5
Modeling Cycle

Selecting the number of models, as well as the tree structure, is a difficult problem. In fact, there are two approaches for the model selection problem: hypothesis testing and the use of a model information criterion. As shown in (51), the log-likelihood ratio tests are not applicable in this case because, under the null hypothesis of fewer basis distributions, the model is non-identified and the test does not have the standard chi-square asymptotic distribution.

There is also an alternative hypothesis testing methodology following (41), which is used by (47), (45) and (46). This methodology is based on a sequence of Lagrange multiplier tests applied to a linearized version of the model. However, adapting this approach to mixture of models is not trivial.

We will introduce a specification algorithm based on two information criteria (IC): Bayesian Information Criterion (BIC) (53) and Akaike Information Criterion (AIC) (1). Both criteria have been used to select the number of experts; see (8), (54), (4), (68), and Wong and Li (1999,2000). There is no consensus about which is the better criterion. The two criteria are defined as

$$BIC = -2\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_t;\widehat{\theta}) + M\log T \tag{5-1}$$

$$AIC = -2\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_t;\widehat{\theta}) + 2M, \tag{5-2}$$

where $M = 3\#\mathbb{J} + (p+2)\#\mathbb{T}$ is the number of estimated parameters.

It is known that, for well behaved models, BIC is consistent for model selection. Furthermore, when the sample size goes to the infinity, the true model will be selected because it has the smallest BIC with probability tending to one. However, when the model is overidentified, the usual regularity conditions to support this result fail, but (68) present some evidence that, even when we have overidentified models, the BIC may still be consistent for model selection.

For instance, let $\mathbb{J}$ and $\mathbb{T}$ define the tree $\mathbb{J}\mathbb{T}$ with $\#\mathbb{T}$ local models and let $k \in \mathbb{T}$ be a node to be split. Then, when we split the terminal node k we have the

new tree $\mathbb{JT}^{(k)}$ defined by $\mathbb{J}^{(k)}$ and $\mathbb{T}^{(k)}$. The expressions for these sets are

$$\mathbb{J}^{(k)} = \mathbb{J} \cup \{k\} \tag{5-3}$$

$$\mathbb{T}^{(k)} = \{2k+1, 2k+2\} \cup (\mathbb{T} \setminus \{k\}), \tag{5-4}$$

where $\mathbb{T} \setminus \{k\}$ means the complement of $\{k\}$ in $\mathbb{T}$. Furthermore, the new parameter vector $\boldsymbol{\theta}^{(k)}$ is defined as

$$\boldsymbol{\theta}^{(k)} = \left[\boldsymbol{\nu}'_{j_1}, \ldots, \boldsymbol{\nu}'_{j_{\#\mathbb{J}^{(k)}}}, \psi'_{t_1}, \ldots, \psi'_{t_{\#\mathbb{T}^{(k)}}}\right]', \tag{5-5}$$

where $j_i \in \mathbb{J}^{(k)}$ and $t_i \in \mathbb{T}^{(k)}$.

The growing algorithm for the first split is the following:

1. Assume that the number of covariates is p and estimate a linear model with all p regressors in $\mathbf{x}_t$ and compute the value of the IC;
2. for each covariate $x_{s_0} \in \mathbf{x}_t$ with $s_0 = 1, \ldots, p$, estimate the model $\mathbb{JT}^{(0)}$, where each terminal node is a linear model with all p regressors, and compute the IC; and
3. select the model with the smallest IC.

The growing algorithm for the k-th split is:

1. For each $k \in \mathbb{T}$ and for each covariate $x_{s_i} \in \mathbf{x}_t$ with $s_i = 1, \ldots, p$:
   (a) Split the node k following (5-3) and (5-4);
   (b) estimate the new parameter vector $\boldsymbol{\theta}^{(k)}$ defined in (5-5); and
   (c) compute the IC.
2. select the tree $\mathbb{JT}^{(k)}$ with the smallest IC;
3. if the smallest IC for the tree $\mathbb{JT}^{(k)}$ is greater than the IC of the tree $\mathbb{JT}$, then we stop growing the tree. Case contrary, repeat the steps above setting $\mathbb{JT} = \mathbb{JT}^{(k)}$.