# Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

Department of mathematics
KTH Royal Institute of Technology
johawes@kth.se

Lecture 15
The EM algorithm
12 May 2017

# Plan of today's lecture

1. MC methods for hypothesis testing (Ch. 8)
2. The expectation-maximisation (EM) algorithm
   - Missing data problems
   - The algorithm
   - Some theory
   - A Monte Carlo EM implementation

# Outline

# Statistical hypotheses

- A statistical hypothesis is a statement about the distributional properties of data.
- The goal of a hypothesis test is to see if data agrees with the statistical hypothesis.
- Rejection of a hypothesis indicates that there is sufficient evidence in the data to make the hypothesis unlikely.
- Strictly speaking, a hypothesis test does not accept a hypothesis; it fails to reject it.

# Testing hypotheses

- The basis of a hypothesis test consist of
  - a null hypothesis $\mathcal{H}_0$ that we wish to test.
  - a test statistic $t(y)$, i.e., a function of the observed data $y$.
  - a critical region R.
- If the test statistic falls into the critical region, then we reject the null hypothesis $\mathcal{H}_0$.

# Important concepts

**Significance** The probability (risk) that the test incorrectly rejects the null hypothesis. Also called the level of the test (not. $\alpha$).

**Power** The probability that the test rejects correctly the null hypothesis. Is a function of the true parameter.

***p*-value** The probability, given the null hypothesis, of observing a result at least as extreme as the test statistic.

**Type I error** Incorrectly rejecting the null hypothesis.

**Type II error** Failing to reject the null hypothesis.

# Testing simple hypotheses

- A simple hypothesis specifies completely a single distribution for the data, e.g., $Y \sim \mathrm{N}(\theta, 1)$ with $\mathcal{H}_0 : \theta = 0$.
- We construct/define a test statistic $t(y)$ such that large values of $t(y)$ indicate evidence against $\mathcal{H}_0$.
- The $p$-value of the test is now $p(y) = \mathbb{P}(t(Y) \geq t(y) \| \mathcal{H}_0)$.
- The critical region is $\mathrm{R} = \{y : p(y) \leq \alpha\}$, where $\alpha$ is the level of the test.
- Thus, to evaluate the $p$-value we need to be able to compute probabilities under the distribution of $t(Y)$ under $\mathcal{H}_0$.
- This can be tricky if this distribution is complex. Use MC!

# MC tests of a simple hypothesis

- An MC-based algorithm for testing simple hypotheses goes as follows:
  1. Draw $N$ samples, $Y_1, \ldots, Y_N$, from the distribution specified by $\mathcal{H}_0$.
  2. Calculate the test statistic $t_i = t(Y_i)$ for each sample.
  3. Estimate the $p$-value using MC integration by letting

$$\widehat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{t_i \geq t(y)\}}.$$

  4. If $\widehat{p}(y) \leq \alpha$, reject $\mathcal{H}_0$.

# Permutation tests

- A random vector $Y = (Y_1, \ldots, Y_n)$ is said to be exchangeable if $Y = (Y_{I_1}, \ldots, Y_{I_n})$ has the same distribution for all all permutations $\{I_1, \ldots, I_n\}$ of $\{1, \ldots, n\}$.
- In this case, the conditional distribution of $Y$ given the ordered sample is the uniform distribution on the set of all permutations of $Y$.
- Conditioning on the ordered variables leads to permutation tests.
- Permutation tests can be very efficient in testing an exchangeable null-hypothesis against a non-exchangeable alternative, e.g., for testing if two samples differ in some way.

# MC permutation test

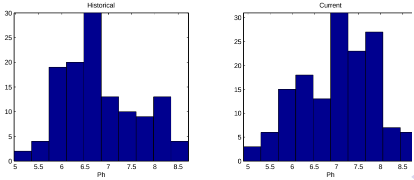
- An MC-based permutation test can be implemented as follows.
  1. Draw $N$ permutations, $Y_1, \ldots, Y_N$, of the vector $y$.
  2. Calculate the test statistic $t_i = t(Y_i)$ for each permutation.
  3. Estimate the $p$-value using MC integration according to

$$\widehat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{t_i \geq t(y)\}}.$$

  4. If $\widehat{p}(y) \leq \alpha$, reject $\mathcal{H}_0$.

# Example: pH data

- We have 273 historical and current pH-measurements $y = (y_1, \dots, y_{273})$ [$(y_1, \dots, y_{124})$ are historical and $(y_{125}, \dots, y_{273})$ are current] of 149 lakes in Wisconsin and want to test if the pH-levels have increased.
- We assume that all measurements are independent and that historical measurements have a distribution $F_0$ and that new measurements have a distribution $G_0$.
- We want to test $\mathcal{H}_0 : F_0 = G_0$ against $\mathcal{H}_1 : F_0 \neq G_0$.



# Example: pH data (cont.)

- Assume that the distribution of the current data can be written as $G_0(y) = F_0(y - \theta)$, i.e, the mean of the current data is the mean of the historical data plus $\theta$.
- We now want to test $\mathcal{H}_0 : \theta = 0$ against $\mathcal{H}_1 : \theta > 0$.
- Under $\mathcal{H}_0$, all data are i.i.d. and thus exchangeable.
- Naturally, we use the difference

$$t(y) = \frac{1}{148} \sum_{i=125}^{273} y_i - \frac{1}{124} \sum_{i=1}^{124} y_i$$

  as a test statistic.
- A permutation test gives $p = .0198$ and we reject $\mathcal{H}_0$ on the level .05.

# Outline

# Outline

# Data augmentation

- We suppose that
  - we are interested in some model parameter $\theta$, but likelihood inference based solely on the observed, but somehow “incomplete”, data $Y$ is intractable.
  - there exists some latent variable $X$ which is not observed, but if observed would make the estimation problem relatively simple.
- The pair $(X, Y)$ is known as the complete data, whereas $Y$ is referred to as incomplete data.
- We suppose that the joint distribution of $(X, Y)$ admits, for a given parameter $\theta$, a density $f_\theta(x, y) = f_\theta(y \mid x) f_\theta(x)$.

# Data augmentation

- We suppose that
  - we are interested in some model parameter $\theta$, but likelihood inference based solely on the observed, but somehow "incomplete", data $Y$ is intractable.
  - there exists some latent variable $X$ which is not observed, but if observed would make the estimation problem relatively simple.
- The pair $(X, Y)$ is known as the complete data, whereas $Y$ is referred to as incomplete data.
- We suppose that the joint distribution of $(X, Y)$ admits, for a given parameter $\theta$, a density $f_\theta(x, y) = f_\theta(y \mid x) f_\theta(x)$.

# Maximum likelihood estimation in latent data models

- Recall that given $Y$, the maximum likelihood estimator (MLE) is given by

$$\theta \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{\theta \in \Theta} \ell(\theta),$$

where

$$\ell(\theta) \stackrel{\text{def}}{=} \log f_\theta(Y) = \log \int f_\theta(Y \mid x) f_\theta(x)\, dx$$

is the log-likelihood function.

- Even though the complete data likelihood

$$f_\theta(x, y) = f_\theta(y \mid x) f_\theta(x)$$

has typically a simple form, the integral may prevent closed-form computation of $\ell(\theta)$.

# Example: radioactive emission

- Assume a radioactive material is known to emit Po(100) particles in unit time. A measurement equipment records each particle with probability $\theta$. If the recorded value was $y = 84$, what is the maximum likelihood estimate of $\theta$?
- Here we failed to observe $X$, the total number of particles emitted. Conditionally on $X$, $y$ is an observation of a $\text{Bin}(\theta, X)$ variable.

# Example: radioactive emission

- Here the complete data likelihood

$$
\begin{aligned}
f_\theta(x, y) = f_\theta(y \mid x) f_\theta(x) &= \binom{x}{y} \theta^y (1 - \theta)^{x-y} e^{-100} \frac{100^x}{x!} \\
&\propto \theta^y (1 - \theta)^{x-y}
\end{aligned}
$$

  has a simple, closed-form expression.
- On the contrary, the observed data likelihood

$$
f_\theta(y) = \sum_{x=y}^{\infty} f_\theta(y \mid x) f_\theta(x) = \sum_{x=y}^{\infty} \binom{x}{y} \theta^y (1 - \theta)^{x-y} e^{-100} \frac{100^x}{x!}
$$

  is complex.

# Outline

# The expectation-maximization (EM) algorithm

- Thus, maximizing $\ell(\theta)$ is a complicated task. Nevertheless, for latent data models the problem of computing the MLE can most often be cast efficiently into the framework of the expectation-maximization (EM) algorithm.
- Let $p$ and $q$ be two probability densities on some common state space. The EM algorithm uses the fact that the Kullback-Leibler divergence

$$K(p\|q) \stackrel{\text{def}}{=} \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\,dx \geq 0$$

is always positive and zero only if and only if $p = q$ (for almost all $x$).

# The EM algorithm (cont'd)

- The algorithm goes as follows.
  **Data:** Initial value $\theta_0$
  **Result:** $\{\theta_\ell; \ell \in \mathbb{N}\}$
  **for** $\ell \leftarrow 0, 1, 2, \ldots$ **do**
  set $\mathcal{Q}_{\theta_\ell}(\theta) \leftarrow \mathbb{E}_{\theta_\ell}(\log f_\theta(X, Y) \mid Y)$;
  set $\theta_{\ell+1} \leftarrow \arg\max_{\theta \in \Theta} \mathcal{Q}_{\theta_\ell}(\theta)$
  **end**
- The two steps within the main loop are referred to as expectation (E-) and maximization (M-) steps, respectively.

# Example: radioactive emission (cont.)

- First, in order to execute the E-step we need the conditional distribution

$$f_\theta(x \mid y) \propto f_\theta(y \mid x) f_\theta(x) = \binom{x}{y} \theta^y (1-\theta)^{x-y} e^{-100} \frac{100^x}{x!}$$

$$\propto \frac{x!}{(x-y)!}(1-\theta)^{x-y}\frac{100^x}{x!} \propto \frac{\{100(1-\theta)\}^{x-y}}{(x-y)!}, \quad x \geq y,$$

which means that $(X \mid Y = y) \overset{\text{d.}}{=} W + y$, where $W \sim \text{Po}(100(1-\theta))$.

# Example: radioactive emission (cont.)

- Thus,

$$\log f_\theta(x, y) = y \log \theta + (x - y) \log(1 - \theta) \quad (+\text{const.}),$$

which implies that

$$\begin{aligned} \mathcal{Q}_{\theta_\ell}(\theta) &= \mathbb{E}_{\theta_\ell}(\log f_\theta(X, Y) \mid Y) \\ &= Y \log \theta + \{\mathbb{E}_{\theta_\ell}(X \mid Y) - Y\} \log(1 - \theta) \\ &= Y \log \theta + 100(1 - \theta_\ell) \log(1 - \theta), \end{aligned}$$

as $\mathbb{E}_{\theta_\ell}(X \mid Y) = 100(1 - \theta_\ell) + Y$.

# Example: radioactive emission (cont.)

- Letting $\theta_{\ell+1}$ be the maximum of $\mathcal{Q}_{\theta_\ell}(\theta)$ with respect to $\theta$ yields the updating formula

$$\theta_{\ell+1} = \frac{Y}{Y + 100(1 - \theta_\ell)}.$$

- In MATLAB:

```
L = 50;
theta = zeros(1,L);
y = 84;
theta(1) = 0.5;
for i = 2:50,
    theta(i) = y/(y + 100*(1 - theta(i - 1)));
end
```

# Example: radioactive emission (cont.)

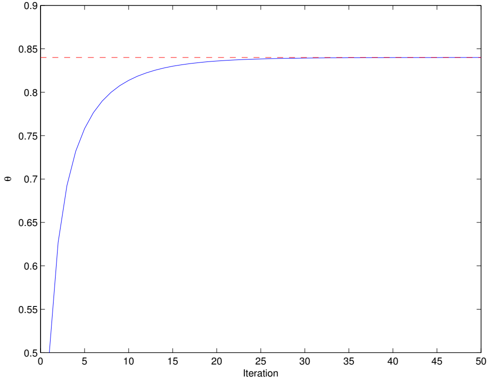



Figure: EM learning trajectory (blue curve) for $\theta$. Red-dashed line indicates the value 0.84.

# Outline

# The EM inequality

- In order to understand the EM algorithm, define the entropy

$$
\begin{aligned}
\mathcal{H}_{\theta'}(\theta) &\stackrel{\text{def}}{=} \ell(\theta) - \mathcal{Q}_{\theta'}(\theta) \\
&= \log f_\theta(Y) - \int \{\log f_\theta(x, Y)\} f_{\theta'}(x \mid Y)\, dx \\
&= -\int \{\log f_\theta(x \mid Y)\} f_{\theta'}(x \mid Y)\, dx.
\end{aligned}
$$

- Consequently,

$$
\begin{aligned}
\mathcal{H}_{\theta'}(\theta) - \mathcal{H}_{\theta'}(\theta') &= \int \log \left\{ \frac{f_{\theta'}(x \mid Y)}{f_\theta(x \mid Y)} \right\} f_{\theta'}(x \mid Y)\, dx \\
&= K(f_{\theta'}(x \mid Y) \| f_\theta(x \mid Y)) \geq 0.
\end{aligned}
$$

# The EM inequality (cont'd)

- By rearranging the terms we obtain the following.

**Theorem (the EM inequality)**

*For all $(\theta, \theta') \in \Theta^2$ it holds that*

$$\ell(\theta) - \ell(\theta') \geq \mathcal{Q}_{\theta'}(\theta) - \mathcal{Q}_{\theta'}(\theta'),$$

*where the equality is strict unless $f_{\theta'}(x \mid Y) = f_{\theta}(x \mid Y)$ (a.s.).*

- Thus, by the very construction of $\{\theta_\ell; \ell \in \mathbb{N}\}$ it is made sure that $\{\ell(\theta_\ell); \ell \in \mathbb{N}\}$ is non-decreasing. Hence, the EM algorithm is a monotone optimization algorithm.

# Convergence of EM

- Under additional differentiability assumptions one may prove that
$$\nabla_\theta \ell(\theta') = \nabla_\theta \mathcal{Q}_{\theta'}(\theta)\big|_{\theta=\theta'}.$$
- Thus, if the algorithm ever stops at $\tilde{\theta}$, then the mapping $\theta \mapsto \mathcal{Q}_{\tilde{\theta}}(\theta)$ must be maximal at $\tilde{\theta}$, which implies that $\nabla_\theta \ell(\tilde{\theta}) = 0$, i.e. $\tilde{\theta}$ is a stationary point of the likelihood.
- The “if the algorithm ever stops”-part has to be established rigorously and some more analysis is thus needed to proof the convergence. This is however possible.

# EM in exponential families

- In order to be practically useful, the E- and M-steps of EM have to be feasible. A rather general context in which this is the case is the following.

**Definition (exponential family)**

The family $\{f_\theta(x, y); \theta \in \Theta\}$ defines an exponential family if the complete data likelihood is of form

$$f_\theta(x, y) = \exp\left(\psi(\theta)^\intercal \phi(x) - c(\theta)\right) h(x),$$

where $\phi$ and $\psi$ are (possibly) vector-valued functions on $\mathbb{R}^d$ and $\Theta$, respectively, and $h$ is a non-negative real-valued function on $\mathbb{R}^d$. All these quantities may depend on $y$.

# EM in exponential families (cont'd)

- The intermediate quantity becomes

$$\mathcal{Q}_{\theta'}(\theta) = \psi(\theta)^\intercal \mathbb{E}_{\theta'}\left(\phi(x) \mid Y\right) - c(\theta) + \underbrace{\mathbb{E}_{\theta'}\left(\log h(x) \mid Y\right)}_{(*)},$$

  where $(*)$ does not depend on $\theta$ and may thus be ignored.
- Consequently, in order to be able to apply EM we need
  1. to be able to compute the "smoothed" sufficient statistics

$$\tau = \mathbb{E}_{\theta'}\left(\phi(X) \mid Y\right) = \int \phi(x) f_{\theta'}(x \mid Y)\, dx.$$

  2. maximization of $\theta \mapsto \psi(\theta)^\intercal \tau - c(\theta)$ to be feasible for all $\tau$.

# Outline

# Example: nonlinear Gaussian model

- Let $(y_i)_{i=1}^n$ be independent observations of a random variable
$$Y = h(X) + \sigma_y \varepsilon_y,$$
where $h$ is a possibly nonlinear function and
$$X = \mu + \sigma_x \varepsilon_x$$
is not observable. Here $\varepsilon_x$ and $\varepsilon_y$ are independent standard Gaussian noise variables.
- The parameters $\theta = (\mu, \sigma_x)$ governing the distribution of the unobservable variable $X$ are unknown while it is known that $\sigma_y = .5$.

# Example: nonlinear Gaussian model (cont.)

- Given observations $Y = (Y_1, \ldots, Y_n)$, the likelihood is

$$\begin{aligned}\ell(\theta) &= \log\left\{\prod_{i=1}^{n}\int f(Y_i \mid x_i) f_\theta(x_i)\,dx_i\right\}\\ &= -\,n\log(2\pi\sigma_x\sigma_y)\\ &+ \sum_{i=1}^{n}\log\int \exp\left(-\frac{1}{2\sigma_y^2}\{Y_i - h(x_i)\}^2 - \frac{1}{2\sigma_x^2}\{x_i - \mu\}^2\right)dx_i,\end{aligned}$$

  which is intractable for a general $h$.
- The complete data log-likelihood $\log\{\prod_{i=1}^{n} f(y_i \mid x_i) f_\theta(x_i)\}$ is however easily computed as it does not contain any integral.

# Example: nonlinear Gaussian model (cont.)

- In this case the complete data likelihood belongs to an exponential family with

$$\phi(x_{1:n}) = \left( \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i \right), \quad \psi(\theta) = \left( -\frac{1}{2\sigma_x^2}, \frac{\mu}{\sigma_x^2} \right),$$

  and

$$c(\theta) = -\frac{n}{2} \log \sigma_x^2 - \frac{n\mu^2}{2\sigma_x^2}.$$

- Letting $\tau_i = \mathbb{E}_{\theta_\ell}(\phi_i(X_{1:n}) \mid Y_{1:n})$, $i \in \{1, 2\}$, leads to

$$\mu_{\ell+1} = \frac{\tau_2}{n},$$
$$(\sigma_x^2)_{\ell+1} = \frac{\tau_1}{n} - \mu_{\ell+1}^2.$$

# Example: nonlinear Gaussian model (cont.)

- Since the "smoothed" sufficient statistics are of additive form it holds that, e.g.,

$$\tau_1 = \mathbb{E}_{\theta_\ell}\left(\sum_{i=1}^{n} X_i^2 \mid Y_{1:n}\right) = \sum_{i=1}^{n} \mathbb{E}_{\theta_\ell}\left(X_i^2 \mid Y_i\right)$$

  (and similarly for $\tau_2$).
- However, computing expectations under

$$f_{\theta_\ell}(x_i \mid y_i) \propto \exp\left(-\frac{1}{2\sigma_y^2}\{y_i - h(x_i)\}^2 - \frac{1}{2(\sigma_x^2)_\ell}\{x_i - \mu_\ell\}^2\right)$$

  is in general infeasible (i.e., when the transformation $h$ is a general nonlinear function).

# Example: nonlinear Gaussian model (cont.)

- Thus, within each iteration of EM we sample each component $f_{\theta_\ell}(x_i \mid y_i)$ using an MH-step.
- For simplicity we use the independent proposal $r(x_i) = f_{\theta_\ell}(x_i)$, yielding the MH acceptance probability

$$\alpha(X_i^{(k)}, X_i^*) = 1 \wedge \frac{f_{\theta_\ell}(Y_i \mid X_i^*)}{f_{\theta_\ell}(Y_i \mid X_i^{(k)})}$$

$$= 1 \wedge \exp\left(-\frac{1}{2\sigma_y^2}\{h^2(X_i^*) - h^2(X_i^{(k)}) + 2Y_i(h(X_i^{(k)}) - h(X_i^*))\}^2\right)$$

# Example: nonlinear Gaussian model (cont.)

**Data:** Initial value $\theta_0$; $Y_{1:n}$
**Result:** $\{\theta_\ell; \ell \in \mathbb{N}\}$
**for** $\ell \leftarrow 0, 1, 2, \dots$ **do**
  set $\hat{\tau}_j \leftarrow 0$, $\forall j$;
  **for** $i \leftarrow 1, \dots, n$ **do**
    run an MH sampler targeting $f_{\theta_\ell}(x_i \mid Y_i) \rightsquigarrow (X_i^{(k)})_{k=1}^{N_\ell}$;
    set $\hat{\tau}_1 \leftarrow \hat{\tau}_1 + \sum_{k=1}^{N_\ell} (X_i^{(k)})^2 / N_\ell$;
    set $\hat{\tau}_2 \leftarrow \hat{\tau}_2 + \sum_{k=1}^{N_\ell} X_i^{(k)} / N_\ell$;
  **end**
  set $\mu_{\ell+1} \leftarrow \hat{\tau}_2 / n$;
  set $(\sigma_x^2)_{\ell+1} \leftarrow \hat{\tau}_1 / n - \mu_{\ell+1}^2$;
**end**

# Example: nonlinear Gaussian model (cont.)

- In order to assess the performance of the MCEM algorithm we focus on the linear case, $h(x) = x$.
- A data record comprising $n = 40$ values was produced by simulation under $(\mu^*, \sigma_x^*) = (1, .4)$ with $\sigma_y = .4$.
- In this case, the true MLE is known and given by (check this!)

$$\hat{\mu}(Y_{1:n}) = \frac{1}{n} \sum_{i=1}^{n} Y_i = 1.04,$$

$$\hat{\sigma}_x^2(Y_{1:n}) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \hat{\mu}(Y_{1:n})\}^2 - .4^2 = .27.$$

# Example: nonlinear Gaussian model (cont.)

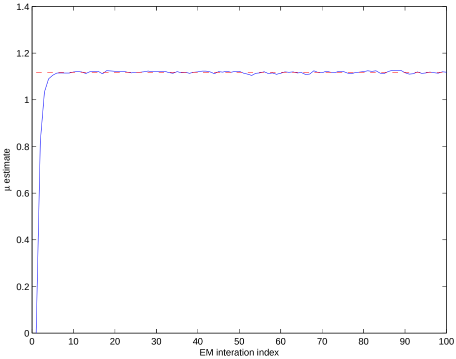



Figure: EM learning trajectory (blue curve) for $\mu$ in the case $h(x) = x$. Red-dashed line indicates true MLE. $N_\ell = 1,000$ for all iterations.

# Example: nonlinear Gaussian model (cont.)

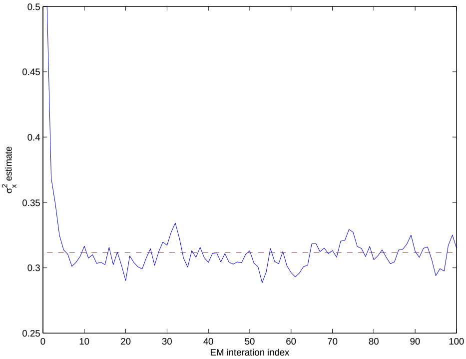



Figure: EM learning trajectory (blue curve) for $\sigma_x^2$ in the case $h(x) = x$. Red-dashed line indicates true MLE. $N_\ell = 1,000$ for all iterations.

# Averaging

- Roughly, for the MCEM algorithm one may prove that the variance of $\theta_\ell - \hat{\theta}$ (where $\hat{\theta}$ is the MLE) is of order $1/N_\ell$.
- In the idealised situation where the estimates $(\theta_\ell)$ are uncorrelated (which is not the case here) one may obtain an improved estimator of $\hat{\theta}$ by combining the individual estimates $\theta_\ell$ in proportion of the inverse of their variance. Starting the averaging at iteration $\ell_0$ leads to

$$\tilde{\theta}_\ell = \sum_{m=\ell_0}^{\ell} \frac{N_m}{\sum_{m'=\ell_0}^{\ell} N_{m'}} \theta_m, \quad \ell \geq \ell_0.$$

- In the idealised situation the variance of $\tilde{\theta}_\ell$ is inversely proportional to $\sum_{m=\ell_0}^{\ell} N_m$ (= total number of simulations).

# Example: nonlinear Gaussian model (cont.)

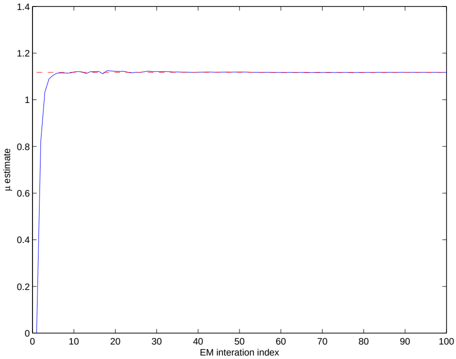



Figure: EM learning trajectory (blue curve) for $\mu$ in the case $h(x) = x$. Red-dashed line indicates true MLE. Averaging after $\ell_0 = 30$ iterations.

# Example: nonlinear Gaussian model (cont.)

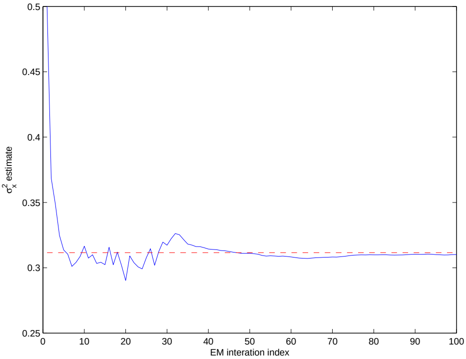



Figure: EM learning trajectory (blue curve) for $\sigma_x^2$ in the case $h(x) = x$. Red-dashed line indicates true MLE. Averaging after $\ell_0 = 30$ iterations.