

Expectation Maximization (EM) Algorithm

Philipp Gschöpf

Wolfgang Karl Härdle

Andrija Mihoci

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E. – Center for Applied Statistics
and Economics

Humboldt-Universität zu Berlin

<http://lvb.wiwi.hu-berlin.de/>

<http://case.hu-berlin.de>

<http://irtg1792.hu-berlin.de>



Known Coin Tossing



Figure 1: Ancient coins

Example 1: Tossing coin type known, estimate probabilities



Known Coin Tossing

Example 1: Two coins - two different distributions

- ▶ Probability p_1 or p_2 for "head"
- ▶ The regime (coin type) is known for every toss
- ▶ Maximum Likelihood (ML) for $\theta = (p_1, p_2)^\top$



Unknown Coin Tossing

Example 2: Current regime (coin) is unknown

- ▶ Probability to select the second coin δ
- ▶ Challenge: unobserved indicator variable (latent)
- ▶ Expectation Maximization (EM) for $\theta = (p_1, p_2, \delta)^\top$



Unknown Coin Tossing



Figure 2: Ancient coins

Example 2: Tossing coin type unknown, estimate probabilities



Known Coin Tossing

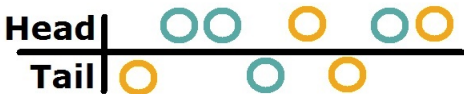


Figure 3: Parameter estimation - Maximum Likelihood (ML)



Unknown Coin Tossing

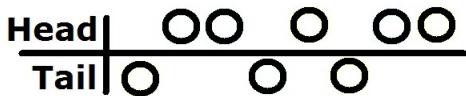


Figure 4: Parameter estimation - Expectation Maximization (EM)



EM Algorithm

- Maximum likelihood estimates under incomplete data
 - ▶ Hartley and Rao (1967)
 - ▶ Dempster et al. (1977)
- Applications
 - ▶ Missing data, grouping, censoring and truncation
 - ▶ Finite mixtures

Hermann Otto Hirschfeld on BBI:



Outline

1. Motivation ✓
2. EM for a Mixture
3. EM Algorithm
4. Conclusions



Mixtures

- Two sequences of (dependent) observations
 - ▶ Latent $X_i, i = 1, \dots, n$
 - ▶ Observable $Y_i, i = 1, \dots, n$

- Linear mixture: Combination of distributions (components)

$$Y = (1 - X)Z_0 + XZ_1, \quad X \in \{0, 1\}, P(X = 1) = \delta$$

with pdfs $f_{Z_j}(\bullet|\theta_j)$ and parameter $\theta_j, j \in \{0, 1\}$



Mixtures

Coin example

- X_i selects one coin:

$$X_i = \begin{cases} 1, & \text{first coin with probability } \delta \\ 0, & \text{second coin} \end{cases}$$

- $Y_i | X_i$ is the observed result

$$Y_i | X_i = \begin{cases} 1, & \text{heads with probability } p_1 \text{ or } p_2 \\ 0, & \text{tails} \end{cases}$$



Mixtures

- Parameter of interest $\theta = (\theta_1, \theta_2, \delta)^\top$
 - ▶ Component parameter, e.g. $\theta_1 = p_1, \theta_2 = p_2$
 - ▶ Probability $\delta = P(X = 1)$

- State X selects a component of Y

$$f_{Y|X}(y|X = x, \theta) = \begin{cases} f_{Z_0}(y|\theta_1), & \text{if } x = 0 \\ f_{Z_1}(y|\theta_2), & \text{if } x = 1 \end{cases}$$



Maximum Likelihood for a Mixture

- Marginal density of Y

$$f_Y(y|\theta) = (1 - \delta) f_{Z_0}(y|\theta_1) + \delta f_{Z_1}(y|\theta_2)$$

- Maximum likelihood (ML) using marginal $f_Y(y|\theta)$
 - ▶ Requires observed data
 - ▶ Challenge: the likelihood function has an additive structure

▶ Proof



Maximum Likelihood for a mixture

- Joint density of X, Y

$$f_{XY}(y, x|\theta) = \{(1 - \delta)f_{Z_0}(y|\theta_1)\}^{1-x} \{\delta f_{Z_1}(y|\theta_2)\}^x$$

- Maximum likelihood using joint $f_{XY}(x, y|\theta)$
 - ▶ The likelihood function has a multiplicative structure
 - ▶ State x is unobserved

▶ Proof



EM for a Mixture

- E-Step: Compute the conditional expectation $E[X|Y, \theta]$
 - ▶ Expectation step
 - ▶ Requires parameter θ

- M-Step: Estimate parameter θ
 - ▶ Use $E[X|Y, \theta]$ instead of X in the likelihood
 - ▶ ML for the joint density $f_{XY}(x, y|\theta)$



E-Step

- Conditional expectation $\gamma(\theta, Y) \stackrel{\text{def}}{=} E[X|Y, \theta]$
 - Parameter $\theta = (\theta_1, \theta_2, \delta)^\top$ is required

$$\gamma(\theta, Y = y) = \frac{\delta f_{Z_1}(y|\theta_2)}{(1 - \delta)f_{Z_0}(y|\theta_1) + \delta f_{Z_1}(y|\theta_2)}$$

► Proof



E-Step

□ Employ (arbitrary) initial parameter $\theta^{(0)}$

- 1) Component parameter $\theta_1^{(0)}, \theta_2^{(0)}$
- 2) Mixture weight $\delta^{(0)}$

$$\gamma(\theta^{(0)}, Y = y) = \frac{\delta^{(0)} f_{Z_1}(y | \theta_2^{(0)})}{(1 - \delta^{(0)}) f_{Z_0}(y | \theta_1^{(0)}) + \delta^{(0)} f_{Z_1}(y | \theta_2^{(0)})}$$



M-Step

- Maximize log-likelihood $\ell(\theta) = \sum_{i=1}^n \log f_{XY}(x_i, y_i | \theta)$

$$\theta^{(1)} = \arg \max_{\theta} \ell \left\{ \theta | x = \gamma(\theta^{(0)}, y_i), y \right\}$$

- Mixture weight δ estimate

$$\delta^{(1)} = n^{-1} \sum_{i=1}^n \gamma(\theta^{(0)}, y_i)$$

► Proof



Iteration

- Iteration of the E- and M-steps

- ▶ Step 1: $\theta^{(1)}$

- ▶ Step 2: $\theta^{(2)}$

- ...

- ▶ Step k : $\theta^{(k)}$

- Repetition of the steps until convergence



EM algorithm - Example

Mixture of normals, e.g. Gentle et al. (2004)

- Mixture with two $N(\mu_j, 1)$ components
- Parameter $\theta = (\mu_1, \mu_2, \delta)^\top$, $\theta_1 = \mu_1$, $\theta_2 = \mu_2$

$$f_{Z_j}(y|\theta_j) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(y - \mu_j)^2}{2} \right\}, \quad j \in \{1, 2\}$$



EM algorithm - Example

Mixture of normals

- Maximum likelihood using the joint density

$$f_{XY}(x, y|\theta) = \sum_{j=0}^1 \mathbf{I}\{x = j\} f_{Z_j}(y|\mu_{j+1})$$

- Component mean

$$\mu_1^{(1)} = \frac{\sum_{i=1}^n \{y_i - y_i \gamma(\theta^{(0)}, y_i)\}}{\sum_{i=1}^n \{1 - \gamma(\theta^{(0)}, y_i)\}}, \quad \mu_2^{(1)} = \frac{\sum_{i=1}^n y_i \gamma(\theta^{(0)}, y_i)}{\sum_{i=1}^n \gamma(\theta^{(0)}, y_i)}$$



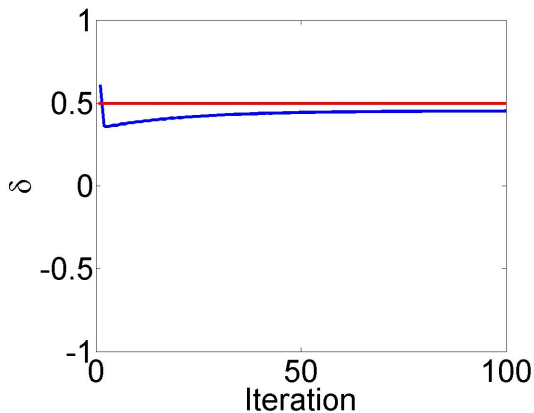


Figure 5: Parameter convergence example, true value in red, $n = 250$

 EM_Normal



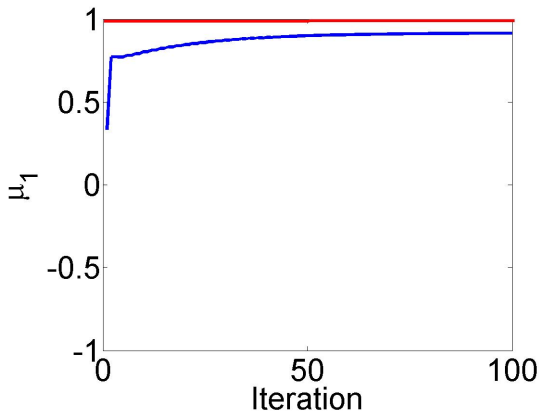


Figure 6: Parameter convergence example, true value in red, $n = 250$

 EM_Normal



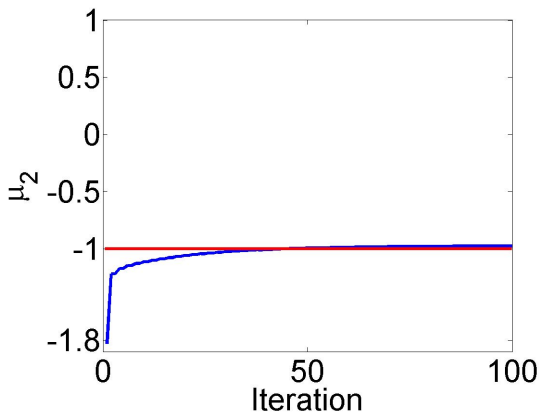


Figure 7: Parameter convergence, true value in red, $n = 250$

► Example 2

► Example 3

► Example 4

 Q-EM_Normal



The general Algorithm

- Blimes (1998): Maximum likelihood estimation
 - ▶ Infeasible likelihood, simplifies with hidden parameter
 - ▶ Hidden or missing values

- Complete data log-likelihood

$$\ell(\theta) = \sum_{i=1}^n \log \{f_{XY}(x_i, y_i | \theta)\}$$



EM algorithm - E-Step

- Expectation step, general algorithm
 - ▶ Expectation of the log-likelihood ℓ
 - ▶ Gentle et al. (2012)

$$\begin{aligned} Q(\theta \mid \theta^{(k)}) &= E_X \left[\ell(\theta) \mid Y, \theta^{(k)} \right] \\ &= \int_{z \in \mathfrak{X}} \ell(\theta) f_{X|Y}(z \mid Y = y, \theta^{(k)}) dz, \\ &\quad \text{with } \mathfrak{X} \text{ the support of } X \end{aligned}$$



EM algorithm - E-Step

Example: Linear mixtures

$$\ell(\theta) = \sum_{i=1}^n \log [\mathbf{I}\{x_i = 0\}(1 - \delta)f_{Z_0}(y_i|\theta_1) + \mathbf{I}\{x_i = 1\}\delta f_{Z_1}(y_i|\theta_2)]$$

- The joint likelihood is a linear function of x
- E-Step: Expectation of X



EM algorithm - M-Step

- Maximization: Estimate the parameter of interest θ
 - ▶ Maximization of $Q(\theta|\theta^{(0)})$ w.r.t. θ
 - ▶ Updated (optimal) estimate $\theta^{(1)}$

$$\theta^{(1)} = \arg \max_{\theta} Q(\theta | \theta^{(0)})$$

► Properties

S. Kullback and R. Leibler on BBI:



Conclusions

- EM Algorithm
 - ▶ Finding parameter estimates
 - ▶ Latent variables, missing data

- Application
 - ▶ Linear Mixtures
 - ▶ Normal mixture



Expectation Maximization (EM) Algorithm

Philipp Gschöpf

Wolfgang Karl Härdle

Andrija Mihoci

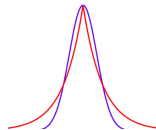
Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics

Humboldt-Universität zu Berlin

<http://lvb.wiwi.hu-berlin.de/>

<http://case.hu-berlin.de>

<http://irtg1792.hu-berlin.de>



References



Blimes, J. A.

A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models

International Computer Science Institute, 1998,
Electronic publication



Barber, D.

Bayesian Reasoning and Machine Learning

Cambridge University Press, 2012, ISIN: 978-0-521-51814-7



References



Dempster, A. P., Laird, N. M. and Rubin, D. B.
Maximum Likelihood from Incomplete Data via the EM Algorithm

Journal of the Royal Statistical Society. Series B
(Methodological), 1977, **39** (1)



Ng, S. K., Krishnan, T. and McLachlan, G. J.
The EM Algorithm
Handbook of Computational Statistics - Concepts and Methods
Edts: Gentle, J. E., Härdle, W. K. and Mori, Y.
Springer Verlag, Heidelberg, **2**: 139–173, 2012



References



Hamilton, J. D.

Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates
Journal of Economic Dynamics and Control, **12**(2): 285–423,
1988



Hasselblad, V.

Estimation of parameters for a mixture of normal distributions
Technometrics, **8**: 431–444, 1966



Härdle, W.K., Okhrin, O. and Wang, W.

HMM and HAC

Advances in Intelligent Systems and Computing **190**: 341–348,
2013



References



Hastie, T., Tibshiriani, R. and Friedman, J.

The Elements of Statistical Learning

Springer Series in Statistics 2, 2008



Hartley, H.O. and Rao, J.N.K.

Maximum-likelihood estimation for the mixed analysis of variance model

Biometrika **54**(1-2): 93–108, 1967



Huber, P.J.

Robust Estimation of a Location Parameter

The Annals of Mathematical Statistics **35**(1): 73–101, 1964



Distributions

- Marginal density of X and $Y|X$ with $\theta = (\theta_1, \theta_2, \delta)^\top$ are given

$$f_X(x|\theta) = \mathbb{I}\{x = 1\}\delta + \mathbb{I}\{x = 0\}(1 - \delta) = \delta^x(1 - \delta)^{1-x}$$

$$f_{Y|X}(y|X = x, \theta) = f_{Z_0}(y|\theta_1)^{1-x} f_{Z_1}(y|\theta_2)^x$$

- Marginal of Y applying the law of total probability

$$f_Y(y|\theta) = \sum_{j=0}^1 P(X = j) f_{Y|X}(y|X = j, \theta)$$

$$f_Y(y|\theta) = (1 - \delta) f_{Z_0}(y|\theta_1) + \delta f_{Z_1}(y|\theta_2)$$



Distributions

- Marginal density of X and $Y|X$ with $\theta = (\theta_1, \theta_2, \delta)^\top$ are given

$$f_X(x|\delta) = \mathbb{I}\{x = 1\}\delta + \mathbb{I}\{x = 0\}(1 - \delta) = \delta^x(1 - \delta)^{1-x}$$

$$f_{Y|X}(y|X = x, \theta) = f_{Z_0}(y|\theta_1)^{1-x} f_{Z_1}(y|\theta_2)^x$$

- Joint distribution of X and Y

$$f_{XY}(x, y|\theta) = f_{Y|X}(y|X = x, \theta) f_X(x|\theta)$$

$$f_{XY}(x, y|\theta) = \{(1 - \delta)f_{Z_0}(y|\theta_1)\}^{1-x} \{\delta f_{Z_1}(y|\theta_2)\}^x$$



Conditional Expectation

- Conditional distribution of $X|Y$ via Bayes rule

$$f_{X|Y}(x|Y=y, \theta) = \frac{f_{XY}(x, y|\theta)}{f_Y(y|\theta)}$$

- Resulting distribution of $X|Y$

$$f_{X|Y}(x|Y=y, \theta) = \frac{\{(1-\delta)f_{Z_0}(y|\theta_1)\}^{1-x} \{\delta f_{Z_1}(y|\theta_2)\}^x}{(1-\delta)f_{Z_0}(y|\theta_1) + \delta f_{Z_1}(y|\theta_2)}$$

► E-Step



Conditional Expectation

- Since X is binomial

$$\gamma(Y, \theta) = E[X|Y, \theta] = P(X = 1|Y, \theta)$$

- Expectation of the unobserved variable

$$E[X|Y, \theta] = \frac{\delta f_{Z_1}(y|\theta_2)}{(1 - \delta)f_{Z_0}(y|\theta_1) + \delta f_{Z_1}(y|\theta_2)}$$

► E-Step



M-Step for mixture probabilities

- Maximize the expectation of the likelihood Q
- M components, following Blimes (1998)
 - ▶ The sum of component probabilities must be equal to 1

$$\arg \max_{\theta} Q(\theta \mid \theta^{(k)}) \quad \text{s.t.} \quad \sum_{j=1}^M \delta_j = 1$$

▶ M-Step



M-Step for mixture probabilities

- Optimization w.r.t. δ_j , Lagrange parameter λ

$$\frac{1}{\delta_j} \sum_{i=1}^n f_X(j|y_i, \theta^{(k)}) = \lambda$$

$\lambda = n$ completes the proof

$$\sum_{i=1}^n \sum_{j=1}^M \frac{1}{\delta_j} f_X(j|y_i, \theta^{(k)}) = M\lambda$$

► M-Step



Properties

- The likelihood of Y has a lower bound, e.g. Hastie (2008)
 - ▶ EM maximizes the bound

$$\mathbb{E}_{f_X(x|\theta)} [\log f_{XY}(x, y | \theta) - \log f_X(x|\theta)] \leq \log\{f_Y(y|\theta)\}$$

▶ M-Step



Lower bound of the marginal likelihood

Kullback-Leibler divergence of a variational distribution $\tilde{f}_x(x_i|\theta)$ and the parametric model $f_X(x|y, \theta)$, e.g., Barber (2012):

$$KL \left\{ \tilde{f}_X(x|\theta) || f_X(x|y, \theta) \right\} \geq 0$$
$$\mathbb{E}_{\tilde{f}_X(x|\theta)} \left[\log \{ \tilde{f}_X(x|\theta) \} - \log \left\{ \frac{f_{XY}(x, y|\theta)}{f_Y(y|\theta)} \right\} \right] \geq 0$$

► M-Step



Kullback-Leibler Divergence

- Also known as relative entropy
 - ▶ Difference measure of distributions P and Q
 - ▶ Not symmetric, not a metric

$$KL(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \geq 0$$

▶ M-Step



Kullback-Leibler Divergence

The Kullback-Leibler divergence is positive

$$\log(z) \leq z - 1$$
$$q(x) \log \frac{p(x)}{q(x)} \leq p(x) - q(x)$$

$p(x)$ and $q(x)$ are densities, thus

$$\int_{-\infty}^{\infty} \{p(x) - q(x)\} dx = 1 - 1 = 0$$

► M-Step



Convergence Example 2: $\delta = 0.5, \mu_1 = 0.5, \mu_2 = -0.5, n = 250$

- True μ_1 and μ_2 closer together
- Components harder to "disentangle"

▶ Back

 EM_Normal



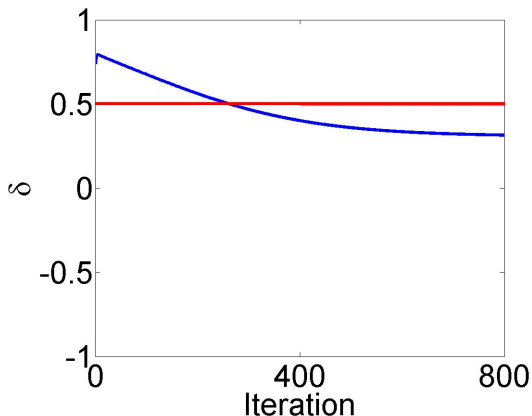
Convergence Example 2: $\delta = 0.5, \mu_1 = 0.5, \mu_2 = -0.5, n = 250$ 

Figure 8: Parameter convergence example, $n = 250$, true value in red

[▶ Back](#) EM_Normal

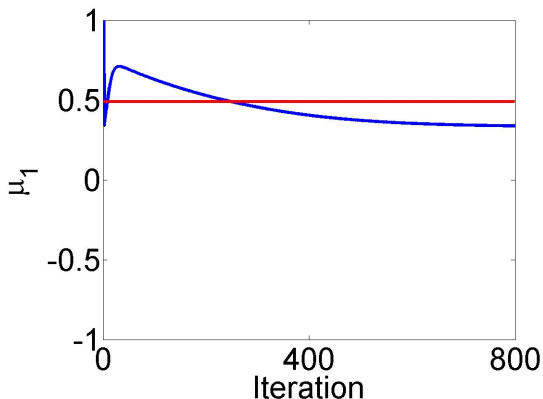
Convergence Example 2: $\delta = 0.5, \mu_1 = 0.5, \mu_2 = -0.5, n = 250$ 

Figure 9: Parameter convergence example, $n = 250$, true value in red

[▶ Back](#) EM_Normal

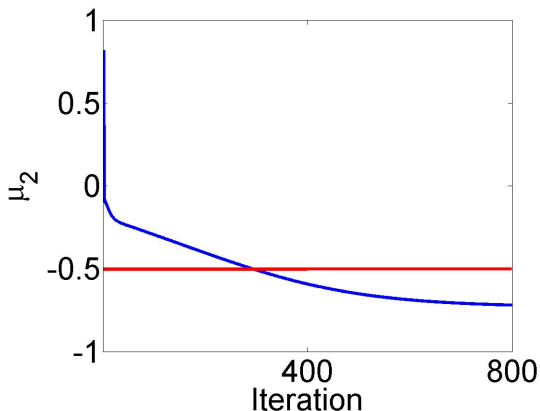
Convergence Example 2: $\delta = 0.5, \mu_1 = 0.5, \mu_2 = -0.5, n = 250$ 

Figure 10: Parameter convergence example, $n = 250$, true value in red

[▶ Back](#) EM_Normal

Convergence Example 3:

$\delta = 0.5, \mu_1 = 0.25, \mu_2 = -0.25, n = 250$

- True μ_1 and μ_2 even closer together
- Components harder to "disentangle"

[▶ Back](#)

EM_Normal



Convergence Ex. 3: $\delta = 0.5, \mu_1 = 0.25, \mu_2 = -0.25, n = 250$

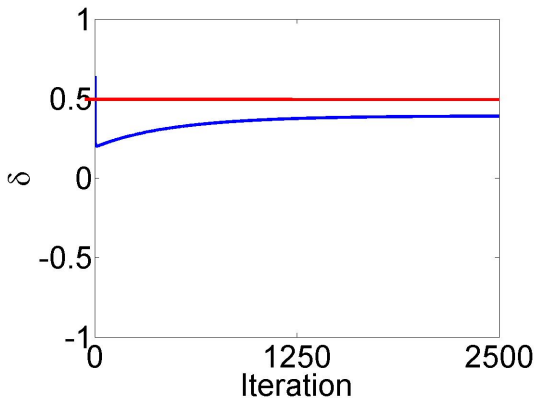


Figure 11: Parameter convergence example, $n = 250$, true value in red

► Back

 EM_Normal



Convergence Ex. 3: $\delta = 0.5, \mu_1 = 0.25, \mu_2 = -0.25, n = 250$

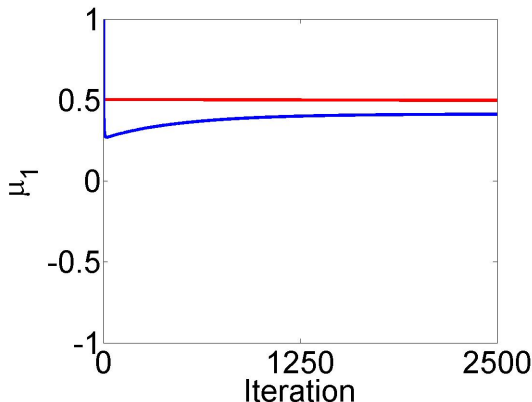


Figure 12: Parameter convergence example, $n = 250$, true value in red

► Back

EM_Normal



Convergence Ex. 3: $\delta = 0.5, \mu_1 = 0.25, \mu_2 = -0.25, n = 250$

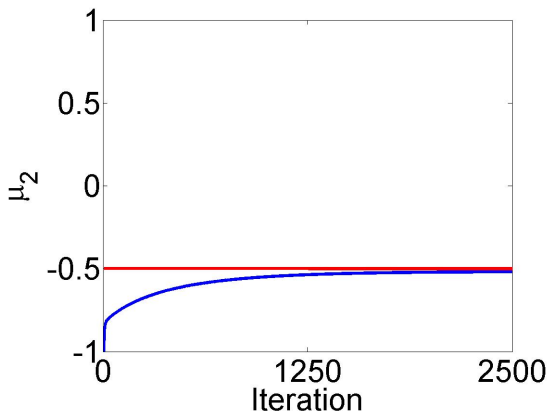


Figure 13: Parameter convergence example, true value in red

► Back

EM_Normal



Convergence Example 2: $\delta = 0.9, \mu_1 = 0.5, \mu_2 = -0.5, n = 250$

- ▣ $\delta = 0.9$, low probability of first component
- ▣ Few observations from first component

► Back

 EM_Normal



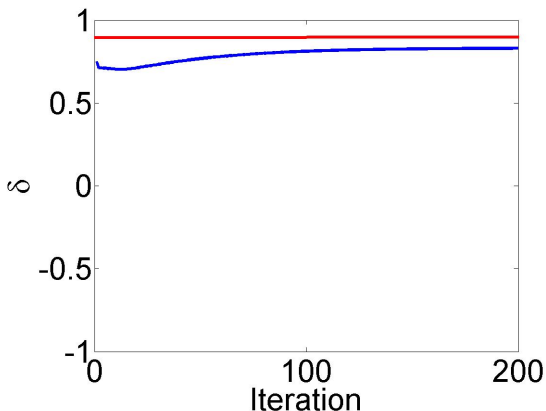
Convergence Example 4: $\delta = 0.9, \mu_1 = 1, \mu_2 = -1, n = 250$ 

Figure 14: Parameter convergence example, true value in red

► Back

EM_Normal



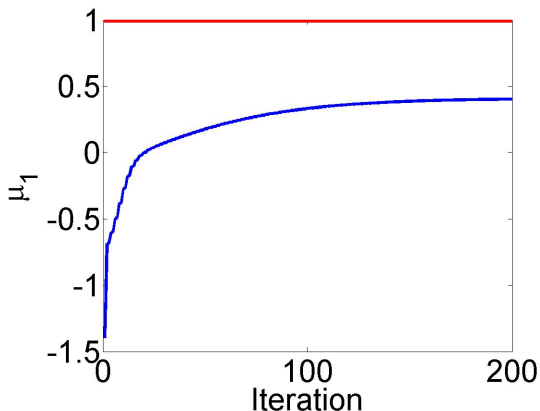
Convergence Example 4: $\delta = 0.9, \mu_1 = 1, \mu_2 = -1, n = 250$ 

Figure 15: Parameter convergence example, true value in red

[▶ Back](#) EM_Normal

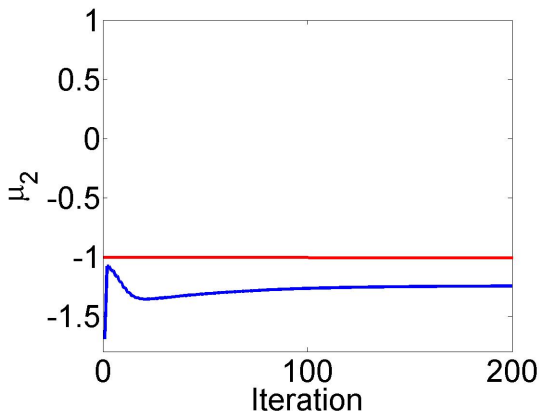
Convergence Example 4: $\delta = 0.9, \mu_1 = 1, \mu_2 = -1, n = 250$ 

Figure 16: Parameter convergence example, true value in red

[▶ Back](#) EM_Normal