THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 13: Gaussian Mixture Models and Density Based Clustering

Instructor: Sergei V. Kalinin

Clustering: Data Structures

- Hierarchical clustering
- K-means clustering
- Gaussian Mixture Models
- Density-based clustering
- Spectral clustering

Data matrix (two modes)

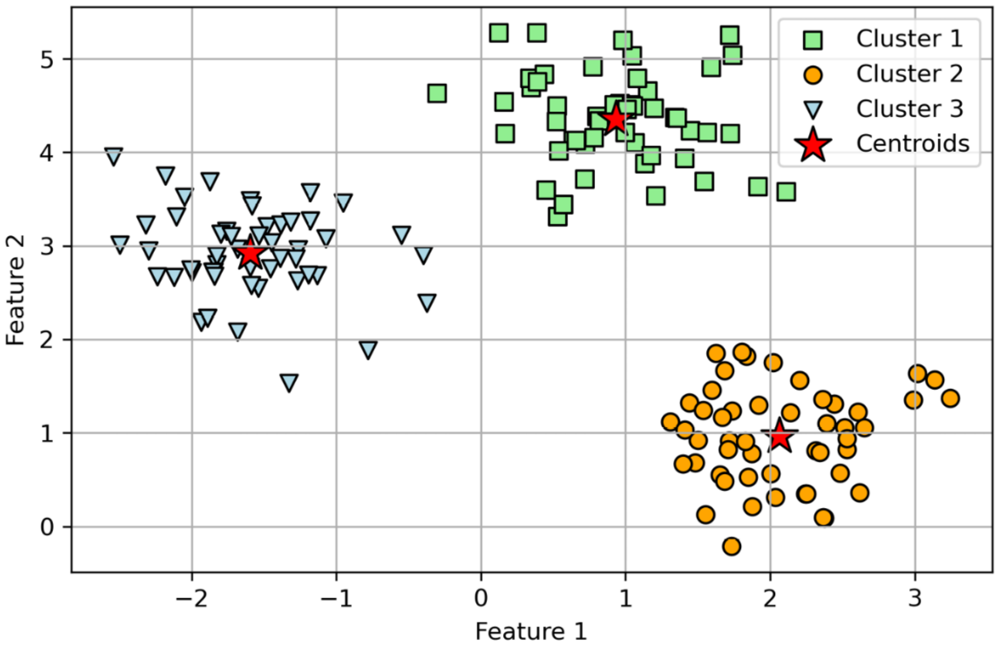
$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

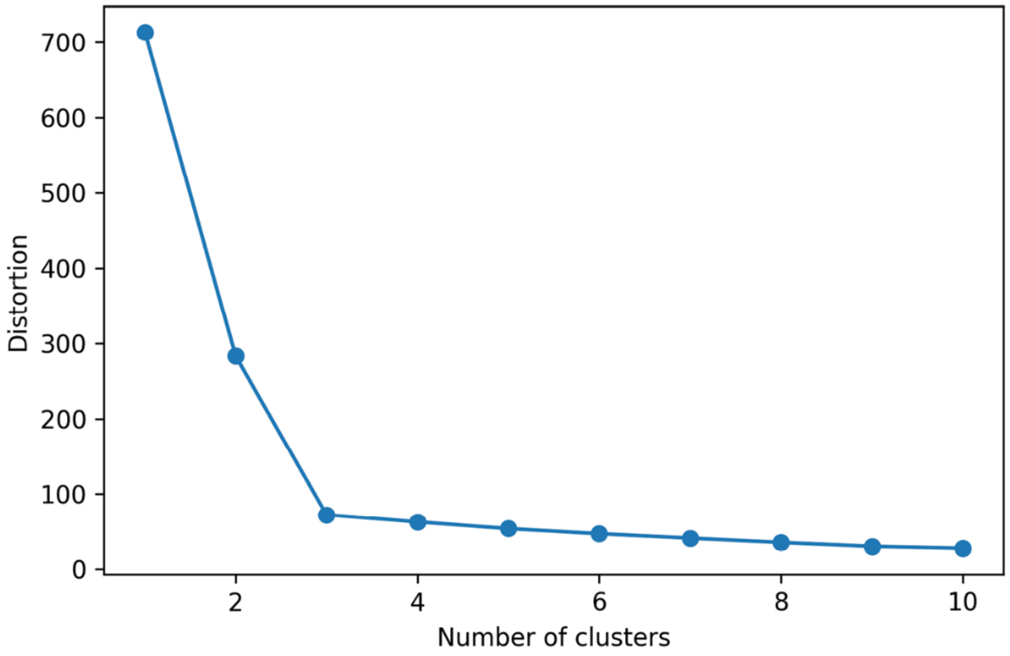
Dissimilarity matrix (one mode)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

From: Han and Kamber, Data Mining: Concepts and Techniques

K-means clustering and elbow methods





Calculate cluster size as a function of number of clusters

Silhouette

We need techniques that find a balance between inter-cluster similarity and intra-cluster dissimilarity

Silhouette:

- Scores any clustering with an arbitrary number of unique clusters. Clustering can come from any clustering algorithm.
- $a(i)$ = average dissimilarity of instance i to all other instances in the cluster to which i is assigned – **Minimize**
  - Dissimilarity could be Euclidian distance, etc.
- $b(i)$ = the smallest average dissimilarity of instance i to all instances in the closest cluster to $b(i)$ – **Maximize**
- $b(i)$ is smallest for the best different cluster that i could be assigned to – the best cluster that you would move i to if needed

Silhouette

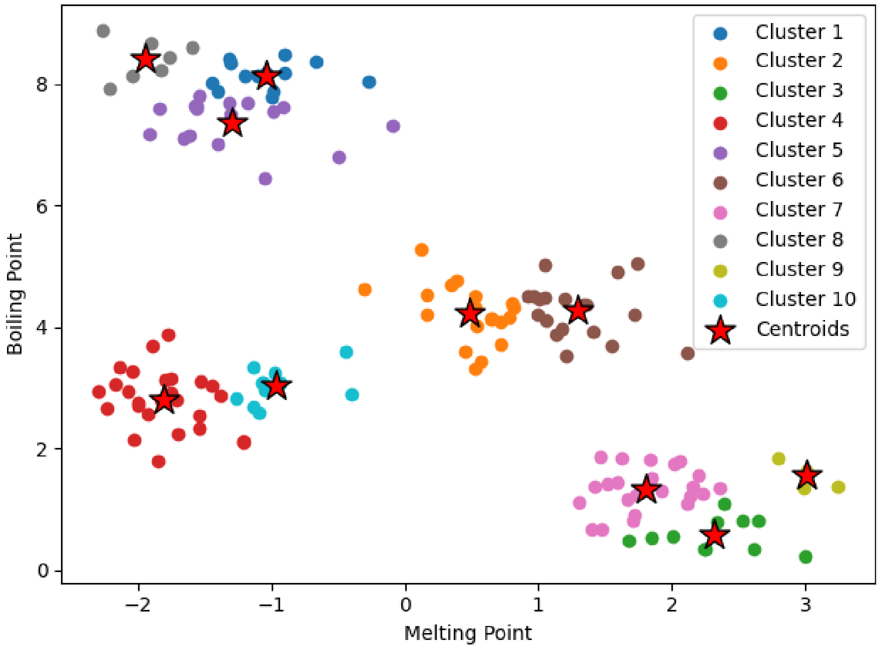
1. Calculate the **cluster cohesion**, $a^{(i)}$, as the average distance between an example, $\boldsymbol{x}^{(i)}$, and all other points in the same cluster.

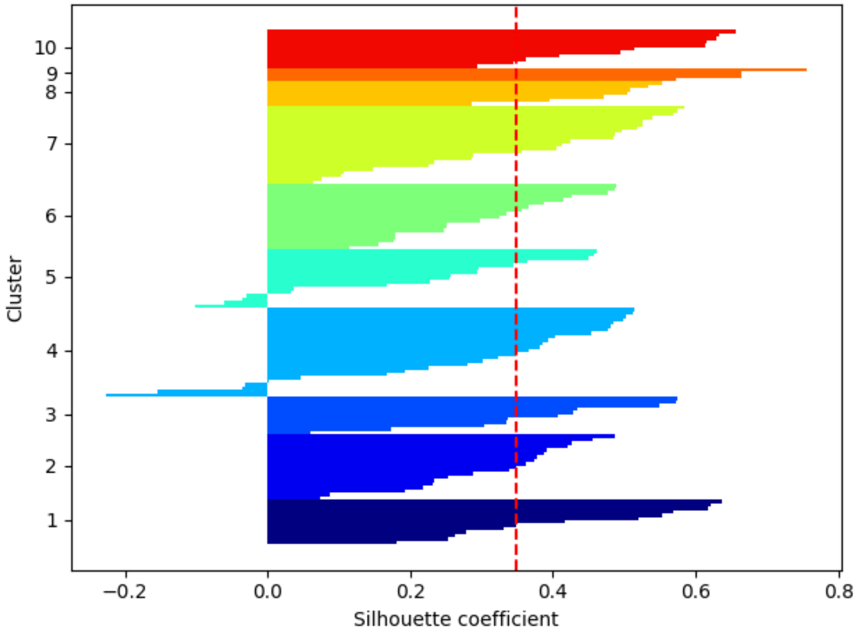
2. Calculate the **cluster separation**, $b^{(i)}$, from the next closest cluster as the average distance between the example, $\boldsymbol{x}^{(i)}$, and all examples in the nearest cluster.

3. Calculate the silhouette, $s^{(i)}$, as the difference between cluster cohesion and separation divided by the greater of the two, as shown here:

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{max\{b^{(i)}, a^{(i)}\}}$$

Silhouette





- The quality of a single cluster can be measured by the average silhouette score of its members, (close to 1 is best)
- The quality of a total clustering can be measured by the average silhouette score of all the instances
- To find best clustering, compare total silhouette scores across clusterings with different *k* values and choose the highest

Summary of k-means clustering

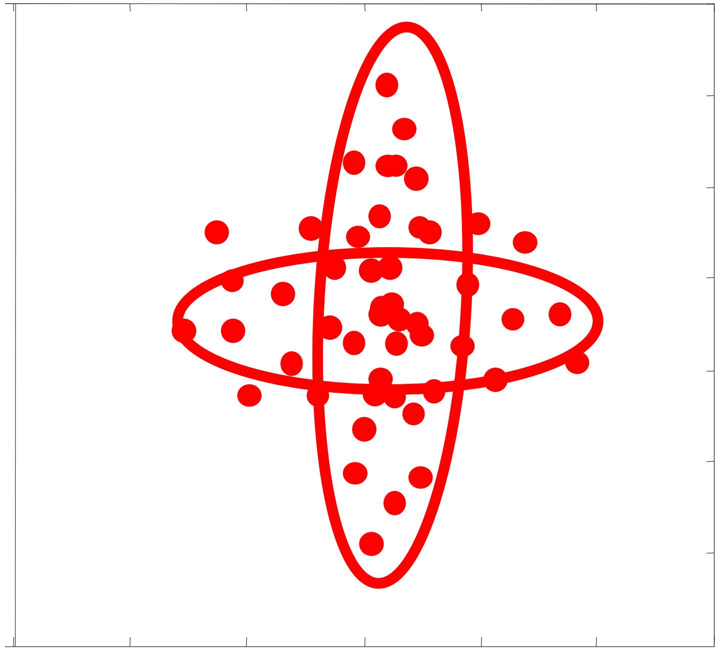
- **Strengths**
  - *Relatively efficient*: $O(tkn)$, where n is number of objects, k is number of clusters, and t is number of iterations. Normally, k, $t << n$.
  - Often terminates at a local optimum

- **Weakness**
  - Applicable only when mean is defined (what about categorical data)?
  - Need to specify k, the number of clusters, in advance
  - Unable to handle noisy data and outliers
  - Not suitable to discover clusters with non-convex shapes
  - Scales matter

Mixture of Gaussians

K-means algorithm

- Assigned each example to exactly one cluster
- What if clusters are overlapping?
  - Hard to tell which cluster is right
  - Maybe we should try to remain uncertain
- Used Euclidean distance
- What if cluster has a non-circular shape?



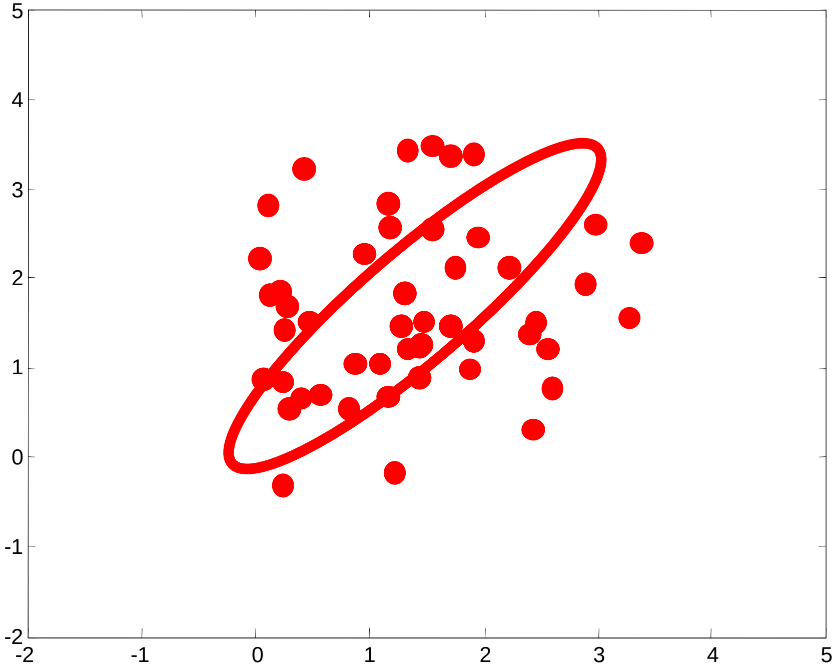
Gaussian mixture models

- Clusters modeled as Gaussian distributions
- EM algorithm: assign data to cluster with some *probability*

(adapted from) Prof. Alexander Ihler

Multivariate Gaussian Model

$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$



Maximum Likelihood estimates

$$\hat{\mu} = \frac{1}{N} \sum_i x^{(i)}$$

$$\hat{\Sigma} = \frac{1}{N} \sum_i (x^{(i)} - \hat{\mu})^T (x^{(i)} - \hat{\mu})$$

We model each cluster using Gaussian distribution

(adapted from) Prof. Alexander Ihler

Expectation Maximization: E-Step

- Initialize parameters of each cluster: mean μ_c, Covariance Σ_c, size π_c

- **E-step ("Expectation")**
  - For each datum (example) x_i,
  - Compute r_{ic}, the probability that it belongs to cluster c
    - Compute its probability under model c
    - Normalize to sum to one (over clusters c)

$$r_{ic} = \frac{\pi_c \mathcal{N}(x_i \; ; \; \mu_c, \Sigma_c)}{\sum_{c'} \pi_{c'} \mathcal{N}(x_i \; ; \; \mu_{c'}, \Sigma_{c'})}$$

  - If x_i is very likely under the c^{th} Gaussian, it gets high weight
  - Denominator just makes probabilities to sum to one

(adapted from) Prof. Alexander Ihler

Expectation Maximization: M-Step

- Start with assignment probabilities r_{ic}
- Update parameters: mean μ_c, Covariance Σ_c, "size" π_c

- M-step ("Maximization")
  - For each Gaussian cluster x_c,
  - Update its parameters using the (weighted) data points

$$N_c = \sum_i r_{ic}$$

Total responsibility allocated to cluster c

$$\pi_c = \frac{N_c}{N}$$

Fraction of total assigned to cluster c

$$\mu_c = \frac{1}{N_c} \sum_i r_{ic} x_i$$

Weighted mean of assigned data

$$\Sigma_c = \frac{1}{N_c} \sum_i r_{ic} (x_i - \mu_c)^T (x_i - \mu_c)$$

Weighted covariance of assigned data (use new weighted means here)

(adapted from) Prof. Alexander Ihler

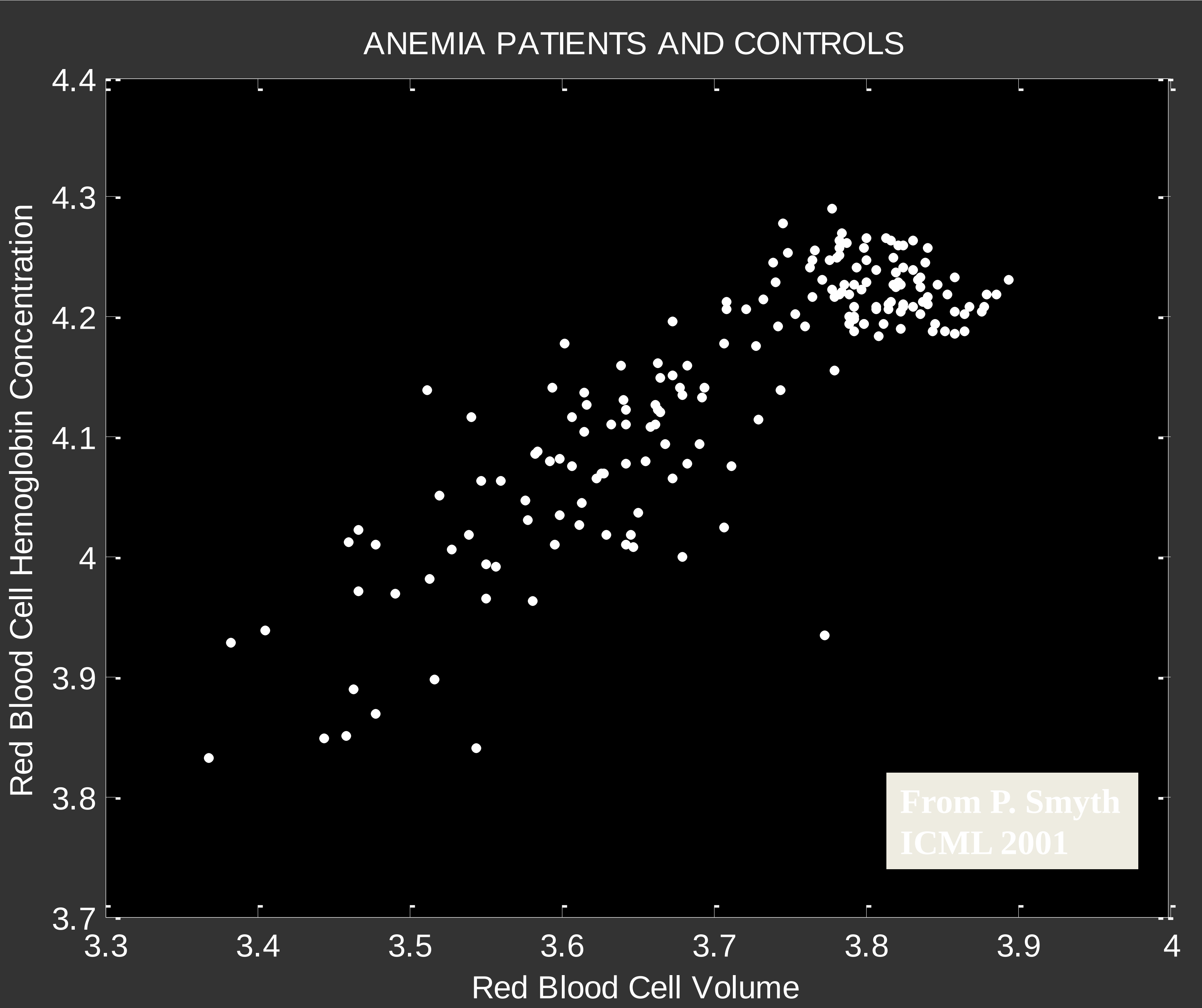
Expectation Maximization

- Each step increases the log-likelihood of our model

$$\log p(\underline{X}) = \sum_i \log \left[\sum_c \pi_c \, \mathcal{N}(x_i \; ; \; \mu_c, \Sigma_c) \right]$$

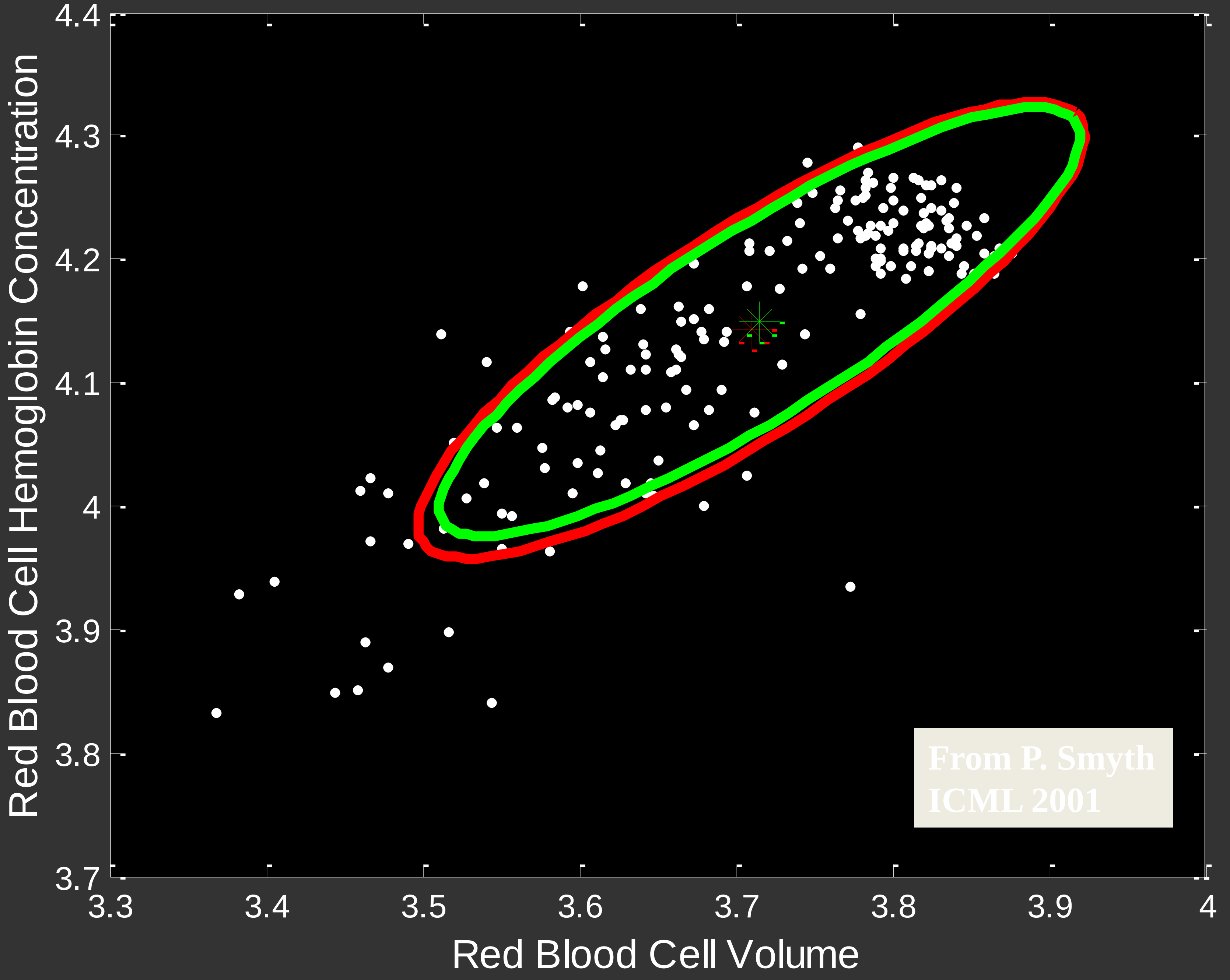
- Iterate until convergence
  – Convergence guaranteed – another ascent method

- What should we do
  – If we want to choose a single cluster for an "answer"?
  – With new data we didn't see during training?



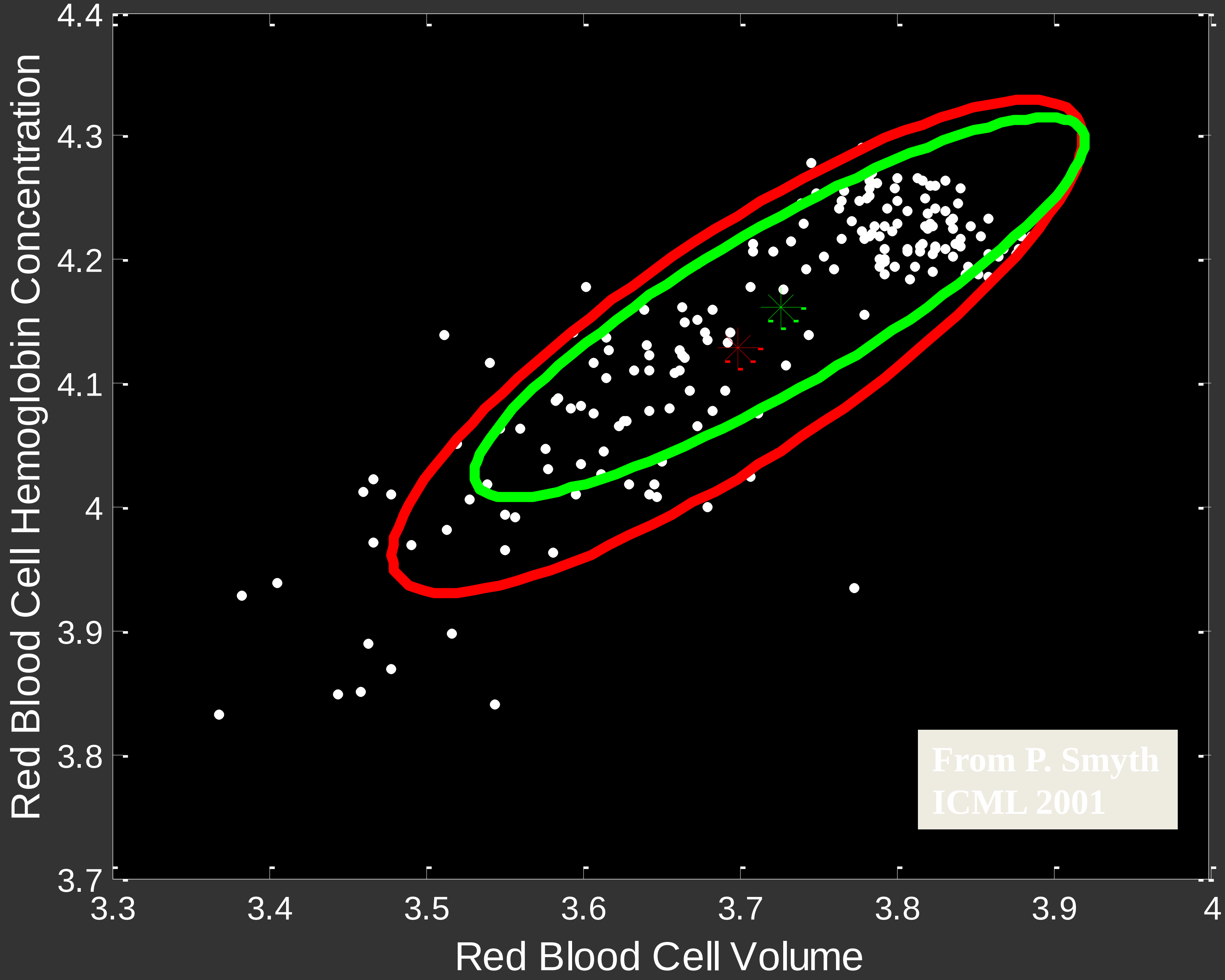
From P. Smyth
ICML 2001

EM ITERATION 1

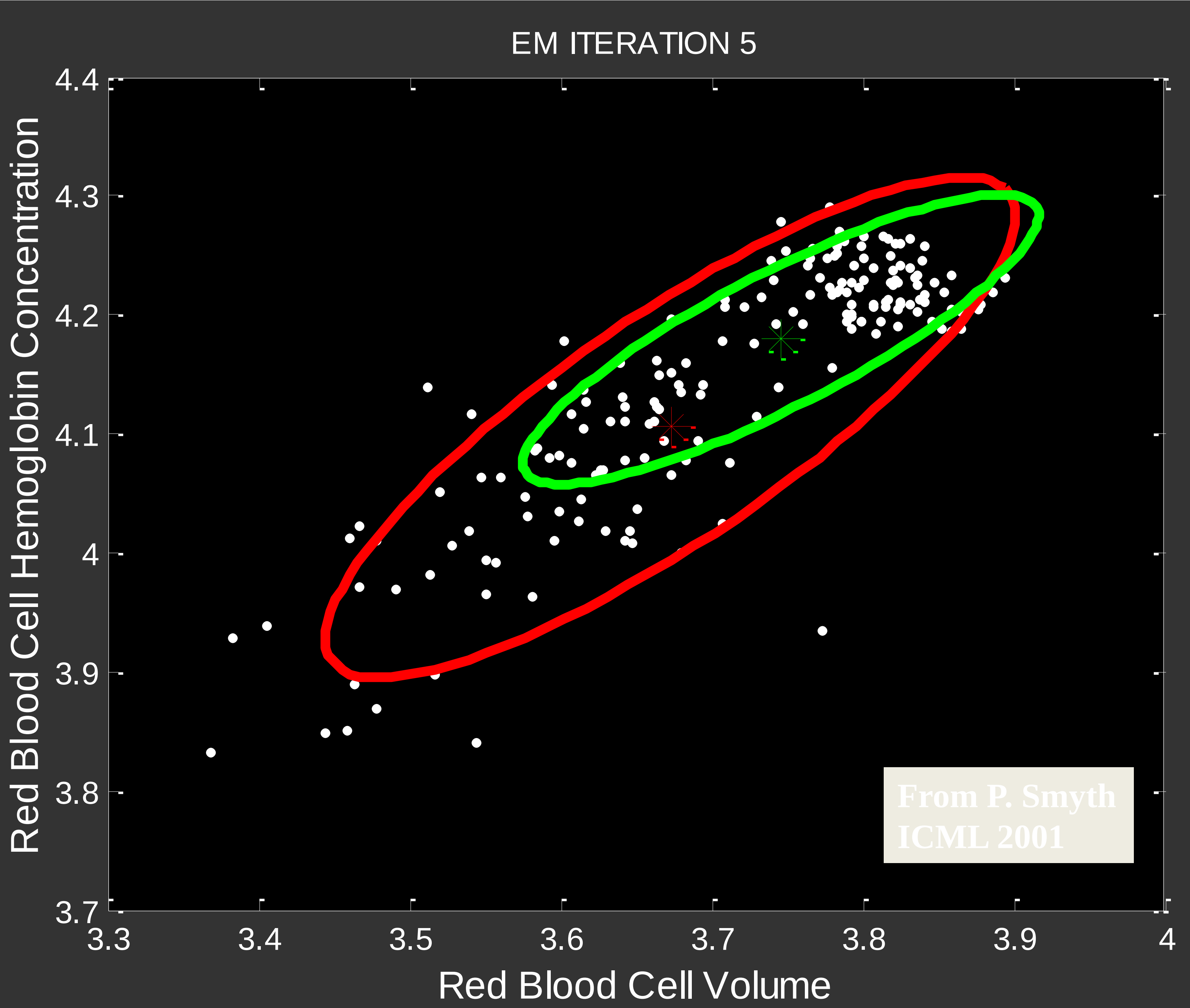


From P. Smyth
ICML 2001

EM ITERATION 3

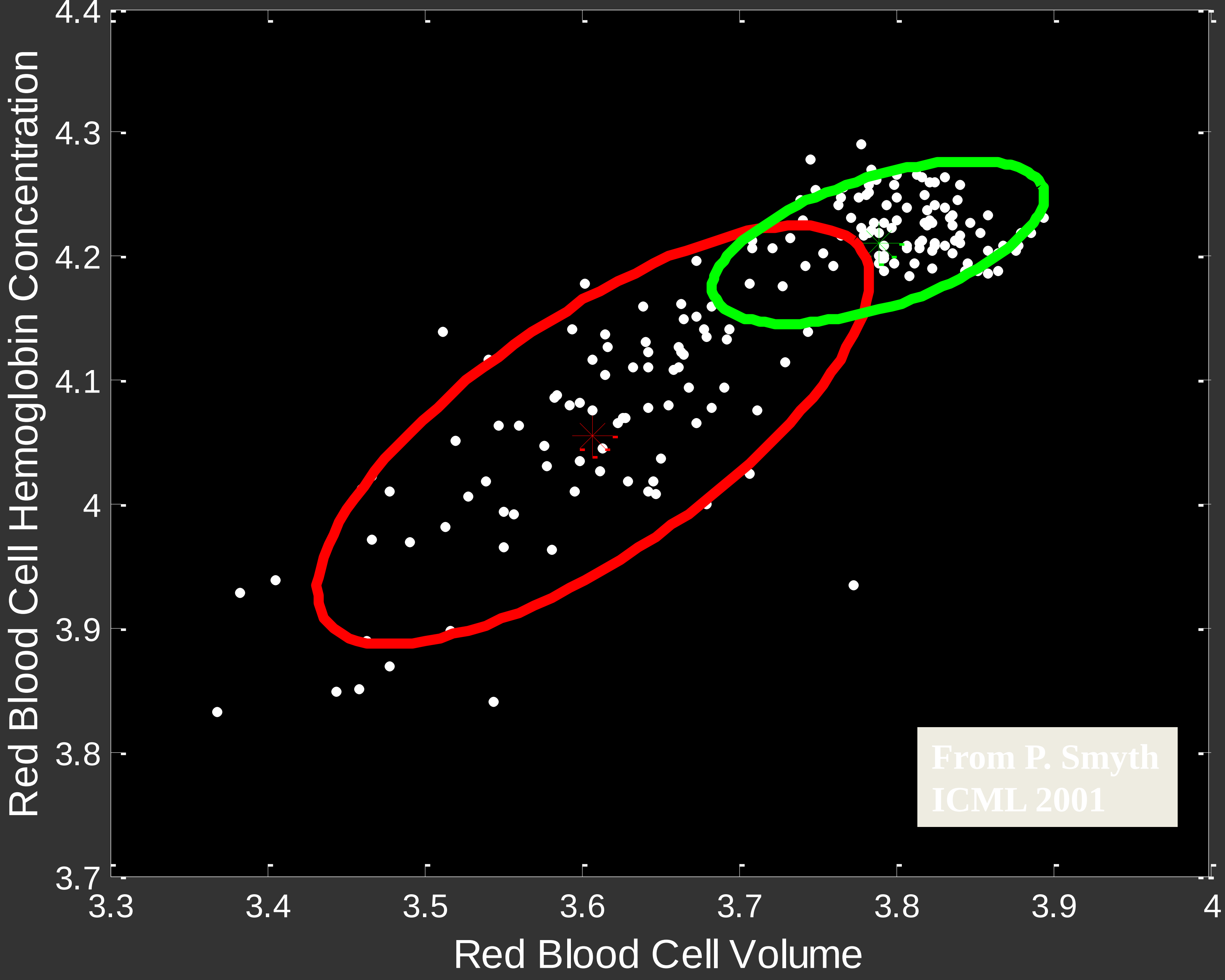


From P. Smyth
ICML 2001



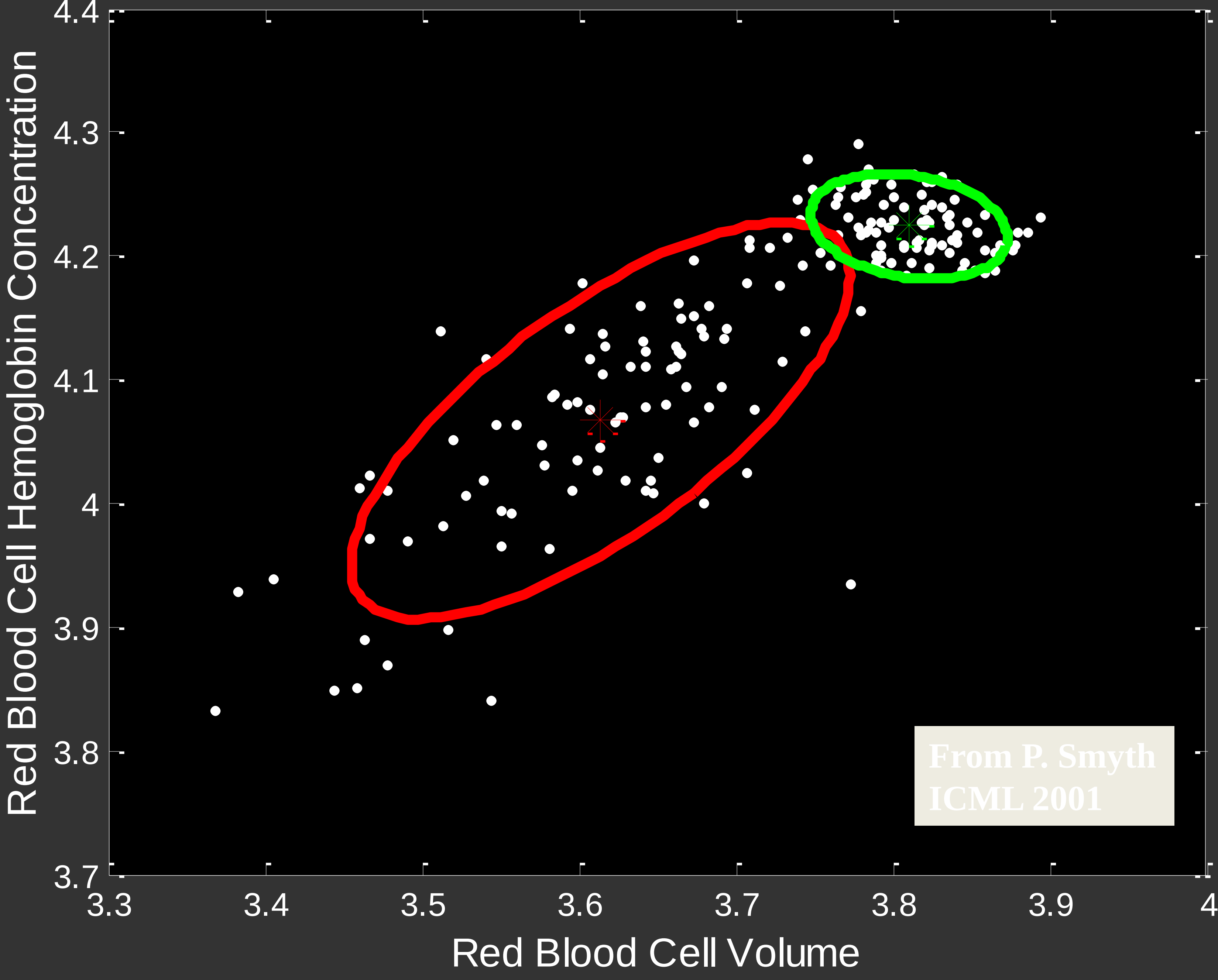
From P. Smyth
ICML 2001

EM ITERATION 10



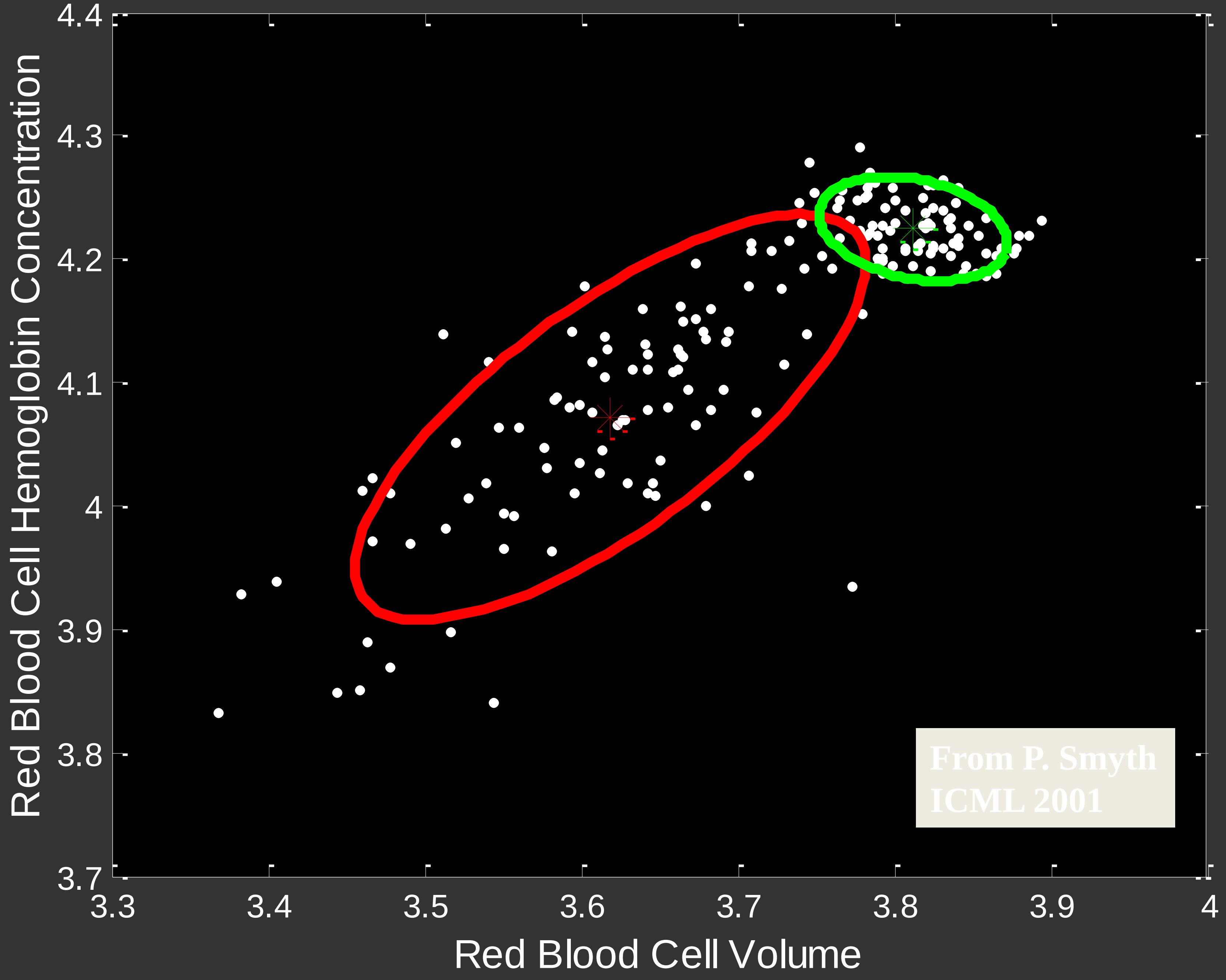
From P. Smyth
ICML 2001

EM ITERATION 15



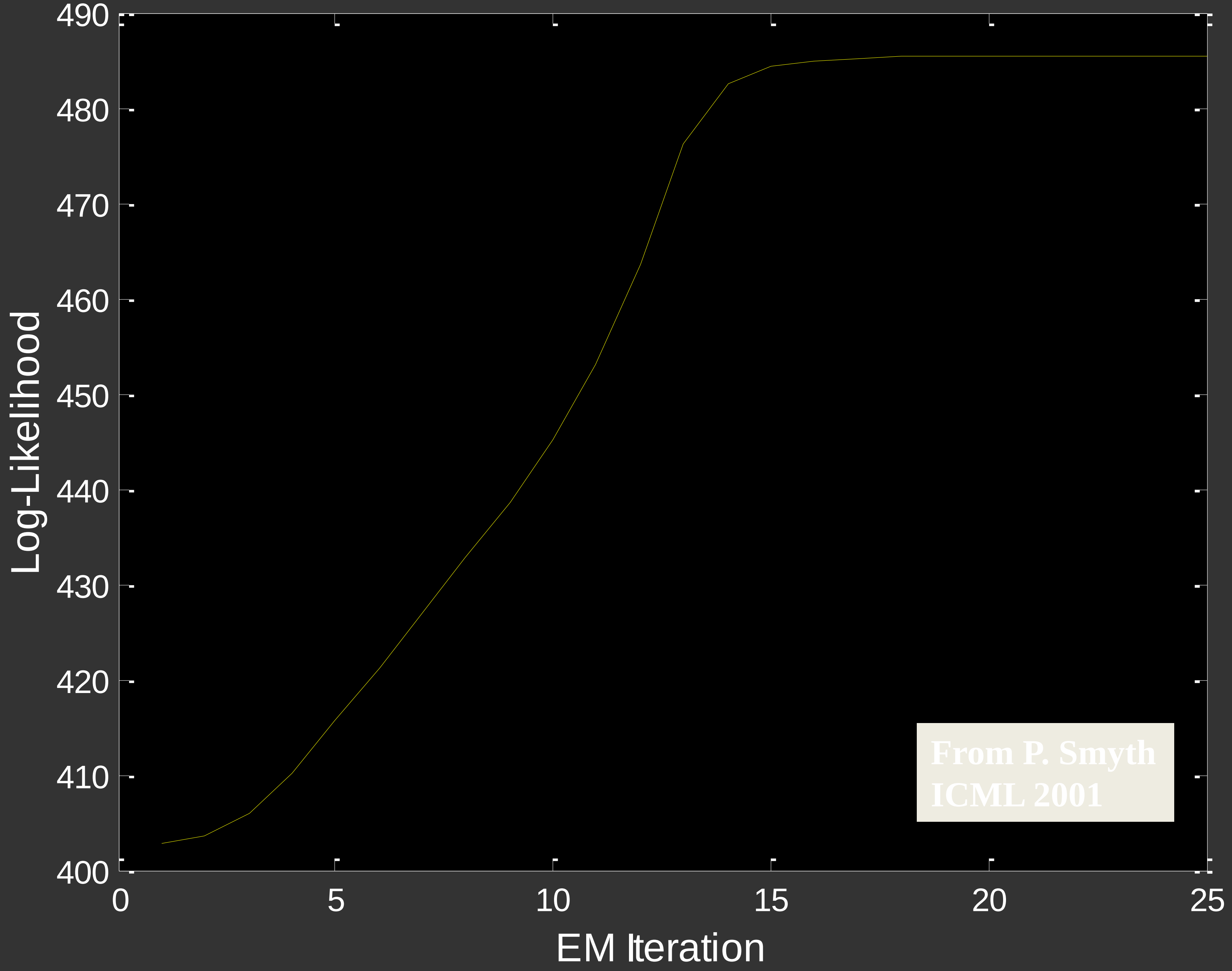
From P. Smyth
ICML 2001

EM ITERATION 25



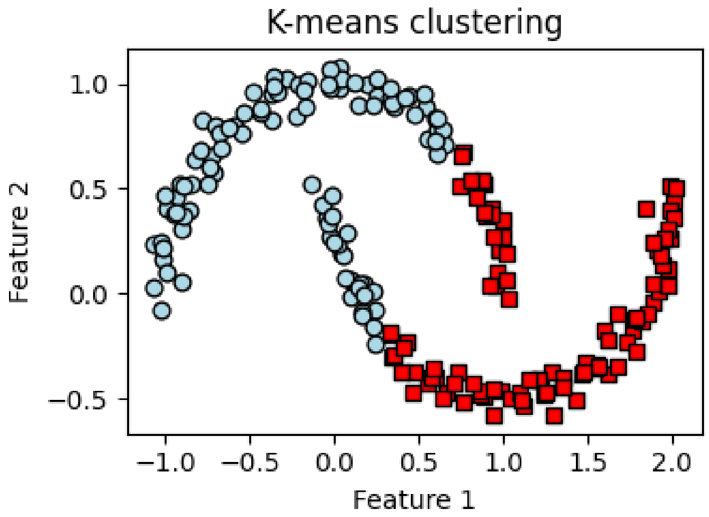
From P. Smyth
ICML 2001

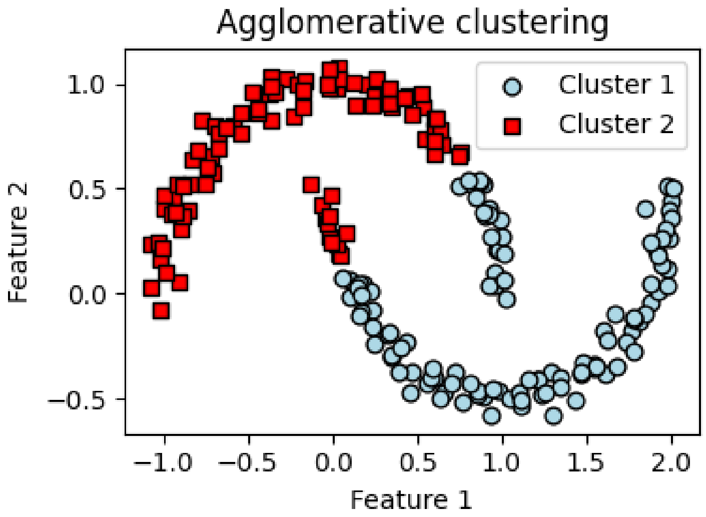
LOG-LIKELIHOOD AS A FUNCTION OF EM ITERATIONS

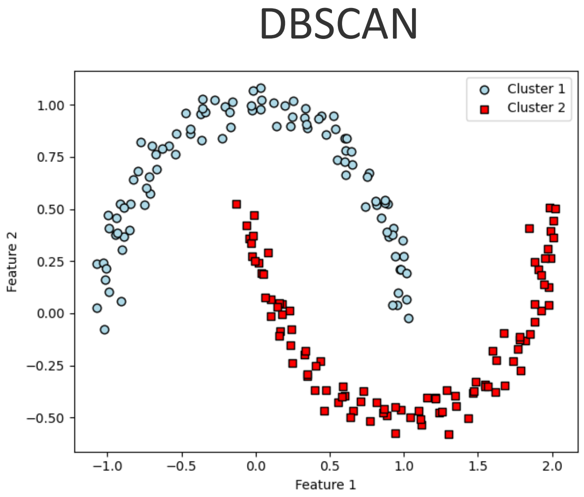


From P. Smyth
ICML 2001

Density Based Clustering


K-means clustering


Agglomerative clustering

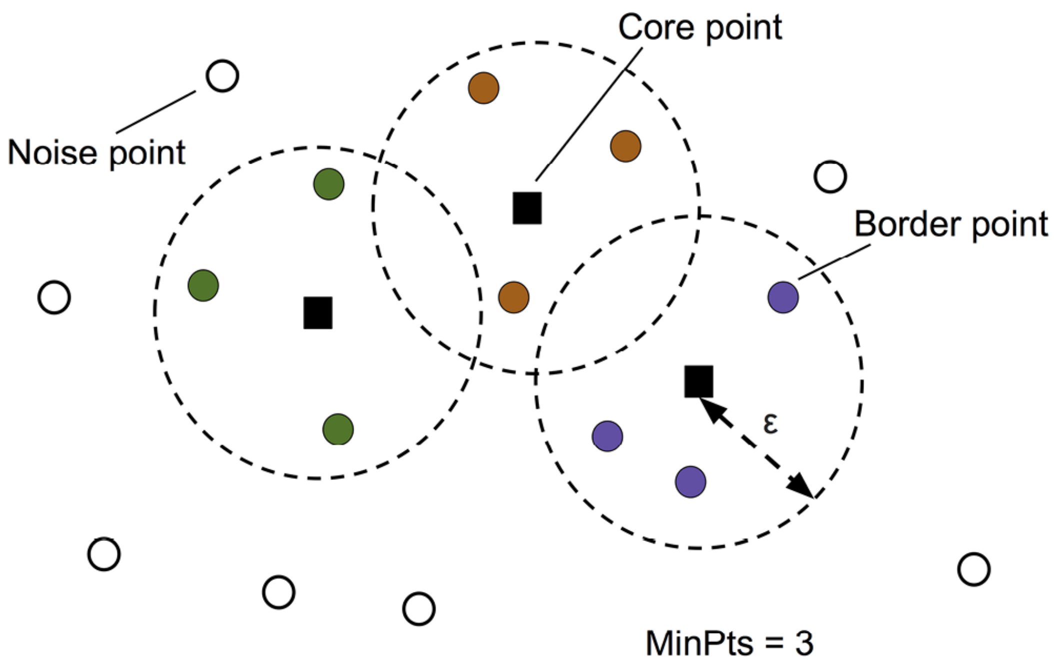

DBSCAN

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise

Density Based Clustering

- Two parameters**:**
  - ***Eps***: Maximum radius of the neighbourhood
  - ***MinPts***: Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid dist(p,q) <= Eps\}$
- Directly density-reachable**:** A point ***p*** is directly density-reachable from a point ***q*** wrt. ***Eps***, ***MinPts*** if
  - ***p*** belongs to $N_{Eps}(q)$
  - core point condition: $|N_{Eps}(q)| >= MinPts$

Density Based Clustering



- Arbitrary select a point ***p***
- Retrieve all points density-reachable from ***p*** wrt ***Eps*** and ***MinPts***.
- If ***p*** is a core point, a cluster is formed.
- If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

Summary

- In clustering, clusters are inferred from the data without human input (unsupervised learning)

- However, in practice, it is very domain specific:
  - Definition of distance in data space
  - Representation of data
  - Defining distance between clusters
  - Number of clusters
  - And so on.

- Practice, practice, practice!