Allasia, María Belén; Branco, Márcia D´Elia; Quaglino, Marta Beatriz
Instituto de Investigaciones Teóricas y Aplicadas en Estadística

REGRESIÓN LASSO BAYESIANA. AJUSTE DE MODELOS LINEALES PENALIZADOS MEDIANTE LA ASIGNACIÓN DE PRIORES NORMALES CON MEZCLA DE ESCALA.

Resumen:

Uno de los desafíos más importantes del análisis estadístico en grandes volúmenes de datos es identificar aquellas variables que provean información valiosa, haciendo una selección de variables predictoras. La estimación Lasso (Least Absolute Shrinkage and Selection Operator) para el modelo de regresión lineal puede ser interpretada desde el enfoque Bayesiano como la moda a posteriori cuando los coeficientes de regresión tienen distribución priori doble exponencial independientes. Al representar dicha distribución como una distribución Normal con mezcla de escala, es factible la construcción de un modelo jerárquico mediante la introducción de un vector de variables latentes, conjugando una distribución priori normal para los parámetros de regresión y prioris exponencial independientes para sus respectivas variancias. Mediante la implementación del algoritmo de simulación de Gibbs a partir de las distribuciones condicionales completas, se obtienen secuencias que permiten estimar cualquier característica de interés de la distribución a posteriori de manera sencilla. La regresión Lasso Bayesiana tiene una enorme ventaja sobre el método clásico, dado que permite mejorar sustancialmente la inferencia, especialmente en el contexto de muchas variables predictoras.

Palabras claves: Data Mining – Regresión penalizada – Selección de variables

Abstract:

One of the most important challenges of statistical analysis in big data is to identify those variables that provide valuable information, making a selection of predictor variables. The Lasso (Least Absolute Shrinkage and Selection Operator) estimate for the linear regression model can be interpreted from the Bayesian approach as a posterior mode estimate when the regression parameters have independent double-exponential priors. Representing such distribution as a scale mixture of normals, it is feasible to construct a hierarchical model by introducing a vector of latent variables, with conjugate normal priors for the regression parameters and independent exponential priors on their variances. By implementing the simulation Gibbs algorithm from complete conditional distributions, the obtained sequences allow to estimate any characteristic of interest based on the posterior distribution in a simple way. The Bayesian Lasso regression has a huge advantage over conventional methods; it substantially improves inference, especially in the context of many predictor variables.

Keywords: Data Mining – Penalized regression – Variable Selection

INTRODUCCIÓN

Uno de los desafíos más importantes del análisis estadístico en grandes volúmenes de datos es identificar aquellas variables que provean información valiosa. Cuando el interés radica en ajustar un modelo de regresión, la elección del mejor subconjunto de variables predictoras es una de las cuestiones clave en la formulación del mismo.

El método usual de estimación de modelos y selección de variables (método de mínimos cuadrados) produce estimadores insesgados, pero en contextos con gran número de predictores se incrementa la variancia de los estimadores.

Con el objetivo de mejorar las predicciones, es decir, reducir dicha variancia de los estimadores, se han propuesto métodos de penalización (o regularización) que consisten en ajustar el modelo eliminando predictores en el proceso de estimación -se fuerza que los coeficientes pequeños sean igual a cero-, controlando el impacto relativo de la penalización sobre la estimación de los coeficientes y admitiendo algo de sesgo. Entre estos métodos, la regresión Lasso (Least Absolute Shrinkage and Selection Operator) de Tibshirani (1996) se ha convertido en una alternativa ampliamente utilizada, principalmente por su capacidad de realizar simultáneamente una estimación y el procedimiento de selección de variables.

En este trabajo se presentan las generalidades de esta metodología, incorporando métodos bayesianos para la estimación de los parámetros de regresión, estrategia que permite mejorar sustancialmente la inferencia, especialmente en el contexto de muchas variables predictoras.

REGRESION LASSO BAYESIANA

Se plantea un modelo:

$$\boldsymbol{Y} = \mu \mathbf{1}_n + \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

donde $\boldsymbol{Y}$ es un vector $n \times 1$ que contiene la variable respuesta para cada individuo, μ es la media general, $\boldsymbol{X}$ es la matriz $n \times p$ de regresores estandarizados y $\boldsymbol{\varepsilon}$ es el vector $n \times 1$ de errores aleatorios independientes e idénticamente distribuidos normalmente con media cero y variancia constante (σ^2).

El método Lasso para estimar los parámetros de regresión $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$, consiste en minimizar la suma de cuadrados residual (SCR) controlando la norma L_1 del vector de coeficientes β.

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{Lasso} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X\beta}\right)' \left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X\beta}\right) + \lambda \|\beta\|_1 \right\} \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\tilde{y}_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \end{aligned} \tag{2}$$

donde, $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \bar{y}\mathbf{1}_n$ y $\lambda \geq 0$ determina la influencia de la penalización en la estimación. Si $\lambda = 0$ el estimador $\widehat{\boldsymbol{\beta}}_{Lasso}$ es idéntico al estimador minímo-cuadrático ($\widehat{\boldsymbol{\beta}}_{MC}$) y un λ suficien-

temente grande, produce un $\widehat{\boldsymbol{\beta}}_{Lasso} = \mathbf{0}$. La selección de este parámetro es clave ya que existe un conjunto $(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$ distinto para cada valor posible de λ. Interesa escoger aquel que provea el mejor ajuste de los datos, aumentando la exactitud de las estimaciones. Esto puede realizarse, por ejemplo, mediante validación cruzada (Efron et al, 2004).

La Regresión Lineal Bayesiana es un enfoque de regresión lineal en el que el análisis estadístico se realiza dentro del contexto de la Inferencia Bayesiana. Cuando el modelo de regresión tiene errores que tienen una distribución normal, si se asume una forma particular de distribución a priori para los parámetros desconocidos del modelo, se podrá disponer de resultados explícitos para las distribuciones de probabilidad a posteriori de los mismos.

El método Lasso tiene una interpretación particular desde el enfoque Bayesiano. El estimador $\widehat{\boldsymbol{\beta}}_{Lasso}$ puede pensarse como la moda de la distribución a posteriori de $\boldsymbol{\beta}$ -esto es $\widehat{\boldsymbol{\beta}}_{Lasso} = \operatorname{argmax}_{\boldsymbol{\beta}} f(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \lambda)$- cuando los parámetros de regresión tienen a priori distribución idéntica e independiente Laplace (doble exponencial)

$$\boldsymbol{\beta} \sim f(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}} \propto e^{-\frac{\lambda}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j|}, \qquad (3)$$

y se considera la priori marginal no informativa $f(\sigma^2) \propto \frac{1}{\sigma^2}$ para σ^2 o cualquier distribución gamma inversa, para mantener la conjugación. El empleo de la distribución condicionada a σ^2 (3), garantiza una distribución unimodal (Park & Casella, 2008).

El primer procedimiento explícito para la regresión Lasso Bayesiana fue proporcionado por Park & Casella (2008), utilizando la representación de la distribución doble exponencial como mezcla de escala de normales (Andrews & Mallows, 1974):

$$\frac{\lambda}{2\sqrt{\sigma^2}} e^{-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}} = \int_0^{+\infty} \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2s}\beta_j^2} \frac{\lambda^2}{2\sigma^2} e^{-\frac{\lambda^2 s}{2\sigma^2}} ds \qquad (4)$$

Esta función permite crear una formulación jerárquica del modelo, mediante la introducción de un vector de variables latentes.

Para obtener la distribución a posteriori, Casella & George en 1992 proponen utilizar una muestra obtenida mediante el algoritmo de Gibbs.

Otros autores (Figueredo, 2003; Bae & Mallick, 2004) han propuesto modelos jerárquicos similares, pero con diferentes algoritmos para la obtención de la muestra a posteriori por simulación. Por su parte, Hans (2009) planteó un nuevo método de Gibbs Sampling para la regresión Lasso Bayesiana, a través de una nueva caracterización de la distribución a posteriori para conseguir mejoras en las predicciones mediante el modelo.

Cuando el número de variables del modelo de regresión es muy grande (incluso mayor que el número de observaciones), o si existe una correlación lineal entre los predictores, la selección de variables mediante los métodos usuales generalmente es ineficiente. Por lo que la importancia de los métodos bayesianos se hace más evidente, ya que al establecer el comportamiento de los coeficientes a posteriori permiten calcular, entre otros, el nivel de incertidumbre para cualquier modelo elegido -esto es, $Var(\hat{\beta}_j)$-. En este sentido, la regresión

Lasso Bayesiana supera a la regresión Lasso estándar.

MODELO JERÁRQUICO

El modelo jerárquico completo propuesto por Park & Casella (2008) es:

$$\boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N_n(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

$$\boldsymbol{\beta} | \sigma^2, \tau_1^2, \tau_2^2, \dots, \tau_p^2 \sim N_p\left(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau\right),$$

$$\boldsymbol{D}_\tau = \begin{pmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_p^2 \end{pmatrix}, \qquad (5)$$

$$\sigma^2, \tau_1^2, \tau_2^2, \dots, \tau_p^2 \sim f(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \tau_j^2}{2}} d\tau_j^2,$$

$$\sigma^2, \tau_1^2, \tau_2^2, \dots, \tau_p^2 > 0$$

Para el parámetro μ puede colocarse una distribución plana independiente de los restantes parámetros, y como las columnas de $\boldsymbol{X}$ están centradas, puede obtenerse analíticamente y de manera fácil la distribución a posteriori de μ. Como raramente es de interés este parámetro, se suele marginalizar -sin afectar la conjugación- para conseguir mayor simplicidad en el modelo y rapidez en la convergencia de los algoritmos computacionales. Por otra parte, como ya se ha enunciado, $f(\sigma^2)$ puede ser una priori impropia (no informativa) o bien gamma inversa.

A partir del modelo (5), es factible la ejecución del algoritmo de Gibbs Sampling (GS) el cual permite generar variables aleatorias provenientes de una distribución marginal indirectamente, esto es, sin necesidad de calcular o conocer dicha densidad. Esta técnica se basa en las propiedades elementales de las Cadenas de Markov tales como la convergencia de la distribución empírica de la secuencia a la distribución de interés y permite realizar cálculos complejos a través de una serie de cálculos más sencillos. Para la aplicación de este método es necesaria la definición de las distribuciones condicionales completas.

En este caso interesa conocer los parámetros de regresión $(\beta_1, \beta_2, \dots, \beta_p)$, el objetivo es estimarlo como la moda a posteriori de la distribución de los mismos ($\widehat{\boldsymbol{\beta}}_{Lasso}$) pero la distribución $f(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \lambda)$ es desconocida, o bien, la integración que debería realizarse para su cálculo es demasiado compleja. Como éste método realiza simultáneamente estimación y selección de variables, se aplica la noción propuesta por George & McCulloch (1993) para la selección vía GS, en la cual el modelo de regresión canónico se rescribe a través de un modelo jerárquico.

Para dar comienzo a la simulación se requiere establecer valores iniciales para los parámetros desconocidos. En general, los autores sugieren que se utilicen los estimadores mínimo-cuadráticos del modelo ($\widehat{\boldsymbol{\beta}}_{MC}$ y $\hat{\sigma}^2_{MC}$) y se realice una simulación sucesiva de las funciones condicionales completas reemplazando los términos requeridos por sus valores actualizados.

Si el número de iteraciones es suficientemente grande, la secuencia obtenida mediante GS puede pensarse como una muestra de observaciones independientes proveniente de la distribución $f(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \lambda)$, permitiendo así estimar cualquier característica de interés con el grado de exactitud que se desee.

ELECCIÓN DEL PARÁMETRO λ

Interesa escoger un valor para el parámetro λ que provea el mejor ajuste de los datos, aumentando la exactitud de las estimaciones, por lo que su elección es un problema de importancia para el ajuste del modelo.

En la regresión Lasso estándar es usual que el parámetro λ del modelo (2) se escoja mediante validación cruzada, validación cruzada generalizada o alguna alternativa basada en la estimación insesgada del riesgo (Tibshirani, 1996), los que poseen cierto grado de intervención del estadístico en los valores del mismo.

Para la regresión Lasso Bayesiana se proponen métodos ligados a este enfoque estadístico, que al utilizarse permiten la elección del parámetro de una manera más automática.

Bayes empírico mediante máxima verosimilitud marginal

Casella (2001) propuso un algoritmo que complementa a GS y provee estimaciones de máximo verosimilitud marginal para hiperparámetros (parámetros de las distribuciones a priori).

Para la regresión Lasso esta técnica consiste en utilizar para cada iteración del algoritmo un valor de λ estimado a partir de la muestra obtenida en la iteración anterior. Se establece un valor inicial $\lambda^{(0)}$, el cual se sugiere como una construcción de los estimadores mínimo-cuadráticos:

$$\lambda^{(0)} = \frac{p\sqrt{\hat{\sigma}^2_{MC}}}{\sum_{j=1}^{p}|\hat{\beta}_j^{MC}|} \qquad (6)$$

y en cada iteración k se usa GS con hiperparámetro $\lambda^{(\mathrm{k}-1)}$ para aproximar una estimación actualizada ideal:

$$\lambda^{(k)} = \frac{\sqrt{2p}}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}\left(\tau_j^2|\tilde{\boldsymbol{y}}\right)} \qquad (7)$$

reemplazando las esperanzas condicionales de (7) por los promedios de la muestra de GS.

Hiperpriori para el parámetro Lasso

Otra alternativa es elegir λ especificando una distribución hiperpriori difusa, que p. Se considera la clase de prioris gamma para λ^2:

$$f(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2}, \qquad (\lambda^2, r, \delta > 0) \qquad (8)$$

Observar que la hiperpriori no se especifica sobre λ, sino sobre λ^2, esto es para mantener el

grado del parámetro en la conjugación y permitir una extensión sencilla del GS.

Al utilizar la priori gamma (8) en el modelo jerárquico, la distribución condicional completa de λ^2 es $G\left(p+r; \sum_{j=1}^{p} \tau_j^2/2 + \delta\right)$; mientras que si se utiliza una priori impropia para λ^2, $f(\lambda^2) = \frac{1}{\lambda^2}$ $(r = 0, \delta = 0)$ conduce a una posteriori impropia, impidiendo la correcta estimación del parámetro.

COMENTARIOS FINALES

La regresión Lasso Bayesiana es fácil de implementar y permite establecer intervalos de credibilidad para todos los parámetros estimados, incluida la variancia de los errores aleatorios. Al conseguir mediante GS una estimación de la distribución a posteriori de los parámetros, puede calcularse cualquier característica de interés bajo dicha distribución, como la esperanza a posteriori o la moda ($\widehat{\boldsymbol{\beta}}_{Lasso}$). Esto le otorga una enorme ventaja sobre el método clásico.

En algunos casos, los valores de las estimaciones producidos por las regresiones Lasso estándar y bayesiana son muy similares. Dependiendo del método de optimización y si se utilizan o no aproximaciones, ambas estimaciones pueden coincidir.

Los mecanismos de elección de λ que se proponen para la regresión Lasso Bayesiana, son aplicables para la regresión Lasso clásica y podrían ayudar a simplificar, otorgando mayor objetividad, la elección del mismo.

Por otra parte, Casella (2008) muestra algunas extensiones del enfoque bayesiano para las regresiones Lasso y plantea la posibilidad de extender las consideraciones a modelos lineales generalizados, mediante algunas modificaciones metodológicas que no deberían requerir mayor esfuerzo computacional que desde el punto de vista clásico.

REFERENCIAS BIBLIOGRÁFICAS

Andrews, D., & Mallows, C. (1974). Scale Mixtures of Normal Distributions. Journal of the Royal Statistical Society, Ser. B(36), 99-102.

Bae, K., & Mallick, B. (2004). Gene Selection Using a Two-Level Hierarchical Bayesian Model. Bioinformatics, 20, 3423-3430.

Casella, G. (2001). Empirical Bayes Gibbs Sampling. Biostatistics, 2, 485-500.

Casella, G., & George, E. (1992). Explaining the Gibbs Sampler. The American Statistician, 46, 167-174.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. The Annals of Statistics, 32, 407-499.

Figueiredo, M. (2003). Adaptative Sparseness for Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25, 1150-1159.

George, E., & McCulloch, R. (1993). Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association, 88(23), 881-889.

Hans, C. (2009). Bayesian Lasso Regression. Biometrika, 96(4), 835-845.

Park, T., & Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical As-

sociation, 103:482, 681-686.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Ser. B(46), 267-288.