# Computación Científica I

## Laboratorio 2

Profesoras: Paola Arce, Raquel Pezoa

Luz Martínez
lmartine@alumnos.inf.usm.cl

Jorge Nacer
jnacer@alumnos.inf.usm.cl

5 de junio de 2012

# 1. Instrucciones generales

El laboratorio es individual y está conformado por dos entregables: un informe y los códigos realizados.

## 1.1. Informe

El informe debe contar con las secciones: Introducción, Análisis de Resultados, Conclusiones y Anexos (si es el caso). Además, se debe tener presente las siguientes consideraciones:

- La ortografía y redacción del informe serán consideradas en la nota final.
- En los anexos deben ir los casos de prueba que se utilizaron, especificando claramente input(s) y output(s) (en caso de que el ejercicio lo requiera).
- El informe debe ser elaborado en LaTeX.

## 1.2. Código

- Se puede trabajar en python, matlab y octave.
- Se evaluará el orden (indentación y claridad) y la documentación del código.
- No se permite el uso parcial o total de códigos encontrados en Internet o en libros.
- Debe respetarse el(los) input(s) solicitado(s) en el ejercicio.
- Los nombres de los archivos deben llevar el mismo nombre de la función, por ejemplo, si se pide una función llamada funcion1, su archivo debe llamarse funcion1.py.

# 2. Factorización QR

La factorización QR de una matriz $A$ permite descomponer $A$ en dos matrices: $Q$ que es una matriz ortogonal y $R$ que corresponde a una matriz triangular superior. Para los ejercicios considere las siguientes matrices:

$$A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix} \qquad B = \begin{bmatrix} 19 & 0 & 0 \\ 0 & 5 & 0 \\ 7 & 0 & 0 \\ 0 & 0 & 23 \\ 0 & 31 & 0 \\ 91 & 0 & 7 \\ 13 & 41 & 19 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 2 & 1 & -1 & 1 \\ -2 & 1 & 3 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$D = \begin{bmatrix} -1 & 1 & -1 & 2 \\ 1 & -2 & 3 & -3 \\ -1 & 1 & 0 & 1 \\ 0 & -1 & 2 & -2 \end{bmatrix} \qquad E = \begin{bmatrix} 2 & 3 & 0 & 1 & 8 & 1 \\ -5 & 1 & 4 & -2 & -2 & 5 \\ 1 & 0 & 0 & -3 & 9 & 2 \end{bmatrix}$$

## 2.1. Ejercicios

1. Utilizando el método de ortogonalización de Gram-Schmidt implemente la factorización QR completa y QR reducida creando las funciones `[Q,R]=mifullQR(A)` y `[Q,R]=miredQR(A)` donde `A` es la matriz de entrada y `Q,R` la matrices de salida.

2. Calcule el *rank* de cada una de las matrices utilizando las funciones que provee octave, matlab (`rank()`) o python (`numpy.linalg.matrix_rank()`). ¿Cuál o cuáles matrices no tienen *full rank*? En las funciones implementadas, usted debe tomar en cuenta el *rank* de la matriz. Explique como usted manejó esta situación en sus funciones.

3. La matriz $E$ tiene más columnas que filas ($E \in Mat(3 \times 6, \mathbb{R}), m < n$). ¿Cuáles son las dimensiones de las matrices entregadas por sus funciones `mifullQR(A)` y `miredQR(A)`. ¿Qué modificación a su algoritmo debió realizar para considerar la factorización QR de la matriz $E$?

# 3. LSI

El método "Latent semantic indexing"[1] [2] permite representar un conjunto de documentos de manera reducida capturando los términos más relevantes de cada documento.

Se tienen $D$ número de documentos y $N$ términos que pueden o no aparecer en los documentos. Para la representación de la ocurrencia de los términos en los documentos se utiliza una matriz de incidencia $A$ donde:

$$A(i,j) = \begin{cases} 1 & \text{si el término } i \text{ se encuentra en el documento } d_j \\ 0 & \text{si el término } i \text{ no se encuentra en el documento } d_j \end{cases}, \forall_{1 \leq i \leq N, 1 \leq j \leq D}$$

Al calcular la SVD de $A$, se construye un "espacio semántico" en el que se asocian entre sí términos y documentos, donde las matrices de U y V representan los términos y documentos respectivamente en este nuevo espacio. Y la matriz S, representa la matriz diagonal de la disminución de valores singulares.

## 3.1. Reducción de la dimensión del espacio

Dado que la matriz $A$ es posible expresarla como:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^*$$

donde $r$ corresponde al $rank(A)$.

Dado que $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_p$ una matriz aproximada $A_p$ se puede obtener calculando los primeros $p$ términos de la sumatoria anterior, es decir, puede ser expresada como:

$$A_p = U_p S_p V_p^* = \sum_{i=1}^{p} \sigma_i u_i v_i^* \,,$$

Donde $U_p$ y $V_p$ son matrices formadas por las primeras $p$ columnas de las matrices $U$ y $V$. Como consecuencia de esta operación, se pasa del espacio vectorial generado por las columnas de la matriz $A$ al espacio generado por las columnas $A_p$.

A esta operación se le llama reducción de la dimensión y al espacio de dimensión $p$ se le llama espacio reducido.

## 3.2. Búsqueda semántica

Utilizando el modelo creado, es posible realizar una consulta que permita determinar qué documentos contienen la información que se busca. Para esto se debe definir un vector $q$ que represente la consulta que desea realizarse:

[1] http://arantxa.ii.uam.es/ castells/docencia/ir/apuntes/3-lsa.pdf

[2] http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

$$q_i = \begin{cases} 1 & \text{si se buscará el término } i \text{ en los documentos} \\ 0 & \text{si no se buscará el término } i \text{ en los documentos} \end{cases} \forall_{1 \leq i \leq N}$$

Para trasladar el vector $q$ al espacio reducido, se realiza como:

$$q_p = (q^T U_2 S_2)$$

Donde las componentes del vector $q_p$ son las coordenadas del vector $q$ en el espacio reducido. A continuación se calcula la similaridad que existe entre el vector de búsqueda y los vectores de los documentos $d_p$, utilizando el coseno del ángulo $\theta$ que forman ambos vectores.

$$cos\theta = \frac{q_p \cdot d_i}{|q_p||d_i|},$$

Si el ángulo $\theta$ que forman los dos vectores es pequeño, el coseno será proximo a 1, y se interpretará que la búsqueda y el documento son semánticamente muy similares. Por otra parte, si el coseno es cercano a 0 se entenderá que el documento es muy distinto a la consulta realizada.

## 3.3. Ejercicios

1. Utilizando los siguientes documentos construya su respectiva matriz de incidencia. Considere sólo los términos subrayados para su creación.

   - $d_1$ : Análisis automático de datos en astronomía.
   - $d_2$ : Minería de datos distribuidos.
   - $d_3$ : Nuevas plataformas para computación en nube.
   - $d_4$ : Bases de datos no relacionales distribuidos en computación nube.
   - $d_5$ : Astronomía en informática.
   - $d_6$ : Implementación de un método molecular para la detección del virus de la influenza humana.
   - $d_7$ : Valoración molecular de esquemas de virus humano.
   - $d_8$ : Actividad del virus de la coriomeningitis.

2. Cree un vector $q_1$ con las palabras: "datos", "distribuidos" e "informática", y un vector $q_1$ con las palabras: "virus", "nube". Luego proceda a encontrar la relación con los documentos. Concluya.

3. Realice una reducción del espacio a $p = 2$ de las matrices entregadas por la SVD y encuentre la relación de los documentos con los vectores $q_1$ y $q_2$. Luego grafique una representación de los documentos en el plano (es decir, las columnas de la matriz V) y

los vectores de las consultas trasladados al espacio reducido. Compare sus resultados con la pregunta anterior. Concluya.

4. Uno de los problemas más frecuentes del método LSI es cuando aparecen sinónimos y polisemia.¿Por qué?, ¿qué recomendaría en dichas situaciones?.

Se puede utilizar el comando facilitado por python, matlab u octave para calcular la SVD.

# 4. Sobre la entrega

- El plazo máximo de entrega es el próximo 19 de Junio, a las 23:55 hrs, vía moodle.
- El normbre del archivo debe ser lab2-InicialnombreApellido1.tar.gz,(ejemplo lab2-lmartinez.tar.gz) que debe contener un directorio llamado Informe que contenga los archivos .pdf y .tex correspondientes y otro directorio llamado Códigos con los archivos correspondientes.
- Se sancionará con 15 puntos menos en la nota final del laboratorio por cada día de atraso.
- Las copias serán sancionadas con nota cero (0) para todos los involucrados.

# 5. Evaluación

| Item | Puntaje |
|---|---|
| QR 1) | 15 |
| QR 2) | 15 |
| QR 3) | 16 |
| LSI 1) | 5 |
| LSI 2) | 15 |
| LSI 3) | 15 |
| LSI 4) | 11 |
| Redacción y Ortografía | 8 |
| Descuento: Código desordenado o no comentado | 10 |
| Descuento: Nombre incorrecto del archivo | 5 |