Econometric Applications of Hierarchical Mixture of Experts

Lucas C. Dowiak

August 16, 2020

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 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*. The authors employ this mixture of experts (ME) framework to model vowel discrimination in a speech recognization context. Shortly after, M. I. Jordan and Robert A. Jacobs 1992 generalize a single-layer multinomial gating network to allow for more than one layer while M. Jordan and R. Jacobs 1993 demonstrate an Expectation-Maximization approach to estimation strategy more suitable to 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. 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 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, we provide a vary simple demonstration of the HME in action and then move on to a more economically relevant example of applying the HME model to a Mincer wage equation in section 7. Section 8 offers a modest conclusion.

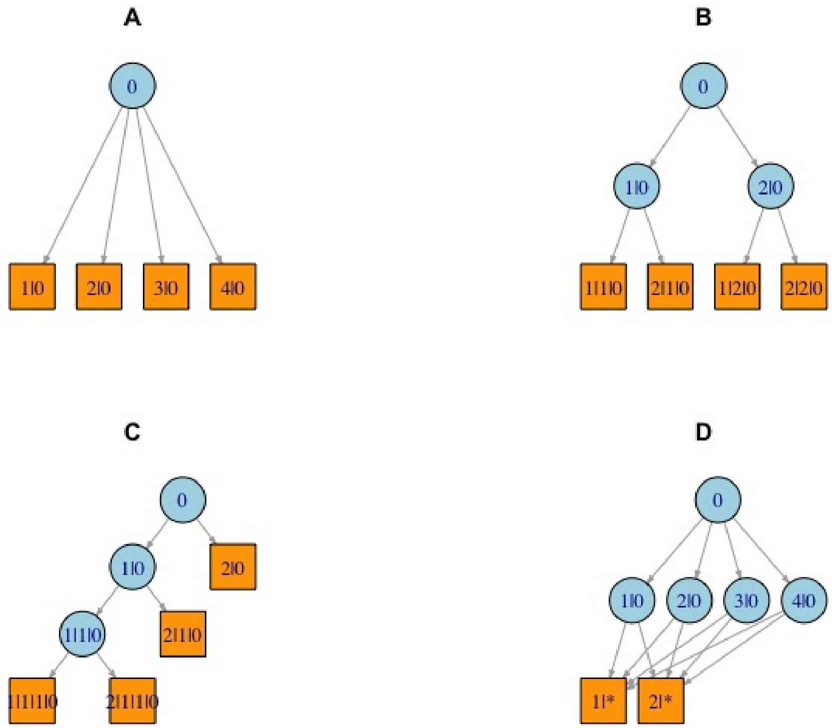

Figure 1: Networks **A** - **D** depict various network architectures that are discussed in this article. For all four networks, gating nodes are represented as blue circles and experts as orange rectangles. Network **A** illustrates the original Mixture of Experts (ME) architexture with a single multinomial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

MoE and HME frameworks have been utilized for both time series and cross-sectional analysis. Within the cross-sectional approach literature, S. Waterhouse and A. 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 consider S. Waterhouse and A. Robinson 1995 and alter their growing algorithm with a mind to scaling the model to handle thousands of experts. M. Jordan and Xu 1995 provide an extended discussion on the convergence of the model used by M. Jordan and R. Jacobs 1993. The authors also suggest algorithmic improvements to help with estimation. Continuing the theoretical discussing, Jiang and M. A. Tanner 1999 cover 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.

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. Huerta, Jiang, and M. A. 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 M. Tanner 2003 lay out the necessary regularity conditions to perform hypthesis tests on stationary 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 take a similar approach to Carvalho and M. Tanner 2003 but apply their analysis to 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.

In this article, estimation and inference is from a maximum likelihood perspective and will remain the primary focues. Estimation of MoE and HME models from a Bayesian has received considerable amount of attention as well.

S. R. Waterhouse, MacKay, and A. J. Robinson 1995 provided an initial approach to estimating a MoE by combining gaussian priors on the gating and expert parameters with gamma hyperparameter priors in an approximating ensemble to the true joint density of the model. Optimizition of the parameter vector for the approximating density occurs a block of parameters at a time. Ueda and Ghahramani 2002 improve on S. R. Waterhouse, MacKay, and A. J. Robinson 1995 by optizimg for the appropriate number of experts in in addtion to model parameters. 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 complete 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.

1.2 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.

Carvalho and Skoulakis 2005

2 Model

We start by presenting the HME as a standard mixture model. For a given input and output pair $(\boldsymbol{X_t}, Y_t)$, each expert provides a probablistic model relating input $\boldsymbol{X_t}$ 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

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

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

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

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

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

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

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

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

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:

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

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:

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

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:

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

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

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

And taking its log yields the log likelihood

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

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 M. 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 resarcher 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

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

leading to the complete-data log-likelihood

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

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 forthcoming articles.

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:

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

Here $\pi_{h^0 \longleftrightarrow h^{k,\ldots|j|i|0}}$ is analogous 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' theorem.

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

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, M. Jordan and R. 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 optmized on their own. For the experts we have:

$$
\underset{\boldsymbol{\beta^m}}{\arg\max} \sum_t \pi_{h_t^0 \longleftrightarrow P_t^m} \log P^m(Y_t|X_t;\boldsymbol{\beta}^m) \tag{17}
$$

And for the gating node:

$$
\underset{\boldsymbol{\omega^a}}{\arg\max} \sum_t \pi_{h_t^0 \longleftrightarrow h_t^a} \log g(Z_t, \boldsymbol{\omega}^a) \tag{18}
$$

4 Inference

When considering inference, it's 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 $Y_i = X_i\boldsymbol{\beta}$ may mask a critical change in relationship depending on membership to some unknown sub-group j of the data $Y_{ij} = X_{ij}\boldsymbol{\beta}_j$. A wide variety of time series, especially those with longer histories, experience changes in behaviour over time. They can be subjected to sharp one-off changes in value or more gradual changes of behavior 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 Z to feed through the gating network. This limited workload may come at a cost, though. By allowing the gating network to find it's own mixture allocations, the odds of arriving at a mispecified model becomes a concern. To guard against this, we use a sandwich estimator for the variance-covariance matrix:

$$\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{H}^{-1}(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{H}^{-1}(\boldsymbol{\theta}) \tag{19}$$

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

$$\boldsymbol{G}(\boldsymbol{\theta}) = \sum_t \boldsymbol{S}_t(\boldsymbol{\theta})\boldsymbol{S}_t(\boldsymbol{\theta})^\top \tag{20}$$

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

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{1}{T}\sum_t \mathbf{H}_t(\boldsymbol{\theta}) \tag{21}$$

We discuss the score vectors and the hessians in more detail in the following two sections.

4.1 The Score

For the score vector, we concatenate the scores of each gating node and those of each local expert regressions.

$$\boldsymbol{S}_t(\boldsymbol{\theta}) = [\boldsymbol{S}_t(\boldsymbol{\Omega})\ \boldsymbol{S}_t(\boldsymbol{\beta})] \tag{22}$$

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

$$\boldsymbol{S}_t(\boldsymbol{\Omega}) = [\boldsymbol{S}_t(\boldsymbol{\omega}^0)\ \boldsymbol{S}_t(\boldsymbol{\omega}^{1|0})\ \boldsymbol{S}_t(\boldsymbol{\omega}^{2|0})\ \ldots] \tag{23}$$

$$\boldsymbol{S}_t(\boldsymbol{\omega}^a) = [\boldsymbol{S}_t(\boldsymbol{\omega}^{a,1})\ \ldots\ \boldsymbol{S}_t(\boldsymbol{\omega}^{a,J-1})] \tag{24}$$

For a generic gating node a we can define the individual score for sample t as:

$$\boldsymbol{S}_t(\boldsymbol{\omega}^{a,i}) = \frac{\partial Q_t}{\partial \boldsymbol{\omega}^{a,i}} = \pi_{h_t^0 \longleftrightarrow h_t^a}(1 - g_t^{a,i})\boldsymbol{Z}_t \tag{25}$$

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 M. A. Tanner 2003 is a notable exception). When all experts share the same functional form, it's standard to accept the restriction that no experts in the HME model produce the same parameter vector $\boldsymbol{\beta}^j \neq \boldsymbol{\beta}^k$. Such an HME is defined by Jiang and M. A. Tanner 2000 as being *irreducible*. The irreducibility of an HME plays a critical role in guaranteeing the convergence of the model.

In this article, each HME discussed will employ set of experts running a standard regression model with gaussian errors. To aide with model optimization, the specification of the parameter vector for each regression $\boldsymbol{\beta}^m = [\beta_0^m\ \ldots\ \beta_k^m\ \phi^m]$ takes on a unique form where we model the log variance explicitly $\phi = \log \sigma^2$.

$$P^m(Y_t|X_t;\, \boldsymbol{\beta}^m, \phi^m) = (2\pi \exp(\phi^m))^{-\frac{1}{2}} \exp\left(-\frac{(Y_t - \boldsymbol{X}_t\boldsymbol{\beta}^m)^2}{2\exp(\phi^m)}\right) \tag{26}$$

In this case the score vector for any particular regression expert is:

$$\boldsymbol{S}_t(\boldsymbol{\beta^m}) = \Big(\frac{\partial Q_t}{\partial \boldsymbol{\beta}^m}\ ,\ \frac{\partial Q_t}{\partial \phi^m}\Big)^\top \tag{27}$$

$$\boldsymbol{S}_t(\boldsymbol{\beta}) = [\boldsymbol{S}_t(\boldsymbol{\beta}^1)\ \ldots\ \boldsymbol{S}_t(\boldsymbol{\beta}^M)] \tag{28}$$

with:

$$\frac{\partial Q_t}{\partial \boldsymbol{\beta}^m} = \pi_{h_t^0 \longleftrightarrow f_t^m}\frac{(Y_t - \boldsymbol{X}_t\boldsymbol{\beta}^m)}{\exp(\phi^m)}\boldsymbol{X}_t \tag{29}$$

and

$$\frac{\partial Q_t}{\partial \phi^m} = \frac{\pi_{h_t^0 \longleftrightarrow f_t^m}}{2}\Big(\frac{(Y_t - \boldsymbol{X}_t\boldsymbol{\beta}^m)^2}{\exp(\phi^m)} - 1\Big) \tag{30}$$

4.2 The Hessian

The hessian is equally straight-forward. Starting with equation (25), the hessian for each gating node is:

$$\mathbf{H}_t(\boldsymbol{\omega}^a) \equiv \frac{\partial^2 Q}{\partial \boldsymbol{\omega}^{a,i} \partial \boldsymbol{\omega}^{a,j}} = \pi_{h_t^0 \longleftrightarrow h_t^a} \boldsymbol{\Gamma}_t^a \otimes \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \tag{31}$$

where $\otimes$ is the kronecker product and:

$$\boldsymbol{\Gamma}_t^a = \begin{bmatrix} -g_t^{a,1}(1-g_t^{a,1}) & g_t^{a,1} g_t^{a,2} & \dots & g_t^{a,1} g_t^{a,J-1} \\ g_t^{a,1} g_t^{a,2} & -g_t^{a,2}(1-g_t^{a,2}) & \dots & g_t^{a,2} g_t^{a,J-1} \\ \vdots & \vdots & \ddots & \vdots \\ g_t^{a,1} g_t^{a,J-1} & g_t^{a,2} g_t^{a,J-1} & \dots & -g_t^{a,J-1}(1-g_t^{a,J-1}) \end{bmatrix} \tag{32}$$

For each expert regression:

$$\mathbf{H}_t(\boldsymbol{\beta}^m) = \frac{\pi_{h_t^0 \longleftrightarrow f_t^m}}{\exp(\phi^m)} \begin{bmatrix} \boldsymbol{X}_t \boldsymbol{X}_t & \boldsymbol{X}_t \epsilon_t^m \\ \boldsymbol{X}_t \epsilon_t^m & \frac{1}{2}(\epsilon_t^m)^2 \end{bmatrix} \tag{33}$$

where we have set $\epsilon_t^m = Y_t - \boldsymbol{X}_t \boldsymbol{\beta}^m$ to ease the notational burden. Staying consistent with the score vector, we sum the hessian matrices across observations:

$$\boldsymbol{H}(\boldsymbol{\omega}^a) = \sum_t^T \mathbf{H}_t(\boldsymbol{\omega}^a) \tag{34}$$

$$\boldsymbol{H}(\boldsymbol{\beta}^m) = \sum_t^T \mathbf{H}_t(\boldsymbol{\beta}^m) \tag{35}$$

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 regression's functional form f_t^m and use the relationship in equation (6) to arrive at:

$$f_t = f(X_t, Z_t; \boldsymbol{\beta}, \boldsymbol{\Omega}) = \sum_{m=1}^M \pi_{g_t^0 \longleftrightarrow f_t^m} f^m(X_t; \boldsymbol{\beta}^m) \tag{36}$$

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:

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

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 collect in $\boldsymbol{W}$. $\boldsymbol{U}_Z$ 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.

$$\frac{\partial f_t}{\partial \boldsymbol{T}} = \sum_{m=1}^{M} \boldsymbol{\Delta}^m = \sum_{m=1}^{M} \left[\frac{\partial f_t^m}{\partial \boldsymbol{U}_Z} \quad \frac{\partial f_t^m}{\partial \boldsymbol{W}} \quad \frac{\partial f_t^m}{\partial \boldsymbol{U}_X} \right] \tag{37}$$

with the functional form of the each covariate group in (37) defined as:

$$\frac{\partial f_t^m}{\partial \boldsymbol{U}_Z} = \frac{\partial \pi_{g_t^0 \longleftrightarrow f_t^m}}{\partial \boldsymbol{U}_Z} f_t^m \tag{38}$$

$$\frac{\partial f_t^m}{\partial \boldsymbol{U}_X} = \pi_{g_t^0 \longleftrightarrow f_t^m} \frac{\partial f_t^m}{\partial \boldsymbol{U}_X} \tag{39}$$

$$\frac{\partial f_t^m}{\partial \boldsymbol{W}} = \frac{\partial \pi_{g_t^0 \longleftrightarrow f_t^m}}{\partial \boldsymbol{W}} f_t^m + \pi_{g_t^0 \longleftrightarrow f_t^m} \frac{\partial f_t^m}{\partial \boldsymbol{W}} \tag{40}$$

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 (38) - (40), three have already been established: f_t^m is the output from local expert m, $\pi_{g_t^0 \longleftrightarrow f_t^m}$ is the prior weight for input t for local expert m, and $\frac{\partial f_t^m}{\partial \boldsymbol{X}}$ is the marginal effect of the local expert m. What is left is the partial derivative of the gating network with respect to a variables in that network $\frac{\partial \pi_{g^0 \longleftrightarrow f^m}}{\partial Z}$. Starting with equation (5), we take the partial with respect to gating matrix $\boldsymbol{Z}$:

$$\boldsymbol{\delta}^m \equiv \frac{\partial \pi_{g^0 \longleftrightarrow f^m}}{\partial \boldsymbol{Z}} = \frac{\partial g^{0,i} g^{i|0,j} \dots g^{k|\cdots|j|i|0,m}}{\partial \boldsymbol{Z}} \tag{41}$$

and applying the product rule gives us:

$$\begin{aligned}\delta^m = & \frac{\partial g^{0,i}}{\partial \boldsymbol{Z}} g^{i|0,j} \cdots g^{k|\cdots|j|i|0,m} \\ & + g^{0,i} \frac{\partial g^{i|0,j}}{\partial \boldsymbol{Z}} \cdots g^{k|\cdots|j|i|0,m} \\ & + \dots \\ & + g^{0,i} g^{i|0,j} \dots \frac{\partial g^{k|\cdots|j|i|0,m}}{\partial \boldsymbol{Z}}\end{aligned} \tag{42}$$

and since:

$$\frac{\partial g^{a,i}}{\partial \boldsymbol{Z}} = g^{a,i} \left(\boldsymbol{\omega}^{a,i} - \sum_j g^{a,j} \boldsymbol{\omega}^{a,j} \right) = g^{a,i} \left(\boldsymbol{\omega}^{a,i} - \bar{\boldsymbol{\omega}}^a \right) \tag{43}$$

we can substitute equation (43) into (42) to arrive at:

$$\boldsymbol{\delta^m} = \pi_{g^0 \longleftrightarrow P^m} \left(\boldsymbol{\omega}^{0,i} + \boldsymbol{\omega}^{i|0,j} + \cdots + \boldsymbol{\omega}^{k|\cdots|j|i|0,m} - \left(\bar{\boldsymbol{\omega}}^0 + \bar{\boldsymbol{\omega}}^{i|0} + \cdots + \bar{\boldsymbol{\omega}}^{k|\cdots|j|i|0} \right) \right) \tag{44}$$

5.1 Delta method

Using the delta method, we can approximate standard errors for the marginal effects of the HME model. Starting with equation (37) 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 regression, $\boldsymbol{\beta}$. These results are collected in table 1.

Again, many of the expressions in table 1 have already been defined in previous sections. The two experssions new to this section are $\frac{\partial^2 f_t^m}{\partial \boldsymbol{X} \partial \boldsymbol{\beta}^m}$ and $\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^{a,i}}$. For the standard OLS regressions that are considered in this paper, $\frac{\partial^2 f_t^m}{\partial \boldsymbol{X} \partial \boldsymbol{\beta}^m} = \mathbf{1}$. Conceptually, $\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^{a,i}}$ describes how the marginal effects of the gating network change in response

	$\underline{\boldsymbol{U}_Z}$	$\underline{\boldsymbol{W}}$	$\underline{\boldsymbol{U}_X}$
$\frac{\partial \boldsymbol{\Delta}_t^m}{\partial \boldsymbol{\omega}^a}$	$\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^a} f_t^m$	$\frac{\partial \boldsymbol{\delta}_t^m}{\partial \boldsymbol{\omega}^a} f_t^m + \frac{\partial \pi_{g_t^0 \longleftrightarrow f_t^m}}{\partial \boldsymbol{\omega}^a} \frac{\partial f_t^m}{\partial \boldsymbol{W}}$	$\mathbf{0}$
$\frac{\partial \boldsymbol{\Delta}_t^m}{\partial \boldsymbol{\beta}^m}$	$\mathbf{0}$	$\boldsymbol{\delta}_t^m \frac{\partial f_t^m}{\partial \boldsymbol{\beta}^m} + \pi_{g_t^0 \longleftrightarrow f_t^m} \frac{\partial^2 f_t^m}{\partial \boldsymbol{W} \partial \boldsymbol{\beta}^m}$	$\pi_{g_t^0 \longleftrightarrow f_t^m} \frac{\partial^2 f_t^m}{\partial \boldsymbol{U}_X \partial \boldsymbol{\beta}^m}$

Table 1: Delta Method Gradient Cases

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. For instance, say that we're at the root node, and it's our mission is to traverse the gating network down to expert m. When we arrive at node a, if the direction we need to take to reach expert m is along path i, then we'll call $\boldsymbol{\omega}^{a,i}$ an *explicit* parameter set with respect to expert m. If taking path i leads to a different expert, then $\boldsymbol{\omega}^{a,i}$ will be refered to as an *implicit* parameter set.

For an explict path

$$\frac{\partial \delta_t^m}{\partial \omega_p^{a,i}} = \pi_{g^0 \longleftrightarrow f^m} \left[(1 - g^{a,i}) + \left[W_p^m (1 - g^{a,i}) - G_p^{a,i} \right] Z_p \right] \tag{45a}$$

$$\frac{\partial \pi_{g^0 \longleftrightarrow f^m}}{\partial \omega_p^{a,i}} = \pi_{g^0 \longleftrightarrow f^m} (1 - g^{a,i}) Z_p \tag{45b}$$

and for an implicit path

$$\frac{\partial \delta_t^m}{\partial \omega_p^{a,j}} = \pi_{g^0 \longleftrightarrow f^m} \left[-g^{a,j} + \left[-W_p^m (1 - g^{a,j}) - G_p^{a,j} \right] Z_p \right] \tag{46a}$$

$$\frac{\partial \pi_{g^0 \longleftrightarrow f^m}}{\partial \omega_p^{a,j}} = -\pi_{g^0 \longleftrightarrow f^m} g^{a,j} Z_p \tag{46b}$$

where

$$W_p^m = \left[\omega_p^{0,i} + \cdots + \omega_p^{k|\cdots|j|i|0,m} - \left(\bar{\omega}_p^0 + \cdots + \bar{\omega}_p^{k|\cdots|j|i|0} \right) \right] \tag{47}$$

$$G_p^{a,i} = \left\{ g^{a,i} (1 - g^{a,i}) \omega_p^{a,i} - \sum_{j \neq i} g^{a,i} g^{a,j} \omega_p^{a,j} \right\} \tag{48}$$

The intermediate step for equation 45a

$$\frac{\partial \delta_t^m}{\partial \omega_p^{a,i}} = (1-g^{a,i})\pi_{g^0 \longleftrightarrow f^m}\left[\omega^{0,i}+\cdots+\omega^{k|\cdots|j|i|0,m}-\left(\bar{\omega}^0+\cdots+\bar{\omega}^{k|\cdots|j|i|0}\right)\right]Z_p \quad + \\ \pi_{g^0 \longleftrightarrow f^m}\left[(1-g^{a,i})-\left\{g^{a,i}(1-g^{a,i})\omega_p^{a,i}-\sum_{j\neq i} g^{a,i}g^{a,j}\omega_p^{a,j}\right\}Z_p\right]$$

Standard errors for the marginal effects for the HME models can then be constructed with the robust variance-covariance matrix from equation (19) and the collection of equations from (37) to (46).

$$Asy.Var\left[\hat{\boldsymbol{\Delta}}\right] = \sum_{n=1}^{M}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \boldsymbol{\Delta}_t}{\partial \boldsymbol{\theta}_n}\right)\boldsymbol{V}(\hat{\boldsymbol{\theta}})\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \boldsymbol{\Delta}_t}{\partial \boldsymbol{\theta}_n}\right)^{\top} \tag{49}$$

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 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.

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).

$$sepal.width_i = \beta_0 + \beta_1 * petal.width_i + \varepsilon_i \;\mid\; \omega_0 + \omega_1 * petal.width_i \tag{50}$$

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.

$$sepal.width_{is} = \beta_{0,s} + \beta_{1,s} * petal.width_{is} + \varepsilon_{is} \tag{51}$$

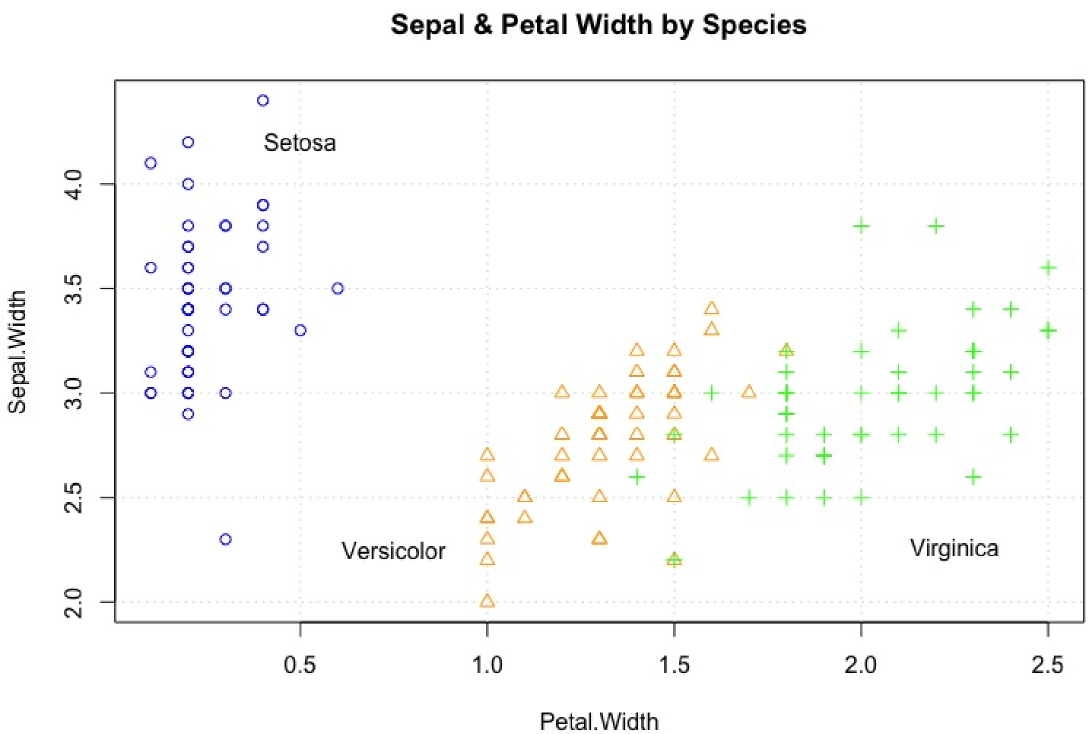

Figure 2: Three different Iris species are presented: Setosa (blue circles), Versicolor (orange triangles), Virginia (green crosses). Sepal width is on the vertical axis and petal width on the horizontal axis.

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 2. 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 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 (51)

ME regressions are modeled using equation (50) and architecture $\boldsymbol{A}$ from figure 1

HME regressions are modeled using equation (50) and architecture $\boldsymbol{C}$ from figure 1

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", our version of it will be:

$$\log(wage) = \beta_0 + \beta_1 * \text{Age} + \beta_2 * \text{Age}^2 + \beta_3 * \text{YrsEdu} + \boldsymbol{\beta_4 X} + \varepsilon \tag{52}$$

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 the hourly (log) wage. In addition to income, we also collect information on age, years of education (YrsEdu), job occupations codes, and a set of demographic identifiers indicating the race of the indivduals contained in census sample. 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 link the occupational codes in the census data to the soc codes used by ONet, we use the cross walk provided by Sarah Porter "citeCrosswalk". This mapping 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. After a quick but careful scan of the job attributes available on ONet, the following four were selected. The footnootes provide the full classification hierarchy listed on the website.

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

The guiding principle for attribute selection was to choose a small but diverse set of attributes that contrast well, with each attribute embodying a human skill valued

[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

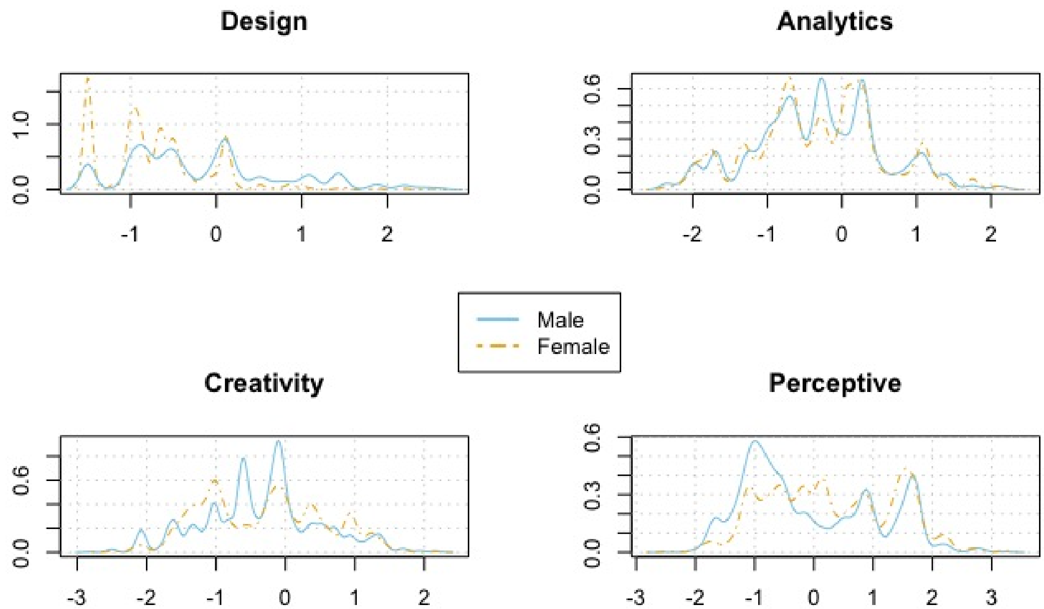

Figure 3: Density estimates of ONet job characteristics broken down by sex. The job characteristics have been mean centered and scaled to have unit variance.

across industry, culture, and society. For these selected attributes, ONet grades their relevance on a 100 point scale. Each attribute contains two 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 cobb-douglass style average with a 2/3's weight for importance and a 1/3 weight for the level scale.

The total number of individuals in the Census data numbers 105,796. After applying our 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 model performance. This leaves 68,642 individuals left as a training set. A summary descriptions of the covariates is provided in table 3.

For comparative purposes, we first estimate equation (52) for a two expert model. At this specification, there is no distinction between the HME and ME. A three

Table 3: 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
Perceptive	41.15	50.49	46.03	59.84

N = 68,642

expert model is then estimated for these two respective architectures to test whether there is an advantage to allowing the gating network to go deep, as opposed to staying shallow with a depth of one. A detail summary for these models will be explored further in this section. To provide context on these models' behaviour has they grow in complexity, each architecture is estimated up to six experts. References for these results also appear below but in less detail.

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:

$$log(wage) = Age + YrsEdu + Sex + Race + Occ \mid Age + YrsEdu + Sex + Race + Occ \tag{53}$$

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:

Table 4: Full Model Information

		Model Comparison			
Depth	Experts	Log-Lik	AIC	BIC	MSE
1	2	-0.541	1.082	1.088	0.182
1	3	-0.524	1.051	1.060	0.182
1	4	-0.537	1.078	1.091	0.181
1	5	-0.535	1.073	1.089	0.182
2	3				
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***

Log-Likelihood and BIC are divided by the sample size (68,642)

$$log(wage) = Age + YrsEdu \mid Age + YrsEdu + Sex + Race + Occ \tag{54}$$

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

$$log(wage) = Age \mid Age + YrsEdu + Sex + Race + Occ \tag{55}$$

Results for these regressions are collected in tables 6, 7, and 8. 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 (), 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.

We also compare the marginal effects across specifications. According to table (), the five-expert HME model performs the best for each of the three specifications. Table 9 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

Table 5: Returns to Years of Education

		Coefficient		
Depth	Experts	Min	Mid	Full
1	2	0.051	0.082	0.075
1	3	0.063	0.080	0.074
1	4	0.039	0.085	0.075
1	5	0.063	0.095	0.076
2	3			
2	4	0.063	0.078	0.073
3	5	0.068	0.075	0.069
3	6	0.067	0.074	0.069

OLS coef: 0.078

Log-Likelihood is divided by the sample size (68,642)

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

Table 6: Two Expert Full Specification Regression Results

	Coef.	SE	Coef.	SE
(Intercept)	1.259	0.017	1.494	0.014
Age16	0.031	0.001	0.068	0.001
Age16sq	-0.000	0.000	-0.002	0.000
YrsEduc	0.081	0.001	0.036	0.001
Sex	-0.243	0.004	-0.032	0.004
Black	-0.078	0.006	-0.045	0.006
Indian	-0.106	0.016	1.390	0.108
Asian	-0.045	0.010	0.036	0.009
Hisp	-0.121	0.006	-0.082	0.005
Creativity	-0.054	0.003	-0.008	0.003
Design	0.081	0.002	0.078	0.002
Analytics	0.134	0.003	0.112	0.003
Perceptive	0.064	0.002	-0.013	0.002
Variance	-1.653	0.007	-2.682	0.008
Share:		0.830		0.170

Log-Likelihood: 0.541

Table 7: Three Expert Deep Full Specification Regression Results

	Coef.	SE	Coef.	SE	Coef.	SE
(Intercept)	1.430	0.014	0.951	0.072	1.506	0.015
age16	0.018	0.001	0.037	0.003	0.051	0.001
age16sq	-0.000	0.000	-0.001	0.000	-0.001	0.000
yreduc	0.082	0.001	0.074	0.004	0.038	0.001
sex	-0.260	0.003	-0.122	0.017	-0.039	0.004
black	-0.089	0.006	-0.017	0.029	-0.047	0.006
indian	-0.118	0.016	0.103	0.055	-0.059	0.016
asian	-0.037	0.008	-0.042	0.029	0.055	0.010
hisp	-0.145	0.005	0.101	0.031	-0.066	0.005
Creativity	-0.045	0.003	-0.121	0.013	-0.022	0.003
Design	0.081	0.002	-0.058	0.013	0.070	0.003
Analytics	0.127	0.003	0.180	0.014	0.116	0.003
Perceptive	0.065	0.002	0.125	0.010	-0.003	0.002
variance	-1.884	0.005	-0.594	0.018	-2.729	0.009
share:		0.771		0.078		0.151

Log-Likelihood:-0.525

Table 8: Three Expert Wide Full Specification Regression Results

	Coef.	SE	Coef.	SE	Coef.	SE
(Intercept)						
age16						
age16sq						
yreduc						
sex						
black						
indian						
asian						
hisp						
Creativity						
Design						
Analytics						
Perceptive						
variance						
share:						

Log-Likelihood:

8 Conclusion

9 Miscellaneous

9.1 HMRE

Score function for HMRE

$$\frac{\partial Q}{\partial \omega_p^{a,i}} = \sum_t \sum_{m=1} \frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}} \sum_l \frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} \tag{56}$$

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 (56) is simply that same product chain but absent of node g^a times the partial of node g^a with respect to $\omega_p^{a,i}$.

$$\frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} = [\pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}(-g^a)] \frac{\partial g^{a,j(l)}}{\partial \omega^{a,i}} \tag{57}$$

Hessian for HMRE

$$\frac{\partial^2 Q}{\partial \omega_p^{a,i} \partial \omega_q^{a',j}} = \sum_t \sum_m \left[\frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}} \frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} - \left(\frac{h_t^{0,m}}{\sum_l \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}} \right)^2 \sum_l \sum_{l'} \frac{\partial \pi_{n_t^0 \overset{l}{\longleftrightarrow} P_t^m}}{\partial \omega_p^{a,i}} \frac{\partial \pi_{n_t^0 \overset{l'}{\longleftrightarrow} P_t^m}}{\partial \omega_q^{a',j}} \right] \tag{58}$$

Posterior node for HMRE

$$h_t^{a,m} = \frac{P_t^m \sum_l \pi_{g_t^a \overset{l}{\longleftrightarrow} P_t^m}}{\sum_k P^k \sum_l \pi_{g_t^a \overset{l}{\longleftrightarrow} P_t^k}} \tag{59}$$

9.2 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 9: 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 10: Full Marginal Effects

	OLS	2	3	4	5	6
Const.		1.321	1.359	1.262	1.359	1.405
Age		0.037	0.032	0.036	0.030	0.035
Age Sq		-0.001	-0.001	-0.001	-0.000	-0.001
Yrs Edu		0.072	0.072	0.073	0.070	0.070
Sex		-0.208	-0.209	-0.201	-0.191	-0.176
Afr Amer		-0.078	-0.079	-0.079	-0.076	-0.074
Indian		-0.096	-0.102	-0.095	-0.107	-0.046
Asian		-0.028	-0.030	-0.020	-0.009	-0.001
Hispanic		-0.118	-0.119	-0.098	-0.090	-0.088
Creativity		-0.044	-0.043	-0.054	-0.069	-0.060
Design		0.081	0.070	0.062	0.043	0.055
Analytics		0.129	0.128	0.131	0.144	0.140
Perceptive		0.055	0.058	0.057	0.051	0.052

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

Table 11: 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 12: 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.

9.3 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.

10 Diagnostics

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

density forecasts evaluations.

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