Universidad Nacional del Sur

Tesis de Magister en Matemática

Métodos de Proyecciones Alternas para Problemas de Optimización Persimétricos.

María Gabriela Eberle

Bahía Blanca Argentina

1999

A Liberato, mi abuelo,
y a mis padres Susana y Pedro.

Contenido

Prefacio

Esta Tesis es presentada como parte de los requisitos para optar al grado académico de Magister en Matemática de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otras. La misma contiene resultados obtenidos en estudios e investigaciones llevadas a cabo en el Departamento de Matemática de la Universidad Nacional de Sur durante el período comprendido entre los meses de septiembre de 1995 y junio de 1999, bajo la dirección de la Dra. María Cristina Maciel, Profesor Asociado del Departamento de Matemática de la Universidad Nacional del Sur.

Agradecimientos

Quiero expresar mi más sincero agradecimiento a la Dra. María Cristina Maciel, por su confianza y paciencia, a la Mg. Flavia Buffo, por su permanente apoyo y a un amigo de siempre, el Lic. Héctor Brunet.

14 de julio de 1999

Departamento de Matemática.

Universidad Nacional del Sur.

Resumen

En este trabajo serán resueltos los problemas persimétrico y Toeplitz de Procrusto, y sus soluciones serán aplicadas a problemas concretos. Se presenta una nueva estrategia que emplea métodos de proyecciones alternas para resolver el problema de aproximar la inversa de una matriz persimétrica. Se sugiere utilizar dicha aproximación en un algoritmo para resolver cierta clase de problemas de programación cuadrática, utilizando la técnica de gradiente reducido.

En el capítulo 1 se hace una introducción, explicando los objetivos y la forma en que ha sido organizado el trabajo. En el capítulo 2 presentamos brevemente tres trabajos que consideramos fundamentales pues han servido de guía y base para nuestros desarrollos, ellos son el problema simétrico de Procrusto, el método de proyecciones alternas de Dykstra y los algoritmos de Escalante-Raydán. El problema persimétrico de Procrusto es presentado en el capítulo 3, se caracterizan las soluciones y se discute la experiencia numérica. El capítulo 4 está dedicado a resolver el problema Toeplitz de Procrusto, y el capítulo 5 a las aplicaciones y conclusiones.

Capítulo 1

Introducción

Consideremos el siguiente problema de minimización:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{P}, \end{array} \right. \tag{1.1}$$

donde $A,\ \ B \in \mathbb{R}^{m\times n}$, con $m > n$,$\mathcal{P}$ es un subconjunto de $\mathbb{R}^{n\times n}$, y $\|.\|_{\mathtt{F}}$ denota la norma de Frobenius. Si $Y \in \mathbb{R}^{m\times n}$, entonces

$$\|Y\|_{\mathtt{F}} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} {Y_{ij}}^2 \right)^{\frac{1}{2}}.$$

A esta clase de problemas se los denomina habitualmente problemas de Procrusto [1]. Si $\mathcal{P}$ es el subespacio de matrices simétricas, al cual llamaremos $\mathcal{S}$, dicho problema es conocido como el problema simétrico de Procrusto, y es el que ha dado nombre a toda esta clase de problemas, pues aparece muy frecuentemente en problemas de elasticidad, siendo precisamente su solución la matriz de deformación de una estructura elástica.

Es claro que si en (1.1) se varía el conjunto de factibilidad el problema cambia y por consiguiente la estrategia para resolverlo. Si en lugar de considerar un subespacio se considera una intersección de subespacios ó de conjuntos convexos, el método por elección es el método de proyecciones alternas de Dykstra, que obtiene la solución sobre la intersección proyectando de manera alternada en cada uno de los convexos involucrados, resolviendo en cada caso un problema de minimización.

Por ejemplo, Escalante-Raydán [5] y [6], aplican el método de proyecciones alternas para resolver problemas de cuadrados mínimos sobre una intersección de convexos, logrando algoritmos sencillos, óptimos y de bajo costo.

[1] Procrusto fue un personaje mitológico de Ática, recordado por la historia griega como un cruel bandido, que luego de asaltar a sus víctimas las sometía a la tortura de colocarlos en un lecho de hierro para estirarlos, si su estatura era menor que la longitud del mismo, ó mutilarlos en caso contrario.

Todo lo mencionado hasta aquí motiva nuestra inquietud por saber cómo serán las soluciones del problema de Procrusto para distintas regiones de factibilidad, e investigar sus posibles aplicaciones. En particular nos interesa el caso en que el conjunto factible es el subespacio de matrices cuadradas persimétricas (son aquellas que son simétricas respecto a la diagonal NE-SO), el subespacio $\mathcal{T}$ de matrices de Toeplitz (matrices constantes por diagonales), y el de las matrices simétricas con respecto a ambas diagonales. En particular nos planteamos resolver los sguientes problemas:
problema persimétrico de Procrusto

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X = EX^T E, \end{array} \right. \tag{1.2}$$

problema Toeplitz de Procrusto

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T}, \end{array} \right. \tag{1.3}$$

problema doblemente simétrico,

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T} \cap \mathcal{S}. \end{array} \right. \tag{1.4}$$

La idea es, por un lado caracterizar las soluciones de cada uno de estos problemas, estudiar las propiedades del conjunto de soluciones y analizar los problemas que aparecen al intersectar los convexos estudiados con otros, aplicando proyecciones alternas. Por otra parte, es también nuestro objetivo presentar aplicaciones de las nuevas técnicas desarrolladas.

Capítulo 2

Fundamentos

Este capítulo incluye una reseña de importantes trabajos que han servido de inspiración para esta contribución: el método de proyecciones alternas de Dykstra [4], la resolución del Problema simétrico de Procrusto [8], y distintas versiones del algoritmo de proyecciones alternas de Escalante-Raydán [5], [6].

El método de proyecciones alternas encuentra su origen en trabajos de Von Neumann [10], que datan de 1932. Von Neumann aborda el problema de hallar la proyección de un punto dado en el espacio de Hilbert, sobre la intersección de dos subespacios del espacio de Hilbert. En 1959 Cheney y Goldstein [3], extienden el análisis de Von Neumann al caso en que la intersección sea de dos conjuntos cerrados convexos. Tres años más tarde, Halperín resuelve el problema de proyectar un punto sobre la intersección de un número finito (eventualmente mayor que dos) de subespacios cerrados de Hilbert, recordemos que hasta ese momento se trabajaba con intersecciones de dos subespacios.

Dykstra [4] en 1983 diseña un método que resuelve el problema de proyectar un punto dado de $\mathbb{R}^n$ sobre la intersección de un número finito de conos cerrados convexos de $\mathbb{R}^n$, y en 1986, junto a Boyle [1], generalizan el método para el caso en que la intersección sea de conjuntos cerrados convexos, no necesariamente conos.

El método de proyecciones alternas de Dykstra y el problema simétrico de Procrusto son aplicados por Escalante y Raydán, quienes desarrollan nuevos algoritmos que resuelven problemas de cuadrados mínimos con restricciones.

2.1 Método de proyecciones alternas de Dykstra

El método de proyecciones alternas de Dykstra resuelve el problema de minimizar la distancia desde un punto dado de $\mathbb{R}^n$ a la intersección de un número finito de conos cerrados convexos.

Consideremos el problema:

$$\left\{\begin{array}{ll} \min & ||g-x|| \\ s.a & \\ & x \in \bigcap K_i, \end{array}\right. \tag{2.1}$$

donde $K_1, K_2, \cdots, K_r$, son conos cerrados convexos de $\mathbb{R}^n$, y la norma es la asociada al producto interno:

$$(x,y) = \sum_{i=1}^{n} x_i y_i,$$

es decir:

$$||x|| = (x,x)^{1/2} = \left(\sum_{i=1}^{n} {x_i}^2\right)^{1/2}.$$

Debe notarse que si K_i, $i = 1, \cdots, r$ es un cono cerrado convexo, entonces la intersección de todos ellos también lo es.

Asumiendo que es posible resolver el problema de hallar un vector en K_i tal que sea solución de

$$\left\{\begin{array}{ll} \min & ||f-x|| \\ s.a & \\ & x \in K_i, \end{array}\right. \tag{2.2}$$

para toda f y para todo i, la solución del problema (2.1) se obtiene resolviendo una sucesión de problemas de tipo (2.2).

Un vector $g^\star \in K_i$ es solución de (2.2) si:

$$(g - g^\star, f - g^\star) \leq 0, \quad \forall \ \ f \in K_i,$$

es decir,

$$\text{si } \theta = ang(g - g^\star, f - g^\star), \text{ entonces para todo } f \in K_i,\ \theta \geq \pi/2.$$

El método consiste de un simple procedimiento iterativo, que minimiza en cada K_i, imponiendo a ellos como únicas restricciones que r sea finito y K_i un cono cerrado convexo para cada i. Si $P_B(A)$ denota la proyección de A sobre el conjunto B, el siguiente algoritmo describe el método de proyecciones alternas.

Algoritmo 2.1.1 *(Proyecciones Alternas)*

Dado $X \in \mathbb{R}^n$, $I_i = 0, \ \forall\, i,$
repetir para $i = 1, 2, \cdots, r$

$Y \leftarrow X - I_i$

$X \leftarrow P_{K_i}(Y)$

$I_i \leftarrow X - Y$

El siguiente teorema garantiza la convergencia del algoritmo anterior.

Teorema 2.1.1 *Los vectores generados por el Algoritmo 2.1.1, convergen a una solución de (2.1), es decir* $X_i \to X^\star$ *cuando* $i \to \infty$.

La demostración puede ser consultada en [4].

2.2 Problema simétrico de Procrusto

Interesa resolver:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X^T = X, \end{array} \right. \tag{2.3}$$

A y $B \in \mathbb{R}^{m\times n}$, $m > n$. La norma utilizada es la norma de Frobenius.

Se considera la descomposición en valores singulares de la matriz A:

$$A = P \left[\begin{array}{c} \sum \\ 0 \end{array} \right] Q^T,$$

donde $P \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$, son matrices ortogonales, y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

Recordemos que dada la matriz A, si P y Q son matrices ortogonales de dimensiones adecuadas, se verifica que:

$$\|QAP\|_{\mathtt{F}} = \|A\|_{\mathtt{F}},$$

por ello, resolver (2.3) es equivalente a encontrar una solución para:

$$\left\{ \begin{array}{ll} \min & \|\sum Y - C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & Y = Y^T, \end{array} \right. \tag{2.4}$$

siendo

$$Y = Q^T X Q,$$

$$C = \left[\begin{array}{c} C_1 \\ C_2 \end{array} \right] = P^T B Q,$$

$$C_1 \in \mathbb{R}^{n\times n}.$$

La solución de (2.4):

$$(Y_*)_{ij} = \left\{ \begin{array}{ll} \dfrac{\sigma_i (C_1)_{ij} + \sigma_j (C_1)_{ji}}{{\sigma_i}^2 + {\sigma_j}^2} & si \quad {\sigma_i}^2 + {\sigma_j}^2 \neq 0 \\ arbitrario & en \quad otro \quad caso. \end{array} \right.$$

y por lo tanto la solución para (2.3) está dada por $X_* = QY_*Q^T$.

Luego de hallar la expresión de la solución el autor caracteriza el conjunto de soluciones y demuestra además un teorema relacionado a la estabilidad de su estrategia. Ambos resultados serán prsentados a continuación y serán de interés cuando se estudie el caso persimétrico.

2.2.1 Caracterización de las soluciones

El teorema siguiente da una caracterización para el conjunto de soluciones del problema (2.3)

$$\mathcal{S} = \left\{ X \in \mathbb{R}^{n\times n} \ : \ X = X^T, \quad \|AX - B\|_{\mathrm{F}}^2 = \min_{Z=Z^T} \|AZ - B\|_{\mathrm{F}}^2 \right\}.$$

Lema 2.2.1

1) $X \in \mathcal{S} \Longleftrightarrow X = X^T$ *y* $A^TAX + XA^TA = A^TB + B^TA$.

2) $\mathcal{S}$ *es convexo.*

3) $\mathcal{S}$ *tiene un único elemento* ${X_\star}^{mn}$ *de norma de Fobenius mínima.*

4) $\mathcal{S} = \{{X_\star}^{mn}\} \Longleftrightarrow Rank(A) = n$.

Observación:

La igualdad que aparece en el inciso 1) del lema anterior,

$$A^TAX + XA^TA = A^TB + B^TA, \tag{2.5}$$

recibe la denominación "*ecuaciones normales*", y de esa forma nos referiremos a ella en adelante.

Teorema 2.2.1 *Sea* $A \in \mathbb{R}^{m\times n}$, $m \geq n$. *Sea* X *solución de (2.3). Sea* $\hat{X}$ *tal que resuelve el problema perturbado:*

$$\begin{cases} \min & \|(A+\delta A)X - (B+\delta B)\|_{\mathrm{F}}^2 \\ s.a & \\ & X = X^T, \end{cases} \tag{2.6}$$

donde δA *y* $\delta B \in \mathbb{R}^{m\times n}$.

Sean:

$\mathcal{R}$=$AX - B$,

$\hat{\mathcal{R}}$=$(A+\delta A)X - (B+\delta B)$,

$$\epsilon_A = \frac{||\delta A||_2}{||A||_2},$$

$$\kappa = \begin{cases} \kappa_2(A) & si \quad rank(A) = n \\ \sqrt{2\kappa_2(A)} & \quad\quad rank(A) < n. \end{cases}$$

Suponiendo que $rank(A) = rank(A + \delta A) = n$, *y* $\kappa\epsilon_A < 1$, *entonces*

$$\|X - \hat{X}\|_{\mathtt{F}} \leq \frac{\kappa}{1 - \epsilon_A} \left(\epsilon_A \|X\|_{\mathtt{F}} + \frac{\|\delta B\|_{\mathtt{F}}}{||A||_2} + \kappa\epsilon_A \frac{\|\mathcal{R}\|_{\mathtt{F}}}{||A||_2} \right) + \kappa\epsilon_A ||\|X\|_{\mathtt{F}}, \tag{2.7}$$

$$\|\mathcal{R} - \hat{\mathcal{R}}\|_{\mathtt{F}} \leq \epsilon_A \|X\|_{\mathtt{F}} \, ||A||_2 + \|\delta B\|_{\mathtt{F}} + \kappa \, \epsilon_A \|\mathcal{R}\|_{\mathtt{F}}. \tag{2.8}$$

Las demostraciones de ambos resultados pueden ser encontradas en [8].

Si $rank(A) = n$, el autor encuentra un expresión para el residuo. Si X_{sp} es la solución de (2.3), entonces:

$$\rho_{ps} = \|AS_\star - B\|_{\mathtt{F}}^2 = \sum_{j>i} \frac{(\sigma_i (C_1)_{ji} - \sigma_j (C_1)_{ij})^2}{{\sigma_i}^2 + {\sigma_j}^2} + \|C_2\|_{\mathtt{F}}^2,$$

siendo:

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = P^T B Q \in \mathbb{R}^{m \times n},$$

P, Q y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, las matrices involucradas en la descomposición en valores singulares de la matriz A.

Debe notarse el hecho que, para el problema de cuadrados mínimos standard, el residuo es exactamente igual a $\|Z_2\|_{\mathtt{F}}^2$, y que si la matriz solución del problema standard, A^+B es simétrica, siendo A^+ la matriz pseudoinversa de A, entonces la sumatoria que aparece en la expresión de ρ_{ps} es cero.

2.3 Algoritmos de Escalante-Raydán

Los algoritmos propuestos por Ecalante- Raydán resuelven el problema de hallar una matriz cuadrada, simétrica, definida positiva que, satisfaciendo restricciones de caja, minimice la norma de Frobenius de $AX - B$. Primero se considerará el caso en el cual $A = I$, y luego el caso general.

2.3.1 Caso cuadrado

Se considera el problema de cuadrados mínimos con restricciones:

$$\left\{\begin{array}{ll} \min & \|X-A\|_{\mathtt{F}}^2 \\ s.a & \\ & L \leq X \leq U, \\ & \\ & X^T = X, \\ & \\ & \lambda_{min}(X) \geq \epsilon > 0, \\ & \\ & X \in \mathcal{P}. \end{array}\right. \tag{2.9}$$

donde

$A, L, U \in \mathbb{R}^{n\times n}$, X matriz cuadrada simétrica.

$\lambda_{min}(X)$ es el menor autovalor de X, y ϵ es una constante dada.

$\mathcal{P} \subseteq \mathbb{R}^{n\times n}$, matrices que tienen una estructura especial.

La notación $A \leq B$, donde A y B son dos matrices significa que $A_{ij} \leq B_{ij}$, para todo i, j con $1 \leq i, j \leq n$. La norma utilizada es la norma de Frobenius:

$$||A||_{\mathtt{F}}^2 = (A, A) = \sum_{ij=1}^{m} (A_{ij})^2,$$

con el producto interno:

$$(A, B) = Traza(A^T B).$$

La región de factibilidad del problema está dada por la intersección de los siguientes conjuntos:

$$\mathcal{B} = \{X \in \mathbb{R}^{n\times n} \ : \ L \leq X \leq U\},$$

$$\epsilon pd = \{X \in \mathbb{R}^{n\times n} \ : \ X^T = X, \ \lambda_{min}(X) \geq \epsilon > 0\},$$

$$\mathcal{P} = \{X \in \mathbb{R}^{n\times n} \ : \ X = \sum_{i=1}^{m} \alpha_i G_i, \ \alpha_i \in \mathbb{R}, 1 \leq i \leq m\}.$$

Es preciso aclarar que en la definición de $\mathcal{P}$ se considera que

$$1 \leq m \leq \frac{n(n+1)}{2},$$

aunque en la práctica en general se tiene $m \leq n$.

Se consideran $G_1, G_2, \cdots, G_m$, matrices no nulas, simétricas, en $\mathbb{R}^{n\times n}$, cuyas entradas valen cero ó uno de acuerdo a la siguiente propiedad:

para cada entrada st, $1 \leq s,t \leq n$ existe un y sólo un k tal que $1 \leq k \leq m$, y $(G_k)_{s,t} = 1$.

De esta forma (2.9) puede ser expresado como sigue:

$$min\{||X - A||_{\mathtt{F}}^2 \ : \ X \in \mathcal{B} \cap \epsilon pd \cap \mathcal{P}\}. \tag{2.10}$$

La región de factibilidad para el problema (2.10) está dada por la intersección de conjuntos cerrados y convexos en el espacio producto interno $\mathbb{R}^{n\times n}$. Debe notarse que $\mathcal{P}$ es un subespacio del subespacio de matrices simétricas y que el conjunto de matrices $\{G_1, G_2, \cdots, G_m\}$ constituye una base para $\mathcal{P}$.

El problema (2.10) se resuelve por medio del método de proyecciones alternas de Dykstra. La idea fundamental es proyectar sobre cada uno de los subconjuntos cerrados y convexos que forman la región de factibilidad, por lo cual es necesario caracterizar las proyecciones sobre los mismos. Los tres teoremas siguientes, cuyas demostraciones pueden consultarse en [5], cumplen tal objetivo.

Teorema 2.3.1 *Si $A \in \mathbb{R}^{n\times n}$, entonces la única solución para el problema*

$$\begin{cases} \min & \|X - A\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{B}, \end{cases}$$

está dada por la matriz cuadrada $P_{\mathcal{B}}(A)$ cuya entrada i,j para todo $1 \leq i,j \leq n$ se define como sigue:

$$(P_{\mathcal{B}}(A))_{ij} = \begin{cases} A_{ij} & si \quad L_{ij} \leq A_{ij} \leq U_{ij}, \\ U_{ij} & si \quad A_{ij} > U_{ij}, \\ L_{ij} & si \quad A_{ij} < L_{ij}. \end{cases}$$

Teorema 2.3.2 *Si $A \in \mathbb{R}^{n\times n}$, entonces la única solución para el problema*

$$\begin{cases} \min & \|X - A\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{P}, \end{cases}$$

está dada por

$$P_{\mathcal{P}}(A) = \sum_{k=1}^{m} \bar{\alpha_k} G_k,$$

con los coeficientes

$$\bar{\alpha_k} = \frac{\sum_{ij=1}^{n} A_{ij}(G_k)_{ij}}{\sum_{ij=1}^{n}(G_k)_{ij}}, \qquad \forall k \ : \ 1 \leq k \leq m.$$

Teorema 2.3.3 *Sea $A \in \mathbb{R}^{n\times n}$ y sean B y C las partes simétrica y antisimétrica de A respectivamente:*

$$B = \frac{(A + A^T)}{2}, \qquad C = \frac{(A - A^T)}{2}.$$

Entonces la solución para el problema

$$\begin{cases} \min & \|X - A\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \epsilon pd, \end{cases}$$

está dada por

$$P_{\epsilon pd}(A) = Z diag(d_i) Z^T,$$

donde

$$d_i = \begin{cases} \lambda_i(B) & si \quad \lambda_i(B) \geq \epsilon, \\ \epsilon & si \quad \lambda_i(B) < \epsilon \end{cases}$$

y la matriz Z es tal que $B = ZDZ^T$ es la descomposición espectral de B es decir:

- $Z^T Z = I$,
- $D = diag(\lambda_i(B))$.

Además:

$$\min_{X\in\epsilon pd} \|X - A\|_{\mathtt{F}}^2 = \sum_{\lambda_i(B)<\epsilon} (\lambda_i(B) - \epsilon)^2 + \|C\|_{\mathtt{F}}{}^2.$$

Corolario 2.3.1 *Si $A \in \mathbb{R}^{n\times n}$ es simétrica, entonces:*

$$P_{\epsilon pd}(A) = Z diag(d_i) Z^T,$$

donde

$$d_i = \begin{cases} \lambda_i(A) & si \quad \lambda_i(A) \geq \epsilon, \\ \epsilon & si \quad \lambda_i(A) < \epsilon \end{cases}$$

y la matriz Z es tal que $A = ZDZ^T$ es la descomposición espectral de A es decir:

- $Z^T Z = I$,
- $D = diag(\lambda_i(A))$.

Algoritmo 2.3.1 *(Algoritmo de Escalente-Raydán)*

Dada $A \in \mathbb{R}^{n\times n}$, sea $A_0 = A$, $I^0_{\epsilon pd} = I^0_{\mathcal{B}} = 0$,
para $i = 1, 2, \cdots$, repetir

$A_i = P_{\mathcal{P}}(A_i) - I^i_{\epsilon pd}$

$$I_{\epsilon pd}^{i+1} = P_{\epsilon pd}(A_i) - A_i,$$

$$A_i = P_{\epsilon pd}(A_i) - I_{\mathcal{B}}^i,$$

$$I_{\mathcal{B}}^{i+1} = P_{\mathcal{B}}(A_i) - A_i,$$

$$A_{i+1} = P_{\mathcal{B}}(A_i).$$

Observaciones:

- La proyección sobre el conjunto ϵpd es realizada en cada iteración, inmediatamente después de realizar la proyección sobre $\mathcal{P}$, de modo de proyectar una matriz simétrica y aplicar el corolario.
- Los incrementos asociados a ϵpd y $\mathcal{B}$ son considerados pues se debe garantizar que la sucesión generada por el algoritmo (2.3.1) converja a una solución de (2.10). Dado que $\mathcal{P}$ es un subespacio, los incrementos asociados no son necesarios.
- En la práctica, el algoritmo es detenido cuando dos proyecciones consecutivas sobre un mismo espacio son muy cercanas, es decir si la diferencia entre ambas alcanza cierta tolerancia prefijada.

El siguiente resultado constituye el teorema de convergencia para el algoritmo 2.3.1.

Teorema 2.3.4 *Si el conjunto cerrado y convexo $\mathcal{B} \cap \epsilon pd \cap \mathcal{P}$, es no vacío, entonces para alguna matriz $A \in \mathbb{R}^{n \times n}$ la sucesiones $\{P_{\mathcal{P}}(A_i)\}$, $\{P_{\mathcal{B}}(A_i)\}$, y $\{P_{\epsilon pd}(A_i)\}$ generadas por el algoritmo 2.3.1 convergen en norma de Frobenius a una solución de (2.10).*

La demostración se sigue de un resultado de convergencia establecido por Boyle y Dykstra [1].

2.3.2 Versión mejorada del algoritmo

Del análisis del costo asociado a cada proyección se advierte que llevar a cabo las proyecciones sobre ϵpd constituye la instancia más costosa, pues supone el cálculo de todos los autovalores y autovectores de una matriz simétrica. Si se piensa en $\mathcal{B}$, sólo se deben efectuar comparaciones entre elementos de A y los correspondientes de L y U. Finalmente la proyección sobre $\mathcal{P}$ sólo implica efectuar el cálculo de medias aritméticas entre la entradas de A. Luego es claro que las proyecciones más costosas son las que se realizan sobre ϵpd. Precisamente con el objeto de reducir el costo computacional es que Escalante y Raydán presentan una nueva versión del algoritmo, la cual proyecta sobre ϵpd y sobre la intersección $\mathcal{B} \cap \mathcal{P}$.

Algoritmo 2.3.2 *(Versión Mejorada)*

Dada $A \in \mathbb{R}^{n\times n}$, *sea* $A_0 = A$, $I^0_{\mathcal{B}\cap\mathcal{P}} = I^0_{\epsilon pd} = 0$,
para $i = 1, 2, \cdots$, *repetir*

$A_i = P_{\mathcal{B}\cap\mathcal{P}}(A_i) - I^i_{\epsilon pd}$

$I^{i+1}_{\epsilon pd} = P_{\epsilon pd}(A_i) - A_i,$

$A_{i+1} = P_{\epsilon pd}(A_i) - I^i_{\mathcal{B}\cap\mathcal{P}},$

$I^{i+1}_{\mathcal{B}\cap\mathcal{P}} = P_{\mathcal{B}\cap\mathcal{P}}(A_{i+1}) - A_{i+1}.$

En el algoritmo anterior, se supone que la proyección $P_{\mathcal{B}\cap\mathcal{P}}(A_i)$ es la única solución del problema:

$$\begin{cases} \min & \|X - A\|^2_{\mathbf{F}} \\ s.a & \\ & X \in \mathcal{B}\cap\mathcal{P}. \end{cases}$$

Para caracterizar dicha solución se harán las siguientes consideraciones:

Sea $\bar{\mathcal{B}} \subseteq \mathcal{B}$:

$$\bar{\mathcal{B}} = \left\{ X \in \mathbb{R}^{n\times n} \quad : \quad \bar{L} \leq X \leq \bar{U} \right\},$$

donde $\bar{L}$, y $\bar{U}$ se obtienen como sigue:

Para $k = 1, 2, \cdots, m$, si $(G_k)_{ij} = 1$ entonces:

$$\bar{L}_{ij} = max_{(G_k)_{rs}=1}\ L_{rs}, \qquad \bar{U}_{ij} = min_{(G_k)_{rs}=1}\ U_{rs}.$$

El siguiente teorema permite caracterizar las proyecciones sobre $\mathcal{B}\cap\mathcal{P}$.

Teorema 2.3.5 *Si* $\mathcal{B}\cap\mathcal{P}$ *es no vacío entonces en la* $i-$*ésima iteración del algoritmo 2.3.2, la proyección de* A_i *sobre* $\mathcal{B}\cap\mathcal{P}$ *está dada por:*

$$P_{\mathcal{B}\cap\mathcal{P}} = P_{\bar{\mathcal{B}}}(P_{\mathcal{P}}(A_i)).$$

La demostración puede consultarse en [5].

Finalmente, en cuanto a la experiencia computacional presentada por los autores, son claros los beneficios de aplicar el algoritmo en su versión mejorada, el cual no sólo se impone al algoritmo 2.3.1 en cuanto a tiempo, sino que es significativa la diferencia en número de iteraciones que resulta de aplicar uno u otro. Debe destacarse lo ingenioso, simple y práctico de este método.

2.3.3 Caso rectangular

Considérese el siguiente problema:

$$\left\{\begin{array}{ll} \min & \|AX-B\|_{\mathtt{F}}^2 \\ s.a & \\ & L \leq X \leq U, \\ & \\ & X^T = X, \\ & \\ & \lambda_{min}(X) \geq \epsilon > 0. \end{array}\right. \tag{2.11}$$

donde

A y B son matrices $m \times n$ dadas, $m \geq n$, $rank(A) = n$,

$L, U, X \in \mathbb{R}^{n\times n}$, X matriz cuadrada simétrica,

$\lambda_{min}(X)$ es el menor autovalor de X, y ϵ es una constante dada.

Para la resolución se aplica descomposición en valores singulares, a la matriz A, con el objeto de transformar el problema original en uno más simple, que sólo involucre matrices cuadradas, para aplicar la técnica desarrollada en el caso anterior.

La región de factibilidad del problema está dada por la intersección de los conjuntos:

$$\mathcal{B} = \{X \in \mathbb{R}^{n\times n} \ : \ L \leq X \leq U\},$$
$$\epsilon pd = \{X \in \mathbb{R}^{n\times n} \ : \ X^T = X, \ \lambda_{min}(X) \geq \epsilon > 0\}.$$

Así (2.11) puede ser expresado como sigue:

$$min\{\|AX-B\|_{\mathtt{F}}^2 \ : \ X \in \mathcal{B} \cap \epsilon pd \cap \mathcal{P}\}. \tag{2.12}$$

Aplicando descomposición en valores singulares, se tiene que resolver el problema (2.12) es equivalente a resolver:

$$\left\{\begin{array}{ll} \min & \|Z-C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & \mathcal{B}' \cap \epsilon pd', \end{array}\right. \tag{2.13}$$

siendo $Z = \sum Y \in \mathbb{R}^{n\times n}$, $Y = Q^T X Q$, y la región de factibilidad la intersección de los dos conjuntos convexos siguientes:

$$\mathcal{B}' = \left\{Z \in \mathbb{R}^{n\times n} \ : \ L \leq Q\Sigma^{-1} Z Q^T \leq U\right\},$$

$$\epsilon pd' = \left\{Z \in \mathbb{R}^{n\times n} \ : \ Z = \sum Y, Y = Y^T, \ \lambda_{min}(Z) \geq 0\right\},$$

donde y es hallada teniendo en cuenta (2.2). Se presenta a continuación la versión del algoritmo de proyecciones alternas que resuelve este caso.

Algoritmo 2.3.3 *(Caso Rectangular)*

Dada $C_1 \in \mathbb{R}^{n\times n}$, *sea* $(C_1)_0 = C_1$, $I^0_{\epsilon pd'} = I^0_{\mathcal{B}'} = 0$,
para $i = 1, 2, \cdots$, *repetir*

$(C_1)_i = P_{\mathcal{B}'}(C_1)_i - I^i_{\epsilon pd'}$

$I^{i+1}_{\epsilon pd'} = P_{\epsilon pd'}(C_1)_i - (C_1)_i,$

$(C_1)_{i+1} = P_{\epsilon pd'}(C_1)_i - I^i_{\mathcal{B}'},$

$I^{i+1}_{\mathcal{B}'} = P_{\mathcal{B}'}(C_1)_{i+1} - (C_1)_{i+1}.$

El siguiente teorema garantiza la convergencia del algoritmo.

Teorema 2.3.6 *Si el conjunto cerrado y convexo* $\mathcal{B}' \cap \epsilon pd'$ *es no vacío entonces, para todo* $C_1 \in \mathbb{R}^{n\times n}$, *las sucesiones de matrices generadas por el algoritmo 2.3.3,* $\{P_{\mathcal{B}'}(C_1)_i\}$ *y* $\left\{P_{\epsilon pd'}(C_1)_i\right\}$, *convergen en norma de Frobenius a una solución del problema (2.13).*

Luego resulta claro que volviendo hacia atrás es posible hallar una solución para el problema (2.9).

Capítulo 3

Problema persimétrico de Procrusto

En este capítulo se considera la clase de todos los problemas de cuadrados mínimos con restricciones del tipo:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{P}, \end{array} \right. \tag{3.1}$$

donde A y B son matrices rectangulares, pertenecientes a $\mathbb{R}^{m\times n}$, con $m > n$; $\mathcal{P}$ es un subconjunto de $\mathbb{R}^{n\times n}$. Nos ocupa ahora el problema que se obtiene si $\mathcal{P}$ es el conjunto de matrices persimétricas, se desea resolver el *problema persimétrico de Procrusto*.

Recordemos la definición de matriz persimétrica:
Definición (según Golub y Van Loan [7]): $B \in \mathbb{R}^{n\times n}$ se dice persimétrica cuando es simétrica respecto a la diagonal NE-SO, es decir, si $B_{ij} = B_{n-j+1,n-i+1}$ para todo i, j. Esto es equivalente a requerir que

$$B = EB^TE,$$

donde E es la matriz que tiene todos los elementos nulos excepto sobre la diagonal NE-SO, en los que los elementos son iguales a 1. A tal matriz la llamaremos "peridentidad"; de modo que el problema queda expresado como sigue:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X = EX^TE. \end{array} \right. \tag{3.2}$$

Consideremos la descomposición en valores singulares de A:

$$A = P \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] Q^T,$$

dónde $P \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$, son matrices ortogonales, y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

En virtud de la invariancia de la norma de Frobenius bajo transformaciones ortogonales se tiene que:

$$\begin{aligned}
\|AX - B\|_{\mathtt{F}}^2 &= \left\| P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - B \right\|_{\mathtt{F}}^2 = \left\| P^T \left(P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - B \right) Q \right\|_{\mathtt{F}}^2 \\
&= \left\| \begin{bmatrix} \sum \\ 0 \end{bmatrix} (Q^T X Q) - P^T B Q \right\|_{\mathtt{F}}^2 = \left\| \begin{bmatrix} \sum \\ 0 \end{bmatrix} Y - Z \right\|_{\mathtt{F}}^2 \\
&= \left\| \sum Y - Z_1 \right\|_{\mathtt{F}}^2 + \|Z_2\|_{\mathtt{F}}^2
\end{aligned}$$

siendo:

$$Y = Q^T X Q,$$

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = P^T B Q,$$

$$Z_1 \in \mathbb{R}^{n\times n}.$$

De modo que resolver (3.2) es equivalente a resolver:

$$\left\{ \begin{array}{ll} \min & \| \sum Y - Z_1 \|_{\mathtt{F}}^2 \\ s.a & \\ & Y \in \mathcal{P}'. \end{array} \right. \tag{3.3}$$

Finalmente, se analiza la región de factibilidad del problema. Esto es, si X es tal que $X = EX^TE$, hay que determinar que relación existe entre Y e Y^T.

Proposición 3.0.1 *Sean X, E, P y Q como arriba, $Y = Q^T\ X\ Q$, entonces existe $U \in \mathbb{R}^{n\times n}$ ortogonal y simétrica tal que*

$$Y = U\ Y^T\ U.$$

Demostración :

$$\begin{aligned}
Y &= Q^T X Q = Q^T (E X^T E) Q = (Q^T E)(X^T)(EQ) = (Q^T E)(Q Y^T Q^T)(EQ) \\
&= (Q^T E Q) Y^T (Q^t E Q) = U Y^T U,
\end{aligned}$$

donde $U = Q^T EQ$ es ortogonal y simétrica pues:

$$\begin{aligned}
UU^T &= (Q^T EQ)(Q^T EQ)^T = (Q^T EQ)(Q^T E^T Q) = Q^T EQQ^T EQ \\
&= Q^T E^2 Q = Q^T Q = I,
\end{aligned}$$

$$\begin{aligned} U^TU &= (Q^TEQ)^T(Q^TEQ) = (Q^TE^TQ)(Q^TEQ) = Q^TEQQ^TEQ \\ &= Q^TE^2Q = Q^TQ = I, \end{aligned}$$

$$U^T = (Q^TEQ)^T = Q^TE^TQ = Q^TEQ = U,$$

pues Q ortogonal, $Q^TQ = QQ^T = I$, y $E^T = E$ y $E^2 = E$. □

De la proposición anterior resulta que Y verifica:

$$Y = UY^TU \Leftrightarrow YU = UY^T \Leftrightarrow YU = (YU)^T,$$

llamando $S = YU$, resulta que $S = S^T$. Luego expresando a (3.3) en términos de S, lo que se debe resolver es un problema en el cual la region de factibilidad es el conjunto de matrices simétricas, lo cual permite resolver el problema (3.2), resolviendo otro de tipo simétrico. Veamos como obtener la función objetivo en términos de S:

$$\begin{aligned} \|\sum Y - Z_1\|_{\mathtt{F}}^2 &= \|\left(\sum Y - Z_1\right)U\|_{\mathtt{F}}^2 = \|\sum YU - Z_1U\|_{\mathtt{F}}^2 \\ &= \|\sum S - \bar{Z}_1\|_{\mathtt{F}}^2. \end{aligned}$$

Luego el problema a resolver es el siguiente:

$$\left\{ \begin{array}{ll} \min & \|\sum S - \bar{Z}_1\|_{\mathtt{F}}^2 \\ s.a & \\ & S = S^T, \end{array} \right. \tag{3.4}$$

entonces, aplicando el resultado de Higham para el caso simétrico [8], se tiene:

$$S_{ij} = \left\{ \begin{array}{ll} \dfrac{\sigma_i(\bar{Z}_1)_{ij} + \sigma_j(\bar{Z}_1)_{ji}}{{\sigma_i}^2 + {\sigma_j}^2} & si \quad {\sigma_i}^2 + {\sigma_j}^2 \neq 0 \\ arbitrario & en \quad otro \quad caso. \end{array} \right.$$

Si S_* es la solución de (3.4), se obtiene $Y_* = S_*U$, y por lo tanto $X_* = QY_*Q^T$, la solución del problema (3.2).

3.1 Caracterización de las soluciones

Sea $\mathcal{S}$ el conjunto de soluciones del problema dado

$$\mathcal{S} = \left\{ X \in \mathbb{R}^{n\times n} \ : \ X = E\,X^T\,E, \quad \|AX - B\|_{\mathtt{F}}^2 = \min_{Z = EZ^TE} \|AZ - B\|_{\mathtt{F}}^2 \right\}.$$

En el siguiente lema se establecen propiedades del conjunto $\mathcal{S}$, de hecho generaliza el lema 2.2.1 al caso persimétrico.

Lema 3.1.1

1) $X \in \mathcal{S} \Longleftrightarrow X = EX^TE$

$y\ A^TA(XE) + (XE)A^TA = A^T(BE) + (BE)^TA.$

2) $\mathcal{S}$ *es convexo.*

3) $\mathcal{S}$ *tiene un único elemento* ${X_\star}^{mn}$ *de norma de Fobenius mínima.*

4) $\mathcal{S} = \{{X_\star}^{mn}\} \Longleftrightarrow rank(A) = n.$

Demostración :

1) Si $X \in \mathcal{S}$, es claramente persimétrica, y solo resta probar la igualdad. Para ello basta con considerar las ecuaciones normales, obtenidas por Higham para el caso simétrico. Pues si X es una solución de (3.2), entonces $X = QYQ^T$, donde $Y = SU$, con $U = Q^TEQ$ y S es tal que se obtiene como solución de un problema simétrico:

$$\left\{ \begin{array}{ll} \min & \|\Sigma S - \bar{Z}_1\|_{\mathtt{F}}^2 \\ s.a & \\ & S = S^T. \end{array} \right. \tag{3.5}$$

Por lo tanto S verifica las ecuaciones normales , es decir:

$$\Sigma^T\Sigma S + S\Sigma^T\Sigma = \Sigma^T\bar{Z}_1 + {\bar{Z}_1}^T\Sigma.$$

Teniendo en cuenta que:

$$\bar{Z}_1 = UZ_1,$$

$$Z = \left[\begin{array}{c} Z_1 \\ Z_2 \end{array} \right] = P^TBQ,$$

$$Z_1 \in \mathbb{R}^{n\times n},$$

$$A = P \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] Q^T,$$

$$\bar{Z}_2 = UZ_2,$$

el primer miembro de la igualdad es:

$$\begin{aligned} \Sigma^T\Sigma S + S\Sigma^T\Sigma &= \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right]^T \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] (YU) + (YU) \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right]^T \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] \\ &= \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right]^T \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] (Q^TXQ)(Q^TEQ) + (Q^TXQ)(Q^TEQ) \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right]^T \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] \end{aligned}$$

$$
\begin{aligned}
&= \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T \left(\begin{bmatrix} \Sigma \\ 0 \end{bmatrix} Q^T \right) (XE)Q + Q^T(XE) \left(Q \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T \right) \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T (P^T A)(XE)Q + Q^T(XE)(A^T P) \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \\
&= \left(\begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T P^T \right) A(XE)Q + Q^T(XE)A^T \left(P \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \right) \\
&= Q^T A^T A(XE)Q + Q^T(XE)A^T AQ \\
&= Q^T(A^T A(XE) + (XE)A^T A)Q.
\end{aligned}
$$

y el segundo miembro se transforma en:

$$
\begin{aligned}
\Sigma^T \bar{Z}_1 + \bar{Z}_1{}^T \Sigma &= \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T \begin{bmatrix} \bar{Z}_1 \\ \bar{Z}_2 \end{bmatrix} + \begin{bmatrix} \bar{Z}_1 \\ \bar{Z}_2 \end{bmatrix}^T \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T \begin{bmatrix} Z_1 U \\ Z_2 U \end{bmatrix} + \begin{bmatrix} Z_1 U \\ Z_2 U \end{bmatrix}^T \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} U + \left(\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} U \right)^T \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} U + U \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}^T \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T (P^T BQ)(Q^T EQ) + (Q^T EQ)(P^T BQ)^T \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \\
&= \left(\begin{bmatrix} \Sigma \\ 0 \end{bmatrix}^T P^T \right) BEQ + Q^T EB^T \left(P \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \right) \\
&= Q^T(A^T BE)Q + Q^T(EB^T)AQ \\
&= Q^T(A^T(BE) + (BE)^T A)Q.
\end{aligned}
$$

Luego se tiene que la igualdad inicial es equivalente a:

$$Q^T(A^T A(XE) + (XE)A^T A)Q = Q^T(A^T(BE) + (BE)^T A)Q,$$

es decir:

$$A^T A(XE) + (XE)A^T A = A^T(BE) + (BE)^T A.$$

Para probar la recíproca, se supone que X es persimétrica y verifica la última igualdad. Se debe probar que X es un minimizador para

$$f(X) = ||AX - B||_{\mathtt{F}}^2 = Traza((AX - B)^T(AX - B)).$$

Para ello basta con verificar que $f(X+\Delta) \geq f(X)$, cualesquiera que sea $\Delta = E\Delta^T E$.

Sea entonces:

$$\begin{aligned}
f(X+\Delta) &= Traza\left[(A(X+\Delta) - B)^T(A(X+\Delta) - B)\right] \\
&= Traza\left[(AX + A\Delta - B)^T(AX + A\Delta - B)\right] \\
&= Traza\left[((AX - B) + A\Delta)^T((AX - B) + A\Delta)\right] \\
&= Traza\left[((AX - B)^T + (A\Delta)^T)((AX - B) + A\Delta)\right] \\
&= Traza((AX - B)^T(AX - B) + (AX - B)^T(A\Delta) \\
&\quad + \ (A\Delta)^T(AX - B) + (A\Delta)^T(A\Delta)) \\
&= Traza((AX - B)^T(AX - B)) + Traza((A\Delta)^T(A\Delta)) \\
&\quad + \ Traza((AX - B)^T(A\Delta) + (A\Delta)^T(AX - B)) \\
&= f(X) + Traza((AX - B)^T(A\Delta) + (A\Delta)^T(AX - B)) + \\
&\quad + \ Traza((A\Delta)^T(A\Delta)).
\end{aligned}$$

Es claro que

$$Traza((A\Delta)^T(A\Delta)) = ||A\Delta||_{\mathtt{F}}^2 \geq 0,$$

además:

$$Traza((AX - B)^T(A\Delta) + (A\Delta)^T(AX - B)) = 0,$$

pues si $D = (AX - B)^T(A\Delta)$,

$$\begin{aligned}
D + D^T &= \left[(AX - B)^T(A\Delta)\right] + \left[(AX - B)^T(A\Delta)\right]^T \\
&= X^T A^T A\Delta - B^T A\Delta + \Delta^T A^T AX - \Delta^T A^T B \\
&= (EXE)A^T A\Delta - E^2 B^T A\Delta) + (\Delta^T A^T AX - \Delta^T A^T B)E^2 \\
&= E\left[(XE)A^T A - EB^T A\right]\Delta + \Delta^T\left[A^T A(XE) - A^T(BE)\right]E \\
&= E\left[(XE)A^T A - (BE)^T A\right]\Delta + \left[E\left[A^T A(XE) - A^T(BE)\right]^T\Delta\right]^T \\
&= EM\Delta + [EM\Delta]^T,
\end{aligned}$$

siendo

$$M = (XE)A^TA - (BE)^TA = \left[A^TA(XE) - A^T(BE)\right]^T$$

de las ecuaciones normales para el caso persimétrico, se tiene que

$$M^T = -M,$$

luego, por ser Δ persimétrica y $Traza(AB) = Traza(BA)$, cualesquiera que sean A y $B \in \mathbb{R}^{m\times m}$, se tiene que:

$$\begin{aligned}
Traza(D + D^T) &= Traza\left[(EM\Delta) + (EM\Delta)^T\right] = Traza\left[(EM\Delta) + (\Delta^T M^T E)\right] \\
&= Traza\left[EM\Delta + (E\Delta E)M^T E\right] = Traza\left[E(M\Delta + \Delta E(-M)E)\right] \\
&= Traza\left[(M\Delta + \Delta E(-M)E)E\right] = Traza\left[M\Delta E - \Delta EME^2\right] \\
&= Traza\left[M(\Delta E) - (\Delta E)M\right] = Traza\left[M(\Delta E)\right] - Traza\left[(\Delta E)M\right] \\
&= 0.
\end{aligned}$$

Queda probado que X es minimizador de f.

2) S es convexo: Sean X_1 y X_2 en $\mathcal{S}$, $\mu \in \mathbb{R}$, entonces:

i) $\mu X_1 + (1-\mu)X_2 = E(\mu X_1 + (1-\mu)X_2)^T E$:

$$\begin{aligned}
E(\mu X_1 + (1-\mu)X_2)^T E &= E(\mu {X_1}^T + (1-\mu){X_2}^T)E \\
&= E(\mu {X_1}^T)E + E((1-\mu){X_2}^T)E \\
&= \mu(E{X_1}^T E) + (1-\mu)(E{X_2}^T E) = \mu X_1 + (1-\mu)X_2
\end{aligned}$$

ii) $A^TA((\mu X_1 + (1-\mu)X_2)E) + ((\mu X_1 + (1-\mu)X_2)E)A^TA = A^T(BE) + (BE)^TA$

$$\begin{aligned}
A^TA((\mu X_1 + (1-\mu)X_2)E) &= A^TA(\mu X_1 E + (1-\mu)X_2 E) \\
&= \mu A^TA(X_1E) + (1-\mu)A^TA(X_2E)
\end{aligned}$$

utilizando 1), la condición necesaria y suficiente para que $X \in S$ es que se verifiquen las correspondientes ecuaciones normales:

$$\begin{aligned}
((\mu X_1 + (1-\mu)X_2)E)A^TA &= (\mu X_1 E + (1-\mu)X_2 E)A^TA \\
&= \mu(X_1E)A^TA + (1-\mu)(X_2E)A^TA
\end{aligned}$$

luego, sumando miembro a miembro, se tiene que la expresión:

$$A^TA((\mu X_1 + (1-\mu)X_2)E) + ((\mu X_1 + (1-\mu)X_2)E)A^TA,$$

es igual a

$$(\mu A^TA(X_1E) + (1-\mu)A^TA(X_2E)) + (\mu(X_1E)A^TA + (1-\mu)(X_2E)A^TA),$$

reagrupando de manera conveniente resulta:

$$\mu(A^T(BE) + (BE)^TA) + (1-\mu)(A^T(BE) + (BE)^TA) = A^T(BE) + (BE)^TA.$$

i) y ii) prueban que $\mathcal{S}$ es convexo.

3) El único elemento de mínima norma se obtiene considerando en la expresión de S_{ij} los valores arbitrarios igual a cero. De modo que S será de mínima norma, y por lo tanto $Y = SU$, y $X = QYQ^T$, ya que U y Q son matrices ortogonales.

4)Si $rank(A) = n$, es claro que los valores singulares serán no nulos en su totalidad, y por lo tanto la única solución es la de mínima norma . □

Observaciones:

- Si $m = n$ y $A = I$, el problema se reduce a encontrar la matriz persimétrica, más cercana en norma de Frobenius a una matriz dada B, es decir:

$$\left\{ \begin{array}{ll} \min & ||X - B||_F{}^2 \\ s.a & \\ & X = EX^TE. \end{array} \right. \tag{3.6}$$

  Luego, a partir del lema anterior se tiene que en este caso particular la solución es:

$$X = \frac{B + (EB^TE)}{2}.$$

- Es posible llevar a cabo la tranformación del problema inicial de manera más sencilla, para convertirlo en un problema simétrico. Sea el problema (3.2), se requiere que la solución del mismo verifique:

$$X = EX^TE,$$

  lo cual es equivalente a:

$$XE = EX^T = (XE)^T.$$

  Con el cambio de variables $T = XE$, sólo se necesita que $T = T^T$, luego ya que E es ortogonal

$$\|AX - B\|_{\mathtt{F}} = \|I(AX - B)E\|_{\mathtt{F}} = \|A(XE) - (BE)\|_{\mathtt{F}} = \|AT - \bar{B}\|_{\mathtt{F}},$$

  por lo tanto, resolver el problema persimétrico es equivalente a resolver el siguiente:

$$\left\{ \begin{array}{ll} \min & \|AT - \bar{B}\|_{\mathtt{F}}^2 \\ s.a & \\ & T = T^T, \end{array} \right. \tag{3.7}$$

  el cual ha sido resuelto por Higham. Luego, si $T^\star$ es su solución, $X^\star = T^\star E$, es la solución del problema (3.2).

Teniendo en cuenta esta observación, es posible demostrar de manera más directa al lema anterior.

Investigaremos cual es el comportamiento de la solución obtenida frente a pequeñas perturbaciones de los datos estableciendo el siguiente teorema:

Teorema 3.1.1 *Sean A, $B \in \mathbb{R}^{m\times n}$, $m \geq n$, $rank(A) = n$.*
Sea X solución de (3.2), y $\hat{X}$ solución

$$\left\{ \begin{array}{ll} \min & \|(A+\delta A)X - (B+\delta B)\|_{\mathtt{F}}^2 \\ s.a & \\ & X = EX^TE, \end{array} \right. \tag{3.8}$$

siendo δA, $\delta B \in \mathbb{R}^{m\times n}$, tal que $rank(A+\delta A) = rank(A) = n$. Definiendo:

$$R = AX - B$$

$$\widehat{R} = (A+\delta A)\hat{X} - (B+\delta B),$$

$$\epsilon_A = \frac{||\delta A||_2}{||A||_2},$$

$$\kappa = \kappa_2(A),$$

si $\kappa\ \epsilon_A < 1$, entonces:

$$||X - \hat{X}\|_{\mathtt{F}} \leq \frac{\kappa}{1-\epsilon_A}\left(\epsilon_A||X\|_{\mathtt{F}} + \frac{||\delta B\|_{\mathtt{F}}}{||A||_2} + \kappa\epsilon_A\frac{||\mathcal{R}\|_{\mathtt{F}}}{||A||_2}\right) + \kappa\epsilon_A||X\|_{\mathtt{F}} \tag{3.9}$$

$$||R - \widehat{R}\|_{\mathtt{F}} \leq \epsilon_A||X\|_{\mathtt{F}}\ ||A||_2 + ||\delta B\|_{\mathtt{F}} + \kappa\ \epsilon_A||R\|_{\mathtt{F}} \tag{3.10}$$

Demostración :

Se sabe que resolver (3.2) es equivalente a resolver (3.7). Si se perturba ligeramente el problema simétrico anterior

$$\left\{ \begin{array}{ll} \min & \|(A+\delta A)T - (B+\delta B)\|_{\mathtt{F}}^2 \\ s.a & \\ & T = T^T, \end{array} \right. \tag{3.11}$$

de solución $\hat{T}$, se tiene que:

$$||T - \hat{T}\|_{\mathtt{F}} \leq \frac{\kappa}{1-\epsilon_A}\left(\epsilon_A||T\|_{\mathtt{F}} + \frac{||\delta B\|_{\mathtt{F}}}{||A||_2} + \kappa\epsilon_A\frac{||R_T\|_{\mathtt{F}}}{||A||_2}\right) + \kappa\epsilon_A||T\|_{\mathtt{F}} \tag{3.12}$$

$$||R_T - \widehat{R_T}\|_{\mathtt{F}} \leq \epsilon_A||X\|_{\mathtt{F}}\ ||A||_2 + ||\delta B\|_{\mathtt{F}} + \kappa\ \epsilon_A||R_T\|_{\mathtt{F}} \tag{3.13}$$

para :

δA, $\delta \bar{B} \in \mathbb{R}^{m \times n}$

$R_T = AT - \bar{B}$

$\widehat{R_T} = (A + \delta A)\hat{T} - (\bar{B} + \delta \bar{B})$,

en particular, para $\delta \bar{B} = \delta BE$, resulta el problema cuya solución es $\hat{T} = \hat{X}E$, y verificandose las dosdesigualdades anteriores; por lo tanto:

$$\begin{aligned}
\|X - \hat{X}\|_{\mathtt{F}} &= \|TE - \hat{T}E\|_{\mathtt{F}} = \|(T - \hat{T})E\|_{\mathtt{F}} \\
&\leq \frac{\kappa}{1 - \kappa\epsilon_A}\left(\epsilon_A||T||_{\mathtt{F}} + \frac{||\delta\bar{B}||_{\mathtt{F}}}{||A||_2} + \kappa\epsilon_A + \frac{||R_T||_{\mathtt{F}}}{||A||_2}\right) \\
&\leq \frac{\kappa}{1 - \kappa\epsilon_A}\left(\epsilon_A||XE||_{\mathtt{F}} + \frac{||\delta BE||_{\mathtt{F}}}{||A||_2} + \kappa\epsilon_A + \frac{||AT - \bar{B}||_{\mathtt{F}}}{||A||_2}\right) \\
&\leq \frac{\kappa}{1 - \kappa\epsilon_A}\left(\epsilon_A||X||_{\mathtt{F}} + \frac{||\delta B||_{\mathtt{F}}}{||A||_2} + \kappa\epsilon_A + \frac{||AT - \delta BE||_{\mathtt{F}}}{||A||_2}\right) \\
&\leq \frac{\kappa}{1 - \kappa\epsilon_A}\left(\epsilon_A||X||_{\mathtt{F}} + \frac{||\delta B||_{\mathtt{F}}}{||A||_2} + \kappa\epsilon_A + \frac{||AT - \delta B||_{\mathtt{F}}}{||A||_2}\right)
\end{aligned}$$

lo cual muestra la primera de las dos desigualdades propuestas. Para demostrar la restante, se tiene que $||R - \hat{R}||_{\mathtt{F}} = ||R_T - \widehat{R_T}||_{\mathtt{F}}$, pues:

$$\begin{aligned}
||R - \hat{R}||_{\mathtt{F}} &= ||(AX - B) - ((A + \delta A)\hat{X} - (B + \delta B))||_{\mathtt{F}} \\
&= ||[(AX - B) - ((A + \delta A)\hat{X} - (B + \delta B))]E||_{\mathtt{F}} \\
&= ||[(AXE - BE) - ((A + \delta A)\hat{X}E - (B + \delta B)E)||_{\mathtt{F}} \\
&= ||[(AXE - BE) - ((A + \delta A)\hat{X}E - (BE + \delta BE)||_{\mathtt{F}} \\
&= ||(AT - \bar{B}) - ((A + \delta A)\hat{T} - (\bar{B} + \delta\bar{B}))||_{\mathtt{F}} = ||R_T - \widehat{R_T}||_{\mathtt{F}}
\end{aligned}$$

además, está demostrado [8] que:

$$||R_T - \widehat{R_T}||_{\mathtt{F}} \leq \epsilon_A||T||_{\mathtt{F}}||A||_2 + ||\delta\bar{B}||_{\mathtt{F}} + \kappa\epsilon_A||R_T||_{\mathtt{F}}.$$

finalmente:

$$\begin{aligned}
||R - \hat{R}||_{\mathtt{F}} &= ||R_T - \widehat{R_T}||_{\mathtt{F}} \leq \epsilon_A||T||_{\mathtt{F}}\,||A||_2 + ||\delta\bar{B}||_{\mathtt{F}} + \kappa\epsilon_A||R_T||_{\mathtt{F}} \\
&\leq \epsilon_A||XE||_{\mathtt{F}}\,||A||_2 + ||\delta BE||_{\mathtt{F}} + \kappa\epsilon_A||AT - \bar{B}||_{\mathtt{F}} \\
&\leq \epsilon_A||X||_{\mathtt{F}}\,||A||_2 + ||\delta B||_{\mathtt{F}} + \kappa\epsilon_A||AXE - BE||_{\mathtt{F}} \\
&\leq \epsilon_A||X||_{\mathtt{F}}\,||A||_2 + ||\delta B||_{\mathtt{F}} + \kappa\epsilon_A||AX - B||_{\mathtt{F}} \\
&\leq \epsilon_A||X||_{\mathtt{F}}\,||A||_2 + ||\delta B||_{\mathtt{F}} + \kappa\epsilon_A||R||_{\mathtt{F}}.
\end{aligned}$$

lo cual demuestra el teorema □

La observación efectuada previo al desarrollo del teorema anterior, es sumamente útil a la hora de hallar una expresión para el residuo.

Teorema 3.1.2 *Sea el problema persimétrico de Procrusto (3.2), y supongamos que* $rank(A) = n$. *Sea* $X_\star$ *la solución del mismo. Entonces el residuo puede expresarse como sigue:*

$$\rho^2{}_{\text{pp}} = \|AX_\star - B\|_{\text{F}}^2 = \sum_{j>i} \frac{(\sigma_i(C_1)_{ji} - \sigma_j(C_1)_{ij})^2}{\sigma_i{}^2 + \sigma_j{}^2} + \|C_2\|_{\text{F}}^2,$$

siendo:

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = P^T(BE)Q \in \mathbb{R}^{m\times n},$$

P, Q y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, *las matrices involucradas en la descomposición en valores singulares de la matriz A.*

Demostración :

En virtud de la observación se tiene que

$$\rho^2{}_{\text{pp}} = \|AX_\star - B\|_{\text{F}}^2 = \|(AX_\star - B)E\|_{\text{F}}^2 = \|A(X_\star E) - (BE)\|_{\text{F}}^2.$$

Ya que $S_\star = (X_\star E) \in \mathbb{R}^{n\times n}$ es simétrica, resulta:

$$\rho^2{}_{\text{pp}} = \|AS_\star - (BE)\|_{\text{F}}^2,$$

Luego, según ha demostrado Higham [8] para el problema simétrico de Procrusto, el residuo puede expresarse:

$$\rho^2{}_{\text{pp}} = \|AS_\star - (BE)\|_{\text{F}}^2 = \sum_{j>i} \frac{(\sigma_i(C_1)_{ji} - \sigma_j(C_1)_{ij})^2}{\sigma_i{}^2 + \sigma_j{}^2} + \|C_2\|_{\text{F}}^2.$$

□

3.2 Experiencia numérica

El procedimiento computacional que permite hallar la solución para el problema (3.2) es simple, implementado en Fortran, utilizando la versión FORTRAN POWER STATION (1993) en una PC Pentium(R2) Processor Intelmmx(TM) Technology, 31.0 MB de RAM. y 32 bits de memoria Virtual. Para la descomposición en valores

singulares se han empleado las subrutinas de LINPACK.

Se considera el problema

$$\begin{cases} \min & \|AX - B\|_{\mathrm{F}}^2 \\ s.a & \\ & X = EX^T E, \end{cases}$$

para distintas dimensiones de las matrices A y B.

Ejemplo 3.2.1 Considerando A y B como sigue:

$$A = \begin{bmatrix} 5 & 3 & 2 \\ 1 & 2 & 4 \\ 6 & 0 & 3 \\ -1 & 2 & -3 \end{bmatrix}, \qquad B = \begin{bmatrix} 15 & 10 & -3 \\ 1 & 5 & 3 \\ 15 & 6 & -3 \\ 2 & 3 & -2 \end{bmatrix}.$$

La matriz solución es la siguiente:

$$X_\star = \begin{bmatrix} -.9896 & 0.9203 & 2.9339 \\ 0.0315 & 1.8791 & 0.9203 \\ .9838 & 0.0315 & -0.9896 \end{bmatrix}.$$

Ejemplo 3.2.2 Para las siguientes matrices A y B

$$A = \begin{bmatrix} -1 & 2 & 3 & -2 \\ 1 & 4 & 5 & -1 \\ 0 & 2 & 3 & 7 \\ -1 & 2 & -1 & 0 \\ 3 & 4 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 12 & 0 & 1 & 3 \\ 1 & -1 & 2 & -2 \\ 1 & 3 & 2 & 1 \\ 0 & 2 & 1 & 10 \\ 1 & 1 & 2 & 3 \\ -4 & 3 & -1 & 2 \\ 10 & 9 & 0 & 1 \end{bmatrix}.$$

La matriz solución es la siguiente:

$$X_\star = \begin{bmatrix} -6.2156 & 6.0507 & 6.6790 & -9.6773 \\ 7.9559 & -1.4131 & 4.7940 & 6.6790 \\ -2.6760 & 4.9199 & -1.4131 & 6.0507 \\ 6.4650 & -2.6760 & 7.9559 & -6.2156 \end{bmatrix}.$$

Ejemplo 3.2.3 Finalmente, considerando A y B como sigue

$$A = \begin{bmatrix} 1 & 1 & 2 & -2 & 2 & 1 \\ 0 & 2 & -1 & -2 & -3 & 2 \\ 0 & 2 & 1 & -1 & 2 & 2 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 2 & 2 & -1 & 2 & 0 & 1 \\ 3 & -1 & 1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 & 0 & 1 \\ 1 & -1 & -2 & 0 & -1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & -1 & 0 & -1 & 1 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & -1 & 0 & 1 \\ -1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 2 & -1 & 2 & 2 \\ 1 & 1 & 2 & 0 & -1 & 1 \\ 1 & 2 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ -1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & -1 & 0 \end{bmatrix}.$$

La matriz matriz solución $X_\star$ es:

$$\begin{bmatrix} 1.81 \times 10^{-1} & -3.53 \times 10^{-1} & -1.27 \times 10^{-1} & 5.11 \times 10^{-1} & -9.63 \times 10^{-2} & 2.17 \times 10^{-1} \\ 3.08 \times 10^{-1} & 3.11 \times 10^{-2} & -1.34 \times 10^{-2} & 2.17 \times 10^{-1} & -1.69 \times 10^{-1} & -9.63 \times 10^{-2} \\ -9.78 \times 10^{-3} & -8.92 \times 10^{-2} & -1.82 \times 10^{-1} & -1.42 \times 10^{-1} & 2.17 \times 10^{-1} & 5.11 \times 10^{-1} \\ 4.51 \times 10^{-2} & 3.63 \times 10^{-1} & 1.19 \times 10^{-3} & -1.82 \times 10^{-1} & -1.34 \times 10^{-2} & -1.27 \times 10^{-1} \\ 2.43 \times 10^{-1} & 1.53 \times 10^{-1} & 3.63 \times 10^{-1} & -8.92 \times 10^{-2} & 3.11 \times 10^{-2} & -3.53 \times 10^{-1} \\ 2.95 \times 10^{-1} & 2.43 \times 10^{-1} & 4.51 \times 10^{-2} & -9.78 \times 10^{-3} & 3.08 \times 1^{-1} & 1.81 \times 10^{-1} \end{bmatrix}.$$

Hemos estimado el valor del residuo ρ_{pp} según la expresión del Teorema 3.1.2. El ejemplo 3.2.1, incluído aquí, que ha sido también presentado por Higham en su trabajo sobre el problema simétrico de Procrusto, ha sido elegido para evaluar el valor del residuo. Hemos encontrado en este caso que $\rho_{\text{pp}} = 1.33 \times 10^{-2}$. Siendo el número de condición de la matriz solución $\kappa(X_\star) = 8.3811$.

Capítulo 4

El problema Toeplitz

En este capítulo se considera la clase de todos los problemas de cuadrados mínimos con restricciones del tipo:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T}, \end{array} \right. \tag{4.1}$$

donde A y B son matrices rectangulares, pertenecientes a $\mathbb{R}^{m\times n}$, con $m > n$; y $\mathcal{T}$ es el subespacio de las matrices de Toeplitz.

Definición (matriz Toeplitz): Una matriz $T \in \mathbb{R}^{n\times n}$ se dice Toeplitz si es constante a lo largo de sus diagonales, es decir si existen escalares, $r_{-n+1}, \cdots, r_0, \cdots, r_{n-1}$, tales que $a_{ij} = r_{j-i}$, para todo i, j:

$$T = \begin{bmatrix} r_0 & r_1 & r_2 & r_3 & \cdots & r_{n-2} & r_{n-1} \\ r_{-1} & r_0 & r_1 & r_2 & r_3 & \cdots & r_{n-2} \\ r_{-2} & r_{-1} & r_0 & r_1 & r_2 & \cdots & r_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{-(n-3)} & r_{-(n-4)} & \cdots & r_{-1} & r_0 & r_1 & r_2 \\ r_{-(n-2)} & r_{-(n-3)} & \cdots & \cdots & \cdots & r_0 & r_1 \\ r_{-(n-1)} & r_{-(n-2)} & \cdots & \cdots & \cdots & r_{-1} & r_0 \end{bmatrix}.$$

Como es claro, si T es Toeplitz, entonces T es persimétrica. El nombre de tales matrices se debe a Otto Toeplitz, quien en trabajos fechados a principios de siglo ha estudiado sus propiedades y aplicaciones.

Llamaremos $\mathcal{T}$, al subespacio de matrices Toeplitz densas, $\mathcal{T}_{\mathtt{s}}$, al subespacio de matrices Toeplitz triangular superior, y $\mathcal{T}_{\mathtt{i}}$, al subespacio de matrices Toeplitz triangular inferior.

4.1 El problema general

Consideramos A, $B \in \mathbb{R}^{m\times n}$, en las condiciones anteriores. Interesa el caso en el cual el conjunto de factibilidad está formado por todas las matrices de Toeplitz. Notemos que $\mathcal{T}$ puede ser representado de la siguiente manera:

$$\mathcal{T} = \left\{ X \in \mathbb{R}^{n\times n} \ : \ X = \sum_{p=-(n-1)}^{(n-1)} \alpha_p G_p, \right\}$$

donde $\alpha_p \in \mathbb{R}$ son escalares y las matrices $G_p \in \mathbb{R}^{n\times n}$ están definidas como sigue:

$$(G_p)_{ij} = \begin{cases} 1 & si \ \ j = i + p \\ \\ 0 & en \ \ otro \ \ caso. \end{cases} \tag{4.2}$$

Debe notarse que las matrices G_p se caracterizan por el hecho de que para cada entrada ij, existe un único p, $-(n-1) \leq p \leq (n-1)$, tal que $(G_p)_{ij} = 1$. Además constituyen una base del subespacio $\mathcal{T}$.

4.1.1 Transformación del problema

Consideremos la descomposición en valores singulares de A:

$$A = P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T,$$

dónde $P \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$, son matrices ortogonales, y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

En virtud de la invariancia de la norma de Frobenius bajo transformaciones ortogonales se tiene que:

$$\begin{aligned} \|AX - B\|_{\mathtt{F}}^2 &= \left\| P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - B \right\|_{\mathtt{F}}^2 = \left\| P^T \left(P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - B \right) Q \right\|_{\mathtt{F}}^2 \\ &= \left\| \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - P^T B \right\|_{\mathtt{F}}^2 = \left\| \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T X - C \right\|_{\mathtt{F}}^2 \\ &= \left\| (\sum Q^T) X - C_1 \right\|_{\mathtt{F}}^2 + \|C_2\|_{\mathtt{F}}^2 = \left\| T - C_1 \right\|_{\mathtt{F}}^2 + \|C_2\|_{\mathtt{F}}^2 \end{aligned}$$

siendo:

$$T = \sum Q^T X \in \mathbb{R}^{n\times n},$$

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = P^T B,$$

$C_1 \in \mathbb{R}^{n\times n}$.

luego, el problema inicial es equivalente a:

$$\left\{\begin{array}{ll} \min & \|T - C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & T \in \mathcal{T}', \end{array}\right. \tag{4.3}$$

donde $\mathcal{T}'$ es el siguiente subespacio:

$$\mathcal{T}' = \left\{ T \in \mathbb{R}^{n\times n} \ : \ T = \sum Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l G_l \right\}.$$

4.1.2 Caracterización de las soluciones sobre $\mathcal{T}'$

El teorema siguiente caracteriza las proyecciones sobre el subespacio de matrices $\mathcal{T}'$.

Teorema 4.1.1 *Si $C_1 \in \mathbb{R}^{n\times n}$, entonces, la única solución del problema (4.3) está dada por:*

$$P_{\mathcal{T}'}(C_1) = \Sigma Q^T \sum_{l=-(n-1)}^{(n-1)} {\alpha^\star}_l G_l, \qquad {\alpha^\star}_l = \frac{\sum_{i=1}^{n}\sum_{j>l}^{n}(C_1)_{ij}Q_{(j-l)i}}{\sum_{i=1}^{n}\sum_{j>l}^{n}(Q_{(j-l)i})^2 {\sigma_i}^2},$$

para $-(n-1) \leq l \leq (n-1)$.

Demostración :

Sea f la función objetivo; $f : \mathbb{R}^{2n-1} \to \mathbb{R}$, dada por:

$$\begin{aligned} f(\alpha) = f(\alpha_{-(n-1)}, \cdots, \alpha_0, \cdots, \alpha_{(n-1)}) &= \frac{1}{2}||T - C_1\|_{\mathtt{F}}^2 = \frac{1}{2}||\sum Q^T X - C_1\|_{\mathtt{F}}^2 \\ &= \frac{1}{2}||\sum Q^T \sum_{l=-(n-1)}^{(n-1)} \bar{\alpha}_l G_l - C_1\|_{\mathtt{F}}^2 \\ &= \frac{1}{2}\sum_{i,j=1}^{n}\left((\sum Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l G_l)_{ij} - (C_1)_{ij} \right)^2. \end{aligned}$$

El gradiente de f,

$$\nabla f = \left(\frac{\partial f}{\partial \alpha_{-(n-1)}}, \frac{\partial f}{\partial \alpha_{-(n-2)}}, \cdots, , \frac{\partial f}{\partial \alpha_0}, \cdots, \frac{\partial f}{\partial \alpha_{(n-1)}} \right)^T.$$

Luego, $\nabla f(\alpha) = 0$, si y solo si para todo p tal que $-(n-1) \leq p \leq (n-1)$ la derivada $\dfrac{\partial f}{\partial \alpha_p}(\alpha) = 0$, esto es:

$$\begin{aligned}
\frac{\partial f}{\partial \alpha_p} &= \frac{\partial}{\partial \alpha_p}\left(\left(\frac{1}{2}\sum_{i,j=1}^{n}\left(\sum Q^T \sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij} - (C_1)_{ij}\right)^2\right) \\
&= \left(\sum_{i,j=1}^{n}\left(\sum Q^T \sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij} - (C_1)_{ij}\right)\frac{\partial}{\partial \alpha_p}\left(\left(\sum Q^T \sum_{l=1}^{n}(\alpha_l G_l)\right)_{ij}\right) = 0.
\end{aligned}$$

La derivada en el segundo factor de la expresión anterior es:

$$\begin{aligned}
\frac{\partial}{\partial \alpha_p}\left(\left(\sum Q^T \sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij}\right) &= \frac{\partial}{\partial \alpha_p}\left(\sum_{k=1}^{n}(\sum Q^T)_{ik}\left(\sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{kj}\right) \\
&= \sum_{k=1}^{n}(\sum Q^T)_{ik}\frac{\partial}{\partial \alpha_p}\left(\sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{kj} \\
&= \left(\sum Q^T G_p\right)_{ij}
\end{aligned}$$

luego se tiene que

$$\frac{\partial f}{\partial \alpha_p} = \left(\sum_{i,j=1}^{n}\left(\sum Q^T(\sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij} - (C_1)_{ij}\right)\left(\sum Q^T G_p\right)_{ij}.$$

Teniendo en cuenta que el elemento

$$\left(\sum Q^T \sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij}$$

es el producto escalar entre la i-ésima fila de $(\sum Q^T)$ y la j-ésima columna de $\sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)$, se obtiene que

$$\left(\sum Q^T \sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij} = \sigma_i \sum_{k=(j-n)}^{(j-1)} Q_{(j-k)i}\alpha_k.$$

De manera similar, el elemento $\left(\sum Q^T G_p\right)_{ij}$, es el producto escalar entre la fila i de $\sum Q^T$ y la columna j de G_p, y por lo tanto:

$$\left(\sum Q^T G_p\right)_{ij} = \sigma_i Q_{(j-p)i}.$$

A partir de la estructura sparse de las matrices G_p, debe notarse que siempre que $(G_p)_{ij} = 1$ es $j > p$, pues:

Si $p > 0$, entonces $(G_p)_{ij} = 1$ para todo j tal que $p+1 \leq j \leq n$.

Si $p \leq 0$, entonces $(G_p)_{ij} = 1$ para todo j tal que $1 \leq j \leq (n+p)$.

Así, el producto:

$$\left(\sum_{i=1}^{n}\sum_{j>p}^{n}\left(\sum Q^T \sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij}\right)\left(\sum Q^T G_p\right)_{ij}$$

$$= \sum_{i=1}^{n}\sum_{j>p}^{n}\left(\sigma_i \sum_{k=(j-n)}^{(j-1)} Q_{(j-k)i}\alpha_k\right)(\sigma_i Q_{(j-p)i}),$$

$$= \sum_{j>p}^{n}\sum_{k=(j-n)}^{(j-1)} \alpha_k \sum_{i=1}^{n}\left(Q_{(j-k)i}Q_{(j-p)i}\right)(\sigma_i)^2.$$

Observemos que, llamando

$$\sum_{i=1}^{n}\left(Q_{(j-k)i}Q_{(j-p)i}\right) = v^T\theta,$$

con:

$$v^T = (Q_{(j-k)1}Q_{(j-p)1}, Q_{(j-k)2}Q_{(j-p)2}, \cdots, Q_{(j-k)n}Q_{(j-p)n}),$$
$$\theta = ((\sigma_1)^2, (\sigma_2)^2, \cdots, (\sigma_n)^2)^T,$$

se verifica que la suma de todas las componentes del vector v es el producto escalar entre dos columnas de la matriz ortogonal Q, y por consiguiente, toda vez que $k \neq p$, dicha suma será cero. Se tiene entonces, para $k \neq p$, que:

$$0 = \left(\sum_{i=1}^{n} v_i\right)(\sigma_n)^2 \leq v^T\theta = \sum_{i=1}^{n} v_i(\sigma_i)^2 \leq \left(\sum_{i=1}^{n} v_i\right)(\sigma_1)^2 = 0,$$

luego, $v^T\theta = 0$ para todo $k \neq p$, y con ello resulta lo siguiente:

$$\begin{aligned}\sum_{j>p}^{n}\sum_{k=(j-n)}^{(j-1)} \alpha_k \sum_{i=1}^{n}\left(Q_{(j-k)i}Q_{(j-p)i}\right)(\sigma_i)^2 &= \sum_{j=1}^{n}\sum_{k=(j-n)}^{(j-1)} \alpha_k v^T\theta \\ &= \sum_{j>p}^{n}\alpha_p \sum_{i=1}^{n}\left(Q_{(j-p)i}\right)^2(\sigma_i)^2.\end{aligned}$$

Luego:

$$\begin{aligned}\frac{\partial f}{\partial \alpha_p} &= \sum_{i,j=1}^{n}\left(\sum Q^T \sum_{l=-(n-1)}^{(n-1)}(\alpha_l G_l)\right)_{ij}\left(\sum Q^T G_p\right)_{ij} - (C_1)_{ij}\left(\sum Q^T G_p\right)_{ij} \\ &= \alpha_p \sum_{j>p}^{n}\sum_{i=1}^{n}\left(Q_{(j-p)i}\right)^2(\sigma_i)^2 - \sum_{j>p}^{n}\sum_{i=1}^{n}(C_1)_{ij}Q_{(j-p)i},\end{aligned}$$

a partir de donde se obtiene el valor de α_p que anula la derivada:

$$\alpha_p = \frac{\sum_{i=1}^{n}\sum_{j>p}^{n}(C_1)_{ij}Q_{(j-p)i}}{\sum_{i=1}^{n}\sum_{j>p}^{n}\left(Q_{(j-p)i}\right)^2(\sigma_i)^2}. \tag{4.4}$$

Debe notarse que el denominador en la expresión anterior, es no nulo, pues si fuera cero, entonces, cualesquiera que sean i, j, sería $Q_{(j-p)i}\sigma_i = 0$, y por lo tanto $\sum Q^T = 0$, y $P\begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T = A = 0$.

Resta verificar que el valor de α_p hallado es un minimizador. Para ello basta con ver que:

$$\frac{\partial^2 f}{\partial {\alpha_p}^2}(\alpha^\star) = \sum_{j=1}^{n}\sum_{i=1}^{n}\left(Q_{(j-p)i}\sigma_i\right)^2 > 0, \qquad \forall\ i,\ j,$$

y además

$$\frac{\partial^2 f}{\partial \alpha_p \alpha_q} = 0, \qquad si\ \ p \neq q,$$

lo que implica que el Hessiano es definido positivo, y α_p dado por (4.4) es minimizador de f. □

4.1.3 Experiencia Numérica

Dado el problema

$$\begin{cases} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T}, \end{cases} \tag{4.5}$$

lo hemos resuelto para el caso particular en que las matrices A y B son las consideradas en los ejemplos de la sección 2 del capítulo anterior. Se han obtenido las siguientes soluciones:

Ejemplo 4.1.1 Considerando A y B como en el ejemplo 3.2.1, hemos obtenido:

$$X_\star = \begin{bmatrix} 1.67 \times 10^{-1} & 0 & 0 \\ 1.29 \times 10^{-1} & 1.67 \times 10^{-1} & 0 \\ 3.76 \times 10^{-1} & 1.29 \times 10^{-1} & 1.67 \times 10^{-1} \end{bmatrix}$$

Ejemplo 4.1.2 Para A y B como en el ejemplo 3.2.2, se obtuvo:

$$X_\star = \begin{bmatrix} 5.05 \times 10^{-3} & 0 & 0 & 0 \\ 4.37 \times 10^{-2} & 5.05 \times 10^{-3} & 0 & 0 \\ 7.34 \times 10^{-2} & 4.37 \times 10^{-2} & 5.05 \times 10^{-3} & 0 \\ 1.57 \times 10^{-1} & 7.34 \times 10^{-2} & 4.37 \times 10^{-2} & 5.05 \times 10^{-3} \end{bmatrix}$$

Ejemplo 4.1.3 Considerando A y B como en el ejemplo 3.2.3, hemos obtenido como matriz solución $X_\star$:

$$\begin{bmatrix} -3.49 \times 10^{-2} & 9.34 \times 10^{-3} & 1.37 \times 10^{-2} & 1.34 \times 10^{-2} & 1.68 \times 10^{-2} & 1.66 \times 10^{-2} \\ -5.43 \times 10^{-2} & -3.49 \times 10^{-2} & 9.34 \times 10^{-3} & 1.37 \times 10^{-2} & 1.34 \times 10^{-2} & 1.68 \times 10^{-2} \\ -7.2 \times 10^{-3} & -5.43 \times 10^{-2} & -3.49 \times 10^{-2} & 9.34 \times 10^{-3} & 1.37 \times 10^{-2} & 1.34 \times 10^{-2} \\ 1.46 \times 10^{-2} & -7.2 \times 10^{-3} & -5.43 \times 10^{-2} & -3.49 \times 10^{-2} & 9.34 \times 10^{-3} & 1.37 \times 10^{-2} \\ 8.75 \times 10^{-4} & 1.46 \times 10^{-2} & -7.2 \times 10^{-3} & -5.43 \times 10^{-2} & -3.49 \times 10^{-2} & 9.34 \times 10^{-3} \\ -2.268 \times 10^{-2} & 8.75 \times 10^{-4} & 1.46 \times 10^{-2} & -7.2 \times 10^{-3} & -5.43 \times 10^{-2} & -3.49 \times 10^{-2} \end{bmatrix}$$

El equipamiento utilizado es el que ya hemos indicado en el capítulo anterior, PC con procesador Pentium II. Las subrutinas fueron trabajadas en FORTRAN y se ha empleado LINPACK para la descomposición en valores singulares.

4.2 El problema triangular

Consideraremos en particular el caso en que el conjunto de factibilidad se compone de matrices Toeplitz triangulares. Recordemos que hemos llamado $\mathcal{T}_{\mathtt{s}}$ al subespacio de matrices Toeplitz triangular superior, y $\mathcal{T}_{\mathtt{i}}$ al subespacio de matrices Toeplitz triangular inferior.

4.2.1 Problema Toeplitz triangular superior

Consideremos el siguiente problema de minimización:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T}_{\mathtt{s}}, \end{array} \right. \tag{4.6}$$

con A y B en las condiciones indicadas al comenzar este capítulo, $\mathcal{T}_{\mathtt{s}}$ es el subespacio de matrices triangulares superiores con estructura Toeplitz. Es decir, si $X \in \mathcal{T}_{\mathtt{s}}$,

entonces

$$X = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \cdots & \cdots & \alpha_n \\ 0 & \alpha_1 & \alpha_2 & \cdots & \cdots & \cdots & \alpha_{n-1} \\ 0 & 0 & \alpha_1 & \alpha_2 & \cdots & \cdots & \alpha_{n-2} \\ : & \cdots & \cdots & \cdots & & & : \\ : & & & \cdots & \cdots & \cdots & : \\ 0 & & & & \cdots & \alpha_1 & \alpha_2 \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & \alpha_1 \end{bmatrix}.$$

Debe observarse que $\mathcal{T}_{\mathtt{s}}$, puede ser expresado:

$$\mathcal{T}_{\mathtt{s}} = \left\{ X \in \mathbb{R}^{n\times n} \ : \ X = \sum_{p=1}^{n} \alpha_p G_p, \right\}$$

donde los $\alpha_p \in \mathbb{R}$ son escalares y las matrices $G_p \in \mathbb{R}^{n\times n}$ están definidas , para $1 \leq i, j, p \leq n$ como sigue:

$$(G_p)_{ij} = \begin{cases} 1 & si \ \ j = i + p - 1 \\ \\ 0 & en \ \ otro \ \ caso. \end{cases} \tag{4.7}$$

Nótese que las matrices G_p se caracterizan por el hecho de que para cada entrada ij, existe un único p, $1 \leq p \leq n$ tal que $(G_p)_{ij} = 1$.

El conjunto de matrices $\left\{G_1, \ G_2, \ \cdots \ , G_n\right\}$, forman una base para el subespacio de las matrices cuadradas, triangulares superiores, con estructura Toeplitz. Esto equivale a decir que $\mathcal{T}_{\mathtt{s}}$ es un subespacio del espacio de matrices cuadradas.

En la misma forma en que lo hicimos en la sección anterior, el problema inicial se transforma, aplicando descomposición en valores singulares, de modo de obtener una solución resolviendo el siguiente problema:

$$\begin{cases} \min & \|T - C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & T \in \mathcal{T}_{\mathtt{s}}', \\ & \end{cases} \tag{4.8}$$

con

$$A = P \begin{bmatrix} \sum \\ 0 \end{bmatrix} Q^T,$$

dónde $P \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$, son matrices ortogonales, y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, y el subespacio

$$\mathcal{T}_{\mathtt{s}}' = \left\{ T \in \mathbb{R}^{n\times n} \ : \ T = \Sigma Q^T \sum_{l=1}^{n} \alpha_l G_l \right\},$$

Caracterización de las soluciones sobre $\mathcal{T}_{\mathtt{s}}{}'$

El teorema que presentamos a continuación, similar al de la sección anterior para matrices Toeplitz densas, caracteriza las proyecciones sobre $\mathcal{T}_{\mathtt{s}}{}'$.

Teorema 4.2.1 *Si $C_1 \in \mathbb{R}^{n\times n}$, entonces, la única solución del problema*

$$\left\{ \begin{array}{ll} \min & \|T - C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & T \in \mathcal{T}_{\mathtt{s}}{}' \end{array} \right. \tag{4.9}$$

está dada por:

$$P_{\mathcal{T}'}(C_1) = \Sigma Q^T \sum_{l=1}^{n} \alpha^\star{}_l G_l, \qquad \alpha^\star{}_l = \frac{\sum_{i=1}^{n}\sum_{j\geq l}^{n}(C_1)_{ij}Q_{(j-l+1)i}}{\sum_{i=1}^{n}\sum_{j\geq l}^{n}(Q_{(j-l+1)i})^2{\sigma_i}^2},$$

para $1 \leq l \leq n$.

Demostración :

Es análoga a la del Teorema 4.1.1, teniendo en cuenta que en la expresión de la derivada

$$\frac{\partial f}{\partial \alpha_p}(\alpha) = \left(\sum_{i,j=1}^{n} \left(\sum Q^T (\sum_{l=1}^{n}(\alpha_l G_l) \right)_{ij} - A_{ij} \right) \left(\sum Q^T G_p \right)_{ij},$$

el elemento $\left(\sum Q^T (\sum_{l=1}^{n}(\alpha_l G_l) \right)_{ij}$ es el producto escalar entre la i-ésima fila de $(\sum Q^T)$ y la j-ésima columna de $\sum_{l=1}^{n}(\alpha_l G_l)$, se obtiene que

$$\left(\sum Q^T \sum_{l=1}^{n}(\alpha_l G_l) \right)_{ij} = \sigma_i \sum_{k=1}^{j} Q_{(j-k+1)i}\alpha_k.$$

De manera similar, el elemento $\left(\sum Q^T G_p \right)_{ij}$, es el producto escalar entre la fila i de $\sum Q^T$ y la columna j de G_p, y por lo tanto:

$$\left(\sum Q^T G_p \right)_{ij} = \left\{ \begin{array}{ll} \sigma_i Q_{(j-p+1)i} & si \quad j \geq p \\ & \\ 0 & si \quad j < p. \end{array} \right.$$

□

4.2.2 Problema Toeplitz triangular inferior

Consideremos el siguiente problema de minimización:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T}_{\mathtt{i}}, \\ & \end{array} \right. \tag{4.10}$$

A y B como en la sección anterior y $\mathcal{T}_{\mathtt{i}}$ el subespacio de matrices cuadradas, triangulares inferiores, con estructura Toeplitz. Es decir si, $X \in \mathcal{T}_{\mathtt{s}}$, entonces

$$X = \left[\begin{array}{ccccccc} \beta_1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \beta_2 & \beta_1 & 0 & \cdots & \cdots & \cdots & 0 \\ \beta_3 & \beta_2 & \beta_1 & 0 & \cdots & \cdots & 0 \\ : & \cdots & \cdots & \cdots & & & : \\ : & & & \cdots & \cdots & \cdots & : \\ \beta_{n-1} & & & & \cdots & \beta_1 & 0 \\ \beta_n & \beta_{n-1} & \cdots & \cdots & \beta_3 & \beta_2 & \beta_1 \end{array} \right],$$

Debe observarse que $\mathcal{T}_{\mathtt{i}}$, puede ser expresado:

$$\mathcal{T}_{\mathtt{i}} = \left\{ X \in \mathbb{R}^{n \times n} \ : \ X = \sum_{p=1}^{n} \beta_p G_p, \right\}$$

donde los $\beta_p \in \mathbb{R}$ son escalares y las matrices $G_p \in \mathbb{R}^{n \times n}$, con similares propiedades que las definidas en secciones anteriores, están definidas para $1 \leq i, j, p \leq n$ como sigue:

$$(G_p)_{ij} = \left\{ \begin{array}{ll} 1 & si \ \ j = i - p + 1 \\ & \\ 0 & en \ \ otro \ \ caso. \end{array} \right. \tag{4.11}$$

El conjunto de matrices $\left\{ G_1, \ G_2, \ \cdots \ , G_n \right\}$, forman una base para el subespacio de las matrices cuadradas, triangulares inferiores, con estructura Toeplitz. Es decir que $\mathcal{T}_{\mathtt{i}}$ es un subespacio del espacio de matrices cuadradas.

También en este caso, se transforma el problema inicial aplicando descomposición en valores singulares, de modo de obtener una solución resolviendo:

$$\left\{ \begin{array}{ll} \min & \|T - C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & T \in \mathcal{T}_{\mathtt{i}}', \\ & \end{array} \right. \tag{4.12}$$

con

$$A = P \left[\begin{array}{c} \sum \\ 0 \end{array} \right] Q^T,$$

dónde $P \in \mathbb{R}^{m\times m}$ y $Q \in \mathbb{R}^{n\times n}$, son matrices ortogonales, y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, y el subespacio

$$\mathcal{T}_{\mathtt{i}}{}' = \left\{ T \in \mathbb{R}^{n\times n} \ : \ T = \Sigma Q^T \sum_{l=1}^{n} \alpha_l G_l \right\},$$

Caracterización de las soluciones sobre $\mathcal{T}_{\mathtt{i}}{}'$

El teorema siguiente, análogo al de las dos secciones anteriores, caracteriza las proyecciones sobre $\mathcal{T}_{\mathtt{i}}{}'$.

Teorema 4.2.2 *Si $C_1 \in \mathbb{R}^{n\times n}$, entonces, la única solución del problema*

$$\left\{ \begin{array}{ll} \min & \|T - C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & T \in \mathcal{T}_{\mathtt{i}}{}' \end{array} \right. \tag{4.13}$$

está dada por:

$$P_{\mathcal{T}'}(C_1) = \Sigma Q^T \sum_{l=1}^{n} \beta^\star{}_l G_l, \qquad \beta^\star{}_l = \frac{\displaystyle\sum_{i=1}^{n}\sum_{j<l} (C_1)_{ij} Q_{(j+l-1)i}}{\displaystyle\sum_{i=1}^{n}\sum_{j<l} (Q_{(j+l-1)i})^2 {\sigma_i}^2},$$

para $1 \leq l \leq n$.

Demostración :

Es análoga a la del Teorema 4.1.1, teniendo en cuenta que en la expresión de la derivada

$$\frac{\partial f}{\partial \alpha_p} = \left(\sum_{i,j=1}^{n} \left(\sum Q^T (\sum_{l=1}^{n} (\beta_l G_l) \right)_{ij} - A_{ij} \right) \left(\sum Q^T G_p \right)_{ij},$$

el elemento $\left(\sum Q^T (\sum_{l=1}^{n} (\beta_l G_l)\right)_{ij}$ es el producto escalar entre la i-ésima fila de $(\sum Q^T)$ y la j-ésima columna de $\sum_{l=1}^{n} (\beta_l G_l)$, se obtiene que

$$\left(\sum Q^T \sum_{l=1}^{n} (\beta_l G_l) \right)_{ij} = \sigma_i \sum_{k=1}^{j} Q_{(j+k-1)i} \beta_k.$$

De manera similar, el elemento $\left(\sum Q^T G_p\right)_{ij}$, es el producto escalar entre la fila i de $\sum Q^T$ y la columna j de G_p, y por lo tanto:

$$\left(\sum Q^T G_p \right)_{ij} = \left\{ \begin{array}{ll} \sigma_i Q_{(j-p+1)i} & si \quad j \leq p \\ & \\ 0 & si \quad j > p. \end{array} \right.$$

□

4.2.3 Experiencia Numérica

Hemos resuelto los ejemplos de la sección 4.1, pero proyectando sobre los conjuntos $\mathcal{T}_{\mathtt{s}}$ y $\mathcal{T}_{\mathtt{i}}$, obteniendo en cada caso:

Ejemplo 4.2.1 Considerando A y B como en el ejemplo 3.2.1, hemos obtenido los siguientes resultados.

a) Proyección sobre el subespacio de las matrices triangulares superiores con estructura Toeplitz:

$$X_\star = \begin{bmatrix} 1.67 \times 10^{-1} & 0 & 0 \\ 0 & 1.67 \times 10^{-1} & 0 \\ 0 & 0 & 1.67 \times 10^{-1} \end{bmatrix}$$

b) Proyección sobre el subespacio de las matrices triangulares inferiores con estructura Toeplitz:

$$X_\star = \begin{bmatrix} 2.62 \times 10^{-1} & 0 & 0 \\ 9.47 \times 10^{-1} & 2.62 \times 10^{-1} & 0 \\ 3.75 \times 10^{-1} & 9.47 \times 10^{-1} & 2.62 \times 10^{-1} \end{bmatrix}$$

Ejemplo 4.2.2 Sean ahora A y B como en el ejemplo 3.2.2, hemos obtenido:

a) Proyección sobre el subespacio de las matrices triangulares superiores con estructura Toeplitz:

$$X_\star = \begin{bmatrix} 5.05 \times 10^{-3} & 0 & 0 & 0 \\ 0 & 5.05 \times 10^{-3} & 0 & 0 \\ 0 & 0 & 5.05 \times 10^{-3} & 0 \\ 0 & 0 & 0 & 5.05 \times 10^{-3} \end{bmatrix}$$

b) Proyección sobre el subespacio de las matrices triangulares inferiores con estructura Toeplitz:

$$X_\star = \begin{bmatrix} 9.42 \times 10^{-1} & 0 & 0 & 0 \\ 1.09 \times 10^{-1} & 9.42 \times 10^{-1} & 0 & 0 \\ 1.08 \times 10^{-1} & 1.09 \times 10^{-1} & 9.42 \times 10^{-1} & 0 \\ 1.57 \times 10^{-1} & 1.08 \times 10^{-1} & 1.09 \times 10^{-1} & 9.42 \times 10^{-1} \end{bmatrix}$$

Ejemplo 4.2.3 Finalmente, considerando A y B como en el ejemplo 3.2.3, hemos obtenido:

a) Proyección sobre el subespacio de las matrices triangulares superiores con estructura Toeplitz:

$$\begin{bmatrix} -3.53\times10^{-2} & 8.49\times10^{-3} & 2.03\times10^{-2} & 1.15\times10^{-2} & 4.34\times10^{-2} & 0 \\ 0 & -3.53\times10^{-2} & 8.49\times10^{-3} & 2.03\times10^{-2} & 1.15\times10^{-2} & 4.34\times10^{-2} \\ 0 & 0 & -3.53\times10^{-2} & 8.49\times10^{-3} & 2.03\times10^{-2} & 1.15\times10^{-2} \\ 0 & 0 & 0 & -3.53\times10^{-2} & 8.49\times10^{-3} & 2.03\times10^{-2} \\ 0 & 0 & 0 & 0 & -3.53\times10^{-2} & 8.49\times10^{-3} \\ 0 & 0 & 0 & 0 & 0 & -3.53\times10^{-2} \end{bmatrix}$$

b) Proyección sobre el subespacio de las matrices triangulares inferiores con estructura Toeplitz:

$$\begin{bmatrix} -4.06\times10^{-2} & 0 & 0 & 0 & 0 & 0 \\ -2.97\times10^{-2} & -4.06\times10^{-2} & 0 & 0 & 0 & 0 \\ -1.79\times10^{-2} & -2.97\times10^{-2} & -4.06\times10^{-2} & 0 & 0 & 0 \\ 6.02\times10^{-2} & -1.79\times10^{-2} & -2.97\times10^{-2} & -4.06\times10^{-2} & 0 & 0 \\ -1.78\times10^{-3} & 6.02\times10^{-2} & -1.79\times10^{-2} & -2.97\times10^{-2} & -4.06\times10^{-2} & 0 \\ -2.27\times10^{-2} & -1.78\times10^{-3} & 6.02\times10^{-2} & -1.79\times10^{-2} & -2.97\times10^{-2} & -4.06\times10^{-2} \end{bmatrix}$$

4.3 Problema simétrico de Toeplitz ó problema doblemente simétrico

Nos ocupa ahora el siguiente problema:

$$\begin{cases} \min & \|AX-B\|_{\mathsf{F}}^2 \\ s.a & \\ & X\in\mathcal{T}\cap\mathcal{S}, \end{cases} \tag{4.14}$$

A y B en las condiciones indicadas al comenzar el capítulo, $\mathcal{T}$ es el subespacio de las matrices de Toeplitz y $\mathcal{S}$ es el subespacio de las matrices simétricas. Es decir que la idea es obtener como solución una matriz que sea, al mismo tiempo, simétrica y Toeplitz, para ser más precisos, una matriz simétrica con respecto a las dos diagonales.

Recordemos que anteriormente hemos presentado el problema simétrico de Procrusto, y resuelto el problema Toeplitz. Se han indicado además distintas versiones de algoritmos de Escalante- Raydán para resolver problemas en los que la región de factibilidad es la intersección entre el subespacio $\mathcal{S}$ con otros conjuntos. Para resolver (4.14), combinaremos la caracterización para la solución del problema simétrico empleada por Escalante- Raydán con nuestros resultados para el problema Toeplitz,

generando una versión propia del algoritmo de proyecciones alternas, que obtiene la solución proyectando, alternativamente, sobre los subespacios $\mathcal{T}$ y $\mathcal{S}$.

Dado el problema

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{S}, \end{array} \right. \tag{4.15}$$

Escalante- Raydán, inspirados por Higham, sugieren transformar el problema aplicando descomposición en valores singulares, obteniendo el problema equivalente:

$$\left\{ \begin{array}{ll} \min & \|Z - C_1\|_{\mathtt{F}}^2 \\ s.a & \\ & Z \in \mathcal{S}', \end{array} \right. \tag{4.16}$$

siendo

$$Z = \sum Y,\ Y = Q^T X Q,$$

$$C = \left[\begin{array}{c} C_1 \\ C_2 \end{array} \right] = P^T B Q,$$

$$C_1 \in \mathbb{R}^{n \times n}$$

y la región de factibilidad

$$\mathcal{S}' = \left\{ Z \in \mathbb{R}^{n \times n},\ :\ Z = \sum Y, Y = Y^T \right\}.$$

de modo que si la solución del último problema es

$$P_{\mathcal{S}'}(C_1) = \sum Y_\star = \sum (Q^T X_\star Q),$$

con $Y_\star$ dada por (2.2); entonces se tiene que la solución del primero es:

$$X_\star = Q \sum^{-1} P_{\mathcal{S}'}(C_1) Q^T.$$

Por otra parte, al comienzo del capítulo nos hemos planteado resolver:

$$\left\{ \begin{array}{ll} \min & \|AX - B\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T}, \end{array} \right. \tag{4.17}$$

siguiendo la estrategia sugerida por Higham hemos transformado el problema, obteniendo:

$$\left\{ \begin{array}{ll} \min & \|T - \bar{C}_1\|_{\mathtt{F}}^2 \\ s.a & \\ & T \in \mathcal{T}'. \end{array} \right. \tag{4.18}$$

con

$T = \sum Q^T X \in \mathbb{R}^{n\times n}$,

$\bar{C} = \begin{bmatrix} \bar{C}_1 \\ \bar{C}_2 \end{bmatrix} = P^T B$,

$\bar{C}_1 \in \mathbb{R}^{n\times n}$.

$$\mathcal{T}' = \left\{ T \in \mathbb{R}^{n\times n} \ : \ T = \sum Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l G_l \right\}.$$

Demostrado el Teorema 4.1.1, que caracteriza las soluciones sobre $\mathcal{T}'$, se sabe que

$$P_{\mathcal{T}'}(C_1) = \Sigma Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha^\star{}_l G_l, \qquad \alpha^\star{}_l = \frac{\sum_{i=1}^{n} \sum_{j>l} (C_1)_{ij} Q_{(j-l)i}}{\sum_{i=1}^{n} \sum_{j>l} (Q_{(j-l)i})^2 {\sigma_i}^2}.$$

Combinando ambos resultados, podemos establecer que resolver el problema (4.14), es equivalente a resolver

$$\left\{ \begin{array}{ll} \min & \|Z - C_1\|_{\mathrm{F}}^2 \\ s.a & \\ & Z \in \mathcal{T}'' \cap \mathcal{S}'. \end{array} \right. \tag{4.19}$$

siendo

$$\mathcal{T}'' = \left\{ Z \in \mathbb{R}^{n\times n} \ : \ Z = \sum Q^T \sum_{l=-(n-1)}^{(n-1)} \alpha_l G_l Q \right\}.$$

Debe observarse, que la necesidad de considerar el conjunto $\mathcal{T}''$, se origina en el hecho de que ambos problemas, el simétrico y el toeplitz, tengan la misma función objetivo.

El algoritmo de proyecciones alternas que hemos diseñado para encontrar una solución de (4.19) es el siguiente:

Algoritmo 4.3.1 *Dada* $X_0 = C_1 \in \mathbb{R}^{n\times n}$,
para i=0,1,2,...,hasta convergencia, repetir

$X_i = P_{\mathcal{S}'}(X_i)$

$X_{i+1} = P_{\mathcal{T}''}(X_i)$.

Es preciso aclarar que el algoritmo termina cuando dos proyecciones consecutivas sobre un mismo subespacio son muy cercanas, y que su convergencia está garantizada por los resultados de convergencia para métodos de proyecciones alternas sobre intersección de subespacios [10].

Capítulo 5

Aplicaciones

Nos proponemos ahora aplicar los resultados del capítulo anterior al problema de aproximar la matriz inversa de una matriz dada, no singular y persimétrica. Para ello debe tenerse en cuenta que el conjunto de factibilidad del problema debe contener matrices no singulares y persimétricas, pues como veremos a continuación, la inversa de toda matriz persimétrica es persimétrica, y por supuesto, no singular. Una vez definida la función objetivo y el conjunto factible, se caracterizarán las soluciones, para establecer entonces el algoritmo de proyecciones alternas que de solución al problema de aproximar la inversa.

Dicha aproximación será finalmente aplicada a cierta clase de problemas de programación cuadrática, a resolver por el método de gradiente reducido.

5.1 Aproximación de la matriz inversa de una matriz persimétrica dada

Dada $A \in \mathbb{R}^{n\times n}$, no singular y persimétrica. Interesa obtener una aproximación de A^{-1}, por lo tanto, necesitamos una matriz $X^\star$ tal que el producto por A se aproxime a la identidad. Representaremos a esa situación exigiendo que $X^\star$ minimice la norma de Frobenius de A por X menos la identidad. Sea entonces el siguiente problema de minimización:

$$\left\{\begin{array}{ll} \min & \|AX - I\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in P. \end{array}\right. \tag{5.1}$$

Ya que el objetivo es aproximar la inversa de A, necesitamos que el conjunto P esté conformado por matrices no singulares y persimétricas, pues la inversa de toda matriz persimétrica es persimétrica, en efecto:

Proposición 5.1.1 *Sea $A \in \mathbb{R}^{n\times n}$ no singular y persimétrica, entonces A^{-1} es persimétrica.*

Demostración :

Si A es persimétrica, entonces

$$A = EA^TE,$$

y ya que para toda A inversible $(A^{-1})^T = (A^T)^{-1}$, se tiene que:

$$A^{-1} = (EA^TE)^{-1} = E(A^T)^{-1}E = E(A^{-1})^TE,$$

lo cual significa que A^{-1} es persimétrica. □

Es decir que necesitamos como solución una matriz no singular y persimétrica. Una opción es elegir $P = \mathcal{T}_{\mathtt{s}}$, pues las matrices de Toeplitz forman una subclase de la clase de las matrices persimétricas, y de tal forma estaríamos seguros de que la solución es persimétricas; pero no existe ninguna garantía de que dichas matrices sean no singulares, a menos que se exija que los elementos sobre la diagonal sean no nulos.

Para ello sea $\epsilon \in \mathbb{R}$, $\epsilon > 0$, y considérese el siguiente subconjunto en $\mathbb{R}^{n\times n}$:

$$\Gamma = \left\{ X \in \mathbb{R}^{n\times n} \ : \ X_{ii} \geq \epsilon \right\},$$

formado por matrices cuadradas con elementos no nulos sobre la diagonal.

Elegiremos entonces $P = \mathcal{T}_{\mathtt{s}} \cap \Gamma$, y nuestro problema queda planteado como sigue:

$$\left\{ \begin{array}{ll} \min & \|AX - I\|_{\mathtt{F}}^2 \\ s.a & \\ & X \in \mathcal{T}_{\mathtt{s}} \cap \Gamma. \end{array} \right. \tag{5.2}$$

Se sabe que $\mathcal{T}_{\mathtt{s}}$ es un subespacio. Estudiemos las propiedades de Γ:

Proposición 5.1.2 *Para cada ϵ, el conjunto Γ es convexo.*

Demostración :

Sea $\epsilon > 0$ y sean X, e Y pertenecientes a Γ, entonces como: $X_{ii} \geq \epsilon$, y $Y_{ii} \geq \epsilon$, la combinación lineal convexa de X e Y, $\mu X + (1-\mu)Y$ es una matriz cuadrada y tal que:

$$\begin{array}{rcl} \mu X_{ii} + (1-\mu)Y_{ii} & \geq & \mu\epsilon + (1-\mu)\epsilon \\ & = & \mu\epsilon + \epsilon - \mu\epsilon \\ & = & \epsilon \end{array}$$

lo que prueba que $\mu X + (1-\mu)Y \in \Gamma$ □

Proposición 5.1.3 *Para cada ϵ, el conjunto Γ es cerrado.*

Demostración :

Sea $\{T_m\}$ una sucesión en Γ tal que converge a T^*. Se probará que $T^* \in \Gamma$:
Si $T_m \in \Gamma$ para todo m, entonces $(T_m)_{ii} \geq \epsilon$ para todo m. Por lo tanto, para cada i tal que $1 \leq i \leq n$, la sucesión $\{(T_m)_{ii}\}$ está acotada inferiormente por ϵ, y además

$$\{(T_m)_{ii}\}_m \to (T^*)_{ii},$$

necesariamente es $(T^*)_{ii} \geq \epsilon$ y por lo tanto $T^* \in \Gamma$. □

5.1.1 Transformación del Problema

El problema inicial se transforma, aplicando descomposición en valores singulares, de modo de obtener una solución a partir del siguiente problema:

$$\left\{ \begin{array}{ll} \min & \|T - P^T\|_{\mathtt{F}}^2 \\ s.a & \\ & T \in \mathcal{T}_{\mathtt{s}}' \cap \Gamma', \\ & \end{array} \right. \tag{5.3}$$

pues se tiene, :

$$\begin{aligned} \|AX - I\|_{\mathtt{F}}^2 &= \|(P\sum Q^T)X - I\|_{\mathtt{F}}^2 = \|P^T(P\sum Q^T)X - I]Q\|_{\mathtt{F}}^2 \\ &= \|\sum(Q^TXQ) - P^TQ\|_{\mathtt{F}}^2 = \|[\sum(Q^TXQ) - P^TQ]Q^T\|_{\mathtt{F}}^2 \\ &= \|\sum(Q^TX) - P^T\|_{\mathtt{F}}^2 = \|T - P^T\|_{\mathtt{F}}^2 \end{aligned}$$

siendo $A = P\sum Q^T$, dónde $P \in \mathbb{R}^{n\times n}$ y $Q \in \mathbb{R}^{n\times n}$, son matrices ortogonales, y $\sum = diag(\sigma_1, \cdots, \sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

Los conjuntos $\mathcal{T}_{\mathtt{s}}'$ y Γ' son respectivamente, los convexos $\mathcal{T}_{\mathtt{s}}$ y Γ "transformados", es decir:

$$\mathcal{T}_{\mathtt{s}}' = \left\{ T \in \mathbb{R}^{n\times n}, \, : \, T = \Sigma Q^T \sum_{l=1}^{n} \alpha_l \sigma_l \right\}, \tag{5.4}$$

$$\Gamma' = \left\{ Z \in \mathbb{R}^{n\times n}, \, : \, (Q\Sigma^{-1}Z)_{ii} \geq \epsilon \right\}. \tag{5.5}$$

Se sabe que $\mathcal{T}_{\mathtt{s}}'$ es un subespacio. En las dos proposiciones siguientes, estudiaremos las propiedades de Γ'.

Proposición 5.1.4 *Γ' es cerrado.*

Demostración :

Sea $\{T_m\}$ una sucesión en Γ' tal que converge a T^*. Se probará que $T^* \in \Gamma'$:
Si $T_m \in \Gamma$ para todo m, entonces

$$\sum_{k=1}^{n} \frac{Q_{ik}(T_m)_{ki}}{\sigma_k} \geq \epsilon, \qquad \forall m$$

luego, como

$$\lim_{n\to\infty} T_m = T^*,$$

tomando límite :

$$\lim_{m\to\infty} \sum_{k=1}^{n} \frac{Q_{ik}(T_m)_{ki}}{\sigma_k} = \sum_{k=1}^{n} \lim_{n\to\infty} \frac{Q_{ik}(T_m)_{ki}}{\sigma_k} = \sum_{k=1}^{n} \frac{Q_{ik}}{\sigma_k} \lim_{n\to\infty} (T_m)_{ki} = \sum_{k=1}^{n} \frac{Q_{ik}(T^*)_{ki}}{\sigma_k}.$$

Por lo tanto, para cada i tal que $1 \leq i \leq n$, la sucesión $\left\{\sum_{k=1}^{n} \frac{Q_{ik}(T_m)_{ki}}{\sigma_k}\right\}$ está acotada inferiormente por ϵ, y además es convergente a $(T^*)_{ii}$, necesariamente el límite es también mayor que ϵ. Luego $T^* \in \Gamma'$. □

Proposición 5.1.5 *El conjunto de matrices Γ' es un conjunto convexo.*

Demostración :

Γ' es convexo.
Si $T_1, T_2 \in \Gamma'$ y $0 \leq \mu \leq 1$, entonces

$$\mu T_1 + (1-\mu)T_2 \in \Gamma'.$$

Si $T_1 \in \Gamma'$ entonces

$$\sum_{k=1}^{n} \frac{Q_{ik}(T_1)_{ki}}{\sigma_k} \geq \epsilon,$$

y si $T_2 \in \Gamma'$ entonces

$$\sum_{k=1}^{n} \frac{Q_{ik}(T_2)_{ki}}{\sigma_k} \geq \epsilon,$$

luego:

$$\begin{aligned}(\mu T_1 + (1-\mu)T_2)_{ii} &= \sum_{k=1}^{n} \frac{Q_{ik}(\mu T_1 + (1-\mu)T_2)_{ki}}{\sigma_k} = \sum_{k=1}^{n} \frac{Q_{ik}[(\mu T_1)_{ki} + ((1-\mu)T_2)_{ki}]}{\sigma_k} \\ &= \sum_{k=1}^{n} \frac{Q_{ik}(\mu T_1)_{kj}}{\sigma_k} + \sum_{k=1}^{n} \frac{Q_{ik}((1-\mu)T_2)_{kj}}{\sigma_k} \geq \epsilon + \epsilon = 2\epsilon > \epsilon\end{aligned}$$

Luego, como $(\mu T_1 + (1-\mu)T_2)_{ii} \geq \epsilon$, $(\mu T_1 + (1-\mu)T_2)$ pertenece a Γ'y por lo tanto Γ' es convexo. □

5.1.2 Caracterización de las soluciones

Ya hemos caracterizado en 4.2.1 las proyecciones sobre $\mathcal{T}_s'$, indicaremos con el siguiente teorema cómo son las proyecciones sobre Γ'.

Teorema 5.1.1 *Sean $\epsilon > 0$ y $C \in \mathbb{R}^{n\times n}$, entonces, la única solución del problema*

$$\begin{cases} \min & \|T - C\|_F^2 \\ s.a & \\ & T \in \Gamma', \end{cases} \tag{5.6}$$

para Γ' dado por (5.5) está dada por:

$$P_{\Gamma'}(C) = \Sigma Q^T M,$$

donde $M \in \mathbb{R}^{n\times n}$ está definida como sigue:

$$(M)_{ij} = \begin{cases} (Q\Sigma^{-1}C)_{ij} & si \quad i \neq j \\ (Q\Sigma^{-1}C)_{ii} & si \quad i = j \quad y \quad (Q\Sigma^{-1}C)_{ii} > \epsilon \\ \epsilon & si \quad i = j \quad y \quad (Q\Sigma^{-1}C)_{ii} < \epsilon \end{cases} \tag{5.7}$$

para $i, j = 1, \cdots, m$.

Demostración :

Puesto que el convexo es

$$\Gamma' = \left\{ T \in \mathbb{R}^{n\times n} \ : \ (Q\Sigma^{-1}C)_{ii} \geq \epsilon \right\}.$$

la demostración es evidente. □

Observación:

- Debe notarse que si $(Q\Sigma^{-1}C)_{ij} \geq \epsilon$, para todo ij, entonces $M = Q\Sigma^{-1}C$ y por lo tanto
$$P_{\Gamma'}(C) = \Sigma Q^T (Q\Sigma^{-1}C) = C.$$

Teniendo ya caracterizadas las proyecciones sobre ambos convexos, daremos nuestra versión del algoritmo de proyecciones alternas para este caso.

5.1.3 El algoritmo

Nuestra versión del algoritmo de proyecciones alternas para el problema

$$min\{||T - P^T||_F^2 \ : \ T \in \mathcal{T}_s' \cap \Gamma'\}. \tag{5.8}$$

es la siguiente:

Algoritmo 5.1.1 *Dada* $P^T \in \mathbb{R}^{n\times n}$, *sea* $X_0 = P^T$, $I_{\Gamma'}^0 = 0$,
para $i = 1, 2, \cdots$, *hasta convergencia, repetir*

$$X_i = P_{\mathcal{T}_s'}(X_i) - I_{\Gamma'}^i$$

$$I_{\Gamma'}^{i+1} = P_{\Gamma'}(X_i) - X_i$$

$$X_{i+1} = P_{\Gamma'}(X_i)$$

El teorema siguiente establece el resultado de convergencia para el algoritmo anterior. Su demostración se sigue directamente de un resultado de Boyle y Dykstra [1]

Teorema 5.1.2 *Si el conjunto cerrado y convexo* $\mathcal{T}_s' \cap \Gamma'$ *es no vacío, para toda* $P^T \in \mathbb{R}^{n\times n}$ *las sucesiones* $\{P_{\mathcal{T}_s'}(X_i)\}$ *y* $\{P_{\Gamma'}(X_i)\}$ *generadas por el algoritmo, convergen en norma de Frobenius a una solución de (5.8).*

Observaciones:

- El algoritmo termina cuando dos proyecciones consecutivas sobre un mismo conjunto son muy cercanas. Por ejemplo, establecida una tolerancia $\mathcal{T}l$ por el usuario, se detiene el proceso iterativo si:

$$\|P_{\mathcal{T}_s'}(X_{i+1}) - P_{\mathcal{T}_s'}(X_i)\|_F < \mathcal{T}l.$$

- Es posible plantear una versión similar del algoritmo considerando como conjunto de factibilidad a la intersección $\mathcal{T}_i \cap \Gamma$.

En la próxima sección se sugiere una aplicación de esta estrategia de aproximación de la inversa de una matriz persimétrica no singular dada, a un problema de optimización con restricciones de igualdad, con la particularidad de que el Jacobiano de las restricciones puede ser particionado en dos bloques uno de los cuales es una matriz no singular y persimétrica.

5.1.4 Aplicación a un problema de optimización

Consideremos el siguiente problema de minimización con restricciones de igualdad

$$(mri) = \begin{cases} \min & f(x) \\ s.a & \\ & C(x) = 0, \end{cases} \tag{5.9}$$

donde $f : \mathbb{R}^n \to \mathbb{R}$, y $C : \mathbb{R}^n \to \mathbb{R}^m$, $m < n$, muy grandes.

Una manera muy difundida de resolverlo consiste en aplicar el método de programación cuadrática sucesiva. Ya que es de interés trabajar con problemas de gran porte, es conveniente efectuar una reducción al espacio tangente de las restricciones,

$$\mathcal{N}(\nabla C^T)= \Big\{ X \in \mathbb{R}^{m\times m} \ : \ \nabla C^T X = 0 \Big\}.$$

Supongamos que $rank(\nabla C^T) = m$ y consideremos la siguiente partición de la matriz ∇C^T:

$$\nabla C^T = [\ B \ \ | \ \ N \ ]$$

$B \in \mathbb{R}^{m\times m}$ es una submatriz no singular de ∇C^T y $N \in \mathbb{R}^{m\times(n-m)}$.

Practicada la reducción cuya técnica aparece descripta en el Apéndice B, el problema a resolver queda expresado de la siguiente manera:

$$(\bar{Q}) = \left\{ \begin{array}{ll} \min & \bar{q}(\bar{s}) = \frac{1}{2}\bar{s}^T \bar{H} \bar{s} + \bar{h}^T \bar{s} \\ s.a & \\ & \bar{s} \in \mathcal{N}(\nabla C^T). \end{array} \right.$$

Este último problema puede ser resuelto aplicando cualquiera de los algoritmos disponibles para la resolución de problemas cuadráticos. Tanto la eficiencia del algoritmo como la transformación definida al efectuar la reducción, dependen de la elección de la matriz $\mathcal{W}$,

$$\mathcal{W} = \left(\begin{array}{c} -B^{-1}N \\ I_{n-m} \end{array} \right).$$

Ya que $\mathcal{W}$, cuyas columnas forman una base del espacio $\mathcal{N}(\nabla C^T)$, depende de B la elección de una dependerá de la forma en que la otra sea seleccionada, en particular de la forma en que B^{-1} sea obtenida.

En la sección anterior hemos desarrollado un algoritmo para obtener una aproximación de B^{-1}, a partir de la matriz cuadrada B no singular y persimétrica. Supongamos un problema de minimización en el cual el Jacobiano ∇C^T tiene un bloque B que es una matriz no singular con estructura persimétrica. Sea $X^\star$ la aproximación de la inversa de B según el algoritmo de la sección anterior, queremos que la matriz

$$\mathcal{W}^\star = \left(\begin{array}{c} -X^\star N \\ I_{n-m} \end{array} \right),$$

sea tal que, en algún sentido, se "aproxime" a la verdadera $\mathcal{W}$.

Para ejemplificar, supongamos que se tiene el siguiente problema de minimización con restricciones de igualdad

$$\left\{ \begin{array}{ll} \min & x_1^2 + x_2^2 + x_3^2 - 2x_3x_4 + 2x_5x_6 + x_7 \\ s.a & \\ & 3x_1 + x_2 + 2x_3^2 + x_4^2 - 2x_5 + x_6 + x_7^2 = 0 \\ & -2x_1 + 2x_2 + \frac{1}{1}(x_3+1)^2 + 2x_4^2 + \frac{1}{2}(x_5-1)^2 + x_6 + (x_7-1)^2 = 0 \\ & (x_1-1)^2 + x_2^2 + \frac{1}{1}(x_3+1)^2 + x_4^2 + x_5 + x_6 = 0 \\ & x_1 + x_2^2 + \frac{1}{1}(x_3+1)^2 + 2x_4^2 + x_5^2 + x_6^2 - x_7^2 = 0 \end{array} \right.$$

siendo el punto inicial $x_0 = (1, 1, 0, 1, 1, 1, 1)$.

Ya que se quiere resolver empleando el método de gradiente reducido, se considera el problema cuadrático

$$\begin{cases} \min \quad \bar{q}(\bar{s}) = \frac{1}{2}\bar{s}^T \bar{H} \bar{s} + \bar{h}^T \bar{s} \\ s.a \\ \qquad\qquad \bar{s} \in \mathcal{N}(\nabla C^T). \end{cases}$$

el cual se obtiene, luego de efectuada la reducción al espacio tangente de las restricciones. Mediante una permutación de columnas adecuada, se tiene en el punto inicial la siguiente partición del Jacobiano de las restricciones

$$\nabla C^T = [\ B \ \mid \ N\ ] = \left[\begin{array}{cccc|ccc} 1 & 0 & -2 & 3 & 2 & 1 & 2 \\ 2 & 1 & 1 & -2 & 4 & 2 & 0 \\ 1 & 1 & 1 & 0 & 2 & 1 & -1 \\ 2 & 1 & 2 & 1 & 4 & 2 & -2 \end{array}\right]$$

El método requiere la matriz inversa del bloque no singular B, que además es, como puede observarse, persimétrico. La idea es emplear, en esta etapa, el algoritmo descripto anteriormente para estimar la inversa de B. Se ha obtenido, al cabo de 10 iteraciones la matriz $\widetilde{B^{-1}}$, aproximación de B^{-1} por nuestro algoritmo:

$$\widetilde{B^{-1}} = E - 005 \left[\begin{array}{cccc} 1.81 & 5.62 & 8.71 & 7.58 \\ 0 & 1.81 & 5.62 & 8.71 \\ 0 & 0 & 1.81 & 5.62 \\ 0 & 0 & 0 & 1.81 \end{array}\right]$$

Creemos que la anterior es una aplicación interesante. Como indicaremos en nuestras conclusiones, otra posible aplicación tiene que ver con precondicionadores para sistemas Toeplitz.

5.2 Conclusiones

Han sido resueltos los problemas persimétrico y Toeplitz de Procrusto, y sus soluciones han sido aplicadas a problemas concretos. Una nueva estrategia que emplea métodos de proyecciones alternas para resolver el problema de aproximar la inversa de una matriz persimétrica ha sido presentada. La aproximación de la inversa fue utilizada concretamente en un algoritmo para resolver cierta clase de problemas de programación cuadrática, utilizando la técnica de gradiente reducido.

La estrategia desarrollada se basa en las propuestas por Escalante y Raydán y dos nuevos algoritmos que resuelven los problemas persimétrico y Toeplitz de Procrusto fueron presentados.

La experiencia numérica presentada es breve pues nuestro trabajo es de interés fundamentalmente teórico, aunque se planea ampliar la experimentación, para poder

obtener más conclusiones sobre la aplicabilidad y utilidad de nuestros resultados. Es nuestro objetivo continuar con esta linea de trabajo, en particular en lo que se refiere a problemas que involucran matrices de Toeplitz.

Los sistemas Toeplitz aparecen en una gran variedad de aplicaciones, por ejemplo en análisis de Series Temporales, en teoría de Ecuaciones Diferenciales ó en Teoría de Control. Pero más nos ha intersado toda la temática relacionada con precondicionadores para sistemas Toeplitz. En base a la lectura de un trabajo de Chan y Ng de 1996 [2], en el cual trabajan con el método de gradiente conjugado para sistemas Toeplitz, conjeturamos que es posible vincular nuestros resultados con el problema de hallar precondicionadores circulantes para sistemas Toeplitz. Basta con tener en cuenta que una matriz C se dice circulante si

$$C = \begin{bmatrix} C_0 & C_{-1} & C_{-2} & \cdots & \cdots & \cdots & C_{1-n} \\ C_1 & C_0 & C_{-1} & \cdots & \cdots & \cdots & \cdots \\ C_2 & C_1 & C_0 & C_{-1} & \cdots & \cdots & \cdots \\ \vdots & \cdots & \cdots & \cdots & & & \vdots \\ \vdots & & & \cdots & \cdots & \cdots & \vdots \\ C_{n-2} & & & & \cdots & C_0 & C_{-1} \\ C_{n-1} & C_{n-2} & \cdots & \cdots & C_2 & C_1 & C_0 \end{bmatrix},$$

verificándose $\forall\, k$, tal que $1 \leq k \leq n-1$ que

$$C_{-k} = C_{n-k}.$$

Por ejemplo una matriz circulante $C \in \mathbb{R}^{4\times 4}$ es de la forma:

$$C = \begin{bmatrix} a & d & c & b \\ b & a & d & c \\ c & b & a & d \\ d & c & b & a \end{bmatrix}.$$

Resulta claro que si C es circulante, entonces C es Toeplitz pues es constante por diagonales. De allí a que nuestro propósito sea vincular nuestros resultados con el problema de hallar precondicionadores circulantes para sistemas de Toeplitz. Nuestro trabajo futuro se centrará en ese punto específicamente, y probablemente sea el tema de una futura tesis doctoral.

Apéndice A

Conceptos básicos

Definición (Matriz Sparse): De acuerdo a J. H. Wilkinson, una matriz se dice sparse si tiene una cantidad suficiente de elementos nulos como para intentar tomar ventajas de ello.

Definición (Transformación lineal): Sea P una tranformación que a cada $x_i \in \mathbb{R}^m$ lo aplica en $\bar{x_i} \in \mathbb{R}^n$. Se dice que P es lineal si para $\alpha_i \in \mathbb{R}$:

$$P\left(\sum_{i=1}^{p} \alpha_i x_i\right) = \sum_{i=1}^{p} \alpha_i P\left(x_i\right).$$

Definición (Proyección): Sea $\mathcal{P}$ una transformación lineal acotada, se dice que $\mathcal{P}$ es una proyección si satisface:

$$\mathcal{P}^2 = \mathcal{P}.$$

La definición anterior puede ser consultada en [9].

Las definiciones que se darán a continuación pueden ser ampliadas consultando [11]:

Definición (Conjunto Convexo): Se denomina conjunto convexo a un conjunto C de puntos tal que:

Si x_1, $x_2 \in C$,

y $0 \leq \mu \leq 1$, entonces:

$$\mu x_1 + (1 - \mu) x_2 \in C.$$

En otros términos, un conjunto es convexo si al contener a un par de puntos, contiene al segmento que ellos determinan. Por ejemplo, en la recta real, un intervalo $[a, b]$ es un conjunto convexo cerrado. En $\mathbb{R}^n$, una recta, un hiperplano, un semiespacio son conjuntos convexos.

Definición (Combinación Lineal Convexa): Se llama combinación lineal convexa de p puntos x_i de $\mathbb{R}^n$ al punto

$$x = \sum_{i=1}^{p} \mu_i x_i,$$

verificando que $\mu_i \leq 0$ para todo i, y la suma $\sum_{i=1}^{p} \mu_i = 1$.

Definición equivalente de conjunto convexo: Un conjunto C es convexo si y solamente si contiene a todas las combinaciones lineales convexas de sus puntos.

Propiedad: La intersección de conjuntos convexos es un conjunto convexo.

Definición (Hiperplano): Sean $a \in \mathbb{R}^n$, $a \neq 0$, y $\alpha \in \mathbb{R}$, se denomina hiperplano al conjunto determinado por:

$$H = \{x \ : \ ax = \alpha\}.$$

Definición (Semiespacio Cerrado): Sean $a \in \mathbb{R}^n$, $a \neq 0$, y $\alpha \in \mathbb{R}$, se denomina semiespacio cerrado al conjunto:

$$H = \{x \ : \ ax \geq \alpha\}.$$

Definición (Cono): Se dice que el conjunto C es un cono si para todo $x \in C$ y $\lambda > 0$, verifica $\lambda x \in C$.

Es decir que un cono es un conjunto tal que si contiene a un punto, entonces contiene a todas las semirrectas que pasan por él. Por ejemplo, en $\mathbb{R}^n$, todo subespacio vectorial es un cono.

Definición (Cono Convexo): Es un cono que además es un conjunto convexo, es decir:

$x \in C$, $\lambda \geq 0$, entonces $\lambda x \in C$.

x_1, $x_2 \in C$, entonces $x_1 + x_2 \in C$.

Propiedades:

1- El conjunto de todas las combinaciones lineales positivas de un número finito de puntos de $\mathbb{R}^n$ es un cono convexo C:

$$C = \left\{ x \ : \ x = \sum_{i=1}^{p} \lambda_i x_i, \lambda_i \geq 0 \right\}.$$

2- La intersección de un número finito de semiespacios cerrados es un cono convexo S:

$$S = \{x \ : \ x = b_i x \leq 0, \ \ i = 1 : m\}.$$

Definición (Matriz de Toeplitz): Sea $T \in \mathbb{R}^{n\times n}$, se dice que T es una matriz de Toeplitz si existen escalares, $r_{-n+1}, \cdots, r_0, \cdots, r_{n-1}$, tales que $a_{ij} = r_{j-i}$, para todo i, j, así la matriz $T \in \mathbb{R}^{n\times n}$ es una matriz Toeplitz:

$$T = \begin{bmatrix} r_0 & r_1 & r_2 & r_3 & \cdots & r_{n-2} & r_{n-1} \\ r_{-1} & r_0 & r_1 & r_2 & r_3 & \cdots & r_{n-2} \\ r_{-2} & r_{-1} & r_0 & r_1 & r_2 & \cdots & r_{n-3} \\ : & : & : & : & : & \cdots & : \\ : & : & : & : & : & \cdots & : \\ r_{-n+3} & r_{-n+4} & \cdots & r_{-1} & r_0 & r_1 & r_2 \\ r_{-n+2} & r_{-n+3} & \cdots & \cdots & \cdots & r_0 & r_1 \\ r_{-n+1} & r_{-n+2} & \cdots & \cdots & \cdots & r_{-1} & r_0 \end{bmatrix}.$$

Las matrices Toeplitz pertenecen a una clase más amplia de matrices, la de las matrices persimátricas, es decir aquellas que son simétricas respecto a la diagonal NE-SO. En otras palabras, diremos que $B \in \mathbb{R}^{n\times n}$, es persimétrica, si para todo i, j se verifica:

$$b_{i,j} = b_{n-j+1,n-i+1}.$$

Lo anterior es equivalente a requerir que $B = EB^T E$, siendo E la matriz de permutación siguiente:

$$E = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & 0 \\ : & : & : & : & : & \cdots & : \\ : & : & : & : & : & \cdots & : \\ 0 & 0 & 1 & 0 & \ldots & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}.$$

Es muy simple verificar que efectivamente, las matrices Toeplitz son Persimétricas. Una propiedad muy importante de las matrices Toeplitz es que la inversa de una matriz Toeplitz no singular es persimétrica.

Apéndice B

Técnica de reducción al espacio tangente

Conocido el iterado en la iteración actual, x_c, se necesita hallar s_c como solución aproximada de:

$$(Q_c) = \begin{cases} \min & q(s) = \frac{1}{2}s^T H_c s + h_c^T s \\ s.a & \\ & \nabla C_c^T s + C_c = 0. \end{cases}$$

Sin pérdida de generalidad, supondremos que $C_c = 0$ pues de no ser así sería posible pasar a ese caso por medio de traslaciones. Eliminemos los subíndices para facilitar la presentación. Nuestro problema cuadrático es:

$$(Q) = \begin{cases} \min & q(s) = \frac{1}{2}s^T H s + h^T s \\ s.a & \\ & \nabla C^T s = 0, \end{cases}$$

dónde $\nabla C^T \in \mathbb{R}^{m\times n}$, con $m < n$, m y n grandes; H simétrica y definida positiva en $\mathcal{N}(\nabla C^T)$.

La idea es transformar a (Q) en un problema sin restricciones y de menor dimensión, para que el problema de gran porte sea más fácil de tratar.

Supongamos que $rank(\nabla C^T) = m$ y consideremos la siguiente partición de la matriz ∇C^T:

$$\nabla C^T = [\ B \ \ | \ \ N\ ]$$

$B \in \mathbb{R}^{m\times m}$ es una submatriz no singular de ∇C^T y $N \in \mathbb{R}^{m\times(n-m)}$.

Realizando una partición análoga para s:

$$s = (s_B, s_N)^T,$$

resulta que:

$$\nabla C^T s = B s_B + N s_N = 0$$

por lo tanto

$$s_B = -B^{-1} N s_N.$$

Con una adecuada permutación de filas y columnas de H se efectua la siguiente partición de la matriz Hessiana:

$$H = \left[\begin{array}{c|c} H_{11} & H_{12} \\ \hline H_{21} & H_{22} \end{array} \right]$$

y para el gradiente:

$$h^T = (h_B \;\; h_N)$$

De modo que es posible escribir a $q(x)$ en términos de la componente tangencial, es decir en $\mathcal{N}(\nabla C^T)$:

$$\begin{aligned}
q(s) &= \frac{1}{2}s^T H s + h^T s = \frac{1}{2}(s_B, s_N) \left[\begin{array}{c|c} H_{11} & H_{12} \\ \hline H_{21} & H_{22} \end{array} \right] (s_B, s_N)^T \\
&\quad +(h_B \;\; h_N)(s_B, s_N)^T \\
&= \left[s_B^T H_{11} + s_N^T H_{21} \quad s_B^T H_{12} + s_N^T H_{22} \right] (s_B, s_N) \\
&\quad +(h_B^T s_B + h_N^T s_N) \\
&= (s_B^T H_{11} + s_N^T H_{21}) s_B + (s_B^T H_{12} + s_N^T H_{22}) s_N \\
&\quad +h_B^T(-B^{-1}N s_N) + h_N^T s_N \\
&= ((-B^{-1}N s_N)^T H_{11}(-B^{-1}N s_N) + s_N^T H_{21}(-B^{-1}N s_N) \\
&\quad +(-B^{-1}N s_N)^T H_{12} s_N + s_N^T H_{22} s_N \\
&\quad +(h_B^T(-B^{-1}N) + h_N^T) s_N \\
&= s_N^T[(-B^{-1}N)^T H_{11}(-B^{-1}N)] s_N + s_N^T[(-B^{-1}N)^T H_{12}] s_N \\
&\quad +s_N^T[(-H_{21}B^{-1}N)] s_N + s_N^T H_{22} s_N \bar{h}^T s_N \\
&= s_N^T[(-B^{-1}N)^T H_{11}(B^{-1}N) - H_{21}(B^{-1}N) \\
&\quad -(B^{-1}N)^T H_{12} + H_{22}] s_N + \bar{h}^T s_N \\
&= s_N^T \bar{H} s_N + \bar{h}^T s_N.
\end{aligned}$$

Siendo:

$$\bar{H} = (-B^{-1}N)^T H_{11}(B^{-1}N) - H_{21}(B^{-1}N) - (B^{-1}N)^T H_{12} + H_{22}.$$

$$\bar{h} = (h_B^T(-B^{-1}N) + h_N^T)$$

luego la forma cuadrática $q(s)$ queda escrita de la siguiente manera:

$$\bar{q}(s) = \frac{1}{2} s_N^T \bar{H} s_N + \bar{h}^T s_N.$$

Dado que:

$$s = \begin{pmatrix} s_B \\ s_N \end{pmatrix} = \begin{pmatrix} -B^{-1}N s_N \\ s_N \end{pmatrix} = \begin{pmatrix} -B^{-1}N \\ I_{n-m} \end{pmatrix} s_N,$$

si definimos $\mathcal{W} \in \mathbb{R}^{n\times(n-m)}$:

$$\mathcal{W} = \begin{pmatrix} -B^{-1}N \\ I_{n-m} \end{pmatrix},$$

entonces

$$s = \mathcal{W} s_N.$$

Luego:

$$\begin{aligned} q(s_N) &= (\mathcal{W}s_N)^T H(\mathcal{W}s_N) + h^T(\mathcal{W}s_N) \\ &= s_N^T(\mathcal{W}^T H \mathcal{W})s_N + (\mathcal{W}^T h)^T s_N \\ &= \frac{1}{2} s_N^T \bar{H} s_N + \bar{h}^T s_N. \end{aligned}$$

y si $s_N \equiv \bar{s}$: resulta:

$$\bar{q}(s) = \frac{1}{2} \bar{s}^T \bar{H} \bar{s} + \bar{h}^T \bar{s}.$$

Es claro que las columnas de $\mathcal{W}$ forman una base de $\mathcal{N}(\nabla C^T)$, pues:

$$\mathcal{N}(\nabla C^T)\mathcal{W} = \begin{bmatrix} B & | & N \end{bmatrix} \begin{pmatrix} -B^{-1}N \\ I_{n-m} \end{pmatrix} = \Big(B(-B^{-1}N) + N I_{n-m}\Big) = -N + N = 0,$$

siendo 0, la matriz nula en $\mathbb{R}^{m\times(n-m)}$.

De esta forma, es posible definir una transformación lineal de $\mathbb{R}^n$ en $\mathcal{N}(\nabla C^T)$, que a cada $s \in \mathbb{R}^n$ lo transforme en $\mathcal{W}^T s \in \mathcal{N}(\nabla C^T)$, con lo cual el problema (Q_c) se reduce a:

$$(\bar{Q}) = \begin{cases} \min & \bar{q}(\bar{s}) = \frac{1}{2}\bar{s}^T \bar{H} \bar{s} + \bar{h}^T \bar{s} \\ s.a & \\ & \bar{s} \in \mathcal{N}(\nabla C^T), \end{cases}$$

siendo $(\bar{Q})$ el problema reducido al espacio tangente.

Bibliografía

[1] J. P. BOYLE and R. L. DYKSTRA. A method for finding projections onto the intersections of convex sets in hilbert spaces. *Lecture Notes in Statistics*, 37:28–47, 1986.

[2] R. H. CHAN and M. K. NG. Conjugate gradient methods for Toeplitz systems. *SIAM Journal on Optimization*, 38(3):427–482, 1996.

[3] W. CHENEY and A. GOLDSTEIN. Proximity maps for convex sets. *Proc. Amer. Math. Soc.*, 10:448–450, 1959.

[4] R.L. DYKSTRA. Restricted least-squares regression. *J. Amer. Stat. Assoc.*, 78:837–842, 1983.

[5] R. ESCALANTE and M. RAYDAN. Dykstra's algorithm for a constrained least-squares matrix problem. *Numerical Linear Algebra with applications*, 9(6):459–471, 1996.

[6] R. ESCALANTE and M. RAYDAN. On Dykstra's algorithm for constrained least-squares rectangular matrix problem. *Computers and Mathematics with Applications*, 35:73–79, 1998.

[7] G.H. GOLUB and C.F. VAN LOAN. *Matrix Computations*. Johns Hopkins University Press, Baltimore, Maryland, 1989.

[8] N. J. HIGHAM. The symmetric Procrustes problem. *BIT*, 28:133–143, 1988.

[9] E. R. LORCH. *Spectral Theory*. N. Y. Oxford University Press, New York, 1962.

[10] J. VON NEUMANN. Functional operators vol. ii, the geometry of orthogonal spaces. *Annals of Math. Studies Princeton University Press*, 22, 1950.

[11] M. SIMONNARD. *Programmation Linéaire*. Dunod Université, París, Francia, 1962.