Review: The Multivariate Normal Distribution

version: 2018-01-09 · 09:55:34

The Multivariate Normal Distribution

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

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

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

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

Simulation

- To generate $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

  1. Find $\mathbf{C}$ such that $\mathbf{C}\mathbf{C}' = \boldsymbol{\Sigma}$.

  2. Generate $\mathbf{Z} = (Z_1, \ldots, Z_d)$ with $Z_i \sim \mathcal{N}(0, 1)$.

  3. Set $\mathbf{X} = \mathbf{C}\mathbf{Z} + \boldsymbol{\mu}$.

- To find $\mathbf{C}$:
  - Note that $\boldsymbol{\Sigma}$ is (i) symmetric and (ii) positive definite: for any vector $\mathbf{a}$ we have $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \geq 0$, with equality $\iff$ $\mathbf{a} = 0$ (since $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X}) \implies \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \mathrm{var}(\mathbf{a}'\mathbf{X}) \geq 0$)
  - Every symmetric +ve definite matrix has a Cholesky definition: $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$ where $\mathbf{L}$ is lower triangular and all diagonal elements $L_{ii} > 0$.
  - *Properties:*
    - The eigenvalues of triangular matrices are the diagonal elements $\implies |\boldsymbol{\Sigma}| = \prod_{i=1}^{d} L_{ii}^2$.
    - $\boldsymbol{\Sigma}^{-1}\mathbf{x} = (\mathbf{L}')^{-1}(\mathbf{L}^{-1}\mathbf{x})$. Solving linear systems with triangular matrices is $\mathcal{O}(d)$, as opposed to $\mathcal{O}(d^3)$ for general matrices. (Cholesky decomposition itself is $\mathcal{O}(d^3)$, but solving system this way is much more accurate.)

Conditional Distribution

- **Block Notation:**

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

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

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

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

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

Parameter Inference

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

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

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

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

Parameter Inference

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

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

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

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

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

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

Parameter Inference

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

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

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

    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 $\hat{\theta}_i = \arg\max_{\theta_i} \ell(\theta_i, \hat{\boldsymbol{\theta}}_{[-i]} \mid \mathbf{X})$ 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:** $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \mathbf{Y}), \qquad \ell(\boldsymbol{\theta} \mid \mathbf{Y}) = \sum_{i=1}^n \log f(Y_i \mid \boldsymbol{\theta})$.
- **Asymptotic Theory:** As $n \to \infty$, we have

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

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

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

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 $\hat{\boldsymbol{\theta}} \approx \mathcal{N}\big(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}_0^{-1}\big)$, where
  - θ_0 is the true parameter value
  - $\boldsymbol{\mathcal{I}}_0 = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}_0 \mid \mathbf{Y})\right]$ 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: $\hat{\boldsymbol{\mathcal{I}}} = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\hat{\boldsymbol{\theta}} \mid \mathbf{Y}) \xrightarrow{n} \boldsymbol{\mathcal{I}}_0$
- **Asymptotic Confidence Intervals**: 95% CI for each element of $\boldsymbol{\theta}$:

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

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

Applications

2. Linear Regression

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

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

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