Referências Bibliográficas

[1] AKAIKE, H.. **A new look at the statistical model identification**. IEEE Transactions on Automatic Control, 19(6):716–723, 1974. 5

[2] ANDERSEN, T.; BOLLERSLEV, T.; CHRISTOFFERSEN, P. ; DIEBOLD, F.. **Practical volatility and correlation modeling for financial market risk management**. In: Carey, M.; Stulz, R., editors, RISKS OF FINANCIAL INSTITUTIONS. University of Chicago Press for NBER, Chicago, IL, 2006. 1

[3] ANDERSEN, T.; BOLLERSLEV, T. ; DIEBOLD, F.. **Parametric and nonparametric measurement of volatility**. In: Elliott, G.; Granger, C. ; Timmermann, A., editors, HANDBOOK OF FINANCIAL ECONOMETRICS. North-Holland, Amsterdam, 2006. 1

[4] AUDRINO, F.; BÜHLMANN, P.. **Tree-structured generalized autoregressive conditional heterocedastic models**. Journal of the Royal Statistal Society, Series B, 63(4):727–744, 2001. 5

[5] BERNDT, E.; HALL, B.; HALL, R. ; HAUSMAN, J.. **Estimation and inference in nonlinear structural models**. Annal of Economic Social Measurements, 3(4):653–665, 1974. A

[6] BOLLERSLEV, T.. **Generalized autoregressive conditional heteroskedasticity**. Journal of Econometrics, 21:307–328, 1986.

[7] CARVALHO, A.; SKOULAKIS, G.. **Ergodicity and existence of moments for local mixture of linear autoregressions**. Technical report, Northwestern University, 2004. 2

[8] CARVALHO, A.; TANNER, M.. **Mixture-of-experts of autoregressive time series: asymptotic normality and model specification**. IEEE Transactions on Neural Networks, Forthcoming, 2005a. 5

[9] CARVALHO, A.; TANNER, M.. **Modeling nonlinear time series with mixture-of-experts of generalized linear models**. The Canadian journal of Statistics, Forthcoming, 2005b. 2, 4

[10] CARVALHO, A. X.; TANNER, M. A.. **Mixtures-of-experts of generalized time series: Asymptotic normality and model specification**. Technical report, University of British Columbia and Northwestern University, 2002.

[11] CARVALHO, A. X.; TANNER, M. A.. **Mixtures-of-experts of generalized time series: Consistency of the maximum likelihood estimator**. Technical report, University of British Columbia and Northwestern University, 2002.

[12] CHAN, K.; TONG, H.. **A note on testing for multi-modality with dependent data**, 1998. 7.1

[13] CHAN, K. S.; TONG, H.. **On estimating thresholds in autoregressive models**. Journal of Time Series Analysis, 7:179–190, 1986. 1, 1

[14] CHEN, X.; SHEN, X.. **Sieve Extremum Estimates for Weakly Dependent Data**. Econometrica, 66:289–314, 1998. 1

[15] CHEN, X.; WHITE, H.. **Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators**. IEEE Transactions on Information Theory, 18:17–39, 1998. 1

[16] CHRISTOFFERSEN, P. F.. **Evaluating interval forecasts**. International Economic Review, 39:841–862, 1998. 6.3, 6.3, 6.7

[17] CYBENKO, G.. **Approximation by superposition of sigmoidal functions**. Mathematics of Control, Signals, and Systems, 2:303–314, 1989. 1

[18] DA ROSA, J. C.; VEIGA, A. ; MEDEIROS, M. C.. **Tree-structured smooth transition regression models based on CART algorithm**. Textos para Discussão 469, Pontifical Catholic University of Rio de Janeiro, 2003. 3

[19] DEMPSTER, A.; LAIRD, N. ; RUBIN, D.. **Maximum likelihood estimation from incomplete data via em algorithm (with discussion)**. J. R. Statist. Soc. B, 39:1–38, 1977. 4, A

[20] ENGLE, R. F.. **Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation**. Econometrica, 50:987–1007, 1982.

[21] FAN, J.; YAO, Q.. **Nonlinear Time Series: Nonparametric and Parametric Methods**. Springer-Verlag, New York, NY, 2003. 1

[22] FUNAHASHI, K.. **On the approximate realization of continuous mappings by neural networks**. Neural Networks, 2:183–192, 1989. 1

[23] GALLANT, A. R.; WHITE, H.. **On learning the derivatives of an unknown mapping with multilayer feedforward networks**. Neural Networks, 5:129–138, 1992. 1

[24] GRANGER, C. W. J.; TERÄSVIRTA, T.. **Modelling Nonlinear Economic Relationships**. Oxford University Press, Oxford, 1993. 1

[25] HASTIE, T.; TIBSHIRANI, R. ; FRIEDMAN, J.. **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. Springer, 2001. 1

[26] HEILER, S.. **A survey on nonparametric time series analysis**. Economics Working Papers at WUSTL 9904005, Washington University, 1999. 1

[27] HORNIK, K.; STINCHOMBE, M. ; WHITE, H.. **Multi-layer Feedforward networks are universal approximators**. Neural Networks, 2:359–366, 1989.

[28] HORNIK, K.; STINCHOMBE, M. ; WHITE, H.. **Universal approximation of an unknown mapping and its derivatives using multi-layer feedforward networks**. Neural Networks, 3:551–560, 1990.

[29] HÄRDLE, W.. **Applied Nonparametric Regression**. Cambridge University Press, Cambridge, 1990. 1

[30] HÄRDLE, W.; LÜTKEPOHL, H. ; CHEN, R.. **A review of nonparametric time series analysis**. International Statistical Review, 65:49–72, 1997. 1

[31] HUERTA, G.; JIANG, W. ; TANNER, M.. **Mixture of time series models**. Journal of Computational and Graphical Statistics, 10(1):82–89, 2001. 4

[32] HUERTA, G.; JIANG, W. ; TANNER, M.. **Mixtures of time series models**. Journal of Computational and Graphical Statistics, 10:82–89, 2001.

[33] HUERTA, G.; JIANG, W. ; TANNER, M.. **Time series modeling via hierachical mixtures**. Statistica Sinica, 13:1097–1118, 2003.

[34] JACOBS, R. A.; JORDAN, M. I.; NOWLAN, S. J. ; HINTON, G. E.. **Adaptive mixtures of local experts**. Neural Computation, 3:79–87, 1991. 1, 2

[35] JIANG, W.; TANNER, M.. **On the identifiability of mixtures-of-experts**. Neural Networks, 12(9):1253–1258, 1999c. 4.1

[36] JORDAN, M. I.; JACOBS, R. A.. **Hierarchical mixtures of experts and the EM algorithm**. Neural Computation, 6:181–214, 1994. 1, 4, A

[37] KUAN, C. M.; WHITE, H.. **Artificial neural networks: An econometric perspective**. Econometric Reviews, 13:1–91, 1994. 1

[38] LE, N.; MARTIN, R. ; RAFTERY, A.. **Modeling flat streches, bursts, and outliers in time series using mixture transition distribution models**. Journal of the American Statistical Association, 91(436):1504–1515, 1996. 4

[39] LUMSDAINE, R.. **Consistency and asymptotic notmality of the quasi-maximum likelihood estimator in igarch(1,1) and covariance stationary garch(1,1) models**. Econometrica, 64(3):575–596, 1996. D.1

[40] LUUKKONEN, R.; SAIKKONEN, P. ; TERÄSVIRTA, T.. **Testing linearity against smooth transition autoregressive models**. Biometrika, 75:491–499, 1988. 1

[41] LUUKKONEN, R.; SAIKKONEN, P. ; TERÄSVIRTA, T.. **Testing linearity against smooth transitions autoregressive models**. Biometrika, 75:491–499, 1988. 5

[42] MACKAY, D. J. C.. **Bayesian interpolation**. Neural Computation, 4:415–447, 1992.

[43] MACKAY, D. J. C.. **A practical Bayesian framework for backpropagation networks**. Neural Computation, 4:448–472, 1992.

[44] MCALEER, M.. **Automated inference and learning in modeling financial volatility**. Econometric Theory, 21:232–261, 2005. 1

[45] MEDEIROS, M.; DA ROSA, J. ; VEIGA, A.. **A tree-structured approach to cross-section and semiparametric regression**. Technical report, PUC-Rio, 2005. 5

[46] MEDEIROS, M.; TERÄSVIRTA, T. ; RECH, G.. **Building neural network models for time series: A statistical approach**. Journal of Forecasting, 25:49–75, 2006. 1, 5

[47] MEDEIROS, M.; VEIGA, A.. **Flexible coeficient smooth transition time series model**. IEEE Transactions on Neural Networks, 16(1):97–113, 2005. 5, 6.1, 7.1

[48] MORAN, P.. **The statistical analysis of the canadian lynx cycle: Structure and prediction**. Aust. J. Zool, 1:163–173, 1953. 7.1

[49] NOWLAN, S. J.. **Maximum likelihood competitive learning**. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, volumen 2, p. 574–582. Morgan Kaufmann, 1990. 1

[50] POON, S.; GRANGER, C.. **Forecasting volatility in financial markets**. Journal of Economic Literature, 41:478–539, 2003. 1

[51] QUINN, B.; MCLACHLAN, G. ; HJORT, L.. **A note on the aitkin-rubin approach to hypothesis testing in mixture models**. Journal of the Royal Statistal Society, Series B, 49(3):311–314, 1987. 5

[52] RECH, G.; TERÄSVIRTA, T. ; TSCHERNIG, R.. **A simple variable selection technique for nonlinear models**. Working Paper Series in Economics and Finance 296, Stockholm School of Economics, 1999. 7.2

[53] SCHWARZ, G.. **Estimating the dimension of a model**. Annals of Statistics, 6:461–464, 1978. 5

[54] SUAREZ-FARIÑAS, M.; PEDREIRA, C. E. ; MEDEIROS, M. C.. **Local-Global Neural Network: A New Approach For Nonlinear Time Series Modeling**. Journal of the American Statistical Association, 99:1092–1107, 2004. 5

[55] TAYLOR, S.. **Modelling Financial Time Series**. Wiley, Chichester, 1986. 1

[56] TERÄSVIRTA, T.. **Specification, estimation, and evaluation of smooth transition autoregressive models**. Journal of the American Statistical Association, 89:208–218, 1994. 1

[57] TONG, H.. **On a threshold model**. In: Chen, C. H., editor, PATTERN RECOGNITION AND SIGNAL PROCESSING, Amsterdam, 1978. Sijthoff and Noordhoff. 1

[58] TONG, H.. **Non-linear Time Series: A Dynamical Systems Approach**, volumen 6 de **Oxford Statistical Science Series**. Oxford University Press, Oxford, 1990. 1, 7.1

[59] TONG, H.; LIM, K.. **Threshold autoregression, limit cycles and cyclical data (with discussion)**. Journal of the Royal Statistical Society, Series B, 42:245–292, 1980. 1

[60] TRAPLETTI, A.; LEISCH, F. ; HORNIK, K.. **Stationary and integrated autoregressive neural network processes**. Neural Computation, 12:2427–2450, 2000. 1

[61] VAN DIJK, D.; TERÄSVIRTA, T. ; FRANSES, P. H.. **Smooth transition autoregressive models - a survey of recent developments**. Econometric Reviews, 21:1–47, 2002. 1

[62] WEIGEND, A. S.; MANGEAS, M. ; SRIVASTAVA, A. N.. **Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting**. International Journal of Neural Systems, 6:373–399, 1995. 1

[63] WHITE, H.. **Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings**. Neural Networks, 3:535–550, 1990. 1

[64] WHITE, H.. **Estimation, Inference and Specification Analysis**. Cambridge University Press, New York, NY, 1992. D, D.1, D.2, D.3

[65] WONG, C. S.; LI, W. K.. **On a generalized mixture autoregressive model**. Research Report 242, Department of Statistics and Actuarial Science, University of Hong Kong, 1999. 2, 7.1

[66] WONG, C. S.; LI, W. K.. **On a mixture autoregressive model**. Journal of the Royal Statistical Society, Series B, 62:91–115, 2000. 1, 2, 7.1

[67] WONG, C. S.; LI, W. K.. **On a mixture autoregressive conditional heterocedastic model**. Journal of the American Statistical Association, 96:982–995, 2001. 1

[68] WOOD, S.; JIANG, W. ; TANNER, M.. **Bayesian mixture of splines for spatially adaptative nonparametric regression**. Biometrika, 89(3):513–528, 2001. 5, 5

[69] ZEEVI, A.; MEIR, R. ; ADLER, R.. **Non-linear models for time series using mixtures of autoregressive models**. Technical report, Technion, 1998. 2, C, C

A
EM Algorithm

The estimation of the parameter vector $\boldsymbol{\theta}$ can be performed by maximizing the quasi-likelihood of the density function in (3-1). We can write the conditional likelihood based on a sample $\{y_t\}_{t=1}^{T}$. The log-likelihood to be maximized over $\boldsymbol{\Theta}$, the parametric space, is given by

$$\log \mathcal{L}_T(\theta) = \sum_{t=1}^{T} \log \left[\sum_{i \in \mathbb{T}} B_i(\mathbf{x}_t; \theta_i) \pi(y_t | \mathbf{x}_t; \boldsymbol{\psi}_i) \right].$$

Numerical optimization can carried out by using the EM algorithm of (19). The idea behind EM algorithm is to maximize a sequence of simple functions which leads to the same solution as maximizing a complex function.

We define the probability that the local model $i \in \mathbb{T}$ has generated the output y_t as

$$p_i(y_t) = \frac{B_i(\mathbf{x}_t; \boldsymbol{\theta}_i)\, \pi(y_t | \mathbf{x}_t; \boldsymbol{\psi}_i)}{\sum_{j \in \mathbb{T}} B_j(\mathbf{x}_t; \boldsymbol{\theta}_j)\, \pi(y_t | \mathbf{x}_t; \boldsymbol{\psi}_j)}. \tag{A-1}$$

It is important to note that a parent node $i \in \mathbb{J}$ has the probability $p_{t,i} = p_{t,2i+1} + p_{t,2i+2}$.

We define our estimation algorithm following (36). The EM algorithm demands the definition of a complete dataset $\mathbb{X}_t$ and an incomplete dataset $\mathbb{Y}_t$. We introduce an indicator variable $z_t = \{z_{it}\}, i \in \mathbb{T}$, such that only one of the z_{it} is equal to one each time to simplify the likelihood function. If the variables z_{it} are known, the maximization problem is divided into a set of regression problems for each model and a classification problem for the functions $B_i(\cdot)$.

Since the indicator variables are not known we shall define a probabilistic model which links these variables with the observed data (i.e. the complete dataset probabilistic model). This model can be written in terms of z_{it} as follows:

$$\begin{aligned} \mathbb{P}(y_t, z_{it} | \mathbf{x}_t; \theta) &= B_i(\mathbf{x}_t; \theta_i) \pi(y_t | \mathbf{x}_t; \boldsymbol{\psi}_i) \\ &= \prod_{i \in \mathbb{T}} [B_i(\mathbf{x}_t; \theta_i) \pi(y_t | \mathbf{x}_t; \boldsymbol{\psi}_i)]^{z_{it}}. \end{aligned} \tag{A-2}$$

The log-likelihood $l_c(\theta) = \log \mathcal{L}_T(\theta)$ of the complete dataset (i.e. considering the equation (A-2)) based on a sample $\{y_t\}_{t=1}^{T}$ is that such

$$l_c(\theta) = \sum_{t=1}^{T} \sum_{i \in \mathbb{T}} z_{it} \ [\log \ B_i(\cdot) + \log \ \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i)] \tag{A-3}$$

Calculating the expected value of the log-likelihood of the complete dataset (E-Step), it follows that:

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) &= \mathbb{E}\left[\mathbb{E}\left[l_c(\theta)|\mathbb{Y}\right]|\mathbb{X}\right] \\ &= \sum_{t=1}^{T} \sum_{i \in \mathbb{T}} \mathbb{E}\left[z_{it} \left(\log \ B_i(\cdot) + \log \ \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i)\right)\right] \\ &= \sum_{t=1}^{T} \sum_{i \in \mathbb{T}} \mathbb{E}[z_{it}] \left(\log \ B_i(\cdot) + \log \ \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i)\right) \\ &= \sum_{t=1}^{T} \sum_{i \in \mathbb{T}} p_{t,i} \left[\log \ B_i(\cdot) + \log \ \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i)\right]. \end{aligned} \tag{A-4}$$

The expected value of z_{it} is the following

$$\begin{aligned} \mathbb{E}[z_{it}|\mathbb{Y}] &= \mathbb{P}(z_{it} = 1|y_t, \mathbf{x}_t, \boldsymbol{\theta}) \\ &= \frac{\mathbb{P}(z_{it} = 1|\mathbf{x}_t, \boldsymbol{\theta}) \mathbb{P}(y_t|z_{it} = 1, \mathbf{x}_t, \boldsymbol{\theta})}{\mathbb{P}(y_t|\mathbf{x}_t, \boldsymbol{\theta})} \\ &= \frac{B_i(\cdot)\pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i)}{\sum_{j \in \mathbb{T}} B_j(\cdot)\pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_j)} \\ &= p_i(y_t). \end{aligned} \tag{A-5}$$

Maximizing (A-4) with respect to the parameter vector $\boldsymbol{\theta}$, we find

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta} \sum_{t=1}^{T} \sum_{i \in \mathbb{T}} h_{t,i}^{(k)} \ \left[\log \ B_i(\mathbf{x}_t; \boldsymbol{\theta}_i^{(k)}) + \log \ \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i^{(k)})\right]. \tag{A-6}$$

Analyzing the equation (A-6), we note that the parameters $\boldsymbol{\theta}_i$ influence $Q(\cdot)$ only through the term $p_{t,i} \log B_i(\cdot)$ and the parameters $\boldsymbol{\psi}_i$ through the term $p_{t,i} \log \pi(\cdot)$. Thus we can split the maximization problem as follows:

$$\boldsymbol{\psi}_i^{(k+1)} = \arg\max_{\boldsymbol{\psi}_i} \sum_{t=1}^{T} p_{t,i} \ \log \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i^{(k)}) \tag{A-7}$$

$$\boldsymbol{\theta}_i^{(k+1)} = \arg\max_{\boldsymbol{\theta_i}} \sum_{t=1}^{T} \sum_{i \in \mathbb{T}} p_{t,i} \ \log B_i(\mathbf{x}_t; \boldsymbol{\theta}_i^{(k)}) \tag{A-8}$$

The maximization problem for the expert parameter vector $\boldsymbol{\psi}_i$ is a weighted least squares problem. The maximization of (A-8) is a complex non-linear optimization problem, which can be solved using numeric algorithms.

In order to introduce a general equation to estimate the parameters $\boldsymbol{\beta}_i$, we

must define the following notation. Let $x_{t,j}$ be the j-th element in $\tilde{\mathbf{x}}_t$ if $j > 0$ and 1 otherwise. Furthermore, $\tilde{\mathbf{x}}_{t,\bar{j}}$ is the vector $\tilde{\mathbf{x}}_t$ without $x_{t,j}$. Following the same notation, we define $\boldsymbol{\beta}_{\bar{i}}$ as the $\boldsymbol{\beta}$ vector without the β_i element.

The general equation for estimate the linear parameters is

$$\beta_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T} p_{t,i}\, x_{t,j} \left(y_t - \boldsymbol{\beta}_{i\bar{j}}^{\prime(k)}\, \tilde{\mathbf{x}}_{t,\bar{j}} \right)}{\sum_{t=1}^{T} p_{t,i}\, x_{t,j}^2} \tag{A-9}$$

Then, we define the general equation to estimate the σ_i, $\forall i \in \mathbb{T}$:

$$\sigma_i^{2(k+1)} = \frac{\sum_{t=1}^{T} p_{t,i} \left(y_t - \boldsymbol{\beta}_i^{\prime(k)} \tilde{\mathbf{x}}_t \right)^2}{\sum_{t=1}^{T} p_{t,i}}. \tag{A-10}$$

The quasi-Newton optimization algorithm is used to estimate the parameters of the logistic function. We can write the equation (A-8) as follows:

$$\boldsymbol{\theta}_i^{(k+1)} = \arg\max_{\boldsymbol{\theta}_i} \sum_{t=1}^{T} \sum_{i \in \mathbb{T}} p_{t,i}\, \log \left(\prod_{j \in \mathbb{J}_i} g((-1)^{\chi_{\mathbb{J}_i \cup \{i\}}(2j+1)}\, \mathbf{x}_t; \boldsymbol{\theta}_j^{(k)}) \right) \tag{A-11}$$

where $\chi_{\mathbb{J}_i \cup \{i\}}(2j+1)$ is equal to 1 if $2j+1 \in \mathbb{J}_i \cup \{i\}$ and 0 otherwise. It is easy to show that this notation is equivalent to that in equation (3-2).

To estimate the parameters of the nodes $\mathbb{J}$ of the tree, we need so select a node $H \in \mathbb{J}$ and then estimate just the parameters of the logistic function g_H. We select the sequence of H in a increasing order, beginning with $H = 0$, to estimate all the logistic function parameters.

Only two sets $\mathbb{J}_i\, \forall i \in \mathbb{J} \cup \mathbb{T}$ have the node H as the last node: $\mathbb{J}_{2H+1}$ and $\mathbb{J}_{2H+1}$. As a result, we can write B_{2H+1} and B_{2H+2} as functions of the previously estimated parameters.

$$B_{2H+1} = [1 - g(\mathbf{x}_t; \boldsymbol{\nu}_H)] \prod_{j \in \mathbb{J}_H} g((-1)^{\chi_{\mathbb{J}_H \cup \{H\}}(2j+1)}\, \mathbf{x}_t; \boldsymbol{\nu}_j) \tag{A-12}$$

$$B_{2H+2} = g(\mathbf{x}_t; \boldsymbol{\nu}_H) \prod_{j \in \mathbb{J}_H} g((-1)^{\chi_{\mathbb{J}_H \cup \{H\}}(2j+1)}\, \mathbf{x}_t; \boldsymbol{\nu}_j) \tag{A-13}$$

Rewriting the equation (A-11) using this result, we have

$$\begin{aligned} \boldsymbol{\nu}_H^{(k+1)} &= \arg\max_{\boldsymbol{\nu}_H} \sum_{t=1}^{T} \sum_{i=2H+1}^{2H+2} p_{t,i}\, \log B_i(\mathbf{x}_t; \boldsymbol{\theta}_i^{(k)}) \qquad \text{(A-14)} \\ &= \arg\max_{\boldsymbol{\nu}_H} \sum_{t=1}^{T} p_{t,2H+1}\, \log B_{2H+1}(\cdot) + p_{t,2H+2}\, \log B_{2H+2}(\cdot) \\ &= \arg\max_{\boldsymbol{\nu}_H} \sum_{t=1}^{T} p_{t,2H+1} \log g_H(\cdot) + p_{t,2H+2} \log(1 - g_H(\cdot)) \end{aligned}$$

Then, the BHHH algorithm (5) is used to find the parameter estimates. The differentiation of the previous equation is then calculated.

$$\frac{\partial}{\partial \boldsymbol{\nu}_H} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \;=\; \sum_{t=1}^{T} \frac{\partial}{\partial \boldsymbol{\nu}_H} \left[\frac{-\gamma_H (x_{s_H,t} - c_H)}{\hat{\sigma}_{s_H}} \right] \left[(p_{t,2H+1} + p_{t,2H+2})\, g_H(\cdot) - p_{t,2H+1} \right].$$

We adopt the following notation: $p_t^{(k)} = [p_{t,2H+1} - g_H^{(k)}\, p_H]$ and $C_t^{(k)} = \frac{(x_{s_H,t} - c_H^{(k)})}{\sigma_{s_H}}$. So we can write the update rule as

$$\boldsymbol{\nu}_H^{(k+1)} = \boldsymbol{\nu}_H^{(k)} + \lambda_k \left\{ \sum_{t=1}^{T} {p_t^{(k)}}^2 \begin{bmatrix} {C_t^{(k)}}^2 & -\frac{C_t^{(k)} \gamma_H^{(k)}}{\hat{\sigma}_{s_H}} \\ -\frac{C_t^{(k)} \gamma_H^{(k)}}{\hat{\sigma}_{s_H}} & \frac{{\gamma_H^{(k)}}^2}{\hat{\sigma}^2_{s_H}} \end{bmatrix} \right\}^{-1} \sum_{t=1}^{T} p_t^{(k)} \begin{bmatrix} -C_t^{(k)} \\ \frac{\gamma_H^{(k)}}{\hat{\sigma}_{s_H}} \end{bmatrix}. \tag{A-15}$$

Then, the parameter estimation algorithm can be summarized as follows:

1. The probabilities $p_{t,i}^{(k)}$ are calculated trough equation (A-1).
2. The local model parameters $\boldsymbol{\psi}_i^{(k+1)}$ are then estimated using (A-9) and (A-10).
3. The gating parameters $\boldsymbol{\nu}_j^{(k+1)}$, $\forall j \in \mathbb{J}$, are calculated through (A-15).
4. These steps are performed until the square error between $\theta^{(k)}$ and $\theta^{(k+1)}$ is small enough.

B
Identifiability

To prove the identifiability of the Tree-MM models, we need to define some concepts and assumptions. First, we define de concept of a sub-tree and then state two assumptions to establish the theorem of identifiability of the Tree-MM models.

Let $\mathbb{JT}$ be a tree with sets $\mathbb{J}$, $\mathbb{T}$ and $\mathbb{S}$, where $\mathbb{S}$ is the set of indexes s_j, $\forall j \in \mathbb{J}$ and parameter vector $\boldsymbol{\theta}$. We define a subtree $\mathbb{JT}^k$ as the tree beginning at node k, with the sets $\mathbb{J}^k \subseteq \mathbb{J}$, $\mathbb{T}^k \subseteq \mathbb{T}$ and $\mathbb{S}^k \subseteq \mathbb{S}$, where $i \in \mathbb{JT}^k \Leftrightarrow k \in \mathbb{J}_i \cup \{i\}$ and parameter vector $\boldsymbol{\theta}^k$. For example, assume the tree $\mathbb{JT} = \{0, 1, 2, 3, 4, 5, 6, 11, 12\}$ then $\mathbb{JT}^2 = \{2, 5, 6, 11, 12\}$.

Assumption 7 *Let $f_k(y_t|\mathbf{x}_t; \boldsymbol{\theta}_k)$ be the conditional p.d.f. of the subtree $\mathbb{JT}^k$. Then $\forall k \in \mathbb{J}$, $f_{2k+1}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+1}) \neq f_{2k+2}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+2})$.*

This assumption guarantees that our tree is irreducible in the sense that any split cannot be changed by a subtree or by a local model.

Assumption 8 *We assume that for any tree $\mathbb{JT}$ and all sub-trees $\mathbb{JT}^k$:*

1. *The parameters $\gamma_j > 0, \forall j \in \mathbb{J}$;*
2. *$\forall j \in \mathbb{J}^{2k+1}$, if $s_j = s_k$ then $c_j < c_k$;*
3. *$\forall j \in \mathbb{J}^{2k+2}$, if $s_j = s_k$ then $c_j \geq c_k$.*

These assumptions together ensure that the sets $\mathbb{J}$, $\mathbb{T}$ and $\mathbb{S}$ uniquely specify any tree.

Lemma B.1 *Under Assumptions (7) and (8), a tree $\mathbb{JT}$ is uniquely specified and the parameter vector $\boldsymbol{\theta}$ has a unique representation.*

PROOF. We start proving irreducibility. Suppose that for any node $k \in \mathbb{J}$, $f_{2k+1}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+1}) = f_{2k+2}(y_t|\mathbf{x}_t; \boldsymbol{\theta}^{2k+2})$. So, $f_k = g_k(\cdot)f_{2k+1} + (1 - g_k(\cdot))f_{2k+2} = f_{2k+1} = f_{2k+2}$. Then we can change the node k by the node $2k + 1$ or $2k + 2$. If $f_{2k+1}(\cdot) \neq f_{2k+2}(\cdot)$, $\forall k \in \mathbb{J}$, then the tree cannot be reduced so it is irreducible.

Now, suppose there is an irreducible tree $\mathbb{JT}$. On the first split at s_0, c_0 can assume any value in $\mathbb{R}$. Now consider the sub-trees $\mathbb{JT}^1$ and $\mathbb{JT}^2$. Following the

condition (8), on the next split at $s_k = s_0, k \in \mathbb{J}^1$, c_k can assume any value in $(-\infty, c_0)$ and on the next split at $s_l = s_0, l \in \mathbb{J}^2$, c_l can assume any value in $[c_0, \infty)$. So, the values of c_k and c_l cannot be interchanged. Repeating this argument for all splits, and considering that the transition has the same shape (which is guaranteed by the constraint over the γs), we show that any irreducible tree under Assumption (8) is uniquely specified.

Q.E.D.

The next theorem gives the conditions under which the Tree-MM model is identifiable.

Teorema B.2 *Under Assumptions (7) and (8), the tree mixture-of-expert structure, as presented in (3-1), is identifiable, in the sense that, for a sample* $\{y_t; \mathbf{x}_t\}_{t=1}^T$*, and for* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$

$$\prod_{t=1}^{T} f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_1) = \prod_{t=1}^{T} f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_2) \quad , a.s.$$

is equivalent to $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

PROOF. Suppose that $f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_1) = f(y_t|\mathbf{x}_t; \boldsymbol{\theta}_2)$, for any sequence $\{y_t; \mathbf{x}_t\}_{t=1}^T$. Therefore, we have

$$\sum_{i \in \mathbb{T}_1} B_i(x_{s_i}; \boldsymbol{\theta}_{1i}) \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_{1i}) = \sum_{i \in \mathbb{T}_2} B_i(x_{s_i}; \boldsymbol{\theta}_{2i}) \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_{2i}). \tag{B-1}$$

Considering the Lemma B.1, $\mathbb{T}_1 = \mathbb{T}_2 = \mathbb{T}$; furthermore, if $\boldsymbol{\psi}_{1i} = \boldsymbol{\psi}_{2i}$ then $\pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_{1i}) = \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_{2i})$. Then we can write this equation as

$$\sum_{i \in \mathbb{T}} (B_i(\cdot; \boldsymbol{\theta}_{1i}) - B_i(\cdot; \boldsymbol{\theta}_{2i})) \pi(y_t|\mathbf{x}_t; \boldsymbol{\psi}_i) = 0, \tag{B-2}$$

where $\boldsymbol{\psi}_i = \boldsymbol{\psi}_{1i} = \boldsymbol{\psi}_{2i}$.

We have to show that $B_i(\cdot; \boldsymbol{\theta}_{1i}) - B_i(\cdot; \boldsymbol{\theta}_{2i}) = 0$. Following the definition of $B_i(\cdot;\cdot)$ in equation (3-2) and the definition of the logistic function, we can write $B_i(\cdot;\cdot)$ as a product of logistic functions, then

$$g_0(\cdot; \boldsymbol{\nu}_{10}) \prod_{k \in \mathbb{J}_i} g_k(\cdot; \boldsymbol{\nu}_{1k}) = g_0(\cdot; \boldsymbol{\nu}_{20}) \prod_{k \in \mathbb{J}_i} g_k(\cdot; \boldsymbol{\nu}_{2k}). \tag{B-3}$$

If we show $g_0(\cdot; \boldsymbol{\nu}_{10}) = g_0(\cdot; \boldsymbol{\nu}_{20})$, then we can show iteractively that $B_i(\cdot; \boldsymbol{\theta}_{1i}) = B_i(\cdot; \boldsymbol{\theta}_{2i})$:

$$g_0(\cdot; \boldsymbol{\nu}_{10}) = g_0(\cdot; \boldsymbol{\nu}_{20}), \tag{B-4}$$

$$\frac{1}{1 + e^{-\gamma_{10}(x_{s_0,t} - c_{10})}} = \frac{1}{1 + e^{-\gamma_{20}(x_{s_0,t} - c_{20})}}, \tag{B-5}$$

which is true only if $(\gamma_{10}, c_{10}) = (\gamma_{20}, c_{20})$.

Concluding, we have shown that $f(y_t|\mathbf{x}_t;\boldsymbol{\theta}_1) = f(y_t|\mathbf{x}_t;\boldsymbol{\theta}_2)$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. Thus,

$$\prod_{t=1}^{T} f(y_t|\mathbf{x}_t;\boldsymbol{\theta}_1) = \prod_{t=1}^{T} f(y_t|\mathbf{x}_t;\boldsymbol{\theta}_2), \quad a.s.$$

is equivalent to $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Q.E.D.

C
Stationarity and Geometric Ergodicity

As in the study of autoregressive processes, the most fundamental issues involve finding sufficient conditions to ensure the stochastic stability of the process. In this case, we need to know under which conditions a Tree-MM process with only autoregressive local models (i.e. $\mathbf{x}_t$ contains only lags of y_t) will be stationary.

Some results on the stability of mixture of experts were shown in (69). In that case, the conditions were proved for a MixAR$(m; d)$ model. We can use the same results because the limiting behavior of the multinomial logits and the $B(\cdot)$ functions are similar.

Set $\alpha_k \equiv \max_{i \in \mathbb{T}} |\beta_{ik}|$, $k = 1, \ldots, p$, where β_{ik} is the k-th component of $\boldsymbol{\beta}_i$. Then we have the following results:

Teorema C.1 *Let $\{y_t\}_{t \geq 0}$ follow a Tree-MM model (3.1) with AR(p) local models. Assume that the polynomial*

$$\mathcal{P}(z) = z^d - \sum_{k=1}^{p} \alpha_k z^{d-k} \quad ; \; z \in C$$

has all its zeros in the open unit disk, $z < 1$. Then the vector process y_t has a unique stationary probability measure, and is geometrically ergodic.

PROOF. To use the results of (69), we need to show some similarities between the multinomial logit function and the $B(\cdot)$ function. We define $B^{(1)}$ as the left most expert of the tree and $B^{(J)}$ as the right most expert of the tree. Obviously, $B^{(1)}$ is a product of $1 - g(\cdot)$ functions and $B^{(J)}$ is a product of $g(\cdot)$ functions. Furthermore, any $B^{(j)}$ for $j = 2, \ldots, J-1$ has at least one term $g(\cdot)$ and one term $1 - g(\cdot)$.

If we satisfy the following conditions, we can show the equivalence of the proofs.

(i) $B^{(1)} \to 1$ for $y_{s^{(1)}} \to -\infty$

(ii) $B^{(1)} \to 0$ for $y_{s^{(1)}} \to \infty$

(iii) $B^{(J)} \to 1$ for $y_{s^{(J)}} \to \infty$

(iv) $B^{(J)} \to 0$ for $y_{s^{(J)}} \to -\infty$

(v) $B^{(j)} \to 0$ for $y_{s^{(j)}} \to \pm\infty$.

We know that $g(\mathbf{x}_t, \boldsymbol{\nu}_k) \to 1$ for $y_{s_k} \to \infty$ and $g(\mathbf{x}_t, \boldsymbol{\nu}_k) \to 0$ for $y_{s_k} \to -\infty$. Consequently, $[1 - g(\mathbf{x}_t, \boldsymbol{\nu}_k)] \to 0$ for $y_{s_k} \to \infty$ and $[1 - g(\mathbf{x}_t, \nu_k)] \to 1$ for $y_{s_k} \to -\infty$. Then

$$\begin{aligned} \lim_{y_{s(1)} \to -\infty} B^{(1)}(\cdot) &= \lim_{y_{s(1)} \to -\infty} \prod [1 - g(\cdot)] \\ &= \prod \lim_{y_{s(1)} \to -\infty} [1 - g(\cdot)] \\ &= 1, \end{aligned}$$

such that Condition (i) holds.

Conditions (ii)–(v) can be verified using the same steps. Consequently, the results of (69) hold for the Tree-MM model.

Q.E.D.

D
Proofs of Theorems

We follow (64), to prove the existence, consistency and asymptotic normality of the QMLE. Besides, we define some notation to make the proofs clearer.

Define $f_t \equiv f(y_t|\mathbf{x}_t;\theta)$ and $f_t^* \equiv f(y_t|\mathbf{x}_t;\boldsymbol{\theta}^*)$. Following this notation, we also define $\pi_{it} \equiv \pi(y_t|\mathbf{x}_t;\boldsymbol{\psi}_i)$, $\pi_{it}^* \equiv \pi(y_t|\mathbf{x}_t;\boldsymbol{\psi}_i^*)$, $B_{it} \equiv B_i(\mathbf{x}_t;\boldsymbol{\theta}_i)$ and $B_{it}^* \equiv B_i(\mathbf{x}_t;\boldsymbol{\theta}_i^*)$. Furthermore define recursively $f_{k,t} = (1 - g(x_{s_k};\boldsymbol{\nu}_k))f_{2k+1,t} + g(x_{s_k};\boldsymbol{\nu}_k)f_{2k+2,t}$, for all k in $\mathbb{J}$, and $f_{k,t} = \pi_{kt}$, for all k in $\mathbb{T}$.

D.1
Proof of Theorem 4.1

We need to satisfy Assumptions 2.1, 2.3 and 2.4 of Theorem 2.13 in (64), show that $|\mathcal{L}_T(\boldsymbol{\theta})| < \infty$, and that the QMLE has a unique maximum at $\boldsymbol{\theta}^*$.

Assumption 2.1 is satisfied by Assumption 1, and Assumption 2.3 is satisfied by Assumption 2 and Lemma E.1. Assumption 2.4 and $|\mathcal{L}(\boldsymbol{\theta})| < \infty$ are satisfied by Lemma E.1. So we need to show that $\mathcal{L}_T(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta}^*$.

To show that $\mathcal{L}_T(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}^*$, we follow (39) writing the maximization problem as follows:

$$\max_{\boldsymbol{\theta}\in\Theta}\left[\mathcal{L}_T(\boldsymbol{\theta}) - \mathcal{L}_T(\boldsymbol{\theta}^*)\right] = \max_{\boldsymbol{\theta}\in\Theta}\mathbb{E}\left[\log\frac{f_t^*}{f_t} - \frac{f_t}{f_t^*} - 1\right].$$

Furthermore, for any $x > 0$, $m(x) = x - log(x) \leq 0$, then

$$\mathbb{E}\left[\log\frac{f_t^*}{f_t} - \frac{f_t}{f_t^*}\right] \leq 0.$$

Given that $m(x)$ archives its maximum at $x = 1$, $\mathbb{E}[m(x)] \leq \mathbb{E}[m(1)]$ with equality holding almost surely only if $f_t^* = f_t$ with probability one. By the mean value theorem, it is equivalent to show that

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'\frac{\partial \log f_t}{\partial \boldsymbol{\theta}} = 0 \tag{D-1}$$

almost surely. A straightforward application of Lemma E.2 shows that it happens if, and only if, $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ with probability one, which completes the proof.

C.Q.D

D.2
Proof of Theorem 4.2

We must satisfy the conditions of Theorem 3.5 in (64), which are Assumptions 2.1, 2.3, 2.4, 3.1 and 3.2'. Assumptions 2.1, 2.3, 2.4 and 3.2' are satisfied by Assumptions 1–3 (see the proof of Theorem 4.1). Assumption 3.1 states that:

(a) $\mathbb{E}_{\mathcal{G}}(\log f_t) < \infty, \forall t = 1, 2, \ldots;$

(b) $\mathbb{E}_{\mathcal{G}}(\log f_t)$ is continuous in $\boldsymbol{\Theta}$; and

(c) $\{\log f_t\}$ obeys the uniform law of large numbers (ULLN).

it is clear that $\mathbb{E}_{\mathcal{G}}(\log f_t) \leq \log \mathbb{E}_{\mathcal{G}}(f_t) \leq \log \mathbb{E}_{\mathcal{G}}(\sup_t f_t)$. But $\sup_t f_t = \Delta < \infty$, then $\log \left[\mathbb{E}_{\mathcal{G}}(\sup_t f_t)\right] = \log \Delta < \infty$. Then, Condition (a) is satisfied. In addition, note that $\log(\cdot)$, $\mathcal{G}_t$ and f_t are continuous, measurable, and integrable functions, so $h_t = \mathcal{G}_t \log f_t$ is also continuous, measurable, and integrable. Then, $\int h_t dy$ is continuous and Condition (b) is satisfied. Finally, Condition (c) is satisfied by Lemma E.7. As a result, $\widehat{\boldsymbol{\theta}}_T \rightarrow \boldsymbol{\theta}^*$ almost surely.

C.Q.D

D.3
Proof of Theorem 4.3

To prove this theorem, we must satisfy Assumptions 2.1, 2.3, 3.1, 3.2', 3.6, 3.7(a), 3.8, 3.9 and 6.1 in (64). Assumptions 2.1, 2.3, 3.1, 3.2' are satisfied by Assumptions 1–6 (see proof of Theorem 4.2). Assumption 3.6 is satisfied by Lemma E.1, Assumption 3.7(a) is satisfied by Lemma E.4, Assumption 3.8 by Lemmas E.5 and E.7, Assumption 3.9 by Lemma E.6, and Assumption 6.1 is shown here.

Assumption 6.1 requires that $\left\{T^{-1/2}\partial_\theta f_t|_{\theta^*}\right\}$ obeys a central limit theorem with covariance matrix $B(\theta^*)$, where $B(\theta^*)$ is $O(1)$ and uniformly positive definite. First note that, from lemma E.8, $\{\partial_\theta f_t|_{\theta^*}\}$ is a martingale difference process. Then we must show the following to satisfy assumption 6.1:

(a) $T^{-1}\sum_{t=1}^{T} \partial_\theta f_t|_{\theta^*}\partial_{\theta'} f_t|_{\theta^*} \xrightarrow{a.s.} \mathbb{E}(\partial_\theta f_t|_{\theta^*}\partial_{\theta'} f_t|_{\theta^*})$;

(b) the sequence is strictly stationary.

Condition (a) is readily verified by Lemmas E.7 and E.4. Condition (b) is satisfied by Assumption 4. Hence, satisfying these assumptions, we can show that

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\right) \xrightarrow{D} \mathsf{N}(0, \mathbf{I}^{-1}),$$

where $\mathbf{I}(\boldsymbol{\theta}^*) \equiv \mathbf{A}^{-1}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}^{-1}(\boldsymbol{\theta}^*)$ and

$$\begin{aligned}
\mathbf{A}(\boldsymbol{\theta}^*) &= \mathbb{E}\left[-\frac{\partial^2 \mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}^*}\right], \\
\mathbf{B}(\boldsymbol{\theta}^*) &= \mathbb{E}\left[\frac{\partial \mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}^*} \frac{\partial \mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}^*}\right],
\end{aligned}$$

where $\mathcal{L}_T(\boldsymbol{\theta}) = \log \sum_{i\in\mathbb{T}} B_i(\mathbf{x}_t;\boldsymbol{\theta}_i)\,\pi(y_t|\mathbf{x}_t;\boldsymbol{\psi}_i)$.

C.Q.D

E
Lemmas

Lemma E.1 *Under Assumptions (2)–(3) $f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$, defined in (3-1), is a measurable function of $\boldsymbol{Y}_t = [y_t, mathbbx_t]'$, limited, positive and continuously differentiable of order n on $\boldsymbol{\Theta}$.*

PROOF. Trivially, $\pi(y_t|\mathbf{x}_t;\boldsymbol{\psi}_i)$ and $g(x_{s_jt};\boldsymbol{\nu}_j)$ are continuous, mensurable, finite, positive and differentiable functions of $\boldsymbol{Y}_t$. The function $f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$ is a sequence of sums and products of these functions. As a result, $f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$ is a continuous, mensurable, finite, positive and differentiable function of $\boldsymbol{Y}_t$.

C.Q.D

Lemma E.2 *Let $\boldsymbol{d}$ be a constant vector with the same dimension of $\boldsymbol{\theta}$. Then it follows that*

$$\boldsymbol{d}'\left(\frac{\partial \log f_t}{\partial \boldsymbol{\theta}}\right) = 0 \quad a.s.$$

if, and only if, $\boldsymbol{d} = \mathbf{0}$.

PROOF. First write

$$\boldsymbol{d}'\left(\frac{\partial \log f_t}{\partial \boldsymbol{\theta}}\right) = \boldsymbol{d}'\frac{1}{f_t}\frac{\partial f_t}{\partial \boldsymbol{\theta}} = 0. \tag{E-1}$$

From Lemma E.1, we know that $f_t > 0$. Rewriting (E-1), and considering that we can write in terms of $\partial/\partial\boldsymbol{\psi}_i$ and $\partial/\partial\boldsymbol{\nu}_k$, for all $i \in \mathbb{T}$ and $k \in \mathbb{J}$,

$$\boldsymbol{d}'\frac{\partial \pi_{it}}{\partial \boldsymbol{\psi}_i} = 0 \quad \text{and} \quad [f_{2k+1,t}g_{kt} - f_{2k+2,t}(1-g_{kt})]\boldsymbol{d}'\frac{\partial[-\gamma_k(y_t - c_k)]}{\partial \boldsymbol{\nu}_k} = 0,$$

which are both functions of y_t. By the non-degeneracy condition, and supposing that y_t is not null for all $t = 1, 2, \ldots, T$, then $\boldsymbol{d}'\frac{\partial f_t}{\partial \boldsymbol{\theta}} = 0$ if, and only if, $\boldsymbol{d} = \mathbf{0}$.

C.Q.D

Lemma E.3 *Under Assumptions 2, 5 and 6, $\mathbb{E}(\log f_t) < \infty$.*

PROOF. First we make

$$\log f_t = \log \sum_{i\in\mathbb{T}} B_{it}\pi_{it} < \log \sum_{i\in\mathbb{T}} \pi_{it} < \log \#\mathbb{T} + \log \sup_{i\in\mathbb{T}} \pi_{it}. \tag{E-2}$$

Let $\pi_{It} = \pi(y_t|\mathbf{x}_t; \mathbf{x}_t'\boldsymbol{\beta}_I, \sigma_I^2) = \sup_{i\in\mathbb{T}} \pi_{it}$, then equation (E-2) becomes

$$\log \pi_{It} = -\frac{1}{2}\log 2\pi\sigma_I^2 - \frac{1}{2\sigma_I^2}(\boldsymbol{x}_t'\boldsymbol{\beta}_I - y_t)^2. \tag{E-3}$$

Taking the expected value of (E-3), and under Assumptions 2 and 5,

$$\mathbb{E}\left[\log \pi_{It}\right] = -\frac{1}{2}\log 2\pi\sigma_I^2 - \frac{1}{2\sigma_I^2}\mathbb{E}\left[(\boldsymbol{x}_t'\boldsymbol{\beta}_I - y_t)^2\right] < \infty. \tag{E-4}$$

C.Q.D

Lemma E.4 *Under Assumptions 2, 4, 5 and 6,*

$$\mathbb{E}\left(\frac{\partial \log f_t}{\partial \boldsymbol{\theta}}\right) < \infty \quad \textit{and}$$
$$\mathbb{E}\left(\frac{\partial \log f_t}{\partial \boldsymbol{\theta}}\frac{\partial \log f_t}{\partial \boldsymbol{\theta'}}\right) < \infty.$$

PROOF. Let $\partial_\theta \equiv \frac{\partial}{\partial \boldsymbol{\theta}}$. Then we make

$$\partial_\theta \log f_t = \frac{1}{f_t}\partial_\theta f_t = \frac{1}{f_t}\sum_{i\in\mathbb{T}} \pi_{it}\partial_\theta B_{it} + B_{it}\partial_\theta \pi_{it} \leq \Delta_{\pi,f}\sum_{i\in\mathbb{T}} \partial_\theta B_{it} + \Delta_B \sum_{i\in\mathbb{T}} \partial_\theta \pi_{it}, \tag{E-5}$$

where $\Delta_{\pi,f} = \sup_i(f_t^{-1}\pi_{it}) < \infty$ and $\Delta_B = \sup_i f_t^{-1} < \infty$.

This equation can be written in terms of $\partial_{\psi_i} \equiv \partial/\partial\psi_i$ and $\partial_{\nu_j} \equiv \partial/\partial\nu_j$. Let $\Delta_\pi = \sup_i \pi_{it}$, then

$$\partial_{\psi_i}\pi_{it} = \pi_{it}\partial_{\psi_i}\log\pi_{it} \leq \Delta_\pi \partial_{\psi_i}\log\pi_{it}, \tag{E-6}$$
$$\partial_{\nu_j}B_{it} = B_{it}(-g_{jt})(1-g_{jt})\partial_{\nu_j}[-\gamma_j(x_{s_j} - c_j)] \leq \left|\partial_{\nu_j}[-\gamma_j(x_{s_j} - c_j)]\right| \tag{E-7}$$

But $\boldsymbol{\psi}_i = [\beta_{0i}, \ldots, \beta_{pi}, \sigma_i^2]'$, thus the right size of equation (E-6) can be written as

$$\Delta_\pi \partial_{\beta_{ki}}\log\pi_{it} = -\Delta_\pi \frac{\tilde{x}_{kt}(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t)}{\sigma_i^2}, \tag{E-8}$$
$$\Delta_\pi \partial_{\sigma_i^2}\log\pi_{it} = \Delta_\pi\left(-\frac{1}{2\sigma_i^2} + \frac{(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t)^2}{2\sigma_i^4}\right), \tag{E-9}$$

where $\tilde{x}_{kt}$ is the k-th element of the vector $\tilde{\mathbf{x}}_t$.

Using the same argument, we can write the right side of equation (E-7) as

$$\left|\partial_{\gamma_j}[-\gamma_j(x_{s_j} - c_j)]\right| = \left|-(x_{s_j} - c_j)\right|, \tag{E-10}$$
$$\left|\partial_{c_j}[-\gamma_j(x_{s_j} - c_j)]\right| = \left|\gamma_j\right|. \tag{E-11}$$

It is readily verified that, under Assumptions 2, 4 and 5, the expected values of equations (E-8) – (E-11) are finite. Furthermore, under Assumption 6, the expected

value of any product between these equations is also finite.

C.Q.D

Lemma E.5 *Under Assumptions 2, 4, 5 and 6,* $\mathbb{E}(\partial^2 \log f_t/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}') < \infty$.

PROOF. Assume that $\partial_{\theta\theta'} \equiv \frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$. Then we make

$$\partial_{\theta\theta'} \log f_t = -\partial_\theta \log f_t \partial_{\theta'} \log f_t + f_t^{-1}\partial_{\theta\theta'} f_t. \tag{E-12}$$

Using the product law of differentiation, we can write $\partial_{\theta\theta'} f_t$ as a sum of products of $\partial_\theta B_{it}$ and $\partial_\theta \pi_{it}$ with $\partial_{\theta\theta'} B_{it}$ and $\partial_{\theta\theta'} \log \pi_{it}$. Using the results of lemma E.4, the expected value of the product of any two of these derivatives is finite. So, we must show that $\mathbb{E}[\partial_{\theta\theta'} B_{it}] < \infty$ and $\mathbb{E}[\partial_{\theta\theta'} \log \pi_{it}] < \infty$. Considering that ψ_i and ψ_j do not have elements in common, and that B_{it} depends only on the vectors ν_j, $j \in \mathbb{J}_i$, we can write these derivatives in terms of $\partial_{\psi_i\psi_i'}$ and $\partial_{\nu_j\nu_k'}$. But $\boldsymbol{\psi}_i = [\beta_{0i}, \ldots, \beta_{pi}, \sigma_i^2]'$ and $\nu_j = [\gamma_j, c_j]'$. Then

$$\begin{aligned}
\partial_{\beta_{li}\beta_{ki}} log\pi_{it} &= -\sigma_i^{-2}\tilde{x}_{kt}\tilde{x}_{lt}, && \text{(E-13)}\\
\partial_{\beta_{li}\sigma_i^2} log\pi_{it} &= \sigma_i^{-4}\tilde{x}_{lt}(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t), && \text{(E-14)}\\
\partial_{\sigma_i^2\sigma_i^2} log\pi_{it} &= (2\sigma_i^4)^{-1}\sigma_i^{-8}(\tilde{\mathbf{x}}_t'\boldsymbol{\beta} - y_t)^2, && \text{(E-15)}\\
\left|\partial_{\nu_k\nu_j'} B_{it}\right| &< \left|\partial_{\nu_k}[-\gamma_k(x_{s_k} - c_k)]\partial_{\nu_j'}[-\gamma_j(x_{s_j} - c_j)]\right|. && \text{(E-16)}
\end{aligned}$$

It is readily verified that, under Assumptions 2, 4 and 5, the expected values of equations (E-13)–(E-16) are finite.

C.Q.D

Lemma E.6 *Under Assumptions 2, 4, 5 and 6,* $\mathbb{E}(\partial^2 \log f_t/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\theta^*})$ *is negative definite.*

PROOF. If $\mathbb{E}(\partial^2 \log f_t/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\theta^*})$ is negative definite, then $\log f_t$ has a maximum in $\boldsymbol{\Theta}$. We know by Lemma E.2 that $\log f_t$ has only one maximum or minimum in $\boldsymbol{\Theta}$; thus we only have to show that f_t must have a maximum.

Trivially, the Gaussian functions π_{it} have a maximum. If we multiply by a constant or monotone functions or add functions with a maximum, the function still has a maximum. The logistic function is a monotone function (in relation to its parameters and the variable). Hence, $B_{it}\pi_{it}$ has a maximum and f_t has a maximum, and $\mathbb{E}(\partial^2 \log f_t/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'|_{\theta^*})$ is negative definite.

C.Q.D

Lemma E.7 *Under Assumptions 2, 4, 5 and 6,*

(a) $T^{-1}\sum_{t=1}^{T} f_t \overset{a.s.}{\to} \mathbb{E}(f_t)$*;*

(b) $T^{-1}\sum_{t=1}^{T} \partial_\theta f_t \overset{a.s.}{\to} \mathbb{E}(\partial_\theta f_t)$*; and*

(c) $T^{-1}\sum_{t=1}^{T} \partial_{\theta\theta'} f_t \overset{a.s.}{\to} \mathbb{E}(\partial_{\theta\theta'} f_t)$.

PROOF. First we must show that $T^{-1}\sum_{t=1}^{T} y_t \overset{a.s.}{\to} \mathbb{E}(y_t)$. Once y_t is a mixing process, we just need to show that $\mathbb{E}\left(T^{-1}\sum_{t=1}^{T} y_t\right) = \mathbb{E}(y_t)$ and that $\mathbb{V}\left(T^{-1}\sum_{t=1}^{T} y_t\right) < \infty$. The first assumption is trivially satisfied because y_t is stationary. The second assumption is satisfied because $\sum \mathbb{E}(y_t y_{t-k}) < \Delta < \infty$.

Lemma E.1 ensures that f_t, $\partial_\theta f_t$ and $\partial_{\theta\theta'} f_t$ are continuous functions of y_t given θ. Besides, Lemmas E.3, E.4 and E.5 guarantee that the expected value is also finite. Once the functions are continuous and the expected value is finite, we can extend the results of y_t for f_t, $\partial_\theta f_t$ and $\partial_{\theta\theta'} f_t$, thereby completing the proof.

C.Q.D

Lemma E.8 *Under Assumptions (2)–(5),* $\partial\mathcal{L}_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$ *is a martingale difference in terms of* $\mathcal{F}_t$*, the* σ*-field generated by* $\{y_t, \mathbf{x}_t, \ldots\}$*, where* $\mathcal{L}_T(\boldsymbol{\theta}) = \log f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$ *and* $f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$ *is defined in (3-1).*

PROOF. We prove following the definition of martingale differences:

(a) $\mathbb{E}\left[\partial\mathcal{L}_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}^*}\right] < \infty$;

(b) $\mathbb{E}\left[\partial\mathcal{L}_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}^*}|\mathcal{F}_{t-1}\right] = 0$

The first condition is satisfied by Lemma E.4. The second condition is satisfied by Theorem 4.1 and Lemma E.4. Satisfying Conditions (a) and (b), $\partial\mathcal{L}_T(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}^*}$ is a martingale difference in terms of $\mathcal{F}_t$.

C.Q.D