CHAPTER 3
DATA PRE-PROCESSING

APPLIED PREDICTIVE MODELING BY KUHN & JOHNSON
COLLATED BY PROF. CHING-SHIH (VINCE) TSOU (PH.D.)
CENTER FOR APPLICATIONS OF DATA SCIENCE (CADS)
GRADUATE INSTITUTE OF INFORMATION AND DECISION SCIENCES (GIIDS)
NATIONAL TAIPEI UNIVERSITY OF BUSINESS (NTUB)
CHINESE ACADEMY OF R SOFTWARE (CARS)
DATA SCIENCE & BUSINESS APPLICATIONS (DSBA) ASSOCIATION OF TAIWAN

AGENDA

- Introduction
- Case Study: Cell Segmentation in High-Content Screening
- Data Transformations for Individual Predictors
- Data Transformations for Multiple Predictors
- Dealing with Missing Values
- Removing Predictors
- Adding Predictors
- Binning Predictors

INTRODUCTION

- Data pre-processing techniques generally refer to the addition, deletion, or transformation of training set data.
- Different models have different sensitivities to the type of predictors in the model; ***how*** the predictors enter the model is also important.
- Transformations of the data to reduce the impact of data skewness or outliers can lead to significant improvements in performance.
- The need for data pre-processing is determined by the type of model being used. Some procedures, such as tree-based models, are notably insensitive to the characteristics of the predictor data.

INTRODUCTION

- For modeling techniques described in subsequent chapters, we will also discuss which, if any, pre-processing techniques can be useful.
- This chapter outlines approaches to unsupervised data processing: the outcome variable is not considered by the pre-processing techniques.
- One piece of information in the data is the submission date of the grant. This date can be represented in myriad ways: (Encoding)
  - The number of days since a reference date
  - Isolating the month, year, and day of the week as separate predictors
  - The numeric day of the year (ignoring the calendar year)
  - Whether the date was within the school year (as opposed to holiday or summer sessions)

CASE STUDY: CELL SEGMENTATION IN HIGH-CONTENT SCREENING

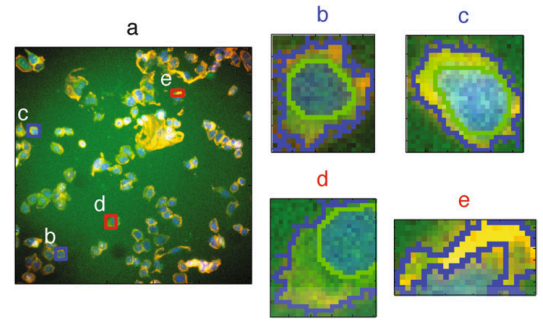
- Medical researchers often seek to understand the effects of medicines or diseases on the size, shape, development status, and number of cells in a living organism or plant.
- To do this, experts can examine the target serum or tissue under a microscope and manually assess the desired cell characteristics.
- This work is tedious and requires expert knowledge of the cell type and characteristics.
- Another way to measure the cell characteristics from these kinds of samples is by using high-content screening (Giuliano et al. 1997).
- Briefly, a sample is first dyed with a substance that will bind to the desired characteristic of the cells.

CASE STUDY: CELL SEGMENTATION IN HIGH-CONTENT SCREENING

If a researcher wants to quantify the size or shape of cell nuclei, then a stain can be applied to the sample that attaches to the cells' DNA.

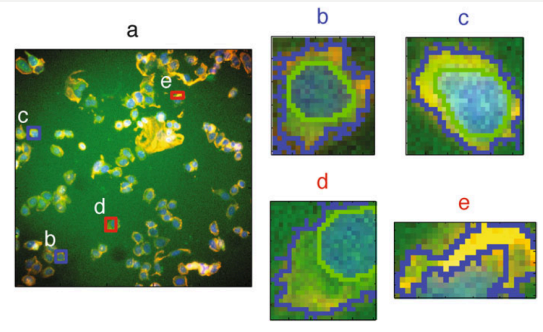
- The cells can be fixed in a substance that preserves the nature state of the cell.
- The sample is then interrogated by an instrument where the dye deflects light and the detectors quantify the degree of scattering for that specific wavelength.
- If multiple characteristics of the cells are desired, then multiple dyes and multiple light frequencies can be used simultaneously.
- The light scattering measurements are then processed through imaging software to quantify the desired cell characteristics.

CASE STUDY: CELL SEGMENTATION IN HIGH-CONTENT SCREENING



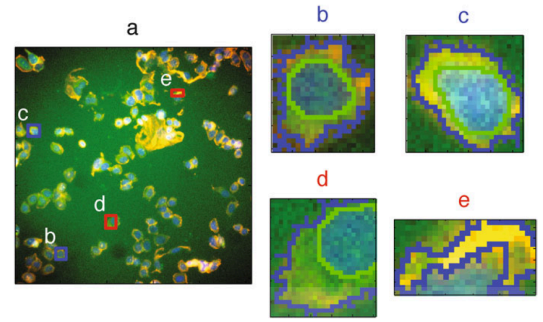
- Using an automated, high-throughput approach to access samples' cell characteristics can sometimes produce misleading results. (d, e)
- Hill et al. (2007) describe a research project that used high-content screening to measure several aspects of cells.
- In these images, the bright green boundaries identify the cell nucleus, while the blue boundaries define the cell perimeter.
- If cell size, shape, and/or quantity are the endpoints of interest in a study, then it is important that the instrument and imaging software can correctly segment cells.

CASE STUDY: CELL SEGMENTATION IN HIGH-CONTENT SCREENING



- For this research, Hill et al. (2007) assembled a data set consisting of 2,019 cells.
- Of these cells, 1,300 were judged to be ***poorly segmented (PS)*** and 719 were ***well segmented (WS)***; 1,009 cells were reserved for the ***training set***.

CASE STUDY: CELL SEGMENTATION IN HIGH-CONTENT SCREENING



- For a particular type of cell, the researchers used different stains that would be visible to different optical channels.
  - Channel one was associated with the cell body and can be used to determine the cell perimeter, area, and other qualities.
  - Channel two interrogated the cell nucleus by staining the nuclear DNA (shown in blue shading in this figure).
  - Channels three and four were stained to detect actin and tubulin, respectively.
- These are two types of filaments that transverse the cells in scaffolds and are part of the cell's cytoskeleton.
- For all cells, 116 features (eg. cell area, spot fiber count) were measured and were used to predict the segmentation quality of cells.

DATA TRANSFORMATIONS FOR INDIVIDUAL PREDICTORS

- Transformations of predictor variables may be needed for several reasons.
  - Some modeling techniques may have strict requirements, such as the predictors having a common scale.
  - In other cases, creating a model may be difficult due to outliers.
- Here we discuss centering, scaling, and skewness transformations.
  - Centering and Scaling (Standardization, normalization)
  - Transformations to Resolve Skewness (BC only for positive obs.)

CENTERING AND SCALING

- The most straightforward and common data transformation is to center and/or scale the predictor variables.
- To center a predictor, the average predictor value is subtracted from all the values. As a result, the predictor has a zero mean.
- Scaling the data coerce the values to have a common standard deviation of one.
- These manipulations are generally used to improve the numerical stability of some calculations. (pros)
- The only real downside to these transformations is a loss of interpretability of the individual values since the data are no longer in the original units. (cons)

TRANSFORMATIONS TO RESOLVE SKEWNESS

- Another common reason for transformations is to remove distributional skewness.
- Un-skewed or roughly symmetric distribution means that the probability of falling on either side of the distribution's mean is roughly equal.
- A right-skewed distribution has a large number of points on the left side of the distribution (smaller values) than on the right side (larger values).

TRANSFORMATIONS TO RESOLVE SKEWNESS

- A general rule of thumb to consider is that skewed data whose ratio of the highest value to the lowest value is greater than 20 have significant skewness.
- Also, the skewness statistic can be used as a diagnostic.
- If the predictor distribution is roughly symmetric, the skewness values will be close to zero.
- As the distribution becomes more right skewed, the skewness statistic becomes larger.

TRANSFORMATIONS TO RESOLVE SKEWNESS

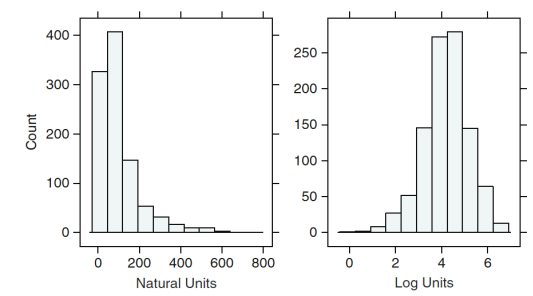
- The formula for the sample skewness statistic is

$$\text{skewness} = \frac{\sum(x_i - \overline{x})^3}{(n-1)v^{3/2}}$$

$$\text{where} \quad v = \frac{\sum(x_i - \overline{x})^2}{(n-1)}$$

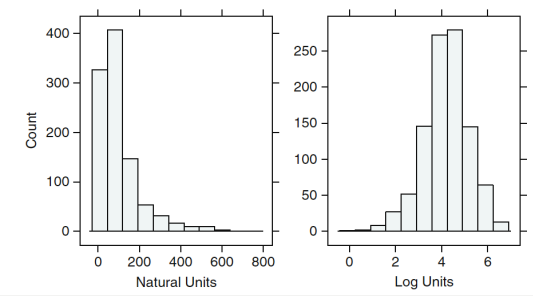
- where x is the predictor variable, n is the number of values, and x *bar* is the sample mean of the predictor.

TRANSFORMATIONS TO RESOLVE SKEWNESS – NATURAL VS. LOG UNITS



- In the natural units, the data exhibit a strong right skewness; there is a greater concentration of data points at relatively small values and small number of large values.
- For the actin filament data shown in this figure , the skewness statistic was calculated to be 2.39 while the ratio to the largest and smallest value was 870.

TRANSFORMATIONS TO RESOLVE SKEWNESS – NATURAL VS. LOG UNITS



- For the data in this figure, the right panel shows the distribution of the data once a log transformation has been applied.
- After the transformation, the distribution is not entirely symmetric but these data are better behaved than when they were in the natural units.

TRANSFORMATIONS TO RESOLVE SKEWNESS

- Box and Cox (1964) propose a family of transformations that are indexed by a parameter, denoted as λ:

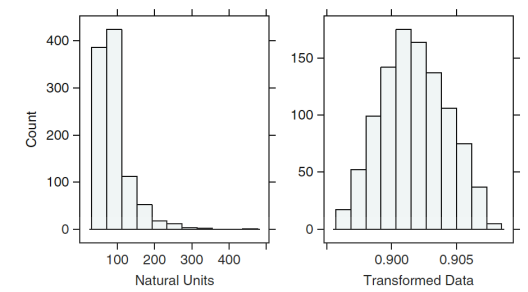
$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases}$$

- In addition to the log transformation, this family can identify square transformation (λ = 2), square root (λ = 0.5), inverse (λ = −1), and others in-between.

TRANSFORMATIONS TO RESOLVE SKEWNESS

- Box and Cox (1964) show how to use maximum likelihood estimation to determine the transformation parameter.
- This procedure would be applied independently to each predictor data that contain values greater than zero.

TRANSFORMATIONS TO RESOLVE SKEWNESS



- Another predictor, the estimated cell perimeter, had a λ estimate of −1. 1. For these data, the original and transformed values are shown in this figure.

DATA TRANSFORMATIONS FOR MULTIPLE PREDICTORS

- Of primary importance are methods to resolve outliers and reduce the dimension of the data.
  - Transformations to Resolve Outliers (Spatial sign trans.)
  - Data Reduction and (critical) Feature Extraction (PCA)

TRANSFORMATIONS TO RESOLVE OUTLIERS

- We will generally define outliers as samples that are exceptionally far from the mainstream of the data.
- Under certain assumptions, there are formal statistical definitions (a formula?) of an outlier. Even with a thorough understanding of the data, outliers can be hard to define.
- When one or more samples are suspected to be outliers, the first step is to make sure that the values are scientifically valid and that no data recording errors have occurred. (First make sure where and why the outliers come from?)
- With small sample sizes, apparent outliers might be a result of a skewed distribution where there are not yet enough data to see the skewness.
- Also, the outlying data may be an indication of a special part of the population under study that is just starting to be sampled.

TRANSFORMATIONS TO RESOLVE OUTLIERS

- There are several predictive models that are resistant to outliers.
  - Tree-based classification models create splits of the training data and the prediction equation is a set of logical statements such as "if predictor A is greater than X, predict the class to be Y," so the outlier does not usually have an exceptional influence on the model.
  - SVM for classification generally disregard a portion of the training set samples when creating a prediction equation (related to support vectors).

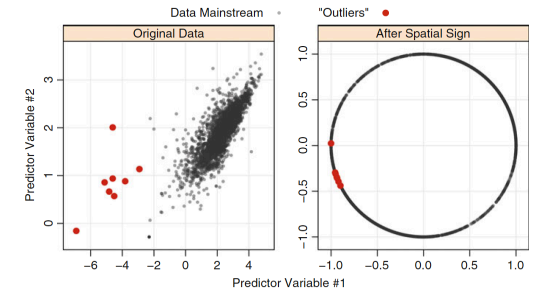
TRANSFORMATIONS TO RESOLVE OUTLIERS

- If a model is considered to be sensitive to outliers, one data transformation that can minimize the problem is the *spatial sign* (Serneels et al. 2006).
- This procedure projects the predictor values onto a multidimensional sphere.
- This has the effect of making all the samples the same distance from the center of the sphere.
- Mathematically, each sample is divided by its squared norm:

$$x_{ij}^{*} = \frac{x_{ij}}{\sum_{j=1}^{P} x_{ij}^{2}}$$

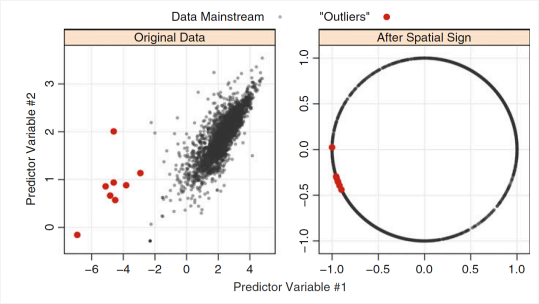
- Note that, unlike centering or scaling, this manipulation of the predictors transforms them as a group (as a whole). (So, this is a transformation for multiple predictors)

TRANSFORMATIONS TO RESOLVE OUTLIERS



- In these data, at least eight samples cluster away from the majority of other data.
- These data points are likely a valid, but poorly sampled subpopulation of the data.
- The modeler would investigate why these points are different; perhaps they represent a group of interest, such as highly profitable customers.

TRANSFORMATIONS TO RESOLVE OUTLIERS



- The spatial sign transformation is shown on the right-hand panel where all the data points are projected to be a common distance away from the origin. (The radius is 1.)
- The outliers still reside in the Southwest section of the distribution but are contracted inwards.
- This mitigates the effect of the outlying samples on model training.

DATA REDUCTION AND FEATURE EXTRACTION

- Data reduction techniques are another class of predictor transformations.
- These methods reduce the data by generating a smaller set of predictors that seek to capture a majority of the information (little information loss) in the original variables (or predictors). (k -> P, k < P)
- For most data reduction techniques, the new predictors are functions of the original predictors; therefore, all the original predictors are still needed to create the surrogate variables.
- This class of methods is often called *signal extraction* or *feature extraction* techniques.

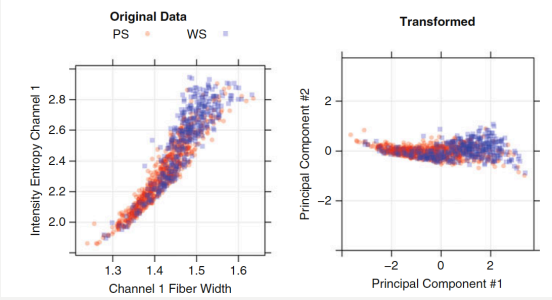
DATA REDUCTION AND FEATURE EXTRACTION

- PCA is a commonly used data reduction technique (Abdi and Williams 2010).
  - This method seeks to find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance.
- The first PC is defined as the linear combination of the predictors that captures the most variability of all possible linear combinations.
- Then, subsequent PCs are derived such that these linear combinations capture the most remaining variability while also being uncorrelated with all previous PCs. Mathematically, the jth PC can be written as:

$$\text{PC}_j = (a_{j1} \times \text{Predictor 1}) + (a_{j2} \times \text{Predictor 2}) + \cdots + (a_{jP} \times \text{Predictor } P)$$

- P is the number of predictors. The coefficients $a_{j1}, a_{j2}, \ldots, a_{jp}$ are called component weights and help us understand which predictors are most important to each PC (check whose absolute weight is higher and its sign).

DATA REDUCTION AND FEATURE EXTRACTION



- This set contains a subset of two correlated predictors, average pixel intensity of channel 1 and entropy of intensity values in the cell (a measure of cell shape), and a categorical response.
- In this example, two PCs can be derived (right plot in this figure); this transformation represents a rotation of the data about the axis of greatest variation.
- The first PC summarizes 97% of the original variability, while the second summarizes 3%.
- Hence, it is reasonable to use only the first PC for modeling since it accounts for the majority of information in the data.

DATA REDUCTION AND FEATURE EXTRACTION

- The primary advantage of PCA, and the reason that it has retained its popularity as a data reduction method, is that it creates components that are uncorrelated. (independent or orthogonal)
- Practitioners must understand that PCA seeks predictor-set variation (only) without regard to any further understanding of the predictors or to knowledge of the modeling objectives.
- Without proper guidance, PCA can generate components that summarize characteristics of the data that are irrelevant to the underlying structure of the data and also to the ultimate modeling objective.

- So, PCA is an unsupervised data reduction technique.

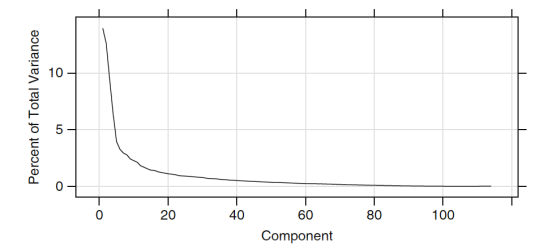
DATA REDUCTION AND FEATURE EXTRACTION – CAUTIONS AND CAVEATS

- To help PCA avoid summarizing distributional differences and predictor scale information, it is best to first transform skewed predictors and then center and scale the predictors prior to performing PCA.

- The second caveat of PCA is that it does not consider the modeling objective or response variable when summarizing variability. Because PCA is blind to the response, it is an *unsupervised technique*.

- In this case, a *supervised technique* will derive components while simultaneously considering the corresponding response.

DATA REDUCTION AND FEATURE EXTRACTION

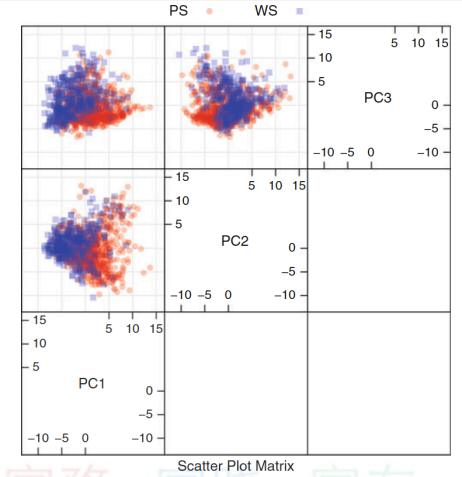


- A heuristic (i.e. rule of thumb) approach for determining the number of components to retain is to create a scree plot, which contains the ordered component number (x-axis) and the amount of summarized variability (y -axis).
- For most data sets, the first few PCs will summarize a majority of the variability, and the plot will show a steep descent; variation will then taper off for the remaining components.
- In this figure, the variation tapers off at component 5. Using this rule of thumb, four PCs would be retained.

DATA REDUCTION AND FEATURE EXTRACTION

- The first few principal components can be plotted against each other (that is a scatterplot of first few PCs) and the plot symbols can be colored by relevant characteristics, such as the class labels.
- If PCA has captured a sufficient amount of information in the data, this type of plot can demonstrate clusters of samples or outliers that may prompt a closer examination of the individual data points.
- For classification problems, the PCA plot can show potential separation of classes.

DATA REDUCTION AND FEATURE EXTRACTION

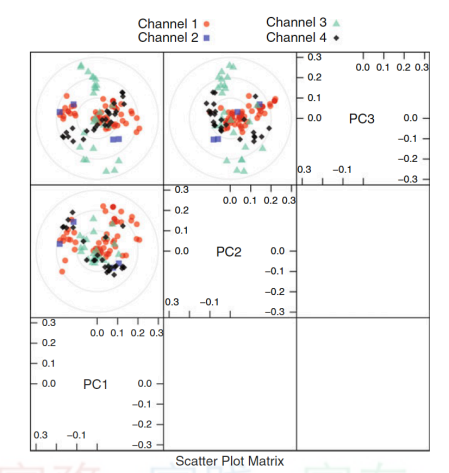


- Scatterplot matrix where the points are colored by class.
- From this plot, there appears to be some separation between the classes when plotting the first and second components.
- One conclusion to infer from this image is that the cell types are not *easily* separated.

DATA REDUCTION AND FEATURE EXTRACTION

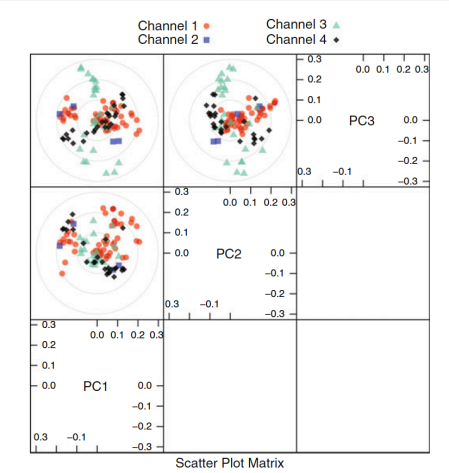
- Another exploratory use of PCA is characterizing which predictors are associated with each component.
- Recall that each component is a linear combination of the predictors and the coefficient for each predictor is called the loading.

DATA REDUCTION AND FEATURE EXTRACTION



- Loadings close to zero indicate that the predictor variable did not contribute much to that component.
- Each (2D) point corresponds to a predictor variable and is colored by the optical channel used in the experiment. (Wow, what an amazing plot !)
- For the first principal component, the loadings for the first channel (Attention to the red points !) are on the extremes.

DATA REDUCTION AND FEATURE EXTRACTION



- This indicates that cell body characteristics have the largest effect on the first principal component and by extension the predictor values.
- The majority of the loadings for the third channel (actin and tubulin) are closer to zero for the first component.
- Conversely, the third principal component is mostly associated with the third channel (actin and tubulin) while the cell body channel (channel 1) plays a minor role here.

DEALING WITH MISSING VALUES

- In many cases, some predictors have no values for a given sample.
- These missing data could be *structurally missing (i,.e. it is not random!)*, such as the number of children a man has given birth to. (Come on ! Can a man give birth to a child ?)
- In other cases, the value cannot or was not determined at the time of model building.
- It is important to understand why the values are missing. And it is important to know if the pattern of missing data is related to the outcome.
- This is called "*informative missingness*" since the missing data pattern is instructional on its own.

DEALING WITH MISSING VALUES

- Informative missingness (missing values at random) can induce significant bias in the model.
- In the introductory chapter, a short example was given regarding predicting a patient's response to a drug.
  - Suppose the drug was extremely ineffective or had significant side effects. The patient may be likely to miss doctor visits or to drop out of the study.
  - In this case, there clearly is a relationship between the probability of missing values and the treatment.
- Another is customer ratings can often have informative missingness; people are more compelled to rate products when they have strong opinions (good or bad).
  - In this case, the data are more likely to be polarized by having few values in the middle of the rating scale.

DEALING WITH MISSING VALUES – CENSORED DATA

- Missing data should not be confused with *censored* data where the exact value is missing but something is known about its value.
- For example, a company that rents movie disks by mail may use the duration that a customer has kept a movie in their models.
  - If a customer has not yet returned a movie, we do not know the actual time span, only that it is at least as long as the current duration.
  - Censored data can also be common when using laboratory measurements. Some assays cannot measure below their limit of detection.
- In such cases, we know that the value is smaller than the limit but was not precisely measured.

DEALING WITH MISSING VALUES

Are censored data treated differently than missing data?

- When building traditional statistical models focused on interpretation or inference, the censoring is usually taken into account in a formal manner by making assumptions about the censoring mechanism.

- For predictive models, it is more common to treat these data as simple missing data or use the censored value as the observed value. (missing or censored as observed or a random number in between)
  - Marked as missing
  - Actual limit can be used in place of the real value.
  - A random number between zero and the limit of detection.

DEALING WITH MISSING VALUES

- For large data sets
  - Removal of samples based on missing values is not a problem, assuming that the missingness is not informative.
- For smaller data sets
  - There is a steep price in removing samples; some of the alternative approaches described below may be more appropriate.

DEALING WITH MISSING VALUES

- If we do not remove the missing data, there are two general approaches.
  - First, a few predictive models, especially tree-based techniques, can specifically account for missing data. (Tree-based models say: I don't care about missing data)
  - Alternatively, missing data can be imputed (estimate or makeup). In this case, we can use information in the training set predictors to, in essence, estimate the values of other predictors.
- This amounts to *a predictive model within a predictive model*.

DEALING WITH MISSING VALUES

- Imputation has been extensively studied in the statistical literature, but in the context of generating correct hypothesis testing procedures in the presence of missing data.
- This is a separate problem; for predictive models we are concerned about accuracy of the predictions rather than making valid inferences.
- It is just another layer of modeling where we try to estimate values of the predictor variables based on other predictor variables.
- The most relevant scheme for accomplishing this is to use the training set to built an imputation model for each predictor in the data set.

DEALING WITH MISSING VALUES

- Prior to model training or the prediction of new samples, missing values are filled in using imputation. This extra layer of models adds uncertainty.
- If we are using resampling to select tuning parameter values or to estimate performance, the imputation should be incorporated within the resampling.
- This will increase the computational time for building models, but it will also provide honest estimates of model performance.

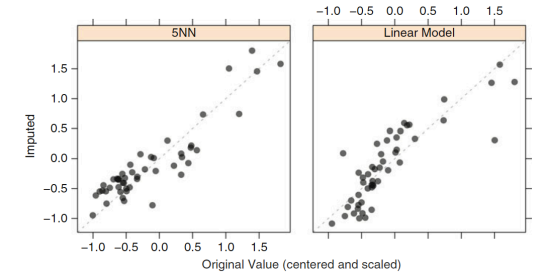
DEALING WITH MISSING VALUES

- If the number of predictors affected by missing values is small, an exploratory analysis of the relationships between the predictors is a good idea.
- For example, visualizations or methods like PCA can be used to determine if there are strong relationships between the predictors.
- If a variable with missing values is highly correlated with another predictor that has few missing values, a focused model can often be effective for imputation. (by linear regression)

DEALING WITH MISSING VALUES

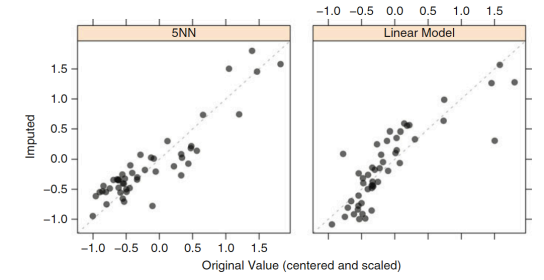
- One popular technique for imputation is a K-nearest neighbor model.
- A new sample is imputed by finding the samples in the training set "closest" to it and averages these nearby points to fill in the value.
  - One advantage of this approach is that the imputed data are confined to be within the range of the training set values.
  - One disadvantage is that the entire training set is required every time a missing value needs to be imputed.
- The number of neighbors is a tuning parameter, as is the method for determining "closeness" of two points.

DEALING WITH MISSING VALUES



- 5NN vs. Linear Model
- The left-hand panel shows the results of the 5-nearest neighbor approach. (centered and scaled)
- This imputation model does a good job predicting the absent samples; the correlation between the real and imputed values is 0.91.
- The cell fiber length, another predictor associated with cell size, has a very high correlation (0.99) with the cell perimeter data.

DEALING WITH MISSING VALUES



- We can create a simple linear regression model using these data to predict the missing values. (Use cell fiber length to predict or impute the cell perimeter)
- These results are in the right-hand panel of this figure.
- For this approach, the correlation between the real and imputed values is 0.85.

REMOVING PREDICTORS

- There are potential advantages to removing predictors prior to modeling.
  - First, fewer predictors means decreased computational time and complexity.
  - Second, if two predictors are highly correlated, this implies that they are measuring the same underlying information.
  - Third, some models can be crippled by predictors with degenerate distributions.
- In these cases, there can be a significant improvement in model performance and/or stability without the problematic variables.

REMOVING PREDICTORS – ZV OR NZV

- Consider a predictor variable that has a single unique value; we refer to this type of data as a zero variance predictor.
- For some models, such an uninformative variable may have little effect on the calculations.
  - A tree-based is impervious to this type of predictor since it would never be used in a split.
  - On the other side, linear regression would find these data problematic and is likely to cause an error in the computations.
- In either case, these data have no information and can easily be discarded.
- Similarly, some predictors might have only a handful of unique values that occur with very low frequencies. These "near-zero variance predictors" may have a single value for the vast majority of the samples.

REMOVING PREDICTORS

	#Documents
Occurrences: 0	523
Occurrences: 2	6
Occurrences: 3	1
Occurrences: 6	1

- Consider a text mining application where keyword counts are collected for a large set of documents.
- After filtering out commonly used "stop words," such as 'the' and 'of', predictor variables can be created for interesting keywords.
- Suppose a keyword occurs in a small group of documents but is otherwise unused. (More differentiated words ! It means that words occur frequently in a small set of docs, but almost disappear in other set of docs.).
- A hypothetical distribution of such a word count distribution is given in this table.

REMOVING PREDICTORS

	#Documents
Occurrences: 0	523
Occurrences: 2	6
Occurrences: 3	1
Occurrences: 6	1

- The majority of the documents (523) **do not have the keyword**; while six documents have **two occurrences**, one document has **three** and another has **six** occurrences.
- Since 98% of the data have values of zero (523/531 = 98.5%), a minority of documents might have an undue influence on the model.

REMOVING PREDICTORS

How can the user diagnose this mode of problematic data?

- First, the number of unique points in the data must be small relative to the number of samples.
- In the document example, there were 531 documents in the data set, but only four unique values, so the percentage of unique values is 0.8 % (4/531 = 0.75%).
- A small percentage of unique values is, in itself, not a cause for concern as many "dummy variables" generated from categorical predictors would fit this description.
- The problem occurs *when the frequency of these unique values is severely disproportionate*.

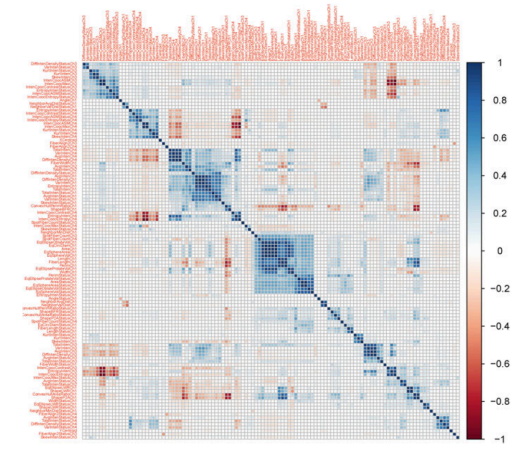
REMOVING PREDICTORS – REMOVING INDIVIDUALLY

- Given this, a rule of thumb for detecting near-zero variance predictors is:
  - The fraction of unique values over the sample size is low (say 10%). (percentUnique: the percentage of unique data points out of the total number of data points)
  - The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20). (freqRatio: the ratio of frequencies for the most common value over the second most common value)
- If both of these criteria are true and the model in question is susceptible to this type of predictor, it may be advantageous to remove the variable from the model.

BETWEEN-PREDICTOR CORRELATIONS

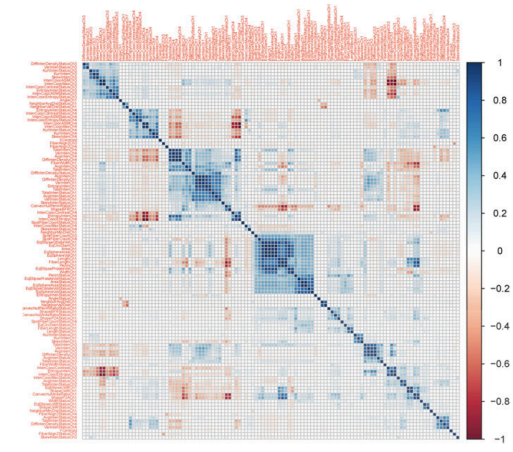
- *Collinearity* is the technical term for the situation where a pair of predictor variables have a substantial correlation with each other.
- It is also possible to have relationships between multiple predictors at once (called *multicollinearity*).
- For example, the cell segmentation data have a number of predictors that reflect the size of the cell.
  - There are measurements of the cell perimeter, width, and length as well as other, more complex calculations.
  - There are also features that measure cell morphology, such as the roughness of the cell.

BETWEEN-PREDICTOR CORRELATIONS



- Each pairwise correlation is computed from the training data and colored according to its magnitude.
- In this figure, the predictor variables have been grouped using a clustering technique (Everitt et al. 2011) so that collinear groups of predictors are adjacent to one another.
- Near the center of the diagonal is a large block of predictors from the first channel. These predictors are related to cell size, such as the width and length of the cell.

BETWEEN-PREDICTOR CORRELATIONS



- This visualization is symmetric:
  - The top and bottom diagonals show identical information.
  - Dark blue colors indicate strong positive correlations.
  - Dark red is used for strong negative correlations.
  - White implies no empirical relationship between the predictors.

BETWEEN-PREDICTOR CORRELATIONS

- In general, there are good reasons to avoid data with highly correlated predictors.
- In situations where obtaining the predictor data is costly (either in time or money), fewer variables is obviously better.
- While this argument is mostly philosophical, there are mathematical disadvantages to having correlated predictor data.
- Using highly correlated predictors in techniques like linear regression can result in highly unstable models, numerical errors, and degraded predictive performance.

BETWEEN-PREDICTOR CORRELATIONS

- Since collinear predictors can impact the variance of parameter estimates in this model, a statistic called the variance inflation factor (VIF) can be used to identify predictors that are impacted (Myers 1994).
- Beyond linear regression, this method may be inadequate for several reasons:
  - It was developed for linear models, it requires more samples than predictor variables
  - While it does identify (only) collinear predictors, it does not determine which should be removed to resolve the problem.
- A less theoretical, more heuristic approach to dealing with this issue is to remove the minimum number of predictors to ensure that all pairwise correlations are below a certain threshold.

BETWEEN-PREDICTOR CORRELATIONS

- While this method only identify collinearities in two dimensions, it can have a significantly positive effect on the performance of some models.
- The algorithm is as follows:
  1. Calculate the correlation matrix of the predictors.
  2. Determine the two predictors associated with the largest absolute pairwise correlation (call them predictors A and B).
  3. Determine the average correlation between A and the other variables. Do the same for predictor B .
  4. If A has a larger average correlation, remove it; otherwise, remove predictor B. (remove predictor, A and B, having higher correlation with other predictors)
  5. Repeat Steps 2–4 until no absolute correlations are above the threshold.
- The idea is to first remove the predictors that have the most correlated relationships.

BETWEEN-PREDICTOR CORRELATIONS

- As previously mentioned, feature extraction methods (e.g., principal components) are another technique for mitigating the effect of strong correlations between predictors.
- These techniques make the connection between the predictors and the outcome more complex.
- Since signal extraction methods are usually unsupervised, there is no guarantee that the resulting surrogate predictors have any relationship with the outcome.

ADDING PREDICTORS

- When a predictor is categorical, such as gender or race, it is common to decompose the predictor into a set of more specific variables.
- Usually, each category get its own dummy variable that is a zero/one indicator for that group.
- Many of the models described in this text automatically generate highly complex, nonlinear relationships between the predictors and the outcome.
- More simplistic models do not be produced unless the user manually specifies which predictors should be nonlinear and in what way.

ADDING PREDICTORS

		Dummy variables				
Value	n	<100	100–500	500–1,000	>1,000	Unknown
<100 DM	103	1	0	0	0	0
100–500 DM	603	0	1	0	0	0
500–1,000 DM	48	0	0	1	0	0
>1,000 DM	63	0	0	0	1	0
Unknown	183	0	0	0	0	1

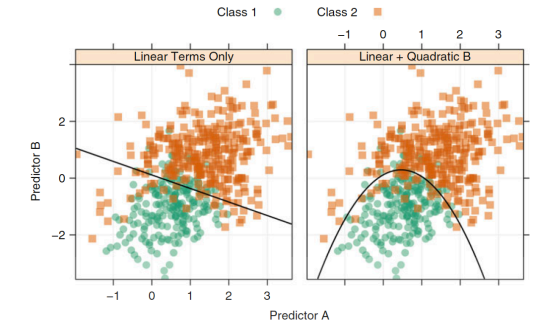
- Only four dummy variables are needed here; once you know the value of four of the dummy variables, the fifth can be inferred. (i.e. they are linearly dependent)
- Models that include an intercept term, such as simple linear regression (Sect. 6.2), would have numerical issues if each dummy variable was included in the model.

ADDING PREDICTORS

		Dummy variables				
Value	n	<100	100–500	500–1,000	>1,000	Unknown
<100 DM	103	1	0	0	0	0
100–500 DM	603	0	1	0	0	0
500–1,000 DM	48	0	0	1	0	0
>1,000 DM	63	0	0	0	1	0
Unknown	183	0	0	0	0	1

- The reason is that, for each sample, these variables all add up to one and this would provide the same information as the intercept.
- If the model is insensitive to this type of issue (eg. tree-like models), using the complete set of dummy variables would help improve interpretation of the model.

ADDING PREDICTORS



- The left-hand panel shows the basic logistic regression classification boundaries when the predictors are added in the usual (linear) manner.
- The right-hand panel shows a logistic model with the basic linear terms and an additional term with the square of predictor B.

ADDING PREDICTORS

- Since logistic regression is a well-characterized and stable model, using this model with some additional nonlinear terms may be preferable to highly complex techniques (which may overfit).
- For classification models, they calculate the "class centroids," which are the centers of the predictor data for each class.
- Then for each predictor, the distance to each class centroid can be calculated and these distances can be added to the model.

BINNING PREDICTORS

- While there are recommended techniques for pre-processing data, there are also methods to *avoid*.
- One common approach to simplifying a data set is to take a numeric predictor and pre-categorize or "bin" it into two or more groups prior to data analysis.
- There are many issues with the manual binning of continuous data.
  - First, there can be a significant loss of performance in the model.
  - Second, there is a loss of precision in the predictions when the predictors are categorized.
  - Third, research has shown (Austin and Brunner 2004) that categorizing predictors can lead to a high rate of false positives. (high FP rate)

BINNING PREDICTORS

- For example, Bone et al. (1992) define a set of clinical symptoms to diagnose Systemic Inflammatory Response Syndrome (SIRS).
  - SIRS can occur after a person is subjected to some sort of physical trauma (e.g., car crash).
- A simplified version of the clinical criteria for SIRS are:
  - Temperature less than 36^{o}C or greater than 38^{o}C.
  - Heart rate greater than 90 beats per minute.
  - Respiratory rate greater than 20 breaths per minute.
  - White blood cell count less than 4,000 cells/mm^3 or greater than 12,000 cells/mm^3.

BINNING PREDICTORS

- The perceived advantages to this approach are:
  - The ability to make seemingly simple statements, either for sake of having a simple decision rule (as in the SIRS example) or the belief that there will be a simple interpretation of the model.
  - The modeler does not have to know the exact relationship between the predictors and the outcome.
  - A higher response rate for survey questions where the choices are binned.
- For example, asking the date of a person's last tetanus shot is likely to have fewer responses than asking for a range (e.g., in the last 2 years, in the last 4 years).

BINNING PREDICTORS

- Unfortunately, the predictive models that are most powerful are usually the least interpretable.
- The bottom line is that the perceived improvement in interpretability gained by manual categorization is usually offset by a significant loss in performance.
- Since this book is concerned with predictive models (where interpretation is not the primary goal), loss of performance should be avoided.

BINNING PREDICTORS

- The argument here is related to the manual categorization of predictors prior to model building.
- There are several models, such as classification / regression trees and multivariate adaptive regression splines, that estimate cut points (automatically) in the process of model building.
- The difference between these methodologies and manual binning is that the models use all the predictors to derive bins based on a single objective (such as maximizing accuracy).
- They evaluate many variables simultaneously and are usually based on statistically sound methodologies.

THANK YOU