Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: March 1, 2021

Solution 1

We are given that $x \sim N(0, \eta)$, hence,

$$p(x \mid \eta) = \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}x^2) \quad x \in \mathbb{R} \tag{1}$$

Also, it is given that $\eta \sim \text{Exp}(\frac{\gamma^2}{2})$ with $\gamma > 0$, hence,

$$p(\eta \mid \gamma) = \frac{\gamma^2}{2} \exp(-\frac{\gamma^2}{2}\eta) \quad \eta > 0 \tag{2}$$

Hence, we have the pdf. for marginal distribution of x given by

$$\begin{aligned} p(x \mid \gamma) &= \int_{\eta>0} p(x, \eta \mid \gamma)\, d\eta \\ &= \int_{\eta>0} p(x, \mid \eta) p(\eta \mid \gamma)\, d\eta \\ &= \int_{\eta>0} \frac{\gamma^2}{2\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}x^2 - \frac{\gamma^2}{2}\eta)\, d\eta \end{aligned} \tag{3}$$

This integral is hard to compute. We will try to identify the marginal distribution of x using the method of **Moment Generating Function** (MGF) and using the fact that MGFs are unique for each distribution.

The Moment Generating Function for the marginal distribution of x is given by

$$\begin{aligned} M_X(t) &= E_X(e^{tx}) \\ &= \int_{x=-\infty}^{\infty} e^{tx} p(x \mid \gamma)\, dx \\ &= \int_{x=-\infty}^{\infty} e^{tx} \Big\{ \int_{\eta>0} \frac{\gamma^2}{2\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}x^2 - \frac{\gamma^2}{2}\eta)\, d\eta \Big\}\, dx && \text{(using (3))} \\ &= \int_{x=-\infty}^{\infty} \int_{\eta>0} \frac{\gamma^2}{2\sqrt{2\pi\eta}} \exp(-\frac{\gamma^2}{2}\eta) \exp(-\frac{1}{2\eta}\big(x^2 - 2x\, t\eta)\big)\, d\eta\, dx \\ &= \int_{x=-\infty}^{\infty} \int_{\eta>0} \frac{\gamma^2}{2\sqrt{2\pi\eta}} \exp\big(-\frac{\eta(\gamma^2 - t^2)}{2}\big) \exp(-\frac{1}{2\eta}\big(x - t\eta)^2\big)\, d\eta\, dx \\ &= \int_{\eta>0} \frac{\gamma^2}{2} \exp\big(-\frac{\eta(\gamma^2 - t^2)}{2}\big) \Big\{ \int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}\big(x - t\eta)^2\big)\, dx \Big\}\, d\eta \end{aligned}$$

Note that

$$\int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}(x-t\eta)^2)\ dx = 1 \tag{4}$$

$$\int_{\eta>0} (\frac{\gamma^2-t^2}{2}) \exp\big(-\frac{\eta(\gamma^2-t^2)}{2}\big)\ d\eta = 1 \tag{5}$$

Hence, we get,

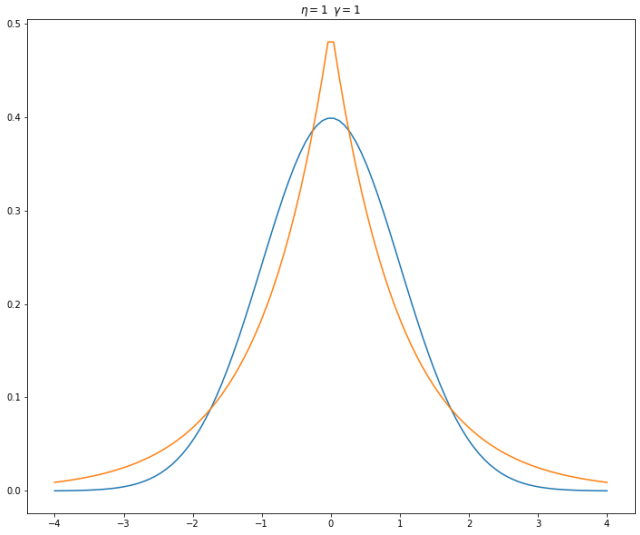
$$\begin{aligned}
M_X(t) &= \int_{\eta>0} \frac{\gamma^2}{2} \exp\big(-\frac{\eta(\gamma^2-t^2)}{2}\big)\Big\{ \int_{x=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\eta}} \exp(-\frac{1}{2\eta}(x-t\eta)^2)\ dx \Big\}\ d\eta \\
&= \frac{\gamma^2}{2}(\frac{\gamma^2-t^2}{2})^{-1} \int_{\eta>0} (\frac{\gamma^2-t^2}{2}) \exp\big(-\frac{\eta(\gamma^2-t^2)}{2}\big)\ d\eta && \text{(using (4))} \\
&= \frac{\gamma^2}{2}(\frac{\gamma^2-t^2}{2})^{-1} \text{ where } \frac{\gamma^2-t^2}{2} > 0 && \text{(using (5))} \\
&= \frac{1}{1-(t/\gamma)^2} \text{ where } |t| < \gamma
\end{aligned}$$

This compares with the MGF of Laplace distribution $L(\mu, b)$ where

$M_L(t) = \frac{e^{t\mu}}{1-b^2t^2}$ where $|t| < \frac{1}{b}$ with $\mu = 0$ and $b = \frac{1}{\gamma}$

and the required marginal distribution of x is Laplacian distribution $L(x \mid 0, \frac{1}{\gamma})$ with

$$p(x \mid \gamma) = \frac{\gamma}{2} \exp(-\gamma\,|x|) \quad \forall\ x \in \mathbb{R} \tag{6}$$



While one would expect the marginal likelihood to follow Gaussian, since the likelihood itself follows Gaussian distribution, this is not the case here due to non conjugacy of Exponential prior with Gaussian likelihood. Also the calculations were fairly involved and generally we look for conjugate priors for this reason.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 2

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: March 1, 2021

Solution 2

Let $M \in \mathbf{R}^{D\times D},\ \boldsymbol{v} \in \mathbf{R}^{D\times 1}$. We are given that,

$$(M + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = M^{-1} - \frac{(M^{-1}\boldsymbol{v})(\boldsymbol{v}^T M^{-1})}{1 + \boldsymbol{v}^T M^{-1}\boldsymbol{v}} \tag{1}$$

Note that,

$$\begin{aligned}
\Sigma_N &= (\beta \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T + \lambda I_D)^{-1} \\
&= \frac{1}{\lambda}(\frac{\beta}{\lambda} \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T + I_D)^{-1} \qquad (2) \\
&= \frac{1}{\lambda}(\frac{\beta}{\lambda} \boldsymbol{x}_N^T \boldsymbol{x}_N + \frac{\beta}{\lambda} \sum_{i=1}^{N-1} \boldsymbol{x}_i \boldsymbol{x}_i^T + \lambda I_D)^{-1} \qquad (3)
\end{aligned}$$

Let $f(N) = (\frac{\beta}{\lambda} \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T + I_D)^{-1}$, then, using (1), we can continue from (3) as follows

$$\Sigma_N = \frac{1}{\lambda}\big[\ f(N-1) - \frac{\big[f(N-1)\boldsymbol{x}_N\big]\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1 + \boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N}\ \big] \tag{4}$$

Hence, using (4) and (2), we can further obtain

$$\begin{aligned}
\Sigma_N - \Sigma_{N-1} &= \frac{1}{\lambda}\big[\ f(N-1) - \frac{\big[f(N-1)\boldsymbol{x}_N\big]\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1 + \boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N}\ \big] - \frac{1}{\lambda}(\frac{\beta}{\lambda} \sum_{i=1}^{N-1} \boldsymbol{x}_i \boldsymbol{x}_i^T + I_D)^{-1} \\
&= \frac{1}{\lambda}\big[\ f(N-1) - \frac{\big[f(N-1)\boldsymbol{x}_N\big]\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1 + \boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N}\ \big] - \frac{1}{\lambda} f(N-1) \\
&= -\frac{1}{\lambda}\big[\ \frac{\big[f(N-1)\boldsymbol{x}_N\big]\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1 + \boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N}\ \big] \\
&= -\frac{1}{\lambda}\big[\ \frac{\big[\boldsymbol{x}_N^T f(N-1)^T\big]^T\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1 + \boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N}\ \big] \qquad (5)
\end{aligned}$$

Now note that $f(N) = (\frac{\beta}{\lambda} \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T + I_D)^{-1} = \lambda \Sigma_N \quad \forall\ N \in \mathbb{N}$

Σ_N being a covariance matrix is symmetric, positive semi-definite and since $\lambda > 0$, we have $f(N)$ as symmetric, positive semi-definite $\quad \forall\ N \in \mathbb{N}$ (6)

Using (5) and (6), we have

$$\Sigma_N - \Sigma_{N-1} = -\frac{1}{\lambda}\Big[\ \frac{\big[\boldsymbol{x}_N^T f(N-1)\big]^T\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1+\boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N}\ \Big] \tag{7}$$

Now, $f(N-1)$ is positive semi-definite from (6), hence, for arbitrary $\boldsymbol{x}_N \in \mathbb{R}^{D\times 1}$, we have,

$$1+\boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N \geq 1 \tag{8}$$

Also, it is easy to see that $\big[\boldsymbol{x}_N^T f(N-1)\big]^T\ \big[\boldsymbol{x}_N^T f(N-1)\big]\ \big(=Y,\ \text{say}\big)$ is positive semi-definite, since for arbitrary $\boldsymbol{x} \in \mathbb{R}^{D\times 1}$, we have (9)

$$\begin{aligned}\boldsymbol{x}^T Y \boldsymbol{x} &= \boldsymbol{x}^T\big[\boldsymbol{x}_N^T f(N-1)\big]^T\ \big[\boldsymbol{x}_N^T f(N-1)\big]\boldsymbol{x}\\ &= \big[\boldsymbol{x}_N^T f(N-1)\boldsymbol{x}\big]^T\ \big[\boldsymbol{x}_N^T f(N-1)\boldsymbol{x}\big]\\ &= \big|\big|\ \boldsymbol{x}_N^T f(N-1)\boldsymbol{x}\ \big|\big|_2 \geq 0\end{aligned}$$

From (8) and (9),

$$\frac{\big[\boldsymbol{x}_N^T f(N-1)\big]^T\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1+\boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N} \text{ is positive semi-definite}$$

$$\text{Since } \frac{-1}{\lambda} < 0,\ \text{we have, } -\frac{1}{\lambda}\Big[\ \frac{\big[\boldsymbol{x}_N^T f(N-1)\big]^T\ \big[\boldsymbol{x}_N^T f(N-1)\big]}{1+\boldsymbol{x}_N^T f(N-1)\boldsymbol{x}_N}\ \Big] \text{ as negative semi-definite}$$

$$\text{Thus, using (7),} \qquad \Sigma_N - \Sigma_{N-1} \text{ is negative semi-definite} \tag{10}$$

Now, we have for arbitrary $\boldsymbol{x}_* \in \mathbb{R}^{D\times 1}$

$$\begin{aligned}\sigma_N^2(\boldsymbol{x}_*) - \sigma_{N-1}^2(\boldsymbol{x}_*) &= \big[\ \beta^{-1} + \boldsymbol{x}_*^T\Sigma_N\boldsymbol{x}_*\ \big] - \big[\ \beta^{-1} + \boldsymbol{x}_*^T\Sigma_{N-1}\boldsymbol{x}_*\ \big]\\ &= \boldsymbol{x}_*^T\big[\Sigma_N - \Sigma_{N-1}\big]\boldsymbol{x}_* \leq 0 \qquad \text{(using (10))}\end{aligned}$$

Thus,

$$\beta^{-1} \leq \sigma_N^2(\boldsymbol{x}_*) \leq \sigma_{N-1}^2(\boldsymbol{x}_*) \tag{11}$$

Thus, the variance of the predictive posterior decreases as the size of the training set, N, increases (12)

Also, we see that $\big\langle\sigma_N^2(\boldsymbol{x}_*)\big\rangle_N = \big\langle\beta^{-1} + \boldsymbol{x}_*^T\Sigma_N\boldsymbol{x}_*\big\rangle_N$ is a monotonically decreasing sequence bounded below by β^{-1}, with $\boldsymbol{x}_*^T\Sigma_N\boldsymbol{x}_* \to 0$ as $N\to\infty$ $\big($because of (10)$\big)$, thus,

$$\big\langle\sigma_N^2(\boldsymbol{x}_*)\big\rangle_N \to \beta^{-1} \text{ as } N\to\infty \tag{13}$$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: March 1, 2021

QUESTION 3

Solution 3

We are given that $x_i \sim N(\mu, \sigma^2) \,\big|_{i=1}^{N}$ are independent and identically distributed, hence

$$E(x_i) = \mu \,\big|_{i=1}^{N} \tag{1}$$
$$V(x_i) = \sigma^2 \,\big|_{i=1}^{N} \tag{2}$$
$$\text{Cov}(x_i, x_j) = 0 \,\big|_{i \neq j} \tag{3}$$

We want to derive the distribution of $\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

We know that any **linear combination** of independent normal random variables is also normally distributed and since it is easy to see that $\overline{x}$ is a linear combination of $x_i \,\big|_{i=1}^{N}$, hence,

$$\overline{x} \sim N(\eta, \gamma^2) \tag{4}$$

where $\eta = E(\overline{x})$ and $\sigma^2 = V(\overline{x})$ are to be evaluated as follows

$$\begin{aligned}
\eta &= E(\overline{x}) \\
&= E(\frac{1}{N} \sum_{i=1}^{N} x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} E(x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} \mu && \text{(from (1))} \\
&= \mu && (5)
\end{aligned}$$

$$
\begin{aligned}
\gamma^2 &= V(\overline{x}) \\
&= V(\frac{1}{N}\sum_{i=1}^{N} x_i) \\
&= \frac{1}{N^2}\{\sum_{i=1}^{N} V(x_i) + \sum_{i\neq j} \mathrm{Cov}(x_i, x_j)\} \\
&= \frac{1}{N^2}\{\sum_{i=1}^{N} \sigma^2 + 0\} \qquad \text{(from (2) and (3))} \\
&= \frac{\sigma^2}{N} \qquad (6)
\end{aligned}
$$

Hence, using (4), (5) and (6), we obtain,

$$\overline{x} \sim N(\mu, \frac{\sigma^2}{N}) \tag{7}$$

This makes intuitive sense since we expect the sample mean to be concentrated near μ and as our sample size (N) increases, we become more and more certain, i.e, $\sigma^2/N \to 0$ as $N \to \infty$

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 4

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: March 1, 2021

Solution 4

We are given that $\boldsymbol{x}^m = (x_1^m, x_2^m \dots x_{N_m}^m)$ satisfies for each $m \in \{1, 2 \ \dots \ M\}$, the following

$$x_j^m \mid \mu_m \overset{i.i.d}{\sim} N(\mu_m, \sigma^2) \qquad \forall\, j \in \{1, 2 \ \dots \ N_m\} \tag{1}$$

Further, it is given that

$$\mu_m \overset{i.i.d}{\sim} N(\mu_0, \sigma_0^2) \qquad \forall\, m \in \{1, 2 \ \dots \ M\} \tag{2}$$

Letting $\bar{\boldsymbol{x}}^m = \frac{1}{N_m} \sum_{i=1}^{N_m} x_i^m$, we have from (7) of Question No. 3,

$$\bar{\boldsymbol{x}}^m \mid \mu_m \sim N(\mu_m, \sigma^2/N_m) \qquad \forall\, m \in \{1, 2 \ \dots \ M\} \tag{3}$$

Hence, with $\boldsymbol{x} = (\boldsymbol{x}^1, \boldsymbol{x}^2 \ \dots \ \boldsymbol{x}^m)$ and $\boldsymbol{\mu} = (\mu_1, \mu_2 \ \dots \ \mu_M)$, we have from (3)

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2, \mu_0, \sigma_0^2) &= \prod_{m=1}^{M} p(\bar{\boldsymbol{x}}^m \mid \mu_m, \sigma^2, \mu_0, \sigma_0^2) \\ &= \prod_{m=1}^{M} N(\bar{\boldsymbol{x}}^m \mid \mu_m, \sigma^2/N_m) \end{aligned} \tag{4}$$

Also, using (2), we get

$$p(\boldsymbol{\mu} \mid \mu_0, \sigma_0^2) = \prod_{m=1}^{M} N(\mu_m \mid \mu_0, \sigma_0^2) \tag{5}$$

Hence,

$$\begin{aligned} p(\boldsymbol{\mu} \mid \boldsymbol{x}, \mu_0, \sigma_0^2) &\propto p(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2, \mu_0, \sigma_0^2)\ p(\boldsymbol{\mu} \mid \mu_0, \sigma_0^2) \\ &\propto \prod_{m=1}^{M} N(\bar{\boldsymbol{x}}^m \mid \mu_m, \sigma^2/N_m) \prod_{m=1}^{M} N(\mu_m \mid \mu_0, \sigma^2) \\ &\propto \prod_{m=1}^{M} N(\bar{\boldsymbol{x}}^m \mid \mu_m, \sigma^2/N_m)\ N(\mu_m \mid \mu_0, \sigma^2) \\ &= \prod_{m=1}^{M} N(\mu_m \mid \mu_{N_m}, \sigma_{N_m}^2) \end{aligned} \tag{6}$$

where

$$\frac{1}{\sigma_{N_m}^2} = \frac{1}{\sigma_0^2} + \frac{N_m}{\sigma^2} \tag{7}$$

$$\mu_{N_m} = \frac{\sigma^2}{N_m \sigma_0^2 + \sigma^2} \mu_0 + \frac{N_m \sigma_0^2}{N_m \sigma_0^2 + \sigma^2} \bar{\boldsymbol{x}}^m \tag{8}$$

Therefore, the required posterior for μ_m is given by

$$N(\mu_m \mid \frac{\sigma^2}{N_m\sigma_0^2+\sigma^2}\mu_0 + \frac{N_m\sigma_0^2}{N_m\sigma_0^2+\sigma^2}\bar{\boldsymbol{x}}^m, (\frac{1}{\sigma_0^2}+\frac{N_m}{\sigma^2})^{-1}) \tag{9}$$

The marginal likelihood is given by

$$\begin{aligned} p(\boldsymbol{x} \mid \sigma^2, \mu_0, \sigma_0^2) &= \int_{\boldsymbol{\mu}} p(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2, \mu_0, \sigma_0^2)\; p(\boldsymbol{\mu} \mid \mu_0, \sigma_0^2)\; d\boldsymbol{\mu} \\ &= \prod_{m=1}^{M} \int_{\mu_m} p(\bar{\boldsymbol{x}}^m \mid \mu_m, \sigma^2, \mu_0, \sigma_0^2)\; p(\mu_m \mid \mu_0.\sigma_0^2)\; d\mu_m \\ &= \prod_{m=1}^{M} \int_{\mu_m} N(\bar{\boldsymbol{x}}^m \mid \mu_m, \sigma^2/N_m)\; N(\mu_m \mid \mu_0, \sigma_0^2)\; d\mu_m \\ &= \prod_{m=1}^{M} N(\bar{\boldsymbol{x}}^m \mid \mu_0, \sigma_0^2 + \sigma^2/N_m) \end{aligned} \tag{10}$$

Now, in order to use MLE 2 for estimation of μ_0, we consider the log of marginal likelihood and try to maximise it w.r.t μ_0

$$\begin{aligned} \ln p(\boldsymbol{x} \mid \sigma^2, \mu_0, \sigma_0^2) &= \ln \prod_{m=1}^{M} N(\bar{\boldsymbol{x}}^m \mid \mu_0, \sigma_0^2 + \sigma^2/N_m) \\ &= \sum_{m=1}^{M} \ln N(\bar{\boldsymbol{x}}^m \mid \mu_0, \sigma_0^2 + \sigma^2/N_m) \\ &\propto -\sum_{m=1}^{M} \frac{(\bar{\boldsymbol{x}}^m - \mu_0)^2}{\sigma_0^2 + \sigma^2/N_m} = g(\mu_0) \quad \text{(say)} \end{aligned}$$

Then,

$$\begin{aligned} \nabla_{\mu_0} g(\mu_0) &= 0 \\ \rightarrow \sum_{m=1}^{M} \frac{2(\bar{\boldsymbol{x}}^m - \mu_0)}{\sigma_0^2 + \sigma^2/N_m} &= 0 \\ \rightarrow \mu_0 \sum_{m=1}^{M} \frac{1}{\sigma_0^2 + \sigma^2/N_m} &= \sum_{m=1}^{M} \frac{\bar{\boldsymbol{x}}^m}{\sigma_0^2 + \sigma^2/N_m} \end{aligned}$$

Observing that, $\nabla^2 g(\mu_0) = -2 \sum_{m=1}^{M} \frac{1}{\sigma_0^2+\sigma^2/N_m} < 0$, we get the required estimate for μ_0 as

$$\hat{\mu_0}\big|_{MLE\ 2} = (\sum_{m=1}^{M} \frac{\bar{\boldsymbol{x}}^m}{\sigma_0^2 + \sigma^2/N_m}) / (\sum_{m=1}^{M} \frac{1}{\sigma_0^2 + \sigma^2/N_m}) \tag{11}$$

Using the MLE 2 estimate for μ_0 in posterior of μ_m by substituting (11) in (9), we observe the resultant posterior mean of μ_m is

$$\frac{\sigma^2}{N_m\sigma_0^2+\sigma^2}\Big(\sum_{m=1}^{M}\frac{\bar{\boldsymbol{x}}^m}{\sigma_0^2+\sigma^2/N_m}\Big)\Big/\Big(\sum_{m=1}^{M}\frac{1}{\sigma_0^2+\sigma^2/N_m}\Big)+\frac{N_m\sigma_0^2}{N_m\sigma_0^2+\sigma^2}\bar{\boldsymbol{x}}^m \tag{12}$$

The benefit of using this, in comparison with the posterior mean of (9), is that the posterior mean in (12) incorporates the data from all M schools whereas the one in (9) doesn't.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: March 1, 2021

QUESTION

5

Solution 5

We have for $\boldsymbol{y}^m \in \mathbb{R}^{N_m \times 1}$ and $X^m \in \mathbb{R}^{N_m \times D}$

$$p(\boldsymbol{y}^m \mid X^m, \boldsymbol{w}_m) = N(\boldsymbol{y}^m \mid X^m \boldsymbol{w}_m, \beta^{-1} I_{N_m}) \qquad \forall m \in \{1, 2 \ \ldots \ M\} \tag{1}$$

Hence, with $\boldsymbol{y}$, X, and $\boldsymbol{w}$ denoting the collective responses, inputs and weight vectors respectively, we have,

$$\begin{aligned} p(\boldsymbol{y} \mid X, \boldsymbol{w}) &= \prod_{m=1}^{M} p(\boldsymbol{y}^m \mid X^m, \boldsymbol{w}_m) \\ &= \prod_{m=1}^{M} N(\boldsymbol{y}^m \mid X^m \boldsymbol{w}_m, \beta^{-1} I_{N_m}) \end{aligned} \tag{2}$$

Further, it is also given that $p(\boldsymbol{w_d}) = N(\boldsymbol{w}_m \mid \boldsymbol{w}_0, \lambda^{-1} I_D)$, therefore,

$$\begin{aligned} p(\boldsymbol{w}) &= \prod_{m=1}^{M} p(\boldsymbol{w_d}) \\ &= \prod_{m=1}^{M} N(\boldsymbol{w}_m \mid \boldsymbol{w}_0, \lambda^{-1} I_D) \end{aligned} \tag{3}$$

We will first find the marginal likelihood $p(\boldsymbol{y} \mid X)$ in order to arrive at the required MLE 2 objective

$$\begin{aligned} p(\boldsymbol{y} \mid X) &= \int_{\boldsymbol{w}} p(\boldsymbol{y} \mid X, \boldsymbol{w}) \, p(\boldsymbol{w}) \, d\boldsymbol{w} \\ &= \prod_{m=1}^{M} \int_{\boldsymbol{w}^m} N(\boldsymbol{y}^m \mid X^m \boldsymbol{w}_m, \beta^{-1} I_{N_m}) \, N(\boldsymbol{w}_m \mid \boldsymbol{w}_0, \lambda^{-1} I_D) \, d\boldsymbol{w}^m \\ &= \prod_{m=1}^{M} N(\boldsymbol{y}^m \mid X^m \boldsymbol{w}_0, \lambda^{-1} X^m X^{mT} + \beta^{-1} I_{N_m}) \end{aligned}$$

Further,

$$\begin{aligned} \ln p(\boldsymbol{w} \mid \boldsymbol{y}, X) &= \ln \prod_{m=1}^{M} N(\boldsymbol{y}^m \mid X^m \boldsymbol{w}_0, \lambda^{-1} X^m X^{mT} + \beta^{-1} I_{N_m}) \\ &= \sum_{m=1}^{M} \ln N(\boldsymbol{y}^m \mid X^m \boldsymbol{w}_0, \lambda^{-1} X^m X^{mT} + \beta^{-1} I_{N_m}) \end{aligned} \tag{4}$$

Thus, the required MLE 2 objective is,

$$\hat{\boldsymbol{w}}_0\big|_{MLE} = \arg\max_{\boldsymbol{w}_0} \sum_{m=1}^{M} \ln N(\boldsymbol{y}^m \mid X^m \boldsymbol{w}_0, \lambda^{-1} X^m X^{mT} + \beta^{-1} I_{N_m}) \tag{5}$$

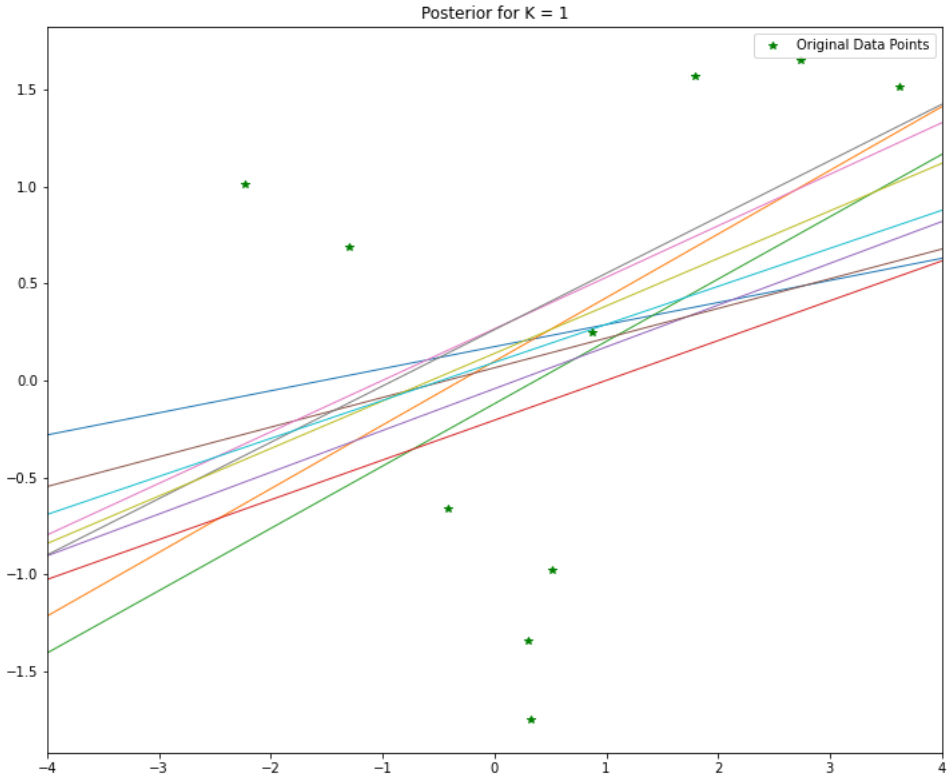
This has obvious benefit over fixing $\boldsymbol{w}_0$ to some value since this approach incorporates the knowledge of complete data, i.e, data from all schools, for estimation of $\boldsymbol{w}_0$. Once having derived the posterior $p(\boldsymbol{w} \mid \boldsymbol{y}, X)$, we can plug in estimation of $\boldsymbol{w}_0$, i.e, $\hat{\boldsymbol{w}}_0\big|_{MLE}$.

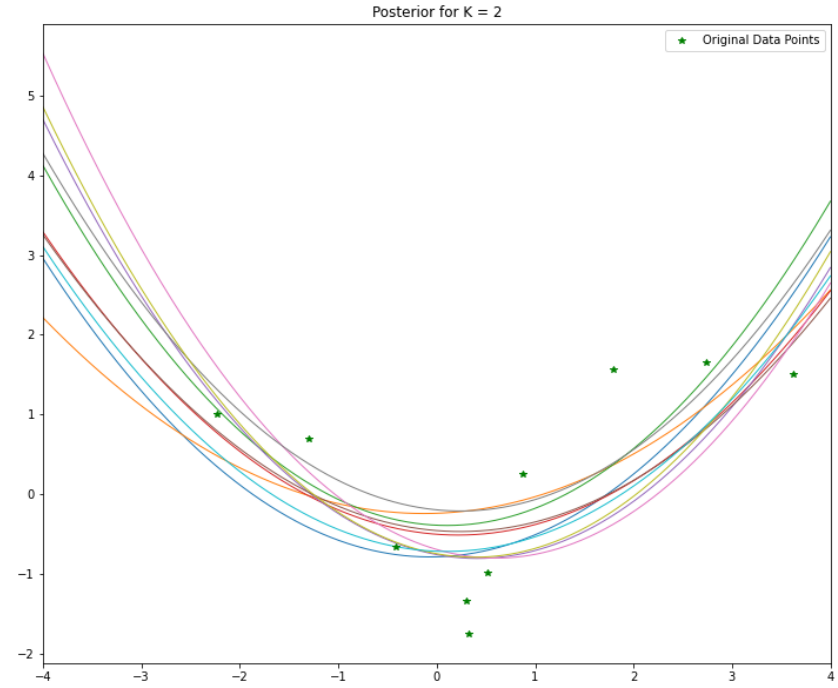
QUESTION
6

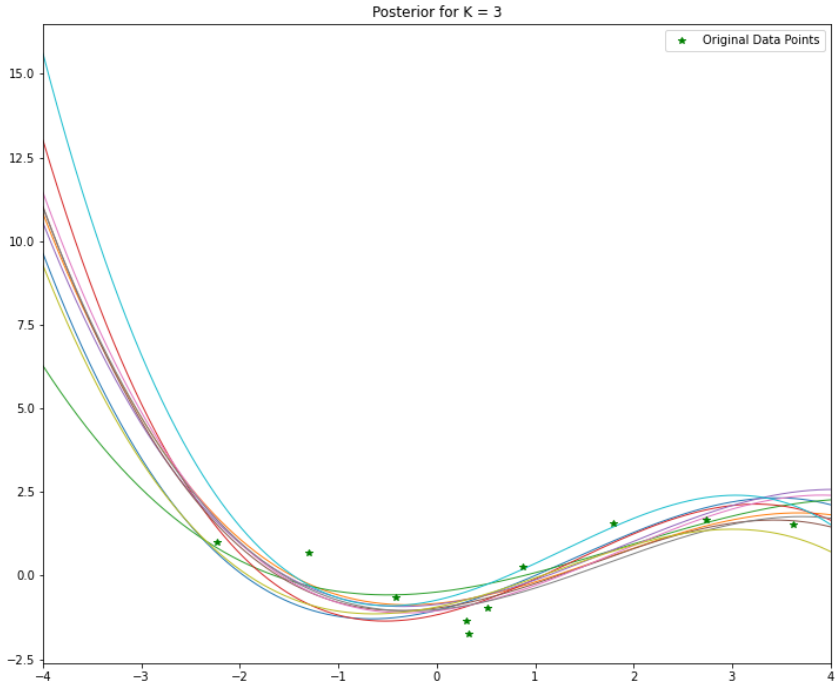
Student Name: Vishweshwar Tyagi
Roll Number: 191173
Date: March 1, 2021

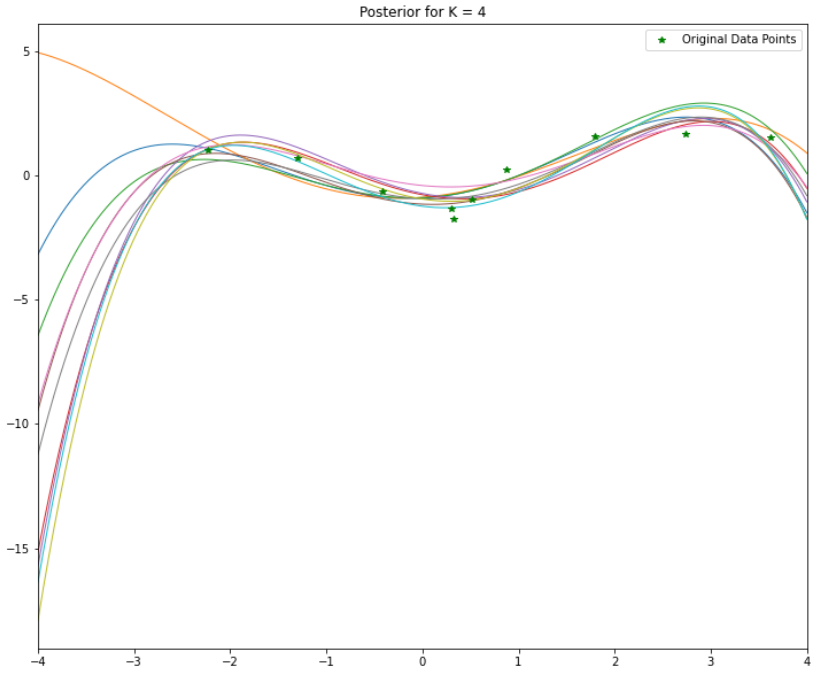
Solution 6

a)

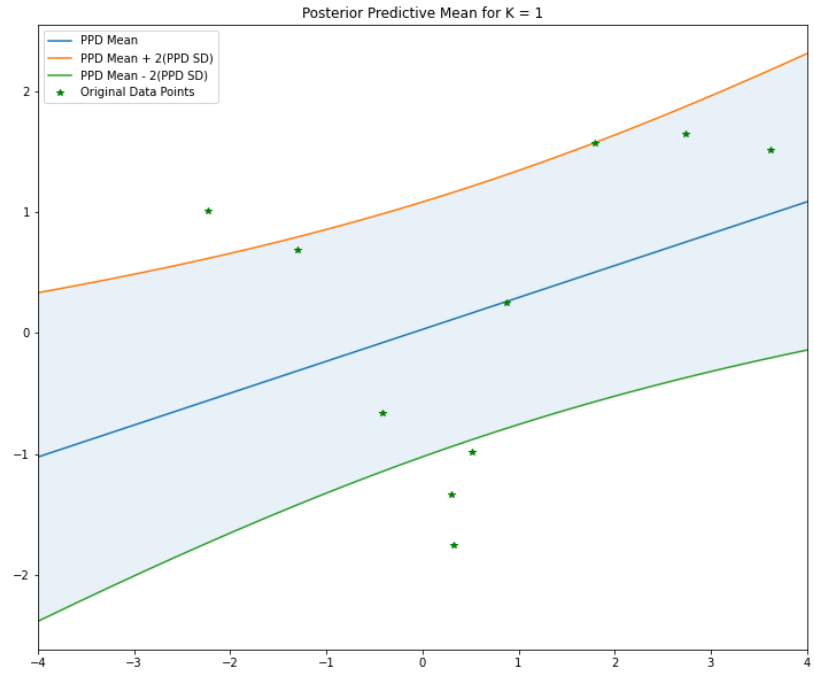


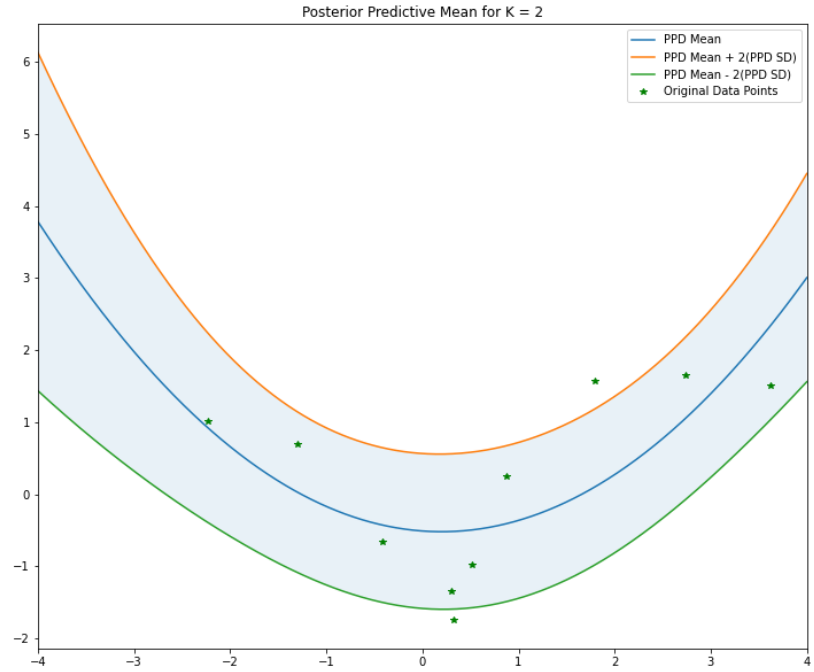

Posterior for K = 2
Original Data Points
5
4
3
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4

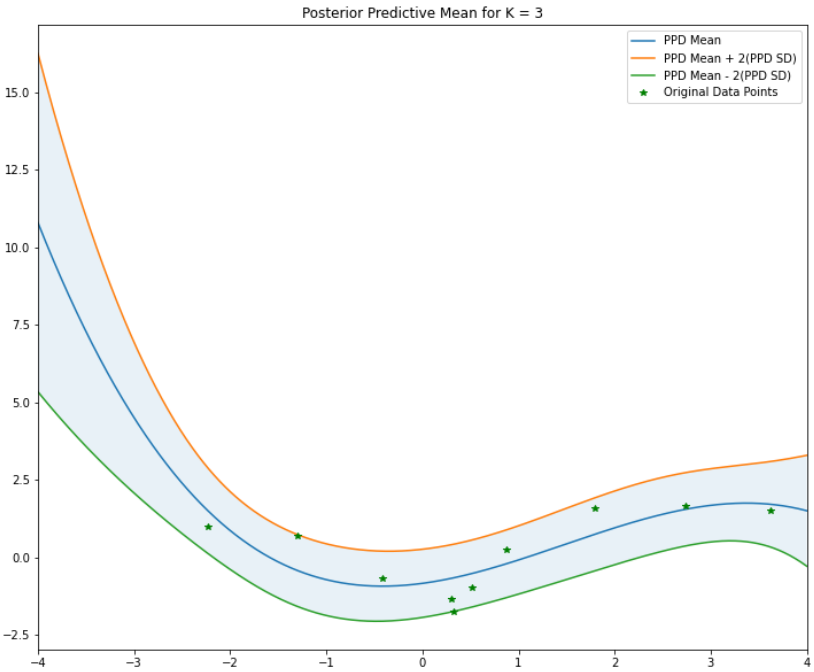

Posterior for K = 3
Original Data Points
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−3
−2
−1
0
1
2
3
4

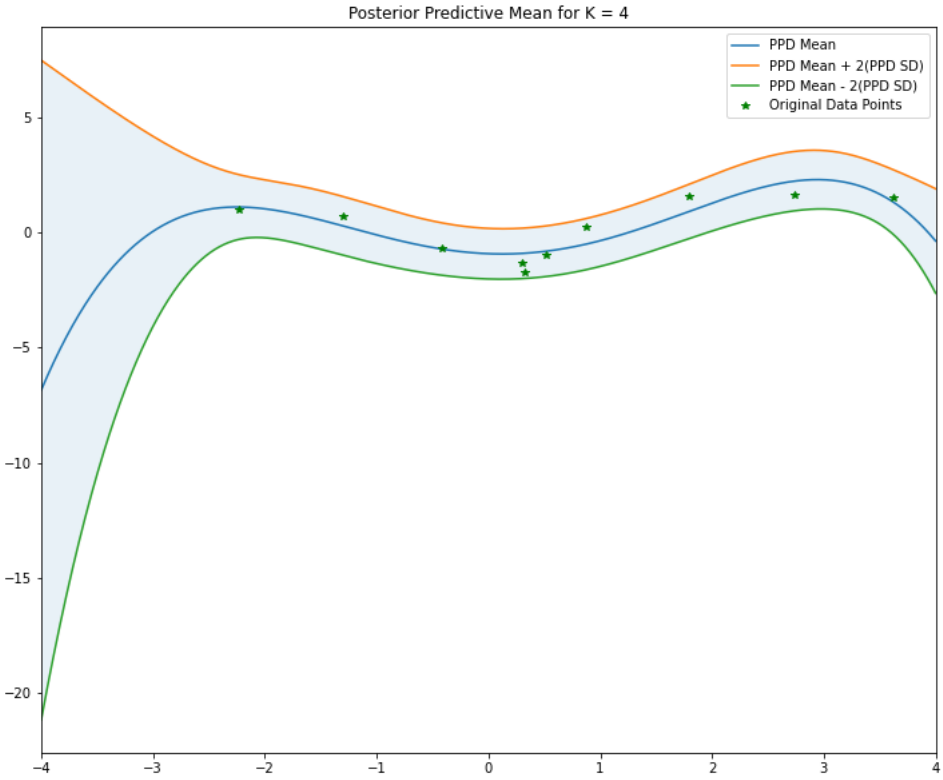

Posterior for K = 4
Original Data Points
5
0
−5
−10
−15
−4
−3
−2
−1
0
1
2
3
4

b)


Posterior Predictive Mean for K = 1
PPD Mean
PPD Mean + 2(PPD SD)
PPD Mean - 2(PPD SD)
Original Data Points
2
1
0
−1
−2
−4
−3
−2
−1
0
1
2
3
4


Posterior Predictive Mean for K = 2
PPD Mean
PPD Mean + 2(PPD SD)
PPD Mean - 2(PPD SD)
Original Data Points


Posterior Predictive Mean for K = 3
PPD Mean
PPD Mean + 2(PPD SD)
PPD Mean - 2(PPD SD)
Original Data Points



c) We have the marginal likelihood given by

$$p(\boldsymbol{y} \mid \phi(\boldsymbol{x})) = N(\boldsymbol{y} \mid \mathbf{0}, \phi(\boldsymbol{x})\phi(\boldsymbol{x})^T + \beta^{-1} I_M)$$

where M denotes the total number of training examples.

We obtained the following values for log marginal likelihood

For $K = 1$, the log marginal likelihood is -32.3520152804452
For $K = 2$, the log marginal likelihood is -22.772153178782283
For $K = 3$, the log marginal likelihood is -22.079070642241916
For $K = 4$, the log marginal likelihood is -22.386776180349212

It is easy to see that the model with $\mathbf{K = 3}$ has the highest log marginal likelihood and therefore seems to best explain the data.

d) The posterior mean for $\boldsymbol{w}$ is given by

$$\left(\phi(\boldsymbol{x})\phi(\boldsymbol{x})^T + \beta^{-1} I_D\right)^{-1} X^T \boldsymbol{y}$$

where $D = K + 1$ is the dimension of ϕ map.

Now, the MAP estimate for $\boldsymbol{w}$ is given by the mode of posterior but since we know that the posterior follows normal distribution due to conjugacy, the mode is equal to the mean and hence,

$$\hat{\boldsymbol{w}}\big|_{MAP} = \left(\phi(\boldsymbol{x})\phi(\boldsymbol{x})^T + \beta^{-1} I_D\right)^{-1} X^T \boldsymbol{y}$$

We obtained the following values for log likelihood using **MAP** estimate for the weight vector $\hat{\boldsymbol{w}}_{MAP}$

For $K = 1$, the log likelihood is -28.094004379075553
For $K = 2$, the log likelihood is -15.360663659052214
For $K = 3$, the log likelihood is -10.935846883615742
For $K = 4$, the log likelihood is -7.225291259028564

Using the **MAP** estimate, we see that the log likelihood of model with $\mathbf{K = 4}$ is the highest. This result doesn't agree with what we obtained above using the log marginal likelihood.

Using log marginal likelihood in order to select the best model would be a better criteria as compared to only using the log likelihood since the former doesn't rely on point based estimate (MAP estimate for weight vector in this case $\hat{\boldsymbol{w}}_{MAP}$) and instead does prior averaging over the weight vector.

e) It can be observed from the graph 'Posterior Predictive Mean for K = 3', the 2σ uncertainity gap seems to be the wides for $x \in [-4, -3]$ which makes sense as it lacks training examples in this region.

Hence, I would prefer $x' \in [-4, -3]$ in order to improve the model by reducing the uncertainity gap in this region.