Random Vectors and Covariance Structure

Statistics · Module 8 — Multivariate Linear Models

Node 8.1

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.1 — Random Vectors and Covariance Structure.

This node establishes the algebraic and geometric framework for multivariate statistics. The mean vector and covariance matrix are the fundamental objects governing linear operations on random vectors. This node depends on the probability foundations (Nodes 1.3, 1.5, 1.6) and enables the multivariate normal (Node 8.2), linear models (Node 8.3), and PCA (Node 8.7).

Formal Definition

Let $X = (X_1, \ldots, X_p)^\top$ be a p-dimensional random vector on $(\Omega, \mathcal{F}, \mathbb{P})$.

Mean vector. $\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p])^\top \in \mathbb{R}^p$.

Covariance matrix. $\Sigma = \mathrm{Cov}(X) = \mathbb{E}[(X - \mu)(X - \mu)^\top] \in \mathbb{R}^{p \times p}$, with $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

Correlation matrix. $R = D^{-1/2} \Sigma D^{-1/2}$ where $D = \mathrm{diag}(\Sigma_{11}, \ldots, \Sigma_{pp})$, with $R_{ij} = \mathrm{Cor}(X_i, X_j) \in [-1, 1]$ and $R_{ii} = 1$.

Intuition

The mean vector locates the center of mass of the distribution. The covariance matrix captures both the spread (diagonal: variances) and the linear dependence structure (off-diagonal: covariances). Geometrically, the covariance matrix defines an ellipsoid: the set $\{x : (x - \mu)^\top \Sigma^{-1} (x - \mu) \leq c\}$ is an ellipsoid whose axes are aligned with the eigenvectors of Σ and whose half-lengths are proportional to the square roots of the eigenvalues. The correlation matrix is the standardized version: it strips out the individual scales and retains only the dependence structure.

Structural Role

The covariance matrix is the central object in multivariate analysis. It encodes the variance and dependence structure, governs linear transformations via the congruence $A\Sigma A^\top$, determines the geometry of confidence ellipsoids, regression projections, and PCA directions. In the Gaussian case, μ and Σ fully determine the distribution. The spectral decomposition of Σ provides the principal component directions (Node 8.7) and the Mahalanobis metric for multivariate inference (Node 8.2).

Mathematical Development

Positive semidefiniteness. For any $a \in \mathbb{R}^p$,

$$a^\top \Sigma a = \mathrm{Var}(a^\top X) \geq 0.$$

Thus $\Sigma \succeq 0$. Equality holds for some $a \neq 0$ if and only if $a^\top X$ is degenerate (a.s. constant), meaning the distribution of X is supported on a proper affine subspace. Σ is positive definite ($\Sigma \succ 0$) if and only if no nontrivial linear combination is degenerate.

Linear transformation rule. For $A \in \mathbb{R}^{q \times p}$ and $b \in \mathbb{R}^q$:

$$\mathbb{E}[AX + b] = A\mu + b, \qquad \mathrm{Cov}(AX + b) = A\Sigma A^\top.$$

Proof. $\mathrm{Cov}(AX + b) = \mathbb{E}[A(X - \mu)(X - \mu)^\top A^\top] = A\,\mathbb{E}[(X - \mu)(X - \mu)^\top]\,A^\top = A\Sigma A^\top$. $\square$

This is the congruence transformation: Σ transforms as a bilinear form under linear maps.

Cross-covariance. For random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, the cross-covariance is $\Sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\top] \in \mathbb{R}^{p \times q}$. For the joint vector $(X^\top, Y^\top)^\top$, the covariance is

$$\mathrm{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix},$$

with $\Sigma_{YX} = \Sigma_{XY}^\top$.

Spectral decomposition. Since Σ is symmetric positive semidefinite, the spectral theorem gives

$$\Sigma = Q\Lambda Q^\top = \sum_{j=1}^{p} \lambda_j q_j q_j^\top,$$

where $Q = [q_1, \ldots, q_p]$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. The eigenvalues are the variances along the principal directions q_j: $\mathrm{Var}(q_j^\top X) = \lambda_j$.

Mahalanobis distance. For $\Sigma \succ 0$, the Mahalanobis distance between x and μ is

$$d_M(x, \mu) = \sqrt{(x - \mu)^\top \Sigma^{-1} (x - \mu)}.$$

This is the Euclidean distance after whitening: if $Z = \Sigma^{-1/2}(X - \mu)$, then $\mathrm{Cov}(Z) = I_p$ and $d_M(x, \mu) = \|z\|$. The Mahalanobis distance accounts for the scale and correlation structure, making it invariant under affine transformations of the data.

Sample covariance matrix. For observations $X_1, \ldots, X_n$, the sample mean and covariance are

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top.$$

$\bar{X}$ is unbiased for μ and S is unbiased for Σ. Under finite fourth moments, the multivariate CLT gives $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ and $S \xrightarrow{p} \Sigma$.

Partial correlation. The partial correlation of X_i and X_j given the remaining variables is

$$\rho_{ij \cdot \mathrm{rest}} = -\frac{(\Sigma^{-1})_{ij}}{\sqrt{(\Sigma^{-1})_{ii}(\Sigma^{-1})_{jj}}}.$$

Partial correlations encode the conditional linear dependence structure and are read from the precision matrix Σ^{-1}, not from Σ itself.

Whitening transformation. The matrix $W = \Sigma^{-1/2}$ transforms X into $Z = W(X - \mu)$ with $\mathrm{Cov}(Z) = I_p$. The whitening transformation decorrelates the components and normalizes their variances to 1. It is not unique: any matrix W satisfying $W\Sigma W^\top = I$ produces a whitened vector. The symmetric whitening $W = \Sigma^{-1/2}$ (via the matrix square root) is the unique whitening that is also symmetric positive definite.

Cauchy-Schwarz for covariance. For any two random variables X_i, X_j in the vector,

$$|\mathrm{Cov}(X_i, X_j)|^2 \leq \mathrm{Var}(X_i)\mathrm{Var}(X_j),$$

which is equivalent to $|R_{ij}| \leq 1$. This follows from the Cauchy-Schwarz inequality in $L^2(\Omega, \mathbb{P})$. Equality holds if and only if $X_j = aX_i + b$ a.s. for constants a, b.

Block matrix inversion. For the partitioned covariance, the precision matrix is

$$\Sigma^{-1} = \begin{pmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix}.$$

The diagonal blocks are the inverses of the Schur complements, connecting covariance structure to conditional distributions (Node 8.2).

Best linear predictor. The best linear predictor of X_1 given X_2 (minimizing $\mathbb{E}[\|X_1 - a - BX_2\|^2]$) is

$$\hat{X}_1 = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),$$

with prediction error covariance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. This holds for any distribution with finite second moments, not just the Gaussian. Under Gaussianity, the best linear predictor coincides with the conditional expectation.

Worked Examples

Example 1. Let $X = (X_1, X_2)^\top$ with $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$. The eigenvalues solve $\det(\Sigma - \lambda I) = (4-\lambda)(3-\lambda) - 4 = \lambda^2 - 7\lambda + 8 = 0$, giving $\lambda_1 = (7 + \sqrt{17})/2 \approx 5.56$ and $\lambda_2 = (7 - \sqrt{17})/2 \approx 1.44$. The correlation is $R_{12} = 2/\sqrt{4 \cdot 3} = 1/\sqrt{3} \approx 0.577$. The Mahalanobis distance from the origin to $(1, 1)^\top$ (assuming $\mu = 0$) is $\sqrt{(1,1)\Sigma^{-1}(1,1)^\top} = \sqrt{(1,1)(3,-2;-2,4)(1,1)^\top/8} = \sqrt{3/8} \approx 0.612$, using $\Sigma^{-1} = (1/8)\begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}$.

Example 2. Let $Y = AX + b$ with $A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and $\Sigma_X = I_2$. Then $\Sigma_Y = AA^\top = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2I_2$. The transformation rotates and scales: the two components of Y are uncorrelated with equal variance 2, even though the transformation is not orthogonal. If instead $\Sigma_X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, then $\Sigma_Y = \begin{pmatrix} 2(1+\rho) & 0 \\ 0 & 2(1-\rho) \end{pmatrix}$ — the transformation A diagonalizes the covariance when the off-diagonal structure aligns with its rows.

Example 3 (Whitening). Let $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$. The Cholesky decomposition gives $\Sigma = LL^\top$ with $L = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{2} \end{pmatrix}$. The whitening matrix is $W = L^{-1} = \begin{pmatrix} 1/2 & 0 \\ -1/(2\sqrt{2}) & 1/\sqrt{2} \end{pmatrix}$, and $Z = W(X - \mu)$ has $\mathrm{Cov}(Z) = I_2$. This is the Cholesky whitening, which produces a lower-triangular transformation. The symmetric whitening $\Sigma^{-1/2}$ produces a different but equally valid decorrelation.

Example 4 (Best linear predictor). With $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$ and $\mu = (1, 2)^\top$, the best linear predictor of X_1 given $X_2 = x_2$ is $\hat{X}_1 = 1 + (2/3)(x_2 - 2)$. The prediction error variance is $4 - 4/3 = 8/3$. The coefficient $\Sigma_{12}/\Sigma_{22} = 2/3$ is the regression coefficient of X_1 on X_2.

Cross-Links

AI. The empirical covariance matrix is central to PCA, linear discriminant analysis, Gaussian process kernels, and whitening in preprocessing pipelines. In neural networks, the Fisher information matrix (Node 7.3) is a covariance matrix of score vectors, used in natural gradient methods.

Economics. Portfolio theory (Markowitz) uses the covariance matrix of asset returns to compute the efficient frontier; the minimum-variance portfolio is $w \propto \Sigma^{-1}\mathbf{1}$, a direct function of the covariance structure.

Linear Algebra. The covariance matrix is a symmetric positive semidefinite linear operator on $\mathbb{R}^p$. Its spectral decomposition is the spectral theorem for self-adjoint operators; the congruence transformation $A\Sigma A^\top$ is the action of A on the quadratic form defined by Σ.

Quantum Mechanics. The covariance matrix of quadrature operators $(\hat{q}_1, \hat{p}_1, \ldots, \hat{q}_n, \hat{p}_n)$ characterizes Gaussian quantum states; the Heisenberg uncertainty principle imposes $\Sigma + i\Omega/2 \succeq 0$ where Ω is the symplectic form, a constraint without classical analogue.

Structural Compression

Invariant. $\mathrm{Cov}(AX) = A\Sigma A^\top$: covariance transforms by congruence under linear maps.

Mechanism. The covariance matrix is the second-moment operator; congruence transformation is the second-order analogue of the first-order pushforward $\mathbb{E}[AX] = A\mu$. Positive semidefiniteness is preserved because variance is nonnegative.

Cross-System Echo. The covariance matrix is a Riemannian metric on the space of linear functionals of X; its eigendecomposition provides the natural coordinate system (principal components). In information geometry, the Fisher information matrix is a covariance matrix (of the score) and defines the Riemannian metric on the statistical manifold.

Exercises

1. Prove that every symmetric positive semidefinite matrix Σ is a valid covariance matrix by constructing a random vector with that covariance.
2. Let $X \in \mathbb{R}^p$ and $Y = AX$ for invertible A. Show that the Mahalanobis distance is invariant: $d_M^Y(Ax, A\mu) = d_M^X(x, \mu)$.
3. Prove that the sample covariance matrix S is unbiased for Σ, and explain why the divisor is $n - 1$ rather than n.
4. For $X = (X_1, X_2, X_3)^\top$ with $\Sigma^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{pmatrix}$, compute the partial correlation $\rho_{13 \cdot 2}$ and interpret the zero in $(\Sigma^{-1})_{13}$.
5. Show that $\mathrm{tr}(\Sigma) = \sum_j \mathrm{Var}(X_j) = \sum_j \lambda_j$, connecting the total variance to the trace and the eigenvalue sum.
6. Prove that $|\mathrm{Cor}(X_i, X_j)| = 1$ if and only if $X_j = aX_i + b$ a.s. for some constants $a \neq 0, b$.

7. Argue, structurally, why the covariance matrix is the natural inner product on the space of linear functionals of X, and why its spectral decomposition provides the coordinate system that diagonalizes this inner product.

Statistics · Module 8 — Multivariate Linear Models · Node 8.1
Concepts: variance-and-covariance, random-variables, vector-space, eigenvalue-decomposition
Prerequisites: 1.3, 1.5, 1.6
Enables: 8.2, 8.3, 8.7

hbar.university — own.your.cognition