CS698S: Bayesian Machine Learning *Your Name & Roll no:*

Mid-sem Exam (2016-2017, Sem-II)

Instructions:

- This question paper is worth a total of 100 marks.
- Total duration is 3 hours.
- Please write your name and roll number at the top of this sheet as well as on the provided answer copy.
- The True/False questions have to be answered on this question sheet itself. For True-False questions, please write T/F in the square bracket printed at the beginning of each question. The question sheet must also be submitted.
- To answer the other parts of the question paper, please use the provided answer copy.
- For rough work, please use pages at the back of answer copy. Extra sheets can be provided if needed.

1 True/False Problems (6 questions, 6 marks)

1. [**T**] The Laplace approximation for Bayesian linear regression with a Gaussian likelihood and Gaussian prior on the weights (assuming no other hyperparameters) would give us the true posterior.
2. [**F**] Given i.i.d. data $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ from some model $p(\boldsymbol{x}|\theta)$, the following always hold: $\int p(\mathbf{X}|\theta)d\theta = 1$. ($p(\mathbf{X}|\theta) = \prod_{n=1}^N p(\boldsymbol{x}_n|\theta)$ is a function of θ, not a *distribution* over θ. Thus it need not integrate to 1.)
3. [**F**] Models for which each parameter/hyperparameter has a prior that is conjugate to its associated likelihood function do not require doing approximate inference to infer the overall posterior distribution. (It only makes the model locally conjugate. Approximate inference is still needed in general to infer the joint posterior of all the unknowns)
4. [**T**] Bayesian inference can still be done when the prior is uninformative.
5. [**F**] The Metropolis algorithm, unlike Metropolis-Hastings, doesn't require an accept/reject step.
6. [**F**] Unlike MCMC based inference, VB is an optimization based approach, which gives a point estimate of the model parameters.

2 Short Answer Problems (14 questions, 28 marks)

1. Briefly explain what "integrating out" means in a Bayesian model. You may use an example.

   Answer: Integrating out means averaging a *random* quantity, say $f(\theta)$ (f can be a distribution, or function in general), that depends on a random variable, say θ, over all possibilities of that random variable. In general, this boils down to computing an expectation of the form $\mathbb{E}_{p(\theta|.)}[f(\theta)] = \int f(\theta)p(\theta|.)d\theta$.

2. Briefly explain the difference between likelihood and marginal likelihood. Give a example scenario where you may be interested in computing and working with the marginal likelihood.

   Answer: The likelihood $p(\boldsymbol{x}|\theta, m)$ tells us that , given some model m, how likely (or "probable") the data $\boldsymbol{x}$ is for a given value of the parameter θ. Marginal likelihood $p(\boldsymbol{x}|m)$ tells us how likely $\boldsymbol{x}$ is under model m, regardless of the value θ might take. This boils down to integrating out θ from $p(\boldsymbol{x}|\theta, m)$.

3. MAP estimation makes use of prior distributions on the model parameters. Does it then make MAP estimation a Bayesian method? If yes, why? If no, why not?

   Answer: No, because MAP estimation only gives us a point estimate of the parameters, not the full posterior distribution over the parameters.

4. Why is assuming a Gaussian prior $\mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D)$ on the parameters $\boldsymbol{\theta} \in \mathbb{R}^D$ equivalent to putting a regularizer on $\boldsymbol{\theta}$. What's this regularizer and what's the role of the precision hyperparameter λ?

   Answer: Because the negative of the log-prior which corresponds to a regularizer has a quadratic term of the form $\lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}$. Here λ controls the extent of regularization (large λ means more regularization).

5. Suppose you want to estimate (using a Bayesian approach) the probabilities of each of the six faces of a dice showing up upon a dice roll, given the past outcomes of a number of rolls of this dice. The parameter to be estimated here is defined by a 6-dimensional probability vector $\pi = (\pi_1, \ldots, \pi_6)$. Suggest which distributions would you choose for the likelihood model and the prior, and justify the reason you chose these distributions.

   Answer: Multinoulli likelihood (each observation is caterorical), Dirichlet prior (since π sums to 1).

6. What are the sufficient statistics for a Bayesian linear regression model? The sufficient statistics, intuitively, should be such that, if you have computed and store those, you can simply *forget* the past training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, and given some new training data, can update the posterior using the stored sufficient statistics and the new data.

   Answer: $\mathbf{A}_N = \sum_{n=1}^N \boldsymbol{x}_n\boldsymbol{x}_n^\top$ and $\boldsymbol{b}_N = \sum_{n=1}^N y_n\boldsymbol{x}_n$. Given $\mathbf{A}_N$, $\boldsymbol{b}_n$, and a new observation, say $(\boldsymbol{x}_{N+1}, y_{N+1})$, we can update the suff-stats as $\mathbf{A}_{N+1} = \mathbf{A}_N + \boldsymbol{x}_{N+1}\boldsymbol{x}_{N+1}^\top$ and $\boldsymbol{b}_{N+1} = \boldsymbol{b}_N + y_{N+1}\boldsymbol{x}_{N+1}$.

7. Suppose the posterior $p(\theta|\mathcal{D})$ (exact or approximated) is a well-known distribution but the predictive posterior $p(\mathcal{D}_{new}|\mathcal{D}) = \int p(\mathcal{D}_{new}|\theta)p(\theta|\mathcal{D})d\theta$ involves an intractable integral. Suggest the *simplest possible* way (that you know of) to approximate it.

   Answer: Simple Monte-Carlo approximation. Since $p(\theta|\mathcal{D})$ is a known distribution, we can directly draw a set of i.i.d. samples $\{\theta^{(s)}\}_{s=1}^S$ from it and approximate the predictive posterior by the following empirical average: $\frac{1}{S}\sum_{s=1}^S p(\mathcal{D}_{new}|\theta^{(s)})$.

8. When using VB to approximate a posterior p using another distribution q, why directly minimizing $KL(q||p)$ w.r.t. q is hard? How does VB get around this issue?

   Answer: Since the posterior p itself is intractable, we can't directly minimize $KL(q||p)$. However, minimizing this KL w.r.t. q is equivalent to maximizing the EBLO w.r.t. q, which does not depend on the the posterior p. VB essentially does the same!

9. State at least two benefits of having a predictive posterior $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ over having a plug-in approximation $p(y_*|\boldsymbol{x}_*, \hat{\boldsymbol{w}})$ for the label distribution. (you may take linear/logistic regression as an example)?

   Answer: (1) Less risk of overfitting since the predictive posterior $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y}) = \int p(y_*|\boldsymbol{w}, \boldsymbol{x}_*)p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})$ doesn't rely on a "single best" value of $\boldsymbol{w}$ but rather averages over all possible values of $\boldsymbol{w}$; (2) The variance of predictive posterior $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ tells us how certain the model is about the prediction y_* for any $\boldsymbol{x}_*$ (which can be useful when doing sequential data acquisition or "active learning").

10. Why a Gaussian distribution may not be an appropriate choice as the prior on variance parameter of some distribution (say another Gaussian)?

    Answer: Because a Gaussian has support over the entire real line whereas the variance is non-negative.

11. Suppose we are generating an observation $\boldsymbol{x} \in \mathbb{R}^D$ by first choosing a cluster-id $\boldsymbol{z} \in \{1, \ldots, K\}$ from a categorical distribution $p(\boldsymbol{z})$ which is a Multinoulli$(\boldsymbol{z}|\pi_1, \ldots, \pi_K)$, and then drawing $\boldsymbol{x}$ *conditioned* on $\boldsymbol{z}$ from a distribution $p(\boldsymbol{x}|\boldsymbol{z})$ which is a Gaussian $\mathcal{N}(\boldsymbol{x}|\mu_{\boldsymbol{z}}, \Sigma_{\boldsymbol{z}})$. Write down the expression for the marginal distribution $p(\boldsymbol{x})$ in terms of the other distributions specified for this problem.

    Answer: Using the sum rule, $p(\boldsymbol{x}) = \sum_{k=1}^K p(\boldsymbol{z} = k)p(\boldsymbol{x}|\boldsymbol{z} = k) = \sum_{k=1}^K \pi_k\mathcal{N}\boldsymbol{x}|\mu_k, \Sigma_k)$.

12. Suppose we have a *count-valued* random variable c drawn from a Poisson distribution $p(c|\lambda) = \frac{\lambda^c e^{-\lambda}}{c!}$. We then generate a binary random var. b as $b = \mathbb{1}[c > 0]$, where $\mathbb{1}$ denotes the indicator function which returns 1 if the condition is true, and 0 otherwise. What will be the marginal distribution $p(b|\lambda)$ of b?

    Answer: $b = 0$ only when $c = 0$. Thus $p(b = 0|\lambda) = p(c = 0|\lambda) = e^{-\lambda}$ and $p(b = 1|\lambda) = 1 - e^{-\lambda}$.

13. Why is VB computationally faster than sampling based methods such as MCMC (assuming both are run for the same number of iterations)?

    Answer: Doing MCMC requires sampling from distributions and sampling in general is computationally expensive. Standard VB doesn't require sampling.

14. Given N i.i.d. observations $\{x_1, \ldots, x_N\}$ drawn from a Poisson $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, and assuming a gamma prior $p(\lambda|a, b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}$ on λ, derive the expression for the posterior of λ.

    Answer: Due to the Poisson-gamma conjugacy, the posterior will be gamma: $p(\lambda|a+\sum_{n=1}^{N} x_n, b+N)$.

3 Medium-Length Answer Problems (5 questions, 50 marks)

1. Suppose we have some model for which the observations are drawn from a distribution $p(x|\theta)$. Let's assume the following prior on the parameters θ: $p(\theta) \propto |\mathbf{I}(\theta)|^{1/2}$ where $\mathbf{I}(\theta) = -\mathbb{E}_{p(x|\theta)}\left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right]$, where $|.|$ denotes the determinant of its argument (or just the value if the argument is a scalar).

    Assume $p(x|\theta) = \mathcal{N}(x|\mu, \sigma^2)$ where the parameters $\theta = (\mu, \sigma^2)$. For this case, use the above definition of $p(\theta)$ to derive the priors $p(\mu)$ (assuming σ^2 is fixed) and $p(\sigma^2)$ (assuming μ is fixed). What are these distributions and what does this imply about the general form of the prior $p(\theta) \propto |\mathbf{I}(\theta)|^{1/2}$?

    Answer: The log-likelihood $\mathcal{L} = -\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2 + \text{const}$.
    - Treating σ^2 as fixed, $\frac{\partial \mathcal{L}}{\partial \mu} = \frac{(x-\mu)^2}{\sigma^2}$ and $\frac{\partial^2 \mathcal{L}}{\partial \mu^2} = -\frac{1}{\sigma^2}$. Therefore $\mathbf{I}(\mu) = -\mathbb{E}_{p(x|\theta)}[-\frac{1}{\sigma^2}] = \frac{1}{\sigma^2}$ and thus $p(\mu) \propto \frac{1}{\sigma}$. Note that $p(\mu)$ does not depend on μ and is basically a uniform prior.
    - Treating μ as fixed, $\frac{\partial \mathcal{L}}{\partial \sigma^2} = \frac{(x-\mu)^2}{2(\sigma^2)^2} - \frac{1}{2\sigma^2}$ and $\frac{\partial^2 \mathcal{L}}{\partial (\sigma^2)^2} = -\frac{(x-\mu)^2}{(\sigma^2)^3} + \frac{1}{2(\sigma^2)^2}$. Therefore $\mathbf{I}(\sigma^2)$ will be $\mathbf{I}(\sigma^2) = -\mathbb{E}_{p(x|\theta)}\left[-\frac{(x-\mu)^2}{(\sigma^2)^3} + \frac{1}{2(\sigma^2)^2}\right]$. Using $\mathbb{E}[(x-\mu)^2] = \sigma^2$, this simplifies to $\mathbf{I}(\sigma^2) = \frac{1}{2(\sigma^2)^2}$, and thus $p(\sigma^2) \propto \frac{1}{\sigma^2}$. Also note that the prior on $\log\sigma^2$ is a uniform prior!

    (The following comment is not required for the answer; useful to know about nevertheless): Note that the final form of these priors are tied to observation model. One good thing is that these priors are uninformative priors (popularly known as Jeffreys prior). These priors often have some nice properties, e.g., $p(\mu)$ above is translation-invariant (adding/subtracting a constant to μ doesn't change our inference) by and $p(\sigma^2)$ above is scale-invariant (scaling σ^2 by a constant doesn't change our inference).

2. Consider a model where each observation $\boldsymbol{x} \in \mathbb{R}^D$ is drawn as follows: $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\epsilon}$, with $\boldsymbol{A} \in \mathbb{R}^{D\times K}$, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \mathbf{I}_K)$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Psi})$. Suppose some entries of the vector $\boldsymbol{x}$ are missing. Denote the vector $\boldsymbol{x}$ by splitting into 2 parts, observed and missing, and write $\boldsymbol{x} = [\boldsymbol{x}_o, \boldsymbol{x}_h]$ where $\boldsymbol{x}_o$ and $\boldsymbol{x}_h$ denote the observed and missing entries of $\boldsymbol{x}$, respectively. Treating the missing entries $\boldsymbol{x}_h$ as latent variables, derive their posterior distribution, i.e., $p(\boldsymbol{x}_h|\boldsymbol{x}_o, \mathbf{A}, \boldsymbol{\Psi})$. Treat $\mathbf{A}$ and $\boldsymbol{\Psi}$ as given. You may use the properties of Gaussian distributions.

    Answer: Note that we basically have to find the conditional distribution of $\boldsymbol{x}_h$, given $\boldsymbol{x}_o$. If we can get the joint distribution of $\boldsymbol{x}_o$ and $\boldsymbol{x}_h$, then we can apply Gaussian conditional from Gaussian joint formula to get our answer. Note that the joint is nothing but simply the marginal $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$. Here we are integrating out $\boldsymbol{z}$ because we want to compute the answer $p(\boldsymbol{x}_h|\boldsymbol{x}_o, \mathbf{A}, \boldsymbol{\Psi})$ without using $\boldsymbol{z}$.

    Using $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z}, \Psi)$ and $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \mathbf{I})$, and recognizing $\boldsymbol{x}$ as a linear transformation of $\boldsymbol{z}$ by $\mathbf{A}$, plus added Gaussian noise, we get the joint distribution $p(\boldsymbol{x}) = p(\boldsymbol{x}|\mathbf{A}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Psi}) = p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that this is only in terms of $\mathbf{A}$ and $\boldsymbol{\Psi}$, with $\boldsymbol{z}$ integrated out.

    Now assume $\boldsymbol{\mu}$ partitioned as $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_h \end{bmatrix}$ and $\boldsymbol{\Sigma}$ partitioning block-wise as $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{oh} \\ \boldsymbol{\Sigma}_{ho} & \boldsymbol{\Sigma}_{hh} \end{bmatrix}$, and apply the identity for Gaussian conditional from Gaussian joint to get the conditional $p(\boldsymbol{x}_h|\boldsymbol{x}_o, \mathbf{A}, \boldsymbol{\Psi})$.

3. Consider a matrix factorization model for a partially observed $N \times M$ matrix $\mathbf{R}$, where $p(r_{ij}|\boldsymbol{u}_i, \boldsymbol{v}_j) = \mathcal{N}(r_{ij}|\boldsymbol{u}_i^\top \boldsymbol{v}_j, \beta^{-1})$, and $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ denote the latent factors of i-th row and j-th column or $\mathbf{R}$, respectively. The predictive posterior of each r_{ij} is defined as $p(r_{ij}|\mathbf{R}) = \int p(r_{ij}|\boldsymbol{u}_i, \boldsymbol{v}_j)p(\boldsymbol{u}_i, \boldsymbol{v}_j|\mathbf{R})d\boldsymbol{u}_i d\boldsymbol{v}_j$, which is in general intractable. Suppose we are given a set of S samples $\{\mathbf{U}^{(s)}, \mathbf{V}^{(s)}\}_{s=1}^S$ generated by a Gibbs sampler for this matrix factorization model, where $\mathbf{U}^{(s)} = \{\boldsymbol{u}_i^{(s)}\}_{i=1}^N$ and $\mathbf{V}^{(s)} = \{\boldsymbol{v}_j^{(s)}\}_{j=1}^M$.

   Given these samples, derive the expressions for the sample based approximation of the *mean* as well as the *variance* of any entry r_{ij} (observed/unobserved) of the matrix $\mathbf{R}$.

   Answer: Note that the Gaussian model $p(r_{ij}|\boldsymbol{u}_i, \boldsymbol{v}_j) = \mathcal{N}(r_{ij}|\boldsymbol{u}_i^\top \boldsymbol{v}_j, \beta^{-1})$ implies that $r_{ij} = \boldsymbol{u}_i^\top \boldsymbol{v}_j + \epsilon_{ij}$ where $\epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1})$. Thus the posterior mean of r_{ij} given $\mathbf{R}$ will be $\mathbb{E}_{p(\mathbf{U},\mathbf{V}|\mathbf{R})}[r_{ij}|\mathbf{R}]$, where I am explicitly writing that the expectation is w.r.t. the posterior distribution of $\mathbf{U}, \mathbf{V}$.

   Given the samples $\{\mathbf{U}^{(s)}, \mathbf{V}^{(s)}\}_{s=1}^S$ from the posterior, we can approximate the mean as

$$\mathbb{E}[r_{ij}|\mathbf{R}] = \mathbb{E}[\boldsymbol{u}_i^\top \boldsymbol{v}_j + \epsilon_{ij}] = \mathbb{E}[\boldsymbol{u}_i^\top \boldsymbol{v}_j] \approx \frac{1}{S}\sum_{s=1}^S {\boldsymbol{u}_i^{(s)}}^\top \boldsymbol{v}_j^{(s)}$$

   The variance can be likewise approximated as

$$\begin{aligned}
\text{Var}[r_{ij}|\mathbf{R}] = \text{Var}[\boldsymbol{u}_i^\top \boldsymbol{v}_j + \epsilon_{ij}] &= \text{Var}[\boldsymbol{u}_i^\top \boldsymbol{v}_j] + \text{Var}[\epsilon_{ij}] \\
&= \mathbb{E}[(\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2] - \left[\mathbb{E}[\boldsymbol{u}_i^\top \boldsymbol{v}_j]\right]^2 + \beta^{-1} \\
&\approx \frac{1}{S}\sum_{s=1}^S \left({\boldsymbol{u}_i^{(s)}}^\top \boldsymbol{v}_j^{(s)}\right)^2 - \left(\frac{1}{S}\sum_{s=1}^S {\boldsymbol{u}_i^{(s)}}^\top \boldsymbol{v}_j^{(s)}\right)^2 + \beta^{-1}
\end{aligned}$$

4. Suppose we are trying to estimate an unknown quantity θ using some data X. Suppose we have come up with a *point estimate* $\hat{\theta}$ of θ (let's call $\hat{\theta}$ our "action") where $\hat{\theta}$ is estimated by minimizing a certain loss function. Define a function $\ell(\hat{\theta}, \theta)$ where $\ell(.,.)$ measures the difference between $\hat{\theta}$ and θ. The Bayesian *expected* loss is defined as $\varphi(\hat{\theta}) = \mathbb{E}_{p(\theta|X)}\ell(\hat{\theta}, \theta)$ where $p(\theta|X)$ denotes the posterior of θ given data X. We would like to estimate $\hat{\theta}$ by minimizing the Bayesian expected loss.

   - What's the optimal action $\hat{\theta}$ when $\ell(\hat{\theta}, \theta)$ is the squared loss, i.e., $\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$? How is $\hat{\theta}$ related to the posterior $p(\theta|X)$?
   - What's the optimal action $\hat{\theta}$ when $\ell(\hat{\theta}, , \theta)$ is the absolute loss, i.e., $\ell(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$? How is $\hat{\theta}$ related to the posterior $p(\theta|X)$?

   If needed, you may use the following result about derivative of a definite integral (The symbol $'$ here denotes the derivative w.r.t. x): $(\int_{a(x)}^{b(x)} f(x,t)dt)' = \int_{a(x)}^{b(x)} f'(x,t)dt + f(x, b(x))b'(x) - f(x, a(x))a'(x)$.

   Answer: For the squared loss case, $\varphi(\hat{\theta}) = \mathbb{E}_{p(\theta|X)}(\hat{\theta} - \theta)^2 = \hat{\theta}^2 - 2\hat{\theta}\mathbb{E}[\theta] + \mathbb{E}[\theta^2]$, where $\mathbb{E}$ denotes expecation w.r.t. the posterior $p(\theta|X)$. The optimal $\hat{\theta}$ is found by minimizing it $\varphi(\hat{\theta})$ w.r.t. $\hat{\theta}$. It is easy to see even without taking derivatives that the minimizer $\hat{\theta} = \mathbb{E}[\theta]$, is simply the posterior mean!

   For the absolute loss case, we can write the expression of the Bayesian expected loss as

$$\varphi(\hat{\theta}) = \int_{-\infty}^{\infty} |\hat{\theta} - \theta| p(\theta|X)\text{d}\theta = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta)p(\theta|X)\text{d}\theta - \int_{\hat{\theta}}^{\infty} (\hat{\theta} - \theta)p(\theta|X)\text{d}\theta$$

   You can take derivative and set it to zero to find the optimal $\hat{\theta}$. But you don't actually have to do all that if you note that if we just set $\hat{\theta}$ to the median of $p(\theta|X)$, due to the symmetry of the above expression, the Bayesian expected loss $\varphi(\hat{\theta})$ will attain its minima (which is equal to 0).

5. The Bayesian Information Criterion (BIC) is sometimes used for model selection, which is the problem of choosing the best model from a set of models $\{\mathcal{M}_i\}$. The BIC for model $\mathcal{M}_i$ is defined as

$$\text{BIC}_i = \log p(\mathcal{D}|\hat{\theta}_i) - \frac{M}{2}\log N$$

where $\hat{\theta}_i$ denotes the MAP estimate for the parameters of model $\mathcal{M}_i$, M is the dimension of $\hat{\theta}_i$, N is the number of observations from this model, $\mathcal{D}$ denotes the observed data, and $p(\mathcal{D}|\hat{\theta}_i)$ denotes the likelihood's value at the MAP estimate.

Show that, under certain assumptions, doing model selection using an approximation of the model evidence $p(\mathcal{D}|\mathcal{M}_i)$ computed using a Laplace approximation of the model's posterior is equivalent to using the BIC criterion defined above. Also state clearly the assumption(s) necessary to establish the equivalence with the BIC criterion.

Answer: Since the posterior is approximated using the Laplace approximation, we have

$$p(\theta_i|\mathcal{D},\mathcal{M}_i) = \frac{p(\mathcal{D}|\theta_i,\mathcal{M}_i)p(\theta_i|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)} \approx \mathcal{N}(\theta_i|\hat{\theta}_i,\mathbf{H}^{-1}) = \frac{1}{\sqrt{(2\pi)^M|\mathbf{H}^{-1}|}}\exp\left[(\theta_i-\hat{\theta}_i)^\top\mathbf{H}(\theta_i-\hat{\theta}_i)\right]$$

where we are explicitly showing that we are conditioning on the model $\mathcal{M}_i$.

Plugging in $\theta_i = \hat{\theta}_i$, we get the following approximation for the model evidence $p(\mathcal{D}|\mathcal{M}_i)$

$$p(\mathcal{D}|\mathcal{M}_i) \approx \sqrt{(2\pi)^M|\mathbf{H}^{-1}|} \times p(\mathcal{D}|\hat{\theta}_i,\mathcal{M}_i)p(\hat{\theta}_i|\mathcal{M}_i)$$

Taking log on both sides (log is monotonic), we can write (ignore conditioning on $\mathcal{M}_i$ on the R.H.S.)

$$\log p(\mathcal{D}|\mathcal{M}_i) \approx \log p(\mathcal{D}|\hat{\theta}_i) + \log p(\hat{\theta}_i) + \frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{H}|$$

The R.H.S. is now beginning to look like the expression for BIC_i. If we further assume a uniform prior, we can also ignore the log-prior term. If we also assume $\mathbf{H}$ to be full-rank, $\mathbf{H}$ is a summation over N, $\log|\mathbf{H}| \propto M\log N$, and we see that $\log p(\mathcal{D}|\mathcal{M}_i)$ roughly gives the same approximation as BIC_i.

4 Long Answer Problems (1 question, 16 marks)

1. The multivariate normal is given by $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$. Express it in form of an exponential family distribution, identify its natural parameters, sufficient statistics, and the log-partition function. You may use the fact $\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \text{trace}(\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}) = \text{trace}(\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1})$.

A conjugate prior for the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is given by the Normal Inverse-Wishart prior $\mathcal{NIW}$, denoted as $\mathcal{NIW}(\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{\mu}_0,\kappa_0,\boldsymbol{\Sigma}_0,\nu_0) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}/\kappa_0)\mathcal{W}^{-1}(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_0,\nu_0)$, where $\mathcal{W}^{-1}$ denotes the inverse Wishart distribution. The overall (somewhat messy looking) expression for the $\mathcal{NIW}$ is given by

$$\frac{\kappa_0^{\frac{D}{2}}|\boldsymbol{\Sigma}_0|^{\frac{\nu_0}{2}}|\boldsymbol{\Sigma}|^{-\frac{\nu_0+D+2}{2}}}{2^{\frac{(\nu_0+1)D}{2}}\pi^{\frac{D}{2}}\Gamma_D(\frac{\nu_0}{2})}\exp\left(-\frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \frac{1}{2}\text{trace}(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1})\right)$$

where κ_0,ν_0 are positive scalars, $\boldsymbol{\mu}_0 \in \mathbb{R}^D$ and $\boldsymbol{\Sigma}_0$ is $D\times D$ a symmetric positive definite matrix, and Γ_D denotes the multivariate gamma function. Express this $\mathcal{NIW}$ distribution in form of an exponential family distribution, identify its natural parameters, sufficient statistics, and the log-partition function.

Now suppose we are given N observations $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$ generated from $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. Derive the form of the posterior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Specifically, write down the expressions for the posterior updates of the hyperparameters $\kappa_0,\nu_0,\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0$. Note that you don't need to do this part all from scratch; you should be able to readily obtain this part of the solution using the exponential family representations obtained for the other parts of this question.