Econometric Applications of Hierarchical Mixture of Experts

Lucas C. Dowiak

November 3, 2021

PhD Program in Economics, City University of New York, Graduate Center, New York, NY, 10016, *Email: ldowiak@gradcenter.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 field. Hamilton (1989) and Goldfeld and Quandt (1973) are a few of the pioneering works for time series and cross sectional regression, respectively. Today, the modern computing environment is dominated by machine learning, and its reigning champion, the artificial neural network, has been successfully adapted and studied in the context of applied econometrics. This essay adds to the small body of literature that employs a novel neural network architecture to model the weights of a mixture model. In doing so, the model leverages 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 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*. The authors employ this mixture of experts (ME) framework to model vowel discrimination in a speech recognization context. Shortly after, Jordan and Jacobs (1992) generalize this single-layer multinomial gating network to one with an arbitrary number of layers. Jordan and Jacobs (1993) then demonstrate an Expectation-Maximization approach to model estimation that is capable of handling the additional complexity the generalization requires during optimization. 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. 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. In the case of a symmetric HME network, this growth is geometric with respect to the network's depth. With this in mind, a further extension can be considered where each expert is not unique, but a member of a fixed set of experts. This additional model is referred to as a Hierarchical Mixture of Repeated Experts (HMRE). Figure (1) provides an example of each of the variations of this class of model.

This essay investigates the adoption of ME and HME 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 essay is as follows: the remainder of this section provides a brief literature review and Section 2 describes the model in formal detail. Section 3 discusses the expectation-maximization approach to estimation while Section 4 concerns itself with robust inference of the estimated parameters. Section 5 provides detail on how to derive the marginal effects of the model's covariates. In Section 6, a vary simple demonstration of the HME in action is presented with a more economically relevant example of applying the HME model to a Mincer wage equation in Section 7. Section 8 concludes.

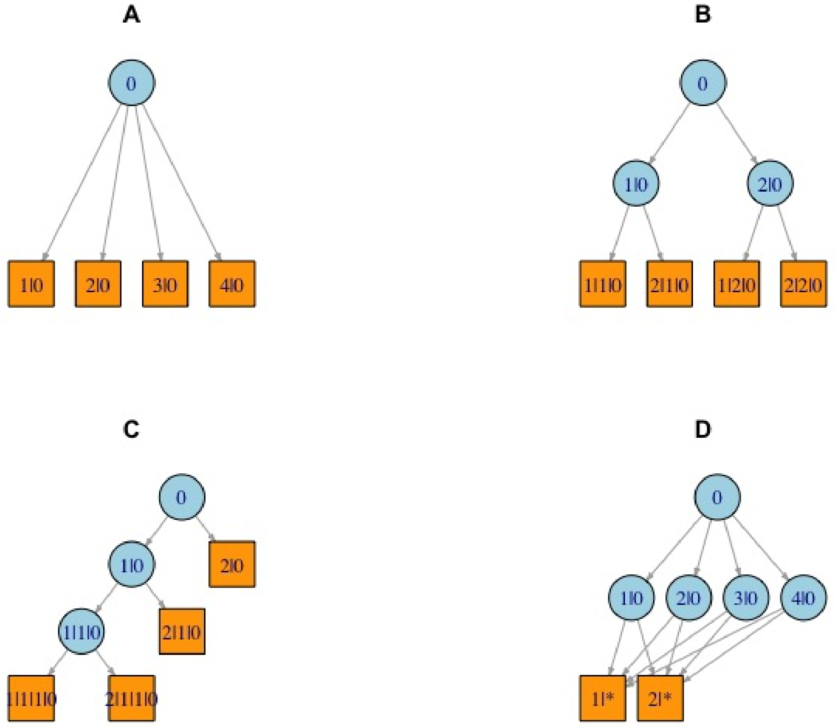


Figure 1: Networks **A** - **D** depict various network architectures that are discussed in this article. For all four networks, gating nodes are represented as blue circles and experts as orange rectangles. Network **A** illustrates the original Mixture of Experts (ME) architecture with a single multinomial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

1.1 Relevant Literature

ME and HME frameworks have been utilized for both time series and cross-sectional analysis. Within the cross-sectional 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 popular 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 suggest an additional trimming algorithm to prevent overfitting. Fritsch, Finke, and Waibel (1997) extend the approach of Waterhouse and Robinson (1995) by altering their growing algorithm with a mind to scaling the model to handle thousands of experts. Jordan and Xu (1995) provide an extended discussion on the convergence of the model used by Jordan and Jacobs (1993). The authors also suggest algorithmic improvements to help with estimation. Continuing the theoretical discussing, Jiang and Tanner (1999) cover convergence rates of an HME model where experts are from the exponential family with generalized linear mean functions. Jiang and 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.

Alternatively, Weigend, Mangeas, and Srivastava (1995) provide a detailed discussion 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. The authors' formulation of the model has close similarities to other non-linear time series models developed in the late 1980's and early 1990's. A ME time series model sits between the markov-switching (MS) model of Hamilton (1989) and the smooth transition auto-regressive (STAR) model of Terasvirta (1994), borrowing a bit from both. From an estimation perspective, the ME time series follows close to the markov-switching model due to the fact that they are both mixture distribution where each (conditional) distribution represents a different "state" of nature. The STAR model, on the other hand, posits only a single distribution and different "states" are represented by unique parameter vectors, and as the name implies, the parameters transition smoothly from one state to another over time. The association between the ME, MS, and STAR models is inverted when it comes to how to frame the time evolution of the states. From this perspective, the ME model is very similar to a STAR model in that it also uses the logistic (or multinomial) function to force the probability of belonging to one state to change over time. For MS models, a discrete state markov process is used to model this dynamic change over the time dimension. Huerta, Jiang, and Tanner (2003) extend (Weigend, Mangeas, and Srivastava, 1995) to an HME framework. Using five and a half decades of monthly US industrial production data, the authors allow 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 Tanner (2003) lay out the necessary regularity conditions to perform hypothesis tests on stationary ME time series of generalized linear models (ME-GLM) using Wald tests. The dual cases of a well-specified and a miss-specified model are considered. The authors restrict their analysis to ME-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 Tanner (2005) take a similar approach to Carvalho and Tanner (2003) but apply their analysis to a purely auto-regressive context restricted to Gaussian models. The authors extend arguments in Carvalho and Tanner (2003) to non-stationary series and provide simulated evidence that the BIC is a helpful statistic for selecting the appropriate number of experts to include. Carvalho and Tanner (2006) re-focus the discussion on ME 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 ME 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.

In this essay estimation and inference is from a maximum likelihood perspective and will remain the primary focus. Estimation of ME and HME models from a Bayesian has received considerable amount of attention as well. Waterhouse, MacKay, and Robinson (1995) provided an initial approach to estimating a ME by combining gaussian priors on the gating and expert parameters with gamma hyper-parameter priors in an approximating ensemble to the true joint density of the model. Optimization of the parameter vector for the approximating density occurs a block of parameters at a time. Ueda and Ghahramani (2002) improve on Waterhouse, MacKay, and Robinson (1995) by optimizing for the appropriate number of experts in addition to model parameters. Bishop and Svenson (2003) find previous bayesian approaches to estimating an HME lacking. Using variational inference, the authors provide a complete bayesian estimation approach to the log marginal likelihood. With an eye to prediction, the authors advocate that their approach makes the HME model easier to estimate without overfitting.

2 Model

To start, the HME is presented as a standard mixture model. For a given input and output pair $(\boldsymbol{x}_t, y_t)$, each expert provides a probabilistic model relating input row $\boldsymbol{x}_t$ to output y_t:

$$P_t^m \equiv P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^m), \quad m = 1, 2, ..., M \tag{1}$$

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

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

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

$$P(y_t|\boldsymbol{x}_t, \boldsymbol{z}_t;\, \boldsymbol{\beta}, \boldsymbol{\Omega}) = \sum_{m=1}^{M} \mathbb{\Pi}(m|\boldsymbol{z}_t; \boldsymbol{\Omega}) P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^m) \tag{3}$$

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

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 arbitrary gating node at address a, we use a multinomial logistic regression to model the split in direction i to be:

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

The parameters in equation (4) are subject to the usual identification restrictions. For the remainder of this essay, the choice is made 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 suppressed, the product path from one node (say the root node 0) to another (say $k|\dots|j|i$) can be defined as

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

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:

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

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

$$\mathbb{\Pi}(m|\boldsymbol{Z}, \boldsymbol{\Omega}) = \sum\nolimits_l \pi_{g^0 \xrightarrow{l} m} \tag{7}$$

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:

$$\begin{aligned} P(y_t|\boldsymbol{x}_t, \boldsymbol{z}_t; \boldsymbol{\Omega}, \boldsymbol{\beta}) &= \sum_m \mathbb{\Pi}(m|\boldsymbol{z}_t, \boldsymbol{\Omega}) P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^m) \\ &= \sum_m \left(\sum\nolimits_l \pi_{g^0 \xrightarrow{l} m} \right) P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^m) \end{aligned} \tag{8}$$

If we concatenate the parameters of the gating network with the parameters of the experts as $\boldsymbol{\theta} = [\boldsymbol{\beta}\ \boldsymbol{\Omega}]$, then the product of these individual probabilities across the full sample size T yields the likelihood function.

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = \prod_t \sum_m \left(\sum\nolimits_l \pi_{g^0 \xrightarrow{l} m} \right) P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^m) \tag{9}$$

And taking its log yields the log likelihood

$$\boldsymbol{l}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = \sum_t \log \left[\sum_m \left(\sum\nolimits_l \pi_{g^0 \xrightarrow{l} m} \right) P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^{\boldsymbol{m}}) \right] \tag{10}$$

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 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 $\boldsymbol{x}_t$ 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

$$\mathcal{L}_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = \prod_t \prod_m \left[\left(\sum\nolimits_l \pi_{g^0 \xrightarrow{l} m} \right) P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^m) \right]^{I_t(m)} \tag{12}$$

leading to the complete-data log-likelihood

$$\boldsymbol{l}_c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = \sum_t \sum_m I_t(m) \left[\log \left(\sum\nolimits_l \pi_{g^0 \xrightarrow{l} m} \right) + \log P^m(y_t|\boldsymbol{x}_t; \boldsymbol{\beta}^m) \right] \tag{13}$$

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 equals 1, simplifying the first log in (13) to $\log(\pi_{g^0 \longrightarrow m})$. Going forward, we will focus our analysis on the ME and HME specifications with work on the HMRE case left for another time.

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:

$$
\begin{aligned}
Q(\boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{l}_c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{X},\boldsymbol{Z})\right] &= \sum_t \sum_m \mathbb{E}\left[I_t(m)\right]\left[\log P^m(y_t|\boldsymbol{x}_t;\boldsymbol{\beta}^m) + \log \pi_{g_t^0 \longrightarrow m}\right] \\
&= \sum_t \sum_m \mathbb{E}\left[I_t(m)\right] \log P^m(y_t|\boldsymbol{x}_t;\boldsymbol{\beta}^m) + \sum_t \sum_m \sum_a \mathbb{E}\left[I_t(a)\right] \log \pi_{g_t^0 \longrightarrow a} \\
&= \sum_t \sum_m \pi_{h_t^0 \longrightarrow m} \log P^m(y_t|\boldsymbol{x}_t;\boldsymbol{\beta}^m) + \sum_t \sum_m \sum_a \pi_{h_t^0 \longrightarrow a} \log \pi_{g_t^0 \longrightarrow a} \\
&= \sum_t Q_t^{(1)}(\boldsymbol{\beta}) + \sum_t Q_t^{(2)}(\boldsymbol{\Omega}) \\
&= \sum_t Q_t(\boldsymbol{\theta})
\end{aligned}
\tag{14}
$$

Here $\pi_{h_t^0 \longrightarrow k,\ldots|j|i|0}$ is analogous to equation (5)

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

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

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

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, Jordan and Jacobs (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

One of the more attractive features of using EM to a optimize a HME is how the log-likelihood function compartmentalizes into a set of independent functions which can be individually optimized. After taking the expectation of the log-likelihood function (14), the parameters governing each expert and each gating network can be grouped together and optimized on their own. For the experts we have:

$$
\arg\max_{\boldsymbol{\beta^m}} \sum_t \pi_{h_t^0 \longrightarrow m} \log P^m(y_t|\boldsymbol{x}_t;\boldsymbol{\beta}^m)
\tag{17}
$$

and for the gating nodes:

$$\arg\max_{\boldsymbol{\omega^a}} \sum_t \pi_{h_t^0 \longrightarrow a} \log g(\boldsymbol{z}_t, \boldsymbol{\omega}^a) \tag{18}$$

It is worth noting that the weights in these optimizations $\pi_{h_t^0 \longrightarrow h_t^a}$ are provided to the M-step by the E-step and should be considered constant values.

3.3 The EM-Algorithm

The EM algorithm iterates back-and-forth between the E-step and the M-step. Given the data ($\boldsymbol{y}_t$, $\boldsymbol{X}_t$, $\boldsymbol{Z}_t$) and the current set of parameters $\boldsymbol{\theta}^k$, the expected value of the complete log-likelihood (eq. (13)) is found, resulting in the deterministic function $\mathrm{Q}(\boldsymbol{\theta}^k)$. In essence, the main objective of the E-step is to derive the values of the posterior weights ($h_t^{a,i}$) using equations (1), (4), (5), (15) and (16). Once the posterior weights have been calculated in the E-step, the M-step holds them constant and then re-estimates the parameter vector:

$$\boldsymbol{\theta}^{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\hat{\boldsymbol{\theta}}^k) = \begin{bmatrix} \arg\max_{\boldsymbol{\beta}} Q^{(1)}(\hat{\boldsymbol{\beta}}^k) & \arg\max_{\boldsymbol{\Omega}} Q^{(2)}(\hat{\boldsymbol{\Omega}}^k) \end{bmatrix} \tag{19}$$

Again, due to the separable nature of Q (see the middle equality of eq (14)), the parameters of each expert regression and each gating node can be updated one-at-a-time with equations (17) and (18), respectively. The separability of the Q function – when applied to finite mixture – was noticed in the original and excellent work of Dempster, Laird, and Rubin (1977). See Section 4.3 for the authors' example. From a computational perspective, this set-up has the additional benefit of being embarrassingly parallel, making it easier to scale to larger and larger data sets.

It is worth mentioning a few more of the remarkable properties of the EM algorithm that are also established in Dempster, Laird, and Rubin (1977):

1. Given a sequence of parameter values produced by the General EM algorithm, $\boldsymbol{\theta}^k \rightarrow \boldsymbol{\theta}^{k+1} \rightarrow \ldots \rightarrow \boldsymbol{\theta}^{k+n}$, the sequence of values are non-decreasing in their log-likelihood values $\boldsymbol{l}(\boldsymbol{\theta}^k|\cdot) \leq \boldsymbol{l}(\boldsymbol{\theta}^{k+1}|\cdot) \leq \ldots \leq \boldsymbol{l}(\boldsymbol{\theta}^{k+n}|\cdot)$ and are strictly increasing in the Q function $Q(\boldsymbol{\theta}^k) < Q(\boldsymbol{\theta}^{k+1}) < \ldots < Q(\boldsymbol{\theta}^{k+n})$.

2. The sequence of parameter values produced by the General EM algorithm converges to a fixed point such that in the limit:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^*) \tag{20}$$

Crucially, the vector that the general EM algorithm converges to is a maximum likelihood estimator of the original log-likelihood equation defined in (10). That is, $\boldsymbol{l}(\boldsymbol{\theta}^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) \geq \boldsymbol{l}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

4 Inference

When considering inference, it is worth thinking about what would motivate a researcher to turn to an HME model in the first place. At times, a researcher may suspect that a latent structure exists within the data and that a single regression $y_t = \boldsymbol{x}_t\boldsymbol{\beta}$ may mask a critical change in relationship depending on membership to some unknown sub-group m of the data $y_{t,m} = \boldsymbol{x}_t\boldsymbol{\beta}^m$. A wide variety of time series, especially those with longer histories, experience changes in behavior over time. They can be subjected to sharp one-off structural breaks or the changes can be more gradual changes over time. Regardless of the context, any latent structural change in the data generating process may also introduce some hidden form of heterogeneity to the error terms. Rather than taking a firm stance on any concealed structure, an HME setup ideally limits the work the researcher needs to do to specifying a set of well-chosen conditioning variables $\boldsymbol{Z}$ to feed through the gating network. This limited workload may come at a cost, though. By allowing the gating network to find its own mixture allocations, the odds of arriving at a misspecified model becomes a concern. To guard against this, we use a sandwich estimator for the variance-covariance matrix:

$$\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{H}^{-1}(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{H}^{-1}(\boldsymbol{\theta}) \tag{21}$$

where $\boldsymbol{G}(\boldsymbol{\theta})$ is the sum of the outer products of the score vectors

$$\boldsymbol{G}(\boldsymbol{\theta}) = \sum_t \boldsymbol{S}_t(\boldsymbol{\theta})\boldsymbol{S}_t(\boldsymbol{\theta})^\top \tag{22}$$

and $\boldsymbol{H}(\boldsymbol{\theta})$ is the empirical Hessian:

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{1}{T}\sum_t \mathbf{H}_t(\boldsymbol{\theta}) \tag{23}$$

The following sections discusses how to form the score and hessian matrices for the log-likelihood described in equation (10).

4.1 The Score

The notation is tedious but the acyclic nature of the gating network makes interpretation of the score vectors clear and straightforward. The full score vector is the concatenated scores of each gating node and those of each local expert.

$$\boldsymbol{S}_t(\boldsymbol{\theta}) = [\boldsymbol{S}_t(\boldsymbol{\beta})^\top \ \boldsymbol{S}_t(\boldsymbol{\Omega})^\top]^\top \tag{24}$$

Starting with parameters of the gating network, they can be partitioned in some logical order into the sub-vectors of each node's individual score.

$$\boldsymbol{S}_t(\boldsymbol{\Omega}) = [\boldsymbol{S}_t(\boldsymbol{\omega}^0)^\top \ \boldsymbol{S}_t(\boldsymbol{\omega}^{1|0})^\top \ \boldsymbol{S}_t(\boldsymbol{\omega}^{2|0})^\top \ \ldots]^\top \tag{25}$$

$$\boldsymbol{S}_t(\boldsymbol{\omega}^a) = [\boldsymbol{S}_t(\boldsymbol{\omega}^{a,1})^\top \ \ldots \ \boldsymbol{S}_t(\boldsymbol{\omega}^{a,J-1})^\top]^\top \tag{26}$$

In what follows, the functions $m(a)$ and $m(a,i)$ will be used to return a subset of experts from a general HME model. The function $m(a)$ will return the set of all experts that are ancestors of node a, while $m(a,i)$ returns the set of experts that are ancestors from branch i of node a. For instance, in network $\boldsymbol{C}$ of Figure 1, $m('1|0') = \{'1|1|1|0', '2|1|1|0', '2|1|0'\}$, $m('1|0', 1) = \{'1|1|1|0', '2|1|1|0'\}$, and $m('1|0', 2) = \{'2|1|0'\}$. For a generic gating node a we can define the individual score for sample t as:

$$\boldsymbol{S}_t(\boldsymbol{\omega}^{a,i}) = \frac{\partial \boldsymbol{l}_t(\boldsymbol{\theta}^* | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})}{\partial \boldsymbol{\omega}^{a,i}} = \left[\frac{\Omega_t^{(0)}(a,i)}{\sum_m \pi_{g_t^0 \longrightarrow m} P_t^m} \right] \boldsymbol{z}_t^\top \tag{27}$$

with

$$\Omega_t^{(0)}(a,i) = \left((1 - g_t^{a,i}) \sum_{k}^{m(a,i)} \pi_{g_t^0 \longrightarrow k} P_t^k - \sum_{j \neq i} g_t^{a,j} \sum_{k'}^{m(a,j)} \pi_{g_t^0 \longrightarrow k'} P^{k'} \right) \tag{28}$$

In expression (27) above, $\Omega_t^{(0)}(a,i) \left[\sum_m \pi_{g_t^0 \longrightarrow P^m}\right]^{-1}$ is the instantaneous rate of change of the t^{th} contribution to the log-likelihood caused by a small perturbation of $\boldsymbol{\omega}^{a,i}$. At the maximum likelihood estimator θ^*, the sum of (28) should be approximately zero. This implies that over the sample T, the optimal $\boldsymbol{\omega}^{a,i}$ balances any gain of moving more weight to the set of experts that can be reached by taking direction i at node a against the loss suffered by removing weight from the experts at the end of any path j that does not equal i.

Turning our attention to the expert regressions, the exact functional form of the score vector depends on the type of regression we wish to run. In most cases, all experts in an HME model are from the same family (Huerta, Jiang, and Tanner (2003) is a notable exception). When all experts share the same functional form, it is standard to accept the restriction that no experts in the HME model produce the same parameter vector $\boldsymbol{\beta}^m \neq \boldsymbol{\beta}^k$. Such an HME is defined by Jiang and Tanner (2000) as being *irreducible*. The irreducibility of an HME plays a critical role in guaranteeing the convergence of the model. In this essay, each HME discussed will employ a set of experts running a standard linear regression model with Gaussian errors. To aid with model optimization, the specification of the parameter vector for each regression, $\boldsymbol{\beta}^m = [\beta_0^m \ \ldots \ \beta_k^m \ \phi^m]^\top$, takes on a unique form where we model the log variance explicitly: $\phi = \log \sigma^2$.

$$P^m(y_t|\boldsymbol{x}_t;\, \boldsymbol{\beta}^m, \phi^m) = (2\pi \exp(\phi^m))^{-\frac{1}{2}} \exp\left(-\frac{(y_t - \boldsymbol{x}_t\boldsymbol{\beta}^m)^2}{2\exp(\phi^m)}\right) \tag{29}$$

To help save space in the sections below, the following shorthand will be used to denote the residual of each local expert: $\epsilon_t^m = y_t - \boldsymbol{x}_t\boldsymbol{\beta}^m$. In this case the score vector for all expert regressions can be expressed as:

$$\boldsymbol{S}_t(\boldsymbol{\beta}) = [\boldsymbol{S}_t(\boldsymbol{\beta}^1)^\top \ \ldots \ \boldsymbol{S}_t(\boldsymbol{\beta}^M)^\top]^\top \tag{30}$$

$$\boldsymbol{S}_t(\boldsymbol{\beta^m}) = \left[\frac{\partial \boldsymbol{l}_t}{\partial \boldsymbol{\beta}^m}^\top \quad \frac{\partial \boldsymbol{l}_t}{\partial \phi^m}\right]^\top \tag{31}$$

with

$$\frac{\partial \boldsymbol{l}_t}{\partial \boldsymbol{\beta}^m} = \pi_{g_t^0 \longrightarrow m} \frac{\epsilon_t^m}{\exp(\phi^m)} \boldsymbol{x}_t^\top \tag{32}$$

and

$$\frac{\partial \boldsymbol{l}_t}{\partial \phi^m} = \frac{1}{2}\pi_{g_t^0 \longrightarrow m}\left(\frac{(\epsilon_t^m)^2}{\exp(\phi^m)} - 1\right) \tag{33}$$

Expressions (32) and (33) are the same as the score vectors for a single OLS regression but with the associated prior weights added on.

4.2 The Hessian

The hessian, admittedly, has a complicated form. At its most general it can be written as $\mathbf{H}_t(\boldsymbol{\theta})$ in the equation below. The exact nature of the full hessian depends critically on the structure of the gating network and the locations of the gate and expert nodes in relation to each other.

$$\mathbf{H}_t(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{H}_t(\boldsymbol{\beta}^1) & \mathbf{0} & \cdots & \mathbf{H}_t(\boldsymbol{\beta}^1,\, \boldsymbol{\omega}^{0,1}) & \cdots & \mathbf{H}_t(\boldsymbol{\beta}^1,\, \boldsymbol{\omega}^{\cdot,J-1}) \\ \mathbf{0} & \mathbf{H}_t(\boldsymbol{\beta}^2) & \cdots & \mathbf{H}_t(\boldsymbol{\beta}^2,\, \boldsymbol{\omega}^{0,1}) & \cdots & \mathbf{H}_t(\boldsymbol{\beta}^2,\, \boldsymbol{\omega}^{\cdot,J-1}) \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \mathbf{H}_t(\boldsymbol{\beta}^1,\, \boldsymbol{\omega}^{0,1})^\top & \mathbf{H}_t(\boldsymbol{\beta}^2,\, \boldsymbol{\omega}^{0,1})^\top & \cdots & \mathbf{H}_t(\boldsymbol{\omega}^{0,1}) & \cdots & \mathbf{H}_t(\boldsymbol{\omega}^{0,1},\, \boldsymbol{\omega}^{\cdot,J-1}) \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ \mathbf{H}_t(\boldsymbol{\beta}^1,\, \boldsymbol{\omega}^{\cdot,J-1})^\top & \mathbf{H}_t(\boldsymbol{\beta}^2,\, \boldsymbol{\omega}^{\cdot,J-1})^\top & \cdots & \mathbf{H}_t(\boldsymbol{\omega}^{0,1},\, \boldsymbol{\omega}^{\cdot,J-1})^\top & \cdots & \mathbf{H}_t(\boldsymbol{\omega}^{\cdot,J-1}) \end{bmatrix} \tag{34}$$

Before getting into the details of matrix (34), it will be worth noting the parameter combinations that will result in zero-valued elements of $\boldsymbol{H}_t$. First, there are no second-order partial derivatives across experts, meaning that for any experts m, k where $m \neq k$, then $\frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\beta}^m \partial \boldsymbol{\beta}^k} = \mathbf{0}$. This can already be seen in equation (34). Not noted in (34) are two additional cases where elements of $\boldsymbol{H}_t$ will equal zero. One, if the ancestors of two gating nodes have no shared experts. That is, for nodes a and b, if $m(a) \cap m(b) = \{\}$, then $\frac{\partial^2 \boldsymbol{l}_t}{\partial \boldsymbol{\omega}^{a,i} \partial \boldsymbol{\omega}^{b,n}} = 0$ for all i, n. Two, if the gating path to expert m does not go through node a then there are also no second-order partial derivatives between any parameters in node a and those that appear in gating path to expert m or in the parameter vector $\boldsymbol{\beta}^m$. That is, if g^a is not a component of $\pi_{g_t^0 \longrightarrow m} P^m$, then $\frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\beta}^m \partial \boldsymbol{\omega}^{a,i}} = \mathbf{0}$ and $\frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\omega}^{a,i} \partial \boldsymbol{\omega}^{b,n}} = \mathbf{0}$ for all g^b in $\pi_{g_t^0 \longrightarrow m}$.

The expressions for the cross-partial derivatives between a gating parameter vector and an expert parameter vector can differ based on the relative position between $\boldsymbol{\omega}^{a,i}$ and $\boldsymbol{\beta}^m$ in the HME structure. For instance, start at the root node and consider what path is needed to traverse the network to expert m. When arriving at node a (which is on the path to expert m), if the direction needed to take to reach expert m is along branch i, then $\boldsymbol{\omega}^{a,i}$ will be called an *explicit* parameter vector with respect to expert m. If taking branch i leads to a different expert than m, then $\boldsymbol{\omega}^{a,i}$ will be referred to as an *implicit* parameter vector. Now, define $\mathbb{1}\{a, i, m\}$ as an indicator function that equals one if $\boldsymbol{\omega}^{a,i}$ is an explicit parameter vector to expert m and zero if it is an implicit parameter vector (it can only be one or the other). With this notation, the details to the hessian in equation (34) can now be tackled. Starting with equation (27), the second-order partial derivatives for a pair of gating vectors is:

$$\frac{\partial^2 \boldsymbol{l}_t}{\partial \boldsymbol{\omega}^{a,i} \partial \boldsymbol{\omega}^{b,n}} = \left[\sum_m \pi_{g_t^0 \longrightarrow m}\right]^{-2} \Omega_t^{(0)}(a,i) \cdot \Omega_t^{(0)}(b,n)\, \boldsymbol{z}_t^\top \boldsymbol{z}_t + \left[\sum_m \pi_{g_t^0 \longrightarrow m}\right]^{-1} \boldsymbol{z}_t^\top \, \frac{\partial \Omega_t^{(0)}(a,i)}{\partial \boldsymbol{\omega}^{b,n}}^\top \tag{35}$$

where

$$\frac{\partial \Omega_t^{(0)}(a,i)}{\partial \boldsymbol{\omega}^{b,n}} = \sum_k^{m(a) \cap m(b)} \sum_j^{J-1} \sum_l^{J-1} \left\{ \left(\mathbb{1}\{a,j,k\} - g_t^{a,j}\right) \left(\mathbb{1}\{b,l,k\} - g_t^{b,l}\right) \cdot \Omega_t^{(1)}(a,j)(b,l) \cdot P^k \right\} \boldsymbol{z}_t^\top \tag{36}$$

and

$$\Omega_t^{(1)}(a,i)(b,n) = \begin{cases} \sum_m^{m(a,i) \cap m(b,n)} \pi_{g_t^0 \longrightarrow m} & \text{if } m(a,i) \cap m(b,n) \neq \{\} \\ 0 & \text{if } m(a,i) \cap m(b,n) = \{\} \end{cases} \tag{37}$$

Starting from equations (32) and (33), the next set of equations express the second-order partial derivatives for parameters of each individual expert:

$$\frac{\partial \boldsymbol{l}_t^2}{\partial (\boldsymbol{\beta}^m)^2} = \frac{\pi_{g_t^0 \longrightarrow m}}{\exp\left(\phi^m\right)} \boldsymbol{x}_t^\top \boldsymbol{x}_t \tag{38}$$

$$\frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\beta}^m \partial \phi^m} = -\pi_{g_t^0 \longrightarrow m} \frac{\epsilon_t^m}{\exp\left(\phi^m\right)} \boldsymbol{x}_t^\top \tag{39}$$

$$\frac{\partial \boldsymbol{l}_t^2}{\partial (\phi^m)^2} = \frac{\pi_{g_t^0 \longrightarrow m} (\epsilon_t^m)^2}{2 \exp\left(\phi^m\right)} \tag{40}$$

And to round it out, the expression for the cross-partial derivatives between a gating parameter vector and an expert parameter vector can be expressed as:

$$\frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\beta}^m \partial \boldsymbol{\omega}^{a,i}} = \pi_{g_t^0 \longrightarrow m} \left(\mathbb{1}\{a,i,m\} - g_t^{a,i}\right) \frac{\epsilon^m}{\exp\left(\phi^m\right)} \boldsymbol{x}_t^\top \boldsymbol{z}_t \tag{41}$$

$$\frac{\partial \boldsymbol{l}_t^2}{\partial \phi^m \partial \boldsymbol{\omega}^{a,i}} = \frac{1}{2} \pi_{g_t^0 \longrightarrow m} \left(\mathbb{1}\{a,i,m\} - g_t^{a,i}\right) \left(\frac{(\epsilon^m)^2}{\exp\left(\phi^m\right)} - 1 \right) \boldsymbol{z}_t^\top \tag{42}$$

Given the collection of equations (35) - (42), the individual elements of the hessian in equation (34) can be expressed as:

$$\mathbf{H}_t(\boldsymbol{\beta}^m) = \begin{bmatrix} \frac{\partial \boldsymbol{l}_t^2}{\partial (\boldsymbol{\beta}^m)^2} & \frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\beta}^m \partial \phi^m} \\ \frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\beta}^m \partial \phi^m}^\top & \frac{1}{2}(\epsilon_t^m)^2 \end{bmatrix} \tag{43}$$

$$\mathbf{H}_t(\boldsymbol{\beta}^m,\, \boldsymbol{\omega}^{a,i}) = \begin{bmatrix} \frac{\partial \boldsymbol{l}_t^2}{\partial \boldsymbol{\beta}^m \partial \boldsymbol{\omega}^{a,i}} & \frac{\partial \boldsymbol{l}_t^2}{\partial \phi^m \partial \boldsymbol{\omega}^{a,i}} \\ \frac{\partial \boldsymbol{l}_t^2}{\partial \phi^m \partial \boldsymbol{\omega}^{a,i}}^\top & \frac{\partial \boldsymbol{l}_t^2}{\partial (\phi^m)^2} \end{bmatrix} \tag{44}$$

$$\mathbf{H}_t(\boldsymbol{\omega}^{a,i},\, \boldsymbol{\omega}^{b,n}) = \frac{\partial^2 \boldsymbol{l}_t}{\partial \boldsymbol{\omega}^{a,i} \partial \boldsymbol{\omega}^{b,n}} \tag{45}$$

5 Marginal Effects

Due to the complexity of the model's structure and the ability to place covariates in either the gating network, the expert regressions, or both, viewing the relationship between the covariates and the dependent variable through their marginal effects may provide a simplifying lens of the model's governing principles. Just as for logistic and multinomial regression, the marginal effects of an HME model have a closed form solution. Starting with equation (3) we replace the expert distributions P_t^m with the expected output for each of the m regressions and use the relationship in equation (6) to arrive at:

$$\mathbb{E}\left[y_t|\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{\theta}\right] = \sum_{m=1}^{M} \pi_{g_t^0 \longrightarrow m} \mathbb{E}\left[y_t|\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{\theta}, m\right] \tag{46}$$

In what follows, $\mathbb{E}\left[y_t\right]$ and $\mathbb{E}^m\left[y_t\right]$ will be used as shorthand for $\mathbb{E}\left[y_t|\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{\theta}\right]$ and $\mathbb{E}\left[y_t|\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{\theta}, m\right]$, respectively. The functional form of the marginal effect depends on where the variables appear in the model. Our existing notation labels the covariates in gating network as $\boldsymbol{Z}$ and the covariates in the expert regressions as $\boldsymbol{X}$. As seen later, the variables belonging to $\boldsymbol{Z}$ and $\boldsymbol{X}$ do not need to be mutually exclusive. There is also no requirement that they differ at all. In light of this, a few more notational definitions are needed to cover all the cases:

- $\boldsymbol{T} = \boldsymbol{Z} \cup \boldsymbol{X}$
- $\boldsymbol{V} = \boldsymbol{Z} \cap \boldsymbol{X}$
- $\boldsymbol{U}_Z = \boldsymbol{Z} \setminus \boldsymbol{X}$
- $\boldsymbol{U}_X = \boldsymbol{X} \setminus \boldsymbol{Z}$

The full list of variables considered in the model is labeled $\boldsymbol{T}$. Covariates that appear in both the gating network and the expert regressions are collected in $\boldsymbol{V}$. $\boldsymbol{U}_Z$ and $\boldsymbol{U}_X$ are used to label variables that appear only in the gating network or only in the expert regressions, respectively. With this notation, we can express the full marginal effects of the HME by where the explanatory variables appear in the model.

$$\frac{\partial \mathbb{E}\left[y_t\right]}{\partial \boldsymbol{T}} \equiv \boldsymbol{\Delta}_t = \sum_{m=1}^{M} \boldsymbol{\Delta}_t^m = \sum_{m=1}^{M} \left[\frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{U}_Z} \quad \frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{V}} \quad \frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{U}_X} \right] \tag{47}$$

with the functional form of the each covariate group in (47) defined as:

$$\frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{U}_Z} = \frac{\partial \pi_{g_t^0 \longrightarrow m}}{\partial \boldsymbol{U}_Z} \mathbb{E}^m\left[y_t\right] \tag{48}$$

$$\frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{U}_X} = \pi_{g_t^0 \longrightarrow m} \frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{U}_X} \tag{49}$$

$$\frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{V}} = \frac{\partial \pi_{g_t^0 \longrightarrow m}}{\partial \boldsymbol{V}} \mathbb{E}^m\left[y_t\right] + \pi_{g_t^0 \longrightarrow m} \frac{\partial \mathbb{E}^m\left[y_t\right]}{\partial \boldsymbol{V}} \tag{50}$$

Not matter how complex the model becomes, the researcher can always interpret the estimated HME through a single vector of marginal effects of $\boldsymbol{T}$. Of the four components in equations (48) - (50), three have already been established: $\mathbb{E}^m\left[y_t\right]$ is the output from local expert m, $\pi_{g_t^0 \longrightarrow m_t}$ is the prior weight for input t for local expert m, and $\frac{\partial \mathbb{E}^m[y_t]}{\partial \boldsymbol{X}}$ is the marginal effect of the local expert m with respect to covariates $\boldsymbol{X}$. What is left is the partial derivative of the gating network with respect to a variable in that network $\frac{\partial \pi_{g_t^0 \longrightarrow m}}{\partial \boldsymbol{z}_t}$. Starting with equation (5), we take the partial with respect to parameters in the gating matrix:

$$\boldsymbol{\delta}_t^m \equiv \frac{\partial \pi_{g_t^0 \longrightarrow m}}{\partial \boldsymbol{z}_t} = \frac{\partial g_t^{0,i} g_t^{i|0,j} \cdots g_t^{k|\cdots|j|i|0,m}}{\partial \boldsymbol{z}_t} \tag{51}$$

Applying the product rule yields:

$$\begin{aligned}\boldsymbol{\delta}_t^m &= \frac{\partial g_t^{0,i}}{\partial \boldsymbol{z}_t} g_t^{i|0,j} \cdots g_t^{k|\cdots|j|i|0,m} \\ &+ g_t^{0,i} \frac{\partial g_t^{i|0,j}}{\partial \boldsymbol{z}_t} \cdots g_t^{k|\cdots|j|i|0,m} \\ &+ \ldots \\ &+ g_t^{0,i} g_t^{i|0,j} \cdots \frac{\partial g_t^{k|\cdots|j|i|0,m}}{\partial \boldsymbol{z}_t}\end{aligned} \tag{52}$$

Since

$$\frac{\partial g_t^{a,i}}{\partial \boldsymbol{z}_t} = g_t^{a,i} \left(\boldsymbol{\omega}^{a,i} - \sum_j g_t^{a,j} \boldsymbol{\omega}^{a,j} \right)^\top = g_t^{a,i} \left(\boldsymbol{\omega}^{a,i} - \bar{\boldsymbol{\omega}}^a \right)^\top \tag{53}$$

we can substitute equation (53) into (52) to arrive at:

$$\begin{aligned}\boldsymbol{\delta}_t^m &= \pi_{g_t^0 \longrightarrow m} \left(\boldsymbol{\omega}^{0,i} + \boldsymbol{\omega}^{i|0,j} + \cdots + \boldsymbol{\omega}^{k|\cdots|j|i|0,m} - \left(\bar{\boldsymbol{\omega}}^0 + \bar{\boldsymbol{\omega}}^{i|0} + \cdots + \bar{\boldsymbol{\omega}}^{k|\cdots|j|i|0} \right) \right)^\top \\ &= \pi_{g_t^0 \longrightarrow m} (\boldsymbol{W}^m)^\top\end{aligned} \tag{54}$$

Looking closely at equation (54), the instantaneous rate of change of $\pi_{g_t^0 \longrightarrow m}$ to small deviations of $\boldsymbol{z}_t$ has an interesting representation. The row vector $(\boldsymbol{W}^m)^\top$ mean differences the parameter values of each edge in the path from the root node to expert m. This path is the *only* path from the root node to expert m. The sum of the mean parameter deviations are then appropriately weighted by the prior gate path $\pi_{g^0 \longrightarrow m}$.

5.1 Delta Method

Using the delta method, we can approximate standard errors for the marginal effects of the HME model. Starting with equation (47) from the previous section, we break down the gradient of the marginal effects with respect to the parameters by those in the gating network, $\boldsymbol{\Omega}$, and the parameters in the expert regressions, $\boldsymbol{\beta}$. These results are collected in Table 1.

Again, many of the expressions in Table 1 have already been defined in previous sections. The three expressions new to this section are $\frac{\partial^2 \mathbb{E}^m[y_t]}{\partial \boldsymbol{X} \partial \boldsymbol{\beta}^m}$, $\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^{a,i}}$, and $\frac{\partial \pi_{g_t^0 \longrightarrow m}}{\partial \boldsymbol{\omega}^{a,i}}$.

	$\underline{\boldsymbol{U}_Z}$	$\underline{\boldsymbol{V}}$	$\underline{\boldsymbol{U}_X}$
$\frac{\partial \boldsymbol{\Delta}_t^m}{\partial \boldsymbol{\omega}^a}$	$\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^a} \mathbb{E}^m \left[y_t \right]$	$\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^a} \mathbb{E}^m \left[y_t \right] + \frac{\partial \pi_{g_t^0 \longrightarrow m}}{\partial \boldsymbol{\omega}^a} \frac{\partial \mathbb{E}^m [y_t]}{\partial \boldsymbol{V}}$	$\mathbf{0}$
$\frac{\partial \boldsymbol{\Delta}_t^m}{\partial \boldsymbol{\beta}^m}$	$\mathbf{0}$	$\boldsymbol{\delta}_t^m \frac{\partial \mathbb{E}^m [y_t]}{\partial \boldsymbol{\beta}^m} + \pi_{g_t^0 \longrightarrow m} \frac{\partial^2 \mathbb{E}^m [y_t]}{\partial \boldsymbol{V} \partial \boldsymbol{\beta}^m}$	$\pi_{g_t^0 \longrightarrow m} \frac{\partial^2 \mathbb{E}^m [y_t]}{\partial \boldsymbol{U}_X \partial \boldsymbol{\beta}^m}$

Table 1: Delta Method Gradient Cases

For the standard OLS regressions that are considered in this paper, $\frac{\partial^2 \mathbb{E}^m [y_t]}{\partial \boldsymbol{X} \partial \boldsymbol{\beta}^m} = \mathbf{1}$. Conceptually, $\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^{a,i}}$ describes how the marginal effects of the gating network change in response to small changes in the parameters of $\boldsymbol{\Omega}$. The value of $\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^{a,i}}$ depends on what role $\boldsymbol{\omega}^{a,i}$ plays in navigating an input pattern from the root node to the expert m. In what follows, the indicator notation introduced in Section 4.2 will be used where $\mathbb{1}\{a,i,m\}$ is equal to one if $\boldsymbol{\omega}^{a,i}$ is an explicit gating vector for expert m and zero if it is an implicit gating vector. With this notation in mind, the partial derivative of the prior weight with respect to gate parameter vector $\boldsymbol{\omega}^{a,i}$ is:

$$\frac{\partial \pi_{g_t^0 \longrightarrow m}}{\partial \boldsymbol{\omega}^{a,i}} = \pi_{g_t^0 \longrightarrow f^m} \left(\mathbb{1}\{a,i,m\} - g^{a,i} \right) \boldsymbol{z}_t^\top \tag{55}$$

The partial derivative of the marginal effects of an HME with respect to a gate parameter vector is expressed as:

$$\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^{a,i}} = \pi_{g_t^0 \longrightarrow m} (\mathbb{1}\{a,i,m\} - g_t^{a,i}) + \pi_{g_t^0 \longrightarrow m} \left[(\mathbb{1}\{a,i,m\} - g_t^{a,i}) (\boldsymbol{W}^m)^\top - (\boldsymbol{G}^{a,i})^\top \right] \boldsymbol{z}_t^\top \tag{56}$$

where $\boldsymbol{W}^m$ was first seen in equation (54) and

$$\boldsymbol{G}^{a,i} = \left\{ g^{a,i} (1 - g^{a,i}) \boldsymbol{\omega}^{a,i} - \sum_{j \neq i} g^{a,i} g^{a,j} \boldsymbol{\omega}^{a,j} \right\} \tag{57}$$

Standard errors for the marginal effects for the HME models can then be constructed with the robust variance-covariance matrix from equation (21) and the collection of equations in this Section that fully defines $\frac{\partial \boldsymbol{\Delta}}{\partial \boldsymbol{\theta}}$.

$$Asy.Var \left[\hat{\boldsymbol{\Delta}} \right] = \sum_{n=1}^{M} \left(\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \boldsymbol{\Delta}_t}{\partial \boldsymbol{\theta}_n} \right) \boldsymbol{V}(\hat{\boldsymbol{\theta}}) \left(\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \boldsymbol{\Delta}_t}{\partial \boldsymbol{\theta}_n} \right)^\top \tag{58}$$

6 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 Anderson (1936) and popularized in the statistics literature by 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. The work below uses the ME and HME architectures to estimate a flower's sepal width using only its petal width as a predictor. The petal width will be used as the sole covariate in the local linear expert regressions ($\boldsymbol{X}$) as well as in the gating network ($\boldsymbol{Z}$).

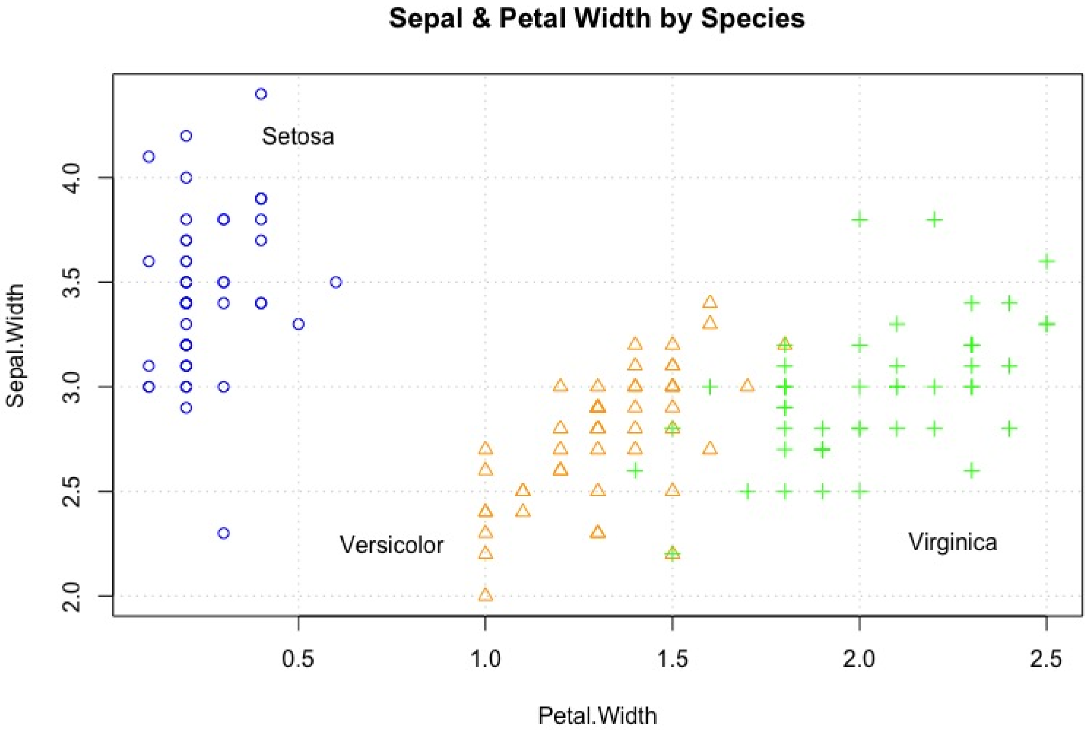


Figure 2: Three different iris species: Setosa (blue circles), Versicolor (orange triangles), Virginia (green crosses). Sepal width is on the vertical axis and petal width on the horizontal axis.

$$sepal.width_i = \beta_0 + \beta_1 * petal.width_i + \varepsilon_i \mid \omega_0 + \omega_1 * petal.width_i \qquad (59)$$

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 its species, resulting in a species-specific estimation of sepal width.

$$sepal.width_{is} = \beta_{0,s} + \beta_{1,s} * petal.width_{is} + \varepsilon_{is} \qquad (60)$$

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. When moving to the three expert model, there is now a choice on what kind of gating architecture to employ. We can go deep by adding a gating network with depth two (HME), or we can go wide by keeping the depth of the gating network at one (ME). Again, for comparative purposes, a benchmark OLS regression is estimated for each species separately. Results are collected in Table 2. Coefficients for local experts in the two expert ME regression match closely with the OLS benchmark. The strong separation between the Setosa and Versicolor/Virginica clusters makes it easy for the ME gating network to discriminate between the two using just the Petal Width dimension. This task becomes a little more complicated when considering all three species at the same time since there exists some overlap between the Versicolor and Virginica clusters. When comparing the coefficients of the local regressions (see Table 2), the HME architecture clearly outperforms the ME architecture. While the ME model does obtain a larger log-likelihood value than the OLS estimate, it fails to identify the three separate species that are known to exist. The HME model, on the other hand, naturally picks up on the three underlying clusters while also providing a superior likelihood value. This speaks to one of the major caveats of using this class of model. The likelihood value of an ME or HME can always been improved by adding more and more experts, but this improvement should not be confused with the model gaining a finer understanding of the underlying data generating process. It simply may start to over-fit the data at hand.

Table 2: Iris Dataset - OLS vs ME vs HME

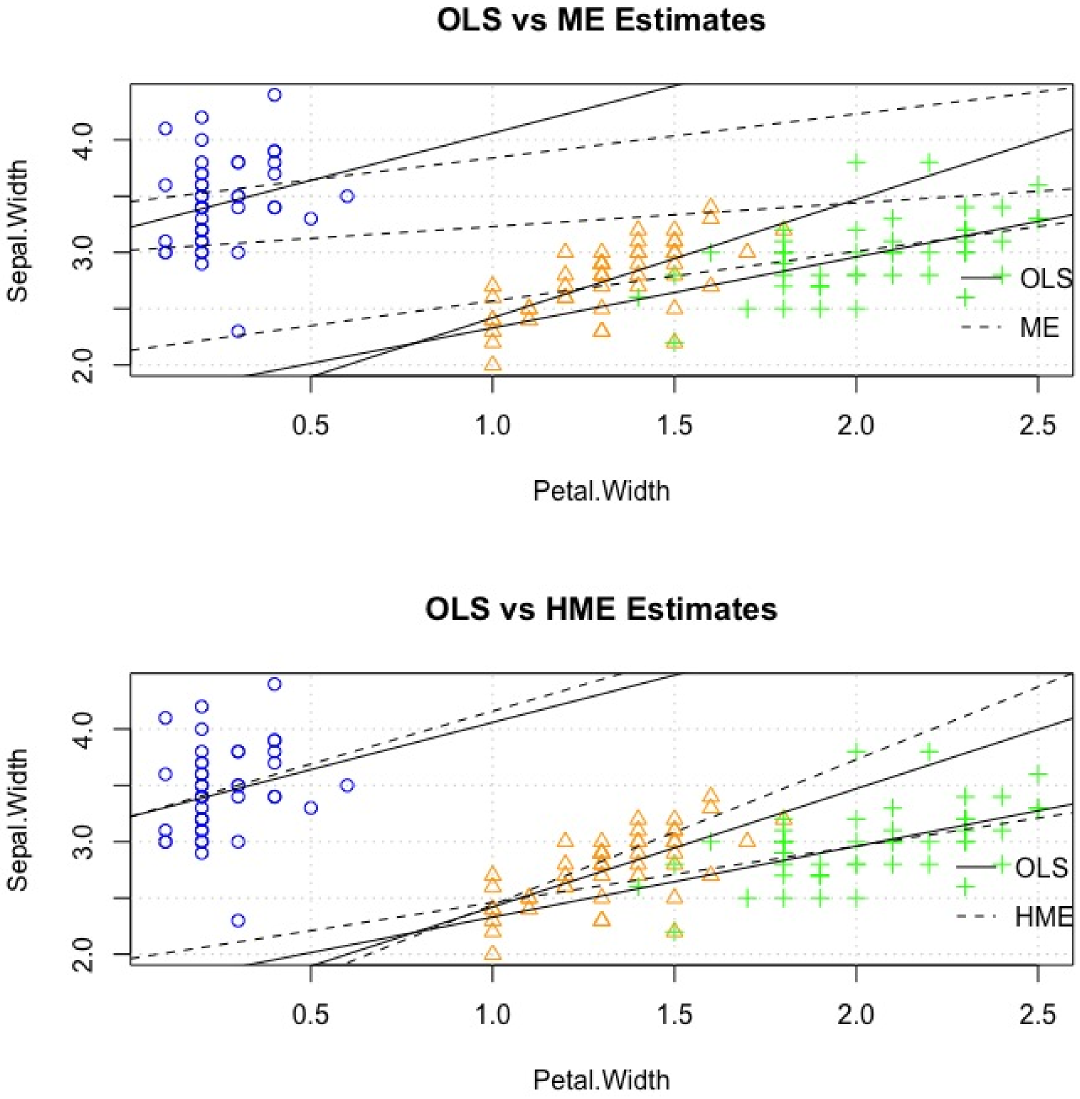
	2 Expert Mixture				3 Expert Mixture					
	OLS		ME		OLS		HME		ME	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Setosa										
Const.	3.22	0.11**	3.22	0.13**	3.22	0.11**	3.22	0.13**	3.45	0.13**
Petal.Width	0.84	0.42*	0.95	0.49**	0.84	0.41*	0.94	0.49	0.39	0.46
Virginica										
Const.	–	–	–	–	1.70	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 (60)

ME regressions are modeled using equation (59) and architecture $\boldsymbol{A}$ from Figure 1

HME regressions are modeled using equation (59) and architecture $\boldsymbol{C}$ from Figure 1


OLS vs ME Estimates

OLS vs HME Estimates

Figure 3: Comparison of the fitted experts between the ME and HME architectures applied to the Iris dataset. OLS regression estimates are drawn in solid lines. Although the HME and ME both achieve superior log-likelihood values compared to OLS, only the HME is able to identify the three iris species clusters.

7 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", this essay's version of it will be:

$$\log(wage) = \beta_0 + \beta_1 * \text{Age} + \beta_2 * \text{Age}^2 + \beta_3 * \text{YrsEdu} + \boldsymbol{\beta_4 X} + \varepsilon \tag{61}$$

with $\boldsymbol{X}$ containing a set of individual-specific variables as well as a set of occupation-specific attributes. The data will come from two sources. First, from the 2000 Census, a measure of the hourly (log) wage is devised. In addition to income, information on age, years of education (YrsEdu), job occupations codes, and a set of demographic identifiers indicating the race of the individuals are also obtained from the Census sample. For the occupational codes, the Standard Occupation Classification (SOC) codes from the Occupation Information Network (ONet) are used. Each occupation is associated with a set of knowledge and skill-based attributes describing which qualities are necessary to perform each job suitably. A federally sponsored source, ONet details, "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). The cross walk provided by Porter (2019) is used to link the occupational codes in the Census data to the SOC codes used by ONet. This mapping is not one-to-one. When more than one SOC code points to a single census code, the average of the SOC codes is taken. After a quick but careful scan of the job attributes available on ONet, the following four were selected to provide a small but diverse set of attributes that contrast well, with each attribute embodying a skill valued across industry, culture, and society: Social Perceptiveness [1], Data Analytics [2], Design [3], and Creative Thinking [4]. The footnotes provide a link to full classification hierarchy listed on the website.

For these selected attributes, ONet grades their relevance on a 100 point scale. Each attribute contains two scales, an "importance" (I) scale and a "level" (L) 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 into a single value, a Cobb-Douglass style average with a 2/3 weight for importance and a 1/3 weight for the level scale is used: $A = L^{\frac{1}{3}} I^{\frac{2}{3}}$. With

[1] https://www.onetonline.org/find/descriptor/result/2.B.1.a

[2] https://www.onetonline.org/find/descriptor/result/4.A.2.a.4

[3] https://www.onetonline.org/find/descriptor/result/2.C.3.c

[4] https://www.onetonline.org/find/descriptor/result/4.A.2.b.2

a unified attribute measure for every occupation in ONet's index, each attribute is mean centered and scaled to have unit variance across all ONet occupations.

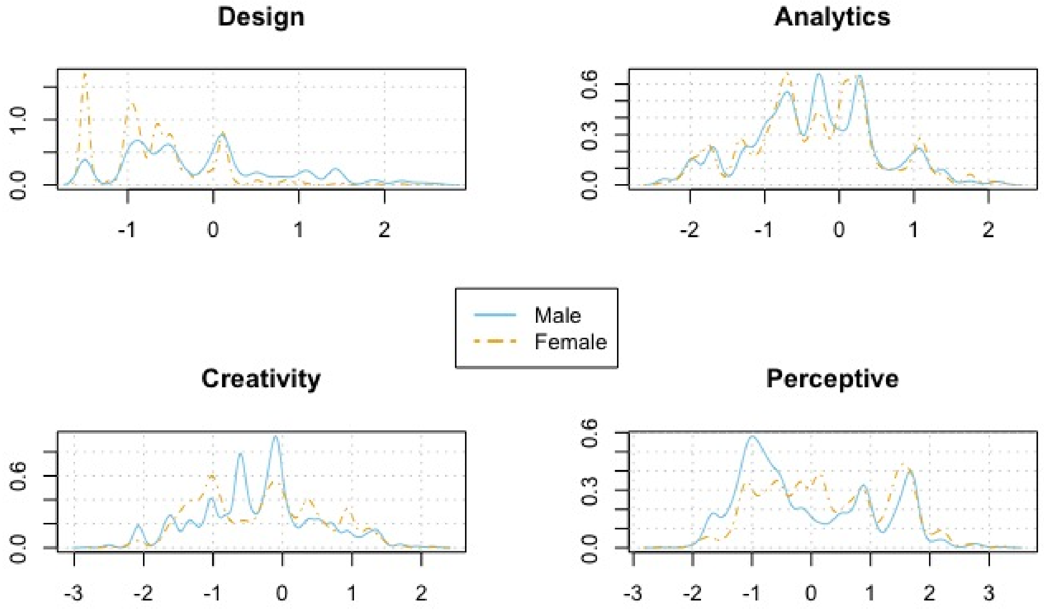


Figure 4: Density estimates of ONet job attributes for the Census sample broken down by sex. **Note:** The job attributes have been mean centered and scaled to have unit variance at the *occupational* level and not at the observation level with respect to sample.

The total number of individuals in the Census data numbers 105,796. After applying the crosswalk, only 75,957 cases remain with complete information across both datasets. Of those 75,957, roughly ten percent (7,315) are randomly held-out and used as a test set to gauge out-of-sample forecast performance across model specifications. This leaves 68,642 individuals left as a training set. A statistical summary of the covariates is provided in Table 3.

A natural question to consider as a researcher is where to put the variable(s) of interest while performing an HME estimation. Jiang and Tanner (2000) provide their proof of model consistency for HME of GLMs for the case where all covariates appear in the gating network as well as the experts. This will be referred to as the

Table 3: Summary Statistics

	25%	Mean	50%	75%
log Wage (hr)	2.22	2.61	2.59	2.96
Yrs Edu	12.00	13.78	14.00	16.00
Age	30.00	39.15	39.00	48.00
Age-16	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	-0.81	-0.23	-0.14	0.33
Design	-0.94	-0.36	-0.54	0.11
Analytic	-0.80	-0.24	-0.26	0.28
Perceptive	-0.82	0.16	0.13	1.08

N = 68,642

full specification:

$$log(wage) = Age + YrsEdu + Sex + Race + Occ \mid Age + YrsEdu + Sex + Race + Occ \tag{62}$$

The *full* specification will be compare to two others. First, a *mid* specification where the local experts contain age and years of education while removing demographic indicators:

$$log(wage) = Age + YrsEdu \mid Age + YrsEdu + Sex + Race + Occ \tag{63}$$

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

$$log(wage) = Age \mid Age + YrsEdu + Sex + Race + Occ \tag{64}$$

For comparative purposes, several different regressions across three different dimensions will be estimated: model architectures (ME vs HME), the number of experts, and the regression specification (equations (62) - (64)). Table 4 presents a view of these results across those dimensions. After looking at the results, two themes

Table 4: Comparing Complexity, Architecture, and Regression Specification

			Performance Metrics			
Specification	Architecture	Experts	Log-Lik	AIC	BIC	MSE
Full	ME	2	-0.541	1.082	1.088	0.182
	ME	3	-0.526	1.053	1.062	0.182
	ME	4	-0.537	1.078	1.091	0.181
	ME	5	-0.535	1.073	1.089	0.182
	HME	3	-0.525	1.052	1.061	0.182
	HME	4	-0.515	1.034	1.047	0.181
	HME	5	*-0.505*	*1.015*	*1.031*	*0.178*
Mid	ME	2	-0.560	1.120	1.123	0.185
	ME	3	-0.558	1.117	1.123	0.186
	ME	4	-0.581	1.163	1.171	0.192
	ME	5	-0.590	1.182	1.192	0.199
	HME	3	-0.541	1.083	1.088	0.184
	HME	4	-0.528	1.057	1.065	0.183
	HME	5	*-0.519*	*1.039*	*1.050*	*0.182*
Min	ME	2	-0.596	1.192	1.195	0.192
	ME	3	-0.587	1.176	1.181	0.192
	ME	4	-0.629	1.260	1.268	0.211
	ME	5	-0.564	1.131	1.140	0.189
	HME	3	-0.581	1.163	1.168	0.190
	HME	4	-0.546	1.094	1.101	0.182
	HME	5	*-0.524*	*1.049*	*1.059*	*0.182*

Note: Log-Likelihood, AIC, and BIC are divided by the sample size of 68,642. Italicized entries are the winning values within specification while underlined entries are the best values across all three specifications.

Note: The MSE is calculated from a hold-out test set with sample size of 7,315

Note: After looking at the results, two themes emerge. **One**, there is a clear advantage to using the HME structure if the aim is to maximize the likelihood value. The HME structure shows consistent improvement across specifications as the number of experts increases, while the ME struggles to match this consistency. **Two**, give the expert regressions as much information as possible. The Full specification clearly outperforms the Mid and Min specifications across the board.

emerge. First, there is a clear advantage to using the HME structure if the aim is to maximize the likelihood value. The HME structure shows consistent improvement across specifications as the number of experts increase, while the ME struggles to improve the likelihood value if there is only one gating split. This increase in efficiency is most likely due to the HME's more refined gating architecture, whose recursive partitioning is more effective at finding the next improvement in the parameter vector than the single multinomial split in the ME. As for the second theme, it is best to give the expert regressions as much information as possible. The Full specification clearly outperforms the Mid specification, which outperforms the Min specification. Referencing Table 4, if one holds the architecture and the number of experts constant, the performance metrics show clear improvement as the regression specification adds more explanatory variables.

Turning attention to the main variable of focus, Table 5 provides a comparison of the average marginal effect for *YrsEdu* across the same dimensions explored for the performance metrics. There is a noticeable change across model specifications. Compared to the OLS coefficient of 0.076, the Min specification, which includes *YrsEdu* only in the gating network, underestimates the returns to education. The Mid specification, which includes *Age* and *YrsEdu* in the expert regressions as well as the gating network, overestimates the returns to education in all the models except the HME with four and five experts. The Full specification, which has the entire suite of variables in both places, matches most closely to the OLS estimate across the estimated models.

For the Census sample, estimating up to five experts is pretty extreme. It is rather unlikely that there exists more than one distinct cluster, let alone two[5]. Because of this, a deeper analysis of the regression results are only explored for the three models that have the least complexity/experts. We first estimate equation (61) for a two expert model. At this specification, there is no distinction between the HME and ME. A three expert model is then estimated for these two respective architectures to assess if different conclusions to the estimated Mincer equations arise. Results for these regressions are collected in Tables 6, 8, and 10 and complimented by Tables 7, 9, and 11, which provide mean and median values for the subset of individuals in the census sample that are attributed to each expert based on the value of their posterior weights[6].

[5] Testing if a (H)ME model is even necessary would be a valuable addition to this paper.

[6] For example, observation i is assigned to expert j if the posterior vector's largest value is the j-th index: $\arg\max \boldsymbol{h}_i = h_{ij}$.

Table 5: Returns to Years of Education

		Avg. Marginal Effect		
Depth	Experts	Min	Mid	Full
ME	2	0.051	0.082	0.076
ME	3	0.051	0.081	0.074
ME	4	0.039	0.085	0.075
ME	5	0.063	0.095	0.076
HME	3	0.063	0.080	0.073
HME	4	0.063	0.078	0.073
HME	5	0.068	0.075	0.069

Note: OLS coef: 0.076

Note: There is a noticeable change across in the marginal return to an extra year of education. Compared to the OLS coefficient of 0.076, the Min specification, which includes *YrsEdu* only in the gating network, underestimates the returns to education. The Mid specification, which includes *Age* and *YrsEdu* in the expert regressions as well as the gating network, overestimates the returns to education in all the models except the HME with four and five experts. The Full specification, which has the entire suite of variables in both places, matches most closely to the OLS estimate across the estimated models.

Broadly speaking, all three models explored share the same macro view of the data. On the right side of Tables 6, 8, and 10 are a group of columns titled '(H)ME Marginal Effects'. Here the marginal effects of the model can be broken down and attributed to the gating network or the expert regressions. "Both", "Experts", and "Gates" refers to marginal effects referenced by equations (50), (49), and (48), respectively. The values are fairly consistent across variables and model architectures with the coefficients for *Age* and its square a modest exception, ranging from 0.028 (HME) to 0.042 (2-Expert ME) for *Age*. Notice also that the marginal effects from the expert regressions are the lion's share of total marginal effect, ranging from one to two orders of magnitude larger than marginal effects for the gating network. When looking at the occupational attributes there is similar agreement between the estimated models. The marginal effects for all three are in close proximity between the ME and HME models. Those individuals who specialize in performing analytics enjoy the greatest hourly rate (0.126 - 0.128). Design (0.074 to 0.081) and Perceptive (0.053 to 0.057) attributes get a smaller bump to the their hourly wage while Creative types (-0.044 to -0.043) clearly have alternative motivation than monetary gain.

When left to segment the data set on its own, the fitted HME models that are returned lead to some interesting conclusions. The first segmentation of the sample is seen by the two expert ME model that estimates two different wage equations. One for a majority of the population that tends to be older (median Age-16 = 25), whiter (78%), more educated (median YrsEduc = 14), and a second smaller population that is more diverse (70% white), significantly younger (median Age-16 = 7) and with less education on average (median YrsEduc = 12) (see Tables 6 and 7). The difference between the average age of the two populations is noticeable and may play a role behind the marginal effects for *Age* moving around as much as it does. Notice also that the members of the younger cohort hold lower-skilled jobs: the mean and median values for their occupational attributes are uniformly lower than their older and more educated counterparts. Finally, notice that the "penalty" for occupying a female or non-white body is less severe (and even turns positive for Indian and Asian) in the younger cohort.

Additional narratives present themselves as the segmentation continues and the number of experts expands. To reduce the chance of confusion the results from the deep three-expert HME model are used in what follows due to its superior likelihood value over the three-expert ME model (see Table 4). The main segmentation discovered by the two-expert ME model is carried over to the three expert HME model while a third latent sub-population emerges. The dominate cluster from the two-expert model is still quite large (78.3% of the posterior weight) compared to the younger cohort (13.3% of the posterior weight) and the new third cohort (8.4%). Three features distinguish this new population:

1. It skews slightly older than dominate cluster (27 vs 25 for median age-16)
2. It is the most educated of the three sub-populations with median years of education equal to 16 (compared to 14 for the dominate cluster and 12 for the younger group).
3. Members of this group are employed in positions where it is critically important to be aware of and understand others individual's behavior (Perceptive).

Just as with the two-expert ME model, the returns to education vary across these sub-groups. The young cohort, whose typical member has a high school diploma, has the lowest returns to education (0.034). The dominate cohort, whose median educational attainment is an Associate's degree, sees the highest returns to their years of schooling (0.082). There is a drop in returns (0.074) for the third and oldest cohort, even though the educational attainment for that group is the highest of the three groups with the median years of education equaling a Bachelor's degree. Taken together, the HME models suggests there is significant heterogeneity to returns in education over an individual's lifetime, across job types, and even by within similar cohorts. cohort.

8 Conclusion

In this article, a novel mixture model is explored that borrows equally from the economic and deep learning fields. A flexible (and optionally deep) gating network is used to learn the latent structure of a dataset and then apply local regressions to that latent structure. Robust standard errors and closed form expressions for marginal effects were developed and demonstrated on two different datasets.

Table 6: Regression Results: Two-Expert, Full Parameter Specification

	ME Regressions[1]				OLS[2]		ME Marginal Effects[3]				
	Coef.		Coef.		Coef.		Both		Experts		Gates
Intercept	1.231	*	1.494	*	1.241	*	1.225	*	1.260	*	-0.040
Age-16	0.032	*	0.068	*	0.035	*	0.042		0.038	*	0.004
Age-16^2	-0.000	*	-0.002	*	-0.001	*	-0.001		-0.001	*	-0.000
YrsEduc	0.082	*	0.036	*	0.076	*	0.076	*	0.075	*	0.000
Female	-0.244	*	-0.032	*	-0.215	*	-0.209	*	-0.207	*	-0.002
Af Amer	-0.076	*	-0.045	*	-0.076	*	-0.076	*	-0.071	*	-0.005
Indian	-0.081	*	1.390	*	-0.091	*	-0.085	+	-0.079	*	-0.005
Asian	-0.045	*	0.036	*	-0.032	*	-0.024		-0.028	*	0.003
Hisp	-0.121	*	-0.082	*	-0.106	*	-0.112	*	-0.112	*	-0.000
Creativity	-0.054	*	-0.008	*	-0.046	*	-0.044	*	-0.045	*	0.002
Design	0.080	*	0.078	*	0.082	*	0.081	*	0.080	*	0.001
Analytics	0.133	*	0.112	*	0.131	*	0.126	*	0.129	*	-0.003
Perceptive	0.063	*	-0.013	*	0.058	*	0.053	*	0.049	*	0.004
Log-Variance	-1.651	*	-2.682	*	–		–		–		–
Share[4]:	0.826		0.174		1.000		–		–		–

Signif. Codes: 0.01 '*', 0.05 '+', 0.1 '-'

Log-Likelihood: ME -0.541, OLS -0.558

[1] Fitted coefficients from the two-expert model with the full parameter specification from equation (62)

[2] Fitted coefficients from an OLS regression. These coefficient values can be compared to the HME coefficients to their left as well as to the marginal values to their right

[3] Marginal effects for the HME model. Standard errors are estimated by equation (58).

[4] The share is calculated by summing the posterior weights across observations for each expert.

Table 7: Sample Mean Comparison: Two-Expert ME

Share:[1]	(0.826)		(0.174)	
	Mean	Median	Mean	Median
log Wage (hr)	2.679	2.681	2.175	2.197
Age-16	25.814	25.000	6.965	7.000
Age-16^2	759.812	625.000	62.478	49.000
Female	0.408	0.000	0.386	0.000
Af Amer	0.084	0.000	0.101	0.000
Indian	0.009	0.000	0.018	0.000
White	0.778	1.000	0.698	1.000
Hispanic	0.037	0.000	0.028	0.000
Asian	0.091	0.000	0.155	0.000
YrsEduc	13.916	14.000	12.974	12.000
Creative	-0.191	-0.137	-0.464	-0.542
Design	-0.344	-0.535	-0.442	-0.635
Analytic	-0.196	-0.247	-0.499	-0.550
Perceptive	0.230	0.127	-0.233	-0.532
N	–	58,939	–	9,703

[1] The share is calculated by summing the posterior weights across observations for each expert.

Note: Mean and median values are applied to individuals in the census sample that are classified based on the value of their posterior weights. For example, observation i is assigned to expert j if the posterior vector's largest value is the j-th index: $\arg\max \boldsymbol{h}_i = h_{ij}$

Table 8: Regression Results: Wide Three-Expert, Full Parameter Specification

	ME Regressions[1]						OLS[2]		ME Marginal Effects[3]			
	Coef.		Coef.		Coef.		Coef.		Both	Experts		Gates
Intercept	1.379	*	1.574	*	0.562	*	1.241	*	1.367	1.340	*	0.032
Age-16	0.021	*	0.045	*	0.060	*	0.035	*	0.029	0.027	*	0.002
Age-16^2	-0.000	*	-0.001	*	-0.001	*	-0.001	*	-0.000	-0.000	*	0.000
YrsEduc	0.082	*	0.032	*	0.080	*	0.076	*	0.074	0.077	*	-0.002
Female	-0.251	*	-0.022	*	-0.149	*	-0.215	*	-0.206	-0.218	*	0.012
Af Amer	-0.084	*	-0.056	*	-0.054	-	-0.076	*	-0.076	-0.078	*	0.002
Indian	-0.105	*	-0.046	*	0.010		-0.091	*	-0.091	-0.090	*	-0.002
Asian	-0.030	*	0.057	*	-0.091	*	-0.032	*	-0.024	-0.025	*	0.001
Hisp	-0.136	*	-0.061	*	0.071	+	-0.106	*	-0.107	-0.111	*	0.004
Creativity	-0.038	*	-0.022	*	-0.177	*	-0.046	*	-0.044	-0.047	*	0.003
Design	0.080	*	0.080	*	-0.037	*	0.082	*	0.075	0.071	*	0.004
Analytics	0.123	*	0.110	*	0.196	*	0.131	*	0.128	0.128	*	0.000
Perceptive	0.060	*	-0.008	*	0.168	*	0.058	*	0.057	0.061	*	-0.004
Log-Variance	-1.893	*	-2.891	*	-0.627	*	–					
Share[4]:	0.809		0.111		0.080		1.000		–	–		–

Signif. Codes: 0.01 '*', 0.05 '+', 0.1 '-'

Log-Likelihood: ME -0.526, OLS -0.558

[1] Fitted coefficients from the three-expert model with the full parameter specification from equation (62)

[2] Fitted coefficients from an OLS regression. These coefficient values can be compared to the HME coefficients to their left as well as to the marginal values to their right

[3] Marginal effects for the HME model. Standard errors are estimated by equation (58).

[4] The share is calculated by summing the posterior weights across observations for each expert.

Table 9: Sample Mean Comparison: Wide Three-Expert HME

Share:[1]	(0.809)		(0.111)		(0.080)	
	Mean	Median	Mean	Median	Mean	Median
log Wage (hr)	2.664	2.667	2.106	2.096	2.549	2.221
Age-16	24.916	24.000	5.830	6.000	27.827	28.000
Age-16^2	722.355	576.000	41.789	36.000	904.103	784.000
Female	0.420	0.000	0.301	0.000	0.250	0.000
Af Amer	0.090	0.000	0.060	0.000	0.064	0.000
Indian	0.010	0.000	0.012	0.000	0.010	0.000
Hispanic	0.036	0.000	0.020	0.000	0.102	0.000
Asian	0.100	0.000	0.114	0.000	0.045	0.000
YrsEduc	13.802	14.000	13.101	12.000	15.837	16.000
Creative	-0.209	-0.141	-0.422	-0.456	-0.195	-0.282
Design	-0.344	-0.535	-0.387	-0.535	-0.765	-0.860
Analytic	-0.218	-0.264	-0.472	-0.412	-0.072	0.049
Perceptive	0.177	0.127	-0.122	-0.455	0.851	0.877
N	–	60,396	–	6,603	–	1,643

[1] The share is calculated by summing the posterior weights across observations for each expert.

Note: Mean and median values are applied to individuals in the census sample that are classified based on the value of their posterior weights. For example, observation i is assigned to expert j if the posterior vector's largest value is the j-th index: $\arg\max \boldsymbol{h}_i = h_{ij}$

Table 10: Regression Results: Deep Three-Expert, Full Parameter Specification

	HME Regressions[1]						OLS[2]		HME Marginal Effects[3]			
	Coef.		Coef.		Coef.		Coef.		Both	Experts		Gates
Intercept	1.404	*	1.559	*	0.898	*	1.241	*	1.393	1.382	*	0.011
Age-16	0.020	*	0.050	*	0.044	*	0.035	*	0.028	0.026	*	0.003
Age-16^2	-0.000	*	-0.001	*	-0.001	*	-0.001	*	-0.000	-0.000	*	0.000
YrsEduc	0.082	*	0.034	*	0.074	*	0.076	*	0.073	0.075	*	-0.001
Female	-0.257	*	-0.034	*	-0.131	*	-0.215	*	-0.209	-0.217	*	0.008
Af Amer	-0.086	*	-0.048	*	-0.041		-0.076	*	-0.076	-0.077	*	0.001
Indian	-0.113	*	-0.057	*	0.043		-0.091	*	-0.100	-0.093	*	-0.007
Asian	-0.033	*	0.058	*	-0.062	+	-0.032	*	-0.025	-0.023	*	-0.001
Hisp	-0.143	*	-0.066	*	0.077	*	-0.106	*	-0.111	-0.114	*	0.003
Creativity	-0.042	*	-0.021	*	-0.136	*	-0.046	*	-0.043	-0.047	*	0.004
Design	0.080	*	0.068	*	-0.048	*	0.082	*	0.074	0.068	*	0.006
Analytics	0.124	*	0.112	*	0.183	*	0.131	*	0.128	0.127	*	0.000
Perceptive	0.063	*	-0.003		0.135	*	0.058	*	0.056	0.061	*	-0.004
Log-Variance	-1.895	*	-2.791	*	-0.622	*	–					
Share[4]:	0.783		0.133		0.084		1.000		–	–		–

Signif. Codes: 0.01 '*', 0.05 '+', 0.1 '-'

Log-Likelihood: HME -0.525, OLS -0.558

[1] Fitted coefficients from the three-expert model with the full parameter specification from equation (62)

[2] Fitted coefficients from an OLS regression. These coefficient values can be compared to the HME coefficients to their left as well as to the marginal values to their right

[3] Marginal effects for the HME model. Standard errors are estimated by equation (58).

[4] The share is calculated by summing the posterior weights across observations for each expert.

Table 11: Sample Mean Comparison: Deep Three-Expert HME

Share:[1]	(0.783)		(0.133)		(0.084)	
	Mean	Median	Mean	Median	Mean	Median
log Wage (hr)	2.683	2.676	2.137	2.140	2.432	2.075
Age-16	25.523	25.000	6.494	7.000	26.913	27.000
Age-16^2	746.250	625.000	50.858	49.000	873.924	729.000
Female	0.414	0.000	0.358	0.000	0.313	0.000
Af Amer	0.088	0.000	0.073	0.000	0.075	0.000
Indian	0.010	0.000	0.016	0.000	0.018	0.000
White	0.770	1.000	0.753	1.000	0.749	1.000
Hispanic	0.036	0.000	0.027	0.000	0.101	0.000
Asian	0.096	0.000	0.131	0.000	0.057	0.000
YrsEduc	13.846	14.000	13.077	12.000	15.378	16.000
Creative	-0.198	-0.137	-0.444	-0.508	-0.201	-0.282
Design	-0.330	-0.530	-0.477	-0.635	-0.757	-0.859
Analytic	-0.206	-0.253	-0.471	-0.412	-0.161	-0.007
Perceptive	0.185	0.127	-0.082	-0.308	0.756	0.877
N	–	58,429	–	8,674	–	1,539

[1] The share is calculated by summing the posterior weights across observations for each expert.

Note: Mean and median values are applied to individuals in the census sample that are classified based on the value of their posterior weights. For example, observation i is assigned to expert j if the posterior vector's largest value is the j-th index: $\arg\max \boldsymbol{h}_i = h_{ij}$

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.

Brieman, L. et al. (1984). *Classification and Regression Trees.* Wadswoth and Brooks/Cole.

Carvalho, Alexandre and Georgios Skoulakis (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.

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, Robert 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.

Jordan, Michael I. and Robert A. Jacobs (1992). “Hierarchies of adaptive experts”. In: *Advances in Neural Information Processing Systems 4*. Ed. by J. E. Moody, S. J. Hanson, and R. P. Lippmann. Morgan-Kaufmann, pp. 985–992. URL: http://papers.nips.cc/paper/514-hierarchies-of-adaptive-experts.pdf.

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

Porter, Sarah (2019). *Census to ONet Mapping*. URL: http://econterms.net/pbmeyer/research/occs/wiki/index.php?title=Crosswalk_by_Sarah_Porter_to_map_1980_codes_forward_in_SAS (visited on 01/28/2019).

Terasvirta, Timo (1994). “Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models”. In: *Journal of the American Statistical Association* 89.425, pp. 208–218. ISSN: 01621459. URL: http://www.jstor.org/stable/2291217.

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*.

Waterhouse, Steve R., David MacKay, and Anthony J. Robinson (1995). “Bayesian Methods for Mixtures of Experts”. In: *NIPS*.

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.