Feature Engineering Chapter 9
Working with Profile Data, pt 1

Simple example of “Unit of Prediction”

- the price of a house in Iowa

“The data structure is straightforward: rows are houses and columns are fields describing them. There is one house per row and, for the most part, we can assume that these houses are statistically independent of one another. In statistical terms, it would be said that the houses are the *independent experimental unit* of data… the properties are the *unit of prediction*.”

SaleType	SaleCondition	SalePrice
WD	Normal	208500
WD	Normal	181500
WD	Normal	223500
WD	Abnorml	140000
WD	Normal	250000

Ex. slightly more challenging data structure

- Chicago train ridership data

"The data set has rows corresponding to specific dates and columns are characteristics of these days: holiday indicators, ridership at other stations a week prior, and so on. Predictions are made daily; this is the *unit of prediction*."

Recall there is also weather measurements:

How does this complicate things?

Time	Temp	Humidity	Wind	Conditions
00:53	27.0	92	21.9	Overcast
01:53	28.0	85	20.7	Overcast
02:53	27.0	85	18.4	Overcast
03:53	28.0	85	15.0	Mostly Cloudy
04:53	28.9	82	13.8	Overcast
05:53	32.0	79	15.0	Overcast

"Since the goal is to make daily predictions, the *profile* of within-day weather measurements should be somehow summarized at the day level in a manner that preserves the potential predictive information."

How could you do this?

Mean, median, range across data, % of day that was "clear"

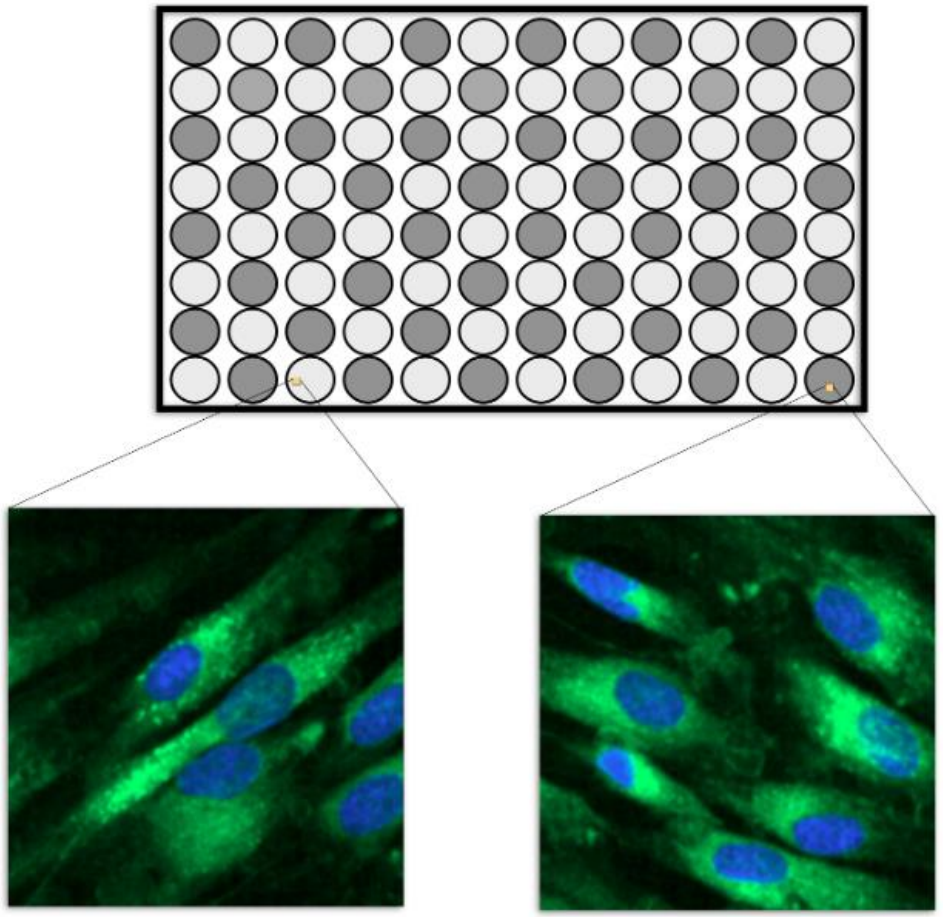
“Are there good and bad ways of summarizing profile data?”

YES

- Ex. Testing drug treatments on batches of cells
- *Unit of measure: well;*
- measure characteristics of cells (cell-within-well)
- Suppose calculating feature from X1 and X2 (e.g. cell length and width) “If an important feature of these two values is their difference, then there are two ways of performing the calculation.” (to get a *well-level* summary)

1. mean(x1 – x2)
2. mean(x1) – mean(x2)

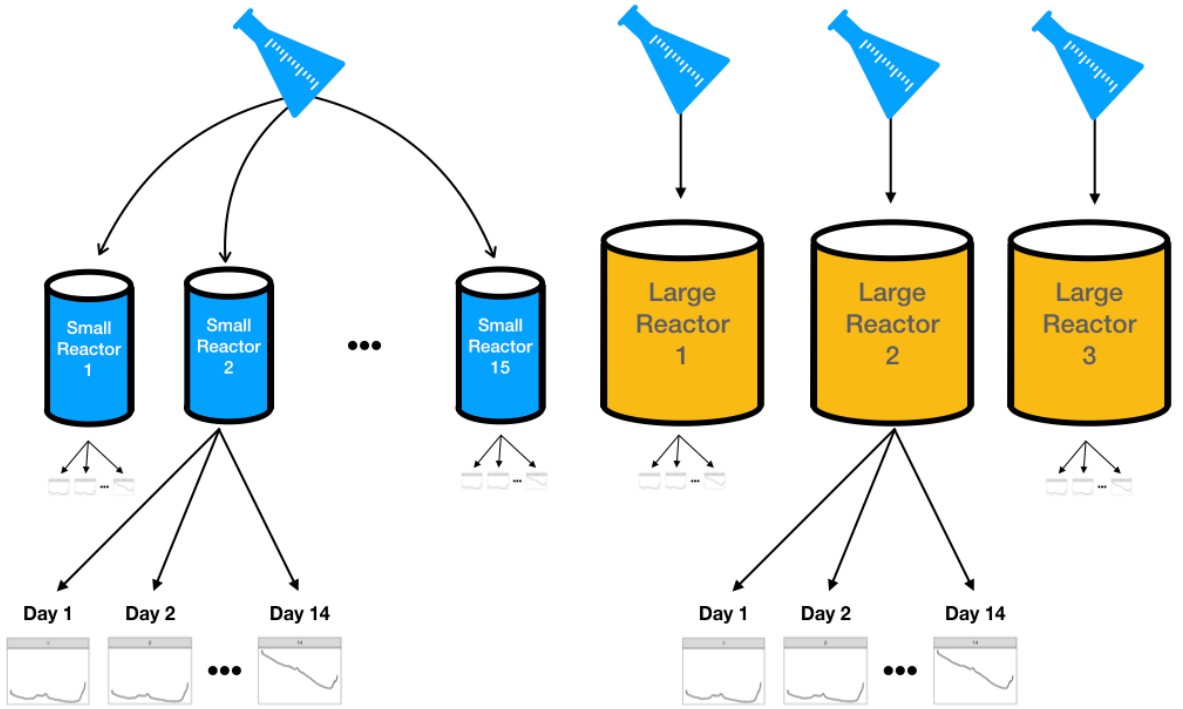
Which seems more appropriate?



The problem with approach 2. is breaks correlational structure of data.

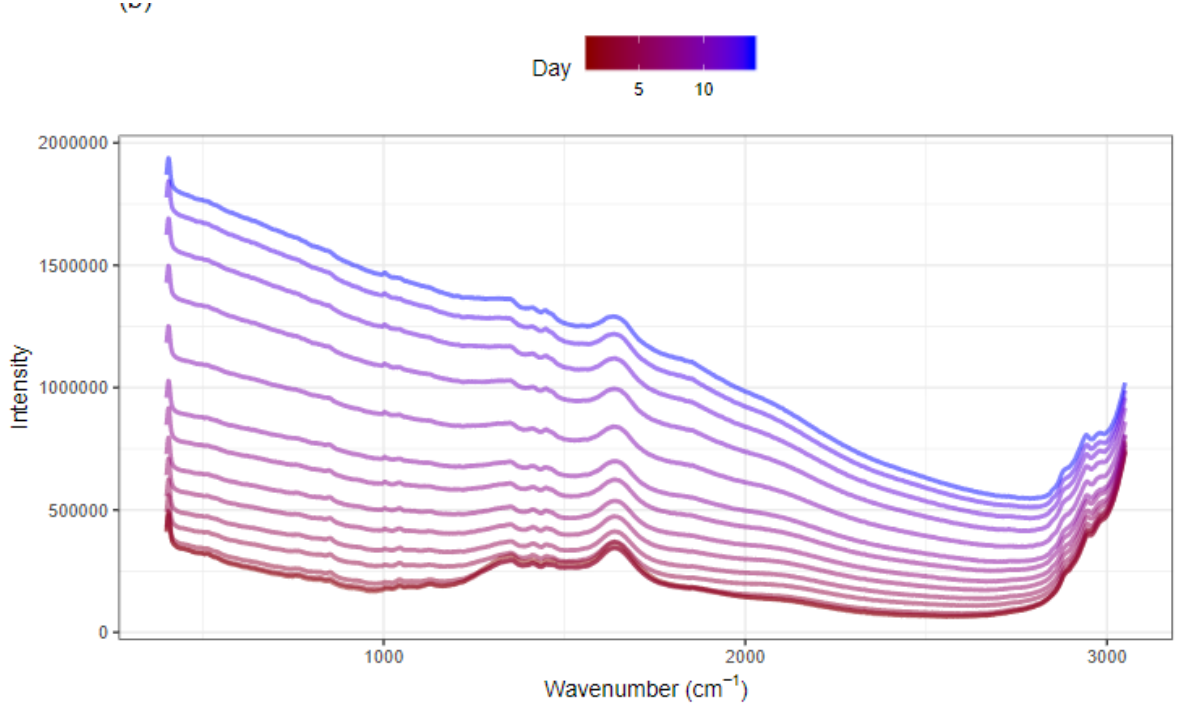
Example Pharmaceutical manufacturing

- “Pharmaceutical companies use spectroscopy measurements to assess critical process parameters during the manufacturing of a biological drug… Models built on this process can be used with real-time data to recommend changes that can increase product yield.”
- (spectroscopy is meant to give faster window to things like ammonia level, glucose level, etc.)



Example Pharmaceutical manufacturing

- “Nearly 2600 wavelengths are measured each day for two weeks for each of 15 small-scale bioreactors. This type of data forms a *hierarchical* structure in which wavelengths are measured within each day and within each bioreactor. Another way to say this is that the wavelength measurements are nested within day which is further nested within bioreactor.”
- “The use case for the model is to make a prediction for a bioreactor for a specific number of days that the cells have been growing. For this reason, **the unit of prediction is day within bioreactor**.”
- HOWEVER THE EXPERIMENTAL UNIT IS AN INDIVIDUAL BIOREACTOR



Correlation and understanding experimental unit

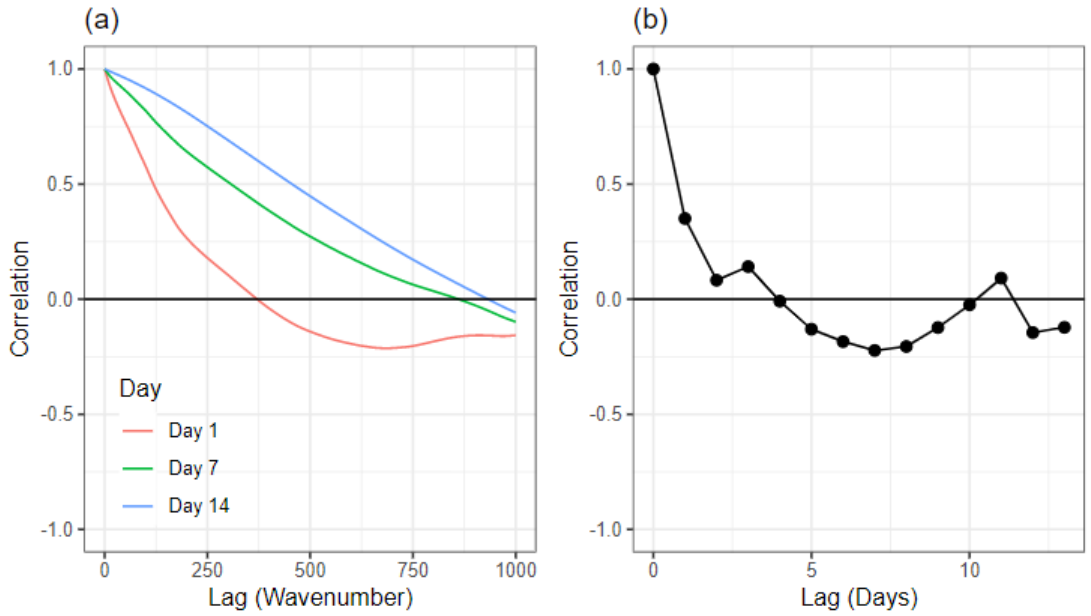


Figure 9.4: (a) Autocorrelations for selected lags of wavelengths for small-scale bioreactor 1 on the first day. (b) Autocorrelations for lags of days for average wavelength intensity for small-scale bioreactor 1.

- "Understanding the units will guide the selection of cross validation method and is crucial for getting an honest assessment of a model's predictive ability on new days."

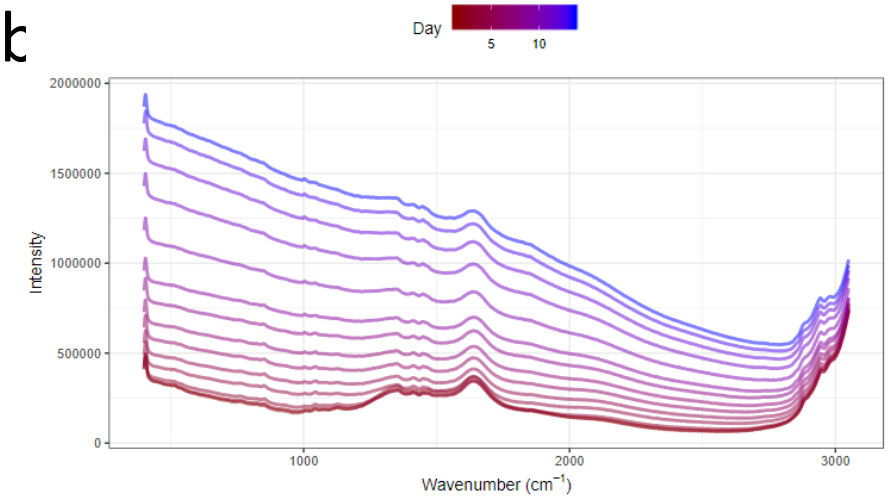
Correlation and understanding experimental unit

- “Consider, for example, if each day (within each bioreactor) was taken to be independent experimental unit and *V*-fold cross-validation was used as the resampling technique. In this scenario, days within the same bioreactor will likely be in both the analysis and assessment sets… Given the amount of correlated data within day, this is a bad idea since it will lead to artificially optimistic characterizations of the model.”

Feature Engineering Chapter 9
Working with Profile Data, pt 2

Reducing background noise

- Some big differences are due to differences in background, rather than differences in molecules of interest (like glucose)
- Smooth polynomial to LOWEST INTENSITY across spectra for each b



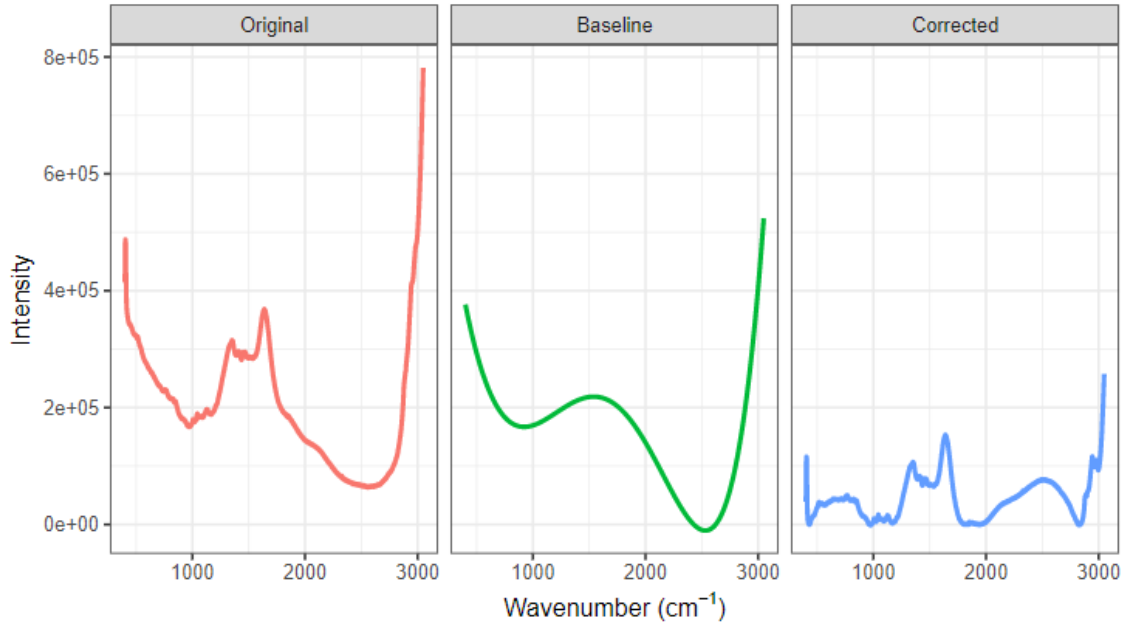


Figure 9.5: A polynomial baseline correction for the small-scale bioreactor 1, day 1 intensities.

Reducing Other Noise

- Standardize intensity (mean = 0, standard deviation = 1)
- In this case, also trimmed extreme values to prevent this from skewing standardization

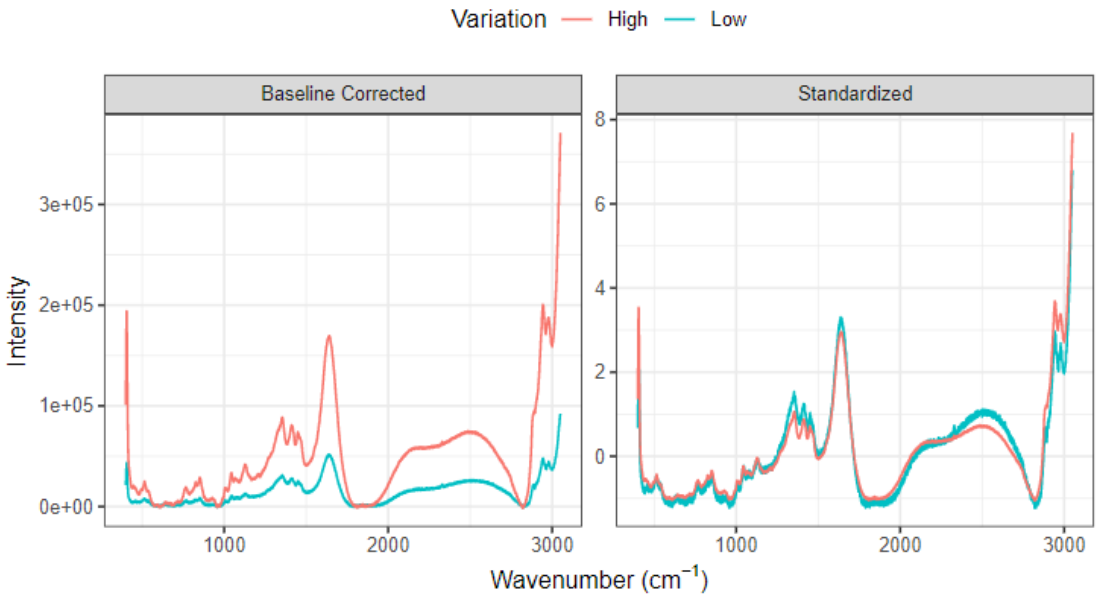


Figure 9.6: (a) The baseline-corrected intensities for the spectra that are the most- and least variable. (b) The spectra after standardizing each to have a mean of 0 and standard deviation of 1.

Reducing other noise

- Can correct for jagged differences in data via moving average or smoothing spline

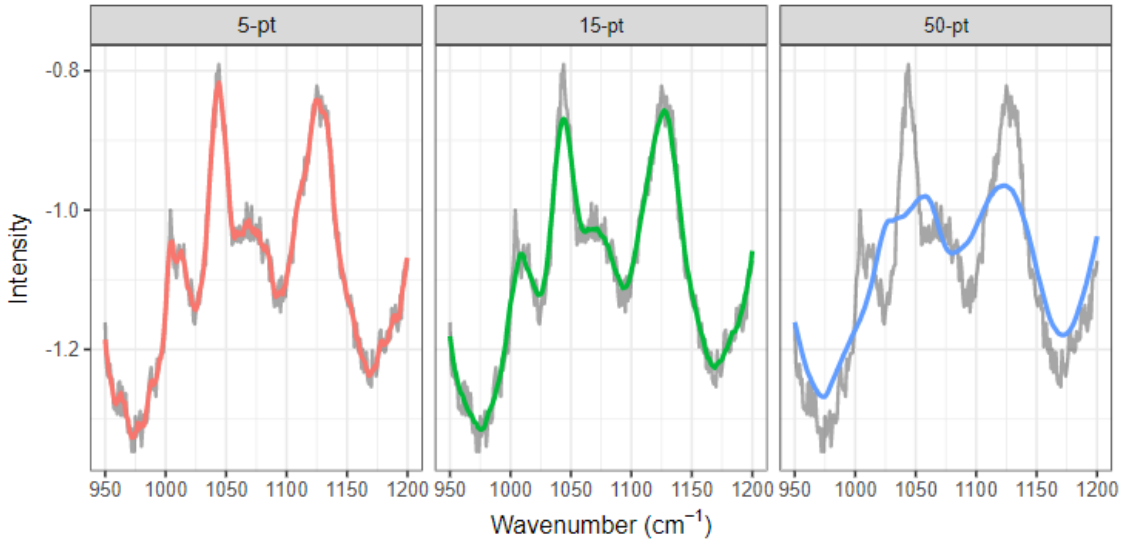


Figure 9.7: The moving average of lengths 5, 15, and 50 applied to the first day of the first small-scale bioreactor for wavelengths 950 through 1200.

Exploiting correlation

- “The previous steps of estimating and reducing baseline and reducing noise help to refine the profiles and enhance the true signal that is related to the response within the profiles. These steps, however, do not reduce the between-wavelength correlation within each sample which is still a problematic characteristic for many predictive models.”
- Try PCA, kernel PCA, ICA, partial least squares… or other dimensionality reduction techniques…

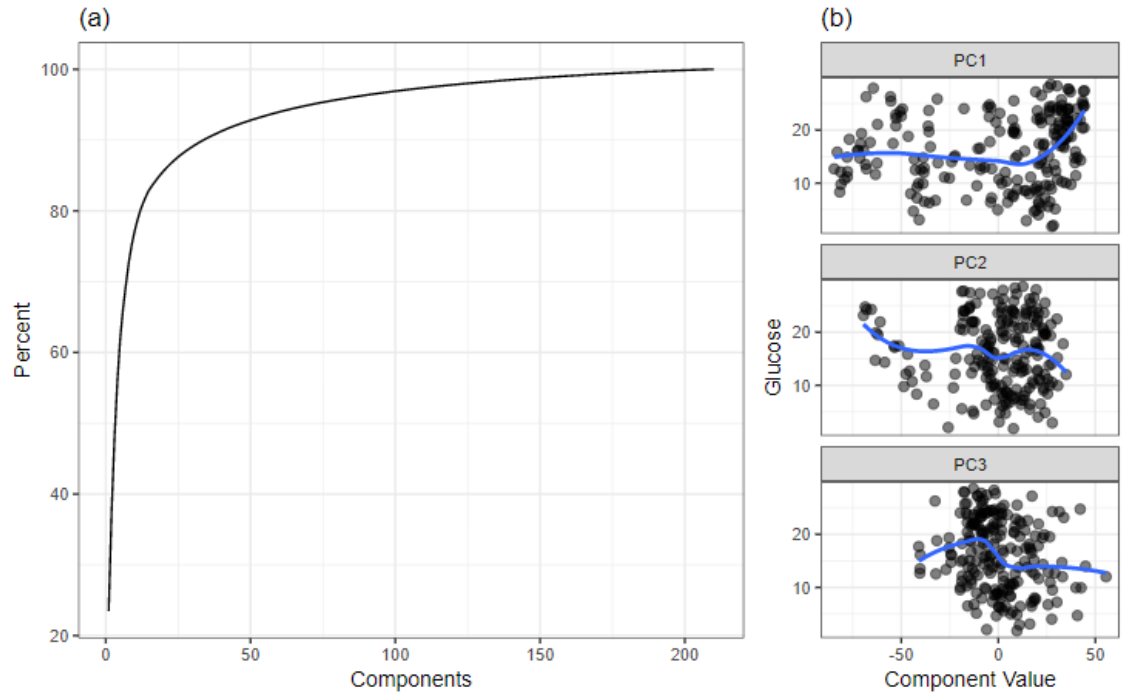


Figure 9.8: PCA dimension reduction applied across all small-scale data. (a) A scree plot of the cumulative variability explained across components. (b) Scatterplots of Glucose and the first three principal components.

Exploiting correlation

- Calculate differences (x_t) – (x_t-1)
- Can substantially reduce autocorrelation
- (Could also filter out highly correlated spectra wavelengths…)

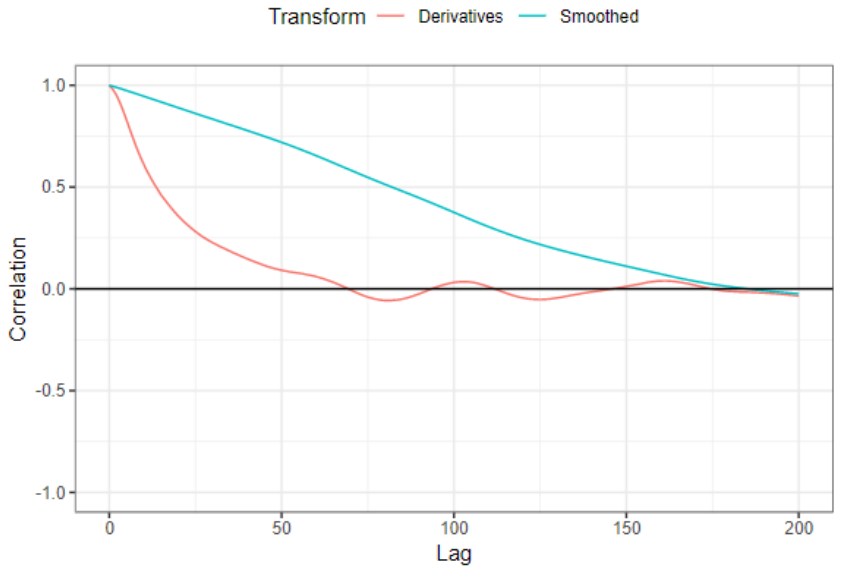


Figure 9.9: Autocorrelations before and after taking derivatives of the spectra.

Exploiting correlation

- “These steps shows how the within-spectra drift has been removed and most of the trends that are unrelated to the peaks have been minimized.”

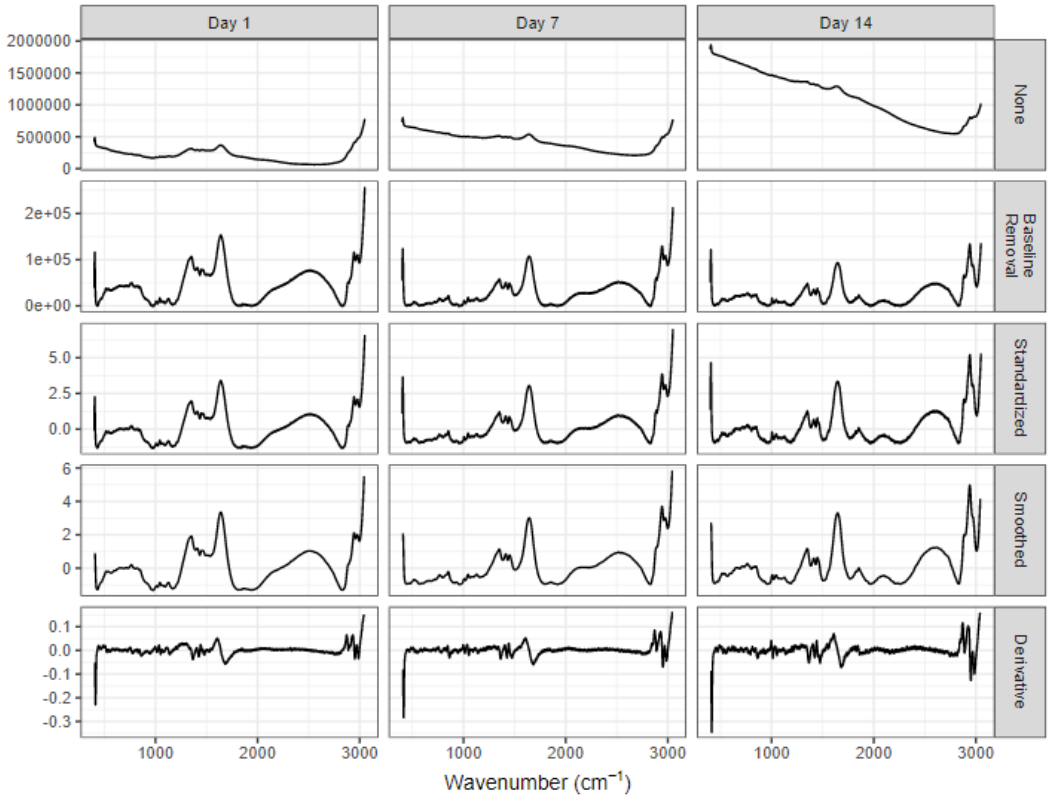


Figure 9.10: Spectra for the first day of the first small-scale bioreactor where the preprocessing steps have been sequentially applied.

Sampling pre-processing

- “The amount of preprocessing could be considered a tuning parameter.”
- Want best model *and* best preprocessing
- 15 small bioreactors in training set
  - Leave one out cross validation
  - V-fold cross validation, e.g. 5-fold (would have 3 reactors in hold-out)

Resample	Heldout Bioreactor
1	5, 9, and 13
2	4, 6, and 11
3	3, 7, and 15
4	1, 8, and 10
5	2, 12, and 14

  - Repeated V-fold cross-validation

Impacts of Data Processing on Modeling

- Notice systematic errors by day until preprocessing is compl

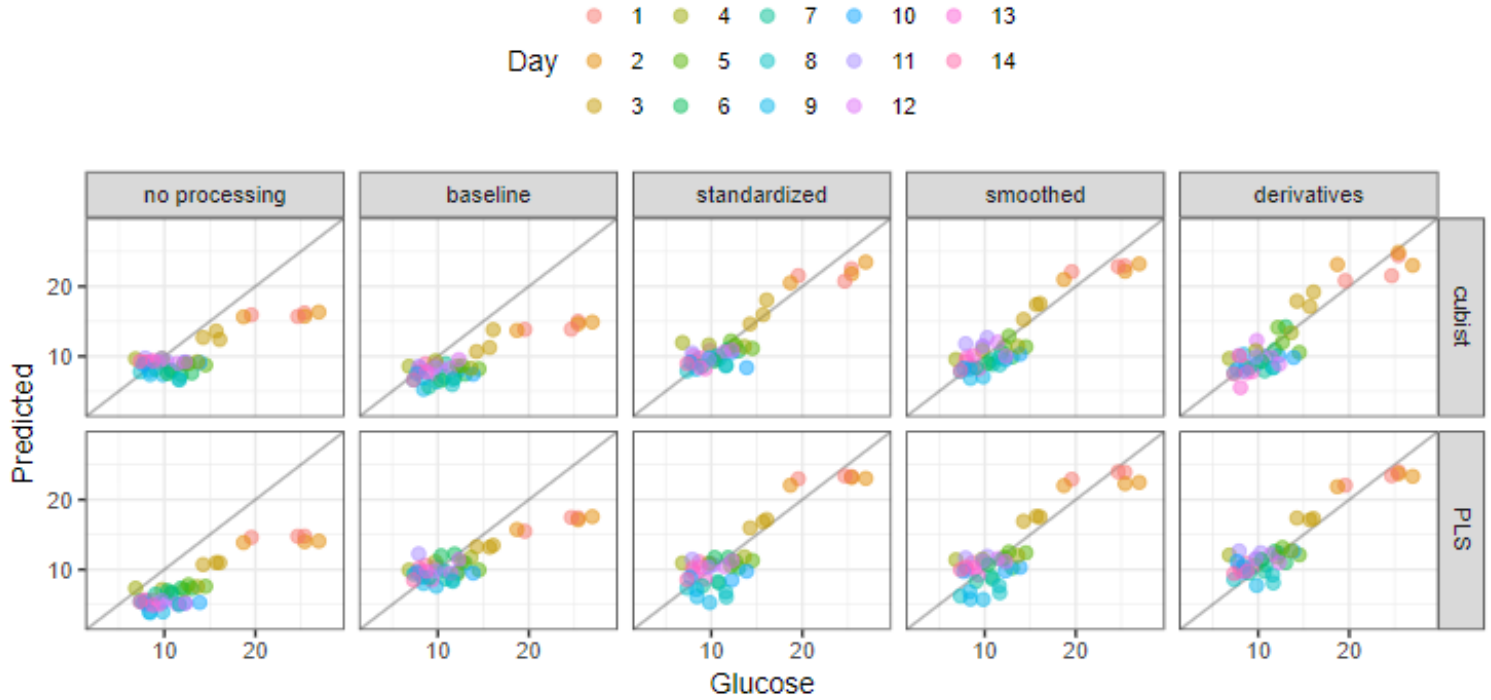


Figure 9.13: A comparison of the observed and predicted glucose values for the large-scale bioreactor data.

Impacts of Data Processing on Modeling

- If had taken naïve assumptior each day-within-bioreactor w experimental unit, would hav naïve performance estimates

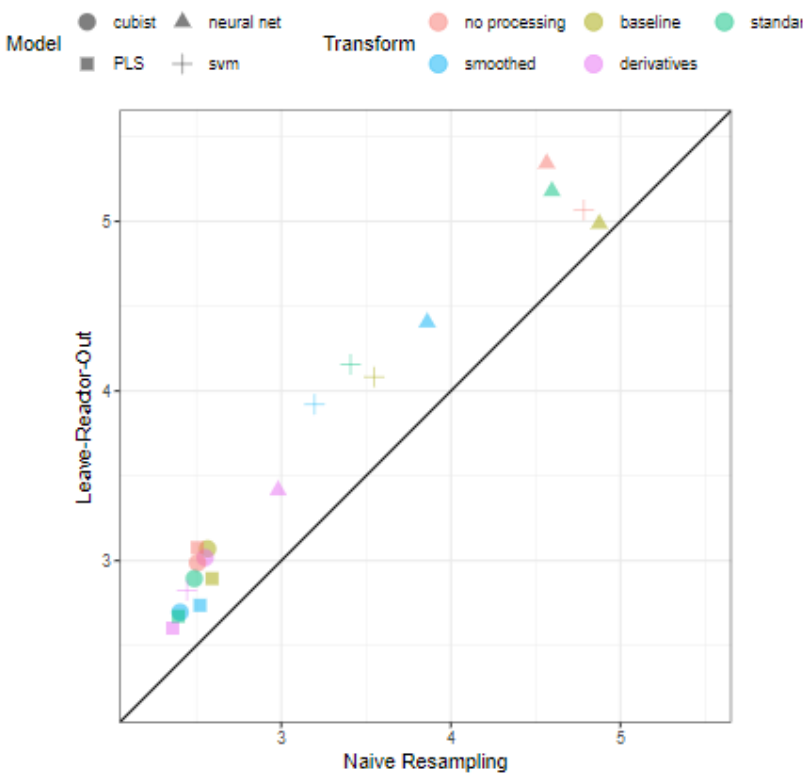


Figure 9.14: Cross-validation performance comparison using bioreactors, the experimental units, or naive resampling of rows.

Summary

- “the analyst needs to be keenly aware of what the experimental unit is. Understanding the unit informs decisions about how the profiles should be preprocessed, how samples should be allocated to training and test sets, and how samples should be allocated during resampling.”
- “Basic preprocessing steps for profiled data can include reducing baseline effect, reducing noise across the profile, and harnessing the information contained in the correlation among predictors. An underlying goal of these steps is to remove the characteristics that prevent this type of data from being used with most predictive models while simultaneously preserving the predictive signal between the profiles and the outcome.”