STA521: Linear Algebra and Some Linear Model Theory

Víctor Peña

Contents

1 Matrix Algebra

Basically taken from Sam Roweis' notes[1]:

$$
\begin{aligned}
A(B+C) &= AB + AC \\
(A+B)^\intercal &= A^\intercal + B^\intercal \\
(AB)^{-1} &= B^{-1}A^{-1} \\
(A^{-1})^\intercal &= (A^\intercal)^{-1} \\
|AB| &= |A||B| \\
|A^{-1}| &= 1/|A| \\
|A| &= \prod \text{evals}(A) \\
|cA_{n\times n}| &= c^n |A_{n\times n}| \\
\text{tr}(A) &= \sum \text{evals}(A) \\
\text{tr}(ABC) &= \text{tr}(BCA) = \text{tr}(CAB) \\
\text{rank}(A) &= \text{rank}(A^\intercal A) = \text{rank}(AA^\intercal) \\
& X^\intercal X \text{ and } XX^\intercal \text{ are positive-semidefinite and symmetric.}
\end{aligned}
$$

2 Linear Algebra

Most of the results here can be found in Appendix B in Christensen.

2.1 Column Space and Rowspace

Let X be $n \times p$.

$$
\begin{aligned}
C(X) &= \{Xa \in \mathbb{R}^n : a \in \mathbb{R}^p\} \\
N(X) &= \{v \in \mathbb{R}^p : Xv = 0\} \\
p &= \dim(C(X)) + \dim(N(X)) \\
C(X^\intercal) &= N(X)^\perp \text{ (they are orthogonal complements)} \\
N(X^\intercal) &= C(X)^\perp \\
C(X) &= C(X^\intercal X)
\end{aligned}
$$

2.2 Orthogonal Matrices

Let Q be a $n \times n$ matrix.

- Q is orthogonal if its columns form an orthonormal basis of $\mathbb{R}^n$ (that is, if they are orthogonal unit vectors).
- Equivalently, Q is orthogonal if $Q^T = Q^{-1}$.
- If Q is orthogonal, then $|Q|^2 = 1$.

[1] Link: `http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf`. Also of interest: `http://www.cs.nyu.edu/~roweis/notes/gaussid.pdf` (Gaussian identities).

2.3 Spectral Theorem and Decompositions

Spectral Theorem: If X is a symmetric matrix, it has real eigenvalues and it can be decomposed as $X = U\Lambda U^\intercal$, where Λ is a diagonal matrix with the eigenvalues of X and U is an orthogonal matrix with the eigenvectors of X. The rank of X equals the number of nonzero eigenvalues of X. Based on this decomposition, one can define matrix powers as $X^p = U\Lambda^p U^\intercal$.

SVD: Let X be a $n \times p$ matrix, then one can find U $(n \times n)$, Σ $(n \times p)$ and V $(p \times p)$ such that $X = U\Sigma V^\intercal$, where U and V are orthogonal matrices. The columns of U are eigenvectors of $XX^\intercal$, and the columns of V are eigenvectors of $X^\intercal X$. Σ is a rectangular diagonal matrix with the square roots of the eigenvalues of $X^\intercal X$ or $XX^\intercal$ (they are the same).

Cholesky: Any positive definite matrix X can be factorized as $X = LL^\intercal$, where L is a lower triangular matrix.

QR: Any square matrix X can be decomposed as $X = QR$, where Q is an orthogonal matrix and R is upper triangular.

2.4 Generalized Inverses

Definition: A generalized inverse of a matrix X is any matrix X^- such that $XX^-X = X$. Generalized inverses exist for arbitrary matrices.

Results:

- Let X be a symmetric matrix. The Moore-Penrose generalized inverse is defined as follows. First, decompose $X = U\Lambda U^\intercal$. Then $X^- = U\Lambda^- U^\intercal$ where $\lambda_i^- = 1/\lambda_i$ if $\lambda_i \neq 0$ and $\lambda_i^- = 0$ if $\lambda_i = 0$. The Moore-Penrose generalized inverse is nice because it is symmetric and reflexive (i.e. $A^-AA^- = A^-$).
- If G and H are generalized inverses of $X^\intercal X$, then $XGX^\intercal X = XHX^\intercal X = X$ and $XGX^\intercal = XHX^\intercal$.

2.5 Orthogonal Projections

Definition: P is a perpendicular projection operator (ppo) onto $C(X)$ if and only if

- $v \in C(X)$ implies $Pv = v$.
- $v \perp C(X)$ implies $Pv = 0$.

Properties:

- $P = X(X^\intercal X)^- X^\intercal$ is the ppo onto $C(X)$. If $X^\intercal X$ is invertible, $P = X(X^\intercal X)^{-1}X^\intercal$.
- P is a ppo onto $C(P)$ if and only if $PP = P$ (idempotent) and $P^\intercal = P$ (symmetry).
- Ppos are unique.
- If P is ppo onto $C(X)$, then $C(X) = C(P)$.
- If P is ppo onto $C(X)$, $(I - P)$ is ppo onto $C(X)^\perp$.
- Let P_1 and P_2 be ppos onto $C(P_1)$, $C(P_2)$, then $P_1 + P_2$ is the ppo onto $C(P_1, P_2)$ if and only if $C(P_1) \perp C(P_2)$.

- If $\{o_1, o_2, \dots, o_r\}$ is an orthonormal basis of $C(X)$ and we construct O such that its columns are $\{o_1, o_2, \dots, o_r\}$, then $OO^\intercal$ is the ppo onto $C(X)$.
- If P is ppo its eigenvalues are either 0 or 1, and $\text{tr}(P) = \sum \text{evals}(P) = r(P)$.

3 Random Vectors

Definition: Let Y be a random vector such that $E(Y) = \mu$. The covariance matrix of Y is

$$\text{Cov}(Y) = E[(Y - \mu)(Y - \mu)^\intercal].$$

The elements of the diagonal are variances, the off-diagonal elements are covariances.

Properties
Let Y be a random vector with $E(Y) = \mu$ and $\text{Cov}(Y) = \Sigma$, and let A and b be a matrix and a vector with constants, respectively:

- $E(AY + b) = A\mu + b$.
- $\text{Cov}(AY + b) = A\Sigma A^\intercal$.
- **Expectation of a quadratic form:** Assume A is symmetric, then $E(Y^\intercal AY) = \text{tr}(A\Sigma) + \mu^\intercal A\mu$.

4 Multivariate Normal

These are notes I took after watching the videos on the Multivariate Normal that Jeff Miller uploaded to his YouTube[2].

Definition: $Y = (Y_1, \dots Y_n)^\intercal$ follows a Multivariate Normal (MVN) with mean μ and covariance Σ if any linear combination of the components is $v^\intercal Y \sim \text{Normal}(v^\intercal\mu, v^\intercal\Sigma v)$, for $v \in \mathbb{R}^n$. If $|\Sigma| = \det(\Sigma) \neq 0$, Y has the following pdf:

$$f(Y) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left\{-\frac{1}{2}\left[(Y - \mu)^\intercal\Sigma^{-1}(Y - \mu)\right]\right\}.$$

Remark: Recall that if $c \in \mathbb{R}$ and A is an $n \times n$ matrix, then $|cA| = c^n|A|$.

4.1 Zero Correlation and Independence

Let $Y \sim \text{MVN}(\mu, \Sigma)$. For an arbitrary partition

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),$$

then $\text{Cov}(Y_1, Y_2) = \Sigma_{12} = 0$ if and only if Y_1 and Y_2 are independent. That is, uncorrelation implies independence.

[2] YouTube channel: `https://www.youtube.com/user/mathematicalmonk`.

Corollary: $Y \sim \text{MVN}(\mu, \sigma^2 I)$ and $AB^\intercal = 0$, then AY and BY are independent.

Caution! X_1 and X_2 normally distributed does not imply $(X_1, X_2) \sim \text{MVN}$. For example, let $X_1 \sim$N$(0, 1)$ and

$$X_2 = \begin{cases} X_1 & \text{if } |X_1| \leq 1 \\ -X_1 & \text{if } |X_1| > 1. \end{cases}$$

X_1 and X_2 are Normal, but (X_1, X_2) is not Multivariate Normal. Also, if two Normal random variables are uncorrelated, it doesn't mean they're independent (unless they are jointly MVN!).

4.2 Affine Property

Any affine transformation of a MVN is MVN. If $X \sim \text{MVN}(\mu, \Sigma)$, then $AX + b \sim \text{MVN}(A\mu + b, A\Sigma A^\intercal)$, for any matrix of constants A and vector b of conformable sizes.

Some operations:

- **Constructing:** If $X_1, X_2, \ldots, X_n \sim \text{N}(0, 1)$ are independent, then $(X_1, X_2, \ldots, X_n) = X \sim \text{MVN}(0, I)$. We have $AX + \mu \sim \text{MVN}(\mu, \Sigma)$, where $\Sigma = AA^\intercal$. Using the spectral theorem, one can find A such that $AA^\intercal = \Sigma$ (i.e. you can construct any MVN from standard Normal rvs). Recall that, by the spectral theorem, $\Sigma = U\Lambda U^\intercal = U\Lambda^{1/2}\Lambda^{1/2}U^\intercal = AA^\intercal$, where U is an orthogonal matrix.
- **Sphering:** If $Y \sim \text{MVN}(\mu, \Sigma)$ and Σ is invertible, then $A^{-1}(Y - \mu) \sim \text{MVN}(0, I)$, where $\Sigma = AA^\intercal$. Alternatively, one can write this as follows: $\Sigma^{-1/2}(Y - \mu) \sim \text{MVN}(0, I)$.

4.3 Marginals and conditionals

Let $Y \sim \text{MVN}(\mu, \Sigma)$. Marginal distributions are Normal with the same parameters as in the MVN.

Let

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

then

$$Y_1 | Y_2 = a \sim \text{MVN}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

5 Chi-square Distribution

Definition: Let Z_j for $j \in \{1, 2, \ldots, p\}$ be iid $N(0, 1)$, and let $Z = (Z_1, Z_2, \ldots, Z_p)^\intercal$. Then, $X = Z^\intercal Z \sim \chi^2_p$.

Noncentral chi-square: Let Z_j for $j \in \{1, 2, \ldots, p\}$ be iid $N(\mu, 1)$ and let $Z = (Z_1, Z_2, \ldots, Z_p)^\intercal$. Then, $Z^\intercal Z \sim \chi^2_p(\sum_{j=1}^p \mu_j^2)$.

Quadratic forms and χ^2 distributions: Let $Y \sim \text{MVN}(0, I)$ and let P be a symmetric $n \times n$ matrix of rank k. Then $Y^\intercal P Y \sim \chi^2_k$ if and only if P is a rank k perpendicular projection operator (ppo). In the noncentral case, if $Y \sim \text{MVN}(\mu, I)$, then $Y^\intercal PY \sim \chi^2_k(\mu^\intercal P\mu/2)$ if and only of P is a ppo. Application: let $Y \sim \text{MVN}(\mu, I)$ with $\mu \in C(X)$ and $(I - P)$ be a rank k ppo onto $C(X)^\perp$. Then $Y^\intercal(I - P)Y \sim \chi^2_k$.

6 A Little Bit of Linear Model Theory

A lot of important results are not even mentioned here. I strongly recommend looking at Faraday's "Practical Regression and ANOVA using R"[3] if you think Christensen is too dry and need some extra intuition.

In matrix notation:

$$Y = X\beta + \varepsilon$$

where

- Y is the response variable, which is observable.
- X is the design matrix, also observable.
- β are the unknown coefficients, which we want to estimate.
- ε are the errors, also unobservable.

Assume X is full rank, then the Ordinary Least Squares (OLS) estimator is

$$\hat\beta_{OLS} = \arg\min_\beta \sum_{i=1}^n (Y_i - x_i^\intercal \beta)^2 = \arg\min_\beta (Y - X\beta)^\intercal(Y - X\beta) = \arg\min_\beta ||Y - X\beta||_2^2 = (X^\intercal X)^{-1}X^\intercal Y.$$

If X is not full rank $\hat\beta_{\text{OLS}} = (X^\intercal X)^- X^\intercal Y$.

Let $Y|\beta \sim \text{MVN}(X\beta, \sigma^2 I)$. Then $\hat\beta_{\text{MLE}} = \hat\beta_{\text{OLS}} = \hat\beta$. Now, let $\mu = X\beta$. Then $\hat\mu = X\hat\beta = P_X Y$ is the MLE of μ, where P_X is the ppo onto $C(X)$. On the other hand, the MLE of σ^2 is $\hat\sigma^2 = ||(I - P_X)Y||^2/n = e^\intercal e/n$, where $e = (I - P_X)Y$ are the residuals and $(I - P_X)$ is the ppo onto the orthogonal complement of $C(X)$.

6.1 Inference

Again, $Y \mid \beta \sim \text{MVN}(X\beta, \sigma^2 I)$.

The MLE of $\mu = X\beta$ is $\hat\mu = P_X Y$, and it is an unbiased estimate:

$$E(\hat\mu) = E(P_X Y) = P_X E(Y) = P_X \mu = \mu,$$

because $\mu \in C(X)$.

The residuals $e = (I - P_X)Y$ have mean zero:

$$E(e) = E[(I - P_X)Y] = (I - P_X)E(Y) = (I - P_X)\mu = 0,$$

[3]Link: `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`

because $\mu \perp C(X)^\perp$.

The residuals are correlated:

$$\text{Cov}(e) = \text{Cov}[(I - P_X)Y] = \sigma^2(I - P_X)(I - P_X) = \sigma^2(I - P_X).$$

Therefore $e \sim \text{MVN}(0, \sigma^2(I - P_X))$ by the affine property of the MVN.

Now we find the expectation of the MLE of σ^2, which is $\hat{\sigma}^2 = e^\intercal e/n = ||(I - P_X)Y||^2/n$:

$$E(||(I - P_X)Y||^2) = E[Y^\intercal(I - P_X)Y] = \sigma^2\text{tr}(I - P_X) + \mu^\intercal(I - P_X)\mu = \sigma^2(n - r(X)),$$

Therefore, the MLE is biased, but $e^\intercal e/(n - r(X))$ is unbiased.

Let $Y|\beta \sim \text{MVN}(X\beta, \sigma^2 I)$. The MLE for β is $\hat{\beta} = (X^\intercal X)^- X^\intercal Y$. We take the Moore-Penrose generalized inverse. Then, the distribution of $\hat{\beta}$ is MVN because it's just a linear combination of Y, which is MVN.

Taking expectations,

$$E(\hat{\beta}) = (X^\intercal X)^- X^\intercal X\beta,$$

and the variance is

$$\text{Cov}(\hat{\beta}) = \sigma^2(X^\intercal X)^- X^\intercal X(X^\intercal X)^- = \sigma^2(X^\intercal X)^-,$$

because if A^- is a Moore-Penrose generalized inverse, then it is symmetric and $A^- A A^- = A^-$ (reflexive).

Therefore

$$\hat{\beta} \sim \text{MVN}((X^\intercal X)^- X^\intercal X\beta, \sigma^2(X^\intercal X)^-)$$

If X full rank,

$$\hat{\beta} \sim \text{MVN}(\beta, \sigma^2(X^\intercal X)^{-1}).$$

The standard error for a particular component is $\text{se}(\hat{\beta}_i) = \sqrt{(X^\intercal X)^{-1}_{ii}}\sigma$.

6.2 Gauss-Markov Theorem

- Gauss-Markov says that OLS has minimum variance in the class of all linear unbiased estimators.
- Requires just first and second moments, no normality required.
- Under normality assumption, OLS = MLE has minimum variance out of all unbiased estimators, not just the linear ones.
- However, we can find estimators with smaller MSE if we allow some bias.

7 Bayesian Linear Regression

As usual, we have

$$Y = X\beta + \varepsilon,$$

and assume that ε is normal, yielding

$$Y \mid \beta, \phi \sim \text{MVN}(X\beta, \phi^{-1}I),$$

where ϕ is the precision (a scalar).

Since we're Bayesians now, we need to specify priors on β and ϕ.

A convenient (conjugate) choice of priors is

$$\begin{aligned} \beta \mid \phi &\sim \text{MVN}(b_0, \Phi_0^{-1}/\phi) \\ \phi &\sim \text{Gamma}(v_0/2, SS_0/2). \end{aligned}$$

Posteriors are

$$\begin{aligned} \beta \mid \phi, Y &\sim \text{MVN}(b_n, \Phi_n^{-1}/\phi) \\ \phi \mid Y &\sim \text{Ga}(v_n/2, SS_n/2), \end{aligned}$$

where

$$\begin{aligned} b_n &= (X^\intercal X + \Phi_0)^{-1}(X^\intercal X\hat{\beta}_{\text{OLS}} + \Phi_0 b_0) = (X^\intercal X + \Phi_0)^{-1}(X^\intercal y + \Phi_0 b_0) \\ \Phi_n &= (X^\intercal X + \Phi_0) \\ v_n &= v_0 + n \\ SS_n &= SS_0 + Y^\intercal Y + b_o^\intercal \Phi_0 b_0 - b_n^\intercal \Phi_n b_n \end{aligned}$$

7.1 Marginal of β and predictive distribution, all t

The following theorem is really useful for deriving marginal distributions.

Theorem: Let $\theta \mid \phi \sim \text{MVN}(m, \Sigma/\phi)$ and $\phi \sim \text{Gamma}(\nu/2, \nu\hat{\sigma}^2/2)$. Then $\theta \sim t_\nu(m, \hat{\sigma}^2\Sigma)$ with density

$$p(\theta) \propto \left[1 + \frac{1}{\nu}\frac{(\theta - m)^\intercal\Sigma^{-1}(\theta - m)}{\hat{\sigma}^2}\right]^{-(p+\nu)/2}.$$

Trick for predictives: Recall that $\beta \mid \phi, Y \sim \text{MVN}(b_n, \phi^{-1}\Phi_n^{-1})$ and $\phi \mid Y \sim \text{Ga}(v_n/2, SS_n/2)$. Then the new data $Y^* = X^*\beta + \varepsilon^* | \phi, Y \sim \text{MVN}(X^*b_n, \phi^{-1}(X^*\Phi_n^{-1}X^{*\intercal} + I))$. Using the previous theorem, $Y^* \mid Y \sim t_{v_n}(X^*b_n, \hat{\sigma}^2(X^*\Phi_n^{-1}X^{*\intercal} + I))$, where $\hat{\sigma}_n^2 = SS_n/v_n$.