Homework 3 Solutions

1 Linear Regression

If we define $\mathbf{R} = diag\,(r_1, \ldots, r_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_D\,(\boldsymbol{w}) = \frac{1}{2}\,(\boldsymbol{t} - \boldsymbol{\Phi w})^T\,\mathbf{R}\,(\boldsymbol{t} - \boldsymbol{\Phi w})\,. \tag{1}$$

Setting the derivative with respect to $\boldsymbol{w}$ to zero and rearranging, we have

$$\boldsymbol{w}^* = \left(\boldsymbol{\Phi}^T\mathbf{R}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}\mathbf{R}\boldsymbol{t}. \tag{2}$$

Note that the above solution reduces to the standard solution if $\mathbf{R} = \mathbf{I}$.

Note that there are two interpretations of the above solution. First, from lecture notes, we know

$$\beta E_D\,(\boldsymbol{w}) = \beta\frac{1}{2}\sum_{n=1}^{N}\left\{y_n - \boldsymbol{w}^T\phi\,(\mathbf{x}_n)\right\}^2. \tag{3}$$

Comparing Eq. (1) in the problem statement with Eq. 3, we can conclude that r_n can be regarded as a precision (inverse variance) parameter, particular to the data point that either replaces or scales β.

Alternatively, r_n can be regarded as an effective number of replicated observations of data points $(\boldsymbol{x}_n, t_n)$. This is particularly true if r_n is a positive integer; although it is true for any $r_n > 0$.

2 Empirical Bayes

We know that the posterior updates for θ_i are

$$\theta_i \sim Beta(\alpha + y_i, \beta + N_i - y_i)$$

and our problem is mainly to estimate the hyperparameters α, β from fixed-point iteration. This can be obtained as

$$\boldsymbol{\alpha}_k^{\text{new}} = \boldsymbol{\alpha}_k\frac{\sum_i \Psi\,(n_{ik} + \boldsymbol{\alpha}_k) - \Psi\,(\boldsymbol{\alpha}_k)}{\sum_i \Psi\,(n_i + \sum_k \boldsymbol{\alpha}_k) - \Psi\,(\sum_k \boldsymbol{\alpha}_k)}, \tag{4}$$

where Ψ is the digamma function. Further details on this can be found in (**?**). In this problem we are provided with the MATLAB code to solve the above optimization problem for computing a and b.

Running the appropriate MATLAB code gives

$$\alpha \approx 0.83 \quad \beta \approx 611$$

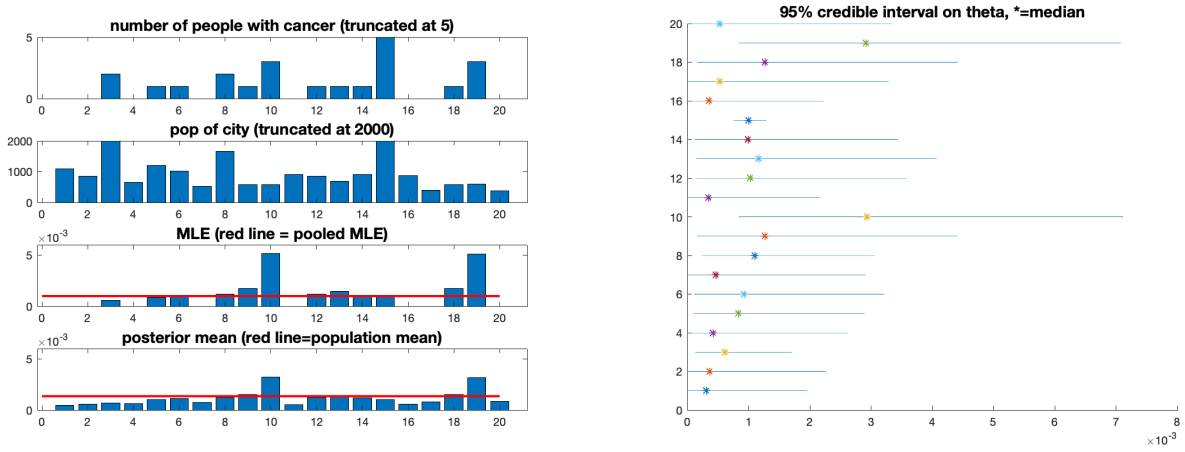
for a final estimate of the posterior distribution

$$\theta_i \mid \mathcal{D}, \alpha, \beta \sim Beta(0.83 + y_i, 611 + N_i + y_i)$$

for which the numerical posterior summary statistics of the first five cities are provided below; the rest are shown in Fig. 1.

Country List			
City Index	Posterior α	Posterior $\beta(\times 10^3)$	Posterior Mean $(\times 10^{-3})$
1	0.83	0.1694	0.4897
2	0.83	0.1466	0.5658
3	2.83	0.4070	0.6948
4	0.83	0.1268	0.6540
5	1.83	0.1818	1.0054

Note that the posterior variances that are several orders of magnitudes smaller and are thus left out of this analysis. A nice comparison of the MLE and Bayesian approaches to this problem are shown in Fig. 1, where we observe the effect the prior distribution has on reducing the regularizing the zero-inflation of the MLE estimate.



(a) (b)

Figure 1: From top to bottom: plot of the population sizes over the cities (N_i). Number of deaths from cancer (y_i). The MLE solution $\hat{\theta}_{\text{MLE}} = y_i/N_i$ (The red line is the pooled MLE, $\hat{\theta} = \frac{\sum_i x_i}{\sum_i N_i}$). The posterior mean $\mathbb{E}[\theta_i|\mathcal{D}]$ using Empirical Bayes on the hyperparameters. The red line is the population level mean computed as the prior mean of θ with the parameters a and b evaluated by maximizing the evidence $p(\mathcal{D})$, i.e. $a/(a+b)\,|\mathcal{D}$. (b) Posterior 95% credible intervals on the cancer rate.

3 Robust linear regression

A. For the first part of this problem we use the MLE estimate with a Gaussian likelihood for this linear regression problem.

Using the code developed in Homework 2, we can see that the MLE estimate is clearly susceptible to influence from out-liers.

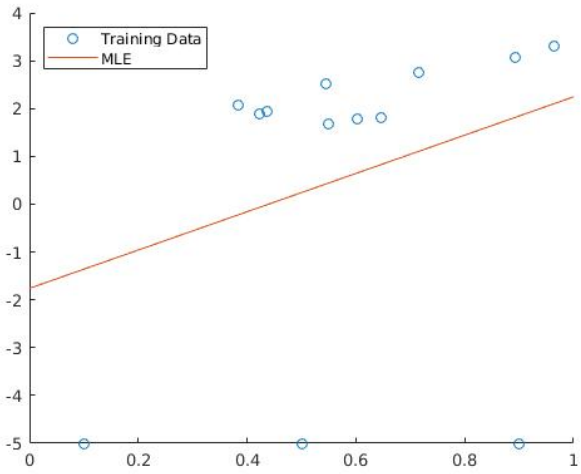


Figure 2: Linear regression using Maximum-Likelihood/Least squares loss function.

B. We now choose to use a Laplace distribution (equivalent to L1 norm loss) and Student-T distribution as potential likelihood models. Both have heavier tails than the Gaussian, thus can perform better with outlier data. Solving the Laplace likelihood is straight forward, however an iterative solver was used for the Student T with provided degree of freedom and variance.

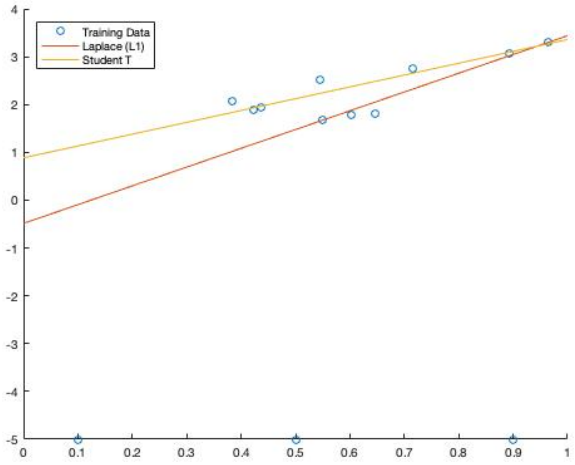


Figure 3: Linear regression using Laplace and Student T likelihood models.

C. Finally, we look at the Huber loss model for two different δ's. By the definition of the loss, as δ is decreased the Huber loss becomes more like the Laplace (L1) loss. Thus we can see that for $\delta = 1$, the prediction is closer to the Laplace prediction from the previous section.

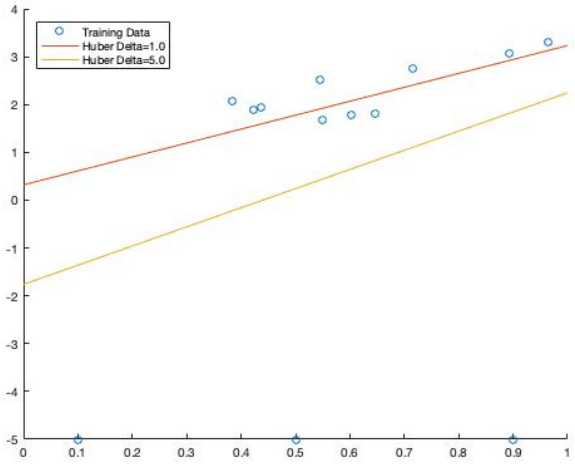


Figure 4: Linear regression using the Huber likelihood model.

Discussion: Over all we can see that just default least squares does a poor job when outliers are present. This can easily be resolved but changing the likelihood model. As we have seen, distributions with heavier tails are more robust in this aspect (Student T and Laplace). Additionally more complex loss functions such as the Huber loss can be used to manually throttle the influence of outliers. The codes for generating the results shown above are provided at this link.

4 Bayesian linear regression

A. Start with the likelihood:

$$p(y|\Phi, w, \sigma^2) = \mathcal{N}(y|\Phi w, \sigma^2 I_n) \tag{5}$$

and the prior as:

$$p(w|\sigma^2, \Phi) = \mathcal{N}(0, \gamma\sigma^2 I) \tag{6}$$

$$p(\sigma^2) = InvGamma(a, b) \tag{7}$$

We start with finding the joint prior:

$$p(w, \sigma^2) = p(w|\sigma^2, \Phi)p(\sigma^2) \tag{8}$$

$$= \frac{1}{(2\pi\gamma\sigma^2)^{\frac{D}{2}}|I|^{\frac{1}{2}}} exp\{\frac{-w^T I^{-1}(w)}{2\gamma\sigma^2}\}\frac{b^a}{\Gamma(a)}(\frac{1}{\sigma^2})^{a+1}exp(-\frac{b}{\sigma^2}) \tag{9}$$

$$= \frac{b^a}{(2\pi\gamma)^{\frac{D}{2}}\Gamma(a)}(\sigma^2)^{-(a+D/2+1)}exp(\frac{-w^T w + 2b\gamma}{2\gamma\sigma^2}) \tag{10}$$

Now we can find the posterior which is a Normal-Inverse Gamma distribution:

$$p(w, \sigma^2|y) = p(y|\Phi w, \sigma^2 I_n)p(w, \sigma^2), \tag{11}$$

$$= \frac{1}{2\pi}(\sigma^2)^{-\frac{N}{2}}exp(-\frac{(y-\Phi w)^T(y-\Phi w)}{2\sigma^2})\frac{b^a}{(2\pi\gamma)^{\frac{D}{2}}\Gamma(a)}(\sigma^2)^{-a+\frac{b}{2}+1}exp(\frac{-w^T w + 2b\gamma}{2\gamma\sigma^2}), \tag{12}$$

$$\boxed{p(w,\sigma^2|y) = \frac{b^a}{(2\pi)^{\frac{N+D}{2}}\gamma^{\frac{D}{2}}\Gamma(a)}(\sigma^2)^{-(a+\frac{N+D}{2}+1)}exp(-\frac{\gamma(y-\Phi w)^T(y-\Phi w)-w^Tw+2b\gamma}{2\gamma\sigma^2}).} \tag{13}$$

Now we drive the expression for the marginal posterior as:

$$p(w|D) = \int p(w,\sigma^2|y)d(\frac{1}{\sigma^2}) \tag{14}$$

$$= \int \frac{1}{(2\pi)^{\frac{N+D}{2}}\gamma^{\frac{D}{2}}\Gamma(a)}(\sigma^2)^{-(a+\frac{N+D}{2}+1)}exp\left\{-\frac{\gamma(y-\Phi w)^T(y-\Phi w)-w^Tw+2b\gamma}{2\gamma\sigma^2}\right\}d\left(\frac{1}{\sigma^2}\right) \tag{15}$$

$$= \frac{b^a}{\Gamma(a)}(\frac{1}{2\pi})^{\frac{(N+D)}{2}}(\frac{1}{\gamma})^{(\frac{D}{2})}\int(\frac{1}{\sigma^2})^{(a+(\frac{N+D}{2})+1)}exp\left\{-\frac{1}{\sigma^2}-\frac{\gamma(y-\Phi w)^T(y-\Phi w)-w^Tw+2b\gamma}{2\gamma}\right\}d\left(\frac{1}{\sigma^2}\right) \tag{16}$$

Recall the definition of the gamma function:

$$\gamma(a) = \int x^{a-1}exp(-x)dx. \tag{17}$$

Using this we can arrive at the form of the marginal posterior which is in the form of a Student T:

$$p(w|D) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{N+D}{2}}\left(\frac{1}{\gamma}\right)^{\frac{D}{2}}\frac{\Gamma(a+\frac{N+D}{2}+2)}{\left(\frac{1}{2}(y-\Phi w)^T(y-\Phi w)-\frac{1}{2\gamma}w^Tw+b\right)^{a+\frac{N+D}{2}+2}} \tag{18}$$

$$= \frac{\Gamma(a+\frac{N+D+1}{2})}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{N+D}{2}}\left(\frac{1}{\gamma}\right)^{\frac{D}{2}}\left(\frac{1}{b}\right)^{\frac{N+D+1}{2}}\left(\frac{1}{1+\frac{1}{2b}(w-w_n)^TV_n^{-1}(w-w_n)}\right)^{-(a+\frac{N+D+1}{2})} \tag{19}$$

$$\boxed{= \mathcal{T}(2a, w_n, \frac{b}{a}V_n)} \tag{20}$$

where we have defined (following Dr. Zabaras' slides):

$$V_n = (\gamma I + \Phi^T\Phi)^{-1}, \quad w_n = V_n(\Phi^T y). \tag{21}$$

To find the predictive distribution we need to marginalize out the weights and noise:

$$p(\tilde{y}|x,D) = \int\int p(\bar{y}|\Phi w,\sigma^2 I)p(w,\sigma^2|y)dwd\sigma^2, \tag{22}$$

$$= \int\int \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{1}{(\sigma^2)^{\frac{M}{2}}}exp(-\frac{(y-\Phi w)^T(y-\Phi w)}{2\sigma^2})\frac{b^a}{(2\pi)^{\frac{N+D}{2}}\gamma^{\frac{D}{2}}\Gamma(a)(\sigma^2)^{-(a+\frac{N+D}{2}+1)}}... \tag{23}$$

$$exp\left(-\frac{\gamma(y-\Phi w)^T(y-\Phi w)-w^Tw+2b\gamma}{2\gamma\sigma^2}\right)dwd(\sigma^2), \tag{24}$$

$$= \int\int \frac{1}{(2\pi)^{N+\frac{D}{2}}}\frac{1}{(\sigma^2)^{-(a+\frac{N+D+M}{2}+1)}}\frac{b^a}{\gamma^{\frac{D}{2}}\Gamma(a)}... \tag{25}$$

$$exp\left(-\frac{\gamma(y-\Phi w)^T(y-\Phi w)+(w-w_n)^TV_n^{-1}(w-w_n)+2b_n}{2\gamma\sigma^2}\right)dwd(\sigma^2). \tag{26}$$

where we have defined:

$$b_n = \gamma b + \frac{1}{2}(y^T y - {w_n}^T V_n^{-1} w_n). \tag{27}$$

Now working a little bit with the stuff in the exponential:

$$\begin{aligned}
&\gamma(y-\Phi w)^T(y-\Phi w) + (w-w_n)^T V_n^{-1}(w-w_n) + 2b_n = \\
&(w-(\Phi^T\Phi + \frac{1}{\gamma}{V_n}^{-1})^{-1}(\Phi^T y + \frac{1}{\gamma}V_n^{-1}w_n))^T(\Phi^T\Phi + \frac{1}{\gamma}V_n^{-1})(w-(\Phi^T\Phi + \frac{1}{\gamma}V_n^{-1})^{-1}(\Phi^T y + \frac{1}{\gamma}V_n^{-1}w_n))) \\
&\qquad - (\Phi^T y + \frac{1}{\gamma}V_n^{-1}w_n)^T(\Phi^T\Phi + \frac{1}{\gamma}{V_n}^{-1})^{-1}(\Phi^T y + V_n^{-1}w_n) + {w_n}^T V_n^{-1} w_n + y^T y + 2b_n
\end{aligned} \tag{28}$$

We define the latter components that are constant in the integral as a single variable:

$$2\beta \;=\; -(\Phi^T y + \frac{1}{\gamma}V_n^{-1}w_n)^T(\Phi^T\Phi + \frac{1}{\gamma}V_n^{-1})^{-1}(\Phi^T y + V_n^{-1}w_n) + {w_n}^T V_n^{-1} w_n + y^T y + 2b_n. \tag{29}$$

Now we can integrate out w since it is in the form of a Gaussian resulting in a constant, thus we need only to be concerned about the 2β term.

$$p(\tilde{y}|x, D) \propto \int (\frac{1}{\sigma^2})^{-(a+\frac{N}{2}+\frac{M}{2})} exp\left(\frac{-2\beta}{2\sigma^2}\right) d(\sigma^2) \tag{30}$$

Using the Gamma normalization constant and completing the square, we have:

$$p(\tilde{y}|x, D) \propto (-(\Phi^T y + \frac{1}{\gamma}V_n^{-1}w_n)^T(\Phi^T\Phi + \frac{1}{\gamma})^{-1}(\Phi^T y + V_n^{-1}w_n) + {w_n}^T V_n^{-1} w_n + y^T y + 2b_n)^{-(a+\frac{N}{2}+\frac{M}{2})} \tag{31}$$

$$\propto (1 + \frac{(y-\Phi w_n)^T(\frac{b_n}{a+\frac{N}{2}})(I_m + \Phi V_n \Phi^T)^{-1}(y-\Phi w_n)}{2(a+\frac{N}{2})})^{-(a+\frac{N}{2}+\frac{M}{2})} \tag{32}$$

We can see that $p(\tilde{y}|\tilde{x}, D)$ can be expressed as a Student-T distribution:

$$\boxed{p(\tilde{y}|\tilde{x}, D) = \mathcal{T}(y|\Phi w_n, \frac{b_n}{a+\frac{N}{2}}(I_m + \Phi V_n \Phi^T), 2a+N)} \tag{33}$$

$$\mu = \Phi w_n, \quad \Sigma = \frac{b_n}{a+\frac{N}{2}}(I_m + \Phi V_n \Phi^T), \quad \nu = 2a + N \tag{34}$$

Finally we calculate the model evidence by marginalizing the posterior:

$$p(D) = \int\int p(y|\Phi w, \sigma^2 I_n) p(w|0, \gamma\sigma^2 I) dw p(\sigma^2) d(\sigma^2) \tag{35}$$

$$= \int\int \frac{1}{(2\pi)^{\frac{N}{2}}(\sigma^2)^{\frac{N}{2}}} exp\left(-\frac{(y-\Phi w)^T(y-\Phi w)}{2\sigma^2}\right) \frac{1}{(2\pi\gamma\sigma^2)^{\frac{D}{2}}} ... \tag{36}$$

$$exp\left(-\frac{w^T w}{2\gamma\sigma^2}\right) dw \frac{b^a}{\Gamma(a)} (\frac{1}{\sigma^2})^{a+1} exp(-\frac{b}{\sigma^2}) d\sigma^2 \tag{37}$$

$$= \frac{b^a}{(2\pi)^{\frac{N+D}{2}}\Gamma(a)\gamma^{\frac{D}{2}}} \int\int exp\left(-\frac{(y-\Phi w)^T(y-\Phi w)}{2\sigma^2}\right) ... \tag{38}$$

$$exp\left(-\frac{w^T w}{2\gamma\sigma^2}\right) dw \left(\frac{1}{\sigma^2}\right)^{a+\frac{N+D}{2}+1} exp\left(-\frac{b}{\sigma^2}\right) d\sigma^2 \tag{39}$$

$$= \frac{b^a}{(2\pi)^{(\frac{N+D}{2})}\Gamma(a)\gamma^{\frac{D}{2}}} \int\int exp\left(-\frac{(w-w_n)^T V_n^- 1 (w-w_n)}{2\sigma^2}\right) dw \left(\frac{1}{\sigma^2}\right)^{a+\frac{N+D}{2}+1} exp\left(-\frac{b}{\sigma^2}\right) d\sigma^2 \tag{40}$$

We can integrate out the first exponential term since it takes the form of a Gaussian, it is simply the normalization constant of that Gaussian.

$$p(D) = \frac{b^a}{(2\pi)^{(\frac{N+D}{2})}\Gamma(a)\gamma^{\frac{D}{2}}} \int (2\pi)^{(\frac{D}{2})} V_n^{\frac{1}{2}} (\sigma^2)^{(\frac{D}{2})} (\frac{1}{\sigma^2})^{a+\frac{N+D}{2}+1} exp\left(-\frac{b}{\sigma^2}\right) d\sigma^2, \tag{41}$$

$$= \frac{b^a (2\pi)^{(\frac{D}{2})} |V_n|^{\frac{1}{2}}}{(2\pi)^{\frac{N+D}{2}}\Gamma(a)\gamma^{\frac{D}{2}}} \int \left(\frac{1}{\sigma^2}\right)^{a+\frac{N}{2}+1} exp\left(-\frac{b}{\sigma^2}\right) d\sigma^2. \tag{42}$$

We can repeat the same trick by for a Gamma distribution, $\mathcal{G}(\alpha = a + \frac{N}{2}, \beta = b)$.

$$\boxed{p(D) = \frac{b^a (2\pi)^{\frac{D}{2}} |V_n|^{\frac{1}{2}} \Gamma(a + \frac{N}{2})}{b^{a+\frac{N}{2}} (2\pi)^{\frac{N+D}{2}} \gamma^{\frac{D}{2}} \Gamma(a)}} \tag{43}$$

B. Using the derivation carried out in part A (Eqs. 69 and 70), a computer code is written. The result obtained using this code are provided in Fig. 5. The behavior is as expected with prediction error more outside the range of the data. Note that the error bar shown is $\pm\sigma$.

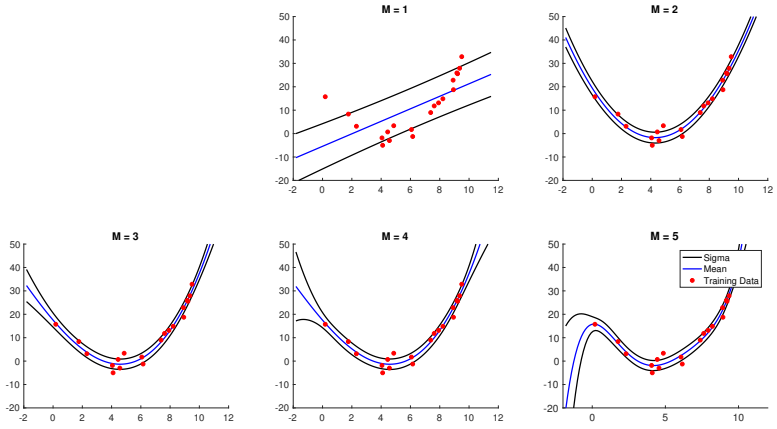


Figure 5: Bayesian linear regression mean and variance.

C. In this problem, we are supposed to draw samples for the weights w_n. This is achieved by drawing a sample from the inversegamma distribution (with updated parameters) and then, using the same to draw samples for w_n. In essence, we are drawing samples from the NIG (normal inverse gamma) distribution. The results obtained are shown in Fig. 6.

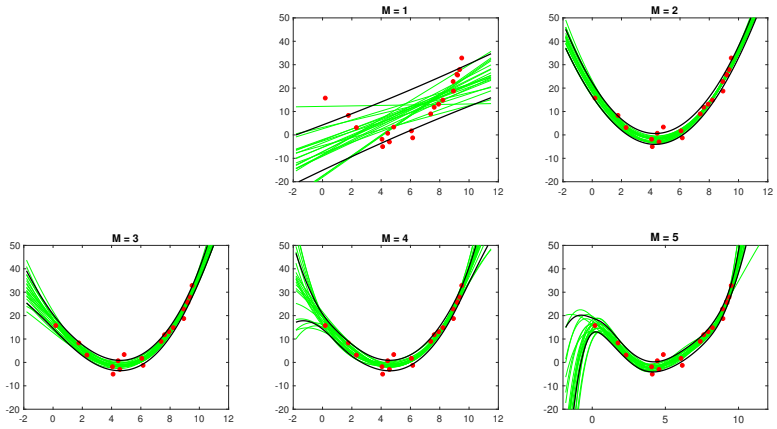


Figure 6: Bayesian linear regression samples.

D. In this question, we are asked to compute the model evidence and hence, select the most appropriate model. The formula for model evidence is shown in Eq. 79. A MATLAB code for implementing this formula has been written. The results are shown in Fig. 7. It is observed that model 2 is the most suitable one.

From the model evidence it seem model with order 2 was the best for this set of data. I would agree from the sample plots above.

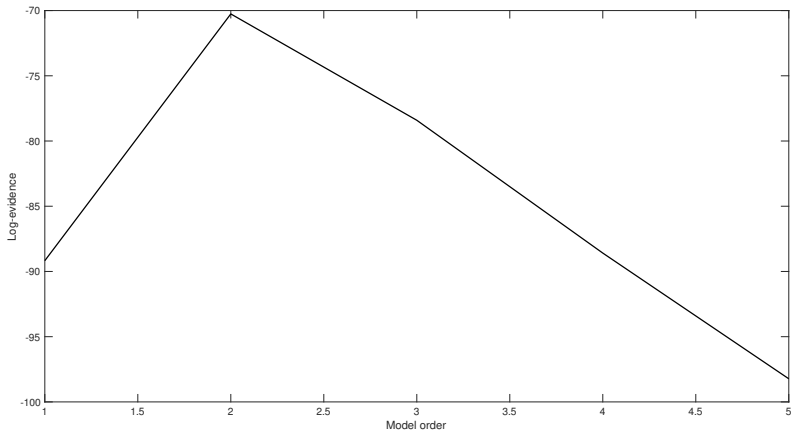


Figure 7: Model evidence of the Bayesian linear regression for various polynomial basis functions.

E. Lastly, we repeated part B by centering the data. The results are shown in Fig 8. The difference with part b is minimal. In fact, it seems, results obtained in Part B is somewhat better for $M = 1$.

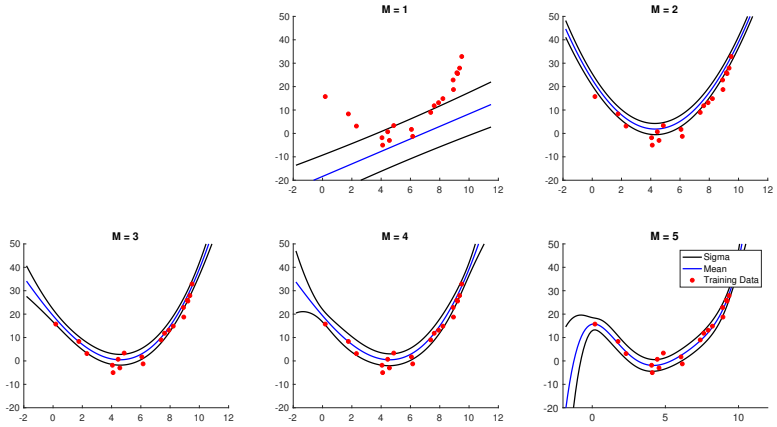


Figure 8: Linear regression with centering the data and computing the bias term in post-processing.

5 Behavior of training set error with increasing sample size, Multi-output regression and Ridge regression

A. Complex models are prone to overfitting, which means when they are paired with little data, they will "memorize" the training data set and give low errors on the training data set. Nevertheless, since the model has a lot of parameters and not enough training data to learn them well, the model don't generalize beyond the training set and give high errors on the test data set and on real data a well. When we increase the data set size, the model starts learning instead of memorizing the data set. The final result is that both the training and test error start to converge to the same answer. In another words, the training/test errors increases/decreases to the best generalization error of the model.

B. Consider a 2 dimensional response vector $\boldsymbol{y}_i \in \mathbb{R}^2$ and binary input data $x_i \in 0, 1$. Also, consider the following basis function.

$$\phi(0) = [1, 0]^T \tag{44}$$
$$\phi(1) = [0, 1]^T \tag{45}$$

Define the inputs/observations as the following:

$$x = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad y = \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix} \tag{46}$$

The MLE of this follow the solution of the MSE linear regression $\hat{\boldsymbol{E}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_{:j}$. Thus everything falls down into linear algebra:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \qquad (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix}, \qquad \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} -4 & -4 \\ 4 & 4 \end{bmatrix} \tag{47}$$

Resulting in the weights:

$$\boldsymbol{W} = \begin{bmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \\ \frac{4}{3} & \frac{4}{3} \end{bmatrix}$$