Cholesky factorization and power method for positive (semi-)definite matrices

Sungsu Lim
Applied Algorithm Lab
Nov. 9, 2011

Outline

Positive definite matrix

- A matrix $A \in \mathbb{R}^{n \times n}$ is positive definite if $\mathbf{x}^T A\mathbf{x} > 0$ for all $x \neq 0$.
- **Theorem** If A is a symmetric matrix, then the following statements are equivalent.
  - **(a)** A is positive definite.
  - **(b)** There is an invertible matrix B such that $A = BB^T$.

Proof

- (a)⇒(b): Since A is real and symmetric, it is orthogonally diagonalizable. $A = PDP^T = PD_1D_1P^T = (PD_1P^T)(PD_1P^T)$, where D_1 is the diagonal matrix whose entries are the square roots of the eigenvalues of A. We can check that $B = PD_1P^T$ is symmetric and positive definite. Thus, $A = BB^T$ since $B = B^T$.
- (b)⇒(a): $\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T BB^T\mathbf{x} = (B^T\mathbf{x})^T(B^T\mathbf{x}) = \| B\mathbf{x} \|^2 > 0$ if $\mathbf{x} \neq 0$.

Cholesky factorization

- Every positive definite matrix A can be factored as $A = LL^T$ where L is lower triangular with positive diagonal elements.

- Cost: $(1/3)n^3$ flops
- L is unique, and it is called the Cholesky factor of A.
- The requirement that L has positive diagonal entries can be dropped to extend the factorization to the PSD case.
  In this case, Cholesky factorizations are not unique in general.
- Cholesky factorizations are important in many kinds of numerical algorithms, and programs such as MATLAB, *Maple*, and *Mathematica* have built-in commands for computing them.

Proof

- Induction on n. For $n = 1$, trivial. If $n \geq 2$, A is

$$\begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} l_{11} & L_{21}^T \\ 0 & L_{22}^T \end{bmatrix}$$
$$= \begin{bmatrix} l_{11}^2 & l_{11} L_{21}^T \\ l_{11} L_{21} & L_{21} L_{21}^T + L_{22} L_{22}^T \end{bmatrix},$$

  where $l_{11} = \sqrt{a_{11}}$, $L_{21} = (1/l_{11})A_{21}$, and L_{22} is a solution of $A_{22} - L_{21}L_{21}^T = L_{22}L_{22}^T$.
- Schur complement $A_{22} - L_{21}L_{21}^T = A_{22} - (1/a_{11})A_{21}A_{21}^T$ is positive definite, because for any $\mathbf{v} \neq 0$ and take $\mathbf{w} = -(1/a_{11})A_{21}^T\mathbf{v}$

$$\mathbf{v}^T(A_{22} - (1/a_{11})A_{21}A_{21}^T)\mathbf{v} = \begin{bmatrix} \mathbf{w}^T & \mathbf{v}^T \end{bmatrix} \begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{v} \end{bmatrix} > 0$$

Applications

- Solving equations $A\mathbf{x} = \mathbf{b}$ with positive definite $A = LL^T$
  $\Rightarrow$ factor A as $A = LL^T$
  $\Rightarrow$ forward substitution $L\mathbf{z} = \mathbf{b}$, back substitution $L^T\mathbf{x} = \mathbf{z}$
- Inverse of a positive definite matrix $A = LL^T$
  $\Rightarrow$ L is invertible (its diagonal elements are nonzero)
  $\Rightarrow$ A is invertible and $A^{-1} = L^{-T}L^{-1}$
- If A is very sparse, then L is often (but not always) sparse. If L is sparse, the cost of the factorization is much less than $(1/3)n^3$.
- Computing the Cholesky decomposition is more efficient and numerically more stable than computing LU decompositions.

Algorithm *Power*

- Input: PSD symmetric matrix $M \in \mathbb{R}^{n \times n}$, positive integer t
- Pick uniformly at random $\mathbf{x}_0 \sim \{-1, 1\}^n$
- for $i = 1, \ldots, t$
  - $\mathbf{x}_i := M\mathbf{x}_{i-1}$
- return $\mathbf{x}_t$

Theorem 1

For every PSD matrix M, positive integer t and parameter $\epsilon > 0$, with probability $\geq 3/16$ over the choice of $\mathbf{x}_0$, the algorithm *Power* outputs a vector $\mathbf{x}_t$ such that

$$\frac{\mathbf{x}_t^T M \mathbf{x}_t}{\mathbf{x}_t^T \mathbf{x}_t} \geq \lambda_1 (1 - \epsilon) \frac{1}{1 + 4n(1 - \epsilon)^{2t}}$$

where λ_1 is the largest eigenvalue of M.

Lemma 2

Let $\mathbf{v} \in \mathbb{R}^n$ be a vector such that $\| \mathbf{v} \| = 1$. Sample uniformly at random $\mathbf{x} \sim \{-1, 1\}^n$. Then

$$P\left[|<\mathbf{x}, \mathbf{v}>| \geq \frac{1}{2} \right] \geq \frac{3}{16}.$$

Lemma 3

Let $\mathbf{x} \in \mathbb{R}^n$ be a vector such that $|<\mathbf{x}, \mathbf{v}_1>| \geq \frac{1}{2}$. Then, for every positive integer t and parameter $\epsilon > 0$, if we define $\mathbf{y} := M^t \mathbf{x}$, we have

$$\frac{\mathbf{y}^T M \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \geq \lambda_1 (1 - \epsilon) \frac{1}{1 + 4 \| x \|^2 (1 - \epsilon)^{2t}}.$$

Proof of Lemma 2

- Let $\mathbf{v} = (v_1, \ldots, v_n)$. Then $S =< \mathbf{x}, \mathbf{v} >$ be a random variable with $E[S] = 0$, $E[S^2] = \sum v_i^2 = 1$, $E[S^4] = 3(\sum v_i^2) - 2\sum v_i^4 \leq 3$.
- Paley-Zygmund inequality: If Z is a non-negative random variable with finite variance, then, for every $0 \leq \delta \leq 1$, we have

$$P[Z \geq \delta E[Z]] \geq (1-\delta)^2 \frac{(E[Z])^2}{E[Z]^2}.$$

  (Proof: by Cauchy-Schwarz inequality)
- Take $Z = S^2$ and $\delta = 1/4$, we have

$$P\left[S^2 \geq \frac{1}{4}\right] \geq \left(\frac{3}{4}\right)^2 \cdot \frac{1}{3} = \frac{3}{16}.$$

Proof of Lemma 3

- Let us write $\mathbf{x}$ as a linear combinatioin of the eigenvectors

$$\mathbf{x} = a_1\mathbf{v}_1 + \ldots + a_n\mathbf{v}_n$$

where $a_i =< \mathbf{x}, \mathbf{v}_i >$. By assumption $|a_1| \geq 1/2$, and by orthonormality of the eigenvectors, $\| \mathbf{x} \|^2 = \sum a_i^2$. We have

$$\mathbf{y} = a_1\lambda_1^t\mathbf{v}_1 + \ldots + a_n\lambda_n^t\mathbf{v}_n$$

and so

$$\mathbf{y}^T M\mathbf{y} = \sum a_i^2\lambda_i^{2t+1} \quad \textit{and} \quad \mathbf{y}^T\mathbf{y} = \sum a_i^2\lambda_i^{2t}.$$

Proof of Lemma 3 (cont'd)

- Let k be the number of eigenvalues larger than $\lambda_1 \cdot (1-\epsilon)$. Then,

$$\mathbf{y}^T M \mathbf{y} \geq \sum_{i=1}^{k} a_i^2 \lambda_i^{2t+1} \geq \lambda_1 (1-\epsilon) \sum_{i=1}^{k} a_i^2 \lambda_i^{2t}.$$

We also see that

$$\begin{aligned} \sum_{i=k+1}^{n} a_i^2 \lambda_i^{2t} &\leq \lambda_1^{2t} (1-\epsilon)^{2t} \sum_{i=k+1}^{n} a_i^2 \\ &\leq \lambda_1^{2t} (1-\epsilon)^{2t} \parallel \mathbf{x} \parallel^2 \\ &\leq 4 a_1^2 \lambda_1^{2t} (1-\epsilon)^{2t} \parallel \mathbf{x} \parallel^2 \\ &\leq 4 \parallel \mathbf{x} \parallel^2 (1-\epsilon)^{2t} \sum_{i=1}^{k} a_i^2 \lambda_i^{2t}. \end{aligned}$$

Proof of Lemma 3 (cont'd)

So we have

$$\mathbf{y}^T\mathbf{y} \leq (1 + 4 \parallel \mathbf{x} \parallel (1-\epsilon)^{2t}) \sum_{i=1}^{k} a_i^2$$

giving

$$\frac{\mathbf{y}^T M \mathbf{y}}{\mathbf{y}^T\mathbf{y}} \geq \lambda_1 (1-\epsilon) \frac{1}{1 + 4 \parallel x \parallel^2 (1-\epsilon)^{2t}}.$$

Application

Graph partitioning: Let M be a matrix with eigenvalues $1 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. If it is a symmetric matrix and all its eigenvalues are nonnegative, then it is positive semi-definite. In some cases, we want to compute the second largest eigenvalue. That is, we want to find a vector $\mathbf{x} \perp \mathbf{1}$ such that $\mathbf{x}^T M\mathbf{x} \leq (\lambda_2 - \epsilon)\mathbf{x}^T\mathbf{x}$. In face, we can modify Theorem 1.

Algorithm *Power2*

- Input: PSD symmetric matrix M, positive integer t, vector $\mathbf{v}_1$
- Pick uniformly at random $\mathbf{x} \sim \{-1, 1\}^n$
- $\mathbf{x}_0 := \mathbf{x} - < \mathbf{v}_1, \mathbf{x} > \cdot \mathbf{v}_1$
- for $i = 1, \ldots, t$
  - $\mathbf{x}_i := M\mathbf{x}_{i-1}$
- return $\mathbf{x}_t$

Theorem 4

For every PSD matrix M, positive integer t and parameter $\epsilon > 0$, if $\mathbf{v}_1$ is a length-1 eigenvalue of M, then with probability $\geq 3/16$ over the choice of $\mathbf{x}_0$, the algorithm *Power2* outputs a vector $\mathbf{x}_t \perp \mathbf{v}_1$ such that

$$\frac{\mathbf{x}_t^T M \mathbf{x}_t}{\mathbf{x}_t^T \mathbf{x}_t} \geq \lambda_2 (1-\epsilon) \frac{1}{1+4n(1-\epsilon)^{2t}}$$

where λ_2 is the second largest eigenvalue of M, counting multiplicities.

References

Some useful lecture notes

- *Cholesky factorization:*
  http://www.ee.ucla.edu/ vandenbe/103/lectures/chol.pdf
- *Power method for PSD matrices:*
  http://theory.stanford.edu/ trevisan/cs359g/lecture07.pdf