Biostatistics 682: Applied Bayesian Inference
Lecture 7: Multivariate Normal Model

Jian Kang

Department of Biostatistics
University of Michigan, Ann Arbor

Multivariate Normal Distribution

- Let $y = (y_1, \ldots, y_d)$ be a d-dimensional column vector.

$$y \mid \mu, \Sigma \overset{\text{iid}}{\sim} \text{MVN}(\mu, \Sigma),$$

  where μ is also a d-dimensional column vector and Σ is a symmetric positive definite (SPD) matrix.

- What is an SPD matrix?

$$A = A^{\mathrm{T}} \text{ and } x^{\mathrm{T}} A x > 0, \text{ for any } x \in \mathbb{R}^d,$$

- The density function of multivariate normal distribution is given by

$$\pi(y \mid \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (y - \mu)^{\mathrm{T}} \Sigma^{-1} (y - \mu) \right\}$$

Different prior models

- Conjugate prior for μ with Σ known.
- Improper, uninformative prior for μ with Σ known.
- Conjugate prior for (μ, Σ).
- Uninformative prior for (μ, Σ).

Conjugate prior for μ with Σ known

- Suppose $y_i \sim \mathrm{MVN}(\mu, \Sigma)$, for $i = 1, \ldots, n$. Let $Y = (y_1^{\mathrm{T}}, \ldots, y_n^{\mathrm{T}})^{\mathrm{T}}$.
- What is the likelihood function of μ?

$$\pi(Y \mid \mu) \propto \exp\left\{\mu^{\mathrm{T}} A(Y) \mu + \mu^{\mathrm{T}} B(Y)\right\}$$

$$A(Y) = -n(2\Sigma)^{-1}, \qquad B(Y) = -n\Sigma^{-1}\bar{y},$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$.

- The conjugate prior for μ when Σ.

$$\pi(\mu \mid Y) \propto \exp\left\{\mu^{\mathrm{T}} A_0 \mu + \mu^{\mathrm{T}} B_0\right\}.$$

We have

$$\mu \sim \mathrm{MVN}(\theta, \Lambda).$$

This implies that

$$\Lambda = -(2A_0)^{-1}, \quad \theta = \Lambda B_0.$$

Conjugate prior for μ with Σ known

- The posterior distribution is given by

$$\pi(\mu \mid Y) \propto \exp\left[\mu^{\mathrm{T}}\{A_0 + A(Y)\}\mu + \mu^{\mathrm{T}}\{B_0 + B(Y)\}\right].$$

This implies that

$$\mu \mid Y \sim \mathrm{MVN}(\mu_p, \Lambda_p)$$

where

$$\mu_p = (\Lambda^{-1} + n\Sigma^{-1})^{-1}(\Lambda^{-1}\theta + n\Sigma^{-1}\bar{y}).$$
$$\Lambda_p = (\Lambda^{-1} + n\Sigma^{-1})^{-1}.$$

- The posterior mean is a weighted average of the sample mean and the prior mean.
- The posterior precision is the sum of the data precision and the prior precision.

Multivariate normal distribution

- Partition the posterior mean $\mu_p = (\mu_{p,1}^{\mathrm{T}}, \mu_{p,2}^{\mathrm{T}})^{\mathrm{T}}$ as well as the posterior covariance matrix
$$\Lambda_p = \begin{pmatrix} \Lambda_{p,11} & \Lambda_{p,12} \\ \Lambda_{p,21} & \Lambda_{p,22} \end{pmatrix}.$$
- Partition the prior mean θ and covariance Λ as well as the posterior mean μ_p and covariance Λ_p.
- The conditional posterior of μ_1 given μ_2 is
$$(\mu_1 \mid \mu_2, Y) \sim \mathrm{MVN}(\mu_{p,1} + \Lambda_{p,12}\Lambda_{p,22}^{-1}(\mu_2 - \mu_{p,2}), \Lambda_{p,1|2}),$$
where the covariance is the Schur complement of Λ_p,
$$\Lambda_{p,1|2} = \Lambda_{p,11} - \Lambda_{p,12}\Lambda_{p,22}^{-1}\Lambda_{p,21}.$$
- What is the marginal posterior of μ_1?
$$\mu_1 \sim \mathrm{MVN}(\mu_{p,1}, \Lambda_{p,11}).$$
- What is the posterior predictive distribution: $\pi(\tilde{Y} \mid Y)$?
$$\tilde{Y} \mid Y \sim \mathrm{MVN}(\mu_p, \Lambda_p + \Sigma).$$

Uninformative prior for μ with Σ known

- Start with the conjugate prior for μ:

$$\mu \sim \mathrm{MVN}(\theta, \Lambda)$$

  and let the determinant of the prior precision go to zero: $|\Lambda^{-1}| \to 0$.
- Write

$$\pi(\mu) \propto 1.$$

- The posterior for μ becomes

$$\mu \mid Y \sim \mathrm{MVN}(\bar{y}, \Sigma/n).$$

Conjugate prior for (μ, Σ)

- The conjugate prior follows along the same lines as for the univariate case. Let
$$\Sigma^{-1} \sim W(\nu, \Lambda) \qquad \mu \mid \Sigma^{-1} \sim \mathrm{N}(\theta, \Sigma/k).$$
- $W(\nu, \Lambda)$ represents a Wishart distribution with ν degrees of freedom and $d \times d$ SPD scale matrix Λ.
- The density of Σ^{-1} is given by
$$\pi(\Sigma^{-1}) = \frac{|\Lambda|^{-\nu/2}|\Sigma^{-1}|^{(\nu-d-1)/2}\exp\{-1/2\mathrm{tr}(\Lambda^{-1}\Sigma^{-1})\}}{2^{\nu d/2}\Gamma_d(\nu/2)}.$$
where
$$\Gamma_d(x) = \int_{S>0} \exp\{-\mathrm{tr}(S)\}|S|^{x-(d+1)/2}dS.$$
and
$$\Gamma_d(\nu/2) = \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma\{(\nu+1-i)/2\}.$$

More on Wishart Distribution

- If $z_1, \ldots, z_\nu \sim \mathrm{MVN}(0, \Lambda)$, then $\Sigma^{-1} = \sum_{l=1}^{\nu} z_l z_l^{\mathrm{T}} \sim W(\nu, \Lambda)$.
- If $\nu > d$, then Σ^{-1} is positive definite with probability one.
- Σ^{-1} is symmetric with probability one.
- $\mathrm{E}(\Sigma^{-1}) = \nu\Lambda$.
- $\Sigma \sim W^{-1}(\nu, \Lambda^{-1})$, which is an inverse-Wishart distribution. The density function is given by

$$\pi(\Sigma) = \frac{|\Lambda|^{-\nu/2}|\Sigma|^{-(\nu+d+1)/2}\exp\{-1/2\mathrm{tr}(\Lambda^{-1}\Sigma^{-1})\}}{2^{\nu d/2}\Gamma_d(\nu/2)}.$$

- $\mathrm{E}(\Sigma) = (\nu - d - 1)^{-1}\Lambda^{-1}$.

Joint posterior distribution

- The joint posterior is then

$$
\begin{aligned}
\pi(\mu, \Sigma^{-1} \mid Y) \propto |\Sigma^{-1}|^{\frac{\nu-d}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}\Sigma^{-1}) - \frac{k}{2}(\mu-\theta)^{\mathrm{T}}\Sigma^{-1}(\mu-\theta)\right\} \\
\times |\Sigma^{-1}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^{\mathrm{T}}\Sigma^{-1}(y_i-\mu)\right\} \\
= |\Sigma^{-1}|^{\frac{n+\nu-d}{2}} \exp\left\{-\frac{1}{2}\left[\mathrm{tr}(\Lambda^{-1}\Sigma^{-1}) + k(\mu-\theta)^{\mathrm{T}}\Sigma^{-1}(\mu-\theta) + \right.\right. \\
\left.\left.\sum_{i=1}^{n}(y_i-\mu)^{\mathrm{T}}\Sigma^{-1}(y_i-\mu)\right]\right\}.
\end{aligned}
$$

Marginal posterior distribution

- Integrating out μ, we have

$$\pi(\Sigma^{-1} \mid Y) \propto |\Sigma^{-1}|^{\frac{n+\nu-d}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Lambda^{-1}\Sigma^{-1})\right\} \\ \times \int \exp\left\{-\frac{1}{2}\left[k(\mu-\theta)^{\mathrm{T}}\Sigma^{-1}(\mu-\theta)+ \right.\right. \\ \left.\left. \sum_{i=1}^{n}(y_i-\mu)^{\mathrm{T}}\Sigma^{-1}(y_i-\mu)\right]\right\} d\mu$$

- After some algebra,

$$\pi(\Sigma^{-1} \mid Y) \propto |\Sigma^{-1}|^{\frac{n+\nu-d-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\left(\Lambda^{-1}+S+\frac{kn}{k+n}(\theta-\bar{y})(\theta-\bar{y})^{\mathrm{T}}\right)\Sigma^{-1}\right]\right.$$

where

$$S=\sum_i (y_i-\bar{y})(y_i-\bar{y})^{\mathrm{T}}.$$

This implies that

$$\Sigma^{-1} \mid Y \sim W(\nu_p, \Lambda_p),$$

$$\nu_p = n+\nu, \qquad \Lambda_p = \left(\Lambda^{-1}+S+\frac{kn}{k+n}(\theta-\bar{y})(\theta-\bar{y})^{\mathrm{T}}\right)^{-1}.$$

Full conditional of μ

- The full conditional of μ given y and Σ^{-1} is given by

$$\pi(\mu \mid Y, \Sigma^{-1}) \propto \exp\left\{-\frac{1}{2}\left[\mu^{\mathrm{T}}(k\Sigma^{-1} + n\Sigma^{-1})\mu - 2\mu^{\mathrm{T}}(k\Sigma^{-1}\theta + n\Sigma^{-1}\bar{y})\right]\right\}.$$

- This implies that

$$[\mu \mid Y, \Sigma^{-1}] \sim \mathrm{MVN}\left(\frac{k}{k+n}\theta + \frac{n}{k+n}\bar{y}, \frac{\Sigma}{k+n}\right).$$

- How about the marginal posterior distribution of μ?

$$\mu \mid Y \sim t_{\nu+n-d+1}\left(\frac{k\theta + n\bar{y}}{k+n}, \frac{\Lambda_p^{-1}}{(k+n)(\nu+n-d+1)}\right).$$

- What is the posterior predictive distribution: $\pi(\tilde{y} \mid Y)$?

$$\tilde{y} \mid Y \sim t_{\nu+n-d+1}\left\{\frac{k\theta + n\bar{y}}{k+n}, \left(1 + \frac{1}{k+n}\right)\frac{\Lambda_p^{-1}}{(\nu+n-d+1)}\right\}.$$

Uninformative prior for (μ, Σ)

- Let $k \to 0$, $\nu \to -1$ and $\Lambda^{-1} \to 0$, then we have

$$\pi(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}.$$

- The corresponding posterior distribution is given by

$$\Sigma^{-1} \mid Y \sim W(n-1, S^{-1}), \qquad \mu \mid Y, \Sigma^{-1} \sim \mathrm{MVN}\left(\bar{y}, n^{-1}\Sigma\right).$$

- How about the marginal posterior distribution of μ?

$$\mu \mid Y \sim t_{n-d}\left\{\bar{y}, \frac{S}{n(n-d)}\right\}.$$

- What is the posterior predictive distribution: $\pi(\tilde{y} \mid Y)$?

$$\mu \mid Y \sim t_{n-d}\left\{\bar{y}, \left(1 + \frac{1}{n}\right)\frac{S}{n-d}\right\}.$$

Example: Reading comprehension (Hoff 2009)

- A sample of twenty-two children are given reading comprehension tests before and after receiving a particular instructional method. Each student i will then have two scores: $Y_{i,1}$ and $Y_{i,2}$ denoting the pre- and post- instructional scores respectively. We denote each student's pair of scores as a 2×1 vector y_i, so that

$$y_i = \begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} = \begin{pmatrix} \text{score on first test} \\ \text{score on second test} \end{pmatrix}$$

  Things we may be interested in include the population mean θ and the population covariance Σ.

$$E(y_i) = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \mathrm{Cov}(y_i) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

- Having information about θ and Σ may help us assess the effectiveness of the teaching method, possibly evaluated with $\theta_2 - \theta_1$, or the consistency of the reading comprehension test, which could be evaluated with the correlation coefficient $\rho_{1,2} = \sigma_{1,2}/(\sigma_1\sigma_2)$

Example:Reading comprehension (Hoff 2009)

- We have $n = 22$, $\bar{y} = (47.18, 53.86)^{\mathrm{T}}$, sample variances are $s_1^2 = 182.16$, $s_2^2 = 243.65$ and sample correlation is $s_{1,2}/(s_1 s_2) = 0.70$.
- Let's consider the uninformative prior for (μ, Σ) then

$$\Sigma \mid Y \sim W^{-1}(n-1, S).$$

$$\mu \mid \Sigma, Y \sim \mathrm{MVN}(\bar{y}, n^{-1}\Sigma).$$

- What are the posterior probabilities for the following events:
  - The average score on the second exam is higher than that on the first.

$$\Pr(\theta_2 > \theta_1 \mid y_1, \ldots, y_n) \approx 0.99$$

  - A student will get lower score on the second

$$\Pr(\tilde{y}_2 > \tilde{y}_1 \mid y_1, \ldots, y_n) \approx 0.71$$

  - The correlation coefficient $\rho_{1,2}$ is greater than 0.5.

$$\Pr(\rho_{1,2} > 0.5 \mid \tilde{y}_1, \ldots, y_n) \approx 0.93$$