CSC424 Take-home final

Jasmine Dumas

March 14, 2016

CSC 324/424 Winter 2016 Take-home Final Exam Due Tue/Wed March 15/16, 2016 @ 10:00pm in D2L Dropbox

The purpose of this final exam is to test your ability to apply the techniques of this course without specific prompting to run a particular technique with a particular data set. This is a course where it is especially true that, “You get out of it what you put in,” because you can do well on the assignments and other tests without digging in and working with the techniques we have discussed. However, digging in and getting practice is the only way to fully understand these. I encourage you to enjoy this assignment and look at some interesting data, but I also understand many people will have other obligations and I will not judge you for doing something straightforward.

Submission: turn in two documents. First (1) written responses to the questions below, labeled by question number and (2) your code and/or output from your analysis.

Questions:

1. Identify a multivariate data set you will use for this assignment. If you haven't already chosen and need ideas, take a look at the last lab.

**YearPredictionMSD Data Set: This data is a subset of the Million Song Dataset. The Million Song Dataset is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. The core of the dataset is the feature analysis and metadata for one million songs, provided by The Echo Nest. The dataset does not include any audio, only the derived features.**

(Source: <http://labrosa.ee.columbia.edu/millionsong/>)

1. Provide a link to the data set (if not possible, add the data as one of the posted files or explain why you cannot, e.g. corporate IP).

[http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD#](http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD)

1. Explain the variables of the data including their meaning and type (nominal, numeric, …)

**There are 90 numeric/interger variables: 12 = timbre average, 78 = timbre covariance and the first variable is the song year (target, y). Timbre is the character or quality of a musical sound or voice as distinct from its pitch and intensity. The first value is the year (target), ranging from 1922 to 2011. Features extracted from the 'timbre' features from The Echo Nest API. We take the average and covariance over all 'segments', each segment being described by a 12-dimensional timbre vector.**

#MSD <- read.table("/Users/jasminedumas/desktop/depaul/CSC424/YearPredictionMSD.txt", sep = ",")  
#save(MSD, file = "MSD.RData")  
load("/Users/jasminedumas/desktop/R-directory/MSD.RData")

1. How many samples are there?

**There are 515,345 samples in the dataset**

dim(MSD)

## [1] 515345 91

1. Are there missing values?

**There are no missing values**

which(is.na(MSD)) # returning a zero indicates no missing values

## integer(0)

1. Provide a question you have about this data that can be answered using a technique we've studied.

**Can audio features predict the song release year?**

1. Explain what analysis technique from class you will use to answer your question, including why and how. I'm happy to see visualization in the results, but this is limited to the core techniques we learned plus SVM if you need it.

**I'm choosing to intially perform a Principal Component Analysis (PCA) on the data to reduce the dimentionality then implore a multiple linear regression to determine if timbre attributes are a linear predictor of song year.**

1. Test that the proposed technique can be used with the chosen data.

**There a four assumptions to satisfy before performing the dimension reduction technique: multiple continuous variables, a linear relationship exists between all the variables, enough samples, enough variables for reduction**

1. Run the appropriate tests we discussed in class to determine if the technique is appropriate. Examples include testing for adequate sampling for PCA and Box's M Test for LDA.

**Yes, there are multiple continuous variables.**

**Yes, the p-value is less than the significance criteria of 0.05 indicated that we reject the null hypothesis and there is no difference between the varaibles and accept the alternative hypothesis that there is a statistically significant difference between the variables.**

# do we have a linear relationship between the variables?  
initial.model = lm(V1~., data=MSD)  
summary(initial.model)

##   
## Call:  
## lm(formula = V1 ~ ., data = MSD)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -75.844 -3.364 1.731 5.822 66.625   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.951e+03 1.895e-01 10294.608 < 2e-16 \*\*\*  
## V2 8.754e-01 3.902e-03 224.342 < 2e-16 \*\*\*  
## V3 -5.633e-02 4.130e-04 -136.380 < 2e-16 \*\*\*  
## V4 -4.365e-02 7.798e-04 -55.975 < 2e-16 \*\*\*  
## V5 3.353e-03 2.224e-03 1.507 0.131699   
## V6 -1.475e-02 7.220e-04 -20.426 < 2e-16 \*\*\*  
## V7 -2.201e-01 2.477e-03 -88.852 < 2e-16 \*\*\*  
## V8 -6.739e-03 1.531e-03 -4.401 1.08e-05 \*\*\*  
## V9 -1.009e-01 2.970e-03 -33.969 < 2e-16 \*\*\*  
## V10 -7.047e-02 1.809e-03 -38.952 < 2e-16 \*\*\*  
## V11 2.507e-02 3.937e-03 6.367 1.93e-10 \*\*\*  
## V12 -1.657e-01 6.518e-03 -25.421 < 2e-16 \*\*\*  
## V13 -1.855e-03 2.089e-03 -0.888 0.374539   
## V14 4.701e-02 9.263e-04 50.756 < 2e-16 \*\*\*  
## V15 3.551e-04 1.274e-05 27.873 < 2e-16 \*\*\*  
## V16 -4.226e-04 2.003e-05 -21.101 < 2e-16 \*\*\*  
## V17 5.992e-04 3.253e-05 18.417 < 2e-16 \*\*\*  
## V18 4.766e-04 4.483e-05 10.629 < 2e-16 \*\*\*  
## V19 1.467e-03 7.080e-05 20.715 < 2e-16 \*\*\*  
## V20 1.924e-03 8.538e-05 22.540 < 2e-16 \*\*\*  
## V21 2.128e-03 1.131e-04 18.821 < 2e-16 \*\*\*  
## V22 7.699e-04 1.412e-04 5.452 4.98e-08 \*\*\*  
## V23 -4.026e-04 2.241e-04 -1.797 0.072395 .   
## V24 7.539e-03 2.198e-04 34.302 < 2e-16 \*\*\*  
## V25 2.812e-03 1.634e-04 17.206 < 2e-16 \*\*\*  
## V26 -3.556e-03 1.466e-04 -24.257 < 2e-16 \*\*\*  
## V27 7.114e-05 2.422e-05 2.937 0.003317 \*\*   
## V28 1.589e-03 4.810e-05 33.043 < 2e-16 \*\*\*  
## V29 5.294e-04 7.802e-05 6.786 1.15e-11 \*\*\*  
## V30 8.745e-04 1.182e-04 7.401 1.35e-13 \*\*\*  
## V31 -3.042e-04 1.500e-04 -2.027 0.042636 \*   
## V32 -1.405e-03 1.924e-04 -7.301 2.86e-13 \*\*\*  
## V33 -1.401e-03 2.511e-04 -5.580 2.41e-08 \*\*\*  
## V34 -5.560e-03 3.166e-04 -17.560 < 2e-16 \*\*\*  
## V35 2.472e-03 3.558e-04 6.949 3.68e-12 \*\*\*  
## V36 1.850e-03 5.179e-04 3.571 0.000355 \*\*\*  
## V37 -5.294e-03 2.061e-04 -25.692 < 2e-16 \*\*\*  
## V38 -2.773e-04 4.381e-05 -6.330 2.46e-10 \*\*\*  
## V39 6.792e-04 4.272e-05 15.900 < 2e-16 \*\*\*  
## V40 1.365e-03 7.354e-05 18.562 < 2e-16 \*\*\*  
## V41 -1.710e-03 8.927e-05 -19.160 < 2e-16 \*\*\*  
## V42 -1.991e-03 1.457e-04 -13.665 < 2e-16 \*\*\*  
## V43 -7.642e-04 1.609e-04 -4.750 2.03e-06 \*\*\*  
## V44 -1.403e-03 2.387e-04 -5.876 4.20e-09 \*\*\*  
## V45 -2.359e-03 4.267e-04 -5.528 3.23e-08 \*\*\*  
## V46 -3.180e-03 4.280e-04 -7.429 1.09e-13 \*\*\*  
## V47 6.813e-03 4.111e-04 16.573 < 2e-16 \*\*\*  
## V48 4.561e-04 3.469e-05 13.145 < 2e-16 \*\*\*  
## V49 -2.075e-03 7.337e-05 -28.282 < 2e-16 \*\*\*  
## V50 2.752e-04 1.003e-04 2.745 0.006060 \*\*   
## V51 1.941e-03 1.375e-04 14.114 < 2e-16 \*\*\*  
## V52 2.201e-04 2.038e-04 1.080 0.280208   
## V53 -1.605e-03 2.635e-04 -6.091 1.12e-09 \*\*\*  
## V54 1.971e-03 2.476e-04 7.961 1.70e-15 \*\*\*  
## V55 4.908e-04 2.253e-04 2.178 0.029395 \*   
## V56 -8.438e-05 3.218e-04 -0.262 0.793179   
## V57 1.629e-04 6.875e-05 2.369 0.017827 \*   
## V58 -1.898e-03 6.029e-05 -31.473 < 2e-16 \*\*\*  
## V59 1.940e-03 8.570e-05 22.643 < 2e-16 \*\*\*  
## V60 -1.304e-03 1.137e-04 -11.474 < 2e-16 \*\*\*  
## V61 2.332e-04 1.727e-04 1.351 0.176826   
## V62 -3.032e-03 3.305e-04 -9.173 < 2e-16 \*\*\*  
## V63 -1.880e-03 3.168e-04 -5.934 2.96e-09 \*\*\*  
## V64 -7.769e-03 6.289e-04 -12.352 < 2e-16 \*\*\*  
## V65 1.190e-03 5.276e-05 22.558 < 2e-16 \*\*\*  
## V66 -2.025e-03 9.507e-05 -21.301 < 2e-16 \*\*\*  
## V67 6.597e-04 1.357e-04 4.862 1.16e-06 \*\*\*  
## V68 -1.934e-04 1.954e-04 -0.990 0.322367   
## V69 -4.275e-04 1.736e-04 -2.462 0.013813 \*   
## V70 -4.251e-03 1.773e-04 -23.978 < 2e-16 \*\*\*  
## V71 -5.084e-03 4.700e-04 -10.817 < 2e-16 \*\*\*  
## V72 -1.061e-03 7.482e-05 -14.186 < 2e-16 \*\*\*  
## V73 2.378e-04 8.168e-05 2.911 0.003603 \*\*   
## V74 6.891e-04 1.321e-04 5.215 1.84e-07 \*\*\*  
## V75 3.987e-03 2.542e-04 15.686 < 2e-16 \*\*\*  
## V76 3.001e-03 2.523e-04 11.895 < 2e-16 \*\*\*  
## V77 1.521e-02 7.459e-04 20.393 < 2e-16 \*\*\*  
## V78 1.996e-04 6.711e-05 2.974 0.002935 \*\*   
## V79 -4.423e-03 1.333e-04 -33.170 < 2e-16 \*\*\*  
## V80 -4.243e-05 1.255e-04 -0.338 0.735364   
## V81 -1.515e-04 1.359e-04 -1.116 0.264631   
## V82 -8.278e-04 6.470e-04 -1.280 0.200707   
## V83 -5.568e-04 1.073e-04 -5.191 2.09e-07 \*\*\*  
## V84 1.372e-03 1.638e-04 8.378 < 2e-16 \*\*\*  
## V85 9.965e-04 1.675e-04 5.950 2.69e-09 \*\*\*  
## V86 2.614e-02 1.119e-03 23.357 < 2e-16 \*\*\*  
## V87 1.074e-04 1.458e-04 0.737 0.461171   
## V88 1.165e-03 9.576e-05 12.164 < 2e-16 \*\*\*  
## V89 -3.120e-02 1.484e-03 -21.018 < 2e-16 \*\*\*  
## V90 -1.381e-03 8.191e-05 -16.854 < 2e-16 \*\*\*  
## V91 -1.615e-03 7.457e-04 -2.167 0.030270 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.549 on 515254 degrees of freedom  
## Multiple R-squared: 0.237, Adjusted R-squared: 0.2369   
## F-statistic: 1778 on 90 and 515254 DF, p-value: < 2.2e-16

**Yes, the overall Measure of Sampling Adequacy (MSA) of factor analytic data matrices is above the threshold of 0.5. This (Barlett's Test in SPSS) KMO function is just a function of the squared elements of the ‘image' matrix compared to the squares of the original correlations. The overall MSA as well as estimates for each item are found and the ideal value is above 0.5. The index is known as the Kaiser-Meyer-Olkin (KMO) index.**

(Source: <http://www.personality-project.org/r/html/KMO.html>)

# do we have enough samples?  
nrow(MSD)

## [1] 515345

library(psych)  
KMO(MSD)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = MSD)  
## Overall MSA = 0.78  
## MSA for each item =   
## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15   
## 0.67 0.88 0.75 0.62 0.61 0.75 0.76 0.39 0.72 0.44 0.56 0.54 0.34 0.75 0.90   
## V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30   
## 0.88 0.93 0.90 0.92 0.88 0.90 0.86 0.89 0.93 0.86 0.64 0.69 0.80 0.60 0.85   
## V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 V41 V42 V43 V44 V45   
## 0.87 0.86 0.87 0.80 0.80 0.89 0.74 0.64 0.75 0.89 0.74 0.65 0.68 0.89 0.73   
## V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57 V58 V59 V60   
## 0.74 0.71 0.78 0.75 0.90 0.77 0.81 0.85 0.65 0.63 0.69 0.80 0.68 0.87 0.75   
## V61 V62 V63 V64 V65 V66 V67 V68 V69 V70 V71 V72 V73 V74 V75   
## 0.87 0.82 0.72 0.71 0.62 0.76 0.80 0.82 0.71 0.67 0.45 0.77 0.59 0.83 0.59   
## V76 V77 V78 V79 V80 V81 V82 V83 V84 V85 V86 V87 V88 V89 V90   
## 0.79 0.75 0.73 0.61 0.85 0.66 0.73 0.70 0.73 0.71 0.55 0.52 0.48 0.55 0.67   
## V91   
## 0.64

**Yes, we have enough variables for PCA as the number of sample should be ~5 times the number of variables**

# do we have enough variables?  
ncol(MSD)

## [1] 91

nrow(MSD) >= 5\*ncol(MSD)

## [1] TRUE

1. Report the results of your test. If the data fail the tests, you must switch data or explain why you are going forward anyway.

**Yes, the tests about the assumptions passed and I'm moving forward with the PCA technique.**

1. Apply the chosen technique to the chosen data.

fit = princomp(MSD[, -1], cor=F, n.obs=.) # using the covaraince matrix  
summary(fit) # print variance accounted for

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4  
## Standard deviation 2114.5220857 1173.3518856 934.39564195 686.28669623  
## Proportion of Variance 0.4691291 0.1444523 0.09160718 0.04941731  
## Cumulative Proportion 0.4691291 0.6135814 0.70518857 0.75460587  
## Comp.5 Comp.6 Comp.7 Comp.8  
## Standard deviation 543.26739815 465.09338211 412.47094603 394.10026558  
## Proportion of Variance 0.03096672 0.02269594 0.01785067 0.01629601  
## Cumulative Proportion 0.78557259 0.80826854 0.82611921 0.84241522  
## Comp.9 Comp.10 Comp.11 Comp.12  
## Standard deviation 384.35239573 347.3185227 331.59175482 3.052774e+02  
## Proportion of Variance 0.01549984 0.0126568 0.01153653 9.778164e-03  
## Cumulative Proportion 0.85791506 0.8705719 0.88210839 8.918866e-01  
## Comp.13 Comp.14 Comp.15 Comp.16  
## Standard deviation 2.846723e+02 2.619850e+02 259.77465030 2.504909e+02  
## Proportion of Variance 8.502726e-03 7.201461e-03 0.00708046 6.583425e-03  
## Cumulative Proportion 9.003893e-01 9.075907e-01 0.91467120 9.212546e-01  
## Comp.17 Comp.18 Comp.19 Comp.20  
## Standard deviation 2.337856e+02 2.289327e+02 2.221275e+02 1.956707e+02  
## Proportion of Variance 5.734603e-03 5.498999e-03 5.176932e-03 4.017165e-03  
## Cumulative Proportion 9.269892e-01 9.324882e-01 9.376652e-01 9.416823e-01  
## Comp.21 Comp.22 Comp.23 Comp.24  
## Standard deviation 1.931382e+02 1.902759e+02 1.861883e+02 1.805806e+02  
## Proportion of Variance 3.913852e-03 3.798704e-03 3.637248e-03 3.421448e-03  
## Cumulative Proportion 9.455962e-01 9.493949e-01 9.530321e-01 9.564536e-01  
## Comp.25 Comp.26 Comp.27 Comp.28  
## Standard deviation 175.90022116 1.704308e+02 1.585041e+02 1.528319e+02  
## Proportion of Variance 0.00324639 3.047641e-03 2.636021e-03 2.450731e-03  
## Cumulative Proportion 0.95969997 9.627476e-01 9.653836e-01 9.678344e-01  
## Comp.29 Comp.30 Comp.31 Comp.32  
## Standard deviation 1.452879e+02 1.395904e+02 1.344358e+02 1.247355e+02  
## Proportion of Variance 2.214762e-03 2.044463e-03 1.896258e-03 1.632481e-03  
## Cumulative Proportion 9.700491e-01 9.720936e-01 9.739898e-01 9.756223e-01  
## Comp.33 Comp.34 Comp.35 Comp.36  
## Standard deviation 122.53965901 1.146924e+02 1.130375e+02 1.098301e+02  
## Proportion of Variance 0.00157551 1.380185e-03 1.340643e-03 1.265642e-03  
## Cumulative Proportion 0.97719784 9.785780e-01 9.799187e-01 9.811843e-01  
## Comp.37 Comp.38 Comp.39 Comp.40  
## Standard deviation 1.086387e+02 1.062048e+02 1.054485e+02 1.009500e+02  
## Proportion of Variance 1.238331e-03 1.183467e-03 1.166673e-03 1.069253e-03  
## Cumulative Proportion 9.824226e-01 9.836061e-01 9.847728e-01 9.858420e-01  
## Comp.41 Comp.42 Comp.43 Comp.44  
## Standard deviation 1.003866e+02 97.698174795 9.359834e+01 8.889137e+01  
## Proportion of Variance 1.057351e-03 0.001001477 9.191876e-04 8.290621e-04  
## Cumulative Proportion 9.868994e-01 0.987900859 9.888200e-01 9.896491e-01  
## Comp.45 Comp.46 Comp.47 Comp.48  
## Standard deviation 8.797705e+01 8.617056e+01 8.311217e+01 7.997741e+01  
## Proportion of Variance 8.120948e-04 7.790866e-04 7.247649e-04 6.711238e-04  
## Cumulative Proportion 9.904612e-01 9.912403e-01 9.919651e-01 9.926362e-01  
## Comp.49 Comp.50 Comp.51 Comp.52  
## Standard deviation 7.894336e+01 7.524455e+01 72.706032794 6.729591e+01  
## Proportion of Variance 6.538817e-04 5.940432e-04 0.000554637 4.751659e-04  
## Cumulative Proportion 9.932901e-01 9.938841e-01 0.994438740 9.949139e-01  
## Comp.53 Comp.54 Comp.55 Comp.56  
## Standard deviation 6.382514e+01 6.349711e+01 6.185714e+01 60.472512066  
## Proportion of Variance 4.274167e-04 4.230345e-04 4.014648e-04 0.000383693  
## Cumulative Proportion 9.953413e-01 9.957644e-01 9.961658e-01 0.996549515  
## Comp.57 Comp.58 Comp.59 Comp.60  
## Standard deviation 5.735726e+01 53.680871747 5.185361e+01 5.064151e+01  
## Proportion of Variance 3.451793e-04 0.000302348 2.821148e-04 2.690799e-04  
## Cumulative Proportion 9.968947e-01 0.997197043 9.974792e-01 9.977482e-01  
## Comp.61 Comp.62 Comp.63 Comp.64  
## Standard deviation 47.968020358 4.717687e+01 4.422675e+01 3.996710e+01  
## Proportion of Variance 0.000241419 2.335211e-04 2.052287e-04 1.675996e-04  
## Cumulative Proportion 0.997989656 9.982232e-01 9.984284e-01 9.985960e-01  
## Comp.65 Comp.66 Comp.67 Comp.68  
## Standard deviation 3.943608e+01 3.848432e+01 3.765334e+01 3.655662e+01  
## Proportion of Variance 1.631757e-04 1.553944e-04 1.487561e-04 1.402168e-04  
## Cumulative Proportion 9.987592e-01 9.989146e-01 9.990633e-01 9.992035e-01  
## Comp.69 Comp.70 Comp.71 Comp.72  
## Standard deviation 3.417079e+01 2.989833e+01 2.954967e+01 2.846360e+01  
## Proportion of Variance 1.225118e-04 9.379110e-05 9.161639e-05 8.500558e-05  
## Cumulative Proportion 9.993261e-01 9.994199e-01 9.995115e-01 9.995965e-01  
## Comp.73 Comp.74 Comp.75 Comp.76  
## Standard deviation 2.613228e+01 24.856155793 2.243647e+01 2.027781e+01  
## Proportion of Variance 7.165103e-05 0.000064824 5.281738e-05 4.314296e-05  
## Cumulative Proportion 9.996681e-01 0.999732949 9.997858e-01 9.998289e-01  
## Comp.77 Comp.78 Comp.79 Comp.80  
## Standard deviation 1.978893e+01 1.854895e+01 1.435932e+01 1.400368e+01  
## Proportion of Variance 4.108776e-05 3.609996e-05 2.163394e-05 2.057559e-05  
## Cumulative Proportion 9.998700e-01 9.999061e-01 9.999277e-01 9.999483e-01  
## Comp.81 Comp.82 Comp.83 Comp.84  
## Standard deviation 1.091287e+01 9.758205e+00 8.965337e+00 7.772034e+00  
## Proportion of Variance 1.249527e-05 9.990972e-06 8.433371e-06 6.337783e-06  
## Cumulative Proportion 9.999608e-01 9.999708e-01 9.999792e-01 9.999856e-01  
## Comp.85 Comp.86 Comp.87 Comp.88  
## Standard deviation 6.975213e+00 6.296718e+00 4.946299e+00 3.459308e+00  
## Proportion of Variance 5.104848e-06 4.160029e-06 2.567016e-06 1.255586e-06  
## Cumulative Proportion 9.999907e-01 9.999948e-01 9.999974e-01 9.999987e-01  
## Comp.89 Comp.90  
## Standard deviation 3.042061e+00 1.898070e+00  
## Proportion of Variance 9.709657e-07 3.780006e-07  
## Cumulative Proportion 9.999996e-01 1.000000e+00

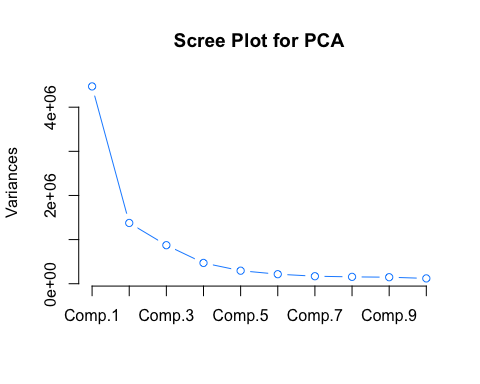
head(loadings(fit)) # pc loadings

## [1] -0.001492918 -0.006293541 -0.004794322 0.002143497 0.003013464  
## [6] 0.001892908

head(fit$scores) # the principal components

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  
## [1,] -2415.709 28.47538 -123.78190 102.25783 99.70768 77.03527  
## [2,] -1439.821 1129.52402 233.77791 -12.31912 301.11478 -173.15825  
## [3,] -2343.676 21.81713 -32.98421 54.36646 -13.07864 385.51776  
## [4,] -1404.432 1243.16977 271.67941 -126.84616 -428.94147 -59.37263  
## [5,] -2409.953 542.54247 -175.95359 12.98134 21.81229 232.40328  
## [6,] -2213.361 399.93903 21.00449 -68.39982 168.57273 129.62801  
## Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12  
## [1,] 248.80142 30.293094 155.356530 150.93510 84.47983 13.78586  
## [2,] -173.54038 -175.964297 -228.313801 30.03707 234.72303 -22.48453  
## [3,] 134.33271 100.919680 133.432762 92.57456 41.08271 -102.31997  
## [4,] 462.26237 -2.776925 -166.185274 -76.91901 -123.55132 -140.67199  
## [5,] 240.80639 116.737325 -1.586518 54.12408 33.90793 154.71439  
## [6,] 75.21068 -10.012731 -6.320626 172.77543 96.35934 -145.15334  
## Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18  
## [1,] 63.08902 -164.73483 25.45680 -70.74487 -7.343463 15.74204  
## [2,] 122.33368 21.68408 96.37927 32.37995 -79.989598 -76.12171  
## [3,] 237.92956 -123.76430 92.18470 139.55759 -59.758875 -135.18287  
## [4,] 123.29307 -44.94703 -300.68790 -78.27814 35.937286 110.50527  
## [5,] 43.34984 -32.69519 18.97230 -35.90610 21.521765 27.62980  
## [6,] -33.11842 -13.20143 79.18613 13.02461 7.926733 -35.00685  
## Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24  
## [1,] -8.246076 -24.698415 -3.085154 -117.65169 22.55309 84.85943  
## [2,] -240.995426 58.283047 -30.895385 -69.76234 196.24534 -100.05097  
## [3,] -11.347982 -10.636988 39.198712 53.04506 95.75500 -72.72447  
## [4,] 53.709943 -45.223933 -110.536638 79.19081 -36.11263 118.47003  
## [5,] -33.235241 6.693865 -89.360313 33.70831 54.06666 -26.37169  
## [6,] -153.605444 2.004224 -10.193277 33.34531 60.22478 -33.90749  
## Comp.25 Comp.26 Comp.27 Comp.28 Comp.29 Comp.30  
## [1,] -55.024694 6.087214 -87.76245 -20.03397 -36.759628 68.33129  
## [2,] 6.150711 -82.646933 18.18080 158.07290 20.469207 70.35834  
## [3,] 40.525514 11.354648 -135.22271 183.81489 -25.193854 31.14818  
## [4,] 101.557039 1.368994 -37.06564 15.14627 -182.863557 54.58167  
## [5,] -77.528592 22.149542 -47.63320 106.73706 -7.827002 54.01395  
## [6,] -48.908233 61.338057 -40.43922 16.31606 -41.095035 43.60779  
## Comp.31 Comp.32 Comp.33 Comp.34 Comp.35 Comp.36 Comp.37  
## [1,] -13.45240 21.49399 82.22864 18.61507 14.36965 81.79943 -45.70160  
## [2,] 38.64021 23.82251 83.82314 -63.67193 -45.70079 64.41781 76.47165  
## [3,] 66.42243 41.06506 30.01828 10.78116 -6.42480 51.42056 -36.69566  
## [4,] 105.35996 79.17994 59.87456 158.70727 34.23778 94.19893 16.73024  
## [5,] -60.00265 15.17597 70.80912 32.93280 16.61543 22.20963 -17.57901  
## [6,] -103.22914 26.83135 49.04138 119.08136 70.29509 17.76745 -59.49825  
## Comp.38 Comp.39 Comp.40 Comp.41 Comp.42 Comp.43  
## [1,] -34.08895 7.562176 -31.27051 3.040074 -25.867203 -10.67583  
## [2,] -168.84234 -17.581348 -15.43898 27.064534 -33.104078 -69.28468  
## [3,] 49.32961 20.687365 -15.81483 -11.366299 -47.944797 -12.40906  
## [4,] 62.46488 -86.035471 29.73853 -7.497750 66.239506 -41.57489  
## [5,] -101.64289 25.809822 12.65396 -14.686977 6.639151 -25.23698  
## [6,] 23.27718 -6.212974 26.45432 -51.990551 -7.745412 23.33202  
## Comp.44 Comp.45 Comp.46 Comp.47 Comp.48 Comp.49  
## [1,] 5.767230 45.410070 -50.239078 -31.888663 69.14667 3.832148  
## [2,] 5.980342 -7.664951 -10.182621 -35.369054 47.90816 31.088305  
## [3,] -16.835381 56.196738 -14.965576 28.362540 -13.72451 -17.190632  
## [4,] -5.862260 -25.948431 23.965215 59.550464 31.38407 -2.337821  
## [5,] -1.175525 -4.218382 -94.342443 34.471530 13.61607 -33.688943  
## [6,] 14.399800 44.690161 2.796596 2.451728 51.91476 12.857129  
## Comp.50 Comp.51 Comp.52 Comp.53 Comp.54 Comp.55  
## [1,] 9.3560628 23.208317 -19.281644 6.582228 30.668956 -26.40556  
## [2,] -41.7338552 9.711946 6.907992 -34.327810 -50.927079 52.19644  
## [3,] -0.5297246 -5.195597 -11.347005 22.319260 11.689724 56.56231  
## [4,] -6.7230053 -24.986900 -22.249427 7.436901 -4.808585 22.31647  
## [5,] -52.7829353 39.262346 -37.988893 6.429447 -6.118891 -3.72652  
## [6,] 2.7911458 12.510026 -5.061006 16.064084 3.314453 -22.93324  
## Comp.56 Comp.57 Comp.58 Comp.59 Comp.60 Comp.61  
## [1,] 28.420751 19.22067 3.692994 16.0991002 -1.535945 5.6210789  
## [2,] 43.540686 26.00501 46.936176 0.5434656 -2.592321 -12.5897108  
## [3,] -2.379424 18.34234 4.253304 -20.7231909 -31.262926 -9.6079384  
## [4,] -39.078518 -40.61611 29.379046 -39.0815575 -18.208925 10.2055168  
## [5,] 8.950697 -22.41101 14.916705 2.2091593 30.976576 -0.3852629  
## [6,] 15.980068 -25.90200 -17.724579 4.2692661 -10.869030 -5.0083681  
## Comp.62 Comp.63 Comp.64 Comp.65 Comp.66 Comp.67  
## [1,] 8.880336 12.43818 -4.743702 -12.92622 -13.180833 -11.716818  
## [2,] 53.314654 -15.08083 -20.013476 -11.02736 -22.729495 4.446455  
## [3,] -7.516742 22.70192 -4.473513 20.16216 3.954833 -18.536860  
## [4,] 36.217669 -10.23111 15.015774 -45.13443 2.552482 3.302985  
## [5,] 1.667873 20.30517 -59.511022 23.73054 -5.932577 6.826938  
## [6,] -11.132238 17.13318 1.976171 25.10394 3.956497 10.799758  
## Comp.68 Comp.69 Comp.70 Comp.71 Comp.72 Comp.73  
## [1,] -7.25922 1.5607513 12.184123 -7.824651 52.636950 6.236415  
## [2,] -16.81297 -42.2438489 -15.974685 -18.458420 43.480011 -5.868230  
## [3,] 25.39526 -13.4551521 -6.258010 4.365960 45.782320 -1.735410  
## [4,] -16.80729 -6.8670631 1.981317 -35.181101 6.435137 19.620780  
## [5,] -14.11892 -17.6841979 -6.407491 -23.597032 36.330177 13.911417  
## [6,] -18.67236 -0.7464889 30.209764 -21.585104 75.473310 -8.512502  
## Comp.74 Comp.75 Comp.76 Comp.77 Comp.78 Comp.79  
## [1,] -3.996042 -0.1321107 -6.660453 2.827692 0.7868649 -10.123275  
## [2,] -12.153738 -0.7751169 27.197557 -2.045888 4.0762761 13.091033  
## [3,] -2.347819 3.6303157 -6.965494 12.041065 7.7614838 -9.532095  
## [4,] -7.261599 12.8157295 20.022540 17.762332 3.0759218 1.484853  
## [5,] -11.997029 5.6768845 10.638468 3.872110 18.2773723 13.341930  
## [6,] -4.055415 1.2712938 2.079572 -5.146352 -9.6231836 -1.538033  
## Comp.80 Comp.81 Comp.82 Comp.83 Comp.84 Comp.85  
## [1,] -25.360009 -1.580761 -2.4114144 -1.5802528 -8.5572569 0.7605201  
## [2,] 6.966768 2.754998 15.4169730 -1.9348489 -2.3228698 3.0801066  
## [3,] -5.819317 -2.561006 10.4791224 -0.1728562 -0.7862868 2.2137636  
## [4,] 6.475354 -1.432794 1.7789375 3.5291053 -3.9688733 1.3571422  
## [5,] -15.382693 5.457887 6.6169916 -6.0710995 -0.1594424 -0.5853492  
## [6,] 21.391365 -1.765160 0.9299168 4.7389373 -3.2664039 4.4779605  
## Comp.86 Comp.87 Comp.88 Comp.89 Comp.90  
## [1,] -3.714952 6.995345 -0.07255931 -0.4640891 1.4836456  
## [2,] -9.092578 -1.761528 0.60480662 -1.5930472 0.1725360  
## [3,] -3.176681 4.124660 1.54873878 -1.8990132 1.8919056  
## [4,] -5.727025 11.579770 2.58767194 -2.1347580 -2.9119601  
## [5,] -3.207202 3.410215 1.49640426 0.1421038 2.1422059  
## [6,] -3.718138 6.852595 0.74292702 -2.0820378 0.6232563

plot(fit,type="lines", main="Scree Plot for PCA", col = "dodgerblue") # scree plot



**The PCA technique revealed that the atleast 90% of the total variance of the data can be explained using 13 principal components (the SPSS method selects 11 principal components which explain ~88% of the total varaince in the dataset). But for real world simplicity the first two components explain abut 61% of the variance. These correlation values between the latent variables and real values are called loadings. The scree plot indicates the "knee" is at Comp.2**

Note: the head() function was used to show the first 6 loadings and score output and the prevent print 500 + pages since the scope of the data set (large amount of observations) is very large.

1. Validate the model constructed by the chosen technique by testing that it is significant and has a reasonable model fit. Use the criteria appropriate for the model you've chosen.

**I'm validating the model by re-incorporating the PC scores into a multiple linear model and using the usual p-value, value to evaluate the result and fit.**

newdat <-fit$scores[, 1:13] # select the first 13 principl components  
newMSD = as.data.frame(newdat) # create a new data frame  
newMSD$year = MSD$V1 # add target to new data frame  
pca.model = lm(year ~., data=newMSD) # new linear model with principal component scores  
summary(pca.model) # includes p-value and R values

##   
## Call:  
## lm(formula = year ~ ., data = newMSD)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -77.057 -3.968 3.084 7.393 70.310   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.998e+03 1.490e-02 134124.236 < 2e-16 \*\*\*  
## Comp.1 2.830e-04 7.046e-06 40.169 < 2e-16 \*\*\*  
## Comp.2 4.368e-04 1.270e-05 34.399 < 2e-16 \*\*\*  
## Comp.3 8.766e-04 1.595e-05 54.972 < 2e-16 \*\*\*  
## Comp.4 1.910e-04 2.171e-05 8.799 < 2e-16 \*\*\*  
## Comp.5 -2.585e-03 2.743e-05 -94.256 < 2e-16 \*\*\*  
## Comp.6 -1.606e-03 3.204e-05 -50.121 < 2e-16 \*\*\*  
## Comp.7 -1.055e-04 3.612e-05 -2.921 0.00349 \*\*   
## Comp.8 3.429e-04 3.781e-05 9.070 < 2e-16 \*\*\*  
## Comp.9 -3.382e-04 3.876e-05 -8.723 < 2e-16 \*\*\*  
## Comp.10 1.486e-03 4.290e-05 34.643 < 2e-16 \*\*\*  
## Comp.11 -9.959e-04 4.493e-05 -22.164 < 2e-16 \*\*\*  
## Comp.12 1.977e-03 4.881e-05 40.516 < 2e-16 \*\*\*  
## Comp.13 -2.406e-03 5.234e-05 -45.976 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.7 on 515331 degrees of freedom  
## Multiple R-squared: 0.04256, Adjusted R-squared: 0.04253   
## F-statistic: 1762 on 13 and 515331 DF, p-value: < 2.2e-16

1. Finally, report on the results and answer your proposed question if possible.

**Based on the linear model criterion, I'm not comfortable in using this model to predict song release year from timbre characteristics. The p-value for the model was very low and the p-value for each principal component was very low which is acceptable but its very suspicious that the Adjusted R-squared value was very low at 0.04253. The original question could be answered by syaying the song charactersitics can predict song release year but it would have to be attached with a accuracy level (low).**

**The reduction of covariates from 90 to 13 is definetly our goal to simplify complexity and develop a parsimonius model that is in-expensive to compute and reduces noise obviously present in the majority of the covariates. The coefficients in the new linear model don't correspond back to the original variables which is a disadvantage of using Principal Component Analysis but for this specific dataset the variables were not specificfied aside from all being extracted timbre characteristics averages and covarainces.**

**I used a linear model function but this could improved on my using a generalized linear model (GLM) to specify error distribution and the link function between the response and covariates.**