



UNIVERSIDAD  
DE LA REPUBLICA  
URUGUAY

# Selección de variables para datos espaciales

Romina Gonella Furtado

Programa de Posgrado en Ingeniería Matemática  
Facultad de Ingeniería  
Universidad de la República

Montevideo – Uruguay  
Octubre de 2021



UNIVERSIDAD  
DE LA REPUBLICA  
URUGUAY

# Selección de variables para datos espaciales

Romina Gonella Furtado

Tesis de Maestría presentada al Programa de Posgrado en Ingeniería Matemática, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magíster en Ingeniería Matemática.

Director:

D.Sc. Prof. Liliane Bel

Codirector:

D.Sc. Prof. Mathias Bourel

Director académico:

D.Sc. Prof. Mathias Bourel

Montevideo – Uruguay

Octubre de 2021

Gonella Furtado, Romina

Selección de variables para datos espaciales / Romina  
Gonella Furtado. - Montevideo: Universidad de la  
República, Facultad de Ingeniería, 2021.

XX, 122 p.: il.; 29, 7cm.

Director:

Liliane Bel

Codirector:

Mathias Bourel

Director académico:

Mathias Bourel

Tesis de Maestría – Universidad de la República,  
Programa en Ingeniería Matemática, 2021.

Referencias bibliográficas: p. 83 – 88.

1. estadística espacial, 2. selección de variables,  
3. correlación, 4. LASSO. I. Bel, Liliane, *et al.*  
II. Universidad de la República, Programa de Posgrado  
en Ingeniería Matemática. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

---

D.Sc. Prof. Tomás Goicoa

---

D.Sc. Prof. José Rafael León

---

D.Sc. Prof. Marco Scavino

Montevideo – Uruguay  
Octubre de 2021

A mi abuelo, que le hubiese  
gustado ver este trabajo  
terminado.

# Agradecimientos

En primer lugar agradezco a mis codirectores de tesis Liliane Bel y Mathias Bourel por aceptar dirigir este trabajo, aportar ideas, brindarme su tiempo y tener mucha paciencia durante todo el proceso. A Mathias también por su rol de director académico y por haberme invitado a realizar esta Maestría.

En el marco de esta tesis viajé a París para trabajar de forma más cercana con Liliane, a quien le agradezco en particular por su tiempo y las gestiones realizadas durante esos días. También agradezco a la Embajada de Francia en Uruguay por proporcionarme una ayuda económica, a AgroParisTech por recibirme, y a la Comisión Nacional Honoraria de MEVIR por otorgarme días de licencia especial por estudio.

Las simulaciones presentadas en este documento se realizaron utilizando el Centro Nacional de Supercomputación ClusterUY<sup>1</sup>, por lo que también extiendo a este mi agradecimiento.

Los datos reales utilizados fueron proporcionados por el Instituto Nacional de Estadística a través de una solicitud de información específica, agradezco a dicha institución y muy especialmente a Laura Nalbarte, quien fuera su Directora Técnica en aquel momento. Laura fue mi docente, dirigió mi tesis de grado y había aceptado la invitación a integrar el tribunal de defensa de este trabajo. Su inesperada y prematura partida no lo permitió, por lo tanto plasmo en estas líneas mi gran agradecimiento a ella.

Finalmente, quiero agradecer a familiares, amigos y compañeros de trabajo, por su valioso apoyo y contención en todo momento.

---

<sup>1</sup>Ver Nesmachnow and Iturriaga [40], <https://cluster.uy>

## RESUMEN

En este trabajo se estudia la selección de variables en modelos espaciales en red, en particular en modelos de regresión espacial, con ubicaciones irregulares y donde la estructura de autocorrelación se modela en los errores aleatorios. Primero se estudia la selección de variables para datos dependientes y luego como caso particular para datos espaciales. Se presenta una estrategia para “eliminar” la dependencia, que consiste en estimar la matriz de covarianzas de los errores, luego transformar el problema en uno equivalente donde los errores ya no presentan autocorrelación y finalmente realizar la selección de variables utilizando un modelo LASSO clásico. Se adapta un teorema que establece las condiciones que deben cumplir tanto la matriz de covarianzas estimada como la matriz de diseño del modelo. Se demuestra que las condiciones de ese teorema se cumplen para un modelo de regresión espacial con errores de tipo CAR o SAR, estructura de vecindad triangular y pesos específicos.

También se compara esta estrategia con otra desarrollada en Zhu et al. [57], denominada  $LARS_m$ . Se comparan ambas estrategias tanto en datos simulados como reales. Se obtiene que el modelo estimado luego de eliminar la dependencia espacial selecciona mejor que el modelo aplicado al problema original. Lo mismo ocurre con el modelo  $LARS_m$ . Al comparar nuestra metodología con el  $LARS_m$  en las simulaciones, se obtiene que en general el primero selecciona mejor las variables que participan del modelo, mientras que el segundo presenta menor sesgo en la estimación de los parámetros asociados a las variables que participan del modelo verdadero.

Palabras claves:

estadística espacial, selección de variables, correlación, LASSO.

## ABSTRACT

In this work we study the variable selection in spatial network models, particularly in spatial regression models, with irregular locations and where the autocorrelation structure is modeled in random errors. First, the variable selection for dependent data is studied and then as a particular case for spatial data. A strategy is presented to “eliminate” the dependence, which consists of estimating the covariance matrix of the errors, then transforming the problem into an equivalent one where the errors no longer present autocorrelation and finally making the variable selection using a classic LASSO model. A theorem is adapted that establishes the conditions that must meet both the estimated covariance matrix and the model design matrix. It is shown that the conditions of this theorem are fulfilled for a spatial regression model with errors of type CAR or SAR, triangular neighborhood structure and specific weights.

This strategy is also compared with another developed in Zhu et al. [57], called  $LARS_m$ . Both strategies are compared in both simulated and real data. It is obtained that the estimated model after eliminating the spatial dependence selects better than the model applied to the original problem. The same goes for the  $LARS_m$  model. When comparing our methodology with the  $LARS_m$  in the simulations, it is obtained that in general the former better selects the variables that participate in the model, while the latter presents less bias in the parameter estimation associated with the variables that participate in the true model.

Keywords:

spatial statistic, model selection, correlation, LASSO.



# Lista de figuras

2.1	Geometría del método LASSO . . . . .	11
3.1	Ejemplo de vecinos de primer y segundo orden . . . . .	38
3.2	Máximo del promedio de vecinos que tiene un vecino de $i$ , con $i \in D = \{1, \dots, n\}$ , según el valor de $n$ . . . . .	40
4.1	$\hat{\beta}_i$ $0 \leq i \leq 20$ para el <i>Problema 1</i> con error CAR. . . . .	57
4.2	$\hat{\beta}_i$ $0 \leq i \leq 20$ para el <i>Problema 1</i> con error SAR. . . . .	58
4.3	$\hat{\beta}_i$ $0 \leq i \leq 20$ para el <i>Problema 2</i> con error CAR. . . . .	59
4.4	$\hat{\beta}_i$ $0 \leq i \leq 20$ para el <i>Problema 2</i> con error SAR. . . . .	60
5.1	Ubicación espacial de los hogares del medio rural ampliado de la ECH considerados, identificando el cuartil de ingreso per cápita al que pertenecen. . . . .	69
5.2	Grafo asociado a la estructura de vecindad triangular para la base ECH considerada. . . . .	69
1.1	Función valor absoluto . . . . .	118

# Lista de tablas

3.1	Condiciones sobre los pesos $w_{ij}$ de la matriz $\mathbf{W}_{\mathcal{R}}$ . . . . .	34
4.1	Resultados de LASSO inicial, LASSO ajustado y LARS $_m$ para el <i>Problema 1</i> con errores CAR. . . . .	53
4.2	Resultados de LASSO inicial, LASSO ajustado y LARS $_m$ para el <i>Problema 1</i> con errores SAR. . . . .	54
4.3	Resultados de LASSO inicial, LASSO ajustado y LARS $_m$ para el <i>Problema 2</i> con errores CAR. . . . .	55
4.4	Resultados de LASSO inicial, LASSO ajustado y LARS $_m$ para el <i>Problema 2</i> con errores SAR. . . . .	56
4.5	Parámetros espaciales estimados en LASSO ajustado y LARS $_m$ para el <i>Problema 1</i> con errores CAR. . . . .	61
4.6	Parámetros espaciales estimados en LASSO ajustado y LARS $_m$ para el <i>Problema 1</i> con errores SAR. . . . .	62
4.7	Parámetros espaciales estimados en LASSO ajustado y LARS $_m$ para el <i>Problema 2</i> con errores CAR. . . . .	63
4.8	Parámetros espaciales estimados en LASSO ajustado y LARS $_m$ para el <i>Problema 2</i> con errores SAR. . . . .	64
4.9	Cantidad de pasos necesarios ( $m$ ) para obtener la convergencia en LARS $_m$ por problema y tipo de error. . . . .	66
5.1	Pruebas de autocorrelación espacial para la variable a explicar: <i>ingreso per cápita del hogar</i> . . . . .	70
5.2	Correlaciones más altas entre variables explicativas. . . . .	72
5.3	Pruebas de autocorrelación espacial para los residuos, modelo CAR. . . . .	73
5.4	Pruebas de autocorrelación espacial para los residuos, modelo SAR. . . . .	73

5.5	Parámetros estimados para el LASSO inicial, LASSO ajustado y LARS <sub>m</sub> , con errores CAR y SAR. . . . .	74
5.6	Distribución de los valores estimados para cada parámetro de LARS <sub>m</sub> en 100 iteraciones con errores CAR. . . . .	77
5.7	Distribución de los valores estimados para cada parámetro de LARS <sub>m</sub> en 100 iteraciones con errores SAR. . . . .	78
5.8	Indicadores de bondad de ajuste de los modelos estimados. . . . .	79
2.1	Matriz de correlaciones entre las variables explicativas en la aplicación a datos reales (Capítulo 5). Parte 1. . . . .	114
2.2	Matriz de correlaciones entre las variables explicativas en la aplicación a datos reales (Capítulo 5). Parte 2. . . . .	115
1.1	Distintas posibilidades para el subdiferencial, de acuerdo al signo de $x$ e $y$ . . . . .	119

# Lista de símbolos

Lista de los símbolos más relevantes de la tesis.

- $\mathbb{R}$  Conjunto de los números reales. 10
- $\mathbf{I}$  Matriz identidad. 5
- $\mathbf{1}$  Vector donde todos sus elementos son 1. 16
- $\mathbf{0}$  Vector nulo, es decir, donde todos sus elementos son 0. 15
- $Y$  Vector aleatorio, variable a explicar del modelo lineal. 4
- $\mathbf{X}$  Matriz de diseño del modelo lineal. 4
- $\beta$  Parámetro desconocido del modelo lineal. 4
- $\varepsilon$  Vector aleatorio que representa el error del modelo lineal. 4
- $n$  Cantidad de observaciones del modelo, se corresponde con la dimensión del vector  $Y$  y la cantidad de filas de la matriz  $\mathbf{X}$ . 4
- $p$  Cantidad total de variables consideradas en el modelo. Se corresponde con la dimensión de la matriz  $\mathbf{X}$ . 4
- $p^*$  Cantidad de variables que participan en el modelo, es decir, donde  $\beta^* \neq 0$ . 5
- $\sigma^2$  Parámetro que participa en la matriz de covarianzas del vector  $\varepsilon$ . 5
- $A^*$  Conjunto de índices de las variables que participan del modelo verdadero. 5
- $A^{*c}$  Conjunto de índices de las variables que no participan del modelo verdadero, es decir, complemento de  $A^*$ . 5
- $\mathbf{X}_{A^*}$  Submatriz de  $\mathbf{X}$  que incluye solamente las variables que participan del modelo. 15
- $\mathbf{X}_{A^{*c}}$  Submatriz de  $\mathbf{X}$  que incluye solamente las variables que no participan del modelo. 15
- $M$  Conjunto de todas las combinaciones de las  $p$  variables, variando de 1 a  $p$  elementos. 6

- $m_M$  Una combinación específica de las  $p$  variables, puede tener entre 1 y  $p$  elementos. 6
- $\hat{f}_{m_M}$  Denota el modelo que se construye a partir del conjunto  $m_M$ . 6
- $R(\hat{f}_{m_M})$  Riesgo  $l_2$  del estimador  $\hat{f}_{m_M}$ :  $R(\hat{f}_{m_M}) = E(\|\hat{f}_{m_M} - f\|_2^2)$ . 6
- $m_0$  Elemento perteneciente al conjunto  $M$  que representa la combinación de variables que minimiza el riesgo  $R(\hat{f}_{m_M})$ . 6
- $f$  El verdadero modelo LASSO, que es desconocido. 4
- $\hat{f}_{m_0}$  Denota el modelo que se construye a partir del conjunto  $m_0$ . 6
- $p_{m_M}$  Cantidad de parámetros del modelo  $f_{m_M}$ . 7
- $\mathcal{L}(\beta)$  Modelo LASSO. 9
- $\hat{\beta}$  Estimador del parámetro  $\beta$  por el método LASSO. 10
- $\lambda$  Parámetro de regularidad del modelo LASSO. 9
- $\mathcal{L}_1(\beta)$  Función que forma parte del modelo LASSO:  $\mathcal{L}_1(\beta) = \|Y - \mathbf{X}\beta\|_2^2$ . 9
- $\mathcal{L}_2(\beta)$  Función que forma parte del modelo LASSO:  $\mathcal{L}_2(\beta) = \lambda\|\beta\|_1$ . 9
- $\mathbf{C}^n$  Matriz definida como:  $\mathbf{C}^n = \frac{1}{n}\mathbf{X}'\mathbf{X}$ . 15
- $\mathbf{C}_{11}^n$  Matriz definida como:  $\mathbf{C}_{11}^n = \frac{1}{n}\mathbf{X}'_{A^*}\mathbf{X}_{A^*}$ . 15
- $\mathbf{C}_{12}^n$  Matriz definida como:  $\mathbf{C}_{12}^n = \frac{1}{n}\mathbf{X}'_{A^*}\mathbf{X}_{A^{*c}}$ . 15
- $\mathbf{C}_{21}^n$  Matriz definida como:  $\mathbf{C}_{21}^n = \frac{1}{n}\mathbf{X}'_{A^{*c}}\mathbf{X}_{A^*}$ . 15
- $\mathbf{C}_{22}^n$  Matriz definida como:  $\mathbf{C}_{22}^n = \frac{1}{n}\mathbf{X}'_{A^{*c}}\mathbf{X}_{A^{*c}}$ . 15
- $W_{A^*}$  Vector definido como:  $W_{A^*} = \frac{1}{\sqrt{n}}\mathbf{X}'_{A^*}\varepsilon$ . 17
- $W_{A^{*c}}$  Vector definido como:  $W_{A^{*c}} = \frac{1}{\sqrt{n}}\mathbf{X}'_{A^{*c}}\varepsilon$ . 17
- $\beta_{A^*}$  Subvector de  $\beta$  que contiene solamente los elementos de  $A^*$ . 15
- $\beta_{A^{*c}}$  Subvector de  $\beta$  que contiene solamente los elementos de  $A^{*c}$ . 15
- $\hat{\beta}_{A^*}$  Subvector de  $\hat{\beta}$  que contiene solamente los elementos de  $A^*$ . 47
- $\hat{\beta}_{A^{*c}}$  Subvector de  $\hat{\beta}$  que contiene solamente los elementos de  $A^{*c}$ . 47
- $S$  Subespacio lineal. 5
- $D$  Conjunto de ubicaciones espaciales. 23
- $d$  Dimensión del conjunto espacial  $D$ . 23
- $\Sigma$  Matriz de covarianzas del vector  $\varepsilon$ . 19
- $\hat{\Sigma}$  Matriz de covarianzas estimada del vector  $\varepsilon$ . 20
- $\mathcal{R}$  Grafo de influencia entre observaciones. 24
- $\mathbf{W}_{\mathcal{R}}$  Matriz de pesos asociada al grafo  $\mathcal{R}$ . 25
- $\theta$  Parámetro asociado a la matriz  $\mathbf{W}_{\mathcal{R}}$ . Puede ser un escalar o un vector, dependiendo del orden de vecindad. 28

- $\hat{\theta}$  Estimador del parámetro  $\theta$ . 40
- $\mathbf{C}_{\mathcal{R}}$  Matriz definida como:  $\mathbf{C}_{\mathcal{R}} = \sum_{k=1}^K \theta_k \mathbf{W}_{\mathcal{R}^k}$ . 27
- $\hat{\sigma}^2$  Estimador del parámetro  $\sigma^2$ . 40
- $\Sigma_{\text{CAR}}$  Matriz de covarianzas del modelo CAR. 27
- $\Sigma_{\text{SAR}}$  Matriz de covarianzas del modelo SAR. 28
- $\hat{\Sigma}_{\text{CAR}}$  Matriz de covarianzas estimada del modelo CAR. 33
- $\hat{\Sigma}_{\text{SAR}}$  Matriz de covarianzas estimada del modelo SAR. 33
- $I^M$  Índice de Moran. 25
- $I^G$  Índice de Geary. 26
- $\dot{Y}$  Vector  $Y$  transformado, definido como:  $\dot{Y} = \Sigma^{-1/2}Y$ . 20
- $\dot{\mathbf{X}}$  Matriz  $\mathbf{X}$  transformada, definida como:  $\dot{\mathbf{X}} = \Sigma^{-1/2}\mathbf{X}$ . 20
- $\dot{\varepsilon}$  Vector  $\varepsilon$  transformado, definido como:  $\dot{\varepsilon} = \Sigma^{-1/2}\varepsilon$ . 20
- $\hat{\beta}$  Estimador de  $\beta$  en el modelo lineal transformado utilizando la matriz  $\Sigma$ . 20
- $\dot{\mathbf{X}}_{A^*}$  Transformación de la matriz  $\mathbf{X}_{A^*}$  para eliminar la dependencia espacial utilizando la matriz  $\Sigma$ . 20
- $\dot{\mathbf{X}}_{A^{*c}}$  Transformación de la matriz  $\mathbf{X}_{A^{*c}}$  para eliminar la dependencia espacial utilizando la matriz  $\Sigma$ . 20
- $\dot{\mathbf{C}}^n$  Matriz definida como:  $\dot{\mathbf{C}}^n = \frac{1}{n}\dot{\mathbf{X}}'\dot{\mathbf{X}}$ . 20
- $\dot{\mathbf{C}}_{11}^n$  Matriz definida como:  $\dot{\mathbf{C}}_{11}^n = \frac{1}{n}\dot{\mathbf{X}}'_{A^*}\dot{\mathbf{X}}_{A^*}$ . 20
- $\dot{\mathbf{C}}_{12}^n$  Matriz definida como:  $\dot{\mathbf{C}}_{12}^n = \frac{1}{n}\dot{\mathbf{X}}'_{A^*}\dot{\mathbf{X}}_{A^{*c}}$ . 20
- $\dot{\mathbf{C}}_{21}^n$  Matriz definida como:  $\dot{\mathbf{C}}_{21}^n = \frac{1}{n}\dot{\mathbf{X}}'_{A^{*c}}\dot{\mathbf{X}}_{A^*}$ . 20
- $\dot{\mathbf{C}}_{22}^n$  Matriz definida como:  $\dot{\mathbf{C}}_{22}^n = \frac{1}{n}\dot{\mathbf{X}}'_{A^{*c}}\dot{\mathbf{X}}_{A^{*c}}$ . 20
- $\dot{W}_{A^*}$  Vector definido como:  $\dot{W}_{A^*} = \frac{1}{\sqrt{n}}\dot{\mathbf{X}}'_{A^*}\varepsilon$ . 97
- $\dot{W}_{A^{*c}}$  Vector definido como:  $\dot{W}_{A^{*c}} = \frac{1}{\sqrt{n}}\dot{\mathbf{X}}'_{A^{*c}}\varepsilon$ . 106
- $\tilde{Y}$  Vector  $Y$  transformado, definido como:  $\tilde{Y} = \hat{\Sigma}^{-1/2}Y$ . 20
- $\tilde{\mathbf{X}}$  Matriz  $\mathbf{X}$  transformada, definida como:  $\tilde{\mathbf{X}} = \hat{\Sigma}^{-1/2}\mathbf{X}$ . 20
- $\tilde{\varepsilon}$  Vector  $\varepsilon$  transformado, definido como:  $\tilde{\varepsilon} = \hat{\Sigma}^{-1/2}\varepsilon$ . 20
- $\tilde{\beta}$  Estimador de  $\beta$  en el modelo lineal transformado utilizando la matriz  $\hat{\Sigma}$ . 20
- $\tilde{\mathbf{X}}_{A^*}$  Submatriz de  $\tilde{\mathbf{X}}$  que incluye solamente las variables que participan del modelo. 20
- $\tilde{\mathbf{X}}_{A^{*c}}$  Submatriz de la matriz  $\tilde{\mathbf{X}}$  que incluye solamente las variables que no participan del modelo. 20
- $\tilde{\mathbf{C}}^n$  Matriz definida como:  $\tilde{\mathbf{C}}^n = \frac{1}{n}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ . 20
- $\tilde{\mathbf{C}}_{11}^n$  Matriz definida como:  $\tilde{\mathbf{C}}_{11}^n = \frac{1}{n}\tilde{\mathbf{X}}'_{A^*}\tilde{\mathbf{X}}_{A^*}$ . 20

$\tilde{\mathbf{C}}_{12}^n$  Matriz definida como:  $\tilde{\mathbf{C}}_{12}^n = \frac{1}{n} \tilde{\mathbf{X}}'_{A^*} \tilde{\mathbf{X}}_{A^{*c}}$ . 20

$\tilde{\mathbf{C}}_{21}^n$  Matriz definida como:  $\tilde{\mathbf{C}}_{21}^n = \frac{1}{n} \tilde{\mathbf{X}}'_{A^{*c}} \tilde{\mathbf{X}}_{A^*}$ . 20

$\tilde{\mathbf{C}}_{22}^n$  Matriz definida como:  $\tilde{\mathbf{C}}_{22}^n = \frac{1}{n} \tilde{\mathbf{X}}'_{A^{*c}} \tilde{\mathbf{X}}_{A^{*c}}$ . 20

$\tilde{W}_{A^*}$  Vector definido como:  $\tilde{W}_{A^*} = \frac{1}{\sqrt{n}} \tilde{\mathbf{X}}'_{A^*} \varepsilon$ . 21

$\tilde{W}_{A^{*c}}$  Vector definido como:  $\tilde{W}_{A^{*c}} = \frac{1}{\sqrt{n}} \tilde{\mathbf{X}}'_{A^{*c}} \varepsilon$ . 22

$\eta$  Vector de parámetros tanto de la regresión como de la matriz de covarianzas

$\Sigma$ , definido como:  $\eta = (\beta, \theta, \sigma^2)$ . 43

$\hat{\eta}$  Estimador del parámetro  $\eta$ . 43

# Lista de notaciones

Solamente se especifican las notaciones más importantes

$X_{.j}$   $j$ -ésima columna de la matriz  $\mathbf{X}$ . 5

$X_i$   $i$ -ésima fila de la matriz  $\mathbf{X}$ . 5

$|A| = \text{card}(A)$  Numero de elementos de  $A$  (si  $A$  es un conjunto). Valor absoluto de  $A$  (si  $A$  es un numero real). 5

$\|X\|_2$  Norma  $l_2$  de  $X$ , donde  $X$  es un vector. Se define como:  $\|X\|_2 = \sqrt{\sum_{i=1}^p X_i^2}$ . 6

$\|X\|_1$  Norma  $l_1$  de  $X$ , donde  $X$  es un vector. Se define como:  $\|X\|_1 = \sum_{i=1}^p |X_i|$ . 10

$\|X\|_\infty$  Norma  $l_\infty$  de  $X$ , donde  $X$  es un vector. Se define como:  $\|X\|_\infty = \max(|X_1|, |X_2|, \dots, |X_p|)$ . 14

$\|\mathbf{X}\|_2$  Norma espectral de  $\mathbf{X}$ , donde  $\mathbf{X}$  es una matriz cuadrada de  $n \times n$ . Se define como:  $\|\mathbf{X}\|_2 = \max_{1 \leq i \leq n} (\sqrt{\rho_i(\mathbf{X}'\mathbf{X})})$ , donde  $\rho_i(\mathbf{X})$  representa el  $i$ -ésimo valor propio de  $\mathbf{X}$ . 94

$\|\mathbf{X}\|_1$  Norma de la máxima suma de columnas de  $X$ , donde  $X$  es una matriz cuadrada de  $n \times n$ . Se define como:  $\|\mathbf{X}\|_1 = \max_{1 \leq j \leq n} \left( \sum_{i=1}^n |X_{ij}| \right)$ . 16

$\|\mathbf{X}\|_\infty$  Norma de la máxima suma de filas de  $\mathbf{X}$ , donde  $\mathbf{X}$  es una matriz cuadrada de  $n \times n$ . Se define como:  $\|\mathbf{X}\|_\infty = \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |X_{ij}| \right)$ . 30

$\rho(\mathbf{X})$  Vector de valores propios de la matriz  $\mathbf{X}$ . En particular  $\rho_{\min}(\mathbf{X})$  y  $\rho_{\max}(\mathbf{X})$  representan el menor y mayor valor propio de  $\mathbf{X}$  respectivamente. 22

$\gamma(\mathbf{X})$  Radio espectral de la matriz  $\mathbf{X}$ , definido como:  $\gamma(\mathbf{X}) = \max_{1 \leq i \leq n} (|\sqrt{\rho_i(\mathbf{X})}|)$ . 30

$\partial f(x)$  Subdiferencial de la funcion  $f$  en  $x$ . Se define como:  $\partial f(x) = \{w \in \mathbb{R}^n : f(y) \geq f(x) + \langle w, y - x \rangle \forall y \in \mathbb{R}^n\}$ . 11



$B_{L_1}(r)$  Esfera  $L_1$ :  $B_{L_1}(r) = \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq r\}$ . 10

$sign(x)$  Función signo. Se define como:  $sign(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x = 0 \\ -1 & \text{si } x < 0 \end{cases}$ . Con  $x \in$

$\mathbb{R}$ . 11

$a =_s b$  Igual en signo. Significa que  $P(sign(a) = sign(b)) = 1$ . 22

$Proj_S$  Proyección ortogonal en  $S$ . 5

$\mathcal{N}(i)$  Conjunto de índices de los vecinos de  $i$ , de acuerdo a una estructura de vecindad dada. 27

$\mathcal{N}_2(i)$  Conjunto de índices de los vecinos de segundo orden de  $i$ , de acuerdo a una estructura de vecindad dada. 37

# Lista de siglas

- AIC** Akaike Information Criterion 1  
**BIC** Bayesian Information Criterion 1  
**CAR** Condicional Autorregresivo 2  
**CART** Classification and Regression Trees 1  
**ECH** Encuesta Continua de Hogares 3  
**INE** Instituto Nacional de Estadística 67  
**LARS** Least-angle Regression 1  
**LASSO** Least Absolute Shrinkage and Selection Operator 1  
**SAR** Simultáneo Autorregresivo 2

# Tabla de contenidos

Lista de figuras	IX
Lista de tablas	X
Lista de símbolos	XV
Notaciones	XVII
Lista de siglas	XVIII
<b>1 Introducción</b>	<b>1</b>
<b>2 Selección de variables para datos independientes</b>	<b>4</b>
2.1 Selección de variables en Modelos Lineales . . . . .	4
2.2 Método LASSO y selección de variables . . . . .	9
2.3 Consistencia en signo del estimador LASSO . . . . .	14
<b>3 Selección de variables para datos dependientes: caso de la de- pendencia espacial</b>	<b>19</b>
3.1 Selección de variables en un modelo lineal dependiente . . . . .	19
3.1.1 Eliminación de la dependencia en los errores . . . . .	20
3.2 Contexto espacial . . . . .	23
3.2.1 Autocorrelación espacial . . . . .	24
3.2.2 Modelos para procesos espaciales . . . . .	27
3.2.3 Modelos con errores espaciales . . . . .	29
3.2.4 Procedimiento para eliminar la dependencia espacial . .	41
3.3 Otra metodología para selección de variables en el caso espacial:	
LARS <sub>m</sub> . . . . .	42
3.3.1 Estimación máximo verosímil penalizada via LARS <sub>m</sub> . .	43

3.3.2	Procedimiento para aplicar $LARS_m$ . . . . .	48
<b>4</b>	<b>Simulaciones</b>	<b>50</b>
<b>5</b>	<b>Aplicación con datos reales de la Encuesta Continua de Hogares</b>	<b>67</b>
<b>6</b>	<b>Conclusiones</b>	<b>80</b>
	<b>Referencias bibliográficas</b>	<b>83</b>
	<b>Apéndices</b>	<b>89</b>
Apéndice 1	Demostraciones . . . . .	90
Apéndice 2	Correlación de $\mathbf{X}$ en la aplicación real . . . . .	114
<b>Anexos</b>		<b>116</b>
Anexo 1	Definiciones y resultados preliminares . . . . .	117
1.0.1	Subdiferenciales y subgradientes . . . . .	117
1.0.2	Subdiferencial de la función valor absoluto . . . . .	118
1.0.3	Subdiferencial de la norma $L_1$ . . . . .	119
Anexo 2	Resultados de Zhu et al utilizados . . . . .	121

# Capítulo 1

## Introducción

Este trabajo se enmarca en modelos para datos espaciales con (potencialmente) un gran número de variables explicativas. Este tipo de datos se encuentran en diversos contextos, como por ejemplo el medio ambiente, el clima o en agronomía, y tienen la particularidad de presentar correlación vinculada a la proximidad de las observaciones. Los métodos de modelización para datos espaciales deben tener en cuenta esta correlación.

Con el desarrollo de las tecnologías se puede considerar un gran número de variables explicativas y seleccionar las más relevantes. Entre los métodos de selección clásicos se encuentran el Akaike Information Criterion (AIC), el Bayesian Information Criterion (BIC) y el Stepwise (ver Akaike [1], Schwarz [46] y Efron [16]). En los últimos años han surgido varios métodos de selección de variables adaptados al tipo de modelo que se usa. En el caso de los modelos paramétricos, como los modelos lineales, la selección de variables se puede efectuar mediante métodos de regularización sobre los coeficientes, por ejemplo Least Absolute Shrinkage and Selection Operator (LASSO), Least-angle Regression (LARS) o ElasticNet (ver Tibshirani [50], Efron et al. [15] y Zou and Hastie [60]). En el contexto de los modelos no paramétricos, como Classification and Regression Trees (CART) o Random Forest (Breiman et al. [10] y Breiman [9]), existen criterios sobre la importancia de las variables que se basan en criterios de homogeneidad en los nodos del árbol o de variabilidad sobre el error de clasificación del estimador. La implementación y la eficiencia de la mayoría de estos métodos se justifica en que los datos considerados son independientes, raras veces se usan para datos correlacionados y aún menos para datos espaciales.

Este trabajo se enfoca en la selección de variables en contextos espaciales con ubicaciones irregulares. En primer lugar, en el Capítulo 2 se estudian los métodos de selección utilizados en modelos lineales para datos independientes: se presentan los métodos más utilizados, con énfasis en LASSO, a su vez se introduce el concepto de consistencia en signo, condiciones de irrepresentabilidad (fuerte y débil) y se presenta un teorema que enuncia las condiciones necesarias para establecer la consistencia fuerte en signo del estimador LASSO.

En el Capítulo 3 se estudia la selección de variables para datos dependientes (como caso particular los espaciales). En primer lugar se presenta una estrategia para “eliminar” la dependencia, luego se adapta un teorema del artículo de Perrot-Dockès et al. [42] para un modelo de regresión lineal con respuesta multivariada y correlación entre variables, a nuestro problema de interés donde la dependencia se da entre observaciones. Se establecen las condiciones que deben cumplir tanto la matriz de covarianzas estimada como la matriz de diseño del modelo (que contiene la información de las variables consideradas). A continuación se introduce el contexto espacial (conceptos de datos espaciales y sus tipos: geoestadística y datos en red, autocorrelación espacial, vecindad, entre otros). Se considera un modelo de regresión lineal donde la estructura espacial se modela en el vector de errores aleatorios, asumiendo que los errores son normales centrados y con una matriz de covarianzas determinada. Se demuestra que las condiciones del teorema para errores dependientes se cumplen en este contexto, para errores de tipo Condicional Autorregresivo (CAR) o Simultáneo Autorregresivo (SAR) (ver Gaetan and Guyon [19]), con una estructura de vecindad triangular, pesos específicos y orden de vecindad igual a uno. En esta estrategia primero se estima la matriz de covarianzas, luego se transforma el problema en uno equivalente donde los errores ya no presentan autocorrelación espacial y finalmente se realiza la selección de variables por el método LASSO clásico.

Por otro lado se presenta otra estrategia, donde se realiza en simultáneo la estimación de parámetros del modelo y la selección de variables mediante un modelo LASSO adaptativo. Se trata de una metodología desarrollada en Zhu et al. [57], la cual considera errores de tipo CAR o SAR, datos regulares y otro tipo de vecindad. A los efectos de poder comparar ambas estrategias, se adapta esta metodología al problema de interés.

Para ambas estrategias se llevan a cabo en el Capítulo 4 simulaciones en distintos contextos (*Problema 1 y 2*) y distintos escenarios: 40 para cada problema,

dependiendo de la cantidad de observaciones, el tipo de errores y los parámetros de las matrices de covarianzas.

Finalmente, en el Capítulo 5 se aplican también estas estrategias a un conjunto de datos reales extraídos de la Encuesta Continua de Hogares (ECH) de Uruguay correspondiente al año 2018, donde el objetivo consiste en seleccionar las variables más importantes para explicar el ingreso per cápita de un subconjunto de hogares del medio rural<sup>1</sup> de un total de 29 variables consideradas. Estas variables contemplan las características socioeconómicas del hogar y sus integrantes, así como también el nivel de confort y el grado de satisfacción de sus necesidades básicas.

---

<sup>1</sup>zonas rurales y localidades de hasta 5000 habitantes, que no son propietarios de la vivienda ni del terreno que ocupan y tienen al menos una necesidad básica insatisfecha

# Capítulo 2

## Selección de variables para datos independientes

Se considera un escenario donde se quiere explicar una variable aleatoria  $Y$  a partir de un conjunto fijo de  $p$  variables explicativas y un conjunto de  $n$  observaciones, independientes entre sí. En este contexto, la selección de variables consiste en buscar el “mejor” subconjunto posible de variables, una vez que se ha fijado el tipo de modelo a estimar.

El desarrollo de este capítulo se basa en Giraud [21] y Zhao and Yu [55] y Perrot-Dockès et al. [42]<sup>1</sup>.

### 2.1. Selección de variables en Modelos Lineales

Se considera el modelo lineal clásico:

$$Y = \underbrace{\mathbf{X}\beta}_{f(X)} + \varepsilon \quad (2.1)$$

donde:

- $Y$  es el vector aleatorio que representa la “respuesta” o “variable a explicar” del modelo, de dimensión  $n \times 1$
- $\mathbf{X}$  es la matriz que contiene en sus columnas a las “variables explicativas” del modelo. Se la denomina matriz de diseño de dimensión  $n \times p$ .

---

<sup>1</sup>Otras fuentes consultadas: González [22], Yuan and Lin [54] y Huang et al. [33]



- $\beta$  es el vector de coeficientes del modelo, de dimensión  $p \times 1$
- $\varepsilon$  es el vector de variables aleatorias independientes e idénticamente distribuidas, tal que  $E(\varepsilon) = \mathbf{0}$  y  $Var(\varepsilon) = \sigma^2 \mathbf{I}$ , donde  $\mathbf{0}$  representa el vector nulo de dimensión  $n$  e  $\mathbf{I}$  la matriz identidad de dimensión  $n \times n$ .

Sin pérdida de generalidad, se asume que  $\mathbf{X}$  e  $Y$  están estandarizadas, es decir, fueron centradas y normalizadas. La  $j$ -ésima columna de  $\mathbf{X}$  se denota como vector  $X_{.j}$ , para  $j = 1, 2, \dots, p$ , mientras que la  $i$ -ésima fila de  $\mathbf{X}$  se denota vector  $X_{i.}$ , para  $i = 1, \dots, n$ .

Asumiendo que entre las  $p$  variables consideradas hay  $p^*$  que “participan” en el modelo ( $0 < p^* \leq p$ ), es decir, los  $\beta_j$  correspondientes a esas variables son distintos de cero, y las restantes  $p - p^*$  no participan (es decir,  $\beta_j = 0$  para esas variables), la selección de variables busca identificar al siguiente conjunto:

$$A^* = \{j : \beta_j \neq 0, 1 \leq j \leq p\} \quad (2.2)$$

donde  $|A^*| = p^*$  se denomina soporte de  $\beta^1$ .

Por lo tanto, asociado a este conjunto, su complemento es el conjunto de variables que no participan del modelo: lo notaremos  $A^{*c}$  y  $|A^{*c}| = p - p^*$ .

En función de lo anterior, el nuevo modelo es:

$$Y = \underbrace{\sum_{j \in A^*} X_{.j} \beta_j}_{f(X)} + \varepsilon$$

Desde ahora hasta el final del capítulo, se asumirá que  $\varepsilon$  sigue una distribución normal de media  $\mathbf{0}$  y matriz de covarianzas  $\sigma^2 \mathbf{I}$ .

Si se sabe que  $f$  proviene de algún subespacio lineal  $S$  de  $\mathbb{R}^n$ , en lugar de estimar  $f$  maximizando la verosimilitud, se puede estimar  $f$  maximizando la verosimilitud restringida a que el estimador provenga de  $S$ .

Por ejemplo, en el contexto que se describió anteriormente, se puede tomar  $S = \text{span}\{X_{.j}, j \in A^*\}^2$ , entonces, el estimador que maximiza la verosimilitud restringida a  $S$  es  $\hat{f} = \text{Proj}_S Y$ , donde  $\text{Proj}_S: \mathbb{R}^n \rightarrow \mathbb{R}^n$  es el operador de proyección ortogonal en  $S$ .

Por otro lado, si no se conoce que  $f$  proviene de un subespacio lineal  $S$  de  $\mathbb{R}^n$ ,

<sup>1</sup> $|A| = \text{card}(A)$  representa el cardinal de  $A$ . Ver lista de notaciones.

<sup>2</sup> $S$  es el subespacio generado por  $\{X_{.j}, j \in A^*\}$

entonces se debe proceder de la siguiente manera:

- considerar una colección  $\{S_{m_M}, m_M \in M\}$  de subespacios lineales de  $\mathbb{R}^n$ , llamados modelos, donde  $M$  representa distintos subconjuntos de las  $p$  variables.
- Asociar a cada subespacio  $S_{m_M}$  el estimador máximo verosímil restringido  $\hat{f}_{m_M} = Proj_{S_{m_M}} Y$ .
- Estimar  $f$  por el “mejor” estimador de la colección  $\{\hat{f}_{m_M}, m_M \in M\}$

Se utiliza como criterio para cuantificar la calidad del estimador al riesgo  $L_2$ , definido como<sup>1</sup>:

$$R(\hat{f}_{m_M}) = E(\|\hat{f}_{m_M} - f\|_2^2) \quad (2.3)$$

Por lo que cada estimador  $\hat{f}_{m_M}$  tiene asociado un riesgo  $R(\hat{f}_{m_M})$  y el mejor estimador  $\hat{f}_{m_M}$  lo notaremos como  $\hat{f}_{m_0}$ , donde  $m_0 \in \arg \min_{m_M \in M} \{R(\hat{f}_{m_M})\}$ .

A pesar de lo anterior, se recalca que  $f$  puede no provenir de ninguno de los modelos  $\{S_{m_M}, m_M \in M\}$ .

Dado que  $f$  es desconocido, el riesgo  $R(\hat{f}_{m_M})$  es desconocido, por lo que debe estimarse. El mejor estimador de  $f$  es  $\hat{f}_{\widehat{m}_{M0}} : \widehat{m}_{M0} \in \arg \min_{m_M \in M} \{\hat{R}(\hat{f}_{m_M})\}$ , para algún estimador  $\hat{R}(\hat{f}_{m_M})$  de  $R(\hat{f}_{m_M})$ .

A continuación se presenta una expresión diferente para el riesgo  $R(\hat{f}_{m_M})$ . Partiendo de  $Y = f(\mathbf{X}) + \varepsilon$ , se obtiene la descomposición  $f - \hat{f}_{m_M} = (\mathbf{I} - Proj_{S_{m_M}})f - Proj_{S_{m_M}} \varepsilon$ . Entonces, se tiene:

$$\begin{aligned} R(\hat{f}_{m_M}) &= E\left(\|f - \hat{f}_{m_M}\|_2^2\right) = E\left(\|(\mathbf{I} - Proj_{S_{m_M}})f\|_2^2\right) + E\left(\|Proj_{S_{m_M}} \varepsilon\|_2^2\right) \\ &\quad - 2 \underbrace{E\left(\langle (\mathbf{I} - Proj_{S_{m_M}})f, Proj_{S_{m_M}} \varepsilon \rangle\right)}_{=0} \\ &= \|(\mathbf{I} - Proj_{S_{m_M}})f\|_2^2 + p_{m_M} \sigma^2 \end{aligned}$$

ya que  $Proj_{S_{m_M}} \varepsilon \sim N(0, \sigma^2 Proj_{S_{m_M}})$ , por tanto  $\|Proj_{S_{m_M}} \varepsilon\|_2^2 \sim \chi^2(p_{m_M} \sigma^2)$ , donde  $p_{m_M} = \dim(S_{m_M})$ . El riesgo  $R(\hat{f}_{m_M})$  incluye dos términos: el primero refleja la calidad de  $S_{m_M}$  para aproximar  $f$  (representa el sesgo), mientras que el segundo se incrementa linealmente con la dimensión de  $S_{m_M}$  (repre-

---

<sup>1</sup> $\|X\|_2$  representa la norma euclídeana. Ver notaciones.

senta la varianza). Ampliando  $S_{m_M}$  se reduce el primer término pero se incrementa el segundo. El modelo  $S_{m_0}$  es entonces el modelo en  $\{S_{m_M}, m_M \in M\}$  que logra la mejor compensación entre el sesgo y la varianza.

Los métodos de selección de variables más conocidos que surgen a partir de estimar el riesgo  $R(\hat{f}_{m_M})$  son el AIC y el BIC.

- **Akaike Information Criterion (AIC, [1]):** Surge como un estimador insesgado del riesgo  $R(\hat{f}_{m_M})$ . Se parte de la descomposición  $Y - \hat{f}_{m_M} = (\mathbf{I} - Proj_{S_{m_M}})(f + \varepsilon)$ , tal que:

$$\begin{aligned} E\left(\|Y - \hat{f}_{m_M}\|_2^2\right) &= E\left(\|(\mathbf{I} - Proj_{S_{m_M}})f\|_2^2 + \underbrace{2\langle (\mathbf{I} - Proj_{S_{m_M}})f, \varepsilon \rangle}_{=0} \right. \\ &\quad \left. + \|(\mathbf{I} - Proj_{S_{m_M}})\varepsilon\|_2^2\right) \\ &= \|(\mathbf{I} - Proj_{S_{m_M}})f\|_2^2 + (n - p_{m_M})\sigma^2 \\ &= R(\hat{f}_{m_M}) + (n - 2p_{m_M})\sigma^2 \end{aligned}$$

donde  $p_{m_M}$  es la cantidad de parámetros del modelo  $\hat{f}_{m_M}$ .

Como consecuencia surge el estimador insesgado  $\hat{R}(\hat{f}_{m_M}) = \|Y - \hat{f}_{m_M}\|_2^2 + (2p_{m_M} - n)\sigma^2$ , que al quitar el término  $-n\sigma^2$  (que no afecta la elección de  $\widehat{m_M}$ ) deriva en el AIC:

$$\hat{m}_{AIC} \in \underset{m_M \in M}{\operatorname{argmín}}\{\|Y - \hat{f}_{m_M}\|_2^2 + 2p_{m_M}\sigma^2\} \quad (2.4)$$

Si bien este criterio es muy popular e intuitivo, puede producir resultados muy pobres en algunos casos debido a no considerar la variabilidad de los riesgos estimados  $\hat{R}(\hat{f}_{m_M})$  alrededor de su media  $R(\hat{f}_{m_M})$ . En particular, cuando  $\hat{R}(\hat{f}_{m_M})$  es muy pequeño (mucho menor que  $\hat{R}(\hat{f}_{m_0})$ ) este criterio selecciona modelos con muchas más variables que el valor óptimo ( $m_0$ ).

- **AIC penalizado ([2]):** Este criterio surge con el fin de corregir el problema del AIC en lo referente a la tendencia de seleccionar modelos con muchas variables. Más precisamente, se sustituye el término  $2p_{m_M}\sigma^2$  por otro que “penalice” la cantidad de variables del modelo, es decir, se busca el  $\widehat{m_M}$  que minimice la función  $\|Y - \hat{f}_{m_M}\|_2^2 + \sigma^2 pen(m_M)$  con  $m_M \in M$ ,

donde  $pen(m_M) : M \rightarrow \mathbb{R}^+$  es la “función de penalidad”.

Lo deseable es que la función de penalidad utilizada verifique que el riesgo  $R(\hat{f}_{\hat{m}})$  sea lo más cercano posible a  $R(\hat{f}_{m_0})$ .

Se asocia a la colección de modelos  $\{S_{m_M}, m_M \in M\}$ , una distribución de probabilidad  $\pi = \{\pi_{m_M}, m_M \in M\}$  en  $M$ . Fijada una probabilidad  $\pi$  y una constante  $K > 1$ , el criterio AIC penalizado selecciona el modelo que verifique:

$$\hat{m}_{AIC_{pen}} \in \underset{m_M \in M}{\operatorname{argmín}} \left\{ \|Y - \hat{f}_{m_M}\|_2^2 + \sigma^2 K \left( \sqrt{p_{m_M}} + \sqrt{2 \log(1/\pi_{m_M})} \right)^2 \right\} \quad (2.5)$$

donde  $pen(m_M) = K \left( \sqrt{p_{m_M}} + \sqrt{2 \log(1/\pi_{m_M})} \right)^2$ .

El criterio  $AIC_{pen}$  depende fuertemente de la probabilidad  $\pi$ , por lo cual es muy importante elegirla de forma apropiada. La misma puede reflejar algún tipo de conocimiento previo, pero la mayoría de las veces es elegida ad hoc.

Un algoritmo computacional que utiliza el AIC penalizado es el método stepwise. En este algoritmo, se parte de un modelo sin variables (sólo con intercepto) y se construye una secuencia de modelos agregando o quitando una variable en cada paso, tratando de minimizar el  $AIC_{pen}$ . Cuando el  $AIC_{pen}$  ya no se reduce significativamente al agregar o quitar ninguna variable, el algoritmo termina, y el modelo seleccionado es el correspondiente al último paso.

- **BIC ([46]):** En la perspectiva bayesiana, la media  $f$  de  $Y$  es asumida como la salida de un esquema de muestreo donde un modelo  $S_{m_M}$  es sorteado de acuerdo a la distribución  $\{\pi_{m_M}, m_M \in M\}$ , entonces  $f$  es sorteado de acuerdo a la distribución  $d\Pi(f|m_M)$  en  $S_{m_M}$ . La media  $f$  de  $Y$  es entonces muestreada de acuerdo a la distribución de:

$$d\Pi(f) = \sum_{m_M \in M} \pi_{m_M} d\Pi(f|m_M)$$

con  $\{\pi_{m_M}, m_M \in M\}$  una distribución en  $M$  y  $d\Pi(f|m)$  una distribución en  $S_{m_M}$ . En este esquema probabilístico, donde  $m_M$ ,  $f$  e  $Y$  son varia-

bles aleatorias, tiene sentido considerar la probabilidad  $\Pi(m_M|Y)$  de un modelo  $S_{m_M}$  dadas las observaciones  $Y$ . Computando las probabilidades condicionales se obtiene:

$$\Pi(m_M|Y) = \frac{\pi_{m_M} d\pi(Y|m_M)}{d\pi(Y)} = \frac{\pi_{m_M} \int_{f \in S_{m_M}} e^{-\|Y-f\|_2^2/(2\sigma^2)} d\Pi(f|m_M)}{\sum_{m'_M} \pi_{m'_M} \int_{f \in S_{m'_M}} e^{-\|Y-f\|_2^2/(2\sigma^2)} d\Pi(f|m'_M)}$$

Bajo ciertas condiciones técnicas sobre la distribución  $d\Pi(f|m_M)$ , una expansión asintótica cuando  $n \rightarrow \infty$  permite obtener para  $m_M, m'_M \in M$ :

$$\log \left( \frac{\Pi(m_M|Y)}{\Pi(m'_M|Y)} \right) \approx \frac{\|Y - \hat{f}_{m'_M}\|_2^2 - \|Y - \hat{f}_{m_M}\|_2^2}{2\sigma^2} + \frac{p_{m'_M} - p_{m_M}}{2} \log(n) + \log \left( \frac{\pi_{m_M}}{\pi_{m'_M}} \right) + O(1)$$

cuando  $n \rightarrow \infty$ . Esta expansión asintótica sugiere elegir  $m_M$  minimizando el criterio

$$crit(m_M) = \|Y - \hat{f}_{m_M}\|_2^2 + \sigma^2 p_{m_M} \log(n) + 2\sigma^2 \log(\pi_{m_M}^{-1})$$

Asumiendo una distribución  $\pi_{m_M}$  uniforme en  $M$ , se obtiene el BIC:

$$\hat{m}_{BIC} \in arg \min_{m_M \in M} \{ \|Y - \hat{f}_{m_M}\|_2^2 + \sigma^2 p_{m_M} \log(n) \} \quad (2.6)$$

El término  $\sigma^2 p_{m_M} \log(n)$  se incrementa más rápidamente que el término  $2p_{m_M}\sigma^2$  que aparece en el criterio AIC, por lo que puede ser demasiado grande cuando hay un número exponencial de modelos por dimensión.

## 2.2. Método LASSO y selección de variables

En el modelo LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani [50]) se quiere minimizar la siguiente función, con respecto al parámetro desconocido  $\beta$ :

$$\mathcal{L}(\beta) = \underbrace{\|Y - \mathbf{X}\beta\|_2^2}_{\mathcal{L}_1(\beta)} + \lambda \underbrace{\|\beta\|_1}_{\mathcal{L}_2(\beta)} \text{ para algún } \lambda > 0 \quad (2.7)$$

El estimador  $\hat{\beta}$  de  $\beta$  se obtiene minimizando la función  $\mathcal{L}(\beta)$  definida en (2.7)<sup>1</sup>. Como esta función es convexa pero no diferenciable (la función  $\mathcal{L}_1(\beta)$  es diferenciable, mientras que la función  $\mathcal{L}_2(\beta)$  es convexa pero no diferenciable), el estimador puede no ser único.

Para interpretar geoméricamente el modelo (2.7), se considera la esfera- $L_1$   $B_{L_1}(r)$ :

$$B_{L_1}(r) = \{\beta \in \mathbb{R}^p : \|\beta\|_1 \leq r\}$$

El estimador LASSO  $\hat{\beta}$  es solución de:

$$\hat{\beta} \in \underset{\beta \in B_{L_1}(r)}{\operatorname{argmín}} \|Y - \mathbf{X}\beta\|_2^2$$

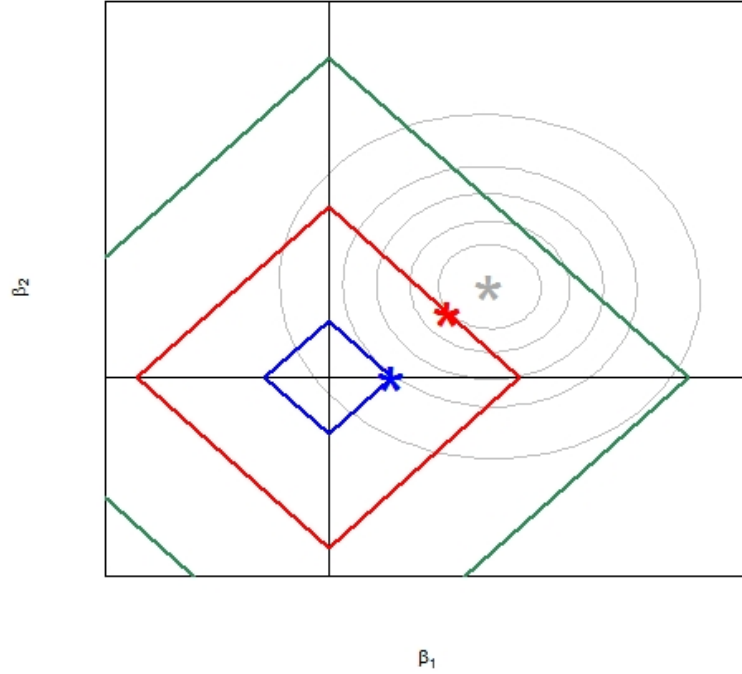
En la Figura 2.1 se grafican algunas curvas de nivel de la función  $\beta \mapsto \|Y - \mathbf{X}\beta\|_2^2$  (líneas en gris) junto a la restricción  $\|\beta\|_1 \leq r$  (líneas continuas), para tres valores distintos de  $r$  y  $\beta \in \mathbb{R}^2$ .

El asterisco gris representa el estimador  $\beta$  sin restricciones, es decir, el estimador mínimo cuadrático. En color verde se grafica la restricción  $\|\beta\|_1 \leq 0.9$ , para este valor de  $r$  el estimador LASSO coincide con el estimador mínimo cuadrático.

Sin embargo, cuando  $r = 0.48$  (línea roja), la restricción no incluye al estimador  $\beta$  por mínimos cuadrados, por lo que el estimador LASSO (asterisco rojo) difiere un poco del anterior. Cuando  $r = 0.16$  (línea azul) ocurre lo mismo, y además el estimador LASSO dista un poco más del estimador mínimo cuadrático. En este caso se puede apreciar que el óptimo ocurre cuando  $\beta_2 = 0$ . En resumen: para valores de  $r$  suficientemente pequeños (lo que equivale a valores suficientemente grandes de  $\lambda$ ), algunas coordenadas de  $\hat{\beta}$  son cero. Esto ilustra el hecho de que el método LASSO seleccione variables para  $\lambda$  suficientemente grande.

---

<sup>1</sup> $\|X\|_1$  representa la norma  $L_1$ . Ver notaciones.



**Figura 2.1:** Geometría del método LASSO

Se consideran las derivadas de primer y segundo orden de  $\mathcal{L}(\beta)$ , denotadas como  $\frac{\partial \mathcal{L}(\beta)}{\partial \beta}$  y  $\frac{\partial^2 \mathcal{L}(\beta)}{\partial^2 \beta}$  respectivamente. La derivada de  $\frac{\partial \mathcal{L}(\beta)}{\partial \beta}$  se compone de la suma de  $\frac{\partial \mathcal{L}_1(\beta)}{\partial \beta}$  y  $\frac{\partial \mathcal{L}_2(\beta)}{\partial \beta}$ . En el primer sumando se tiene  $\frac{\partial \mathcal{L}_1(\beta)}{\partial \beta} = -2\mathbf{X}'(Y - \mathbf{X}\beta)$ , mientras que el segundo equivale a  $\frac{\partial \mathcal{L}_2(\beta)}{\partial \beta} = \lambda z$ , donde  $z$  es un vector subgradiente del subdiferencial de  $\|\beta\|_1$ , es decir  $z \in \partial\|\beta\|_1$ <sup>1</sup>.

Por otro lado, se tiene que  $\frac{\partial^2 \mathcal{L}(\beta)}{\partial^2 \beta} = 2\mathbf{X}'\mathbf{X}$ . Al ser la derivada segunda estrictamente positiva, para hallar el mínimo alcanza con igualar  $\frac{\partial \mathcal{L}(\beta)}{\partial \beta}$  a cero y despejar  $\beta$ . Dicho de otra forma, para hallar  $\beta$  que minimice  $\mathcal{L}(\beta)$ , se busca el  $\hat{\beta}$  que verifique:

$$\frac{\partial \mathcal{L}(\hat{\beta})}{\partial \beta} = -2\mathbf{X}'(Y - \mathbf{X}\hat{\beta}) + \lambda \hat{z} = \mathbf{0}$$

donde se sustituyó  $z$  por  $\hat{z}$ , con  $\hat{z} \in \partial\|\hat{\beta}\|_1$  y  $\hat{z} = (\hat{z}_1, \dots, \hat{z}_n)'$ , por el resultado (1.5) del Anexo 1, se obtiene que  $\hat{z}_j = \text{sign}(\hat{\beta}_j)$ <sup>2</sup> si  $\hat{\beta}_j \neq 0$ , y  $\hat{z}_j \in [-1, 1]$  si  $\hat{\beta}_j = 0$ .

<sup>1</sup> $\partial f(x)$  representa el subdiferencial de  $f(x)$ . Ver notaciones y Anexo 1

<sup>2</sup> $\text{sign}(x)$  representa a función signo. Ver notaciones.

Ahora bien, aplicando la Propiedad 1 (sección 1.0.1 del Anexo 1) se obtiene:

$$\begin{aligned}\frac{\partial \mathcal{L}(\hat{\beta})}{\partial \beta} &= -2\mathbf{X}'(Y - \mathbf{X}\hat{\beta}) + \lambda \hat{z} = \mathbf{0} \\ \mathbf{X}'(Y - \mathbf{X}\hat{\beta}) &= \frac{\lambda}{2} \hat{z} \\ \mathbf{X}'Y - \mathbf{X}'\mathbf{X}\hat{\beta} &= \frac{\lambda}{2} \hat{z} \\ \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'Y - \frac{\lambda}{2} \hat{z}\end{aligned}\quad (2.8)$$

Si  $\mathbf{X}$  es ortogonal, entonces  $\mathbf{X}'\mathbf{X} = \mathbf{I}$  y el resultado anterior se reduce a:

$$\hat{\beta} = \mathbf{X}'Y - \frac{\lambda}{2} \hat{z}\quad (2.9)$$

Volviendo al resultado (2.9), cuando  $\hat{\beta}_j \neq 0$ :

$$\hat{\beta}_j = X'_{\cdot j}Y - \frac{\lambda}{2} \text{sign}(\hat{\beta}_j)\quad (2.10)$$

Reordenando y tomando el signo en ambos lados de la igualdad, queda:

$$\text{sign}\left(\hat{\beta}_j + \frac{\lambda}{2} \text{sign}(\hat{\beta}_j)\right) = \text{sign}(X'_{\cdot j}Y)\quad (2.11)$$

$$\text{sign}(\hat{\beta}_j) = \text{sign}(X'_{\cdot j}Y)\quad (2.12)$$

El paso de (2.11) a (2.12) se cumple dado que  $\lambda > 0$ . Utilizando este resultado en (2.10):

$$\hat{\beta}_j = X'_{\cdot j}Y - \frac{\lambda}{2} \text{sign}(X'_{\cdot j}Y)$$

Existen dos posibilidades:

1. Si  $X'_{\cdot j}Y < 0$   $\Rightarrow \text{sign}(X'_{\cdot j}Y) = -1 \Rightarrow \hat{\beta}_j = X'_{\cdot j}Y + \frac{\lambda}{2}$ .

Tomando signo en ambos lados de la ecuación y considerando el resultado



(2.12):

$$\begin{aligned} \text{sign}(\hat{\beta}_j) &= \text{sign}\left(X'_{\cdot j}Y + \frac{\lambda}{2}\right) = \text{sign}(X'_{\cdot j}Y) = -1 \\ &\Rightarrow X'_{\cdot j}Y + \frac{\lambda}{2} < 0 \Rightarrow X'_{\cdot j}Y < -\frac{\lambda}{2} \end{aligned}$$

2. Si  $X'_{\cdot j}Y > 0 \Rightarrow \text{sign}(X'_{\cdot j}Y) = 1 \Rightarrow \hat{\beta}_j = X'_{\cdot j}Y - \frac{\lambda}{2}$ .

Nuevamente tomando signo en ambos lados y considerando el resultado (2.12):

$$\begin{aligned} \text{sign}(\hat{\beta}_j) &= \text{sign}\left(X'_{\cdot j}Y - \frac{\lambda}{2}\right) = \text{sign}(X'_{\cdot j}Y) = 1 \\ &\Rightarrow X'_{\cdot j}Y - \frac{\lambda}{2} > 0 \Rightarrow X'_{\cdot j}Y > \frac{\lambda}{2} \end{aligned}$$

Se obtiene que si  $\hat{\beta}_j \neq 0 \Rightarrow |X'_{\cdot j}Y| > \frac{\lambda}{2}$ .

Por otro lado, de la ecuación (2.9) y considerando el valor apropiado de  $\hat{z}$ , se obtiene que  $\hat{\beta}_j = 0 \Leftrightarrow X'_{\cdot j}Y - \frac{\lambda}{2}z = 0$ , con  $z \in [-1, 1]$ . Y eso sólo se verifica para  $-\frac{\lambda}{2} \leq X'_{\cdot j}Y \leq \frac{\lambda}{2}$ .

Entonces, se concluye que

$$\boxed{\hat{\beta}_j \neq 0 \Leftrightarrow |X'_{\cdot j}Y| > \frac{\lambda}{2}}$$

Este resultado es muy importante, ya que en el contexto de una matriz  $\mathbf{X}$  ortogonal, el valor de  $\lambda$  determina directamente la selección de variables. Esto es, se seleccionarán las variables que tengan producto interno con  $Y$ , en valor absoluto, mayor a  $\frac{\lambda}{2}$ .

Por otro lado, cuando  $\mathbf{X}$  no es ortogonal, no hay una forma analítica para  $\hat{\beta}$ . Se define el conjunto  $\hat{A} = \{j : \hat{\beta}_j \neq 0, 1 \leq j \leq p\}$ , es decir, el conjunto de índices de las variables que participan del modelo estimado. Este conjunto no debe confundirse con  $A^*$ , el cual representa el conjunto de índices que participan del verdadero modelo.

Partiendo de la ecuación (2.8), premultiplicando por  $\hat{\beta}'$ , se obtiene:

$$0 \leq \hat{\beta}' \mathbf{X}' \mathbf{X} \hat{\beta} = \langle \hat{\beta}, X'Y - \frac{\lambda}{2} \hat{z} \rangle = \sum_{j \in \hat{A}} \hat{\beta}_j \left( X'_{\cdot j} Y - \frac{\lambda}{2} \text{sign}(\hat{\beta}_j) \right)$$

Para que el resultado anterior se verifique, se debe cumplir que  $\hat{\beta}_j \left( X'_{\cdot j} Y - \frac{\lambda}{2} \text{sign}(\hat{\beta}_j) \right) \geq 0$  para todo  $j \in \hat{A}$ . Si  $\hat{\beta}_j > 0$ , entonces  $\lambda \leq 2X'_{\cdot j} Y$ . Por otro lado, si  $\hat{\beta}_j < 0$ , entonces  $\lambda \geq -2X'_{\cdot j} Y$ . En resumen,  $\lambda \leq |2X'_{\cdot j} Y|$  para todo  $j \in \hat{A}$ .

Si  $\|\mathbf{X}'Y\|_{\infty} \leq \frac{\lambda}{2}$ , entonces  $\hat{\beta}$  debe ser igual a  $\mathbf{0}$  para que se verifique la desigualdad anterior. Cuando  $\|\mathbf{X}'Y\|_{\infty} \geq \frac{\lambda}{2}$ , el estimador LASSO  $\hat{\beta}$  es distinto de  $\mathbf{0}$ , pero no es posible obtener una forma analítica para el mismo.

Además de las referencias mencionadas al principio de este capítulo, para profundizar en el método LASSO pueden consultarse las siguientes: Tibshirani [50], Tibshirani [51], Huang et al. [32], Hastie et al. [27], Knight and Fu [34], Zou [59] y Simon et al. [48].

## 2.3. Consistencia en signo del estimador LASSO

En esta sección se define la consistencia en signo del estimador  $\hat{\beta}$  utilizando LASSO y se definen las “condiciones de irrepresentabilidad”, necesarias para que se verifique la consistencia en signo. A su vez, se presenta un teorema que reúne el conjunto de condiciones necesarias para dicha consistencia.

**Definición 1.** *El estimador  $\hat{\beta}$  es **igual en signo** al verdadero vector de parámetros  $\beta$  (se denota  $\hat{\beta} =_s \beta$ ), si y sólo si:*

$$\text{sign}(\hat{\beta}) = \text{sign}(\beta)$$

donde la igualdad es coordenada a coordenada.

**Definición 2.** *El estimador LASSO es **fuertemente consistente en signo** si existe  $\lambda$ , independiente de  $Y$  y  $\mathbf{X}$ , tal que:*

$$\lim_{n \rightarrow \infty} P(\hat{\beta} =_s \beta) = 1$$

---

<sup>1</sup> $\|X\|_{\infty}$  representa la norma infinito. Ver notaciones.

Cabe recordar que  $\hat{\beta}$  depende de  $\lambda$  y de  $n$ .

La consistencia en signo garantiza que se identifican las variables que participan en el modelo (reunidas en el conjunto  $A^*$ ) identificando correctamente su signo, lo que significa que  $\text{sign}(\hat{\beta}_j) = \text{sign}(\beta_j)$  para todo  $j \in A^*$ . Esto es muy importante para la interpretación del modelo: un modelo estimado con los signos cambiados puede ser engañoso y es cuestionable si puede calificarse como un modelo correctamente estimado.

Sin pérdida de generalidad, el modelo (2.1) puede reescribirse, reordenando de la siguiente manera:

$$Y = [\mathbf{X}_{A^*} \ \mathbf{X}_{A^{*c}}] \begin{bmatrix} \beta_{A^*} \\ \beta_{A^{*c}} \end{bmatrix} + \varepsilon \quad (2.13)$$

Donde:

- $\beta_{A^*} = (\beta_1, \dots, \beta_{p^*})'$ , tal que  $\beta_j \neq 0$  para  $j = 1, \dots, p^*$
- $\beta_{A^{*c}} = (\beta_{p^*+1}, \dots, \beta_p)'$ , tal que  $\beta_j = 0$  para  $j = p^* + 1, \dots, p$ .  $\beta_{A^{*c}} = \mathbf{0}$
- $\mathbf{X}_{A^*} = [X_{.1} \ \dots \ X_{.p^*}]$
- $\mathbf{X}_{A^{*c}} = [X_{.p^*+1} \ \dots \ X_{.p}]$

Se define  $\mathbf{C}^n = \frac{1}{n} \mathbf{X}' \mathbf{X}$ , que puede escribirse como:

$$\mathbf{C}^n = \begin{pmatrix} \mathbf{C}_{11}^n & \mathbf{C}_{12}^n \\ \mathbf{C}_{21}^n & \mathbf{C}_{22}^n \end{pmatrix} \quad (2.14)$$

con:

$$\mathbf{C}_{11}^n = \frac{1}{n} \mathbf{X}'_{A^*} \mathbf{X}_{A^*} \quad \mathbf{C}_{12}^n = \frac{1}{n} \mathbf{X}'_{A^*} \mathbf{X}_{A^{*c}}$$

$$\mathbf{C}_{21}^n = \frac{1}{n} \mathbf{X}'_{A^{*c}} \mathbf{X}_{A^*} \quad \mathbf{C}_{22}^n = \frac{1}{n} \mathbf{X}'_{A^{*c}} \mathbf{X}_{A^{*c}}$$

Asumiendo que  $\mathbf{C}_{11}^n$  es invertible se definen a continuación las **condiciones de irrepresentabilidad fuerte y débil**.

**Definición 3. Condición de irrepresentabilidad fuerte:** Existe un vector  $\delta$  constante y con todas sus coordenadas positivas, tal que:

$$|\mathbf{C}_{21}^{\mathbf{n}}(\mathbf{C}_{11}^{\mathbf{n}})^{-1}\text{sign}(\beta_{A^*})| \leq \mathbf{1} - \delta$$

donde  $\mathbf{1}$  es un “vector de unos” de dimensión  $p - p^*$  y la desigualdad es coordenada a coordenada.

**Definición 4. Condición de irrepresentabilidad débil:**

$$|\mathbf{C}_{21}^{\mathbf{n}}(\mathbf{C}_{11}^{\mathbf{n}})^{-1}\text{sign}(\beta_{A^*})| \leq \mathbf{1}$$

donde al igual que en el caso anterior, la desigualdad es coordenada a coordenada.

A continuación se proporciona una interpretación para las condiciones de irrepresentabilidad definidas anteriormente.

Para que se cumpla cualquiera de las dos condiciones anteriores, independientemente del signo de  $\beta_{A^*}$ , se debe verificar que para cada fila de la matriz  $\mathbf{C}_{21}^{\mathbf{n}}(\mathbf{C}_{11}^{\mathbf{n}})^{-1}$  la suma de sus elementos en valor absoluto sea menor a  $1 - \delta$ .

Definiendo como  $c_i$  a cada columna de la matriz  $\mathbf{C}_{21}^{\mathbf{n}}(\mathbf{C}_{11}^{\mathbf{n}})^{-1}$ , la afirmación anterior se justifica en las propiedades del valor absoluto:

$$\begin{aligned} |c_1\text{sign}(\beta_1) + \cdots + c_{p^*}\text{sign}(\beta_{p^*})| &\leq |c_1\text{sign}(\beta_1)| + \cdots + |c_{p^*}\text{sign}(\beta_{p^*})| \\ &= |c_1||\text{sign}(\beta_1)| + \cdots + |c_{p^*}||\text{sign}(\beta_{p^*})| \\ &= |c_1| + \cdots + |c_{p^*}| \end{aligned}$$

Entonces, para que se cumplan las condiciones de irrepresentabilidad, independientemente del signo de  $\beta_{A^*}$ , se debe verificar la siguiente condición:

$$\|(\mathbf{C}_{11}^{\mathbf{n}})^{-1}\mathbf{C}_{12}^{\mathbf{n}}\|_1 \leq \mathbf{1} - \delta$$

donde  $(\mathbf{C}_{11}^{\mathbf{n}})^{-1}\mathbf{C}_{12}^{\mathbf{n}}$  es la traspuesta de  $\mathbf{C}_{21}^{\mathbf{n}}(\mathbf{C}_{11}^{\mathbf{n}})^{-1}$  y la norma 1 matricial se define como  $\|X\|_1 = \max_{1 \leq j \leq J} \sum_{i=1}^{i=I} |X_{ij}|$ , para una matriz  $\mathbf{X}$  de dimensión  $I \times J$ , es decir, es la máxima suma absoluta de las columnas de la matriz  $\mathbf{X}$ . Nuevamente la desigualdad es coordenada a coordenada.

A su vez, al expresar  $(\mathbf{C}_{11}^{\mathbf{n}})^{-1}\mathbf{C}_{12}^{\mathbf{n}}$ , en función de  $\mathbf{X}_{A^*}$  y  $\mathbf{X}_{A^{*c}}$ , se obtiene:

$$\left\| \left( \mathbf{X}'_{A^*} \mathbf{X}_{A^*} \right)^{-1} \mathbf{X}'_{A^*} \mathbf{X}_{A^{*c}} \right\|_1 \leq \mathbf{1} - \delta$$

Cada columna de la matriz anterior puede interpretarse como el vector de coeficientes estimados de la regresión de  $X_{.j}$  con  $j \in A^{*c}$  en función de  $\mathbf{X}_{A^*}$ , es decir:  $X_{.j} = \mathbf{X}_{A^*} \theta^j + e^j$ , donde  $\hat{\theta}^j = \left[ \left( \mathbf{X}'_{A^*} \mathbf{X}_{A^*} \right)^{-1} \mathbf{X}'_{A^*} \mathbf{X}_{A^{*c}} \right]_j$ , y  $j \in A^{*c}$ . Dicho de otra manera, la matriz anterior contiene los coeficientes de regresión de las variables irrelevantes (con índices en  $A^{*c}$ ) en función de las relevantes (con índices en  $A^*$ ), por lo que las condiciones definidas anteriormente implican que la suma en valor absoluto de dichos coeficientes debe ser menor a uno, y de allí proviene la “irrepresentabilidad”.

La siguiente proposición presenta un resultado importante para verificar la consistencia fuerte.

**Proposición 1** (Cota inferior a la probabilidad de seleccionar el modelo correcto). *Sea  $\hat{\beta}$  la solución del problema LASSO, entonces:*

$$P(\hat{\beta} =_s \beta) \geq P(A_n \cap B_n) \tag{2.15}$$

siendo

- $A_n = \left\{ \left| \left( \mathbf{C}_{\mathbf{11}}^n \right)^{-1} W_{A^*} \right| < \sqrt{n} \left( \left| \beta_{A^*} \right| - \frac{\lambda}{2n} \left| \left( \mathbf{C}_{\mathbf{11}}^n \right)^{-1} \text{sign}(\beta_{A^*}) \right| \right) \right\}$
- $B_n = \left\{ \left| \mathbf{C}_{\mathbf{21}}^n \left( \mathbf{C}_{\mathbf{11}}^n \right)^{-1} W_{A^*} - W_{A^{*c}} \right| < \frac{\lambda}{2\sqrt{n}} \left( \mathbf{1} - \left| \mathbf{C}_{\mathbf{21}}^n \left( \mathbf{C}_{\mathbf{11}}^n \right)^{-1} \text{sign}(\beta_{A^*}) \right| \right) \right\}$

con  $W_{A^*} = \frac{1}{\sqrt{n}} \mathbf{X}'_{A^*} \varepsilon$  y  $W_{A^{*c}} = \frac{1}{\sqrt{n}} \mathbf{X}'_{A^{*c}} \varepsilon$ , donde las desigualdades en  $A_n$  y  $B_n$  son coordinada a coordinada.

El Teorema 1 presenta las condiciones que deben cumplirse para que el estimador LASSO  $\hat{\beta}$  sea fuertemente consistente en signo.

**Teorema 1** (Condiciones para la consistencia fuerte en signo). *Suponiendo que  $\varepsilon_i \sim N(0, \sigma^2)$  iid, y si existen constantes  $M_1$ ,  $M_2$  y  $M_3$  estrictamente positivas y  $c_1, c_2$  con  $0 \leq c_1 \leq c_2 \leq 1$  tales que:*

1.  $\frac{1}{n} X'_{.j} X_{.j} \leq M_1 \forall j$
2.  $\rho_{\min}(\mathbf{C}_{11}^n) \geq M_2$ , donde  $\rho_{\min}(\mathbf{C}_{11}^n)$  representa el menor valor propio de la matriz  $\mathbf{C}_{11}^n$
3.  $p^* = O(n^{c_1/2})$
4.  $n^{\frac{1-c_2}{2}} \min_{1 \leq j \leq p^*} |\beta_j| \geq M_3$
5. se cumple la condición de irrepresentabilidad fuerte (Definición 3).

Entonces,  $\forall \lambda > 0$  ( $\lambda$  depende de  $n$ ), tal que  $\frac{\lambda}{\sqrt{n}} \xrightarrow{n \rightarrow \infty} \infty$  y  $\frac{\lambda}{\sqrt{n}} = o(n^{(c_2-c_1)/2})$ , se verifica la consistencia fuerte en signo.

El Teorema 1 se demuestra gracias a la Proposición 1. La hipótesis 2 del Teorema 1 implica que  $\mathbf{C}_{11}^n$  es definida positiva. La hipótesis 3 establece que el número de variables que participan del modelo tiene que ser acotado y no demasiado grande. La hipótesis 4 indica que  $\min_{1 \leq j \leq p^*} |\beta_j|$  no puede acercarse demasiado a 0. Por otra parte, se requiere que si  $\lambda$  crece con  $n$ , no lo haga más rápido que  $n^{(1+c_2-c_1)/2} < n$ . La demostración de la Proposición 1 y el Teorema 1 se encuentran en el Apéndice 1.

# Capítulo 3

## Selección de variables para datos dependientes: caso de la dependencia espacial

En este capítulo se presenta un teorema que enuncia las condiciones necesarias para que el estimador  $\hat{\beta}$  sea fuertemente consistente en signo cuando el modelo no tiene errores independientes, y se plantea una forma de realizar la selección de variables en este contexto. Se demuestra que en el caso espacial, bajo ciertas condiciones se cumplen las hipótesis de dicho teorema y se sugiere un procedimiento para eliminar la dependencia en ese caso. También se presenta otra metodología para la selección de variables en el caso espacial, que fue desarrollada en el artículo de Zhu et al [57].

### 3.1. Selección de variables en un modelo lineal dependiente

Se considera el modelo lineal (2.1) definido en la sección 2.1 del capítulo anterior:

$$Y = \mathbf{X}\beta + \varepsilon \tag{3.1}$$

con la diferencia de que en este caso  $\varepsilon_i$  no es independiente de  $\varepsilon_j$  para  $i \neq j$ . El vector  $\varepsilon$  sigue una distribución normal multivariada centrada con matriz de covarianzas  $\Sigma$ , es decir  $\varepsilon \sim N(\mathbf{0}, \Sigma)$ , donde  $\Sigma_{ij} = Cov(\varepsilon_i, \varepsilon_j)$   $i = 1, \dots, n$ ,

$j = 1, \dots, n$ , y a priori se supone que  $\Sigma_{ij} \neq 0$  para algunos  $i, j = 1, \dots, n$ , con  $i \neq j$ .

En este contexto no se cumple la hipótesis de independencia de los errores del Teorema 1, por lo que no puede garantizarse la consistencia del estimador LASSO estudiado en el capítulo anterior.

Para superar esta limitante, se propone una alternativa: transformar los datos en independientes para poder aplicar el método LASSO estudiado en el capítulo anterior.

### 3.1.1. Eliminación de la dependencia en los errores

Esta estrategia parte del supuesto de que los errores siguen una distribución normal centrada en cero y con matriz de covarianzas  $\Sigma$ . Cuando esta matriz es conocida, basta con aplicar la siguiente transformación para obtener datos independientes<sup>1</sup>:

$$\underbrace{\Sigma^{-1/2}Y}_{\dot{Y}} = \underbrace{\Sigma^{-1/2}X}_{\dot{X}}\beta + \underbrace{\Sigma^{-1/2}\varepsilon}_{\dot{\varepsilon}} \quad (3.2)$$

donde  $\dot{\varepsilon} \sim N(\mathbf{0}, \mathbf{I})$  ya que  $Var(\dot{\varepsilon}) = Var(\Sigma^{-1/2}\varepsilon) = \Sigma^{-1/2}\Sigma(\Sigma^{-1/2})' = \Sigma^{-1/2}\Sigma^{1/2}\Sigma^{1/2}(\Sigma^{-1/2})' = \mathbf{I}$ . Asociado a estos se definen también  $\dot{X}_{A^*}$ ,  $\dot{X}_{A^{*c}}$  y  $\dot{C}^n = \begin{pmatrix} \dot{C}_{11}^n & \dot{C}_{12}^n \\ \dot{C}_{21}^n & \dot{C}_{22}^n \end{pmatrix}$ , al igual que en la Sección 2.3 del capítulo anterior.

Una vez aplicada esta transformación, se realiza la selección de variables estimando un modelo LASSO al igual que en el Capítulo 2. El estimador LASSO de  $\beta$  asociado a este modelo se denota  $\dot{\beta}$ .

En la práctica, conocer la verdadera matriz de covarianzas  $\Sigma$  no es posible, por lo que será necesario estimarla previamente.

Denotando como  $\hat{\Sigma}$  al estimador de  $\Sigma$ , se define el nuevo modelo transformado:

$$\underbrace{\hat{\Sigma}^{-1/2}Y}_{\hat{Y}} = \underbrace{\hat{\Sigma}^{-1/2}X}_{\hat{X}}\beta + \underbrace{\hat{\Sigma}^{-1/2}\varepsilon}_{\hat{\varepsilon}} \quad (3.3)$$

donde  $\tilde{\beta}$  representa el estimador LASSO de  $\beta$  para este modelo. A su vez se definen también en este caso  $\tilde{X}_{A^*}$ ,  $\tilde{X}_{A^{*c}}$  y  $\tilde{C}^n = \begin{pmatrix} \tilde{C}_{11}^n & \tilde{C}_{12}^n \\ \tilde{C}_{21}^n & \tilde{C}_{22}^n \end{pmatrix}$ .

<sup>1</sup>Esta estrategia se denomina “Whitening” en el artículo de referencia (Perrot-Dockès et al. [42]).



Para poder estimar  $\Sigma$  es necesario suponer que la misma tiene una estructura particular conocida. La estimación de la matriz  $\Sigma$  no es sencilla, ya que requiere del sustento teórico que permita determinar en qué condiciones la estimación es lo suficientemente adecuada para garantizar la consistencia del estimador  $\tilde{\beta}$ .

Basándose en el artículo de Perrot-Dockès et al. [42], se establecen a continuación dos resultados teóricos en este sentido: la Proposición 2 y el Teorema 2, que son análogos a la Proposición 1 y el Teorema 1 del capítulo anterior pero esta vez para el caso dependiente.

Se usará la Proposición 2 para la demostración del Teorema 2, mientras que el Teorema 2 presenta las condiciones mínimas que debe cumplir la matriz de diseño  $\mathbf{X}$  y las matrices de varianzas y covarianzas de los errores, tanto reales como estimados ( $\Sigma$  y  $\hat{\Sigma}$ ), para que el estimador  $\tilde{\beta}$  sea fuertemente consistente en signo para el caso dependiente. Si bien este teorema es muy similar al Teorema 5 de Perrot-Dockès et al. [42], es importante resaltar que no es exactamente igual, dado que en Perrot-Dockès et al. [42] el problema consiste en realizar la selección de variables en un modelo lineal multivariado (es decir, cuando  $Y$  es una matriz de  $n \times q$ ) donde puede existir dependencia entre las  $p$  variables, pero no se considera el caso donde existe dependencia en los errores. Dicho de otro modo, si se consideran los resultados de Perrot et al ([42]) para  $q = 1$ , se obtiene un modelo lineal con errores independientes.

No obstante, tanto el enunciado como la estrategia que seguimos para la demostración del Teorema 2 siguen la estructura del Teorema 5 de Perrot-Dockès et al. [42], realizando las modificaciones necesarias para ajustarlo a nuestro problema de interés.

**Proposición 2** (Cota inferior a la probabilidad de seleccionar el modelo correcto en el caso de errores dependientes). *Sea  $\tilde{\beta}$  el estimador LASSO del modelo (3.3). Entonces,*

$$P(\tilde{\beta} =_s \beta) \geq P(\tilde{A}_n \cap \tilde{B}_n) \quad (3.4)$$

siendo

$$\tilde{A}_n = \left\{ |(\tilde{\mathbf{C}}_{11}^n)^{-1} \tilde{W}_{A^*}| < \sqrt{n} \left( |\beta_{A^*}| - \frac{\lambda}{2n} |(\tilde{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})| \right) \right\} \quad (3.5)$$

$$\tilde{B}_n = \left\{ |\tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \tilde{W}_{A^*} - \tilde{W}_{A^{*c}}| < \frac{\lambda}{2\sqrt{n}} (\mathbf{1} - |\tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})|) \right\} \quad (3.6)$$

con

$$\tilde{\mathbf{C}}_{11}^n = \frac{1}{n} \tilde{\mathbf{X}}'_{A^*} \tilde{\mathbf{X}}_{A^*} \quad \tilde{\mathbf{C}}_{21}^n = \frac{1}{n} \tilde{\mathbf{X}}'_{A^{*c}} \tilde{\mathbf{X}}_{A^*}$$

$$\tilde{W}_{A^*} = \frac{1}{\sqrt{n}} \tilde{\mathbf{X}}'_{A^*} \tilde{\varepsilon} \quad \tilde{W}_{A^{*c}} = \frac{1}{\sqrt{n}} \tilde{\mathbf{X}}'_{A^{*c}} \tilde{\varepsilon}$$

(las desigualdades en  $\tilde{A}_n$  y  $\tilde{B}_n$  son coordenada a coordenada.)

**Teorema 2** (Condiciones para la consistencia fuerte en signo en el caso de errores dependientes). *Al considerar el modelo (3.3), si existen constantes  $M_1, M_2, M_3, M_4, M_5, M_6$  y  $M_7$  estrictamente positivas y  $c_1, c_2$  con  $0 \leq c_1 \leq c_2 \leq 1/2$  tales que:*

1.  $\frac{1}{n} \mathbf{X}'_j \boldsymbol{\Sigma}^{-1} \mathbf{X}_j \leq M_1 \quad \forall j$
2.  $\rho_{\min}(\frac{1}{n} \mathbf{X}'_{A^*} \boldsymbol{\Sigma}^{-1} \mathbf{X}_{A^*}) \geq M_2$
3.  $p^* = O(n^{c_1/2})$ , donde  $p^*$  representa la cantidad de variables que participan del verdadero modelo.
4.  $n^{(1-c_2)/2} \min_{1 \leq j \leq p^*} (|\beta_j|) \geq M_3$
5. existe un vector  $\delta$  constante y positivo tal que  $|\frac{1}{n} \mathbf{X}'_{A^{*c}} \boldsymbol{\Sigma}^{-1} \mathbf{X}_{A^*} (\frac{1}{n} \mathbf{X}'_{A^*} \boldsymbol{\Sigma}^{-1} \mathbf{X}_{A^*})^{-1} \text{sign}(\beta_{A^*})| \leq \mathbf{1} - \delta$  donde  $\mathbf{1}$  es un vector de unos de dimensión  $p - p^*$  y la desigualdad es coordenada a coordenada.
6.  $\frac{\lambda}{\sqrt{n}} \xrightarrow{n \rightarrow \infty} \infty$  y  $\frac{\lambda}{\sqrt{n}} = o(n^{(c_2 - c_1)/2})$
7.  $\|(\mathbf{X}'\mathbf{X})/n\|_{\infty} \leq M_4$
8.  $\rho_{\max}(\boldsymbol{\Sigma}^{-1}) \leq M_6$
9.  $\rho_{\min}(\boldsymbol{\Sigma}^{-1}) \geq M_7$
10.  $\left\| \boldsymbol{\Sigma}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1} \right\|_{\infty} = O_P(\frac{1}{\sqrt{n}})$  cuando  $n \rightarrow \infty^1$

donde  $\rho(\mathbf{X})$  representa el vector de valores propios de  $\mathbf{X}$ , en particular  $\rho_{\min}(\mathbf{X})$  y  $\rho_{\max}(\mathbf{X})$  representan el menor y mayor valor propio respectivamente.

Entonces se cumple la consistencia fuerte en signo del estimador  $\tilde{\beta}$ , esto es,  $\lim_{n \rightarrow \infty} P(\tilde{\beta} =_s \beta) = 1^2$ .

<sup>1</sup> $X_n = O_P(a_n)$  significa que para todo  $\xi > 0$  existen  $M > 0$  y  $N > 0$  finitos, tal que  $P(|X_n/a_n| > M) < \xi \quad \forall n > N$

<sup>2</sup> $a =_s b$  significa que  $P(\text{sign}(a) = \text{sign}(b)) = 1$

Las hipótesis 3, 4 y 6 del Teorema 2 están presentes en el Teorema 1, mientras que las hipótesis 1, 2 y 5 son análogas a las del Teorema 1 para  $\dot{\mathbf{X}}$  en lugar de  $\mathbf{X}$ .

Las hipótesis 8 y 9 indican que los valores propios de  $\Sigma$  son positivos y acotados y la hipótesis 10 establece que  $\hat{\Sigma}$  estima bien a  $\Sigma$ .

Las demostraciones de la Proposición 2 y del Teorema 2 se encuentran en el Apéndice 1.

## 3.2. Contexto espacial

En el contexto espacial<sup>1</sup>, los datos provienen de un proceso aleatorio  $Y = \{Y_s : s \in D\}$ , donde  $D \subset \mathbb{R}^d$  es el conjunto de ubicaciones espaciales. Cuando  $d = 2$  el proceso se denomina espacial y cada coordenada representa una dimensión en el espacio (latitud y longitud), pero también podría ser de dimensión 3 (longitud, latitud y altitud) o superior.

Asimismo, el conjunto  $D$  puede ser un subespacio continuo o discreto de  $\mathbb{R}^d$ , aleatorio o no. Cuando  $D$  es continuo y el proceso  $Y$  toma valores en el conjunto de los números reales, estamos en el contexto de datos espaciales “geoestadísticos” (ver Giraldo [20]).

Por otro lado, cuando  $D$  es discreto y no aleatorio, estamos en el contexto de datos espaciales “en red”. En este caso, el proceso  $Y$  puede ser real o no, mientras que las ubicaciones pueden ser regulares (o también llamadas de tipo grilla, en inglés “lattice”) o no regulares.

El desarrollo de este capítulo se enmarca en procesos espaciales en red de segundo orden, es decir, aquellos donde  $Y$  tiene varianza finita, con  $d = 2$  y ubicaciones no regulares.

Una característica importante de los procesos espaciales es que las observaciones no son independientes sino que existe una dependencia espacial (o autocorrelación espacial como se definirá más adelante). Para trabajar con este tipo de datos, en principio podrían aplicarse las mismas técnicas que se utilizan para datos independientes, pero esta estrategia tiene un inconveniente: los modelos tradicionales no están pensados para recoger la estructura espacial

---

<sup>1</sup>Una de las principales referencias en este tema es Cressie [14], a nivel nacional la publicación de Riaño [45] refiere a este tema.

subyacente en los datos, por lo que la misma quedará contenida en los residuos, los cuales no verificarán en general la hipótesis de independencia que sustenta dichos métodos. Por este motivo se han desarrollado modelos para trabajar específicamente con este tipo de problemas<sup>1</sup>, como los modelos de regresión espacial que serán utilizados más adelante.

Otro aspecto que se ve afectado en el contexto espacial es la selección de variables, en particular al utilizar el método LASSO en la regresión espacial, ya que se realiza en simultáneo la estimación y selección de parámetros de la regresión. Esta sección tiene como objetivo establecer las condiciones para poder utilizar el Teorema 2 en el contexto espacial de datos en red, con una estructura específica para la matriz de covarianzas de los errores. Se propone un procedimiento para realizar la selección de variables mediante LASSO.

### 3.2.1. Autocorrelación espacial

Como se mencionó previamente, en el contexto de datos espaciales en red se utiliza el concepto de *autocorrelación espacial* para referirse a la correlación que existe entre distintas observaciones del conjunto  $D$  al considerar una misma variable (recordar que  $D$  es discreto en esta situación)<sup>2</sup>. Este concepto está asociado al de *vecindad*, ya que la autocorrelación no se mide entre todas las observaciones, sino entre aquellas que están más próximas de acuerdo a algún criterio predefinido. Entre los criterios de vecindad más utilizados se encuentran:

- la vecindad *triangular* que conforma un grafo a partir de la triangulación de las ubicaciones,
- *vecinos más cercanos* que asume como vecinos a los  $k$  vecinos más cercanos (donde  $k$  es un parámetro previamente elegido) y
- la vecindad *basada en distancias* que calcula la distancia euclídea entre los puntos y considera como vecinos a aquellas observaciones que se encuentren dentro de cierto rango predefinido.

Se considera un grafo de influencia  $\mathcal{R}$ , en principio dirigido, donde  $(i, j) \in$

---

<sup>1</sup>Algunos ejemplos: Wang and Zhu [53], Cai and Maiti [11], Hoeting et al. [28], Huang and Chen [30], Huang et al. [31], Chu et al. [13], Fu et al. [18], Reyes et al. [44], Zhu and Y.Liu [58], Nandy [37], Nandy et al. [38] y su material complementario Nandy et al. [39]

<sup>2</sup>para profundizar en este concepto consultar Siabato and Guzmán-Manrique [47] y Goodchild [23]

$\mathcal{R}$  implica que  $j$  influye sobre  $i$ , con  $j \neq i$ . También se define una *matriz de pesos*  $\mathbf{W}_{\mathcal{R}} = \{w_{ij}, (i, j) \in \mathcal{R}\}$  que cuantifica las influencias entre observaciones, tal que  $w_{ii} = 0$  y  $w_{ij} = 0$  si  $(i, j) \notin \mathcal{R}$ . La elección de  $\mathbf{W}_{\mathcal{R}}$  es muy importante y depende del problema en cuestión.

La matriz  $\mathbf{W}_{\mathcal{R}}$  se denomina *matriz de contigüidad* del grafo  $\mathcal{R}$  si  $w_{ij} = 1$  cuando  $(i, j) \in \mathcal{R}$  y 0 en otro caso. A su vez cuando los pesos están normalizados (sus filas suman uno) la matriz  $\mathbf{W}_{\mathcal{R}}$  se denomina *matriz de contigüidad normalizada*.

En la literatura se destacan dos índices para medir la dependencia espacial global de una red: el índice de Moran y el índice de Geary. A continuación se define cada uno de estos índices, de acuerdo a Gaetan and Guyon [19].

**Definición 5** (Índice de Moran). *Se considera una muestra de datos espaciales con media y varianza constantes, de segundo orden  $Y = \{Y_1, \dots, Y_n\}$  y una matriz de pesos conocidos  $\mathbf{W}_{\mathcal{R}}$  de dimensión  $n \times n$  con elementos  $w_{ij}$ . La media y varianzas desconocidas se estiman por la media y varianza muestrales:  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$  y  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$  respectivamente. En este contexto el índice de Moran se define como:*

$$I^M = \frac{n \sum_{i,j=1}^n w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i,j=1}^n w_{ij} \sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (3.7)$$

Adicionalmente se definen:

- $s_0 = \sum_{i,j=1}^n w_{ij}$
- $s_1 = \sum_{i,j=1}^n (w_{ij}^2 + w_{ij}w_{ji})$
- $M = \sum_{i,j=1}^n w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})$ .
- $I_M = \frac{M}{\sqrt{Var(M)}}$ .

El índice de Moran calcula la correlación espacial entre observaciones. Es fácil probar que bajo la hipótesis de independencia espacial  $E(M) = 0$  y  $Var(M) = \sum_{i,j=1}^n (w_{ij}^2 + w_{ij}w_{ji})(\sigma^2)^2$ . A su vez, se obtiene directamente que

$E(I_M) = 0$  y  $Var(I_M) = 1$ , por lo que el índice de Moran puede reescribirse como:

$$I^M = \frac{\sqrt{s_1}M}{s_0\sqrt{\widehat{Var}(M)}} = \frac{\sqrt{s_1}}{s_0}\hat{I}_M \quad (3.8)$$

donde  $\widehat{Var}(M) = \sum_{i,j=1}^n (w_{ij}^2 + w_{ij}w_{ji})(\hat{\sigma}^2)^2$  y  $\hat{I}_M = \frac{M}{\sqrt{\widehat{Var}(M)}}$ . Bajo la hipótesis de independencia espacial se tiene:

$$\begin{aligned} E(I^M) &= o(1) \\ Var(I^M) &= \frac{s_1}{s_0^2}(1 + o(1)) \end{aligned}$$

Si bien el índice de Moran se parece a un coeficiente de correlación,  $I^M$  puede tomar valores fuera del intervalo  $[-1, 1]$ . Los valores grandes de  $I^M$  indican agregación o cooperación espacial (los valores se parecen a los de sus vecinos), mientras que los valores pequeños indican repulsión o competencia espacial, los valores cercanos a cero indican independencia espacial.

Por otro lado, el índice de Geary mide la dependencia espacial en el mismo sentido que el variograma<sup>1</sup> para el caso de geoestadística.

**Definición 6** (Índice de Geary). *Considerando la misma muestra  $Y$  y matriz de pesos  $\mathbf{W}_{\mathcal{R}}$  de la definición anterior, el índice de Geary se define como:*

$$I^G = \frac{(n-1) \sum_{i,j=1}^n w_{ij}(Y_i - Y_j)^2}{2 \sum_{i,j=1}^n w_{ij} \sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (3.9)$$

Asintóticamente se tiene que  $\sqrt{\frac{s_0}{2s_1+s_2}}(I^G - 1) \sim N(0, 1)$ , donde  $s_2 = \sum_{i=1}^n \left( \sum_{j=1}^n w_{ij} + \sum_{j=1}^n w_{ji} \right)$ .

Valores pequeños de  $I^G$  indican la presencia de autocorrelación espacial. Este índice es sensible a grandes diferencias entre puntos vecinos, así como  $I^M$  es sensible a valores extremos de  $Y$ .

<sup>1</sup>El variograma para un proceso  $Y$  se define como  $Variog(h) = \frac{1}{2}Var(Y_{s+h} - Y_s)$   $s \in D$

Estos índices son utilizados en pruebas de hipótesis para determinar si un conjunto de datos presenta dependencia o autocorrelación espacial. En el presente trabajo serán utilizados por un lado para verificar si la variable a explicar proviene de un proceso espacial, y por otro lado para corroborar que el procedimiento empleado efectivamente eliminó la dependencia espacial.

### 3.2.2. Modelos para procesos espaciales

A continuación se definen los modelos CAR y SAR para un proceso espacial  $\{Z_i, i \in D\}$  con  $D = \{1, \dots, n\}$ .

**Definición 7** (Modelo condicional autorregresivo: CAR). *Este modelo puede formularse a través de la distribución condicional de  $Z_i$ :*

$$E(Z_i | Z_j : j \neq i) = \sum_{j=1}^n c_{ij} Z_j \quad \text{Var}(Z_i | Z_j : j \neq i) = \sigma_i^2$$

Para que la distribución conjunta de  $Z = (Z_1, \dots, Z_n)$  sea válida es necesario que  $c_{ii} = 0$ ,  $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$  y  $c_{ij} = 0$  si  $j \notin \mathcal{N}(i)$ , donde  $\mathcal{N}(i)$  representa el conjunto de índices de los vecinos de  $i$  de acuerdo a la estructura de vecindad considerada.

La matriz de varianzas y covarianzas de  $Z$  es igual a

$$\Sigma_{\text{CAR}} = (\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1} \mathbf{V}$$

donde  $\mathbf{I}$  es la matriz identidad,  $\mathbf{C}_{\mathcal{R}} = [c_{ij}]_{i,j=1}^n$ ,  $\mathbf{I} - \mathbf{C}_{\mathcal{R}}$  es no singular,  $\mathbf{V}$  es una matriz diagonal con  $V_{ii} = \sigma_i^2$  y  $\Sigma_{\text{CAR}}$  es simétrica y definida positiva.

**Definición 8** (Modelo simultáneo autorregresivo: SAR). *El modelo puede formularse como:*

$$Z = \mathbf{C}_{\mathcal{R}} Z + \nu$$

donde  $\nu \sim N(\mathbf{0}, \mathbf{V})$  representa un vector de errores independientes, con  $\mathbf{V} = \text{diag}(\sigma_i^2)$ .

La matriz de varianzas y covarianzas de  $Z$  es igual a

$$\Sigma_{\text{SAR}} = (\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1} \mathbf{V} (\mathbf{I} - \mathbf{C}'_{\mathcal{R}})^{-1}$$

donde las matrices  $\mathbf{C}_{\mathcal{R}}$ ,  $\mathbf{V}$  e  $\mathbf{I} - \mathbf{C}_{\mathcal{R}}$  son iguales al modelo CAR. La matriz  $\Sigma_{\text{SAR}}$  así definida siempre es simétrica y definida positiva<sup>1</sup>.

En Zhu et al. [57], se sugiere utilizar la siguiente estructura para la matriz  $\mathbf{C}_{\mathcal{R}}$ :

$$\mathbf{C}_{\mathcal{R}} = \sum_{k=1}^K \theta_k \mathbf{W}_{\mathcal{R}}^k$$

donde  $\mathbf{W}_{\mathcal{R}}^k$  representa la matriz de pesos asociada a la vecindad de orden  $k$ : las observaciones  $i$  y  $j$  tienen vecindad de orden  $k$  si existen  $k-1$  observaciones  $a_1, \dots, a_{k-1}$  tales que  $a_1 \in \mathcal{N}(i), a_2 \in \mathcal{N}(a_1), a_3 \in \mathcal{N}(a_2), \dots, a_{k-1} \in \mathcal{N}(j)$  y, si se considera un grafo donde los nodos están representados por las observaciones y los arcos están representados por las relaciones de vecindad, el camino  $i \rightarrow a_1 \rightarrow a_2 \rightarrow \dots \rightarrow a_{k-1} \rightarrow j$  es el más corto para llegar de  $i$  a  $j$ .

Se considera el caso en que el orden de vecindad es igual a uno ( $K = 1$ ) y  $\sigma_i^2 = \sigma^2 \forall i \in 1, \dots, n$ . En este contexto, se tiene:

$$\mathbf{V} = \begin{bmatrix} \sigma^2 & 0 & \dots & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \dots & \dots & 0 & \sigma^2 \end{bmatrix} \quad \mathbf{W}_{\mathcal{R}} = \begin{bmatrix} w_{11} & \dots & \dots & \dots & w_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & \dots & \dots & \dots & w_{nn} \end{bmatrix} \quad \mathbf{C}_{\mathcal{R}} = \begin{bmatrix} \theta w_{11} & \dots & \dots & \dots & \theta w_{1n} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta w_{n1} & \dots & \dots & \dots & \theta w_{nn} \end{bmatrix}$$

donde los pesos  $w_{ij}$  cumplen:

- $w_{ii} = 0 \forall i \in 1, \dots, n$
- $w_{ij} = w_{ji} \forall i, j \in 1, \dots, n$
- $w_{ij} = 0$  si  $j \notin \mathcal{N}(i)$

Por lo tanto, la matriz  $\mathbf{W}_{\mathcal{R}}$ , y también la matriz  $\mathbf{C}_{\mathcal{R}}$ , son simétricas y con ceros en la diagonal principal.

Calcular las matrices  $\Sigma_{\text{CAR}}$  y  $\Sigma_{\text{SAR}}$  en función de sus parámetros  $\sigma^2$ ,  $\theta$  y

<sup>1</sup>Si bien no es necesario que  $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$ , se mantiene esta condición para el modelo SAR



$\{w_{ij} : i, j \in 1, \dots, n\}$  es complejo, ya que implica conocer la inversa de  $\mathbf{I} - \mathbf{C}_{\mathcal{R}}$ . Sin embargo, determinar los elementos de  $\Sigma_{\text{CAR}}^{-1}$  y  $\Sigma_{\text{SAR}}^{-1}$  es muy sencillo, ya que:

$$\begin{aligned}\Sigma_{\text{CAR}}^{-1} &= \mathbf{V}^{-1}(\mathbf{I} - \mathbf{C}_{\mathcal{R}}) \\ \Sigma_{\text{SAR}}^{-1} &= (\mathbf{I} - \mathbf{C}_{\mathcal{R}})\mathbf{V}^{-1}(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\end{aligned}$$

y la única inversa que se necesita conocer es la de  $\mathbf{V}$ , y ya que es una matriz diagonal su inversa es inmediata.

Se tiene:

$$\mathbf{I} - \mathbf{C}_{\mathcal{R}} = \begin{bmatrix} 1 - \theta \overbrace{w_{11}}^{=0} & -\theta w_{12} & \cdots & -\theta w_{1n} \\ -\theta w_{21} & 1 - \theta \overbrace{w_{22}}^{=0} & \cdots & -\theta w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta w_{n1} & \cdots & -\theta w_{n,n-1} & 1 - \theta \overbrace{w_{nn}}^{=0} \end{bmatrix}$$

por lo que:

$$\Sigma_{\text{CAR}}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & -\frac{\theta w_{12}}{\sigma^2} & \cdots & -\frac{\theta w_{1n}}{\sigma^2} \\ -\frac{\theta w_{21}}{\sigma^2} & \frac{1}{\sigma^2} & \cdots & -\frac{\theta w_{2n}}{\sigma^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\theta w_{n1}}{\sigma^2} & \cdots & -\frac{\theta w_{n,n-1}}{\sigma^2} & \frac{1}{\sigma^2} \end{bmatrix}$$

y

$$\Sigma_{\text{SAR}}^{-1} = \frac{\theta^2}{\sigma^2} \begin{bmatrix} \sum_{k=1}^n w_{1k}^2 + \frac{1}{\theta^2} & \sum_{k=1}^n w_{1k}w_{k2} - \frac{2}{\theta}w_{12} & \cdots & \sum_{k=1}^n w_{1k}w_{kn} - \frac{2}{\theta}w_{1n} \\ \sum_{k=1}^n w_{2k}w_{k1} - \frac{2}{\theta}w_{21} & \sum_{k=1}^n w_{2k}^2 + \frac{1}{\theta^2} & \cdots & \sum_{k=1}^n w_{2k}w_{kn} - \frac{2}{\theta}w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n w_{nk}w_{k1} - \frac{2}{\theta}w_{n1} & \cdots & \cdots & \sum_{k=1}^n w_{nk}^2 + \frac{1}{\theta^2} \end{bmatrix}$$

### 3.2.3. Modelos con errores espaciales

Se considera el modelo 3.1:

$$Y = \mathbf{X}\beta + \varepsilon$$

donde  $\varepsilon$  sigue un modelo CAR o SAR de orden uno (denominados CAR(1) y SAR(1) respectivamente).

A continuación se estudia en qué situaciones las matrices de covarianzas  $\Sigma_{\text{CAR}}$  y  $\Sigma_{\text{SAR}}$  verifican las hipótesis del Teorema 2, en particular las hipótesis 8, 9 y 10.

Para verificar estas hipótesis, se utiliza la definición de norma infinito<sup>1</sup> y sus propiedades, el Teorema 5.6.9 y el Corolario 5.6.16 (Horn and Johnson [29]).

La **hipótesis 8** establece que  $\rho_{\max}(\Sigma^{-1}) \leq M_6$ , donde  $M_6$  es una constante estrictamente positiva.

Para verificar este punto, se acota  $\rho_{\max}(\Sigma^{-1})$  utilizando el Teorema 5.6.9 de Horn and Johnson [29], considerando que  $\Sigma$  es simétrica y definida positiva, por lo que también lo es  $\Sigma^{-1}$ .

$$\rho_{\max}(\Sigma^{-1}) \leq \gamma(\Sigma^{-1}) \leq \|\|\Sigma^{-1}\|\|_{\infty} \quad (3.10)$$

donde  $\gamma(\mathbf{X})$  representa el radio espectral de  $\mathbf{X}$ :  $\gamma(\mathbf{X}) = \max_{1 \leq i \leq n} (|\sqrt{\rho_i(\mathbf{X})}|)$ , con  $\rho_i(\mathbf{X})$  igual al  $i$ -ésimo valor propio de  $\mathbf{X}$ . Luego se descompone  $\Sigma^{-1}$  en un producto de matrices, de acuerdo a la definición de  $\Sigma$ . Para el caso de una matriz  $\Sigma$  proveniente de un modelo CAR:

$$\|\|\Sigma_{\text{CAR}}^{-1}\|\|_{\infty} = \|\|\mathbf{V}^{-1}(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\|\|_{\infty} \leq \|\|\mathbf{V}^{-1}\|\|_{\infty} \|\|\mathbf{I} - \mathbf{C}_{\mathcal{R}}\|\|_{\infty}$$

donde la desigualdad se debe a la propiedad de submultiplicatividad de las normas matriciales. Se tiene:

$$\|\|\mathbf{V}^{-1}\|\|_{\infty} = \left\| \begin{array}{ccc} 1/\sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma^2 \end{array} \right\|_{\infty} = \frac{1}{\sigma^2}$$

---

<sup>1</sup>Se define como:  $\|\|\mathbf{X}\|\|_{\infty} = \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |X_{ij}| \right)$

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\|_{\infty} &= \left\| \begin{array}{cccc} 1 & -\theta w_{12} & \cdots & -\theta w_{1n} \\ -\theta w_{21} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -\theta w_{n1} & \cdots & \cdots & 1 \end{array} \right\|_{\infty} = \max_{1 \leq i \leq n} \left( 1 + \sum_{j=1}^n |\theta w_{ij}| \right) \\
&= 1 + |\theta| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)
\end{aligned}$$

Por lo que:

$$\rho_{max}(\Sigma_{\text{CAR}}^{-1}) \leq \frac{1}{\sigma^2} \left( 1 + |\theta| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right) \right)$$

y se verifica la hipótesis 8 si  $\max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)$  no depende de  $n$ .

Para el caso de una matriz SAR, el razonamiento es análogo:

$$\begin{aligned}
\|\Sigma_{\text{SAR}}^{-1}\|_{\infty} &= \|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\mathbf{V}^{-1}(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\|_{\infty} \\
&\leq \|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\|_{\infty} \|\mathbf{V}^{-1}\|_{\infty} \|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\|_{\infty} \\
&= \frac{1}{\sigma^2} \left( 1 + |\theta| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right) \right)^2
\end{aligned}$$

y nuevamente se obtiene como condición que  $\max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)$  no dependa de  $n$ .

La **hipótesis 9** indica que  $\rho_{min}(\Sigma^{-1}) \geq M_7$ , con  $M_7$  constante y mayor a cero.

Para verificar este punto, como  $\rho_{min}(\Sigma^{-1}) = \frac{1}{\rho_{max}(\Sigma)}$ , verificar la hipótesis 9 equivale a demostrar que existe  $M_7$  tal que  $\rho_{max}(\Sigma) \leq M_7^{-1}$ .

Para el caso de la matriz CAR, utilizando la propiedad de submultiplicatividad de normas matriciales, se tiene:

$$\|\|\Sigma_{\text{CAR}}\|\|_{\infty} = \|\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\mathbf{V}\|\|_{\infty} \leq \|\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\|\|_{\infty} \|\|\mathbf{V}\|\|_{\infty}$$

Se deduce directamente que  $\|\|\mathbf{V}\|\|_{\infty} = \sigma^2$ . Para determinar  $\|\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\|\|_{\infty}$  se utiliza el siguiente resultado que surge del Corolario 5.6.16 de Horn and Johnson [29]:

$$\|\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\|\|_{\infty} \leq \frac{1}{1 - \|\|\mathbf{C}_{\mathcal{R}}\|\|_{\infty}} \text{ sí } \|\|\mathbf{C}_{\mathcal{R}}\|\|_{\infty} < 1$$

Como  $\|\|\mathbf{C}_{\mathcal{R}}\|\|_{\infty} = \|\|\theta\mathbf{W}_{\mathcal{R}}\|\|_{\infty} = |\theta|\|\|\mathbf{W}_{\mathcal{R}}\|\|_{\infty}$ , con  $\|\|\mathbf{W}_{\mathcal{R}}\|\|_{\infty} = \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)$ , entonces:

$$\|\|\Sigma_{\text{CAR}}\|\|_{\infty} \leq \frac{\sigma^2}{1 - |\theta| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)} \text{ y } \rho_{\min}(\Sigma_{\text{CAR}}^{-1}) \geq \frac{1 - |\theta| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)}{\sigma^2}$$

Para que se verifique la hipótesis 9 es suficiente que  $\max_{1 \leq i \leq n} \sum_{j=1}^n |w_{ij}|$  no dependa de  $n$  y que  $|\theta| \max_{1 \leq i \leq n} \sum_{j=1}^n |w_{ij}| < 1$ . Para el caso de la matriz SAR,

$$\begin{aligned} \|\|\Sigma_{\text{SAR}}\|\|_{\infty} &= \|\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\mathbf{V}(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\|\|_{\infty} \\ &\leq \frac{\|\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\|\|_{\infty} \|\|\mathbf{V}\|\|_{\infty} \|\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})^{-1}\|\|_{\infty}}{\sigma^2} \\ &\leq \frac{1}{\left( 1 - |\theta| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right) \right)^2} \end{aligned}$$

por lo que

$$\rho_{\min}(\Sigma_{\text{SAR}}^{-1}) \geq \frac{\left( 1 - |\theta| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right) \right)^2}{\sigma^2}$$

Para que se verifique la hipótesis 9 en matrices de tipo SAR, se deben cumplir las mismas condiciones sobre los pesos  $w_{ij}$  que en las matrices de tipo CAR.

Finalmente, en la **hipótesis 10** se especifica que  $\|\|\Sigma^{-1} - \hat{\Sigma}^{-1}\|\|_{\infty} = O_P\left(\frac{1}{\sqrt{n}}\right)$  cuando  $n \rightarrow \infty$ . Para verificar esta hipótesis, se supone que  $\hat{\Sigma}^{-1}$  se construye a partir de  $\hat{\mathbf{C}}_{\mathcal{R}} = \hat{\theta}\mathbf{W}_{\mathcal{R}}$  y  $\hat{\mathbf{V}} = \text{diag}(\hat{\sigma}^2)$ . Cabe recordar que los pesos  $w_{ij}$  son conocidos. En el caso de la matriz CAR se tiene:

$$\begin{aligned}
\Sigma_{\text{CAR}}^{-1} - \hat{\Sigma}_{\text{CAR}}^{-1} &= \begin{bmatrix} \frac{1}{\sigma^2} & -\frac{\theta w_{12}}{\sigma^2} & \cdots & -\frac{\theta w_{1n}}{\sigma^2} \\ -\frac{\theta w_{21}}{\sigma^2} & \frac{1}{\sigma^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\theta w_{n1}}{\sigma^2} & \cdots & \cdots & \frac{1}{\sigma^2} \end{bmatrix} - \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & -\frac{\hat{\theta} w_{12}}{\hat{\sigma}^2} & \cdots & -\frac{\hat{\theta} w_{1n}}{\hat{\sigma}^2} \\ -\frac{\hat{\theta} w_{21}}{\hat{\sigma}^2} & \frac{1}{\hat{\sigma}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\hat{\theta} w_{n1}}{\hat{\sigma}^2} & \cdots & \cdots & \frac{1}{\hat{\sigma}^2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} & w_{12} \left( \frac{\hat{\theta}}{\hat{\sigma}^2} - \frac{\theta}{\sigma^2} \right) & \cdots & w_{1n} \left( \frac{\hat{\theta}}{\hat{\sigma}^2} - \frac{\theta}{\sigma^2} \right) \\ w_{21} \left( \frac{\hat{\theta}}{\hat{\sigma}^2} - \frac{\theta}{\sigma^2} \right) & \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} \left( \frac{\hat{\theta}}{\hat{\sigma}^2} - \frac{\theta}{\sigma^2} \right) & \cdots & \cdots & \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} \end{bmatrix}
\end{aligned}$$

Entonces, utilizando la definición de norma infinito, se llega al siguiente resultado:

$$\begin{aligned}
\left\| \Sigma_{\text{CAR}}^{-1} - \hat{\Sigma}_{\text{CAR}}^{-1} \right\|_{\infty} &= \max_{1 \leq i \leq n} \left( \left| \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} \right| + \sum_{j=1}^n \left| w_{ij} \left( \frac{\hat{\theta}}{\hat{\sigma}^2} - \frac{\theta}{\sigma^2} \right) \right| \right) \\
&= \left| \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} \right| + \left| \frac{\hat{\theta}}{\hat{\sigma}^2} - \frac{\theta}{\sigma^2} \right| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)
\end{aligned}$$

Bajo el supuesto de que  $\sqrt{n}(\hat{\theta} - \theta) = O_P(1)$  y  $\sqrt{n} \left( \frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2} \right) = O_P(1)$ , y que  $\max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)$  está acotado, se verifica la hipótesis 10 para el caso CAR.

En el caso SAR se tiene que  $\Sigma_{\text{SAR}}^{-1} - \hat{\Sigma}_{\text{SAR}}^{-1}$  es igual a:

$$\begin{bmatrix} \sum_{k=1}^n w_{1k}^2 \left( \frac{\theta^2}{\sigma^2} - \frac{\theta^2}{\hat{\sigma}^2} \right) + \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} & \sum_{k=1}^n w_{1k} w_{k2} \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + w_{12} \left( \frac{2\theta}{\sigma^2} - \frac{2\hat{\theta}}{\hat{\sigma}^2} \right) & \cdots & \sum_{k=1}^n w_{1k} w_{kn} \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + w_{1n} \left( \frac{2\theta}{\sigma^2} - \frac{2\hat{\theta}}{\hat{\sigma}^2} \right) \\ \sum_{k=1}^n w_{2k} w_{k1} \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + w_{21} \left( \frac{2\theta}{\sigma^2} - \frac{2\hat{\theta}}{\hat{\sigma}^2} \right) & \sum_{k=1}^n w_{2k}^2 \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n w_{nk} w_{k1} \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + w_{n1} \left( \frac{2\theta}{\sigma^2} - \frac{2\hat{\theta}}{\hat{\sigma}^2} \right) & \cdots & \cdots & \sum_{k=1}^n w_{nk}^2 \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} \end{bmatrix}$$

Por lo que se obtiene que la norma infinito de esta matriz es:

$$\begin{aligned}
\left\| \Sigma_{\text{SAR}}^{-1} - \hat{\Sigma}_{\text{SAR}}^{-1} \right\|_{\infty} &= \max_{1 \leq i \leq n} \left( \left| \sum_{k=1}^n w_{ik}^2 \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} \right| \right. \\
&\quad \left. + \sum_{\substack{j=1 \\ j \neq i}}^n \left| \sum_{k=1}^n w_{ik} w_{kj} \left( \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right) + w_{ij} \left( \frac{2\hat{\theta}}{\hat{\sigma}^2} - \frac{2\theta}{\sigma^2} \right) \right| \right) \\
&\leq \left| \frac{1}{\sigma^2} - \frac{1}{\hat{\sigma}^2} \right| + \left| \frac{\theta^2}{\sigma^2} - \frac{\hat{\theta}^2}{\hat{\sigma}^2} \right| \max_{1 \leq i \leq n} \left( \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{k=1}^n |w_{ik} w_{kj}| \right) + \\
&\quad \left| \frac{2\hat{\theta}}{\hat{\sigma}^2} - \frac{2\theta}{\sigma^2} \right| \max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)
\end{aligned}$$

Nuevamente bajo el supuesto de que  $\sqrt{n}(\hat{\theta} - \theta) = O_P(1)$  y  $\sqrt{n}(\frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2}) = O_P(1)$ , y que tanto  $\max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right)$  como  $\max_{1 \leq i \leq n} \left( \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{k=1}^n |w_{ik} w_{kj}| \right)$  están acotados, se verifica la hipótesis 10 para el caso SAR.

En el siguiente cuadro se resumen las condiciones que deben imponerse sobre los pesos  $w_{ij}$  para garantizar que se verifiquen las hipótesis 8 a 10 del Teorema 2 en matrices de tipo CAR y SAR con  $K = 1$ , donde  $c$  representa una constante mayor a cero.

Identificador	Condición	H.8	H.9	H.10	Aplica a
1	$\max_{1 \leq i \leq n} \left( \sum_{j=1}^n  w_{ij}  \right) < c$	sí	sí	sí	CAR-SAR
2	$ \theta  \max_{1 \leq i \leq n} \left( \sum_{j=1}^n  w_{ij}  \right) < 1$	no	sí	no	CAR-SAR
3	$\max_{1 \leq i \leq n} \left( \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{k=1}^n  w_{ik} w_{kj}  \right) < c$	no	no	sí	sólo SAR

**Tabla 3.1:** Condiciones sobre los pesos  $w_{ij}$  de la matriz  $\mathbf{W}_{\mathcal{D}}$ .

Junto con los pesos  $w_{ij}$  a utilizar se debe definir la estructura de vecindad a emplear. Al principio de esta sección se mencionó que existen al menos tres tipos de estructuras de vecindad: vecindad triangular,  $k$  vecinos más cercanos y vecindad basada en distancias. También se debe considerar cómo es la distribución espacial de los datos, ya que no es lo mismo que las ubicaciones estén equiespaciadas o que no sigan un patrón. El primer caso se trata de ubicaciones regulares o de tipo grilla, mientras que el segundo caso se denomina no regular. A los efectos de este trabajo se consideran ubicaciones no regulares. Esto permite acotar el tipo de estructura de vecindad a utilizar, ya que no todas verifican las condiciones del Cuadro 3.1 en este contexto.

La estructura de vecindad basada en distancias tiene algunas limitaciones, como por ejemplo la necesidad de definir apropiadamente el rango (distancia mínima - distancia máxima) para que no queden observaciones desconectadas del grafo, a su vez cuando existe cierta variabilidad en las distancias entre observaciones puede resultar muy difícil determinar la distancia apropiada que define la vecindad (como sucede en el problema real que se verá más adelante). La estructura de vecindad basada en  $k$  vecinos más cercanos tiene el inconveniente, cuando se trata de ubicaciones irregulares, de que no garantiza la simetría en la vecindad, es decir si  $j$  se encuentra entre los  $k$  vecinos más cercanos de  $i$ , puede suceder que  $i$  no se encuentre entre los  $k$  vecinos más cercanos de  $j$ .

Finalmente, la estructura de vecindad triangular no presenta las limitaciones de las otras estructuras mencionadas anteriormente en el caso de ubicaciones irregulares, e incluso definiendo apropiadamente los pesos, se verifican las condiciones del Cuadro 3.1 para errores de tipo SAR o CAR.

Ahora bien, una vez identificada la estructura de vecindad apropiada, deben elegirse los pesos tomando en cuenta las definiciones y condiciones detalladas en esta sección.

Una estructura de pesos binarios, es decir con  $w_{ij} = 1$  si  $i$  y  $j$  son vecinos y  $w_{ij} = 0$  en caso contrario, no verifica las condiciones anteriores, ya que el máximo de la suma de los pesos puede aumentar con  $n$ .

Una estructura normalizada por fila, es decir donde  $w_{ij} = \frac{1}{|\mathcal{N}(i)|}$  si  $j \in \mathcal{N}(i)$  y  $w_{ij} = 0$  en otro caso, no verifica la condición de que  $w_{ij} = w_{ji} \forall i, j = 1, \dots, n$ . A continuación se presenta una estructura de pesos que sí verifica las 3 condiciones del Cuadro 3.1 y a su vez la definición dada al principio de esta sección.

**Definición 9** (Matriz  $\mathbf{W}_{\mathcal{R}}$  que verifica las hipótesis del Teorema 2 para modelos CAR o SAR con estructura de vecindad triangular). *Para cada par de observaciones  $i \in \{1, \dots, n\}$  y  $j \in \{1, \dots, n\}$ , con  $i \neq j$ , se definen los pesos de la siguiente manera:*

$$w_{ij} = \begin{cases} \min\left(\frac{1}{|\mathcal{N}(i)|}, \frac{1}{|\mathcal{N}(j)|}\right) & \text{si } j \in \mathcal{N}(i) \text{ (equiv. } i \in \mathcal{N}(j)) \\ 0 & \text{en otro caso} \end{cases} \quad (3.11)$$

A continuación se demostrará que se verifican las condiciones del Cuadro 3.1 para los pesos definidos anteriormente.

Primero se verificará la **condición 1**, partiendo de:

$$\max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right) = \max_{1 \leq i \leq n} \left( \sum_{j=1}^n \frac{1}{\max(|\mathcal{N}(i)|, |\mathcal{N}(j)|)} \right)$$

existen 3 situaciones posibles:

1.  $|\mathcal{N}(i)| \geq |\mathcal{N}(j)| \forall j \in \mathcal{N}(i)$
2.  $|\mathcal{N}(i)| < |\mathcal{N}(j)| \forall j \in \mathcal{N}(i)$
3. hay  $|\mathcal{N}^+(i)|$  vecinos de  $i$  con  $|\mathcal{N}(j)| \leq |\mathcal{N}(i)|$  y  $|\mathcal{N}^-(i)|$  vecinos de  $i$  con  $|\mathcal{N}(j)| > |\mathcal{N}(i)|$ , donde  $|\mathcal{N}^+(i)| + |\mathcal{N}^-(i)| = |\mathcal{N}(i)|$  y  $1 \leq |\mathcal{N}^+(i)| \leq |\mathcal{N}(i)| - 1$

Suponiendo que estamos ante la situación 1, se tiene:

$$\sum_{j=1}^n |w_{ij}| = \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(i)|} = |\mathcal{N}(i)| \frac{1}{|\mathcal{N}(i)|} = 1$$

Ahora suponiendo que estamos ante la situación 2:

$$\sum_{j=1}^n |w_{ij}| = \sum_{j \in \mathcal{N}(i)} \frac{1}{|\mathcal{N}(j)|} \leq |\mathcal{N}(i)| \frac{1}{\min_{j \in \mathcal{N}(i)} |\mathcal{N}(j)|} < 1$$

Por otro lado, suponiendo la situación 3:



$$\begin{aligned}
\sum_{j=1}^n |w_{ij}| &= \sum_{j \in \mathcal{N}^+(i)} \frac{1}{|\mathcal{N}(i)|} + \sum_{j \in \mathcal{N}^-(i)} \frac{1}{|\mathcal{N}(j)|} \leq \frac{|\mathcal{N}^+(i)|}{|\mathcal{N}(i)|} + \frac{|\mathcal{N}^-(i)|}{\min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)} \\
&= \frac{(|\mathcal{N}(i)| - |\mathcal{N}^-(i)|) \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|) + |\mathcal{N}^-(i)| |\mathcal{N}(i)|}{|\mathcal{N}(i)| \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)} \\
&= \frac{|\mathcal{N}(i)| \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)}{|\mathcal{N}(i)| \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)} - \frac{|\mathcal{N}^-(i)| \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)}{|\mathcal{N}(i)| \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)} \\
&\quad + \frac{|\mathcal{N}^-(i)| |\mathcal{N}(i)|}{|\mathcal{N}(i)| \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)} \\
&= 1 - \frac{|\mathcal{N}^-(i)|}{|\mathcal{N}(i)|} + \frac{|\mathcal{N}^-(i)|}{\min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)} < 1
\end{aligned}$$

donde la última desigualdad se debe a que  $|\mathcal{N}(i)| < \min_{j \in \mathcal{N}^-(i)} (|\mathcal{N}(j)|)$ .

Se obtiene así que:

$$\max_{1 \leq i \leq n} \left( \sum_{j=1}^n |w_{ij}| \right) = \max_{1 \leq i \leq n} \left( \sum_{j=1}^n \frac{1}{\max(|\mathcal{N}(i)|, |\mathcal{N}(j)|)} \right) \leq 1$$

Para que se verifique la **condición 2** basta imponer que  $|\theta| < 1$ .

Ahora se verificará la **condición 3** del Cuadro 3.1. Esta condición involucra a los vecinos de segundo orden de  $i$ , es decir, a los vecinos de los vecinos de  $i$ . Los vecinos de segundo orden de  $i$  se denotan  $\mathcal{N}_2(i)$ , cada observación  $i$  tiene  $|\mathcal{N}(i)|$  vecinos y cada uno de estos a su vez tiene una cantidad a priori desconocida de vecinos, que varía entre 1 y  $n - 1$ . El conjunto  $\mathcal{N}_2(i)$  está formado por los vecinos de primer orden de los vecinos de  $i$ , sin considerarse a sí mismo (es decir, sin considerar a la observación  $i$ ) ni a sus vecinos de primer orden.

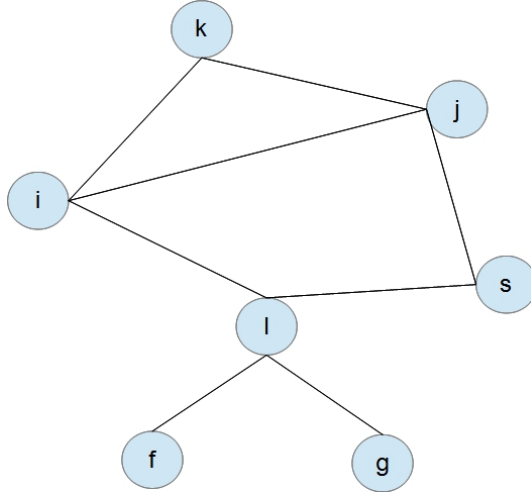
A su vez, se define la cantidad  $|\mathcal{N}'_2(i)| = \sum_{j \in \mathcal{N}(i)} |\mathcal{N}(j) - \{i\}| = \sum_{j \in \mathcal{N}(i)} (|\mathcal{N}(j)| - 1)$  que incluye a los vecinos de segundo orden de  $i$  pero no de forma única, sino que algunos elementos pueden considerarse varias veces: la cantidad de

---

<sup>1</sup> $\mathcal{N}'_2(i)$  no es un conjunto, ya que por definición de conjunto no pueden haber elementos repetidos, por lo que solamente se define lo que sería su cardinal (la cantidad de elementos).

veces que se sume el elemento  $j$  se corresponderá con la cantidad de vecinos de primer orden de  $i$  que a su vez sean vecinos de primer orden de  $j$ . Esta suma también incluye a los vecinos de segundo orden de  $i$  que también son sus vecinos de primer orden.

En el ejemplo de la Figura 3.1 el conjunto  $\mathcal{N}(i)$  está compuesto por  $\{k, j, l\}$ ,  $\mathcal{N}_2(i)$  está compuesto por  $\{s, f, g\}$ , entonces  $|\mathcal{N}(i)| = 3$ ,  $|\mathcal{N}_2(i)| = 3$ , pero  $|\mathcal{N}'_2(i)| = 6$ , ya que la observación  $s$  se considera dos veces y las observaciones  $k$  y  $j$  también son consideradas, aunque sean también vecinos de primer orden de  $i$ .



**Figura 3.1:** Ejemplo de vecinos de primer y segundo orden

Por definición  $w_{ik} \leq \frac{1}{|\mathcal{N}(i)|}$  para todo  $k \in \mathcal{N}(i)$ . Sin embargo, no es posible establecer a priori una cota para  $w_{kj}$  con  $k \in \mathcal{N}(i)$  y  $j \in \mathcal{N}(k)$ . Lo único que por definición debe cumplirse es que  $w_{kj} \leq 1$  para todo  $k \in \mathcal{N}(i)$  y  $j \in \mathcal{N}(k)$ . Con esta información se intentará encontrar una cota superior que asegure el cumplimiento de la condición 3.

$$\begin{aligned}
 \max_{1 \leq i \leq n} \left( \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{k=1}^n |w_{ik} w_{kj}| \right) &= \max_{1 \leq i \leq n} \left( \sum_{k \in \mathcal{N}(i)} \sum_{\substack{j \in \mathcal{N}(k) \\ j \neq i}} |w_{ik} w_{kj}| \right) \\
 &\leq \max_{1 \leq i \leq n} \left( \sum_{k \in \mathcal{N}(i)} \sum_{\substack{j \in \mathcal{N}(k) \\ j \neq i}} \frac{1}{|\mathcal{N}(i)|} \right) \\
 &\leq \max_{1 \leq i \leq n} \left( \frac{|\mathcal{N}'_2(i)|}{|\mathcal{N}(i)|} \right)
 \end{aligned}$$

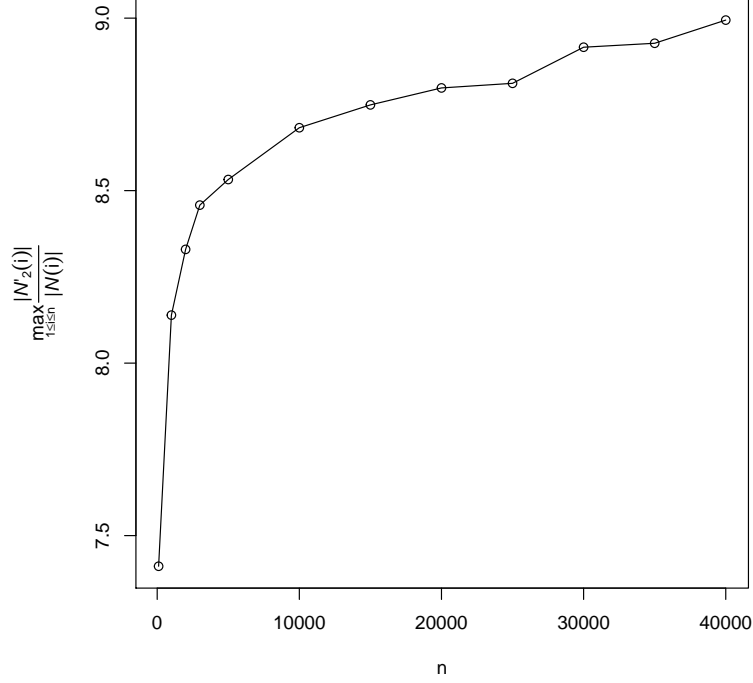
$\frac{|\mathcal{N}'_2(i)|}{|\mathcal{N}(i)|}$  puede interpretarse como la cantidad promedio de vecinos que tiene un vecino de  $i$  (sin contar a  $i$ ), por lo que  $\max_{1 \leq i \leq n} \left( \frac{|\mathcal{N}'_2(i)|}{|\mathcal{N}(i)|} \right)$  sería el máximo de ese promedio variando  $i$ . Si bien es esperable que este valor sea superior a uno, es posible inferir (por las características de la vecindad triangular) que cada observación tiene en promedio un número acotado de vecinos.

Si bien en principio este promedio puede aumentar con  $n$ , se supone que no superará cierto umbral, por lo que bajo este supuesto se admite el cumplimiento de la tercera condición del Cuadro 3.1.

Para verificarlo de forma empírica, se simularon ubicaciones espaciales para  $n$  entre 100 y 40000<sup>1</sup>, para cada  $n$  se simularon 100 conjuntos de datos, considerando en cada caso su estructura de vecindad triangular con los pesos definidos en 3.11. Para cada conjunto de datos simulados se calculó  $\max_{1 \leq i \leq n} \left( \frac{|\mathcal{N}'_2(i)|}{|\mathcal{N}(i)|} \right)$  y luego se consideró el promedio de las 100 repeticiones para cada valor de  $n$ . En la Figura 3.2 se presentan los resultados de estas simulaciones. Como puede observarse, si bien el máximo del promedio de vecinos aumenta con  $n$ , sobre todo para  $n$  menor a 2000, nunca es superior a 9 (varía entre 7.4 y 8.99), por lo que puede concluirse que  $\max_{1 \leq i \leq n} \left( \frac{|\mathcal{N}'_2(i)|}{|\mathcal{N}(i)|} \right)$  es acotado, es decir, no tiende a infinito cuando  $n$  tiende a infinito.

---

<sup>1</sup>los procesadores utilizados soportaron hasta  $n = 40000$ .



**Figura 3.2:** Maximo del promedio de vecinos que tiene un vecino de  $i$ , con  $i \in D = \{1, \dots, n\}$ , segun el valor de  $n$ .

Una vez establecida la estructura de vecindad a utilizar y los pesos asociados, resta determinar como se estimaran los parametros de la matriz de varianzas y covarianzas de los errores espaciales.

Basandose en Gaetan and Guyon [19], a partir del modelo (3.1) y definiendo  $\vartheta = (\theta, \sigma^2)$ , el estimador maximo verosımil  $\hat{\vartheta} = (\hat{\theta}, \hat{\sigma}^2)$  de  $\vartheta$  maximiza:

$$l(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n, \vartheta) = -\frac{1}{2}(\log(|\Sigma|) + \hat{\varepsilon}'\Sigma^{-1}\hat{\varepsilon})$$

donde  $\Sigma = \Sigma(\vartheta)$ , es decir,  $\Sigma$  depende de  $\vartheta$ .

Para estimar  $\hat{\vartheta}$  se requiere aplicar una tecnica de optimizacion iterativa. En los modelos CAR y SAR se conoce la forma parametrica de  $\Sigma^{-1}$  por lo que es posible calcular  $\hat{\varepsilon}'\Sigma^{-1}\hat{\varepsilon}$ . A su vez, tambien es posible conocer  $|\Sigma|$ , que en el caso de CAR(1) es  $|\Sigma_{\text{CAR}}| = \sigma^{2n}|\mathbf{I} - \theta\mathbf{W}_{\mathcal{R}}|^{-1}$  y en el caso de SAR(1) es  $|\Sigma_{\text{SAR}}| = \sigma^{2n}|\mathbf{I} - \theta\mathbf{W}_{\mathcal{R}}|^{-2}$ . En ambos casos  $|\mathbf{I} - \theta\mathbf{W}_{\mathcal{R}}| = \prod_{i=1}^n (1 - \theta\rho_i(\mathbf{W}_{\mathcal{R}}))$ , donde  $\rho_i(\mathbf{W}_{\mathcal{R}})$  representa el  $i$ -esimo valor propio de la matriz  $\mathbf{W}_{\mathcal{R}}$ .

$\Sigma$  es definida positiva si:

- $\hat{\theta} < \frac{1}{\rho_{max}(\mathbf{W}_{\mathcal{R}})}$  cuando  $\rho_{min}(\mathbf{W}_{\mathcal{R}}) \geq 0$
- $\hat{\theta} > \frac{1}{\rho_{min}(\mathbf{W}_{\mathcal{R}})}$  cuando  $\rho_{max}(\mathbf{W}_{\mathcal{R}}) \leq 0$
- $\frac{1}{\rho_{min}(\mathbf{W}_{\mathcal{R}})} < \hat{\theta} < \frac{1}{\rho_{max}(\mathbf{W}_{\mathcal{R}})}$  cuando  $\rho_{min}(\mathbf{W}_{\mathcal{R}}) < 0 < \rho_{max}(\mathbf{W}_{\mathcal{R}})$

En el caso CAR,  $\hat{\varepsilon}'\Sigma^{-1}\hat{\varepsilon} = \frac{1}{\sigma^2}\hat{\varepsilon}'(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\hat{\varepsilon}$  y el estimador máximo verosímil es:

$$\begin{aligned}\hat{\vartheta} &= \underset{\vartheta}{\operatorname{argmín}} \left( \frac{-1}{n} \log(|\mathbf{I} - \mathbf{C}_{\mathcal{R}}|) + \log(\sigma^2) \right) \\ \hat{\sigma}^2(\hat{\vartheta}) &= \frac{1}{n} \hat{\varepsilon}'(\mathbf{I} - \mathbf{C}_{\mathcal{R}}(\hat{\vartheta}))\hat{\varepsilon}\end{aligned}\tag{3.12}$$

donde  $\mathbf{C}_{\mathcal{R}} = \mathbf{C}_{\mathcal{R}}(\vartheta)$  y  $\sigma^2 = \sigma^2(\vartheta)$ . En el caso SAR,  $\hat{\varepsilon}'\Sigma^{-1}\hat{\varepsilon} = \frac{1}{\sigma^2}\|(\mathbf{I} - \mathbf{C}_{\mathcal{R}})\hat{\varepsilon}\|_2^2$  y el estimador máximo verosímil es:

$$\begin{aligned}\hat{\vartheta} &= \underset{\vartheta}{\operatorname{argmín}} \left( \frac{-2}{n} \log(|\mathbf{I} - \mathbf{C}_{\mathcal{R}}|) + \log(\sigma^2) \right) \\ \hat{\sigma}^2(\hat{\vartheta}) &= \frac{1}{n} \|(\mathbf{I} - \mathbf{C}_{\mathcal{R}}(\hat{\vartheta}))\hat{\varepsilon}\|_2^2\end{aligned}\tag{3.13}$$

### 3.2.4. Procedimiento para eliminar la dependencia espacial

Se propone el siguiente procedimiento para estimar las matrices  $\hat{\Sigma}_{\text{CAR}}$  y  $\hat{\Sigma}_{\text{SAR}}$  de un conjunto de datos, de modo de poder “eliminar” la dependencia espacial de los mismos y aplicar el método LASSO para datos independientes.

- Paso 1: Estimar un modelo LASSO al problema original, como si no hubiese autocorrelación espacial. El valor de  $\lambda$  será elegido por validación cruzada<sup>1</sup>. Este modelo se denomina “LASSO inicial”.
- Paso 2: Aplicar pruebas de autocorrelación espacial a los residuos, como las de Moran y Geary definidas en la sección 3.2.1.

---

<sup>1</sup>el valor más grande de la secuencia  $\{\lambda_1, \dots, \lambda_T\}$  de modo que el error esté dentro de  $\pm 1$  error estándar del mínimo.

- Paso 3: Si se rechaza la hipótesis nula de ausencia de autocorrelación, se continúa el procedimiento, de lo contrario termina (los datos no serían espaciales).
- Paso 4: Considerando estructuras CAR(1) y SAR(1), se requiere estimar únicamente los parámetros  $\theta$  y  $\sigma^2$ .
- Paso 5: Estimar  $\hat{\theta}$  y  $\hat{\sigma}^2$  por máxima verosimilitud de acuerdo a (3.13) o (3.14), según corresponda.
- Paso 6: Estimar  $\hat{\Sigma}_{\text{CAR}}$  o  $\hat{\Sigma}_{\text{SAR}}$  utilizando  $\hat{\theta}$  y  $\hat{\sigma}^2$  estimados en el punto anterior.
- Paso 7: Eliminar la dependencia espacial de los datos, utilizando la matriz estimada en el punto anterior, aplicando 3.3.
- Paso 8: Estimar modelo LASSO a los datos transformados en el punto anterior, nuevamente eligiendo el valor de  $\lambda$  por validación cruzada. Este modelo se denomina “LASSO ajustado”.

### 3.3. Otra metodología para selección de variables en el caso espacial: LARS<sub>m</sub>

En esta sección se utiliza la metodología desarrollada en el artículo de Zhu, Huang y Reyes (Zhu et al. [57]), el cual refiere a selección de variables en modelos lineales espaciales para datos en grillas regulares.

En el mismo se realiza en simultáneo la selección de variables y la estimación de parámetros en una regresión espacial, por máxima verosimilitud penalizada, usando un modelo LASSO espacial adaptativo con errores de tipo CAR y SAR. Se considera a su vez un orden de vecindad  $K$  (a priori mayor a uno), y las mismas restricciones en los pesos  $\mathbf{W}_{\mathcal{F}}^{(k)}$  que las explicitadas en la sección 3.2.2 de este capítulo, esto es,  $w_{ii}^{(k)} = 0$ ,  $w_{ij}^{(k)} = w_{ji}^{(k)}$  y  $w_{ij}^{(k)} = 0$  si  $j \notin \mathcal{N}_k(i)$ , para  $0 \leq i, j \leq n$ , siendo  $k$  el  $k$ -ésimo orden de vecindad ( $1 \leq k \leq K$ ).

Se parte del modelo definido en (3.1), con errores de tipo CAR o SAR. Sin pérdida de generalidad se asume que  $\mathbf{X}$  está estandarizada e  $Y$  centrado. A los efectos de compatibilizar con la sección anterior, se considera el caso en que el orden de vecindad es igual a uno. Al igual que en la sección anterior, se utiliza la estructura de vecindad triangular con los pesos definidos en (9) para datos irregulares y la condición de que  $|\theta| < 1$ .

A continuación se presentan los principales resultados de Zhu et al. [57] ajus-

tados al problema de interés.

Sea  $\eta = (\beta, \theta, \sigma^2)$  el vector de parámetros a estimar, con  $\beta = (\beta_1, \dots, \beta_p)$ . Con el modelo (3.1), el logaritmo de la función de verosimilitud es:

$$\begin{aligned} \log(L(\eta, Y, \mathbf{X})) &= \text{constante} - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (Y - \mathbf{X}\beta)' \Sigma^{-1} (Y - \mathbf{X}\beta) \\ &= \text{constante} + l(\eta) \end{aligned} \quad (3.14)$$

Se nota por  $\hat{\eta} = \underset{\eta}{\operatorname{argm\acute{a}x}}(l(\eta))$  al estimador máximo verosímil de  $\eta$ .

Se considera el logaritmo de la función de verosimilitud penalizada:

$$\begin{aligned} Q(\eta) &= l(\eta) - n \sum_{j=1}^p \lambda_j |\beta_j| - n\tau |\theta| \\ &= -\frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (Y - \mathbf{X}\beta)' \Sigma^{-1} (Y - \mathbf{X}\beta) - n \sum_{j=1}^p \lambda_j |\beta_j| - n\tau |\theta| \end{aligned} \quad (3.15)$$

donde  $\{\lambda_j\}_{j=1}^p$  son los parámetros de regularización para los coeficientes de regresión  $\beta$  y  $\tau$  es el parámetro de regularización para el coeficiente autorregresivo de primer orden  $\theta$ . Se denota  $\hat{\eta}_P = \underset{\eta}{\operatorname{argm\acute{a}x}}(Q(\eta))$  al estimador máximo verosímil penalizado de  $\eta$ .

### 3.3.1. Estimación máximo verosímil penalizada via $\mathbf{LARS}_m$

Sea  $\hat{\eta}^{(0)}$  el valor inicial de  $\eta$  al comienzo de las iteraciones ( $\hat{\eta}^{(0)}$  coincide con  $\hat{\eta}$ , es decir el estimador máximo verosímil sin penalizar). En la  $m$ -ésima iteración, el logaritmo de la función de verosimilitud definido en (3.15) es aproximado cuando  $\eta \approx \hat{\eta}^{(m-1)}$  (salvo por una constante) por:

$$\begin{aligned}
Q^*(\eta) &= (\eta - \hat{\eta}^{(m-1)})' \frac{\partial l(\hat{\eta}^{(m-1)})}{\partial \eta} - \frac{1}{2} (\eta - \hat{\eta}^{(m-1)})' \mathcal{I}(\hat{\eta}^{(m-1)}) (\eta - \hat{\eta}^{(m-1)}) \\
&\quad - n \sum_{j=1}^p \lambda_j |\beta_j| - n\tau |\theta|
\end{aligned} \tag{3.16}$$

donde  $\mathcal{I}(\cdot)$  es la matriz de información definida como  $\mathcal{I}(\eta) = E \left( -\frac{\partial^2 l(\eta)}{\partial \eta \partial \eta'} \right)$ . En cada iteración se actualiza  $\hat{\eta}^{(m-1)}$  a través de  $\hat{\eta}^{(m)} = \underset{\eta}{\operatorname{argm\acute{a}x}} (Q^*(\eta))$ , repitiendo hasta la convergencia.

Dado que la matriz  $\mathcal{I}(\eta)$  es diagonal por bloques, se actualiza  $\hat{\beta}^{(m-1)}$  y  $\hat{\omega}^{(m-1)} = (\hat{\theta}^{(m-1)}, (\hat{\sigma}^2)^{(m-1)})$  por separado.

Por un lado se actualiza  $\hat{\beta}^{(m-1)}$  por  $\hat{\beta}^{(m)}$ :

$$\begin{aligned}
\hat{\beta}^{(m)} &= \underset{\beta}{\operatorname{argm\acute{m}n}} \left( -(\beta - \hat{\beta}^{(m-1)})' \frac{\partial l(\hat{\eta}^{(m-1)})}{\partial \beta} + \frac{1}{2} (\beta - \hat{\beta}^{(m-1)})' \mathcal{I}(\hat{\beta}^{(m-1)}) (\beta - \hat{\beta}^{(m-1)}) \right. \\
&\quad \left. + n \sum_{j=1}^p \lambda_j |\beta_j| \right)
\end{aligned}$$

La solución del problema anterior puede obtenerse de forma equivalente como:

$$\hat{\beta}^{*(m)} = \underset{\beta^*}{\operatorname{argm\acute{m}n}} \left( \frac{1}{2} (Y^* - \mathbf{X}^* \beta^*)' (Y^* - \mathbf{X}^* \beta^*) + n \sum_{j=1}^p |\beta_j^*| \right) \tag{3.17}$$

donde:

- $\mathbf{G}$  es tal que  $\mathcal{I}(\hat{\beta}^{(m-1)}) = \mathbf{G}' \mathbf{G}$
- $\mathbf{X}^* = \mathbf{G} \operatorname{diag} \left( \left\{ \frac{1}{\lambda_j} \right\}_{j=1}^p \right)$
- $Y^* = (\mathbf{G}^{-1})' \left( \frac{\partial l(\hat{\eta}^{(m-1)})}{\partial \beta} + \mathcal{I}(\hat{\beta}^{(m-1)}) \hat{\beta}^{(m-1)} \right)$
- $\beta^* = \operatorname{diag} \left( \left\{ \lambda_j \right\}_{j=1}^p \right) \beta$

Entonces,  $\hat{\beta}^{(m)} = \operatorname{diag} \left( \left\{ \frac{1}{\lambda_j} \right\}_{j=1}^p \right) \hat{\beta}^{*(m)}$ .

Por otra parte se actualiza  $\hat{\omega}^{(m-1)}$  por:



$$\hat{\omega}^{(m)} = \underset{\omega}{\operatorname{argmín}} \left( -(\omega - \hat{\omega}^{(m-1)})' \frac{\partial l(\hat{\eta}^{(m-1)})}{\partial \omega} + \frac{1}{2}(\omega - \hat{\omega}^{(m-1)})' \mathcal{I}(\hat{\omega}^{(m-1)})(\omega - \hat{\omega}^{(m-1)}) + n\tau|\theta| \right)$$

Dado que  $\sigma^2$  no está sujeto a ninguna penalidad, se actualizan los términos de una manera diferente. La solución de  $\sigma^2$  en la ecuación anterior puede obtenerse a partir de

$$(\hat{\sigma}^{*2})^{(m)} = \frac{\mathbf{X}_2^{**'} Y^{**}}{\mathbf{X}_2^{**'} \mathbf{X}_2^{**}}$$

donde:

- $\mathbf{H}$  es tal que  $\mathcal{I}(\hat{\omega}^{(m-1)}) = \mathbf{H}'\mathbf{H}$
- $\mathbf{X}_2^{**} = \mathbf{H}_2$
- $Y^{**} = (\mathbf{H}^{-1})' \left( \frac{\partial l(\hat{\eta}^{(m-1)})}{\partial \omega} + \mathcal{I}(\hat{\omega}^{(m-1)})\hat{\omega}^{(m-1)} \right)$
- $\sigma^{*2} = c\theta + \sigma^2$
- $c = \frac{\mathbf{H}'_2 \mathbf{H}_1}{\mathbf{H}'_2 \mathbf{H}_2}$

La solución de este problema puede obtenerse de forma equivalente por:

$$\hat{\theta}^{*(m)} = \underset{\theta^*}{\operatorname{argmín}} \left( \frac{1}{2} (Y^{**} - \mathbf{X}_1^{**} \theta^*)' (Y^{**} - \mathbf{X}_1^{**} \theta^*) + n|\theta^*| \right) \quad (3.18)$$

donde:

- $\mathbf{X}_1^{**} = \frac{1}{\tau}(\mathbf{H}_1 - c\mathbf{H}_2)$
- $\theta^* = \tau\theta$

Entonces,  $\hat{\theta}^{(m)} = \frac{\hat{\theta}^{*(m)}}{\tau}$  y  $(\hat{\sigma}^2)^{(m)} = (\hat{\sigma}^{*2})^{(m)} - c\hat{\theta}^{(m)}$ .

Cuando se obtiene la convergencia,  $\hat{\eta}_{PA}$  representa el estimador máximo verosímil penalizado aproximado de  $\eta$ . En cada iteración las ecuaciones (3.17) y (3.18) se resuelven mediante un algoritmo LARS<sup>1</sup>, por lo que Zhu et al lo denominan “algoritmo LARS de varios pasos” (LARS<sub>m</sub>).

---

<sup>1</sup>sigla en inglés, de Least Angle Regression. Por mayor información consultar Efron et al. [15], Hastie [25] y Canu [12]

Para aplicar esta metodología al presente trabajo se implementó un algoritmo en R, utilizando tanto los resultados anteriores, como también los de la sección 5 de Zhu et al. [57] (que se detallan en el Anexo 2). Para hallar  $\lambda_1, \dots, \lambda_p$  y  $\tau$ , los autores sugieren partir de valores iniciales, que denominaremos  $\lambda_0$  y  $\tau_0$ , y calcularlos de la siguiente manera:

$$\lambda_j = \frac{\lambda_0 \log(n)}{n \hat{\beta}_j} \quad (3.19)$$

$$\tau = \frac{\tau_0 \log(n)}{n \hat{\theta}} \quad (3.20)$$

Para elegir a  $\lambda_0$  y  $\tau_0$ , se calcula el BIC (Bayesian Information Criterion):

$$BIC(\lambda_0, \tau_0) = -2l(\hat{\eta}, \lambda_0, \tau_0) + \log(n) \left( \sum_{j=1}^p I_{\{\hat{\beta}_j \neq 0\}} + I_{\{\hat{\theta} \neq 0\}} \right) \quad (3.21)$$

para todas las combinaciones de  $\lambda_0$  y  $\tau_0$  consideradas. En cada iteración se selecciona la combinación con el menor valor de BIC. Por las características del algoritmo LARS, es posible obtener valores de  $\lambda_0$  y  $\tau_0$  de forma computacionalmente eficiente.

El algoritmo LARS está relacionado al clásico método de selección de variables denominado “stagewise” o “forward stepwise regression”, y se caracteriza por requerir solamente  $p$  pasos para obtener una solución completa, siendo  $p$  la cantidad de variables consideradas. En este algoritmo se empieza con todos los coeficientes iguales a cero, y se encuentra en primer lugar a la variable más correlacionada con la variable dependiente  $Y$ . Se efectúa el paso más largo posible en la dirección de ese predictor, hasta que otra variable tenga tanta correlación con el residuo como el primer predictor. A continuación se procede en una dirección equiangular entre los dos predictores hasta que una tercera variable entre al conjunto de variables más correlacionadas con el correspondiente residuo. Se procede entonces en la dirección de menor ángulo entre las 3 variables más correlacionadas, y así sucesivamente hasta haber seleccionado todas las variables.

En cada uno de estos pasos el algoritmo calcula un valor de  $\lambda$  y produce una estimación  $\hat{\beta}$  asociada a ese valor de  $\lambda$ . De esta manera, es posible utilizar estos resultados parciales que proporciona el algoritmo LARS (aplicado una sola

vez por iteración) para elegir los mejores valores de  $\lambda_0$  y  $\tau_0$ , y poder estimar  $\{\lambda_1, \dots, \lambda_p\}$ ,  $\tau$  y los respectivos  $\hat{\beta}^{(m)}$ ,  $\hat{\theta}^{(m)}$  y  $\hat{\sigma}^{2(m)}$  en cada iteración  $m$ .

Los autores presentan también algunas propiedades asintóticas de estos estimadores, que se enuncian en los teoremas 3 y 4<sup>1</sup>.

**Teorema 3.** *Bajo las siguientes condiciones:*

1. La matriz  $\mathbf{X}$  es de rango completo y  $\eta = (\beta, \theta, \sigma^2) \in \Omega$ , donde  $\Omega$  es un conjunto abierto de  $\mathbb{R}^{p+2}$  que contiene a  $\eta$ .
2.  $\Sigma_\omega$  es una matriz definida positiva dos veces diferenciable respecto a  $\omega = (\theta, \sigma^2)$ , con derivadas de segundo orden continuas.
3.  $\mathcal{I}(\eta)^{-1/2} \left( -\frac{\partial^2 l(\eta)}{\partial \eta \partial \eta'} \right) \mathcal{I}(\eta)^{-1/2} \xrightarrow{P} \mathbf{I} \in \mathbb{R}^{(p+q+1) \times (p+q+1)}$  cuando  $n \rightarrow \infty$
4.  $\frac{1}{n} \mathcal{I}(\beta) \rightarrow \mathbf{J}(\beta)$  y  $\frac{1}{n} \mathcal{I}(\omega) \rightarrow \mathbf{J}(\omega)$  cuando  $n \rightarrow \infty$

Sin pérdida de generalidad se asume que las primeras  $p^*$  variables participan del modelo, y las siguientes  $p - p^*$  no participan. Se define entonces  $a_n = \max(\lambda_1, \dots, \lambda_{p^*}, \tau)$  y  $b_n = \min(\lambda_{p^*+1}, \dots, \lambda_p)$ . Asumiendo que  $a_n = O(n^{-1/2})$  cuando  $n \rightarrow \infty$  se tiene:

- (a) Con probabilidad tendiendo a 1, existe un maximizador local  $\hat{\eta}$  de  $Q(\eta)$  definido en (3.15) tal que  $\|\hat{\eta} - \eta\| = O_P\left(\frac{1}{\sqrt{n}} + a_n\right)$ ,
- (b) Si adicionalmente  $\sqrt{nb_n} \rightarrow \infty$  cuando  $n \rightarrow \infty$ , entonces con probabilidad tendiendo a 1,  $\hat{\beta}_{A^*c} = \mathbf{O}$ ,
- (c) Si adicionalmente  $a_n = o\left(\frac{1}{\sqrt{n}}\right)$ , entonces  $\sqrt{n}(\hat{\eta}_{A^*} - \eta_{A^*}) \xrightarrow{D} N(\mathbf{O}, \mathbf{J}(\eta_{A^*})^{-1})$ , donde  $\eta_{A^*} = (\beta_{A^*}, \theta, \sigma^2)$ ,

$$\hat{\eta}_{A^*} = (\hat{\beta}_{A^*}, \hat{\theta}, \hat{\sigma}^2), \quad \mathbf{J}(\eta_{A^*}) = \begin{bmatrix} \mathbf{J}(\beta_{A^*}) & \mathbf{O} \\ \mathbf{O} & \mathbf{J}(\omega) \end{bmatrix}$$

La parte (a) del teorema anterior establece la existencia del estimador máximo verosímil penalizado,  $\hat{\eta}_{PA}$ . La parte (b) asegura que se identifican correctamente los parámetros que valen cero y la parte (c) es un teorema central del límite para el estimador máximo verosímil penalizado para los coeficientes distintos de cero.

**Teorema 4.** *Suponiendo que se verifican las condiciones 1 a 4 del Teorema 3,  $\sqrt{nb_n} \rightarrow \infty$  cuando  $n \rightarrow \infty$  y  $a_n = o\left(\frac{1}{\sqrt{n}}\right)$ . Entonces, con probabilidad*

---

<sup>1</sup>Los enunciados fueron adaptados al problema de interés, es decir  $K = 1$ .

tendiendo a 1,  $\hat{\eta}_{A^{*c}} = \mathbf{O}$  y  $\sqrt{n}(\hat{\eta}_{A^*}^{(m)} - \eta_{A^*}) \xrightarrow{D} N(\mathbf{O}, \mathbf{J}(\eta_{A^*})^{-1})$  para todo  $m \in \{1, 2, \dots\}$ .

En particular con probabilidad tendiendo a 1,  $\hat{\eta}_{PA, A^{*c}} = \mathbf{O}$  y  $\sqrt{n}(\hat{\eta}_{PA, A^*}^{(m)} - \eta_{A^*}) \xrightarrow{D} N(\mathbf{O}, \mathbf{J}(\eta_{A^*})^{-1})$ .

El Teorema 4 establece que las propiedades asintóticas se mantienen para cada paso.

Las demostraciones de los Teoremas 3 y 4 se encuentran en Zhu et al. [56].

### 3.3.2. Procedimiento para aplicar LARS<sub>m</sub>

Se propone el siguiente procedimiento para aplicar el método LARS<sub>m</sub> detallado anteriormente, para poder estimar el modelo penalizado.

- Se estima el valor inicial  $\hat{\eta}^{(0)}$ . El valor inicial se corresponde con el estimador máximo verosímil de  $\eta$  sin penalizar. El método de optimización utilizado requiere asimismo de un valor inicial, el cual es simulado aleatoriamente a partir de  $p$  variables normales estándar para  $\hat{\beta}$ , 1 variable normal truncada para  $\hat{\theta}$  (considerando las mismas condiciones que al estimar  $\hat{\theta}$  en el método de eliminación de la dependencia espacial) y 1 variable uniforme en el intervalo  $[0, 10]$  para  $\hat{\sigma}^2$ .
- Se define una tolerancia ( $tol = 0.01$ ) y una máxima cantidad de pasos a considerar ( $m.max = 100$ ). En cada paso  $m = 1, 2, \dots$ :
  1. A partir de  $\hat{\eta}^{(m-1)} = (\hat{\beta}^{(m-1)}, \hat{\theta}^{(m-1)}, \hat{\sigma}^{2(m-1)})$  se calculan  $\frac{\partial l(\hat{\eta}^{(m-1)})}{\partial \beta}$ ,  $\frac{\partial l(\hat{\eta}^{(m-1)})}{\partial \omega}$ ,  $\mathcal{I}(\hat{\beta}^{(m-1)})$ ,  $\mathcal{I}(\hat{\omega}^{(m-1)})$ ,  $\mathbf{G}$ ,  $\mathbf{H}$ ,  $Y^*$  y  $Y^{**}$ .
  2. Dado que  $\mathbf{X}^* \beta^* = \mathbf{G} \beta$ , se estima el modelo  $\|Y^* - \mathbf{G} \beta\|_2^2 + \lambda \|\beta\|_1$  usando el algoritmo LARS. Este algoritmo devuelve  $p$  valores distintos de  $\lambda$  asociados a  $p$  vectores distintos  $\hat{\beta}$ , estos valores de  $\lambda$  se consideran los  $\lambda_0$  definidos anteriormente.
  3. Dado que  $\mathbf{X}_1^{**} \theta^* = (\mathbf{H}_1 - c\mathbf{H}_2) \theta$ , se estima el modelo  $\|Y^{**} - (\mathbf{H}_1 - c\mathbf{H}_2) \theta\|_2^2 + \tau \|\theta\|_1$  usando nuevamente el algoritmo LARS. En este caso el algoritmo devuelve un único valor de  $\tau$  asociado a un valor de  $\hat{\theta}$ . Ese valor de  $\tau$  se considera el  $\tau_0$  mencionado anteriormente. Con ese valor de  $\tau$  es posible calcular de forma cerrada  $\hat{\sigma}^{2(m)}$ .
  4. para cada uno de los  $p$  valores  $\lambda_0$  se define el vector  $(\lambda_1, \dots, \lambda_p)$ , de acuerdo a (3.19). Lo mismo para  $\tau_0$ , de acuerdo a (3.20).

5. Se elige la combinación de  $\lambda_0$  y  $\tau_0$  con menor BIC, de acuerdo a (3.21).
  6.  $\hat{\beta}^{(m)}$ ,  $\hat{\theta}^{(m)}$  y  $\hat{\sigma}^2(m)$  equivalen a  $\hat{\beta}$ ,  $\hat{\theta}$  y  $\hat{\sigma}^2$  de la combinación con menor BIC.
- El proceso termina cuando  $\|\hat{\eta}^{(m-1)} - \hat{\eta}^{(m)}\|_2 < tol$  o  $m = m.max$ .

Al comparar los dos métodos de selección de variables considerados en este capítulo, es decir, el método de eliminación de la dependencia espacial y la estimación máximo verosímil penalizada via LARS en  $m$  pasos, se obtiene que si bien ambos métodos son útiles para seleccionar variables en el contexto espacial, proceden de manera diferente.

Por un lado el método de eliminación de la dependencia espacial estima una matriz de covarianzas de los errores y luego la utiliza para quitar la estructura espacial de los mismos, transformando el problema en uno con errores independientes. Por otro lado, el método  $LARS_m$  considera un modelo penalizado y estima en simultáneo tanto los parámetros de la regresión como los de la matriz de covarianzas, pero no transforma los errores en independientes.

En la sección 3.2.3 se demostró que con covarianzas CAR o SAR, estructura de vecindad triangular y pesos específicos, el método de eliminación de la dependencia espacial optimiza la selección de variables (estimación consistente en signo), mientras que el método  $LARS_m$  optimiza la estimación de parámetros (convergencia del estimador al verdadero parámetro). Esto significa que realizan la selección de variables con distintos enfoques, y es de esperar que los resultados obtenidos difieran en este sentido. Este punto se verificará en los siguientes capítulos.

# Capítulo 4

## Simulaciones

En este capítulo se ponen en práctica los resultados del capítulo anterior a través de simulaciones utilizando distintos escenarios.

Se consideran 2 situaciones distintas, denominadas *Problema 1* y *Problema 2*, para los cuales se simulan distintos escenarios de acuerdo a los valores de  $n$ ,  $\sigma^2$ ,  $\theta$  (que se mantiene fijo) y el modelo de error utilizado (CAR o SAR).

En todos los escenarios se consideran ubicaciones espaciales irregulares. Se simula una matriz  $\mathbf{X}$  de dimensión  $n \times p$ , donde  $p$  es fijo e igual a 20 y  $n$  varía dependiendo del escenario. A su vez se define  $Y = X_1 + X_2 + X_3 + X_4 + \varepsilon$ , con  $\varepsilon \sim N(\mathbf{0}, \Sigma)$ , es decir  $p^* = 4$ , y se fijó el parámetro autorregresivo  $\theta$  en 0.9. Para cada escenario se ejecutan 100 réplicas independientes.

Los 2 problemas considerados varían únicamente en la simulación de la matriz  $\mathbf{X}$ , a saber:

- *Problema 1*: Variables normales estándar independientes ( $X_j \sim N(0, 1)$ ,  $1 \leq j \leq 20$ )
- *Problema 2*: Variables normales con media cero y matriz de varianzas y covarianzas tal que  $Cov(X_i, X_j) = 0.9^{|i-j|}$  ( $(X_1, \dots, X_{20})' \sim N(\mathbf{0}, \Sigma_X)$ , con  $[\Sigma_X]_{ij} = 0.9^{|i-j|}$ ).

Para los distintos escenarios se consideran valores de  $n$  iguales a 100, 200, 400 y 800, y valores de  $\sigma^2$  iguales a 0.5, 1, 1.5, 2, 5, 10, 20, 40, 60 y 80. Cada escenario surge de la combinación de:

1. un problema,
2. un valor de  $n$ ,
3. un valor de  $\sigma^2$ ,

#### 4. un modelo de error (CAR o SAR).

A su vez, en cada escenario se estima por un lado un modelo LASSO como si los datos fueran independientes, es decir, según el modelo (2.1) del Capítulo 2, a este modelo se lo denomina “LASSO inicial”. Luego (dentro del mismo escenario) se estima un modelo considerando la metodología de la Sección 3.1.1 del Capítulo 3, el cual se denomina “LASSO ajustado”<sup>1</sup>. En cada escenario simulado se aplica a su vez el método  $LARS_m$ , el cual se desarrolló en la Sección 3.3 del Capítulo 3.

Asimismo, en cada escenario se verifica el cumplimiento de las hipótesis relativas a la matriz  $\mathbf{X}$  en el Teorema 2, es decir, las hipótesis 1, 2, 5 y 7. Se obtiene que en ambos problemas se verifican todas las hipótesis que involucran a  $\mathbf{X}$ . Las simulaciones se realizaron en R utilizando los paquetes `glmnet`, `lars`, `MASS`, `spdep`, `expm` y `TruncatedNormal` (ver [43], [17], [26], [52], [6], [24] y [7]). Para la estructura de vecindad se utilizó como referencia Bivand et al. [4].

A continuación se presentan los principales resultados para cada problema. Por un lado en las Tablas 4.1 a 4.4 se exponen los resultados relativos a la selección de variables para el LASSO inicial, el LASSO ajustado y el  $LARS_m$ . Para cada escenario se presenta el total de variables seleccionadas (VARS), la cantidad de variables seleccionadas que participan del verdadero modelo (VP, de “verdaderos positivos”) y la cantidad de variables seleccionadas que no participan del verdadero modelo (FP, de “falsos positivos”)<sup>2</sup>. Se recuerda que la situación ideal consiste en seleccionar 4 variables en total, siendo  $VP = 4$  y  $FP = 0$ .

A su vez se presenta en las Figuras 4.1 a 4.4 el valor de  $\hat{\beta}$  para los distintos escenarios. En cada figura se muestran los resultados para distintos valores de  $n$  (filas), y para el LASSO inicial, LASSO ajustado y  $LARS_m$  (columnas).

En el eje  $x$  se representa el índice  $i$  (0 a 20) mientras que en el eje  $y$  se representa el valor de  $\hat{\beta}_i$ . La línea continua en color negro representa el verdadero valor de  $\beta$ , mientras que cada punto representa la media de  $\hat{\beta}_i$  para un escenario en particular. Es decir, para cada abscisa  $i$  existen 10 ordenadas que corresponden a los distintos escenarios que surgen al variar  $\sigma^2$ . Cada escenario se identifica con un color distinto.

---

<sup>1</sup>Cabe destacar que en las simulaciones se asume que siempre se identifica correctamente el tipo de errores, por lo que no se estima un modelo SAR para errores CAR y viceversa.

<sup>2</sup>Los resultados de cada escenario surgen de 100 réplicas, por lo que se trata de promedios.

La media de cada  $\hat{\beta}_i$  se calcula solamente sobre las réplicas que seleccionan a esa variable, es decir, las que tienen  $\hat{\beta}_i \neq 0$ , por lo que si en algún escenario algún  $\hat{\beta}_i = 0$  en todas las réplicas, el símbolo asociado a ese escenario no estará representado en la figura correspondiente. En algunos casos puntuales el punto correspondiente a  $\hat{\beta}_i$  no se representa en el gráfico por exceder el rango utilizado para la figura, lo cual se especifica oportunamente.

Por otra parte, las Tablas 4.5 a 4.8 incluyen los valores estimados de  $\hat{\theta}$  y  $\hat{\sigma}^2$  para LASSO ajustado y LARS $_m$  (se recuerda que en el LASSO inicial no se estima ninguna matriz de covarianzas, por lo que no se estima  $\hat{\theta}$  y  $\hat{\sigma}^2$ ).

De forma muy general se puede afirmar que a medida que aumenta  $n$  (para un mismo valor de  $\sigma^2$ ) se obtienen mejores resultados en todos los métodos considerados, tanto en la selección de variables como en la estimación de parámetros. Mientras que para un mismo valor de  $n$ , al aumentar  $\sigma^2$  se obtienen peores resultados en ambos (selección y estimación), y esto ocurre para los tres métodos. La performance en selección al aumentar  $\sigma^2$  decae más rápidamente en modelos SAR que en CAR.

Es razonable que los métodos seleccionen peor a medida que aumente  $\sigma^2$ , porque este parámetro incrementa la magnitud del error aleatorio, incrementando su peso relativo en la variable dependiente  $Y$  y por lo tanto, reduciendo el peso relativo de las variables que lo conforman ( $X_1$  a  $X_4$ ).

Al comparar los resultados según el problema, se observa para valores pequeños de  $\sigma^2$  (menores a 5, 10 o 20 dependiendo de  $n$ ), el *Problema 1* selecciona más verdaderos positivos que el *Problema 2*, y para valores grandes de  $\sigma^2$  la situación se invierte, siendo el *Problema 2* el que selecciona más verdaderos positivos. Incluso se observan algunos escenarios donde la cantidad de falsos positivos del *Problema 1* es superior a la del *Problema 2*.

Al comparar los métodos considerados, en general se aprecia que LARS $_m$  selecciona más verdaderos positivos que LASSO ajustado, y este a su vez selecciona más que LASSO inicial, aunque para valores pequeños de  $\sigma^2$  en general todos seleccionan muy bien.

La mayor diferencia entre los métodos se aprecia en valores medios y grandes de  $\sigma^2$ , donde si bien ninguno de los métodos logra identificar a todas las variables que participan del verdadero modelo, LARS $_m$  obtiene resultados sustancialmente mejores que los otros.



$n$	$\sigma^2$	LASSO inicial			LASSO ajustado			LARS $_m$		
		VARS	VP	FP	VARS	VP	FP	VARS	VP	FP
100	0.5	5.12	4.00	1.12	5.13	4.00	1.13	4.87	3.68	1.19
	1	5.15	4.00	1.15	5.19	4.00	1.19	5.53	4.00	1.53
	1.5	5.15	4.00	1.15	5.46	4.00	1.46	5.75	4.00	1.75
	2	5.20	4.00	1.20	5.15	4.00	1.15	5.54	4.00	1.54
	5	4.64	3.64	1.00	4.71	3.83	0.88	5.63	3.92	1.71
	10	3.02	2.16	0.86	2.67	2.02	0.65	4.86	3.07	1.79
	20	1.19	0.79	0.40	1.27	0.98	0.29	2.56	1.85	0.71
	40	0.40	0.26	0.14	0.63	0.45	0.18	1.75	1.08	0.67
	60	0.45	0.20	0.25	0.37	0.20	0.17	1.30	0.67	0.63
80	0.47	0.26	0.21	0.55	0.22	0.33	1.48	0.74	0.74	
200	0.5	4.69	4.00	0.69	4.74	4.00	0.74	5.11	3.90	1.21
	1	4.49	4.00	0.49	4.46	4.00	0.46	5.37	4.00	1.37
	1.5	4.76	4.00	0.76	4.61	4.00	0.61	5.16	4.00	1.16
	2	4.62	4.00	0.62	4.57	4.00	0.57	5.20	4.00	1.20
	5	4.72	3.99	0.73	4.65	4.00	0.65	5.79	4.00	1.79
	10	3.69	3.36	0.33	3.80	3.52	0.28	5.62	3.97	1.65
	20	1.80	1.57	0.23	2.41	2.05	0.36	3.88	3.00	0.88
	40	0.64	0.51	0.13	0.79	0.58	0.21	2.16	1.69	0.47
	60	0.38	0.24	0.14	0.60	0.46	0.14	1.44	1.02	0.42
80	0.33	0.23	0.10	0.33	0.25	0.08	1.21	0.79	0.42	
400	0.5	4.18	4.00	0.18	4.22	4.00	0.22	5.07	3.98	1.09
	1	4.27	4.00	0.27	4.32	4.00	0.32	5.02	4.00	1.02
	1.5	4.26	4.00	0.26	4.27	4.00	0.27	5.01	4.00	1.01
	2	4.26	4.00	0.26	4.25	4.00	0.25	5.05	4.00	1.05
	5	4.23	4.00	0.23	4.31	4.00	0.31	5.15	4.00	1.15
	10	4.28	3.99	0.29	4.09	4.00	0.09	5.13	4.00	1.13
	20	3.32	3.02	0.30	3.74	3.51	0.23	5.06	3.91	1.15
	40	1.41	1.24	0.17	1.53	1.42	0.11	3.24	2.60	0.64
	60	0.54	0.47	0.07	0.67	0.60	0.07	2.26	1.90	0.36
80	0.50	0.34	0.16	0.70	0.52	0.18	1.74	1.40	0.34	
800	0.5	4.12	4.00	0.12	4.06	4.00	0.06	4.92	4.00	0.92
	1	4.29	4.00	0.29	4.12	4.00	0.12	5.03	4.00	1.03
	1.5	4.16	4.00	0.16	4.09	4.00	0.09	4.74	4.00	0.74
	2	4.15	4.00	0.15	4.09	4.00	0.09	4.86	4.00	0.86
	5	4.12	4.00	0.12	4.09	4.00	0.09	4.78	4.00	0.78
	10	4.11	4.00	0.11	4.15	4.00	0.15	4.99	4.00	0.99
	20	4.03	3.93	0.10	4.00	3.93	0.07	4.82	4.00	0.82
	40	2.55	2.48	0.07	2.98	2.91	0.07	4.76	3.88	0.88
	60	1.21	1.04	0.17	1.65	1.59	0.06	4.02	3.42	0.60
80	0.74	0.71	0.03	1.00	0.90	0.10	3.27	2.78	0.49	
Verdadero valor		4.00	4.00	0.00	4.00	4.00	0.00	4.00	4.00	0.00

**Tabla 4.1:** Resultados de LASSO inicial, LASSO ajustado y LARS $_m$  para el *Problema 1* con errores CAR.

$n$	$\sigma^2$	LASSO inicial			LASSO ajustado			LARS $_m$		
		VARS	VP	FP	VARS	VP	FP	VARS	VP	FP
100	0.5	5.03	4.00	1.03	5.03	4.00	1.03	4.82	3.51	1.31
	1	5.31	4.00	1.31	4.96	4.00	0.96	5.81	4.00	1.81
	1.5	5.32	3.99	1.33	5.18	4.00	1.18	5.39	4.00	1.39
	2	4.84	3.86	0.98	4.69	3.98	0.71	5.10	4.00	1.10
	5	3.29	2.38	0.91	4.53	3.75	0.78	5.98	3.76	2.22
	10	1.90	1.22	0.68	3.14	2.52	0.62	4.65	2.70	1.95
	20	0.52	0.35	0.17	1.27	0.90	0.37	2.59	1.58	1.01
	40	0.52	0.26	0.26	0.53	0.40	0.13	1.57	0.72	0.85
	60	0.50	0.20	0.30	0.49	0.30	0.19	1.26	0.51	0.75
	80	0.46	0.19	0.27	0.54	0.23	0.31	1.44	0.63	0.81
200	0.5	4.60	4.00	0.60	4.64	4.00	0.64	5.20	3.92	1.28
	1	4.51	4.00	0.51	4.46	4.00	0.46	5.22	4.00	1.22
	1.5	4.65	4.00	0.65	4.39	4.00	0.39	5.17	4.00	1.17
	2	4.51	4.00	0.51	4.49	4.00	0.49	5.13	4.00	1.13
	5	4.22	3.67	0.55	4.79	4.00	0.79	6.06	4.00	2.06
	10	2.53	2.12	0.41	4.39	3.75	0.64	6.01	3.79	2.22
	20	0.90	0.71	0.19	2.26	1.91	0.35	4.38	3.06	1.32
	40	0.37	0.26	0.11	0.89	0.73	0.16	2.17	1.59	0.58
	60	0.45	0.27	0.18	0.59	0.47	0.12	1.56	1.03	0.53
	80	0.17	0.10	0.07	0.42	0.30	0.12	1.23	0.75	0.48
400	0.5	4.46	4.00	0.46	4.33	4.00	0.33	5.43	4.00	1.43
	1	4.45	4.00	0.45	4.28	4.00	0.28	4.95	4.00	0.95
	1.5	4.24	4.00	0.24	4.20	4.00	0.20	4.86	4.00	0.86
	2	4.25	4.00	0.25	4.27	4.00	0.27	5.16	4.00	1.16
	5	4.35	3.99	0.36	4.28	4.00	0.28	5.59	4.00	1.59
	10	3.88	3.56	0.32	4.28	4.00	0.28	5.65	4.00	1.65
	20	1.74	1.52	0.22	4.05	3.69	0.36	5.31	3.94	1.37
	40	0.78	0.65	0.13	2.12	1.93	0.19	3.69	2.93	0.76
	60	0.51	0.35	0.16	0.90	0.81	0.09	2.17	1.86	0.31
	80	0.38	0.28	0.10	0.52	0.50	0.02	1.89	1.55	0.34
800	0.5	4.10	4.00	0.10	4.13	4.00	0.13	5.16	4.00	1.16
	1	4.13	4.00	0.13	4.14	4.00	0.14	4.66	4.00	0.66
	1.50	4.11	4.00	0.11	4.12	4.00	0.12	4.81	4.00	0.81
	2	4.15	4.00	0.15	4.19	4.00	0.19	4.82	4.00	0.82
	5	4.20	4.00	0.20	4.19	4.00	0.19	5.10	4.00	1.10
	10	4.11	3.99	0.12	4.11	4.00	0.11	5.11	4.00	1.11
	20	3.40	3.18	0.22	4.09	3.99	0.10	5.14	4.00	1.14
	40	1.12	1.08	0.04	3.56	3.46	0.10	4.80	3.95	0.85
	60	0.67	0.60	0.07	2.07	1.94	0.13	4.44	3.54	0.90
	80	0.33	0.27	0.06	0.99	0.91	0.08	3.20	2.73	0.47
Verdadero valor		4.00	4.00	0.00	4.00	4.00	0.00	4.00	4.00	0.00

**Tabla 4.2:** Resultados de LASSO inicial, LASSO ajustado y LARS $_m$  para el *Problema 1* con errores SAR.

$n$	$\sigma^2$	LASSO inicial			LASSO ajustado			LARS <sub>m</sub>		
		VARS	VP	FP	VARS	VP	FP	VARS	VP	FP
100	0.5	4.28	4.00	0.28	4.44	4.00	0.44	4.41	3.66	0.75
	1	4.32	3.99	0.33	4.32	4.00	0.32	4.97	4.00	0.97
	1.5	4.23	3.93	0.30	4.26	3.96	0.30	4.53	3.98	0.55
	2	4.38	3.85	0.53	4.43	3.91	0.52	4.50	3.91	0.59
	5	3.86	3.41	0.45	3.93	3.55	0.38	3.97	3.46	0.51
	10	3.09	2.74	0.35	3.31	2.87	0.44	3.62	2.99	0.63
	20	2.46	2.19	0.27	2.63	2.41	0.22	3.21	2.53	0.68
	40	1.53	1.32	0.21	1.67	1.47	0.20	2.53	1.94	0.59
	60	1.24	1.05	0.19	1.38	1.21	0.17	2.01	1.49	0.52
	80	0.92	0.69	0.23	1.02	0.86	0.16	1.94	1.38	0.56
200	0.5	4.35	4.00	0.35	4.36	4.00	0.36	4.36	3.57	0.79
	1	4.21	4.00	0.21	4.22	4.00	0.22	4.85	4.00	0.85
	1.5	4.26	4.00	0.26	4.28	4.00	0.28	4.71	4.00	0.71
	2	4.41	3.99	0.42	4.33	3.99	0.34	4.61	4.00	0.61
	5	4.12	3.71	0.41	4.26	3.84	0.42	4.53	3.87	0.66
	10	3.45	3.11	0.34	3.64	3.31	0.33	4.10	3.60	0.50
	20	2.95	2.66	0.29	3.21	2.81	0.40	3.73	3.07	0.66
	40	2.20	1.88	0.32	2.34	2.00	0.34	3.50	2.77	0.73
	60	1.83	1.56	0.27	1.89	1.71	0.18	2.80	2.22	0.58
	80	1.34	1.17	0.17	1.61	1.45	0.16	2.40	1.94	0.46
400	0.5	4.16	4.00	0.16	4.18	4.00	0.18	4.47	3.70	0.77
	1	4.32	4.00	0.32	4.27	4.00	0.27	4.64	4.00	0.64
	1.5	4.32	4.00	0.32	4.31	4.00	0.31	4.49	4.00	0.49
	2	4.18	4.00	0.18	4.20	4.00	0.20	4.52	4.00	0.52
	5	4.28	3.99	0.29	4.39	4.00	0.39	4.66	4.00	0.66
	10	3.92	3.78	0.14	3.96	3.79	0.17	4.66	3.96	0.70
	20	3.52	3.29	0.23	3.66	3.43	0.23	4.18	3.64	0.54
	40	2.72	2.56	0.16	2.86	2.78	0.08	3.80	3.18	0.62
	60	2.27	2.09	0.18	2.39	2.22	0.17	3.45	2.85	0.60
	80	2.10	1.89	0.21	2.20	2.06	0.14	2.97	2.46	0.51
800	0.5	4.18	4.00	0.18	4.20	4.00	0.20	4.27	3.64	0.63
	1	4.10	4.00	0.10	4.09	4.00	0.09	4.68	4.00	0.68
	1.5	4.16	4.00	0.16	4.23	4.00	0.23	4.48	4.00	0.48
	2	4.17	4.00	0.17	4.17	4.00	0.17	4.45	4.00	0.45
	5	4.19	4.00	0.19	4.21	4.00	0.21	4.58	4.00	0.58
	10	4.10	3.97	0.13	4.15	3.99	0.16	4.66	3.99	0.67
	20	3.81	3.70	0.11	3.97	3.79	0.18	4.52	3.85	0.67
	40	3.38	3.22	0.16	3.52	3.37	0.15	4.45	3.76	0.69
	60	2.61	2.54	0.07	2.99	2.88	0.11	3.86	3.29	0.57
	80	2.49	2.35	0.14	2.79	2.66	0.13	3.55	3.08	0.47
Verdadero valor		4.00	4.00	0.00	4.00	4.00	0.00	4.00	4.00	0.00

**Tabla 4.3:** Resultados de LASSO inicial, LASSO ajustado y LARS<sub>m</sub> para el *Problema 2* con errores CAR.

$n$	$\sigma^2$	LASSO inicial			LASSO ajustado			LARS $_m$		
		VARS	VP	FP	VARS	VP	FP	VARS	VP	FP
100	0.5	4.66	3.99	0.67	4.41	4.00	0.41	4.34	3.58	0.76
	1	4.53	3.95	0.58	4.39	3.99	0.40	5.01	4.00	1.01
	1.50	4.37	3.79	0.58	4.40	3.98	0.42	4.95	3.98	0.97
	2	4.05	3.60	0.45	4.44	3.95	0.49	4.40	3.90	0.50
	5	3.53	3.07	0.46	3.98	3.53	0.45	3.35	3.10	0.25
	10	2.78	2.39	0.39	3.26	2.90	0.36	3.31	2.72	0.59
	20	1.86	1.57	0.29	2.56	2.28	0.28	3.27	2.31	0.96
	40	1.17	0.94	0.23	1.90	1.59	0.31	2.58	1.79	0.79
	60	0.85	0.62	0.23	1.21	1.07	0.14	2.22	1.50	0.72
80	0.58	0.40	0.18	1.02	0.88	0.14	2.03	1.31	0.72	
200	0.5	4.32	4.00	0.32	4.39	4.00	0.39	4.59	3.86	0.73
	1	4.40	4.00	0.40	4.36	4.00	0.36	4.57	4.00	0.57
	1.5	4.30	3.94	0.36	4.26	4.00	0.26	4.59	4.00	0.59
	2	4.24	3.89	0.35	4.24	3.99	0.25	4.50	4.00	0.50
	5	3.68	3.41	0.27	4.25	3.90	0.35	4.36	3.81	0.55
	10	3.26	3.01	0.25	3.78	3.58	0.20	4.56	3.64	0.92
	20	2.36	2.08	0.28	3.35	3.09	0.26	3.77	3.06	0.71
	40	1.55	1.41	0.14	2.48	2.18	0.30	3.35	2.49	0.86
	60	1.01	0.90	0.11	1.87	1.75	0.12	2.65	2.03	0.62
80	0.77	0.67	0.10	1.63	1.48	0.15	2.37	1.82	0.55	
400	0.5	4.17	4.00	0.17	4.16	4.00	0.16	4.74	3.91	0.83
	1	4.31	4.00	0.31	4.29	4.00	0.29	4.60	4.00	0.60
	1.5	4.20	4.00	0.20	4.18	4.00	0.18	4.38	4.00	0.38
	2	4.25	4.00	0.25	4.26	4.00	0.26	4.77	4.00	0.77
	5	4.06	3.82	0.24	4.16	3.99	0.17	4.75	3.98	0.77
	10	3.58	3.34	0.24	4.02	3.83	0.19	4.73	3.86	0.87
	20	2.90	2.69	0.21	3.59	3.36	0.23	4.36	3.67	0.69
	40	2.00	1.93	0.07	2.96	2.81	0.15	3.91	3.15	0.76
	60	1.59	1.51	0.08	2.67	2.51	0.16	3.46	2.81	0.65
80	1.02	0.93	0.09	2.26	2.05	0.21	2.99	2.53	0.46	
800	0.5	4.19	4.00	0.19	4.15	4.00	0.15	4.63	3.90	0.73
	1	4.12	4.00	0.12	4.11	4.00	0.11	4.52	4.00	0.52
	1.5	4.14	4.00	0.14	4.11	4.00	0.11	4.43	4.00	0.43
	2	4.08	4.00	0.08	4.14	4.00	0.14	4.38	4.00	0.38
	5	4.10	3.97	0.13	4.15	4.00	0.15	4.75	4.00	0.75
	10	3.92	3.77	0.15	4.11	4.00	0.11	4.53	3.99	0.54
	20	3.46	3.30	0.16	3.93	3.80	0.13	4.60	3.93	0.67
	40	2.57	2.49	0.08	3.41	3.31	0.10	4.34	3.64	0.70
	60	2.33	2.15	0.18	3.15	2.91	0.24	4.02	3.38	0.64
80	1.81	1.72	0.09	2.79	2.68	0.11	3.72	3.27	0.45	
Verdadero valor		4.00	4.00	0.00	4.00	4.00	0.00	4.00	4.00	0.00

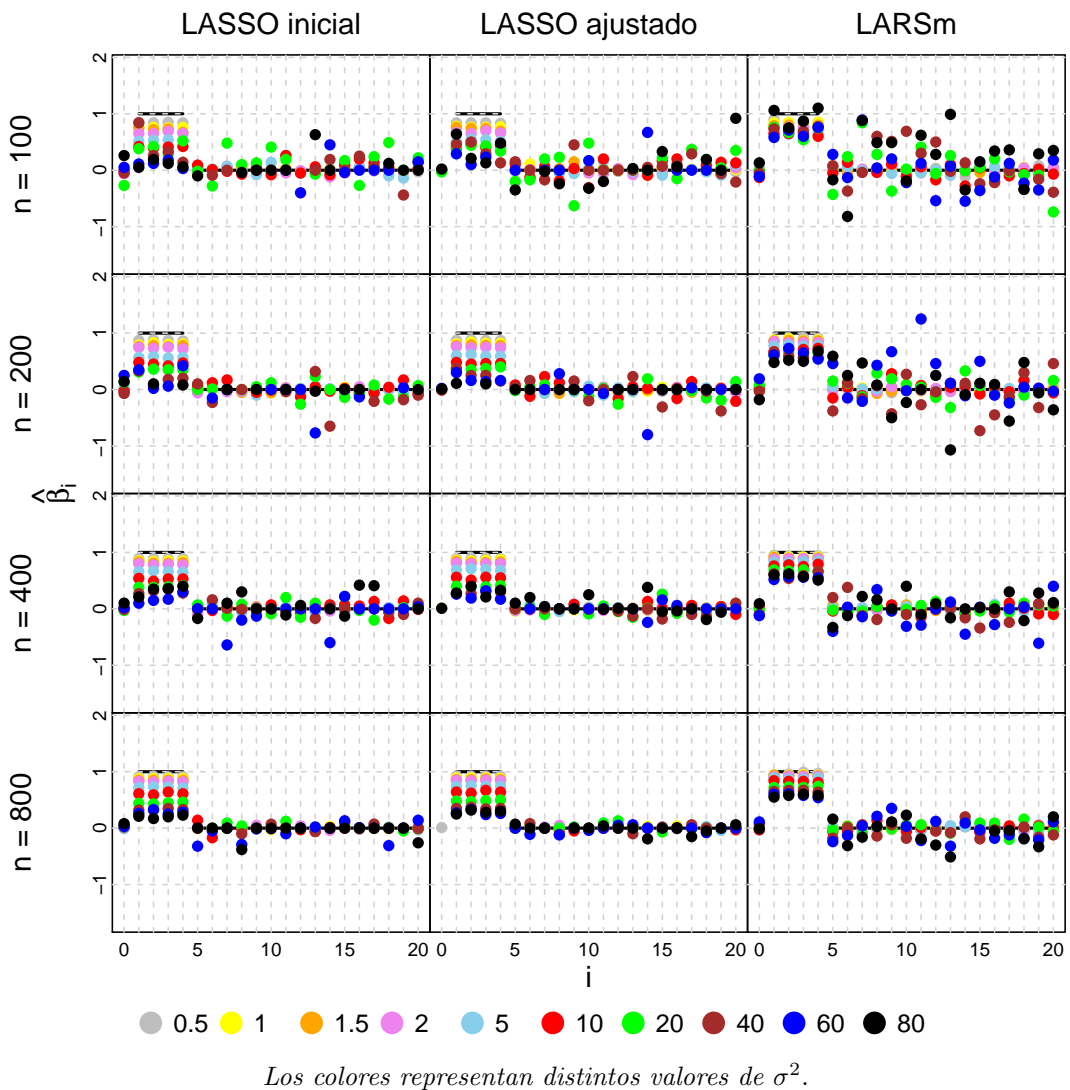
**Tabla 4.4:** Resultados de LASSO inicial, LASSO ajustado y LARS $_m$  para el *Problema 2* con errores SAR.

En lo que refiere a la estimación de  $\beta$ , al analizar las figuras también se observa que la estimación puntual tiene menor sesgo a medida que aumenta  $n$ , y para cada valor de  $n$  es mejor para valores pequeños de  $\sigma^2$ . El método LARS $_m$  estima mejor que sus pares a los verdaderos positivos ( $\beta_1$  a  $\beta_4$ ), pero en general estima peor a los falsos positivos ( $\beta_5$  a  $\beta_{20}$ ). El parámetro  $\beta_0 = 0$

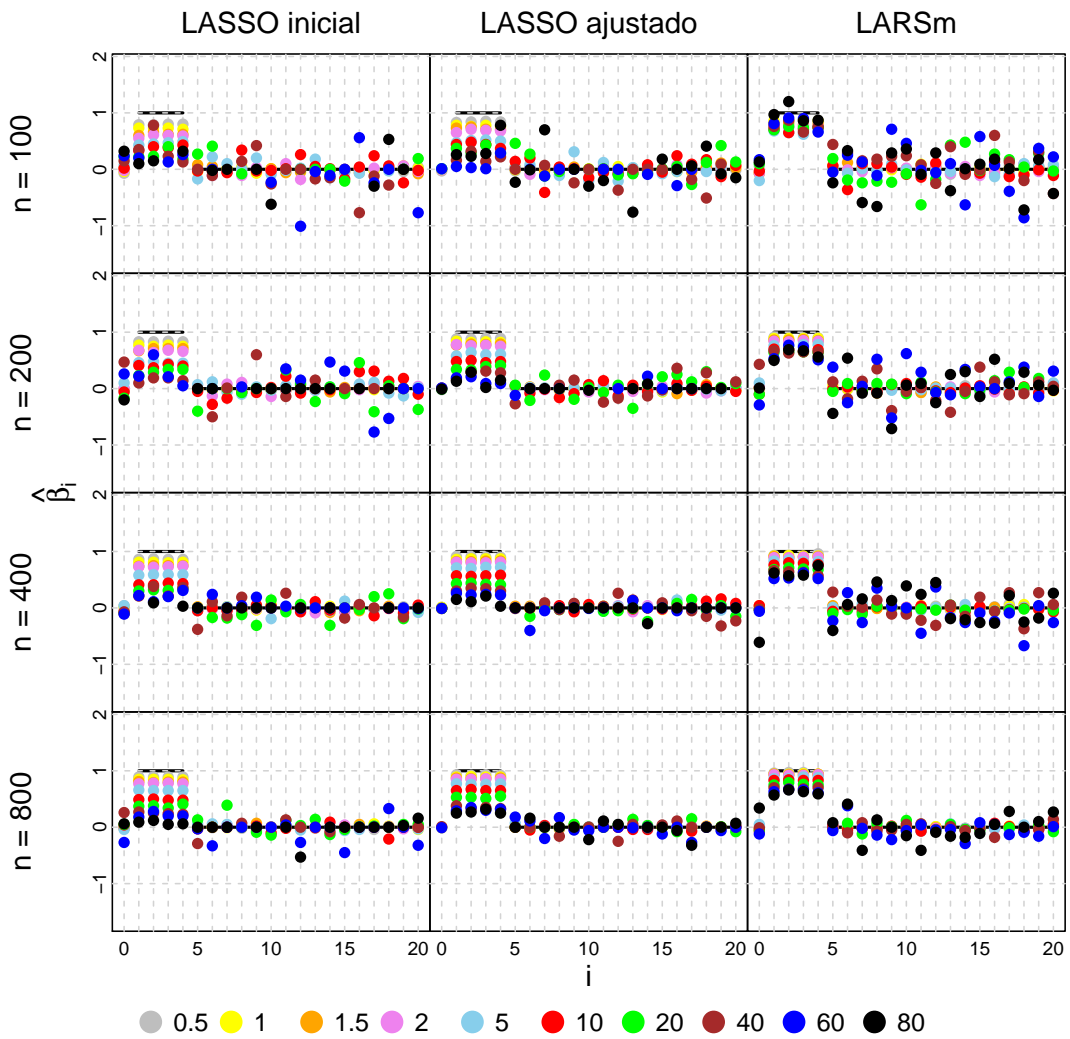
es estimado correctamente en todos los escenarios ( $\hat{\beta}_0 \approx 0$ ).

También se puede ver que CAR estima mejor que SAR, y que el *Problema 2* estima con menor sesgo a las variables que participan en el verdadero modelo que el *Problema 1*.

Los escenarios correspondientes al *Problema 2*, con  $n = 800$ , cualquier valor de  $\sigma^2$  y ambos modelos de error, visualmente parecen ser los que estiman mejor al vector  $\beta$ .

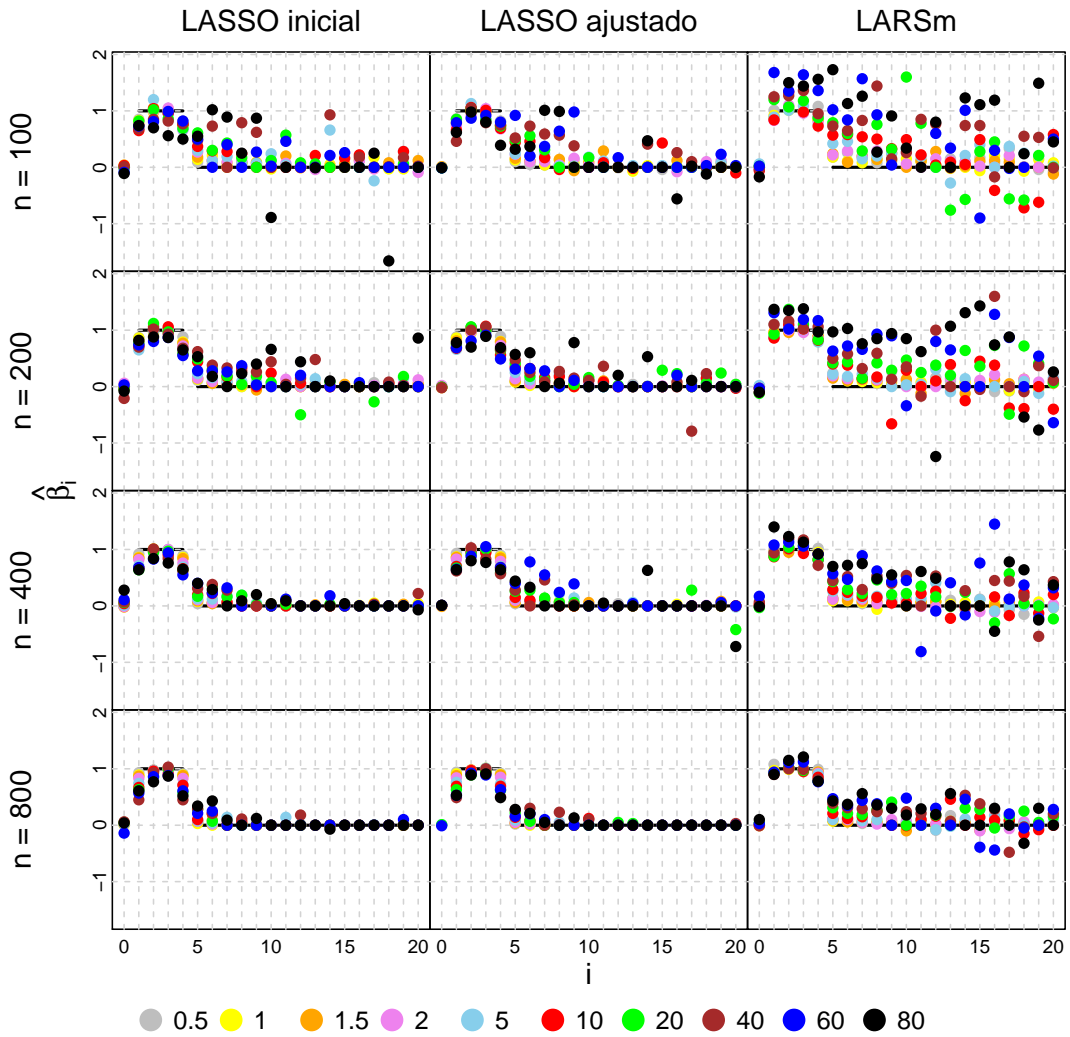


**Figura 4.1:**  $\hat{\beta}_i$   $0 \leq i \leq 20$  para el *Problema 1* con error CAR.



*Los colores representan distintos valores de  $\sigma^2$ .*

**Figura 4.2:**  $\hat{\beta}_i$   $0 \leq i \leq 20$  para el *Problema 1* con error SAR.

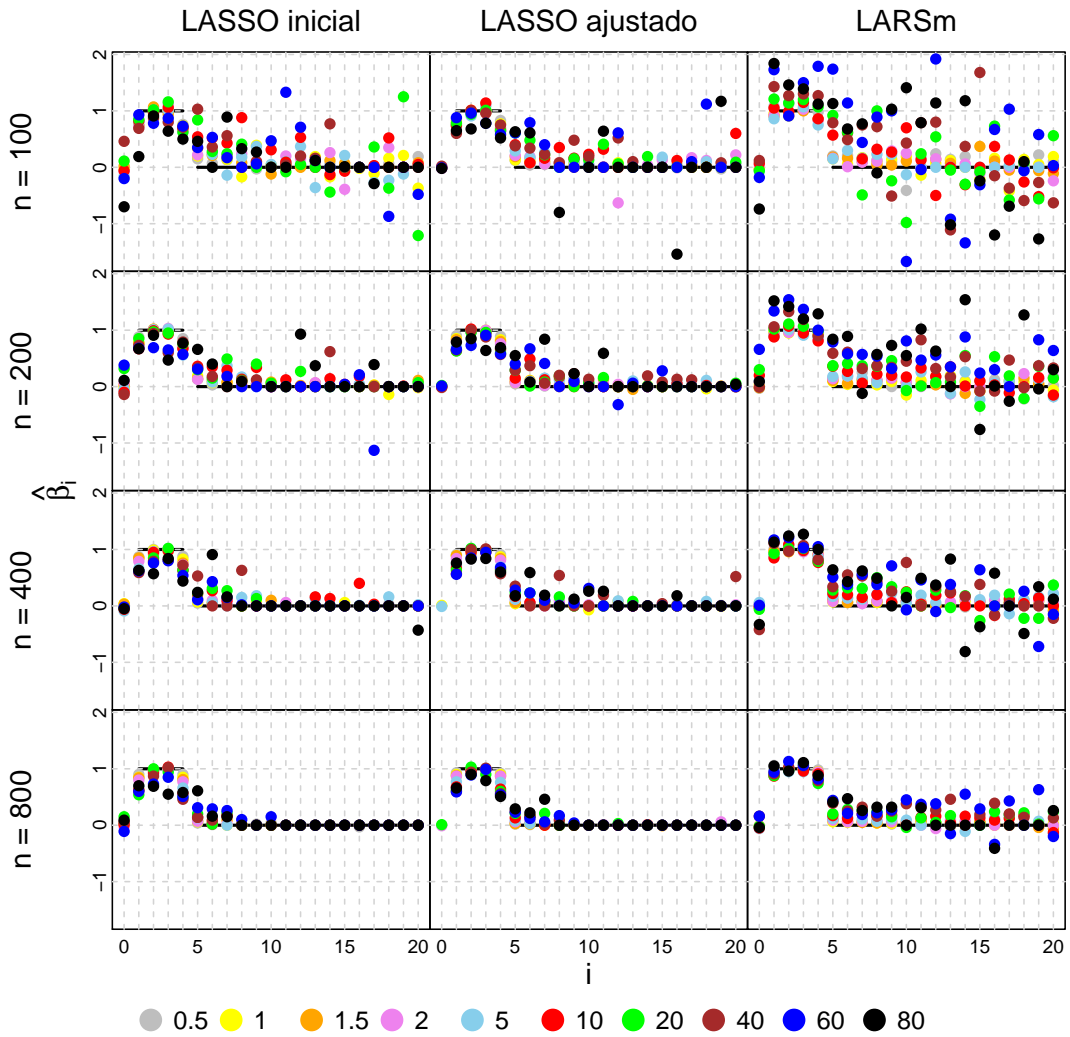


Los colores representan distintos valores de  $\sigma^2$ .

**Figura 4.3:**  $\hat{\beta}_i$   $0 \leq i \leq 20$  para el Problema 2 con error CAR.

1

<sup>1</sup>Algunos  $\hat{\beta}_i$  no figuran en los gráficos por exceder el rango utilizado:  $\hat{\beta}_{10} = 3.36$  en  $\text{LARS}_m$  para  $n = 100$  y  $\sigma^2 = 60$ ,  $\hat{\beta}_1 = 2.13$  ( $\text{LARS}_m$ ),  $\hat{\beta}_{12} = -2.57$ ,  $\hat{\beta}_{14} = 3.10$  y  $\hat{\beta}_{19} = 2.55$  en  $\text{LASSO inicial}$  para  $n = 100$  y  $\sigma^2 = 80$ .



Los colores representan distintos valores de  $\sigma^2$ .

**Figura 4.4:**  $\hat{\beta}_i$   $0 \leq i \leq 20$  para el *Problema 2* con error SAR.

1

Por último, al comparar las estimaciones puntuales de  $\theta$  y  $\sigma^2$ , se obtiene que en general todos los escenarios estiman razonablemente bien ambos parámetros, el sesgo de  $\hat{\theta}$  se reduce al aumentar  $n$ , pero no hay diferencias sustanciales al variar  $\sigma^2$ . El método que produce mejores estimaciones puntuales de estos parámetros es  $\text{LARS}_m$ .

<sup>1</sup>En  $\text{LARS}_m$ ,  $\hat{\beta}_{20}$  con  $n = 100$  y  $\sigma^2 = 80$  no figura en el gráfico por exceder el rango utilizado ( $\hat{\beta}_{20} = 2.57$ ).



$n$	$\theta$	$\sigma^2$	LASSO ajustado		LARS $_m$	
			$\hat{\theta}$	$\hat{\sigma}^2$	$\hat{\theta}$	$\hat{\sigma}^2$
100	0.9	0.5	0.85 (0.12)	0.58 (0.11)	0.62 (0.58)	1.23 (1.10)
	0.9	1	0.85 (0.12)	1.15 (0.20)	0.90 (0.13)	1.00 (0.16)
	0.9	1.5	0.86 (0.11)	1.75 (0.29)	0.89 (0.16)	1.50 (0.24)
	0.9	2	0.84 (0.12)	2.33 (0.41)	0.86 (0.14)	2.04 (0.34)
	0.9	5	0.80 (0.13)	5.93 (1.20)	0.86 (0.15)	5.01 (0.90)
	0.9	10	0.77 (0.13)	12.16 (2.64)	0.84 (0.14)	10.22 (1.96)
	0.9	20	0.79 (0.14)	22.85 (3.16)	0.83 (0.14)	21.28 (3.84)
	0.9	40	0.79 (0.14)	44.30 (6.78)	0.79 (0.16)	41.70 (6.65)
	0.9	60	0.76 (0.15)	64.71 (9.23)	0.83 (0.14)	60.64 (9.51)
	0.9	80	0.82 (0.16)	81.5 (11.79)	0.82 (0.14)	78.92 (13.93)
200	0.9	0.5	0.88 (0.10)	0.55 (0.05)	0.64 (0.51)	0.90 (0.65)
	0.9	1	0.87 (0.10)	1.13 (0.11)	0.90 (0.10)	1.01 (0.12)
	0.9	1.5	0.85 (0.11)	1.69 (0.17)	0.89 (0.10)	1.53 (0.18)
	0.9	2	0.83 (0.12)	2.23 (0.25)	0.90 (0.10)	1.98 (0.23)
	0.9	5	0.79 (0.14)	5.70 (0.65)	0.87 (0.11)	4.94 (0.54)
	0.9	10	0.79 (0.13)	11.45 (1.40)	0.83 (0.12)	9.90 (1.22)
	0.9	20	0.79 (0.13)	23.23 (2.53)	0.83 (0.13)	20.43 (2.52)
	0.9	40	0.76 (0.15)	43.93 (4.64)	0.81 (0.12)	41.06 (4.46)
	0.9	60	0.80 (0.13)	63.72 (6.57)	0.84 (0.12)	61.13 (6.68)
	0.9	80	0.80 (0.13)	84.68 (9.73)	0.82 (0.13)	82.05 (8.48)
400	0.9	0.5	0.88 (0.07)	0.55 (0.04)	0.65 (0.30)	0.89 (0.49)
	0.9	1	0.85 (0.09)	1.09 (0.09)	0.89 (0.08)	1.00 (0.08)
	0.9	1.5	0.86 (0.08)	1.62 (0.11)	0.89 (0.08)	1.52 (0.13)
	0.9	2	0.86 (0.09)	2.16 (0.17)	0.88 (0.09)	2.02 (0.14)
	0.9	5	0.82 (0.10)	5.46 (0.44)	0.85 (0.10)	5.10 (0.42)
	0.9	10	0.82 (0.10)	10.94 (0.77)	0.85 (0.10)	10.04 (0.77)
	0.9	20	0.83 (0.10)	21.94 (2.03)	0.87 (0.09)	20.09 (1.53)
	0.9	40	0.81 (0.09)	43.89 (3.08)	0.87 (0.09)	41.23 (3.33)
	0.9	60	0.84 (0.09)	64.38 (4.9)	0.85 (0.10)	62.14 (4.99)
	0.9	80	0.84 (0.09)	84.52 (5.57)	0.85 (0.09)	83.13 (6.62)
800	0.9	0.5	0.88 (0.08)	0.53 (0.03)	0.71 (0.23)	0.77 (0.38)
	0.9	1	0.88 (0.06)	1.07 (0.06)	0.90 (0.06)	1.01 (0.05)
	0.9	1.5	0.86 (0.06)	1.60 (0.09)	0.88 (0.06)	1.51 (0.08)
	0.9	2	0.85 (0.06)	2.14 (0.13)	0.89 (0.06)	2.02 (0.11)
	0.9	5	0.85 (0.08)	5.30 (0.33)	0.89 (0.06)	4.98 (0.25)
	0.9	10	0.86 (0.07)	10.57 (0.59)	0.89 (0.06)	10.03 (0.53)
	0.9	20	0.86 (0.07)	21.19 (1.18)	0.88 (0.06)	20.29 (1.10)
	0.9	40	0.87 (0.07)	42.70 (2.63)	0.88 (0.07)	40.35 (1.98)
	0.9	60	0.86 (0.06)	63.71 (3.20)	0.88 (0.07)	60.89 (3.09)
	0.9	80	0.87 (0.06)	84.28 (4.38)	0.89 (0.06)	81.07 (5.14)

**Tabla 4.5:** Parámetros espaciales estimados en LASSO ajustado y LARS $_m$  para el *Problema 1* con errores CAR.

$n$	$\theta$	$\sigma^2$	LASSO ajustado		LARS $_m$	
			$\hat{\theta}$	$\hat{\sigma}^2$	$\hat{\theta}$	$\hat{\sigma}^2$
100	0.9	0.5	0.78 (0.08)	0.71 (0.17)	0.65 (0.71)	1.89 (5.32)
	0.9	1	0.78 (0.10)	1.33 (0.28)	0.88 (0.08)	1.02 (0.17)
	0.9	1.5	0.75 (0.10)	2.09 (0.44)	0.86 (0.10)	1.57 (0.31)
	0.9	2	0.74 (0.10)	2.95 (0.83)	0.83 (0.11)	2.11 (0.39)
	0.9	5	0.74 (0.10)	7.47 (1.72)	0.83 (0.11)	5.44 (1.21)
	0.9	10	0.76 (0.11)	13.91 (2.33)	0.83 (0.10)	10.75 (2.28)
	0.9	20	0.80 (0.11)	24.63 (4.08)	0.83 (0.11)	22.02 (3.82)
	0.9	40	0.80 (0.11)	45.90 (6.37)	0.83 (0.11)	41.38 (6.55)
	0.9	60	0.84 (0.11)	65.32 (9.66)	0.82 (0.10)	61.59 (9.35)
	0.9	80	0.83 (0.10)	84.76 (13.34)	0.83 (0.10)	82.73 (12.89)
200	0.9	0.5	0.83 (0.05)	0.63 (0.09)	0.76 (0.22)	0.83 (0.52)
	0.9	1	0.80 (0.07)	1.30 (0.18)	0.87 (0.06)	1.05 (0.12)
	0.9	1.5	0.80 (0.08)	1.93 (0.30)	0.87 (0.06)	1.54 (0.15)
	0.9	2	0.81 (0.07)	2.53 (0.37)	0.86 (0.06)	2.03 (0.22)
	0.9	5	0.79 (0.07)	6.62 (1.04)	0.86 (0.07)	5.11 (0.57)
	0.9	10	0.81 (0.08)	12.98 (1.77)	0.87 (0.06)	10.19 (1.25)
	0.9	20	0.82 (0.07)	24.44 (2.81)	0.86 (0.07)	20.58 (2.67)
	0.9	40	0.84 (0.07)	45.40 (4.98)	0.86 (0.07)	41.74 (4.93)
	0.9	60	0.86 (0.07)	64.50 (6.55)	0.86 (0.07)	62.14 (7.24)
	0.9	80	0.84 (0.07)	85.50 (8.87)	0.86 (0.06)	81.65 (7.82)
400	0.9	0.5	0.84 (0.05)	0.6 (0.06)	0.82 (0.13)	0.69 (0.30)
	0.9	1	0.84 (0.05)	1.19 (0.12)	0.89 (0.04)	1.00 (0.08)
	0.9	1.5	0.84 (0.05)	1.83 (0.18)	0.88 (0.05)	1.52 (0.11)
	0.9	2	0.84 (0.05)	2.36 (0.24)	0.89 (0.04)	2.03 (0.15)
	0.9	5	0.84 (0.05)	5.94 (0.58)	0.89 (0.04)	5.04 (0.38)
	0.9	10	0.83 (0.05)	11.86 (1.18)	0.88 (0.05)	10.25 (0.85)
	0.9	20	0.84 (0.05)	23.94 (2.06)	0.88 (0.05)	20.17 (1.57)
	0.9	40	0.86 (0.05)	44.22 (3.60)	0.88 (0.04)	40.85 (3.34)
	0.9	60	0.87 (0.04)	64.66 (4.46)	0.88 (0.04)	63.13 (4.76)
	0.9	80	0.87 (0.04)	85.57 (5.69)	0.88 (0.05)	83.25 (6.15)
800	0.9	0.5	0.86 (0.03)	0.56 (0.04)	0.84 (0.08)	0.63 (0.15)
	0.9	1	0.87 (0.03)	1.13 (0.07)	0.89 (0.03)	1.01 (0.06)
	0.9	1.5	0.86 (0.03)	1.70 (0.12)	0.89 (0.03)	1.53 (0.09)
	0.9	2	0.86 (0.03)	2.26 (0.15)	0.89 (0.03)	2.01 (0.10)
	0.9	5	0.87 (0.03)	5.63 (0.40)	0.89 (0.03)	4.99 (0.23)
	0.9	10	0.86 (0.04)	11.30 (0.84)	0.90 (0.03)	9.97 (0.51)
	0.9	20	0.86 (0.04)	22.62 (1.68)	0.90 (0.03)	20.09 (1.11)
	0.9	40	0.87 (0.03)	44.06 (2.35)	0.89 (0.03)	40.60 (2.19)
	0.9	60	0.87 (0.03)	64.61 (3.21)	0.89 (0.03)	60.95 (3.17)
	0.9	80	0.88 (0.03)	84.74 (4.39)	0.89 (0.03)	81.40 (4.04)

**Tabla 4.6:** Parámetros espaciales estimados en LASSO ajustado y LARS $_m$  para el Problema 1 con errores SAR.

$n$	$\theta$	$\sigma^2$	LASSO ajustado		LARS $_m$	
			$\hat{\theta}$	$\hat{\sigma}^2$	$\hat{\theta}$	$\hat{\sigma}^2$
100	0.9	0.5	0.85 (0.14)	0.55 (0.09)	0.68 (0.31)	0.75 (0.34)
	0.9	1	0.84 (0.14)	1.11 (0.17)	0.86 (0.17)	0.97 (0.18)
	0.9	1.5	0.86 (0.12)	1.66 (0.32)	0.86 (0.17)	1.45 (0.23)
	0.9	2	0.88 (0.12)	2.19 (0.38)	0.84 (0.16)	2.00 (0.46)
	0.9	5	0.87 (0.11)	5.51 (0.95)	0.88 (0.14)	5.02 (0.87)
	0.9	10	0.86 (0.11)	11.66 (1.84)	0.85 (0.14)	10.30 (1.76)
	0.9	20	0.83 (0.13)	23.07 (3.60)	0.87 (0.13)	20.08 (3.32)
	0.9	40	0.81 (0.13)	46.64 (7.72)	0.84 (0.12)	40.09 (6.29)
	0.9	60	0.78 (0.14)	70.12 (11.26)	0.81 (0.13)	61.36 (10.10)
	0.9	80	0.80 (0.13)	88.13 (13.33)	0.82 (0.15)	80.48 (12.24)
200	0.9	0.5	0.87 (0.10)	0.53 (0.06)	0.74 (0.23)	0.72 (0.25)
	0.9	1	0.86 (0.11)	1.08 (0.12)	0.86 (0.14)	0.98 (0.12)
	0.9	1.5	0.89 (0.10)	1.61 (0.20)	0.87 (0.12)	1.48 (0.17)
	0.9	2	0.86 (0.11)	2.12 (0.23)	0.90 (0.10)	1.99 (0.21)
	0.9	5	0.86 (0.09)	5.55 (0.63)	0.89 (0.10)	5.07 (0.66)
	0.9	10	0.84 (0.11)	10.93 (1.28)	0.88 (0.11)	10.13 (1.17)
	0.9	20	0.82 (0.13)	22.06 (2.53)	0.84 (0.12)	20.01 (2.21)
	0.9	40	0.80 (0.13)	44.84 (4.60)	0.83 (0.12)	39.90 (4.04)
	0.9	60	0.80 (0.13)	66.33 (7.15)	0.83 (0.11)	60.17 (6.84)
	0.9	80	0.80 (0.14)	88.92 (10.21)	0.84 (0.13)	80.30 (8.35)
400	0.9	0.5	0.86 (0.09)	0.53 (0.04)	0.77 (0.21)	0.69 (0.26)
	0.9	1	0.88 (0.08)	1.07 (0.08)	0.90 (0.08)	1.00 (0.07)
	0.9	1.5	0.88 (0.08)	1.59 (0.12)	0.89 (0.08)	1.50 (0.12)
	0.9	2	0.87 (0.09)	2.14 (0.17)	0.89 (0.08)	2.01 (0.14)
	0.9	5	0.87 (0.09)	5.33 (0.44)	0.89 (0.08)	4.95 (0.39)
	0.9	10	0.84 (0.09)	10.87 (0.78)	0.89 (0.08)	10.04 (0.68)
	0.9	20	0.84 (0.10)	21.34 (1.61)	0.87 (0.09)	20.10 (1.56)
	0.9	40	0.83 (0.10)	43.50 (3.67)	0.87 (0.09)	40.32 (3.19)
	0.9	60	0.86 (0.09)	65.25 (4.39)	0.88 (0.09)	59.97 (4.41)
	0.9	80	0.83 (0.10)	86.68 (6.29)	0.87 (0.08)	79.83 (5.87)
800	0.9	0.5	0.89 (0.06)	0.53 (0.03)	0.77 (0.16)	0.70 (0.23)
	0.9	1	0.89 (0.05)	1.04 (0.06)	0.91 (0.05)	0.99 (0.05)
	0.9	1.5	0.89 (0.06)	1.57 (0.08)	0.89 (0.06)	1.49 (0.08)
	0.9	2	0.89 (0.06)	2.09 (0.11)	0.90 (0.06)	1.99 (0.09)
	0.9	5	0.87 (0.05)	5.28 (0.27)	0.89 (0.06)	4.97 (0.28)
	0.9	10	0.87 (0.07)	10.53 (0.55)	0.89 (0.06)	10.06 (0.54)
	0.9	20	0.87 (0.07)	21.19 (1.38)	0.88 (0.06)	19.99 (0.99)
	0.9	40	0.86 (0.06)	42.55 (2.35)	0.90 (0.06)	39.97 (2.46)
	0.9	60	0.86 (0.07)	64.18 (3.78)	0.89 (0.06)	59.74 (3.33)
	0.9	80	0.86 (0.06)	85.58 (4.41)	0.89 (0.06)	79.54 (3.71)

**Tabla 4.7:** Parámetros espaciales estimados en LASSO ajustado y LARS $_m$  para el Problema 2 con errores CAR.

$n$	$\theta$	$\sigma^2$	LASSO ajustado		LARS $_m$	
			$\hat{\theta}$	$\hat{\sigma}^2$	$\hat{\theta}$	$\hat{\sigma}^2$
100	0.9	0.5	0.82 (0.09)	0.62 (0.14)	0.76 (0.26)	1.02 (1.44)
	0.9	1	0.82 (0.09)	1.26 (0.24)	0.89 (0.10)	0.97 (0.15)
	0.9	1.5	0.82 (0.09)	1.94 (0.38)	0.90 (0.08)	1.46 (0.25)
	0.9	2	0.81 (0.08)	2.62 (0.56)	0.89 (0.09)	2.13 (0.82)
	0.9	5	0.77 (0.10)	6.48 (1.39)	0.82 (0.13)	6.32 (2.57)
	0.9	10	0.78 (0.11)	12.99 (3.19)	0.82 (0.13)	11.44 (2.72)
	0.9	20	0.76 (0.11)	26.36 (5.35)	0.83 (0.12)	20.36 (3.89)
	0.9	40	0.78 (0.10)	50.07 (9.85)	0.83 (0.11)	41.41 (6.80)
	0.9	60	0.80 (0.11)	73.51 (11.98)	0.85 (0.11)	62.38 (10.09)
	0.9	80	0.79 (0.11)	96.76 (14.46)	0.84 (0.09)	82.32 (14.09)
200	0.9	0.5	0.85 (0.07)	0.59 (0.07)	0.81 (0.14)	0.67 (0.32)
	0.9	1	0.85 (0.06)	1.21 (0.16)	0.89 (0.06)	0.98 (0.10)
	0.9	1.5	0.84 (0.06)	1.84 (0.24)	0.91 (0.05)	1.48 (0.17)
	0.9	2	0.83 (0.06)	2.45 (0.37)	0.89 (0.06)	1.99 (0.26)
	0.9	5	0.81 (0.07)	6.25 (0.82)	0.85 (0.07)	5.33 (1.27)
	0.9	10	0.80 (0.08)	12.51 (1.72)	0.86 (0.07)	10.13 (1.41)
	0.9	20	0.82 (0.07)	24.68 (3.49)	0.87 (0.07)	19.83 (2.40)
	0.9	40	0.81 (0.08)	49.01 (6.77)	0.87 (0.06)	39.61 (4.20)
	0.9	60	0.81 (0.08)	73.12 (8.71)	0.87 (0.06)	59.45 (6.18)
	0.9	80	0.82 (0.08)	94.84 (9.29)	0.87 (0.06)	81.12 (8.43)
400	0.9	0.5	0.86 (0.05)	0.57 (0.05)	0.85 (0.07)	0.61 (0.10)
	0.9	1	0.86 (0.04)	1.14 (0.10)	0.89 (0.04)	0.99 (0.07)
	0.9	1.5	0.85 (0.05)	1.72 (0.16)	0.88 (0.05)	1.50 (0.11)
	0.9	2	0.85 (0.05)	2.29 (0.22)	0.89 (0.04)	2.02 (0.15)
	0.9	5	0.84 (0.05)	5.78 (0.58)	0.88 (0.05)	5.08 (0.41)
	0.9	10	0.84 (0.04)	11.74 (1.07)	0.88 (0.05)	10.01 (0.90)
	0.9	20	0.85 (0.05)	23.53 (2.57)	0.89 (0.04)	19.90 (1.35)
	0.9	40	0.83 (0.05)	47.48 (4.47)	0.89 (0.05)	40.40 (2.97)
	0.9	60	0.85 (0.05)	69.88 (6.74)	0.89 (0.05)	60.49 (4.75)
	0.9	80	0.85 (0.05)	92.36 (7.17)	0.88 (0.05)	81.28 (6.50)
800	0.9	0.5	0.88 (0.03)	0.55 (0.04)	0.86 (0.07)	0.59 (0.10)
	0.9	1	0.87 (0.03)	1.13 (0.08)	0.90 (0.03)	1.00 (0.05)
	0.9	1.5	0.87 (0.03)	1.67 (0.12)	0.89 (0.03)	1.51 (0.08)
	0.9	2	0.86 (0.04)	2.23 (0.14)	0.90 (0.03)	2.01 (0.11)
	0.9	5	0.86 (0.04)	5.57 (0.43)	0.90 (0.03)	4.96 (0.25)
	0.9	10	0.87 (0.04)	11.06 (0.82)	0.90 (0.03)	9.99 (0.54)
	0.9	20	0.87 (0.03)	22.26 (1.45)	0.89 (0.03)	20.13 (1.09)
	0.9	40	0.87 (0.04)	44.77 (3.02)	0.89 (0.03)	40.03 (1.97)
	0.9	60	0.87 (0.03)	67.13 (4.42)	0.89 (0.03)	60.24 (3.15)
	0.9	80	0.87 (0.03)	89.91 (6.28)	0.90 (0.03)	80.13 (4.88)

**Tabla 4.8:** Parámetros espaciales estimados en LASSO ajustado y LARS $_m$  para el Problema 2 con errores SAR.

Para comparar los distintos métodos de selección de una forma cuantificable, se consideró la distancia euclídea entre los resultados obtenidos en las

simulaciones de los distintos escenarios y la situación óptima, es decir, la cantidad de verdaderos positivos y falsos positivos identificados en cada escenario versus la cantidad que debería haberse identificado respectivamente. Utilizando esta medida, se puede determinar en cada caso el método que menos se alejó de la situación óptima (menor distancia euclídea), considerando en simultáneo las variables seleccionadas y no seleccionadas.

De esta forma, **se obtiene que LASSO ajustado seleccionó globalmente mejor en el 46 % de los escenarios, mientras que LARS<sub>m</sub> lo hizo en el 33 % y LASSO inicial en el 21 % restante.** Estos resultados no se diferencian sustancialmente según el problema ni el valor de  $n$  considerado, pero sí de acuerdo al valor de  $\sigma^2$  y el tipo de error. En los escenarios con error de tipo CAR, LASSO ajustado y LARS<sub>m</sub> empatan siendo mejores en el 36 % de los casos (LASSO inicial es mejor en el 28 % restante), mientras que en los escenarios con error SAR, LASSO ajustado selecciona mejor en el 55 % de los casos, LARS<sub>m</sub> en el 30 % y LASSO inicial solamente en el 15 % restante.

Por otra parte, en escenarios con  $\sigma^2 = 0.5$  el método que selecciona mejor es LASSO inicial (66 % de los casos, versus 34 % LASSO ajustado, y en ninguno es mejor LARS<sub>m</sub>). En aquellos con  $1 \leq \sigma^2 \leq 20$  LASSO ajustado selecciona mejor (67 %, versus 24 % LASSO inicial y apenas 8 % LARS<sub>m</sub>) y en los que tienen  $\sigma^2 \geq 40$  es LARS<sub>m</sub> el método que mejor selecciona (94 % de los casos, frente a 6 % LASSO ajustado y 0 % LASSO inicial).

Es importante resaltar que los resultados del método LARS<sub>m</sub> son sensibles a la tolerancia ( $tol$ ) y la cantidad máxima de pasos consideradas. Los resultados presentados en este apartado consideran una tolerancia de 0.01 y una cantidad máxima de pasos por iteración igual a 100<sup>1</sup>. Ambos parámetros afectan la convergencia del estimador  $\hat{\eta}$ .

En las simulaciones presentadas anteriormente la cantidad de pasos  $m$  varía en función de los escenarios, es decir, en función de la combinación de parámetros utilizados. En general se obtiene que la cantidad de pasos no se ve afectada por el valor de  $n$  ni por el tipo de error. El parámetro  $\sigma^2$  afecta a la cantidad de pasos necesarios de una forma particular, ya que cuando  $\sigma^2 = 0.5$  la cantidad de pasos necesarios para obtener la convergencia es bastante superior al resto de los casos, e incluso en algunos casos no se obtiene la convergencia del

---

<sup>1</sup>Cada iteración termina cuando se obtuvo la convergencia ( $\|\hat{\eta}^{(m)} - \hat{\eta}^{(m-1)}\|_2 < tol$ ) o cuando se llegó al máximo de pasos estipulados ( $m = 100$ ).

método. Más allá de este caso, no se aprecia una relación clara entre  $\sigma^2$  y la cantidad de pasos necesarios.

Al considerar el problema también se notan diferencias en la cantidad de pasos, siendo levemente menor en el *Problema 1* (variables independientes) respecto al *Problema 2* (variables correlacionadas). En la Tabla 4.9 se presentan los resultados por Problema y modelo de error (excluyendo los casos en que  $\sigma^2 = 0.5$ ). En promedio, se requieren 6 pasos para obtener la convergencia de  $\hat{\eta}$ .

	Problema 1		Problema 2		Total
	CAR	SAR	CAR	SAR	
Mínimo	3	3	3	3	3
Media	4	5	5	7	6
Máximo	11	23	21	37	37

**Tabla 4.9:** Cantidad de pasos necesarios ( $m$ ) para obtener la convergencia en  $LARS_m$  por problema y tipo de error.

## Capítulo 5

# Aplicación con datos reales de la Encuesta Continua de Hogares

En este capítulo se aplican las metodologías propuestas a los microdatos de la ECH que elabora el Instituto Nacional de Estadística (INE) de Uruguay, correspondiente al año 2018. Esta encuesta, que se lleva a cabo de forma ininterrumpida desde el año 1968, brinda los indicadores oficiales del mercado laboral (actividad, empleo y desempleo) y de ingresos de los hogares y las personas del país. Es de tipo transversal, por lo que se aplica a cada hogar en una única oportunidad.

En este trabajo se utiliza una sub-base conformada por 933 hogares del medio rural ampliado (localidades de hasta 5.000 habitantes y zona rural dispersa) que no son propietarios de la vivienda ni del terreno que ocupan, y a su vez presentan al menos una necesidad básica insatisfecha (NBI)<sup>1</sup>. Además de los paquetes mencionados en el Capítulo 5, en este capítulo se utilizan shapefiles, sp y rgdal (ver [49], [41] y [3]).

Los microdatos de la ECH son de acceso libre y están disponibles en la página web del INE, sin embargo por razones de confidencialidad no se dispone de la georreferenciación de los hogares encuestados. Para poder utilizar esta base para análisis espacial, y por tratarse de un fin académico, se decidió imputar a cada hogar una ubicación lo más aproximada a la real que fuera posible. Esto se logró mediante la asignación aleatoria de ubicaciones dentro del área de la

---

<sup>1</sup>“Las NBI miden la falta de acceso de la población a determinados bienes y servicios que se consideran críticos para el desarrollo humano como son, el acceso a una vivienda decorosa, energía eléctrica, agua potable, servicios sanitarios, artículos de confort y acceso a la educación” (Fuente: [www.ine.gub.uy](http://www.ine.gub.uy)).

sección censal a la que pertenece cada hogar<sup>1</sup>.

Con la información disponible en la encuesta se busca estimar el ingreso per cápita de los hogares (sin valor locativo y sin servicio doméstico), expresado en Unidades Reajustables (UR), asumiendo a priori las siguientes premisas:

- el ingreso de los hogares no se distribuye equitativamente en el territorio, sino que hay zonas donde se concentran hogares de mayores ingresos y otras donde se concentran hogares de menores ingresos,
- existe autocorrelación espacial en la variable “ingreso per cápita del hogar” para los hogares rurales en cuestión, asumiendo que los hogares que están más próximos entre sí comparten algunas características como por ejemplo la actividad laboral que desempeñan sus integrantes.

Es importante aclarar que si bien la base utilizada pertenece a una muestra aleatoria que cumple todas las condiciones de representatividad del muestreo estadístico, no se utilizan ponderadores para expandir los resultados a toda la población por no ser el objetivo del presente trabajo. El objetivo que se persigue consiste en realizar la selección de variables en un contexto espacial utilizando un conjunto de datos reales, y no hacer inferencia sobre la población a partir de la muestra.

En la Figura 5.1 se presenta la ubicación espacial de los hogares considerados, identificando el cuartil de ingreso per cápita al que pertenecen. Como puede observarse, la dispersión de las observaciones es lo suficientemente grande como para cubrir todo el territorio, constantándose la presencia de hogares en todos los departamentos, con mayor concentración en el departamento de Canelones. Asimismo, se aprecia una mayor concentración de hogares de menor ingreso per cápita en los departamentos del noreste del país (Rivera, Cerro Largo y Treinta y Tres) y una mayor concentración de hogares de mayor ingreso per cápita en los departamentos del litoral y suroeste (Salto, Paysandú, Río Negro, Soriano, Colonia, Flores y Florida).

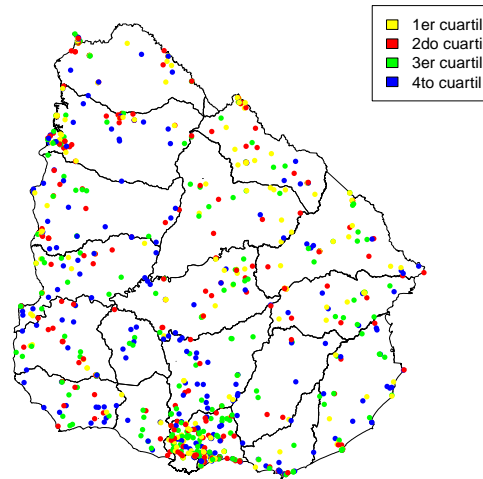
La estructura de vecindad considerada para estos datos (triangular) produce el grafo que se muestra en la Figura 5.2. De acuerdo a esta estructura, cada hogar tiene entre 3 y 12 hogares vecinos, siendo este guarismo en promedio igual a 5.94.

---

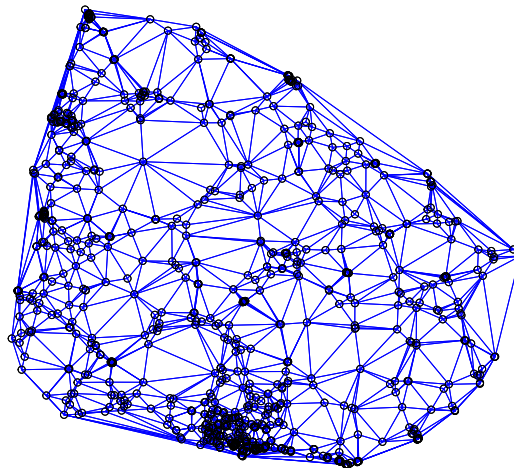
<sup>1</sup>La identificación del segmento y sección censal para los hogares del área rural ampliada no estaba disponible para todo público, por lo que fue necesario realizar una solicitud de información específica.



Las pruebas de autocorrelación espacial sobre la variable de interés (ingreso per cápita del hogar) se presentan en la Tabla (5.1).



**Figura 5.1:** Ubicación espacial de los hogares del medio rural ampliado de la ECH considerados, identificando el cuartil de ingreso per cápita al que pertenecen.



**Figura 5.2:** Grafo asociado a la estructura de vecindad triangular para la base ECH considerada.

Para las pruebas de autocorrelación se utilizan los índices de Moran y Geary definidos en el Capítulo 3 (definiciones (3.7) y (3.9)). La hipótesis nula consiste en la ausencia de autocorrelación espacial. El estadístico utilizado es  $\frac{I-E(I)}{\sqrt{Var(I)}}$ , donde  $I$  representa el índice de autocorrelación correspondiente. De acuerdo a la Proposición 5.4 de Gaetan and Guyon [19], este estadístico tiene asintóticamente una distribución normal estándar. Las distintas variantes de la Tabla (5.1) surgen de cómo se calcula la varianza de los índices: mediante randomización (remuestra), asumiendo normalidad (gaussiano) o realizando permutaciones (Monte Carlo). Para cada uno de los casos se presenta el estadístico y el p-valor correspondiente<sup>1</sup>. Se rechaza la hipótesis de ausencia de autocorrelación espacial en todos los casos.

Tipo	Moran test		Geary test	
	Estadístico	p-valor	Estadístico	p-valor
remuestra	4.5428	2.775e-06	3.9167	4.489e-05
gaussiano	4.5262	3.003e-06	4.7795	8.788e-07
Monte Carlo	-	0.0099	-	0.0099

**Tabla 5.1:** Pruebas de autocorrelación espacial para la variable a explicar: *ingreso per cápita del hogar*.

Por otro lado, se considera un conjunto de variables cuantitativas buscando que reflejen las principales características socioeconómicas del hogar así como también su nivel de confort. Estas variables son:

- $X_1$ : cantidad de integrantes del hogar,
- $X_2$ : proporción de perceptores de ingresos en el hogar<sup>2</sup>,
- $X_3$ : proporción de perceptores de ingresos que son de sexo masculino,
- $X_4$ : edad promedio de no perceptores de ingresos,
- $X_5$ : edad promedio de perceptores de ingresos de sexo masculino,
- $X_6$ : edad promedio de perceptores de ingresos de sexo femenino,
- $X_7$ : años de educación formal en promedio de los perceptores de ingresos,
- $X_8$ : años de educación formal en promedio de los perceptores de ingresos de sexo masculino,
- $X_9$ : años de educación formal en promedio de los perceptores de ingresos de sexo femenino,

<sup>1</sup>El test de permutaciones por ser no paramétrico, no provee un estadístico

<sup>2</sup>Perceptores de ingresos incluye ocupados, jubilados, pensionistas y rentistas.

- $X_{10}$ : proporción de integrantes del hogar que reciben prestaciones sociales,
- $X_{11}$ : edad promedio de los integrantes del hogar que reciben prestaciones sociales,
- $X_{12}$ : proporción de ocupados en el hogar,
- $X_{13}$ : edad promedio de los ocupados de sexo masculino,
- $X_{14}$ : edad promedio de los ocupados de sexo femenino,
- $X_{15}$ : proporción de los ocupados que son de sexo masculino,
- $X_{16}$ : años de educación formal en promedio de los ocupados,
- $X_{17}$ : años de educación formal en promedio de los ocupados de sexo masculino,
- $X_{18}$ : años de educación formal en promedio de los ocupados de sexo femenino,
- $X_{19}$ : promedio de horas trabajadas por semana entre los integrantes del hogar que están ocupados,
- $X_{20}$ : promedio de horas trabajadas por integrantes ocupados de sexo masculino,
- $X_{21}$ : promedio de horas trabajadas por integrantes ocupados de sexo femenino,
- $X_{22}$ : cantidad de necesidades básicas insatisfechas presentes en el hogar,
- $X_{23}$ : cantidad de elementos de confort presentes en el hogar,
- $X_{24}$ : cantidad de automóviles en el hogar,
- $X_{25}$ : cantidad de ciclomotores en el hogar,
- $X_{26}$ : cantidad de computadoras portátiles (excluyendo las del Plan Ceibal) en el hogar,
- $X_{27}$ : cantidad de aires acondicionados en el hogar,
- $X_{28}$ : cantidad de televisores a color en el hogar,
- $X_{29}$ : proporción de integrantes con celular en el hogar.

Es fácil percibir que por definición algunas de las variables explicativas tienen que estar correlacionadas entre sí, de todos modos esto se verifica estimando la matriz de correlaciones. Se obtienen 41 casos donde la correlación es mayor a 0.5 en valor absoluto, mientras que en 13 de estos es superior a 0.7 en valor absoluto. En la Tabla 5.2 se presentan las correlaciones que son mayores a 0.7 en valor absoluto. La mayor correlación se da entre las variables años de

educación formal de los perceptores de ingresos y años de educación formal de los ocupados, tanto para hombres como para mujeres y en el total ( $X_7$ ,  $X_8$  y  $X_9$  versus  $X_{16}$ ,  $X_{17}$  y  $X_{18}$  respectivamente). A su vez, el promedio de horas trabajadas en total ( $X_{19}$ ) está muy correlacionado con el promedio de horas trabajadas por hombres ( $X_{20}$ ).

Por su parte, la proporción de integrantes del hogar que reciben prestaciones ( $X_{10}$ ) está muy correlacionada con la edad promedio de estos ( $X_{11}$ ). Otras variables que también presentan correlación importante son la proporción de perceptores de ingreso que son hombres con la proporción de ocupados que son hombres ( $X_3$  y  $X_{15}$ ), la edad promedio de los hombres ocupados y la proporción de ocupados que son hombres ( $X_{13}$  y  $X_{15}$ ), los años de educación de los ocupados en total y de los hombres ( $X_{16}$  y  $X_{17}$ ), la proporción de ocupados que son hombres y el promedio de horas trabajadas por hombres ( $X_{15}$  y  $X_{20}$ ) y la edad promedio, los años de educación y las horas trabajadas de las mujeres ocupadas ( $X_{14}$ ,  $X_{18}$  y  $X_{21}$  respectivamente). Por otro lado, la proporción de perceptores de sexo masculino ( $X_3$ ) y la edad promedio de perceptores de sexo femenino ( $X_6$ ) tienen correlación alta pero negativa.

En el Apéndice 2 se presenta la matriz de correlaciones para las 29 variables consideradas.

	$X_6$	$X_{11}$	$X_{15}$	$X_{16}$	$X_{17}$	$X_{18}$	$X_{20}$	$X_{21}$
$X_3$	-0.70	.	0.72	.	.	.	.	.
$X_7$	.	.	.	0.85	.	.	.	.
$X_8$	.	.	.	.	0.89	.	.	.
$X_9$	.	.	.	.	.	0.89	.	.
$X_{10}$	.	0.80	.	.	.	.	.	.
$X_{13}$	.	.	0.77	.	.	.	.	.
$X_{14}$	.	.	.	.	.	0.76	.	0.77
$X_{15}$	.	.	.	.	.	.	0.74	.
$X_{16}$	.	.	.	.	0.75	.	.	.
$X_{18}$	.	.	.	.	.	.	.	0.73
$X_{19}$	.	.	.	.	.	.	0.84	.

**Tabla 5.2:** Correlaciones más altas entre variables explicativas.

Siguiendo el procedimiento para eliminar la dependencia espacial, se estima en primer lugar el LASSO inicial asumiendo que los datos fueran independientes. A continuación se aplica el procedimiento de eliminación de la autocorrelación espacial tanto para el modelo de errores CAR como SAR y luego se estima el LASSO ajustado.

Tanto para el modelo CAR como para SAR el valor de  $\lambda$  óptimo, estimado por validación cruzada dejando uno afuera ( $n\text{folds} = n$ ), es igual a 0.575 en el LASSO inicial y 0.082 en el LASSO ajustado.

Por otro lado, se aplica el método  $\text{LARS}_m$  para ambos modelos de error.

Para los tres métodos considerados se realizan pruebas de autocorrelación espacial sobre los residuos. En la Tabla 5.3 se muestran los resultados de dichas pruebas para el LASSO inicial, LASSO ajustado y  $\text{LARS}_m$  con errores CAR, y en la Tabla 5.4 lo mismo para errores SAR<sup>1</sup>.

Como puede apreciarse, en todos los casos se rechaza la hipótesis de ausencia de autocorrelación en los residuos del LASSO inicial y no se rechaza en los residuos del LASSO ajustado, es decir, el LASSO ajustado logró captar la estructura espacial subyacente en los datos para ambos modelos de error. Sin embargo, en el método  $\text{LARS}_m$  los residuos aún presentan estructura espacial, ya que en este caso se rechaza la hipótesis nula de ausencia de autocorrelación. Esto era esperable dado que este método no quita la autocorrelación, sino que estima los parámetros teniendo en cuenta esta autocorrelación.

Método	Tipo	Moran test		Geary test	
		Estadístico	p-valor	Estadístico	p-valor
LASSO inicial	remuestra	3.6689	0.0001	2.9659	0.0015
	gaussiano	3.6528	0.0001	3.7359	9.3e-05
	Monte Carlo	-	0.0099	-	0.0099
LASSO ajustado	remuestra	0.0760	0.4697	0.3335	0.3694
	gaussiano	0.0757	0.4698	0.4195	0.3374
	Monte Carlo	-	0.5248	-	0.4158
$\text{LARS}_m$	remuestra	3.2256	0.0006	2.5989	0.0047
	gaussiano	3.2139	0.0006	3.1635	0.0008
	Monte Carlo	-	0.0099	-	0.0099

**Tabla 5.3:** Pruebas de autocorrelación espacial para los residuos, modelo CAR.

Método	Tipo	Moran test		Geary test	
		Estadístico	p-valor	Estadístico	p-valor
LASSO inicial	remuestra	3.6689	0.0001	2.9659	0.0015
	gaussiano	3.6528	0.0001	3.7359	9.3e-05
	Monte Carlo	-	0.0099	-	0.0099
LASSO ajustado	remuestra	0.0818	0.4674	0.3327	0.3697
	gaussiano	0.0815	0.4675	0.4184	0.3378
	Monte Carlo	-	0.4455	-	0.3861
$\text{LARS}_m$	remuestra	3.8352	6.2e-05	3.0866	0.0010
	gaussiano	3.8186	6.7e-05	3.8805	5.2e-05
	Monte Carlo	-	0.0099	-	0.0099

**Tabla 5.4:** Pruebas de autocorrelación espacial para los residuos, modelo SAR.

<sup>1</sup>El LASSO inicial es idéntico para CAR y SAR por eso los resultados coinciden en el  $\lambda$  óptimo, en las pruebas de autocorrelación y en la estimación de parámetros.

En la Tabla 5.5 se presentan los valores estimados de los parámetros de acuerdo a los tres métodos y los dos modelos de error espacial considerados.

Parámetro	LASSO inicial	CAR		SAR	
		LASSO ajustado	LARS <sub>m</sub>	LASSO ajustado	LARS <sub>m</sub>
$\hat{\beta}_0$	6.14	0.64	-0.91	0.65	5.80
$\hat{\beta}_1$	-0.90	-0.80	-	-0.81	-0.31
$\hat{\beta}_2$	8.57	9.00	10.15	8.99	4.33
$\hat{\beta}_3$	-	-	-	-	0.17
$\hat{\beta}_4$	-	-	-	-	-
$\hat{\beta}_5$	-	-	-	-	0.01
$\hat{\beta}_6$	-	-	-	-	-
$\hat{\beta}_7$	0.01	0.005	-	0.01	-
$\hat{\beta}_8$	-	-	-	-	-
$\hat{\beta}_9$	-	-	-	-	-
$\hat{\beta}_{10}$	-0.21	-0.19	-	-0.19	-
$\hat{\beta}_{11}$	-	-	-	-	-
$\hat{\beta}_{12}$	0.51	0.36	6.10	0.37	5.46
$\hat{\beta}_{13}$	-	-	-	-	-
$\hat{\beta}_{14}$	-	-	-	-	-
$\hat{\beta}_{15}$	-	-	-	-	-
$\hat{\beta}_{16}$	-	-	-	-	-
$\hat{\beta}_{17}$	-	-	-	-	-
$\hat{\beta}_{18}$	-	-	-	-	-
$\hat{\beta}_{19}$	0.09	0.09	-	0.09	-
$\hat{\beta}_{20}$	-	-	-	-	-
$\hat{\beta}_{21}$	-	-	-	-	-
$\hat{\beta}_{22}$	-0.39	-0.35	-	-0.35	-
$\hat{\beta}_{23}$	0.06	0.06	-	0.06	-
$\hat{\beta}_{24}$	2.18	2.08	2.02	2.08	0.43
$\hat{\beta}_{25}$	-	-	-	-	-
$\hat{\beta}_{26}$	2.06	1.96	1.30	1.96	1.50
$\hat{\beta}_{27}$	-	-	-	-	-
$\hat{\beta}_{28}$	-	-	-	-	-
$\hat{\beta}_{29}$	1.20	1.32	5.86	1.32	3.65
$\hat{\theta}$	-	0.41	0.36	0.21	0.22
$\hat{\sigma}^2$	-	48.02	55.05	48.29	57.17

**Tabla 5.5:** Parámetros estimados para el LASSO inicial, LASSO ajustado y LARS<sub>m</sub>, con errores CAR y SAR.

En el LASSO inicial se seleccionan 11 variables de las 29 disponibles a priori, las variables seleccionadas son: la cantidad de integrantes ( $X_1$ ), la proporción de integrantes que perciben ingresos ( $X_2$ ), el promedio de años de educación formal de quienes perciben ingresos ( $X_7$ ), la proporción de integrantes que reciben prestaciones sociales ( $X_{10}$ ), la proporción de ocupados ( $X_{12}$ ), el promedio de horas trabajadas por semana ( $X_{19}$ ), la cantidad de necesidades básicas insatisfechas ( $X_{22}$ ), la cantidad de elementos de confort ( $X_{23}$ ), la cantidad de automóviles ( $X_{24}$ ), la cantidad de computadoras portátiles ( $X_{26}$ ) y la proporción de integrantes con celular ( $X_{29}$ ).

Con respecto al signo de los parámetros estimados en el LASSO inicial, se obtienen en general los resultados esperados; los parámetros asociados a la proporción de perceptores de ingreso, los años de educación formal de los perceptores, la proporción de ocupados, las horas trabajadas por semana y la cantidad de elementos de confort tienen signo positivo. Por otro lado, los parámetros asociados a la cantidad de personas en el hogar, la cantidad de integrantes que reciben prestaciones y la cantidad de necesidades básicas insatisfechas en el hogar tienen signo negativo.

En el LASSO ajustado, con ambos modelos de error se seleccionan las mismas variables que en el LASSO inicial. Además, todos los parámetros mantienen su signo respecto al método inicial, observándose sin embargo algunas variaciones en la magnitud de los mismos, principalmente en el parámetro  $\hat{\beta}_0$ .

En el método LARS $_m$ , se seleccionan 5 variables con el modelo CAR y 8 variables con el modelo SAR. Las variables seleccionadas con el modelo CAR para este método son: la proporción de integrantes que perciben ingresos ( $X_2$ ), la proporción de ocupados ( $X_{12}$ ), la cantidad de automóviles ( $X_{24}$ ), la cantidad de computadoras portátiles ( $X_{26}$ ) y la proporción de integrantes con celular ( $X_{29}$ ). Para el modelo SAR, además de estas variables, se agregan: la cantidad de integrantes del hogar ( $X_1$ ), la proporción de perceptores de ingresos que son de sexo masculino ( $X_3$ ) y la edad promedio de perceptores de ingresos de sexo masculino ( $X_5$ ). Con respecto al signo y magnitud de los parámetros estimados por este método, se obtiene que en el modelo CAR todos los parámetros son positivos (salvo  $\hat{\beta}_0$ ) y las variaciones más importantes en la magnitud ocurren en  $\hat{\beta}_0$  (se reduce),  $\hat{\beta}_{12}$  (aumenta) y  $\hat{\beta}_{29}$  (aumenta). Por otro lado, en el modelo SAR no hay cambios de signo respecto a los otros métodos, y los cambios más importantes en la magnitud tienen lugar en  $\hat{\beta}_2$  (se reduce) y en  $\hat{\beta}_{12}$  (aumenta). En ambos casos (CAR y SAR) se necesitaron 4 pasos para obtener la conver-

gencia en  $LARS_m$ .

Tanto LASSO ajustado como  $LARS_m$  dependen de valores iniciales de los parámetros. En LASSO ajustado se necesitan valores iniciales para la optimización de  $\hat{\theta}$  y  $\hat{\sigma}^2$ , mientras que en  $LARS_m$  se necesita un valor inicial para optimizar  $\hat{\eta}^{(0)} = (\hat{\beta}^{(0)}, \hat{\theta}^{(0)}, \hat{\sigma}^{2(0)})$ . En ambos casos estos valores son simulados, por lo que existe un componente aleatorio que puede afectar los resultados presentados anteriormente. Para poder evaluar la incidencia de estos valores iniciales en los resultados, se aplicó 100 veces cada método para cada modelo de error, cada vez con un valor inicial diferente. Se obtuvo como resultado principal que los valores estimados por LASSO ajustado no se ven afectados por los valores iniciales elegidos, mientras que en  $LARS_m$  sí varían de una iteración a otra. En las Tablas 5.6 y 5.7 se muestra la distribución de los valores estimados en dichas iteraciones con el método  $LARS_m$ . Para cada parámetro se consideran los cuartiles de su distribución siempre que sea distinto de cero, es decir, se considera el valor del parámetro solamente cuando la variable correspondiente es seleccionada.

Se puede observar que en algunos casos las estimaciones presentan una variabilidad considerable, incluso en algunos casos cambian de signo de una iteración a otra. También se pueden apreciar distintas situaciones en la proporción de veces que cada  $\hat{\beta}_i$  es seleccionado.

Los parámetros distintos de cero con mayor frecuencia son 5 en CAR y 6 en SAR. En CAR se trata de  $\hat{\beta}_2$  (proporción de perceptores de ingresos en el hogar),  $\hat{\beta}_{12}$  (proporción de ocupados en el hogar),  $\hat{\beta}_{24}$  (cantidad de automóviles),  $\hat{\beta}_{26}$  (cantidad de computadoras portátiles) y  $\hat{\beta}_{29}$  (proporción de integrantes que tienen celular). En SAR se agrega a la lista anterior  $\hat{\beta}_1$ , la cantidad de integrantes del hogar. A su vez, se destaca que estas variables fueron seleccionadas por los tres métodos considerados en la Tabla 5.5.

Por otro lado, hay 6 variables que fueron seleccionadas en LASSO ajustado con modelo CAR, pero no fueron seleccionadas en  $LARS_m$  (5 para el caso SAR). Al analizar en cuantas iteraciones de  $LARS_m$  fueron seleccionadas estas variables, se obtiene que en CAR  $X_1$  fue seleccionada en la mitad de las iteraciones,  $X_7$ ,  $X_{10}$  y  $X_{23}$  fueron seleccionadas en menos del 20% de las iteraciones, y  $X_{19}$  y  $X_{22}$  no fueron seleccionadas nunca. Lo mismo ocurre en SAR, donde a su vez la estimación de  $LARS_m$  presentada en la Tabla 5.5 selecciona a  $X_3$  y  $X_5$  (que no son seleccionados en LASSO). Para estas variables se obtiene al iterar el método que son seleccionadas solamente en 7 de las 100 iteraciones realizadas.



En los que respecta a  $\hat{\theta}$  y  $\hat{\sigma}^2$  no se observa una variabilidad considerable en las estimaciones.

Parámetro	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo	# veces $\neq 0$
$m$	3	4	5	5	7	100
$\hat{\beta}_0$	-3.70	-0.17	1.08	5.02	15.00	100
$\hat{\beta}_1$	-1.09	-0.61	-0.45	-0.30	-0.09	54
$\hat{\beta}_2$	0.58	6.84	9.85	11.72	17.13	99
$\hat{\beta}_3$	-2.48	-1.40	-0.67	0.71	2.09	9
$\hat{\beta}_4$	-	-	-	-	-	0
$\hat{\beta}_5$	0.002	0.002	0.003	0.004	0.007	4
$\hat{\beta}_6$	-	-	-	-	-	0
$\hat{\beta}_7$	0.03	0.03	0.03	0.03	0.03	1
$\hat{\beta}_8$	-	-	-	-	-	0
$\hat{\beta}_9$	-0.03	0.06	0.07	0.09	0.14	15
$\hat{\beta}_{10}$	-15.44	-4.60	-2.76	-1.00	4.28	19
$\hat{\beta}_{11}$	-0.04	-0.02	-0.01	-0.01	0.001	25
$\hat{\beta}_{12}$	0.34	3.11	3.94	5.64	9.06	85
$\hat{\beta}_{13}$	-	-	-	-	-	0
$\hat{\beta}_{14}$	0.001	0.003	0.01	0.01	0.02	9
$\hat{\beta}_{15}$	-3.82	-1.73	-1.15	-0.73	-0.36	6
$\hat{\beta}_{16}$	-	-	-	-	-	0
$\hat{\beta}_{17}$	-	-	-	-	-	0
$\hat{\beta}_{18}$	0.03	0.04	0.06	0.09	0.10	10
$\hat{\beta}_{19}$	-	-	-	-	-	0
$\hat{\beta}_{20}$	-	-	-	-	-	0
$\hat{\beta}_{21}$	0.01	0.01	0.01	0.01	0.01	3
$\hat{\beta}_{22}$	-	-	-	-	-	0
$\hat{\beta}_{23}$	0.03	0.08	0.11	0.21	0.24	6
$\hat{\beta}_{24}$	0.38	1.42	1.98	2.52	3.87	87
$\hat{\beta}_{25}$	-0.97	-0.97	-0.97	-0.97	-0.97	1
$\hat{\beta}_{26}$	0.36	1.44	2.13	2.63	4.51	79
$\hat{\beta}_{27}$	-0.49	0.74	1.44	1.98	2.88	9
$\hat{\beta}_{28}$	-	-	-	-	-	0
$\hat{\beta}_{29}$	0.58	3.76	4.69	5.97	9.67	98
$\hat{\theta}$	0.31	0.36	0.38	0.40	0.46	100
$\hat{\sigma}^2$	52.49	54.86	55.73	57.65	78.10	100

**Tabla 5.6:** Distribución de los valores estimados para cada parámetro de  $LARS_m$  en 100 iteraciones con errores CAR.

Parámetro	Mínimo	Primer cuartil	Mediana	Tercer cuartil	Máximo	# veces $\neq 0$
$m$	3	4	5	5	8	100
$\hat{\beta}_0$	-2.17	1.43	2.72	4.54	13.02	100
$\hat{\beta}_1$	-1.10	-0.55	-0.44	-0.32	-0.05	73
$\hat{\beta}_2$	3.25	6.41	8.69	10.71	18.46	100
$\hat{\beta}_3$	-0.97	-0.84	1.00	1.69	3.44	7
$\hat{\beta}_4$	-	-	-	-	-	0
$\hat{\beta}_5$	0.00002	0.002	0.004	0.01	0.02	7
$\hat{\beta}_6$	-	-	-	-	-	0
$\hat{\beta}_7$	0.03	0.04	0.05	0.06	0.07	2
$\hat{\beta}_8$	-0.00001	-0.00001	-0.00001	-0.00001	-0.00001	1
$\hat{\beta}_9$	0.00001	0.01	0.04	0.06	0.09	13
$\hat{\beta}_{10}$	-8.65	-5.88	-4.51	-1.54	5.73	16
$\hat{\beta}_{11}$	-0.04	-0.02	-0.01	-0.01	-0.003	31
$\hat{\beta}_{12}$	0.59	2.76	3.93	5.59	9.23	87
$\hat{\beta}_{13}$	0.01	0.01	0.01	0.01	0.02	3
$\hat{\beta}_{14}$	-0.0001	0.004	0.01	0.01	0.02	3
$\hat{\beta}_{15}$	-2.09	-1.13	-1.11	-0.53	0.41	5
$\hat{\beta}_{16}$	-	-	-	-	-	0
$\hat{\beta}_{17}$	0.07	0.07	0.07	0.07	0.07	1
$\hat{\beta}_{18}$	0.02	0.04	0.06	0.10	0.13	15
$\hat{\beta}_{19}$	-	-	-	-	-	0
$\hat{\beta}_{20}$	-	-	-	-	-	0
$\hat{\beta}_{21}$	0.02	0.03	0.03	0.03	0.03	2
$\hat{\beta}_{22}$	-	-	-	-	-	0
$\hat{\beta}_{23}$	0.00001	0.04	0.06	0.07	0.10	8
$\hat{\beta}_{24}$	0.40	1.10	1.68	2.34	3.62	87
$\hat{\beta}_{25}$	-1.16	-0.93	-0.70	-0.49	-0.28	3
$\hat{\beta}_{26}$	0.27	1.37	1.96	2.48	3.81	76
$\hat{\beta}_{27}$	-1.63	0.39	1.23	1.62	2.27	4
$\hat{\beta}_{28}$	-	-	-	-	-	0
$\hat{\beta}_{29}$	0.51	3.44	4.63	5.49	9.08	96
$\hat{\theta}$	0.17	0.19	0.20	0.21	1.14	100
$\hat{\sigma}^2$	52.55	54.26	55.87	57.30	72.42	100

**Tabla 5.7:** Distribución de los valores estimados para cada parámetro de LARS $_m$  en 100 iteraciones con errores SAR.

A los efectos de poder comparar los modelos estimados previamente, se

considera el pseudo  $R^2$  de Nagelkerke ([36]), el AIC y el BIC <sup>1</sup> para cada método y tipo de error. Para calcular estos indicadores se reestiman los modelos de acuerdo al siguiente criterio: para el LASSO inicial se estima un modelo de regresión lineal con las variables seleccionadas en el LASSO, mientras que tanto para el LASSO ajustado como para  $LARS_m$  se estiman modelos SAR y CAR con las variables seleccionadas en cada caso. Se utilizan los paquetes `spatialreg` y `rcompanion` (ver [5] y [35]).

El modelo estimado mediante LASSO ajustado (tanto CAR como SAR) obtiene el mayor pseudo  $R^2$ , y el menor valor tanto de AIC como de BIC respecto a los otros métodos.

Indicador	CAR			SAR	
	LASSO inicial	LASSO ajustado	$LARS_m$	LASSO ajustado	$LARS_m$
pseudo $R^2$	0.442	0.447	0.373	0.447	0.392
AIC	6277	6270	6375	6270	6352
BIC	6340	6338	6413	6338	6405

**Tabla 5.8:** Indicadores de bondad de ajuste de los modelos estimados.

El método LASSO (tanto el inicial como el ajustado por ambos modelos de error) seleccionó un tercio de las variables disponibles, mientras que el método  $LARS_m$  seleccionó 5 variables con el modelo SAR y 8 variables con el modelo CAR.

Salvo casos puntuales, en general no se aprecia un cambio sustancial en la estimación puntual de los parámetros por los tres métodos. La comparación de acuerdo a indicadores como el pseudo  $R^2$ , AIC y BIC sugiere que el LASSO ajustado tiene mejor ajuste con respecto a los demás.

Si bien todas las variables a priori podrían ser elegibles para explicar el ingreso per cápita de los hogares en el medio rural ampliado, se entiende que las seleccionadas por el LASSO conforman un subconjunto razonable de las que tienen mayor poder explicativo, aún en un contexto de variables que no son independientes entre sí.

El método  $LARS_m$  acota aún más el conjunto de variables seleccionadas por LASSO, pero tiene la desventaja de que los resultados son sensibles a los valores iniciales, la tolerancia y la cantidad máxima de pasos considerados.

<sup>1</sup>Por más detalles sobre estos indicadores consultar la Sección 2.1.

# Capítulo 6

## Conclusiones

En este trabajo se presentó una estrategia para llevar a cabo la selección de variables en un modelo de regresión lineal cuando los datos no son independientes, en particular cuando tienen una estructura espacial que es modelada a través de los errores aleatorios.

La estrategia consiste en usar LASSO para “eliminar” la dependencia (espacial) mediante la transformación del modelo original en uno equivalente donde los errores son independientes (LASSO ajustado). Cuando la matriz de covarianzas de los errores es estimada y se verifican las hipótesis del Teorema 2, esta estrategia permite estimar los parámetros LASSO de forma consistente en signo, es decir, identificando correctamente las variables que no pertenecen al modelo, las que sí pertenecen al modelo y para estas últimas estimando correctamente su signo.

Por otro lado se considera el método  $LARS_m$ , desarrollado en Zhu et al. [57], el cual consiste en estimar en simultáneo los parámetros tanto de la regresión como de la matriz de covarianzas. Cuando se cumplen las condiciones del Teorema 3 se asegura que el estimador identifica correctamente las variables que no participan del modelo, y para las variables que participan del modelo existe convergencia en distribución.

Ambas estrategias fueron evaluadas y comparadas, entre sí y respecto al LASSO tradicional, tanto en simulaciones como en una aplicación con datos reales. Para ambas estrategias fue necesario elaborar los algoritmos al inicio, sin referencia de las fuentes, los cuales quedan a disposición para poder ser utilizados en posteriores investigaciones.

En las simulaciones realizadas se obtuvo en general buenos resultados tanto en

el LASSO ajustado como en  $LARS_m$ , en comparación con estimar el método LASSO tradicional asumiendo que los datos fueran independientes. Con este último, en la mayoría de los casos no se logra identificar a las 4 variables que participan del verdadero modelo, salvo algunos escenarios en los que funcionan igualmente bien los tres métodos considerados.

El LASSO ajustado destaca en identificar los verdaderos positivos (variables que participan del verdadero modelo) sin seleccionar a su vez demasiados falsos positivos (variables que no participan del modelo) para la mayoría de los escenarios. Es el método que logra la mejor combinación entre verdaderos y falsos positivos en la mayoría de los casos. Presenta como desventaja que en algunos escenarios contrae la estimación de los parámetros (coeficientes) de los verdaderos positivos, situándolos por debajo de su verdadero valor. Además no logra identificar correctamente a las variables que participan del verdadero modelo en los escenarios con errores grandes (valores altos de  $\sigma^2$ ).

Por otro lado,  $LARS_m$  identifica una mayor cantidad de verdaderos positivos en los escenarios donde el error tiene un peso relativo mayor que las variables para conformar la variable dependiente, es decir, aquellos con valores grandes de  $\sigma^2$ . Además, estima con menor sesgo los verdaderos positivos y los parámetros asociados a la estructura espacial, con respecto a los otros. En contraparte se observa que estima peor a los falsos positivos.

En lo que refiere a las simulaciones, las posibilidades de explorar otros problemas son infinitas. A modo de ejemplo, podría ampliarse la dimensión del problema, variar la relación entre la cantidad de verdaderos y falsos positivos, considerar situaciones con mayor cantidad de variables que observaciones ( $p > n$ ), problemas donde no se verifique la condición de irrepresentabilidad, como también variar el valor de  $\theta$  que en nuestro caso se mantuvo fijo, etcétera. Asimismo, se probaron los tres métodos mencionados anteriormente a un conjunto de datos reales provenientes de la Encuesta Continua de Hogares, donde el objetivo era seleccionar las variables que explicaran mejor al ingreso per cápita de los hogares. El primer resultado a destacar es que en el LASSO ajustado y para ambos modelos de error (tanto CAR como SAR), el método logró captar la estructura de autocorrelación subyacente, lo que se demuestra al aplicar las pruebas de autocorrelación sobre los residuos.

De las 29 variables consideradas, tanto en el LASSO inicial como en el ajustado se seleccionan 11 de ellas (y son las mismas). Por otra parte, en  $LARS_m$  se seleccionan 5 variables en el modelo CAR y 8 variables en el modelo SAR.

En CAR las variables seleccionadas son un subconjunto de las seleccionadas en LASSO ajustado, mientras que en SAR hay dos variables seleccionadas en  $LARS_m$  que no habían sido seleccionadas en LASSO ajustado. Hay un conjunto de variables que son seleccionadas por los tres métodos considerando ambos modelos de error, se trata de la proporción de perceptores de ingresos en el hogar ( $X_2$ ), la proporción de ocupados ( $X_{12}$ ), la cantidad de automóviles ( $X_{24}$ ), la cantidad de computadoras portátiles ( $X_{26}$ ) y la proporción de integrantes con celular ( $X_{29}$ ).

Por su parte, debido a que el método  $LARS_m$  depende de un vector de parámetros iniciales que es aleatorio, presenta algunas variaciones en los resultados dependiendo de la inicialización. Para evaluar cuanto afecta esta variabilidad a los resultados se realizaron 100 iteraciones para cada tipo de error, variando los valores iniciales utilizados en cada caso. Los resultados mostraron que hay un conjunto de variables que son seleccionados con alta frecuencia, mientras que otras variables tienen pocas chances de ser seleccionadas.

Los métodos de selección de variables considerados para el caso de errores dependientes presentan ventajas y desventajas. A modo de resumen puede resaltarse lo siguiente:

- cuando los errores del modelo son dependientes no es conveniente aplicar el método LASSO tradicional,
- el Teorema 2 constituye un resultado teórico importante para proceder en este contexto,
- el procedimiento de eliminación de la dependencia espacial propuesto en este trabajo, bajo las condiciones detalladas en el Capítulo 3, representa una alternativa viable, con propiedades teóricas demostradas y eficacia probada de forma empírica mediante simulaciones y una aplicación real,
- el método  $LARS_m$  es efectivo para estimar los parámetros tanto del modelo como de la matriz de covarianzas en un contexto espacial. También es efectivo para identificar las variables que participan del verdadero modelo, pero requiere de una calibración precisa para mejorar la estimación de los falsos positivos.

Futuras investigaciones podrían incluir otros tipos de modelos para los errores, o incluso poder prescindir de tener que determinar a priori una estructura de error en particular, y también ampliar el orden de vecindad considerando otras estructuras de vecindad y pesos asociados.

# Referencias bibliográficas

- [1] Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on automatic control*, 19(6):716–723.
- [2] Akaike, H. (1998). *Selected Papers of Hirotugu Akaike*, chapter Information Theory and an Extension of the Maximum Likelihood Principle, pages 199–213. Springer New York.
- [3] Bivand, R., Keitt, T., and Rowlingson, B. (2019). *rgdal: Bindings for the “Geospatial” Data Abstraction Library*. R package version 1.4-8. <https://CRAN.R-project.org/package=rgdal>.
- [4] Bivand, R., Pebesma, E., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*, chapter Creating Neighbours, pages 239—251. Springer-Verlag, first edition.
- [5] Bivand, R. and Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, 63(18):1–36. <https://www.jstatsoft.org/v63/i18/>.
- [6] Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R*. Springer, New York, second edition. <http://www.asdar-book.org/>.
- [7] Botev, Z. and Belzile, L. (2020). *TruncatedNormal: Truncated Multivariate Normal and Student Distributions*. R package version 2.2. <https://CRAN.R-project.org/package=TruncatedNormal>.
- [8] Boyd, S. and Vandenberghe, L. (2008). *Subgradients*. Notes for EE364b, Stanford University, Winter 2006-07.
- [9] Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

- [10] Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. (1984). *Classification And Regression Trees*. Taylor & Francis Group.
- [11] Cai, L. and Maiti, T. (2020). Variable selection and estimation for high-dimensional spatial autoregressive models. *Scandinavian Journal of Statistics*, 47(2):587–607.
- [12] Canu, S. (2006). *Le LAR(s) et autres méthodes de régression parcimonieuse*. Presentación.
- [13] Chu, T., Zhu, J., and Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in Geostatistics. *The Annals of Statistics*, 39(5):2607–2625.
- [14] Cressie, N. (1993). *Statistics for Spatial Data, Revised Edition*. John Wiley & Sons, Inc.
- [15] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2):407–451.
- [16] Efron, B., Tibshirani, R., and Wainwright, M. (1996). *Mathematical Methods for Digital Computers*, chapter Multiple Regression Analysis. John Wiley, New York.
- [17] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22. <http://www.jstatsoft.org/v33/i01/>.
- [18] Fu, R., Thurman, A., Chu, T., Steen-Adams, M., and Zhu, J. (2013). On Estimation and Selection of Autologistic Regression Models via Penalized Pseudolikelihood. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):429–449.
- [19] Gaetan, C. and Guyon, X. (2010). *Spatial Statistics and Modeling*. Springer.
- [20] Giraldo, R. (2002). *Introducción a la Geoestadística. Teoría y Aplicación*. Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística, Bogotá, DC.
- [21] Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. CRC Press, first edition.



- [22] González, A. (2015). Selección de variables: Una revisión de métodos existentes. *Universidade da Coruña. Facultad de Informática*.
- [23] Goodchild, M. (1986). *Spatial Autocorrelation*. Geo Books, first edition.
- [24] Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., and Stadelmann, M. (2019). *expm: Matrix Exponential, Log, "etc"*. R package version 0.999-4. <https://CRAN.R-project.org/package=expm>.
- [25] Hastie, T. (2007). Least Angle Regression. *Stanford Statistics, Stanford University*. Presentación sobre el artículo homónimo.
- [26] Hastie, T. and Efron, B. (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 1.2. <https://CRAN.R-project.org/package=lars>.
- [27] Hastie, T., Tibshirani, R., and Wainwright, M. (2016). *Statistical Learning with Sparsity. The Lasso and Generalizations*. CRC Press, first edition.
- [28] Hoeting, J., Davis, R., Merton, A., and Thompson, S. (2006). Model selection for Geostatistical models. *Ecological Applications*, 16(1):87–98.
- [29] Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis*. Cambridge University Press, second edition.
- [30] Huang, H. and Chen, C. (2007). Optimal Geostatistical Model Selection. *Journal of the American Statistical Association*, 102(479):1009–1024.
- [31] Huang, H., Hsu, N., Theobald, D., and Breidt, F. (2009). Variable Selection and Model Choice in Geoaddivitive Regression Models. *Biometrics*, 65(2):626–634.
- [32] Huang, H., Hsu, N., Theobald, D., and Breidt, F. (2010a). Spatial Lasso with Applications to GIS Model Selection. *Journal of Computational and Graphical Statistics*, 19(4):963–983.
- [33] Huang, J., Horowitz, J., and Wei, F. (2010b). Variable selection in non-parametric additive models. *The Annals of Statistics*, 38(4):2282–2313.
- [34] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*, 28(5):1356–1378.

- [35] Mangiafico, S. (2021). *rcompanion: Functions to Support Extension Education Program Evaluation*. R package version 2.4.1. <https://CRAN.R-project.org/package=rcompanion>.
- [36] Nagelkerke, N. (1991). A note on a General Definition of the Coefficient of Determination. *Biometrika*, 78(3):691–692.
- [37] Nandy, S. (2016). High-dimensional variable selection for spatial regression and covariance estimation. A dissertation.
- [38] Nandy, S., Lim, C. Y., and Maiti, T. (2017a). Additive model building for spatial regression. *Journal of the Royal Statistical Society. Statistical Methodology. Series B.*, 79(3):779–800.
- [39] Nandy, S., Lim, C. Y., and Maiti, T. (2017b). *Supplementary material to “Additive Model Building for Spatial Regression”*.
- [40] Nesmachnow, S. and Iturriaga, S. (2019). Cluster-uy: Collaborative Scientific High Performance Computing in Uruguay. *Supercomputing. ISUM 2019. Communications in Computer and Information Science*, 1151:188–202.
- [41] Pebesma, E. and Bivand, R. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13. <https://CRAN.R-project.org/doc/Rnews/>.
- [42] Perrot-Dockès, M., Lévy-Leduc, C., Sansonet, L., and Chiquet, J. (2018). Variable selection in multivariate linear models with high-dimensional covariance matrix estimation. *Journal of Multivariate Analysis*, 166:78–97.
- [43] R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- [44] Reyes, P., Zhu, J., and Aukema, B. (2012). Selection of spatial-temporal lattice models: assessing the impact of climate conditions on a mountain pine beetle outbreak. *Journal of Agricultural, Biological and Environmental Statistics*, 17(3):508–525.
- [45] Riaño, M. (2018). Imputación de datos faltantes del Censo de Población y Vivienda utilizando técnicas de estadística espacial. Technical report, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración,

Universidad de la República, Uruguay. Serie Documentos de Trabajo, DT (18/1).

- [46] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.
- [47] Siabato, W. and Guzmán-Manrique, J. (2019). La autocorrelación espacial y el desarrollo de la geografía cuantitativa. *Cuadernos de Geografía: Revista Colombiana de Geografía*, 28:1–28.
- [48] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.
- [49] Stabler, B. (2013). *shapefiles: Read and Write ESRI Shapefiles*. R package version 0.7. <https://CRAN.R-project.org/package=shapefiles>.
- [50] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.
- [51] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B*, 73(3):273–282.
- [52] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. <http://www.stats.ox.ac.uk/pub/MASS4>.
- [53] Wang, H. and Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *The Canadian Journal of Statistics*, 37(4):607–624.
- [54] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67.
- [55] Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7(90):2541–2563.
- [56] Zhu, J., Huang, H., and Reyes, P. (2009). *Web-based Supplementary Materials for “On selection of spatial linear models for lattice data”*.

- [57] Zhu, J., Huang, H., and Reyes, P. (2010). On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society Series B*, 72(3):389–402.
- [58] Zhu, Z. and Y.Liu (2009). Estimating spatial covariance using penalised likelihood with weighted  $L_1$  penalty. *Journal of Nonparametric Statistics*, 21:925–942.
- [59] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.
- [60] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320.

# APÉNDICES

# Apéndice 1

## Demostraciones

*Proposición 1.* Definiendo  $\hat{u} = \hat{\beta} - \beta$  y desarrollando la función a minimizar evaluada en  $\hat{\beta}$ :

$$\begin{aligned} \|Y - \mathbf{X}\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 &= \|\mathbf{X}\beta + \varepsilon - \mathbf{X}\hat{\beta}\|_2^2 + \lambda\|\hat{u} + \beta\|_1 \\ &= \|\varepsilon - \mathbf{X}\hat{u}\|_2^2 + \lambda\|\hat{u} + \beta\|_1 \\ &= \|\varepsilon\|_2^2 - 2\hat{u}'\mathbf{X}'\varepsilon + \hat{u}'\mathbf{X}'\mathbf{X}\hat{u} + \lambda\|\hat{u} + \beta\|_1 \end{aligned}$$

por lo que  $\hat{u}$  puede pensarse como el vector que minimiza la función  $V_n(u)$ , definida como:

$$V_n(u) = -2(\sqrt{n}u)'W + (\sqrt{n}u)'\mathbf{C}^n(\sqrt{n}u) + \lambda\|u + \beta\|_1$$

con  $W = \frac{1}{\sqrt{n}}\mathbf{X}'\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{C})$  y  $\mathbf{C}^n$  definido en la sección anterior. Al derivar  $V_n$  respecto a  $u$  se obtiene:

$$\frac{\partial V_n(u)}{\partial u} = 2\sqrt{n}(\mathbf{C}^n(\sqrt{n}u) - W) + \lambda\partial\|u + \beta\|_1$$

Reescribiendo el resultado anterior matricialmente de acuerdo a (2.13), (2.14) y las definiciones de  $W_{A^*}$  y  $W_{A^{*c}}$ , se tiene:

$$\begin{aligned} \frac{\partial V_n(u)}{\partial u} &= 2\sqrt{n} \left( \sqrt{n} \begin{bmatrix} \mathbf{C}_{11}^n & \mathbf{C}_{12}^n \\ \mathbf{C}_{21}^n & \mathbf{C}_{22}^n \end{bmatrix} \begin{bmatrix} u_{A^*} \\ u_{A^{*c}} \end{bmatrix} - \begin{bmatrix} W_{A^*} \\ W_{A^{*c}} \end{bmatrix} \right) \\ &\quad + \lambda\partial \left\| \begin{bmatrix} u_{A^*} \\ u_{A^{*c}} \end{bmatrix} + \begin{bmatrix} \beta_{A^*} \\ \beta_{A^{*c}} \end{bmatrix} \right\|_1 \end{aligned} \quad (1.1)$$

Para que se cumpla que  $\hat{\beta} =_s \beta$ , deben verificarse las siguientes condiciones:

■ **Condición 1:**  $\hat{\beta}_{A^*} \neq \mathbf{0}$

El subvector de parámetros estimados correspondiente a los índices de las variables que participan del modelo verdadero,  $\hat{\beta}_{A^*} = \hat{u}_{A^*} + \beta_{A^*}$ , debe ser distinto de cero, para que el modelo estimado seleccione las variables correctas (las que tienen  $\beta_j \neq 0$ ).

■ **Condición 2:**  $\hat{\beta}_{A^{*c}} = \mathbf{0}$

El subvector de parámetros estimados correspondiente a los índices de las variables que no participan del modelo verdadero,  $\hat{\beta}_{A^{*c}} = \hat{u}_{A^{*c}} + \beta_{A^{*c}}$ , deben ser cero para que el modelo estimado no seleccione las variables que no participan del modelo verdadero (las que tienen  $\beta_j = 0$ , implica que  $\hat{u}_{A^{*c}} = \mathbf{0}$ , ya que  $\beta_{A^{*c}} = \mathbf{0}$ )

■ **Condición 3:**  $|\hat{\mathbf{u}}_{A^*}| \leq |\beta_{A^*}|$ , para asegurar que  $\text{sign}(\hat{\beta}_{A^*}) = \text{sign}(\beta_{A^*})$

Incorporando estas condiciones a (1.1), evaluado en  $u = \hat{u}$ , se obtienen dos resultados:

- **Resultado 1:**

$$\begin{aligned}
& 2\sqrt{n}(\underbrace{\sqrt{n}\mathbf{C}_{11}^n \hat{u}_{A^*} + \sqrt{n}\mathbf{C}_{12}^n \hat{u}_{A^{*c}}}_{=0 \text{ (cond. 2)}} - W_{A^*}) + \lambda \partial \|\underbrace{\hat{u}_{A^*} + \beta_{A^*}}_{\neq 0 \text{ (cond. 1)}}\|_1 = \\
& 2\sqrt{n}(\underbrace{\sqrt{n}\mathbf{C}_{11}^n \hat{u}_{A^*} - W_{A^*}}_{= \text{sign}(\beta_{A^*}) \text{ (cond. 3)}}) + \lambda \text{sign}(\hat{u}_{A^*} + \beta_{A^*}) = \mathbf{0} \\
& \Rightarrow \boxed{(\sqrt{n}\mathbf{C}_{11}^n \hat{u}_{A^*} - W_{A^*}) = -\frac{\lambda}{2\sqrt{n}} \text{sign}(\beta_{A^*})} \quad (1.2)
\end{aligned}$$

- **Resultado 2:**

$$\begin{aligned}
& 2\sqrt{n} \left( \underbrace{\sqrt{n}\mathbf{C}_{21}^n \hat{u}_{A^*} + \sqrt{n}\mathbf{C}_{22}^n \hat{u}_{A^{*c}}}_{=0 \text{ (cond. 2)}} - W_{A^{*c}} \right) + \lambda \partial \|\underbrace{\hat{u}_{A^{*c}} + \beta_{A^{*c}}}_{=0 \text{ (cond. 2)}}\|_1 = \\
& 2\sqrt{n} [\underbrace{\sqrt{n}\mathbf{C}_{21}^n \hat{u}_{A^*} - W_{A^{*c}}}_{z \in [-1,1]}] + \lambda z = \mathbf{0} \\
& -\frac{\lambda}{2\sqrt{n}} \leq [\sqrt{n}\mathbf{C}_{21}^n \hat{u}_{A^*} - W_{A^{*c}}] \leq \frac{\lambda}{2\sqrt{n}} \\
& \Rightarrow \boxed{|\sqrt{n}\mathbf{C}_{21}^n \hat{u}_{A^*} - W_{A^{*c}}| \leq \frac{\lambda}{2\sqrt{n}}} \quad (1.3)
\end{aligned}$$

A continuación se va a demostrar que si se cumplen  $A_n$  y  $B_n$  se verifican los dos resultados anteriores.

En primer lugar, partiendo del Resultado 1, premultiplicando por  $(\mathbf{C}_{11}^n)^{-1}$  y despejando, se obtiene:

$$(\mathbf{C}_{11}^n)^{-1}W_{A^*} = \sqrt{n}\hat{u}_{A^*} + \frac{\lambda}{2\sqrt{n}}(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})$$

A su vez, tomando valor absoluto y aplicando la condición 3, se obtiene:

$$\begin{aligned} |(\mathbf{C}_{11}^n)^{-1}W_{A^*}| &\leq \underbrace{\sqrt{n}|\hat{u}_{A^*}|}_{\leq |\beta_{A^*}|} + \sqrt{n} \left| \frac{\lambda}{2n}(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*}) \right| \\ &\leq \sqrt{n} \left( |\beta_{A^*}| + \left| \frac{\lambda}{2n}(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*}) \right| \right) \\ &= \underbrace{\sqrt{n}|\beta_{A^*}|}_{a_1} + \underbrace{\frac{\lambda}{2\sqrt{n}}|(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})|}_{b_1} \end{aligned}$$

Considerando que la desigualdad  $A_n$  puede reescribirse como  $A_n = \{|(\mathbf{C}_{11}^n)^{-1}W_{A^*}| < a_1 - b_1\}$  con  $a_1 \geq 0$  y  $b_1 \geq 0$ , entonces, si se verifica  $A_n$  se verifica el Resultado 1.

Por otro lado, partiendo del Resultado 2, se opera utilizando el Resultado 1:

$$\begin{aligned} |\underbrace{\mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} [\mathbf{C}_{11}^n \sqrt{n}\hat{u}_{A^*}]}_{=I} - W_{A^{*c}}| &= \left| \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \left[ W_{A^*} - \frac{\lambda}{2\sqrt{n}}\text{sign}(\beta_{A^*}) \right] - W_{A^{*c}} \right| \\ &\leq \underbrace{|\mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} W_{A^*} - W_{A^{*c}}|}_{a_2} + \frac{\lambda}{2\sqrt{n}} \underbrace{|\mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})|}_{b_2} \leq \frac{\lambda}{2\sqrt{n}} \end{aligned}$$

Como el conjunto  $B_n$  por definición verifica  $a_2 \leq \frac{\lambda}{2\sqrt{n}}(1 - b_2)$ , entonces si se cumple  $B_n$  se cumple el Resultado 2.

En resumen, si se cumple  $A_n$ , se verifica el Resultado 1 y si se cumple  $B_n$  se verifica el Resultado 2. Por lo que si se cumplen simultáneamente  $A_n$  y  $B_n$  se cumplen simultáneamente los resultados 1 y 2.

Cabe recordar que estos resultados surgen de la definición de consistencia en



signo (a través de las condiciones 1, 2 y 3), por lo que verificar los resultados 1 y 2 implica verificar la consistencia en signo.

Finalmente, hemos demostrado que es posible obtener una solución  $\hat{\beta}$  que verifique  $\hat{\beta} =_s \beta$ , ya que hemos demostrado que  $\{A_n \cap B_n\} \subset \{\hat{\beta} =_s \beta\}$ .  $\square$

*Teorema 1.* De la Proposición 1 se tiene:

$$P(\hat{\beta} =_s \beta^*) \geq P(A_n \cap B_n) = 1 - P(A_n^c \cup B_n^c) \geq 1 - P(A_n^c) - P(B_n^c)$$

por lo tanto, para demostrar el teorema es suficiente con acotar  $P(A_n^c)$  y  $P(B_n^c)$ .

Se tiene que:

$$P(A_n^c) = P\left(|(\mathbf{C}_{11}^n)^{-1}W_{A^*}| \geq \sqrt{n} \left(|\beta_{A^*}| - \frac{\lambda}{2n}|(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})|\right)\right)$$

y

$$P(B_n^c) = P\left(|\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}W_{A^*} - W_{A^{*c}}| \geq \frac{\lambda}{2\sqrt{n}} \left(1 - |\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})|\right)\right)$$

Bajo la condición de irrepresentabilidad se tiene que:

$$|\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})| \leq 1 - \delta$$

por lo tanto, el término  $1 - |\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})|$  puede acotarse como:

$$\underbrace{1 - |\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}\text{sign}(\beta_{A^*})|}_b \geq 1 - (1 - \delta) = \underbrace{\delta}_a$$

y por monotonía<sup>1</sup>, se obtiene que:

$$P(B_n^c) \leq P\left(|\mathbf{C}_{21}^n(\mathbf{C}_{11}^n)^{-1}W_{A^*} - W_{A^{*c}}| \geq \frac{\lambda}{2\sqrt{n}}\delta\right)$$

Ahora bien, definiendo:

---

<sup>1</sup>utilizando el argumento de que si  $a < b$ , entonces considerando una variable aleatoria  $X$ ,  $P(|X| \geq b) \leq P(|X| \geq a)$

- $\xi = (\xi_1, \dots, \xi_p)' = (\mathbf{C}_{11}^n)^{-1} W_{A^*} = \frac{1}{\sqrt{n}} (\mathbf{C}_{11}^n)^{-1} \mathbf{X}'_{A^*} \varepsilon = \mathbf{H}_A \varepsilon$
- $\zeta = (\zeta_1, \dots, \zeta_{p-p^*})' = \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} W_{A^*} - W_{A^{*c}} = \frac{1}{\sqrt{n}} (\mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \mathbf{X}'_{A^*} - \mathbf{X}'_{A^{*c}}) \varepsilon = \mathbf{H}_B \varepsilon$
- $b = (b_1, \dots, b_p)' = (\mathbf{C}_{11}^n)^{-1} \text{sign}(\beta_{A^*})$

entonces se tiene:

$$P(A_n^c) \leq \sup_{1 \leq j \leq p^*} P \left( |\xi_j| \geq \sqrt{n} \left( |\beta_j| - \frac{\lambda}{2n} |b_j| \right) \right)$$

y

$$P(B_n^c) \leq \sup_{1 \leq j \leq p-p^*} P \left( |\zeta_j| \geq \frac{\lambda}{2\sqrt{n}} \delta \right)$$

El término  $|b_j|$  puede acotarse de la siguiente forma:

$$|b_j| \leq \sum_{j=1}^{p^*} |b_j| \leq \sqrt{p^*} \left( \sum_{j=1}^{p^*} b_j^2 \right)^{1/2} = \sqrt{p^*} \|b\|_2$$

además, acotando el término  $\|b\|_2$ :

$$\begin{aligned} \|b\|_2 &= \|(\mathbf{C}_{11}^n)^{-1} \text{sign}(\beta_{A^*})\|_2 \leq \|(\mathbf{C}_{11}^n)^{-1}\|_2 \|\text{sign}(\beta_{A^*})\|_2 \\ &\leq \rho_{\max}((\mathbf{C}_{11}^n)^{-1}) \sqrt{p^*} \end{aligned}$$

donde la primera desigualdad se debe al Teorema 5.6.2 parte (b) de [29] que establece que  $\|Xy\|_2 \leq \|X\|_2 \|y\|_2$ , y la segunda desigualdad se desprende de las definiciones de ambas normas. El término  $\rho_{\max}(\mathbf{A})$  designa el mayor valor propio de la matriz  $\mathbf{A}$ .

A su vez, utilizando la hipótesis número 2 del teorema:

$$\rho_{\max}((\mathbf{C}_{11}^n)^{-1}) = \frac{1}{\rho_{\min}(\mathbf{C}_{11}^n)} \leq \frac{1}{M_2}$$

por lo que se deduce que:

$$|b_j| \leq \frac{p^*}{M_2}$$

sustituyendo  $|b_j|$  y  $|\beta_j|$  considerando la hipótesis número 4:

$$\begin{aligned}
P(A_n^c) &\leq \sup_{1 \leq j \leq p^*} P \left( |\xi_j| \geq \sqrt{n} \left( M_3 n^{-\frac{1-c^2}{2}} - \frac{\lambda p^*}{2n M_2} \right) \right) \\
&= \sup_{1 \leq j \leq p^*} P \left( |\xi_j| \geq M_3 n^{c_2/2} - \frac{p^* \lambda}{2M_2 \sqrt{n}} \right)
\end{aligned}$$

Por otro lado, como  $\varepsilon$  es un vector gaussiano centrado, con matriz de varianzas y covarianzas  $\sigma^2 \mathbf{I}$ , entonces el vector  $\xi$  es un vector gaussiano centrado con matriz de varianzas y covarianzas igual a  $\sigma^2 \mathbf{H}_A \mathbf{H}'_A$ , con:

$$\begin{aligned}
\mathbf{H}_A \mathbf{H}'_A &= \left( \frac{1}{\sqrt{n}} (\mathbf{C}_{11}^n)^{-1} \mathbf{X}'_{A^*} \right) \left( \frac{1}{\sqrt{n}} (\mathbf{C}_{11}^n)^{-1} \mathbf{X}_{A^*} \right)' \\
&= \frac{1}{n} (\mathbf{C}_{11}^n)^{-1} \mathbf{X}'_{A^*} \mathbf{X}_{A^*} (\mathbf{C}_{11}^n)^{-1} \\
&= (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{11}^n (\mathbf{C}_{11}^n)^{-1} \\
&= (\mathbf{C}_{11}^n)^{-1}
\end{aligned}$$

Entonces se tiene que la varianza de  $\xi_j$  puede acotarse, utilizando nuevamente la hipótesis número 2:

$$Var(\xi_j) = \sigma^2 (\mathbf{H}_A \mathbf{H}'_A)_{j,j} \leq \sigma^2 \rho_{max}(\mathbf{H}_A \mathbf{H}'_A) = \sigma^2 \rho_{max}((\mathbf{C}_{11}^n)^{-1}) \leq \frac{\sigma^2}{M_2}$$

y considerando una variable aleatoria  $Z$ , con distribución normal centrada y varianza  $\sigma^2$  ( $Z \sim N(0, \sigma^2)$ ), se obtiene que para todo  $1 \leq j \leq p^*$ :

$$P \left( |\xi_j| \geq M_3 n^{c_2/2} - \frac{p^* \lambda}{2M_2 \sqrt{n}} \right) \leq P \left( |Z| \geq \sqrt{M_2} \left( M_3 n^{c_2/2} - \frac{p^* \lambda}{2M_2 \sqrt{n}} \right) \right)$$

Por la desigualdad de Chernoff se sabe que  $\forall t \in \mathbb{R}$ ,  $P(|Z| > t) \leq 2 \exp \left( -\frac{t^2}{2\sigma^2} \right)$ , por lo que se deduce que para todo  $j \in \{1, \dots, p^*\}$ ,

$$P \left( |\xi_j| \geq M_3 n^{c_2/2} - \frac{p^* \lambda}{2M_2 \sqrt{n}} \right) \leq 2 \exp \left( -\frac{M_2}{2\sigma^2} \left( M_3 n^{c_2/2} - \frac{p^* \lambda}{2M_2 \sqrt{n}} \right)^2 \right)$$

Por las hipótesis número 3 y 6 se tiene que  $\frac{p^* \lambda}{\sqrt{n}} = o(n^{c_2/2})$  y sustituyéndolo

en el último término de la desigualdad anterior, se obtiene:

$$\begin{aligned} P\left(|\xi_j| \geq M_3 n^{c_2/2} - \frac{p^* \lambda}{2M_2 \sqrt{n}}\right) &\leq 2 \exp\left(-\frac{M_2}{2} \left(M_3 n^{c_2/2} - \frac{n^{c_2/2}}{2M_2}\right)^2\right) \\ &= 2 \exp\left(-n^{c_2/2} \frac{M_2}{2} \left(M_3 - \frac{1}{2M_2}\right)^2\right) \end{aligned}$$

que tiende a 0 cuando  $n$  tiende a  $\infty$ , por lo que queda demostrado que  $P(A_n^c) \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Resta demostrar que  $P(B_n^c)$  también tiende a 0 cuando  $n$  tiende a  $\infty$ . Se recuerda que:

$$P(B_n^c) \leq \sup_{1 \leq j \leq p-p^*} P\left(|\zeta_j| > \frac{\lambda}{2\sqrt{n}} \delta\right)$$

siendo  $\zeta = \mathbf{H}_B \varepsilon$ , por lo que  $\zeta$  es un vector gaussiano con esperanza 0 y matriz de varianzas y covarianzas igual a  $\sigma^2 \mathbf{H}_B \mathbf{H}_B'$ :

$$\begin{aligned} \mathbf{H}_B \mathbf{H}_B' &= \frac{1}{n} (\mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \mathbf{X}'_{A^*} - \mathbf{X}'_{A^{*c}}) (\mathbf{X}_{A^*} (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{12}^n - \mathbf{X}_{A^{*c}}) \\ &= \frac{1}{n} \left[ \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \underbrace{\mathbf{X}'_{A^*} \mathbf{X}_{A^*}}_{n\mathbf{C}_{11}^n} (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{12}^n - \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \underbrace{\mathbf{X}'_{A^*} \mathbf{X}_{A^{*c}}}_{n\mathbf{C}_{12}^n} \right. \\ &\quad \left. - \underbrace{\mathbf{X}'_{A^{*c}} \mathbf{X}_{A^*}}_{n\mathbf{C}_{21}^n} (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{12}^n + \underbrace{\mathbf{X}'_{A^{*c}} \mathbf{X}_{A^*}}_{n\mathbf{C}_{22}^n} \right] \\ &= \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{12}^n - \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{12}^n - \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{12}^n + \mathbf{C}_{22}^n \\ &= \mathbf{C}_{22}^n - \mathbf{C}_{21}^n (\mathbf{C}_{11}^n)^{-1} \mathbf{C}_{12}^n \\ &= \frac{1}{n} \mathbf{X}'_{A^{*c}} \mathbf{X}_{A^{*c}} - \frac{1}{n} \mathbf{X}'_{A^{*c}} \mathbf{X}_{A^*} \left( \frac{1}{n} \mathbf{X}'_{A^*} \mathbf{X}_{A^*} \right)^{-1} \frac{1}{n} \mathbf{X}'_{A^*} \mathbf{X}_{A^{*c}} \\ &= \frac{1}{n} \mathbf{X}'_{A^{*c}} (\mathbf{I} - \mathbf{X}_{A^*} (\mathbf{X}'_{A^*} \mathbf{X}_{A^*})^{-1} \mathbf{X}'_{A^*}) \mathbf{X}_{A^{*c}} \end{aligned}$$

Sabemos que  $\mathbf{X}_{A^*} (\mathbf{X}'_{A^*} \mathbf{X}_{A^*})^{-1} \mathbf{X}'_{A^*}$  es semidefinida positiva, por lo que los elementos de su diagonal son mayores o iguales a cero, y por ende,  $(\mathbf{I} - \mathbf{X}_{A^*} (\mathbf{X}'_{A^*} \mathbf{X}_{A^*})^{-1} \mathbf{X}'_{A^*})_{jj} \leq 0 \forall j \in A^*$ . Entonces, de acuerdo a lo ante-

rior y utilizando la hipótesis número 1, se obtiene para todo  $j$ :

$$\text{Var}(\zeta_j) = \sigma^2(\mathbf{H}_B \mathbf{H}'_B)_{jj} \leq \frac{\sigma^2}{n} (\mathbf{X}_{A^*c} \mathbf{X}'_{A^*c})_{jj} \leq \sigma^2 M_1$$

Utilizando este resultado en la desigualdad (1), y utilizando nuevamente Chernoff, se obtiene:

$$P\left(|\zeta_j| > \frac{\lambda}{2\sqrt{n}}\delta\right) \leq P\left(|Z| \geq \frac{\lambda}{2\sqrt{n}\sqrt{M_1}}\delta\right) \leq 2 \exp\left(-\frac{\delta^2}{2\sigma^2} \left(\frac{\lambda}{2\sqrt{n}\sqrt{M_1}}\right)^2\right)$$

Dado que  $\frac{\lambda}{\sqrt{n}} \rightarrow \infty$  cuando  $n \rightarrow \infty$ , se deduce que  $P(B_n^c) \rightarrow 0$  cuando  $n \rightarrow \infty$ , lo que concluye la prueba.  $\square$

*Proposición 2.* La demostración es idéntica a la de la Proposición (1) sustituyendo  $\mathbf{C}_{11}^n$ ,  $\mathbf{C}_{21}^n$ ,  $W_{A^*}$  y  $W_{A^*c}$  por  $\tilde{\mathbf{C}}_{11}^n$ ,  $\tilde{\mathbf{C}}_{21}^n$ ,  $\tilde{W}_{A^*}$  y  $\tilde{W}_{A^*c}$ .  $\square$

*Teorema 2.* De la Proposición 2 se tiene que:

$$P(\tilde{\beta} =_s \beta) \geq P(\tilde{A}_n \cap \tilde{B}_n) = 1 - P(\tilde{A}_n^c \cup \tilde{B}_n^c) \geq 1 - P(\tilde{A}_n^c) - P(\tilde{B}_n^c)$$

con  $\tilde{A}_n$  y  $\tilde{B}_n$  definidas en (3.5) y (3.6).

La demostración consiste en probar que  $P(\tilde{A}_n^c)$  y  $P(\tilde{B}_n^c)$  tienden a 0 cuando  $n \rightarrow \infty$ .

En primer lugar se considera  $P(\tilde{A}_n^c)$ :

$$P(\tilde{A}_n^c) = P\left(|(\tilde{\mathbf{C}}_{11}^n)^{-1}\tilde{W}_{A^*}| \geq \sqrt{n} \left(|\beta_{A^*}| - \frac{\lambda}{2n}|(\tilde{\mathbf{C}}_{11}^n)^{-1}\text{sign}(\beta_{A^*})|\right)\right)$$

Ahora bien, es posible descomponer  $((\tilde{\mathbf{C}}_{11}^n)^{-1}\tilde{W}_{A^*})$  y  $((\tilde{\mathbf{C}}_{11}^n)^{-1}\text{sign}(\beta_{A^*}))$  en los siguientes sumandos:

$$\begin{aligned} (\tilde{\mathbf{C}}_{11}^n)^{-1}\tilde{W}_{A^*} &= (\dot{\mathbf{C}}_{11}^n)^{-1}\dot{W}_{A^*} + (\dot{\mathbf{C}}_{11}^n)^{-1}(\tilde{W}_{A^*} - \dot{W}_{A^*}) + ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1})\dot{W}_{A^*} \\ &\quad + ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1})(\tilde{W}_{A^*} - \dot{W}_{A^*}) \end{aligned}$$

$$(\tilde{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}) = (\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}) + ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*})$$

Donde  $\dot{\mathbf{C}}_{11}^n = \frac{1}{n} \dot{\mathbf{X}}_{A^*}' \dot{\mathbf{X}}_{A^*}$  y  $\dot{\mathbf{X}}_{A^*} = \Sigma^{-1/2} \mathbf{X}_{A^*}$ .

Al sustituir los resultados anteriores en  $P(\tilde{A}_n^c)$  y utilizando la desigualdad triangular, se obtiene:

$$P(\tilde{A}_n^c) \leq PA_1 + PA_2 + PA_3 + PA_4 + PA_5$$

con

$$\begin{aligned} PA_1 &= P\left(\left|(\dot{\mathbf{C}}_{11}^n)^{-1} \dot{W}_{A^*}\right| \geq \frac{\sqrt{n}}{5} \left(|\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})|\right)\right) \\ PA_2 &= P\left(\left|(\dot{\mathbf{C}}_{11}^n)^{-1} (\tilde{W}_{A^*} - \dot{W}_{A^*})\right| \geq \frac{\sqrt{n}}{5} \left(|\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})|\right)\right) \\ PA_3 &= P\left(\left|((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \dot{W}_{A^*}\right| \geq \frac{\sqrt{n}}{5} \left(|\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})|\right)\right) \\ PA_4 &= P\left(\left|((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) (\tilde{W}_{A^*} - \dot{W}_{A^*})\right| \geq \frac{\sqrt{n}}{5} \left(|\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})|\right)\right) \\ PA_5 &= P\left(\left|\frac{\lambda}{2\sqrt{n}} \left|((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*})\right|\right| \geq \frac{\sqrt{n}}{5} \left(|\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})|\right)\right) \end{aligned}$$

En la demostración del Teorema 1 ya se probó que  $P(A_n^c) \rightarrow 0$  cuando  $n \rightarrow \infty$ , y dado que  $PA_1$  es casi idéntico a  $P(A_n^c)$ , salvo por el término  $\frac{1}{5}$  a la derecha de la desigualdad y porque  $PA_1$  se aplica a  $\dot{\mathbf{X}}_{A^*}$  en lugar de  $\mathbf{X}_{A^*}$ , por lo que basta con probar que  $\dot{\mathbf{X}}_{A^*}$  verifica las hipótesis impuestas a  $\mathbf{X}_{A^*}$  en el Teorema 1. Esto se verifica por las hipótesis 1, 2, 4 y 5, por lo que se concluye entonces que  $PA_1 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

A continuación se demostrará que los demás términos ( $PA_2$  a  $PA_5$ ) también tienden a 0.

El término  $PA_5$  puede acotarse teniendo en cuenta que la probabilidad de que todas las coordenadas cumplan la desigualdad es menor o igual a que cualquiera de ellas la cumpla por sí sola y que  $\left[|\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})|\right]_j \geq M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2}$ , con  $M_2$  constante mayor a cero (ver demostración del Teorema 1), por lo que vale lo siguiente:

$$\begin{aligned}
P \left( \left| ((\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} - (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1}) \text{sign}(\beta_{A^*}) \right| \geq \frac{2n}{5\lambda} \left( |\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \text{sign}(\beta_{A^*})| \right) \right) \\
\leq P \left( \left| \left( ((\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} - (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1}) \text{sign}(\beta_{A^*}) \right)_j \right| \geq \frac{2n}{5\lambda} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right)
\end{aligned}$$

Sea  $U = (\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} - (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1}$  y  $s = \text{sign}(\beta_{A^*})$ , entonces por la desigualdad de Cauchy-Schwarz se tiene que para todo  $j \in A^*$ ,

$$|(Us)_j| < \max_{i \in A^*} |(Us)_i| = \|Us\|_\infty \leq \|Us\|_2 \leq \|U\|_2 \|s\|_2 = \|U\|_2 \sqrt{p^*} \quad (1.4)$$

donde la segunda desigualdad se debe a que para todo vector  $x$ ,  $\|x\|_2^2 = \sum_i x_i^2 \geq \max_i (x_i^2) = \|x\|_\infty^2$ , mientras que la tercera desigualdad se debe al Teorema 5.6.2 de Horn and Johnson [29], parte (b). A su vez,

$$\begin{aligned}
\left\| (\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} - (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \right\|_2 &= \left\| (\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} (\dot{\mathbf{C}}_{11}^{\mathbf{n}} - \tilde{\mathbf{C}}_{11}^{\mathbf{n}}) (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \right\|_2 \\
&\leq \left\| (\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \right\|_2 \left\| (\dot{\mathbf{C}}_{11}^{\mathbf{n}} - \tilde{\mathbf{C}}_{11}^{\mathbf{n}}) \right\|_2 \left\| (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \right\|_2 \\
&\leq \frac{\left\| \dot{\mathbf{C}}_{11}^{\mathbf{n}} - \tilde{\mathbf{C}}_{11}^{\mathbf{n}} \right\|_2}{\rho_{\min}(\tilde{\mathbf{C}}_{11}^{\mathbf{n}}) \rho_{\min}(\dot{\mathbf{C}}_{11}^{\mathbf{n}})}
\end{aligned}$$

Por la hipótesis 2,  $\rho_{\min}(\dot{\mathbf{C}}_{11}^{\mathbf{n}}) > M_2$ . Por otro lado:

$$\begin{aligned}
\left\| \dot{\mathbf{C}}_{11}^n - \tilde{\mathbf{C}}_{11}^n \right\|_2 &= \left\| \frac{1}{n} \mathbf{X}'_{A^*} (\hat{\Sigma}^{-1} - \Sigma^{-1}) \mathbf{X}_{A^*} \right\|_2 \\
&= \gamma \left( \frac{1}{n^2} \mathbf{X}'_{A^*} (\hat{\Sigma}^{-1} - \Sigma^{-1}) \mathbf{X}_{A^*} \mathbf{X}'_{A^*} (\hat{\Sigma}^{-1} - \Sigma^{-1}) \mathbf{X}_{A^*} \right)^{1/2} \\
&= \gamma \left( \frac{1}{n^2} \mathbf{X}'_{A^*} \mathbf{X}_{A^*} (\hat{\Sigma}^{-1} - \Sigma^{-1}) (\hat{\Sigma}^{-1} - \Sigma^{-1}) \mathbf{X}_{A^*} \mathbf{X}'_{A^*} \right)^{1/2} \\
&\leq \gamma \left( \frac{1}{n} \mathbf{X}'_{A^*} \mathbf{X}_{A^*} \right) \gamma \left( \hat{\Sigma}^{-1} - \Sigma^{-1} \right) \\
&\leq \gamma \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right) \gamma \left( \hat{\Sigma}^{-1} - \Sigma^{-1} \right) \\
&\leq \left\| \frac{1}{n} \mathbf{X}' \mathbf{X} \right\|_\infty \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_\infty \\
&\leq M_4 \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_\infty
\end{aligned} \tag{1.5}$$

donde  $\gamma(\mathbf{X})$  representa el radio espectral de  $\mathbf{X}$ :  $\gamma(\mathbf{X}) = \max_{1 \leq i \leq n} (|\sqrt{\rho_i(\mathbf{X})}|)$ , con  $\rho_i(\mathbf{X})$  igual al  $i$ -ésimo valor propio de  $\mathbf{X}$ .

La tercera igualdad se debe al Teorema 1.3.22, la segunda desigualdad se debe al Teorema 4.3.28, la tercera desigualdad se debe al Teorema 5.6.9 de Horn and Johnson [29] y la cuarta igualdad se debe a la hipótesis 7.

Por su parte,  $\rho_{\min}(\tilde{\mathbf{C}}_{11}^n) = \min_{\|u\|_2=1} (\frac{1}{n} u' \mathbf{X}'_{A^*} \hat{\Sigma}^{-1} \mathbf{X}_{A^*} u)$  puede acotarse teniendo en cuenta lo siguiente:

$$\frac{1}{n} u' \mathbf{X}'_{A^*} \hat{\Sigma}^{-1} \mathbf{X}_{A^*} u = \frac{1}{n} u' \mathbf{X}'_{A^*} (\hat{\Sigma}^{-1} - \Sigma^{-1}) \mathbf{X}_{A^*} u + \frac{1}{n} u' \mathbf{X}'_{A^*} \Sigma^{-1} \mathbf{X}_{A^*} u$$

$$\begin{aligned}
\frac{1}{n} u' \mathbf{X}'_{A^*} (\hat{\Sigma}^{-1} - \Sigma^{-1}) \mathbf{X}_{A^*} u &\leq \|u\|_2 \left\| \frac{1}{n} \mathbf{X}'_{A^*} \mathbf{X}_{A^*} \right\|_2 \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_2 \\
&\leq \|u\|_2 \left\| \frac{1}{n} \mathbf{X}' \mathbf{X} \right\|_\infty \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_\infty
\end{aligned}$$

donde la primera desigualdad se debe a la desigualdad de Cauchy-Schwarz y la segunda se debe a los teoremas 4.3.28 y 5.6.9 de Horn and Johnson [29]. Entonces se tiene:



$$\frac{1}{n}u' \mathbf{X}'_{A^*} \hat{\Sigma}^{-1} \mathbf{X}_{A^*} u \geq \frac{1}{n}u' \mathbf{X}'_{A^*} \Sigma^{-1} \mathbf{X}_{A^*} u - \|u\|_2 \left\| \left\| \frac{1}{n} \mathbf{X}' \mathbf{X} \right\|_{\infty} \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_{\infty} \right\|$$

por lo que:

$$\begin{aligned} \rho_{\min}(\tilde{\mathbf{C}}_{11}^{\mathbf{n}}) &= \min_{\|u\|_2=1} \left( \frac{1}{n} u' \mathbf{X}'_{A^*} \hat{\Sigma}^{-1} \mathbf{X}_{A^*} u \right) \\ &\geq \min_{\|u\|_2=1} \left( \frac{1}{n} u' \mathbf{X}'_{A^*} \Sigma^{-1} \mathbf{X}_{A^*} u \right) - \left\| \left\| \frac{1}{n} \mathbf{X}' \mathbf{X} \right\|_{\infty} \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_{\infty} \right\| \\ &\geq \rho_{\min}(\dot{\mathbf{C}}_{11}^{\mathbf{n}}) - M_4 \left\| \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_{\infty} \right\| \\ &\geq M_2 - M_4 \left\| \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_{\infty} \right\| \end{aligned} \quad (1.6)$$

donde en las últimas dos desigualdades se utilizan las hipótesis 2 y 7. Utilizando los resultados (1.5), (1.6) y la hipótesis 10, se obtiene:

$$\left\| \left\| (\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} - (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \right\|_2 \right\| \leq \frac{M_4 \left\| \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_{\infty} \right\|}{M_2 \left( M_2 - M_4 \left\| \left\| \hat{\Sigma}^{-1} - \Sigma^{-1} \right\|_{\infty} \right\| \right)} = O_P \left( \frac{1}{\sqrt{n}} \right) \quad (1.7)$$

Entonces, por (1.4) se tiene que para todo  $j \in A^*$

$$\begin{aligned} P \left( \left| \left( (\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} - (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \right) \text{sign}(\beta_{A^*}) \right|_j \geq \frac{2n}{5\lambda} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right) &\leq \\ P \left( \sqrt{p^*} \left\| \left\| (\tilde{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} - (\dot{\mathbf{C}}_{11}^{\mathbf{n}})^{-1} \right\|_2 \right\| \geq \frac{2\sqrt{n}}{5\lambda} \left( M_3 n^{c_2/2} - \frac{\lambda}{2\sqrt{n}} \frac{p^*}{M_2} \right) \right) & \end{aligned}$$

Para probar que el segundo término de la desigualdad tiende a 0 cuando  $n \rightarrow \infty$ , utilizando que  $\frac{p^* \lambda}{\sqrt{n}} = o(n^{c_2/2})$  y (1.7), es suficiente probar que:

$$\lim_{n \rightarrow \infty} P(n^{c_1/4} n^{-1/2} \geq n^{1/2} n^{c_2/2} / \lambda) = 0$$

lo cual se verifica ya que

$$P(n^{c_1/4}n^{-1/2} \geq n^{1/2}n^{c_2/2}/\lambda) = P(\underbrace{n^{c_1/4}n^{-1/2}n^{-c_1/2}}_{=n^{-c_1/4}n^{-1/2} \rightarrow 0} \geq \underbrace{n^{1/2}n^{c_2/2}n^{-c_1/2}/\lambda}_{=\frac{n^{1/2}/\lambda}{n^{-(c_2-c_1)/2}} \rightarrow \infty})$$

dado que  $\frac{\lambda/\sqrt{n}}{n^{(c_2-c_1)/2}} \rightarrow 0$  por la hipótesis 6.

Ahora se probará que  $PA_2 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Se tiene que:

$$\begin{aligned} \tilde{W}_{A^*} - \dot{W}_{A^*} &= \frac{1}{\sqrt{n}} \left( \tilde{\mathbf{X}}'_{A^*} \tilde{\varepsilon} - \dot{\mathbf{X}}'_{A^*} \dot{\varepsilon} \right) \quad (1.8) \\ &= \frac{1}{\sqrt{n}} \left( (\hat{\Sigma}^{-1/2} \mathbf{X}_{A^*})' \hat{\Sigma}^{-1/2} \varepsilon - (\Sigma^{-1/2} \mathbf{X}_{A^*})' \Sigma^{-1/2} \varepsilon \right) \\ &= \frac{1}{\sqrt{n}} \left( \mathbf{X}'_{A^*} \underbrace{\hat{\Sigma}^{-1/2} \hat{\Sigma}^{-1/2}}_{\hat{\Sigma}^{-1}} \varepsilon - \mathbf{X}'_{A^*} \underbrace{\Sigma^{-1/2} \Sigma^{-1/2}}_{\Sigma^{-1}} \varepsilon \right) \\ &= \frac{1}{\sqrt{n}} \mathbf{X}'_{A^*} (\hat{\Sigma}^{-1} - \Sigma^{-1}) \varepsilon = \frac{1}{\sqrt{n}} \left[ \mathbf{X}' (\hat{\Sigma}^{-1} - \Sigma^{-1}) \varepsilon \right]_{A^*} \stackrel{d}{=} AZ \end{aligned}$$

Donde  $\mathbf{A} = \frac{1}{\sqrt{n}} \left[ \mathbf{X}' (\hat{\Sigma}^{-1} - \Sigma^{-1}) \Sigma^{1/2} \right]_{A^*}$  y  $Z$  es un vector aleatorio con distribución normal centrado y con matriz de varianzas y covarianzas igual a la matriz identidad.

Por la desigualdad de Cauchy-Schwarz, se tiene para cada matriz  $\mathbf{B}$  de dimensión  $K \times n$ , para cada vector  $U$  de dimensión  $n \times 1$  y para cada  $k \in \{1, \dots, K\}$

$$|(\mathbf{B}U)_k| \leq \|\mathbf{B}U\|_\infty \leq \|\mathbf{B}U\|_2 \leq \|\mathbf{B}\|_2 \|U\|_2 \quad (1.9)$$

Entonces, para cada  $j \in A^*$ , para cada  $\zeta \in \mathbb{R}$  y para cada matriz  $\mathbf{D}$  de dimensión  $A^* \times A^*$ :

$$\begin{aligned} P(|(\mathbf{D}(\tilde{W}_{A^*} - \dot{W}_{A^*}))_j| \geq \zeta) &= P(|(\mathbf{D}AZ)_j| \geq \zeta) \quad (1.10) \\ &\leq P(\|\mathbf{D}\|_2 \|\mathbf{A}\|_2 \|Z\|_2 \geq \zeta) \end{aligned}$$

y

$$P \left( \left| ((\dot{\mathbf{C}}_{11}^n)^{-1}(\tilde{W}_{A^*} - \dot{W}_{A^*}))_j \right| \geq \frac{\sqrt{n}}{5} \left( |\beta_j| - \frac{\lambda}{2n} |((\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}))_j| \right) \right) \leq$$

$$P \left( \left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A} \right\|_2 \|Z\|_2 \geq \frac{\sqrt{n}}{5} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right)$$

En primer lugar se acota  $\left\| \mathbf{A} \right\|_2$ :

$$\begin{aligned} \left\| \mathbf{A} \right\|_2 &= \gamma(\mathbf{A}'\mathbf{A})^{1/2} & (1.11) \\ &= \gamma \left( \frac{1}{n} [\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{X}\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}^{1/2}]_{A^*, A^*} \right)^{1/2} \\ &\leq \gamma \left( \frac{1}{n} \boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{X}\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}^{1/2} \right)^{1/2} \\ &\leq \gamma(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})\gamma(\boldsymbol{\Sigma})^{1/2}\gamma(\mathbf{X}'\mathbf{X}/n)^{1/2} \\ &\leq \frac{\left\| \hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} \right\|_\infty \left\| (\mathbf{X}'\mathbf{X})/n \right\|_\infty^{1/2}}{\rho_{\min}(\boldsymbol{\Sigma}^{-1})^{1/2}} \end{aligned}$$

donde la primera desigualdad se justifica en el Teorema 4.3.28 y la tercera en el Teorema 5.6.9 de [29].

Utilizando las hipótesis 7, 9 y 10 se prueba que

$$\left\| \mathbf{A} \right\|_2 = O_P \left( \frac{1}{\sqrt{n}} \right) \quad (1.12)$$

El término  $\left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2$  puede acotarse utilizando la hipótesis 2:

$$\begin{aligned} \left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 &= \gamma((\dot{\mathbf{C}}_{11}^n)^{-1}) = \rho_{\max}((\dot{\mathbf{C}}_{11}^n)^{-1}) & (1.13) \\ &= \frac{1}{\rho_{\min}(\dot{\mathbf{C}}_{11}^n)} \leq \frac{1}{M_2} \end{aligned}$$

Por lo que se tiene:

$$\begin{aligned} &\lim_{n \rightarrow \infty} P \left( \left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A} \right\|_2 \|Z\|_2 \geq \frac{\sqrt{n}}{5} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right) = \\ &\lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \|Z\|_2 \geq \sqrt{n} \left[ n^{-\frac{1-c_2}{2}} - \frac{\lambda}{n} p^* \right] \right) = \lim_{n \rightarrow \infty} P \left( \|Z\|_2 \geq n^{\frac{1+c_2}{2}} - \lambda n^{\frac{c_1}{2}} \right) \end{aligned}$$

donde en la última igualdad se utiliza la hipótesis 3. Ahora bien,  $\|Z\|_2 \stackrel{d}{=} \sqrt{\chi}$ , donde  $\chi$  representa una variable aleatoria con distribución  $\chi_n^2$ .

Utilizando la desigualdad de Markov se puede demostrar que  $P(\|Z\|_2 \geq a) = P(\chi \geq a^2) \leq \frac{E(\chi)}{a^2}$ .

Por otro lado, la hipótesis 6 indica que  $\lambda = \iota n^{\frac{1+c_2-c_1}{2}}$  con  $\iota \rightarrow 0$ . Se obtiene:

$$\begin{aligned} \lim_{n \rightarrow \infty} P \left( \|(\dot{\mathbf{C}}_{11}^n)^{-1}\|_2 \|\mathbf{A}\|_2 \|Z\|_2 \geq \frac{\sqrt{n}}{5} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right) &\leq \\ \lim_{n \rightarrow \infty} \frac{n}{\left( n^{\frac{1+c_2}{2}} - \lambda n^{\frac{c_1}{2}} \right)^2} &= \lim_{n \rightarrow \infty} \frac{n}{\left( n^{\frac{1+c_2}{2}} - \iota n^{\frac{1+c_2}{2}} \right)^2} = \lim_{n \rightarrow \infty} \frac{n}{n^{1+c_2}} \frac{1}{(1-\iota)^2} = 0 \end{aligned}$$

Lo que permite demostrar que  $PA_2 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PA_3 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Se observa que:

$$\dot{W}_{A^*} = (\Sigma^{-1/2} \mathbf{X}_{A^*})' \Sigma^{-1/2} \varepsilon \stackrel{d}{=} \mathbf{A}_1 Z$$

donde  $\mathbf{A}_1$  es una matriz de  $p^* \times n$  definida como  $\mathbf{A}_1 = (\Sigma^{-1/2} \mathbf{X}_{A^*})'$  y  $Z$  es un vector aleatorio de dimensión  $n \times 1$  con distribución normal, centrado y con matriz de varianzas y covarianzas igual a la matriz identidad.

Usando (1.9) se tiene para cada  $j \in A^*$ , para cada  $\zeta \in \mathbb{R}$  y para cada matriz  $\mathbf{D}$  de dimensión  $A^* \times A^*$ :

$$P(|(\mathbf{D}\dot{W}_{A^*})_j| \geq \zeta) = P(|(\mathbf{D}\mathbf{A}_1 Z)_j| \geq \zeta) \leq P(\|\mathbf{D}\|_2 \|\mathbf{A}_1\|_2 \|Z\|_2 \geq \zeta) \quad (1.14)$$

entonces,

$$\begin{aligned} P \left( |((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \dot{W}_{A^*}|_j \geq \frac{\sqrt{n}}{5} \left( |\beta_j| - \frac{\lambda}{2n} |((\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}))_j| \right) \right) &\leq \\ P \left( \left\| ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \right\|_2 \|\mathbf{A}_1\|_2 \|Z\|_2 \geq \frac{\sqrt{n}}{5} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right) \end{aligned}$$

ahora bien, acotando  $\|\mathbf{A}_1\|_2$ :

$$\begin{aligned}
\|\mathbf{A}_1\|_2 &= \gamma(\mathbf{A}_1' \mathbf{A}_1)^{1/2} = \gamma \left( \frac{1}{n} \mathbf{X}'_{A^*} \underbrace{\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2}}_{\boldsymbol{\Sigma}^{-1}} \mathbf{X}_{A^*} \right)^{1/2} \\
&\leq \gamma \left( \frac{1}{n} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X} \right)^{1/2} \leq \gamma (\boldsymbol{\Sigma}^{-1})^{1/2} \gamma \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{1/2} \\
&\leq \sqrt{M_6 M_4}
\end{aligned} \tag{1.15}$$

donde en la primera desigualdad se utiliza el Teorema 4.3.28 de Horn and Johnson [29] y en la última desigualdad se utilizan las hipótesis 7 y 8.

Utilizando (1.7), (1.15) y los cálculos realizados en  $PA_2$ , se concluye que:

$$\begin{aligned}
\lim_{n \rightarrow \infty} P \left( \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \|\mathbf{A}_1\|_2 \|Z\|_2 \geq \frac{\sqrt{n}}{5} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right) = \\
\lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \|Z\|_2 \geq \sqrt{n} \left[ n^{-\frac{1-c_2}{2}} - \frac{\lambda}{n} p^* \right] \right) = 0
\end{aligned}$$

Por lo que queda demostrado que  $PA_3 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PA_4 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Por (1.10), para todo  $j \in A^*$ :

$$\begin{aligned}
P \left( \left| ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1})(\tilde{W}_{A^*} - \dot{W}_{A^*}) \right| \geq \frac{\sqrt{n}}{5} \left( |\beta_{A^*}| - \frac{\lambda}{2n} |(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})| \right) \right) \\
\leq P \left( \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \|\mathbf{A}\|_2 \|Z\|_2 \geq \frac{\sqrt{n}}{5} \left( M_3 n^{-\frac{1-c_2}{2}} - \frac{\lambda}{2n} \frac{p^*}{M_2} \right) \right)
\end{aligned}$$

por lo que utilizando (1.7) y (1.12) es suficiente con probar que:

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} \|Z\|_2 \geq \sqrt{n} \left( n^{-\frac{1-c_2}{2}} - \frac{\lambda}{n} p^* \right) \right) = 0$$

se sabe por la desigualdad de Markov que:

$$\begin{aligned}
P\left(\|Z\|_2 \geq n^{3/2} \left(n^{\frac{1-c_2}{2}} - \frac{\lambda}{n} p^*\right)\right) &= P\left(\|Z\|_2 \geq \sqrt{n} \left(n^{\frac{3-c_2}{2}} - \lambda p^*\right)\right) \\
&\leq \frac{n}{\left(\sqrt{n} \left(n^{\frac{3-c_2}{2}} - \lambda p^*\right)\right)^2} \\
&= \frac{1}{\left(n^{\frac{3-c_2}{2}} - \lambda p^*\right)^2}
\end{aligned}$$

entonces,

$$\lim_{n \rightarrow \infty} \frac{1}{\left(n^{\frac{3-c_2}{2}} - \lambda p^*\right)^2} = \lim_{n \rightarrow \infty} \frac{1}{n^{3-c_2} \left(1 - n^{-\frac{3-c_2}{2}} \lambda p^*\right)^2} = 0$$

ya que por las hipótesis 3 y 6,  $n^{-\frac{3-c_2}{2}} \lambda p^* = o\left(\frac{1}{n}\right)$ . Queda demostrado que  $PA_4 \rightarrow 0$  cuando  $n \rightarrow 0$ .

A continuación se estudiará  $\tilde{B}_n^c$ . Por definición se tiene que:

$$\tilde{B}_n^c = \left\{ \left| \tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \tilde{W}_{A^*} - \tilde{W}_{A^{*c}} \right| \geq \frac{\lambda}{2\sqrt{n}} \left(1 - \left| \tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}) \right| \right) \right\}$$

Ahora bien, es posible descomponer  $(\tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \tilde{W}_{A^*} - \tilde{W}_{A^{*c}})$  y  $(\tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}))$  en los siguientes sumandos:

$$\begin{aligned}
\tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \tilde{W}_{A^*} - \tilde{W}_{A^{*c}} &= \dot{\mathbf{C}}_{21}^n (\dot{\mathbf{C}}_{11}^n)^{-1} \dot{W}_{A^*} - \dot{W}_{A^{*c}} \\
&+ \dot{\mathbf{C}}_{21}^n (\dot{\mathbf{C}}_{11}^n)^{-1} (\tilde{W}_{A^*} - \dot{W}_{A^*}) \\
&+ \dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \dot{W}_{A^*} \\
&+ \dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) (\tilde{W}_{A^*} - \dot{W}_{A^*}) \\
&+ (\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) (\dot{\mathbf{C}}_{11}^n)^{-1} \dot{W}_{A^*} \\
&+ (\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) (\dot{\mathbf{C}}_{11}^n)^{-1} (\tilde{W}_{A^*} - \dot{W}_{A^*}) \\
&+ (\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \dot{W}_{A^*} \\
&+ (\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) (\tilde{W}_{A^*} - \dot{W}_{A^*}) \\
&+ \dot{W}_{A^{*c}} - \tilde{W}_{A^{*c}}
\end{aligned}$$

$$\begin{aligned}
\tilde{\mathbf{C}}_{21}^n (\tilde{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}) &= \dot{\mathbf{C}}_{21}^n (\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}) \\
&+ \dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*}) \\
&+ (\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) (\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*}) \\
&+ (\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*})
\end{aligned}$$

Por la hipótesis 5 y la desigualdad triangular, se tiene que:

$$\begin{aligned}
P(\tilde{B}_n^c) &\leq PB_1 + PB_2 + PB_3 + PB_4 + PB_5 + PB_6 + PB_7 + PB_8 + PB_9 \\
&+ PB_{10} + PB_{11} + PB_{12}
\end{aligned}$$

con

$$\begin{aligned}
PB_1 &= P\left(|\dot{\mathbf{C}}_{21}^n (\dot{\mathbf{C}}_{11}^n)^{-1} \dot{W}_{A^*} - \dot{W}_{A^{*c}}| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_2 &= P\left(|\dot{\mathbf{C}}_{21}^n (\dot{\mathbf{C}}_{11}^n)^{-1} (\tilde{W}_{A^*} - \dot{W}_{A^*})| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_3 &= P\left(|\dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \dot{W}_{A^*}| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_4 &= P\left(|\dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) (\tilde{W}_{A^*} - \dot{W}_{A^*})| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_5 &= P\left(|(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) (\dot{\mathbf{C}}_{11}^n)^{-1} \dot{W}_{A^*}| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_6 &= P\left(|(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) (\dot{\mathbf{C}}_{11}^n)^{-1} (\tilde{W}_{A^*} - \dot{W}_{A^*})| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_7 &= P\left(|(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \dot{W}_{A^*}| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_8 &= P\left(|(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) (\tilde{W}_{A^*} - \dot{W}_{A^*})| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_9 &= P\left(|\dot{W}_{A^{*c}} - \tilde{W}_{A^{*c}}| \geq \frac{\lambda}{24\sqrt{n}} \delta\right) \\
PB_{10} &= P\left(|\dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*})| \geq \frac{\delta}{12}\right) \\
PB_{11} &= P\left(|(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) (\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})| \geq \frac{\delta}{12}\right) \\
PB_{12} &= P\left(|(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n) ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*})| \geq \frac{\delta}{12}\right)
\end{aligned}$$

En la demostración del Teorema 1 se probó que  $P(B_n^c) \rightarrow 0$  cuando  $n \rightarrow \infty$ , y  $PB_1$  es casi idéntico, salvo por el término  $1/12$  y porque se utiliza  $\dot{\mathbf{X}}_{A^*}$  y  $\dot{\mathbf{X}}_{A^{*c}}$  en lugar de  $\mathbf{X}_{A^*}$  y  $\mathbf{X}_{A^{*c}}$ .

Las hipótesis 1, 2 y 5 aseguran que  $\dot{\mathbf{X}}_{A^*}$  y  $\dot{\mathbf{C}}_{11}^n$  verifican las hipótesis del Teo-

rema 1, lo que permite concluir que  $PB_1 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_2 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Utilizando (1.10), se tiene para todo  $j \in A^*$ ,

$$\begin{aligned} & P \left( |\dot{\mathbf{C}}_{21}^n (\dot{\mathbf{C}}_{11}^n)^{-1} (\tilde{W}_{A^*} - \dot{W}_{A^*})| \geq \frac{\lambda}{24\sqrt{n}} \delta \right) \\ & \leq P \left( \left\| \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| A \right\|_2 \left\| Z \right\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right) \end{aligned}$$

A su vez,

$$\begin{aligned} \left\| \dot{\mathbf{C}}_{21}^n \right\|_2 &= \gamma \left( \frac{1}{n} \dot{\mathbf{X}}'_{A^*c} \dot{\mathbf{X}}_{A^*} \frac{1}{n} \dot{\mathbf{X}}'_{A^*} \dot{\mathbf{X}}_{A^*c} \right)^{1/2} \leq \gamma \left( \frac{\dot{\mathbf{X}}'_{A^*c} \dot{\mathbf{X}}_{A^*c}}{n} \right)^{1/2} \gamma \left( \frac{\dot{\mathbf{X}}'_{A^*} \dot{\mathbf{X}}_{A^*}}{n} \right)^{1/2} \\ &= \gamma (\dot{\mathbf{C}}_{22}^n)^{1/2} \gamma (\dot{\mathbf{C}}_{11}^n)^{1/2} \leq \gamma (\dot{\mathbf{C}}^n) \\ &\leq \rho_{max}(\Sigma^{-1}) \rho_{max}(\mathbf{X}'\mathbf{X}/n) \leq M_6 M_4 \end{aligned} \quad (1.16)$$

donde en la segunda desigualdad se utiliza el Teorema 4.3.28 de Horn and Johnson [29] y en la última desigualdad se utilizan las hipótesis 7 y 8.

Utilizando (1.13) y (1.16) se tiene:

$$\begin{aligned} \lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \|Z\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right) &= \lim_{n \rightarrow \infty} P \left( \chi \geq \left( \sqrt{n} \frac{\lambda}{24\sqrt{n}} \delta \right)^2 \right) \\ &\leq \lim_{n \rightarrow \infty} \frac{n}{n \left( \frac{\lambda}{24\sqrt{n}} \delta \right)^2} \\ &= \lim_{n \rightarrow \infty} \frac{1}{\left( \frac{\lambda}{\sqrt{n}} \frac{\delta}{24} \right)^2} = 0 \end{aligned}$$

Donde en el resultado anterior se utiliza la desigualdad de Markov y la hipótesis 6, y  $\chi \stackrel{d}{=} \sum_{i=1}^n Z_i^2 \sim \chi_n^2$ . Por lo que queda demostrado que  $PB_2 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_3 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Por (1.14) se tiene que para todo  $j \in A^*$ ,



$$P \left( |\dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \dot{W}_{A^*}| \geq \frac{\lambda}{24\sqrt{n}} \delta \right) \leq \\ P \left( \left\| \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A}_1 \right\|_2 \|Z\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right)$$

Por (1.7), (1.15) y (1.16), es suficiente con probar que:

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) = 0$$

lo que es directo ya que fue demostrado para  $PB_2$ . Se concluye que  $PB_3 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_4 \rightarrow 0$  cuando  $n \rightarrow \infty$ , lo que por (1.10) equivale a probar:

$$\lim_{n \rightarrow \infty} P \left( \left\| \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A} \right\|_2 \|Z\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right) = 0$$

Por (1.7), (1.12) y (1.16), es suficiente con probar que:

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{n} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) = 0$$

por la desigualdad de Markov se tiene que:

$$\begin{aligned} \lim_{n \rightarrow \infty} P \left( \frac{1}{n} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) &\leq \lim_{n \rightarrow \infty} \frac{n}{n^2 \left( \frac{\lambda}{\sqrt{n}} \right)^2} \\ &= \lim_{n \rightarrow \infty} \frac{1}{n \left( \frac{\lambda}{\sqrt{n}} \right)^2} \\ &= \lim_{n \rightarrow \infty} \frac{1}{\lambda^2} = 0 \end{aligned}$$

lo que permite concluir que  $PB_4 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_5 \rightarrow 0$  cuando  $n \rightarrow \infty$ , lo que por (1.14) equivale a probar que:

$$\lim_{n \rightarrow \infty} P \left( \left\| \tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A}_1 \right\|_2 \|Z\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right) = 0$$

A continuación se acotará  $\left\| \tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n \right\|_2$ :

$$\begin{aligned} \left\| \dot{\mathbf{C}}_{21}^n - \tilde{\mathbf{C}}_{21}^n \right\|_2 &= \gamma((\dot{\mathbf{C}}_{21}^n - \tilde{\mathbf{C}}_{21}^n)(\dot{\mathbf{C}}_{21}^n - \tilde{\mathbf{C}}_{21}^n)')^{1/2} \\ &\leq \gamma((\dot{\mathbf{C}}^n - \tilde{\mathbf{C}}^n)(\dot{\mathbf{C}}^n - \tilde{\mathbf{C}}^n)')^{1/2} \\ &\leq \gamma(\dot{\mathbf{C}}^n - \tilde{\mathbf{C}}^n) = \gamma \left( \frac{1}{n} \mathbf{X}'(\boldsymbol{\Sigma}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1})\mathbf{X} \right) \\ &\leq \gamma \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right) \gamma(\boldsymbol{\Sigma}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1}) \\ &\leq \left\| \frac{\mathbf{X}'\mathbf{X}}{n} \right\|_\infty \left\| \boldsymbol{\Sigma}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1} \right\|_\infty \\ &= O_P \left( \frac{1}{\sqrt{n}} \right) \end{aligned} \tag{1.17}$$

Para obtener el resultado anterior se utilizaron los Teoremas 4.3.28 y 5.6.9 de [29] y las hipótesis 7 y 10.

Por (1.13), (1.15) y (1.17), es suficiente probar que

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) = 0$$

lo que es directo ya que fue demostrado para  $PB_2$ . Se concluye que  $PB_5 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_6 \rightarrow 0$  cuando  $n \rightarrow \infty$ , lo que por (1.10) equivale a probar que:

$$\lim_{n \rightarrow \infty} P \left( \left\| \tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A} \right\|_2 \|Z\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right) = 0$$

Por (1.12), (1.13) y (1.17) es suficiente probar que:

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{n} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) = 0$$

lo que ya fue demostrado en  $PB_4$ , por lo que es directo probar que  $PB_6 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

A continuación se probará que  $PB_7 \rightarrow 0$  cuando  $n \rightarrow \infty$ , lo que por (1.14) equivale a probar que:

$$\lim_{n \rightarrow \infty} P \left( \left\| \tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A}_1 \right\|_2 \|Z\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right) = 0$$

Utilizando (1.7), (1.15) y (1.17), es suficiente con probar que:

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{n} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) = 0$$

lo que ya fue demostrado en  $PB_4$ , por lo que es directo probar que  $PB_7 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_8 \rightarrow 0$  cuando  $n \rightarrow \infty$ , lo que por (1.10) equivale a probar que:

$$\lim_{n \rightarrow \infty} P \left( \left\| \tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \left\| \mathbf{A} \right\|_2 \|Z\|_2 \geq \frac{\lambda}{24\sqrt{n}} \delta \right) = 0$$

Utilizando (1.7), (1.12) y (1.17), es suficiente con probar que:

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{n^{3/2}} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) = 0$$

utilizando la desigualdad de Markov y la hipótesis 6 se tiene:

$$\begin{aligned} \lim_{n \rightarrow \infty} P \left( \frac{1}{n^{3/2}} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) &= \lim_{n \rightarrow \infty} P \left( \|Z\|_2 \geq n^{3/2} \frac{\lambda}{\sqrt{n}} \right) \\ &\leq \frac{n}{n^3 \left( \frac{\lambda}{\sqrt{n}} \right)^2} = 0 \end{aligned}$$

Por lo que queda demostrado que  $PB_8 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_9 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Reemplazando  $A^*$  por  $A^{*c}$  en (1.8), (1.10) y (1.11), se obtiene:

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{\sqrt{n}} \|Z\|_2 \geq \frac{\lambda}{\sqrt{n}} \right) = 0$$

lo que es directo ya que fue demostrado para  $PB_2$ . Se concluye que  $PB_9 \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_{10} \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Utilizando (1.4) al igual que se utilizó para demostrar  $PA_5$ , se tiene que:

$$\begin{aligned} P \left( |\dot{\mathbf{C}}_{21}^n ((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*})| \geq \frac{\delta}{12} \right) &\leq \\ P \left( \sqrt{p^*} \left\| \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \geq \frac{\delta}{12} \right) \end{aligned}$$

Lo que según (1.7), (1.16) y la hipótesis 3 es equivalente a probar:

$$\lim_{n \rightarrow \infty} P \left( n^{c_1/4} n^{-1/2} \geq \frac{\delta}{12} \right) = 0$$

lo cual se verifica por el hecho de que  $c_1 < 1/2$ , por lo que se concluye que  $PB_{10} \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Ahora se probará que  $PB_{11} \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Utilizando la misma idea que para la demostración de  $PA_5$ , se tiene que:

$$P \left( |(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n)(\dot{\mathbf{C}}_{11}^n)^{-1} \text{sign}(\beta_{A^*})| \geq \frac{\delta}{12} \right) \leq \\ P \left( \sqrt{p^*} \left\| \tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \geq \frac{\delta}{12} \right)$$

Por (1.13), (1.17) y la hipótesis 3, es suficiente probar que:

$$\lim_{n \rightarrow \infty} P \left( n^{c_1/4} n^{-1/2} \geq \frac{\delta}{12} \right) = 0$$

lo cual ya fue demostrado para  $PB_{10}$ , por lo que se concluye que  $PB_{11} \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Finalmente, se probará que  $PB_{12} \rightarrow 0$  cuando  $n \rightarrow \infty$ .

Utilizando las mismas ideas que en  $PB_{10}$  y  $PB_{11}$ , se tiene:

$$P \left( |(\tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n)((\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1}) \text{sign}(\beta_{A^*})| \geq \frac{\delta}{12} \right) \leq \\ P \left( \sqrt{p^*} \left\| \tilde{\mathbf{C}}_{21}^n - \dot{\mathbf{C}}_{21}^n \right\|_2 \left\| (\tilde{\mathbf{C}}_{11}^n)^{-1} - (\dot{\mathbf{C}}_{11}^n)^{-1} \right\|_2 \geq \frac{\delta}{12} \right)$$

Lo cual equivale (por (1.7), (1.17) e hipótesis 3) a:

$$\lim_{n \rightarrow \infty} P \left( n^{c_1/4} n^{-1} \geq \frac{\delta}{12} \right) = 0$$

lo que se verifica al considerar que  $c_1 < 1/2$ . □

## Apéndice 2

# Correlación de X en la aplicación real

A continuación se presenta la matriz de correlaciones entre las variables explicativas utilizadas en el modelo para explicar el ingreso per cápita del hogar, utilizando datos de la Encuesta Continua de Hogares

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$X_{11}$	$X_{12}$	$X_{13}$	$X_{14}$	$X_{15}$
$X_1$	1.00	-0.64	0.01	0.22	-0.14	0.02	0.13	0.20	0.17	0.16	0.39	-0.22	0.11	0.19	0.21
$X_2$	.	1.00	0.05	-0.59	0.35	0.31	0.05	0.02	0.12	-0.19	-0.28	0.44	0.04	0.09	-0.15
$X_3$	.	.	1.00	0.07	0.65	-0.70	0.05	0.50	-0.59	-0.07	-0.06	0.17	0.49	-0.47	0.72
$X_4$	.	.	.	1.00	-0.09	-0.27	-0.11	-0.02	-0.19	0.03	0.07	-0.24	0.08	-0.15	0.16
$X_5$	.	.	.	.	1.00	-0.24	-0.05	0.34	-0.23	-0.15	-0.16	0.14	0.50	-0.14	0.26
$X_6$	.	.	.	.	.	1.00	0.04	-0.20	0.63	-0.02	-0.00	-0.03	-0.22	0.56	-0.48
$X_7$	.	.	.	.	.	.	1.00	0.67	0.48	-0.06	-0.06	0.29	0.14	0.20	0.18
$X_8$	.	.	.	.	.	.	.	1.00	0.00	-0.10	-0.09	0.28	0.44	-0.02	0.52
$X_9$	.	.	.	.	.	.	.	.	1.00	-0.02	0.01	0.16	-0.11	0.65	-0.39
$X_{10}$	.	.	.	.	.	.	.	.	.	1.00	0.80	-0.13	-0.07	-0.02	-0.01
$X_{11}$	.	.	.	.	.	.	.	.	.	.	1.00	-0.16	-0.05	0.04	0.02
$X_{12}$	.	.	.	.	.	.	.	.	.	.	.	1.00	0.59	0.38	0.43
$X_{13}$	.	.	.	.	.	.	.	.	.	.	.	.	1.00	0.02	0.77
$X_{14}$	.	.	.	.	.	.	.	.	.	.	.	.	.	1.00	-0.32
$X_{15}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.00
$X_{16}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{17}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{18}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{19}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{20}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{21}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{22}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{23}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{24}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{25}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{26}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{27}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{28}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$X_{29}$	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

**Tabla 2.1:** Matriz de correlaciones entre las variables explicativas en la aplicación a datos reales (Capítulo 5). Parte 1.

	$X_{16}$	$X_{17}$	$X_{18}$	$X_{19}$	$X_{20}$	$X_{21}$	$X_{22}$	$X_{23}$	$X_{24}$	$X_{25}$	$X_{26}$	$X_{27}$	$X_{28}$	$X_{29}$
$X_1$	0.28	0.27	0.24	0.26	0.29	0.20	-0.07	0.29	0.07	0.26	0.03	0.08	0.29	-0.24
$X_2$	-0.12	-0.08	-0.00	-0.14	-0.14	0.06	0.05	-0.16	0.05	-0.11	0.06	-0.00	-0.18	0.22
$X_3$	0.08	0.41	-0.47	0.32	0.48	-0.40	0.10	-0.05	0.10	0.14	-0.05	-0.06	-0.09	0.01
$X_4$	0.02	0.04	-0.12	0.09	0.10	-0.11	-0.02	0.10	-0.01	0.08	-0.00	-0.02	0.12	0.05
$X_5$	-0.10	0.15	-0.19	0.05	0.21	-0.13	0.07	0.01	0.21	0.05	0.08	-0.01	-0.05	0.04
$X_6$	-0.04	-0.18	0.40	-0.21	-0.26	0.42	-0.10	0.13	0.05	-0.06	0.10	0.08	0.15	0.02
$X_7$	0.85	0.62	0.46	0.31	0.22	0.25	-0.22	0.27	0.18	0.08	0.22	0.18	0.11	0.17
$X_8$	0.65	0.89	0.05	0.36	0.50	0.02	-0.14	0.24	0.25	0.20	0.18	0.12	0.06	0.12
$X_9$	0.40	0.02	0.89	0.03	-0.07	0.63	-0.20	0.29	0.14	0.04	0.22	0.17	0.19	0.10
$X_{10}$	-0.06	-0.06	-0.03	-0.10	-0.07	-0.08	0.16	-0.12	-0.18	-0.08	-0.11	-0.08	-0.00	-0.13
$X_{11}$	-0.01	-0.04	0.01	-0.06	-0.03	-0.06	0.15	-0.08	-0.19	-0.04	-0.13	-0.09	0.05	-0.18
$X_{12}$	0.48	0.42	0.30	0.51	0.44	0.33	-0.05	0.08	0.16	0.11	0.00	0.04	-0.04	0.31
$X_{13}$	0.35	0.58	-0.01	0.55	0.69	0.03	-0.04	0.16	0.25	0.15	0.00	0.01	0.06	0.12
$X_{14}$	0.30	0.03	0.76	0.16	0.02	0.77	-0.13	0.27	0.15	0.11	0.09	0.12	0.18	0.14
$X_{15}$	0.41	0.67	-0.31	0.63	0.74	-0.26	0.01	0.08	0.12	0.19	-0.08	-0.02	0.01	0.04
$X_{16}$	1.00	0.75	0.51	0.53	0.42	0.32	-0.19	0.30	0.19	0.16	0.16	0.18	0.13	0.14
$X_{17}$	.	1.00	0.08	0.50	0.64	0.07	-0.15	0.26	0.23	0.21	0.13	0.11	0.09	0.11
$X_{18}$	.	.	1.00	0.17	0.05	0.73	-0.19	0.30	0.16	0.10	0.19	0.17	0.19	0.10
$X_{19}$	.	.	.	1.00	0.84	0.35	-0.19	0.29	0.25	0.24	0.02	0.10	0.13	0.19
$X_{20}$	.	.	.	.	1.00	0.09	-0.16	0.27	0.28	0.28	0.01	0.06	0.12	0.13
$X_{21}$	.	.	.	.	.	1.00	-0.18	0.30	0.18	0.13	0.15	0.16	0.18	0.15
$X_{22}$	.	.	.	.	.	.	1.00	-0.57	-0.21	-0.11	-0.19	-0.17	-0.37	-0.19
$X_{23}$	.	.	.	.	.	.	.	1.00	0.48	0.36	0.42	0.39	0.52	0.30
$X_{24}$	.	.	.	.	.	.	.	.	1.00	0.07	0.27	0.24	0.14	0.19
$X_{25}$	.	.	.	.	.	.	.	.	.	1.00	0.07	0.06	0.13	0.12
$X_{26}$	.	.	.	.	.	.	.	.	.	.	1.00	0.26	0.10	0.14
$X_{27}$	.	.	.	.	.	.	.	.	.	.	.	1.00	0.19	0.05
$X_{28}$	.	.	.	.	.	.	.	.	.	.	.	.	1.00	0.07
$X_{29}$	.	.	.	.	.	.	.	.	.	.	.	.	.	1.00

**Tabla 2.2:** Matriz de correlaciones entre las variables explicativas en la aplicación a datos reales (Capítulo 5). Parte 2.

# ANEXOS



# Anexo 1

## Definiciones y resultados preliminares

Este apartado se basa en la sección 4.1 de Giraud [21]. Para profundizar puede consultarse Boyd and Vandenberghe [8].

### 1.0.1. Subdiferenciales y subgradientes

Para una cierta función convexa  $F$ , se define el subdiferencial de la función  $F$  en  $x$ , como el siguiente conjunto:

$$\partial F(x) = \{w \in \mathbb{R}^n : F(y) \geq F(x) + \langle w, y - x \rangle \forall y \in \mathbb{R}^n\} \quad (1.1)$$

A cada uno de los vectores  $w$  se los denomina subgradientes de  $F$  en el punto  $x$ .

Se desprende de la definición la siguiente propiedad:

#### Propiedad 1.

$$x_* \in \operatorname{argmín}_{x \in \mathbb{R}^n} F(x) \Leftrightarrow 0 \in \partial F(x_*) \quad (1.2)$$

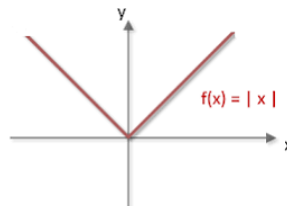
*Demostración.* Si  $x_* \in \operatorname{argmín}_{x \in \mathbb{R}^n} F(x) \Rightarrow F(x_*) \leq F(x), \forall x \in \mathbb{R}^n$ . Y se tiene que  $F(x) = F(x) - \langle 0, x - x_* \rangle, \forall x \in \mathbb{R}^n$  por lo que se verifica la definición 1.1 y 0 es subgradiente de  $F(x)$  en  $x = x_*$ .

Por otro lado, si 0 es subgradiente de  $F(x)$  para  $x = x_*$ , por definición  $F(x) \geq F(x_*) + \langle 0, x - x_* \rangle = F(x_*), \forall x \in \mathbb{R}^n$ , por lo que  $x_*$  es el  $\operatorname{argmín}(F(x))$ .  $\square$

## 1.0.2. Subdiferencial de la función valor absoluto

La función valor absoluto se define de la siguiente manera:

$$f(x) = \begin{cases} x & \text{si } x \geq 0 \\ -x & \text{si } x < 0 \end{cases}$$



**Figura 1.1:** Función valor absoluto

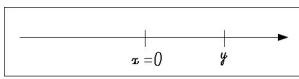
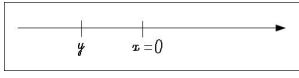
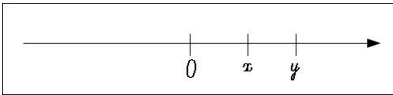
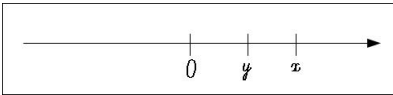
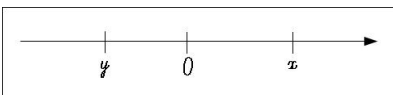
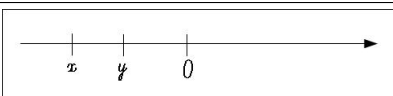
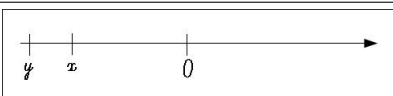
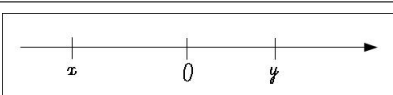
De acuerdo a la definición (1.1), el subdiferencial de la función valor absoluto es:

$$\partial|x| = \{w \in \mathbb{R} : |y| \geq |x| + w(y - x) \forall y \in \mathbb{R}\}$$

Para hallar los  $w$  que verifican la ecuación anterior, es necesario diferenciar según el signo de  $x$  e  $y$ . En particular, se tiene:

$$\partial|x| = \left\{ w \in \mathbb{R} : \begin{array}{l} \frac{|y|-|x|}{y-x} \geq w \quad \text{si } y - x > 0 \\ \text{o} \\ \frac{|y|-|x|}{y-x} \leq w \quad \text{si } y - x < 0 \end{array} , y \in \mathbb{R} \right\}$$

En el cuadro (1.1) se muestran las distintas posibilidades.

CASO 1: $x = 0$		
1)		$\left. \begin{array}{l} x = 0 \\ y > 0 \\ (y - x) > 0 \end{array} \right\} \Rightarrow \frac{ y - x }{y-x} = \frac{y}{y} = 1 \geq w$
2)		$\left. \begin{array}{l} x = 0 \\ y < 0 \\ (y - x) < 0 \end{array} \right\} \Rightarrow \frac{ y - x }{y-x} = \frac{-y}{y} = -1 \leq w$
CASO 2: $x > 0$		
3)		$\left. \begin{array}{l} x > 0 \\ y > 0 \\ (y - x) > 0 \end{array} \right\} \Rightarrow \frac{ y - x }{y-x} = \frac{y-x}{y-x} = 1 \geq w$
4)		$\left. \begin{array}{l} x > 0 \\ y > 0 \\ (y - x) < 0 \end{array} \right\} \Rightarrow \frac{ y - x }{y-x} = \frac{y-x}{y-x} = 1 \leq w$
5)		$\left. \begin{array}{l} x > 0 \\ y < 0 \\ (y - x) < 0 \end{array} \right\} \Rightarrow \begin{array}{l} \frac{ y - x }{y-x} = \frac{-y-x}{y-x} \leq w \\ -1 \leq \frac{-y-x}{y-x} \leq 1 \\ w \geq 1 \end{array}$
CASO 3: $x < 0$		
6)		$\left. \begin{array}{l} x < 0 \\ y < 0 \\ (y - x) > 0 \end{array} \right\} \Rightarrow \frac{ y - x }{y-x} = \frac{-y+x}{y-x} = -1 \geq w$
7)		$\left. \begin{array}{l} x < 0 \\ y < 0 \\ (y - x) < 0 \end{array} \right\} \Rightarrow \frac{ y - x }{y-x} = \frac{-y+x}{y-x} = -1 \leq w$
8)		$\left. \begin{array}{l} x < 0 \\ y > 0 \\ (y - x) > 0 \end{array} \right\} \Rightarrow \begin{array}{l} \frac{ y - x }{y-x} = \frac{y+x}{y-x} \geq w \\ -1 \leq \frac{y+x}{y-x} \leq 1 \\ w \leq -1 \end{array}$

**Tabla 1.1:** Distintas posibilidades para el subdiferencial, de acuerdo al signo de  $x$  e  $y$

En resumen, el subdiferencial de la función valor absoluto es:

$$\partial|x| = \left\{ w \in \mathbb{R} : \begin{array}{ll} w = \text{sign}(x) & \text{si } x \neq 0 \\ -1 \leq w \leq 1 & \text{si } x = 0 \end{array}, x \in \mathbb{R} \right\} \quad (1.3)$$

donde  $\text{sign}(x) = 1$  si  $x > 0$ ,  $\text{sign}(x) = -1$  si  $x < 0$  y  $\text{sign}(x) = 0$  si  $x = 0$ .

### 1.0.3. Subdiferencial de la norma $L_1$

La norma  $L_1$  de un vector  $x \in \mathbb{R}^n$  se define como:

$$\|x\|_1 = \sum_{i=1}^n |x_i| \quad (1.4)$$

El subdiferencial de la norma  $L_1$  es:

$$\begin{aligned} \partial\|x\|_1 &= \{w \in \mathbb{R}^n : \|y\|_1 \geq \|x\|_1 + \langle w, y - x \rangle \forall y \in \mathbb{R}^n\} \\ &= \left\{ w \in \mathbb{R}^n : \sum_{i=1}^n |y_i| \geq \sum_{i=1}^n |x_i| + w_i(y_i - x_i) \forall y \in \mathbb{R}^n \right\} \end{aligned}$$

Entonces, los  $w_i$  deben verificar que  $|y_i| \geq |x_i| + w_i(y_i - x_i)$ ,  $i = 1, \dots, n$ , y de acuerdo a la sección anterior esto se cumple para  $w_i = \text{sign}(x_i)$  si  $x_i \neq 0$  o  $w_i \in [-1, 1]$  si  $x_i = 0$ . En resumen:

$$\partial\|x\|_1 = \{w \in \mathbb{R}^n : w_i = \text{sign}(x_i) \text{ si } x_i \neq 0, w_i \in [-1, 1] \text{ si } x_i = 0\} \quad (1.5)$$

Cuando la función  $F : \mathbb{R}^n \rightarrow \mathbb{R}$  es convexa, se tiene que:

- El subdiferencial es monótono:

$$\langle w_k - w_y, x - y \rangle \geq 0 \quad \forall w_x \in \partial F(x) \text{ y } w_y \in \partial F(y) \quad (1.6)$$

La demostración es inmediata, considerando que por definición  $F(y) \geq F(x) + \langle w_k, y - x \rangle$  y  $F(x) \geq F(y) + \langle w_y, x - y \rangle$ . Sumando ambas desigualdades se obtiene:

$$F(x) + F(y) \geq F(x) + F(y) + \langle w_x, y - x \rangle + \langle w_y, x - y \rangle \Rightarrow \underbrace{-\langle w_x, y - x \rangle - \langle w_y, x - y \rangle}_{\langle w_k - w_y, x - y \rangle} \geq 0$$

- si  $F$  además es diferenciable  $\Rightarrow \partial F(x) = \{\nabla F(x)\}$

## Anexo 2

# Resultados de Zhu et al utilizados

En este apartado se presentan los resultados de la sección 5 del artículo de Zhu et al ([57]) que fueron utilizados para generar el código R.

$$\frac{\partial l(\eta)}{\partial \beta} = \mathbf{X}'\Sigma^{-1}(Y - \mathbf{X}\beta)$$

$$\frac{\partial l(\eta)}{\partial \theta} = \frac{1}{2}tr(\Sigma_1\Sigma) - \frac{1}{2}(Y - \mathbf{X}\beta)'\Sigma_1(Y - \mathbf{X}\beta) \text{ con } \Sigma_1 = \frac{\partial \Sigma^{-1}}{\partial \theta}$$

$$\frac{\partial l(\eta)}{\partial \sigma^2} = \frac{1}{2}tr(\Sigma_2\Sigma) - \frac{1}{2}(Y - \mathbf{X}\beta)'\Sigma_2(Y - \mathbf{X}\beta) \text{ con } \Sigma_2 = \frac{\partial \Sigma^{-1}}{\partial \sigma^2}$$

$$\mathcal{I}(\beta) = \mathbf{X}'\Sigma^{-1}\mathbf{X}$$

$$\mathcal{I}(\omega) = \begin{bmatrix} \frac{1}{2}tr(\Sigma_1\Sigma\Sigma_1\Sigma) & \frac{1}{2}tr(\Sigma_1\Sigma\Sigma_2\Sigma) \\ \frac{1}{2}tr(\Sigma_2\Sigma\Sigma_1\Sigma) & \frac{1}{2}tr(\Sigma_2\Sigma\Sigma_2\Sigma) \end{bmatrix} \text{ con } \omega = (\theta, \sigma^2)$$

Para errores de tipo CAR,  $\Sigma^{-1} = \mathbf{V}^{-1}(\mathbf{I} - \mathbf{C}_{\mathcal{R}}) = \frac{1}{\sigma^2}(\mathbf{I} - \theta\mathbf{W}_{\mathcal{R}})$ , por lo que

$$\Sigma_1 = \frac{\partial \Sigma^{-1}}{\partial \theta} = -\frac{1}{\sigma^2}\mathbf{W}_{\mathcal{R}}, \quad \Sigma_2 = \frac{\partial \Sigma^{-1}}{\partial \sigma^2} = -\frac{1}{(\sigma^2)^2}\mathbf{I} + \frac{\theta}{(\sigma^2)^2}\mathbf{W}_{\mathcal{R}}$$

Por otro lado, para errores de tipo SAR,  $\Sigma^{-1} = (\mathbf{I} - \mathbf{C}_{\mathcal{R}})\mathbf{V}^{-1}(\mathbf{I} - \mathbf{C}_{\mathcal{R}}) = \frac{1}{\sigma^2}(\mathbf{I} - \theta\mathbf{W}_{\mathcal{R}})(\mathbf{I} - \theta\mathbf{W}_{\mathcal{R}}) = \frac{1}{\sigma^2}(\mathbf{I} - 2\theta\mathbf{W}_{\mathcal{R}} + \theta^2\mathbf{W}_{\mathcal{R}}\mathbf{W}_{\mathcal{R}})$ , por lo que

$$\Sigma_1 = \frac{\partial \Sigma^{-1}}{\partial \theta} = -\frac{2}{\sigma^2}\mathbf{W}_{\mathcal{R}} + \frac{2\theta}{\sigma^2}\mathbf{W}_{\mathcal{R}}\mathbf{W}_{\mathcal{R}}$$

$$\Sigma_2 = \frac{\partial \Sigma^{-1}}{\partial \sigma^2} = -\frac{1}{(\sigma^2)^2}(\mathbf{I} - 2\theta\mathbf{W}_{\mathcal{R}} + \theta^2\mathbf{W}_{\mathcal{R}}\mathbf{W}_{\mathcal{R}})$$