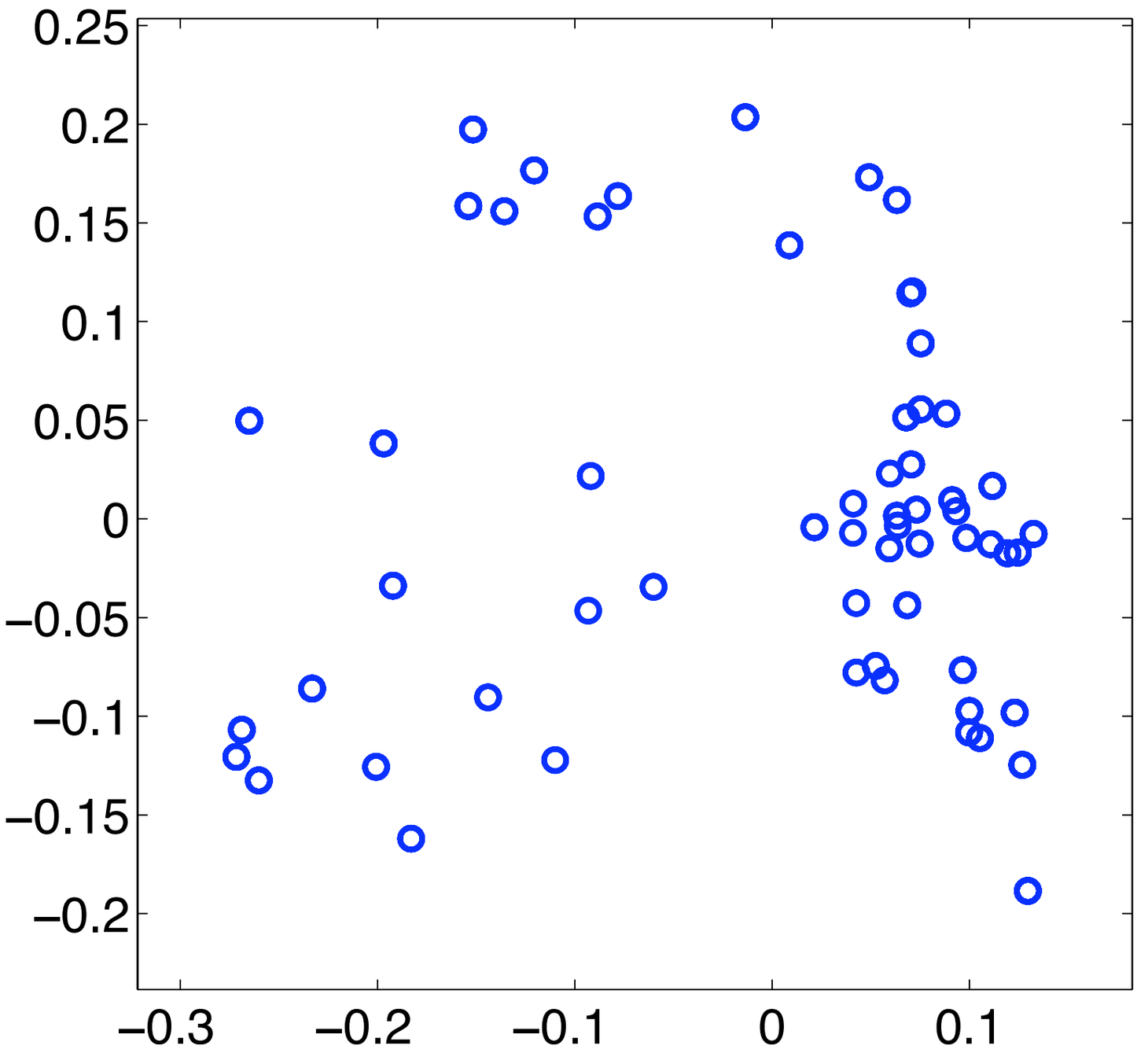
Clustering



- Given a set of points, break them into groups of "similar" points
- Why?
  -
  -
  -

Ex: vector quantization

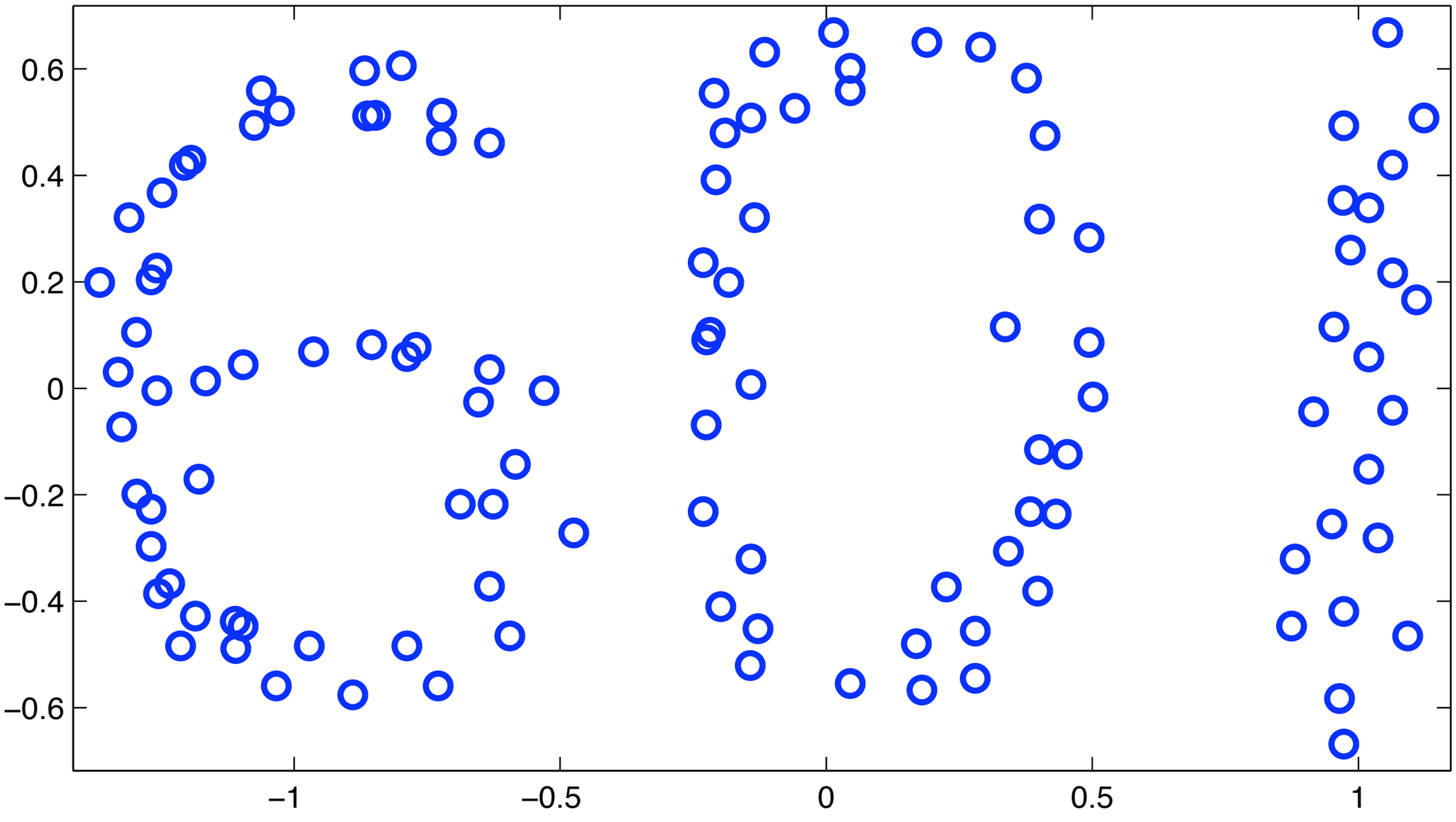
$K = 2$ $K = 3$ $K = 10$ Original image

figure 9.3 from Bishop

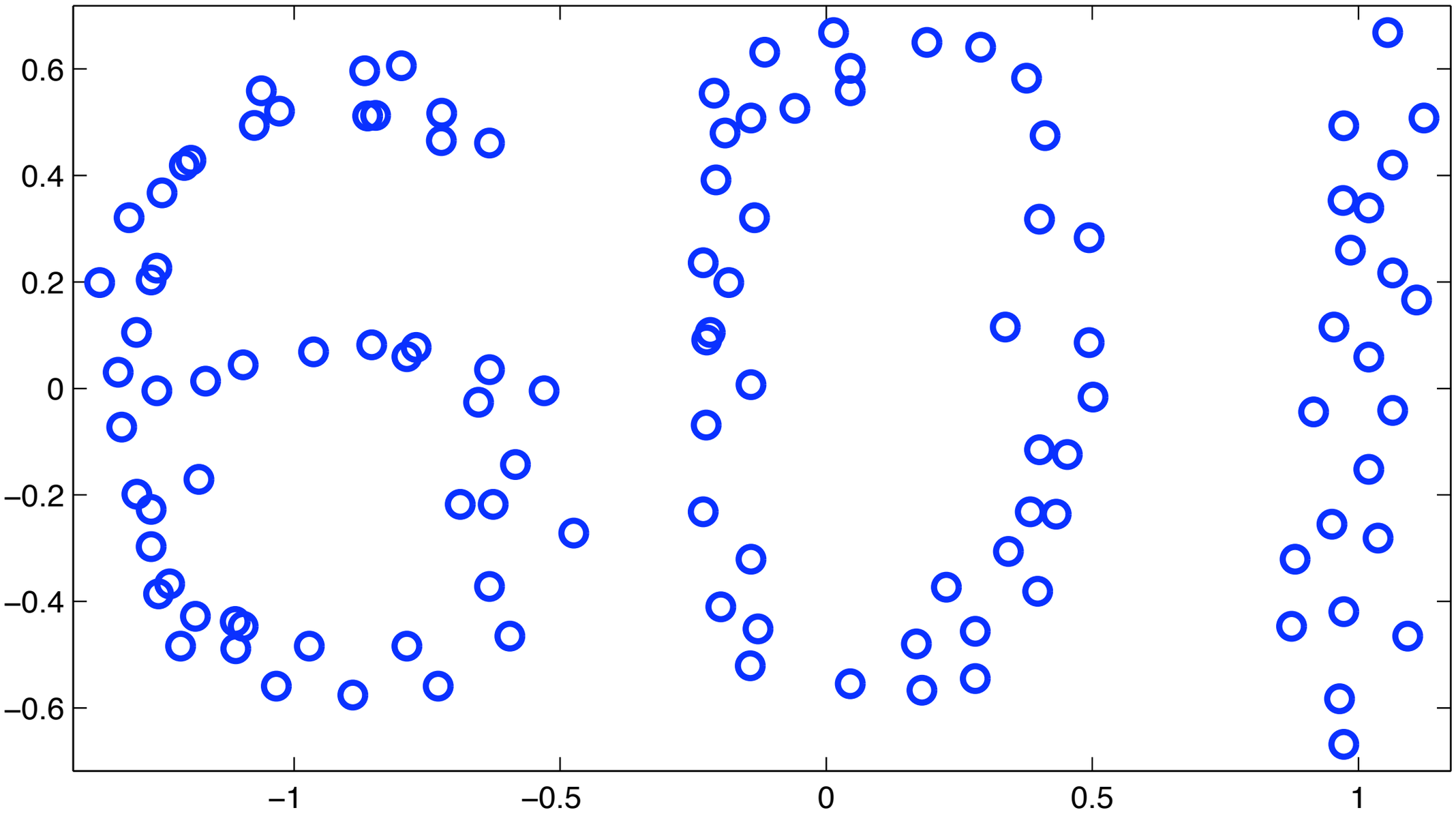
points = RGB tuples, 1 per pixel

cluster centers = the best RGB colors if we can use only K

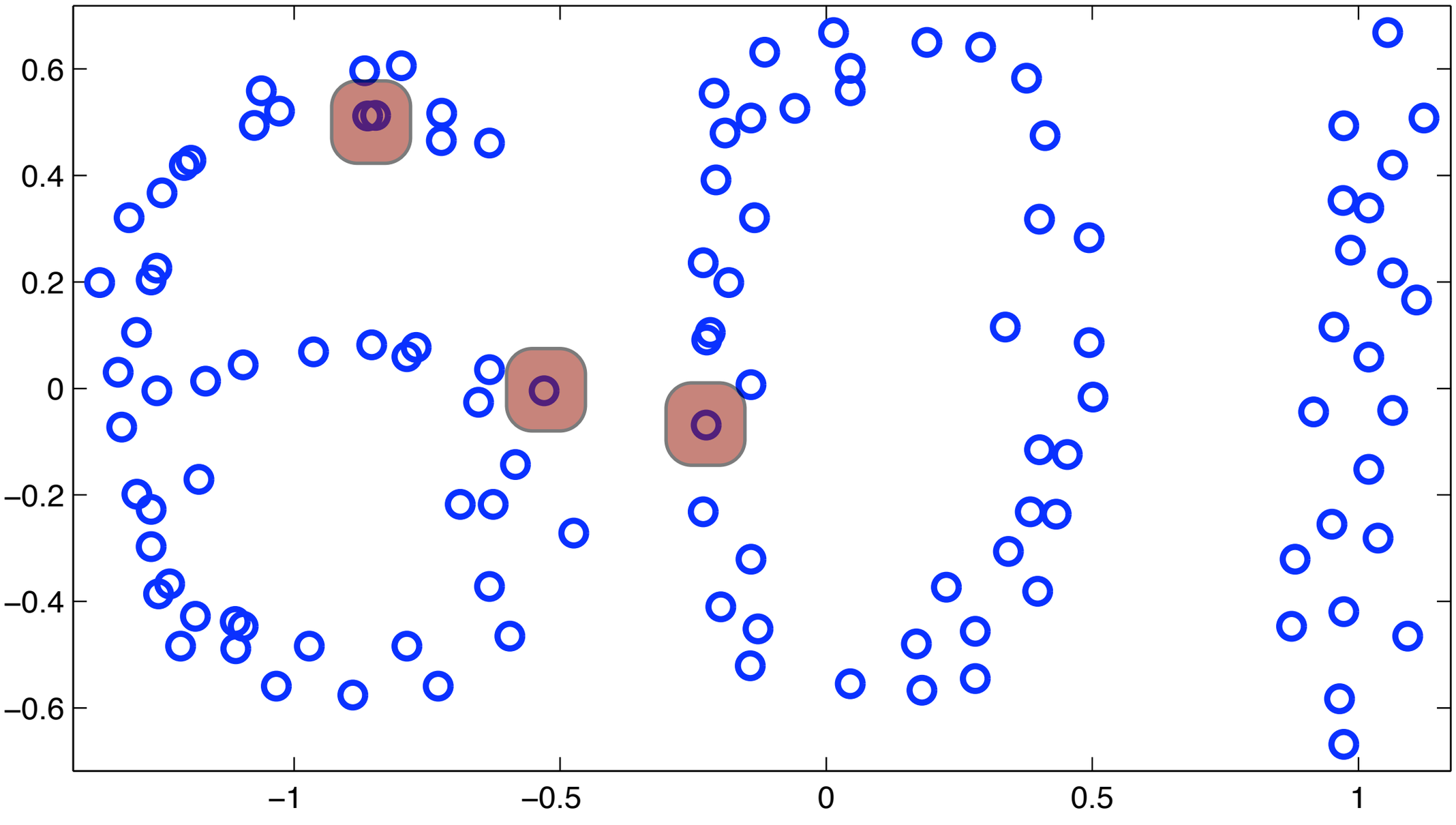
Clustering



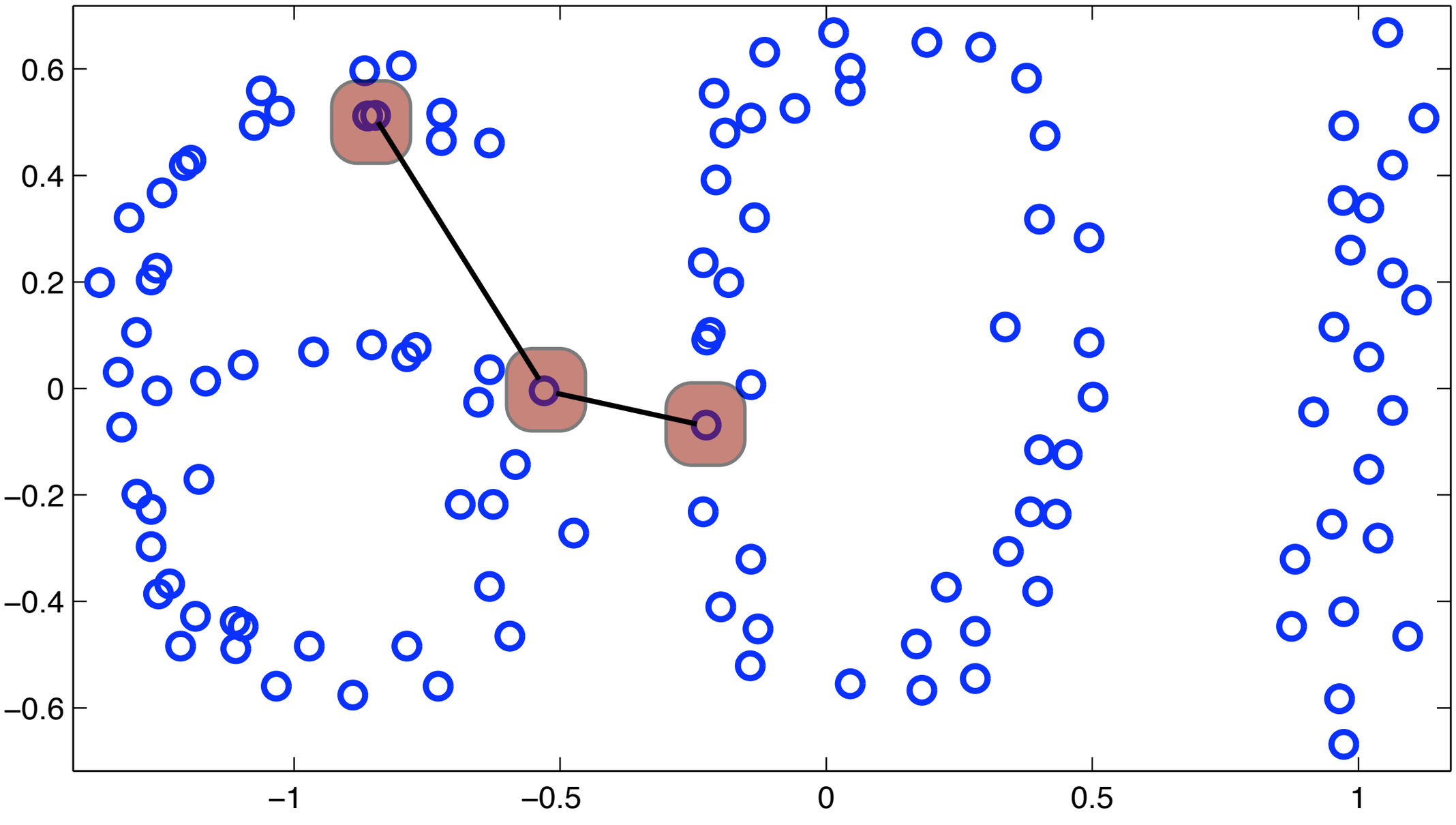
What do we mean by "similar"?



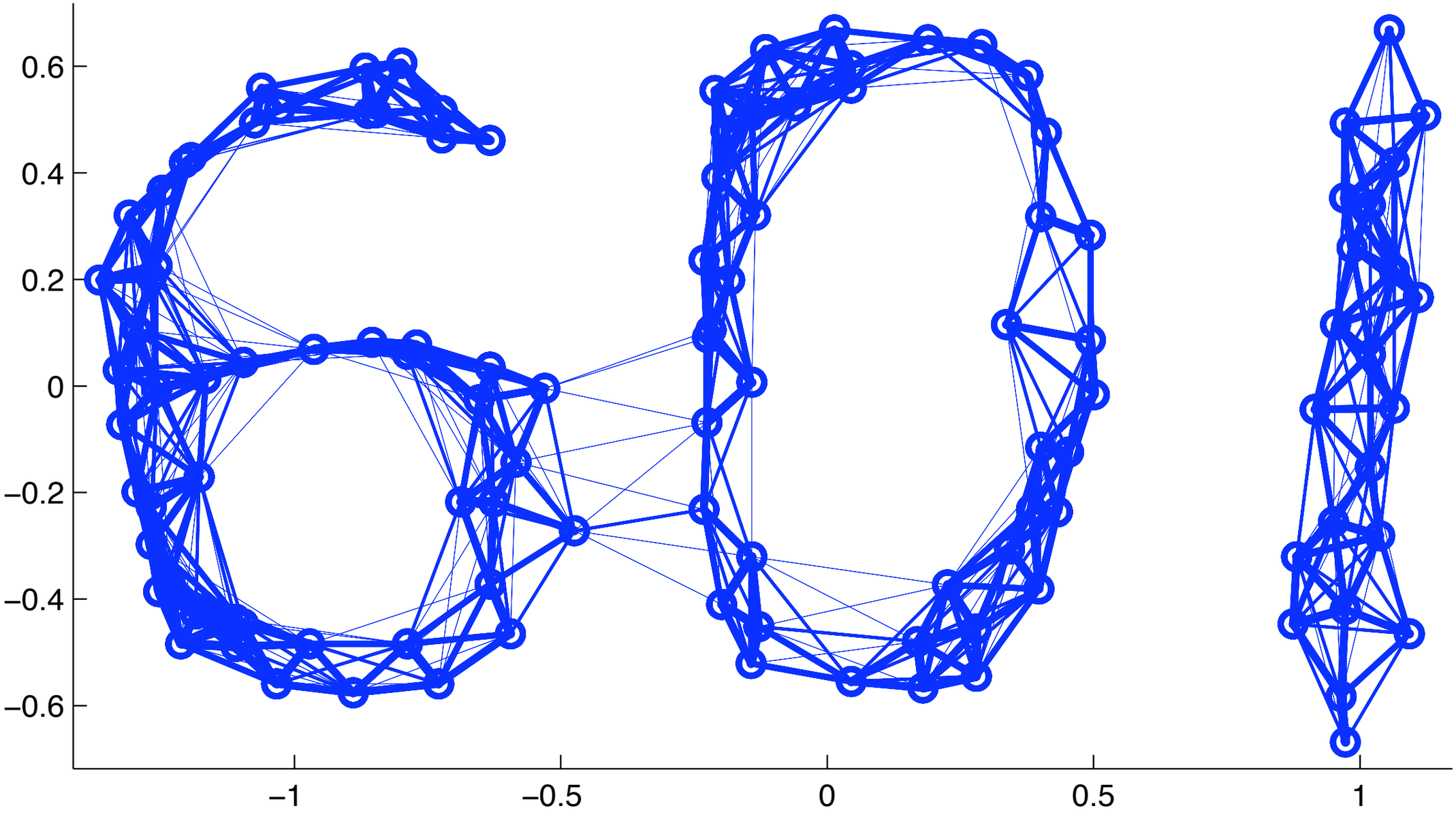
What do we mean by "similar"?



What do we mean by "similar"?



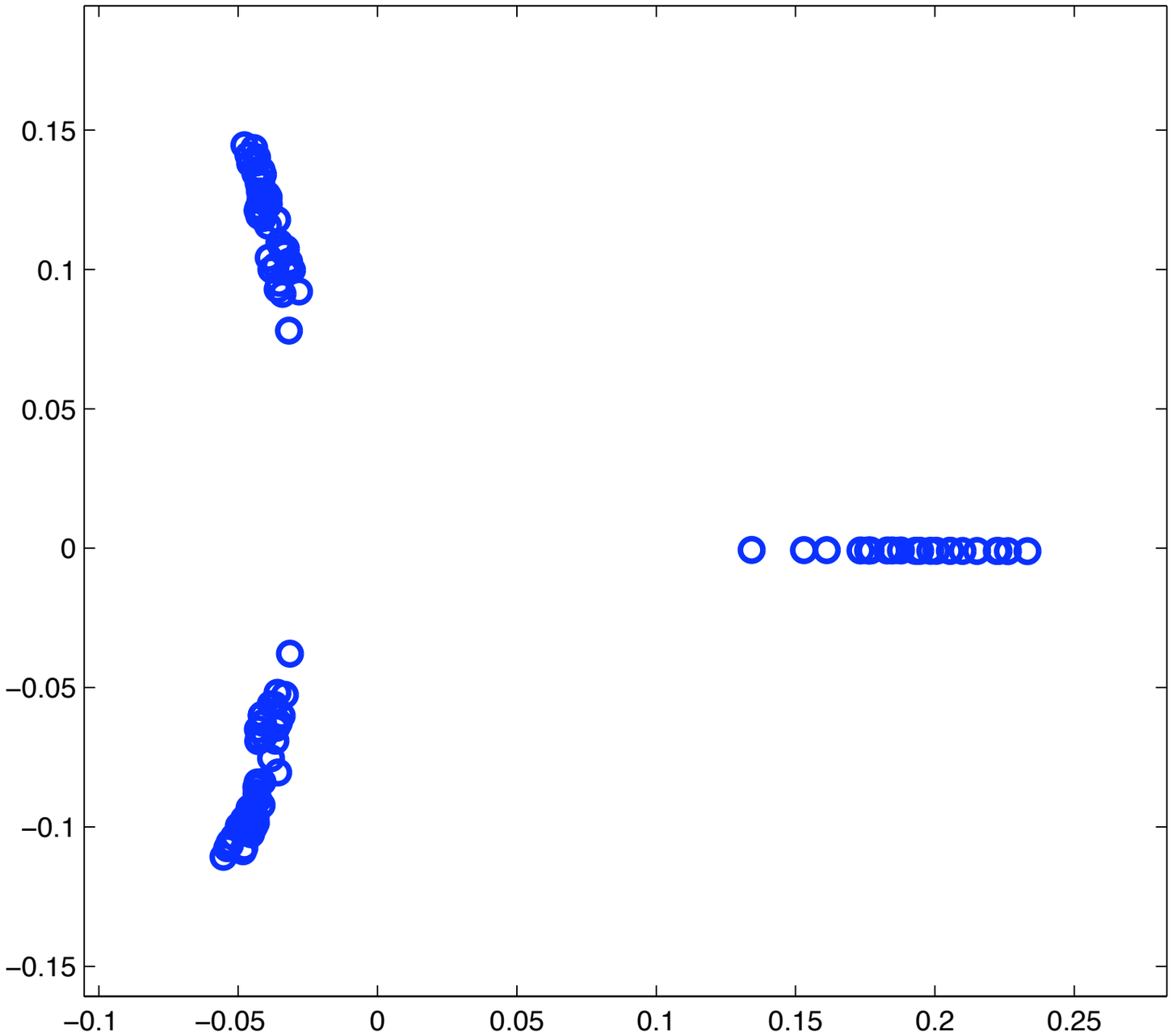
Spectral embedding: adjacency graph



Spectral embedding

$$
\begin{aligned}
A_{ij} &= e^{\|x_i - x_j\|^2 / 2\sigma^2} \\
d_i &= \sum_j A_{ij} \\
T_{ij} &= \frac{1}{\sqrt{d_i d_j}} A_{ij} \\
U \Sigma V^T &\approx T \\
y_i &= \Sigma U^T_{i,2:k+1}
\end{aligned}
$$

After embedding



Most common similarity metric

- …is just plain L_2 distance
  - perhaps with diagonal scaling, sqrt($x^T D x$)
  - rarely, sqrt($x^T A x$) for full, symmetric, positive definite A
- Many, many fancier metrics…
- But, after spectral embedding, L_2 is usually fine

A clustering objective

- Data $x_1, \ldots, x_n$
- Suppose k clusters
  - centers $\mu_1, \ldots, \mu_k$
- Let z_{ij} =
- L(z, μ) =

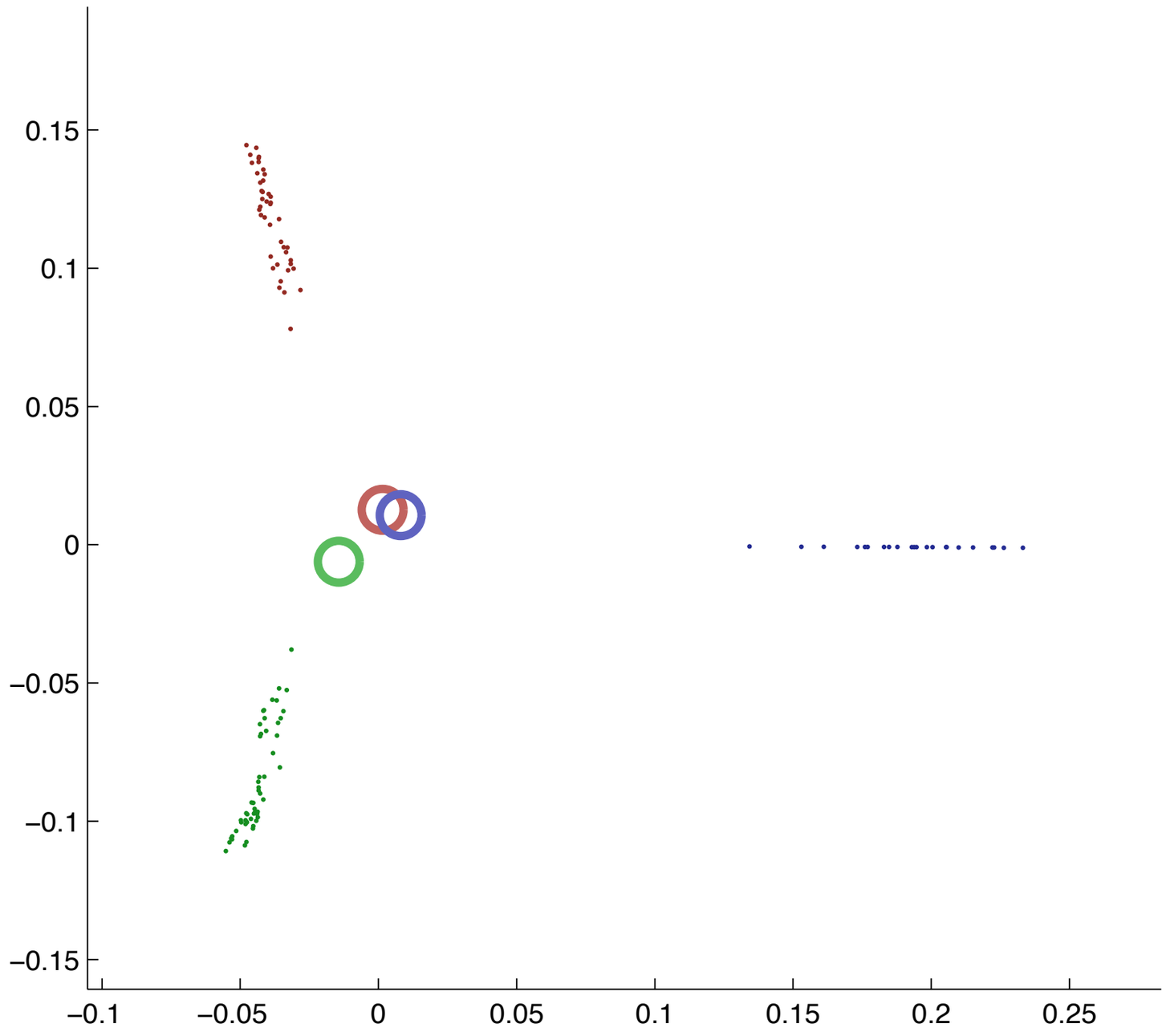
Optimization

- Finding $\min_{z,\mu} L(z, \mu)$ is an integer quadratic program
  - NP-complete
- Many common heuristics with varying degrees of success
  - in some circumstances, provable approximations

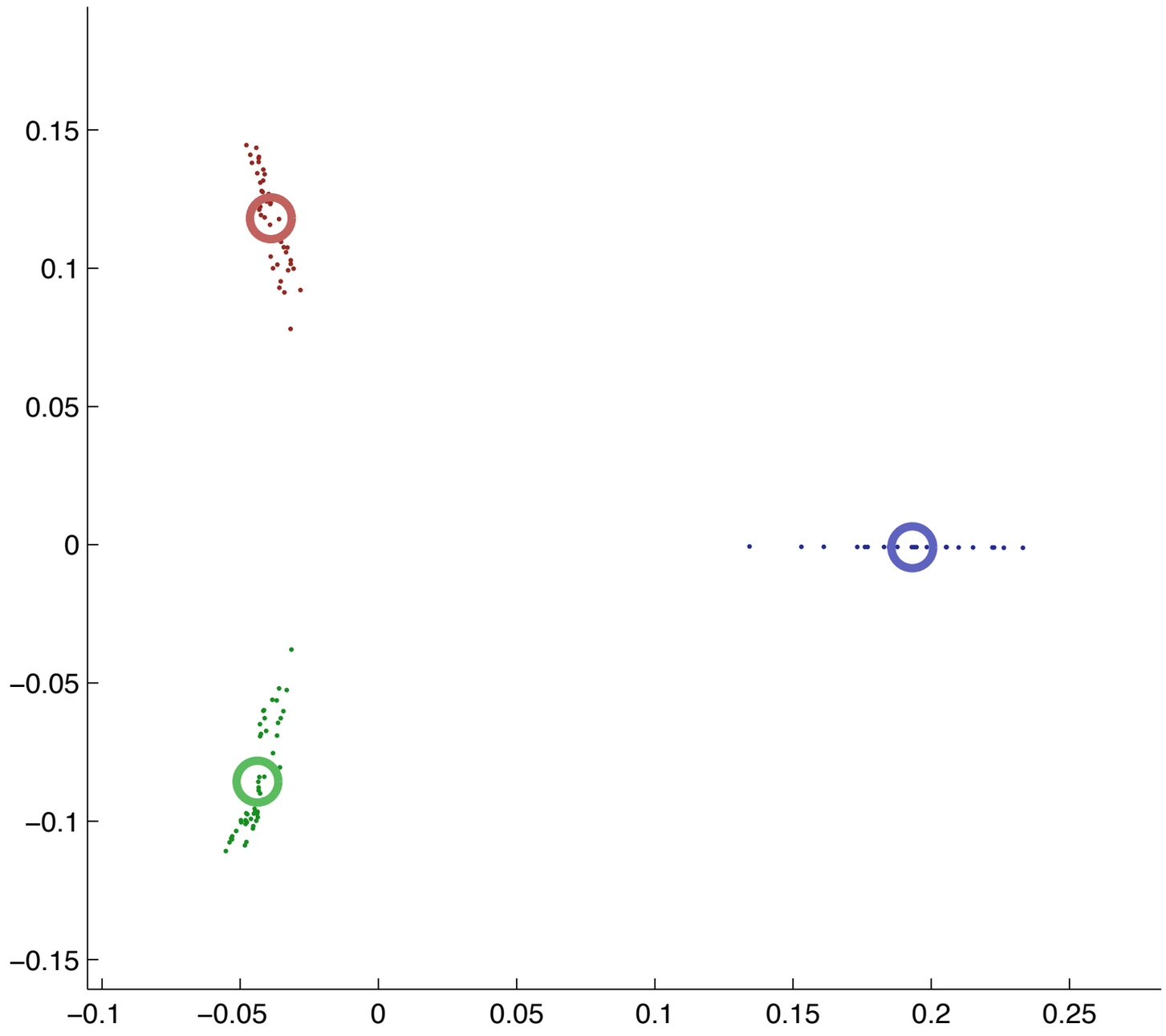
Alternating optimization for k-means

- Initialize z (or μ)
- Alternately minimize wrt. μ:
- and z, subject to

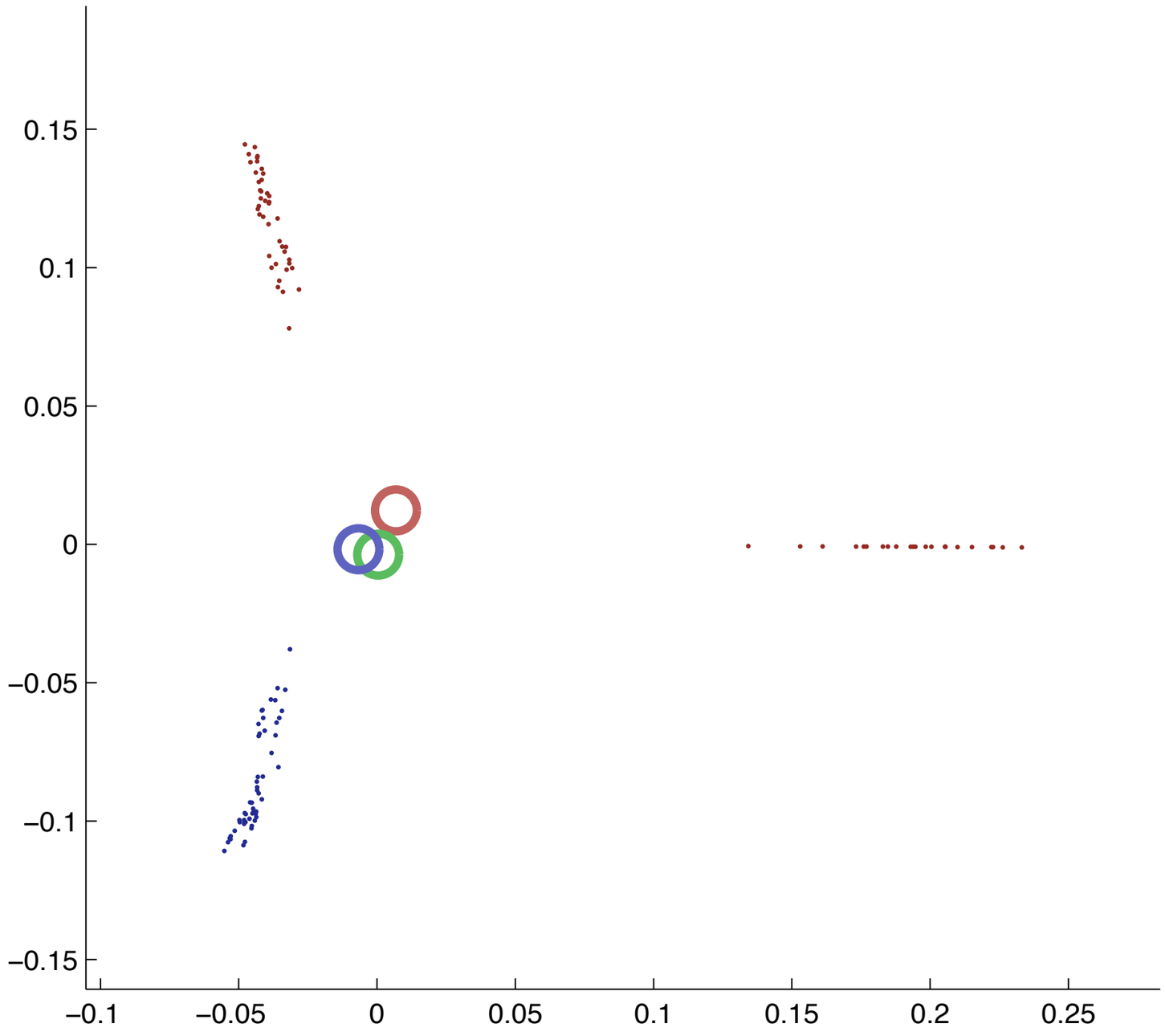
Example



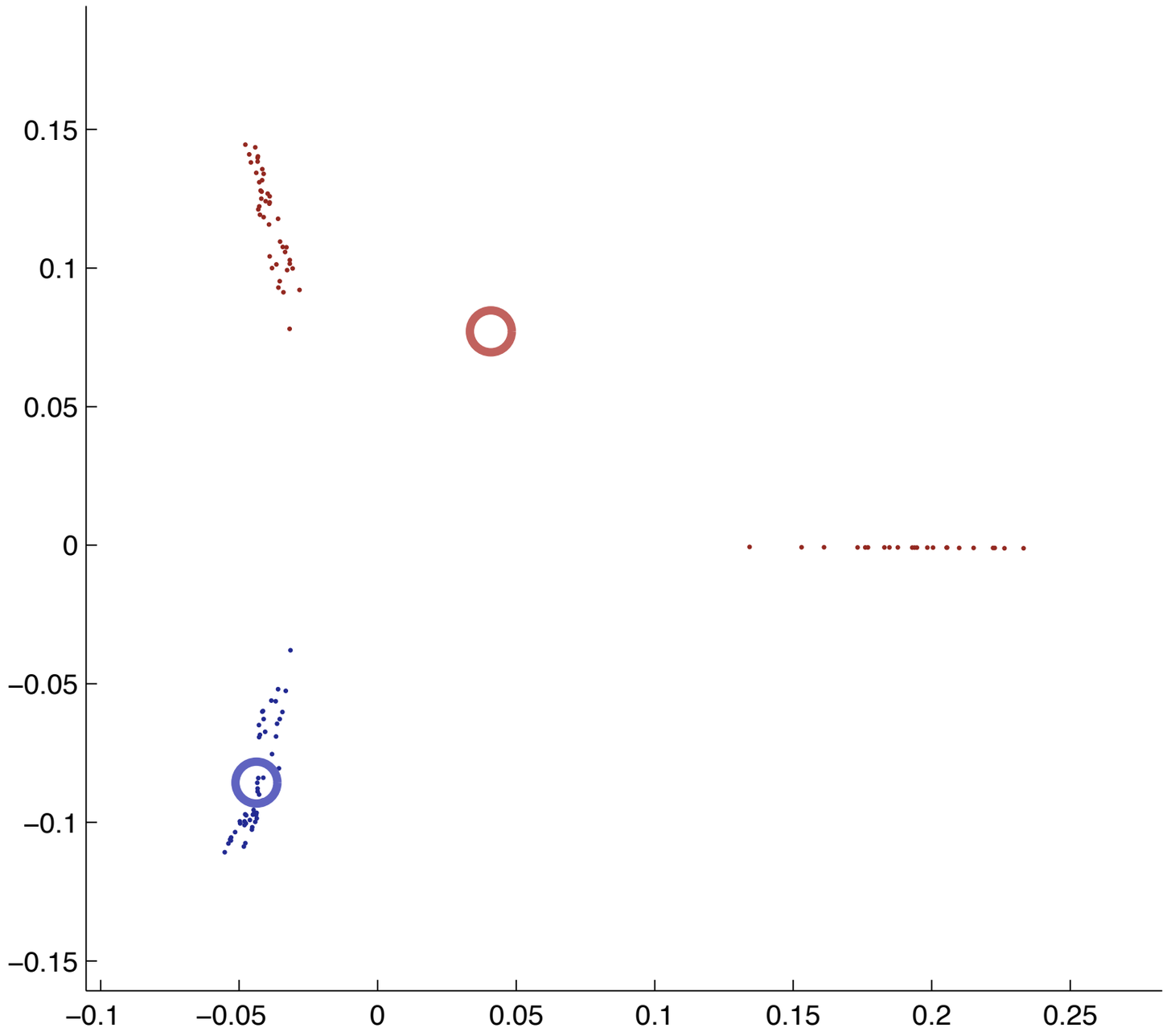
Example



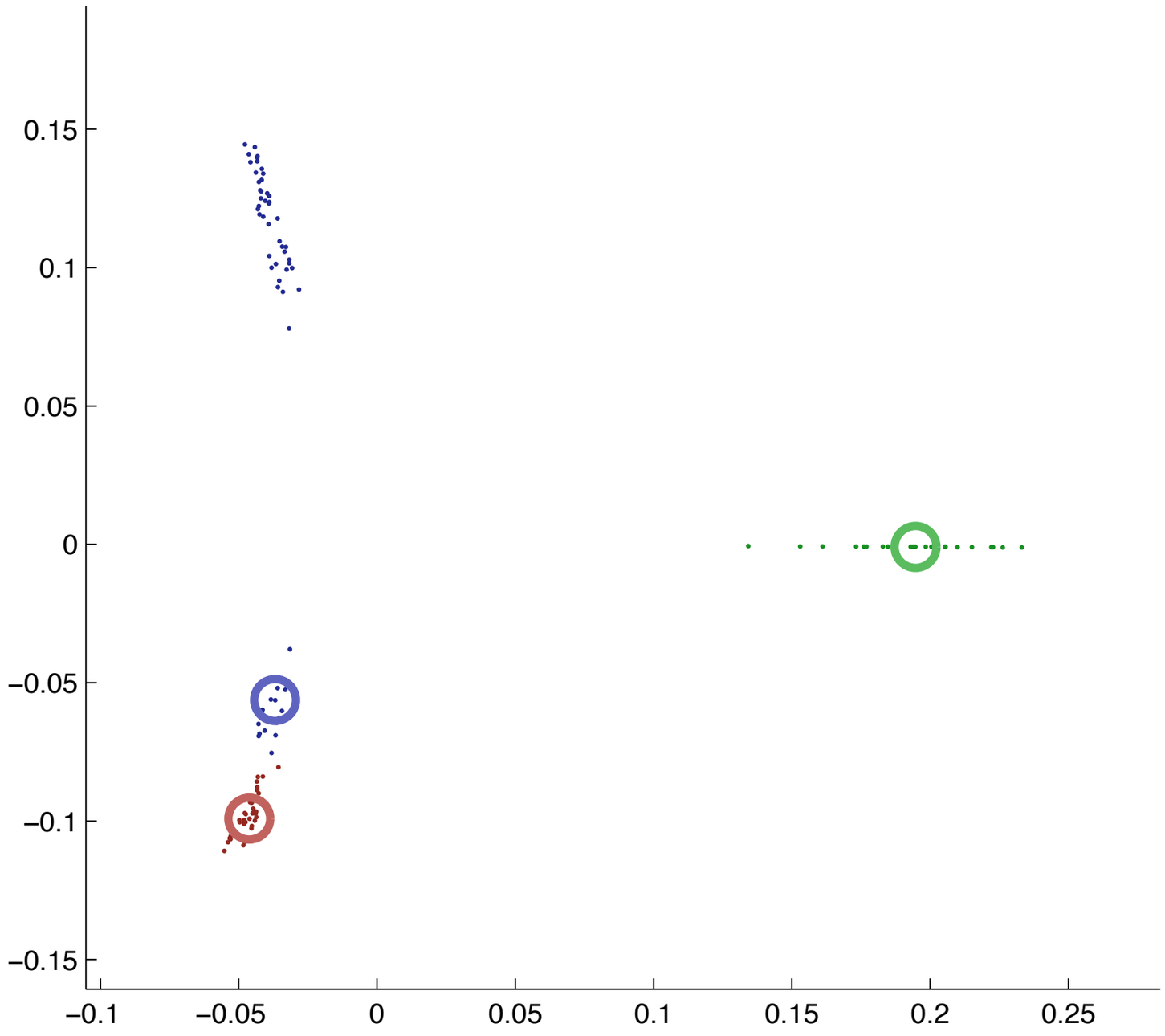
Example 2



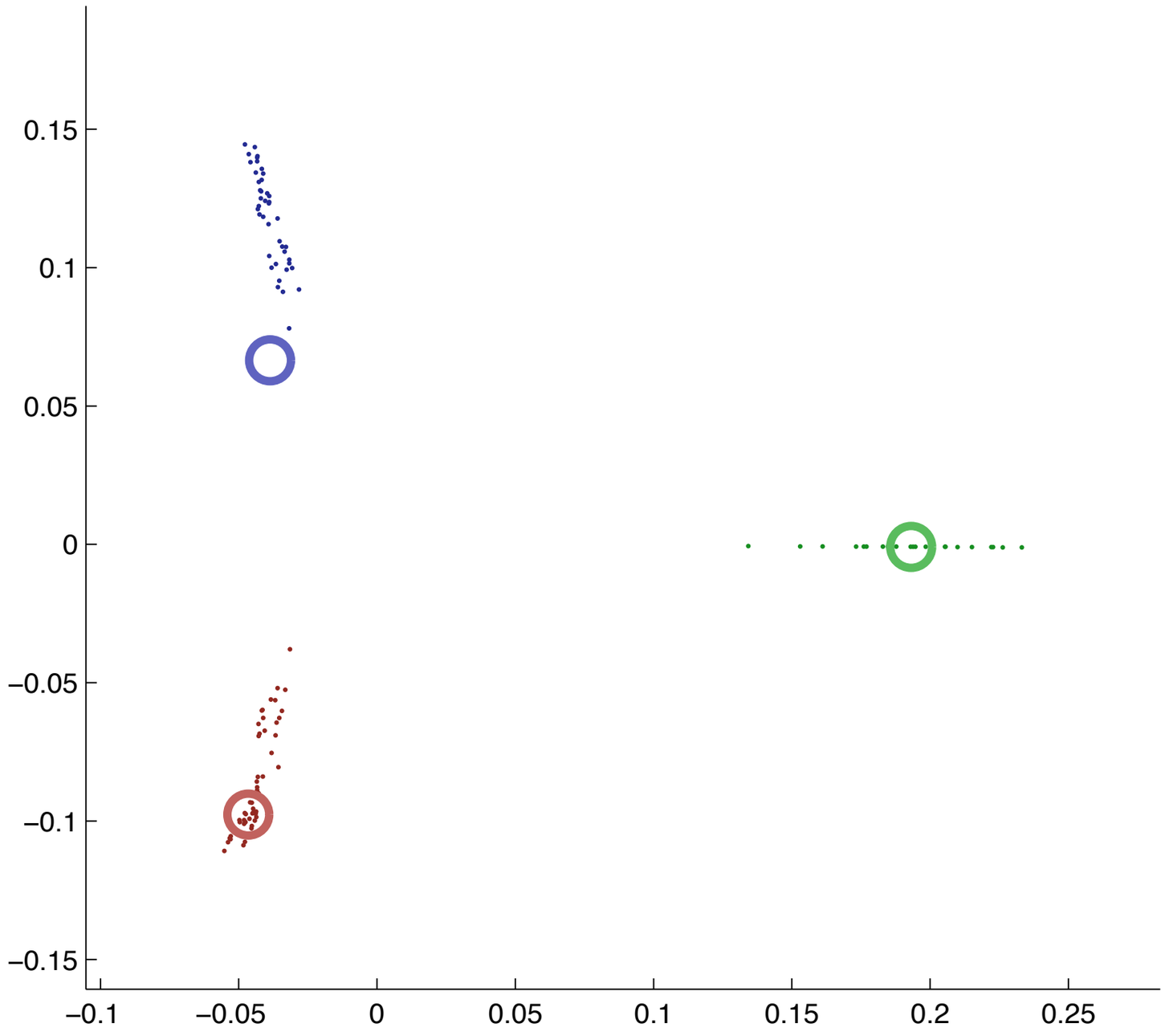
Example 2



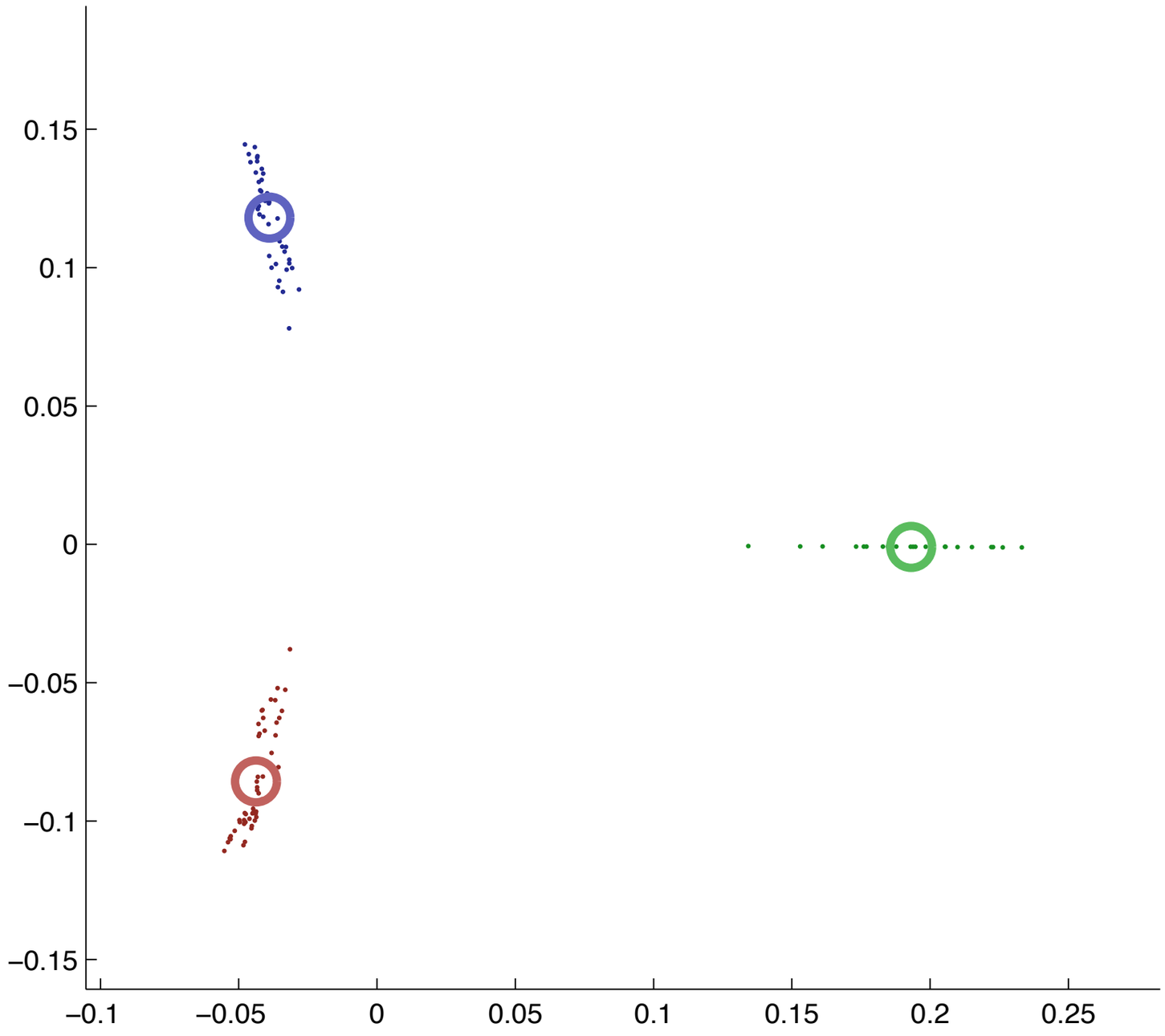
Example 3



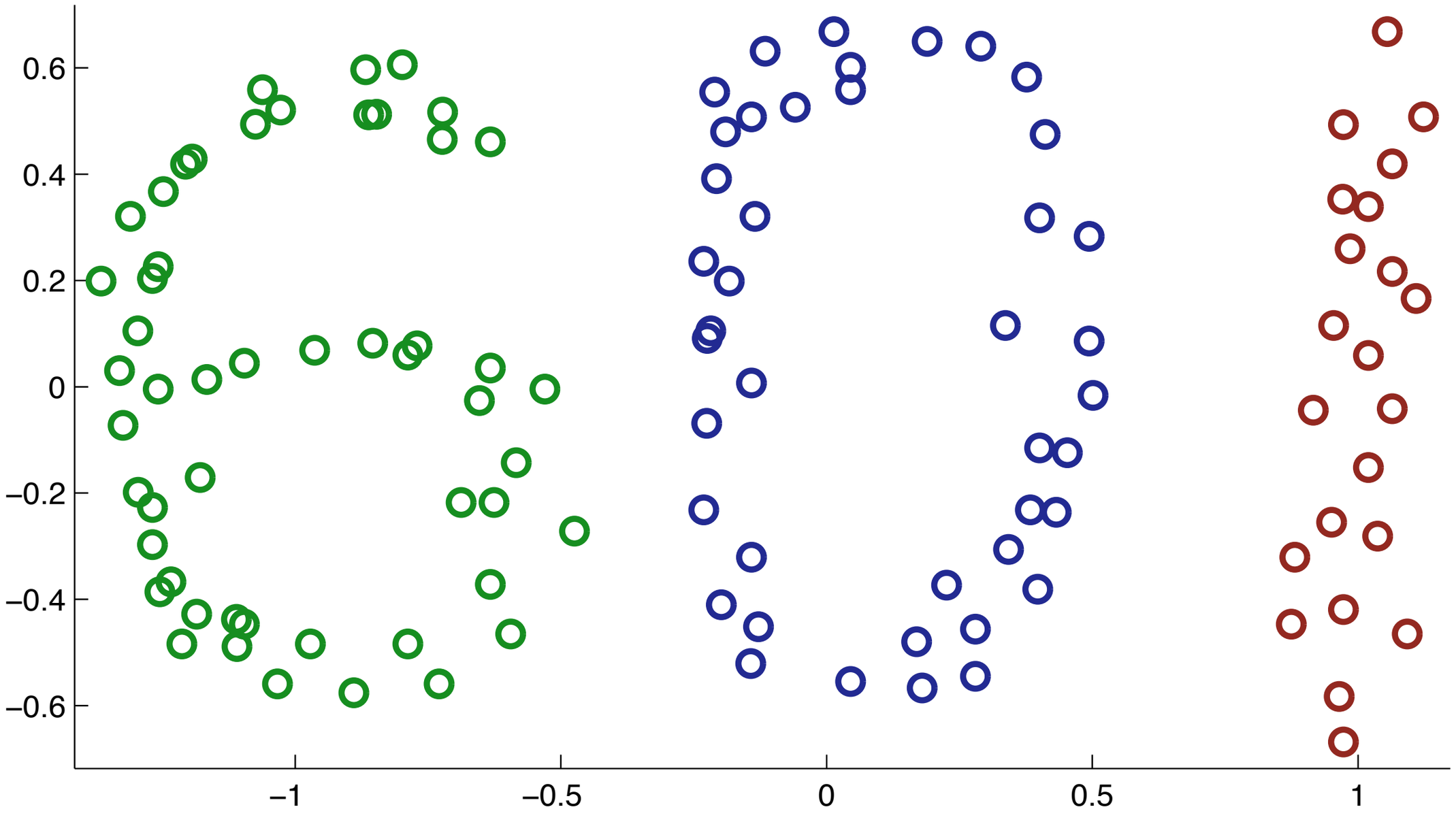
Example 3



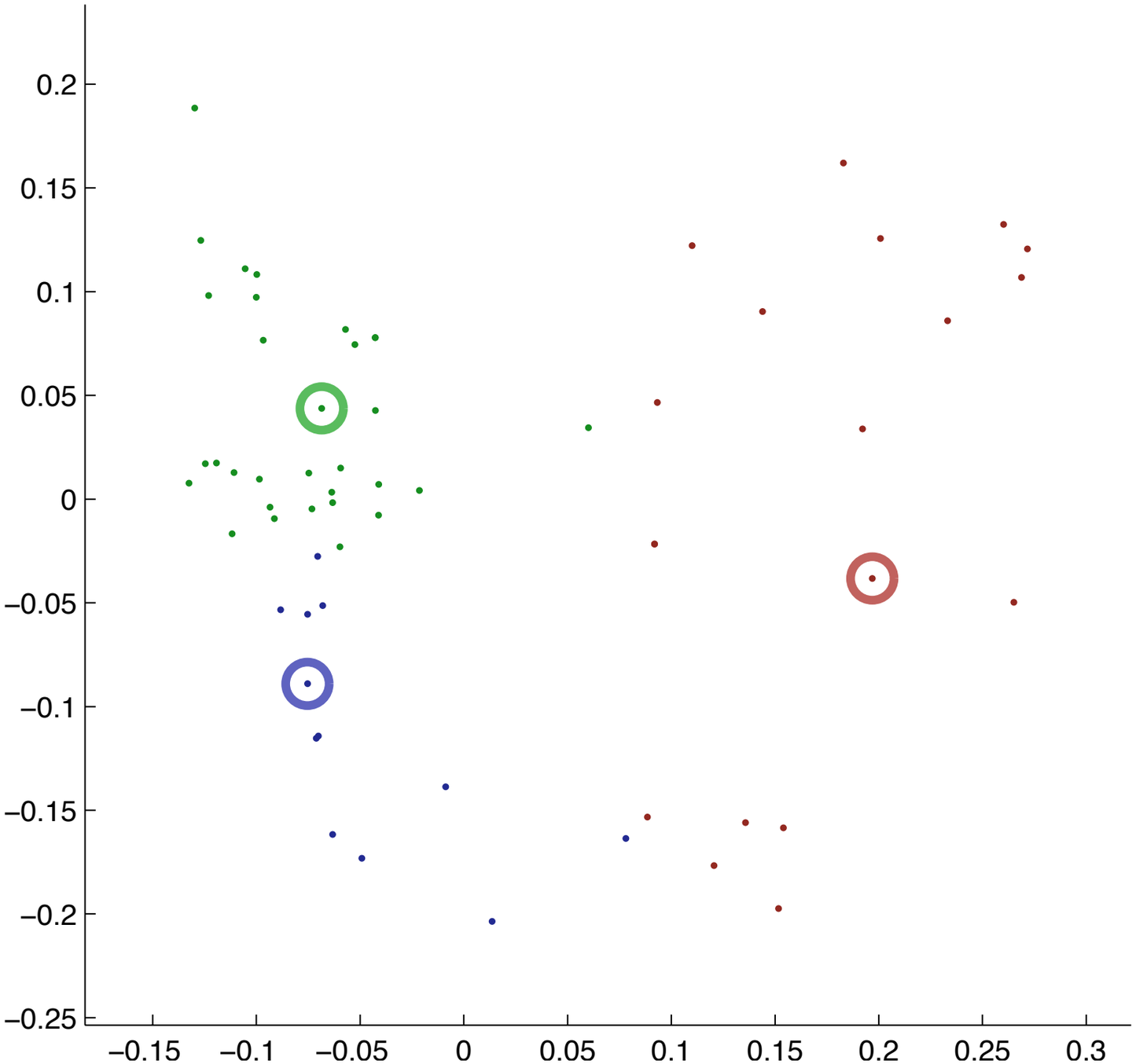
Example 3



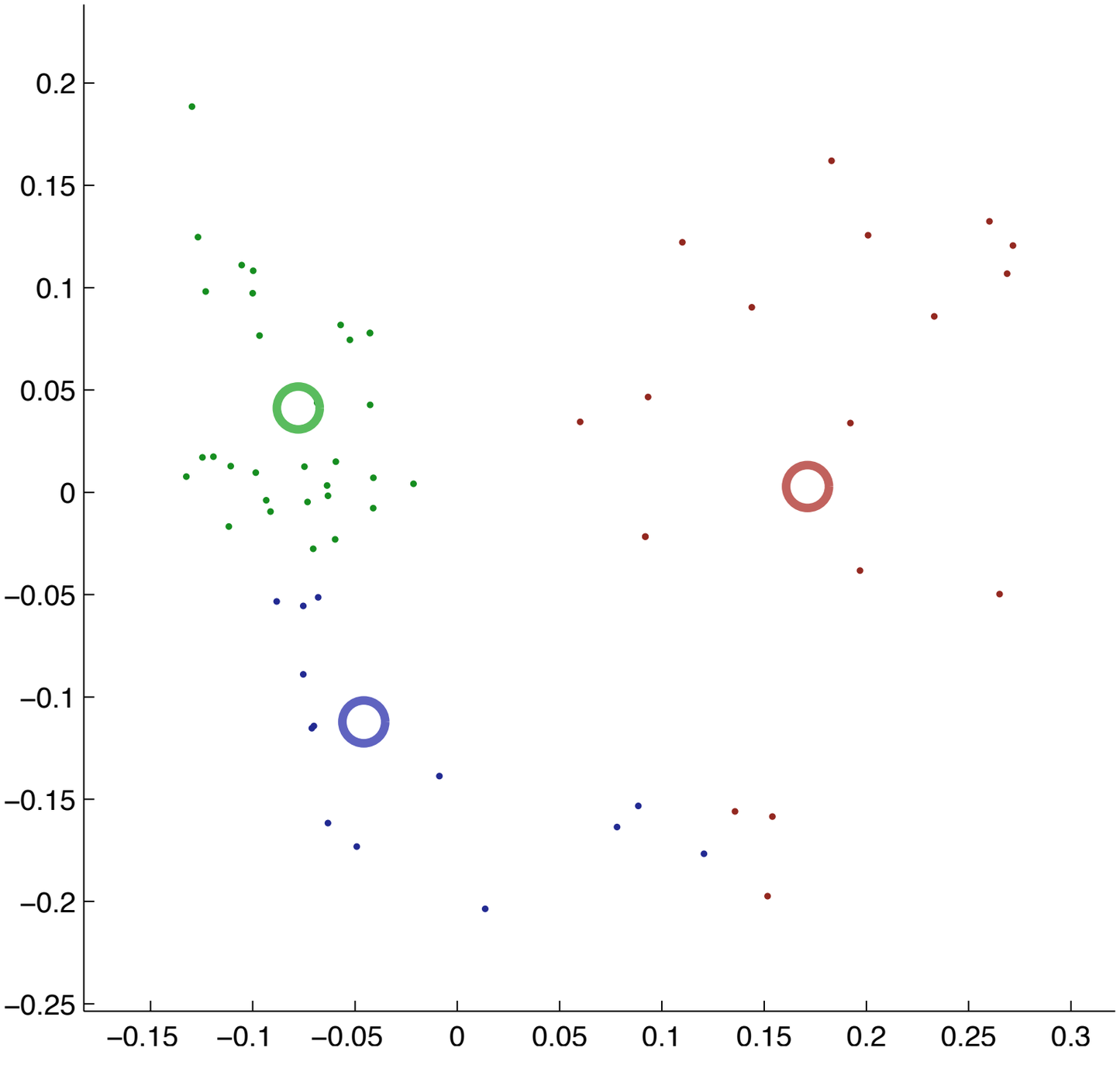
In original space



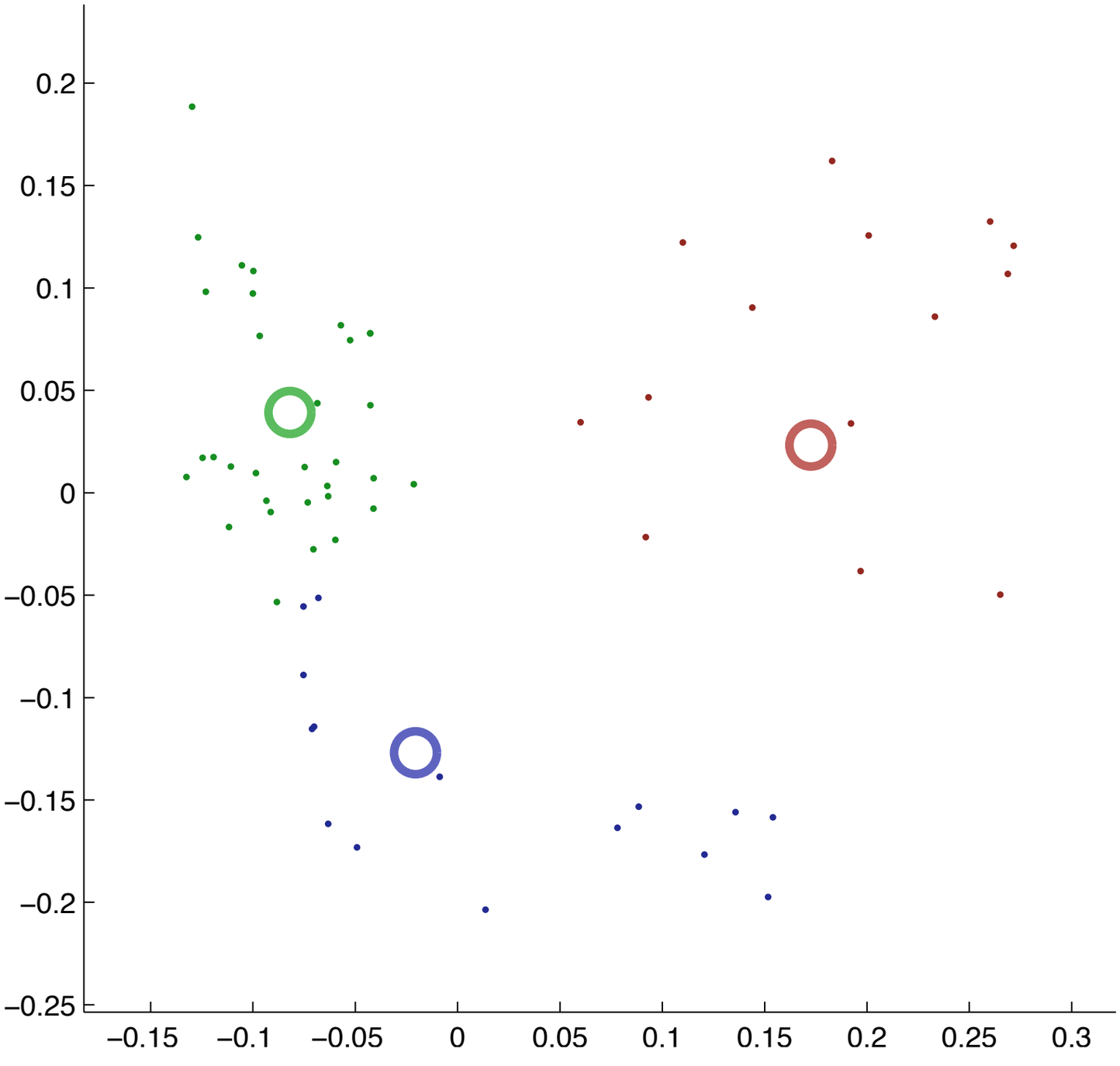
Example 4



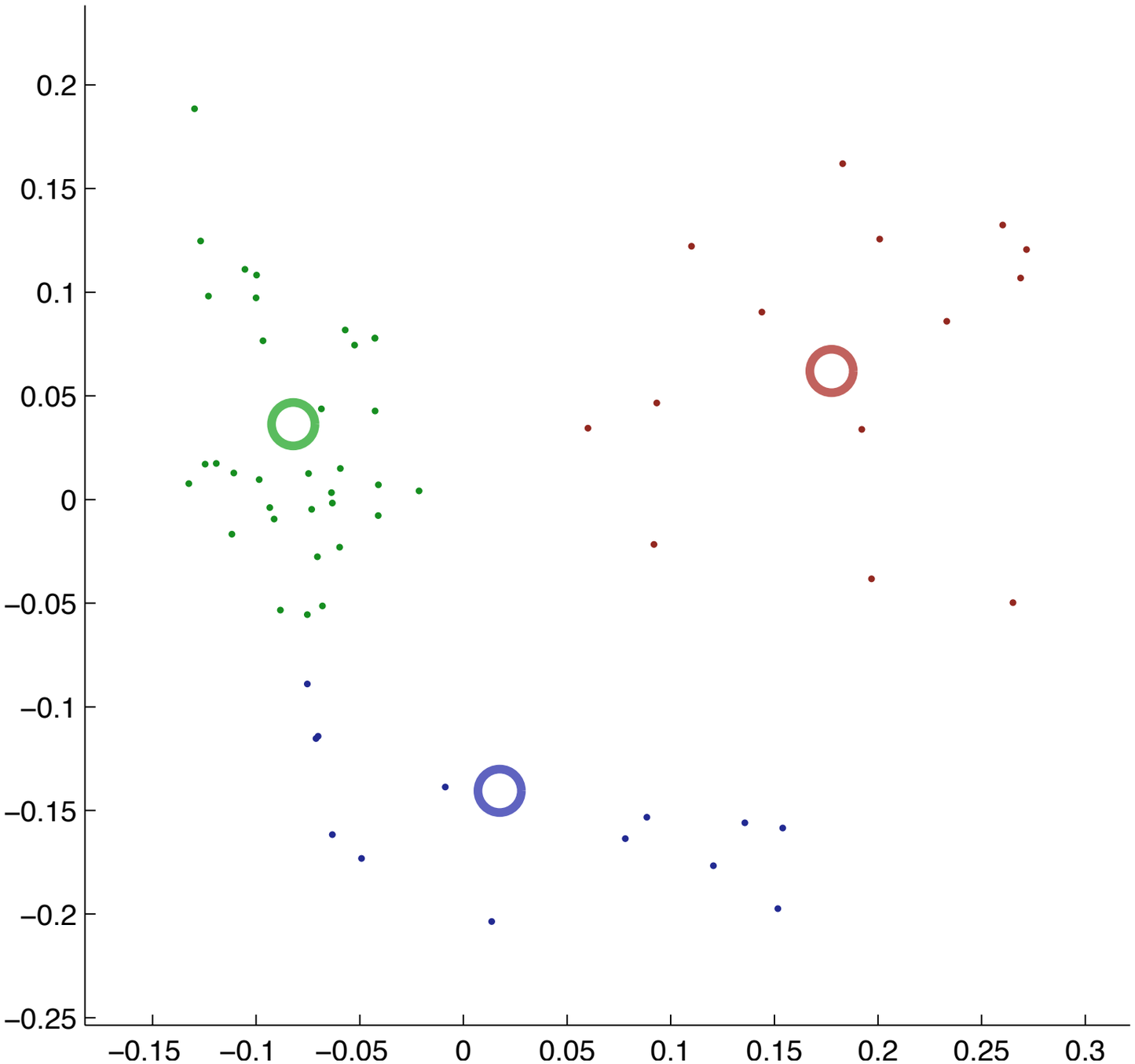
Example 4



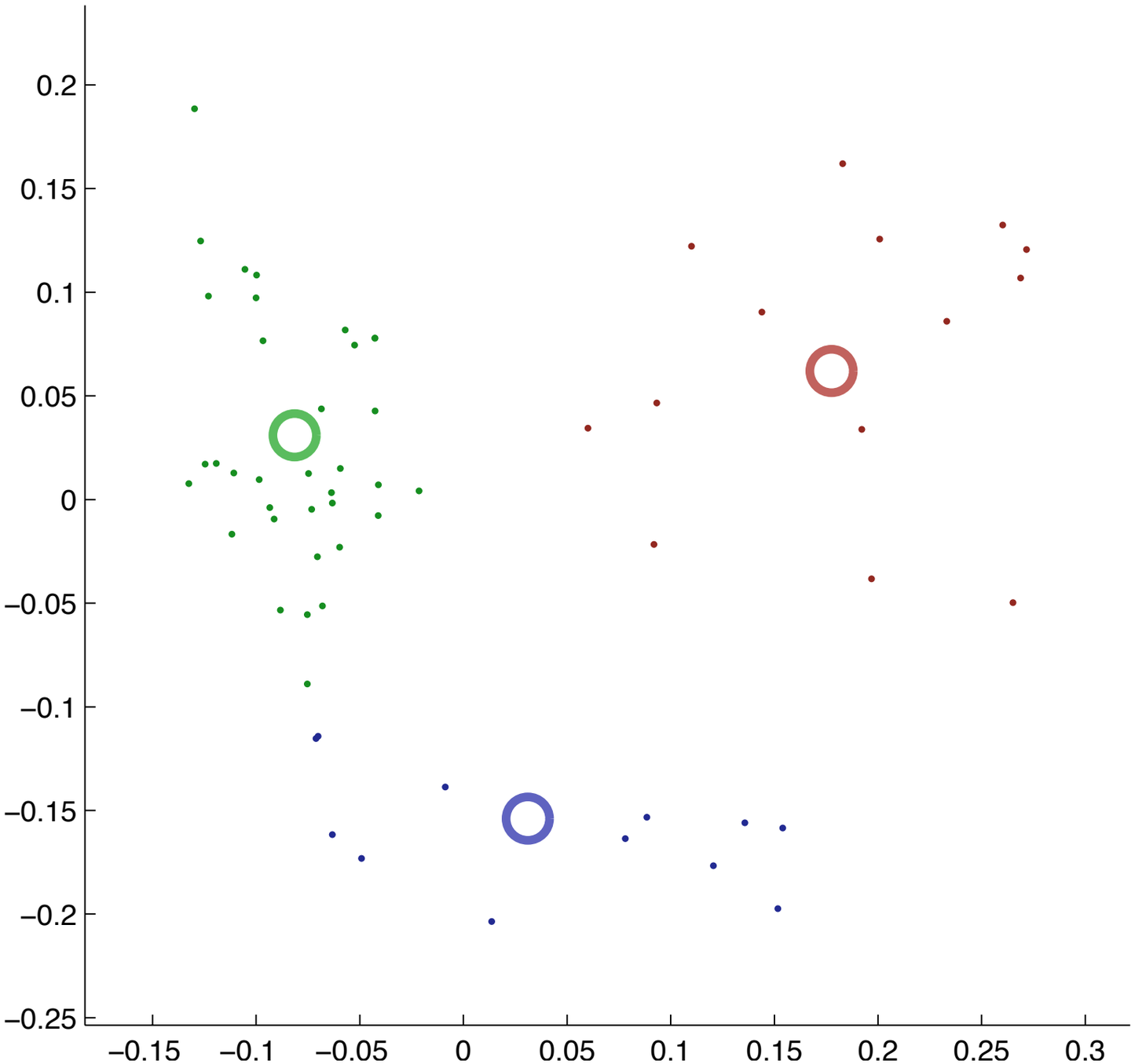
Example 4



Example 4



Example 4



Optimization strategies

- Initialization:
  - want to get a "diverse" set of starting centers
  - simplest: select K at random
  - smarter: select too many, then remove centers that are close to others
  - or: select one at a time by maximizing minimum distance to previous centers
- Optimization:
  - want to avoid local optima
  - split / merge
  - multiple restarts

Beyond k-means

- Suppose:
  - $P(X_i \mid Z_{ij} = 1, \theta) = \text{Gaussian}(\mu_j, \sum_j)$
  - $P(Z_{ij} = 1 \mid \theta) = p_j$
- Mixture of Gaussians
- Leads to an algorithm called "soft" k-means or "fuzzy" k-means
  - vs. "hard" k-means

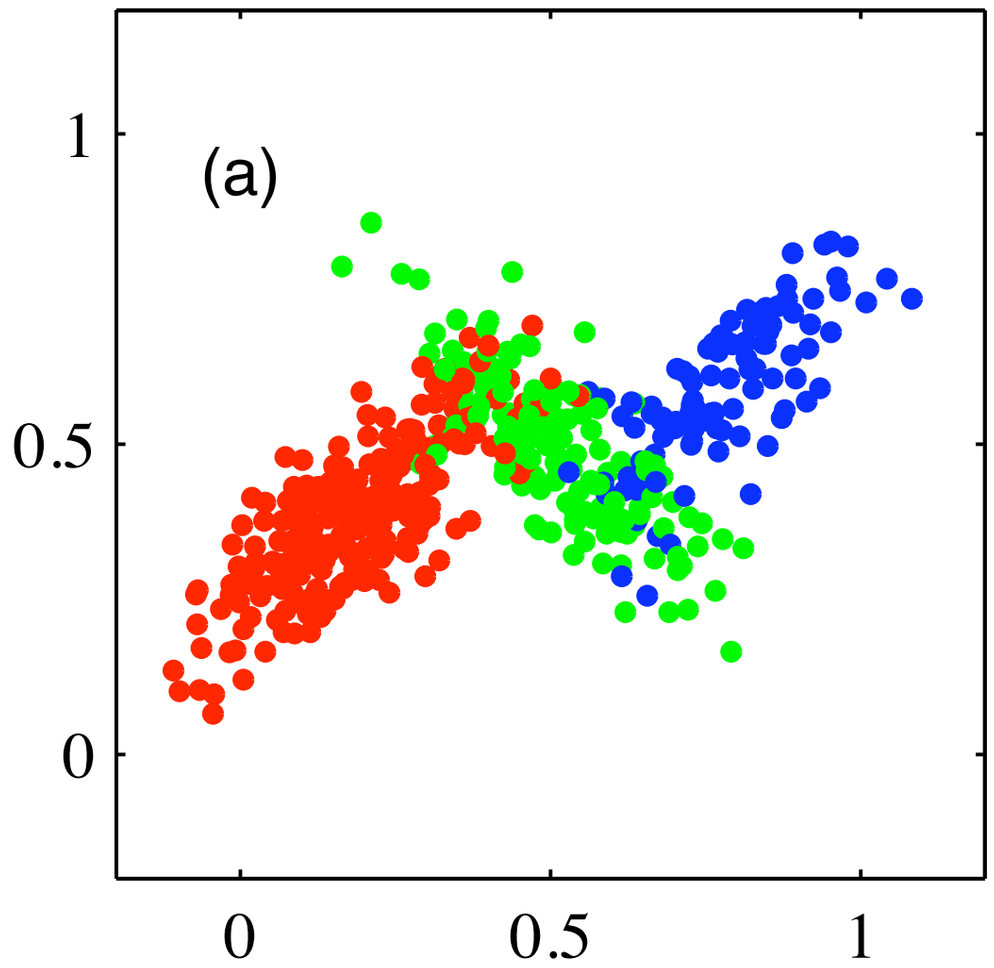


fig 9.5a from Bisho

“Soft” k-means

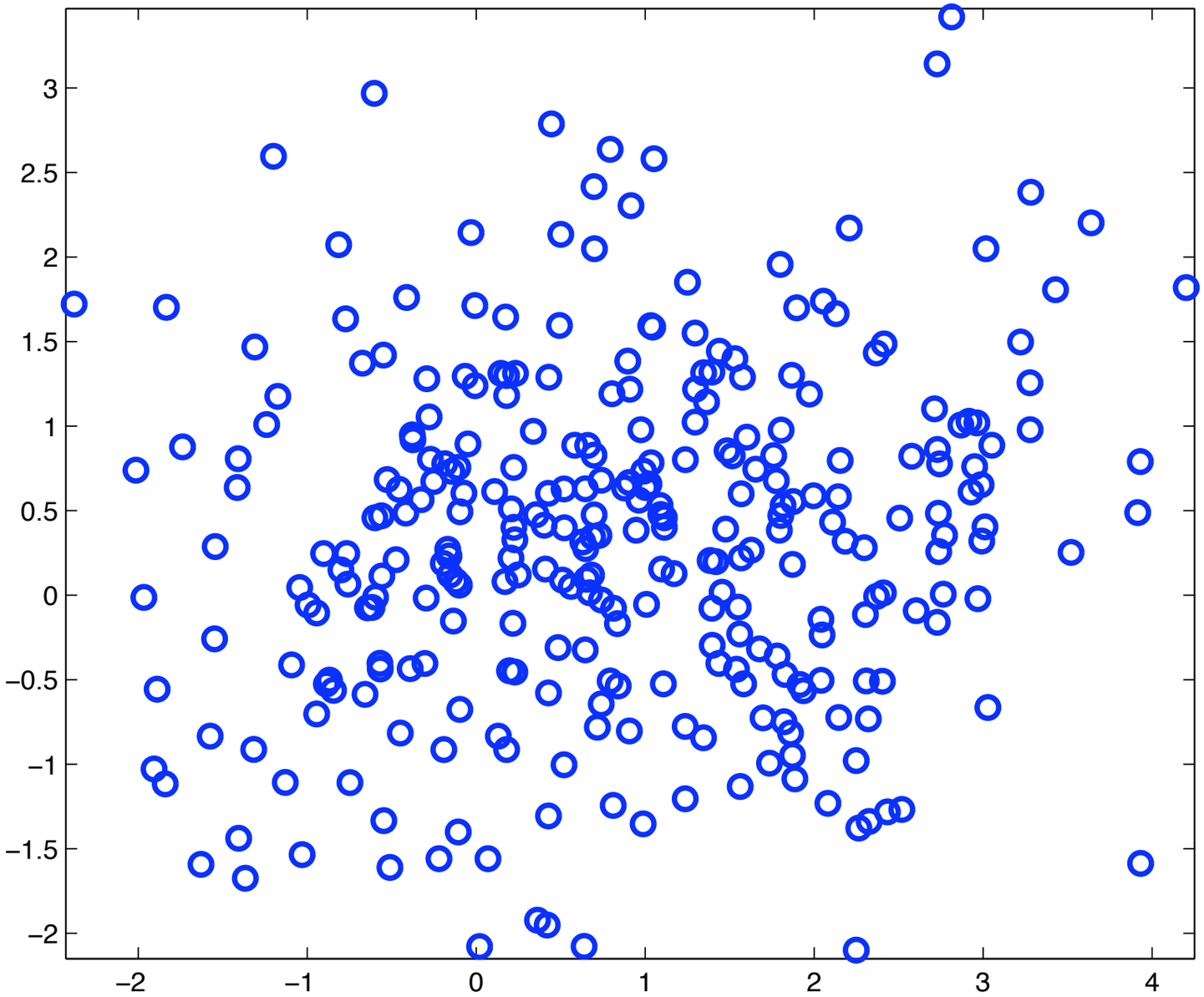
- Find soft assignments:

- Update means:

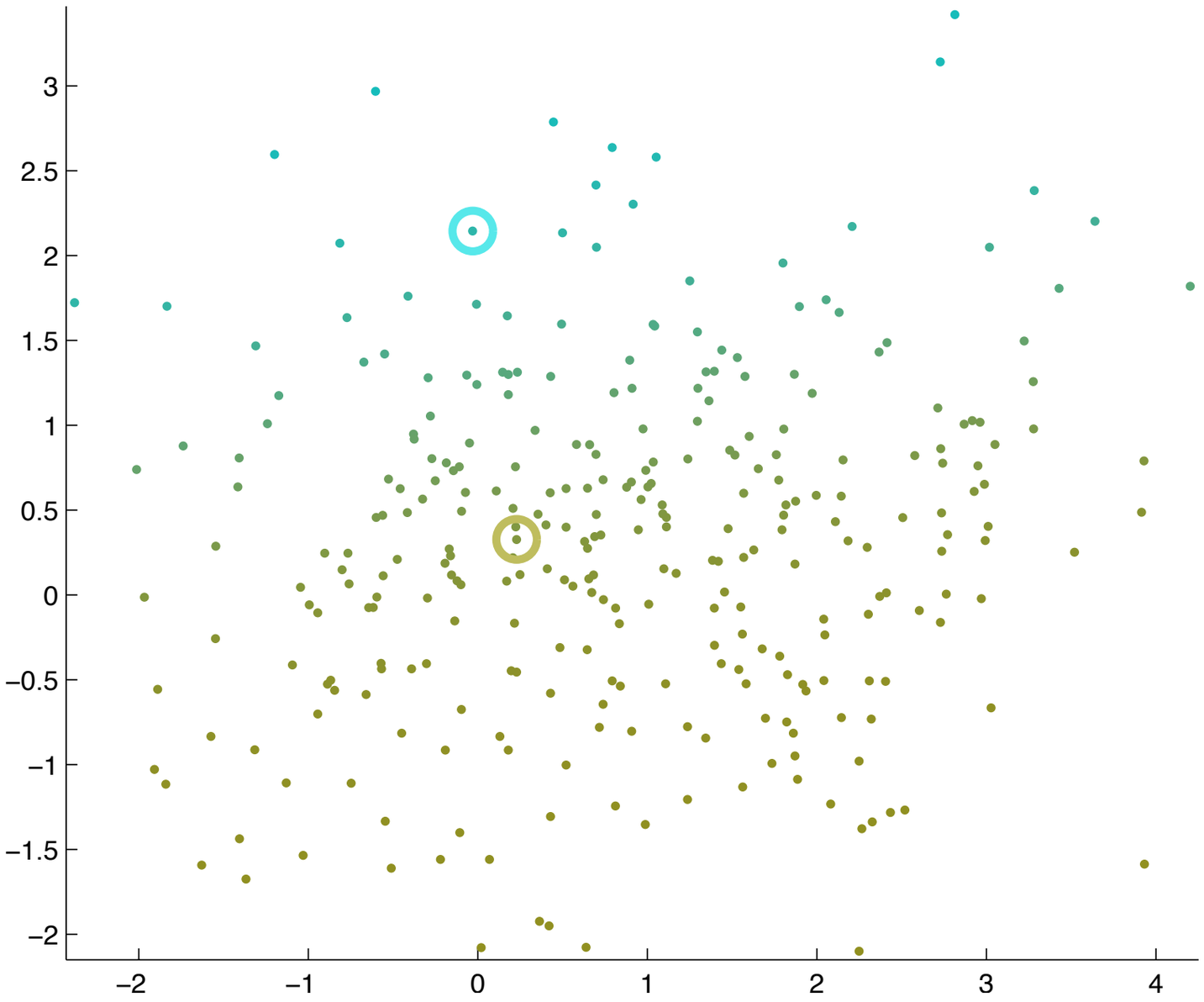
- Possibly: update covariances

- Repeat

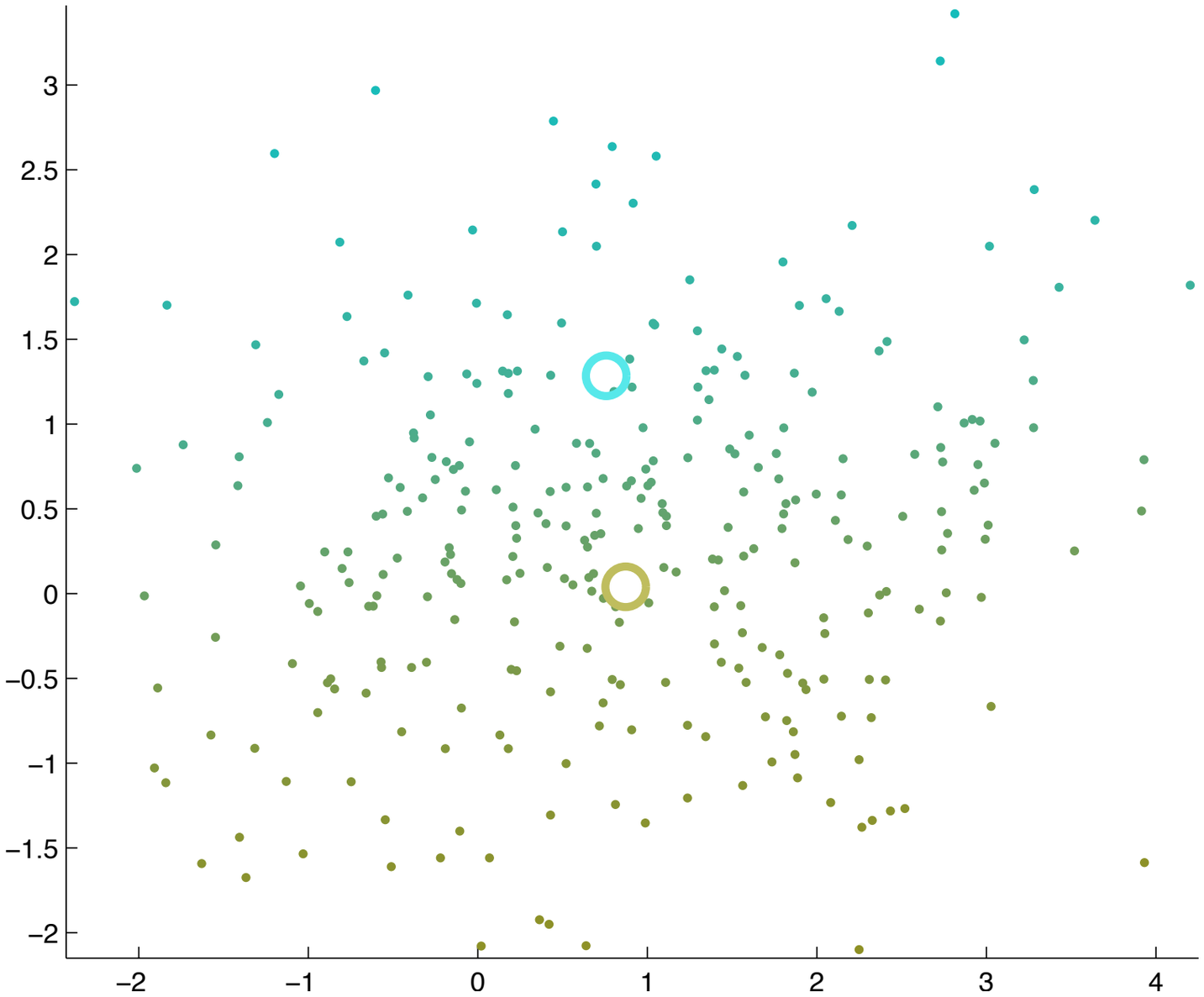
Example



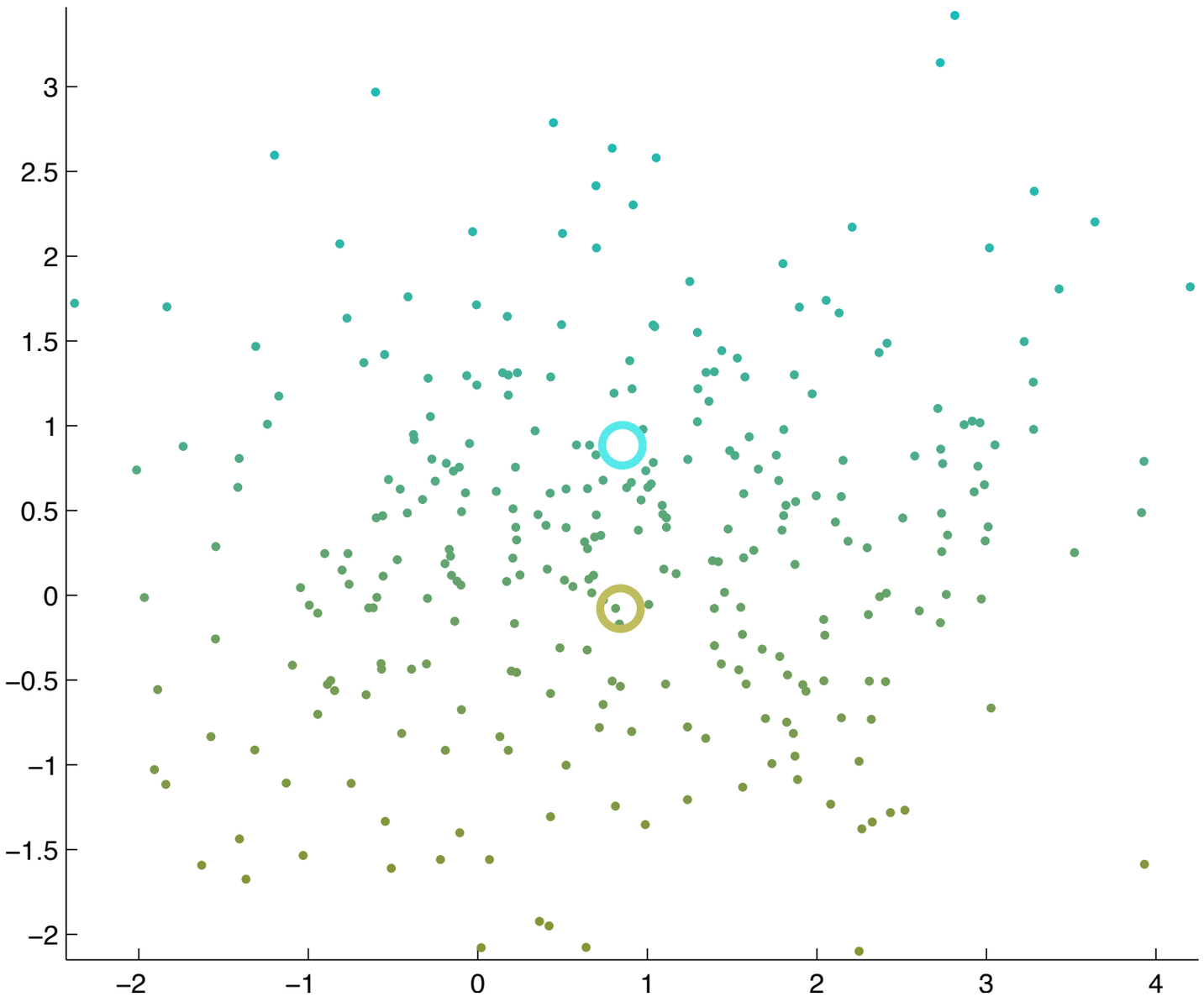
Example



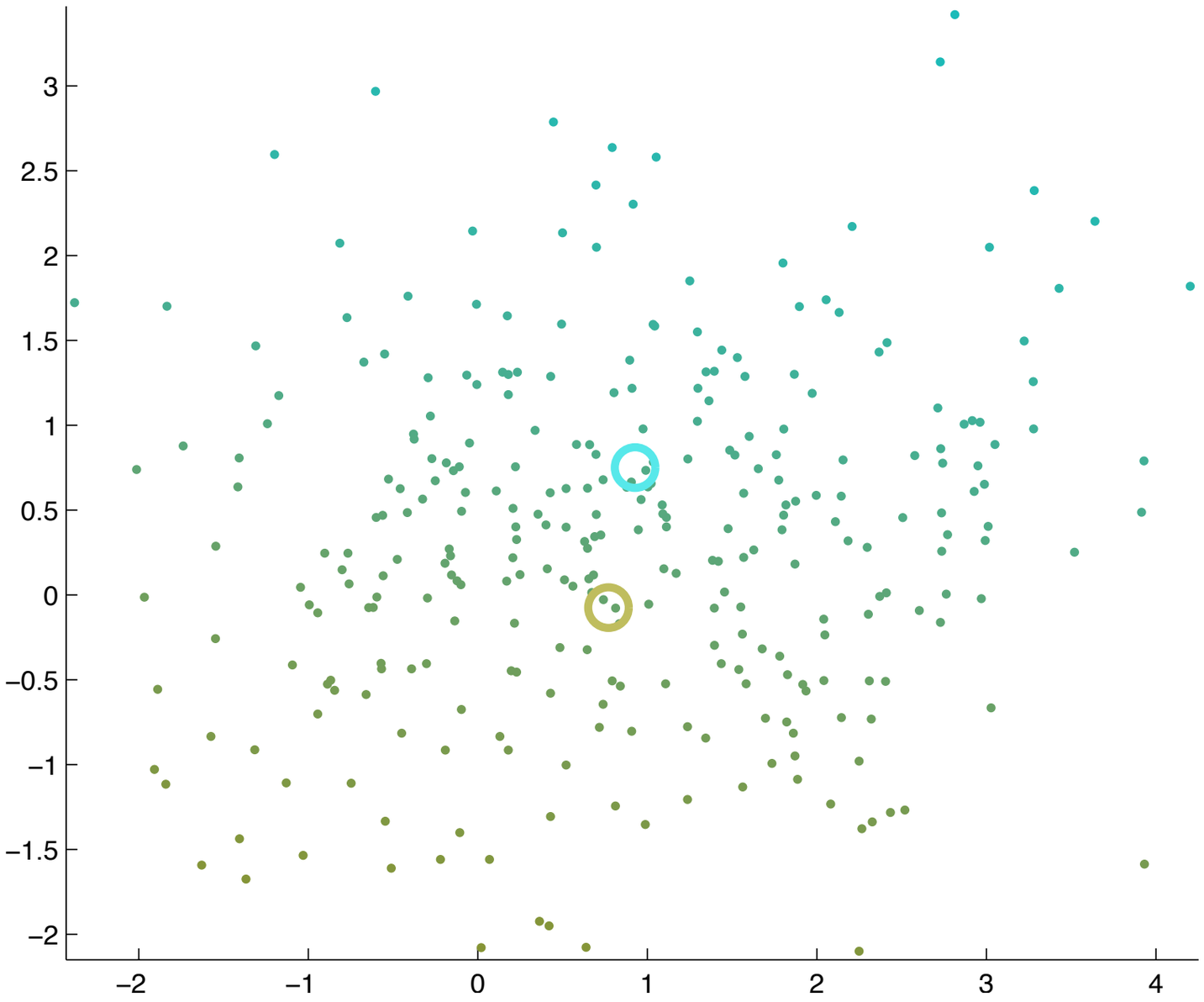
Example



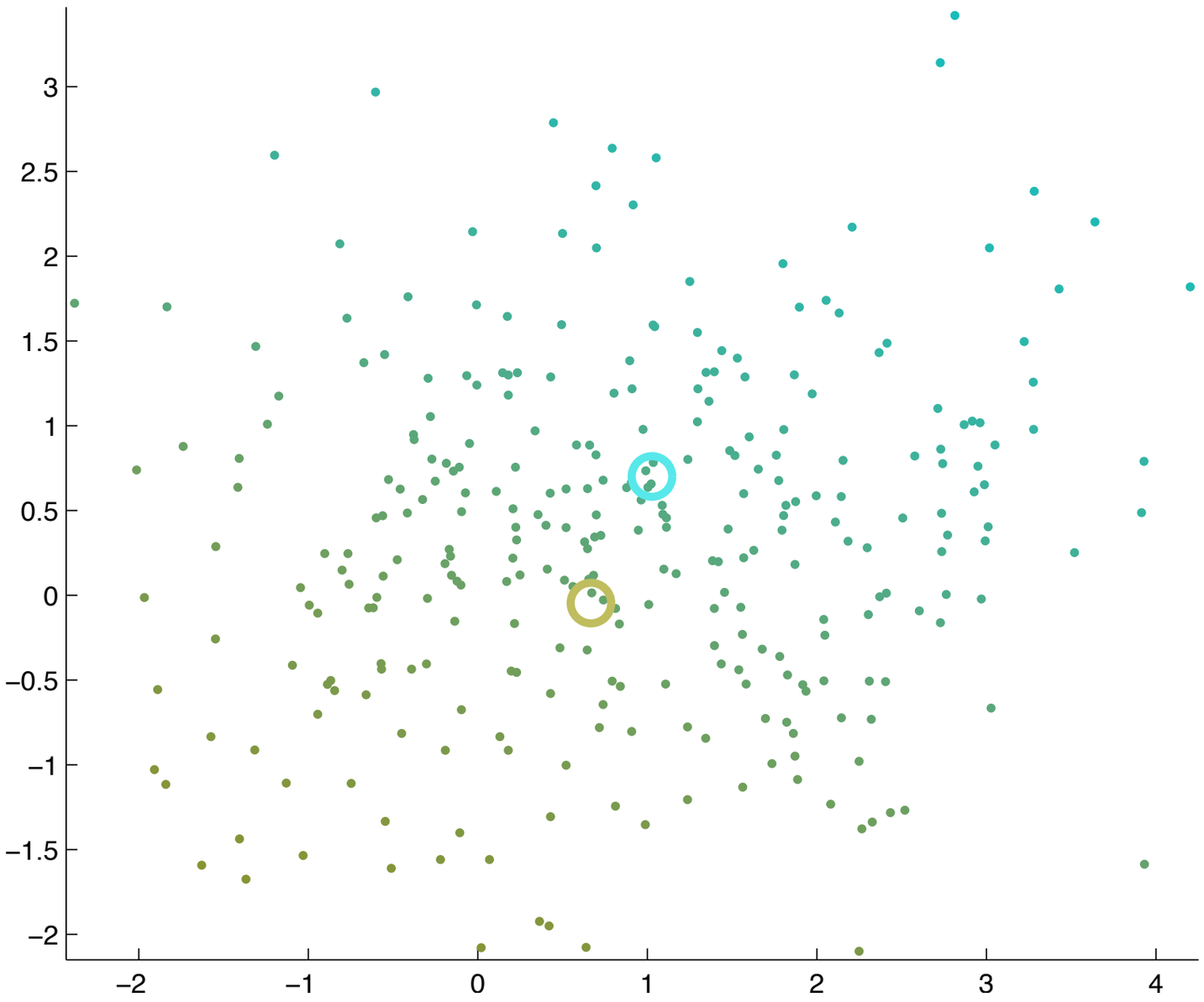
Example



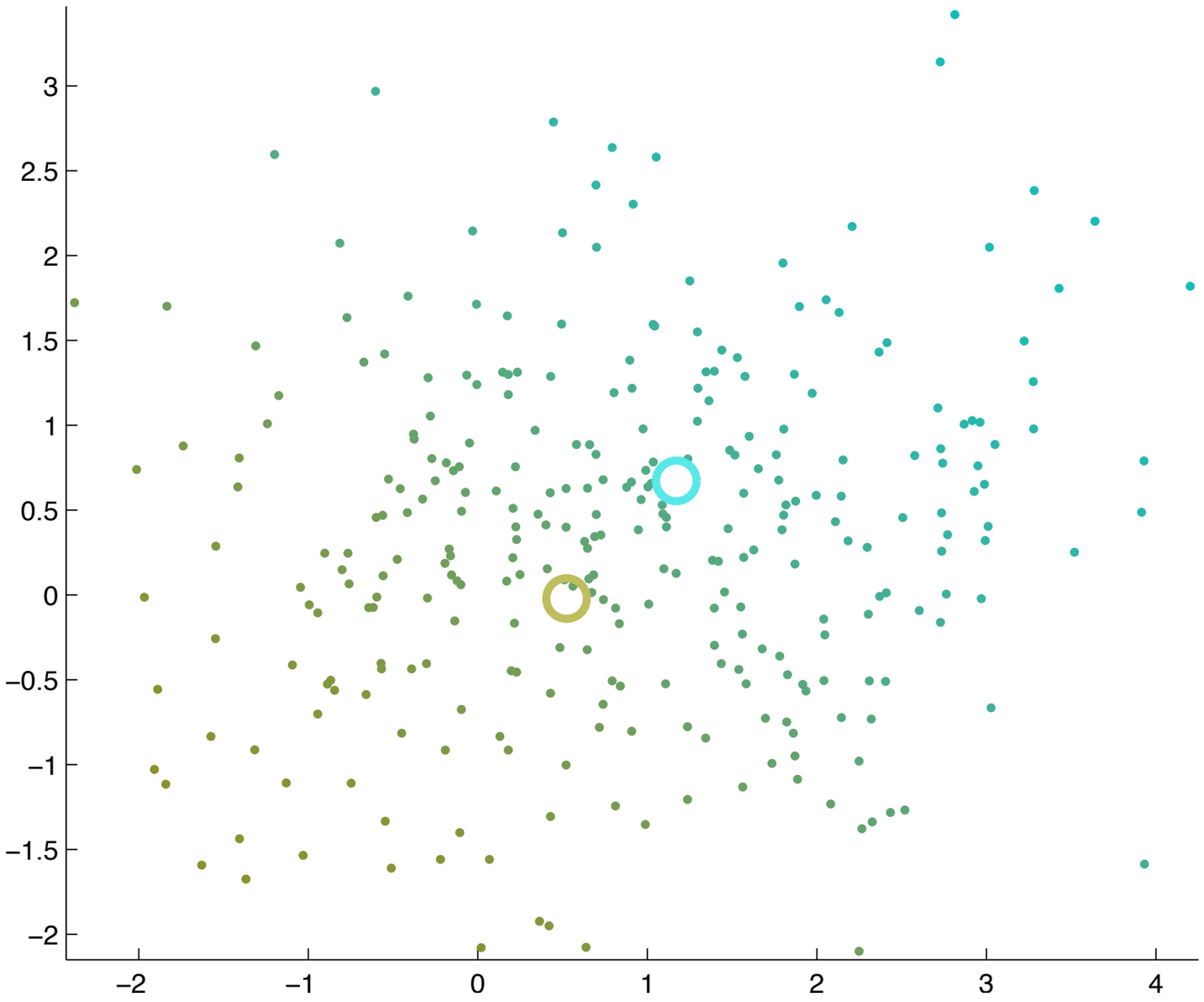
Example



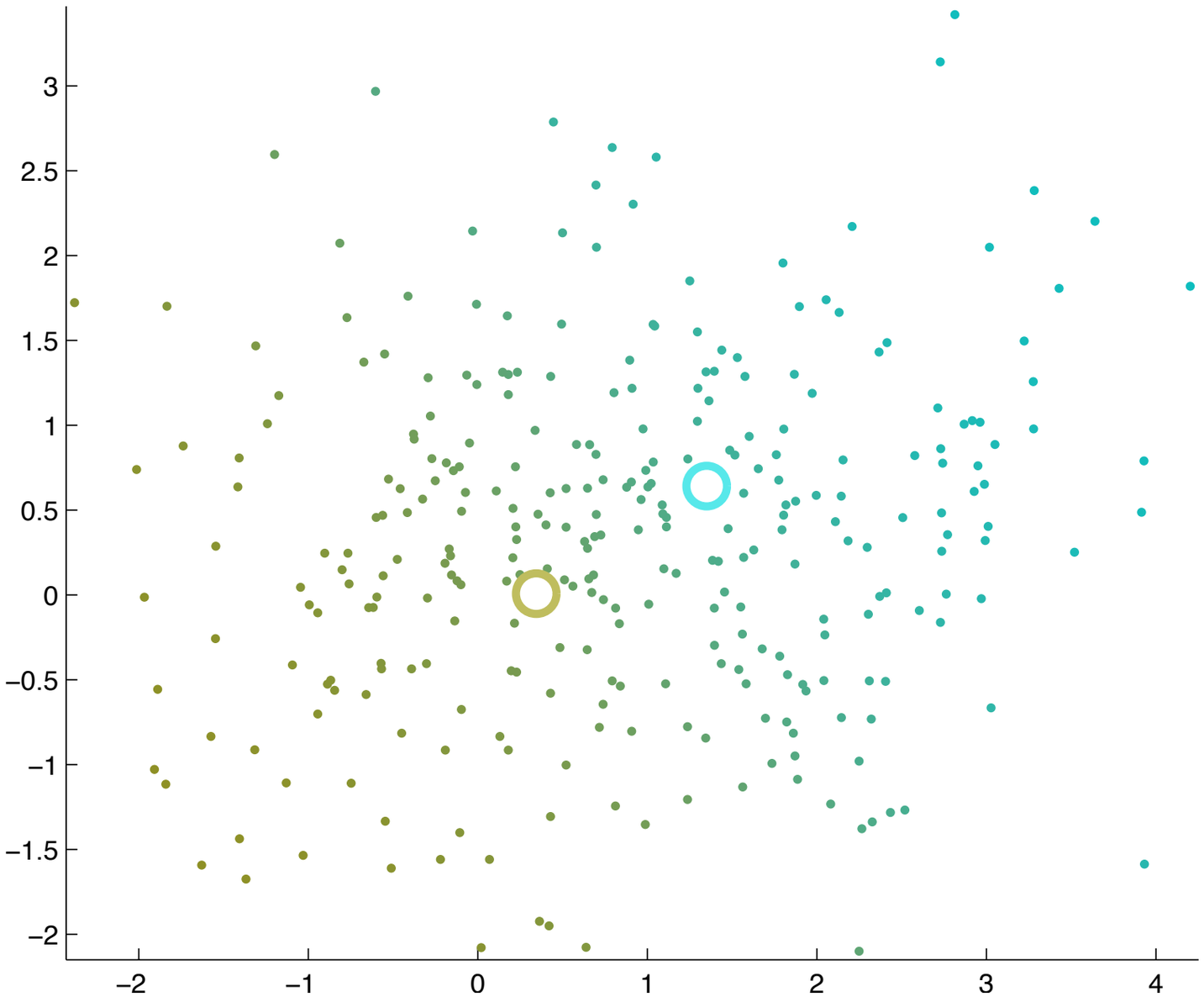
Example



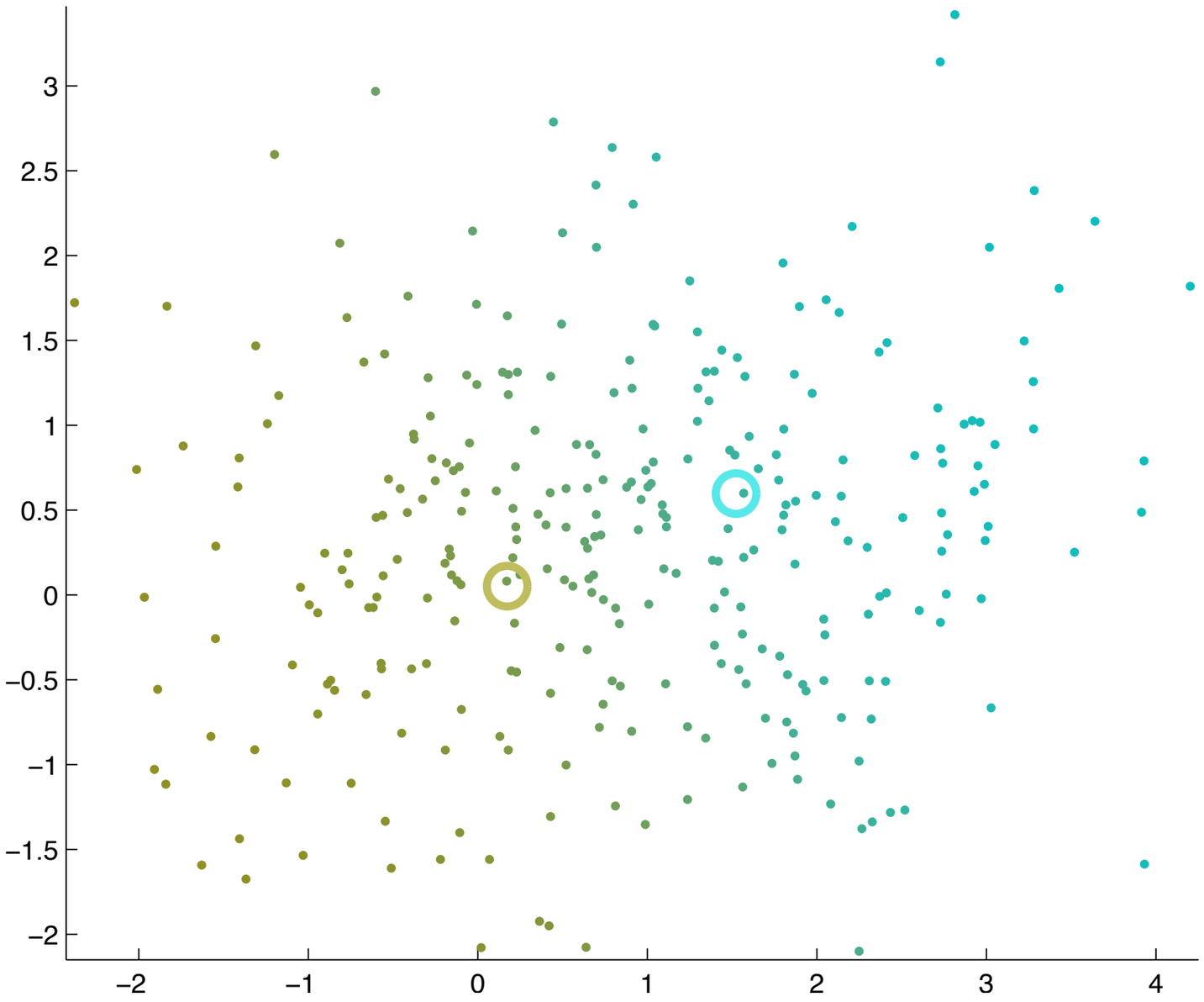
Example



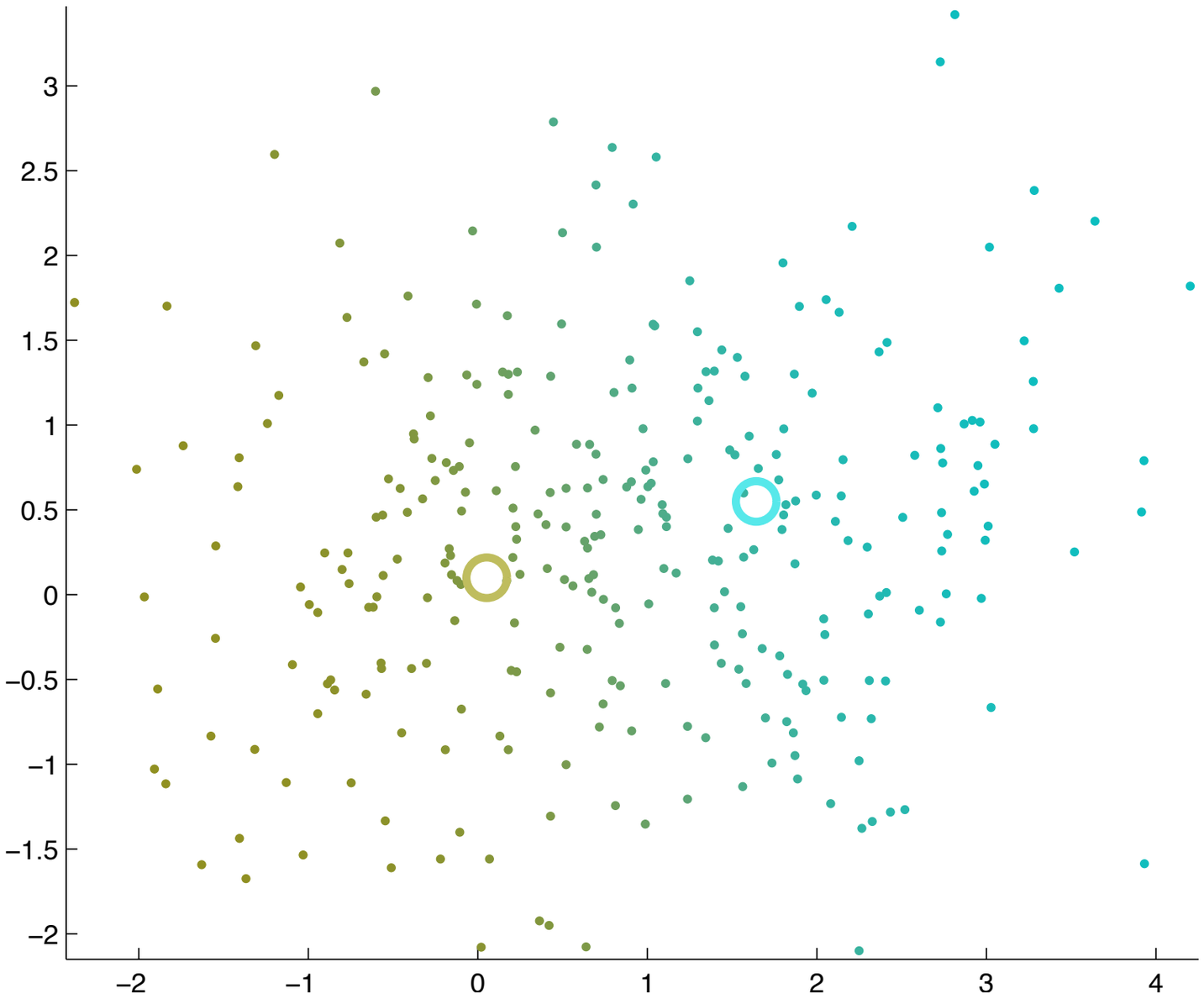
Example



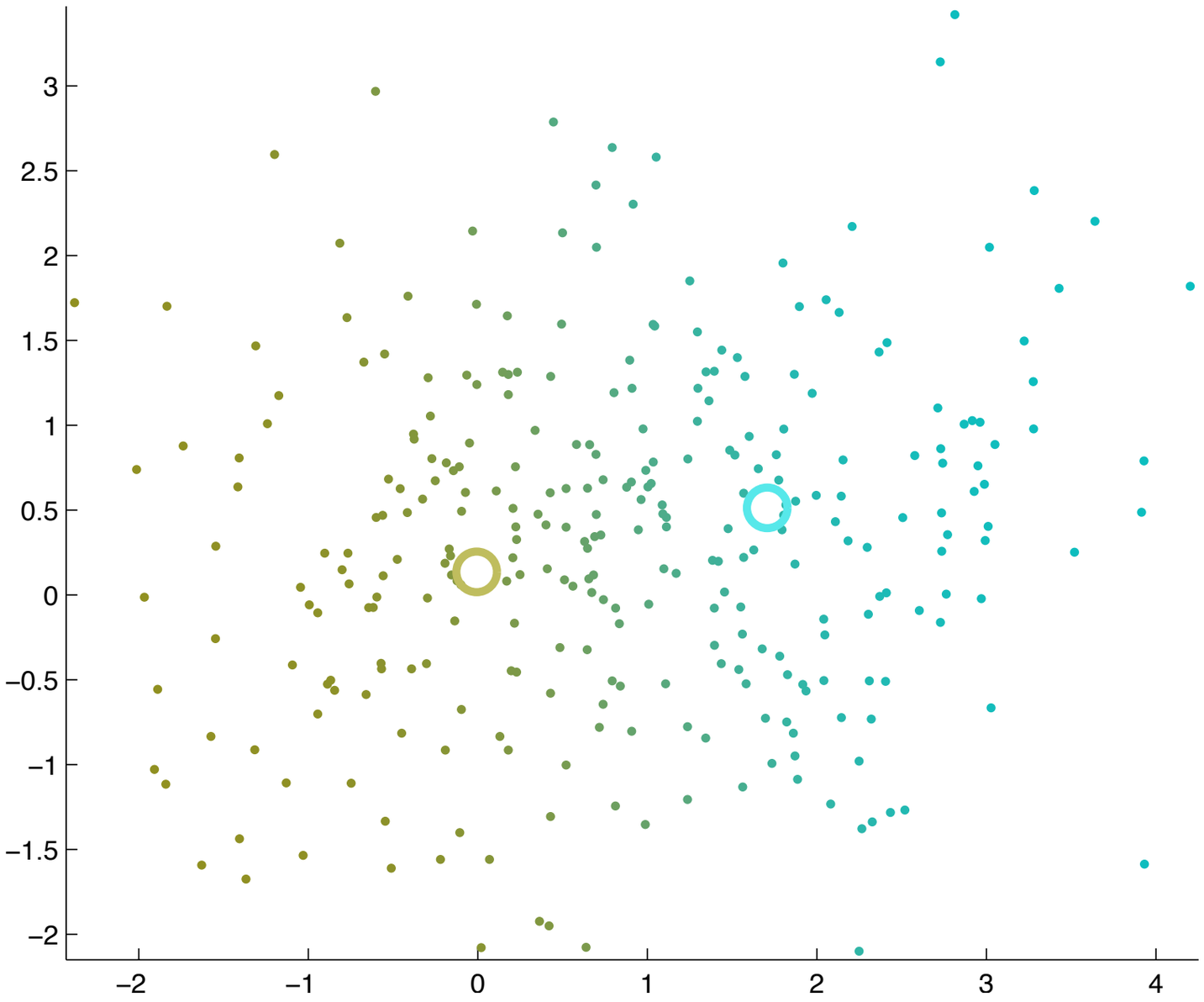
Example



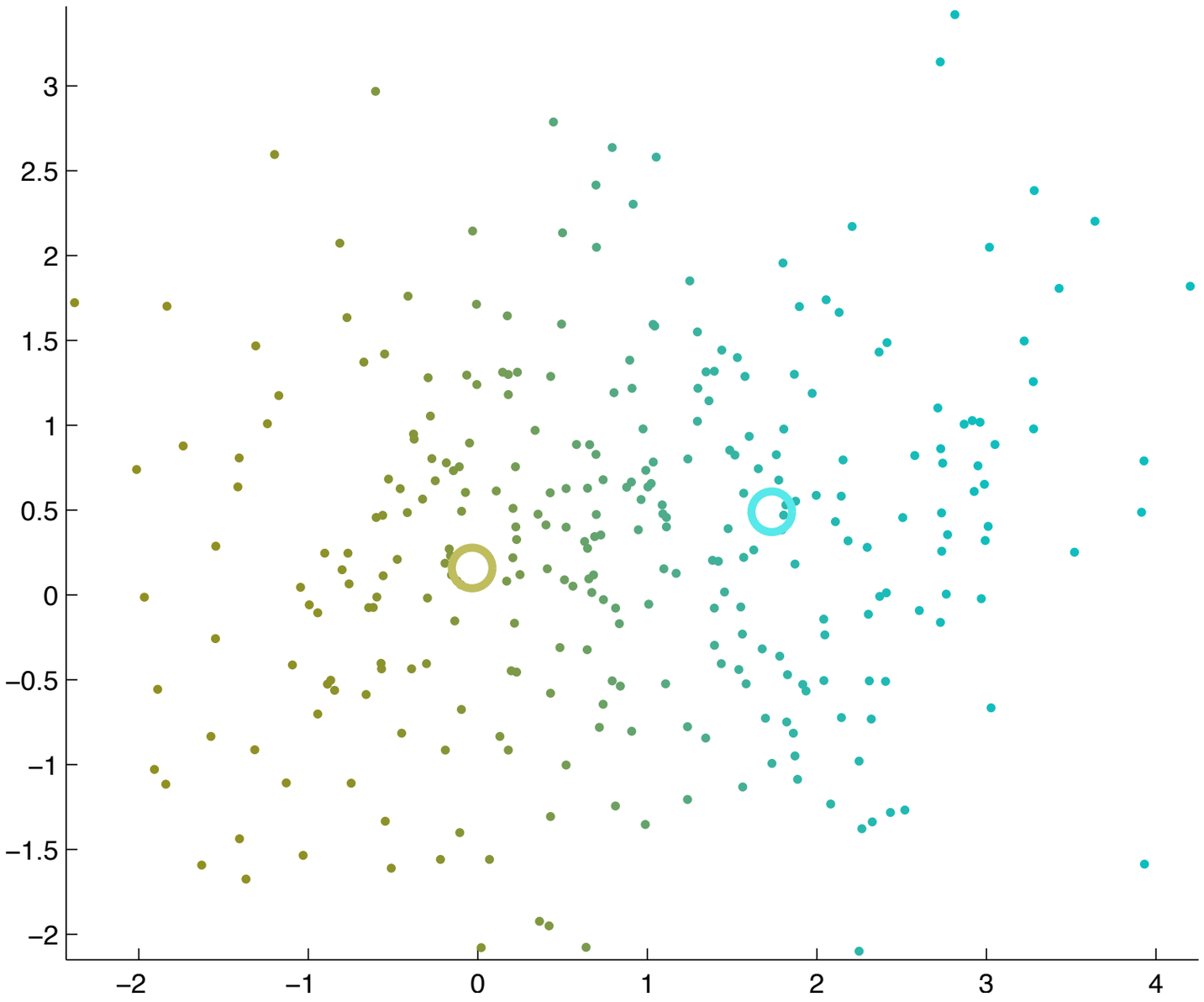
Example



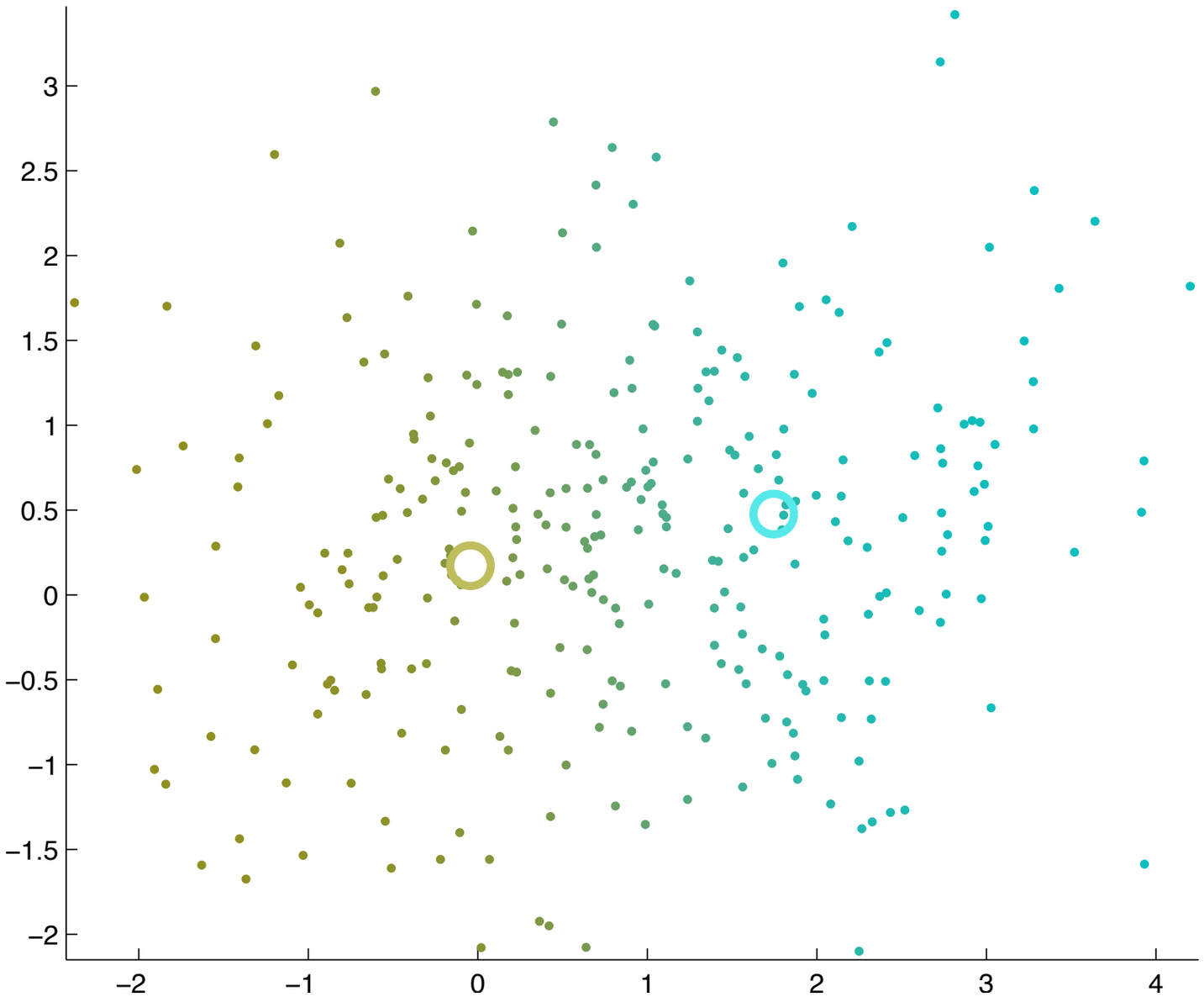
Example



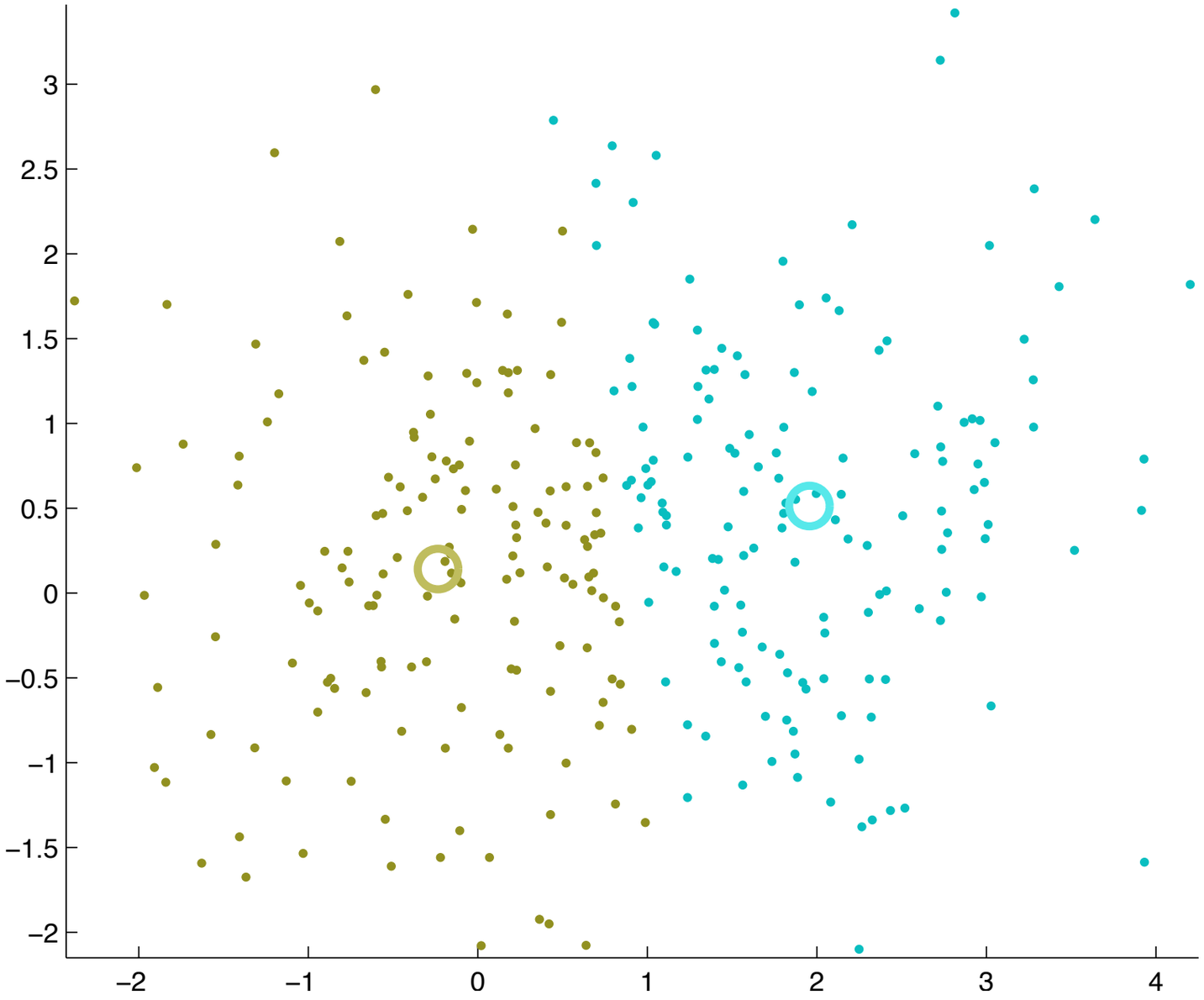
Example



Example



Comparison: "hard" k-means



Comparison

- Soft k-means:
  - converges slower than hard k-means
  - more expensive per iteration
  - but asymptotically unbiased (unlike hard k-means), ***if*** we restart often enough
- Often, can initialize soft w/ hard

Hard k-means is a special case

- Let:
- Then:
- E-step:
- M-step:

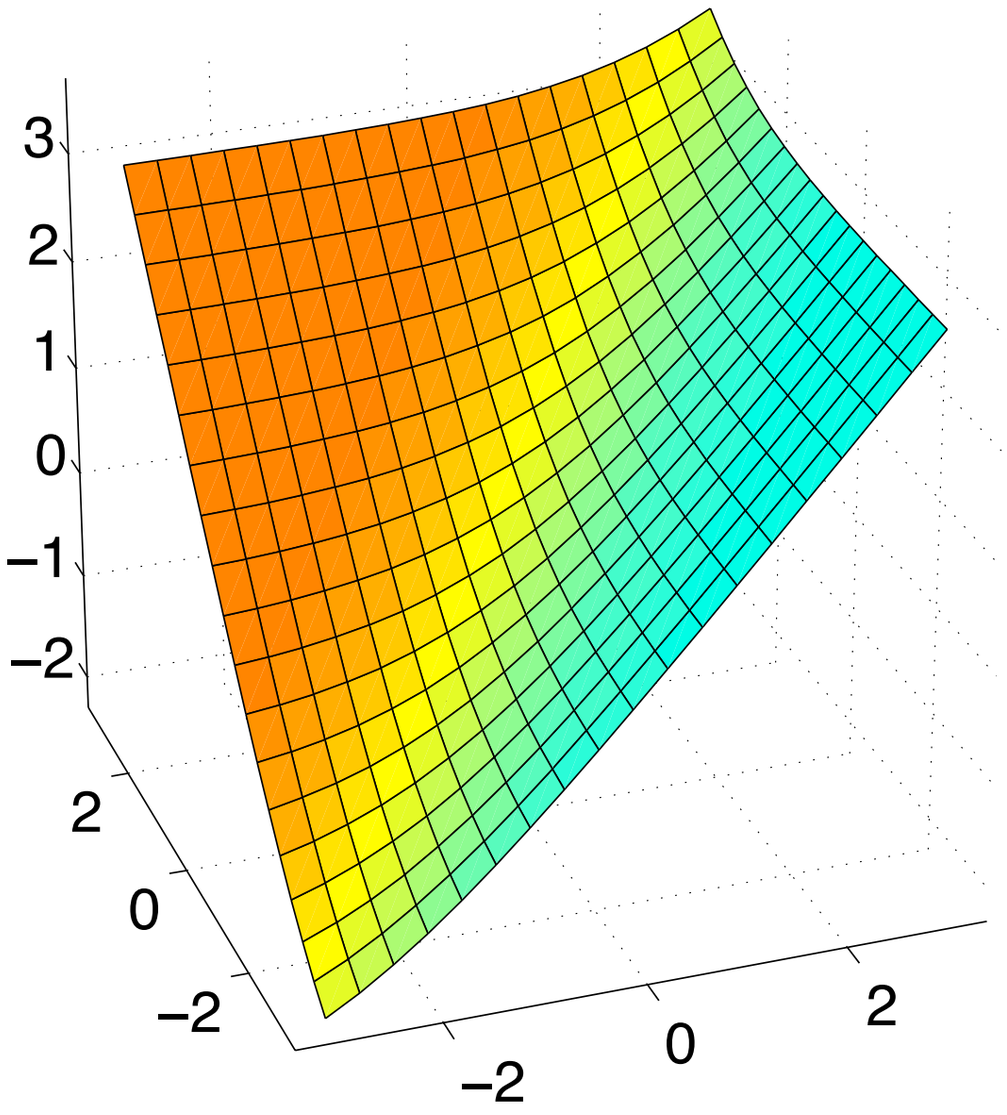
Deriving soft k-means

- Soft k-means is an example of an ***EM algorithm***
  - general strategy for MLE or MAP with ***hidden variables*** (in our case, Z_{ij})
- We'll derive soft k-means first, then generalize

Deriving soft k-means

- $P(X_i \mid Z_{ij} = 1, \theta) = \text{Gaussian}(\mu_j, \sum_j)$
- $P(Z_{ij} = 1 \mid \theta) = p_j$
- $P(X_i, Z_{i\cdot} \mid \theta) =$
- $L = \ln P(X \mid \theta) =$

$f(x,y) = \ln(e^x+e^y)$

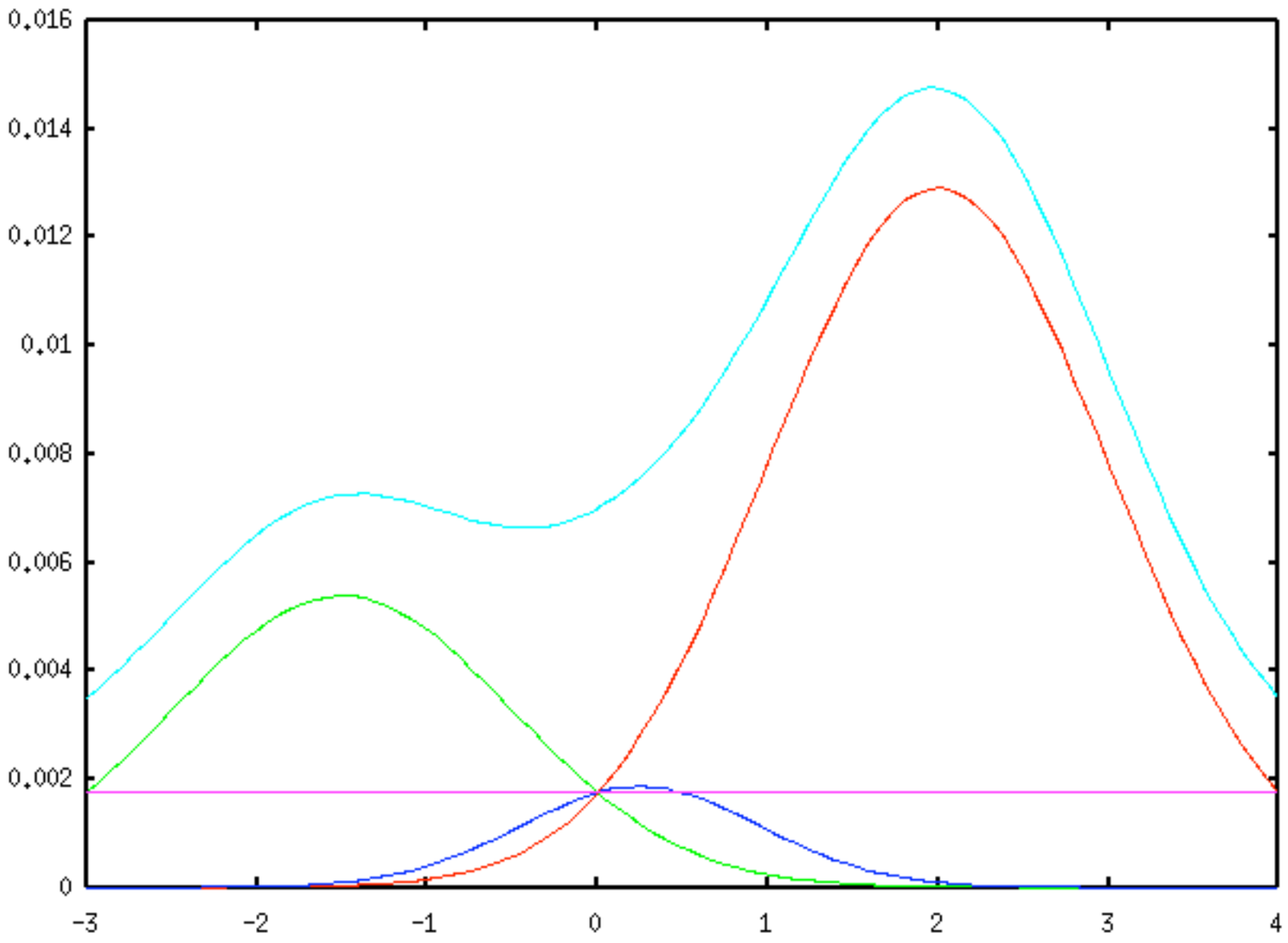


Optimizing a bound

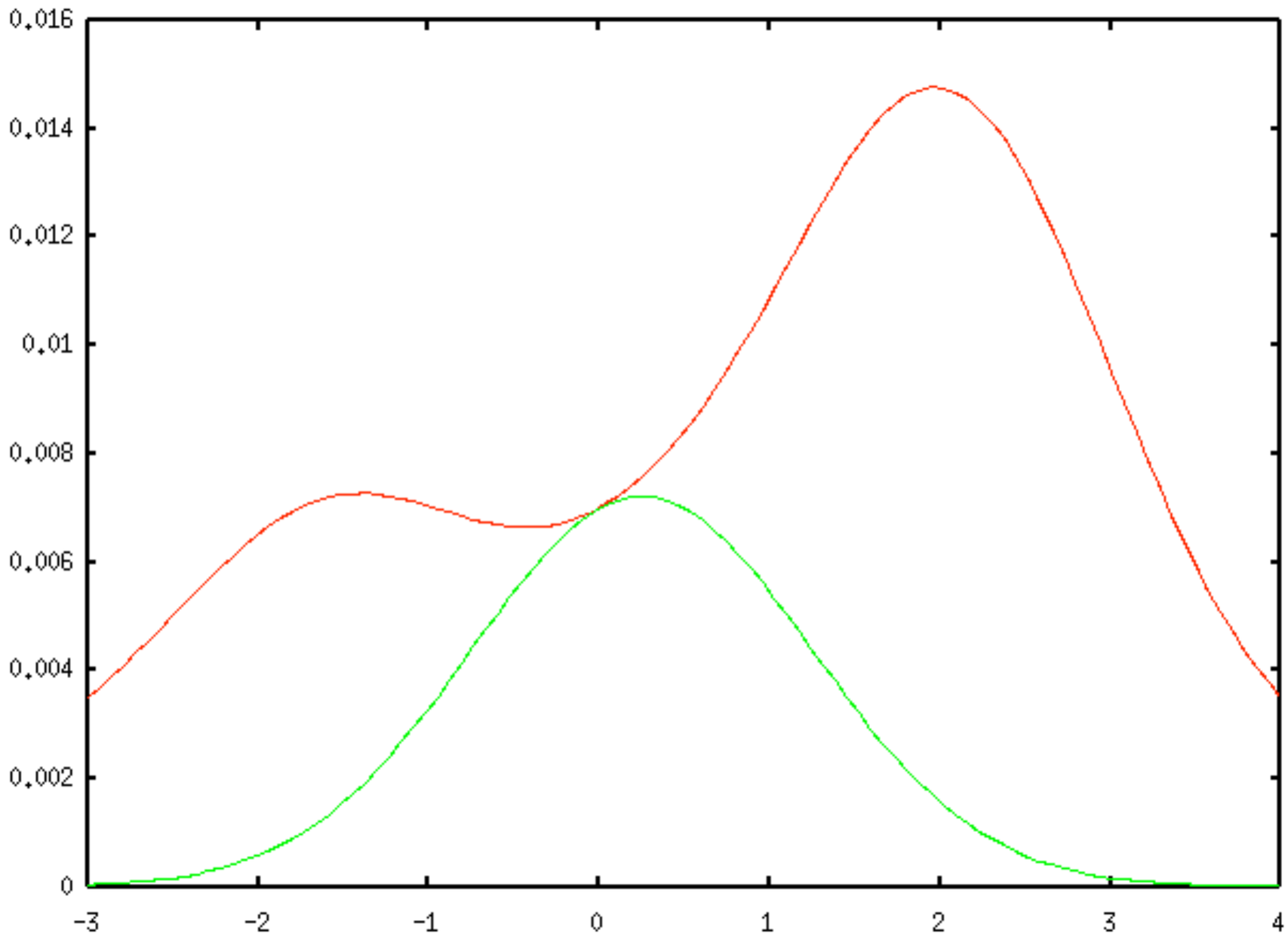
- $L(\theta) = \ln \sum_z \exp(\ln P(X,z \mid \theta)) \geq$
  - for any distribution $q = \langle q_z \rangle$

- Maximizing $L(\theta)$ is hard
- So, maximize $L(\theta, q)$ instead
  - start w/ arbitrary q, max wrt θ
  - then max wrt q to get a tighter bound
  - repeat

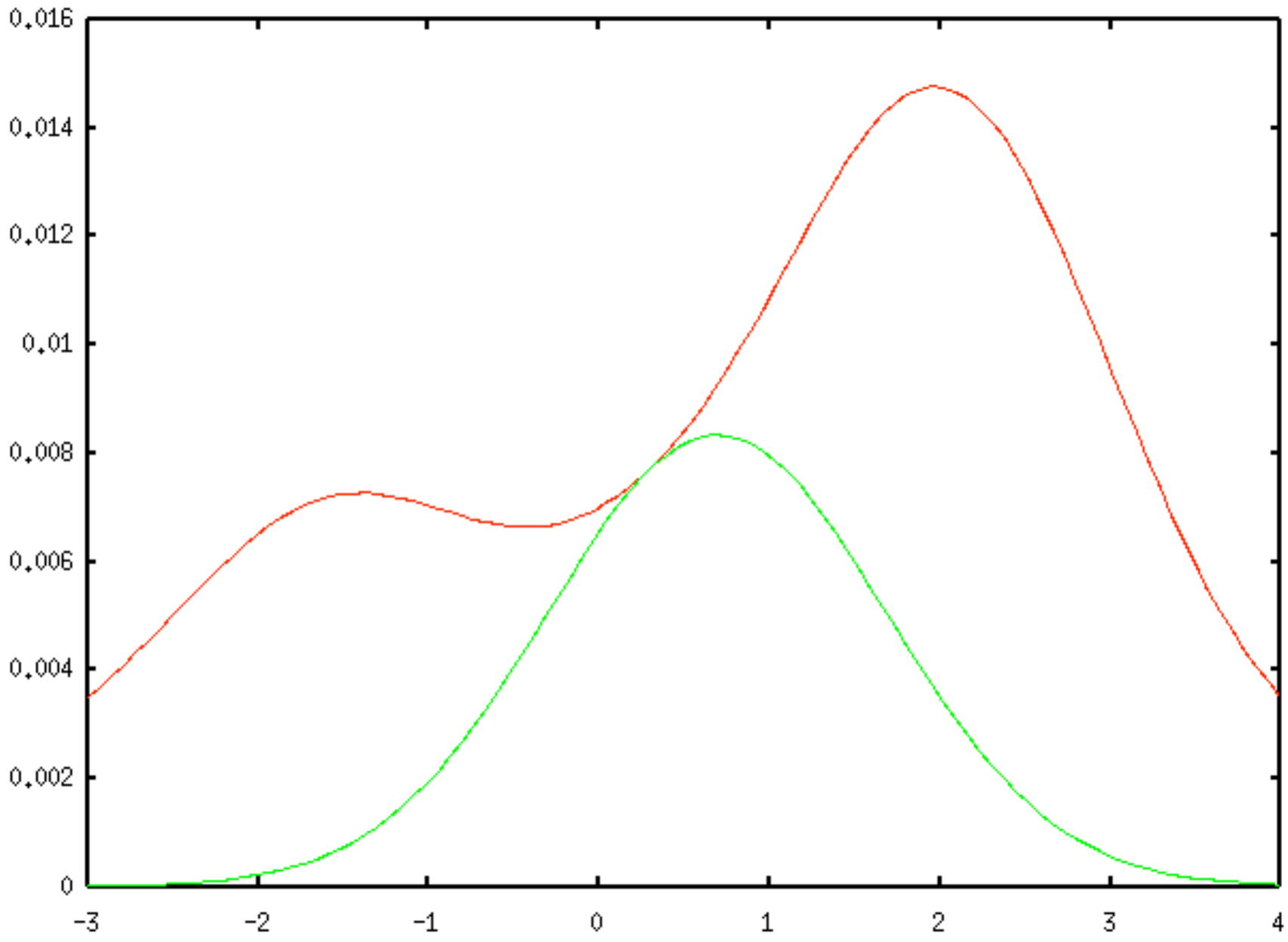
Example



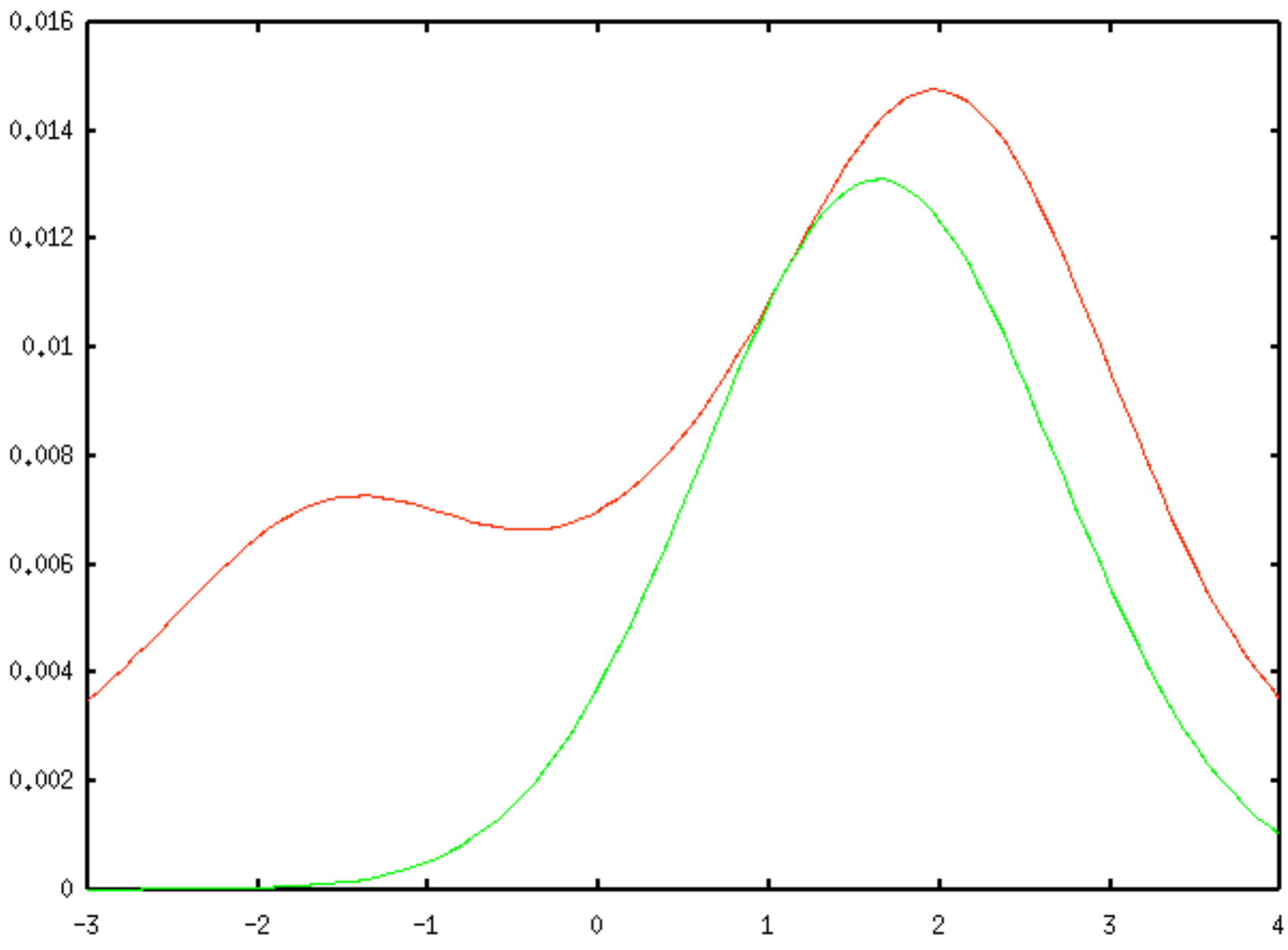
Example



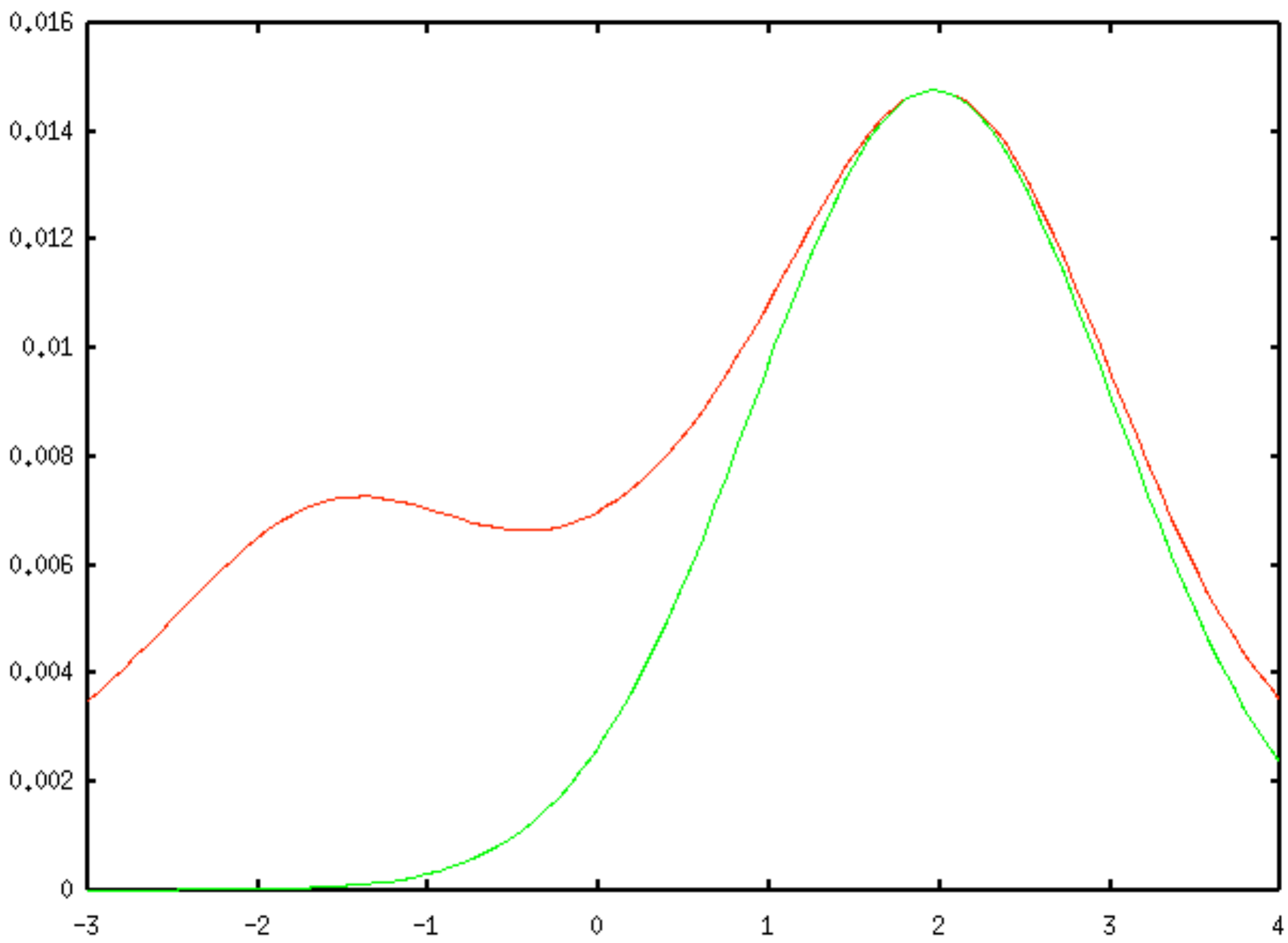
Example



Example



Example



Optimizing the bound

- $L(\theta, q) = \sum q_z \ln P(X, z \mid \theta) - \sum q_z \ln q_z$