Module 8 — Multivariate Linear Models

Statistics

hbar.university

Random Vectors and Covariance Structure

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 $\{x : (x-\mu)^\top \Sigma^{-1}(x-\mu) \le c\}$ 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.

Multivariate Normal Distribution

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.2 — Multivariate Normal Distribution.

The multivariate normal (MVN) is the canonical multivariate distribution, completely determined by its first two moments. It is the reference distribution for all finite-sample inference in linear models. This node depends on the covariance structure (Node 8.1) and distribution theory (Node 1.4), and enables linear models (Node 8.3), inference (Node 8.5), and PCA (Node 8.7).

Formal Definition

A random vector $X \in \mathbb{R}^p$ has the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ if every linear combination is univariate normal:

$$a^\top X \sim \mathcal{N}(a^\top \mu,\ a^\top \Sigma a) \quad \forall\, a \in \mathbb{R}^p.$$

When $\Sigma \succ 0$, the density is

$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\!\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

The moment generating function is $M(t) = \exp(t^\top \mu + \tfrac{1}{2} t^\top \Sigma t)$.

Intuition

The MVN generalizes the bell curve to multiple dimensions. The density is constant on ellipsoids centered at μ, with axes aligned along the eigenvectors of Σ and half-lengths proportional to the square roots of the eigenvalues. The key structural property is that the entire dependence structure is captured by Σ: for jointly normal vectors, uncorrelatedness implies independence, and all conditional and marginal distributions remain normal. This makes the MVN the only distribution where second-moment analysis is sufficient for complete probabilistic characterization.

Structural Role

The MVN is the reference distribution for multivariate statistics. Its closure under affine transformations, the normality of marginals and conditionals, and the characterization of quadratic forms underlie all finite-sample distributional results in the normal linear model — t-tests, F-tests, confidence ellipsoids. The Wishart distribution (the sampling distribution of the sample covariance) is derived from the MVN. In asymptotic theory, the MVN appears as the limit distribution in the multivariate CLT.

Mathematical Development

Affine closure. If $X \sim \mathcal{N}(\mu, \Sigma)$ and $Y = AX + b$ with $A \in \mathbb{R}^{q \times p}$, then

$$Y \sim \mathcal{N}(A\mu + b,\ A\Sigma A^\top).$$

Proof. For any $c \in \mathbb{R}^q$, $c^\top Y = c^\top AX + c^\top b = (A^\top c)^\top X + c^\top b$, which is univariate normal since X is MVN. The mean is $c^\top(A\mu + b)$ and variance is $c^\top A\Sigma A^\top c$. $\square$

Marginal distributions. Partition $X = (X_1^\top, X_2^\top)^\top$ with $X_1 \in \mathbb{R}^{p_1}$, $X_2 \in \mathbb{R}^{p_2}$, and correspondingly

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Theorem. $X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$.

Proof. $X_1 = (I_{p_1}, 0)X$, so by affine closure, $X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$. $\square$

Conditional distributions. The conditional distribution of $X_1 \mid X_2 = x_2$ is

$$X_1 \mid X_2 = x_2 \;\sim\; \mathcal{N}\big(\mu_{1|2},\, \Sigma_{1|2}\big)\,,$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Proof. Define $Z = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$. Then $\mathrm{Cov}(Z, X_2) = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$. Since (Z, X_2) is jointly normal (as an affine function of X) and uncorrelated, $Z \perp\!\!\!\perp X_2$. Therefore $X_1 \mid X_2 = x_2$ has the same distribution as $Z + \Sigma_{12}\Sigma_{22}^{-1}x_2$, giving the stated mean and variance. $\square$

The conditional variance $\Sigma_{1|2}$ is the Schur complement of Σ_{22} in Σ. It does not depend on the conditioning value x_2 — this is the homoskedasticity of the MVN conditional.

Independence and uncorrelatedness. For jointly normal X_1, X_2:

$$\Sigma_{12} = 0 \iff X_1 \perp\!\!\!\perp X_2.$$

The forward direction follows because uncorrelated jointly normal vectors have independent MGFs: $M_{X_1,X_2}(t_1, t_2) = M_{X_1}(t_1)M_{X_2}(t_2)$.

Quadratic forms. If $X \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$, then

$$(X - \mu)^\top \Sigma^{-1}(X - \mu) \sim \chi^2_p.$$

Proof. Let $Z = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I_p)$. Then $(X - \mu)^\top\Sigma^{-1}(X - \mu) = Z^\top Z = \sum_{j=1}^p Z_j^2 \sim \chi^2_p$. $\square$

More generally, if $X \sim \mathcal{N}(0, I_n)$ and A is symmetric idempotent with rank r, then $X^\top AX \sim \chi^2_r$. This is the fundamental result connecting projection matrices to chi-squared distributions (used in Node 8.5).

Cochran's theorem. Let $X \sim \mathcal{N}(0, \sigma^2 I_n)$ and let $A_1, \ldots, A_m$ be symmetric matrices with $\sum_j A_j = I_n$ and $\mathrm{rank}(A_j) = r_j$ with $\sum_j r_j = n$. If each A_j is idempotent (equivalently, each $A_j \succeq 0$ and the rank condition holds), then $X^\top A_1 X/\sigma^2, \ldots, X^\top A_m X/\sigma^2$ are independent $\chi^2_{r_j}$ random variables.

Proof sketch. Since A_j is symmetric idempotent with rank r_j, it has eigendecomposition $A_j = U_j U_j^\top$ where $U_j \in \mathbb{R}^{n \times r_j}$ has orthonormal columns. The condition $\sum A_j = I_n$ implies the columns of $U_1, \ldots, U_m$ form an orthonormal basis of $\mathbb{R}^n$. Define $Z_j = U_j^\top X/\sigma \sim \mathcal{N}(0, I_{r_j})$. Then $X^\top A_j X/\sigma^2 = Z_j^\top Z_j \sim \chi^2_{r_j}$, and the Z_j are independent because $U_j^\top U_k = 0$ for $j \neq k$. $\square$

Characterization via MGF. A random vector X is MVN if and only if its MGF has the form $M(t) = \exp(t^\top\mu + t^\top\Sigma t/2)$. This characterization is useful for proving closure properties: if X has this MGF and $Y = AX + b$, then $M_Y(t) = \exp(t^\top(A\mu + b) + t^\top A\Sigma A^\top t/2)$, confirming $Y \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$.

Singular MVN. When Σ is positive semidefinite but not positive definite (rank $r < p$), X is supported on an r-dimensional affine subspace and has no density with respect to Lebesgue measure on $\mathbb{R}^p$. The distribution is still well-defined: $X = \mu + BZ$ where $Z \sim \mathcal{N}(0, I_r)$ and $B \in \mathbb{R}^{p \times r}$ satisfies $BB^\top = \Sigma$. All linear combination properties, marginal/conditional formulas, and the Portmanteau characterization remain valid.

Stein's lemma. If $X \sim \mathcal{N}(\mu, \Sigma)$ and $g : \mathbb{R}^p \to \mathbb{R}$ is differentiable with $\mathbb{E}[|\nabla g(X)|] < \infty$, then

$$\mathrm{Cov}(X_j, g(X)) = \sum_{k=1}^{p} \Sigma_{jk} \mathbb{E}\left[\frac{\partial g}{\partial x_k}(X)\right].$$

This identity characterizes the normal distribution and is the basis for Stein's method for proving normal approximation. In the scalar case: $\mathrm{Cov}(X, g(X)) = \sigma^2 \mathbb{E}[g'(X)]$.

Multivariate CLT. If $X_1, \ldots, X_n$ are i.i.d. in $\mathbb{R}^p$ with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Cov}(X_i) = \Sigma$, then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. The limit is MVN regardless of the distribution of X_i — the Gaussian character is inherited from the CLT, not assumed. This is the reason the MVN appears throughout asymptotic statistics.

Wishart distribution. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$. The matrix $W = \sum_{i=1}^{n} X_i X_i^\top$ follows the Wishart distribution $W_p(n, \Sigma)$. The density (for $n \geq p$) is

$$f(W) = \frac{|W|^{(n-p-1)/2}}{2^{np/2} |\Sigma|^{n/2} \Gamma_p(n/2)} \exp\!\left(-\tfrac{1}{2}\mathrm{tr}(\Sigma^{-1} W)\right),$$

where Γ_p is the multivariate gamma function. The Wishart is the multivariate generalization of the chi-squared distribution: when $p = 1$, $W_1(n, \sigma^2) = \sigma^2 \chi^2_n$.

Worked Examples

Example 1. Let $X \sim \mathcal{N}\left(\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}4 & 2\\2 & 3\end{pmatrix}\right)$. The conditional distribution of $X_1 \mid X_2 = 5$ is $\mathcal{N}(1 + (2/3)(5 - 2),\ 4 - 4/3) = \mathcal{N}(3,\ 8/3)$. The conditional mean is a linear function of the conditioning value: a unit increase in X_2 shifts the conditional mean of X_1 by $\Sigma_{12}/\Sigma_{22} = 2/3$. The conditional variance $8/3$ is less than the marginal variance 4, reflecting the information gained from observing X_2.

Example 2. Let $Z \sim \mathcal{N}(0, I_3)$. The quadratic form $Q = Z_1^2 + Z_2^2 + Z_3^2 \sim \chi^2_3$. Let H be the projection onto $\mathrm{span}(1, 1, 1)^\top/\sqrt{3}$, so $H = \mathbf{1}\mathbf{1}^\top/3$. Then $Z^\top H Z/1 = (\bar{Z})^2 \cdot 3 \sim \chi^2_1$ and $Z^\top (I - H) Z \sim \chi^2_2$, with the two quadratic forms independent by Cochran's theorem. This decomposition is the simplest instance of the ANOVA F-test structure.

Example 3 (Conditional distribution computation). Let $X \sim \mathcal{N}\left(0, \begin{pmatrix}1 & 0.5 & 0.3\\0.5 & 1 & 0.4\\0.3 & 0.4 & 1\end{pmatrix}\right)$. Partition as $X_1 = X_1$ (scalar) and $X_2 = (X_2, X_3)^\top$. Then $\Sigma_{12} = (0.5, 0.3)$, $\Sigma_{22} = \begin{pmatrix}1 & 0.4\\0.4 & 1\end{pmatrix}$, $\Sigma_{22}^{-1} = \frac{1}{0.84}\begin{pmatrix}1 & -0.4\\-0.4 & 1\end{pmatrix}$. The conditional mean of $X_1 \mid X_2 = x_2, X_3 = x_3$ is $\Sigma_{12}\Sigma_{22}^{-1}(x_2, x_3)^\top = \frac{1}{0.84}(0.5 - 0.12)x_2 + (0.3 - 0.2)x_3) = \frac{1}{0.84}(0.38 x_2 + 0.1 x_3)$. The conditional variance is $1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - \frac{1}{0.84}(0.5 \cdot 0.38 + 0.3 \cdot 0.1) \approx 0.738$.

Cross-Links

AI. Gaussian processes use the MVN as the finite-dimensional distribution of function values; the conditional distribution formula gives the posterior predictive distribution. Variational autoencoders use the MVN as the latent space prior and the reparametrization trick exploits affine closure.

Economics. The MVN is the standard assumption in portfolio theory (Markowitz) and CAPM; the mean-variance efficient frontier is derived from the MVN assumption on asset returns.

Linear Algebra. The MVN density is an exponential of a negative definite quadratic form; level sets are ellipsoids determined by the spectral decomposition of Σ^{-1}. The Schur complement in the conditional variance formula is the same Schur complement that appears in block matrix inversion.

Quantum Mechanics. Gaussian quantum states are characterized by the first two moments of the quadrature operators, with the covariance matrix subject to the Heisenberg uncertainty constraint $\Sigma + i\Omega/2 \succeq 0$. The conditional distribution formula has a direct quantum analogue in Gaussian state conditioning (homodyne detection).

Structural Compression

Invariant. The MVN is determined by its first two moments; all higher-order cumulants vanish. Marginals, conditionals, and affine transformations of MVN vectors are MVN.

Mechanism. The density is an exponential of a negative definite quadratic form in x; level sets are ellipsoids aligned with the eigenvectors of Σ^{-1}. The MGF $\exp(t^\top \mu + t^\top \Sigma t/2)$ encodes all moments and makes closure under affine maps immediate.

Cross-System Echo. The MVN is the maximum-entropy distribution with fixed mean and covariance (information-theoretic characterization). In information geometry, the Gaussian family forms a statistical manifold with Fisher metric determined by the precision matrix. The Wishart distribution generalizes the chi-squared to matrix-valued sufficient statistics.

Exercises

1. Let $X \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$. Derive the density by applying the change-of-variables formula to $Z = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I_p)$.
2. Prove that if $X \sim \mathcal{N}(\mu, \Sigma)$ and $\Sigma_{12} = 0$, then $X_1 \perp\!\!\!\perp X_2$, using the MGF factorization.
3. Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Compute the conditional distribution $X_1 \mid X_2 = x_2$ and show that the regression of X_1 on X_2 has slope ρ and residual variance $1 - \rho^2$.
4. Prove Cochran's theorem for the case $m = 2$: if $A_1 + A_2 = I_n$ with A_1, A_2 symmetric and idempotent, then $X^\top A_1 X$ and $X^\top A_2 X$ are independent chi-squared.
5. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$. Show that $\bar{X}$ and $S = (n-1)^{-1} \sum (X_i - \bar{X})(X_i - \bar{X})^\top$ are independent, and that $(n-1)S \sim W_p(n-1, \Sigma)$.
6. Compute the entropy of $\mathcal{N}(\mu, \Sigma)$ and show it depends only on $|\Sigma|$, confirming the maximum-entropy characterization.
7. Argue, structurally, why the equivalence of uncorrelatedness and independence under joint normality is a special property of the Gaussian and fails for general distributions, connecting this to the role of higher-order cumulants.

Linear Models: Geometry and OLS

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.3 — Linear Models: Geometry and OLS.

The linear model is the most fundamental regression framework in statistics. This node develops OLS as an orthogonal projection, establishes the geometric interpretation, and derives the properties of the hat matrix and residuals. It depends on the covariance structure (Node 8.1) and the MVN (Node 8.2), and enables Gauss-Markov (Node 8.4), inference (Node 8.5), and GLMs (Node 8.6).

Formal Definition

The linear model is

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

where $Y \in \mathbb{R}^n$ is the response, $X \in \mathbb{R}^{n \times p}$ is the design matrix with $\text{rank}(X) = p \leq n$, $\beta \in \mathbb{R}^p$ is the coefficient vector, and $\varepsilon \in \mathbb{R}^n$ satisfies $\mathbb{E}[\varepsilon] = 0$, $\text{Cov}(\varepsilon) = \sigma^2 I_n$.

The **OLS estimator** is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Intuition

The column space $\mathcal{C}(X) = \{X\beta : \beta \in \mathbb{R}^p\}$ is a p-dimensional subspace of $\mathbb{R}^n$. The model says the mean response $\mathbb{E}[Y] = X\beta$ lies in this subspace. OLS finds the point in $\mathcal{C}(X)$ closest to the observed Y in Euclidean distance — this is orthogonal projection. The residuals are the component of Y perpendicular to $\mathcal{C}(X)$, and the fitted values are the projection. This geometric view makes the algebraic properties of OLS transparent.

Structural Role

The linear model separates signal $X\beta \in \mathcal{C}(X)$ from noise $\varepsilon \perp \mathcal{C}(X)$ via orthogonal projection. The hat matrix H is the central object: it encodes both the estimator and the residuals, and its rank determines degrees of freedom. Optimality of OLS under second-moment assumptions is established in Node 8.4. Under normality, OLS is also the MLE, and exact finite-sample inference follows (Node 8.5).

Mathematical Development

Derivation of the normal equations. The OLS criterion is $\min_\beta \|Y - X\beta\|^2$. Expanding:

$$\|Y - X\beta\|^2 = Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta.$$

Taking the gradient and setting to zero:

$$\nabla_\beta \|Y - X\beta\|^2 = -2X^\top Y + 2X^\top X\beta = 0.$$

This gives the **normal equations** $X^\top X\hat{\beta} = X^\top Y$. Since $\text{rank}(X) = p$, $X^\top X$ is invertible and $\hat{\beta} = (X^\top X)^{-1}X^\top Y$.

Geometric interpretation. The normal equations $X^\top(Y - X\hat{\beta}) = 0$ state that the residual $Y - X\hat{\beta}$ is orthogonal to every column of X, hence orthogonal to $\mathcal{C}(X)$. This is the defining property of orthogonal projection.

Hat matrix. The fitted values are

$$\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y = HY,$$

where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix.

Properties of H: - Symmetric: $H^\top = H$. - Idempotent: $H^2 = H$ (projecting twice is the same as projecting once). - $\text{rank}(H) = \text{tr}(H) = p$. - $HX = X$ (columns of X are in the range of H). - Eigenvalues of H are 0 or 1, with exactly p ones.

Residuals. The residuals are

$$\hat{\varepsilon} = Y - \hat{Y} = (I_n - H)Y.$$

The matrix $I_n - H$ is also symmetric and idempotent, with rank $n - p$. Key property: $(I - H)X = 0$, so the residuals are orthogonal to every column of X.

Orthogonal decomposition. The observation vector decomposes as

$$Y = HY + (I - H)Y = \hat{Y} + \hat{\varepsilon},$$

with $\hat{Y} \perp \hat{\varepsilon}$. The Pythagorean theorem gives

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2.$$

After centering (subtracting the mean), this becomes the ANOVA decomposition: TSS = ESS + RSS.

Moments of the OLS estimator.

$$\mathbb{E}[\hat{\beta}] = (X^\top X)^{-1}X^\top \mathbb{E}[Y] = (X^\top X)^{-1}X^\top X\beta = \beta.$$

$$\text{Cov}(\hat{\beta}) = (X^\top X)^{-1}X^\top \text{Cov}(Y)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}.$$

OLS is unbiased for any error distribution with mean zero and any σ^2.

Residual sum of squares.

$$\text{RSS} = \|\hat{\varepsilon}\|^2 = Y^\top(I - H)Y.$$

Since $\mathbb{E}[Y^\top(I - H)Y] = \text{tr}((I - H)\sigma^2 I) + \beta^\top X^\top(I - H)X\beta = \sigma^2(n - p)$ (using $(I - H)X = 0$), we get

$$\mathbb{E}[\text{RSS}] = \sigma^2(n - p),$$

so $\hat{\sigma}^2 = \text{RSS}/(n - p)$ is unbiased for σ^2.

Leverage scores. The diagonal elements $h_{ii} = H_{ii}$ are the leverage scores. They satisfy $0 \le h_{ii} \le 1$ and $\sum_i h_{ii} = p$. The leverage h_{ii} measures the influence of the i-th observation on its own fitted value: $\hat{Y}_i = h_{ii}Y_i + \sum_{j \ne i} h_{ij}Y_j$. High-leverage points have h_{ii} close to 1 and dominate their own fit.

Residual properties. Under homoskedasticity, $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2(I - H)$, so $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$. The residuals are heteroskedastic even when the errors are not. Standardized residuals $r_i = \hat{\varepsilon}_i / (\hat{\sigma}\sqrt{1 - h_{ii}})$ correct for this.

Connection to the Moore-Penrose pseudoinverse. When X has full column rank, $(X^\top X)^{-1}X^\top = X^+$ is the Moore-Penrose pseudoinverse. The OLS solution is $\hat{\beta} = X^+ Y$, the minimum-norm least-squares solution.

Worked Examples

Example 1 (Simple linear regression). Let $X = (\mathbf{1}, x) \in \mathbb{R}^{n \times 2}$ with $x = (x_1, \ldots, x_n)^\top$. Then

$$X^\top X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad X^\top Y = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}.$$

Solving gives $\hat{\beta}_1 = \sum (x_i - \bar{x})(Y_i - \bar{Y}) / \sum (x_i - \bar{x})^2$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$. The hat matrix has $h_{ii} = 1/n + (x_i - \bar{x})^2 / \sum (x_j - \bar{x})^2$, showing that leverage increases with distance from $\bar{x}$.

Example 2 (One-way ANOVA). For k groups with n_j observations each, X is a block-diagonal indicator matrix. The hat matrix projects onto group means: $\hat{Y}_{ij} = \bar{Y}_{j\cdot}$. The residual is $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{j\cdot}$. RSS $= \sum_j \sum_i (Y_{ij} - \bar{Y}_{j\cdot})^2$ with $n - k$ degrees of freedom, and ESS $= \sum_j n_j (\bar{Y}_{j\cdot} - \bar{Y})^2$ with $k - 1$ degrees of freedom.

Cross-Links

AI. Linear regression is empirical risk minimization under squared loss. The hat matrix appears in leave-one-out cross-validation via the shortcut $\mathrm{CV} = n^{-1} \sum (\hat{\varepsilon}_i / (1 - h_{ii}))^2$, avoiding n refits. Ridge regression replaces H with $X(X^\top X + \lambda I)^{-1} X^\top$, shrinking the projection.

Economics. The linear model is the workhorse of empirical economics. OLS with heteroskedasticity-robust standard errors (Huber-White) replaces $\sigma^2 (X^\top X)^{-1}$ with the sandwich covariance (Node 7.5). Instrumental variables estimation projects onto a different column space.

Linear Algebra. OLS is the orthogonal projection theorem in $\mathbb{R}^n$ with the standard inner product. The normal equations are the first-order optimality conditions for a convex quadratic. The hat matrix is a rank-p orthogonal projector, decomposing $\mathbb{R}^n = \mathcal{C}(X) \oplus \mathcal{C}(X)^\perp$.

Quantum Mechanics. In quantum state tomography, linear inversion estimators reconstruct the density matrix by projecting measurement outcomes onto the space of valid states, the same projection structure as OLS but on the cone of positive semidefinite matrices.

Structural Compression

Invariant. $\hat{Y} = HY$ is the orthogonal projection of Y onto $\mathcal{C}(X)$; the OLS estimator is the unique solution to the normal equations $X^\top X \hat{\beta} = X^\top Y$.

Mechanism. Minimizing $\|Y - X\beta\|^2$ over β is equivalent to finding the closest point in $\mathcal{C}(X)$ to Y; the projection theorem yields the hat matrix as the unique orthogonal projector. The orthogonal decomposition $Y = \hat{Y} + \hat{\varepsilon}$ separates signal from noise.

Cross-System Echo. OLS is the Moore-Penrose pseudoinverse solution. In numerical linear algebra, the QR decomposition provides a numerically stable computation: $X = QR$ gives $\hat{\beta} = R^{-1}Q^\top Y$ and $H = QQ^\top$. In AI, the leverage scores h_{ii} determine the effective number of parameters and the generalization error via the optimism.

Exercises

1. Derive the OLS estimator by completing the square in $\|Y - X\beta\|^2$ rather than by differentiation.
2. Show that H is the unique symmetric idempotent matrix with $\mathcal{C}(H) = \mathcal{C}(X)$.
3. Prove that $\sum_{i=1}^{n} h_{ii} = p$ using $\mathrm{tr}(H) = \mathrm{tr}(X(X^\top X)^{-1}X^\top)$ and the cyclic property of trace.
4. For simple linear regression with equally spaced $x_i = i$, compute the leverage scores and identify the points with maximum leverage.
5. Show that adding a predictor to the model (increasing p by 1) always decreases RSS, and relate this to the geometry of the projection.
6. Prove the leave-one-out formula: the i-th cross-validation residual is $\hat{\varepsilon}_i^{(-i)} = \hat{\varepsilon}_i/(1 - h_{ii})$ using the Sherman-Morrison-Woodbury formula.
7. Argue, structurally, why the hat matrix is the fundamental object in the linear model — more fundamental than $\hat{\beta}$ — by showing that all inferential quantities (fitted values, residuals, RSS, leverage, degrees of freedom) are determined by H alone.

Gauss-Markov Theorem

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.4 — Gauss-Markov Theorem.

The Gauss-Markov theorem establishes OLS as optimal within the class of linear unbiased estimators under second-moment assumptions. This is a finite-sample result requiring no distributional assumptions beyond the first two moments. This node depends on the linear model geometry (Node 8.3) and estimation theory (Node 3.1), and enables inference in linear models (Node 8.5).

Formal Definition

Let $Y = X\beta + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$, and $\mathrm{rank}(X) = p$.

An estimator $\tilde{\beta} = CY$ is **linear** (in Y) and **unbiased** if $CX = I_p$.

Gauss-Markov Theorem. The OLS estimator $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ is the Best Linear Unbiased Estimator (BLUE): for any linear unbiased estimator $\tilde{\beta} = CY$,

$$\mathrm{Cov}(\tilde{\beta}) - \mathrm{Cov}(\hat{\beta}) \succeq 0 \quad \text{(positive semidefinite)}.$$

Equivalently, for every $a \in \mathbb{R}^p$, $\mathrm{Var}(a^\top \tilde{\beta}) \geq \mathrm{Var}(a^\top \hat{\beta})$.

Intuition

Among all estimators that are linear functions of Y and unbiased for β, OLS has the smallest variance in every direction. The reason is geometric: OLS uses the minimum-norm linear map from Y to $\hat{\beta}$, namely the pseudoinverse $X^+ = (X^\top X)^{-1} X^\top$. Any other linear unbiased estimator adds a component D that is orthogonal to the column space of X (to preserve unbiasedness) but contributes additional variance $\sigma^2 DD^\top \succeq 0$.

Structural Role

Gauss-Markov establishes the optimality of OLS within a restricted class (linear, unbiased) under minimal assumptions (first two moments only). It is a finite-sample result — no asymptotics needed. It does not claim OLS is the best among all estimators: biased estimators (ridge, LASSO) or nonlinear estimators can have lower MSE. Under normality, OLS is additionally the MLE and achieves the Cramer-Rao bound, giving a stronger optimality. When the homoskedasticity assumption fails, GLS replaces OLS as the BLUE.

Mathematical Development

Proof of the Gauss-Markov theorem.

Let $\tilde{\beta} = CY$ be any linear unbiased estimator. Write $C = (X^\top X)^{-1} X^\top + D$ where $D = C - (X^\top X)^{-1} X^\top$.

Unbiasedness constraint: $CX = I_p$ requires $(X^\top X)^{-1} X^\top X + DX = I_p + DX = I_p$, so $DX = 0$.

Covariance computation:

$$\mathrm{Cov}(\tilde{\beta}) = \sigma^2 CC^\top = \sigma^2[(X^\top X)^{-1}X^\top + D][(X^\top X)^{-1}X^\top + D]^\top.$$

Expanding, using $DX = 0$ (so $D \cdot X(X^\top X)^{-1} = 0$, meaning the cross terms vanish):

$$\mathrm{Cov}(\tilde{\beta}) = \sigma^2(X^\top X)^{-1} + \sigma^2 DD^\top = \mathrm{Cov}(\hat{\beta}) + \sigma^2 DD^\top.$$

Since $DD^\top \succeq 0$, we have $\mathrm{Cov}(\tilde{\beta}) \succeq \mathrm{Cov}(\hat{\beta})$. $\square$

Why the cross terms vanish: $(X^\top X)^{-1}X^\top D^\top = [(X^\top X)^{-1}(DX)^\top]^\top$. Since $DX = 0$, this is zero. Geometrically, the rows of D lie in $\mathcal{C}(X)^\perp$ while the rows of $(X^\top X)^{-1}X^\top$ lie in $\mathcal{C}(X)$, making them orthogonal.

Scalar estimand version. For any linear functional $\psi = a^\top\beta$, the BLUE is $a^\top\hat{\beta}$ with variance $\sigma^2 a^\top(X^\top X)^{-1}a$. Any other linear unbiased estimator of ψ has variance at least this large.

Assumptions revisited. The Gauss-Markov theorem requires: 1. Linearity: $\mathbb{E}[Y] = X\beta$. 2. Homoskedasticity: $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$. 3. Full column rank of X.

It does not require: normality of ε, independence of observations (only uncorrelatedness and equal variance), or any knowledge of σ^2.

When assumptions fail: Generalized Least Squares.

If $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega$ with known $\Omega \succ 0$ and $\Omega \neq I$, OLS is no longer BLUE. The BLUE is the GLS estimator:

$$\hat{\beta}_{\mathrm{GLS}} = (X^\top\Omega^{-1}X)^{-1}X^\top\Omega^{-1}Y.$$

Derivation. Transform the model by $\Omega^{-1/2}$: let $Y^* = \Omega^{-1/2}Y$, $X^* = \Omega^{-1/2}X$, $\varepsilon^* = \Omega^{-1/2}\varepsilon$. Then $\mathrm{Cov}(\varepsilon^*) = \sigma^2 I_n$, so OLS applied to the transformed model gives

$$\hat{\beta}_{\mathrm{GLS}} = (X^{*\top}X^*)^{-1}X^{*\top}Y^* = (X^\top\Omega^{-1}X)^{-1}X^\top\Omega^{-1}Y.$$

By Gauss-Markov applied to the transformed model, this is BLUE.

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}}) = \sigma^2(X^\top\Omega^{-1}X)^{-1}.$$

Feasible GLS. When Ω is unknown, replace it by a consistent estimator $\hat{\Omega}$. The feasible GLS estimator $\hat{\beta}_{\mathrm{FGLS}} = (X^\top\hat{\Omega}^{-1}X)^{-1}X^\top\hat{\Omega}^{-1}Y$ is asymptotically equivalent to GLS but not exactly BLUE in finite samples.

Beyond BLUE: bias-variance tradeoff. Ridge regression $\hat{\beta}_\lambda = (X^\top X + \lambda I)^{-1}X^\top Y$ is biased but has MSE

$$\mathrm{MSE}(\hat{\beta}_\lambda) = \sigma^2\mathrm{tr}[(X^\top X + \lambda I)^{-2}X^\top X] + \lambda^2\beta^\top(X^\top X + \lambda I)^{-2}\beta.$$

For any $\beta \neq 0$ and $\lambda > 0$ small enough, $\mathrm{MSE}(\hat{\beta}_\lambda) < \mathrm{MSE}(\hat{\beta}) = \sigma^2\mathrm{tr}(X^\top X)^{-1}$. Gauss-Markov does not contradict this because ridge regression is biased; BLUE is optimal only within the unbiased class.

Under normality: stronger optimality. If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, then OLS is the MLE and achieves the Cramer-Rao lower bound among all unbiased estimators (not just linear ones). Under normality, BLUE = UMVUE for $a^\top\beta$.

Weighted least squares as GLS. When $\mathrm{Var}(\varepsilon_i) = \sigma^2/w_i$ with known weights $w_i > 0$, the covariance is $\mathrm{Cov}(\varepsilon) = \sigma^2 W^{-1}$ where $W = \mathrm{diag}(w_1, \ldots, w_n)$. GLS gives

$$\hat{\beta}_{\mathrm{WLS}} = (X^\top WX)^{-1}X^\top WY,$$

which minimizes $\sum_i w_i(Y_i - x_i^\top \beta)^2$. Observations with smaller variance (larger w_i) receive more weight — they are more informative.

Efficiency loss of OLS under heteroskedasticity. When $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega \neq \sigma^2 I$, OLS remains unbiased ($\mathbb{E}[\hat{\beta}] = \beta$) but its covariance is

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) = \sigma^2 (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1},$$

which differs from the naive $\sigma^2(X^\top X)^{-1}$. The difference $\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) - \mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}}) \succeq 0$, quantifying the efficiency loss from ignoring the error structure.

Gauss-Markov for estimable functions. When X does not have full column rank, β is not identifiable, but linear functions $a^\top\beta$ that lie in $\mathcal{C}(X^\top)$ are still estimable. The Gauss-Markov theorem extends: for any estimable $a^\top\beta$, the OLS estimate (via the pseudoinverse) is BLUE.

Worked Examples

Example 1. Consider the model $Y_i = \mu + \varepsilon_i$ with $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$. Here $X = \mathbf{1}_n$, so $\hat{\mu} = \bar{Y}$ with $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$. Any other linear unbiased estimator is $\tilde{\mu} = \sum c_i Y_i$ with $\sum c_i = 1$, and $\mathrm{Var}(\tilde{\mu}) = \sigma^2 \sum c_i^2 \geq \sigma^2/n$ by Cauchy-Schwarz with equality iff $c_i = 1/n$ for all i.

Example 2 (Heteroskedastic errors). Let $Y_i = \beta x_i + \varepsilon_i$ with $\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2$ (no intercept). Here $\Omega = \mathrm{diag}(x_1^2, \ldots, x_n^2)$. The OLS estimator is $\hat{\beta}_{\mathrm{OLS}} = \sum x_i Y_i / \sum x_i^2$, but the GLS estimator is $\hat{\beta}_{\mathrm{GLS}} = \sum Y_i/x_i / \sum 1 = n^{-1}\sum Y_i/x_i$, which weights each observation inversely proportional to its variance. GLS has smaller variance than OLS in this setting.

Example 3 (Ridge vs. OLS). Consider $X^\top X = I_p$, $n = p$, $\sigma^2 = 1$, $\|\beta\|^2 = B^2$. The MSE of OLS is $\mathrm{tr}(I_p) = p$. The MSE of ridge with λ is $p/(1+\lambda)^2 + \lambda^2 B^2/(1+\lambda)^2$. Setting the derivative to zero: optimal $\lambda^* = p/B^2$, giving MSE $= pB^2/(p+B^2) < p$. Ridge strictly dominates OLS for every $B > 0$ when $p \geq 3$ (Stein's phenomenon). The Gauss-Markov theorem is not violated because ridge is biased.

Cross-Links

AI. Gauss-Markov justifies OLS in the unbiased regime; regularized methods (ridge, LASSO) trade bias for variance reduction, stepping outside the BLUE framework. The bias-variance tradeoff is the central concept in statistical learning theory.

Economics. GLS is the basis for correcting heteroskedasticity (weighted least squares) and serial correlation (Cochrane-Orcutt, Prais-Winsten) in time series econometrics. The Hausman test compares OLS and GLS to detect misspecification.

Linear Algebra. The Gauss-Markov proof uses the decomposition of the linear map $C = X^+ + D$ into the pseudoinverse plus a perturbation orthogonal to the column space. The Loewner order $\succeq$ on covariance matrices is the partial order on symmetric matrices induced by the cone of positive semidefinite matrices.

Quantum Mechanics. The BLUE concept has an analogue in quantum estimation: the locally unbiased estimator with minimum variance under the symmetric logarithmic derivative inner product achieves the quantum Cramer-Rao bound, with the same structure of projecting onto a minimum-variance subspace.

Structural Compression

Invariant. Among all linear unbiased estimators of β, OLS has the smallest covariance matrix in the Loewner order.

Mechanism. Any competing linear unbiased estimator adds a matrix D satisfying $DX = 0$; the cross terms vanish by orthogonality and the covariance increment $\sigma^2 DD^\top \succeq 0$ is nonnegative definite. The constraint $DX = 0$ forces the perturbation into the orthogonal complement of the column space.

Cross-System Echo. Gauss-Markov is the second-moment analogue of the Cramer-Rao bound: both establish lower bounds on estimator variance within a restricted class. Under normality, the two coincide. The BLUE property is a finite-sample optimality result; the Cramer-Rao bound can be either finite-sample (under regularity) or asymptotic.

Exercises

1. Verify the Gauss-Markov theorem for the intercept-only model $Y_i = \mu + \varepsilon_i$ by directly showing $\bar{Y}$ has minimum variance among all $\sum c_i Y_i$ with $\sum c_i = 1$.
2. Derive the GLS estimator by transforming the model with $\Omega^{-1/2}$ and applying OLS to the transformed model.
3. Show that if $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega$ with $\Omega \neq I$, OLS is still unbiased but no longer BLUE, and compute its (incorrect) variance and its true variance.
4. Prove that ridge regression has lower MSE than OLS for λ sufficiently small, by showing $d(\mathrm{MSE}(\hat{\beta}_\lambda))/d\lambda|_{\lambda=0} < 0$ when $\beta \neq 0$.
5. In a model with two predictors and $X^\top X = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}$, show that Gauss-Markov implies $\hat{\beta}_1 = \bar{X}_1$ and $\hat{\beta}_2 = \bar{X}_2$ are individually optimal, and explain why orthogonality of predictors simplifies the result.
6. Construct an example where a biased estimator has strictly lower MSE than the BLUE for every value of β in a bounded parameter space.
7. Argue, structurally, why Gauss-Markov is both powerful and limited: powerful because it requires only second-moment assumptions, limited because the class of linear unbiased estimators excludes the most effective modern methods (regularization, nonlinear methods).

Inference in Linear Models

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.5 — Inference in Linear Models.

This node develops the exact finite-sample distributional theory for inference in the normal linear model: t-tests, F-tests, confidence intervals, and prediction intervals. These results depend on the geometry of OLS (Node 8.3), Gauss-Markov (Node 8.4), the MVN (Node 8.2), and the general testing and estimation frameworks (Nodes 5.1, 4.1). This node enables the extension to GLMs (Node 8.6).

Formal Definition

Let $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\text{rank}(X) = p$. Under normality:

$$\hat{\beta} \sim \mathcal{N}(\beta,\ \sigma^2(X^\top X)^{-1}), \qquad \frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}, \qquad \hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2.$$

Intuition

Under Gaussian errors, the OLS estimator is exactly normal (not just approximately). The residual sum of squares follows a chi-squared distribution because it is a quadratic form in normal variables via the idempotent matrix $I - H$. The independence of $\hat{\beta}$ and $\hat{\sigma}^2$ — which follows from the orthogonality of H and $I - H$ — allows us to form exact pivotal quantities. The ratio of a normal to a chi-squared produces a t-distribution; the ratio of two chi-squareds produces an F-distribution. These are exact for any sample size, not asymptotic approximations.

Structural Role

Under normality, the linear model admits exact finite-sample inference. The t-test for single coefficients and the F-test for general linear restrictions are both instances of likelihood ratio testing (Node 5.4) specialized to the Gaussian linear model, where the chi-squared approximation happens to be exact. Cochran's theorem (Node 8.2) provides the distributional foundation, connecting the rank of projection matrices to chi-squared degrees of freedom.

Mathematical Development

Distribution of $\hat{\beta}$. Since $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ is a linear function of $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, it is exactly normal: $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$.

Distribution of RSS. Write $\hat{\varepsilon} = (I - H)Y = (I - H)\varepsilon$ (since $(I - H)X\beta = 0$). Then $\hat{\varepsilon}/\sigma \sim \mathcal{N}(0, I - H)$ and

$$\frac{\text{RSS}}{\sigma^2} = \frac{\varepsilon^\top (I - H)\varepsilon}{\sigma^2}.$$

Since $I - H$ is symmetric idempotent with rank $n - p$, this is a chi-squared with $n - p$ degrees of freedom: $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$.

Independence of $\hat{\beta}$ and RSS. $\hat{\beta}$ depends on Y through HY, and RSS depends on Y through $(I-H)Y$. Since $H(I-H)=0$ and Y is normal, $HY \perp\!\!\!\perp (I-H)Y$. Therefore $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$.

Cochran's theorem application. The decomposition $\varepsilon = H\varepsilon + (I-H)\varepsilon$ corresponds to projection onto $\mathcal{C}(X)$ and $\mathcal{C}(X)^\perp$. By Cochran's theorem, $\varepsilon^\top H\varepsilon/\sigma^2 \sim \chi^2_p$ and $\varepsilon^\top(I-H)\varepsilon/\sigma^2 \sim \chi^2_{n-p}$, independently. Under the alternative $\mathbb{E}[Y] = X\beta \neq 0$, the first quadratic form has a noncentral chi-squared distribution, while RSS remains $\sigma^2\chi^2_{n-p}$.

t-test for a single coefficient. To test $H_0 : \beta_j = \beta_j^0$:

$$T_j = \frac{\hat{\beta}_j - \beta_j^0}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}} \sim t_{n-p} \quad \text{under } H_0.$$

Derivation. $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2(X^\top X)^{-1}_{jj})$, so $(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{(X^\top X)^{-1}_{jj}}) \sim \mathcal{N}(0,1)$. Dividing by $\sqrt{\hat{\sigma}^2/\sigma^2}$, where $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ independently, gives a t_{n-p} distribution.

Confidence interval for β_j:

$$\hat{\beta}_j \pm t_{n-p,\,\alpha/2} \cdot \hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}.$$

F-test for general linear hypotheses. For $H_0 : R\beta = r$ with $R \in \mathbb{R}^{q\times p}$ of rank q:

$$F = \frac{(R\hat{\beta} - r)^\top [R(X^\top X)^{-1}R^\top]^{-1}(R\hat{\beta} - r)/q}{\hat{\sigma}^2} \sim F_{q,\,n-p} \quad \text{under } H_0.$$

Derivation. Under H_0, $R\hat{\beta} - r \sim \mathcal{N}(0, \sigma^2 R(X^\top X)^{-1}R^\top)$. The numerator quadratic form divided by σ^2 is χ^2_q. Dividing by the independent $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$ gives $F_{q,n-p}$.

Nested model comparison. To compare the full model $Y = X\beta + \varepsilon$ with a reduced model $Y = X_0\beta_0 + \varepsilon$ where $\mathcal{C}(X_0) \subseteq \mathcal{C}(X)$:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p - p_0)}{\mathrm{RSS}/(n-p)} \sim F_{p-p_0,\,n-p} \quad \text{under the reduced model.}$$

This is the ANOVA F-test.

Confidence ellipsoid for β:

$$\left\{ b \in \mathbb{R}^p : (b-\hat{\beta})^\top (X^\top X)(b-\hat{\beta}) \leq p\hat{\sigma}^2 F_{p,\,n-p,\,\alpha} \right\}.$$

This is the inversion of the F-test for $H_0 : \beta = b$.

Prediction interval. For a new observation $Y_{\text{new}} = x_0^\top\beta + \varepsilon_{\text{new}}$:

$$\hat{Y}_{\text{new}} = x_0^\top\hat{\beta}, \quad \mathrm{Var}(Y_{\text{new}} - \hat{Y}_{\text{new}}) = \sigma^2(1 + x_0^\top(X^\top X)^{-1}x_0).$$

The $(1-\alpha)$ prediction interval is

$$x_0^\top\hat{\beta} \pm t_{n-p,\,\alpha/2} \cdot \hat{\sigma}\sqrt{1 + x_0^\top(X^\top X)^{-1}x_0}.$$

The extra σ^2 accounts for the irreducible noise in the new observation.

Coefficient of determination. $R^2 = 1 - \text{RSS}/\text{TSS}$ where $\text{TSS} = \|Y - \bar{Y}\mathbf{1}\|^2$. The adjusted $R^2 = 1 - (1-R^2)(n-1)/(n-p)$ penalizes for additional predictors. Under the null model (no predictors beyond the intercept), $R^2 \cdot (n-p)/((1-R^2)(p-1)) \sim F_{p-1,n-p}$.

Simultaneous confidence intervals. The Bonferroni correction produces simultaneous $(1-\alpha)$ intervals for all p coefficients by using $t_{n-p,\alpha/(2p)}$ in place of $t_{n-p,\alpha/2}$. The Scheffe method uses the F-distribution: $\hat{\beta}_j \pm \sqrt{pF_{p,n-p,\alpha}} \cdot \hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}$, which provides coverage for all linear combinations $a^\top\beta$ simultaneously.

Power of the F-test. Under the alternative $R\beta \neq r$, the F-statistic has a noncentral F distribution $F_{q,n-p,\delta}$ with noncentrality parameter

$$\delta = \frac{(R\beta - r)^\top [R(X^\top X)^{-1}R^\top]^{-1}(R\beta - r)}{\sigma^2}.$$

The power increases with δ, which depends on the effect size, the sample size (through $X^\top X$), and the error variance. Power analysis uses this noncentrality parameter to determine the required sample size.

Residual diagnostics. Under the model, the standardized residuals $r_i = \hat{\varepsilon}_i/(\hat{\sigma}\sqrt{1-h_{ii}})$ are approximately t_{n-p-1}. The externally studentized residuals $t_i = \hat{\varepsilon}_i/(\hat{\sigma}_{(i)}\sqrt{1-h_{ii}})$, where $\hat{\sigma}_{(i)}$ is estimated from the data excluding observation i, follow t_{n-p-1} exactly. Cook's distance $D_i = r_i^2 h_{ii}/(p(1-h_{ii}))$ combines leverage and residual magnitude to measure influence.

Analysis of variance (ANOVA) decomposition. The total sum of squares decomposes as

$$\text{TSS} = \text{ESS} + \text{RSS}, \quad \text{where ESS} = \|\hat{Y} - \bar{Y}\mathbf{1}\|^2,\ \text{RSS} = \|Y - \hat{Y}\|^2.$$

Under $H_0 : \beta_2 = \cdots = \beta_p = 0$ (only intercept), $\text{ESS}/\sigma^2 \sim \chi^2_{p-1}$ and is independent of $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$, yielding the overall F-test $F = (\text{ESS}/(p-1))/(\text{RSS}/(n-p)) \sim F_{p-1,n-p}$.

Connection to likelihood ratio. Under normality, the LR statistic for $H_0 : R\beta = r$ is

$$\Lambda = n\log(\text{RSS}_0/\text{RSS}),$$

where RSS_0 is the residual sum of squares under the restricted model. The F-statistic is a monotone transformation: $F = ((e^{\Lambda/n} - 1)(n-p))/q$. Thus the F-test is equivalent to the likelihood ratio test, with the F-distribution providing the exact (not asymptotic) null distribution.

Multiple testing. When testing many coefficients simultaneously, the individual t-test p-values require adjustment. The Bonferroni correction controls the family-wise error rate by using α/m as the per-test level. The Benjamini-Hochberg procedure controls the false discovery rate, which is less conservative and more appropriate for exploratory analysis with many predictors.

Worked Examples

Example 1. In simple linear regression with $n = 30$ observations, $\hat{\beta}_1 = 2.5$, $\text{SE}(\hat{\beta}_1) = 0.8$. The t-statistic is $T = 2.5/0.8 = 3.125$ on 28 df. The p-value is $2\mathbb{P}(t_{28} > 3.125) \approx 0.004$. The 95% CI is $2.5 \pm 2.048 \times 0.8 = (0.86, 4.14)$.

Example 2. Test $H_0 : \beta_2 = \beta_3 = 0$ in a model with $p = 5$ predictors and $n = 50$. Here $q = 2$, $R = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$, $r = 0$. If $\text{RSS}_{\text{full}} = 120$ and $\text{RSS}_{\text{reduced}} = 160$, then $F = ((160-120)/2)/(120/45) = 20/2.667 = 7.5$ on $(2, 45)$ df. The critical value $F_{2,45,0.05} \approx 3.20$, so we reject H_0.

Cross-Links

AI. The F-statistic for nested model comparison is the basis for feature selection in linear models. In deep learning, the degrees-of-freedom concept generalizes via effective degrees of freedom $\text{tr}(H_\lambda)$ for regularized estimators, connecting to generalization bounds.

Economics. The t-test and F-test are the primary tools in empirical economics. The Chow test for structural breaks is an F-test comparing pooled and split regressions. Robust inference under heteroskedasticity replaces the exact F with asymptotic chi-squared/F using the sandwich covariance (Node 7.5).

Linear Algebra. The distributional results rest on the spectral properties of idempotent matrices: a symmetric idempotent matrix of rank r has r eigenvalues equal to 1 and the rest zero. Quadratic forms in standard normals with such matrices give chi-squared distributions.

Quantum Mechanics. Hypothesis testing in quantum state discrimination uses analogous chi-squared statistics for goodness-of-fit testing of quantum states, with degrees of freedom determined by the dimension of the state space minus the number of estimated parameters.

Structural Compression

Invariant. Under Gaussian errors, all finite-sample inference in the linear model is determined by $\hat{\beta}$, $\hat{\sigma}^2$, and the hat matrix H; the t and F distributions arise from the chi-squared distribution of RSS/σ^2 and the independence of $\hat{\beta}$ and $\hat{\sigma}^2$.

Mechanism. Orthogonal decomposition $Y = HY + (I - H)Y$ separates signal from noise; independence follows from orthogonality under normality (Cochran's theorem); quadratic forms in standard normals yield chi-squared distributions; ratios of independent chi-squareds yield F; a standard normal divided by the square root of an independent chi-squared yields t.

Cross-System Echo. The F-statistic is the likelihood ratio statistic for the Gaussian linear model; the exact F-distribution replaces the asymptotic χ^2 approximation that holds in general parametric models. In ANOVA, the same F-test compares group means, decomposing total variation into between-group and within-group components.

Exercises

1. Derive the distribution of the t-statistic T_j from the joint distribution of $\hat{\beta}_j$ and $\hat{\sigma}^2$, verifying the t_{n-p} result.
2. Show that the F-test for $H_0 : \beta_j = 0$ is equivalent to T_j^2 by verifying $F_{1,n-p} = t_{n-p}^2$.
3. Prove that $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$ using the fact that HY and $(I - H)Y$ are functions of Y through orthogonal projections and Y is normal.
4. For the prediction interval, explain why the variance $\sigma^2(1 + x_0^\top (X^\top X)^{-1} x_0)$ has two components and identify which component is reducible with more data.
5. In a one-way ANOVA with $k = 3$ groups of size $n_j = 10$, derive the F-statistic as a ratio of between-group and within-group mean squares, and state the null distribution.
6. Show that $R^2 = 0$ if and only if $\hat{\beta}_j = 0$ for all $j \geq 2$ (assuming an intercept is included), and explain why R^2 always increases when adding predictors.
7. Argue, structurally, why the exact distributional results of this node (t-tests, F-tests) require Gaussian errors while the Gauss-Markov theorem (Node 8.4) does not, identifying exactly which properties of the normal distribution are needed and which are not.

Generalized Linear Models

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.6 — Generalized Linear Models.

GLMs extend the normal linear model to responses from any exponential family distribution, replacing the identity link with a general link function. This node depends on the linear model geometry (Node 8.3), exponential family theory (Node 2.6), and M-estimation (Node 7.5). It is the natural generalization of the Gaussian linear model to non-Gaussian responses.

Formal Definition

A generalized linear model specifies three components:

1. **Random component.** $Y_i \mid x_i$ independently follows an exponential family distribution with density

$$f(y_i; \eta_i, \phi) = h(y_i, \phi) \exp\left(\frac{y_i \eta_i - b(\eta_i)}{a(\phi)} \right),$$

where η_i is the canonical parameter, $b(\cdot)$ is the cumulant function, ϕ is the dispersion parameter, and $a(\phi) = \phi / w_i$ for known weights w_i.

2. **Systematic component.** A linear predictor $\eta_i^* = x_i^\top \beta$.

3. **Link function.** A monotone differentiable function g such that $g(\mu_i) = x_i^\top \beta$, where $\mu_i = \mathbb{E}[Y_i \mid x_i] = b'(\eta_i)$.

The **canonical link** sets $g = (b')^{-1}$, so that $\eta_i = x_i^\top \beta$ directly.

Intuition

The normal linear model ties the mean directly to the linear predictor: $\mu_i = x_i^\top \beta$. For binary or count data, this is inappropriate — the mean must be constrained (to $[0, 1]$ or $[0, \infty)$). The link function bridges the gap: it maps the constrained mean space to the unconstrained linear predictor space. The canonical link is special because it makes the sufficient statistics linear in the linear predictor, simplifying the score equations and guaranteeing concavity of the log-likelihood.

Structural Role

GLMs unify regression models for continuous, binary, and count responses under a single framework. The exponential family structure (Node 2.6) determines the variance function $V(\mu) = b''(\eta)$, the canonical link, and the form of the log-likelihood. The IRLS algorithm reduces all GLM fitting to a sequence of weighted least squares problems, specializing to OLS in the Gaussian case. Asymptotic inference follows from M-estimation theory (Node 7.5).

Mathematical Development

Mean-variance relationship. From exponential family theory, $\mu = b'(\eta)$ and $\mathrm{Var}(Y) = a(\phi)b''(\eta) = a(\phi)V(\mu)$, where $V(\mu) = b''((b')^{-1}(\mu))$ is the variance function. The variance function determines the distribution:

Distribution	$b(\eta)$	$V(\mu)$	Canonical link
Normal	$\eta^2/2$	1	Identity: $g(\mu) = \mu$
Bernoulli	$\log(1 + e^\eta)$	$\mu(1-\mu)$	Logit: $g(\mu) = \log \frac{\mu}{1-\mu}$
Poisson	e^η	μ	Log: $g(\mu) = \log \mu$
Gamma	$-\log(-\eta)$	μ^2	Reciprocal: $g(\mu) = 1/\mu$

Log-likelihood and score equations. For the canonical link ($\eta_i = x_i^\top \beta$):

$$\ell(\beta) = \sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i x_i^\top \beta - b(x_i^\top \beta) \right] + \text{const.}$$

The score is

$$\nabla_\beta \ell(\beta) = \frac{1}{\phi} \sum_{i=1}^{n} w_i (y_i - \mu_i) x_i = \frac{1}{\phi} X^\top W^{1/2} (y - \mu),$$

where $\mu_i = b'(x_i^\top \beta)$. Setting the score to zero gives the score equations $X^\top (y - \mu(\beta)) = 0$ (up to weights and ϕ), which generalize the normal equations of OLS.

For a general link, the score is

$$\nabla_\beta \ell = \sum_{i=1}^{n} \frac{w_i (y_i - \mu_i)}{a(\phi) V(\mu_i) g'(\mu_i)} x_i.$$

Fisher information. The expected Fisher information is

$$\mathcal{I}(\beta) = \frac{1}{\phi} X^\top W X,$$

where $W = \mathrm{diag}(w_1/(V(\mu_1)[g'(\mu_1)]^2), \ldots, w_n/(V(\mu_n)[g'(\mu_n)]^2))$. For the canonical link, $W = \mathrm{diag}(w_i V(\mu_i))$.

Fisher scoring / IRLS. The MLE is computed iteratively. Fisher scoring updates:

$$\beta^{(t+1)} = \beta^{(t)} + [\mathcal{I}(\beta^{(t)})]^{-1} \nabla_\beta \ell(\beta^{(t)}).$$

This is equivalent to Iteratively Reweighted Least Squares: at each step, solve

$$\hat{\beta}^{(t+1)} = (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} z^{(t)},$$

where the working response is $z_i^{(t)} = \hat{\eta}_i^{(t)} + (y_i - \hat{\mu}_i^{(t)}) g'(\hat{\mu}_i^{(t)})$ and $W^{(t)}$ is the weight matrix evaluated at $\hat{\mu}^{(t)}$.

Convergence. For the canonical link, the log-likelihood is concave (since b is convex), so IRLS converges to the unique global maximum. For non-canonical links, local convergence is guaranteed under smoothness conditions.

Logistic regression (Bernoulli GLM). For binary $Y_i \in \{0, 1\}$ with $\mathbb{P}(Y_i = 1|x_i) = \mu_i$, the canonical link is the logit: $\log(\mu_i/(1-\mu_i)) = x_i^\top \beta$, so $\mu_i = 1/(1+e^{-x_i^\top \beta})$. The log-likelihood is $\ell(\beta) = \sum_i [y_i x_i^\top \beta - \log(1+e^{x_i^\top \beta})]$. The variance function is $V(\mu) = \mu(1-\mu)$, and the weights are $w_i = \mu_i(1-\mu_i)$.

Poisson regression. For count $Y_i \in \{0, 1, 2, \ldots\}$ with $\mu_i = \mathbb{E}[Y_i|x_i]$, the canonical link is the log: $\log \mu_i = x_i^\top \beta$. The log-likelihood is $\ell(\beta) = \sum_i [y_i x_i^\top \beta - e^{x_i^\top \beta}] + \text{const}$. The variance equals the mean: $V(\mu) = \mu$, and the weights are $w_i = \mu_i$.

Quasi-likelihood. When only the mean-variance relationship $\text{Var}(Y_i) = \phi V(\mu_i)$ is specified (without a full distributional assumption), the quasi-likelihood score equations are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_i}{V(\mu_i) g'(\mu_i)} = 0.$$

These coincide with the GLM score equations, so the same IRLS algorithm applies. The quasi-likelihood estimator is consistent and asymptotically normal with the sandwich covariance, even without specifying the full distribution — only the first two conditional moments matter.

Negative binomial regression. For overdispersed count data where $\text{Var}(Y) > \mu$, the negative binomial GLM uses $V(\mu) = \mu + \mu^2/\kappa$ where κ is the overdispersion parameter. This nests Poisson regression ($\kappa \to \infty$) and accommodates extra-Poisson variation.

Probit model. An alternative to logistic regression for binary data uses the link $g(\mu) = \Phi^{-1}(\mu)$ where Φ is the standard normal CDF. The probit model arises naturally from a latent variable formulation: $Y_i = \mathbf{1}(x_i^\top \beta + \varepsilon_i > 0)$ with $\varepsilon_i \sim \mathcal{N}(0, 1)$. Logit and probit give nearly identical fitted probabilities in practice (after rescaling by $\pi/\sqrt{3} \approx 1.81$).

Deviance. The deviance is twice the log-likelihood ratio between the saturated model (one parameter per observation) and the fitted model:

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i(\tilde{\eta}_i - \hat{\eta}_i) - b(\tilde{\eta}_i) + b(\hat{\eta}_i) \right],$$

where $\tilde{\eta}_i$ is the canonical parameter of the saturated model ($\tilde{\mu}_i = y_i$). The deviance plays the role of RSS in the normal linear model. Under regularity conditions, the difference in deviances between nested models follows χ^2_q asymptotically, enabling likelihood ratio tests for model comparison.

Asymptotic inference. By the M-estimation theory of Node 7.5, the MLE satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0,\ \mathcal{I}(\beta)^{-1}).$$

Wald tests use $(\hat{\beta}_j/\text{SE}(\hat{\beta}_j))^2$; Score tests use $\nabla \ell(\beta_0)^\top \mathcal{I}(\beta_0)^{-1} \nabla \ell(\beta_0)$; LR tests use the deviance difference. All three converge to χ^2 under the null.

Worked Examples

Example 1 (Logistic regression). Data: 100 patients, binary outcome (disease/no disease), predictor $x =$ age. Fitted model: $\text{logit}(\hat{\mu}) = -5.0 + 0.08 \cdot \text{age}$. At age 50: $\hat{\mu} = 1/(1 + e^{-(-5+4)}) = 1/(1+e) \approx 0.269$. The odds ratio per year of age is $e^{0.08} \approx 1.083$: each year increases the odds of disease by 8.3%. The Wald 95% CI for the log-odds-ratio is $0.08 \pm 1.96 \times \text{SE}(\hat{\beta}_1)$.

Example 2 (Poisson regression). Data: counts of insurance claims by age group. Model: $\log \hat{\mu}_i = 1.2 + 0.03 \cdot \text{age}_i$. The rate ratio per year is $e^{0.03} \approx 1.030$. Deviance goodness-of-fit: if the residual deviance

is 45.2 on 48 df, the model fits adequately (p-value from χ^2_{48} is ≈ 0.59). If overdispersion is detected ($D/(n-p) \gg 1$), a quasi-Poisson model with estimated $\phi > 1$ inflates standard errors.

Cross-Links

AI. Logistic regression is the canonical binary classifier and the basis for the output layer of neural networks for classification. The cross-entropy loss in deep learning is the negative Bernoulli GLM log-likelihood. Softmax regression extends logistic regression to multinomial responses.

Economics. Probit and logit models are the primary binary choice models in microeconomics. Poisson regression is used for count data in labor economics (number of patents, doctor visits). The quasi-likelihood framework extends GLMs when the full distributional assumption is relaxed.

Linear Algebra. IRLS reduces each GLM iteration to a weighted least squares problem: $(X^\top W X)^{-1} X^\top W z$. The convergence of IRLS is analyzed via the contraction mapping theorem on the Fisher scoring iteration.

Quantum Mechanics. In quantum detector tomography, the detection probabilities follow a multinomial model that can be expressed as a GLM with a log-link on the POVM elements. The Fisher scoring algorithm applies to reconstruct the detector from count data.

Structural Compression

Invariant. The score equations $X^\top(y - \mu(\beta)) = 0$ generalize the normal equations of OLS to any exponential family response; the canonical link aligns the natural parameter with the linear predictor.

Mechanism. The link function maps the constrained mean space to the unconstrained linear predictor; the exponential family structure provides the variance function, score, and Fisher information. IRLS iteratively constructs a locally quadratic approximation via working responses and weights, converging to the MLE.

Cross-System Echo. Logistic regression is maximum entropy classification (AI connection). The deviance is the likelihood ratio statistic comparing fitted and saturated models (Node 5.4). The quasi-likelihood approach relaxes the full distributional assumption to the first two conditional moments, paralleling the Gauss-Markov approach in the linear model.

Exercises

1. Derive the canonical link for the Bernoulli distribution by writing the PMF in exponential family form and identifying $\eta = \log(\mu/(1-\mu))$.
2. Show that the score equations for logistic regression reduce to $X^\top(y - \mu(\beta)) = 0$ and verify that the log-likelihood is concave.
3. For Poisson regression with canonical link, derive the Fisher information matrix and the IRLS weight matrix.
4. Prove that the deviance for the normal linear model equals RSS/σ^2, recovering the residual sum of squares as a special case.
5. Fit a logistic regression to the data $(x, y) = \{(1,0), (2,0), (3,1), (4,0), (5,1), (6,1), (7,1)\}$ by writing the log-likelihood and performing one Newton-Raphson iteration starting from $\beta = (0, 0)$.
6. Explain the concept of overdispersion in the Poisson GLM, derive the quasi-Poisson variance $\text{Var}(Y_i) = \phi\mu_i$, and describe how standard errors are adjusted.

7. Argue, structurally, why the canonical link occupies a privileged position among all possible link functions, connecting its role to sufficiency, concavity of the log-likelihood, and the information matrix equality.

Principal Component Analysis

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.7 — Principal Component Analysis.

PCA is the canonical method for dimensionality reduction via variance maximization. It connects the spectral theorem from linear algebra to the statistical problem of finding the most informative low-dimensional representation of high-dimensional data. This node depends on the covariance structure (Node 8.1) and the multivariate normal (Node 8.2).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} (\mu, \Sigma)$ with $X_i \in \mathbb{R}^p$. The k-th **principal component direction** $v_k \in \mathbb{R}^p$ solves

$$v_k = \arg \max_{\|v\|=1,\ v \perp v_1, \ldots, v_{k-1}} v^\top \Sigma v.$$

The k-th **principal component score** is $Z_k = v_k^\top (X - \mu)$.

The solution is the eigendecomposition $\Sigma = \sum_{j=1}^{p} \lambda_j v_j v_j^\top$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, where v_k is the k-th eigenvector.

Intuition

PCA rotates the coordinate system to align with the directions of greatest variability. The first principal component captures the most variance; each subsequent component captures the maximum remaining variance orthogonal to all previous ones. The eigenvalues are the variances along these directions. If the data lies approximately in a low-dimensional subspace, the first few components explain most of the total variance and the rest can be discarded with minimal information loss. PCA is a descriptive procedure — it requires no probabilistic model — but it is the optimal linear dimensionality reduction under the mean-squared reconstruction error criterion.

Structural Role

PCA identifies the directions of greatest variance, providing a basis for low-rank representation, visualization, and preprocessing. It is the bridge between the algebraic spectral theorem and statistical data analysis. PCA connects to probabilistic PCA (a latent variable model with Gaussian noise), factor analysis, and the SVD of the data matrix. In high-dimensional statistics, PCA is both a tool and an object of study (spiked covariance models, random matrix theory).

Mathematical Development

PCA as variance maximization.

Theorem. The maximum of $v^\top \Sigma v$ subject to $\|v\| = 1$ is λ_1, attained at $v = v_1$.

Proof. By Lagrange multipliers, the stationary condition is $\Sigma v = \lambda v$, so the critical points are eigenvectors with critical values equal to eigenvalues. The maximum is the largest eigenvalue λ_1.

For the k-th component, the constraint $v \perp v_1, \ldots, v_{k-1}$ restricts to the subspace orthogonal to the first $k-1$ eigenvectors. On this subspace, Σ acts with eigenvalues $\lambda_k, \ldots, \lambda_p$, so the maximum is λ_k at v_k. $\square$

PCA as minimum reconstruction error.

Theorem. The rank-r linear subspace minimizing the expected reconstruction error is $\text{span}(v_1, \ldots, v_r)$:

$$\min_{\dim(S)=r} \mathbb{E}\left[\|X - \mu - \text{proj}_S(X-\mu)\|^2\right] = \sum_{j=r+1}^{p} \lambda_j.$$

Proof. Let $V_r = [v_1, \ldots, v_r]$. The projection is $\text{proj}_S(X-\mu) = V_r V_r^\top (X-\mu)$ and the reconstruction error is

$$\begin{aligned}\mathbb{E}[\|X - \mu - V_r V_r^\top (X-\mu)\|^2] &= \mathbb{E}[\text{tr}((I - V_r V_r^\top)(X-\mu)(X-\mu)^\top (I - V_r V_r^\top))] \\ &= \text{tr}((I - V_r V_r^\top)\Sigma) = \sum_{j=1}^{p} \lambda_j - \sum_{j=1}^{r} \lambda_j = \sum_{j=r+1}^{p} \lambda_j.\end{aligned}$$

For any other rank-r subspace, the Eckart-Young theorem (or the Fan-Poincare minimax principle) shows the error is at least $\sum_{j=r+1}^{p} \lambda_j$. $\square$

Variance explained. The proportion of variance explained by the first r components is

$$\frac{\sum_{j=1}^{r} \lambda_j}{\text{tr}(\Sigma)} = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

The scree plot displays λ_j vs. j; a sharp drop ("elbow") suggests the appropriate number of components. The cumulative proportion curve $\sum_{j=1}^{r} \lambda_j / \text{tr}(\Sigma)$ provides a threshold-based selection rule (e.g., retain components until 95% of variance is explained).

Connection to SVD. Let $\tilde{X} \in \mathbb{R}^{n \times p}$ be the centered data matrix with rows $(X_i - \bar{X})^\top$. The singular value decomposition $\tilde{X} = UDV^\top$ (with $U \in \mathbb{R}^{n \times p}$, $D = \text{diag}(d_1, \ldots, d_p)$, $V \in \mathbb{R}^{p \times p}$ orthogonal) gives:

- V: columns are the principal component directions (right singular vectors = eigenvectors of $\hat{\Sigma}$).
- $d_j^2/(n-1) = \hat{\lambda}_j$: the sample eigenvalues.
- UD: the matrix of principal component scores.
- $\hat{\Sigma} = V(D^2/(n-1))V^\top$.

The SVD is computationally preferable to forming $\hat{\Sigma}$ explicitly, avoiding the squaring of condition numbers.

Sample PCA. Replace Σ by $\hat{\Sigma} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top$. The sample principal directions $\hat{v}_k$ are eigenvectors of $\hat{\Sigma}$. When population eigenvalues are distinct ($\lambda_j \neq \lambda_k$ for $j \neq k$), $\hat{v}_k \xrightarrow{p} v_k$ (up to sign) by the continuous mapping theorem applied to the eigendecomposition map.

When eigenvalues coincide ($\lambda_j = \lambda_{j+1}$), the eigenvectors are not uniquely identified — any orthonormal basis of the eigenspace is valid — and the sample eigenvectors converge only to the eigenspace, not to a specific direction.

Asymptotic distribution of eigenvalues. Under normality or finite fourth moments, for distinct eigenvalues,

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0,\ 2\lambda_j^2) \quad \text{(under MVN)}.$$

The asymptotic variance of $\hat{v}_j$ depends on the eigenvalue gaps: $\text{Var}(\hat{v}_j^\top v_k) \approx \lambda_j \lambda_k / (n(\lambda_j - \lambda_k)^2)$ for $k \neq j$. Small gaps produce unreliable eigenvector estimates.

Probabilistic PCA (Tipping-Bishop). Assume the generative model $X = Wz + \mu + \varepsilon$ where $z \sim \mathcal{N}(0, I_r)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$, and $W \in \mathbb{R}^{p \times r}$. The marginal is $X \sim \mathcal{N}(\mu, WW^\top + \sigma^2 I)$. The MLE for W satisfies $\hat{W} = V_r(\Lambda_r - \sigma^2 I)^{1/2} R$ where V_r contains the first r eigenvectors, Λ_r the corresponding eigenvalues, and R is an arbitrary orthogonal matrix. As $\sigma^2 \to 0$, this recovers classical PCA.

PCA on the correlation matrix. When variables have different units, PCA is applied to the correlation matrix R rather than Σ. This is equivalent to standardizing each variable to unit variance before computing PCA. The choice between Σ and R is a modeling decision: Σ-PCA preserves the original scale; R-PCA treats all variables equally.

High-dimensional PCA and spiked models. When $p/n \to \gamma \in (0, \infty)$, the sample eigenvalues of $\hat{\Sigma}$ are inconsistent for the population eigenvalues due to the Marchenko-Pastur law. In the spiked covariance model $\Sigma = I_p + \sum_{j=1}^{r} \ell_j v_j v_j^\top$, a phase transition occurs: the j-th sample eigenvalue separates from the bulk only when $\ell_j > \sqrt{\gamma}$ (the BBP transition). Below this threshold, the signal is undetectable by PCA.

Kernel PCA. Classical PCA finds directions in the input space $\mathbb{R}^p$. Kernel PCA maps data to a feature space $\phi(x) \in \mathcal{H}$ via a kernel $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ and performs PCA in $\mathcal{H}$. The principal components are computed from the $n \times n$ kernel matrix $K_{ij} = K(X_i, X_j)$ without explicitly constructing ϕ. This enables nonlinear dimensionality reduction.

Relationship to factor analysis. Factor analysis posits $X = \Lambda F + \varepsilon$ with $F \sim \mathcal{N}(0, I_r)$, $\varepsilon \sim \mathcal{N}(0, \Psi)$ where Ψ is diagonal, giving $\Sigma = \Lambda\Lambda^\top + \Psi$. Unlike PCA, factor analysis separates common variance ($\Lambda\Lambda^\top$) from unique variance (Ψ). PCA is equivalent to factor analysis only when $\Psi = \sigma^2 I$ (the probabilistic PCA model).

Courant-Fischer minimax characterization. The eigenvalues of Σ admit the variational characterization:

$$\lambda_k = \min_{\dim(S) = p-k+1} \max_{v \in S, \|v\|=1} v^\top \Sigma v = \max_{\dim(S) = k} \min_{v \in S, \|v\|=1} v^\top \Sigma v.$$

This provides the theoretical basis for PCA: the top k eigenvalues represent the maximum achievable total variance captured by any k-dimensional subspace.

Sparse PCA. In high dimensions, classical PCA eigenvectors are dense and difficult to interpret. Sparse PCA adds an ℓ_1 penalty:

$$\max_v v^\top \Sigma v - \lambda \|v\|_1 \quad \text{subject to } \|v\|_2 = 1,$$

producing sparse loading vectors. This loses optimality in reconstruction error but gains interpretability and consistency in high-dimensional settings where $p \gg n$.

Johnson-Lindenstrauss vs. PCA. Random projections preserve pairwise distances up to distortion ε in $O(\log n/\varepsilon^2)$ dimensions, regardless of the data's intrinsic structure. PCA optimally preserves variance but requires $O(np^2)$ computation. For very high-dimensional problems, random projections provide a computational alternative with different statistical guarantees.

Worked Examples

Example 1. Let $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$. The eigenvalues satisfy $\lambda^2 - 7\lambda + 1 = 0$, giving $\lambda_1 = (7 + 3\sqrt{5})/2 \approx 6.854$, $\lambda_2 = (7 - 3\sqrt{5})/2 \approx 0.146$. The first PC explains $6.854/7 \approx 97.9\%$ of the variance. The eigenvector v_1 satisfies $(5 - \lambda_1)v_{11} + 3v_{12} = 0$, giving $v_1 \propto (3, \lambda_1 - 5)^\top \approx (3, 1.854)^\top$, normalized to unit length. The data is nearly one-dimensional: the principal direction captures almost all variability.

Example 2 (SVD computation). Given centered data matrix $\tilde{X} = \begin{pmatrix} 2 & 1 \\ -1 & 1 \\ -1 & -2 \end{pmatrix}$, compute $\tilde{X}^\top \tilde{X} = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$. Eigenvalues: 9 and 3, with eigenvectors $(1,1)^\top/\sqrt{2}$ and $(1,-1)^\top/\sqrt{2}$. Sample variances: $\hat{\lambda}_1 = 9/2 = 4.5$, $\hat{\lambda}_2 = 3/2 = 1.5$. First PC direction: $\hat{v}_1 = (1,1)^\top/\sqrt{2}$, the 45-degree line. Scores: $Z_1 = \tilde{X}\hat{v}_1 = (3/\sqrt{2}, 0, -3/\sqrt{2})^\top$. Proportion of variance: $4.5/6 = 75\%$.

Cross-Links

AI. PCA is the basis for whitening, feature extraction, and linear autoencoders. A linear autoencoder with r hidden units recovers the PCA projection. In NLP, PCA applied to word embedding matrices identifies semantic dimensions. Kernel PCA extends to nonlinear dimensionality reduction via the kernel trick.

Economics. PCA on macroeconomic time series extracts latent factors (diffusion indices) used in forecasting. In finance, PCA on bond yield curves identifies level, slope, and curvature factors that explain >95% of yield variation.

Linear Algebra. PCA is the spectral theorem applied to the covariance matrix. The Eckart-Young theorem establishes the optimality of the truncated SVD for low-rank approximation in Frobenius norm. The Davis-Kahan theorem quantifies how perturbations in Σ affect the eigenvectors.

Quantum Mechanics. PCA on quantum state tomography data decomposes the noise covariance of reconstructed states. The eigenvalues of the density matrix ρ are the quantum analogue of PCA eigenvalues: they determine the effective dimensionality (rank) of the quantum state.

Structural Compression

Invariant. The principal components are the eigenvectors of the covariance matrix; they diagonalize Σ and provide the variance-maximizing (equivalently, reconstruction-error-minimizing) orthonormal basis.

Mechanism. The spectral theorem for symmetric positive semidefinite matrices decomposes Σ into rank-one projections $\lambda_j v_j v_j^\top$; each successive eigenvector maximizes residual variance subject to orthogonality. The SVD of the data matrix provides a computationally stable path to the same decomposition.

Cross-System Echo. PCA is the method of moments estimator in probabilistic PCA (Tipping-Bishop). In information theory, PCA maximizes the mutual information between the data and its linear projection under Gaussianity. The connection between variance maximization and reconstruction error minimization is an instance of the duality between primal and dual optimization.

Exercises

1. Prove that the k-th principal component direction maximizes $v^\top \Sigma v$ subject to $\|v\| = 1$ and $v \perp v_1, \ldots, v_{k-1}$ using Lagrange multipliers.
2. Show that PCA on Σ and PCA on $\hat{\Sigma}$ are equivalent to performing the SVD of the centered data matrix, by relating the eigendecomposition of $\hat{\Sigma}$ to the right singular vectors of $\tilde{X}$.
3. For a 3×3 covariance matrix with eigenvalues $\lambda_1 = 10$, $\lambda_2 = 2$, $\lambda_3 = 0.5$, compute the proportion of variance explained by the first two components and the reconstruction error of the rank-2 approximation.
4. Prove that the total variance $\mathrm{tr}(\Sigma)$ is invariant under orthogonal transformations $Q^\top \Sigma Q$, explaining why PCA redistributes but does not create or destroy variance.

5. In the probabilistic PCA model $X = Wz + \mu + \varepsilon$, derive the marginal distribution of X and show that the MLE for μ is $\bar{X}$.

6. Explain why PCA on the correlation matrix vs. the covariance matrix can give qualitatively different results, and give a concrete example where one is appropriate and the other is not.

7. Argue, structurally, why PCA is simultaneously the optimal linear dimensionality reduction under two different criteria (maximum variance and minimum reconstruction error), and explain why this duality breaks down for nonlinear methods.