IMPLEMENTACIÓN DE ANÁLISIS DE COMPONENTES PRINCIPALES APLICADO AL CONSUMO DE PROTEÍNAS EN LAS DIETAS DE 25 PAÍSES

TRABAJO FINAL

EQUIPO 11

Mariana Godina 113682

Sonia Mendizábal 105720

Primavera 2017

17 de abril de 2017

Índice

Introducción **3**

Componentes Principales **4**

Método Jacobi **6**
Paralelización del Método Jacobi 6

Referencias **7**

Introducción

El objetivo de este trabajo es la implementación del análisis de componentes principales y la descomposición de valores singulares (SVD) en el lenguaje de programación C usando metodologías computacionales de optimización.

En el primer capítulo del trabajo se presenta una introducción al análisis de componentes principales y la descomposición en valores singulares.

En los tres capítulos siguientes se presenta el método Jacobi como método computacional para obtener la descomposición y la paralelización del método.

Finalmente, se concluye el trabajo con la implementación del método para detectar patrones de consumo de proteínas en las dietas de 25 países.

Componentes Principales

El método de componentes principales tiene como objetivo explicar parte de la variación de un conjunto de variables basándose en dimensiones subyacentes.

Si se tiene un conjunto de n observaciones o individuos y p variables. Las dimensiones permiten simplificar la estructura de los datos de tal forma que se puede tener una reducción en el número de variables de p variables originales a un conjunto menor k de variables, donde $p < k$.

La forma en que se obtiene esto es mediante la construcción de combinaciones lineales no correlacionada que capturan la mayoría de la información de las variables. Es decir, una componente principal o dependencia lineal se define como:

$$y_1 = a_{11}x_1 + a_{21}x_2 + \ldots + a_{p1}x_p$$

donde a_{ij} son los pesos estimados al maximizar la varianza o suma de cuadrados de las correlaciones de la componentes con las variables originales.

En particular, si existen p variables, máximo hay p componentes. Pero, si hay dependencias lineales entre variables, forzosamente $k < p$.

Sea $R \in \mathbf{R^{n \times n}}$ la matriz de correlación simétrica de X. Entonces se obtiene la primera componente maximizando la varianza de la siguiente forma,

$$max \sum_{i=1}^{p} a_{1i}x_i \quad s.a. \quad \sum_{i=1}^{p} a_{1i}^2 = 1.$$

La segunda componente se obtiene maximizando la varianza con las siguientes restricciones,

$$max \sum_{i=1}^{p} a_{2i}x_i \quad s.a. \quad \sum_{i=1}^{p} a_{2i}^2 = 1 \quad y \quad \sum_{i=1}^{p} a_{1i}a_{2i} = 0.$$

La restricción $\sum_{i=1}^{p} a_{1i}a_{2i} = 0$ es necesaria dado que una particularidad del método es la ortogonalidad entre las componentes principales.

En el caso de la tercera componente se obtiene maximizando la varianza con las siguientes restricciones,

$$max \sum_{i=1}^{p} a_{3i}x_i \quad s.a. \quad \sum_{i=1}^{p} a_{3i}^2 = 1 \quad y \quad \sum_{i=1}^{p} a_{1i}a_{2i}a_{3i} = 0.$$

Esta maximización se repite para las p componentes principales, considerando la ortogonalidad de las componentes previas.

Además, $\sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} \sigma_i^2$ donde λ_i es la varianza de la i-ésima componente.

La proporción de la varianza original explicada por la componente se define como $\sum_{i=1}^{k} \frac{\lambda_i}{p}$ donde $k < p$.

Este sistema de ecuaciones se traduce en la descomposición de valores singulares $Ra = \lambda a$ donde λ son las raíces latentes o los eigen valores de la matriz R y a los vectores latentes o eigen vectores.

En particular, si R es no singular, entonces existen p raíces latentes λ_i y p vectores latentes asociados a_i. Esto cumple que $\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_p$ y cada uno tiene asociado el eigen vector a_i para $i = 1, \ldots, p$.

Sea $A_{(p \times p)}$ la matriz de vectores latentes o eigen vectores, $x_{(p \times 1)}$ las covariables originales y $y_{(p \times 1)}$ el vector de scores de las componentes, tal que:

$$y = A'X$$

Entonces,

$$RA = A\Lambda$$

donde Λ es la matriz diagonal de eigen valores.

La meta es descomponer la matriz de correlación y explicar la variación en términos de vectores de pesos a_i o vectores latentes de las componentes y las varianzas o raíces latentes de cada componente.

En el siguiente capítulo, se describe el método computacional de Jacobi para obtener la descomposición en valores singulares de una matriz.

Método Jacobi

El Método Jacobi es un método iterativo que realiza rotaciones hasta que la matriz sea casi diagonal. De tal forma, que los elementos de la diagonal son aproximaciones de los eigen valores de la matriz.

En específico, se busca reducir sistemáticamente la norma de los elementos no diagonales, definida a continuación:

$$off(A) = ||A||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2}$$

Esto se logra realizando rotaciones Jacobi con la matriz de rotación de Givens, que se presenta a continuación:

$$J(p,q,\theta) = \begin{bmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 & \\ \vdots & \ddots & \vdots & \dots & \vdots & \dots & \vdots & \\ 0 & \dots & c & \dots & s & \dots & \vdots & p \\ \vdots & \dots & \vdots & \ddots & \vdots & \dots & \vdots & \\ 0 & \dots & -c & \dots & -s & \dots & \vdots & q \\ \vdots & \dots & \vdots & \dots & \vdots & \ddots & \vdots & \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 & \\ & & p & & q & & & \end{bmatrix}$$

El procedimiento consiste en:

1. Escoger un par de renglones (p, q) tal que $1 \leq p \leq q \leq n$.
2. Se calcula el par (c, s) donde $c = cos(\theta)$ y $s = sin(\theta)$ y se obtiene la matriz B:

$$\begin{bmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{bmatrix} = \begin{bmatrix} c & s \\ -s & -c \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{bmatrix} \begin{bmatrix} c & s \\ -s & -c \end{bmatrix}$$

3. Finalmente, se reescribe $A \leftarrow B = J(p,q,\theta)^T A\, J(p,q,\theta)$.

Es por esto que se dice que A se acerca a la diagonal con cada paso del método Jacobi. Esto se puede ver en la norma de los elementos no diagonales de ambas matrices $off(B)^2 = off(A)^2 - 2a_{pq}^2$. Finalmente, el objetivo del método es maximizar la reducción de la norma $off(A)$, es decir, maximizar el término a_{pq}^2.

A la descomposición de la matriz A en la selección del par de índices (p, q) se le conoce como la *Descomposición Simétrica de Schur 2 por 2*.

Paralelización del Método Jacobi

El algoritmo del Método Jacobi, gracias al método de Descomposición Simétrica de Schur 2 por 2 y la selección de índices, es altamente paralelizable. Las rotaciones dentro de cada conjunto de pares no tienen conflicto entre sí al realizar las rotaciones.

Para la paralelización del método se genera un *ordenamiento* $(i_1, j_1), (i_2, j_2), \dots, (i_N, j_N)$ del conjunto o set $\{(i, j) | 1 \leq i \leq j \leq n\}$ tal que para s el set de rotación $rot.set(s) = \{(i_r, j_r) : r = (1 + \frac{n(s-1)}{2}) : (\frac{ns}{2})\}$ no tiene conflictos, donde $N = \frac{(n-1)n}{2}$ y $s = 1 : (n-1)$.

Referencias

- B. B. Zhou, R. P. Brent. On Parallel Implementation of the One-sided Jacobi Algorithm for Singular Value Descompositions
- R. P. Brent, F. Luk. The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays.
- Capítulo 8, Jacobi Methods, del libro: G. H. Golub, C. F. Van Loan, Matrix Computations. John Hopkins University Press, 2013.
- Sobre rotaciones Givens: capítulo 5 del libro Carl D. Meyer. Matrix analysis and applied linear Algebra. En este libro también encuentran teoría sobre eigenvalores-vectores, diagonalización y descomposición en valores singulares.
- G. H. Dunteman. Principal Component Analysis.
- Referencia para los cyclic distribution (round robin), capítulo 8 de P.Pacheco. Parallel Programming with MPI.