Appendix: Derivations

Sophie L.

May 2025

1 Introduction

Hi! These are some motivations and derivations I write up as I develop on my project. No definite structure to this, but I think it's important as a mathematician and researcher to understand the theory behind what I implement in code (although some of this is just interesting theory I randomly discover along the way).

2 Cholesky Decomposition

The Cholesky decomposition is a computational linear algebra technique. We have $C, L \in M_{n\times n}(R)$, with entries indexed as such:

$$C = \begin{bmatrix} C_{0,0} & \cdots & C_{0,n-1} \\ \vdots & \ddots & \vdots \\ C_{n-1,0} & \cdots & C_{n-1,n-1} \end{bmatrix}, L = \begin{bmatrix} L_{0,0} & \cdots & L_{0,n-1} \\ \vdots & \ddots & \vdots \\ L_{n-1,0} & \cdots & L_{n-1,n-1} \end{bmatrix}$$

We seek a lower triangular matrix L such that

$$C = LL^T$$

Since L is lower-triangular, any entry l_{ij} with $i < j$ is 0. So we can write it as

$$L = \begin{bmatrix} L_{0,0} & 0 & \cdots & 0 \\ L_{1,0} & L_{1,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{n-1,0} & L_{n-1,1} & \cdots & L_{n-1,n-1} \end{bmatrix}$$

Things get weird when entries are complex. For simplicity, I'll assume the decomposition matrix L is real-valued. Note, this is guaranteed when C is a symmetric positive-definite matrix (see section 3).

2.1 Motivation on Gaussian Distributions

For motivation, I am using this technique in the context of sampling a d-dimensional multivariate Gaussian distribution, whose parameters are the mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and the covariance matrix $C \in \mathbb{R}^{d \times d}$. We seek to sample a random vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, C)$. Doing this directly is difficult since the correlation structure can be complicated. Instead, we can define a variable $\mathbf{U}$ that follows the standard normal, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Let L be the matrix resulting from the Cholesky Decomposition of C. It turns out, if we sample $\mathbf{U}$ and set

$$\mathbf{X} = \boldsymbol{\mu} + L\mathbf{U}$$

then $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, C)$ as desired. I believe this is actually how Pytorch does it internally.

It's worth proving this fact. Doing so definitely helped solidify the intuition for me since the multivariate case is highly non-obvious at first glance. We write out $\boldsymbol{\mu} = (\mu_0, .., \mu_{d-1}), \mathbf{U} = (U_0, ..., U_{d-1})$. Expanding out $\mathbf{X} = \boldsymbol{\mu} + L\mathbf{U}$ yields

$$X_0 = \mu_0 + L_{00}U_0$$

$$X_1 = \mu_2 + L_{10}U_0 + L_{11}U_1$$

$$\dots$$

Consider the general

$$X_k = \mu_k + L_{k0}U_0 + L_{k1}U_1 + ... + L_{kk}U_k$$

Applying linearity of expectation,

$$E[X_k] = \mu_k + L_{k0}E[U_0] + ... + L_{kk}E[U_k]$$

By assumption, each U_0 follows the standard normal, which is centered at 0. Thus, we've shown $E[X_k] = \mu_k$. Also, it's known that the sum of normally distributed random variables is normal, so X_k is normally distributed. No worries there!

It remains to prove the desired covariance structure. Let us take $i, j < d$ and without loss of generality, let $i \leq j$.

$$Cov(X_i, X_j) = Cov(L_{i0}U_0 + L_{i1}U_1 + ... + L_{ii}U_i, L_{j0}U_0 + L_{j1}U_1 + ... + L_{jj}U_j)$$

Applying bi-linearity, we get a nasty looking expression:

$$\sum_{k=0}^{i}[Cov(L_{ik}U_k, L_{j0}U_0) + Cov(L_{ik}U_k, L_{j1}U_1) + ... + Cov(L_{ik}U_k, L_{jj}U_j)]$$

$$\sum_{k=0}^{i}[L_{ik}L_{j0}Cov(U_k, U_0) + L_{ik}L_{j1}Cov(U_k, U_1) + ... + L_{ik}L_{jj}Cov(U_k, U_j)]$$

Thankfully, all terms here except $Cov(U_k, U_k) = 1$ equal 0, so we get

$$\sum_{k=0}^{i} L_{ik}L_{jk}Cov(U_k, U_k) = \sum_{k=0}^{i} L_{ik}L_{jk}$$

This shows that

$$Cov(X_i, X_j) = \sum_{k=0}^{i} L_{ik} L_{jk}$$

Now, going back to our matrix $C = LL^T$.

$$C_{ij} = \sum_{k=0}^{n-1} L_{ik} L_{kj}^T = \sum_{k=0}^{n-1} L_{ik} L_{jk}$$

In other words, the (i, j)th entry of C is just the dot product of row i and row j of matrix L, whose entries we must solve for. Once $k > i$, then $L_{ik} = 0$ and analogously for j. We re-write the above sum as

$$C_{ij} = \sum_{k=0}^{min(i,j)} L_{ik} L_{jk}$$

Here, we assumed $i \leq j$ so indeed,

$$C_{ij} = \sum_{k=0}^{i} L_{ik} L_{jk} = Cov(X_i, X_j)$$

proving that the covariance matrix is indeed correct! This shows that in the multivariate case, any Gaussian can be reduced to sampling from the standard form.

The takeaway here is that intuitively, the L matrix makes matrix multiplication work in a way very similar to that of Covariance "splits" across terms. Therefore the structure is preserved. The explicit density function can be written as such (but it is really nasty):

$$\frac{1}{(2\pi)^{d/2}} (\det C)^{-1/2} \exp\{-(1/2)(\mathbf{x} - \boldsymbol{\mu}) C^{-1} (\mathbf{x} - \boldsymbol{\mu})^T\}$$

Anyways, onto the actual derivation of Cholesky. We have C, L defined as above. Recall we derived

$$C_{ij} = \sum_{k=0}^{min(i,j)} L_{ik} L_{jk}$$

Since $C^T = (LL^T)^T = (L^T)^T L^T = LL^T = C$, then C is symmetric. Without loss of generality, assume $i \leq j$.

Case 1 (Equality): $i = j$**:**

$$C_{00} = L_{00}^2$$

$$C_{11} = L_{00}^2 + L_{11}^2$$

$$\cdots$$

$$C_{(n-1),(n-1)} = \sum_{i=0}^{n-1} L_{ii}^2$$

We easily solve, always taking the positive root, to get

$$L_{00} = \sqrt{C_{00}}$$

$$L_{11} = \sqrt{C_{11} - C_{00}}$$

$$\dots$$

$$L_{kk} = \sqrt{C_{kk} - C_{(k-1),(k-1)}}$$

This initializes all n diagonal entries of L.
Case 2 (Strict Inequality): $i < j$

$$C_{ij} = \sum_{k=0}^{i} L_{ik}L_{jk} = \sum_{k=0}^{i-1} L_{ik}L_{jk} + L_{ii}L_{ji}$$

We isolate for L_{ji}:

$$L_{ji} = \frac{1}{L_{ii}}(C_{ij} - \sum_{k=0}^{i-1} L_{ik}L_{jk})$$

This expresses L_{ji} in terms of $C_{i,j}$ and entries L_{km}, where $k \leq j, m \leq i$. This actually gets us the numerical solution – not in closed form, but in dynamic-programming style.

We go row by row. First, we set $L_{00} = \sqrt{C_{00}}$. Then, we can set L_{10}, L_{11} using the formula above. L_{10} only relies on L_{00}, and then once we get that value, L_{11} only relies on L_{00}, L_{10}. We go to the next row, and so forth, until the last entry we solve is $L_{(n-1),(n-1)}$ giving us the entire L matrix!

3 Positive Semi-definiteness

3.1 Remarks

As mentioned before, the Cholesky Decomposition requires a matrix that is positive semidefinite. Fortunately, the covariance matrix turns out to have this property so we don't need to check separately. I first lay out some definitions of positive semi-definite and prove that they are equivalent, from which it follows that the covariance matrix (of the form XX^T) must be positive semi-definite (abbrev: PSD).
Again, for my purposes, the complex case isn't too relevant so I'll focus on proving core results over $\mathbb{R}$ without worrying about doing so in full generality. We will culminate in the Spectral Theorem for Singular Value Decomposition.

Theorem 3.1. *Let a symmetric $n \times n$ matrix A be positive semi-definite. Then, the following definitions are equivalent characterizations.*

1. $\forall x \in \mathbb{R}^n, x^T Ax \geq 0$
2. *All eigenvalues λ_i of A are non-negative*
3. *There exists a matrix B such that $A = B^T B$*

4. There exists a lower triangular matrix L such that $A = LL^T$

Quick justification on why we're restricting A to a symmetric matrix. Generally, they're easier to work with and. Definitions (3) and (4) only apply to symmetric matrices. Also, each matrix A can be represented uniquely as the sum of a symmetric and skew-symmetric matrix where $A = B + B'$, with $B = \frac{A+A^T}{2}$ being symmetric and $B' = \frac{A-A^T}{2}$ being skew symmetric. The quadratic form equals

$$x^T Ax = x^T Bx + x^T B'x$$

where one can verify that $x^T B'x$ is 0 due to B' being skew symmetric. So, this shows for each matrix, only the symmetric part determines the quadratic form.

Proof. In this proof, we only show that definition 1 and 2 are equivalent.

Assume 1 holds. For the sake of contradiction, assume there exists a negative eigenvalue λ_i corresponding to eigenvector v_i. Then,

$$v_i^T Av_i = v_i^T(\lambda_i v_i) = \lambda_i \langle v_i, v_i \rangle = \lambda_i |v_i|^2 < 0$$

This contradicts our quadratic form assumption, so all eigenvalues must be non-negative.

For the backwards direction, assume 2 holds. The standard proof involves the Spectral Theorem for Symmetric Matrices. But first, we must state one fact. Each symmetric matrix A has an orthonormal eigenbasis (pairwise orthogonal, unit length).
I will give a partial justification (for the orthogonal part only). Suppose we have v_1, v_2 in the eigenbasis with corresponding λ_1, λ_2. If $\lambda_1 = \lambda_2$, then we can guarantee that v_1, v_2 are orthogonal since every subspace (here, $ker(A - \lambda_1 I)$) has a basis, which can be turned into an orthogonal basis via the Gram-Schmidt process. Now, suppose $\lambda_1 \neq \lambda_2$, then

$$\lambda_1 v_1^T v_2 = (\lambda v_1)^T v_2 = (Av_1)^T v_2 = v_1^T A^T v_2$$

Since A is symmetric, $A = A^T$ so this becomes $v_1^T Av_2 = \lambda_2 v_1^T v_2$ Since $\lambda_1 v_1^T v_2 = \lambda_2 v_1^T v_2$ for distinct lambdas, then $v_1^T v_2 = 0$ as desired. It is a little more work to count the dimensions of each eigenspace and show they indeed form a basis. (must show each eigenvalue's algebraic multiplicity, in the characteristic polynomial, is equal to the geometric multiplicity, the dimension of the eigenspace, too much work right now). I'll revisit later if I have time. Onto the Spectral theorem!

Theorem 3.2. *Let A be an $n \times n$ symmetric matrix. Then, A is diagonalizable as $A = PDP^{-1} = PDP^T$, where P is an orthonormal matrix and D is a diagonal matrix.*

(Recall I am not proving this, but rather using it in service of proving definition 2 implies definition 1 of PSD.)
Here, P is the matrix with column vectors $v_1, ..., v_n$ with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$.

Let

$$PDP^T = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \end{bmatrix}$$

$$= \begin{bmatrix} | & | & & | \\ \lambda_1 v_1 & \lambda_2 v_2 & \cdots & \lambda_n v_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \end{bmatrix}$$

Claim: Formal proof won't be given, but from here it can be worked out that $PDP^T = \lambda_1 v_1 v_1^T + ... + \lambda_n v_n v_n^T$. Evaluating any arbitrary quadratic form yields:

$$x^T Ax = x^T(PDP^T)x = \lambda_1 x^T v_1 v_1^T x + ... + \lambda_n x^T v_n v_n^T x$$

Consider arbitrary $\lambda_i x^T v_i v_i^T x = \lambda_i \langle x, v_i \rangle \cdot \langle v_i, x \rangle = \lambda_i \langle x, v_i \rangle^2$. Since $\lambda_i \geq 0$, each term $\lambda_i x^T v_i v_i^T x \geq 0$ in the summation, thus the quadratic form is always non-negative as desired.

□

We've now shown definitions (1) and (2) are equivalent. For thoroughness, I quickly sketch how we include definition (3) in this.

Proof. Assume (3) holds so $A = B^T B$ for some matrix B. Then,

$$x^T Ax = x^T B^T Bx = (Bx)^T (Bx) = \langle Bx, Bx \rangle = |Bx|^2 \geq 0$$

Hence, (3) implies (1). We've shown (1) implies (2). Now, assume (2) holds. Applying the spectral theorem, we the decomposition

$$A = PDP^T$$

Since all the eigenvalues are non-negative, we have

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}^2 = Q^2$$

So $A = (PQ)(QP^T) = (PQ)(Q^T P^T)$ since Q is a diagonal matrix. Thus,

$$A = (PQ)(Q^T P^T) = (Q^T P^T)^T (Q^T P^T)$$

, so setting $B = Q^T P^T$ satisfies $A = B^T B$, so (3) holds true. This completes our cycle of (1) implies (2) implies (3) implies (1), so these are all equivalent. □

3.2 Applications of the Spectral Theorem

Since the Spectral Theorem is about diagonalization, this naturally makes us think about changing coordinates from our regular $\mathbb{R}^d$ to the eigenbasis. Let $X \in \mathbb{R}^{d \times n}$ so each column vector is an observation in random variables $X_1, ..., X_d$. For convenience, assume X has already been normalized (so $E(X_1) = ... = E(X_d) = 0$ so the covariance matrix is $XX^T = PDP^T$, where P has column vectors consisting of an orthonormal eigenbasis (possible since XX^T is symmetric).

$$P = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix}$$

We change coordinates by multiplying $P^{-1}X$. Each column vector can be thought of an observation in $E_1, ..., E_d$ where E_i is the ith eigen-coordinate. We can get the covariance matrix of our new variables:

$$(P^{-1}X)(P^{-1}X)^T$$

Again, since P holds an orthonormal basis we have $P^{-1} = P^T$ so this equals

$$(P^{-1}X)X^TP = P^{-1}(PDP^{-1})P = D$$

We have discovered two facts. From the diagonal entries, $Var(E_i) = \lambda_i$ which means the eigenvector axis with the largest eigenvalue has the highest variance. WLOG, let λ_1 be the largest. Then, E_i explains $\frac{\lambda_1}{\sum_{i=1}^{d}\lambda_i}$ of the variance in the dataset. For $i \neq j, Cov(E_i, E_j) = 0$, so different axes are uncorrelated.

This is a very well-known fact in data analysis, made even more satisfying by understanding the linear algebra behind it!

Diagonalization also enables us to uncover beautiful properties of the matrix. The Cayley-Hamilton Theorem states the following:

Theorem 3.3. *Let A be a square $n \times n$ matrix with a geometric polynomial $p_A(\lambda) = \det(A - \lambda I)$. Then, A is a root of the p_A.*

Even though this holds for all matrices. we will restrict A to a diagonalizable matrix since that is the essence of the idea. To prove the general form, one would leverage the fact that any matrix can be approximated by diagonalizable matrices over the complex numbers, then apply a limiting argument (omitted here).

Proof. Based on how determinants are calculated (omitted here), the characteristic polynomial must have degree n. Let $p_A(\lambda) = c_n\lambda^n + c_{n-1}\lambda^{n-1} + ... + c_0$. We seek to compute

$$p_A(A) = c_nA^n + ... + c_1A + c_0I$$

We can diagonalize $A = PDP^{-1}$, so the matrix powers are $A^k = PD^kP^{-1}$. Substituting in, we get

$$p_A(A) = P(c_nD^n)P^{-1} + P(c_{n-1}D^{n-1})P^{-1} + ... + P(c_1D)P^{-1} + c_0I$$

We re-write $c_0I = P(c_0I)P^{-1}$ and factor:

$$= P(c_nD^n + c_{n-1}D^{n-1} + ... + c_1D + c_0I)P^{-1}$$

We realize

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Exponentiating D and merging the constants, the above expression equals

$$P \begin{bmatrix} \sum_{i=0}^{n} c_i\lambda_1^i & 0 & \cdots & 0 \\ 0 & \sum_{i=0}^{n} c_i\lambda_2^i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i=0}^{n} c_i\lambda_n^i \end{bmatrix} P^{-1}$$

The center matrix equals 0 since $\lambda_1, ..., \lambda_n$ are eigenvalues, thus roots of p_A. This proves that $P_A(A) = 0_n$, the 0 matrix! $\square$

3.3 Applications in Stochastic Matrices and Markov Chains