STAT 840 A4

In this question, we build a latent variable formulation to fit a variance component model through EM and MCML respectively. The formulation is:

$y|u \sim N(X\beta + Zu, \sigma^2 I_N)$

$u \sim N(0, \sigma^2 \tau I_m)$

, where $y \in R^N$ is the observed data, $u \in R^m$ are latent variables, $X \in R^{N\times p}$ is the desgin matrix for the mean, $Z \in R^{N\times m}$ is the design matrix for the variance, $\beta \in R^p$, $\sigma^2 > 0, \tau \leq 0$ are unknown parameters.

(a)

First, we prove the equivalence between the hierarchical conditional distribution and a joint distribution:

$$\begin{pmatrix} y \\ u \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} I_N + \tau ZZ^T & \tau Z \\ \tau Z^T & \tau I_m \end{pmatrix}\right)$$

First, we take advantage of the property of a partitioned multivariate normal distribution to ensure that $y|u$ indeed follows another normal distribution. Then, we obtain the expectation and covariance matrix for the conditional distribution, again using the property of a partitioned multivariate normal distribution:

$$E[y|u] = X\beta + [\sigma^2 \tau Z][\sigma^2 \tau I_m]^{-1}(u - 0) = X\beta + \sigma^2 \tau Z \frac{1}{\sigma^2} \frac{1}{\tau} I_m u = X\beta + Zu$$

I followed the formula provided by Wiki. The proof of the formula is not shown here, but here is my understanding: It first converts the anomaly of the realization of u (i.e., distance between the realization and the expected value of u) from a deterministic number to the "distributional distance" that is probabilistically weighted to account for the variability of the distribution of u . This is because the same deterministic anomaly distance can be rare in a distance in a distribution with less variability (e.g., above 3 standard deviation), but can be very common in a distribution with larger variability (e.g., within 0.5 standard deviation). Then, it measures how the anomaly of u will drag y to by right multiplying the covariance between y and u .

Then, the conditional covariance matrix can be obtained:

$$\frac{1}{\sigma^2} Cov(y|u) = (I_N + \tau ZZ^T) - (\tau Z)(\tau I_m)^{-1}(\tau Z^T) = I_N + \tau ZZ^T - \tau Z \frac{1}{\tau} I_m \tau Z^T = I_N + \tau ZZ^T - \tau ZZ^T = I_N$$

$$Cov(y|u) = \sigma^2 I_N$$

The insight here is that conditional probability is equivalent to shrinking the sample space, namely, reducing some uncertainty. If u follows a highly variable distribution, then providing one realization from it doesn't give much information, i.e., reducing much uncertainty. If u follow a less variable distribution, then having one realization provides more information.

Also, note that trivially, $u \sim N(0, \sigma^2 \tau I_m)$ from the lower part of this joint distribution.

For a more detailed proof, we use a trivial linear mapping to attain:

$u = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} y \\ u \end{pmatrix}$, so according to the property of a multivariate normal distribution,

$$E[u] = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} X\beta \\ 0 \end{pmatrix} = 0$$

,and

$$Cov(u) = \begin{pmatrix} 0 & 1 \end{pmatrix} \sigma^2 \begin{pmatrix} I_N + \tau ZZ^T & \tau Z \\ \tau Z^T & \tau I_m \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sigma^2 \begin{pmatrix} \tau Z^T & \tau I_m \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sigma^2 \tau I_m$$

, so $u \sim N(0, \sigma^2 \tau I_m)$

Finally, we draw the conclusion that the joint distribution is indeed equivalent to the hierarchical formulation:

$$y|u \sim N(X\beta + Zu, \sigma^2 I_N)$$

$$u \sim N(0, \sigma^2 \tau I_m)$$

(b)

In this question, we again apply the trivial linear mapping to show the equivalence between the variance component model and the hierarchical model.

From the joint distribution $\begin{pmatrix} y \\ u \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} I_N + \tau ZZ^T & \tau Z \\ \tau Z^T & \tau I_m \end{pmatrix}\right)$ that is equivalent to the hierarchical model, we have:

$$y = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} y \\ u \end{pmatrix}$$

Hence, by using the property of a multivariate normal distribution,

$$E[y] = \begin{pmatrix} 1 & 0 \end{pmatrix} E\left[\begin{pmatrix} y \\ u \end{pmatrix}\right] = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} X\beta \\ 0 \end{pmatrix} = X\beta$$

, and

$$Cov(y) = \begin{pmatrix} 1 & 0 \end{pmatrix} Cov\left[\begin{pmatrix} y \\ u \end{pmatrix}\right] \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix} \sigma^2 \begin{pmatrix} I_N + \tau ZZ^T & \tau Z \\ \tau Z^T & \tau I_m \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} =$$
$$\sigma^2 \begin{pmatrix} I_N + \tau ZZ^T & \tau Z \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \sigma^2 (I_N + \tau ZZ^T)$$

Thus, $y \sim N(X\beta, \sigma^2 (I_N + \tau ZZ^T))$

Therefore, the marginal distribution of y from the hierarchical model is just the variance component model.

Intuitively, I think the latent variable u stands for the group difference of each observation of y . For this model setting, we expect the between-group mean difference of y to be 0, but the variance to be $\sigma^2\tau$. The conditional distribution tells us that we expect the in-group variance of $y|u$ given a membership of a group to be σ^2 and i.i.d. The hierarchical model's idea follows identically with the variance component model.

(c)

Now, we derive the conditional distribution of $u|y$ from the joint distribution $\begin{pmatrix} y \\ u \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} I_N + \tau ZZ^T & \tau Z \\ \tau Z^T & \tau I_m \end{pmatrix}\right)$.

Again, by using the formula provided by Wiki, we have

$$E[u|y] = 0 + (\sigma^2\tau Z^T)[\sigma^2(I_N + \tau ZZ^T)]^{-1}(y - X\beta) = \sigma^2\tau Z^T \frac{1}{\sigma^2}(I_N + \tau ZZ^T)^{-1}(y - X\beta) = \tau Z^T(I_N + \tau ZZ^T)^{-1}(y - X\beta)$$

We can also get the conditional covariance matrix:

$$Cov(u|y) = \sigma^2\tau I_m - (\sigma^2\tau Z^T)[\sigma^2(I_N + \tau ZZ^T)]^{-1}(\sigma^2\tau Z) = \sigma^2\tau I_m - \sigma^2\tau Z^T \frac{1}{\sigma^2}(I_N + \tau ZZ^T)^{-1}\sigma^2\tau Z = \sigma^2\tau I_m - \sigma^2\tau^2 Z^T(I_N + \tau ZZ^T)^{-1}Z = \sigma^2\tau\left(I_m - \tau Z^T(I_N + \tau ZZ^T)^{-1}Z\right)$$

Therefore, the conditional distribution for $u|y$ is:

$$u|y \sim N\left(\tau Z^T(I_N + \tau ZZ^T)^{-1}(y - X\beta), \sigma^2\tau\left(I_m - \tau Z^T(I_N + \tau ZZ^T)^{-1}Z\right)\right)$$

(d)

Now, since we have both the conditional distribution $u|y$ and the joint distribution y, u , we can devise an EM algorithm to find the MLE iteratively.

The E-step calculates the expectation of the "complete data" joint log-likelihood $Q(\tau|\tau_j)$ for any τ with respect to the conditional density of $u|y$ parameterized by the last iteration of τ_j . It can be understood that, assuming that the hierarchical model describes a correct relationship between the latent variable and the observed data, a bad guess of τ_j will yield a bad joint log-likelihood expectation.

Here, we can derive the expression for the expectation of the joint log-likelihood function:

$$Q(\tau|\tau_j) = E_{U|y;\tau_j}[\log f_{Y,U}(y, u; \tau)] = \int f_{U|y;\tau_j}(u) \log f_{Y,U}(y, u; \tau) du$$

, where $f_{U|y;\tau_j}$ is the conditional density of $u|y$ parameterized by τ_j , and $f_{Y,U}$ is the joint density of the complete data.

Since the function in the integral is just a product of a normal distribution and a log-normal distribution, i.e., an exponential function of u multiplied by a quadratic form of u , the integral should be tractable by using the second generating moment. Nonetheless, the technical details, e.g., the derivation of the determinant/inverses of the non-diagonal/lower-triangular covariance matrices, are too cumbersome, so I choose to use a Monte Carlo to approximate the integral in the E-step.

Now, we just sample $u_1, ..., u_L$ from $u|y; \tau_j$ as many as possible, then we use the joint density $f_{Y,U}$ to obtain $\log f_{Y,U}(y, u_l; \tau)$ for any $l = 1, 2, ..., L$. Finally, we have the approximation:

$$Q(\tau|\tau_j) \approx \frac{1}{L}\sum_{l=1}^{L} \log f_{Y,U}(y, u_l; \tau)$$

After I have done several trials, I verify that by using a reasonable sample size (e.g., 1000), the above MCEM algorithm converges stably and accurately without significant variance. Therefore, I don't bother doing any importance-sampling/change-of-vairables for a better performance.

Then, the M-step finds the optimal $\tau_{j+1} = \arg\max_{\tau} Q(\tau|\tau_j)$, and the algorithm runs iteratively until convergence.

For the M-step, we use the R built-in descent gradient algorithm to find the maximizer. The reason why I don't use Newton's method is that the numerical Hessian can be problematic with high error, and the easiest algorithm using only the first derivative works well.

(e)

In this question, we derive a Monte Carlo approximation of the marginal likelihood. First, we derive the expression of the marginal likelihood:

$$L(\tau) = f_Y(y; \tau) = \int f_{Y,U}(y, u; \tau)du = \int f_{Y|u}(y|u)f_U(u; \tau)du$$

This formula tells us that we can approximate the marginal likelihood by: 1) sample as many as possible u from the latent distribution, 2) compute the conditional probability $y|u$ for each sampled u , and 3) take the average.

The latter two are trivial, given that we already know the conditional density function. For the first task, we do a change-of-variable to replace u with a standard normal random variable z that is independent of the parameter τ :

Let $z = (\sigma^2\tau)^{-1/2}u$, then by the property of a multivariate normal distribution, we have the following holds:

$$E[z] = (\sigma^2\tau)^{-1/2}E[u] = 0$$

, and

$$Cov(z) = [(\sigma^2\tau)^{-1/2}I_m]Cov(u)[(\sigma^2\tau)^{-1/2}I_m]^T = [(\sigma^2\tau)^{-1/2}I_m](\sigma^2\tau I_m)[(\sigma^2\tau)^{-1/2}I_m]^T = (\sigma^2\tau)^{-1}(\sigma^2\tau)I_m = I_m$$

, thus $z \sim N(0, I_m)$.

There are also some useful derivatives that we may use later:

$$u = (\sigma^2 \tau)^{1/2} z$$

$$du = |(\sigma^2 \tau)^{1/2} I_m| dz = (\sigma^2 \tau)^{m/2} dz$$

Then, we first reform the latent density function:

$$f_U(u;\tau) = (2\pi)^{-m/2} |\sigma^2 \tau I_m|^{-1/2} e^{-\frac{1}{2}[u^T (\sigma^2 \tau I_m)^{-1} u]} = (2\pi)^{-m/2} (\sigma^2 \tau)^{-m/2} e^{-\frac{1}{2\sigma^2 \tau} u^T u} = (2\pi)^{-m/2} (\sigma^2 \tau)^{-m/2} e^{-\frac{1}{2\sigma^2 \tau}((\sigma^2 \tau)^{1/2} z)^T ((\sigma^2 \tau)^{1/2} z)} = (2\pi)^{-m/2} (\sigma^2 \tau)^{-m/2} e^{-\frac{\sigma^2 \tau}{2\sigma^2 \tau} z^T z} = (2\pi)^{-m/2} (\sigma^2 \tau)^{-m/2} e^{-\frac{1}{2} z^T I_m z}$$

Then, we use the above identities to reform the expression of the marginal likelihood to:

$$L(\tau) = \int f_{Y|u}(y|u) f_U(u;\tau) du = \int f_{Y|z}(y|(\sigma^2 \tau)^{1/2} z)(2\pi)^{-m/2} (\sigma^2 \tau)^{-m/2} e^{-\frac{1}{2} z^T I_m z} (\sigma^2 \tau)^{m/2} dz = \int f_{Y|z}(y|(\sigma^2 \tau)^{1/2} z)(2\pi)^{-m/2} |I_m|^{-1/2} e^{-\frac{1}{2} z^T I_m z} dz = \int f_{Y|z}(y|\sigma\tau^{1/2} z)\phi(z) dz$$

, where $\phi(z)$ is the PDF of $z \sim N(0, I_m)$.

Hence, the MCML algorithm now is: 1) Sample $z_1, ..., z_N$ from the standard normal distribution $N(0, I_m)$ as many as possible, 2) compute the conditional probability $f_{Y|z}(y|\sigma\tau^{1/2} z;\tau)$ for each sample using the PDF of $y|z \sim N(X\beta + \sigma\tau^{1/2} Zz, \sigma^2 I_N)$, 3) average the conditional probability values. The explicit form following which we do the approximation is the empirical integral:

$$L(\tau) \approx \frac{1}{L} \sum_{l=1}^{L} f_{Y|z}(y|\sigma\tau^{1/2} z_l),\ z_1, ..., z_L \sim^{i.i.d} N(0, I_m)$$

, where $f_{Y|z}(y|\sigma\tau^{1/2} z_l)$ is the PDF of the distribution $y|z \sim N(X\beta + \sigma\tau^{1/2} Zz_l, \sigma^2 I_N)$.

Even though this form is already easy enough for computing, we can derive a more closed form since the R package "dmvnorm" for computing multivariate normal densities is too time-consuming:

$$f_{Y|z}(y|\sigma\tau^{1/2} z_l) = (2\pi)^{-N/2} |\sigma^2 I_N|^{-1/2} e^{-\frac{1}{2}(y - X\beta - \sigma\tau^{1/2} Zz_l)^T (\sigma^2 I_N)^{-1} (y - X\beta - \sigma\tau^{1/2} Zz_l)} = (2\pi)^{-N/2} \sigma^{-N} e^{-\frac{1}{2\sigma^2}\|y - X\beta - \sigma\tau^{1/2} Zz_l\|_2^2},\ l = 1, 2, ..., L$$

$$L(\tau) \approx \frac{1}{L} \sum_{l=1}^{L} f_{Y|z}(y|\sigma\tau^{1/2} z_l) = L(\tau) \approx \frac{(2\pi)^{-N/2} \sigma^{-N}}{L} \sum_{l=1}^{L} e^{-\frac{1}{2\sigma^2}\|y - X\beta - \sigma\tau^{1/2} Zz_l\|_2^2},\ z_1, ..., z_L \sim^{i.i.d} N(0, I_m)$$

Finally, we apply the log operation to make it a negative log-likelihood function:

$$l(\tau) = -\log(L(\tau)) = \frac{N}{2} log(2\pi) + N \log \sigma + \log L - \log\left(\sum_{l=1}^{L} e^{-\frac{1}{2\sigma^2}\|y - X\beta - \sigma\tau^{1/2} Zz_l\|_2^2}\right)$$

(f)

First, we implement the EM algorithm following part (d). In each iteration j, we sample 1000 latent observations $u_1, \ldots, u_{1000}$ from the conditional latent density $u|y \sim N\left(\tau Z^T \left(I_N + \tau Z Z^T\right)^{-1}(y - X\beta), \sigma^2 \tau \left(I_m - \tau Z^T \left(I_N + \tau Z Z^T\right)^{-1} Z\right)\right)$ using the R package "mvrnorm". The sample size is verified to be stable enough after a number of trials. Then, we implement the complete data joint PDF to calculate $Q(\tau|\tau_j) \approx \frac{1}{1000}\sum_{l=1}^{1000} \log f_{Y,U}(y, u_l; \tau)$. The R package "dmvnorm" worked but turned out to be extremely slow. Hence, I typed in the analytical mathematical formula for the numtivariate normal distribution PDF using vectorized operations only. The speed eventually increased several times and reached satisfaction. Then, we try to search for the optimal τ_{j+1} on the interval $[0, 5]$ using gradient descent algorithm. The stopping criteria is when the last iterations don't update larger than 0.0001, which seem to be sufficiently accurate for this task.

I input $\tau_0 = 10$, then the EM algorithm converges in 13 iterations with the convergence plot:

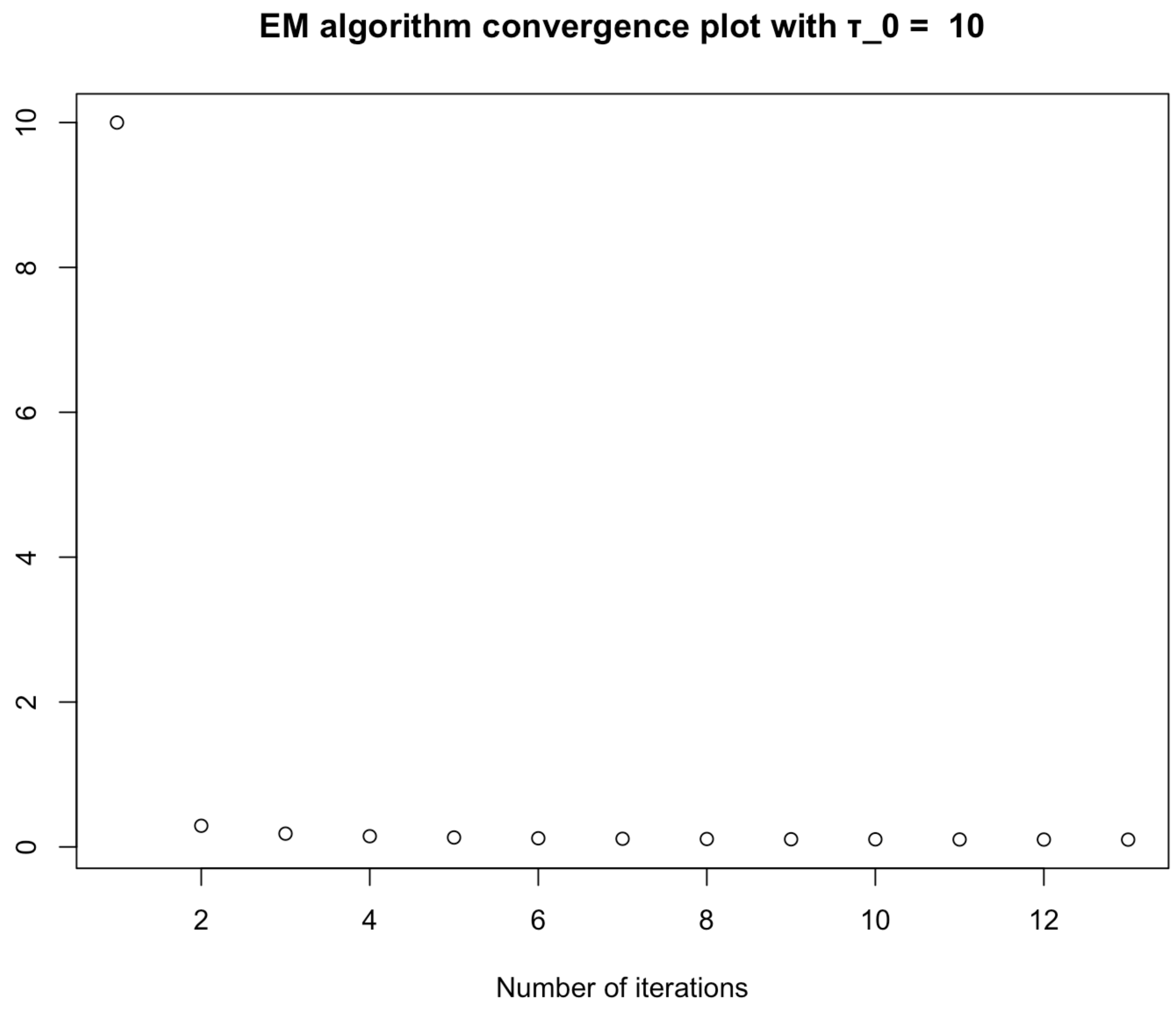

Fig.1 EM algorithm: Number of iterations vs. tau values

The EM algorithm converges to $\hat{\tau} \approx 0.1006944$, which is approximately the true parameter. I also tried more initial values, and the algorithm is robust to all large ones (e.g., even for $\tau_0 = 1000$, it converges to $\hat{\tau} = 0.1012296$), but converged to a wrong MLE (much smaller) when the initial value is below 0.01. I believe this is partially because of numerical stability. The extremely small numbers might have been treated as 0 numerically, even though they are not actually. Another reason might be because my stopping criteria was set loosely.

Then, we implement an MCML algorithm to directly estimate the marginal log-likelihood based on the hierarchical model setting following exactly part (e). I again used a self-typed multivariate PDF (for $y|u;\tau$ this time) instead of using the R "dmvnorm" package to speed up multiple times. All the operations in the MCML algorithm are efficient vectorized ones rather than using any for-loops. Also, to overcome the numerical stability problem, the last term in the marginal log-lihood $\log\left(\sum_{l=1}^{L} e^{-\frac{1}{2\sigma^2}\|y-X\beta-\sigma\tau^{1/2}Zz_l\|_2^2}\right)$ is calculated using the "logSumExp" package. This is because some bad τ values will yield very large residuals, and some rare z_l samples might also have very extreme residuals. They all generate very small probabilities within the summation. Hence, we operate the computings on a log-scaled level of order to make the algorithm numerically stable.

By setting the MC sample size of u to 2000, the algorithm has a good balance between performance and efficiency. The yielded marginal log-likelihood function curve is here:

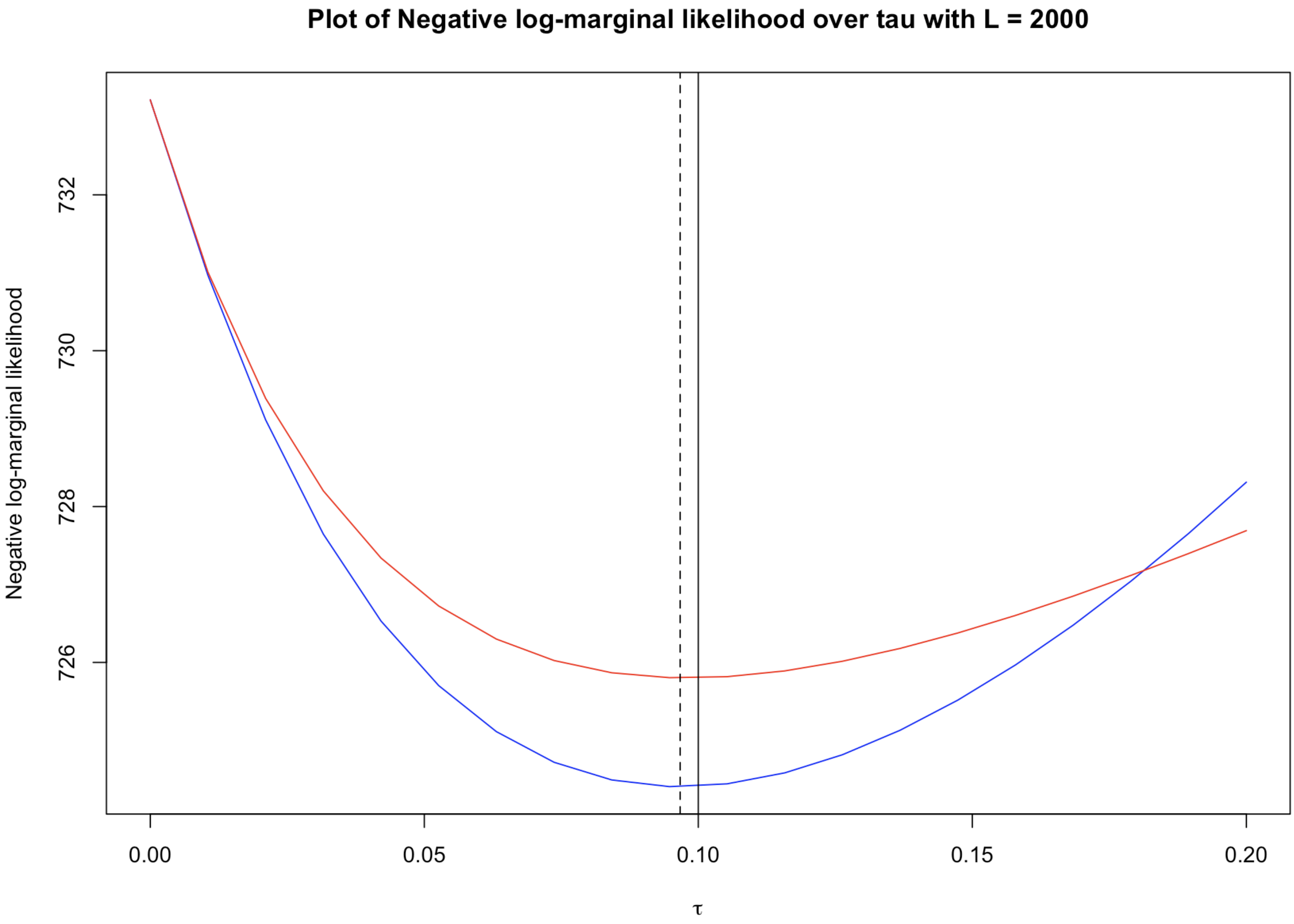

Fig.2 tau vs. negative marginal log-likelihood value. Red curve: The closed-form analytical formula for the marginal log-likelihood; Blue curve: The MC approximated marginal lieklihood. Solid line: The true parameter; Dashed line: The MCMLE

We can see from Fig.2 that the curves of the MC approximated marginal likelihood is fairly close to the analytical one at the interval $\tau \in [0, 0.2]$ that is close to the true parameter. The MCMLE $\hat{\tau} \approx 0.09668646$ is also close to the true parameter.

In general, the EM algorithm is harder to implement, since it requires drawing a new set of latent samples in each iteration. It also requires a lot of performence optimization operations, especitaly for computing the joint PDF for each latent sample. The problems are all from the computing burden of my solution. The trouble can be avoided if I derive the mathematical form of the joint likelihood expectation rather than approximating it through MC. Nonetheless, the mathematics can be extremely cumbersome to derive, so we have to pick one poison eventually. The MCML algorithm is more intuitive to implement. The hardest part is to take numerical stability into account, since we will generate a lot of small probability values during the algorithm and compute them. According to some of my other trials, my MCML algorithm can be more variable and sensitive to a small latent sample size. Also, even though it's not that relevant, the MC approximated marginal log-likelihood value deviates from the analytical one worse and worse as it goes farther from the true parameter. I believe it's because we fail to sample any latent value that will yield a non-zero $y|u$ probability with a finite sample size when τ goes too far from the true parameter. The analytical form, however, can be regarded as having an infinite sample size and adding up all possibilities, which makes the analytical margin likelihood larger than the MC approximate one for those uncommon parameters.

The final conclusion is that, the EM gives an MLE of $\hat{\tau} \approx 0.1006944$, and the MCML gives an MLE $\hat{\tau} \approx 0.09668646$, which means that they all yield an accurate MLE as the true parameter $\tau = 0.1$. EM is harder to design and implement, but it yields a more accurate MLE than the MCML.