Problem 1)

Francisco Arceo
Homework #3
p. 1

a) $\{(x_i, y_i)\}_{i=1}^{N}$; $y_i \in \mathbb{R}$; $x_i \in \mathbb{R}^d$ $d > N$

$y_i \overset{ind}{\sim} N(x_i^T w, \lambda^{-1})$ $w \sim N(0, \mathrm{diag}(\alpha_1, \ldots, \alpha_d)^{-1})$

$\alpha_K \sim \Gamma(a_0, b_0)$ $\lambda \sim \Gamma(e_0, f_0)$

Derive optimal $q(\cdot)$ of each distribution using

$$q(w, \alpha_1, \ldots, \alpha_d, \lambda) \approx p(w, \alpha_1, \ldots, \alpha_d, \lambda \mid y, x) = q(w) q(\lambda) \prod_{k=1}^{d} q(\alpha_k)$$

The optimal $q(\lambda)$

$$q(\lambda) \propto \exp\left[\mathbb{E}_q\left[\sum_{i=1}^{N} \ln p(y_i \mid w, \lambda, x)\right]\right] p(\lambda)$$

$$\propto \exp\left[\mathbb{E}_q\left[\sum_{i=1}^{N} -\frac{1}{2}\ln 2\pi + \frac{1}{2}\ln\lambda - \frac{\lambda}{2}(y_i - x_i^T w)^2\right]\right] \times \frac{f_0}{\Gamma(e_0)} \lambda^{e_0 - 1} e^{-f_0 \lambda}$$

$$\propto \left(\frac{\lambda}{2\pi}\right)^{N/2} \exp\left[-\frac{\lambda}{2}\mathbb{E}\left[(y_i^2 - 2 y_i x_i^T w + w^T x_i x_i^T w)\right]\right] \frac{f_0}{\Gamma(e_0)} \lambda^{e_0 - 1} e^{-f_0 \lambda}$$

$$\propto \lambda^{N/2 + e_0 - 1} \exp\left[-\frac{\lambda}{2}\left(\sum_{i=1}^{N} y_i^2 - 2 y_i x_i^T \mathbb{E}[w] + \mathbb{E}[\mathrm{tr}(x_i x_i^T w w^T)]\right) - f_0 \lambda\right]$$

Note that: $\mathbb{E}[w] = \mu$ $\mathbb{E}[w^T w] = \Sigma + \mu\mu^T$

$\mathrm{tr}(x_i x_i^T \mathbb{E}[w w^T]) = \mathrm{tr}(x_i x_i^T (\Sigma - \mu\mu^T))$

$$q(\lambda) \propto \exp\left[-\lambda\left(f_0 + \frac{1}{2}\left(\sum_{i=1}^{N} x_i^T \Sigma x_i + y_i^2 - 2 y_i x_i^T \mu + \mu^T x_i x_i^T \mu\right)\right)\right] \lambda^{e_0 + \frac{N}{2} - 1}$$

$$\propto \lambda^{e_0 + \frac{N}{2} - 1} \exp\left[-\lambda\left[f_0 + \frac{1}{2}\left(\sum_{i=1}^{N} (y_i - x_i^T \mu)^2 + x_i^T \Sigma x_i\right)\right]\right]$$

Which is Gamma. Thus, $q(\lambda) \sim \Gamma(e_1, f_1)$

where $e_1 = e_0 + N/2$

$$f_1 = f_0 + \frac{1}{2}\underbrace{\left(\sum_{i=1}^{N} (y_i - x_i^T \mu)^2 + x_i^T \Sigma x_i\right)}_{\mathbb{E}[y_i - x_i^T w]}$$

and

$$\mathbb{E}_{q(\lambda)}[\lambda] = \frac{e_1}{f_1}$$

The optimal $q(\alpha_k)$

$$q(\alpha_k) \propto \exp\Big[\mathbb{E}_{-q}\Big[\sum_{i=1}^{N} \ln p(y_i|w,X,\lambda) + \ln p(\lambda) + \ln p(w|\alpha_1,\dots,\alpha_k) + \sum_{k=1}^{d} \ln p(\alpha_1,\dots,\alpha_k)\Big]\Big]$$

can be ignored

$$\propto \alpha_k^{a_0-1} e^{-\alpha_k b_0} \times \exp\Big[\mathbb{E}_{-q}\Big[\tfrac{1}{2}\sum_{k=1}^{d} \ln\alpha_k - \tfrac{1}{2} w^T \mathrm{diag}(\alpha_1,\dots,\alpha_d) w\Big]\Big]$$

$$\propto \alpha_k^{a_0-1} e^{-\alpha_k b_0} \exp\Big[\mathbb{E}\Big[\tfrac{1}{2}\ln\alpha_k - \tfrac{1}{2} w^T \mathrm{diag}(\alpha_1,\dots,\alpha_d) w\Big]\Big]$$

$$\propto \alpha_k^{a_0+\frac{1}{2}-1} e^{-\alpha_k b_0} \exp\Big[-\tfrac{1}{2}\,\mathrm{trace}\big(\mathrm{diag}(\alpha_1,\dots,\alpha_d)\,\mathbb{E}[ww^T]\big)\Big]$$

$$\propto \alpha_k^{a_0-\frac{1}{2}} e^{-\alpha_k b_0} \exp\Big[-\tfrac{1}{2}\,\mathrm{trace}\big(\mathrm{diag}(\alpha_1,\dots,\alpha_d)(\Sigma + \mu\mu^T)\big)\Big]$$

$$\propto \alpha_k^{a_0-\frac{1}{2}} \exp\Big[-\alpha_k b_0 - \tfrac{1}{2}\Big(\alpha_k\big[\Sigma_{(k,k)} + \mu_k\mu_k^T\big]\Big)\Big]$$

Where $\Sigma_{(k,k)}$ denotes the k^{th} row of k^{th} column of Σ and the k^{th} element of μ.

This is obviously a Gamma and so, $q(\alpha_k) \sim \Gamma(a_1, b_1)$
where

$$a_1 = a_0 + \tfrac{1}{2}$$

$$b_1 = b_0 + \tfrac{1}{2}\big(\Sigma_{(k,k)} + \mu_k\mu_k^T\big)$$

and

$$\mathbb{E}_{q(\alpha_k)}[\alpha_k] = \frac{a_1}{b_1}$$

The optimal $q(w)$

$$q(w) \propto \exp\Big[\mathbb{E}_{-q}\Big[\sum_{i=1}^{N} \ln p(y_i|w,x_i,\lambda) + \ln p(\lambda) + \ln p(w|\alpha_1,\dots,\alpha_d) + \sum_{k=1}^{d} \ln p(\alpha_k)\Big]\Big]$$

Don't depend on w

$$q(w) \propto \exp\Big[\mathbb{E}_{-q}\Big[\sum_{i=1}^{N}\Big(\tfrac{1}{2}\ln\big(\tfrac{\lambda}{2\pi}\big) - \tfrac{\lambda}{2}(y_i - x_i^T w)^2\Big) - \tfrac{d}{2}\ln 2\pi + \tfrac{1}{2}\sum_{k=1}^{d}\ln\alpha_k - \tfrac{1}{2} w^T \mathrm{diag}(\alpha_1,\dots,\alpha_d) w\Big]\Big]$$

$$\propto \exp\Big[\mathbb{E}_{-q}\Big[-\tfrac{\lambda}{2}\sum_{i=1}^{N}(y_i - x_i^T w)^2 - \tfrac{1}{2} w^T \mathrm{diag}(\alpha_1,\dots,\alpha_d) w\Big]\Big]$$

$$\propto \exp\Big[-\tfrac{1}{2}\mathbb{E}\Big[\lambda\sum_{i=1}^{N}(y_i^2 - 2y_i x_i^T w + w^T x_i x_i^T w) - w^T \mathrm{diag}(\alpha_1,\dots,\alpha_d) w\Big]\Big]$$

$$\propto \exp\Big[-\tfrac{1}{2}\mathbb{E}\Big[\lambda\sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\alpha_1,\dots,\alpha_d) ww^T - 2\lambda\sum_{i=1}^{N} y_i x_i^T w\Big]\Big]$$

$$\propto \exp\Big[-\tfrac{1}{2}\mathbb{E}\Big[\Big(w - \big(\lambda \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\alpha_1,\dots,\alpha_d)\big)\big(\lambda \sum_{i=1}^{N} y_i x_i\big)\Big)^T$$
$$\times\Big(\lambda + \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\alpha_1,\dots,\alpha_d)\Big) \times \Big(w - \big(\lambda \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\alpha_1,\dots,\alpha_d)\big)^{-1}\big(\lambda \sum_{i=1}^{N} y_i x_i\big)\Big)\Big]\Big]$$

$$\propto \exp\Big[-\tfrac{1}{2}\Big(\mathbb{E}[w] - \big[\mathbb{E}[\lambda] \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\mathbb{E}[\alpha_1],\dots,\mathbb{E}[\alpha_d])\big]^{-1}\big(\mathbb{E}[\lambda] \sum y_i x_i\big)\Big)^T$$
$$\times\Big(\mathbb{E}[\lambda] \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\mathbb{E}[\alpha_1],\dots,\mathbb{E}[\alpha_d])\Big)$$
$$\times\Big(\mathbb{E}[w] - \big(\mathbb{E}[\lambda] \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\mathbb{E}[\alpha_1],\dots,\mathbb{E}[\alpha_d])\big)^{-1}\big(\mathbb{E}[\lambda] \sum_{i=1}^{N} y_i x_i\big)\Big)\Big]$$

which is Normal. Thus,

$q(w) \sim N(\mu_1, \Sigma_1)$ where $\mu_1 = \Sigma_1\big(\mathbb{E}[\lambda] \sum_{i=1}^{N} y_i x_i\big)$

$$\Sigma_1 = \Big[\mathbb{E}[\lambda] \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}(\mathbb{E}[\alpha_1],\dots,\mathbb{E}[\alpha_d])\Big]^{-1}$$

D) <u>VI for Bayesian Linear Regression</u>

<u>Input</u>: $\{(x_i, y_i)\}_{i=1}^N$ $y \in \mathbb{R}$; $x \in \mathbb{R}^d$

$q(\lambda) \sim \Gamma(\lambda | e_1, f_1)$; $q(w) \sim N(w|\mu_1, \Sigma_1)$; $q(\alpha) \sim \Gamma(a_1, b_1)$

<u>Output</u>: Values for $a_1, \dots, a_d, b_1, \dots, b_d, e_1, f_1, \mu_1, \Sigma_1$

1.) Initialize parameters $a_{10},\dots,a_{d0}, b_{10},\dots,b_{d0}, e_0, f_0, \mu_0, \Sigma_0$ in some way.

2.) for $t = 1,\dots,T$

- Update $q(\lambda)$ with $e_{1t} = e_0 + \frac{N}{2}$ & $f_{1t} = f_0 + \frac{1}{2}\Big[\sum_{i=1}^{N}(y_i - x_i^T\mu_{t-1})^2 + x_i^T \Sigma_{t-1} x_i\Big]$
- Update $q(\alpha_k)$ with $a_{1k_t} = a_{0k_t} + \frac{1}{2}$ & $b_{1k_t} = b_{0k_t} + \frac{1}{2}\Big[\Sigma_{t-1}(k,k) + \mu_{k_{t-1}}\mu_{k_{t-1}}^T\Big]$
- Update $q(w)$ with $\mu_1 = \Sigma_t\Big(\frac{e_{1t}}{f_{1t}} \sum_{i=1}^{N} y_i x_i\Big)$ & $\Sigma_t = \Big(\frac{e_{1t}}{f_{1t}} \sum_{i=1}^{N} x_i x_i^T + \mathrm{diag}\big(\frac{a_{1t}}{b_{1t}},\dots,\frac{a_{dt}}{b_{dt}}\big)\Big)^{-1}$

then evaluate $\mathcal{L}(\vec{a}_{1t}, \vec{b}_{1t}, e_1, f_1, \mu_1, \Sigma_1)$ to assess convergence ($< \varepsilon$).

c) Calculate the Variational objective function

$$\mathcal{L}(\vec{a}_1, \vec{b}_1, e_1, f_1, \mu_1, \Sigma_1) = \int q(\lambda) \ln p(\lambda)\, d\lambda$$
$$+ \int \cdots \int q(w) \prod_{k=1}^{d} q(\alpha_k) \ln p(w|\alpha_1, \dots, \alpha_d)\, dw\, d\alpha_1, \dots, d\alpha_d$$
$$+ \sum_{k=1}^{d} \int q(\alpha_k) \ln p(\alpha_k)\, d\alpha_k$$
$$+ \sum_{i=1}^{N} \int \cdots \int \prod_{k=1}^{d} q(\alpha_k) q(\lambda) q(w) \ln p(y_i | x_i, w, \lambda)\, dw\, d\lambda\, d\alpha_1, \dots, d\alpha_d$$
$$- \int q(\lambda) \ln q(\lambda)\, d\lambda - \int q(w) \ln q(w)\, dw - \left(\sum_{k=1}^{d} \int q(\alpha_k) \ln q(\alpha_k)\, d\alpha_k \right)$$

$$\mathcal{L}(\vec{a}_1, \vec{b}_1, e_1, f_1, \mu_1, \Sigma_1) = \mathbb{E}[\ln p(\lambda)] + \mathbb{E}[\ln p(w|\alpha_1, \dots, \alpha_d)] + \sum_{k=1}^{d} \mathbb{E}[\ln p(\alpha_k)]$$
$$+ \sum_{i=1}^{N} \mathbb{E}[\ln p(y_i | x_i, w, \lambda)] - \mathbb{E}[\ln q(\lambda)] - \mathbb{E}[\ln q(w)]$$
$$- \left(\sum_{k=1}^{d} \mathbb{E}[\ln q(\alpha_k)] \right)$$

Since these expectations are with respect to the $q(\cdot)$ parameters, we can see that 3 terms are just Entropies with constant terms.

Mainly

$$\mathbb{E}_{q(w)}[\ln q(w)] = \frac{N}{2} + \frac{N}{2} \ln 2\pi + \frac{1}{2} \ln |\Sigma_1| = \text{constants} + \frac{1}{2} \ln |\Sigma_1|$$

$$\mathbb{E}_{q(\lambda)}[\ln q(\lambda)] = e_1 - \ln f_1 + \ln \Gamma(e_1) + (1 - e_1)\psi(e_1)$$

$$\mathbb{E}_{q(\alpha_k)}[\ln q(\alpha_k)] = a_{1k} - \ln b_{1k} + \ln \Gamma(a_{1k}) + (1 - a_{1k})\psi(a_{1k})$$

where $\psi(x)$ is the digamma function $\left(\frac{\partial \ln \Gamma(x)}{\partial x}\right)$ and the definitions of entropy have been used for the Multivariate Normal and the Gamma distributions.

Francesco Arezzo

p. 5

$$\mathbb{E}[\ln p(w\mid\alpha_1,\dots,\alpha_d)] = \mathbb{E}\Big[-\frac{d}{2}\ln 2\pi + \frac{1}{2}\sum_{k=1}^{d}\ln\alpha_i - \frac{1}{2}w^T \operatorname{diag}(\alpha_1,\dots,\alpha_d)w\Big]$$

$$= -\frac{d}{2}\ln 2\pi + \frac{1}{2}\sum_{k=1}^{d}\mathbb{E}[\ln\alpha_k] - \frac{1}{2}\operatorname{trace}\big(\operatorname{diag}(\mathbb{E}[\alpha_1],\dots,\mathbb{E}[\alpha_d])\,\mathbb{E}[ww^T]\big)$$

$$= \text{constants} + \frac{1}{2}\sum_{k=1}^{d}[\psi(a_{1k}) - \ln b_{1k}] - \frac{1}{2}\operatorname{trace}\Big(\operatorname{diag}\Big(\frac{a_{11}}{b_{11}},\dots,\frac{a_{1d}}{b_{1d}}\Big)(\Sigma_1 + \mu_1\mu_1^T)\Big)$$

$$= \text{constants} + \frac{1}{2}\sum_{k=1}^{d}(\psi(a_{1k}) - \ln b_{1k}) - \frac{1}{2}\Big[\operatorname{trace}\Big(\operatorname{diag}\Big(\frac{a_{11}}{b_{11}},\dots,\frac{a_{1d}}{b_{1d}}\Big)\Sigma_1\Big) + \mu_1^T \operatorname{diag}\Big(\frac{a_{11}}{b_{11}},\dots,\frac{a_{1d}}{b_{1d}}\Big)\mu_1\Big]$$

$$\mathbb{E}_{q(\lambda)}[\ln p(\lambda)] = e_0\ln f_0 - \ln\Gamma(e_0) + (e_0-1)\mathbb{E}_{q(\lambda)}[\ln\lambda] - f_0\,\mathbb{E}_{q(\lambda)}[\lambda]$$

$$= \text{constants} + (e_0-1)[\psi(e_1) - \ln f_1] - f_0\frac{e_1}{f_1}$$

where we've used $\mathbb{E}[\ln\lambda]$ = Entropy of a Gamma distribution.

$$\mathbb{E}[\ln p(\alpha_k)] = a_{0k}\ln b_{0k} - \ln\Gamma(a_{0k}) + (a_{0k}-1)\mathbb{E}[\ln\alpha_k] - b_{0k}\mathbb{E}[\alpha_k]$$

$$= \text{constants} + (a_{0k}-1)[\psi(a_{1k}) - \ln b_{1k}] - b_{0k}\frac{a_{1k}}{b_{1k}}$$

$$\mathbb{E}[\ln p(y_i\mid x_i,w,\lambda)] = -\frac{1}{2}\ln 2\pi + \frac{\mathbb{E}[\ln\lambda]}{2} - \frac{\mathbb{E}[\lambda]}{2}\mathbb{E}\Big[\sum_{i=1}^{N}(y_i - x_i^T w)^2\Big]$$

$$= \text{constants} + \frac{1}{2}(\psi(e_1) - \ln f_1) - \frac{1}{2}\times\frac{e_1}{f_1}\Big(\sum_{i=1}^{N}(y_i - x_i^T\mu_1)^2 + x_i^T\Sigma_1 x_i\Big)$$

Combining all of the terms yields:

$$
\begin{aligned}
\mathcal{L}(a_1, b_1, e_1, f_1, \mu_1, \Sigma_1) = {} & \mathbb{E}[\ln p(\lambda)] \\
& + \mathbb{E}[\ln p(\omega \mid \alpha_1, \ldots, \alpha_d)] \\
& + \sum_{k=1}^{d} \mathbb{E}[\ln p(\alpha_k)] \\
& + \sum_{i=1}^{N} \mathbb{E}[\ln p(y_i \mid x_i, \omega, \lambda)] \\
& - \mathbb{E}[\ln q(\lambda)] \\
& - \mathbb{E}[\ln q(\omega)] \\
& - \sum_{k=1}^{d} \mathbb{E}[\ln q(\alpha_k)]
\end{aligned}
$$

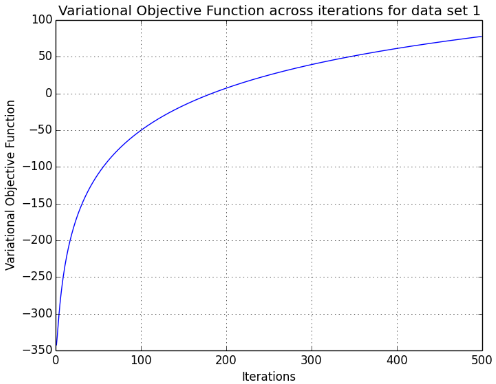
$$
\begin{aligned}
\mathcal{L}(a_1, b_1, e_1, f_1, \mu_1, \Sigma_1) = {} & \text{constants} + (e_0 - 1)[\psi(e_1) - \ln f_1] - f_0 \times \frac{e_1}{f_1} \\
& + \text{constants} + \frac{1}{2}\sum_{k=1}^{d}(\psi(a_{1k}) - \ln b_{1k}) - \frac{1}{2}\Big[\text{trace}\Big(\text{diag}\Big(\frac{a_{11}}{b_{11}}, \ldots, \frac{a_{1d}}{b_{1d}}\Big)\Sigma_1\Big) + \mu_1^T \text{diag}\Big(\frac{a_{11}}{b_{11}}, \ldots, \frac{a_{1d}}{b_{1d}}\Big)\mu\Big] \\
& + \text{constants} + \sum_{k=1}^{d}\Big[(a_{0k} - 1)[\psi(a_{1k}) - \ln b_{1k}] - b_{0k}\frac{a_{1k}}{b_{1k}}\Big] \\
& + \Big[\text{constants} + \frac{1}{2}(\psi(e_1) - \ln f_1) - \frac{1}{2} \times \frac{e_1}{f_1}\Big(\sum_{i=1}^{N}(y_i - x_i^T\mu_1)^2 + x_i^T \Sigma_1 x_i\Big)\Big] \\
& - [e_1 - \ln f_1 + \ln \Gamma(e_1) + (1 - e_1)\psi(e_1)] \\
& - \Big[\text{constants} + \frac{1}{2}\ln|\Sigma_1| \\
& - \sum_{k=1}^{d}[a_{1k} - \ln b_{1k} + \ln \Gamma(a_{1k}) + (1 - a_{1k})\psi(a_{1k})]
\end{aligned}
$$

Note, in the actual calculation of the objective function, all of the constants can be ignored.

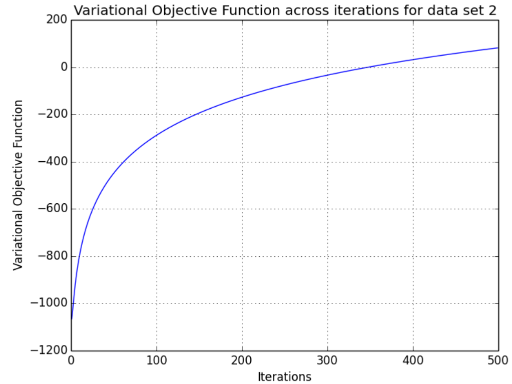
Problem 2

a)

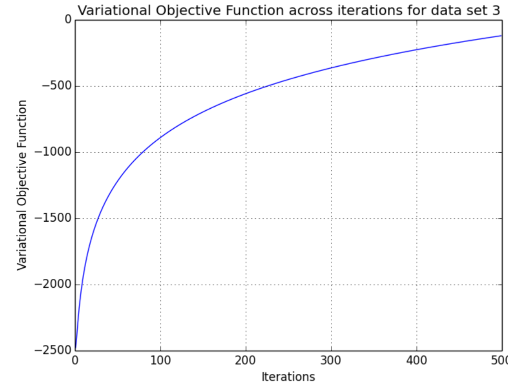
Data Set 1



Data Set 2

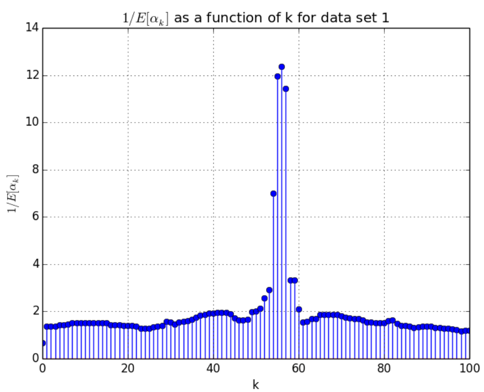


Data Set 3

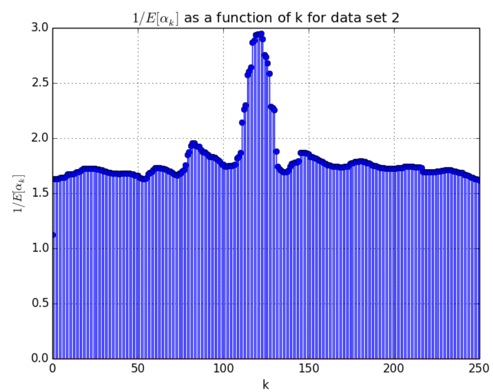


b)

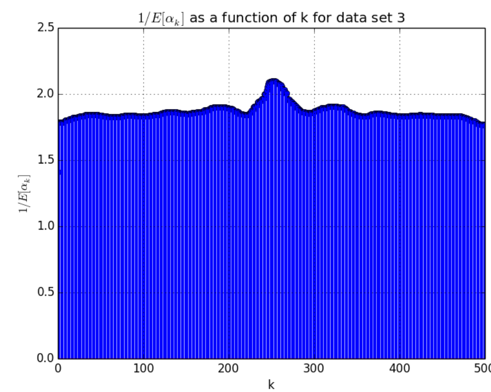
Data Set 1



Data Set 2



Data Set 3



c)

Data Set 1

$1/\mathbb{E}_q[\lambda]$ = 3.67329015015

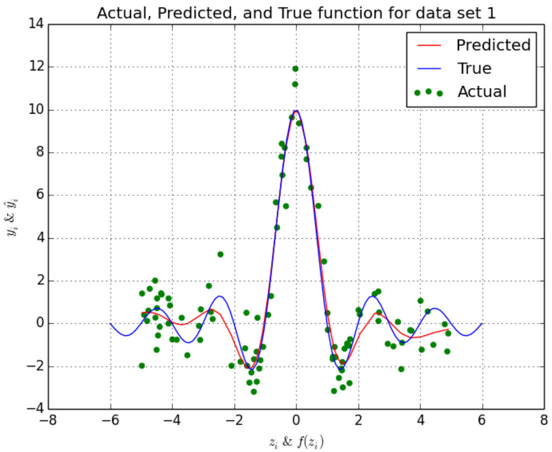
Data Set 2

$1/\mathbb{E}_q[\lambda]$ = 10.2601623179

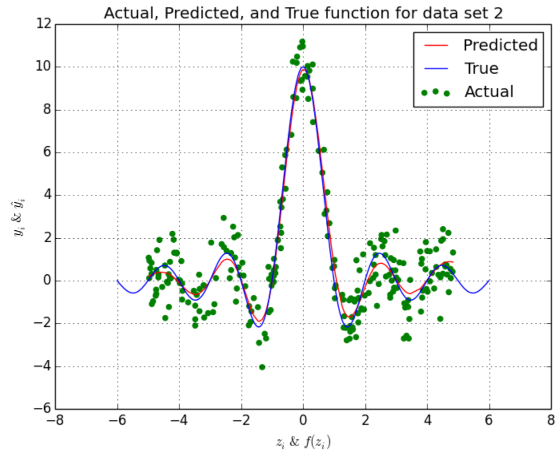
Data Set 3

$1/\mathbb{E}_q[\lambda]$ = 33.4137810438 d)

Data Set 1



Data Set 2



Data Set 3