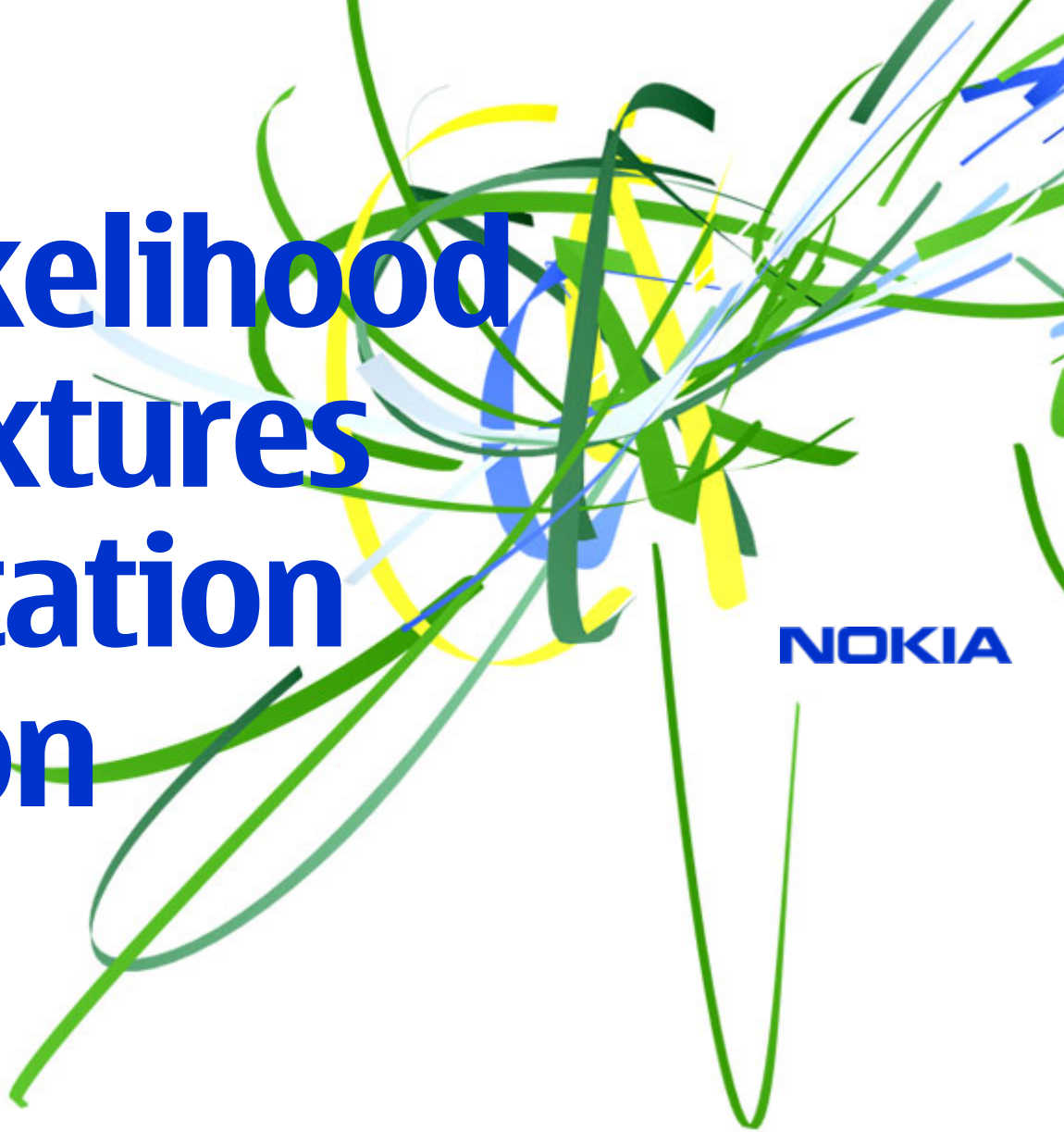
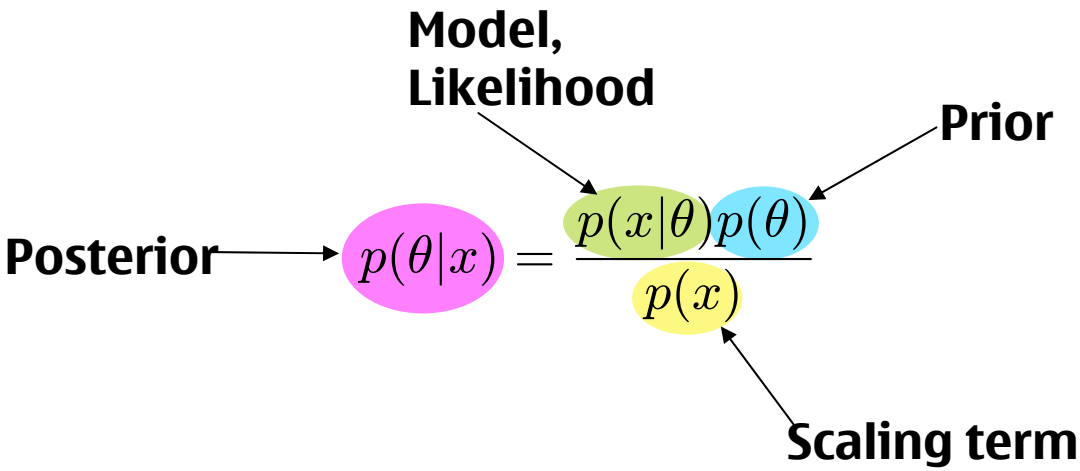
Maximum likelihood Gaussian mixtures using expectation maximization



Jarno Seppänen
2008-02-27

Bayes vs. maximum likelihood



Point estimate

$$\hat{\theta} = \text{argmax}_\theta \; p(x|\theta)$$

- **Model**: what we are observing (data generating process)
- **Likelihood**: model applied to specific data
- **Prior distribution**: before-the-fact beliefs of different outcomes
- **Posterior distribution**: after-the-fact distribution of model parameters
- **Maximum likelihood**: point estimate of parameters, ignoring prior and posterior
- **Maximum likelihood** $\neq$ **Bayesian**

Example: flipping two coins

- Flipped coin is picked randomly with probability $p(z_k = 1) = \pi_k$

Which coin was flipped (unknown)

z_1	z_2
0	1
0	1
1	0
1	0
0	1
1	0
0	1
0	1
1	0
0	1
1	0
1	0
0	1
1	0

Example: flipping two coins

- Two coins with unknown probabilities θ_1 and θ_2 of landing heads

$$p(x|z_1, z_2, \theta_1, \theta_2) = \mathrm{B}(x|\theta_1)^{z_1}\mathrm{B}(x|\theta_2)^{z_2}$$

$$\mathrm{B}(x|\theta) = \theta^x(1-\theta)^{1-x}$$

- Flipped coin is picked randomly with probability $p(z_k = 1) = \pi_k$

Outcome (heads or tails)

z_1	z_2	x
0	1	1
0	1	0
1	0	0
1	0	1
0	1	1
1	0	0
0	1	0
0	1	1
1	0	0
0	1	0
1	0	1
1	0	0
0	1	1
1	0	0

Example: flipping two coins

- Two
and
$p($

- Flip
pro

Problem: *since each flipped coin z_{ik} is hidden, how can we learn anything from just the observations x_i?*

Solution: *estimate the* ***responsibility*** *$\gamma(z_{ik})$ of each coin pick z_{ik} for each observation x_i:*

$$\begin{aligned}\gamma(z_k) &\equiv p(z_k = 1|x, \pi_1, \pi_2, \theta_1, \theta_2) \\ &= \frac{p(x|z_k = 1, \theta_k)p(z_k = 1|\pi_k)}{p(x|\pi, \theta)} \\ &= \frac{\pi_k \mathrm{B}(x|\theta_k)}{\sum_j \pi_j \mathrm{B}(x|\theta_j)}\end{aligned}$$

1	0	0

Example: flipping two coins

- Two coins with unknown probabilities θ_1 and θ_2 of landing heads

$$p(x|z_1, z_2, \theta_1, \theta_2) = \mathrm{B}(x|\theta_1)^{z_1}\mathrm{B}(x|\theta_2)^{z_2}$$

$$\mathrm{B}(x|\theta) = \theta^x(1-\theta)^{1-x}$$

- Flipped coin is picked randomly with probability $p(z_k = 1) = \pi_k$
- Unknowns:
  - Which coin was flipped for each *i*?

$$\gamma(z_k) \equiv p(z_k = 1|x, \pi_1, \pi_2, \theta_1, \theta_2) = \frac{\pi_k \mathrm{B}(x|\theta_k)}{\sum_j \pi_j \mathrm{B}(x|\theta_j)}$$

  - What were the parameters?

$$\hat{\pi}_k = \frac{1}{N}\sum_i \gamma(z_{ik}), \quad \hat{\theta}_k = \frac{\sum_i \gamma(z_{ik})x_i}{\sum_i \gamma(z_{ik})}$$

Responsibility (estimate)

z_1	z_2	x	$\gamma(z_1)$	$\gamma(z_2)$
0	1	1	0.2	0.8
0	1	0	0.4	0.6
1	0	0	0.9	0.1
1	0	1	0.6	0.4
0	1	1	0.2	0.8
1	0	0	0.8	0.2
0	1	0	0.3	0.7
0	1	1	0.2	0.8
1	0	0	0.7	0.3
0	1	0	0.4	0.6
1	0	1	0.7	0.3
1	0	0	0.6	0.4
0	1	1	0.3	0.7
1	0	0	0.7	0.3

Example: flipping two coins

- Two coins with unknown probabilities θ_1 and θ_2 of landing heads

$$p(x|z_1, z_2, \theta_1, \theta_2) = \mathrm{B}(x|\theta_1)^{z_1}\mathrm{B}(x|\theta_2)^{z_2}$$

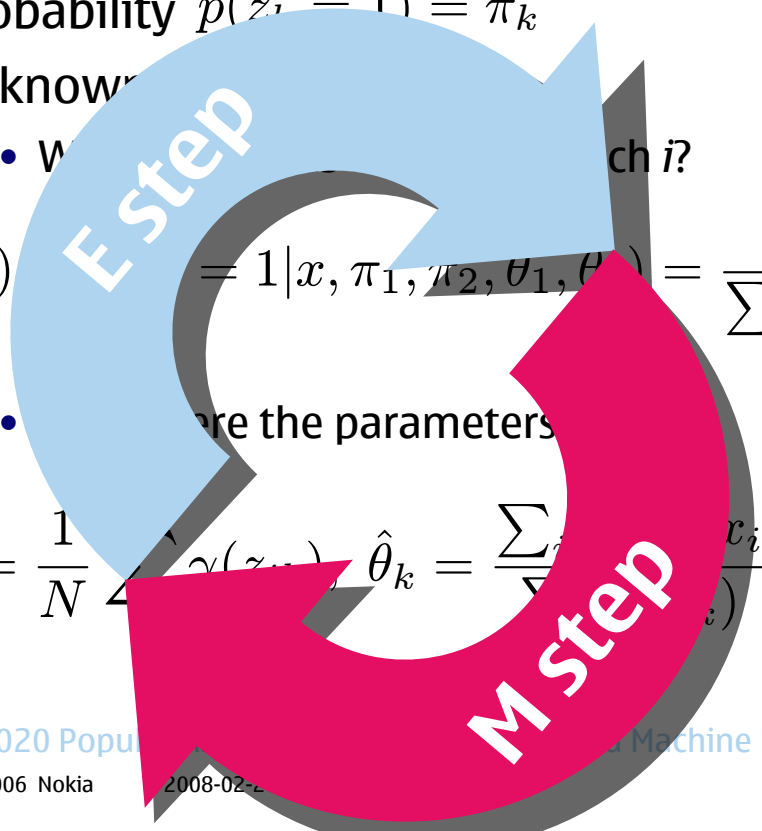
$$\mathrm{B}(x|\theta) = \theta^x(1-\theta)^{1-x}$$

- Flipped coin is picked randomly with probability $p(z_k = 1) = \pi_k$
- Unknow
  - W ch *i*?

$$\gamma(z_k) = 1|x, \pi_1, \pi_2, \theta_1, \theta = \frac{\pi_k \mathrm{B}(x|\theta_k)}{\sum_j \pi_j \mathrm{B}(x|\theta_j)}$$

  - re the parameters

$$\hat{\pi}_k = \frac{1}{N} \quad \hat{\theta}_k = \frac{\sum_i \quad x_i}{\quad)}$$



z_1	z_2	x	$\gamma(z_1)$	$\gamma(z_2)$
0	1	1	0.2	0.8
0	1	0	0.4	0.6
1	0	0	0.9	0.1
1	0	1	0.6	0.4
0	1	1	0.2	0.8
1	0	0	0.8	0.2
0	1	0	0.3	0.7
0	1	1	0.2	0.8
1	0	0	0.7	0.3
0	1	0	0.4	0.6
1	0	1	0.7	0.3
1	0	0	0.6	0.4
0	1	1	0.3	0.7
1	0	0	0.7	0.3

Expectation maximization

- The EM algorithm is used to solve maximum likelihood problems in general
- Applies when the likelihood has unknown *hidden (latent) variables*
  - Latent variables are model assumptions
  - Mixture problems can be stated using discrete latent variables
  - Hidden Markov model state is a discrete latent variable
  - Kalman filter state is a continuous latent variable
- EM not tied to mixture problems however
  - Any kind of probabilistic model with unknown variables
  - EM for Hidden Markov models is also known as *Baum-Welch* or *forward-backward recursion*
  - Very widely used

Expectation maximization

- EM finds local likelihood maxima by iterative optimization
  - Alternating between *E step* and *M step* until convergence
  - **E step:** compute expected values of hidden variables, given parameters
  - **M step:** re-estimate parameters, given values for hidden variables
  - Analogous to *k*-means
  - Not guaranteed to find global likelihood maximum

Mixtures of Gaussians

- Multivariate Gaussian (normal) distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ parametrized by mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
- Gaussian mixture model is a linear combination of *K* Gaussian densities

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

  - π_k Component weights or mixing coefficients
  - $\boldsymbol{\mu}_k$ Component means
  - $\boldsymbol{\Sigma}_k$ Component covariances
  - $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$ Model parameters

Covariance structures

- Covariance matrix structure can be controlled
- Full covariance matrices
  - Unrestricted covariance (symmetric)
  - Rotated ellipsoid shape
  - $(D+1)D/2$ parameters
- Diagonal covariance $\boldsymbol{\Sigma}_{ij} = 0,\ i \neq j$
  - Zeros except on diagonal
  - Ellipsoid shape
  - *D* parameters
- Scaled identity covariance $\boldsymbol{\Sigma} = \lambda \boldsymbol{I}$
  - Zeros except on diagonal
  - Diagonal values are equal
  - Spherical shape
  - One parameter

Gaussian mixture likelihood

- Log-likelihood of Gaussian mixture

$$\log p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

- Has no closed-form solution

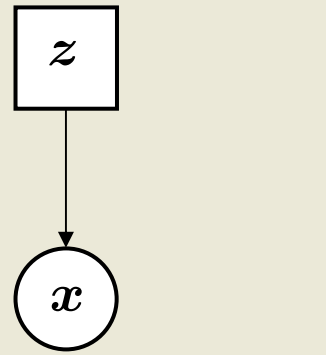
Latent variable formulation of Gaussian mixtures

- To apply EM, we introduce latent *K*-dimensional indicator variable $\boldsymbol{z}$ to Gaussian mixtures
  - $z_k = 1$ iff k^{th} component is active
  - $z_k = 0$ otherwise
- Then $p(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \pi_k^{z_k}$ and $p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\theta}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$
- Marginalizing over $\boldsymbol{z}$,

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Therefore, we obtain the same Gaussian mixture density as earlier

Factorization



$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z})$$

EM for Gaussian mixtures

- Complete-data log likelihood

$$\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) = \log \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$
$$= \sum_n \sum_k z_{nk} [\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]$$

- Responsibilities

$$\gamma(z_k) \equiv p(z_k = 1|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

EM algorithm for Gaussian mixtures

E step

- Evaluate responsibilities, given current parameters

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j^K \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

M step

- Re-estimate parameters, given current responsibilities

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nk})$$

$$\boldsymbol{\mu}_k = \frac{\sum_n \gamma(z_{nk})\boldsymbol{x}_n}{\sum_n \gamma(z_{nk})}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_n \gamma(z_{nk})}$$

EM algorithm for Gaussian mixtures

E step

- Evaluate responsibilities, given current parameters

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Mean responsibility

M step

- Re-estimate parameters using the current responsibilities

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\boldsymbol{\mu}_k = \frac{\sum_n \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_n \gamma(z_{nk})}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_n \gamma(z_{nk})}$$

EM algorithm for Gaussian mixtures

E step

- Evaluate responsibilities, given current parameters

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Mean responsibility

M step

- Re-estimate parameters using the current responsibilities

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\boldsymbol{\mu}_k = \frac{\sum_n \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_n \gamma(z_{nk})}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_n \gamma(z_{nk})}$$

Weighted mean of observations

EM algorithm for Gaussian mixtures

E step

- Evaluate responsibilities, given current parameters

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Mean responsibility

M step

- Re-estimate parameters using the current responsibilities

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})$$

Weighted mean of observations

$$\boldsymbol{\mu}_k = \frac{\sum_n \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_n \gamma(z_{nk})}$$

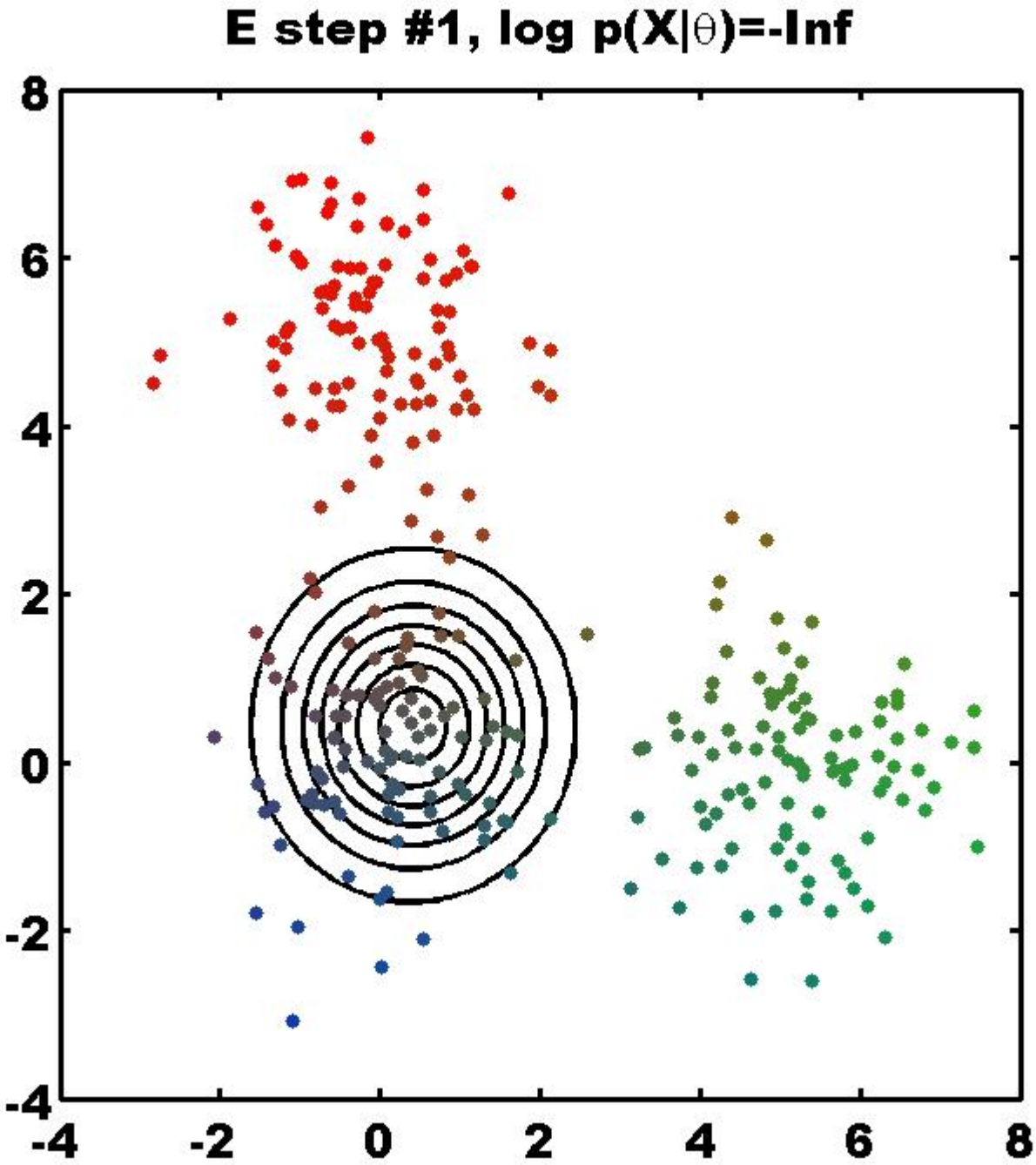
$$\boldsymbol{\Sigma}_k = \frac{\sum_n \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_n \gamma(z_{nk})}$$

Weighted mean of variances

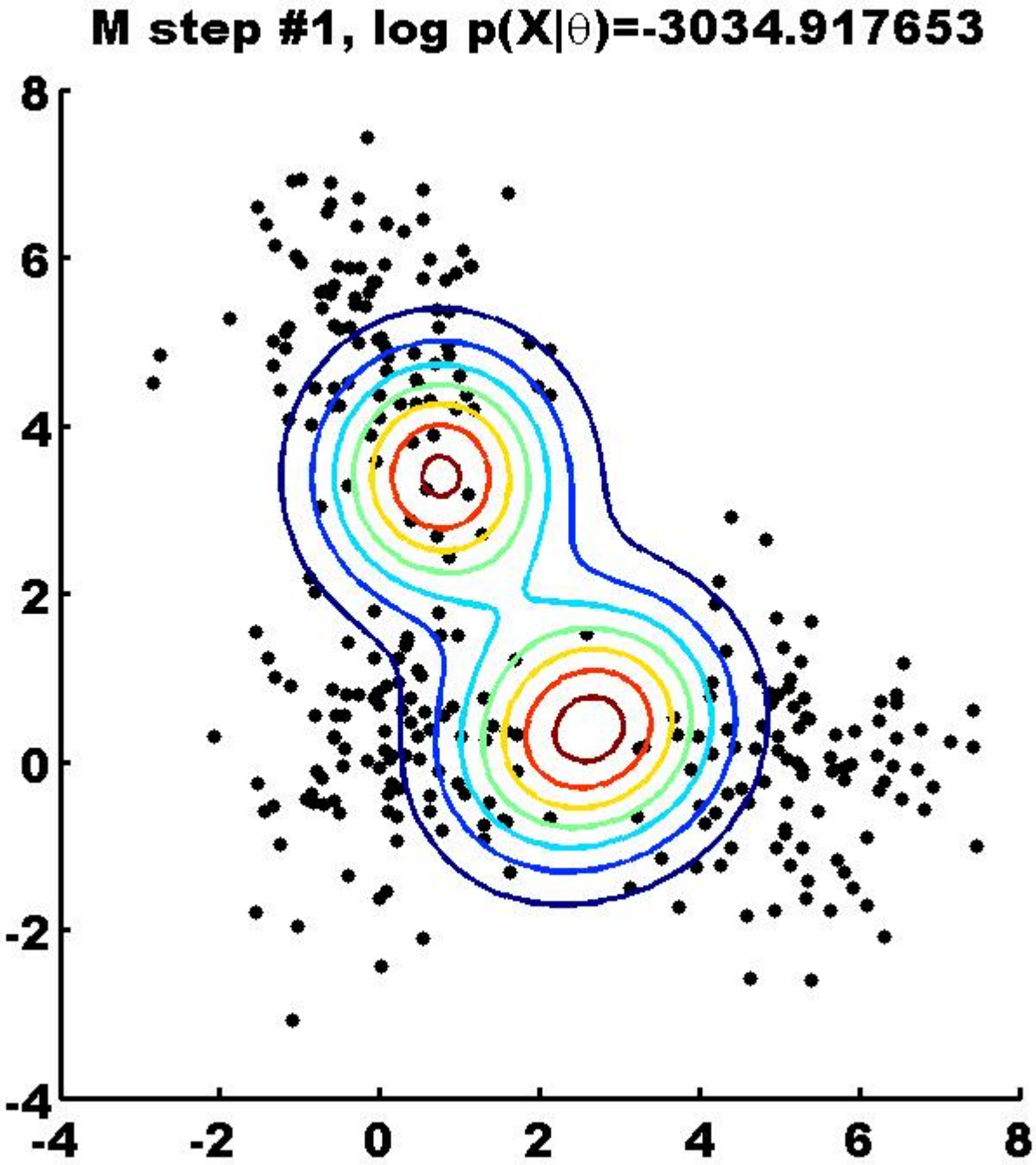
Demo: mixture of three Gaussians

- Two-dimensional data
- Find max likelihood estimate for means of a mixture of k=3 Gaussians
  - $\Sigma = I$
- E step
  - Show responsibilities in color
  - Red, green, blue color corresponding to each of three components
- M step
  - Show Gaussian mixture density in color

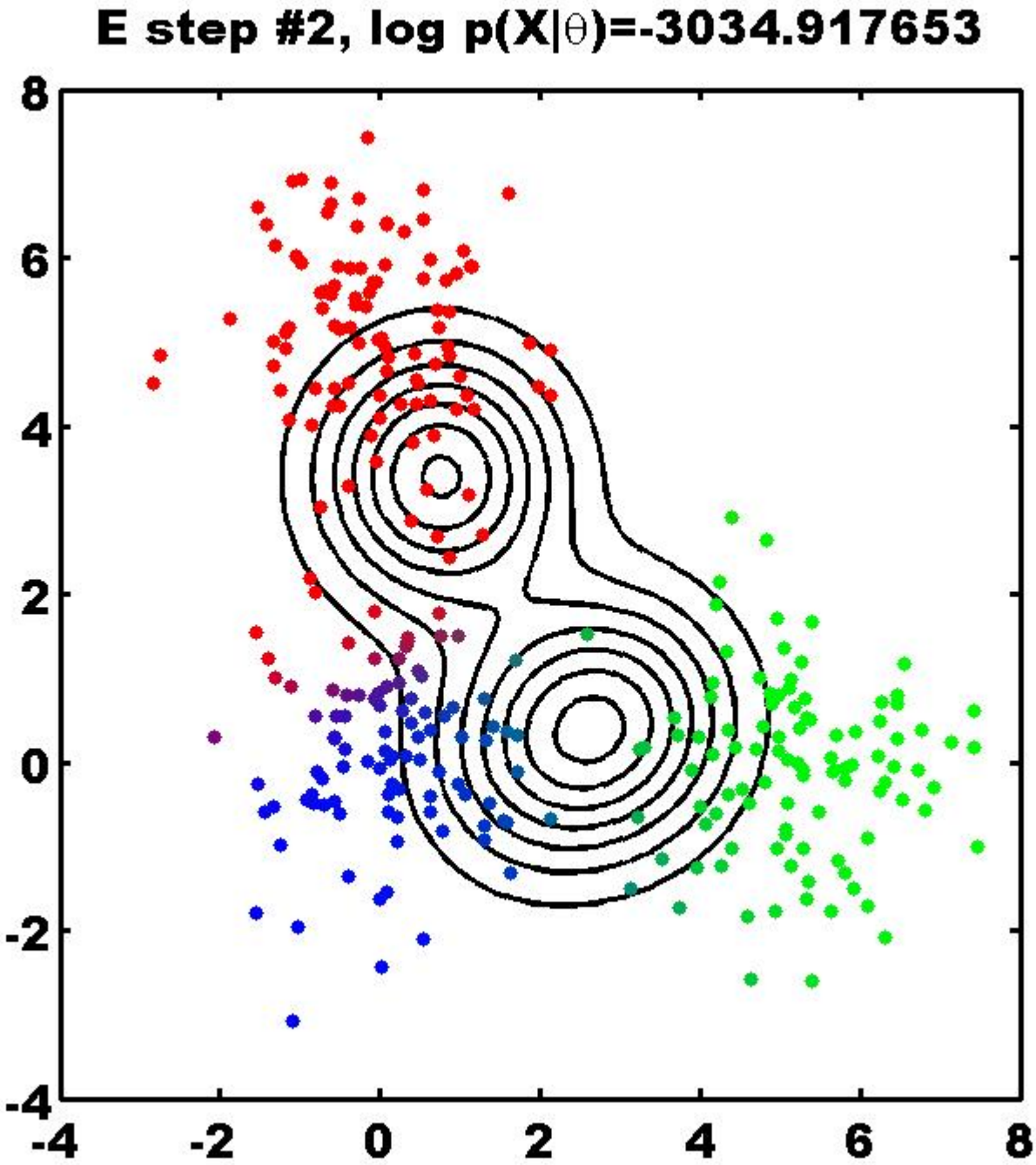
Demo



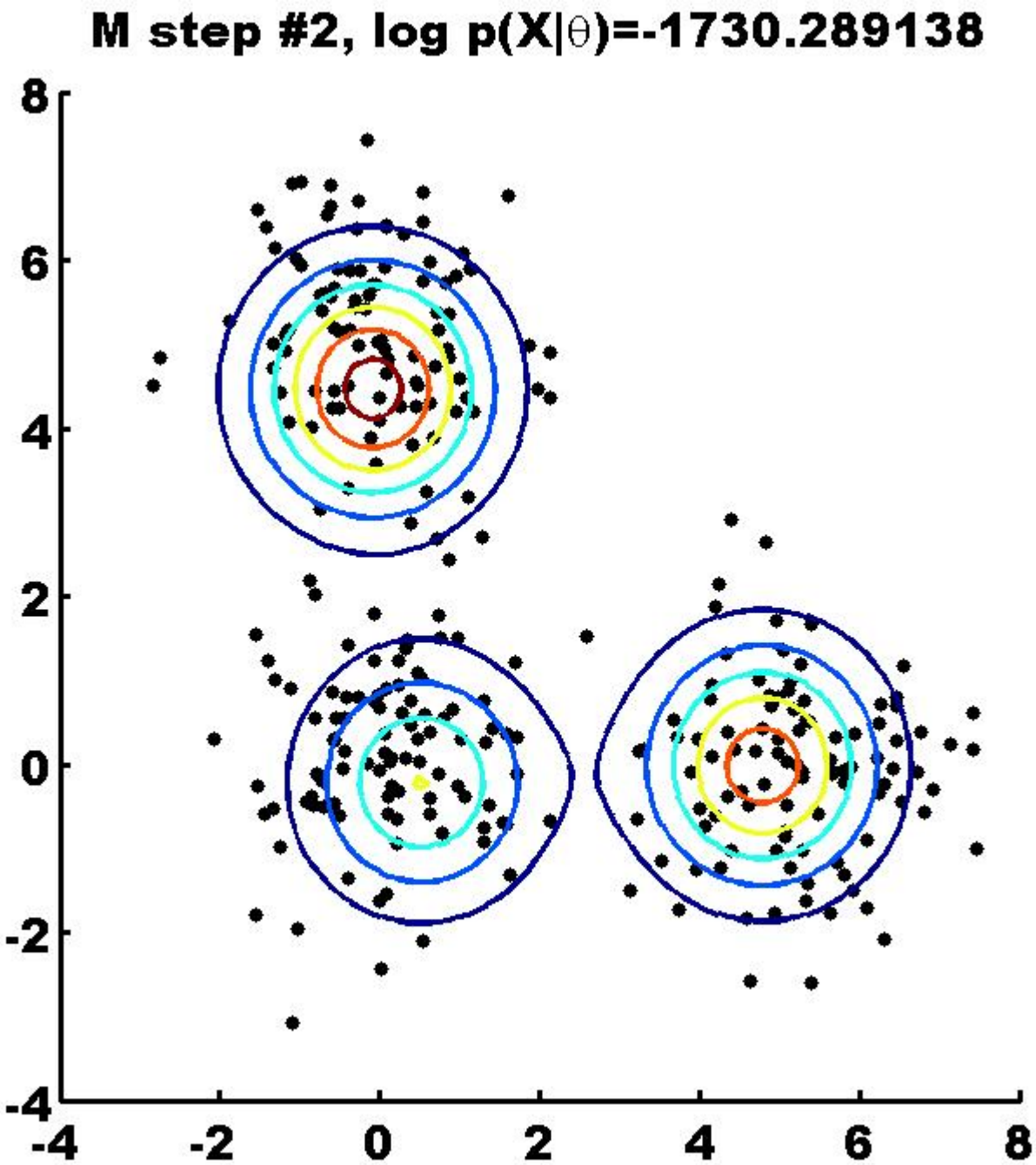
Demo



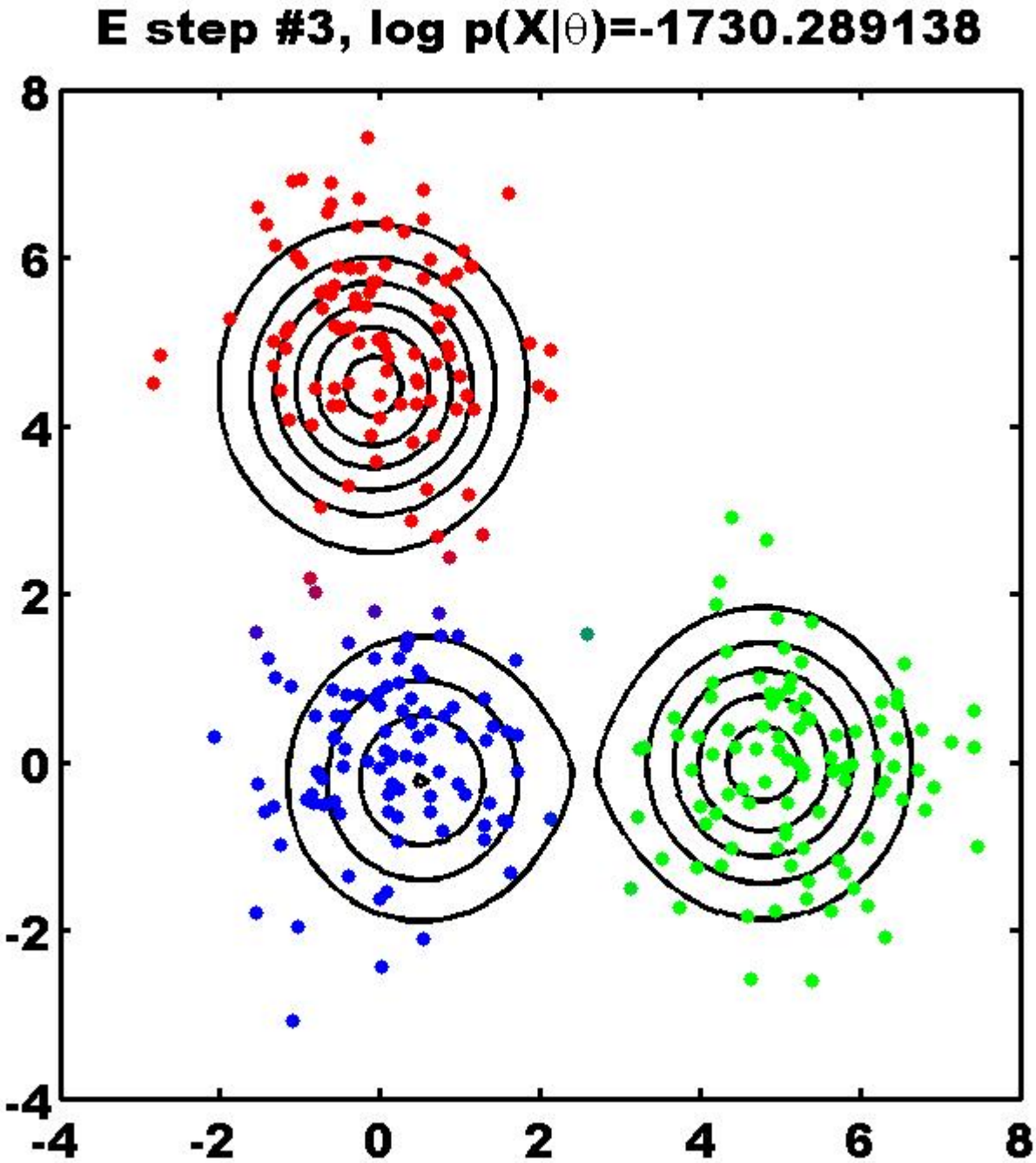
Demo



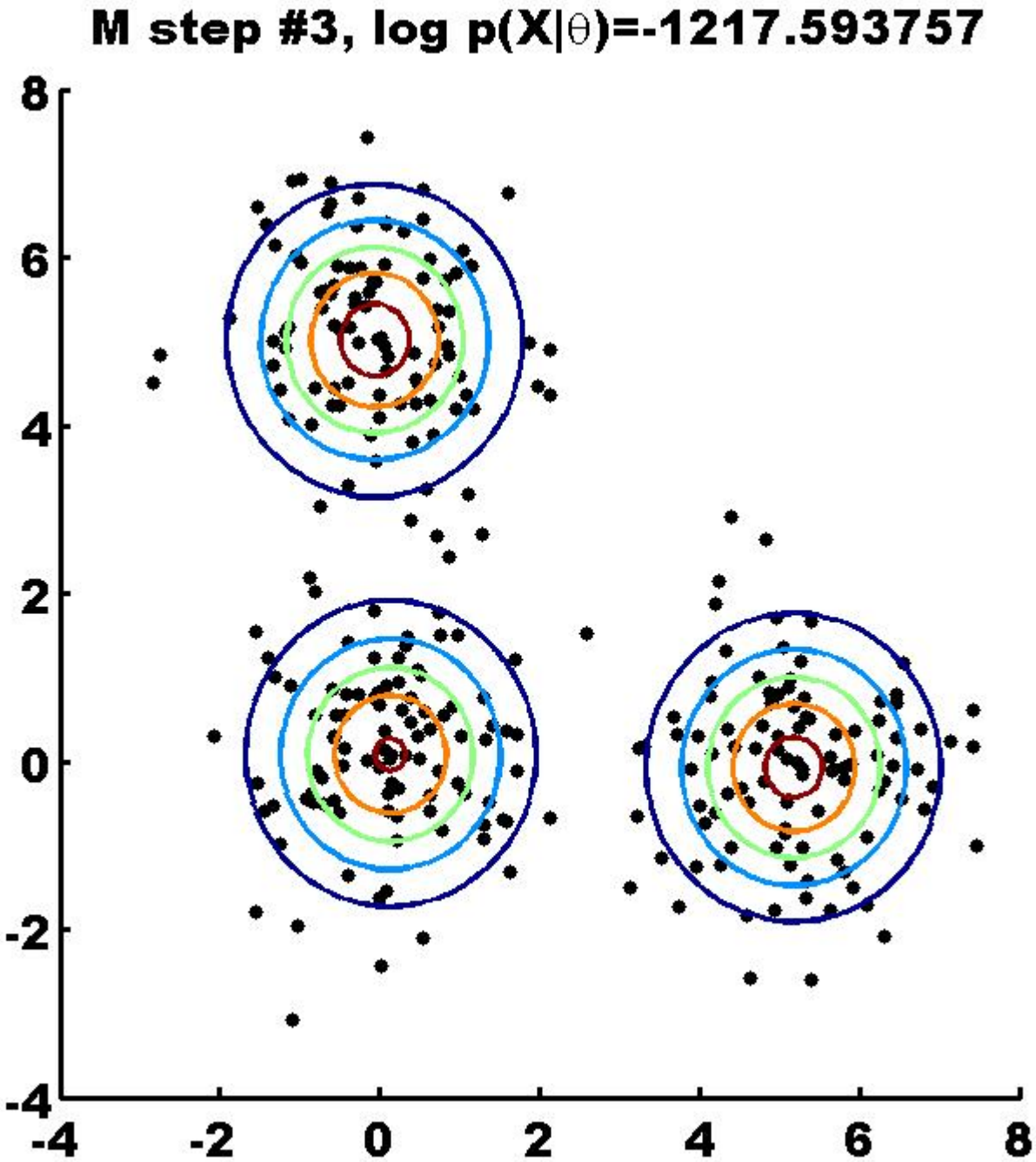
Demo



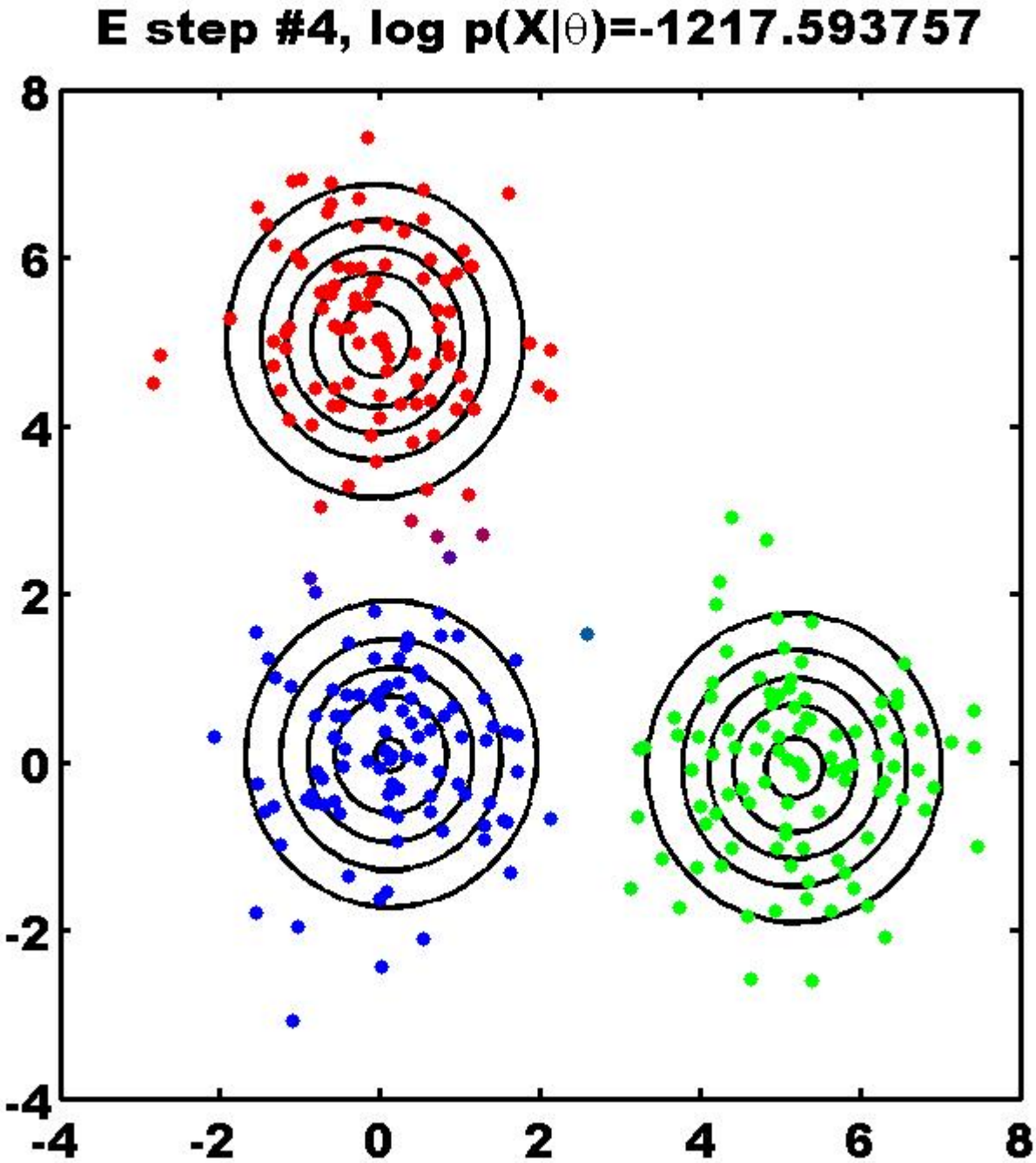
Demo



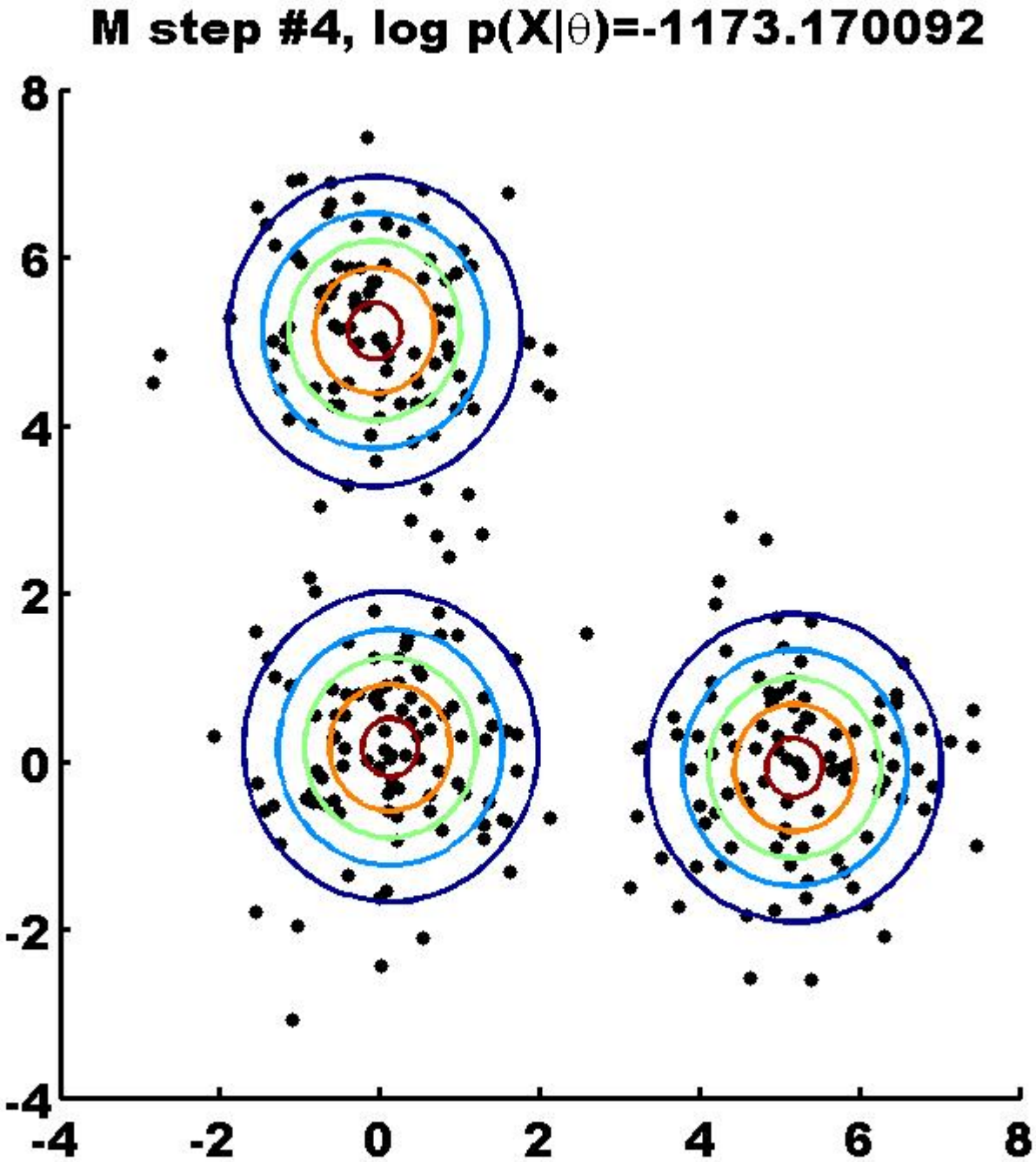
Demo



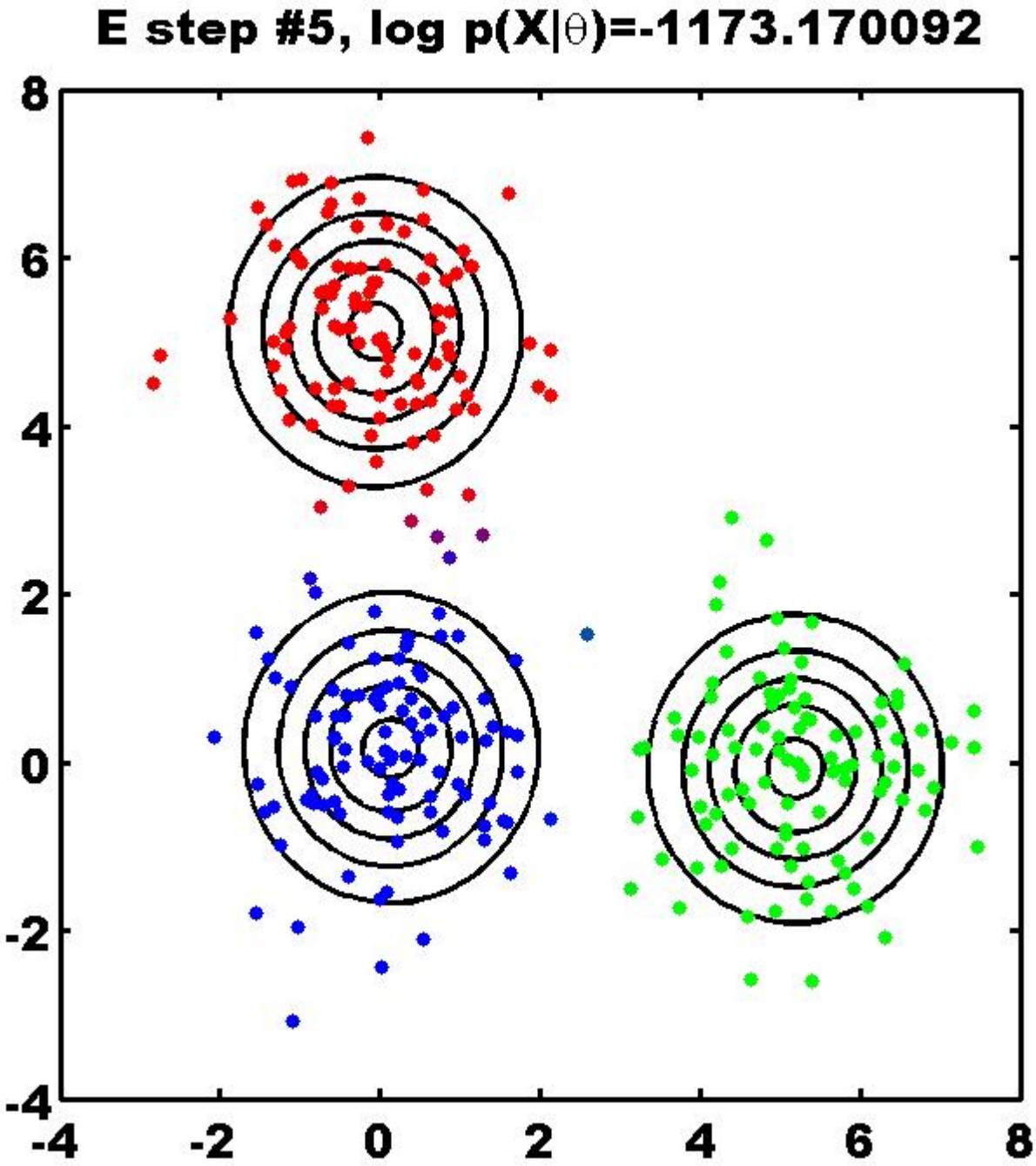
Demo



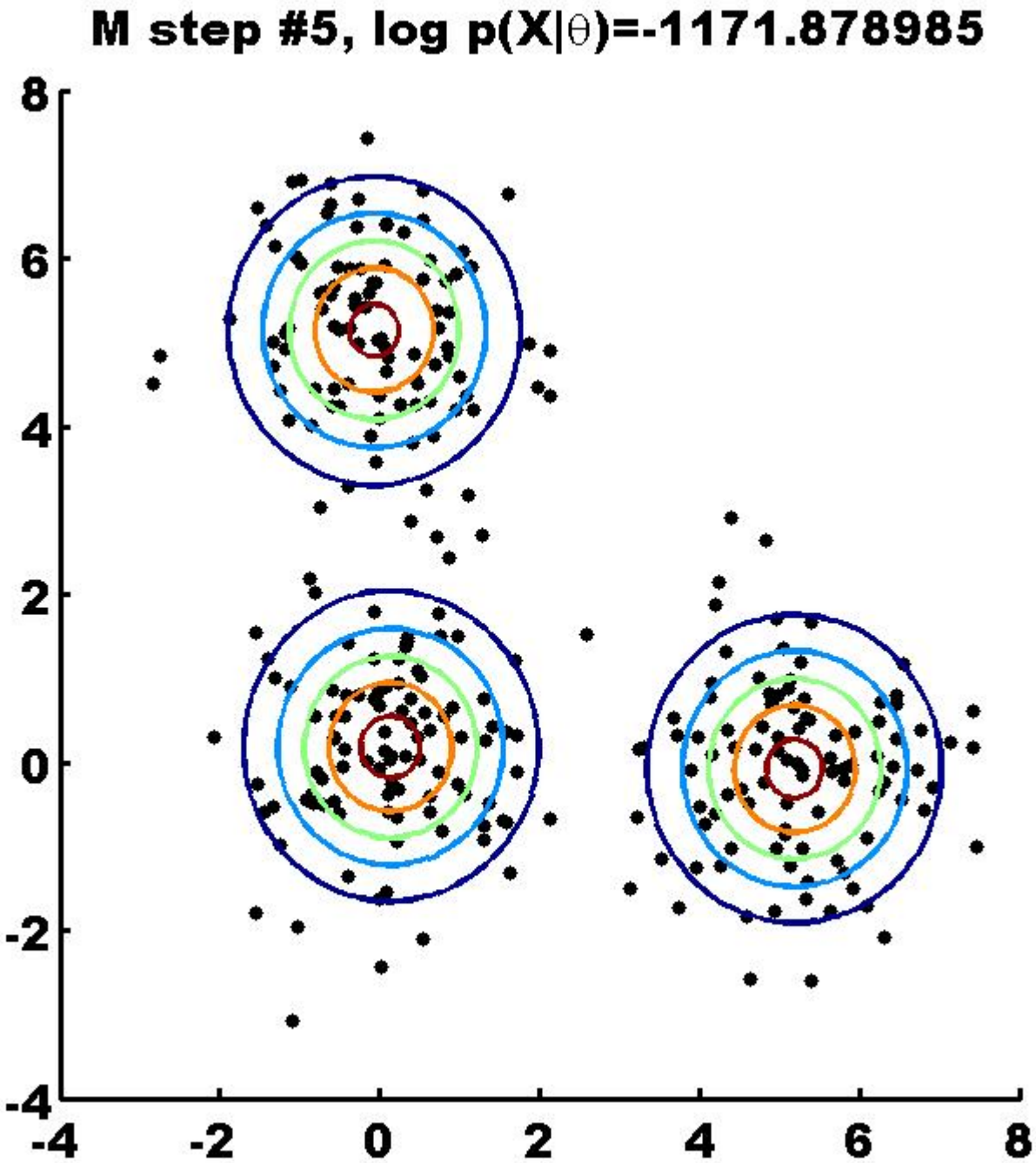
Demo



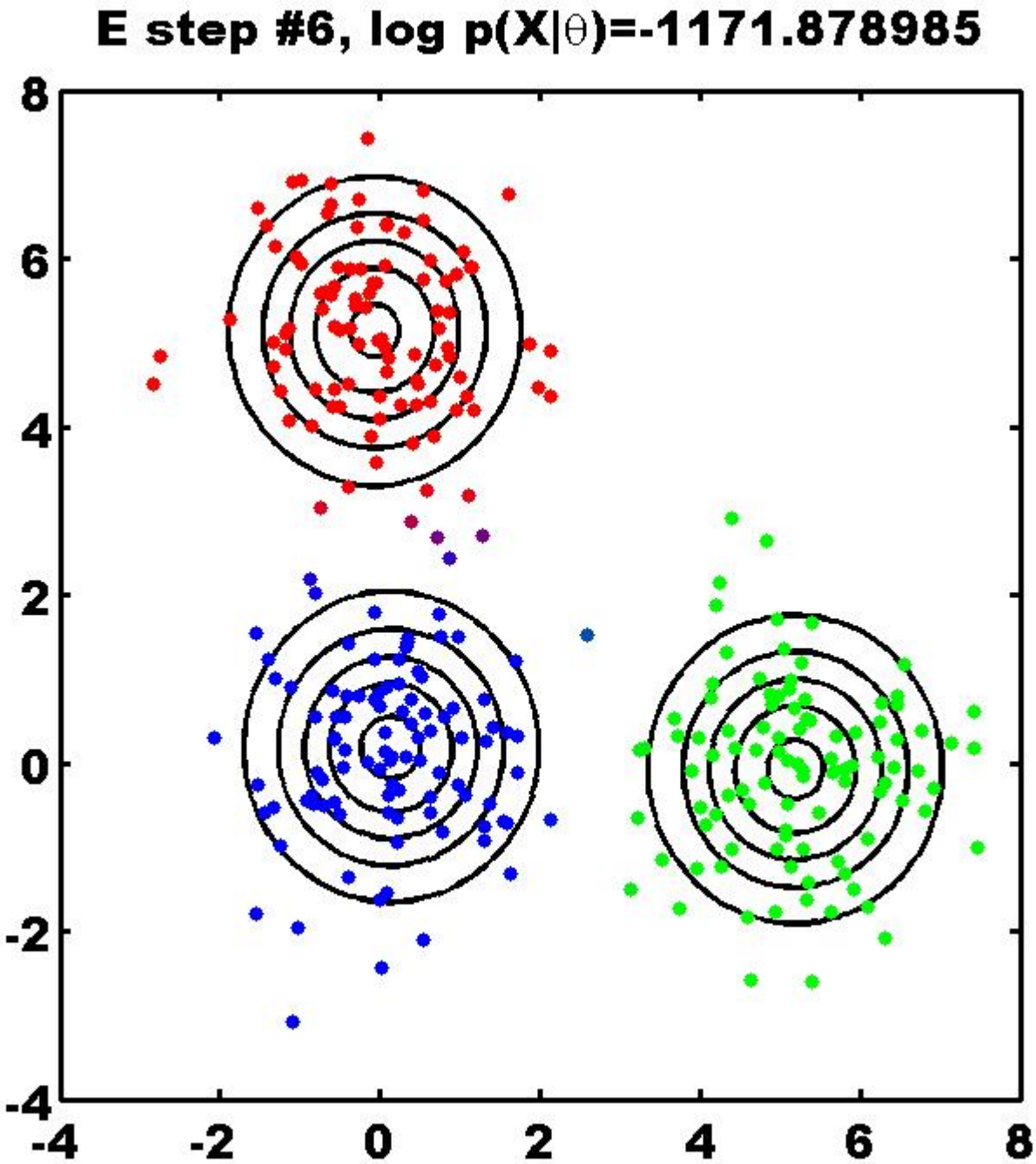
Demo



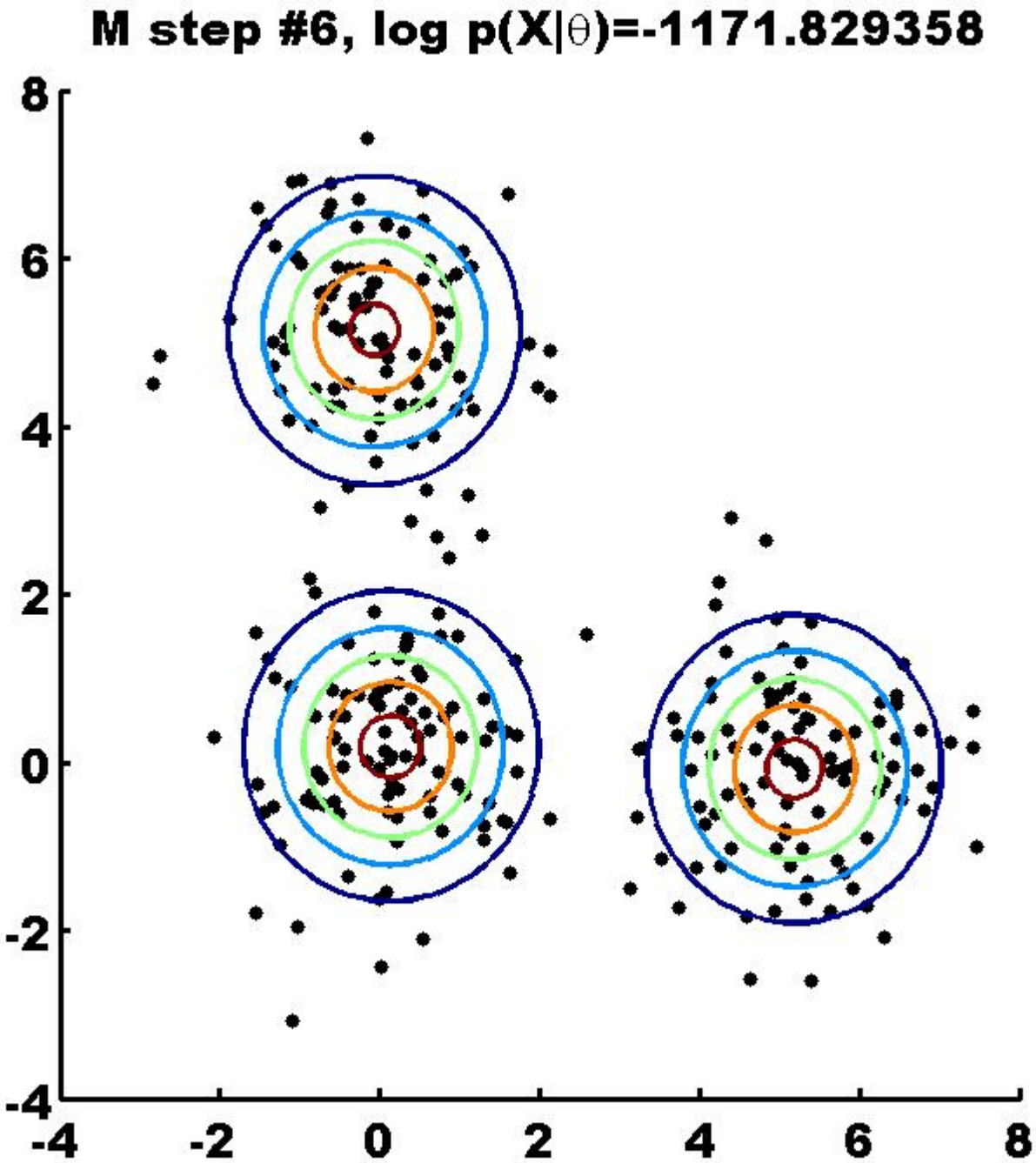
Demo



Demo



Demo



General EM problem

- *How to find maximum likelihood estimates of latent variables in a probabilistic model?*
- Assumptions
  - Latent variables are random variables
  - Complete-data likelihood is known
  - Complete-data likelihood expectation can be maximized
- Requirements
  - Observed data
  - Complete-data likelihood
  - Initial parameters
- EM is not a single algorithm, but a recipe of an algorithm

General EM solution

- Given
  - Observed data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$
  - Complete-data likelihood $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$
  - Initial parameters $\boldsymbol{\theta}^{\text{old}}$

Conditional expectation

$$\mathrm{E}_x[f(x)|y] = \int f(x)p(x|y)\mathrm{d}x$$

- **E step**
  - Compute expected value of complete-data log likelihood $\log\, p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$ over hidden variables ***z***, given parameters $\boldsymbol{\theta}^{\text{old}}$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \mathrm{E}_{\boldsymbol{z}}[\log\, p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) \,|\, \boldsymbol{x}, \boldsymbol{\theta}^{\text{old}}] = \sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}^{\text{old}}) \log\, p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$$

  - Assuming ***z*** is a random variable, $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}^{\text{old}})$
- **M step**
  - Re-estimate parameters $\boldsymbol{\theta}^{\text{new}}$, given values for hidden variables $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$

$$\boldsymbol{\theta}^{\text{new}} = \text{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

- Repeat E and M steps until convergence (and restart with random parameters)

General EM algorithm properties

- Convergence to local maximum guaranteed
  - Each iteration can only increase the observed data likelihood
  - Not necessarily global maximum likelihood
- Benefits
  - Applicable to a wide variety of problems (latent probabilistic models)
  - Faster than Monte Carlo simulations
- Drawbacks
  - Maximum likelihood point estimates are often misleading, e.g., Gaussian singularities
  - Not strictly maximum likelihood (local maxima)
  - Slow convergence for big data
- Bayesian alternatives
  - Monte Carlo simulation
  - Variational approximation

Summary

- Expectation maximization is an **algorithm recipe** for solving **maximum likelihood problems** having **latent variables**
  - Algorithm recipe: family of related algorithms, not a single algorithm
  - Maximum likelihood: point estimate, not fully Bayesian
  - Latent variables: estimate arbitrary hidden parameters, not only Gaussian mixtures
- EM algorithm is iterative and converges to local likelihood maxima
- Full Bayesian alternatives are often computationally heavier
  - Monte Carlo, etc.
- EM for Gaussian mixtures
  - Estimate mixture weights, means, and covariances, given observations
  - Soft version of *k*-means
  - How to choose *k*?

References

- Bilmes, ”A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models.” Tech report, UC Berkeley, CA, USA, 1998.
- Bishop, ”Pattern recognition and machine learning.” p. 430–439, Springer, 2006.