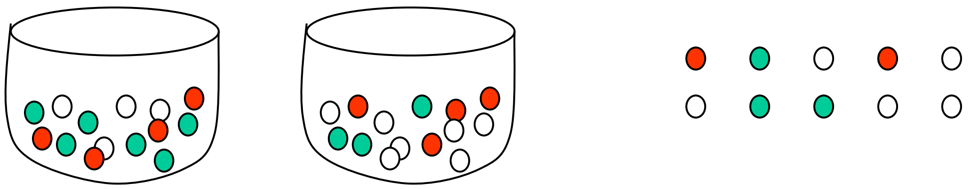
Inference with Hidden Variables

Summary

- EM Method
- Estimation of Gaussian mixtures
- Identification of Multiple Dynamic Systems

Challenge



We extract n pairs of balls, each pair from one box (we don't know which). Each box is randomly selected with equal probability.

Is it possible to guess the color content of each box ?

1st Try

The ML method can be used to estimate B.

Variables:

- $k = k_1, \ldots, k_n$ sequence of chosen boxes
- $x = x_1, \ldots, x_n$ sequence of 1st balls
- $y = y_1, \ldots, y_n$ sequence 2nd balls
- B_{ij} probability of extracting ball j from box i

Log likelihood function:

$$l(B) = c + \sum_t \log\left(B_{1x_t} B_{1y_t} + B_{2x_t} B_{2y_t}\right)$$

The optimization of this function is difficult. Alternative methods ?

EM Method

The EM method is used when there is incomplete observations:

- y observed variables
- x hidden variables (missing)
- θ vector of parameters to estimate

and a probabilistic model p(x,y|θ) is known.

The estimation of q can be solved using the ML method i.e., maximizing the likelihood function

$$p(y \mid \theta) = \int p(x, y \mid \theta) dx$$

This task is unfeasible in many problems.

EM Method

The ML estimate of θ, knowing *x and y* is

$$\hat{\theta} = \arg\max_{\theta} \log p(y, x \mid \theta)$$

If x is unknown the EM method replaces the log likelihood function of x,y by the expected value, using the conditional distribution of x

$$p(x \mid y, \theta^{old})$$

The auxiliary function

$$U(\theta, \theta^{old}) = E\{\log p(x, y \mid \theta) \mid y, \theta^{old}\} = \int \log p(x, y \mid \theta) p(x \mid y, \theta^{old}) dx$$

is then optimized with respect to θ.

EM Method

(Dempster, Laird, Rubin, 1977)

The EM (Expectation-Maximization) method is an iterative method based on two steps:

E step: $$U(\theta,\theta^{t-1}) = E\{\log p(x,y|\theta)\,|y,\theta^{t-1}\}$$

M step: $$\theta^{t} = \arg\max_{\theta} U(\theta,\theta^{t-1})$$

The E step computes the conditional distribution of the hidden variables, knowing the available information y and the best estimate of the unknown parameters: $p(x|y,\theta^{t-1})$.

- the likelihood function does not decrease in each iteration
- if the algorithm converges, it converges to a local maximum of the likelihood function.

Proof

$$\log p(y|\theta) = \log p(x,y|\theta) - \log p(x|y,\theta)$$

Taking the expected value assuming that $x \sim q(x) = p(x|y,\theta^{old})$

$$\log p(y|\theta) = \int q(x) \log p(x,y|\theta)\, dx - \int \mathrm{q(x)} \log p(x|y,\theta)\, dx$$

$$\log p(y|\theta) = \mathrm{U}\left(\theta,\theta^{old}\right) + \mathrm{H}\left(\theta,\theta^{old}\right)$$

It can be shown that $\mathrm{H}\left(\theta,\theta^{old}\right) \geq \mathrm{H}\left(\theta^{old},\theta^{old}\right)$

$$\log p(y|\theta) - \log p\left(y\middle|\theta^{old}\right) \geq U(\theta,\theta^{old}) - U(\theta^{old},\theta^{old})$$

Therefore, maximizing U wrt θ improves the likelihood function.

Challenge (revisited)

x,y observed variables
k sequence of boxes (hidden)
B parameters to estimate.

Total log-likelihood function

$$l = \log p(x,y,k \mid B) = c + \sum_t \log B_{k_t x_t} + \log B_{k_t y_t}$$

E step

$$U(B,B^{t-1}) = c + \sum_t \sum_i \alpha_t(i)\left(\log B_{ix_t} + \log B_{iy_t}\right)$$

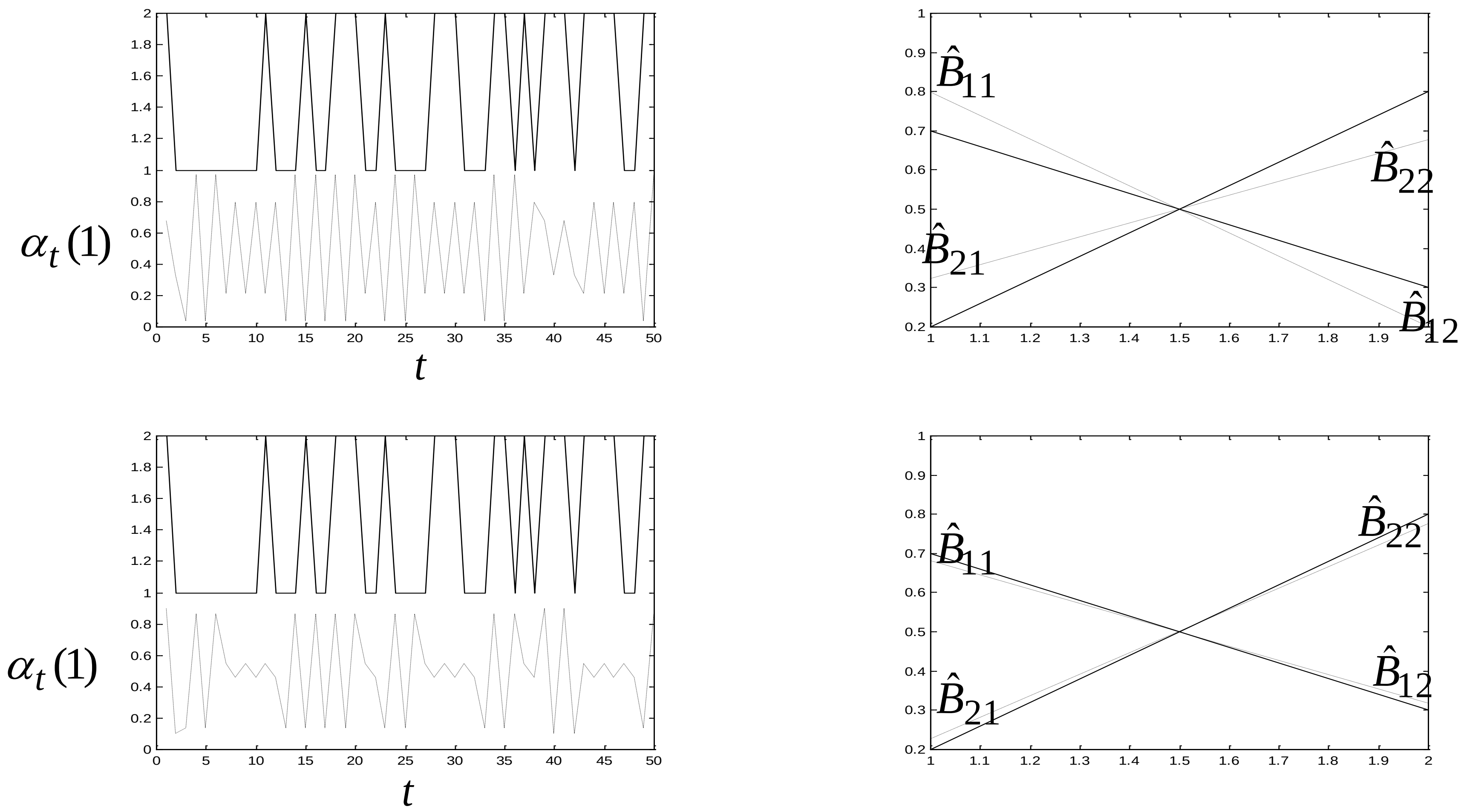
$$\alpha_t(i) = P(k_t / x_t, y_t) = \frac{B_{ix_t}^{t-1} B_{iy_t}^{t-1}}{\sum_j B_{jx_t}^{t-1} B_{jx_t}^{t-1}}$$

M step

$$B_{pq}^t = \frac{\sum_{t:x_t=q} \alpha_t(p) + \sum_{t:y_t=q} \alpha_t(p)}{2\sum_t \alpha_t(p)}$$

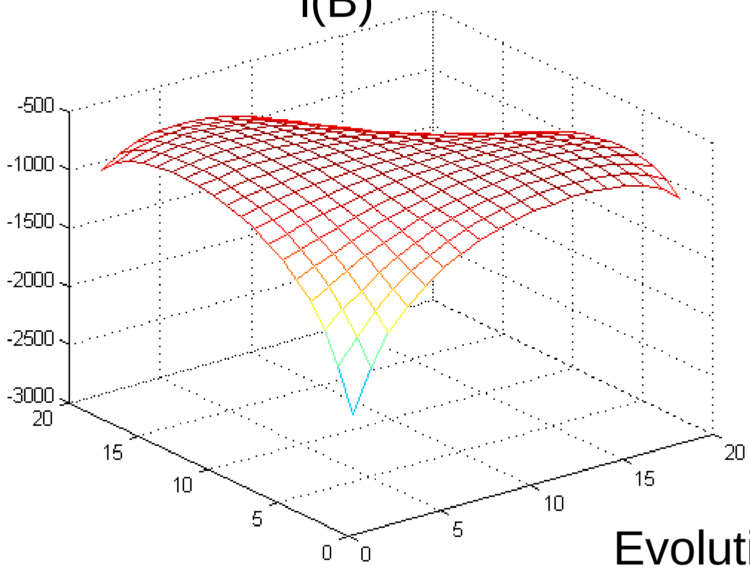
$\alpha_t(i)$ is the membership degree of the observation t to the class i

Results

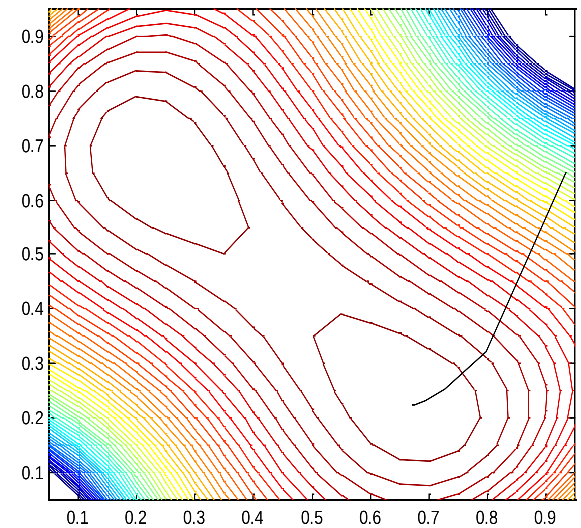

$\alpha_t(1)$
t
$\hat{B}_{11}$
$\hat{B}_{22}$
$\hat{B}_{21}$
$\hat{B}_{12}$
$\alpha_t(1)$
t
$\hat{B}_{22}$
$\hat{B}_{11}$
$\hat{B}_{12}$
$\hat{B}_{21}$

Result

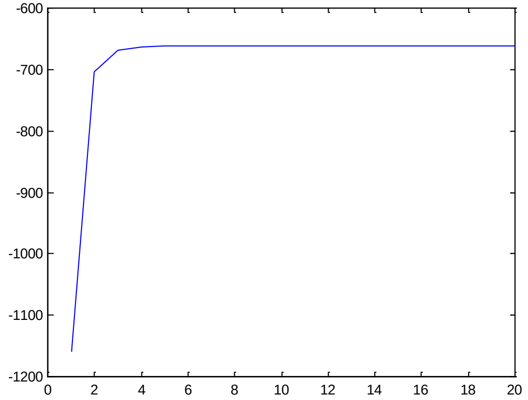
l(B)



l(B)



Evolution of l(B)

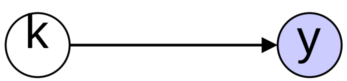


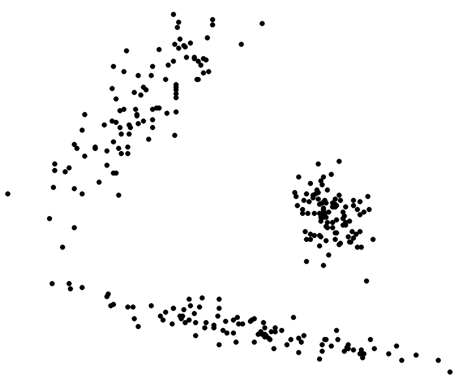
Mixtures

Mixture of distributions

$$p(y) = \sum_{k=1}^{m} c_k p_k(y) \qquad c_k \geq 0 \ \forall k, \quad \sum_{k=1}^{m} c_k = 1$$

Model





K is a discrete hidden variable with distribution P{k=i}=c_i; k selects which distribution p_k generates y. Only y is observed.

Estimation of Gaussian Mixtures

Given n observations $y_1,\ldots, y_n$, generated by a mixture of Gaussian distributions, we wish to estimate the mixture parameters: mixture coefficients, mean vectors and covariance matrices.

In this case,

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) \qquad p(y_i / \theta) = \sum_{i=1}^{m} c_i N(y_i ; \mu_i, R_i)$$

Log likelihood function :

$$l(\theta) = \sum_i \log\left\{ \sum_k c_k \frac{1}{(2\pi)^{\dim x/2} |R_k|^{1/2}} \exp\{-\tfrac{1}{2}(y_i - \mu_k)' R_k^{-1} (y_i - \mu_k)\} \right\}$$

The optimization of l is difficult.

Mixture Estimation

E step – compute the distribution of the hidden variables

$$w_{ij} = P\{k_i = j / y_i, \theta\} = \alpha N(y_i; \hat{\mu}_i, \hat{R}_i)\hat{c}_j$$

a – normalization factor

M step – parameter update

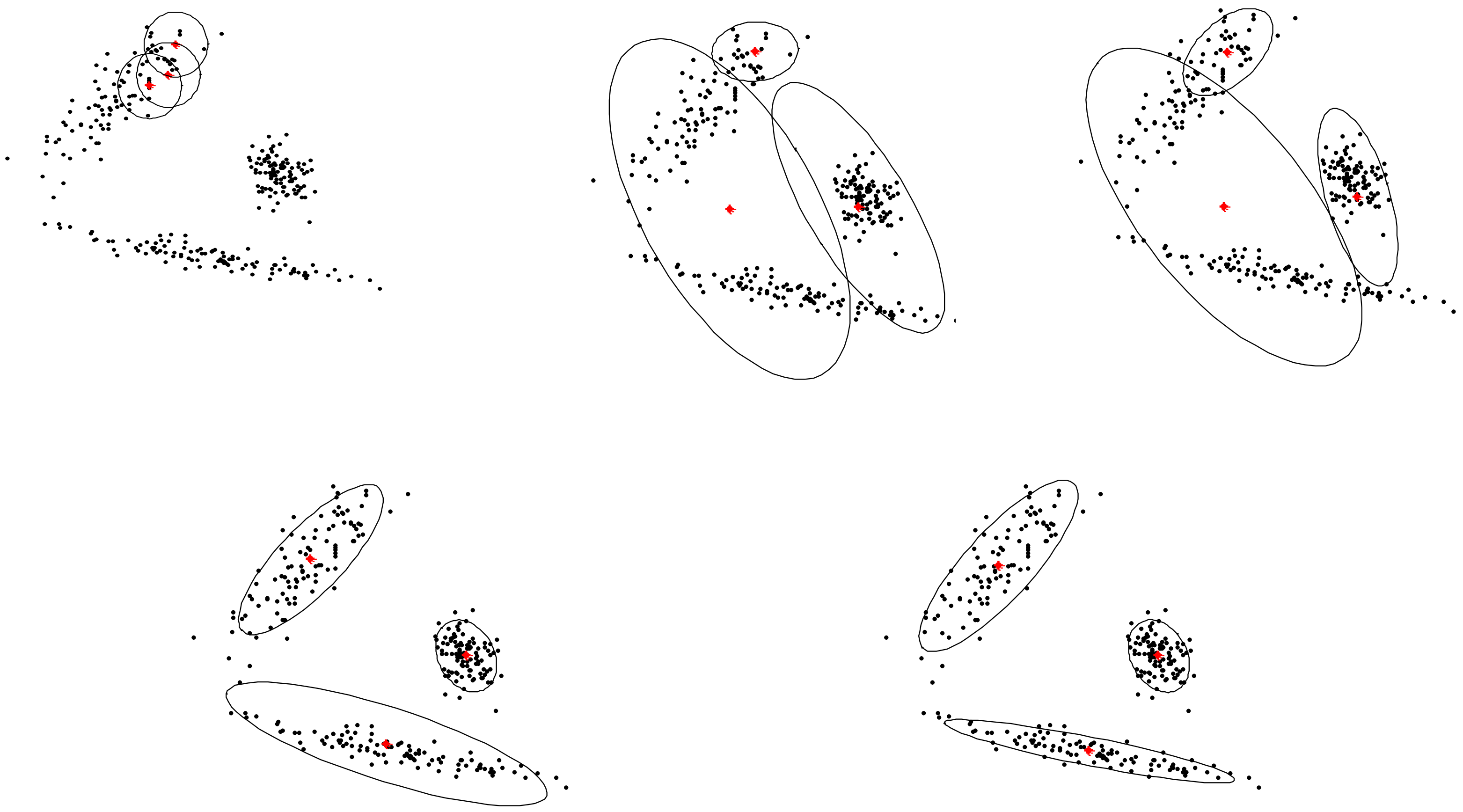
$$\hat{c}_j = \frac{1}{n}\sum_i w_{ij}$$

$$\hat{\mu}_j = \frac{1}{n}\sum_i w_{ij} y_i \qquad \hat{R}_j = \frac{1}{n}\sum_i w_{ij}(y_i - \hat{\mu}_j)(y_i - \hat{\mu}_j)'$$

The mean vectors can be initialized with the first *m* observations.

Example

Estimation of a mixture of Gaussians with the EM algorith; iterations 0, 1, 5, 10, 15



k-Means

The last problem is a clustering problem if we associate a cluster to each mixture mode.

The EM method is related to the k-means algorithm which performs a hard classification of the data *y*, replacing the unknown variables *k* by its most probable values.

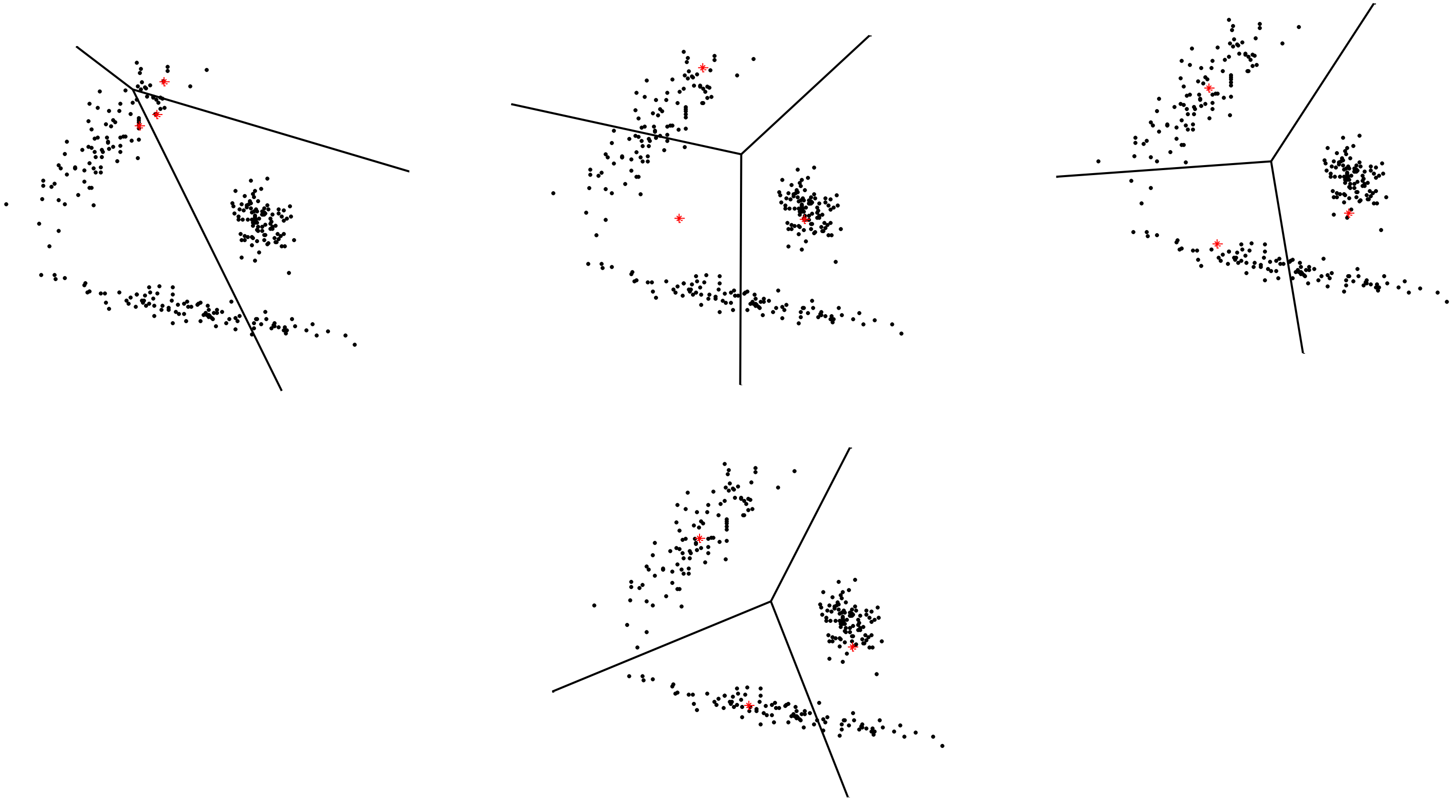
k-means algorithm

1. Initialize the mean vectors
2. repeat until convergence is achieved:
   - classify data patterns y in the class with closest mean vector
   - update the mean vectors using the patterns classified in each class

Note: the k-means algorithm assumes that the covariance matrices are all equal to the identity I.

Example

Clustering with the k-means algorithm; iterations 0, 1, 5, 10



Multi-Predictors

Sometimes a signal is described by several models. A single predictor is not enough to cope with this situation.

Example: Let us consider two predictors:

$$y_i = \phi_i' \theta^{k_i} + v_i \qquad \mathrm{k_i} = 1,2,...$$

$$\theta^k = [a_1^k \ldots a_p^k]' \qquad \phi_i = [y_{i-1} \ldots y_{i-p}]' \qquad v_i \sim N(0, \sigma_k^2)$$

How can we estimate the parameters of multiple predictors ?

The difficulty lies in the fact that we do not know which predictor is valid at each instant of time since k_1 ,..., k_n are unknown.

Learning Multi-Preditors

Learning multi-preditors parameters can be done using the EM method.

E step – distribution of hidden variables

$$w_{ij} = P\{k_i = j \mid y_i, \theta\} = \alpha N(y_i - \phi_i \theta^j ; 0, \sigma_j^2)\hat{c}_j$$

M step – parameter update

$$\hat{c}_j = \frac{1}{n} \sum_i w_{ij}$$

$$\theta^j = (A^j)^{-1} b^j \quad A^j = \sum_i w_{ij} \varphi_i \varphi_i' \quad b^j = \sum_i w_{ij} \varphi_i y$$

$$\sigma_j^2 = \frac{1}{n c_j} \sum_i w_{ij} (y_i - \phi_i \theta^j)^2$$

Robust Estimation

Many real signals contain outliers (data points which are not described by the model).

The estimation methods based on Gaussian distribution assumptions have a poor performances in the presence of outliers (outliers have a strong influence in the estimates).

One way to avoid this problem consists of using outlier models to describe invalid data (p.ex., using a Gaussian distributions with large covariance matrix).

The distribution of the data with outliers is a mixture of both distributions.

F. Girosi, Models of Noise and Robust Estimates, MIT AI Memo 1287, 1991

Variational EM

The EM method is still too difficult to be applied in many complex problems.

A more general framework consists of approximating the a posteriori distribution of the unknown parameters by a simpler distribution q(x).

Auxiliary function: $F(\mathrm{q}, \theta) = \int q(x) \log \frac{p(x,y|\theta)}{q(x)} \mathrm{dx}$

Iteração: passo E:

maximizar $F(q, \theta^{old})$ em ordem a $q(x)$

passo M:

maximizar $F(q^{old}, \theta)$ em ordem a θ

Proof

$$\log p(y|\theta) = F(q,\theta) + D_{KL}(q||p)$$

q – auxiliary distribution

where

$$F(q,\theta) = \int p(x) \log \frac{p(x,y|\theta)}{q(x)} dx \qquad D(q||p) = -\int p(x) \log \frac{p(x|y,\theta)}{q(x)} dx$$

Since $D_{KL}(q,p) \geq 0$, then $$\log p(y|\theta) \geq F(q,\theta)$$

Idea: iteratively optimize $F(q,\theta)$ wrt probability distribution $q(x)$ and parameter θ.

Choice of auxiliary distribution

1. Unconstrained q(x): the minimization of *F(q,θ) w.r.t. q(x)* leads to

$$q(x) = p(x|y,\theta)$$

the *a posteriori* distribution of the hidden variables.
This is the choice of classic EM method and it leads to integrals that may not be analytically evaluated.

2. Constrained q(x): choose a parametric model for q trying to simplify the calculation of F(q,θ).

Exercises

1. Derive the EM algorithm for the estimation of a mixture of Gaussians .

2. Derive an algorithm to approximate 3D data points by two vertical planes. Make appropriate hypothesis about the observation noise.

3. The flow of a river has two different regimes, depending on a nearby factory being active or not

$$x_t = c_1 x_{t-1} + c_0 w_t$$
$$x_t = d_1 x_{t-1} + d_0 w_t$$

x_t – flow, $w_t \sim N(0,1)$ random perturbation

We know the flow at several consecutive days but we don't know which model is active. Define an algorithm to identify the system parameters and to detect which model is active.

Computer work

1. Consider a process generated by $y_t = .9y_{t-1} + w_t$, w_t, N~(0,1). Suppose the output signal is given by:

$$z_t = \begin{cases} y_t & \text{with probability } .95 \\ v_t \sim N(0,100) & \text{with probability } .05 \end{cases}$$

Apply the EM method to estimate the parameters of the model from the sensor measurements and experimentally evaluate the algorithm .

2. Apply the EM method to estimate a straight line from experimental data, assuming that α % of all observations are outliers. Evaluate the algorithm performance for different values of α.