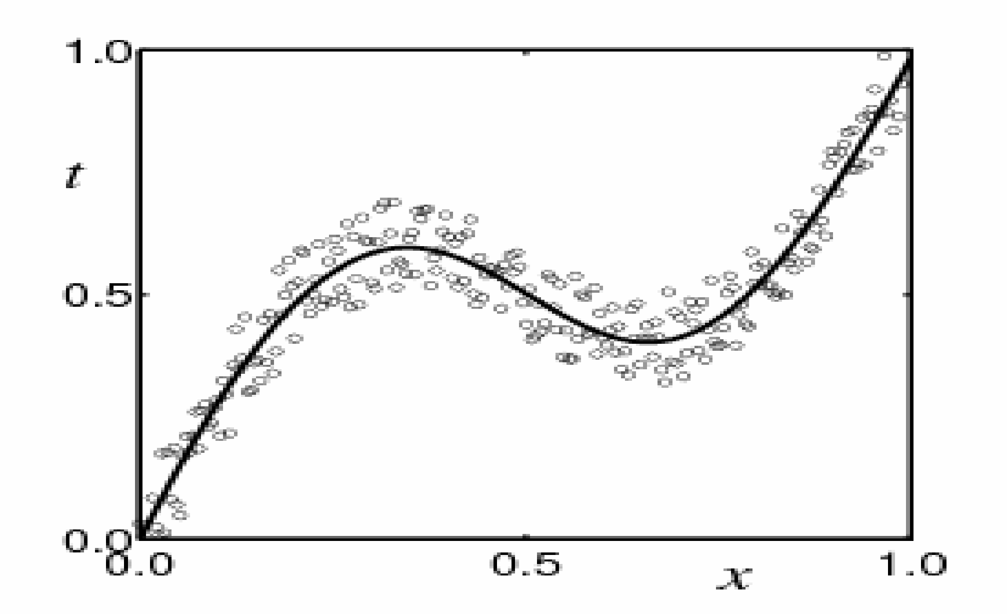
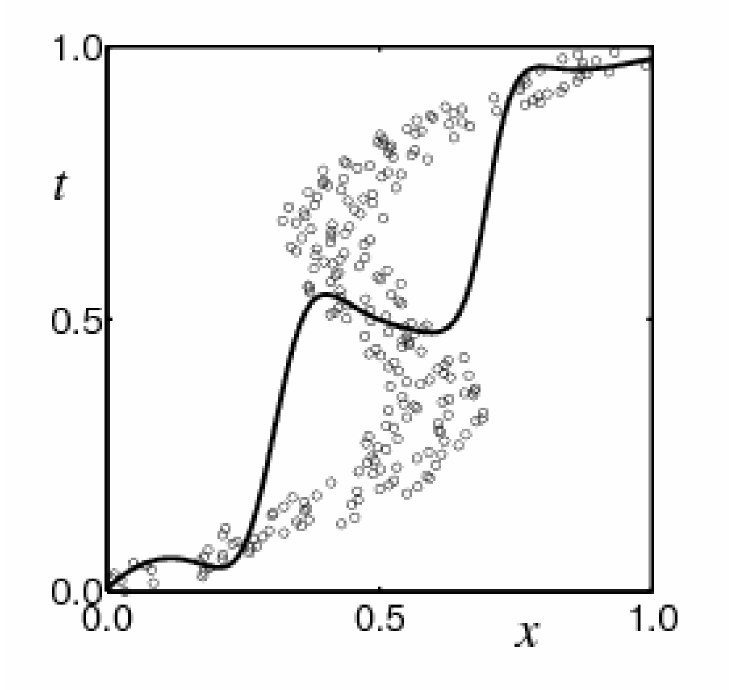
CS540 Machine learning
Lecture 16
EM: theory and applications

Outline

- Conditional mixture models
- EM for Empirical Bayes
- “Sparse Bayesian learning”
- EM theory

One to many "functions"

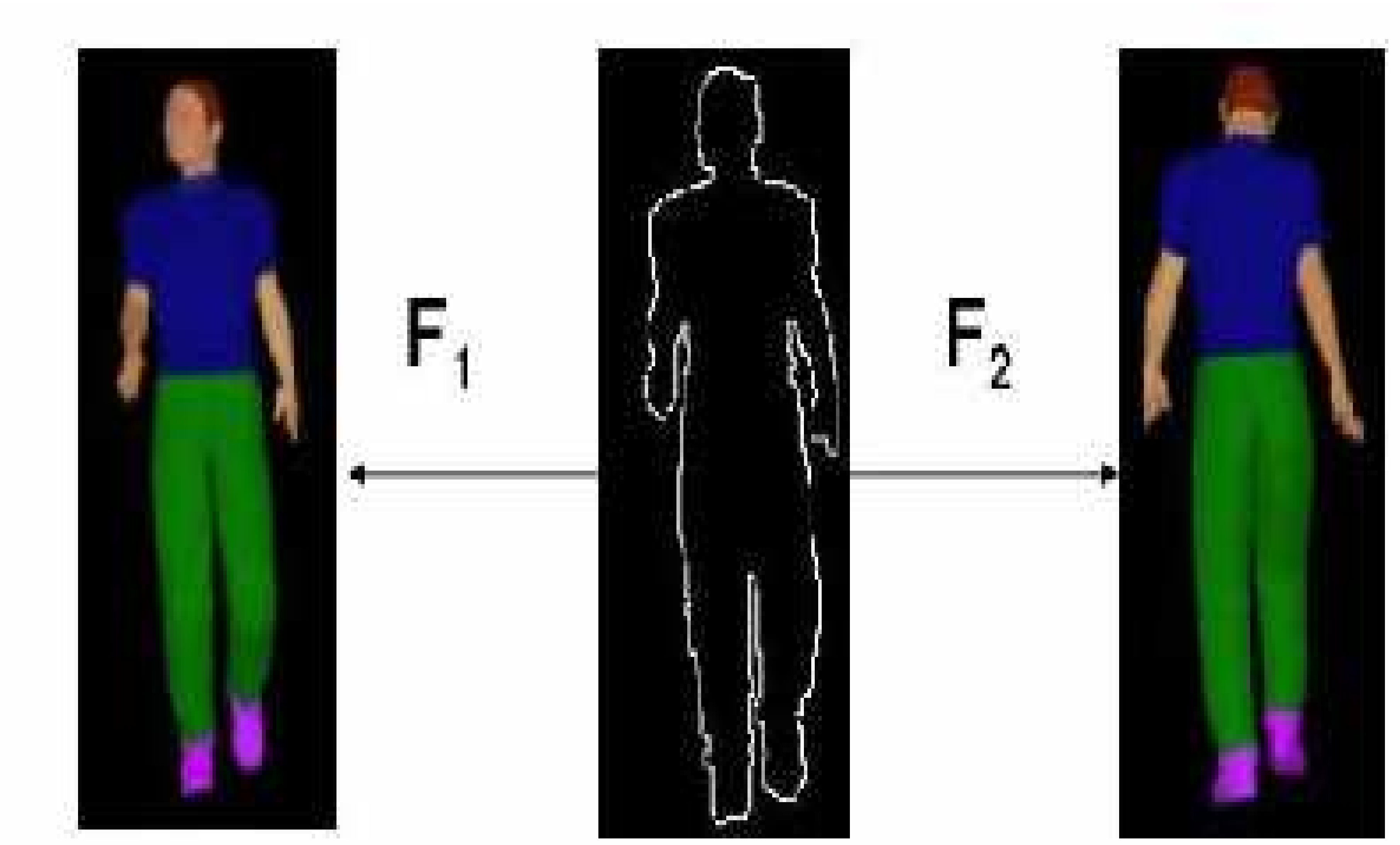




Neural net models E[y|x]

Need to model p(y|x)

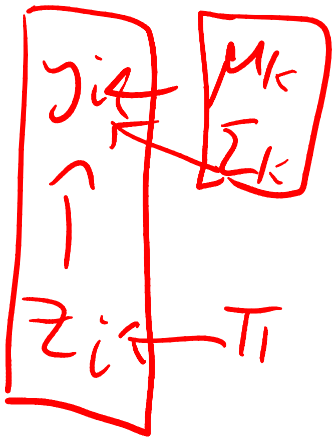
Ambiguity in inferring 3d from 2d

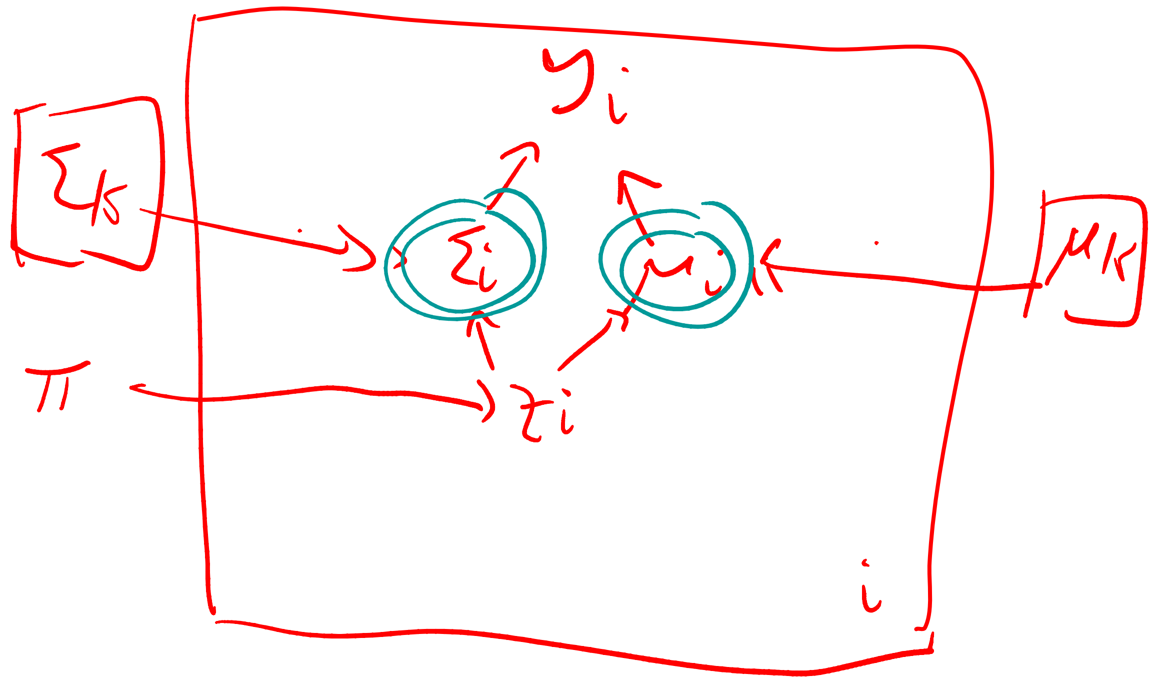


Sminchisescu

Mixture of gaussians

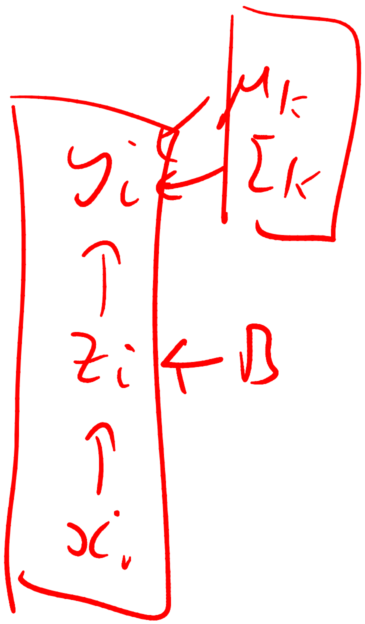
Deterministic nodes in green double circles





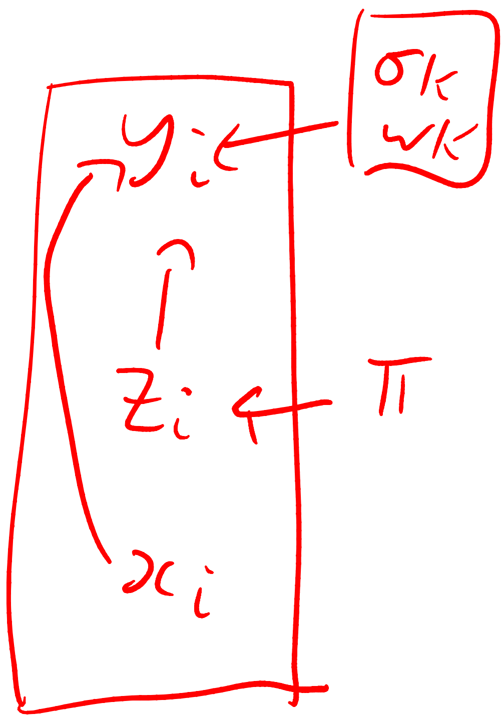
$$
\begin{aligned}
p(\mathbf{y}_i, z_i = k|\boldsymbol{\theta}) &= p(z_i = k|\boldsymbol{\theta})(\mathbf{y}_i|z_i = k, \boldsymbol{\theta}) \\
&= \mathrm{Mu}(z_i = k|\boldsymbol{\pi}, 1)\mathcal{N}(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned}
$$

Conditional mixture of gaussians



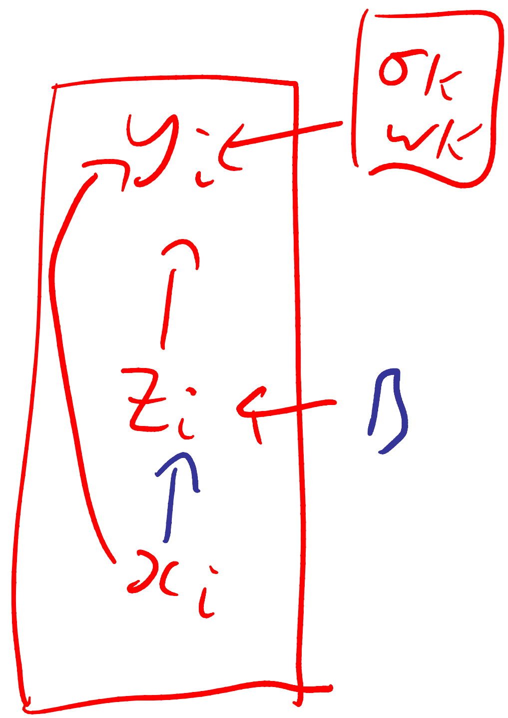
$$
\begin{aligned}
p(\mathbf{y}_i, z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) &= p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})p(\mathbf{y}_i|z_i = k, \boldsymbol{\theta}) \\
&= \mathrm{Mu}(z_i = k|\mathcal{S}(\mathbf{x}_i, \mathbf{B}), 1)\mathcal{N}(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned}
$$

mixture of linear regression



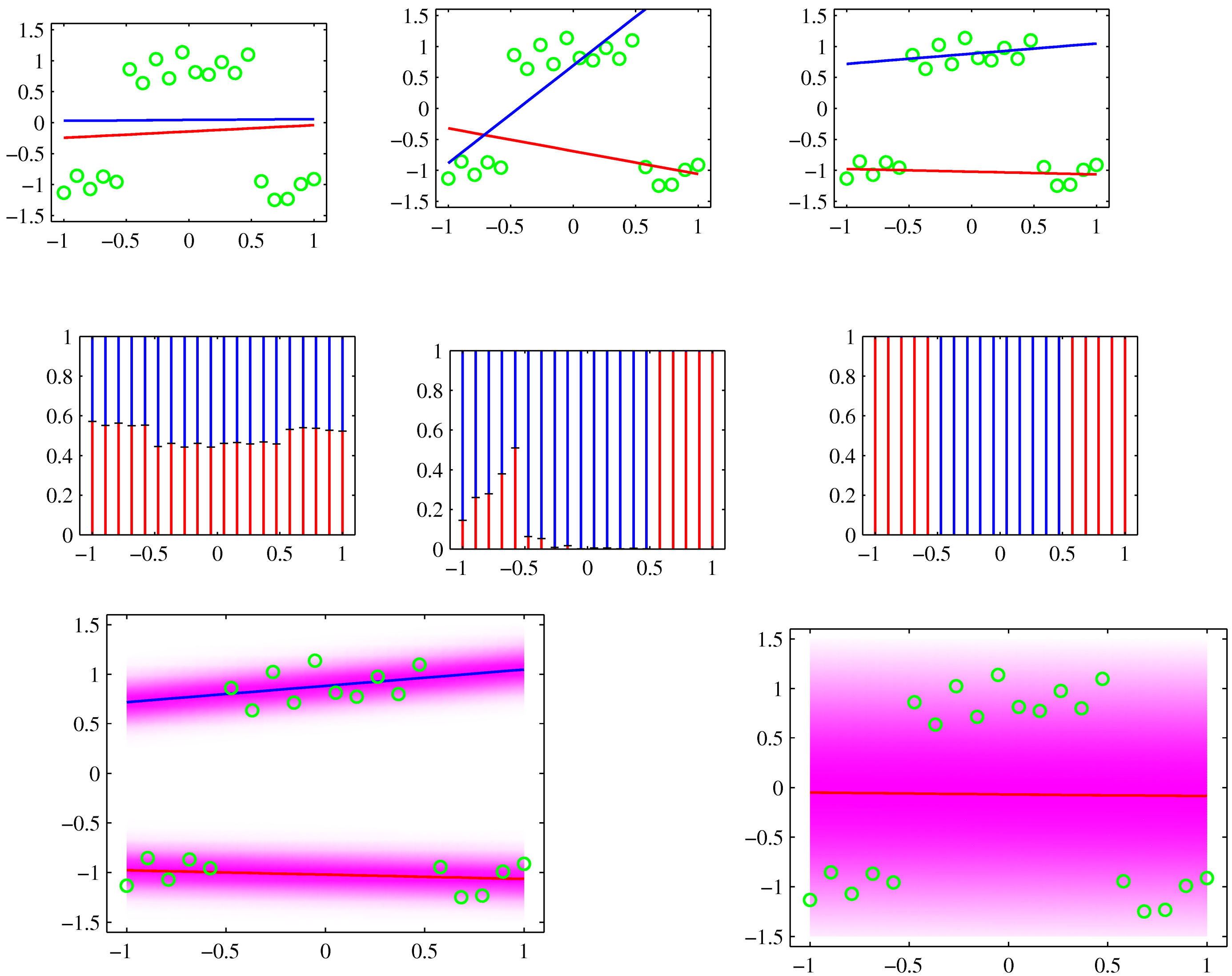
$$\begin{aligned} p(y_i, z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) &= p(z_i = k|\boldsymbol{\theta}) p(y_i|\mathbf{x}_i, z_i = k, \boldsymbol{\theta}) \\ &= \mathrm{Mu}(z_i = k|\boldsymbol{\pi}, 1)\mathcal{N}(y_i|\mathbf{x}_i^T \mathbf{w}_k, \sigma_k^2) \end{aligned}$$

Conditional mixture of linear regression



$$
\begin{aligned}
p(y_i, z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) &= p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta})p(y_i|\mathbf{x}_i, z_i = k, \boldsymbol{\theta}) \\
&= \mathbf{Mu}(z_i = k|\mathcal{S}(\mathbf{x}_i, \mathbf{B}), 1)\mathcal{N}(y_i|\mathbf{x}_i^T\mathbf{w}_k, \sigma_k^2)
\end{aligned}
$$

Mixtures of linear regression

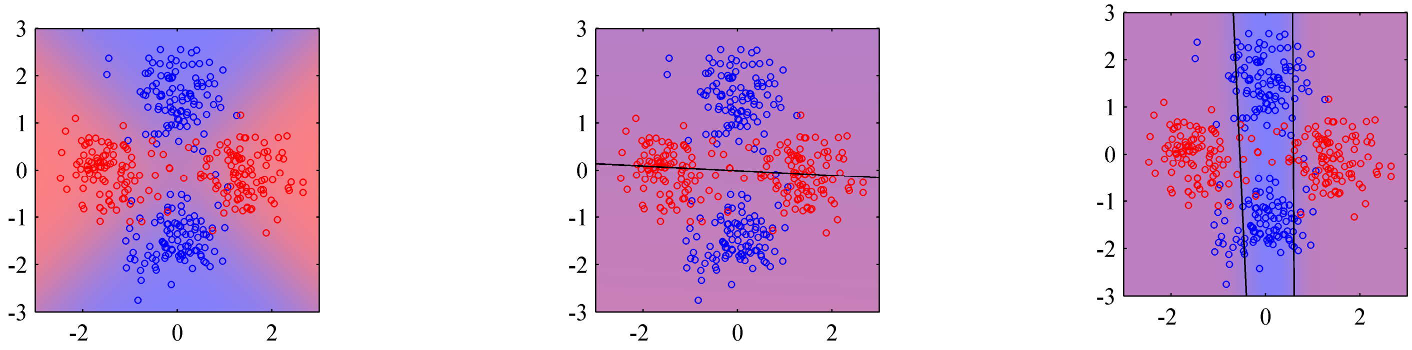


LL=-3

LL=-27.6

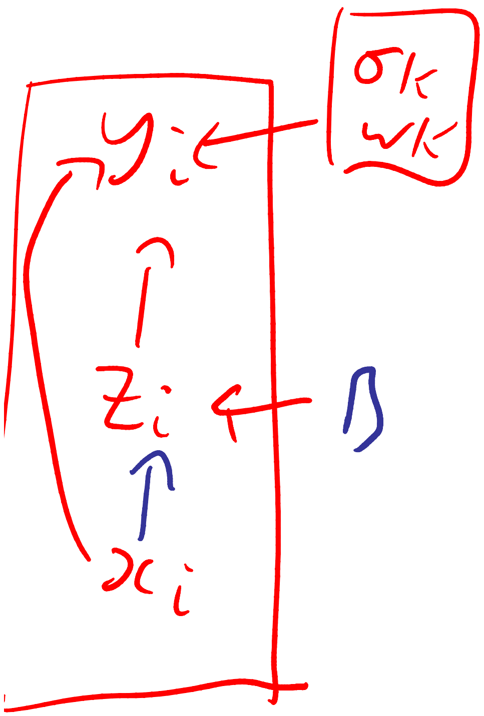
Bishop

Mixtures of logistic regression



EM for CondMixLinReg

- Expected complete data log likelihood

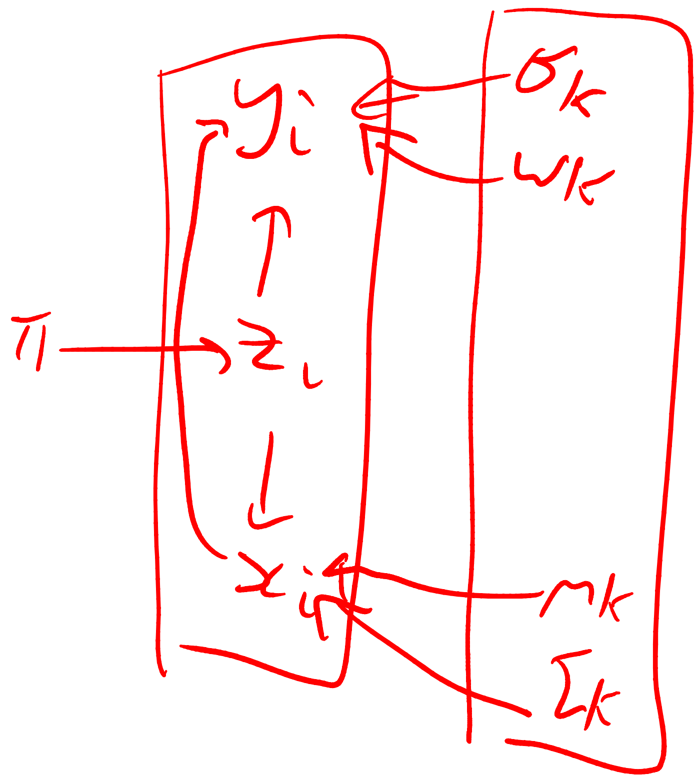


$$
\begin{aligned}
p(y_i, z_i|\mathbf{x}_i, \boldsymbol{\theta}) &= \prod_{k=1}^{K} p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) p(y_i|\mathbf{x}_i, z_i = k, \boldsymbol{\theta})^{I(z_i=k)} \\
\ell_c(\mathbf{y}, \mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} I(z_i = k) \log \mathcal{S}(k|\mathbf{x}_i, \mathbf{B}) \\
&\quad + I(z_i = k) \log \mathcal{N}(y_i|\mathbf{x}_i^T \mathbf{w}_k, \sigma_k^2) \\
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) &= E_{\mathbf{z}|\boldsymbol{\theta}^t} \ell_c(\mathbf{y}, \mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} p(z_i = k|\mathbf{x}_i, y_i, \boldsymbol{\theta}^t) \log \mathcal{S}(k|\mathbf{x}_i, \mathbf{B}) \\
&\quad + p(z_i = k|\mathbf{x}_i, y_i, \boldsymbol{\theta}^t) \log \mathcal{N}(y_i|\mathbf{x}_i^T \mathbf{w}_k, \sigma_k^2)
\end{aligned}
$$

E step: compute responsibilities $p(z_i = k|\mathbf{x}_i, y_i, \boldsymbol{\theta}^t)$

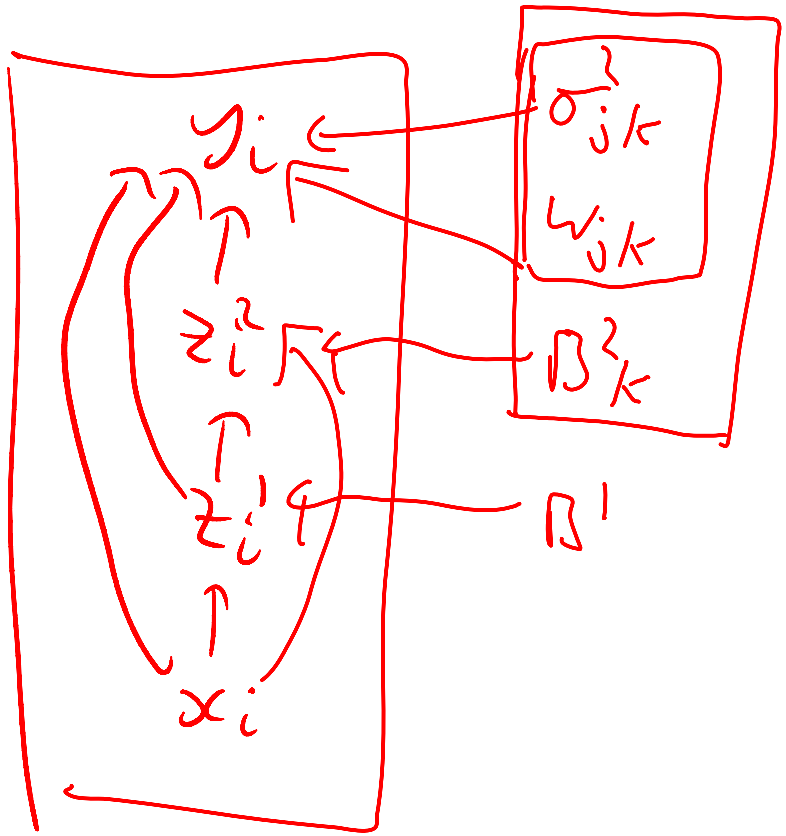
M step: weighted IRLS for B, weighted LS for w, residual for σ

Cluster weighted regression



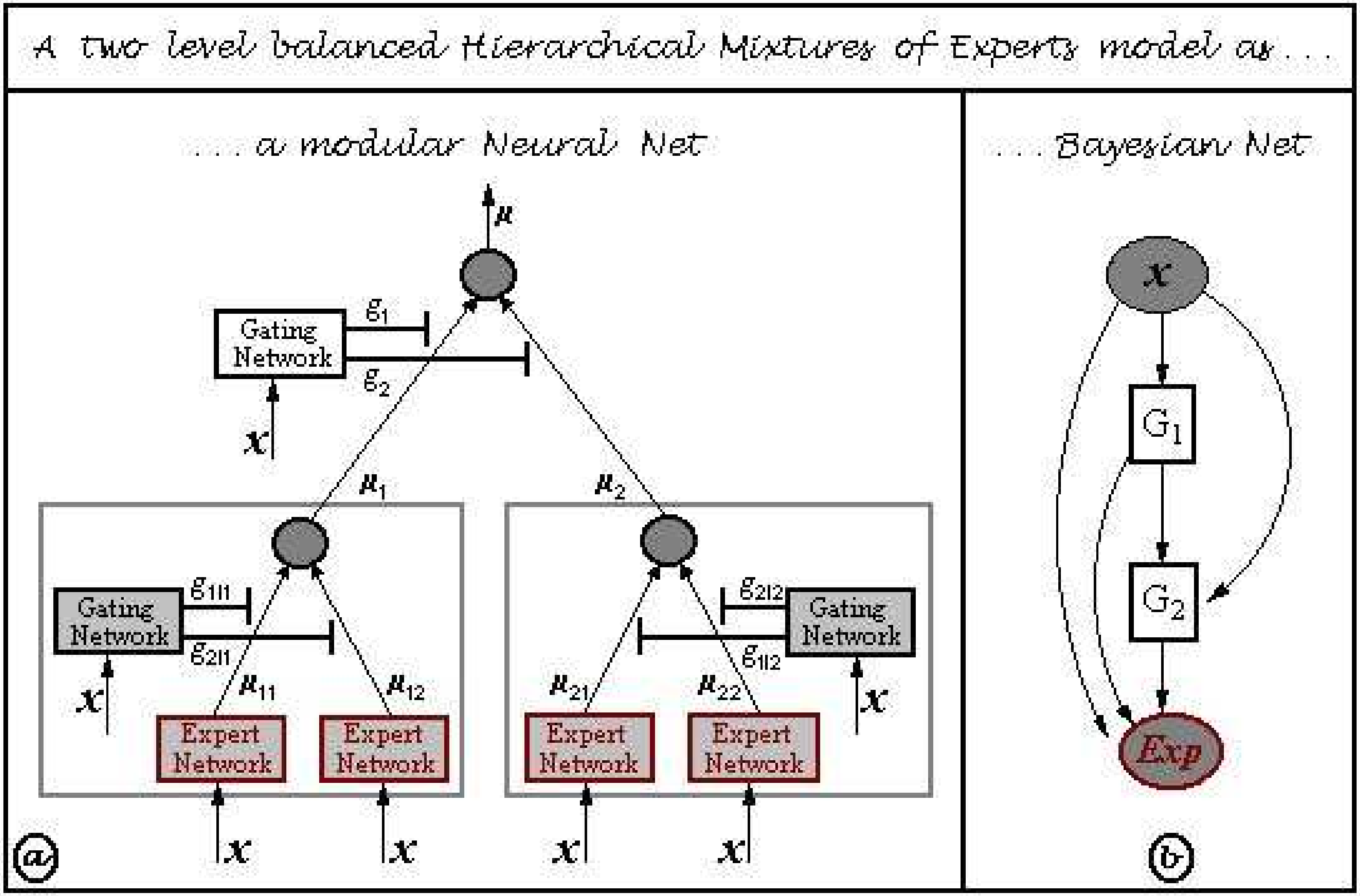
$$
\begin{aligned}
p(y_i, \mathbf{x}_i, z_i = k|\boldsymbol{\theta}) &= p(z_i = k|\boldsymbol{\theta})p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})p(y_i|\mathbf{x}_i, z_i = k, \boldsymbol{\theta}) \\
&= Mu(z_i = k|\boldsymbol{\pi}, 1)\mathcal{N}(\mathbf{x}_i|\mathbf{m}_k, \boldsymbol{\Sigma}_k)\mathcal{N}(y_i|\mathbf{x}_i^T\mathbf{w}_k, \sigma_k^2)
\end{aligned}
$$

Hierarchical mixture of experts



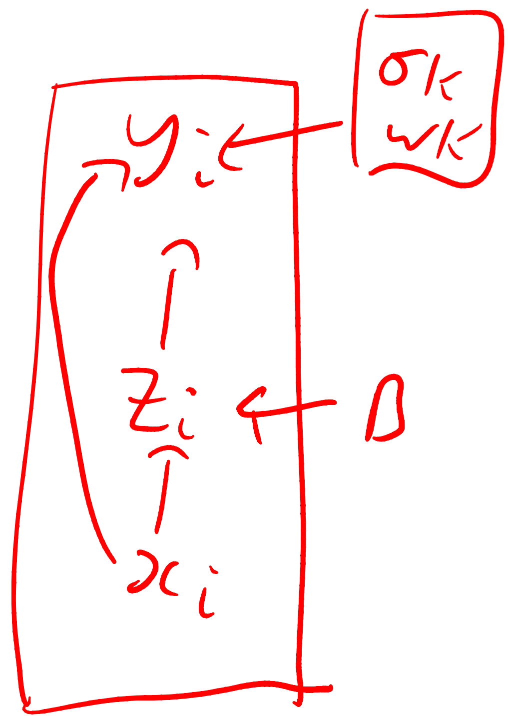
Probabilistic regression tree of fixed depth

Hierarchical mixtures of experts



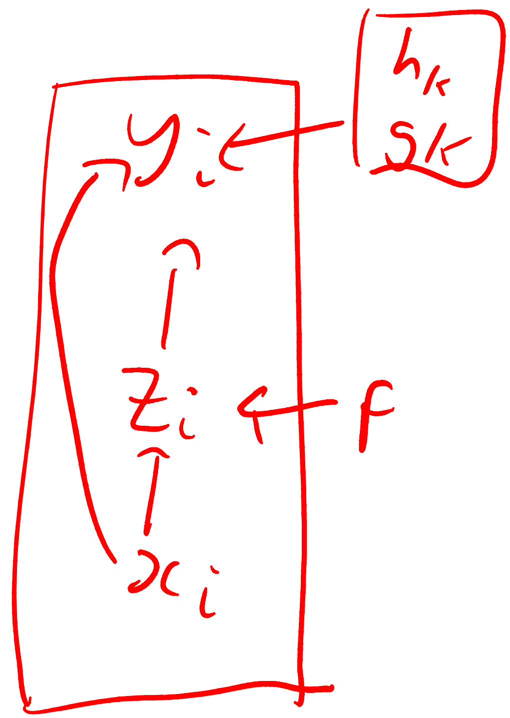
Brutti

CondMixLinReg



$$
\begin{aligned}
p(y_i, z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) &= p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) p(y_i|\mathbf{x}_i, z_i = k, \boldsymbol{\theta}) \\
&= \mathbf{Mu}(z_i = k|\mathcal{S}(\mathbf{x}_i, \mathbf{B}), 1) \mathcal{N}(y_i|\mathbf{x}_i^T \mathbf{w}_k, \sigma_k^2)
\end{aligned}
$$

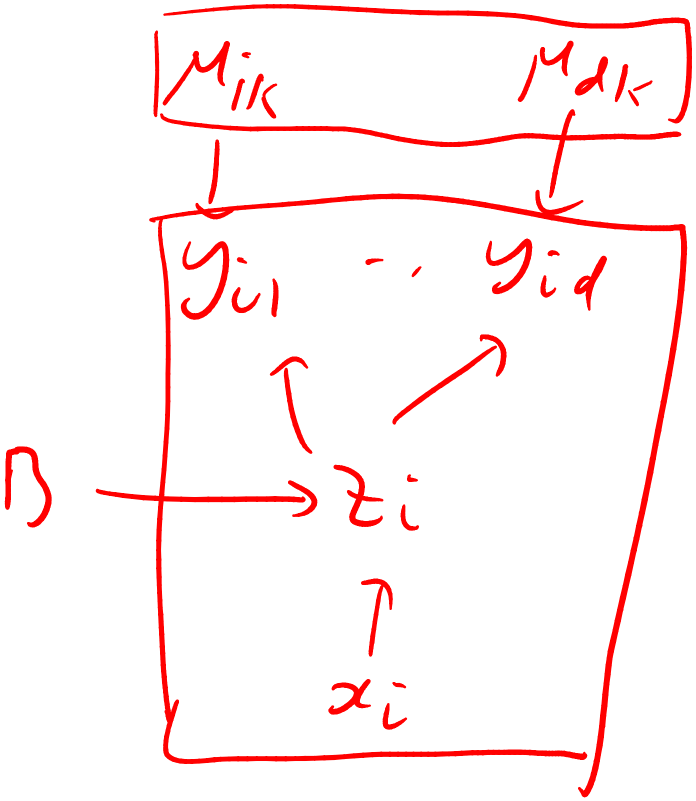
Mixture density networks



$$
\begin{aligned}
p(y_i, z_i = k | \mathbf{x}_i, \boldsymbol{\theta}) &= p(z_i = k | \mathbf{x}_i, \boldsymbol{\theta}) p(y_i | \mathbf{x}_i, z_i = k, \boldsymbol{\theta}) \\
&= \mathbf{Mu}(z_i = k | f(\mathbf{x}_i), 1) \mathcal{N}(y_i | g_k(\mathbf{x}_i), \exp(h_k(\mathbf{x}_i)))
\end{aligned}
$$

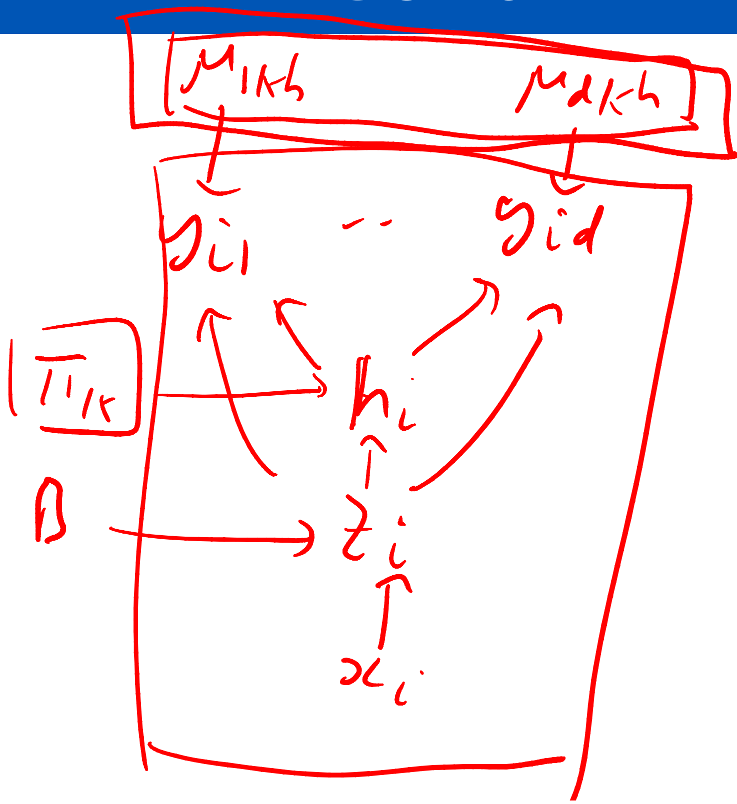
Have to use gradient descent or generalized EM

CondMixBernoulli



$$
\begin{aligned}
p(\mathbf{y}_i, z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) &= p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) p(\mathbf{y}_i|z_i = k, \boldsymbol{\theta}) \\
&= \mathrm{Mu}(z_i = k|\mathcal{S}(\mathbf{x}_i, \mathbf{B}), 1) \prod_{j=1}^{d} \mathrm{Ber}(y_{i,j}|\mu_{j,k})
\end{aligned}
$$

CondMixBernoulliMix



$$p(\mathbf{y}_i | z_i = k, \boldsymbol{\theta}) = \sum_{h=1}^{H} p(h_i = h | \boldsymbol{\theta}) \prod_{j=1}^{d} \mathrm{Ber}(y_{i,j} | \mu_{j,h,k})$$

$$\begin{aligned} p(\mathbf{y}_i, z_i = k, h_i = h | \mathbf{x}_i, \boldsymbol{\theta}) &= p(z_i = k | \mathbf{x}_i, \boldsymbol{\theta}) p(h_i = h | z_i = k, \boldsymbol{\theta}) p(\mathbf{y}_i | h_i = h, z_i = k, \boldsymbol{\theta}) \\ &= \mathrm{Mu}(z_i = k | \mathcal{S}(\mathbf{x}_i, \mathbf{B}), 1) \mathrm{Mu}(h_i = h | \pi_k, 1) \prod_{j=1}^{d} \mathrm{Ber}(y_{i,j} | \mu_{j,h,k}) \end{aligned}$$

Outline

- Conditional mixture models
- EM for Empirical Bayes
- “Sparse Bayesian learning”
- EM theory

Empirical Bayes

Method	Definition
Maximum likelihood	$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D}\vert\theta)$
MAP estimation	$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D}\vert\theta)p(\theta\vert\alpha)$
Empirical Bayes	$\hat{\alpha} = \arg\max_{\alpha} p(\mathcal{D}\vert\alpha) = \arg\max_{\alpha} \int p(\mathcal{D}\vert\theta)p(\theta\vert\alpha)d\theta$

EB = Type II maximum likelihood
= evidence approximation

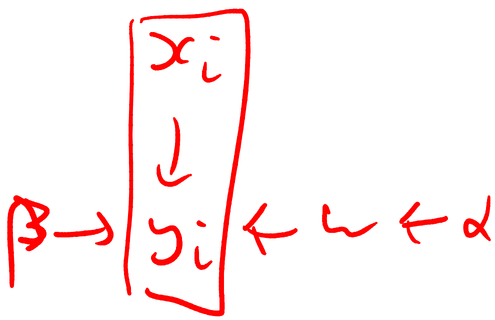
EM for EB

$$
\begin{aligned}
\text{E step} &= p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\theta}) \\
\text{M step} &= \max_{\boldsymbol{\alpha}} E \log p(\mathcal{D}, \boldsymbol{\theta}|\boldsymbol{\alpha})
\end{aligned}
$$

Outline

- Conditional mixture models
- EM for Empirical Bayes
- “Sparse Bayesian learning”
- EM theory

Automatic Relevancy Determination (ARD)

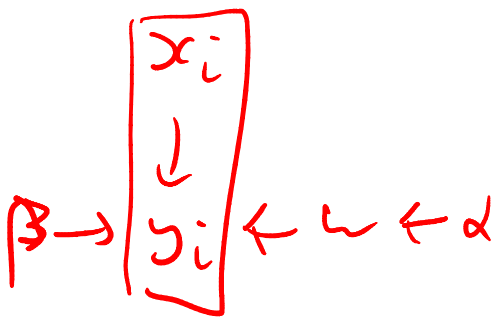


$$
\begin{aligned}
p(y_i|\mathbf{x}_i, \mathbf{w}, \beta) &= \mathcal{N}(y_i|\mathbf{x}_i^T \mathbf{x}, \beta^{-1}) \\
p(\mathbf{w}|\boldsymbol{\alpha}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})
\end{aligned}
$$

Expected complete data log likelihood

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= E \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \alpha, \beta) \\
&= \frac{d}{2} \log \frac{\alpha}{2\pi} - \frac{\alpha}{2} E[\mathbf{w}^T \mathbf{w}] + \frac{n}{2} \log \frac{\beta}{2\pi} - \frac{\beta}{2} \sum_{i=1}^{n} E[(y_i - \mathbf{w}^T \mathbf{x})i)^2]
\end{aligned}
$$

EM for ARD



$$\begin{aligned} p(y_i|\mathbf{x}_i, \mathbf{w}, \beta) &= \mathcal{N}(y_i|\mathbf{x}_i^T\mathbf{x}, \beta^{-1}) \\ p(\mathbf{w}|\boldsymbol{\alpha}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \end{aligned}$$

E step

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \alpha, \beta) &\propto \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}_d)\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}_n) \\ &= \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}) \\ \mathbf{S} &= \alpha\mathbf{I}_d + \beta\mathbf{X}^T\mathbf{X} \\ \mathbf{m} &= \beta\mathbf{S}\mathbf{X}^T\mathbf{y} \end{aligned}$$

M step

$$\frac{\partial}{\partial\alpha}J(\boldsymbol{\theta}) = 0 \quad \Rightarrow \quad \alpha = \frac{d}{E[\mathbf{w}^T\mathbf{w}]} = \frac{d}{\mathbf{m}^T\mathbf{m} + \mathrm{trace}(\mathbf{S})} = \frac{d}{\sum_{j=1}^d m_j^2 + S_{jj}}$$

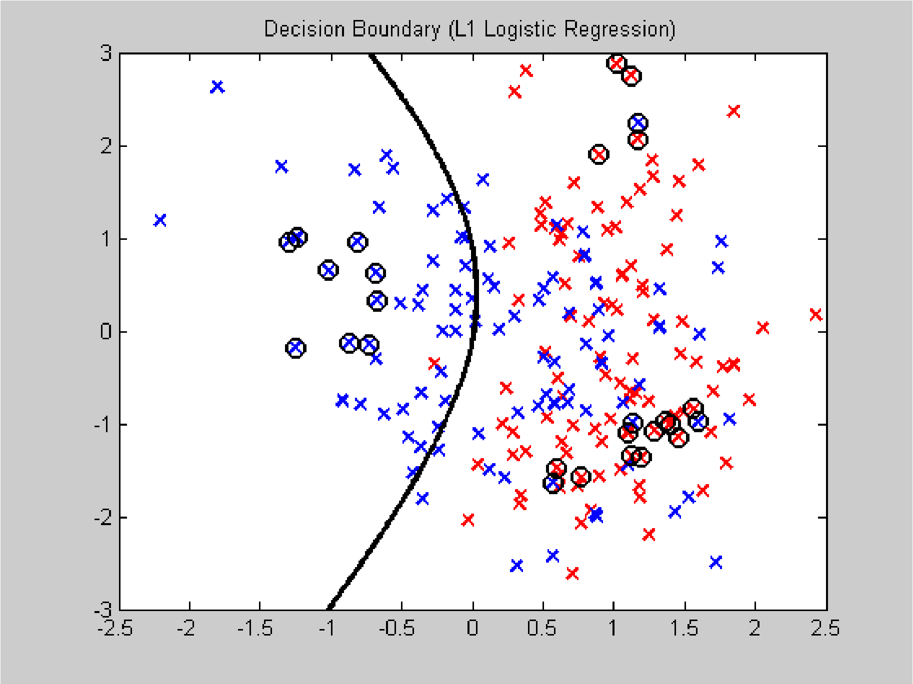
Relevance vector machines (RVMs)

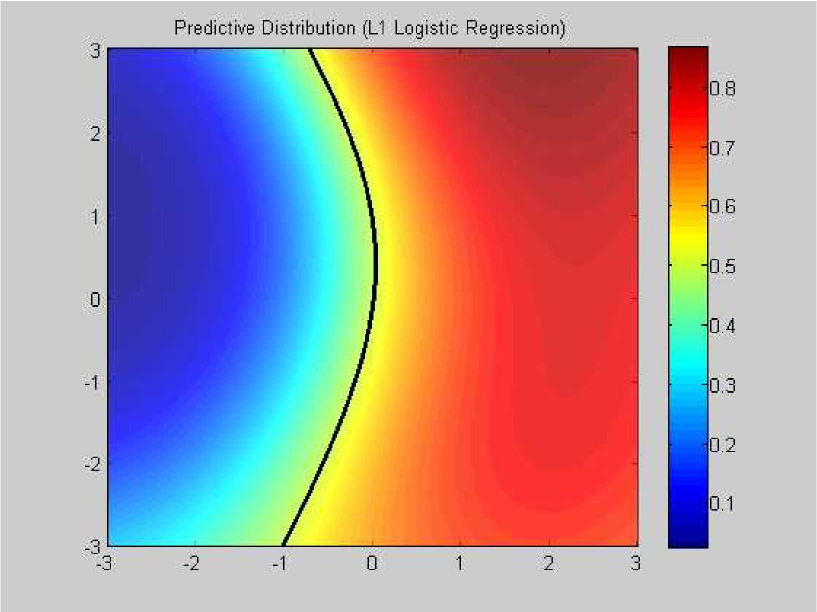
- Perform a kernel expansion of the input data
  eg using RBFs

$$\phi_i(\mathbf{x}_i) \quad = \quad [K(\mathbf{x}_i, \mathbf{x}_1), \ldots, K(\mathbf{x}_i, \mathbf{x}_n)]$$

- Then apply ARD to select a subset of the input features

L1 penalized logreg with RBF expansion





Outline

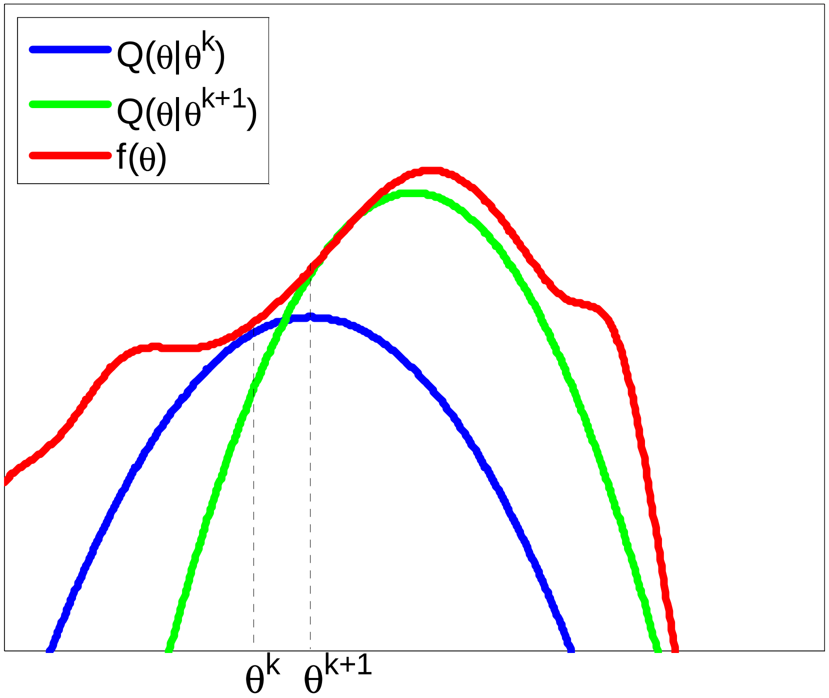
- Conditional mixture models
- EM for Empirical Bayes
- “Sparse Bayesian learning”
- EM theory

Bound optimization algorithms

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k)$$

Key condition: $Q(\theta_k|\theta_k)$ touches $f(\theta_k)$
So pushing up on Q will actually push up on f

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k) \quad = \quad f(\boldsymbol{\theta}_k) - Q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k) \leq f(\boldsymbol{\theta}_{k+1}) - Q(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k)$$



MM algorithm

- In general, if Q is a lower bound on f that satisfies the key condition, we say Q minorizes f.
- The algorithm is called the minorize-maximize (MM) algorithm.
- We can also create majorize-minimize algorithms.

MM monotonically increases objective

$$\begin{aligned} f(\boldsymbol{\theta}_{k+1}) &= f(\boldsymbol{\theta}_{k+1}) - Q(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k) + Q(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k) && (1) \\ &\geq f(\boldsymbol{\theta}_k) - Q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k) + Q(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k) && (2) \end{aligned}$$

which follows from the key condition

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k) = f(\boldsymbol{\theta}_k) - Q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k) \leq f(\boldsymbol{\theta}_{k+1}) - Q(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k) \qquad (3)$$

Also, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_k)$ is maximized when $\boldsymbol{\theta} = \boldsymbol{\theta}_{k+1}$, by definition, so

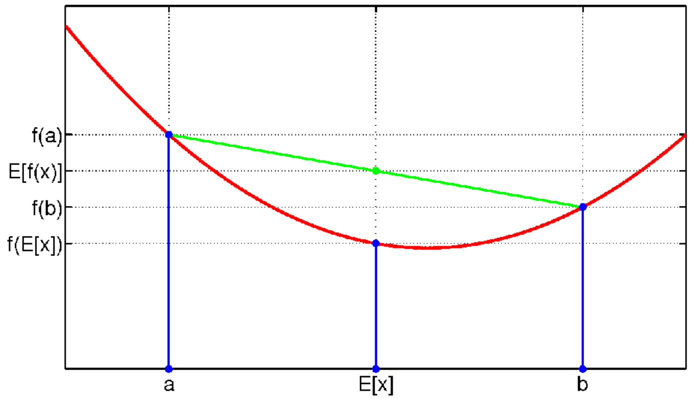
$$\begin{aligned} f(\boldsymbol{\theta}_{k+1}) &\geq f(\boldsymbol{\theta}_k) - Q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k) + Q(\boldsymbol{\theta}_{k+1}|\boldsymbol{\theta}_k) && (4) \\ &\geq f(\boldsymbol{\theta}_k) - Q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k) + Q(\boldsymbol{\theta}_k|\boldsymbol{\theta}_k) && (5) \\ &= f(\boldsymbol{\theta}_k) && (6) \end{aligned}$$

EM is an MM algorithm

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log[\sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})]$$

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \frac{p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})}{q_i(\mathbf{z}_i)}$$

$$\sum_i \log \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \frac{p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})}{q_i(\mathbf{z}_i)} \geq \sum_i \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log \frac{p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})}{q_i(\mathbf{z}_i)}$$



Jensen's inequality

Lower bound on log lik!
What q value?

What q function?

$$
\begin{aligned}
\boldsymbol{\theta}_{k+1} &= \arg\max_{\boldsymbol{\theta}} \sum_i \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log \frac{p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta})}{q_i(\mathbf{z}_i)} \\
L(q_i, \boldsymbol{\theta}) &= \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log \frac{p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta})}{q_i(\mathbf{z}_i)} \\
&= \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log \frac{p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\theta}) p(\mathbf{x}_i | \boldsymbol{\theta})}{q_i(\mathbf{z}_i)} \\
&= \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log \frac{p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\theta})}{q_i(\mathbf{z}_i)} - \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log p(\mathbf{x}_i | \boldsymbol{\theta}) \\
&= KL(q_i(\mathbf{z}_i) || p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\theta})) - \log p(\mathbf{x}_i | \boldsymbol{\theta})
\end{aligned}
$$

To make bound tight, set $q_i(z_i) = p(z_i|x_i,\theta)$

$$
Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) \stackrel{\text{def}}{=} \sum_i \sum_{\mathbf{z}_i} p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\theta}_k) \log p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta})
$$

EM

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k) \quad \stackrel{\text{def}}{=} \quad \sum_i \sum_{\mathbf{z}_i} p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}_k) \log p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) \qquad ($$

which we recognize as the expected complete data log likelihood.

- E step: compute $p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}_k)$
- M step: compute $\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_k)$

Variational EM: set $q_i(z_i)$ to be an approximate $p(z_i|x_i,\theta)$

Outline

- EM theory
- Conditional mixture models
- Empirical Bayes for linear regression