Missing data and IPCA

Dataset

We will use the Breast Cancer Wisconsin Diagnostic Dataset from the UCI Machine Learning Repository. The full dataset is available from the `dslabs` package, and it was adapted for this data analysis to include missing data.

The dataset contains 30 biopsy features for classification of 569 malignant (cancer) and benign (not cancer) breast masses. The dataset was randomly separated into a training and testing dataset. Around 1% of the observations in the training dataset were randomly removed to create missing data.

We can import the two datasets as follows:

```r
library(readr)
library(dplyr)

data_test <- read_csv("brca_test.csv")
data_train <- read_csv("brca_train.csv")

# Turn response variable into factor
data_test <- mutate(data_test, class = factor(class))
data_train <- mutate(data_train, class = factor(class))
```

Complete-case analysis

We will measure the classification performance of logistic regression on this dataset. As a benchmark, we will start with a complete-case analysis, i.e. every row with missing data will be omitted.

```r
data_cc <- na.omit(data_train)
```

Next, we can fit the K-Nearest Neighbours algorithm that will predict whether a mass is benign or malignant based on the biopsy features:

```r
library(kknn)
model_cc <- train.kknn(class ~ ., data_cc)
```

We can now use this model to predict the tumour status on the test data, and we can use the predictions to measure the performance of our algorithm:

```r
library(yardstick)

data_perf <- data.frame(
    "estimate" = predict(model_cc, newdata = data_test),
    "truth" = data_test$class
)
```

```r
# Create function to compute metrics
multi_metric <- metric_set(accuracy, kap, f_meas)
multi_metric(data_perf, truth = truth,
             estimate = estimate)
```

```
## # A tibble: 3 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.965
## 2 kap      binary         0.925
## 3 f_meas   binary         0.972
```

These metrics are the accuracy, Kappa, and the F-score. For each metric, a higher score means a better model. You can read more about these metrics here:

- **Accuracy and F-score**: https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers#Single_metrics
- **Kappa**: https://en.wikipedia.org/wiki/Cohen%27s_kappa

Iterative PCA

Your task is to use iterative PCA to impute the missing data. Once the data has been imputed, you can re-run the analysis to see if the metrics have improved.

For this task, you will need to find the optimal number of components K to retain in order to perform the imputation. You can use two approaches:

- Retain the number of components necessary to explain at least 90% of the variation in the data.
- Use the Generalized Cross-Valiation (GCV) criterion.