EE 5111: Estimation

Jan - May 2025

Mini Project 2

April 17, 2025

The aim of this exercise is to study the importance of conjugate priors while performing Bayesian estimation.
Consider the estimation of the covariance of a bivariate Gaussian distribution. We have access to n observations $\boldsymbol{y_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ for $i = 1, ...n$. Here $\boldsymbol{y_i}$ is a 2×1 vector; $\mathbf{\Sigma}$ is a 2×2 matrix. We denote by $\vec{y}$ the set of all observations, $\boldsymbol{y_i}$. Perform the following experiments using $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ as the underlying covariance for $n = 10, 100, 1000$.

1. Estimate the covariance using Maximum Likelihood

2. Perform Bayesian estimation and provide a point estimate for the covariance. The conjugate prior distribution is the inverse Wishart distribution. The d-dimensional distribution is given by

$$InvWishart(\nu, \mathbf{\Delta}) : p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{\nu+d+1}{2}} \exp\left(-\frac{1}{2} Tr(\mathbf{\Delta}\mathbf{\Sigma}^{-1})\right) \tag{1}$$

   Consider the following hyperparameters for the prior: $\mathbf{\Delta_0} = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}$ and $\nu_0 = 5$; here $d = 2$. (Refer to Section 3.6 in Gelman)
   The posterior density is of the same density with parameters

$$\nu_n = \nu_0 + n \tag{2}$$

$$\mathbf{\Delta_n} = \mathbf{\Delta_0} + \sum_{i=1}^{n} \boldsymbol{y_i} \boldsymbol{y_i}^T \tag{3}$$

3. Use of Non-informative prior:
   As an alternative to the conjugate prior, use the non-informative Jeffrey's prior given by

$$p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-2}, \tag{4}$$

   and the independence-Jeffreys prior given by

$$p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-3/2}. \tag{5}$$

   What are the differences in the inferences using non-informative priors as compared to the conjugate prior?

4. Monte Carlo Bayesian estimation:
   This method is useful when the posterior is not available in closed form. Note that we require the mean of the posterior distribution.

$$p(\mathbf{\Sigma}|\vec{y}) = \frac{p(\vec{y}|\mathbf{\Sigma})p(\mathbf{\Sigma})}{\int p(\vec{y}|\mathbf{\Sigma})p(\mathbf{\Sigma})d\mathbf{\Sigma}} \tag{6}$$

$$\mathbb{E}_{\mathbf{\Sigma}|\vec{y}}[\mathbf{\Sigma}|\vec{y}] = \frac{\mathbb{E}_{\mathbf{\Sigma}}[\mathbf{\Sigma}p(\vec{y}|\mathbf{\Sigma})]}{\mathbb{E}_{\mathbf{\Sigma}}[p(\vec{y}|\mathbf{\Sigma})]} \tag{7}$$

   Note that the likelihood is

$$p(\vec{y}|\mathbf{\Sigma}) \propto det(\mathbf{\Sigma})^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{y_i}^T \mathbf{\Sigma}^{-1} \boldsymbol{y_i}\right). \tag{8}$$

   Instead of using the closed form expression for the posterior update, find the posterior using Monte Carlo integration using the following equation

$$A = \frac{\frac{1}{m}\sum_{j=1}^{m}\left[\mathbf{\Sigma}_j det(\mathbf{\Sigma}_j)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{y_i}^T \mathbf{\Sigma}_j^{-1} \boldsymbol{y_i}\right)\right]}{\frac{1}{m}\sum_{j=1}^{m}\left[det(\mathbf{\Sigma}_j)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{y_i}^T \mathbf{\Sigma}_j^{-1} \boldsymbol{y_i}\right)\right]} \tag{9}$$

   where each $\mathbf{\Sigma}_j \sim p(\mathbf{\Sigma})$ (a sample drawn from the prior distribution). Report the values of A for $n = 10, 100, 1000$ and for $m = 10^3, 10^4, 10^5$ for $p(\mathbf{\Sigma}) \sim InvWishart_{\nu_0}(\Delta_0)$ for the following parameters:

   (a) $\nu_0 = 5$, $\Delta_0 = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}$.

   (b) $\nu_0 = 5$, $\Delta_0 = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$.

   Which prior performs better? Why do you think it happens? Can you justify why modeling the prior is important? Note that you can now model your prior distribution as any non-conjugate distribution as well.

5. Hierarchical Bayes estimation and Gibbs sampling:
   Consider the following formulation of the prior for covariance.

$$\mathbf{\Sigma} \sim InvWishart\left(\nu + d - 1, 2\nu \boldsymbol{Diag}(\frac{\mathbf{1}}{\boldsymbol{a_1}}, \frac{\mathbf{1}}{\boldsymbol{a_2}})\right) \tag{10}$$

$$a_k \sim InvGamma(\frac{1}{2}, \frac{1}{A_k^2}) \tag{11}$$

   For performing Gibbs sampling, use the following equations to draw samples iteratively from one distribution and use the drawn samples in the next equation:

$$p(\mathbf{\Sigma}|\bar{y}, a_k) \sim InvWishart\left(\nu + d + n - 1, 2\nu \begin{bmatrix} 1/a_1 & 0 \\ 0 & 1/a_2 \end{bmatrix} + \sum_{i=1}^{n} \boldsymbol{y_i}\boldsymbol{y_i}^T\right) \tag{12}$$

$$p(a_k|\bar{y}, \mathbf{\Sigma}) \sim InvGamma\left(\frac{\nu + n}{2}, \nu(\Sigma^{-1})_{kk} + \frac{1}{A_k^2}\right) \tag{13}$$

   Use $A_1 = 0.05$ and $A_2 = 0.05$. Report the covariance estimate after 10^3 iterations of Gibbs sampling.

6. Empirical Bayes:
   For empirical Bayes, we consider an inverse Wishart prior

$$p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\frac{\nu+d+1}{2}} \exp\left(-\frac{1}{2} Tr(\mathbf{\Delta}\mathbf{\Sigma}^{-1})\right). \tag{14}$$

   However, instead of using a distribution over the parameters of the Wishart distribution, the marginal likelihood is computed as follows:

$$p(\bar{y}|\nu, \mathbf{\Delta}) = \int p(\bar{y}|\mathbf{\Sigma}) p(\mathbf{\Sigma}|\nu, \mathbf{\Delta}) d\mathbf{\Delta} \tag{15}$$

   The obtained $p(\bar{y}|\nu, \mathbf{\Delta})$ is then used to maximizing the log likelihood with respect to ν and $\mathbf{\Delta}$ to obtain ν_{opt} and $\mathbf{\Delta}_{opt}$,

$$\mathbf{\Delta}_{opt} = \frac{\nu}{n} \sum_{i=1}^{n} \boldsymbol{y_i} \boldsymbol{y_i}^T \tag{16}$$

$$\nu_{opt} = \arg\max_{\nu} \left[\nu \log\left(\frac{\nu+n}{\nu}\right) + n \log\left(\frac{\nu+n}{n}\right) + \log \frac{\Gamma_2(\nu/2)}{\Gamma_2((\nu+n)/2)} \right] \tag{17}$$

   where $\Gamma_d(a)$ is the multivariate gamma function,

$$\Gamma_d(a) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left(a - \frac{i-1}{2}\right). \tag{18}$$

   You may employ an iterative optimization algorithm to solve (17). The posterior is given by,

$$p(\mathbf{\Sigma}|\bar{y}, \nu_{opt}, \mathbf{\Delta}_{opt}) = InvWishart(\nu_{opt} + n, \mathbf{\Delta}_{opt} + \sum_{i=1}^{n} \boldsymbol{y_i} \boldsymbol{y_i}^T) \tag{19}$$

   Consider the following hyperparameters for the prior: $\mathbf{\Delta_0} = \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}$ and $\nu_0 = 5$. You can refer to `https://emtiyaz.github.io/Writings/wishart.pdf` for more details. Compare the estimate with the one obtained using the conjugate prior, non-informative prior and the hierarchical Bayes method.

Questions:

- Which of the six methods listed above would you advocate for this problem and why?
- Here, we deal with a dimension of $d = 2$. For a problem of a higher dimension, which method would you recommend? Justify.