ANLY-601 Homework 2

Guanzhi Wang, collaborate with Shaoyu Feng

Question 1

1.

Yes, it's convex.

First, by Triangle Inequality of norm, we have $||v + w|| \leq ||v|| + ||w||$ for any v, w. So for $f(x)$, we have $||\lambda x_1 + (1-\lambda)x_2|| \leq ||\lambda x_1|| + ||(1-\lambda)x_2|| = \lambda||x_1|| + (1-\lambda)||x_2||$. So each norm is convex. Thus, the sum of norm is convex.

2.

It's not convex. We can find some examples to prove that. Assume $k(x, x') = k(d) = d \; and \; k'(x, x') = k(d) = d^2$, then $f(x) = d - d^2$. This is obvious not convex. for example, let $x_1 = 0, x_2 = 1, t = (1-t) = 0.5, f(tx_1 + (1-t)x_2) > tf(x_1) + (1-t)f(x_2)$

3.

It's not convex. Assume $k(d) = k(x, x') = d = -k'(x, x'), f(x) = -d^2 - b$. Thus, easy to find when $t = 0.5, x_1 = 10, x_2 = 0, b = 100, f(tx_1 + (1-t)x_2) > tf(x_1) + (1-5)f(x_2)$.

4.

It's convex.

$$f(tx_1 + (1-t)x_2) = ||tx_1 + (1-t)x_2||_p + min(0, -tx_1 - (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) = t(||x_1$$

5.

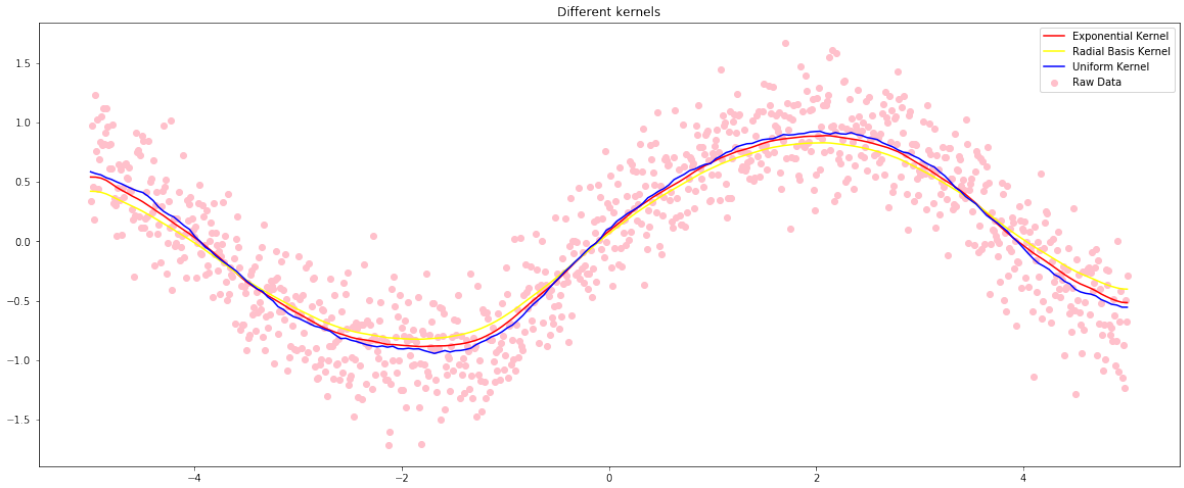
It's convex. From (1), we have f(x)=||x|| is convex. Easy to know f(y)=max(0,y) is also convex. so f(h)=f(x)+f(y) is also convex since a positive combination of convex functions is convex.

Question 2

(1)

In [87]:

Out[87]: Text(0.5,1,'Different kernels')



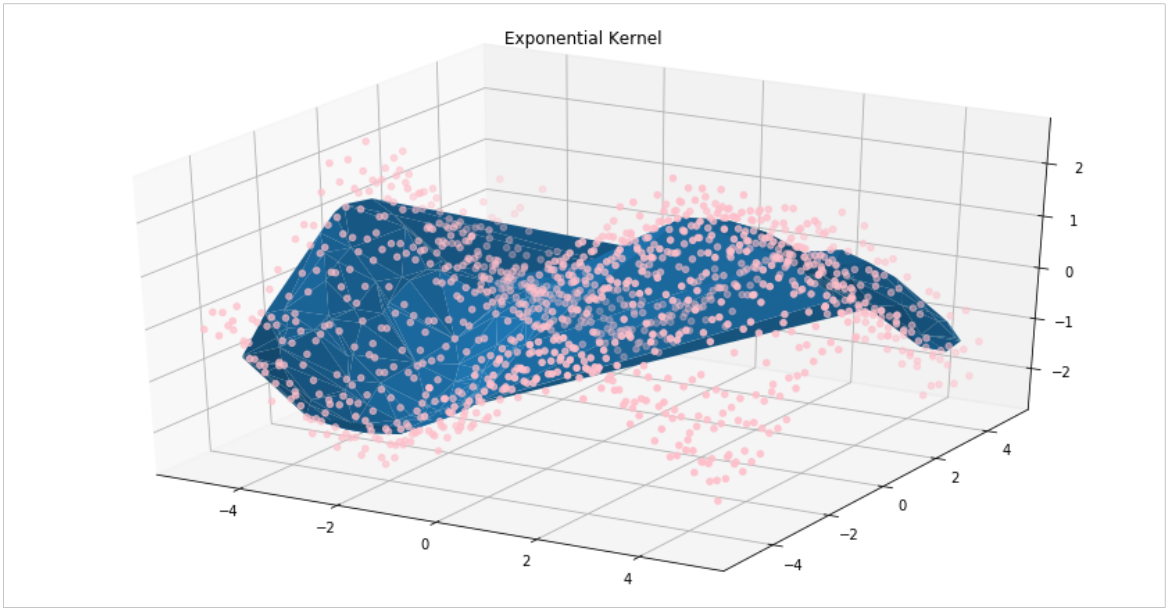
(2)

From the output, we can see basically all three kernels fit the data well. Furthermore, Radial and Exponential kernel performance better than Uniform kernel.

(3)

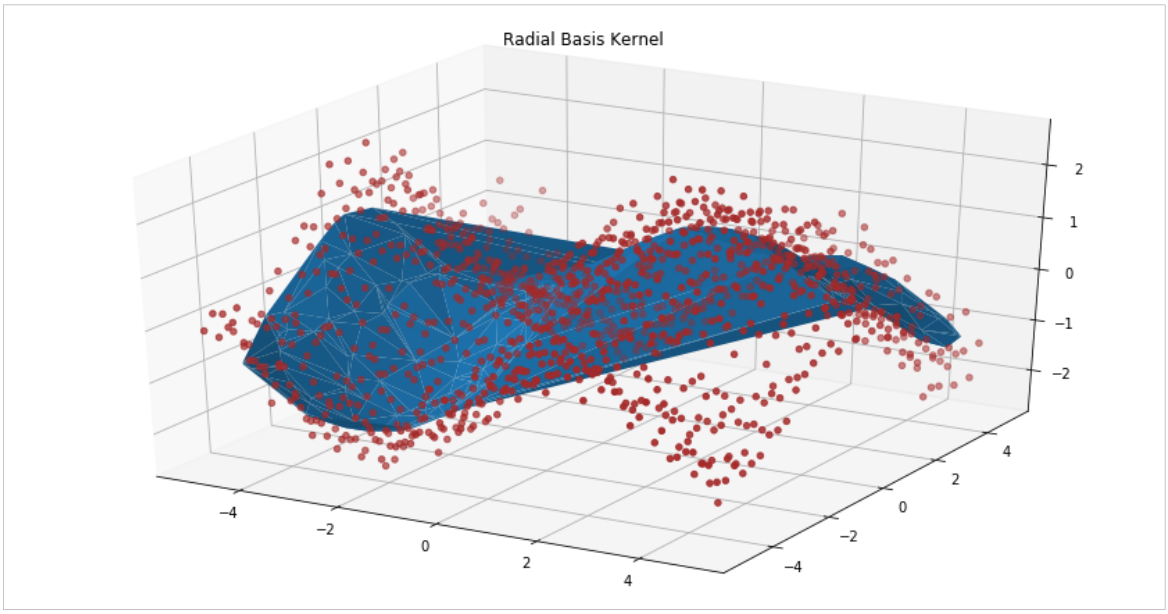
In [36]:

Out[36]: Text(0.5,0.92,'Exponential Kernel')



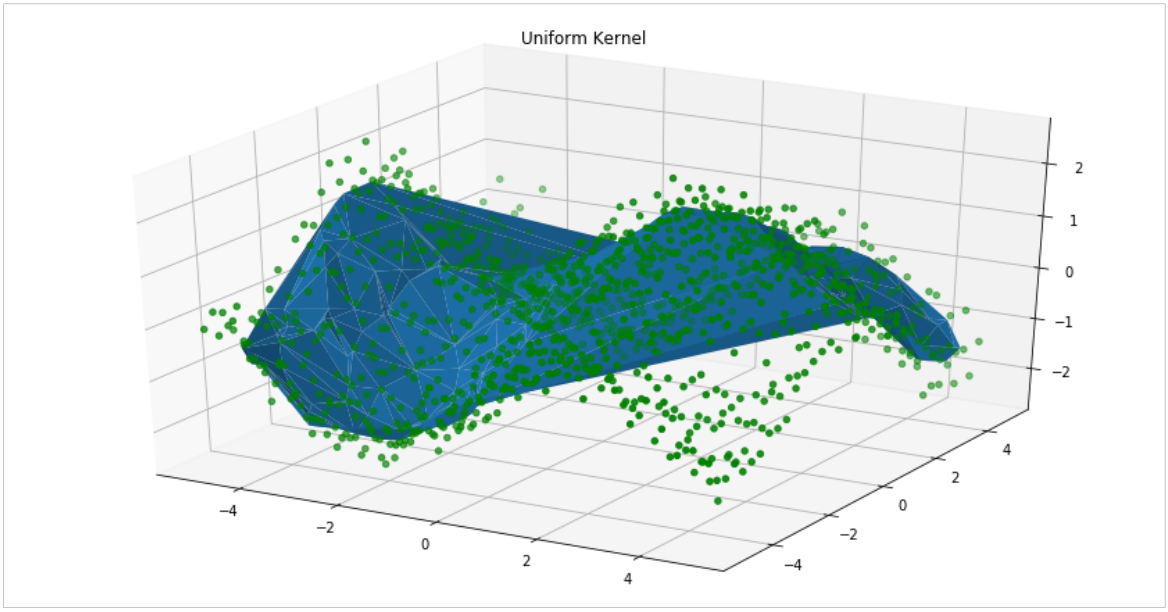
In [37]:

Out[37]: Text(0.5,0.92,'Radial Basis Kernel')

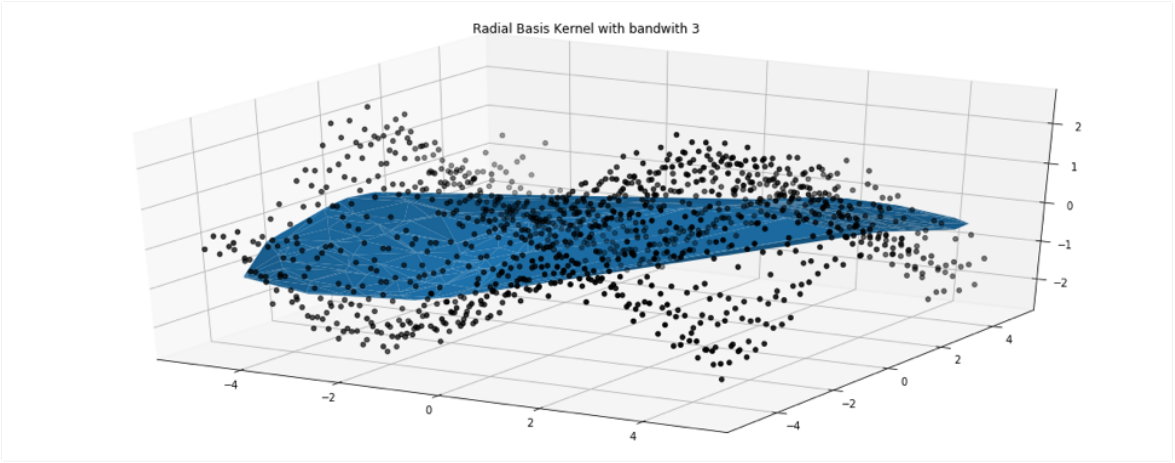


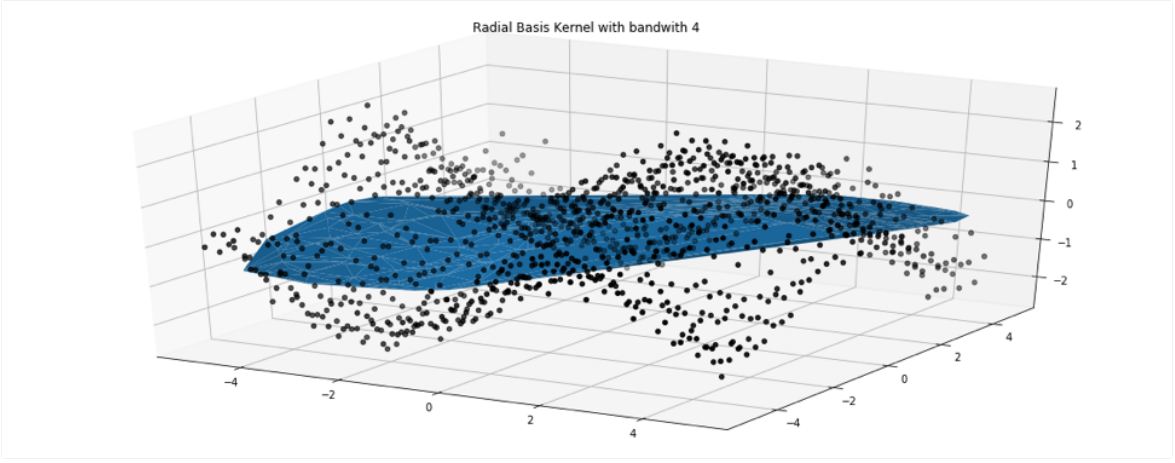
In [38]:

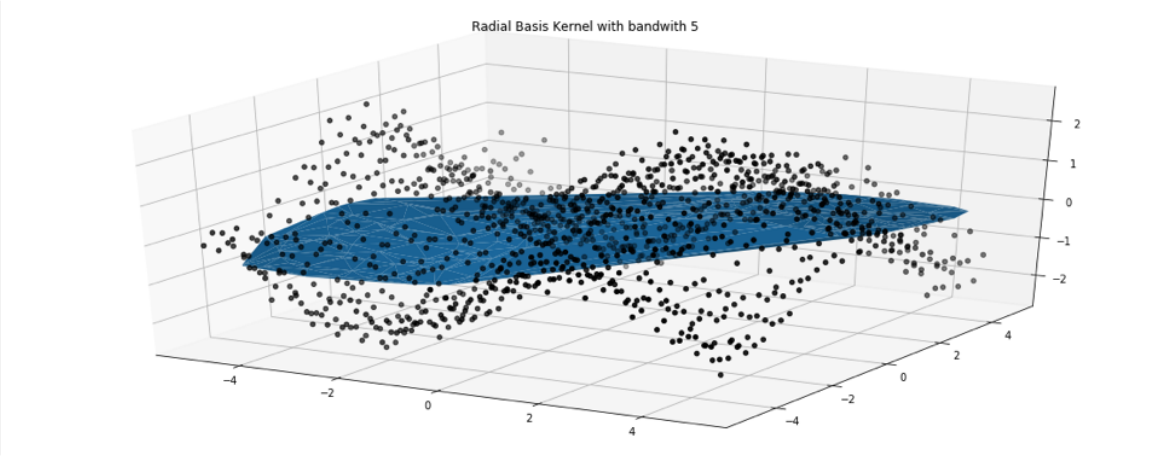
Out[38]: Text(0.5,0.92,'Uniform Kernel')

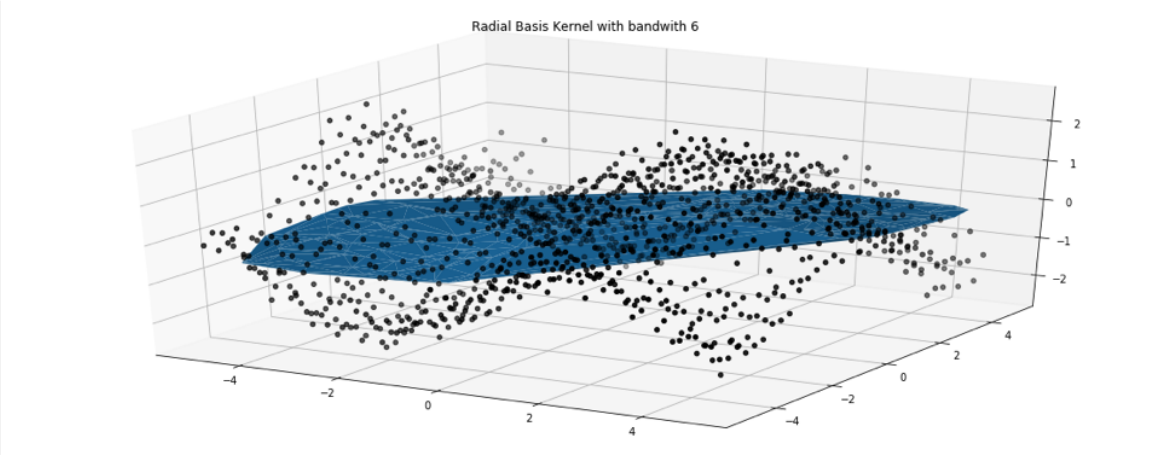


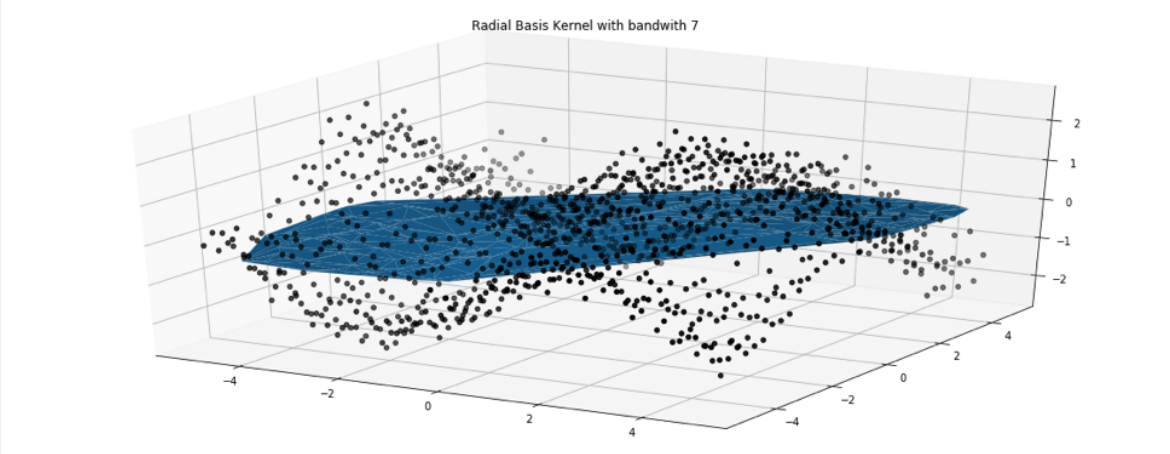
In [39]:

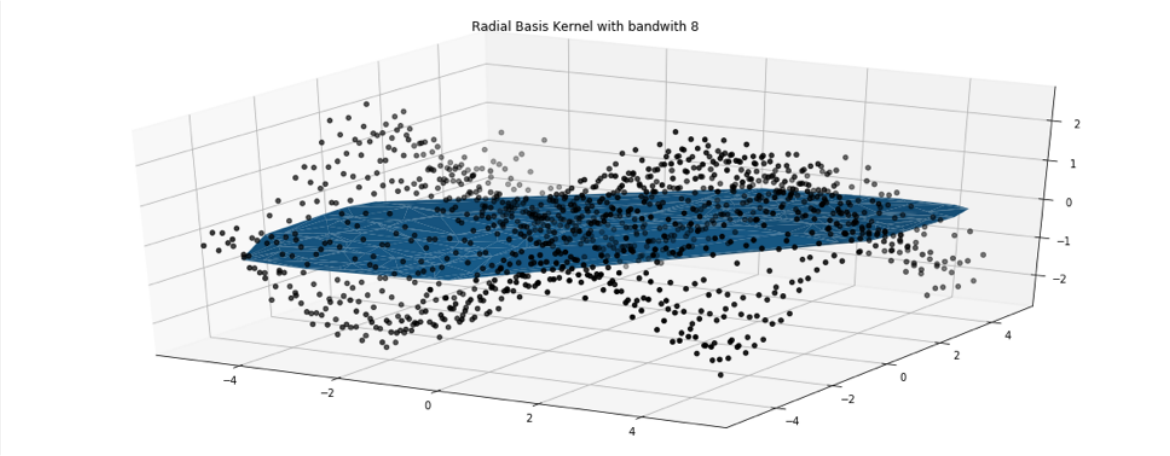





Radial Basis Kernel with bandwith 5


Radial Basis Kernel with bandwith 6


Radial Basis Kernel with bandwith 7

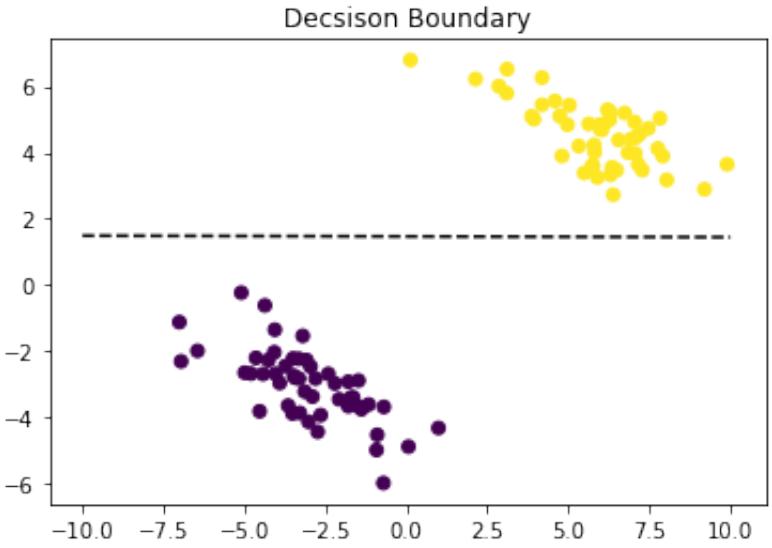


From bandwith with 3-8, the kernel still fit the data well. However, the planes are flatter with the increase of bandwith.

Question 3

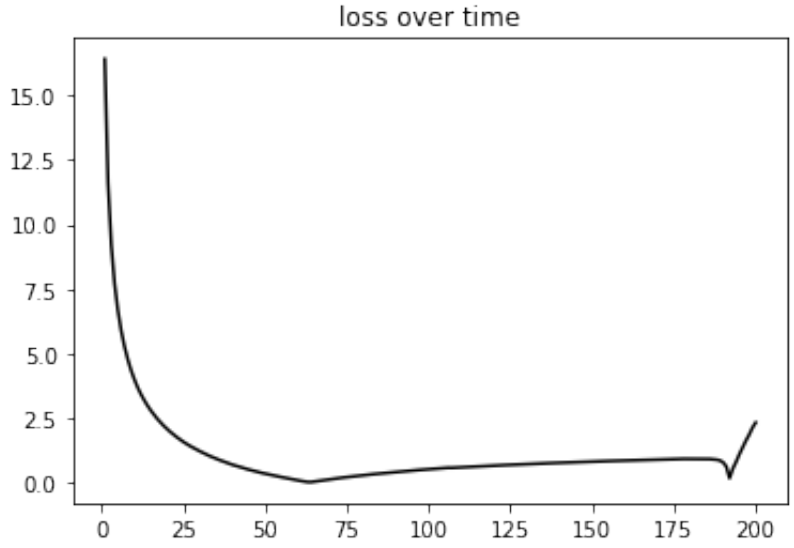
To make the model runs faster, we can use the subset of data instead of full dataset for each epoch. The following graphs show that the boundery divided the data well, the loss is getting smaller and smaller, and the run time is become longer but still in a linear form.

In [83]:



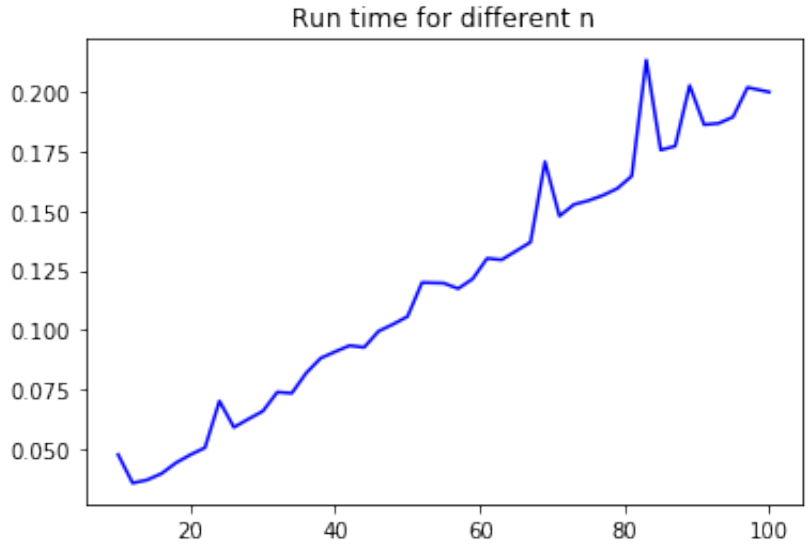
In [84]:

Out[84]: Text(0.5,1,'loss over time')



In [85]:

Out[85]: Text(0.5,1,'Run time for different n')



Question 4

(a)

First, find the conjugate posteriors for μ:

$$P(\mu|\tau,\upsilon,\sigma,X) \propto \frac{1}{\sqrt{2\pi}\upsilon}e^{-\frac{(\mu-\tau)^2}{2\upsilon}} * \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma}}$$

Thus,

$$P(\mu|\tau,\upsilon,\sigma,X) \propto C * e^{\frac{(\mu-(\tau\sigma+n\bar{X}\upsilon)/(\sigma+n\upsilon))^2}{2\upsilon\sigma/(\sigma+n\upsilon)}}$$

So

$$\mu|\tau,\upsilon,\sigma,X \sim N(\frac{\tau\sigma+n\bar{X}\upsilon}{\sigma+n\upsilon},\frac{\upsilon\sigma}{\sigma+n\upsilon}).$$

Then, find the conjugate posteriors for σ^2:

$$P(\sigma|\tau,\upsilon,\sigma,X) \propto (\frac{1}{\sigma})^{\alpha+1+n/2}e^{(2\beta+\sum_1^n(x_i-\mu)^2)/2\sigma^2}$$

Thus,

$$\sigma|\tau,\upsilon,\sigma,X \sim InverseGamma(\alpha+n/2,\sum_1^n(x_i-\mu)^2/2)$$

(b)

$$P(p_1,p_2,\dots p_j|x_{ij}) \propto \prod_{i=1}^{n}p_i^{\alpha_i-1}\frac{n!}{x_{i1}!\dots x_{ik}!}p_1^{x_{i1}}\dots p_k^{x_{ik}}$$

$$\propto \prod_{i=1}^{n}p_i^{\alpha_i+\sum_{j=1}^{n}x_{ji}-1}$$

Thus,

$$p_1,p_2,\dots p_j|x_{ij} \sim Dirichlet(\alpha_1+\sum_{j=1}^{n}x_{j1}-1,\dots,\alpha_k+\sum_{j=1}^{n}x_{jk}-1,\dots)$$

(3)

$$P(\lambda|x_1,x_2,\dots x_n) \propto \lambda^{\alpha-1}e^{-\lambda/\beta}\prod_{i}^{n}=1\lambda^{x_i}e^{-\lambda}$$

$$\propto \lambda^{n\bar{x}+\alpha-1}e^{\lambda(-1/\beta-1)}$$

Thus,

$$\lambda|x_1,x_2,\dots x_n \sim Gamma(n\bar{x}+\alpha,\frac{\beta}{n\beta+1})$$

Question 5

(1)

First regularise parameter β with $N(\beta|0,\lambda^{-1})$. The Gaussian likelihood is

$$\prod_{n=1}^{N}\mathcal{N}(y_n|\beta x_n,\sigma^2).$$

So, the combination of the likelihood and the prior is

$$\prod_{n=1}^{N}\mathcal{N}(y_n|\beta x_n,\sigma^2)\mathcal{N}(\beta|0,\lambda^{-1}).$$

Take the log, we have

$$\sum_{n=1}^{N}-\frac{1}{\sigma^2}(y_n-\beta x_n)^2-\lambda\beta^2+\text{const.}$$

From the above, we can find that the L2 penalty (ridge) is equivalent to a Normal prior.

(2)

$$argmax_\beta[log\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(y_i-(\beta_0+\beta_1x_{i,1}+\ldots\beta_jx_{i,j}))^2}{2\sigma^2}}+log\prod_{i=1}^{n}\frac{1}{2b}e^{-\frac{\beta_j}{2b}}]$$

$$=argmax_\beta\frac{1}{2\sigma^2}[\sum_{i-1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+\ldots\beta_jx_{i,j}))^2+\frac{\sigma^2}{b}\sum_{j=0}^{p}|\beta_j|]$$

$$=argmax_\beta[\sum_{i-1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+\ldots\beta_jx_{i,j}))^2+\lambda\sum_{j=0}^{p}|\beta_j|].$$

Thus, the L1 penalty is equivalent to a Laplace prior since the target function of maximum posterior estimation is equivalent to LASSO regression.

Question 6

(1)

Firstly, The posterior distribution refers to the distribution of the parameter, while the predictive posterior distribution refers to the distribution of observations of data. The posterior distribution is the distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained while the posterior predictive distribution is basically used to predict new data values.

(2)

Posterior predictive distribution. Since it is the distribution based on the data you have already have. The posterior predictive distribution is basically used to predict new data values.

(3)

First we have,

$$f(x_1, x_2, \dots x_n|\mu, \sigma) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}.$$

Thus,

$$log(f(x_1, x_2, \dots x_n|\mu, \sigma)) = -\frac{n}{2}log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2.$$

Take derivative of μ:

$$-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 = 0$$

Take derivative of σ:

$$-\frac{n}{\sigma} + \sum_{i=1}^n (x_i-\mu)^2\sigma^{-3} = 0.$$

Thus,

$$\hat{\mu_{MLE}} = \bar{x}, \hat{\sigma^2_{MLE}} = \sigma_x^2.$$

From Question 4, we have

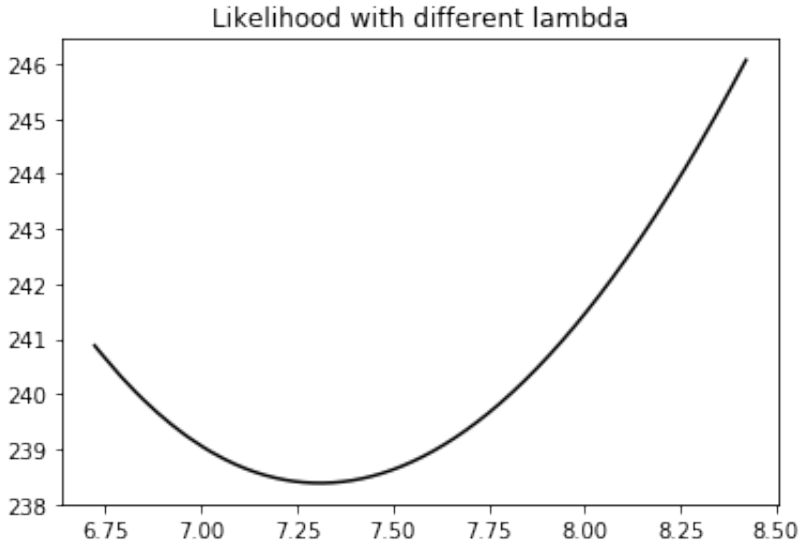
$$\hat{\mu_{MAP}} = \frac{\alpha\sigma + n\bar{x}\beta}{\sigma + n\beta}, \hat{\sigma^2_{MAP}} = \frac{\upsilon + \sum_{i=1}^n (x_i-\mu)^2/2}{\tau + n/2 - 1}.$$

So when n becomes larger and larger, we can see MAP for μ, σ^2 are more and more close to the MLE for μ, σ^2.

Question 7

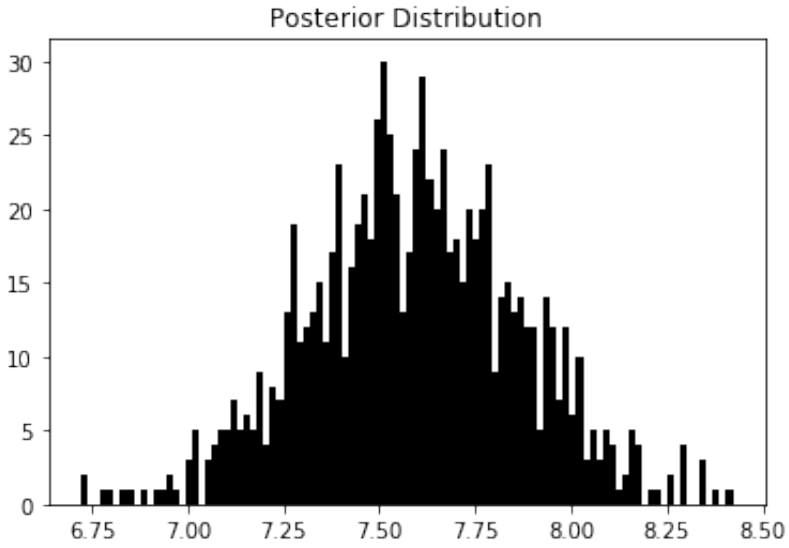
In [90]:

Out[90]: Text(0.5,1,'Likelihood with different lambda')



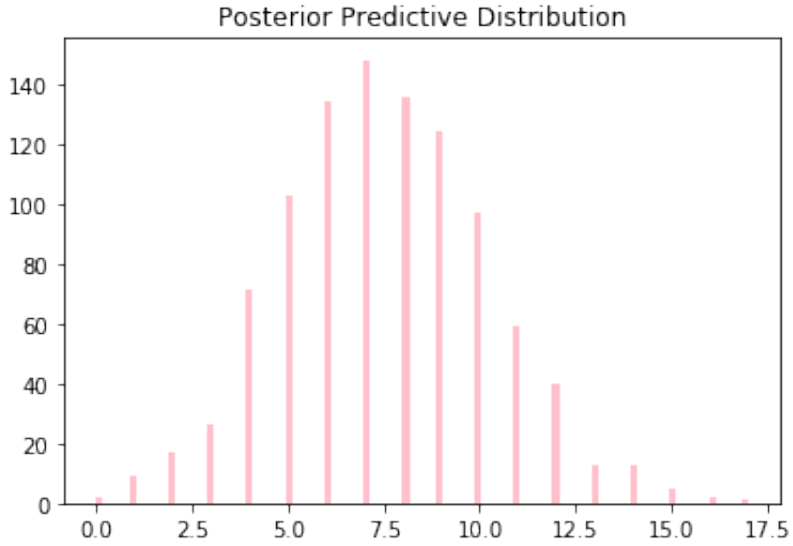
In [91]:

Out[91]: Text(0.5,1,'Posterior Distribution')



In [92]:

Out[92]: Text(0.5,1,'Posterior Predictive Distribution')

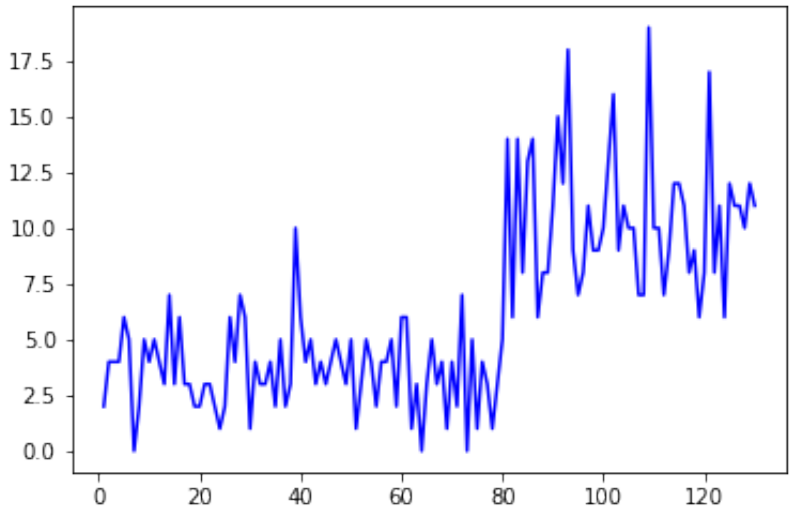


From the graph, we can see how likelihood functions' value change with lambda and how the Posterior Distribution and Posterior Predictive Distribution look like. Also, there is no convergence problem since there is no dependency.

Question 8

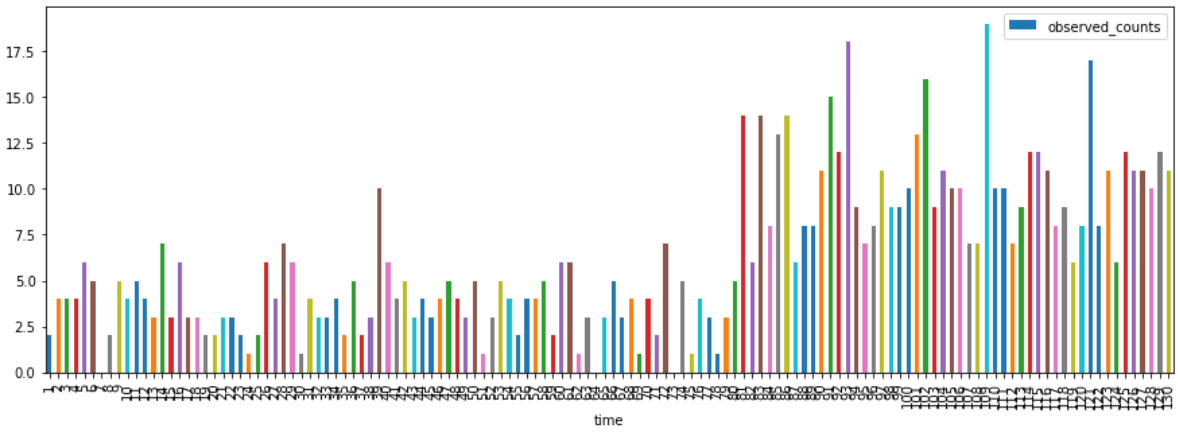
First we plot the dataset out. And Visualize the data.

In [136]:

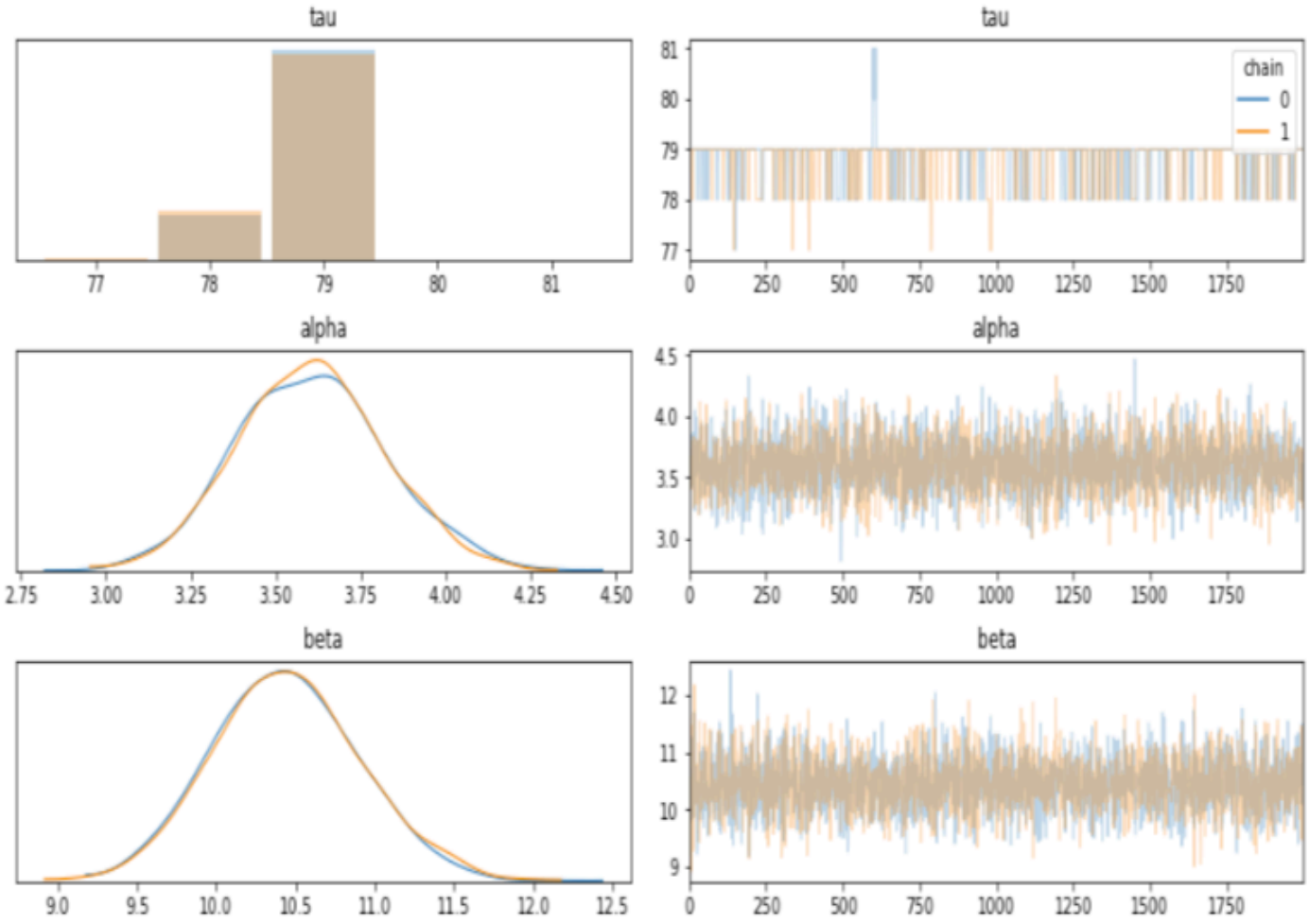


In [101]:

Out[101]: <matplotlib.axes._subplots.AxesSubplot at 0x1c28117a20>

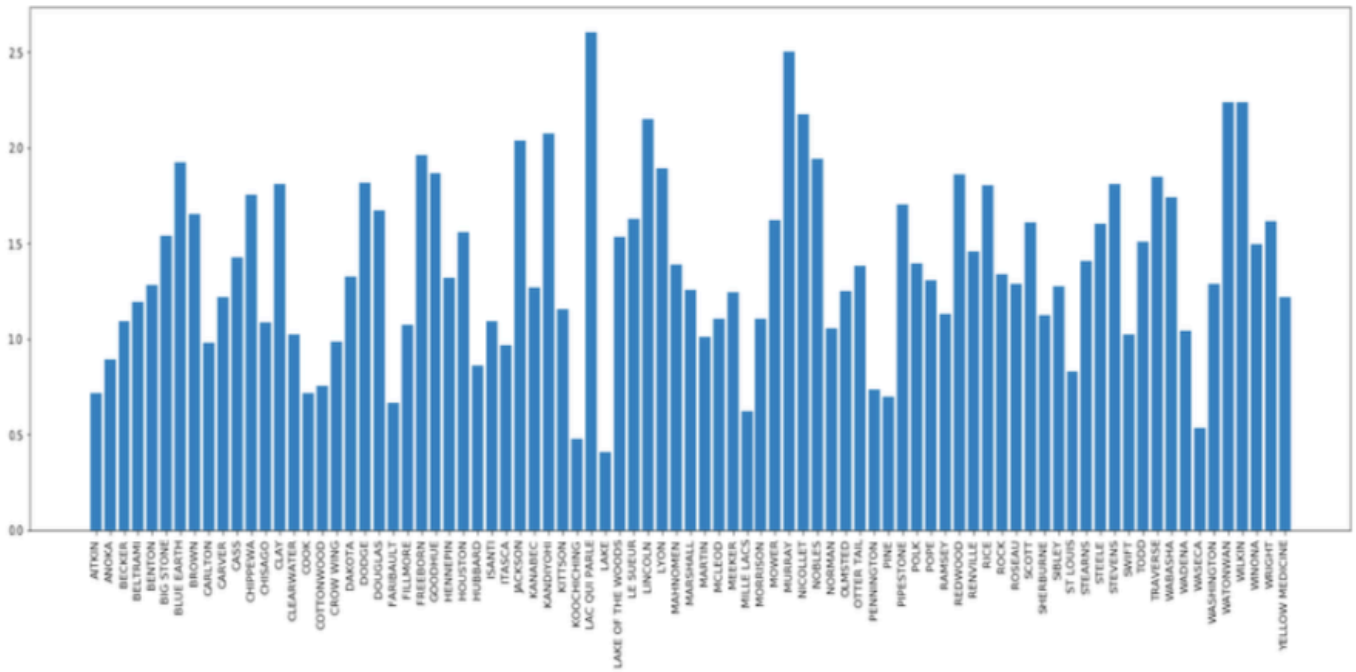


By using the Bayesian approach, we can estimate all patameters through MCMC. The following is the results.

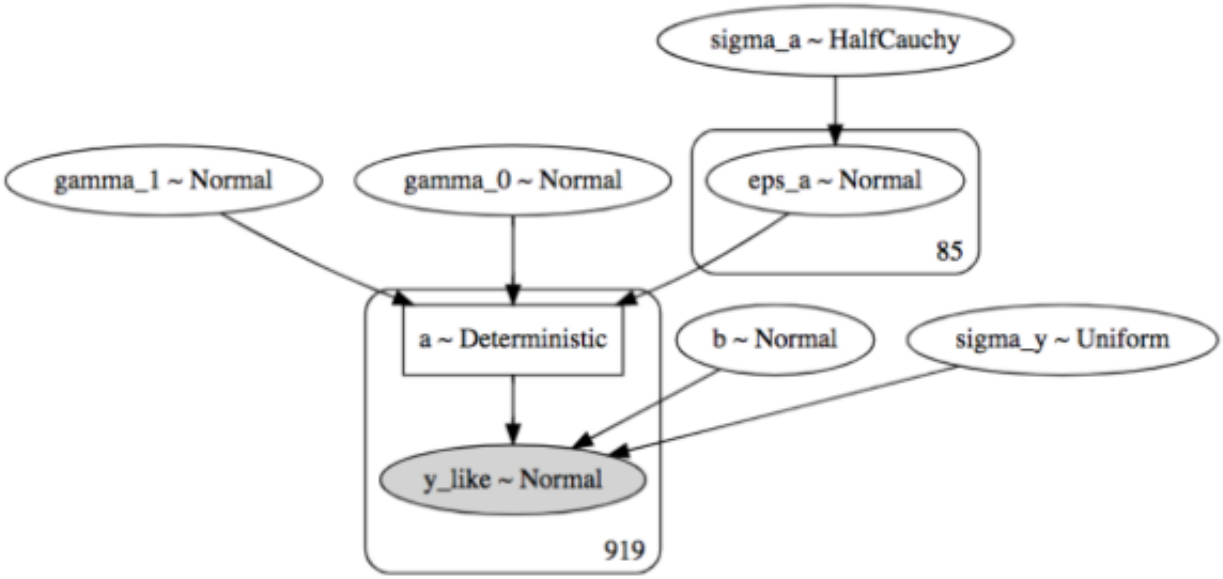


From above, we observe that the τ was mostly identified as 79 through the histogram. and multiple-chains results showed similar results. Thus, the sampler is converged.

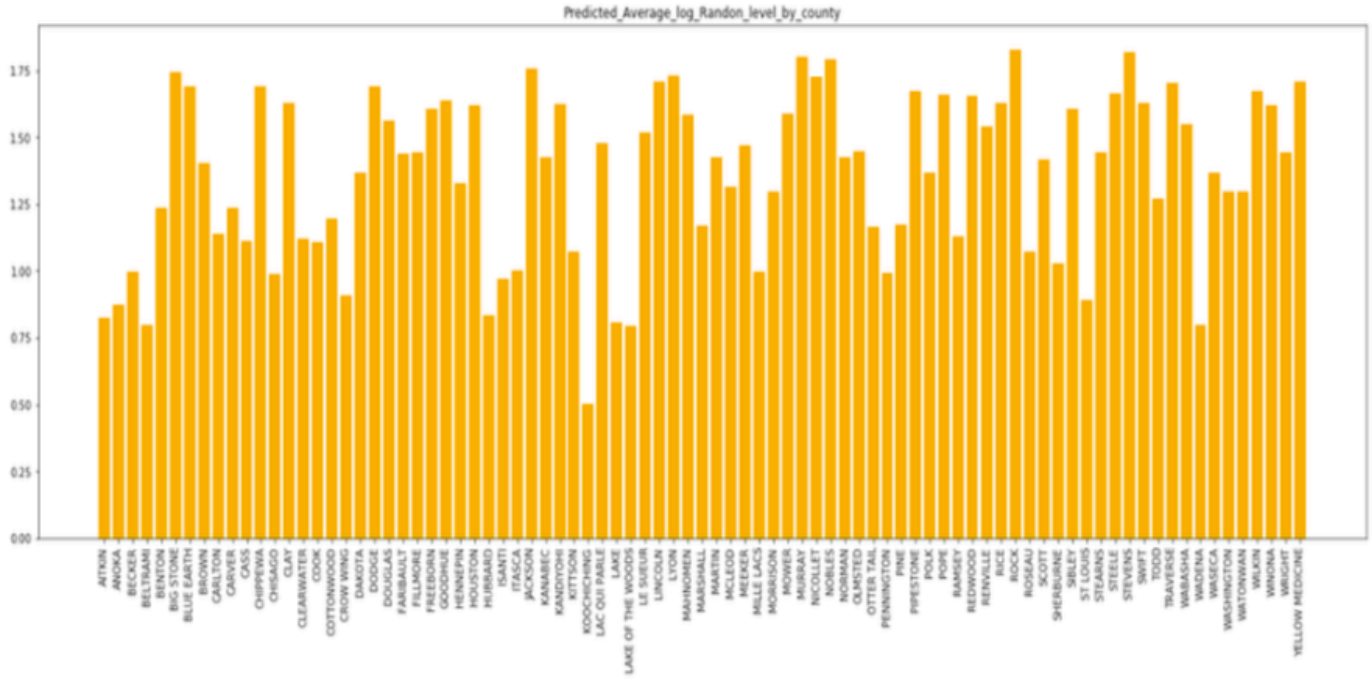
Question 9



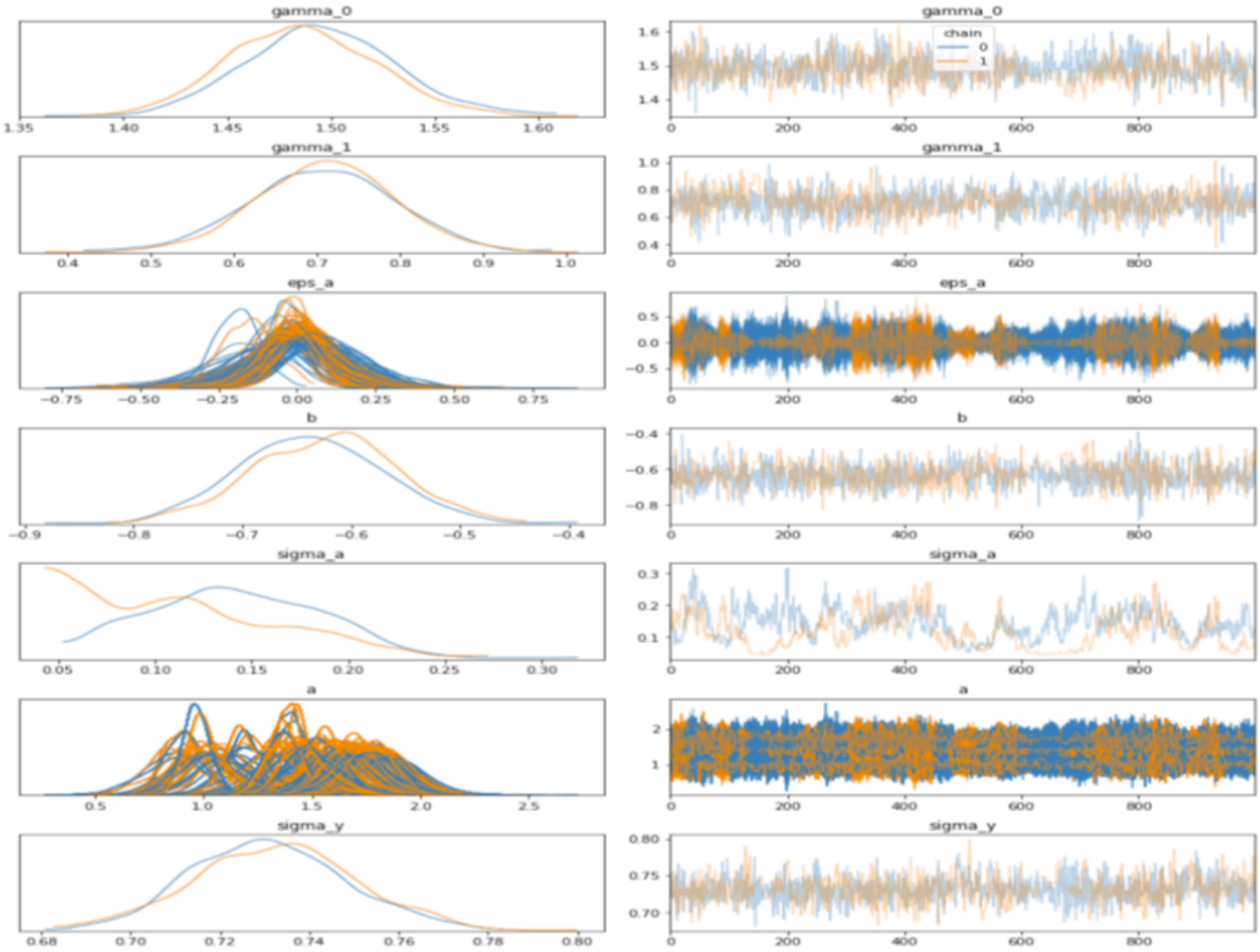
From the pic, we have two important predictors: measurement in basement or first floorand county uranium level. So we first add another dataset and then we take log. For the modeling, we use varying-intercept model, the structure of model shows below:



The predicted lowest county is 'Koochiching' and highest county is 'Rock'.



Then we plot the traces out. The trace plot shows that the sampled has converged.



Interview Question 1

(a)

SVM algorithms use a set of mathematical functions that are defined as the kernel. The kernel function is what is applied on each data instance to map the original non-linear observa- tions into a higher-dimensional space in which they become separable. Different SVM algorithms use different types of kernel functions. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

(b)

In an optimization problem, a slack variable is a variable that is added to an inequality constraint to transform it into an equality. Slack variables are positive, local quantities that relax the stiff condition of linear separability, where each training point is seeing the same marginal hyper plane. slack variables can be geometrically defined as the ratio between the distance from a training point to a marginal hyperplane, and half of the margin.

(c)

Since the value of the RBF kernel decreases with distance and ranges between zero and one, it has a ready interpretation as a similarity measure. It is a stationary kernel. The kernel function can be thought of as a cheap way of computing an infinite dimensional inner product.

(d)

We can formulate the primal optimization problem of the SVM as $min||w^2|| + c_i = 1nk_i$, the Let's rewrite $||w^2||$ as

$$||w^2|| = \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j) + b.$$

Then This leads to the dual form of the SVMs

$$w_{\alpha \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j).$$

s.t.

$0 \leq \alpha_i \leq c, \sum_i \alpha_i y_i = 0$

This optimization problem can be solved using quadratic programming, which is simple enough. The α terms can be interpreted as support vectors. Finally, since we can express the optimization problem and classification function in terms of dot products with the training data, SVMs lend themselves naturally to kernels.

(e) n^2 when C is small and n^3 when C is large.

Interview Question 3

We add a set of convex constraints to the lasso to produce sparse interaction models that honour the hierarchy restriction that interaction only be included in a model if one or both variables are marginally important. We give a precise characterization of the effect of this hierarchy constraint, prove that hierarchy holds with probability one and derive an unbiased estimate for the degrees of freedom of our estimator. A bound on this estimate reveals the amount of fitting "saved" by the hierarchy constraint. We distinguish between parameter sparsity - the number of nonzero coefficients - and practical sparsity - the number of raw variables one must measure to make a new prediction. Hierarchy focuses on the latter, which is more closely tied to important data collection concerns such as cost, time and effort.