2.874/2.884/10.354/10.554 Process Data Analytics Solution for Homework 3

1. Your objective is to construct models for estimating paracetamol1 concentration from ATRFTIR spectra and temperature collected in a batch purification process. The original training, validation, and testing datasets are 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. Although respective of a common practice in applications, the training, validation, and testing sets were poorly chosen in HW1, as the use of random sampling resulted in highly biased data appearing in all three datasets. Also, in HW1 you only implemented a single cross-validation, which can lead to inaccurate estimates of the prediction error. In this homework, your tasks are to (a) perform high-quality splits of the complete dataset into training, validation, and testing datasets; (b) build models for various sets and estimate the model stability, and (c) apply rigorous cross-validation procedures to estimate the prediction error for both unbiased and biased data. As before, 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. You are allowed to feed into your analyses any learning from your HW1 solution and the HW1 solution key.

(a) Combine the data into a single dataset, and then split the dataset into a good choice for the original training, validation, and testing datasets. An allowable splitting should consider that the complete dataset consists of six groups, and that the data in one part of the groups are biased. Visualize the data to identify which data are too biased to be used for training and validation. Biased data should be used in test data to estimate the prediction error for the biased data, which should be compared to the prediction error for unbiased data. Construct a linear regression model based on the absorbance at frequencies 1635, 1581, 1516, and 1244 cm^{-1}, and temperature. 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. Compare these prediction errors to the prediction error obtained for the original dataset split in HW 1. Explain your results, that is, why the relative magnitudes of the prediction errors make sense.

Solution (20 points): As discussed in the homework 1 solution, the data for $C = 0.014$ display significant bias.

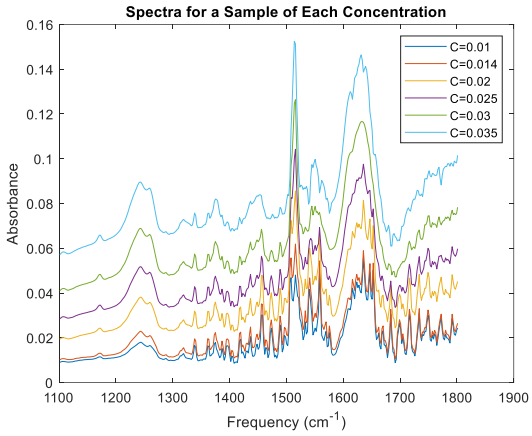


Figure a.1. Some spectra at different concentrations. All of the spectral data other than between the C = 0.14 and C = 0.01 spectra are evenly stratified, suggesting biased data for at least one concentration.

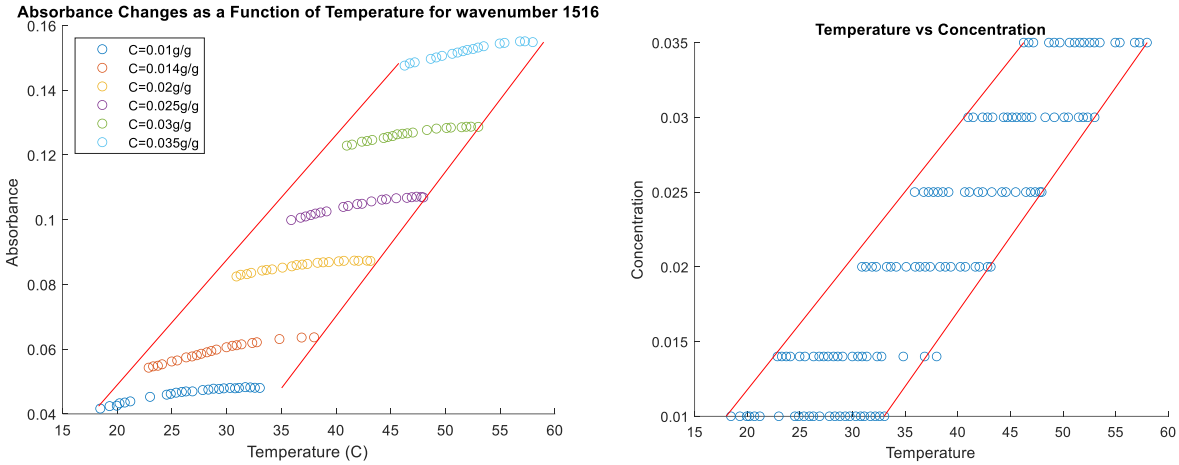


Figure a.2. Preliminary data analysis in the form of data visualization indicates that a linear calibration model may provide adequate prediction accuracy for most of the data, and that data for C = 0.014 g/g appear to be biased, especially at low temperature.

The original data are redistributed into training and two sets of test data by concentration for the regression fit in Table a.1.

Training Dataset	**C = 0.01, 0.02, 0.025, 0.035**
Unbiased Test Dataset	**C = 0.03**
Biased Test Dataset	**C = 0.014**

Table a.1: Data classification for regression model

The biased dataset should be separated as a test dataset for estimating the prediction error for that data, and the largest and smallest concentrations should be included in the training data, as extrapolating data from such fits is generally poor practice. There are no hyperparameters needed for a regression model, so a validation dataset is unnecessary. [Alternative solutions would be to use some other intermediate concentration for the unbiased test dataset, as long as C = 0.014 is not included in the unbiased test dataset.]

Using this data grouping results in the prediction errors in Table a.2.

	Original uniformly randomly sampled dataset (HW1)	Grouped datasets as described in the text: unbiased dataset	Grouped datasets as described in the text: biased dataset
Training RMSE	2.9734×10^{-4}	1.6159×10^{-4}	
Test RMSE	4.4875×10^{-4}	1.0×10^{-3}	1.8×10^{-3}

Table a.2: Model performance for an OLS regression model trained using the validation and test datasets as described in the text.

The average prediction error for the training data for uniformly sampled data in HW1 is larger than for the grouped data, which is consistent with the observation that the former is fitting a linear model to a dataset that includes biased data which deviate from linearity (e.g., Figure a.2). The test RMSE (average prediction error estimate) for the uniformly sampled data in HW1 is larger than for the training data but is still relatively small (that the RMSE for the test dataset is about 50% larger than for the training dataset is suspicious, and would motivate spending more time in constructing visualizations of the data in more detail, and/or assessing whether the RMSEs change significantly when a different random seed is used to select which data to include in the training dataset). The test RMSE for the uniformly sampled dataset is likely underestimating the RMSE that would be obtained by applying the model to new data; the model constructed for the training dataset is biased (due to the biased data), but that bias is largely unobservable in the calculated test RMSE which also includes the biased data, and the uniform sampling would result in any datapoint in the test dataset being very close to having the same values as some datapoint in the training dataset (e.g., Figure a.2).

Removing the group of data that are observed to be biased in Figure a.2 results in a smaller RMSE for the training dataset, since a linear model is better able to fit the data when the biased data are removed (compare the training RMSEs on the first row of the table). The RMSE for the biased test dataset is much larger than for the training RMSE, and is a more reliable realistic estimate of the RMSE that would occur if the model was fed data that are similarly biased. The RMSE for the unbiased test dataset is about a factor of two lower than for the unbiased data, but is also larger than the training RMSE, since this error estimate now includes the additional variability associated with datasets between different groups (that variability between different groups was largely hidden when the data was uniformly sampled as in HW1, as both its training and test datasets included a large number of datapoints from all six groups). The test RMSE of 1.0×10^{-3} is much more reliable estimate of the test RMSE that would occur if the model was fed data that are relatively unbiased, which is about a factor of two larger than the test RMSE computed based on the model constructed from uniformly sampled data. The new model will also be less biased, that is, less prone to making concentration predictions that are too high or too low compared to the true concentrations.

Next, we compute the residual errors (aka residuals) by $\varepsilon = y - \hat{y}$, and visualize the results, as a check on whether the underlying assumptions on the model errors are reasonable, e.g., that the error variance is approximately uniform and random (Figures a.3 and a.4).

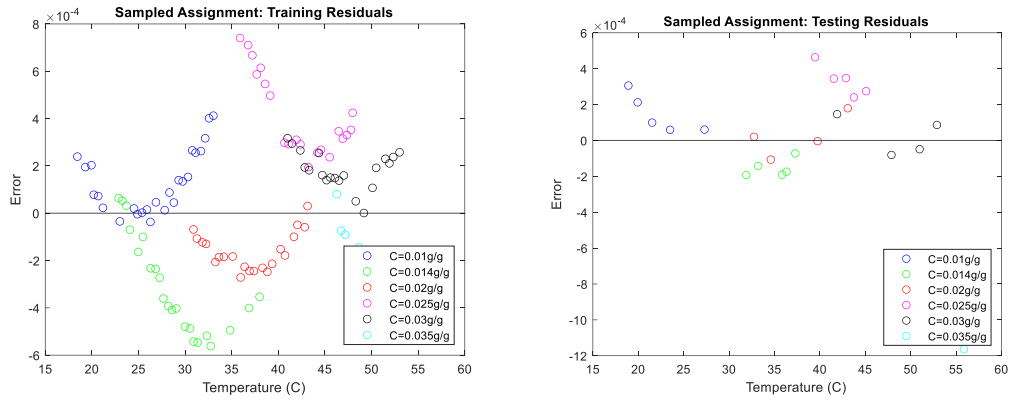


Figure a.3: Residuals for data sampled across all runs as in HW1

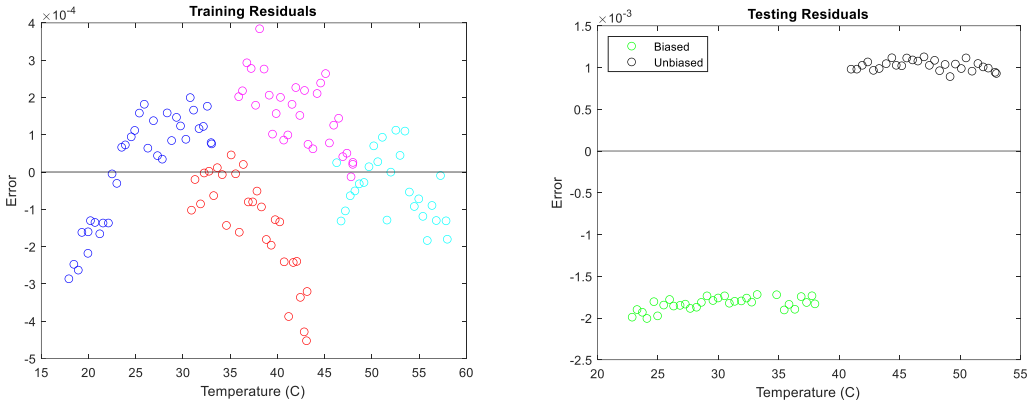


Figure a.4: Residuals for data grouped by concentration as described above. Concentration colorations are consistent across Figures a.3 and a.4.

The first observation is that the residuals do not look like "zero-mean with constant variation" for either model (which is commonly observed in spectral data, largely due to spectral baseline drifts and the much higher between-group variability compared to within-group variability). The residuals are more highly structured for the model constructed from uniformly sampled data (cf. the left plots in Figures a.3 and a.4), with what appears to be quadratic dependency on temperature within each group, with similar curvature for each group. The residuals for the group-trained data do not show any such quadratic dependency on temperature in each group, and instead have a linear dependency in each group. The residuals for the test datasets for both models show a dependency on which data comes from which group (cf. the right plots in Figures a.3 and a.4). For the group-trained data (right plot in Figure a.4), the residuals are very random within each group but have nonzero means, which indicates that the variability between groups is much larger than the variability of data within each group. This observation is consistent with examination of the data variability in Figure a.2. The large difference in within-group and between-group variability was not observable in the residuals for the test set for the model constructed from uniformly sampled data (right plot in Figure a.3), meaning that an important property of the model errors was missed when using uniformly sampled data. [That information can be used in model building, e.g., by designing future experiments so that the concentration is known exactly at the start of an experiment and used to correct the calibration model by a constant term, which would have the

effect of shifting the residuals in a group in the right plot in Figure a.3 to have zero mean error (as each group corresponding to a single experiment operating across a range in temperature).]

Pitfalls:

Did not attempt the question:	(−25 pts)
Did not compare group model to the model trained with uniform sampling	(−10 pts)
Did not compare biased predictions to unbiased predictions	(−5 pts)
Provided inadequate graphical or numerical support for conclusions	(−5 pts)

(b) For your split in part a, 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. Compare these prediction errors to the prediction errors obtained by these methods for the original dataset split in HW 1.

Solution (30 pts): As in part a, the biased C = 0.014 g/g and unbiased C = 0.03 g/g data are set aside for testing. The C = 0.025 g/g data are designated as the validation dataset because it is centrally located in the desired range, and the remaining data are used for training, as summarized in Table b.1.

Training Data	**C = 0.01, 0.02, 0.035**
Validation Data	**C = 0.025**
Testing Data	**C = 0.03**
Biased Testing Data	**C = 0.014**

Table b.1: Data classification for regularized regression modelling

Uniformly Randomly Sampled Data Results (HW1)			
	Ridge	Lasso	Elastic Net
Validation RMSE	1.7837×10^{-5}	2.3595×10^{-5}	2.0138×10^{-5}
Test RMSE	1.8430×10^{-5}	2.0893×10^{-5}	2.2216×10^{-5}

Grouped Data Results			
	Ridge	Lasso	Elastic Net
Validation RMSE	2.837×10^{-5}	2.086×10^{-5}	2.025×10^{-5}
Unbiased Test RMSE	2.247×10^{-5}	5.117×10^{-5}	4.829×10^{-5}
Biased Test RMSE	9.037×10^{-5}	3.203×10^{-4}	3.157×10^{-4}

Table b.2 and b.3: Validation and testing error for sampled and grouped regression

While predictive models based on spectral data generally perform better when input as raw absorbances, all of the above implementations used normalized inputs, for compatibility with the Matlab and SpaSM back-end implementations. Overall, the grouped-data models had higher test RMSE than for the model trained using uniformly sampled data; the latter would underestimate the RMSE that would be obtained if applied to data from a new experiment for the reasons discussed in part a, and the grouped-data models provide a more accurate estimate of the RMSE for such a dataset. Within each model-fitting method, the RMSE for the biased test data are much higher than for the unbiased test data, which provides additional evidence that the data identified in part a as being biased are truly biased.

The test residuals are visualized in Figure b.1 to gain a qualitative understanding of the model fit.

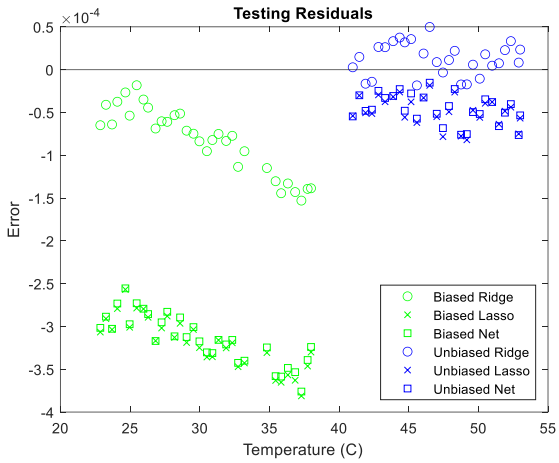


Figure b.1: Testing residuals for ridge regression, lasso, and elastic net on biased and unbiased data groups

It is interesting to note the non-horizontal slope of the biased residuals when plotted as a function of temperature, which could be related to the much sharper temperature vs. absorbance shift in the biased data compared to the other data groupings. Also, note that the residuals for ridge regression, lasso, and elastic net for the unbiased data are much closer to "zero-mean with constant variance" than for OLS in part a. As such, these methods have residuals that much more closely satisfy the assumptions made on the noise in the data.

Figure b.2 plots the coefficient fits for the new group-trained data compared to the models trained from uniformly sampled data in HW1. These are not required to receive full credit.

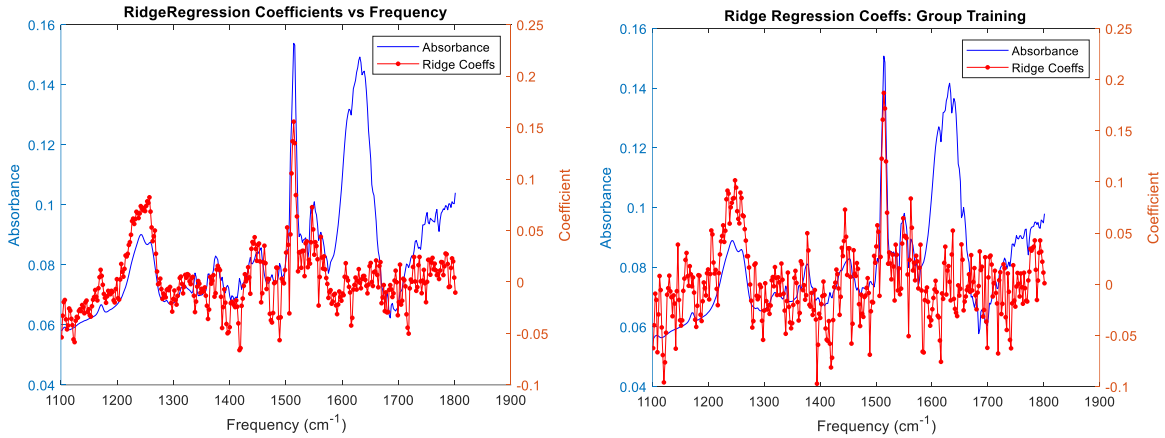


Figure b.2: Ridge regression coefficients compared to the sample absorbance as a function of frequency (the left plot is for the model constructed from uniformed randomly sampled data, and the right plot is for the model constructed from grouped data).

The ridge regression model trained with the grouped data displays more noise in the coefficient fits, with larger oscillations in coefficients in regions not corresponding to specific absorbance peaks than the model trained with uniformly sampled data. The ridge regression model more heavily weighs the peaks 1250 cm^{-1} and around 1520 cm^{-1}. As discussed in the HW1 solution, these two peaks are associated with vibrations in the molecule, whereas the peak at around 1250

cm^{-1} is associated with water. By removing the biased data from the training dataset, the ridge regression model was better able to identify the key peaks to weigh heavily in the regression model.

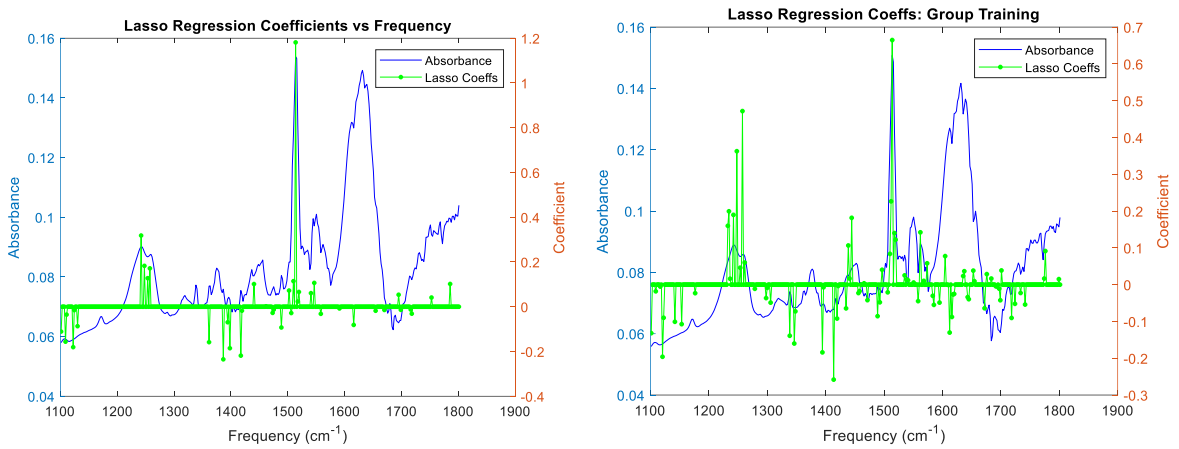


Figure b.3: Lasso regression coefficients compared to the sample absorbance as a function of frequency (the left plot is for the model constructed from uniformed randomly sampled data, and the right plot is for the model constructed from grouped data).

Examining the lasso coefficients, we observe more nonzero coefficients in the model trained with grouped data, as well as much larger coefficients in the region of the first main spectral peak at about 1250 cm^{-1} and more coefficients in the peak at around 1520 cm^{-1}. By removing the biased data from the training dataset, the lasso model was better able to identify the key peaks to weigh heavily in the regression model, and the largest regression coefficient are associated with the important peaks in the spectrum.

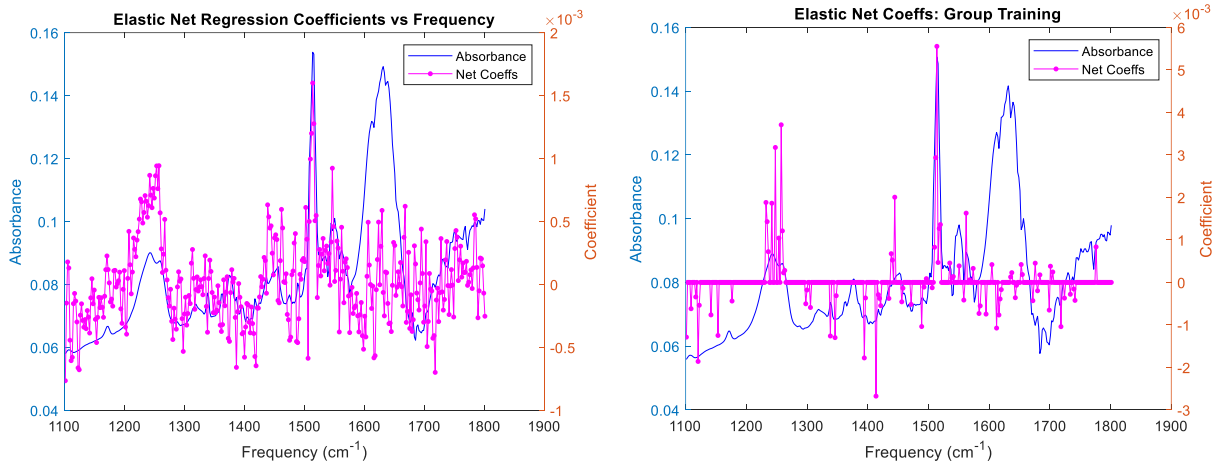


Figure b.4: Elastic net regression coefficients compared to the sample absorbance as a function of frequency (the left plot is for the model constructed from uniformed randomly sampled data, and the right plot is for the model constructed from grouped data).

It is interesting to note that, while the elastic net model trained with uniformly sampled data more closely resembles the ridge regression model, the elastic net model trained with grouped data is nearly identical to the lasso model, which both have the largest regression coefficients associated with the two important peaks in the spectrum.

Pitfalls:

Did not attempt the question:	(−30 pts)
Assigned training and testing data without grouping	(−20 pts)
Provided no explanation for their choice of grouping	(−5 pts)
Did not provide graphical analysis (residuals or predicted vs actual)	(−5 pts)
Did not provide analysis of root mean squared errors	(−5 pts)

(c) Repeat the model constructions in parts a and b but instead use a high-quality nested cross-validation procedure which involves multiple choices of splitting the data. Analyze the model stability and discuss whether the model stability is high enough that you are confident in the quality of your models. For this application, how much additional value was there in carrying out the nested cross-validation procedure compared to using the single well-chosen split of the dataset in parts a and b.

Solution (50 pts): Cross validation was performed by designating each of the different non-biased groupings as the testing dataset and then repeating the training and testing procedure used in part a. Using simple regression with a subset of frequencies and cross-validated grouping yields the Table c.1.

	Randomly Sampled Data	Test $C = 0.01$	Test $C = 0.02$	Test $C = 0.025$	Test $C = 0.03$	Test $C = 0.035$
Training RMSE	2.9734×10^{-4}	1.24×10^{-4}	8.48×10^{-5}	9.76×10^{-5}	1.62×10^{-4}	1.61×10^{-4}
Unbiased Test RMSE	4.4875×10^{-4}	4.0×10^{-4}	4.0×10^{-4}	7.78×10^{-4}	1.025×10^{-3}	7.07×10^{-4}
Biased Test RMSE	–	1.0×10^{-3}	1.2×10^{-3}	1.5×10^{-3}	1.8×10^{-3}	1.3×10^{-3}

Table c.1: Training and testing error for the uniformly sampled data, compared to a cross-validation of the group regression model.

The training RMSE is lower for all models constructed from the data selected based on groups, regardless of which group was assigned for unbiased test set, than for the model constructed from uniformly sampled data (cf. elements in the first row of Table c.1). This result is consistent with the fact that biased data were included in the training RMSE for the model constructed from the uniformly sampled data, and the biased data are more poorly fit by a linear model (Figure a.2). Among the models constructed from the grouped datasets, selecting the group of $C = 0.03$ g/g as the unbiased test dataset, as in part a, gave the highest RMSE for both the unbiased and biased test datasets (cf. elements in the second and third rows of Table c.2). The RMSE for the unbiased test datasets are highly variable with respect to which group is used as the unbiased test dataset, indicating model stability that is poorer than desired. The lowish model stability is consistent with most of the variability being between different groups, and the number of groups in the training dataset is low (only four), which is not enough groups to be able to obtain a stable estimate for the average prediction error. We can be reasonably confident that the average predictive error should be in the range of 4×10^{-4} to 10^{-3} for unbiased data. Among the models constructed from grouped datasets, the RMSE for the biased test dataset is consistently larger than the RMSE for the training data and the unbiased test dataset, which is consistent with the $C = 0.014$ g/g dataset being biased (cf. elements in the third row of Table c.2 with its first and second rows). Among the models constructed from grouped datasets, the RMSEs for the biased test datasets are more consistent, on average approximately double the RMSE for the unbiased test datasets. Before doing the analysis, it would normally be expected that predictive performance for test concentrations of $C = 0.01$ g/g

and C = 0.035 g/g would be significantly worse than the others because fitting to those data would require the model to extrapolate from the training data; interestingly, no such trend is observed. A likely explanation for the lack of higher error during extrapolation is that the variability between groups is much larger than the magnitude of any deviation from linearity that would occur from extrapolating beyond the training dataset. Figure c.1 is a plot for examining the residuals to obtain a better understanding of the fit quality.

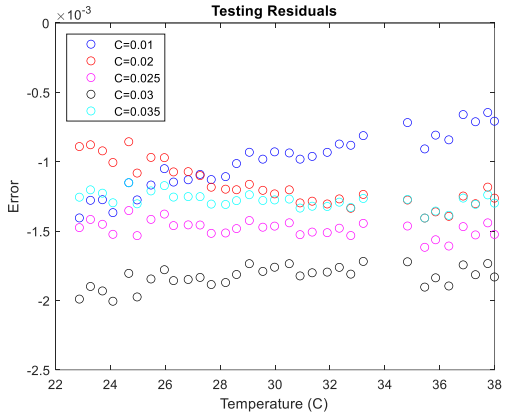


Figure c.1: Regression test residuals for each cross-validation case.

Regardless of the grouping selected in the cross validation, the regression model consistently overestimates the testing concentration, as all residuals are negative. As seen earlier, most of the residuals are associated with a constant shift in the concentration prediction for each group, rather than fluctuations within each group. Performing a similar cross validation on lasso, ridge regression, and elastic net yields the results in Tables c.2 to c.5.

	Val:C=0.01	Val:C=0.02	Val:C=0.025	Val:C=0.03	Val:C=0.035
Test: C=0.01		5.55×10^{-5}	2.54×10^{-5}	1.42×10^{-5}	1.65×10^{-5}
		1.15×10^{-3}	8.34×10^{-4}	7.76×10^{-4}	1.00×10^{-3}
		1.20×10^{-3}	7.3×10^{-4}	8.2×10^{-4}	9.57×10^{-4}
Test: C=0.02	9.997×10^{-4}		3.17×10^{-5}	1.83×10^{-5}	2.44×10^{-5}
	9.51×10^{-5}		1.66×10^{-4}	5.96×10^{-5}	6.28×10^{-5}
	1.06×10^{-3}		1.07×10^{-4}	9.54×10^{-5}	1.54×10^{-4}
Test: C=0.025	8.34×10^{-4}	9.03×10^{-5}		2.06×10^{-5}	4.56×10^{-5}
	2.54×10^{-5}	4.42×10^{-5}		3.05×10^{-5}	4.51×10^{-5}
	7.30×10^{-4}	1.34×10^{-4}		1.11×10^{-4}	1.11×10^{-4}
Test: C=0.03	4.96×10^{-4}	1.74×10^{-5}	2.83×10^{-5}		9.02×10^{-4}
	4.85×10^{-5}	2.34×10^{-5}	2.19×10^{-5}		2.92×10^{-4}
	7.46×10^{-4}	1.79×10^{-4}	9.4×10^{-5}		1.67×10^{-4}
Test: C=0.035	9.89×10^{-4}	2.04×10^{-5}	4.37×10^{-5}	2.92×10^{-4}	
	3.05×10^{-5}	$4.02\times\times10^{-5}$	1.55×10^{-4}	9.08×10^{-4}	
	9.96×10^{-4}	5.58×10^{-5}	5.73×10^{-5}	1.76×10^{-4}	

Table c.2: Ridge regression validation and unbiased testing root-mean-squared errors in units of g/g. The values in yellow are validation errors, in blue are unbiased test errors, and in red are biased test errors.

	Val:C=0.01	Val:C=0.02	Val:C=0.025	Val:C=0.03	Val:C=0.035
Test: C=0.01		1.23×10^{-4}	1.89×10^{-5}	5.92×10^{-5}	2.87×10^{-4}
		1.49×10^{-3}	8.95×10^{-4}	1.32×10^{-3}	9.61×10^{-4}
		1.52×10^{-3}	8.72×10^{-4}	1.56×10^{-3}	1.41×10^{-3}
Test: C=0.02	8.52×10^{-4}		7.11×10^{-5}	6.39×10^{-5}	9.29×10^{-4}
	1.32×10^{-4}		1.24×10^{-4}	1.09×10^{-4}	1.78×10^{-4}
	9.77×10^{-4}		2.76×10^{-4}	3.52×10^{-4}	5.96×10^{-4}
Test: C=0.025	8.90×10^{-4}	1.20×10^{-4}		1.79×10^{-5}	6.49×10^{-5}
	1.91×10^{-5}	7.31×10^{-5}		4.77×10^{-5}	2.73×10^{-5}
	8.75×10^{-4}	2.77×10^{-4}		2.39×10^{-4}	2.94×10^{-4}
Test: C=0.03	4.14×10^{-4}	3.27×10^{-5}	2.09×10^{-5}		4.1×10^{-5}
	7.35×10^{-5}	6.39×10^{-5}	5.12×10^{-5}		3.83×10^{-5}
	7.86×10^{-4}	3.26×10^{-4}	3.20×10^{-4}		2.74×10^{-4}
Test: C=0.035	2.8×10^{-4}	1.43×10^{-5}	1.38×10^{-5}	2.3×10^{-5}	
	9.04×10^{-5}	1.277×10^{-4}	1.20×10^{-4}	1.14×10^{-4}	
	8.07×10^{-4}	3.71×10^{-4}	3.69×10^{-4}	3.18×10^{-4}	

Table c.3: Lasso validation and unbiased testing root mean squared errors in units of g/g. The values in yellow are validation errors, in blue are unbiased test errors, and in red are biased testing errors.

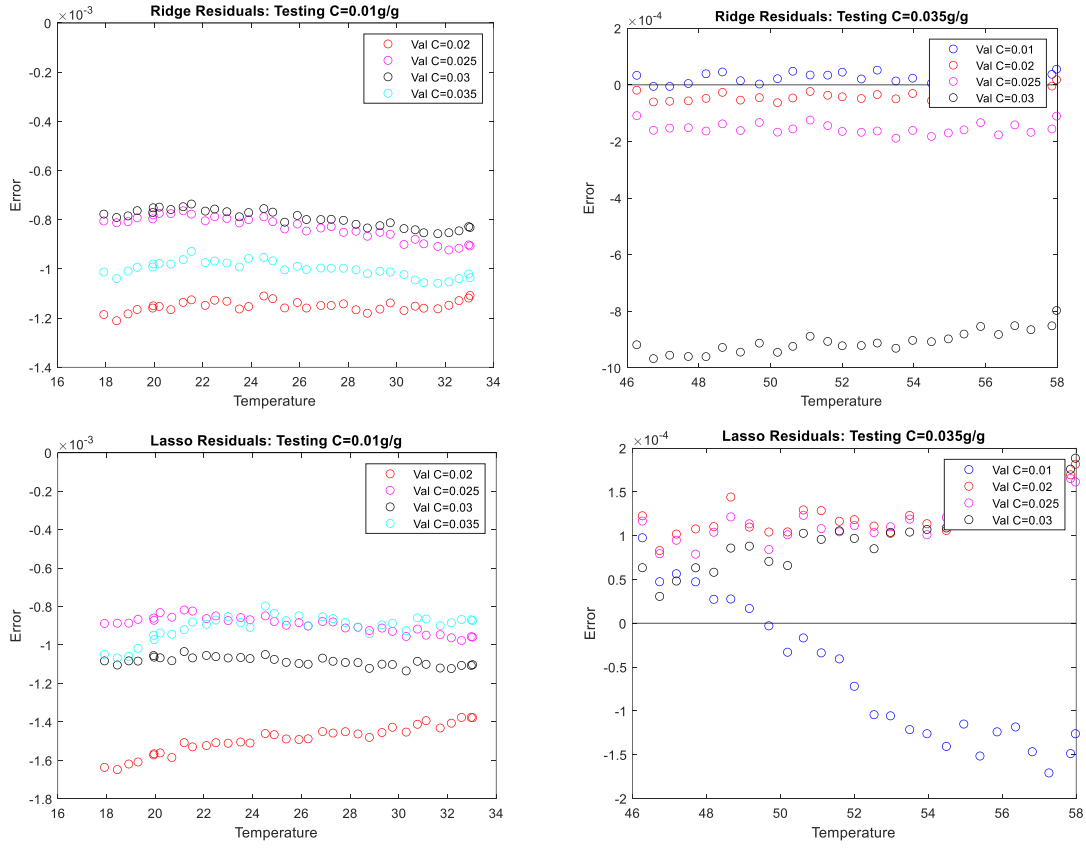
	Val:C=0.01	Val:C=0.02	Val:C=0.025	Val:C=0.03	Val:C=0.035
Test: C=0.01		1.23×10^{-4}	1.89×10^{-5}	5.92×10^{-5}	2.87×10^{-4}
		1.18×10^{-3}	9.42×10^{-4}	1.04×10^{-3}	9.93×10^{-4}
		1.22×10^{-3}	9.13×10^{-4}	1.02×10^{-3}	9.70×10^{-4}
Test: C=0.02	8.52×10^{-4}		7.11×10^{-5}	6.39×10^{-5}	9.29×10^{-5}
	1.32×10^{-4}		9.71×10^{-5}	2.06×10^{-5}	3.98×10^{-4}
	9.83×10^{-4}		1.36×10^{-4}	1.93×10^{-4}	1.87×10^{-4}
Test:C=0.025	8.90×10^{-4}	1.20×10^{-4}		1.79×10^{-5}	6.49×10^{-5}
	2.54×10^{-4}	2.36×10^{-4}		4.67×10^{-5}	2.28×10^{-5}
	7.30×10^{-4}	6.01×10^{-4}		2.35×10^{-4}	2.63×10^{-4}
Test: C=0.03	4.14×10^{-4}	3.27×10^{-5}	2.09×10^{-5}		4.10×10^{-5}
	5.98×10^{-5}	4.47×10^{-5}	4.83×10^{-5}		4.40×10^{-5}
	7.05×10^{-4}	2.54×10^{-4}	3.16×10^{-4}		2.84×10^{-4}
Test:C=0.035	2.80×10^{-4}	1.43×10^{-5}	1.38×10^{-5}	2.33×10^{-5}	
	1.14×10^{-4}	9.75×10^{-5}	1.22×10^{-4}	1.14×10^{-4}	
	6.61×10^{-4}	2.77×10^{-4}	3.71×10^{-4}	3.18×10^{-4}	

Table c.4: Elastic net validation and unbiased testing root mean squared errors in units of g/g. The values in yellow are validation errors, in blue are unbiased test errors, and in red are biased testing errors.

	Mean Ridge Regression RMSE	Mean Lasso RMSE	Mean Elastic Net RMSE
Test: C = 0.01	9.46×10^{-4}	1.12×10^{-3}	1.04×10^{-3}
Test: C = 0.02	8.35×10^{-5}	1.36×10^{-4}	7.23×10^{-5}
Test: C = 0.025	3.63×10^{-5}	4.18×10^{-5}	8.27×10^{-5}
Test: C = 0.03	9.65×10^{-5}	5.67×10^{-5}	3.92×10^{-5}
Test: C = 0.035	2.84×10^{-4}	1.13×10^{-4}	1.12×10^{-4}

Table c.5: Average unbiased testing RMSEs for each choice of testing data in the cross-validation.

From these RMSE values, several trends can be observed. For all three fitting techniques, the largest testing RMSEs had magnitudes on the order of 8×10^{-4} to 1×10^{-3} when testing on the C = 0.01 g/g dataset. While not as far removed from the other errors, the C = 0.035 g/g testing data also features a notable uptick compared to the models validated using the C = 0.025 g/g and C = 0.03 g/g data groups. These trends suggest a noticeable improvement in predictive power when validating using runs close to the center of the overall concentration range than validating with edge cases; because we removed the biased C = 0.014 g/g data, the C = 0.01 g/g data are much further removed from the training set when used as testing data than the C = 0.035 g/g data are, explaining the relative magnitudes of the resulting errors. Another trend is that the mean RMSE for the elastic net is the lowest of the three methods except when C = 0.025 is the test case, in which case ridge regression gave the lowest mean RMSE. The mean RMSE for lasso was worse than elastic net for all five cases, but are very close to elastic net for two of the cases.



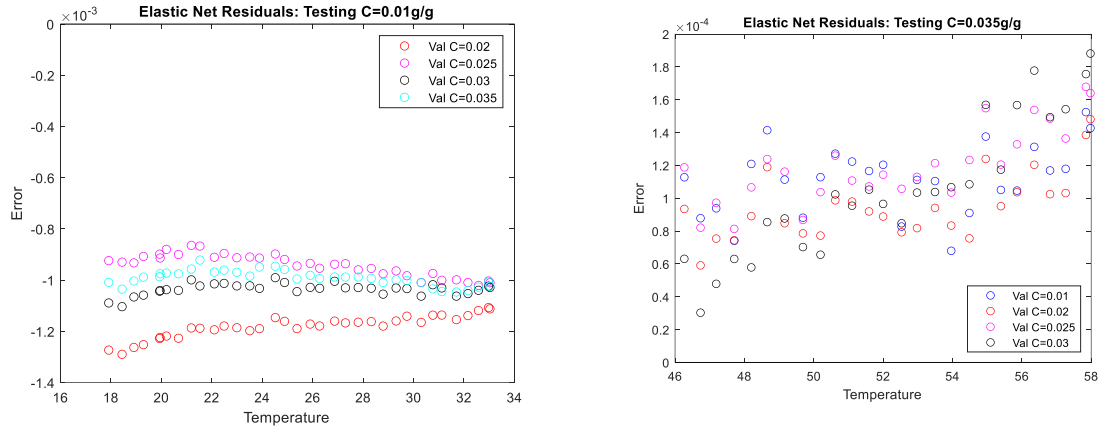


Figure c.2: Residual plots for ridge regression, lasso, and elastic net for testing data with C = 0.01 g/g and C = 0.035 g/g.

All three fitting techniques result in overestimates of the C = 0.01 dataset; while lasso and elastic net underestimated the C = 0.035 data, ridge regression overestimated the concentrations for all validation sets except C = 0.01 g/g. Ridge regression gave the higher RMSE when the closest dataset was chosen for validation rather than training, indicating that proximity to the training data gave a lower RMSE than proximity to the validation data when extrapolating from such a model.

Averaging instead over the choice of validation data gives the RMSE values in Table c.5.

	Mean Ridge Regression RMSE	Mean Lasso RMSE	Mean Elastic Net RMSE
Val: C = 0.01	4.99×10^{-5}	8.5×10^{-5}	8.27×10^{-5}
Val: C = 0.02	3.15×10^{-4}	4.40×10^{-4}	3.90×10^{-4}
Val: C = 0.025	2.83×10^{-4}	2.97×10^{-4}	3.03×10^{-4}
Val: C = 0.03	4.49×10^{-4}	3.39×10^{-4}	3.04×10^{-4}
Val: C = 0.035	3.50×10^{-4}	2.88×10^{-4}	2.75×10^{-4}

Table c.5: Average unbiased testing RMSEs for each choice of validation data in the cross-validation. All entries are in concentration units of g/g.

The significantly lower RMSE values for data validated using C = 0.01 g/g can be explained by the fact that this dataset was therefore never used for testing, and the C = 0.01 g/g data were shown to be predicted significantly less accurately than the other data values. Aside from this dataset, the other RMSE values are all approximately 3×10^{-4}, suggesting that our choice of validation dataset did not significantly impact the model's predictive power.

Pitfalls

Did not attempt the question (-50 pts)
Provided no quantitative metrics of relative model performance in the cross validation (e.g. RMSE, residuals). These can be tabulated or graphical. (-25 pts)
Included biased data in the training dataset (-5 pts)
Reported values where testing and validation were performed on the same group (–5 pts)

Did not draw conclusions on the effect of validation or testing data selection on the model performance (-10 pts, 5 for each missing)