Research Notes

Ruoyang Xu

Last Edited: January 25, 2021

Contents

1 Notations **1**

2 Todo List **1**

3 Linear Algebra **1**
3.1 Definitions . . . 1
3.1.1 Hermitian Matrix . . . 1
3.1.2 Definiteness . . . 1
3.1.3 Inner Product . . . 1
3.2 Methods . . . 2
3.2.1 Whitening Transformation . . . 2
3.2.2 Cholesky decomposition . . . 2
3.2.3 QR Decomposition . . . 3
3.2.4 Eigendecomposition of a Matrix . . . 4
3.2.5 Gram–Schmidt Process . . . 4
3.2.6 Householder Reflection . . . 5

4 Optimization **5**
4.1 Methods . . . 5
4.1.1 Steepest Descent . . . 5
4.1.2 Gauss-Newton . . . 5
4.1.3 Levenberg-Maquardt . . . 5
4.1.4 Dogleg . . . 5

1 Notations

Red texts: follow up notes.
Blue texts: commentary notes.

2 Todo List

- Optimization
- Mahalanobis Distance

3 Linear Algebra

3.1 Definitions

3.1.1 Hermitian Matrix

Introduction Otherwise known as a self-adjoint matrix, a Hermitian matrix is a *complex square* matrix that is equal to its own *conjugate transpose*. In otherwords

$$A \text{ Hermitian } \Leftrightarrow A = \overline{A^T}$$

Concisely written as $A = A^H$

Properties Entries on the main diagonal of any hermitian matrix are real
forgot why I included this definition, need investigation
Perhaps the definition of a more general "Symmetric matrix"

3.1.2 Definiteness

Introduction Symmetri $n \times n$ real matrix is said to be positive-definite if scalar $z^T M z$ is strictly positive for every non-zero column vector $z \in \mathcal{R}^n$. Semi-definite is defined similarly, except above scalar must be non-negative.

Properties Squre root and Cholesky decomposition might be two relevant properties here.

A matrix is positive semidefinite iff. there is a positive semidefinite matrix B satisfying $M = BB$. B is unique, and is called the non-negative square root of M, denoted with $B = M^{\frac{1}{2}}$.

B is hermitian, so $B* = B$, complex conjugate of itself.

Some use square root and $\sqrt{M}$ for any such dcomposition, or specifically for the Cholesky decomposition, or any decomposition, other only use for the non-negative square root.

Related to Cholesky decomposition as $M = LL*$, where L is lower triangular with non-negative diagonal. See more for section 3.2.2 on Cholesky decomposition.

3.1.3 Inner Product

An inner product operation contains the following properties:

1. Positivity: $\langle v, v \rangle \geq 0$
2. Definiteness: $\langle v, v \rangle iff v = 0$
3. Linearity in first slot: $\langle \lambda v + u, w \rangle = \lambda \langle v, w \rangle + \langle u, w \rangle$
4. Conjugate symmetry: $\langle u, v \rangle = \overline{\langle v, u \rangle}$

Inner product space is a vector space V along with an inner product on V.

3.2 Methods

3.2.1 Whitening Transformation

Introduction Otherwise known as **Sphering Transformation**, is a linear transformation that transforms a vector of random variables with a known covariance matrix into a set of new RVs whose covariance is the identity matrix, **meaning uncorrelated**[1] **and each have variance 1**. Transformation is called whitening because it changes the input vector to a "white noise vector", zero mean and finite variance, and statistically independent.

$$Cov(X, Y) = 0 \Rightarrow \text{ Uncorrelated vs Independent } P(X, Y) = P(X)P(Y)$$

For column vector X with non-singular covariance matrix Σ and mean $\mathbf{0}$, then

$$Y = WX, W^T W = \Sigma^{-1}$$

gives whitened random vector Y with unit diagonal covariance. There are infinitely many possible whitening matrices W. Common coices are

1. $W = \Sigma^{-1/2}$ (Mahalanobis or ZCA whitening)
2. Cholesky decomposition of Σ^{-1}
3. Eigen-system of Σ

Optimal whitening transform can be singled out via cross-covariance and cross-correlation of X and Y. The unique optimal whitening transformation achieving maximal component-wise correlation between original X and whitened Y is produced by the whitening matrix

$$W = P^{-1/2} V^{-1/2}$$

where P is the correlation matrix and V is the variance matrix. This seems more of a verification, rather than a method to obtain the result, need to check the original citation [1].

Related transformation

- Decorrelation transform: removes only the correlations but leaves variances intact
- Standardization transform sets variances to 1 but leaves correlations intact
- Coloring transformation transforms a vector of white random variables into a random vector with a specified covariance matrix

3.2.2 Cholesky decomposition

Introduction Otherwise known as Cholesky factorization, is a decomposition of a Hermitian positive-definite matrix to the product of a lower triangular and its conjugate transpose, which is useful for efficient numerical solutions, e.g. Monte Carlo simulations [2].

Statement The cholesky decomposition of a Hermitian positive-definite matrix

$$A = LL^*$$

L is lower triangular matrix with real and positive diagonal entries, and L^* denotes the conjugate tranpose of L. Every A has a unique Cholesky decomposition. Converse holds trivially. Positive semi-definite matrix holds similarly, except the diagonal are allowed to be zero.

[1] Not necessarily independent

[2] When applicable, roughly twice as efficient compared to LU decomposition for solving linear systems [2]

Relevant Methods: LDL

$$A = LDL^* = LD^{1/2}(D^{1/2})^*L^* = \left(LD^{1/2}\right)\left(LD^{1/2}\right)^*$$

L is a lower triangular matrix, and D is a diagonal matrix. Main advantage being that LDL decomposition can be computed and used with essentially the same algorith, but **avoids extracting square roots**. Therefore often called *square-root-free* Cholesky decomposition.

Otherwise called as LDLT decomposition for real matrices. Closely related to the eigen decomposition of real symmetric matrices, $\mathbf{A} = \mathbf{Q\Lambda Q^T}$, see section 3.2.4.

Applications Mainly used for the numerical solution of linear equations $\mathbf{Ax} = \mathbf{b}$. If $\mathbf{A}$ is symmetric and positive definite, then we can solve it by computing Cholesky decomposition $\mathbf{A} = \mathbf{LL}^*$, then $\mathbf{Ly} = \mathbf{b}$ by forward substitution, and finally $\mathbf{L}^*\mathbf{x} = \mathbf{y}$ by back substitution.

Alternative way to eliminate taking square roots in the $\mathbf{LL}^*$ is via LDLT decomposition.

Computation The cholesky algorithm is a modified version of Gaussian elimination. It recursively computes matrix A in the form that

$$A^{(i)} = LA^{(i+1)}L^*$$

Where $A^{(1)} = A$, and at each step, $A^{(i)}$ gives meaning in the following form

$$A^{(i)} = \begin{bmatrix} I & 0 & 0 \\ 0 & a_{i,i} & \mathbf{b}^* \\ 0 & \mathbf{b} & \mathbf{B} \end{bmatrix}, L = \begin{bmatrix} I & 0 & 0 \\ 0 & \sqrt{a_{i,i}} & 0 \\ 0 & \frac{1}{\sqrt{a_{i,i}}}\mathbf{b} & I \end{bmatrix}$$

We compute recursively in this manner until the following $\mathbf{A}^{i+1}$ matrix becomes identity matrix.

$$A^{i+1} = \begin{bmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \mathbf{B} - \frac{1}{a_{i,i}}\mathbf{bb}^* \end{bmatrix}$$

This operation in total operates in about $O(n^3)$, roughly about $n^3/3$ FLOPs.

3.2.3 QR Decomposition

Introduction Decomposition of a matrix A into a product $A = QR$ of orthogonal matrix Q and upper triangular matrix R.

Square Matrix For real square matrices, $A = QR$, if A is invertible, then factorization is unique if we require the diagonal elements of R to be positive. For n linearly independent columns in A, first n columns of Q form an orthonormal basis for the column space of A.

Rectangular Matrix More generally, we factor a complex $m \times n, (m \geq n)$ matrix A. Product would be $m \times mQ$ and $m \times n$ uppper triangular matrix R.

$$A = QR = Q\begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

Computation Gram-Schmidt method, Householder reflection. numerical instability and add others, explain how to use householder reflection

Application Least squares problem

$$A^*Ax = A^*b$$
$$R^*Rx = (QR)^*QRx = R^*Q^*b \Leftrightarrow Rx = Q^*b$$

There are a couple of advantages to this method

1. Numerical stability (A^TA is worse than A, since the ratio of eigen vector is bigger)
2. Faster

3.2.4 Eigendecomposition of a Matrix

Introduction Factorization of a matrix into a canonical form, whereby matrix is represented in terms of its eigenvalues and eigenvectors. Only diagonalizable[3] matrices can be factorized in this way.

Statement Let A be square matrix $n \times n$ with n linearly independent eigenvectors q_i, then A can be factorized as

$$A = Q\Lambda Q^{-1}$$

where Q is the square matrix $n \times n$ whoce ith column is the ith eigenvector of A and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. The decomposition can be derived from the fundamental property of eigenvectors:

$$Av = \lambda v \Rightarrow AQ = Q\Lambda \Rightarrow A = Q\Lambda Q^{-1}$$

Other characteristics Eigendecomposition can be used for example as matrix inversion.

$$\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{-1}$$

If the original matrix is symmetric, then since Q is formed from the eigenvectors, it is guaranteed to be an orthogonal matrix, therefore $Q^{-1} = Q^T$.

3.2.5 Gram–Schmidt Process

Introduction The GS process is a method for orthonormalizing a set of vectors in an inner product space. The process takes a finite, linearly independent set $S = \{v_1, \cdots v_k\}$ and generates an orthogonal set $S' = \{u_1, \cdots, u_k\}, k \leq n$ that spans the same k-dimensional subspace of $\mathcal{R}^n$ as S.

Application of the GS process to the column vectors of a full column rank matrix yields the QR decomposition.

Computation Naïvely, we compute the the set via $\text{proj}_u(v) = \frac{\langle u,v\rangle}{\langle u,u\rangle}u$

$$\begin{aligned}
u_1 &= v_1 \\
u_2 &= v_2 - \text{proj}_{u_1}(v_2) \\
u_3 &= v_3 - \text{proj}_{u_1}(v_3) - \text{proj}_{u_2}(v_3) \\
&\vdots \\
u_k &= v_k - \sum_{j=1}^{k-1} \text{proj}_{u_j} v_k
\end{aligned}$$

[3] there exists an invertible matrix P and a diagonal matrix D such that $P^{-1}AP = D$

Numerical Stability When implemented on a computer, the vectors are often not quite orthogonal, due to rounding errors. For the classical GS process, the loss of orthogonality is particularly bad, therefore it is said that the classical GFS process is numerically unstable.

This process *can* be stablized by a small modification, sometimes referred to as modified GS (MGS). Gives the same result as the original formula and introduces smaller errors in finite precision arithmetic.

We do this by iteratively project onto the previous computed vector

$$\begin{aligned} u_k^{(1)} &= v_k - \text{proj}_{u_1}(v_k) \\ u_k^{(2)} &= u_k^{(1)} - \text{proj}_{u_1}(u_k^{(1)}) \\ &\vdots \\ u_k^{(k-1)} &= u_k^{(k-2)} - \text{proj}_{u_{k-1}}(u_k^{k-2}) \\ u_k &= \frac{u_k^{(k-1)}}{||u_k^{(k-1)}||} \end{aligned}$$

Alternatives Other orthogonalization algorithms includes householder transformation or Givens rotations, or the use of Cholesky decomposition.

Householder are more stable than the MGS, but Householder produces all vector only at the end, making GS process applicable for other methods.

3.2.6 Householder Reflection

Introduction A linear transformation that describes a reflection about a plane or hyperplane containing the origin.

4 Optimization

4.1 Methods

4.1.1 Steepest Descent

4.1.2 Gauss-Newton

4.1.3 Levenberg-Maquardt

4.1.4 Dogleg

References

[1] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. "Optimal Whitening and Decorrelation". In: *The American Statistician* 72.4 (Jan. 2018), pp. 309–314. ISSN: 1537-2731. DOI: 10.1080/00031305.2016.1277159. URL: https://arxiv.org/abs/1512.00809.

[2] William H.; Saul A. Teukolsky; William T. Vetterling; Brian P. Flannery Press. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University England EPress, 1992, p. 994. ISBN: 0-521-43108-5.