Chapter 4
Examples of Mass Functions and Densities
Covariance and Correlation

Outline

Covariance

Recall that for $X_i, i = 1, \ldots, n$,

$$\text{Var}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \text{Cov}(X_i, X_j)$$

We can write this more compactly by introducing the vector $\mathbf{b} = (b_1, \ldots, b_n)^T$ and the $n \times n$ covariance matrix $\text{Var}(X)$ with i, j entry $\text{Cov}(X_i, X_j)$, then

$$\text{Var}\left(\sum_{i=1}^{n} b_i X_i\right) = \mathbf{b}^{\mathbf{T}}\text{Var}(X)\mathbf{b}.$$

Linear Transformations

For $X = (X_1, \ldots, X_n)$, let $Y = AX$ be a linear transformation X. Then,

$$\begin{aligned} \mathrm{Cov}(Y)_{ij} &= \mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}((AX)_i, (AX)_j) \\ &= \mathrm{Cov}\left(\sum_{k=1}^{n} A_{ik} X_k, \sum_{\ell=1}^{n} A_{\ell j} X_\ell\right) \\ &= \sum_{\ell=1}^{n} \sum_{k=1}^{n} A_{ik} \mathrm{Cov}\left(X_k, X_\ell\right) A^T_{j\ell} = (A \ \mathrm{Cov}(X) A^T)_{ij} \end{aligned}$$

In particular, if $X = (X_1, \ldots, X_n)$ are independent with common variance σ^2, then $\mathrm{Var}(X) = \sigma^2 I$ and $\mathrm{Var}(Y) = \sigma^2 A A^T$

Multivariate Normal Distributions

Returning to a multivariate normal variable with density

$$f_Y(\mathbf{y}) = \frac{1}{|\det(A)|(2\pi)^{n/2}} \exp\left(-\frac{\mathbf{y}^T(AA^T)^{-1}\mathbf{y}}{2}\right)$$

For $Z = (Z_1, \ldots, Z_n)$ consisting of independent $N(0,1)$. $\mathrm{Cov}(Z) = I$, the identity matrix, $\Sigma = \mathrm{Cov}(Y) = AA^T$. Note that

$$\det\Sigma = \det(AA^T) = \det A \cdot \det A^T = (\det A)^2.$$

Thus, $|\det A| = \sqrt{\det\Sigma}$, and

$$f_Y(\mathbf{y}) = \frac{1}{\sqrt{\det\Sigma(2\pi)^n}} \exp\left(-\frac{\mathbf{y}^T\Sigma^{-1}\mathbf{y}}{2}\right)$$

Covariance Matrices

Covariance matrices have two important properties.

- Σ is symmetric

$$\Sigma_{ij} = \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) = \Sigma_{ji}$$

- Σ is non-negative definite

$$\mathbf{b}^T \Sigma \mathbf{b} \geq 0, \quad \mathbf{b} \in \mathbb{R}^n.$$

This leads to a variety of properties for Σ

- The eigenvalues are non-negative. Let $\mathbf{u}$ be a unit eigenvector for Σ with eigenvalue λ

$$0 \leq \mathbf{u}^T \Sigma \mathbf{u} = \lambda \mathbf{u}^T \mathbf{u} = \lambda$$

Covariance Matrices

- Eigenvectors corresponding to distinct eigenvalues are orthogonal . Let $\mathbf{u_i}, i = 1, 2$ be a unit eigenvectors with distinct eigenvalues $\lambda_i, i = 1, 2$

$$\lambda_2 \mathbf{u}_1^T \mathbf{u}_2 = \mathbf{u_2}^T \Sigma^T \mathbf{u_1} = \mathbf{u}_2^T \Sigma \mathbf{u}_1 = \lambda_1 \mathbf{u}_1^T \mathbf{u}_2$$

Thus, $(\lambda_2 - \lambda_1)\mathbf{u}_1^T \mathbf{u}_2$ and $\mathbf{u_i}, i = 1, 2$ are orthogonal

Consequently, we can find a orthonormal basis U of eigenvectors.

$$U = (\mathbf{u}_1 | \cdots | \mathbf{u}_n)$$

and a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1 \ldots, \lambda_v)$ so that

$$\Sigma U = U \Lambda.$$

Covariance Matrices

Because the rows of U are orthogonal, $UU^T = U^U = I$ and

$$\Sigma = U\Lambda U^T.$$

It is common practice to order the eigenvalues (and corresponding eigenvectors) from largest to smallest and call the eigen vectors principle components

Again, let $Z = (Z_1, \ldots, Z_n)^T$ be independent $N(0,1)$ and let $\Lambda^{1/2}$ be the diagonal matrix whose entries are the square roots of the entries in Λ. Set $Y = U\Lambda^{1/2}Z$. Then,

$$\mathrm{Cov}(Y) = U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = \Sigma$$

Thus, given an non-negative symmetric positive definite matrix, we can construct a set of random variables $Y_1, \ldots, Y_n$ whose covariance matrix is Σ

Covariance Matrices

Example In two dimensions, orthogonal matrices are rotations by an angle θ

$$U = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = U \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

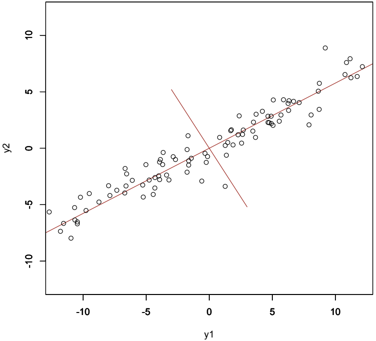


Figure: $\lambda_1 = 5, \lambda_2 = 1, \theta = \pi/6$

Covariance Matrices

- Non-negative symmetric matrices are closed under addition and multiplication by a positive constant. So for such matrices Σ_1, Σ_2 and non-negative constants c_1, c_2

$$c_1\Sigma_1 + c_2\Sigma_2$$

is a non-negative symmetric matrix.

- If A is an $n \times p$ matrix. Then AA^T is a symmetric $n \times n$ matrix. Also, for $\mathbf{b} \in \mathbb{R}^n$, then $A^T\mathbf{b} \in \mathbb{R}^p$

$$\mathbf{b}^T(AA^T)\mathbf{b} = (A^T\mathbf{b})^T(A^T\mathbf{b}) = ||A^T\mathbf{b}|| \geq 0$$

showing that AA^T is non-negative definite.

Principal Component Analysis

- Many modern statistical questions can start out with a hundreds to thousands number of correlated variables.
- We have used the structure of the covariance matrix Σ to determine
  - A set of non-negative eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

  - An orthogonal matrix U whose rows are the corresponding eigenvectors

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

- the original random variables Y_i can be written as a linear combination of $\sigma_i Z_i$, $\sigma_i^2 = \lambda_i$

Principal Component Analysis

Principal Component Analysis is a dimension-reduction tool that can be used advantageously in such situations. In this situation, we keep enough of the ransom variables $\sigma_i Z_i$ to capture a large fraction of the variance in the Y_i.

In many applications $n \gg p$. Thus, AA^T is $n \times n$ and the computational demand can be very high. However, $A^T A$ is much smaller, namely, $p \times p$ and thus is much more manageable for our needs. To show how this happens, let $v_1, \ldots, v_p$ be eigenvectors for $A^T A$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_p$. Then

$$A^T A v_i = \lambda_i v_i, \quad AA^T A v_i = \lambda_i A v_i, \quad AA^T u_i = \lambda_i u_i$$

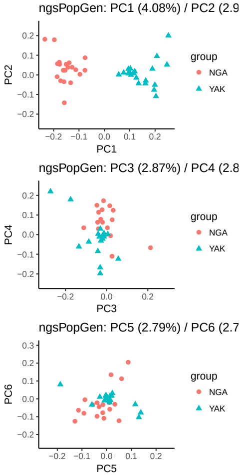
where $u_i = A v_i$ are eigenvectors for AA^T.

Principal Component Analysis

Example. We consider two indigenous Siberian populations, the Nganasan (nomadic hunters, NGA, $n = 21$) from the Taymyr Peninsula in the Arctic Ocean, and the Yakut (herders, YAK, $n = 21$) of North-Central Siberia.

The data at each DNA site is either 0, 1, or 2 depending on the number of alleles that match a reference allele.

The estimated covariance matrix $\hat{\Sigma}$ is determined from from an $n \times p$ data matrix where p is tens of thousands.



Multinomial Distribution

We first recall the multinomial theorem

$$(a_1 + \cdots + a_k)^n = \sum_{|\mathbf{x}|=n} \binom{n}{\mathbf{x}} a_1^{x_1} \cdots a_k^{x_k}$$

where $\mathbf{x} = (x_1, \cdots, x_k)$, $|\mathbf{x}| = x_1 + \cdots + x_k$, and

$$\binom{n}{\mathbf{x}} = \frac{n!}{x_1! \cdots x_k!}$$

.A vector-valued random variable $X = (X_1, \ldots, X_k)$ is said to have a multinomial distribution with parameters n and $\mathbf{p} = (p_1, \ldots, p_k), \sum_{i=1}^{k} p_i = 1$ ($X \sim Multi(n, \mathbf{p})$) if its mass function

$$f_X(\mathbf{x}|n, \mathbf{p}) = \binom{n}{\mathbf{x}} p_1^{x_1} \cdots p_k^{x_k}$$

Multinomial Distribution

Exercise. X_i $Bin(n, p_i)$ We check the case $i = 1$,

$$
\begin{aligned}
f_{X_1}(x_1) &= \sum_{x_2+\cdots x_k=n-x_1} \binom{n}{\mathbf{x}} p_1^{x_1} \cdots p_k^{x_k} \\
&= \binom{n}{x_1} p_1^{x_1} \sum_{x_2+\cdots x_k=n-x_1} \binom{n-x_1}{x_1, \ldots, x_k} p_1^{x_2} \cdots p_k^{x_k} \\
&= \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1}
\end{aligned}
$$

Thus,

$$EX_i = np_i, \quad \text{and} \quad \mathrm{Var}(X_i) = np_(1 - ip_i).$$

Multinomial Distribution

The probability generating function

$$
\begin{aligned}
E[z_1^{X_1}\cdots z_k^{X_k}] &= \sum_{|\mathbf{x}|=n} \binom{n}{\mathbf{x}} (z_1^{x_1}\cdots z_k^{x_k}) p_1^{x_1}\cdots p_k^{x_k} \\
&= \sum_{|\mathbf{x}|=n} \binom{n}{\mathbf{x}} (p_1 z_1)^{x_1}\cdots (p_k z_k)^{x_k} = (p_1 z_1 + \cdots + p_k z_k)^n
\end{aligned}
$$

Thus, we can write

$$X = Y_1 + \cdots + Y_n$$

as n independent $Multi(1, \mathbf{p})$ random variables. Then $X_i = Y_{1i} + \cdots + Y_{n,i}$. Then, for $i \neq j$, $Y_{\ell i} Y_{\ell j} = 0$ (Both cannot differ from zero in one with only one trial. Thus,

$$\mathrm{Cov}(Y_{\ell i}, Y_{\ell j}) = EY_{\ell i} Y_{\ell j} - EY_{\ell i} \cdot EY_{\ell j} = -p_i p_j.$$

Multinomial Distribution

Exercise. $\text{Cov}(X_i, X_j) = -np_ip_j$

For $\ell \neq m$, $Y_{\ell i}$ and Y_{mj} are independent,

$$\text{Cov}(X_i, X_j) = \sum_{\ell=1}^{n}\sum_{m=1}^{n} \text{Cov}(Y_{\ell i}, Y_{mj}) = \sum_{\ell=1}^{n} \text{Cov}(Y_{\ell i}, Y_{\ell j}) = -np_ip_j$$

The correlation

$$\begin{aligned} \rho_{X_i,X_j} &= \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)}\sqrt{\text{Var}(X_j)}} = \frac{-np_ip_j}{\sqrt{np_i(1-p_i)}\sqrt{np_j(1-p_j)}} \\ &= -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}} \end{aligned}$$

which depends on the odds but not on the number of trials. If each of the $p_i = 1/k$ then the correlation is $-1/(k-1)$