LETTER Communicated by Marcelo Medeiros

On Convergence Rates of Mixtures of Polynomial Experts

Eduardo F. Mendes
eduardo.mendes@northwestern.edu
Wenxin Jiang
wjiang@northwestern.edu
Department of Statistics, Northwestern University, Evanston, IL 60208, U.S.A.

In this letter, we consider a mixture-of-experts structure where m experts are mixed, with each expert being related to a polynomial regression model of order k. We study the convergence rate of the maximum likelihood estimator in terms of how fast the Hellinger distance of the estimated density converges to the true density, when the sample size n increases. The convergence rate is found to be dependent on both m and k, while certain choices of m and k are found to produce near-optimal convergence rates.

1 Introduction

Mixture-of-experts models (ME) (Jacobs, Jordan, Nowlan, & Hinton, 1991) and hierarchical mixture-of-experts models (HME) (Jordan & Jacobs, 1994) are powerful tools for estimating the density of a random variable Y conditional on a known set of covariates X. The idea is to "divide and conquer." We split the space of covariates and approximate the conditional density within each subspace. Additionally, it can be seen as a generalization of the classical mixture models, whose weights are constant across the covariate space. Mixture-of-experts have been widely used on a variety of fields, including image recognition and classification, medicine, audio classification, and finance. Such flexibility has also inspired a series of distinct models, those of including Wood, Jiang, and Tanner (2002), Carvalho and Tanner (2005a), Geweke and Keane (2007), Wood, Kohn, Cottet, Jiang, and Tanner (2008), Villani, Kohn, and Giordani (2009), Young and Hunter (2010), and Wood, Rosen, and Kohn (2011), among many others.

We consider a framework similar to Jiang and Tanner (1999a) and others. Assume each expert is a member of a one-parameter exponential family with mean $\varphi(h_k)$, where h_k is a kth-degree polynomial on the conditioning variables X (hence, a linear function of the parameters) and $\varphi(\cdot)$ is the inverse link function. In other words, each expert is a generalized linear model on a one-dimensional exponential family (GLM1). We allow the target density to be in the same family of distributions, but with conditional mean

$\varphi(h)$ where $h \in \mathcal{W}^{\infty}_{\alpha,K_0}$, a Sobolev class with α derivatives and bounding constant K_0 (see section 2.1) . Some examples of target densities include the Poisson, binomial, Bernoulli, and exponential distributions with unknown mean. Normal, gamma, and beta distributions also fall in this class if the dispersion parameter is known.

One might be skeptical about using (H)ME models with polynomial experts since it leads to more complex models as the degree k of the polynomials increases. We justify the use of such models through the approximation and estimation errors. We show that some choices of k and m lead to better convergence rates. This discussion about whether it is better to mix many simple models or fewer complex models is not new in the literature of mixture of experts. Earlier in the literature, Jacobs et al. (1991) and Peng, Jacobs, and Tanner (1996) proposed mixing many simple models; more recently, Wood et al. (2002) and Villani et al. (2009) considered using only a few complex models. Celeux, Hurn, and Robert (2000) and Geweke (2007) advocate for mixing fewer complex models, claiming that mixture models can be very difficult to estimate and interpret.

This work extends Jiang and Tanner (1999a) in some directions. We show that by including polynomial terms, one is able to improve the approximation rate on sufficiently smooth classes. This rate is sharp for the piecewise polynomial approximation with a fixed degree k and increasing number of "pieces," as shown in Windlund (1977). Moreover, we contribute to the literature by providing rates of convergence of the maximum likelihood estimator to the true density. We emphasize that such rates have never been developed for this class of models and the method used can be easily generalized to more general classes of mixture of experts. Convergence of the estimated density function to the true density and parametric consistency of the maximum likelihood estimator are also obtained.

Zeevi, Meir, and Maiorov (1998) show approximation in the L^p norm and estimation error for the conditional expectation of the ME with generalized linear experts. Jiang and Tanner (1999a) show consistency and approximation rates for the HME with generalized linear model as experts and a general specification for the gating functions. They consider the target density to belong to the exponential family with one parameter. Their approximation rate of the Kullback-Leibler divergence between the target density and the model is $O(1/m^{4/s})$, where m is the number of experts and s is the number of covariates or independent variables. Norets (2010) shows the approximation rate for the mixture of gaussian experts where both the variance and the mean can be nonlinear and the weights are given by multinomial logistic functions. The target density is considered to be a smooth, continuous function and the dependent variable Y to be continuous and satisfy some moment conditions. The approximation rate is $O(1/m^{s+2+1/(q-2)+\varepsilon})$, where Y is assumed to have at least q moments and ε is a sufficiently small number. Despite these findings, there are no convergence rates for the maximum likelihood estimator of mixture-of-experts class of models in the literature.

We show that under conditions similar to those of Jiang and Tanner (1999a), the approximation rate in Kullback-Leibler divergence is uniformly bounded by $((cs)^{k^*} m^{-k^*/s}/k^*!)^2$, where c is some positive constant not depending on k or m, $k^* = (k+1) \wedge \alpha$. This is a generalization of the rate found in Jiang and Tanner (1999a), who assume $\alpha = 2$ and $k = 1$. In squared Hellinger distance, the convergence rate of the maximum likelihood estimator to the true density is $O_p(((cs)^{k^*} m^{-k^*/s}/k^*!)^2 + d_{m,k} n^{-1} \log(d_{m,k} n))$, where $d_{m,k}$ is the total number of parameters in the model. To show the previous results, we do not assume identifiability of the model as it is natural for mixture of experts to be nonidentifiable under permutation of the experts (Jiang & Tanner, 1999c). Near-optimal nonparametric rates of convergence can be attained for some choices of k and m.

Throughout the letter, we use the notation $\|h(\cdot)\|_{p,\lambda(S)} = [\int_S |h(z)|^p d\lambda(z)]^{1/p}$ for all $0 < p < \infty$ and $\|h(\cdot)\|_{\infty,\lambda(S)} = \inf\{a \in \mathbb{R} : \lambda(\{z \in S : |h(z)| > a\}) = 0\}$. The former is the $L^p(\lambda)$ function norm of $h(\cdot)$ with respect to the measure λ over the set S and the latter $L^\infty(\lambda)$ the function norm of $h(\cdot)$ with respect to the measure λ over the set S. In the case that the set is not specified, we consider the entire support of the measure. Similarly, if the measure is omitted, $\|\cdot\|_p$ is the L^p vector norm, for $0 < p \leq \infty$.

The remainder of the letter is organized as follows. In the next section, we introduce the target density and mixture-of-experts models. We also demonstrate that the maximum likelihood estimator is consistent. Section 3 establishes the main results of the letter: approximation rate, convergence rate, and nonparametric consistency. Section 4 discusses model specification and the trade-off that we unveil between the number of experts and the degree of the polynomials. In the concluding remarks, we compare our results with Jiang and Tanner (1999a) and provide some directions for future research. The appendix differs from the main body of the letter for being more technical. In appendix A, we present the main steps in showing convergence rate in Hellinger distance. Appendix B has a set of useful lemmas required for proving the main results of the letter in appendix C.

2 Preliminaries

In this section we introduce the target class of density functions and the mixture-of-experts model with GLM1 experts.

2.1 Target Family of Densities. Consider a sequence of random vectors $\{(X_i', Y_i)'\}_{i=1}^n$ defined on $((\Omega \times A)^n, \mathcal{B}_{(\Omega\times A)^n}, P_{XY}^n)$ where $X \in \Omega \subset \mathbb{R}^s$, $Y \in A \subseteq \mathbb{R}$ and $\mathcal{B}_S$ is the Borel σ-algebra generated by the set S. We assume that P_{XY} has a density $p_{xy} = p_{y|x} p_x$ with respect to some measure λ. More precisely, we assume that p_x is known and $p_{y|x}$ belongs to a one-dimensional exponential family; the target density is such that

$$p_{y|x} = \exp\{ya(h(x)) + b(h(x)) + c(y)\}, \tag{2.1}$$

where $a(\cdot)$ and $b(\cdot)$ are known functions, three times continuously differentiable, with first derivative bounded away from zero, and $a(\cdot)$ has a nonnegative second derivative; $c(\cdot)$ is a known measurable function of Y. The function $h(\cdot)$ is an element of $\mathcal{W}^{\infty}_{\alpha,K_0}$, a Sobolev class of order α.[1] Throughout the letter, we denote the class of target density functions $p_{xy} = p_{y|x}p_x$ as $\mathcal{P}(\mathcal{W}^{\infty}_{\alpha,K_0})$.

The one-parameter exponential family of distributions includes the Bernoulli, exponential, Poisson, and binomial distributions. It also includes the gaussian, gamma, and Weibull distributions if the dispersion parameter is known. In this work, we focus only on the one-parameter case, but we conjecture that the results still hold in the case where the dispersion parameter has to be estimated from the data.

2.2 Mixture-of-Experts Model. The mixture-of-experts model with GLM1 experts is defined as

$$\begin{aligned} f_{m,k}(x, y; \zeta) &= \sum_{j=1}^{m} g_j(x; \nu)\, \pi(h_k(x, \theta_j), y) \cdot p_x \\ &= \sum_{j=1}^{m} g_j(x; \nu)\ \exp\{ya(h_k(x; \theta_j)) + b(h_k(x; \theta_j)) + c(y)\} \cdot p_x. \end{aligned} \tag{2.2}$$

Each $g_j(\cdot; \nu)$ is a positive function of x indexed by a v_m-dimensional parameter vector $\nu \in V_m \subset \ell^{\infty}(\mathbb{R}^{v_m})$,[2] and $\sum_{j=1}^{m} g_j(\cdot; \cdot) = 1$. The function $h_k(x; \theta_j)$ is a kth-degree polynomial on x indexed by a J_k-dimensional parameter vector $\theta_j \in \Theta_k \subset \ell^{\infty}(\mathbb{R}^{J_k})$. The parameter vector of the model is $\zeta = (\nu', \theta_1', \ldots, \theta_m')'$ and is defined on $Z_{m,k} = V_m \times \Theta_k^m \subset \ell^{\infty}(\mathbb{R}^{d_{m,k}})$, with $d_{m,k} = v_m + mJ_k$. Throughout the letter, we denote by $\mathcal{F}_{m,k}$ the class of (approximant) densities $f_{m,k}$.

To derive consistency and convergence rates, one needs to impose some restrictions on the functions π and g_j to avoid abnormal cases. This condition is not restrictive and is satisfied by the multinomial logistic weight functions and the Bernoulli, binomial, Poisson, and exponential experts, among many other classes of distributions and weight functions.

[1] Suppose $1 \le p \le \infty$ and $\alpha > 0$ is an integer. We define $\mathcal{W}^{p}_{\alpha,K_0}$ as the collection of measurable functions h with all partial derivatives $D^r h$, $|r| \le \alpha$, on $L^p(P_X)$, satisfying $\|D^r h\|_{p,P_X} \le K_0$. Here $D^r = \partial^{|r|}/(\partial^{r_1}x_1 \ldots \partial^{r_s}x_s)$ and $|r| = r_1 + \cdots + r_s$ for $r = (r_1, \ldots, r_s)$.

[2] We denote $\ell^{\infty}(\mathbb{R}^k) \equiv \{x \in \mathbb{R}^k : \|x\|_{\infty} < l\}$, for some finite l.

Assumption 1. *The next conditions hold jointly:*

i. The parameter space V_m is contained inside a hypercube of sufficiently large side l_1, possibly depending polynomially on m. Each parameter space Θ_k is contained in a hypercube of sufficiently large side l_2.
ii. For all $m' > m$, $\mathcal{F}_{m,k} \subseteq \mathcal{F}_{m',k}$.
iii. For each $\mathcal{F}_{m,k}$, there exists a square integrable function $F(x, y)$, such that for each (x, y),

$$\sup_{\zeta \in Z_{m,k}} \left| \frac{\partial}{\partial \zeta} \log f_{m,k}(x, y) \right|_\infty \leq F(x, y)$$

and

$$\int F(x, y)^2 f_{m,k}^{\infty} \, d\lambda \leq C(l),$$

where $f_{m,k}^{\infty}(x, y) = \sup_{\zeta \in Z_{m,k}} f_{m,k}(x, y; \zeta)$ and $C(l)$ is at most a polynomial function of l, where $l = l_1 \vee l_2$.

Remark 1. Note that if $\| f_{m,k}^{\infty} / p_{xy} \|_\infty < c$ for some finite c,

$$\int F(x, y)^2 f_{m,k}^{\infty} \, d\lambda(x, y) \leq c \int F(x, y)^2 \, dP_{XY}(x, y).$$

We present two examples of mixtures of experts satisfying the previous assumption. In the examples, we assume the gating functions are multinomial logistic functions and two distinct distributions: Bernoulli and Poisson. For simplicity, we take the number of covariates to be one (i.e., $s = 1$) and $\| \sup_\zeta f_{m,k} / p_{xy} \|_\infty < c$.

Example 1 (mixture of Bernoulli experts). A mixture of Bernoulli experts with multinomial logistic gating functions is given by:

$$f_m(y|x; \zeta) = \sum_{i=1}^{m} \frac{e^{\alpha_i + \beta_i x}}{\sum_{j=1}^{m} e^{\alpha_j + \beta_j x}} \phi(\theta_i x)^y \, (1 - \phi(\theta_i x))^{1-y},$$

where $\nu = (\alpha_1, \beta_1, \ldots, \alpha_m, \beta_m)'$, and $\phi(\cdot)$ is the logistic function. Condition i is satisfied by choosing an appropriate parameter space; condition ii is satisfied by the multinomial logistic functions if the parameter space for the α's increases logarithmic with m (see Ge & Jiang, 2006). The final condition can be shown by taking the derivatives of log $f_{m,k}$, which leads to a choice of $F(x, y) = |x|$.

Example 2 (mixture of Poisson experts). A mixture of Poisson experts with multinomial logistic gating functions is given by

$$f_m(y|x;\zeta) = \sum_{i=1}^{m} \frac{e^{\alpha_i+\beta_i x}}{\sum_{j=1}^{m} e^{\alpha_j+\beta_j x}} \frac{e^{y(\theta_i x)-e^{\theta_i x}}}{y!},$$

where $\nu = (\alpha_1, \beta_1, \ldots, \alpha_m, \beta_m)'$. Conditions i and ii have already been discussed. By taking the derivatives of $\log f_{m,k}$, one finds that it is sufficient to take $F(x, y) = \sup_\theta |x(y - e^{\theta x})|$.

2.3 Maximum Likelihood Estimation. We want to find the parameter vector $\hat{\zeta}_n = (\hat{\nu}_n', \hat{\theta}_n')'$ that maximizes the log-likelihood function of the data,

$$L_n(\zeta) = n^{-1} \sum_{i=1}^{n} \log \left\{ f_{m,k}(X_i, Y_i; \zeta)/\varphi_0(X_i, Y_i) \right\}, \tag{2.3}$$

where $\varphi_0(X, Y) = \exp(c(Y))p_x(X)$, that is,

$$\hat{\zeta}_n = \arg \max_{\zeta \in Z_{m,k}} L_n(\zeta). \tag{2.4}$$

The maximum likelihood estimator is not necessarily unique. In general, mixture-of-experts models are not identifiable under permutation of the experts. To circumvent this issue, one must impose restrictions on the experts and the weighting (or the parameter vector of the model), as shown in Jiang and Tanner (1999c).

Define the Kullback-Leibler (KL) divergence between p_{xy} and $f_{m,k}$ as

$$KL(p_{xy}, f_{m,k}) = \int_\Omega \int_A \log \frac{p_{xy}}{f_{m,k}} dP_{Y|X} dP_X. \tag{2.5}$$

The log-likelihood function in equation 2.3 converges to its expectation with probability one as the number of observations increases. Therefore, in the limit, the maximizer $\hat{f}_{m,k}$ of equation 2.3 (indexed by $\hat{\zeta}_n$) also minimizes the Kullback-Leibler divergence between the true density and the estimated density.

As stated earlier, this work considers only independent and identically distributed (i.i.d.) observations, but it is straightforward to extend the results to more general data-generating processes (e.g., martingales). The next assumption formalizes it.

Assumption 2 (data-generating process). *The sequence* $(X_i, Y_i)_{i=1}^n$, $n = 1, 2, \ldots$ *is an i.i.d. sequence of random vectors with common distribution* P_{XY}.

The next result ensures the existence of such estimator:

Theorem 1 (existence). *For a sequence $\{(Z_{m,k})_n\}$ of compact subsets of $Z_{m,k}$, $n = 1, 2, \ldots$, there exists a $\mathcal{B}(\Omega \times A)$−measurable function $\hat{\zeta}_n : \Omega \times A \to (Z_{m,k})_n$, satisfying equation 2.4 P_{XY}-almost surely.*

The maximum likelihood estimator $\hat{\zeta}$ consistently estimates ζ^*, where $\zeta^* \in Z_{m,k}$ is a minimizer of equation 2.5. We require the classic conditions: parametric identifiability and the existence of a unique minimizer of equation 2.5. If the i.i.d. condition fails, it can be shown that ergodicity of $(\log f_{m,k}(X_i, Y_i; \zeta))_{i=1}^n$ is a sufficient condition for consistency. However, conditions to ensure ergodicity of the log-likelihood function are out of the scope of this letter.

Assumption 3 (identifiability). *For any distinct ζ_1 and ζ_2 in $Z_{m,k}$, the two corresponding densities $f_{m,k}(\cdot, \cdot; \zeta_1)$ and $f_{m,k}(\cdot, \cdot; \zeta_2)$ are not almost everywhere equal.*

Jiang and Tanner (1999c) find sufficient conditions for identifiability of the parameter vector for the HME with one layer and Mendes, Veiga, and Medeiros (2006) for a binary tree structure. Both cases can be adapted to more general specifications. Although one can show consistency to a set, we adopt a more traditional approach requiring identifiability of the parameter vector.

Assumption 4 (unique maximizer). *Let $\zeta = (\nu', \theta')'$ and let ζ^* be the argument that minimizes $KL(p_{xy}, f_{m,k})$ over $\zeta \in Z_{m,k}$. Then*

$$det\left(\int \frac{\partial^2}{\partial\zeta\,\partial\zeta'} log\, f_{m,k}\big|_{\zeta=\zeta^*} dP_{XY}\right) \neq 0. \tag{2.6}$$

This assumption follows from a second-order Taylor expansion of the expected likelihood around the parameter vector that minimizes equation 2.5, denoted by ζ^*. We require the Hessian matrix in equation 2.6 to be invertible at ζ^*. The requirement for an identifiable unique maximizer is technical only in the sense that the objective function is not allowed to become too flat around the maximum. (For more discussion on this topic, see Bates & White, 1985, and White, 1996.) A similar assumption was made in the series of papers from Carvalho and Tanner (2005a, 2005b, 2006, 2007) and Zeevi et al. (1998) and is a usual assumption in the estimation of misspecified models.

Theorem 2 (parametric consistency of misspecified models). *Under assumptions 1, 2, 3, and 4, the maximum likelihood estimate $\hat{\zeta} \to \zeta^*$ as $n \to \infty$ P_{XY}-almost surely.*

Huerta, Jiang, and Tanner (2003) and the series of papers by Carvalho and Tanner (2005a, 2005b, 2006, 2007) derive similar results for time series processes.

3 Main Results

3.1 Approximation Rate. We follow Jiang and Tanner (1999a) to bound the approximation error. Before presenting the main conditions, we introduce some key concepts.

Definition 1 (fine partition). *For $m = 1, 2, \ldots$, let $Q^m = \{Q_j^m\}_{j=1}^{r_m}$ be a partition of Ω. If $m \to \infty$ and if for all $x_1, x_2 \in Q_j^m$, $\|x_1 - x_2\|_\infty \leq c_0 / r_m^{1/s}$, for some constant c_0 independent of x_1, x_2, m or j. Then $\{Q^m,\ m = 1, 2, \ldots\}$ is called a sequence of fine partitions with cardinality r_m and bounding constant c_0.*

The key idea behind the approximation rate is to control the approximation rate inside each fine partition of the space. More precisely, bound the approximation inside the "worst" (i.e., most difficult to approximate) partition. We need the following assumption:

Assumption 5. *For a fine partition Q^m of Ω, with bounding constant c_0 and cardinality sequence $r_m = \lfloor m^{1/s} \rfloor^s$, $m = 1, 2, \ldots$, there exists a constant $c_1 > 0$, and a parameter vector $v_{c_1} \in V_m$ such that*

$$\max_{1 \leq j \leq r_m} \|g_j(\cdot; v_{c_1}) - I_{Q_j^m}(\cdot)\|_{1, P_X} \leq \frac{c_1}{r_m}, \tag{3.1}$$

where $I_Q(x) = 1$ if $x \in Q$ and 0 otherwise.

This assumption is similar to the one employed in Jiang and Tanner (1999a) and requires that the vector $g = (g_1, \ldots, g_{r_m})$ approximate the vector of indicator functions $(I_{Q_1^m}, \ldots, I_{Q_{r_m}^m})$ at a rate not slower than $O(r_m)$.

The cardinality r_m is similar to Jiang and Tanner (1999b), who essentially used $r_m = \lfloor m^{1/s} \rfloor^s$ so as to form a regular hypercube partition of an s-dimensional domain of x.

We show the approximation rate in a divergence measure stronger than the Kullback-Leibler divergence. Let p and q denote densities. The χ^2 divergence (see, e.g., Wong & Shen, 1995) between p and q is

$$\chi^2(p, q) = \int \frac{(p - q)^2}{q} d\lambda = \int p \left(\frac{p}{q} - 1 \right) d\lambda, \tag{3.2}$$

where the equivalence is obtained by expanding the squares and integrating the densities to 1.

Theorem 3 (approximation rate). *Let* $p \in \mathcal{P}(\mathcal{W}^{\infty}_{\alpha,K_0})$ *and* $f_{m,k} \in \mathcal{F}_{m,k}$. *If assumptions 1 and 5 hold, then*

$$\sup_{p} \inf_{f_{m,k}} \chi^2(p, f_{m,k}) \leq \left(\frac{(cs)^{k^*}}{m^{k^*/s}\, k^*!} \right)^2, \tag{3.3}$$

where $k^* = \alpha \wedge (k+1)$, *and some positive constant* c *not depending on* m *or* k. *It also follows that*

$$\sup_{p} \inf_{f_{m,k}} KL(p, f_{m,k}) \leq \left(\frac{(cs)^{k^*}}{m^{k^*/s}\, k^*!} \right)^2, \tag{3.4}$$

This result is a generalization of Jiang and Tanner (1999a) in three directions. First, we allow the target function $h(\cdot)$ to be in a Sobolev class with α derivatives; second, we consider a polynomial approximation to the target function in each expert (in fact, their result is a special case when $\alpha = 2$ and $k = 1$); finally, we consider a divergence measure stronger than the Kullback-Leibler divergence. The result also holds under more general specifications of densities and experts. If a dispersion parameter has to be estimated from the data, we have to modify lemma 4 accordingly, and the same result holds.

The approximation rate also agrees with the optimal one for approximating functions on $\mathcal{W}^{\infty}_{\alpha,K_0}$ by piecewise polynomials of a fixed degree (Windlund, 1977). Under assumption 5, it is exactly what we are doing, and therefore this approximation rate is sharp, meaning one cannot derive faster rates for fixed k.

3.2 Convergence Rate. In the previous section, we found a bound for the approximation error. In this section, we will find the estimation error and combine it with the approximation error to derive the rate of convergence. The estimation error describes how far the estimated function is from the best approximant in the class. In this work, the convergence rates are derived with respect to the squared Hellinger distance, defined below. For any two densities p and q with respect to λ, the squared Hellinger distance is

$$d_h^2(p, f) = \frac{1}{2} \int \left(\sqrt{p} - \sqrt{q} \right)^2 d\lambda.$$

The estimation error, measured in the squared Hellinger distance, is $O_p(d_{m,k}n^{-1}\log(d_{m,k}n))$. We also show that some choices of k and m achieve near-optimal convergence rates.

The next theorem summarizes the convergence rate of the maximum likelihood estimator $\hat{f}_{m,k}$ with respect to the squared Hellinger distance between the true density p_{xy} and the estimated density.

Theorem 4 (convergence rate). *Let $p_{xy} \in \mathcal{P}(\mathcal{W}^{\infty}_{\alpha,K_0})$ and $\hat{f}_{m,k}$ denote its maximum likelihood estimator on $\mathcal{F}_{m,k}$. Let $m = m_n$ be such that $m(n^{-1}\log n) \to 0$ as n increases. Under assumptions 1, 2, and 5,*

$$d_h^2(p_{xy}, \hat{f}_{m,k}) = O_p\left(\frac{(cs)^{2k^*}}{(k^*!)^2 m^{2k^*/s}} + d_{m,k}\frac{\log(d_{m,k}n)}{n}\right), \tag{3.5}$$

where $k^* = \alpha \wedge (k+1)$ and c is a positive constant. In particular, if we assume $v_m = O(m)$ and let m be proportional to $(n/\log n)^{s/(2k^*+s)}$, then

$$d_h^2(p_{xy}, \hat{f}_{m,k}) = O_p\left(\left(\frac{\log n}{n}\right)^{\frac{2k^*}{2k^*+s}}\right). \tag{3.6}$$

Remark 2. Although the previous result is derived for the i.i.d. case, the result also holds for a more general data-generating process. This convergence rate is close to the optimal rate found in the sieves literature if $k^* = \alpha$ (see, e.g., Stone, 1980, and Barron & Sheu, 1991). To derive the convergence rate, we do not assume $f^*_{m,k}$ is a unique, identifiable maximizer of the log-likelihood function 2.3. Here, $f^*_{m,k}$ is allowed to be any of such maximizers. The price to pay for such generality is the inclusion of the "$\log n$" term in the convergence rates.

3.3 Consistency. We apply the previous result to show that the maximum likelihood estimator is consistent, that is, the Hellinger distance between the true density and the estimated model approaches zero as the sample size n and the index of the approximation class m go to infinity.

Corollary 1 (consistency) *Let $p_{xy} \in \mathcal{P}(\mathcal{W}^{\infty}_{\alpha,K_0})$ and $\hat{f}_{m,k}$ denote its maximum likelihood estimator on $\mathcal{F}_{m,k}$. Allow $m = m_n$ and $m(n^{-1}\log n) \to 0$ as n increases. Under assumptions 1, 2, and 5, $d_h(p_{xy}, \hat{f}_{m,k}) \to 0$ as n increases.*

4 Effects of m and k

Two important problems in the area of ME are: (1) What number of experts m should be chosen, given the size n of the training data, and (2) Given the total number of parameters, whether it is better to use a few complex

experts or combine many simple experts. Our results will not be able to answer these questions completely, but they can provide some qualitative insights. We provide a related theoretical result.

Start by setting $d_h^2 \equiv d_h^2(p_{xy}, \hat{f}_{m,k})$ and $d = m(k+1)^s$, an upper bound in the number of parameters $d_{m,k}$. The convergence rate of d_h^2 in equation 3.5 can be upper-bounded by a simpler expression

$$U \equiv U(m,k,n) = \left(\frac{cs}{k^* m^{1/s}}\right)^{2k^*} + \frac{mk_1^s \log(mk_1^s n)}{n},$$

such that $k_1 = k+1$, $k^* = k_1 \wedge \alpha$, and c is a positive constant. This assumes that $v(m) = O(m)$ and uses the fact that the number of parameters needed in s-dimensional polynomials of order k is bounded by $J_k \leq (k+1)^s$. We have also used a lower bound of the factorial based on Stirling's formula. We now study the upper bound U.

Proposition 1. *Let c be a positive constant that does not depend on n but can take different values at different places. Let $k_1 = k+1$, $k^* = k_1 \wedge \alpha$, and*

$$U \equiv \left(\frac{cs}{k^* m^{1/s}}\right)^{2k^*} + \frac{mk_1^s \log(mk_1^s n)}{n}, \tag{4.1}$$

(which is an upper bound for the d_h^2 convergence rate derived in theorem 4). Let $d = mk_1^s$, which is a bound for the approximate order of the total number of parameters.

Then the following statements are true:

I. Consider the case where α is finite:

Ia. As $n \to \infty$, we have $U \to 0$ if $m \to \infty$ and $d = o(n/\log n)$.

Ib. U achieves a near optimal rate $O(n^{-2\alpha/(s+2\alpha)}(\log n)^c)$ for some $c > 0$, under the following choices:

- *$k_1 \geq \alpha$ and $k_1 = O((\log n)^c)$ for some $c > 0$.*
- *m is of order $n^{s/(s+2\alpha)}(\log n)^{c'}$ for any constant $c' \in \mathbb{R}$.*

II. Consider the case where $\alpha = \infty$ (or $\alpha \geq k_1$):

IIa. As $n \to \infty$, we have $U \to 0$ if $d \to \infty$ and $d = o(n/\log n)$.

IIb. The following choices will make U to have a "near-parametric rate" $U = O((\log n)^c/n)$ for some $c > 0$:[3]

- *$m \geq 1$ and $m = O((\log n)^c)$ for some $c > 0$.*
- *$k_1 \geq c \log n$ for any constant $c > 0$, and $k_1 = O((\log n)^c)$ for some $c > 0$.*

Remark 3. (a) The results above do not completely answer the earlier questions and on how to choose m and k in practice. For example, the

[3] "Near-parametric rate" stands for "close to the parametric rate $O(1/n)$."

results on m and k are known only up to some order in n. In addition, the convergence rates may depend on the smoothness parameter α, which may be unknown in practice. A practical method of the choice of m and k may involve a complexity penalty or cross-validation and is outside of the scope of this letter. On the other hand, some qualitative insights could be useful from our convergence rate analysis.

(b) For the very smooth situation, result IIa suggests that for the purpose of consistency (which means the convergence of d_h^2 to 0 in probability), question ii about the ratio between $(m, k+1)$ is not relevant as long as $d = m(k+1)^s$ grows to infinity at a rate slower than $n/\log n$. However, consistency is not enough to guarantee a good performance. For example, equation 4.1 suggests that for $s = 1$, $(m, k+1) = (\log n, 1)$ will lead to a very slow rate $O((\log n)^{-2})$, and $(m, k+1) = (1, \log n)$ will lead to a very fast rate $O((\log n)^2/n)$, and in both cases the total number of parameters $d = m(k+1)$ are the same. It is therefore important to look into the convergence rates.

(c) Results Ib and IIb imply that smoother target functions (with large α) and lower dimensions (s) generally encourage using fewer experts. For finite α, the near-optimal rates described in result Ib are achieved when $m \gg k$ in order. For the very smooth situation $\alpha = \infty$, even $m = 1 (\ll k)$ can lead to near-optimal performances.

(d) We note that *near*-optimal convergence rates can always be achieved with k_1 not being too large compared to the sample size n. This is summarized in the two situations in results Ib and IIb, where we see that even in the case $\alpha = \infty$, we only need about $k_1 \sim \log n$ for us to achieve a near-parametric convergence rate.

(e) Although in result Ib (with finite α) we have used $m \gg k$ to achieve near-optimal rates, we conjecture that even with $m = 1$, a good (but perhaps suboptimal) convergence rate can be attained. For example, for $s = 1$, using the Legendre approximation technique 7.5 of Barron and Sheu (1991), we conjecture that a convergence rate is of the form $d_h^2 = O_p((k_1^2(c/d)^{2k^*} + (d/n)\log(dn))$, where $d = mk_1$ and c is a positive constant. Therefore (denoting $\alpha^* = \alpha - 1$), even when $m = 1$, we can still take k_1 to be of order $n^{1/(2\alpha^*+1)}$ and get $d_h^2 = O_p(n^{-2\alpha^*/(2\alpha^*+1)}\log n)$, which is suboptimal compared to result Ib but is still converging to 0 if $\alpha > 1$. [Similarly, we conjecture that $m \to \infty$ is not necessary for the consistency result Ia; we need only $d \to \infty$ and $d = o(n/\log n)$.]

5 Conclusion

In this letter, we study the mixture-of-experts model with m experts in a one-exponential family with conditional mean $\varphi(h_k)$, where h_k is a kth order polynomial and $\varphi(\cdot)$ is the inverse link function. We derive the approximation rate and convergence rate of the maximum likelihood

estimator to densities in a one-parameter exponential family with mean $\varphi(h)$ with $h \in \mathcal{W}^{\infty}_{\alpha,K_0}$, a Sobolev class with α derivatives, and bounding constant K_0. We found that the convergence rate of the maximum likelihood estimator to the true density in squared Hellinger distance is $O_p((cs)^{2k^*}/(k^*!m^{k^*/s})^2 + d_{m,k}n^{-1}\log(d_{m,k}n))$, for $k^* = (k+1) \wedge \alpha$ and c some positive constant.

We discuss choices of k and m for achieving good convergence rates. The results of this letter can be generalized to more complex target densities (from, e.g., a Besov or a piecewise Besov class) and models (e.g., mixture of trigonometric polynomials or wavelets) with simple modifications to the proofs.

We generalize Jiang and Tanner (1999a) in several directions: (1) we assume one can include polynomial terms of the variables on the GLM1 experts; (2) we assume the target density is in a $\mathcal{W}^{\infty}_{\alpha,K_0}$ class, for $\alpha > 0$, instead of $\mathcal{W}^{\infty}_{2,K_0}$; (3) we show consistency of the maximum likelihood estimator for a fixed number of experts; (4) we calculate convergence rates of the maximum likelihood estimator in squared Hellinger distance; (5) we show consistency when the number of experts and the sample size increase; and, finally, (6) we find that using polynomials in the experts, one can yield better estimation and error bounds. These developments have shed light on the important question of how the number of experts and the complexity of the experts jointly affect the convergence rate.

Appendix A: Showing the Convergence Rate

In this appendix we explain and justify the main steps in proving the convergence rate.

One of the drawbacks of working with the Kullback-Leibler divergence is that it is not bounded. We will use the Hellinger distance:

Definition 2 (Hellinger distance). *Let P and Q denote two probability measures absolutely continuous with respect to some measure λ. The Hellinger distance between P and Q is given by*

$$d_h(P, Q) = \left\{ \frac{1}{2} \int \left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}} \right)^2 d\lambda \right\}^{1/2}. \tag{A.1}$$

Alternatively, the Hellinger distance between two densities p and q with respect to λ is given by

$$d_h(p, q) = \left\{ \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, d\lambda \right\}^{1/2}. \tag{A.2}$$

The next lemma summarizes basic inequalities well known in the literature (e.g., Wong & Shen, 1995) relating the Hellinger distance, the Kullback-Leibler divergence, and the χ^2 divergence.

Lemma 1. *Let $p_{xy} = dP/d\lambda$. For $f \in \mathcal{F}_{m,k}$ we have*

$$2d_h^2(p_{xy}, f) \leq KL(p_{xy}, f) \leq \chi^2(p_{xy}, f).$$

In order to bound the estimation error, we use results from the theory of empirical processes. The convergence rate theorem presented below is derived for the i.i.d. case; however, the same result holds for martingales (see van der Geer, 2000).

The control of estimation rate inside a class of functions requires the knowledge of the complexity of the functional class. Denote by $N_B(\varepsilon, \mathcal{F}, \|\cdot\|)$ the number of ε-brackets, with respect to the distance $\|\cdot\|$, needed to cover the set $\mathcal{F}$ and $H_B(\varepsilon, \mathcal{F}, \|\cdot\|) = \log N_B(\varepsilon, \mathcal{F}, \|\cdot\|)$—the respective bracketing entropy.[4] The use of a bracketing entropy to assess the complexity of a class of mixture of regressions is not new. Genovese and Wasserman (2000) and Viele and Tong (2002) use the entropy with bracketing to measure the complexity in a class of mixture models and mixture of regressions, respectively. Applying their method in our setting gives a bracketing entropy of the same order as if one employs the method used in this letter. We use the latter for the ease of exposition.

We are particularly interested in the class of functions

$$\bar{\mathcal{F}}_{m,k}^{1/2}(\delta) = \left\{ \sqrt{\frac{f + f^*}{2}} : f \in \mathcal{F}_{m,k},\ d_h\left(\frac{f + f^*}{2}, f^*\right) \leq \delta \right\},$$

for some fixed $f^* \in \mathcal{F}_{m,k}$, and $0 < \delta \leq 1$.

Let $d_{m,k} = v_m + mJ_k$, $Z_{m,k} = V_m \times \Theta_{mk}$, and use c for any arbitrary positive constant that may change its value every time it appears.

Lemma 2 (bracketing entropy). *Under assumption 1, for any $0 < \delta \leq d_{m,k}^a/\sqrt{e}$ and some $a \geq 1$,*

$$H_B(\delta, \bar{\mathcal{F}}_{m,k}^{1/2}, \|\cdot\|_2) \leq c\, d_{m,k} \log \frac{d_{m,k}^a}{\delta}, \tag{A.3}$$

where

$$\int_{0^+}^{\delta} H_B^{1/2}(u, \bar{\mathcal{F}}_{m,k}^{1/2}, \|\cdot\|_2)\, du \leq c\, d_{m,k}^{1/2} \delta \log^{1/2} \frac{d_{m,k}^a}{\delta}. \tag{A.4}$$

[4] For a formal definition of bracketing numbers, see van der Vaart and Wellner (1996).

Proof. The proof of equation A.3 makes use of lemma 6. Set $g_i \in \bar{\mathcal{F}}^{1/2}_{m,k}(\delta)$, $i = 1, 2$. Each function g_i can be written as $\sqrt{(f_{\zeta_i} + f^*)/2}$, for $\zeta_i \in Z_{m,k}$, which depends on ζ only through f_{ζ_i}. Then for each (x, y),

$$|g_1(x, y) - g_2(x, y)| \leq d_{m,k} \, \|\partial_\zeta g(x, y)|_{\zeta=\tilde{\zeta}}\|_\infty \, \|\zeta_1 - \zeta_2\|_\infty.$$

The derivative on the right-hand side can be bounded by

$$\|\partial_\zeta g(x, y)|_{\zeta=\tilde{\zeta}}\|_\infty \leq \sup_{\zeta \in Z_m} \sqrt{\frac{f_\zeta}{2}} \|\partial_\zeta \log f_\zeta\|_\infty.$$

By applying lemma 6 with $\|\cdot\| = \|\cdot\|_2$ and $F = d_{m,k} \sup_{\zeta \in Z_m} \sqrt{\frac{f_\zeta}{2}}$ $\|\partial_\zeta \log f_\zeta\|_\infty$, we have that

$$\|F(x, y)\|_2 = \frac{d_{m,k}}{\sqrt{2}} \left(\int \sup_{\zeta \in Z_{m,k}} f_\zeta \left| \partial_\zeta \log f_\zeta \right|^2_\infty d\lambda \right)^{1/2},$$

which is bounded by $C(l)\, d_{m,k}/\sqrt{2}$ by assumption 1. The number of ε-balls with respect to L_∞ needed to cover Z_m is

$$N(\varepsilon, Z_{m,k}, |\cdot|_\infty) \leq \left(\frac{l}{\varepsilon} + 1 \right)^{d_{m,k}},$$

because by assumption 1, Z_m is a hypercube with side l. It follows from lemma 6 that

$$N_B\left(\delta, \bar{\mathcal{F}}^{1/2}_{m,k}, \|\cdot\|_2\right) \leq \left(\sqrt{8} \frac{l\, C(l)\, d_{m,k}}{\delta} \right)^{d_{m,k}}.$$

Since l is polynomial in $d_{m,k}$, we can take $l\, C(l) \leq c\, d^{a-1}_{m,k}/\sqrt{8}$, for some $a \geq 1$ and $c > 0$. Taking the log, we obtain equation A.3. Then we apply lemma 5.

We use a modified version of theorem 10.13 in van der Geer (2000) to show the rate of convergence of the Hellinger distance between the maximum likelihood estimator and the true density. This modification allows

for unbounded likelihood ratios, that is, it relaxes the assumption that $\|p_{xy}/f^*_{m,k}\|_{\infty,\lambda}$ is bounded.

Theorem 5 (modified version of theorem 10.13 in van der Geer, 2000). *Let $\hat{f}_{m,k}$ denote the maximum likelihood estimator of p_{xy} over $\mathcal{F}_{m,k}$. Set*

$$\bar{\mathcal{F}}^{1/2}_{m,k}(\delta) = \left\{ \sqrt{\frac{f+f^*}{2}} : f \in \mathcal{F}_{m,k},\, d_h\left(\frac{f+f^*}{2}, f^*\right) \le \delta \right\},$$

for some fixed $f^* \in \mathcal{F}_{m,k}$ satisfying $f^* > 0$ λ-a.e. Choose

$$\Psi(\delta) \ge \int_{0^+}^{\delta} H_B^{1/2}\left(u, \bar{\mathcal{F}}^{1/2}_{m,k}(\delta), \|\cdot\|_2\right) du \vee \delta,$$

in such a way that $\Psi(\delta)/\delta^2$ is a nonincreasing function of δ. Then, for $\sqrt{n}\delta_n^2 \ge const.\,\Psi(\delta_n)$, we have

$$d_h^2(p_{xy}, \hat{f}_{m,k}) = O_p\left(\delta_n^2 + \chi^2(p_{xy}, f^*)\right).$$

Sketch Proof. The proof is parallel to the one of theorem 10.13 in van der Geer (2000). At line 3 of the displayed equations on p. 191, the second term,

$$\int_{f^*>0} \left(\sqrt{f^*} - \sqrt{\frac{\hat{f}_{m,k}+f^*}{2}}\right)\left(\sqrt{p_{xy}} - \sqrt{f^*}\right)\left(1+\sqrt{\frac{p_{xy}}{f^*}}\right) d\lambda,$$

which is lower-bounded by, using the Cauchy-Schwarz inequality,

$$-2(1+c)d_h\left(f^*, \frac{\hat{f}_{m,k}+f^*}{2}\right) d_h(p_{xy}, f^*),$$

using the uniform bound condition on the likelihood ratio, that is, $\|p_{xy}/f^*_{m,k}\|_{\infty,\lambda} \le c$ (van der Geer, 2000, eq. 10.69). We modify the proof as follows:

$$\int_{f^*>0} \left(\sqrt{f^*} - \sqrt{\frac{\hat{f}_{m,k}+f^*}{2}}\right)\left(\sqrt{p_{xy}} - \sqrt{f^*}\right)\left(1+\sqrt{\frac{p_{xy}}{f^*}}\right) d\lambda$$

$$\ge -\left(\int_{f^*>0} \left(\sqrt{f^*} - \sqrt{\frac{\hat{f}_{m,k}+f^*}{2}}\right)^2 d\lambda\right)^{1/2}$$

$$\times \left(\int_{f^*>0} \left(\sqrt{p_{xy}} - \sqrt{f^*} \right)^2 \left(1 + \sqrt{\frac{p_{xy}}{f^*}} \right)^2 d\lambda \right)^{1/2}$$

$$= -\sqrt{2} d_h \left(f^*, \frac{\hat{f}_{m,k} + f^*}{2} \right) \left(\chi^2(p_{xy}, f^*) \right)^{1/2}.$$

This allows us to proceed with the χ^2-divergence without needing to bound the densities as in equation 10.69 of van der Geer (2000).

Appendix B: Auxiliary Results

In the next lemma, we use the notation $\partial_\theta = \partial/\partial\theta$, $\partial_{\theta\theta'} = \partial^2/\partial\theta\partial\theta'$, $a_j = a(h_k(x;\theta_j))$, $\dot{a}_j = \partial_\theta a_j$, $\ddot{a}_j = \partial_{\theta\theta'} a_j$ and so on.

Lemma 3. *Let $f \in \mathcal{F}_{m.k}$. Under assumption 1,*

- $\mathbb{E}|log\ f| \leq \infty$
- $\mathbb{E}|\nabla log\ f| \leq \infty$
- $\mathbb{E}\|\nabla log\ f\|_2^2 \leq \infty$
- *if we further assume 3 and 4, then $\mathbb{E}|\nabla^2 log\ f| \leq \infty$ and is nonsingular at ζ^*.*

Proof. This theorem is proved by calculating the derivatives and bounding it. First, note that a_j and b_j are continuous differentiable functions of $h_k(x;\theta_j)$. Since $|h_k(x;\theta_j)| \leq |\theta_j| < \sqrt{J_k}\|\theta_j\|_2 < \infty$ for any fixed k, then both a_j and b_j are also bounded. The same reasoning can be applied to $\dot{a}_j$, $\dot{b}_j$, $\ddot{a}_j$, and $\ddot{b}_j$. Also, by definition, $\mathbb{E}|Y|^p < \infty$ for any $p \geq 0$. Then

$$\mathbb{E} \log f \leq \mathbb{E} \left[\log \sum_{j=1}^{m} g_j \pi_j \right] \leq \mathbb{E} |\max_j [Y a_j + b_j + c(Y)]| < \infty.$$

Let $\delta_j = g_j e^{Y a_j + b_j + c(Y)} p_x / f \leq 1$ and $c^* = \max_j \|\partial_\nu \log g_j\|_{\infty,\Omega}$. Then

$$\mathbb{E}|\partial_{\theta_j} \log f| = \mathbb{E}|\delta_j (Y \dot{a}_j + \dot{b}_j) X| < \infty,$$

$$\mathbb{E}|\partial_\nu \log f| = \mathbb{E} \left| \frac{\sum \dot{g}_j e^{Y a_j + b_j}}{f} \right| \leq m c^* < \infty.$$

The same follows for $\mathbb{E}|\partial_\theta \log f|_2^2$, $\mathbb{E}|\partial_\nu \log f|_2^2$ and $\mathbb{E}|\partial_\theta \log f \partial_\nu \log f|$. Let $\dot{c}^* = \|\partial_\nu \log \dot{g}_j\|_{\infty,\Omega}$, and choose any vector α with appropriate dimensions

satisfying $\alpha'\alpha = 1$. Then

$$\begin{aligned}
\mathbb{E}\alpha'|\partial_{\theta_j\theta_j'}\log f|\alpha &= \mathbb{E}\alpha|\delta_j(1-\delta_j)(Y\dot{a}_j+\dot{b}_j)^2XX' + \delta_j(Y\ddot{a}_j+\ddot{b}_j)XX'|\alpha \\
&\le 0.25\mathbb{E}|Y\dot{a}_j+\dot{b}_j|^2 + \mathbb{E}\max_j|Y\ddot{a}_j+\ddot{b}_j| < \infty, \\
\mathbb{E}\alpha'|\partial_{\theta_j\theta_k'}\log f|\alpha &= \mathbb{E}\alpha'|-\delta_j\delta_k(Y\dot{a}_k+\dot{b}_k)(Y\dot{a}_j+\dot{b}_j)XX'|\alpha \\
&\le \mathbb{E}|(Y\dot{a}_k+\dot{b}_k)(Y\dot{a}_j+\dot{b}_j)| < \infty, \\
\mathbb{E}\alpha'|\partial_{\theta_j\nu'}\log f|\alpha &= \mathbb{E}\alpha'\left|\frac{e^{Ya_j+b_j}(Y\dot{a}_j+\dot{b}_j)X\dot{g}'p_x}{f}\right|\alpha \\
&\le \mathbb{E}\alpha'|\delta_j(Y\dot{a}_j+\dot{b}_j)X1_{\boldsymbol{v}_m}{}'|\alpha c^* \\
&\le \mathbb{E}|Y\dot{a}_j+\dot{b}_j|c^* < \infty, \\
\mathbb{E}\alpha'|\partial_{\nu\nu'}\log f|\alpha &= \mathbb{E}\alpha'\left|\frac{\sum_j\ddot{g}_je^{Ya_j+b_j}p_x}{f} - \frac{\sum_j\dot{g}_je^{Ya_j+b_j}p_x}{f}\frac{\sum_j\dot{g}_j'e^{Ya_j+b_j}p_x}{f}\right|\alpha \\
&\le c^*|\dot{c}^*| + c^{*2} < \infty.
\end{aligned}$$

Since ζ^* is a maximizer of $\mathbb{E}\log f$ over $\mathcal{F}_{m,k}$, $\mathbb{E}|\nabla^2\log f|$ has to be nonnegative definite. Assumption 4 tells us it is also invertible; therefore, $\mathbb{E}|\nabla^2\log f|$ is positive definite.

We use the next lemma to bound uniformly the approximation rate of the family of functions $\mathcal{F}_{m,k}$ with respect to the χ^2 divergence. Define the upper divergence between $p \in \mathcal{P}(\mathcal{W}^\infty_{\alpha,K_0})$ and $f_{m,k} \in \mathcal{F}_{m,k}$ as

$$\mathcal{D}(p, f_{m,k}) = \int_\Omega \sum_{j=1}^m g_j(x,\nu)(h_k(x;\theta_j) - h(x))^2 dP_X(x). \tag{B.1}$$

We can use the upper divergence to bound the χ^2 divergence.

Lemma 4. *Let $p \in \mathcal{P}(\mathcal{W}^\infty_{\alpha,K_0})$ and $f_{m,k} \in \mathcal{F}_{m,k}$, then, under condition 1(i),*

$$\chi^2(p, f_{m,k}) \le M_\infty \mathcal{D}(p, f_{m,k}),$$

where M_∞ is a finite, positive constant (see the proof of the lemma for a closed-form expression).

Proof. It follows from the definition of χ^2 divergence and concavity of the logarithm that for any $f \in \mathcal{F}_{m,k}$ and $p \in \mathcal{P}(\mathcal{W}^\infty_{\alpha,K_0})$,

$$\chi^2(p, f_{m,k}) \leq \int_\Omega \sum_{j=1}^m g_j(x;\nu) \int_A \left(e^{y(2a-a_j)+2b^*(a)-b^*(a_j)+c(y)} - 1\right) dy dP_X(x),$$

where $a = a(h(x))$, $a_j = a(h_k(x;\theta_j))$, and $b^*(a) = b(h(x))$. Consider the identity

$$\int_A \exp\{ya + c(y)\}\, dy = \exp\{-b^*(a)\};$$

hence,

$$\int_A e^{y(2a-a_j)+2b^*(a)-b^*(a_j)+c(y)} dy = e^{-b^*(2a-a_j)+2b^*(a)-b^*(a_j)},$$

which does not depend on y. A second-order Taylor expansion of $b^*(2a - a_j)$ and $b^*(a_j)$ gives us, respectively,

$$b^*(2a-a_j) = b^*(a) + \dot{b}^*(a)(a-a_j) + \frac{1}{2}\ddot{b}^*(\tilde{a}_j)(a-a_j)^2,$$
$$b^*(a_j) = b^*(a) - \dot{b}^*(a)(a-a_j) + \frac{1}{2}\ddot{b}^*(\bar{a}_j)(a-a_j)^2,$$

for $\tilde{a}_j$ and $\bar{a}_j$ on the line connecting a and a_j.

Adding up these equations and subtracting from $2b^*(a)$ gives

$$e^{-b^*(2a-a_j)+2b^*(a)-b^*(a_j)} = e^{(a-a_j)^2\frac{1}{2}\left[-\ddot{b}^*(\tilde{a}_j)-\ddot{b}^*(\bar{a}_j)\right]}.$$

Call $M^j \equiv M^j(x) \geq |(1/2)[-\ddot{b}^*(\tilde{a}_j) - \ddot{b}^*(\bar{a}_j)]|$. Use the inequality $e^{|x|} - 1 \leq |x|e^{|x|}$ and the mean value theorem to show that for some $\tilde{h}$,

$$e^{(a-a_j)^2\frac{1}{2}\left[-\ddot{b}^*(\tilde{a}_j)-\ddot{b}^*(\bar{a}_j)\right]} - 1 \leq |h - h_k(x,\theta_j)|^2 |\dot{a}(\tilde{h}_j) M^j(x) e^{M^j(x)(a-a_j)^2}|.$$

Choose $M_\infty \geq \max_j \|\dot{a}(\tilde{h}_j)M^j(x)e^{M^j(x)(a-a_j)^2}\|_{\infty,P_X}$ to conclude that

$$\chi^2(p, f_{m,k}) \leq M_\infty \int_\Omega \sum_{j=1}^{m} g_j(x; v)(h(x) - h_k(x; \theta_j))^2 dP_X(x).$$

Lemma 5. *For any $0 < a < b \leq C/\sqrt{e}$ and a positive constant C,*

$$\int_a^b log^{1/2}\frac{C}{u}du \leq 2b\, log^{1/2}\frac{C}{b}. \tag{B.2}$$

Proof. For any $0 < a < b \leq C$,

$$\begin{aligned}\int_a^b \log^{1/2}\frac{C}{u}du &= C\int_{\log^{1/2}(C/a)}^{\log^{1/2}(C/b)} v^2 e^{-v^2}dv \\ &\leq C\int_{\log(C/b)}^{\infty} t^{3/2-1}e^{-t}dt \\ &= C\Gamma(3/2, \log(C/b)) \\ &\leq b\left(\log^{-1/2}(C/b)/2 + \log^{1/2}(C/b)\right).\end{aligned}$$

The bound on the gamma function follows from the definition of the incomplete gamma function and the Mill's ratio:

$$\begin{aligned}\Gamma(3/2, x) &= \frac{1}{2}\Gamma(1/2, x) + x^{1/2}e^{-x} \\ &= \sqrt{\pi}\Phi(-\sqrt{2x}) + x^{1/2}e^{-x} \\ &\leq (x^{-1/2}/2 + x^{1/2})e^{-x},\end{aligned}$$

The result follows from the fact that $x > 1/2$.

The next lemma provides a bound on the bracketing number of functional classes that are Lipschitz in a parameter:

Lemma 6 (theorem 2.7.11 in van der Vaart & Wellner, 1996). *Let $\mathcal{F} = \{f_t : t \in T\}$ be a class of functions satisfying*

$$|f_s(x) - f_t(x)| \leq \|s - t\|_\infty F(x),$$

for some metric d on T, function F on the sample space, and every x. Then for any norm $\|\cdot\|$,

$$N_B(2\varepsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\varepsilon, T, \|\cdot\|_\infty), \tag{B.3}$$

where $N(\varepsilon, T, \|\cdot\|_\infty)$ *is the* ε*-covering number of T with respect to the metric* L_∞.

Appendix C: Proof of the Main Results

Proof of Theorem 1. The data-generating process of (X, Y) and the structure of the model $\mathcal{F}_{m,k}$ are enough to satisfy the measurability assumptions (i.e., it is a weighted sum of measurable functions).

The approximating density $f_m(x, y; \zeta)$ is a continuous function of the parameter vector ζ P_{XY}-almost everywhere. We verify this claim by choosing $(x, y) \in \Omega \times A$ from a set with positive probability and noting that (1) $\pi(h_k(x; \theta), y)$ is a continuous function of θ and (2) $(g_1(x; \nu), \ldots, g_m(x; \nu))$ is a vector of continuous functions of ν; both imply that $f_m(x, y; \zeta) = \sum_i g_i(x; \nu)\pi(h_k(x; \theta_i), y)$ is also a continuous function of the parameter vector $\zeta = (\nu', \theta_1', \ldots, \theta_m')'$.

The result follows from theorem 2.12 in White (1996).

Proof of Theorem 2. There are different approaches to show the consistency of the estimate We verify the conditions of theorem 3.5 in White (1996).

The first assumptions regarding the existence of the estimate are already shown to be satisfied in theorem 1. Assumption 3.2 in White (1996), regarding identifiability is satisfied by assumptions 3 and 4. It remains to satisfy assumption 3.1 in White (1996), regarding boundedness and uniform convergence of the log-likelihood function. We can show continuity of $\mathbb{E}\log f_{m,k}(X, Y; \zeta)$ by noting that we can interchange integration with limits and a first-order Taylor expansion:

$$\begin{aligned}
&\mathbb{E}\left[\log \frac{f_{m.k}(X, Y; \zeta)}{f_{m,k}(X, Y; \zeta - \varepsilon)}\right] \\
&\leq \sup_\zeta \mathbb{E}\left[\left|\varepsilon' \frac{\partial}{\partial \zeta} f_{m,k}(X, Y; \zeta)\right|\right] \\
&\leq \sup_\zeta \mathbb{E}\left[\left|\frac{\partial}{\partial \zeta} f_{m,k}(X, Y; \zeta)' \frac{\partial}{\partial \zeta} f_{m,k}(X, Y; \zeta)\right|\right]^{1/2} (\varepsilon'\varepsilon)^{1/2},
\end{aligned}$$

which is bounded by lemma 3 and by the fact that ε is arbitrary.

To show uniform convergence of the likelihood function, we satisfy the conditions of theorem 2 in Jennrich (1969). By assumption, $Z_{m,k}$ is a compact subset of $\mathbb{R}^{d_{m.k}}$. Measurability and continuity conditions are already satisfied;

thus, it remains to show that $\log f_{m,k}$ is bounded by an integrable function. We can bound the log-likelihood function by

$$\left|\log \frac{f_{m,k}(X,Y;\zeta)}{\varphi(X,Y)}\right| = \left|\log \sum_{i=1}^{m} g_i(X;\nu)(\pi(h_k(X;\theta_i),y) - c(y))\right|$$
$$\leq \sum_i g_i(X;\nu)|[a(h_k(X;\theta_i))Y + b(h_k(X;\theta_i))]|$$
$$\leq \max_{1\leq i\leq m} \|a(h_k(\cdot;\theta_i))Y + b(h_k(\cdot;\theta_i))\|_{\infty,P_X}.$$

Define the bounding function $D(X,Y) = \max_{1\leq i\leq m} \|a(h_k(\cdot;\theta_i))Y + b(h_k(\cdot;\theta_i))\|_{\infty,P_X}$. The function $|h_k(x;\theta)| \leq \sum_{i=1}^{J_k} |\theta_i| < \infty$ because $\max_i x_i = 1$ and $\sum_i |\theta_i| < \infty$; then both $a(h_k)$ and $b(h_k)$ are finite. Thus, it is straightforward to show that $\mathbb{E}D(X,Y) \leq \infty$, given that $\mathbb{E}_{Y|X}(Y) \leq \infty$, which is satisfied by assumption about $p_{y|x}$. As a conclusion, $\log f_{m,k}(X,Y;\hat{\zeta}) \to_{a.s.} \log f_{m,k}(X,Y;\zeta^*)$ as $n \to \infty$.

It follows from theorem 3.5 in White (1996) that $\hat{\zeta}_n \to \zeta^*$ P_{XY}-a.s. as $n \to \infty$.

Proof of Theorem 3. It follows from lemma 4 that it is enough to bound the upper divergence $\mathcal{D}(f_{m,k}, p)$ defined as

$$\mathcal{D}(f_{m,k}, p) = \int_\Omega \sum_{j=1}^{r_m} g_j(x;\nu)\{h_k(x,\theta_j) - h(x)\}^2 dP_X(x).$$

Assumption 5 ensures the existence of a ν_{c_1} such that $\max_j \|g_j(\cdot;\nu_{c_1}) - I_{Q_j^m}(\cdot)\|_{1,P_X} \leq c_1/r_m \|dP_X/d\lambda\|_{\infty,P_X}$, where $\|dP_X/d\lambda\|_{\infty,P_X}$ is finite because P_X has continuous density function with respect to the finite measure λ on Ω. Consider

$$\mathcal{D}(f_{m,k}, p) \leq \underbrace{\int_\Omega \left|\sum_{j=1}^{r_m} \{g_j(x;\nu_\varepsilon) - I_{Q_j^m}(x)\}\{h_k(x;\theta_j) - h(x)\}^2\right| dP_X(x)}_{(A_1)}$$
$$+ \underbrace{\int_\Omega \left|\sum_{j=1}^{r_m} I_{Q_j^m}(x)\{h_k(x;\theta_j) - h(x)\}^2\right| dP_X(x)}_{(A_2)}. \tag{C.1}$$

Now we just have to find bounds for both terms in the right-hand side of equation C.1 (A_1 and A_2). The second term can be written as

$$(A_2) = \int_\Omega \sum_{j=1}^{r_m} I_{Q_j^m}(x)\{h_k(x;\theta_j) - h(x)\}^2 dP_X(x)$$

$$= \int_\Omega \left\{ \sum_{j=1}^{r_m} I_{Q_j^m}(x)[h_k(x;\theta_j) - h(x)] \right\}^2 dP_X(x),$$

where the equality follows from the fact that $I_{Q_j^m} I_{Q_i^m} = I_{Q_j^m} I_{i=j}$ and $\sum_j I_{Q_j^m}(\cdot) = 1$.

If $k < \alpha$, one can choose θ_j such that

$$\sup_{x \in Q_j^m} |h_k(x,\theta_j) - h(x)| \le K_0 \sum_{|\boldsymbol{k}|=k+1} \frac{1}{\boldsymbol{k}!} \operatorname{diam}\left(Q_j^m\right)^{k+1}$$

$$= \frac{K_0 \sup_j \operatorname{diam}\left(Q_j^m\right)^{k+1}}{(k+1)!} \sum_{|\boldsymbol{k}|=k+1} \frac{(k+1)!}{k_1! \ldots k_s!}$$

$$= K_0 \sup_j \operatorname{diam}(Q_j^m)^{k+1} \frac{s^{k+1}}{(k+1)!},$$

where $\boldsymbol{k} = (k_1, \ldots, k_s)$ is a vector of positive integers satisfying $|\boldsymbol{k}| = k+1$, $\boldsymbol{k}! = k_1!, \ldots, k_s!$ and $\operatorname{diam}(Q_j^m) = \sup_{x_1, x_2 \in Q_j^m} \|x_1 - x_2\|_\infty$. This claim follows from a Taylor expansion of $h(x)$ around fixed points $x_j \in Q_j^m$ and the fact that $h \in \mathcal{W}^\infty_{\alpha, K_0}$. Similarly, if $k \ge \alpha$, we can use the expansion only up to α terms. By assumption 5, $\sup_j \operatorname{diam}(Q_j^m) \le c_0 / r_m^{1/s}$. Then

$$\sup_j \sup_{x \in Q_j^m} |h_k(x;\theta_j) - h(x)| \le \frac{c_2^{k^*}}{r_m^{k^*/s}} \frac{s^{k^*}}{k^*!}, \tag{C.2}$$

where $c_2 = c_0 K_0^{1/s}$, and $k^* = \alpha \wedge (k+1)$.

Therefore, $(A_2) \le c_2^{2k^*} / r_m^{2k^*/s} (k^*!)^2$. Note that

$$(A_1) \le \sum_{j=1}^m \|\{g_j(\cdot;\nu_\varepsilon) - I_{Q_j^m}(\cdot)\} \cdot \{h_k(\cdot;\theta_j) - h(\cdot)\}^2\|_{1,P_X}$$

$$\le \sup_j \sup_{x \in Q_j^m} |h_k(x;\theta_j) - h(x)|^2 \sum_{j=1}^{r_m} \|g_j(\cdot;\nu_\varepsilon) - I_{Q_j^m}(\cdot)\|_{1,P_X}$$

$$\le c_1 \left(\frac{(c_2 s)^{k^*}}{k^* \, r_m^{k^*/s}} \right)^2,$$

where the last inequality is due to equation C.2 and assumption 5.

Combining the results for (A_1) and (A_2),

$$\mathcal{D}(p, f_{m,k}) \leq \left(\frac{c_2^{k^*}}{r_m^{k^*/s}} \frac{s^{k^*}}{k^*!} \right)^2 (c_1 + 1). \tag{C.3}$$

It follows from lemma 4 that

$$\chi^2(p, f_{m,k}) \leq \left(\frac{c_2^{k^*}}{r_m^{k^*/s}} \frac{s^{k^*}}{k^*!} \right)^2 M_\infty (c_1 + 1). \tag{C.4}$$

We choose $m^{1/s} \geq r_m^{1/s} = \lfloor m^{1/s} \rfloor \geq m^{1/s}/2$. By assumption 1(ii), $\mathcal{F}_{r_m,k} \subseteq \mathcal{F}_{m,k}$. Hence,

$$\begin{aligned} \inf_{f_{m,k} \in \mathcal{F}_{m,k}} \chi^2(p, f_{m,k}) &\leq \inf_{f_{m,k} \in \mathcal{F}_{r_m,k}} \chi^2(p, f_{m,k}) \\ &\leq c_3 \left(\frac{c_2^{k^*}}{r_m^{k^*/s}} \frac{s^{k^*}}{k^*!} \right)^2 \\ &\leq c_3 \left(\frac{(2c_2 s)^{k^*}}{m^{k^*/s}\, k^*!} \right)^2, \end{aligned}$$

where $c_3 = M_\infty(c_1 + 1)$ does not depend on f. Therefore,

$$\sup_{p \in \mathcal{P}(\mathcal{W}^\infty_{\alpha, K_0})} \inf_{f_{m,k} \in \mathcal{F}^*_{m,k}} \chi^2(p, f_{m,k}) \leq c_3 \left(\frac{(2c_2 s)^{k^*}}{m^{k^*/s}\, k^*!} \right)^2,$$

proving the first result. The second result follows from lemma 1.

Proof of Theorem 4. We use theorem 5 in appendix A setting $f^* = f^*_{m,k}$. By lemma 2, we can choose $\Psi(\delta) = d_{m,k}^{1/2} \delta \log^{1/2}(\frac{d^a_{m,k}}{\delta})$. This choice of function that satisfies $\Psi(\delta)/\delta^2$ is nonincreasing, and we can take $\delta_n = \sqrt{d_{m,k} \log(d_{m,k} n)/n}$. To appreciate that this choice of δ_n is valid, note that for all n sufficiently large and some positive constant c,

$$\begin{aligned} \frac{\Psi(\delta_n)}{\sqrt{n}\delta_n} &= \sqrt{\frac{d_{m,k}}{n} \log \frac{d^a_{m,k}}{\delta_n}} \\ &= \sqrt{\frac{d_{m,k}}{n} \log \frac{d^{a-1/2}_{m,k} n^{1/2}}{\log^{1/2}(d_{m,k} n)}} \end{aligned}$$

$$\leq c \sqrt{\frac{d_{m,k}}{n} \log(d_{m,k} n)}$$
$$= c\, \delta_n.$$

Then, $\sqrt{n}\delta_n^2 \geq c\, \Psi(\delta_n)$.

We use theorems 5 and 3 to arrive at our result, equation 3.5:

$$d_h^2(p_{x,y}, \hat{f}_{m,k}) = O_p\left(\frac{(cs)^{2k^*}}{(k^*!)^2 (m)^{2k^*/s}} + \frac{d_{m,k}}{n} \log(d_{m,k} n) \right). \tag{C.4}$$

Proof of Proposition 1. (Ia) Write $U = (cs/(k^* m^{1/s}))^{2k^*} + d \log(dn)/n$ where $d = m(k+1)^s$ and $k^* = (k+1) \wedge \alpha$, $s = \dim(x) \geq 1$. Note that $k^* \geq 1$ since α is a positive integer for the Sobolev space introduced in section 2.1. Therefore, the first term of U converges to 0 as $m \to \infty$. For the second term, apply the condition $d = o(n/\log n)$, and we have $d \log(dn)/n = o(n/\log n) \log(o(n/\log n)n))/n = o(1)$. This shows $U \to 0$.

(Ib) In our notation, $k_1 = k+1$. When $(k_1 =) k+1 \geq \alpha, k^* = (k+1) \wedge \alpha = \alpha$. Then $U = (cs/(\alpha m^{1/s}))^{2\alpha} + m(k+1)^s \log(m(k+1)^s n)/n$. We plug in the choice $k+1 = O((\log n)^c)$ for some positive constant c, and the choice m being of order $n^{s/(s+2\alpha)} (\log n)^{c'}$ for some constant power c'; then we have that both terms in U are at most of order $O(n^{-2\alpha/(s+2\alpha)} (\log n)^c)$ for some positive power c.

(IIa) When $\alpha = \infty$ (or at least $k+1$, where $k \geq 0$ is the degree of the polynomial model), we have $k^* = (k+1) \wedge \alpha = k+1$. Then we can write $U = (cs/(m^{1/s}(k+1)))^{2(k+1)} + d \log(dn)/n = (cs/d^{1/s})^{2(k+1)} + d \log(dn)/n$. The first term converges to 0 as $d \to \infty$. The second term converges to 0 due to $d = o(n/\log n)$ (the same as in the proof of Ia).

(IIb) Consider the expression in the proof of IIa: $U = (cs/(m^{1/s}(k+1)))^{2(k+1)} + d \log(dn)/n$, where $d = m(k+1)^s$. The second term in U is at most $O(n^{-1}(\log n)^c)$ for some $c > 0$, when m and $k+1$ are both at most some powers of $\log n$ in order. When $m \geq 1$ and $(k+1) \geq c \log n$ for some positive constant c, the first term in U is at most $(cs/(k+1))^{2(k+1)} \leq (c_1(\log n))^{-c_2 \log n} = n^{-c_2 \log(c_1 \log n)}$ for large n, for some positive constants c_1 and c_2, which is negligible for large n compared to the order $O(n^{-1}(\log n)^c)$ of the second term of U.

Acknowledgments

We are grateful to the referees for their useful comments that have substantially improved the overall presentation of our letter. Also, we thank Martin Tanner, Thomas Severini, Robert Kohn, Marcelo Fernandes, and Marcelo

Medeiros for insightful discussions about mixture of experts and/or comments on previous versions of this letter.

References

Barron, A., & Sheu, C. (1991). Approximation of density functions by sequences of exponential families. *Annals of Statistics*, *19*(3), 1347–1369.

Bates, C., & White, H. (1985). A unified theory of consistent estimation for parametric models. *Econometric Theory*, *1*(2), 151–178.

Carvalho, A., & Tanner, M. (2005a). Modeling nonlinear time series with local mixtures of generalized linear models. *Canadian Journal of Statistics*, *33*(1), 97–113.

Carvalho, A., & Tanner, M. (2005b). Mixtures-of-experts of autoregressive time series: Asymptotic normality and model specification. *IEEE Transactions on Neural Networks*, *16*(1), 39–56.

Carvalho, A., & Tanner, M. (2006). Modeling nonlinearities with mixtures-of-experts of time series models. *International Journal of Mathematics and Mathematical Sciences*, *9*, 1–22.

Carvalho, A., & Tanner, M. (2007). Modelling nonlinear count time series with local mixtures of Poisson autoregressions. *Computational Statistics and Data Analysis*, *51*(11), 5266–5294.

Celeux, G., Hurn, M., & Robert, C. (2000). Computation and inferential difficulties with mixture distributions. *Journal of the American Statistical Association*, *99*, 957–970.

Ge, Y., & Jiang, W. (2006). On consistency of Bayesian inference with mixtures of logistic regression. *Neural Computation*, *18*(1), 224–243.

Genovese, C., & Wasserman, L. (2000). Rates of convergence for the gaussian mixture sieve. *Annals of Statistics*, *128*, 1105–1127.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics and Data Analysis*, *51*, 3529–3550.

Geweke, J., & Keane, M. (2007). Smoothly mixing regressions. *Journal of Econometrics*, *138*(1), 252–290.

Huerta, G., Jiang, W., & Tanner, M. (2003). Time series modeling via hierarchical mixtures. *Statistica Sinica*, *13*(4), 1097–1118.

Jacobs, R., Jordan, M., Nowlan, S., & Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*(1), 79–87.

Jennrich, R. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, *40*(2), 633–643.

Jiang, W., & Tanner, M. (1999a). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Annals of Statistics*, *27*, 987–1011.

Jiang, W., & Tanner, M. (1999b). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural Computation*, *11*(5), 1183–1198.

Jiang, W., & Tanner, M. (1999c). On the identifiability of mixtures-of-experts. *Neural Networks*, *12*(9), 1253–1258.

Jordan, M., & Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*(2), 181–214.

Mendes, E., Veiga, A., & Medeiros, M. (2006). *Estimation and asymptotic theory for a new class of mixture models*. Unpublished manuscript, Pontifical Catholic University of Rio de Janeiro.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *Annals of Statistics*, *38*(3), 1733–1766.

Peng, F., Jacobs, R., & Tanner, M. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, *91*, 953–960.

Stone, C. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, *8*(6), 1348–1360.

van der Geer, S. (2000). *Empirical processes in M-estimation*. Cambridge, UK: Cambridge University Press.

van der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes*. New York: Springer-Verlag.

Viele, K., & Tong, B. (2002). Modeling with mixture of linear regressions. *Statistics and Computing*, *12*, 315–330.

Villani, M., Kohn, R., & Giordani, P. (2009). Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics*, *153*(2), 155–173.

White, H. (1996). *Estimation, inference and specification analysis*. Cambridge, UK: Cambridge University Press.

Windlund, O. (1977). On best error bounds for approximation by piecewise polynomial functions. *Numerische Mathematik*, *27*, 327–338.

Wong, W., & Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieves MLEs. *Annals of Statistics*, *23*(2), 339–362.

Wood, S., Jiang, W., & Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika*, *89*(3), 513–528.

Wood, S., Kohn, R., Cottet, R., Jiang, W., & Tanner, M. (2008). Locally adaptive nonparametric binary regression. *Journal of Computational and Graphical Statistics*, *17*(2), 352–372.

Wood, S., Rosen, O., & Kohn, R. (2011). Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics*, *20*(1), 174–195.

Young, D., & Hunter, D. (2010). Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics and Data Analysis*, *54*(10), 2253–2266.

Zeevi, A., Meir, R., & Maiorov, V. (1998). Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Transactions on Information Theory*, *44*(3), 1010–1025.

Received November 1, 2011; accepted May 18, 2012.