# Bayesian Machine Learning Heating Load Dataset Analysis

Adam El Kholy

Bachelor of Science in Computer Science with Artificial Intelligence
The University of Bath
2024

## 0.1 Exploratory Analysis and Linear Least Squares Model

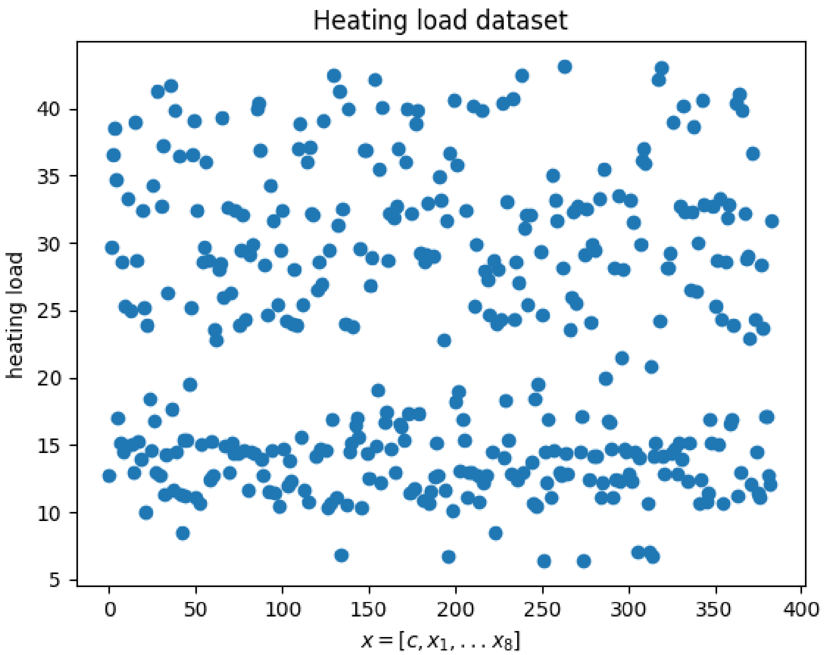


Figure 1: Graph showing all datapoints $x$ against heating load values $y$ for the heating load dataset

We conducted an initial exploratory analysis of the variables affecting heating load so as to appropriately inform our modelling decisions. Figure 1 shows a graph of training set data points (where $x_i = c, x_1, ...x_8$) against heating load, in which the prediction of heating load appears to be a non-linear task. Figure 2 shows a breakdown of the effects of each individual variable on heating load.

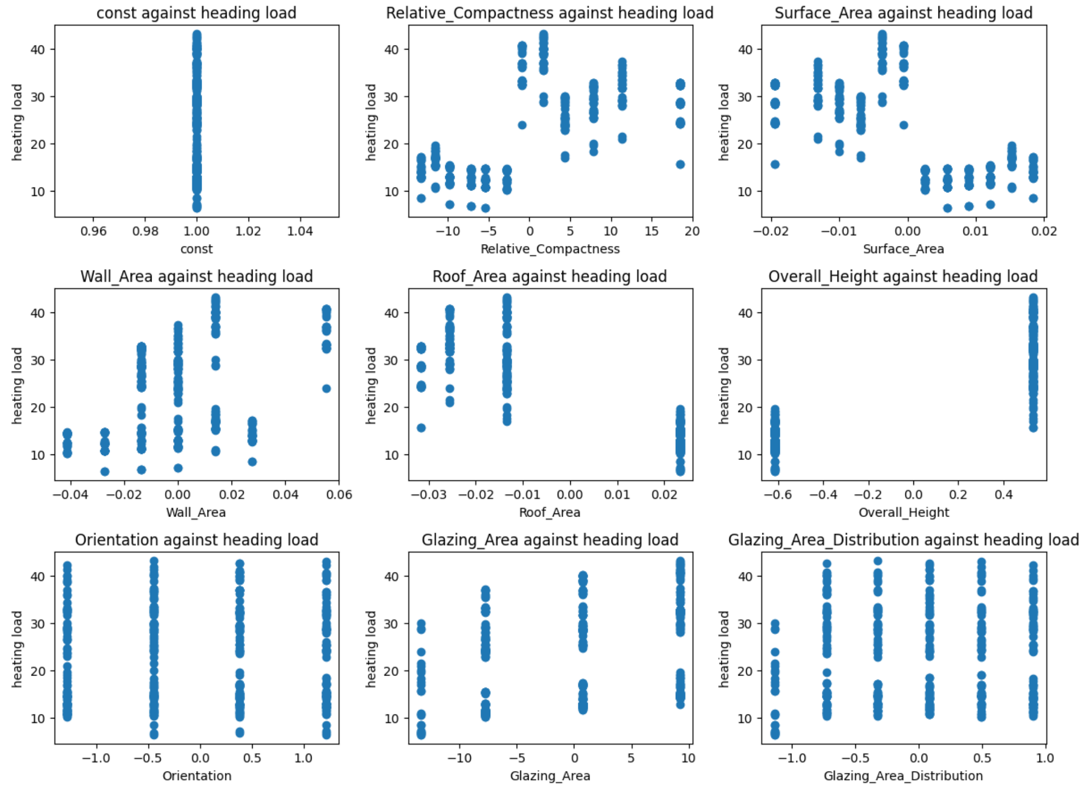


Figure 2: Graphs showing each individual variable comprising $x$ against heating load $y$

From Figure 2 we observe that orientation and glazing area distribution appear to be the least informative variables for predicting heating load, as the distribution of data points is uniform over all x values. Overall height appears to be one of the most informative variables as we observe an apparently linear relationship between overall height and heating load (lower overall height results in lower heating load and vice versa). Our analysis of individual variables is somewhat limited as we have not considered any relationships between individual variables and we note that there is likely a large degree of co-linearity between individual variables.

Throughout our analysis we use three performance metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE), defined respectively as follows (where $y_i$ is the true value, $\hat{y}_i$ is the predicted value and $n$ is the number of samples)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

As a benchmark for further analysis, we fit a linear least squares model to the train and test set, shown in Table 1.

| | Train | Test |
|---|---|---|
| MAE | 2.131 | 2.388 |
| MAPE | 0.097 | 0.125 |
| RMSE | 3.012 | 3.122 |

Table 1: MAE, MAPE and RMSE for the linear least squares model on training and test data

The linear model performs reasonably well when predicting heating load, with an admissible MAE and MAPE on both the test and train data. An MAPE of 0.125 on the test set represents an average percentage difference of 12.5% between the predicted and true value, a moderate MAPE which can likely be improved upon through use of a more complex model.

## 0.2 Bayesian Linear Regression

We calculated the log posterior distribution $log p(y|\alpha, \beta)$ where $\alpha = 1/\sigma_w^2$ and $\beta = 1/\sigma_e^2$ over a range of values for $\alpha$ and $\beta$ in order to find the most probable hyperparameter values for Bayesian linear regression using Type II maximum likelihood. $\alpha$ here denotes the precision of the prior and $\beta$ denotes the precision of the noise. We tested 100 values linearly spaced over the range [-5, 0] for both $\alpha$ and $\beta$ as shown in Figure 3.

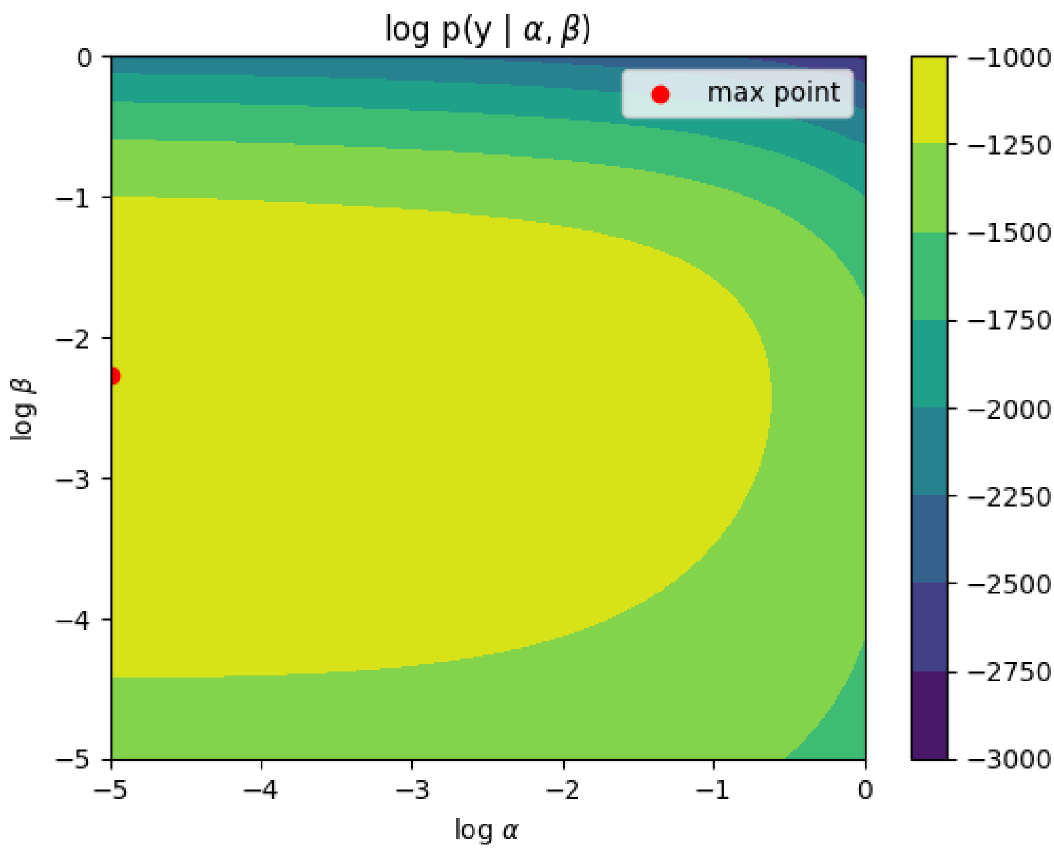


Figure 3: Contour plot of the log posterior distribution over log($\alpha$) and log($\beta$)

While the shape of the contour plot for our log posterior distribution is consistent with our expectations, our maximum point value appears to be potentially erroneous. We would not expect the maximum $log\alpha$ value to be on the edge of the graph, suggesting that we have either defined our range of values poorly or miscalculated the log marginal likelihood. As such we calculated the log posterior of a different range for $\alpha$; 100 linearly spaced values over the range [-10, -5].

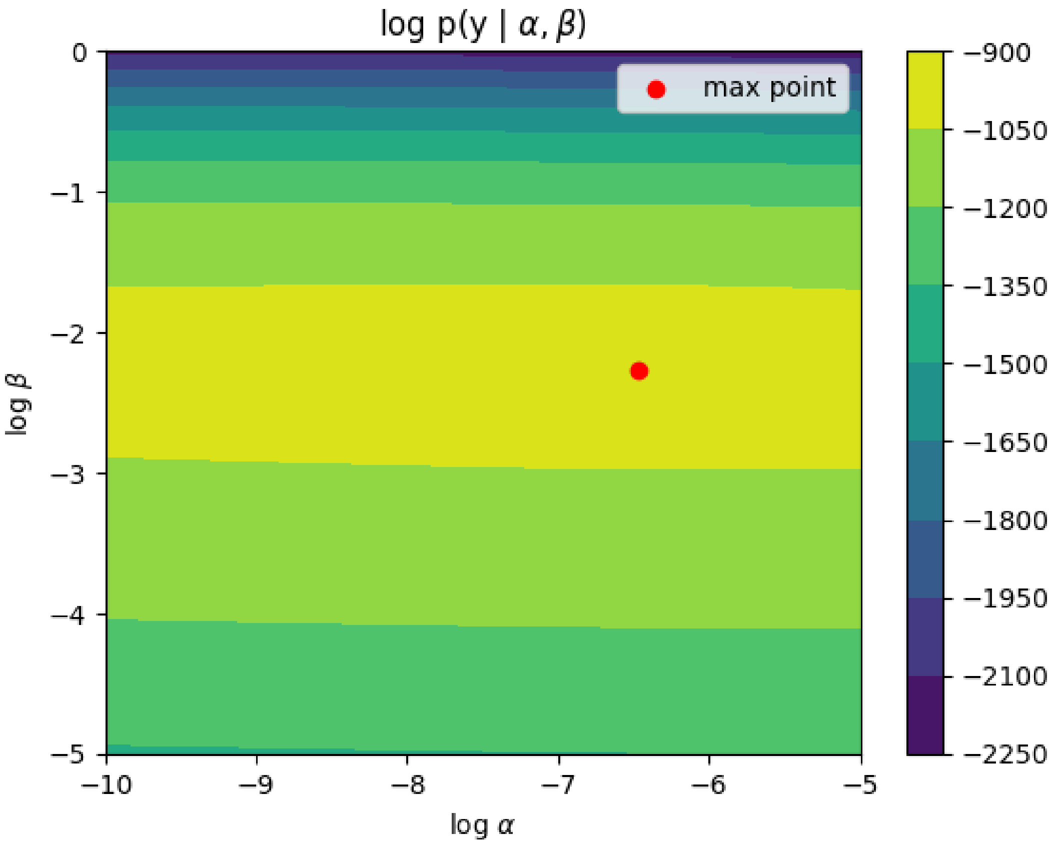


Figure 4: Contour plot of the log posterior distribution over log($\alpha$) and log($\beta$) with an adjusted range [-10, -5] for log($\alpha$)

This new range appears to include the true maximum value for log($\alpha$). We then proceeded to evaluate a Bayesian linear regression model using the most probable hyperparameters $\alpha$ and $\beta$, as shown in Table 2.

| | Train | Test |
|---|---|---|
| MAE | 2.221 | 2.338 |
| MAPE | 0.104 | 0.123 |
| RMSE | 3.056 | 3.125 |

Table 2: MAE, MAPE and RMSE for the unoptimised Bayesian linear regression model on training and test data

These results are extremely similar to those of the linear least squares model (Table 1). This may suggest an error in our calculation of the log posterior distribution, as we would expect the Bayesian linear regression model to outperform the least squares model.

## 0.3 HMC with Standard 2D Gaussian

We then set about implementing an HMC sampler for later use in optimising our Bayesian linear regression hyperparameters. We used the following standard 2D Gaussian distribution to verify our HMC implementation

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2 + y^2}{2\sigma^2})$$

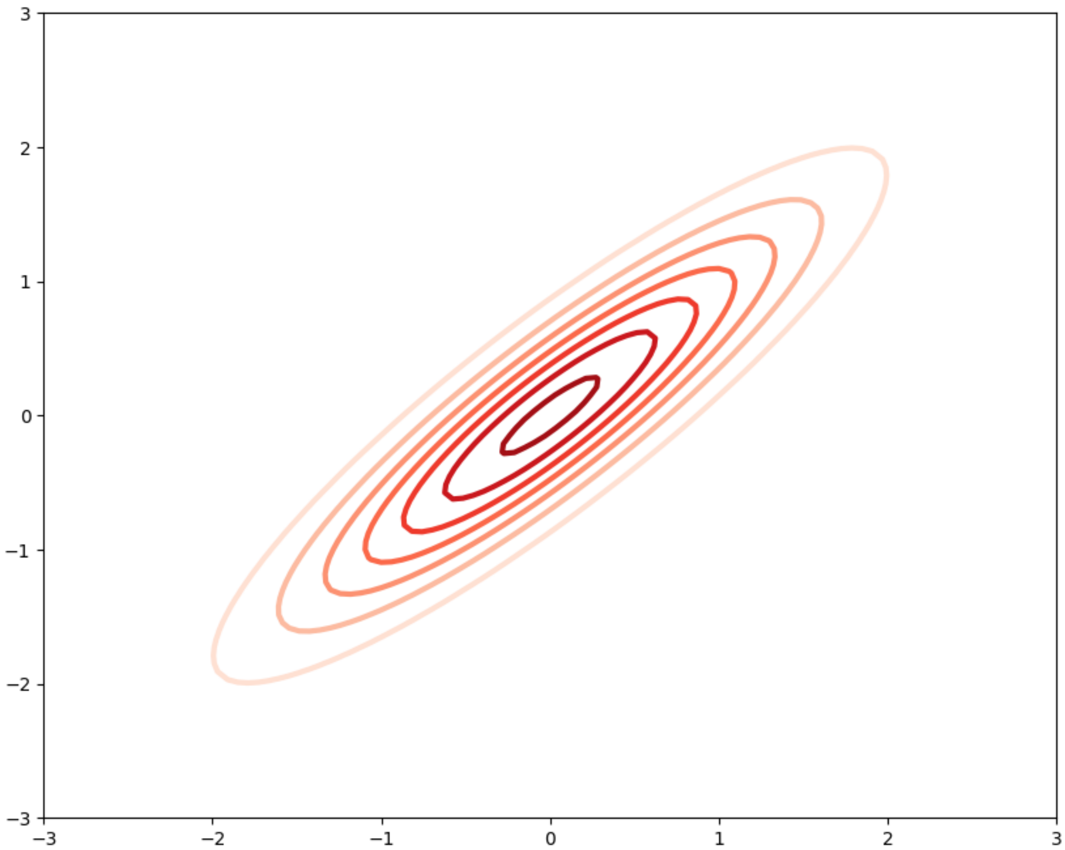


Figure 5: Contour plot of the standard 2D Gaussian distribution

We defined the following energy function for our HMC sampler for the standard 2D Gaussian distribution

$$g(x, \Sigma) = -\frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2} x^\intercal \Sigma^{-1} x\right)$$

where $\Sigma$ is the covariance matrix and $k$ is the rank of $\Sigma$. We then calculated the partial derivative of the energy function with respect to $\Sigma$, used to update our HMC sampler, as follows

$$\frac{\partial g}{\partial \Sigma} g(x, \Sigma) = \frac{1}{2}(x^\intercal \cdot (\Sigma^{-1})^\intercal) + (x^\intercal \cdot (\Sigma^{-1})^\intercal)$$

We used the following hyperparameters in our HMC implementation for the standard 2D Gaussian distribution $R = 5,000, L = 20, \epsilon = 0.5$, where $R$ is the number of leapfrog steps, $L$ is the leapfrog step size and $\epsilon$ is the step size, resulting in an acceptance rate of 78.0%.

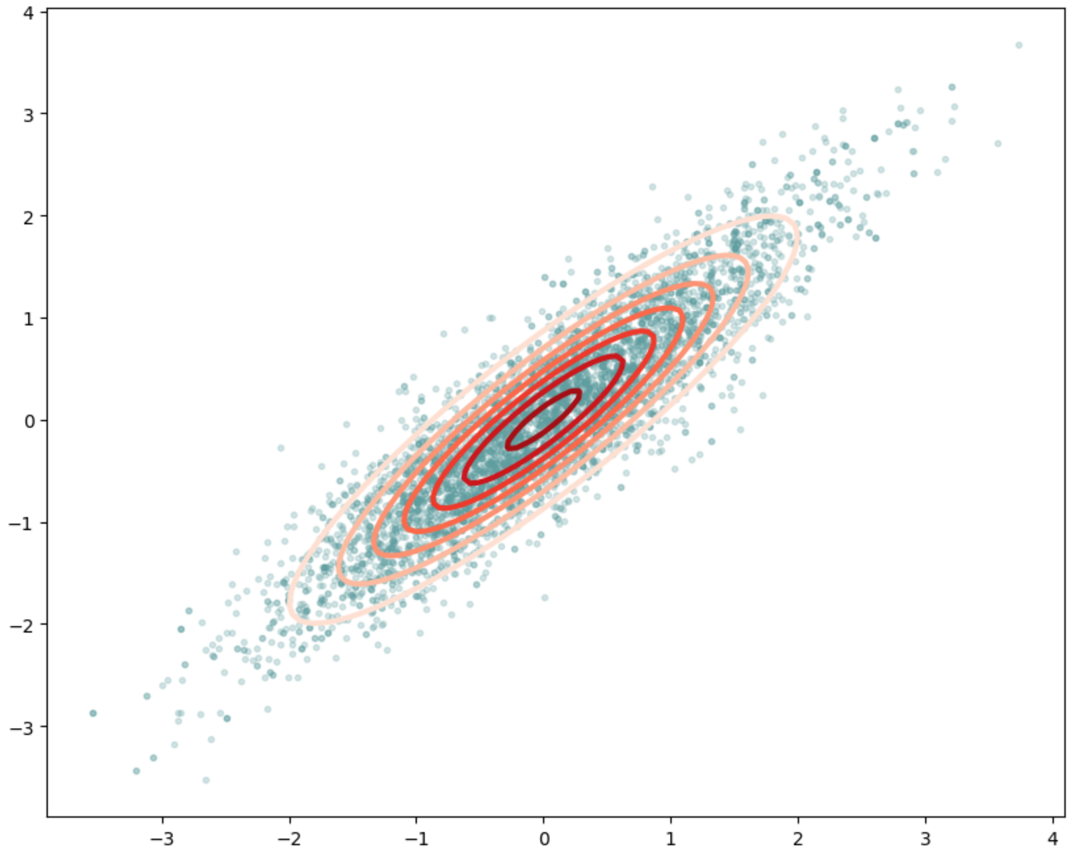


Figure 6: Contour plot of the standard 2D Gaussian distribution with the accepted values from the HMC sampler overlayed

Evidently our HMC implementation was working correctly, as the points accepted by the sampler closely resemble the shape of the original 2D Gaussian distribution. We acknowledge the limitations of our implementation and identify room for improvement in the accuracy of our sampler, perhaps through further tuning of the HMC hyperparameters $R, L, \epsilon$. We then endeavoured to implement a HMC sampler to optimise the hyperparameters of our Bayesian linear regression model.

## 0.4 HMC with Bayesian Linear Regression

We then implemented an HMC sampler in order to optimise our Bayesian linear regression model, namely the weights $w$ and the hyperparameters $\alpha$ and $\beta$, which affect the calculation of the weights $w$. We verified that our energy and gradient functions were consistent with one another, as shown by the output of our gradient consistency check below, where $Acc.$ denotes the degree of precision to which our calculation matches the true gradient $(1 \times 10^{-Acc})$.

```
 Calc.         Numeric       Delta          Acc.
 29.5076       29.5076      -7.866684e-07   8
 94213.3       94213.3      -2.100292e-06   11
-4886.63      -4886.63      -8.113351e-06   9
-32642.7      -32642.7       3.765224e-06   10
 44.674        44.674       -3.153689e-07   9
-23.06        -23.06         3.901517e-06   7
 91.7814       91.7814       1.017210e-06   8
-2113.3       -2113.3       -1.076671e-05   9
-49.6927      -49.6927       7.549343e-06   7
-6738.53      -6738.53      -6.445938e-06   10
-27.1075      -27.1076      -8.339655e-06   7
```

We used the following hyperparameter values in our HMC sampler for linear regression: $R = 10,000$, $L = 20$, $\epsilon = 0.014$, resulting in an acceptance rate of 94.1%. Our final optimised weights and hyperparameters were as follows

$$\begin{aligned}
\alpha &= -9.694004386315664 \\
\beta &= -2.22523522062341 \\
w_0 &= \phantom{-}22.921442221601687 \\
w_1 &= -0.8411908621976004 \\
w_2 &= -171.02450704194484 \\
w_3 &= -24.812910122815605 \\
w_4 &= -312.2043868001974 \\
w_5 &= \phantom{-}12.517084738329297 \\
w_6 &= -0.13135510402158124 \\
w_7 &= \phantom{-}0.3680597741878445 \\
w_8 &= \phantom{-}0.32502395397473216
\end{aligned}$$

Using the optimised weights in our Bayesian linear regression model we evaluated the model on the test data (Table 3) and compared its predictions against the ground truth values (Figure 7).

| | Train | Test |
|---|---|---|
| MAE | 11.651 | 2.399 |
| MAPE | 0.627 | 0.125 |
| RMSE | 14.194 | 3.125 |

Table 3: MAE, MAPE and RMSE for the HMC sampler optimised Bayesian linear regression model on training and test data

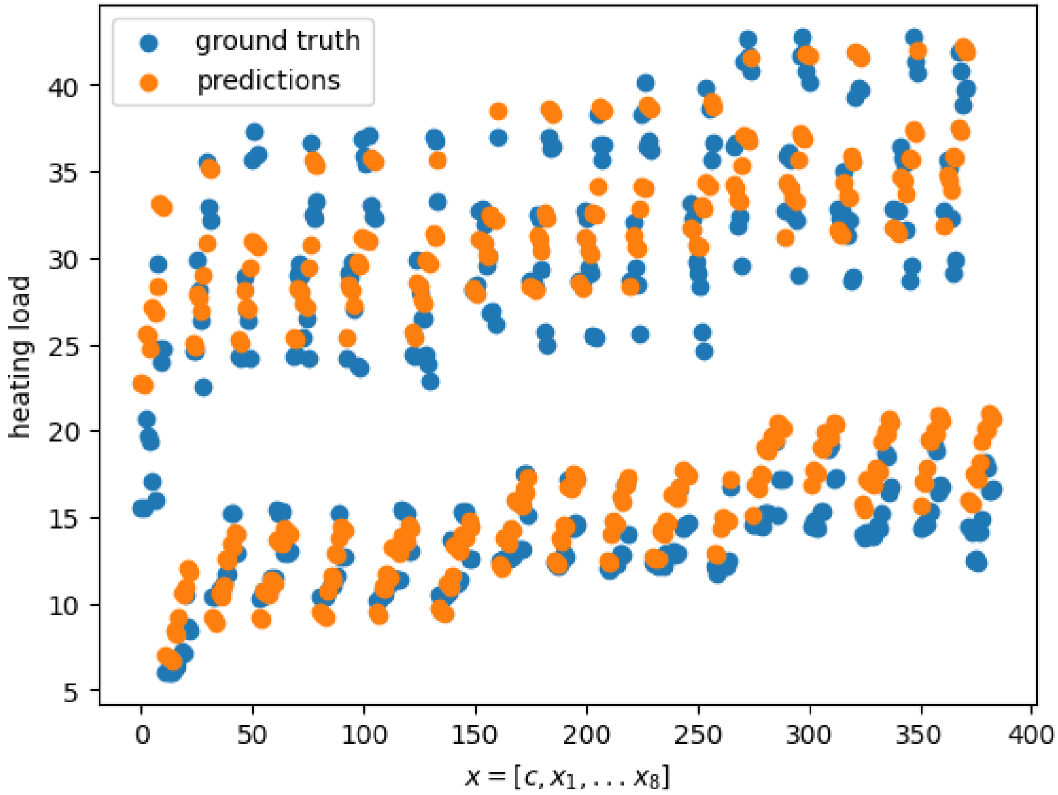


Figure 7: Predictions of the HMC sampler optimised Bayesian linear regression model against ground truth values for heating load

We then conducted a comparison of all of our models' performances on the test set.

| | Linear Least Squares | Linear Regression | Optimised Linear Regression |
|---|---|---|---|
| MAE | **2.388** | 2.338 | 2.399 |
| MAPE | 0.125 | **0.123** | 0.125 |
| RMSE | **3.122** | 3.125 | 3.125 |

Table 4: MAE, MAPE and RMSE for all three models on the test set

Interestingly, all of our models achieve very similar performances on the test set, with a minimal difference in MAE, MAPE and RMSe across all models. In our exploratory analysis we commented that the data appeared to be non-linear, suggesting a high level of co-linearity between the variables. This may serve to explain the minimal variance in performance between our linear models, as the non-linear data may therefore be challenging to predict to a low error rate using a linear approach. Given further scope we would have liked to investigate the performance of a non-linear model in order to test this hypothesis. It is also possible that there is an error in our calculations in our HMC Bayesian linear regression model, which we would theoretically expect to see outperforming the unoptimised model (which it does not). Our final test set RMSEs of around 3.12 demonstrate a moderate performance across all models, making them potentially useful in predicting heating load to a lower degree of precision.

Word count: 997