Review: The Multivariate Normal Distribution

version: 2020-01-07 · 10:58:30

STAT 440/840 – CM 761: Computational Inference

The Multivariate Normal Distribution

- **Definition:** $\boldsymbol{X} = (X_1, \ldots, X_d)$ is multivariate normal if and only if it a linear combination of iid normals:

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{Z} + \boldsymbol{\mu}, \qquad \boldsymbol{Z} = (Z_1, \ldots, Z_d), \quad Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1).$$

- **Mean and Variance:**
  - $E[\boldsymbol{X}] = \boldsymbol{C}\, E[\boldsymbol{Z}] + \boldsymbol{\mu} = \boldsymbol{\mu}$
  - $\text{var}(\boldsymbol{X}) = \boldsymbol{C}\, \text{var}(\boldsymbol{Z}) \boldsymbol{C}' = \boldsymbol{C}\boldsymbol{C}' := \boldsymbol{\Sigma}$ (many different $\boldsymbol{C}$ give the same variance)
- **Notation:** $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **PDF:**

$$f(\boldsymbol{x}) = (2\pi)^{-d/2} \cdot \exp\left\{ -\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) - \tfrac{1}{2} \log|\boldsymbol{\Sigma}| \right\}$$

Simulation

- To generate $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:
  1. Find $\boldsymbol{C}$ such that $\boldsymbol{C}\boldsymbol{C}' = \boldsymbol{\Sigma}$.
  2. Generate $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$ with $Z_i \sim \mathcal{N}(0, 1)$.
  3. Set $\boldsymbol{X} = \boldsymbol{C}\boldsymbol{Z} + \boldsymbol{\mu}$.
- To find $\boldsymbol{C}$:
  - Note that $\boldsymbol{\Sigma}$ is (i) symmetric and (ii) positive definite: for any vector $\boldsymbol{a}$ we have $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \geq 0$, with equality $\iff \boldsymbol{a} = 0$.
  - Every symmetric positive-definite matrix has a Cholesky definition: $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is lower triangular and all diagonal elements $L_{ii} > 0$.
  - *Properties:*
    - The eigenvalues of triangular matrices are the diagonal elements $\implies |\boldsymbol{\Sigma}| = \prod_{i=1}^{d} L_{ii}^2$.
    - $\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = (\boldsymbol{L}')^{-1}(\boldsymbol{L}^{-1}\boldsymbol{x})$. Solving linear systems with triangular matrices is $\mathcal{O}(d)$, as opposed to $\mathcal{O}(d^3)$ for general matrices.

Conditional Distribution

- **Block Notation:**

$$\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim \mathcal{N}\left\{ \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right\}$$

- **Marginal Distribution:** $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
- **Conditional Distribution:**

$$\boldsymbol{X}_2 \,|\, \boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_2^\star, \boldsymbol{\Sigma}_2^\star), \qquad \begin{aligned} \boldsymbol{\mu}_2^\star &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1) \\ \boldsymbol{\Sigma}_2^\star &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \end{aligned}$$

- **Verify Calculations:** $f(\boldsymbol{x}_1, \boldsymbol{x}_2) = f(\boldsymbol{x}_1) \times f(\boldsymbol{x}_2 \,|\, \boldsymbol{x}_1)$ for any pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$.

  Can do this analytically (harder) or computationally (easier)

Parameter Inference

- **Data:** $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n), \qquad \boldsymbol{X}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- **Loglikelihood function:**

$$\begin{aligned}
\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \,|\, \boldsymbol{X}) &= \log \prod_{i=1}^{n} f(\boldsymbol{X}_i \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - \{\text{terms not involving } \boldsymbol{\mu} \text{ or } \boldsymbol{\Sigma}\} \\
&= -\frac{1}{2}\Big\{ n \log |\boldsymbol{\Sigma}| + \textstyle\sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}) \Big\} \\
&= -\frac{1}{2}\Big\{ n \log |\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}) + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \Big\},
\end{aligned}$$

  where $\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i, \qquad \boldsymbol{S}_{d \times d} = \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$

- **MLE:** $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/n$.

Parameter Inference

- **Data:** $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n), \qquad \boldsymbol{X}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$
- **Loglikelihood function:**

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) = -\frac{1}{2}\Big\{ n \log |\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \Big\},$$

  where $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i, \qquad \boldsymbol{S}_{d\times d} = \sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$

- **Verify Calculations:** Can do this analytically (harder) or computationally (easier):
  - Check that

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X}) = \log f(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \text{CONST}$$

    for fixed $\boldsymbol{X}$ and varying $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Parameter Inference

- **Data:** $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n), \qquad \boldsymbol{X}_i \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$
- **MLE:** $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/n$,

  where $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i, \qquad \boldsymbol{S}_{d\times d} = \sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$
- **Verify Calculations:** Can do this analytically (harder) or computationally (easier):
  - For differentiable loglikelihood $\ell(\boldsymbol{\theta} \mid \boldsymbol{X})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, each component of MLE is the maximum of a 1-d function:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \boldsymbol{X}) \implies \hat{\theta}_i = \arg\max_{\theta_i} \ell(\theta_i, \hat{\boldsymbol{\theta}}_{[-i]} \mid \boldsymbol{X}),$$

    where $\hat{\boldsymbol{\theta}}_{[-i]} = \hat{\boldsymbol{\theta}} \setminus \{\hat{\theta}_i\}$.
  - Converse is false (as $\hat{\boldsymbol{\theta}}$ could be a saddlepoint). However, loglikelihoods are generally well-behaved, so if $\hat{\theta}_i = \arg\max_{\theta_i} \ell(\theta_i, \hat{\boldsymbol{\theta}}_{[-i]} \mid \boldsymbol{X})$ for $i = 1, \ldots, p$, then very likely that $\hat{\boldsymbol{\theta}}$ is the MLE.

Applications

1. Confidence Intervals

- **Model:** $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta}), \qquad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_d).$
- **MLE:** $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \boldsymbol{Y}), \qquad \ell(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \sum_{i=1}^{n} \log f(Y_i \mid \boldsymbol{\theta}).$
- **Asymptotic Theory:** As $n \to \infty$, we have

$$\hat{\boldsymbol{\theta}} \approx \mathcal{N}\left(\boldsymbol{\theta}_0, \mathcal{I}_0^{-1}\right), \qquad \text{where}$$

  - $\boldsymbol{\theta}_0$ is the true parameter value
  - $\mathcal{I}_0$ is the (expected) Fisher Information:

$$\mathcal{I}_0 = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}_0 \mid \boldsymbol{Y})\right] = -n \int \left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log f(y \mid \boldsymbol{\theta}_0)\right] \cdot f(y \mid \boldsymbol{\theta}_0)\, \mathrm{d}y$$

Applications

1. Confidence Intervals

- **Model:** $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y \mid \boldsymbol{\theta})$
- **Asymptotic Theory:** As $n \to \infty$, we have $\hat{\boldsymbol{\theta}} \approx \mathcal{N}\left(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}_0^{-1}\right)$, where
  - θ_0 is the true parameter value
  - $\boldsymbol{\mathcal{I}}_0 = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}_0 \mid \boldsymbol{Y})\right]$ is the (expected) Fisher Information.

    Typically $\boldsymbol{\mathcal{I}}_0$ is impossible to calculate because (i) expectation is usually intractable and (ii) true $\boldsymbol{\theta}_0$ is unknown.

    Observed Fisher Information is a consistent estimator: $\hat{\boldsymbol{\mathcal{I}}} = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{Y}) \overset{n}{\to} \boldsymbol{\mathcal{I}}_0$
- **Asymptotic Confidence Intervals**: 95% CI for each element of $\boldsymbol{\theta}$:

$$\hat{\theta}_i \pm 1.96 \cdot \text{se}(\hat{\theta}_i), \qquad \text{se}(\hat{\theta}_i) = \sqrt{[\hat{\boldsymbol{\mathcal{I}}}^{-1}]_{ii}}.$$

  Such CI's are often valid even without iid data.

Applications

2. Linear Regression

- **Model:** $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{V})$, where
  - $\boldsymbol{y} = (y_1, \ldots, y_n)$ is multivariate normal (random)
  - $\boldsymbol{X}_{n \times p}$ and $\boldsymbol{V}_{n \times n}$ are known (nonrandom)
  - $\beta = (\beta_1, \ldots, \beta_p)$ and σ are unknown (parameters)
- **Loglikelihood:**

$$\begin{aligned}\ell(\beta, \sigma \,|\, \boldsymbol{y}) &= -\frac{1}{2}\Big\{(\boldsymbol{y} - \boldsymbol{X}\beta)'[\sigma^2 \boldsymbol{V}]^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta) + \log|\sigma^2 \boldsymbol{V}|\Big\} \\ &= -\frac{1}{2}\left\{\frac{(\beta - \hat{\beta})'\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}(\beta - \hat{\beta}) + n\hat{\sigma}^2}{\sigma^2} + n\log\sigma^2\right\},\end{aligned}$$

  where $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$ and $\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$.
- **Inference:** The MLE of $\boldsymbol{\theta} = (\beta, \sigma)$ is $\hat{\boldsymbol{\theta}} = (\hat{\beta}, \hat{\sigma})$.