STA 360/602L: Module 4.2

Multivariate normal model II

Dr. Olanrewaju Michael Akande

Multivariate normal likelihood recap

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

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

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

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

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

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

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

  where $\boldsymbol{S}_\theta = \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T$ 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

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

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* Λ_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 Λ_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) $\boldsymbol{\theta}|\Sigma, \boldsymbol{Y}$, is then

$$
\begin{aligned}
\pi(\boldsymbol{\theta}|\Sigma, \boldsymbol{Y}) &\propto p(\boldsymbol{Y}|\boldsymbol{\theta}, \Sigma) \cdot \pi(\boldsymbol{\theta}) \\
&\propto \underbrace{\exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T(n\Sigma^{-1})\boldsymbol{\theta} + \boldsymbol{\theta}^T(n\Sigma^{-1}\bar{\boldsymbol{y}})\right\}}_{p(\boldsymbol{Y}|\boldsymbol{\theta},\Sigma)} \cdot \underbrace{\exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right\}}_{\pi(\boldsymbol{\theta})} \\
&= \exp\left\{\underbrace{-\frac{1}{2}\boldsymbol{\theta}^T(n\Sigma^{-1})\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta}}_{\text{First parts from } p(\boldsymbol{Y}|\boldsymbol{\theta},\Sigma) \text{ and } \pi(\boldsymbol{\theta})} + \underbrace{\boldsymbol{\theta}^T(n\Sigma^{-1}\bar{\boldsymbol{y}}) + \boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0}_{\text{Second parts from } p(\boldsymbol{Y}|\boldsymbol{\theta},\Sigma) \text{ and } \pi(\boldsymbol{\theta})}\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\left[n\Sigma^{-1} + \Lambda_0^{-1}\right]\boldsymbol{\theta} + \boldsymbol{\theta}^T\left[n\Sigma^{-1}\bar{\boldsymbol{y}} + \Lambda_0^{-1}\boldsymbol{\mu}_0\right]\right\},
\end{aligned}
$$

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,

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

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

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

and

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

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:

$$\begin{aligned}\boldsymbol{\mu}_n &= \Lambda_n \left[\Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{y}}\right] \\ &= \overbrace{\left[\Lambda_n\Lambda_0^{-1}\right]}^{\text{weight on prior mean}} \underbrace{\boldsymbol{\mu}_0}_{\text{prior mean}} + \overbrace{\left[\Lambda_n(n\Sigma^{-1})\right]}^{\text{weight on sample mean}} \underbrace{\bar{\boldsymbol{y}}}_{\text{sample mean}}\end{aligned}$$

- 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 σ^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 τ.
- In the multivariate normal case, we have a covariance matrix Σ 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 Σ 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 Σ (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 Σ 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 Σ (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 $\Sigma \sim \mathrm{IW}_p(\nu_0, \boldsymbol{S}_0)$, where Σ is positive definite and $p \times p$, has pdf

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

  where

  - $\nu_0 > p - 1$ is the "degrees of freedom", and
  - $\boldsymbol{S}_0$ is a $p \times p$ positive definite matrix.

- For this distribution, $\mathbb{E}[\Sigma] = \dfrac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$, for $\nu_0 > p + 1$.

- Hence, $\boldsymbol{S}_0$ is the scaled mean of the $\mathrm{IW}_p(\nu_0, \boldsymbol{S}_0)$.

Inverse-Wishart distribution

- If we are very confident in a prior guess Σ_0, for Σ, then we might set
  - ν_0, the degrees of freedom to be very large, and
  - $\boldsymbol{S}_0 = (\nu_0 - p - 1)\Sigma_0$.

  In this case, $\mathbb{E}[\Sigma] = \frac{1}{\nu_0 - p - 1}\boldsymbol{S}_0 = \frac{1}{\nu_0 - p - 1}(\nu_0 - p - 1)\Sigma_0 = \Sigma_0$, and Σ is tightly (depending on the value of ν_0) centered around Σ_0.

- If we are not at all confident but we still have a prior guess Σ_0, we might set
  - $\nu_0 = p + 2$, so that the $\mathbb{E}[\Sigma] = \frac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$ is finite.
  - $\boldsymbol{S}_0 = \Sigma_0$

  Here, $\mathbb{E}[\Sigma] = \Sigma_0$ as before, but Σ is only loosely centered around Σ_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 Σ^{-1} in a multivariate normal model.

- Specifically, if $\Sigma \sim \mathrm{IW}_p(\nu_0, \boldsymbol{S}_0)$, then $\Phi = \Sigma^{-1} \sim \mathrm{W}_p(\nu_0, \boldsymbol{S}_0^{-1})$.

- A random variable $\Phi \sim \mathrm{W}_p(\nu_0, \boldsymbol{S}_0^{-1})$, where Φ has dimension $(p \times p)$, has pdf

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

- Here, $\mathbb{E}[\Phi] = \nu_0 \boldsymbol{S}_0$.

- Note that the textbook writes the inverse-Wishart as $\mathrm{IW}_p(\nu_0, \boldsymbol{S}_0^{-1})$. I prefer $\mathrm{IW}_p(\nu_0, \boldsymbol{S}_0)$ instead. Feel free to use either notation but try not to get confused.

Conditional posterior for covariance

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

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

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

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

Conditional posterior for covariance

- We once again see that the "posterior sample size" or "posterior degrees of freedom" ν_n is the sum of the "prior degrees of freedom" ν_0 and the data sample size n.
- $\boldsymbol{S}_n$ 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 $\Sigma \sim \mathrm{IW}_p(\nu_0, \boldsymbol{S}_0)$, then $\mathbb{E}[\Sigma] = \frac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$.
- $\Rightarrow$ the conditional posterior expectation of the population covariance is

$$\begin{aligned}\mathbb{E}[\Sigma|\boldsymbol{\theta}, \boldsymbol{Y}] &= \frac{1}{\nu_0 + n - p - 1}[\boldsymbol{S}_0 + \boldsymbol{S}_\theta] \\ &= \underbrace{\frac{\nu_0 - p - 1}{\nu_0 + n - p - 1}}_{\text{weight on prior expectation}} \overbrace{\left[\frac{1}{\nu_0 - p - 1}\boldsymbol{S}_0\right]}^{\text{prior expectation}} + \underbrace{\frac{n}{\nu_0 + n - p - 1}}_{\text{weight on sample estimate}} \overbrace{\left[\frac{1}{n}\boldsymbol{S}_\theta\right]}^{\text{sample estimate}},\end{aligned}$$

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

What's next?

Move on to the readings for the next module!