Data pre-processing

Michela Mulas

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for individual predictors.

  Centering and scaling: To **improve the numerical stability** of some calculation.

  Center: The average predictor value is subtracted from all the values.

  Scale: each value of the predictor variable is divided by its standard deviation.

  Downside: **loss of interpretability** of the individual values since the data are no longer in the original units.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for individual predictors.

  Some predictive models prefer that the predictors have un-skewed distributions.

  Sample skewness statistic:

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

  x is the predictor variable.

  n is the number of values.

  $\bar{x}$ is the sample mean of the predictor.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for individual predictors.

  For **resolving skewness** we can:
  - Replace the data with the log, square root, or inverse (it may help).
  - Use the Box and Cox transformation:

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

  - Box and Cox[1] show how to use maximum likelihood estimation to determine the transformation parameter.

[1] Box G, Cox D (1964). *An Analysis of Transformations.* Journal of the Royal Statistical Society. Series B (Methodological), pp. 211-252.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors.

  Transformation to **resolve outliers**.

  We can use the **spatial sign** transformation, which has the effect of making all the samples the same distance from the center of the sphere.

  Mathematically, each sample is divided by its Euclidean norm.

  It does not remove outliers.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.

  PCA is a multivariate statistical technique for learning a low-dimensional representation of a set of data.

  PCA extracts the dominant patterns in the data by eliminating information redundancy due to variables cross-correlation.

  PCA searches in the original data space for new directions that are maximally independent in a linear sense, hence uncorrelated.

  PCA identifies the principal directions in which the data varies.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.

  PCA factorizes the $K \times D$ data matrix $\mathbf{X}$ using the eigen-decomposition, to obtain:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

  - $\mathbf{T}$ is a $K \times S$ **score matrix**.
  - $\mathbf{P}$ is a $D \times S$ **loading matrix**.
  - $\mathbf{E}$ is a $K \times D$ **residual matrix**.
  - S is the number of the retained **principal components**.
  - Each of the K measurements at time k is modelled as S-dimensional point $\mathbf{t}(k) = \mathbf{x}(k)\mathbf{P}$.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.

  PCA factorizes the $K \times D$ data matrix $\mathbf{X}$ using the eigen-decomposition, to obtain:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

  The scores are understood as the **new coordinates** of the point in a (sub)space whose directions are defined by the set of loadings $\{p_1, \dots, p_s, \dots, p_S\}$, or PCs, which are eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$.

  Typically, most of the variation in the data can be explained by **retaining a small number of PCs** compared with the original dimension of the data matrix $\mathbf{X}$.

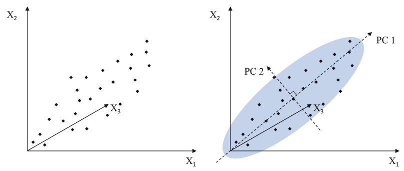
  The discarded PCs are associated with the smallest eigenvalues and they represent the directions with the least variance among the data.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.



Left: Observations are shown in the original space defined by X_1, X_2 and X_3.

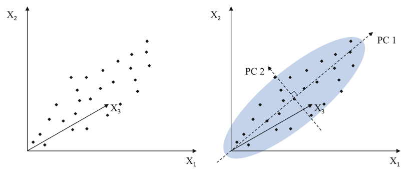
Right: A plane defined by two PCs (PC_1 and PC_2) is depicted in the original space.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.



The direction of PC_1 is associated with the largest variation among the observations. PC_2 is orthogonal to PC_1 and, within that constraint, it describes the largest possible variation left in the data.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.

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**.

Some predictive models prefer predictors to be uncorrelated (or at least low correlation) in order to find solutions and to improve the model's numerical stability.

PCA pre-processing creates new predictors with desirable characteristics for these kinds of models but **it must be used with understanding and care**.

PCA seeks predictor-set variation without regard to any further understanding of the predictors (i.e., measurement scales or distributions) or to knowledge of the modeling objectives (i.e., response variable).

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.

If the original predictors are on **measurement scales that differ in orders of magnitude** [e.g., demographic predictors such as income level (in dollars) and height (in m)], then the first few components will focus on summarizing the higher magnitude predictors (e.g., income), while latter components will summarize lower variance predictors (e.g., height).

The PC weights will be larger for the higher variability predictors on the first few PCs.

PCA will be focusing its efforts on identifying the data structure based on measurement scales rather than based on the important relationships within the data for the current problem.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.

  In addition, predictors may have skewed distributions.

  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.

  Centering and scaling enables PCA to find the underlying relationships in the data without being influenced by the original measurement scales.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Data transformation for multiple predictors: **Principal component analysis**.

  PCA does not consider the modeling objective or response variable when summarizing variability.

  That is, PCA is blind to the response: it is an **unsupervised technique**.

  If the predictive relationship between the predictors and response is not connected to the predictors' variability, then the derived PCs will not provide a suitable relationship with the response.

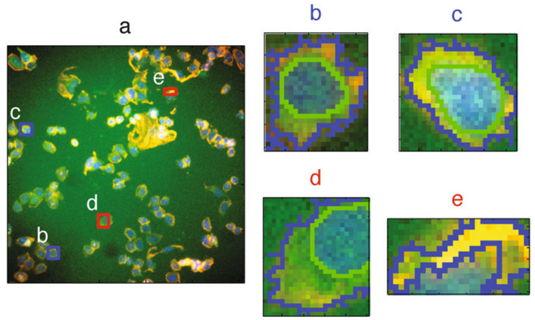
  In this case, a **supervised technique, like Partial Least Squares (PLS)**, will derive components while simultaneously considering the corresponding response.

A quick recap

During the last lecture, we did...

Introduce data pre-processing techniques.

- Case study: Cell segmentation in high-content screening.



An introduction to R

The R language is a platform for mathematical and statistical computations.

It is **free in two senses**:

1. R can be obtained free of charge (although commercial versions exist).
2. Anyone can examine or modify the source code
   (released under the General Public License)

R is an extremely powerful and flexible tool for data analysis, and it contains extensive capabilities for predictive modeling.

`https://cran.r-project.org`

- For the basic commands on R, check the script `TI0077_intro.R` on SIGAA.

An introduction to R

The packages

The AppliedPredictiveModeling package serves as a companion to the Kuhn and Johnson book:

- It includes many of the data sets not already available in other R packages.
- It also includes the R code used throughout the chapters and R functions.
- The package is available on CRAN and at (http://appliedpredictivemodeling.com).

The caret package (short for Classification And REgression Training) was created to streamline the process for building and evaluating predictive models.

- Using the package, a practitioner can quickly evaluate many different types of models to find the more appropriate tool for their data.
- The package provides many options for data pre-processing and resampling-based parameter tuning techniques.

Today's goal

Today, we going to ...

Continue with the data pre-processing case study.

- Data transformation for multiple predictors.
- Dealing with missing values.
- Removing predictors.

Reading list

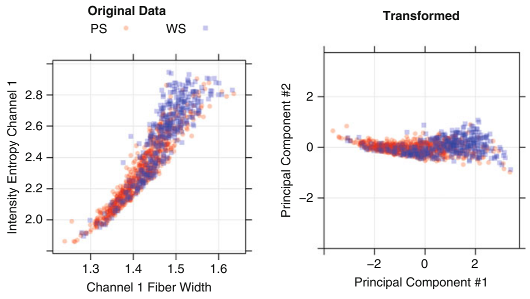
Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*, Springer, 2014[1]

W. N. Venables, D. M. Smith and the R Core Team. *An Introduction to R – Notes on R: A Programming Environment for Data Analysis and Graphics*, 2017[2]

[1] Today's lecture is mainly based on Chapter 3 of the book

[2] Available at: https://cran.r-project.org

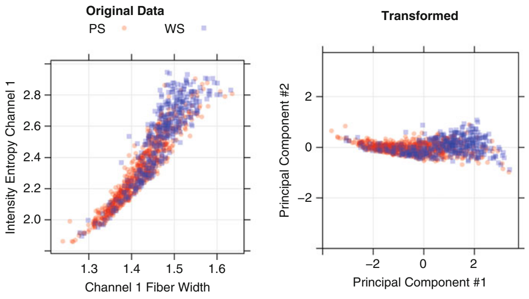
Case study: Cell segmentation in high-content screening



Example of the PC transformation for the cell segmentation data. This set contains a subset of two correlated predictors:

- Average pixel intensity of entropy of intensity in the cell (a measure of cell shape).
- Fiber width (a categorical response).

Case study: Cell segmentation in high-content screening

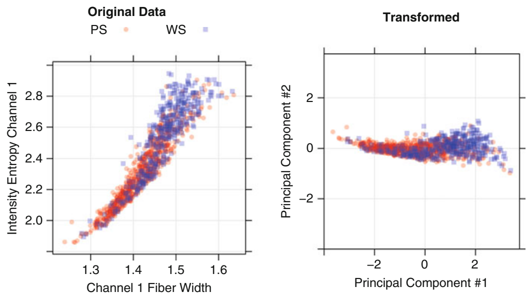


There is a **high correlation between the predictors** (0.93)

They measure redundant information about the cells.

Either predictor or a linear combination of these predictors could be used in place of the original predictors.

Case study: Cell segmentation in high-content screening



Two PCs can be derived: this transformation represents a rotation of the data about the axis of greatest variation.

We found that PC_1 summarizes 97% of the original variability and PC_2 summarizes 3%.

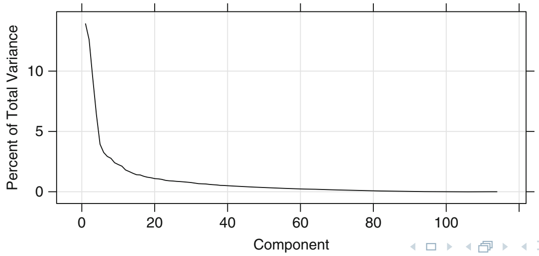
Hence, it is reasonable to **use only the first PC** for modeling since it accounts for the majority of information in the data.

Case study: Cell segmentation in high-content screening

We must decide **how many components to retain**.

- A heuristic approach is to create a **scree plot**.

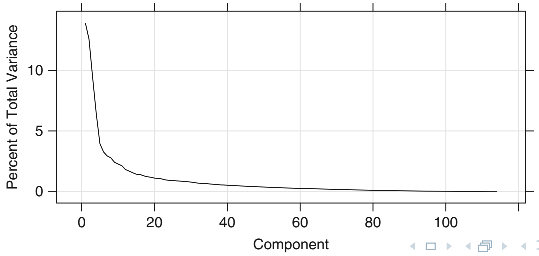
  It contains the ordered component number (x-axis) and the amount of summarized variability (y-axis).
- For most data sets, **the first few PCs summarize a majority of the variability**.
- The plot show a steep descent and variation then tapers off.
- Generally, the component number prior to the tapering off of variation is the maximal component that is retained.



Case study: Cell segmentation in high-content screening

We must decide **how many components to retain**.

- In the figure below, the first three components accounted for 14%, 12.6%, and 9.4% of the total variance, respectively.
- After four components, there is a sharp decline in the percentage of variation being explained, although these four components describe only 42.4% of the information in the data set.
- In an automated model building process, the **optimal number of components can be determined by cross-validation**.

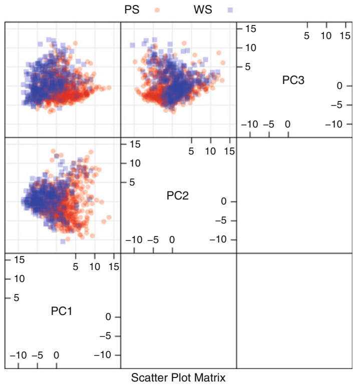


Case study: Cell segmentation in high-content screening

Visually examining the principal components is a critical step for assessing data quality and gaining intuition for the problem.

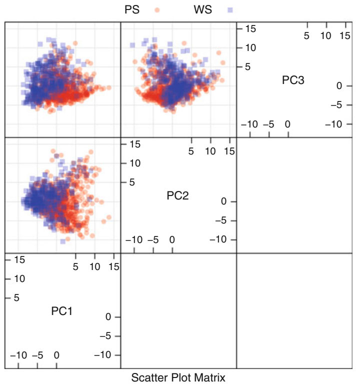
- **The first few principal components can be plotted against each other** and the plot symbols can be colored by relevant characteristics, such as the class labels.
- If PCA has captured a sufficient amount of information in the data, **this type of plot can demonstrate clusters of samples or outliers** that may prompt a closer examination of the individual data points.
- For classification problems, the PCA plot can show potential separation of classes (if there is a separation). This can set the initial expectations of the modeler; if there is little clustering of the classes, the plot of the principal component values will show a significant overlap of the points for each class.
- **Care should be taken when plotting the components**: the scale of the components tend to become smaller as they account for less and less variation in the data.

Case study: Cell segmentation in high-content screening



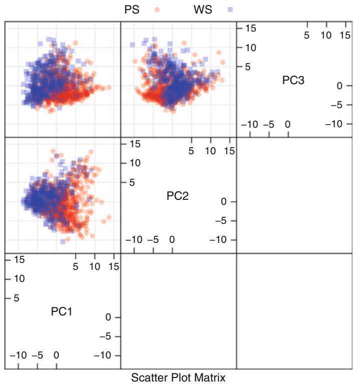
- The figure shows a scatter plot matrix for the first three principal components.
- The points are colored by class (segmentation quality).

Case study: Cell segmentation in high-content screening



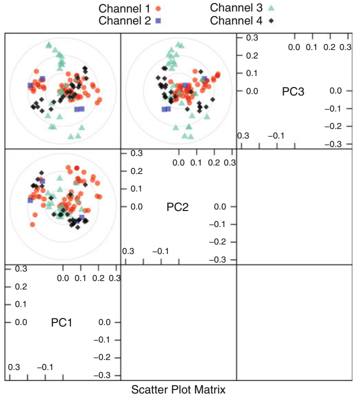
- Since the variations explained are not large for the first three components, it is important not to over-interpret the resulting image.
- From this plot, there appears to be some separation between the classes when plotting the first and second components.

Case study: Cell segmentation in high-content screening



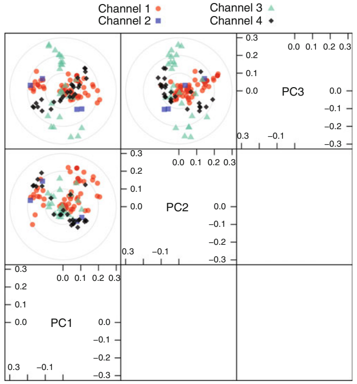
- However, the distribution of the well-segmented cells is roughly contained within the distribution of the poorly identified cells.
- One conclusion to infer is that the cell types are not easily separated.

Case study: Cell segmentation in high-content screening



- We can use of PCA to characterize which predictors are associated with each PC.
- Remember that each component is a linear combination of the predictors and the coefficients for each predictor are the loadings.
- Loadings close to zero indicate that the predictor variable did not contribute much to that component.

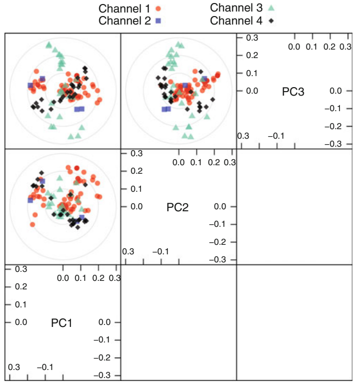
Case study: Cell segmentation in high-content screening



Scatter Plot Matrix

- The figure shows the loadings for the first three components in the cell data.
- Ch1 is associated with the cell body and Ch2 with the cell nucleus.
- Ch3 is associated with actin and Ch4 with tubulin.

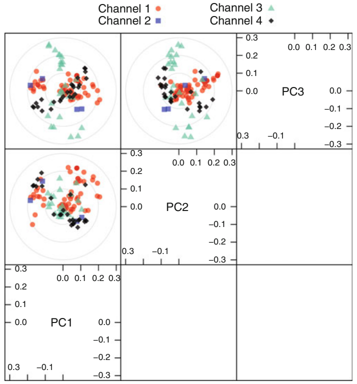
Case study: Cell segmentation in high-content screening



Scatter Plot Matrix

- The 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 are closer to zero for the first component.

Case study: Cell segmentation in high-content screening



Scatter Plot Matrix

- Conversely, the third principal component is mostly associated with the third channel while the cell body channel plays a minor role here.
- Even though the cell body measurements account for more variation in the data, this does not imply that these variables will be associated with predicting the segmentation quality.

Dealing with missing values

In many cases, **some predictors have no values for a given sample**.

- These missing data could be structurally missing.

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.

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.
- Informative missingness can induce significant bias in the model.

  For example, 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

Missing data should **not be confused with censored data** where the exact value is missing but something is known about its value.

- Censored data can 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.

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. For example, when a sample has a value below the limit of detection, the actual limit can be used in place of the real value.

Dealing with missing values

If we do not remove the missing data, there are two general approaches.

1. A few predictive models, especially tree-based techniques, can specifically account for missing data.
2. Missing data can be imputed. 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**.

   Imputation 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.

   Prior to model training or the prediction of new samples, missing values are filled in using imputation.

   Note that **this extra layer of models adds uncertainty**.

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.

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.

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.

- **Advantage** is that the imputed data are confined to be within the range of the training set values.
- **Disadvantage** is that the entire training set is required every time a missing value needs to be imputed.

Removing predictors

Potential advantages in removing predictors prior to modeling are:

1. Fewer predictors means decreased computational time and complexity.
2. If two predictors are highly correlated, this implies that they are measuring the same underlying information.
3. Some models can be crippled by predictors with degenerate distributions.

Removing predictors: Between-Predictor Correlations

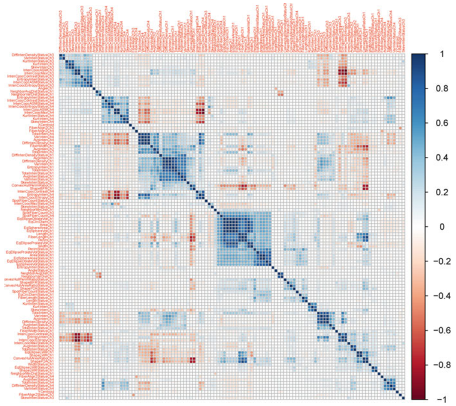
Collinearity is the technical term for the situation where a pair of predictor variables have a substantial correlation with each other.

Multicollinearity: We have relationships between multiple predictors at once.

- The cell data we 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 (i.e., shape), such as the roughness of the cell.

Removing predictors: Between-Predictor Correlations

Visualization of the cell segmentation correlation matrix.



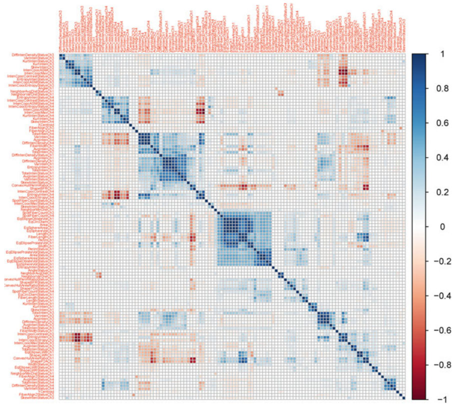
Each pairwise correlation is computed from the training data and colored according to its magnitude.

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.

Removing predictors: Between-Predictor Correlations

Visualization of the cell segmentation correlation matrix.



The predictor variables have been grouped using a clustering technique[1] so that collinear groups of predictors are adjacent to one another.

Looking along the diagonal, there are blocks of strong positive correlations that indicate “clusters” of collinearity.

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.

[1] Everitt B, Landau S, Leese M, Stahl D (2011). *Cluster Analysis*. Wiley.

Removing predictors: Between-Predictor Correlations

In general, there are **good reasons to avoid data with highly correlated predictors**.

Redundant predictors frequently add more complexity to the model than information they provide to the model.

- In situations where obtaining the predictor data is costly (either in time or money), fewer variables is obviously better.
- Using highly correlated predictors in techniques like linear regression can result in highly unstable models, numerical errors, and degraded predictive performance.
- A statistic called the **variance inflation factor** can be used to identify predictors that are impacted. 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 collinear predictors, it does not determine which should be removed to resolve the problem.

Removing predictors: Between-Predictor Correlations

A **more heuristic approach** is to remove the minimum number of predictors to ensure that all pairwise correlations are below a certain threshold.

While this method only identify collinearities in two dimensions, it can have a significantly positive effect on the performance of some models.

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.
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.

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**.

Example: The German credit data set is a popular tool for benchmarking machine learning algorithms.

- It contains 1000 samples that have been given labels of good and bad credit. In the data set, 70% were rated as having good credit.
- Data were collected related to credit history, employment, account status, ….
- Some predictors are **numeric**: the loan amount.
- Most of the predictors are **categorical** in nature: purpose of the loan, gender, or marital status.
- These data were encoded into several groups, including a group of “unknown”.
- To use these data in models, the categories are re-encoded into smaller bits of information called **dummy variables**.

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 table shows the possible dummy variables for the German data[1].

- Only four dummy variables are needed here; once you know the value of four of the dummy variables, the fifth can be inferred.
- However, the decision to include all of the dummy variables can depend on the choice of the model.

[1] The values are the amount in the savings account (in Deutsche Marks)