Multivariate normal model

Dr. Olanrewaju Michael Akande

Feb 19, 2020

Announcements

- Take Survey I
- Link: https://duke.qualtrics.com/jfe/form/SV_54rrMwDxp3hmagt
- Responses are anonymized.

Outline

- Wrap up exercise from last class
- Multivariate normal/Gaussian model
  - Motivating example
  - Inference for mean
  - Inference for covariance

Recap of conditional distributions

- Partition $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$ as

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \sim \mathcal{N}_p \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right],$$

  where

  - $\boldsymbol{Y}_1$ and $\boldsymbol{\mu}_1$ are $q \times 1$,
  - $\boldsymbol{Y}_2$ and $\boldsymbol{\mu}_2$ are $(p-q) \times 1$,
  - Σ_{11} is $q \times q$, and
  - Σ_{22} is $(p-q) \times (p-q)$, with $\Sigma_{22} > 0$.

- Then,

$$\boldsymbol{Y}_1 | \boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim \mathcal{N}_q \left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right).$$

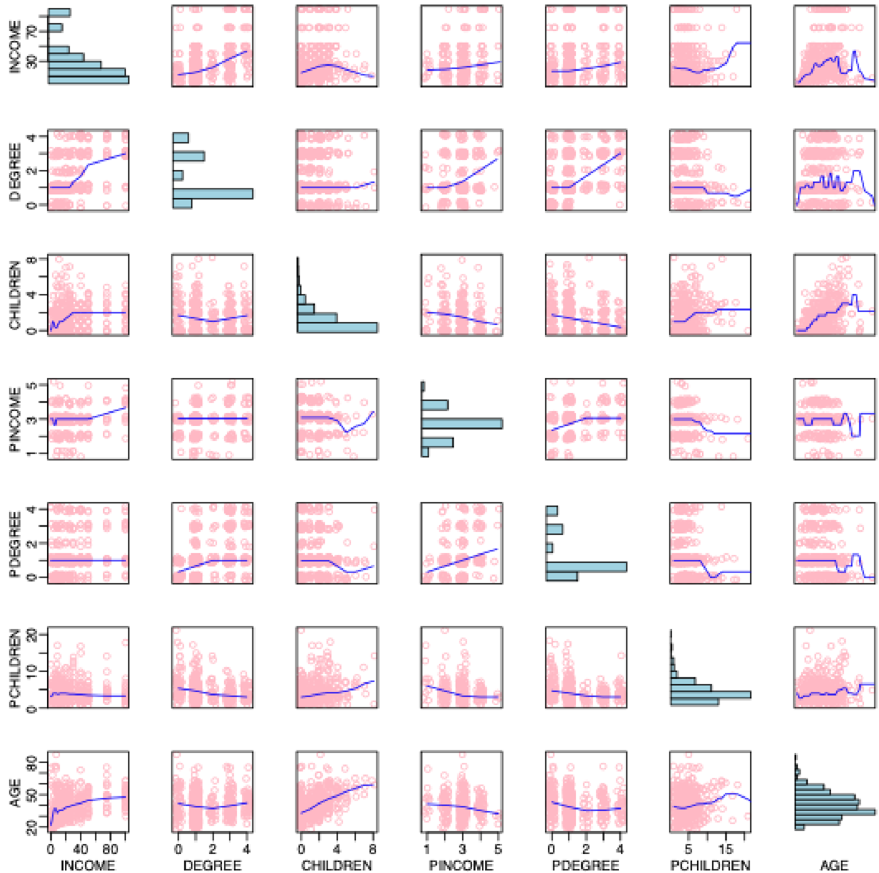
Working with normal distributions

- Three real (univariate) random quantities x, y and z have a joint normal distribution given by $p(x, y, z) = p(y|x)p(x|z)p(z)$.
- Suppose
  - $p(y|x) = \mathcal{N}(x, w)$ independently of z, for some known variance w;
  - $p(x|z) = \mathcal{N}(\theta z, v)$ for some known parameter θ, and known variance v; and
  - $p(z) = \mathcal{N}(m, M)$, with some known mean m, and known variance M.
- What is
  - $p(x)$? $p(y)$?
  - $p(x|y)$? $p(z|x)$?
- **To be done on the board.**

Multivariate data

- Survey data often yield multivariate data of varied types.
- **Typical survey data:** response vector $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{ip})^T$ for each person i in a sample of survey respondents, $i = 1, \ldots, n$. For example, we could have
  - $y_{i1} =$ income
  - $y_{i2} =$ level of education
  - $y_{i3} =$ number of children
  - $y_{i4} =$ age
  - $y_{i5} =$ attitude
- Interest is then often on inferring the potential associations among these variables.
- See *https://www.stat.washington.edu/people/pdhoff/public/coptalk.pdf*

GSS DATA



See *https://www.stat.washington.edu/people/pdhoff/public/coptalk.pdf*

Conditional models

- Interest is often in conditional relationships between pairs of variables, accounting for heterogeneity in other variables of less interest.
- Consider the following models.
- GSS data:
  - **Model 1**

    $$\mathrm{INC}_i = \beta_0 + \beta_1 \mathrm{CHILD}_i + \beta_2 \mathrm{DEG}_i + \beta_3 \mathrm{AGE}_i + \beta_4 \mathrm{PCHILD}_i + \beta_5 \mathrm{PINC}_i + \beta_6 \mathrm{PDEG}_i + \epsilon_i$$

    p-value for β_1 here is 0.11: "little evidence" that $\beta_1 \neq 0$.
  - **Model 2**

    $$\mathrm{CHILD}_i \sim \mathrm{Poisson}\left(\exp\left[\beta_0 + \beta_1 \mathrm{INC}_i + \beta_2 \mathrm{DEG}_i + \beta_3 \mathrm{AGE}_i + \beta_4 \mathrm{PCHILD}_i + \beta_5 \mathrm{PINC}_i + \beta_6 \mathrm{PDEG}_i\right]\right)$$

    p-value for β_1 here is 0.01: "strong evidence" that $\beta_1 \neq 0$.
- Not satisfactory; better to use multivariate models instead to do this jointly.
- See *https://www.stat.washington.edu/people/pdhoff/public/coptalk.pdf*

Multivariate normal distribution recap

- Recall that if $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$, then

$$f(\boldsymbol{y}) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\theta}) \right\}.$$

- $\boldsymbol{\theta}$ is the $p \times 1$ mean vector, that is, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$.
- Σ is the $p \times p$ **positive definite** covariance matrix, that is, $\Sigma = \{\sigma_{jk}\}$, where σ_{jk} denotes the covariance between Y_j and Y_k.
- For each $j = 1, \ldots, p$, $Y_j \sim \mathcal{N}(\theta_j, \sigma_{jj})$.
- How to do posterior inference if this is our sampling model?

Reading comprehension example

- Twenty-two children are given a reading comprehension test before and after receiving a particular instruction method.
  - Y_{i1}: pre-instructional score for student i.
  - Y_{i2}: post-instructional score for student i.
- Vector of observations for each student: $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^T$.
- Clearly, we should expect some correlation between Y_{i1} and Y_{i2}.

Reading comprehension example

- Questions of interest:
  - Do students improve in reading comprehension on average?
  - If so, by how much?
  - Can we predict post-test score from pre-test score?
  - If there is a "significant" improvement, does that mean the instructional method is good?
  - If we have students with missing pre-test scores, can we predict the scores?
- We will come back to this example. First, let's specify priors and see what the implied (conditional) posteriors look like.

Multivariate normal likelihood

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

$$\begin{aligned} L(\boldsymbol{Y}; \boldsymbol{\theta}, \Sigma) &= \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}) \right\} \\ &\propto |\Sigma|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}) \right\}. \end{aligned}$$

- It will be super useful to be able to write the likelihood in two different formulations depending on whether we about the posterior of $\boldsymbol{\theta}$ or Σ.

Multivariate normal likelihood

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

$$\begin{aligned}
L(\boldsymbol{Y};\boldsymbol{\theta},\Sigma) &\propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i^T-\boldsymbol{\theta}^T)\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{y}_i^T\Sigma^{-1}\boldsymbol{y}_i \underbrace{-\boldsymbol{y}_i^T\Sigma^{-1}\boldsymbol{\theta}-\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{y}_i}_{\text{same term}}+\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}-2\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{y}_i\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}-\frac{1}{2}\sum_{i=1}^{n}(-2)\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{y}_i\right\} \\
&= \exp\left\{-\frac{1}{2}n\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}^T\Sigma^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T(n\Sigma^{-1})\boldsymbol{\theta}+\boldsymbol{\theta}^T(n\Sigma^{-1}\bar{\boldsymbol{y}})\right\},
\end{aligned}$$

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

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 conjugate 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

$$
\begin{aligned}
\pi(\boldsymbol{\theta}) &= (2\pi)^{-\frac{p}{2}} |\Lambda_0|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^T \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0) \right\} \\
&\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^T \Lambda_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0) \right\} \\
&= \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} \underbrace{- \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_0^T \Lambda_0^{-1} \boldsymbol{\theta}}_{\text{same term}} + \boldsymbol{\mu}_0^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right] \right\} \\
&= \exp\left\{ -\frac{1}{2} \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right\}
\end{aligned}
$$

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

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

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:

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

  - 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.

Positive definite and symmetric

- One complication is that the covariance matrix Σ must be **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 $\Sigma \sim \text{IW}_p(\nu_0, \boldsymbol{S}_0)$, where Σ is positive definite and $p \times p$, has pdf

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

  where

  - $\text{tr}(\cdot)$ is the **trace function** (sum of diagonal elements),
  - $\nu_0 > p - 1$ 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 $\text{IW}_p(\nu_0, \boldsymbol{S}_0)$.

STA 602L

Wishart distribution

- If we are very confidence 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.

Back to inference on covariance

- For inference on $\boldsymbol{\Sigma}$, we need to rewrite the likelihood a bit to match the inverse-Wishart kernel.

- First a few results from matrix algebra:

  1. $\text{tr}(\boldsymbol{A}) = \sum_{j=1}^{p} a_{jj}$, where a_{jj} is the jth diagonal element of a square $p \times p$ matrix $\boldsymbol{A}$.
  2. Cyclic property:
     $$\text{tr}(\boldsymbol{ABC}) = \text{tr}(\boldsymbol{BCA}) = \text{tr}(\boldsymbol{CAB}),$$
     given that the product $\boldsymbol{ABC}$ is a square matrix.
  3. If $\boldsymbol{A}$ is a $p \times p$ matrix, then for a $p \times 1$ vector $\boldsymbol{x}$,
     $$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \text{tr}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})$$
     holds by (1), since $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ is a scalar.
  4. $\text{tr}(\boldsymbol{A} + \boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$.

Multivariate normal likelihood again

- It is thus convenient to rewrite $L(\boldsymbol{Y};\boldsymbol{\theta},\Sigma)$ as

$$
\begin{aligned}
L(\boldsymbol{Y};\boldsymbol{\theta},\Sigma) &\propto \underbrace{|\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right\}}_{\text{no algebra/change yet}} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\underbrace{\text{tr}\left[(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right]}_{\text{by result 3}}\right\} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\underbrace{\text{tr}\left[(\boldsymbol{y}_i-\boldsymbol{\theta})(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}\right]}_{\text{by cyclic property}}\right\} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{\underbrace{-\frac{1}{2}\text{tr}\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\theta})(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}\right]}_{\text{by result 4}}\right\} \\
&= |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\boldsymbol{S}_\theta\Sigma^{-1}\right]\right\},
\end{aligned}
$$

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.

Conditional posterior for covariance

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

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

which is $\mathrm{IW}_p(\nu_n, \boldsymbol{S}_n)$, or using the notation in the book, $\mathrm{IW}_p(\nu_n, \boldsymbol{S}_n^{-1})$, 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 $\mathbb{E}[\Sigma] = \dfrac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$.
- $\Rightarrow$ 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.