MAST90104 - Lecture 10

Weichang Yu

Room 108, Old Geology South Bldg
School of Mathematics and Statistics, University of Melbourne

Motivating examples

Video streaming: Suppose you have survey responses from n subscribers who rated $P \gg n$ videos. You want group your subscribers into $K \ll P$ groups using their responses.

Image recognition: Suppose you have n pictures each with $P \gg n$ pixel intensity data. You want use the raw pixel data to engineer a small number of composite features (predictors) for classification.

Temperature tracking: There are up to 8 different types of earth temperature measurements that climate scientists are actively tracking. Plotting a time series of 8 different temperatures can be hard to interpret visually. You want to create composite scores based on the different measurement types.

Principal Components Analysis (PCA)

Suppose $\mathbf{\Sigma}$ is the covariance matrix of a p random vector of $\mathbf{X}$, i.e., $\text{Var}(\mathbf{X}) = \mathbf{\Sigma}$

From results on matrices in Lecture 2, $\mathbf{\Sigma} = \mathbf{P}\Lambda\mathbf{P}^T$ where

- $\mathbf{P}$ is an orthogonal matrix with the eigenvectors of Σ as its columns
- Λ is a diagonal matrix with the (non-negative) eigenvalues of Σ along the diagonal
- The eigenvalues can be - and from now on are assumed to be - organised in descending order from the top left hand entry of Λ

Recall that the eigenvectors in $\mathbf{P}$ must have length one and be orthogonal - $\mathbf{P}$ is called a rotation matrix as a result.

For any column vector $\mathbf{a}$, if $\mathbf{a}$ has unit length $\mathbf{a}^T\mathbf{a} = 1$, the column vector $\mathbf{b} = P^T\mathbf{a}$ also has unit length, since

$$\mathbf{b}^T\mathbf{b} = \mathbf{a}^T PP^T\mathbf{a} = \mathbf{a}^T\mathbf{I}\mathbf{a} = 1$$

recalling that the inverse of an orthogonal matrix is its transpose.

Further, from variance properties in Lecture 3,

$$Var(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\Sigma\mathbf{a} = \mathbf{b}^T\Lambda\mathbf{b}.$$

And so

$$Var(\mathbf{a}^T\mathbf{X}) = \mathbf{b}^T\Lambda\mathbf{b} = \sum_{i=1}^{p}\Lambda_{ii}b_i^2 \leq \Lambda_{11}\sum_{i=1}^{p}b_i^2 = \Lambda_{11} \qquad (1)$$

If $\mathbf{a}$ is taken to be the first column of P, that is the eigenvector corresponding to the largest eigenvalue of Σ, then $\mathbf{b}^T = \mathbf{a}^T P = (1, 0, \cdots, 0)$.

The bound in inequality (1) is then achieved so that $\mathbf{a}^T\mathbf{X}$ has the greatest variance amongst all linear combinations of elements of $\mathbf{X}$.

This linear combination $\mathbf{a}^T\mathbf{X}$ is called the *first principal component* .

For data, the *first principal component* is the linear combination of the variables with weights equal to the first eigenvector.

The *first principal component* is often said to be the linear combination of variables which most *explains* the variation in the multivariate data.

If $\mathbf{a}$ is taken to be the second column of $\mathbf{P}$, it is the eigenvector corresponding to the second largest eigenvalue.

The random variable $\mathbf{a}^T\mathbf{X}$ is called the *second principal component*.

For data, the *second principal component* is the linear combination of the variables with weights equal to the second eigenvector of $\boldsymbol{\Sigma}$.

This has maximal variance amongst all linear combinations of the variables that are orthogonal to the first linear combination.

For data, the orthogonality is uncorrelatedness of the components.

Often principal components is performed with the correlation matrix - the covariance matrix for the standardised random variables.

The reason for this is that it eliminates the scale of the random variables as a determinant of the principal components.

The first two principal components are often plotted against each other as this can reveal structure in the data.

This is particularly appropriate if the sum of their variances accounts for a large proportion of the total of variances of the principal components.

Using an estimate for Σ

In practice, we don't know Σ. So we would need to estimate Σ.

We can collect iid data $\{\mathbf{x}_i\}_{i=1}^n$ and construct the sample covariance matrix estimator:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T.$$

We use $\widehat{\boldsymbol{\Sigma}}$ instead of Σ.

Example: Global Temperatures

The Kaggle web site, Climate Change Earth Surface Temperature Data, contains data from the organisation Berkeley Earth.

Berkeley Earth is affiliated with Lawrence Berkeley National Laboratory in the US.

The Berkeley Earth Surface Temperature Study combines 1.6 billion temperature reports from 16 pre-existing archives.

Figure 1 shows their estimates of land-surface average temperatures together with a 95% confidence interval.

Also shown are estimates from the most US Government National Oceanigraphic and Atmospheric Administration, NASA and the UK Government Hadley Centre of the Meterological Office.

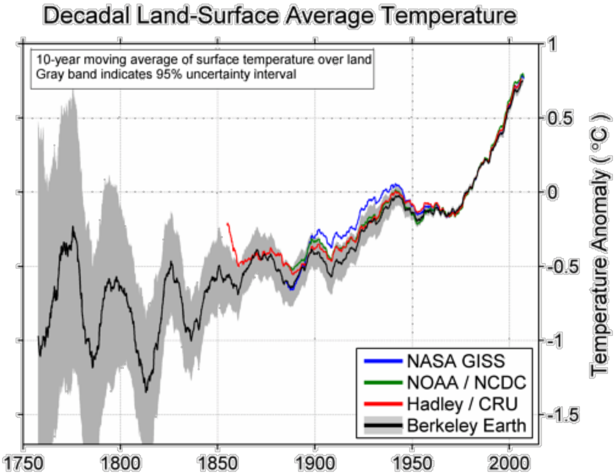


Figure: Berkeley Earth Global Temperatures From 1750

PCA can be illustrated from one of the Berkley Earth data sets, Global Land and Ocean-and-Land Temperatures (GlobalTemperatures.csv), which is available in the data folder on the LMS.

The data set gives average land temperature from 1850, max and min land temperatures and global ocean and land temperatures.

It also gives uncertainty estimates for each average, so that the average plus and minus the uncertainty estimate gives a 95% confidence interval for the average.

The next slide gives descriptions of each of the variables.

1. Date: the date at the start of the month
2. L_Ave: global average land temperature in celsius
3. L_AveCI: half the width of the 95% confidence interval around the average
4. L_Max: global average maximum land temperature in celsius
5. L_MaxCI: half the width of the 95% confidence interval around the max-average
6. L_Min: global average minimum land temperature in celsius
7. L_MinCI: half the width of the 95% confidence interval around the min-average
8. LO_Ave: global average land and ocean temperature in celsius
9. LO_AveCI: half the width of the 95% confidence interval around the LO average

```
GT <- read.csv("../data/GlobalTemperatures.csv")
GT$Date <- as.Date(GT$Date)
GT$Year <- as.integer(format(GT$Date,"%Y"))
GT$YearGroup <-
cut(GT$Year, breaks = c(1849,1899,1939,1979,1999,2020),
labels = c("pre 1900","pre 1940",
"pre 1980","pre 2000","after 2000"))
GT$Month = format(GT$Date,’%b’)
```

```
#  do Principal Components on Columns 2 to 9
#  summary gives the proportion of variance for each component
PCAtemp <- prcomp(GT[2:9])
summary(PCAtemp)

## Importance of components%s:
##                           PC1     PC2     PC3     PC4     PC5     PC6
## Standard deviation     7.4453 0.74341 0.33829 0.25820 0.17922 0.17140
## Proportion of Variance 0.9857 0.00983 0.00203 0.00119 0.00057 0.00052
## Cumulative Proportion  0.9857 0.99555 0.99758 0.99877 0.99934 0.99986
##                            PC7     PC8
## Standard deviation     0.08660 0.01549
## Proportion of Variance 0.00013 0.00000
## Cumulative Proportion  1.00000 1.00000

#  PCAtemp£ rotation has the component weightings
PCAtemp$rotation

##                   PC1         PC2          PC3           PC4         PC5
## L_Ave     0.571946823  0.04001266  0.039868197  0.7436785600 -0.13625963
## L_AveCI  -0.004512822  0.27501118 -0.045303280 -0.0055524370 -0.15597650
## L_Max     0.577714155  0.08747253 -0.708251464 -0.3749478512  0.11347921
## L_MaxCI  -0.009022128  0.76075681  0.180640746  0.0008976024  0.59460736
## L_Min     0.557041711 -0.07698389  0.676395111 -0.4656649725 -0.08395457
## L_MinCI  -0.010408586  0.55934918 -0.008052543 -0.0046540451 -0.68469486
## LO_Ave    0.169161866 -0.09745747  0.064428657  0.2990879162  0.33460064
## LO_AveCI -0.001370918  0.08983038 -0.019511093 -0.0023589538 -0.05344326
##                  PC6         PC7          PC8
## L_Ave    -0.31008624 -0.04301755 -0.007607443
## L_AveCI  -0.04266753  0.90723977 -0.270244467
## L_Max     0.04250666 -0.04106506 -0.004265789
## L_MaxCI  -0.13651667 -0.12767968 -0.006243556
## L_Min     0.01821578  0.04493341  0.004960284
## L_MinCI   0.37956797 -0.27203576 -0.007162697
## LO_Ave    0.85793382  0.14038710  0.023771881
## LO_AveCI -0.03358999  0.24768133  0.962399015
```

The summary shows that the first two Principal Components account for most of the variation in all eight components

The component weightings show that the First Principal Component is mostly an average of the Land Ave, Max and Min with a smaller weighting on the Land and Ocean Average, and very small weightings on the CI components.

The Second Principal Component is mostly and average of the CI for Max and Min on Land with a smaller weighting for the land average and very small weightings on the other variables.

```r
# for easy plots of the first two principal components
library(ggplot2)
library(ggfortify)
# Plot of first two Principal Components
autoplot(PCAtemp)
# colour by Year Group
autoplot(PCAtemp,data= GT,colour="YearGroup")
# colour by Month
autoplot(PCAtemp,data= GT,colour="Month")
```

The next three slides shows the plots in Figures 2, 3 and 4. The non-coloured plot shows that there is structure and the coloured ones show that both month and year produce clusters.

Figure: Plot of First Two Principal Components against each other

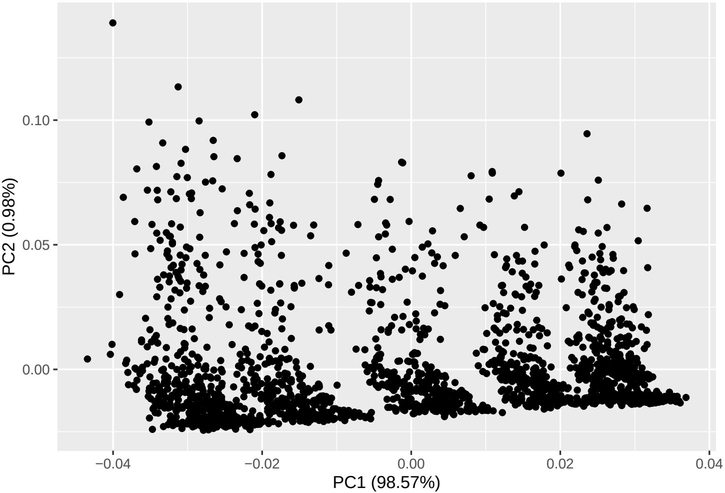


Figure: PCA Plot with Year Group coloured

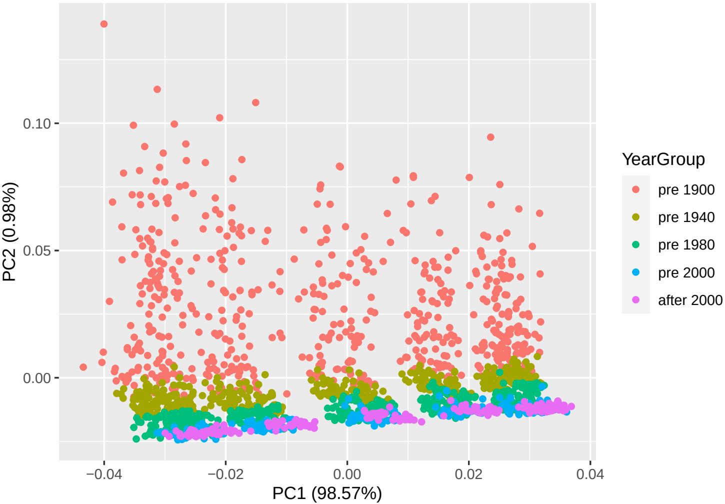
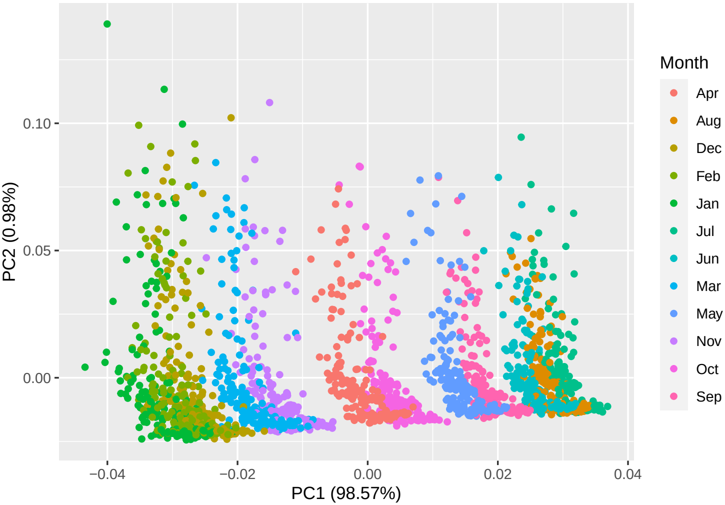


Figure: PCA Plot with Month coloured



Clearly the CI columns are on a different scale to the various temperature columns:

```
> apply(GT[2:9],2,var)
L_Ave      L_AveCI        L_Max      L_MaxCI
18.174811792  0.050189229 18.572470864  0.340125690
L_Min      L_MinCI       LO_Ave     LO_AveCI
17.270967207  0.198771377  1.623312857  0.005415017
```

This might explain why the weights on `L_Ave`, `L_Max` and `L_Min` when perform PCA are larger than the weights of the uncertainty variables.

But it is not ideal to have the result depends on the choice of scaling .

It would be better to do the plots by scaling: the mean is subtracted from each variable and the result is divided by the standard deviation.

This is just doing the PCA with the correlation matrix rather than the covariance matrix.

```
# standardise variables by subtracting mean and dividing SD
PCAtempcor <- prcomp(GT[2:9],scale = TRUE)
summary(PCAtempcor)
```

```
## Importance of components%s:
##                          PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.113 1.7838 0.42115 0.35380 0.16829 0.11880
## Proportion of Variance 0.558 0.3977 0.02217 0.01565 0.00354 0.00176
## Cumulative Proportion  0.558 0.9558 0.97794 0.99359 0.99713 0.99889
##                            PC7     PC8
## Standard deviation     0.07554 0.05633
## Proportion of Variance 0.00071 0.00040
## Cumulative Proportion  0.99960 1.00000
```

```
PCAtempcor$rotation
```

```
##                 PC1       PC2         PC3           PC4         PC5
## L_Ave     0.3883726 0.3190540 -0.02014057  0.0005663997 -0.06476292
## L_AveCI  -0.3157703 0.3977348 -0.44317094 -0.1254976876  0.71341443
## L_Max     0.3855833 0.3223742 -0.05419762  0.0114395136 -0.10248587
## L_MaxCI  -0.2927403 0.3955458  0.69087016 -0.5282526974 -0.01581318
## L_Min     0.3934632 0.3089703  0.01519964 -0.0045519228 -0.04887134
## L_MinCI  -0.3169789 0.3772258  0.27841818  0.8234703267 -0.03621378
## LO_Ave    0.4036124 0.2864475  0.04937042  0.0346052614  0.22533309
## LO_AveCI -0.3112757 0.3997697 -0.49271672 -0.1604575077 -0.64932068
##                   PC6          PC7           PC8
## L_Ave     0.132165675 -0.008086470 -0.8516090093
## L_AveCI   0.143846268  0.004429598 -0.0165687360
## L_Max     0.324679898  0.711601274  0.3493355744
## L_MaxCI   0.008702339  0.034181359  0.0002258244
## L_Min     0.357743364 -0.699281260  0.3607066271
## L_MinCI   0.016555557 -0.002012846 -0.0039242571
## LO_Ave   -0.824301453 -0.008014694  0.1452505946
## LO_AveCI -0.220500413 -0.057512440  0.0350681032
```

The results show that the first two principal components still account for most of the total variance in the principal components.

The first principal component is close to the difference between the average of the temperature variables and the uncertainty variables.

The second principal component is close to an average of all variables.

From 1980 onwards the temperature variables are high and the uncertainties lower as can be seen in the plots.

Now do the plots using the PCA from the Correlation Matrix

```
autoplot(PCAtempcor)
autoplot(PCAtempcor,data= GT,colour="YearGroup")
autoplot(PCAtempcor,data= GT,colour="Month")
```

The next three slides shows these plots in Figures 5, 6 and 7.

Figure: Plot of First Two Principal Components on Correlation Matrix

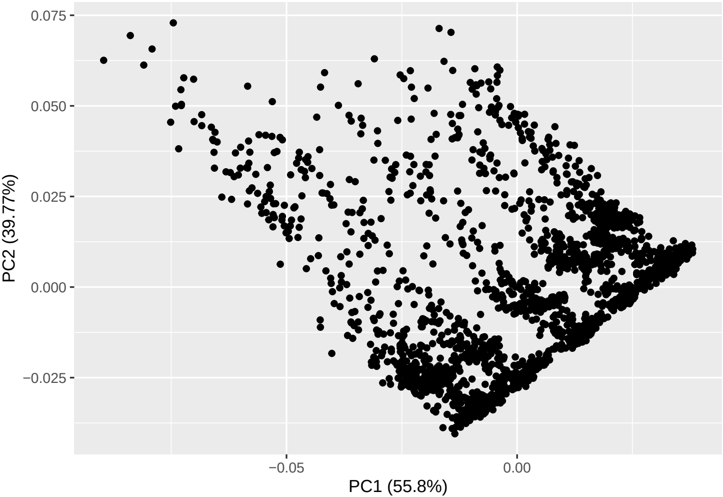


Figure: PCA Plot on Correlation Matrix with Year Group coloured

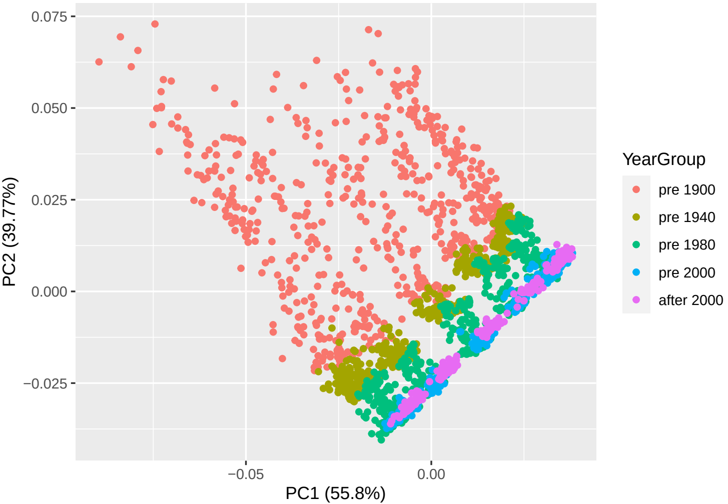
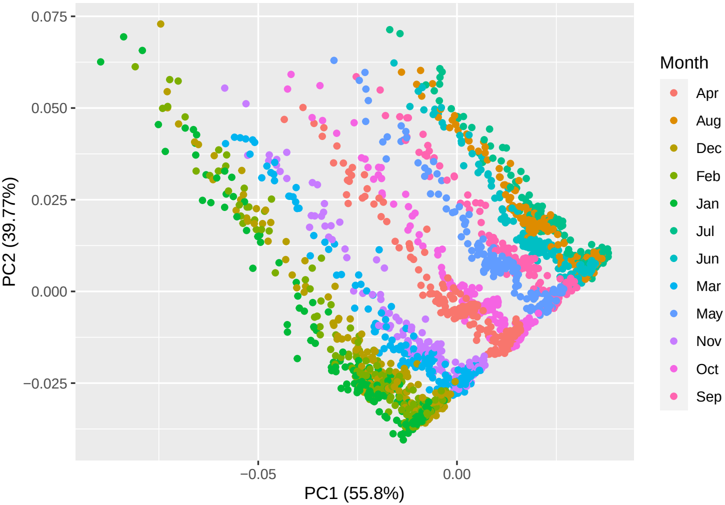


Figure: PCA Plot on Correlation Matrix with Month coloured



Clearly from the plots of the Principal Components, both for the covariance or the correlation matrix, Month is important - probably reflecting more Northern Hemisphere collecting stations than Southern Hemisphere.

This is not surprising since 90% of the world's population lives in the Northern Hemisphere (Wikipedia), collecting stations are more on land than ocean and the Northern Hemisphere has two thirds of the land (Wikipedia).

It makes sense, therefore, to aggregate the data by year, recording the mean of each of the variables for the year.

For uncertainty estimates, the additional averaging will make a uniform difference for all variables.

Scaling, that is using the correlation matrix, will remove this difference and so all the results from now on are done with correlation matrices.

```r
# Select Year and the temperature and CI variables
# Aggregate by Year using the mean of the monthly variables
# Set up grouping by Year
GTYear <- GT[2:10]
GTYear <- aggregate(. ~ Year,data=GTYear, FUN= "mean")
GTYear$YearGroup <-
cut(GTYear$Year, breaks = c(1849,1899,1939,1979,1999,2020),
labels = c("pre 1900","pre 1940","pre 1980",
"pre 2000","after 2000"))
```

```
# Redo previous analysis with yearly averages
PCAYear <- prcomp(GTYear[2:9],scale=TRUE)
summary(PCAYear)

## Importance of components%s:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.5265 1.1425 0.38174 0.25521 0.20372 0.16954
## Proportion of Variance 0.7979 0.1631 0.01822 0.00814 0.00519 0.00359
## Cumulative Proportion  0.7979 0.9611 0.97928 0.98742 0.99261 0.99620
##                            PC7     PC8
## Standard deviation     0.15572 0.07827
## Proportion of Variance 0.00303 0.00077
## Cumulative Proportion  0.99923 1.00000

PCAYear$rotation

##                 PC1       PC2         PC3         PC4        PC5
## L_Ave     0.3554530 0.3613321 -0.05044360  0.29848929 -0.2794324
## L_AveCI  -0.3682680 0.3063994 -0.15365629  0.16644851 -0.1575010
## L_Max     0.3526953 0.3083686 -0.66532373 -0.38757566  0.4227651
## L_MaxCI  -0.3395425 0.4104530  0.40763061 -0.35393838  0.2920075
## L_Min     0.3531396 0.3418129  0.42024091 -0.49627869 -0.3483652
## L_MinCI  -0.3502887 0.3840346  0.04215379  0.22224333  0.3915426
## LO_Ave    0.3450857 0.3979668  0.15917660  0.55024318  0.1095643
## LO_AveCI -0.3631183 0.2988301 -0.40163448 -0.09622665 -0.5885950
##                    PC6           PC7         PC8
## L_Ave     0.3261299537  0.6538040165  0.19886930
## L_AveCI   0.2327987972  0.0029660579 -0.80010755
## L_Max     0.0354899344 -0.0522254081 -0.07019235
## L_MaxCI   0.5202142853 -0.0615202830  0.25520196
## L_Min    -0.3837787905 -0.0006713813 -0.25868223
## L_MinCI  -0.6353229423  0.3381582994  0.08575749
## LO_Ave    0.0001920529 -0.6160751895  0.05367076
## LO_AveCI -0.1289965010 -0.2685500822  0.41601985
```

The first two principal components now account for 96% of the total variation in principal components.

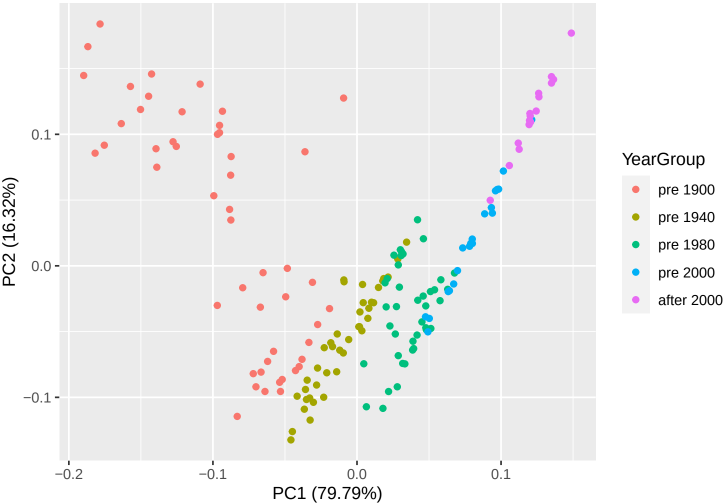
The first principal component is very close to the difference between the average of the temperature variables and the uncertainty variables.

The second principal component is verl close to the average of all the variables.

Figure 8 shows the plot of the first two principal components in the yearly averages with Year Group coloured.

The contrasts in accuracy and level over time are not simple.

Figure: PCA Plot on Yearly Averages with Year Group coloured


PC2 (16.32%)
PC1 (79.79%)
YearGroup
● pre 1900
● pre 1940
● pre 1980
● pre 2000
● after 2000