

# STA 360/602L: MODULE 4.2

## MULTIVARIATE NORMAL MODEL II

DR. OLANREWaju MICHAEL AKANDE

# MULTIVARIATE NORMAL LIKELIHOOD RECAP

- For data  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})^T \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$ , the likelihood is

$$p(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\mathbf{y}_i - \boldsymbol{\theta}) \right\}.$$

- For  $\boldsymbol{\theta}$ , it is convenient to write  $p(\mathbf{Y}|\boldsymbol{\theta}, \Sigma)$  as

$$p(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T (n\Sigma^{-1}) \boldsymbol{\theta} + \boldsymbol{\theta}^T (n\Sigma^{-1} \bar{\mathbf{y}}) \right\},$$

where  $\bar{\mathbf{y}} = (\bar{y}_1, \dots, \bar{y}_p)^T$ .

- For  $\Sigma$ , it is convenient to write  $p(\mathbf{Y}|\boldsymbol{\theta}, \Sigma)$  as

$$p(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\mathbf{S}_\theta \Sigma^{-1}] \right\},$$

where  $\mathbf{S}_\theta = \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\theta})(\mathbf{y}_i - \boldsymbol{\theta})^T$  is the residual sum of squares matrix.

# PRIOR FOR THE MEAN

- A convenient specification of the joint prior is  $\pi(\boldsymbol{\theta}, \Sigma) = \pi(\boldsymbol{\theta})\pi(\Sigma)$ .
- As in the univariate case, a convenient prior distribution for  $\boldsymbol{\theta}$  is also normal (multivariate in this case).
- Assume that  $\pi(\boldsymbol{\theta}) = \mathcal{N}_p(\boldsymbol{\mu}_0, \Lambda_0)$ .
- The pdf will be easier to work with if we write it as

$$\begin{aligned}\pi(\boldsymbol{\theta}) &= (2\pi)^{-\frac{p}{2}} |\Lambda_0|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)^T \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0) \right\} \\ &\propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)^T \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0) \right\} \\ &= \exp \left\{ -\frac{1}{2} \left[ \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} - \underbrace{\boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^T \Lambda_0^{-1} \boldsymbol{\theta}}_{\text{same term}} + \underbrace{\boldsymbol{\mu}_0^T \Lambda_0^{-1} \boldsymbol{\mu}_0}_{\text{does not involve } \boldsymbol{\theta}} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} [\boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0] \right\} \\ &= \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right\}\end{aligned}$$

# PRIOR FOR THE MEAN

- So we have

$$\pi(\boldsymbol{\theta}) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right\}.$$

- **Key trick for combining with likelihood:** When the normal density is written in this form, note the following details in the exponent.
  - In the first part, the inverse of the *covariance matrix*  $\Lambda_0^{-1}$  is "sandwiched" between  $\boldsymbol{\theta}^T$  and  $\boldsymbol{\theta}$ .
  - In the second part, the  $\boldsymbol{\theta}$  in the first part is replaced (sort of) with the *mean*  $\boldsymbol{\mu}_0$ , with  $\Lambda_0^{-1}$  keeping its place.
- The two points above will help us identify **updated means** and **updated covariance matrices** relatively quickly.

# CONDITIONAL POSTERIOR FOR THE MEAN

- Our conditional posterior (full conditional)  $\theta|\Sigma, \mathbf{Y}$ , is then

$$\pi(\theta|\Sigma, \mathbf{Y}) \propto p(\mathbf{Y}|\theta, \Sigma) \cdot \pi(\theta)$$

$$\propto \underbrace{\exp \left\{ -\frac{1}{2} \theta^T (n\Sigma^{-1}) \theta + \theta^T (n\Sigma^{-1} \bar{\mathbf{y}}) \right\}}_{p(\mathbf{Y}|\theta, \Sigma)} \cdot \underbrace{\exp \left\{ -\frac{1}{2} \theta^T \Lambda_0^{-1} \theta + \theta^T \Lambda_0^{-1} \mu_0 \right\}}_{\pi(\theta)}$$

$$= \exp \left\{ \underbrace{-\frac{1}{2} \theta^T (n\Sigma^{-1}) \theta - \frac{1}{2} \theta^T \Lambda_0^{-1} \theta}_{\text{First parts from } p(\mathbf{Y}|\theta, \Sigma) \text{ and } \pi(\theta)} + \underbrace{\theta^T (n\Sigma^{-1} \bar{\mathbf{y}}) + \theta^T \Lambda_0^{-1} \mu_0}_{\text{Second parts from } p(\mathbf{Y}|\theta, \Sigma) \text{ and } \pi(\theta)} \right\}$$

$$= \exp \left\{ -\frac{1}{2} \theta^T [n\Sigma^{-1} + \Lambda_0^{-1}] \theta + \theta^T [n\Sigma^{-1} \bar{\mathbf{y}} + \Lambda_0^{-1} \mu_0] \right\},$$

which is just another multivariate normal distribution.

# CONDITIONAL POSTERIOR FOR THE MEAN

- To confirm the normal density and its parameters, compare to the prior kernel

$$\pi(\boldsymbol{\theta}) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right\}$$

and the posterior kernel we just derived, that is,

$$\pi(\boldsymbol{\theta} | \Sigma, \mathbf{Y}) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T [\Lambda_0^{-1} + n\Sigma^{-1}] \boldsymbol{\theta} + \boldsymbol{\theta}^T [\Lambda_0^{-1} \boldsymbol{\mu}_0 + n\Sigma^{-1} \bar{\mathbf{y}}] \right\}.$$

- Easy to see (relatively) that  $\boldsymbol{\theta} | \Sigma, \mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}_n, \Lambda_n)$ , with

$$\Lambda_n = [\Lambda_0^{-1} + n\Sigma^{-1}]^{-1}$$

and

$$\boldsymbol{\mu}_n = \Lambda_n [\Lambda_0^{-1} \boldsymbol{\mu}_0 + n\Sigma^{-1} \bar{\mathbf{y}}]$$

# BAYESIAN INFERENCE

- As in the univariate case, we once again have that
  - Posterior precision is sum of prior precision and data precision:

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

- Posterior expectation is weighted average of prior expectation and the sample mean:

$$\mu_n = \Lambda_n [\Lambda_0^{-1} \mu_0 + n\Sigma^{-1} \bar{y}]$$

$$= \overbrace{[\Lambda_n \Lambda_0^{-1}]}^{\text{weight on prior mean}} \underbrace{\mu_0}_{\text{prior mean}} + \overbrace{[\Lambda_n (n\Sigma^{-1})]}^{\text{weight on sample mean}} \underbrace{\bar{y}}_{\text{sample mean}}$$

- Compare these to the results from the univariate case to gain more intuition.

# WHAT ABOUT THE COVARIANCE MATRIX?

- In the univariate case with  $y_i \sim \mathcal{N}(\mu, \sigma^2)$ , the common choice for the prior is an inverse-gamma distribution for the variance  $\sigma^2$ .
- As we have seen, we can rewrite as  $y_i \sim \mathcal{N}(\mu, \tau^{-1})$ , so that we have a gamma prior for the precision  $\tau$ .
- In the multivariate normal case, we have a covariance matrix  $\Sigma$  instead of a scalar.
- Appealing to have a matrix-valued extension of the inverse-gamma (and gamma) that would be conjugate.
- One complication is that the covariance matrix  $\Sigma$  must be **positive definite and symmetric**.



# POSITIVE DEFINITE AND SYMMETRIC

- "Positive definite" means that for all  $x \in \mathcal{R}^p$ ,  $x^T \Sigma x > 0$ .
- Basically ensures that the diagonal elements of  $\Sigma$  (corresponding to the marginal variances) are positive.
- Also, ensures that the correlation coefficients for each pair of variables are between -1 and 1.
- Our prior for  $\Sigma$  should thus assign probability one to set of positive definite matrices.
- Analogous to the univariate case, the **inverse-Wishart distribution** is the corresponding conditionally conjugate prior for  $\Sigma$  (multivariate generalization of the inverse-gamma).
- The textbook covers the construction of Wishart and inverse-Wishart random variables. We will skip the actual development in class but will write code to sample random variates.

# INVERSE-WISHART DISTRIBUTION

- A random variable  $\Sigma \sim \text{IW}_p(\nu_0, \mathbf{S}_0)$ , where  $\Sigma$  is positive definite and  $p \times p$ , has pdf

$$p(\Sigma) \propto |\Sigma|^{\frac{-(\nu_0+p+1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S}_0 \Sigma^{-1}) \right\},$$

where

- $\nu_0 > p - 1$  is the "degrees of freedom", and
- $\mathbf{S}_0$  is a  $p \times p$  positive definite matrix.
- For this distribution,  $\mathbb{E}[\Sigma] = \frac{1}{\nu_0 - p - 1} \mathbf{S}_0$ , for  $\nu_0 > p + 1$ .
- Hence,  $\mathbf{S}_0$  is the scaled mean of the  $\text{IW}_p(\nu_0, \mathbf{S}_0)$ .

# WISHART DISTRIBUTION

- If we are very confidence in a prior guess  $\Sigma_0$ , for  $\Sigma$ , then we might set
  - $\nu_0$ , the degrees of freedom to be very large, and
  - $S_0 = (\nu_0 - p - 1)\Sigma_0$ .

In this case,  $\mathbb{E}[\Sigma] = \frac{1}{\nu_0 - p - 1} S_0 = \frac{1}{\nu_0 - p - 1} (\nu_0 - p - 1) \Sigma_0 = \Sigma_0$ ,  
and  $\Sigma$  is tightly (depending on the value of  $\nu_0$ ) centered around  $\Sigma_0$ .

- If we are not at all confident but we still have a prior guess  $\Sigma_0$ , we might set
  - $\nu_0 = p + 2$ , so that the  $\mathbb{E}[\Sigma] = \frac{1}{\nu_0 - p - 1} S_0$  is finite.
  - $S_0 = \Sigma_0$

Here,  $\mathbb{E}[\Sigma] = \Sigma_0$  as before, but  $\Sigma$  is only loosely centered around  $\Sigma_0$ .

# WISHART DISTRIBUTION

- Just as we had with the gamma and inverse-gamma relationship in the univariate case, we can also work in terms of the **Wishart distribution** (multivariate generalization of the gamma) instead.
- The **Wishart distribution** provides a conditionally-conjugate prior for the precision matrix  $\Sigma^{-1}$  in a multivariate normal model.
- Specifically, if  $\Sigma \sim IW_p(\nu_0, \mathbf{S}_0)$ , then  $\Phi = \Sigma^{-1} \sim W_p(\nu_0, \mathbf{S}_0^{-1})$ .
- A random variable  $\Phi \sim W_p(\nu_0, \mathbf{S}_0^{-1})$ , where  $\Phi$  has dimension  $(p \times p)$ , has pdf

$$f(\Phi) \propto |\Phi|^{\frac{\nu_0 - p - 1}{2}} \exp \left\{ -\frac{1}{2} \text{tr}(\mathbf{S}_0 \Phi) \right\}.$$

- Here,  $\mathbb{E}[\Phi] = \nu_0 \mathbf{S}_0$ .
- Note that the textbook writes the inverse-Wishart as  $IW_p(\nu_0, \mathbf{S}_0^{-1})$ . I prefer  $IW_p(\nu_0, \mathbf{S}_0)$  instead. Feel free to use either notation but try not to get confused.

# CONDITIONAL POSTERIOR FOR COVARIANCE

- Assuming  $\pi(\Sigma) = \text{IW}_p(\nu_0, \mathbf{S}_0)$ , the conditional posterior (full conditional)  $\Sigma|\boldsymbol{\theta}, \mathbf{Y}$ , is then

$$\begin{aligned}\pi(\Sigma|\boldsymbol{\theta}, \mathbf{Y}) &\propto p(\mathbf{Y}|\boldsymbol{\theta}, \Sigma) \cdot \pi(\boldsymbol{\theta}) \\ &\propto \underbrace{|\Sigma|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\mathbf{S}_\theta \Sigma^{-1}] \right\}}_{p(\mathbf{Y}|\boldsymbol{\theta}, \Sigma)} \cdot \underbrace{|\Sigma|^{\frac{-(\nu_0+p+1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr} (\mathbf{S}_0 \Sigma^{-1}) \right\}}_{\pi(\boldsymbol{\theta})} \\ &\propto |\Sigma|^{\frac{-(\nu_0+p+n+1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\mathbf{S}_0 \Sigma^{-1} + \mathbf{S}_\theta \Sigma^{-1}] \right\}, \\ &\propto |\Sigma|^{\frac{-(\nu_0+n+p+1)}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [(\mathbf{S}_0 + \mathbf{S}_\theta) \Sigma^{-1}] \right\},\end{aligned}$$

which is  $\text{IW}_p(\nu_n, \mathbf{S}_n)$ , or using the notation in the book,  $\text{IW}_p(\nu_n, \mathbf{S}_n^{-1})$ , with

- $\nu_n = \nu_0 + n$ , and
- $\mathbf{S}_n = [\mathbf{S}_0 + \mathbf{S}_\theta]$

# CONDITIONAL POSTERIOR FOR COVARIANCE

- We once again see that the "posterior sample size" or "posterior degrees of freedom"  $\nu_n$  is the sum of the "prior degrees of freedom"  $\nu_0$  and the data sample size  $n$ .
- $S_n$  can be thought of as the "posterior sum of squares", which is the sum of "prior sum of squares" plus "sample sum of squares".
- Recall that if  $\Sigma \sim \text{IW}_p(\nu_0, S_0)$ , then  $\mathbb{E}[\Sigma] = \frac{1}{\nu_0 - p - 1} S_0$ .
- $\Rightarrow$  the conditional posterior expectation of the population covariance is

$$\begin{aligned} \mathbb{E}[\Sigma | \theta, \mathbf{Y}] &= \frac{1}{\nu_0 + n - p - 1} [S_0 + S_\theta] \\ &= \underbrace{\frac{\nu_0 - p - 1}{\nu_0 + n - p - 1}}_{\text{weight on prior expectation}} \underbrace{\left[ \frac{1}{\nu_0 - p - 1} S_0 \right]}_{\text{prior expectation}} + \underbrace{\frac{n}{\nu_0 + n - p - 1}}_{\text{weight on sample estimate}} \underbrace{\left[ \frac{1}{n} S_\theta \right]}_{\text{sample estimate}}, \end{aligned}$$

which is a weighted average of prior expectation and sample estimate.

# WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!