MESA-UKB-LVAtlas

Avan Suinesiaputra

2024-09-29

Table of contents

About

This online notebook details atlas-based shape score analysis conducted for the comparison between two left ventricular (LV) atlases derived from MESA and UK Biobank cohorts. It serves as an online source code material for the following paper:

> Avan Suinesiaputra, Kathleen Gilbert, Charlène Mauger, David A Bluemke, Colin Wu, Nay Aung, Stefan Neubauer, Stefan Piechnik, Steffen E Petersen, Joao A Lima, Bharath Ambale-Venkatesh, and Alistair Young, "Relationship between Left Ventricular Shape and Cardiovascular Risk Factors: Comparison between the Multi-Ethnic Study of Atherosclerosis and UK Biobank", *in review.*

Data availability

The MESA CMR images and their clinical and demographic data used in this study were available on request to the MESA Coordinating Centre at https://www.mesa-nhlbi.org. The UK Biobank CMR images and their clinical and demographic data used in this study were available on request to the UK Biobank at https://www.ukbiobank.ac.uk. The principal components of both MESA and UK Biobank derived in this study to build the PLSR model are available from the Cardiac Atlas Project website https://www.cardiacatlas.org.

Funding

This research was funded by the Health Research Council of New Zealand (17/608 and 17/234), and supported by the Innovate UK (104691) London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare with the core funding from the Wellcome/EPSRC Centre for Medical Engineering [WT203148/Z/16/Z] and Wellcome Trust Innovator award number 222678/Z/21/Z.

1 PLSR Training

In Partial Least Square Regression (PLSR), we want to estimate a linear combination of $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$ that are good predictors for both the input $\mathbf{X} \in \mathbb{R}^{n \times p}$ and the response $\mathbf{y} \in \mathbb{R}^n$. The PLSR linear relationships can be written as

$$\begin{aligned} \mathbf{X} &= \mathbf{Z}\mathbf{V}^T + \mathbf{E} \\ \mathbf{y} &= \mathbf{Z}\mathbf{b} + \mathbf{e} \end{aligned}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times k}$ are the *PLS scores*, $\mathbf{V} \in \mathbb{R}^{k \times p}$ are the *PLS loadings*, and $\mathbf{b} \in \mathbb{R}^k$ are the *PLS coefficients*. The terms $\mathbf{E}$ and $\mathbf{e}$ are the residual matrices for both input and response, respectively. The scalars n, p, and k denote the number of samples, input predictors, and **PLS components**.

Hence, to build a PLSR model, we need the number of PLS components, or k, and that's what we do the PLSR training.

In this paper,

- The input predictor $\mathbf{X}$ contains `age`, `sex`, and the first m principal components from the LV shape atlas. The number of m depends on the cohort: MESA or UKBB.
- The response $\mathbf{y}$ is a binary variable to denote the presence of a risk (=0) or not (=1). Five cardiovacular risk factors were used: hypertension, diabetes, obesity, hypercholesterolemia, and smoking.

1.1 k-Folds Cross Validation

We used five-fold cross validation to determine the optimal number of PLSR components. We did this for each cohort and for each risk factor. The general function to perform the k-fold cross validation for PLSR training is given below:

```r
train_pls <- function(form, dt, n_folds=5, n_comps=30,
                      prep=c("center"), probMethod="softmax")
{
  # create frequency table to calculate the weights
  response <- model.frame(form, data=dt)[[form[[2]]]]

  # create cross-validation folds
  cvIndex <- createFolds(factor(response), n_folds, returnTrain = T)

  # create caret's training controller
  ctrl <- trainControl(method = "cv",
                       index = cvIndex,
                       classProbs = TRUE,
                       verboseIter=TRUE,
                       summaryFunction = twoClassSummary,
                       savePredictions = TRUE,
                       allowParallel = TRUE)

  # train using PLS, metric is ROC.
  # Note that the number of PLS modes is given in the tuneLength argument.
  model <- train(form=form,
                 data=dt,
                 method="pls",
                 probMethod=probMethod,
                 metric="ROC",
                 tuneLength = n_comps,
                 preProc = prep,
                 trControl = ctrl)

  return(model)
}
```

1.2 Training results

1. MESA atlas

    a. Hypertension
    b. Diabetes
    c. Obesity
    d. Hypercholesterolemia
    e. Smoking

2. UKBB atlas

    a. Hypertension
    b. Diabetes
    c. Obesity
    d. Hypercholsterolemia
    e. Smoking

2 PLSR Validation

There are two types of validation: *internal* (within-cohort), and *external* (cross-cohort) validations. For the internal validation, a leave-one-out cross validation was used to build the final PLSR model across the whole cohort population. For the external validation, the PLSR model was built using all training case samples. For both internal and external validations, we used the optimal number of PLSR component computed during the traing.

Validation results:

1. MESA atlas

   a. Hypertension
   b. Diabetes
   c. Obesity
   d. Hypercholesterolemia
   e. Smoking

2. UKBB atlas

   a. Hypertension
   b. Diabetes
   c. Obesity
   d. Hypercholsterolemia
   e. Smoking