Andrew's Curves

MOD16

0:00 / 2:01 1.0x

Start of transcript. Skip to the end.

An Andrew's plot, also known as an Andrew's curve,

helps you visualize higher-dimensionality,

multivariate data by plotting your data set samples as curves.

The feature values of your samples act as the coefficients

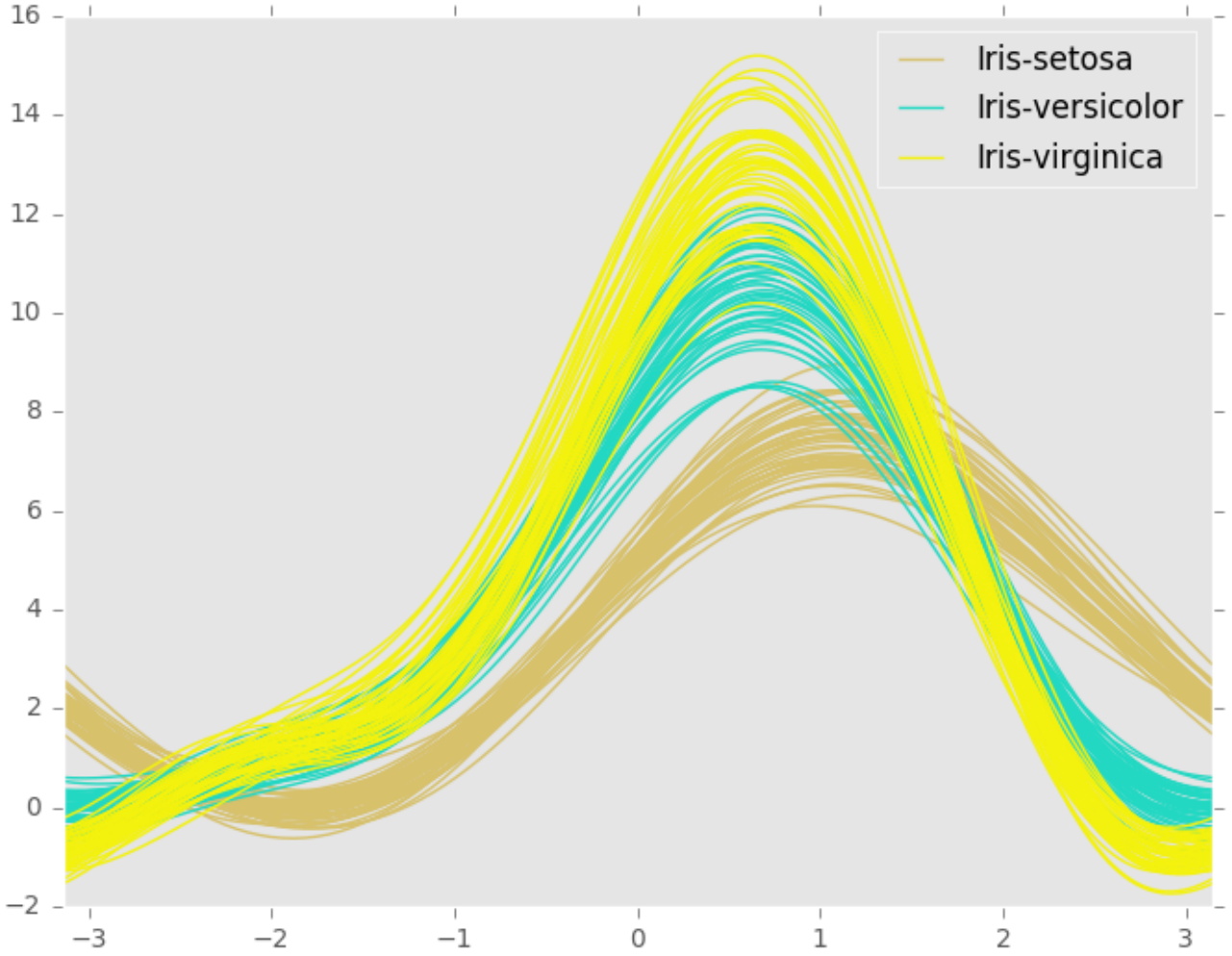
of the curve.

So just as with parallel

Download video Download transcript .srt

An Andrews plot, also known as Andrews curve, helps you visualize higher dimensionality, multivariate data by plotting each of your dataset's observations as a curve. The feature values of the observation act as the coefficients of the curve, so observations with similar characteristics tend to group closer to each other. Due to this, Andrews curves have some use in outlier detection.

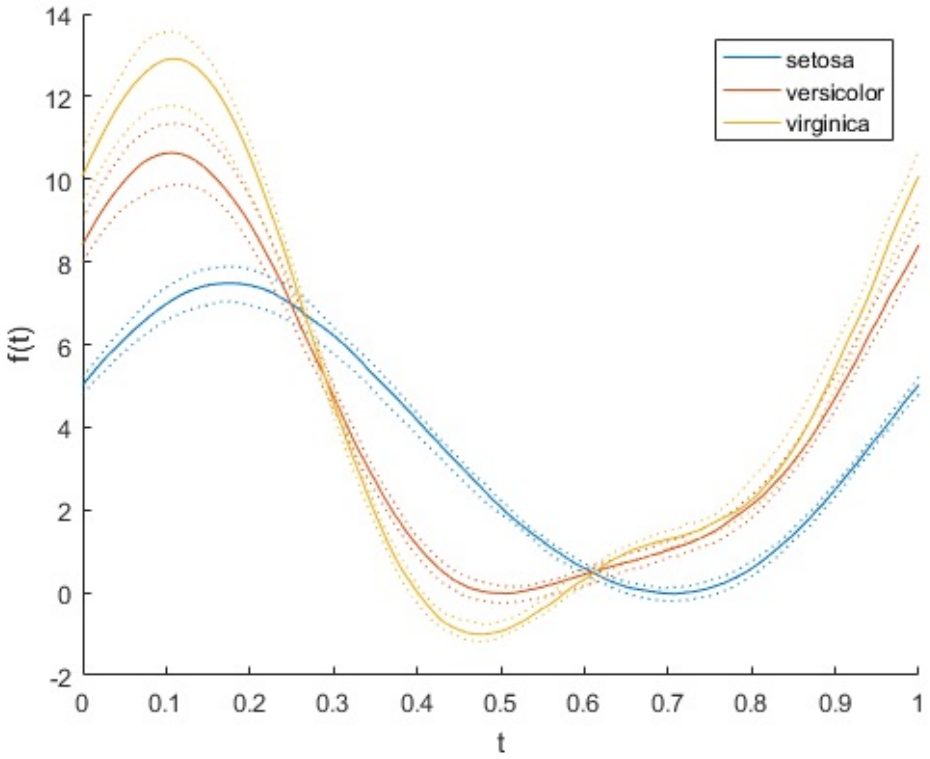
Just as with Parallel Coordinates, every plotted feature must be numeric since the curve equation is essentially the product of the observation's features vector (transposed) and the vector: (1/sqrt(2), sin(t), cos(t), sin(2t), cos(2t), sin(3t), cos(3t), ...) to create a Fourier series.



The Pandas implementation requires you once again specify a GroupBy feature, which is then used to color code the curves as well as produce as chart legend:

```
from sklearn.datasets import load_iris
from pandas.tools.plotting import andrews_curves

import pandas as pd
import matplotlib.pyplot as plt
import matplotlib

# Look pretty...
matplotlib.style.use('ggplot')

# Load up SKLearn's Iris Dataset into a Pandas Dataframe
data = load_iris()
df = pd.DataFrame(data.data, columns=data.feature_names)
df['target_names'] = [data.target_names[i] for i in data.target]

# Andrews Curves Start Here:
plt.figure()
andrews_curves(df, 'target_names')
plt.show()
```

One of the current weaknesses with the Pandas implementation (and this goes for Parallel Coordinates as well) is that every single observation is charted. In the MATLAB version, you can specify a quantile or probability distribution cutoff. This way, only the mean feature values for a specific group are plotted, with a transparent boundary around the cutoffs. If you feel up to the challenge, a straightforward bonus assignment for you is to take the existing Pandas Andrews curve implementation and extend it with said functionality.



(Desired MatLab Implementation)