Data Preparation

Modify Assignment

Allen Rahrooh

23 March 2020

Table of Contents

Table of Figures

Part I Transforming the Predictors:

Create a table listing the predictor 1-90, and the transform you selected. Create a summary table showing how many times each transform was used.

Predictors	Transformations	Times Applied
1	scale and center, square, exponential, log, no	1
2	scale and center, square, exponential, log, no	1
3	scale and center, square, exponential, log, no	1
4	scale and center, square, exponential, log, no	1
5	scale and center, square, exponential, log, no	1
6	scale and center, square, exponential, log, no	1
7	scale and center, square, exponential, log, no	1
8	scale and center, square, exponential, log, no	1
9	scale and center, square, exponential, log, no	1
10	scale and center, square, exponential, log, no	1
11	scale and center, square, exponential, log, no	1
12	scale and center, square, exponential, log, no	1
13	scale and center, square, exponential, log, no	1
14	scale and center, square, exponential, log, no	1
15	scale and center, square, exponential, log, no	1
16	scale and center, square, exponential, log, no	1
17	scale and center, square, exponential, log, no	1
18	scale and center, square, exponential, log, no	1
19	scale and center, square, exponential, log, no	1
20	scale and center, square, exponential, log, no	1
21	scale and center, square, exponential, log, no	1
22	scale and center, square, exponential, log, no	1
23	scale and center, square, exponential, log, no	1
24	scale and center, square, exponential, log, no	1
25	scale and center, square, exponential, log, no	1
26	scale and center, square, exponential, log, no	1
27	scale and center, square, exponential, log, no	1
28	scale and center, square, exponential, log, no	1
29	scale and center, square, exponential, log, no	1
30	scale and center, square, exponential, log, no	1
31	scale and center, square, exponential, log, no	1
32	scale and center, square, exponential, log, no	1
33	scale and center, square, exponential, log, no	1
34	scale and center, square, exponential, log, no	1
35	scale and center, square, exponential, log, no	1
36	scale and center, square, exponential, log, no	1
37	scale and center, square, exponential, log, no	1
38	scale and center, square, exponential, log, no	1
39	scale and center, square, exponential, log, no	1
40	scale and center, square, exponential, log, no	1
41	scale and center, square, exponential, log, no	1
42	scale and center, square, exponential, log, no	1
43	scale and center, square, exponential, log, no	1
44	scale and center, square, exponential, log, no	1
45	scale and center, square, exponential, log, no	1
46	scale and center, square, exponential, log, no	1
47	scale and center, square, exponential, log, no	1
48	scale and center, square, exponential, log, no	1
49	scale and center, square, exponential, log, no	1
50	scale and center, square, exponential, log, no	1
51	scale and center, square, exponential, log, no	1
52	scale and center, square, exponential, log, no	1
53	scale and center, square, exponential, log, no	1
54	scale and center, square, exponential, log, no	1
55	scale and center, square, exponential, log, no	1
56	scale and center, square, exponential, log, no	1
57	scale and center, square, exponential, log, no	1
58	scale and center, square, exponential, log, no	1
59	scale and center, square, exponential, log, no	1
60	scale and center, square, exponential, log, no	1
61	scale and center, square, exponential, log, no	1
62	scale and center, square, exponential, log, no	1
63	scale and center, square, exponential, log, no	1
64	scale and center, square, exponential, log, no	1
65	scale and center, square, exponential, log, no	1
66	scale and center, square, exponential, log, no	1
67	scale and center, square, exponential, log, no	1
68	scale and center, square, exponential, log, no	1
69	scale and center, square, exponential, log, no	1
70	scale and center, square, exponential, log, no	1
71	scale and center, square, exponential, log, no	1
72	scale and center, square, exponential, log, no	1
73	scale and center, square, exponential, log, no	1
74	scale and center, square, exponential, log, no	1
75	scale and center, square, exponential, log, no	1
76	scale and center, square, exponential, log, no	1
77	scale and center, square, exponential, log, no	1
78	scale and center, square, exponential, log, no	1
79	scale and center, square, exponential, log, no	1
80	scale and center, square, exponential, log, no	1
81	scale and center, square, exponential, log, no	1
82	scale and center, square, exponential, log, no	1
83	scale and center, square, exponential, log, no	1
84	scale and center, square, exponential, log, no	1
85	scale and center, square, exponential, log, no	1
86	scale and center, square, exponential, log, no	1
87	scale and center, square, exponential, log, no	1
88	scale and center, square, exponential, log, no	1
89	scale and center, square, exponential, log, no	1
90	scale and center, square, exponential, log, no	1

Figure 1 Summary of Transformations Applied

```
library(readr)
music <- read.table("YearPredictionMSD.txt", sep=",")
#data is read in
#now we apply a transformation for all the predictors
#v1 is the target variable
predictors <- music[,2:91]
target <- music[,1]
target <- as.data.frame(target)
```

Scale and Center Transformation

```
scaling <- scale(predictors, center = TRUE, scale = TRUE)
scaling <- as.data.frame(scaling)
head(scaling)
```

	V2	V3	V4	V5	V6	V7
1	1.0805738	0.39126500	1.8265307	0.46465620	-0.474729066	-0.2782038
2	0.8809185	0.33229215	1.7485375	0.72182731	-0.164944871	-1.1911721
3	1.2476212	0.59259903	1.3371720	0.75065633	-0.001110167	-0.7020998
4	0.8010429	-0.06180501	0.7836825	0.08721766	0.329179447	-1.2984272
5	1.2497736	0.79333361	1.6570354	0.44745939	-0.406774707	-0.5671376
6	1.1801361	-0.01888072	2.3729989	1.30019658	-0.829590889	-0.7409775

Figure 2 First 6 Rows of Scale and Center Transformation

Square Transformation

```
square <- predictors^2
head(square)
```

	V2	V3	V4	V5	V6	V7
1	2494.360	4.610099e+02	5340.321	76.538177	302.9785834	171.5851
2	2374.822	3.396391e+02	4945.857	167.608237	106.5926159	616.9148
3	2596.630	1.014806e+03	3115.706	180.014011	43.2829778	344.0802
4	2327.821	3.603809e+00	1317.524	6.696502	0.9442009	687.3222
5	2597.961	1.781682e+03	4502.362	71.705500	251.3109508	282.7136
6	2555.067	9.965386e-02	8528.644	501.175978	651.2040497	362.8751

Figure 3 First 6 Rows of Square Transformation

Square Root Transformation

Does not work since we have negative values

```
square_root <- sqrt(predictors)

head(square_root)
```

	V2	V3	V4	V5	V6	V7
1	7.067076	4.6336961	8.548538	2.957805	NaN	NaN
2	6.980842	4.2929361	8.386107	3.598105	NaN	NaN
3	7.138427	5.6441137	7.471179	3.662913	NaN	NaN
4	6.946042	NaN	6.024759	1.608652	0.9857484	NaN
5	7.139342	6.4969208	8.191437	2.909967	NaN	NaN
6	7.109688	0.5618541	9.609925	4.731486	NaN	NaN

Figure 4 First 6 Rows of Square Root Transformation

Log Transformation

```
#does not work since we have values outside the log domain

logging <- log(predictors)

head(logging)
```

	V2	V3	V4	V5	V6	V7
1	3.910894	3.066710	4.291521	2.1688948	NaN	NaN
2	3.886339	2.913942	4.253153	2.5608147	NaN	NaN
3	3.930985	3.461226	4.022106	2.5965173	NaN	NaN
4	3.876344	NaN	3.591755	0.9507926	-0.02870816	NaN
5	3.931241	3.742657	4.206179	2.1362837	NaN	NaN
6	3.922917	-1.153026	4.525593	3.1084786	NaN	NaN

Figure 5 First 6 Rows of Log Transformation

No Transformation

```
#I will just take the raw data and apply a regression model in part 3

head(predictors)
```

	V2	V3	V4	V5	V6	V7
1	49.94357	21.47114	73.07750	8.74861	-17.40628	-13.09905
2	48.73215	18.42930	70.32679	12.94636	-10.32437	-24.83777
3	50.95714	31.85602	55.81851	13.41693	-6.57898	-18.54940
4	48.24750	-1.89837	36.29772	2.58776	0.97170	-26.21683
5	50.97020	42.20998	67.09964	8.46791	-15.85279	-16.81409
6	50.54767	0.31568	92.35066	22.38696	-25.51870	-19.04928

Figure 6 First 6 Rows of No Transformation

I decided to use the square transformation for my regression model because all the values would be positive and no negative values to lower the regression performance metrics.

Part II Principal Component Analysis:

Statistics:

```
Importance of components:
                          PC1       PC2       PC3       PC4       PC5       PC6       PC7       PC8       PC9      PC10
Standard deviation     2114.5241 1173.3530 934.39655 686.28736 543.26793 465.0938 412.47135 394.1006 384.3528 347.31886
Proportion of Variance    0.4691    0.1444   0.09161   0.04942   0.03097   0.0227   0.01785   0.0163   0.0155   0.01266
Cumulative Proportion     0.4691    0.6136   0.70519   0.75461   0.78557   0.8083   0.82612   0.8424   0.8579   0.87057
                         PC11      PC12      PC13      PC14      PC15      PC16      PC17      PC18      PC19      PC20
Standard deviation     331.59208 305.27773 284.6725 261.9852 259.77490 250.49117 233.78581 228.9330 222.12769 195.67094
Proportion of Variance   0.01154   0.00978   0.0085   0.0072   0.00708   0.00658   0.00573   0.0055   0.00518   0.00402
Cumulative Proportion    0.88211   0.89189   0.9004   0.9076   0.91467   0.92125   0.92699   0.9325   0.93767   0.94168
                         PC21      PC22      PC23      PC24      PC25      PC26      PC27      PC28      PC29      PC30
Standard deviation     193.13842 190.2761 186.18852 180.58075 175.90039 170.43092 158.50426 152.83201 145.28808 139.59058
Proportion of Variance   0.00391    0.0038   0.00364   0.00342   0.00325   0.00305   0.00264   0.00245   0.00221   0.00204
Cumulative Proportion    0.94560    0.9494   0.95303   0.95645   0.95970   0.96275   0.96538   0.96783   0.97005   0.97209
                         PC31      PC32      PC33      PC34      PC35      PC36      PC37      PC38      PC39      PC40
Standard deviation     134.4359 124.73564 122.53978 114.69252 113.03762 109.83025 108.63877 106.20492 105.44865 100.95009
Proportion of Variance   0.0019   0.00163   0.00158   0.00138   0.00134   0.00127   0.00124   0.00118   0.00117   0.00107
Cumulative Proportion    0.9740   0.97562   0.97720   0.97858   0.97992   0.98118   0.98242   0.98361   0.98477   0.98584
                         PC41     PC42     PC43     PC44     PC45     PC46     PC47     PC48     PC49     PC50     PC51
Standard deviation     100.38665 97.6983 93.59843 88.89145 87.97714 86.17064 83.11225 79.97749 78.94344 75.24462 72.70610
Proportion of Variance   0.00106  0.0010  0.00092  0.00083  0.00081  0.00078  0.00072  0.00067  0.00065  0.00059  0.00055
Cumulative Proportion    0.98690  0.9879  0.98882  0.98965  0.99046  0.99124  0.99197  0.99264  0.99329  0.99388  0.99444
                         PC52     PC53     PC54     PC55     PC56     PC57     PC58     PC59     PC60     PC61     PC62
Standard deviation     67.29597 63.82520 63.49717 61.8572 60.47257 57.35732 53.6809 51.85366 50.64156 47.96807 47.17692
Proportion of Variance  0.00048  0.00043  0.00042  0.0004  0.00038  0.00035  0.0003  0.00028  0.00027  0.00024  0.00023
Cumulative Proportion   0.99491  0.99534  0.99576  0.9962  0.99655  0.99689  0.9972  0.99748  0.99775  0.99799  0.99822
                         PC63     PC64     PC65     PC66     PC67     PC68     PC69     PC70     PC71     PC72     PC73
Standard deviation     44.22680 39.96713 39.43612 38.48435 37.65337 36.55665 34.17082 29.89835 29.54970 28.46362 26.13230
Proportion of Variance  0.00021  0.00017  0.00016  0.00016  0.00015  0.00014  0.00012  0.00009  0.00009  0.00009  0.00007
Cumulative Proportion   0.99843  0.99860  0.99876  0.99891  0.99906  0.99920  0.99933  0.99942  0.99951  0.99960  0.99967
                         PC74     PC75     PC76     PC77     PC78     PC79     PC80     PC81     PC82     PC83     PC84
Standard deviation     24.85618 22.43649 20.27783 19.78895 18.54897 14.35933 1.4e+01 10.91288 9.75821 8.96535 7.77204
Proportion of Variance  0.00006  0.00005  0.00004  0.00004  0.00004  0.00002 2.0e-05  0.00001 0.00001 0.00001 0.00001
Cumulative Proportion   0.99973  0.99979  0.99983  0.99987  0.99991  0.99993 1.0e+00  0.99996 0.99997 0.99998 0.99999
                          PC85  PC86  PC87  PC88  PC89  PC90
Standard deviation     6.97522 6.297 4.946 3.459 3.042 1.898
Proportion of Variance 0.00001 0.000 0.000 0.000 0.000 0.000
Cumulative Proportion  0.99999 1.000 1.000 1.000 1.000 1.000
```

Figure 7 PCA Statistics

Visualizations

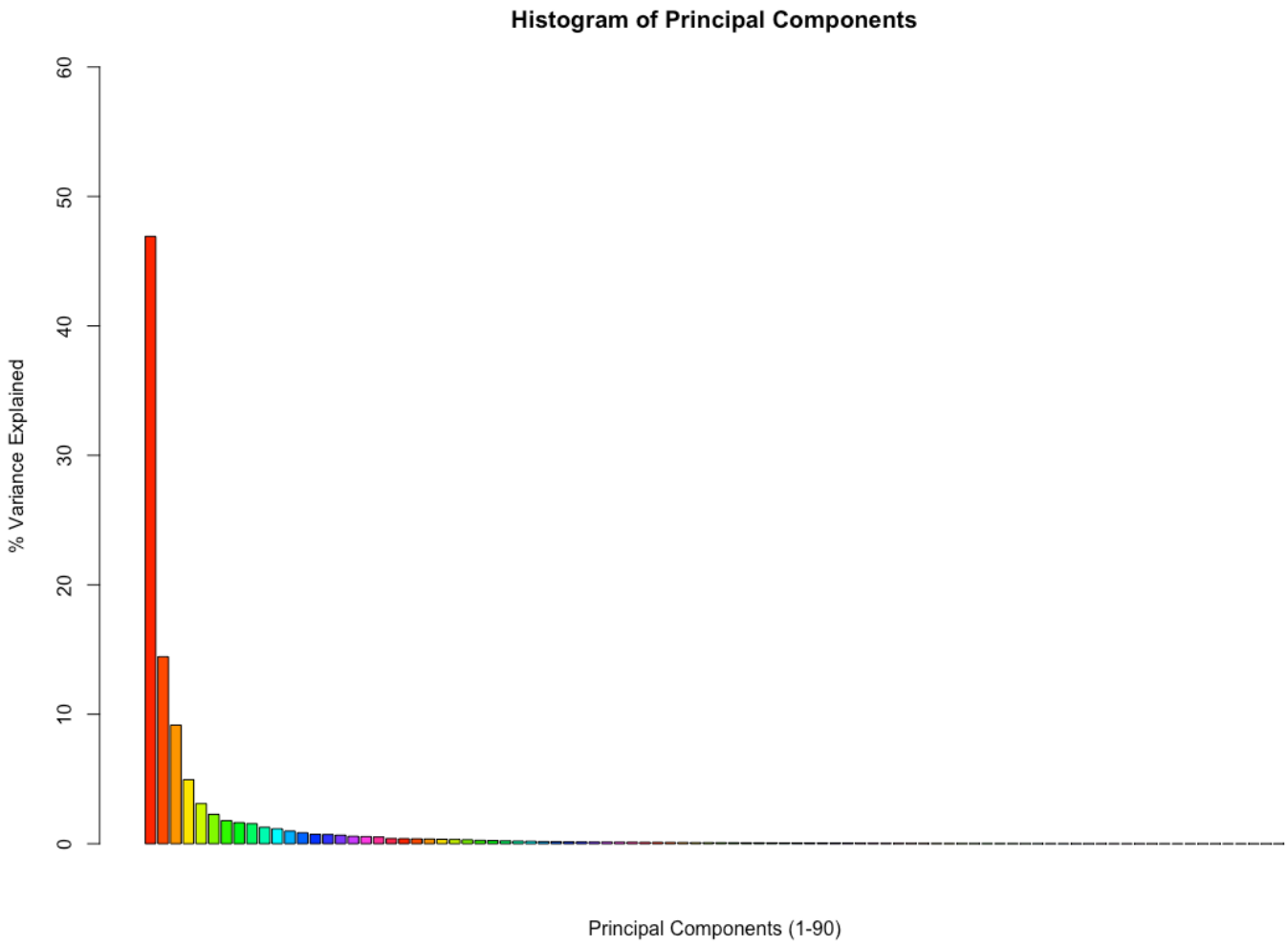

Figure 8 Variance Histogram of Principal Components

Looking at Figure 8 it looks like either 4 or 5 components is the optimal amount.

I now will cut the histogram to show the top 7 components.

```
barplot(100*p.variance, las = 3 , ylim = c(0,60),
    ylab = '% Variance Explained', main = "Histogram of Principal Components",
    col = rainbow(20), xlab = 'Principal Components (1-6)', xlim = c(0,6))
```

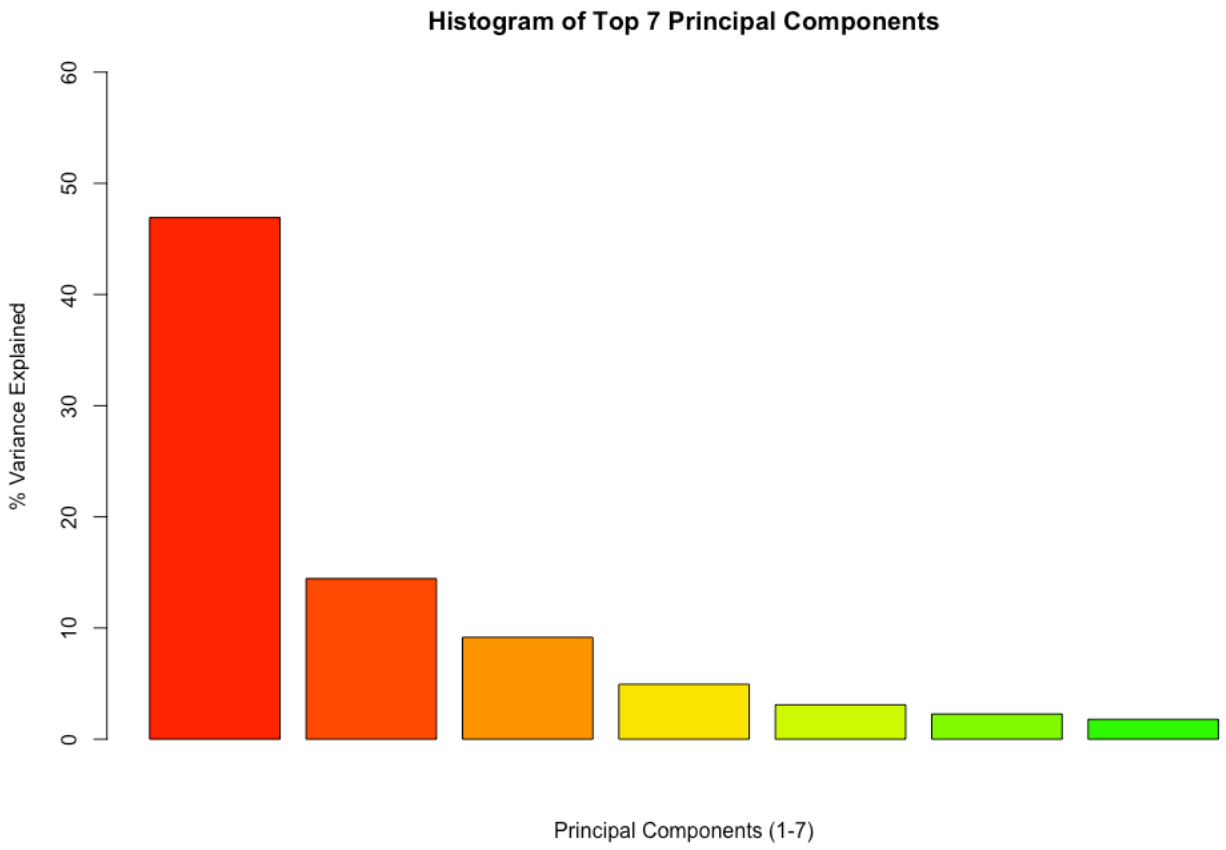

Figure 9 Histogram of Top 7 Principal Components

I decided to use four Principal Components since at the fifth component the variance starts to level out as shown in Figure 9.

Part III Regression Models:

```
# Function that returns Root Mean Squared Error
rmse <- function(error)
{
  sqrt(mean(error^2))
}
```

Untransformed model RMSE: 9.479

```
#running regression for untransformed predictors

untransformed_test <- music[463716:515345,]

model1 <- lm(V1 ~., data = untransformed_test)

summary(model1)

rmse_model1 <- rmse(model1$residuals)
```

```
rmse_model1
[1] 9.47909
```

Transformed model RMSE: 10.23

```
#running regression for transformed predictors using square transformations
transformed_test <- cbind(target, square)
transformed_test <- transformed_test[463716:515345,]
model2 <- lm(target ~., data = transformed_test)
summary(model2)
rmse_model2 <- rmse(model2$residuals)
rmse_model2
[1] 10.23846
```

Principal component model RMSE: 10.66

```
library(pls)
train_music <- music[1:463715,]
y_test <- music[463716:515345,1]
test_music <- music[463716:515345, 2:91]
pcr_model <- pcr(V1 ~., data = train_music, scale = TRUE, validation = "CV")
validationplot(pcr_model, main = "Root Mean Squared Error Prediction Plot")
```

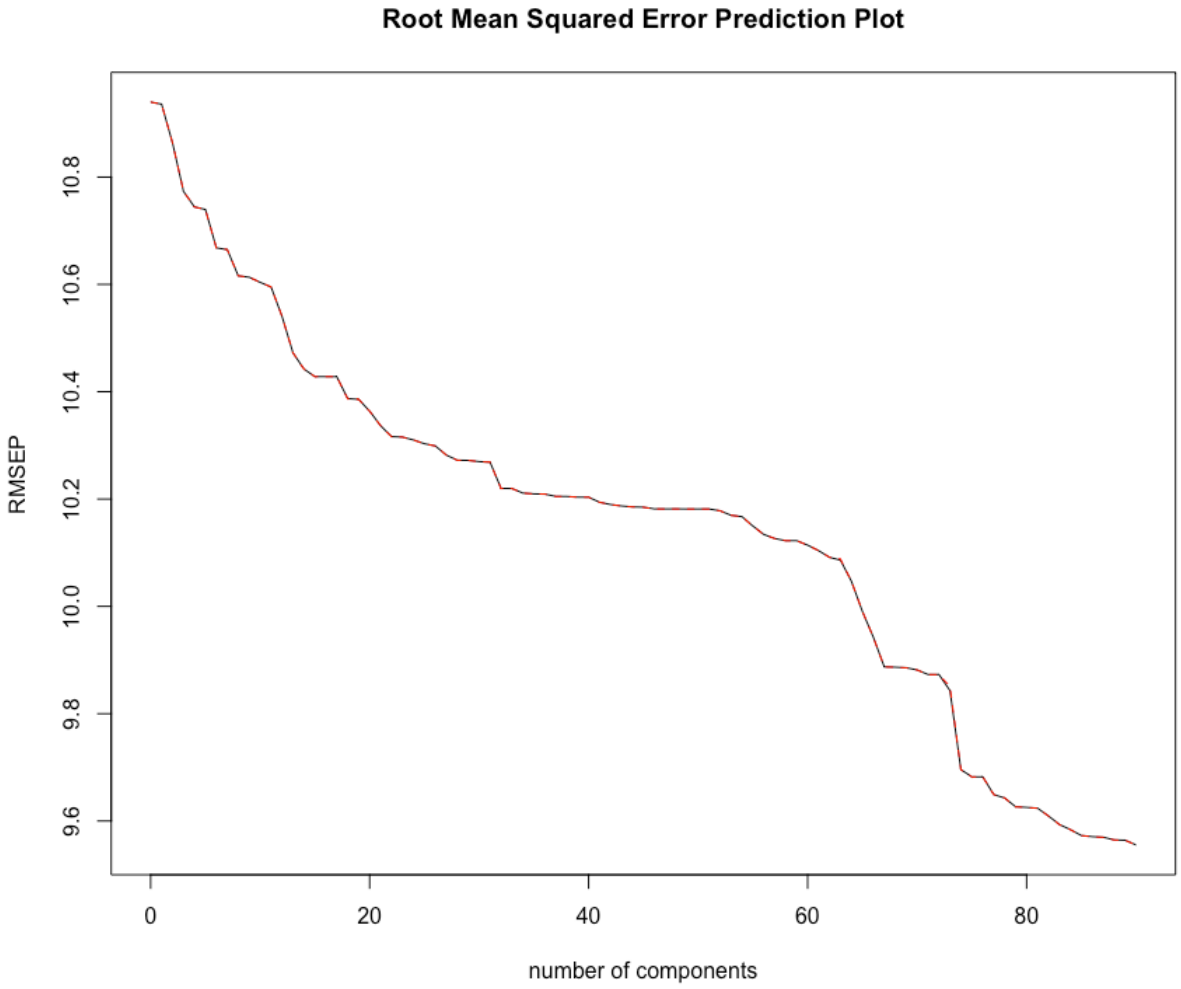

Figure 10 Root Mean Squared Prediction Plot for Principal Component Regression

```
pcr_pred <- predict(pcr_model, test_music, ncomp = 4)
sqrt(mean((pcr_pred - y_test)^2))
[1] 10.66351
```