Stats 202B: Matrix Algebra and Optimization

Melody Y. Huang

University of California, Los Angeles
Department of Statistics

"Matrices are just large vectors" (Amini, 2019).

1 Definitions

Definition 1.1 (Trace)
The trace of a square matrix X is the sum of the diagonal elements. For a general matrix (not necessarily square), we can take the trace of the cross product (i.e., $X^\top X$) or the transpose of the cross product (i.e., $XX^\top$), both of which are square.

Properties of Traces

1. For an arbitrary matrix X,

$$\text{tr}(XX^\top) = \text{tr}(X^\top X) = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{i,j}^2 := ||X||_F^2$$

   Note that an implication of this is that given a square matrix, the trace of the transpose of a square matrix is equivalent to the trace of the original matrix. In this case, the square matrix we are dealing with $XX^\top$. $X^\top X$ is the transpose of $XX^\top$. Therefore, the trace is equivalent. Intuitively, this is because the trace is the sum of the diagonal elements of a matrix, and the diagonal elements of a transposed matrix do not change.

2.

$$\begin{aligned}\text{tr}(A^\top B) &= \sum_{j=1}^{m}(A^\top B)_{j,j} \\ &= \sum_{j=1}^{m}\left(\sum_{i=1}^{n}(A^\top)_{j,i}B_{i,j}\right) \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} A_{i,j}B_{i,j}\end{aligned}$$

3. For arbitrary matrices A, B, C:

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$$

   Note that this requires A, B, C to be compatible dimension wise. This implies that we can permute the order of the matrix multiplication within a trace, as long as the dimensionality works out. It is worth noting that most of the time, the multiplication will not commute.

Definition 1.2 (Permutation Matrix)
A permutation matrix is a binary square matrix, where the rows and columns sum to one (i.e., each row has exactly one element equal to one, and each column has exactly one element equal to one).

The identity matrix is an example of a permutation matrix. (We can think about this as an example in which nothing in the original matrix gets moved around.) Furthermore, permutation matrices are all orthonormal matrices.

Properties of Permutation Matrices

1. For a matrix X with n rows, there exist $n!$ permutation matrices.
2. If P_1 and P_2 are permutation matrices, then P_1P_2 and P_2P_1 are also permutation matrices.

   Justification: Since P_1 is a permutation matrix, it will permute any matrix it is multiplied by. Since P_2 is also a permutation matrix, P_1P_2 is simply a permuted version of P_2, which remains a permutation matrix. The same is true for P_2P_1.
3. A permutation matrix is non-singular (i.e., invertible).

   Justification: This comes from the fact that permutation matrices are orthonormal. Anything that is orthogonal is invertible, due to the fact that the inverse of an orthogonal matrix is simply the transpose of the matrix.
4. The inverse of a permutation matrix is defined as the transpose of itself:

   $$P^{-1} = P^\top$$

   A basic implication of this is that:

   $$P^\top P = PP^\top = I$$

Definition 1.3 (Linear Dependence)
A set of vectors $\{\mathbf{v}_i\}$ is linearly dependent if there exists a set $\{\theta_i\}$ where at least one $\theta_i \neq 0$ such that:

$$\sum_{i=1}^{n} \theta_i \mathbf{v}_i = 0$$

When no such set $\{\theta_i\}$ exists, then we say that $\{\mathbf{v}_i\}$ is linearly independent.

In the context of matrices, we treat the columns of a matrix, or the rows of a matrix, as the set of vectors of interest. If the columns of a matrix are linearly independent, we say that the matrix if **column non-singular**.

Properties of Column Singularity

1. If X is column singular (i.e., there exists some linear dependence between the columns of X), then the augmented matrix $[X|Y]$ is also column singular.

   Justification: Because X is column singular, there exists some nonzero a s.t. $Xa = 0$ (by definition). Therefore, because a is nonzero, no matter what Y is, we can create a modified a' defined as the original nonzero a, padded with zeros for all entries corresponding to Y, such that a' is also nonzero, but $[X|Y]a' = 0$. Therefore, since a' is nonzero, $[X|Y]$ is also column singular.
2. If X is column non-singular, then $[X|Y]$ will be column singular if and only if there is a B such that $Y = XB$.

Justification:
$\Longrightarrow$: Since X is non-singular, $Xv = 0$ iff $v = 0$. Assume $[X|Y]$ is column singular. Let v' be a non-zero vector. We can write this as:

$$\begin{bmatrix} X & Y \end{bmatrix} \begin{bmatrix} v' \\ -b \end{bmatrix} = 0$$

Note that $\begin{bmatrix} v' \\ -b \end{bmatrix}$ is a non-zero vector (since v' is nonzero). As such:

$$Xv' - Yb = 0$$

If $b = 0$, then this would imply that Xv' is zero, but v' is non-zero. Therefore, $b \neq 0$ (or this would contradict X's non-singularity. As such:

$$Xv' = Yb \Longrightarrow Xv'b^{-1} = Y$$

Defining $B = v'b^{-1}$, we have shown $XB = Y$.
$\Longleftarrow$: Let $Y = XB$. Then:

$$\begin{bmatrix} X & Y \end{bmatrix} = \begin{bmatrix} X & XB \end{bmatrix}$$

$$\begin{bmatrix} X & XB \end{bmatrix} \begin{bmatrix} B \\ -1 \end{bmatrix} = XB - XB = 0$$

As such, $[X|Y]$ is a non-singular matrix.

3. If X has more columns than rows, then X is column singular. In other words, if X has more columns than rows, there must exist linear dependence between the columns.

4. A square matrix is non-singular if it is row non-singular and column non-singular. A square matrix will be singular if it is not non-singular.

5. If A is a triangular matrix with non-zero diagonal, then A is non-singular (i.e., it must have linearly independent columns and rows).

   Justification:
   Consider this low dimensional example:

$$\begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

   If we rewrite this as a system of equations, we see that it implies:

$$\begin{aligned} a_{11}x_1 = 0 &\Longrightarrow x_1 = 0 \\ a_{21}x_1 + a_{22}x_2 = 0 &\Longrightarrow x_2 = 0 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = 0 &\Longrightarrow x_3 = 0 \end{aligned}$$

Lemma 1.1
Let A be a non-singular matrix. Then YA is column non-singular if and only if Y is column non-singular.

Proof: Proof accidentally shown under properties (above). $\Longrightarrow$: Let Y be column non-singular. This implies that all of the columns of Y are linearly independent. It is given that A also has linearly independent columns (and rows). As such, the only time $Ab = 0$ is when b is zero. ... show later □

Definition 1.4 (Rank)
The column-rank of an $n \times m$ matrix X is the number of linearly independent columns of X. Alternatively, we can think of it as the size of the basis spanning X.

We can identify the rank of a matrix using a full-rank decomposition, in which we take the matrix X and identify an $n \times p$ matrix Y and a $m \times p$ matrix A such that:

$$X = YA^\top$$

Both Y and A will be column non-singular (i.e., all the columns are linearly independent). Using the full-rank decomposition, the rank of X is thus equal to p. Examples of full-rank decompositions include the LDU decomposition, QR decomposition, Cholesky decomposition, and SVD.

Row-rank is defined similarly, but instead of using a full-rank decomposition on X, we perform a full rank decomposition on $X^\top$. Because of this definition, it is simple to see that the row-rank and the column-rank of a matrix is equivalent:

$$X = YA^\top \implies X^\top = (YA^\top)^\top = AY^\top$$

The dimensions of A are $m \times p$, and the dimensions of $Y^\top$ will be $p \times n$.

Rank Nullity Theorem

Properties

1. If a matrix $X \in \mathbb{R}^{n \times m}$ is column-singular, then the following is true:

$$\text{Rank}(X) < m$$

$$\text{Nullity}(X) > 0$$

   This is a basic consequence of the Rank-Nullity theorem.

2. Let X be an $m \times n$ matrix, where $m > n$ (more rows than columns). If X is of full column-rank, then there exists some $m \times (n - m)$ matrix Y of full column rank such that $[X|Y]$ is non-singular.

   Justification: Since X has full column rank, it has column rank n, and nullity $m - n$. Append an $m - n \times m - n$ identity matrix to a bunch of zeros to make Y. Basis for the nullity??? rewrite later

3. If $X = 0$, then $\text{Rank}(X) = 0$.

Theorem 1.1 *Let X be an arbitrary $m \times n$ matrix. Multiplying a matrix X with a square non-singular matrix does not change its rank or nullity. Therefore, if S and T are non-singular matrices, then:*

$$\text{Rank}(SXT) = \text{Rank}(X)$$

Proof: Let S be a non-singular matrix. Let X also be a non-singular matrix. Consider the case of:

$$SXa = 0$$

If we treat Xa as a vector b, then:

$$Sb = 0$$

Since S is non-singular, the only time this would be true is if $b = 0$. However, $b = Xa$. As a result, the only time $Xa = 0$ is when $a = 0$, since X is also a non-singular matrix. Therefore, $SXa = 0$ only occurs when $a = 0$, indicating that SX is a non-singular matrix as well. The same can be shown for XT.

□

Theorem 1.2 *Let X be an arbitrary $m \times n$ matrix. Then the following is true:*

$$\text{Rank}(X) = \text{Rank}(X^\top X) = \text{Rank}(XX^\top)$$

Proof: Let $X = YA^\top$ be a full rank decomposition. Then:

$$X^\top X = (YA^\top)^\top YA^\top = AY^\top YA^\top$$

Note that the matrix $AY^\top Y$ is column non-singular due to the fact that both Y and A are non-singular matrices (and the product of two non-singular matrices will be non-singular, by Lemma 1.1). Furthermore, $AY^\top Y$ will be dimension $m \times p$. Therefore, the rank of $X^\top X$ will still be p. The same logic follows for $XX^\top$.

□

Definition 1.5 (Orthonormal Matrices)
An $n \times m$ matrix X is column-wise orthonormal if $X^\top X$ is diagonal. Likewise, a matrix X is row-wise orthonormal if $XX^\top$ is diagonal.

This implies that the inner product of all of the columns of X will equal 0, except for the inner product of a column with itself, which is equal to 1. In other words:

$$\langle X_i, X_j \rangle = 0, \ \forall \ i \neq j$$

Theorem 1.3 *Let X be a column-wise orthonormal $n \times m$ matrix. Then $Xa = 0$ if and only if $a = 0$.*

Proof:
Let a be such that $Xa = 0$. Assume that a is non-zero:

$$Xa = 0 \implies X^\top Xa = 0$$

However, by orthonormality, $X^\top X = I$. This would imply then that $a = 0$, which is a contradiction.

□

The implication of this theorem is that the columns of an orthonormal matrix will be linearly independent.

Properties of Orthonormal Matrices

1. If X is a column-wise orthonormal $n \times m$ matrix, then $m \leq n$. In other words, X has more columns than rows (or is square). This is an implication of Theorem 1.3, which states that orthonormal matrices are linearly independent. As such, we know that matrices with less columns than rows are not linearly independent (i.e., singular).

2. If X is an $n \times m$ column-wise orthonormal matrix, then there is some $n \times (n - m)$ column-wise orthonormal matrix Y such that $X^\top Y = 0$. In this case, we call Y the **orthonormal complement** to X.

   The matrix of $[X|Y]$ is referred to as **square orthonormal**.

   - If X is square orthonormal, then $X^\top X = XX^\top = I$.
   - The transpose of a square orthonormal matrix is also orthonormal.
   - If X and Y are square orthonormal matrices, then so are XY and YX.

Definition 1.6 (Orthogonal Projector)
A symmetric matrix P is a projector if $P^2 = P$. (By extension, $P^n = P$.) If X is column-wise orthonormal, then $XX^\top$ is a projector.

Definition 1.7 (System of Linear Equations)
A system of linear equations is represented as:

$$Ax = b$$

*When $b = 0$, we say the system is **homogenous**. (Thus, when $b \neq 0$, the system is inhomogenous.) We can represent any inhomogenous system as a homogenous system. All solutions take on the form of $x + d$.*

Lemma 1.2 *If x and z are solutions to the linear equations $Ax = b$, then $z = x + d$, where d is defined such that $Ad = 0$.*

Proof: We can re-write d as $d = z - x$. As such, then:

$$Ad = A(z - x) = Az - Ax = b - b = 0$$

□

Furthermore, $Ax = b$ has at most one solution iff $Ax = 0$ has the unique solution of $x = 0$. In other words, if A is of full column-rank, the system of equations has at most one solution.

Definition 1.8 (Matrix Inverse)
An $n \times n$ matrix X has an inverse if there exists some matrix A where:

$$XA = AX = I_n$$

Usually, A is written as X^{-1}.

It is important to note that not all matrices are invertible. In fact the majority of matrices are not invertible! Common matrices that we are forced to deal with that are not invertible include any non-square matrices and any singular matrices (i.e., matrices with linearly dependence between columns/rows).

As a random note: We can identify singular matrices by looking at the determinant. If the determinant is 0, then the matrix is singular.

Definition 1.9 (Generalized Inverse)
A generalized inverse of a matrix A is defined as any matrix G such that

$$AGA = A$$

It is worth noting that the generalized inverse of a singular matrix A will not be unique. In the case that A *is* in fact non-singular, the generalized inverse will be unique and will be equal to the inverse of A (i.e., $G = A^{-1}$). The Moore-Penrose inverse is a version of generalized inverse that *is* unique.

Definition 1.10 (Moore-Penrose Inverse)
The Moore-Penrose inverse of a matrix X is defined as any matrix X^+ such that the following four conditions are met:

1. $XX^+X = X$
2. $X^+XX^+ = X^+$
3. $(X^+X)^\top = X^+X$
4. $(XX^+)^\top = XX^+$

Definition 1.11 (Positive Semi-Definite)

2 Matrix Decompositions

2.1 QR Decomposition

The QR decomposition takes an $n \times m$ matrix X and decomposes into QR, where Q is an $n \times m$ orthonormal matrix, and R is an upper-triangular matrix. This effectively takes the matrix X and identifies a basis, Q, for the space that X spans. This can be done using the Gram-Schmidt Algorithm.

Definition 2.1 (Gram-Schmidt Algorithm)

Step 1. *Set q_1 equal to X_1:*

$$q_1 = x_1$$

Step 2. *For $i = 2, ..., n$:*

$$q_i = x_i - \sum_{j=2}^{i} \underbrace{\frac{\langle q_{j-1}, x_i \rangle}{||q_{j-1}||^2} \cdot q_{j-1}}_{\textit{Projection of } x_i \textit{ on } q_{j-1}}$$

Step 3. *Normalize all q_i s.t. they are of unit length:*

$$e_i = \frac{q_i}{||q_i||}$$

Intuitively, Gram-Schmidt takes a set of vectors and gets rid of the components that overlap with one another. This is effectively what is happening when subtracting the projection of each x_i with the q_{i-1} vectors. As a result, the resulting q_i vectors only contain the component of x_i that is orthogonal with the q_{i-1} vectors. By construction, each of the q_i vectors will be orthogonal to each other, since each one recursively subtracts out the overlapping projections.

We can also think of each step as running a regression between x_i and the previous $\{q_{i-1}\}$ vectors. The resulting q_i vector is the residual from this regression which is the leftover component of x_i that is unexplained by the other $\{q_{i-1}\}$ vectors. This is a helpful way to think about things because the contents in the R matrix hold all of these "regression coefficients":

$$r_{ij} = \begin{cases} \frac{x_j^\top q_i}{||q_i||^2}, & j \geq i \\ \\ 0 & \text{else} \end{cases}$$

Need to check if the denominator is squared or not

QR decompositions are guaranteed to exist, regardless of whether or not a matrix is full-rank or not. When a matrix is not full-rank, columns that are linearly dependent on others will result in $q_i = 0$. As such, the number of non-zero q_i is equal to the rank of the matrix X. When we encounter these zero q_i elements, we usually just drop them from the resulting matrix Q. As a result, Q will be of dimension $n \times r$ and R will be of dimension $r \times r$. From here, it is clear that a QR decomposition is a **full rank decomposition**, where r tells us the rank of matrix X.

What happens if we divide by zero? Intuitively, this is not a big issue because we can simply drop all the q_i. However, from an algorithmic standpoint, how does this work? To account for this, we can think of the QR decomposition in terms of:

$$XP = QR,$$

where we can permute the rows of X such that the diagonal elements of R are ordered in descending order: $|r_{11}| \geq ... \geq |r_{nn}|$. When X is not full rank, the last $n - r$ elements of the diagonal element of R will simply be zero. To reconstruct X, we use $X = QRP^\top$.

Why do we like QR decompositions? The factorization of a matrix X into QR allows for simpler calculations. Take the following example:

$$\begin{bmatrix} \vdots & \vdots & \vdots \\ \mathbf{x_1} & \mathbf{x_2} & \mathbf{x_3} \\ \vdots & \vdots & \vdots \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \mathbf{q_1} & \mathbf{q_2} & \mathbf{q_3} \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix}$$

Expanding this out into a system of equations gives us:

$$x_1 = r_{11} q_1$$

What

2.2 LDU Decomposition

The LDU decomposition takes square matrices and rewrites them as a product of a lower triangular matrix L, a diagonal matrix D, and an upper triangular matrix U. L and U are often times written such that they have 1's on the diagonal, making them non-singular. (Why does this make them non-singular?)

Theorem 2.1 *If A is a non-singular matrix, then there exists a permutation matrix P such that PA has an LDU-factorization:*

$$PA = LDU$$

We like LDU decompositions because they make it very easy to invert a matrix:

$$A = LDU \implies A^{-1} = U^{-1}D^{-1}L^{-1}$$

This means all we have to do is invert triangular and diagonal matrices, which is easier than inverting a regular matrix.

2.3 Eigenvalue (Spectral) Decomposition

We will introduce eigenvalues and eigenvectors with respect to something known as the **Raleigh Quotient**.

Definition 2.2 (Raleigh Quotient)
For a real symmetric matrix A of order n, define the Rayleigh Quotient as:

$$\lambda(x) = \frac{x^\top Ax}{x^\top x} = \left(\frac{x}{||x||}\right)^\top A \left(\frac{x}{||x||}\right)$$

Note that $\lambda(x) : \mathbb{R}^{n\times 1} \to \mathbb{R}$ (i.e., maps a vector x to a scalar value).

The first derivative of the Rayleigh Quotient is given as:

$$\frac{\partial \lambda(x)}{\partial x} = \frac{\partial}{\partial x}\frac{x^\top Ax}{x^\top x} = \frac{2}{x^\top x}(Ax - \lambda(x)x)$$

Setting the derivative equal to zero gives us the following:

$$Ax = \lambda(x)x$$

Therefore, we see that solution to the stationary condition given by optimizing the Rayleigh Quotient is equal to the eigenvalues of matrix A.

A different way to think about eigenvalues is to think of the following constrained maximization problem:

$$\begin{cases} \max x^\top Ax \\ \text{where } x^\top x = 1 \end{cases}$$

In other words, we are trying to find the stationary values of $x^\top Ax$ on the unit sphere $x^\top x = 1$. This can be obtained by using Lagrange Multipliers, which gives us the following system of equations:

$$\begin{cases} Ax = \lambda x \\ x^\top x = 1 \end{cases}$$

Solving for (x, λ), we can an eigen-pair, where x is an eigenvector of A and λ is an eigenvalue. Eigen-pairs are stationary values of the Rayleigh quotient.

Lemma 2.1 *Let X be a symmetric matrix with eigenvector v_1 and v_2, with respective eigenvalues λ_1 and λ_2. If $\lambda_1 \neq \lambda_2$, then v_1 and v_2 are orthogonal.*

Proof:

$$\begin{aligned}\lambda_1 v_1^\top v_2 &= (Av_1)^\top v_2 \\ &= v_1^\top A^\top v_2 \\ &\text{By symmetry of } A\text{:} \\ &= v_1^\top A v_2 \\ &= v_1^\top (\lambda_2 v_2) \\ &= \lambda_2 v_1^\top v_2\end{aligned}$$

Therefore, since $\lambda_1 v_1^\top v_2 = \lambda_2 v_1^\top v_2$:

$$(\lambda_1 - \lambda_2) v_1^\top v_2 = 0$$

Since $\lambda_1 \neq \lambda_2$, this implies that $v_1^\top v_2 = 0$, which means that v_1 and v_2 are orthogonal. □

2.4 Singular Value Decomposition

In general, any real $m \times n$ matrix X, where $m \geq n$ (i.e., more rows than columns), can be decomposed into:

$$X = \underbrace{U}_{m \times n} \underbrace{\Lambda}_{n \times n} \underbrace{V^\top}_{n \times n},$$

where U is a column orthonormal matrix ($U^\top U = I$) containing eigenvectors of the symmetric matrix $XX^\top$, Λ is a diagonal matrix containing the singular values, and $V^\top$ is a row orthonormal matrix ($V^\top V = I$) containing the eigenvectors of $X^\top X$.

The number of non-zero singular values in Λ (i.e., number of non-zero elements on the diagonal) is going to be equal to the rank of matrix X. As such, we have the following theorem.

Theorem 2.2 (Singular Value Decomposition)
Let $X \in \mathbb{R}^{m \times n}$ of rank r where $r \leq n \leq m$. Then there exists some $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{r \times n}$, and diagonal matrix $\Lambda \in \mathbb{R}^{r \times r}$ with positive diagonal elements such that:

$$X = U \Lambda V^\top$$

In other words, if X is not full rank, we can represent it with fewer dimensions using a singular value decomposition. As it turns out, the singular value decomposition of a matrix is going to be the *best rank p approximation of the full matrix X*. In other words, if we want a p-dimension representation of X, using the SVD with p singular values will give us the best reconstruction! More formally:

Theorem 2.3 (Singular Value Decomposition as Best Rank p Approximation)
If $X = U\Lambda V^\top$ is the SVD of X, and the singular values are sorted as $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$, then for any $p < r$, the best rank p approximation to X is:

$$\tilde{X} = \sum_{i=1}^{p} \lambda_i u_i v_i^\top,$$

where $||X - \tilde{X}||_F^2 = \sum_{i=p+1}^{r} \lambda_i^2$.

Note that $||X - \tilde{X}||_F^2$ is the amount of error that we have incurred from using the approximation with respect to the Frobenius norm. In other words, it is the unexplained component of X (the residual) that the rank-p approximation did not capture.

To understand why the error term is given by the sum of the remaining $r - p$ singular values squared:

Proof:

$$
\begin{aligned}
||X - \tilde{X}||_F^2 &= \text{tr}((X - \tilde{X})^\top (X - \tilde{X})) \\
&= \text{tr}((U\Lambda V^\top - U\Lambda_p V^\top)^\top (U\Lambda V^\top - U\Lambda_p V^\top)) \\
&= \text{tr}((V\Lambda U^\top - V\Lambda_p U^\top)(U\Lambda V^\top - U\Lambda_p V^\top)) \\
&= \text{tr}(V\Lambda U^\top U\Lambda V^\top - V\Lambda U^\top U\Lambda_p V^\top - V\Lambda_p U^\top U\Lambda V^\top + V\Lambda_p U^\top U\Lambda_p V^\top) \\
&\text{Note that } U^\top U = I\text{:} \\
&= \text{tr}(V\Lambda^2 V^\top - V\Lambda\Lambda_p V^\top - V\Lambda_p\Lambda V^\top + V\Lambda_p^2 V^\top) \\
&= \text{tr}(V(\Lambda - \Lambda_p)(\Lambda - \Lambda_p)V^\top) \\
&= \text{tr}(VV^\top(\Lambda - \Lambda_p)(\Lambda - \Lambda_p)) \\
&\text{Note that } VV^\top = I\text{:} \\
&= \text{tr}((\Lambda - \Lambda_p)(\Lambda - \Lambda_p)) \\
&= \sum_{i=p+1}^{r} \lambda_i^2
\end{aligned}
$$

□

3 Multivariate Models as Matrix Approximations

3.1 Ordinary Least Squares

The objective function we want to minimize is:

$$
\begin{aligned}
L(B) = ||Y - XB||^2 &= \text{tr}((Y - XB)^\top (Y - XB)) \\
\frac{\partial}{\partial B} &= -2X^\top Y + 2X^\top XB = 0 \\
\implies B &= (X^\top X)^{-1} X^\top Y
\end{aligned}
$$

We can use QR decomposition to compute $(X^\top X)^{-1}$. Let $X = QR$ be the QR decomposition, where $Q \in \mathbb{R}^{m \times r}$ and $R \in \mathbb{R}^{r \times r}$.

$$
\begin{aligned}
B &= (X^\top X)^{-1} X^\top Y \\
&= ((QR)^\top QR)^{-1} (QR)^\top Y \\
&= (R^\top Q^\top QR)^{-1} (QR)^\top Y \\
&= (R^\top R)^{-1} R^\top Q^\top Y \\
&= R^{-1} R^{-T} R^\top Q^\top Y \\
&= R^{-1} Q^\top Y
\end{aligned}
$$

Using this decomposition, we can rewrite $Y - XB$ as:

$$
\begin{aligned}
Y - XB &= Y - QRR^{-1}Q^\top Y \\
&= (I - QQ^\top)Y
\end{aligned}
$$

Therefore, we can think of $QQ^\top$ as the classic projection matrix that takes Y and projects it into the column space of X, and $I - QQ^\top$ is the residual maker (which projects Y into the space orthogonal to X). Ordinary least squares is nice because we can re-write Y as a function of the parts of Y that can be explained by X, and the parts of Y that cannot:

$$
||Y||^2 = ||Y - XB||^2 + ||XB||^2
$$

$$= ||(I - QQ^\top)Y||^2 + ||QQ^\top Y||^2$$

Alternatively, we can use SVD. Let $X = U\Lambda V^\top$ be the SVD, where $U \in \mathbb{R}^{m\times r}$, $\Lambda, L \in \mathbb{R}^{r\times r}$.

$$\begin{aligned}
B &= (X^\top X)^{-1}X^\top Y \\
&= ((U\Lambda V^\top)^\top (U\Lambda V^\top))^{-1} \cdot (U\Lambda V^\top)^\top Y \\
&= (V\Lambda U^\top U\Lambda V^\top)^{-1}(V\Lambda U^\top)Y \\
&= (V\Lambda^2 V^\top)^{-1}V\Lambda U^\top Y \\
&= V^{-\top}\Lambda^{-2}V^{-1}V\Lambda U^\top Y \\
&= V^{-\top}\Lambda^{-1}U^\top Y \\
&= V\Lambda^{-1}U^\top Y
\end{aligned}$$

3.2 Orthogonal Least Squares

In ordinary least squares, we attempt to minimize the squared, vertical distances between the actual points and a line that we've fit. However, a more geometrically intuitive way to think about this is instead of minimizing the vertical distance, we want to minimize the the distance between the actual points and the projection of it on the line (i.e., the projection of the point on the fitted line).

Using just a 2-dimensional example (i.e., with only 1 covariate X_i), we want to represent the n original points, given by (X_i, Y_i) with the line $(Z_i, a + bZ_i)$, with the loss function:

$$L(a, b) = \sum_{i=1}^{n}(X_i - Z_i)^2 + \sum_{n=1}^{n}(Y_i - a - bZ_i)^2$$

Z_i is a function of the parameters a, b:

$$Z_i(a, b) = \frac{X_i + (Y_i - a)b}{1 + b^2}$$

Moving into a more general regime with more than 1 covariate, the loss function becomes:

$$L(Z, B) = ||X - Z||^2 + ||Y - ZB||^2$$

Here, akin with the multivariate case of OLS, we lose the intercept term a. We can re-write the loss function to be more concise using a padded matrix $[X|Y]$:

$$L(Z, B) = \left|\left|\left[X|Y\right] - Z\left[I|B\right]\right|\right|^2$$

Using this formulation, we can think about the minimization process with respect to Z and $[I|B]$:

$$\frac{\partial}{\partial Z} = (Z[I|B] - [X|Y])[I|B]^\top = 0 \implies Z[I|B][I|B]^\top = [X|Y][I|B]^\top$$

$$\frac{\partial}{\partial [I|B]} = Z^\top([X|Y] - Z[I|B]) = 0 \implies Z^\top[X|Y] = Z^\top Z[I|B]$$

We can take the transpose on both sides of the second stationary condition: $[X|Y]^\top Z = [I|B]^\top Z^\top Z$. This gives us the following stationary conditions:

$$\begin{cases} Z\left[I|B\right]\left[I|B\right]^\top = \left[X|Y\right]\left[I|B\right]^\top \\ \\ \left[X|Y\right]^\top Z = \left[I|B\right]^\top Z^\top Z \end{cases}$$

We see that Z and $[I|B]$ are equivalent to the singular value decomposition of $[X|Y]$.

3.3 Total Least Squares

Sometimes we want to solve a linear system $Y = XB$ that may not necessarily be consistent. One way to still obtain estimates to the solution to this linear system is to perturb Y and X by small amounts. We can formally write this as the following optimization problem:

$$\min_{B,\varepsilon_X,\varepsilon_Y} \kappa||\varepsilon_X||^2 + \lambda||\varepsilon_Y||^2$$
$$\text{s.t. } Y + \varepsilon_Y = (X + \varepsilon_X)B$$

κ and λ are tuning parameters that determine the weight of how much we care about perturbations with respect to X and Y.

We can re-write this slightly. Let $\tilde{X}$ be the perturbed version of X:

$$\tilde{X} = X + \varepsilon_X \implies \varepsilon_X = \tilde{X} - X$$

Then it follows that:

$$\varepsilon_Y = \tilde{X}B - Y$$

We can then re-write the objective function as:

$$L(\tilde{X}, B) = \kappa||\tilde{X} - X||^2 + \lambda||Y - \tilde{X}B||^2$$

If we make κ very large, this implies that we are penalizing more for any perturbations made to X. As a result, as $\kappa \to \infty$, $\tilde{X} \to X$, and the estimate of B will be equivalent to the OLS estimator.

We can solve for the solution using singular value decomposition. To do so, we rewrite the objective function once more as:

$$L(\tilde{X}, A) = \kappa \cdot ||U - \tilde{X}A||^2,$$

where:

$$U = \begin{bmatrix} X & \frac{\sqrt{\lambda}}{\sqrt{\kappa}}Y \end{bmatrix}$$

and

$$A = \begin{bmatrix} I & \frac{\sqrt{\lambda}}{\sqrt{\kappa}}B \end{bmatrix}$$

As it turns out, the solution for the optimal B and $\tilde{X}$ values that minimizes the loss function can be obtained by computing the SVD of U.

3.4 Ridge Regression

Ridge regression is nice because we can obtain solutions for the linear system $Y = XB$ even when there exists singularities in X. The loss function takes on this form:

$$L(B) = ||Y - XB||^2 + \kappa||B||^2$$

Minimizing this function gives us:

$$B(\kappa) = (X^\top X + \kappa I)^{-1} X^\top Y$$

Now since κ is restricted to be greater than zero, $X^\top X + \kappa I$ is always positive definite (since $X^\top X$ is positive semi-definite, and κI is PD).

Even when X is singular, the following is true:

$$\lim_{\kappa \to 0} B(\kappa) = X^+ Y$$

As it turns out, the MSE of the estimated B value (given by $B(\kappa)$) can be shown with the following property:

$$\frac{\partial \mathbb{E}(||B(\kappa) - B||^2)}{\partial \kappa} < 0$$

This implies that as κ decreases, the MSE also decreases.

3.5 PCA Regression

Recall from earlier that we can express the solution to OLS by using the SVD of X:

$$\begin{aligned} B_{OLS} &= (X^\top X)^{-1} X^\top Y \\ &= ((U\Lambda V^\top)^\top U\Lambda V^\top)^{-1}(U\Lambda V^\top)^\top Y \\ &= (V\Lambda U^\top U\Lambda V^\top)^{-1}(U\Lambda V^\top)^\top Y \\ &= (V\Lambda^2 V^\top)^{-1}(U\Lambda V^\top)\top Y \\ &= V^{-\top}\Lambda^{-2}V^{-1}V\Lambda U^\top Y \\ &= V\Lambda^{-1}U^\top Y \end{aligned}$$

What is nice about this formulation is that we can use the fact that SVD is the best approximation of a matrix, given a fixed rank r. Therefore, we can use only r singular values and vectors in order to have a truncated B vector:

$$B_{PCA} = V_r\Lambda_r^{-1}U_r^\top Y$$

How does this effect the MSE? Let $Y = XB + \varepsilon$, where the error is assumed to be centered at zero, and the covariance of the error term is given by σ^2.

Looking at the variance of B_{PCA}:

$$\begin{aligned} \mathbb{E}(||B_r - \mathbb{E}(B_r)||^2) &= \mathbb{E}(||V_r\Lambda_r^{-1}U_r^\top Y - \mathbb{E}(V_r\Lambda_r^{-1}U_r^\top Y)||^2) \\ &\text{Making use of the fact that } \mathbb{E}(Y) = XB\text{:} \\ &= \mathbb{E}(||V_r\Lambda_r^{-1}U_r^\top Y - V_r\Lambda_r^{-1}U_r^\top XB||^2) \\ &= \mathbb{E}(||V_r\Lambda_r^{-1}U_r^\top (Y - XB)||^2) \\ &= \mathbb{E}(||V_r\Lambda_r^{-1}U_r^\top \varepsilon||^2) \\ &\text{By definition of the Frobenius norm:} \\ &= \mathbb{E}(\mathrm{tr}((V_r\Lambda_r^{-1}U_r^\top \varepsilon)^\top (V_r\Lambda_r^{-1}U_r^\top \varepsilon))) \\ &= \mathbb{E}(\mathrm{tr}(\varepsilon^\top U_r\Lambda_r^{-1}V_r^\top V_r\Lambda_r^{-1}U_r^\top \varepsilon)) \\ &= \mathbb{E}(\mathrm{tr}(\varepsilon^\top U_r\Lambda_r^{-2}U_r^\top \varepsilon)) \\ &= \mathbb{E}(\mathrm{tr}(\varepsilon^\top \varepsilon)\mathrm{tr}(U_rU_r^\top)\mathrm{tr}(\Lambda_r^{-2})) \\ &= \sigma^2\sum_{i=1}^{r}\frac{1}{\lambda_i^2} \end{aligned}$$

Looking at the bias:

$$||B - \mathbb{E}(B_r)||^2 = \mathrm{tr}(B^\top(I - U_rU_r^\top)B)$$

4 Low Rank Approximations

We want to approximate X as a product of $AB^\top$, where $A \in \mathbb{R}^{n\times r}$, and $B \in \mathbb{R}^{m\times r}$. In other words, we want to provide a rank r approximation of X. (Usually, $r << \min\{m, n\}$.) The loss function in this case is:

$$L(A, B) = ||X - AB^\top||_F^2$$

Minimizing the loss function requires us to take the partial derivative of it w.r.t. A and B and setting it equal to zero:

$$\frac{\partial}{\partial A} = (AB^\top - X)B = 0 \Longrightarrow AB^\top B = XB$$

$$\frac{\partial}{\partial B} = A^\top(AB^\top - X) = 0 \Longrightarrow A^\top AB^\top = A^\top X$$

As a result, we have the following stationary conditions:

$$\begin{cases} X^\top A = B(A^\top A) \\ XB = A(B^\top B) \end{cases}$$

We note that simply solving for matrices A, B that meet these conditions would not result in a unique solution. As such, we can place restrictions that require $A^\top A = I$, or $B^\top B = I$, or that both $A^\top A$ and $B^\top B$ are diagonal.

If we go with the restriction that both $A^\top A$ and $B^\top B$ are diagonal containing the singular values (i.e., $A^\top A = B^\top B = \Lambda_r$), then we can rewrite the stationary conditions as:

$$\begin{cases} X^\top A = BA^\top A & \implies X^\top A = B\Lambda_r \\ XB = AB^\top B & \implies XB = A\Lambda_r \end{cases}$$

Without placing restrictions on A and B, directly solving for this is difficult because the problem is non-convex. One way to do this is to use **alternating least squares**.

Definition 4.1 (Alternating Least Squares Algorithm)

Step 1. *Initialize $A^{(1)}$ randomly.*

Step 2. *For $i = 1, 2, ..., n_{iter}$:*

$$B^{(k)} = X^\top A^{(k)}((A^{(k)})^\top A^{(k)})^{-1}$$
$$A^{(k+1)} = X^\top B^{(k)}((B^{(k)})^\top B^{(k)})^{-1}$$

4.1 Case: Missing Data

Assume we can only observe some of the values of X. In that case, we do not care about the approximation $AB^\top$ with respect to the missing values, and only want to penalize approximations for actually observed values. As a result, the loss function becomes:

$$L(A, B, Z) = ||Z - AB^\top||^2$$

where $Z_{i,j}$ is defined as being equal to X_{ij} if X_{ij} is observed. We can then do alternating least squares by solving first for $AB^\top$ (as listed above), and then solving for Z.