# Matrix Algebra and Multivariate Probability

Ben Goodrich

April 25, 2022

# Vectors and Transposition

- A vector will be written in lowercase boldface, like $\mathbf{x}$
- A (column) vector of size $K$ is given by

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}$$

- The transpose operator $(\top)$ changes a column vector into a row vector and vice versa

$$\mathbf{x}^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \end{bmatrix}$$

- Unless otherwise indicated, vectors are column vectors

# Matrices and Transposition

- Matrices are collections of (row) vectors of the same size and are written in capital boldface letters like $\mathbf{X}$, where the first index pertains to the row

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \cdots & \vdots \\ x_{K1} & x_{K2} & \cdots & x_{KP} \end{bmatrix}$$

- $\mathbf{x}_p = \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{Kp} \end{bmatrix}$ is the $p$th column of $\mathbf{X}$
- $\mathbf{x}_k^\top = \begin{bmatrix} x_{k1} & x_{k2} & \cdots & x_{kP} \end{bmatrix}$ is the $k$th row of $\mathbf{X}$

# Things that Are Easy

- Multiplying or dividing a vector or matrix by a scalar, $a$; just multiply or divide each element of the vector or matrix by $a$ to create a new vector or matrix
- Adding or subtracting two vectors or matrices of the same size
- Subtracting or adding a scalar from or to a vector or matrix, which is defined as $\mathbf{y} = \mathbf{x} + a \equiv \mathbf{x} + a\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^\top$
- Elementwise multiplication or division of two vectors or matrices of the same size; just multiply or divide each element on the left by the corresponding element on the right to form a new vector or matrix of the same size. In Stan, these elementwise operators start with a period, e.g. `.*` and `./`

# Vector Multiplication

- If $\mathbf{x}$ and $\mathbf{y}$ are both vectors of size $K$

$$\mathbf{x}^\top \mathbf{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} \equiv \sum_{k=1}^{K} x_k y_k$$

- Called the dot product, inner product, vector product, etc.
- In R, do `t(x) %*% y` or equivalently `crossprod(x,y)`
- Common construction: $\mathbf{x}^\top \mathbf{x} = \sum_{k=1}^{K} x_k^2$
- The length (not size) of a $K$-vector $\mathbf{x}$ is $\sqrt{\mathbf{x}^\top \mathbf{x}}$, which is confusing because `length(x)` in R returns its size, $K$. Better to use `NROW`.

# Matrix Multiplication

- If $\mathbf{X}$ is $K \times M$ and $\mathbf{Y}$ is $M \times P$, then $\mathbf{Z} = \mathbf{XY}$ is a $K \times P$ matrix such that for all $k$ and $p$: $Z_{kp} = \mathbf{x}_k^\top \mathbf{y}_p = \sum_{m=1}^M x_{km} y_{mp}$

$$\mathbf{Z} = \mathbf{XY} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{y}_1 & \mathbf{x}_1^\top \mathbf{y}_2 & \cdots & \mathbf{x}_1^\top \mathbf{y}_P \\ \mathbf{x}_2^\top \mathbf{y}_1 & \mathbf{x}_2^\top \mathbf{y}_2 & \cdots & \mathbf{x}_2^\top \mathbf{y}_P \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_K^\top \mathbf{y}_1 & \mathbf{x}_K^\top \mathbf{y}_2 & \cdots & \mathbf{x}_K^\top \mathbf{y}_P \end{bmatrix}$$

- Matrix multiplication is not commutative but $(\mathbf{XY})^\top = \mathbf{Y}^\top \mathbf{X}^\top$, i.e. a column vector
- Common construction: If $\mathbf{X}$ is $N \times K$ and $\boldsymbol{\beta}$ is $K \times 1$

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{x}_1^\top \boldsymbol{\beta} \\ \mathbf{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ \mathbf{x}_N^\top \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^K x_{1k}\beta_k \\ \sum_{k=1}^K x_{2k}\beta_k \\ \vdots \\ \sum_{k=1}^K x_{Nk}\beta_k \end{bmatrix} = \begin{bmatrix} x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1K}\beta_K \\ x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2K}\beta_K \\ \vdots \\ x_{N1}\beta_1 + x_{N2}\beta_2 + \cdots + x_{NK}\beta_K \end{bmatrix} = \boldsymbol{\eta}$$

# Multivariate CDFs, PDFs, and Expectations

- If $\mathbf{x}$ is a $K$-vector of continuous random variables

$$F(\mathbf{x}) = \Pr\left(X_1 \leq x_1 \bigcap X_2 \leq x_2 \bigcap \cdots \bigcap X_K \leq x_K\right)$$

$$f(\mathbf{x}) = \frac{\partial^K F(\mathbf{x})}{\partial x_1 \partial x_2 \cdots \partial x_K} = f_1(x_1) \prod_{k=2}^{K} f_k(x_k | x_1, \ldots, x_{k-1})$$

$$F(\mathbf{x}) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(\mathbf{x})\, dx_1 dx_2 \cdots dx_K$$

$$\mathbb{E}g(\mathbf{x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x})\, dx_1 dx_2 \cdots dx_K$$

$$\boldsymbol{\mu}^\top = \mathbb{E}\mathbf{x}^\top = \begin{bmatrix} \mathbb{E}X_1 & \mathbb{E}X_2 & \cdots & \mathbb{E}X_K \end{bmatrix}$$

$$\boldsymbol{\Sigma}^\top = \boldsymbol{\Sigma} = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{12} & \sigma_2^2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & \sigma_{(K-1)K} \\ \sigma_{1K} & \cdots & \sigma_{(K-1)K} & \sigma_K^2 \end{bmatrix}$$

# Special Matrices

- A square matrix has the same number of rows as columns
- A square matrix $\mathbf{X}$ is symmetric iff $\mathbf{X} = \mathbf{X}^\top$
- Triangular matrices
  - Lower triangular square matrix has $X_{kp} = 0 \, \forall k < p$
  - Upper triangular square matrix has $X_{kp} = 0 \, \forall k > p$
- Diagonal matrix is a square matrix that is simultaneously lower and upper triangular and thus has $X_{kp} = 0 \, \forall k \neq p$
- The identity matrix, $\mathbf{I}$, is the diagonal matrix with only ones on its diagonal — i.e. $I_{kp} = \begin{cases} 1 & \text{if } k = p \\ 0 & \text{if } k \neq p \end{cases}$ — and is the matrix analogue of the scalar $1$
- If $\mathbf{X}$ is square, then $\mathbf{XI} = \mathbf{X} = \mathbf{IX}$
- An orthogonal matrix $\mathbf{Q}$ is such that $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I} = \mathbf{Q}\mathbf{Q}^\top$
- A zero vector / matrix is a vector / matrix with $0$ in each cell

# Matrix Inversion

- If $\mathbf{X}$ is $K \times K$, then the inverse of $\mathbf{X}$ — if it exists — is denoted $\mathbf{X}^{-1}$ and is the unique $K \times K$ matrix such that $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I} = \mathbf{X}^{-1}\mathbf{X}$
- Don't worry about how R finds the elements of $\mathbf{X}^{-1}$, just use `solve`
  - If $\mathbf{X}$ is diagonal, then $\left[\mathbf{X}^{-1}\right]_{kp} = \begin{cases} \frac{1}{X_{kp}} & \text{if } k = p \\ 0 & \text{if } k \neq p \end{cases}$
  - If $\mathbf{X}$ is only triangular, $\mathbf{X}^{-1}$ is also triangular and easy to find
- There is no vector or matrix "division" but multiplying $\mathbf{X}$ by $\mathbf{X}^{-1}$ is the matrix analogue of scalar multiplying $a$ by $\frac{1}{a}$. Also, $\left(\mathbf{X}a\right)^{-1} = \frac{1}{a}\mathbf{X}^{-1}$.
- An inverse of a product of square matrices equals the product of the inverses in reverse order: $\left(\mathbf{X}\mathbf{Y}\right)^{-1} = \mathbf{Y}^{-1}\mathbf{X}^{-1}$. Also, the inverse of a transpose of a square matrix is the transpose of the inverse: $\left(\mathbf{X}^\top\right)^{-1} = \left(\mathbf{X}^{-1}\right)^\top$

# Matrix Factorization

- How many ways can $24$ be factored over the positive integers?

  1. $1 \times 24$
  2. $2 \times 12$
  3. $3 \times 8$
  4. $4 \times 6$
  5. $2^3 \times 3$

- Matrices can be factored into the product of two (or more) special matrices, and the restrictions on the special matrices can make the factorization unique

- An example is the QR factorization $\underbrace{\mathbf{X}}_{N\times K} = \underbrace{\mathbf{Q}}_{N\times K}\underbrace{\mathbf{R}}_{K\times K}$, where $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ and $\mathbf{R}$ is upper triangular with non-negative diagonal elements

# What Does `QR = TRUE` Do?

- Let the vector of linear predictions in a GLM be $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$
- If we apply the QR decomposition to $\mathbf{X}$,

$$\boldsymbol{\eta} = \overbrace{\mathbf{Q}\mathbf{R}}^{\mathbf{X}}\boldsymbol{\beta} = \overbrace{\mathbf{Q}\frac{R_{KK}}{R_{KK}}\mathbf{R}}^{\mathbf{X}}\boldsymbol{\beta} = \overbrace{\mathbf{Q}^{*}\mathbf{R}^{*}}^{\mathbf{X}}\boldsymbol{\beta} = \mathbf{Q}^{*}\overbrace{\boldsymbol{\theta}}^{\mathbf{R}^{*}\boldsymbol{\beta}}$$

- When you specify `QR = TRUE` in `stan_glm` (or use `stan_lm` or `stan_polr`), rstanarm internally does a GLM using $\mathbf{Q}^{*} = \mathbf{Q}R_{KK}$ as the matrix of predictors instead of $\mathbf{X}$ to get the posterior distribution of $\boldsymbol{\theta}$ and then pre-multiplies each posterior draw of $\boldsymbol{\theta}$ by $\frac{1}{R_{KK}}\mathbf{R}^{-1}$ to get a posterior draw of $\boldsymbol{\beta}$
- Doing so makes it easier for NUTS to sample from the posterior distribution (of $\boldsymbol{\theta}$) efficiently because the columns of $\mathbf{Q}$ are orthogonal, whereas the columns of $\mathbf{X}$ are not

# Determinants

- A determinant is "like" a multivariate version of the absolute value operation and is denoted with the same symbol, $|\mathbf{X}|$
- Iff $|\mathbf{X}| \neq 0$, then $\mathbf{X}^{-1}$ exists and $\left|\mathbf{X}^{-1}\right| = \frac{1}{|\mathbf{X}|}$
- Statisticians mostly worry about determinants of triangular (inclusive of a diagonal) matrices and the determinant of a triangular matrix is the product of its diagonal entries, so $|\mathbf{R}| = \prod_{k=1}^{K} R_{kk}$
- Determinant of a product of square matrices is the product of their determinants
- $|\mathbf{X}| = |\mathbf{Q}|\,|\mathbf{R}|$. Since $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ and $|\mathbf{I}| = 1$, $|\mathbf{Q}| = \mp 1$. Thus, $|\mathbf{X}| = \mp|\mathbf{R}|$.

# Covariance and Correlation Matrices

- Recall that if $g\left(X_1, X_2\right) = \left(X_1 - \mu_1\right)\left(X_2 - \mu_2\right)$, then

$$\mathbb{E}g\left(X_1, X_2\right) = \int_{\Omega_{X_2}} \int_{\Omega_{X_1}} \left(x_1 - \mu_1\right)\left(x_2 - \mu_2\right) f\left(x_1, x_2\right) dx_1 dx_2 = \sigma_{12}$$

is the covariance between $X_1$ and $X_2$, while $\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \in [-1, 1]$ is their correlation, which is a measure of LINEAR dependence

- Let $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ be $K \times K$, such that $\Sigma_{ij} = \sigma_{ij}\, \forall i, j$ and $\Lambda_{ij} = \rho_{ij}\, \forall i \neq j$
  - Since $\sigma_{ij} = \sigma_{ji}\, \forall i, j$, $\mathbf{\Sigma} = \mathbf{\Sigma}^\top$ is symmetric
  - Since $\sigma_{ij} = \sigma_i^2$ iff $i = j$, $\Sigma_{ii} = \sigma_i^2 > 0$
  - Hence, $\mathbf{\Sigma}$ is called the variance-covariance matrix of $\mathbf{x}$
  - $\mathbf{\Sigma} = \mathbf{\Delta\Lambda\Delta}$ where $\Delta_{ij} = \begin{cases} \sigma_i & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$ is a diagonal matrix

# Cholesky Factors and Positive Definiteness

- Let $\mathbf{L}$ be lower triangular w/ positive diagonal entries such that $\mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}$, which is a Cholesky factor of $\boldsymbol{\Sigma}$ and can uniquely be defined via recursion:

$$L_{ij} = \begin{cases} \sqrt[+]{\Sigma_{jj} - \sum_{k=1}^{j-1} L_{kj}^2} & \text{if } i = j \\ \frac{1}{L_{jj}} \left( \Sigma_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk} \right) & \text{if } i > j \\ 0 & \text{if } i < j \end{cases}$$

- Positive definiteness of $\boldsymbol{\Sigma}$ implies $L_{jj}$ is real and positive for all $j$ and implies the existence of $\boldsymbol{\Sigma}^{-1} = \mathbf{L}^{-1} \left(\mathbf{L}^{-1}\right)^\top$, which is called a “precision matrix”. But not all symmetric matrices are positive definite, so $\Theta \subset \mathbb{R}^{K + \binom{K}{2}}$ in this case
- A Cholesky factor is “like” a square root of a positive definite matrix
- The `chol` function in R outputs $\mathbf{L}^\top$ instead

# Properties of the Multi(variate) Normal

- Univariate and bivariate normal are special cases where $K = 1$ and $K = 2$
- All margins of a multivariate normal distribution are multivariate normal
- All conditional distributions derived from a multivariate normal are multivariate normal
- A multivariate normal distribution stays in the multivariate normal family under shift, scale, and rotation transformations
- You are often going to have to estimate $\boldsymbol{\Sigma}$

# The LKJ Distribution for Correlation Matrices

- Let $\mathbf{\Delta}$ be a $K \times K$ diagonal matrix such that $\Delta_{kk}$ is the $k$-th standard deviation, $\sigma_k$, and let $\mathbf{\Lambda}$ be a correlation matrix
- Formulating a prior for $\mathbf{\Sigma} = \mathbf{\Delta\Lambda\Delta}$ is harder than putting a prior on $\mathbf{\Delta}$ & $\mathbf{\Lambda}$
- LKJ PDF is $f\left(\mathbf{\Lambda} | \eta\right) = \frac{1}{c(K,\eta)} \left|\mathbf{\Lambda}\right|^{\eta-1} = \left|\mathbf{L}\right|^{2(\eta-1)}$ where $\mathbf{\Lambda} = \mathbf{L}\mathbf{L}^\top$ with $\mathbf{L}$ a Cholesky factor and $c\left(K, \eta\right)$ is the normalizing constant that forces the PDF to integrate to $1$ over the space of correlation matrices
  - Iff $\eta = 1$, $f\left(\mathbf{\Lambda} | \eta\right) = \frac{1}{c(K,\eta)}$ is constant
  - If $\eta > 1$, the mode of $f\left(\mathbf{\Lambda} | \eta\right)$ is at $\mathbf{I}$ and as $\eta \uparrow \infty$, $\mathbf{\Lambda} \rightarrow \mathbf{I}$
  - If $0 < \eta < 1$, trough of $f\left(\mathbf{\Lambda} | \eta\right)$ is at $\mathbf{I}$, which is an odd thing to believe
- Can also derive the distribution of the Cholesky factor $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^\top$ is a correlation matrix with an LKJ$(\eta)$ distribution