Zoltan Onodi-Szucs
zo2131

Homework 3
Bayesian Machine Learning

Problem 1. (50 points)

We have a data set of the form $\{(x_i, y_i)\}_{i=1}^N$, where $y \in \mathbb{R}$ and $x \in \mathbb{R}^d$. We assume d is large and not all dimensions of x are informative in predicting y. Consider the following regression model for this problem:

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \lambda^{-1}), \quad w \sim \text{Normal}(0, \text{diag}(\alpha_1, \dots, \alpha_d)^{-1}),$$

$$\alpha_k \overset{iid}{\sim} \text{Gamma}(a_0, b_0), \quad \lambda \sim \text{Gamma}(e_0, f_0).$$

Use the density function $\text{Gamma}(\eta|\tau_1, \tau_2) = \frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)} \eta^{\tau_1 - 1} \mathrm{e}^{-\tau_2 \eta}$. In this homework, you will derive a variational inference algorithm for approximating the posterior distribution with

$$q(w, \alpha_1, \dots, \alpha_d, \lambda) \approx p(w, \alpha_1, \dots, \alpha_d, \lambda | y, x)$$

a) Using the factorization $q(w, \alpha_1, \dots, \alpha_d, \lambda) = q(w)q(\lambda)\prod_{k=1}^d q(\alpha_k)$, derive the optimal form of each q distribution. Use these optimal q distributions to derive a variational inference algorithm for approximating the posterior.

b) Summarize the algorithm derived in Part (a) using pseudo-code in a way similar to how algorithms are presented in the notes for the class.

c) Using these q distributions, calculate the variational objective function. You will need to evaluate this function in the next problem to show the convergence of your algorithm.

ZOLTAN ONODI-SZUCS
ZO2131

Data: $\{(x_i, y_i)\}_{i=1}^N$, $y \in \mathbb{R}$, $x \in \mathbb{R}^d$

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

$\alpha_k \sim \mathrm{Gamma}(a_0, b_0)$, $\lambda \sim \mathrm{Gamma}(e_0, f_0)$

a) $q(w, \lambda, \alpha_1 \ldots \alpha_d) = q(w)\, q(\lambda) \prod_{k=1}^{d} q(\alpha_k)$

$$\mathcal{L}(w, \lambda, \alpha_1 \ldots \alpha_d) = \int \ldots \int q(w, \lambda, \alpha_1 \ldots \alpha_d) \cdot \ln \frac{P(y, w, \lambda, \alpha_1 \ldots \alpha_d | X)}{q(w, \lambda, \alpha_1 \ldots \alpha_d)} \, dw\, d\lambda\, d\alpha_1 \ldots d\alpha_k$$

$$P(y, w, \lambda, \alpha_1 \ldots \alpha_k) = P(y | \alpha, w, \lambda, X) \cdot P(\alpha, w, \lambda | X)$$
$$= P(y | w, \lambda, x) \cdot P(w | \alpha, \lambda, x)\, P(\alpha, \lambda | x)$$
$$= P(y | w, \lambda, x) \cdot P(w | \alpha) \cdot P(\alpha)\, P(\lambda)$$

$P(y | w, \lambda, x) = \prod_{i=1}^{N} P(y_i | w, \lambda x_i)$

$P(\alpha) = \prod_{i=1}^{d} P(\alpha_k)$

$$q(\lambda) \propto \exp\{E_{q(\Theta \neq \lambda)}[\ln P(y, \lambda, w, \alpha \mid x)]$$

$$\propto \exp\{E_{q(\Theta\neq\lambda)}[\ln P(Y \mid w, \lambda, X) + \ln P(\lambda)]$$

$$\propto \exp\{E_{q(w)}[\ln P(Y \mid w, \lambda, X)] + \ln P(\lambda)]$$

$$\propto \prod_{i=1}^{N} \lambda^{1/2} \exp\{-\frac{\lambda}{2} E_{q(w)}[(y_i - x_i^T w)^2]\} \cdot \lambda^{e_0-1} \cdot e^{-f_0\lambda}$$

$$\propto \lambda^{\frac{N}{2}+e_0-1} \cdot e^{-\lambda(f_0 + \frac{1}{2}\sum_{i=1}^{N} E_{q(w)}[(y_i - x_i^T w)^2])}$$

$$\sim \text{gamma}(e', f') \qquad e' = \frac{N}{2} + e_0$$

$$f' = f_0 + \frac{1}{2}\sum_{i=1}^{N} E_{q(w)}[(y_i - x_i^T w)^2]$$

$$q(w) \propto \exp\{E_{q(\Theta\neq w)}[\ln P(Y, w, \lambda, \alpha \mid x)]\}$$

$$\propto \exp\{E_{q(\Theta\neq w)}[\ln P(Y \mid w, \lambda, X) + \ln P(w \mid \alpha)]\}$$

$$\propto \exp\{E_{q(\lambda)}[\ln P(Y \mid w, \lambda, X] + E_{q(\alpha)}[\ln P(w \mid \alpha)]\}$$

$$\propto \prod_{i=1}^{N} \exp\{-\frac{1}{2}E_{q(\lambda)}[\lambda](y_i - x_i^T w)^2\} \cdot \cancel{\exp\{-\frac{1}{2}E_{q(\alpha)}[\alpha] w^T w\}}$$

$$\cdot \exp\{-\frac{1}{2}E_{q(\alpha)}[w^T \cdot [\text{diag}(\alpha_k)] w)\}$$

$$\sim \mathcal{N}(\mu', \Sigma') \qquad \Sigma' = (E_{q(\lambda)}[\lambda]\sum_{i=1}^{N} x_i x_i^T + E_{q(\alpha)}[\text{diag}(\alpha_k)])^{-1}$$

$$\mu' = \Sigma'(E_{q(\lambda)}[\lambda]\sum_{i=1}^{N} y_i x_i)$$

$$q(\lambda_k) \propto E_{q(\Theta \neq \lambda_k)}[\ln P(Y, w, \lambda, L \mid X)]$$

$$\propto \exp\{E_{q(\Theta \neq \lambda_k)}[\ln P(w \mid L) + \ln P(\lambda_k)]$$

$$\propto \exp\{E_{q(\Theta \neq \lambda_k)}[\sum_{\ell \neq k} \ln \lambda_\ell^{1/2} \cdot e^{-\frac{1}{2} w_\ell^2 \lambda_\ell}] + E_{q(\lambda_k)}[\ln \lambda_k^{1/2} e^{-\frac{1}{2} w_k^2 \lambda_k}] +$$

$$+ \ln \lambda_k^{a_0 - 1} e^{-b_0 \lambda_k}$$

$$\propto \lambda_k^{a_0 + \frac{1}{2} - 1} e^{-\lambda_k (b_0 + \frac{1}{2} E_{q(w)}[w_k^2])}$$

$$\sim \text{Gamma}(a_k', b_k') \qquad a_k' = a_0 + \frac{1}{2}$$

$$b_k' = b_0 + \frac{1}{2} E_{q(w)}[w_k^2]$$

Expectations:

$$E_{q(\lambda)}[\lambda] = \frac{e'}{f'}$$

$$E_{q(w)}[(y_i - x_i^T w)^2] = (y_i - x_i^T \mu')^2 + x_i^T \Sigma' x_i$$

$$E_{q(w)}[w_k^2] = \Sigma'_{k,k} + \mu_k'^2$$

$$E_{q(\lambda)}[\text{diag}(\lambda_k)] = \text{diag}\left(\frac{a_k'}{b_k'}\right)$$

Algorithm:

1. Initialize $a'_0, b'_0, e'_0, f'_0, \mu'_0, \Sigma'_0$ in some way

2. For iterations $t \ldots T$:

a. Update $q(w) = \mathcal{N}(\mu'_{(t)}, \Sigma'_{(t)})$ by setting

$$\Sigma'_{(t)} = \left(\frac{e'_{(t-1)}}{f'_{(t-1)}} \sum_{i=1}^{N} x_i x_i^T + diag\left(\frac{a'_{(t-1)}}{b'_{k(t-1)}} \right) \right)^{-1}, \quad \mu'_t = \Sigma'_{(t)} \left(\frac{e'_{(t-1)}}{f'_{(t-1)}} \sum_{i=1}^{N} y_i x_i \right)$$

b. Update $q(\lambda) = \text{Gamma}(e', f')$ by setting

$$e'_t = \frac{N}{2} + e_0, \quad f'_t = f_0 + \frac{1}{2} \sum_{i=1}^{N} \left((y_i - x_i^T \mu'_{(t)})^2 + x_i^T \Sigma'_{(t)} x_i \right)$$

c. Update $q(\alpha_k) = \text{Gamma}(a', b'_k)$ by setting

$$a'_t = a_0 + \frac{1}{2} \quad b'_{k(t)} = b_0 + \frac{1}{2} \left(\Sigma'_{kk(t)} + \mu'^{2}_{k(t)} \right)$$

d. Evaluate the variational objective function $\mathcal{L}$ to assess convergence

$$\mathcal{L} = \int \dots \int q(w, \lambda, \lambda_1 \dots \lambda_d) \ln \frac{p(y, w, \lambda, \lambda_1 \dots \lambda_d | x)}{q(w, \lambda, \lambda_1 \dots \lambda_d)} \, dw \, d\lambda \, d\lambda_1 \dots d\lambda_d$$

$$= E_q[\ln p(y, w, \lambda, \lambda_1 \dots \lambda_d | x)] - E_q[\ln q(w, \lambda, \lambda_1 \dots \lambda_d)]$$

$$= E_q[\ln p(y | w, \lambda, x) + \ln p(w | \lambda) + \ln p(\lambda) + \ln p(\lambda)] -$$

$$- E_q[\ln q(w)] - E_q[\ln q(\lambda)] - E_q[\ln q(\lambda)]$$

$$= \frac{N}{2} E_{q(\lambda)}[\ln(\lambda)] - \frac{E_q[\lambda]}{2} \sum_{i=1}^{N} E_{q(w)}[(y_i - x_i^T w)^2] +$$

$$+ \frac{1}{2} \sum_{k=1}^{d} E_{q(\lambda_k)}[\ln(\lambda_k)] - \frac{1}{2} E_{q(w)}[w^T [\mathrm{diag}(E_{q(\lambda_k)}[\lambda_k])] w] +$$

$$+ (a_0 - 1) \sum_{k=1}^{d} E_{q(\lambda_k)}[\ln \lambda_k] - b_0 \sum_{k=1}^{d} E_{q(\lambda_k)}[\lambda_k] +$$

$$+ (e_0 - 1) E_{q(\lambda)}[\ln(\lambda)] - f_0 E_{q(\lambda)}[\lambda] +$$

$$+ \frac{1}{2} \ln(|\Sigma'|)$$

$$+ \left(e' - \ln f' + \ln(\Gamma(e')) + (1 - e') \Psi(e')\right) +$$

$$+ \sum_{k=1}^{d} \left[a'_k - \ln b'_k + \ln(\Gamma(a'_k)) + (1 - a'_k) \Psi(a'_k)\right]$$

$$+ \text{const.}$$

$$
\begin{aligned}
\mathcal{L} = {} & \frac{N}{2}\left(\Psi(e') - \ln(f')\right) - \frac{e'}{2f'}\sum_{i=1}^{N}\left(y_i - x_i^T\mu'\right)^2 + x_i^T\Sigma' x_i \\
& + \frac{1}{2}\sum_{k=1}^{d}\left(\Psi(a') - \ln(b_k')\right) - \frac{1}{2}\sum_{k=1}^{d}\left(\Sigma'_{kk} + \mu_{kk}'^{2}\right)\frac{a'}{b_k'} \\
& + (a_0 - 1)\sum_{k=1}^{d}\left(\Psi(a') - \ln(b_k')\right) - b_0\sum_{k=1}^{d}\frac{a'}{b_k'} \\
& + (e_0 - 1)\left(\Psi(e') - \ln(f')\right) - f_0\frac{e'}{f'} \\
& + \frac{1}{2}\ln(|\Sigma'|) \\
& + \left(e' - \ln f' + \ln\Gamma(e') + (1 - e')\Psi(e')\right) \\
& + \sum_{k=1}^{d}\left[a_k' - \ln b_k' + \ln\Gamma(a') + (1 - a')\Psi(a_k')\right]
\end{aligned}
$$

Problem 2. (50 points)

Implement the algorithm derived in Problem 1 and run it on the three data sets provided. Set the prior parameters $a_0 = b_0 = 10^{-16}$ and $e_0 = f_0 = 1$. We will not discuss sparsity-promoting "ARD" priors in detail in this course, but setting a_0 and b_0 in this way will encourage only a few dimensions of w to be significantly non-zero since many α_k should be extremely large according to $q(\alpha_k)$.

For each of the three data sets provided, show the following:

a) Run your algorithm for 500 iterations and plot the variational objective function.

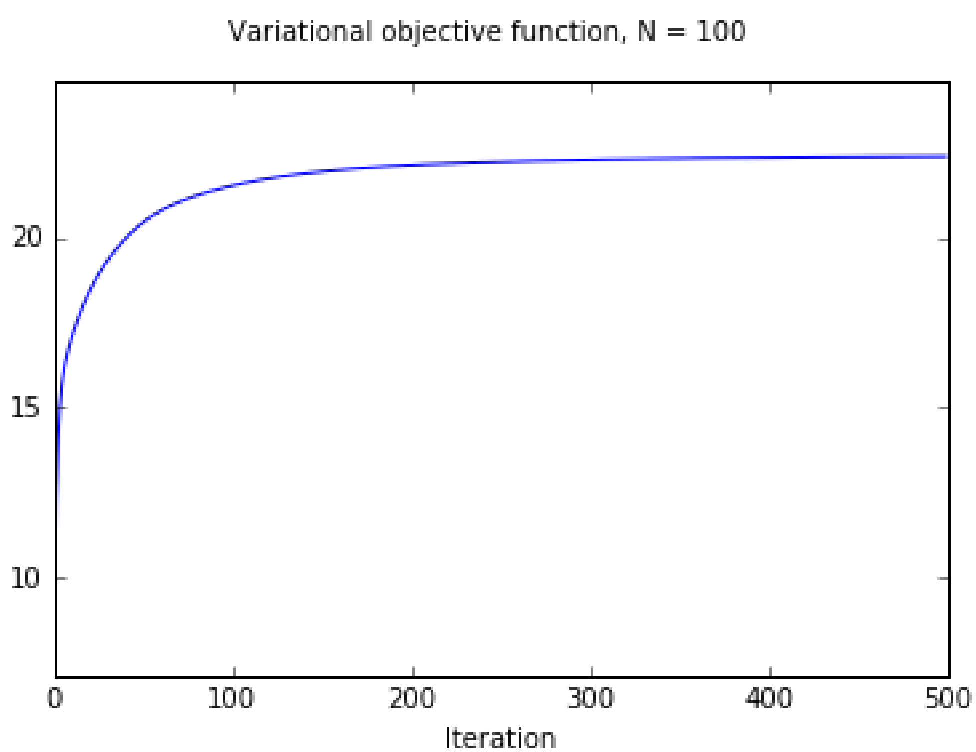
b) Using the final iteration, plot $1/\mathbb{E}_q[\alpha_k]$ as a function of k.

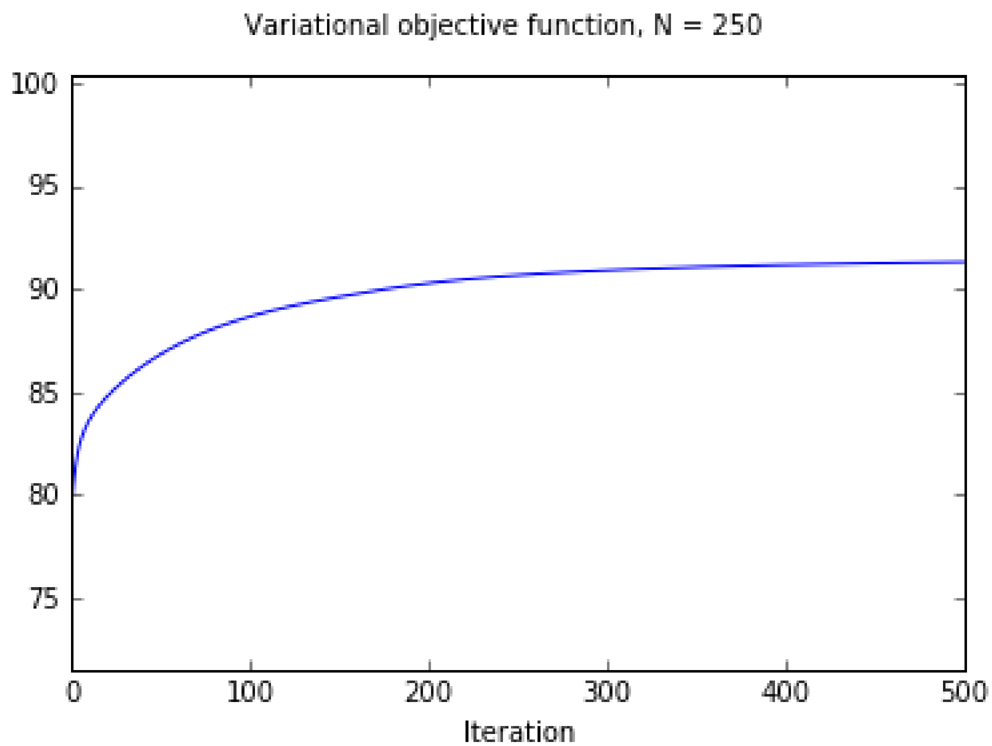
c) Give the value of $1/\mathbb{E}_q[\lambda]$ for the final iteration.

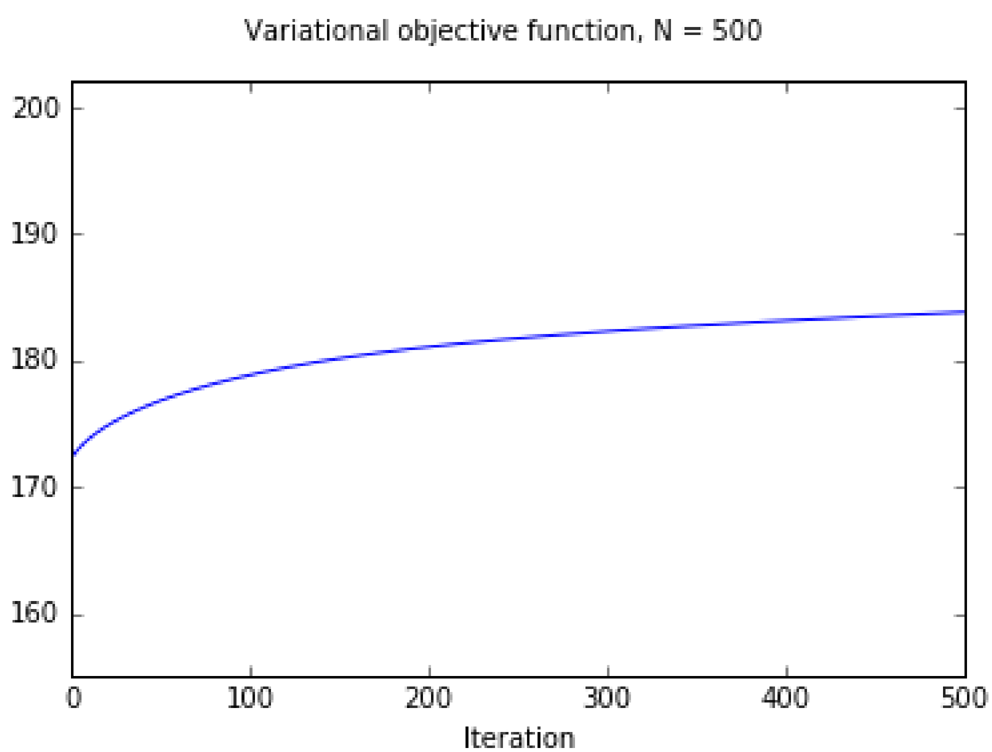
d) Using $\hat{w} = \mathbb{E}_{q(w)}[w]$, calculate $\hat{y}_i = x_i^T \hat{w}$ for each data point. Using the z_i associated with y_i (see below), plot $\hat{y}_i$ vs z_i as a solid line. On the same plot show (z_i, y_i) as a scatter plot. Also show the function $(z_i, 10 * \text{sinc}(z_i))$ as a solid line in a different color.

   Hint about Part (d): z is the horizonal axis and y the vertical axis. Both solid lines should look like a function that smoothly passes through the data. The second line is ground truth.

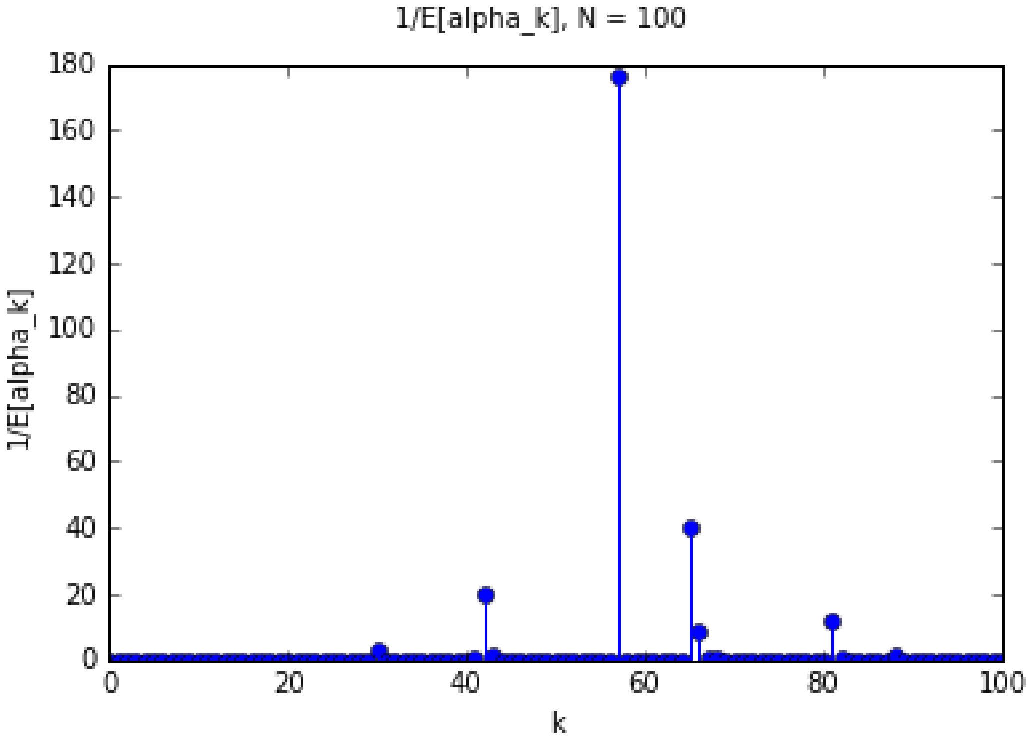
a, Note that the variational objective functions are positive, because I drop the constant terms. If I include them, it's the same shape but a negative number.

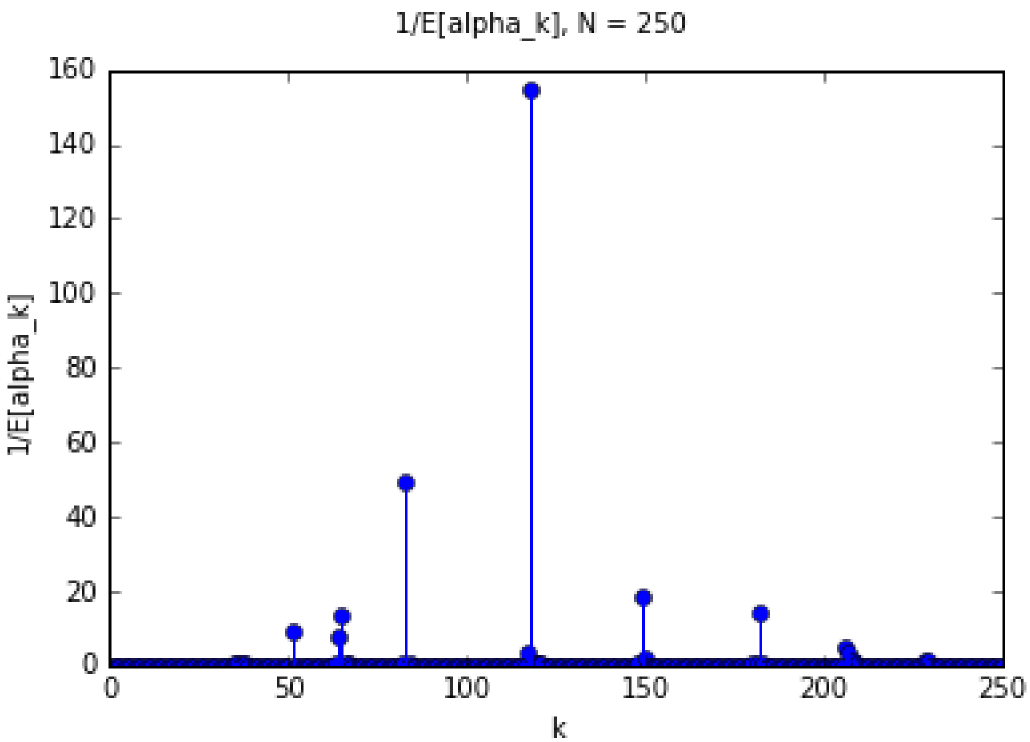


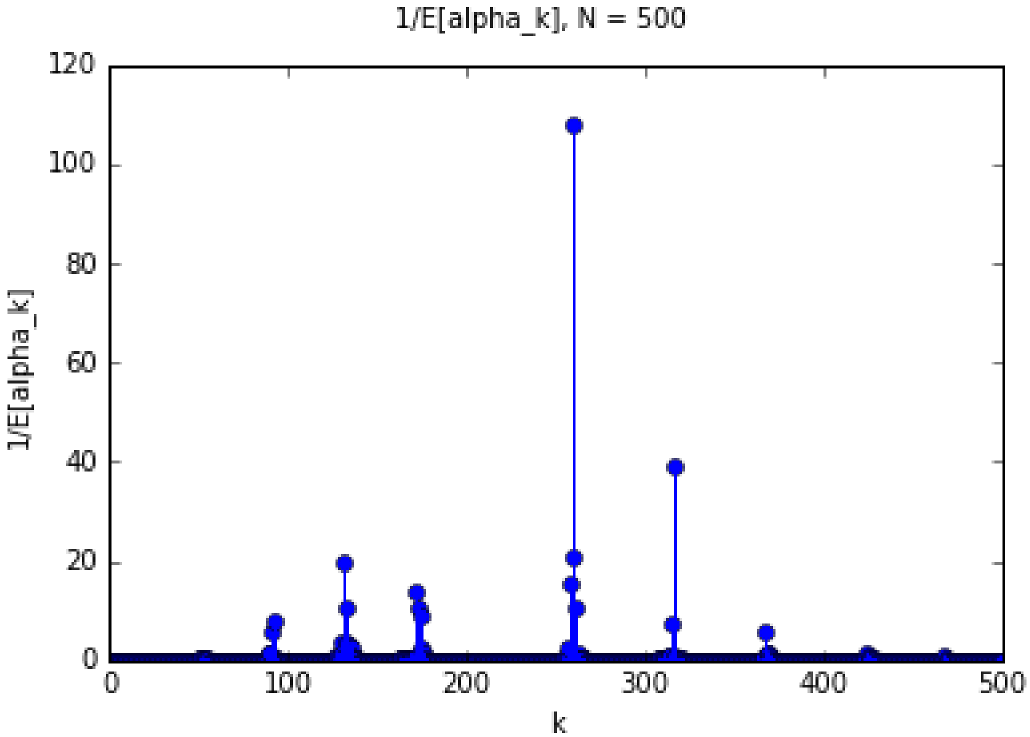

Variational objective function, N = 250
100
95
90
85
80
75
0
100
200
300
400
500
Iteration


Variational objective function, N = 500
200
190
180
170
160
0
100
200
300
400
500
Iteration

b,


1/E[alpha_k], N = 100
1/E[alpha_k]
k
0
20
40
60
80
100
120
140
160
180


1/E[alpha_k], N = 250
1/E[alpha_k]
k
0
50
100
150
200
250
20
40
60
80
120
140
160



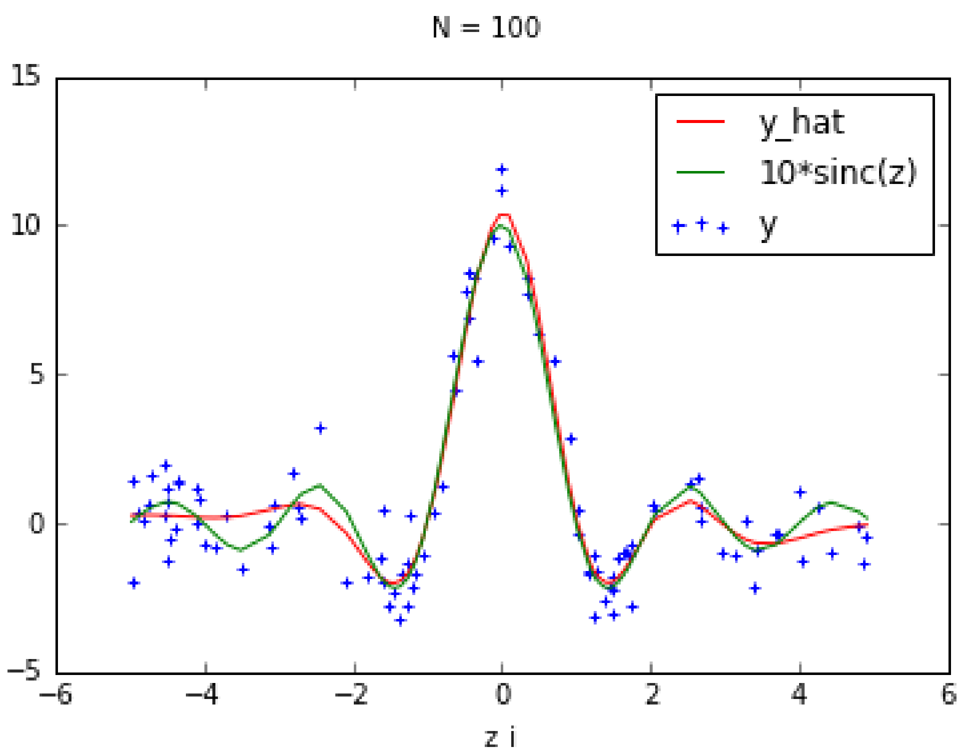
c,

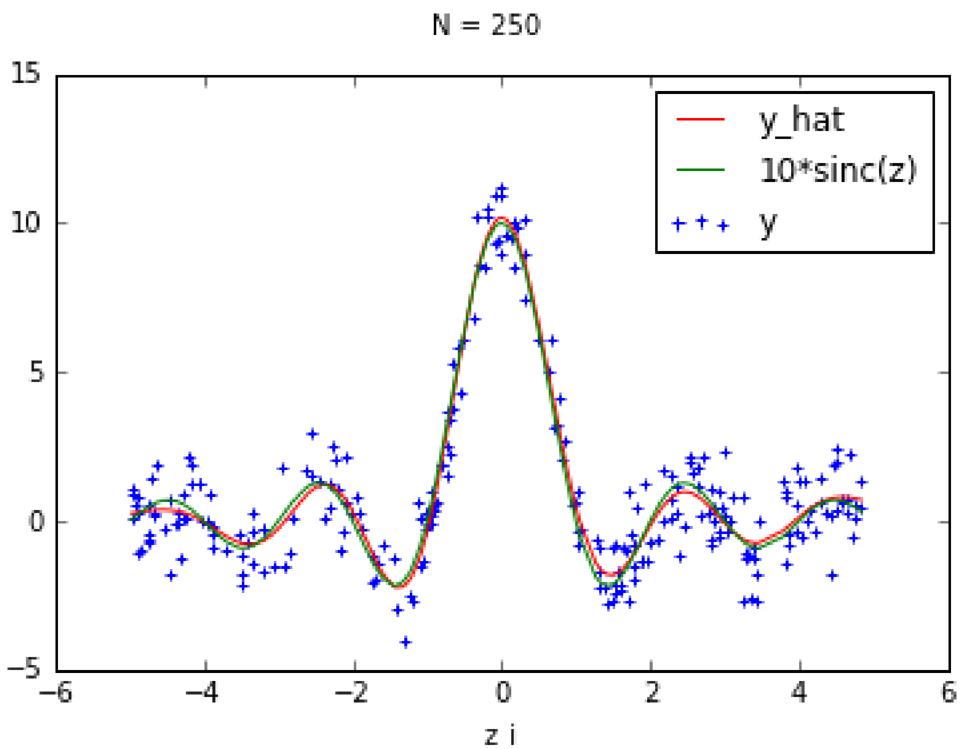
N = 100: 1/E[λ] = 1.07981360709

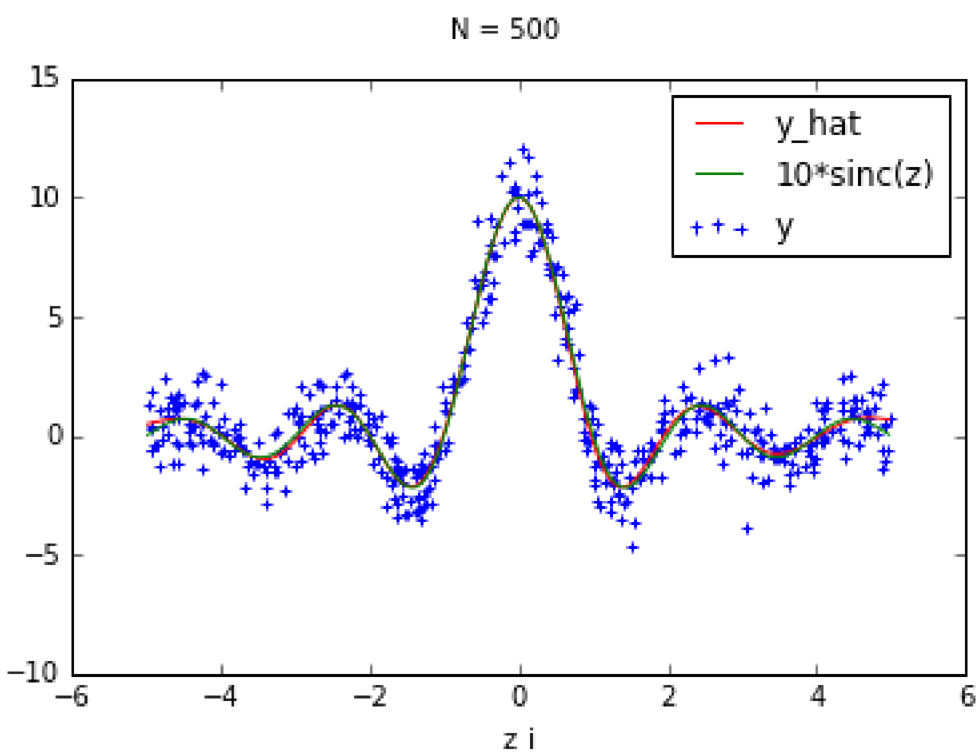
N = 250: 1/E[λ] = 0.899449415713

N = 500: 1/E[λ] = 0.978100451976

d,




N = 250
y_hat
10*sinc(z)
y
z_i


N = 500
y_hat
10*sinc(z)
y
z_i