

Machine Learning for Computer Vision (EE5177)

Programming Assignment 1 : Fitting Probability Distribution to Data

Problem #1

Akshith Kumar
EE14B127

5th February 2017

Contents

1	Calculation of ML, MAP and Bayesian Estimate	1
1.1	Goal	1
1.2	Calculation of ML Estimate of mean	1
1.3	Calculation of MAP Estimate of mean	2
1.4	Calculation of Posterior Distribution using Bayesian Method	2
2	Plots of Bhattacharya Distance of ML,MAP and Bayesian	2
2.1	Goal	2
2.2	Plot of ML and MAP for Prior-1	3
2.3	Plot of MAP and Bayesian for Prior-1	3
2.4	Plot of ML and MAP for Prior-2	4
2.5	Plot of MAP and Bayesian for Prior-2	4
3	Inferences	4
3.1	Comparison of ML and MAP estimate from d vs. n plots	4
3.2	Comparison of MAP and Bayesian estimate from d vs. n plots	5

1 Calculation of ML, MAP and Bayesian Estimate

1.1 Goal

We are given a set X of D -dimensional data points generated using a Gaussian distribution of mean μ and covariance matrix Σ . We are also given the distribution of prior on the mean to be a Gaussian with $P(\mu) = N_\mu(\mu_o, \Sigma_o)$. Given the covariance, we are required to calculate the ML estimate of mean, MAP estimate of mean and posterior distribution using Bayesian Methods.

1.2 Calculation of ML Estimate of mean

The Multivariate Gaussian Distribution is defined as :

$$p(x|\mu, \Sigma) = \frac{\exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\}}{\sqrt{|2\pi \Sigma|}}$$

Given a set of i.i.d data $X = \{x_1, x_2, x_3, \dots, x_I\}$ drawn from $N(x; \mu, \Sigma)$, we can estimate (μ, Σ) by *Maximum Likelihood Estimation*. The loglikelihood function is

$$L = \ln(p(x|\mu, \Sigma)) = \frac{-I}{2} \ln|\Sigma| - \frac{1}{2} \sum_i (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + c$$

On taking derivative with respect to μ and setting to 0 gives :

$$\frac{dL}{d\mu} = \frac{-1}{2} \sum_i \frac{d(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}{d\mu} = 0$$

$$\frac{dL}{d\mu} = \frac{-1}{2} \sum_i (-2\Sigma^{-1}(x_i - \mu)) = 0$$

$$\Rightarrow \mu_{ML} = \frac{1}{I} \sum_i x_i$$

1.3 Calculation of MAP Estimate of mean

The Multivariate Gaussian Distribution is defined as :

$$p(x|\mu, \Sigma) = \frac{\exp\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\}}{\sqrt{|2\pi \Sigma|}}$$

We define the μ_{MAP} as follows

$$\mu_{MAP} = \underset{\mu}{\operatorname{argmax}} [\prod_{i=1}^I p(x_i|\mu, \Sigma)p(\mu)]$$

Similar to the *Maximum Likelihood Estimate*, we can maximise the logarithm of the function, which implies

$$\mu_{MAP} = \underset{\mu}{\operatorname{argmax}} [\sum_{i=1}^I \log(p(x_i|\mu, \Sigma) + \log(p(\mu))]$$

Let the cost function be $L = \sum_{i=1}^I \log(p(x_i|\mu, \Sigma) + \log(p(\mu))$ It can be simplified as

$$L = \sum_{i=1}^I (x_i - \mu)^T \Sigma^{-1}(x_i - \mu) + (\mu - \mu_o)^T \Sigma_o^{-1}(\mu - \mu_o)$$

On differentiating L with respect to μ and setting to 0, we get :

$$\frac{dL}{d\mu} = \sum_{i=1}^I \Sigma^{-1}(x_i - \mu) + \Sigma_o^{-1}(\mu - \mu_o) = 0$$

On re-arranging the terms, we get:

$$\begin{aligned} \Rightarrow (I\Sigma^{-1} + \Sigma_o^{-1})\mu_{MAP} &= \sum_{i=1}^I \Sigma^{-1}x_i + \Sigma_o^{-1}\mu_o \\ \Rightarrow \mu_{MAP} &= (I\Sigma^{-1} + \Sigma_o^{-1})^{-1}(\sum_{i=1}^I \Sigma^{-1}x_i + \Sigma_o^{-1}\mu_o) \end{aligned}$$

1.4 Calculation of Posterior Distribution using Bayesian Method

As given in the question, we assume that $p(x|\mu) = N(\mu, \Sigma)$ and $p(\mu) = N(\mu_o, \Sigma_o)$ where Σ , Σ_o and μ_o are known. After observing a set D of n independent samples x_1, x_2, \dots, x_n , we use the Bayes' Theorem to obtain

$$\begin{aligned} p(\mu|D) &= \alpha \prod_{i=1}^n p(x_i|\mu)p(\mu) \\ p(\mu|D) &= \beta \exp\left[-\frac{1}{2}(\mu^T(n\Sigma^{-1} + \Sigma_o^{-1})\mu - 2\mu^T(\Sigma^{-1}\sum_{i=1}^n x_i + \Sigma_o^{-1}\mu_o))\right] \end{aligned}$$

,which has the form

$$p(\mu|D) = \gamma \exp\left[-\frac{1}{2}(\mu - \mu_n)^T \Sigma_n^{-1}(\mu - \mu_n)\right]$$

Thus we can say that $p(\mu|D) = N(\mu_n, \Sigma_n)$ where we have

$$\begin{aligned} \mu_n &= \Sigma_o(\Sigma_o + \frac{1}{n}\Sigma)^{-1}\mu_n^* + \frac{1}{n}\Sigma(\Sigma_o + \frac{1}{n}\Sigma)^{-1}\mu_o \\ \Sigma_n &= \Sigma_o(\Sigma_o + \frac{1}{n}\Sigma)^{-1}\frac{1}{n}\Sigma \end{aligned}$$

Observe that x can be viewed as the sum of two mutually independent random variables, a random vector μ with $p(\mu|D) = N(\mu_n, \Sigma_n)$ and an independent random vector y with $p(y) = N(0, \Sigma)$. Since the sum of two independent, normally distributed random vectors is again a normally distributed vector whose mean is the sum of the means and whose covariance matrix is the sum of covariance matrices. we have

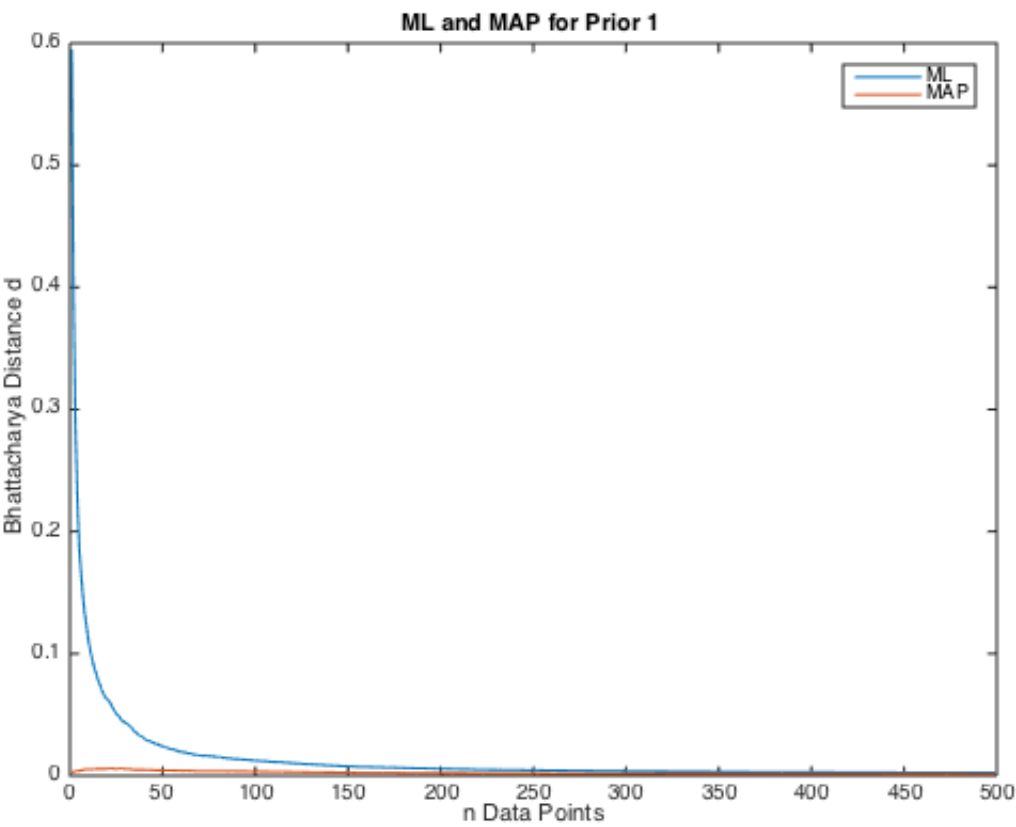
$$p(x|D) = N(\mu_n, \Sigma + \Sigma_n)$$

2 Plots of Bhattacharya Distance of ML,MAP and Bayesian

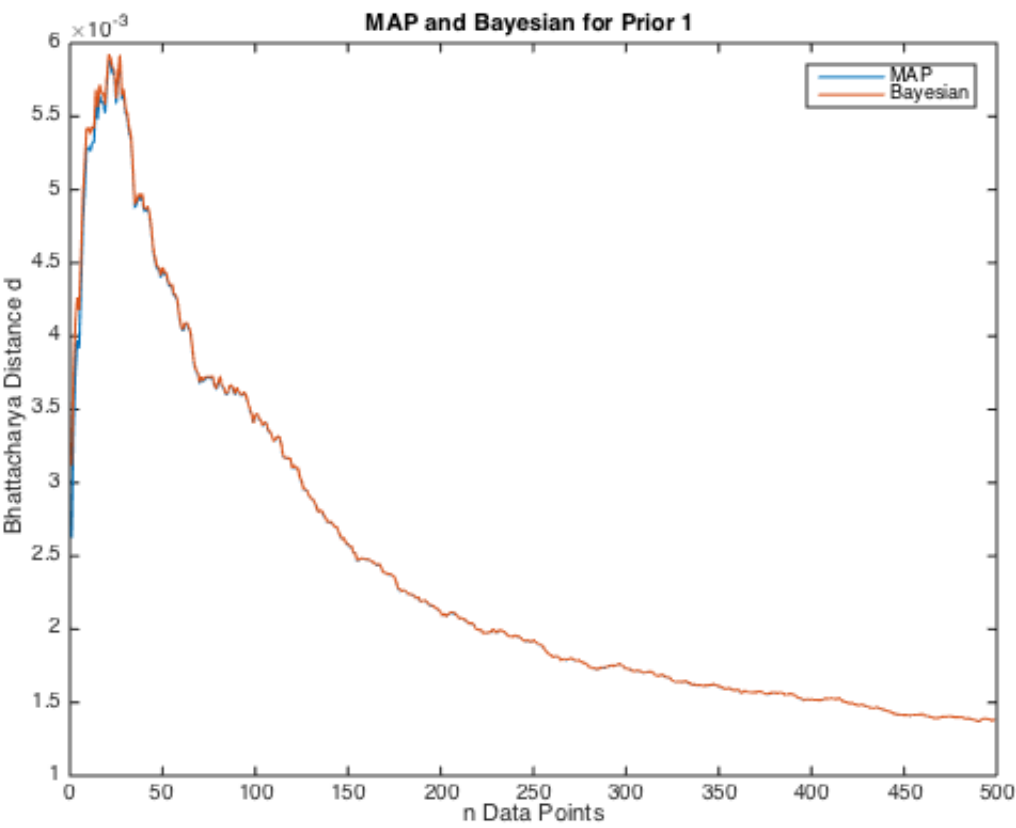
2.1 Goal

The estimates derived using Maximum Likelihood, Maximum A Posteriori and Bayesian Methods can be analysed quantitatively using the Bhattacharya distance as a distance metric between the ground truth values and the estimated values.

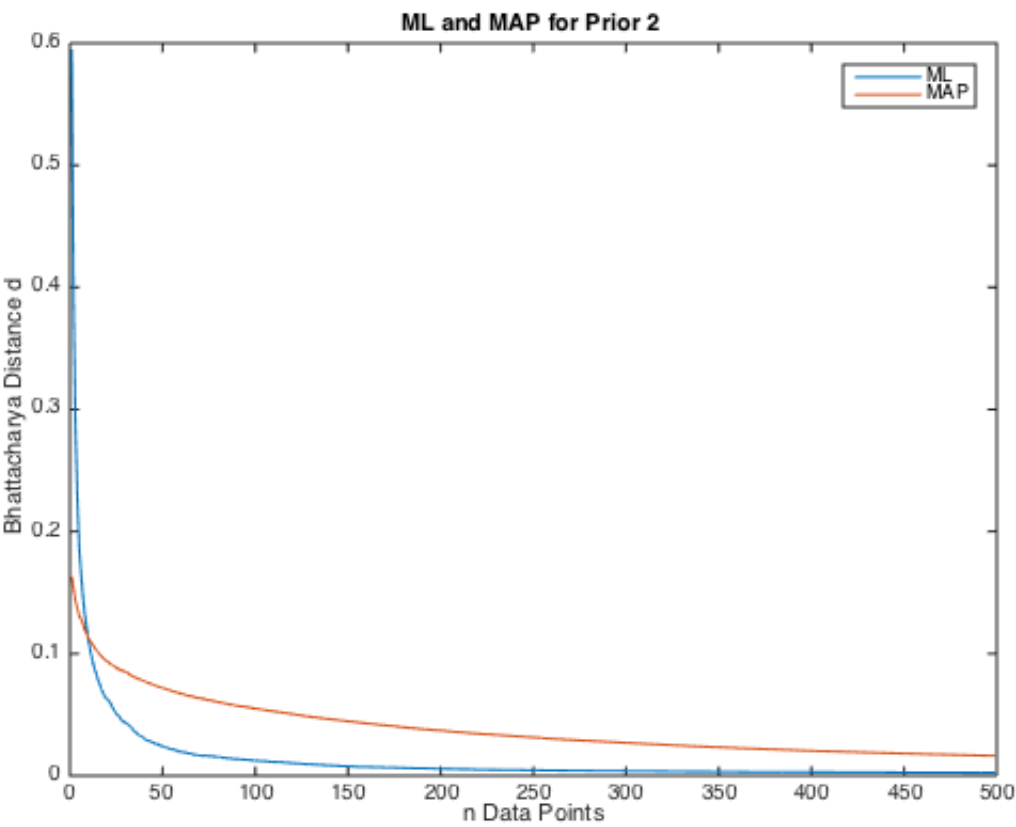
2.2 Plot of ML and MAP for Prior-1



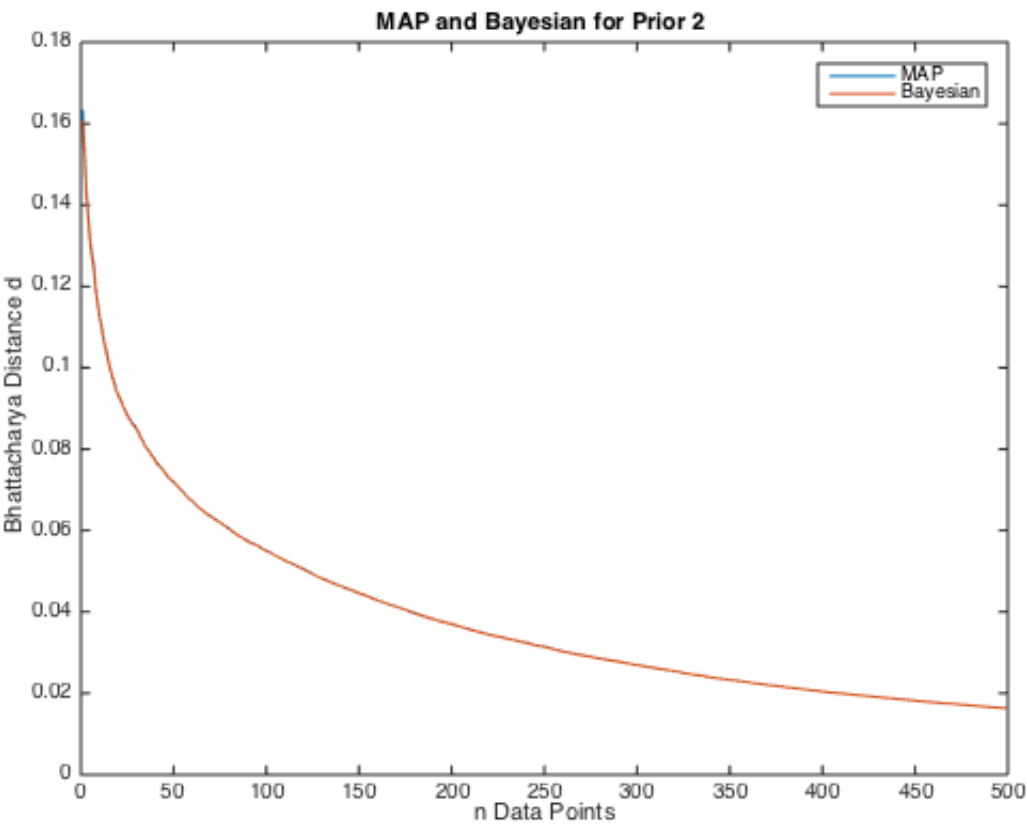
2.3 Plot of MAP and Bayesian for Prior-1



2.4 Plot of ML and MAP for Prior-2



2.5 Plot of MAP and Bayesian for Prior-2



3 Inferences

3.1 Comparison of ML and MAP estimate from d vs. n plots

From the d vs. n plot for Prior-1 shows that initially ML estimate is worse compared to MAP estimate but as n increases, the ML estimate converges to the MAP estimate, performing as well as MAP estimate as n increases (goes to 500). Since the MAP estimate

is so low to start with, it implies that Prior-1 is close to the true prior.

From the d vs. n plot for Prior-2 shows that initially ML estimate is worse compared to MAP estimate, for some 20 odd iterations and then after that quickly converges to the true value, performing better than the MAP estimate. The plot shows that MAP estimate will require a lot data to converge to the true values of the mean and covariance.

3.2 Comparison of MAP and Bayesian estimate from d vs. n plots

From both the plots, the MAP and the Bayesian estimates seem to be performing the same as the plot for MAP and Bayesian overlap with each other.

For prior-1, MAP and Bayesian estimates both increase initially and start decreasing after some 50 odd points.

For prior-2, MAP and Bayesian estimates continuously decrease as n increases but lot of data points to converge to the true values.