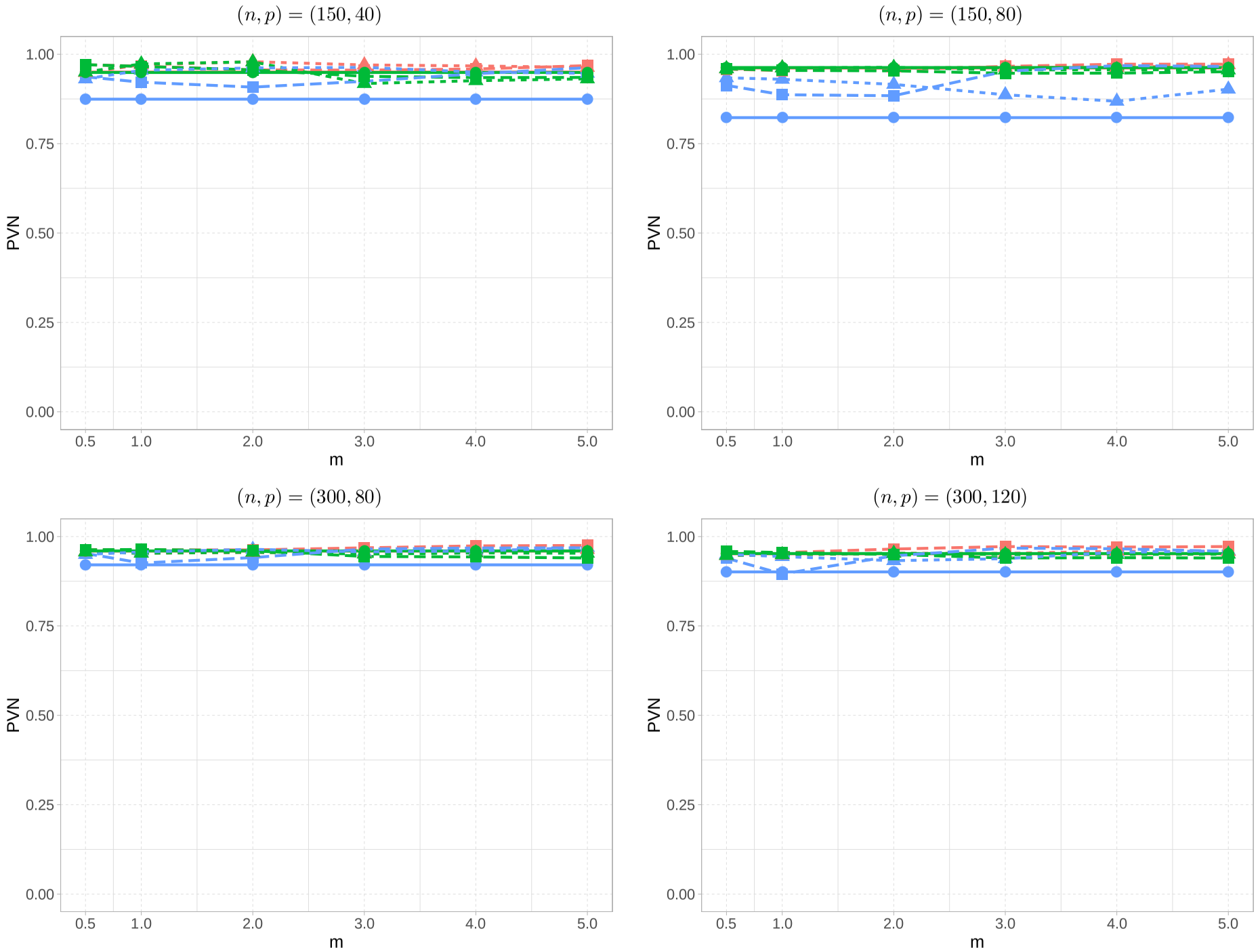


Figura 7.13: Proporción de Verdaderos Negativos para la penalización Signo en los escenarios **C0**, **CB1** y **CB2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CB1** y **CB2**, respectivamente. Por otra parte, las líneas roja, verde y celeste corresponden a $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{M}}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{WM}}$ y $\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\mathrm{MV}}$.

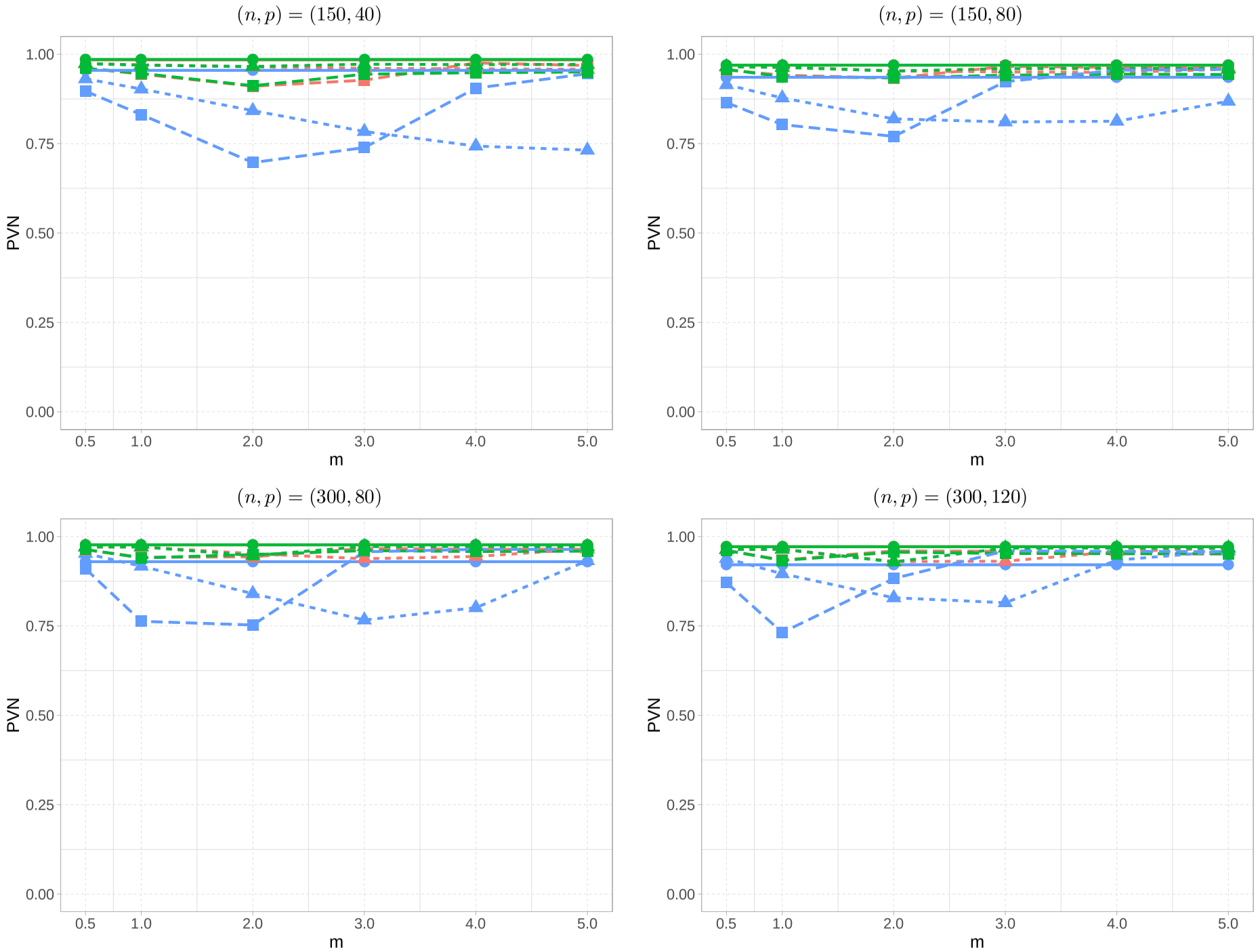


Figura 7.14: Proporción de Verdaderos Negativos para la penalización MCP en los escenarios **C0**, **CB1** y **CB2**. La línea sólida corresponde a **C0**, mientras que los guiones cortos con triángulos y los guiones largos con cuadrados a **CB1** y **CB2**, respectivamente. Por otra parte, las líneas roja, verde y celeste corresponden a $\widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{MCP}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{MCP}}$ y $\widehat{\boldsymbol{\beta}}_{\mathrm{MV}}^{\mathrm{MCP}}$.

7.3. Análisis de datos reales

7.3.1. Imágenes de SPECT

Las imágenes SPECT (Single Positron Emission Computed Tomography) son usadas como una herramienta de diagnóstico para la perfusión miocárdica. Esta técnica es muy popular debido a su alta relación señal/ruido y su relativo bajo costo. Sin embargo, las interpretaciones subjetivas de estas imágenes pueden ser equivocadas. Esto sugiere que un método computacional automático puede ser adecuado para complementar la apreciación del médico.

Con el objetivo de semi-automatizar el proceso de diagnóstico, $p = 44$ covariables fueron generadas a partir de cada imagen, tal como se describe en Kurgan *et al.* (2001). Se quiere clasificar la situación cardíaca de cada paciente en dos categorías: "Normal" y "Anormal". Este conjunto de datos consiste de $n = 267$ pacientes, 212 de ellos fueron clasificados en la categoría "Normal" y los restantes 55 en la categoría "Anormal". Los datos también están disponibles en el repositorio UCI (`https://archive.ics.uci.edu/ml/datasets/SPECT+Heart`).

Se realiza un análisis implementando un procedimiento de convalidación cruzada tal como se describe a continuación. El conjunto de datos completo fue dividido en 10 subconjuntos de tamaños aproximadamente iguales. Para cada subconjunto i ($1 \leq i \leq 10$) y cada estimador $\widehat{\boldsymbol{\beta}}$ del parámetro de regresión y $\widehat{\gamma}$ de la ordenada al origen, computamos $(\widehat{\boldsymbol{\beta}}^{(-i)}, \widehat{\gamma}^{(-i)})$, la estimación correspondiente utilizando todos los datos salvo los del subconjunto i. Luego, se clasifica cada observación s perteneciente al subconjunto i de acuerdo al signo de $\mathbf{X}_s^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(-i)} + \widehat{\gamma}^{(-i)}$. Denotamos como t_i a la proporción de observaciones correctamente clasificadas del subconjunto i. Más aún, definimos a m_i como la proporción de observaciones bien clasificadas del subconjunto i que pertenecen a la categoría "Anormal". Se define b_i de forma análoga para la categoría "Normal". Por otra parte, denotamos como a_i al número de coordenadas activas (es decir, distintas de cero) del vector $\widehat{\boldsymbol{\beta}}^{(-i)}$. Finalmente, promediamos las 10 cantidades $t_1, \ldots, t_{10}$; $m_1, \ldots, m_{10}$; $b_1, \ldots, b_{10}$ y $a_1, \ldots, a_{10}$, obteniendo valores PCC_{total}, PCC_{nor}, PCC_{anor} y $\mathrm{N}_{activas}$, respectivamente. Para cada estimador, estas medidas se muestran en la Tabla 7.14. Cabe mencionar los valores de los estimadores de máxima verosimilitud penalizados con penalidad LASSO y MCP se obtuvieron utilizando los paquetes `glmnet` y `cvplogistic`, respectivamente. Los valores entre paréntesis corresponden al desvío estándar de los 10 elementos que dieron origen a cada uno de los valores. Las medidas PCC_{nor} y PCC_{anor} también se muestran en la Figura 7.15 en la que las barras de color gris claro corresponden a las imágenes clasificadas como "Normales"y las negras a las clasificadas como "Anormales".

Los resultados reportados en la Tabla 7.14 permiten observar que los valores más altos de PCC_{total} se alcanzan al utilizar los $M-$estimadores pesados con la función de pérdida introducida en Croux y Haesbroeck (2003). Es interesante notar que al utilizar la penalización LASSO, la proporción de clasificaciones correctas en la categoría "Normal"es siempre muy baja, con valores iguales a 0 en el caso del $M-$estimador basado en las pérdidas de mínimos cuadrados y $\rho = \rho_c$ dada en (3.5), es decir, estos métodos tienden a clasificar a casi todas las observaciones como normales, de allí los valores altos de PCC_{nor} reportados en la Tabla 7.14 y Figura 7.15. Las penalizaciones Signo y MCP conducen a una mayor proporción de clasificaciones correctas entre las observaciones de categoría

"Anormal", a costa de solo un leve decrecimiento de la proporción de clasificaciones correctas en la otra categoría. Por ejemplo, el estimador $\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{WM}}$ clasifica correctamente a más del 50 % de las observaciones de la categoría "Anormal"y a más del 90 % de las de categoría "Normal", siendo el método con mejores resultados para este ejemplo.

Respecto a la cantidad de variables elegidas, al igual que en los otros ejemplos y simulaciones de esta tesis, puede observarse que la penalización MCP es la que arroja estimadores más ralos, seguida por Signo, mientras que LASSO es la penalidad que selecciona mayor cantidad de variables, en todos los casos.

	PCC_{total}	PCC_{anor}	PCC_{nor}	$\text{N}_{activas}$
$\widehat{\boldsymbol{\beta}}^{\text{L}}_{\text{MV}}$	0.802 (0.03)	0.060 (0.13)	0.995 (0.02)	9.700 (1.70)
$\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{MV}}$	0.790 (0.07)	0.473 (0.18)	0.873 (0.09)	7.100 (1.97)
$\widehat{\boldsymbol{\beta}}^{\text{L}}_{\text{MC}}$	0.794 (0.01)	0.000 (0.00)	1.000 (0.00)	18.500 (2.01)
$\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{MC}}$	0.794 (0.05)	0.390 (0.19)	0.901 (0.04)	9.600 (3.20)
$\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{MC}}$	0.798 (0.04)	0.350 (0.24)	0.915 (0.06)	5.000 (1.15)
$\widehat{\boldsymbol{\beta}}^{\text{L}}_{\text{DIV}}$	0.806 (0.05)	0.277 (0.13)	0.944 (0.05)	23.000 (2.26)
$\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{DIV}}$	0.805 (0.05)	0.440 (0.19)	0.901 (0.06)	15.600 (3.92)
$\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{DIV}}$	0.817 (0.04)	0.493 (0.15)	0.901 (0.06)	8.700 (1.25)
$\widehat{\boldsymbol{\beta}}^{\text{L}}_{\text{M}}$	0.794 (0.01)	0.000 (0.00)	1.000 (0.00)	17.700 (2.31)
$\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{M}}$	0.798 (0.03)	0.457 (0.20)	0.887 (0.05)	10.800 (3.29)
$\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{M}}$	0.791 (0.05)	0.330 (0.23)	0.910 (0.06)	5.100 (1.10)
$\widehat{\boldsymbol{\beta}}^{\text{L}}_{\text{WM}}$	0.802 (0.04)	0.227 (0.18)	0.953 (0.04)	21.600 (1.58)
$\widehat{\boldsymbol{\beta}}^{\text{S}}_{\text{WM}}$	0.821 (0.06)	0.443 (0.15)	0.920 (0.06)	16.800 (3.39)
$\widehat{\boldsymbol{\beta}}^{\text{MCP}}_{\text{WM}}$	0.840 (0.06)	0.570 (0.20)	0.911 (0.05)	10.600 (2.22)

Tabla 7.14: Proporción de clasificaciones correctas y número de variables seleccionadas para el conjunto de datos sobre imágenes SPECT. Los desvíos estándar se muestran entre paréntesis

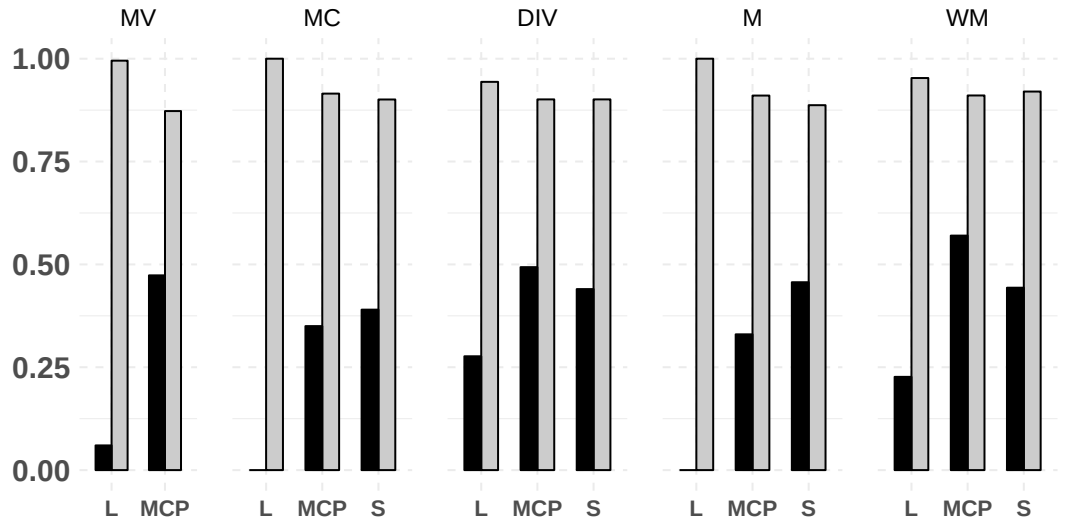


Figura 7.15: Gráfico de barras de las medidas PCC_{nor} (en gris claro) y PCC_{anor} (en negro) para el conjunto de datos sobre imágenes SPECT.

7.3.2. Diagnóstico de cáncer de mama

En este ejemplo, se analizan datos provenientes de $n = 569$ imágenes digitalizadas de biopsias de tumores localizados en la zona mamaria. Se medieron las siguientes 10 características para cada núcleo celular presente en cada imagen: radio, textura, perímetro, área, suavidad, compacidad, intensidad de la concavidad, número de puntos de concavidad, simetría y dimensión fractal. Para cada una de estas 10 características, se consideró la media, el desvío estándar y el máximo entre todos los núcleos presentes en la imagen. De este modo, se generaron $p = 30$ covariables por imagen. Los datos pueden obtenerse en el repositorio UCI `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29`.

De los n tumores biopsiados, 357 eran benignos y 212 malignos. Se quiere predecir la clase de tumor en función de las 30 características de cada imagen.

Para este conjunto de datos, analizaremos cuál es el cambio en la selección de variables de los distintos métodos al agregar n_0 datos atípicos artificialmente. Cada dato atípico $(\widetilde{Y}, \widetilde{\mathbf{X}})$ se genera del modo que se describe a continuación. Primero, se calcula el $M-$estimador con penalidad MCP, $(\widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{MCP}}, \widehat{\gamma}_{\mathrm{M}}^{\mathrm{MCP}})$ obtenido con el conjunto de datos original. Luego, se toma $\widetilde{\mathbf{X}} \sim N(\mathbf{0}, 100\,\mathbf{I})$ y se define una respuesta mal clasificada de la siguiente forma

$$\widetilde{Y} = \begin{cases} 1 & \text{si } \widetilde{\mathbf{X}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{MCP}} + \widehat{\gamma}_{\mathrm{M}}^{\mathrm{MCP}} < 0 \\ 0 & \text{si } \widetilde{\mathbf{X}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\mathrm{M}}^{\mathrm{MCP}} + \widehat{\gamma}_{\mathrm{M}}^{\mathrm{MCP}} \geq 0. \end{cases} \tag{7.2}$$

Se comparan los resultados obtenidos con los procedimientos basados en la pérdida $\rho(t) = t$ con penalidad LASSO y MCP y con los estimadores de mínimos cuadrados, de divergencia y los $M-$estimadores con función $\rho = \rho_c$ dada en (3.5) con pesos y sin ellos, en todos los casos con penalidades Signo o MCP.

Se repitió el procedimiento de generación de los datos atípicos calculando el $M-$estimador pesado con penalidad MCP, $(\widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{MCP}}, \widehat{\gamma}_{\mathrm{WM}}^{\mathrm{MCP}})$ obtenido con el conjunto de datos original y definiendo una respuesta mal clasificada de la siguiente forma

$$\widetilde{Y} = \begin{cases} 1 & \text{si } \widetilde{\mathbf{X}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{MCP}} + \widehat{\gamma}_{\mathrm{WM}}^{\mathrm{MCP}} < 0 \\ 0 & \text{si } \widetilde{\mathbf{X}}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{\mathrm{WM}}^{\mathrm{MCP}} + \widehat{\gamma}_{\mathrm{WM}}^{\mathrm{MCP}} \geq 0. \end{cases} \tag{7.3}$$

Para cada método considerado, la cantidad n_0 de observaciones mal clasificadas artificialmente generadas varía en el conjunto $\{0, 20, 40, 80\}$. Para evaluar las propiedades de selección de variables de los procedimientos considerados, se divide a cada conjunto de datos (posiblemente con datos atípicos agregados de este modo) en 10 subconjuntos. Para cada subconjunto i $(1 \leq i \leq 10)$ y cada estimador $\widehat{\boldsymbol{\beta}}$, computamos $\widehat{\boldsymbol{\beta}}^{(-i)}$, la estimación correspondiente utilizando todos los datos salvo los del subconjunto i. Luego, para cada coordenada $1 \leq j \leq 30$, se define $\Pi_{a,j}$ como la proporción de veces que la coordenada j es estimada como activa por el estimador $\widehat{\boldsymbol{\beta}}$, es decir,

$$\Pi_{a,j} = \frac{\#\{i : \widehat{\boldsymbol{\beta}}_j^{(-i)} \neq 0\}}{10}.$$

Notemos que esta cantidad depende del estimador utilizado y de n_0. En cada fila de los gráficos de las Figuras 7.16 y 7.17 se muestran, para cada estimador, una representación en escala de grises

de los valores $\Pi_{a,1} \ldots, \Pi_{a,30}$ en función del valor n_0. La primer figura corresponde al caso de los datos atípicos generados como se indica en (7.2), mientras que la Figura 7.17 al caso de datos generados según (7.3). Se puede observar que la selección de variables de los estimadores no robustos $\widehat{\boldsymbol{\beta}}^{\mathrm{L}}_{\mathrm{MV}}$ y $\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{MV}}$ es muy sensible ante esta contaminación, lo que refleja una selección de variables muy inestable. Por otra parte, los métodos robustos basados en $\rho = \rho_c$ seleccionan aproximadamente el mismo subconjunto de covariables, sin importar cuál sea el valor de n_0 mostrando una identificación de variables activas más estable. Al igual que en los ejemplos anteriores, los estimadores con penalización MCP son más ralos que al utilizar la penalización Signo, lo que puede explicarse por las propiedades de selección de variables obtenidas en el Capítulo 5. Finalmente, si bien los datos atípicos generados por ambos mecanismos no son los mismos, el patrón de comportamiento de los distintos estimadores es semejante en ambas perturbaciones. Cabe destacar que los outliers generados a partir del $M-$estimador pesado penalizado parecen ser más dañinos ya que producen estimaciones menos ralas para los estimadores no robustos.

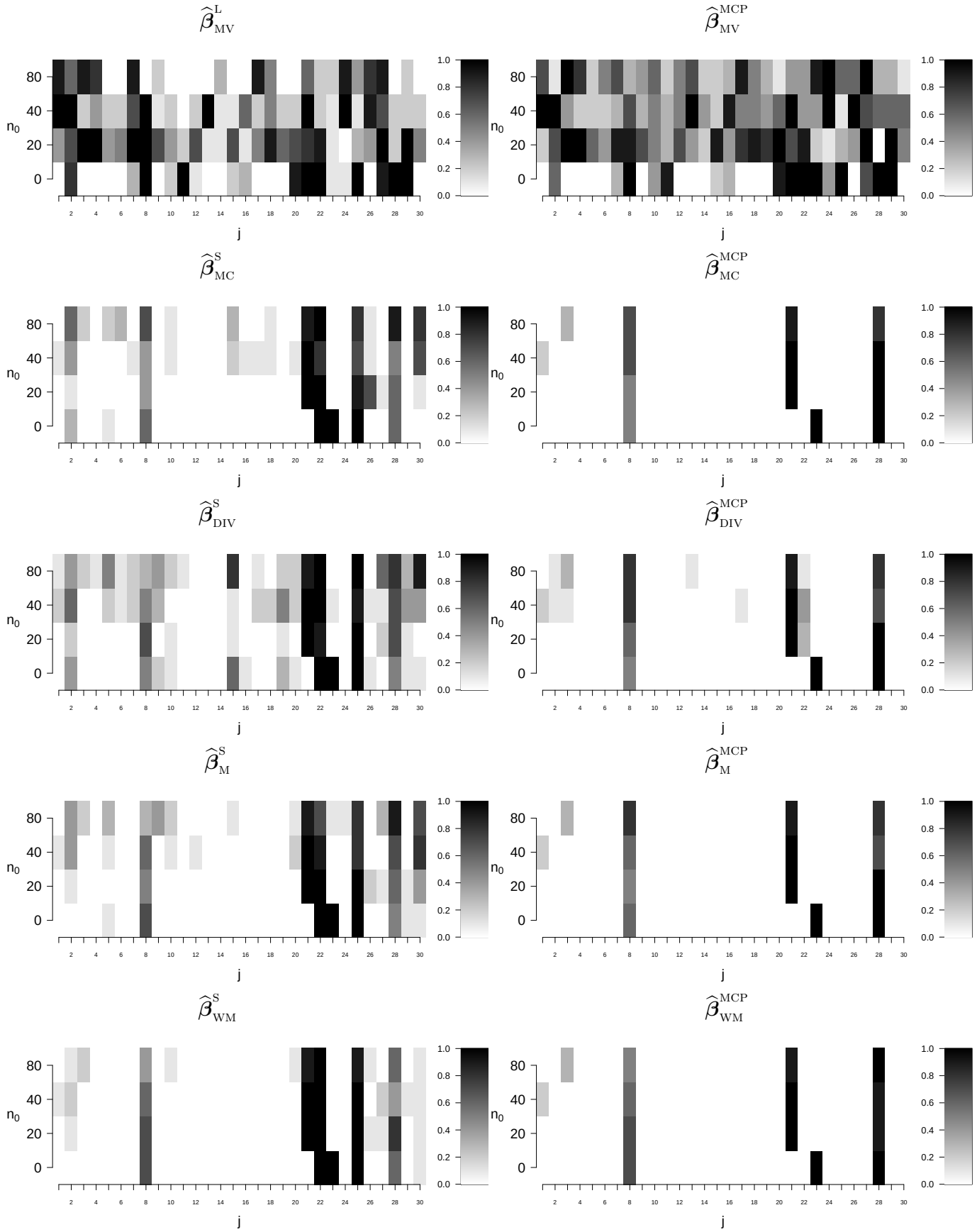


Figura 7.16: Escala de grises de las medidas $\Pi_{a,1}, \ldots, \Pi_{a,30}$ para cada método y número de datos atípicos agregados artificialmente usando el estimador $(\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{M}}, \widehat{\gamma}^{\mathrm{MCP}}_{\mathrm{M}})$.

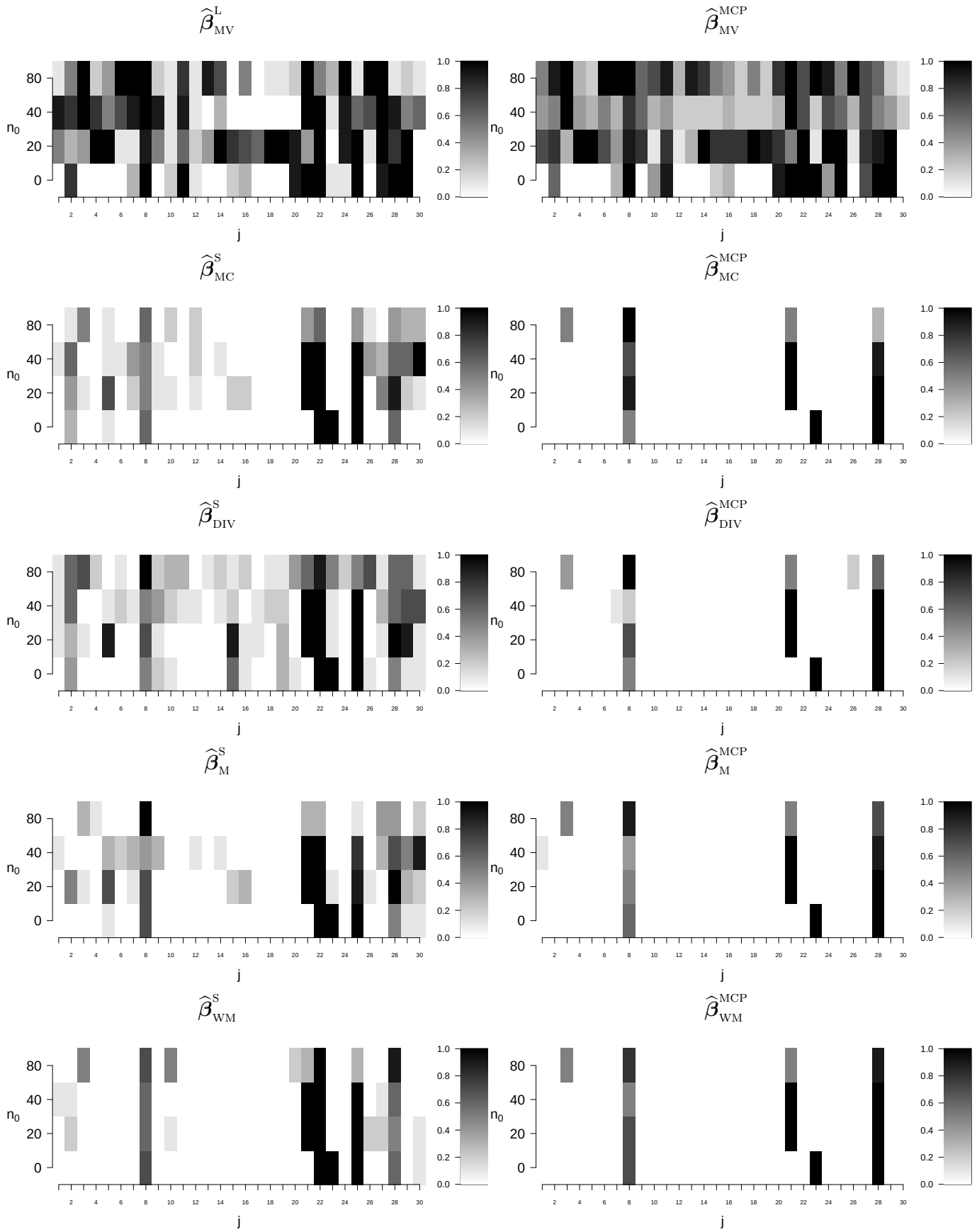


Figura 7.17: Escala de grises de las medidas $\Pi_{a,1}, \ldots, \Pi_{a,30}$ para cada método y número de datos atípicos agregados artificialmente usando el estimador $(\widehat{\boldsymbol{\beta}}^{\mathrm{MCP}}_{\mathrm{WM}}, \widehat{\gamma}^{\mathrm{MCP}}_{\mathrm{WM}})$.

7.4. Apéndice A: Tablas adicionales

En esta sección, incluímos las tablas correspondientes a las medidas PECM y PCC. Observemos que dichas medidas dependen no sólo del estimador del parámetro de regresión, sino que también dependen del estimador de la ordenada al origen. Por esta razón, indicaremos por $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\gamma})$.

n	150			300				150			300		
p	40	80	120	40	80	120		40	80	120	40	80	120
$\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}$	0.095	0.162	0.170	0.032	0.091	0.160	$\widehat{\boldsymbol{\theta}}_{\mathrm{WMV}}$	0.096	0.162	0.170	0.033	0.092	0.160
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MV}}$	0.022	0.027	0.033	0.012	0.014	0.015	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMV}}$	0.022	0.027	0.034	0.012	0.014	0.015
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.025	0.043	0.089	0.007	0.009	0.010	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMV}}$	0.026	0.044	0.092	0.007	0.009	0.010
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.020	0.024	0.031	0.004	0.005	0.005	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	0.020	0.024	0.032	0.004	0.005	0.005
$\widehat{\boldsymbol{\theta}}_{\mathrm{MC}}$	0.134	0.146	0.155	0.113	0.125	0.134	$\widehat{\boldsymbol{\theta}}_{\mathrm{WMC}}$	0.134	0.146	0.156	0.114	0.126	0.134
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MC}}$	0.046	0.047	0.051	0.037	0.038	0.038	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMC}}$	0.046	0.047	0.051	0.037	0.038	0.038
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MC}}$	0.046	0.060	0.078	0.009	0.012	0.013	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMC}}$	0.046	0.061	0.079	0.009	0.012	0.013
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	0.027	0.033	0.046	0.005	0.007	0.009	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	0.027	0.033	0.046	0.005	0.007	0.008
$\widehat{\boldsymbol{\theta}}_{\mathrm{DIV}}$	0.137	0.152	0.162	0.065	0.131	0.140	$\widehat{\boldsymbol{\theta}}_{\mathrm{WDIV}}$	0.137	0.151	0.162	0.068	0.131	0.140
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.025	0.030	0.036	0.015	0.017	0.018	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	0.025	0.030	0.036	0.015	0.017	0.018
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.032	0.047	0.076	0.008	0.010	0.011	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.032	0.047	0.077	0.008	0.011	0.011
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.022	0.027	0.037	0.005	0.006	0.006	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.022	0.026	0.037	0.005	0.006	0.006
$\widehat{\boldsymbol{\theta}}_{\mathrm{M}}$	0.141	0.159	0.166	0.038	0.139	0.147	$\widehat{\boldsymbol{\theta}}_{\mathrm{WM}}$	0.142	0.159	0.166	0.039	0.139	0.149
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{M}}$	0.051	0.052	0.055	0.042	0.042	0.043	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.051	0.052	0.055	0.042	0.043	0.043
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.029	0.038	0.058	0.008	0.010	0.011	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.029	0.038	0.057	0.008	0.010	0.011
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.021	0.027	0.038	0.005	0.007	0.008	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.022	0.027	0.038	0.005	0.007	0.008

Tabla 7.15: Promedio podado al 10 % del error cuadrático medio de la probabilidades (PECM), bajo **C0**.

n	150			300				150			300		
p	40	80	120	40	80	120		40	80	120	40	80	120
$\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}$	0.898	0.873	0.862	0.950	0.901	0.876	$\widehat{\boldsymbol{\theta}}_{\mathrm{WMV}}$	0.897	0.874	0.861	0.951	0.900	0.876
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MV}}$	0.956	0.948	0.934	0.983	0.977	0.975	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMV}}$	0.956	0.948	0.934	0.983	0.977	0.975
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.952	0.934	0.911	0.985	0.980	0.979	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMV}}$	0.951	0.935	0.907	0.986	0.980	0.979
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.961	0.954	0.941	0.990	0.987	0.986	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	0.960	0.955	0.939	0.990	0.987	0.987
$\widehat{\boldsymbol{\theta}}_{\mathrm{MC}}$	0.908	0.889	0.878	0.936	0.920	0.908	$\widehat{\boldsymbol{\theta}}_{\mathrm{WMC}}$	0.910	0.889	0.878	0.936	0.919	0.907
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MC}}$	0.936	0.935	0.921	0.973	0.970	0.971	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMC}}$	0.937	0.933	0.920	0.973	0.970	0.971
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MC}}$	0.921	0.898	0.867	0.982	0.976	0.974	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMC}}$	0.922	0.898	0.862	0.982	0.977	0.975
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	0.952	0.942	0.923	0.988	0.984	0.981	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	0.953	0.941	0.924	0.988	0.983	0.981
$\widehat{\boldsymbol{\theta}}_{\mathrm{DIV}}$	0.907	0.884	0.869	0.942	0.913	0.901	$\widehat{\boldsymbol{\theta}}_{\mathrm{WDIV}}$	0.908	0.885	0.870	0.940	0.913	0.900
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.955	0.947	0.930	0.983	0.977	0.974	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	0.955	0.946	0.931	0.982	0.977	0.974
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.942	0.931	0.917	0.983	0.979	0.978	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.942	0.931	0.915	0.983	0.978	0.977
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.958	0.951	0.934	0.989	0.986	0.986	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.959	0.951	0.934	0.989	0.986	0.985
$\widehat{\boldsymbol{\theta}}_{\mathrm{M}}$	0.900	0.876	0.865	0.948	0.902	0.892	$\widehat{\boldsymbol{\theta}}_{\mathrm{WM}}$	0.899	0.876	0.865	0.947	0.902	0.890
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{M}}$	0.927	0.925	0.912	0.967	0.964	0.965	$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.926	0.924	0.912	0.967	0.963	0.964
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.948	0.925	0.881	0.984	0.978	0.978	$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.946	0.926	0.882	0.984	0.978	0.978
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.958	0.947	0.930	0.989	0.984	0.982	$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.959	0.949	0.931	0.989	0.984	0.982

Tabla 7.16: Promedio podado al 10 % de la proporción de clasificaciones correctas (PCC), bajo **C0**.

	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
	$n = 150$			$n = 300$			$n = 150$			$n = 300$		
p	40	80	120	40	80	120	40	80	120	40	80	120
$\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}$	0.093	0.193	0.204	0.058	0.093	0.179	0.104	0.217	0.230	0.079	0.103	0.157
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MV}}$	0.077	0.079	0.082	0.069	0.071	0.071	0.105	0.106	0.107	0.104	0.104	0.105
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.083	0.096	0.143	0.078	0.084	0.080	0.102	0.105	0.130	0.097	0.102	0.104
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.074	0.078	0.086	0.062	0.068	0.070	0.106	0.108	0.109	0.106	0.108	0.109
$\widehat{\boldsymbol{\theta}}_{\mathrm{MC}}$	0.146	0.165	0.177	0.116	0.138	0.150	0.173	0.194	0.206	0.100	0.162	0.175
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MC}}$	0.063	0.066	0.070	0.049	0.051	0.053	0.096	0.100	0.102	0.083	0.086	0.089
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MC}}$	0.068	0.080	0.090	0.021	0.026	0.031	0.089	0.097	0.100	0.053	0.061	0.076
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	0.046	0.061	0.078	0.016	0.020	0.025	0.079	0.096	0.106	0.037	0.047	0.064
$\widehat{\boldsymbol{\theta}}_{\mathrm{DIV}}$	0.153	0.171	0.181	0.052	0.145	0.156	0.168	0.197	0.211	0.063	0.170	0.180
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.047	0.051	0.057	0.032	0.036	0.039	0.091	0.093	0.095	0.078	0.079	0.082
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.059	0.075	0.105	0.023	0.025	0.030	0.088	0.096	0.109	0.066	0.067	0.075
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.045	0.055	0.067	0.020	0.021	0.026	0.080	0.089	0.097	0.049	0.053	0.062
$\widehat{\boldsymbol{\theta}}_{\mathrm{M}}$	0.141	0.178	0.189	0.047	0.151	0.166	0.123	0.203	0.219	0.063	0.126	0.191
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{M}}$	0.077	0.078	0.082	0.066	0.068	0.068	0.103	0.105	0.106	0.101	0.101	0.102
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.073	0.082	0.091	0.031	0.034	0.045	0.095	0.100	0.103	0.090	0.092	0.098
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.052	0.064	0.084	0.023	0.024	0.033	0.097	0.107	0.110	0.079	0.089	0.101
$\widehat{\boldsymbol{\theta}}_{\mathrm{WMV}}$	0.095	0.170	0.186	0.032	0.092	0.165	0.095	0.178	0.198	0.032	0.091	0.169
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMV}}$	0.022	0.028	0.033	0.012	0.014	0.015	0.022	0.028	0.034	0.012	0.014	0.016
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMV}}$	0.027	0.047	0.088	0.007	0.009	0.010	0.030	0.048	0.084	0.008	0.009	0.010
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	0.020	0.029	0.038	0.004	0.006	0.006	0.022	0.034	0.046	0.004	0.008	0.008
$\widehat{\boldsymbol{\theta}}_{\mathrm{WMC}}$	0.135	0.160	0.174	0.114	0.132	0.145	0.139	0.171	0.192	0.115	0.137	0.155
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMC}}$	0.046	0.047	0.051	0.037	0.038	0.039	0.046	0.047	0.051	0.037	0.038	0.039
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMC}}$	0.047	0.067	0.079	0.009	0.012	0.013	0.050	0.071	0.083	0.010	0.012	0.014
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	0.027	0.040	0.053	0.005	0.009	0.010	0.029	0.048	0.063	0.005	0.010	0.014
$\widehat{\boldsymbol{\theta}}_{\mathrm{WDIV}}$	0.139	0.164	0.178	0.067	0.136	0.150	0.141	0.173	0.194	0.066	0.140	0.157
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	0.025	0.030	0.036	0.015	0.017	0.018	0.025	0.030	0.037	0.015	0.017	0.019
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.032	0.049	0.074	0.008	0.011	0.011	0.036	0.052	0.069	0.009	0.011	0.011
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.022	0.033	0.043	0.005	0.007	0.007	0.024	0.037	0.050	0.005	0.008	0.010
$\widehat{\boldsymbol{\theta}}_{\mathrm{WM}}$	0.142	0.169	0.181	0.039	0.142	0.155	0.144	0.177	0.196	0.038	0.145	0.161
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.051	0.052	0.055	0.042	0.043	0.043	0.051	0.052	0.055	0.041	0.043	0.044
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.034	0.047	0.061	0.008	0.010	0.011	0.047	0.054	0.066	0.014	0.011	0.011
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.022	0.035	0.048	0.005	0.008	0.010	0.023	0.040	0.056	0.005	0.010	0.013

Tabla 7.17: Promedio podado al 10 % del error cuadrático medio de la probabilidades (PECM), para los escenarios de contaminación **CA1** y **CA2**.

	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
	$n = 150$			$n = 300$			$n = 150$			$n = 300$		
p	40	80	120	40	80	120	40	80	120	40	80	120
$\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}$	0.864	0.835	0.825	0.878	0.861	0.833	0.821	0.805	0.791	0.822	0.819	0.813
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MV}}$	0.813	0.799	0.790	0.875	0.856	0.864	0.679	0.673	0.658	0.696	0.680	0.674
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.804	0.831	0.845	0.822	0.794	0.807	0.713	0.712	0.749	0.764	0.708	0.696
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.833	0.825	0.811	0.857	0.832	0.836	0.672	0.665	0.661	0.667	0.641	0.641
$\widehat{\boldsymbol{\theta}}_{\mathrm{MC}}$	0.892	0.867	0.850	0.922	0.905	0.888	0.859	0.833	0.813	0.901	0.876	0.858
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{MC}}$	0.880	0.868	0.854	0.930	0.926	0.923	0.730	0.708	0.694	0.803	0.781	0.770
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MC}}$	0.866	0.827	0.801	0.957	0.950	0.939	0.783	0.748	0.733	0.886	0.863	0.820
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MC}}$	0.915	0.881	0.833	0.966	0.961	0.950	0.808	0.732	0.680	0.926	0.903	0.843
$\widehat{\boldsymbol{\theta}}_{\mathrm{DIV}}$	0.884	0.861	0.847	0.924	0.898	0.883	0.851	0.831	0.811	0.891	0.864	0.852
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{DIV}}$	0.901	0.887	0.874	0.940	0.931	0.925	0.751	0.734	0.724	0.805	0.791	0.785
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{DIV}}$	0.887	0.879	0.872	0.952	0.948	0.942	0.784	0.774	0.780	0.851	0.843	0.823
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{DIV}}$	0.912	0.891	0.871	0.956	0.956	0.949	0.815	0.785	0.759	0.896	0.884	0.861
$\widehat{\boldsymbol{\theta}}_{\mathrm{M}}$	0.875	0.852	0.837	0.916	0.880	0.867	0.840	0.825	0.800	0.874	0.849	0.838
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{M}}$	0.838	0.824	0.808	0.898	0.889	0.893	0.690	0.678	0.658	0.712	0.703	0.699
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.829	0.802	0.777	0.930	0.926	0.899	0.749	0.727	0.704	0.787	0.763	0.739
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.890	0.847	0.786	0.948	0.947	0.933	0.710	0.660	0.637	0.772	0.723	0.676
$\widehat{\boldsymbol{\theta}}_{\mathrm{WMV}}$	0.897	0.865	0.842	0.947	0.899	0.872	0.899	0.856	0.830	0.952	0.901	0.868
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMV}}$	0.958	0.943	0.932	0.980	0.978	0.978	0.958	0.950	0.932	0.984	0.980	0.976
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMV}}$	0.951	0.926	0.907	0.983	0.981	0.982	0.943	0.923	0.903	0.984	0.980	0.981
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMV}}$	0.963	0.942	0.927	0.988	0.987	0.989	0.960	0.936	0.909	0.991	0.984	0.982
$\widehat{\boldsymbol{\theta}}_{\mathrm{WMC}}$	0.908	0.874	0.855	0.934	0.913	0.894	0.904	0.863	0.829	0.936	0.906	0.883
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WMC}}$	0.942	0.932	0.921	0.970	0.972	0.974	0.941	0.941	0.923	0.974	0.972	0.972
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WMC}}$	0.924	0.886	0.858	0.980	0.979	0.978	0.924	0.884	0.844	0.983	0.977	0.976
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WMC}}$	0.954	0.926	0.902	0.986	0.982	0.982	0.950	0.916	0.879	0.989	0.980	0.974
$\widehat{\boldsymbol{\theta}}_{\mathrm{WDIV}}$	0.905	0.871	0.852	0.938	0.908	0.890	0.905	0.863	0.829	0.943	0.903	0.880
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WDIV}}$	0.957	0.941	0.930	0.980	0.978	0.978	0.958	0.949	0.930	0.984	0.980	0.976
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WDIV}}$	0.945	0.920	0.905	0.981	0.981	0.981	0.936	0.914	0.899	0.983	0.977	0.980
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WDIV}}$	0.962	0.939	0.917	0.987	0.986	0.987	0.960	0.931	0.904	0.991	0.983	0.980
$\widehat{\boldsymbol{\theta}}_{\mathrm{WM}}$	0.896	0.864	0.849	0.944	0.901	0.885	0.898	0.858	0.829	0.948	0.898	0.877
$\widehat{\boldsymbol{\theta}}^{\mathrm{L}}_{\mathrm{WM}}$	0.931	0.922	0.915	0.964	0.965	0.966	0.931	0.929	0.914	0.968	0.964	0.965
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.936	0.902	0.874	0.981	0.981	0.982	0.907	0.889	0.854	0.971	0.977	0.978
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.960	0.933	0.909	0.987	0.982	0.982	0.960	0.929	0.889	0.991	0.980	0.975

Tabla 7.18: Promedio podado al 10 % de la proporción de clasificaciones correctas (PCC), para los escenarios de contaminación **CA1** y **CA2**.

$n = 150$	$p = 40$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.044	0.080	0.094	0.103	0.111	0.115	0.078	0.101	0.118	0.119	0.119	0.120
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.039	0.068	0.102	0.120	0.128	0.128	0.066	0.100	0.133	0.134	0.119	0.118
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.045	0.072	0.087	0.093	0.093	0.089	0.082	0.102	0.117	0.119	0.119	0.119
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.036	0.049	0.067	0.068	0.060	0.053	0.069	0.103	0.118	0.118	0.116	0.118
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.047	0.072	0.089	0.032	0.035	0.035	0.082	0.103	0.117	0.074	0.081	0.080
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.036	0.050	0.068	0.023	0.023	0.024	0.069	0.103	0.118	0.025	0.023	0.023
	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.075	0.096	0.112	0.122	0.131	0.121	0.098	0.122	0.125	0.120	0.120	0.121
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.060	0.097	0.130	0.135	0.135	0.124	0.095	0.134	0.153	0.118	0.117	0.118
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.075	0.090	0.099	0.098	0.096	0.097	0.101	0.120	0.123	0.120	0.119	0.119
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.056	0.078	0.077	0.066	0.058	0.062	0.103	0.117	0.119	0.113	0.114	0.115
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.076	0.091	0.100	0.051	0.051	0.050	0.103	0.121	0.120	0.090	0.093	0.091
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.057	0.078	0.077	0.040	0.039	0.040	0.103	0.117	0.115	0.042	0.033	0.034

Tabla 7.19: Promedio podado al 10 % de la proporción de clasificaciones correctas (PCC), para los escenarios de contaminación **CB1** y **CB2** con $n = 150$.

$n = 300$	$p = 80$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.030	0.072	0.098	0.111	0.114	0.115	0.078	0.114	0.116	0.119	0.121	0.121
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.028	0.080	0.112	0.122	0.120	0.114	0.068	0.112	0.130	0.117	0.119	0.120
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.019	0.031	0.037	0.027	0.027	0.041	0.081	0.113	0.116	0.117	0.118	0.118
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.017	0.018	0.012	0.010	0.009	0.013	0.065	0.112	0.114	0.116	0.115	0.114
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.019	0.031	0.037	0.010	0.010	0.010	0.082	0.113	0.115	0.057	0.049	0.035
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.017	0.019	0.013	0.009	0.009	0.009	0.064	0.112	0.110	0.007	0.008	0.009
	$p = 120$											
	$\varepsilon = 0.05$						$\varepsilon = 0.10$					
m	0.5	1	2	3	4	5	0.5	1	2	3	4	5
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{MV}}$	0.049	0.081	0.109	0.113	0.114	0.115	0.096	0.117	0.117	0.120	0.121	0.121
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{MV}}$	0.044	0.099	0.121	0.123	0.114	0.113	0.088	0.132	0.120	0.117	0.119	0.118
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{M}}$	0.022	0.046	0.042	0.034	0.052	0.063	0.099	0.117	0.117	0.118	0.118	0.118
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{M}}$	0.022	0.025	0.011	0.010	0.018	0.043	0.095	0.114	0.112	0.114	0.114	0.114
$\widehat{\boldsymbol{\theta}}^{\mathrm{S}}_{\mathrm{WM}}$	0.022	0.047	0.043	0.011	0.012	0.011	0.098	0.117	0.085	0.051	0.043	0.030
$\widehat{\boldsymbol{\theta}}^{\mathrm{MCP}}_{\mathrm{WM}}$	0.023	0.026	0.011	0.011	0.011	0.011	0.096	0.114	0.070	0.010	0.012	0.013

Tabla 7.20: Promedio podado al 10 % de la proporción de clasificaciones correctas (PCC), para los escenarios de contaminación **CB1** y **CB2** con $n = 300$.

Capítulo 8

Conclusiones

El modelo de regresión logística puede ser usado con el propósito de clasificación cuando existen covariables con capacidad predictiva para cada una de las clases. Sin embargo, en la actualidad tratar con datos de alta dimensión se ha vuelto un problema recurrente que atraviesa toda la Estadística contemporánea. Una característica frecuente de este tipo de datos es que el número de covariables, digamos p, suele ser alto, mientras que el tamaño muestral, digamos n, es relativamente pequeño. Una manera popular de abordar este problema es asumiendo que el vector de los coeficientes de regresión es ralo, es decir, asumiendo que sólo unos pocos coeficientes son distintos de cero. Covariables ralas son también frecuentes en el problema de clasificación y en este contexto el problema de selección de variables puede resultar de interés. Por esta razón, en esta tesis abordamos el problema de estimación y selección de variables bajo un modelo de regresión logística, mediante la utlización de métodos robustos penalizados de modo a obtener estimaciones más fiables ante la presencia de datos atípicos.

La familia de estimadores considerada es una versión penalizada del $M-$estimador propuesto en Bianco y Yohai (1996) y de su versión pesada definida en Croux y Haesbroeck (2003). Entre otros resultados, mostramos que la familia de $M-$estimadores contiene a varios estimadores ya definidos en la literatura, como por ejemplo los estimadores de mínima divergencia definidos en Basu et al. (2017). Por otra parte, los $M-$estimadores penalizados objeto de esta tesis incluyen los estimadores mínimos cuadrados penalizados utilizados en Chi y Scott (2014). Los estimadores pesados reducen la influencia de los puntos de alta palanca y para ello, proponemos un método heurístico para definir los pesos basado en el estimador de la matriz de precisión definido en Öllerer y Croux (2015).

En esta tesis consideramos una familia amplia de funciones de penalización, que incluyen las penalidades LASSO, Ridge, SCAD y MCP. Además de estas funciones, definimos una nueva penalización llamada Signo, que tiene una motivación intuitiva y una expresión simple. Esta penalidad depende de un sólo parámetro de ajuste, mientras que las penalizaciones cóncavas SCAD y MCP poseen dos.

Como parte de nuestro aporte, se presenta un profundo estudio de las propiedades teóricas de los métodos propuestos, tanto cuando la dimensión p es fija como cuando crece con el tamaño de muestra. En particular, para el caso en que la cantidad de covariables p es fija, mostramos que los $M-$estimadores pesados penalizados son consistentes y, para algunas familias de penalidades, seleccionan variables consistentemente cuando el tamaño de muestra n tiende a infinito. Además,

obtenemos expresiones para su distribución asintótica. Se muestra que la elección de la función de penalización juega un papel fundamental en este caso. Específicamente, mostramos que al utilizar penalidades constantes a partir de un punto (como SCAD o MCP), se obtienen estimadores con la propiedad oráculo. Las hipótesis necesarias para probar estos resultados son muy poco exigentes, lo cual muestra que estos métodos pueden ser aplicados en contextos muy diversos. En este sentido, nuestros resultados cubren el vacío existente en cuanto al estudio del comportamiento asintótico de algunas de las propuestas previas dadas en el contexto de modelos de regresión logística ralos.

Cuando la cantidad de covariables p tiende a infinito junto con n, mostramos primero, bajo condiciones muy similares al caso en que p es fijo, que las probabilidades predichas de los estimadores propuestos son consistentes cuando $p/n \to 0$. Además, probamos que, bajo condiciones sobre la distribución del vector de covariables, la distancia ℓ_2 entre el vector de coeficientes estimado y el verdadero parámetro del modelo de regresión logística converge a 0. Por otra parte, mostramos que, al utilizar las penalizaciones SCAD y MCP, los $M-$estimadores penalizados asintóticamente seleccionan variables de forma consistente y obtenemos que las proyecciones unidimensionales de los estimadores de las covariables activas son asintóticamente normales. Para probar estos resultados, esencialmente se requiere que $pk/n \to 0$, donde k es el número de covariables activas del verdadero vector de coeficientes del modelo. A diferencia de trabajos previos sobre $M-$estimadores penalizados, las hipótesis requeridas para estos resultados son muy fáciles de interpretar.

Un punto importante es la implementación de los estimadores propuestos. Por ello, proponemos un algoritmo genérico para obtener una solución aproximada de los problemas de minimización que surgen para distintas elecciones de la pérdida y la penalización. El $M-$estimador pesado penalizado depende de la elección del parámetro de regularización λ y dicha elección puede ser sensible a la presencia de observaciones atípicas si se utiliza el procedimiento clásico. Por esta razón, proponemos un método robusto de convalidación cruzada y mostramos su ventaja por sobre el clásico. A través de un extenso estudio de simulación, comparamos el comportamiento de los estimadores clásicos y robustos para distintas elecciones de la función de pérdida y penalizaciones. En general, mostramos que los métodos robustos tienen rendimientos similares al clásico cuando no hay contaminación y se comportan mucho mejor que los clásicos en escenarios contaminados, mostrando mayor confiabilidad. Por otra parte, mostramos que los resultados obtenidos al utilizar penalizaciones acotadas como Signo o MCP son notablemente mejores respecto de los obtenidos al utilizar penalidades convexas como LASSO. Finalmente, los $M-$estimadores pesados basados en la función $\rho = \rho_c$ dada en (3.5) combinados con las penalizaciones MCP y Signo, resultan ser los más estables y fiables.

Finalmente, se aplican los métodos propuestos en algunos ejemplos de datos médicos y se muestra que los resultados obtenidos con funciones de pérdida y penalizaciones acotadas son mejores respecto a las que consideran procedimientos no robustos ya propuestos en la literatura. Además, se observa que la selección de variables resulta mucho más estable ante la presencia de datos atípicos al utilizar $M-$estimadores pesados penalizados.

Referencias

1. Avella Medina, M. A. (2016). Robust penalized M-estimators for generalized linear and additive models. PhD. Thesis, University of Geneva.

2. Basu, A., Gosh, A., Mandal, A., Martin, N. & Pardo, L. (2017). A Wald–type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electronic Journal of Statistics*, **11**, 2741-2772.

3. Basu, A., Harris, I. R., Hjort, N. L. & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85(3)**, 549-559.

4. Bianco, A. (1990) $M-$estimadores robustos para regresión binominal y multinominal. Tesis doctoral. Disponible en `https://digital.bl.fcen.uba.ar/download/tesis/tesis_n2316_Bianco.pdf`

5. Bianco, A. & Boente, G. (2002). On the asymptotic behavior of one-step estimates in heteroscedastic regression models. *Statistics and Probabilty Letters*, **60**, 33-47.

6. Bianco, A. & Martínez, E. (2009). Robust testing in the logistic regression model. *Computational Statistics & Data Analysis* **53**, 4095-4105.

7. Bianco, A. & Yohai, V. (1996). Robust estimation in the logistic regression model. *Lecture Notes in Statistics*, **109**, 17-34. Springer-Verlag, New York.

8. Bühlmann, P. & van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

9. Cantoni, E. & Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**, 1022-1030.

10. Chi, E. C. & Scott, D. W. (2014). Robust parametric classification and variable selection by a minimum distance criterion. *Journal of Computational and Graphical Statistics*, **23**, 111-128.

11. Clarke, F. H. (1975). Generalized gradients and applications. *Transactions of the American Mathematical Society*, **205**, 247-262.

12. Croux, C. & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational statistics & Data Analysis*, **44**, 273-295.

13. Dasarathy, G. (2011). A simple probability trick for bounding the expected maximum of n random variables. `http://www.public.asu.edu/~gdasarat/files/maxGaussians.pdf`

14. Efron, B. & Hastie, T. (2016). *Computer age statistical inference*, (Vol. 5). Cambridge University Press.

15. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, **32**, 407-499.

16. Elsener, A. & van de Geer, S. (2018). Sharp oracle inequalities for stationary points of nonconvex penalized M-estimators. *IEEE Transactions on Information Theory*, **65**, 1452-1472.

17. Fan, J., & Li, R. (2001). Variable selection via non–concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

18. Fan, J., & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, **57**, 5467-5484.

19. Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics,* **32**, 928-961.

20. Feng, J., Xu, H., Mannor, S. & Yan, S. (2014). Robust logistic regression and classification. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. and Weinberger, K. Q. (eds.) *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 253-261.

21. Fernandes, K., Cardoso, J. S. & Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. In: Alexandre L., Salvador Sánchez J., Rodrigues J. (eds) *Pattern Recognition and Image Analysis. IbPRIA 2017. Lecture Notes in Computer Science*, vol. 10255. Springer.

22. Frank, L. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-135.

23. Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

24. Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1.

25. Gourieroux, C. & Monfort, A. (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics*, **17**, 83-97.

26. Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

27. Huang, J. & Xie, H. (2007). Asymptotic oracle properties of SCAD–penalized least squares estimators. *IMS Lecture Notes, Monograph Series. Asymptotics: Particles, Processes and Inverse Problems*, **55**, 149-166.

28. Jiang, D., & Huang, J. (2014). Majorization minimization by coordinate descent for concave penalized generalized linear models. *Statistics and Computing*, **24**, 871-883.

29. Kim, Y., Choi, H., & Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, **103**, 1665-1673.

30. Kim, Y., & Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, **99**, 315-325.

31. Kim, J., & Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, **18**, 191-219.

32. Knight, K. & Fu, W. (2000). Asymptotics for LASSO-type estimators. *Annals of Statistics*, **28**, 1356-1378.

33. Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer–Verlag, New York.

34. Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. *Artificial Intelligence in Medicine*, **23**, 149-169.

35. Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, **172**, 211-222.

36. Li, G., Peng, H., & Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica*, **21**, 391-419.

37. Loh, P. L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics*, **45**, 866-896.

38. Loh, P. L., & Tan, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, **12**, 1429-1467.

39. Loh, P. L., & Wainwright, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Annals of Statistics*, **45**, 2455-2482.

40. Maronna, R., Martin, R., Yohai, V. & Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. Wiley, New York.

41. Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, **52**, 374-393.

42. Öllerer, V., & Croux, C. (2015). Robust high-dimensional precision matrix estimation. In: *Modern Nonparametric, Robust and Multivariate Methods* (pp. 325-350). Springer, Cham.

43. Park, H., & Konishi, S. (2016). Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation*, **86**, 1450-1461.

44. Pollard, D. (1984). *Convergence of stochastic processes*, Springer.

45. Pollard, D. (1989). Asymptotics via Empirical Processes. *Statistical Science*, **4**, 341-354.

46. Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, **38**, 485-498.

47. Smucler, E. (2016). *Estimadores robustos para el modelo de regresión lineal con datos de alta dimensión*. Tesis doctoral, Universidad de Buenos Aires. Disponible en `http://cms.dm.uba.ar/academico/carreras/doctorado/Tesis%20Smucler.pdf`

48. Smucler, E., & Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, **111**, 116-130.

49. Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, July). Nuclear feature extraction for breast tumor diagnosis. In: *Biomedical Image Processing and Biomedical Visualization, International Society for Optics and Photonics*, **Vol. 1905**, pp. 861-871.

50. Tarr, G., Müller, S., & Weber, N. C. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis*, **93**, 404-420.

51. Tibshirani, J., & Manning, C. D. (2013). Robust Logistic Regression using Shift Parameters. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 124-129.

52. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**, 267-288.

53. Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., & Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso–type problems. *Journal of the Royal Statistical Society, Series B (Methodological)*, **74**, 245-266.

54. Tibshirani, R., Wainwright, M., & Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

55. van de Geer, S. (2000). *Empirical processes in $M-$estimation*. Cambridge Series in Statistical and Probabilistic Mathematics.

56. van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics*, **36**, 614-645.

57. van de Geer, S., & Müller, P. (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Statistical science*, **27**, 469-480.

58. Van Der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer, New York, NY.

59. Wang, F. & Scott, D. (1994). The L_1 method for robust nonparametric regression. *Journal of the American Statistical Association*, **89**, 65-76.

60. Xie, H. & Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Annals of Statistics*, **37**, 673-696.

61. Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38**, 894-942.

62. Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

63. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301-320.