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

QUESTION 1

Student Name: Abhas Kumar
Roll Number: 20111001
Date: February 26, 2021

Given, $p(\mathbf{x}|\eta) = \mathcal{N}(\mathbf{x}|0,\eta)$ and $p(\eta|\gamma) = Exp(\eta|\frac{\gamma^2}{2})$

The Marginal distribution of x, i.e., $p(\mathbf{x}|\gamma) = \int p(\mathbf{x}|\eta)p(\eta|\gamma)d\eta$

Moment generating function of any random variable $\mathbf{X}$ having pdf $f(x)$,$\mathcal{M}_X[t] = \mathbb{E}[e^{tx}] = \int_{-\infty}^{\infty} e^{tx}f(x)dx$

Hence the moment generating funtion for the given marginal likelihood can be written as

$$\int_\eta \int_x e^{tx}p(\mathbf{x}|\eta)p(\eta|\gamma)dxd\eta = \int_0^\infty \int_{-\infty}^\infty e^{tx}\mathcal{N}(\mathbf{x}|0,\eta)Exp(\eta|\frac{\gamma^2}{2})dxd\eta$$

$$= \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\eta}} \exp(tx) \exp(-\frac{x^2}{2\eta}) * \frac{\gamma^2}{2} \exp(-\frac{\gamma^2\eta}{2}) dxd\eta \quad .$$

$$= \tfrac{\gamma^2}{2} \int_0^\infty \int_{-\infty}^\infty \tfrac{1}{\sqrt{2\pi\eta}} \exp\left[tx - \tfrac{x^2}{2\eta} - \tfrac{\gamma^2\eta}{2}\right] dxd\eta$$

$$= \frac{\gamma^2}{2} \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\eta}} \exp\left[\frac{-(-2tx\eta + x^2 + \gamma^2\eta^2 - t^2\eta^2 + t^2\eta^2)}{2\eta}\right] dxd\eta$$

$$= \frac{\gamma^2}{2} \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\eta}} \exp\left(\frac{-(x-t\eta)^2}{2\eta}\right) \exp\left(\frac{-(\gamma^2-t^2)\eta^2}{2\eta}\right) dxd\eta$$

$$= \frac{\gamma^2}{2} \int_0^\infty \int_{-\infty}^\infty \mathcal{N}(x|t\eta,\eta) \exp\left(\frac{-(\gamma^2-t^2)\eta^2}{2\eta}\right) dxd\eta$$

$$= \frac{\gamma^2}{2} \int_0^\infty \exp\left(\frac{-(\gamma^2-t^2)\eta}{2}\right) d\eta \qquad as\left(\int_{-\infty}^\infty \mathcal{N}(x|t\eta,\eta)dx = 1\right)$$

$$= \frac{\gamma^2}{2} \frac{\left|\exp\left(\frac{-(\gamma^2-t^2)\eta}{2}\right)\right|_0^\infty}{\frac{-(\gamma^2-t^2)}{2}} = \frac{\gamma^2}{\gamma^2-t^2} = \frac{1}{1-(\frac{1}{\gamma})^2t^2}$$

Comparing this with moment generating function of Laplace, $\mathcal{L}(\mu,b)$ i.e $\frac{e^t\mu}{1-b^2t^2}$ we can observe that marginal likelihood, $p(\mathbf{x}|\gamma)$ has Laplace distribution $\mathcal{L}(0,\frac{1}{\gamma})$.

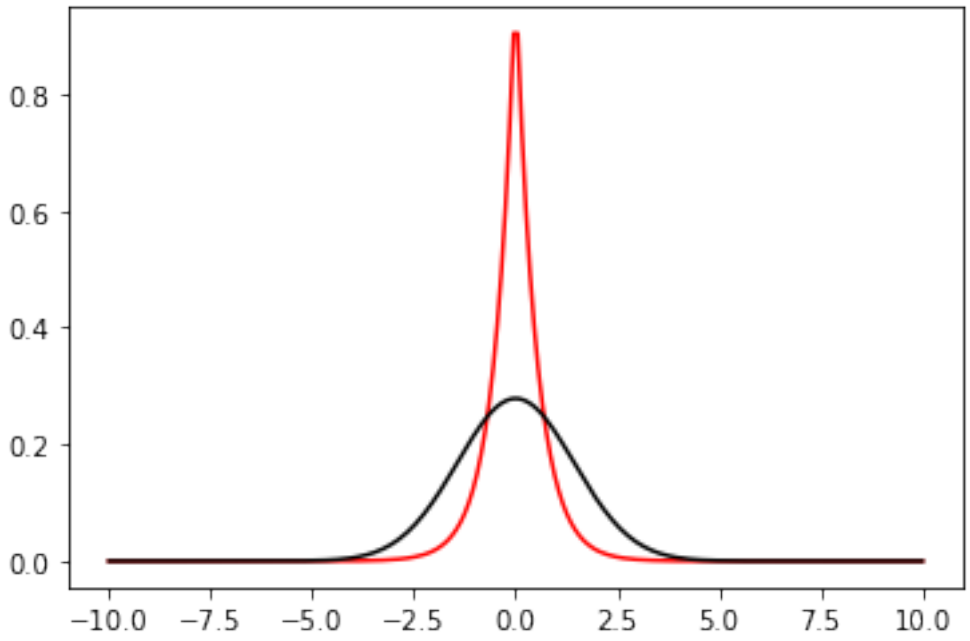

Figure 1: distribution of $p(x|\eta)$(red curve) and $p(\eta|\gamma)$(black curve) for $\gamma = 2$.

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

QUESTION 2

Student Name: Abhas Kumar
Roll Number: 20111001
Date: February 26, 2021

Since posterior can be regarded as the prior for the next observation,by considering an additional data point $(x_{n+1}, y_{n+1}) = (x_*, y_*)$ the resulting posterior can be estimated as .

$$p(w) = \mathcal{N}(w|\mu_n, \sum\nolimits_n)$$

$$p(y_{n+1}|x_{n+1}, w) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\beta}{2}(y_{n+1} - w^T x_{n+1})^2\right)$$

$$\therefore \text{posterior } p(w|y_{n+1}, x_{n+1}, \mu_n, \sum\nolimits_n)$$

$$\propto \exp\left(\frac{-1}{2}(w-\mu_n)^T(\sum\nolimits_n)^{-1}(w-\mu_n) - \frac{1}{2}\beta(y_{n+1} - w^T x_{n+1})\right)$$

ignoring the factor $\frac{-1}{2}$

$$\propto \exp\left[(w-\mu_n)^T(\sum\nolimits_n)^{-1}(w-\mu_n) + \beta(y_{n+1} - w^T x_{n+1})^2\right]$$

$$\propto \exp\left[w^T(\sum\nolimits_n)^{-1}w - 2w^T(\sum\nolimits_n)^{-1}\mu_n + \beta w^T x_{n+1}^T x_{n+1} - 2\beta w^T x_{n+1} x_{n+1}^T + \text{Constant}\right]$$

$$\propto \exp\left[w^T\left((\sum\nolimits_n)^{-1} + \beta x_{n+1}x_{n+1}^T\right)w - 2w^T\left((\sum\nolimits_n)^{-1}\mu_n + \beta x_{n+1}y_{n+1}\right)\right]$$

$$\therefore p(w|y_{n+1}, x_{n+1}, \mu_n, \sum\nolimits_n) = \mathcal{N}(w|\mu_{n+1}, \sum\nolimits_{n+1})$$

$$(\sum\nolimits_{n+1})^{-1} = (\sum\nolimits_n)^{-1} + \beta x_{n+1}x_{n+1}^T \text{ - - - - - - - - - - - - - - - - - - (1)}$$

Now , we can write ,

$$\sigma_{n+1}^2 x = \frac{1}{\beta} + x_{n+1}^T(\sum\nolimits_{n+1})x_{n+1}$$

Using equation-(1) we have ,

$$\sum\nolimits_{n+1} = \left((\sum\nolimits_n)^{-1} + \beta x_{n+1}x_{n+1}^T\right)^{-1}$$

Using the identity ,

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

Where M is a square matrix and v is a column vector , we can write

$$\sum_{n+1} = \sum_{n} - \frac{\left(\sum_n x_{n+1}\beta^{\frac{1}{2}}\right)\left(\beta^{\frac{1}{2}} x_{n+1}^T \sum_n\right)}{1 + \beta(x_{n+1}^T)\sum_n x_{n+1}}$$

$$= \sum_{n} - \frac{\beta \sum_n (x_{n+1})(x_{n+1}^T)\sum_n}{1 + \beta x_{n+1}^T \sum_n x_{n+1}}$$

Hence ,

$$\sigma_{n+1}^2(x) = \frac{1}{\beta} + x^T \left(\sum_{n} - \frac{\beta \sum_n x_{n+1} x_{n+1}^T \sum_n}{1 + \beta x_{n+1}^T \sum_n x_{n+1}} \right) x$$

$$= \frac{1}{\beta} + x^T \sum_n x - \frac{\beta x^T \sum_n x_{n+1} x_{n+1}^T \sum_n x}{1 + \beta x_{n+1}^T \sum_n x_{n+1}}$$

$$= \sigma_n^2(x) - \frac{\beta x^T \sum_n x_{n+1} x_{n+1}^T \sum_n x}{1 + \beta x_{n+1}^T \sum_n x_{n+1}}$$

Since $\sum_n$ is positive definite matrix,

$\beta x^T \sum_n x_{n+1} x_{n+1}^T \sum_n x >= 0$ and $1 + \beta x_{n+1}^T \sum_n x_{n+1} > 0$

$\sigma_{n+1}^2(x) <= \sigma_n^2(x)$, as the training examples increases variance of predicate posterior decreases.

As $n \to \infty$, variance is only due to the reciprocal of noise (β).

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

QUESTION 3

Student Name: Abhas Kumar
Roll Number: 20111001
Date: February 26, 2021

Empirical mean $\tilde{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = [\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}]^T[x_1, x_2, ..., x_N]$. Let $\mathbf{k} = [\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}]$ and random variable $\mathbf{v} = [x_1, x_2, ..., x_N]$ be $N \times 1$ vector.

Clearly,$\tilde{x}$ can be written as linear transformation of random variable $\mathbf{v}$ as $\tilde{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = \mathbf{k}^T\mathbf{v}$.

and , Since it is given that $x_i \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2), i = 1, 2..., N$.$\mathbf{v}$ is a Gaussian r.v with mean vector, $\mathbf{E}[\mathbf{v}] = \mu$ and co-variance matrix,$\mathbf{V}[\mathbf{v}] = \Sigma = \sigma^2 I_N$.

$\mathbf{E}[\tilde{x}] = \mathbf{E}[\mathbf{k}^T\mathbf{v}] = \mathbf{k}^T\mathbf{E}[\mathbf{v}] = \mathbf{k}^T\mu$

$= [\frac{1}{N}, ..., \frac{1}{N}]^T[\mu, \mu, ..., \mu] = \frac{N\mu}{N} = \mu$

$$\mathbf{V}[\tilde{x}] = \mathbf{V}[\mathbf{k}^T\mathbf{v}] = \mathbf{k}^T\mathbf{V}[\mathbf{v}]\mathbf{k} = [\frac{1}{N}, ..., \frac{1}{N}] \begin{pmatrix} \sigma^2 & 0 & . & . & 0 \\ 0 & \sigma^2 & . & . & 0 \\ . & 0 & . & . & . \\ 0 & 0 & . & . & \sigma^2 \end{pmatrix} [\frac{1}{N}, ..., \frac{1}{N}]^T$$

$= [\frac{\sigma^2}{N}, \frac{\sigma^2}{N}, ..., \frac{\sigma^2}{N}][\frac{1}{N}, ..., \frac{1}{N}]^T$

$= \frac{\sigma^2 N}{N^2} = \frac{\sigma^2}{N}$

As $\tilde{x}$ is a linear transformation of a Gaussian r.v , its probability distribution will also be a Gaussian with above derived mean and variance i.e $\mathcal{N}(\tilde{x}|\mu, \frac{\sigma^2}{N})$.

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

QUESTION

4

Student Name: Abhas Kumar
Roll Number: 20111001
Date: February 26, 2021

---

1. Posterior of u_m with given problem statement can be written as:

$$p(\mu_m|x^{(m)}, \mu_0, \sigma_0^2) = \frac{p(x^{(m)}|\mu_m, \sigma^2) * p(\mu_m|u_0, \sigma_0^2)}{\int p(x^{(m)}|\mu_m, \sigma^2) * p(\mu_m|\mu_0, \sigma_0^2) d\mu_m}$$
$$\propto p(x^{(m)}|\mu_m, \sigma^2) * p(\mu_m|\mu_0, \sigma_0^2)$$

likelihood: $p(x_n^{(m)}|\mu_m, \sigma^2) = \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2)$

Since $x_n^{(m)} \overset{i.i.d}{\sim} \mathcal{N}(\mu_m, \sigma^2)$

$p(x^{(m)}|\mu_m, \sigma^2) = \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2)$

Prior : $p(\mu_m|\mu_0, \sigma_0^2) = \mathcal{N}(\mu_m, \mu_0, \sigma^2)$

$$\therefore p(\mu_m|x^{(m)}, \mu_0, \sigma_0^2) \propto \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) * \mathcal{N}(\mu_m, \mu_0, \sigma_0^2)$$
$$\propto \prod_{n=1}^{N_m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x_n^{(m)} - \mu_m)^2}{2\sigma^2}\right] * \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[\frac{-(\mu_m - \mu_0)^2}{2\sigma_0^2}\right]$$
$$\propto \frac{1}{2\pi\sigma\sigma_0} \exp\left[\frac{-\sum(x_n^{(m)} - \mu_m)^2}{2\sigma^2}\right] \exp\left[\frac{-(\mu_m - \mu_0)^2}{2\sigma_0^2}\right]$$

By using the "Completing the square- trick" and ignoring the constant terms, and the fact that likelihood and prior are conjugate, we get :

$p(\mu_m|x^{(m)}, \sigma^2) = \mathcal{N}(\mu_m|\mu_e, \sigma_e^2)$ Where,

$$\mu_e = \frac{\sigma^2}{\sigma^2 + N_m\sigma_0^2} u_0 + \frac{1}{N}\left(\frac{N_m\sigma_0^2}{\sigma^2 + N_m\sigma_0^2}\right) \sum_{n=1}^{N_m} x_n^{(m)} \ ,$$

$$\frac{1}{\sigma_e^2} = \frac{N_m}{\sigma^2} + \frac{1}{\sigma_0^2}$$

2. Marginal Likelihood : $p(x|\mu_0, \sigma^2, \sigma_0^2) = \int p(x|\mu, \sigma^2) p(\mu|\mu_0, \sigma_0^2) d\mu$

$$p(x|\mu, \sigma^2) = \prod_{m=1}^{M} p(x^{(m)}|\mu_m, \sigma^2)]$$
$$= \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(x_n^{(m)}|\mu_m, \sigma^2)$$
$$= \prod_{m=1}^{M} \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2)$$

also, $p(\mu|\mu_0, \sigma_0^2) = \prod_{m=1}^{M} p(\mu_m|\mu_0, \sigma_0^2)$

$$= \prod_{m=1}^{M} \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)$$

$$\therefore p(x|\mu_0, \sigma^2, \sigma_0^2) = \int \prod_{m=1}^{M} \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \prod_{m=1}^{M} \mathcal{N}(\mu_m|\mu_0, \sigma_0^2) d\mu_m$$

$$= \prod_{m=1}^{M} \int \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2) d\mu_m$$

from part-1 , we know that $p(\mu_m|x^{(m)}, \sigma^2) = \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)$

$$\mathcal{N}(\mu_m|\mu_e, \sigma_e^2) = \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)}{\int \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2) d\mu_m}$$

$$\int \prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2) d\mu_m = \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)}{\mathcal{N}(\mu_m|\mu_e, \sigma_e^2)}$$

$$\therefore p(x|\mu_0, \sigma^2, \sigma_0^2) = \prod_{m=1}^{M} \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)}{\mathcal{N}(\mu_m|\mu_e, \sigma_e^2)}$$

We know MLE-II is ,

$$\hat{\mu_0} = \underset{\mu_0}{\text{argmax}} \, (\text{Marginal likelihood})$$
$$= \underset{\mu_0}{argmax} \left\{ \prod_{m=1}^{M} \frac{\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_m, \sigma^2) \mathcal{N}(\mu_m|\mu_0, \sigma_0^2)}{\mathcal{N}(\mu_m|\mu_e, \sigma_e^2)} \right\}$$

On taking log marginal likelihood and ignoring the constants we get

$$= \underset{\mu_0}{argmin} \left\{ \sum_{m=1}^{M} \left[\frac{(\mu_m - \mu_0)^2}{2\sigma_0^2} - \frac{(\mu_m - \mu_e)^2}{2\sigma_e^2} \right] \right\}$$

Differentiating w.r.t μ_0 we get ,

$$\sum_{m=1}^{M} \mu_0 - \sum_{m=1}^{M} \left(\frac{\sigma^2}{\sigma^2 + N_m\sigma_0^2}\mu_0 + \frac{N_m\sigma_0^2}{\sigma^2 + N_m\sigma_0^2} * \frac{1}{N}\sum_{n=1}^{N_m} x_n^{(m)} \right) = 0$$

$$\sum_{m=1}^{M} \frac{N_m\sigma_0^2}{\sigma^2 + N_m\sigma_0^2}\mu_0 = \sum_{m=1}^{M} \frac{N_m\sigma_0^2}{\sigma^2 + N_m\sigma_0^2} * \frac{1}{N}\sum_{n=1}^{N_m} x_n^{(m)}$$

let $\bar{x}^{(m)} = \frac{1}{N}\sum_{n=1}^{N_m} x_n^{(m)}$,

$$\hat{\mu_0} = \frac{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m\sigma_0^2}\bar{x}^{(m)}}{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m\sigma_0^2}}$$

3. Posterior estimates on using $\hat{\mu_0}$

$$\mu_e = \left(\frac{\sigma^2}{\sigma^2 + N_m\sigma_0^2} \right) * \left(\frac{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m\sigma_0^2}\bar{x}^{(m)}}{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m\sigma_0^2}} \right) + \left(\frac{N_m\sigma_0^2}{\sigma^2 + N_m\sigma_0^2} \right) \bar{x}^{(m)}$$

$$\frac{1}{\sigma_e^2} = \frac{N_m}{\sigma^2} + \frac{1}{\sigma_0^2}$$

With MLE-II , the suitable prior μ_0 can be chosen according to the data observed, rather having a fixed prior μ_0, independent of the given data. This leads to choosing prior for μ_m which is in accordance with data which in turn helps in generalisation of our model.

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

QUESTION 5

Student Name: Abhas Kumar
Roll Number: 20111001
Date: February 26, 2021

We have linear regression model for the scores .i.e

$p(y_n^{(m)}|x_n^{(m)}, w_m) = \mathcal{N}(y_n^{(m)}|w_m^T x_n^{(m)}, \beta^{-1})$ .i.e

$p(y^{(m)}|X^{(m)}, w_m) = \mathcal{N}(y^{(m)}|X^{(m)} w_m, \beta^{-1} I_N)$, Where $y^{(m)}$ is $N_m * 1$ and $X^{(m)}$ is $N_m * D$.

Also given prior $p(w_m) = \mathcal{N}(w_m|w_0, \lambda^{-1} I_D)$, where w_0 is unknown.

Marginal distribution :

$$p(Y|X, w_0, \lambda, \beta) = \prod_{m=1}^{M} p(y^{(m)}|X^{(m)}, w_0, \lambda, \beta)$$

Using the properties of Gaussian Model we have ,

$p(y^{(m)}|X^m, w_0, \lambda, \beta) = \mathcal{N}(y^{(m)}|X^{(m)} w_0, \beta^{-1} I_{N_m} + \lambda^{-1} X^{(m)} X^{(m)^T})$

$p(Y|X, w_0, \lambda, \beta) = \prod_{m=1}^{M} \mathcal{N}(y^{(m)}|X^{(m)} w_0, \beta^{-1} I_{N_m} + \lambda^{-1} X^{(m)} X^{(m)^T})$

Log Marginal,

$$\begin{aligned}\log(p(Y|X, w_0, \lambda, \beta)) &= \log(\prod_{m=1}^{M} \mathcal{N}(y^{(m)}|X^{(m)} w_0, \beta^{-1} I_{N_m} + \lambda^{-1} X^{(m)} X^{(m)^T}) \\ &= \sum_{m=1}^{M} \log(\mathcal{N}(y^{(m)}|X^{(m)} w_0, \beta^{-1} I_{N_m} + \lambda^{-1} X^{(m)} X^{(m)^T}))\end{aligned}$$

Let $\sum_{N_m} = \beta^{-1} I_{N_m} + \lambda^{-1} X^{(m)} X^{(m)^T}$

Then ,

$$\log(p(Y|X, w_0, \lambda, \beta)) = \sum_{m=1}^{M} \big(y^{(m)} - X^{(m)} w_0\big)^T \big(\sum\nolimits_{N_m}\big)^{-1} \big(y^{(m)} - X^{(m)} w_0\big) + \text{Constant terms}$$

MLE- II for unknown w_0 can be given as :

$$\hat{w_0} = \underset{w_0}{argmin} \log(p(Y|X, w_0, \lambda, \beta))$$

$$\hat{w_0} = \underset{w_0}{argmin} \left(\sum_{m=1}^{M} \left(y^{(m)} - X^{(m)} w_0\right)^T \left(\sum\nolimits_{N_m}\right)^{-1} \left(y^{(m)} - X^{(m)} w_0\right) \right)$$

Above is the required objective function.

The value of w_0 obtained by optimising the above equation will give the better estimate of the hyper-parameter. because it is obtained by considering the data from all schools instead having it as a fixed value for w or estimating it from only the data of particular school.

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

QUESTION
6

Student Name: Abhas Kumar
Roll Number: 20111001
Date: February 26, 2021

Part 01: Required plots in figure 01

Part 02: Required plots in figure 02

Part 03: Since the maximum log marginal likelihood is for k=3, this model seems to explain the data the best.

Part 04

Based on the likelihood model 4 is the best whereas based on the log marginal likelihood model 3 is the best. But between the highest log likelihood and highest log marginal likelihood, log marginal likelihood is more reasonable to select the best model because it captures the true picture of standard deviation(variance) resulting due to uncertainty in the data , which the log likelihood method fails to capture. Log marginal likelihood also generalises the data better.

Part 05

Since the best model that is 3rd model has maximum uncertainty in the range of [-4,-3], i would want to include an additional training input in this region.

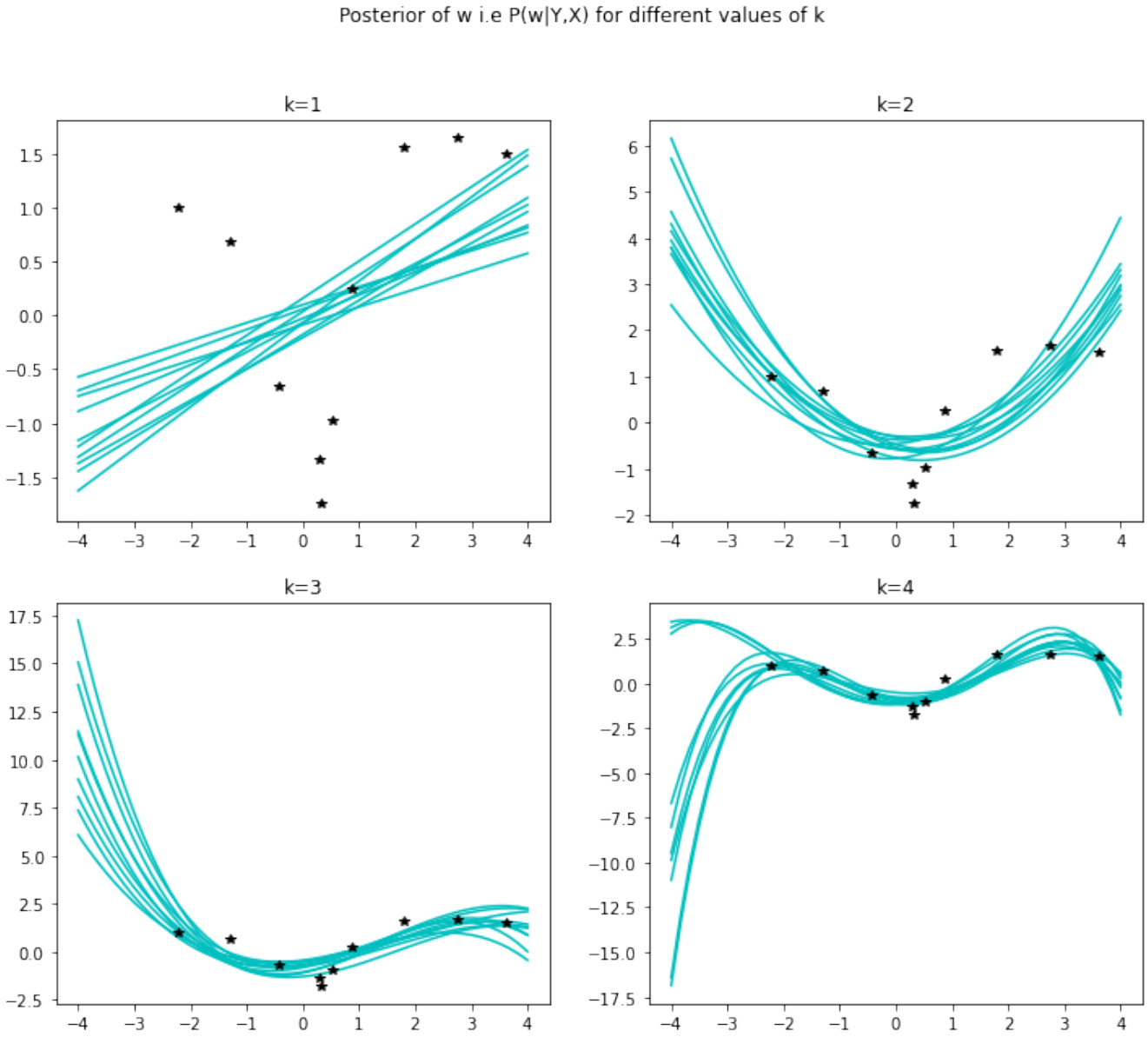

Figure 2: Plots to illustrate how well the functions fit the training data

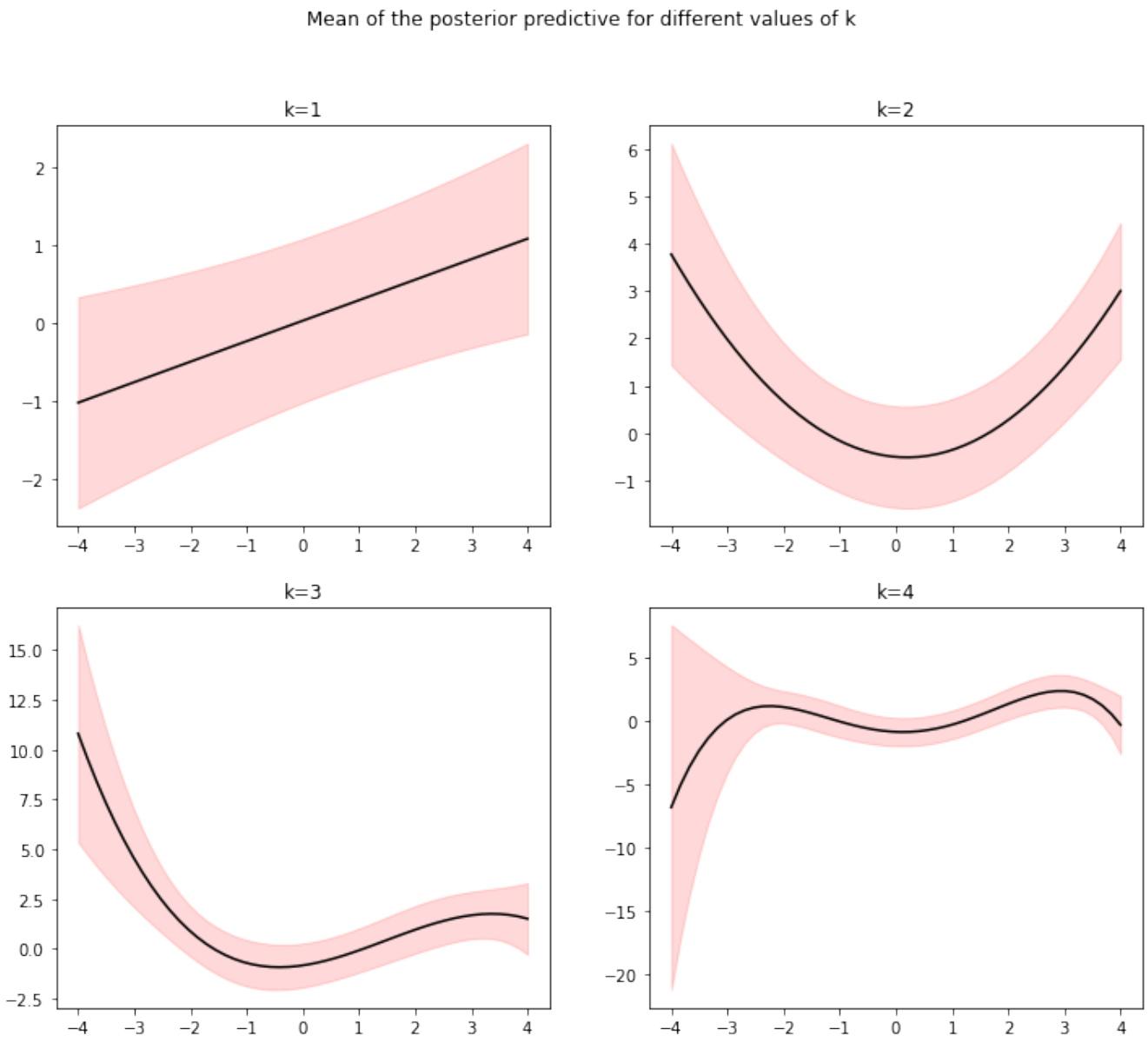

Figure 3: Plots to show mean of the posterior predictive and predictive posterior mean plus-and-minus two times the predictive posterior standard deviation