What They forgot to Tell You about Creating Predictive Models with an Application to Pharmaceutical Manufacturing

Kjell Johnson & Max Kuhn

Background

TO DOs:

- Add information in the introduction and throughout manuscript to provide the setting for "What they forgot to tell you"
- Add conclusion
- Put in Journal format

The manufacturing process of a biological drug is a complex process that requires careful monitoring to ensure that the cells are efficiently creating the drug product. This process can be very challenging to systematically control since the incubation process can take many days and cells are complex biological entities that are affected by slight changes to environmental conditions. To ensure that the bioreactor conditions are conducive to the cells producing product, key attributes are measured by sampling the contents of the bioreactor daily. If attributes are not in an acceptable range, then steps must be taken to alter the conditions of the bioreactor. Generally, the sooner the conditions can be adjusted, the better the quantity and quality of the final drug product. Measuring the attributes takes time. Therefore, there is usually a lag between the attribute measurements and corresponding adjustment. This lag can lead to less and lower quality product.

Raman spectroscopy is a tool that can measure chemical characteristics (i.e. a chemical fingerprint) of samples in real time (Jesus, Löbenberg, and Bou-Chacra 2020; Esmonde-White, Cuellar, and Lewis 2022; Silge et al. 2022). Using the spectra in a predictive model of the characteristics of interest would enable real-time knowledge of and corresponding adjustments to the bioreactor, thus generating higher quality, larger volume drug product.

In the example outlined in this tutorial, several key input parameters were varied systematically across their operating ranges within each of 60 small-scale bioreactors for producing a biological drug. At seven days after the start of the experiment, a sample was collected and was analyzed by Raman spectroscopy. The concentration of the drug product in the sample was also measured. The goal in this analysis is to understand how predictive Raman spectra can be of the drug product concentration. If there is a relationship, then the model could be used to signal if the bioreactor was insufficiently producing product and prompting remedial steps to increase production.

In this tutorial, we will discuss the process of constructing a predictive model. This process starts with understanding the predictors and the response of the available data. After this initial understanding, we must then determine how to spend the data for the model building process. Specifically, some data will need to be used to learn the generalizable characteristics that relate the predictors with the response (i.e. the training set). And other data will need to be used to assess how well the model predicts new data (i.e. the test set). After splitting the data, the predictors and/or the response may need to be preprocessed prior to modeling to better enable models to extract the predictive signal. After preprocessing, we can determine which types of predictive models to build. Each model has one or more parameters that determine how predictors are related to the response. In general, we do not know a priori which values of the tuning parameters are best. Therefore, we search a range of values to identify an optimal value. After identifying an optimal model, this model is then evaluated on the test data to determine if the model can be trusted to reliably predict new, yet-to-be-seen samples.

Understanding the Data

The first step in any modeling process is to understand the available data. In this application, there is one sample from each of 60 bioreactors. Raman spectroscopy has been applied to each sample, and the drug product concentration has been measured. Figure 1 displays the original Raman spectra. From this figure, we can see that there is an initial downward trend towards the middle of the wavenumbers, then an upward trend towards the higher wavenumbers. The intensities are not randomly scattered. Instead, there is a relationship across wavenumbers with intensity. This relationship indicates that wavenumber intensities are correlated with each other. In fact, the correlation between the majority of adjacent wavenumbers is greater than 0.99.

To illustrate this more clearly, let's examine the relationship among wavenumber measurements for the first sample. For the first sample, the first 3000 lags are created. To create a lag, the data is shifted by a specified number of rows to create a new variable. For example, to create the first lag, the wavenumber measurements are shifted over by one wavenumber. To create the second lag, the measurements are shifed by two wavenumberes, and so on. Figure 2 illustrates the correlation between each subquent lag for the first 1000 lags. Clearly, close

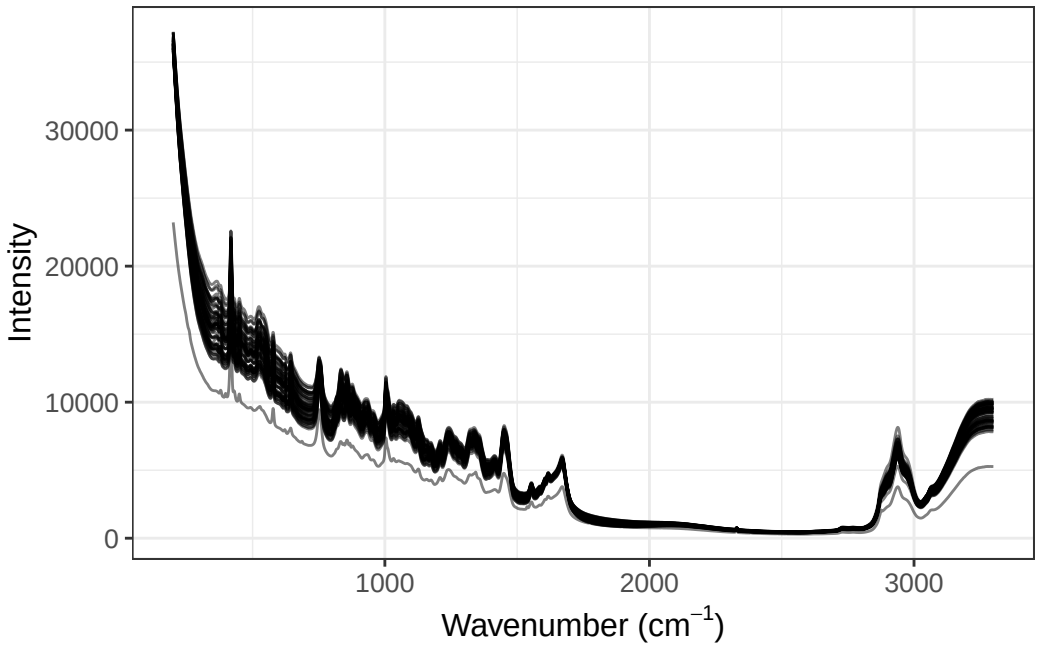

Figure 1: Raman spectra profiles for each of the 60 samples.

wavenumbers have high correlation, whereas far wavenumbers have low correlation. As we will see, understanding this characteristic will turn out to be very important when making decisions about how to pre-process the data prior to modeling and which models to train.

In addition to understanding the predictors, we should also understand characteristics of the response. Examining the response distribution can help determine if a transformation may be necessary or if there are samples that are unusual with respect to the majority of the data. Figure 3 presents the histogram drug product concentration across the samples. For this data, the distribution is approximately symmetric and has a range of 85 to 115. Based on this figure, a transformation does not appear to be necessary, and there are no samples that are unusual.

Data Splting

The primary objective of predictive modeling is to use the existing data to develop a model that predicts new samples as accurately as possible. To achieve this objective, a process must be implemented that avoids overfitting to the existing data (Kuhn, Johnson, et al. 2013; Hawkins 2004). An overfit model is one that accurately predicts the response for the data on which the model was trained, but does not accurately predict new data. To avoid overfitting, we must construct a model building process that mimics the prediction process for new samples. One

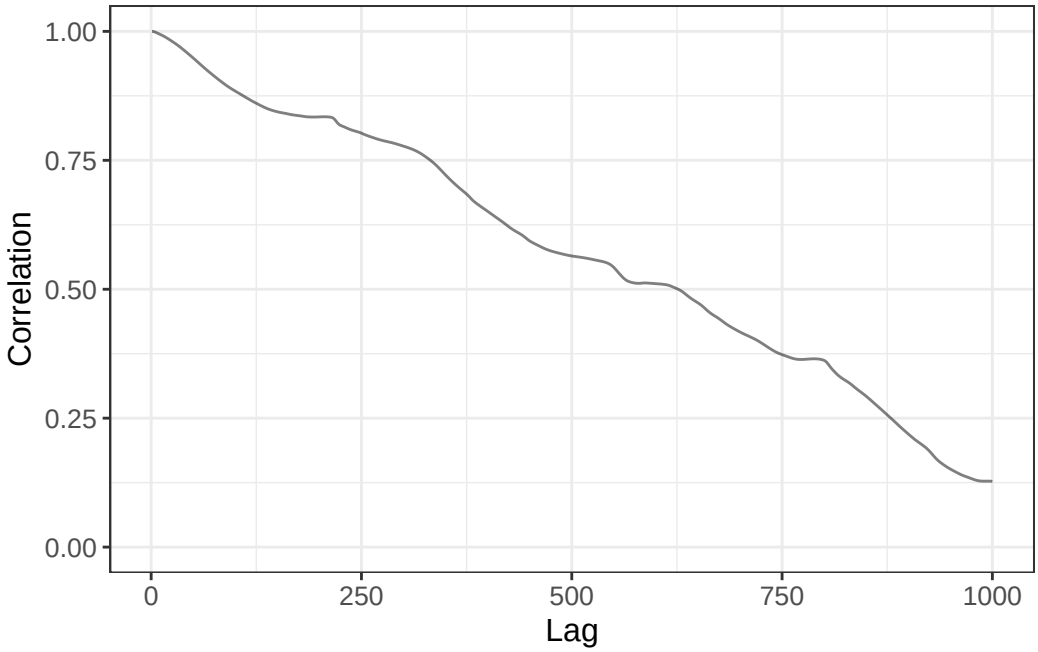

Figure 2: The correlation between the original intensities and lagged intensities for the first sample. As wavenumbers depart, the correlation of the intensities decreases.

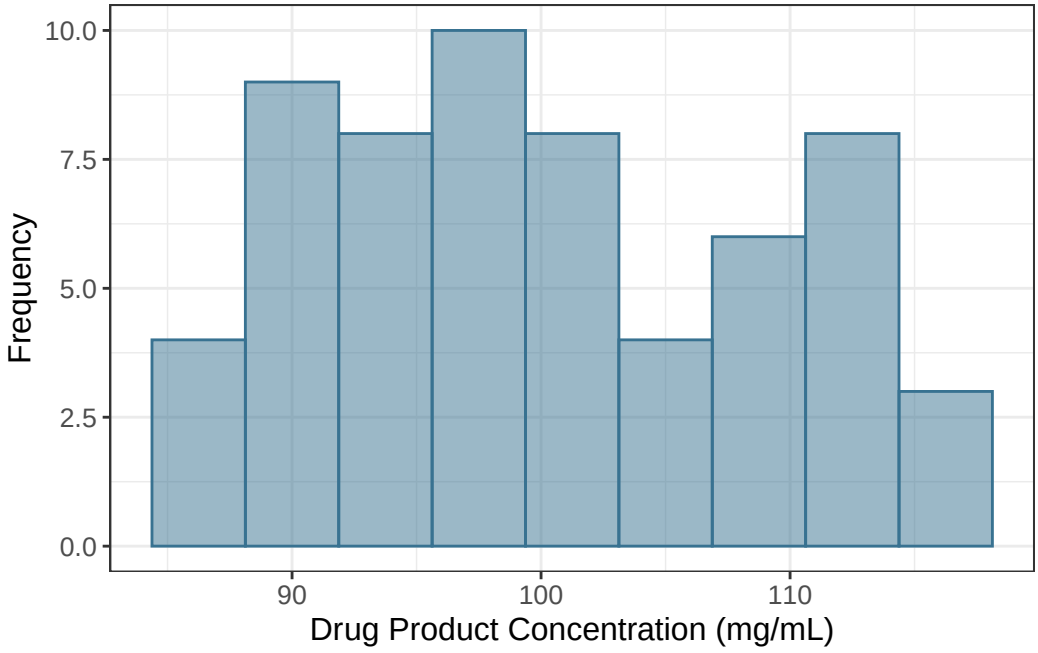

Figure 3: The distribution of drug product concentration across samples.

way to do this would be to split the data into a training set and a test set as illustrated in Figure 4. A model could be constructed with the training set, then predictive performance could be evaluated with the test set. However, most predictive models must be constructed using a variety of tuning parameter values. The test set would then need to be evaluated multiple times to assess predictive performance. When the test set is evaluated multiple times, we are essentially finding a model that fits the test set. This process leads to overfitting, and the model performance cannot be trusted as an accurate evaluation of the predictive performance on new samples. Therefore, a single training/test split will not be adequate for building predictive models. Moreover, it is important to understand that the test set should only be used once to evaluate the final selected models.

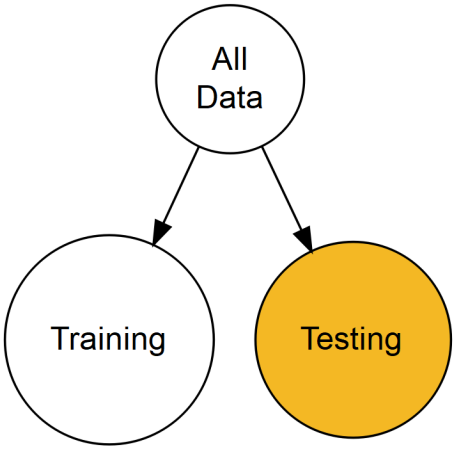

Figure 4: Diagram of an intitial training and testing split.

Instead of a single training/test split, we need a process that can be used to evaluate many tuning parameter values for each of many different models. Figure Figure 5 illustrates a two-layered process that incorporates the use of resampling. In the first layer, the entire data set is split into a training and test set. In general, anywhere between 50% to 80% of the data is randomly selected for the training data, while the remaining data is placed in the test set. A random split may be adequate. However, we may desire that the data in the training set and testing split have similar characteristics. For example, it may be advantageous for the training and test sets to have a similar distribution of the response. If the response distribution is skewed, then it would be important that the training and test sets reflect the entirety of the distribution. Likewise, if there are characteristics or covariates in the data that should be proportional represented, then the data should be split into the training and testing set using a stratified random approach.

In the second layer of Figure Figure 5, the training data is split using resampling. Cross-validation could be used in this layer, where the data is split into k-folds. For example, if 10-fold cross-validation was used in this layer, then the training data would be partitioned into 10 folds. The analysis set for the first resample would contain 9 folds of the data, while

the assessment set would contain 1 fold of the data. A model would be constructed using the 9 folds and would evaluated using the hold-out fold. To create the analysis set for the second resample, a different combination of 9-folds would be used to construct the model. The model would then be evaluated on the fold that was not used in the modeling. Figure 6 provides an illustration of 10-fold cross-validation.

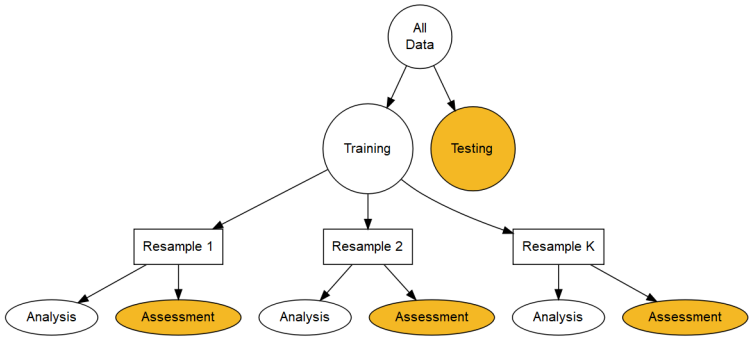

Figure 5: Illustration of data splitting for cross-validation.

In addition to k-fold cross-validation, there are quite a few different resampling approaches for the second layer of the data splitting process. These include group, repeated k-fold, bootstrapping, Monte Carlo, and cluster-based, to name a few (Kuhn and Silge 2022; Kohavi 1995). Determining which approach to take depends on factors such as the structure and size of the data, and the models that will be trained. When constructing initial models, it may be sufficient to use a less computationally intensive method like k-fold cross-validation to gain some level of understanding about the potential signal in the data and the range of optimal tuning parameter values. However, more computationally intensive methods like repeated k-fold, bootstrapping, or Monte Carlo often yield a more accurate estimate of a model's predictive ability and help narrow down the tuning parameter estimates.

For the example presented here, a stratified random approach will be used to split the data into a training (75%) and a test (25%) set. The distribution of the response will be used as the stratification variable such that an equal proportion of samples will be randomly selected within each quartile of the distribution. When training models, we will compare the performance of 10-fold cross-validation as well as repeated 10-fold cross-validation.

Pre-processing

The predictors and response, in their original form, are usually not in the best form for enabling models to find an optimal predictive relationship. The original data may contain highly correlated predictors, predictors that lack information, missing values, multi-category predictors, or highly skewed predictors. Some models, such as those based on recursive partitioning

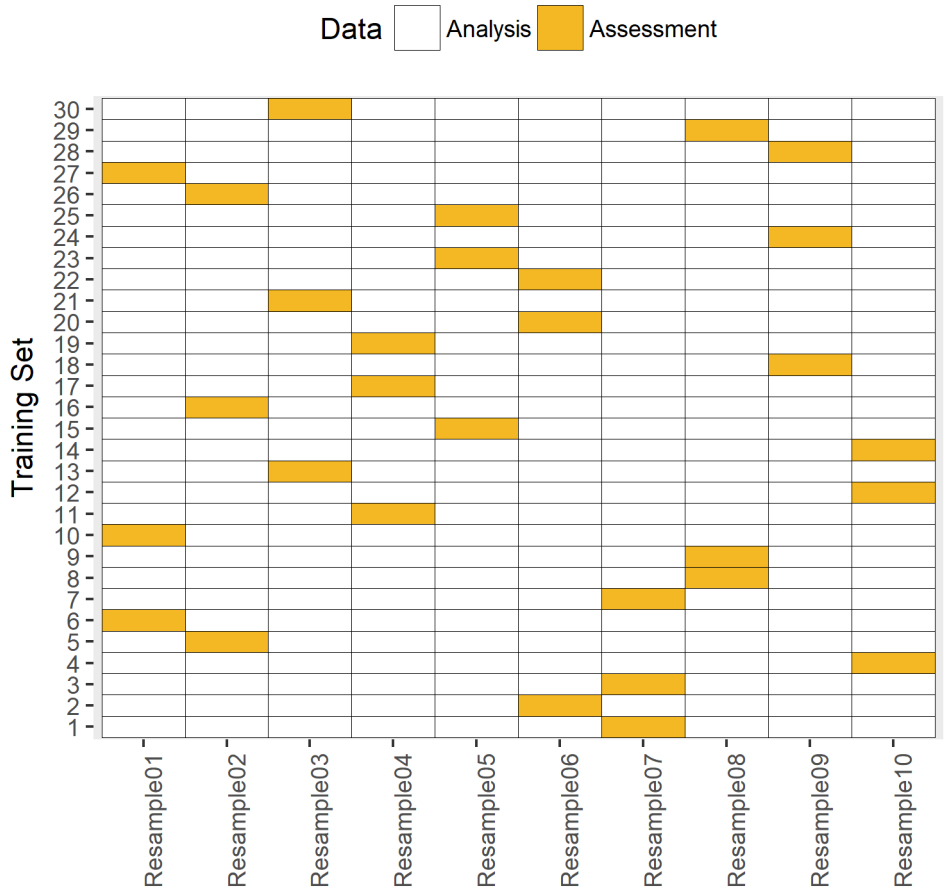

Figure 6: Diagram of an intitial training and testing split.

algorithms, can handle most of these challenging characteristics. However, many models either cannot be built or the predictive performance will be detrimentally impacted when one or more of these characteristics are present. As a simple example, consider a predictor that has three categories such as Low, Mid, and High. The information, in this form, cannot be ingested by most models. Instead, the information needs to be converted into either an ordinal-scaled predictor, or two binary variables. Missing data also wreaks havoc on predictive models, because the models require non-missing information. Therefore, appropriate pre-processing steps must be taken prior to the model training process.

A common problem across most data sets is that predictors that lack information, which is characterized by variability. Figure 7 illustrates the distribution of the standard deviation of intensity measurements across wavenumbers. For this data, all wavenumbers have positive standard deviation. Therefore, there is no reason to omit any wavenumbers from the data set prior to modeling.

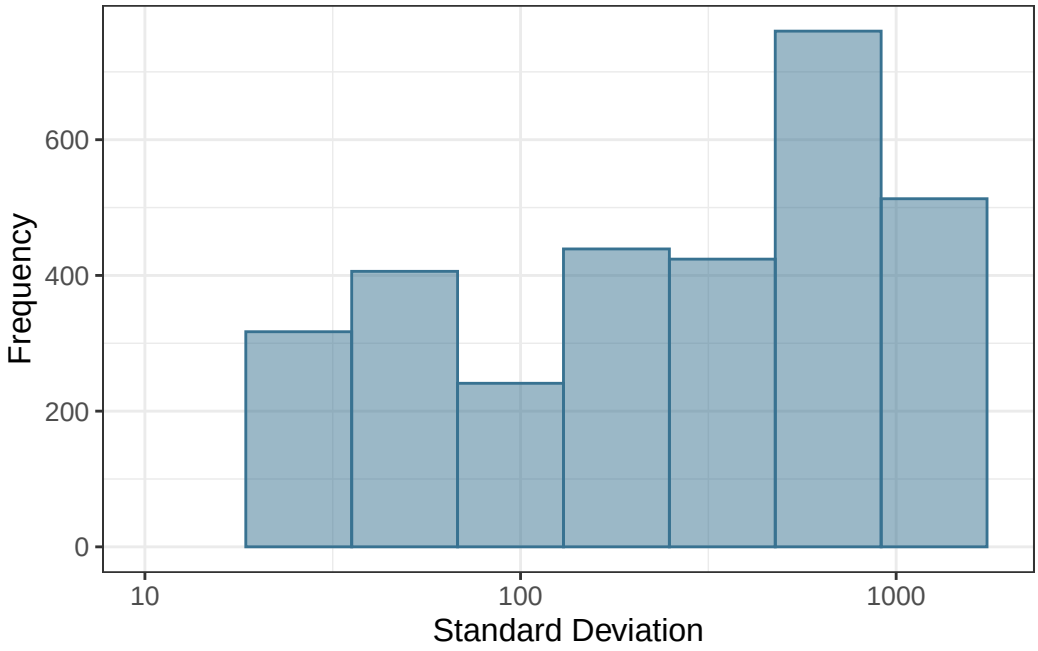

Figure 7: The distribution standard deviation of intensity measurements across wavenumbers.

A second common problem across data sets is that predictors contain redundant information. We saw earlier in Figure 1 that the intensity measurements were highly correlated across wavenumbers. We will therefore need to implement pre-processing to address this characteristic. Spectral data also provide some unique pre-processing challenges. Recall from Figure 1 that the intensity values across samples have an initial downward trend towards about wavenumber 2500, then begin to trend upward. In spectroscopy data, deviations in intensity from zero are commonly referred to as baseline drift, typically stemming from factors

such as measurement system noise, interference, or fluorescence (Rinnan, Van Den Berg, and Engelsen 2009). Importantly, these deviations are not indicative of the chemical composition of the sample itself.

Baseline drift is a notable source of measurement variability, where the vertical variability surpasses that associated with spectral peaks. The excess variability, originating from extraneous sources contributing to the background, can detrimentally affect models reliant on predictor variability, such as principal component regression and partial least squares.

It would be ideal if all background could be completely removed. A value of zero intensity for a wavenumber would theoretically mean that no molecules were present that respond to that specific wavenumber. Although measures can be implemented to mitigate interference, fluorescence, and noise, it remains exceedingly challenging to completely eliminate background through experimental means. Therefore, the background patterns must be approximated and this approximation must be removed from the observed intensities. Therefore, the approximate the background patterns need to be approximated and subsequently subtracted from the observed intensities.

A polynomial smoother (Cleveland and Devlin 1988; Luers and Wenning 1971) is one tool that can be used to approximate the background. Figure 8 illustrates the original spectra for the first sample, the background as modeled by a polynomial smoother, and the corrected spectra. Notice that the corrected spectra is now more anchored with intensities at or near zero.

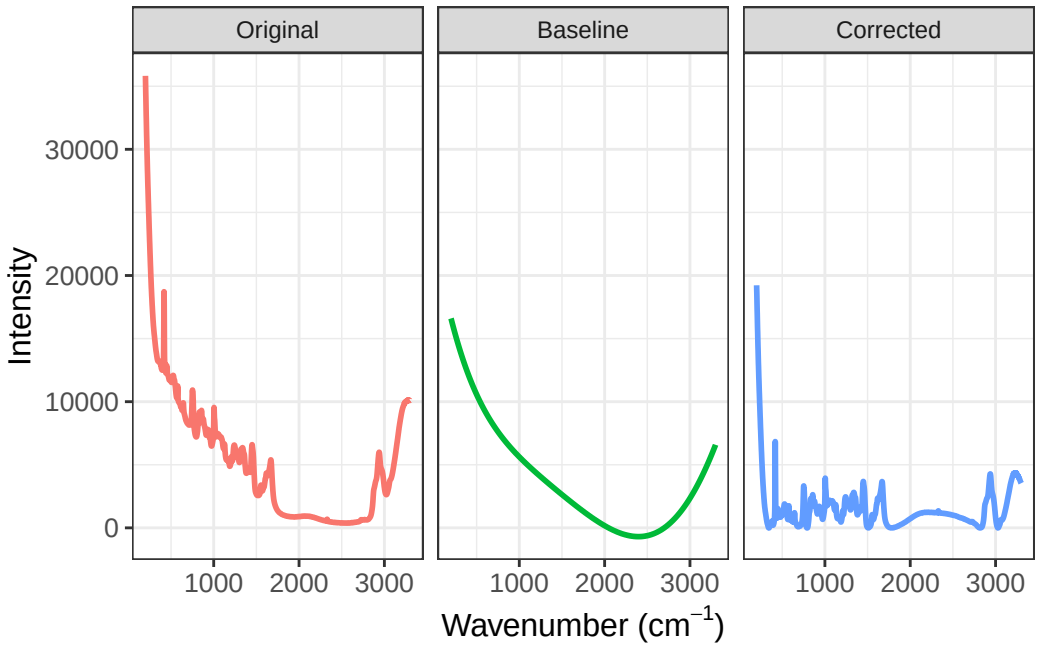

Figure 8: The distribution standard deviation of intensity measurements across wavenumbers.

A second source of noise is apparent in the variation of the intensity measurements across wavelengths within a spectrum. This is illustrated by the jagged profile illustrated in the "Original" and "Corrected" panels of Figure Figure 8. Smoothing splines and moving averages are two commonly used tools for reducing this type of noise. The moving average is computed at each point by averaging a specified number of values about that point. For example, the moving average of size 10 would replace each point by the average of the ten points before and after the selected point. The original curve becomes smoother as the number of points averaged together becomes larger. Therefore we need to be careful with the number of points chosen for the smoothing process. Too few points may not remove enough noise. While too many points may remove important signal.

The Savitzky-Golay procedure (Savitzky and Golay 1964; Stevens and Ramirez-Lopez 2022) is designed to remove spurious signal by simultaneously smoothing the data while also centering the overall signal and dampening variability. The procedure is governed by the order of differentiation, degree of polynomial, and window size. Figure 9 compares the impact of this procedure for differentiation order of 1 or 2, polynomial order of 2, and a small (15) or large (49) window size.

Figure 10 displays the correlation across the first 1000 wavenumbers for the original data as well as the each of the selected Savitzky-Golay transformations. The effect of differentiation and window size on the correlation across the transformed intensities is clear. When comparing first order differentiation to second order differentiation, second order differentiation more rapidly reduces correlation among close wavenumbers up to about the nearest 100 wavenumbers. Increasing the smoothing window also helps smooth the correlation profiles, but does not further reduce correlation. We will examine the impact of each of these different smoothing parameter selections on the model performance in following sections.

Modeling

Over the past half century, the number and types of models for relating a set of predictors to a response has rapidly grown. Improvements in computational power as well as mathematical complexity have been the primary drivers of this increase. Model complexity is generally tied to the number of parameters of a model. That is, as the number of model parameters increases, the ability of a model to adapt and morph to the relationship between predictors and the response also increases. For example, partial least squares has one tuning parameter and is effective at finding a linear relationship between predictors and the response. However, this method is ineffective at finding non-linear relationships. In constrast, consider a simple single layer, feed forward neural network. This model can easily have many more parameters than the number of predictors. For this data, the number of predictors already exceeds the number of samples. Therefore, even the simplest of neural network models can overfit to the available data without appropriate precautions.

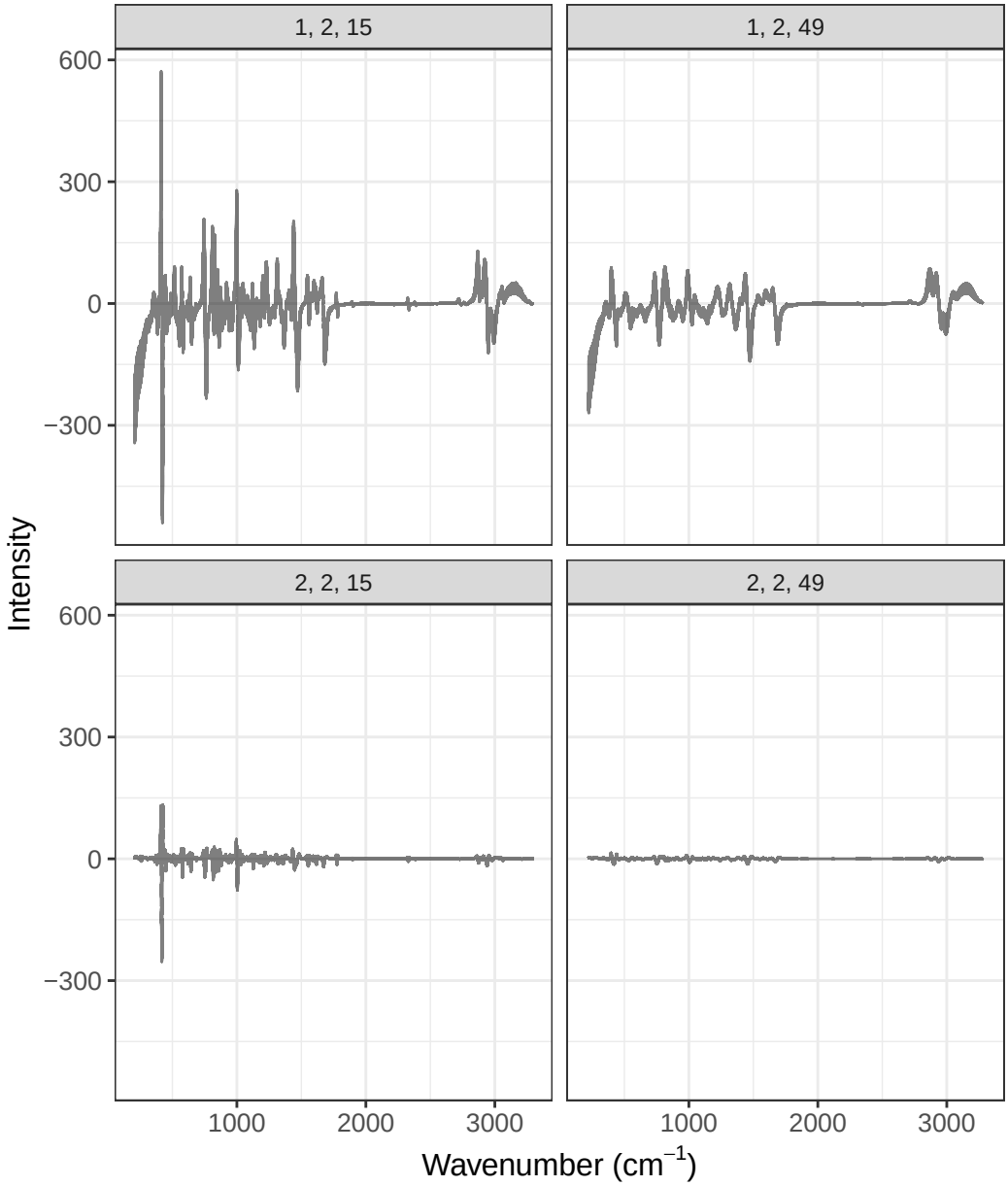

Figure 9: The impact of the Savitzky-Golay procedure on the Raman spectra. Each facet corresponds to a different parameterization of the procedue, where the first number represnts the differentiation order (1 or 2), the second number represents the polynomial order (2), and the third number the smoothing window (15 or 49).

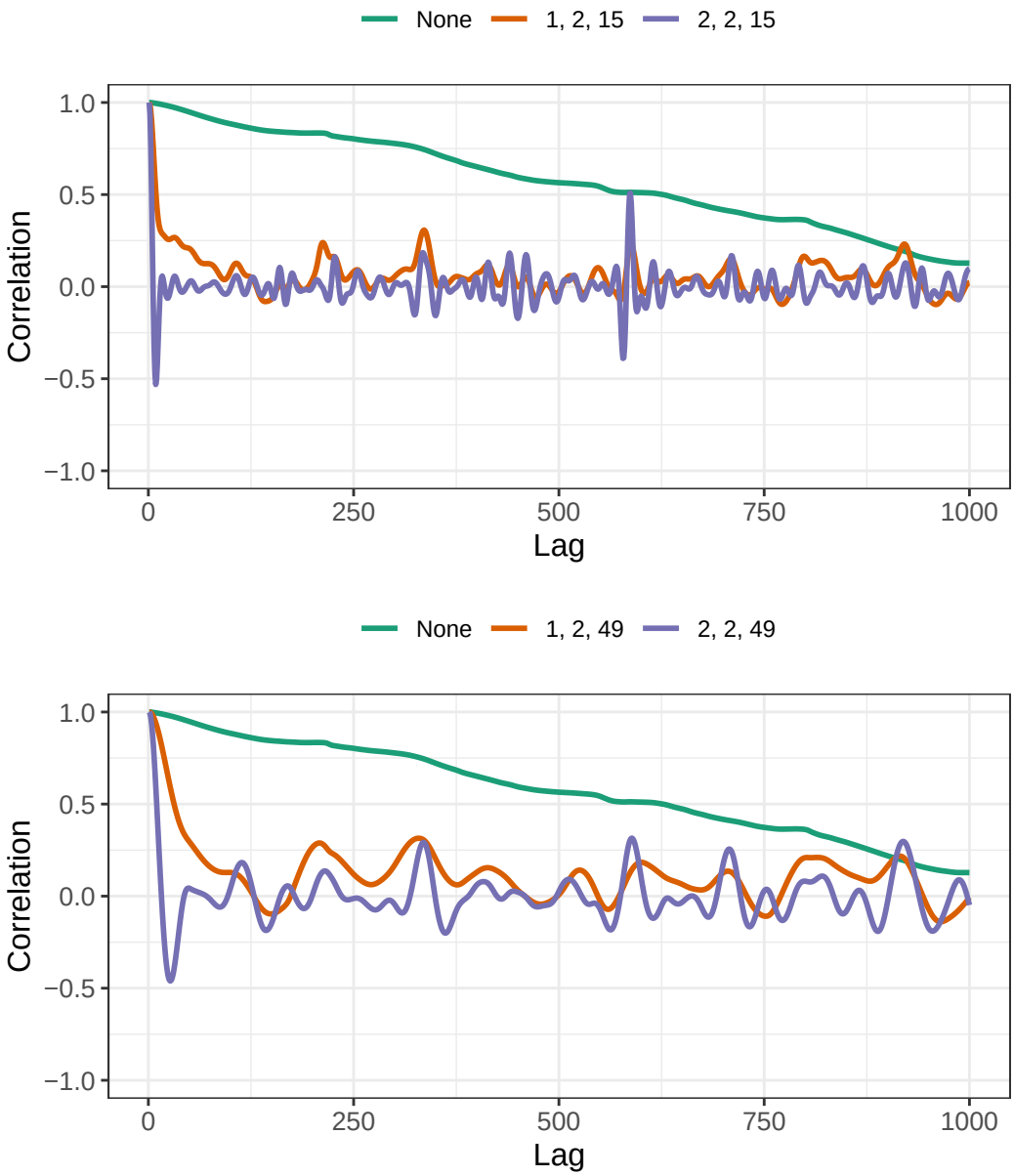

Figure 10: The impact of the Savitzky-Golay procedure on the correlation between lagged wavenumbers.

As part of the modeling process, we need to find a set of values for the tuning parameters of each model that effectively uncovers an optimal predictor-response relationship. As mentioned in the section on data splitting, the search for an optimal model must be done in the context of cross-validation to protect the model building process from overfitting to the available data. The next question we must address is what values of the tuning parameters should be evaluated. A brute-force approach would be to evaluate many different values of the tuning parameter and select the optimal one. More sophisticated techniques are also available that utilize gradient descent, genetic algorithms, or principles of experimental design to more effectively find an optimal set of parameter values (Ali et al. 2023; Ippolito 2022).

How should the parameter sets be evaluated? Answering this question depends on the response. When the response is continuous, then the two most common performance metrics are R^2 and root mean square error (RMSE) (Neter et al. 1996). Many more options are available when the response is categorical, and the user must be keenly aware of response characteristics when selecting the performance metric. For example, if a categorical outcome is highly imbalanced, then selecting accuracy as the metric is not advisable. Specifically, it is possible to get high accuracy simply by classifying all samples into the majority class. Instead, a metric like the Kappa statistic (Cohen 1960) or area under the receiver operating characteristic curve (Nahm 2022) may be better choices for a performance metric since these measurements force a model to more accurately predict the minority class.

In this manufacturing example, the response is continuous and the metric of RMSE will be used to assess predictive performance. While there are many models to choose from, we will compare four modeling techniques for this data: partial least squares (PLS), random forest (RF), Cubist, and support vector machines (SVM). These models were selected to illustrate a range of types of models. We will now provide a high-level explanation of each of these models. Please see the references to learn more.

Spectroscopy data has traditionally been modeled using PLS (Htet et al. 2021; Esmonde-White et al. 2017). PLS is a logical technique to use for this type of data because it naturally handles highly correlated predictors. This model seeks to find linear combinations of the original predictors that have optimal correlation with the response by using as few linear combinations as possible (Wold, Sjöström, and Eriksson 2001). Specifically, PLS finds linear combinations that summarize variability across the predictors, while simultaneously finding the combinations that are optimally correlated with the response. There is one tuning parameter for PLS, which is the number of linear combinations, or latent variables, to retain.

Random forest is a recursive partitioning, or tree-based method which is built on an ensemble of trees (Breiman 2001; Seifert 2020). A single tree is constructed by recursively splitting the data into subsets that have greater purity with respect to the response. The RF model provides an improvement over a single tree by reducing variance through an ensemble of trees. Specifically, an RF model does the following process many times: selects a bootstrap sample of the data and builds a tree on the bootstrap sample. To construct each tree, a randomly selected number of predictors is chosen at each split. An optimal predictor within the sample is selected and the routine proceeds to the next split. Prediction for a new sample is the average

value across the entire ensemble of trees. There are two tuning parameters for RF, which are the number of trees in the ensemble and the number of randomly selected predictors for each split.

The Cubist model is also constructed from an ensemble of trees, but in a very different, more complex way than RF (Quinlan 1987). It uses a model tree rather than a partitioning tree as its foundation. The primary difference between a partitioning tree and a model tree is that a model tree constructs a linear model in each terminal node. A sequential ensemble of these trees can then be constructed where at each step in the sequence the response is adjusted to help improve model prediction for samples that are difficult to predict. Once the ensemble has been completed, the model predictions can be further adjusted by predictions from samples' closest neighbors. Cubist has two tuning parameters, which are the number of committees and the number of neighbors.

Support vector machines is a modeling technique which uncovers the relationship between the predictors and the response using samples that lie outside of a margin (a boundary about the optimal relationship) (Drucker et al. 1996; Ullah et al. 2018). Several versions of SVMs exist; the one implemented in this analysis uses a radial basis function. For the radial-basis SVM the number of samples allowed to be outside of the margin is controlled by the cost parameter and the flexibility of the surface is controlled by the sigma parameter. Therefore, the radial basis SVM has the flexibility to identify a non-linear relationship between the predictors and the response.

For each model a range of possible tuning parameter values was selected and evaluated using 5 repeats of 10-fold cross validation. The tuning parameter profiles for each model can then be examined to determine the values that correspond to the optimal model performance. Figure 11 (a) illustrates the tuning parameter profiles across the number of components that were evaluated for the original spectra as well as the different Savitzky-Golay pre-processed spectra. RMSE generally decreases as the number of components increases regardless of whether or not the spectra were pre-processed. The model with the lowest RMSE uses the SG filter with parameterization of first order differentiation, second order polynomial, and a small window size (1, 2, 15), and 12 components.

Figure 11 (b) displays the tuning parameter profile for the RF model. To tune this model, 8 values of the *mtry* parameter were evaluate for each predictor set. We can see from this figure that as the value of the *mtry* parameter increases, the RMSE value decreases. Also, predictor set that has the best RMSE is the same as the set that was identified using PLS. However, the optimal RF model (RMSE = 3.41) does not perform as well as the optimal PLS model (RMSE = 1.77).

The next step in the process is to select which predictor set, model, and tuning parameter settings are optimal for the data. Figure 12 displays the cross-validated RMSE values for each model's optimal tuning parameter settings and predictor set. There several important findings that this figure reveals. First, the SVM and RF models have improved predictive performance when the SG pre-processing is applied. On the other hand, the PLS and Cubist models are

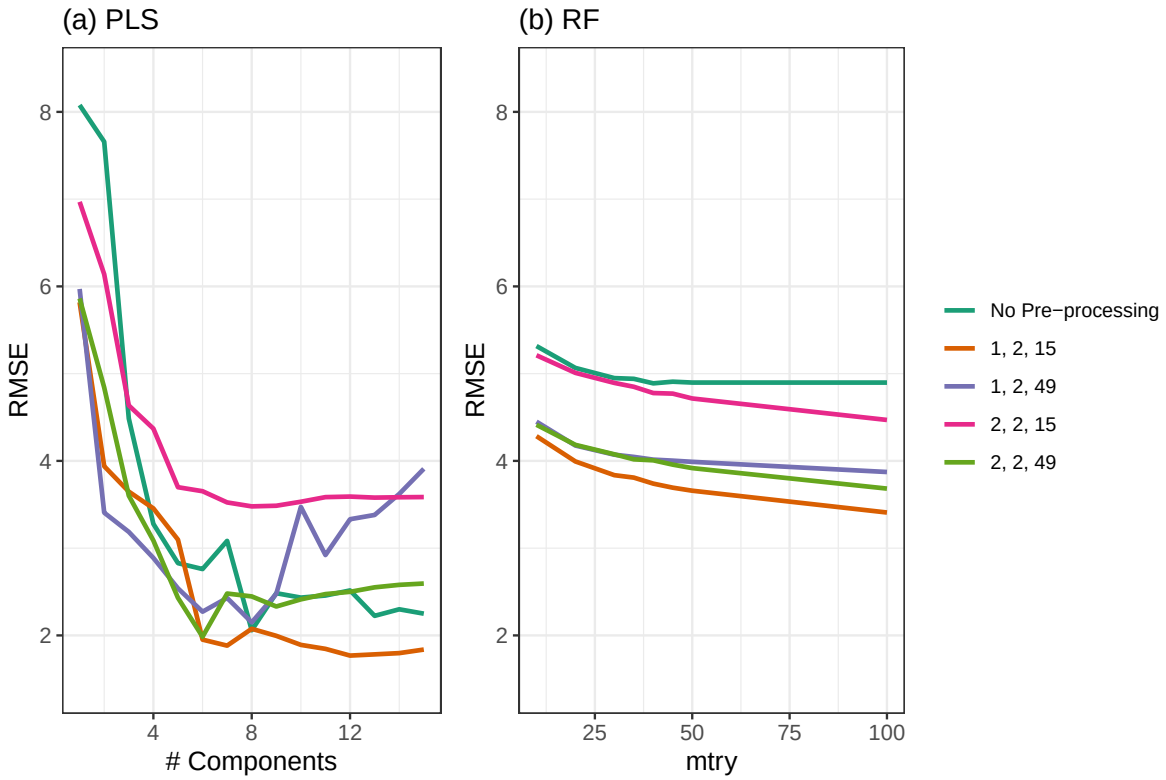

Figure 11: The tuning parameter profiles for partial least squares and random forest.

not substantially improved with this pre-processing. In fact, the performance can be made worse if second order differentiation is applied.

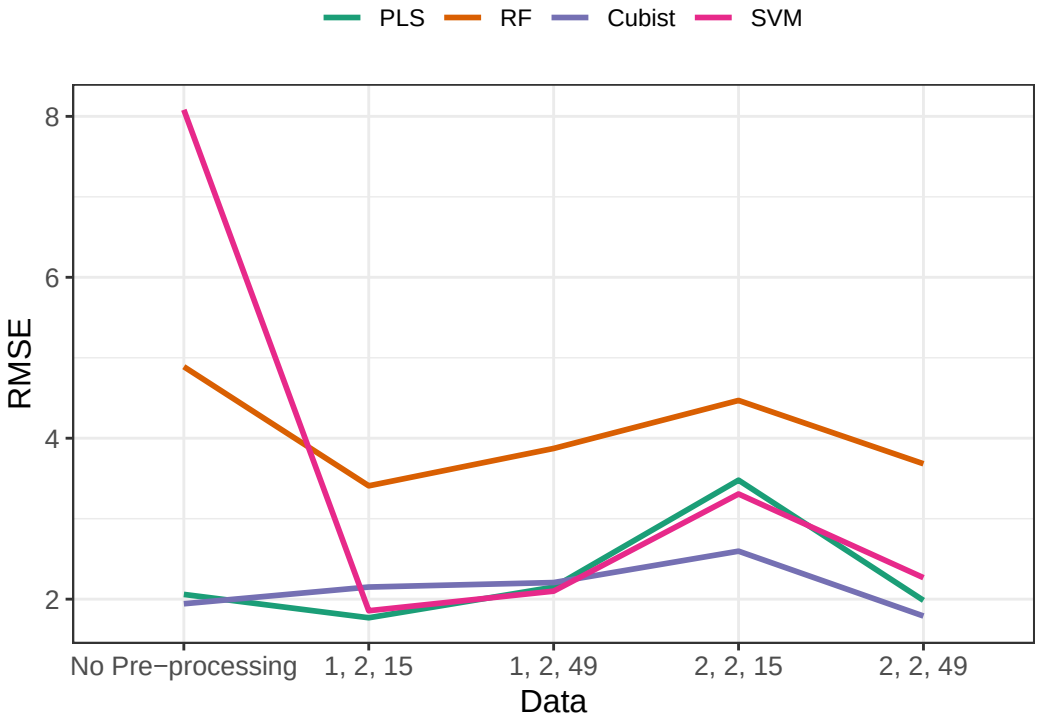

Figure 12: Cross-validation predictive performance using the optimal tuning parameter settings for each predictor set and model.

Across the five pre-processing sets and four models, let's compare the the observed versus predicted values for the PLS, Cubist, and SVM models with the unprocessed, and two variations of the SG pre-processed data (1, 2, 15 and 2, 2, 15). Figure 13 highlights some interesting characteristics across the models and predictor sets. First, we see that the SVM model cannot decipher the relationship between predictors and the response when the data is not pre-processed. In fact, it is challenging for SVM models to find predictive signal when there are many correlated predictors in the data [REFS]. Next, there is one sample that is challenging for the PLS model to fit when the predictors are not pre-processed or when there is some pre-processing (2, 2, 15). For the Cubist model, this same sample is better predicted by the model with no pre-processing, and less well predicted with either of the pre-processed data sets. What this means is that the same pre-processing approach is rarely effective across all models. Instead, many model types evaluated across a range of tuning parameters along with a variety of pre-processing conditions need to be evaluated in order to find the optimal model,

The RF model is popular model to use because it often performs well regardless of application,

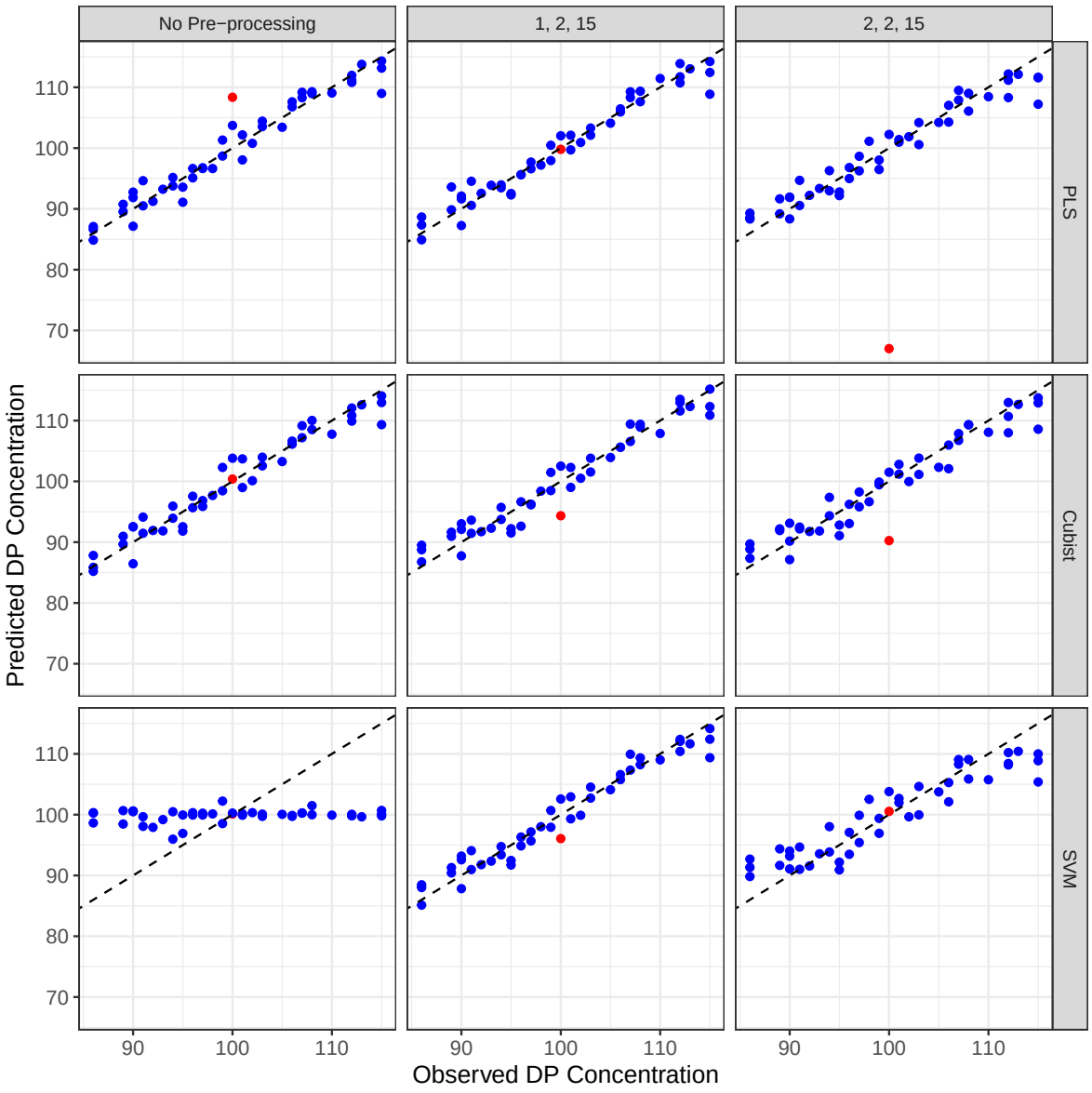

Figure 13: Comparison of observed versus hold-out predicted values from cross-validation for the optimal tuning parameter settings for three models and three pre-processing combinations. One challenging sample to predict is highlighted in red.

and becuase of its implementation in many software libraries and packages. In this example, however, it is the worst performing model. Why is it difficult for this model to predict the drug product concentration? Figure 14 presents the observed versus predicted values for the best and worst trained RF models. A smoothing line is overlayed for each predictor set. These lines help to elucidate the challenges that the RF model has with this data. For either predictor set, RF is much less accurate at predicting very low or very response values. The loss of accuracy in prediction at the extremes of the response is a problem with the RF model (Pang, Chang, and Chen 2022). This is primarily due to the fact that the terminal nodes of the trees within the ensembles are comprised of the existing data. Therefore, samples with extreme response values (either small or large) will be constrained to be either greater than the smallest value or less than the largest value. One potential remedy to this problem is to use a different ensemble model, like Cubist, that can better predict the extreme response values.

Again, not all models will perform similarly on all data sets. Instead, we should take an approach that allows many models to be evaluated prior to selecting a model that will be evaluated on the test set.

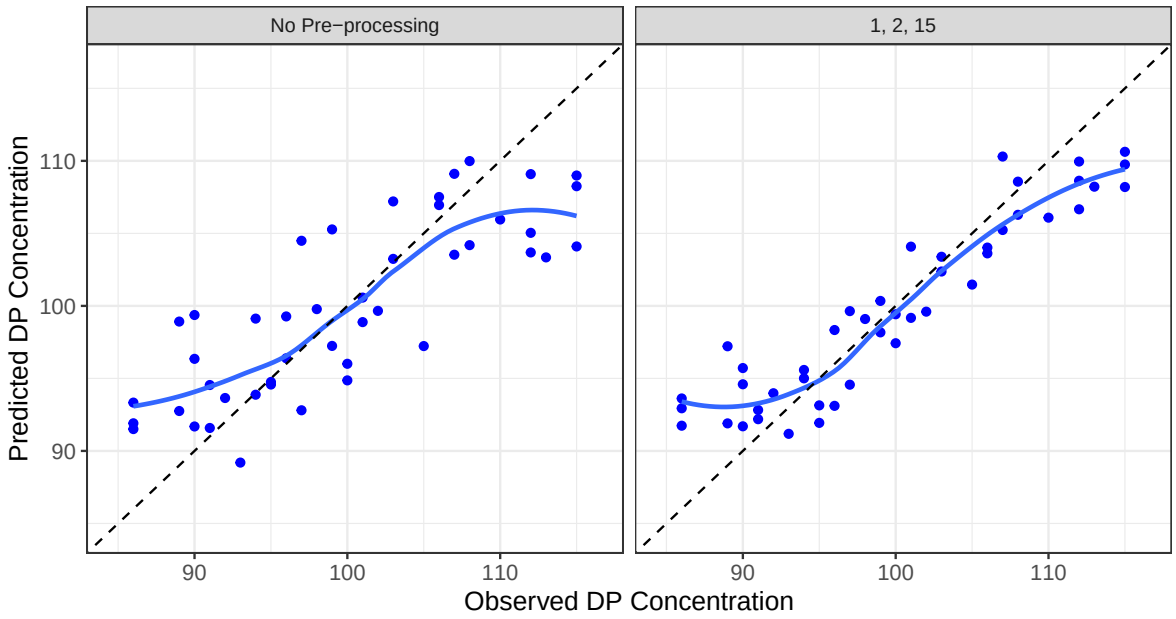

Figure 14: Comparison of observed versus hold-out predicted values from cross-validation for the optimal tuning parameter settings for two random forest models. A scatterplot smoother is overlayed for each pre-processed data set.

The final step in this tutorial will be to select the optimal model and apply it to the test set to assess predictive performance. Two model, pre-processing combinations are nearly identical in their cross-validation performance. The PLS model with SG pre-processing of 1, 2, 15 has an RMSE of 1.77 and corresponding standard error of 0.1. The cubist model with SG

pre-processing of 2, 2, 49 has an RMSE of 1.79 with standard error of 0.07. Either of these models would be acceptable to choose as a final model. At this point, the rationale for picking one model over the other may fall to practical considerations. For example, the PLS model is easier to interpret and implemented in some on-board software within the manufacturing community. On the other hand, it may be more practical to use the original spectra rather than passing it through the SG filer. If this is the case, then we would favor the Cubist model with no-preprocessing as the final model. Figure 15 displays the obsered versus predicted values of drug product concentration for the final PLS model. For this model, the RMSE of the test set is 1.93, which is comparable to the RMSE of the training set.

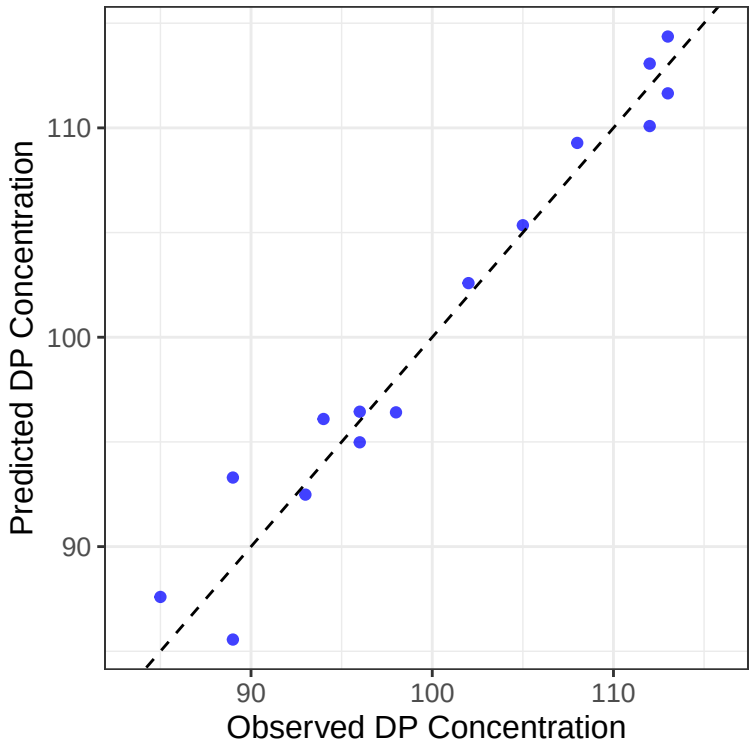

Figure 15: Comparison of observed versus predicted values of the test set for the optimal PLS model.

Conclusions

Ali, Yasser A, Emad Mahrous Awwad, Muna Al-Razgan, and Ali Maarouf. 2023. "Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity." *Processes* 11 (2): 349.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Cleveland, William S, and Susan J Devlin. 1988. "Locally Weighted Regression: An Approach

to Regression Analysis by Local Fitting." *Journal of the American Statistical Association* 83 (403): 596–610.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46.

Drucker, Harris, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. "Support Vector Regression Machines." *Advances in Neural Information Processing Systems* 9.

Esmonde-White, Karen A, Maryann Cuellar, and Ian R Lewis. 2022. "The Role of Raman Spectroscopy in Biopharmaceuticals from Development to Manufacturing." *Analytical and Bioanalytical Chemistry*, 1–23.

Esmonde-White, Karen A, Maryann Cuellar, Carsten Uerpmann, Bruno Lenain, and Ian R Lewis. 2017. "Raman Spectroscopy as a Process Analytical Technology for Pharmaceutical Manufacturing and Bioprocessing." *Analytical and Bioanalytical Chemistry* 409 (3): 637–49.

Hawkins, D. 2004. "The Problem of Overfitting." *Journal of Chemical Information and Computer Sciences* 44 (1): 1–12.

Htet, Tar Tar Moe, Jordi Cruz, Putthiporn Khongkaew, Chaweewan Suwanvecho, Leena Suntornsuk, Nantana Nuchtavorn, Waree Limwikrant, and Chutima Phechkrajang. 2021. "PLS-Regression-Model-Assisted Raman Spectroscopy for Vegetable Oil Classification and Non-Destructive Analysis of Alpha-Tocopherol Contents of Vegetable Oils." *Journal of Food Composition and Analysis* 103: 104119.

Ippolito, Pier Paolo. 2022. "Hyperparameter Tuning: The Art of Fine-Tuning Machine and Deep Learning Models to Improve Metric Results." In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, 231–51. Springer.

Jesus, José Izo Santana da Silva de, Raimar Löbenberg, and Nadia Araci Bou-Chacra. 2020. "Raman Spectroscopy for Quantitative Analysis in the Pharmaceutical Industry." *Journal of Pharmacy and Pharmaceutical Sciences* 23 (1): 24–46.

Kohavi, R. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *International Joint Conference on Artificial Intelligence* 14: 1137–45.

Kuhn, Max, Kjell Johnson, et al. 2013. *Applied Predictive Modeling*. Vol. 26. Springer.

Kuhn, Max, and Julia Silge. 2022. *Tidy Modeling with r.* " O'Reilly Media, Inc.".

Luers, James K, and Robert H Wenning. 1971. "Polynomial Smoothing—Linear Vs Cubic." *Technometrics* 13 (3): 589–600.

Nahm, Francis Sahngun. 2022. "Receiver Operating Characteristic Curve: Overview and Practical Use for Clinicians." *Korean Journal of Anesthesiology* 75 (1): 25–36.

Neter, John, Michael H Kutner, Christopher J Nachtsheim, William Wasserman, et al. 1996. "Applied Linear Statistical Models."

Pang, Alexis, Melissa WL Chang, and Yang Chen. 2022. "Evaluation of Random Forests (RF) for Regional and Local-Scale Wheat Yield Prediction in Southeast Australia." *Sensors* 22 (3): 717.

Quinlan, J. Ross. 1987. "Simplifying Decision Trees." *International Journal of Man-Machine Studies* 27 (3): 221–34.

Rinnan, Åsmund, Frans Van Den Berg, and Søren Balling Engelsen. 2009. "Review of the

Most Common Pre-Processing Techniques for Near-Infrared Spectra." *TrAC Trends in Analytical Chemistry* 28 (10): 1201–22.

Savitzky, Abraham, and Marcel JE Golay. 1964. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." *Analytical Chemistry* 36 (8): 1627–39.

Seifert, Stephan. 2020. "Application of Random Forest Based Approaches to Surface-Enhanced Raman Scattering Data." *Scientific Reports* 10 (1): 5436.

Silge, A, Karina Weber, D Cialla-May, L Müller-Bötticher, D Fischer, and J Popp. 2022. "Trends in Pharmaceutical Analysis and Quality Control by Modern Raman Spectroscopic Techniques." *TrAC Trends in Analytical Chemistry* 153: 116623.

Stevens, Antoine, and Leornardo Ramirez-Lopez. 2022. *An Introduction to the Prospectr Package*.

Ullah, Rahat, Saranjam Khan, Samina Javaid, Hina Ali, Muhammad Bilal, and Muhammad Saleem. 2018. "Raman Spectroscopy Combined with a Support Vector Machine for Differentiating Between Feeding Male and Female Infants Mother's Milk." *Biomedical Optics Express* 9 (2): 844–51.

Wold, Svante, Michael Sjöström, and Lennart Eriksson. 2001. "PLS-Regression: A Basic Tool of Chemometrics." *Chemometrics and Intelligent Laboratory Systems* 58 (2): 109–30.