Selección, contracción y destilación

Germán González
Santiago Plata

Pronósticos
Abril de 2018

Índice

Problemas a resolver

- Aproximación a un caso real

Selección

- Criterios de información
- Consistencia y eficiencia asintótica
- Validación cruzada
- Stepwise Selection

Contracción

- Contracción Bayesiana
- Lasso: Combinación, selección y contracción

Destilación

- Componentes principales
- Regresión de componentes principales

Problemas a resolver

- Grandes dimensionalidades de datos
- Altos costos computacionales
- Posibles correlaciones entre variables (Multicolinealidad)
- Intuiciones confusas y distorsionadas

Aproximación a un caso real

Objetivo: Predicción de la tendencia del EMBI+ dado un conjunto de variables de atributo

$X = \{X_1, X_2, \ldots, X_n\}$, que son indicadores económicos.

- En este ejemplo, se utilizó una red bayesiana que aprende de un conjunto de datos (D) que consiste en muestras sobre (X, Y). Existen varios enfoques de aprendizaje, en la práctica un **clasificador Bayesiano *ingenuo*** se ajusta bien.
- Las Redes bayesianas tienen la capacidad de combinar opiniones expertas y datos experimentales, a diferencia de las redes neuronales.

Bayesian Network Classifiers for Country Risk Forecasting.

Autores: Ernesto Coutinho Colla, Jaime Shinsuke Ide1, & Fabio Gagliardi Cozman.

Grandes dimensionalidades

- El conjunto de variables explicativas se derivó de estudios empíricos anteriores sobre riesgo cambiario, riesgo país, capacidad de servicio de la deuda de los países y modelos teóricos de préstamos internacionales en presencia de riesgo de incumplimiento.
  - Responsabilidad fiscal del país.
  - Flujos de exportaciones e importaciones.
  - Niveles de deuda, niveles de intercambio y volatilidad.
  - Expectativas de los mercados nacionales e internacionales.
  - Se utilizaron 117 variables cuantitativas oficiales entre 01/01/1999 - 01/03/2006.

¿Cómo abordar este problema haciendo uso de herramientas estadística?

Selección de modelo de subconjuntos

- Examinar todas las posibles combinaciones de K regresores y a partir de unos criterios (AIC , BIC) seleccionar la mejor combinación.
- Los criterios de información se pueden utilizar en una amplia variedad de pronósticos. ¿Cómo seleccionar entre estos?
- ¿Cuáles son las consecuencias de seleccionar únicamente el modelo con el R cuadrado más alto? ¿Hay una mejor manera?

Criterios de selección

- $El\ R^2$ ajusta dentro de la muestra. Seleccionar un modelo bajo este criterio no significa que se produzcan buenos pronósticos por fuera de muestra.
- Se necesitan criterios de información que incorporen la información del pronóstico por fuera de la muestra.
- La mayoría de los criterios de selección se fundamentan en un modelo que incorpora el mínimo **error cuadrático medio (**MSE**) del pronóstico h=1, por fuera de la muestra.**
- Las diferencias entre los criterios de selección se deben principalmente a los grados de libertad que se utilizan para estimar un modelo.

Error cuadrático medio

$$MSE = \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{T} = \frac{\sum_{t=1}^{T} e_t^2}{T}$$

Dada la relación de los criterios de selección con MSE, por construcción entre más pequeño el valor del criterio este es mejor.

MSE se relacionada con otras estadísticas de diagnóstico:

- Suma de cuadrados de los residuos (SSR)
- R^2

R cuadrado

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2} = \frac{MSE}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_t)^2}$$

El modelo con el MSE más pequeño también es el modelo con la menor suma de cuadrados de los residuos (SSR).

Seleccionar el modelo que minimiza la SSR es equivalente a seleccionar el modelo que:

I. Minimiza MSE
II. Maximiza R^2

Criterios de selección

- A partir del MSE dentro muestra se puede estimar el MSE h=1 fuera de la muestra. La selección de modelos de pronóstico sobre esta medida o cualquiera de las formas equivalentes no es una buena práctica.
- El MSE **caerá continuamente** a medida que se agreguen más variables al modelo (los parámetros se eligen para minimizar SSR).
- En algunos casos, la inclusión de nuevas variables podrían generar estimadores no significativos que aumenten ($\uparrow \hat{Y}$) y genere que SSR sea menor $\sum_{t=1}^{T} \downarrow (y_{t-} \uparrow \hat{y}_t)^2$.
- Es poco probable que al incluir una nueva variable se obtenga un estimador igual a cero ($\hat{\beta}_k = 0$), incluso si el coeficiente es cero en la población.

Criterios de selección

- Incluir más variables en un modelo de pronóstico no mejorará necesariamente su desempeño de pronóstico por fuera de la muestra, aunque si mejorará el "ajuste" del modelo en información histórica.
- Incluir más variables en modelo de pronóstico genera:
  I. SSR más pequeños.
  II. R^2 más alto.
- El MSE en la muestra proporciona una evaluación excesivamente optimista (sesgo negativo) de MSE fuera de la muestra. El tamaño del sesgo aumenta con el número de variables incluidas en el modelo.

Error cuadrático medio corregido

El Error cuadrático medio se puede corregir por grados de libertad (GL):

$$S^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K}$$

K= GL utilizados en el ajuste del modelo (k ≠ 0).

S^2 también es la estimación insesgada de la varianza del error de regresión.

Seleccionar el modelo que minimiza S^2 es equivalente a seleccionar el modelo que minimiza los errores estándar de la regresión.

R cuadrado ajustado

$$\bar{R}^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2/(T-K)}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2/(T-K)} = \frac{S^2}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_t)^2/(T-K)}$$

K= GL utilizados en el ajuste del modelo. (k ≠ 0)

I. El R cuadrado también se puede ajustar por GL y reescribir en términos de S^2.

II. El denominador depende únicamente de los datos, no del ajuste del modelo particular, por lo que el modelo que minimiza S^2 y maximiza $\bar{R}^2$.

III. A valores más altos de k el $\bar{R}^2$ será bajo que el del R Cuadrado normal.

Error cuadrático medio corregido

$$S^2 = \left(\frac{T}{T-K}\right) \frac{\sum_{t=1}^{T} e_t^2}{T} = \left(\frac{T}{T-K}\right)\text{MSE}$$

I. La penalización de grado de libertad se resalta cundo se escribe S^2 como un factor de penalización por el MSE.

II. La inclusión de más variables en una regresión no necesariamente disminuirá el S^2 ni elevará el $\bar{R}^2$. Aunque el MSE se reduzca, la penalización de GL aumentará, por lo que el producto podría ir en cualquier dirección.

III. Para obtener una estimación precisa del pronóstico por fuera de muestra (h=1) del MSE, se debe penalizar el MSE en la muestra para reflejar los grados de libertad utilizados.

Criterios de información

Al igual que S^2 muchos de los criterios de selección de modelo de pronóstico utilizan un factor de penalización por MSE. Entre estos se encuentran:

I. Criterio de información de Akaike (AIC)

$$\text{AIC} = e^{\left(\frac{2K}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T} = \boldsymbol{e^{\left(\frac{2K}{T}\right)}} MSE$$

II. Criterio de información Schwarz (SIC)

$$\text{SIC} = T^{\left(\frac{K}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T} = \boldsymbol{T^{\left(\frac{K}{T}\right)}} MSE$$

En donde $\boldsymbol{e^{\left(\frac{2K}{T}\right)}}$ y $\boldsymbol{T^{\left(\frac{K}{T}\right)}}$ son penalizaciones a los MSE. (Estás formulas aplican para el caso gaussiano)

Socrative: Criterios de información

Tanto el AIC y el SIC son funciones de un ratio de parámetros estimados por observación de la muestra (K/T), donde K es el número de parámetros del modelo y T es el número de observaciones en el tiempo.

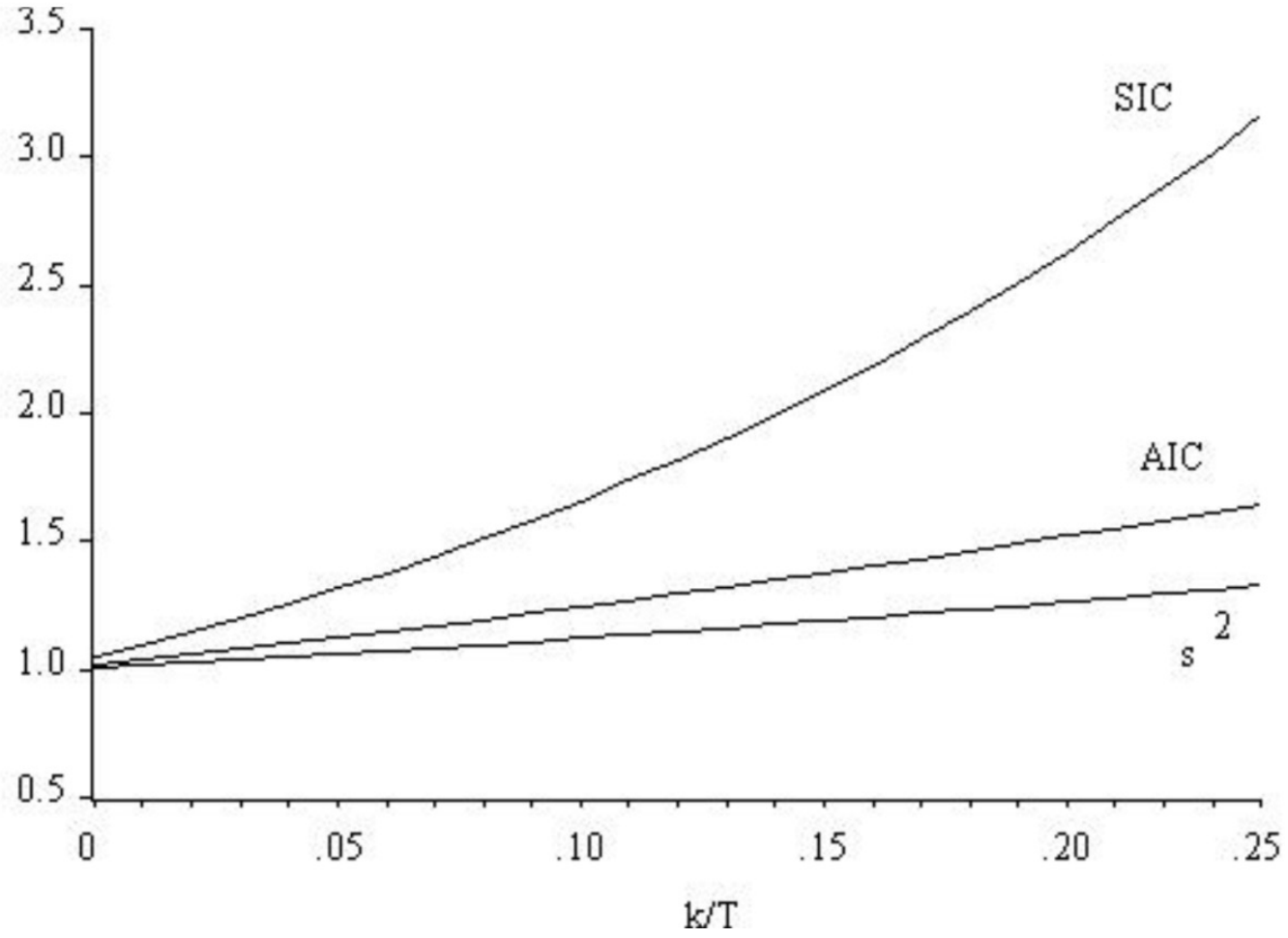
Cuando el ratio (K/T) es más grande, ¿Cuál criterio crece más rápido?

A. S^2
B. AIC
C. SIC
D. Todos crecen igual

¡Correcto!

El SIC crece más rápido

$$\text{SIC} = T^{\left(\frac{K}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T}$$



Criterios de información

Manteniendo constante MSE y valores de T = 100 y K = 25, el Ratio (K/T)es igual 0.25

I. **Penalización baja (S^2):** es pequeña y aumenta lentamente con K/T.
II. **Penalización Media (AIC):** es más grande que la de S^2 e incrementa lentamente con K / T.
III. **Penalización Alta (SIC**): es sustancialmente mayor que la del S^2 y la del AIC y crece a un ritmo mayor con K / T.

SIC > AIC > S^2

Criterios de información

Dado lo que los criterios de información manejan diferentes penalizaciones surgen las siguientes preguntas:

I. ¿Cómo saber cuál de estos criterios tenemos que seleccionar?
II. ¿Qué métricas podríamos utilizar para comparar estos criterios?
III. ¿Qué propiedades podríamos esperar que tenga un criterio de selección de modelo "bueno"?
IV. ¿Los criterios de selección de modelo S^2, AIC y SIC son "buenos"?

La respuesta a estas preguntas no son triviales, pero las propiedades de los criterios de selección sirven como una buena aproximación.

Consistencia

Un criterio de selección modelo es consistente si:

Caso 1: Si el modelo verdadero (Proceso generador de datos -DGP-) **se encuentra** entre los modelos de **conjuntos fijos considerados**, la probabilidad de seleccionar el verdadero proceso generador de datos se a próxima a 1 a medida que el tamaño de la muestra aumenta.

Caso 2: Si el modelo verdadero (DGP) **no se encuentra** entre el **conjunto fijo de modelos a considerar,** es imposible seleccionar el verdadero DGP, pero la probabilidad de seleccionar la mejor aproximación al verdadero DGP se aproxima a uno a medida que el tamaño de muestra aumenta.

Consistencia

- Dado que mayoría de los criterios de selección de modelos -incluidos los presentados aquí- evalúan la bondad de ajuste en términos del pronóstico del MSE fuera de la muestra, la "mejor aproximación" al verdadero DGP, también se evalúa bajo este criterio.
- La consistencia es deseable, en el caso en el que DGP se encuentre entre los modelos considerados, pues a medida que el tamaño de la muestra aumenta, eventualmente se seleccionará el DGP.
- Sin embargo, todos nuestros modelos son falsos, y son simplificaciones intencionales de una realidad mucho más compleja.

Consistencia

- **MSE**: Dado que no penaliza por GL, a medida que crece la muestra este no converge al GDP **(Inconsistente).**
- $\mathbf{S^2}$: Aunque penalice por GL, esto no es suficiente para lograr una convergencia al GDP cuando la muestra es muy grande **(Inconsistente).**
- **AIC:** Aunque el penaliza los GL más fuertemente que S^2 no es suficiente para garantizar la convergencia al GDP, incluso cuando el tamaño de muestra es grande. Dada la penalización del AIC se selecciona modelos sobre-parametrizados **(Inconsistente).**
- **SIC**: Dado que la penalización de los GL es la más grande todas, este criterio logra la convergencia cuando la muestra es muy grande **(Consistente).**

Eficiencia asintótica

Objetivo: Se busca **expandir el conjunto de modelos** a medida que **crece el tamaño de la muestra**, para así obtener progresivamente mejores aproximaciones al DGP.

Un criterio de selección de modelo es asintóticamente eficiente:

Si se elige una **secuencia de modelos**, y a medida que el tamaño de la **muestra aumenta**, el pronóstico de MSE fuera de muestra se acerca al que se obtendría utilizando el DGP a una velocidad al menos tan rápida como la de **cualquier otro criterio de selección.**

Resumen: propiedades

Criterio de selección	Consistencia	Eficiencia asintótica
MSE	No	No
s^2	No	No
AIC	No	Si
SIC	Si	No

El AIC y SIC son muy utilizados, pero no son universalmente aplicables, y aún se está explorando su desempeño en situaciones específicas.

El principio universal de los criterios consiste en inflar de alguna manera las estimaciones de pérdida en la muestra para obtener buenas estimaciones de pérdidas fuera de la muestra.

En la práctica

I. Casi siempre al examinar el AIC y el SIC, se selecciona **el mismo modelo.**

II. Cuando no lo hacen, a pesar de la propiedad teórica de eficiencia asintótica de AIC, la literatura recomienda que en igualdad de condiciones se seleccione el modelo más parsimonioso indicado por el SIC.

III. Muchos autores recomiendan usar el modelo más parsimonioso que selecciona el BIC en igualdad de circunstancias (Diebold, 1999, pág. 75).

Criterios de información caso general:

Los criterios AIC y SIC presentados y las propiedades en términos del pronóstico del MSE fuera de la muestra sólo aplican para el caso gaussiano.

Estos criterios se pueden escribir en términos de la log-verosimilitud:

AIC = $-2\ln L + 2K$

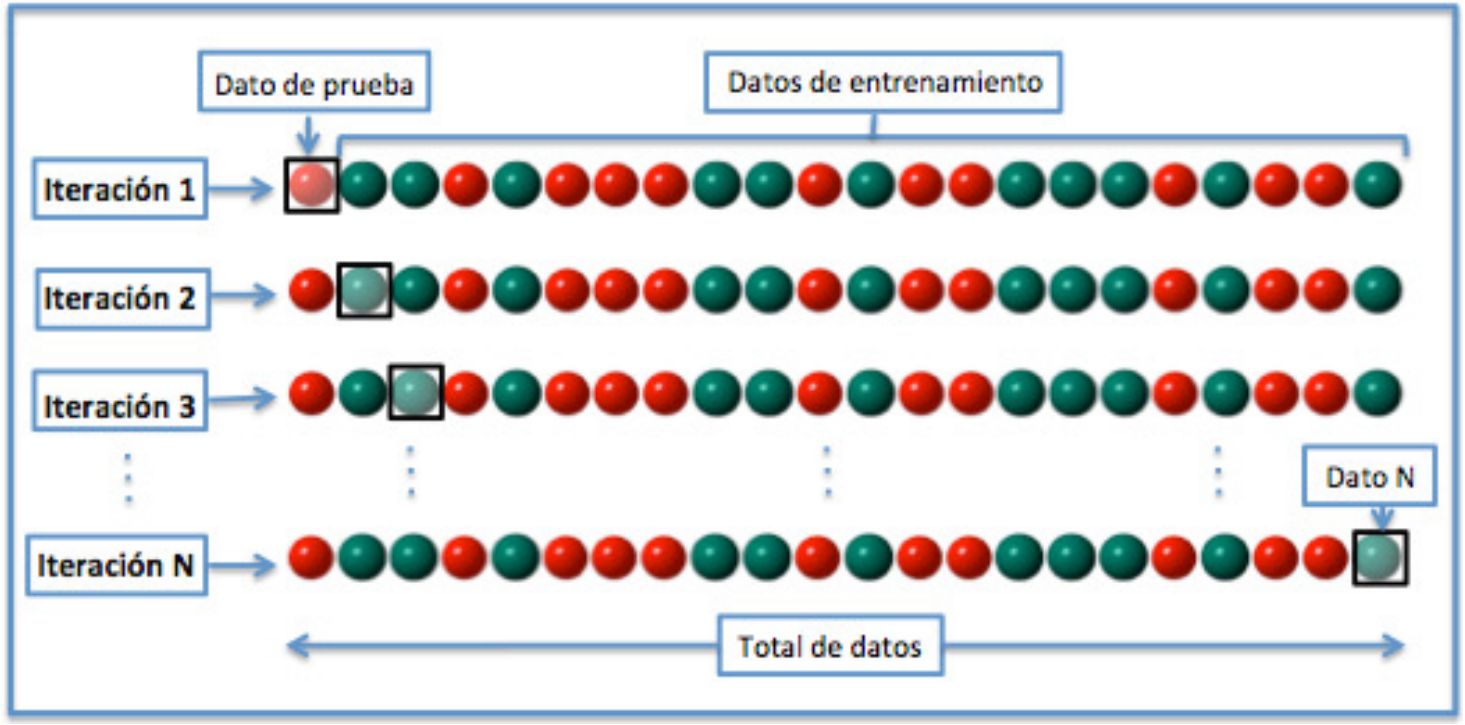
SIC = $-2\ln L + K\ln T$

Siendo aplicables para todos los casos, sin importar si es Gaussiano o no.

Validación Cruzada

Técnica para evaluar los resultados de un modelo estadístico y garantizar que el análisis es independiente de la partición entre datos de entrenamiento y prueba.

1. Validación cruzada dejando un dato por fuera
2. Validación cruzada T−Fold
3. Validación cruzada VS SIC



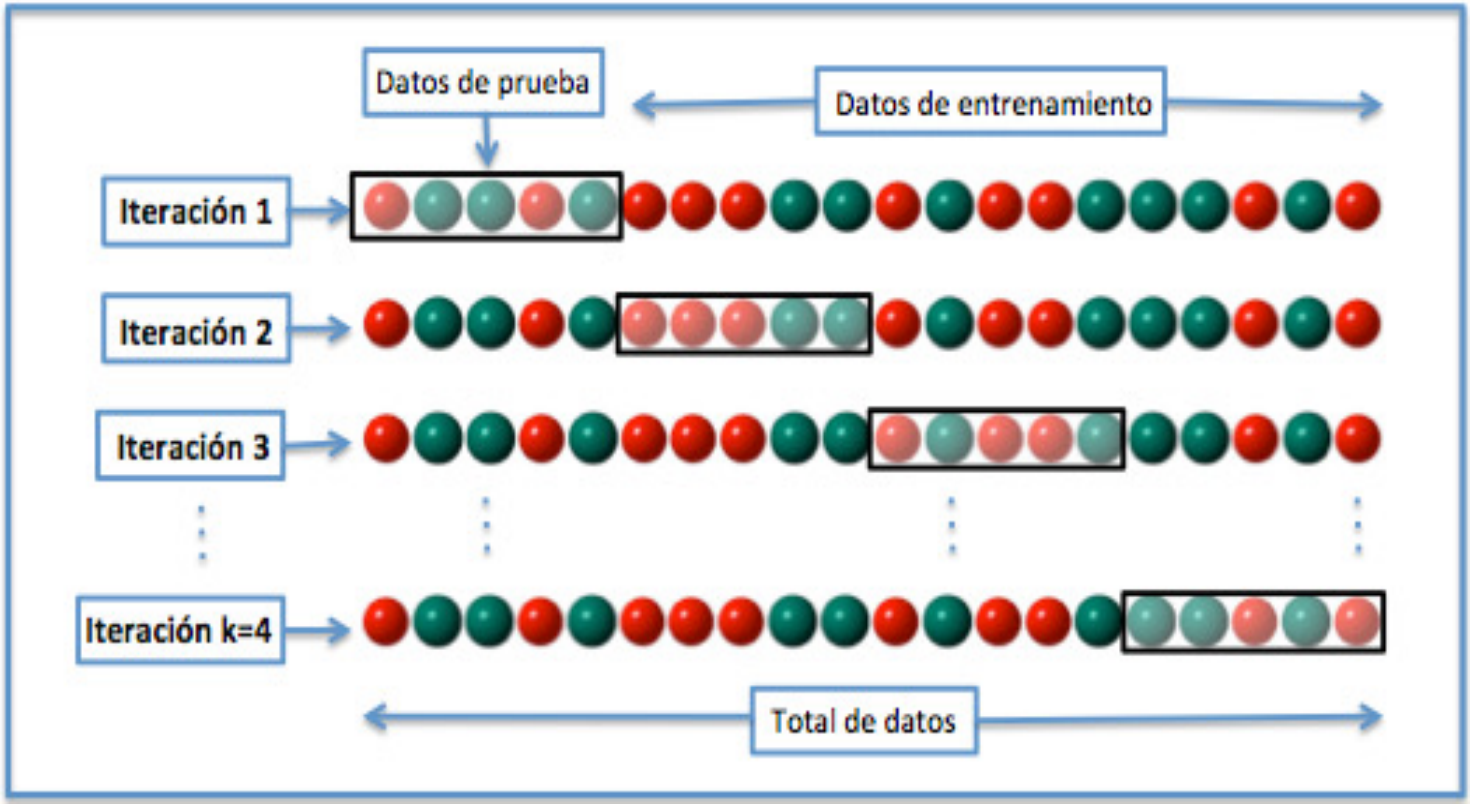
Fuente*: Wikipedia*

Validación cruzada dejando un dato por fuera

i. Estime el primer modelo usando todas las observaciones, a excepción de la primera. Utilice el modelo para predecir la observación eliminada y calcule el error de predicción al cuadrado.

ii. Repita este proceso para cada observación, y promedie los errores al cuadrado al predecir cada una observaciones eliminadas.

iii. Realice este procedimiento con todos los modelos y escoja el que dé menor promedio de errores cuadráticos



Fuente*: Wikipedia*

Validación cruzada T –fold

I. Divida en M muestras aleatorias la muestra original de tamaño T (M<T). Dada la partición M entrene el modelo sin los datos de esta muestra, y extraiga el error de predicción de esa partición.

II. Repita este proceso para cada división M y promedie los errores al cuadrado.

III. Realice este procedimiento con todos los modelos y escoja el que dé menor promedio de errores cuadráticos

(El óptimo de divisiones depende del número de datos. Un M grande con pocos datos sobre estima el error de prueba (varianza alta). Un M bajo subestima el error (sesgo alto))

* El autor recomienda M=10

Validación Cruzada vs SIC

I. Tanto la validación cruzada como el SIC son consistentes
II. SIC penaliza los MSE en la muestra para obtener una estimación insesgada de MSE fuera de la muestra.
III. La Validación Cruzada por el contrario obtiene directamente un MSE insesgado estimado por fuera de la muestra.

Beneficios de la Validación cruzada:

- Se puede usar incluso cuando los grados de libertad del modelo no son claros.
- Se pueden utilizar otras funciones de pérdida diferente a la cuadrática.
- Generalizaciones a otros contextos.

Stepwise Selection

La selección de un modelo en un subconjuntos de modelos, ya sea utilizando AIC, SIC o Validación Cruzada, es una tarea dispendiosa pues requiere 2^k subconjuntos de k regresores.

Stepwise es un algoritmo que a partir de iteraciones acota los posibles subconjuntos de modelos y no explora todas las combinaciones de estos. (Es útil en casos específicos)

I. Forward
II. Backward

Forward

i. Iniciar con una regresión que solo contenga el **intercepto**.
ii. Incluya un primer regresor con el p-valor más pequeño.
iii. Incluya un segundo regresor con el p-valor más pequeño.
iv. Realice este procedimiento hasta que la inclusión de una nueva variable explicativa ya no sea significativa.

Generalmente, se utilizan criterios de información o Validación Cruzada para seleccionar el mejor modelo a partir de la secuencia escalonada de modelos.

Backward

i. Iniciar con una regresión que incluya todas las variables
ii. Pase a un modelo de K-1 variables, eliminando el regresor con el p-valor más grande.
iii. Pase a un modelo de K-2 variables, eliminando el regresor con el p-valor más grande.
iv. Realice este procedimiento hasta que el p-valor de la variable excluida sea estadísticamente significativo.

Generalmente, se utilizan criterios de información o validación cruzada para seleccionar el mejor modelo a partir de la secuencia escalonada de modelos.

Stepwise Selection

Ventajas:

i. Buena aproximación en casos muy específicos.

ii. Reduce tiempos y costos computacionales.

Desventajas:

I. Es un "algoritmo ambicioso" que produce una secuencia decreciente de modelos candidatos.

II. Este algoritmo no garantiza que las propiedades de optimalidad se cumplan en el modelo seleccionado.

Estimadores Bayesianos

Ventajas

i. Simplicidad conceptual.
ii. No requiere pensar en experimentos repetibles.
iii. Uniformidad de aplicación.

Desventajas

- Si a una hipótesis se le asigna una probabilidad inicial de 0 nunca podrá establecerse *a posteriori*.
- Si a una hipótesis se le asigna una probabilidad de 1 nunca podrá rechazarse.

	Bayesiana	Clásica
θ	Aleatorio (=desconocimiento)	Constante, aunque desconocida
$\hat{\theta}$	Fijo (datos)	Aleatorio
Aleatoriedad	(conocimiento parcial) Subjetivo	Viene del muestreo
Distribución importante	A posteriori	D. muestral

La inferencia Bayesiana $\hat{\theta}$ es un valor no-aleatorio que está en función de los datos disponibles.

El parámetro poblacional θ (desconocido) se asume aleatorio y desconocido (incierto). Dado que es desconocido se asigna una distribución de probabilidad sobre θ, que varía ante los datos a priori y después *a posteriori*.

Estimadores Bayesianos

- La contracción es una característica genérica de la estimación bayesiana.
- La regla de Bayes bajo una función de pérdida cuadrática es la media *a posteriori*, que es un promedio ponderado del estimador obtenido por Máxima Verosimilitud y la media *a priori*.

$$\hat{\beta}_{Bayes} = \omega_1 \hat{\beta}_{MLE} + \omega_2 \beta_0$$

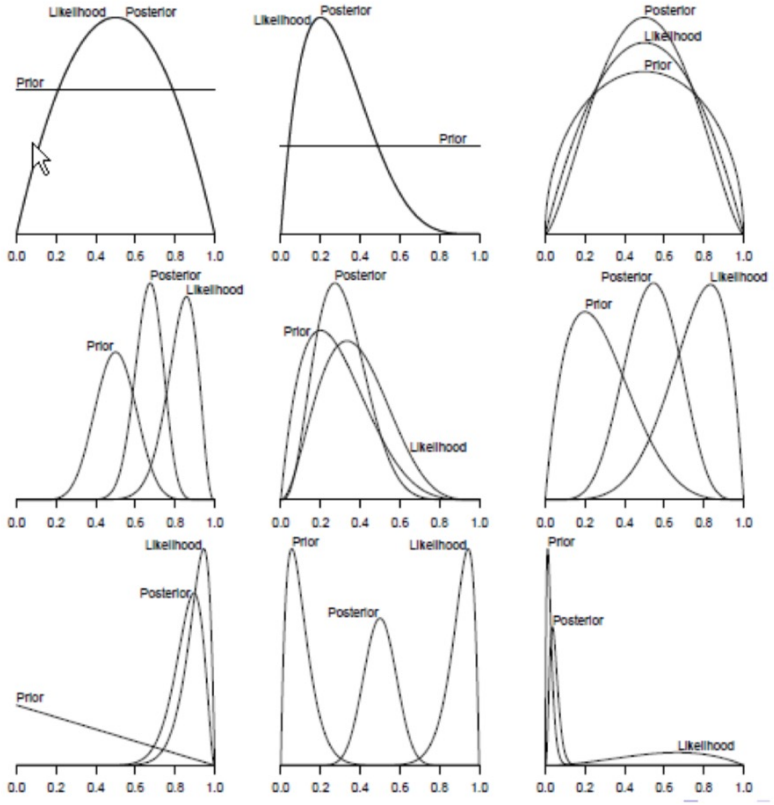
- La regla de Bayes "contrae" el estimador de máxima verosimilitud basado en la información conocida. Los ponderadores dependen de la precisión de la media *a priori*, si esta es más precisa, se le asigna un mayor peso.

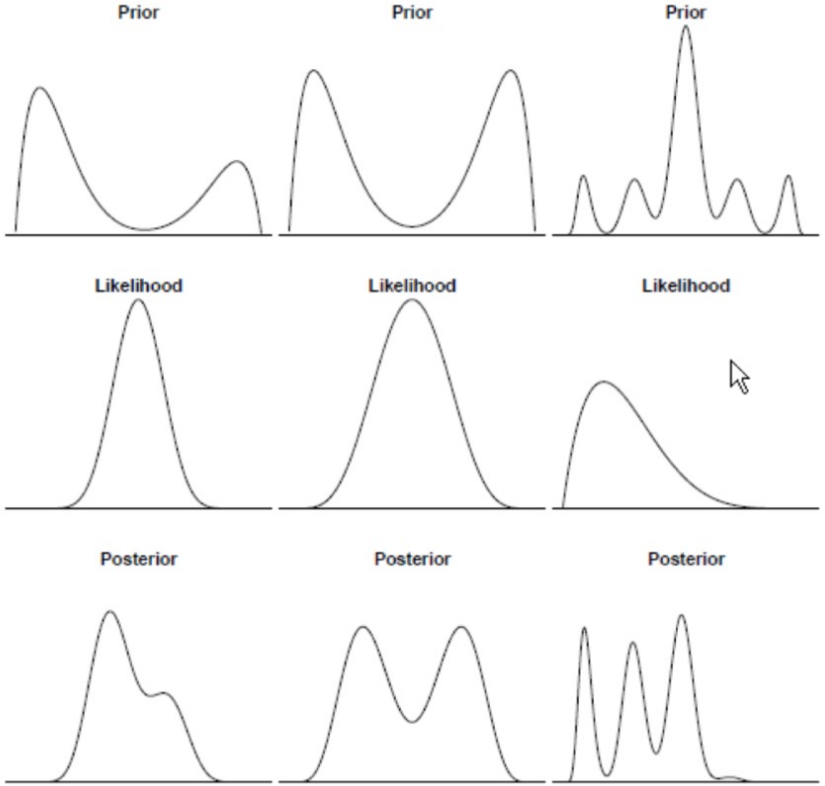
Estimadores Bayesianos

- MCO permite obtener estimadores M.E.L.I
- Sin embargo, la correlación entre variables (Multicolinealidad) incrementa la varianza de los estimadores y afecta su significancia individual
- Ejemplo 2 en dos variables:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_{1i}^2(1 - r_{1,2})}$$

- El objetivo de los Estimadores Bayesianos es reducir esta varianza.





Distribución

Estimadores Bayesianos

El estimador de contracción bayesiana clásico es la regresión de Ridge. Un método alternativo que busca estimadores que pueden ser sesgados, pero con menor error cuadrático medio.

$$\hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1} X'Y$$

I. Si $\lambda = 0$ el estimador Ridge es idéntico al MCO

II. Si $\lambda = \infty$ el estimador de Ridge se contrae a 0.

λ no puede ser elegido por criterios de información, debido a que los regresores (K) se incluyen independientemente de λ. Se puede utilizar Validación Cruzada para seleccionar diferentes parámetros de ajuste.

Estimación One-Shot

$$\hat{\beta}_{PEN} = \ argmin_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2$$

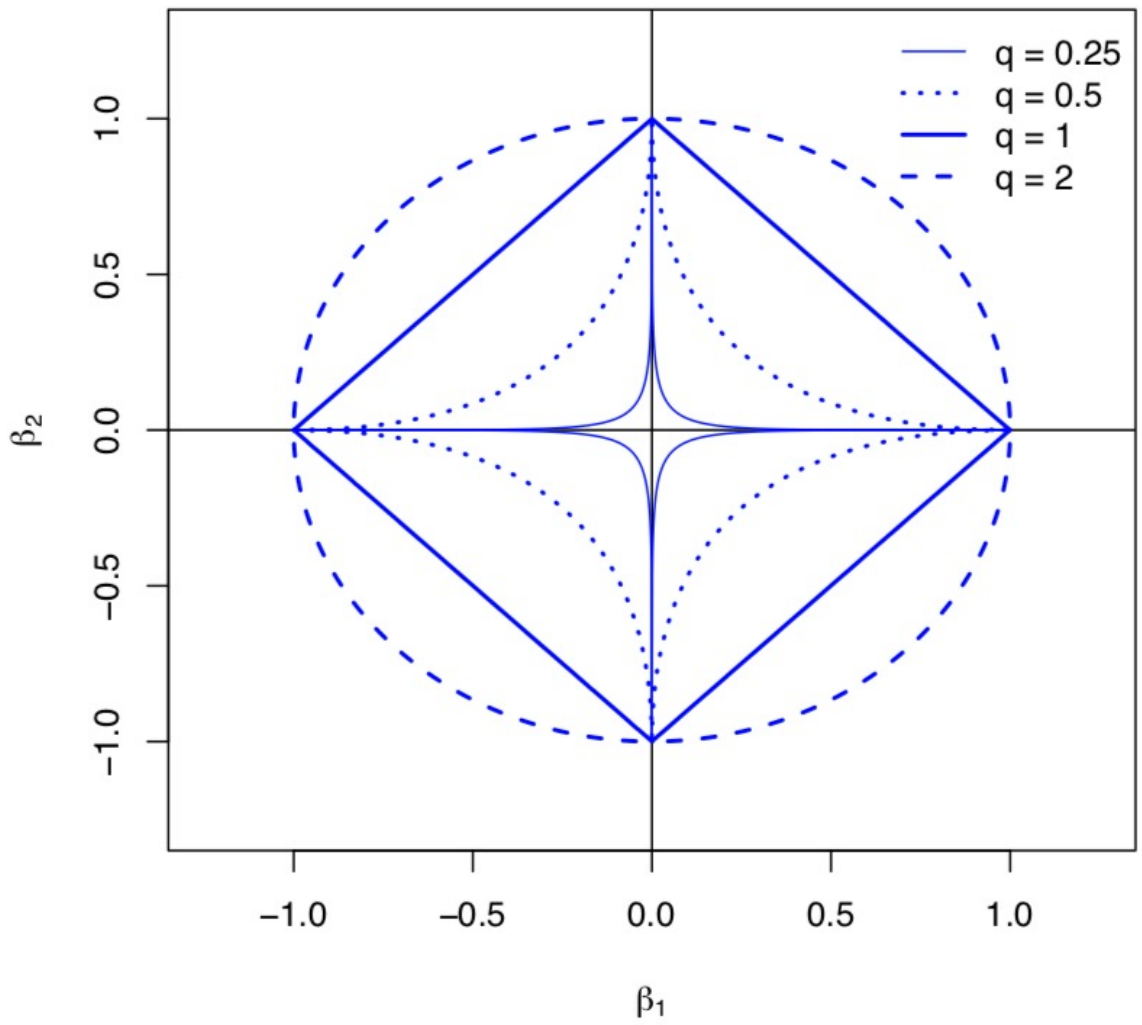
Sujeto a $\sum_{i=1}^{K} |\beta_i|^q \ \leq c$

Funciones de penalización cóncavas no diferenciables desde el origen producen selección. Las sanciones suaves y convexas producen contracción.

- Cuando q = 0 métodos de penalización por el **número de variables**.
- Cuando q = 1 se produce una selección de **Lasso**
- Cuando q = 2 se trata de una regresión de **Ridge**.

Curvas de nivel de la penalización

- Cuando q = 0 métodos de penalización por el **número de variables**.
- Cuando q = 1 -> **Lasso**
- Cuando q = 2 -> **Ridge**.



Lasso

$$\hat{\beta}_{Lasso} = argmin_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2$$

Sujeto a $\sum_{i=1}^{K} |\beta_i| \leq c$

Lasso al igual Ridge es una técnica de regresión lineal regularizada, con una diferencia en la penalización (norma) que trae consecuencias en la selección.

i. Ridge contrae
ii. Lasso contrae y selecciona.
iii. Lasso es más eficiente que Pen. (**q = 1 → mínima convexidad con una única estimación**)

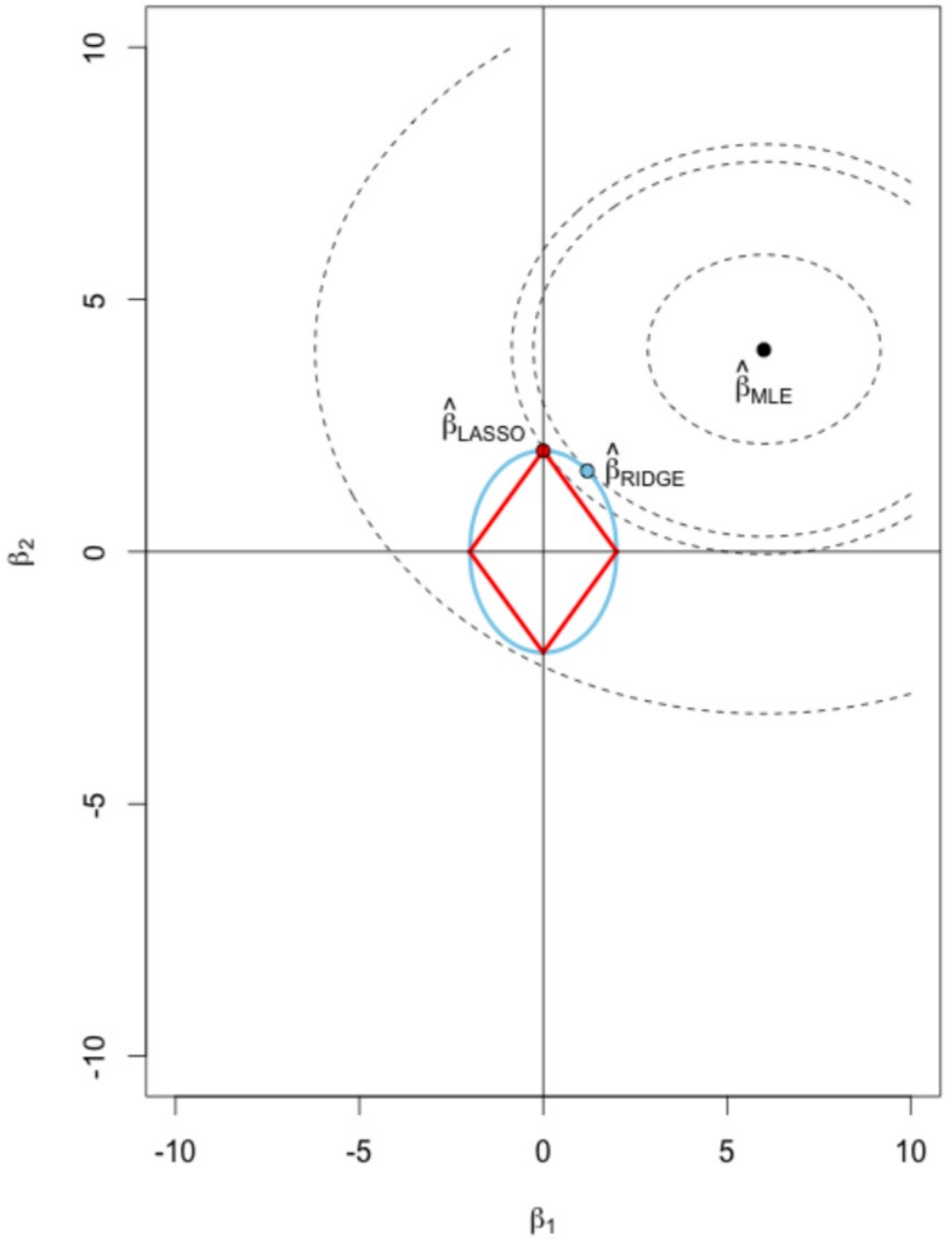
Resultados Lasso

- El valor óptimo de parámetros es el mismo número de variables seleccionadas (β diferentes de cero).

- Dado lo anterior, se pueden utilizar criterios de información para seleccionar diferentes "modelos de lasso" para diferentes valores de λ.

- Lasso produce una secuencia **"creciente"** de modelos candidatos (cuando λ aumenta). El "mejor" valor de λ puede ser elegido por medio de criterios de información o Validación cruzada.

Comparación de Lasso y Ridge



Mezcla: Red elástica

$$\hat{\beta}_{EN} = \ argmin_{\beta}\left(\sum_{t=1}^{T}\left(y_t - \sum_{i}\beta_i x_{it}\right)^2 + \lambda\sum_{i=1}^{K}\left(\alpha|\beta_i| + (1-\alpha)\,\beta_i^2\right)\right)$$

- Se trata de una mezcla de Lasso y Ridge que combina las penalidades de L1 y L2.
- Esta aproximación produce una mejor predicción que Lasso, en el caso de predictores fuertemente correlacionados.
- Mueve los predictores fuertemente correlacionados dentro o fuera de ambos modelos con la intención de producir una mejor precisión en la predicción, en relación con Lasso.
- A diferencia de Lasso, hay dos parámetros de ajuste: λ y α
- α = 1 → Lasso, cuando α = 0, → Ridge.

Lasso adaptativo

$$\hat{\beta}_{ALASSO} = argmin_{\beta}\left(\sum_{t=1}^{T}\left(y_t - \sum_{i}\beta_i x_{it}\right)^2 + \lambda\sum_{i=1}^{K} w_i|\beta_i|\right)$$

Donde $w_i = 1/\hat{\beta}_i^{v}$, $\hat{\beta}_i$ es la estimación MCO y $v > 0$

- Propiedad de consistencia
- Los pesos se calcula por medio de MCO
- Cada uno de los parámetros de la función de penalización se pondera de manera diferente, en contraste con el Lasso normal.

Red elástica adaptativa

$$\hat{\beta}_{ALN} = \ argmin_{\beta}\left(\sum_{t=1}^{T}\left(y_t - \sum_{i}\beta_i x_{it}\right)^2 + \lambda\sum_{i=1}^{K} w_i(\alpha|\beta_i| + (1-\alpha)\,\beta_i^2\right)$$

Donde $w_i = 1/\hat{\beta}_i^{v}$, $\hat{\beta}_i$ es la estimación MCO y $v > 0$

- Propiedad de consistencia.
- Combinación de Red Elástica y Lasso adaptativo

Resumen Contracciones

Contracción	Problema Optimización
Lasso	$\hat{\beta}_{Lasso} = argmin_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2$
Red elástica	$\hat{\beta}_{EN} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} (\alpha \lvert\beta_i\rvert + (1-\alpha)\, \beta_i^2 \right)$
Lasso adaptativo	$\hat{\beta}_{ALASSO} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i \lvert\beta_i\rvert \right)$
Red elástica adaptativa	$\hat{\beta}_{ALN} = argmin_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{K} w_i (\alpha \lvert\beta_i\rvert + (1-\alpha)\, \beta_i^2 \right)$

Destilación: Componentes principales

Método estadístico que permite simplificar la complejidad de espacios muéstrales con amplias dimensiones sin pérdida de Información de los mismos.

Sea X una matriz de Tamaño KxT

Descomposición propia: $X = V\ D^2 V'$

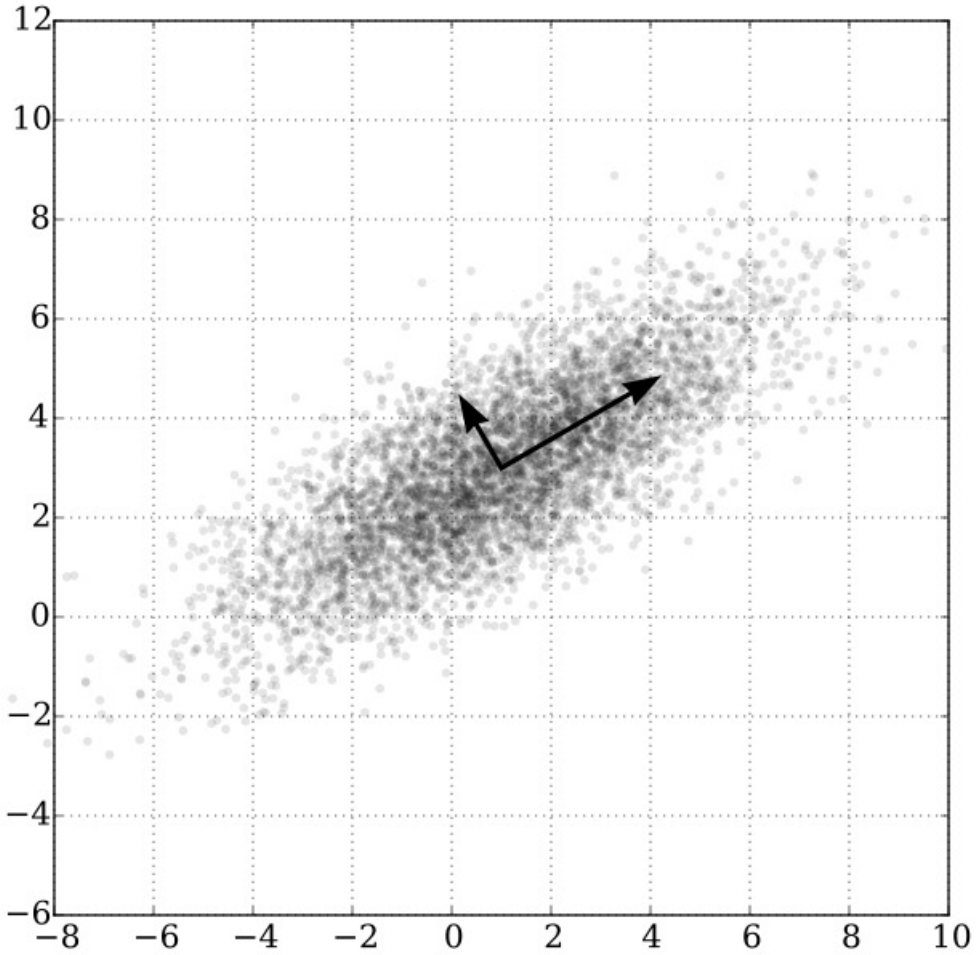
I. La j-ésima columna de V, v_j, es el j-ésimo auto-vector de X'X (Cargas)
II. La matriz diagonal D^2 contiene los valores descendentes de los auto-valores de X'X (Valores)

El primer componente principal se puede escribir como:

$$z_1 = X v_1$$

Donde la varianza se puede escribir como Var(z_1) = d_1^2/T

Componentes principales



Componentes principales

- Los componentes principales son, por construcción, ortogonales entre si mismos.

- Cada componente explica una proporción de la varianza. Para reducir dimensionalidad se puede seleccionar un número de componentes (N) que expliquen un umbral de varianza (generalmente el 95%).

- Con menos componentes se logra un modelo más parsimonioso y a cambio de una pequeña perdida de información.

Componentes principales

- Es posible realizar econometría (Regresiones, ARIMA, GARCH, VAR) sobre los auto-valores y después devolverse a la escala original (Ejemplo: Curva TES).
- La regresión Ridge y la PRC (*Principal Regression Components)* son procedimientos de contracción que involucran Componentes principales.
- Ridge incluye de manera efectiva todas los componentes y los reduce de acuerdo con el tamaño de los auto-valores asociados con los componentes.
- La regresión por componentes principales reduce efectivamente algunos componentes a cero (no incluidos) y no reduce otros (incluidas).

Referencias consultadas

Castro, S. (10 de junio de 2013). Estimación y selección de variables en grandes dimensiones. [Diapositivas]. Recuperado de http://www.iesta.edu.uy/wpcontent/uploads/2014/05/CursoPosgrado_Aprendizaje_Automatico_SCastro_2013.pdf

Diebold, F. (2017). *Forecasting in Economics, Bussines, Finance and Beyond*. Pensilvania, Estados Unidos: University of Pennsylvania.

Lejarza, I. (s.f.). Introducción a la inferencia bayesiana. [Diapositivas]. Recuperado de https://www.uv.es/mlejarza/actuariales/iibayes.pdf

Riascos, A. (Marzo de 2018). Selección, Riesgo Esperado y Validación de Modelos. [Diapositivas] Recuperado de https://www.uv.es/mlejarza/actuariales/iibayes.pdf

http://www.dm.uba.ar/materias/estadistica_teorica_Mae/2006/2/practicas/bayes.PDF

https://ocw.ehu.eus/pluginfile.php/3145/mod_resource/content/1/estadistica/tema-10-regresion-sesgada.pdf