TRABAJO FIN DE MASTER EN MATEMÁTICAS AVANZADAS

Estimación de Matrices de Covarianzas: Nuevas Perspectivas

Autor: Fernando Godino Gómez

Director del TFM: Dr. Hilario Navarro Veguillas

DPTO. ESTADÍSTICA, INVESTIGACIÓN OPERATIVA Y CÁLCULO NUMÉRICO
FACULTAD DE CIENCIAS-UNED

Curso 2013-2014

ÍNDICE

1. INTRODUCCIÓN

La estimación de las matrices de covarianzas poblacional Σ y de su inversa Σ^{-1}, llamada matriz de precisión o de concentración, es de suma importancia en múltiples aplicaciones estadísticas, y por tanto, de gran importancia en muchos campos del conocimiento, como en economía (mercados financieros), biología (genética), ingeniería (curvas de datos), computación (datos de imágenes) y en general en todos aquellos campos donde es necesario manejar grandes cantidades de datos.

La $p \times p$ matriz de covarianzas Σ de un vector aleatorio $X = (x_1,...,x_p)$ con al menos $p(p+1)/2$ parámetros juega un papel central dentro de la estadística multivariante. Algunas de las aplicaciones estadísticas donde es necesaria una buena estimación de las matrices de covarianzas son:

- El estadístico T^2 de Hotelling requiere una estimación de la matriz de precisión (Σ^{-1}) (Johnson and Wichern, 2007, Capítulo 5).
- Análisis Factorial (Johnson and Wichern, 2007, Capítulo 8; Anderson, 2003, Capítulo 14).
- Componentes principales (Johnson and Wichern, 2007, Capítulo 9; Anderson, 2003, Capítulo 11).
- Discriminación y Clasificación (Anderson, 2003, Capítulo 6).
- Análisis de Series-Temporales (Box, Jenkins and Reinsel, 1994; Shumway and Stoffer, 2010).
- Modelos gráficos Gaussianos (Wong, Carter and Kohn, 2003; Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007).

Como los valores reales de los elementos de la matriz de covarianzas poblacional Σ son desconocidos, se hace imprescindible una buena estimación de la matriz de covarianzas. El estimador más habitual que se suele tomar cuando la población de partida sigue una distribución normal $N(\mu,\Sigma)$ o sobre todo, cuando el número de observaciones n supera ampliamente a la dimensión poblacional p, es la matriz de covarianza muestral S.

Veremos con mayor profundidad en el punto 2, que los *estimadores de máxima verosimilitud* para μ y Σ son $\overline{X}$ y S, los cuales se definen de la siguiente forma:

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de una población normal de media μ y covarianza Σ, entonces los *estimadores de máxima verosimilitud* para μ y Σ *serán*:

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t \text{ y } S = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \overline{X})(X_t - \overline{X})'$$

Usar la covarianza muestral S cuando el número de observaciones n supera al número de variables p, es decir $n > p$, tiene una serie de ventajas:

- Es un estimador muy fácil de construir.
- Insesgado, es decir, el valor esperado es igual a la verdadera matriz de covarianzas ($E(S) = \sum$).
- Está basada en el estimador de máxima verosimilitud de Σ.
- S^{-1} se puede usar para estimar Σ^{-1}.

En el trabajo en análisis multivariante se tiene tradicionalmente asumido que n/p, es decir, el número de observaciones por variable, es grande. Sin embargo, actualmente es muy común que el número de variables sea grande y el cociente n/p sea pequeño o incluso que sea menor de uno. Por ejemplo:

- *Estudios sobre el clima*. Aquí n es el número de observaciones en un instante, y p el número de estaciones de observación. El análisis de componentes principales es ampliamente usado bajo el nombre en inglés de "*empirical orthogonal functions*" Preisendorfer (1988), donde se tiene un cociente n/p moderado.
- *Mercados financieros.* La estimación de la matriz de covarianzas de alta dimensión es una parte fundamental de las carteras de valores (portfolio selection), minimización de riesgos (risk managament) y del estudio del precio de los activos (asset pricing). Cada día son publicados gran cantidad de indicadores financieros, que sería el tamaño poblacional. Por ejemplo, se puede ver en *riskmetrics.com*.

- *Buscadores de Internet.* Aunque es un caso alejado de ser Gaussiano, la estrategia de los buscadores es otro ejemplo más, de matrices donde se utilizan enormes cantidades de datos (n y p suelen ser miles cada uno).

- *Estudios en biología:* Como indican Dobbin y Simon (2007), y Yao et al. (2008), estos estudios a menudo emplean un pequeño número de observaciones, debido a que se disponen de pocos datos experimentales, a que su recolección es costosa o a que tenemos una serie de restricciones temporales para el estudio.

- *Estudios genéticos.* El agrupamiento de genes usando datos de un método de microarray, (por ejemplo, Eisen et al., 1998). Está basado en las medidas de las distancias relativas a la correlación muestral. Entonces, si p genes son analizados, con p del orden de entre 1.000 y 10.000, la matriz de covarianzas de tamaño $p \times p$ tiene que ser calculada.

- *Estudios sobre el cáncer.* El estudio y control del cáncer, implica un gran número de marcadores y genes a revisar. Por razones obvias, el número de muestras suele ser limitado en comparación con los anteriores, siendo n/p un número menor de uno. Por ejemplo, en el trabajo realizado por Beerenwinkel et al. (2007), se estudió la correlación parcial entre 78 genes cancerígenos (para un total de $78 \times 77 \div 2 = 3.003$ correlaciones parciales) con únicamente 35 muestras de tumores disponibles.

- *Análisis funcional de datos.* Cada dato es una curva y normalmente de alta dimensión. En el ejemplo de la Figura 1, extraído de, Hastie, Buja y Tibshirani (1995), un conjunto de datos de un pequeño discurso consistente en 162 ejemplos de un fonema "dcl" hablado por 50 varones. Cada ejemplo está calculado como un periodogram sobre 256 puntos. Aquí n=162 and p=256.

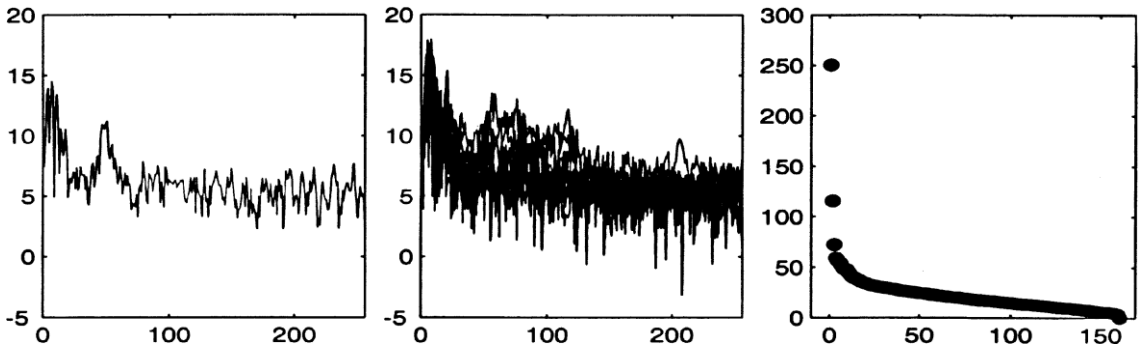


Figura 1. (a) Un ejemplo aislado de un *periodogram del conjunto de datos del fonema;* (b) Diez ejemplos*, que indican la variabilidad;* (c) S*creeplot de autovalores del ejemplo del fonema.*

- Y en otros múltiples casos, como espectrografía, tratamiento de los datos de imágenes en computación, ecología, etc., es muy común tener situaciones donde $p >> n$.

El hecho de que n/p sea pequeño o que $p > n$, trae consigo una serie de consecuencias indeseables para el estimador S:

- Cuando $p > n$ la matriz de covarianzas muestral S no es de rango máximo por el aumento de los autovalores que se hacen cero, lo que hace que S sea singular y no pueda ser definida positiva, lo que conlleva que no sea invertible, lo que supone un gran problema ya que en muchas aplicaciones la necesidad de calcular la matriz de precisión Σ^{-1} es incluso mayor que la de calcular Σ.

- Aunque la matriz de covarianzas muestral S es un estimador insesgado y definido positivo Johnstone (2001), es un pobre estimador cuando el cociente p/n es grande. Los autovalores de S divergen de los autovalores de Σ. El mayor de los autovalores de S tendrá un alto sesgo hacia arriba, y el menor de los autovalores de S tendrá un alto sesgo hacia abajo.

- Cuando el cociente n/p es pequeño, ya que el número de observaciones es algo mayor a la dimensión poblacional o los dos valores son de alta dimensión, podremos calcular su matriz de covarianzas muestral S, aunque esta tendrá una significativa cantidad de error muestral, y sobre todo, su inversa S^{-1} será un pobre estimador de Σ^{-1}. Por ejemplo, como aparece en Bai and Shi (2011), bajo el supuesto de normalidad, el valor esperado de la inversa será $E\left(S^{-1}\right) = \frac{n}{n-p-2}\sum^{-1}$. De tal forma que aunque S es insesgada para Σ, S^{-1} es altamente sesgada para Σ^{-1} si el valor de p está próximo al de n. En particular, si $p = n/2 + 2$ tendremos que $E\left(S^{-1}\right) = 2\sum^{-1}$.

Teniendo en cuenta estos problemas con el uso como estimador de Σ de la matriz de covarianzas muestral S cuando n/p es pequeño o $p > n$, es decir, en casos de alta dimensionalidad, se lleva estudiando desde hace tiempo y más ahora con la existencia de ordenadores cada vez más potentes, estimadores de Σ y de Σ^{-1} que no tengan los defectos de S.

Las soluciones que se han buscado en este entorno de alta dimensionalidad, están basadas principalmente en la regularización o sparsity y en los métodos que se engloban en el término “shrinkage”. Estos métodos lo que buscan es disminuir la dimensión de la matriz a estimar y en transformar nuestra matriz de covarianzas muestral en una estimación que sea insesgada para Σ y además que tenga una estimación de error pequeña.

El esquema de este trabajo será como sigue. El punto 2 está dedicado a los métodos clásicos de estimación, su fundamentación y a estudiar el porqué no son adecuados cuando nos movemos en alta dimensionalidad. En el punto 3 estudiaremos algunas de las nuevas perspectivas de estimación. Por último en el punto 4 daremos las conclusiones del trabajo y dedicaremos un anexo al entorno computacional que se ha usado en el trabajo, explicando algunas de las funciones que se han usado del paquete estadístico R, que un software gratuito de gran uso en estadística.

2. ESTIMACIÓN CLÁSICA DE LA MATRIZ DE COVARIANZAS

En este punto vamos a recordar cómo obtenemos la matriz de covarianzas muestral S y algunas características sobre ella; por qué es un buen estimador de Σ cuando $n > p$; por qué deja de estar bien condicionada cuando $p > n$. Por último, aunque en el punto 3 desarrollaremos distintos métodos para encontrar estimadores que resuelvan los problemas que encontramos con S, vamos a dedicar un último apartado a encontrar otro estimador para Σ sin tanto "ruido" en el apartado 2.3.

2.1 LA MATRIZ DE COVARIANZAS MUESTRAL

Estimador de máxima verosimilitud de Σ

Ya hemos comentado en la introducción que la matriz de covarianzas muestral S es el estimador natural de la matriz de covarianzas. Recordaremos que esto es así porque es el estimador de máxima verosimilitud de Σ.

Como guía del desarrollo que vamos a seguir, o para completar la explicación se puede consultar, por ejemplo, Johnson and Wichern, 2007, Capítulo 4, o en Anderson, 2003.

Todo el desarrollo que vamos a realizar a continuación, se basa en la suposición de que nos movemos en una población normal multivariante de vector de medias μ y matriz de covarianzas Σ, o al menos, en que el valor de las observaciones n es mucho mayor que la dimensión poblacional p. Esto, generalmente, no es cierto para poblaciones que no sean normales. Por eso es conveniente usar alguna de las técnicas multivariantes para chequear la presunción de normalidad multivariante.

Supongamos que los $p \times 1$ vectores $X_1, X_2, ..., X_n$ representan una muestra aleatoria de una población normal multivariante de vector de medias μ y matriz de covarianzas Σ. Entonces al ser $X_1, X_2, ..., X_n$ independientes entre sí y cada una con una distribución

$N_p(\mu,\Sigma)$, la función de densidad conjunta de todas las observaciones es el producto de las densidades marginales normales:

$$L(\mu,\Sigma)=\prod_{j=1}^{n}\left\{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-(x_j-\mu)'\Sigma^{-1}(x_j-\mu)/2}\right\}=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}e^{-\sum_{j=1}^{n}(x_j-\mu)'\Sigma^{-1}(x_j-\mu)/2}$$

La expresión que resulta, ahora considerada en función de μ y Σ para el conjunto de observaciones fijas $x_1, x_2,..., x_n$ es llamada la *función de verosimilitud.* La técnica que mejor explica los datos observados es la *estimación de máxima verosimilitud*, que consiste en seleccionar los valores de μ y Σ que maximicen la densidad conjunta evaluada en las observaciones. Los estimadores que encontraremos de esta forma se llaman *estimadores de máxima verosimilitud.*

Aunque vamos a omitir la mayoría de los pasos, que se pueden consultar en cualquier libro de estadística multivariante, por ejemplo en los señalados al comienzo de este apartado, vamos a comentar los más importantes.

Utilizando las propiedades de las matrices simétricas y la definición de traza de una matriz (suma de los elementos de su diagonal), podemos escribir la *función de verosimilitud* de la siguiente forma:

$$L(\mu,\Sigma)=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\times\exp\left\{-tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'+n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right]/2\right\}$$

El exponente de la *función de verosimilitud*, obviando $-1/2$, es

$$tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'+n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right]=tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'\right)\right]+n(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)$$

Se sabe que Σ^{-1} es definida positiva, entones $(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)>0$ a menos que $\mu=\bar{x}$. Entonces la función se maximiza con respecto a μ en $\hat{\mu}=\bar{x}$. Sólo resta maximizar la *función de verosimilitud* sobre Σ, obteniendo, que el máximo se produce cuando

$$\hat{\Sigma}=\frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'=\frac{n-1}{n}S\,.$$

Recordar que los estimadores de máxima verosimilitud son consistentes, es decir, cuanto mayor sea el tamaño de la muestra, la estimación será más precisa. Dando la definición formal diremos que $\hat{\Sigma}$ (o S) es consistente al estimar Σ porque para cualquier $\xi > 0$ se cumple $\underset{n\to\infty}{Lim}\, P\left(\left|\hat{\Sigma} - \Sigma\right| \le \xi\right) = 1$.

Estimador insesgado

Por tanto, los estimadores de máxima verosimilitud para μ y Σ serán $\hat{\mu} = \bar{x}$ y $\hat{\Sigma} = \frac{n-1}{n} S$. Sin embargo, se prefiere tomar como estimador de Σ a S, ya que S es un estimador insesgado, es decir $E(S) = \sum$ y $\hat{\Sigma}$ es un estimador sesgado, ya que $E\left(\hat{\Sigma}\right) = \frac{n-1}{n}\sum$ (Podemos observar que para un número grande de observaciones n el sesgo es insignificante).

Demostración.

Vamos a demostrar de dónde salen estos resultados para $E\left(\hat{\Sigma}\right)$ y $E(S)$. Nos vamos a guiar por Johnson and Wichern, 2007, Capítulo 3, resultado 3.1.

Empecemos remarcando que $\left(X_{ji} - \bar{X}_i\right)\left(X_{jk} - \bar{X}_k\right)$ es el elemento $(i,k) - ésimo$ de $\left(X_j - \bar{X}\right)\left(X_j - \bar{X}\right)'$.

Podemos escribir

$$\sum_{j=1}^{n}\left(X_j - \bar{X}\right)\left(X_j - \bar{X}\right)' = \sum_{j=1}^{n}\left(X_j - \bar{X}\right)X_j + \left(\sum_{j=1}^{n}\left(X_j - \bar{X}\right)\left(-\bar{X}\right)\right) = \sum_{j=1}^{n} X_j X_j' - n\bar{X}\bar{X}',$$

dado que $\sum_{j=1}^{n}\left(X_j - \bar{X}\right) = 0$ y $n\bar{X}' = \sum_{j=1}^{n} X_j'$. Por lo tanto, su valor esperado es

$$E\left(\sum_{j=1}^{n} X_j X_j' - n\bar{X}\bar{X}'\right) = \sum_{j=1}^{n} E\left(X_j X'\right) - nE\left(\bar{X}\bar{X}'\right)$$

Para cualquier vector aleatorio V con $E(V)=\mu_V$ y $Cov(V)=\Sigma_V$, tenemos $E(VV')=\Sigma_V+\mu_V\mu_V'$. Entonces, $E(X_jX')=\Sigma+\mu\mu'$ y $E(\overline{XX}')=\frac{1}{n}\Sigma+\mu\mu'$. Usando estos resultados, tendremos lo siguiente

$$\sum_{j=1}^{n}E(X_jX')-nE(\overline{XX}')=n\Sigma+n\mu\mu'-n\left(\frac{1}{n}\Sigma+\mu\mu'\right)=(n-1)\Sigma.$$

Y por tanto, dado que $\hat{\Sigma}=\frac{1}{n}\left(\sum_{j=1}^{n}X_jX_j'-n\overline{XX}'\right)$, se sigue inmediatamente que $E(\hat{\Sigma})=\frac{n-1}{n}\Sigma$ y que, como $S=\frac{n}{n-1}\hat{\Sigma}$, entonces $E(S)=\Sigma$.

Estimador suficiente

De la expresión

$L(\mu,\Sigma)=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\times\exp\left\{-tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'+n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right]/2\right\}$, podemos ver que la densidad conjunta depende del conjunto de observaciones $x_1,x_2,...,x_n$, sólo a través de la media muestral $\overline{X}$ y de $\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'=(n-1)S$. Esto implica que $\overline{X}$ y $(n-1)S$ o simplemente S, son estadísticos suficientes.

La importancia de los estadísticos suficientes para una población normal es que toda la información sobre μ y Σ en la matriz de datos X está contenida en $\overline{X}$ y S, independientemente del tamaño de la muestra n.

Estimador bien condicionado

Recordemos que una matriz A está mal condicionada si tiene un número condicional $k(A)$ muy grande o incluso singular. El número condicional de una matriz A se define como $k(A)=\|A\|\cdot\|A^{-1}\|$, donde $\|\cdot\|$ es la norma de Frobenius (*esta norma se define en la página 48 de este trabajo*) que cumple para el caso de ser A simétrica que

$\|A\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} A_{ij}^2 = \sum_{i=1}^{p} \lambda_i^2$, donde λ_i son los autovalores de A. En este caso de simetría se cumple que $k(A) = \frac{\lambda_{máx}}{\lambda_{mín}}$. Por tanto, para estudiar si nuestra matriz de covarianzas muestral S está bien condicionada, o en general la matriz que tomemos como estimación de Σ, calcularemos el cociente entre su mayor autovalor y el menor.

Bai and Yin (1993) demuestran que cuando $n \geq p$ el autovalor más pequeño de una matriz de covarianzas de la forma $\frac{1}{n} XX'$ tiende casi seguro al límite $\left(1-\sqrt{y}\right)^2$ cuando $n \to \infty$ y $p/n \to y \in (0,\infty)$, donde X es una matriz $p \times n$ de media 0 y varianza 1 y donde $E|X_{11}|^4 < \infty$. De igual forma se obtiene se tiene que el autovalor más grande tiende casi seguro al límite $\left(1+\sqrt{y}\right)^2$, resultado obtenido por Bai, Yin y Krishnaiah (1988).

De esta forma, podemos decir que cuando $n \geq p$ tenemos que $Lim.\lambda_{\min} = \left(1-\sqrt{y}\right)^2$ y $Lim.\lambda_{\max} = \left(1+\sqrt{y}\right)^2$, y por tanto entender que cuando el número de observaciones es mayor que la dimensión poblacional, la diferencia entre el mayor y el menor autovalor es mínima, y por tanto, el cociente $k(A) = \frac{\lambda_{máx}}{\lambda_{mín}}$ es un número pequeño y podremos decir que el estimador S está bien condicionado.

Ahora podemos entender fácilmente por qué cuando $p >> n$ (y entonces, el autovalor más pequeño es cero) $k(A)$ es singular, y por tanto, el estimador S no está bien condicionado en ese caso.

Distribución de S

El hecho de que $X_1, X_2, ..., X_n$ representan una muestra aleatoria de una población normal multivariante de vector de medias μ y matriz de covarianzas Σ determina completamente las distribuciones muestrales de $\bar{X}$ y S.
No vamos a presentar el desarrollo completo, que se puede ver en cualquiera de las referencias ya dadas. Simplemente nos limitaremos a dar los resultados que se obtienen.

La distribución de la matriz de covarianzas muestral S es llamada *distribución de Wishart.* Esta está definida como la suma de los productos independientes de vectores aleatorios normales multivariante, concretamente:

$W_m(\cdot|\Sigma) =$ Distribución de Wishart con m grados de libertad= distribución de $\sum_{j=1}^{m} Z_j Z_j'$,

donde la Z_j están independientemente distribuidas como $N_p(0,\Sigma)$.

Podemos resumir los principales resultados para las distribuciones muestrales como siguen:

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño n de una normal con p variables con media μ y matriz de covarianzas Σ. Entonces

- $\overline{X}$ está distribuida como una $N_p\left(\mu, \frac{1}{n}\Sigma\right)$
- $(n-1)S$ está distribuida como una Wishart con n-1 grados de libertad
- $\overline{X}$ y S son independientes.

Como Σ es desconocida, la distribución de $\overline{X}$ no puede ser usada directamente para hacer inferencias sobre μ, sin embargo, S provee de información independiente sobre Σ, y la distribución de S no depende de μ. Esto nos permite construir un estadístico para hacer inferencias sobre μ.

2.2 EJEMPLOS DEL COMPORTAMIENTO DE S PARA DISTINTOS VALORES DEL COCIENTE *n/p*

Ejemplo 1

Vamos a comprobar con un ejemplo muy sencillo, el comportamiento de los autovalores de la matriz de covarianzas muestral S y de su número condicional $k(A)$, para distintos valores del cociente *n/p*.

- ***n/p=2***

```
# En este caso el número de variables es 3 y el tamaño muestral es 6.
> X1<-matrix(c(4,4.2,3.9,4.3,4.1,3.8,2,2.1,2,2.1,2.2,2,0.6,0.59,0.58,0.62,0.63,0.56),nrow=6,ncol=3)
> X1
     [,1] [,2]  [,3]
[1,]  4.0  2.0  0.60
[2,]  4.2  2.1  0.59
[3,]  3.9  2.0  0.58
[4,]  4.3  2.1  0.62
[5,]  4.1  2.2  0.63
[6,]  3.8  2.0  0.56
# Calculamos la matriz de covarianzas muestral S1 de nuestra matriz de datos X1
> S1<-var(X1)
> S1
      [,1]       [,2]         [,3]
[1,] 0.0350 0.010000000 0.0036000000
[2,] 0.0100 0.006666667 0.0016666667
[3,] 0.0036 0.001666667 0.0006666667
# Calculamos los autovalores  y autovectores de la matriz de covarianzas muestral S1
> eigen(S1)
$values
[1] 0.0385824445 0.0035673123 0.0001835765
$vectors
       [,1]       [,2]        [,3]
[1,] 0.9476033  0.3151545 -0.05220711
[2,] 0.3022998 -0.9375025 -0.17234810
[3,] 0.1032606 -0.1475354  0.98365164

# Calculamos el número condicional
```

$$k(S1)=\frac{\lambda_{máx}}{\lambda_{mín}}=210.1709342$$

Observamos que el número condicional calculado es grande. Esto es debido a que el número de observaciones, aun siendo mayor, es cercano al número de variables. Sin

embargo, podemos ver que ningún autovalor se anula y que por tanto sería posible calcular S^{-1}.

- ***n/p=1***

```
# En este caso el número de variables es 3 y el tamaño muestral es 3.
> X2<-matrix(c(4,4.2,3.9,2,2.1,2,0.6,0.59,0.58),nrow=3,ncol=3)
> X2
     [,1] [,2]  [,3]
[1,]  4.0  2.0 0.60
[2,]  4.2  2.1 0.59
[3,]  3.9  2.0 0.58

# Calculamos la matriz de covarianzas muestral S2 de nuestra matriz de datos X2
> S2<-var(X2)
> S2
          [,1]       [,2]  [,3]
[1,] 0.023333333 0.008333333 5e-04
[2,] 0.008333333 0.003333333 0e+00
[3,] 0.000500000 0.000000000 1e-04
# Calculamos los autovalores y autovectores de S2
> eigen(S2)
$values
[1]  2.635883e-02  4.078329e-04 -4.117958e-19
$vectors
          [,1]       [,2]       [,3]
[1,] 0.9401607  0.2917020  0.1760902
[2,] 0.3402607 -0.8309177 -0.4402255
[3,] 0.0179018  0.4737992 -0.8804509

# Calculamos el número condicional
```

$$k(S2)=\frac{\lambda_{máx}}{\lambda_{mín}}=6.400959215\cdot 10^{16}$$

El número condicional se hace gigantesco debido a que los autovalores más pequeños son prácticamente cero y además la matriz S2 no es definida positiva.

- ***n/p=2/3***

```
# En este caso el número de variables es 2 y el tamaño muestral es 3.
> X3<-matrix(c(4,4.2,2,2.1,0.6,0.59),nrow=2,ncol=3)
> X3
     [,1] [,2] [,3]
[1,]  4.0  2.0 0.60
[2,]  4.2  2.1 0.59
# Calculamos la matriz de covarianzas muestral S3 de nuestra matriz de datos X3
> S3<-var(X3)
> S3
       [,1]   [,2]   [,3]
[1,]  0.020  1e-02 -1e-03
[2,]  0.010  5e-03 -5e-04
[3,] -0.001 -5e-04  5e-05
# Calculamos los autovalores y autovectores de S3
> eigen(S3)
$values
[1] 2.505000e-02 2.081668e-17 3.497320e-21
$vectors
           [,1]       [,2]       [,3]
[1,]  0.89353410  0.44899533 0.00000000
[2,]  0.44676705 -0.88909966 0.09950372
[3,] -0.04467671  0.08890997 0.99503719
# Calculamos el número condicional
```

$$k(S3)=\frac{\lambda_{máx}}{\lambda_{mín}}=7.162627383\cdot 10^{18}$$

El número condicional se hace gigantesco debido a que los autovalores son aún más pequeños que en el caso anterior, y son prácticamente cero.

Por tanto, vemos cómo disminuyendo el valor del cociente n/p el autovalor más pequeño se va haciendo cero y su número condicional crece de forma muy notable, con lo que S se convierte en un estimador mal condicionado. Para $n/p=1$ ya nos encontramos con un autovalor que prácticamente se anula y para $n/p=2/3$ es todavía más notable que

el autovalor más pequeño puede considerarse como cero, con lo que S no sería definida positiva y por tanto no invertible.

Ejemplo 2

Vamos a ver otro ejemplo con una base de datos real sobre el estudio de ciertas propiedades del vino. Estos datos están tomados de la página Web http://archive.ics.uci.edu. En este ejemplo el número de variables será mayor que en el ejemplo anterior (14 características del vino), lo cual hará que sea más significativa la aproximación de los autovalores hacia cero. Vamos a omitir escribir la matriz de datos original, las matrices de covarianzas muestrales y los autovectores en aquellos casos en que su tamaño sea muy grande, y nos limitaremos a escribir los comandos y los autovalores de las diferentes matrices de covarianzas muestrales.

```
# Cargamos los datos tamaño muestral de 175 y número de variables 14
> wine <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data",
+ sep=",")
```

- **Primer caso**: tamaño muestral de 175 y número de variables 14

```
# Calculamos la matriz de covarianzas muestral S1 de la matriz de datos wine
> S1<-var(wine)

# Calculamos los autovalores y autovectores de S1
> eigen(S1)
$values
 [1] 9.920203e+04 1.725366e+02 9.531196e+00 5.100810e+00 1.285788e+00
8.681666e-01 2.870070e-01 1.552790e-01 1.137333e-01 8.637843e-02 4.620262e-02
[12] 3.492730e-02 2.076267e-02 8.092659e-03

$vectors
…….

# Calculamos el número condicional
```

$$k(S1) = \frac{\lambda_{máx}}{\lambda_{mín}} = 12258273.83$$

- **Segundo caso**: tamaño muestral de 14 y número de variables 14

```
> X2<-wine[c(1,2,3,4,5,6,7,8,9,10,11,12,13,14), ]
```

```
> X2

   V1    V2   V3   V4   V5  V6   V7   V8   V9  V10  V11  V12  V13  V14
1   1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29 5.64 1.04 3.92 1065
2   1 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.05 3.40 1050
3   1 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.03 3.17 1185
4   1 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.86 3.45 1480
5   1 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82 4.32 1.04 2.93  735
6   1 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.05 2.85 1450
7   1 14.39 1.87 2.45 14.6  96 2.50 2.52 0.30 1.98 5.25 1.02 3.58 1290
8   1 14.06 2.15 2.61 17.6 121 2.60 2.51 0.31 1.25 5.05 1.06 3.58 1295
9   1 14.83 1.64 2.17 14.0  97 2.80 2.98 0.29 1.98 5.20 1.08 2.85 1045
10  1 13.86 1.35 2.27 16.0  98 2.98 3.15 0.22 1.85 7.22 1.01 3.55 1045
11  1 14.10 2.16 2.30 18.0 105 2.95 3.32 0.22 2.38 5.75 1.25 3.17 1510
12  1 14.12 1.48 2.32 16.8  95 2.20 2.43 0.26 1.57 5.00 1.17 2.82 1280
13  1 13.75 1.73 2.41 16.0  89 2.60 2.76 0.29 1.81 5.60 1.15 2.90 1320
14  1 14.75 1.73 2.39 11.4  91 3.10 3.69 0.43 2.81 5.40 1.25 2.73 1150

# Calculamos la matriz de covarianzas muestral S2 de la matriz de datos X2
> S2<-var(X)

# Calculamos los autovalores y autovectores de S2
> eigen(S2)
$values
 [1]  4.476594e+04  1.391049e+02  5.617707e+00  9.200965e-01  2.923981e-01
1.999927e-01 1.007961e-01 4.296059e-02 1.789065e-02 5.083378e-03 1.922637e-03
[12] 2.234128e-04 2.241956e-06 0.000000e+00

$vectors
……..

# Calculamos el número condicional
```

El número $k(S2)$ es singular, ya que hay autovalores que se anulan.

- **Tercer caso**: tamaño muestral de 7 y número de variables 14

```
> X3<-wine[c(1,2,3,4,5,6,7), ]
> X3

  V1    V2   V3   V4   V5  V6   V7   V8   V9  V10  V11  V12  V13  V14
1  1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29 5.64 1.04 3.92 1065
2  1 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.05 3.40 1050
3  1 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.03 3.17 1185
4  1 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.86 3.45 1480
5  1 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82 4.32 1.04 2.93  735
6  1 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.05 2.85 1450
7  1 14.39 1.87 2.45 14.6  96 2.50 2.52 0.30 1.98 5.25 1.02 3.58 1290


# Calculamos la matriz de covarianzas muestral S3 de la matriz de datos X3
> S3<-var(X3)

# Calculamos los autovalores y autovectores de S3
> eigen(S3)
$values
```

[1] 6.720576e+04 1.171289e+02 7.920291e+00 2.602831e-01 1.575257e-01 1.042540e-01 2.863582e-17 0.000000e+00 -2.844232e-18 -1.437018e-17
[11] -8.144487e-17 -6.730798e-15 -2.685001e-14 -2.581863e-12

$vectors
……..

Calculamos el número condicional

El número $k(S3)$ es singular, ya que hay autovalores que se anulan.

En el primer caso donde el tamaño muestral es muy superior al número de variables, vemos cómo aunque el número condicional (cociente entre el autovalor más grande y el más pequeño) es alto, la matriz de covarianzas muestral es definida positiva, por tanto invertible, y su autovalor más pequeño está lejos de ser cero. Sin embargo, tanto para $n = p$ (14 variables con 14 muestras) como para $n < p$ (14 variables con 7 muestras) ya tenemos autovalores que se anulan y por tanto, el número condicional es singular y la matriz de covarianzas muestral será sesgada y no invertible.

2.3 APROXIMACIÓN A UNA MATRIZ DE COVARIANZAS SIN RUIDO USANDO LA TEORÍA DE LAS MATRICES ALEATORIAS

La matriz de covarianzas muestral S como estimador de Σ, contiene una notable cantidad de ruido cuando el número de observaciones n es mucho más pequeño que la dimensión poblacional p. La teoría sobre matrices aleatorias nos ofrece un camino para quitar el ruido de S .La covarianza muestral sin ruido es usada como estimador de Σ.

Empezaremos presentando algunas propiedades de los autovalores de una matriz de correlación muestral bajo la premisa de variables aleatorias independientes e idénticamente distribuidas (iid).

Sea X una matriz muestral de tamaño $p \times n$ con elementos iid (por ejemplo, iid variables aleatorias normales), y sea S su correspondiente covarianza muestral. Sea $D = diag(S)$ y $C = D^{-1/2} S D^{-1/2}$, donde C es la matriz de correlación muestral. Supongamos que $n/p \to q$. Bajo estas premisas, los autovalores de C tienen la siguiente función de densidad cuando $n, p \to \infty$:

$$P(\lambda)=\frac{q}{2\pi\lambda}\sqrt{(\lambda_{máx}-\lambda)(\lambda-\lambda_{mín})}\text{, con } \lambda_{mín}<\lambda<\lambda_{máx},$$

donde $\lambda_{máx}=\left(1+\sqrt{q^{-1}}\right)^2$ y $\lambda_{mín}=\left(1-\sqrt{q^{-1}}\right)^2$ cuando $q>1$. Para $q<1$, la densidad tiene una masa de $1-q$ en cero. Todos los autovalores están limitados por $\lambda_{máx}$ (ver Laloux et al. 1999, Plerou et al. 1999, Bai Z. 1999). Por ejemplo, cuando $q=1$ tenemos que $\lambda_{máx}=4$, y en este caso bajo las presunciones de matrices aleatorias, se espera que la mayoría de los autovalores sean menores de 4, y los que sean mayores no deberían estar lejos de 4, suponiendo muestras finitas.
La existencia de autovalores que exceden $\lambda_{máx}$ indica la presencia de ruido.

Consideremos la descomposición espectral

$$C=\sum_{i=1}^{p}\tau_i\xi_i\xi_i'$$

donde $\tau_1\geq\tau_2\geq...\tau_p$ son los autovalores de C y $\{\xi_k\}$ sus correspondientes autovectores. Supongamos que tenemos k autovalores mayores que $\lambda_{máx}$, Laloux et al. 2001 definió una matriz de correlación limpia de ruido de la forma

$$\overline{C}=\sum_{i=1}^{p}\lambda_i\xi_i\xi_i'+aI_p$$

donde I_p es la matriz identidad de $p\times p$, y a es una constante tal que la traza de $\overline{C}$ es igual a la de C. Esto implica que

$$a=\frac{\lambda_{k+1}+...+\lambda_p}{p}$$

Del hecho de que $\xi_i'\xi_i=1$ para todo i, es inmediato que $tr\left(\overline{C}\right)=\sum_{i=1}^{p}\lambda_i=tr(C)=p$. La última igualdad viene del hecho de que los elementos de la diagonal de C son 1.

Una vez obtenida la matriz de correlación limpia de ruido, la matriz de covarianzas muestral limpia de ruido nos queda

$$\bar{S} = D^{1/2}\bar{C}D^{1/2}$$

Se debe apuntar que mientras $\bar{C}$ es definida positiva, no es una matriz de correlación ya que los elementos de la diagonal no son necesariamente la unidad. Sin embargo, $\bar{S}$ es una matriz de covarianzas debido a que está positivamente definida.

Por último, dada la importancia de calcular la matriz inversa del estimador, ya que como hemos visto en la introducción, la necesidad de usar la matriz de precisión Σ^{-1} en muchas aplicaciones es incluso mayor que la de la propia matriz de covarianzas Σ, vamos a ver cómo quedaría.

Dado el estimador $\bar{S} = D^{1/2}\bar{C}D^{1/2}$ siendo D una matriz diagonal y con $\bar{C}$ teniendo una estructura de factores, tenemos que $\bar{S}^{-1} = D^{-1/2}\bar{C}^{-1}D^{-1/2}$ donde $\bar{C}^{-1}$ se calcula fácilmente en vista de su estructura de factores.

3. NUEVAS PERSPECTIVAS DE ESTIMACIÓN

3.1 MÉTODOS DE REGULARIZACIÓN. ALGORITMO GRAPHICAL LASSO

Como acabamos de ver, la alta dimensionalidad supone un gran problema a la hora de estimar la matriz de covarianzas Σ y la matriz de precisión Σ^{-1}. Por tanto, para poder trabajar con esta dimensión, se hace necesario el desarrollo de técnicas de selección de variables y de reducción de dimensiones, ya que modelos de menor dimensión serán más fáciles de interpretar y los errores al estimar serán menores. Métodos tradicionales de selección de variables en regresión, como los métodos secuenciales (forward selection y bacward elimination), suelen ser bastante inestables y su aplicación no es posible cuando el número de variables p es similar o ampliamente superior al número de observaciones n, es decir $p >> n$. Como consecuencia, han surgido en estos últimos años metodologías que intentan resolver el problema de la alta dimensión.

Una forma de plantearse la resolución de la alta dimensionalidad consiste en añadir un término de penalización o regularización a la función objetivo que mide el ajuste de los datos a un determinado modelo.

El problema de selección de variables en regresión surge cuando se quiere modelar la relación entre una (o más) variable(s) de interés y un conjunto de potenciales variables explicativas $X_1,...,X_p$, pero existe incertidumbre acerca de qué subconjunto de las variables X_j utilizar. Esto se vuelve bastante complejo cuando el tamaño poblacional p es grande y parte de las variables $X_1,...,X_p$, son redundantes o irrelevantes (George, 2000). Incluyendo cada vez más variables en un modelo de regresión aumentará la cantidad de parámetros a estimar, el ajuste a los datos mejorará pero también aumentará su varianza y como consecuencia, la de la función de regresión estimada.

Recordemos que cuando estamos trabajando con p variables, tendremos un número de 2^p modelos que permitan compararlas, lo cual, incluso para un número moderado de variables p, hace que la evaluación de esos modelos sea muy costosa o imposible dado su

enorme tamaño. Sheather, (2009) y Hastie et al., (2009) demuestran que la selección del mejor subconjunto sólo puede ser implementada con p cercano a 40 ($2^{40} > 10^{12}$).

Esto nos demuestra por qué la reducción del espacio de modelos es necesaria, ya que en múltiples problemas actuales el número de variables con las que nos manejamos puede rondar fácilmente el millar, por ejemplo en genética (Eisen et al., 1998).

Para reducir la dimensión, uno de los enfoques más sencillos es a través de los modelos de regresión lineal: $Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon$, con $X' = (X_1,...,X_p)$. La estimación habitual de los parámetros de regresión β, se realiza por el método de mínimos cuadrados. Sin embargo, este método no es eficaz para alta dimensionalidad. Para corregirlo, se aplican técnicas de regresión lineal, que en nuestro caso serán regresión Ridge y regresión Lasso. Estas dos técnicas aplican la estimación de un modelo de regresión lineal. Sin embargo, se distinguirán por la penalización que aplican. La regresión Ridge aplicará una penalización de norma-L_2 y la regresión Lasso una penalización de norma-L_1.

La diferencia entre el uso de una norma u otra en la penalización tiene gran importancia. Mientras que el uso de la norma-L_2 produce un estimador lineal en y del vector de parámetros β, la penalización L_1 no produce un estimador lineal en y, y no se obtiene una fórmula cerrada para su expresión, sino que debe encontrarse la solución a través de un algoritmo de optimización. Otra característica muy importante que las diferencia, es que la penalización L_2 utiliza todas las variables predictoras en el modelo de regresión final, debido a que valores mayores del parámetro de complejidad λ contraen los coeficientes hacia cero pero sin alcanzar dicho valor en general; sin embargo, la penalización L_1 produce algunos coeficientes de regresión exactamente nulos.

Es esta característica de la regulación Lasso, es decir, que la penalización L_1 produzca algunos coeficientes de regresión exactamente nulos, lo que ha hecho que a lo largo de estos últimos años numerosos autores hayan propuesto su uso conjunto con modelos gráficos para la estimación de la matriz de covarianzas Σ y más concretamente de la matriz de precisión Σ^{-1}. El método se nombra como Graphical Lasso, y estimará la inversa de la matriz de covarianzas.

Un modelo gráfico consiste en un conjunto de vértices acompañados de un conjunto de aristas que unen algunos pares de vértices. En los modelos gráficos, cada vértice representará una variable y el gráfico nos dará una idea visual de la relación entre conjuntos

de variables, ya que la ausencia de arista entre dos vértices nos indicará la independencia condicional entre esas dos variables. Un gráfico *sparse* debe tener un número relativamente pequeño de aristas. Este tipo de gráficos no tendrán flechas de dirección. Estos gráficos indirectos también reciben el nombre de *Campos Aleatorios de Markov* o *Redes de Markov.* Los principales retos que se plantean al trabajar con este tipo de gráficos son elegir una estructura para el gráfico, estimar los parámetros para las aristas a partir de los datos y el cálculo de las probabilidades marginales de los vértices desde la distribución conjunta. En este trabajo nos centraremos únicamente en el uso de este tipo de gráficos conjuntamente con la regulación Lasso para la estimación de Σ y Σ^{-1}.

3.1.1. Técnicas de regularización en modelos lineales

Estimación por mínimos cuadrados

Comenzaremos recordando los modelos de regresión lineal y la estimación de los parámetros de regresión β por mínimos cuadrados.
Los modelos de regresión lineal tienen la siguiente forma:

$$Y = f(X) = \sum_{j=1}^{p} \beta_j X_j + \varepsilon \text{, con } X' = (X_1, ..., X_p)$$

El método de estimación más usado es el de mínimos cuadrados, que minimiza la suma de los residuos al cuadrado o en inglés, *residual sum of squares (RSS).*

$$RSS(\beta) = \sum_{j=1}^{n} \varepsilon_i^2 = \sum_{j=1}^{n} (y_i - f(x_i))^2 = \sum_{j=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = (y - X\beta)'(y - X\beta)$$

También se puede escribir como

$$RSS(\beta) = \|y - X\beta\|^2$$

Para encontrar el valor de β que minimiza esta función, derivamos con respecto a β e igualamos a cero

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2X'(y - X\beta) = 0$$

Asumiendo que X es de rango completo y $X'X$ es definido positivo, tenemos como solución única

$$X'(y - X\beta) = 0 \Rightarrow \hat{\beta} = (X'X)^{-1} X'y$$

El estimador de máxima verosimilitud de β es el obtenido por los mínimos cuadrados. Sin embargo, existen un par de razones por la que este estimador no es el más adecuado en algunas ocasiones (Tibshirani, 1996). La primera es la precisión de las predicciones, ya que aunque presenta bajo sesgo, tiene una gran varianza. Esta precisión en la predicciones puede ser mejorada en ciertas ocasiones llevando algunos coeficientes a cero, con lo cual, aunque aumentamos el sesgo se reduce la varianza. La segunda razón es la interpretabilidad o más bien, la falta de interpretabilidad, ya que sería conveniente que a partir de un gran número de predictores pudiéramos determinar un subconjunto más pequeño que fuera lo más representativo posible, conservando el mayor poder de predicción.

Además, si las columnas no son linealmente independientes o $p >> n$, X no será de rango máximo y por tanto, $X'X$ será singular y los coeficientes $\hat{\beta}$ no tendrán solución única.
Estas deficiencias de rango que ocurren cuando $p >> n$ se podrán reducir por regulación.

Entonces, los métodos de regularización que aplicaremos reducirán la varianza de los estimadores, posiblemente con menor error predictivo. Podemos decir, por tanto, que los métodos de regularización buscarán transformar problemas mal condicionados, ya que los coeficientes $\hat{\beta}$ no tienen solución única, en problemas bien condicionados.

Podemos describir las técnicas de regulación como la reducción de los coeficientes de la regresión mediante alguna restricción. De tal forma, en general los coeficientes obtenidos por regulación se obtendrán minimizando, adecuándolo a un parámetro de regularización λ, la siguiente suma:

$$\hat{\beta} = \arg\min_{\beta_0, \beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \sum_{j=1}^{p} \phi_\lambda \left(\left| \beta_j \right| \right) \right\} \qquad (1)$$

donde $\beta = (\beta_1,...,\beta_p)$, $\lambda \geq 0$ y $\phi_\lambda\left(\left|\beta_j\right|\right)$ es una función creciente de penalización sobre el tamaño de β, que depende a su vez de λ. Este parámetro controlará el peso que demos a la penalización. Si crece también crecerá la penalización en los coeficientes de regresión, y por tanto, estos coeficientes se contraerán más hacia cero. La búsqueda del valor adecuado de λ habrá que particularizarlo a cada problema, sabiendo que $0 < \lambda < 1$.

Si la penalización se realiza mediante la norma-L_q, la solución al problema de regresión lineal (regularizado) se expresa como:

$$\hat{\beta} = \arg\min_{\beta_0,\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|^q\right\} \quad (2)$$

Los métodos de regresión Ridge y Lasso son casos particulares de (2), con valores de $q = 2$ (norma-L_2) y $q = 1$ (norma-L_1) respectivamente, y por tanto, se distinguirán en el tipo de penalización. Esta diferencia de norma marcará notables diferencias entre las dos regresiones, siendo especialmente interesante para el caso de $p >> n$, que motiva este trabajo, la regresión Lasso; ya que, como veremos más adelante, hace que algunos coeficientes de β se anulen, reduciendo notablemente la dimensión.

Regresión Ridge

La regresión Ridge encoge los coeficientes de regresión imponiendo una penalización. Representa un caso particular de las técnicas de regularización donde se aplica una norma L_2. Los coeficientes ridge minimizarán la suma de la suma de los cuadrados de los residuos

$$\hat{\beta}^{ridge} = \arg\min_{\beta_0,\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} \quad (3)$$

Aquí $\lambda \geq 0$ es un parámetro de complejidad que controla la cantidad de encogimiento: a mayor valor de λ mayor encogimiento. Los coeficientes son encogidos hacia cero, aunque sin llegar a tomar el valor cero. La regresión Ridge producirá predicciones más precisas que los modelos obtenidos por mínimos cuadrados, a menos que en el verdadero modelo la mayoría de los coeficientes sean nulos.

Una forma equivalente de escribir el problema Ridge es

$$\min_{\beta_0,\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\}, \text{ sujeto a } \sum_{j=1}^{p}\beta_j^2\le t \qquad (4)$$

donde aparece de forma explícita una limitación en el tamaño de los parámetros. Existe una correspondencia directa entre los parámetros λ en (3) y t en (4).

Cuando hay muchas variables correlacionadas en un modelo de regresión lineal, sus coeficientes pueden ser pobremente determinados y además tener una gran varianza. Puede ocurrir que un coeficiente positivo grande se compense con un coeficiente negativo de tamaño similar. Se puede mitigar este problema imponiendo una limitación sobre el tamaño de los parámetros como hemos hecho en (4).

Para la estimación por mínimos cuadrados de β, que minimiza la suma de los residuos al cuadrado (*residual sum of squares, RSS),* teníamos que $\hat{\beta}=(X'X)^{-1}X'y$. De forma similar, a partir de (3) escrito en forma matricial tenemos:

$$RSS(\lambda)=(y-X\beta)'(y-X\beta)+\lambda\beta'\beta \qquad (5)$$

Y de forma similar a como hemos resuelto antes las soluciones de la regresión Ridge a partir de minimizar (5) serán

$$\hat{\beta}^{ridge}=(X'X+\lambda I_p)^{-1}X'y \qquad (6)$$

donde I_p la matriz identidad de dimensión $p\times p$. Observemos que eligiendo la penalización cuadrática $\beta'\beta$ la solución de la regresión Ridge es otra vez una función lineal de y. La solución añade una constante $\lambda\ge 0$ a la diagonal de $X'X$ antes de invertir la matriz. Esto hace nuestro problema no singular aunque $X'X$ no sea de rango máximo, lo cual corrige la inestabilidad que posee el estimador por mínimos cuadrados. Esta fue la principal motivación para el desarrollo de la regresión Ridge (Hoerl-Kennard, 1970).

Regresión Lasso (lasso)

La regresión Lasso (Least Absolute Shrinkage and Selection Operator) es un método de selección de variable, y al igual que el método Ridge, es un método de encogimiento (shrinkage). Como ya hemos comentado, es una técnica de regresión lineal regularizada

como Ridge, con la diferencia en la penalización usada, ya que en el método Lasso usaremos una penalización L_1 en vez de la penalización L_2 que usábamos en Ridge.

Este cambio tiene importantes consecuencias: el estimador de Lasso producirá estimaciones nulas para algunos coeficientes y no nulas para otros, de esta forma el estimador Lasso realizará una selección de variables. Con el estimador de Ridge esto no ocurre, ya que este estimador encoge todos los coeficientes hacia cero pero sin llegar a alcanzar este valor.

El uso de Lasso se ha generalizado debido a su buen comportamiento para el caso $p >> n$, es decir, su facilidad de uso y relativo poco coste computacional (Tibshirani, 2011).

El estimador de lasso $\hat{\beta}^{lasso}$, se define como solución del problema de optimización con restricciones:

$$\underset{\beta_0,\beta}{mín}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\}, \text{ sujeto a } \sum_{j=1}^{p}|\beta|\le t \qquad (7)$$

Aquí t es parámetro a optimizar, donde $t \ge 0$. Se puede calcular con validación cruzada, una técnica de aprendizaje supervisado. Debido a la naturaleza de la limitación, haciendo t suficientemente pequeña tendremos que algunos de los coeficientes son exactamente cero. Si t es elegida mayor que $t_0 = \sum_{i=1}^{p}\left|\hat{\beta}_j\right|$, donde $\hat{\beta}_j = \hat{\beta}_j^{mco}$ la estimación por mínimos cuadrados, entonces la estimación Lasso son las $\hat{\beta}_j$. Por otro lado, si $t = t_0/2$ por ejemplo, entonces los coeficientes de los mínimos cuadrados son encogidos a la mitad de media.

Podemos escribir también el problema Lasso mediante la equivalente forma lagrangiana como:

$$\hat{\beta}^{LASSO} = \arg\underset{\beta_0,\beta}{mín}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \qquad (8)$$

Observamos la similitud con la regresión Ridge, donde la penalización L_2 Ridge $\sum_{i=1}^{p}\beta_j^2$ es sustituida por la penalización L_1 Lasso $\sum_{i=1}^{p}\left|\beta_j\right|$. Esta última penalización hace

que las soluciones en la y_i sean no lineales y no hay en general una forma cerrada para la expresión como en los mínimos cuadrados y Ridge.

Calcular la solución del problema Lasso (7), es un problema de programación cuadrática. Como veremos más adelante, para resolverlo usaremos un algoritmo de ajuste y de selección de variables para modelos lineales llamado LARS (*Least Angle Regression*), que es un algoritmo eficiente que calcula el camino completo de soluciones cuando variamos λ, con el mismo coste computacional que para la regresión ridge.

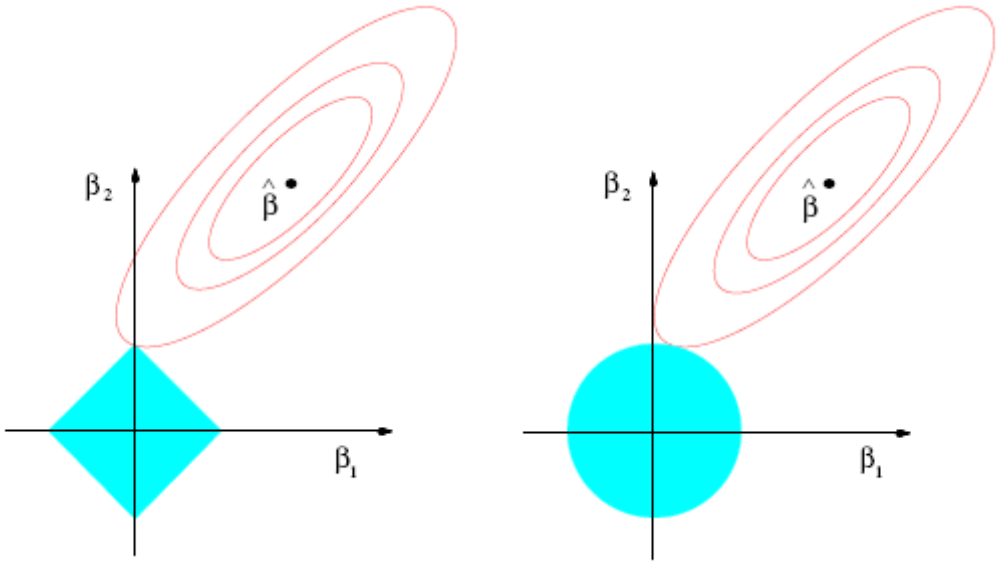


Figura 2: Gráfica obtenida de Hastie., Tibshirani y Friedman (2009). *Representación de la estimación lasso a la izquierda y Ridge a la derecha para el caso de dos dimensiones. Se muestran los contornos de error y las funciones límite .Las áreas sólidas son las regiones límite* $|\beta_1| + |\beta_2| \leq t$ *y* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectivamente, mientras las elipses son los contornos de la función de error mínimo cuadrático. En la restricción Lasso la solución suele estar en los vértices del cuadrado donde alguno de los* β_j *es cero.*

LARS (Least Angle Regression)

El algoritmo LARS (*Least Angle Regression*) es un algoritmo de ajuste y de selección de variables para modelos lineales (Efron et al., 2004). LARS está intimamente relacionado con el método lasso, y de hecho provee de un algoritmo muy eficiente para calcular el camino completo lasso. LARS usa un estrategia muy similar al algoritmo *forward stepwise.*

Es un modelo de selección de variable que realiza los procedimientos por etapas. Computacionalmente es equivalente al ajuste de un modelo por mínimos cuadrados. Los

pasos que sigue el algoritmo LARS para la selección de variable serán de la siguiente manera (Efron et al., 2004):

a) Se normalizan los datos con media cero y desviación uno.

b) Se comienza con todos los coeficientes igualados a cero.

c) Se busca la variable más correlacionada con la respuesta.

d) Se hace un salto en la dirección de la variable seleccionada y entonces en ese momento en vez de seguir la dirección de esta variable, se sigue una dirección equiangular entre ambas variables.

e) Se repite el proceso hasta que entren todas las variables y al finalizar tenemos la solución de mínimos cuadrados.

Este algoritmo presenta una serie de ventajas que hace que su uso sea aconsejable. Una ventaja que nos concierne en este trabajo en particular, es que funciona bastante bien en los casos de alta dimensionalidad $p >> n$. Es computacionalmente muy eficiente, ya que como ya hemos mencionado, requiere un esfuerzo similar que con los mínimos cuadrados. Una ventaja que nos resulta fundamental, es que con una simple modificación del algoritmo se puede obtener otro modelo de regresión ya estudiado como Lasso. Por último, se obtiene una aproximación simple de los grados de libertad del estimador LARS que permite derivar una estimación del error de predicción (Efron et al., 2004; Hastie et al., 2009).

El ***algoritmo LARS,*** que aparece en Hastie et al. (2009), se puede secuenciar de la siguiente forma:

1. Estandarizamos los valores predictores y comenzamos con los valores residuales $r = y - \bar{y}$, con $\beta_1, \beta_2, ..., \beta_p = 0$.
2. Buscamos los predictores x_j más correlacionados con el valor residual r.
3. Cambiamos β_j de cero al coeficiente $\langle X_j, r \rangle$ estimado por mínimos cuadrados, hasta otra variable X_k que tenga tanta correlación con el residuo actual, $r = y - \beta_j X_j$ como X_j.
4. Cambiamos β_j y β_k en la dirección definida por su coeficiente conjunto estimado por mínimos cuadrados del residuo actual sobre (X_j, X_k), hasta otra variable X_l que tenga tanta correlación con el residuo actual $r = y - \beta_j X_j - \beta_k X_k$.

5. Continuamos de esta manera hasta que los p predictores se hayan introducido. Después de un mínimo de $(n-1, p)$ pasos, llegamos hasta solución completa por mínimos cuadrados.

Si $p > n-1$ el algoritmo LARS alcanza la solución de residuo cero después de $n-1$ pasos (el -1 viene del hecho de haber centrado los datos).

Ya hemos comentado que una pequeña modificación en este algoritmo nos da el camino completo lasso. La modificación se producirá en el cuarto paso y será como sigue:

El ***algoritmo LARS: Lasso modificado:***

4a Si un coeficiente β_j no se anula nunca, se elimina su variable correspondiente y se vuelve a recalcular el modelo por mínimos cuadrados.

Este algoritmo es extremadamente eficiente, requiriendo el mismo orden de cálculo que un ajuste por mínimos cuadrados usando p pasos. LARS siempre toma p pasos para completar las estimaciones por mínimos cuadrados. El camino lasso puede tomar más de p pasos, aunque los dos métodos son muy similares. El algoritmo LARS lasso modificado es una forma muy eficiente de calcular la solución de cualquier problema lasso, especialmente cuando $p >> n$.

Un método alternativo al algoritmo LARS para la estimación de lasso es el de Coordenada Descendente (Hastie et al., 2009). La idea principal consiste en fijar el parámetro de penalización λ y optimizar sucesivamente respecto de cada parámetro β_j, dejando los restantes parámetros $\beta_k, k \neq j$, fijos en sus valores actuales. Este método puede ser más eficiente que LARS, especialmente cuando $p >> n$ (Friedman et al., 2010).

Penalizaciones no convexas

Para terminar el apartado de las penalizaciones, mencionaremos un tipo de penalización no convexa desarrollada por Fan y Li (2001) y que se está utilizando cada vez más estos últimos años para salvar las situaciones donde las penalizaciones L_q en general, y Ridge y Lasso en particular, no trabajan bien, ya sea por la *esparcisidad* o por la

insesgadez. El único problema que nos plantea la implantación de este método es que es un problema no convexo. Fan y Li (2001) sugieren usar aproximaciones locales convexas de la función objetivo.

Fan y Li (2001) proponen la penalización SCAD (*Smoothly Clipped Absolute Deviation*):

$$\phi_\lambda(\beta_j)=\begin{cases}\lambda|\beta_j| & si \quad 0\le|\beta_j|<\lambda \\ -\left(\beta_j^2-2a\lambda|\beta_j|+\lambda^2\right)/(2(a-1)) & si \quad \lambda\le|\beta_j|<a\lambda \\ (a+1)\lambda^2/2 & si \quad |\beta_j|\ge a\lambda\end{cases}$$

donde $a>2$ y $\lambda>0$ son parámetros de ajuste, que suelen ser seleccionados mediante validación cruzada. Los autores aconsejan utilizar $a\approx 3.7$ para mejorar la eficiencia computacional. La penalización SCAD es muy parecido a la penalización L_1 (lasso) para valores pequeños de β_j, y para valores grandes la primera es constante y la última no.

El estimador de SCAD, $\hat{\beta}^{SCAD}$, se define para a y λ fijos, como el que minimiza:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\sum_{j=1}^{p}\phi_\lambda(\beta_j) \qquad (9)$$

3.1.2. Graphical Lasso

El hecho de que el número de parámetros a estimar crezca cuadráticamente respecto a la dimensión, sugiere que es importante tener una alternativa robusta a la matriz de covarianzas muestrales estándar.

Para subsanar esta situación, muchos autores han optado por reducir el número efectivo de parámetros a través de imponer *sparsity* en la inversa de la matriz de covarianzas. Dempster (1972) sugiere ajustar elementos de la matriz inversa de covarianzas a cero. Meinshausen y Bühlmann (2006) proponen usar series de la regresión lasso para identificar los ceros de la inversa de la matriz de covarianzas. Y más recientemente, Yuan y Lin (2007), Banerjee et al. (2008) y Friedman et al. (2007) enmarcan esto como un problema de estimación *sparse*, implementando la máxima verosimilitud penalizada con la

penalización lasso sobre la inversa de la matriz de covarianzas. Esto es conocido como ***graphical lasso***.

Los ceros de la matriz de covarianzas inversa Σ^{-1} son de interés porque corresponden con la independencia condicional entre las variables. Es decir, si la componente $ij-$ésima de $\Theta=\Sigma^{-1}$ es cero, entonces la variable i y la j son condicionalmente independientes.

Asumiendo que las observaciones siguen una distribución multivariante Gaussiana con media μ y matriz de covarianzas Σ, es interesante estudiar la distribución condicional de una variable respecto al resto, donde Θ tiene un papel explícito. Supongamos que tenemos la partición $X=(Z,Y)$ donde $Z=(X_1,...,X_{p-1})$ consiste en las $p-1$ primeras variables e $Y=X_p$ es la última. Entonces tenemos la distribución condicional de Y dado Z (Mardia et al., 1979)

$$Y|Z=z\sim N\left(\mu_Y+(z-\mu_Z)'\cdot\Sigma_{ZZ}^{-1}\cdot\sigma_{ZY},\sigma_{YY}-\sigma_{ZY}'\cdot\Sigma_{ZZ}^{-1}\cdot\sigma_{ZY}\right) \quad (10)$$

donde hemos partido Σ como

$$\Sigma=\begin{pmatrix}\Sigma_{ZZ} & \sigma_{ZY}\\ \sigma_{ZY}' & \sigma_{YY}\end{pmatrix}$$

La media condicional en (10) tiene exactamente la misma forma que la regresión lineal múltiple de Y sobre Z, con los coeficientes de regresión $\beta=\Sigma_{ZZ}^{-1}\cdot\sigma_{ZY}$ (Friedman et al., 2010). Si partimos $\Theta=\Sigma^{-1}$ de la misma manera y sabiendo que $\Sigma\Theta=I$ nos queda lo siguiente:

$\theta_{ZY}=-\theta_{YY}\cdot\Sigma_{ZZ}^{-1}\cdot\sigma_{ZY}$, donde $\frac{1}{\theta_{YY}}=\sigma_{YY}-\sigma_{ZY}'\cdot\Sigma_{ZZ}^{-1}\cdot\sigma_{ZY}>0$. Entonces

$$\beta=\Sigma_{ZZ}^{-1}\cdot\sigma_{ZY}=-\frac{\theta_{ZY}}{\theta_{YY}}$$

Esto establece una relación directa entre los coeficientes de la regresión lineal y $\Theta=\Sigma^{-1}$, y podemos ver que si los coeficientes β tienen valor cero también será cero θ_{ZY},

lo cual implica que los correspondientes elementos de Z son condicionalmente independientes de Y.

Entonces $\Theta = \Sigma^{-1}$ captura toda la información de segundo orden, tanto la estructural como la cuantitativa, necesaria para describir las distribuciones condicionales de cada variable (vértice) dado el resto, y es el llamado parámetro "natural" del modelo gráfico Gaussiano que veremos a continuación.

Modelos gráficos Gaussianos

Un modelo gráfico o un campo aleatorio de Markov (*Markov random field*) es una familia de distribuciones de probabilidad para las cuales la independencia condicional y las propiedades de factorización están reflejadas en una gráfica.

Sea $X = \left(X_1,...,X_p\right)$ un vector de media cero gaussiano.

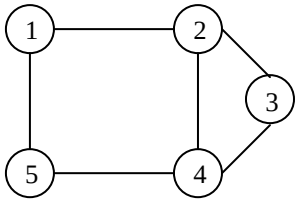
Su densidad puede ser parametrizada por la inversa de la covarianza $\Theta^* = \left(\Sigma^*\right)^{-1} \in S_+^p$, donde $S_+^p = \left\{A \in R^{p\times p} \middle| A = A', A \geq 0\right\}$ y pude escribirse

$$f\left(x_1,...,x_p,\Theta^*\right) = \frac{1}{\sqrt{\left(2\pi\right)^p \det\left(\Theta^*\right)^{-1}}} \exp\left\{-\frac{1}{2} X' \cdot \Theta' \cdot X\right\}$$

Podemos ver reflejada esta distribución Gaussiana del vector aleatorio X en un modelo gráfico de la siguiente forma:

Suponemos que tenemos el grafo $G = \left(V, E\right)$ con un conjunto de vértices $V = \left\{1,2,...,p\right\}$ y aristas E, de tal forma que cada variable X_i está asociada con un vértice $i \in V$. El *Gaussian Markov random field* (GMRF) asociado al grafo G sobre el vector X, es entonces la familia de distribución Gaussiana con matrices de *concentración* Θ^* que respeta las aristas del grafo en el sentido que $\theta_{ij}^* = 0$ si $\left(i, j\right) \notin E$.

Ejemplo:



$$\Theta^* = \begin{pmatrix} \theta_{11}^* & \theta_{12}^* & 0 & 0 & \theta_{15}^* \\ \theta_{21}^* & \theta_{22}^* & \theta_{23}^* & \theta_{24}^* & 0 \\ 0 & \theta_{32}^* & \theta_{33}^* & \theta_{34}^* & 0 \\ 0 & \theta_{42}^* & \theta_{34}^* & \theta_{44}^* & \theta_{45}^* \\ \theta_{51}^* & 0 & 0 & \theta_{45}^* & \theta_{55}^* \end{pmatrix}$$

En este ejemplo presentamos unja gráfica de cinco variables. Los valores de θ_{ij}^* para los que no hay aristas entre las variables son igual a cero. Por ejemplo, como la variable 1 y la 3 no están unidas por una arista entonces $\theta_{13}^* = \theta_{31}^* = 0$

Método propuesto

Con el fin de obtener las aristas que podemos omitir en nuestra gráfica, numerosos autores en los últimos años han propuesto el uso de la regularización lasso para este propósito.

Meinshausen y Bühlmann (2006) afrontan el problema de tal forma que en vez de tratar de estimar totalmente Σ o $\Theta = \Sigma^{-1}$, estiman sólo las componentes de θ_{ij} que no se anulan. Para hacer esto, aplican una regresión lasso usando cada variable como la respuesta y las otras como predictores.

Banerjee et al. (2008) y Friedman et al. (2007) enmarcan el problema de estimación *sparse*, implementando la máxima verosimilitud penalizada con la penalización lasso sobre la inversa de la matriz de covarianzas. Supongamos que tenemos n observaciones multivariantes normales de dimensión p, con media μ y covarianza Σ. Sea $\Theta = \Sigma^{-1}$ y S la matriz de covarianzas muestrales, el problema será maximizar la función *log-verosimilitud (log-likelihood)*

$$L(\Theta) = \log|\Theta| - tr(S \cdot \Theta) - \lambda\|\Theta\|_1 \qquad (11)$$

donde $\|\Theta\|_1$ es la norma L_1, la suma de los valores absolutos de los valores de Σ^{-1}.

Banerjee et al. (2008) muestran que el problema (11) es convexo y consideran la estimación de Σ como sigue. Sea W el estimador de Σ. Demuestran que se puede

resolver el problema por optimización sobre cada fila y correspondiente columna de W y haciendo las particiones de W y S

$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}' & w_{22} \end{pmatrix}$, $S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}' & s_{22} \end{pmatrix}$, muestran que la solución para w_{12} satisface

$$w_{12} = \arg mín_y \left\{ y' \cdot W_{11}^{-1} \cdot y : \left\| y - s_{12} \right\|_\infty \leq \lambda \right\} \quad (12)$$

Banerjee et al. (2008) siguen demostrando que resolver (12) es equivalente a resolver el problema dual

$$mín_\beta \left\{ \frac{1}{2} \left\| W_{11}^{1/2} \cdot \beta - b \right\|^2 + \lambda \left\| \beta \right\|_1 \right\} \quad (13)$$

donde $b = W_{11}^{-1/2} \cdot s_{12}$. Si β resuelve (13), entonces $w_{12} = W_{11} \cdot \beta$ resuelve (12).

Primero verificaremos la equivalencia entre las soluciones para (11) y (13). Expandiendo la relación $W \cdot \Theta = I$ tenemos la siguiente expresión:

$$\begin{pmatrix} W_{11} & w_{12} \\ w_{12}' & w_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}' & \theta_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0' & 1 \end{pmatrix} \quad (14)$$

Ahora la ecuación sub-gradiente que maximiza la *log-verosimilitud* queda

$$\Theta^{-1} - S - \lambda \cdot Sign(\Theta) = 0 \quad (15)$$

Aquí usamos la notación sub-gradiente, con $Sign(\theta_{jk}) = sign(\theta_{jk})$ si $\theta_{jk} \neq 0$, y si no $Sign(\theta_{jk}) \in [-1,1]$ si $\theta_{jk} = 0$, y $\Theta^{-1} = W$. Ahora el bloque superior derecho de la ecuación (15), recordando que β y θ_{12} tienen signos opuestos, es

$$W_{11}\beta - s_{12} + \lambda \cdot Sign(\beta) = 0 \quad (16)$$

Veamos que este sistema es equivalente a la estimación de ecuaciones por una regresión lasso. Consideremos la regresión usual con resultado la variable dependiente Y y matriz predictora Z. La minimización lasso es:

$$\frac{1}{2}(y - Z\beta)' \cdot (y - Z\beta) + \lambda \cdot \|\beta\|_1 \quad (17)$$

El gradiente de esta expresión es

$$Z' \cdot Z \cdot \beta - Z' \cdot y + \lambda \cdot Sign(\beta) = 0 \quad (18)$$

$Z' \cdot y$ es el análogo de s_{12}, y reemplazando $Z' \cdot Z$ por W_{11}. El procedimiento resultante es llamado *Graphical lasso*, propuesto por Friedman et al. (2008b), construido sobre el trabajo de Banerjee et al. (2008). Todo esto se resume en el algoritmo que exponemos a continuación.

El algoritmo *Graphical Lasso* se puede secuenciar de la siguiente forma:

1. Iniciamos con $W = S + \lambda I$. La diagonal de W permanecerá sin cambios en los siguientes pasos.
2. Repetimos para $j = 1,2,...,p,1,2,...,p$ hasta que converja:
   a) Dividimos W en dos partes. En la parte 1 dejamos todas las filas y columnas menos la fila y columna j *-ésima*. En la parte 2 la j *-ésima* fila y columna.
   b) Resolvemos las ecuaciones estimadas $W_{11} \cdot \beta - s_{12} + \lambda \cdot sign(\beta) = 0$ usando el método de descenso coordinado para la regresión lasso.
   c) Actualizamos $w_{12} = W_{11} \cdot \hat{\beta}$.
3. En el ciclo final (para cada j) resolvemos $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, con $\frac{1}{\hat{\theta}_{22}} = w_{22} - w_{12}' \cdot \hat{\beta}$.

Friedman et al. (2008b) usan el método redescenso coordinado para resolver el problema modificado lasso en cada paso. Tomando $V = W_{11}$ y $u = s_{12}$, la solución quedará de la siguiente forma:

$$\hat{\beta}_j \leftarrow S\left(u_j - \sum_{k \neq j} V_{kj} \cdot \hat{\beta}_k, \lambda\right) / V_{jj}$$

para $j = 1,2,..., p-1,1,2,..., p-1$, donde S es el operador soft-threshold:

$S(x,t) = sign(x)(|x| - t)_+$

El proceso pasa cíclicamente por los predictores hasta la convergencia. Es fácil ver que los elementos de la diagonal de la matriz solución W son simplemente $s_{jj} + \lambda$, y estos

son establecidos en el paso 1 del algoritmo. El método *Graphical lasso* es extremadamente rápido, y puede resolver un problema con 1000 variables en menos de un minuto.

3.1.3. Estimación *sparse* de la matriz de covarianzas

Otro tipo diferente de modelo gráfico es el grafo de covarianzas o red de relevancia, en el cual los vértices están conectados con aristas bidireccionales si la covarianza entre las respectivas variables no es nula. El grafo de covarianzas es el modelo gráfico para independencias marginales. Entonces, estimación sparse de la matriz de covarianzas corresponde a estimar un grafo de covarianzas teniendo un pequeño número de aristas. Es bastante popular en estudios de genética, ver especialmente Butte et al. (2000) que introduce la noción de red de relevancia (*relevante network*). Chaudhuri et al., (2007) consideran el problema de estimar la matriz de covarianzas dando un patrón preespecífico de ceros.

Bien J. y Tibshirani (2012) proponen un método, que en contraste con los anteriores, estima tanto las covarianzas no nulas como la estructura gráfica, es decir, las localizaciones de los ceros simultáneamente. En su método no asumen un orden de las variables.

Este método hace para las matrices de covarianzas lo que el método graphical lasso hace con la inversa de la matriz de covarianzas, y al igual que el método graphical lasso este método propone maximizar una verosimilitud penalizada. Aunque no vamos a desarrollarlo completamente, esbozaremos algunas ideas sobre este método.

Supongamos que tenemos una muestra de p variables aleatorias de distribución normal de media cero. La función de *log-verosimilitud* es:

$$L(\Sigma) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{n}{2}tr\left(\Sigma^{-1}S\right)$$

Bien J. y Tibshirani (2012) sugieren añadir a la verosimilitud una penalización lasso sobre $P*\Sigma$, donde P es una matriz arbitraria sin elementos negativos. Una opción común es tomar como P una matriz con todos unos en su diagonal o todos ceros. Entonces, proponen que el estimador salga de resolver la siguiente ecuación:

$$\underset{\Sigma>0}{Mín}\left\{\log|\Sigma| + tr\left(\Sigma^{-1}\cdot S\right) + \lambda\left\|P*\Sigma\right\|_1\right\} \qquad (19)$$

El problema graphical lasso es idéntico a (19) excepto que la penalización toma la forma $\left\|\Sigma^{-1}\right\|_1$ y la variable a optimizar es Σ^{-1}.

Sin entrar en detalles, simplemente concluiremos este apartado dando el algoritmo que resuelve (19).

1. $\Sigma \leftarrow S$
2. Repetir
3. $\Sigma_0 \leftarrow \Sigma$
4. Repetir
5. $\Sigma \leftarrow \Im\left\{\Sigma - t\left(\Sigma_0^{-1} - \Sigma^{-1} S \Sigma^{-1}\right), \lambda t P\right\}$ donde $\Im$ denota *elementwise soft-thresholding* definido por $\Im(A,B)_{ij} = sign(A_{ij})(A_{ij} - B_{ij})_+$
6. Hasta la convergencia.

3.1.4. Ejemplo

Vamos a concluir con la aplicación del paquete de R *glasso* para implementar el método graphical lasso. La función nos devolverá una estimación de la matriz de covarianzas y de su inversa, la matriz de precisión.

Utilizaremos los datos de uno de los ejemplos anteriores en el que tenemos un tamaño muestral de 2 y un número de variables de 3. Empezamos cargando el paquete *glasso* (ver ANEXO 1).

```
> X3<-matrix(c(4,4.2,2,2.1,0.6,0.59),nrow=2,ncol=3)

> X3

     [,1]  [,2]  [,3]
[1,]  4.0   2.0  0.60
[2,]  4.2   2.1  0.59

# Calculamos la matriz de covarianzas muestral S3 de la matriz de datos X3

> S3<-var(X3)

> S3
```

```
      [,1]      [,2]      [,3]
[1,]  0.020    1e-02    -1e-03
[2,]  0.010    5e-03 -   5e-04
[3,] -0.001    -5e-04    5e-05
```

Calculamos el algoritmo lasso con $\lambda = 0.1$

```
> a<-glasso(S3, rho=.01)
> a
```

Estimación de la matriz de covarianzas

```
$w
             [,1]              [,2]             [,3]
[1,] 3.000000e-02  1.561251e-17   0.00000
[2,] 1.561251e-17  1.500000e-02   0.00000
[3,] 0.000000e+00  0.000000e+00   0.01005
```

Estimación de la matriz inversa de covarianzas

```
$wi
             [,1]              [,2]             [,3]
[1,]  3.333333e+01  -3.469447e-14   0.00000
[2,] -3.469447e-14   6.666667e+01   0.00000
[3,]  0.000000e+00   0.000000e+00  99.50249

$loglik
[1] 13.95967
$errflag
[1] 0
$approx
[1] FALSE
$del
[1] 0
# Número de iteraciones
$niter
[1] 2
```

En este caso, con parámetro $\lambda = 0.1$, observando la estimación de la matriz inversa de covarianzas, podemos observar cómo la variable X_1 es condicionalmente independiente con respecto a X_3 y X_2 también es condicionalmente independiente con respecto a X_3,

ya que los coeficientes de la matriz *wi* (matriz inversa de la matriz de coeficientes) presentan ceros en los términos θ_{13}^{*} y θ_{23}^{*}.

```
# Calculamos el algoritmo lasso con λ = 0.2
> aa<-glasso(S3,rho=.02, w.init=a$w, wi.init=a$wi)
> aa

# Estimación de la matriz de covarianzas
$w
     [,1]    [,2]    [,3]
[1,] 0.04   0.000   0.00000
[2,] 0.00   0.025   0.00000
[3,] 0.00   0.000   0.02005

# Estimación de la matriz inversa de covarianzas
$wi
     [,1]  [,2]    [,3]
[1,]   25    0    0.00000
[2,]    0   40    0.00000
[3,]    0    0   49.87531

$loglik
[1] 11.72592
$errflag
[1] 0
$approx
[1] FALSE
$del
[1] 0
# Número de iteraciones
$niter
[1] 0
```

En este caso, con parámetro $\lambda = 0.2$, observando la estimación de la matriz inversa de covarianzas, podemos observar cómo existe independencia condicional entre todas las variables, ya que los coeficientes de la matriz *wi* (matriz inversa de la matriz de coeficientes) presentan ceros en todos los términos fuera de la diagonal.

3.2 ESTIMACIÓN SHRINKAGE

La estimación shrinkage o de encogimiento de la matriz de covarianzas, es uno de los métodos más comunes que se utiliza para salvar las debilidades ya observadas de la matriz de covarianzas muestral. Esta nueva estimación estará bien condicionada y la matriz será definida positiva, con lo que podremos invertirla.

En un principio fue introducido para estudios financieros y desde entonces, la aproximación ha recibido considerable interés en diversos campos científicos, como biología, donde el número de observaciones es mucho menor que el de variables. La aproximación que propone este método es crear un nuevo estimador que salga de la combinación lineal de la matriz de covarianzas muestral y una matriz objetivo (target matriz), de tal manera que se pueda alcanzar un promedio ponderado que se aproxime lo más posible a la matriz de covarianzas real.

Ya hemos visto, que aunque la matriz de covarianzas muestral es un estimador insesgado, distorsiona sistemáticamente la autoestructura de Σ y no minimiza el error cuadrático medio (mean-squared error, MSE). Stein (1956, 1975) demostró este hecho, que se producía especialmente cuando *p/n* es grande. El propio Stein ya demostró que una mejor aproximación se podía obtener encogiendo (shrinking) la covarianza muestral. Desde entonces, muchos estimadores shrinkage han sido propuestos. Por ejemplo, Haff (1980) introdujo un estimador inspirado en la aproximación empírica Bayesiana; Yang and Berger (1994) obtuvieron expresiones para estimadores Bayesianos bajo un grupo de a prioris para la covarianza y más recientemente, Daniela y Kass (2001) volvieron a proponer un estimador inspirado en la aproximación empírica Bayesiana. Estos trabajos estaban dirigidos para los casos de matrices de covarianzas muestral invertibles cuando $n \geq p$, y no son buenos estimadores cuando $p >> n$, o son computacionalmente muy pesados (como por ejemplo, los métodos basados en las Cadenas de Markov Montecarlo muestrales, o MCMC sampling), o asumen una determinada distribución.

Fueron Ledoit y Wolf (2003, 2004a, 2004b) los que propusieron un estimador shrinkage para el caso $n < p$, el cual minimiza asintóticamente el MSE, aunque el estudio en su forma original no intentaba ser una medida para solucionar la escasez de observaciones. Este estimador está bien condicionado para muestras pequeñas y puede ser aplicado a problemas de alta dimensión.

Fue desarrollado para estimar la matriz de covarianzas en estudios financieros. Dada la multitud de casos que se presentan en distintas áreas de las ciencias donde el número de observaciones es mucho menor que el tamaño poblacional, Schäfer y Strimmer (2005), han extendido el estudio de Ledoit-Wolf a otros muchos campos de las ciencias.

Dando una idea general, podemos decir que el estimador shrinkage propuesto originalmente por Ledoit y Wolf es la combinación lineal óptima asintóticamente convexa $\hat{\Sigma}^* = \lambda T + (1-\lambda)S$, donde $\lambda \in [0,1]$ es analíticamente determinada y se denomina intensidad óptima shrinkage, T es la matriz objetivo estructurada, y S es la "in estructurada" e insesgada matriz de covarianzas muestral. El estimador shrinkage de la matriz de covarianzas Σ resultante es invertible (eligiendo una adecuada matriz T) y estabilizada.

Aunque dedicaremos el apartado 3.2.3 para hablar sobre el cálculo de la intensidad óptima shrinkage λ, esta se calculará analíticamente a partir de minimizar la función cuadrática de pérdida.
La matriz de covarianzas objetivo T juega un papel fundamental en el cálculo del estimador shrinkage. Su elección puede volverse algo compleja. Por un lado, T debe ser definida positiva y conllevar sólo un pequeño número de parámetros libres. Por otro lado, debe reflejar importantes características de la estructura de covarianzas entre las variables. Las matrices objetivo más comúnmente usadas son las dadas por Schäfer y Strimmer (2005), que veremos en el apartado 3.2.2.

3.2.1 Estimador shrinkage de Ledoit-Wolf

Ledoit y Wolf desarrollan este tipo de estimadores para el estudio de optimización de carteras en el mundo financiero, al ser conscientes de la incorrecta estimación que se producía con la matriz de estimación muestral. En su lugar, sugirieron usar la matriz obtenida de la matriz de covarianzas muestral tras una transformación llamada shrinkage o de encogimiento. Esta trata de llevar los coeficientes extremos hacia un valor central, reduciendo de este modo el error de estimación. El motivo del método es que esos coeficientes en la matriz de covarianzas muestral que son extremadamente altos, tienden a contener un gran error positivo y necesitan ser llevados hacia abajo para compensarlo. Igual ocurre con los coeficientes extremadamente bajos, que deben ser llevados hacia arriba para compensar el error negativo.

El estimador de la covarianza construido por Ledoit y Wolf (2003) funcionará bien para valores pequeños del tamaño muestral y gran número de variables, y además, es poco costoso de calcularlo.

Volviendo a referirnos a S, recordemos que tiene una serie de ventajas como su fácil cálculo y que es un estimador insesgado y que su principal desventaja es el hecho de que contiene una gran estimación de error cuando el número de datos (data points) es o comparable o menor que el número de variables (stocks), lo cual es muy normal en aplicaciones financieras. Alternativamente uno podría considerar un estimador muy estructurado, como el modelo single-factor de Sharpe (1963). Este modelo sería de la siguiente forma:

$X_{it} = \alpha_i + \beta_i X_{0t} + \varepsilon_{it}$, donde X_{it} es el rendimiento de la acción (stock) i-ésima, X_{0t} es el rendimiento del mercado (market return), y ε_{it} es el intrínsico rendimiento del stock i en el periodo t y es incorrelacionado con el rendimiento del mercado. Esto implica la matriz de covarianzas

$\Phi = \beta\beta'\sigma_{00}^2 + \Omega_\varepsilon$, donde $\beta = \left(\beta_1,...,\beta_p\right)'$ es $N \times 1$, y σ_{00}^2 es la varianza de la cartera de valores (portfolio) del mercado.

Un Estimador para Φ es

$\hat{\Phi} = BB'\hat{\sigma}_{00}^2 + \hat{\Omega}_\varepsilon$, donde $B = \left(b_1,...,b_p\right)'$ siendo b_i es el estimador de mínimos cuadrados para β_i, y $\hat{\Omega}_\varepsilon = diag\left(\hat{\sigma}_{1,\varepsilon}^2,...,\hat{\sigma}_{p,\varepsilon}^2\right)$, y cada $\hat{\sigma}_{i,\varepsilon}^2$ está basada en los residuos OLS. Más específicamente,

$$b_i = \left(\sum_{t=1}^{N} X_{0t}^2\right)^{-1} \sum_{t=1}^{N} X_{0t} X_{it}, \ (i = 1,...,N)$$

$$\hat{\sigma}_i^2 = \frac{1}{N-1}\sum_{t=1}^{N} \hat{\varepsilon}_{it}^2, \text{ donde } \hat{\varepsilon}_{it} = X_{it} - b_i X_{ot}$$

Finalmente, $\hat{\sigma}_{00}^2$ es la varianza muestral del rendimiento del mercado (market return).

Este estimador contiene relativamente poco error pero, por otro lado, tiende a ser inespecífico y bastante sesgado.

La idea del estimador de Ledoit y Wolf, fue considerar la matriz de covarianzas muestral S y un estimador altamente estructurado F. El resultado es que el nuevo

estimador es una combinación lineal convexa de la forma $\lambda F + (1-\lambda)S$, donde λ es un número entre 0 y 1.
La técnica se denomina shrinkage (o de encogimiento) ya que la matriz de covarianzas muestral es "shrunk", es decir "encogida" hacia el estimador estructurado.

Un estimador obtenido por este método tiene tres ingredientes: un estimador no estructurado, un estimador estructurado (matriz objetivo o shrinkage target), y una constante shrinkage (intensidad óptima shrinkage).

- **Matriz objetivo (shrinkage target)**

  La matriz objetivo debe cumplir dos requerimientos al mismo tiempo: que esté altamente estructurada (sólo tiene un número pequeño de parámetros) y que a pesar de ello, refleje importantes características de la cantidad que está siendo estimada. Ledoit y Wolf (2003) sugieren la matriz "single-factor" de Sharpe (1963) como matriz objetivo. Sin embargo, Ledoit y Wolf (2004) sugieren una nueva matriz objetivo: el modelo de correlación constante (*the constant correlation model*), que da resultados similares y es más fácil de aplicar. El modelo dice que todas las correlaciones son iguales. La estimación del modelo es directa. La media de todas las correlaciones muestrales es el estimador de la correlación constante común. Este número junto con el vector de las varianzas muestrales, nos da la matriz objetivo, denotada por F anteriormente.
  La descripción de esta matriz objetivo es la siguiente:

  Sea y_{it}, $1 \le i \le N$, $1 \le t \le T$, el retorno de la acción (return *stock*) i durante el periodo t. El análisis asume que el retorno de la acción es independiente e idénticamente distribuida (*iid*) sobre el tiempo y tiene sus cuartos momentos finitos. La media muestral del retorno de la acción (return *stock*) i está dada por $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$.
  Sea Σ la matriz de covarianzas poblacional y S la matriz de covarianzas muestrales. Las entradas de las matrices Σ y S se escriben como σ_{ij} y s_{ij}, respectivamente.
  Las correlaciones poblacionales y muestrales entre los retornos de las acciones i y j están dadas por

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \qquad \text{y} \qquad r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

Las medias de las correlaciones poblacionales y muestrales están dadas por

$$\bar{\rho} = \frac{2}{(N-1)N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\rho_{ij} \qquad \text{y} \qquad \bar{r} = \frac{2}{(N-1)N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}r_{ij}$$

Se define la matriz de correlación poblacional Φ por medio de las varianzas poblacionales y la media de correlación poblacional:

$$\phi_{ii} = \sigma_{ii} \qquad \text{y} \qquad \phi_{ij} = \bar{\rho}\sqrt{\sigma_{ii}\sigma_{jj}}$$

Y se define la matriz de correlación muestral F por medio de las varianzas muestrales y la media de correlación muestral:

$$f_{ii} = s_{ii} \qquad \text{y} \qquad f_{ij} = \bar{r}\sqrt{s_{ii}s_{jj}}$$

Esta matriz F, es nuestra matriz objetivo.

- **Constante shrinkage (intensidad óptima shrinkage)**

A la hora de calcular el estimador shrinkage, un problema que surge es el de la elección de la constante shrinkage. Cualquier elección de λ entre 0 y 1 nos daría una media ponderada entre S y F. Pero este resultado tiene infinitas posibilidades.
Obviamente, existe un valor óptimo para la constante shrinkage. Esta minimizará la distancia esperada entre el estimador shrinkage y la verdadera matriz de covarianzas. Llamemos a este número λ^*. La estimación de constante shrinkage óptima la escribimos como $\hat{\lambda}^*$. Desarrollaremos la fórmula en el apartado 3.2.3.

Por tanto, la matriz shrinkage para estimar la matriz de covarianzas poblacional Σ nos quedará de la siguiente forma:

$$\hat{\Sigma}_{Shrink} = \hat{\lambda}^* F + \left(1-\hat{\lambda}^*\right)S$$

3.2.2 Estimadores shrinkage de Schäfer y Strimmer

Partiendo de los trabajos de Ledoit-Wolf, Schäfer y Strimmer (2005) compilan y desarrollan seis matrices objetivo (*Target matriz*). Para todos estos casos, el estimador shrinkage que resulta, presentará el mismo orden de complejidad algorítmica que la estimación estándar S.

A partir de este momento, para nombrar la matriz objetivo vamos a utilizar la letra T en vez de la letra F que hemos usado en el punto anterior, ya que una de las matrices objetivo la nombraremos como "target F" y podría conducir a confusión. Los elementos de esta matriz objetivo los nombraremos como $T = (t_{ij})$.

Las matrices objetivo más comúnmente usadas son la matriz identidad y sus múltiplos escalares. Estas corresponden a las matrices objetivo A y B, target A: "Diagonal, unit variante" y target B: "Diagonal, common variante". Otra matriz objetivo que podemos incluir en este grupo es la matriz objetivo C, target C: "common (co)variante". Esta matriz tendrá dos parámetros del modelo de covarianzas, y además de la varianza común (common variante) de B se le añade la covarianza común (common covariance). Estas tres matrices objetivo se definen de la siguiente forma:

Target A "Diagonal, unit variante"	**Target B** "Diagonal, common variante"	**Target C** "Common (co)variante"
0 parámetros estimados	*1 parámetros estimados:* v	*2 parámetros estimados:* v, c
$t_{ij} = \begin{cases} 1 & si \;\; i = j \\ 0 & si \;\; i \neq j \end{cases}$	$t_{ij} = \begin{cases} v = avg(s_{ij}) & si\; i = j \\ 0 & si\; i \neq j \end{cases}$	$t_{ij} = \begin{cases} v = avg(s_{ij}) & si\; i = j \\ c = avg(s_{ij}) & si\; i \neq j \end{cases}$

Tabla 1*: Las tres primeras matrices objetivo (target) usadas en la estimación shrinkage de la matriz de covarianzas. La letra* v *, es la abreviatura de la media de las varianzas muestrales,* c *de la media de las covarianzas muestrales.*

Estas tres matrices comparten bastantes propiedades. Primero, tienen una dimensión muy baja (de 0 a 2 parámetros), lo cual nos da unas matrices fuertemente estructuradas y con pocos datos para completarlas. Segundo, el estimador resultante encoge (shrink) todos

los componentes de la matriz de covarianzas muestral, es decir, las diagonales y todas las entradas o elementos de fuera de la diagonal.

Estas matrices objetivo son empleadas en diferentes estudios. Por ejemplo, la target A, es usada en alguno de los estudios ya expuestos en este trabajo, como son la regresión ridge y la regularización de Tikonov (Hastie et aal., 2001). La target B es utilizada por Friedman (1989), quien estimó λ por medio de la validación cruzada, por Leung y Chan (1998), quienes usan $\lambda = 2/(n+2)$, por Dobra et al. (2004), como un parámetro en la inversa de a priori de Wishart para la matriz de covarianzas, y finalmente también por Ledoit y Wolf (2004b). La target C no se suele usar demasiado.

Otras clases de matrices objetivo son: target D, "diagonal, unequal variante", target E, "perfect positive correlation" y target F, "constant correlation". Estas matrices objetivo se definen de la siguiente forma:

Target D "Diagonal, unequal variante"	**Target E** "Perfect positive correlation"	**Target F** "Constant correlation"
p parámetros estimados: s_{ii}	*p parámetros estimados:* s_{ii}	*p+1 parámetros estimados:* $s_{ii}, \bar{r}$
$t_{ij} = \begin{cases} s_{ii} & si\ i = j \\ 0 & si\ i \neq j \end{cases}$	$t_{ij} = \begin{cases} s_{ii} & si\ i = j \\ \sqrt{s_{ii}s_{jj}} & si\ i \neq j \end{cases}$	$t_{ij} = \begin{cases} s_{ii} & si\ i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & si\ i \neq j \end{cases}$

***Tabla 2**: Las tres siguientes matrices objetivo (target) usadas en la estimación shrinkage de la matriz de covarianzas. La letra $\bar{r}$, es la abreviatura de la media de las correlaciones muestrales.*

Una característica compartida por estas tres matrices objetivo es que tienen abundancia de parámetros, y que sólo inducen encogimiento (shrinkage) de los elementos que están fuera de la diagonal de la matriz de covarianzas muestrales S . Las dos últimas matrices objetivo (target E y target F) fueron introducidas con el propósito de modelar el rendimiento de acciones (stock returns). Estas tienden a ser fuertemente correlacionados positivamente (Ledoit y Wolf, 2003, 2004b).

Para su estudio, Schäfer y Strimmer utilizaron la matriz objetivo target D para la estimación de las matrices de covarianzas y correlación surgidas en problemas de genética. Esta matriz "diagonal, unequal variante", supone un punto de encuentro entre las matrices de baja dimensión target A, B y C, y los modelos correlacionados E y F. Como las matrices objetivos más simples, A y B, Encoge (shrink) los elementos de fuera de la diagonal de la

matriz a cero. Sin embargo, a diferencia de las matrices A y B, deja los elementos de la diagonal intactos, es decir, no encoge las varianzas. Por tanto este modelo asume que los parámetros de la matriz de covarianzas son de dos clases, y son tratados de forma diferente en el proceso shrinkage.

3.2.3 Intensidad óptima shrinkage

Como ya hemos comentado, la elección de la intensidad óptima shrinkage λ es fundamental para poder aplicar con éxito nuestro estimador shrinkage. A lo largo de estos años se han planteado diversos métodos para su estimación. Todos los estimadores shrinkage para muestras finitas en la teoría estadística de la decisión y también en Frost and Savarino's (1986), se rompen cuando el número de observaciones es menor que el de variables ($p > n$), ya que su función de pérdida cuadrática utiliza la matriz de covarianzas inversa. Una forma muy común pero computacionalmente muy pesada, es usar la validación cruzada, como en Friedman (1989). Otro método es la estimación en un contexto empírico Bayesiano (Morris (1983); Greenland (2000)). Sin embargo, dada la rapidez del método, es preferible determinar λ analíticamente. Concretamente, Ledoit-Wolf (2003) enunciaron un teorema para elegir λ que garantiza un mínimo MSE sin la necesidad de tener una distribución específica ni ser muy costoso computacionalmente, tal como ocurre en otros métodos como con la validación cruzada, el bootstrap o el MCMC (Markow Chain Monte Carlo).

Schäfer y Strimmer (2005) calcularon un estimador insesgado para λ, y bajo la presunción de normalidad de los datos Chen, Wiesel y Hero (2010) nos dan un estimador insesgado de λ usando el teorema de Rao-Blackwell.

Vamos a desarrollar la estimación que dieron Ledoit-Wolf (2003) y las distintas estimaciones de λ para cada matriz objetivo dadas por Schäfer y Strimmer (2005).

- **Intensidad óptima shrinkage de Ledoit-Wolf**

  Los autores proponen una matriz de pérdida que no depende de la inversa. Consiste en la medida de la distancia entre la matriz de covarianzas real y la muestral basada en la norma de Frobenius. Dicha norma se define de la siguiente forma:

La norma de Frobenius para la matriz simétrica Z de tamaño $p \times p$ con elementos $\left(z_{ij}\right)_{i,j=1,\ldots,p}$ y autovalores $\lambda_{i=1,\ldots,p}$, se define como

$$\|Z\|^2 = Traza\left(Z^2\right) = \sum_{i=1}^{p}\sum_{j=1}^{p} z_{ij}^2 = \sum_{i=1}^{p} \lambda_i$$

Nuestra función de pérdidas cuadrática quedará de la siguiente forma:

$$L(\lambda) = \|\lambda T + (1-\lambda)S - \Sigma\|^2$$

La cual nos da la función de riesgo

$$R(\lambda) = E(L(\lambda)) = \sum_{i=1}^{p}\sum_{j=1}^{p} E\left(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij}\right)^2 =$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p} Var\left(\lambda t_{ij} + (1-\lambda)s_{ij}\right) + \left[E\left(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij}\right)\right]^2 =$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p} \lambda^2 Var\left(t_{ij}\right) + (1-\lambda)^2 Var\left(s_{ij}\right) + 2\lambda(1-\lambda)Cov\left(t_{ij}, s_{ij}\right) + \lambda^2\left(\phi_{ij} - \sigma_{ij}\right)^2$$

El objetivo ahora es minimizar $R(\lambda)$ con respecto a λ. Para lo cual calculamos la dos primeras derivadas de $R(\lambda)$.

$$R'(\lambda) = 2\sum_{i=1}^{p}\sum_{j=1}^{p} \lambda Var\left(t_{ij}\right) - (1-\lambda)Var\left(s_{ij}\right) + (1-\lambda)Cov\left(t_{ij}, s_{ij}\right) + \lambda\left(\phi_{ij} - \sigma_{ij}\right)^2$$

$$R''(\lambda) = 2\sum_{i=1}^{p}\sum_{j=1}^{p} Var\left(t_{ij} - s_{ij}\right) + \left(\phi_{ij} - \sigma_{ij}\right)^2$$

Tomando $R''(\lambda) = 0$ y resolviendo, tenemos

$$\lambda^* = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} Var\left(t_{ij} - s_{ij}\right) - Cov\left(t_{ij}, s_{ij}\right)}{\sum_{i=1}^{p}\sum_{j=1}^{p} Var\left(t_{ij} - s_{ij}\right) + \left(\phi_{ij} - \sigma_{ij}\right)^2}$$

Como $R''(\lambda) > 0$ para cualquier valor de λ, esta solución verifica que es un mínimo para nuestra función de riesgo.

Es fácil ver que $\lambda^* = O(1/n)$. Efectivamente, el siguiente Teorema muestra el comportamiento asintótico de primer orden que tiene la intensidad shrinkage óptima λ^*. La demostración se puede ver en Ledoit-Wolf (2003).

Teorema. Sea π la suma de las varianzas asintóticas de los elementos de la matriz de covarianzas muestral modificados por la escala por $\sqrt{n}$: $\pi = \sum_{i=1}^{p}\sum_{j=1}^{p} AsyVar\left[\sqrt{n}s_{ij}\right]$. Similarmente, sea ρ la suma de las varianzas asintóticas de los elementos de la matriz de covarianzas *single-index* con los datos de la matriz de covarianzas muestral modificados por la escala $\sqrt{n}$: $\rho = \sum_{i=1}^{p}\sum_{j=1}^{p} AsyCov\left[\sqrt{n}t_{ij}, \sqrt{n}s_{ij}\right]$. Finalmente, sea $\gamma = \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\phi_{ij} - \sigma_{ij}\right)^2$ la medida del modelo *single-index.* Entonces la intensidad shrinkage óptima λ^* satisface:

$$\lambda^* = \frac{1}{n}\frac{\pi - \rho}{\gamma} + O\left(\frac{1}{n^2}\right)$$

- **Intensidad óptima shrinkage de Schäfer y Strimmer**

Vamos a estudiar la estimación de la intensidad óptima a la que llegaron Schäfer y Strimmer para cada una de las matrices objetivo que plantearon (tablas 1 y 2). Estas se obtienen partiendo de los trabajos de Ledoit-Wolf (2003).
Vamos a esbozar rápidamente el estudio para la estimación de la intensidad óptima, ya que en el apartado anterior ya hemos hecho el desarrollo completo.

Sean $\Psi = (\psi_1, ..., \psi_p)$ los parámetros de alta dimensión de nuestro modelo, y $\Theta = (\theta_i)$ los parámetros de baja dimensión del submodelo. Para adaptar cada uno de los dos diferentes modelos a los datos asociados a la estimación, obtenemos $U = \hat{\Psi}$ y $T = \hat{\Theta}$. La lineal aproximación shrinkage sugiere combinar ambos estimadores en una media ponderada de la siguiente forma:

$$U^* = \lambda T + (1-\lambda)U$$

donde $\lambda \in [0,1]$ es la intensidad shrinkage. Como ya hemos visto, la forma para seleccionar un valor óptimo de la intensidad shrinkage se puede obtener calculando el valor de λ que minimiza la función de riesgo

$$R(\lambda) = E(L(\lambda)) = E\left(\sum_{i=1}^{p} (u_i^* - \psi_i)^2 \right)$$

Operando para obtener el mínimo valor de la MSE $R(\lambda^*)$, obtenemos la siguiente expresión para la intensidad óptima

$$\lambda^* = \frac{\sum_{i=1}^{p} Var(u_i) - Cov(t_i, u_i) - Bias(u_i)E(t_i - u_i)}{\sum_{i=1}^{p} E\left[(t_i - u_i)^2\right]}$$

Para la aplicación práctica de esta ecuación, necesitamos obtener una estimación de la intensidad óptima shrinkage λ^*. Ledoit-Wolf (2003) enfatizan en su trabajo en que los parámetros de la ecuación deben ser estimados consistentemente. Sin embargo, este es un requerimiento que deberían tener todos los estimadores. Schäfer y Strimmer (2005) sugieren reemplazar todas las varianzas y covarianzas esperadas, por sus correspondientes muestras insesgadas. Por tanto la ecuación quedaría

$$\lambda^* = \frac{\sum_{i=1}^{p} \hat{Var}(u_i) - \hat{Cov}(t_i, u_i) - \hat{Bias}(u_i)(t_i - u_i)}{\sum_{i=1}^{p} (t_i - u_i)^2}$$

Con muestras finitas λ^* podría exceder el valor de uno, e incluso en algunos casos ser negativo. Para evitarlo truncamos la intensidad estimada correspondiente usando $\lambda^{**} = \max(0, \min(1, \hat{\lambda}^*))$ cuando construyamos el estimador shrinkage.

Para calcular la intensidad óptima shrinkage λ^* que usaremos en las diferentes matrices objetivo (tabla 3) es necesario obtener unos estimadores insesgados de las varianzas y covarianzas de los elementos de $S = (s_{ij})$.

Sea x_{ki} la $k-ésima$ observación de la variable X_i y $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ su media empírica.

Sea $w_{kij} = (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$ y $\overline{w}_{ij} = \frac{1}{n}\sum_{k=1}^{n} w_{kij}$. Entonces la covarianza empírica insesgada será

$$\hat{Cov}(x_i, x_j) = s_{ij} = \frac{n}{n-1}\overline{w}_{ij}$$

Y la varianza correspondiente será

$$\hat{Var}(x_i) = s_{ii} = \frac{n}{n-1}\overline{w}_{ii}$$

Las varianzas y covarianzas empíricas de los elementos de la matriz S, serán calculadas de forma similar.

$$\hat{Var}(s_{ij}) = \frac{n^2}{(n-1)^2}\hat{Var}(\overline{w}_{ij}) = \frac{n}{(n-1)^2}\hat{Var}(w_{ij}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \overline{w}_{ij})^2$$

$$\hat{Cov}(s_{ij}, s_{lm}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \overline{w}_{ij})(w_{klm} - \overline{w}_{lm})$$

Momentos de orden alto de $\hat{Var}(s_{ij})$, concretamente varianzas y covarianzas de las medias de s_{ij}, son abandonados en la estimación del valor óptimo de λ^* en la tabla 3.

Los estimadores de la intensidad óptima λ^* para las diferentes matrices objetivo expuestas anteriormente las escribiremos en la siguiente tabla.

Target A "Diagonal, unit variante"	**Target B** "Diagonal, common variante"	**Target C** "Common (co)variante"
0 parámetros estimados	*1 parámetros estimados:* ν	*2 parámetros estimados:* ν, c
$t_{ij} = \begin{cases} 1 & si\ \ i = j \\ 0 & si\ \ i \neq j \end{cases}$	$t_{ij} = \begin{cases} \nu = avg(s_{ij}) & si\ i = j \\ 0 & si\ i \neq j \end{cases}$	$t_{ij} = \begin{cases} \nu = avg(s_{ij}) & si\ i = j \\ c = avg(s_{ij}) & si\ i \neq j \end{cases}$
$\lambda^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij}) + \sum_i \hat{Var}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - 1)^2}$	$\lambda^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij}) + \sum_i \hat{Var}(s_{ii})}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - \nu)^2}$	$\lambda^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij}) + \sum_i \hat{Var}(s_{ii})}{\sum_{i \neq j} (s_{ij} - c)^2 + \sum_i (s_{ii} - \nu)^2}$
Target D "Diagonal, unequal variante"	**Target E** "Perfect positive correlation"	**Target F** "Constant correlation"
p parámetros estimados: s_{ii}	*p parámetros estimados:* s_{ii}	*p+1 parámetros estimados:* $s_{ii}, \bar{r}$
$t_{ij} = \begin{cases} s_{ii} & si\ \ i = j \\ 0 & si\ \ i \neq j \end{cases}$	$t_{ij} = \begin{cases} s_{ii} & si\ i = j \\ \sqrt{s_{ii} s_{jj}} & si\ i \neq j \end{cases}$	$t_{ij} = \begin{cases} s_{ii} & si\ i = j \\ \bar{r}\sqrt{s_{ii} s_{jj}} & si\ i \neq j \end{cases}$
$\lambda^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij})}{\sum_{i \neq j} s_{ij}^2}$	$\lambda^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij}) - f_{ij}}{\sum_{i \neq j} (s_{ij} - \sqrt{s_{ii} s_{jj}})^2}$	$\lambda^* = \frac{\sum_{i \neq j} \hat{Var}(s_{ij}) - \bar{r} f_{ij}}{\sum_{i \neq j} (s_{ij} - \bar{r}\sqrt{s_{ii} s_{jj}})^2}$

Tabla 3*: Las seis matrices objetivo (target) usadas en la estimación shrinkage de la matriz de covarianzas y sus estimadores de la intensidad óptima shrinkage correspondientes. La expresión* t_{ij} *se define como:* $f_{ij} = \frac{1}{2}\left\{\sqrt{\frac{s_{jj}}{s_{ii}}}\hat{Cov}(s_{ii}, s_{ij}) + \sqrt{\frac{s_{ii}}{s_{jj}}}\hat{Cov}(s_{jj}, s_{ij})\right\}$.

3.2.4 Estimador shrinkage "SHIP"

Quiero concluir el estudio de los métodos shrinkage haciendo referencia al trabajo sobre este método de M. Jelizarow et al., (2010) y V. GuillemoT et al., (2011), aplicado al campo de la bioinformática. En él apuestan por la integración de información estructurada dentro del análisis estadístico en el cual al menos uno de los pasos necesita la estimación de una matriz de covarianzas de alta dimensión.

Para dirigir los cambios metodológicos que han surgido de la situación, ya estudiada a lo largo del trabajo, de tener un número de variables mucho mayor que el tamaño muestral,

$p >> n$, este método propone una estimación de la covarianza a través del procedimiento SHIP, que es la abreviatura de **SH**rinkage and **I**ncorporating **P**rior knowledge, es decir, un método shrinkage que incorpora un conocimiento a priori. El estimador resultante $\hat{\Sigma}_{SHIP}$ está basado en el estimador shrinkage introducido por Ledoit-Wolf (2003, 2004) y aplicado por Schäfer-Strimmer (2005) en el contexto de los datos de alta dimensionalidad en el campo de la genética.

Además, este nuevo estimador incorpora conocimiento biológico a priori sobre grupos funcionales de genes de la base de datos del KEGG, abreviatura de Kyoto enciclopedia of genes and genomes.

Recordemos que el estimador shrinkage propuesto por Ledoit y Wolf era de la forma $\hat{\Sigma}^* = \lambda T + (1-\lambda)S$, donde $\lambda \in [0,1]$ es la intensidad óptima shrinkage, T es una matriz estructurada llamada matriz objetivo y S es la matriz no estructurada correspondiente a la matriz de covarianzas muestral. El estimador shrinkage resultante es más eficiente y exacto que el muestral, está definido positivo y es invertible, tiene garantizada que minimiza el MSE (mean squared error) resultante de la función de pérdida cuadrática y en general todas las propiedades ya estudiadas. La matriz objetivo T es elegida y la intensidad óptima shrinkage λ es calculada.

La matriz objetivo T que elijamos debe ser positivamente definida y tener un pequeño número de parámetros, y además, debe reflejar importantes características entre las variables, en este caso entre los genes. Refiriéndonos a las tablas 1 y 2 donde aparecen las matrices de covarianzas objetivo de la A a la F dadas por Schäfer-Strimmer (2005), este tipo de estimador utilizará la target D y la F.

Para incorporar información procedente del KEGG PATHWAY, se propone trabajar con una modificación de la matriz objetivo target F donde los pares de genes conectados (es decir, genes de el mismo grupo funcional) tienen la correlación común $\bar{r}$ no nula. Recordemos que la target F es

$$t_{ij} = \begin{cases} s_{ii} & si\ i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & si\ i \neq j \end{cases}$$

En el caso en que un gen no esté en ningún grupo funcional de genes, asumiremos que forman su propio grupo con tamaño de grupo igual a uno como en Tai y Pan (2007a).

La matriz objetivo de esta modificación la denominaremos target G y servirá para calcular el nuevo estimador $\hat{\Sigma}_{SHIP} = \lambda G + (1-\lambda)S$, donde G es nuestra matriz objetivo y la intensidad óptima shrinkage λ será calculada analíticamente.

La nueva matriz objetivo target G se define de la siguiente forma:

Target G

$$t_{ij} = \begin{cases} s_{ii} & si \quad i = j \\ \bar{r}\sqrt{s_{ii}s_{jj}} & si \quad i \neq j, i \sim j \\ 0 & si \quad \text{Resto} \end{cases}$$

La notación $i \sim j$ significa que los genes i y j están conectados, es decir, pertenecen al mismo grupo funcional.

El estimador de covarianzas shrinkage $\hat{\Sigma}_{SHIP}$ está desarrollado en el paquete 'SHIP' de R, y lo utilizaremos en uno de los ejemplos que veremos a continuación.

3.2.5 Ejemplos

Ejemplo 1

En este caso usaremos el paquete de R *corpcor* para estimar la matriz de covarianzas (ver ANEXO 1). Para estimar la matriz de covarianzas usamos la función *cov.shrink.*

Cargamos el paquete R *corpcor*. Vamos a utilizar una nueva matriz de datos de seis variables y cinco muestras:

```
> X1<-matrix(c(4.2,3.9,4.3,4.1,3.8,2,2.1,2,2.2,2,0.6,0.59,0.58,0.63,0.56,4,3.7,3.9,3.8,
+3.8,1.9,2,2,2.1,1.8,1.6,1.59,1.58,1.62,1.63),nrow=5,ncol=6)
> X1
     [,1]  [,2] [,3] [,4] [,5] [,6]
[1,]  4.2  2.0 0.60  4.0  1.9 1.60
[2,]  3.9  2.1 0.59  3.7  2.0 1.59
[3,]  4.3  2.0 0.58  3.9  2.0 1.58
[4,]  4.1  2.2 0.63  3.8  2.1 1.62
[5,]  3.8  2.0 0.56  3.8  1.8 1.63
```

Calculamos la matriz de covarianzas muestral S para compararla después con la estimación shrinkage.

```
> S<-var(X1)
> S
          [,1]         [,2]        [,3]         [,4]         [,5]         [,6]
[1,]  0.04300 -0.00200 0.002100  0.01700  0.01050 -0.002550
[2,] -0.00200  0.00800 0.001850 -0.00550  0.00800  0.000450
[3,]  0.00210  0.00185 0.000670  0.00015  0.00235  0.000015
[4,]  0.01700 -0.00550 0.000150  0.01300 -0.00300 -0.000450
[5,]  0.01050  0.00800 0.002350 -0.00300  0.01300 -0.000800
[6,] -0.00255  0.00045 0.000015 -0.00045 -0.00080  0.000430

# Calculamos sus autovalores
> eigen(S)
$values
[1]  5.283010e-02  2.255665e-02  2.577584e-03  1.356683e-04  2.684093e-20 -
1.233148e-18
# Calculamos el número condicional
> rank.condition(S)
$condition
[1] Inf
```

Se puede observar cómo los autovalores prácticamente se anulan o son negativos, y el número condicional es singular, lo que nos dice que la matriz de covarianzas muestral no es invertible y está mal condicionada.

Vamos a calcular ahora la estimación shrinkage para distintos valores de la intensidad óptima shrinkage λ.

- $\underline{\lambda = 0.2}$

```
> S1<-cov.shrink(X1, lambda=0.2, lambda.var=0)
Specified shrinkage intensity lambda.var (variance vector): 0
Specified shrinkage intensity lambda (correlation matrix): 0.2

> S1
```

```
      [,1]          [,2]       [,3]        [,4]        [,5]         [,6]
[1,]  0.04300 -0.00160 0.001680  0.01360  0.00840 -0.002040
[2,] -0.00160  0.00800 0.001480 -0.00440  0.00640  0.000360
[3,]  0.00168  0.00148 0.000670  0.00012  0.00188  0.000012
[4,]  0.01360 -0.00440 0.000120  0.01300 -0.00240 -0.000360
[5,]  0.00840  0.00640 0.001880 -0.00240  0.01300 -0.000640
[6,] -0.00204  0.00036 0.000012 -0.00036 -0.00064  0.000430
```

```
# Calculamos sus autovalores
> eigen(S1)

$values
[1] 0.0498378530 0.0203300784 0.0047045248 0.0027103384 0.0002915062
0.0002256992
```

```
# Calculamos el número condicional
> rank.condition(S1)
$condition
[1] 220.8154
```

Ahora ya podemos observar que ningún autovalor se anula y todos son positivos, y su número condicional es finito, aunque sea grande. Esta matriz será invertible.

- $\lambda = 0.4$

```
> S2<-cov.shrink(X1, lambda=0.4, lambda.var=0)
Specified shrinkage intensity lambda.var (variance vector): 0
Specified shrinkage intensity lambda (correlation matrix): 0.4

> S2
      [,1]          [,2]       [,3]        [,4]        [,5]         [,6]
[1,]  0.04300 -0.00120 0.001260  0.01020  0.00630 -0.001530
[2,] -0.00120  0.00800 0.001110 -0.00330  0.00480  0.000270
[3,]  0.00126  0.00111 0.000670  0.00009  0.00141  0.000009
[4,]  0.01020 -0.00330 0.000090  0.01300 -0.00180 -0.000270
[5,]  0.00630  0.00480 0.001410 -0.00180  0.01300 -0.000480
[6,] -0.00153  0.00027 0.000009 -0.00027 -0.00048  0.000430
```

```
# Calculamos sus autovalores
> eigen(S2)
$values
[1] 0.0471749159 0.0181477664 0.0074458836 0.0045633518 0.0004365288
0.0003315535
# Calculamos el número condicional
> rank.condition(S2)
$condition
[1] 142.2845
```

- $\lambda = 0.6$

```
> S3<-cov.shrink(X1, lambda=0.6, lambda.var=0)
Specified shrinkage intensity lambda.var (variance vector): 0
Specified shrinkage intensity lambda (correlation matrix): 0.6

> S3
         [,1]        [,2]       [,3]         [,4]         [,5]          [,6]
[1,]  0.04300 -0.00080 8.4e-04  0.00680  0.00420 -0.001020
[2,] -0.00080  0.00800 7.4e-04 -0.00220  0.00320  0.000180
[3,]  0.00084  0.00074 6.7e-04  0.00006  0.00094  0.000006
[4,]  0.00680 -0.00220 6.0e-05  0.01300 -0.00120 -0.000180
[5,]  0.00420  0.00320 9.4e-04 -0.00120  0.01300 -0.000320
[6,] -0.00102  0.00018 6.0e-06 -0.00018 -0.00032  0.000430

# Calculamos sus autovalores
> eigen(S3)
$values
[1] 0.0449999151 0.0160567664 0.0099889418 0.0061124734 0.0005512046
0.0003906987
# Calculamos el número condicional
> rank.condition(S3)
$condition
[1] 115.1781
```

Podemos volver a observar que ningún autovalor se anula y todos son positivos, y su número condicional es finito y cada vez es menor, con lo que seguimos mejorando la estimación.

Con el paquete R *corpcor* también existe una función para estimar la inversa de la matriz de covarianzas Σ^{-1}. Esta función se define de la siguiente manera:

invcov.shrink(x, lambda, lambda.var, w, verbose=TRUE)

Los argumentos son los mismos que los de la anterior función.
Calculándolo para la última estimación con $\lambda = 0.6$ nos queda:

```
#Estimación de la inversa de la matriz de covarianzas
> InvS<-invcov.shrink(X1, lambda=0.6, lambda.var=0)
Specified shrinkage intensity lambda.var (variance vector): 0
Specified shrinkage intensity lambda (correlation matrix): 0.6

> InvS
          [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
[1,]  28.114245    3.436801  -27.14661 -13.985158  -7.930961   53.873371
[2,]   3.436801  157.490485 -138.10588  21.810919 -29.569419  -68.721777
[3,] -27.146606 -138.105883 1814.15756 -27.768920 -93.763979 -113.298379
[4,] -13.985158   21.810919  -27.76892  88.960145   9.426591    2.337493
[5,]  -7.930961  -29.569419  -93.76398   9.426591  96.146200   70.369929
[6,]  53.873371  -68.721777 -113.29838   2.337493  70.369929 2537.069011
attr(,"lambda")
[1] 0.6
```

Ejemplo 2

Cargamos el paquete de R '*SHIP'*. Los creadores de este paquete son Monika Jelizarow and Vincent Guillemo, a los que ya hemos nombrado al estudiar el estimador de covarianzas shrinkage $\hat{\Sigma}_{SHIP}$. Para este ejemplo nos ayudaremos de la información que aparece en la página Web http://cran.r-project.org/web/packages/SHIP/SHIP.pdf.

Vamos a utilizar la matriz de datos del anterior ejercicio que tiene seis variables y cinco muestras.

```
> X2<-matrix(c(4.2,3.9,4.3,4.1,3.8,2,2.1,2,2.2,2,0.6,0.59,0.58,0.63,0.56,4,3.7,3.9,3.8,
+3.8,1.9,2,2,2.1,1.8,1.6,1.59,1.58,1.62,1.63),nrow=5,ncol=6)
```

- Target D

```
# Usamos para la estimación shrinkage la matriz objetivo target D.
> sig1 <- shrink.estim(X1,targetD(X1))
> sig1
[[1]]
             [,1]          [,2]          [,3]          [,4]          [,5]
[,6]
[1,]  0.0430000000 -0.0004485418 4.709689e-04  3.812605e-03  0.0023548443 -
5.718908e-04
[2,] -0.0004485418  0.0080000000 4.149011e-04 -1.233490e-03  0.0017941671
1.009219e-04
[3,]  0.0004709689  0.0004149011 6.700000e-04  3.364063e-05  0.0005270366
3.364063e-06
[4,]  0.0038126050 -0.0012334899 3.364063e-05  1.300000e-02 -0.0006728127 -
1.009219e-04
[5,]  0.0023548443  0.0017941671 5.270366e-04 -6.728127e-04  0.0130000000 -
1.794167e-04
[6,] -0.0005718908  0.0001009219 3.364063e-06 -1.009219e-04 -0.0001794167
4.300000e-04

#Calcula el valor de la intensidad shrinkage para la target D.

[[2]]
[1] "The shrinkage intensity lambda is:" "0.7757"
```

- Target F

```
> sig2 <- shrink.estim(X1,targetF(X1))

> sig2
```

```
[[1]]
              [,1]          [,2]          [,3]          [,4]          [,5]          [,6]
[1,] 4.300000e-02 0.0019180547 1.042447e-03 0.0060018205 0.0048192327
6.510493e-05
[2,] 1.918055e-03 0.0080000000 6.214254e-04 0.0002540472 0.0027101910
3.100642e-04
[3,] 1.042447e-03 0.0006214254 6.700000e-04 0.0003903951 0.0007906556
6.876711e-05
[4,] 6.001820e-03 0.0002540472 3.903951e-04 0.0130000000 0.0010536232
2.090184e-04
[5,] 4.819233e-03 0.0027101910 7.906556e-04 0.0010536232 0.0130000000
1.453406e-04
[6,] 6.510493e-05 0.0003100642 6.876711e-05 0.0002090184 0.0001453406
4.300000e-04

#Calcula el valor de la intensidad shrinkage para la target D.
[[2]]
[1] "The shrinkage intensity lambda is:" "0.8181"
```

4. CONCLUSIONES

Con este estudio se ha pretendido hacer una puesta al día sobre los recientes avances en la estimación de la matriz de covarianzas cuando el número de variables es mayor o comparable al número de observaciones. Por ejemplo, en finanzas o estudios genéticos el número de variables (acciones o genes) suele ser mayor que el número de observaciones.

El uso de la matriz de covarianzas Σ y de su inversa Σ^{-1} es de vital importancia en múltiples usos en la estadística multivariante y por tanto, en muchos campos del conocimiento científico, financiero, etc. Una buena estimación de estas matrices se torna fundamental. Por desgracia, se comprueba que cuando el número de variables es mayor o comparable al número de observaciones, el estimador usado habitualmente, que es la matriz de covarianzas muestral S, deja de ser definido positivo, por lo que no es invertible, y además, tiene un gran error de estimación.

Se han desarrollado a través de estos años, nuevos métodos de estimación que salvasen las desventajas que nos presenta la matriz de covarianzas muestral.
Hemos agrupado estos métodos en dos bloques: las estimaciones de la inversa de la matriz de covarianzas usando una penalización L_1 para el algoritmo Graphical Lasso, y los métodos shrinkage para la estimación de la matriz de covarianzas a través de un estimador shrinkage, tal que este nuevo estimador saldrá de la combinación lineal de la matriz de covarianzas muestral y una matriz objetivo (target matriz), de tal manera que se pueda alcanzar un promedio ponderado que se aproxime lo más posible a la matriz de covarianzas real.

Cada año se desarrollan nuevos métodos relacionados con los diferentes campos científicos donde surge la necesidad de estudios estadísticos de gran dimensión. Por tanto, es un tema que sigue abierto para futuros desarrollos.

ANEXOS

ANEXO 1: Funciones usadas de R

Vamos a desarrollar y explicar los principales paquetes de **R** que hemos empleado en algunos de nuestros ejemplos.

glasso

Este paquete de R *glasso* es usado para implementar el método graphical lasso y por tanto, calcular la estimación de Σ^{-1}.

Los creadores de este paquete son: Jerome Friedman, Trevor Hastie and Rob Tibshirani. Nos ayudaremos de la información que aparece en la página Web http://cran.r-project.org/web/packages/glasso/glasso.pdf.

La función *glasso* estimará la inversa de la matriz de covarianzas con una penalización lasso (L_1), usando la aproximación de Friedman, Hastie y Tibshirani (2007). La aproximación de Meinhausen-Buhlmann (2006) también está implementada. Este algoritmo también puede ser utilizado para estimar una gráfica con ejes ausentes, especificando que ejes omitir en el argumento *zero*, y tomando *rho*=0.

Se define de la siguiente forma:

glasso(s, rho, zero=NULL, thr=1.0e-4, maxit=1e4, approx=FALSE, penalize.diagonal=TRUE, start=c("cold","warm"), w.init=NULL,wi.init=NULL, trace=FALSE)

Los **argumentos** que aparecen son:

- **s** Matriz de covarianzas
- **rho** Parámetro de regulación para lasso.

- **zero** Índice de entradas de la covarianza inversa para ser restringido a cero (opcional).
- **thr** Threshold para convergencia.
- **Maxit** Máximo número de iteracione. 10.000 por defecto.
- **approx** Si TRUE computa la aproximación de Meinhausen-Buhlmann(2006)
- **penalize.diagonal** ¿Será penalizada la diagonal de la inversa? Por defecto TRUE
- **w.init** Tipo de comienzo.
- **wi.init** Valores opcionales para comenzar.
- **trace** Marca para mostrar las iteraciones del procedimiento. Por defecto FALSE

Los **valores** que devuelve son:

- **w** Estimación de la matriz de covarianzas
- **wi** Estimación de la matriz inversa de covarianzas
- **loglik** Valor de la maximización de la log-verosimilitud+penalización
- **errflag** Marcador de error de colocación (*allocation error flag*). Si 0 significa que no hay error
- **approx** Valor del argumento de entrada aproximado
- **del** Cambio en el valor del parámetro de convergencia
- **niter** Número de iteraciones usadas por el algoritmo

corpcor

El paquete de R *corpcor* es usado para estimar la matriz de covarianzas. Nos ayudaremos de la información que aparece en la página Web http://www.rdocumentation.org/packages y http://cran.r-project.org/web/packages/corpcor/corpcor.pdf. El nombre del nombre de este paquete hace referencia a las palabras en inglés ***cor****relations and* ***p****artial* ***cor****relations.*

Este paquete implementa un estimador shrinkage tipo James-Stein para la matriz de covarianzas, con shrinkage separado para varianzas y correlaciones. Esta aproximación es tanto estadísticamente como computacionalmente muy eficiente, y se puede aplicar sin problemas para datos de muchas variables y pocas observaciones, y siempre nos devuelve

una matriz definida positiva y bien condicionada. Este paquete también provee las correlaciones y varianzas parciales, y los coeficientes de regresión. La matriz inversa de la matriz de covarianzas y la matriz de correlaciones también puede ser calculada eficientemente

Para estimar la matriz de covarianzas usamos del paquete de R *corpcor* la función *cov.shrink,* que se define de la siguiente manera:

cov.shrink(x, lambda, lambda.var, w, verbose=TRUE)

Los **argumentos** que aparecen son:

- ***x*** Es la matriz de datos.
- ***lambda*** Intensidad shrinkage de correlación (entre 0 y 1). Si no se especifica se estimará usando la fórmula analítica de Schäfer-Strimmer (2005).
- ***lambda.var*** Intensidad shrinkage de la varianza (entre 0 y 1). Si no se especifica se estimará usando la fórmula analítica de Opgen-Rhein (2007).
- ***w*** Opcional: si queremos especificar los pesos de cada dato.
- ***verbose*** Muestra mensajes de estado mientras se opera (por defecto: TRUE).

ANEXO 2: Test de esfericidad

Los recientes avances en análisis de datos en alta dimensionalidad requieren muchas veces que ciertas premisas sean aceptadas de forma implícita o explícita. Por ejemplo, Dudoit et al. (2002) en su análisis de datos microarrays sobre genes asume que las matrices de covarianzas son matrices diagonales, y entonces su función de distancias utiliza únicamente los elementos de la diagonal de la matriz de covarianzas muestral.

Para determinar estas premisas sobre la matriz de covarianzas, el test de razón de verosimilitudes (LRT, abreviatura en inglés de *likelihood ratio tests*) no puede ser usado cuando el tamaño de la muestra n es más pequeña que la dimensión p .

Recordemos en qué consiste el LRT y por qué no puede ser aplicado cuando $p > n$.Para el test de hipótesis de esfericidad $H_0 : \Sigma = \sigma^2 I$ vs $H_1 : \Sigma \neq \sigma^2 I$, se define el test de razón de verosimilitudes (LRT) de la siguiente manera:

$$\Lambda(x) = \left(\frac{\prod_{i=1}^{p} l_i^{1/p}}{\sum_{i=1}^{p} l_i / p} \right)^{\frac{1}{2}pN},$$

donde $l_1,...,l_p \geq 0$ son los autovalores de el estimador de máxima verosimilitud para Σ. Podemos observar que si $p > n$, la LRT es degenerada, incluso cuando $n > p$ pero $p \to n$, la LRT es computacionalmente degenerada, y por tanto, el test de razón de verosimilitudes (LRT) no se puede usar en alta dimensionalidad.

Un test estadístico no degenerado fue propuesto por John (1971) de la siguiente forma:

$$U = \frac{1}{p} tr\left[\left(\frac{S}{\frac{1}{p} tr(S)} - I \right)^2 \right] \quad \text{y} \quad V = \frac{1}{p} tr\left[(S - I)^2 \right]$$

John (1971) probó que el test basado en U es test invariante local (LBI, abreviatura en inglés de *locally best invariant*) más potente para la esfericidad. Nagao (1973) propone el estadístico V como el equivalente de U para el test de $\Sigma = I$. Sin embargo, estos tests asumen que la muestra n tiende a infinito mientras que el tamaño poblacional p permanece fijo. Por tanto, el uso de este tipo de test es inapropiado cuando $p > n$, es decir, en casos de alta dimensionalidad.

Ledoit-Wolf (2002) proponen el estadístico modificado

$$W = \frac{1}{p} tr\left[(S - I)^2 \right] - \frac{p}{n} \left[\frac{1}{p} tr(S) \right]^2 + \frac{p}{n}$$

W tiene las mismas propiedades n-asintóticas que V. Este test es robusto incluso para un tamaño poblacional p grande. Ledoit-Wolf (2002) desarrollaron este test que podemos considerar LBI y dieron su distribución asintótica cuando $(p/n) \to c$, una constante, pero la distribución no nula no está disponible.

Srivastava (2005) propone un nuevo estimador para salvar estos problemas. Para el test de hipótesis de esfericidad $H_0 : \Sigma = \sigma^2 I$ vs $H_1 : \Sigma \neq \sigma^2 I$, tomando, sin pérdida de generalidad, que $\Sigma = diag(\lambda_1,...,\lambda_p)$, una matriz diagonal $p \times p$. Si $\lambda_i = \lambda$ para alguna constante λ, entonces

$$\gamma_1 \equiv \frac{\left(\sum_{i=1}^{p} \lambda_i^2 / p\right)}{\left(\sum_{i=1}^{p} \lambda_i / p\right)^2} \geq 1$$

Entonces γ_1 es igual a uno si y sólo si $\lambda_i = \lambda$ para alguna constante λ. Entonces podemos considerar nuestro test de hipótesis como $H_0 : \gamma_1 - 1 = 0$ vs $H_0 : \gamma_1 - 1 > 0$

Un estimador consistente del estimador γ_1 está dado por

$$\hat{\gamma}_1 = \frac{n^2}{(n-1)(n+2)} \frac{1}{p}\left[tr\left(S^2\right) - \frac{1}{n}\left(tr(S)\right)^2 \right] \Big/ \left(tr\left(S^2\right)/p\right)^2 = \frac{\hat{a}_2}{\hat{a}_1^2}$$

Y por tanto, un test de esfericidad puede estar basado en el estadístico

$$T_1 = \hat{\gamma}_1 - 1$$

La distribución no nula del estadístico T_1 se puede calcular y será:

$$\left(\frac{n}{2}\right)\left(T_1 - \hat{\gamma}_1 + 1\right) \sim N\left(0, \tau_1^2\right)$$

donde $\tau_1^2 = \frac{2n\left(a_4 a_1^2 - 2a_1 a_2 a_3 + a_2^3\right)}{p a_1^6} + \frac{a_2^2}{a_1^4}$, con $a_i = \left(tr\Sigma^i / p\right)$.

REFERENCIAS

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

BAI, Z., YIN AND KRIHNAIAH (1988). *On the limit of the largest eigenvalue of the large dimensional sample covariance matrix.*

BAI, Z., (1999). *Methodologies in spectral analysis of large dimensional random matrices, a review*. Statistica Sinica 9, 611-677.

BAI, J. AND NG. S., (2011). *Principal components estimation and identification of factors*. Unpublished manuscript, Department of Economics, Columbia University.

BAI, J. AND SHI, S., (2011). *Estimating High Dimensional Covariance Matrices and Its Applications.* Annals of Economics and finance 20011. 12-2, 199-215

BAI, Z., D. AND. YIN Y. Q (1993). *Limit of the smallest eigenvalue of a large dimensional sample covariance matrix*, Ann. Probab. 21

BANERJEEN O., EL GHAOUI LE., D'ASPREMONT A., (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J Mach Learn Res. 2008; 9:485-516

BEERENWINKEL N, ANTAL T, DINGLI D, TRAULSEN A, KINZLER KW, ET AL. (2007) Genetic Progression and the Waiting Time to Cancer. PLoS Comput Biol 3(11): e225. doi:10.1371/journal.pcbi.0030225

BIEN J. Y TIBSHIRANI R.J. (2011). *Sparse estimation of a covariance matriz.* Biometrika. Dec 2011; 98(4): 807–820. Doi: 10.1093/biomet/asr054

BOX, G. E. P., JENKINS, G.M. and REINSEL, G. C. (1994). *Time Series Analysis*: *Forecasting and Control*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ. MR1312604

BUTTE, A., TAMAYO, P., SLONIM, D., GOLUB, T. AND KOHANE, I. (2000). *Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks*, Proceedings of the National Academy of Sciences pp. 12182–12186.

CHAUDHURI, S., DRTON, M. AND RICHARDSON, T. S. (2007). *Estimation of a covariance matrix with zeros*, Biometrika 94(1): 1–18.

CLARKE, B, FOKOU_E, E., ZHANG, H. (2009). *Principles and Theory for Data Mining and Machine Learning*, Springer.

DEMPSTER AP. (1972). *Covariance selection. Bometrics*.1972; 28: 157-75

DOBBIN K AND SIMON R.(2007). *Sample size planning for developing classifiers using high dimensional DNA microarray data*. *Biostatistics* 8:101-117, 2007.

DOBRA, A., C. HANS, B. JONES, J. R. NEVINS, G. YAO, AND M. WEST (2004). *Sparse graphical models for exploring gene expression data. J. Multiv. Anal. 90*, 196–212.

DUDOIT, S., FRIDLYAND, J. AND SPEED, T. P. (2002). *Comparison of discrimination methods for the classification of tremors using gene expression data*, J. Amer. Statist. Assoc., 97, 77–87.

EFRON, B., HASTIE, T., JOHNSTONE, I., TIBSHIRANI, R. (2004). *Least Angle Regression*, Ann. Stat. Vol. 32, No. 2, 407-499.

EISEN M., SPELLMAN P., BROWN P., BOTSTEIN D. (1998). *Cluster analysis and display of genome-wide expression patterns*

FAN, J., LI, R. (2001). *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties*, Journal of the American Statistical Association, Vol. 96, No. 456, 1348-1360.

FRIEDMAN, J. H. (1989). *Regularized discriminant analysis*. *J. Amer. Statist. Assoc. 84*, 165–175.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics. 2007; 9:432-41

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2008b). Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9: 432–441.

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descendent*, Journal of Statistical Software, Vol. 33, Issue 1, 1-22.

GEORGE, E. (2000). The Variable Selection Problem, Journal of the American Statistical Association. Vol. 95, No. 452, 1304-1308.

GUILLEMOT, V, JELIZAROW, M., TENENHAUS, A, BOULESTEIX, A. (2011). *Shrinkage Covariance Estimation Incorporating Prior Biological Knowledge with Applications to High Dimensional Data*. Technical Report Number 107, 2011 Department of Statistics University of Munich.

HAFF, L. R. (1980). "*Empirical Bayes estimation of the multivariate normal covariance matrix," Ann. Statist.*, vol. 8, no. 3, pp. 586–597, 1980.

HASTIE, T., BUJA, A. Y TIBSHIRANI, R. (1995). *Penalized discriminant analysis.* The Annals of Statistic, 1995, Vol 23, nº 1, 73-102.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HOERL, A. E. AND KENNARD, R. (1970). *Ridge regression: biased estimation for nonorthogonal problems*, Technometrics 12: 55–67.

JELIZAROW, M, GUILLEMOT, V, TENENHAUS, A, STRIMMER, K, AND BOULESTEIX. A.L. *Over-optimism in bioinformatics: an illustration. Bioinformatics*, 26:1990–1998, 2010.

JOHN, S. (1971). *Some optimal multivariate tests, Biometrika,* 58, 123–127.

JOHNSON, R. AND WICHERN, D., (2007) *Applied Multivariate Statistical Analysis.* 6ª ed. Pearson International Edition. ISBN: 0-13-514350-0.

JOHNSTONES, I, (2001). *On the distribution of the largest eigenvalue in principal components analysis. The Annals of Statistics* 2001, Vol. 29, No. 2, 295–327

LALOUX, L., P. CIZEAU, J. P. BOUCHAUD, AND M. POTTERS, (1999). *Noise Dressing of Financial Correlation Matrices*. Phys. Rev. Lett 83, 1467.

LALOUX, L., P. CIZEAU, J. P. BOUCHAUD, AND M. POTTERS, (2000). *Random matrix theory and financial correlations.* International Journal of Theoretical and Applied Finance 3, 391-397.

LEDOIT, O. AND WOLF, M. (2002), "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size", Ann. Statist., Vol. 30, pp1081-1102

LEDOIT, O., AND WOLF, M., (2003). *Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection.* Journal of Empirical Finance, 10, 603-621.

LEDOIT, O., AND WOLF, M., (2004a). *A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices.* Journal of Multivariate Analysis, 88, 365-411.

LEDOIT, O., AND WOLF, M., (2004b). *Honey, I Shrunk the Sample Covariance Matrix.* Journal of Portfolio Management, Summer, 110-119.

LEUNG, P. L. AND W. Y. CHAN (1998). *Estimation of the scale matrix and its eigenvalues in the Wishart and the multivariate F distributions. Ann. Inst. Statist. Math. 50*, 523–530

MARDIA, K., KENT, J. AND BIBBY, J. (1979). Multivariate Analysis, Academia Press.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). *Highdimensional graphs and variable selection with the lasso*. *Ann*. *Statist*. **34** 1436–1462. MR2278363

MEINSHAUSEN, N. (2006). *Relaxed Lasso, Computational Statistics and Data Analysis.* Vol. 52, Issue 1, 374-393.

NAGAO, H. (1973). On some test criteria for covariance matrix, Ann. Math. Statist, 1, 700–709.

PLEROU, V., P. GOPIKRISHNAN, B. ROSENOW, L. A. NUNES AMARAL, AND H. E. STANLEY, (1999). *Universal and Non-universal Properties of Cross Correlations in Financial Time Series*. Phys. Rev. Lett. 83, 1471.

PREISENDOFER R. (1988). Principal component analysis in meteorology and oceanography. Elsevier, Amsterdam, 436 pp.

SCHÄFER, J., AND STRIMMER, K., (2005). *A Shrinkage Approach to Large-Scale Covariance Matrix*
Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology, 4(1), Article 32.

SHARPE, W. F. (1963). *A simplified model for portfolio analysis*. Management Science, 9(1):277{293.

SHEATHER, J. (2009). *A Modern Approach to Regression with R,* Springer

SHUMWAY, R.H. & STOFFER, D.S. (2010). *Time Series Analysis and Its Applications: With R Examples*. 3rd ed., New York: Springer. December, 2010.

SRIVASTAVA, M. (2005), "*Some tests concerning the covariance matrix in high dimensional data*", J. Japan Statist. Soc., Vol.35, pp251-272
STEIN, C., (1975) "*Estimation of a covariance matrix,"* in Rietz Lecture, 39th Ann. Meet., Atlanta, GA, 1975.

TIBSHIRANI, R. (1996). *Regression Shrinkage and Selection via the Lasso*, J. R. Statist. Soc., Serie B., Vol. 58, No. 1, 267-288.

TIBSHIRANI, R. (2011). *Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology*). *Volume 73, Issue 3, pages 273–282, June 2011*

WONG, F., CARTER, C. K. and KOHN, R. (2003). *Efficient estimation of covariance selection models. Biometrika* **90** 809–830. MR2024759

YANG, R. AND. BERGER J. O (1994), "*Estimation of a covariance matrix using the reference prior," Ann. Statist.*, vol. 22, pp. 1195–1211, 1994.

YAO ET AL., (2008). *A bistable Rb-E2F switch underlies the restriction point*.

YUAN, M. and LIN, Y. (2007). *Model selection and estimation in the Gaussian graphical model. Biometrika* **94** 19–35. MR2367824

ZOU, H., HASTIE, T. (2005). *Regularization and variable selection via the elastic net*,J. R. Statist. Soc., Serie B, Vol. 67, Part 2, 301-320.

ZOU, H. (2006). *The adaptive Lasso and its oracles properties*, J. Am. Statist., Vol. 101, 1418-1429.