Sistemas de Ecuaciones Lineales II

- **Métodos Iterativos:** Matrices dispersas. Esquema general. Métodos de Jacobi y de Gauss-Seidel.

Matrices dispersas

En muchas aplicaciones los sistemas de ecuaciones lineales que deben resolverse involucran matrices de gran tamaño, pero la mayor parte de sus entradas son nulas. Estas matrices se denominan **dispersas** o **ralas** (en inglés y en OCTAVE, **sparse**) y existen técnicas para almacenarlas que sólo requieren una cantidad de posiciones de memoria aproximadamente igual al número de entradas no nulas de la matriz. Por ejemplo:

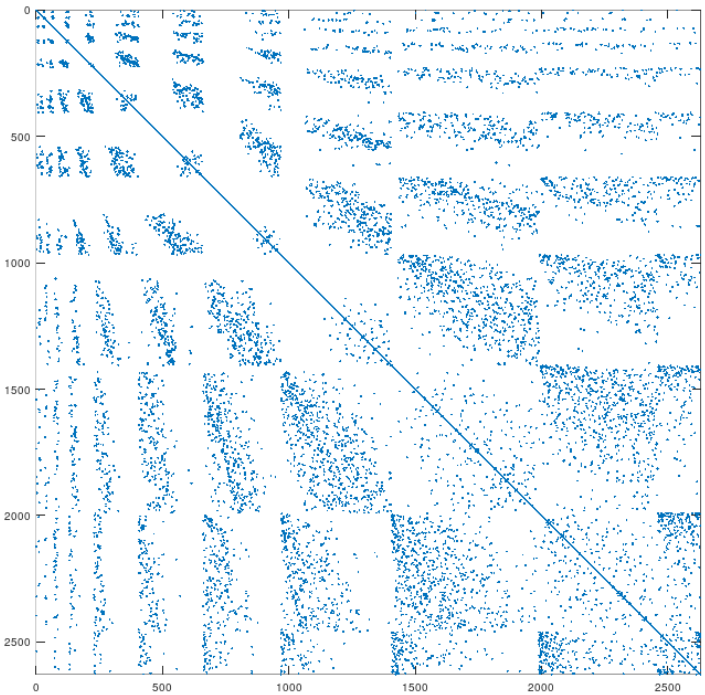
```
>> n = 7;
>> A = diag(ones(n-1,1),-1)+2*diag(ones(n,1))+diag(ones(n-1,1),1);
>> A
A =

   2   1   0   0   0   0
   1   2   1   0   0   0
   0   1   2   1   0   0
   0   0   1   2   1   0
   0   0   0   1   2   1
   0   0   0   0   1   2
```

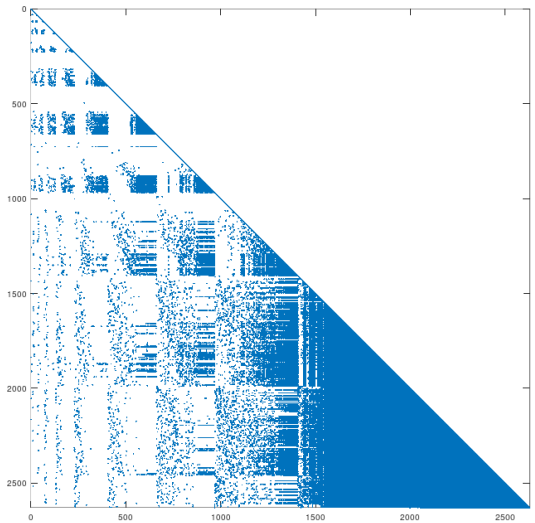
Para visualizar la gráficamente la estructura de la matriz podemos usar el comando `spy(A)`.

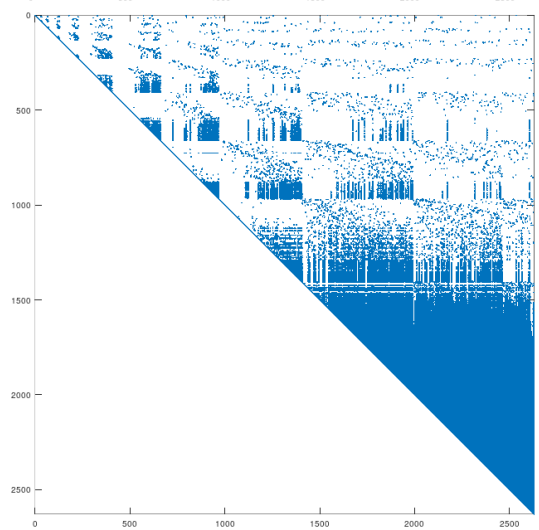
Ejemplo llenado de matrices dispersas

```
>> load DATA_025.mat
>> [L,U,P]=lu(A);
>> whos
   Attr Name        Size                     Bytes  Class
   ==== ====        ====                     =====  =====
        A        2629x2629                  314928  double
        L        2629x2629                 9048128  double
        P        2629x2629                   21032  double
        U        2629x2629                 8804992  double
```

Ejemplo llenado de matrices dispersas (cont.)







- Cuando la matriz $\boldsymbol{A}$ del sistema a resolver es dispersa, pero no banda, los métodos (directos) estudiados hasta ahora (eliminación de Gauss, factorización LU o Cholesky) presentan el defecto denominado **llenado** (**fill-in**).
- El llenado consiste en que, a medida que el proceso de eliminación avanza, se van creando elementos no nulos en posiciones de $\boldsymbol{L}$ y $\boldsymbol{U}$ en donde la matriz $\boldsymbol{A}$ tiene ceros.
- Como consecuencia del llenado se tiene, por una parte, el aumento del número de flop y con ello el aumento del error de redondeo. Por otra parte se tiene el aumento en las necesidades de memoria para almacenar las matrices $\boldsymbol{L}$ y $\boldsymbol{U}$, lo que puede llegar a hacer imposible aplicar estos métodos cuando $\boldsymbol{A}$ es de gran tamaño.
- Los métodos que estudieremos en el siguiente capitulo, llamados **iterativos**, evitan el llenado y sus consecuencias, al trabajar resolviendo reiteradamente sistemas con matriz diagonal o triangular-dispersa.

Esquema general

- Considere el sistema de ecuaciones

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

con $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ no singular y $\boldsymbol{b} \in \mathbb{R}^n$.
Un **método iterativo** para resolver el sistema construye, a partir de un vector inicial $\boldsymbol{x}^{(0)}$, una sucesión de vectores $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(k)}, \ldots$ la que, bajo condiciones apropiadas, resultará convergente a $\boldsymbol{x}$.

- Si suponemos $\boldsymbol{A} = \boldsymbol{N} - \boldsymbol{P}$, donde $\boldsymbol{N}$ debe ser invertible, entonces

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} &\iff (\boldsymbol{N} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{b} \\ &\iff \boldsymbol{N}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} + \boldsymbol{b} \\ &\iff \boldsymbol{x} = \boldsymbol{N}^{-1}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{N}^{-1}\boldsymbol{b}. \end{aligned}$$

- Se usa la igualdad $\boldsymbol{N}\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} + \boldsymbol{b}$ para definir un esquema general para construir la sucesión $\{\boldsymbol{x}^{(k)}\}$. Este esquema se da en el siguiente algoritmo.

Esquema general (cont.)

- **Algoritmo del esquema general:**

  Dado el vector inicial $\boldsymbol{x}^{(0)}$,
  para $k = 1, 2, \ldots$ resolver:
  $$\boldsymbol{N}\boldsymbol{x}^{(k)} = \boldsymbol{P}\boldsymbol{x}^{(k-1)} + \boldsymbol{b},$$
  hasta que se satisfaga un criterio de detención.

- Definiendo $\boldsymbol{M} := \boldsymbol{N}^{-1}\boldsymbol{P}$ (**matriz de iteración**) y $\boldsymbol{e}^{(k)} := \boldsymbol{x} - \boldsymbol{x}^{(k)}$ (**error de $\boldsymbol{x}^{(k)}$**), para cada $k = 1, 2, \ldots$ se tiene

$$\begin{aligned}
\boldsymbol{e}^{(k)} &= \boldsymbol{x} - \boldsymbol{x}^{(k)} \\
&= (\boldsymbol{N}^{-1}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{N}^{-1}\boldsymbol{b}) - (\boldsymbol{N}^{-1}\boldsymbol{P}\boldsymbol{x}^{(k-1)} + \boldsymbol{N}^{-1}\boldsymbol{b}) \\
&= \boldsymbol{N}^{-1}\boldsymbol{P}(\boldsymbol{x} - \boldsymbol{x}^{k-1}) \\
&= \boldsymbol{M}\boldsymbol{e}^{(k-1)}
\end{aligned}$$

y, recursivamente,

$$\boldsymbol{e}^{(k)} = \boldsymbol{M}^k \boldsymbol{e}^{(0)}, \qquad k = 1, 2, \ldots$$

Convergencia de métodos iterativos.

- **Teorema.** (Convergencia) La sucesión $\left\{\boldsymbol{x}^{(k)}\right\}$ converge a la solución $\boldsymbol{x}$ de $\boldsymbol{Ax} = \boldsymbol{b}$, si y sólo si, $\rho(\boldsymbol{M}) < 1$.

- **Observación.** Si la sucesión $\left\{\boldsymbol{x}^{(k)}\right\}$ converge, necesariamente lo hace a la solución $\boldsymbol{x}$ de $\boldsymbol{Ax} = \boldsymbol{b}$.

- **Lema.** (Cota para el radio espectral) Sea $\boldsymbol{A}$ una matriz cuadrada. Para cualquier norma matricial se tiene que

$$\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\| .$$

- **Corolario.** (Condición suficiente de convergencia) Una condición suficiente para que la sucesión $\left\{\boldsymbol{x}^{(k)}\right\}$ sea convergente a la solución $\boldsymbol{x}$ de $\boldsymbol{Ax} = \boldsymbol{b}$ es que

$$\|\boldsymbol{M}\| < 1,$$

donde $\boldsymbol{M}$ es la matriz de iteración del método que genera a $\left\{\boldsymbol{x}^{(k)}\right\}$.

Criterio de detención

- **Detención del proceso.** Cuando el proceso iterativo es convergente, éste se debe detener para un $\boldsymbol{x}^{(k)}$ tal que

$$\frac{\|\boldsymbol{M}\|}{1-\|\boldsymbol{M}\|}\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\right\| \leq \texttt{tol},$$

donde `tol` indica un nivel de tolerancia prefijado para el error.

- **Lema.** Para $\|\boldsymbol{M}\| < 1$, se tiene que $\rho(\boldsymbol{M}) < 1$ y

$$\left\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\right\| \leq \frac{\|\boldsymbol{M}\|}{1-\|\boldsymbol{M}\|}\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\right\|.$$

- **Consecuencia**: Si detenemos el algortimo cuando

$$\frac{\|\boldsymbol{M}\|}{1-\|\boldsymbol{M}\|}\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\right\| \leq \texttt{tol},$$

nos aseguramos que el error $\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|$ es menor que `tol`.

Criterio de detención (cont.)

- El criterio de detención implica calcular $\|\boldsymbol{M}\|$, lo que en general es difícil. En lugar de ello, calculamos

$$m_k := \frac{\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\|}{\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k-2)}\|}.$$

  que constituye una estimación de $\|\boldsymbol{M}\|$. En efecto,

- **Lema.** Para $k = 2, 3, \ldots$ se tiene que $m_k \leq \|\boldsymbol{M}\|$.

Demostración.

$$m_k = \frac{\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\|}{\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k-2)}\|} = \frac{\left\|\left[\boldsymbol{x}^{(k)} - \boldsymbol{x}\right] - \left[\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\right]\right\|}{\|\left[\boldsymbol{x}^{(k-1)} - \boldsymbol{x}\right] - \left[\boldsymbol{x}^{(k-2)} - \boldsymbol{x}\right]\|}$$

$$= \frac{\left\|\boldsymbol{e}^{(k)} - \boldsymbol{e}^{(k-1)}\right\|}{\|\boldsymbol{e}^{(k-1)} - \boldsymbol{e}^{(k-2)}\|} = \frac{\left\|\boldsymbol{M}\left[\boldsymbol{e}^{(k-1)} - \boldsymbol{e}^{(k-2)}\right]\right\|}{\|\boldsymbol{e}^{(k-1)} - \boldsymbol{e}^{(k-2)}\|} \leq \max_{\boldsymbol{y}\in\mathbb{R}^n:\, \boldsymbol{y}\neq 0} \frac{\|\boldsymbol{M}\boldsymbol{y}\|}{\|\boldsymbol{y}\|} = \|\boldsymbol{M}\|.$$

- Así, proceso iterativo detenemos cuando:

$$\frac{m_k}{1 - m_k}\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\| \leq \texttt{tol}.$$

Métodos de Jacobi y Gauss-Seidel

- Se considera resolver un sistema $\boldsymbol{Ax} = \boldsymbol{b}$ con $a_{ii} \neq 0$, para $i = 1, \ldots, n$. Sea $\boldsymbol{x}^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^{\mathrm{t}}$ arbitrario y escribamos la matriz $\boldsymbol{A}$ en la forma

$$\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{E} - \boldsymbol{F},$$

donde $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{A})$, $-\boldsymbol{E}$ es la matriz triangular inferior de $\boldsymbol{A}$ y $-\boldsymbol{F}$ es la matriz triangular superior de $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{pmatrix} \ddots & & -\boldsymbol{F} \\ & \boldsymbol{D} & \\ -\boldsymbol{E} & & \ddots \end{pmatrix}$$

- Notemos que tanto $\boldsymbol{D}$ como $\boldsymbol{D} - \boldsymbol{E}$ son matrices inversibles, ya que $a_{ii} \neq 0$ para $i = 1, \ldots, n$.

- El **método de Jacobi** corresponde al esquema iterativo general con

$$\boldsymbol{N} := \boldsymbol{D} \qquad \text{y} \qquad \boldsymbol{P} := \boldsymbol{E} + \boldsymbol{F}.$$

- El **método de Gauss-Seidel** corresponde al esquema iterativo general con

$$\boldsymbol{N} := \boldsymbol{D} - \boldsymbol{E} \qquad \text{y} \qquad \boldsymbol{P} := \boldsymbol{F}.$$

Convergencia del Método de Jacobi

- La matriz de iteración del método de Jacobi verifica:

$$\boldsymbol{M} = \boldsymbol{D}^{-1}(\boldsymbol{E}+\boldsymbol{F}) = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & -\frac{a_{n-1\,n}}{a_{n-1\,n-1}} \\ -\frac{a_{n1}}{a_{nn}} & \cdots & \cdots & -\frac{a_{n\,n-1}}{a_{nn}} & 0 \end{pmatrix}$$

- En este caso $\|\boldsymbol{M}\|_\infty = \max\limits_{1\le i\le n} \left\{ \dfrac{1}{|a_{ii}|} \sum\limits_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\}$.

- Cuando $\boldsymbol{A}$ es de **diagonal dominante estricta**, es decir,

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n,$$

entonces $\|\boldsymbol{M}\|_\infty < 1$ y el método de Jacobi resulta convergente.

Convergencia del Método de Gauss-Seidel

- El método de Gauss-Seidel es convergente cuando $\boldsymbol{A}$ es de **diagonal dominante estricta**, es decir, cuando

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n.$$

- El método de Gauss-Seidel también es convergente cuando $\boldsymbol{A}$ es **definida positiva**.

Resumen

- **Teorema.** Si $\boldsymbol{A}$ es de diagonal dominante estricta, entonces los métodos de Jacobi y de Gauss-Seidel convergen.

- **Observación.** Para una matriz arbitraria $\boldsymbol{A}$, la convergencia de uno de estos métodos no implica la convergencia del otro.

- **Teorema.** Si $\boldsymbol{A}$ es simétrica y definida positiva, el método de Gauss-Seidel es convergente.

- **Observación.** Aunque $\boldsymbol{A}$ sea simétrica y definida positiva, el método de Jacobi puede ser divergente.

Convergencia de los métodos (cont.)

- **Ejemplo.** Para $s \in \mathbb{R}$, considere la matriz simétrica

$$\boldsymbol{A} = \begin{pmatrix} 1 & s & s \\ s & 1 & s \\ s & s & 1 \end{pmatrix},$$

cuyos valores propios son: $1 - s$ (con multiplicidad 2) y $1 + 2s$.
La matriz $\boldsymbol{A}$ es definida positiva cuando $s \in (-1/2, 1)$ y es de diagonal dominante estricta para $s \in (-0.5, 0.5)$.

- Se resolvió el sistema $\boldsymbol{Ax} = \boldsymbol{b}$ para un par de valores de s, usando los métodos de Jacobi y de Gauss-Seidel, con $\boldsymbol{b} = (1, 1, 1)^{\mathrm{t}}$ y $\boldsymbol{x}^{(0)} = (0.5, 0.5, 0.5)^{\mathrm{t}}$.
- Para $s = 0.3$, Jacobi itera 37 veces y Gauss-Seidel 12 veces (en ambos casos se implementó el criterio de detención visto en clase, con una tolerancia de 10^{-8}).
Ambos métodos entregan como solución $\boldsymbol{x} = (0.6250, 0.6250, 0.6250)^{\mathrm{t}}$, que es la solución exacta.

Convergencia de los métodos (cont.)

- Para $s = 0.8$, en las mismas condiciones anteriores, Gauss-Seidel converge en la iteración 53 a

$$(0.384\,615\,391\,735, 0.384\,615\,381\,035, 0.384\,615\,381\,784)^{\mathrm{t}}$$

que difiere de la solución exacta

$$\boldsymbol{x} = (0.384\,615\,384\,615, 0.384\,615\,384\,615, 0.384\,615\,384\,615)^{\mathrm{t}}$$

en menos de $\texttt{tol} = 10^{-8}$ en cada componente.

- En cambio, al cabo de 100 iteraciones Jacobi entrega

$$10^{19} \times (-1.862\,199\,431\,313, -1.862\,199\,431\,313, -1.862\,199\,431\,313)^{\mathrm{t}},$$

vector que no tiene ninguna relación con la solución del sistema.
Se nota claramente que, en este caso, el método de Jacobi diverge.