Yunjia Zeng, collaborate with Shaoyu Feng, Mengtong Zhang
yz682

Assignment 2

ANLY 601
February 29, 2020

Question 1: Convexity

1. To show $f(x) = \sum_{i=1}^{\inf} ||x||_p; p > 0$ is convex, we first show $g(x) = ||x||_p$ is convex. By applying triangle inequality of norm, we have

$$||\lambda_v + (1-\lambda)w||_p \leq ||\lambda v||_p + ||(1-\lambda)w||_p = \lambda||v||_p + (1-\lambda)||w||_p.$$

   Therefore $g(x)$ is convex for arbitrary $p > 0$. Since $f(x)$ is a linear combination of norm, by theorem positive weighted sum of convex is convex, we conclude that $f(x)$ is also convex.

2. Suppose $k(d) = k(x, x') = x - x' = d$, $k(d) = k(x, x') = (x - x')^2 = d^2$, therefore

$$f(d) = d - d^2.$$

   By letting $v = 0, w = 1, \lambda = 0.5$, we have $f(\lambda v + (1-\lambda)w) = f(0.5) = 0.25 > f(0) + f(1) = 0$. Therefore, $f(d)$ is not convex.

3. Suppose $k(d) = k(x, x') = x - x' = d$, $k(d) = k(x, x') = (x - x')^2 = -d$, therefore

$$f(d) = -d^2 - b.$$

   By letting $b = 0, v = -1, w = 1, \lambda = 0.5$, we have $f(\lambda v + (1-\lambda)w) = f(0) = 0 > f(-1) + f(1) = -2$. Therefore, $f(d)$ is not convex.

4. First of all, $f(x) = ||x||_p - min(0, x); p > 0$ is equivalent to $f(x) = ||x||_p + min(0, -x); p > 0$. Therefore, by applying triangle inequality,

$$||\lambda_v + (1-\lambda)w||_p + min(0, -\lambda v - (1-\lambda)w) \leq ||\lambda v||_p + ||(1-\lambda)w||_p + min(0, -\lambda v) + min(0, -(1-\lambda)w),$$

   and the right hand side equals to

$$\lambda(||v||_p + min(0, -v)) + (1-\lambda)(||w||_p + min(0, -w)),$$

   by definition, $f(x)$ is convex.

5. First of all, $f(x) = ||x||_p - max(0, x); p > 0$ is equivalent to $f(x) = ||x||_p + max(0, -x); p > 0$. Therefore, by applying triangle inequality,

$$||\lambda_v + (1-\lambda)w||_p + max(0, \lambda v + (1-\lambda)w) \leq ||\lambda v||_p + ||(1-\lambda)w||_p + max(0, \lambda v) + max(0, (1-\lambda)w),$$

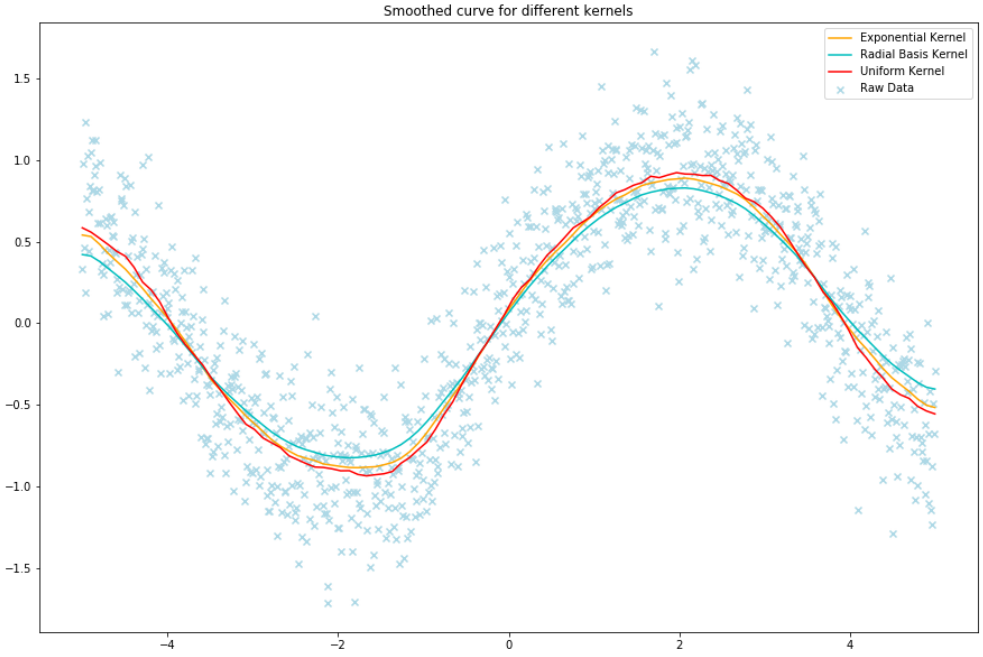
and the right hand side equals to

$$\lambda(||v||_p + max(0, v)) + (1 - \lambda)(||w||_p + max(0, w)),$$

by definition, $f(x)$ is convex.

Question 2: Kernel Regression

Part 1

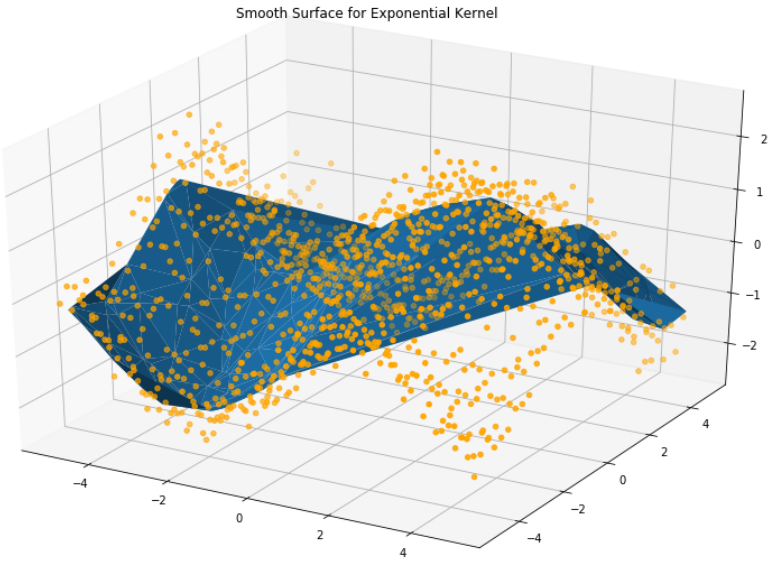


Part 2

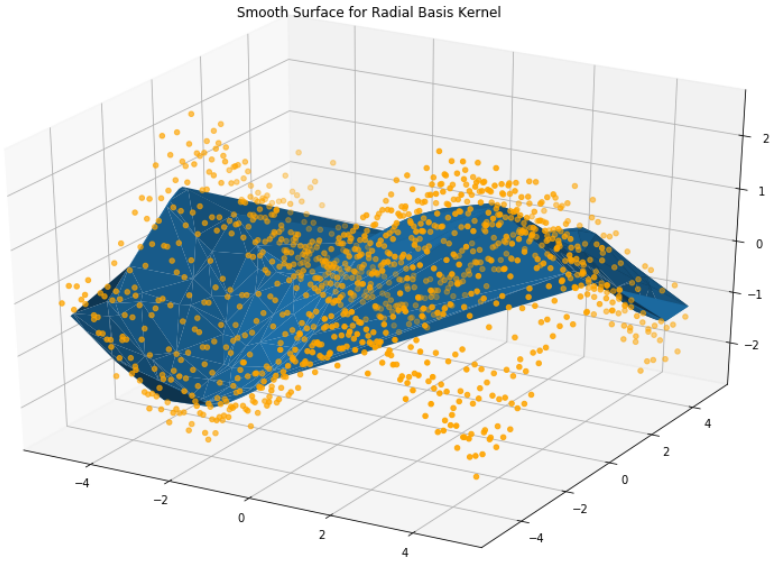
From the plot above, we can see that all three kernels fit the samples very well. There are slightly difference between them since Exponential Kernel and Radial Basis Kernel generate smoother curves rather than Uniform Kernel curve. But the Uniform Kernel is composed with piecewise polylines.

Part 3

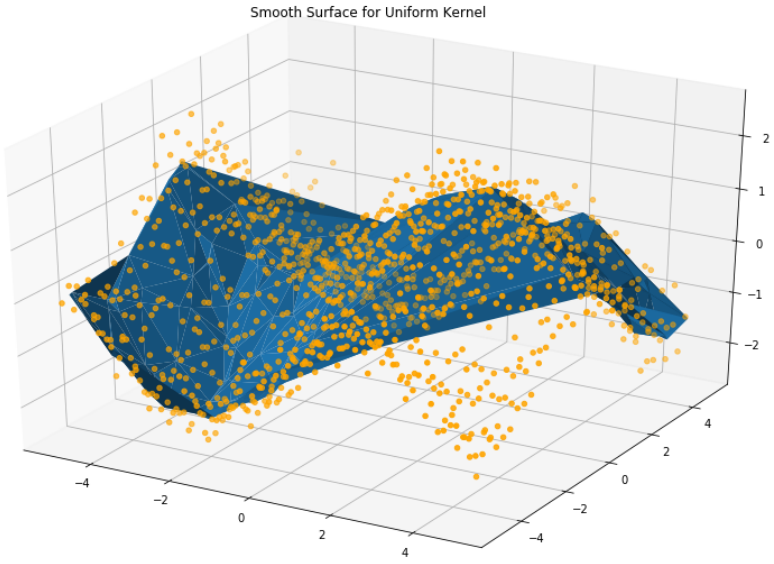
The Exponential Kernel



The Radial Basis Kernel

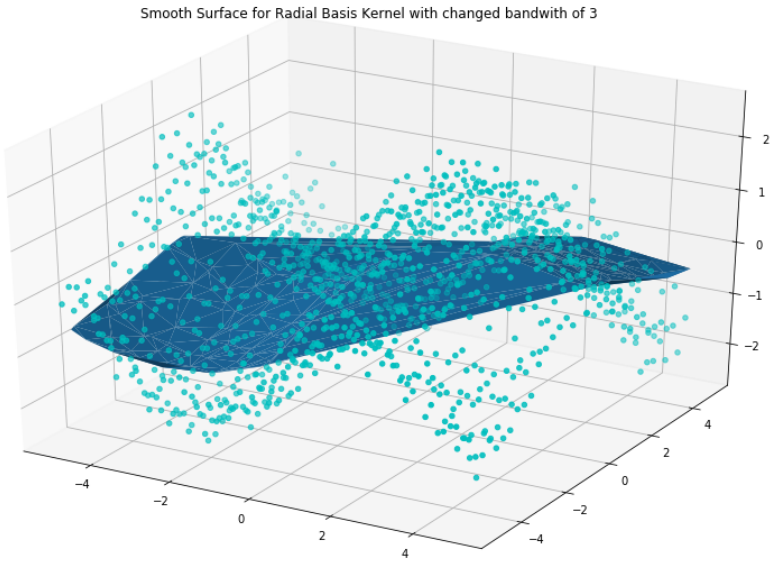


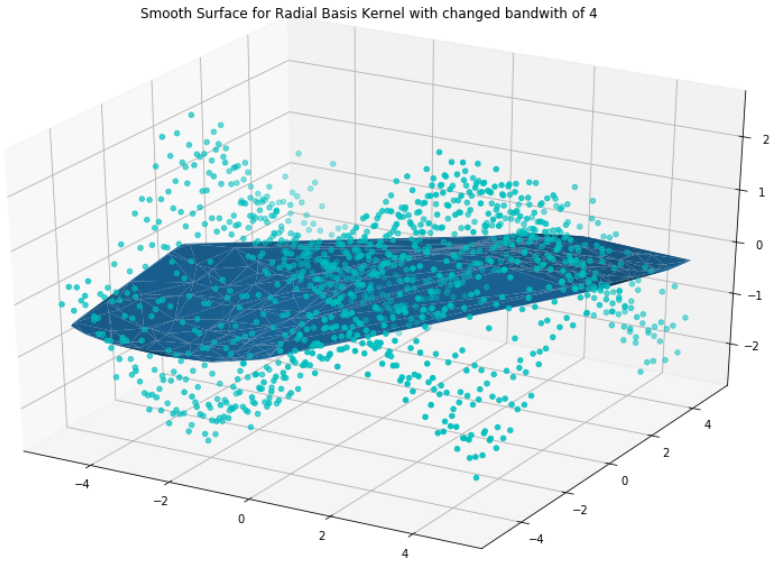
The Uniform Kernel

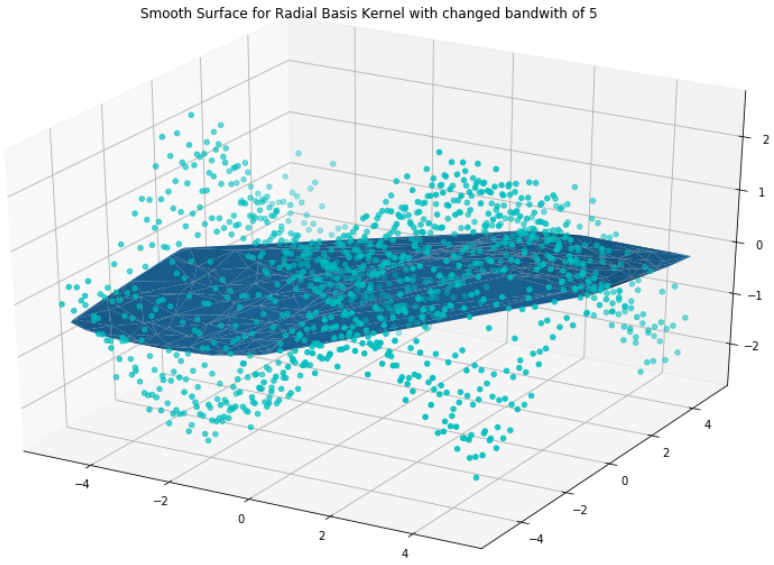


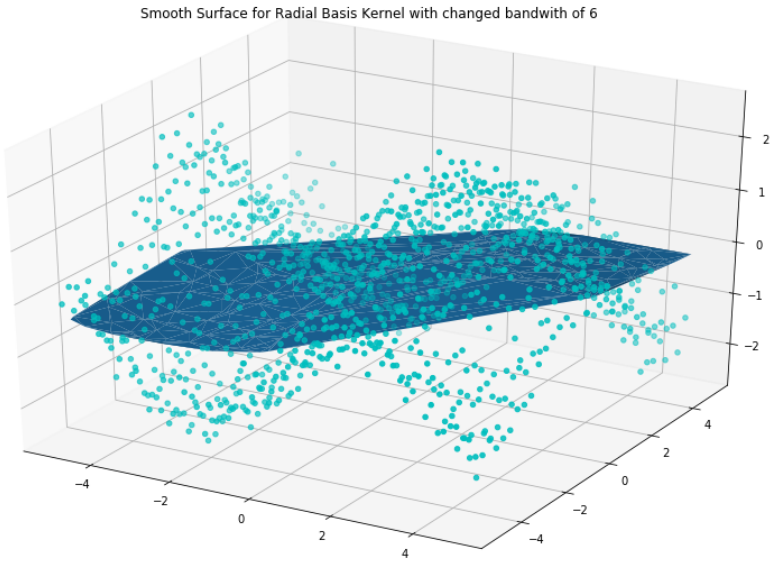
Bandwidth Change

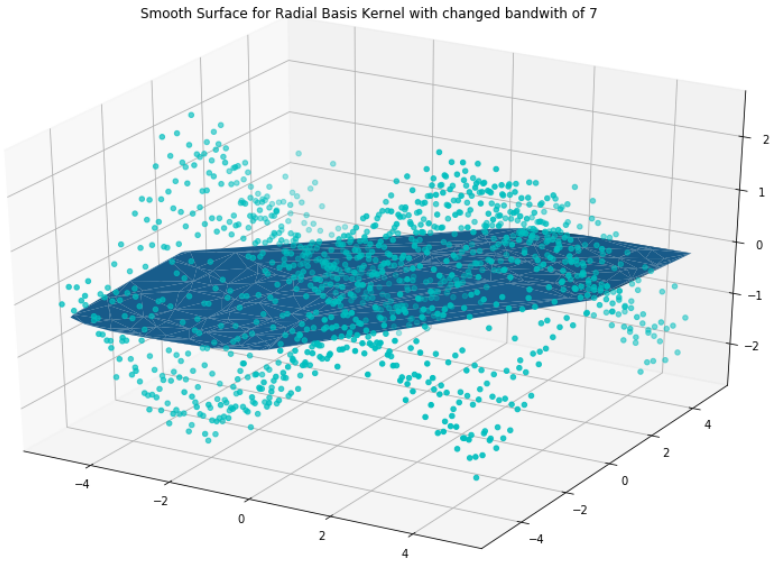
In the following plots the bandwidth of radial basis function has changed from 3 to 8. The cyan dots represent the sample data points and the blue surface represents the kernel. We can see that the kernel surface is going flatter and flatter such that the bandwidth change leads to under-fit.

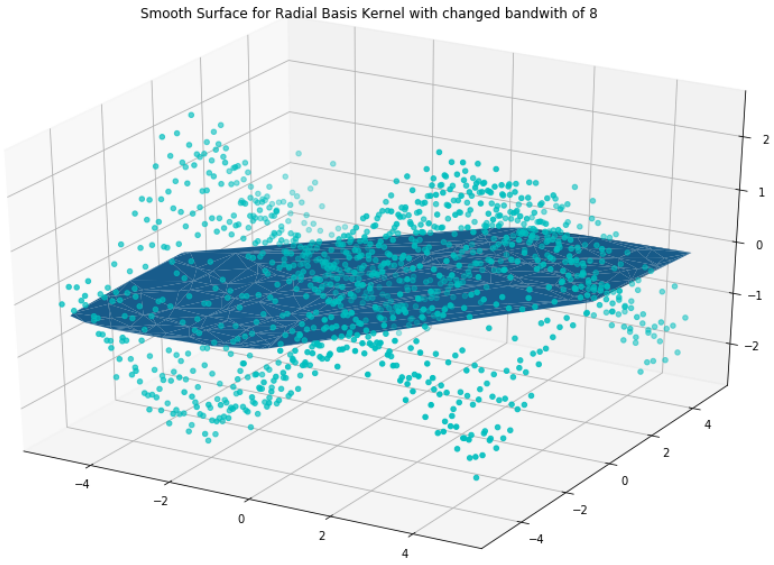

Smooth Surface for Radial Basis Kernel with changed bandwith of 3
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4


Smooth Surface for Radial Basis Kernel with changed bandwith of 4
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4


Smooth Surface for Radial Basis Kernel with changed bandwith of 5
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4

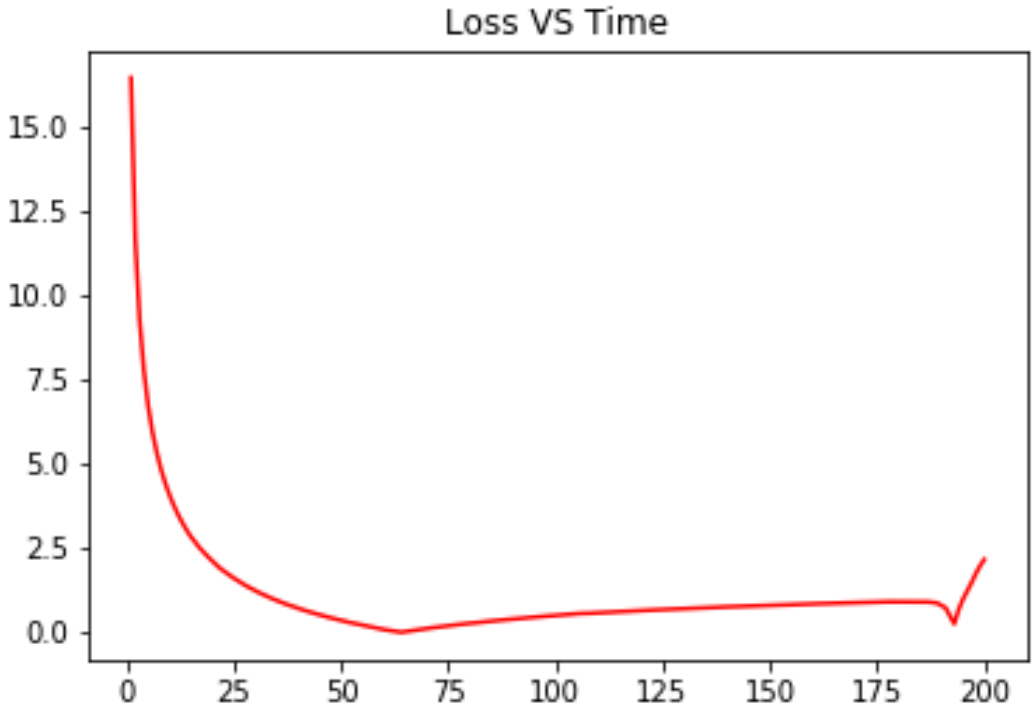

Smooth Surface for Radial Basis Kernel with changed bandwith of 6
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4


Smooth Surface for Radial Basis Kernel with changed bandwith of 7
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4

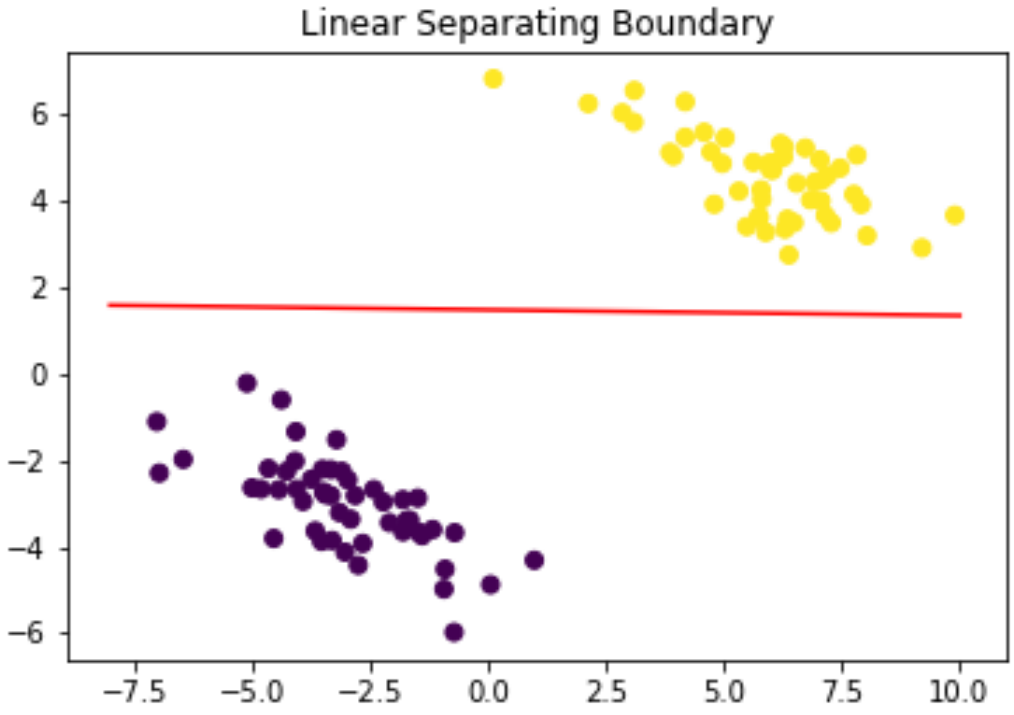

Smooth Surface for Radial Basis Kernel with changed bandwith of 8
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4

Question 3: Programming: Stochastic Subgradient Descent

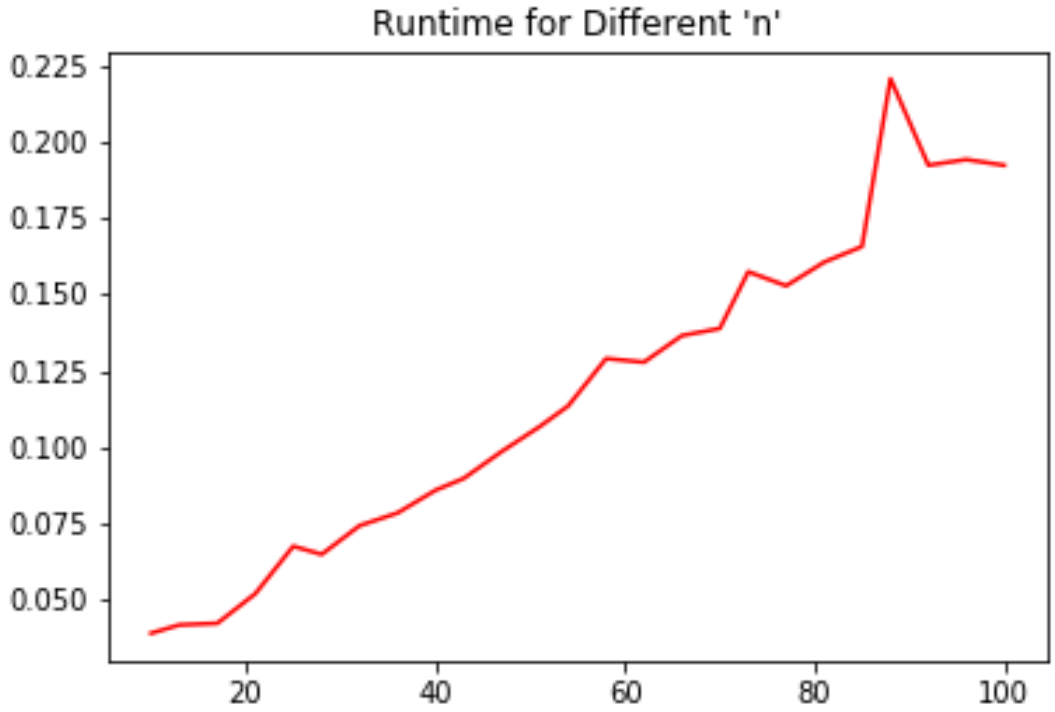
Loss Over Time



Linear Separating Boundary



Runtime Difference for Different Size of n



Question 4: Calculating the conjugate distributions

1. The posterior distribution

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

   is directly proportional to

$$e^{\frac{(\mu-(\tau\sigma+n\bar{x}\nu)/(\sigma+n\nu))^2}{2\nu\sigma/(\sigma+n\nu)}},$$

   since by given conditions, we have

$$\mu|\tau,\nu,\sigma,\mathbf{x} \sim \mathbf{Normal}(\frac{\tau\sigma+\mathbf{n}\bar{\mathbf{x}}\nu}{\sigma+\mathbf{n}\nu},\frac{\nu\sigma}{\sigma+\mathbf{n}\nu}).$$

   The posterior distribution

$$P(\sigma^2|\alpha,\beta,\mu,\mathbf{x}) \propto \frac{\beta^\alpha}{\mathbf{\Gamma}(\alpha)}(\frac{\mathbf{1}}{\sigma^\mathbf{2}})^{\alpha+\mathbf{1}}\mathbf{e}^{-\beta/\sigma^\mathbf{2}}\prod_{\mathbf{i=1}}^{\mathbf{n}}\frac{\mathbf{1}}{\sqrt{\mathbf{2}\pi}\sigma}\mathbf{e}^{-\frac{(\mathbf{x_i}-\mu)^\mathbf{2}}{\mathbf{2}\sigma^\mathbf{2}}}$$

is directly proportional to

$$(\frac{1}{\sigma^2})^{\alpha+1+n/2}e^{-(2\beta+\Sigma_{i=1}^n(x_i-\mu)^2)/2\sigma^2},$$

since by given conditions, we have

$$\sigma^2|\alpha,\beta,\mu,\mathbf{x} \sim \mathbf{InverseGamma}(\alpha+\mathbf{n/2},\beta+\mathbf{\Sigma_{i=1}^n}(\mathbf{x_i}-\mu)^\mathbf{2}/\mathbf{2}).$$

2. The posterior distribution

$$P(p_1,p_2,...,p_k|x_{ij},i=1,2,...k,j=1,2,...,k) \propto \prod_{i=1}^n p_i^{\alpha_i-1}\prod_{i=1}^n\frac{n!}{x_{i1}!...x_{ik}!}p_1^{x_{i1}}...p_k^{x_{ik}}$$

is directly proportional to

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

since by given conditions, we have

$$p_1,p_2,...p_k|x_{ij}(i=1,2,...k,j=1,2,...,k) \sim Dirichlet(\alpha_1+\Sigma_{j=1}^n x_{j1}-1,...,\alpha_k+\Sigma_{j=1}^n x_{jk}-1).$$

3. The posterior distribution

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

is directly proportional to

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

since by given conditions, we have

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

Question 5: Prior as regularizers

1. To show that $L2$ penalty is equivalent to a Normal prior, we first suppose we are estimating $\beta = (\beta_1, ..., \beta_p)$ with prior distribution of β_j as $N(0, \tau^2)$, therefore

$$\begin{aligned}\hat{\beta}_{MAP} &= \arg\max_{\beta} P(\beta|y) \\ &= \arg\max_{\beta} \frac{P(y|\beta)P(\beta)}{P(y)} \\ &= \arg\max_{\beta} P(y|\beta)P(\beta) \\ &= \arg\max_{\beta} \log(P(y|\beta)P(\beta)) \\ &= \arg\max_{\beta} \log P(y|\beta) + \log P(\beta)\end{aligned}$$

and

$$\begin{aligned}&\arg\max_{\beta} \Big[\log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p}))^2}{2\sigma^2}} + \log \prod_{j=0}^{p} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{\beta_j^2}{2\tau^2}} \Big] \\ &= \arg\max_{\beta} \Big[-\sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))^2}{2\sigma^2} - \sum_{j=0}^{p} \frac{\beta_j^2}{2\tau^2} \Big] \\ &= \arg\min_{\beta} \frac{1}{2\sigma^2} \big[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}))^2 + \frac{\sigma^2}{\tau^2} \sum_{j=0}^{p} \beta_j^2 \big] \\ &= \arg\min_{\beta} \big[\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}))^2 + \lambda \sum_{j=0}^{p} \beta_j^2 \big]\end{aligned}$$

Therefore, from the results above, we can see the target function of maximum posterior estimation is equivalent to ridge regression. Thus, The L2 penalty (ridge) is equivalent to a Normal prior.

2. To show that $L1$ penalty is a LaPlace priors, we show that

$$\begin{aligned}
&\arg\max_{\beta}\Big[\log\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y_i-(\beta_0+\beta_1x_{i,1}+\ldots+\beta_px_{i,p}))^2}{2\sigma^2}}\\
&=\arg\max_{\beta}\Big[-\sum_{i=1}^{n}\frac{(y_i-(\beta_0+\beta_1x_{i,1}+...+\beta_px_{i,p}))^2}{2\sigma^2}-\sum_{j=0}^{p}\frac{|\beta_j|}{2b}\Big]\\
&=\arg\min_{\beta}\frac{1}{2\sigma^2}\big[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+...+\beta_px_{i,p}))^2+\frac{\sigma^2}{b}\sum_{j=0}^{p}|\beta_j|\big]\\
&=\arg\min_{\beta}\big[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+\ldots+\beta_px_{i,p}))^2+\lambda\sum_{j=0}^{p}|\beta_j|\big]+\log\prod_{j=0}^{p}\frac{1}{2b}e^{-\frac{|\beta_j|}{2b}}\Big]
\end{aligned}$$

Therefore, from the results above we see that the target function of maximum posterior estimation is equivalent to LASSO regression. Thus, The $L1$ penalty (LASSO) is equivalent to a Laplace prior.

Question 6: General questions?

1. Firstly, the most obvious difference, is that the posterior distribution depends on the unknown parameter θ,

$$p(\theta|x)=c\cdot p(x|\theta)p(\theta),$$

where c is the normalizing constant. But the posterior predictive distribution does not depend on the unknown parameter θ,

$$p(x^*|x)=\int_\theta c\cdot p(x^*,\theta|x)d\theta=\int_\theta c\cdot p(x^*|\theta)p(\theta|x)d\theta.$$

The posterior distribution is the distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained, therefore its the distribution that explains unknown, random, parameter. But the posterior predictive distribution is the distribution for future predicted data based on the data that we have already seen, therefore it's used to predict new data values.

2. As explained in the previous question, I will use the posterior predictive distribution to predict future values of X.

3. To show the change of μ and the change of σ, we first calculate the μ and σ^2 of MLE

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

By taking log on both sides, we obtain the following

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

Therefore,

$$\frac{d\mathcal{L}}{d\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \mid_{\mu}= 0 \implies \frac{1}{2\sigma^2}\sum_{i=1}^{n}(2\hat{\mu} - 2x_i) = 0,$$

and

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

such that

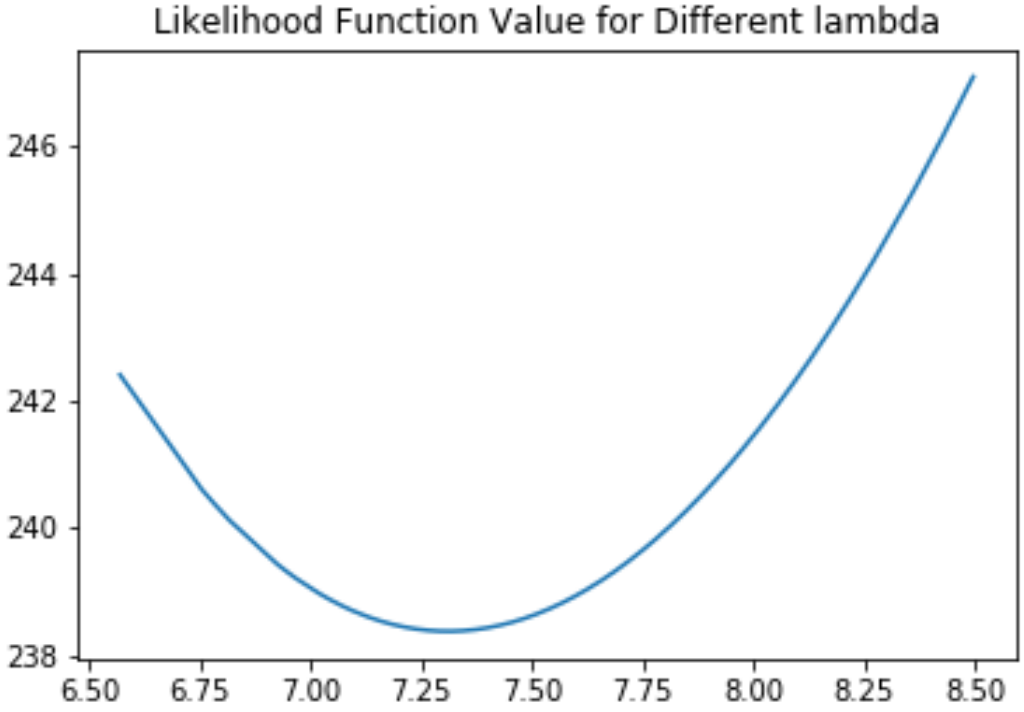
$$\begin{aligned}
\hat{\mu}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\hat{\sigma}^2_{MLE} &= \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n}.
\end{aligned}$$

According to the problem 1 of Question 4, we can then obtain

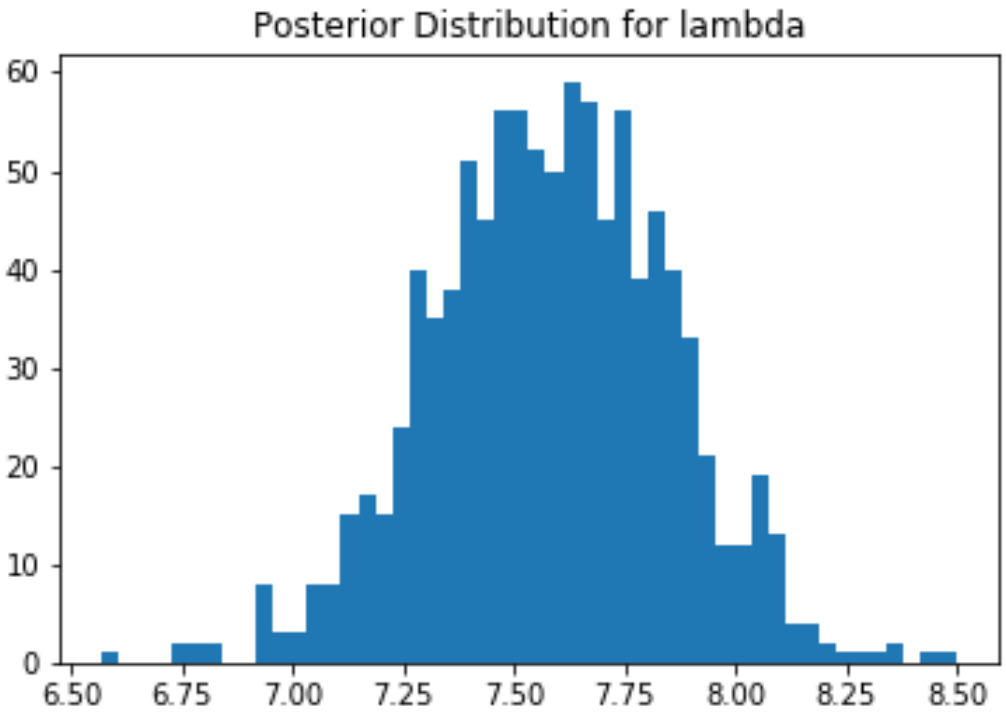
$$\begin{aligned}
\hat{\mu}_{MAP} = \frac{\alpha\sigma + n\bar{x}\beta}{\sigma + n\beta} &\implies \bar{x} = \hat{\mu}_{MLE}(n \uparrow) \\
\hat{\sigma^2}_{MAP} = \frac{\nu + \Sigma_{i=1}^{n}(x_i-\mu)^2/2}{\tau + n/2 - 1} &\implies \frac{\Sigma_{i=1}^{n}(x_i-\mu)^2}{n} = \hat{\sigma^2}_{MLE}(n \uparrow)
\end{aligned}$$

Question 7: Programming a Gibbs Sampler

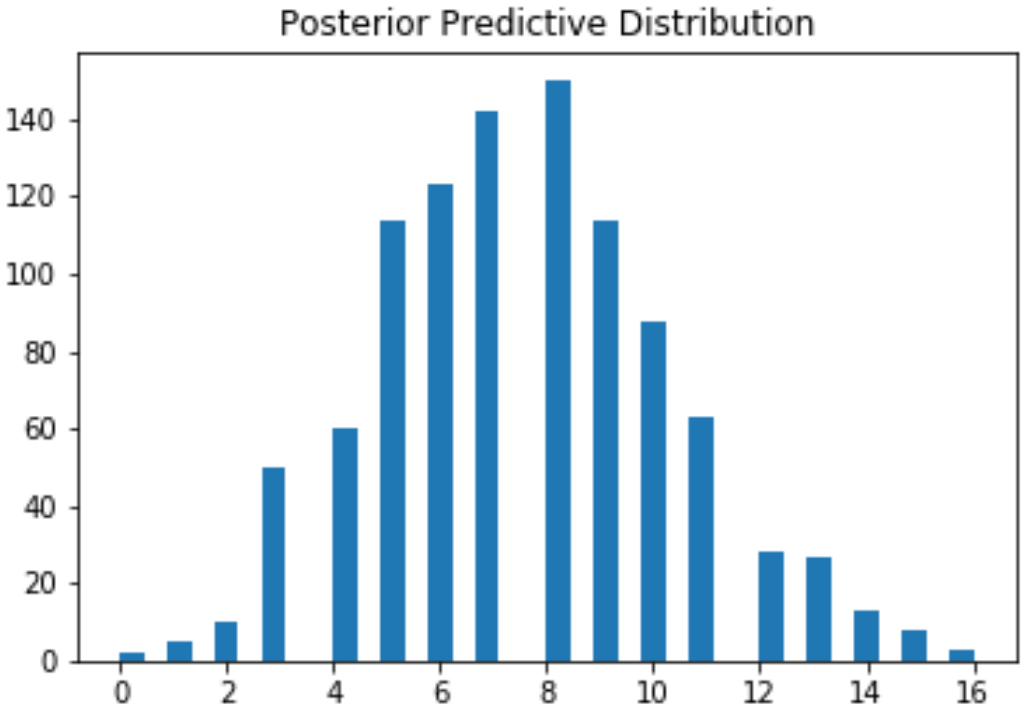
Likelihood



Plot of Posterior Distribution

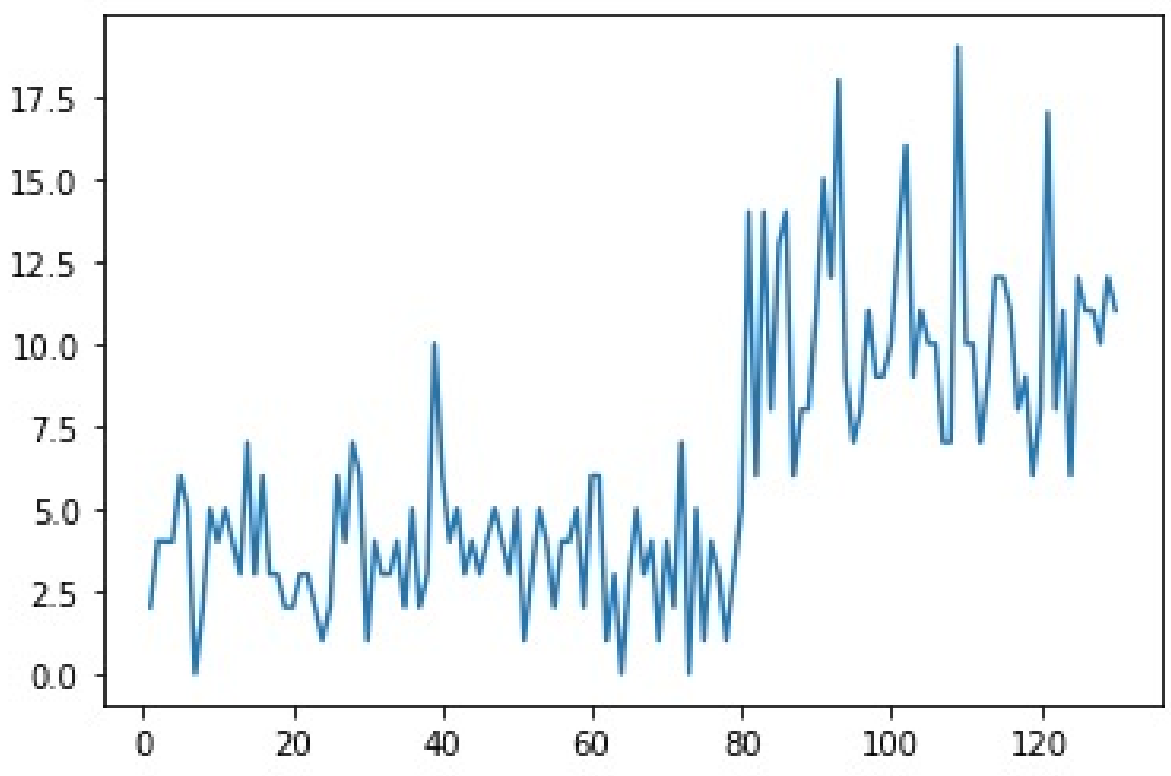


Plot of Posterior Predictive Distribution



Question 8: Change points models

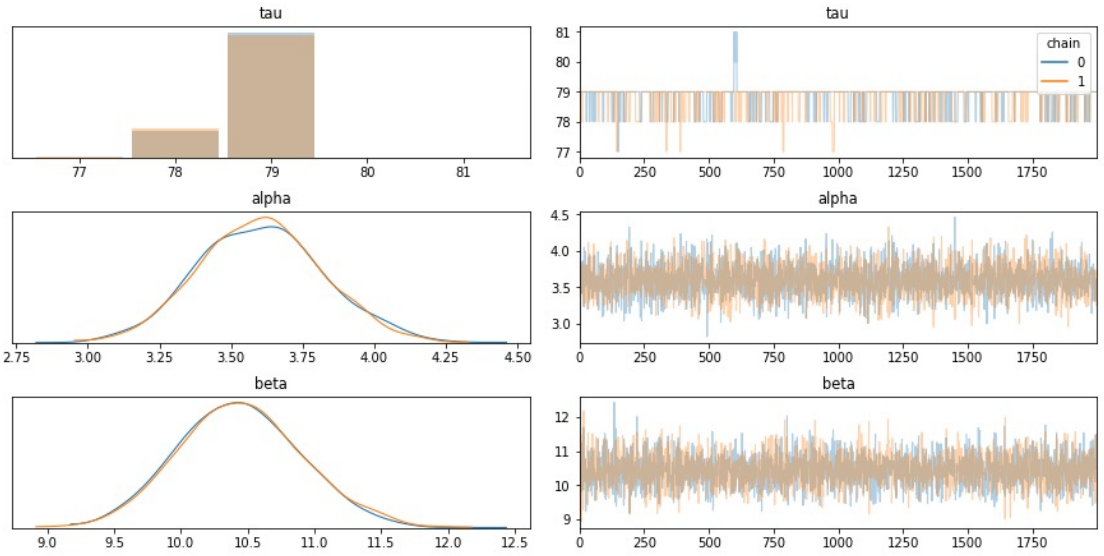
Basic Change Point Model



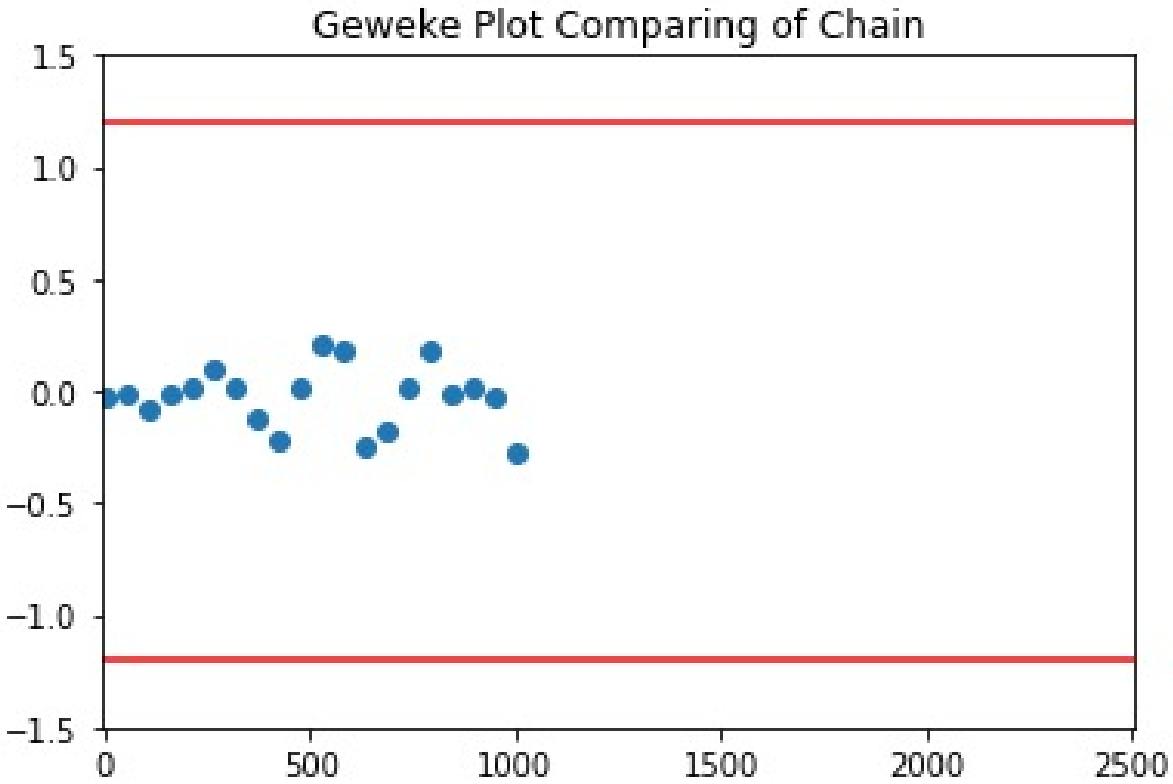
Originally we have $\lambda_t \sim exponential(\alpha)$. After change point

$$f(\alpha|x) \propto const * e^{-\tau\lambda}\lambda^{\sum_{i=1}^{\tau} x_i} * e^{-\alpha\lambda}$$
$$f(\alpha|x) \propto \lambda^{\sum_{i=1}^{\tau} x_i} e^{-(\tau+\alpha)\lambda}$$

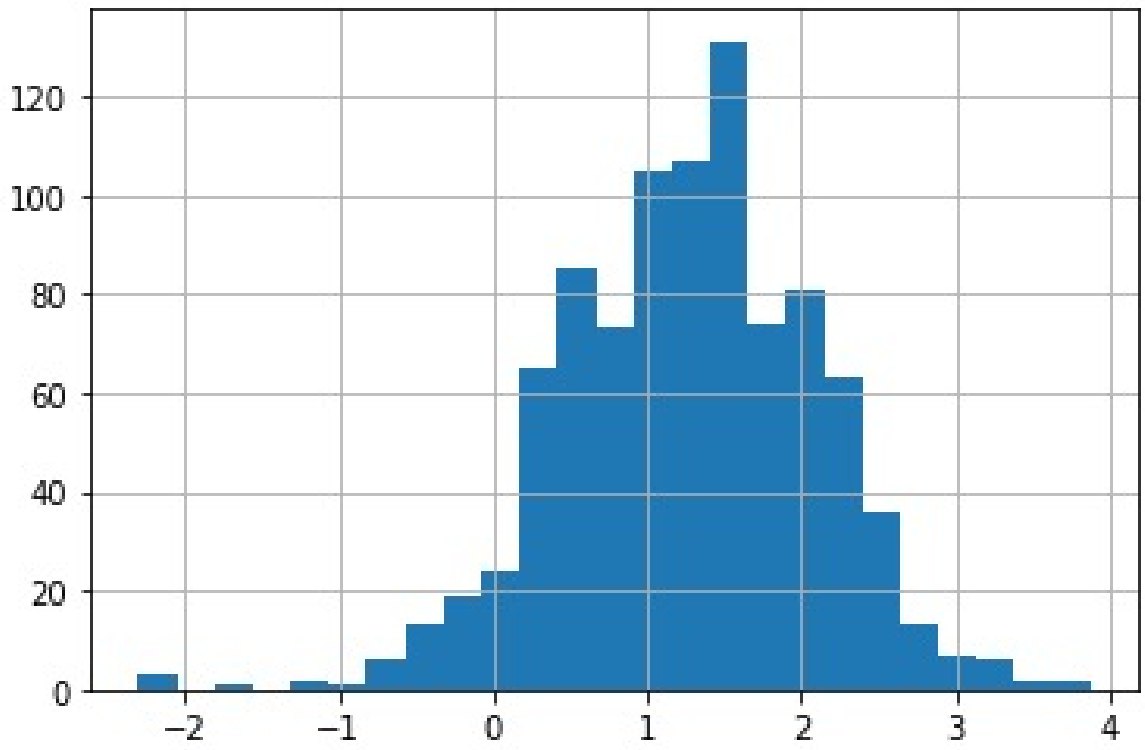
Although It follows a Gamma Distribution $(\sum x_i + 1, \alpha + \tau)$, we can estimate all α, β, τ through MCMC by using the Bayesian approach so that



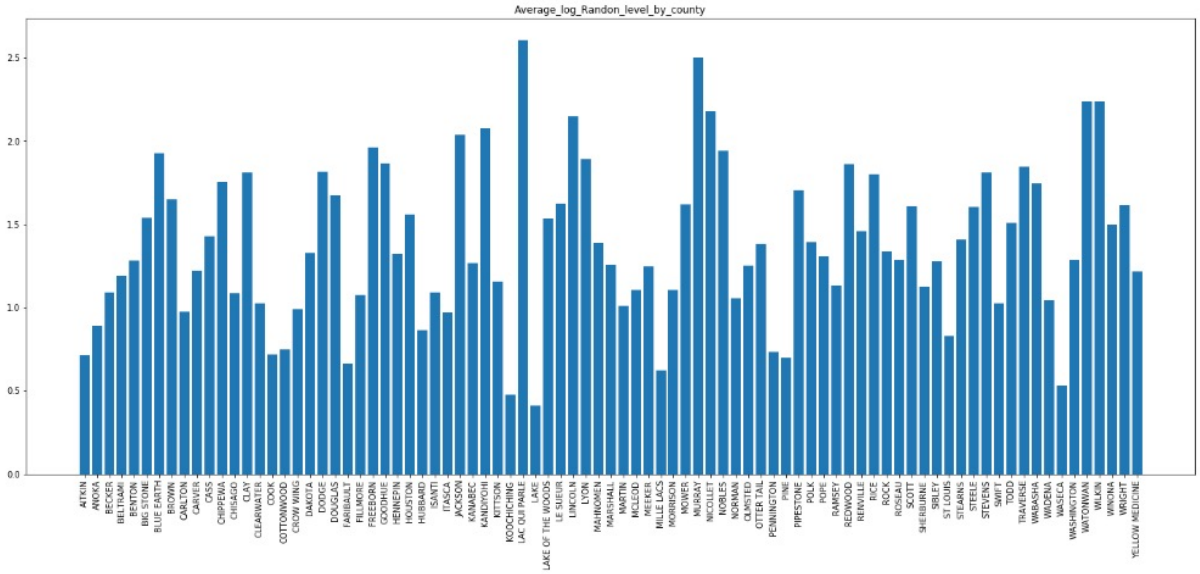
We observe from the histogram that the τ is mostly identified as 79, and multiple-chains plot shows similar results. Therefore the sampler is converged. We can also show the convergence through the Geweke Plot below, since the test value is stable and have a small value, which suggests that the sampler is converged.



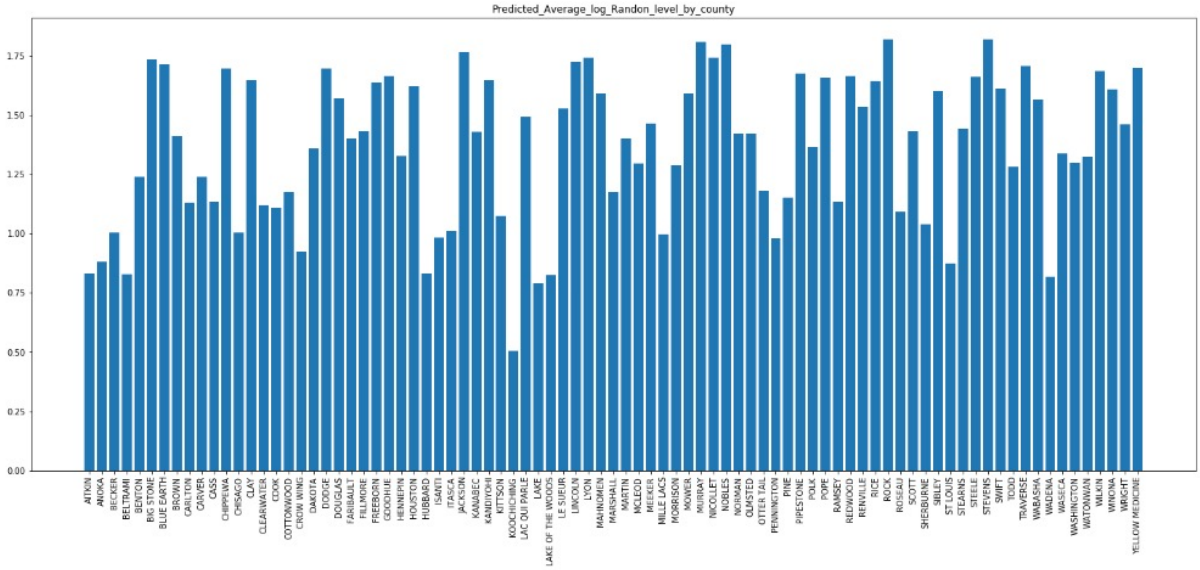
Question 9: Programming a hierarchical model using PYMC3



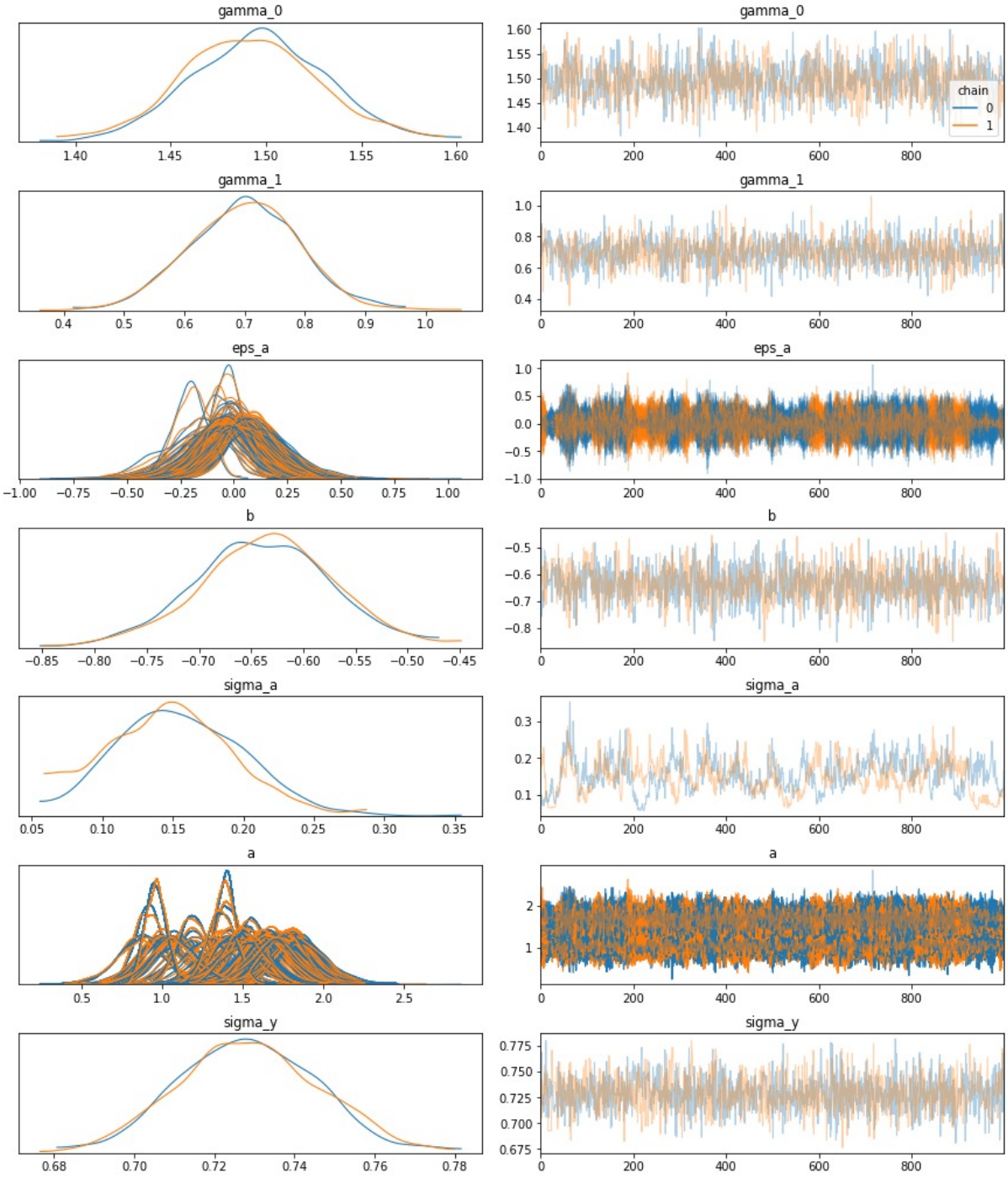
The two important predictors are first the measurement in basement or first floor (radon higher in basements); second the county uranium level (positive correlation with radon levels). We first get county uranium by adding another dataset and then take log of the radon for response variable.



From the results above the County 'Lake' has the lowest radon count while 'Lac Qui Parle' has the highest average radon level in raw data.



From the results above we use the varying-intercept model as described in the post, The predicted lowest county is 'Koochiching' and highest county is 'Rock', which does not match the results from raw data. We plot the trace and observe the sampled has converged, as different chains demonstrated similar results and variations is relatively small.


gamma_0
gamma_0
chain
0
1
gamma_1
gamma_1
eps_a
eps_a
b
b
sigma_a
sigma_a
a
a
sigma_y
sigma_y

Interview Questions

1

a. The function of kernel is to take data as input and transform it into the required form. For example there are linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. The most used type of kernel function for SVMs is RBF, since it has localized and finite response along the entire x-axis. The kernel functions return the inner product between two points in a suitable feature space, thus by defining a notion of similarity, with little computational cost even in very high-dimensional spaces.

b. The slack variables are the variable that added to an inequality constraint to transform it into an equality. They relax the stiff condition of linear separability, where each training point is seeing the same marginal hyperplane. Large penalty will reduce the margin while small penalty tends to increase the margin, and slack variables can be geometrically defined as the ratio between 1/2 margin and the distance from a training point to a marginal hyperplane.

c. This is because the value of the RBF kernel decreases with distance and ranges in the limit between zero and one, it has a ready interpretation as a similarity measure. RBF is invariant to translation since it is a stationary kernel. The kernel function of RBF can be thought of as a cheap way of computing an infinite dimensional inner product.

d. The objective of SVM is to minimize $||w^2|| + c\sum_{i=1} n\zeta_i$ where ζ is the slack variable. We can rewrite the form as

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

and therefore the dual form of SVMs will be

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

The optimization problem above can be solved by quadratic programming. The α terms can be interpreted as support vectors, which in practice is usually a sparse subset of the data. Since we defined the optimization problem and classification function in terms of dot products with the training data, SVMs lend themselves naturally to kernels so that we can perform linear classification in finite high dimensional spaces.

e. The runtime for optimization grows like n^2 when C is small and n^3 when C gets large.

2

3

Hierarchical models are more flexible in modeling the continuum from all groups have the same parameters to all groups that have completely different parameters compare to other models. If the mean of each group are similar or identical then we will get small σ^2 and the resulting inference for the individual θ will be very close to the same as just assumed a common mean for all groups. In contrast, if the groups have very different means, we will have large σ^2 and the resulting inference for the individual θ will be very close to the same as not having the hierarchical model at all. Therefore, we don't need to choose whether to use a model with a common mean for all groups or a completely independent mean for all groups, the hierarchical model allows the data to tell where we fell along that continuum.

Another advantage of hierarchical models is that when the number of observations in each group varies a lot, the groups with smaller numbers of observations will have improved inference about their group parameters by borrowing information via the hierarchical model about the group specific parameters. This is also the reason why hierarchical models provide better model fits and regularization when data is sparse.