# 2.874/2.884/10.354/10.554 Process Data Analytics Solution for Homework 1

1. Your objective is to construct models for estimating paracetamol[[1]](#footnote-2) concentration from ATR-FTIR spectra and temperature collected in a batch purification process. Training, validation, and testing data are provided in FTIR\_train.mat, FTIR\_val.mat, and FTIR\_test.mat, respectively, in the file HW-1-data.zip. Each file contains a predictor matrix **X** and the corresponding paracetamol concentration **y**. Each row in the predictor matrix denotes a different observation, while each column denotes absorbance at different frequencies (see Figure 1 for some representative spectra) and the last column contains temperature data. The frequencies are contained in the vector freq. Training data should be used to train the model parameters, validation data should be used to find optimal hyperparameter(s), and test data should be used to report the average prediction error.
2. Visualize the data and justify whether linear models are appropriate for this dataset. Report on whether any data that are likely outliers or biased. [Hints: analyze the two-variable correlation plot; analyze how a peak absorbance changes as a function of temperature and concentration]

**Solution:** This solution key contains much more information than needed to receive all of the points. The solution key contains multiple correct ways of answering the same question, and has additional information to convey best practices when approaching a data analytics problem.

A good practice is to first plot the raw data (Figure 1) to gain some perspective and an intuitive feeling for the dataset, look for biases and outliers, assess whether there are any peculiarities in the data, and determine conditions where a model would likely have validity. Inspection of Figure 1 indicates that each spectrum is highly correlated with other spectra, and the absorbances within each spectrum are highly correlated. Each spectrum is evenly spaced from the nearest-neighboring spectrum except for the spectra at the two lowest concentrations, suggesting that at least one of those spectra are biased.

  
**Figure 1.** Some spectra at different concentrations (first discussed in Lecture 2).

Insights into a dataset can be obtained by plotting well-chosen subsets of the data. Since the largest signal-to-noise ratio is expected to be at the peak absorbances, those absorbances would be expected to play a significant role in a model for predicting concentration. Figure 2(left) plots temperature and a peak absorbance on the axes, with each concentration marked by a different color. The six concentrations nearly span uniformly over the concentration space, which is a good experimental design for building a model for predicting the concentration from temperature and spectra measurements with the 2D region of temperature and concentration shown in Figure 2(right). The dataset at 0.014 g/g appears to have a large bias at low temperatures, in that the dataset are much closer to the data at 0.01 g/g than the data at 0.02 g/g – although 0.014 is near the average of 0.01 and 0.02. [The observation that the rest of the data appear to be roughly linearly behaved is why it is expected that much of the 0.014 g/g data are biased, mostly likely by some type of experimental error. If we were working with an experimentalist who gave us these data, we would advise that the person repeats the 0.014 g/g experiment.]



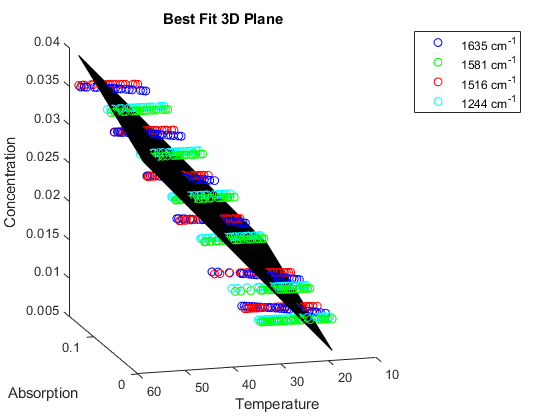
**Figure 2.** Preliminary data analysis indicates that a linear calibration model may have adequate prediction accuracy for most of the data, and that data for 0.0145 g/g appear to be biased.

Another visual way to assess whether a linear model is appropriate is to generate the variable correlation plot for all of the data (Figure 3, which was discussed in detail in Lecture 2).[[2]](#footnote-3) A high degree of linearity is observed between all two-variable combinations, indicating that linear models would be expected to describe most of the data.



**Figure 3.** Correlation plot for the concentration, absorbance for the frequencies (1635, 1581, 1516, and 1244 cm−1) and the temperature for the training data.

Yet another way to assess linearity is to see whether plots of the versus temperature and a peak absorbance lie within a plane for several choices of peak absorbance. Each set of points approximately falls into a plane, which suggests that a linear multivariate model would be appropriate.



**Figure 4.** Planar approximation of concentration using temperature and the absorbances at 4 frequencies.

**Pitfalls:**

Did not try, i.e., the solution is missing a figure (−25 points)

Unconvincing plot or explanation for linearity (−15 points)

**Kudos:**

Linearity shown for most of the data points and pointing out that some of the data appear biased

1. Given the amount of data, is it recommended to use ordinary least-squares regression to construct a model that depends on all of the absorbances? Why or why not?

**Solution:** It is **not** recommended to apply ordinary least-squares regression for the entire range of frequencies; since the number of predictors () exceeds the number of observations (), the solution to such a regression is not unique.

**Pitfalls:**

Did not try (−15 points)

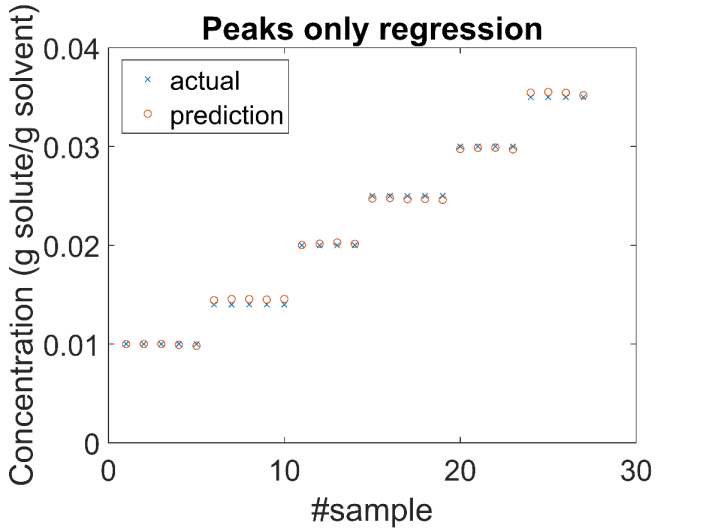
Answered “yes” (−10 points)

Poorly reasoned discussion (−5 points)

1. Construct a linear regression model based on the absorbance at frequencies 1635, 1581, 1516, and 1244 cm−1, and temperature. Specify the objective function to be minimized and find optimal weights (either through the closed-form solution or solving the objective function directly). Report the average prediction error for the test data and plot the training, validation, and test data with associated model predictions. Visualize the results and comment on the quality of the model predictions.

**Solution:** The vector of regression coefficients obtained by ordinary least squares can be computed analytically from where the data matrix for this problem only contains the relevant columns corresponding to temperature and absorbance at the four frequencies, and has a column of ones corresponding to the “y-intercept”. The column of ones is not needed if and are normalized using the z-scores of the training data (subtract the mean and divide by the standard deviation of the training data for each variable), in which case the y-intercept is always zero. Numerical solution of the OLS objective is an expensive way to compute the same result. Whether z-score is used here does not matter, as z-score normalization does not affect the results of ordinary least squares.

The RMSE should be reported for the test data, which for this problem was 3.25×10−4. The training and/or validation data should not be included in the computation of RMSE, since that will make the model accuracy look better than it is. If the data were normalized by the means and standard deviations of the training data, the data need to be unnormalized to correctly calculate the testing RMSE. The student should generate a plot of all of the data and associated model predictions (Figure 5). The model predictions are systematically lower or higher than the data for some concentrations, indicating that the model assumption of independent errors does not strictly hold. In spite of this, the residuals are relatively small compared to the absolute value of the concentration, so the model predictions are potentially useful, depending on the needs of the specific application.



**Figure 5.** Prediction results using absorbances at the peaks only.

**Pitfalls:**

Did not try (−20 points)

Did not extract the correct columns (−5 points)

Did not normalize and using the training data set only *and*   
did not add the column of ones into (−5 points)

Did not solve the correct equation (−5 points)

Calculated RMSE using the wrong data set, or reported an RMSE   
based on normalized data instead of the actual data (−5 points)

1. Apply ridge regression, lasso, and elastic net to the temperature and the absorbance at all frequencies. Report the average prediction error for the test data for each method. [NOTE: Matlab has a *ridge* function, in addition to the commands given in Lecture 2. The built-in *lasso* command in Matlab is unreliable for both lasso and elastic net, whereas the associated commands in the SpaSM Matlab toolbox are reliable.]

**Solution:** Usually data are z-scored before applying ridge regression, lasso, and elastic net, to account for widely different ranges of values for each data type (e.g., a temperature could more than 100 degrees for a system in which absorbance spans 0.01 absorbance units). This particular dataset consists of spectra, however, in which case skipping z-scoring can result in better models (as briefly discussed in Lecture 2). The reason for this different behavior is that typically many absorbances are not associated with the molecule of interest and contain nearly all noise; z-scoring normalizes those absorbances to have the same magnitude of importance in the matrix in the optimization objective, which feeds a bunch of noise into the parameter estimation and degrades the quality of the model. Also, the benefits of z-scoring are lower for spectral data, as most of the absorbance values are already of a similar order of magnitude. Temperature can be measured extremely accurately even with inexpensive thermocouples, and so model quality is not reduced by using nonnormalized large temperature values in the model construction. Some data analytics experts do not z-score spectral data, and other experts compare models constructed from both z-scored and nonnormalized data to see which generates smaller prediction errors for the test data.

All of the data analytics methods in part d use the optimization objective

where is a penalty term that depends on the parameters and a hyperparameter . By adding the penalty term, the regression is well-defined even if *p* > *n*, and as a consequence, the absorbance can be included at every frequency. *Ridge regression* uses the 2-norm penalty . *Lasso* uses the 1-norm penalty which usually returns a sparser model (that is, some regression coefficients equal to zero) useful if interpretability is of importance or physical/chemical/biological evidence indicates that the underlying model should be sparse.

Data analytics follows the reasoning:

Training data → Determine parameters given values of hyperparameters

Validation data → Choose values of the hyperparameters ()

Test data → Show that the optimal values of the model parameters   
(aka regression coefficients) perform well

Usually the hyperparameters () are chosen to minimize the validation RMSE (or MSE, or SSE, or some other metric of error).

For this particular dataset, using nonnormalized data to build the model results in a lower test error (e.g., 1.73×10−5 for nonnormalized vs. 1.84×10−5 for z-scored data for ridge regression), which commonly occurs when building models that are fed spectral data.

**Ridge Regression**

The analytical expression for the regression coefficients in ridge regression is . Either using this equation or correctly using *ridge* in MATLAB to z-scored data results in (see Figure 6) and a corresponding RMSE of the test set of 1.84×10−5.

All of the data and associated model predictions are plotted in Figure 7. The model predictions are observed to have very small errors and no noticeable biases for all data. The test RMSE is one order of magnitude smaller than that of ordinary linear squares. In the ordinary linear squares of part (c), the most important frequencies were guessed from knowledge of which points would have the highest signal-to-noise ratios, and discarded the data from all other frequencies. When building models based on spectral data, such discarding of nearly all of the frequencies before carrying out any data analytics nearly always produces much less accurate models than using elastic net to remove the unimportant frequencies. One case where removing the absorbances from entire frequency ranges is a good practice is when the molecules of interest are known from first principles to have zero absorbance in some frequency ranges.

Some data analytics software applies z-scoring to but not (e.g., *ridge* in Matlab). You should always make sure that your results such as RMSE calculations correctly account for any normalizations, e.g., to make sure that your RMSE for the test set is in terms of the original units.



**Figure 6.** Cross-validation results for ridge regression when the data fed to the algorithm is z-scored (left) and nonnormalized (right).



**Figure 7.** Raw data and model predictions from ridge regression fed z-scored (left) and nonnormalized (right) absorbances and temperature.

A comparison of the plot of the regression coefficients as a function of frequency with an example spectrum indicates that the regression coefficients from ridge regression seem somewhat reasonable and somewhat noisy (Figure 8). Some parts of the plot of the regression coefficients are qualitatively have a similar shape as the spectrum, such as for the peak at about 1250 cm−1. Regression coefficients are small for the huge peak at about 1600 cm−1 but a quick Google search indicates the peak is from water rather than paracetamol.[[3]](#footnote-4) Some parts of the plot of regression coefficients are noisier than the spectrum, suggesting that the ridge regression is fitting some noise in the data. Since we applied ridge regression correctly, we should suspect that something may be fishy about the dataset, e.g., may include some biased data and/or may have poor splitting of the data into the training, validation, and test datasets. [These suspicions would be correct. Biased data was pointed out in point a, which would result in a poor model if included in the training and/or validation datasets, and would result in a poor estimate of the prediction error if included in the test dataset. We will discuss ways to address these problems later in the course.]



**Figure 9.** Cross-validation results for lasso run using the SpaSM toolbox.



**Figure 8.** Comparison of regression coefficients from ridge regression with absorbances for the spectrum for training sample #120.

**Lasso**

Lasso requires a numerical solution such as by using the SpaSM toolbox on MATLAB. The implementation of lasso in Matlab is not high quality and can sometimes give wrong results. Using the SpaSM toolbox, the optimal is found to be 1.14×10−4(see Figure 9). At the optimal , the RMSE on the test dataset is 2.09×10−5. At the optimal , 41 predictors were selected. Figure 10 shows accurate model predictions with negligible bias. Lasso places many nonzero regression coefficients near two of the peaks in the spectrum as would be expected, but also included many peaks where the absorbances are small, which is not expected and suggests that there could be overfitting. Regression coefficients are small for the big water peak at about 1600 cm–1, which is reasonable.

 **

**Figure 10.** Prediction results using lasso run using the SpaSM toolbox (left) and corresponding regression coefficients with a spectrum shown for comparison purposes (right).

If you have used *lasso* in Matlab, the lowest validation RMSE is found at equal to 0, as illustrated in Figure 11, which is clearly incorrect because that value turns off the regularization which results in a linear regression problem that is not well-posed.



**Figure 11.** Validation RMSE obtained using the Matlab *lasso* function. No minimum is found down to the precision limit.

**Elastic Net**

Elastic net uses the penalty function . Whereas is often referred to as in the literature and in Lecture 2, this solution key uses since that symbol is used in the elasticnet function in the SpaSM package that was used here. Choose 𝛿 and using the validation data. The elasticnet function accepts and as before, but uses 𝛿 values as an independent input rather than . An optimal is determined on the backend when the code is run.

Figure 12 indicates the resulting validation RMSEs and s for different values of 𝛿.





**Figure 12.** Cross-validation results for elastic net, showing error as a function of 𝛿 with optimal , of for the optimal 𝛿, and comparing the optimal s found as a function of 𝛿.

**Figure 13.** Experimental data and model predictions using elastic net (left) and regression coefficients with a spectrum for comparison purposes (right).

Figure 13 shows biased predictions, overestimating sub-mean concentrations and underestimating greater-than-mean concentrations. While the minimum validation RMSE obtained for elastic net is comparable to the errors obtained using ridge and lasso (2.03×10−5), the corresponding test RMSE for these coefficients is significantly higher at 8.62×10−4. This large error is likely a result of underfitting, as seen in the plot of regression coefficients which are only nonzero at the highest peak absorbance.

**Pitfall:**

Did not try ridge regression (−10 points)

Did not try lasso (−10 points)

Did not try elastic net (−10 points)

Did not discuss potential normalization by z-score (−5 points)

Did not show cross-validation result (−10 points)

Calculated RMSE using the wrong data set, or reported an RMSE   
based on normalized data instead of raw data (−5 points)

1. Compare and discuss the results from (a), (c), and (d).

**Solution:** The RMSE of the test data are compared in Table 1. Ordinary linear regression (OLS) using only the peak absorbances gave the worst prediction error, which was an order of magnitude higher than ridge regression and lasso. Those two regularization methods produce better predictions by using all available information in the spectra. The worse performance of elastic net over ridge regression and lasso is unexpected, since elastic net reduces to each of those methodsfor particular choices of one of the hyperparameters. The large RMSE on the test data for elastic net suggests that there is something fishy about the data and/or its splitting into the training, validation, and test datasets. Part a pointed out that some of the data appear to be biased, and part d pointed out that including that data into any of the datasets would result in biased models or biased estimates of the prediction error. Since the biased data appear in all three datasets, all of the data analytics methods construct models that fit the highly biased data; the RMSE on the test data are relatively low for all of the methods because biased data also appear in the test data.

This example dataset illustrates how some visualizations of raw data can more easily identify biased data than others (which was seen in part a), and how biased data can result in misleading and/or unexpected results unless identified and removed from the analysis. In practice, based on the visualizations in part a, you would have informed the experimentalist of the biased data and advised repeating the experiment containing the biased data. If repeating the experiment is not an option, then you would remove the biased data from the original full dataset, and then split the resulting abridged data into training, validation, and test datasets. The region of missing data means that the splitting would need to be done with some forethought and that a more sophisticated cross-validation procedure would be needed, which are topics discussed later in this course.

**Table 1.** RMSE on the test data for different modeling techniques.

|  |  |  |  |
| --- | --- | --- | --- |
| **OLS** | **ridge regression** | **lasso** | **elastic net** |
| 3.25×10−4 | 1.85×10−5 | 2.09×10−5 | 8.62×10−4 |

**Pitfalls:**

Did not try (−10 points)

Lack of comparison or poorly reasoned discussion (−5 points)

1. Paracetamol is a painkiller also known as acetaminophen and is the active ingredient in Tylenol. [↑](#footnote-ref-2)
2. <https://www.mathworks.com/help/econ/corrplot.html> [↑](#footnote-ref-3)
3. <http://www1.lsbu.ac.uk/water/water_vibrational_spectrum.html> (the spectra are from ATR-FTIR spectroscopy) [↑](#footnote-ref-4)