Computer vision: models, learning and inference

Chapter 7

Modeling Complex Densities

Structure

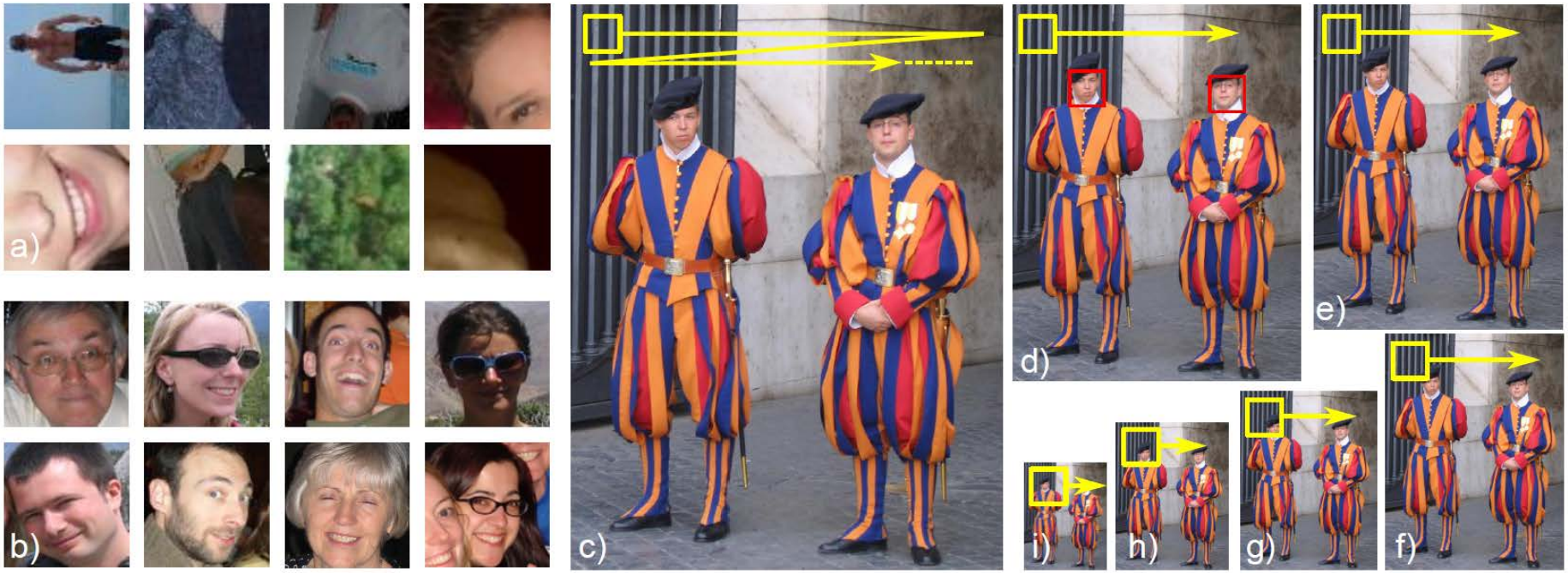
- Densities for classification
- Models with hidden variables
  - Mixture of Gaussians
  - t-distributions
  - Factor analysis
- EM algorithm in detail
- Applications

Models for machine vision

	Model $Pr(w\|x)$	Model $Pr(x\|w)$
Regression $x \in [-\infty, \infty], w \in [-\infty, \infty]$	Linear regression	Linear regression
Classification $x \in [-\infty, \infty], w \in \{0, 1\}$	Logistic regression	Probability density function

Table 5.1: Example models in this chapter. These can be categorized into those that are based on modelling probability density functions, those that are based on linear regression and those that are based on logistic regression.

Face Detection



Type 3: Pr(**x**|**w**) - Generative

How to model Pr(**x**|**w**)?

- Choose an appropriate form for Pr(**x**)
- Make parameters a function of **w**
- Function takes parameters θ that define its shape

Learning algorithm: learn parameters θ from training data **x**,**w**

Inference algorithm: Define prior Pr(**w**) and then compute Pr(**w**|**x**) using Bayes' rule

$$Pr(w=1|\mathbf{x}) = \frac{Pr(\mathbf{x}|w=1)Pr(w=1)}{\sum_{k=0}^{1} Pr(\mathbf{x}|w=k)Pr(w=k)}$$

Classification Model

$$Pr(\mathbf{x}|w) = \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w]$$

Or writing in terms of class conditional density functions

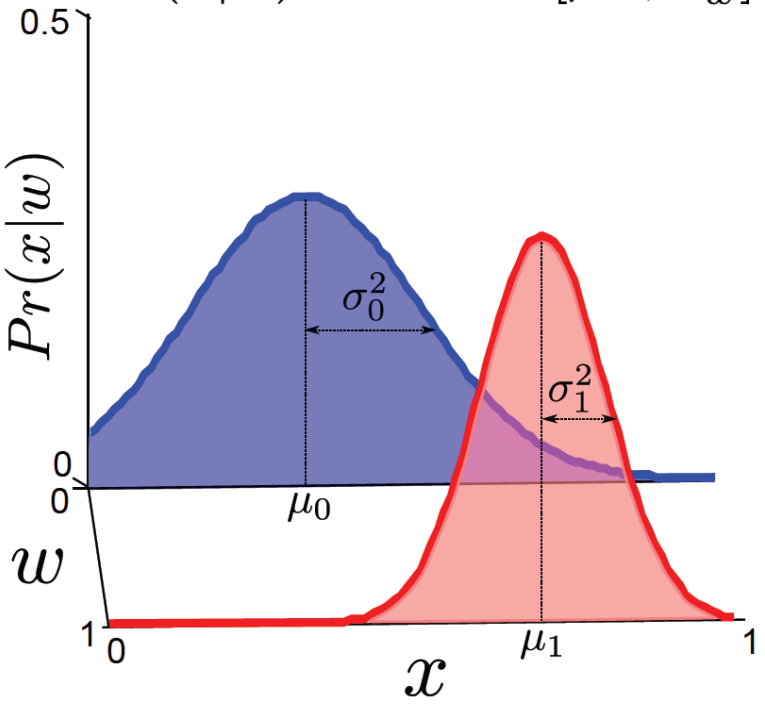
$$\begin{aligned} Pr(\mathbf{x}|w=0) &= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0] \\ Pr(\mathbf{x}|w=1) &= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1] \end{aligned}$$

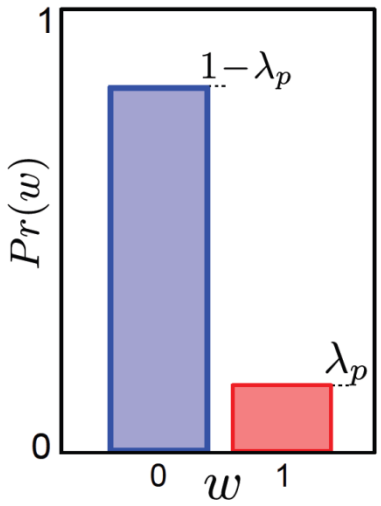
Parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$ learnt just from data $\mathcal{S}_0$ where w=0

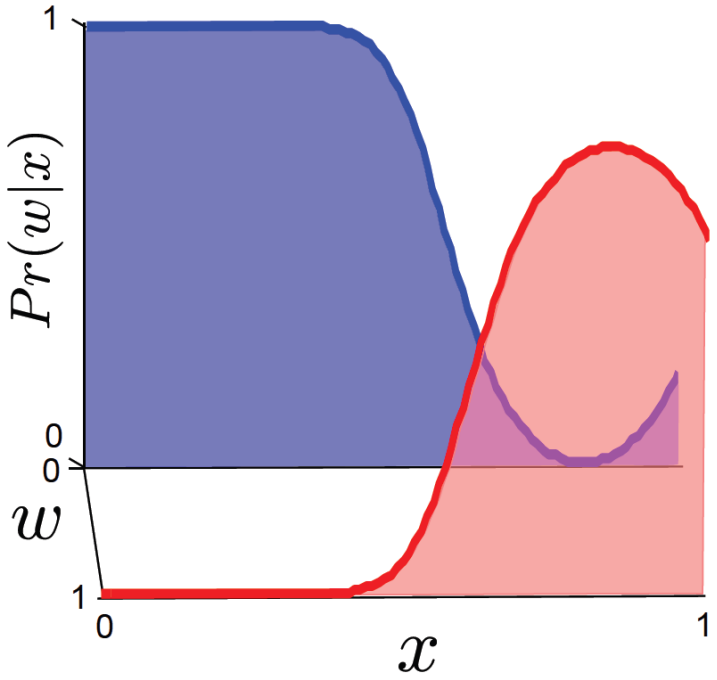
$$\begin{aligned} \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0 &= \underset{\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0}{\text{argmax}} \left[\prod_{i \in \mathcal{S}_0} Pr(\mathbf{x}_i|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \right] \\ &= \underset{\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0}{\text{argmax}} \left[\prod_{i \in \mathcal{S}_0} \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0] \right] \end{aligned}$$

Similarly, parameters μ_1, Σ_1 learnt just from data $\mathcal{S}_1$ where w=1

$$Pr(x|w) = \text{Norm}_x[\mu_w, \sigma_w^2]$$







Inference algorithm: Define prior Pr(**w**) and then compute Pr(**w**|**x**) using Bayes' rule

$$Pr(\mathbf{w}|\mathbf{x}) = \frac{Pr(\mathbf{x}|\mathbf{w})Pr(\mathbf{w})}{\int Pr(\mathbf{x}|\mathbf{w})Pr(\mathbf{w})d\mathbf{w}}$$

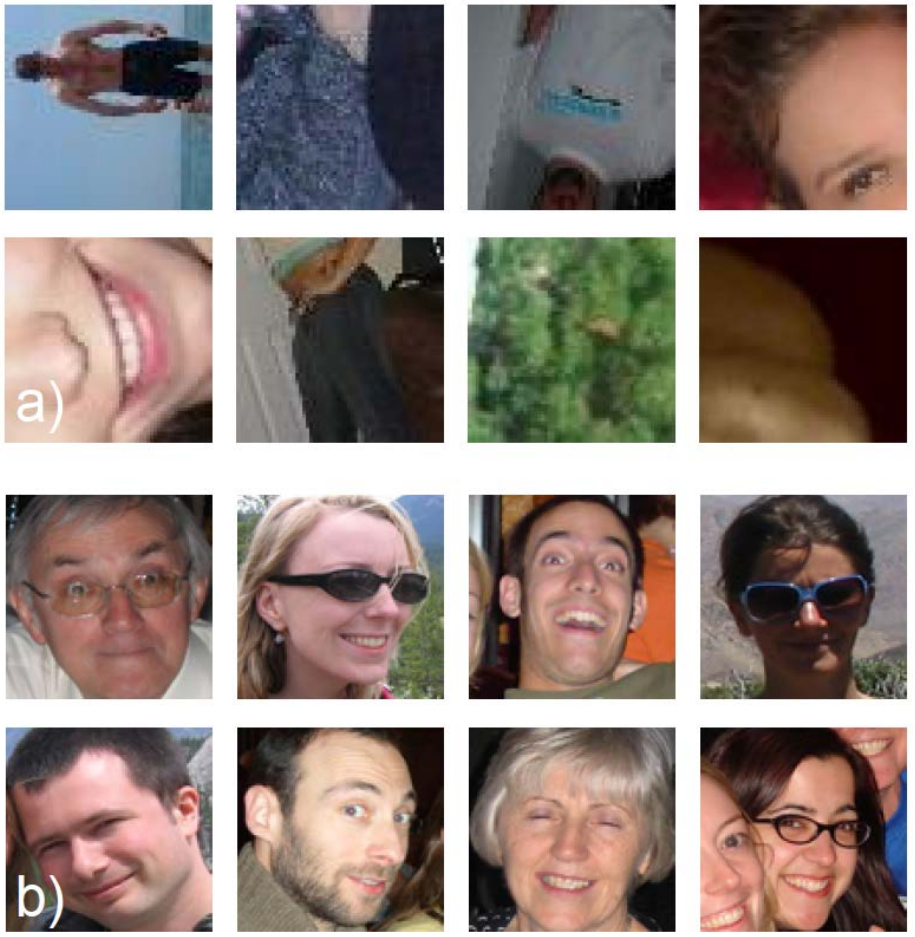
Experiment

1000 non-faces
1000 faces

60x60x3 Images =10800 x1 vectors

Equal priors Pr(y=1)=Pr(y=0) = 0.5

75% performance on test set. Not very good!



Results (diagonal covariance)

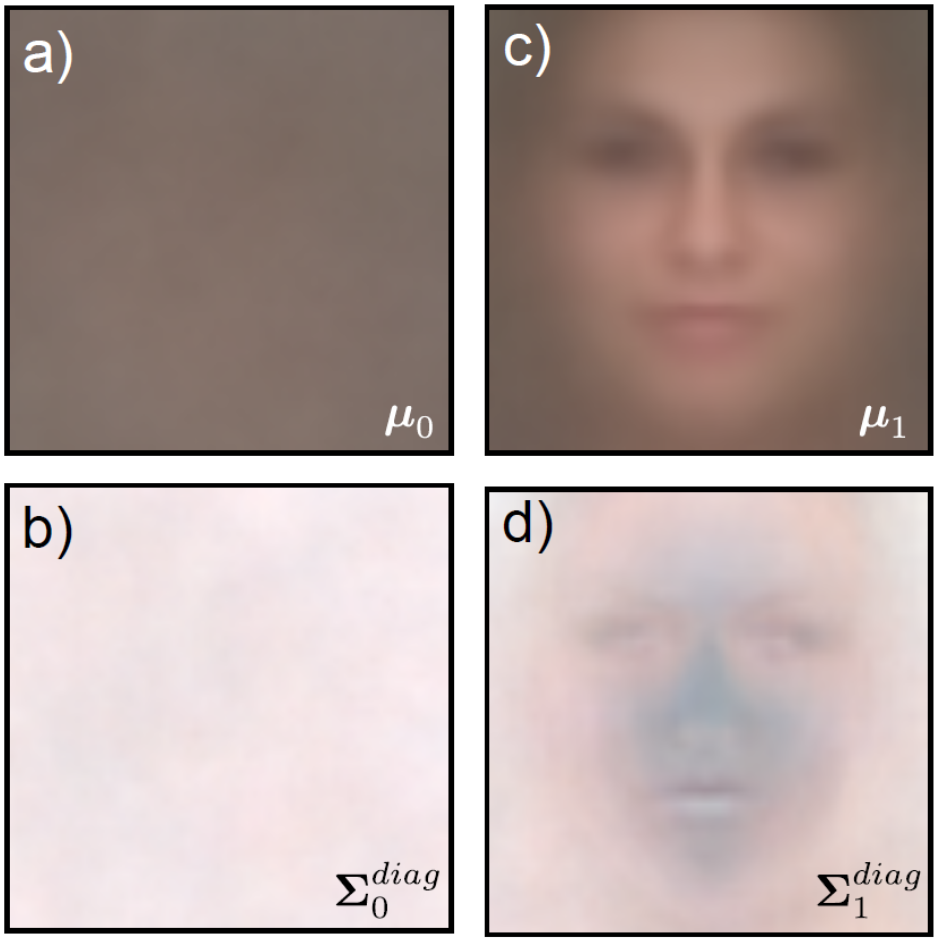
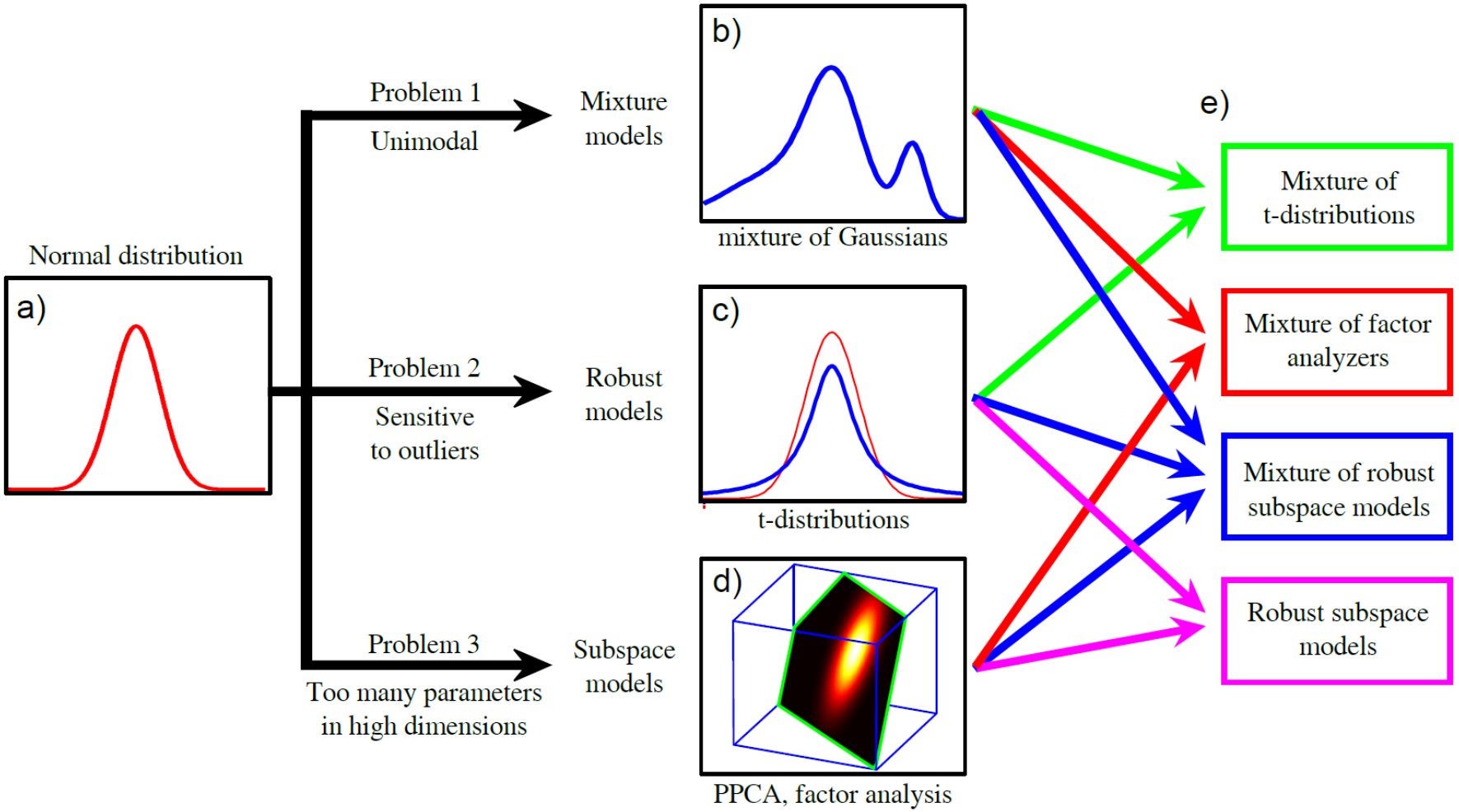


Figure 7.2 Class conditional density functions for normal model with diagonal covariance. Maximum likelihood fits based on 1000 training examples per class. a) Mean for background data $\boldsymbol{\mu}_0$ (reshaped from 10800×1 vector to 60×60 RGB image). b) Reshaped square root of diagonal covariance for background data $\boldsymbol{\Sigma}_0$. c) Mean for face data $\boldsymbol{\mu}_1$ d) Covariance for face data $\boldsymbol{\Sigma}_1$. The background model has little structure: the mean is uniform and the variance is high everywhere. The mean of the face model clearly captures class-specific information. The covariance of the face is larger at the edges of the image which usually contain hair or background.

The plan...


Normal distribution
a)
Problem 1
Unimodal
Problem 2
Sensitive
to outliers
Problem 3
Too many parameters
in high dimensions
Mixture
models
Robust
models
Subspace
models
b)
mixture of Gaussians
c)
t-distributions
d)
PPCA, factor analysis
e)
Mixture of
t-distributions
Mixture of factor
analyzers
Mixture of robust
subspace models
Robust subspace
models

Structure

- Densities for classification
- Models with hidden variables
  - Mixture of Gaussians
  - t-distributions
  - Factor analysis
- EM algorithm in detail
- Applications

Hidden (or latent) Variables

Key idea: represent density Pr(x) as marginalization of joint density with another variable h that we do not see

$$Pr(\mathbf{x}) = \int Pr(\mathbf{x}, \mathbf{h}) \, d\mathbf{h}$$

Will also depend on some parameters:

$$Pr(\mathbf{x}|\boldsymbol{\theta}) = \int Pr(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) \, d\mathbf{h}$$

Hidden (or latent) Variables

$$Pr(\mathbf{x}|\boldsymbol{\theta}) = \int Pr(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta})\, d\mathbf{h}$$

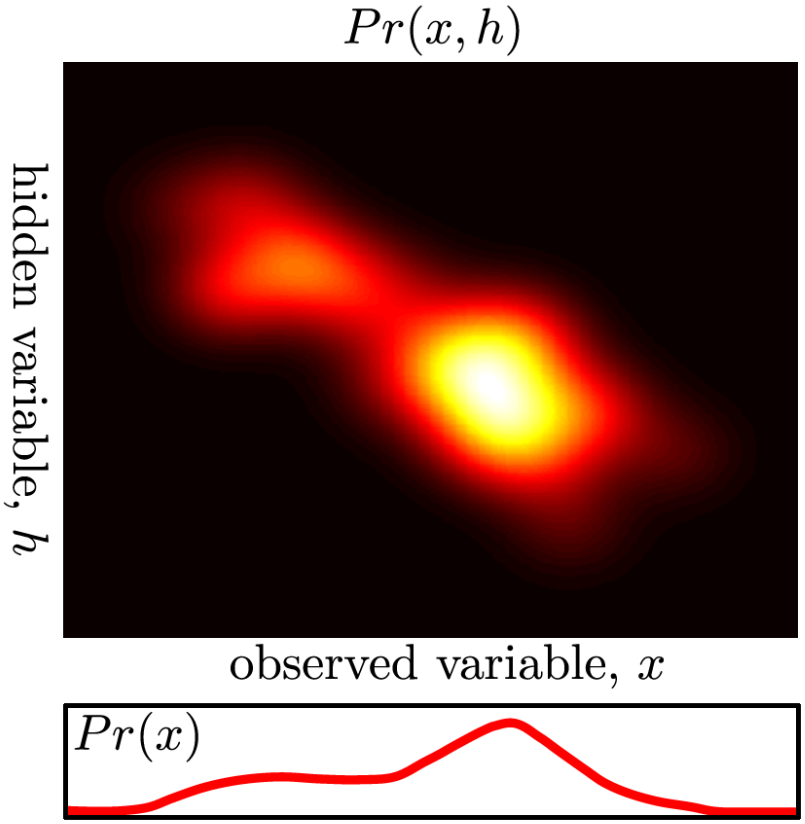


Figure 7.4 Using hidden variables to help model complex densities. One way to model the density $Pr(x)$ is to consider the joint probability distribution $Pr(x,h)$ between the observed data x and a hidden variable h. The density $Pr(x)$ can be considered as the marginalization of (integral over) this distribution with respect to the hidden variable h. As we manipulate the parameters $\boldsymbol{\theta}$ of this joint distribution, the marginal changes and the agreements with the observed data $\{x_i\}_{i=1}^{I}$ increases or decreases. Sometimes it is easier to fit the distribution in this indirect way than to directly manipulate $Pr(x)$.

Expectation Maximization

An algorithm specialized to fitting pdfs which are the marginalization of a joint distribution

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \, d\mathbf{h}_i \right] \right]$$

Defines a lower bound on log likelihood and increases bound iteratively

$$\begin{aligned} \mathcal{B}\left[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}\right] &= \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_{1\ldots I} \\ &\leq \sum_{i=1}^{I} \log \left[\int Pr(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}) d\mathbf{h}_{1\ldots I} \right]. \end{aligned}$$

Lower bound

$$\mathcal{B}\left[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}\right] = \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right] d\mathbf{h}_{1\ldots I}$$

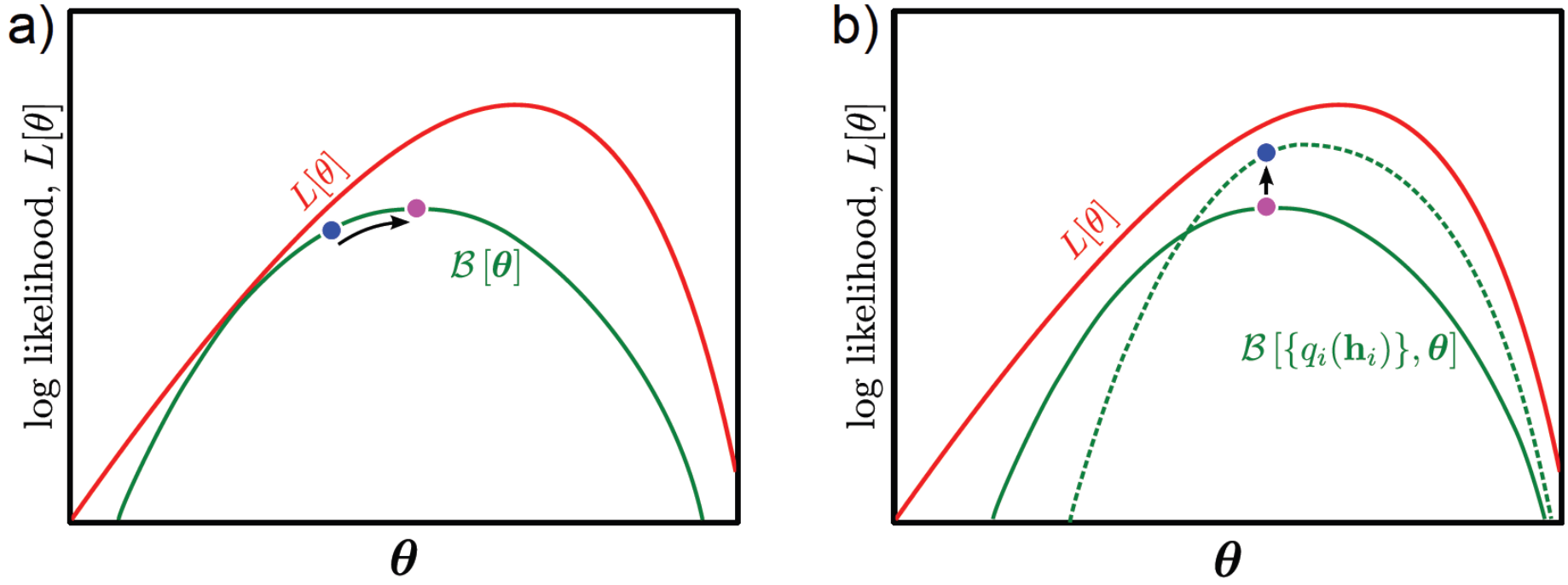
Lower bound is a *function* of parameters $\boldsymbol{\theta}$ and a set of probability distributions $q_i(\mathbf{h}_i)$

Expectation Maximization (EM) algorithm alternates E-steps and M-Steps

E-Step – Maximize bound w.r.t. distributions $q(\mathbf{h}_i)$
M-Step – Maximize bound w.r.t. parameters $\boldsymbol{\theta}$

Lower bound

$$\mathcal{B}\left[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}\right] = \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right] d\mathbf{h}_{1\ldots I}$$



E-Step & M-Step

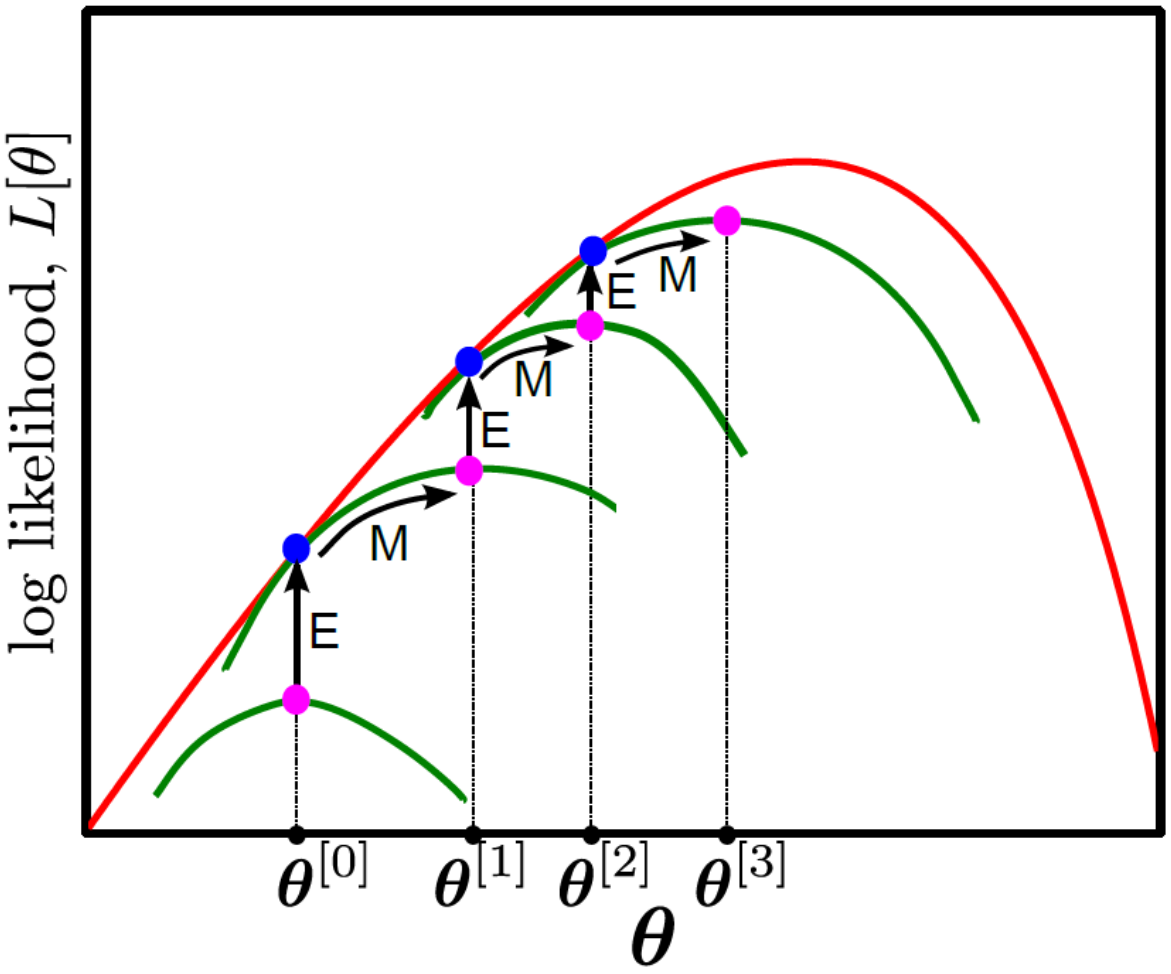
E-Step – Maximize bound w.r.t. distributions $q_i(\mathbf{h}_i)$

$$\hat{q}_i(\mathbf{h}_i) = Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta}^{[t]}) = \frac{Pr(\mathbf{x}_i|\mathbf{h}_i, \boldsymbol{\theta}^{[t]})Pr(\mathbf{h}_i|\boldsymbol{\theta}^{[t]})}{Pr(\mathbf{x}_i)}$$

M-Step – Maximize bound w.r.t. parameters θ

$$\hat{\boldsymbol{\theta}}^{[t+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int \hat{q}_i(\mathbf{h}_i) \log \left[Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \right] \, d\mathbf{h}_i \right]$$

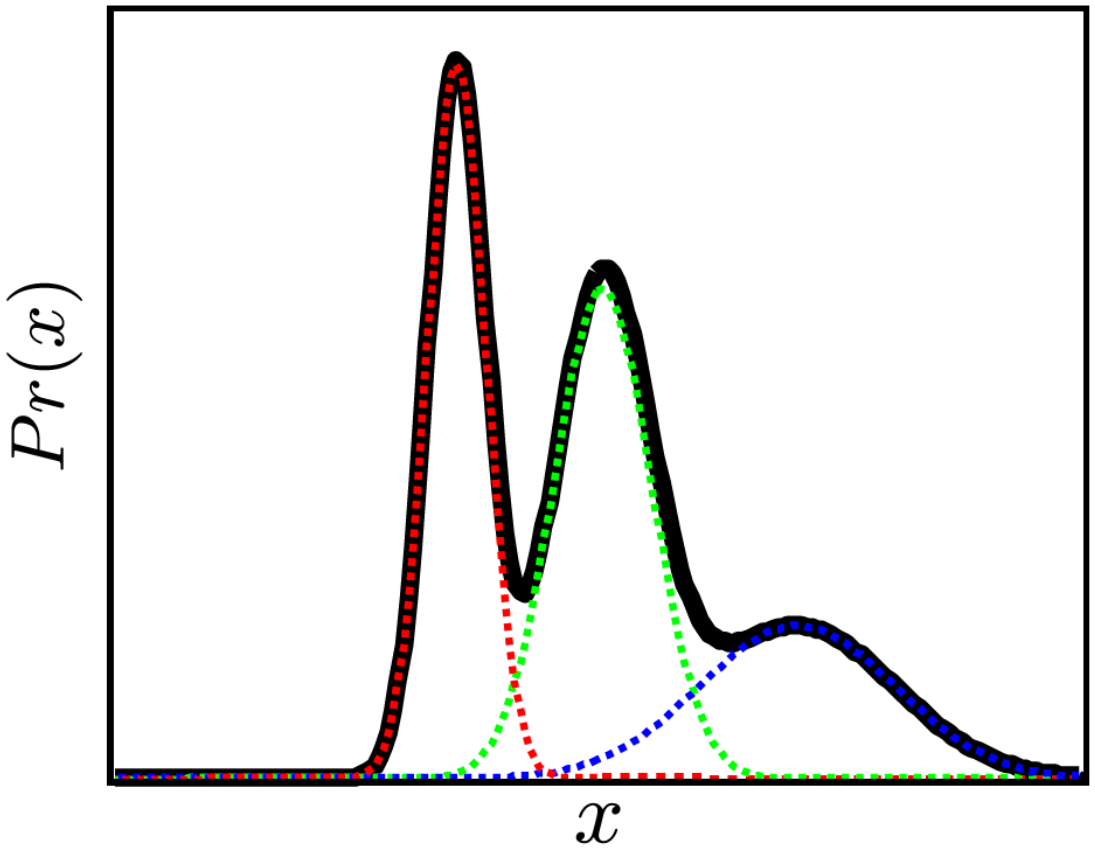
E-Step & M-Step


log likelihood, $L[\theta]$
E
M
E
M
E
M
$\theta^{[0]}$
$\theta^{[1]}$
$\theta^{[2]}$
$\theta^{[3]}$
θ

Structure

- Densities for classification
- Models with hidden variables
  - Mixture of Gaussians
  - t-distributions
  - Factor analysis
- EM algorithm in detail
- Applications

Mixture of Gaussians (MoG)



$$Pr(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \lambda_k \mathrm{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$$

ML Learning

Q. Can we learn this model with maximum likelihood?

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log \left[Pr(\mathbf{x}_i | \boldsymbol{\theta}) \right] \right]$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log \left[\sum_{k=1}^{K} \lambda_k \mathrm{Norm}_{\mathbf{x}_i} [\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k] \right] \right]$$

A. Yes, but using brute force approach is tricky
   - If you take derivative and set to zero, can't solve
     -- the log of the sum causes problems
   - Have to enforce constraints on parameters
     - covariances must be positive definite
     - weights must sum to one

MoG as a marginalization

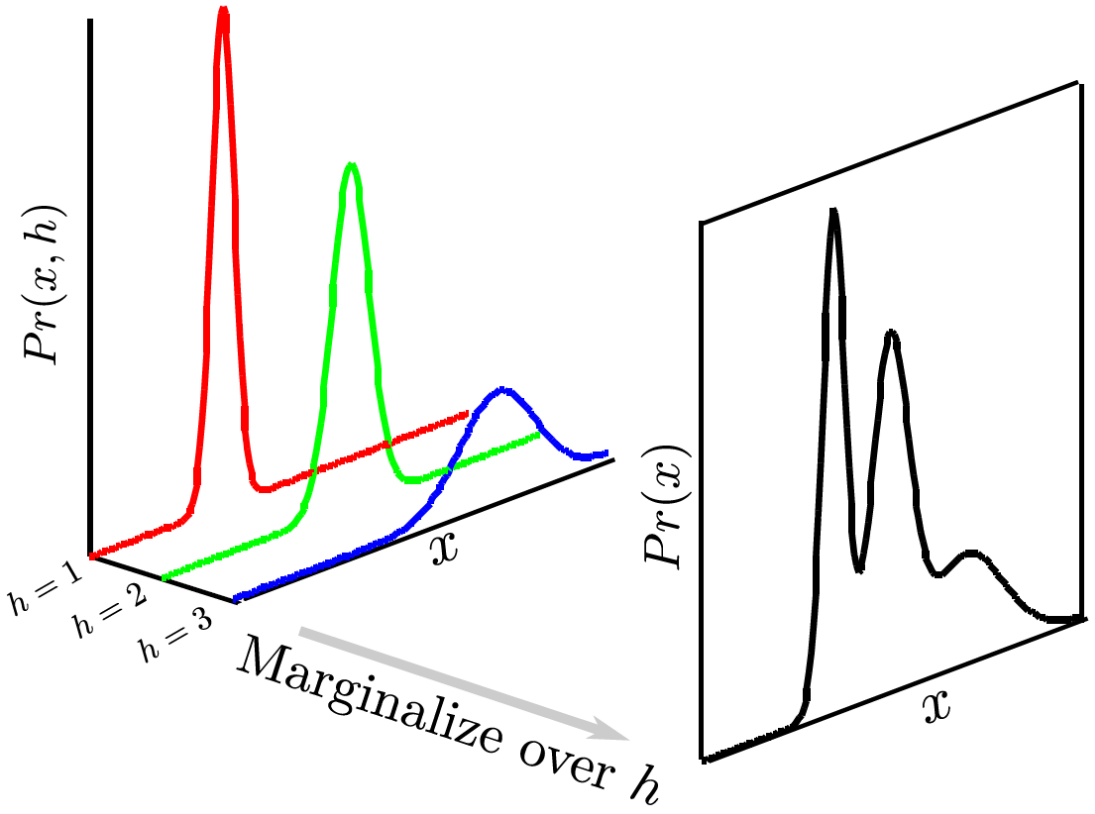
Define a variable $h \in \{1 \ldots K\}$ and then write

$$\begin{aligned} Pr(\mathbf{x}|h, \boldsymbol{\theta}) &= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h] \\ Pr(h|\boldsymbol{\theta}) &= \text{Cat}_h[\boldsymbol{\lambda}] \end{aligned}$$

Then we can recover the density by marginalizing Pr(**x**,h)

$$\begin{aligned} Pr(\mathbf{x}|\boldsymbol{\theta}) &= \sum_{k=1}^{K} Pr(\mathbf{x}, h=k|\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} Pr(\mathbf{x}|h=k, \boldsymbol{\theta}) Pr(h=k|\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \lambda_k \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]. \end{aligned}$$

MoG as a marginalization



$$Pr(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} Pr(\mathbf{x}, h=k|\boldsymbol{\theta})$$

MoG as a marginalization

Define a variable $h \in \{1 \ldots K\}$ and then write

$$\begin{aligned} Pr(\mathbf{x}|h, \boldsymbol{\theta}) &= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h] \\ Pr(h|\boldsymbol{\theta}) &= \text{Cat}_h[\boldsymbol{\lambda}] \end{aligned}$$

Note :

- This gives us a method to generate data from MoG
  - First sample Pr(h), then sample Pr(**x**|h)
- The hidden variable h has a clear interpretation – it tells you which Gaussian created data point **x**

Expectation Maximization for MoG

GOAL: to learn parameters $\boldsymbol{\theta} = \{\lambda_{1\ldots K}, \boldsymbol{\mu}_{1\ldots K}, \boldsymbol{\Sigma}_{1\ldots K}\}$ from training data $\mathbf{x}_{1\ldots I}$

E-Step – Maximize bound w.r.t. distributions q($\mathbf{h}_i$)

$$\hat{q}_i(\mathbf{h}_i) = Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta}^{[t]}) = \frac{Pr(\mathbf{x}_i|\mathbf{h}_i, \boldsymbol{\theta}^{[t]})Pr(\mathbf{h}_i|\boldsymbol{\theta}^{[t]})}{Pr(\mathbf{x}_i)}$$

M-Step – Maximize bound w.r.t. parameters θ

$$\hat{\boldsymbol{\theta}}^{[t+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \sum_{k=1}^{K} \hat{q}_i(\mathbf{h}_i = k) \log \left[Pr(\mathbf{x}_i, \mathbf{h}_i = k|\boldsymbol{\theta}) \right] \right]$$

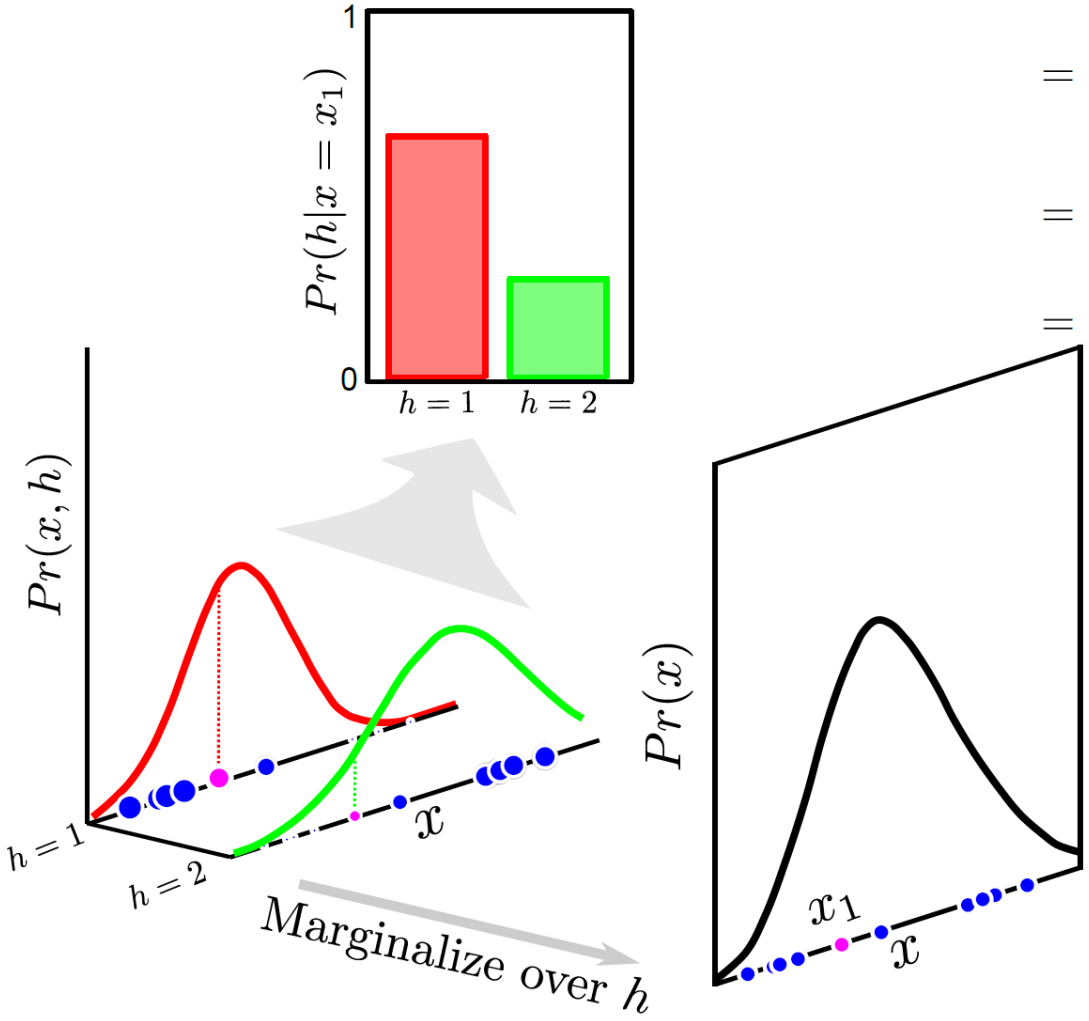
E-Step

$$Pr(h_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{[t]})$$

$$= \frac{Pr(\mathbf{x}_i|h_i = k, \boldsymbol{\theta}^{[t]})Pr(h_i = k, \boldsymbol{\theta}^{[t]})}{\sum_{j=1}^{K} Pr(\mathbf{x}_i|h_i = j, \boldsymbol{\theta}^{[t]})Pr(h_i = j, \boldsymbol{\theta}^{[t]})}$$

$$= \frac{\lambda_k \mathrm{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]}{\sum_{j=1}^{K} \lambda_j \mathrm{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j]}$$

$$= r_{ik}$$



We'll call this the responsibility of the k^{th} Gaussian for the i^{th} data point

Repeat this procedure for every datapoint!

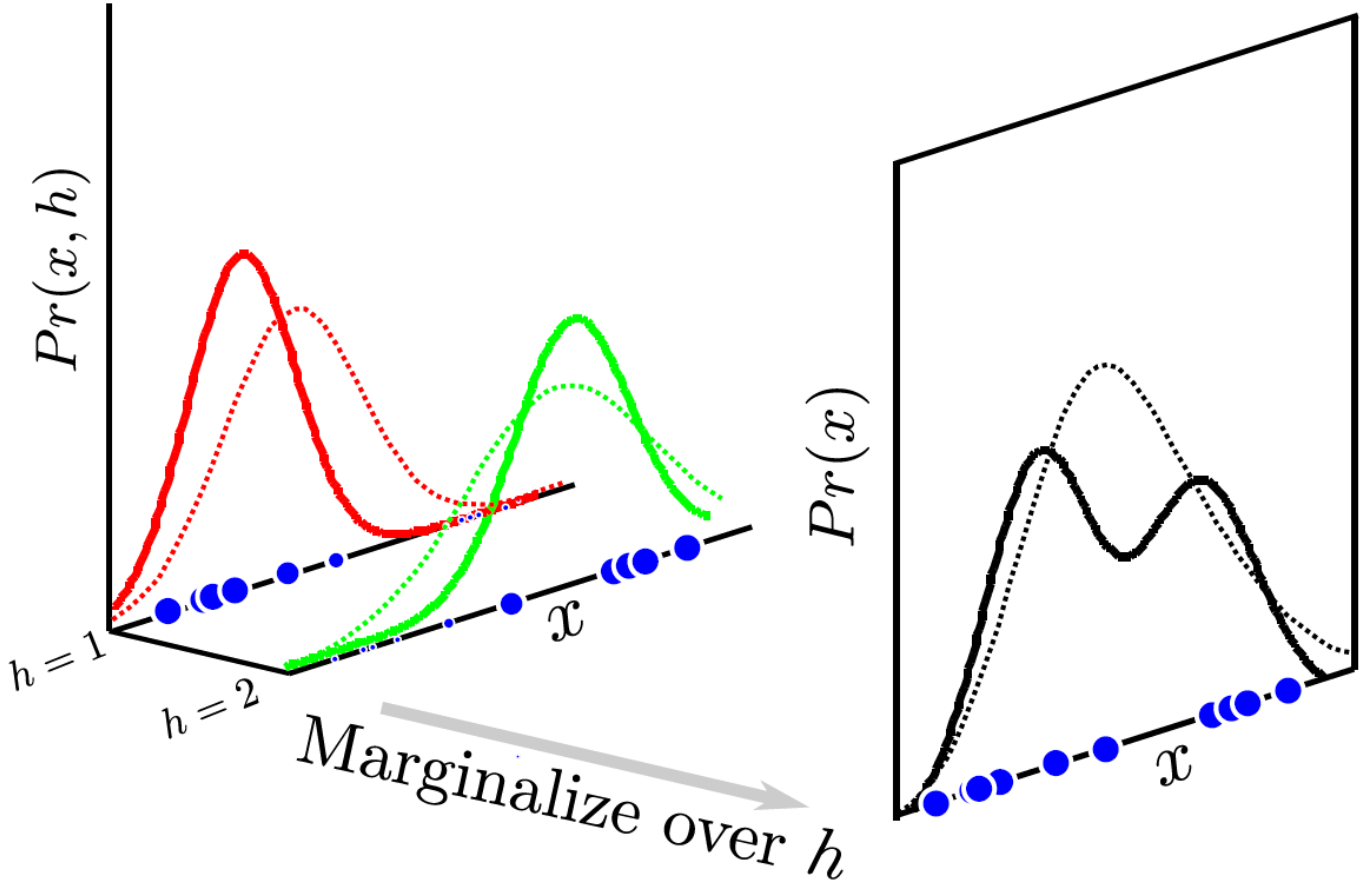
M-Step

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{[t+1]} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\sum_{i=1}^{I}\sum_{k=1}^{K}\hat{q}_i(h_i=k)\log\left[Pr(\mathbf{x}_i,h_i=k|\boldsymbol{\theta})\right]\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\sum_{i=1}^{I}\sum_{k=1}^{K}r_{ik}\log\left[\lambda_k\mathrm{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k]\right]\right].\end{aligned}$$

Take derivative, equate to zero and solve (Lagrange multipliers for λ)

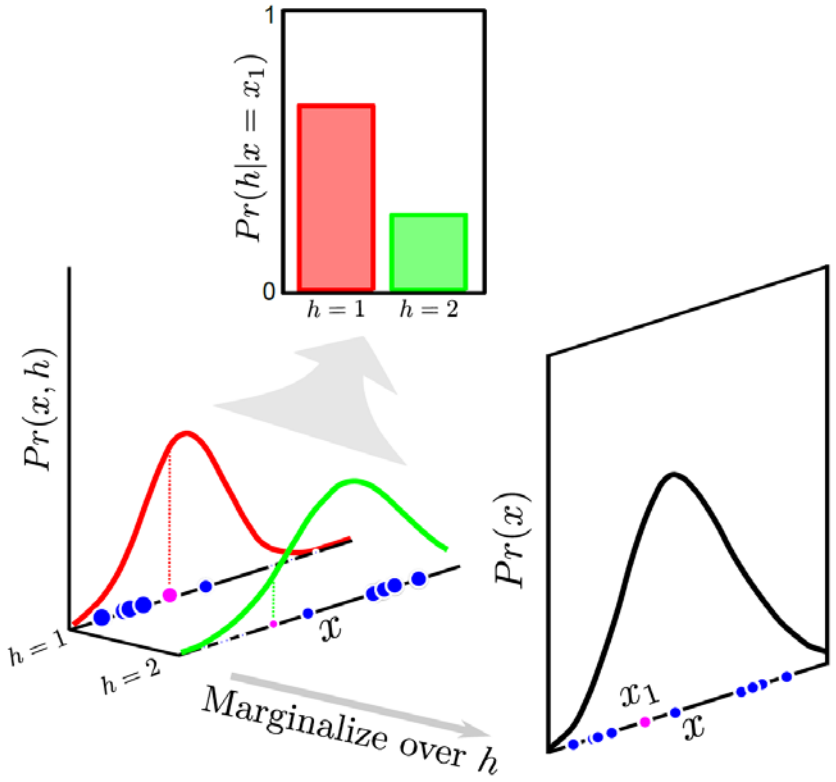
$$\begin{aligned}\lambda_k^{[t+1]} &= \frac{\sum_{i=1}^{I} r_{ik}}{\sum_{j=1}^{K}\sum_{i=1}^{I} r_{ij}} \\ \boldsymbol{\mu}_k^{[t+1]} &= \frac{\sum_{i=1}^{I} r_{ik}\mathbf{x}_i}{\sum_{i=1}^{I} r_{ik}} \\ \boldsymbol{\Sigma}_k^{[t+1]} &= \frac{\sum_{i=1}^{I} r_{ik}(\mathbf{x}_i-\boldsymbol{\mu}_k^{[t+1]})(\mathbf{x}_i-\boldsymbol{\mu}_k^{[t+1]})^T}{\sum_{i=1}^{I} r_{ik}}\end{aligned}$$

M-Step

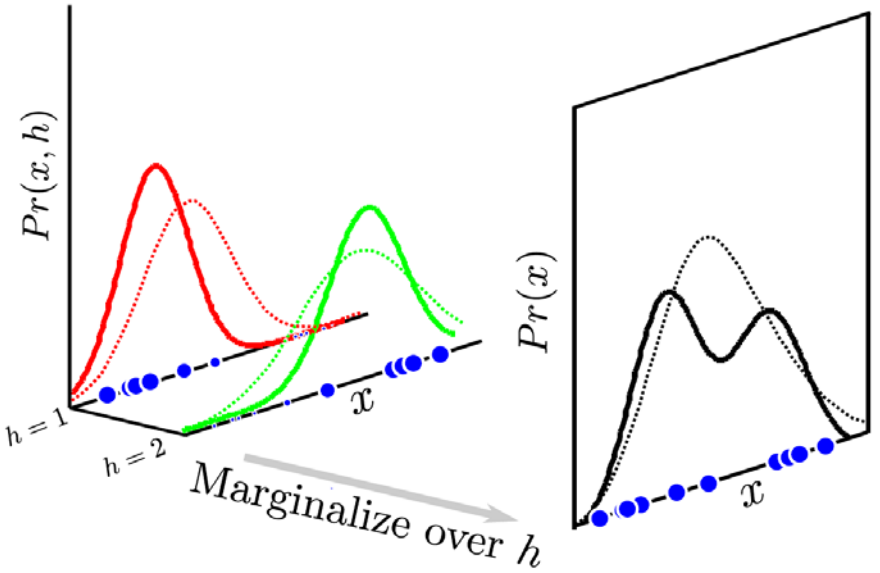


Update means , covariances and weights according to responsibilities of datapoints

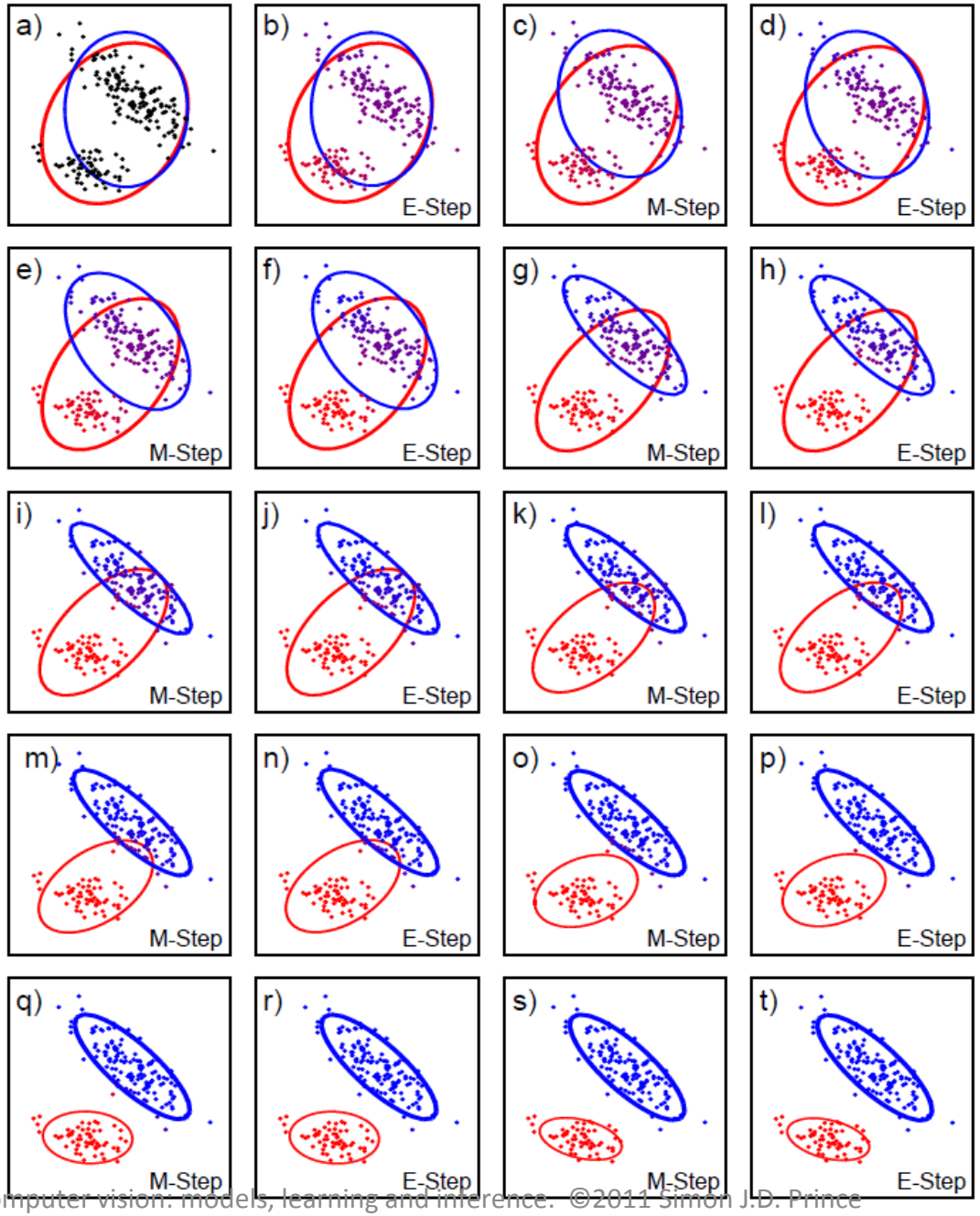
Iterate until no further improvement



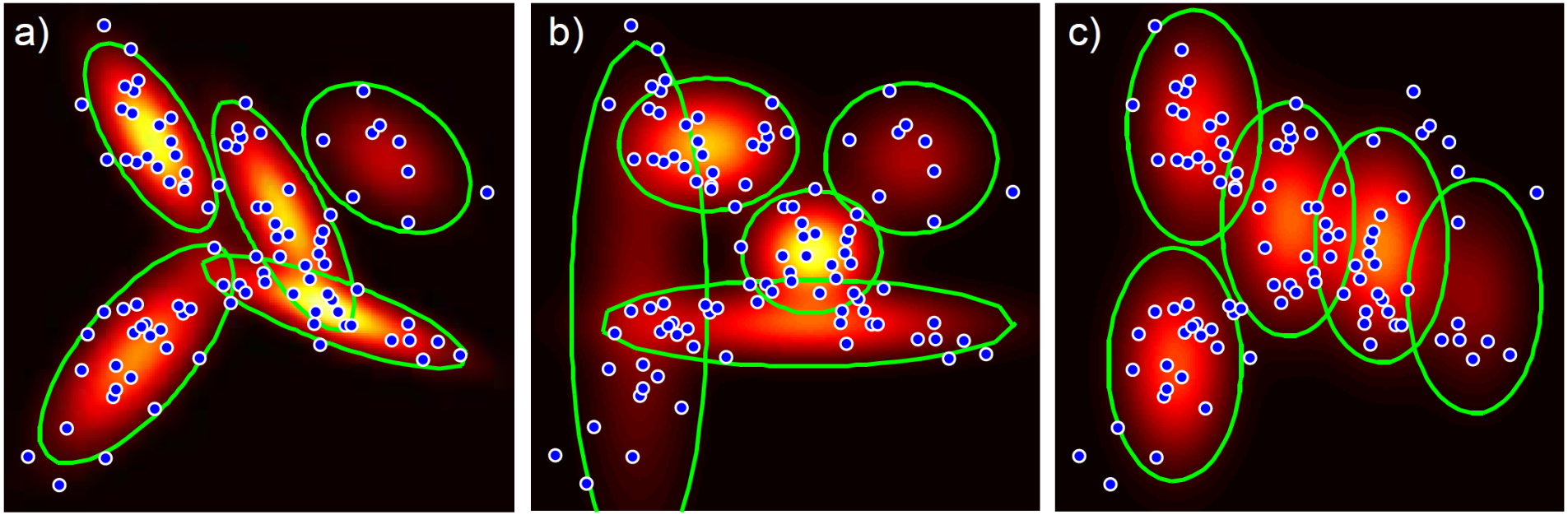
E-Step



M-Step


a)
b)
E-Step
c)
M-Step
d)
E-Step
e)
M-Step
f)
E-Step
g)
M-Step
h)
E-Step
i)
M-Step
j)
E-Step
k)
M-Step
l)
E-Step
m)
M-Step
n)
E-Step
o)
M-Step
p)
E-Step
q)
M-Step
r)
E-Step
s)
M-Step
t)
E-Step

Different flavours...



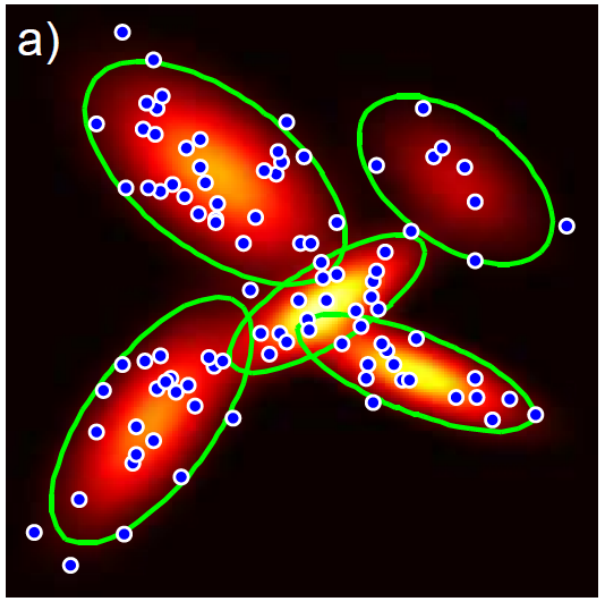
Full covariance

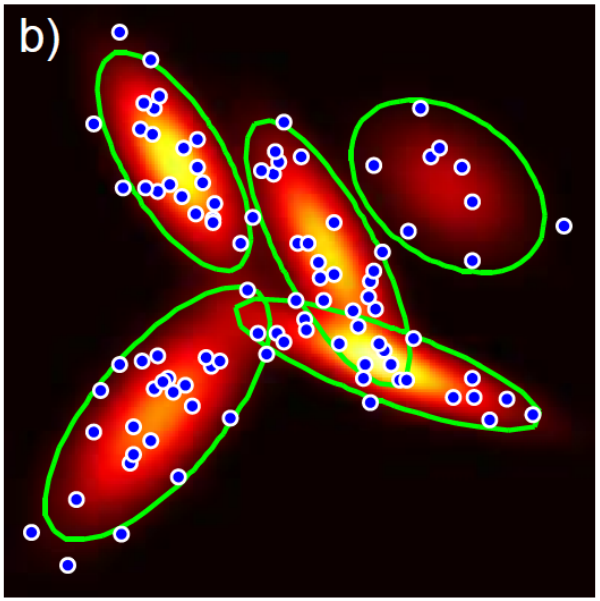
Diagonal covariance

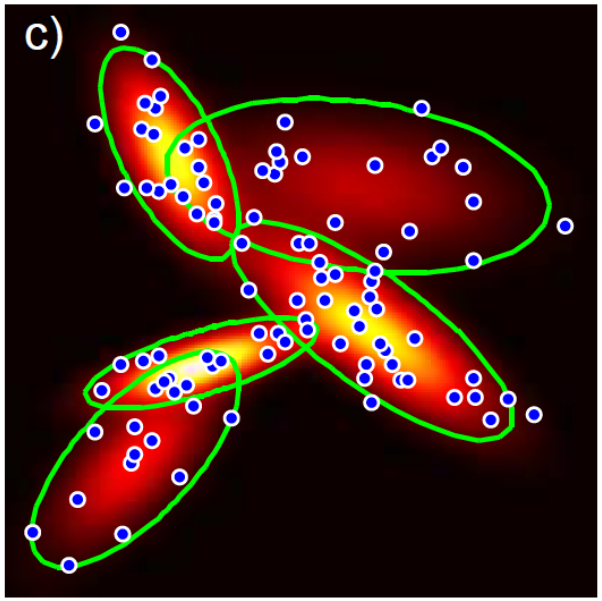
Same covariance

Local Minima

Start from three random positions







L = 98.76

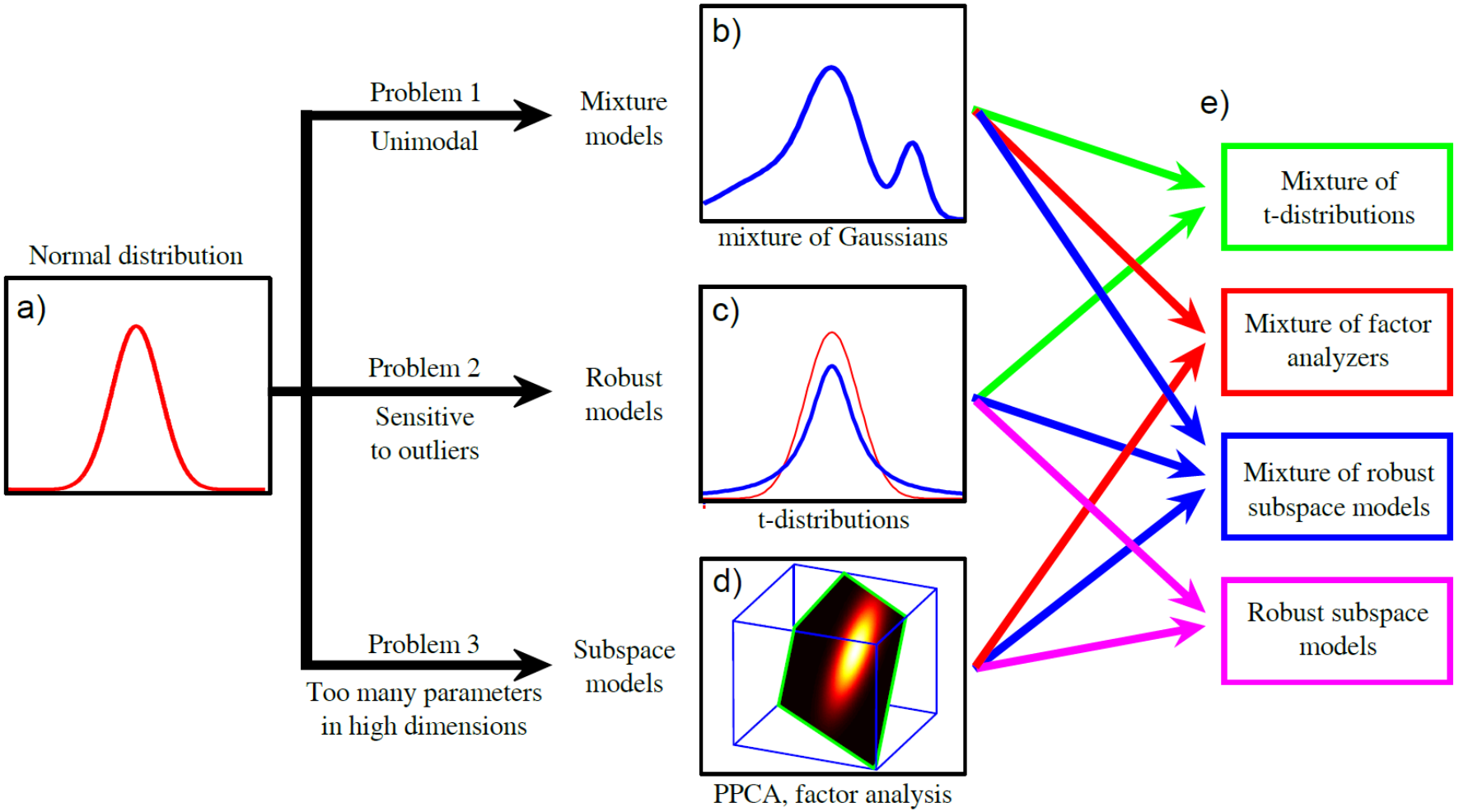
L = 96.97

L = 94.35

Means of face/non-face model

Classification → 84% (9% improvement!)

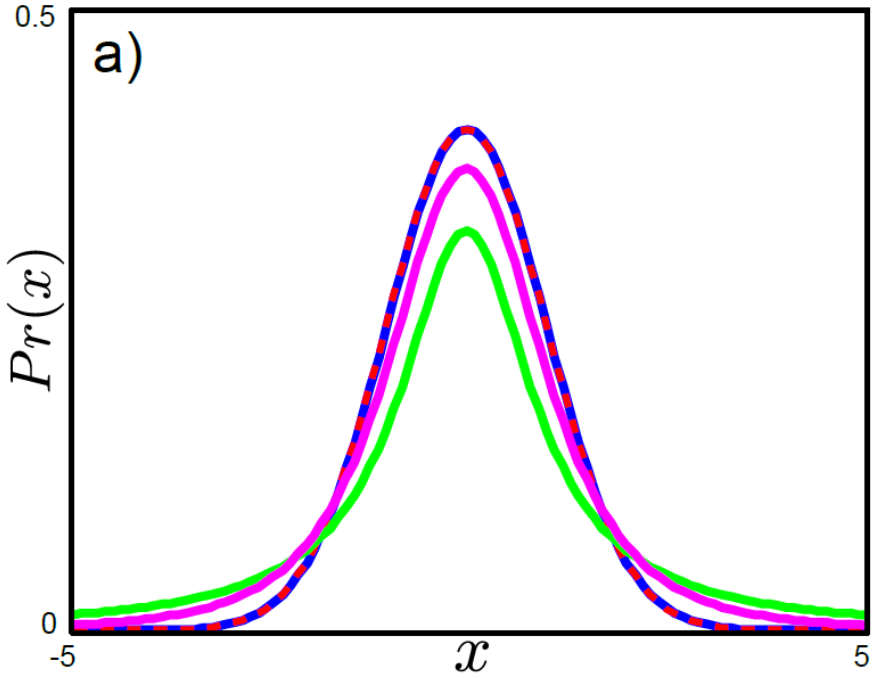
The plan...

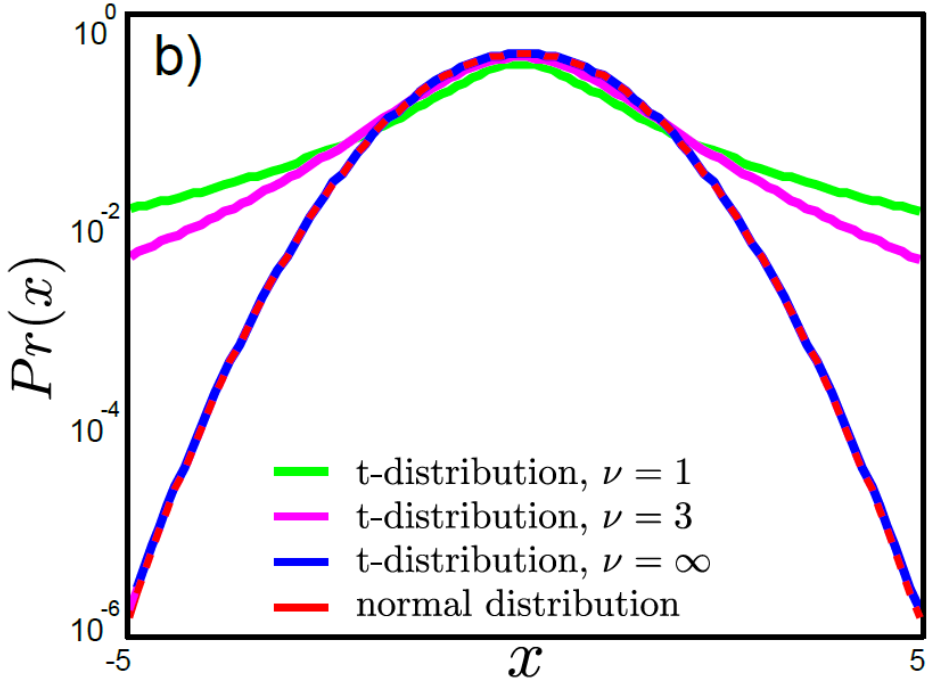

Normal distribution
a)
Problem 1
Unimodal
Mixture models
b)
mixture of Gaussians
Problem 2
Sensitive to outliers
Robust models
c)
t-distributions
Problem 3
Too many parameters in high dimensions
Subspace models
d)
PPCA, factor analysis
e)
Mixture of t-distributions
Mixture of factor analyzers
Mixture of robust subspace models
Robust subspace models

Structure

- Densities for classification
- Models with hidden variables
  - Mixture of Gaussians
  - t-distributions
  - Factor analysis
- EM algorithm in detail
- Applications

Student t-distributions



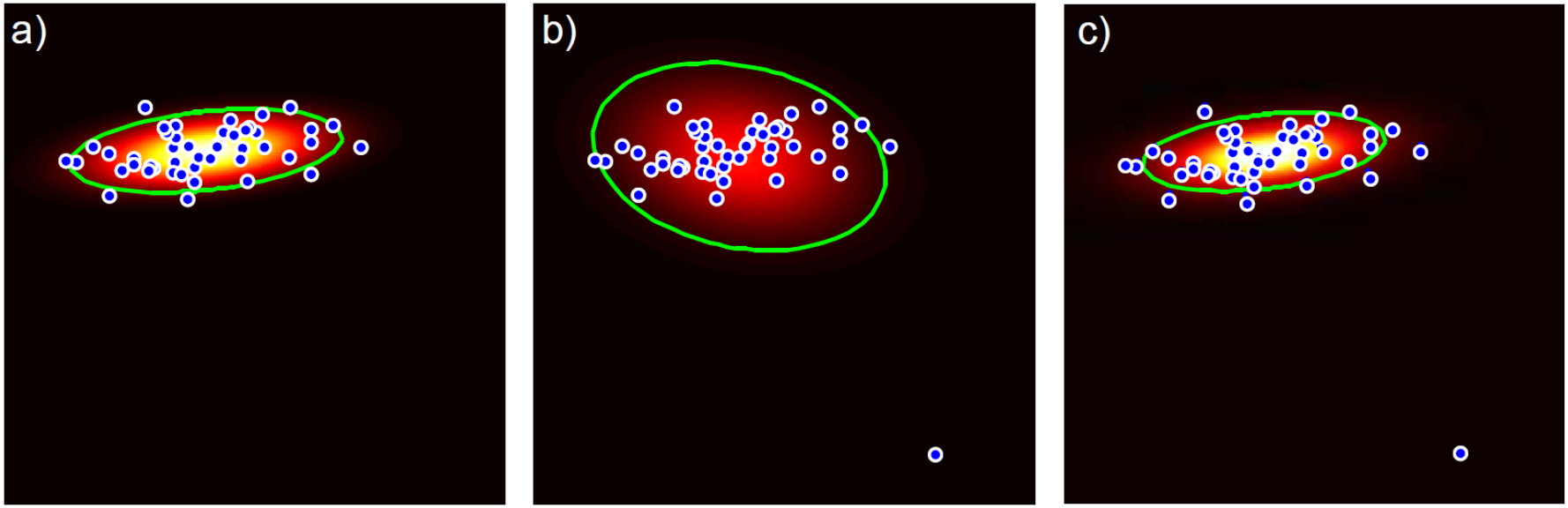


$$Pr(x) = \text{Stud}_{\mathbf{x}}\left[\mu, \sigma^2, \nu\right]$$

$$= \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\nu\pi\sigma^2}\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

Student t-distributions motivation

The normal distribution is not very robust – a single outlier can completely throw it off because the tails fall off so fast...



Normal distribution

Normal distribution w/ one extra datapoint!

t-distribution

Student t-distributions

Univariate student t-distribution

$$\begin{aligned} Pr(x) &= \text{Stud}_{\mathbf{x}}\left[\mu, \sigma^2, \nu\right] \\ &= \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\nu\pi\sigma^2}\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \end{aligned}$$

Multivariate student t-distribution

$$\begin{aligned} Pr(\mathbf{x}) &= \text{Stud}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right] \\ &= \frac{\Gamma\left[\frac{\nu+D}{2}\right]}{(\nu\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+D}{2}} \end{aligned}$$

t-distribution as a marginalization

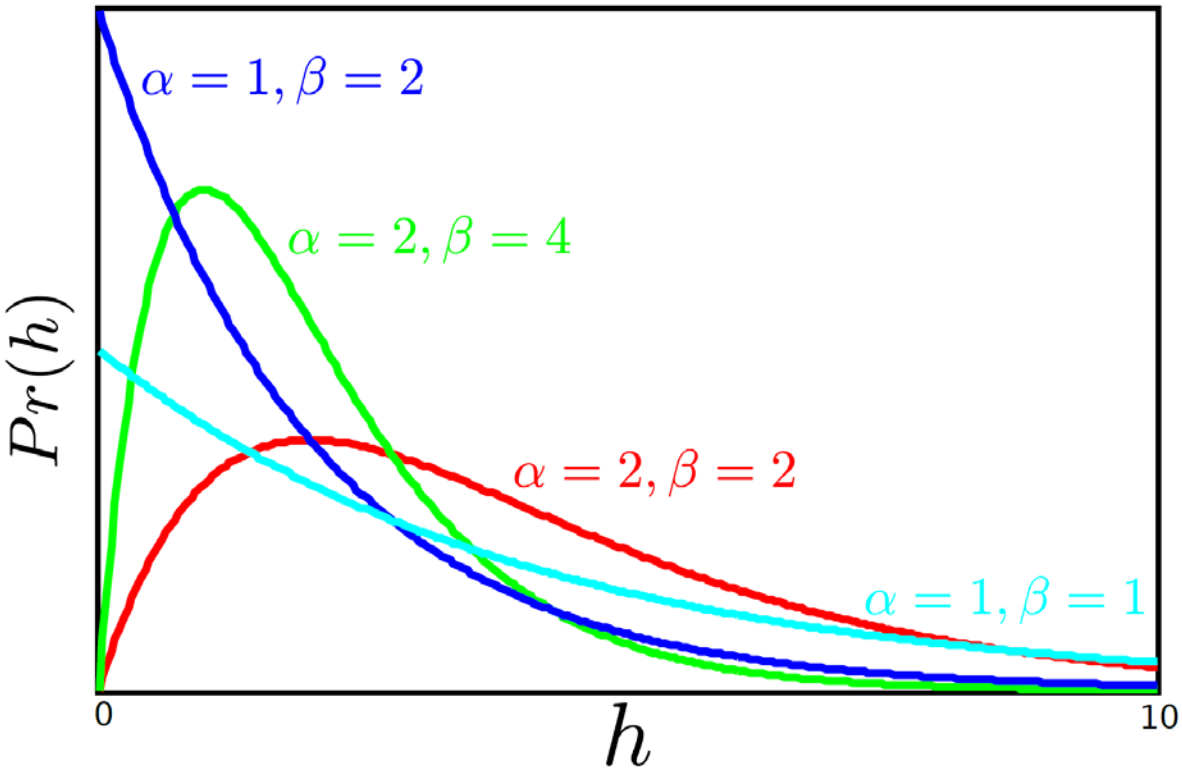
Define hidden variable h

$$\begin{aligned} Pr(\mathbf{x}|h) &= \text{Norm}_x[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h] \\ Pr(h) &= \text{Gam}_h[\nu/2, \nu/2] \end{aligned}$$

Can be expressed as a marginalization

$$\begin{aligned} Pr(\mathbf{x}) = \int Pr(\mathbf{x}, h)dh &= \int Pr(\mathbf{x}|h)Pr(h)dh \\ &= \int \text{Norm}_x[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h]\text{Gam}_h[\nu/2, \nu/2]dh \\ &= \text{Stud}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu]. \end{aligned}$$

Gamma distribution



$$\mathrm{Gam}_h[\alpha, \beta] = \frac{\beta^\alpha}{\Gamma[\alpha]} \exp[-\beta h] h^{\alpha-1}$$

t-distribution as a marginalization

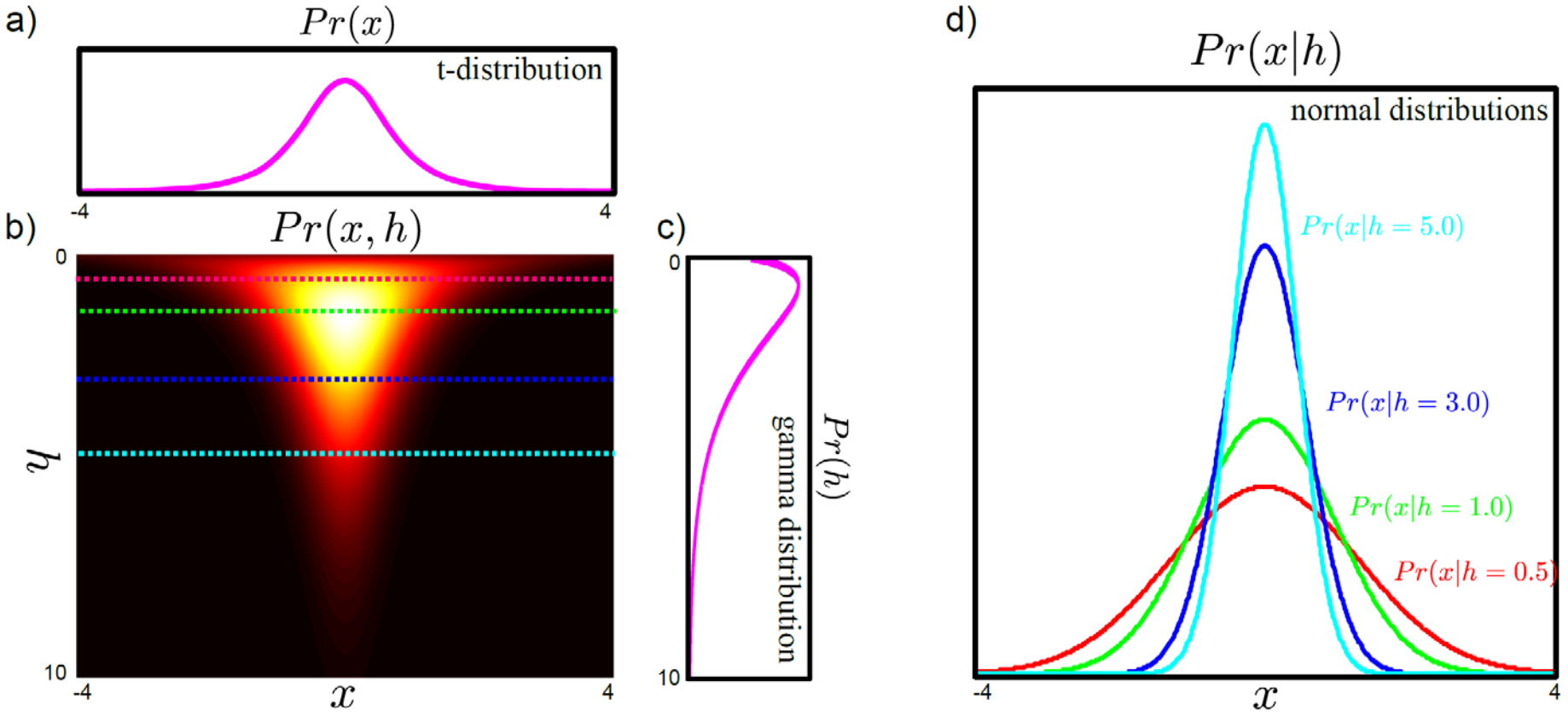
Define hidden variable h

$$\begin{aligned} Pr(\mathbf{x}|h) &= \text{Norm}_x[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h] \\ Pr(h) &= \text{Gam}_h[\nu/2, \nu/2] \end{aligned}$$

Things to note:

- Again this provides a method to sample from the t-distribution
- Variable h has a clear interpretation:
  - Each datum drawn from a Gaussian, mean $\boldsymbol{\mu}$
  - Covariance depends inversely on h
- Can think of this as an infinite mixture (sum becomes integral) of Gaussians w/ same mean, but different variances

t-distribution



EM for t-distributions

GOAL: to learn parameters $\theta = \{\mu, \sigma^2, \nu\}$ from training data $\mathbf{x}_{1\ldots I}$

E-Step – Maximize bound w.r.t. distributions $q(h_i)$

$$q_i(h_i) = Pr(h_i|\mathbf{x}_i, \boldsymbol{\theta}^{[t]}) = \frac{Pr(\mathbf{x}_i|h_i, \boldsymbol{\theta}^{[t]})Pr(h_i)}{Pr(\mathbf{x}_i|\boldsymbol{\theta}^{[t]})}$$

M-Step – Maximize bound w.r.t. parameters θ

$$\hat{\boldsymbol{\theta}}^{[t+1]} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{I} \int \hat{q}_i(h_i) \log\left[Pr(\mathbf{x}, h_i, \boldsymbol{\theta})\right] dh_i$$

E-Step

$$q_i(h_i) = Pr(h_i|\mathbf{x}_i, \boldsymbol{\theta}^{[t]}) = \frac{Pr(\mathbf{x}_i|h_i, \boldsymbol{\theta}^{[t]})Pr(h_i)}{Pr(\mathbf{x}_i|\boldsymbol{\theta}^{[t]})}$$

$$= \frac{\mathrm{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h_i]\mathrm{Gam}_{h_i}[\nu/2, \nu/2]}{Pr(\mathbf{x}_i)}$$

$$= \mathrm{Gam}_{h_i}\left[\frac{\nu + D}{2}, \frac{(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})}{2} + \frac{\nu}{2}\right]$$

Extract expectations

$$\mathrm{E}[h_i] = \frac{(\nu + D)}{\nu + (\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})}$$

$$\mathrm{E}[\log[h_i]] = \Psi\left[\frac{\nu + D}{2}\right] - \log\left[\frac{\nu + (\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})}{2}\right]$$

M-Step

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{[t+1]} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int \hat{q}_i(h_i) \log\left[Pr(\mathbf{x}_i, h_i|\boldsymbol{\theta})\right] dh_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int \hat{q}_i(h_i) \left(\log\left[Pr(\mathbf{x}_i|h_i, \boldsymbol{\theta})\right] + \log\left[Pr(h_i)\right]\right) dh_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int Pr(h_i|\mathbf{x}_i, \boldsymbol{\theta}^{[t]}) \left(\log\left[Pr(\mathbf{x}_i|h_i, \boldsymbol{\theta})\right] + \log\left[Pr(h_i)\right]\right) dh_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} E\left[\log\left[Pr(\mathbf{x}_i|h_i, \boldsymbol{\theta})\right]\right] + E\left[\log\left[Pr(h_i)\right]\right] \right]
\end{aligned}
$$

Where...

$$
E\left[\log\left[Pr(\mathbf{x}_i|h_i, \boldsymbol{\theta})\right]\right] = \frac{D\mathrm{E}[\log h_i] - D\log 2\pi - \log|\boldsymbol{\Sigma}| - (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\mathrm{E}[h_i]}{2}
$$

$$
E\left[\log\left[Pr(h_i)\right]\right] = \frac{\nu}{2}\log\left[\frac{\nu}{2}\right] - \log\Gamma\left[\frac{\nu}{2}\right] + \left(\frac{\nu}{2} - 1\right)\mathrm{E}[\log h_i] - \frac{\nu}{2}\mathrm{E}[h_i].
$$

Updates

$$\boldsymbol{\mu}^{[t+1]} = \frac{\sum_{i=1}^{I} \mathrm{E}[h_i]\mathbf{x}_i}{\sum_{i=1}^{I} \mathrm{E}[h_i]}$$

$$\boldsymbol{\Sigma}^{[t+1]} = \frac{\sum_{i=1}^{I} \mathrm{E}[h_i](\mathbf{x}_i - \boldsymbol{\mu}^{[t+1]})(\mathbf{x}_i - \boldsymbol{\mu}^{[t+1]})^T}{\sum_{i=1}^{I} \mathrm{E}[h_i]}$$

No closed form solution for ν. Must optimize bound – since it is only one number we can optimize by just evaluating with a fine grid of values and picking the best

EM algorithm for t-distributions

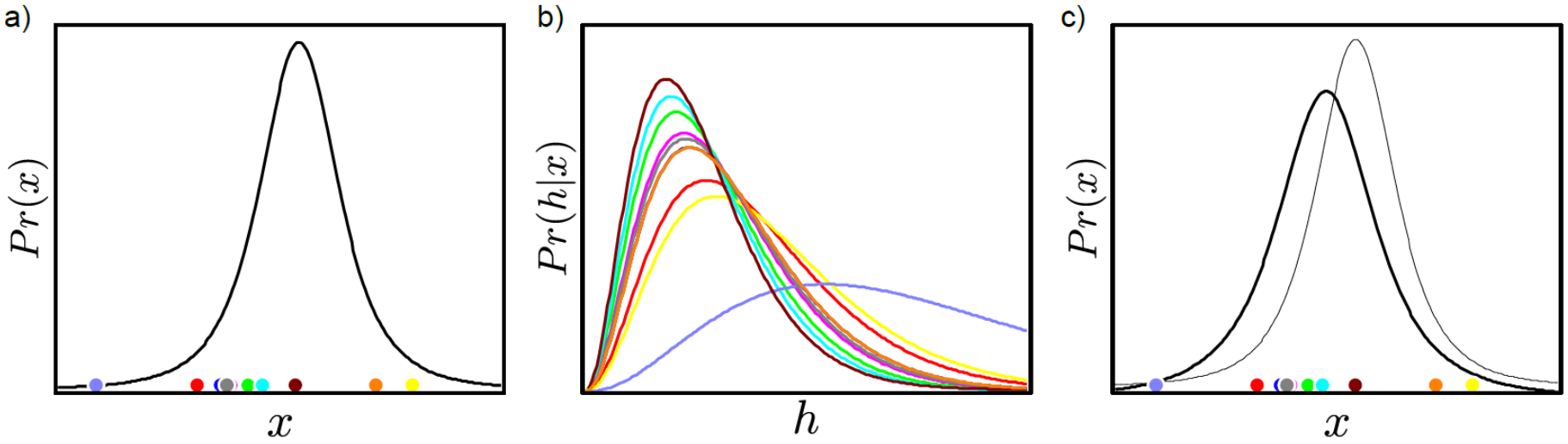
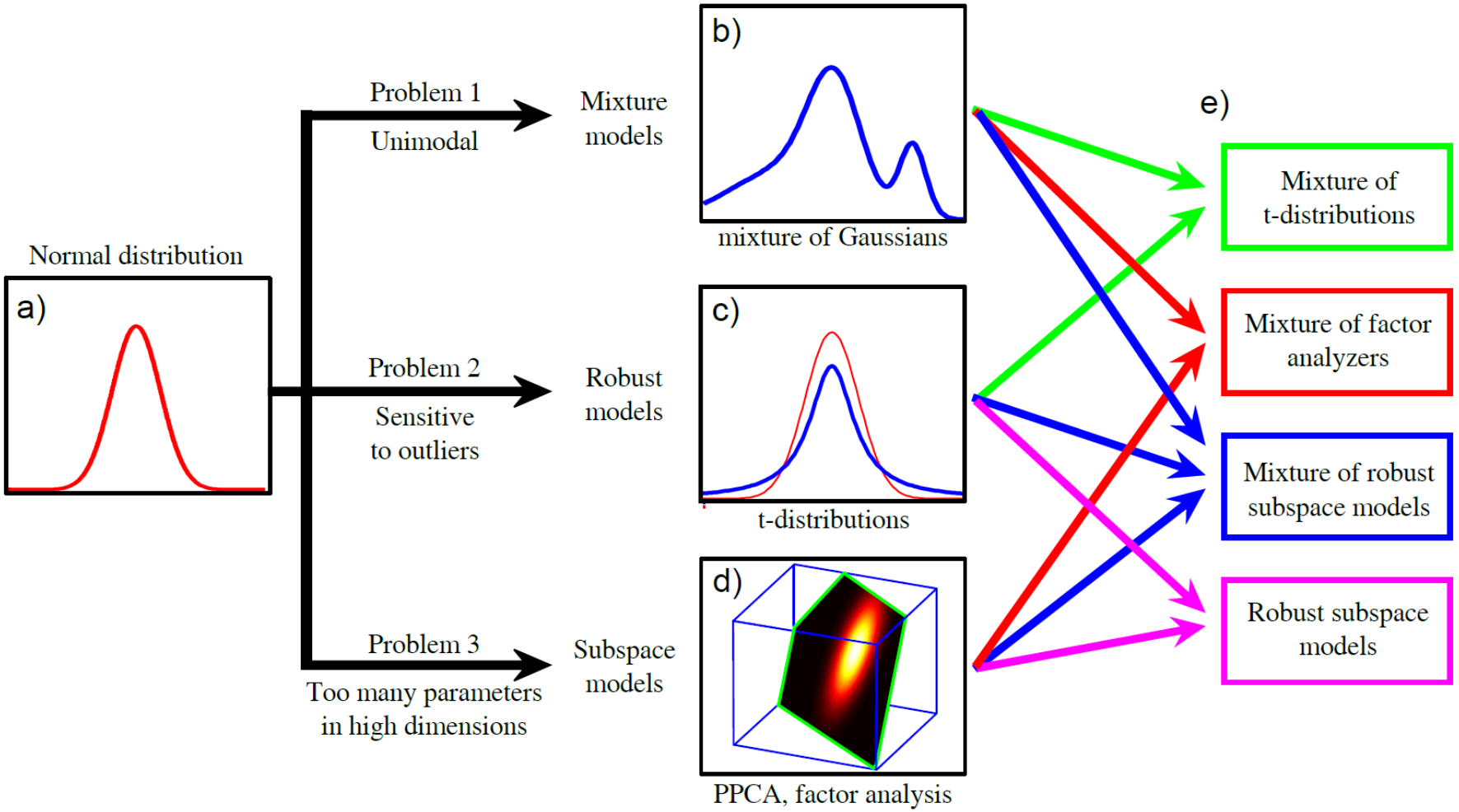


Figure 7.18 Expectation maximization for fitting t-distributions. a) Estimate of distribution before update. b) In the E-Step we calculate the posterior distribution $Pr(h_i|x_i)$ over the hidden variable h_i for each data point x_i. The color of each curve corresponds to that of the original data point in (a). c) In the M-Step we use these distributions over h to update the estimate of the parameters $\theta = \{\mu, \sigma^2, \nu\}$.

The plan...


Normal distribution
a)
Problem 1
Unimodal
Mixture models
b)
mixture of Gaussians
Problem 2
Sensitive to outliers
Robust models
c)
t-distributions
Problem 3
Too many parameters in high dimensions
Subspace models
d)
PPCA, factor analysis
e)
Mixture of t-distributions
Mixture of factor analyzers
Mixture of robust subspace models
Robust subspace models

Structure

- Densities for classification
- Models with hidden variables
  - Mixture of Gaussians
  - t-distributions
  - Factor analysis
- EM algorithm in detail
- Applications

Factor Analysis

Compromise between

- Full covariance matrix (desirable but many parameters)
- Diagonal covariance matrix (no modelling of covariance)

Models full covariance in a subspace $\boldsymbol{\Phi}$

Mops everything else up with diagonal component $\boldsymbol{\Sigma}$

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Phi}^{T} + \boldsymbol{\Sigma}]$$

Data Density



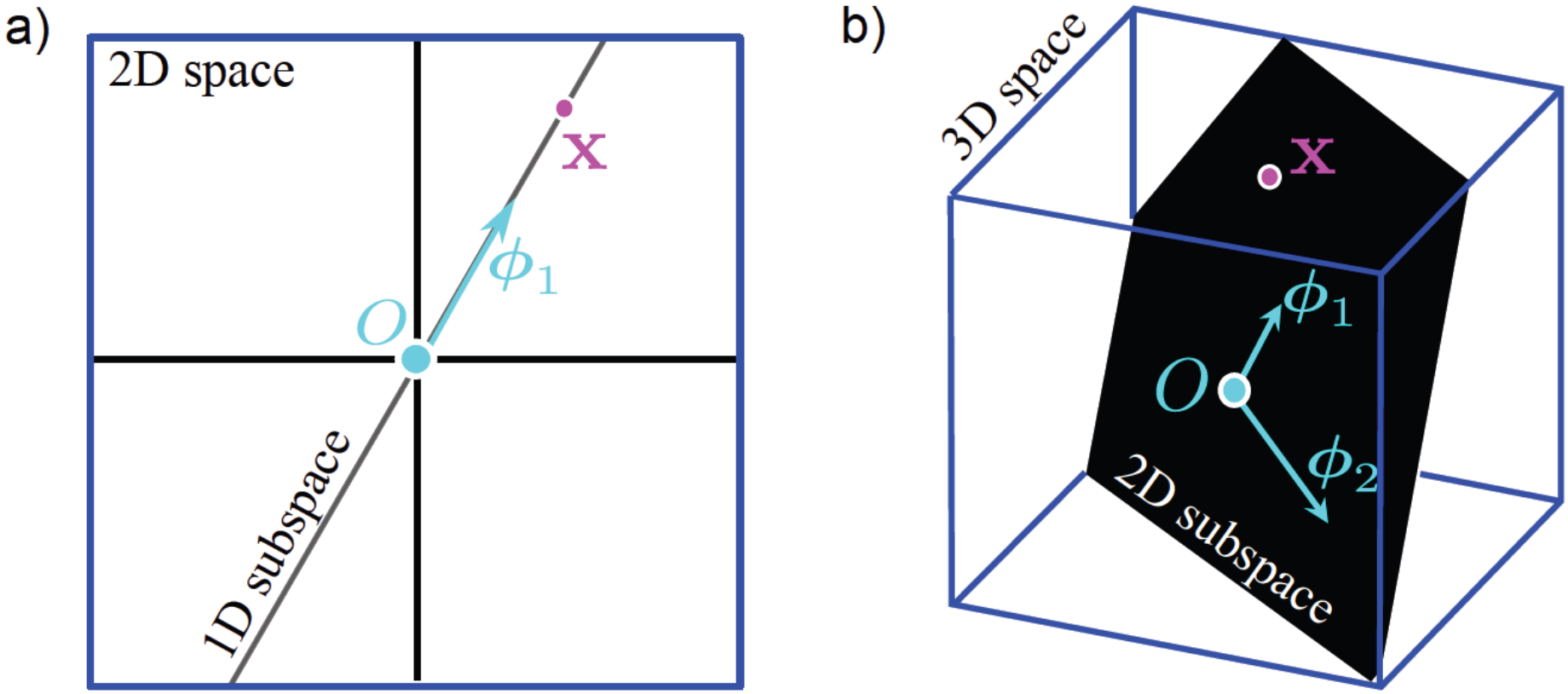


Consider modelling distribution of 100x100x3 RGB Images

Gaussian w/ spherical covariance:	1	covariance matrix parameters
Gaussian w/ diagonal covariance:	D_x	covariance matrix parameters
Full Gaussian:	$\sim D_x^2$	covariance matrix parameters
PPCA:	$\sim D_x D_h$	covariance matrix parameters
Factor Analysis:	$\sim D_x (D_h+1)$	covariance matrix parameters

(full covariance in subspace + diagonal)

Subspaces


a)
2D space
x
ϕ_1
O
1D subspace
b)
3D space
x
ϕ_1
O
ϕ_2
2D subspace

Factor Analysis

Compromise between

- Full covariance matrix (desirable but many parameters)
- Diagonal covariance matrix (no modelling of covariance)

Models full covariance in a subspace $\boldsymbol{\Phi}$

Mops everything else up with diagonal component Σ

$$Pr(\mathbf{x}) = \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Phi}^{T} + \boldsymbol{\Sigma}]$$

Factor Analysis as a Marginalization

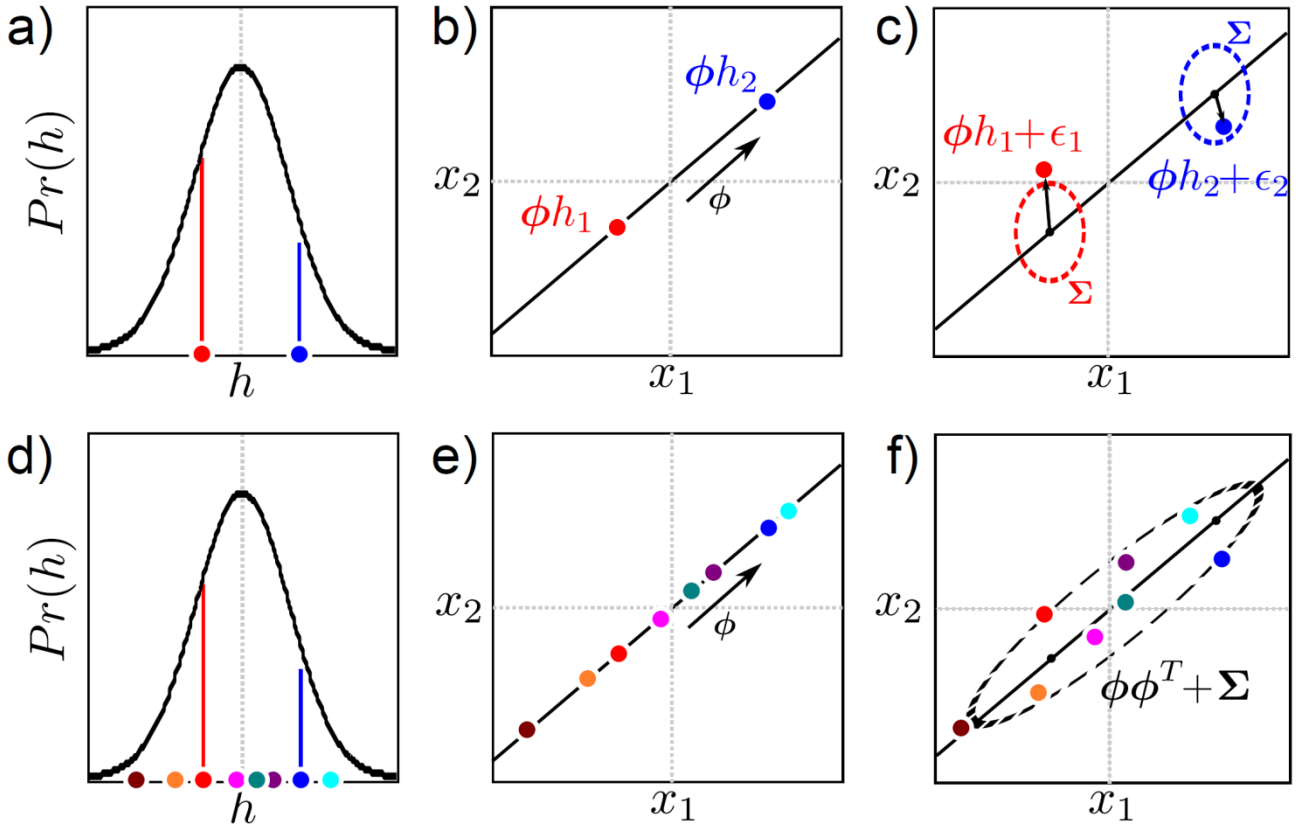
Let's define

$$\begin{aligned} Pr(\mathbf{x}|\mathbf{h}) &= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}, \boldsymbol{\Sigma}] \\ Pr(\mathbf{h}) &= \text{Norm}_{\mathbf{h}}[\mathbf{0}, \mathbf{I}] \end{aligned}$$

Then it can be shown (not obvious) that

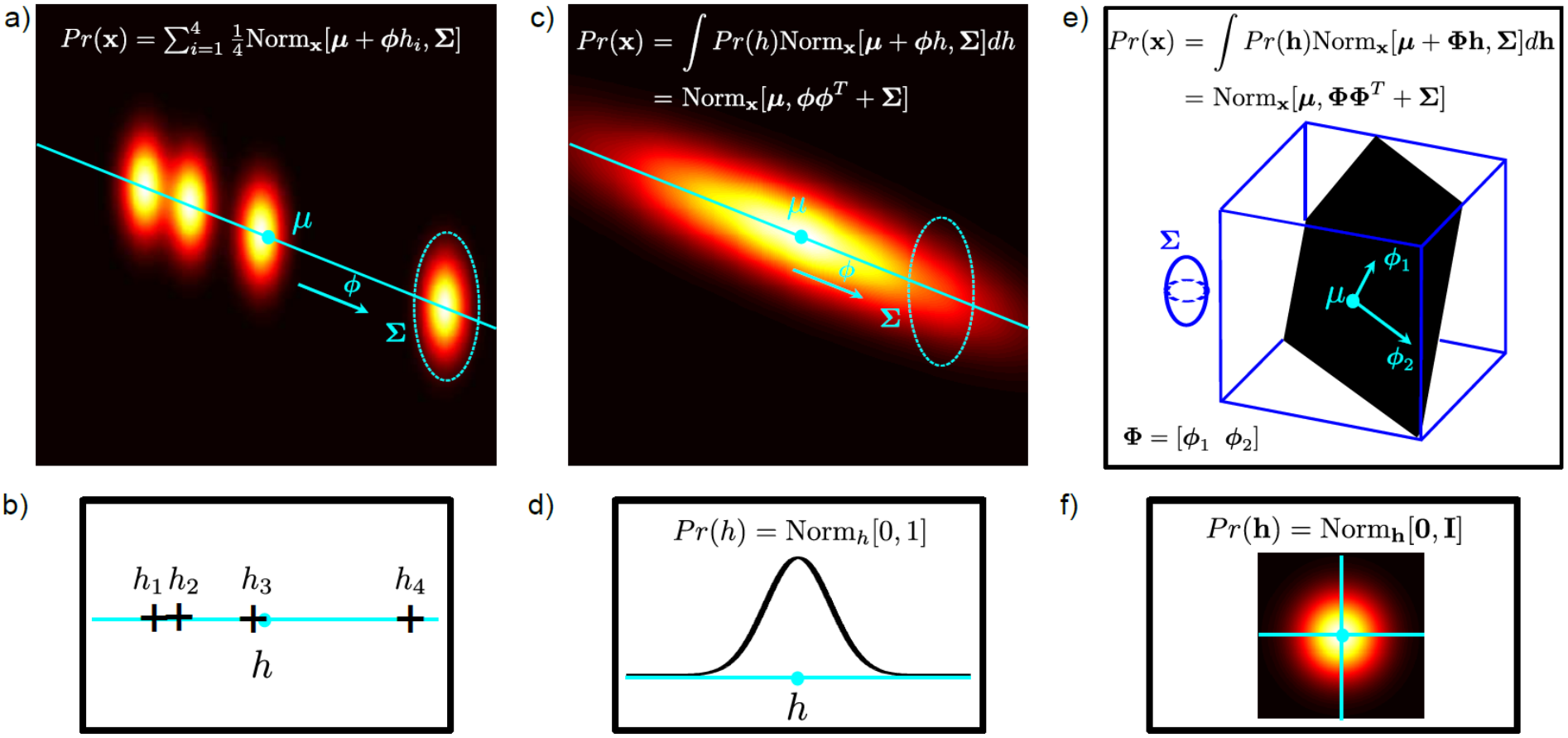
$$\begin{aligned} Pr(\mathbf{x}) = \int Pr(\mathbf{x}, \mathbf{h}) dh &= \int Pr(\mathbf{x}|\mathbf{h}) Pr(\mathbf{h}) d\mathbf{h} \\ &= \int \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}, \boldsymbol{\Sigma}] \text{Norm}_{\mathbf{h}}[\mathbf{0}, \mathbf{I}] d\mathbf{h} \\ &= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \Sigma] \end{aligned}$$

Sampling from factor analyzer



$$Pr(\mathbf{x}|\mathbf{h}) = \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}, \boldsymbol{\Sigma}]$$
$$Pr(\mathbf{h}) = \text{Norm}_{\mathbf{h}}[\mathbf{0}, \mathbf{I}]$$

Factor analysis vs. MoG



E-Step

- Compute posterior over hidden variables

$$\begin{aligned}
\hat{q}(\mathbf{h}_i) &= Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta}^{[t]}) \\
&= \frac{Pr(\mathbf{x}_i|\mathbf{h}_i, \boldsymbol{\theta}^{[t]})Pr(\mathbf{h}_i)}{Pr(\mathbf{x}_i|\boldsymbol{\theta}^{[t]})} \\
&= \frac{\mathrm{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}_i, \boldsymbol{\Sigma}]\mathrm{Norm}_{\mathbf{h}_i}[\mathbf{0}, \mathbf{I}]}{Pr(\mathbf{x}_i|\boldsymbol{\theta}^{[t]})} \\
&= \mathrm{Norm}_{\mathbf{h}_i}[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi} + \mathbf{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi} + \mathbf{I})]
\end{aligned}$$

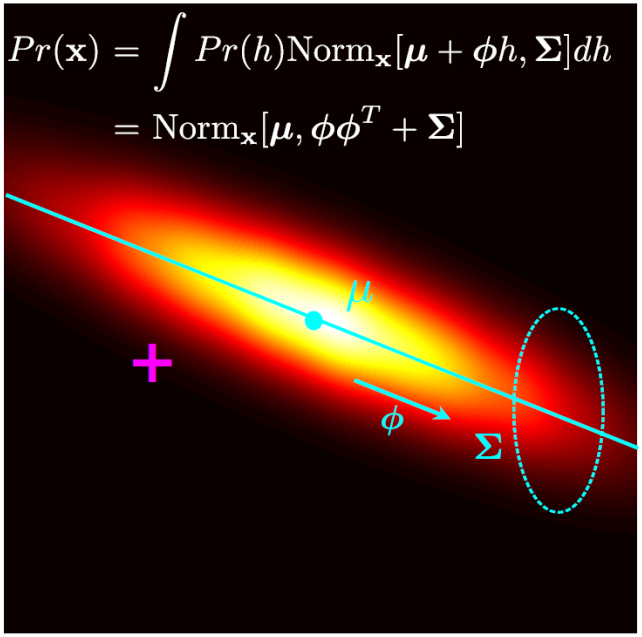
- Compute expectations needed for M-Step

$$\begin{aligned}
\mathrm{E}[\mathbf{h}_i] &= (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi} + \mathbf{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \\
\mathrm{E}[\mathbf{h}_i\mathbf{h}_i^T] &= E\left[(\mathbf{h}_i - \mathrm{E}[\mathbf{h}_i])(\mathbf{h}_i - \mathrm{E}[\mathbf{h}_i])^T\right] + \mathrm{E}[\mathbf{h}_i]\mathrm{E}[\mathbf{h}_i]^T \\
&= (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi} + \mathbf{I})^{-1} + \mathrm{E}[\mathbf{h}_i]\mathrm{E}[\mathbf{h}_i]^T.
\end{aligned}$$

E-Step

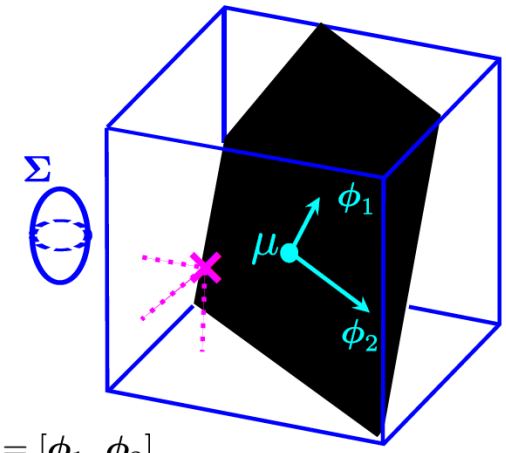
a)

$$Pr(\mathbf{x}) = \int Pr(h)\text{Norm}_{\mathbf{x}}[\boldsymbol{\mu} + \boldsymbol{\phi}h, \boldsymbol{\Sigma}]dh$$
$$= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\phi}\boldsymbol{\phi}^T + \boldsymbol{\Sigma}]$$



c)

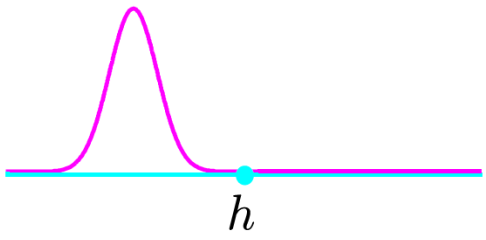
$$Pr(\mathbf{x}) = \int Pr(\mathbf{h})\text{Norm}_{\mathbf{x}}[\boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{h}, \boldsymbol{\Sigma}]d\mathbf{h}$$
$$= \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}]$$



$$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \quad \boldsymbol{\phi}_2]$$

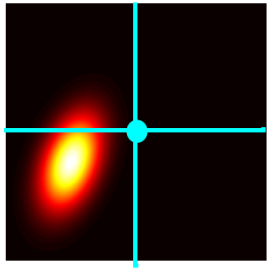
b)

$$Pr(h|\mathbf{x}) = \text{Norm}_h[(\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi} + 1)^{-1}\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\boldsymbol{\phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi})^{-1}]$$



d)

$$Pr(h|\mathbf{x}) = \text{Norm}_h[(\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi} + \mathbf{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi})^{-1}]$$



M-Step

- Optimize parameters w.r.t. EM bound

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{[t+1]} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int \hat{q}_i(\mathbf{h}_i) \log\left[Pr(\mathbf{x}, \mathbf{h}_i, \boldsymbol{\theta})\right] \; d\mathbf{h}_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int \hat{q}_i(\mathbf{h}_i) \left[\log\left[Pr(\mathbf{x}|\mathbf{h}_i, \boldsymbol{\theta})\right] + \log\left[Pr(\mathbf{h}_i)\right]\right] \; d\mathbf{h}_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int \hat{q}_i(\mathbf{h}_i) \log\left[Pr(\mathbf{x}|\mathbf{h}_i, \boldsymbol{\theta})\right] \; d\mathbf{h}_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} E\left[\log Pr(\mathbf{x}|\mathbf{h}_i, \boldsymbol{\theta})\right] \right],
\end{aligned}
$$

where....

$$
\log Pr(\mathbf{x}_i|\mathbf{h}_i) = -\frac{D\log(2\pi) + \log|\boldsymbol{\Sigma}| + (\mathbf{x}_i - \boldsymbol{\mu} - \boldsymbol{\Phi}\mathbf{h}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu} - \boldsymbol{\Phi}\mathbf{h}_i)}{2}
$$

M-Step

- Optimize parameters w.r.t. EM bound

$$\hat{\boldsymbol{\theta}}^{[t+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} E\left[\log Pr(\mathbf{x}|\mathbf{h}_i, \boldsymbol{\theta})\right] \right]$$

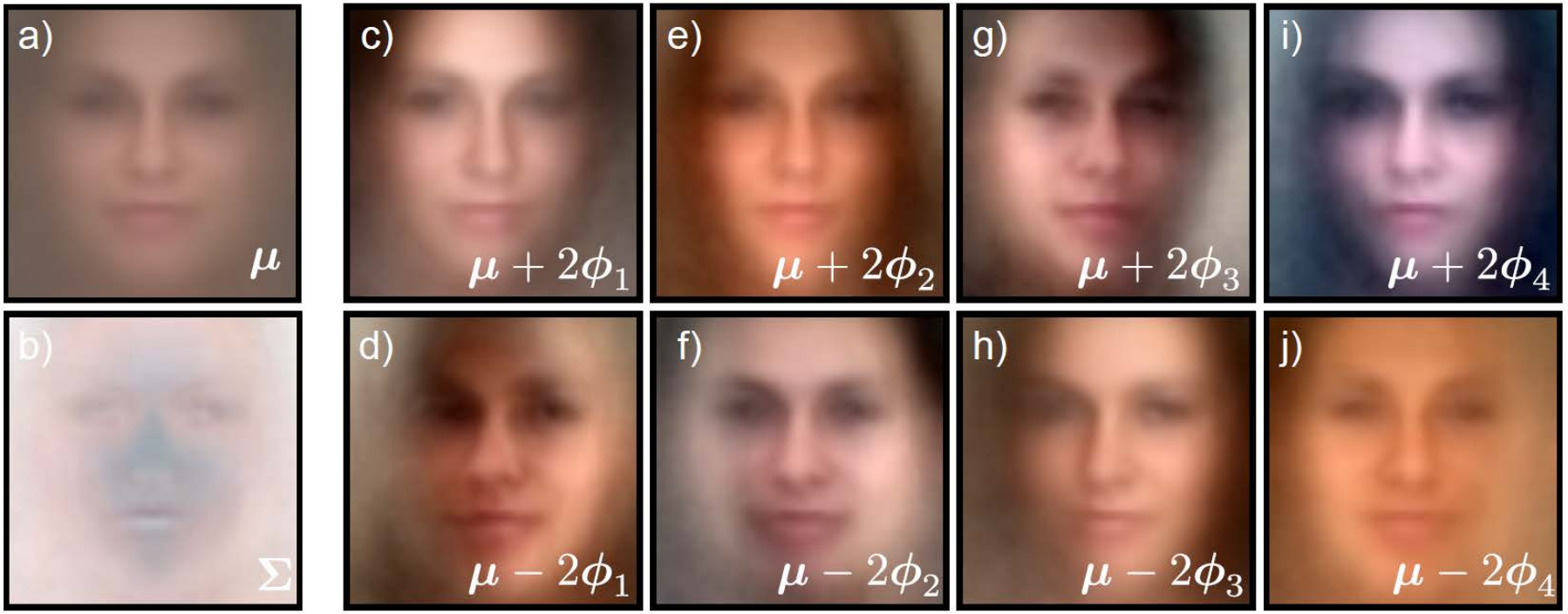
where...

$$\log Pr(\mathbf{x}_i|\mathbf{h}_i) = -\frac{D\log(2\pi) + \log|\boldsymbol{\Sigma}| + (\mathbf{x}_i - \boldsymbol{\mu} - \boldsymbol{\Phi}\mathbf{h}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu} - \boldsymbol{\Phi}\mathbf{h}_i)}{2}$$

giving expressions:

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \frac{\sum_{i=1}^{I}\mathbf{x}_i}{I} \\
\hat{\boldsymbol{\Phi}} &= \left(\sum_{i=1}^{I}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})\mathrm{E}[\mathbf{h}_i]^T\right)\left(\sum_{i=1}^{I}\mathrm{E}[\mathbf{h}_i\mathbf{h}_i^T]\right)^{-1} \\
\hat{\boldsymbol{\Sigma}} &= \frac{1}{I}\sum_{i=1}^{I}\mathrm{diag}\left[(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\Phi}}\mathrm{E}[\mathbf{h}_i]\mathbf{x}^T\right]
\end{aligned}$$

Face model


a)
μ
c)
$\mu + 2\phi_1$
e)
$\mu + 2\phi_2$
g)
$\mu + 2\phi_3$
i)
$\mu + 2\phi_4$
b)
Σ
d)
$\mu - 2\phi_1$
f)
$\mu - 2\phi_2$
h)
$\mu - 2\phi_3$
j)
$\mu - 2\phi_4$

Sampling from 10 parameter model

To generate:

- Choose factor loadings, $\mathbf{h}_i$ from standard normal distribution
- Multiply by factors, $\boldsymbol{\Phi}$
- Add mean, $\boldsymbol{\mu}$
- (should add random noise component ε_i w/ diagonal cov $\boldsymbol{\Sigma}$)

Combining Models

Can combine three models in various combinations

- Mixture of t-distributions = mixture of Gaussians + t-distributions
- Robust subspace models = t-distributions + factor analysis
- Mixture of factor analyzers = mixture of Gaussians + factor analysis

Or combine all models to create mixture of robust subspace models

$$Pr(\mathbf{x}) = \sum_{k=1}^{K} \lambda_k \text{Stud}_{\mathbf{x}}\left[\boldsymbol{\mu}_k, \boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^T + \boldsymbol{\Sigma}_k, \nu_k\right]$$

Structure

- Densities for classification
- Models with hidden variables
  - Mixture of Gaussians
  - t-distributions
  - Factor analysis
- EM algorithm in detail
- Applications

Expectation Maximization

Problem: Optimize cost functions of the form

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{I} \log \left[\sum_{h} Pr(\mathbf{x}_i, h_i) \right] \quad \leftarrow \text{Discrete case}$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{I} \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i) d\mathbf{h}_i \right] \quad \leftarrow \text{Continuous case}$$

Solution: Expectation Maximization (EM) algorithm

(Dempster, Laird and Rubin 1977)

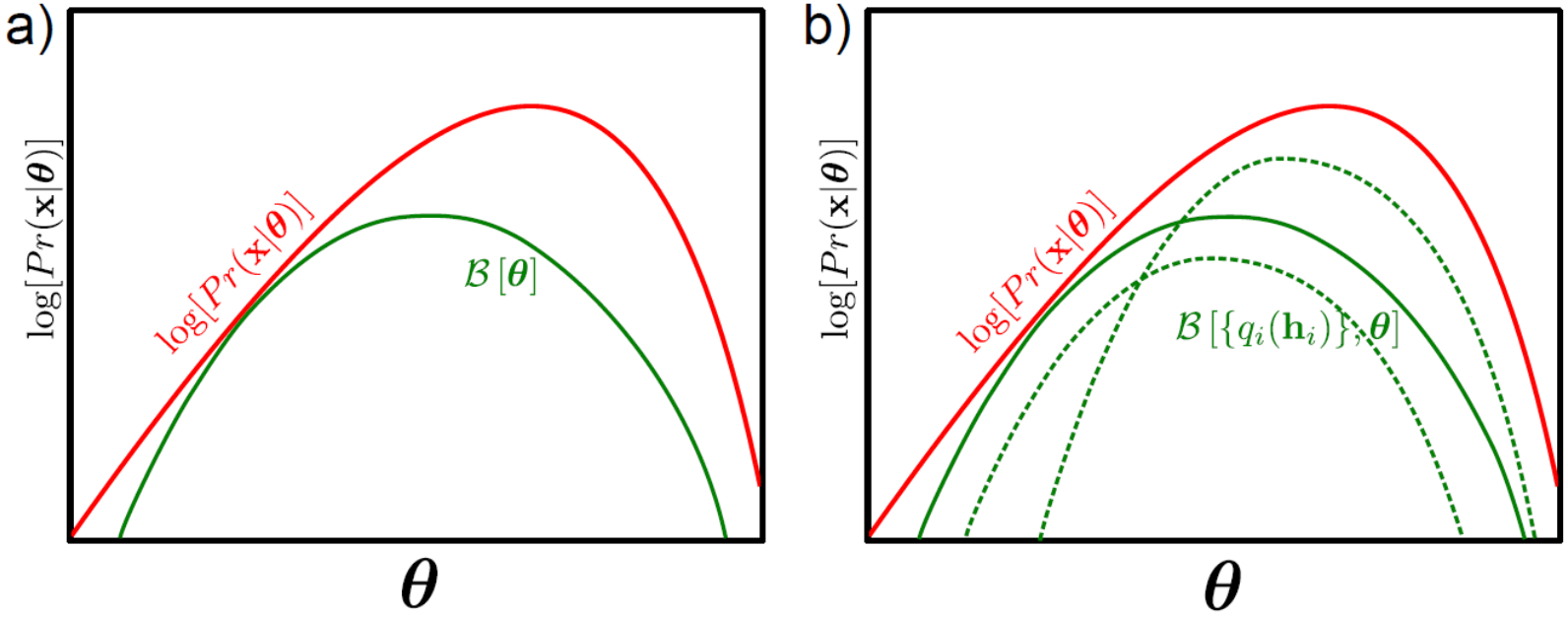
Key idea: Define lower bound on log-likelihood and increase at each iteration

Expectation Maximization

Defines a lower bound on log likelihood and increases bound iteratively

$$\begin{aligned}
\mathcal{B}[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}] &= \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_i \\
&\leq \sum_{i=1}^{I} \log \left[\int q_i(\mathbf{h}_i) \frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)} d\mathbf{h}_i \right] \\
&= \sum_{i=1}^{I} \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \, d\mathbf{h}_i \right],
\end{aligned}$$

Lower bound



$$\mathcal{B}[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}] = \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_i$$

Lower bound is a *function* of parameters θ and a *set* of probability distributions $\{q_i(\mathbf{h}_i)\}$

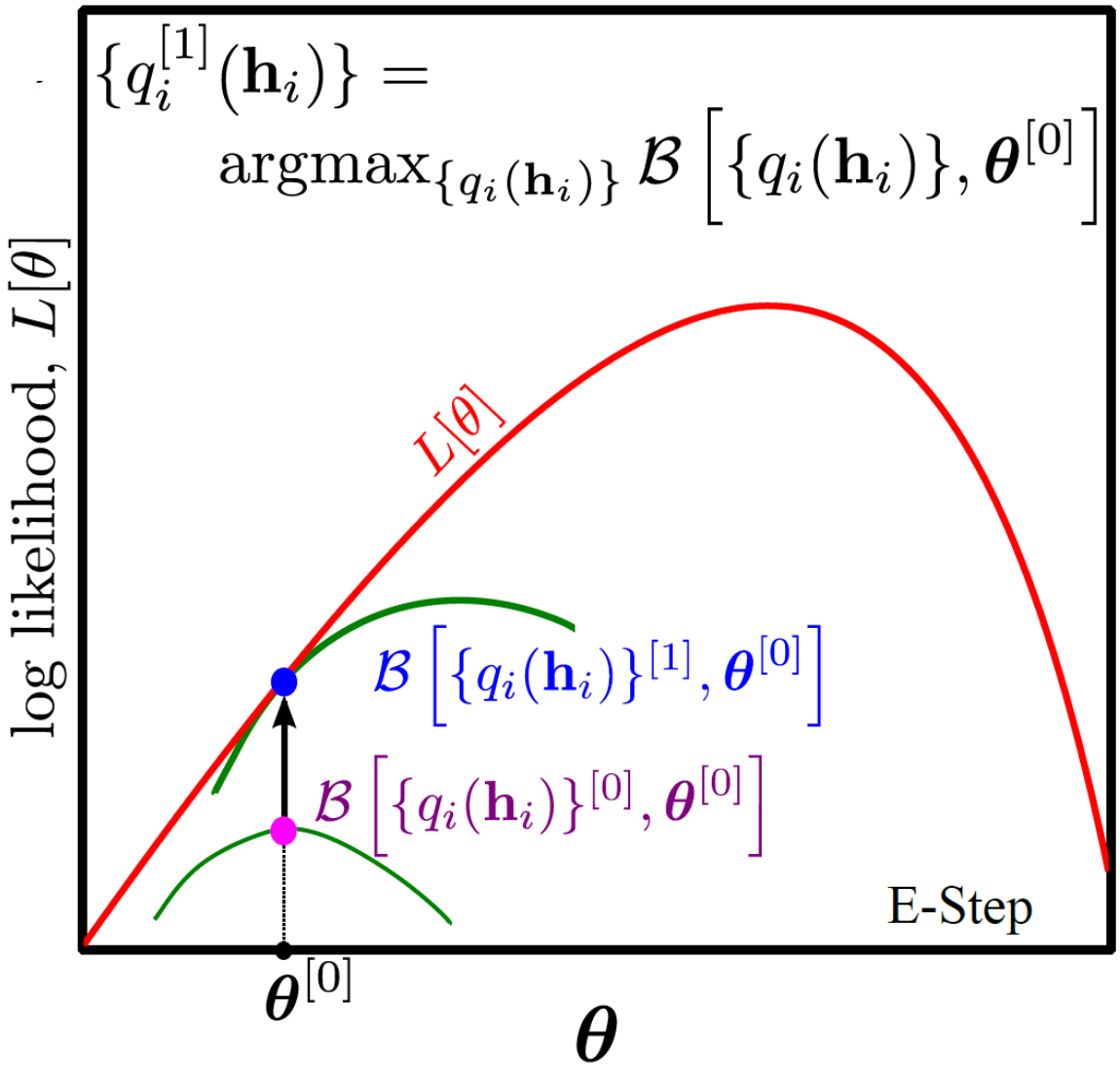
E-Step & M-Step

E-Step – Maximize bound w.r.t. distributions $\{q_i(\mathbf{h}_i)\}$

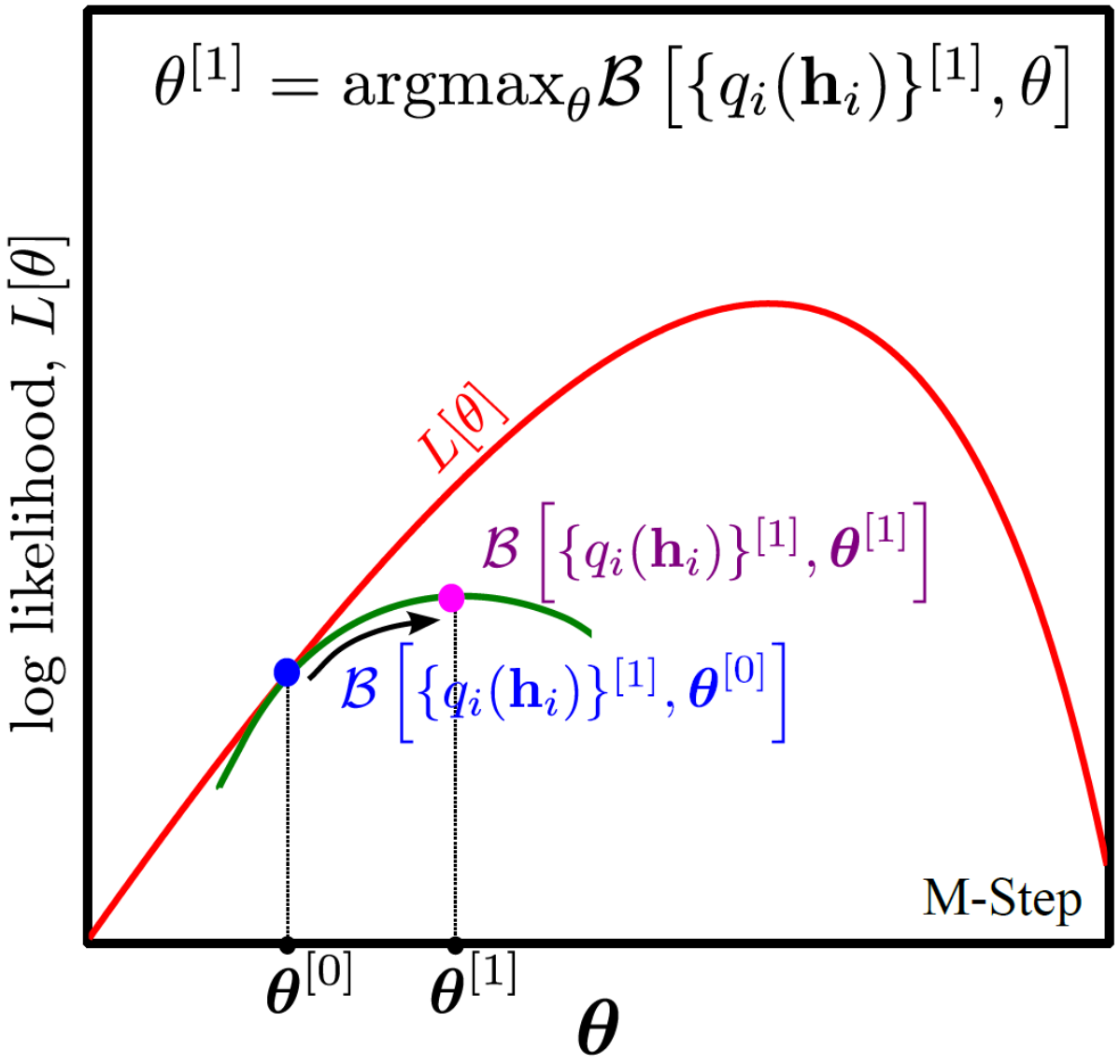
$$q_i^{[t]}[\mathbf{h}_i] = \underset{q_i[\mathbf{h}_i]}{\operatorname{argmax}} \left[\mathcal{B}[\{q_i(\mathbf{h}_i)\}, \theta^{[t-1]}] \right]$$

M-Step – Maximize bound w.r.t. parameters θ

$$\boldsymbol{\theta}^{[t]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\mathcal{B}[\{q_i^{[t]}(\mathbf{h}_i)\}, \boldsymbol{\theta}] \right]$$

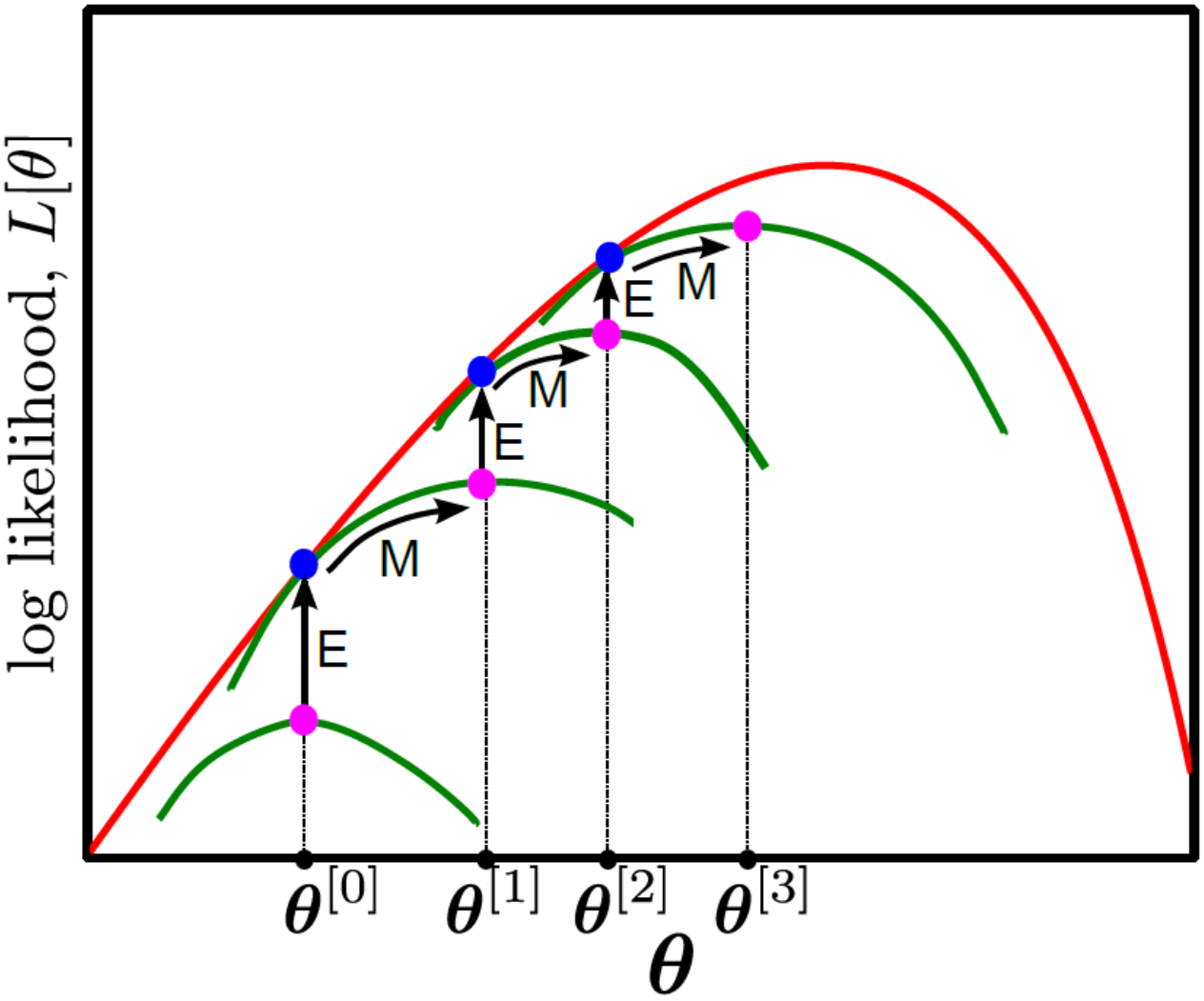


E-Step: Update $\{q_i[\mathbf{h}_i]\}$ so that bound equals log likelihood for this θ



M-Step: Update θ to maximum

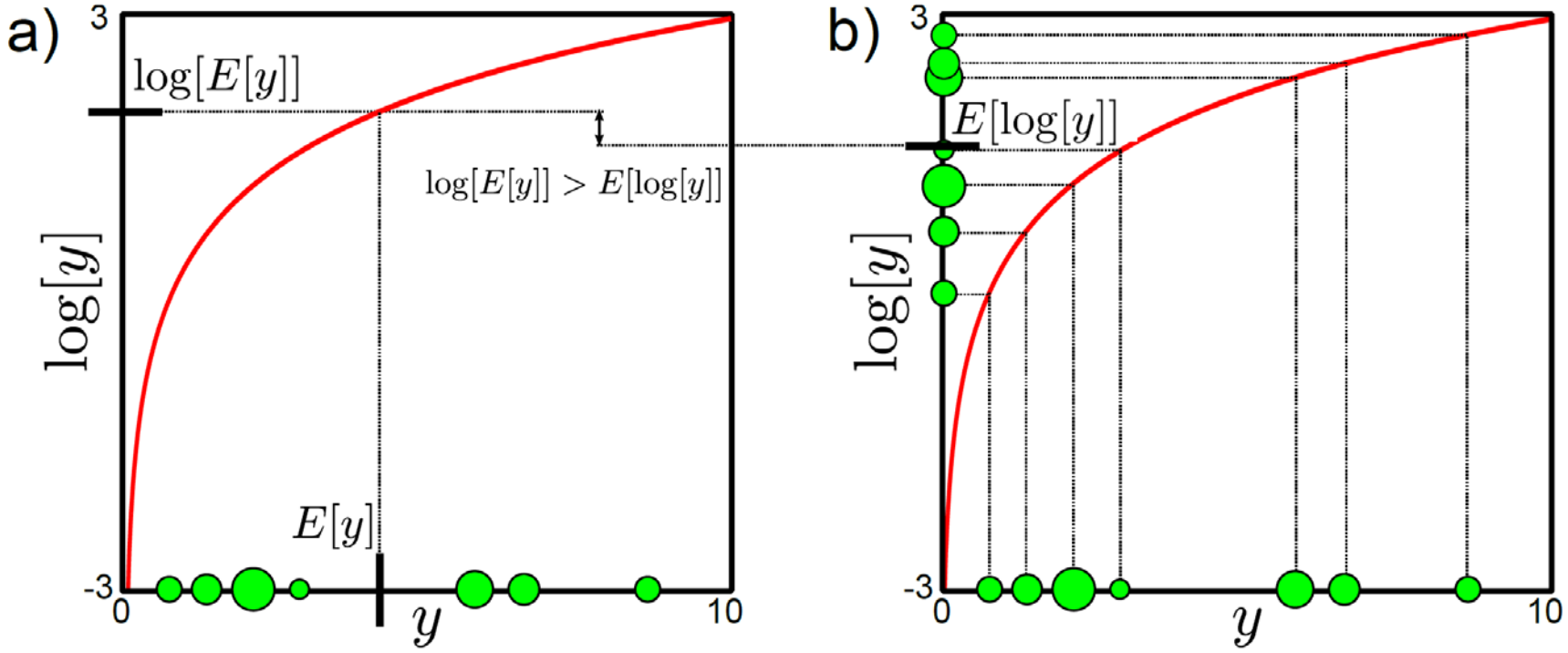
Missing Parts of Argument



Missing Parts of Argument

1. Show that this is a lower bound
2. Show that the E-Step update is correct
3. Show that the M-Step update is correct

Jensen's Inequality



$$E\left[\log[y]\right] \leq \log\left[E[y]\right] \quad \text{or...} \quad \sum_y q(y)\log[y] \leq \log\left[\sum_y q(y)y\right]$$

similarly $$\int q(y)\log[y]dy \leq \int \log[q(y)]dy$$

Lower Bound for EM Algorithm

$$\begin{aligned}
\mathcal{B}[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}] &= \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log\left[\frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right] d\mathbf{h}_i \\
&\leq \sum_{i=1}^{I} \log\left[\int q_i(\mathbf{h}_i) \frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)} d\mathbf{h}_i\right] \\
&= \sum_{i=1}^{I} \log\left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \, d\mathbf{h}_i\right],
\end{aligned}$$

Where we have used Jensen's inequality

$$\int q(y) \log[y] dy \leq \int \log[q(y)] dy$$

E-Step – Optimize bound w.r.t $\{q_i(\mathbf{h}_i)\}$

$$\begin{aligned}
\mathcal{B}[\{q_i(\mathbf{h}_i)\},\boldsymbol{\theta}] &= \sum_{i=1}^{I}\int q_i(\mathbf{h}_i)\log\left[\frac{Pr(\mathbf{x}_i,\mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right]\,d\mathbf{h}_i \\
&= \sum_{i=1}^{I}\int q_i(\mathbf{h}_i)\log\left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i,\boldsymbol{\theta})Pr(\mathbf{x}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right]\,d\mathbf{h}_i \\
&= \sum_{i=1}^{I}\int q_i(\mathbf{h}_i)\log\left[Pr(\mathbf{x}_i|\boldsymbol{\theta})\right]\,d\mathbf{h}_i - \sum_{i=1}^{I}\int q_i(\mathbf{h}_i)\log\left[\frac{q_i(\mathbf{h}_i)}{Pr(\mathbf{h}_i|\mathbf{x}_i,\boldsymbol{\theta})}\right]\,d\mathbf{h}_i \\
&= \sum_{i=1}^{I}\log\left[Pr(\mathbf{x}_i|\boldsymbol{\theta})\right] - \sum_{i=1}^{I}\int q_i(\mathbf{h}_i)\log\left[\frac{q_i(\mathbf{h}_i)}{Pr(\mathbf{h}_i|\mathbf{x}_i,\boldsymbol{\theta})}\right]\,d\mathbf{h}_i \qquad (7.46)
\end{aligned}$$

Constant w.r.t. q(h)

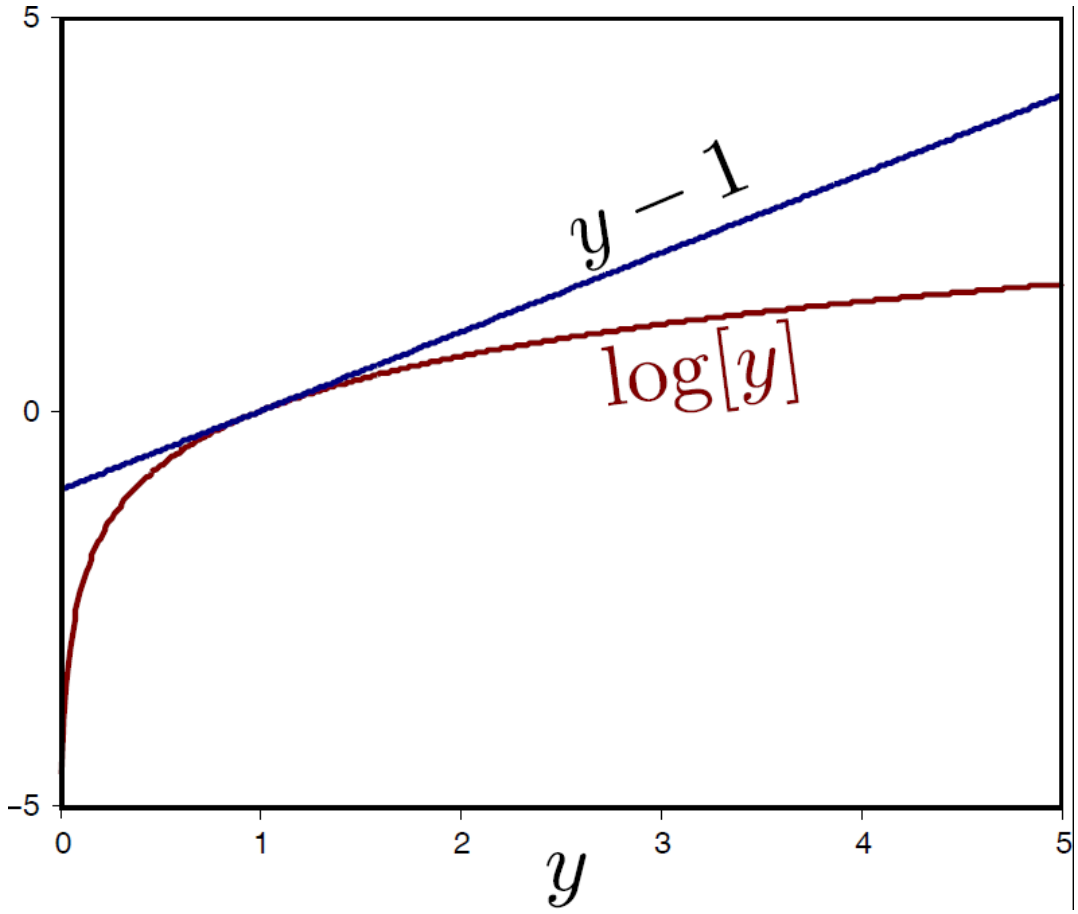
Only this term matters

E-Step

$$\begin{aligned}
\hat{q}_i(\mathbf{h}_i) &= \underset{q_i(\mathbf{h}_i)}{\text{argmax}} \left[-\int q_i(\mathbf{h}_i) \log \left[\frac{q_i(\mathbf{h}_i)}{Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta})} \right] d\mathbf{h}_i \right] \\
&= \underset{q_i(\mathbf{h}_i)}{\text{argmax}} \left[\int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_i \right] \\
&= \underset{q_i(\mathbf{h}_i)}{\text{argmin}} \left[-\int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_i \right]
\end{aligned}$$

Kullback Leibler divergence – distance between probability distributions . We are maximizing the negative distance (i.e. Minimizing distance)

Use this relation



$\log[y] \leq y-1$

Kullback Leibler Divergence

$$\begin{aligned}\int q_i(\mathbf{h}_i)\log\left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i,\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right]d\mathbf{h}_i &\leq \int q_i(\mathbf{h}_i)\left(\frac{Pr(\mathbf{h}_i|\mathbf{x}_i,\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}-1\right)d\mathbf{h}_i \\ &= \int Pr(\mathbf{h}_i|\mathbf{x}_i,\boldsymbol{\theta})-q_i(\mathbf{h}_i)\,d\mathbf{h}_i \\ &= 1-1=0,\end{aligned}$$

So the cost function must be positive

$$\hat{q}_i(\mathbf{h}_i) = \underset{q_i(\mathbf{h}_i)}{\operatorname{argmin}}\left[-\int q_i(\mathbf{h}_i)\log\left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i,\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right]d\mathbf{h}_i\right]$$

In other words, the best we can do is choose $q_i(\mathbf{h}_i)$ so that this is zero

E-Step

So the cost function must be positive

$$\hat{q}_i(\mathbf{h}_i) = \underset{q_i(\mathbf{h}_i)}{\operatorname{argmin}} \left[-\int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_i \right]$$

The best we can do is choose $q_i(\mathbf{h}_i)$ so that this is zero.

How can we do this? Easy – choose posterior $Pr(\mathbf{h}|\mathbf{x})$

$$\begin{aligned} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_i &= \int Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta}) \log \left[\frac{Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta})}{Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta})} \right] d\mathbf{h}_i \\ &= \int Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta}) \log [1] \; d\mathbf{h}_i = 0. \end{aligned}$$

M-Step – Optimize bound w.r.t. θ

$$
\begin{aligned}
\boldsymbol{\theta}^{[t]} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\mathcal{B}[\{q_i^{[t]}(\mathbf{h}_i)\}, \boldsymbol{\theta}] \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int q_i^{[t]}(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i^{[t]}(\mathbf{h}_i)} \right] \, d\mathbf{h}_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int q_i^{[t]}(\mathbf{h}_i) \log \left[Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \right] - q_i^{[t]}(\mathbf{h}_i) \log \left[q_i^{[t]}(\mathbf{h}_i) \right] \, d\mathbf{h}_i \right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int q_i^{[t]}(\mathbf{h}_i) \log \left[Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \right] \, d\mathbf{h}_i \right]
\end{aligned}
$$

In the M-Step we optimize expected joint log likelihood with respect to parameters θ (Expectation w.r.t distribution from E-Step)

E-Step & M-Step

E-Step – Maximize bound w.r.t. distributions $q_i(\mathbf{h}_i)$

$$\hat{q}_i(\mathbf{h}_i) = Pr(\mathbf{h}_i|\mathbf{x}_i, \boldsymbol{\theta}^{[t]}) = \frac{Pr(\mathbf{x}_i|\mathbf{h}_i, \boldsymbol{\theta}^{[t]})Pr(\mathbf{h}_i)}{Pr(\mathbf{x}_i)}$$

M-Step – Maximize bound w.r.t. parameters θ

$$\hat{\boldsymbol{\theta}}^{[t+1]} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \int \hat{q}_i(\mathbf{h}_i) \log \left[Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \right] \, d\mathbf{h}_i \right]$$

Structure

- Densities for classification
- Models with hidden variables
  - Mixture of Gaussians
  - t-distributions
  - Factor analysis
- EM algorithm in detail
- Applications

Face Detection

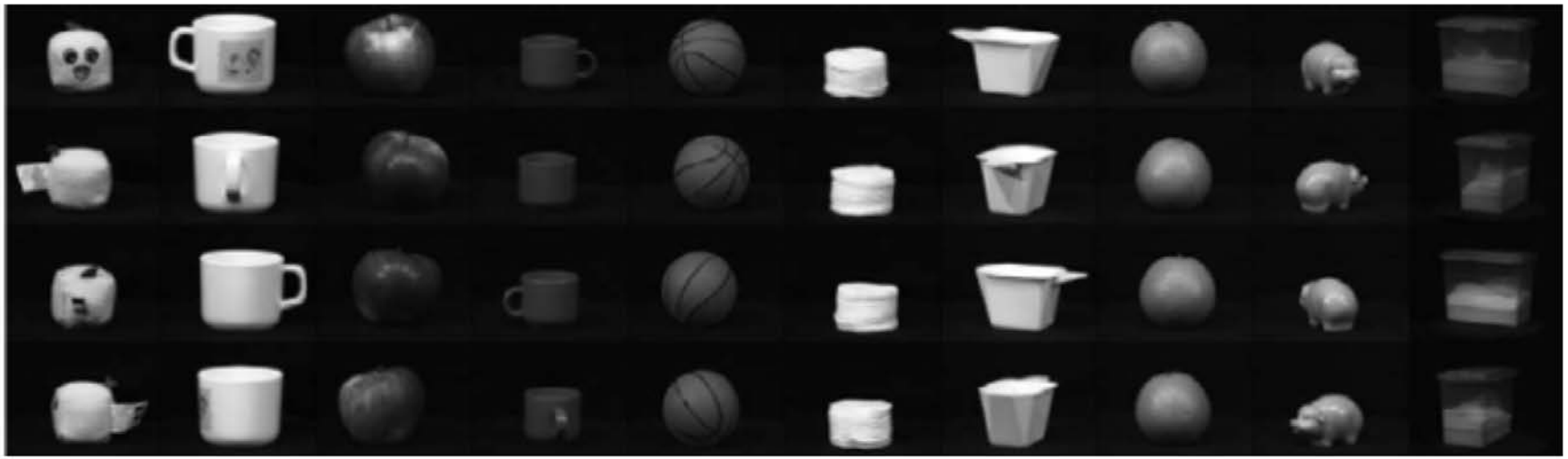
	Color	Grayscale	Equalized
Single Gaussian	76%	79%	80%
Mixture of 10 Gaussians	81%	85%	89%

(not a very realistic model – face detection really done using discriminative methods)

Object recognition

Aeschliman *et al.* (2010)

a)



b)

74	66	66	90	61	97	47	62	43	71
74	25	44	91	43	80	63	24	08	55

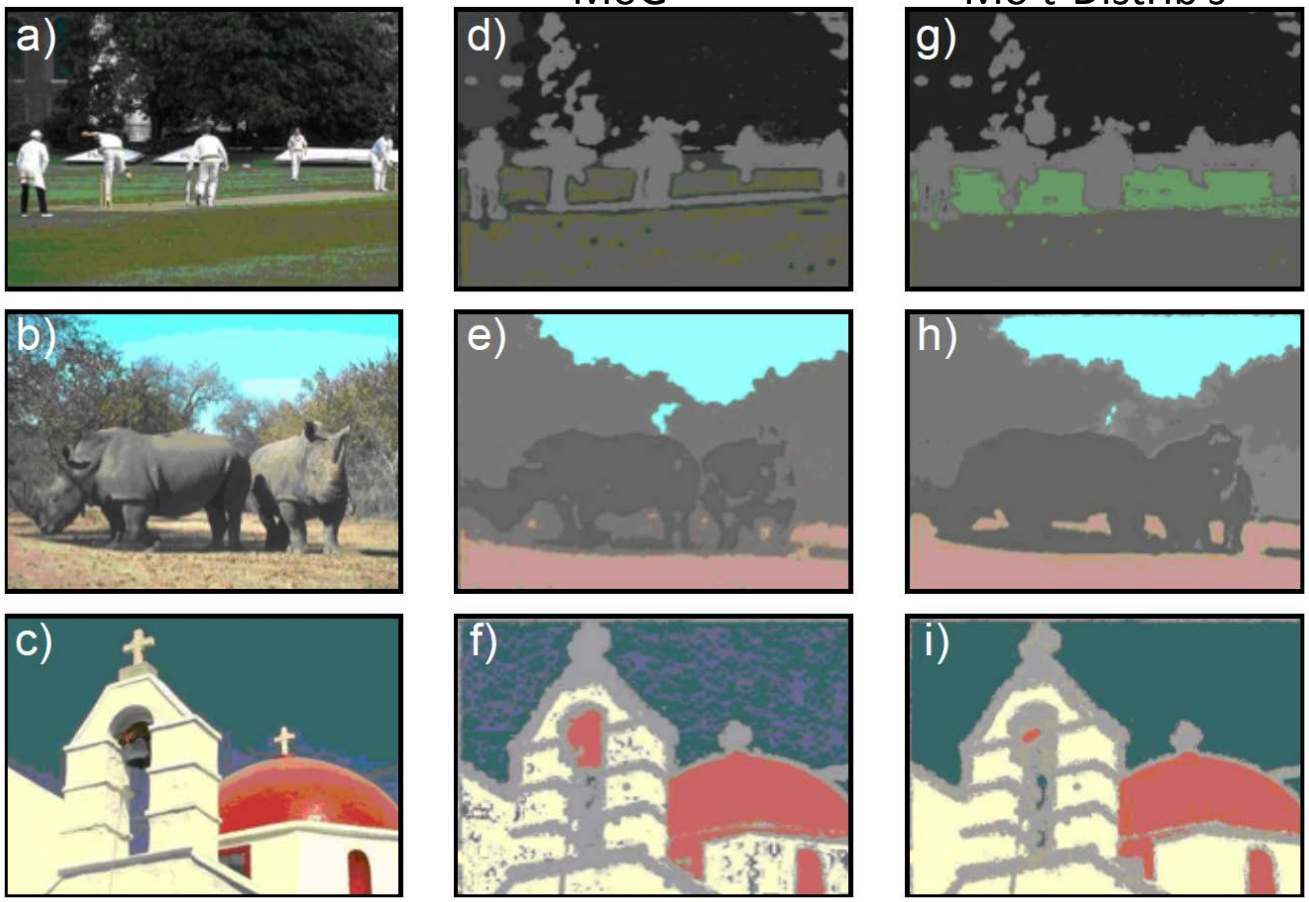
t-distributions

normal distributions

Model as J independent patches

$$Pr(\mathbf{x}_i|w=m) = \prod_{j=1}^{J} \text{Stud}_{\mathbf{x}_{ij}}[\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}, \nu_{jm}]$$

Segmentation

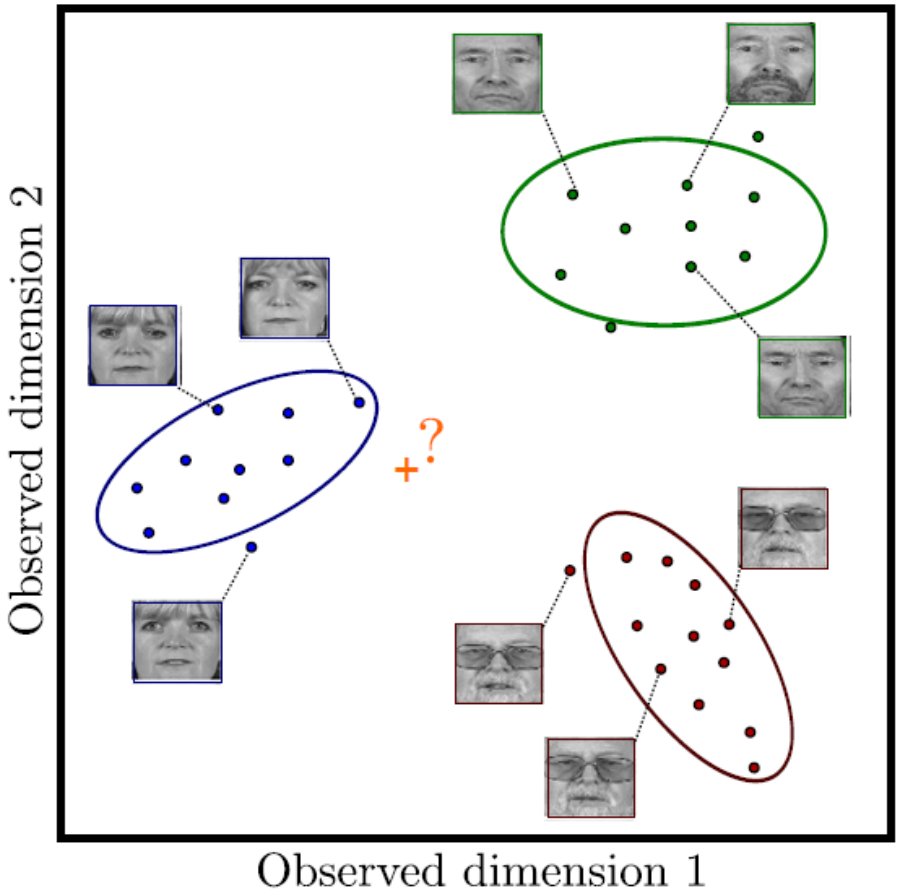


$$Pr(w_n) = \text{Cat}_{w_n}[\boldsymbol{\lambda}]$$
$$Pr(\mathbf{x}_i|w_i = k) = \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$$

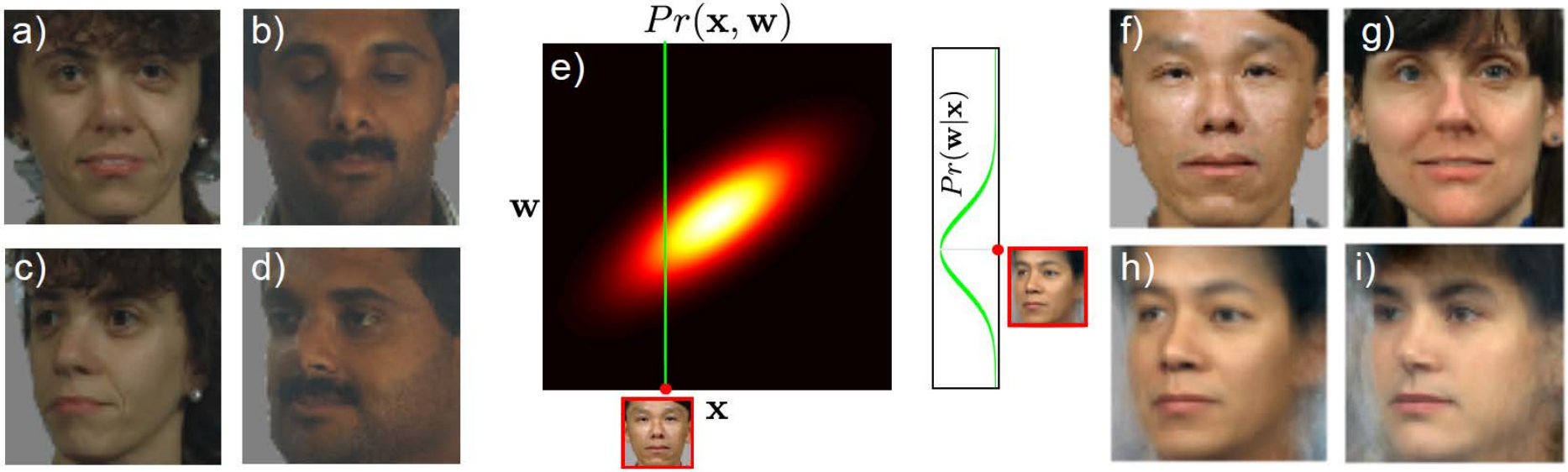
Sfikas *et al.* (2007) ©IEEE 2007

Face Recognition

Figure 6.28 Face recognition. Our goal is to take the RGB values of a facial image $\mathbf{x}$ and assign a label $y \in \{1 \ldots K\}$. Since the data is high dimensional , we model the class conditional density function $Pr(\mathbf{x}|y=k)$ for each individual in the database as a factor analyzer. To classify a new face, we apply Bayes' rule with suitable priors $Pr(y)$ to compute the posterior distribution $Pr(y|\mathbf{x})$. Finally, we choose the label $\hat{y} = \arg\max_k Pr(y=k|\mathbf{x})$ that maximizes the posterior. This is not suitable if there are not plentiful training examples of each individual or where there are significant pose or lighting changes.

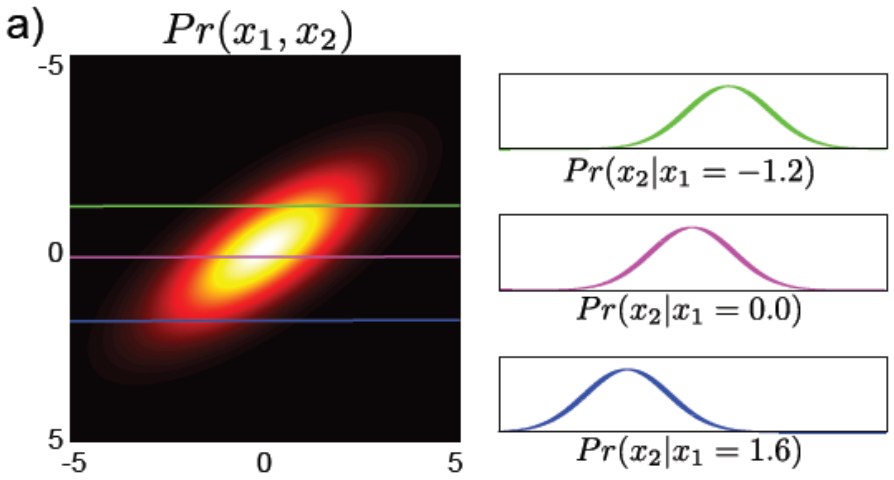


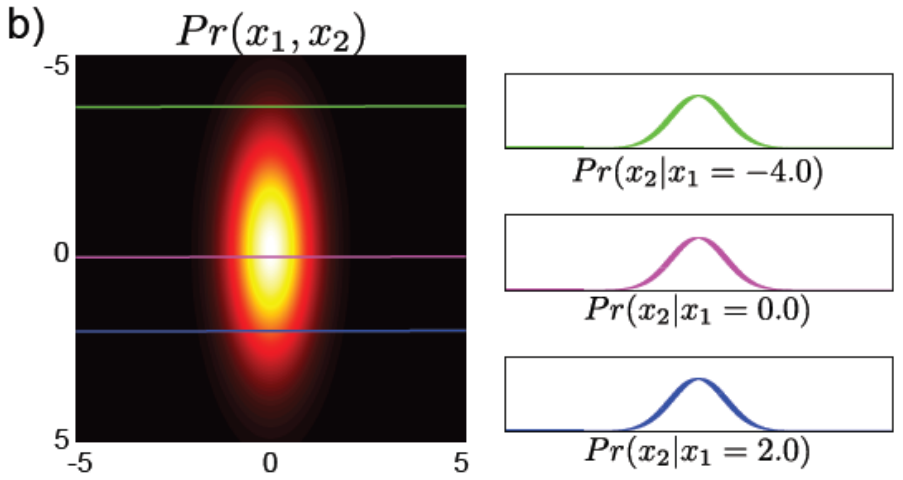
Pose Regression



Training data	Learning	Inference
Frontal faces x Non-frontal faces w	Fit factor analysis model to joint distribution of x,w	Given new x, infer w (cond. dist of normal)

Conditional Distributions





If

$$Pr(\mathbf{x}) = Pr\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right) = \text{Norm}_{\mathbf{x}}\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}^T\\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right)$$

then

$$\begin{aligned} Pr(\mathbf{x}_1|\mathbf{x}_2) &= \text{Norm}_{\mathbf{x}_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\right)\\ Pr(\mathbf{x}_2|\mathbf{x}_1) &= \text{Norm}_{\mathbf{x}_2}\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}^T\right) \end{aligned}$$

Modeling transformations with hidden variables

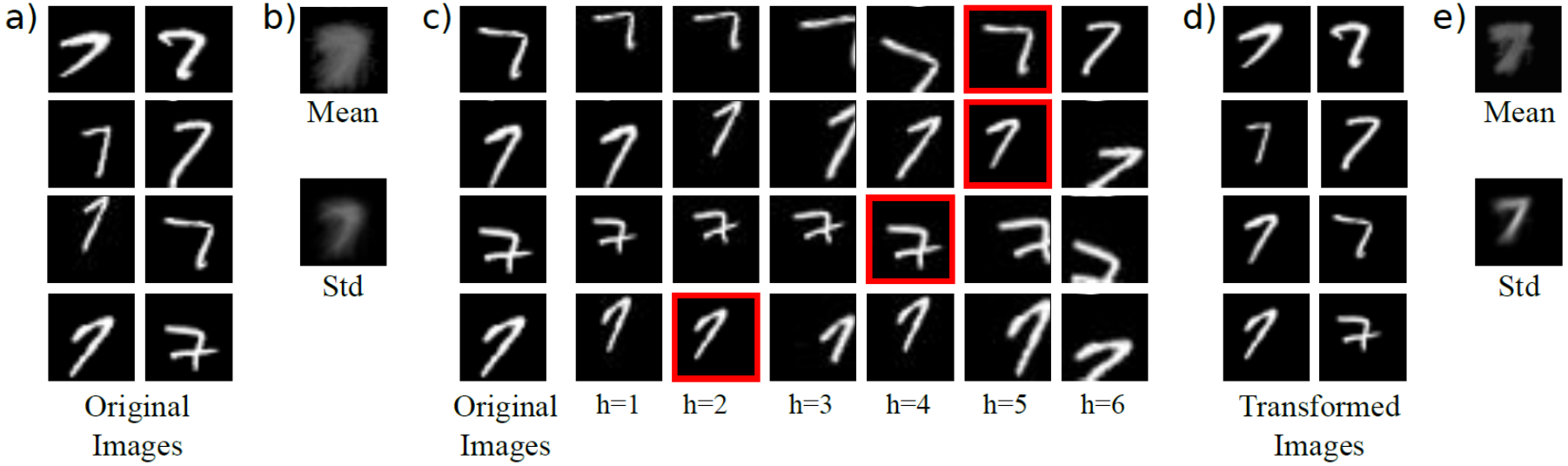


Figure 7.30 Modeling transformations with hidden variables. a) Original set of digit images are only weakly aligned. b) Mean and standard deviation images are consequently blurred out. The probability density model does not fit well. c) Each possible value of a discrete hidden variable represents a different transformation (here inverse transformations are shown). Red square highlights most likely choice of hidden variable after 10 iterations. d) Transformed digits (based on most likely hidden variable). d) New mean and standard deviation images are more focussed: the probability density function fits better.