Econometric Applications of Hierarchical Mixture of Experts

Lucas C. Dowiak

February 23, 2019

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

Abstract

In this article, a novel mixture model is studied. Named the hierarchical mixture of experts (HME) in the machine learning literature, the mixture model utilizes a set of covariates and a tree-based architecture to efficiently allocate each observation to the appropriate local regression. The nature of the conditional weighting scheme provides the researcher a natural interpretation of how the local (and latent) sub-populations are formed. The model is demonstrated by estimating a Mincer earning function using census data. Marginal effects, robust standard errors, a tree-growing algorithm, and a modest extension are also discussed.

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 successfuly 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 probablistic 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 refered 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 recognization 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 it's 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 discuss 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 proceedure to grow the gating network in an objective manner.

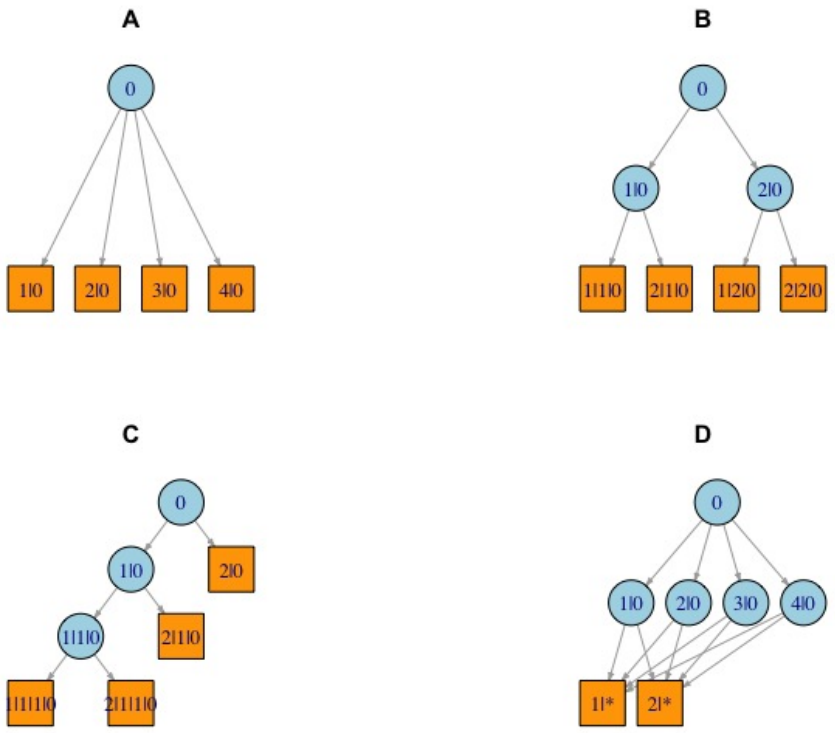

Figure 1: Networks **A** - **D** depict various network architectures that are discussed in this article. For all four networks, gating nodes are represented as blue circles and experts as orange rectangles. Network **A** illustrates the original Mixture of Experts (ME) architexture with a single multinomial 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.

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 (Brieman 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 hypthesis tests on stationarity MoE time series of generalized linear models (HoE-GLM) using Wald tests. The dual cases of a well-specificed 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. Generazlization 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 approprite 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 variabels 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 probablistic model relating input X_t to output Y_t:

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

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

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

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

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

2.1 Gating Network and $\Pi(m|Z, \boldsymbol{\omega})$

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 $1|0, 2|0, ..., J|0$. 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 arbitary node at address a, we use a multinomial logistic regression to model the split in direction i to be:

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

The parameters in equation (4) are subject to the usual identification restrictions. For the remainder of the article, we choose to set $\boldsymbol{\omega}^{a,J} = \mathbf{0}$ 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 supressed, the product path from one node (say the root node 0) to another (say $m|\ldots|j|i$) can be defined as

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

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:

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

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:

$$\mathbb{\Pi}(m|Z_t, \boldsymbol{\omega}) = \sum\nolimits_l \pi_{g^0 \overset{l}{\longleftrightarrow} P^m} \tag{7}$$

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:

$$\begin{aligned} P(Y_t|X_t, Z_t; \boldsymbol{\omega}, \boldsymbol{\beta}) &= \sum_m \mathbb{\Pi}(m|Z_t, \boldsymbol{\omega}) P^m(Y_t|X_t, \beta^m) \\ &= \sum_m P^m(Y_t|X_t, \beta^m) \sum\nolimits_l \pi_{g^0 \overset{l}{\longleftrightarrow} P^m} \end{aligned} \tag{8}$$

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

$$\mathcal{L}(\boldsymbol{\theta}|Y, X, Z) = \prod_t \sum_m P^m(Y_t|X_t, \beta^m) \sum\nolimits_l \pi_{g^0 \overset{l}{\longleftrightarrow} P^m} \tag{9}$$

And taking its log yields to log likelihood

$$\boldsymbol{l}(\boldsymbol{\theta}|Y, X, Z) = \sum_t \log \sum_m P^m(Y_t|X_t, \beta^m) \sum\nolimits_l \pi_{g^0 \overset{l}{\longleftrightarrow} P^m} \tag{10}$$

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:

$$I_t(m) = \begin{cases} 1 & \text{if observation } t \text{ is generated by expert } m \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

We can then reformulate the likelihood equation

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

leading to the complete-data log-likelihood

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

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:

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

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

$$\pi_{h^0 \longleftrightarrow h^{k|\ldots|j|i|0}} = \begin{cases} h^{0,i}\ h^{i|0,j} \ldots h^{\cdots|j|i|0,k} & \text{if path is feasible} \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

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

$$h_t^{a,i} = \frac{g^{a,i} \sum_k P_t^k \pi_{g_t^{i|a} \longleftrightarrow P^k}}{\sum_j g^{a,j} \sum_m P_t^m \pi_{g_t^{j|a} \longleftrightarrow P^m}} \tag{16}$$

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.

3.2 M-Step

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

The expert regressions are weighted regressions

4 Inference

The score vector:

$$\boldsymbol{S}_t(\boldsymbol{\omega}^{a,i}) \equiv \frac{\partial Q}{\partial \boldsymbol{\omega}^{a,i}} = \pi_{h_t^0 \longleftrightarrow h_t^a} \left[h_t^{a,i} - g_t^{a,i} \right] Z_t \tag{17}$$

With

$$\boldsymbol{S}_t(\boldsymbol{\omega}^a) = \left[\boldsymbol{S}_t(\boldsymbol{\omega}^{a,1}), \boldsymbol{S}_t(\boldsymbol{\omega}^{a,2}), \ldots, \boldsymbol{S}_t(\boldsymbol{\omega}^{a,J-1}) \right] \tag{18}$$

And

$$\boldsymbol{S}(\boldsymbol{\omega}^a) = \sum_t^T \boldsymbol{S}_t(\boldsymbol{\omega}^a) \tag{19}$$

The hessian:

$$\mathbf{H}_t(\boldsymbol{\omega}^a) \equiv \frac{\partial^2 Q}{\partial \boldsymbol{\omega}^{a,i} \partial \boldsymbol{\omega}^{a,j}} = \pi_{h_t^0 \longleftrightarrow h_t^a} Z_t \boldsymbol{\Gamma}_t^a Z_t^\top \tag{20}$$

With

$$\mathbf{\Gamma}_t^a = \begin{bmatrix} g_t^{a,1}(1-g_t^{a,1}) & -g_t^{a,1}g_t^{a,2} & \dots & -g_t^{a,1}g_t^{a,J-1} \\ -g_t^{a,1}g_t^{a,2} & g_t^{a,2}(1-g_t^{a,2}) & \dots & -g_t^{a,2}g_t^{a,J-1} \\ \vdots & \vdots & \ddots & \vdots \\ -g_t^{a,1}g_t^{a,J-1} & -g_t^{a,2}g_t^{a,J-1} & \dots & -g_t^{a,J-1}(1-g_t^{a,J-1}) \end{bmatrix} \tag{21}$$

And

$$\boldsymbol{H}(\boldsymbol{\omega}^a) = \sum_t^T \mathbf{H}_t(\boldsymbol{\omega}^a) \tag{22}$$

Leading to the sandwich estimator:

$$\boldsymbol{V}(\boldsymbol{\omega}^a) = \boldsymbol{H}^{-1}(\boldsymbol{\omega}^a)\boldsymbol{S}(\boldsymbol{\omega}^a)\boldsymbol{S}(\boldsymbol{\omega}^a)^\top \boldsymbol{H}^{-1}(\boldsymbol{\omega}^a) \tag{23}$$

5 HMRE

Score function for HMRE

$$\frac{\partial Q}{\partial \omega_p^{a,i}} = \sum_t \sum_{m=1} \frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}} \sum_l \frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} \tag{24}$$

Where we have the ratio of the posterior $h^{0,m}$ and prior $\sum_l \pi_{n^a \overset{l}{\longleftrightarrow} P^m}$ weights with resepect 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 $\pi_{n^a \overset{l}{\longleftrightarrow} P^m}$ is a product of values, the partial in equation (24) is simply that same product chain but absent of node g^a times the partial of node g^a with respect to $\omega_p^{a,i}$.

$$\frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} = [\pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}(-g^a)] \frac{\partial g^{a,j(l)}}{\partial \omega^{a,i}} \tag{25}$$

Hessian for HMRE

$$\frac{\partial^2 Q}{\partial \omega_p^{a,i} \partial \omega_q^{a',j}} = \sum_t \sum_m \left[\frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}} \frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} - \left(\frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}} \right)^2 \sum_l \sum_{l'} \frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} \frac{\partial \pi_{n_t^0 \overset{l'}{\longleftrightarrow} P_t^m}}{\partial \omega_q^{a',j}} \right] \tag{26}$$

Posterior node for HMRE

$$h_t^{a,m} = \frac{P_t^m \sum_l \pi_{g_t^a \longleftrightarrow^l P_t^m}}{\sum_k P^k \sum_l \pi_{g_t^a \longleftrightarrow^l P_t^k}} \tag{27}$$

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).

$$f_t = f(Y_t|X_t, Z_t; \boldsymbol{\beta}, \boldsymbol{\omega}) = \sum_{m=1}^{M} \pi_{g_t^0 \longleftrightarrow f_t^m} f^m(Y_t|X_t, \boldsymbol{\beta}^m) \tag{28}$$

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 $W \subset X$. For covariates exclusive to the gating network, their marginal effects are:

$$\frac{\partial f_t}{\partial Z} = \sum_{m=1}^{M} \frac{\partial \pi_{g_t^0 \longleftrightarrow f_t^m}}{\partial Z} f_t^m \tag{29}$$

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

$$\frac{\partial f_t}{\partial X} = \sum_{m=1}^{M} \pi_{g_t^0 \longleftrightarrow f_t^m} \frac{\partial f_t^m}{\partial X} \tag{30}$$

And covariates in both:

$$\frac{\partial f_t}{\partial W} = \sum_{m=1}^{M} \left\{ \frac{\partial \pi_{g_t^0 \longleftrightarrow f_t^m}}{\partial W} f_t^m + \pi_{g_t^0 \longleftrightarrow f_t^m} \frac{\partial f_t^m}{\partial W} \right\} \tag{31}$$

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 local expert regressions. Starting with equation (5), we take the partial with respect to gating matrix Z_t.

$$\delta^m \equiv \frac{\partial \pi_{g^0 \longleftrightarrow f^m}}{\partial Z} = \frac{\partial g^{0,i} g^{i|0,j} \cdots g^{k|\cdots|j|i|0,m}}{\partial Z} \tag{32}$$

Applying the product rule gives us:

$$\begin{aligned}\delta^m = &\frac{\partial g^{0,i}}{\partial Z} g^{i|0,j} \cdots g^{k|\cdots|j|i|0,m} \\ &+ g^{0,i} \frac{\partial g^{i|0,j}}{\partial Z} \cdots g^{k|\cdots|j|i|0,m} \\ &+ \dots \\ &+ g^{0,i} g^{i|0,j} \dots \frac{\partial g^{k|\cdots|j|i|0,m}}{\partial Z}\end{aligned} \tag{33}$$

and since:

$$\frac{\partial g^{a,i}}{\partial Z} = g^{a,i} \left(\boldsymbol{\omega}^{a,i} - \sum_j g^{a,j} \boldsymbol{\omega}^{a,j} \right) = g^{a,i} \left(\boldsymbol{\omega}^{a,i} - \bar{\boldsymbol{\omega}}^a \right) \tag{34}$$

we can substitute equation (34) into (33) to arrive at:

$$\delta^m = \pi_{g^0 \longleftrightarrow P^m} \left(\boldsymbol{\omega}^{0,i} + \boldsymbol{\omega}^{i|0,j} + \cdots + \boldsymbol{\omega}^{k|\cdots|j|i|0,m} - \left(\bar{\boldsymbol{\omega}}^0 + \bar{\boldsymbol{\omega}}^{i|0} + \cdots + \bar{\boldsymbol{\omega}}^{k|\cdots|j|i|0} \right) \right) \tag{35}$$

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)*

$$\boldsymbol{V}(\boldsymbol{\delta^m}) = \left(\frac{\partial \boldsymbol{\delta}^m}{\partial \boldsymbol{\omega}} \right) \begin{bmatrix} \boldsymbol{V}(\boldsymbol{\omega}^0) & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \boldsymbol{V}(\boldsymbol{\omega}^{1|0}) & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{V}(\boldsymbol{\omega}^{2|0}) & \\ \vdots & \vdots & & \ddots \end{bmatrix} \left(\frac{\partial \boldsymbol{\delta}^m}{\partial \boldsymbol{\omega}} \right)^\top \tag{36}$$

where $\boldsymbol{V}(\boldsymbol{\omega}^a)$ is the sandwich estimator in equation (23) and:

$$\frac{\partial \boldsymbol{\delta}^m}{\partial \boldsymbol{\omega}^{a,p}} = [1 - g^{a,p}](\delta^m Z + \pi_{g^0 \longleftrightarrow P^m}) - g^{a,p} \pi_{g^0 \longleftrightarrow P^m} [\omega^{a,p} - \bar{\omega}^a] Z \tag{37}$$

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

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

7 Growing and Pruning the Gating Network

8 A simple example

In order to provide a concrete example of the concepts discussed previously, the ME and HME models are demonstrated on a small and well known dataset collected by Edgar Anderson (Anderson 1936) and popularized in the statistics literature by Ronald Fisher (Fisher 1936). Anderson collected 50 measurements each from three different species of iris flowers; the width and length of both the petal and the sepal. Figure 2 provides a basic view of the species specific clustering inherent in the data.

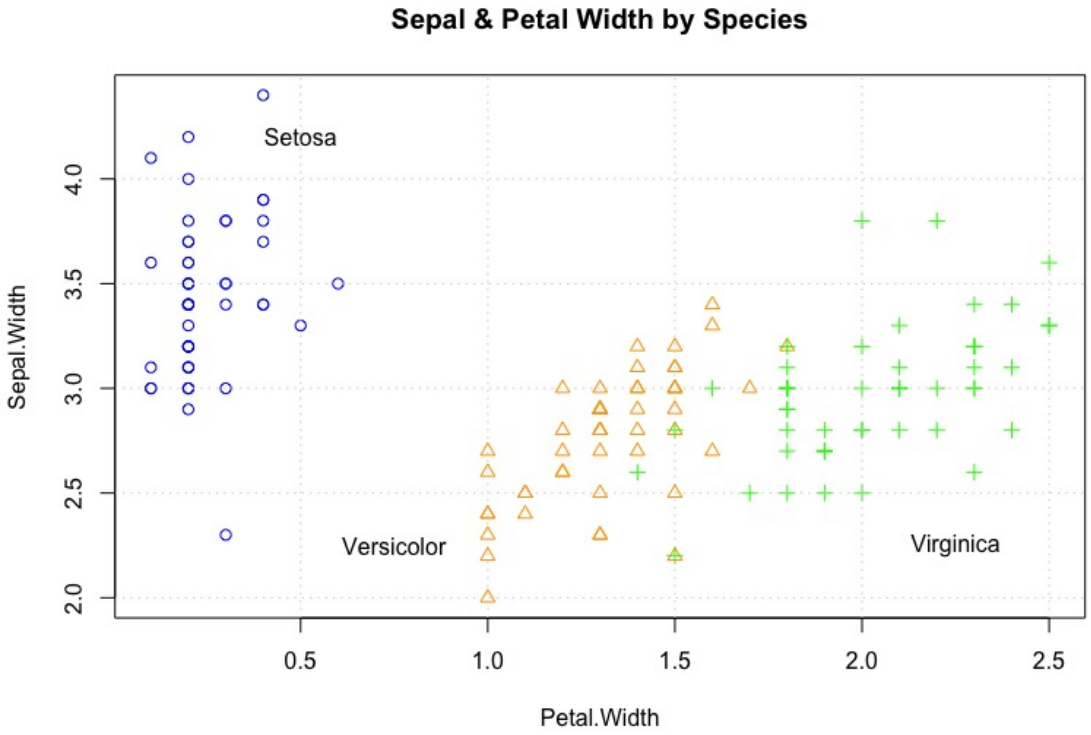

Figure 2: Three different Iris species are presented: Setosa (blue circles), Versicolor (orange triangles), Virginia (green crosses). Sepal width is on the vertical axis and petal width on the horizontal axis.

The work below uses the ME and HME architectures to estimate a flower's sepal width using only it's petal width as a predictor. The petal width will be used as

the sole covariate in the local linear expert regressions (X) as well as in the gating network (Z).

$$sepal.width_i = \beta_0 + \beta_1 * petal.width_i + \varepsilon_i \mid \omega_0 + \omega_1 * petal.width_i \tag{39}$$

The goal is to have the gating network of the models identify the inherent species-specific clustering without explicit knowledge of each observation's species classification, and then fit an appropriate local regression to the self-identified clusters. As a benchmark, an OLS model is run where a flower's petal width is interacted with it's species, resulting in a species-specific estimation of sepal width.

$$sepal.width_{is} = \beta_{0,s} + \beta_{1,s} * petal.width_{is} + \varepsilon_{is} \tag{40}$$

Two sets of regressions are run. Since the Versicolor and Virginica species can be viewed as one larger cluster, a two-expert ME model is run and compared to a benchmark OLS where Versicolor and Virginica are labelled as the same species. A second set of regressions are run with three mixture experts for ME and HME, as well as their corresponding three species OLS regression. Results are collected in table 1. Coefficients for local experts in the two expert ME regression match closely with the OLS benchmark. When moving to the three expert models, there is now a choice on what kind of gating scheme to employ. We can go deep (HME) by adding a gating network with depth two, or we can go wide (ME) by keeping the depth of the gating network at one. When comparing the coefficients of the local regressions, the HME architecture clearly outperforms the ME architecture. The ME model fails to identify the three seperate species that are known to exist.

Table 1: Iris Dataset - OLS vs ME vs HME

	2 Expert Mixture				3 Expert Mixture					
	OLS		ME		OLS		HME		ME	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Setosa										
Const.	3.22	0.11**	3.22	0.13**	3.22	0.11**	3.22	0.13**	3.45	0.13**
Petal.Width	0.84	0.42*	0.95	0.49**	0.84	0.41*	0.94	0.49	0.39	0.46
Virginica										
Const.	–	–	–	–	1.70	0.32**	1.96	0.12**	3.02	0.05**
Petal.Width	–	–	–	–	0.63	0.16**	0.50	0.06**	0.21	0.31
Versicolor										
Const.	–	–	–	–	1.37	0.29**	1.15	0.12**	2.13	0.09**
Petal.Width	–	–	–	–	1.05	0.22**	1.29	0.09**	0.44	0.06**
Virg + Versi										
Const.	2.13	0.13**	2.13	0.09**	–	–	–	–	–	–
Petal.Width	0.44	0.07**	0.44	0.06**	–	–	–	–	–	–
AME										
Petal.Width	0.57	–	0.49	–	0.84	–	0.57	–	0.62	–
Log-Like	-35.5	–	-31.9	–	-29.3	–	-21.8	–	-27.8	–
N	150	–	150	–	150	–	150	–	150	–

** $p < 0.01$, * $p < 0.05$

OLS regressions are modeled using equation (40)

ME regressions are modeled using equation (39) and architecture $\boldsymbol{A}$ from figure 1

HME regressions are modeled using equation (39) and architecture $\boldsymbol{C}$ from figure 1

9 A Mincer Wage Equation

For a more economically relevant example, we turn our attention to a common topic in labor economics: the income return on an additional year of education. At times called the "Mincer wage equation", our version of it will be:

$$\log(wage) = \beta_0 + \beta_1 * \text{Age} + \beta_2 * \text{Age}^2 + \beta_3 * \text{YrsEdu} + \boldsymbol{\beta_4 X} + \varepsilon \tag{41}$$

with $\boldsymbol{X}$ containing a set of individual-specific variables as well as a set of occupation-specific attributes. Our data will come from two sources. First, from the 2000 Census, we devise a measure of hourly wages in addition to age, years of education (YrsEdu), job occupations codes, and a set of demographic identifiers indicating the race of the indivduals contained in census sample. Second, we pull from the Occupation Information Network (ONet) a set of knowledge and skill-based attributes describing what qualities are necessary to perform each job suitably. ONet is a federally sponsered source of occupational information. It details, on a per occupation basis, "the knowledge, skills, and abilities required as well as how the work is performed in terms of tasks, work activities, and other descriptors" (*Occupational Information Network (O*NET)* 2019).

To provide an interesting space to classify occupational codes from the census data, we've chosen the following four attributes to include in the covariates $\boldsymbol{X}$ of our wage equations:

1. Social Perceptiveness [1]
2. Design [2]
3. Data Analytics [3]
4. Creative Thinking [4]

For these selected attributes, ONet grades their relevance on a 100 point scale. Each attribute contains two of scales, an "importance" scale and a "level" scale. The importance scale denotes how critical the attribute is to the occupation while the level indicates how much the skill is required or needed to perform the occupation. To unify the two measures, we follow the paper Prof Wijverberg gave me and take a

[1] Skills - Social Skills - Social Perceptiveness
[2] Work Activities - Mental Processes - Analyzing Data or Information
[3] Knowledge - Design
[4] Work Activities - Mental Processes - Thinking Creatively

Table 2: Summary Statistics

	25%	Mean	50%	75%
Wage (hr)	9.20	15.82	13.32	19.44
Yrs Edu	12.00	13.78	14.00	16.00
Age	30.00	39.15	39.00	48.00
Age16	14.00	23.15	23.00	32.00
Female	–	40.47	–	–
Af Amer	–	8.62	–	–
Indian	–	1.05	–	–
White	–	77.00	–	–
Hispanic	–	10.00	–	–
Asian	–	3.36	–	–
Creative	46.65	53.30	53.82	58.94
Design	15.00	30.58	26.33	38.97
Analytic	44.01	52.63	52.68	62.18
Perseptive	41.15	50.49	46.03	59.84

N = 68,642

cobb-douglass style average with a 2/3's weight for importance and a 1/3 weight for the level scale.

To link the occupational codes in the census data to the soc codes used by ONet, we use the cross walk provided by Sarah Porter. This matching is not one-to-one. When more than one soc code points to a single census code, we take the average of the soc codes.

A summary descriptions of the covariates are provided in table 2. For comparative purposes, we estimate equation (41) for both the HME and ME architectures, testing whether there is any inherent advantage to allowing the gating network to go deep (HME), as opposed to staying shallow with a depth of one. For both architectures, we start with two experts, and then continue to add them until a minimal information criterion is found. In particular, we choose to minimize the Bayesian information criterion (BIC). Compared to the Akaike information criterion (AIC), the BIC places a higher penalty on the number of parameters in a model. We find this feature preferable since the growth in the number of paramaters can be considerable in both models as the number of experts increase.

A natural question to consider as a researcher is where to put the variable(s) of interest while performing an HME estimation. Jiang and M. A. Tanner 2000

provide their proof of model consistency for HME of GLM's for the case where all covariates appear in the gating network as well as the experts. We will call this the *full* specifications:

$$log(wage) = Age + YrsEdu + Sex + Race + Occ \mid Age + YrsEdu + Sex + Race + Occ \tag{42}$$

We will compare this *full* specification to two others. A *mid* specification where the local experts contain age and years of education while removing demographic indicators:

$$log(wage) = Age + YrsEdu \mid Age + YrsEdu + Sex + Race + Occ \tag{43}$$

And finally a *minimal* specification where our core variable of interest, years of education, appears solely in the gating network.

$$log(wage) = Age \mid Age + YrsEdu + Sex + Race + Occ \tag{44}$$

Results for these regressions are collected in table 3. There are a few points worth noting. First, for this dataset, increasing the number of experts yields increasingly better log-likelihood and BIC values. Second, there is a clear advantage to using the HME architecture. In table 3, compare the rows with different gating types (col 1) but with the same number of experts (col 3).The HME approach outperforms the ME across the board. Third, for both ME and HME types, models with more covariates in the local expert regressions lead to higher log-likeihood values.

The coefficients on years of education (or variable of interests), which are summarized in table 4, are pretty stable across model specifications. With the exception of the two-expert *minimal* model, coefficient values range from 0.083 to 0.092.

We also compare the marginal effects across specifications. According to table 3, the five-expert HME model performs the best for each of the three specifications. Table 5 summarizes the marginal effects across specifications and benchmarks them to an standard OLS regresssion.

The marginal effects for the constant term in the HME are not comparable to the OLS model. These marginal effects include the constant terms in the gating network, invalidating them as an intercept term. The return to education does appear to be materially lower for the HME models as compared to the standard OLS while the quadratic wage curve is similar in the *full* specification but somewhat different in the *mid* and *minimal* specifications. When it comes to the coefficients for race, a bit of variations starts to appear across specifications. Interestingly, the *mid* specification seems to deviate the most from the OLS benchmark while the *minimal*

Table 3: Model Information

				Model Comparison			
Model	Stop Criterion	Depth	Experts	Log-Lik	AIC	BIC	MSE
Full	Log-Lik	1	2	-0.540	1.081	1.086	0.182
		2	3	-0.524	1.050	1.059	0.181
		2	4	-0.515	1.034	1.047	0.181
		3	5	-0.505	1.015	1.031	0.178
		3	6	***-0.503***	***1.011***	***1.031***	***0.178***
	MSE	1	2	-0.540	1.081	1.086	0.182
		2	3	-0.524	1.050	1.059	0.181
		2	4	-0.515	1.034	1.047	0.181
		3	5	-0.505	1.015	1.031	0.178
		3	6	*-0.503*	*1.011*	*1.031*	***0.178***
Mid	Log-Lik	1	2	-0.559	1.119	1.122	0.185
		2	3	-0.540	1.082	1.088	0.184
		2	4	-0.527	1.056	1.064	0.183
		3	5	*-0.518*	*1.039*	*1.049*	0.182
		3	6	-0.519	1.041	1.053	*0.181*
	MSE	1	2	-0.559	1.119	1.122	0.185
		2	3	-0.540	1.082	1.088	0.184
		2	4	-0.529	1.060	1.067	0.183
		3	5	*-0.518*	*1.039*	*1.050*	0.182
		3	6	-0.519	1.041	1.053	*0.181*
Min	Log-Lik	1	2	-0.595	1.192	1.195	0.192
		1	3	-0.552	1.106	1.111	0.183
		2	3	-0.580	1.162	1.168	0.190
		2	4	-0.546	1.093	1.101	0.182
		3	5	-0.523	1.049	1.058	0.182
		3	6	*-0.521*	*1.045*	*1.057*	*0.181*
	MSE	1	2	-0.596	1.192	1.195	0.192
		2	3	-0.580	1.162	1.168	0.190
		2	4	-0.546	1.094	1.101	0.182
		3	5	*-0.523*	*1.049*	*1.059*	0.182
		3	6	-0.531	1.064	1.076	*0.181*

Log-Likelihood and BIC are divided by the sample size (54,829)

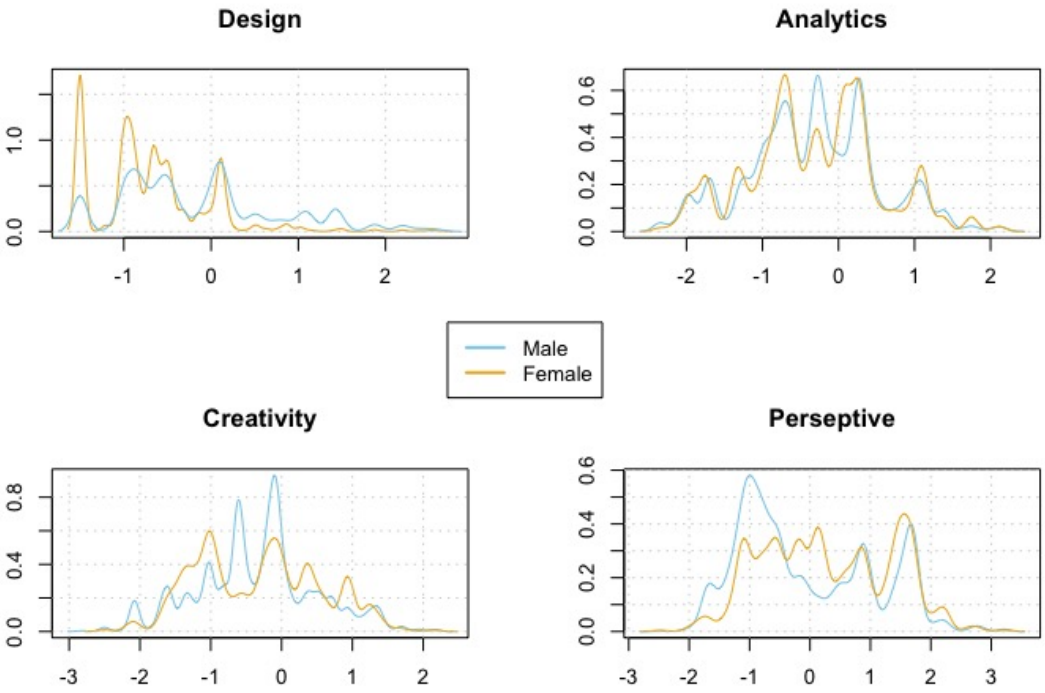

Figure 3: Density estimates of ONet job characteristics broken down by sex. The job characteristics have been mean centered and scaled to have unit variance.

specification maintains comparable marginal effects. There doesn't seem to be an obvious explanation for this discrepency.

10 Miscellaneous

10.1 Odd Facts

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

Table 4: Returns to Years of Education

OLS: 0.096		Coefficient		
Depth	Experts	Min	Mid	Full
1	2	0.051	0.083	0.072
1	3	–	–	–
1	4	–	–	–
1	5	–	–	–
1	6	–	–	–
2	3	0.048	0.082	0.072
2	4	0.063	0.078	0.073
3	5	0.068	0.076	0.070
3	6	0.067	0.075	0.070

OLS coef: 0.96

Log-Likelihood is divided by the sample size (54,829)

10.2 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 $mathbf{\hat{\beta}}$ can be much less than near $\beta_i = 0$, and so the Walk approximation understimates the change in log-likelihood on setting $\beta_i = 0$. This happens in such a way that as $|\hat{\beta}_i| \to \infty$. 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.

11 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

Table 5: Full Marginal Effects

	OLS	Full	Mid	Min
Const.	0.819	1.036	1.035	1.057
	0.013**	–	–	–
Yrs Edu	0.096	0.091	0.091	0.088
	0.008**	–	–	–
Age	0.044	0.044	0.041	0.041
	0.001**	–	–	–
Age Sq	-0.0006	-0.0007	-0.0006	-0.0006
	0.0000**	–	–	–
Afr Amer	-0.175	-0.166	-0.120	-0.155
	0.007**	–	–	–
Indian	-0.118	-0.091	-0.098	-0.116
	0.019**	–	–	–
Hispanic	-0.165	-0.174	-0.117	-0.132
	0.006**	–	–	–
Asian	-0.070	-0.049	-0.043	-0.070
	0.010**	–	–	–

[1] Each non-OLS model is a five-expert HME.

Table 6: Log Wage Expert Equations - Full Specification

	OLS	1	2	3	4	5
Share	–	0.625	0.193	0.117	0.046	0.018
Const.	0.819	1.048	1.493	1.687	1.341	1.643
	0.013**	0.014**	0.010**	0.014**	0.084**	0.013**
Yrs Edu	0.096	0.105	0.032	0.034	0.040	0.0126
	0.008**	0.001**	0.001**	0.001**	0.005**	0.0011**
Age	0.044	0.024	0.0616	0.003	0.044	0.011
	0.001**	0.001**	0.001**	0.0006**	0.004**	0.0008**
Age Sq	-0.0006	-0.0003	-0.001	0.00008	-0.0006	-0.0005
	0.0000**	0.00001**	0.00001**	0.00001**	0.00008**	0.0000**
Afr Amer	-0.175	-0.209	-0.135	-0.036	-0.022	-0.078
	0.007**	0.007**	0.006**	0.006**	0.045	0.007**
Indian	-0.118	-0.131	-0.181	-0.122	0.654	0.432
	0.019**	0.018**	0.016**	0.014**	0.138**	0.014**
Hispanic	-0.165	-0.108	1.590	-0.032	-1.037	-0.041
	0.006**	0.006**	0.041**	0.004**	0.072**	0.005**
Asian	-0.070	-0.051	0.059	-0.068	-0.468	0.145
	0.010**	0.010**	0.008**	0.009**	0.076**	0.013**
log Var	–	-1.824	-2.512	-2.776	-0.409	-4.586
	–	0.008**	0.015**	0.022**	0.005**	0.095**

[1] Five-expert *full* HME.

Table 7: Log Wage Expert Equations - Mid specification

	OLS	1	2	3	4	5
Share	–	0.632	0.168	0.126	0.056	0.017
Const.		1.021	1.291	1.676	1.449	1.684
		0.013**	0.013**	0.011**	0.073**	0.011**
Yrs Edu		0.102	0.059	0.014	0.040	0.012
		0.001**	0.001**	0.001**	0.005**	0.001**
Age		0.029	0.010	0.072	0.036	-0.032
		0.001**	0.001**	0.001**	0.004**	0.001**
Age Sq		-0.0004	-0.0002	-0.001	-0.0005	0.005
		0.0001**	0.0000**	0.0000**	0.0001**	0.000**
log Var	–	-1.886	-2.493	-2.772	-0.435	-4.571
	–	0.008**	0.017**	0.021**	0.005**	0.085**

[1] Five-expert *mid* HME.

AIC/BIC/VOUNG test being prefered.

References

Anderson, Edgar (1936). "The species problem in iris". In: *Annals of the Missouri Botanical Gardens* 23.3, pp. 457–509.

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*. Wadswoth 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.

Carvalho, Alexandre and Georgios Skoulakis (2010). "Time Series Mixutres of Generalized t Experts: ML Estimation and an Application to stock return density fore-

casting". 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.

Fisher, R.A. (1936). "The use of multiple measurements in taxonomic problems". In: *Annals of Eugenics* 7.2, pp. 179–188.

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: *Nueral 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). "Hierarchhical 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: *Nueral 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.

*Occupational Information Network (O*NET)* (2019). URL: `https://www.doleta.gov/programs/onet/` (visited on 01/28/2019).

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: discoverging regimes and avoiding overfitting". In: *International Journal of Neural Systems* 6, pp. 373–399.