STA 630: Bayesian Inference - Chapter 5

Zeya Wang

University of Kentucky

Spring 2025

Lecture 10: Multivariate Normal Model

- Multivariate normal and Inverse Wishart distributions
- Conjugate prior
- Semi-conjugatee prior
- Predictive distribution
- Bivariate case
- Example from Hoff

Data type: Multiple measurements for each experimental unit.
Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k)^T$ be a k-dimensional vector following a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$ and covariance matrix $\boldsymbol{\Sigma}$.

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

$$\mu_j = E\left[Y_j\right] \quad \sigma_{jj}^2 = \mathrm{Var}\left(Y_j\right) \quad \sigma_{jl} = \mathrm{Cov}\left(Y_j, Y_l\right)$$

Multivariate Normal Distribution (Hoff Chapter 7)

- When $k = 2$, $\boldsymbol{Y} = (Y_1, Y_2)^T$, we have a bivariate normal distribution with covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

and correlation $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$.

$$f(\boldsymbol{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-1} \left(\sigma_1 \sigma_2 \sqrt{1 - \rho^2} \right)^{-1}$$
$$\times \exp\left\{ -\frac{1}{2\left(1-\rho^2\right)} \left[\left(\frac{y_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{y_1 - \mu_1}{\sigma_1} \right) \left(\frac{y_2 - \mu_2}{\sigma_2} \right) + \left(\frac{y_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

- If $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{iid}{\sim} \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, their joint density is given by

$$\begin{aligned} f\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = (2\pi)^{-nk/2} |\boldsymbol{\Sigma}|^{-n/2} \\ \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\mu}\right) \right\} \end{aligned}$$

Wishart Distribution

- Recall that when $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(0,1)$, then $\chi^2 = \sum_{i=1}^{n} Y_i^2$ has a chi-squared distribution with n degrees of freedom

$$f\left(\chi^2\right) = \frac{2^{-n/2}}{\Gamma(n/2)} \left(\chi^2\right)^{(n-2)/2} \exp\left(-\chi^2/2\right), \quad \chi^2 > 0$$

- If $\boldsymbol{Y}_i \overset{iid}{\sim} \mathcal{N}_k\left(\mathbf{0}, \boldsymbol{S}^{-1}\right)$ $(i = 1, \ldots, n)$, then the symmetric positive definite matrix $\boldsymbol{\Phi} = \sum_{i=1}^{n} \boldsymbol{Y}_i \boldsymbol{Y}_i^T$ has a ***Wishart*** $\left(n, S^{-1}\right)$ distribution

$$f\left(\boldsymbol{\Phi} \mid \boldsymbol{S}^{-1}, n\right) = \frac{|\boldsymbol{\Phi}|^{(n-k-1)/2} \left|\boldsymbol{S}^{-1}\right|^{-n/2} \exp\left\{-\frac{1}{2} \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Phi})\right\}}{2^{nk/2} \pi^{k(k-1)/4} \prod_{j=1}^{k} \Gamma\left(\frac{n+1-j}{2}\right)}$$

$$E[\boldsymbol{\Phi}] = n\boldsymbol{S}^{-1}$$

Inverse-Wishart Distribution

If $\boldsymbol{\Phi} \sim \text{Wishart}\left(n, \boldsymbol{S}^{-1}\right)$ then $\boldsymbol{W} = \boldsymbol{\Phi}^{-1} \sim \text{inverse-Wishart}\left(n, \boldsymbol{S}^{-1}\right)$

$$f\left(\boldsymbol{W} \mid n, \boldsymbol{S}^{-1}\right) = \frac{|\boldsymbol{W}|^{-(n+k+1)/2}|\boldsymbol{S}|^{n/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{S}\boldsymbol{W}^{-1}\right)\right\}}{2^{nk/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma\left(\frac{n+1-j}{2}\right)}$$

$$E[\boldsymbol{W}] = E\left[\boldsymbol{\Phi}^{-1}\right] = \frac{1}{n-k-1}\boldsymbol{S}$$

The inverse-Wishart is the conjugate prior distribution for the multivariate normal covariance matrix.

Multivariate t Distribution

If $\boldsymbol{Y} \sim t(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, a multivariate t-distribution with degrees of freedom $\nu > 0$, location $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$, and a symmetric positive definite $k \times k$ scale matrix Σ

$$
\begin{aligned}
f(\boldsymbol{y} \mid \boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = & \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \nu^{k/2} \pi^{k/2}} |\boldsymbol{\Sigma}|^{-1/2} \\
& \times \left[1 + \frac{1}{\nu}(\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})\right]^{-(\nu+k)/2}
\end{aligned}
$$

Review of univariate normal with conjugate prior

Recall that for the univariate normal case, the conjugate prior density is given by

$$
\begin{aligned}
\mu \mid \sigma^2 &\sim \mathcal{N}\left(\mu_0, \sigma^2/\kappa_0\right) \\
\sigma^2 &\sim \text{Inv-Gamma}\left(\nu_0/2, \sigma_0^2/2\right) \equiv \frac{1}{\sigma^2} \sim \text{Gamma}\left(\nu_0/2, \sigma_0^2/2\right)
\end{aligned}
$$

which corresponds to the joint prior density

$$
p\left(\mu, \sigma^2\right) \propto \left(\sigma^2\right)^{-\left(\frac{\nu_0+1}{2}+1\right)} \exp\left\{-\frac{\kappa_0}{2\sigma^2}\left(\frac{\sigma_0^2}{\kappa_0} + (\mu-\mu_0)^2\right)\right\}
$$

which we refer to as the Normal-Inverse Gamma $\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right)$ density.

Multivariate normal with conjugate prior

$$\boldsymbol{\mu} \,\Big|\, \Sigma \sim \mathcal{N}_k\left(\mu_0, \tfrac{1}{\kappa_0}\Sigma\right)$$

$$\Sigma \sim \text{Inv-Wishart}\left(\nu_0, \boldsymbol{\Phi}_0^{-1}\right) \quad \text{or equivalently } \Sigma^{-1} \sim \text{Wishart}\left(\nu_0, \boldsymbol{\Phi}_0^{-1}\right)$$

which corresponds to the joint prior density

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\left(\frac{\nu_0+k}{2}+1\right)} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Phi}_0\boldsymbol{\Sigma}^{-1}\right) - \frac{\kappa_0}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right\}$$

which we refer to as the Normal-Inverse Wishart $\left(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Phi}_0, \nu_0, \boldsymbol{\Phi}_0\right)$ density.

The joint posterior distribution, $p\left(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\right)$ is then given by

$$f\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\right) \propto |\boldsymbol{\Sigma}|^{-\left(\frac{\nu_n+k}{2}+1\right)} \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Phi}_n \boldsymbol{\Sigma}^{-1}\right) - \frac{\kappa_n}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_n\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right\}$$

i.e., $\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim$ Normal-Inverse Wishart $\left(\boldsymbol{\mu}_n, \frac{1}{\kappa_n}\boldsymbol{\Phi}_n, \nu_n, \boldsymbol{\Phi}_n\right)$, where

$$\begin{aligned}
\boldsymbol{\mu}_n &= \frac{\kappa_0}{\kappa_0+n}\boldsymbol{\mu}_0 + \frac{n}{\kappa_0+n}\overline{\boldsymbol{y}} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\boldsymbol{\Phi}_n &= \boldsymbol{\Phi}_0 + \sum_i \left(\boldsymbol{y}_i - \overline{\boldsymbol{y}}\right)\left(\boldsymbol{y}_i - \overline{\boldsymbol{y}}\right)^T + \frac{\kappa_0 n}{\kappa_0+n}\left(\overline{\boldsymbol{y}}-\boldsymbol{\mu}_0\right)\left(\overline{\boldsymbol{y}}-\boldsymbol{\mu}_0\right)^T
\end{aligned}$$

The marginal posterior distribution, $f\left(\boldsymbol{\mu} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\right)$

$$\boldsymbol{\mu} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim t\left(\nu_n - k + 1, \boldsymbol{\mu}_n, \frac{1}{\kappa_n\left(\nu_n - k + 1\right)}\boldsymbol{\Phi}_n\right)$$

The marginal posterior distribution, $f\left(\boldsymbol{\Sigma} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\right)$

$$\boldsymbol{\Sigma} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim \text{Inv-Wishart}\left(\nu_n - 1, \boldsymbol{\Phi}_n^{-1}\right)$$

The posterior predictive distribution for a future observation

$$z \Big| y \sim t\left(\nu_n - k + 1, \mu_n, \frac{\Phi_n + \kappa_n + 1}{\kappa_n\left(\nu_n - k + 1\right)}\right)$$

Non-informative Priors

- Multivariate Jeffreys prior

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(k+1)/2}$$

- limit of the conjugate prior for $\kappa_0 \to \infty, |\Phi_0| \to 0$ and $\nu_0 \to -1$
- $\Sigma \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim IW(n-1, S)$ with $S = \sum_i (\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$
- $\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim N(\overline{\boldsymbol{y}}, \Sigma/n)$

Multivariate normal with semi-conjugate prior

$$\boldsymbol{\mu} \sim \mathcal{N}_k\left(\boldsymbol{\mu}_0, \Delta_0\right)$$
$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}\left(\nu_0, \boldsymbol{S}_0^{-1}\right)$$

- In this case, the posterior density does not follow a standard parametric form.
- Gibbs sampling can be used to obtain posterior samples of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The full conditionals are given by:

$$\boldsymbol{\mu} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{\Sigma} \sim \mathcal{N}_k \left(\boldsymbol{\mu}_n, \boldsymbol{\Delta}_n\right)$$
$$\Sigma \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{\mu} \sim \text{Inv-Wishart}\left(\nu_n, \boldsymbol{S}_n^{-1}\right)$$

where

$$\boldsymbol{\Delta}_n = \left(\boldsymbol{\Delta}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}$$
$$\boldsymbol{\mu}_n = \boldsymbol{\Delta}_n \left(\boldsymbol{\Delta}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{y}}\right)$$
$$\boldsymbol{S}_n = \boldsymbol{S}_0 + \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)^T$$

Example: Reading comprehension

Twenty-two children are given a reading comprehension test before and after receiving a particular instruction method.

- $Y_{i,1}$: pre-instructional score for student i
- $Y_{i,2}$: post-instructional score for student i
- vector of observations for each student $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^T$

Questions:

- Does the instructional method lead to improvements in reading comprehension, on average?
- If so, by how much?
- Can improvements be predicted based on the first test?

Example: Reading comprehension

- We could model the data as bivariate normal $\boldsymbol{Y}_i \overset{iid}{\sim} \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- For Bayesian inference, we need to specify prior distributions and use Bayes' theorem to obtain posterior distributions.
- Let us specify semi-conjugate priors

$$\boldsymbol{\mu} \sim \mathcal{N}_2\left(\boldsymbol{\mu}_0, \boldsymbol{\Delta}_0\right)$$
$$\boldsymbol{\Sigma} \sim \mathrm{Inv}-\mathrm{Wishart}\left(\nu_0, \boldsymbol{S}_0^{-1}\right)$$

- The test is designed to have a mean of 50 - set $\boldsymbol{\mu}_0 = (50, 50)^T$
- The true mean is constrained to be between 0 and 100 , i.e., $\mu_j \pm 2\sigma_j = (0, 100)$, which implies $\sigma_j^2 = (50/2)^2 = 625$.
- We assume a prior correlation of 0.5 , so $\sigma_{12} = 0.5 \times 625 = 312.5$.
- Recall that the mean of the inverse-Wishart is $\frac{1}{\nu-k-1}\boldsymbol{S}_0$, so setting $\nu_0 = k+2 = 4$ leads to $E[\Sigma] = \boldsymbol{S}_0$.
- Loosely centered at the same covariance as in the normal mean prior distribution.

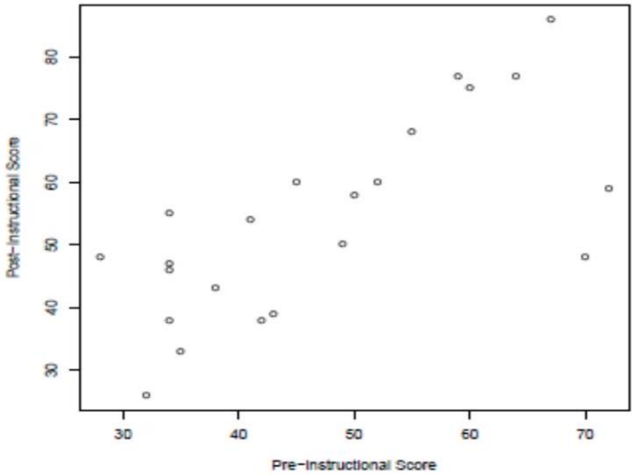
$$\boldsymbol{\mu} \sim N\left((50, 50)^T, \mathbf{S}_0\right), \quad \boldsymbol{\Sigma} \sim IW\left(4, \mathbf{S}_0^{-1}\right), \quad \mathbf{S}_0 = \begin{pmatrix} 625 & 312.5 \\ 312.5 & 625 \end{pmatrix}$$

```r
# install and call package with functions for
# Multivariate Normal & Wishart
install.packages("MCMCpack")
library(MCMCpack)

# read in and plot data
Y = dget("Hoff/Data/Y.reading.dat")
plot(Y, xlab="Pre-instructional Score", ylab="Post-instructional Score")

# set hyperparameters
mu0 = c(50,50); D0 = matrix(c(625,312.5,312.5,625),2,2)
nu0 = 4; S0 = matrix(c(625,312.5,312.5,625), 2, 2)


n = dim(Y)[1]; ybar = apply(Y,2,mean); Sigma = cov(Y)
MU = SIGMA= NULL; YS = NULL
```



Post−Instructional Score
30
40
50
60
70
80
Pre−Instructional Score
30
40
50
60
70

```
set.seed(1); T = 10000
for(t in 1:T) {
    ###update mu
    Dn = solve( solve(D0) + n*solve(Sigma) )
    mun = Dn%*%(solve(D0)%*%mu0 + n*solve(Sigma)%*%ybar)
    mu = mvrnorm(1, mun, Dn)
    ###

    #update Sigma
    Sn=S0+(t(Y)-c(mu)) %*% t(t(Y)-c(mu))
    Sigma = riwish(n+nu0, Sn)

    ###
    YS = rbind(YS, mvrnorm(1,mu,Sigma))

    MU=rbind(MU,mu); SIGMA = rbind(SIGMA, c(Sigma))
    }
```

```
nburn = 1000
# 95% credible intervals for rho and mu2-mu1
> quantile(SIGMA[(nburn+1):T,2]/sqrt(SIGMA[(nburn+1):T,1]),
        prob=c(.025,.5,.975) )
        2.5% 50% 97.5%
0.4048507 0.6877680 0.8476522
> quantile(MU[(nburn+1):T,2]-MU[(n.burn+1):T,1], prob=c(.025,.5,.975))
    2.5% 50% 97.5%
    1.391037 6.611074 11.817720

> mean( MU[(nburn+1):T,2]-MU[(nburn+1):T,1])
[1] 6.605529
> mean( MU[(nburn+1):T,2]>MU[(nburn+1):T,1])
[1] 0.9921111
> mean(YS[(nburn+1):T,2] > YS[(nburn+1):T,1])
[1] 0.7017778
```

```r
# Plots as in Fig. 7.2 of textbook
source("Hoff/Code/hdr2d.r")

install.packages("ash")
library(ash)

plot.hdr2d(MU,xlab=expression(mu[1]),ylab=expression(mu[2]))
abline(0,1)
```

Highest posterior density contour for (μ_1, μ_2)