Distribución previa *horseshoe* en la regresión lineal: estudio de su comportamiento como método de selección de variables en condiciones de alta dimensionalidad

Pablo Jiménez de Yarza

Máster en Metodología de las Ciencias de la Salud y del Comportamiento

MÁSTERES DE LA UAM

2019 - 2020

Facultad de Psicología

UAM Universidad Autónoma de Madrid

UAM Universidad Autónoma de Madrid

Distribución previa *horseshoe* en la regresión lineal: estudio de su comportamiento como método de selección de variables en condiciones de alta dimensionalidad

Horseshoe prior distribution in linear regression: a study of its behavior as a variable selection method in high-dimensional settings

Máster en Metodología de las Ciencias del Comportamiento y de la Salud

Autor: Pablo Jiménez de Yarza

Tutor: José Héctor Lozano Bleda

Modalidad: Investigación

Fecha: Junio, 2020

Resumen

En modelos de regresión donde el número de predictores (p) excede el de observaciones (n) es habitual asumir el supuesto de que solo unos pocos coeficientes van a ser distintos de cero. Son las llamadas *estimaciones escasas*. La distribución previa *horseshoe* es un método bayesiano que tiene unos parámetros de contracción global (τ) y local (λ_j) que propician la selección de unas pocas variables. El objetivo del trabajo es estudiar los efectos de la decisión previa sobre la varianza de τ al estimar modelos con diferentes características. Para ello, se han ajustado modelos de regresión escasos y no escasos en una situación $p > n$ con coeficientes de diferente valor y definiendo varianzas de τ de más a menos constrictivas. Se han comparado los resultados obtenidos con dos criterios de selección distintos. Los resultados muestran que, si se cumple el supuesto de escasez y se utiliza un criterio conservador, la *horseshoe* obtiene distribuciones posteriores precisas de los coeficientes independientemente del valor de estos, y una varianza de τ pequeña consigue contraer el valor de las variables irrelevantes mientras detecta las relevantes. Se destaca la importancia del cumplimiento del supuesto de escasez del modelo verdadero para el buen funcionamiento de la *horseshoe*.

Palabras clave: estimaciones escasas, previas de contracción, *horseshoe*, problemas de alta dimensionalidad, regresión bayesiana

Abstract

In regression models where the number of predictors (p) exceeds the number of observations (n), it is usual to believe the sparsity assumption, where only a few coefficients will be different from zero. They are the so-called sparse estimations. The horseshoe prior distribution is a Bayesian method which has global (τ) and local (λ_j) shrinkage parameters that allow for solutions with only a few relevant variables. The goal of this paper is to study the effects of prior choice for the scale of τ on the estimations of different models. We have fitted sparse and non-sparse regression models in a $p > n$ situation with coefficients of different value and defining different scales for the global parameter with different shrinking strength. Results have been compared using two different variable selection criteria. Results show that, if the sparsity assumption is fulfilled and by using a conservative criterion, the horseshoe obtains accurate posterior distributions of the coefficients regardless of coefficients' real values, and a small scale for τ is enough to shrink the noise variables while leaving signal variables unshrunk. We highlight the relevance of a true sparsity for the horseshoe to properly function in a $p > n$ situation.

Keywords: sparse estimations, shrinkage priors, horseshoe, high-dimensional problems, Bayesian regression.

Índice

1. Introducción

Uno de los procedimientos más extendidos en estadística, tanto frecuentista como bayesiana, es el de regresión lineal. Se estudia y aplica en todo tipo de disciplinas, incluyendo la Psicología, cuyas publicaciones en la rama bayesiana se han incrementado en términos absolutos y relativos en los últimos años (Van de Schoot, Winter, Ryan, Zondervan y Depaoli, 2017). En palabras de Pardo y San Martín (1994), el modelo de regresión lineal permite "valorar el impacto individual y colectivo de las variables independientes sobre la dependiente y efectuar pronósticos sobre la variable dependiente" (p. 372). Sobre esta base, el modelo viene dado por la expresión

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad (1)$$

donde el subíndice i hace referencia a la i-ésima observación de las n totales ($i = 1, \ldots, n$). Por su parte, $x_1, \ldots, x_p$ hacen referencia a las p variables predictoras; estas se multiplican por los coeficientes, que podemos agrupar en el vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. Por lo tanto, la observación y_i de la variable dependiente se obtiene como una suma ponderada de los valores de las variables independientes a la que añadimos un término de error aleatorio ε_i que se distribuye normalmente según $\varepsilon_i \sim N(0, \sigma)$, donde σ es la desviación típica del error, común a todas las observaciones. La estimación de los parámetros determinará el peso que tiene cada uno de los regresores en las predicciones. En función del problema a tratar, existen diferentes aproximaciones para llevar a cabo la selección de variables.

El problema que nos ocupa en el presente trabajo es el conocido como *sparse estimations* en *high dimensional problems* o *estimaciones escasas* (o *dispersas*) en situaciones de *alta dimensionalidad*. Las situaciones de alta dimensionalidad se producen cuando el número de variables predictoras es elevado y sobrepasa al número de observaciones, es decir, $p > n$. Esta clase de problemas son frecuentes en la actualidad en muchos ámbitos dada la creciente accesibilidad a grandes cantidades de datos. En estos casos, la incertidumbre de las estimaciones es elevada (Piironen, 2019), así como la inestabilidad; esto puede llevar a obtener modelos muy complejos con un mal ajuste (Carvalho, Polson y Scott, 2009). Además, se da la posibilidad de que los investigadores no posean argumentos que apoyen una combinación concreta sobre otra y lo que se busque sea un conjunto de variables que aporte información relevante de entre todas las

posibilidades (véanse los ejemplos propuestos en Johnstone y Silverman, 2004; Scott y Berger, 2010).

Para evitar los problemas mencionados, la solución más extendida en las situaciones $p > n$ es llevar a cabo una reducción de la dimensionalidad asumiendo el supuesto de que el modelo verdadero está compuesto por pocos predictores (Clarke, Fokoue y Zhang, 2009) y que, por tanto, muchos de los coeficientes de las variables independientes van a ser cercanos o iguales a cero, lo que se conoce como *estimaciones escasas*. En palabras de O'Hara y Sillanpää (2009), el conocimiento a priori acerca del grado de escasez del modelo es muy específico de dominio y depende de varios factores, como el objetivo (predicción, inferencia) o el número de observaciones n. Con las *estimaciones escasas*, a cambio de sacrificar algo de sesgo (Tibshirani, 1996), se facilita la interpretación del modelo manteniendo una capacidad predictiva adecuada (Piironen, Paasiniemi y Vehtari, 2018; Tibshirani, 1996). Existen muchos procedimientos diferentes, pero según Bhadra, Datta, Polson y Willard (2019a), hay cuatro puntos que son comunes a todos ellos: estimación de parámetros subyacentes, realización de comparaciones múltiples, selección de un conjunto de variables y predicción de datos no observados.

En la rama frecuentista puede destacarse el Lasso (*Least absolute shrinkage and selection operator*), introducido por Tibshirani (1996), en el que se propone un parámetro que penaliza el tamaño del modelo y fuerza algunos coeficientes a que tengan el valor cero, asegurando la escasez de predictores. Con el tiempo se han propuesto alternativas al Lasso original que permiten tratar con las mencionadas situaciones $p > n$ (por ejemplo, Paul, Bair, Hastie y Tibshirani, 2008). Estos métodos incluyen penalizaciones para controlar la tasa de error al realizar las comparaciones múltiples. En Hastie, Tibshirani y Wainwright (2015) puede encontrarse una extensa recopilación de los métodos clásicos.

Por su parte, en la literatura bayesiana se opta por establecer a priori un parámetro -*sparsifying prior*-, que propicie las soluciones con pocas variables relevantes, donde el valor posterior de muchos de los coeficientes no sea distinto de cero. Aplicaciones bayesianas de *estimaciones escasas* con bases de datos reales pueden encontrarse en estudios genéticos (Ishwaran y Rao, 2005a; Peltola et al., 2012; Lee, Sha, Dougherty, Vannucci y Mallick., 2003; Ročková y George, 2014), así como de neuroimagen (Sanyal y Ferreira, 2012), con pacientes clínicos (Malsiner-Walli y Wagner, 2011) o sobre el cáncer (Piironen y Vehtari, 2016, 2017a).

Debido a la complejidad que entrañan los problemas de dimensionalidad elevada, desde el enfoque bayesiano las estimaciones suelen realizarse mediante simulaciones Monte Carlo con Cadenas de Markov (e.g., Brooks, Gelman, Jones y Meng, 2011; Robert y Casella, 2013), también llamadas MCMC por sus siglas en inglés. Además, ya hemos mencionado que los problemas de alta dimensionalidad conllevan una mayor incertidumbre en la estimación de los parámetros de los modelos (Piironen, 2019), así como una importante carga computacional al utilizar herramientas como MCMC (Johndrow, Mattingly, Mukherjee y Dunson, 2015). Definir adecuadamente las distribuciones a priori de los parámetros puede ayudar a aliviar ambas consecuencias (Piironen et al., 2018; Piironen, 2019).

El trabajo se estructurará de la siguiente forma: en la Sección 1.1 vamos a comenzar definiendo y desarrollando las distribuciones previas ya mencionadas. Posteriormente trataremos el concepto de selección de variables y se revisarán los principales procedimientos existentes en el área (Sección 1.2). Después se expondrán los objetivos del presente estudio (Sección 1.3) y el Método (Sección 2), donde se desglosarán las condiciones del trabajo y las decisiones tomadas acerca de la programación y simulación de los modelos. En la Sección 3 se presentarán los principales resultados, y en la Discusión (Sección 4) se pondrán en relación estos resultados con la teoría, incluyendo los hallazgos más relevantes. Las principales limitaciones, propuestas de futuras líneas de investigación y algunas consideraciones finales se realizarán en la Sección 5.

1.1. Distribuciones previas

Las distribuciones a priori más destacadas y comúnmente utilizadas al definir los predictores de este tipo de problemas son las llamadas *spike-and-slab* (Mitchell y Beauchamp, 1988), así como alternativas que han surgido más recientemente, de entre las que destaca la *horseshoe* (Carvalho et al., 2009), que es en la que nos vamos a centrar en el presente trabajo.

1.1.1. *Spike-and-slab*

Para facilitar la selección de predictores relevantes dentro de un conjunto grande de variables, el *spike-and-slab* define una distribución previa que se compone de dos elementos, el *spike* y el *slab*. El *spike*, en la formulación original de Mitchell y Beauchamp (1988), es una concentración de la masa de probabilidad en un único punto, generalmente el cero; posteriormente, George y McCulloch (1993) proponen en su lugar

una distribución normal con una varianza muy pequeña. En cualquiera de los dos casos, el *spike* permite que los coeficientes de las variables con efectos pequeños sean fijados a cero y éstas no sean tenidas en cuenta para el modelo. Por su parte, el *slab* es una distribución vaga que se define para acotar los valores que tomarán los parámetros en el caso de que no sean cero. Según Piironen y Vehtari (2017a) ambas formulaciones pueden obtenerse a partir del mismo planteamiento: consiste en definir una mezcla discreta de dos distribuciones normales de la siguiente forma,

$$\beta_j|\lambda_j, c, \varepsilon \sim \big(1-\lambda_j\big)N(0,\varepsilon^2) + \lambda_j N(0,c^2), \qquad (2)$$

La mitad izquierda de la suma, $N(0,\varepsilon^2)$, corresponde al *spike*. Fijando $\varepsilon = 0$ se obtiene la propuesta original de Mitchell y Beauchamp (1988), donde este elemento es una línea vertical que concentra toda la masa, mientras que en la vertiente de George y McCulloch (1993) se fijaría $\varepsilon > 0$, lo que lo convertiría en una suma de normales con distinta varianza. La mitad derecha de la expresión, $N(0,c^2)$, es el *slab*.

El subíndice j toma los valores $j = 1,\ldots,p$ siendo p el total de predictores. El valor del coeficiente β_j dependerá, por tanto, de si la información la aporta el *spike*, el *slab* o ambos. Para determinar el peso de uno y otro se utiliza el parámetro λ_j. Se le puede asignar un valor a este parámetro en base a nuestras creencias previas o mediante algún tipo de estimador (ver Scott y Berger, 2006). Lo más habitual, no obstante, es que este parámetro se distribuya según $\lambda_j \sim Ber(\pi)$ y que tome los valores 0 o 1. Para ello hay que definir la probabilidad, π, de obtener uno u otro valor, algo que suele hacerse mediante el establecimiento de una determinada distribución previa. Por ejemplo, Malsiner-Walli y Wagner (2011) lo definen como $\pi \sim Beta(a,b)$; Ishwaran y Rao (2005b) optan por

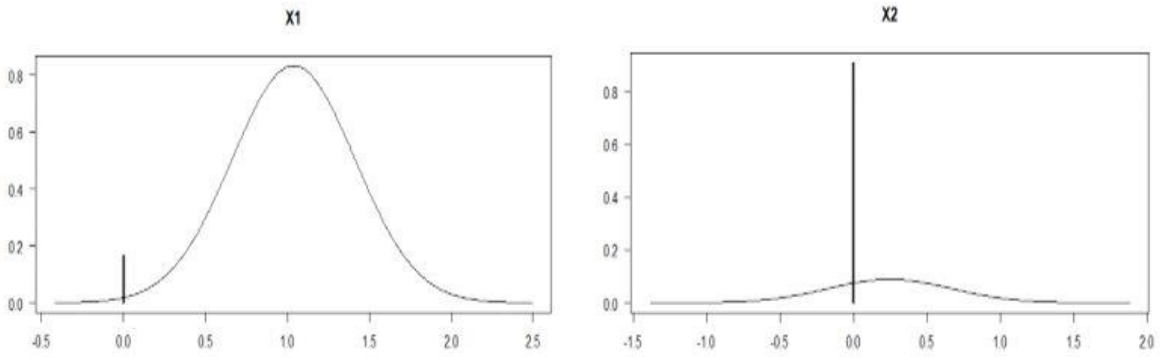

Figura 1. Spike-and-slab: Distribuciones marginales posteriores de un ejemplo de regresión con dos variables cuyo vector de coeficientes simulados es $\boldsymbol{\beta} = [1,0]$. La imagen de la izquierda muestra la variable relevante, con un *spike* poco probable y un *slab* centrado en torno al valor real. La derecha muestra la variable ruido, con un *spike* y un *slab* que señala la alta probabilidad de que el coeficiente sea cero.

$\pi \sim U(0,1)$. Una vez que se actualiza la información a la luz de los datos, el *spike* y el *slab* muestran sus distribuciones posteriores (ver Figura 1). El interés reside en obtener las probabilidades de inclusión a posteriori de los coeficientes y de esta forma identificar aquellos que son relevantes para el modelo, descartando los que no lo son.

Posteriormente, Ishwaran y Rao (2005a, 2005b) introducen una definición alternativa en la que reescalan los valores de la variable independiente y añaden un término inflador de la varianza como penalizador por este reescalamiento; el resultado es una mayor precisión y robustez. Para una descripción exhaustiva y fundamentada, véase el estudio comparativo de Malsiner-Walli y Wagner (2011) sobre las diferentes variaciones de previas del tipo *spike-and-slab*.

1.1.2. *Horseshoe priors*

El término *shrinkage priors* – que podríamos traducir como "previas de contracción"- hace referencia a un conjunto de distribuciones totalmente continuas, a diferencia de la mezcla discreta que supone el *spike-and-slab*. Estas propuestas poseen la ventaja de que son más eficientes computacionalmente (Carvalho et al., 2009; Polson y Scott, 2010; Piironen y Vehtari, 2016), y además evitan los problemas derivados de la sensibilidad que las probabilidades a posteriori pueden mostrar a la elección previa de los parámetros c y π (Piironen y Vehtari, 2017a; Piironen et al., 2018). Previas de contracción es una denominación que engloba distribuciones como Dirichlet-Laplace (Bhattacharya, Pati, Pillai y Dunson, 2015), la Normal-Gamma (Griffin y Brown, 2010) o la más extendida, la *horseshoe*.

Tanto el término como la formulación original de la *horseshoe* fueron propuestas por Carvalho et al. (2009), consistente en una mezcla de distribuciones normales

$$\left(\beta_j \middle| \lambda_j, \tau\right) \sim N\left(0, \lambda_j^2 \tau^2\right), \tag{3}$$

donde

$$\lambda_j \sim C^+(0,1), \tag{4}$$

en la que los coeficientes dependen de un parámetro de contracción global, τ, que "tira" de todos ellos hacia el valor cero, y del parámetro local λ_j que permite que algunos de estos coeficientes escapen a la contracción gracias a la forma de la distribución Cauchy, que concentra mucha masa en torno al cero pero tiene una cola ancha; de esta forma, las

variables ruido quedan "atrapadas" en valores muy próximos a cero, y a su vez posee la propiedad de que es robusta frente a las señales grandes y permite que se manifiesten (Carvalho, Polson y Scott, 2010). La distribución a priori half-Cauchy $\mathrm{C}^{+}(0,1)$ está definida únicamente para los valores reales positivos. En Polson y Scott (2012) y Bhadra, Datta, Li, Polson y Willard (2019b) se observa cómo la introducción de este parámetro local, cuyo valor puede variar entre coeficientes, mejora los resultados de predicción con nuevas muestras en comparación con métodos previos que contienen únicamente un parámetro global.

La media posterior de los coeficientes en la *horseshoe* (Datta y Ghosh, 2013; también en Van der Pas, Kleijn y Van der Vaart, 2014 y Piironen y Vehtari, 2017a, entre otros) se obtiene mediante

$$EAP(\beta_j|y_i,\lambda_j,\tau,\sigma^2) = (1-\frac{1}{1+\lambda_j^2\tau^2})y_i \tag{5}$$

siendo $\kappa_j = {1}\Big/{1+\lambda_j^2\tau^2}$ el denominado *factor de contracción* o *shrinkage factor* correspondiente al coeficiente β_j. Este tiene la siguiente distribución previa:

$$p(\kappa_j|\tau) = \frac{\tau}{\pi}\frac{1}{1-(1-\tau^2)\kappa_j}\frac{1}{\sqrt{\kappa_j}\sqrt{1-\kappa_j}} \tag{6}$$

y, cuando $\tau = 1$, su distribución adquiere la característica forma que le proporciona el nombre -*horseshoe*: herradura-. El factor de contracción puede definirse como la fuerza con la que β_j se ve atraído hacia el cero. Como puede apreciarse en la Figura 2, su formulación favorece los valores muy cercanos a 0 o 1, donde $\kappa = 0$ es la completa ausencia de contracción y $\kappa = 1$ significa contracción total. También puede verse que el parámetro global τ tiene un peso importante en el modelo: cuando toma valores pequeños, la densidad de κ se acumula en torno al 1 y todos los coeficientes experimentarán una fuerte contracción; por el contrario, si τ es grande, el valor de κ se encontrará próximo al 0 y la contracción será débil. En un contexto de regresión con un elevado número de variables, como el que nos ocupa, la decisión previa sobre τ influye en κ. La elección de τ se discute en detalle en el siguiente apartado. Dada la propiedad de κ de tomar valores cercanos a 0 o 1 y, atendiendo a la Ecuación 5, la expresión $1-\hat{\kappa}_j$ (siendo $\hat{\kappa}_j$ la media posterior de κ_j) cumple un papel equivalente a λ_j, la probabilidad de inclusión a posteriori del *spike-and-slab*, mencionada anteriormente (Ecuación 2).

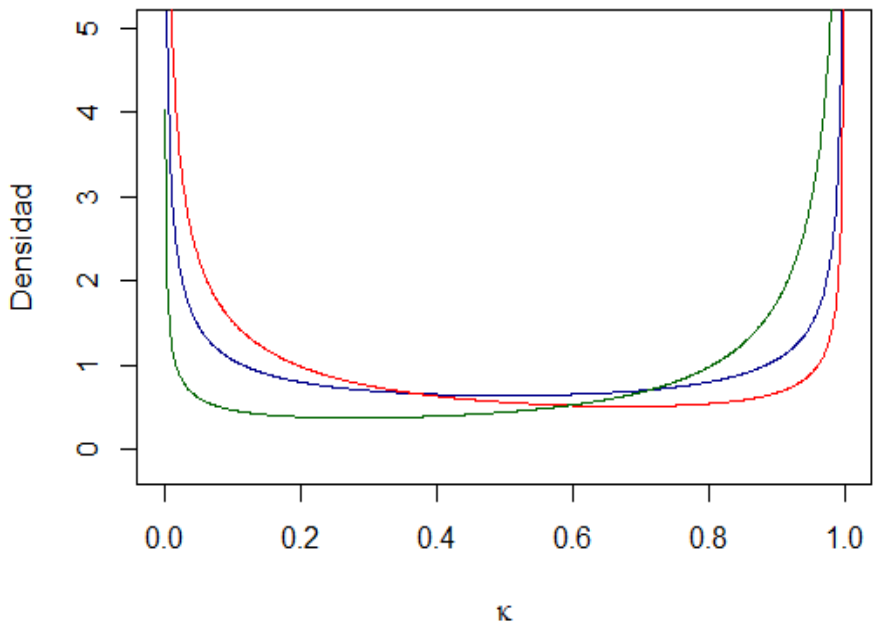

Figura 2. Densidad del factor de contracción κ para $\tau = 1$ (azul), $\tau = 1.7$ (roja) y $\tau = 0.4$ (verde)

La propuesta que Piironen y Vehtari (2017a) han denominado *regularized horseshoe* (RHS) introduce un nuevo parámetro, *c*, cuya función es equivalente a la que cumple el *slab* en el modelo *spike-and-slab* (Ecuación 2): determinar el rango de valores que toman aquellos coeficientes que sí son relevantes. El valor de los coeficientes condicionado al resto de parámetros vendrá dado en este caso por

$$(\beta_j|\lambda_j,\tau,c)\sim N(0,\tau^2\tilde{\lambda}_j^2), \tag{7}$$

donde

$$\tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{c^2+\tau^2\lambda_j^2}, \tag{8}$$

y donde el parámetro local se distribuye

$$\lambda_j\sim C^+(0,1) \qquad j = 1,\ldots,p \tag{9}$$

De esta forma puede imponerse una contracción extra incluso a aquellos coeficientes con valores grandes. Si bien es cierto que la propiedad de detectar valores grandes es uno de los puntos fuertes de la *horseshoe*, Piironen y Vehtari (2017a) encuentran que esto puede llevar a que las distribuciones posteriores de esas variables sean difusas, especialmente en modelos complejos cuando la incertidumbre del problema es grande. Los autores argumentan que este efecto regularizador es especialmente útil en aquellos

modelos de regresión logística donde se produce el fenómeno de la separación, es decir, aquel que ocurre cuando no existe superposición entre los casos de ambos grupos. Esto provoca que las estimaciones por máxima verosimilitud de los coeficientes sean infinitas (Ghosh, Li y Mitra, 2018), dificultando la estimación posterior. Existen soluciones frecuentistas a este problema (Heinze y Schemper, 2002), así como bayesianas (Ghosh et al., 2018). Piironen y Vehtari (2017a) encuentran, tras su estudio de simulación, que una elección adecuada -informativa o vagamente informativa- de c aporta robustez y eficiencia al proceso de estimación, tanto si existe separación como si no; la regresión lineal también se beneficiaría del uso de este elemento, controlando el ancho de la distribución posterior y disminuyendo los problemas de transiciones divergentes en las cadenas de Markov con el método No-U-Turn Sampling (Hoffman y Gelman, 2014), sin comprometer la capacidad predictiva, como ocurriría si se produjese una sobrecontracción en las estimaciones de los parámetros. La distribución que Piironen y Vehtari (2017a) proponen para este parámetro es la siguiente

$$c^2 \sim IG(\alpha, \beta) \tag{10}$$

Cuando $c \to \infty$ se obtiene la *horseshoe* original.

1.1.3. Decisión sobre el parámetro τ

Detengámonos ahora en la elección del parámetro de contracción global τ. Pueden encontrarse dos corrientes principales en torno a esta decisión: el *full Bayes approach*, donde τ es un hiperparámetro que sigue una determinada distribución y la aproximación empírica o *empirical Bayes*, de donde se obtiene un valor fijo a partir de un determinado indicador.

1.1.3.1. Full Bayes

Los autores originales de la distribución *horseshoe* (Carvalho et al., 2009) introducen en primer lugar que

$$\tau \sim C^+(0,1) \tag{11}$$

y basan su elección en los argumentos presentados por Gelman (2006) respecto a la elección de las distribuciones previas de los parámetros de la varianza en modelos jerárquicos. Esta primera propuesta ha sido utilizada habitualmente, y también criticada en ocasiones, como puede verse en Piironen y Vehtari (2016), donde defienden que la

distribución previa $\tau \sim C^+(0,1)$ no es poco informativa, sino que puede tener bastante peso en la inferencia posterior y que, además, este valor suele otorgar demasiada probabilidad a valores muy altos de τ.

Los mismos autores (Carvalho et al., 2010; Polson y Scott, 2010) proponen más tarde una nueva distribución para el parámetro global $\tau \sim C^+(0, a^2)$, abriendo así la posibilidad de flexibilizar el rango de τ según el tipo de problema.

Más recientemente se ha introducido una propuesta que busca integrar las ideas de las dos inmediatamente anteriores (véase su desarrollo en Piironen y Vehtari, 2016 y 2017a):

$$\tau \sim C^+(0, \tau_0^2), \tag{12}$$

$$\tau_0 = \frac{p_0}{p - p_0} \frac{\sigma}{\sqrt{n}}, \tag{13}$$

donde p_0 es nuestra creencia a priori del número de variables con coeficientes distintos de cero y p es el total de variables. Si imaginamos una situación en la que un investigador lleva a cabo un análisis de regresión con total ausencia de conocimiento previo acerca de la cantidad de variables que conforman el modelo subyacente, esto significa que podría decantarse por un valor dentro del intervalo $p_0 = [1, \ldots, p]$ -partiendo de que una expectativa razonable, antes de emprender un análisis, es que al menos una de las potenciales variables explicativas contribuirá a explicar la variable observada y, en el otro extremo, todas aportarán información-. Mediante la Ecuación 13 se obtendrán distintos valores τ_{0j} donde $j = 1, \ldots, p$, que podrán ser utilizados como varianza de la distribución de τ resultando en $\tau \sim C^+(0, \tau_{0j}^2)$. El término $\sigma/\sqrt{n}$ serviría para escalar este parámetro y evitar que se favorezcan modelos de mayor o menor tamaño en función de la desviación típica error σ y del número de observaciones n (Piironen y Vehtari, 2016). Puede intuirse la utilidad que puede tener una propuesta de esta índole en contextos aplicados con datos reales donde los investigadores posean razones previas para estipular un determinado número de variables relevantes. No obstante, dadas las características de los estudios de este tipo, en ocasiones no se dispondrá de la información necesaria para tomar una decisión basada en las creencias previas (por ejemplo, en estudios exploratorios). Sería, entonces, interesante conocer si el efecto de una decisión previa sobre p_0 que coincida o se aproxime al número real de variables aumenta la precisión en las estimaciones de diferentes conjuntos de datos o si, por el contrario, su influencia en los resultados es menor; en definitiva, ¿cuánto mejora la estimación con respecto a la formulación original?

Obtener esta información en condiciones diversas a las del estudio de Piironen y Vehtari (2016) sería útil puesto que permitiría saber si la aplicación de la *horseshoe* necesita de condiciones fuertes como la noción a priori de las variables relevantes o si, por el contrario, puede seleccionarse una cantidad, si bien no universal, sí adecuada y con resultados satisfactorios.

Recopilemos la información para tener una visión de conjunto de las ideas hasta ahora expuestas. Por un lado, cuanto mayor sea p_0 mayor será τ_0 (ver Ecuación 13). En la Figura 3 se representa dicha relación:

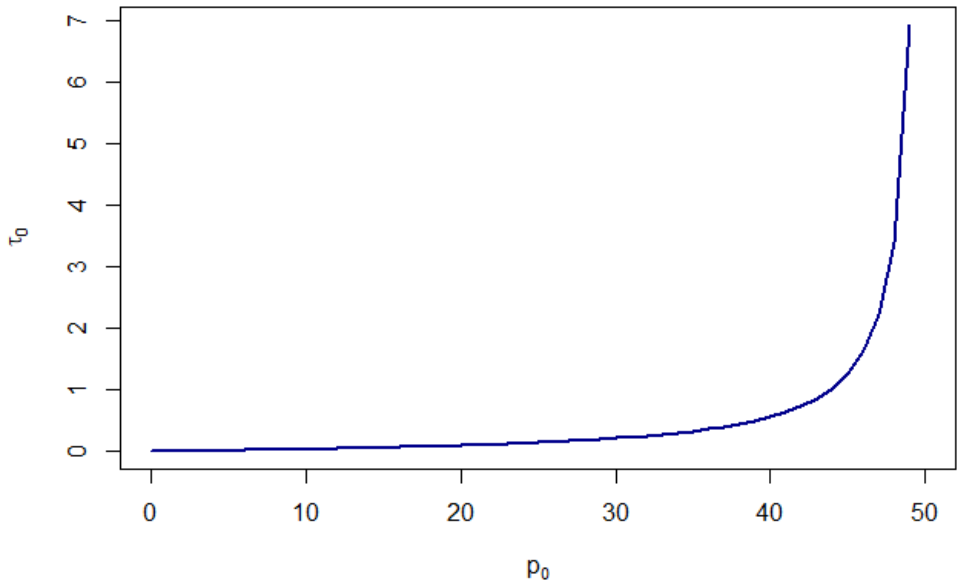

Figura 3. Incremento de valores de τ_0 respecto de p_0.

Este incremento provocará que la cola de la distribución de τ se haga más ancha, lo que significa que se otorgará más probabilidad a valores más altos. En la Figura 4 se muestran las colas de esta distribución para distintos valores de su varianza. Este aumento de la varianza de τ tiene como consecuencia que su potencia como parámetro de contracción global disminuirá; por tanto, en concordancia con lo observado en la Figura 2, es más probable que los diferentes valores de κ sean cercanos a cero y, de esta manera, se favorecerá el hecho de que la media posterior de un determinado coeficiente, obtenida mediante $E(\beta_j|y_i) = (1 - \kappa_j)y_i$, sea distinta de cero.

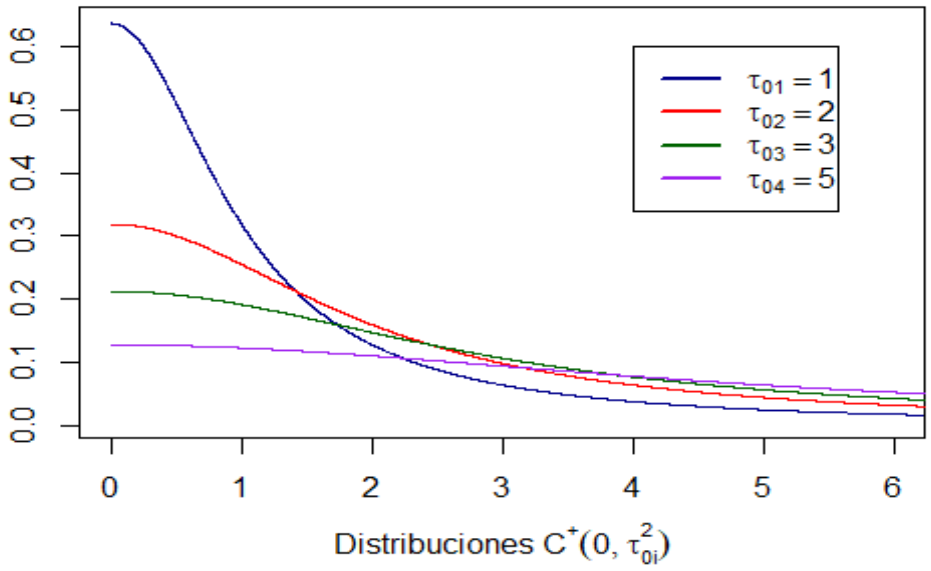

Figura 4. Colas de la distribución half-Cauchy para distintos valores de τ_{0j}. Conforme aumenta su varianza, la distribución se aplana y se da mayor probabilidad a valores más altos.

También es interesante poner en relación la Ecuación 13 con la distribución original $\tau \sim C^+(0,1)$. Cuando la varianza $\tau_0^2 = 1$, podemos obtener el p_0 equivalente, puesto que mediante una pequeña transformación de la Ecuación 13 se llega a

$$p_0 = \frac{\sqrt{n}p}{\sigma+\sqrt{n}}. \tag{14}$$

Esta expresión permite apreciar que establecer $\tau \sim C^+(0,1)$ hace que el componente p_0 sea dependiente del número de observaciones, del total de variables y de la desviación típica error. Esto, según la formulación de Piironen y Vehtari (2017a), equivale a seleccionar un p_0 muy elevado; por ejemplo, para $p = 200$, $n = 100$ y $\sigma = 1$, tendríamos $p_0 = 181.82$. Volviendo al ejemplo, significaría que el investigador tiene la creencia de que la mayoría de las potenciales variables predictoras del modelo terminarán siendo relevantes; esto concordaría con la crítica de que, con la distribución del parámetro global definida como $\tau \sim C^+(0,1)$, se concede una mayor probabilidad a valores grandes. De nuevo, parece pertinente poder conocer el alcance de estas decisiones de cara a futuras aplicaciones con datos reales.

1.1.3.2. Empirical Bayes

Como defensores de esta segunda opción pueden destacarse Van der Pas, Szabó y Van der Vaart (2017a, 2017b). En estos artículos no se enfrentan a problemas de regresión sino a la estimación de parámetros de un vector de medias donde la mayoría son cero,

siguiendo la forma $y_i = \theta_i + \varepsilon_i$. En ellos, proponen obtener en primer lugar el MMLE (Maximum Marginal Likelihood Estimator) de τ, restringiendo su intervalo a $\left[\frac{1}{p}, 1\right]$, cumpliendo también de esta forma con el supuesto de que al menos una y como máximo todas las variables contribuyen al modelo (Van der Pas et al., 2017a).

La discusión entre estos dos frentes continúa abierta. Si bien ambos métodos efectúan un ajuste por la multiplicidad, disminuyendo la probabilidad posterior de inclusión de las variables conforme aumenta el número de predictores ruido (Scott y Berger, 2010), los defensores del "full Bayes" creen que un estimador empírico es habitualmente demasiado pequeño (Datta y Ghosh, 2013) y ofrece problemas en los casos $p > n$ (Piironen et al., 2017a), además de no dar cuenta de la incertidumbre existente (Scott y Berger, 2010); mientras, los seguidores del "empirical Bayes" defienden su eficiencia y un desempeño equiparable a la utilización de un hiperparámetro (Van der Pas et al., 2017a, 2017b).

1.2. Selección de variables

Seleccionar un determinado subconjunto de variables en los modelos de estimación escasa es un tema que cuenta con una bibliografía muy extendida tanto en el ámbito frecuentista (Paul et al., 2008) como en el bayesiano (e.g., Barbieri y Berger, 2004; O'Hara y Sillanpää, 2009; Piironen et al., 2018). Con p predictores pueden construirse hasta 2^p modelos, por lo que evaluar cada uno de estos se convierte en algo virtualmente imposible en situaciones de alta dimensionalidad; además, estas situaciones ponen de relieve que, por muy fuertes que sean las creencias previas del investigador acerca de una o unas pocas combinaciones de variables determinadas, evaluar individualmente un único modelo o un grupo concreto de ellos no sería adecuado desde un punto de vista teórico puesto que esta decisión no refleja la incertidumbre que existe en todo el espacio de modelos (Raftery, Madigan y Hoeting, 1997). Pueden darse dos objetivos diferentes al estudiar modelos de regresión. En palabras de Piironen et al. (2018, p. 4):

> es útil distinguir entre dos problemas diferentes que pueden ser considerados como "selección de variables":
>
> 1. Encontrar un subconjunto mínimo de variables que conduzca a un buen modelo predictivo de *y*, de forma que añadir más variables no mejore considerablemente la capacidad predictiva.

2. Identificar todas las variables que predicen (es decir, que están estadísticamente relacionadas con) la variable objetivo *y*.

Por lo tanto, el propio objeto de estudio puede determinar el método a emplear, bien sea uno que se centre en la capacidad predictiva del modelo final o en la identificación de todas las variables que contribuyen al modelo.

Dentro del primer grupo pueden destacarse métodos que implican una reducción previa de dimensiones de los datos (Bair, Hastie, Paul y Tibshirani, 2006; Piironen y Vehtari, 2017b) o el método *projection predictive* (Piironen y Vehtari, 2017c), cuya idea subyacente es: dado que no podemos conocer cuál es la combinación real de predictores que ha generado la variable dependiente, en primer lugar construimos un modelo de referencia, que puede ser complejo y utilizar muchas variables (Piironen, 2019) y que asumimos es la mejor aproximación al modelo real en términos predictivos; posteriormente tomamos la información proporcionada por la distribución posterior del modelo de referencia y la "proyectamos" sobre un grupo de potenciales modelos restringidos, escogiendo aquel que más se asemeje al modelo de referencia. Puede verse que el *projection predictive* no aspira a discernir el conjunto total de variables relevantes, sino a seleccionar un modelo que sea lo más pequeño posible sin perder ajuste. La sola decisión sobre el proceso a utilizar para obtener el modelo de referencia es todo un campo de estudio actualmente.

El segundo tipo de objetivo (la identificación de todas las variables relevantes), incumbe directamente a las previas de contracción. Hasta ahora se han descrito las propiedades constrictivas de las distribuciones previas. El *spike-and-slab* original (Mitchell y Beauchamp, 1988) permite que se den casos donde el valor posterior de los coeficientes sea cero con probabilidad uno, dado que posee un componente que concentra la probabilidad a priori en un único punto. Sin embargo, en casos de alta dimensionalidad la carga computacional convierte este método en intratable (por ejemplo, Ishwaran y Rao, 2005b; Carvalho et al., 2009), por lo que su modalidad de *spike* continua (George y McCulloch, 1993), junto con sus variantes (Ishwaran y Rao, 2005a y 2005b; Malsiner-Walli y Wagner, 2011, entre otros), es la más extendida. Llegados a este punto, es importante tener en cuenta que ni *spike-and-slab* ni *horseshoe* hacen una selección de variables propiamente dicha, ya que sus distribuciones previas facilitan la contracción hacia el cero de los valores de muchos de los coeficientes, pero su continuidad conlleva que el valor posterior no pueda ser exactamente cero (Bhadra et al., 2019a; Carvalho et

al., 2009; Piironen et al., 2018). Esta continuidad es preferible en algunos casos reales, como estudios genéticos (Bhadra, Datta, Polson y Willard, 2017) donde los genes pueden tener efectos muy pequeños pero distintos de cero. Es necesario, por tanto, establecer un determinado criterio para poder efectuar la selección.

1.2.1. Métodos previos

En cuanto al *spike-and-slab*, George y McCulloch (1993), además de definir este tipo de distribución, la usaron como método de selección del mejor subconjunto de variables -aquel con una mayor probabilidad posterior-, que denominaron *Stocastic Search Variable Selection* (SSVS). Para una revisión del funcionamiento de este y otros métodos, véase el estudio de O'Hara y Sillanpää, (2009), donde se comparan también otras distribuciones que se asemejan más a la *horseshoe* (como el *laplacian shrinkage*) y que no extenderemos más aquí. No obstante, ya se ha mencionado que en los casos con un número de variables p muy grande la probabilidad de inclusión posterior de todos los coeficientes disminuye, por lo que Barbieri y Berger (2004) proponen en su lugar el *Median Probability Model* (MPM), mediante el cual se selecciona aquel modelo cuyos coeficientes excedan en promedio el umbral de probabilidad de .5. Este método ha sido utilizado posteriormente en Ishwaran y Rao (2005a, 2005b), entre otros.

La opción más popular hasta el momento dentro de esta área es el método conocido como *Bayesian Model Averaging* (BMA), que puede encontrarse en Raftery et al. (1997) y Hoeting, Madigan, Raftery y Volinsky (1999) y que consiste en utilizar MCMC para recorrer el espacio de todos los posibles modelos $\mathcal{M} = \{M_1, \ldots, M_J\}$ y estimar la distribución posterior de futuros datos no observados ($\tilde{y}$) dados los datos disponibles D, en un sumatorio ponderado por la probabilidad a posteriori de cada uno de los modelos

$$Pr(\tilde{y}|D) = \sum_{j=1}^{J} \Pr(\tilde{y}|M_j, D)\Pr(M_j|D) \tag{15}$$

Las distribuciones previas en los modelos de este tipo han sido típicamente de la forma *spike-and-slab*, bien sea con la probabilidad del *spike* concentrada en un único punto, bien permitiendo que esta tenga una pequeña varianza (Ecuación 2 con $\varepsilon = 0$ y $\varepsilon > 0$, respectivamente).

1.2.2. *Horseshoe*

En lo referente a la *horseshoe*, Carvalho et al. (2010) proponen un criterio análogo al del *spike-and-slab* basándose en la definición del factor de contracción κ. Si $\hat{\kappa_j}$ es la fuerza de contracción a posteriori que experimenta un coeficiente β_j, este se incluirá en el modelo si $1 - \hat{\kappa_j} \geq .5$. En la literatura sobre la *horseshoe* es habitual encontrar que se compara su desempeño con el BMA por ser el método más extendido. En estos artículos los autores han mostrado que la *horseshoe* con el umbral obtiene resultados equivalentes o incluso mejores que el BMA en algunas condiciones (Carvalho et al., 2009 y 2010; Polson y Scott, 2010 y 2012), a lo que puede sumarse el ahorro de tiempo de computación respecto a la mezcla discreta.

Otro procedimiento se basa en estudiar directamente las distribuciones posteriores, seleccionando aquellas variables cuyos intervalos no incluyan el valor cero. Van der Pas et al. (2017b) llevan a cabo un estudio de simulación donde obtienen τ mediante MMLE y comparan el criterio del umbral y el de las distribuciones. Sus resultados apuntan a que, mientras que ambos detectan correctamente una gran mayoría de las medias fijadas en cero, el criterio del umbral tiende a seleccionar más elementos (tanto verdaderamente relevantes como falsos positivos). En lo tocante a los intervalos posteriores, crecen conforme aumenta el valor simulado de los parámetros, aunque no indefinidamente (Van der Pas et al., 2017b), lo que provoca que algunos de ellos con valores medios no sean seleccionados como relevantes. En cuanto a la decisión de qué amplitud del intervalo utilizar, hay autores que se han decantado por una opción conservadora del 95% (Castillo, Schmidt-Hieber y Van der Vaart, 2015; Van der Pas et al., 2017b) o del 90% (Piironen et al., 2018), mientras que otros prefieren un intervalo de credibilidad liberal, del 50% (Li, Craig y Bhadra, 2019), pretendiendo de esta forma controlar los falsos negativos (variables relevantes no detectadas). Según los autores, la propia naturaleza conservadora de la distribución *horseshoe* mantiene a raya, a su vez, los falsos positivos. En base a los trabajos consultados, nos parece que la información proporcionada por otros autores en lo referente a la amplitud del intervalo está poco justificada o extendida, y entendemos que es conveniente atender a esta decisión, puesto que según Van der Pas et al. (2017b) puede ofrecer mejores resultados que utilizar un umbral y la decisión sobre el intervalo se convierte en algo de gran trascendencia.

El último elemento de interés que vamos a tratar en relación con la selección de variables es el llamado *universal threshold* o umbral universal. Introducido en primer lugar por Donoho y Johnstone (1994), es aplicado por Johnstone y Silverman (2004) al problema del vector escaso (*sparse*) de medias. Ellos proponen seleccionar parámetros basándose en un umbral que se establece en función de la escasez del vector. Si hay poca escasez -es decir, si hay una cantidad grande de medias con valores distintos de cero-, el umbral a partir del cual las seleccionaremos tras la estimación será también pequeño. Dicho umbral aumenta junto con la escasez y, en situaciones con muy pocas medias diferentes de cero, encontraron que el valor más adecuado para seleccionarlas es el umbral universal, que es igual a $U = \sqrt{2\log p}$, siendo p el total de parámetros. Todas aquellas medias por debajo de ese valor no serán incluidas.

Lo que nos interesa de este concepto es su aplicación en el contexto de las estimaciones escasas con *horseshoe*, pues ha sido utilizado por Van der Pas et al. (2014, 2017a, 2017b) para mostrar que, cuando el valor simulado de los parámetros está alrededor o por debajo del umbral universal, sus distribuciones posteriores pueden verse sobrecontraídas, lo que, unido al aumento de los intervalos en función del valor simulado de los parámetros, aumenta el error de las estimaciones. Asimismo, encuentran que la estimación es tanto mejor cuanto mayor es el valor simulado de los parámetros puesto que es más fácil discernir entre aquellos que son cero y los que no.

La forma de obtener τ en estos trabajos es tanto mediante el método *empirical Bayes* como *full Bayes* (Van der Pas et al., 2014, 2017a, 2017b), pero siempre en el contexto de la estimación de un vector de medias escaso (recordemos, de la forma $y_i = \theta_i + \varepsilon_i$) y no en modelos de regresión. Por tanto, se desconoce el efecto de la dimensionalidad (relación entre el número de observaciones y de variables) y, en ese caso, cómo influye la varianza a priori de τ y si muestra robustez a la hora de estimar coeficientes cuyos valores se encuentren en torno o por debajo del umbral universal. De lo expuesto hasta ahora se desprende que, dado este caso, una distribución previa del parámetro global que implique una contracción laxa, es decir, establecer una mayor varianza de τ a priori, podría contrarrestar una posible sobrecontracción cuando el valor de los coeficientes está situado en torno al umbral.

1.2.3. Consideraciones sobre la escasez y la dimensionalidad

Antes de desarrollar en detalle el estudio que nos ocupa, conviene hacer algún comentario al respecto del tratamiento que han recibido en la literatura la alta dimensionalidad y la escasez de los modelos, dos de los principales conceptos con los que estamos trabajando. Esto ayudará a comprender el foco de interés del presente trabajo.

En cuanto al tratamiento de la escasez, cabe mencionar que los estudios se han realizado generalmente con vectores de medias o coeficientes escasos como, por ejemplo, el 20% (Carvalho et al., 2009; Polson y Scott, 2011 y 2012) e incluso mucho más escasos (Bhattacharya, Chakraborty y Mallick, 2016). Estas son decisiones tomadas por los autores y podrían deberse, entre otras cosas, a criterios teóricos, razones de espacio, disponibilidad de recursos o porque las distribuciones previas no fueran el objeto de estudio principal como, por ejemplo, en los artículos sobre el método *projection predictive* (Piironen et al., 2018; Piironen y Vehtari, 2017c). De cara a posibles estudios aplicados a datos reales, el análisis del cumplimiento o incumplimiento del supuesto de escasez y sus efectos en la estimación en modelos de regresión nos parecen temas de gran importancia, especialmente cuando se desconoce el modelo real subyacente, si bien la decisión de aplicar una previa de contracción implica la creencia en que el modelo será escaso, y por tanto la asunción de este supuesto.

La dimensionalidad del modelo en los estudios que se han ido mencionando a lo largo del trabajo es también muy variada. En varios de ellos los modelos de regresión ajustados son del tipo $p < n$ o $p = n$. Cuando el número de observaciones supera al de parámetros (Carvalho et al., 2009) o existe una igualdad como en los casos del vector de medias (por ejemplo, Van der Pas et al., 2014), la *horseshoe* muestra buenos resultados. Sin embargo, no hemos encontrado estudios que traten exhaustivamente los problemas de regresión donde $p > n$ en conjunción con la distribución *horseshoe*. De hecho, es en el reciente artículo de Bhadra et al. (2019b) donde evalúan el comportamiento de la *horseshoe* en diferentes condiciones $p > n$ en comparación con otros métodos. No obstante, no se estudian variaciones en el conjunto de parámetros (e.g. τ) que definen la *horseshoe*.

Resumiendo, dada una situación de regresión con $p > n$, no se ha llevado a cabo aún una comparativa del efecto que tiene la elección de diferentes varianzas previas del parámetro global τ en la estimación de coeficientes de regresión de datos simulados, ni

de la robustez que muestra la distribución *horseshoe* así definida en diferentes niveles de escasez para distintos valores simulados de los coeficientes relevantes.

1.3. Objetivos

1.3.1. Objetivo general

El presente trabajo tiene como objetivo general el estudio, dada una situación específica de dimensionalidad elevada, de las propiedades de estimación de la distribución previa *horseshoe* en modelos de regresión lineal. Se pretende, así, analizar las consecuencias que tiene la decisión previa sobre la varianza del parámetro global en la estimación de las distribuciones posteriores de los coeficientes ante distintas características de escasez, de los valores de los parámetros y de criterio de selección de variables.

1.3.2. Objetivos específicos

Podemos dividir este objetivo general en cuatro objetivos específicos, que recogen las condiciones que vamos a manipular y los resultados esperados de cada uno de ellos:

a) Varianza de τ: queremos conocer si establecer dicha varianza partiendo de la creencia a priori sobre el número de variables relevantes (p_0, Ecuación 13), asumiendo que este número es acertado, mejora la estimación con respecto tanto a una creencia errónea de p_0 como a la formulación original $\tau \sim C^+(0,1)$. Esto permitirá saber cuál es la mejor opción cuando no se tengan conocimientos previos sobre el tamaño del modelo.

   Hipótesis: una decisión acertada respecto a este parámetro aumentará la precisión y el número de aciertos respecto a decisiones extremas. Valores muy bajos de p_0 se traducirán en varianzas pequeñas con alta potencia constrictiva que aumentarán el número de falsos negativos especialmente en casos de baja escasez donde haya muchas variables que contribuyan al modelo; por el contrario, valores muy altos de p_0 darán lugar a varianzas elevadas provocando el efecto contrario, puesto que podrán producir desplazamientos en las distribuciones de ciertas variables ruido, con lo que aumentará el número de falsos positivos.

b) Escasez del modelo: analizaremos la estimación de la distribución *horseshoe* cuando el modelo subyacente es escaso y lo compararemos con la estimación

cuando está conformado por muchas variables predictoras y no se cumple el supuesto de escasez, para comprobar su robustez en este último caso.

Hipótesis: en la situación de no escasez, debido al elevado número de predictores, la penalización automática tenderá a hacer que las distribuciones posteriores sean próximas al cero. Esto, unido a la propia naturaleza de un *shrinkage prior* como la *horseshoe*, provocará que las estimaciones de las distribuciones posteriores de los coeficientes sean más inestables y con una mayor varianza puesto que los parámetros de contracción tirarán de muchas de ellas hacia el cero. Esperamos encontrar, por lo tanto, una mayor cantidad de falsos negativos y, en definitiva, resultados que ofrecen peores estimaciones que los modelos escasos, en los que la estimación será más precisa. Hipotetizamos que, en casos de no escasez, una varianza elevada de τ permitirá aliviar la contracción y aumentará el número de aciertos.

c) Valor de los coeficientes relevantes: basándonos en el concepto de "umbral universal", queremos conocer cómo afecta el valor de los coeficientes a la estimación y selección de variables en función de la varianza de τ y dadas las condiciones de escasez ya mencionadas (escasez y baja escasez).

Hipótesis: esperamos observar que los coeficientes grandes son más fácilmente estimados y, en consecuencia, seleccionados, encontrando menos falsos positivos que aquellos cuyos valores estén cercanos al umbral o por debajo del mismo, donde observaremos una mayor incertidumbre en las estimaciones. Esta inestabilidad será mayor en situaciones de baja escasez, donde el número de falsos negativos será mayor en los modelos con coeficientes bajos. Por el contrario, la identificación de variables relevantes será más precisa en contextos con una alta escasez y coeficientes con valores elevados. No se espera que distintas varianzas de τ afecten a la media posterior de los coeficientes relevantes, dada la capacidad de la *horseshoe* a detectar señales grandes.

d) Intervalos de credibilidad: por último, estudiaremos -de manera exploratoria- el número de variables seleccionadas con diferentes intervalos, para de esta forma percibir diferencias y encontrar en qué condiciones es más adecuado el uso de un intervalo liberal u otro conservador. Si bien es esperable encontrar que un intervalo liberal selecciona más variables que uno conservador puesto que su

amplitud es menor y, por tanto, es menos probable que incluya el cero, la influencia de los factores de contracción o la inestabilidad derivada de la escasez (*sparsity*) y el tamaño de los coeficientes pueden generar situaciones donde existan grandes diferencias en el número de variables seleccionadas por ambos intervalos.

Antes de continuar, hay que mencionar que, en cuanto a la relación entre las variables que conforman la matriz de datos, el estudio de los efectos de la colinealidad sigue siendo un tema activo (Barbieri, Berger, George y Ročková, 2018), así como el de otros supuestos y propiedades de los modelos de alta dimensionalidad (véase Bhadra et al., 2019a). Sin embargo, una parte importante de los artículos han preferido centrarse en las propiedades de las distribuciones previas en situaciones de idoneidad, es decir, con matrices de variables no correlacionadas, para así estudiar su comportamiento sin que este se vea contaminado por los efectos y problemas derivados de la colinealidad (Carvalho et al., 2009; Scott y Berger, 2010; Polson y Scott, 2010). Puesto que nuestro objetivo es someter estas distribuciones a nuevas condiciones, trabajaremos simulando de manera similar, creando matrices de datos con variables no correlacionadas

2. Método

Se van a simular 50 matrices de datos que tendrán $n = 100$ observaciones y $p = 150$ variables. El umbral universal para estos datos será por tanto $U = \sqrt{2\log 150} = 3.166$. Simularemos según

$$Y = \beta X + \varepsilon \tag{16}$$

En este caso, todos los predictores que conforman la matriz de variables independientes $\boldsymbol{X}$ (cuyo tamaño será $n \times p$) se simulan aleatoriamente a partir de la distribución $x_j \sim N(0,1)$; $\boldsymbol{\beta}$ es un vector de longitud p que contiene ceros y los valores de los coeficientes relevantes. De esta manera, sólo se utilizan aquellas variables que queremos que contribuyan a formar el vector respuesta $\boldsymbol{Y}$, al que añadimos un término de error aleatorio, ε. Este error se distribuirá normalmente $\varepsilon_i \sim N(0, \sigma)$ con media 0 y desviación típica ruido $\sigma = 1$, donde $i = 1, \dots, n$. La correlación existente en este caso entre las variables predictoras es únicamente producida por el azar.

Trabajaremos con dos niveles distintos de escasez (*sparsity*), del 10% y del 40%, que llamaremos SP1 y SP2, respectivamente. Por tanto, SP1 contará con $p_{rel1} = 15$ variables

relevantes y con $p_{ru1} = p - p_{rel1} = 135$ variables ruido. Por su parte, SP2 tendrá $p_{rel2} = 60$ y $p_{ru2} = 90$. Esta condición nos permitirá comparar la estimación en condiciones de escasez y de baja escasez.

Vamos a trabajar con tres valores diferentes de los coeficientes relevantes, aplicados a cada uno de los dos niveles de escasez, que serán la condición A1: A_1[1] $= 1$(*i.e.* todos ellos por debajo del umbral); la A2: $A_2 = U + \varepsilon_j$(*i.e.* con valores situados en torno al umbral); y la A3: $A_3 = 3U + \varepsilon_j$(*i.e.* valores muy por encima del umbral). El componente aleatorio se distribuirá $\varepsilon_j \sim N(0, a)$ con $a = 0.5$, para garantizar que sus valores no se alejan demasiado del umbral universal (en el caso de la condición A2), y a su vez asegurar que existe cierta variabilidad (condiciones A2 y A3); el subíndice cubre el rango $j = 1, \dots, p_{rel}$. Para obtener los vectores de coeficientes relevantes de SP2, se añadirán 45 valores nuevos a los de SP1. De esta forma compartirán los coeficientes de las 15 primeras variables. Los demás elementos de los vectores $\boldsymbol{\beta}$ serán iguales a cero. Así podremos estudiar si la estimación de los coeficientes varía según el valor de estos.

Para estudiar la varianza de τ se va a implementar la distribución *regularized horseshoe* (RHS), para la que utilizaremos cuatro varianzas (τ_{0j}^2) diferentes $\tau \sim C^+(0, \tau_{0j}^2)$ partiendo de cuatro valores p_{0j} escogidos de entre todo el espacio $p_0 = [1, \dots, p]$, de más a menos restrictivo. El primero obedecerá a la creencia de que existen muy pocas variables que contribuyen a explicar la variable dependiente, $p_{01} = 3$. El segundo y el tercero coincidirán con el número de variables relevantes simuladas en SP1 ($p_{02} = 15$) y SP2 ($p_{03} = 60$). Según Piironen y Vehtari (2016) serán las decisiones óptimas en sus condiciones. Por último, obtenemos el p_{04} equivalente a la varianza original, $\tau \sim C^+(0,1)$, mediante $p_0 = \frac{\sqrt{n}p}{\sigma + \sqrt{n}}$. Con nuestros datos obtenemos que $p_{04} = \frac{\sqrt{100}150}{1+\sqrt{100}} = 136.364$. La Tabla 1 recoge estos valores junto con las varianzas consiguientes, obtenidas según la Ecuación 13. En cuanto al parámetro c, se distribuirá $c^2 \sim IG(2,8)$, la propuesta poco informativa según Piironen y Vehtari (2017a).

[1] La nomenclatura para los vectores de coeficientes distintos de cero es la utilizada en varias ocasiones en la literatura de este ámbito. Por ejemplo, Piironen y Vehtari (2017b); Van der Pas et al. (2014).

Tabla 1

Equivalencia entre los valores de p_0 y la varianza de τ

	Tau1	Tau2	Tau3	Tau4
$\boldsymbol{p_{0j}}$	3	15	60	136.364
$\boldsymbol{\tau_{0j}^2}$	$.002^2$	$.011^2$	$.067^2$	1^2

Todos los modelos resultantes se ajustarán para cada una de las τ_{0j}^2, por lo que estudiaremos los resultados de un total de 24 condiciones. Las variables se considerarán seleccionadas siempre que el intervalo de credibilidad de su distribución posterior no incluya el valor cero. Los dos criterios que se van a utilizar son uno liberal (intervalo al 50%) y uno conservador (al 90%). Los resultados de las 50 matrices de datos se promediarán; por lo tanto, tendremos la frecuencia media de las veces que una variable ha sido seleccionada en cada una de estas condiciones. Por último, si la variable seleccionada es una variable relevante en el modelo simulado, se clasificará como "acierto" (*HIT*). Si, por el contrario, se ha seleccionado pese a ser una variable ruido, se clasificará como "falso positivo" (FP).

En muchos de los apartados de simulación de los artículos mencionados hasta el momento se utiliza MCMC con Gibbs Sampling y el Metropolis-Hastings (Brooks et al., 2011; Gelman et al., 2013 p.275). Sin embargo, como puede verse en Ghosh et al. (2018), este método es lento en términos de tiempo de computación en comparación con algoritmos más novedosos como el No-U-Turn Sampling (Hoffman y Gelman, 2014) implementado mediante el software R (R Core Team, 2018) en Stan (Stan Development Team, 2018a). Debido a esto, emplearemos el paquete *rstanarm* (Stan Development Team, 2018b), que es muy reciente y, además de implementar las herramientas de muestreo ya mencionadas, incluye la distribución RHS para la manipulación de sus parámetros y cuenta con una sintaxis en R similar a los modelos de regresión frecuentistas que es muy fácil de utilizar. Todo ello hace más eficiente el proceso, simplificando la manipulación de condiciones. Todos los modelos se han ajustado con 4 cadenas y 2000 iteraciones. La convergencia de las cadenas se ha evaluado mediante el estadístico $\hat{R}$ (*potencial scale reduction statistic*), que compara las estimaciones inter e intra-cadenas para ver si estas se han mezclado adecuadamente (Vehtari, Gelman, Simpson, Carpenter

y Bürkner, 2019). Es habitual utilizar el criterio $\hat{R} < 1.05$ para indicar que hay convergencia.

3. Resultados

En este apartado analizaremos los principales resultados obtenidos del ajuste de los modelos. La Tabla 2 muestra el porcentaje de aciertos (*HITS*) y de falsos positivos (FP) obtenidos en cada condición, en función de si el criterio de selección es un intervalo liberal (*i.e.* 50%) o uno conservador (*i.e.* 90%). En el Apéndice A se recogen estos resultados expresados en frecuencias absolutas.

Tabla 2

Porcentaje de aciertos (HITS) y de falsos positivos (FP) en cada una de las condiciones del estudio

			Tau1		**Tau2**		**Tau3**		**Tau4**	
			HITS	**FP**	**HITS**	**FP**	**HITS**	**FP**	**HITS**	**FP**
SP1	A1	50%	100	8.78	100	8.78	100	10.13	100	11.45
		90%	100	**.31**	100	.32	100	.37	100	.46
	A2	50%	100	7.80	100	7.90	100	9.23	100	10.68
		90%	100	**.27**	100	.29	100	.32	100	.50
	A3	50%	100	8.07	100	8.07	100	9.38	100	11.14
		90%	100	**.21**	100	.21	100	.25	100	.36
SP2	A1	50%	58.60	13.64	60.40	13.58	65.63	15.78	76.10	22.47
		90%	11.13	.33	11.87	**.31**	13.90	.47	23.47	.93
	A2	50%	81.80	27.33	81.53	27.44	82.10	27.42	84.70	30.84
		90%	32.27	1.36	32.03	**1.22**	32.90	1.42	38.03	1.87
	A3	50%	83.50	29.31	83.33	29.44	83.60	29.51	85.60	32.04
		90%	31.70	**1.40**	32	1.44	32.50	1.42	37	1.60

La Tabla muestra los resultados en forma de porcentaje para cada una de las varianzas de τ (columnas). Está dividida en las condiciones de escasez (SP1) y no escasez (SP2), para los tres niveles de coeficientes simulados (A1, A2 y A3) y según los intervalos utilizados (50% y 90%). Las cantidades marcadas en negrita muestran el porcentaje más bajo de FP en cada nivel de A.

Esta información nos permite apreciar algunos elementos a simple vista. En primer lugar, si se cumple el supuesto de escasez (pocas variables relevantes, condición SP1), el porcentaje de aciertos es del 100% en todas las condiciones independientemente del valor de los coeficientes y de la potencia restrictiva del parámetro global. Parece que, en la situación de alta dimensionalidad con la que estamos trabajando ($p = 150, n = 100$), ni un tamaño crítico de los coeficientes, es decir, por debajo del umbral universal o en torno

al mismo (A1 y A2, respectivamente), ni una varianza de τ muy restrictiva han provocado falsos negativos (variables simuladas como relevantes pero que no han sido seleccionadas); la *horseshoe* en estas condiciones hace una detección perfecta de las variables señal.

En segundo lugar, podemos observar que el intervalo liberal (50%) aumenta sensiblemente el número de variables ruido seleccionadas o, lo que es lo mismo, los falsos positivos (FP) en todas las condiciones, frente al intervalo al 90% (ver Tabla 2).

En tercer lugar, y dados los dos puntos anteriores, observemos las diferencias existentes en la selección en función del tamaño simulado de los coeficientes en modelos escasos (SP1, mitad superior de la Tabla 2). El porcentaje más bajo de FP se encuentra en la condición Tau1 en los tres niveles de A (señalados en negrita). Puede apreciarse que, con el intervalo al 90%, cuanto mayor es el valor de los coeficientes, mayor es la precisión y disminuye, por tanto, el porcentaje de FP entre A1, A2 y A3 en todos los niveles de varianza de τ. Con el intervalo al 50%, y tomando Tau1 como ejemplo, dicho porcentaje disminuye entre A1 (8.78%) y A2 (7.80%) y vuelve a incrementarse ligeramente en A3 (8.07%).

Lo mismo ocurre con el resto de las varianzas. Observando la mitad superior de la Figura 5, el total de variables seleccionadas con el intervalo al 90% se encuentra por encima, pero cerca de la cantidad simulada en cualquier nivel de A y Tau (barras azules). El total seleccionado por el intervalo liberal (50%, barras rojas) es siempre superior.

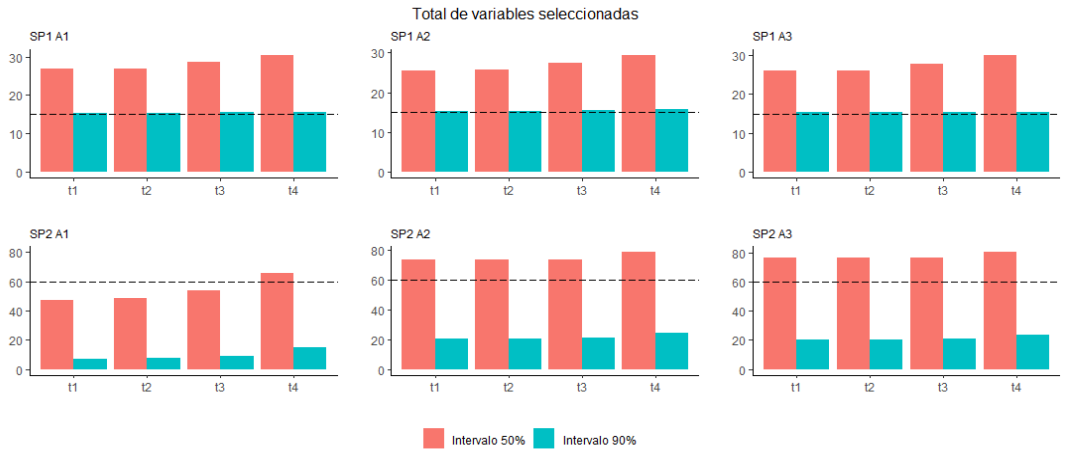

Figura 5. Número total de variables seleccionadas en cada condición (promediando todas las réplicas). Las barras rojas muestran el criterio de selección con un intervalo al 50%, mientras que las azules representan el intervalo al 90%. La línea discontinua horizontal representa el número de variables relevantes simulado en cada condición.

En cuarto lugar, atendiendo ahora a SP2 (mitad inferior de la Tabla 2), dos elementos nos indican la enorme inestabilidad derivada del incumplimiento del supuesto de escasez: el bajo porcentaje de aciertos y la enorme diferencia en dicho porcentaje entre las dos amplitudes del intervalo posterior. La condición A3 es la que mejor ejemplifica este hecho puesto que los valores de los coeficientes se han simulado según $A_3 = 3U + \varepsilon_j$ y, por tanto, están bastante alejados del cero. Con el intervalo al 50% en la condición Tau1 se identificaría el 76.10% de las variables frente al 23.47% del intervalo al 90%; en Tau4, un 85.60% frente al 37%. Esta variabilidad redunda también en el porcentaje de falsos positivos, variando en función del valor real de los coeficientes y del nivel de τ, siendo el mínimo 13.58% (SP2 A1, Tau2 al 50%) y el máximo 32.04% (SP2 A3, Tau4 al 50%). En la Figura 5 puede verse que esto se traduce en la decisión de seleccionar, en términos absolutos, más variables de las realmente relevantes utilizando el intervalo al 50%. Por el contrario, el total seleccionado para el intervalo 90% queda muy por debajo del número real.

En lo referente al tamaño de los coeficientes en SP2 (Tabla 2), se habrá podido advertir que el tanto por ciento de FP, al contrario que en SP1, aumenta conforme lo hace el tamaño de los coeficientes. En consecuencia, podemos ver que, en cualquiera de las varianzas de τ, el paso entre A1 y A2 produce un desplazamiento generalizado de las distribuciones hacia el lado positivo, resultando en un considerable aumento tanto de aciertos como de FP. Entre A2 y A3 este cambio es mucho menos acusado; esto concordaría con los resultados encontrados por Van der Pas et al (2017b) y recogidos en la Sección 1.2.: conforme aumenta el valor de los parámetros a estimar aumenta la amplitud de sus intervalos posteriores, aunque no de forma indefinida, por lo que las mayores diferencias en esta amplitud se producen entre valores pequeños e intermedios de los coeficientes. Esto puede apreciarse también en el total de variables seleccionadas entre A1 y A2 y entre A2 y A3 (Figura 5, mitad inferior).

Para poder observar con más detalle los efectos del incumplimiento del supuesto de escasez y del tamaño de los coeficientes, se han seleccionado una variable señal y una variable ruido (x_4 y x_{90}, respectivamente) para una de las varianzas del parámetro global (Tau3) en una de las réplicas (la número 17); todas estas decisiones se han tomado aleatoriamente. Las distribuciones de estas variables pueden verse en el Apéndice B. De esta forma pueden compararse visualmente las diferencias en la amplitud de los intervalos entre SP1 y SP2 en cada uno de los niveles de A.

Dentro del Apéndice B, en la condición A1 (Figura B.1, Tabla B.1) observamos que en esta realización el valor real del coeficiente ($A_1 = 1$) es poco plausible en SP1 dada la distribución posterior de x_4 (izquierda arriba). No obstante, tampoco se incluye el valor cero, por lo que esta variable ha sido seleccionada con el criterio liberal y con el conservador. En SP2 (izquierda abajo) la media posterior está centrada en cero indicando que, en circunstancias de poca escasez y coeficientes pequeños, pueden aparecer situaciones de sobrecontracción, como en ese ejemplo. La varianza de Tau3 no ha favorecido que la variable haya podido ser seleccionada. Por otro lado, las distribuciones de x_{90} en SP1 y SP2 tienen medias próximas a cero, pero la varianza en SP2 es mucho mayor. En A2 (Figura B.2, Tabla B.2) y A3 (Figura B.3, Tabla B.3) se da una situación similar, donde en condiciones de escasez (SP1, arriba) la precisión es mayor que en SP2.

Por último, el aumento de la varianza $\tau \sim C^+(0, \tau_{0j}^2)$ se traduce en una relajación de la fuerza de contracción del parámetro global, lo cual afecta a las distribuciones posteriores y es acompañada por un incremento en el porcentaje de FP (Tabla 2). En el caso de SP1 con el intervalo al 90%, siempre aumenta entre Tau1 y Tau4 en los tres niveles de A, es decir, Tau1 es la condición con menor ocurrencia de FP. Una elección correcta de p_0 (Tau2) obtiene más FP que en Tau1, pero menos que en Tau3 y Tau4. Eso sí, ninguno supera el .5% (SP1 A2, Tau4 al 90%). En SP2 crece el porcentaje tanto aciertos como de FP al aumentar la varianza de τ (solo disminuyen los FP entre Tau1 y Tau2 en A1 y A2 al 90%; los porcentajes de Tau2 están marcados en negrita en la Tabla 2). El porcentaje de FP más reducido es .31% (SP2 A1, Tau2 al 90%), aunque su porcentaje de aciertos es de un 11.87%. Por lo tanto, la elección adecuada de p_0 (en la condición SP2 es Tau3 con $p_{03} = 60$) ofrece más aciertos y FP que Tau1 y Tau2, pero menos que Tau4.

En cuanto al criterio de convergencia, las estimaciones de todos los parámetros en la condición SP1 han obtenido un $\hat{R} < 1.05$. Esto también ocurre en la condición SP2 (en el 99% de las réplicas, la media de $\hat{R}$ está por debajo de 1.05), aunque se han encontrado parámetros aislados que no cumplen este criterio y muestran un valor un algo superior ($\hat{R} < 1.1$). Los más elevados corresponden al logaritmo posterior de la densidad de la condición SP2 A2, Tau1 (réplica 47, $\hat{R} = 2.44$) y de SP2 A3, Tau3 (réplica 13, $\hat{R} = 1.72$).

4. Discusión

En vista de los resultados obtenidos, parece que basar la selección utilizando una amplitud de los intervalos posteriores del 50% no es adecuado en ninguna de las condiciones propuestas. Pese a que la *horseshoe* es una previa de contracción, pequeñas desviaciones en las distribuciones posteriores de algunos coeficientes provocan que, con este intervalo, determinadas variables ruido sean seleccionadas como relevantes incluso cuando la varianza de τ es muy pequeña y ejerce una fuerte contracción. El uso de un intervalo al 90% permite abarcar la incertidumbre sobre la distribución posterior del parámetro de una forma más adecuada, evitando que estos pequeños desplazamientos conduzcan a seleccionar una variable que tenga muy poco peso en el modelo.

También hemos podido observar en los resultados que el porcentaje de falsos positivos disminuye conforme aumenta el tamaño de los coeficientes (es decir, entre las condiciones A1, A2 y A3) en la condición SP1 y, en cambio, en SP2 dicho porcentaje aumenta al aumentar el tamaño de los coeficientes. Creemos que este comportamiento opuesto puede deberse a la conjunción de una serie de factores: por una parte, en condiciones de escasez (SP1) se cumple la hipótesis de que la precisión en la estimación es mayor cuanto mayor es el valor de los coeficientes; por otra parte, en condiciones de baja escasez (SP2) ocurre que la distribución *horseshoe* se muestra "incapaz" de estimar adecuadamente las distribuciones posteriores cuando existen muchos coeficientes alejados del valor cero. Mientras que la forma de la distribución previa tiende a "arrastrar" muchos coeficientes hacia el cero, la información aportada por la variable observable dice lo contrario. Esto conduce, como puede verse en las distribuciones posteriores de SP2 en el Apéndice B, a un aumento considerable de la varianza de cada coeficiente. De esta manera los datos son congruentes con la hipótesis respecto al incumplimiento del supuesto de escasez dado que se obtienen coeficientes con distribuciones posteriores muy grandes, lo que es indicativo de estimaciones inestables, produciendo además sobrecontracción en la condición A1 y aumento de falsos positivos en A2 y A3. La aparición de falsos negativos es un fenómeno que ocurre sea cual sea el valor de los coeficientes (Tabla 2, SP2). En cambio, podemos decir que la hipótesis de que una varianza grande de τ ayuda a contrarrestar estos efectos no se ha cumplido ya que los porcentajes de Tau1 y Tau4 en SP2 son similares y, además, una varianza más grande obtiene más aciertos, pero también más FP. Los problemas de convergencia según el

estadístico $\hat{R}$ solo han aparecido en la condición SP2, lo cual es otro indicador de la inestabilidad de estas estimaciones.

En lo referente a los efectos de la varianza de τ, los resultados de nuestro trabajo están en consonancia con la teoría (Sección 1.1.3): escoger valores más grandes a priori conlleva que la fuerza de contracción será más débil y la probabilidad de obtener valores más altos de los coeficientes es mayor. Se cumple, por tanto, la hipótesis de que el tanto por ciento de FP se incrementa según aumenta la varianza. Sin embargo, la hipótesis de que un valor muy restrictivo de τ produciría más falsos negativos ha ocurrido únicamente en SP2, donde la ausencia de escasez de los modelos hace más inestables las estimaciones. Además, como hemos visto en los resultados, el aumento en el porcentaje de aciertos y FP en Tau4 respecto de Tau1 es especialmente notable en SP2 A1, condición en la que hay baja escasez y el valor de los coeficientes relevantes está por debajo del umbral universal, por lo que se produce el fenómeno de sobrecontracción.

Recordemos que la varianza de Tau2 ($\tau_{02} = 0.011$) se obtenía a partir de $p_{02} = 15$, es decir, el número de variables relevantes en SP1; la de Tau3 ($\tau_{03} = 0.067$) a partir de $p_{03} = 60$, coincidiendo con el número simulado en SP2. Sin embargo, estas decisiones previas no han resultado ser la mejor estimación en sus respectivas condiciones en ninguno de los niveles de A, por lo que no se cumple la hipótesis de que la decisión adecuada de p_0 mostraría los mejores resultados. Esto, unido a las pequeñas diferencias en FP observadas en las demás condiciones (sobre todo al 90%), nos devuelve a la problemática planteada en la Sección 1.1.3: la propuesta de Piironen y Vehtari (2017a; ver Ecuación 13), ¿mejora sustancialmente las estimaciones respecto a la propuesta original de Carvalho et al. (2009; ver Ecuación 11)? La recomendación que modestamente se desprende de los resultados de este trabajo es que siempre se establezca una varianza de τ por debajo de 1. En nuestra simulación, las condiciones Tau1, Tau2 y Tau3 tenían varianzas por debajo de .1, y pese a ello la *horseshoe* ha mostrado robustez y ha detectado las variables relevantes (en SP1). Esta sería, por lo tanto, una solución a medio camino entre la propuesta original y la de Piironen y Vehtari (2017a): establecer una varianza menor que 1, pero sin necesidad de poseer un conocimiento a priori sobre el tamaño del modelo. Este hecho puede ser de utilidad cuando se aplique este método en análisis con datos reales.

Conviene recordar que, en situaciones aplicadas, la escasez del modelo real subyacente se desconoce y, en tal caso, son razones teóricas las que conducen al investigador a asumir el supuesto de escasez y a utilizar una previa de contracción como la *horseshoe*. Creemos que en estos casos siempre es recomendable estudiar las distribuciones posteriores de los coeficientes con detalle. Si estas muestran una varianza inusualmente grande, esto será un indicativo de la inestabilidad y del posible incumplimiento del supuesto. Esto puede llevar a una reformulación del modelo o a su sustitución (o combinación) con modelos que reduzcan previamente las dimensiones, como los que se han mencionado en la Sección 1.2.

El análisis del criterio de convergencia es también necesario para conocer si se están produciendo estimaciones inestables. En el caso de nuestro estudio de simulación se han encontrado algunas de ellas en pequeña medida. Estos comportamientos son indeseados, pero pueden ocurrir cuando se estiman modelos complejos o de alta dimensionalidad, como en nuestro caso. Además, la distribución del log-posterior (parámetro en el que más problemas se han observado en las diferentes réplicas de SP2) no suele ser Gaussiana, por lo que pueden aparecer valores elevados de $\hat{R}$ (Stan Development Team, 2019). Promediar los resultados a través de varias réplicas permite reducir los posibles efectos adversos de esta inestabilidad ocasional. Un aumento en el número de iteraciones habría ayudado probablemente a reducir alguno de estos problemas, pero el tiempo de computación se habría visto incrementado como consecuencia de ello.

Un factor crucial, que matiza y limita el alcance de la interpretación de estos resultados, es el de la dimensionalidad. Como se ha mencionado en las Secciones 1 y 1.2.3., la incertidumbre de las estimaciones aumenta cuanto mayor es el número de variables respecto al número de observaciones. En una situación $p < n$ o $p = n$ la precisión es más alta y los algoritmos tienen mayor facilidad para converger en torno a un valor incluso cuando el vector de parámetros no es escaso (como en Van der Pas et al., 2014, donde el porcentaje de parámetros relevantes del vector de medias alcanza el 50%). Sin embargo, en situaciones de alta dimensionalidad, donde la complejidad del modelo es mayor, la distribución *horseshoe* se vuelve más sensible a fenómenos como el incumplimiento del supuesto de escasez y el tamaño de los coeficientes. Si el modelo verdadero es escaso, la forma de la distribución representa más fielmente los distintos valores que pueden tomar los coeficientes (como en la condición SP1) y las estimaciones son precisas a pesar de la alta dimensionalidad de nuestra simulación. Si, en cambio,

muchos de ellos se alejan del cero, en condiciones $p > n$ la *horseshoe* deja de ser una distribución adecuada puesto que, si el número de observaciones es pequeño, la información proporcionada por la variable observable puede no ser suficiente, y en esos casos la distribución previa puede tener mucho peso en la inferencia posterior (Kruschke, 2014), por lo que aumenta la varianza de las distribuciones posteriores y se produce una sobrecontracción si el tamaño de los coeficientes es pequeño (como en la condición SP2 A1, por debajo del umbral universal).

5. Conclusiones

El trabajo ha mostrado que el uso de la distribución previa *horseshoe* en modelos de regresión $p > n$ es una herramienta útil que ofrece estimaciones precisas de las distribuciones posteriores de los coeficientes. La elección cuidadosa de la varianza del parámetro global y de la amplitud del intervalo posterior como criterio de selección contribuyen al aumento de la precisión en las estimaciones y a la disminución de falsos negativos y FP en la selección de variables.

Cuando el modelo verdadero es escaso, la *horseshoe* ha demostrado robustez y capacidad para seleccionar un número adecuado de variables independientemente del valor real de los coeficientes incluso en condiciones de alta dimensionalidad.

A partir de los resultados obtenidos y algunas pruebas realizadas, pensamos que parece existir un equilibrio entre dimensionalidad y escasez que separa las estimaciones precisas de las inestables o, dicho de otra manera, un límite en la dimensionalidad del modelo a partir del cual la *horseshoe* es sensible al incumplimiento del supuesto de escasez y sus estimaciones empeoran. La principal limitación del trabajo ha sido el no abordar esta problemática en profundidad. Para poder dar una respuesta más firme a esta cuestión habría sido necesario comparar los resultados de la actual dimensionalidad ($n = 100, p = 150$) con otros donde variase esta condición, incluyendo casos $p < n$. Una segunda limitación es que, dado que el uso del *spike-and-slab* está más extendido, los resultados se verían reforzados de haber ajustado los modelos con esta distribución en las diferentes condiciones, y ver las diferencias en las estimaciones entre esta y la *horseshoe*. En tercer lugar, un estudio más detallado de los problemas de convergencia permitiría conocer las circunstancias en las que se producen, y las características de los datos y los parámetros en esas situaciones.

El estudio de los efectos tanto de la colinealidad como de la desviación típica error exceden los objetivos de este trabajo. No obstante, las decisiones tomadas en lo referente a estos dos elementos (predictores no correlacionados y $\sigma = 1$, respectivamente) limitan la generalizabilidad del trabajo, puesto que la existencia de multicolinealidad o valores extremos de σ pueden producir importantes alteraciones en las estimaciones y desconocemos el comportamiento que mostraría la *horseshoe*. La falta de tiempo y recursos computacionales han sido los motivos por los que no se han podido explorar estas cuestiones.

En futuras investigaciones sería interesante, por tanto, profundizar en la relación entre la dimensionalidad y el supuesto de escasez para complementar la información que se posee sobre esta distribución y conocer las situaciones en la que puede ser, o no, convenientemente utilizada. Otras líneas de investigación pertinentes son el estudio de los efectos que provoca la colinealidad en las estimaciones, las consecuencias de valores extremos de σ y de los coeficientes relevantes, y los resultados que ofrece la *horseshoe* en combinación con métodos como *projection predictive*.

Por la flexibilidad que permite en su definición, el ahorro en tiempo de computación frente a otros métodos y su sencilla implementación a través herramientas como *rstanarm*, la *horseshoe* se presenta como una alternativa atractiva para aquellos analistas que se enfrenten a un problema de regresión en condiciones $p > n$.

Referencias

Bair, E., Hastie, T., Paul, D. y Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473), 119-137.

Barbieri, M. M. y Berger, J. O. (2004). Optimal predictive model selection. *The annals of statistics*, 32(3), 870-897.

Barbieri, M., Berger, J. O., George, E. I. y Ročková, V. (2018). The median probability model and correlated variables. *arXiv preprint arXiv:1807.08336*.

Betancourt, M., Byrne, S., Livingstone, S. y Girolami, M. (2017). The geometric foundations of hamiltonian monte carlo. *Bernoulli*, 23(4A), 2257-2298.

Bhadra, A., Datta, J., Polson, N. G. y Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, *12*(4), 1105-1131.

Bhadra, A., Datta, J., Polson, N. G. y Willard, B. (2019a). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3), 405-427.

Bhadra, A., Datta, J., Li, Y., Polson, N. G. y Willard, B. T. (2019b). Prediction Risk for the Horseshoe Regression. *Journal of Machine Learning Research*, 20(78), 1-39.

Bhattacharya, A., Pati, D., Pillai, N. S. y Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512), 1479-1490.

Bhattacharya, A.K., Chakraborty, A. y Mallick, B.K. (2015). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *Biometrika*, 103 4, 985-991.

Brooks, S., Gelman, A., Jones, G. y Meng, X. L. (2011). *Handbook of markov chain monte carlo*. CRC Press.

Carvalho, C. M., Polson, N. G. y Scott, J. G. (2009). Handling sparsity via the horseshoe. *In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 73-80.

Carvalho, C. M., Polson, N. G. y Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465-480.

Castillo, I., Schmidt-Hieber, J. y Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5), 1986-2018.

Clarke, B., Fokoue, E. y Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Nueva York, NY: Springer Science & Business Media.

Datta, J. y Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1), 111-132.

Donoho, D. L. y Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425-455.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3), 515-534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. y Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

George, E. I. y McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881-889.

Ghosh, J., Li, Y. y Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2), 359-383.

Griffin, J. E. y Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1), 171-188.

Hastie, T., Tibshirani, R. y Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Boca Raton, FL: CRC press.

Heinze, G. y Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16), 2409-2419.

Hoeting, J. A., Madigan, D., Raftery, A. E. y Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382-401.

Hoffman, M. D. y Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593-1623.

Ishwaran, H. y Rao, J. S. (2005a). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association*, *100*(471), 764-780.

Ishwaran, H. y Rao, J. S. (2005b). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730-773.

Johndrow, J. E., Mattingly, J. C., Mukherjee, S. y Dunson, D. (2015). Approximations of Markov chains and high-dimensional Bayesian inference. *arXiv preprint arXiv:1508.03387*.

Johnstone, I. M. y Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4), 1594-1649.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M. y Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, *19*(1), 90-97.

Li, Y., Craig, B. A. y Bhadra, A. (2019). The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3), 747-757.

Malsiner-Walli, G. y Wagner, H. (2011). Comparing Spike and Slab Priors for Bayesian Variable Selection. *Austrian Journal of Statistics*, 40(4), 241-264.

Mitchell, T. J. y Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023-1032.

O'Hara, R. B. y Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, *4*(1), 85-117.

Pardo, A. y San Martín, R. (1994). *Análisis de datos en psicología II*. Madrid: Pirámide.

Paul, D., Bair, E., Hastie, T. y Tibshirani, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, *36*(4), 1595-1618.

Peltola, T., Marttinen, P., Jula, A., Salomaa, V., Perola, M. y Vehtari, A. (2012). Bayesian variable selection in searching for additive and dominant effects in genome-wide data. *PLoS One*, *7*(1), e29115.

Piironen, J. (2019). *Bayesian Predictive Inference and Feature Selection for High-Dimensional Data*. Aalto University.

Piironen, J. y Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *arXiv preprint arXiv:1610.05559*.

Piironen, J. y Vehtari, A. (2017a). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018-5051.

Piironen, J. y Vehtari, A. (2017b). Iterative supervised principal components. *arXiv preprint arXiv:1710.06229.*

Piironen, J. y Vehtari, A. (2017c). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711-735.

Piironen, J., Paasiniemi, M. y Vehtari, A. (2018). Projective inference in high-dimensional problems: prediction and feature selection. *arXiv preprint arXiv:1810.02406*.

Polson, N. G. y Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian statistics*, 9(501-538), 105.

Polson, N. G. y Scott, J. G. (2012). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, 74(2), 287-311.

Raftery, A. E., Madigan, D. y Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437), 179-191.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/

Robert, C. y Casella, G. (2013). *Monte Carlo statistical methods*. Nueva York, NY: Springer Science & Business Media.

Ročková, V. y George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506), 828-846.

Sanyal, N. y Ferreira, M. A. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *NeuroImage*, 63(3), 1519-1531.

Scott, J. G. y Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of statistical planning and inference*, 136(7), 2144-2162.

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

Stan Development Team (2019) Stan Reference Manual. Version 2.23.

Stan Development Team (2018a) RStan: The R interface to Stan. R package version 2.17.3. URL http://mc-stan.org

Stan Development Team (2018b) RStanArm: Bayesian applied regression modeling via Stan. R package version 2.17.4. URL http://mc-stan.org

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Van Der Pas, S. L., Kleijn, B. J. y Van Der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2), 2585-2618.

Van der Pas, S., Szabó, B. y van der Vaart, A. (2017a). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics*, 11(2), 3196-3225.

Van der Pas, S., Szabó, B. y van der Vaart, A. (2017b). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4), 1221-1274.

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M. y Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., Bürkner, P. (2019). Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC. *arXiv:1903.08008*

APÉNDICE A

Tabla A.1

Promedio de variables seleccionadas en cada condición en frecuencias absolutas

SP1		**Tau1**		**Tau2**		**Tau3**		**Tau4**	
$p_{rel1} = 15$		**HITS**	**FP**	**HITS**	**FP**	**HITS**	**FP**	**HITS**	**FP**
A1	50%	15	11.86	15	11.86	15	13.68	15	15.46
		(0)	(4.32)	(0)	(4.29)	(0)	(4.25)	(0)	(5.26)
	90%	15	.42	15	.44	15	.50	15	.62
		(0)	(.83)	(0)	(.73)	(0)	(.89)	(0)	(1.00)
A2	50%	15	10.54	15	10.66	15	12.46	15	14.42
		(0)	(4.54)	(0)	(4.39)	(0)	(4.89)	(0)	(6.26)
	90%	15	.36	15	.40	15	.44	15	.68
		(0)	(.83)	(0)	(.83)	(0)	(1.07)	(0)	(1.33)
A3	50%	15	10.90	15	10.90	15	12.66	15	15.04
		(0)	(6.15)	(0)	(6.10)	(0)	(6.39)	(0)	(7.98)
	90%	15	.28	15	.28	15	.34	15	.48
		(0)	(0.71)	(0)	(0.71)	(0)	(0.71)	(0)	(0.84)

SP2		**Tau1**		**Tau2**		**Tau3**		**Tau4**	
$p_{rel2} = 60$		**HITS**	**FP**	**HITS**	**FP**	**HITS**	**FP**	**HITS**	**FP**
A1	50%	35.16	12.28	36.24	12.22	39.38	14.20	45.66	20.22
		(11.03)	(6.46)	(9.35)	(6.15)	(6.61)	(5.33)	(4.59)	(5.17)
	90%	6.68	.30	7.12	.28	8.34	.42	14.08	.84
		(5.28)	(.61)	(5.08)	(.57)	(4.78)	(0.83)	(4.57)	(1.20)
A2	50%	49.08	24.60	48.92	24.52	49.26	24.68	50.82	27.76
		(3.58)	(5.34)	(3.92)	(5.53)	(3.46)	(5.28)	(2.84)	(4.90)
	90%	19.36	1.22	19.22	1.10	19.74	1.28	22.82	1.68
		(6.02)	(1.20)	(5.65)	(1.15)	(5.49)	(1.16)	(4.21)	(1.41)
A3	50%	50.10	26.38	50.00	26.50	50.16	26.56	51.36	28.84
		(3.73)	(6.26)	(3.42)	(6.03)	(3.45)	(5.97)	(2.85)	(5.54)
	90%	19.02	1.26	19.20	1.30	19.50	1.28	22.20	1.44
		(5.74)	(1.40)	(5.53)	(1.50)	(5.36)	(1.38)	(4.88)	(1.57)

La Tabla muestra la frecuencia de selección de variables en cada condición tras promediar las 50 réplicas, junto con la desviación típica (entre paréntesis). La mitad superior corresponde a la condición SP1 y la inferior a SP2.

APÉNDICE B

A modo de ejemplo, en este Apéndice se presentan las distribuciones posteriores de una variable señal (X4) y una variable ruido (X90) seleccionadas dentro de una de las réplicas (la número 17) para un valor de τ (Tau3); todas las decisiones han sido al azar.

Tabla B.1

Valores simulados y medias estimadas de los coeficientes de las variables X4 y X90 en la condición A1

A1	X4 = 1		X90 = 0	
	SP1	SP2	SP1	SP2
Media	.75	-.01	.04	.13
Desv. Típica	.12	.38	.07	.35

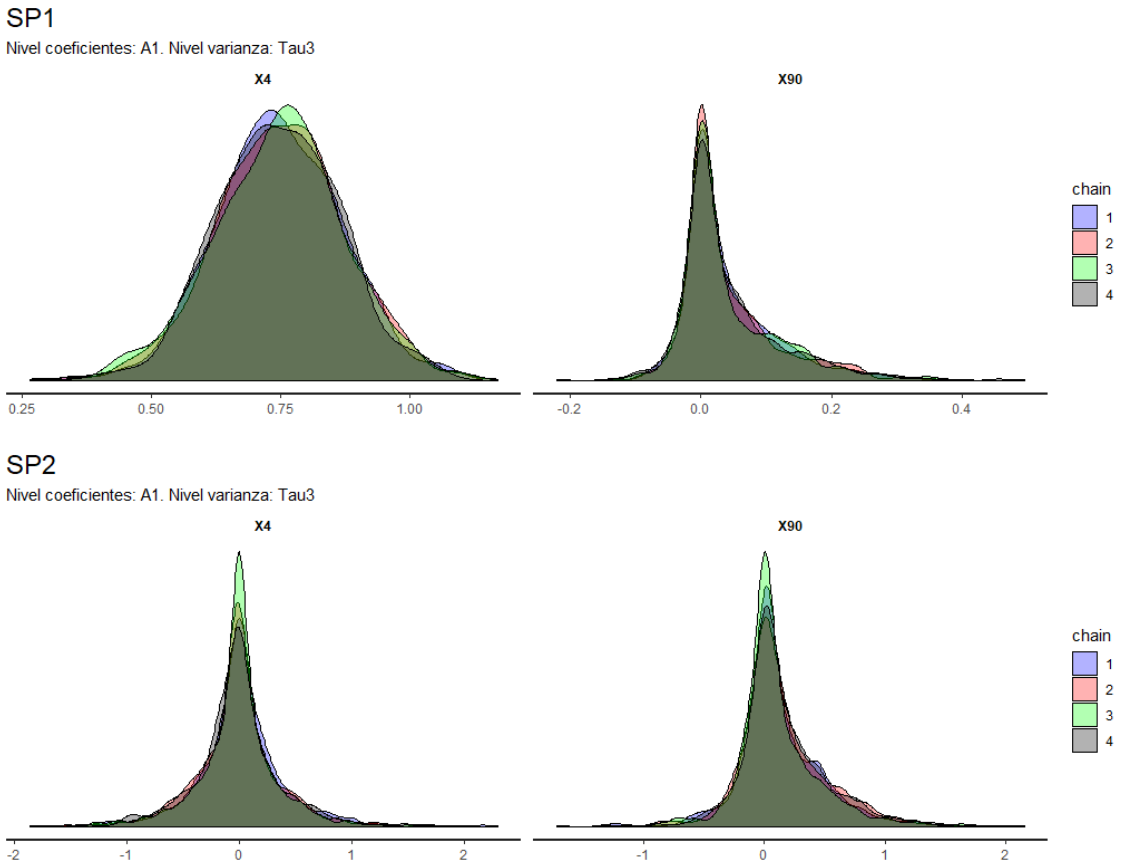

Figura B.1. Distribuciones posteriores de las variables X4 y X90 en las condiciones SP1 A1 (arriba) y SP2 A1 (abajo). Los colores muestran las cuatro cadenas de Markov utilizadas.

Tabla B.2

Valores simulados y medias estimadas de los coeficientes de las variables X4 y X90 en la condición A2

A2	X4 = 2,956		X90 = 0	
	SP1	SP2	SP1	SP2
Media	2.79	2.04	-.02	-.25
Desv. Típica	.13	1.45	.06	1.12

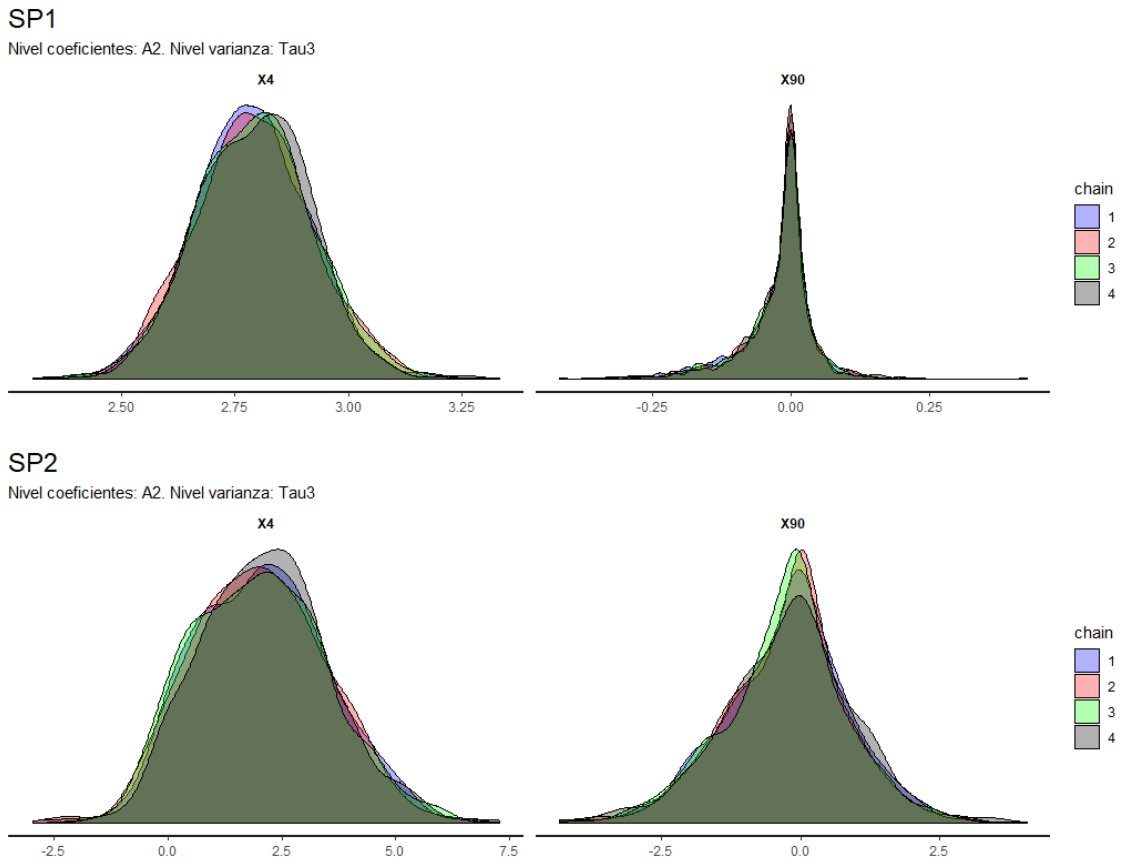

Figura B.2. Distribuciones posteriores de las variables X4 y X90 en las condiciones SP1 A2 (arriba) y SP2 A2 (abajo). Los colores muestran las cuatro cadenas de Markov utilizadas.

Tabla B.3

Valores simulados y medias estimadas de los coeficientes de las variables X4 y X90 en la condición A3

A3	X4 = 9,53		X90 = 0	
	SP1	SP2	SP1	SP2
Media	9.33	9.45	0	-2.26
Desv. Típica	.12	4.24	.11	3.72

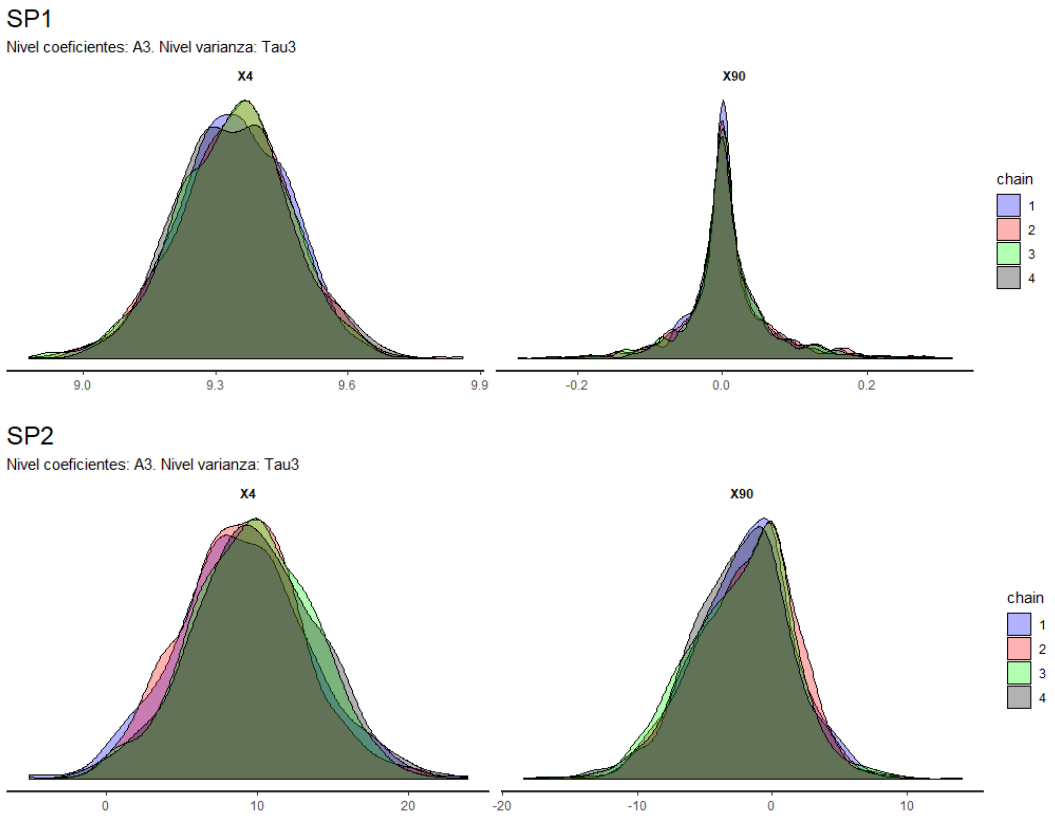

Figura B.3. Distribuciones posteriores de las variables X4 y X90 en las condiciones SP1 A3 (arriba) y SP2 A3 (abajo). Los colores muestran las cuatro cadenas de Markov utilizadas.