Graphical Models, Inference and Learning

Brendan J. Frey
University of Toronto

Tutorial paper available at:

www.psi.toronto.edu/~frey/stuff/tutorial.ps.gz

Acknowledgements

- Nebojsa Jojic, Microsoft Research
- Resources:
  - Neal and Hinton 93, A new view of the EM algorithm
  - Jordan 98 (ed), Inference and Learning in Graphical Models
  - Wiberg, Loeliger, Koetter 95, The sum-product algorithm for error-correcting decoding
  - Neal 93, Probabilistic inference using Markov chain Monte Carlo techniques
  - … (see tutorial paper)

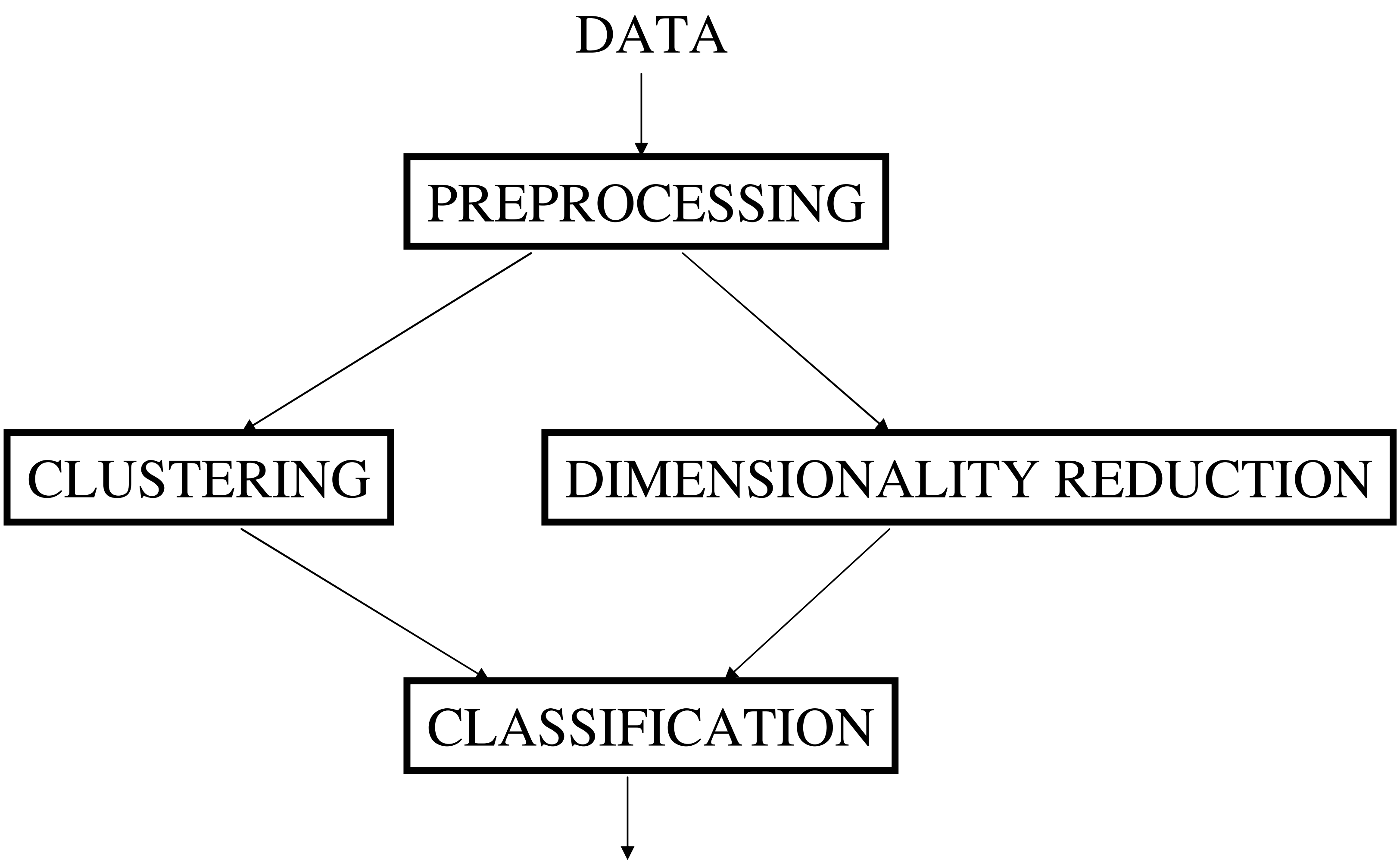
Motivation

- Computer vision
- Communications systems
- Molecular biology
- Microphone array processing
- …

Discriminative models and Generative models

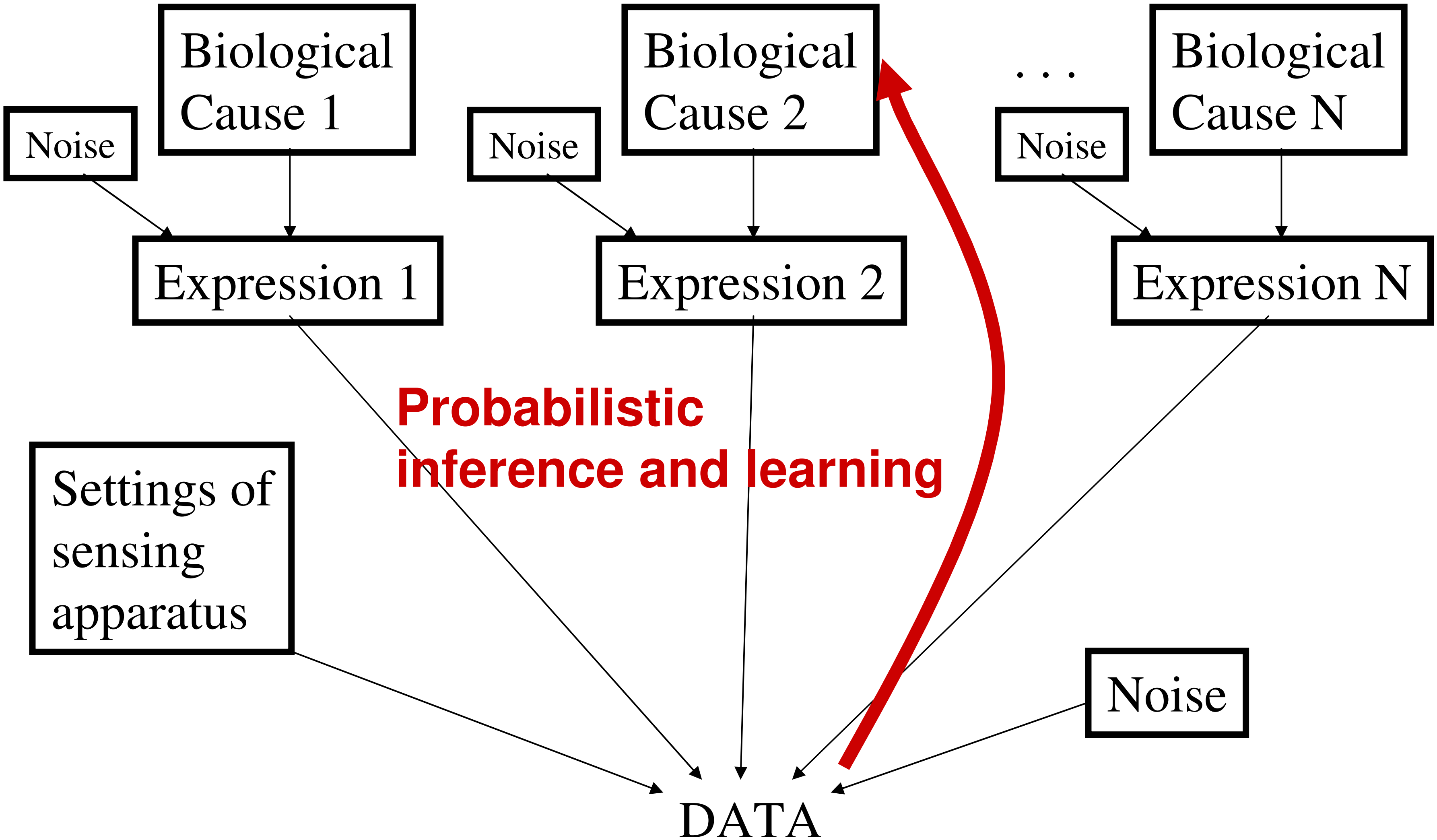
- Discriminative: P(class|input)
  - Supervised learning
- Generative: P(class,input)
  - Supervised or unsupervised learning (class unobserved)
- Hidden variables:
  - $P(\text{class}|\text{input}) = \Sigma_{\text{hidden}}\, P(\text{class},\text{hidden}|\text{input})$
  - $P(\text{class},\text{input}) = \Sigma_{\text{hidden}}\, P(\text{class},\text{hidden},\text{input})$

Sequential (block-diagram) data analysis



B. J. FREY & N. JOJIC

The generative approach (on a genomics problem)



B. J. FREY & N. JOJIC

Modularity and Graphical Models

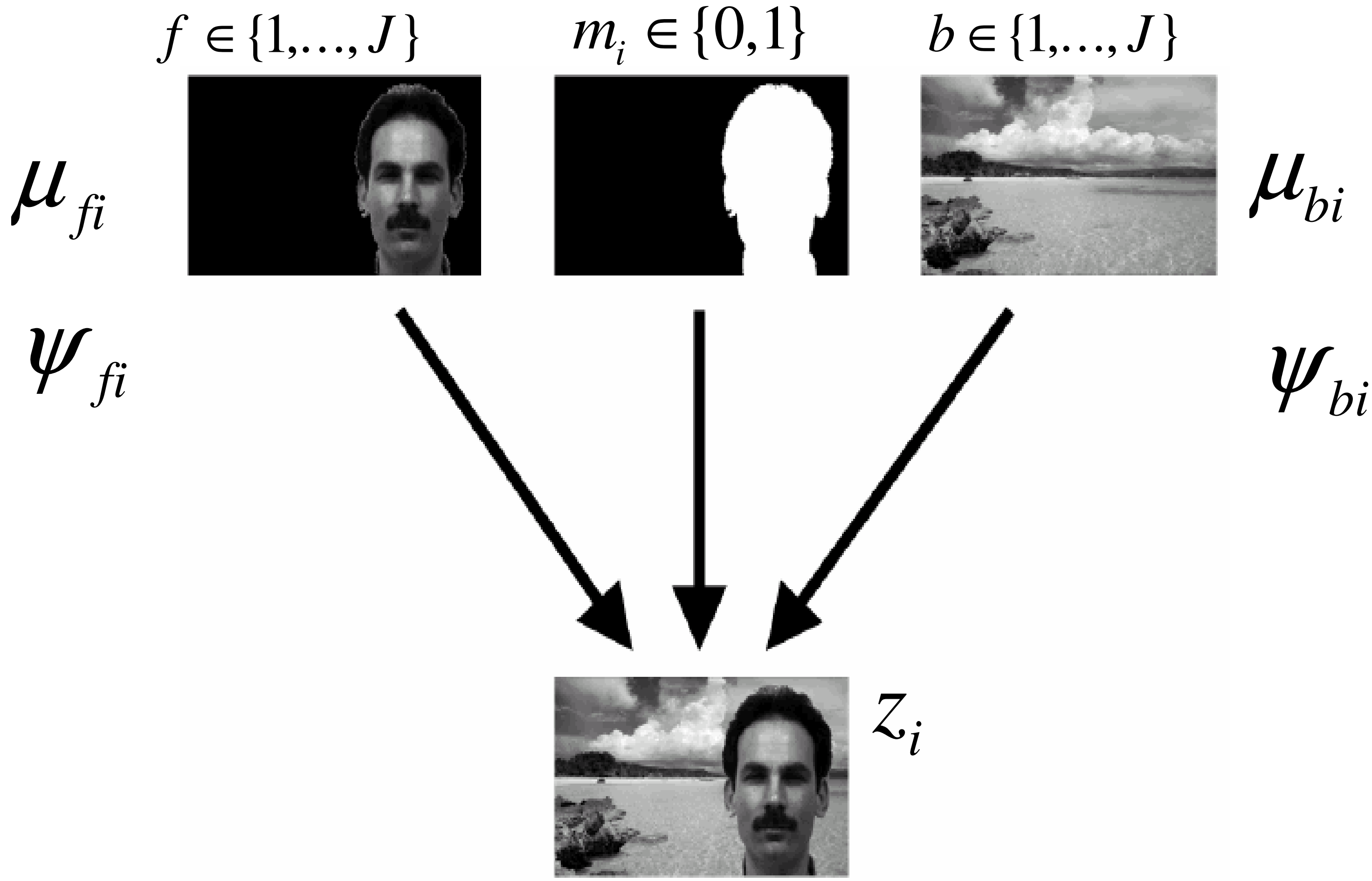
- P(class,hidden,input) is complex for real-world problems
  - How do we specify constraints on variables?
  - How do we modify sub-components of the model?
  - How do we cope with computational intractability?
  - How do we cope with learnability?
- Graphical models: Modular descriptions of complex probability models

Case study: Occlusion Model

A sample from the dataset

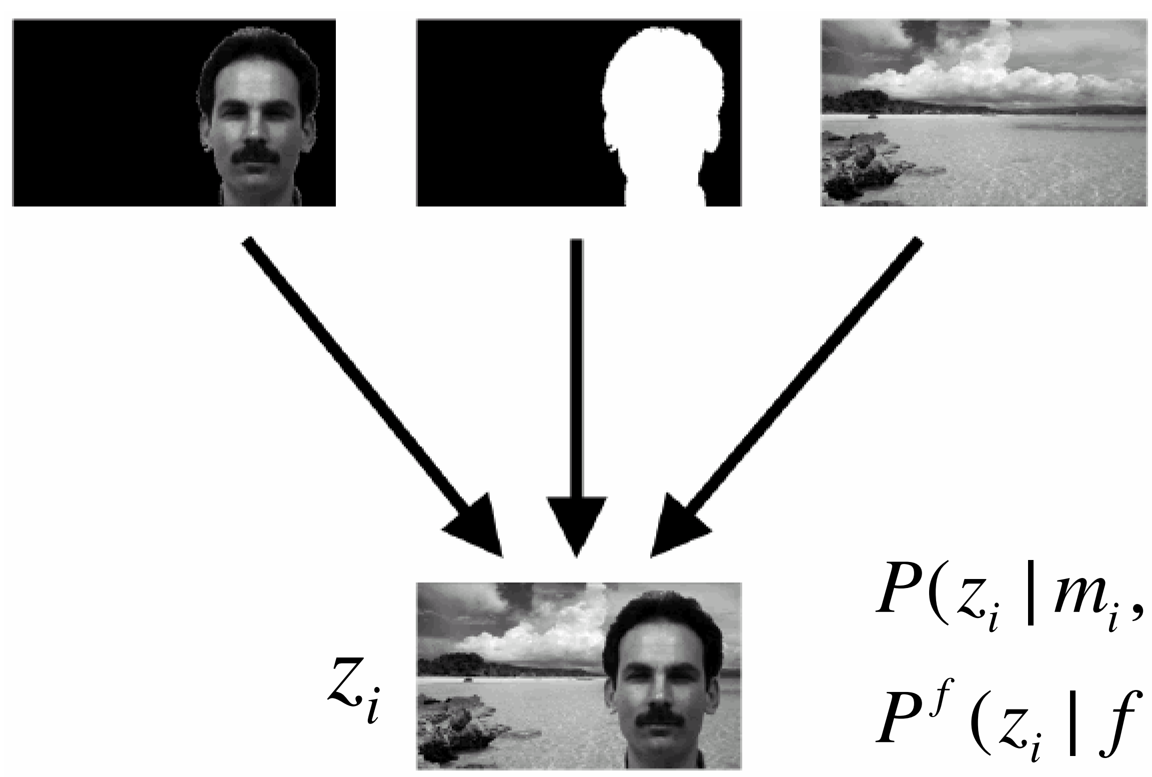
(5 different faces against 7 different backgrounds)

Generative model



Let α_i be the probability that m_i=1 given that the foreground class is f,

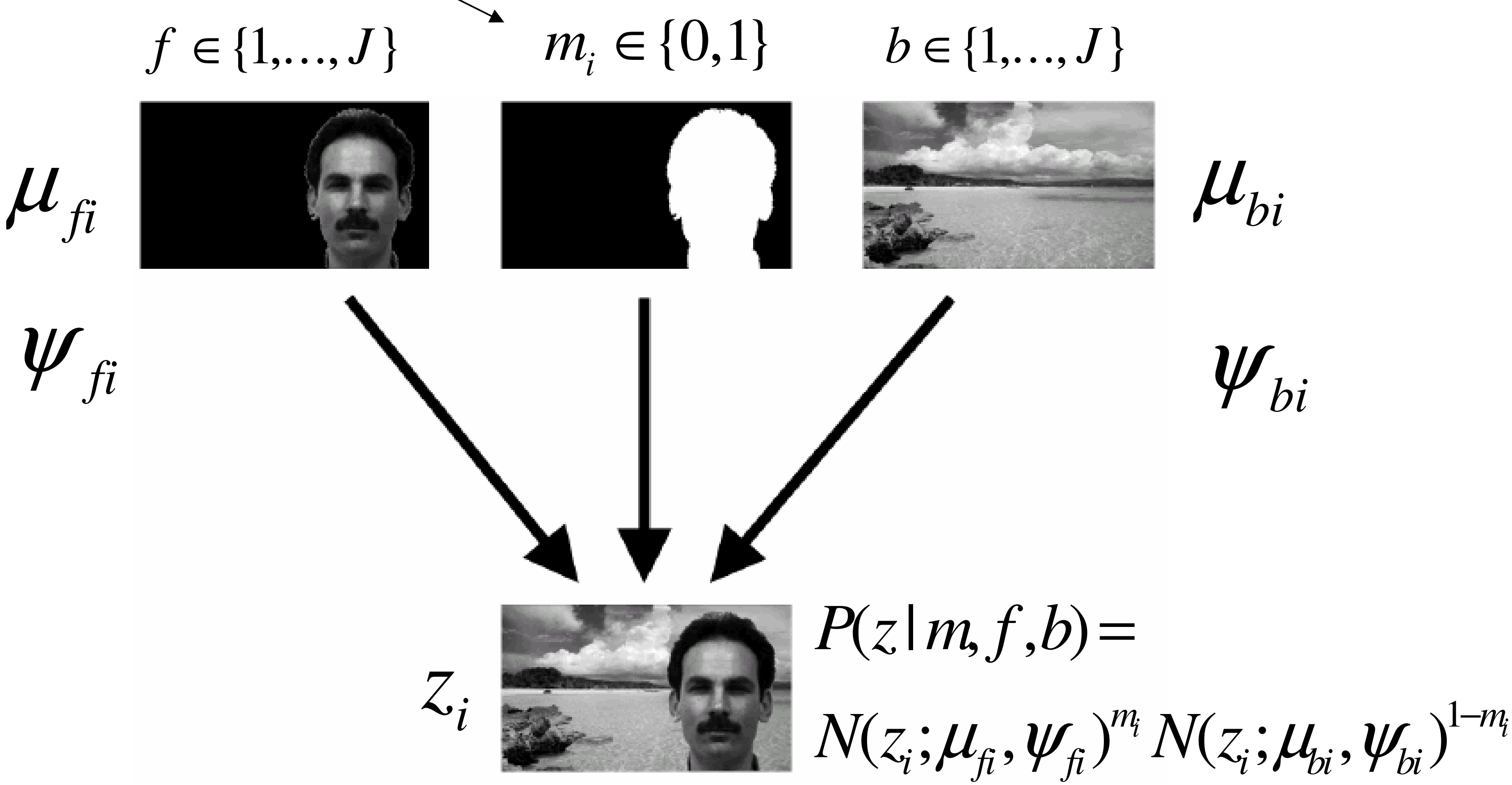
i.e., $P(m_i = 1 \mid f) = \alpha_i$, $P(m_i = 0 \mid f) = 1 - \alpha_i$,



$$P(z_i \mid m_i, f, b) =$$
$$P^f(z_i \mid f)^{m_i} P^b(z_i \mid b)^{1-m_i}$$

Let α_i be the probability that m_i=1 given that the foreground class is f,

i.e., $P(m_i = 1 \mid f) = \alpha_i$, $P(m_i = 0 \mid f) = 1 - \alpha_i$,



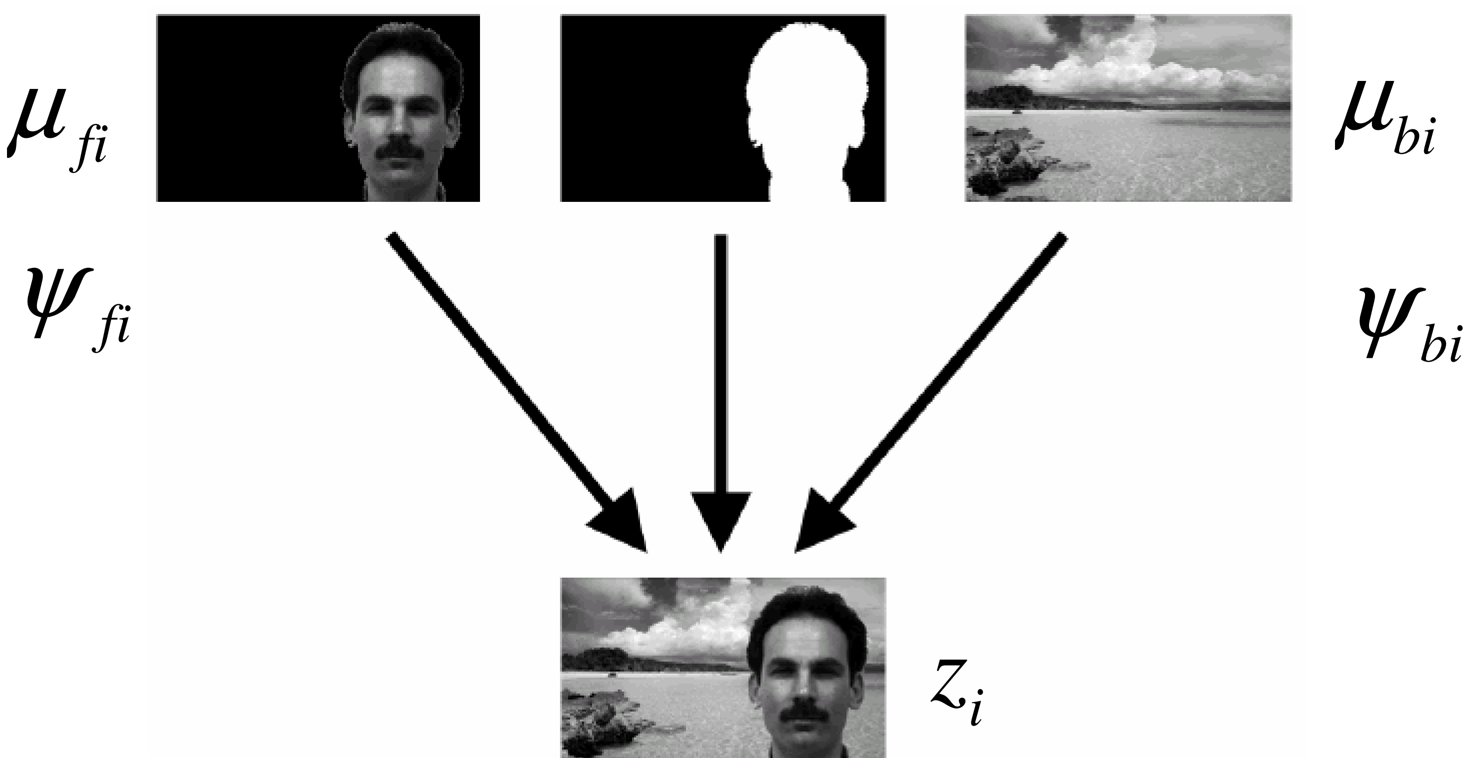
B. J. FREY & N. JOJIC

The joint probability distribution

$$P(z,m,f,b) = P(b)P(f)\left(\prod_{i=1}^{K} P(m_i \mid f)\right)\left(\prod_{i=1}^{K} P(z_i \mid m_i, f, b)\right).$$

$\alpha_i = P(m_i = 1 \mid f)$

$f \in \{1,\ldots,J\}$ $m_i \in \{0,1\}$ $b \in \{1,\ldots,J\}$



The joint probability distribution

$$P(z,m,f,b) = P(b)P(f)\left(\prod_{i=1}^{K} P(m_i \mid f)\right)\left(\prod_{i=1}^{K} P(z_i \mid m_i, f, b)\right)$$

Because m is binary, we can write:

$$P(z_i \mid m_i, f, b) = P^f(z_i \mid f)^{m_i} P^b(z_i \mid b)^{1-m_i}$$

Mask at pixel i

Observed intensity of the pixel i

$$P(z,m,f,b) = \pi_b \pi_f \left(\prod_{i=1}^{K} \alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i} N(z_i;\mu_{fi},\psi_{fi})^{m_i} N(z_i;\mu_{bi},\psi_{bi})^{1-m_i}\right).$$

Prior distribution of foreground and background classes

Probability of occlusion

Means and variances of the layer classes

B. J. FREY & N. JOJIC

Joint distribution over all variables and parameters in the dataset

- We assume uniform prior over the parameters

$$P(\mu, \psi, \pi, \alpha, f^{(1)}, b^{(1)}, m^{(1)}, \ldots, f^{(T)}, b^{(T)}, m^{(T)} | z^{(1)}, \ldots, z^{(T)})$$

$$\propto \prod_{t=1}^{T} \Big(\pi_{f^{(t)}} \pi_{b^{(t)}} \Big(\prod_{i=1}^{K} \alpha_{f^{(t)}i}^{m_i^{(t)}} (1 - \alpha_{f^{(t)}i})^{1-m_i^{(t)}} \mathcal{N}(z_i^{(t)}; \mu_{f^{(t)}i}, \psi_{f^{(t)}i})^{m_i^{(t)}} \mathcal{N}(z_i^{(t)}; \mu_{b^{(t)}i}, \psi_{b^{(t)}i})^{1-m_i^{(t)}} \Big) \Big)$$

Inference

$$P(m,f,b\,|\,z)\propto \pi_b\pi_f\,(\prod_{i=1}^{K}\alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i}N(z_i;\mu_{fi},\psi_{fi})^{m_i}N(z_i;\mu_{bi},\psi_{bi})^{1-m_i}).$$

$$P(m,f,b\,|\,z)=P(f,b\,|\,z)P(m\,|\,f,b,z)=P(f,b\,|\,z)\prod_{i=1}^{K}P(m_i\,|\,f,b,z).$$

$$P(f,b\,|\,z)\propto P(f,b,z)=\sum_{m_1}\cdots\sum_{m_K}P(m,f,b,z)$$

$$=\pi_b\pi_f\prod_{i=1}^{K}(\alpha_{fi}N(z_i;\mu_{fi},\psi_{fi})+(1-\alpha_{fi})N(z_i;\mu_{bi},\psi_{bi})).$$

B. J. FREY & N. JOJIC

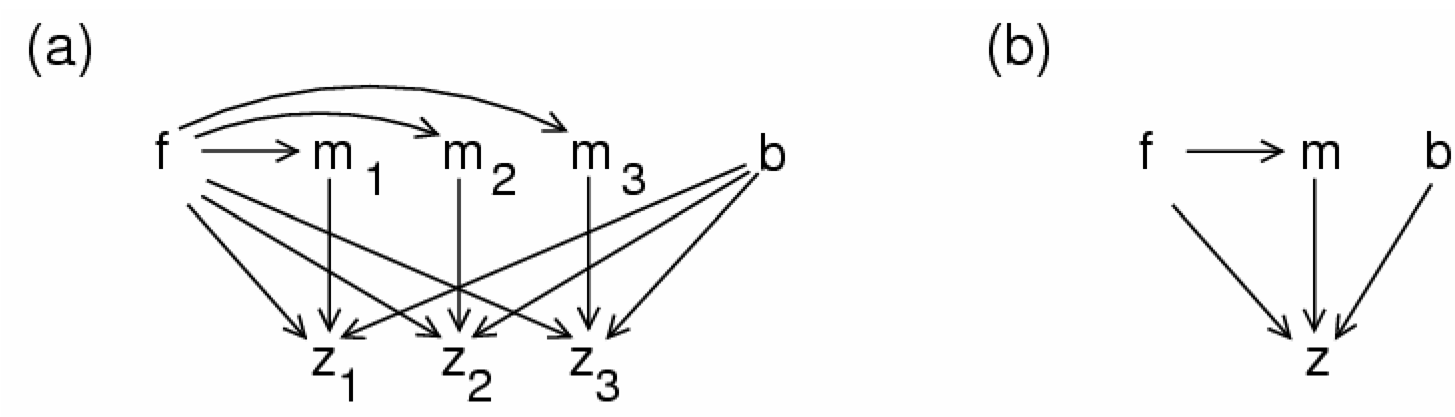
Bayesian Networks

Bayesian network

- *MAY* be constructed using knowledge of causal relationships
- Quickly conveys the factorization of a distribution
- Clearly expresses dependencies and independencies between variables
- Can be used to derive fast inference algorithms

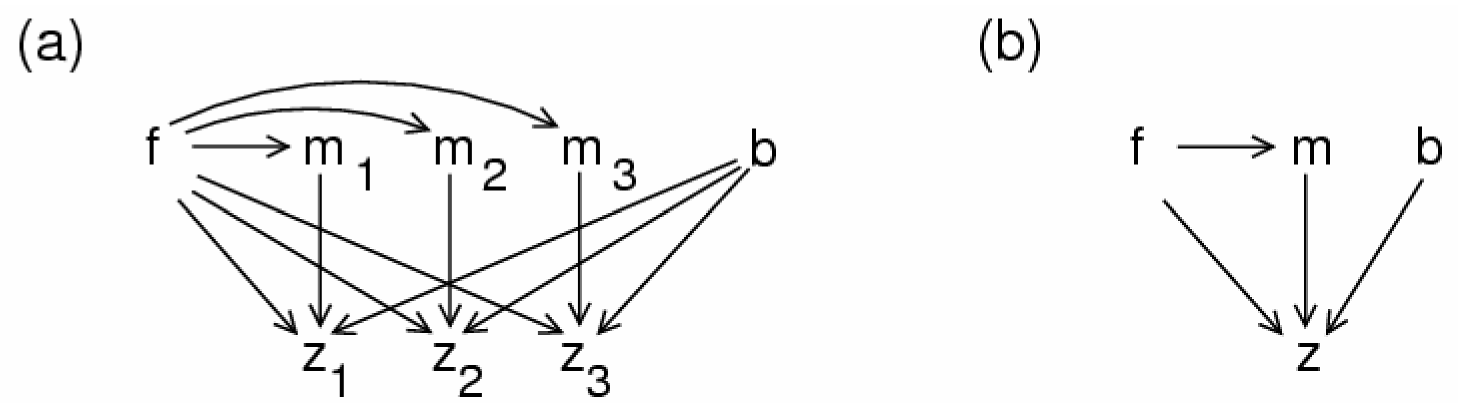
Bayes net for occlusion model

$$P(z,m,f,b) = P(b)P(f)\left(\prod_{i=1}^{K} P(m_i \mid f)\right)\left(\prod_{i=1}^{K} P(z_i \mid m_i, f, b)\right)$$

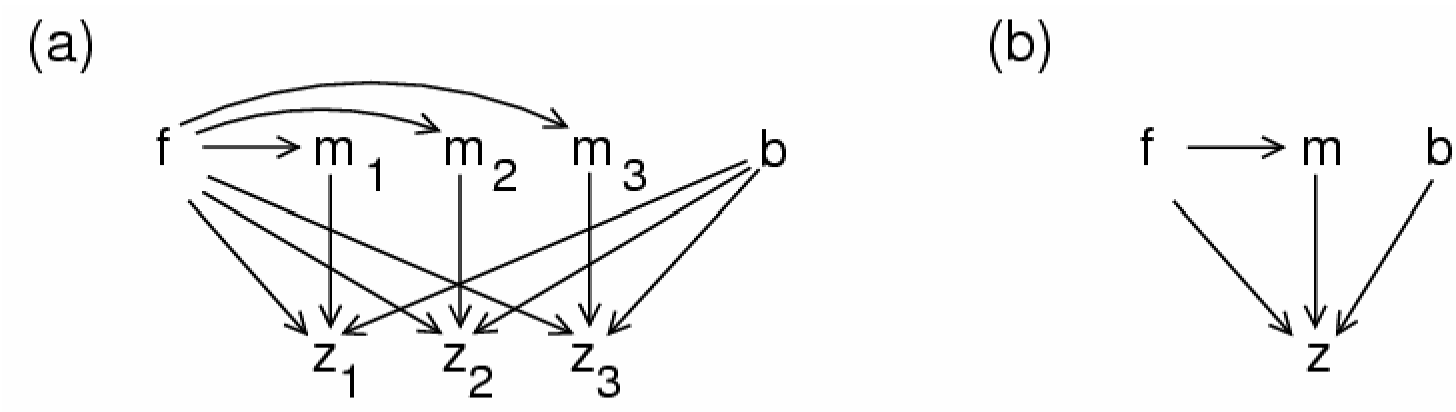


B. J. FREY & N. JOJIC

Simulating Bayes nets



Conditional independence



B. J. FREY & N. JOJIC

A greedy inference and learning algorithm: ICM

Iterative Conditional Modes (ICM)

- For $t = 1, \ldots, T$

  { $f^{(t)} \leftarrow \mathrm{argmax}_{f^{(t)}} \left[\pi_{f^{(t)}} \prod_{i:m_i^{(t)}=1} \mathcal{N}(z_i^{(t)}; \mu_{f^{(t)}i}, \psi_{f^{(t)}i}) \right]$

  { $b^{(t)} \leftarrow \mathrm{argmax}_{b^{(t)}} \left[\pi_{b^{(t)}} \prod_{i:m_i^{(t)}=0} \mathcal{N}(z_i^{(t)}; \mu_{b^{(t)}i}, \psi_{b^{(t)}i}) \right]$

  { For $i = 1, \ldots, K$: $\quad m_i^{(t)} \leftarrow \begin{cases} 1 & \text{if } \alpha_{f^{(t)}i} \mathcal{N}(z_i^{(t)}; \mu_{f^{(t)}i}, \psi_{f^{(t)}i}) > (1 - \alpha_{f^{(t)}i}) \mathcal{N}(z_i^{(t)}; \mu_{b^{(t)}i}, \psi_{b^{(t)}i}) \\ 0 & \text{otherwise} \end{cases}$

- For $j = 1, \ldots, J$

  { $\pi_j \leftarrow (\sum_{t=1}^T [f^{(t)} = j] + \sum_{t=1}^T [b^{(t)} = j])/2T$

- For $j = 1, \ldots, J$, for $i = 1, \ldots, K$

  { $\alpha_{ji} \leftarrow (\sum_{t=1}^T [f^{(t)} = j] m_i^{(t)})/(\sum_{t=1}^T [f^{(t)} = j])$

  { $\mu_{ji} \leftarrow (\sum_{t=1}^T [f^{(t)} = j \text{ or } b^{(t)} = j] z_i^{(t)})/(\sum_{t=1}^T [f^{(t)} = j \text{ or } b^{(t)} = j])$

  { $\psi_{ji} \leftarrow (\sum_{t=1}^T [f^{(t)} = j \text{ or } b^{(t)} = j] (z_i^{(t)} - \mu_{ji})^2)/(\sum_{t=1}^T [f^{(t)} = j \text{ or } b^{(t)} = j])$

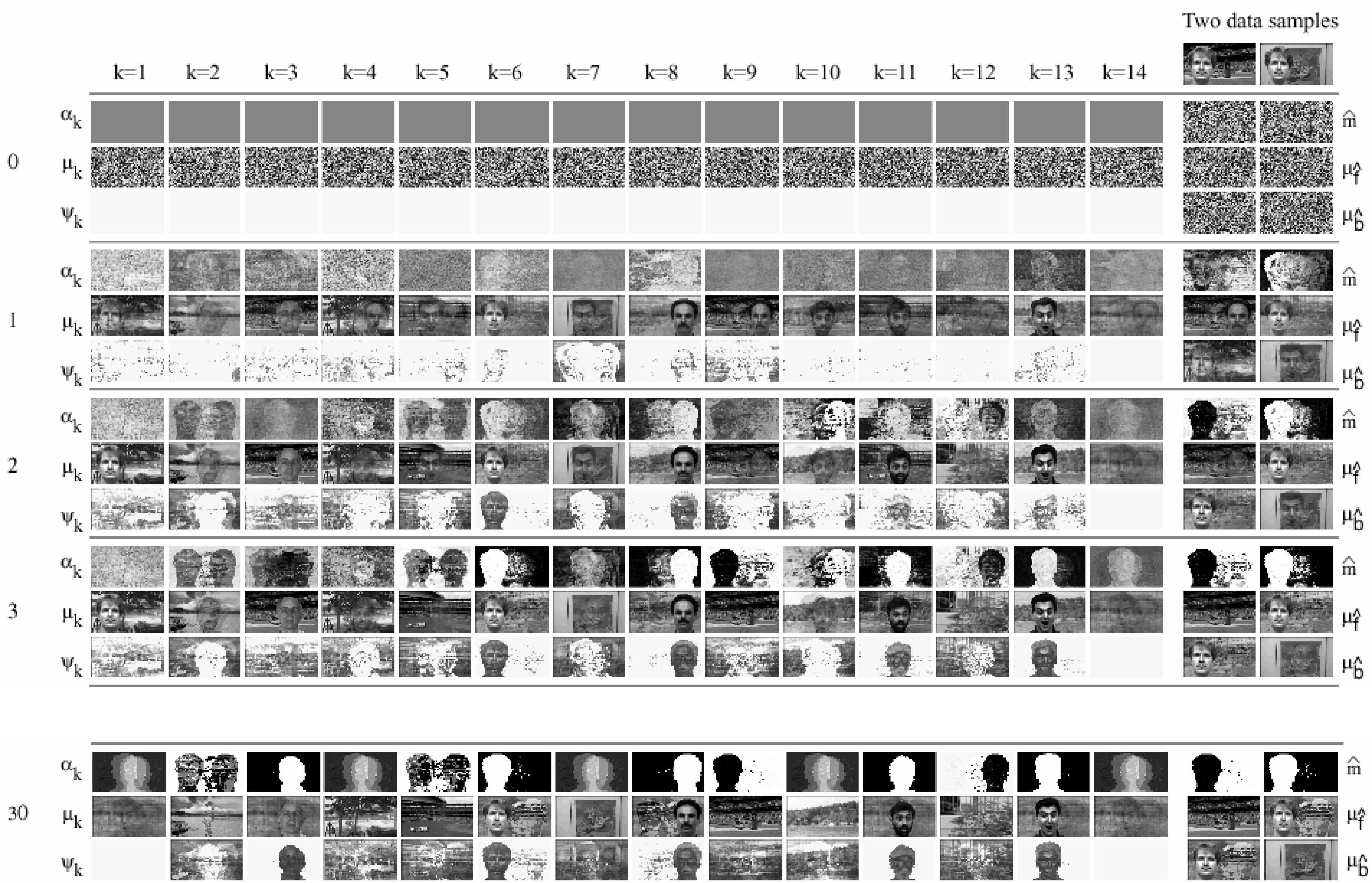
Here, the Iverson notation is used where [True] $= 1$ and [False] $= 0$.

B. J. FREY & N. JOJIC

A sample from the dataset

(5 different faces against 7 different backgrounds)

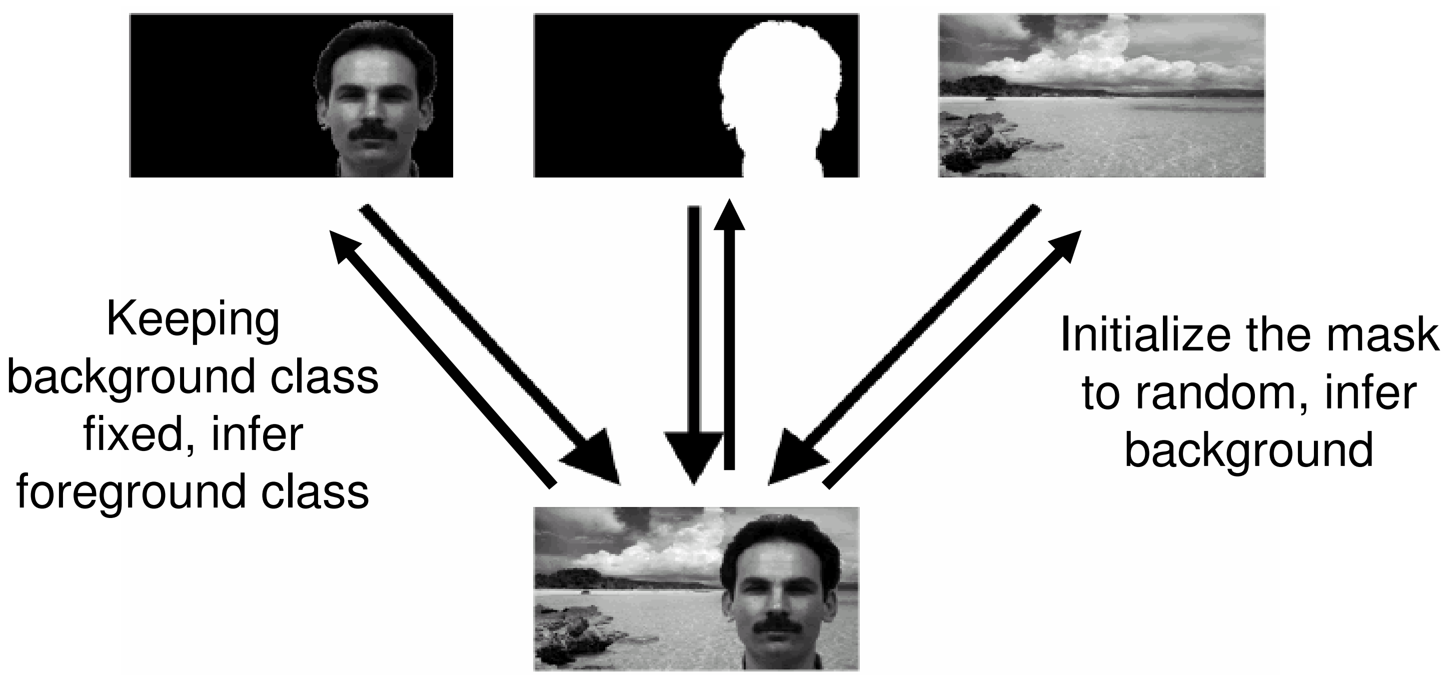
Iterative model optimization



B. J. FREY & N. JOJIC

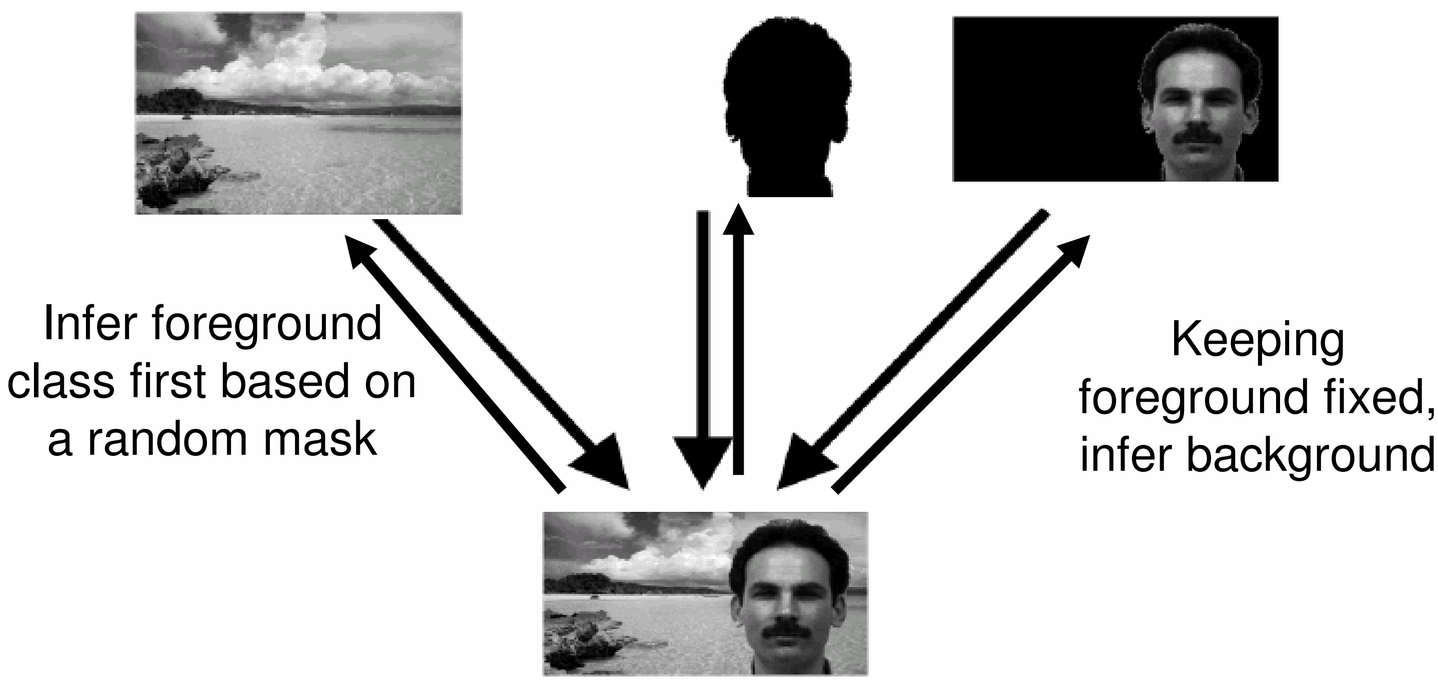
Inferring hidden variables using ICM

- Given the class means and variances and mask probabilities for each class, for the shown dataset, it is (sometimes) possible to invert the generative process



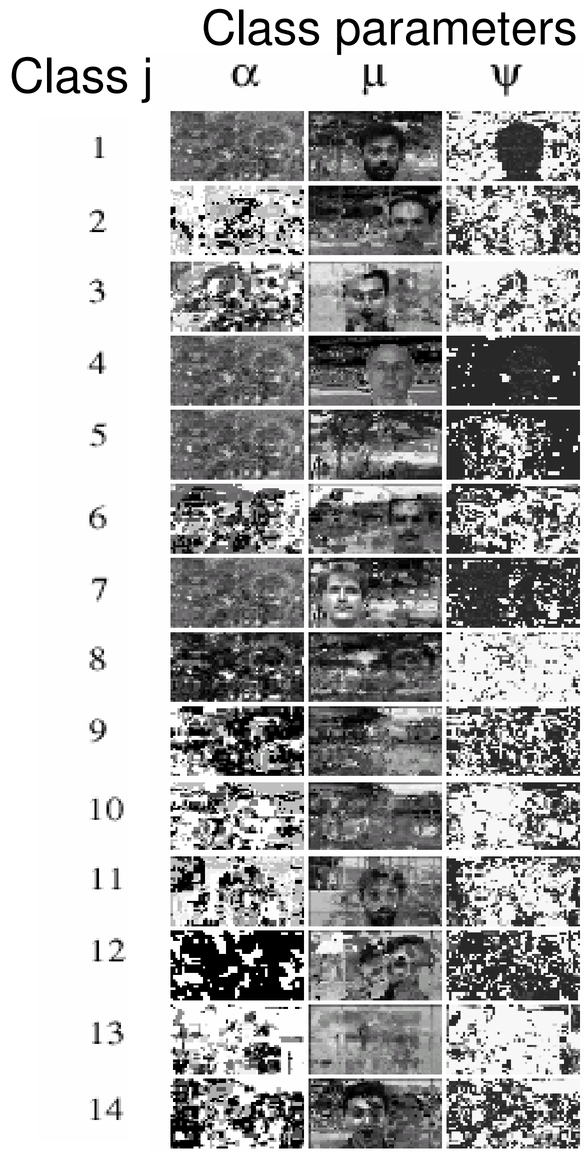
However…

- Order of updates is important



B. J. FREY & N. JOJIC

Inferring the hiddens *and* estimating the parameters of a 14-class model using ICM



B. J. FREY & N. JOJIC

ICM

- Simple, fast
- Prone to local maxima
- Can be seen as iteratively employing simplified conditional posteriors, e.g.,

$$P(m,f,b\,|\,z) \propto \pi_b \pi_f \prod_{i=1}^{K} \alpha_{fi}^{m_i} (1-\alpha_{fi})^{1-m_i} N(z_i;\mu_{fi},\psi_{fi})^{m_i} N(z_i;\mu_{bi},\psi_{bi})^{1-m_i}$$

$$P(b\,|\,\hat{f},\hat{m},z) \propto \pi_b \pi_{\hat{f}} \prod_{i=1}^{K} \alpha_{\hat{f}i}^{\hat{m}_i} (1-\alpha_{\hat{f}i})^{1-\hat{m}_i} N(z_i;\mu_{\hat{f}i},\psi_{\hat{f}i})^{\hat{m}_i} N(z_i;\mu_{bi},\psi_{bi})^{1-\hat{m}_i}$$

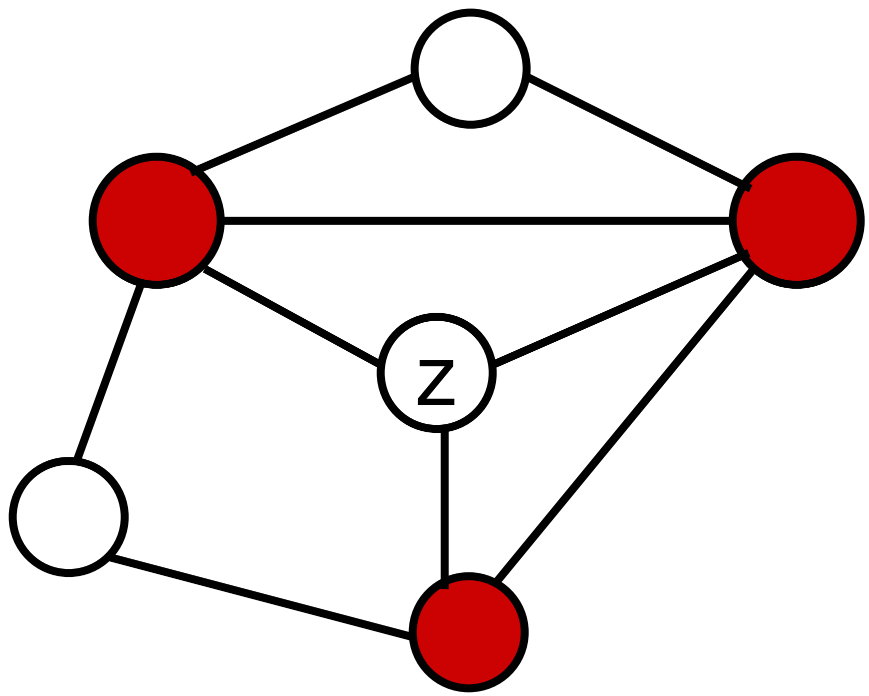
$$\hat{b} = \arg\max P(b\,|\,\hat{f},\hat{m},z)$$

- Related to variational methods that use less severe approximations of the posterior

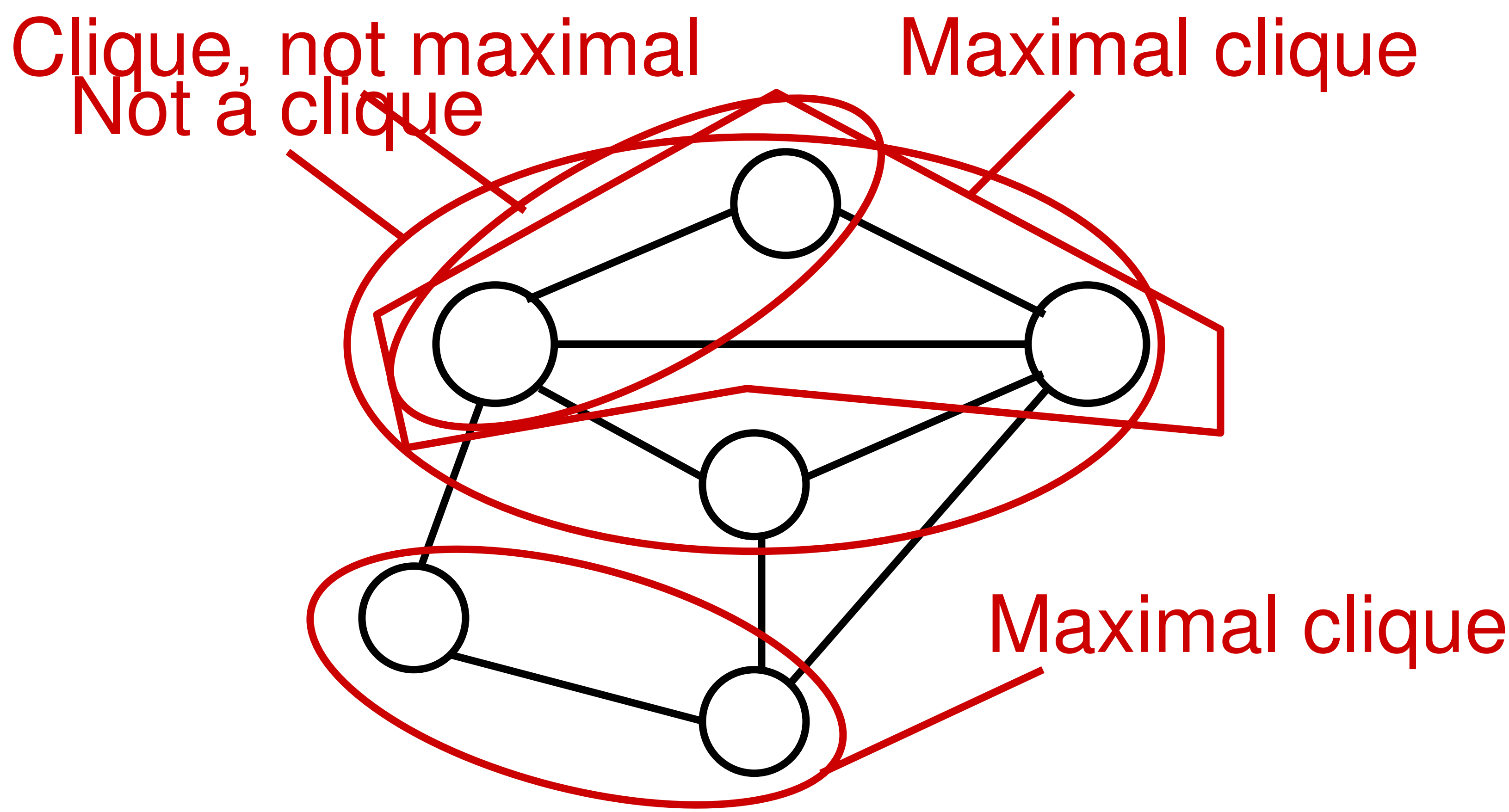
Markov Random Fields (MRFs)

Markov random fields (MRFs)

- Undirected graph on variables
- Each variable is independent of all other variables, given its neighbors



Cliques and maximal cliques

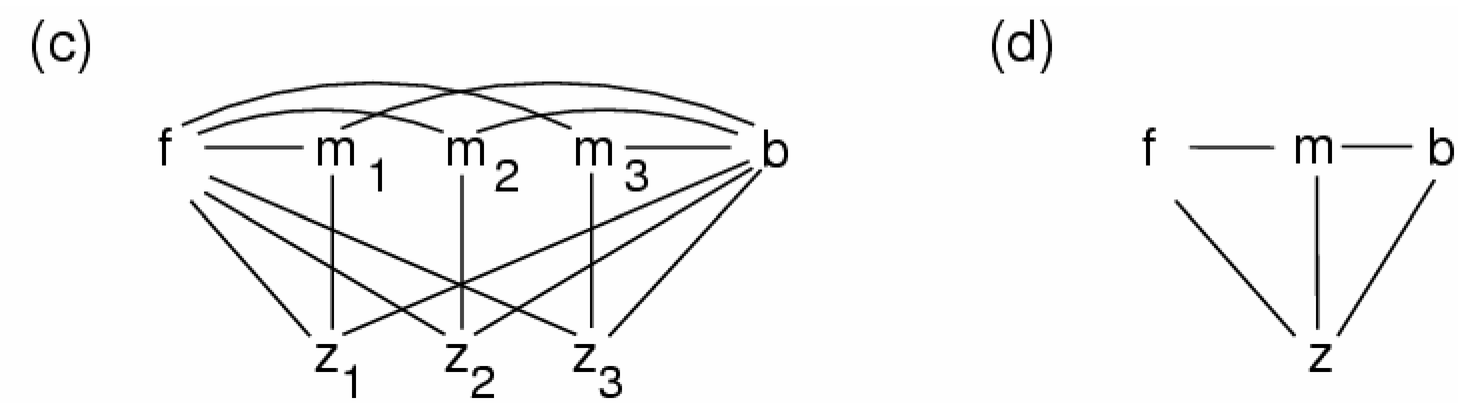


The distribution for an MRF

$$P(x_1, \ldots, x_N) = \alpha \prod_i \Psi_i(x_{C_i})$$

- i = index of maximal clique
  - C_i = set of indices of variables
  - x_{C_i} = set of variables
- $\Psi_i(x_{C_i})$ is a “potential” (“local function”)
- α is a normalizing constant

Occlusion model



$$P(z,m,f,b) = P(b)P(f)\left(\prod_{i=1}^{K} P(m_i \mid f)\right)\left(\prod_{i=1}^{K} P(z_i \mid m_i, f, b)\right).$$

B. J. FREY & N. JOJIC

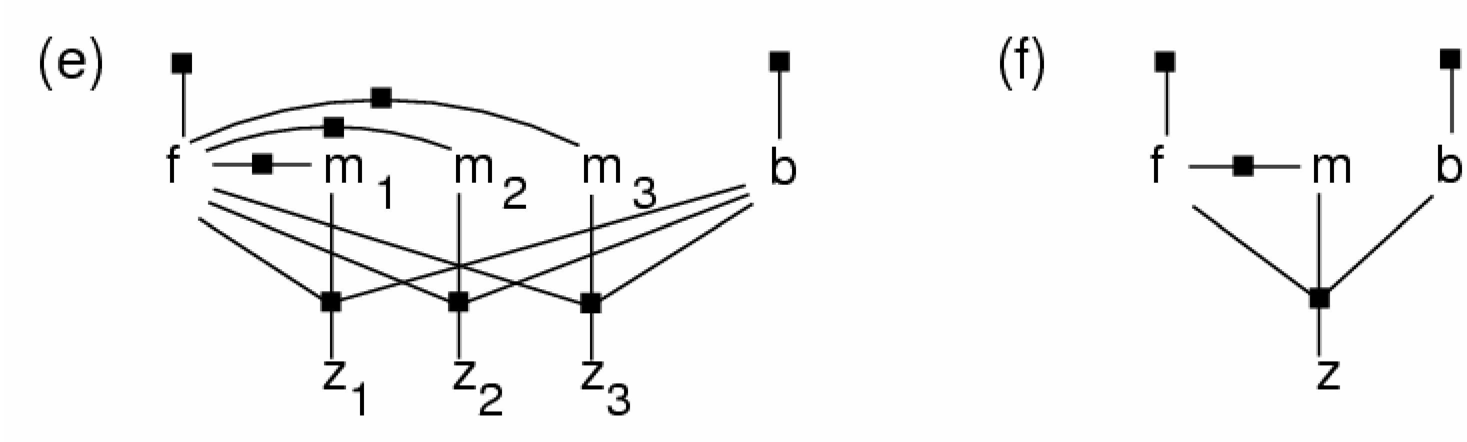
Factor Graphs

Factor graphs

(Kschischang, Frey and Loeliger, 1999)

- Bipartite graph: nodes for variables and functions
- A *potential* (*local function*) is associated with each function node – this function depends on the neighboring variables
- The joint distribution is given by the product of the local functions
- Edges can be directed or undirected (a directed edge indicates a conditional probability)

Occlusion model



$$P(z,m,f,b) = P(b)P(f)\left(\prod_{i=1}^{K} P(m_i \mid f)\right)\left(\prod_{i=1}^{K} P(z_i \mid m_i, f, b)\right).$$

B. J. FREY & N. JOJIC

Parameterized Models and Bayesian Learning

Parameters as variables

- Recall for our toy problem:

$$P(z,m,f,b)=\pi_b\pi_f\left(\prod_{i=1}^{K}\alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i}N(z_i;\mu_{fi},\psi_{fi})^{m_i}N(z_i;\mu_{bi},\psi_{bi})^{1-m_i}\right).$$

- Interpret the parameters as variables, where the joint distribution is

$$P(z,m,f,b,\pi,\alpha,\mu,\psi)=P(b\,|\,\pi)P(f\,|\,\pi)P(m\,|\,f,\alpha)P(z\,|\,m,f,b,\mu,\psi)P(\pi)P(\alpha)P(\mu)P(\psi).$$

where

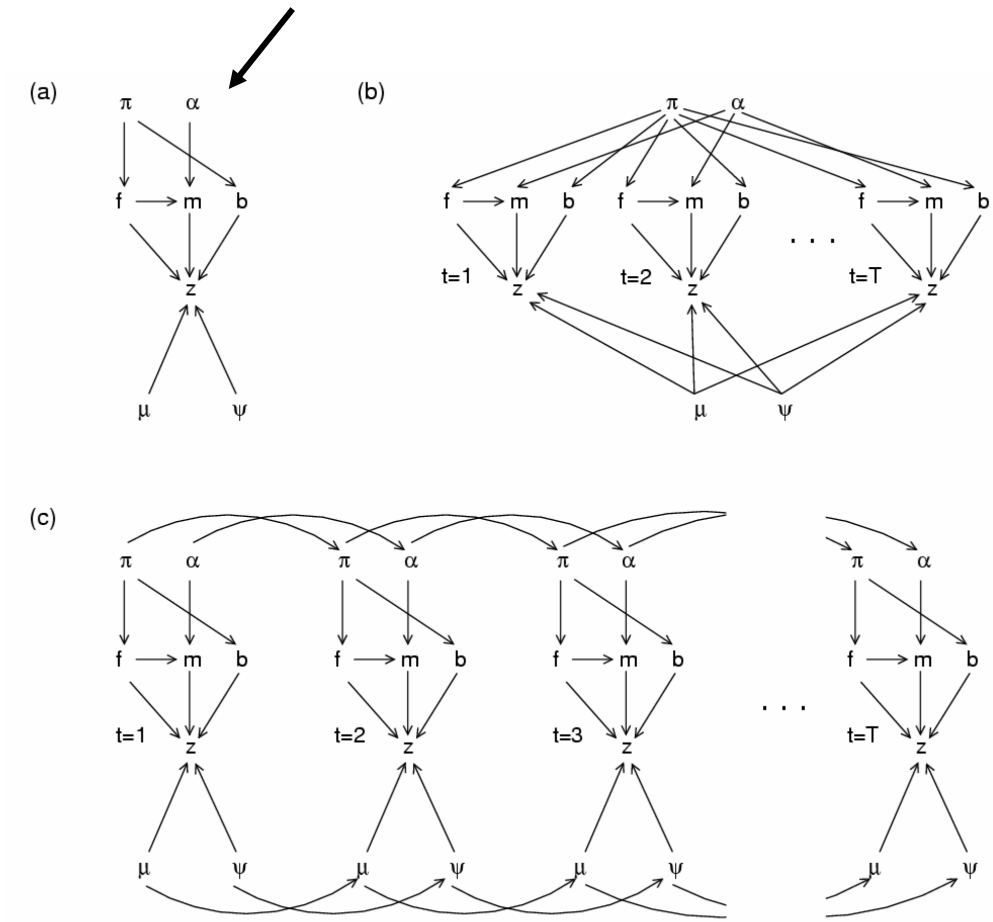
$$P(b\,|\,\pi)=\pi_b \quad P(f\,|\,\pi)=\pi_f \quad P(m_i\,|\,f,\alpha_{1i},\ldots,\alpha_{Ji})=\alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i}$$

$$P^f(z_i\,|\,f,\mu_{1i},\ldots,\mu_{Ji},\psi_{1i},\ldots,\psi_{Ji})=N(z_i;\mu_{fi},\psi_{fi})$$

$$P^b(z_i\,|\,b,\mu_{1i},\ldots,\mu_{Ji},\psi_{1i},\ldots,\psi_{Ji})=N(z_i;\mu_{bi},\psi_{bi})$$

Bayes nets for Bayesian version of toy problem

$$P(z,m,f,b,\pi,\alpha,\mu,\psi) = P(b\,|\,\pi)P(f\,|\,\pi)P(m\,|\,f,\alpha)P(z\,|\,m,f,b,\mu,\psi)P(\pi)P(\alpha)P(\mu)P(\psi).$$



B. J. FREY & N. JOJIC

Introducing training data

- Visible variables: $v = (v^{(1)},\ldots,v^{(T)})$
  - $^{(t)}$ denotes the tth training case
- Hidden variables: $h = (h^{\theta}, h^{(1)},\ldots,h^{(T)})$
  - Parameters: h^{θ}
  - Variables for the tth training case: $h^{(t)}$
- Joint distribution: $P(h,v) = P(h^{\theta})\prod_{t=1}^{T} P(h^{(t)}, v^{(t)} \mid h^{\theta}).$
- Parameter prior: $P(h^{\theta})$
- Parameter likelihood: $\prod_{t=1}^{T} P(h^{(t)}, v^{(t)} \mid h^{\theta})$

Parameter priors

- Uniform priors
  - Computationally attractive
  - Problem: “Uniform” is inconsistent w.r.t. reparameterization
- Conjugate priors
  - Prior x Likelihood has same form as likelihood
  - Conjugate prior can be thought of as likelihood for “fake” data, eg, prior counts in a coin toss experiment

Algorithms for Inference and Learning

B. J. FREY & N. JOJIC

Problem set-up

- Probabilistic inference and learning entail computing the intractable distribution,

$$P(h \mid v) = \frac{P(h,v)}{\int_h P(h,v)},$$

- Note that w.r.t. h, $P(h \mid v) \propto P(h,v)$.

- In a graphical model, $P(h,v)$ factorizes

General “brute force” inference

- Suppose $x_1, x_2, \ldots, x_N$ are binary

$$P(x_1) = \Sigma_{x_2}\Sigma_{x_3} \ldots \Sigma_{x_N} P(x_1, x_2, \ldots, x_N)$$

- This takes about 2^N operations

- Generally, computing $P(x_i|\text{Observed x's})$ takes $2^{(N\text{ - \#observed x's})}$ operations

Exact posterior for occlusion model

$$P(\mu, \psi, \pi, \alpha, f^{(1)}, b^{(1)}, m^{(1)}, \ldots, f^{(T)}, b^{(T)}, m^{(T)} | z^{(1)}, \ldots, z^{(T)})$$
$$\propto \prod_{t=1}^{T} \Big(\pi_{f^{(t)}} \pi_{b^{(t)}} \Big(\prod_{i=1}^{K} \alpha_{f^{(t)}i}^{m_i^{(t)}} (1 - \alpha_{f^{(t)}i})^{1-m_i^{(t)}} \mathcal{N}(z_i^{(t)}; \mu_{f^{(t)}i}, \psi_{f^{(t)}i})^{m_i^{(t)}} \mathcal{N}(z_i^{(t)}; \mu_{b^{(t)}i}, \psi_{b^{(t)}i})^{1-m_i^{(t)}} \Big) \Big)$$

Approximate inference

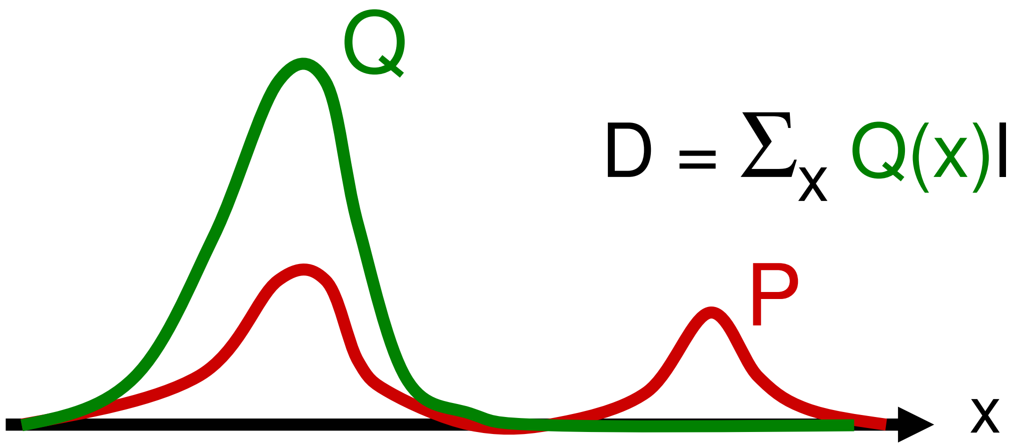
- Since P(h|v) is intractable, we search for a surrogate, Q(h) that is tractable
- Measure of quality of Q: Kullback-Leibler divergence between Q and P,

$$D(Q,P) = \int_h Q(h) \log \frac{Q(h)}{P(h \mid v)}.$$

- Properties:
  - $D \geq 0$
  - $D = 0$ iff $Q(h) = P(h|v)$

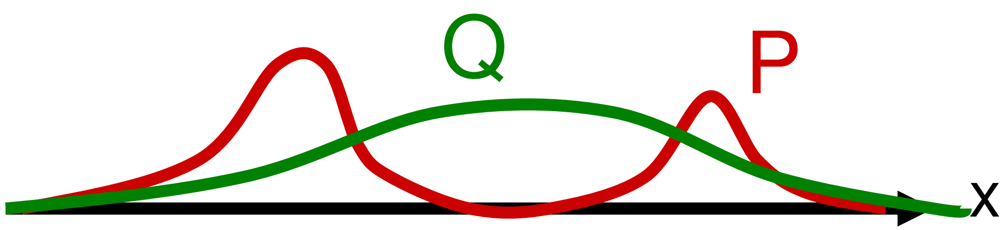
More on the “distance”

$$D = \Sigma_x Q(x)\log[Q(x)/P(x)]$$



For unimodal Q, minimizing D favors this Q

$$D = \Sigma_x P(x) \log[P(x)/Q(x)]$$



For unimodal Q, minimizing D favors this Q

Free energy

- In fact, D is intractable
- We can make it tractable by subtracting log P(v)

$$F(Q,P) = D(Q,P) - \log P(v)$$

$$= \int_h Q(h) \log \frac{Q(h)}{P(h|v)} - \int_h Q(h) \log P(v)$$

$$= \int_h Q(h) \log \frac{Q(h)}{P(h|v)P(v)}$$

$$= \int_h Q(h) \log \frac{Q(h)}{P(h,v)}.$$

Factorizes, for a graphical model

B. J. FREY & N. JOJIC

Alternative derivation of free energy

- log P(v) is the log-likelihood of the data
- Computing log P(v) is intractable, but we can bound it using Jensen's inequality:

$$\log P(v) = \log(\int_h P(h,v))$$

$$= \log(\int_h Q(h)\frac{P(h,v)}{Q(h)})$$

$$\geq \int_h Q(h)\log(\frac{P(h,v)}{Q(h)}) = -F(Q,P).$$

Properties of free energy

- $F \geq - \log P(v)$
- The minimum of F w.r.t Q gives

  $F = - \log P(v)$

  $Q(h) = P(h|v)$

Approaches to constructing Q

- Parameterize Q directly
  - Iterative conditional modes
  - Variational methods (mean field)
  - Structured variational methods
  - Hybrids (the EM algorithm)
- Parameterize marginals of Q
  - Bethe approximation (probability propagation)
  - Kikuchi approximation (generalized prob prop)
- Represent Q using samples
  - Monte Carlo
  - Markov chain Monte Carlo

I.I.D. Training cases

Free energy for i.i.d. training cases

From (5), for a training set of T i.i.d. training cases with hidden variables $h = (h^\theta, h^{(1)}, \ldots, h^{(T)})$ and visible variables $v = (v^{(1)}, \ldots, v^{(T)})$, we have $P(h, v) = P(h^\theta) \prod_{t=1}^{T} P(h^{(t)}, v^{(t)} | h^\theta)$. The free energy is

$$F(Q, P) = \int_h Q(h) \log Q(h) - \int_h Q(h) \log P(h, v)$$
$$= \int_h Q(h) \log Q(h) - \int_{h^\theta} Q(h^\theta) \log P(h^\theta) - \sum_{t=1}^{T} \int_{h^{(t)}, h^\theta} Q(h^{(t)}, h^\theta) \log P(h^{(t)}, v^{(t)} | h^\theta). \quad (8)$$

The decomposition of F into a sum of one term for each training case simplifies learning.

B. J. FREY & N. JOJIC

Point inference

- Discrete hidden variables: $\hat{h}$ is a point estimate

$$F(Q,P) = \sum_h [h=\hat{h}] \log \frac{[h=\hat{h}]}{P(h,v)} = -\log P(\hat{h},v).$$

  – Minimizing F w.r.t. $\hat{h}$ Maximizes $P(\hat{h},v)$ w.r.t. h

- Continuous hidden variables:

$$F(Q,P) = \int_h \delta(h-\hat{h}) \log \frac{\delta(h-\hat{h})}{P(h,v)} = -\log P(\hat{h},v) - H_\delta,$$

  – Very common in engineering and science
  – Problem: $H_\delta \to -\infty$

Iterative Conditional Modes (ICM)

Initialization. Pick values for all hidden variables h (randomly, or cleverly).

ICM Step. Select one of the hidden variables, h_i. Holding all other variables constant, set h_i to its MAP value:

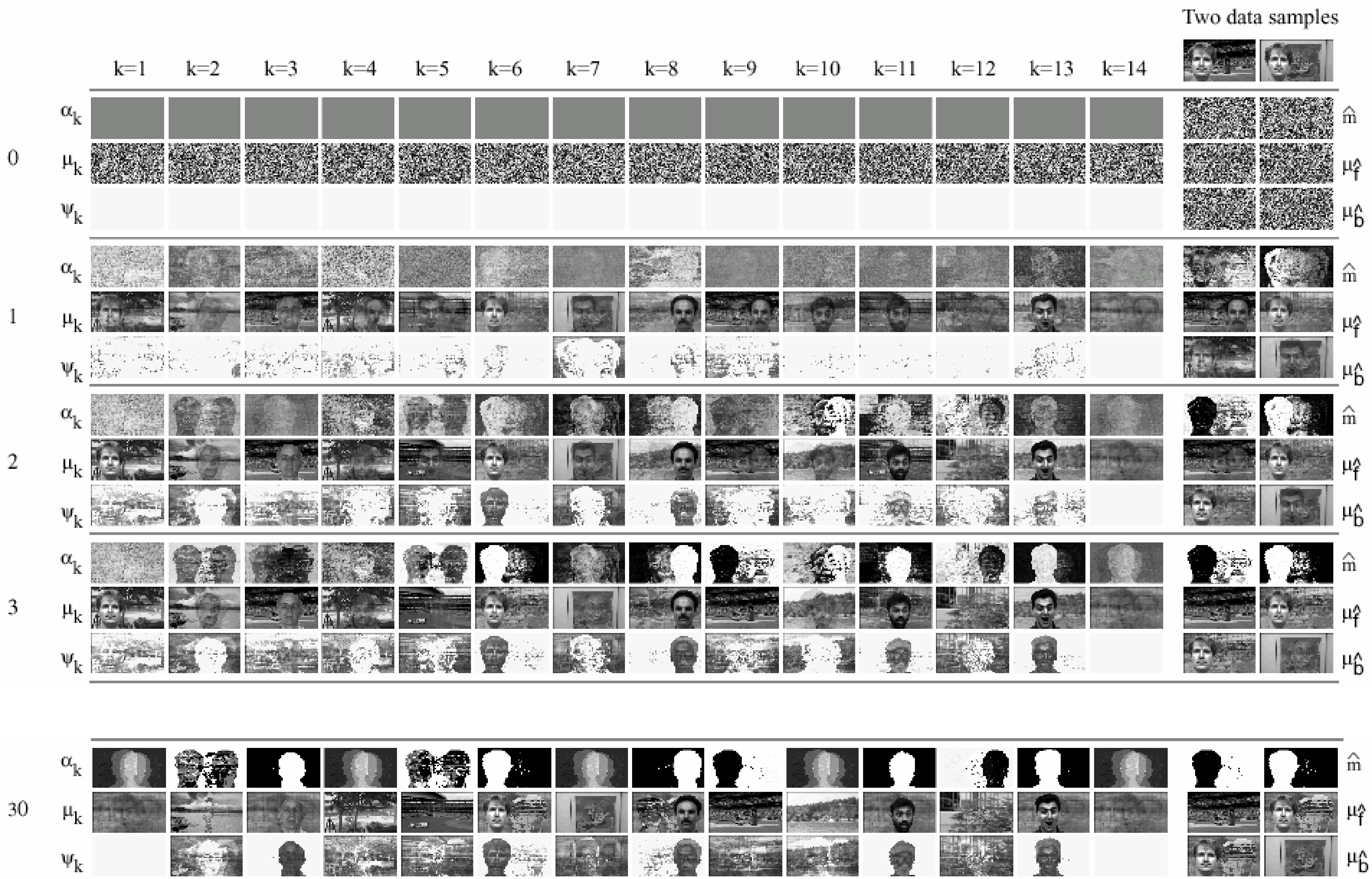
$$h_i \leftarrow \text{argmax}_{h_i} P(h_i | h \setminus h_i, v) = \text{argmax}_{h_i} P(h, v),$$

where $h \setminus h_i$ is the set of all hidden variables other than h_i.

Repeat for a fixed number of iterations or until convergence.

- Examples
  - ICM in mixture of Gaussians = k-means clustering
  - Clever ICM in M of G = sequential k-means clustering

Recall ICM for occlusion model



B. J. FREY & N. JOJIC

Properties of ICM

- Fast, greedy
- Can get stuck in local minima
- Does not account for mass of mode
- Does not account for multiple modes
- Speed depends on choice of substructures
- Usually easy to implement

The EM Algorithm

- Use the following Q-distribution:

$$Q(h) = \delta(h^\theta - \hat{h}^\theta) Q(h^{(1)}, \ldots, h^{(T)}).$$

- IID data: The training cases are independent, so

$$Q(h) = \delta(h^\theta - \hat{h}^\theta) \prod_{t=1}^{T} Q(h^{(t)}).$$

- Free energy:

$$F(Q,P) = -\log P(\hat{h}^\theta) + \sum_{t=1}^{T} \left(\int_{h^{(t)}} Q(h^{(t)}) \log \frac{Q(h^{(t)})}{P(h^{(t)}, v^{(t)} \mid \hat{h}^\theta)} \right).$$

The EM Algorithm

Initialization. Choose values for the parameters, $\hat{h}^\theta$ (randomly, or cleverly).

E Step. Minimize $F(Q, P)$ w.r.t. Q by setting

$$Q(h^{(t)}) \leftarrow P(h^{(t)}|v^{(t)}, \hat{h}^\theta),$$

for each training case, given the parameters $\hat{h}^\theta$ and the data $v^{(t)}$.

M Step. Minimize $F(Q, P)$ w.r.t. the model parameters $\hat{h}^\theta$ by solving

$$-\frac{\partial}{\partial \hat{h}^\theta} \log P(\hat{h}^\theta) - \sum_{t=1}^{T} \Big(\int_{h^{(t)}} Q(h^{(t)}) \frac{\partial}{\partial \hat{h}^\theta} \log P(h^{(t)}, v^{(t)}|\hat{h}^\theta) \Big) = 0. \qquad (9)$$

For M parameters, this is a system of M equations. Often, the prior on the parameters is assumed to be uniform, $P(\hat{h}^\theta) = const$, in which case the first term in the above expression vanishes.

Repeat for a fixed number of iterations or until convergence.

B. J. FREY & N. JOJIC

Occlusion model:
EM as an algorithm for minimizing the free energy

$$F(Q,P) = \int_h Q(h)\log Q(h) - \int_h Q(h)\log P(h,v)$$

$$= \int_h Q(h)\log Q(h) - \int_{h^\theta} Q(h^\theta)\log P(h^\theta) -$$

$$-\sum_{t=1}^{T}\int_{h^{(t)},h^\theta} Q(h^{(t)},h^\theta)\log P(h^{(t)},v^{(t)}\mid h^\theta).$$

$$P \propto \prod_{t=1}^{T}(\pi_{f^{(t)}}\pi_{b^{(t)}}(\prod_{i=1}^{K}\alpha_{f^{(t)}i}^{m_i^{(t)}}(1-\alpha_{f^{(t)}i})^{1-m_i^{(t)}} N(z_i^{(t)};\mu_{f^{(t)}i},\psi_{f^{(t)}i})^{m_i^{(t)}} N(z_i^{(t)};\mu_{b^{(t)}i},\psi_{b^{(t)}i})^{1-m_i^{(t)}})).$$

$$Q = (\prod_k \delta(\pi_k - \hat{\pi}_k))(\prod_{k,i}\delta(\mu_{ki}-\hat{\mu}_{ki}))(\prod_{k,i}\delta(\psi_{ki}-\hat{\psi}_{ki}))(\prod_{k,i}\delta(\alpha_{ki}-\hat{\alpha}_{ki}))$$

$$\cdot Q(b,f)\prod_i Q(m_i\mid b,f).$$

B. J. FREY & N. JOJIC

Occlusion model:
EM as an algorithm for minimizing the free energy

E

$$Q(m_i=1|b,f) \leftarrow \frac{\alpha_{fi} N(z_i;\mu_{fi},\psi_{fi})}{\alpha_{fi} N(z_i;\mu_{fi},\psi_{fi}) + (1-\alpha_{fi}) N(z_i;\mu_{bi},\psi_{bi})},$$

$$Q(b,f) \leftarrow c\pi_b\pi_f \exp\{-\sum_i (Q(m_i=1|b,f)(\frac{(z_i-\mu_{fi})^2}{2\psi_{fi}} + \frac{\log 2\pi\psi_{fi}}{2}) + (1-Q(m_i=1|b,f))(\frac{(z_i-\mu_{bi})^2}{2\psi_{bi}} + \frac{\log 2\pi\psi_{bi}}{2}))\},$$

M

$$\pi_k \leftarrow (\sum_t Q(f^{(t)}=k) + \sum_t Q(b^{(t)}=k))/(2T),$$

$$\alpha_{ki} \leftarrow \frac{\sum_t Q(m_i^{(t)}=1, f^{(t)}=k)}{\sum_t Q(f^{(t)}=k)},$$

$$\mu_{ki} \leftarrow \frac{\sum_t (Q(m_i^{(t)}=1, f^{(t)}=k) + Q(m_i^{(t)}=0, b^{(t)}=k)) z_i^{(t)}}{\sum_t (Q(m_i^{(t)}=1, f^{(t)}=k) + Q(m_i^{(t)}=0, b^{(t)}=k))}.$$

$$\psi_{ki} \leftarrow \frac{\sum_t (Q(m_i^{(t)}=1, f^{(t)}=k) + Q(m_i^{(t)}=0, b^{(t)}=k)) (z_i^{(t)}-\mu_{ki})^2}{\sum_t (Q(m_i^{(t)}=1, f^{(t)}=k) + Q(m_i^{(t)}=0, b^{(t)}=k))}.$$

$$Q(b) \leftarrow \sum_f Q(b,f)$$

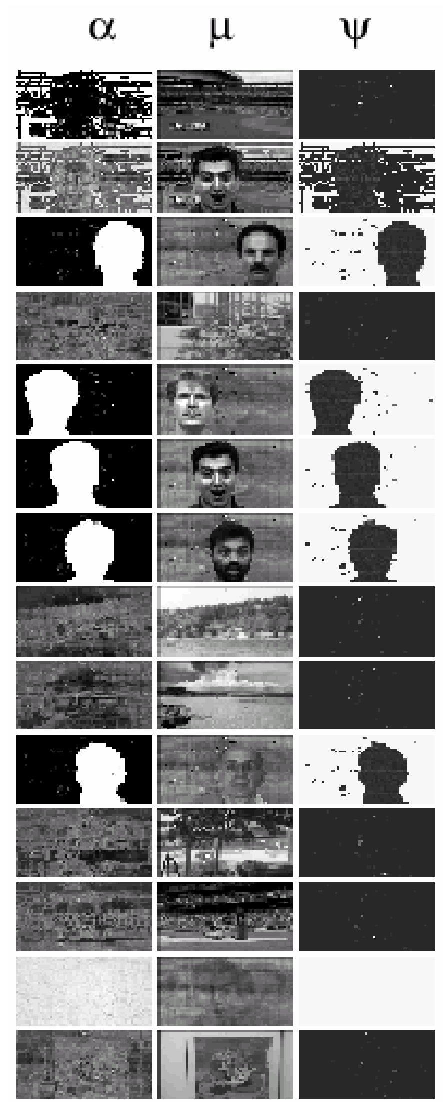
$$Q(f) \leftarrow \sum_b Q(b,f)$$

$$Q(m_i=1,b) \leftarrow \sum_f Q(m_i=1|b,f) Q(b,f)$$

$$Q(m_i=1,f) \leftarrow \sum_b Q(m_i=1|b,f) Q(b,f)$$

B. J. FREY & N. JOJIC

Toy problem:
Result of EM learning



B. J. FREY & N. JOJIC

Generalizations of EM

- Update h^{θ} so as to only decrease F (not minimize F)
- Instead of updating Q using $Q(h^{(t)}) \leftarrow P(h^{(t)} \mid v^{(t)}, \hat{h}^{\theta})$ update Q so as to only decrease F
- Use another inference technique for the Q-distribution over hidden variables…

Variational techniques

- Parameterize the Q-distribution: $Q(h;\phi)$
- Now, the *variational* free energy is

$$F(Q,P) = \int_h Q(h;\phi)\log\frac{Q(h;\phi)}{P(h,v)}.$$

- Computational task: Min F w.r.t ϕ
- Choose form of Q so math works out
- Example: Mean Field, $Q(h) = \prod_{i=1}^{M} Q(h_i),$

Variational inference

Initialization. Pick values for the variational parameters, ϕ (randomly, or cleverly).

Optimization Step. Decrease $F(Q, P)$ by adjusting the parameter vector ϕ, or a subset of ϕ.

Repeat for a fixed number of iterations or until convergence.

Variational EM

Initialization. Pick values for the variational parameters $\phi^{(1)}, \ldots, \phi^{(T)}$ and the model parameters $\hat{h}^{\theta}$ (randomly, or cleverly).

Generalized E Step. Starting from the variational parameters from the previous iteration, modify $\phi^{(1)}, \ldots, \phi^{(T)}$ so as to decrease F.

Generalized M Step. Starting from the model parameters from the previous iteration, modify $\hat{h}^{\theta}$ so as to decrease F.

Repeat for a fixed number of iterations or until convergence.

Occlusion model:
Variational EM for minimizing the free energy

$$F(Q,P) = \int_h Q(h)\log Q(h) - \int_h Q(h)\log P(h,v)$$

$$= \int_h Q(h)\log Q(h) - \int_{h^\theta} Q(h^\theta)\log P(h^\theta) -$$

$$-\sum_{t=1}^{T}\int_{h^{(t)},h^\theta} Q(h^{(t)},h^\theta)\log P(h^{(t)},v^{(t)}\mid h^\theta).$$

$$P \propto \prod_{t=1}^{T}(\pi_{f^{(t)}}\pi_{b^{(t)}}(\prod_{i=1}^{K}\alpha_{f^{(t)}i}^{m_i^{(t)}}(1-\alpha_{f^{(t)}i})^{1-m_i^{(t)}} N(z_i^{(t)};\mu_{f^{(t)}i},\psi_{f^{(t)}i})^{m_i^{(t)}} N(z_i^{(t)};\mu_{b^{(t)}i},\psi_{b^{(t)}i})^{1-m_i^{(t)}})).$$

$$Q = (\prod_k \delta(\pi_k-\hat{\pi}_k))(\prod_{k,i}\delta(\mu_{ki}-\hat{\mu}_{ki}))(\prod_{k,i}\delta(\psi_{ki}-\hat{\psi}_{ki}))(\prod_{k,i}\delta(\alpha_{ki}-\hat{\alpha}_{ki}))$$

$$\cdot Q(b)Q(f)\prod_i Q(m_i).$$

VEM

B. J. FREY & N. JOJIC

Occlusion model:
Variational EM for minimizing the free energy

$$F = \sum_b Q(b)\log\frac{Q(b)}{\pi_b} + \sum_f Q(f)\log\frac{Q(f)}{\pi_f}$$

$$+\sum_i (q_i \log q_i + (1-q_i)\log(1-q_i)) - \sum_i (q_i(\sum_f Q(f)\log\alpha_{fi}) + (1-q_i)(\sum_f Q(f)\log(1-\alpha_{fi})))$$

$$+\sum_i \sum_f Q(f) q_i (\frac{(z_i-\mu_{fi})^2}{2\psi_{fi}} + \frac{\log 2\pi\psi_{fi}}{2}) + \sum_i \sum_b Q(b)(1-q_i)(\frac{(z_i-\mu_{bi})^2}{2\psi_{bi}} + \frac{\log 2\pi\psi_{bi}}{2}).$$

$$P \propto \prod_{t=1}^{T} (\pi_{f^{(t)}} \pi_{b^{(t)}} (\prod_{i=1}^{K} \alpha_{f^{(t)}i}^{m_i^{(t)}} (1-\alpha_{f^{(t)}i})^{1-m_i^{(t)}} N(z_i^{(t)}; \mu_{f^{(t)}i}, \psi_{f^{(t)}i})^{m_i^{(t)}} N(z_i^{(t)}; \mu_{b^{(t)}i}, \psi_{b^{(t)}i})^{1-m_i^{(t)}})).$$

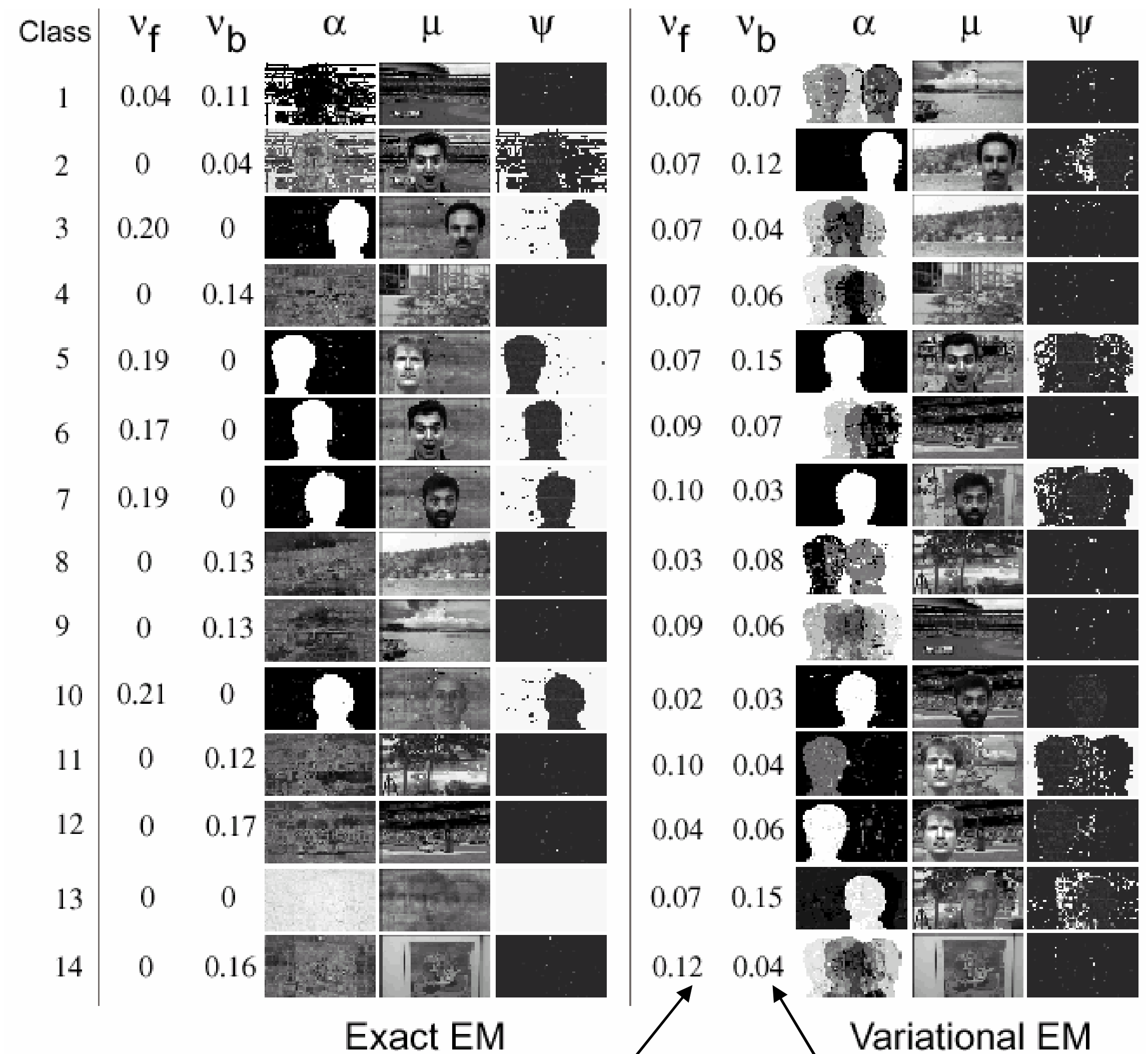
$$Q = (\prod_k \delta(\pi_k - \hat{\pi}_k))(\prod_{k,i} \delta(\mu_{ki} - \hat{\mu}_{ki}))(\prod_{k,i} \delta(\psi_{ki} - \hat{\psi}_{ki}))(\prod_{k,i} \delta(\alpha_{ki} - \hat{\alpha}_{ki}))$$

$$\cdot Q(b)Q(f)\prod_i Q(m_i).$$

VEM

B. J. FREY & N. JOJIC

Occlusion model:
Comparison of exact EM with Variational EM



B. J. FREY & N. JOJIC

Structured variational techniques

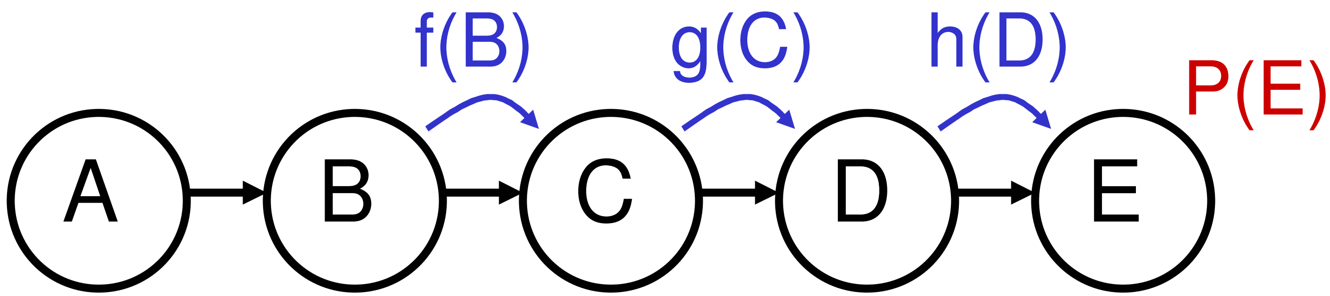
- We can also describe Q using a parameterized graphical model
- Choose the structure and form of Q so that

$$F(Q,P) = \int_h Q(h;\phi) \log \frac{Q(h;\phi)}{P(h,v)}.$$

is tractable

- Structured variational technique for toy problem: See tutorial paper

The sum-product algorithm (belief propagation)



$$P(A,B,C,D,E) = P(E|D)P(D|C)P(C|B)P(B|A)P(A)$$

$$P(E) = \Sigma_D \Sigma_C \Sigma_B \Sigma_A P(E|D)P(D|C)P(C|B)P(B|A)P(A)$$

$$= \Sigma_D P(E|D)\Big[\Sigma_C P(D|C)\Big[\Sigma_B P(C|B)\underbrace{\Big[\Sigma_A P(B|A)P(A)\Big]}_{f(B)}\Big]\Big]$$

f(B) = $[\Sigma_A P(B|A)P(A)]$

g(C) = $[\Sigma_B P(C|B)[\Sigma_A P(B|A)P(A)]]$

h(D) = $[\Sigma_C P(D|C)[\Sigma_B P(C|B)[\Sigma_A P(B|A)P(A)]]]$

B. J. FREY & N. JOJIC

SP Algorithm as Free Energy Minimization

For a graphical model with potentials, $\Psi_i(x_{C_i})$ on cliques x_{C_i},

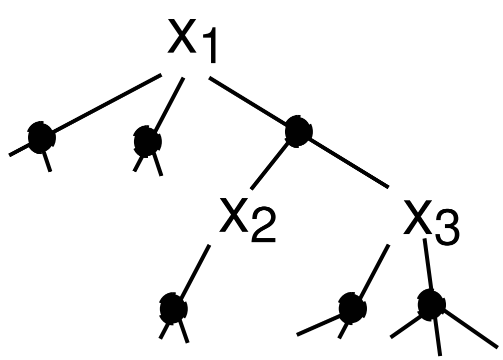
$$F = -\Sigma_i \Sigma_{x_{C_i}} Q(x_{C_i}) \log[\Psi_i(x_{C_i})] + \Sigma_x Q(x) \log[Q(x)]$$

- For a general Q-distribution, if $Q(x_{C_i})$ are known, the 1st term is easy to compute
- Problem: 2nd term is generally intractable
- Idea: Represent Q(x) by $Q(x_j)$, $\forall j$, and $Q(x_{C_i})$, $\forall i$, and approximate the 2nd term

Bethe approximation

- For a tree

$$Q(x) = \frac{\prod_i Q(x_{C_i})}{\prod_j Q(x_j)^{d_j-1}}$$



- d_j = degree of x_j
- Use this expreesion for the entropy, even when the graph is not a tree
- Note: For a tree, the minimizer of F gives $Q(x) = P(x)$: Exact inference

Solving for Q

$$F = -\Sigma_i \Sigma_{x_{C_i}} Q(x_{C_i}) \log[\Psi_i(x_{C_i})]$$

$$+\Sigma_i \Sigma_{x_{C_i}} Q(x_{C_i}) \log[Q(x_{C_i})] - \Sigma_j (d_j - 1) \Sigma_{x_j} Q(x_j) \log[Q(x_j)]$$

Solve $\partial D / \partial Q(x_{C_i}) = 0$ and $\partial D / \partial Q(x_j) = 0$, subject to the constraints

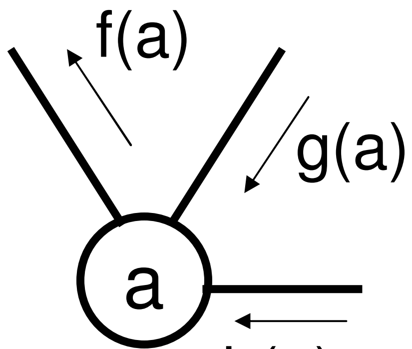
- $\Sigma_{x_j} Q(x_j) = 1$
- $\Sigma_{x_{C_i}} Q(x_{C_i}) = 1$
- For $j \in C_i$, $\Sigma_{x_{C_i \setminus j}} Q(x_{C_i}) = Q(x_j)$

Result (Yedidia et al 2001): Sum-product algorithm

The sum-product algorithm

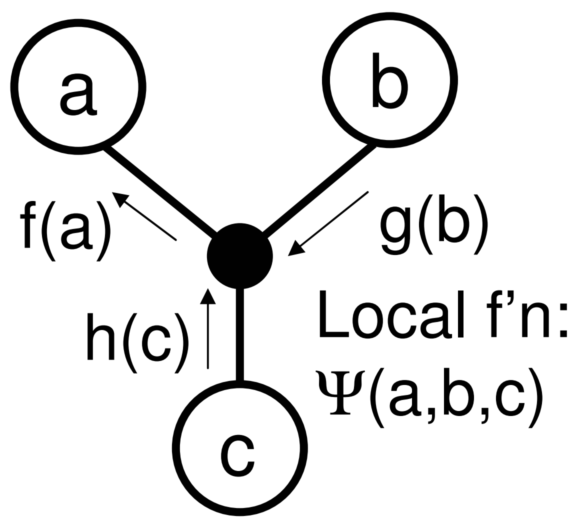
Each message is a function of its neighboring variable

Out of variable



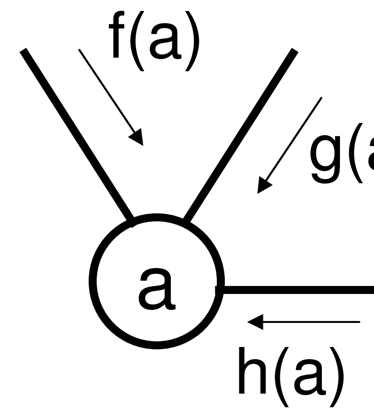
$f(a) = g(a)h(a)$

Out of function



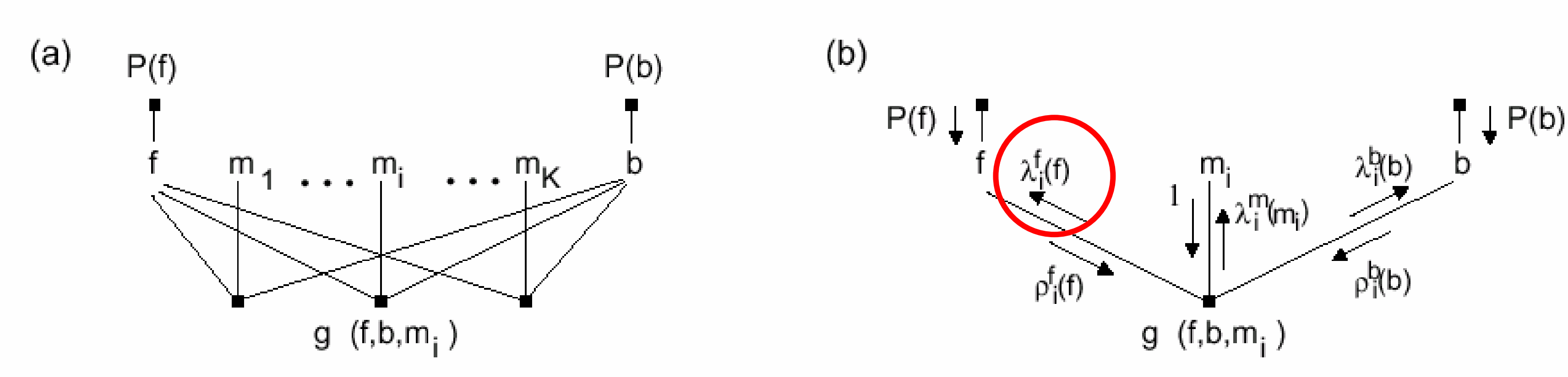
$f(a) = \sum_b \sum_c \Psi(a,b,c) g(b) h(c)$

Fusion



$P(a) \approx f(a)g(a)h(a)$

Occlusion model:
The sum-product algorithm



$$g_i(f,b,m_i) = P(z_i \mid m_i, f, b)P(m_i \mid f) =$$

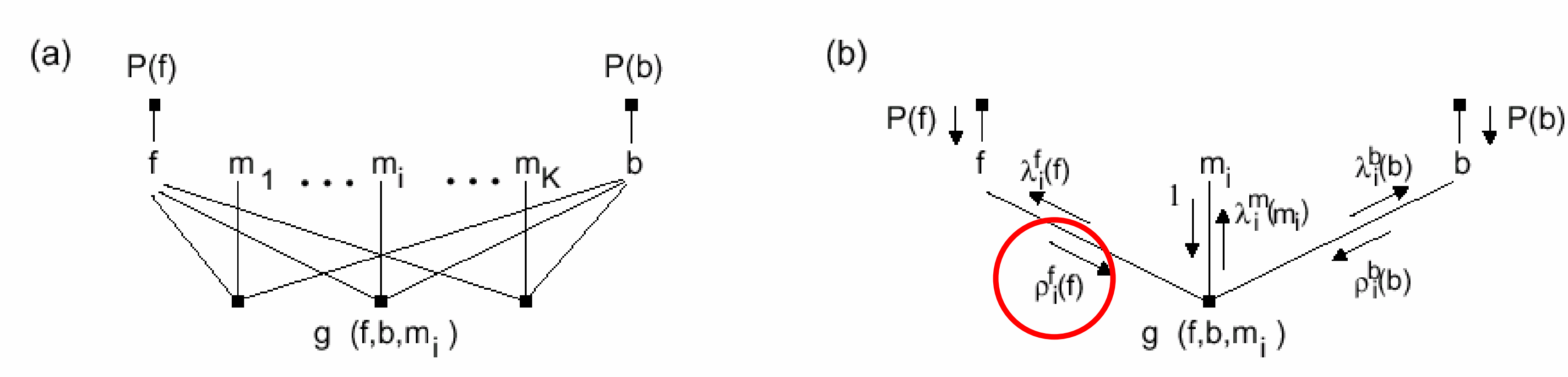
$$N(z_i;\mu_{fi},\psi_{fi})^{m_i} N(z_i;\mu_{bi},\psi_{bi})^{1-m_i} \alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i}.$$

$$\lambda_i^f(f) \leftarrow \alpha_{fi} N(z_i;\mu_{fi},\psi_{fi}) + (1-\alpha_{fi})\sum_b N(z_i;\mu_{bi},\psi_{bi})\rho_i^b(b).$$

$$\lambda_i^f(f) \leftarrow \lambda_i^f(f)/(\textstyle\sum_f \lambda_i^f(f))$$

B. J. FREY & N. JOJIC

Occlusion model:
The sum-product algorithm



$$g_i(f,b,m_i) = P(z_i \mid m_i, f, b) P(m_i \mid f) =$$

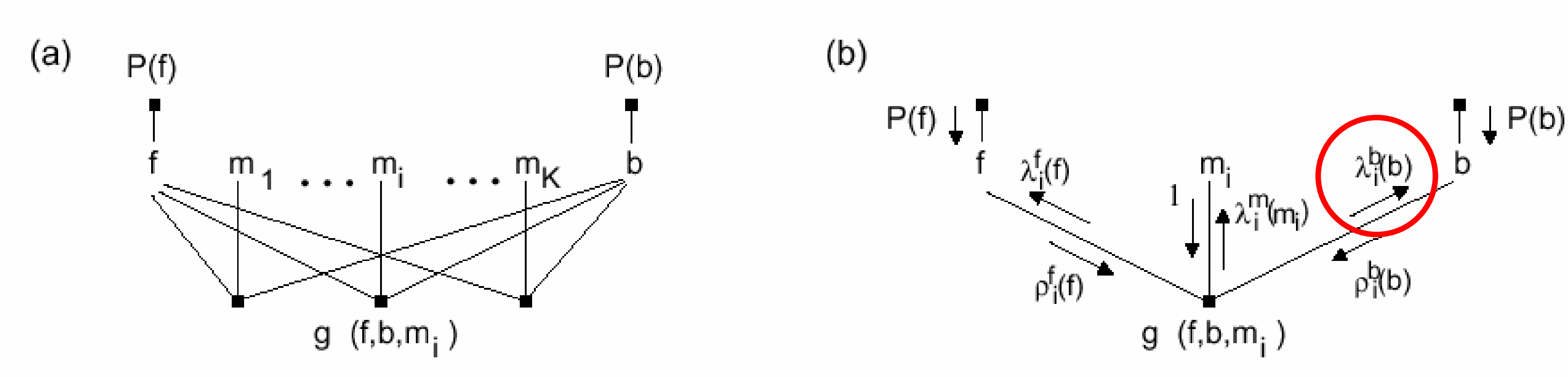
$$N(z_i;\mu_{fi},\psi_{fi})^{m_i} N(z_i;\mu_{bi},\psi_{bi})^{1-m_i} \alpha_{fi}^{m_i} (1-\alpha_{fi})^{1-m_i}.$$

$$\rho_i^f(f) \leftarrow P(f) \prod_{j \neq i} \lambda_i^f(f),$$

$$\rho_i^f(f) \leftarrow \rho_i^f(f) / (\textstyle\sum_f \rho_i^f(f))$$

B. J. FREY & N. JOJIC

Occlusion model:
The sum-product algorithm



$$g_i(f,b,m_i) = P(z_i \mid m_i, f, b)P(m_i \mid f) =$$

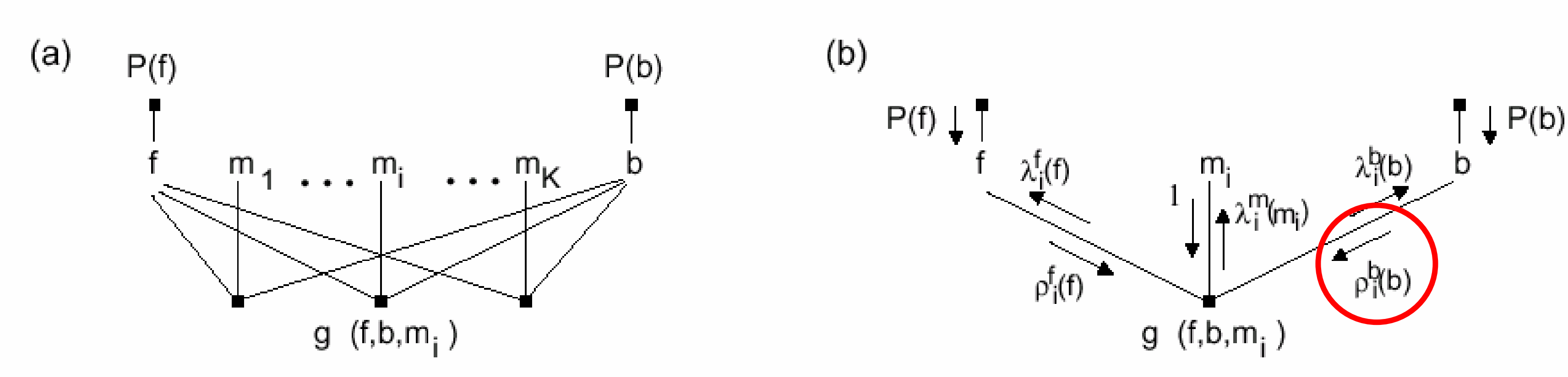
$$N(z_i;\mu_{fi},\psi_{fi})^{m_i} N(z_i;\mu_{bi},\psi_{bi})^{1-m_i} \alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i}.$$

$$\lambda_i^b(b) \leftarrow \sum_f \sum_{m_i} g_i(f,b,m_i)\cdot 1 \cdot \rho_i^f(f)$$

$$\lambda_i^b(b) \leftarrow \lambda_i^b(b)/(\sum_b \lambda_i^b(b))$$

B. J. FREY & N. JOJIC

Occlusion model: The sum-product algorithm



$$g_i(f,b,m_i) = P(z_i \mid m_i, f, b) P(m_i \mid f) =$$

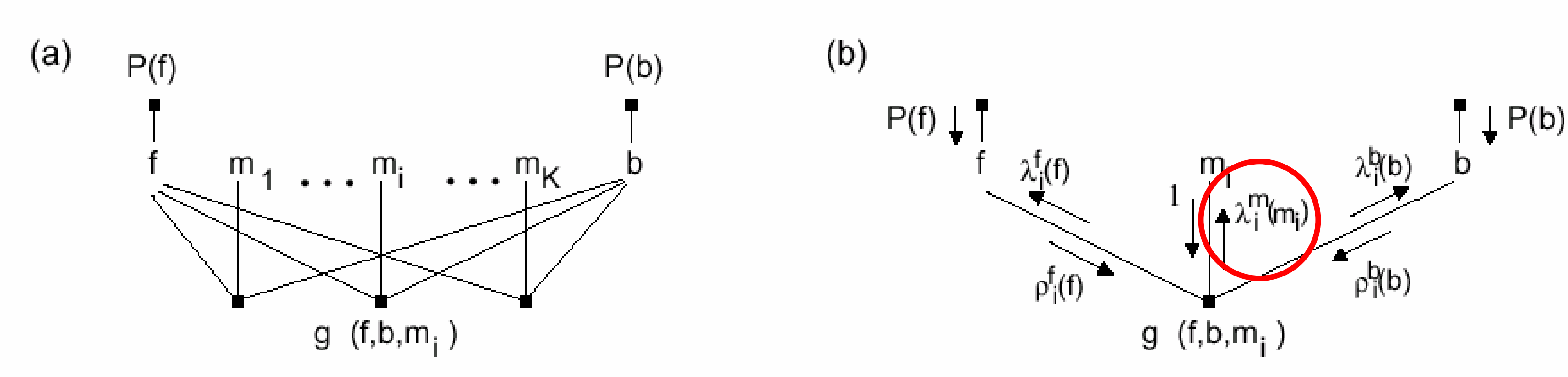
$$N(z_i;\mu_{fi},\psi_{fi})^{m_i} N(z_i;\mu_{bi},\psi_{bi})^{1-m_i} \alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i}.$$

$$\rho_i^b(b) \leftarrow P(b)\prod_{j\neq i}\lambda_i^b(b),$$

$$\rho_i^b(b) \leftarrow \rho_i^b(b)/(\sum\nolimits_b \rho_i^b(b)$$

B. J. FREY & N. JOJIC

Occlusion model: The sum-product algorithm



$$g_i(f,b,m_i) = P(z_i \mid m_i, f, b)P(m_i \mid f) =$$

$$N(z_i;\mu_{fi},\psi_{fi})^{m_i} N(z_i;\mu_{bi},\psi_{bi})^{1-m_i} \alpha_{fi}^{m_i}(1-\alpha_{fi})^{1-m_i}.$$

$$\lambda_i^m(m_i) \leftarrow \sum_f \sum_b g_i(f,b,m_i)\cdot\rho_i^f(f)\cdot\rho_i^b(b):$$

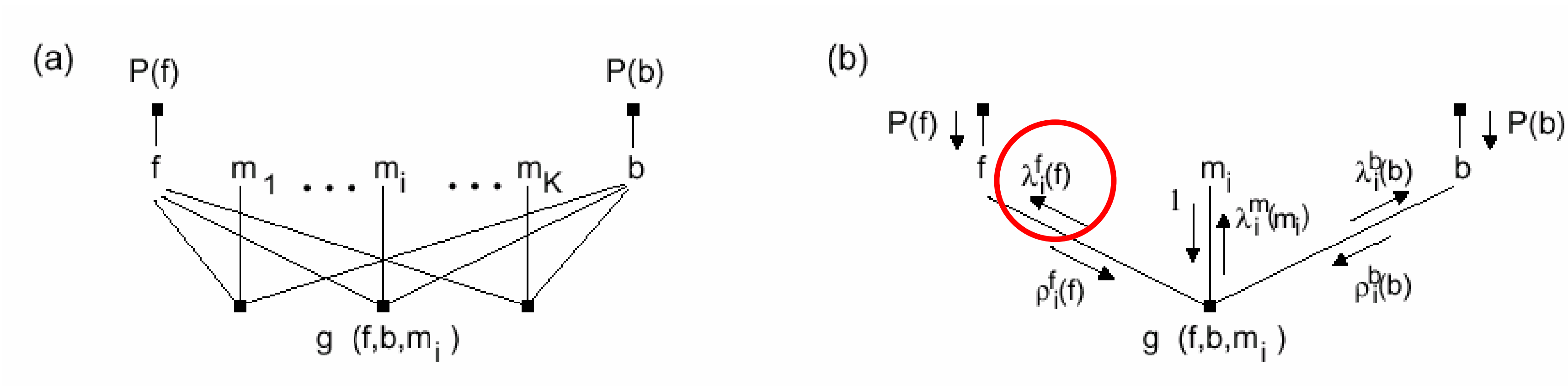
$$\lambda_i^m(1) \leftarrow \sum_f N(z_i;\mu_{fi},\psi_{fi})\alpha_{fi}\rho_i^f(f),$$

$$\lambda_i^m(0) \leftarrow (\sum_b N(z_i;\mu_{bi},\psi_{bi})\rho_i^b(b))(\sum_f (1-\alpha_{fi})\rho_i^f(f)).$$

$$\lambda_i^m(m_i) \leftarrow \lambda_i^m(m_i)/(\lambda_i^m(0)+\lambda_i^m(1))$$

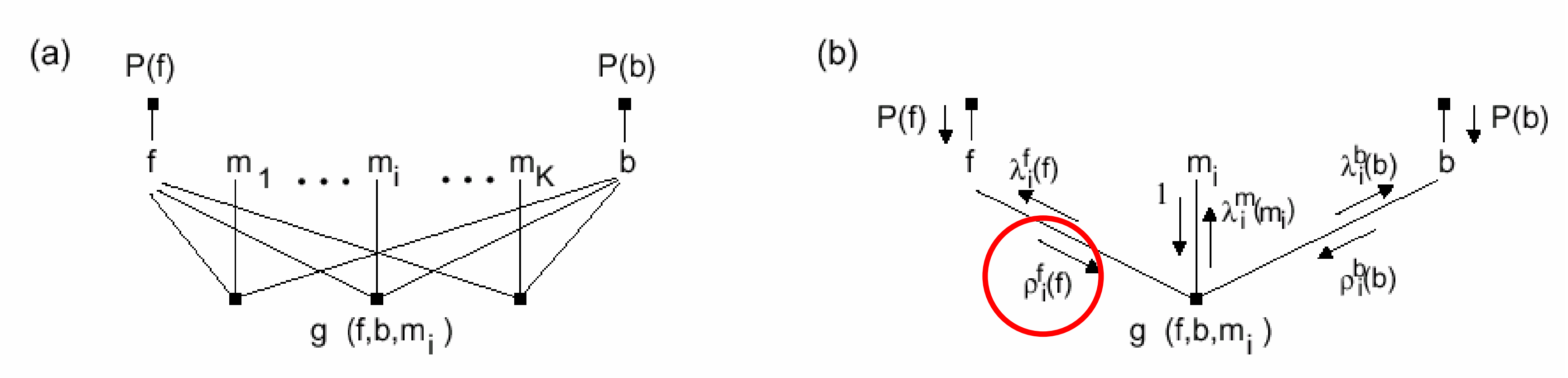
B. J. FREY & N. JOJIC

Occlusion model:
The sum-product algorithm



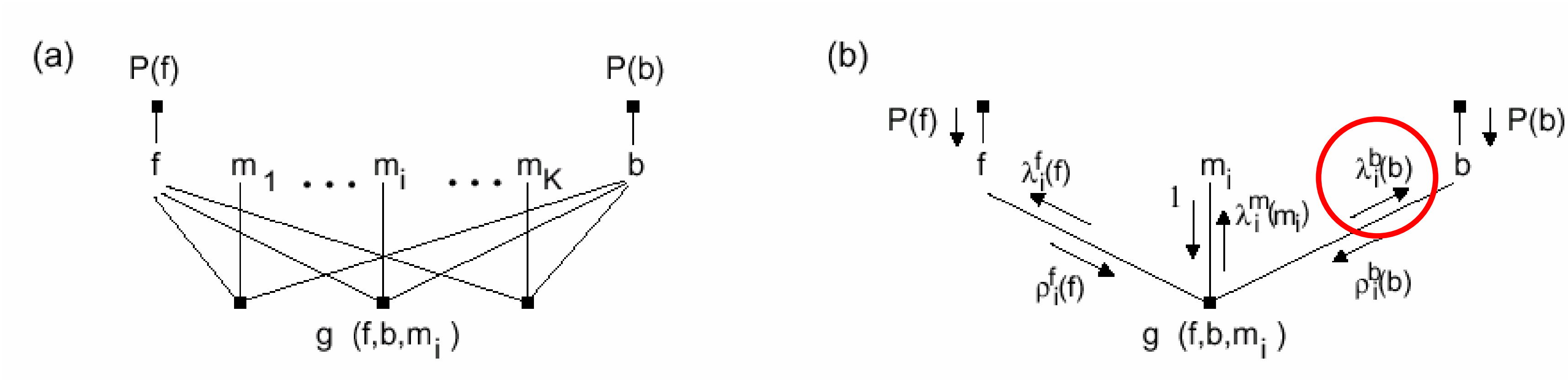
B. J. FREY & N. JOJIC

Occlusion model: The sum-product algorithm



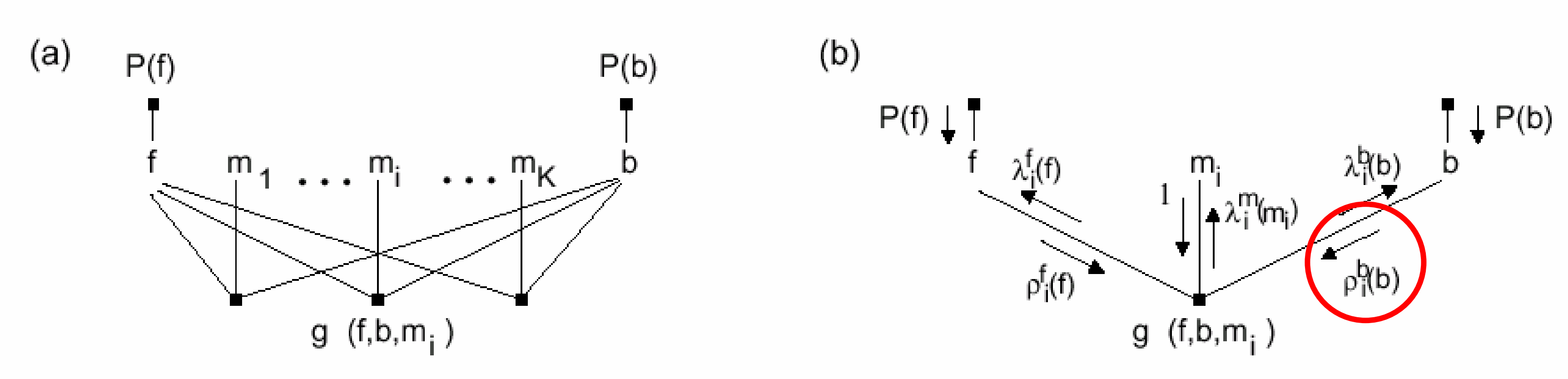
B. J. FREY & N. JOJIC

Occlusion model:
The sum-product algorithm



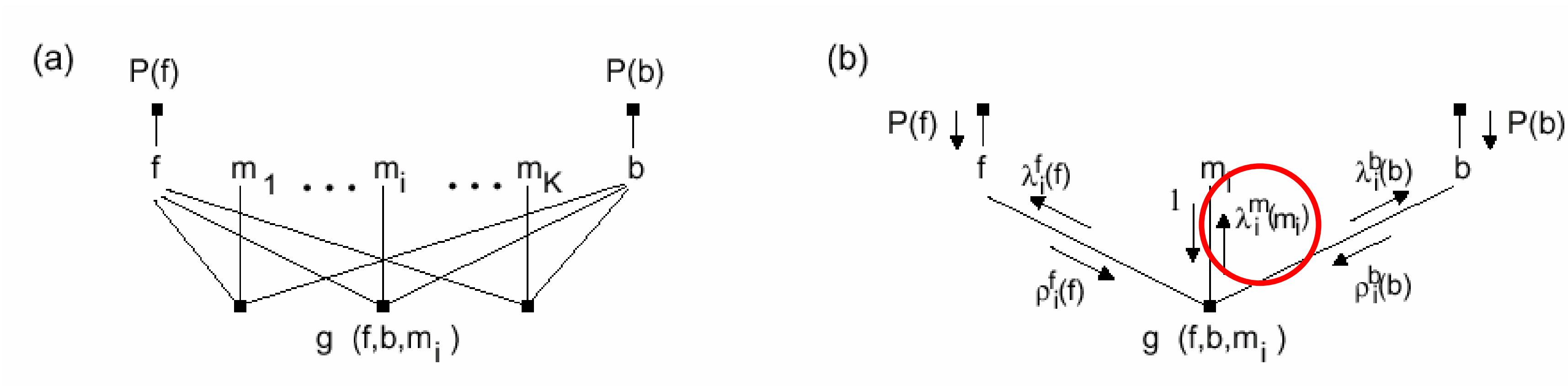
B. J. FREY & N. JOJIC

Occlusion model:
The sum-product algorithm



B. J. FREY & N. JOJIC

Occlusion model:
The sum-product algorithm



B. J. FREY & N. JOJIC

Gibbs sampling

- The posterior $P(h_1, h_2, \ldots, h_K \mid v)$ is not tractable, but often, the conditionals, $P(h_i | h \setminus h_i, v)$, can be sampled from
- Gibbs sampling:

Initialization. Pick values for all hidden variables (randomly, or cleverly).

Sampling Step. Choose a variable h_i at random or in order, and then sample it from $P(h_i | h \setminus h_i)$.

Repeat for a fixed number of iterations or until convergence.

Gibbs sampling as free energy minimization

- Imagine running an ensemble of Gibbs chains in parallel
- Let $Q^n(h)$ describe the distribution of the h's at step n of the sampling procedure
- Suppose at step n, we sample h_i in every chain: $$Q^{n+1}(h) = Q^n(h \setminus h_i)P(h_i|h \setminus h_i, v)$$
- Substituting these into the expression for F, we find that $F^{n+1} \leq F^n$

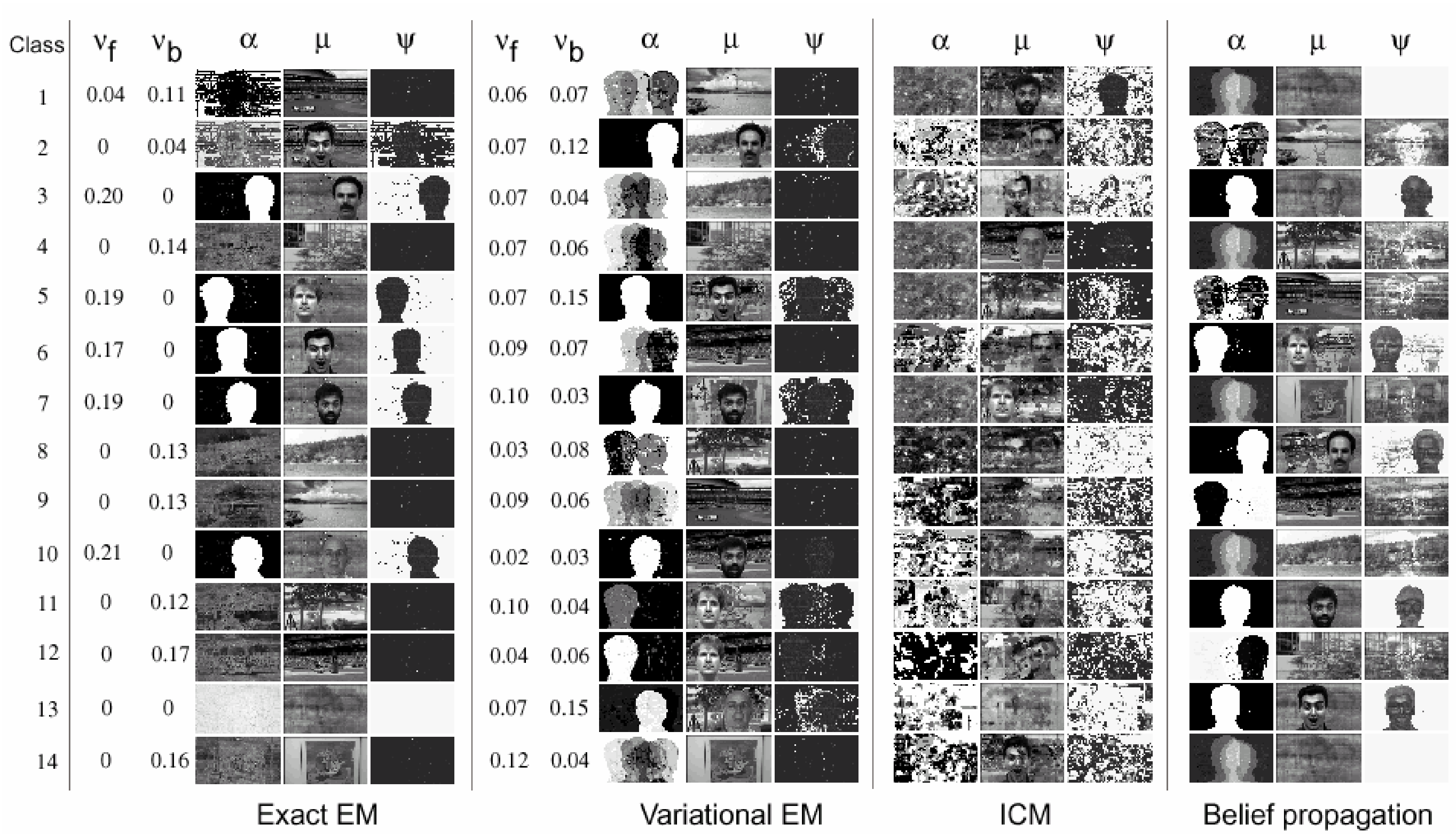
Discussion of Algorithms

Comparison of algorithms for occlusion model

Method	Update for mask variables	Complexity
Exact inference (used in EM)	$\frac{Q(m_i=1\|b,f)}{Q(m_i=0\|b,f)} \leftarrow \frac{\alpha_{fi}\mathcal{N}(z_i;\mu_{fi},\psi_{fi})}{(1-\alpha_{fi})\mathcal{N}(z_i;\mu_{bi},\psi_{bi})}$	J^2K
Iterative conditional modes	$m_i \leftarrow \begin{cases} 1, & \text{if } \frac{\alpha_{fi}\mathcal{N}(z_i;\mu_{fi}\psi_{fi})}{(1-\alpha_{fi})\mathcal{N}(z_i;\mu_{bi},\psi_{bi})} > 1 \\ 0, & \text{otherwise} \end{cases}$	K
Gibbs sampling	$m_i \leftarrow \text{sample}_{m_i} \left\{ \begin{array}{ll} \alpha_{fi}\mathcal{N}(z_i;\mu_{fi},\psi_{fi}) & \text{if } m_i=1 \\ (1-\alpha_{fi})\mathcal{N}(z_i;\mu_{bi},\psi_{bi}) & \text{if } m_i=0 \end{array} \right\}$	K
Fully-factorized variational	$\frac{Q(m_i=1)}{Q(m_i=0)} \leftarrow \frac{\prod_f(\alpha_{fi}\mathcal{N}(z_i;\mu_{fi},\psi_{fi}))^{Q(f)}}{(\prod_f(1-\alpha_{fi})^{Q(f)})(\prod_b \mathcal{N}(z_i;\mu_{bi},\psi_{bi})^{Q(b)})}$	JK
Structured variational	$\frac{Q(m_i=1\|f)}{Q(m_i=0\|f)} \leftarrow \frac{\alpha_{fi}\mathcal{N}(z_i;\mu_{fi},\psi_{fi})}{(1-\alpha_{fi})\prod_b \mathcal{N}(z_i;\mu_{bi},\psi_{bi})^{Q(b)}}$	JK
Sum-product algorithm	$\frac{Q(m_i=1)}{Q(m_i=0)} \leftarrow \frac{\sum_f \rho_i^f(f)\alpha_{fi}\mathcal{N}(z_i;\mu_{fi},\psi_{fi})}{(\sum_f \rho_i^f(f)(1-\alpha_{fi}))(\sum_b \rho_i^b(b)\mathcal{N}(z_i;\mu_{bi},\psi_{bi}))}$	JK

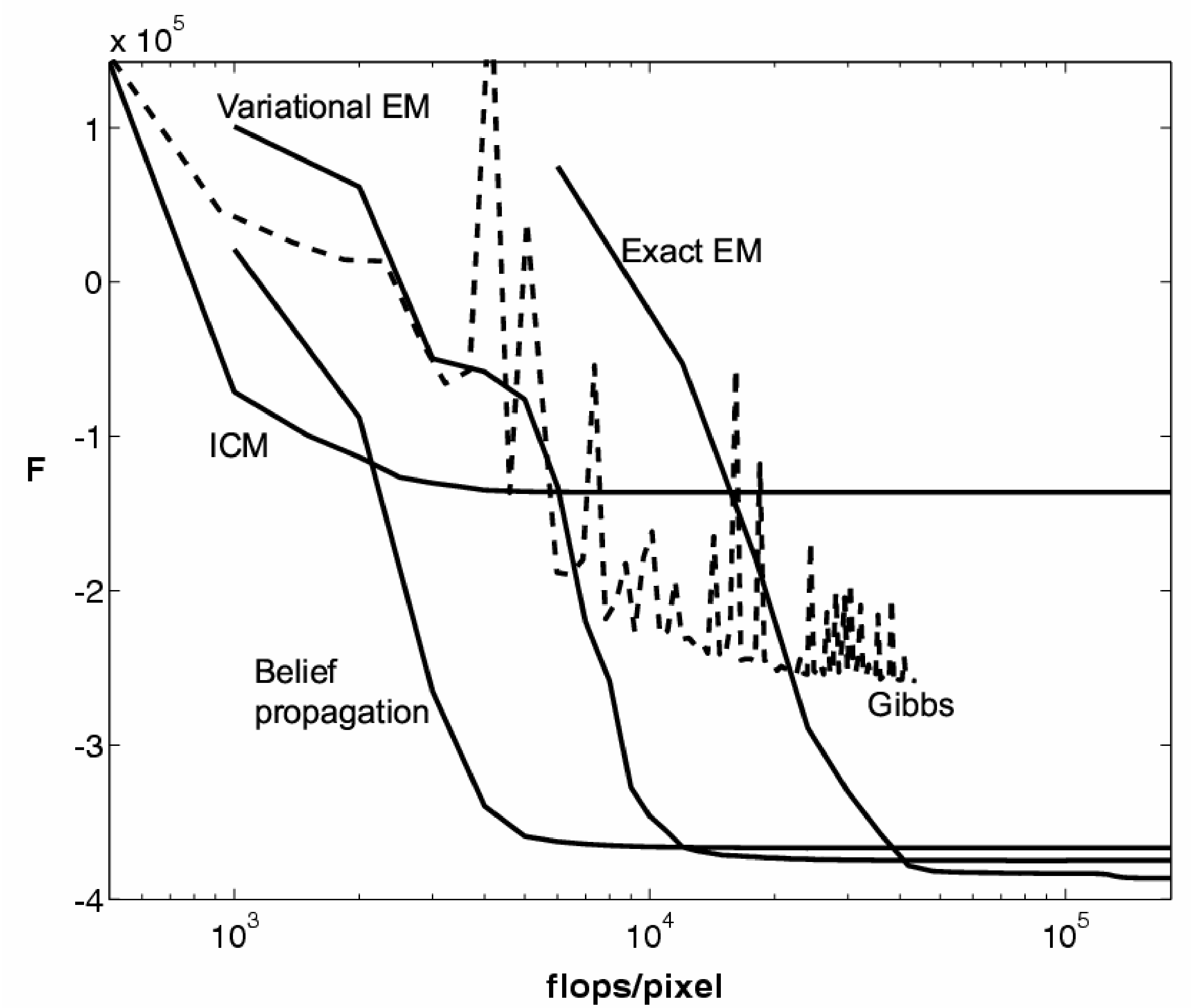
B. J. FREY & N. JOJIC

Experimental results for occlusion model



B. J. FREY & N. JOJIC

Speed of convergence



B. J. FREY & N. JOJIC

Recall: From ICM to Gibbs sampling

Sample

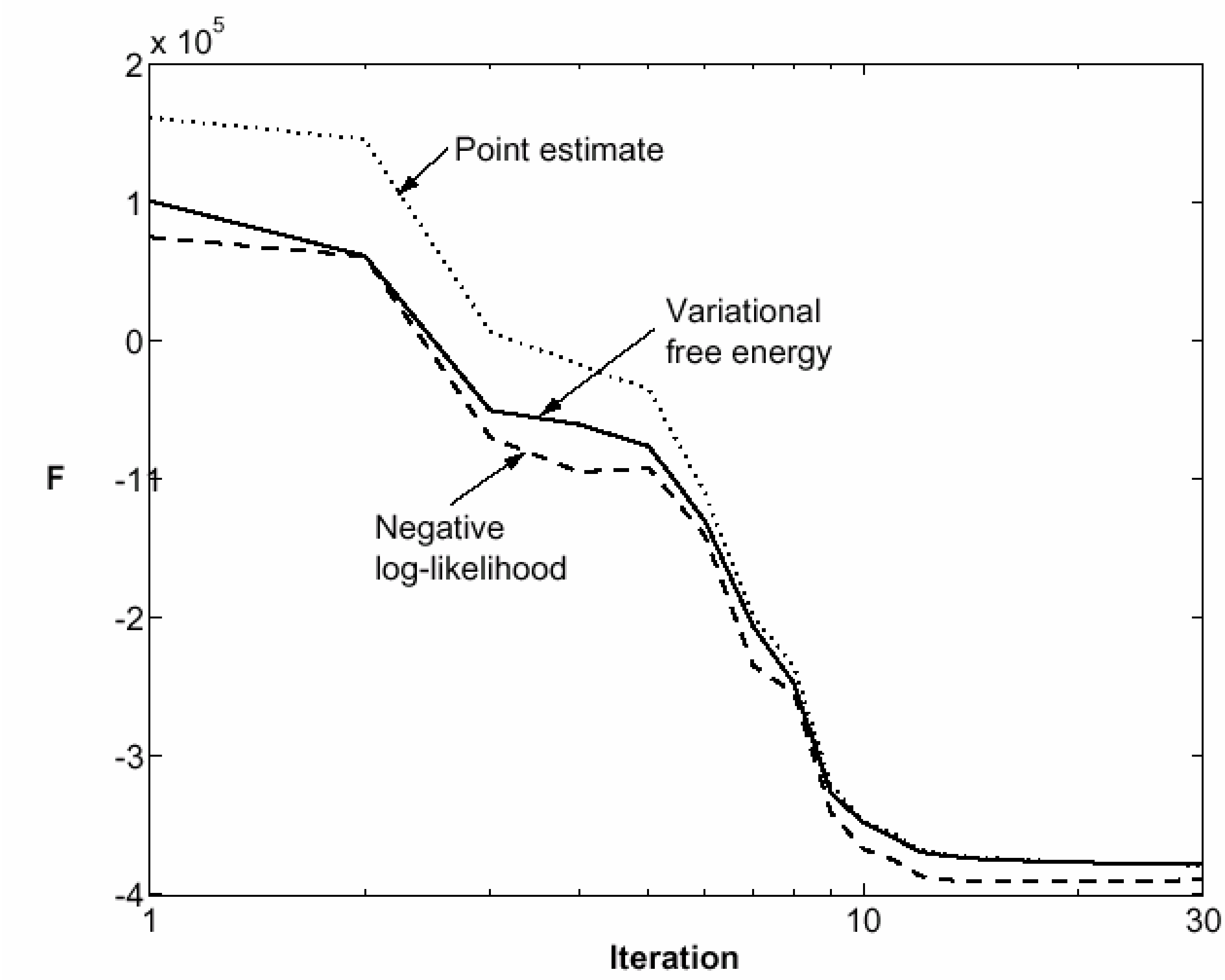
- For $t = 1, \ldots, T$
  - $f^{(t)} \leftarrow \text{argmax}_{f^{(t)}} \left[\pi_{f^{(t)}} \prod_{i:m_i^{(t)}=1} \mathcal{N}(z_i^{(t)}; \mu_{f^{(t)}i}, \psi_{f^{(t)}i})\right]$
  - $b^{(t)} \leftarrow \text{argmax}_{b^{(t)}} \left[\pi_{b^{(t)}} \prod_{i:m_i^{(t)}=0} \mathcal{N}(z_i^{(t)}; \mu_{b^{(t)}i}, \psi_{b^{(t)}i})\right]$
  - For $i = 1, \ldots, K$: $m_i^{(t)} \leftarrow \begin{cases} 1 & \text{if } \alpha_{f^{(t)}i}\mathcal{N}(z_i^{(t)}; \mu_{f^{(t)}i}, \psi_{f^{(t)}i}) > (1-\alpha_{f^{(t)}i})\mathcal{N}(z_i^{(t)}; \mu_{b^{(t)}i}, \psi_{b^{(t)}i}) \\ 0 & \text{otherwise} \end{cases}$
- For $j = 1, \ldots, J$
  - $\pi_j \leftarrow (\sum_{t=1}^{T}[f^{(t)} = j] + \sum_{t=1}^{T}[b^{(t)} = j])/2T$
- For $j = 1, \ldots, J$, for $i = 1, \ldots, K$
  - $\alpha_{ji} \leftarrow (\sum_{t=1}^{T}[f^{(t)} = j]m_i^{(t)})/(\sum_{t=1}^{T}[f^{(t)} = j])$
  - $\mu_{ji} \leftarrow (\sum_{t=1}^{T}[f^{(t)} = j \text{ or } b^{(t)} = j]z_i^{(t)})/(\sum_{t=1}^{T}[f^{(t)} = j \text{ or } b^{(t)} = j])$
  - $\psi_{ji} \leftarrow (\sum_{t=1}^{T}[f^{(t)} = j \text{ or } b^{(t)} = j](z_i^{(t)} - \mu_{ji})^2)/(\sum_{t=1}^{T}[f^{(t)} = j \text{ or } b^{(t)} = j])$

Here, the Iverson notation is used where [True] $= 1$ and [False] $= 0$.

B. J. FREY & N. JOJIC

How tight are the bounds?

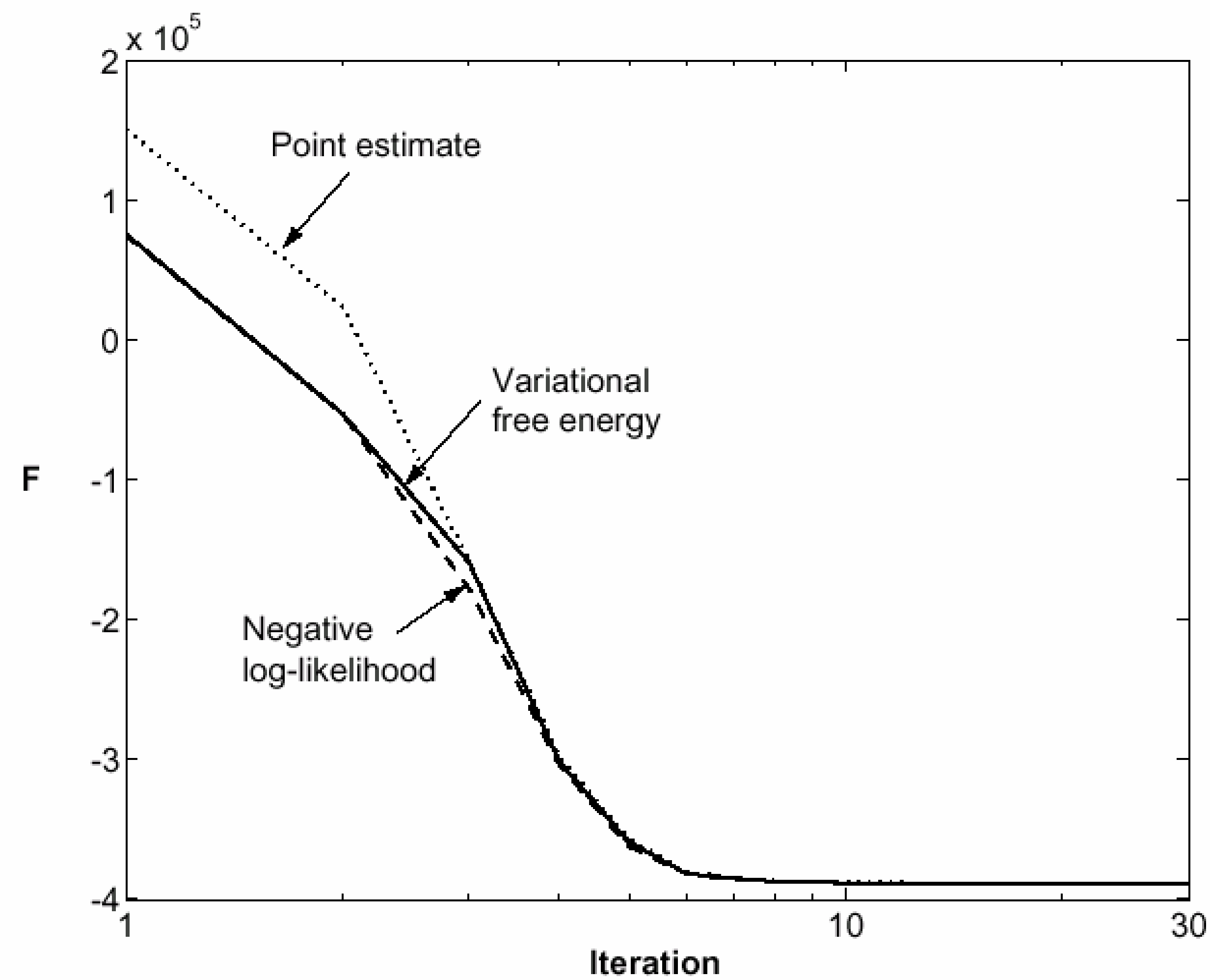
Various bounds during variational learning



B. J. FREY & N. JOJIC

How tight are the bounds?

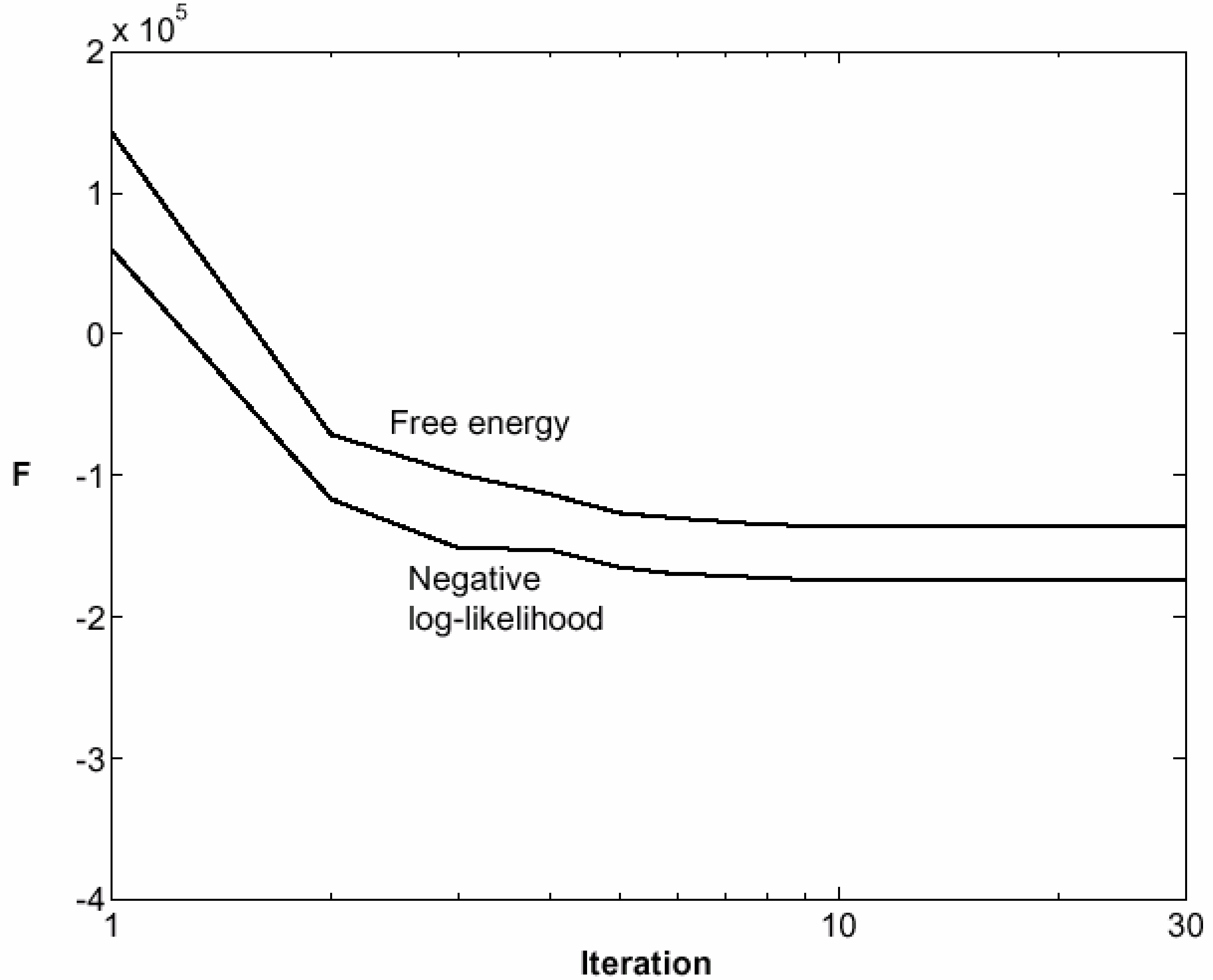
Various bounds during EM learning



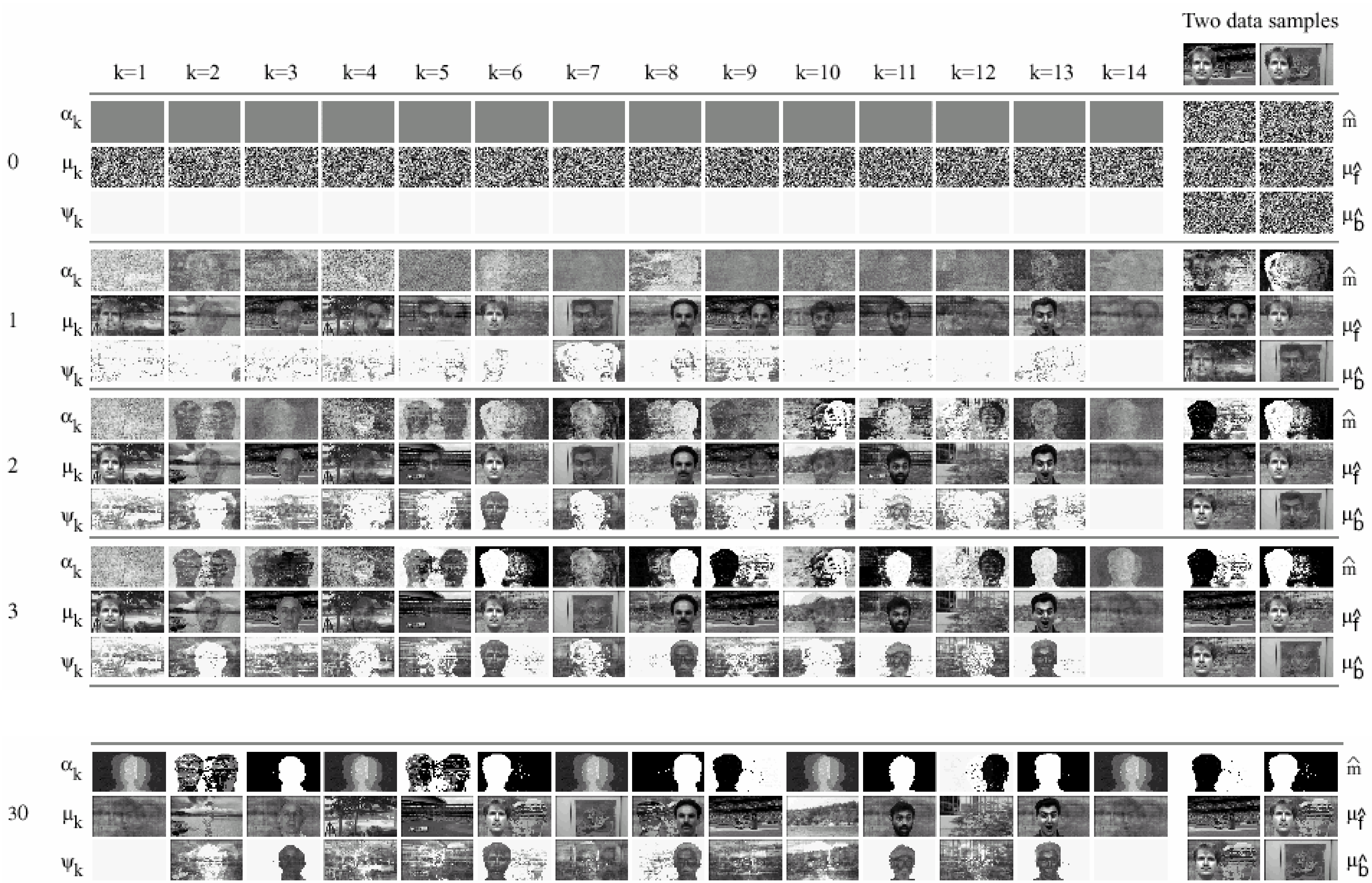
B. J. FREY & N. JOJIC

How tight are the bounds?

Various bounds during ICM learning



Iterative model optimization (sum-product algorithm)



B. J. FREY & N. JOJIC

Improving approximate methods with additional EM iterations

B. J. FREY & N. JOJIC

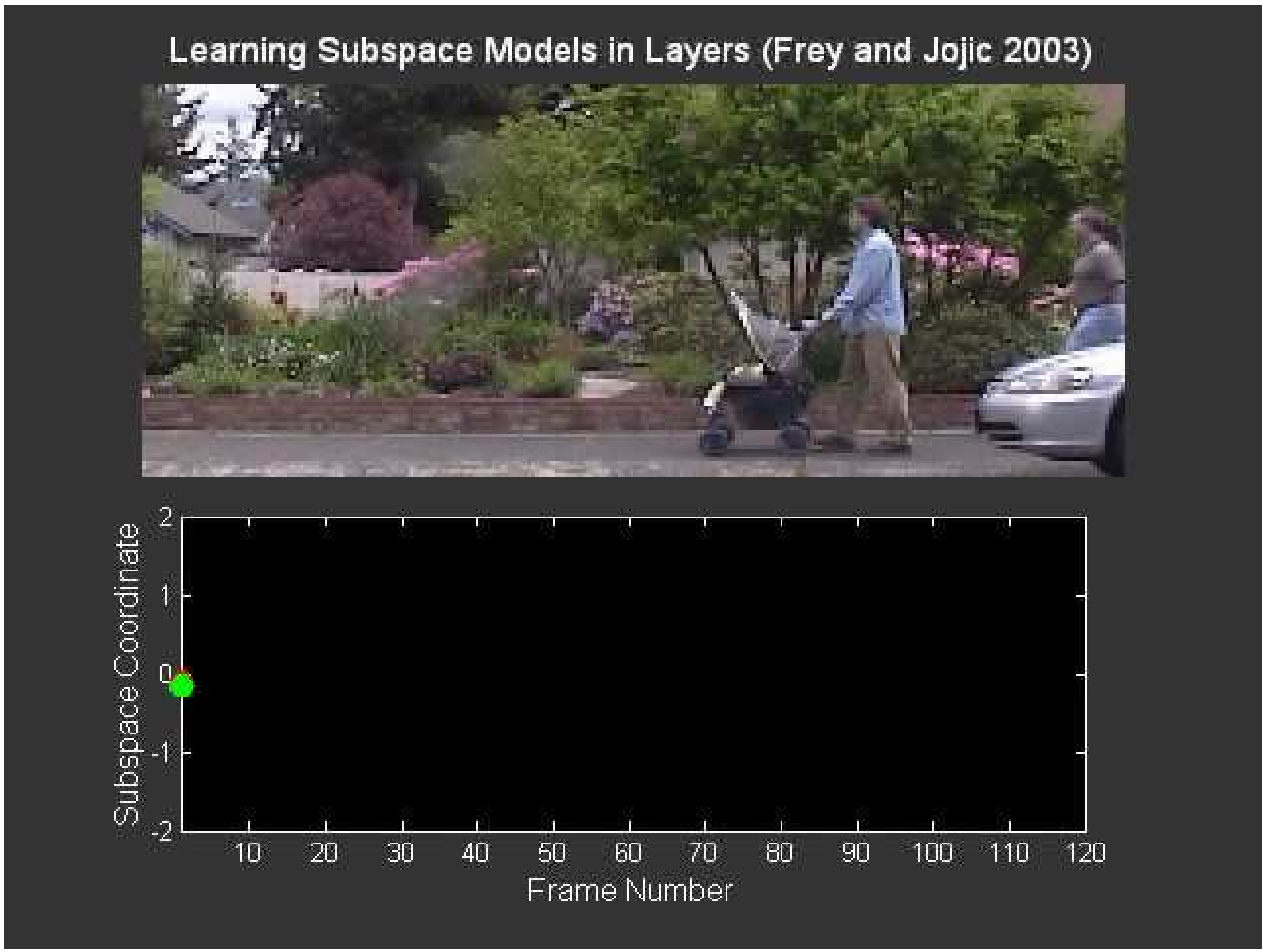
Wrap-up

The Art

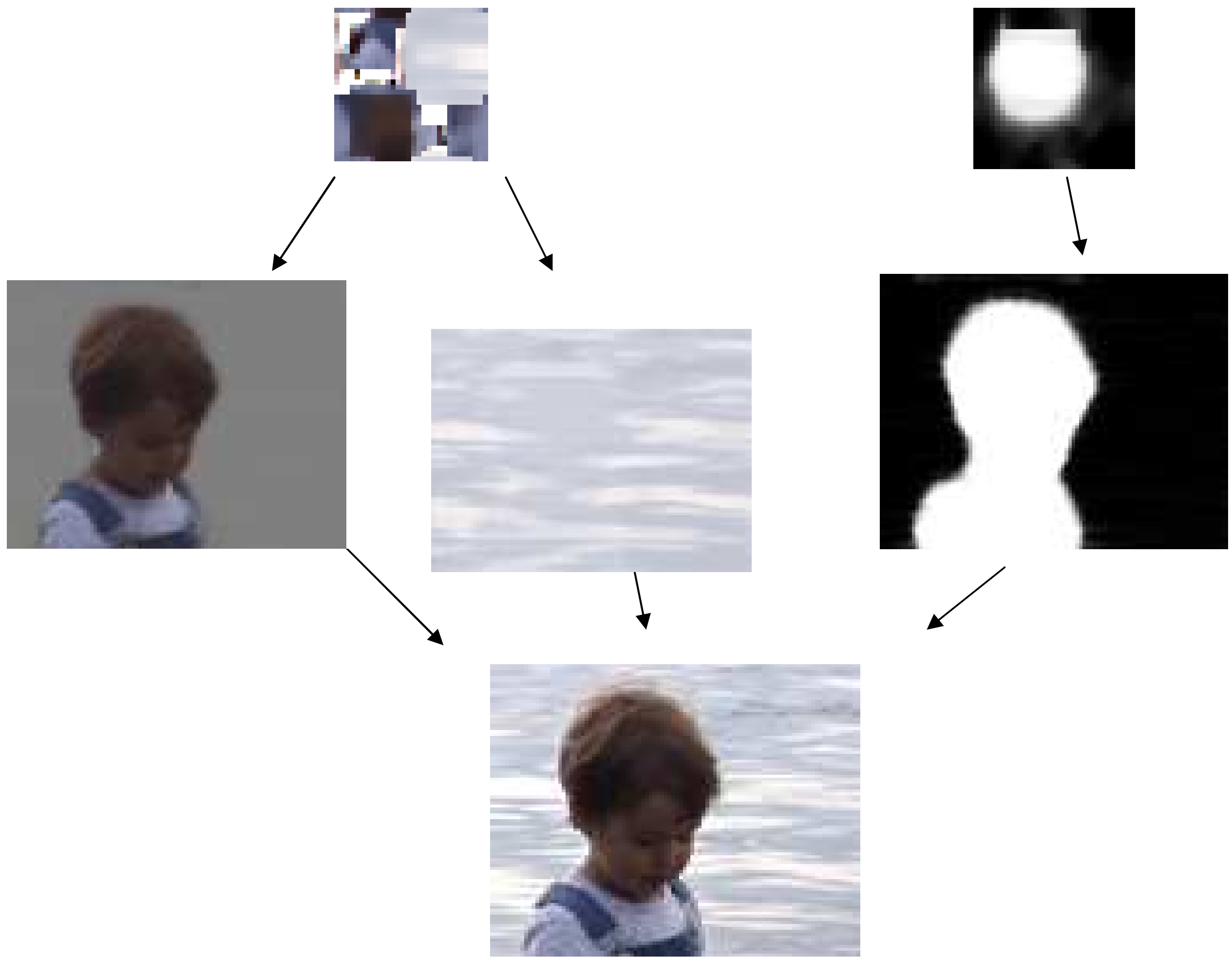
- Representation is critical
- Structure versus forms of probability functions
- What computation should the graph be used for (eg, recursive updates or coordinated optimization)?
- Which is better: Exact inference in inaccurate models or approximate inference in accurate models?
- …

Application: Cardboard cut-out scene analysis
Jojic and Frey, CVPR 2001

Application: Subspace models of occluding objects in 3-D scenes
Frey, Jojic and Kannan 2003



Application: Scene interpretation from single images
Jojic, Frey and Kannan 2003



- Those are the basics

- Go forth, model, perform approximate inference, and have fun!