Principal component analysis

Anonymous

2021-10-26

Contents

1. Setting up

1.1 Load the required packages

```
using Pkg, Weave, CSV, DataFrames, Plots, MultivariateStats, Statistics
```

1.2 Import the data

```
data = DataFrame(CSV.File(joinpath(path, "03_final-data",
                                        "03_final-data_2021",
                                        "final-data.csv")))
```

1.3 Transform variables

```
# Square root of ambush time and chase duration
data[:, :sqrt_ambush_time] = sqrt.(data[:, :ambush_time_close])
data[:, :sqrt_total_chase_duration] = sqrt.(data[:, :total_chase_duration])
```

```
# Log+1 of the latency and chase count
data[:, :log_latency_1st_capture] = log1p.(data[:, :latency_1st_capture])
data[:, :log_chase_count] = log1p.(data[:, :chase_count])
```

1.4 Select the variables and rearrange the table

```
# Subset the columns used for the analysis
select!(data, [:pred_speed, :pred_amount_tiles_visited,
               :sqrt_ambush_time, :log_latency_1st_capture,
               :log_chase_count, :sqrt_total_chase_duration])

# Re-arrange the table in wide format
data_matrix = Array(data)'
```

2. Run the principal component analysis

2.1 Run the PCA on the data matrix

```
# Fit the pca to the data matrix
pca = fit(PCA, data_matrix; pratio = 1, maxoutdim = 4)
```

```
PCA(indim = 6, outdim = 4, principalratio = 0.9953922286008744)
```

2.2 Check some diagnostics

```
# Total variance
principalvars(pca)

# % of variance
principalvars(pca) ./ tvar(pca) * 100
```

```
4-element Vector{Float64}:
 74.87038460871871
 16.657100394028987
  7.397625179339015
  0.6141126780007244
```

2.3 Transform observations to principal components

```
transformed = projection(pca)' * (data_matrix .- mean(pca))
```

```
4×80840 Matrix{Float64}:
 -15.7302    -3.51424    6.16461   …  -11.7483    -9.83942    3.60072
  -6.01508   -6.69037   -4.15033        1.80566    2.24093    0.0646238
  -1.75506   -2.12256   -3.53524        1.59429    0.399181  -0.86747
```

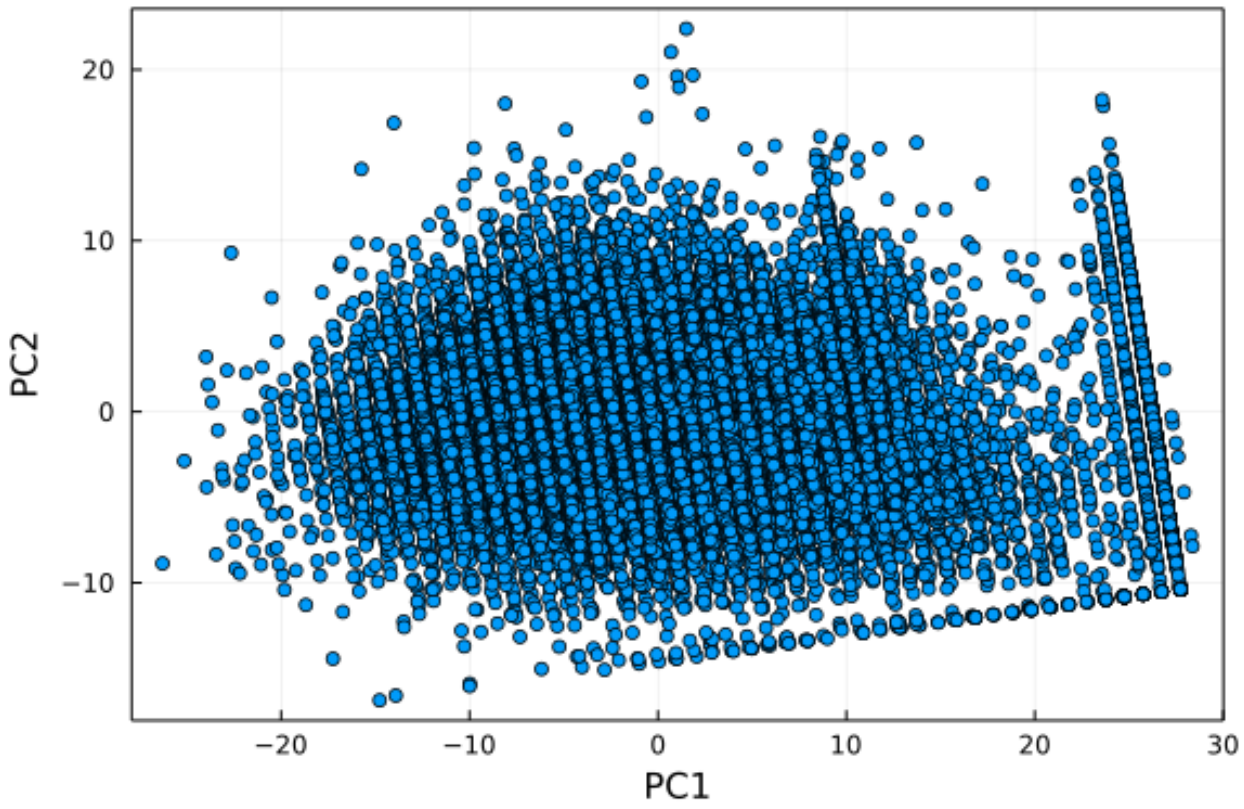
```
-0.641731    0.439463    0.349572        0.111684   -0.226801    0.331564
```

3. Plot the principal component analysis

3.1 Biplot of PC1 and PC2

```
# Compute the plot
biplot = plot(transformed[1, :],
              transformed[2, :],
              seriestype = :scatter,
              label = "")

# Format the plot
plot!(xlabel = "PC1",
      ylabel = "PC2",
      framestyle = :box)
```




```
# Compute the projection matrix
proj = projection(pca)

# Add arrows (variables) to the biplot
for i = 1:6; plot!([0, proj[i, 1]],
                   [0, proj[i, 2]],
                   arrow = true,
                   label = names(data)[i],
```

```
                    legend = :bottomleft);
end

display(biplot)
```