Data analysis
spatio-temporal data and hierarchical models

Bayesian Inference (Part two, Ch3)

Djemel Ziou

Plan

- Probability of things
- Generative learning
- Discriminative learning
- Bayesian inference
- Point estimation methods
- Approximations
- Feature selection
- References

Approximations

Why approximation?

- The central task of Bayesian inference is to specify the posterior $p(\Theta \mid X)$ and the expectations with regards to this distribution.
- It can be done by using exact algorithms.
- However, in many cases the time and space complexity of exact Bayesian inference algorithms are inacceptable
  - high dimensional space, intractable integral, exponential increasing of hidden state (discrete distribution).
- In these cases, approximations are required
  - Laplace approximation
  - variational approximation
  - Monte Carlo
  - search-based algorithms
  - …

Laplace approximation

- The idea is to approximate a given distribution of continuous variables with a Gaussian around a given value of parameters.
- Example
  - let us consider the case of the posterior $p(\Theta|D)=p(D|\Theta)p(\Theta)/\int_{\Omega}p(D|\Theta)p(\Theta)d\Theta$. We would like to approximate $p(D|\Theta)$. Let us assume that Θ^* is the maximum likelihood.
  - by using 2nd order Taylor expansion of $\ln p(D|\Theta)$ around Θ^*, then $\ln p(D|\Theta)\approx \ln p(D|\Theta^*)+(\Theta-\Theta^*)^t\nabla_{\Theta}\ln p(D|\Theta)|_{\Theta=\Theta^*}+(\Theta-\Theta^*)^t\nabla_{\Theta}\nabla_{\Theta}\ln p(D|\Theta)|_{\Theta=\Theta^*}(\Theta-\Theta^*)$.
  - because $\nabla_{\Theta}\ln p(D|\Theta^*)=0$, we can rewrite $p(D|\Theta)\approx p(D|\Theta^*)\exp(-(\Theta-\Theta^*)^t(-\nabla_{\Theta}\nabla_{\Theta}p(D|\Theta))|_{\Theta=\Theta^*}(\Theta-\Theta^*))$.

Laplace approximation

- Example
  - we would like to estimate p(D)

    $p(D)=\int_\Omega p(D|\Theta)p(\Theta)d\Theta$
  - let us set $f(\Theta) = p(D|\Theta)p(\Theta)$. According to the previous development $p(D) \approx \int_\Omega f(\Theta^*)G(\Theta, \Theta^*)\, d\Theta = f(\Theta^*) \int_\Omega G(\Theta, \Theta^*)\, d\Theta = f(\Theta^*) (2\pi)^{d/2}/|H|^{1/2}$, where H is the Hessian and G() the Gaussian form.
  - it follows that $\ln p(D) \approx \ln p(D|\Theta^*) + \ln p(\Theta^*) + d \ln(2\pi)/2 - \ln|H|/2$
  - this is the Laplace Empirical Criterion (LEC) used for the model selection (see Model Selection).

Laplace approximation

- Laplace approximation is easy.
- It can be used only for continuous variables.
- It fails if there are several maxima.

Variational approximation

- The idea of variational calculus is to specify a functional and to study it by varying the functions
  - example : $\min_{p(x)} \int_{\Omega} p(x) \log_2 p(x)\, dx + \lambda(1 - \int_{\Omega} p(x))dx$

- Variational calculus is used to find optimal distributions, functions, lines, curves, surfaces… Even if it is not intended to approximations, we can approximate the distributions or deal with bounds.

Variational approximation

- The variational approximation of the posterior can be summarized by
  - Let X be the incomplete data and z a latent random vector, the likelihood:

$$\ln p(\mathrm{X} \mid \Theta) = \underbrace{\int q(z) \ln\left(\frac{p(\mathrm{X}, z \mid \Theta)}{q(z)}\right) dz}_{F(q,\Theta)} \underbrace{- \int q(z) \ln\left(\frac{p(z \mid \mathrm{X}, \Theta)}{q(z)}\right) dz}_{KL(q\|p)}$$

where q(z) is some chosen distribution of the latent vector z.

  - since $KL() \geq 0$, it follows that $\ln p(\mathrm{X} \mid \Theta) \geq F(q,\Theta)$, then we can just maximize $F(q,\Theta)$ with regards to both q and Θ

Variational approximation

- Example of EM algorithm
  - E step: $q^{t}_{ML} = \arg\max_{q} F(q, \Theta^{t})$, it happens when KL()=0 it follows that $q^{t}_{ML}(z) = p(z \mid X, \Theta^{t})$
  - M Step: $\Theta^{t+1}_{ML} = \arg\max_{\theta} F(q^{t}_{ML}, \Theta)$
  - drawback: at convergence $p(z \mid X, \Theta^{t+1}_{ML}) \neq q^{t}_{ML}(z)$

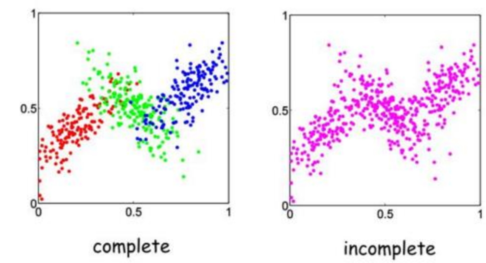
Variational approximation

- EM revisited
  - replacing $q^t(z) = p(z|\mathrm{X}, \Theta^t)$ in F() gives

$$F(q^t, \Theta) = \underbrace{\int p(z|\mathrm{X}, \Theta^t) \ln(p(\mathrm{X}, z|\Theta)) dz}_{Q(\Theta,\Theta^t) = E_{p(z|\mathrm{X},\Theta^t)}(\ln(p(\mathrm{X}, z|\Theta)))} + cte$$

  - E step: compute $p(z|\mathrm{X}, \Theta^t)$
  - M Step: $\Theta^{t+1} = \arg\max_\theta Q(\Theta, \Theta^t)$
  - drawback: $p(z|\mathrm{X}, \Theta)$ can be unknown.

Variational approximation-mixture of pdfs



- Let us consider the complete data
- d-dimensional observations D=(X,Z),
  - observed data X={x_1, x_2,…x_N}

$$X=\begin{bmatrix} x_{11} & \cdots & x_{1d} \\ . & \cdots & . \\ x_{N1} & \cdots & x_{Nd} \end{bmatrix}$$

  - we consider discreet the labels Z={z_1, z_2,…z_N}

$$Z=\begin{bmatrix} z_{11} & \cdots & z_{1d} \\ . & \cdots & . \\ z_{N1} & \cdots & z_{Nd} \end{bmatrix}$$

  only one element of the raw is non zero (=1).

  - complete likelihood

$$p(D, Z \mid \theta)= \prod_{n=1}^{N} \prod_{k=1}^{M} (\alpha_k f(x_n|\theta_k))^{z_{nk}}$$

$$\ln(p(Z \mid D, \theta)) \propto \sum_{n=1}^{N} \sum_{k=1}^{M} z_{nk}\big(\ln(\alpha_k) + \ln(f(x_n|\theta_k)\big)$$

Variational approximation-mixture of pdfs

- But z is unknown,
  - we need to use the posterior. Let us consider the n[th] raw

$p(z_{nk}=1 \mid x_n,\ \theta)=p(x_n, z_{nk}=1 \mid \theta)\, p(z_{nk}=1) \,/\, \sum_{k=1}^{M} \ p(x_n, z_{nk}=1 \mid \theta)$
$= \alpha_k f(x_n|\theta_k)/\sum_{j=1}^{M} \ \alpha_j f\left(x_n \middle| \theta_j\right)$

  - then

$q_{ML}^t(z_{nk}=1) = p(z_{nk}=1|x_n,\Theta^t)=\alpha_k f(x_n|\theta_k^t)/\sum_{j=1}^{M} \ \alpha_j f\left(x_n\middle|\theta_j^t\right)$

$F(q_{ML}^t,\Theta) = \sum_{n=1}^{N} \ E_z(\ln(p(x_n,z|\Theta)))=$
$\sum_n \sum_k p(z_{nk}=1|x_n,\Theta^t)(\ln\ \alpha_k+\ln f(x_n|\theta_k))$

  - Example

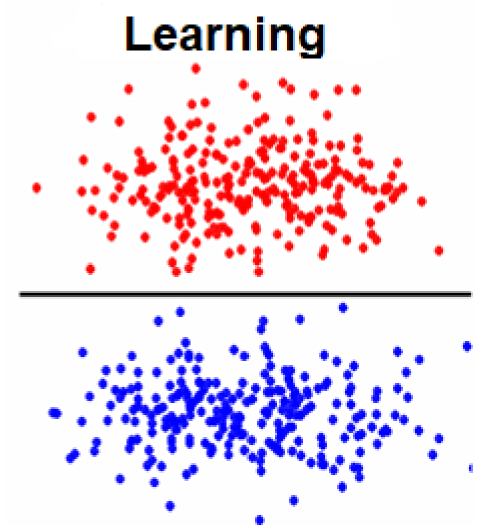
$$\frac{\partial F}{\partial \mu_k} = \sum_n p(z_{nk}=1|x_n,\Theta^t)\frac{\frac{\partial f(x_n|\theta_k)}{\partial \mu_k}}{f(x_n|\theta_k)} = 0$$

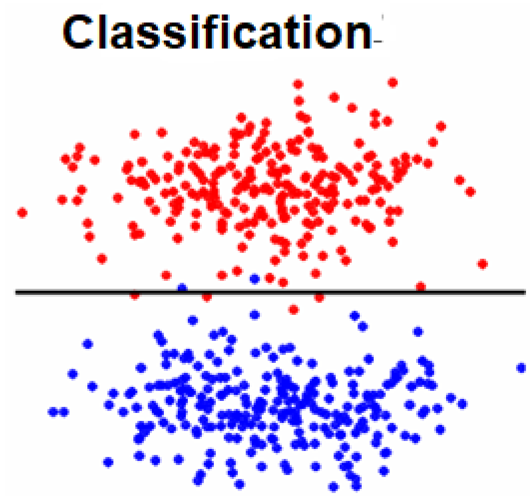
if $f(x_n|\theta_k)$ is a Gaussian

$$\sum_n p(z_{nk}=1|x_n,\Theta^t)(x_n - \mu_k) = 0$$

Variational approximation-mixture of pdfs

- K-means implement a clustering of data
  Two-phase process:
  (1) Label identification: building a model that describes a predetermined set of classes of data.
  (2) Classification: Use the model to assign a class to a new object.





Variational approximation-mixture of pdfs

- Label identification: Partition the set of samples into K separate clusters, each of which is represented by a center, the number K is chosen in advance.

  The best partition is obtained by minimizing the dispersion within clusters defined by the following function:

$$J_{KM} = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} d(x_i, m_k)$$

where

- $\boldsymbol{N}$ number of samples
- K number de clusters
- $\boldsymbol{x_i}$ i^{th} sample
- m_k centre of the k^{th} cluster
- $\boldsymbol{u_{ik}}$ membership of x_i to the k^{th} cluster (0 or 1)
- ***d()*** a similarity measure (e.g. distance).

Variational approximation-mixture of pdfs

- Algorithm

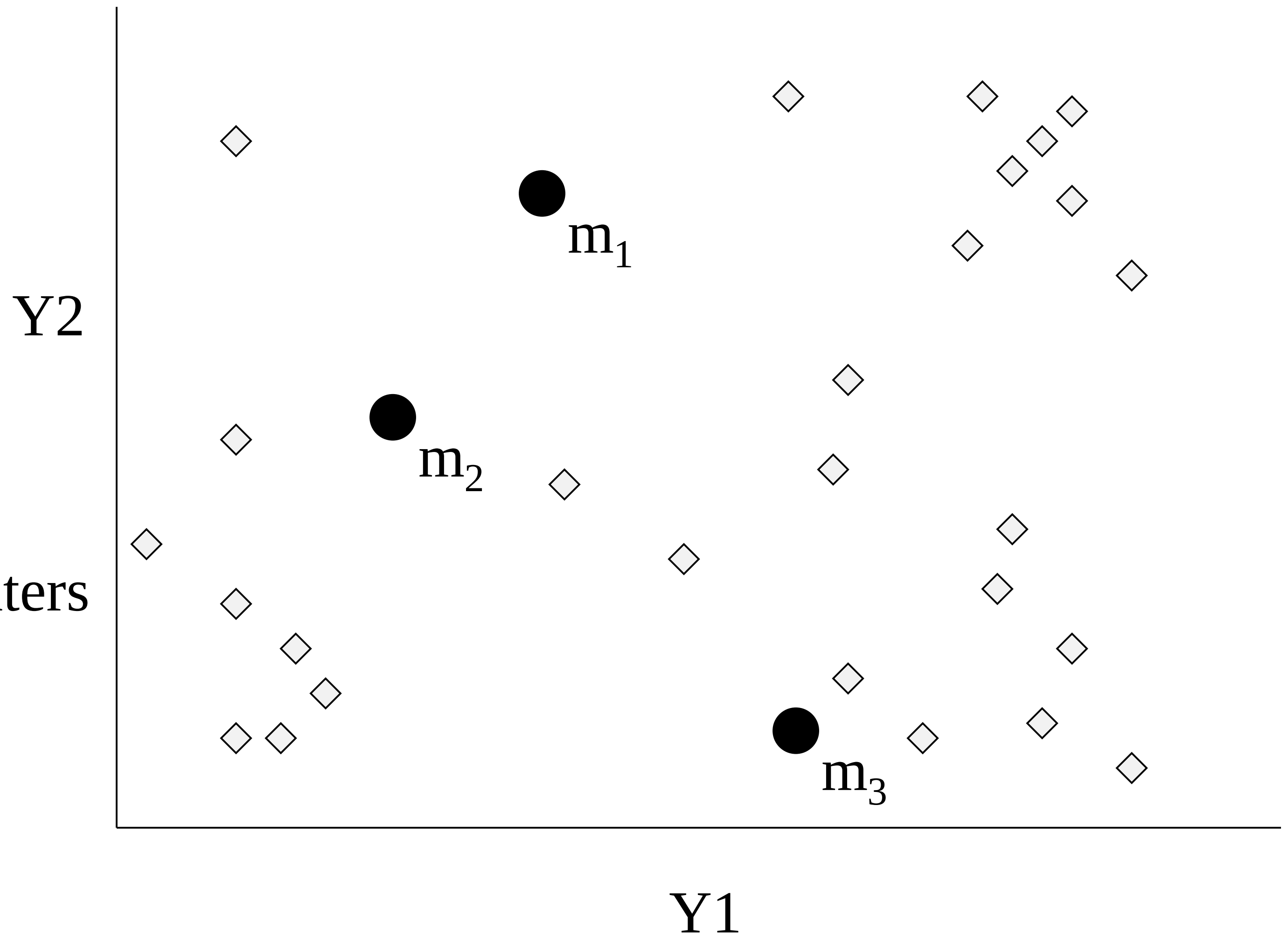
  Input: Data set X, number of clusters K

  Output: A data partition (K clusters)

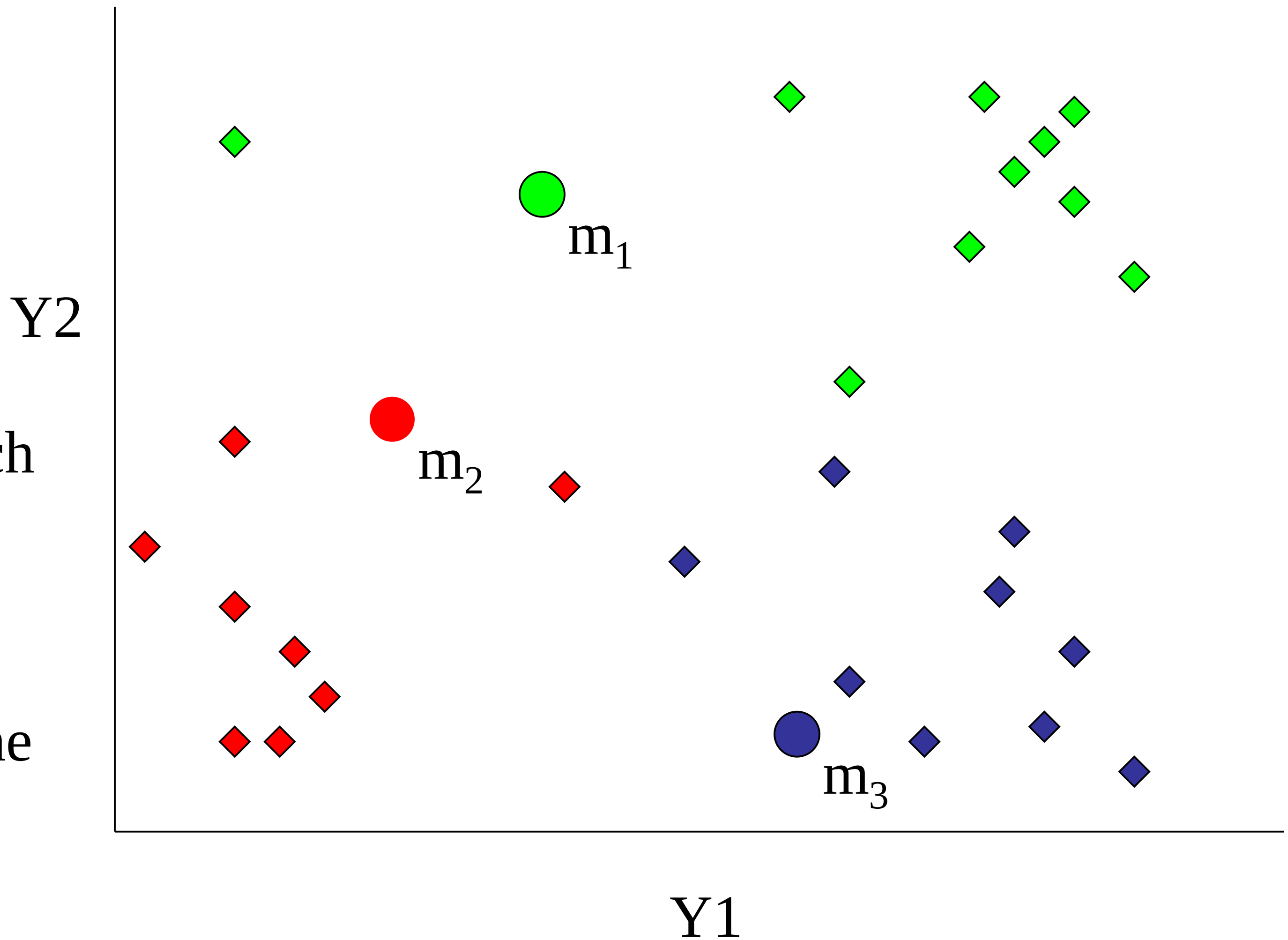
  1. Randomly choose a center for each of the K clusters
  2. Assign each sample to the cluster with the nearest center (using Euclidean distance)
  3. Move each center to the average of the samples of the cluster
  4. Repeat 2 to 3 until convergence (i.e. until all centers remain unchanged)

K-means (Example)

randomly
choose 3
cluster centers



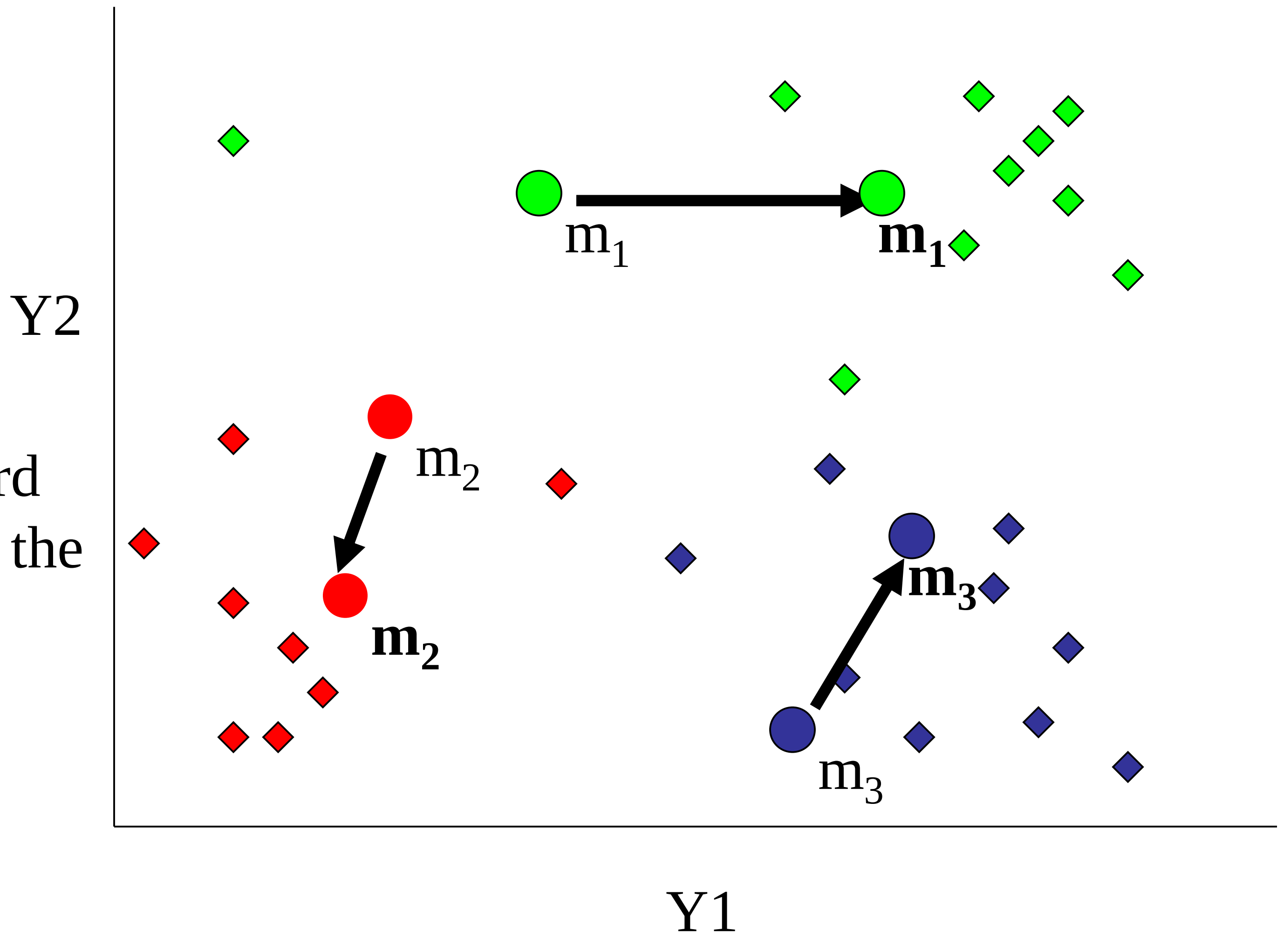
K-means (Example)



Assign each sample to the cluster where the center is the closest

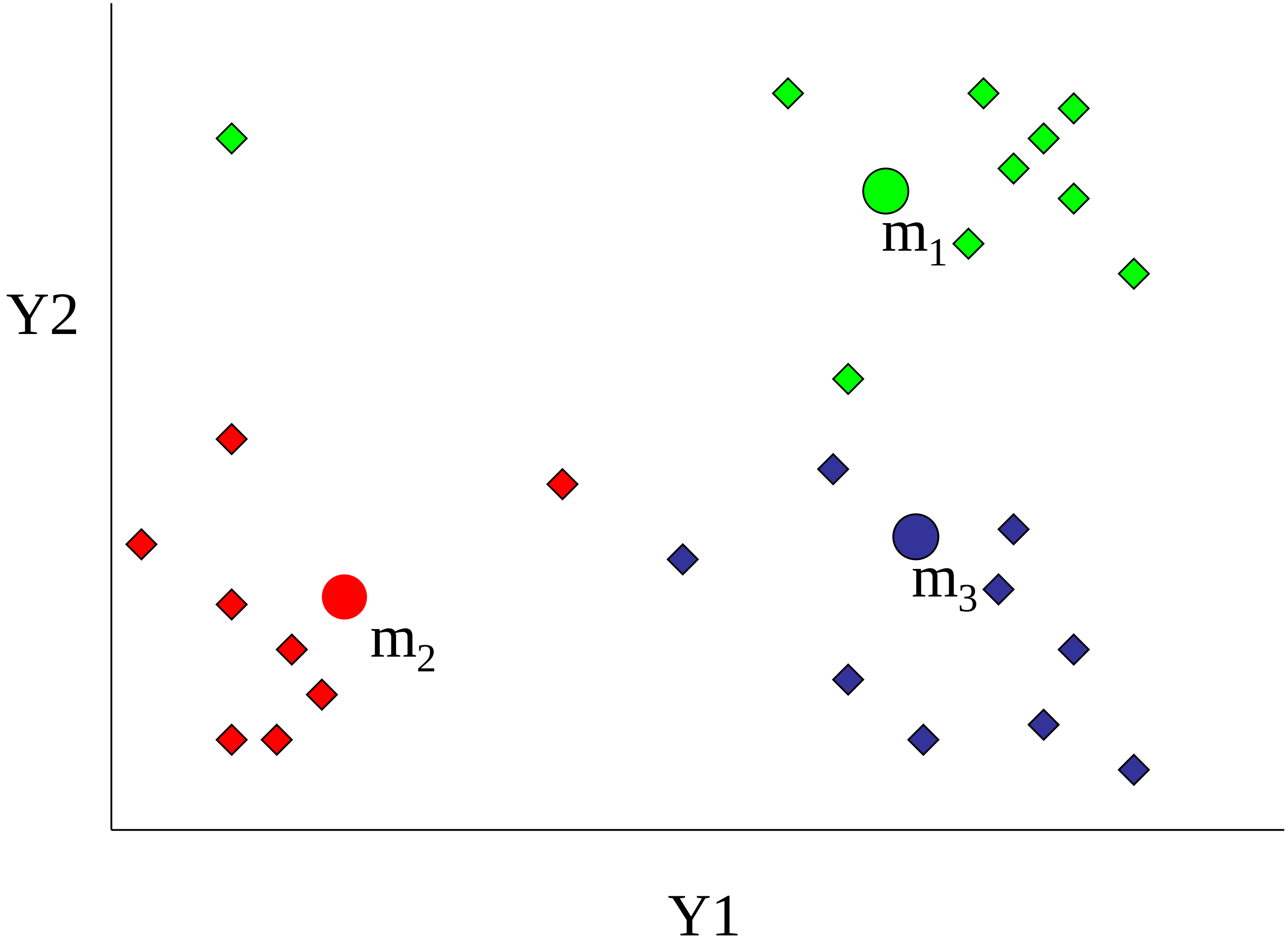
K-means (Example)

Move each center toward the mean of the cluster



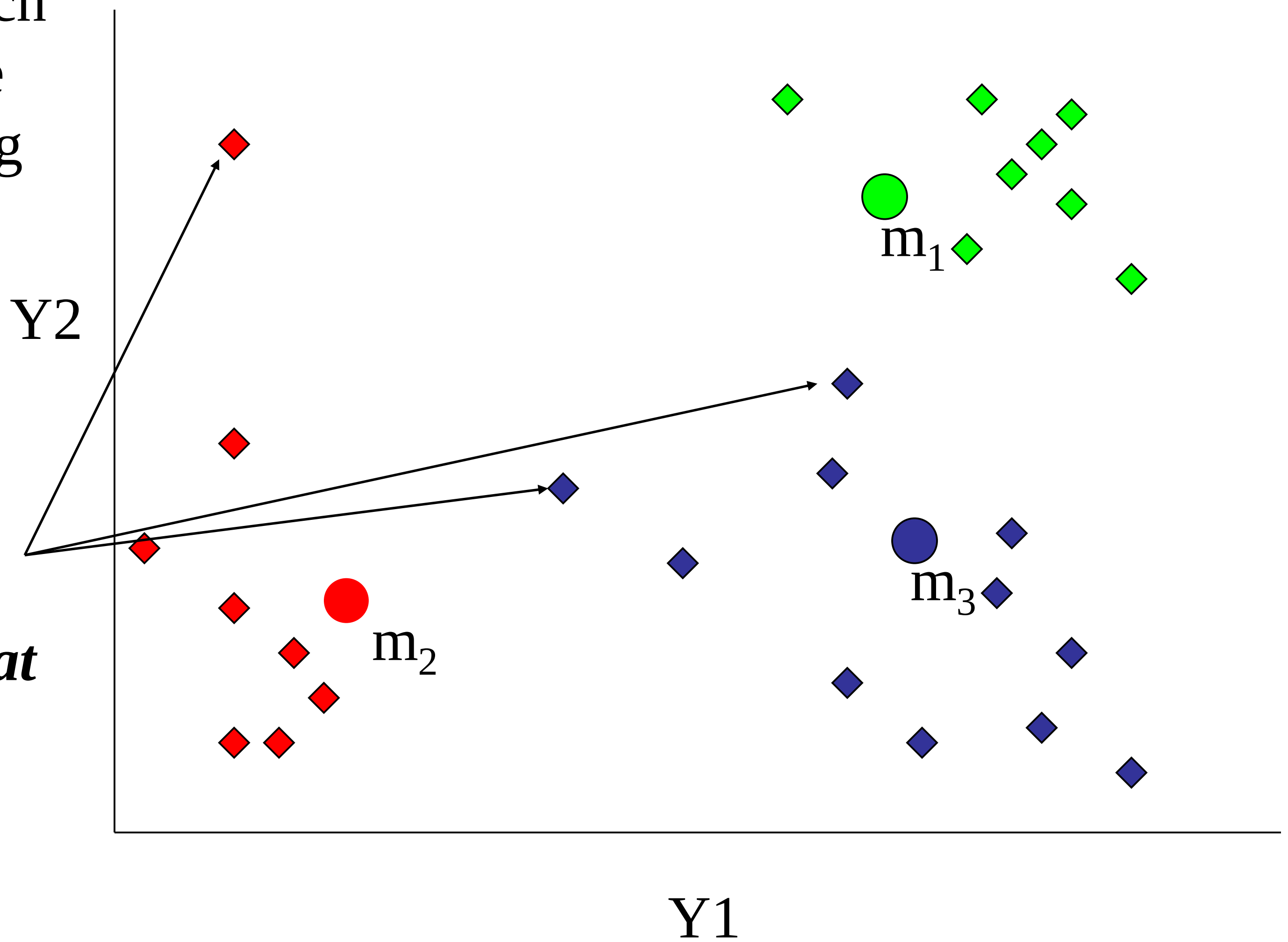
K-means (Example)

The new centers



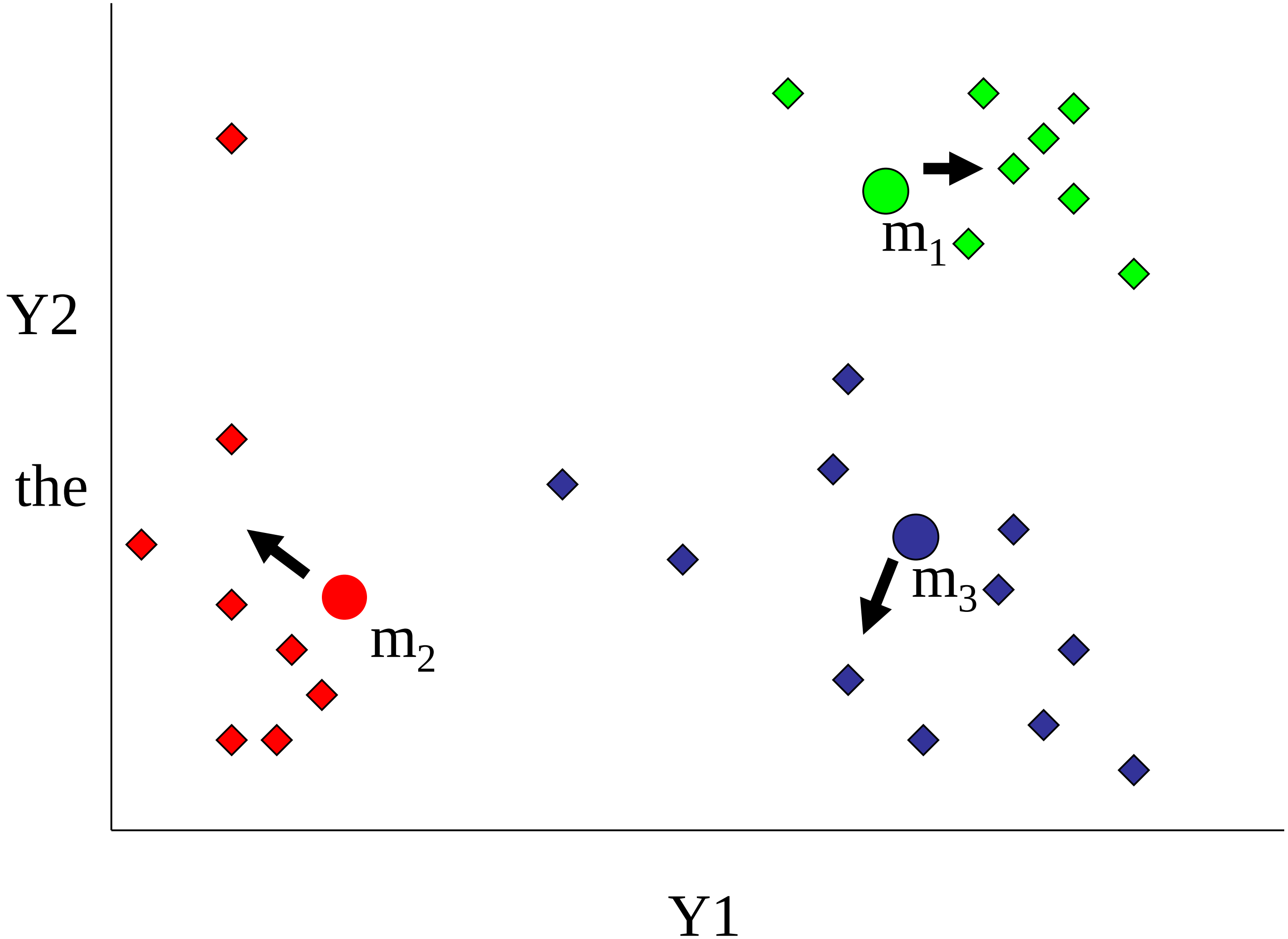
K-means (Example)

Re-assign each sample to the cluster having the closest center.



R : three samples that change the cluster

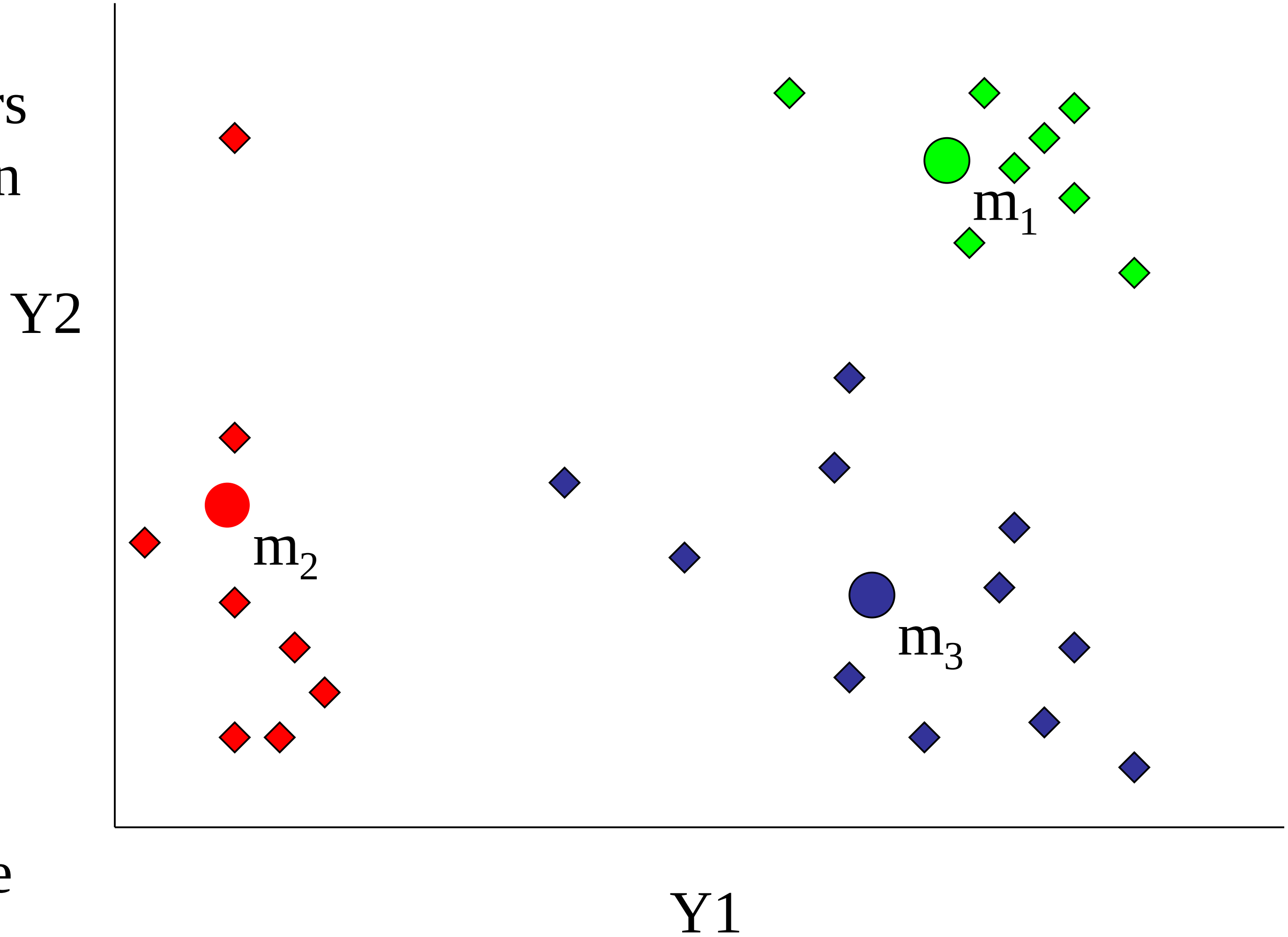
K-means (Example)



Re-compute the mean of clusters

K-means (Example)

Move centers and re-assign samples



No change, then convergence

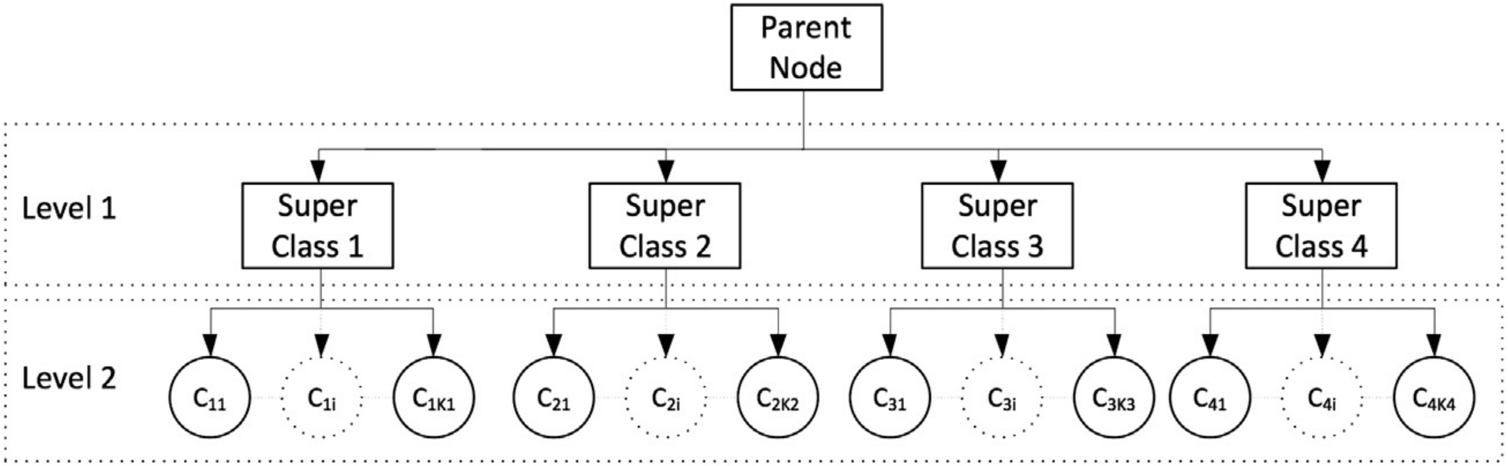
Variational approximation-mixture of pdfs

- Classification: Assign each sample to the closest cluster by using the rule

  x_i *belong the the* k^{th} *cluster if* $k = \min d(x_i, m_k)$

Variational approximation-mixture of pdfs

- Hierarchical mixture and hierarchical EM

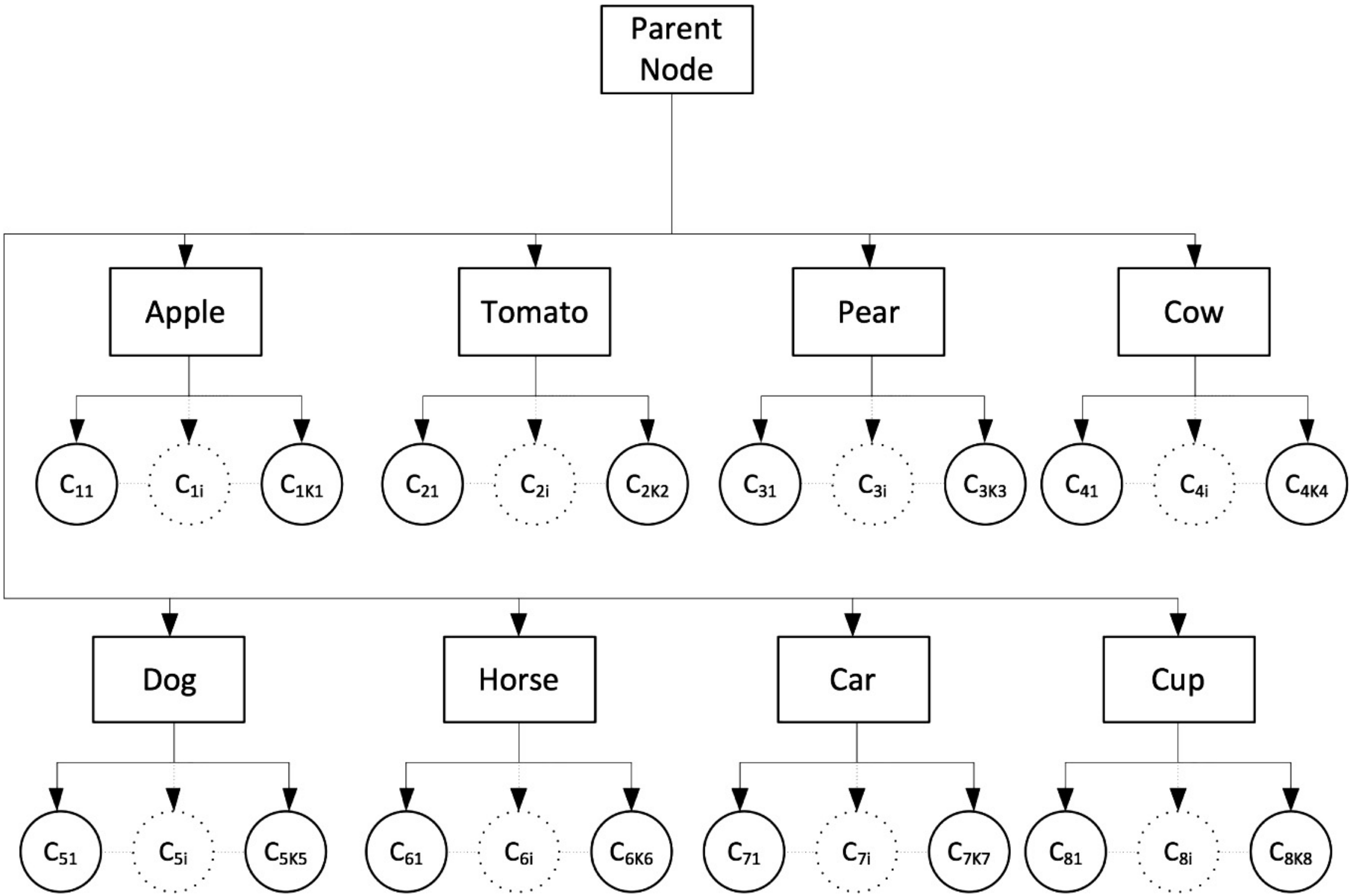


Variational approximation-mixture of pdfs

- Hierarchical mixture and hierarchical EM

Variational approximation-mixture of pdfs

- Hierarchical mixture and hierarchical EM



Variational approximation-mixture of pdfs

Hierarchical EM

- $D=\{X_1, \ldots, X_N\}$ and M the number of superclass
- Superclasses $p(X|\Theta) = \sum_{k=1}^{M} \alpha_k f(X|\theta_k)$ and $\sum_{k=1}^{M} \alpha_k = 1$
- Classes $f(X|\theta_k) = \sum_{j=1}^{M_k} \beta_{kj} f(X|\varphi_{kj})$ and $\sum_{k=1}^{M} \sum_{j=1}^{k_j} \beta_{kj} \quad = 1$
- The parameters $\Theta=(\theta_1, \ldots, \theta_M, \alpha_1, \ldots, \alpha_{M-1})$, where

$\theta_k=(\varphi_{k1}, \ldots, \varphi_{kM_k} \ \beta_{k1}, \ldots, \beta_{kM_k-1})$ for all $k \in \{1, M\}$

How many parameters to estimate?

- Likelihood $\prod_{n=1}^{N} \sum_{k=1}^{M} \alpha_k \sum_{j=1}^{M_k} \beta_{kj}\, f(X_N | \varphi_{kj})$
- EM
  - E-Step $p(k, j|X_n, \theta_k)=\alpha_k \beta_{kj}\, p(X_n| \Theta)\ /p(X_n| \Theta)$ and $p(k)=\sum_{j=1}^{M_j} p(k,j|X_n, \theta_k)$
  - M-Step: compute $\alpha_k, \beta_{kj}, \varphi_{kj}$ for all j and all k

Variational approximation - methods

- Approximation of q(z)
  - accurate approximation of q(z) must leads to a an accurate approximation of the posterior.
  - two strategies for the choice of q(z) are used
    - a variational parameter w is added to q(z) yielding q(z,w) and the optimization is carried out with regards to w.
    - when z is a d-dimensional vector q(z) is factorized; $q(z)=\Pi_{i=1}^{d} q_i(z_i)$, which make possible the approximation.

Local variational approximation

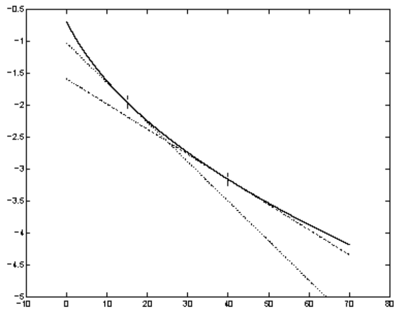
- $q(z)$ is replaced by $q(z,w)$ where w is a variational parameter.
- $q(z,w)$ can be chosen depending on the tackled problem (e.g., a Gaussian with unknown parameter w) or obtained by the approximation of q(z).
- Example, EM revisited
  - let us rewrite

$$p(D\,|\,\theta) = \underbrace{\int q(z,w)\ln\left(\frac{p(D,z\,|\,\Theta)}{q(z,w)}\right)dz}_{F(w,\Theta)} - \underbrace{\int q(z,w)\ln\left(\frac{p(z\,|\,D,\Theta)}{q(z,w)}\right)dz}_{KL(q\|p)}$$

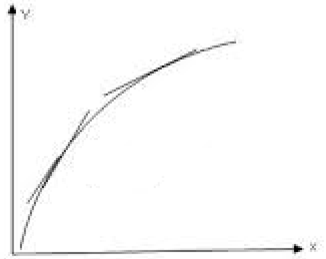
  - in this case, the E-step is $\max_w F(w,\Theta)$ and the M-step is $\max_\Theta F(w,\Theta)$
  - note that $p(D,z\,|\,\Theta)$ is the complete likelihood

Local variational approximation

- In the case of approximations, convex and concave functions properties are often used.
- Convex functions
  - Jensen inequality: $f(\Sigma_{n=1}^{N} a_i x_i) \leq \Sigma_{n=1}^{N} a_i f(x_i)$
  - first order condition: $f(x) \geq L(x,w) = f(w) + (x-w) f'(w)$, where the lower bound L(x,w) is the tangent line at x=w
  - second order condition $\forall x \in \Omega \quad f''(x) \geq 0$
  - x^2, $|x|$, and e^x are convex functions.



- Concave functions
  - change $\leq$ by $\geq$ and $\geq$ by $\leq$ in the above equations
  - x^{-2}, sine and ln are concave.



Local variational approximation

- Example - logistic function,
  - the posterior that an event is affected (z=1) by a variable x is

    $q(z) = p(z=1 | x, a, b) = 1/(1+e^{-(ax+b)})$

    a and b are unknowns
  - to approximate q(z), we just need to approximate

    $\ln f(x) = -\ln(1+e^{-x}) = x/2 - \ln(e^{-x/2}+e^{x/2})$
  - the function $-\ln(e^{-x/2}+e^{x/2})$ is a convex function in x^2
  - then

$$f(x) = -\ln(e^{-x/2} + e^{x/2}) \geq L(x, w) = -0.5w - \ln(1+e^{-w}) + \frac{(x^2 - w^2)}{4w}\tanh(0.5w)$$

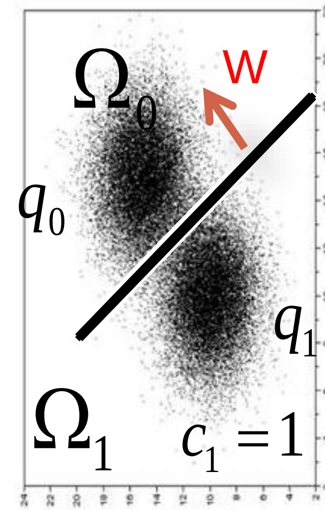
Local variational approximation

- Example of Bayesian logistic regression (BLR, discriminative learning) $$p(W \mid c_0 = 0, c_1 = 1) \propto p(c_0 = 0, c_1 = 1 \mid W)\pi(W)$$
  - the posterior
  - straightforward manipulations leads to

$$p(W \mid c_0 = 0, c_1 = 1) \propto \pi(W) \sum_{x_0 \in \Omega_0, x_1 \in \Omega_1} \prod_{i=0}^{1} p(c_i = i \mid x_i, W) q_i(x_i)$$

$$\pi(W) = N(\mu, \Sigma)$$

$$p(c_i = i \mid x_i, W) = \frac{e^{(2i-1)x_i^T W}}{1 + e^{(2i-1)x_i^T W}}$$



  - the computation of the posterior is intractable.

Local variational approximation

- first approximation (finding a lower bound by the approximation of logistic function)

$$p(c_i = i \mid x_i, W) = \frac{e^{(2i-1)x_i^T W}}{1+e^{(2i-1)x_i^T W}} \geq F(\varepsilon_i) e^{0.5(H_i - \varepsilon_i)\varphi(\varepsilon_i)(H_i^2 - \varepsilon_i^2)} = p(c_i = i \mid x_i, W, \varepsilon_i)$$

$$\varepsilon_i > 0, \quad \varphi(\varepsilon_i) = \frac{\tanh(0.5\varepsilon_i)}{4\varepsilon_i}, \quad F(\varepsilon_i) = \frac{e^{\varepsilon_i}}{1+e^{\varepsilon_i}}, \quad H_i \text{ the Hessian}$$

$$p(W \mid c_0 = 0, c_1 = 1) \geq \pi(W) \sum_{x_0 \in \Omega_0, x_1 \in \Omega_1} \prod_{i=0}^{1} F(\varepsilon_i) e^{0.5(H_i - \varepsilon_i)\varphi(\varepsilon_i)(H_i^2 - \varepsilon_i^2)} q_i(x_i)$$

- second approximation of exp(x) (Jensen inequality)

$$p(W \mid c_0 = 0, c_1 = 1) \geq \pi(W) \prod_{i=0}^{1} F(\varepsilon_i) e^{\sum_{i=0}^{1} 0.5(E_{q_i}(H_i) - \varepsilon_i) - \varphi(\varepsilon_i)(E_{q_i}(H_i^2) - \varepsilon_i^2)}$$

Local variational approximation

- because $\pi(W)$ is a Gaussian, it has been proven that the posterior is a Gaussian with

$$\mu^{post} = \Sigma_{post}(\Sigma^{-1}\mu + \sum_{i=0}^{1}(i-0.5)E_{q_i}(x_i)) \qquad \Sigma_{post}^{-1} = \Sigma^{-1} + 2\sum_{i=0}^{1}\varphi(\varepsilon_i)E_{q_i}(x_i x_i^t)$$

- the optimal variational parameter is obtained by using EM algorithm

$$\varepsilon_i^2 = E_{q_i}(x_i^t \Sigma_{post} x_i) + \mu_{post}^t E_{q_i}(x_i x_i^t)\mu_{post}$$

- the hyperplan is given by

$$W \approx \mu^{post}$$

BLR outperforms SVM and many of its extended versions.

Local variational approximation

- Algorithm
  - compute Σ_{post} and μ_{post}

    Do

    $$\Sigma_{post}^{-1} = \Sigma^{-1} + 2\sum_{i=0}^{1} \varphi(\varepsilon_i) E_{q_i}(x_i x_i^t) \qquad \mu_{post} = \Sigma_{post}(\Sigma^{-1}\mu + \sum_{i=0}^{1}(i-0.5)E_{q_i}(x_i))$$

    For i=0, 1 do $\varepsilon_i^2 = E_{q_i}(x_i^t \Sigma_{post} x_i) + \mu_{post}^t E_{q_i}(x_i x_i^t)\mu_{post}$

    Until converge

    $W \approx \mu_{post}$
  - the computational complexity of this algorithm is dominated by the inversion of the Σ_{post} , which requires around $O(d^3)$ operations at each iteration.
- Initialization
  - compute the parameters of the distributions q0 and q1
  - initialize Σ_{post} to the identity matrix and the mean μ_{post} to a vector with components equal to 1
  - initialize the variational parameters as follows

    For each i=0,1do $\varepsilon_i^2 = E_{q_i}(x_i^t \Sigma_{post} x_i) + \mu_{post}^t E_{q_i}(x_i x_i^t)\mu_{post}$

Factorization

- Let us recall
  - We have $\ln p(D\,|\,\Theta) = \underbrace{\int q(z)\ln(\frac{p(D,z\,|\,\Theta)}{q(z)})dz}_{F(q,\Theta)} \underbrace{- \int q(z)\ln(\frac{p(z\,|\,D,\Theta)}{q(z)})dz}_{KL(q\|p)}$
  - we need to maximize $F(q,\Theta)$ with regards to both q and Θ
- Factorization of q(z); $q(z)=\prod_{i=1}^{d} q_i(z_i)$, which make possible the approximation of F(). Indeed, denoting $q_i(z_i)$ by q_i

$$F(q,\Theta) = \int_{\Omega} \prod_{i=1}^{d} q_i (\ln(p(D,z\,|\,\Theta)) - \sum_{l=1}^{d} \ln(q_l))dz$$

$$F(q,\Theta) = \int_{\Omega} q_j (\int_{\Omega} \ln(p(D,z\,|\,\Theta)) \prod_{i\neq j}^{d} q_i dz_i)dz_j - \int_{\Omega} \sum_{l=1}^{d} \ln(q_l) \prod_{i\neq j}^{d} q_i dz_i)dz_j$$

$$F(q,\Theta) = \int_{\Omega} q_j \int_{\Omega} \ln(p(D,z\,|\,\Theta)) \prod_{i\neq j}^{d} q_i dz_i dz_j - \int_{\Omega} q_j \ln(q_j)dz_j + cste$$

Factorization

$$\ln \tilde{p}(D, z_j \mid \Theta) = E_{\prod_{i \neq j} q_i}(\ln(p(D, z \mid \Theta))) = \int_{\Omega} \ln(p(D, z \mid \Theta)) \prod_{i \neq j}^{d} q_i dz_i$$

$$F(q, \Theta) = \int_{\Omega} q_j \ln \frac{\tilde{p}(D, z_j \mid \Theta)}{q_j} dz_j - \sum_{i \neq j} \int_{\Omega} q_i \ln(q_i) dz_i$$

- because the second term is non negative (entropy) then

$$F(q, \Theta) = \underbrace{-KL(q_j \,\|\, \tilde{p}(D, z_j \mid \Theta))}_{\leq 0} + entropy$$

- For which value of q, F() is maximal?

Factorization

- EM revisited (variational EM)
  - let assume $q(z) = \prod_{i=1}^{d} q(z_i)$, then

$$F(q,\Theta) = \underbrace{-KL(q_j \,\|\, \tilde{p}(D, z_j \mid \Theta))}_{\leq 0} - \sum_{i \neq j} \int q_i \ln q_i dz$$

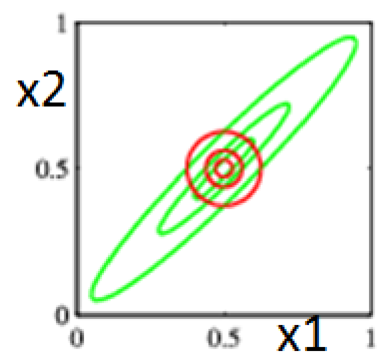
  - $F(q,\Theta)$ is maximal when $q_j^*(z_j) = \tilde{p}(D, z_j \mid \Theta)$
  - after normalization

$$q_j^*(z_j) = \frac{\tilde{p}(D, z_j \mid \Theta)}{\int \tilde{p}(D, z_j \mid \Theta) dz_j}$$

  - E-step: compute $\forall j \quad q_j^*(z_j)$
  - M-step: $\Theta^{t+1} = \arg\max_{\Theta} F(q,\Theta)$

Factorization - Example

- Example of 2D Gaussian factorization
  - the pdf $p(X) = e^{-\frac{1}{2}(X-\mu)^t A^{-1}(X-\mu)} / \sqrt{|2\pi A|}$ where $X=(x_1,x_2)$, $\mu=(\mu_1, \mu_2)$, and
    $$A^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$
  - $\ln q(x_1) = E_{x_2}(\ln(p(X))+cst = -E_{x_2}(0.5a_{11}(x_1-\mu_1)^2 + a_{12}(x_1-\mu_1)(x_2-\mu_2)+cst = 0.5a_{11}(x_1-\mu_1)^2 + a_{12}(x_1-\mu_1)\, E_{x_2}(x_2-\mu_2)+cst$
  - $q(x_1)=G(x_1|m_1,a_{11}^{-1})$, where $m_1=\mu_1 - a_{11}^{-1}a_{12}(E_{x_2}(x_2)-\mu_2)$
  - Idem $q(x_2)=G(x_2|m_2,a_{22}^{-1})$, where $m_2=\mu_2 - a_{22}^{-1}a_{12}(E_{x_1}(x_1)-\mu_1)$
  - because m_1 (resp. m_2) depends on $q(x_2)$ (used in $E_{x_2}(x_2)$) (resp. $q(x_1)$, so the estimation of $q(x_1)$ and $q(x_2)$ is done by using an iterative algorithm.
  - Note that, if $E_{x_2}(x_2)= \mu_2$ and $E_{x_1}(x_1)=\mu_1$, the solution is non iterative.
  - The factorization of three Gaussians



Factorization - Example

- Example of 1D posterior Gaussian factorization
  - the pdf $p(x) = e^{-\frac{1}{2\sigma^2}(x-\mu)^2} / \sqrt{2\pi}\sigma$, $p(\mu | \sigma^2) = G(\mu | \mu_0, \sigma^2 / \lambda_0)$
    $$p(\sigma^2) = gamma(1/\sigma^2 | a_0, b_0)$$
  - the unique used assumption $q(\mu,\sigma)=q(\mu)q(\sigma)$
  - $\ln q(\mu) = E_\sigma(\ln(p(D|\mu,\sigma))+\ln(p(\mu|\sigma))+cst=-E_\sigma(\sigma)(\lambda(\mu-\mu_0)^2+\Sigma_{n=1}^{N}(x_n-\mu)^{2)}+cst= G(\mu|\mu_N,\lambda_N^{-1})$.
  - $\ln q(\sigma) = E_\mu(\ln(p(D|\mu,\sigma))+\ln(p(\mu|\sigma)+\ln(p(\sigma))+cst=Gamma(\sigma|a_N,b_N)$
  - find a_N, b_N
  - Algorithm
    - initialize $E_\mu(\sigma)$
    - estimate iteratively $q(\mu)$, $E_\sigma(\mu)$, $E_\sigma(\mu^2)$, $q(\sigma)$, and $E_\sigma(\mu)$ until convergence
  - The lower band F() should not decrease and it can be used as convergence criteria

Factorization - Example

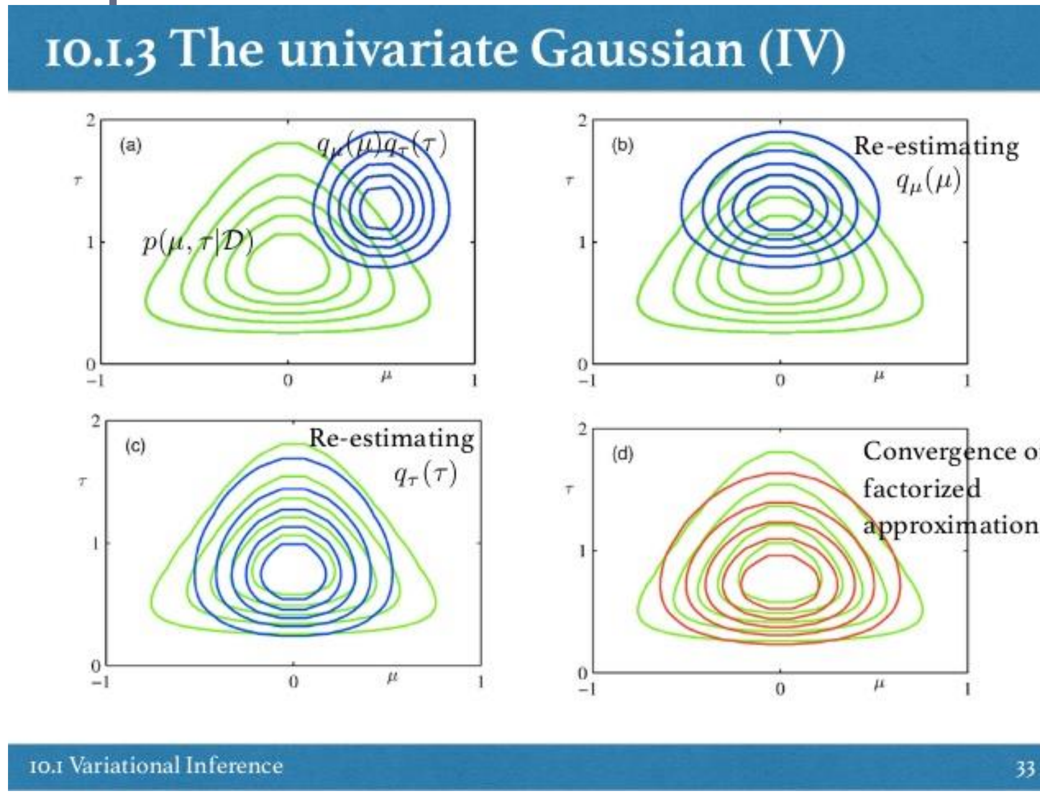


Illustration of variational inference for the mean μ and precision τ of a univariate Gaussian distribution. Contours of the true posterior distribution $p(\mu, \tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

Application to linear regression

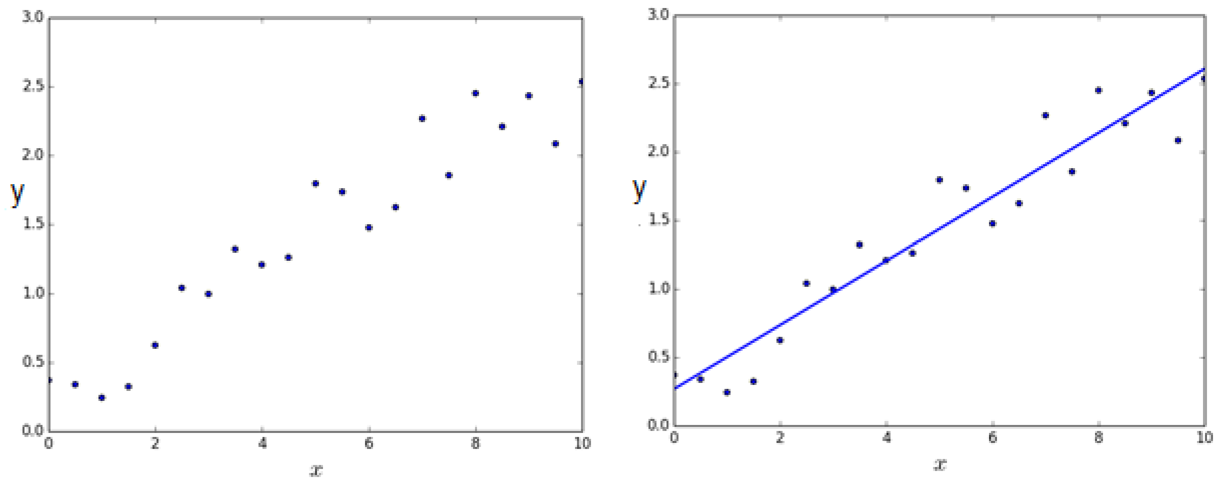
- Le consider the data from the Italian clothing company Benetton
- The company would like to understand the effects of advertising on sales
- Answer: Sales = 168 + 23 Advertising.
  if advertising expenditure is increased by one Euro, then sales will be expected to increase by 23 million Euro
- How they find that?

  linear regression

Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58

Application to linear regression

- Let us consider a random vector x (input) and a random variable y (output)
- Given one input observation of x, we want to predict the value of y
- Often,
  - we assume that $y = w_0 x_0 + w_1 x_1 + \ldots + w_d x_d + \varepsilon$, where $\varepsilon \sim N(\varepsilon | 0, \sigma)$ and $x_0 = 1$

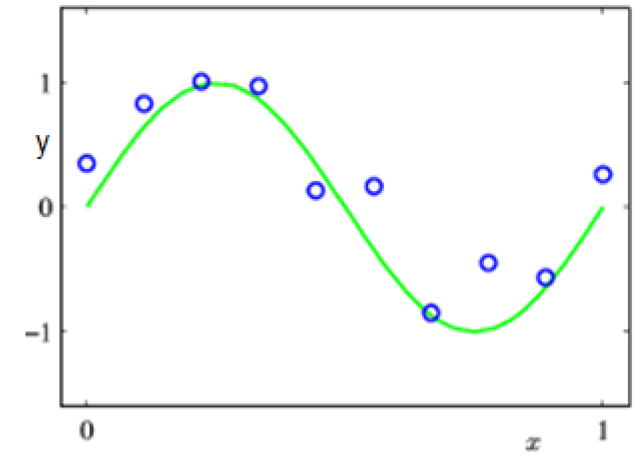


$$y = w_0 + w_1 x_1 + \varepsilon$$

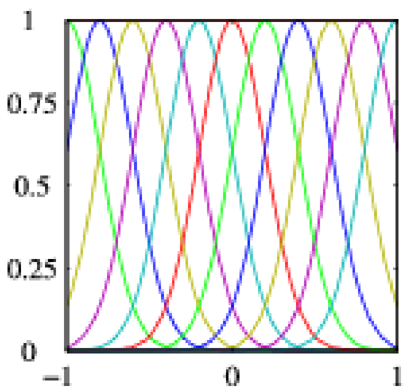
Application to linear regression

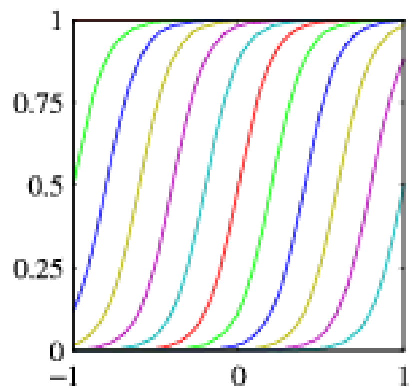
- we can do better by using basis functions

$y = w_0\,\varphi_0(x) + w_1\,\varphi_1(x) + \ldots + w_d\,\varphi_d(x) + \varepsilon$



- examples of $\varphi_k(x) = e^{-\frac{(x-\mu_k)^2}{s^2}}$, $\varphi_k(x) = \frac{1}{1+e^{-\frac{(x-\mu_k)}{s}}}$, ...





Application to linear regression

- To find the posterior pdf p(w | y, $\sigma, \ldots$), we need
  - the prior p(w | …)
  - the conditional pdf p(y | x)
  - the conditional pdf p(y)

- Priors
  - p(w | α) $= \prod_{m=1}^{d} N(w_m|0,\alpha)$
  - β a constantTapez une équation ici.

- Conditional pdf
  - rewrite y= φ^tw+ ε, where $\varphi = (\varphi_0(\mathrm{x}), \ldots, \varphi_d(\mathrm{x}))$ and $w = (w_0, \ldots, w_d)$
  - p(y | w,β) $= N(y|\varphi^t$w,β)

Application-linear regression

- The marginal pdf

  $$p(y\,|\,\beta,\alpha)=\int p(w|\beta)p(y|w,\beta)dw = N(y\,|\,0,\beta^{-1}I+\alpha^{-1}\varphi\varphi^{t})$$

- Posterior pdf
  - Let us consider the data $D=X \times Y=\{(x_1,y_1),\ldots,(x_N,y_N)\}$

$$p(w|D,\alpha,\beta)=\frac{p(w|\alpha)\prod_{n=1}^{N}p(y_n|w,\beta)}{\int \; p(w|\alpha)\prod_{n=1}^{N}p(y_n|w,\beta)dw}=N(w\,|\,\mu,\Sigma)$$

where $\mu=\beta\Sigma\phi^{t}Y$ and $\Sigma=(\beta\phi^{t}\phi+\alpha I)^{-1}$

$$\phi=(\varphi_1,\ldots,\varphi_d) \text{ and } \varphi_k=(\varphi_k(x_1),\ldots,\varphi_k(x_N))$$

- The posterior mean μ is an estimator of w.
- Because ϕ and Y are known, then knowing α and β, we can estimate μ and Σ

Application-linear regression

- To estimate μ and Σ, we need to find (w is replaced by μ)

  Q(α,β, $\alpha^{(t)}$, $\beta^{(t)}$)=$E_{p(w|\alpha^{(t)},\beta^{(t)})}$(ln(p(w, D | α,β))=

$\frac{N}{2}\ln(\beta)-\frac{\beta}{2}(||\text{Y-}\ \phi\mu^{(t)}||^2+\text{tr}(\phi^t\Sigma^{(t)}\phi))+\frac{d}{2}\ln(\alpha)-\frac{\alpha}{2}(||\mu^{(t)}||^2+\text{tr}(\Sigma^{(t)}))+cte$

- EM algorithm
  - E-step Compute $\mu^{(t)}$, $\Sigma^{(t)}$, and Q(α,β, $\alpha^{(t)}$, $\beta^{(t)}$)
  - M-step $(\alpha^{(t+1)},\beta^{(t+1)})=argmax_{\alpha,\beta}\ \alpha^{(t)},\beta^{(t)}$
- In E-step, the expected value of the logarithm of the complete likelihood is used instead of the posterior, why?
- What is the computational complexity of this algorithm?

Application to linear regression

- Let now use variational approximation
- Priors $p(w \mid \alpha) = \prod_{m=1}^{d} N(w_m|0,\alpha)$

  $p(\alpha|a,b) = \prod_{m=1}^{d} Gamma(\alpha_m|a,b)$

  $p(\beta|c,h) = \mathrm{Gamma}(\beta|c,h)$
- Posterior $p(w, \alpha, \beta \mid D, a, b, c, d) =$

$$\frac{p(w|\alpha)p(\alpha)p(\beta)\prod_{n=1}^{N} \quad p(y_n|w,\beta)}{\int p(w|\alpha)p(\alpha)p(\beta)\prod_{n=1}^{N} \quad p(y_n|w,\beta)d\alpha d\beta}$$

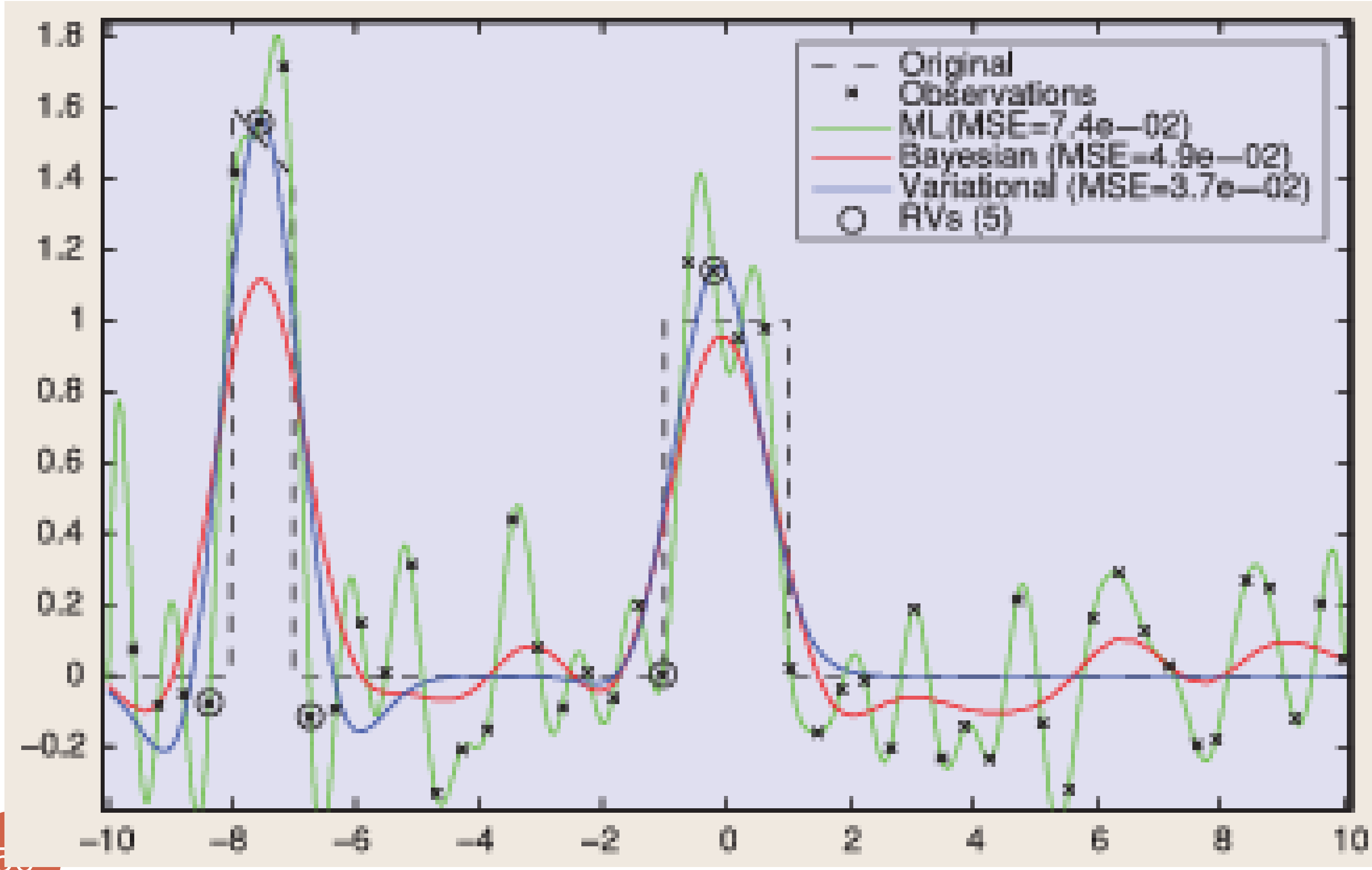
- Factorization $p(w, \alpha, \beta \mid D, a, b, c, h) \simeq q(w)\, q(\alpha)\, q(\beta)$
- Straightforward manipulations

$q(w) = N(w|\mu,\Sigma)$, $q(\alpha) = \prod_{m=1}^{d} Gamma(\alpha_m \mid a+0.5, b+0.5\, E(w_m^2))$,
$q(\beta) = Gamma(\beta, c+0.5N, h+0.5\,||Y-\phi w||^2)$

$\mu = E(\beta)\Sigma\phi^t Y$ and $\Sigma = (E(\beta)\phi^t\phi + E(A))^{-1}$ and $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$

Application to linear regression

- EM
  - E-step: estimate $q^{(t)}(w), q^{(t)}(\alpha), q^{(t)}(\beta)$
  - M-step: $(\alpha^{(t+1)}, \beta^{(t+1)}) = argmax_{(\alpha,\beta)} \mathrm{F}(q^{(t)}, \alpha, \beta)$



$\varphi(x)$ is a Gaussian kernel
N=50
a=b=0
c=h=10^{-6}

Application to linear regression

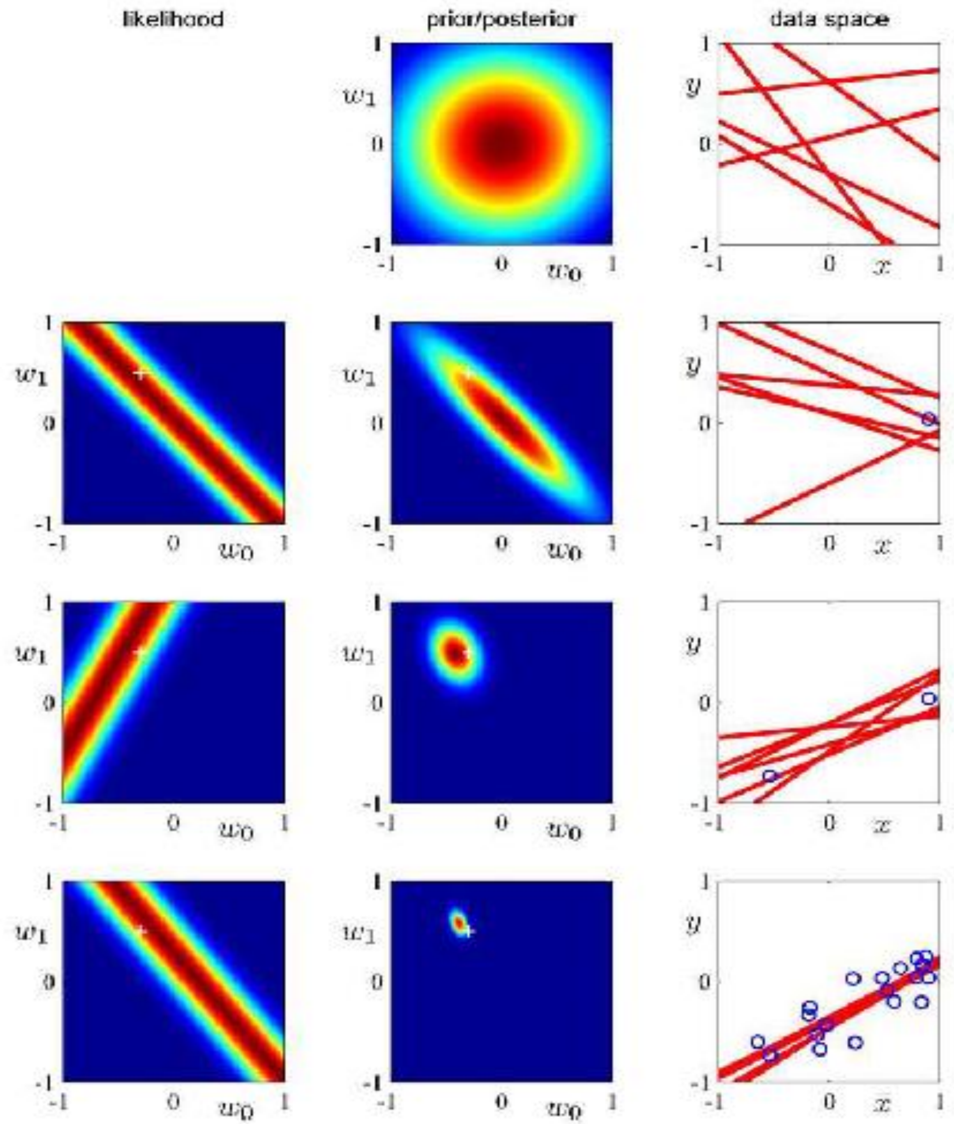


Illustration of sequential Bayesian learning for a simple linear model of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. A detailed description of this figure is given in the text.

High dimensional data

52

High-dimensional data

- Image colors: 256^3 features
- Faces: *128*x*128* pixels= *16384* features/face
- Text: number of terms in a corpus ~*10 000*
- ….

Texte


Visages

High-dimensional data

- Consequences
  - accuracy: noisy features reduces the performance of models
  - more data
  - computation: floating point overflow, matrices inversion, …
- Solutions
  - dimension reduction
  - feature selection
  - …

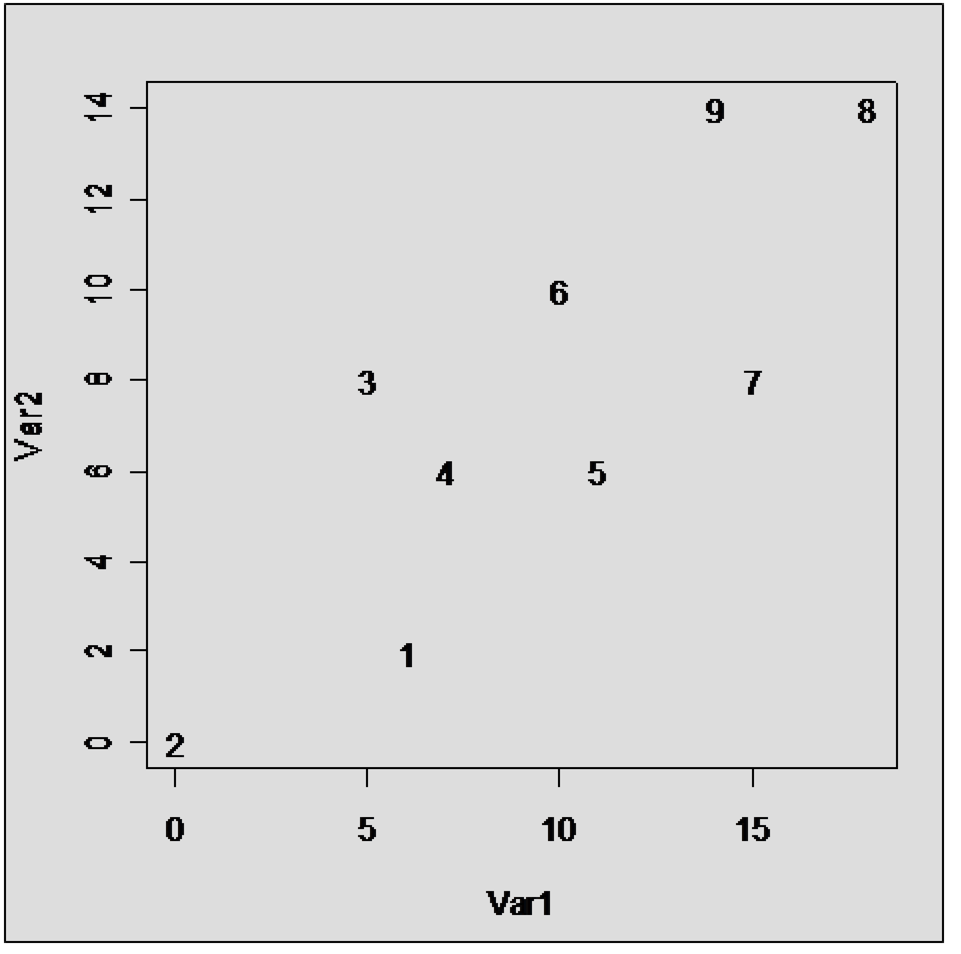
Dimension reduction

- Data projection on a subspace
- Principal Component Analysis (PCA) is a well-known method for analyzing data in statistics and experimental sciences.
- Consists of looking for the directions of space that best represent the correlations between samples.
- It is used for:
  - A reduction of the dimension of the characteristics to a reduced dimension.
  - A selection of features
  - An interpretation and analysis of correlations between the data.
  - Visualization of data in a 2 or 3 dimensional space.

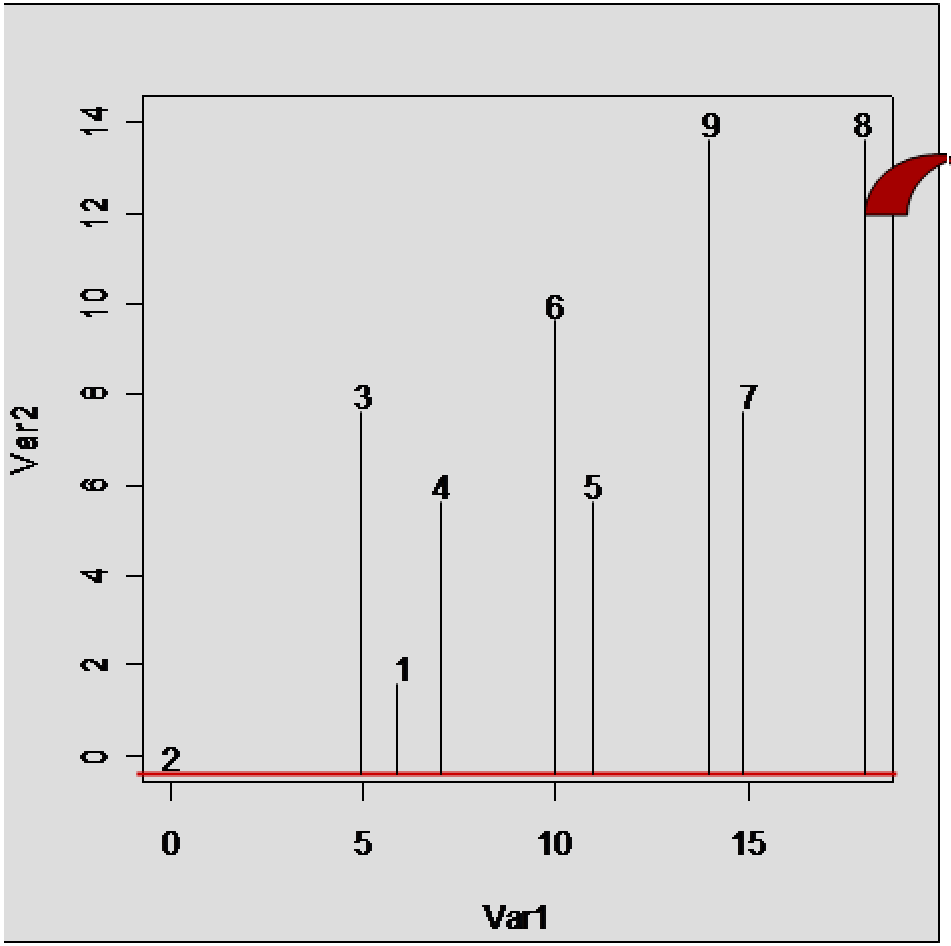
Dimension reduction- Example

Data

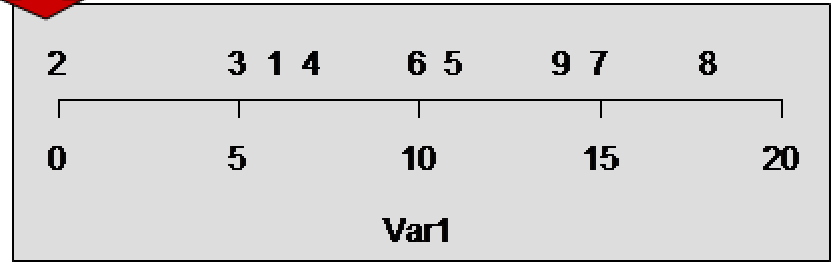
Station	Var1	Var2
1	6	2
2	0	0
3	5	8
4	7	6
5	11	6
6	10	10
7	15	8
8	18	14
9	14	14



Dimension reduction - Example



Solution 1: elimination of a variable (e.g., Var 2)



Bad solution: loss of information;
7 and 9 too close, 9 should be closer to 8 than to 7….

Solution 2: elimination of Var1 leads to loss of information

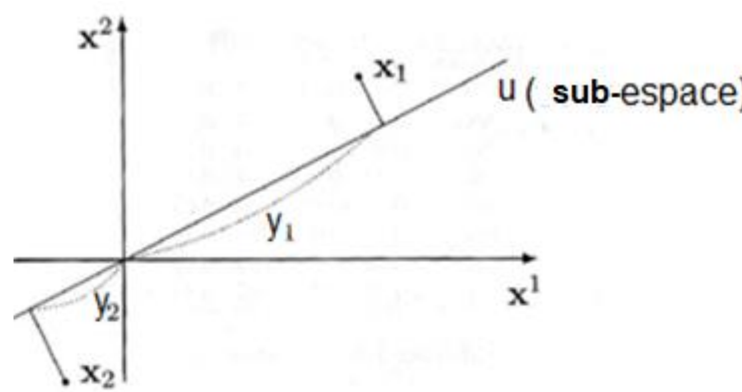
Dimension reduction

- Example
  - D={(85,80), (40,80), …, (95,80)}
  - D={(1,1), (2,2), …, (N, N)}
- Nuisance variables increase the computational times and leads to the increase of errors (classification, recognition, estimation…)

Dimension reduction-Formulation

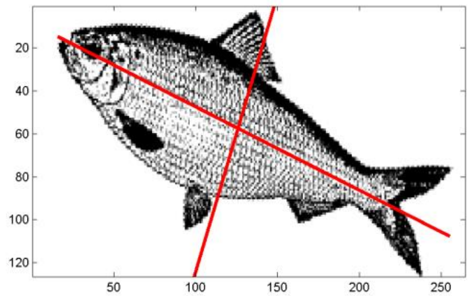
- Data $D = \{x_1, \cdots, x_N\}$, where $x_i \in \Re^d$
- Arithmetic mean $\mu = \sum_{n=1}^{N} x_n / N$
- Covariance $C = \sum_{n=1}^{N} (x_n - \mu)(x_n - \mu)^t / N$
- Variance $v = trace(C) = \sum_{n=1}^{N} \sum_{i=1}^{d} (x_{ni} - \mu_i)^2$
- The projection of x_n on the unitary vector u is
- The variance of the data project on u is $y_n = u^t x_n$

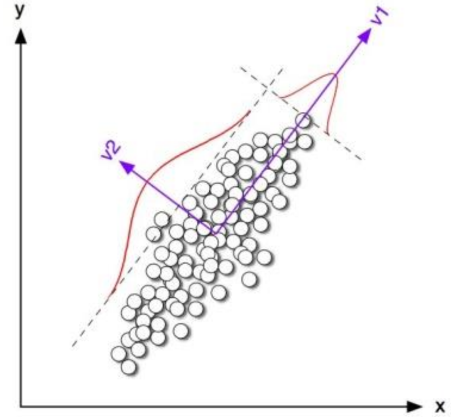
$$v_u = \sum_{n=1}^{N} \sum_{i=1}^{d} (u^t x_{ni} - u^t \mu_i)^2 = u^t C u$$

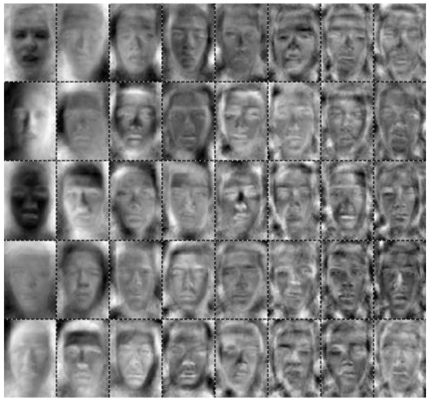


Dimension reduction-Formulation

- Let consider a subspace spanned by $B^r = \{u_1, \cdots, u_r\} \quad r \leq d$







- The variance of the data projected on B^r

$$v_B = \sum_{j=1}^{r} u_j^t C u_j$$

- How to find B^r ?
  - Yielding the highest variance under the constraints

$$\max_{u_1, \cdots, u_r} v_B = \sum_{j=1}^{r} u_j^t C u_j \qquad u_j^t u_j = 1, \quad \forall j \in \{1, r\}$$

Dimension reduction-Formulation

- Lagrange formulation

$$\max_{u_1,\cdots,u_d} \varphi = \sum_{j=1}^{d} u_j^t C u_j - 2\sum_{j=1}^{d} \lambda_j (u_j^t u_j - 1)$$

- First order condition leads to $\varphi_{u_k} = Cu_k - \lambda_k u_k = 0$

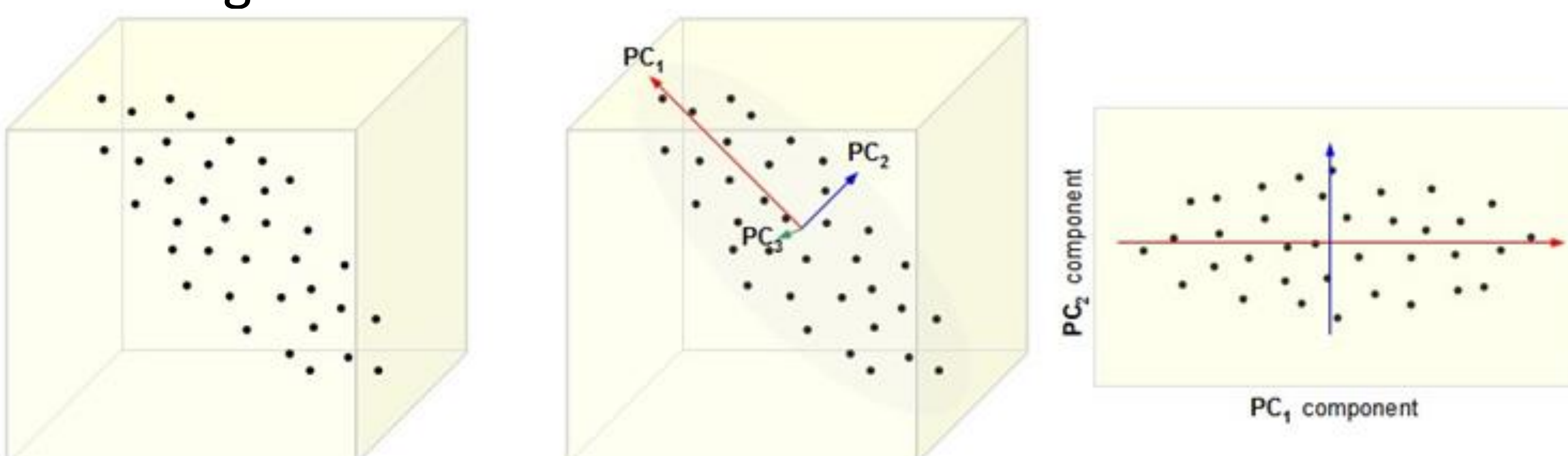
- The vector u_k is a eigenvector of C. Which one?

Dimension reduction-Formulation

- The objective function can be rewritten $v_B = \lambda_k + \sum_{j \neq k}^{d} u_j^t C u_j$
- Repeating the reasoning for each vector $v_B = \lambda_1 + \cdots + \lambda_k + \lambda_d$
- Ranking $\lambda_{(1)} \geq \cdots \geq \lambda_{(d)}$ and choose the r highest eigenvalues. $\lambda_{(1)} \geq \cdots \geq \lambda_{(r)}$
- The percentage of the variance of projected data

$$(\lambda_{(1)} + \cdots \lambda_{(r)}) / v_B$$

- The vector u_k is the eigenvector of C having the k^{th} highest eigenvalue.



Dimension reduction-Formulation

- Algorithm
  - Input: Data D and the dimension r of the subspace
  - Output : Projected data
  - Estimated the covariance of D
  - Compute the highest *r* eigenvalues and the corresponding eigenvectors
  - Project the data ($y_{nj} = u_j^t x_n, \quad \forall j \in \{1, \cdots, r\}$)

- Computational complexity of the algorithm O(Nd^2)

Dimension reduction-Formulation

- The variables y_j are not correlated.

$$y_{jn}^t y_{in} = [0 \ldots u_j^t x_n 0 \ldots][0 .. u_i^t x_n 0 \ldots]^t = 0, \quad \forall i \neq j$$

- PCA consists in transforming the correlated initial variables x_j into new uncorrelated variables y_j of maximum variance, called the principal components.

- y_j a linear combinations of x_j

- x can be reconstructed from y_j

$$x = \sum_{j=1}^{r} y_j u_j^t$$

Dimension reduction-Example

- Data table containing 57 brands of Bottles of water described by 11 variables (G. Govaert, Data Analysis)

- Data provided on the labels of bottles.
- Pays=Country
- M=mineral, S=source
- P=Still water, S=sparkling water

		Pays	Type	PG	CA	MG	NA	K	SUL	NO3	HCO3	CL
1	Evian	F	M	P	78	24	5	1	10	3,8	357	4,5
2	Montagne des pynénées	F	S	P	48	11	34	1	16	4	183	50
3	Cristaline	F	S	P	71	5,5	11,2	3,2	5	1	250	20
4	Fiée des lois	F	S	P	89	31	17	2	47	0	360	28
5	Volcania	F	S	P	4,1	1,7	2,7	0,9	1,1	0,8	25,8	0,9
6	Saint Diéry	F	M	G	85	80	385	65	25	1,9	1350	285
7	Luchon	F	M	P	26,5	1	0,8	0,2	8,2	1,8	78,1	2,3
8	Volvic	F	M	P	9,9	6,1	9,4	5,7	6,9	6,3	65,3	8,4
9	Alpes	F	S	P	63	10,2	1,4	0,4	51,3	2	173,2	1
10	Orée du bois	F	M	P	234	70	43	9	635	1	292	62
11	Arvie	F	M	G	170	90	650	130	31	0	2195	387
12	Roche des Ecrins	F	S	P	63	10,2	1,4	4	51,3	2	173,2	10
13	Ondine	F	S	P	46,1	4,3	6,3	3,5	9	0	163,5	3,5
14	Thonton	F	M	P	108	14	3	1	13	12	350	9
15	Aix des bains	F	M	P	84	23	2	1	27	0,2	341	3
16	Contrex	F	M	P	486	84	9,1	3,2	1187	2,7	403	8,6
17	La bondoire	F	S	P	86	3	17	1	7	19	256	21
18	Dax	F	M	P	125	30,1	126	19,4	365	0	164,7	156
19	Quézac	F	M	G	241	95	255	49,7	143	1	1685,4	38
20	Salvetat	F	M	G	253	11	7	3	25	1	820	4
21	Stamna	GRC	M	P	48,1	9,2	12,6	0,4	9,6	0	173,3	21,3
22	Iolh	GR	M	P	54,1	31,5	8,2	0,8	15	6,2	267,5	13,5
23	Avra	GR	M	P	110,8	9,9	8,4	0,7	39,7	35,6	308,8	8
24	Rouvas	GRC	M	P	25,7	10,7	8	0,4	9,6	3,1	117,2	12,4
25	Alisea	IT	M	P	12,3	2,6	2,5	0,6	10,1	2,5	41,6	0,9
26	San Benedetto	IT	M	P	46	28	6,8	1	5,8	6,6	287	2,4
...	...	...	...	...	...	...	...	...	...	...	...	...
57	Montclar	F	S	P	41	3	2	0	2	3	134	3

Dimension reduction-Example

- Studied variables and correlation between them

	CA	MG	NA	K	SUL	NO3	HCO3	CL
CA	1	0.7041	0.1179	0.1246	0.9131	-0.0634	0.1349	0.2761
MG	0.7041	1	0.6058	0.6561	0.6076	-0.2123	0.6179	0.4793
NA	0.1179	0.6058	1	0.8361	0.0643	-0.1162	0.8562	0.5872
K	0.1246	0.6561	0.8361	1	-0.0259	-0.1668	0.8813	0.3997
SUL	0.9131	0.6076	0.0643	-0.0259	1	-0.1565	-0.0691	0.3176
NO3	-0.0634	-0.2123	-0.1162	-0.1668	-0.1565	1	-0.0604	-0.1205
HCO3	0.1349	0.6179	0.8562	0.8813	-0.0691	-0.0604	1	0.1902
CL	0.2761	0.4793	0.5872	0.3997	0.3176	-0.1205	0.1902	1

Dimension reduction-Example

- Eigenvalues

Numéro	Valeur propore	%	% cumulé
lambda_1	3.8126	47.6577	47.6577
lambda_2	2.0701	25.8765	73.5342
lambda_3	0.9729	12.1614	85.6957
lambda_4	0.7969	9.9615	95.6572
lambda_5	0.1781	2.2257	97.8829
lambda_6	0.095	1.1872	99.0701
lambda_7	0.074	0.9255	100.00
lambda_8	0.0003	0.0044	100.00

- Eigenvectors

Variables	1	2	3	4	5	6	7	8
CA	0.2819	0.5392	-0.1729	-0.1984	0.0098	-0.3169	0.5818	0.3486
MG	0.4657	0.1784	-0.0381	-0.1668	-0.4603	0.6801	-0.166	0.1418
NA	0.4378	-0.2887	-0.0339	0.1729	0.6219	0.0874	-0.1703	0.5201
K	0.4271	-0.3199	0.0056	-0.1189	-0.4522	-0.6297	-0.3128	0.047
SUL	0.2309	0.6026	-0.0316	-0.0324	0.3339	-0.1471	-0.5361	-0.4012
NO3	-0.12	-0.062	-0.9714	0.1483	-0.0632	0.0126	-0.1095	-0.0085
HCO3	0.4013	-0.3473	-0.1328	-0.3486	0.2403	0.1052	0.3834	-0.6028
CL	0.3175	0.0653	0.0725	0.8627	-0.1532	-0.0109	0.2463	-0.2474

Dimension reduction-Example

Which subspace?

Total of eigenvalues =7.9999

% of the variance for 1st variable: 3.8126/7.9999= 0.48 (48%)

% of the variance for 2nd variable: (3.8126+2.0701)/7.9999= 0.74 (74%)

% of the variance for 3rd variable: 0.86 (86%)

% of the variance for 4rd variable: 0.96 (96%)

% of the variance for 5th variable: 0.98 (98%)

….

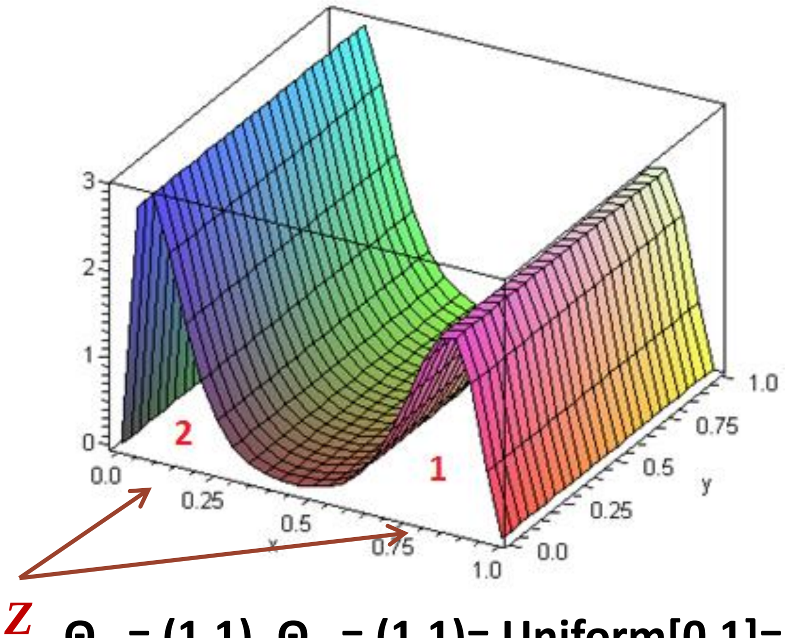
Feature selection

- Useful even for low-dimensional data
- Reduction of the contribution of irrelevant features
- Relevance is not binary
- Related work
  - many approaches for supervised learning, but strong assumptions such that independent features and Gaussian features
  - few approaches for unsupervised generative learning

Feature selection

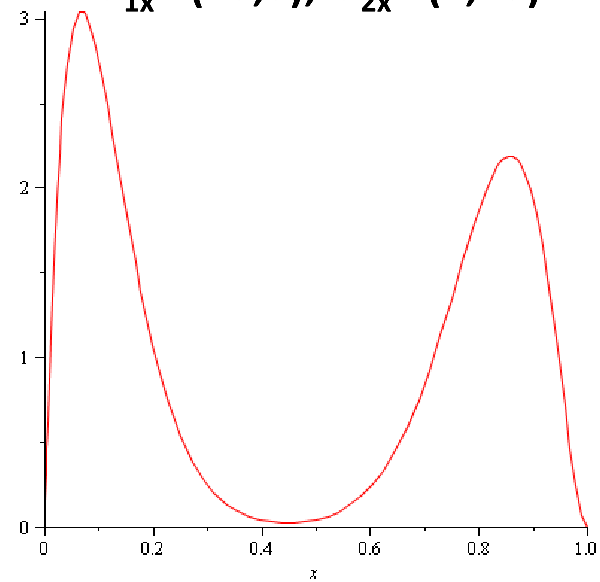
- Let us consider the mixture $p(Y_i|\Theta)=\Sigma_{j=1}^{M} p_j \Pi_{l=1}^{d} p(Y_{il}|\theta_{jl})$
- For complete data $p(Y_i,Z_i|\Theta)=\Pi_{j=1}^{M}(\Pi_{l=1}^{d} p(Y_{il}|\theta_{jl}))^{Z_{ij}}$
- Relevance criteria: independence of *Y* from class labels *Z*

$$\forall n,m,j \in \{1,\ldots,M\}\ \ KL(p(Y_{il}|\ \theta_{nl}),\ p(Y_{il}|\ \theta_{ml}))\approx 0 \Rightarrow\ \theta_{jl}=\xi_l$$



Θ_{1y}= (1,1), Θ_{2y}= (1,1)= Uniform[0,1]=ξ_y
KL(Θ_{1y}, Θ_{2y}) = 0

Θ_{1x}= (13,3), Θ_{2x}= (2,15)



$KL(p(Y_{il}|\ \theta_{1x}),\ p(Y_{il}|\ \theta_{2x}))\neq 0$

Feature selection

- Label Y with hidden Bernoulli variable ϕ_y : $\phi_y = 0$ if $Y \sim \xi_y$ and one if $Y \sim \Theta_y$

  $p(Y_{il} \mid \theta_{nl}, \phi_{il}) = p(Y_{il} \mid \theta_{nl})^{\phi_{il}} \, p(Y_{il} \mid \xi_l)^{1-\phi_{il}}$
- Instead of uniform pdf, let us consider that ξ_y is a model for a mixture of K pdfs with hidden multinomial variables $W = (W_{y1}, W_{y2}, \ldots, W_{yK})$
- It follows that

$$p(\vec{X}_i|\vec{Z}_i,\Theta^*) \simeq p(\vec{X}_i|\vec{Z}_i,\vec{\phi}_i,\vec{W}_i) = \prod_{j=1}^{M}\left[\prod_{l=1}^{d} p(X_{il}|\theta_{jl})^{\phi_{il}}\left(\prod_{k=1}^{K} p(X_{il}|\xi_{kl})^{W_{ilk}}\right)^{1-\phi_{il}}\right]^{Z_{ij}}$$

- New mixture including the feature selection is obtained by marginalization over Z, ϕ, and W.

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} p_j \prod_{l=1}^{d}\left(\epsilon_{l_1} p(X_{il}|\theta_{jl}) + \epsilon_{l_2}\sum_{k=1}^{K}\eta_{kl}\, p(X_{il}|\xi_{kl})\right)$$

References

- Finite and Infinite mixture
  - G. McLachlan and D. Peel. Finite Mixture Models. Wiley, 2000
  - C.M. Bishop. Pattern Recognition and Machine Learning. Springer 2006.
  - W. Fan, N. Bouguila, D. Ziou. Variational Learning for Finite Dirichlet Mixture Models and Applications, IEEE TNN, 2012.
  - M. Cordunneanu and C.M. Bishop. Variational Bayesian Model Selection for Mixture Distribution. Proc. Int. Conf. on Artificial Intelligence and Statistics, 2001.
- Feature selection
  - S. Boutemedjet, N. Bouguila, D. Ziou: A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering. IEEE TPAMI, 31(8): 1429-1443, 2009.
- Discriminative learning
  - R. Ksantini, D. Ziou, B. Colin, F. Dubeau: Weighted Pseudometric Discriminatory Power Improvement Using a Bayesian Logistic Regression Model Based on a Variational Method. IEEE TPAMI, 30(2): 253-266, 2008.
- Graphical models
  - M.J. Jordan et al. An Introduction to Variational Methods for Graphical Models. Machine Learning 37, 183–233,1999

References

- Model selection

  - N. Bouguila and D. Ziou. High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, pp. 1716 - 1731 2007.

- Maximum likelihood

  - In Jae Myung, Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology 47, pp. 90–100, 2003.

- Hierarchical mixture

  - T. Bdiri, N. Bouguila, and D. Ziou. Object clustering and recognition using multi-finite mixtures for semantic classes and hierarchy modeling. Expert Systems with Applications 41, pp.1218–1235, 2014.

Merci!