Homework 3 for **EECS E6720**

submitted to Professor John Paisley

Advait Rajagopal

19 November 2017

1 Answer 1

We have data $\{(x_i, y_i)\}_{i=1}^N$ data points and the following regression problem;

$$
\begin{aligned}
y_i &\overset{ind}{\sim} \text{Normal}(x_i^T w, \lambda^{-1}) \\
w &\sim \text{Normal}(0, \text{diag}(\alpha_1, ..., \alpha_d)^{-1}) \\
\lambda &\sim \text{Gamma}(e_0, f_0) \\
\alpha_k &\overset{iid}{\sim} \text{Gamma}(a_0, b_0)
\end{aligned}
$$

The goal is to approximate the posterior $p(w, \alpha_1, ..., \alpha_d, \lambda | y, x) \approx q(w, \alpha_1, ..., \alpha_d, \lambda)$

1.1 Part A

I use the factorization $q(w, \alpha_1, ..., \alpha_d, \lambda) = q(w)q(\lambda)\prod_{k=1}^d q(\alpha_k)$ as provided in the question.
We follow 3 steps to find the optimal q distribution for the required parameters;

1. Take log of joint likelihood
2. Take expectation of this using all other q distributions except the one of interest
3. Exponentiate the result and normalize over the variable of interest

The first step is to write the joint likelihood function and take its log;

$$
p(y, w, \alpha_1, ..., \alpha_d, \lambda | x) = p(w|0, \text{diag}(\alpha_1, ..., \alpha_d)^{-1}) p(\lambda) \prod_{k=1}^d p(\alpha_k) \prod_{i=1}^N p(y_i | x_i^T w, \lambda^{-1})
$$

For the purpose of this assignment I write $p(w|0, \text{diag}(\alpha_1, ..., \alpha_d)^{-1})$ as $p(w|\alpha)$ which is the prior distribution of w with parameter α. Taking the log of the joint likelihood above gives us the following equation;

$$
\ln(p(y, w, \alpha_1, ..., \alpha_d, \lambda | x)) = \ln p(w|\alpha) + \ln p(\lambda) + \sum_{k=1}^d \ln p(\alpha_k) + \sum_{i=1}^N \ln p(y_i | x_i^T w, \lambda^{-1})
$$

Now I follow the second step of taking expectation over this equation with respect to all other q distributions except the one of interest. I ignore terms that are not functions of the parameter under consideration and write the form where they have been dropped with the normalizing constant.

Deriving $q(w)$

$$
\begin{aligned}
q(w) &\propto \exp\Big\{\mathbb{E}_{q(\lambda)}\mathbb{E}_{q(\alpha)}[\ln p(w|\alpha) + \ln p(\lambda) + \sum_{k=1}^{d}\ln p(\alpha_k) + \sum_{i=1}^{N}\ln p(y_i|x_i^T w, \lambda^{-1})]\Big\} \\
&\propto \exp\Big\{\mathbb{E}_{q(\lambda)}\mathbb{E}_{q(\alpha)}\Big[-\frac{\sum_{i=1}^{N}(y_i - x_i^T w)^2}{2\lambda^{-1}} - \frac{w^T\Sigma_\alpha w}{2}\Big]\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\Big[\mathbb{E}_{q(\lambda)}[\lambda]\sum_{i=1}^{N}(y_i^2 - 2w^T x_i y_i + w^T x_i x_i^T w) + w^T\mathbb{E}_{q(\alpha)}[\Sigma_\alpha]w\Big]\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\Big[-2\mathbb{E}_{q(\lambda)}[\lambda]w^T\sum_{i=1}^{N}x_i y_i + \mathbb{E}_{q(\lambda)}[\lambda]w^T\sum_{i=1}^{N}x_i x_i^T w + w^T\mathbb{E}_{q(\alpha)}[\Sigma_\alpha]w\Big]\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\Big[w^T(\mathbb{E}_{q(\alpha)}[\Sigma_\alpha] + \mathbb{E}_{q(\lambda)}[\lambda]\sum_{i=1}^{N}x_i x_i^T)w - 2\mathbb{E}_{q(\lambda)}[\lambda]w^T\sum_{i=1}^{N}x_i y_i\Big]\Big\}
\end{aligned}
$$

This last equation looks like step 3 in the derivation of the posterior of w in Lecture 2. So I now write the final form of the $q(w)$ distribution.

$$
\begin{aligned}
q(w) &\sim \mathcal{N}(\mu_w, \Sigma_w) \\
\text{where;}\quad \mu_w &= \Sigma_w\Big(\mathbb{E}_{q(\lambda)}[\lambda]\sum_{i=1}^{N}x_i y_i\Big) \\
\Sigma_w &= \Big(\mathbb{E}_{q(\alpha)}[\Sigma_\alpha] + \mathbb{E}_{q(\lambda)}[\lambda]\sum_{i=1}^{N}x_i x_i^T\Big)^{-1}
\end{aligned}
$$

Deriving $q(\lambda)$

$$
\begin{aligned}
q(\lambda) &\propto \exp\Big\{\mathbb{E}_{q(w)}\mathbb{E}_{q(\alpha)}[\ln p(w|\alpha) + \ln p(\lambda) + \sum_{k=1}^{d}\ln p(\alpha_k) + \sum_{i=1}^{N}\ln p(y_i|x_i^T w, \lambda^{-1})]\Big\} \\
&\propto \exp\Big\{\mathbb{E}_{q(w)}\mathbb{E}_{q(\alpha)}\Big[(e_0 - 1)\ln\lambda - f_0\lambda + \frac{N}{2}\ln\lambda - \frac{\sum_{i=1}^{N}(y_i - x_i^T w)^2}{2\lambda^{-1}}\Big]\Big\} \\
&\propto \exp\Big\{(e_0 - 1)\ln\lambda - f_0\lambda + \frac{N}{2}\ln\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(y_i^2 - 2y_i x_i^T\mathbb{E}_{q(w)}[w] + x_i^T\mathbb{E}_{q(w)}[ww^T]x_i)\Big\} \\
&\propto \exp\Big\{\ln\lambda^{(e_0 + \frac{N}{2} - 1)} - f_0\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(y_i^2 - 2y_i x_i^T\mathbb{E}_{q(w)}[w] + x_i^T\mathbb{E}_{q(w)}[ww^T]x_i)\Big\} \\
&\propto \exp\Big\{\ln\lambda^{(e_0 + \frac{N}{2} - 1)} - \lambda\Big[f_0 + \frac{1}{2}\sum_{i=1}^{N}(y_i^2 - 2y_i x_i^T\mathbb{E}_{q(w)}[w] + x_i^T\mathbb{E}_{q(w)}[ww^T]x_i\Big]\Big\}
\end{aligned}
$$

This last equation starts to look like a Gamma distribution. I write the density function below.

$$
\begin{aligned}
q(\lambda) &\sim \text{Gamma}(e', f') \\
\text{where;} \quad e' &= e_0 + \frac{N}{2} \\
f' &= f_0 + \frac{1}{2}\sum_{i=1}^{N}(y_i^2 - 2y_i x_i^T \mathbb{E}_{q(w)}[w] + x_i^T \mathbb{E}_{q(w)}[ww^T]x_i
\end{aligned}
$$

Deriving $q(\alpha_k)$

I now solve for one $q(\alpha_k)$;

$$
\begin{aligned}
q(\alpha_k) &\propto \exp\Big\{\mathbb{E}_{q(w)}\mathbb{E}_{q(\lambda)}\mathbb{E}_{q(\alpha_{-k})}[\ln p(w|\alpha) + \ln p(\lambda) + \sum_{k=1}^{d}\ln p(\alpha_k) + \sum_{i=1}^{N}\ln p(y_i|x_i^T w, \lambda^{-1})]\Big\} \\
&\propto \exp\Big\{\frac{1}{2}\ln\alpha_k - \frac{\alpha_k \mathbb{E}_{q(w)}[w^2]}{2} + (a_0 - 1)\ln\alpha_k - b_0\alpha_k\Big\}
\end{aligned}
$$

This looks like a Gamma distribution for a single α_k. I write the density function below.

$$
\begin{aligned}
q(\alpha_k) &\sim \text{Gamma}(a', b') \\
\text{where;} \quad a' &= a_0 + \frac{1}{2} \\
b' &= b_0 + \frac{\mathbb{E}_{q(w)}[w^2]}{2}
\end{aligned}
$$

Therefore all $q(\alpha_k)$;

$$
q(\alpha_1, ..., \alpha_d) = \prod_{k=1}^{d} q(\alpha_k)
$$

1.2 Part B

Variational Inference algorithm pseudo-code

Input: Data and definitions of $q(w) = \mathcal{N}(w|\mu_w, \Sigma_w)$, $q(\lambda) = \text{Gamma}(e', f')$ and $q(\alpha_k) = \text{Gamma}(a', b')$
Output: Values for $\mu_w, \Sigma_w, e', f', a'$ and b'

1. Initialize values for $e_{0_0}, f_{0_0}, a_{0_0}$ and b_{0_0}
2. For iteration t = 1,. . .,T
   - Update $q(\lambda)$ by setting

$$
\begin{aligned}
e'_t &= e_0 + \frac{N}{2} \\
f'_t &= f_0 + \frac{1}{2}\sum_{i=1}^{N}(y_i^2 - 2y_i x_i^T \mu_{w_{t-1}} + x_i^T \Sigma_{w_{t-1}} x_i)
\end{aligned}
$$

- Update $q(\alpha_k)$ by setting

$$a'_t = a_0 + \frac{1}{2}$$
$$b'_t = b_0 + \frac{\mu^2_{w_{t-1}} + \Sigma_{w_{t-1}}}{2}$$

- Update $q(w)$ by setting

$$\Sigma_{w_t} = \Big(diag(a'_t/b'_t) + \frac{e'_t}{f'_t}\sum_{i=1}^{N} x_i x_i^T\Big)^{-1}$$
$$\mu_{w_t} = \Sigma_{w_t}\Big(\frac{e'_t}{f'_t}\sum_{i=1}^{N} x_i y_i\Big)$$

  where $diag(a'_t/b'_t)$ is a d dimensional diagonal matrix.

- Evaluate $\mathcal{L}(\mu_w, \Sigma_w, e', f', a', b')$ to assess convergence (i.e. decide T). Note that the variational inference objective function $\mathcal{L}$ is calculated in Part C.

1.3 Part C

The variational inference objective function is given by the following equation.

$$\begin{aligned}
\mathcal{L}(\mu_w, \Sigma_w, e', f', a', b') &= \mathbb{E}_q[\ln\, p(y, w, \alpha_1, ...\alpha_d, \lambda|x)] - \mathbb{E}_q[\ln\, q(w, \alpha_1, ...\alpha_d, \lambda)] \\
&\propto \frac{1}{2}\sum_{k=1}^{d}\mathbb{E}_{q(\alpha_k)}[\ln\,\alpha_k] - \frac{1}{2}\sum_{k=1}^{d}\mathbb{E}_{q(\alpha_k)}\alpha_k\mathbb{E}_{q(w)}[w^2] \\
&+ (e_0 - 1)\mathbb{E}_{q(\lambda)}[\ln\,\lambda] - f_0\mathbb{E}_{q(\lambda)}[\lambda] \\
&+ (a_0 - 1)\sum_{k=1}^{d}\mathbb{E}_{q(\alpha_k)}[\ln\,\alpha_k] - b_0\sum_{k=1}^{d}\mathbb{E}_{q(\alpha_k)}[\alpha_k] \\
&+ \frac{N}{2}\mathbb{E}_{q(\lambda)}[\ln\,\lambda] - \frac{1}{2}\mathbb{E}_{q(\lambda)}[\lambda]\sum_{i=1}^{N}\mathbb{E}_{q(w)}[(y_i - x_i^T w)^2] \\
&- (a_0 - \frac{1}{2})\sum_{k=1}^{d}\mathbb{E}_{q(\alpha_k)}[\ln\,\alpha_k] + b_0\sum_{k=1}^{d}\mathbb{E}_{q(\alpha_k)}[\alpha_k] \\
&+ \frac{1}{2}\sum_{k=1}^{d}\mathbb{E}_{q(\alpha_k)}\alpha_k\mathbb{E}_{q(w)}[w^2] - (e_0 - 1)\mathbb{E}_{q(\lambda)}[\ln\,\lambda] - \frac{N}{2}\mathbb{E}_{q(\lambda)}[\ln\,\lambda] \\
&+ f_0\mathbb{E}_{q(\lambda)}[\lambda] + \frac{1}{2}\mathbb{E}_{q(\lambda)}[\lambda]\sum_{i=1}^{N}\mathbb{E}_{q(w)}[(y_i - x_i^T w)^2] \\
&+ \frac{1}{2}\mathbb{E}_{q(w)}[\ln\,|\Sigma_w|] + \frac{1}{2}\mathbb{E}_{q(w)}[(w - \mu_w)^T\Sigma_w^{-1}(w - \mu_w)]
\end{aligned}$$

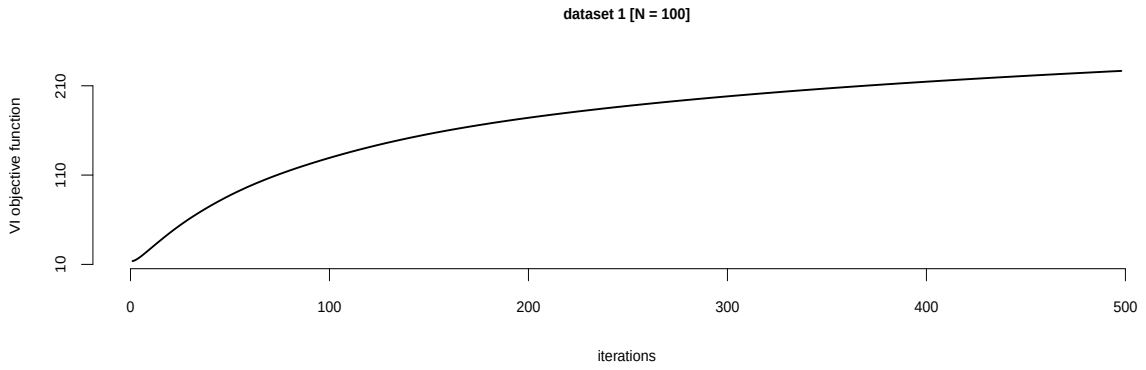
The first 4 lines after the proportionality sign are the expectation of the log joint likelihood. The next four lines are the entropy terms. I observe that all terms except the last line cancel out with the other terms. The terms from the entropy of $q(w)$ are the only ones that remain. The second term in the last line reduces to the trace of an identity matrix and is a constant. The final variational inference objective function is given below.

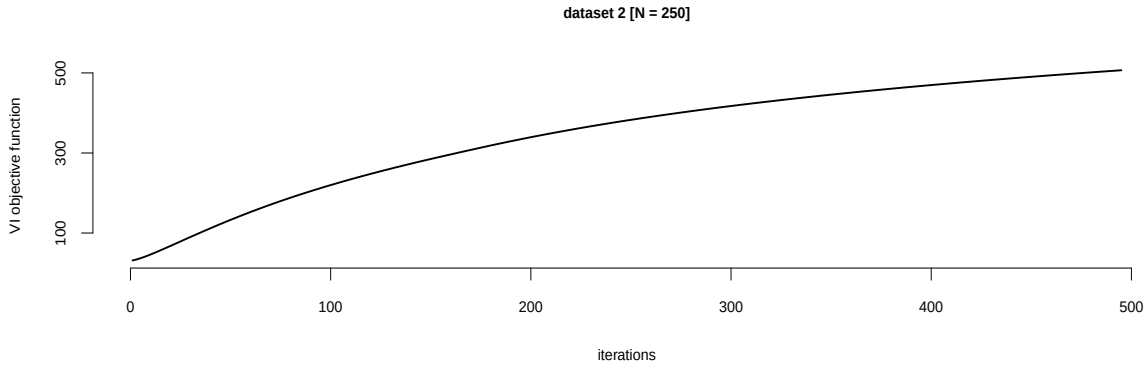
$$\mathcal{L}(\mu_w, \Sigma_w, e', f', a', b') = \frac{1}{2}\ln |\Sigma_w| + constant$$

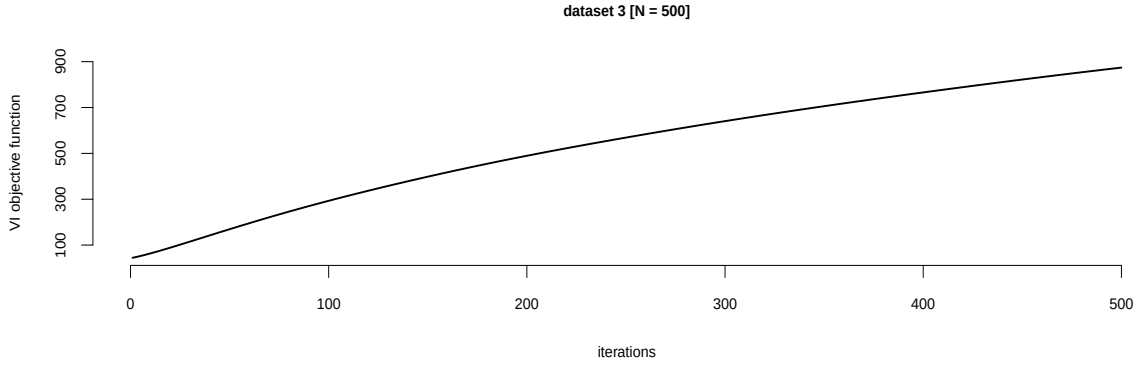
2 Answer 2

2.1 Part A

Here I have plotted the VI objective function for the 3 datasets. It is monotonically increasing.

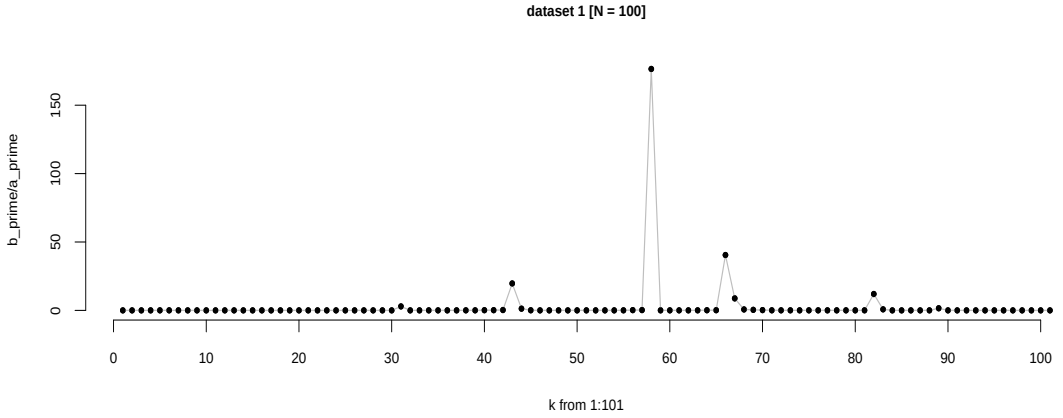


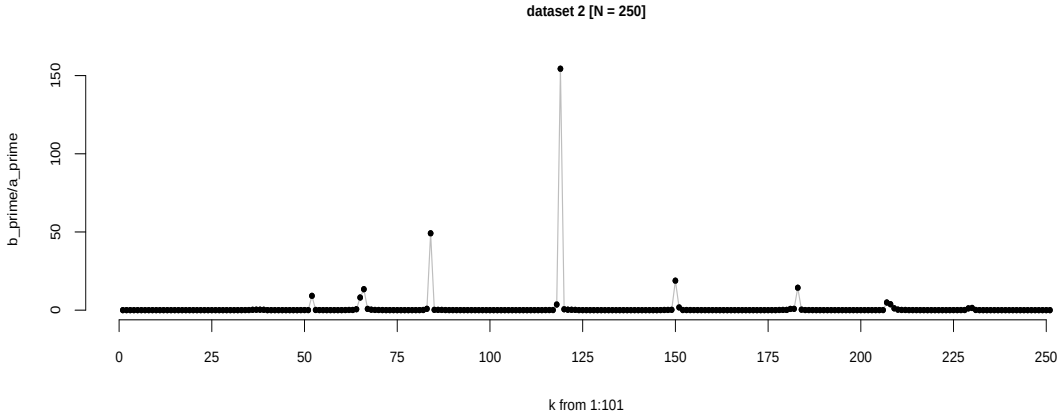


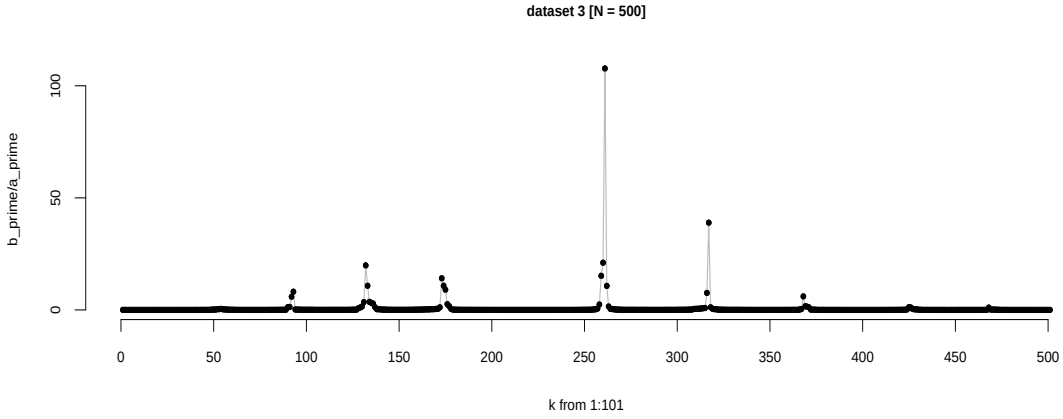


2.2 Part B

Here are the values of $1/\mathbb{E}_{q(\alpha_k)}[\alpha_k]$ from the last iteration for 3 datasets. This value is b'/a'.







2.3 Part C

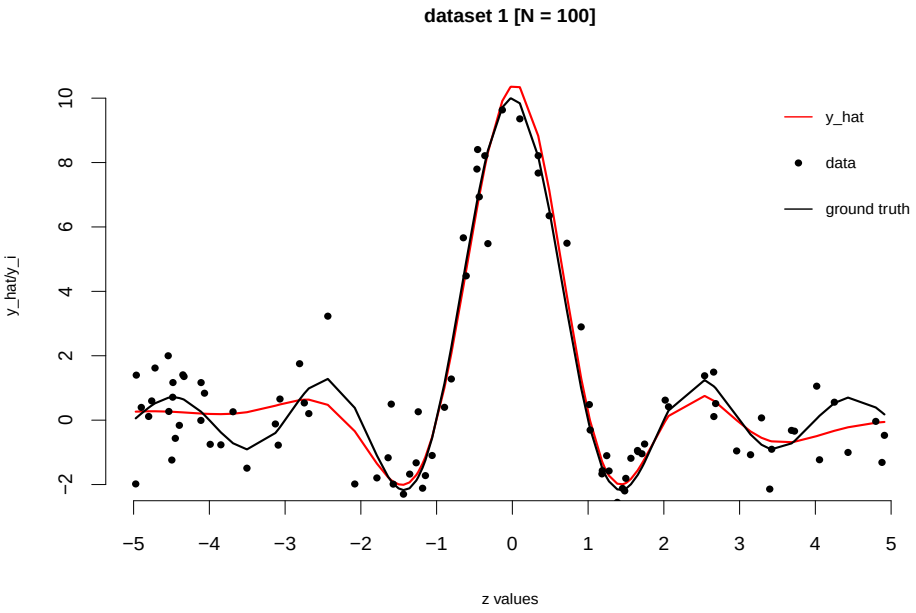
Here I present in tabular form the values of $1/\mathbb{E}_{q(\lambda)}[\lambda]$ which is computed as f'/e'

Table 1: $1/\mathbb{E}_{q(\lambda)}[\lambda]$

Dataset	N	value
1	100	1.0798
2	250	0.8994
3	500	0.9781

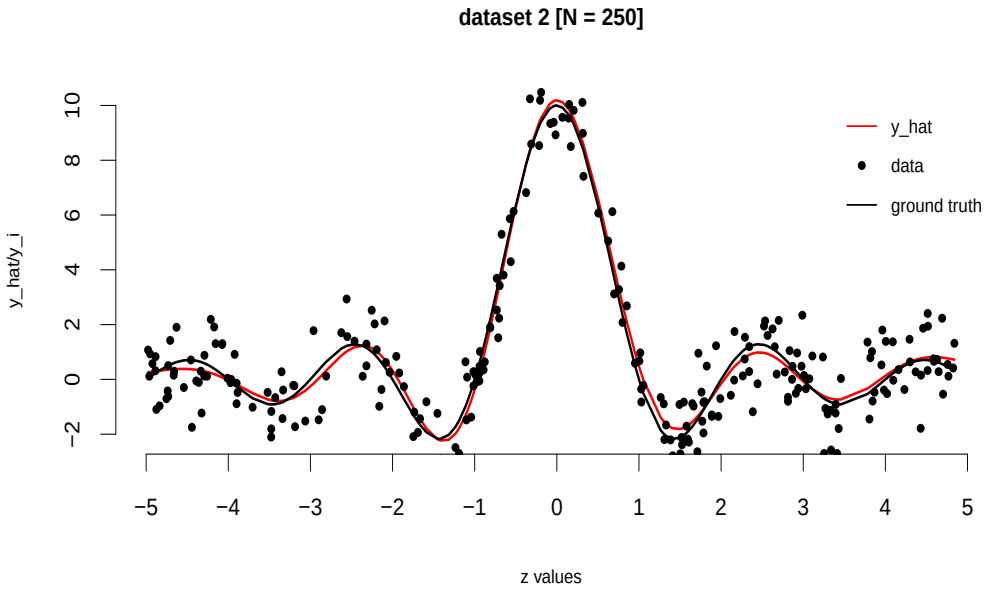
2.4 Part D

2.4.1 The first dataset N = 100



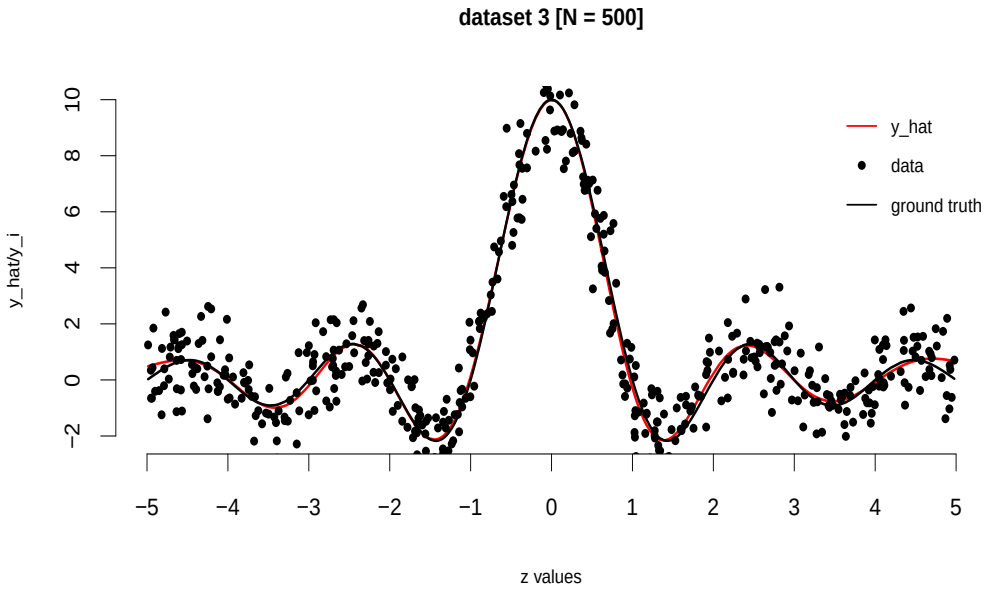
In this figure the solid red line is $\hat{y}_i = x_i^T \hat{w}$, which are the predicted values. The black dots are the scatter plot of y_i. The solid black line is the function $10 * sinc(z_i)$, this indicates the "ground truth".

2.4.2 The second dataset N = 250



In this figure the solid red line is $\hat{y}_i = x_i^T \hat{w}$*, which are the predicted values. The black dots are the scatter plot of* y_i*. The solid black line is the function* $10 * sinc(z_i)$*, this indicates the "ground truth".*

2.4.3 The third dataset N = 500



In this figure the solid red line is $\hat{y}_i = x_i^T \hat{w}$*, which are the predicted values. The black dots are the scatter plot of* y_i*. The solid black line is the function* $10 * sinc(z_i)$*, this indicates the "ground truth".*