Hierarchical Mixture of Experts

Presented by Yongqia Shao and Juan Wu

Academic English for Statistics

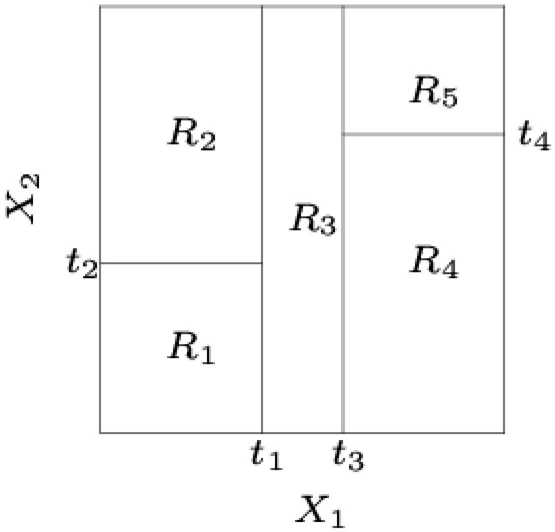
12/09/2014

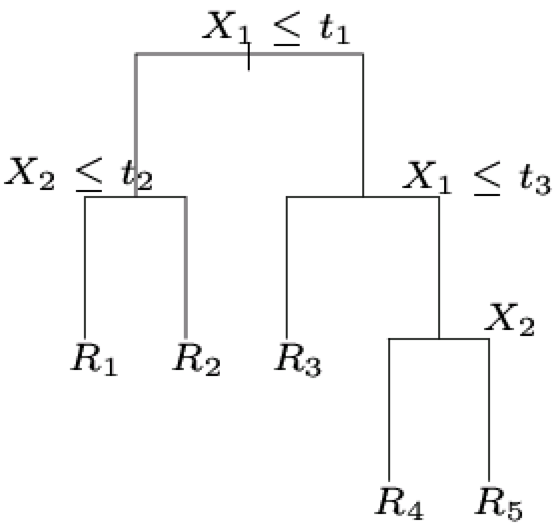
Outline

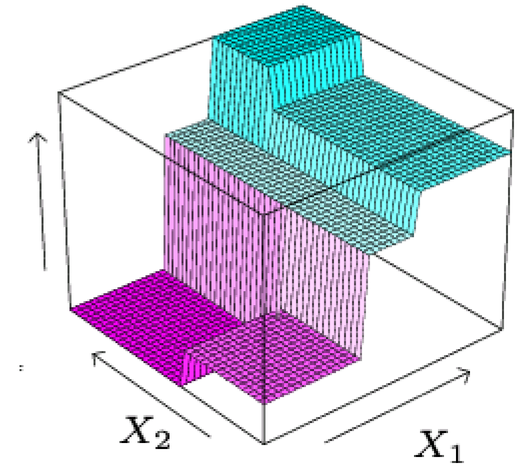
- Introduction
- Hierarchical mixture of experts
- E-M algorithm
- Experimental results

TREE

- An example

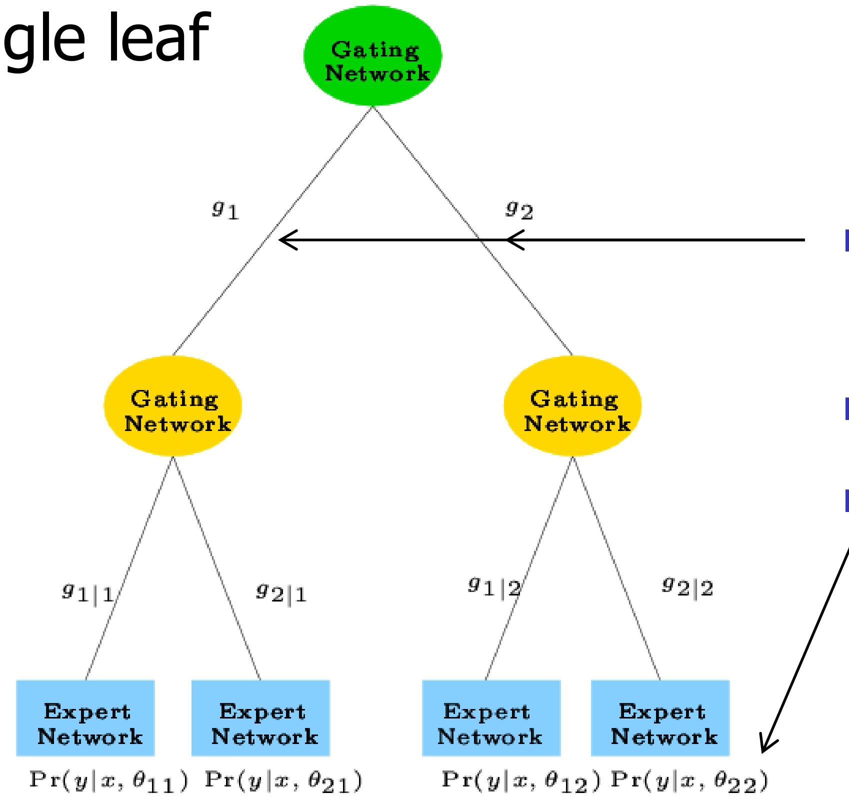






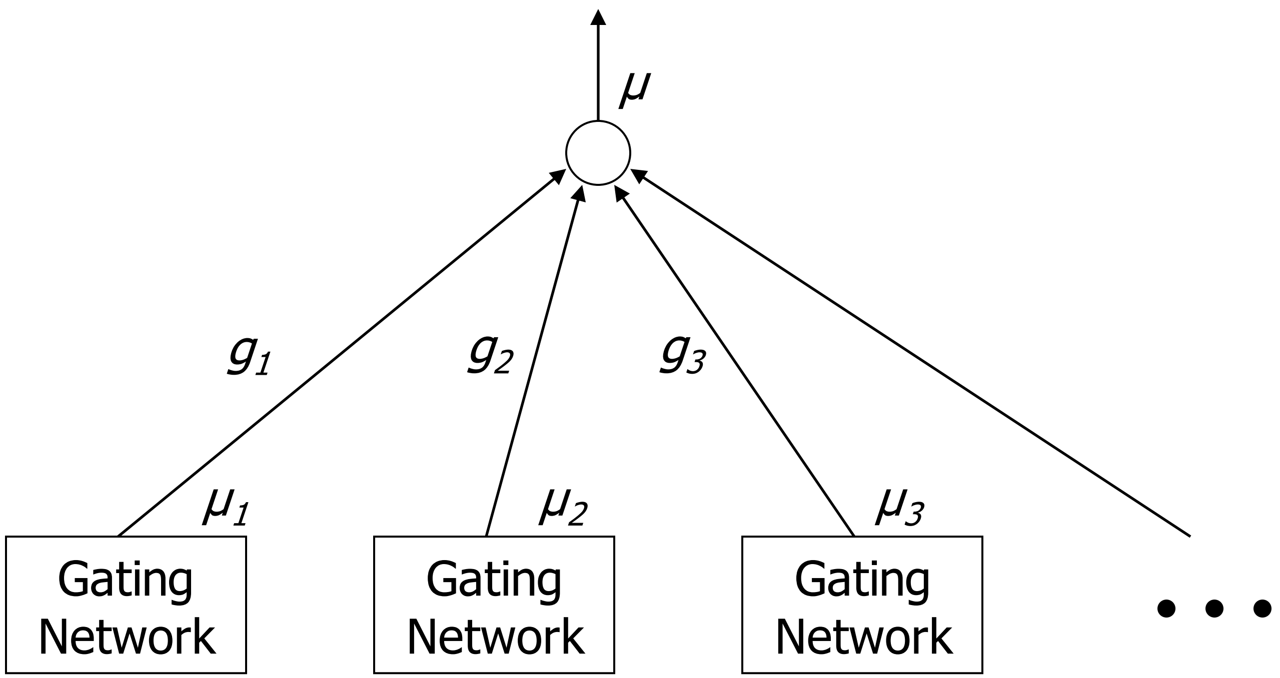
Hierarchical Mixture of Experts

Soft decision tree: Takes a weighted (gating) average of all leaves (experts), as opposed to using a single path and a single leaf



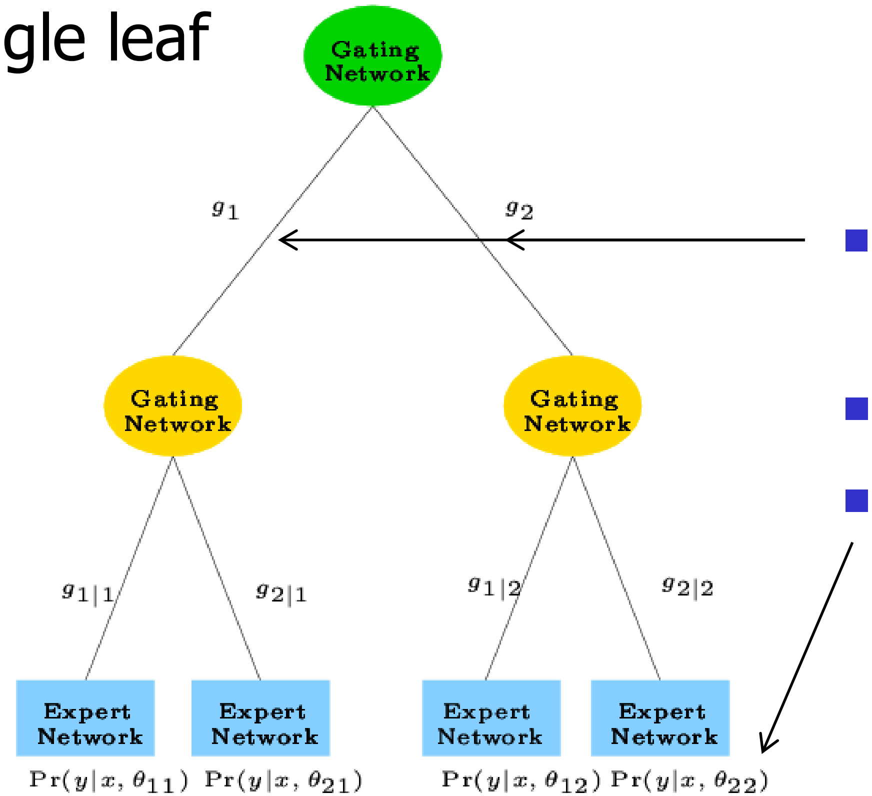
- "soft" partition: tree splits are probabilistic
- Splits can be multiway
- A linear model is fit in each terminal node

Hierarchical Mixture of Experts



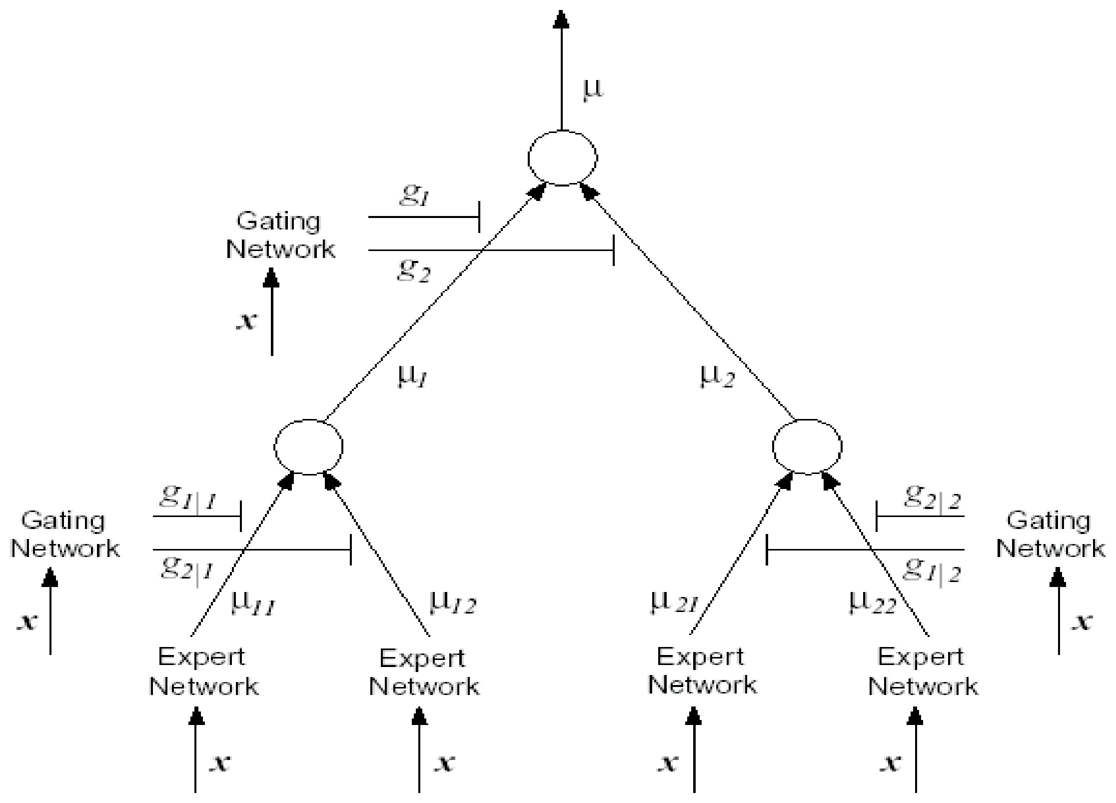
Hierarchical Mixture of Experts

Soft decision tree: Takes a weighted (gating) average of all leaves (experts), as opposed to using a single path and a single leaf



- "soft" partition: tree splits are probabilistic
- Splits can be multiway
- A linear model is fit in each terminal node

Hierarchical Mixture of Experts



Expert Network Output

- At the leaves of trees

for each expert:

$$\mu_{ij} = f(U_{ij}x)$$

output of the expert

Model for response variable

Expert Network Output

- For each expert, assume the true output *y* is chosen from a distribution *P* with parameters θ_{ij}

$$Y \sim P(y \mid x, \theta_{ij})$$

Regression: The Gaussian linear regression model is used:

$$Y = \beta_{ij}^T x + \varepsilon, \varepsilon \sim N(0, \sigma_{ij}^2)$$

Classification: The linear logistic regression model is used:

$$P\left(Y = 1 \mid x, \theta_{ij}\right) = \frac{1}{1 + e^{-\theta_{ij}^T x}}$$

Gating network output

- At the nonterminal of the tree

top level:

$$\xi_i = v_i^T x$$

$$g_i = \frac{\exp(\xi_i)}{\sum_k \exp(\xi_k)}$$

$$\sum_i g_i = 1$$

other level:

$$\xi_{ij} = v_{ij}^T x$$

$$g_{j|i} = \frac{\exp(\xi_{ij})}{\sum_k \exp(\xi_{ik})}$$

$$\sum_j g_{j|i} = 1$$

Gating Network Output

- At the non-leaves nodes

top node:

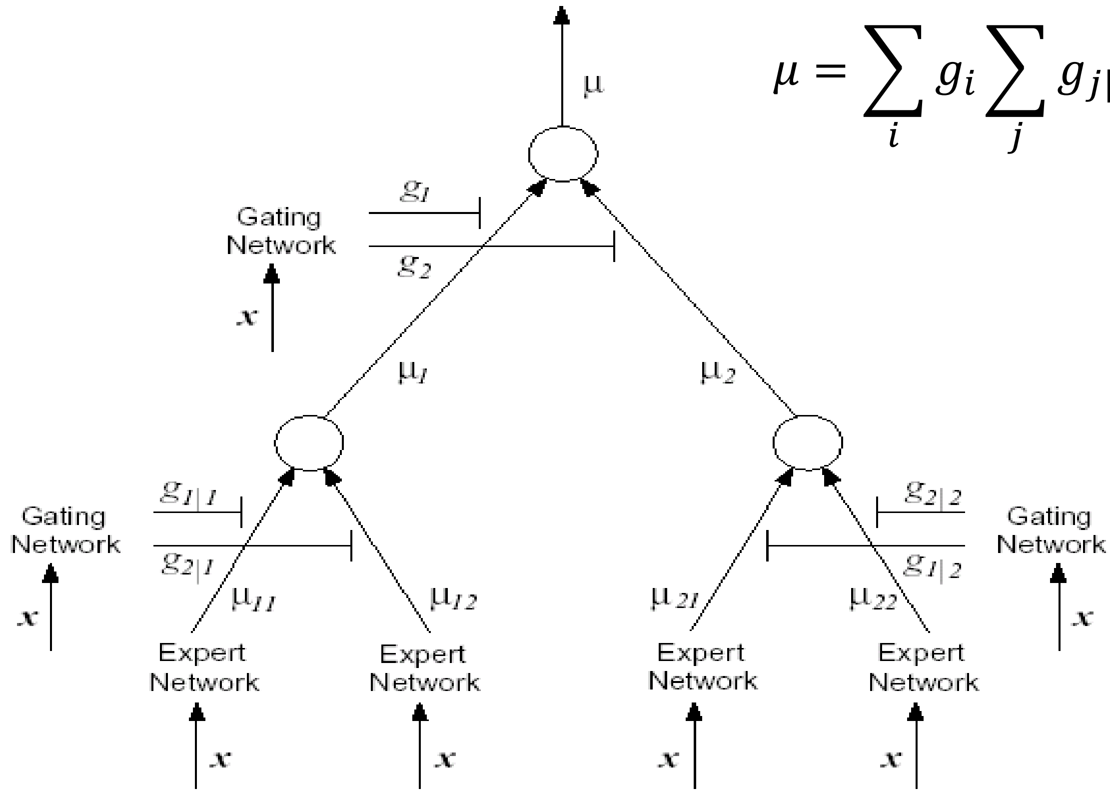
$$\mu = \sum_i g_i \mu_i$$

other nodes:

$$\mu_i = \sum_j g_{j|i} \mu_{ij}$$

Hierarchical Mixture of Experts

$$\mu = \sum_i g_i \sum_j g_{j|i} f(U_{ij} x)$$



Probability model

- Therefore, for data set $\mathrm{X} = \{x^{(t)}, y^{(t)}\}_1^N$, the total probability of generating *y* from *x* is given by

$$P(Y|X,\theta) = \prod_t \sum_i g_i^{(t)}(x, v_i) \sum_j g_{j|i}^{(t)}(x, v_{ij}) P^{(t)}(y|x, \theta_{ij})$$

$$\ln P(Y|X,\theta) = \sum_t \ln\left(\sum_i g_i^{(t)}(x, v_i) \sum_j g_{j|i}^{(t)}(x, v_{ij}) P^{(t)}(y|x, \theta_{ij}) \right)$$

E-M algorithm

- Introduce latent variables z_{ij} which have an interpretation as the labels that corresponds to the experts.
- The probability model can be simplified with the knowledge of latent variables

$$P(y^{(t)}, z_{ij}^{(t)} \middle| x^{(t)}, \theta) = g_i^{(t)} g_{j|i}^{(t)} P_{ij}(y^{(t)}) = \prod_i \prod_j \left\{ g_i^{(t)} g_{j|i}^{(t)} P_{ij}(y^{(t)}) \right\}^{z_{ij}^{(t)}}$$

E-M algorithm

- Log-likelihood function:

$$l_c(\theta; y) = \sum_t \sum_i \sum_j z_{ij}^{(t)} \left\{ \ln g_i^{(t)} + \ln g_{j|i}^{(t)} + \ln P_{ij}(y^{(t)}) \right\}$$

E-M algorithm

Define posterior probabilities … and we get …

$$h_{ij} = \frac{g_i g_{j|i} P_{ij}(\mathbf{y})}{\sum_i g_i \sum_j g_{j|i} P_{ij}(\mathbf{y})} \qquad E[z_{ij}^{(t)} | \mathcal{X}] = h_{ij}^{(t)}$$

$$h_i = \frac{g_i \sum_j g_{j|i} P_{ij}(\mathbf{y})}{\sum_i g_i \sum_j g_{j|i} P_{ij}(\mathbf{y})} \qquad E[z_i^{(t)} | \mathcal{X}] = h_i^{(t)}$$

$$h_{j|i} = \frac{g_{j|i} P_{ij}(\mathbf{y})}{\sum_j g_{j|i} P_{ij}(\mathbf{y})}. \qquad E[z_{j|i}^{(t)} | \mathcal{X}] = h_{j|i}^{(t)},$$

E-M algorithm

- The E-step

$$Q(\theta, \theta^{(p)}) = E_z(l_c(\theta; y)) = \sum_t \sum_i \sum_j \mathrm{h}_{ij}^{(t)} \left\{ \ln g_i^{(t)} + \ln g_{j|i}^{(t)} + \ln P_{ij}(y^{(t)}) \right\}$$

where we have used the fact that:

$$\begin{aligned} E[z_{ij}^{(t)}|\mathcal{X}] &= P(z_{ij}^{(t)} = 1|\mathbf{y}^{(t)}, \mathbf{x}^{(t)}, \boldsymbol{\theta}^{(p)}) \\ &= \frac{P(\mathbf{y}^{(t)}|z_{ij}^{(t)} = 1, \mathbf{x}^{(t)}, \boldsymbol{\theta}^{(p)})P(z_{ij}^{(t)} = 1|\mathbf{x}^{(t)}, \boldsymbol{\theta}^{(p)})}{P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \boldsymbol{\theta}^{(p)})} \\ &= \frac{P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \boldsymbol{\theta}_{ij}{}^{(p)})g_i^{(t)}g_{j|i}^{(t)}}{\sum_i g_i^{(t)} \sum_j g_{j|i}^{(t)} P(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \boldsymbol{\theta}_{ij}{}^{(p)})} \\ &= h_{ij}^{(t)}. \end{aligned}$$

E-M algorithm

- The M-step

$$\theta_{ij}^{p+1} = \arg\max_{\theta_{ij}} \sum_t \mathrm{h}_{ij}^{(t)} \ln P_{ij}(y^{(t)})$$

$$v_i^{p+1} = \arg\max_{v_i} \sum_t \sum_k h_k^{(t)} \ln g_k^{(t)}$$

$$v_{ij}^{p+1} = \arg\max_{v_{ij}} \sum_t \sum_k h_k^{(t)} \sum_l h_{l|k}^{(t)} \ln g_{l|k}^{(t)}$$

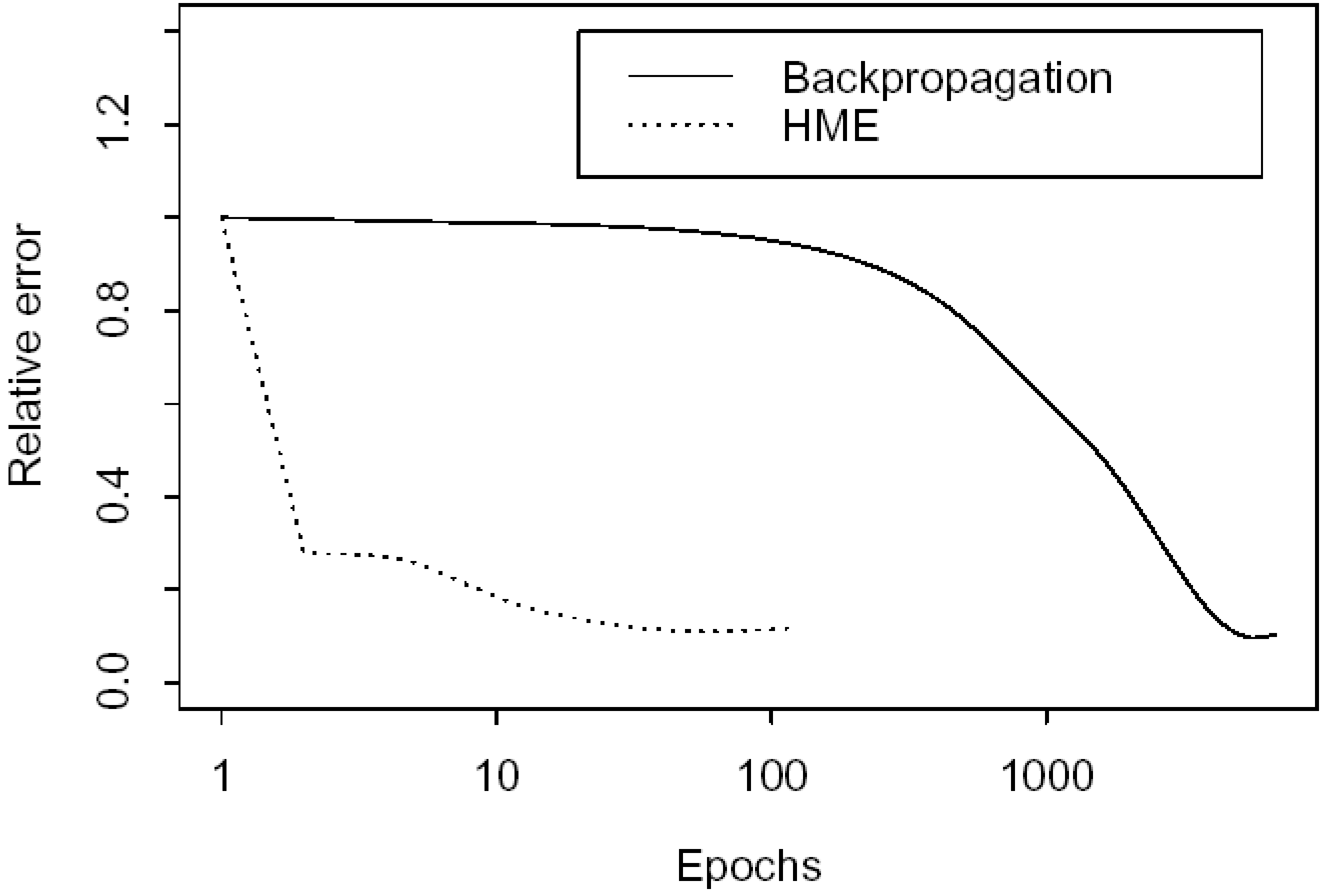
Results

- Simulated data of a four-joint robot arm moving in three-dimensional space

EM ←

Architecture	Relative Error	# Epochs
linear	.31	1
backprop	.09	5,500
HME (Algorithm 1)	.10	35
HME (Algorithm 2)	.12	39
CART	.17	NA
CART (linear)	.13	NA
MARS	.16	NA

Results



Backpropagation
HME
Relative error
0.0
0.4
0.8
1.2
1
10
100
1000
Epochs

Thank you

Reference: Michael.I.Jordan, Hierarchical mixtures of experts and the EM algorithm, Neural Computation, 1994

Hierarchical Mixture of Experts