ET Mini Project 2

Shiva Surya, Jayagowtham, Anuj, Praveen Prasanth

April 2025

1 MLE

Expression

The Maximum Likelihood Estimator (MLE) for the covariance matrix Σ given n samples $y_1, \ldots, y_n \in \mathbb{R}^d$ is:

$$\hat{\Sigma}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean.

Experimental Estimates and Errors

Sample Size	**Covariance Estimate**	**Error**
$N = 10$	$\begin{bmatrix} 0.9564, & 0.5115 \\ 0.5115, & 1.7069 \end{bmatrix}$	0.782
$N = 100$	$\begin{bmatrix} 1.0909 & 0.0677 \\ 0.0677 & 1.4259 \end{bmatrix}$	0.589
$N = 1000$	$\begin{bmatrix} 1.0031 & -0.0091 \\ -0.0091 & 1.9243 \end{bmatrix}$	0.077

Table 1: MLE covariance estimates and corresponding error values.

2 Bayesian Estimate with Inverse Wishart Conjugate Prior

2.1 Expression

The point estimate (analytical mean of the posterior distribution) is:

$$\Sigma_{\text{est}} = \mathbb{E}[\Sigma \mid \mathbf{y}] = \frac{\Delta_n}{\nu_n - d - 1}$$

- Δ_n = updated scale matrix (prior scale + data contribution)
- ν_n = updated degrees of freedom ($\nu_0 + n$)
- d = dimensionality of the data

Experimental Estimates and Errors

Sample Size	Covariance Estimate	Error
$N = 10$	$\begin{bmatrix} 1.1303, & 0.4263 \\ 0.4263, & 1.839 \end{bmatrix}$	0.637
$N = 100$	$\begin{bmatrix} 1.1087, & 0.0663 \\ 0.0663, & 1.4469 \end{bmatrix}$	0.571
$N = 1000$	$\begin{bmatrix} 1.0051, & -0.0091 \\ -0.0091, & 1.9254 \end{bmatrix}$	0.076

Table 2: Bayesian covariance estimates using Inverse Wishart prior and corresponding error values.

3 Jeffreys Priors

3.1 Expression

Non-Informative Jeffreys Prior
The Jeffreys non-informative prior is proportional to $|\Sigma|^{-(d+2)/2} = (|\Sigma|^{-2})$.

The posterior mean estimate is:

$$\Sigma_{\text{non-info}} = \frac{\Delta_n}{\nu_n - d - 1} = \frac{1}{n-2}\sum_{i=1}^{n} y_i y_i^\top$$

where:
- $\Delta_n = \sum_{i=1}^{n} y_i y_i^T$
- $\nu_n = \nu_0 + n$, with $\nu_0 = d - 1$; $\nu_n = n - 2$ here, for non-informative prior.

Jeffreys Independence Prior
The Jeffreys independence prior is proportional to $|\Sigma|^{-(d+1)/2} = (|\Sigma|^{-3/2})$.

The posterior mean estimate is:

$$\Sigma_{\text{ind}} = \frac{\Delta_n}{\nu_n - d - 1}$$

where:
- $\Delta_n = \sum_{i=1}^{n} y_i y_i^T$
- $\nu_n = \nu_0 + n$; $\nu_0 = 2k - d - 1$, $k = 3/2\,; \nu_0 = 0$ for independence prior.
- $\nu_n - d - 1 = n - 3$ in our case

3.2 Experimental Estimates and Errors

Non-informative Jeffreys Prior

Sample Size	**Covariance Estimate**	**Error**
$N = 10$	$\begin{bmatrix} 1.1955, & 0.6394 \\ 0.6394, & 2.1336 \end{bmatrix}$	0.935
$N = 100$	$\begin{bmatrix} 1.1132, & 0.0690 \\ 0.0690, & 1.4550 \end{bmatrix}$	0.565
$N = 1000$	$\begin{bmatrix} 1.0051, & -0.0092 \\ -0.0092, & 1.9282 \end{bmatrix}$	0.073

Jeffreys Independence Prior

Sample Size	**Covariance Estimate**	**Error**
$N = 10$	$\begin{bmatrix} 1.3663, & 0.7307 \\ 0.7307, & 2.4384 \end{bmatrix}$	1.181
$N = 100$	$\begin{bmatrix} 1.1246, & 0.0697 \\ 0.0697, & 1.4700 \end{bmatrix}$	0.553
$N = 1000$	$\begin{bmatrix} 1.0061, & -0.0092 \\ -0.0092, & 1.9301 \end{bmatrix}$	0.071

Table 3: Comparison of covariance estimates using Non-informative Jeffreys prior and Jeffreys Independence prior.

3.3 Comparison: Non-informative vs Conjugate Priors

- **Prior Knowledge Representation:** Non-informative priors (e.g., Jeffreys) encode minimal or no prior belief, often used to represent ignorance. Conjugate priors (e.g., Inverse-Wishart) encode specific prior beliefs, making them subjective but informative.
- **Effect on Posterior Distribution:** In non-informative priors, the posterior is largely driven by the likelihood, especially as the sample size increases. Conjugate priors have more influence when data is sparse and offer regularization.
- **Analytical Tractability:** Conjugate priors lead to posteriors in the same distribution family, allowing for closed-form expressions. Non-informative priors may lead to improper or non-conjugate posteriors, complicating inference.
- **Practical Use Case:** Non-informative priors are suitable when there is little to no prior knowledge. Conjugate priors are more appropriate when prior information is available and reliable, especially with smaller datasets.
- **Error rates** Non-Informative priors have high error with small samples compared to Conjugate priors, but with larger sample size, the errors are almost similar

4 Monte Carlo Bayesian Estimation

4.1 Expression

To get the mean of the posterior estimate, we use the sample average

$$A = \frac{\frac{1}{m}\sum_{j=1}^{m} \Sigma_j \left[\det(\Sigma_j)\right]^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^T \Sigma_j^{-1} y_i\right)}{\frac{1}{m}\sum_{j=1}^{m} \left[\det(\Sigma_j)\right]^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^T \Sigma_j^{-1} y\right)}$$

a) Prior $\nu = 5$, $\quad$ Scale $= \begin{bmatrix} 4 & 0 \\ 0 & 5 \end{bmatrix}$

Sample Size	**Covariance Estimate**	**Error**
N = 10, m = 10^3	$\begin{bmatrix} 1.141 & 0.439 \\ 0.439 & 1.855 \end{bmatrix}$	0.653
N = 10, m = 10^4	$\begin{bmatrix} 1.129 & 0.425 \\ 0.425 & 1.859 \end{bmatrix}$	0.631
N = 10, m = 10^5	$\begin{bmatrix} 1.133 & 0.428 \\ 0.428 & 1.846 \end{bmatrix}$	0.639
N = 100, m = 10^3	$\begin{bmatrix} 1.100 & 0.029 \\ 0.029 & 1.458 \end{bmatrix}$	0.553
N = 100, m = 10^4	$\begin{bmatrix} 1.118 & 0.066 \\ 0.066 & 1.464 \end{bmatrix}$	0.557
N = 100, m = 10^5	$\begin{bmatrix} 1.115 & 0.066 \\ 0.066 & 1.447 \end{bmatrix}$	0.572
N = 1000, m = 10^3	$\begin{bmatrix} 1.061 & -0.121 \\ -0.121 & 1.917 \end{bmatrix}$	0.199
N = 1000, m = 10^4	$\begin{bmatrix} 0.977 & -0.016 \\ -0.016 & 1.911 \end{bmatrix}$	0.095
N = 1000, m = 10^5	$\begin{bmatrix} 1.003 & -0.011 \\ -0.011 & 1.921 \end{bmatrix}$	0.080

Table 4: Covariance estimates with first prior

b) Prior $\nu = 5$, $\quad$ Scale = $\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$

Sample Size	**Covariance Estimate**	**Error**
N = 10, m = 10^3	$\begin{bmatrix} 0.972 & 0.443 \\ 0.443 & 1.802 \end{bmatrix}$	0.658
N = 10, m = 10^4	$\begin{bmatrix} 0.962 & 0.425 \\ 0.425 & 1.746 \end{bmatrix}$	0.653
N = 10, m = 10^5	$\begin{bmatrix} 0.965 & 0.428 \\ 0.428 & 1.761 \end{bmatrix}$	0.652
N = 100, m = 10^3	$\begin{bmatrix} 1.109 & 0.100 \\ 0.100 & 1.371 \end{bmatrix}$	0.654
N = 100, m = 10^4	$\begin{bmatrix} 1.087 & 0.064 \\ 0.064 & 1.448 \end{bmatrix}$	0.566
N = 100, m = 10^5	$\begin{bmatrix} 1.088 & 0.065 \\ 0.065 & 1.446 \end{bmatrix}$	0.568
N = 1000, m = 10^3	$\begin{bmatrix} 0.949 & 0.072 \\ 0.072 & 1.893 \end{bmatrix}$	0.156
N = 1000, m = 10^4	$\begin{bmatrix} 1.012 & -0.007 \\ -0.007 & 1.883 \end{bmatrix}$	0.119
N = 1000, m = 10^5	$\begin{bmatrix} 0.992 & 0.003 \\ 0.003 & 1.917 \end{bmatrix}$	0.084

Table 5: Covariance estimates with second prior

We notice very similar errors in both the priors. For smaller samples tho, the first prior is better than second in terms of the Frobenius norm of the error. This happens maybe because the first prior is more representative of the system than the second. The example demonstrates why modelling the prior is key, as deceptively looking priors yield estimates with notable differences.

5 Gibbs sampling

5.1 Expression

Gibbs sampling is a Markov Chain Monte Carlo (MCMC) method where each parameter is sampled conditionally given all others:

$$\Sigma_i^{(t+1)} \sim p(\Sigma \mid a_{-i}^{(t)}, y)$$

$$a_i^{(t+1)} \sim p(a \mid \Sigma_{-i}^{(t)}, y)$$

This sequential updating eventually generates samples from the joint posterior distribution.

5.2 Experimental Estimates and Error

Sample Size	**Covariance Estimate**	**Error**
$N = 10$	$\begin{bmatrix} 0.7611, & 0.3995 \\ 0.3995, & 1.3455 \end{bmatrix}$	0.897
$N = 100$	$\begin{bmatrix} 1.0781, & 0.0716 \\ 0.0716, & 1.4068 \end{bmatrix}$	0.6069
$N = 1000$	$\begin{bmatrix} 1.0142, & -0.0080 \\ -0.0080, & 1.9318 \end{bmatrix}$	0.070

Table 6: Covariance estimates obtained using Gibbs sampling.

6 Empirical Bayes Estimates

6.1 Expression

In Empirical Bayes, the optimal scale matrix Δ_{opt} and optimal degrees of freedom ν_{opt} are estimated by maximizing the marginal likelihood. They are given by:

$$\Delta_{\text{opt}} = \frac{\nu}{n} \sum_{i=1}^{n} y_i y_i^T$$

$$\nu_{\text{opt}} = \arg\max_{\nu} \left(\nu \log \nu + n \log \nu - \nu\, n + n\, d + \log \Gamma_2 \left(\frac{\nu}{2} \right) - \log \Gamma_2 \left(\frac{\nu + n}{2} \right) \right)$$

where:

- d is the dimension of the data,
- $\Gamma_2(\cdot)$ denotes the multivariate Gamma function.

6.2 Experimental Estimates and Errors

Sample Size	Covariance Estimate	Error
$N = 10$	$\begin{bmatrix} 1.0416, & 0.5570 \\ 0.5570, & 1.8589 \end{bmatrix}$	0.801
$N = 100$	$\begin{bmatrix} 1.1247, & 0.0697 \\ 0.0697, & 1.4700 \end{bmatrix}$	0.553
$N = 1000$	$\begin{bmatrix} 1.0061, & -0.0092 \\ -0.0092, & 1.9301 \end{bmatrix}$	0.071

Table 7: Covariance estimates obtained using Empirical Bayes.

7 Questions

7.1 Best of the 6 methods:

- Looking at the errors, it makes sense for us to look at the method which gave the lowest error for the smallest sample size (10) because that's where the prior actually stands out.
- The Bayesian estimate with inverse Wishart prior and the Monte Carlo versions of it provide us with the least error in estimates (≈ 0.63)

7.2 For large dimension , d >> 2

- For higher dimensions, we may expect sparse representations of data, and hence we expect the estimator to regularize the estimates.
- We can go with Bayesian with inverse Wishart as the prior as it is computationally light (the posterior mean), is a conjugate prior, hence we get the posterior distibution and also, regularizes the estimate to an extent.