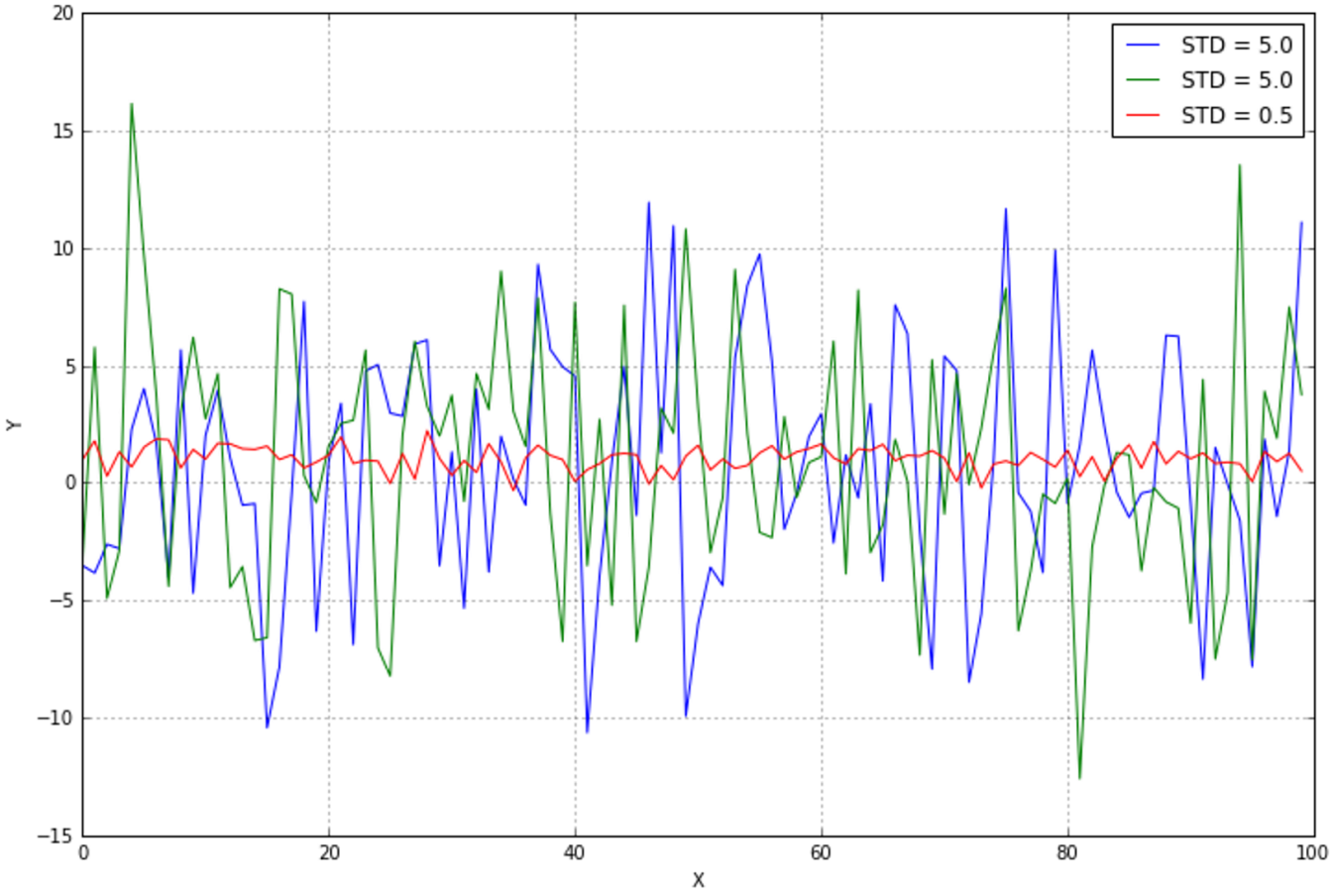
Feature selection and filtering

An unnormalized dataset with many features contains information proportional to the independence of all features and their variance. Let's consider a small dataset with three features, generated with random Gaussian distributions:



Even without further analysis, it's obvious that the central line (with the lowest variance) is almost constant and doesn't provide any useful information. If you remember the previous chapter, the entropy H(X) is quite small, while the other two variables carry more information. A variance threshold is, therefore, a useful approach to remove all those elements whose

contribution (in terms of variability and so, information) is under a predefined level. scikit-learn provides the class `VarianceThreshold` that can easily solve this problem. By applying it on the previous dataset, we get the following result:

Copy

```
from sklearn.feature_selection import VarianceThreshold

>>> X[0:3, :]
array([[-3.5077778 , -3.45267063,  0.9681903 ],...
```


‹ Previous Section (/book/big_data_and_business_intelligence/9781785889622/3/ch03lvl1s

Next Section › (/book/big_data_and_business_intelligence/9781785889622/3/ch03lvl1sec2