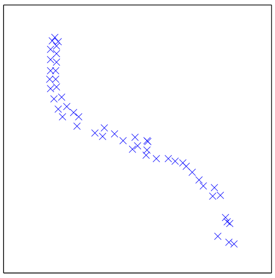
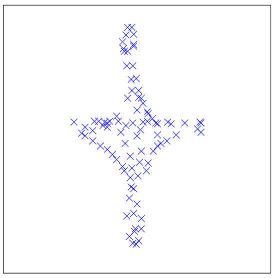
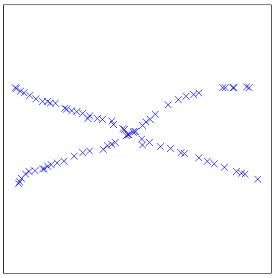
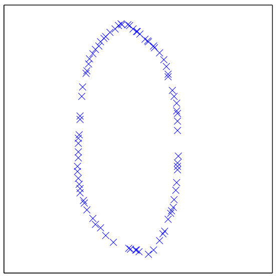
Assignment 7

Machine Learning, Summer term 2014, Ulrike von Luxburg

To be discussed in exercise groups on June 2-4

Exercise 1 (Direction of principal components, 1 point) Below are a number of 2D-data sets. Plot the two principal components.

Exercise 2 (Interpreting principal components, 2 points) A carsharing service runs a survey among 1000 students, who provide information concerning their 1- income, 2- distance they cover by car per month, 3- distance they cover by bike per month, 4- distance they cover by public transport per month, 5- distance they cover by foot per month. Then they run a PCA on the data. Provide answers to the following questions:

- What would it mean if a single eigenvector covered 95% of the total data variance?
- How would you interpret the result if the eigenvector $v_1 = [0, 0, 1, -1, 0]$ covers 90% of the total data variance?
- Why might it be necessary to rescale the data before running PCA in order to obtain a sensible result?

Exercise 3 (Generating samples from a Gaussian distribution, 0.5+0.5+0.5+1+0.5 points) You are given the mean μ and the covariance matrix Σ of a d-dimensional normal density $\mathcal{N}(\mu, \Sigma)$ and you want to sample n points from this density. Assuming that Σ is positive definite, the following MATLAB code will do this for you:

```
S1 = chol(Sigma); X = repmat(mu,n,1) + randn(n,d)*S1;
```

The command `S1 = chol(Sigma)` generates an upper triangular matrix `S1` which satisfies `Sigma=S1'*S1`. This decomposition is called the Cholesky decomposition. An alternative method, which also works when Σ is only positive semi-definite, is to decompose Σ to eigenvectors and eigenvalues by `[V,D] = eig(Sigma)` and then form `S2` by `S2=V*sqrt(D)`. However, the Cholesky decomposition is numerically more stable and computationally faster than eigen decomposition method.

(a) Show that in eigen decomposition, $\Sigma = S2 \cdot S2'$.

(b) Generate $n = 2000$ points in 3 dimensional space from a Gaussian distribution with mean `mu=[0,0,0]` and Covariance `Sigma=[2 0 0;0 1 0;0 0 4]`. Plot it with `plot3`.

(c) What are the eigenvalues and eigenvectors of the covariance matrix `Sigma`?

(d) Assume you know eigenvalues and eigenvectors of your covariance matrix:

$$\Lambda = \begin{bmatrix} 3 & 0 \\ & \end{bmatrix}$$