# Predictive Modeling

*Department of Mathematical Sciences*
*Qiuying Sha, Professor*

Michigan Tech

# Chapter 3. Data Pre-processing

- Data pre-processing techniques refer to the addition, deletion, or transformation of data set.

- This chapter outlines approaches to ***unsupervised* data processing**: **the outcome variable is not considered** by the pre-processing techniques.

- In other chapters, ***supervised* methods**, where **the outcome is utilized to pre-process the data**, are also discussed.

# 3.1 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 **expert knowledge** of the cell type and characteristics.
- Another way to measure the cell characteristics is by **using high-content screening** (Giuliano et al. 1997)
  - A sample is first **dyed** with a substance that will bind to the desired characteristic of the cells. The **light scattering measurements** are then processed through **imaging software** to **quantify the desired cell characteristics.**

- Hill et al. (2007) observed that the imaging software used to determine the location and shape of the cell had **difficulty segmenting cells** (i.e., defining cells' boundaries).

  - In these images, the **bright green** boundaries identify the cell **nucleus,** while the **blue boundaries** define the **cell perimeter.**
  - Clearly some cells are ***well segmented***, while others are **poorly segmented**.

- Hill et al. (2007) assembled a data set consisting of 2,019 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.

  - For all cells, 116 features (e.g., cell area, spot fiber count) were measured and were used to predict the segmentation quality of cells.

- In this chapter, we will use training set samples to demonstrate data preprocessing techniques.

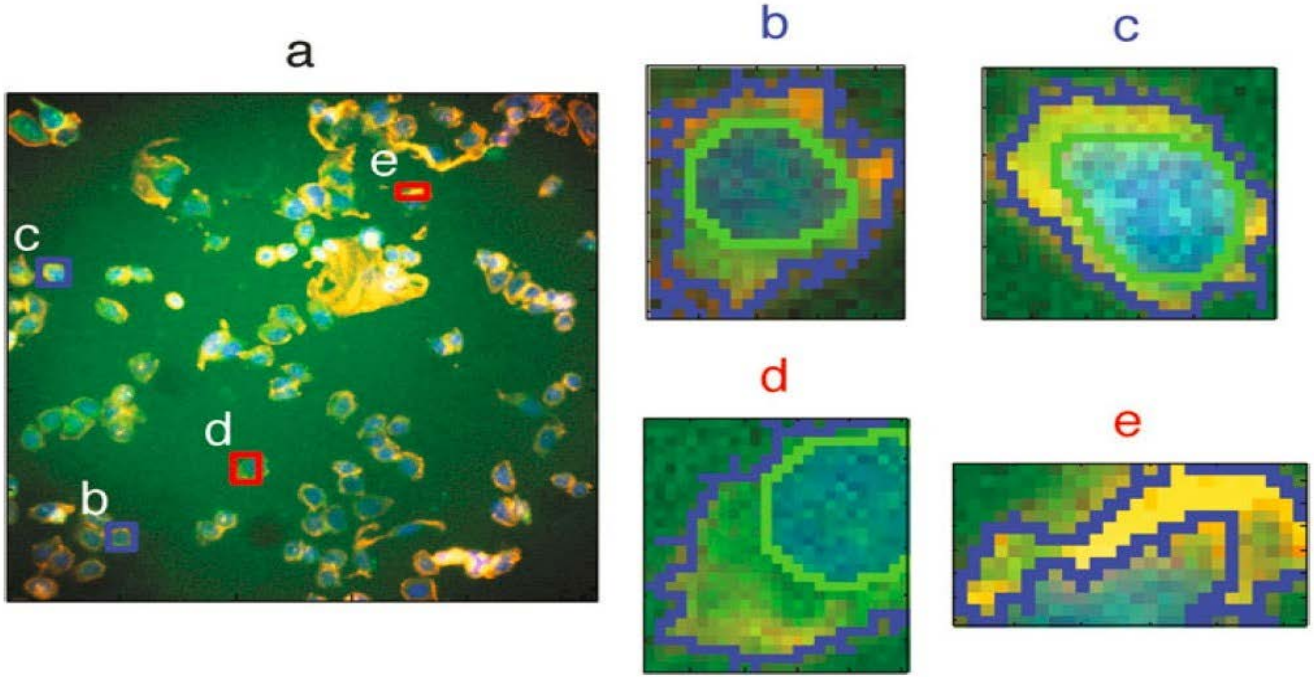



Fig. 3.1: An image showing cell segmentation from Hill et al. (2007). The *red boxes* [panels (**d**) and (**e**)] show poorly segmented cells while the cells in the *blue boxes* are examples of proper segmentation

## 3.2 Data Transformations for Individual Predictors

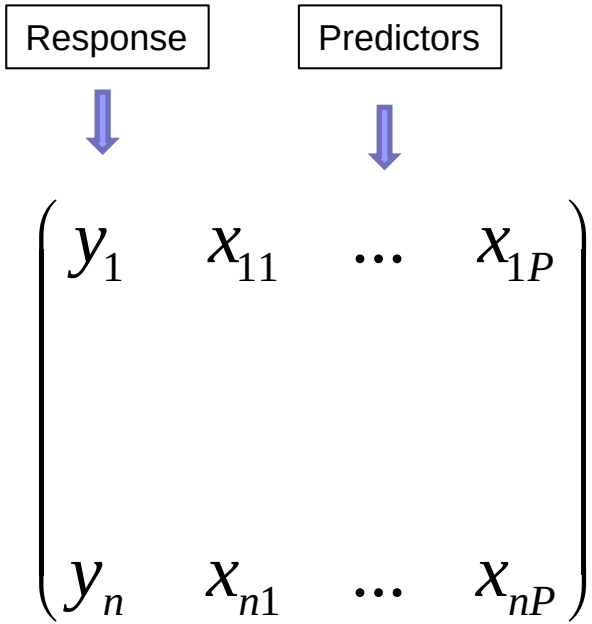
We will discuss centering, scaling, and skewness transformations.

- ***Centering and Scaling***
  - To center a predictor variable, the average predictor value is subtracted from all the values. As a result of centering, the predictor will have a zero mean.
  - To scale the data, each value of the predictor variable is divided by its standard deviation. Scaling the data coerce the values to have a common standard deviation of one.

Some models, such as PLS, benefit from the predictors being on a common scale.

The only real **downside is a loss of interpretability** of the individual values since the data are **no longer in the original units**

Data can be represented using a matrix:



Response

Predictors

$$\begin{pmatrix} y_1 & x_{11} & \cdots & x_{1P} \\ & & & \\ & & & \\ y_n & x_{n1} & \cdots & x_{nP} \end{pmatrix}$$

Sample mean $\quad \bar{y} \quad \bar{x}_1 \quad \bar{x}_P$

Sample variance $\quad S_y^2 \quad S_1^2 \quad S_P^2$

$$\begin{pmatrix} y_1 & x_{11} & \cdots & x_{1P} \\ & & & \\ y_n & x_{n1} & \cdots & x_{nP} \end{pmatrix} \xrightarrow{\text{Center}} \begin{pmatrix} y_1 - \bar{y} & x_{11} - \bar{x}_1 & \cdots & x_{1P} - \bar{x}_P \\ & & & \\ y_n - \bar{y} & x_{n1} - \bar{x}_1 & \cdots & x_{nP} - \bar{x}_P \end{pmatrix}$$

$$\xrightarrow{\text{Scale}} \begin{pmatrix} y_1 / S_y & x_{11} / S_1 & \cdots & x_{1P} / S_P \\ & & & \\ y_n / S_y & x_{n1} / S_1 & \cdots & x_{nP} / S_P \end{pmatrix}$$

- ***Transformations to Resolve Skewness***

**Diagnose skewness:**

➢ The skewness statistic

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

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

✓ If roughly symmetric, the skewness is around 0;

✓ If right skewed, the skewness is positive.

✓ If left skewed, the skewness becomes negative.

## How can you interpret the skewness number?

Bulmer, M. G., *Principles of Statistics* (Dover, 1979) — suggests this rule of thumb:

- ✓ If skewness is less than −1 or greater than +1, the distribution is **highly skewed**.
- ✓ If skewness is between −1 and −½ or between +½ and +1, the distribution is **moderately skewed**.
- ✓ If skewness is between −½ and +½, the distribution is **approximately symmetric.**

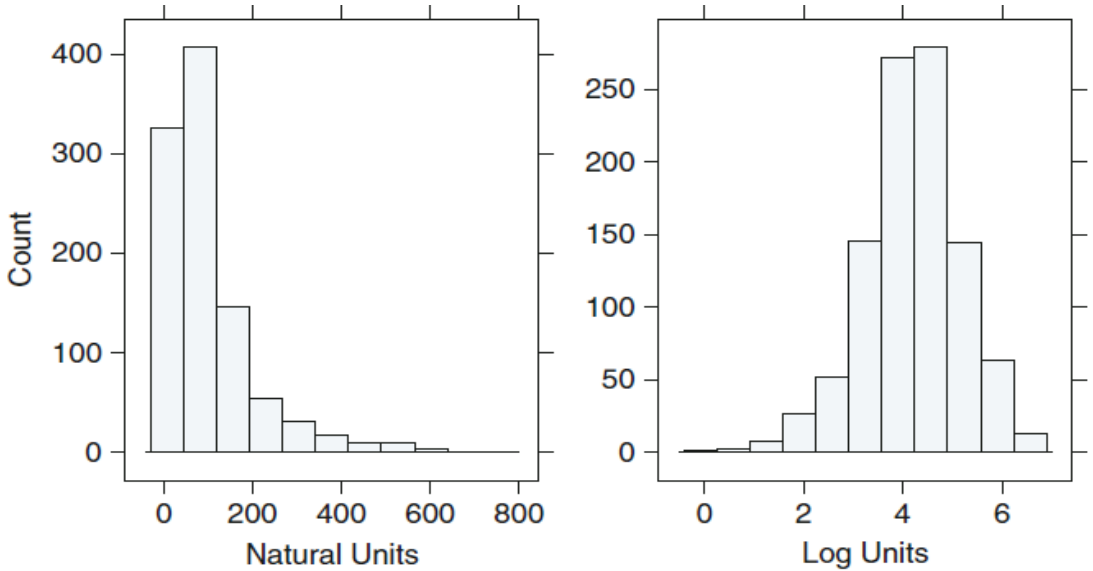



Fig. 3.2: *Left*: a histogram of the standard deviation of the intensity of the pixels in actin filaments. This predictor has a strong right skewness with a concentration of points with low values. For this variable, the ratio of the smallest to largest value is 870 and a skewness value of 2.39. *Right*: the same data after a log transformation. The skewness value for the logged data was −0.4

**Most common transformation to remove skewness**: log, square root, or inverse.

**Alternatively, statistical methods can be used to empirically identify an appropriate transformation**.

➢ Box and Cox (1964) propose a *family* of transformations indexed by $\lambda$:

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

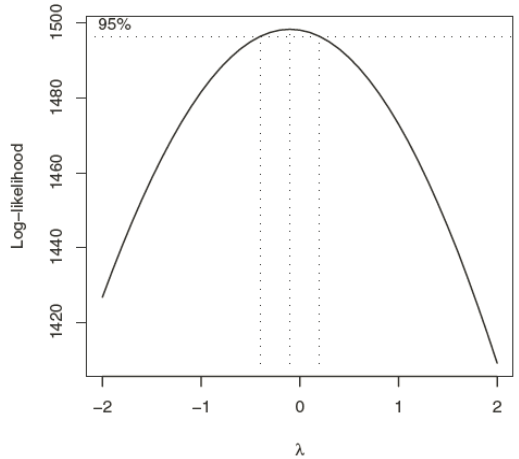
✓ Using the training data, $\lambda$ can be estimated using maximum likelihood Estimation (MLE).

✓ R: boxcox {MASS} – function boxcox is in MASS package. boxcox(object, ...)

$$\begin{pmatrix} x_1 \\ \\ x_n \end{pmatrix} \Rightarrow \begin{pmatrix} (x_1^{\lambda} - 1)/\lambda \\ \\ (x_n^{\lambda} - 1)/\lambda \end{pmatrix} \text{ for } \lambda \neq 0;$$

$$\begin{pmatrix} x_1 \\ \\ x_n \end{pmatrix} \Rightarrow \begin{pmatrix} \log x_1 \\ \\ \log x_n \end{pmatrix} \text{ for } \lambda = 0$$

**Exhibit 5.11 Log-likelihood versus Lambda**




```
> BoxCox.ar(electricity)
```

In this figure, the 95% confidence interval for λ contains the value of λ = 0 quite near its center (the estimated transformation value is 0.1)

This figure strongly suggests a logarithmic transformation (λ = 0) for these data.

For the segmentation data,

- 58 predictors are categorical variables,
- 11 predictors do not have available *λ* values,
- The remaining 47 predictors can be transmitted using Box-Cox.

## 3.3 Data Transformations for Multiple Predictors

Act on **groups of predictors**, typically the entire set under consideration.

To **resolve outliers** and **reduce the dimension** of the data.

- ***Transformations to Resolve Outliers***

  Outliers are exceptionally far from the mainstream of the data.

  Under certain assumptions, there are formal statistical definitions of an outlier.

  When one or more samples are suspected to be outliers,

  - The first step is to make sure that the values are scientifically valid and
  - no data recording errors.

Great care should be taken not to hastily remove or change values, especially if the sample size is small.

With small sample sizes,

- Outliers might be a result of a **skewed distribution** where there are not yet enough data to see the skewness.
- A “cluster” of valid points that reside outside the mainstream of the data might belong to a different population than the other samples.

Some models are resistant to outliers, e.g. tree-based classification, support vector machines for classification.

Some models are sensitive to outliers, e. g. linear model.

**Transformation on the data to minimize the effects of outliers**: ***spatial sign*** (Serneels et al. 2006).

Project the predictor values onto a multidimensional sphere.

- First, center and scale the predictor data,
- Then each sample is divided by its norm:

Spatial sign transformation is to transform $x_{ij}$ to $x_{ij}^*$ , where

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

$$\begin{pmatrix} x_{11} & \cdots & x_{1P} \\ \cdot & \cdot & \cdot \\ x_{n1} & \cdots & x_{nP} \end{pmatrix}$$

Calculate

$$\begin{pmatrix} \sum_{j=1}^{P} x_{1j}^2 \\ \cdot \\ \sum_{j=1}^{P} x_{nj}^2 \end{pmatrix}$$

Distance to the origin (0,…,0) is 1

Spatial sign transformation

$$\begin{pmatrix} x_{11} / \sqrt{\sum_{j=1}^{P} x_{1j}^2} & \cdots & x_{1P} / \sqrt{\sum_{j=1}^{P} x_{1j}^2} \\ \cdot & \cdots & \cdot \\ x_{n1} / \sqrt{\sum_{j=1}^{P} x_{nj}^2} & \cdots & x_{nP} / \sqrt{\sum_{j=1}^{P} x_{nj}^2} \end{pmatrix}$$

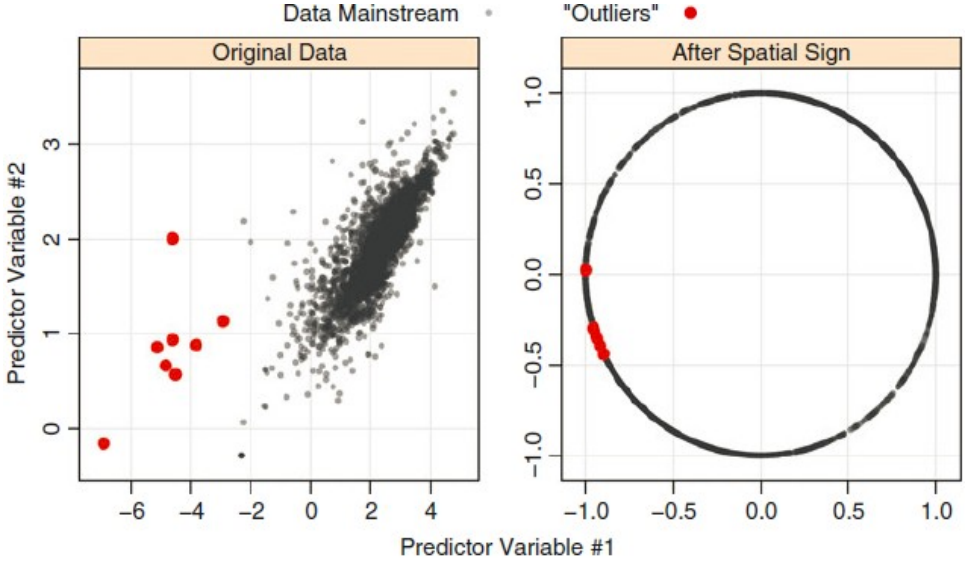



Fig. 3.4: *Left*: An illustrative example with a group of outlying data points. *Right*: When the original data are transformed, the results bring the outliers towards the majority of the data

**Caution:** Removing predictor variables after applying the spatial sign transformation may be problematic.

- ***Data Reduction and Feature Extraction***

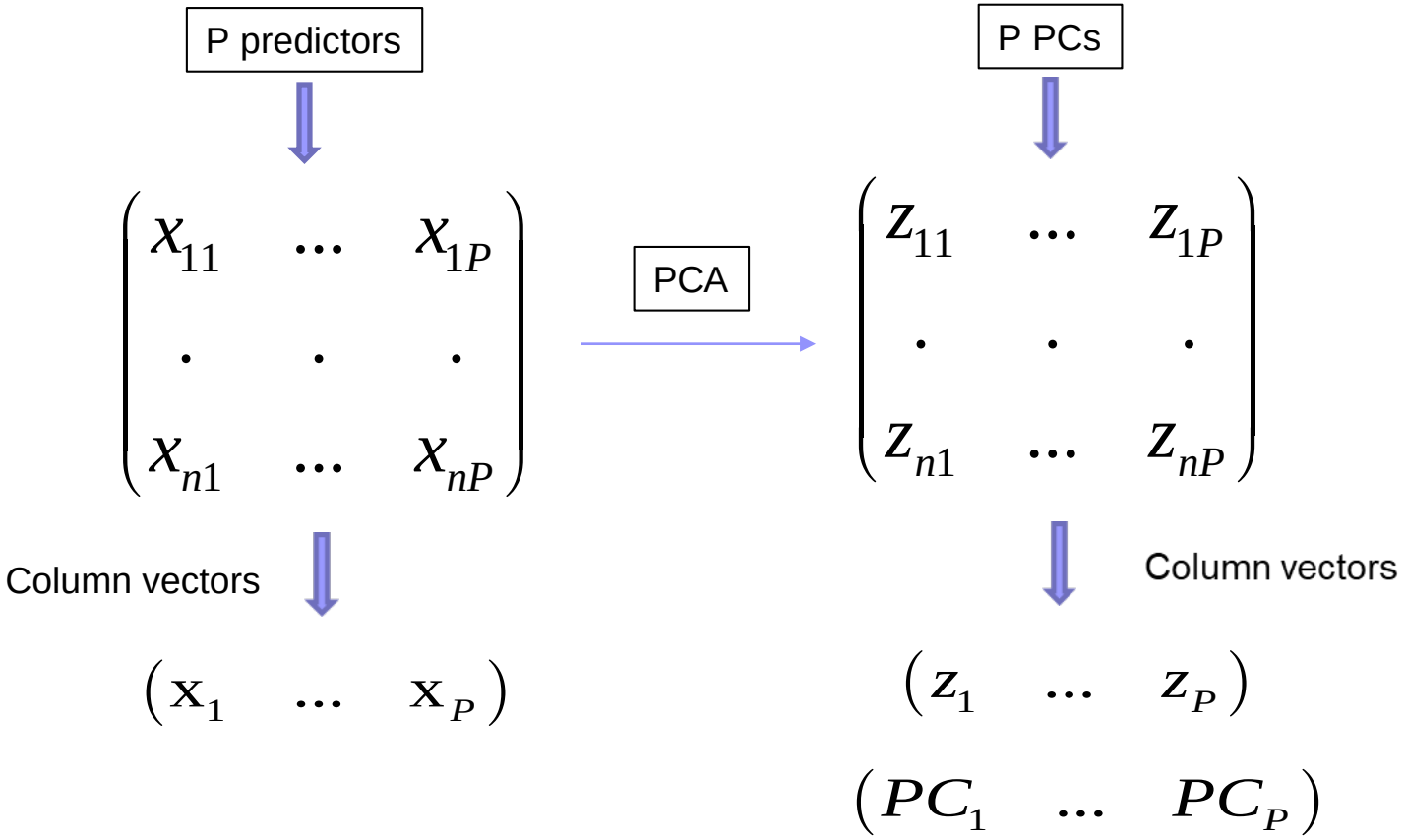
  Data reduction techniques are another class of predictor transformations.

  **Data reduction**: reduce the data by generating a smaller set of predictors that seek to capture a majority of the information in the original variables.

  E. g.: *signal extraction* or *feature extraction* techniques.

  **Principal Component Analysis (PCA)** is a commonly used data reduction technique. See the handout from the link:

  http://www.iasri.res.in/ebook/EB_SMAR/e-book_pdf%20files/Manual%20II/9-data_reduction.pdf


P predictors
$\begin{pmatrix} x_{11} & \cdots & x_{1P} \\ . & . & . \\ x_{n1} & \cdots & x_{nP} \end{pmatrix}$
PCA
P PCs
$\begin{pmatrix} z_{11} & \cdots & z_{1P} \\ . & . & . \\ z_{n1} & \cdots & z_{nP} \end{pmatrix}$
Column vectors
$\begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_P \end{pmatrix}$
Column vectors
$\begin{pmatrix} z_1 & \cdots & z_P \end{pmatrix}$
$\begin{pmatrix} PC_1 & \cdots & PC_P \end{pmatrix}$

P predictors

$$\begin{pmatrix} x_{11} & \cdots & x_{1P} \\ . & . & . \\ x_{n1} & \cdots & x_{nP} \end{pmatrix}$$

Variance-covariance matrix

$$\Sigma$$

$$\begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1P}^2 \\ . & . & . \\ \sigma_{P1}^2 & \cdots & \sigma_{PP}^2 \end{pmatrix}$$

Eigen vectors

$$a_1, \ldots, a_P$$

Eigen values

$$\lambda_1, \lambda_2, \ldots, \lambda_P$$

where $a_1 = \begin{pmatrix} a_{11} \\ . \\ a_{P1} \end{pmatrix}, \ldots, a_P = \begin{pmatrix} a_{P1} \\ . \\ a_{PP} \end{pmatrix}$

P PCs

Predictor matrix

Eigen vector matrix

$$\begin{pmatrix} z_{11} & \cdots & z_{1P} \\ . & . & . \\ z_{n1} & \cdots & z_{nP} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1P} \\ . & . & . \\ x_{n1} & \cdots & x_{nP} \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{P1} \\ . & . & . \\ a_{1P} & \cdots & a_{PP} \end{pmatrix}$$

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

**Facts about PCA**:

$$Var(\mathrm{x}_1)+\ldots+Var(\mathrm{x}_P)=Var(\mathrm{z}_1)+\ldots+Var(\mathrm{z}_P)$$

$$Var(\mathrm{z}_1)=\lambda_1,\ldots,Var(\mathrm{z}_P)=\lambda_P,\ \lambda_1\geq\lambda_2,\ldots,\geq\lambda_P$$

$\mathrm{cov}(\mathrm{z}_i,\mathrm{z}_j)=0$ for $i\neq j$. PCs are unrelated.

**PCA:**

- Find linear combinations of the predictors, known as principal components (PCs), which capture the most possible variance.
- Mathematically, the *j*th PC can be written as:

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

where *P* is the number of predictors, $a_{j1}, a_{j2}, \ldots, a_{jP}$ are called component weights and help us understand which predictors are most important to each PC.

In Figure 3.5,

- Two predictors are high correlation (0.93).
- The first PC summarizes 97% of the original variability.
- It is reasonable to use only the first PC for modeling since it accounts for the majority of information in the data.

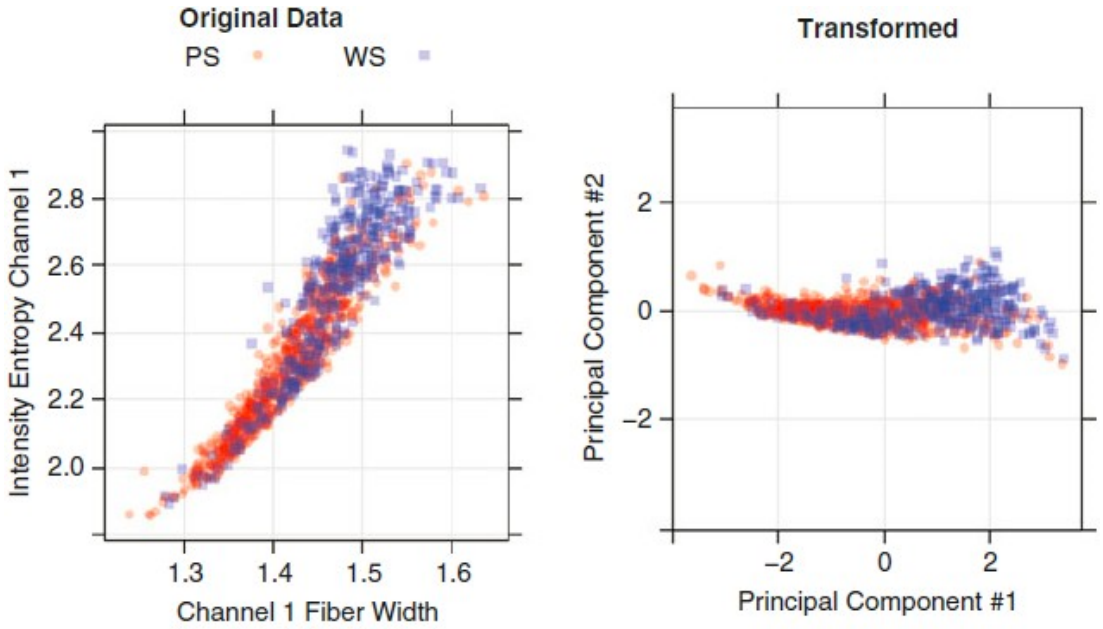



Fig. 3.5: An example of the principal component transformation for the cell segmentation data. The *shapes* and *colors* indicate which cells were poorly segmented or well segmented

- The **primary advantage of PCA**: it creates components that are **uncorrelated**.
- PCA does not consider the response variable (***unsupervised technique***).
- To use PCA, we need to

  ➢ First transform skewed predictors, and then center and scale the predictors prior to performing PCA.

  ➢ Decide how many components to retain in PCA.

- How to decide # of components to retain?

  ➢ A heuristic approach: a scree plot containing the ordered component number (*x*- axis) and the amount of summarized variability (*y*-axis) (Fig. 3.6).

  The component number prior to the tapering off of variation is the maximal component that is retained.

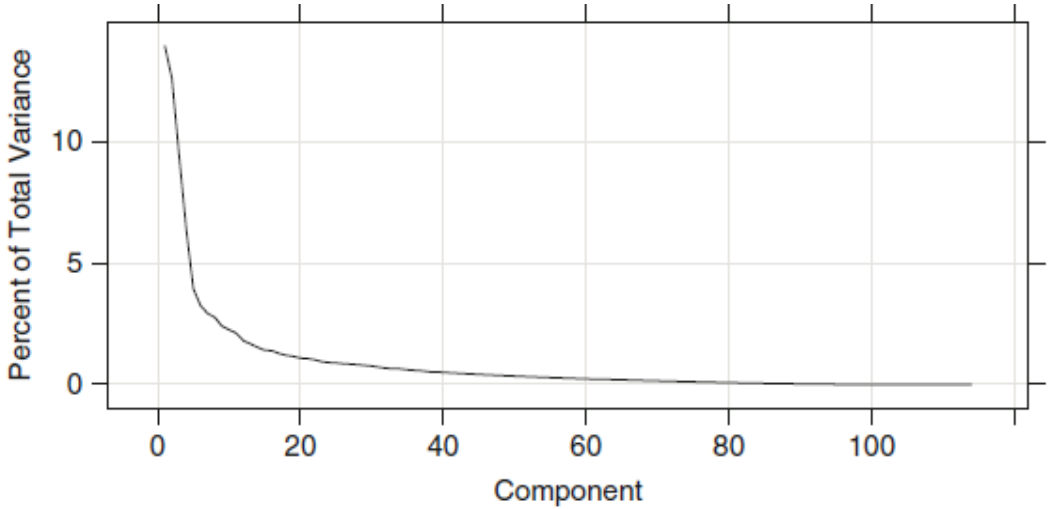



Fig. 3.6: A "scree plot" where the percentage of the total variance explained by each component is shown

In this figure, the variation tapers off at component 5. Using this rule of thumb, four PCs would be retained.

- Choose k components such that the total variation explained by the first k components is greater than C.

$$\frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_P} \geq C$$

where $\lambda_1, ..., \lambda_P$ are the eigenvalues of the variance-covariance matrix of the predictors X, C is the cut-off value.

- In an automated model building process, the optimal number of components can be determined by cross-validation (consider this as a tuning parameter)

- Visually examining the PCs: plot the first few PCs against each other,
  - ✓ If PCA has captured a sufficient amount of information in the data, this type of plot can demonstrate clusters of samples or outliers.
  - ✓ If there is little clustering of the classes, the plot of the PCs will show a significant overlap of the points for each class.

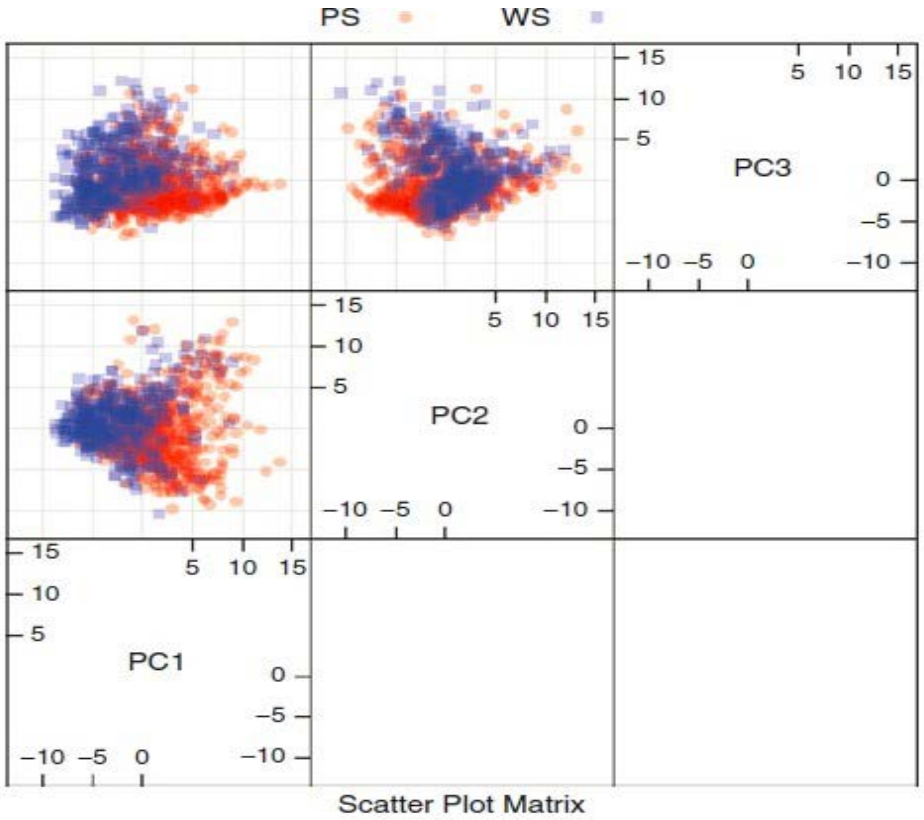



Fig. 3.7: A plot of the first three principal components for the cell segmentation data, colored by cell type

**Figure 3.7**:

- Since the percentages of variation explained are not large for the first three components, it is important not to over-interpret the resulting image.
- There appears to be some separation between the classes when plotting the first and second components.
- However, the distribution of the well-segmented cells is roughly contained within the distribution of the poorly identified cells.
- One conclusion to infer from this image is that the cell types are not *easily* separated.
- However, this does not mean that other models will reach the same conclusion.

Another exploratory use of PCA is characterizing which predictors are associated with each component.

- Loadings close to zero indicate that the predictor variable did not contribute much to that component.

## 3.4 Dealing with Missing Values

- Understand *why* the values are missing
- **Informative missingness** -- the pattern of missing data is related to the outcome.
  - ➢ Remove the predictor or sample can induce significant bias in the model.
- If the **percentage of missing data** is large, we can remove this predictor or sample from subsequent modeling activities.
  - ➢ For large data sets, removal of samples based on missing values is not a problem, assuming that the missingness is not informative.
  - ➢ In smaller data sets, there is a steep price in removing samples.

There are two general approaches to deal with missing:

- First, a few predictive models, especially tree-based techniques, can specifically account for missing data.

- Alternatively, missing data can be **imputed**.

  In this case, we can use information in the training set predictors to estimate the values of other predictors.

  Imputation is just another layer of modeling.

  ✓ We try to estimate values of the predictor variables based on other predictor variables.

One popular technique for imputation is a *K*-nearest neighbor (KNN) 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.
- **The number of neighbors** is a **tuning parameter**.
- Troyanskaya et al. (2001) found the nearest neighbor approach to **be fairly robust** to the tuning parameters.

## For example:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ NaN & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{pmatrix}$$

We calculate the distances using the 2nd, 3rd and 4th columns

Using 2nd, 3rd, and 4th predictors to calculate the distance matrix.

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ & d_{22} & d_{23} & d_{24} \\ & & d_{33} & d_{34} \\ & & & d_{44} \end{pmatrix}$$

We can have $d_{12}, d_{23}, \text{and } d_{24}$

If we use 2-NN and $d_{12}$ and $d_{24}$ are the smallest, then we use

$$(x_{11} + x_{41}) / 2$$

to impute the missing value.

Michigan Tech

## 3.5 Removing Predictors

Advantages to removing predictors:

- Fewer predictors mean decreased computational time and complexity.
- Removing predictors may lead to a more parsimonious and interpretable model.
- Removing predictors amy improve in model performance and/or stability without the problematic variables.

***Zero variance predictor***: a predictor variable with a single unique value.

- A tree-based model (Sects. 8.1 and 14.1) is impervious to this type of predictor.
- Linear regression would find these data problematic and is likely to cause an error in the computations

***Near-zero variance predictors:*** predictors have only a handful of unique values that occur with very low frequencies.

Table 3.1: A predictor describing the number of documents where a keyword occurred

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

- Since 98% of the data have values of zero, a minority of documents might have an undue influence on the model.

- Also, if any resampling is used, there is a strong possibility that one of the resampled data sets will only contain documents without the keyword, so this predictor would only have one unique value.

How to diagnose this mode of problematic data?

A rule of thumb to detect near-zero variance predictors is:

- the fraction of unique values over the sample size is low (say 10 %; 4/531 = 0.8%) and

- the ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20; 523/6 = 87).

***Between-Predictor Correlations***

- *Collinearity (multicollinearity)*: a pair of (multiple) predictor variables have a substantial correlation with each other.
- In the cell segmentation data, there are a number of predictors that reflect the size of the cell, the cell perimeter, width, and length.

***How to diagnose multicollinearity:***

- Correlation plot

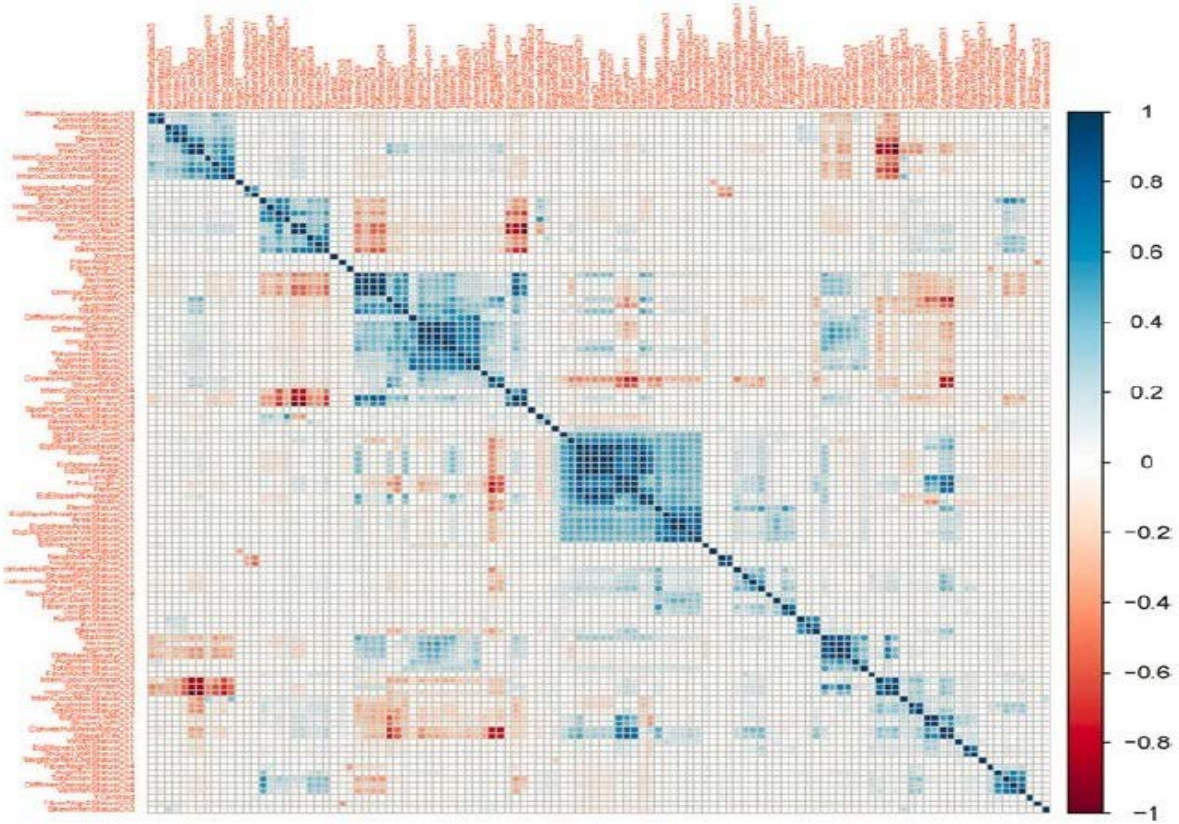



Fig. 3.10: A visualization of the cell segmentation correlation matrix. The order of the variables is based on a clustering algorithm

In figure 3.10,

- 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.
- There are blocks of strong positive correlations that indicate “clusters” of collinearity.

PCA can be used to characterize the magnitude of the problem.

For example, if the first principal component accounts for a large percentage of the variance, this implies that there is at least one group of predictors that represent the same information.

- Classical regression analysis has several tools to diagnose multicollinearity for linear regression.

  The variance inflation factor (VIF) can be used to identify predictors that are impacted (Myers 1994) (See supplemental file: lecture3-sup-VIF).

  - Beyond linear regression, this method may be inadequate.
  - It identifies collinear predictors, it does not determine which should be removed to resolve the problem.

***How to handle collinearity (multicollinearity)*:**

- Remove the minimum number of predictors to ensure that all pairwise correlations are below a certain threshold.
  - ➢ This method only identify collinearities in two dimensions.
  - ➢ The algorithm is as follows:

1) Calculate the correlation matrix of the predictors.

1) Determine the two predictors associated with the largest absolute pairwise correlation (call them predictors *A* and *B*).

1) Determine the average correlation between *A* and the other variables. Do the same for predictor *B*.

1) If *A* has a larger average correlation, remove it; otherwise, remove predictor *B*.

1) Repeat Steps 2–4 until no absolute correlations are above the threshold.

Predictors

$$\begin{pmatrix} x_{11} & \cdots & x_{1P} \\ . & . & . \\ x_{n1} & \cdots & x_{nP} \end{pmatrix}$$

Correlation Matrix

$$\begin{pmatrix} r_{11} & \cdots & r_{1P} \\ . & . & . \\ r_{P1} & \cdots & r_{PP} \end{pmatrix}$$

Find the largest value among

$$r_{ij}, i = 1,...,P; j = 1,...,P; \text{ and } i \neq j$$

For example, if the largest absolute value is $r_{24}$

If $|r_{24}| < C$

Keep all predictors

If $|r_{24}| \geq C$

Calculate the average correlations between each of the two predictors with other predictors:

$$U = (r_{21} + r_{23} + r_{25} + ... + r_{2P}) / (P-2)$$

and

$$V = (r_{41} + r_{43} + r_{45} + ... + r_{4P}) / (P-2)$$

If $|U| > |V|$, delete the 2nd predictor;

If $|U| < |V|$, delete the 4th predictor.

Repeat till all absolute correlations are less than C.

- Apply a threshold of 0.75 to the cell segmentation data, this algorithm would suggest removing 43 predictors.

- Feature extraction methods (e.g., principal components) are another technique for mitigating the effect of strong correlations between predictors. However, these techniques make the connection between the predictors and the outcome more complex (unsupervised).

## 3.6 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

Table 3.2: A categorical predictor with five distinct groups from the credit scoring case study. The values are the amount in the savings account (in Deutsche Marks)

| | | 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 categories are re-encoded into smaller bits of information called “dummy variables.”

- Each category gets its own dummy variable (a zero/one indicator for each group).

## 3.7 Binning Predictors

- **Binning predictor:** take a numeric predictor and pre-categorize or “bin” it into two or more groups prior to data analysis.
- The perceived advantages to this approach are:
  - The ability to make seemingly simple statements, either for sake of having a simple decision rule 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.

- There are many issues with the manual binning of continuous data.
  - There can be a significant loss of performance in the model.
  - There is a loss of precision in the predictions when the predictors are categorized.
  - Categorizing predictors can lead to a high rate of false positives.

Since this course is concerned with predictive models (where interpretation is not the primary goal), loss of performance should be avoided.

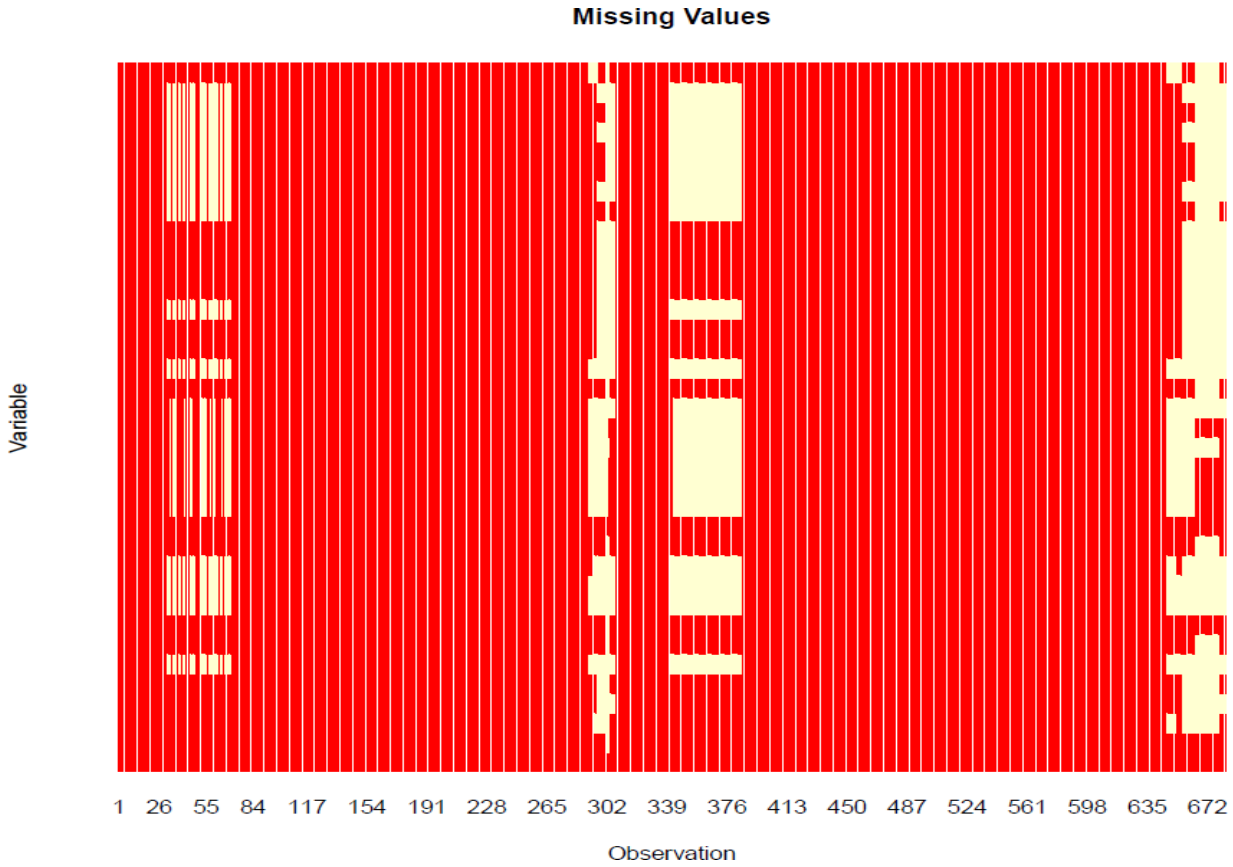
## In Example 12.3: There are four predictors that are not numeric. You need to add dummy variables (dummyVar in R) for these predictors.

Create dummy variables for state, area_code, international_plan, voice_mail_plan

```
library(caret)
dummRes <- dummyVars("~state+area_code+international_plan+voice_mail_plan",  data=data name,
                     fullRank=TRUE)
Add_dumm <- data.frame(predict(dummRes, newdata=data name))
```

## In Exercise 3.2: Get an imaging for missing data

```
image(is.na(Soybean), main = "Missing Values", xlab = "Observation", ylab = "Variable",  xaxt = "n", yaxt =
"n", bty = "n")
axis(1, seq(0, 1, length.out = nrow(Soybean)), 1:nrow(Soybean), col = "white")
```




Michigan Tech