MATH 531 Homework 7
Bayesian linear model and Maximum likelihood
March 1, 2025

Problem 1

Following the Wikipedia derivation that has been included in the canvas *Books* folder we have the formula for the posterior density function as:

With some re-arrangement,[4] the posterior can be re-written so that the posterior mean $\boldsymbol{\mu_n}$ of the parameter vector $\boldsymbol{\beta}$ can be expressed in terms of the least squares estimator $\hat{\boldsymbol{\beta}}$ and the prior mean $\boldsymbol{\mu_0}$, with the strength of the prior indicated by the prior precision matrix $\boldsymbol{\Lambda_0}$

$$\boldsymbol{\mu_n} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda_0})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\Lambda_0}\boldsymbol{\mu_0}).$$

To justify that $\boldsymbol{\mu_n}$ is indeed the posterior mean, the quadratic terms in the exponential can be re-arranged as a quadratic form in $\boldsymbol{\beta} - \boldsymbol{\mu_n}$.[5]

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\mu_0})^\mathsf{T}\boldsymbol{\Lambda_0}(\boldsymbol{\beta} - \boldsymbol{\mu_0}) = (\boldsymbol{\beta} - \boldsymbol{\mu_n})^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda_0})(\boldsymbol{\beta} - \boldsymbol{\mu_n}) + \mathbf{y}^\mathsf{T}\mathbf{y} - \boldsymbol{\mu_n}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda_0})\boldsymbol{\mu_n} + \boldsymbol{\mu_0}^\mathsf{T}\boldsymbol{\Lambda_0}\boldsymbol{\mu_0}.$$

Now the posterior can be expressed as a normal distribution times an inverse-gamma distribution:

$$\rho(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}) \propto (\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu_n})^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda_0})(\boldsymbol{\beta} - \boldsymbol{\mu_n})\right)(\sigma^2)^{-\frac{n+2a_0}{2}-1}\exp\left(-\frac{2b_0 + \mathbf{y}^\mathsf{T}\mathbf{y} - \boldsymbol{\mu_n}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Lambda_0})\boldsymbol{\mu_n} + \boldsymbol{\mu_0}^\mathsf{T}\boldsymbol{\Lambda_0}\boldsymbol{\mu_0}}{2\sigma^2}\right).$$

You should simplify this further as we did in class by assuming $\mu_0 = 0$ and $\Lambda_0 = \lambda I_k$.

1(a) For fixed σ^2 (BTW we are using γ in lecture for σ^2) maximize this posterior over β. and show that your answer only depends on the ratio λ.

Note that this posterior is a bit different than the one in the lectures because we used a different prior for β. For the Wikipedia version the prior for β is $MN(0, \sigma^2/\lambda I_k)$. We used a prior of $MN(0, \lambda I_k)$ to make the derivation easier to follow. In is not clear to me which of these would be preferred in practice.

Proof. We have the expression for our posteriro distribution above. To maximize this posterior distribution over β. To maximize this expression, we take the first-order partial derivative with respect to β and set it equal to 0. Before differentiating, we can simplify the expression by first taking the natural log of each side, and then eliminating terms which are not dependent on β. Thus, we can formulate this maximization problem as

$$\max_{\beta}\left(-\frac{1}{2\sigma^2}(\beta - \mu_n)^T(X^TX + \Lambda_0)(\beta - \mu_n)\right)$$
$$\to 0 = \frac{\partial}{\partial\beta}\left[\left(-\frac{1}{2\sigma^2}(\beta - \mu_n)^T(X^TX + \Lambda_0)(\beta - \mu_n)\right)\right]$$

We can consider the well known result of differentiating a quadratic form expression ($\frac{\partial}{\partial B} B^T AB = 2AB$) and apply this to the expression above.

$$\frac{\partial}{\partial \beta}\left[\left(-\frac{1}{2\sigma^2}(\beta-\mu_n)^T(X^TX+\Lambda_0)(\beta-\mu_n)\right)\right] = -\frac{1}{\sigma^2}(X^TX+\Lambda_0)(\beta-\mu_n) = 0$$

We can expand this expression by distributing the terms inside of the parentheses. Note that we can drop the term $-\frac{1}{\sigma^2}$ from the expression after dividing both sides of the expression of that term. Hence,

$$(X^TX+\Lambda_0)(\beta-\mu_n) = X^TX\beta + \Lambda_0\beta - X^TX\mu_n - \Lambda_0\mu_n = 0$$

Now, we isolate each side by subtracting terms with μ_n from each side. We have

$$X^TX\beta + \Lambda_0\beta = X^TX\mu_n + \Lambda_0\mu_n$$

Clearly, this expression directly implies that $\beta = \mu_n$. Therefore, the posterior distribution is maximized by choosing $\beta = \mu_n$. To ensure that this is indeed a maximum solution, we can take the second order derivative of the expression to determine the behavior of the critical point.

$$\frac{\partial}{\partial \beta}\left[-\frac{1}{\sigma^2}\left(X^TX\beta + \Lambda_0\beta - X^TX\mu_n - \Lambda_0\mu_n\right)\right] = -\frac{1}{\sigma^2}(X^TX+\Lambda_0)$$

We already know that X^TX and Λ_0 are both positive semidefinite, so we can say that $(X^TX+\Lambda_0)$ is positive semidefinite as well. When multiplied by the negative term $-\frac{1}{\sigma^2}$, then this yields a negative semidefinite Hessian matrix, which means that this solution is indeed a local maximum.

Finally, we examine why the optimal solution is only dependent on λ. When we optimize the posterior distribution over β, we obtain the solution $\beta = \mu_n$. The posterior mean can be written as follows:

$$\mu_n = (X^TX + \lambda I_k)^{-1}(X^Ty)$$

since $\mu_0 = 0$. X and y are fixed vectors, so we conclude that the expression for the posterior mean (and therefore the optimization solution) is only dependent on λ. $\square$

1(b) Show that your posterior mode (for fixed σ^2) is a also the solution to the frequentist, ridge regression problem:

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \omega \sum_{i=1}^{k} \beta^2$$

What is the relationship between ω in this case and σ^2 and λ in 1(a)?

Proof. We can rewrite this expression in terms of square norms

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \omega \sum_{i=1}^{k} \beta^2 = \min_{\beta} \left[\|Y - X\beta\|^2 + \omega\|\beta\|^2\right]$$

Again, we differentiate with respect to β and set the expression equal to 0 to solve the optimization problem.

$$0 = \frac{\partial}{\partial\beta} \left[\|Y - X\beta\|^2 + \omega\|\beta\|^2\right]$$

We apply the chain rule to each term to get

$$\begin{aligned}
\frac{\partial}{\partial\beta} \left[\|Y - X\beta\|^2 + \omega\|\beta\|^2\right] &= -2X^T(Y - X\beta) + 2\omega\beta \\
-2X^TY + 2X^TX\beta + 2\omega\beta &= 0 \\
2X^TX\beta + 2\omega\beta &= 2X^TY \\
(X^TX + \omega I_k)\beta &= X^TY \\
\beta &= (X^TX + \omega I_k)^{-1}X^TY
\end{aligned}$$

In the ridge regression derivation, ω in this context is equivalent to $\frac{\lambda}{\sigma^2}$ in the Bayesian linear model. $\square$

Problem 2

Assume a linear model

$$\mathbf{y} = X\beta + \mathbf{e}$$

where $\mathbf{y}$ is an vector of length n and X is a full rank, $n \times p$ matrix. $\mathbf{e}_i$ are independent $N(0, \sigma^2\Omega)$. Also suppose that β_T is "true" value for β and $\hat{\beta}$ the Generalized Least Squares (GLS) /MLE estimate and $\hat{\sigma}$ the MLE for σ. Here Ω is a known correlation matrix.

2(a) Identify the formulas for the MLEs $\hat{\beta}$ and $\hat{\sigma}_{MLE}$.

Proof. First, we derive the MLE for $\widehat{\beta}$. Let $\Sigma = \sigma^2\Omega$ First, we define the optimization problem for maximizing the log-likelihood function. Note that when we differentiate with respect to $\widehat{\beta}$

$$\max_{\widehat{\beta}} \ln L(\widehat{\beta}, \Sigma) \to 0 = \frac{\partial}{\partial \widehat{\beta}} \left[-\frac{1}{2}(y - X\widehat{\beta})^T \Sigma^{-1} (y - X\widehat{\beta}) \right]$$

This derivative expression simplifies to

$$\begin{aligned} 0 &= -\frac{1}{2} \cdot -2X^T\Sigma^{-1}(y - X\widehat{\beta}) \\ 0 &= X^T\Sigma^{-1}y - X^T\Sigma^{-1}X\widehat{\beta} \\ X^T\Sigma^{-1}X\widehat{\beta} &= X^T\Sigma^{-1}y \\ \widehat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \end{aligned}$$

Next, we set up a similar optimization problem for the MLE of $\widehat{\sigma}^2$. First, let $\widehat{\sigma}^2 = \gamma$. We can simplify the expression of the log-likelihood function as such:

$$\begin{aligned} \ln L(\widehat{\beta}, \gamma\Omega) &= \ln \left[(2\pi\gamma\Omega)^{-n/2} \exp\left(-\frac{1}{2\gamma}(y - X\widehat{\beta})^T \Omega^{-1} (y - X\widehat{\beta}) \right) \right] \\ &= \ln[(2\pi\gamma\Omega)^{-n/2}] + \left(-\frac{1}{2\gamma}(y - X\widehat{\beta})^T \Omega^{-1} (y - X\widehat{\beta}) \right) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\gamma\Omega) + \left(-\frac{1}{2\gamma}(y - X\widehat{\beta})^T \Omega^{-1} (y - X\widehat{\beta}) \right) \end{aligned}$$

Again, we maximize over γ by differentiating the log-likelihood function with respect to γ and fix this quantity to 0. We have

$$\begin{aligned}
\max_{\gamma} \ln L(\widehat{\beta}, \gamma\Omega) \to 0 &= \frac{\partial}{\partial\gamma}\left[-\frac{n}{2}\ln(\gamma\Omega) + \left(-\frac{1}{2\gamma}(y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta})\right)\right] \\
0 &= -\frac{n}{2}\cdot\Omega\cdot\frac{\Omega^{-1}}{\gamma} + \frac{1}{2\gamma^2}\left((y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta})\right) \\
\frac{n}{2\gamma} &= \frac{1}{2\gamma^2}\left((y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta})\right) \\
\gamma = \widehat{\sigma}^2 &= \frac{1}{n}\left((y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta})\right)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\widehat{\beta} &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \\
\widehat{\sigma}^2 &= \frac{1}{n}\left((y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta})\right)
\end{aligned}$$

□

2(b) Derive the Fisher information for this model when evaluated at the true values of β and σ^2. (You might reparametrize σ^2 as γ as in the lectures to make derivatives easier.)

Note that I will skip a fair bit of algebra for this problem, so as to be succinct. First, we calculate the score functions S_β and S_γ, where each score function is simply the respective partial derivative of the log-likelihood function, $\ell(\beta, \gamma)$.

$$\begin{aligned}
S_\beta &= \frac{\partial}{\partial\beta}\left[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\gamma\Omega) + \left(-\frac{1}{2\gamma}(y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta})\right)\right] \\
&= \frac{1}{\gamma}X^T\Omega^{-1}(y - X\beta) \\
S_\gamma &= \frac{\partial}{\partial\gamma}\left[-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\gamma\Omega) + \left(-\frac{1}{2\gamma}(y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta})\right)\right] \\
&= -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}(y - X\beta)^T\Omega^{-1}(y - X\beta)
\end{aligned}$$

We use these score functions to compute the Hessian matrix.

$$J(\beta, \gamma) = \begin{bmatrix} \frac{\partial S_\beta}{\partial \beta} & \frac{\partial S_\beta}{\partial \gamma} \\ \frac{\partial S_\gamma}{\partial \beta} & \frac{\partial S_\gamma}{\partial \gamma} \end{bmatrix}$$

We compute each element of the Hessian below. The algebra and calculus will be skipped to avoid unnecessary details. Also note that Hessian matrices are symmetric, which means the off-diagonal elements are equal to each other. We have

$$\begin{aligned} \frac{\partial S_\beta}{\partial \beta} &= -\frac{1}{\gamma} X^T \Omega^{-1} X \\ \frac{\partial S_\beta}{\partial \gamma} = \frac{\partial S_\gamma}{\partial \beta} &= -\frac{1}{\gamma^2} X^T \Omega^{-1} (y - X\beta) \\ \frac{\partial S_\gamma}{\partial \gamma} &= \frac{n}{2\gamma^2} - \frac{1}{\gamma^3} (y - X\beta)^T \Omega^{-1} (y - X\beta) \end{aligned}$$

Now, the Fisher information matrix is defined as

$$I(\beta, \gamma) = -\mathbb{E}[J(\beta, \gamma)]$$

We can apply the negative expectation to each element of the Hessian. We start first with the diagonal elements. The term $\frac{\partial S_\beta}{\partial \beta}$ is a constant term that does not depend on y, so the expectation of this term is simply

$$-\mathbb{E}\left[\frac{\partial S_\beta}{\partial \beta}\right] = -(-\frac{1}{\gamma} X^T \Omega^{-1} X) = \frac{1}{\gamma} X^T \Omega^{-1} X$$

For the other diagonal element, $\frac{\partial S_\gamma}{\partial \gamma}$, we note that the $\mathbb{E}[y - X\beta] = 0$ since we are evaluating the expression at the true value of the parameters. So when we take the negative expectation of this term, we have

$$-\mathbb{E}\left[\frac{\partial S_\gamma}{\partial \gamma}\right] = -\mathbb{E}\left[\frac{n}{2\gamma^2}\right] - 0 = \frac{n}{2\gamma^2}$$

Lastly, we examine the cross-derivative terms.

$$-\mathbb{E}\left[\frac{\partial S_\gamma}{\partial \gamma}\right] = \mathbb{E}\left[-\frac{1}{\gamma^2} X^T \Omega^{-1}(y - X\beta)\right] = 0$$

Thus, we have the Fisher information matrix:

$$I(\beta, \gamma) = \begin{bmatrix} \frac{1}{\gamma} X^T \Omega^{-1} X & 0 \\ 0 & \frac{n}{2\gamma^2} \end{bmatrix}$$

2(c) Do the GLS estimates for β have a covariance matrix that achieves the Cramer-Rao lower bound (the inverse Fisher information matrix)?

Proof. The Cramer-Rao inequality states that

$$\text{Var}[\widehat{\beta}] \geq I^{-1}(\beta, \gamma)$$

We will first derive the variance of $\widehat{\beta}$ to inspect if this estimator satisfies the Cramer-Rao lower bound.

$$\begin{aligned} \text{Var}[\widehat{\beta}] &= \text{Var}[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y] \\ &= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\text{Var}[y]((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1})^T \\ &= (X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}\Sigma\Sigma^{-1}X)(X^T\Sigma^{-1}X)^{-1} \\ &= (X^T\Sigma^{-1}X)^{-1} \end{aligned}$$

Note that we can pull the variance term γ out of the matrix Σ^{-1}. Therefore this expression becomes

$$\text{Var}[\widehat{\beta}] = \sigma^2(X^T\Omega^{-1}X)^{-1}$$

Now, we examine the inverse Fisher matrix. Note that the only element of $I(\beta, \gamma)$ that pertains to $\widehat{\beta}$ is the upper-left term of the 2×2 matrix. We will refer to this as $I_{\beta\beta}$. Therefore, we have

$$I_{\beta\beta}^{-1} = \left(\frac{1}{\gamma} X^T \Omega^{-1} X\right)^{-1}$$
$$= \gamma (X^T \Omega^{-1} X)^{-1}$$

Therefore, we have

$$\text{Var}[\widehat{\beta}] = I^{-1}(\beta, \gamma)$$

So the GLS estimate of β satisfies the Cramer-Rao lower bound. $\square$

2(d) The GLS estimate for σ^2 derived in previous homework differs slightly from the MLE but it is unbiased. It is given by

$$\hat{\sigma}^2_{GLS} = \frac{1}{(n-p)} (\mathbf{y} - X\hat{\beta})^T \Omega^{-1} (\mathbf{y} - X\hat{\beta})$$

and has variance $\frac{2\sigma^4}{(n-p)}$

Show that that the variance of this estimate achieves the Cramer-Rao lower bound in the limit as $n \to \infty$.

Proof. We have the variance of the GLS estimator of σ^2 as

$$\text{Var}[\widehat{\sigma}^2_{GLS}] = \frac{2\sigma^4}{n-p}$$

For the Cramer-Rao boun, we simply use the MLE derivation of $I_{\gamma\gamma}$ (the diagonal element of the inverse Fisher information matrix pertaining to γ) and substitute $(n-p)$ for n. We have

$$I^{-1}_{\gamma\gamma GLS} = \frac{2\gamma^2}{n-p} = \frac{2\sigma^4}{n-p}$$

We substitute these expressions back into the Cramer-Rao inequality.

$$\begin{aligned}\text{Var}[\widehat{\sigma}^2_{GLS}] &\geq I^{-1}_{\gamma\gamma GLS} \\ \frac{2\sigma^4}{n} &\geq \frac{2\sigma^4}{n-p}\end{aligned}$$

Clearly, in finite samples, the variance GLS estimator for σ^2 is greater than the Cramer-Rao lower bound. However, if we consider the behavior of the expression as $n \to \infty$, we have

$$\begin{aligned}\lim_{n\to\infty} \frac{2\sigma^4}{n} &\geq \lim_{n\to\infty} \frac{2\sigma^4}{n-p} \\ &\to 0 = 0\end{aligned}$$

Hence, as the sample size tends to infinity, then the GLS estimator satisfies the Cramer-Rao lower bound. $\square$