

# Econometric Applications of Hierarchical Mixture of Experts

Lucas C. Dowiak

October 27, 2018

Department of Economics, City University of New York, Graduate Center,  
New York, NY, 10016, *Email: ldowiak@gc.cuny.edu*

**Abstract:**

**Keywords:** Hierarchical mixture of experts, expectation maximization

**JEL Classification:**

## 1 Introduction

The concepts of mixture models and mixture distributions are old hat in the economics business. Hamilton 1989 and Goldfeld and Quandt 1973 are a few of the pioneering works for time series and cross sectional regression, respectively. We are also knee deep in the age of machine learning, and it's reigning champion, the artificial neural network, has been successfully adapted and studied in the context of applied econometrics. This article adds to the small body of literature that employs a specific neural network architecture to model the weights of a mixture model. In doing so, we leverage the highly flexible nature of a neural network but maintain interpretability and the means to quantify marginal effects. The model under investigation is called the Hierarchical Mixture of Experts (HME), a class of mixture models whose defining feature is its conditional weighting scheme. The model's origin story traces back to

R. A. Jacobs et al. 1991. The authors use a single multinomial classifier to assign, in a probabilistic sense, input patterns to local *experts*. These *experts* are almost always some flavor of regression or classification model. The multinomial structure that assigns inputs to experts is referred to as the *gating network*. R. A. Jacobs et al. 1991 employ this mixture of experts (ME) framework to model vowel discrimination in a speech recognition context. Extending this approach, Jordan and R. Jacobs 1993 propose a gating network that allows for additional layers of multinomial partitioning of the input space, occurring in a recursive manner. The result of this extension is a gating network that takes on a tree-like structure, stemming from an initial multinomial split and filtering down through additional multinomial partitions of the input space. The hierarchical nature of this gating network is what gives this class of model its name: the Hierarchical Mixture of Experts (HME). HME models nest ME models as special case. Pushing a little further, one additional case is studied as well. As the depth of an HME grows, so too must the number of experts. If we have a symmetric HME network, this growth is geometric with respect to the network's depth. With this in mind, we propose a model where each expert is not unique, but a member of a fixed set of experts that are allowed to repeat at different terminal nodes of the network. We refer to this additional model as a Hierarchical Mixture of Repeated Experts (HMRE). Figure (1) provides an example of each of these models studied in this article.

This article investigates the adoption of ME, HME, and HMRE models to an applied econometric framework, with particular attention focused on interpretation of the gating network and robust inference of parameter estimates. The outline for the rest of this manuscript is as follows: the remainder of this section fills out the literature review and section 2 describes the model in formal detail. Section 3 discusses approaches for estimation while section 4 concerns itself with robust inference of the estimated parameters. Section 6 provides detail on marginal effects of the gating network. Section 7 develops a procedure to grow the gating network in an objective manner.

## 1.1 Relevant Literature

Waterhouse and Robinson 1995 puts forth a method to grow an HME from a single split from the root node. The authors are influenced by the growing technique used for classification and regression trees (Breiman et al. 1984) and apply it to an HME structure. Once the gating structure to an HME tree has been grown, the authors put forth an additional trimming algorithm as well. Fritsch, Finke, and Waibel 1997

consider (Waterhouse and Robinson 1995) and alter their growing algorithm with a mind to scaling the model to handle thousands of experts.

Jordan and Xu 1995 An extended discussion on the convergence of the model used by Jordan and R. Jacobs 1993 ? (VERIFY). The authors also suggest algorithmic improvements to help with estimation.

Jiang and M. A. Tanner 1999 discuss convergence rates of an HME model where experts are from the exponential family with generalized linear mean functions.

Jiang and M. A. Tanner 2000 provide regularity conditions on the HME structure for for a mixture of general linear models estimated by maximum likelihood to produce consistent and asymptotically normal estimates of the mean response. The conditions are validated for poisson, gamma, gaussian, and binomial experts.

## 1.2 Don't leave behind time series analysis

Weigend, Mangeas, and Srivastava 1995 provides a detailed discussion about examining ME applied in a time series context and provide valuable insights to avoid overfitting the model to the data, a common problem in neural network applications.

Huerta, Jiang, and M. A. Tanner 2003 Extends Weigend, Mangeas, and Srivastava 1995 to an HME. Five and a half decades of monthly US Industrial Production Index data. They allowed the series to choose between two models, one modeled as a random walk and the other as trend stationary. In addition, they present a Bayesian approach to estimation.

Carvalho and M. Tanner 2003 lay out the necessary regularity conditions to perform hypothesis tests on stationarity MoE time series of generalized linear models (HoE-GLM) using Wald tests. The dual cases of a well-specified and a miss-specified model are considered. The authors restrict their analysis to MoE-GLM models involving lagged dependent and lagged external covariate variables only. Generalization to include lagged conditional mean values is left for another time.

Carvalho and M. Tanner 2005 is similar to Carvalho and M. Tanner 2003 but applied in a purely auto-regressive context restricted to gaussian models. The authors extend arguments in Carvalho and M. Tanner 2003 to non-stationary series and provide simulated evidence that using the BIC is helpful in selecting the appropriate number of experts to include.

Carvalho and M. Tanner 2006 refocus the discussion on MoE of time series regressions restricted to exponential family distributions. Distilling the available literature at the time, the authors cover the important topics of estimation and asymptotic properties in the maximum likelihood framework, selection of the number of experts, model validation and fitting.

Carvalho and Skoulakis 2010 Applies mixture-of-experts of a single time series. Using stock returns, the authors structure the gating network using lagged dependent variables and an 'external' covariate capturing a measure of the trade volume at that time. THIS NEEDS A LOT MORE...Mention Simulations...READ THIS ONE AGAIN

### 1.3 Additional Articles to Include

Neal and Pfeiffer 2001 cross section

Blei, Kucukelbir, and McAuliffe 2016 A review of variational inference applied to generalized linear models and basic examples.

Ueda and Ghahramani 2002

Bishop and Svenson 2003 find previous bayesian approaches to estimating an HME lacking [Huerta, Jiang, and M. A. Tanner 2003, Ueda and Ghahramani 2002]. Using variational inference, the authors provide a bayesian estimation approach to the log marginal likelihood. With an eye to prediction, the author's advocate that their approach makes the HME model easier to estimate without overfitting. [Discuss how the authors approach model selection]

Carvalho and Skoulakis 2005

## 2 Model

We start by presenting the HME as a standard mixture model. For a given input and output pair  $(X_t, Y_t)$ , each expert provides a probabilistic model relating input  $X_t$  to output  $Y_t$ :

$$P_t^m \equiv P^m(Y_t|X_t, \boldsymbol{\beta}^m), \quad m = 1, 2, \dots, M \quad (1)$$

where  $m$  is one of the  $M$  component experts in the mixture. The experts are combined with associated weights into a mixture distribution

$$P(Y_t|X_t; \boldsymbol{\beta}) = \sum_{m=1}^M \mathbb{I}(m|t) P^m(Y_t|X_t, \boldsymbol{\beta}^m) \quad (2)$$

Here,  $\mathbb{I}_t(m)$  is the probability that the input unit  $t$  belongs to expert  $m$  and has the usual restrictions:  $0 \leq \mathbb{I}(m|t) \leq 1$  for each  $m$  and  $\sum_m \mathbb{I}(m|t) = 1$ . The gating network of the model applies a particular functional form to model  $\mathbb{I}(m|t)$ , which includes a second set of covariates  $Z_t$  and parameter vector  $\boldsymbol{\omega}$ :

$$P(Y_t|X_t, Z_t; \beta, \omega) = \sum_{m=1}^M \mathbb{P}(m|Z_t; \omega) P^m(Y_t|X_t, \beta^m) \quad (3)$$

## 2.1 Gating Network and $\mathbb{P}(m|Z, \omega)$

The gating network model is structured as a collection of nodes in a tree structure that branches out in successive layers. The location of these nodes will be referred to by their address  $a$ . The root node resides at the apex of the tree and has the address 0. The root node then splits into  $J$  different nodes, one level down the tree. The addresses for these  $J$  new nodes are  $1|0, 2|0, \dots, J|0$ . This type of naming convention continues as the rest of network is traversed. At its most general, each gating node can yield an arbitrary number of splits. While a fully generalized gating network is conceptually attractive, it presents practical challenges for implementation. In this paper we address several architectures for the gating network, each with its own set of structural restrictions on the shape of the network and the number of splits each gating node can take. For arbitrary node at address  $a$ , we use a multinomial logistic regression to model the split in direction  $i$  to be:

$$g_t^{a,i} \equiv g_t^{a,i}(Z_t, \omega^a) = \frac{\exp(Z_t \omega^{a,i})}{\sum_{j=1}^J \exp(Z_t \omega^{a,j})} \quad (4)$$

The parameters in equation (4) are subject to the usual identification restrictions. For the remainder of the article, we choose to set  $\omega^{a,J} = \mathbf{0}$  for every gating node. It is important to keep track of the product path an input vector travels from one node to another. If the observation index is suppressed, the product path from one node (say the root node 0) to another (say  $m|\dots|j|i$ ) can be defined as

$$\pi_{g^0 \longleftrightarrow g^k|\dots|j|i|0} = \begin{cases} g^{0,i} g^{i|0,j} \dots g^{\dots|j|i|0,k} & \text{if path is feasible} \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

If one of the nodes is an expert, then we can define the mixture weight of expert  $m$  for input pattern  $i$  to be the product of the path taken from the root node to expert  $m$ :

$$\mathbb{P}(m|Z, \omega) = \pi_{g^0 \longleftrightarrow P^m} \quad (6)$$

For network architectures with multiple paths from the root node to the same expert (see bottom right of figure (1)), we can index these multiples paths by  $l$  so that:

$$\mathbb{P}(m|Z_t, \boldsymbol{\omega}) = \sum_l \pi_{g^0 \xleftrightarrow{l} P^m} \quad (7)$$

By collecting—and summing—all possible paths from the root node to each expert, the conditional probability given in equation (3) can be expanded and expressed as:

$$\begin{aligned} P(Y_t|X_t, Z_t; \boldsymbol{\omega}, \boldsymbol{\beta}) &= \sum_m \mathbb{P}(m|Z_t, \boldsymbol{\omega}) P^m(Y_t|X_t, \boldsymbol{\beta}^m) \\ &= \sum_m P^m(Y_t|X_t, \boldsymbol{\beta}^m) \sum_l \pi_{g^0 \xleftrightarrow{l} P^m} \end{aligned} \quad (8)$$

The product of these individual probabilities across the full sample size  $T$  yields the likelihood function.

$$\mathcal{L}(\boldsymbol{\theta}|Y, X, Z) = \prod_t \sum_m P^m(Y_t|X_t, \boldsymbol{\beta}^m) \sum_l \pi_{g^0 \xleftrightarrow{l} P^m} \quad (9)$$

And taking its log yields to log likelihood

$$\mathcal{l}(\boldsymbol{\theta}|Y, X, Z) = \sum_t \log \sum_m P^m(Y_t|X_t, \boldsymbol{\beta}^m) \sum_l \pi_{g^0 \xleftrightarrow{l} P^m} \quad (10)$$

The functional form of the log likelihood (10) does not lend itself easily to direct optimization, but a well established technique using expectation maximization (Dempster, Laird, and Rubin 1977) to estimate mixture models is available. This was the primary insight of Jordan and R. Jacobs 1993’s original paper.

### 3 The EM Set-Up

The EM approach to estimating an HME model starts by suggesting that if a researcher had perfect information, each input vector  $X_t$  could be matched to the expert  $P^m$  that generated it with certainty. If a set of indicator variables is introduced that captures this certainty, an *augmented* version of the likelihood in equation (9) can be put forward. Define the indicator set as:

$$I_t(m) = \begin{cases} 1 & \text{if input vector } X_t \text{ is generated by expert } m \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

We can then reformulate the likelihood equation

$$\mathcal{L}_c(\boldsymbol{\theta}|Y, X, Z) = \prod_t \prod_m \left[ P^m(Y_t|X_t, \beta^m) \sum_l \pi_{g^0 \longleftrightarrow P^m}^l \right]^{I_t(m)} \quad (12)$$

leading to the complete-data log-likelihood

$$\mathbf{l}_c(\boldsymbol{\theta}|Y, X, Z) = \sum_t \sum_m I_t(m) \left[ \log P^m(Y_t|X_t, \beta^m) + \log \sum_l \pi_{g^0 \longleftrightarrow P^m}^l \right] \quad (13)$$

As mentioned previously, summing over multiple paths  $l$  in equation (13) is only necessary in the HMRE case. For the ME and HME cases,  $l = 1$ , simplifying the second log in (13) to  $\log(\pi_{g^0 \longleftrightarrow P^m})$ . Going forward, we will focus our analysis on the ME and HME specifications with work on the HMRE case arriving in subsequent iterations of the article.

### 3.1 E-Step

The E-step of the algorithm performs an expectation over the complete log-likelihood equation (13), where the expectation includes the additional information contained in the expert regressions. One of the results of this expectation is the creation of second set of weights  $h^a$  that parallel the weights from the gating network  $g^a$  discussed in section (2.1). For an HME model:

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \mathbb{E}[\mathbf{l}_c(\boldsymbol{\theta}|Y, X, Z)] = \sum_t \sum_m \mathbb{E}[I_t(m)] [\log P^m(Y_t|X_t, \beta^m) + \log \pi_{g_t^0 \longleftrightarrow P_t^m}] \\ &= \sum_t \sum_m \pi_{h_t^0 \longleftrightarrow P_t^m} [\log P^m(Y_t|X_t, \beta^m) + \log \pi_{g_t^0 \longleftrightarrow P_t^m}] \end{aligned} \quad (14)$$

Here  $\pi_{h^0 \longleftrightarrow h^k, \dots, |j|i|0}$  is analagous to equation (5)

$$\pi_{h^0 \longleftrightarrow h^k, \dots, |j|i|0} = \begin{cases} h^{0,i} h^{i|0,j} \dots h^{\dots|j|i|0,k} & \text{if path is feasible} \\ 1 & \text{otherwise} \end{cases} \quad (15)$$

and the  $h^{a,i}$  are arrived at using bayes' theorem.

$$h_t^{a,i} = \frac{P_t^m \pi_{g_t^{a,i} \longleftrightarrow P_t^m}}{\sum_k P_t^k \pi_{g_t^{a,j \neq i} \longleftrightarrow P_t^k}} \quad (16)$$

So, now we have two different forms of weights,  $g$ 's and  $h$ 's. The way the  $g$ 's are formed in equation (4), they are only functions of the nodes in the gating network, separate from the expert regressions and the information they contain. For this reason, the authors in Jacob et. al. 1993 refer to  $g$ 's as *priors*. The  $h$ 's draw from both the gating network and the expert regressions and are referred to as *posterior* weights.

One property that  $\pi_{h^0 \longleftrightarrow P^m}$  enjoys is for any node  $a$  that belongs to the gate path from the root node to expert  $m$ , then:

$$\pi_{h^0 \longleftrightarrow P^m} = \pi_{h^0 \longleftrightarrow h^a} h^{a,i} \quad (17)$$

so that when at node  $a$ , split  $i$  is the direction needed in order to reach expert  $m$ .

## 3.2 M-Step

Note that the parameters governing the experts and the gating network in equation (14) are additively separable. Taking each in turn, the Jacobian of the gating portion of equation (14) is:

## 4 Inference

The score vector:

$$\mathbf{S}_t(\boldsymbol{\omega}^{a,i}) \equiv \frac{\partial Q}{\partial \boldsymbol{\omega}^{a,i}} = \pi_{h_t^0 \longleftrightarrow h_t^a} [h_t^{a,i} - g_t^{a,i}] Z_t \quad (18)$$

With

$$\mathbf{S}_t(\boldsymbol{\omega}^a) = [\mathbf{S}_t(\boldsymbol{\omega}^{a,1}), \mathbf{S}_t(\boldsymbol{\omega}^{a,2}), \dots, \mathbf{S}_t(\boldsymbol{\omega}^{a,J-1})] \quad (19)$$

And

$$\mathbf{S}(\boldsymbol{\omega}^a) = \sum_t^T \mathbf{S}_t(\boldsymbol{\omega}^a) \quad (20)$$

The hessian:

$$\mathbf{H}_t(\boldsymbol{\omega}^a) \equiv \frac{\partial^2 Q}{\partial \boldsymbol{\omega}^{a,i} \partial \boldsymbol{\omega}^{a,j}} = \pi_{h_t^0 \longleftrightarrow h_t^a} Z_t \boldsymbol{\Gamma}_t^a Z_t^\top \quad (21)$$

With



$$\mathbf{\Gamma}_t^a = \begin{bmatrix} g_t^{a,1}(1 - g_t^{a,1}) & -g_t^{a,1}g_t^{a,2} & \dots & -g_t^{a,1}g_t^{a,J-1} \\ -g_t^{a,1}g_t^{a,2} & g_t^{a,2}(1 - g_t^{a,2}) & \dots & -g_t^{a,2}g_t^{a,J-1} \\ \vdots & \vdots & \ddots & \vdots \\ -g_t^{a,1}g_t^{a,J-1} & -g_t^{a,2}g_t^{a,J-1} & \dots & -g_t^{a,J-1}(1 - g_t^{a,J-1}) \end{bmatrix} \quad (22)$$

And

$$\mathbf{H}(\omega^a) = \sum_t^T \mathbf{H}_t(\omega^a) \quad (23)$$

Leading to the sandwich estimator:

$$\mathbf{V}(\omega^a) = \mathbf{H}^{-1}(\omega^a) \mathbf{S}(\omega^a) \mathbf{S}(\omega^a)^\top \mathbf{H}^{-1}(\omega^a) \quad (24)$$

## 5 HMRE

Score function for HMRE

$$\frac{\partial Q}{\partial \omega_p^{a,i}} = \sum_t \sum_{m=1} \frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}} \sum_l \frac{\partial \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}}{\partial \omega_p^{a,i}} \quad (25)$$

Where we have the ratio of the posterior  $h^{0,m}$  and prior  $\sum_l \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}$  weights with respect to the root node and the gradient of the posterior weight with respect to the node in questions  $\frac{\partial g^{a,m}}{\partial \omega_p^{a,i}}$ . Since  $\pi_{n_t^0 \leftarrow l \rightarrow P_t^m}$  is a product of values, the partial in equation (25) is simply that same product chain but absent of node  $g^a$  times the partial of node  $g^a$  with respect to  $\omega_p^{a,i}$ .

$$\frac{\partial \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}}{\partial \omega_p^{a,i}} = [\pi_{n_t^0 \leftarrow l \rightarrow P_t^m} (-g^a)] \frac{\partial g^{a,j(l)}}{\partial \omega^{a,i}} \quad (26)$$

Hessian for HMRE

$$\frac{\partial^2 Q}{\partial \omega_p^{a,i} \partial \omega_q^{a',j}} = \sum_t \sum_m \left[ \frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}} \frac{\partial \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}}{\partial \omega_p^{a,i}} - \left( \frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}} \right)^2 \sum_l \sum_{l'} \frac{\partial \pi_{n_t^0 \leftarrow l \rightarrow P_t^m}}{\partial \omega_p^{a,i}} \frac{\partial \pi_{n_t^0 \leftarrow l' \rightarrow P_t^m}}{\partial \omega_q^{a',j}} \right] \quad (27)$$

Posterior node for HMRE

$$h_t^{a,m} = \frac{P_t^m \sum_l \pi_{g_t^a \leftrightarrow P_t^m}}{\sum_k P_t^k \sum_l \pi_{g_t^a \leftrightarrow P_t^k}} \quad (28)$$

## 6 Marginal Effects

To explore what kind of marginal effects each variable has on the mixture's output, we start with equation (3) but replace the expert distributions  $P_t^m$  with the functional form of the expert regression  $f_t^m$  and use the relationship in equation (6).

$$f_t = f(Y_t|X_t, Z_t; \boldsymbol{\beta}, \boldsymbol{\omega}) = \sum_{m=1}^M \pi_{g_t^0 \leftrightarrow f_t^m} f^m(Y_t|X_t, \boldsymbol{\beta}^m) \quad (29)$$

The functional form of the marginal effect depends if the variables are exclusive to the gating network,  $Z$ , exclusive to the expert regressions,  $X$ , or is present in both sets of covariates,  $W \subset X$  and  $W \subset Z$ . For covariates exclusive to the gating network, their marginal effects are:

$$\frac{\partial f_t}{\partial Z} = \sum_{m=1}^M \frac{\partial \pi_{g_t^0 \leftrightarrow f_t^m}}{\partial Z} f_t^m \quad (30)$$

For covariates exclusive to the expert regressions, their marginal effects are:

$$\frac{\partial f_t}{\partial X} = \sum_{m=1}^M \pi_{g_t^0 \leftrightarrow f_t^m} \frac{\partial f_t^m}{\partial X} \quad (31)$$

And covariates in both:

$$\frac{\partial f_t}{\partial W} = \sum_{m=1}^M \left\{ \frac{\partial \pi_{g_t^0 \leftrightarrow f_t^m}}{\partial W} f_t^m + \pi_{g_t^0 \leftrightarrow f_t^m} \frac{\partial f_t^m}{\partial W} \right\} \quad (32)$$

Just as for logistic and multinomial regressions, the marginal effect of the entire gating network has a closed form solution. Independently, looking at the network's marginal effects provides a sense of what gating variables play a decisive role in directing input patterns to the appropriate expert regressions. Starting with equation (5), we take the partial with respect to gating matrix  $Z_t$ .

$$\delta^m \equiv \frac{\partial \pi_{g_t^0 \leftrightarrow f_t^m}}{\partial Z} = \frac{\partial g^{0,i} g^{i|0,j} \dots g^{k|\dots|j|i|0,m}}{\partial Z} \quad (33)$$

Applying the product rule gives us:

$$\begin{aligned}\delta^m &= \frac{\partial g^{0,i}}{\partial Z} g^{i|0,j} \dots g^{k|\dots|j|i|0,m} \\ &\quad + g^{0,i} \frac{\partial g^{i|0,j}}{\partial Z} \dots g^{k|\dots|j|i|0,m} + \dots \\ &\quad + g^{0,i} g^{i|0,j} \dots \frac{\partial g^{k|\dots|j|i|0,m}}{\partial Z}\end{aligned}\tag{34}$$

and since:

$$\frac{\partial g^{a,i}}{\partial Z} = g^{a,i} \left( \omega^{a,i} - \sum_j g^{a,j} \omega^{a,j} \right) = g^{a,i} (\omega^{a,i} - \bar{\omega}^a)\tag{35}$$

we can substitute equation (35) into (34) to arrive at:

$$\delta^m = \pi_{g^0 \longleftrightarrow P^m} (\omega^{0,i} + \omega^{i|0,j} + \dots + \omega^{k|\dots|j|i|0,m} - (\bar{\omega}^0 + \bar{\omega}^{i|0} + \dots + \bar{\omega}^{k|\dots|j|i|0}))\tag{36}$$

Standard errors for these marginal effects can be estimated using the delta method.  
(*Edit: These are actually only for the gating network, not the entire mixture output.*  
*Also, this is only appropriate for MoE, not an HME)*

$$\mathbf{V}(\delta^m) = \left( \frac{\partial \delta^m}{\partial \omega} \right) \begin{bmatrix} \mathbf{V}(\omega^0) & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{V}(\omega^{1|0}) & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{V}(\omega^{2|0}) & \dots \\ \vdots & \vdots & & \ddots \end{bmatrix} \left( \frac{\partial \delta^m}{\partial \omega} \right)^\top\tag{37}$$

where  $\mathbf{V}(\omega^a)$  is the sandwich estimator in equation (24) and:

$$\frac{\partial \delta^m}{\partial \omega^{a,p}} = [1 - g^{a,p}] (\delta^m Z + \pi_{g^0 \longleftrightarrow P^m}) - g^{a,p} \pi_{g^0 \longleftrightarrow P^m} [\omega^{a,p} - \bar{\omega}^a] Z\tag{38}$$

if  $\omega^{a,p}$  appears in the path of  $\pi_{g^0 \longleftrightarrow P^m}$ , or if not, then:

$$\frac{\partial \delta^m}{\partial \omega^{a,p}} = -g^{a,p} (\delta^m Z + \pi_{g^0 \longleftrightarrow P^m}) - g^{a,l} \pi_{g^0 \longleftrightarrow P^m} [\omega^{a,p} - \bar{\omega}^a] Z\tag{39}$$

Table 1: Linear Regression vs HME

	Regression		HME		MoE		
	Const.	AR	MA	Model	ARCH	GARCH	
Setosa	3.222	–	3.217	–	Const. 1	–	–
Virginica	1.695	–	1.956	–	Slope 1	–	
Versicolor	1.373	–	1.148	–	Const. 2	–	
Setosa x Petal.Width	0.837	–	0.943	–	Slope 2	–	
Virginica x Petal.Width	0.631	–	0.497	–	Const. 3	–	
Versicolor x Petal.Width	1.054	–	1.292	–	Slope 3	–	
AME	0.947	–	0.574				

<sup>1</sup> This is a footnote.

## 7 Growing and Pruning the Gating Network

## 8 A simple example

Using the iris data set we can test if the HME model able to capture structure of latent variables while maintaing accurate regression

## 9 Miscellaneous

### 9.1 Odd Facts

Variables that appear in the gating network but not in the expert regressions are sometimes called Concomitant Variables (see R FlixMix package).

### 9.2 Wald Test is Invalid

S+plus GLM section on problems with binomial GLMs – Hauck and Donner (1977) JASA. Quoting S+plus: If there are some  $\hat{\beta}_i$  that are large, the curvature of the log-likelihood at  $\hat{\beta}$  can be much less than near  $\beta_i = 0$ , and so the Walk approximation underestimates the change in log-likelihood on setting  $\beta_i = 0$ . This happens in such a way that as  $|\hat{\beta}_i| \rightarrow \infty$ . Thus highly significant coefficients according to the likelihood ratio test may have non-significant t ratios ..... There is one fairly common circumstance in which both convergence problems and the Hauek-Donner phenomenon can occur. This is when the fitted probs are extremely close to zero or one.

## 10 Diagnostics

pg 7 of Weigend, Mangeas, and Srivastava 1995 suggested observing the distribution of the terminal  $g_i$ . If only one expert is responsible for each observations,  $g_i$  will be close to one for a single expert and near zero for all other experts. Can we formalize this comparison in to a specific test?

density forecasts evaluations.

Standard likelihood-ratio test is not valid (Carvalho and M. Tanner 2006) with AIC/BIC/VOUNG test being preferred.

## References

- Bishop, Christopher and Markus Svenson (2003). “Bayesian Hierarchical Mixtures of Experts”. In: *Proceedings Nineteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, pp. 57–64.
- Blei, David M., Alp Kucukelbir, and McAuliffe (2016). “Variational Inference: A Review for Statisticians”. In: *ArXiv e-prints*. eprint: 1601.00670.
- Brieman, L. et al. (1984). *Classification and Regression Trees*. Wadsworth and Brooks/Cole.
- Carvalho, Alexandre and Georgios Skoulakis (2005). “Ergodicity and existence of moments for local mixtures of linear autoregressions”. In: *Statistics and Probability Letters* 71.3, pp. 313–322.
- (2010). “Time Series Mixutres of Generalized t Experts: ML Estimation and an Application to stock return density forecasting”. In: *Econometric Reviews* 29.5-6, pp. 642–687. DOI: 10.1080/07474938.2010.481987.
- Carvalho, Alexandre and Martin Tanner (2003). “Hypothesis testing in mixture-of-experts of generalized linear time series”. In: *2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings*. Pp. 285–292.
- (2005). “Mixtures-of-experts of autoregressive time series: asymptotic normality and model specification”. In: *IEEE Transactions on Neural Networks* 16.1, pp. 39–56. ISSN: 1045-9227.
- (2006). “Modeling nonlinearities with mixtures-of-experts of time series models”. In: *International Journal of Mathematics and Mathematical Sciences* 2006.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). “Maximum likelihood from incomplete data via the em algorithm”. In: *Journal of the Royal Statistical Society. Series B*. 39.1, pp. 1–38.

- Fritsch, Jürgen, Michael Finke, and Alex Waibel (1997). “Adaptively Growing Hierarchical Mixtures of Experts”. In: *Advances in Neural Information Processing Systems 9*. Ed. by M. C. Mozer, M. I. Jordan, and T. Petsche. MIT Press, pp. 459–465. URL: <http://papers.nips.cc/paper/1279-adaptively-growing-hierarchical-mixtures-of-experts.pdf>.
- Goldfeld, Stephan M. and Richard E. Quandt (1973). “A Markov Model for Regime Switching”. In: *Journal of Econometrics* 1 (1), pp. 3–16.
- Hamilton, J.D. (1989). “A new approach to the economic analysis of nonstationary time series and the business cycle”. In: *Econometrica* 57, pp. 357–384.
- Huerta, Gabriel, Wenxin Jiang, and Martin A. Tanner (2003). “Time series modeling via hierarchical mixtures”. In: *Statistica Sinica* 13.
- Jacobs, R. A. et al. (1991). “Adaptive mixture of local experts”. In: *Neural Computation* 3, pp. 79–82.
- Jiang, Wenxin and Martin A. Tanner (1999). “Hierarchical Mixture-of-Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation”. In: *The Annals of Statistics* 27.3, pp. 987–1011.
- (2000). “On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models”. In: 46.3, pp. 1005–1013. ISSN: 0018-9448.
- Jordan, M. and R. Jacobs (1993). “Hierarchical mixtures of experts and the EM algorithm”. In: *Proceedings of 1993 International Joint Conference on Neural Networks*.
- Jordan, M. and L. Xu (1995). “Convergence results for the em approach to mixtures-of-experts architectures”. In: *Neural Networks* 8 (9), pp. 1409–1431.
- Neal, Jeffries and Ruth Pfeiffer (2001). “A mixture model for the probability distribution of rain rate”. In: *Environmetrics* 12.1, pp. 1–10.
- Ueda, N. and Z. Ghahramani (2002). “Bayesian model search for mixture models based on optimizing variational bounds”. In: *Neural Networks* 15.10, pp. 1223–1241.
- Waterhouse, S.R. and A.J. Robinson (1995). “Constructive Algorithms for Hierarchical Mixture of Experts”. In: *Advances in Neural Information Processing Systems* 8.
- Weigend, A., M. Mangeas, and A. Srivastava (1995). “Nonlinear gated experts for time series: discovering regimes and avoiding overfitting”. In: *International Journal of Neural Systems* 6, pp. 373–399.

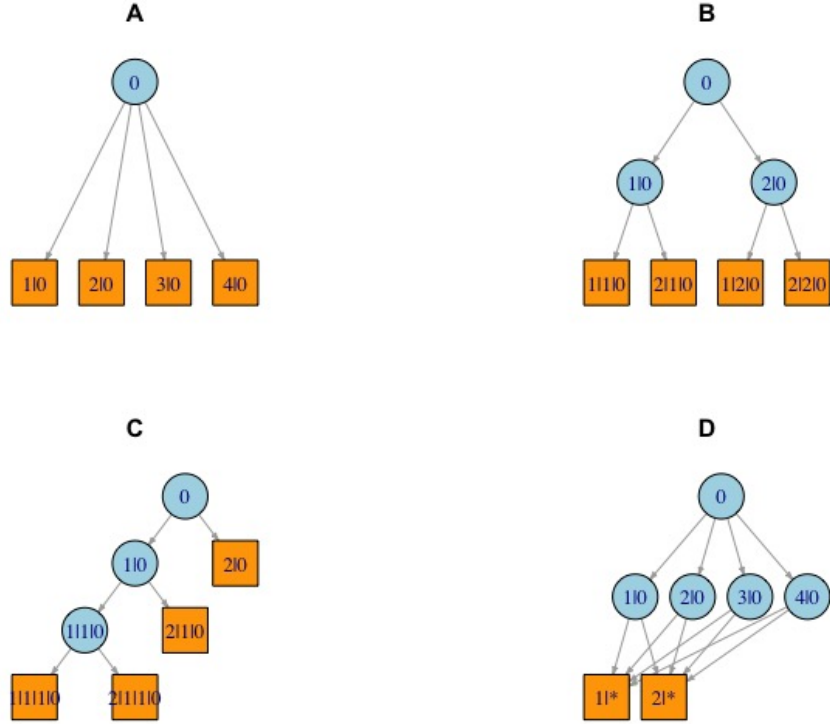


Figure 1: Networks **A** - **D** depict various network architectures that are discussed in this article. For all four networks, gating nodes are represented as blue circles and experts as orange rectangles. Network **A** illustrates the original Mixture of Experts (ME) architecture with a single multinomial split leading to a set of experts one layer down. Networks **B** and **C** both represent different flavors of a Hierarchical Mixture of Experts (HME). Network **B** is a symmetric network of depth 2 with successive binary splits. Network **C** is an asymmetric network of depth 3 with successive binary splits. Network **D** is an example of the Hierarchical Mixture of Repeated Experts (HMRE) architecture. Notice that multiple paths exist from the root node 0 to each expert. Compare this to networks **A** - **C**, where there is only one unique path from the root node to each expert.

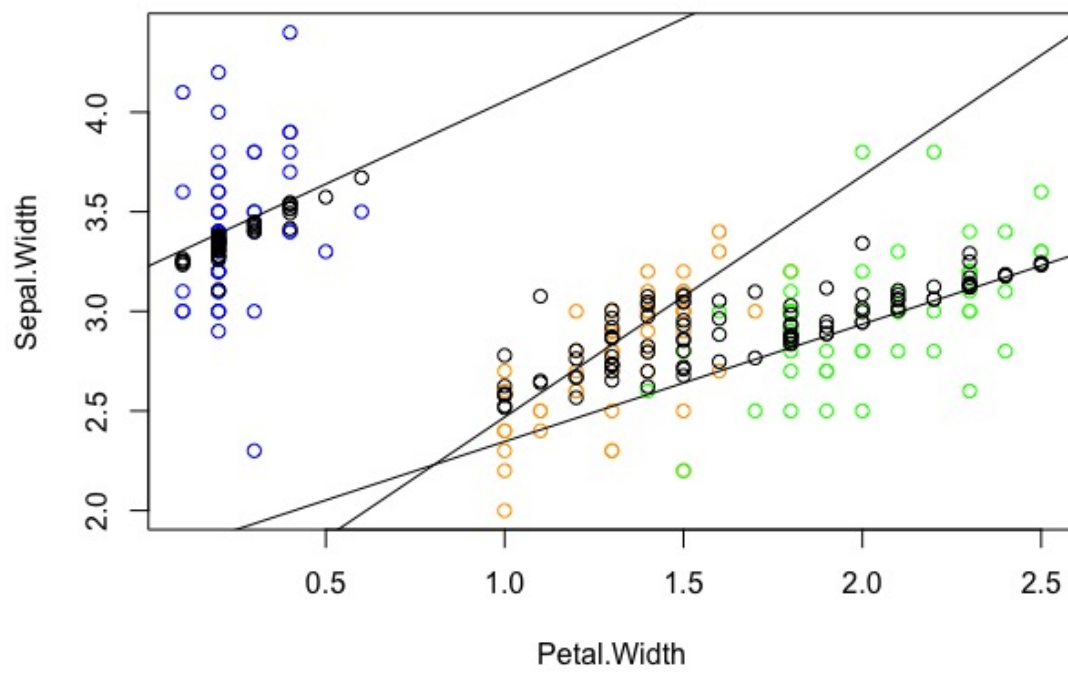


Figure 2: Does it Work?