# Likelihood Ratio Based Tests for Markov Regime Switching*

Zhongjun Qu† Fan Zhuo‡

Boston University Boston University

October 30, 2015

**Abstract**

Markov regime switching models are widely considered in economics and finance. Although there have been persistent interests (see e.g., Hansen, 1992, Garcia, 1998, and Cho and White, 2007), the asymptotic distributions of likelihood ratio based tests have remained unknown. This paper considers such tests and establishes their asymptotic distributions in the context of nonlinear models allowing for multiple switching parameters. The analysis simultaneously addresses three difficulties: (i) some nuisance parameters are unidentified under the null hypothesis, (ii) the null hypothesis yields a local optimum, and (iii) conditional regime probabilities follow stochastic processes that can only be represented recursively. Addressing these issues permits substantial power gains in empirically relevant situations. Besides obtaining the tests' asymptotic distributions, this paper also obtains four sets of results that can be of independent interest: (1) a characterization of conditional regime probabilities and their high order derivatives with respect to the model's parameters, (2) a high order approximation to the log likelihood ratio permitting multiple switching parameters, (3) a refinement to the asymptotic distribution, and (4) a unified algorithm for simulating the critical values. For models that are linear under the null hypothesis, the elements needed for the algorithm can all be computed analytically. The above results also shed light on why some bootstrap procedures can be inconsistent and why standard information criteria, such as the Bayesian information criterion (BIC), can be sensitive to the hypothesis and the model's structure. When applied to the US quarterly real GDP growth rates, the methods suggest fairly strong evidence favoring the regime switching specification, which holds consistently over a range of sample periods.

**Keywords**: Hypothesis testing, likelihood ratio, Markov switching, nonlinearity.
**JEL codes**: C12, C22, E32.

*We thank James Hamilton, Hiroaki Kaido, Pierre Perron and seminar participants at Brown, BU statistics, 2014 JSM, the 23rd SNDE, and the 11th World Congress of the Econometric Society for valuable suggestions, and Carrasco, Hu and Ploberger for making their code available.

†Department of Economics, Boston University, 270 Bay State Rd., Boston, MA, 02215 (qu@bu.edu).

‡Department of Economics, Boston University, 270 Bay State Rd., Boston, MA, 02215 (zhuo@bu.edu).

# 1 Introduction

Markov regime switching models are widely considered in economics and finance. Hamilton (1989) is a seminal contribution, which provides not only a framework for describing economic recessions, but also a general algorithm for filtering, smoothing and maximum likelihood estimation while building on the work of Goldfeld and Quandt (1973) and Cosslett and Lee (1985). Surveys of this voluminous literature can be found in Hamilton (2008, 2014).

Three approaches have been considered for detecting regime switching. The first approach translates this issue into testing for parameter homogeneity against heterogeneity. Neyman and Scott (1966) studied the $C(\alpha)$ test. Chesher (1984) derived a score test and showed that it is closely related to the information matrix test of White (1982). Lancaster (1984) and Davidson and MacKinnon (1991) are related contributions. Carrasco, Hu and Ploberger (2014) have further developed this approach by allowing for an alternative hypothesis where the heterogeneity takes the form of a weakly dependent process. Their test is asymptotically locally optimal in the sense that there exists no test that is more powerful for a specific alternative characterized in their paper. The above tests have two common features. First, they only require estimating the model under the null hypothesis. Second, they are designed for detecting general random coefficients, not particularly Markov regime switching. As a result, the tests can have power against a broad class of alternatives, meanwhile the power can be lower than what is achievable if the parameters indeed follow a finite state Markov chain.

The second approach, due to Hamilton (1996), is to conduct generic tests of the hypothesis that a $K$-regime model (e.g., $K = 1$) adequately describes the data. The insight is that if a $K$-regime specification is accurate, then the score function should have mean zero and form a martingale difference sequence. Otherwise, the model should be enriched to allow for additional features, in some situations by introducing an additional regime. Hamilton (1996) demonstrated how to implement such tests as a by-product of calculating the smoothed probability that a given observation is from a particular regime. This makes the tests simple and widely applicable. Meanwhile, it remains important and useful to have testing procedures that focus specifically on detecting Markov switching alternatives.

The third approach proceeds under the (quasi) likelihood ratio principle. The (quasi) likelihood functions are constructed assuming a single regime under the null and two regimes under the alternative hypothesis. The analysis faces three challenges. (i) Some nuisance parameters are unidentified under the null hypothesis. As a result, the log likelihood ratio is locally non-quadratic, causing the Chi-square approximation to its distribution to break down. This gives rise to the Davies (1977) problem. (ii) The null hypothesis yields a local optimum (c.f. Hamilton, 1990), making the score function identically zero when evaluated at the null parameter estimates. Consequently, a second order Taylor approximation to the likelihood ratio is insufficient for analyzing its asymptotic properties. (iii) The conditional regime probability, i.e., the probability of being in a particular regime at time $t$ given the information up to the time $t-1$, follows a stochastic process that can only be represented recursively. The first two difficulties are also present when testing for mixtures. It is the simultaneous occurrence of all three difficulties that plagues the study of the likelihood ratio in the current context. For example, when analyzing high order expansions of the likelihood ratio, it is necessary to study high order derivatives of the conditional regime probability with respect to the model's parameters. So far, the statistical properties of the latter have remained elusive. Consequently, the asymptotic distribution of the log likelihood ratio has also remained unknown.

Meanwhile, several important progresses have been made by Hansen (1992), Garcia (1998), Cho and White (2007), and Carter and Steigerwald (2012). Specifically, Hansen (1992) clearly documented why the difficulties (i) and (ii) cause the conventional approximation to the likelihood ratio to break down. Further, he treated the likelihood function as a stochastic process indexed by the transition probabilities (i.e., the probabilities of remaining in the first regime $p$ and remaining in the second regime $q$) and the switching parameters, and derived a bound for its asymptotic distribution. His result provides a platform for conducting conservative inference. Garcia (1998) suggested an approximation to the log likelihood ratio that would follow if the score had a positive variance at the null estimates. Results in the current paper will show that this distribution is in general different from the actual limiting distribution. Recently, Cho and White (2007) made a significant progress. They suggested a *quasi* likelihood ratio (QLR) test against a two-component mixture alternative (i.e., a model where the current regime arrives independently of its past values). There, the difficulty (iii) is avoided because the conditional regime probability is reduced to a constant, which can further be treated as an additional unknown parameter. Carter and Steigerwald (2012) further discussed a consistency issue related to QLR test. The current paper makes use of several important techniques in Cho and White (2007). At the same time, it goes beyond their framework to confront directly Markov switching alternatives. As will be seen, the power gains from doing so can be quite substantial.

Specifically, this paper considers a family of likelihood ratio based tests and establishes their asymptotic distributions in the context of nonlinear models allowing for *multiple* switching parameters. The framework encompasses the important special cases of testing for regime switching in autoregressive models and in autoregressive distributed lags models. Throughout the analysis, the model has two regimes under the alternative hypothesis. Some parameters can remain constant across the two regimes. The analysis is structured in five steps:

1. We characterize the dynamics of the conditional regime probability (i.e., the probability of being in a particular regime at time $t$ given the information up to the time $t-1$) and its high order derivatives with respect to the model's parameters. We show that, when evaluated at the null parameter estimates, the former reduces to a constant while the latter can all be represented as linear first order difference equations with the lagged coefficients equal to $p+q-1$. Because $0 < p, q < 1$, these equations are all stable and amenable to the applications of uniform laws of large numbers and functional central limit theorems. This novel characterization is a critical step that makes the subsequent analysis feasible.

2. We fix $p$ and $q$ and derive a fourth order Taylor approximation to the likelihood ratio. This step builds on the analysis in Cho and White (2007), but accounts for the effect of the time variation in the conditional regime probability. The results are informative about why substantial power gains relative to the QLR test are possible when the data are not generated by simple mixtures.

3. We view the likelihood ratio as an empirical process indexed by $p$ and $q$ and derive its limiting distribution. The values of $p$ and $q$ are required to be strictly between 0 and 1 satisfying $p+q \geq 1+\epsilon$ with $\epsilon$ being some arbitrarily small constant. These requirements are compatible with applications in macroeconomics and finance; see the discussion in Section 3. The empirical process perspective undertaken here follows a rich array of studies, including Hansen (1992), Garcia (1998), Cho and White (2007), and Carrasco, Hu and Ploberger (2014).

4. While the above limiting distributions are adequate for a broad class of models, they can lead to over-rejections when a further singularity (the source of which is specified later) is present. To overcome this problem, we analyze a sixth order expansion of the likelihood ratio along the line $p+q=1$ and an eighth order expansion at $p=q=1/2$. The leading terms are then incorporated into the asymptotic distribution to safe guard against their effects. This leads to a refined distribution that delivers reliable approximations throughout our experimentations. This refinement is valid whether or not this singularity is truly present.

5. We provide a unified algorithm for simulating the refined asymptotic distribution. For models that are linear under the null hypothesis, the elements needed for this algorithm can all be computed analytically. This permits developing a computer program, which mainly requires the researcher writing down the model under the null hypothesis, specifying which parameters are allowed to switch, and providing the permissible values for the two transition probabilities.

The asymptotic distribution shows some uncommon features. First, nuisance parameters, though constrained to be constant across the regimes, can affect the limiting distribution. Second, properties of the regressors (i.e., whether they are strictly or weakly exogenous) also affect the distribution. Third, the distribution depends on which parameter (i.e., the intercept, the slope or the residual variance) is allowed to switch. These features imply that some bootstrap procedures can be inconsistent and that standard information criteria, such as BIC, can be sensitive to the hypothesis and the model's structure. The above implications are further discussed in Section 6.

We conduct simulations using a data generating process (DGP) considered in Cho and White (2007). The results show that the power difference can be large when the regimes are persistent, a situation that is common in practice. We also apply the testing procedure to the US quarterly real GDP growth rates, over the period 1960:I-2014:IV and a range of subsamples. The results consistently favor the regime switching specification. In addition, the smoothed regime probabilities closely mirror NBER's recession dating. To our knowledge, this is the first time such consistent evidence for regime switching in mean output growth is documented through hypothesis testing.

From a methodological perspective, this paper contributes to the literature that studies hypothesis testing when some regularity conditions fail to hold. Besides the works mentioned above, closely related studies include the following. Davies (1987), King and Shively (1993), Andrews and Ploberger (1994, 1995), and Hansen (1996) considered tests when a nuisance parameter is unidentified under the null hypothesis. Andrews (2001) studied tests when, in addition to the above feature, some parameter lies on the boundary of the maintained hypothesis. Hartigan (1985), Ghosh and Sen (1985), Lindsay (1995), Liu and Shao (2003), Chen and Li (2009), and Gu, Koenker and Volgushev (2013) tackled the issues of zero score and/or unidentified nuisance parameters in the context of mixture models. Chen, Ponomareva and Tamer (2014) considered uniform inference on the mixing probability in mixture models when nuisance parameters are present. Rotnitzky, Cox, Bottai and Robins (2000) developed a theory for deriving the asymptotic distribution of the likelihood ratio statistic when the information matrix has rank one less than full; also see the discussions in their paper (page 244) for other studies on the same issue in various contexts. Dovonon and Renault

(2013) studied distributions of tests for moment restrictions when the associated Jacobian matrix is degenerate at the true parameter value. The current work is the first that simultaneously tackles the difficulties (i) to (iii) in the hypothesis testing literature. We conjecture that the techniques developed can have implications for hypothesis testing in other related contexts that involve models with hidden Markov structures.

The paper is structured as follows. Section 2 presents the model and the hypotheses. Section 3 introduces a family of test statistics. Section 4 studies the asymptotic properties of the log likelihood ratio for prespecified $p$ and $q$. Section 5 presents four sets of results. It establishes the weak convergence of the second order derivative of the concentrated log likelihood. It provides the limiting distribution of the test statistic. It introduces a finite sample refinement. Finally, it outlines an algorithm for obtaining the relevant critical values. Section 6 discusses some implications of the theory for bootstrapping and information criteria. Section 7 examines the test's finite sample properties. Section 8 considers an application to the US real GDP growth rates. Section 9 concludes. All proofs are in the appendix.

The following notation is used. $||x||$ is the Euclidean norm of a vector $x$. $||X||$ is the vector induced norm of a matrix $X$. $x^{\otimes k}$ and $X^{\otimes k}$ denote the $k$-fold Kronecker product of $x$ and $X$, respectively. The expression $\text{vec}(A)$ stands for the vectorization of a $k$ dimensional array $A$. For example, for a three dimensional array $A$ with $n$ elements along each dimension, $\text{vec}(A)$ returns a $n^3$-vector whose $(i + (j-1)n + (k-1)n^2)$-th element equals $A(i, j, k)$. $\mathbf{1}_{\{\cdot\}}$ is the indicator function. For a scalar valued function $f(\theta)$ of $\theta \in R^p$, $\nabla_\theta f(\theta_0)$ denotes a $p$-by-1 vector of partial derivatives evaluated at $\theta_0$, $\nabla_{\theta'} f(\theta_0)$ equals the transpose of $\nabla_\theta f(\theta_0)$, and $\nabla_{\theta_j} f(\theta_0)$ denotes its $j$-th element. In addition, $\nabla_{\theta_{j_1}} \nabla_{\theta_{j_2}} \cdots \nabla_{\theta_{j_k}} f(\theta_0)$ denotes the $k$-th order partial derivative of $f(\theta)$ taken sequentially with respect to the $j_1, j_2, ..., j_k$-th element of $\theta$ evaluated at $\theta_0$. The symbols "$\Rightarrow$", "$\to^d$" and "$\to^p$" denote weak convergence under the Skorohod topology, convergence in distribution and in probability, and $O_p(\cdot)$ and $o_p(\cdot)$ is the usual notation for the orders of stochastic magnitude.

## 2 Model and hypotheses

We sequentially discuss the following issues: the model, the log likelihood functions under the null (i.e., one regime) and the alternative (i.e., two regimes) hypothesis, and some assumptions related to these two aspects.

The model is as follows. Let $\{(y_t, x_t')\}$ be a sequence of random vectors with $y_t$ being a scalar and $x_t$ a finite dimensional vector. Let $s_t$ be an unobserved binary variable, whose value determines the regime at time $t$. Define the information set at time $t-1$ as

$$\Omega_{t-1} = \sigma\text{-field}\left\{..., x'_{t-1}, y_{t-2}, x'_t, y_{t-1}\right\}. \tag{1}$$

Let $f(\cdot|\Omega_{t-1};\beta,\delta)$ denote the conditional density of $y_t$, satisfying

$$y_t|(\Omega_{t-1}, s_t) \backsim \begin{cases} f(\cdot|\Omega_{t-1};\beta,\delta_1), & \text{if } s_t = 1, \\ f(\cdot|\Omega_{t-1};\beta,\delta_2), & \text{if } s_t = 2, \end{cases} \quad (t = 1, ...., T). \tag{2}$$

This specification allows the vector $\delta$ to switch between $\delta_1$ and $\delta_2$, while restricting the vector $\beta$ to remain constant across the regimes. Henceforth, we abbreviate the two densities on the right hand side of (2) as $f_t(\beta,\delta_1)$ and $f_t(\beta,\delta_2)$, respectively.

The regimes are Markovian, i.e., $p(s_t = 1|\Omega_{t-1}, s_{t-1} = 1, s_{t-2}, ...) = p(s_t = 1|s_{t-1} = 1) = p$ and $p(s_t = 2|\Omega_{t-1}, s_{t-1} = 2, s_{t-2}, ...) = p(s_t = 2|s_{t-1} = 2) = q$. The resulting stationary (or invariant) probability for $s_t = 1$ is given by

$$\xi_* \equiv \xi_*(p,q) = \frac{1-q}{2-p-q}. \tag{3}$$

Evaluated at $0 < p, q < 1$, the log likelihood function associated with (2) is such that

$$\begin{aligned} &\mathcal{L}^A(p,q,\beta,\delta_1,\delta_2) \qquad (4)\\ = &\sum_{t=1}^{T} \log\left\{f_t(\beta,\delta_1)\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2) + f_t(\beta,\delta_2)(1-\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2))\right\}, \end{aligned}$$

where $\xi_{t|t-1}(\cdot)$ denotes the probability of $s_t = 1$ given $\Omega_{t-1}$, i.e.,

$$\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2) = p(s_t = 1|\Omega_{t-1}; p,q,\beta,\delta_1,\delta_2) \quad (t = 1, ..., T), \tag{5}$$

which satisfies the following recursive relationship

$$\xi_{t|t}(p,q,\beta,\delta_1,\delta_2) = \frac{f_t(\beta,\delta_1)\,\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2)}{f_t(\beta,\delta_1)\,\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2) + f_t(\beta,\delta_2)\,(1-\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2))}, \tag{6}$$

$$\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2) = p\xi_{t|t}(p,q,\beta,\delta_1,\delta_2) + (1-q)(1-\xi_{t|t}(p,q,\beta,\delta_1,\delta_2)). \tag{7}$$

Throughout the paper, we set the initial value $\xi_{1|0} = \xi_*$. As shown later, using a different initial value does not affect the asymptotic results. When $\delta_1 = \delta_2 = \delta$, the log likelihood reduces to

$$\mathcal{L}^N(\beta,\delta) = \sum_{t=1}^{T} \log f_t(\beta,\delta). \tag{8}$$

This paper studies tests based on (8) and (4) for the single regime specification against the two regimes specification given in (2). To proceed, we impose the following restrictions on the DGP and the parameter space. Let $n_\beta$ and $n_\delta$ denote the dimensions of $\beta$ and $\delta$.

**Assumption 1** *(i) The random vector* $(x_t', y_t)$ *is strict stationary, ergodic and* $\beta$*-mixing with the mixing coefficient* $\beta_\tau$ *satisfying* $\beta_\tau \leq c\rho^\tau$ *for some* $c > 0$ *and* $\rho \in [0, 1)$. *(ii) Under the null hypothesis,* $y_t$ *is generated by* $f(\cdot|\Omega_{t-1}; \beta_*, \delta_*)$*, where* $\beta_*$ *and* $\delta_*$ *are interior points of* $\Theta \subset \mathbb{R}^{n_\beta}$ *and* $\Delta \subset \mathbb{R}^{n_\delta}$ *with* $\Theta$ *and* $\Delta$ *being compact.*

Part (i) is the same as Assumption A.1(i) in Cho and White (2007). As discussed there, the $\beta$-*mixing* condition is commonly used when analyzing Markov processes. It allows $x_t$ to be affected by regime switching under the null hypothesis. Part (ii) specifies the true parameter values. The interior point requirement ensures that the asymptotic expansions considered later are well defined.

**Assumption 2** *Under the null hypothesis: (i)* $(\beta_*, \delta_*)$ *uniquely solves* $\max_{(\beta,\delta)\in\Theta\times\Delta} E[\mathcal{L}^N(\beta, \delta)]$*; (ii) for any* $0 < p, q < 1$*,* $(\beta_*, \delta_*, \delta_*)$ *uniquely solves* $\max_{(\beta,\delta_1,\delta_2)\in\Theta\times\Delta\times\Delta} E\mathcal{L}^A(p, q, \beta, \delta_1, \delta_2)$.

Part (i) implies that $(\beta, \delta)$ is globally identified at $(\beta_*, \delta_*)$ under the null hypothesis. Part (ii) implies that there does not exist a two-regime specification (i.e., with $\delta_1 \neq \delta_2$) that is observationally equivalent to the single-regime specification (i.e., with $\delta_1 = \delta_2 = \delta_*$). The next assumption relates the identification properties in Assumption 2 to some asymptotic properties of the estimators.

**Assumption 3** *Under the null hypothesis, we have: (i)* $T^{-1}[\mathcal{L}^N(\beta, \delta) - E\mathcal{L}^N(\beta, \delta)] = o_p(1)$ *holds uniformly over* $(\beta, \delta) \in \Theta \times \Delta$ *with* $T^{-1}\sum_{t=1}^T (\nabla_{(\beta',\delta')'} \log f_t(\beta, \delta)(\nabla_{(\beta',\delta')} \log f_t(\beta, \delta))$ *being positive definite in an open neighborhood of* $(\beta_*, \delta_*)$ *for sufficiently large T; (ii) for any* $0 < p, q < 1$*,* $T^{-1}[\mathcal{L}^A(p, q, \beta, \delta_1, \delta_2) - E\mathcal{L}^A(p, q, \beta, \delta_1, \delta_2)] = o_p(1)$ *holds uniformly over* $(\beta, \delta_1, \delta_2) \in \Theta \times \Delta \times \Delta$.

The above assumption states that (8) and (4) both satisfy uniform laws of large numbers. Along with Assumption 2, it implies that, under the null hypothesis, the maximizers of (8) and (4) for $0 < p, q < 1$ converge in probability to $(\beta_*, \delta_*)$ and $(\beta_*, \delta_*, \delta_*)$ respectively. This assumption allows (4) to have multiple local maximizers. The latter feature will be accounted for when analyzing the likelihood expansions.

Assumptions 1 to 3 are similar to those used in Cho and White (2007), with two important differences. First, the likelihood (4) corresponds to a Markov switching model, not a mixture model. Second, *multiple* parameters are allowed to be affected by the regime switching.

Using the above notation, the null and alternative hypotheses can be more formally stated as:

$$H_0 \; : \; \delta_1 = \delta_2 = \delta_* \text{ for some unknown } \delta_*;$$

$$H_1 \; : \; (\delta_1, \delta_2) = (\delta_1^*, \delta_2^*) \text{ for some unknown } \delta_1^* \neq \delta_2^* \text{ and } (p, q) \in (0, 1) \times (0, 1).$$

Technically, as discussed in Cho and White (2007), the null hypothesis can also be formulated as: $H_0' : p = 1$ and $\delta_1 = \delta_*$ or $H_0'' : q = 1$ and $\delta_2 = \delta_*$. In $H_0'$, because the model remains in the first regime with probability 1, any statement about the second regime becomes irrelevant. The reversed holds for $H_0''$.

Below, we introduce a model that will be used throughout the paper to illustrate the main components of the theory.

**An illustrative model.** An important application of regime switching is to linear models with Gaussian errors:

$$y_t = z_t'\alpha + w_t'\gamma_1 \mathbf{1}_{\{s_t=1\}} + w_t'\gamma_2 \mathbf{1}_{\{s_t=2\}} + u_t, \tag{9}$$

where $\alpha, \gamma_1$ and $\gamma_2$ are unknown finite dimensional parameter vectors and $u_t$ are i.i.d. Normally distributed whose unknown variance can also potentially switch. The variables $z_t$ and $w_t$ can include lagged values of $y_t$. Therefore, the specification encompasses finite order autoregressive models and autoregressive distributed lags models as special cases. In relation to (1) and (2), we have $\Omega_{t-1} = \sigma\text{-field}\{..., z_{t-1}', w_{t-1}', y_{t-2}, z_t', w_t', y_{t-1}\}$ and $x_t' = (z_t', w_t')$. Three situations can arise depending on which parameters are allowed to be affected by regime switching: (a) Only the variance of $u_t$ is affected. Let $\sigma_1^2$ and $\sigma_2^2$ denote its variances under the two regimes. Then, in relation to (2), we have $\delta_1 = \sigma_1^2, \delta_2 = \sigma_2^2$ and $\beta' = (\alpha', \gamma')$ with $\gamma = \gamma_1 = \gamma_2$. (b) Only the regression coefficients are affected. Let $\sigma^2$ denote the variance of $u_t$. Then, we have $\delta_1 = \gamma_1$, $\delta_2 = \gamma_2$ and $\beta' = (\alpha', \sigma^2)$. (c) Both components are affected. We have $\delta_1' = (\gamma_1', \sigma_1^2)$, $\delta_2' = (\gamma_2', \sigma_2^2)$ and $\beta = \alpha$. The results in this paper will encompass all three situations. In the most general situation (c), the densities corresponding to (2) are given by

$$\begin{bmatrix} f_t(\beta, \delta_1) \\ f_t(\beta, \delta_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{(y_t - z_t'\alpha - w_t'\gamma_1)^2}{2\sigma_1^2}\right\} \\ \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{(y_t - z_t'\alpha - w_t'\gamma_2)^2}{2\sigma_2^2}\right\} \end{bmatrix}.$$

Note that the normality assumption in this model can be replaced by other distributional assumptions, provided that $f_t(\beta, \delta_1)$ and $f_t(\beta, \delta_2)$ are replaced by the appropriate densities.

We now illustrate Assumptions 1-3 using this model. Regarding Assumption 1, because of the linearity, the *$\beta$-mixing* requirement of $(x_t', y_t)$ reduces to that of $x_t$. The latter is satisfied if $x_t$ follows a stationary VARMA(P,Q) process $\sum_{j=0}^{P} B_j x_{t-j} = \sum_{j=0}^{Q} A_j \varepsilon_{t-j}$ with $\varepsilon_t$ being mean zero i.i.d. random vectors whose density is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{\dim(\varepsilon_t)}$; see Mokkadem (1988). Other processes that are *$\beta$-mixing* with a geometric rate of decay, as reviewed in Chen (2013), include those generated by threshold autoregressive models, functional coefficient autoregressive models, and GARCH and stochastic volatilities models. Regarding Assumption 2, part (i) is satisfied if $Ex_t x_t'$ has full rank. Part (b) requires that, if the data are generated by $\delta_1 \neq \delta_2$ with $0 < p, q < 1$, the conditional distribution of $y_t$ will exhibit features that are not captured by the single regime linear specification. That is, the resulting Kullback-Leibler divergence will be positive. Finally, in Assumption 3, the rank requirement essentially requires $T^{-1}\sum_{t=1}^{T} x_t x_t'$ to be positive definite in large samples. The rest of the assumption requires uniform laws of large numbers to hold. Because $\xi_{t|t-1}(p, q, \beta, \delta_1, \delta_2)$ are bounded between 0 and 1, the latter hold under Assumption 1 and mild conditions on the moments of $y_t$ and $x_t$. ■

# 3 The test statistic

This section studies three issues. First, it considers a family of test statistics based on the log likelihood ratio. Second, it previews the difficulties involved in deriving the limiting distribution and outlines the strategies for addressing them. Third, it examines empirically relevant values for the transition probabilities $p$ and $q$. The latter is important not only for making the tests practically relevant, but also for the technical analysis needed later in the paper.

Let $\widetilde{\beta}$ and $\widetilde{\delta}$ denote the maximizer of the null log likelihood:

$$(\widetilde{\beta}, \widetilde{\delta}) = \arg\max_{\beta,\delta} \mathcal{L}^N(\beta, \delta). \tag{10}$$

The log likelihood ratio evaluated at some $0 < p, q < 1$ then equals

$$LR(p, q) = 2\left[\max_{\beta,\delta_1,\delta_2} \mathcal{L}^A(p, q, \beta, \delta_1, \delta_2) - \mathcal{L}^N(\widetilde{\beta}, \widetilde{\delta})\right]. \tag{11}$$

This leads to the following test statistic:

$$\text{SupLR}(\Lambda_\epsilon) = \underset{(p,q)\in\Lambda_\epsilon}{\text{Sup}} LR\left(p, q\right),$$

where $\Lambda_\epsilon$ is a compact set specified below and the supremum is taken to obtain the strongest evidence against the null hypothesis. Operators other than the supremum can also be used. For example, following Andrews and Ploberger (1994) and Carrasco, Hu and Ploberger (2014), one can consider $\text{ExpLR}(\Lambda_\epsilon) = \int_{\Lambda_\epsilon} LR\left(p, q\right) dJ(p, q)$, where $J(p, q)$ is a function that assigns weights on $p$ and $q$. Such considerations lead to a family of test statistics based on $LR(p, q)$. This paper focuses on $\text{SupLR}(\Lambda_\epsilon)$; the results extend immediately to $\text{ExpLR}(\Lambda_\epsilon)$.

The test statistic $\mathrm{SupLR}(\Lambda_\epsilon)$ is not new. For example, it has been studied by Hansen (1992) and Garcia (1998). The contribution of this paper is in obtaining an adequate approximation to its finite distribution and in providing an algorithm for simulating it. We now discuss nonstandard features associated with this statistic and highlight our strategy for tackling them.

First, the time varying regime probability $\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2)$ present challenges. On the one hand, it can only be expressed recursively; see (6) and (7). On the other hand, when obtaining asymptotic expansions, it is essential to study its high order derivatives with respect to $\beta,\delta_1$ and $\delta_2$. So far, its effect on the log likelihood ratio has remained unknown, even for the simplest situation where $p$ and $q$ are prespecified. In an important contribution, Cho and White (2007, p. 1675) suggested to avoid this difficulty by replacing the likelihood (4) with that for a mixture (i.e., a model where the regime arrives independently of the past with $\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2)=\pi$ throughout the sample): $\sum_{t=1}^{T}\log\{\pi f_t(\beta,\delta_1)+(1-\pi)f_t(\beta,\delta_2)\}$. However, this *quasi* log likelihood function behaves differently from the actual likelihood when $p+q-1$ is different from zero. This can translate into large power differences as seen later in this paper. This is troubling because in economic and financial applications the regimes are typically substantially serially dependent. In this paper, we make progresses by observing that $\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2)$ and its derivatives can all be characterized as first order difference equations, whose properties further simplify drastically once we evaluate them at $(\widetilde{\beta},\widetilde{\delta})$. This is a critical step that makes the subsequent analysis feasible.

Second, the log likelihood ratio has three nonstandard features as in mixture models: (i) The values of $p$ and $q$ are unidentified under the null hypothesis. Consequently, there are infinite directions to approach any one distribution in the null hypothesis (i.e., the score space is infinite dimensional). This complicates matters because a key step in establishing the asymptotic property of the likelihood ratio lies in determining what happens to the score function as we approach the null hypothesis. To address this, we treat the log likelihood ratio as an empirical process indexed by $p$ and $q$, such that once they are fixed, the score space becomes finite dimensional. Such an empirical process perspective follows from a rich array of studies, with the most closely related being Hansen (1992), Garcia (1998), Cho and White (2007), and Carrasco, Hu and Ploberger (2014). (ii) The score of (4) is identically zero when evaluated at the null parameter estimates. Consequently, a second order Taylor expansion is insufficient for analyzing the likelihood ratio. To address this, we obtain likelihood expansions of the fourth order, and in some specifications, of the eighth order. The obstacles for deriving such expansions are substantial, especially given that we allow for multiple switching parameters. (iii) The values $p=1$ and $q=1$ fall on the boundary of parameter space.

Cho and White (2007) addressed this issue by considering the surfaces specified by $H_0, H_0'$ and $H_0''$ and then combined the results to obtain the null limiting distribution. Here, such an approach is no longer feasible because of the additional challenge introduced by $\xi_{t|t-1}(p, q, \beta, \delta_1, \delta_2)$. We pursue a different route. That is, when defining the test statistic, we restrict the supports of $p$ and $q$ to be closed subsets of $(0, 1)$. This approach has also been used when testing for structural changes (e.g., Hawkins, 1987, Andrews, 1993, Andrews and Ploberger, 1994, and Bai and Perron, 1998) and threshold effects (e.g., Hansen, 1996). It is also used in Hansen (1992) and Garcia (1998).

We now examine empirically relevant values for the transition probabilities $p$ and $q$. Hamilton (2008, the first paragraph in p.1) reviewed 12 articles that applied regime switching models in a wide range of contexts. Among them, 10 articles considered two-regime specifications with constant transition probabilities. These studies are related to: exchange rates (Jeanne and Masson, 2000), output growth (Hamilton, 1989 and Chauvet and Hamilton, 2006), interest rates (Hamilton, 1988, 2005, Ang and Bekaert, 2002b), debt-output ratio (Davig, 2004), bond prices (Dai, Singleton and Yang, 2007), equity returns (Ang and Bekaert, 2002a) and consumption and dividend processes (Garcia, Luger and Renault, 2003). Eighteen sets of estimates are reported. The values of the transition probabilities are between 0.855 and 0.998 for the more persistent regime and 0.740 and 0.997 for the other. These estimates are representative of applications in economics and finance and they strongly suggest two features. First, none of the values correspond to mixtures. That is, the values of $p + q$ are all substantially above 1.0. Second, at least one regime is fairly persistent. That is, the value of $p$ (and $q$) can be fairly close to 1.0.

Motivated by the above observations, we suggest to specify $\Lambda_\epsilon$ as follows

$$\Lambda_\epsilon = \{(p, q) : p + q \geq 1 + \epsilon \text{ and } \epsilon \leq p, q \leq 1 - \epsilon \text{ with } \epsilon > 0\}. \tag{12}$$

This set can be generalized to allow for different trimming proportions (e.g., replacing $p+q \geq 1+\epsilon$ and $\epsilon \leq p, q \leq 1-\epsilon$ with $p+q \geq 1+\epsilon_1$ and $\epsilon_2 \leq p, q \leq 1-\epsilon_3$ with $\epsilon_1, \epsilon_2, \epsilon_3 > 0$). The set can also be narrowed if additional information about $p$ and $q$ is available. For example, if their values are both expected to be higher than 0.5, then we can consider

$$\{(p, q) : 0.5 + \epsilon \leq p, q \leq 1 - \epsilon \text{ with } \epsilon > 0\}. \tag{13}$$

The specification (13) is in fact consistent with all the 10 studies mentioned in the previous paragraph. In this paper, we focus on (12); the results continue to hold for the latter two specifications, provided that the set $\Lambda_\epsilon$ in the limiting distribution is changed accordingly.

As will be seen in the next section, for certain models and hypotheses, the asymptotic distributions of $LR(p,q)$ at $p+q=1$ and $p+q>1$ can be different. This arises when

$$\nabla_{\delta_{i_1}}\nabla_{\delta_{i_2}} f_t(\widetilde{\beta},\widetilde{\delta}) = \alpha'_{i_1 i_2}\nabla_{(\beta',\delta'_1)'} f_t(\widetilde{\beta},\widetilde{\delta}) \tag{14}$$

holds for some $i_1, i_2 \in \{1, ..., n_\delta + n_\beta\}$, where $\alpha_{i_1 i_2}$ is a known vector of constants. Because $p+q=1$ falls out of the set (12), such a change in the distribution does not interfere with the first order asymptotic approximation to the likelihood ratio. However, the issue of approximation adequacy when $\epsilon$ is small will arise and we shall account for it using a high order refinement as follows. First, we derive an asymptotic approximation to the likelihood ratio that is valid over (12) whether or not (14) holds. Then, we study the adequacy of this approximation when (14) holds. The analysis will show that the approximation becomes less adequate when $p+q$ is close to 1. Next, we derive an higher order expansion of the likelihood ratio under $p+q=1$. Finally, the additional terms in this expansion are incorporated into the original asymptotic distribution to obtain a refined approximation. Note that whether or not (14) holds, as well as the values of $\alpha_{i_1 i_2}$, will be known once the model and the hypotheses are specified.

# 4 The log likelihood ratio under prespecified $p$ and $q$

The conditional regime probability $\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2)$ represents the key difference between Markov switching and mixture models. We therefore begin by studying this quantity as well as its derivatives with respect to $\beta, \delta_1$ and $\delta_2$. The results will further enable us to develop expansions of the concentrated log likelihood under the null hypothesis. The results reported in this section all hold uniformly over $(p,q) \in [\epsilon, 1-\epsilon] \times [\epsilon, 1-\epsilon]$ with $\epsilon$ being an arbitrary constant satisfying $0 < \epsilon < 1/2$.

## 4.1 The conditional regime probability

We first make the following two observations. (a) The expressions (6) and (7) can be combined to represent $\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2)$ recursively as $(t = 1, 2, ...)$:

$$\begin{aligned} & \xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2) \\ = \;& p + (p+q-1)\frac{f_t(\beta,\delta_2)(\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2)-1)}{f_t(\beta,\delta_1)\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2) + f_t(\beta,\delta_2)(1-\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2))}. \end{aligned} \tag{15}$$

This is a first order difference equation that relates $\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2)$ to $\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2)$. Immediately, this relationship implies that the derivatives of $\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2)$ with respect to $\beta, \delta_1, \delta_2$ must also follow first order difference equations. (b) Although these difference equations are nonlinear when evaluated at general values of $\delta_1$ and $\delta_2$, they simplify substantially once we let $\delta_1 = \delta_2$. Because the asymptotic expansions considered later are around the null parameter estimates, analyzing the latter case will be sufficient.

The next lemma contains the details on $\xi_{t+1|t}(p, q, \beta, \delta_1, \delta_2)$ and its derivatives evaluated at $\delta_1 = \delta_2 = \delta$, where $\delta$ represents an arbitrary value in $\Delta$. Define an augmented parameter vector

$$\theta = (\beta', \delta_1', \delta_2')' \tag{16}$$

and three sets of integers (they index the elements in $\beta, \delta_1$ and $\delta_2$, respectively)

$$I_0 = \{1, ..., n_\beta\}, I_1 = \{n_\beta + 1, ..., n_\beta + n_\delta\}, I_2 = \{n_\beta + n_\delta + 1, ..., n_\beta + 2n_\delta\}.$$

To ease the notation, let $\bar{\xi}_{t+1|t}$ and $\bar{f}_t$ denote $\xi_{t+1|t}(p, q, \beta, \delta_1, \delta_2)$ and $f_t(\beta, \delta_1)$ (or $f_t(\beta, \delta_2)$) evaluated at some $\beta$ and $\delta_1 = \delta_2 = \delta$. Also, let $\nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_k}} \bar{\xi}_{t|t-1}$, $\nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_k}} \bar{f}_{1t}$ and $\nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_k}} \bar{f}_{2t}$ denote the $k$-th order partial derivatives of $\xi_{t|t-1}(p, q, \beta, \delta_1, \delta_2)$, $f_t(\beta, \delta_1)$ and $f_t(\beta, \delta_2)$ with respect to the $j_1$-th,..., $j_k$-th elements of $\theta$ evaluated at some $\beta$ and $\delta_1 = \delta_2 = \delta$. The following relationships hold: $\nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_k}} \bar{f}_{1t} = \nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_k}} \bar{f}_{2t}$ if $j_1, ..., j_k$ all belong to $I_0$, $\nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_k}} \bar{f}_{1t} = 0$ if any of $j_1, ..., j_k$ belongs to $I_2$, and $\nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_k}} \bar{f}_{2t} = 0$ if any of $j_1, ..., j_k$ belongs to $I_1$.

**Lemma 1** *Let $\rho = p + q - 1$ and $r = \rho\xi_*(1 - \xi_*)$ with $\xi_*$ defined in (3). Then, for $t \geq 1$, we have, under $\delta_1 = \delta_2 = \delta$:*

1. $\bar{\xi}_{t+1|t} = \xi_*$.

2. $\nabla_{\theta_j} \bar{\xi}_{t+1|t} = \rho \nabla_{\theta_j} \bar{\xi}_{t|t-1} + \bar{\mathcal{E}}_{j,t}$, *where*

$$\bar{\mathcal{E}}_{j,t} = \begin{cases} 0 & \text{if } j \in I_0 \\ r\nabla_{\theta_j} \log \bar{f}_{1t} & \text{if } j \in I_1 \\ -r\nabla_{\theta_j} \log \bar{f}_{2t} & \text{if } j \in I_2 \end{cases}.$$

3. $\nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t+1|t} = \rho \nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t|t-1} + \bar{\mathcal{E}}_{jk,t}$, *where $\bar{\mathcal{E}}_{jk,t}$ is given by (with $(I_a, I_b)$ denoting a case with $j \in I_a$ and $k \in I_b$; $a, b = 0, 1, 2$):*

$$\begin{aligned}
&(I_0, I_0) : 0 \\
&(I_0, I_1) : -\frac{r\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t}\frac{\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t} + \frac{r\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t} \\
&(I_0, I_2) : \frac{r\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t}\frac{\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t} - \frac{r\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t} \\
&(I_1, I_1) : \frac{\rho(1-2\xi_*)\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t} + \frac{\rho(1-2\xi_*)\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t} + \frac{r\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t} - \frac{2r\xi_*\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t}\frac{\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t} \\
&(I_1, I_2) : \frac{\rho(2\xi_*-1)\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t} - \frac{\rho(2\xi_*-1)\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t} + \frac{r(2\xi_*-1)\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t}\frac{\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t} \\
&(I_2, I_2) : \frac{\rho(2\xi_*-1)\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t} + \frac{\rho(2\xi_*-1)\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t} - \frac{r\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t} - \frac{2r(\xi_*-1)\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t}\frac{\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t}
\end{aligned}.$$

*4.* $\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{\xi}_{t+1|t} = \rho\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{\xi}_{t|t-1} + \bar{\mathcal{E}}_{jkl,t}$*, where the expressions for* $\bar{\mathcal{E}}_{jkl,t}$ *with* $j,k,l \in \{I_a, I_b, I_c\}$ *and* $a,b,c = 0,1,2$ *are given in the appendix.*

**Remark 1** *The lemma holds for samples of any size. It shows that, when* $\delta_1 = \delta_2$*, the conditional regime probability* ($\bar{\xi}_{t+1|t}$) *equals the stationary probability* ($\xi_*$)*, while its derivatives up to third order all follow first order linear difference equations. The lagged coefficients always equal* $\rho = p + q - 1$*. Because* $0 < p, q < 1$*, these difference equations are always stable. As seen below, these features allow us to apply properties of first order linear systems to analyze the properties of the log likelihood. They are the key elements that make the subsequent analysis feasible.*

Now we take a closer look at the four results in the lemma. The first result is intuitive. Because the two regimes are identical when $\delta_1 = \delta_2$, observing the data brings no information about the regime probability. The second to fourth results reflect the first to third order effects of a unit change in the parameters' values on the regime probability. For the first order, changing the value of $\beta$ has no effect; $\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2)$ remains equal to $\xi_*$. Meanwhile, changing the values of $\delta_1$ and $\delta_2$ have exactly the opposite effects, i.e., $\nabla_{\theta_j}\bar{\xi}_{t+1|t} = -\nabla_{\theta_{j+n_\delta}}\bar{\xi}_{t+1|t}$ for any $j \in I_1$. The results concerning the second order derivatives have a similar structure. In particular, changing $\beta$ only has no effect, while changing $\delta_1$ and $\delta_2$ after a change in $\beta$ still have equal opposite effects, as indicated by the cases $(I_0, I_1)$ and $(I_0, I_2)$. The remaining three cases are more complex, but they all show that $\bar{\mathcal{E}}_{jk,t}$ depend only on $\nabla_{\theta_j}\bar{\xi}_{t|t-1}$ ($j \in I_1 \cup I_2$) and quantities related to the density functions. The third order effects consist of ten different cases corresponding to different combinations of $j, k$ and $l$. For the analysis later, the exact expressions of $\bar{\mathcal{E}}_{jkl,t}$ will be unimportant. What matters is that they depend only on lower order derivatives of $\bar{\xi}_{t|t-1}$ and quantities related to the density functions.

The recursive structure within the results, i.e., the higher order derivatives depend successively on the lower orders with the first order depending only on $\nabla_{\theta_j}\log\bar{f}_{1t}$ and $\nabla_{\theta_j}\log\bar{f}_{2t}$, suggests a strategy for analyzing their statistical properties. That is, we can start with the first order derivatives, and then use the results cumulatively to study the second order followed by the third order derivatives. Such a strategy is implemented in Lemma A.1 in the appendix.

Using $\xi_*$ as the initial value for $\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2)$ is not restrictive. With a generic finite initial value, the first result in the lemma becomes $\bar{\xi}_{t+1|t} = (1-q) + \rho\bar{\xi}_{t|t-1}$, while the other results also hold, but with $\xi_*$ and $r$ replaced by $\bar{\xi}_{t|t-1}$ and $\rho\bar{\xi}_{t|t-1}(1-\bar{\xi}_{t|t-1})$, respectively. Because $|\rho| < 1$, $\bar{\xi}_{t|t-1}$ converges at an exponential rate to $\xi_*$ as $t$ increases. Consequently, the first to the fourth order derivatives all converge to their counterparts in the lemma at an exponential rate. This fast rate of convergence implies that using a different finite initial value will not affect the asymptotic results presented later.

Below, we further illustrate the results in the lemma using the linear model (9).

**The illustrative model (cont'd).** Consider the general case where the regression coefficients and the error variance are both allowed to switch. Lemma 1.2 implies:

$$\nabla_\alpha \bar{\xi}_{t+1|t} = 0, \ \text{(w.r.t. the non-switching parameters)}$$

$$\left.\begin{aligned} \nabla_{\gamma_1}\bar{\xi}_{t+1|t} &= \rho\nabla_{\gamma_1}\bar{\xi}_{t|t-1} + r\tfrac{w_t}{\sigma^2}\left(y_t - z_t'\alpha - w_t'\gamma\right), \\ \nabla_{\sigma_1^2}\bar{\xi}_{t+1|t} &= \rho\nabla_{\sigma_1^2}\bar{\xi}_{t|t-1} + r\tfrac{1}{2\sigma^2}\left(\tfrac{(y_t - z_t'\alpha - w_t'\gamma)^2}{\sigma^2} - 1\right), \end{aligned}\right\} \text{(w.r.t. the parameters in the first regime)}$$

$$\left.\begin{aligned} \nabla_{\gamma_2}\bar{\xi}_{t+1|t} &= -\nabla_{\gamma_1}\bar{\xi}_{t+1|t}, \\ \nabla_{\sigma_2^2}\bar{\xi}_{t+1|t} &= -\nabla_{\sigma_1^2}\bar{\xi}_{t+1|t}. \end{aligned}\right\} \text{(w.r.t. the parameters in the second regime)}$$

When evaluated at the true parameter value, the derivatives with respect to $\gamma_1$ and $\sigma_1^2$ all reduce to stationary AR(1) processes with mean zero. Their variances are finite and satisfy (with $\sigma_*^2$ denoting the true value of $\sigma^2$ and $\nabla_{\gamma_{1j}}$ the first order derivative w.r.t. the $j$-th element of $\gamma_1$)

$$E\left(\nabla_{\gamma_{1j}}\bar{\xi}_{t+1|t}\right)^2 = \tfrac{r^2}{(1-\rho^2)\sigma_*^2}Ew_{jt}^2, \quad E(\nabla_{\sigma_1^2}\bar{\xi}_{t+1|t})^2 = \tfrac{r^2}{2(1-\rho^2)\sigma_*^4}.$$

The processes specified by Lemma 1.3-1.4, although more complex, also have finite means and variances when evaluated at the true parameter values, provided that the relevant moments of $w_t, z_t$ and $u_t$ exist. Such results for the general model (2) are established in Lemma A.1 in the appendix.■

## 4.2 Concentrated log likelihood and its expansion

To obtain an asymptotic approximation to the log likelihood ratio (11), a standard approach would be to expand $\mathcal{L}^A(p,q,\beta,\delta_1,\delta_2)$ around the null estimates $(\widetilde{\beta},\widetilde{\delta},\widetilde{\delta})$. However, this is infeasible here due to the complex dependence between the estimates of $\delta_1$ and $\delta_2$ as $\mathcal{L}^A(p,q,\beta,\delta_1,\delta_2)$ can have multiple local maxima. Cho and White (2007) encountered a similar problem and proceeded by working with the concentrated likelihood. We adopt their insightful strategy. This allows us to break the analysis into two steps. In the first step, we quantify the dependence between the estimates of $\delta_1$ and $\delta_2$ using the first order conditions that define the concentrated likelihood (see Lemma 2 below). This effectively removes $\beta$ and $\delta_1$ from the subsequent analysis. In the second step, we expand the concentrated likelihood around $\delta_2 = \widetilde{\delta}$ (see Lemma 3 below) and obtain an approximation to $LR(p, q)$.

Specifically, let $\hat{\beta}(\delta_2)$ and $\hat{\delta}_1(\delta_2)$ be the maximizer of the log likelihood for a given value $\delta_2 \in \Delta$ (here the dependence of $\hat{\beta}$ and $\hat{\delta}_1$ on $p$ and $q$ is suppressed to simplify the notation), i.e.,

$$(\hat{\beta}(\delta_2), \hat{\delta}_1(\delta_2)) = \arg\max_{\beta,\delta_1} \mathcal{L}^A(p, q, \beta, \delta_1, \delta_2). \tag{17}$$

Let $\mathcal{L}(p, q, \delta_2)$ denote the concentrated log likelihood, i.e., $\mathcal{L}(p, q, \delta_2) = \mathcal{L}^A(p, q, \hat{\beta}(\delta_2), \hat{\delta}_1(\delta_2), \delta_2)$. Then, the two terms in the likelihood ratio (11) satisfy: $\max_{\beta,\delta_1,\delta_2} \mathcal{L}^A(p, q, \beta, \delta_1, \delta_2) = \max_{\delta_2} \mathcal{L}(p, q, \delta_2)$ and $\mathcal{L}^N(\widetilde{\beta}, \widetilde{\delta}) = \mathcal{L}(p, q, \widetilde{\delta})$. Consequently:

$$LR(p, q) = 2 \max_{\delta_2} \left[\mathcal{L}(p, q, \delta_2) - \mathcal{L}(p, q, \widetilde{\delta})\right]. \tag{18}$$

For $k \geq 1$, let $\mathcal{L}^{(k)}_{i_1 \ldots i_k}(p, q, \delta_2)$ $(i_1, ..., i_k \in \{1, ..., n_\delta\})$ denote the $k$-th order derivative of $\mathcal{L}(p, q, \delta_2)$ with respect to the $(i_1, ..., i_k)$-th elements of $\delta_2$. Let $d_j$ $(j \in \{1, ..., n_\delta\})$ denote the $j$-th element of $(\delta_2 - \widetilde{\delta})$. Then, a fourth order Taylor expansion of $\mathcal{L}(p, q, \delta_2)$ around $\widetilde{\delta}$ is given by

$$\begin{aligned} \mathcal{L}(p, q, \delta_2) - \mathcal{L}(p, q, \widetilde{\delta}) \quad = \quad & \sum_{j=1}^{n_\delta} \mathcal{L}^{(1)}_j(p, q, \widetilde{\delta}) d_j + \frac{1}{2!} \sum_{j=1}^{n_\delta} \sum_{k=1}^{n_\delta} \mathcal{L}^{(2)}_{jk}(p, q, \widetilde{\delta}) d_j d_k \\ & + \frac{1}{3!} \sum_{j=1}^{n_\delta} \sum_{k=1}^{n_\delta} \sum_{l=1}^{n_\delta} \mathcal{L}^{(3)}_{jkl}(p, q, \widetilde{\delta}) d_j d_k d_l \\ & + \frac{1}{4!} \sum_{j=1}^{n_\delta} \sum_{k=1}^{n_\delta} \sum_{l=1}^{n_\delta} \sum_{m=1}^{n_\delta} \mathcal{L}^{(4)}_{jklm}(p, q, \bar{\delta}) d_j d_k d_l d_m, \end{aligned} \tag{19}$$

where in the last term $\bar{\delta}$ is a value that lies between $\delta_2$ and $\widetilde{\delta}$.

Below, we provide two lemmas to analyze this expansion. The next assumption is needed in order to apply a law of large numbers and a central limit theorem to the terms in (19). It is similar to but slightly stronger than Assumption A5 (iii) in Cho and White (2007). There, instead of $\alpha(k)/k$, the respective values are $4, 2, 2$ and $1$ for $k = 1, 2, 3$ and $4$.

**Assumption 4** *There exists an open neighborhood of $(\beta_*, \delta_*)$, denoted by $B(\beta_*, \delta_*)$, and a sequence of positive, strictly stationary and ergodic random variables $\{\upsilon_t\}$ satisfying $E\upsilon_t^{1+c} < L < \infty$ for some $c > 0$, such that*

$$\sup_{(\beta,\delta_1) \in B(\beta_*,\delta_*)} \left| \frac{\nabla_{\theta_{i_1}} ... \nabla_{\theta_{i_k}} f_t(\beta, \delta_1)}{f_t(\beta, \delta_1)} \right|^{\frac{\alpha(k)}{k}} < \upsilon_t$$

*for all $i_1, ..., i_k \in \{1, ..., n_\beta + n_\delta\}$, where $1 \leq k \leq 5$; $\alpha(k) = 6$ if $k = 1, 2, 3$ and $\alpha(k) = 5$ if $k = 4, 5$.*

The next lemma characterizes the derivatives of $\hat{\beta}(\delta_2)$ and $\hat{\delta}_1(\delta_2)$ with respect to $\delta_2$ evaluated at $\delta_2 = \widetilde{\delta}$. To shorten the expressions, let $\widetilde{\xi}_{t+1|t}$ and $\widetilde{f}_t$ denote $\xi_{t+1|t}(p, q, \beta, \delta_1, \delta_2)$ and $f_t(\beta, \delta_1)$ evaluated at $(\beta, \delta_1, \delta_2) = (\widetilde{\beta}, \widetilde{\delta}, \widetilde{\delta})$. Also, let $\nabla_{\delta_{1i_1}} ... \nabla_{\delta_{1i_k}} \widetilde{\xi}_{t|t-1}$ and $\nabla_{\delta_{1i_1}} ... \nabla_{\delta_{1i_k}} \widetilde{f}_{1t}$ denote the $k$-th order derivative of $\xi_{t+1|t}(p, q, \beta, \delta_1, \delta_2)$ and $f_t(\beta, \delta_1)$ with respect to the $i_1$-th, ..., $i_k$-th elements of $\delta_1$ evaluated at $(\beta, \delta_1, \delta_2) = (\widetilde{\beta}, \widetilde{\delta}, \widetilde{\delta})$. Finally, define

$$\begin{aligned}
\widetilde{U}_{jk,t} &= \frac{1}{\widetilde{f}_t} \left\{ \left( \frac{1-\xi_*}{\xi_*} \right) \nabla_{\delta_{1j}} \nabla_{\delta_{1k}} \widetilde{f}_{1t} + \frac{1}{\xi_*^2} \nabla_{\delta_{1j}} \widetilde{\xi}_{t|t-1} \nabla_{\delta_{1k}} \widetilde{f}_{1t} + \frac{1}{\xi_*^2} \nabla_{\delta_{1j}} \widetilde{f}_{1t} \nabla_{\delta_{1k}} \widetilde{\xi}_{t|t-1} \right\}, \qquad (20) \\
\widetilde{D}_{jk,t} &= \frac{\nabla_{(\beta', \delta_1')'} \widetilde{f}_{1t}}{\widetilde{f}_t} \widetilde{U}_{jk,t}, \quad \widetilde{I}_t = \frac{\nabla_{(\beta', \delta_1')'} \widetilde{f}_{1t}}{\widetilde{f}_t} \frac{\nabla_{(\beta', \delta_1')} \widetilde{f}_{1t}}{\widetilde{f}_t}, \\
\widetilde{V}_{jklm} &= T^{-1} \sum_{t=1}^{T} \widetilde{U}_{jk,t} \widetilde{U}_{lm,t}, \quad \widetilde{D}_{lm} = T^{-1} \sum_{t=1}^{T} \widetilde{D}_{lm,t}, \quad \widetilde{I} = T^{-1} \sum_{t=1}^{T} \widetilde{I}_t.
\end{aligned}$$

Note that $\widetilde{U}_{jk,t}$ involves the first and second order derivatives with respect to the $j$-th and $k$-th elements of $\delta_1$. The term inside the curly brackets can also be represented as $((1-\xi_*)/\xi_*) \nabla_{\delta_{2j}} \nabla_{\delta_{2k}} \widetilde{f}_{2t} - (1/\xi_*^2) \nabla_{\delta_{2j}} \widetilde{\xi}_{t|t-1} \nabla_{\delta_{2k}} \widetilde{f}_{2t} - (1/\xi_*^2) \nabla_{\delta_{2j}} \widetilde{f}_{2t} \nabla_{\delta_{2k}} \widetilde{\xi}_{t|t-1}$. As will be seen, $\widetilde{U}_{jk,t}$ determines $\mathcal{L}_{jk}^{(2)}(p, q, \widetilde{\delta})$ while $\widetilde{D}_{jk,t}$ and $\widetilde{I}_t$ appear within $\mathcal{L}_{jklm}^{(4)}(p, q, \widetilde{\delta})$.

**Lemma 2** *Under the null hypothesis and Assumptions 1-4, for all $k, l, m \in \{1, ..., n_\delta\}$, we have:*

1. *Let $e_k$ be an $n_\delta$-dimensional unit vector whose $k$-th element equals 1, then*

$$\begin{bmatrix} \nabla_{\delta_{2k}} \hat{\beta}(\widetilde{\delta}) \\ \xi_* \nabla_{\delta_{2k}} \hat{\delta}_1(\widetilde{\delta}) \end{bmatrix} = (\xi_* - 1) \begin{bmatrix} 0 \\ e_k \end{bmatrix} + O_p(T^{-1/2}).$$

2. *The second order derivatives satisfy*

$$\begin{bmatrix} \nabla_{\delta_{2k}} \nabla_{\delta_{2l}} \hat{\beta}(\widetilde{\delta}) \\ \xi_* \nabla_{\delta_{2k}} \nabla_{\delta_{2l}} \hat{\delta}_1(\widetilde{\delta}) \end{bmatrix} = -\widetilde{I}^{-1} \frac{1}{T} \sum_{t=1}^{T} \widetilde{D}_{kl,t} + O_p(T^{-1/2}).$$

3. *The third order derivatives satisfy*

$$\begin{bmatrix} \nabla_{\delta_{2k}} \nabla_{\delta_{2l}} \nabla_{\delta_{2m}} \hat{\beta}(\widetilde{\delta}) \\ \xi_* \nabla_{\delta_{2k}} \nabla_{\delta_{2l}} \nabla_{\delta_{2m}} \hat{\delta}_1(\widetilde{\delta}) \end{bmatrix} = O_p(1).$$

The above results generalize Lemma B2(a)-(d) in Cho and White (2007) to Markov switching models and will be used as inputs to analyze $\mathcal{L}_{i_1, ..., i_k}^{(k)}(p, q, \delta_2)$ in the expansion (19). They show how the parameters $\delta_1$ and $\beta$ need to change in order to maximize the likelihood when $\delta_2$ is moved away from $\widetilde{\delta}$. Specifically, consider a unit change in the $j$-th element of $\delta_2$. Then, Lemma 2.1 shows that, in the first order, $\hat{\delta}_{1j}(\delta_2)$ will change by $(\xi_* - 1)/\xi_* + O_p(T^{-1/2})$, while all the other elements of $\hat{\delta}_1(\delta_2)$ and $\hat{\beta}(\delta_2)$ will only change by a factor of order $O_p(T^{-1/2})$. Interestingly, the quantity $(1 - \xi_*)/\xi_*$ is simply the stationary probability of the second regime divided by the first. Lemma 2.2 pertains to changes in the second order. There, the time variation in the conditional regime probability enters explicitly. Finally, the expression for the third order derivative is not needed for the limiting distribution and therefore omitted.

**Assumption 5** *There exists $\eta > 0$, such that $\sup_{p,q\in[\epsilon,1-\epsilon]} \sup_{|\delta-\widetilde{\delta}|<\eta} T^{-1}|\mathcal{L}^{(5)}_{jklmn}(p, q, \delta)| = O_p(1)$ for all $j, k, l, m, n \in \{1, ..., n_\delta\}$, where $\epsilon$ is an arbitrary small constant satisfying $0 < \epsilon < 1/2$.*

In a standard problem, we would need the second order derivative $\mathcal{L}^{(2)}_{jk}(p, q, \delta)$ to be continuous in $\delta$ (e.g., Amemiya, 1985, p.111), or the third order derivative $T^{-1}\mathcal{L}^{(3)}_{jkl}(p, q, \delta)$ to be $O_p(1)$ to ensure that a local quadratic expansion is an adequate approximation to the log likelihood. Here, $\mathcal{L}^{(4)}_{jklm}(p, q, \delta)$ plays the same role as the second order derivative in a standard problem. This is why the above assumption on the fifth order derivative is needed. The next lemma provides the leading terms of $\mathcal{L}^{(k)}_{i_1,\dots,i_k}(p, q, \delta)$ $(k = 1, 2, 3, 4)$ in the expansion (19).

**Lemma 3** *Under the null hypothesis and Assumptions 1-5, for all $j, k, l, m \in \{1, ..., n_\delta\}$, we have*

1. $\mathcal{L}^{(1)}_j(p, q, \widetilde{\delta}) = 0.$
2. $T^{-1/2}\mathcal{L}^{(2)}_{jk}(p, q, \widetilde{\delta}) = T^{-1/2}\sum_{t=1}^T \widetilde{U}_{jk,t} + o_p(1).$
3. $T^{-3/4}\mathcal{L}^{(3)}_{jkl}(p, q, \widetilde{\delta}) = O_p\left(T^{-1/4}\right).$
4. $T^{-1}\mathcal{L}^{(4)}_{jklm}(p, q, \bar{\delta}) = -\{\widetilde{V}_{jklm} - \widetilde{D}'_{jk}\widetilde{I}^{-1}\widetilde{D}_{lm} + \widetilde{V}_{jmkl} - \widetilde{D}'_{jm}\widetilde{I}^{-1}\widetilde{D}_{kl} + \widetilde{V}_{jlkm} - \widetilde{D}'_{jl}\widetilde{I}^{-1}\widetilde{D}_{km}\} + o_p(1).$

The first order derivative $\mathcal{L}^{(1)}_j(p, q, \widetilde{\delta})$ is identically zero for any sample size. Consequently, the MLE of $\delta_2$ will converge at a rate slower than $T^{-1/2}$. The second order derivative $\mathcal{L}^{(2)}_{jk}(p, q, \widetilde{\delta})$ is of order $O_p(T^{1/2})$, rather than $O_p(T)$. As seen below, its leading term $T^{-1/2}\sum_{t=1}^T \widetilde{U}_{jk,t}$ converges to a multivariate normal distribution, whose property depends explicitly on the time varying conditional regime probability. The third order derivative $\mathcal{L}^{(3)}_{jkl}(p, q, \widetilde{\delta})$ is also of order $O_p(T^{1/2})$. The expression of its leading term is not needed to derive the limiting distribution, but we will further analyze it when providing a finite sample refinement. Finally, the fourth order derivative $\mathcal{L}^{(4)}_{jklm}(p, q, \bar{\delta})$ is of order $O_p(T)$. Its leading term provides a consistent estimator of the asymptotic variance of $T^{-1/2}\sum_{t=1}^T \widetilde{U}_{jk,t}$.

**Remark 2** *The first component of $\widetilde{U}_{jk,t}$, $((1-\xi_*)/\xi_*)\,\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\widetilde{f}_{1t}/\widetilde{f}_t$, is also present when testing against mixture alternatives; see Cho and White (2007, Lemma 2(a)). It is sometimes called the dispersion score; see Lindsay (1995, p.71). The remaining two components are new and are due to the Markov switching structure. They can be rewritten as $((1-\xi_*)/\xi_*)\sum_{s=1}^{t-1}\rho^s(\nabla_{\delta_{1j}}\log\widetilde{f}_{1(t-s)})(\nabla_{\delta_{1k}}\log\widetilde{f}_{1t})$ and $((1-\xi_*)/\xi_*)\sum_{s=1}^{t-1}\rho^s(\nabla_{\delta_{1k}}\log\widetilde{f}_{1(t-s)})(\nabla_{\delta_{1j}}\log\widetilde{f}_{1t})$. Intuitively, among the three components, the first picks up overdispersion, while the remaining two pick up serial dependence introduced by the Markov regimes. Further, the magnitudes of the latter two components become more pronounced relative to the first as $\rho$ approaches 1. This follows because the first component is independent of $\rho$ after division by $((1-\xi_*)/\xi_*)$ while the latter two components involve weights $\rho^s$. This feature suggests that the power difference between testing against Markov switching alternatives and mixture alternatives can be substantial when the regimes are persistent, i.e., when $\rho$ is close to 1. This is confirmed by the simulations reported later.*

**The illustrative model (cont'd).** We illustrate the leading terms of $T^{-1/2}\mathcal{L}^{(2)}_{jk}(p,q,\widetilde{\delta})$ and $T^{-1}\mathcal{L}^{(4)}_{jklm}(p,q,\widetilde{\delta})$ in Lemma 3 using (9). Suppose only $\gamma$ is allowed to switch. Then, $\widetilde{U}_{jk,t}$ and $\widetilde{D}_{jk,t}$ are, respectively,

$$\left(\tfrac{1-\xi_*}{\xi_*}\right)\left\{\tfrac{w_{jt}w_{kt}}{\widetilde{\sigma}^2}\left(\tfrac{\widetilde{u}_t^2}{\widetilde{\sigma}^2}-1\right)+\sum_{s=1}^{t-1}\rho^s\left(\tfrac{w_{j(t-s)}\widetilde{u}_{t-s}}{\widetilde{\sigma}^2}\right)\left(\tfrac{w_{kt}\widetilde{u}_t}{\widetilde{\sigma}^2}\right)+\sum_{s=1}^{t-1}\rho^s\left(\tfrac{w_{k(t-s)}\widetilde{u}_{t-s}}{\widetilde{\sigma}^2}\right)\left(\tfrac{w_{jt}\widetilde{u}_t}{\widetilde{\sigma}^2}\right)\right\} \tag{21}$$

and

$$\begin{bmatrix}\tfrac{z_t'\widetilde{u}_t}{\widetilde{\sigma}^2} & \tfrac{1}{2\widetilde{\sigma}^2}\left(\tfrac{\widetilde{u}_t^2}{\widetilde{\sigma}^2}-1\right) & \tfrac{w_t'\widetilde{u}_t}{\widetilde{\sigma}^2}\end{bmatrix}'\widetilde{U}_{jk,t}, \tag{22}$$

where $\widetilde{u}_t$ denote the residuals under the null and $\widetilde{\sigma}^2=T^{-1}\sum_{t=1}^T\widetilde{u}_t^2$. These two expressions show that $\widetilde{U}_{jk,t}$ and $\widetilde{D}_{jk,t}$ depend only on the regressors and the estimates under the null hypothesis. This makes the covariance function of $T^{-1/2}\sum_{t=1}^T\widetilde{U}_{jk,t}$, and therefore of $T^{-1/2}\mathcal{L}^{(2)}_{jk}(p,q,\widetilde{\delta})$, consistently estimable. This feature will be used when deriving the relevant critical values. ■

# 5 Asymptotic approximations

Let $\mathcal{L}^{(2)}(p,q,\widetilde{\delta})$ be a square matrix whose $(j,k)$-th element is given by $\mathcal{L}^{(2)}_{jk}(p,q,\widetilde{\delta})$ for $j,k\in\{1,2,...,n_\delta\}$. This section consists of four sets of results. (1) It establishes the weak convergence of $T^{-1/2}\mathcal{L}^{(2)}(p,q,\widetilde{\delta})$ over $\epsilon\le p,q\le 1-\epsilon$. (2) It obtains the limiting distribution of $\mathrm{SupLR}(\Lambda_\epsilon)$. (3) It develops a finite sample refinement that improves the asymptotic approximation when a singularity is present. (4) It develops an algorithm to obtain the relevant critical values.

### 5.1 Weak convergence of $\mathcal{L}^{(2)}(p,q,\widetilde{\delta})$

For $0 < p_r, q_r, p_s, q_s < 1$ and $j,k,l,m \in \{1,2,...,n_\delta\}$, define

$$\omega_{jklm}(p_r,q_r;p_s,q_s) = V_{jklm}(p_r,q_r;p_s,q_s) - D'_{jk}(p_r,q_r)I^{-1}D_{lm}(p_s,q_s), \tag{23}$$

where $V_{jklm}(p_r,q_r;p_s,q_s) = E\left[U_{jk,t}(p_r,q_r)U_{lm,t}(p_s,q_s)\right], D_{jk}(p_r,q_r) = ED_{jk,t}(p_r,q_r)$, and $I = EI_t$. Here, $U_{jk,t}(p_r,q_r)$, $D_{jk,t}(p_r,q_r)$ and $I_t$ are defined as $\widetilde{U}_{jk,t}$, $\widetilde{D}_{jk,t}$ and $\widetilde{I}_t$ in (20) but evaluated at $(p_r,q_r,\beta_*,\delta_*)$ instead of $(p_r,q_r,\widetilde{\beta},\widetilde{\delta})$.

**Proposition 1** *Under the null hypothesis and Assumptions 1-5, we have, over $\epsilon \le p,q \le 1-\epsilon$ :*

$$T^{-1/2}\mathcal{L}^{(2)}(p,q,\widetilde{\delta}) \Rightarrow G(p,q),$$

*where the elements of $G(p,q)$ are mean zero continuous Gaussian processes satisfying* $\mathrm{Cov}[G_{jk}(p_r,q_r), G_{lm}(p_s,q_s)]{=}\omega_{jklm}(p_r,q_r;p_s,q_s)$ *for $j$,$k$,$l$,$m \in \{1,2,...,n_\delta\}$, where $\omega_{jklm}(p_r,q_r;p_s,q_s)$ is given by (23).*

In the appendix, the result is proved by first showing the finite dimensional convergence and then the stochastic equicontinuity.

Below, we illustrate some important features of $\omega_{jklm}(p_r,q_r;p_s,q_s)$. It will emerge that this function depends on: (1) the model's dynamic properties (e.g., whether the regressors are strictly exogenous or predetermined), (2) which parameters are allowed to switch (e.g., regressions coefficients or the variance of the errors), and (3) whether nuisance parameters are present. Consequently, to make the test easy to apply in practice, we will need a procedure that can adapt to these features to obtain critical values without requiring laborious derivations from the practitioner. Such a procedure is developed in Section 5.4.

**The illustrative model (cont'd).** We consider a simpler version of (9) for which the covariance function $\omega_{jklm}(p_r,q_r;p_s,q_s)$ can be computed analytically:

$$y_t = w_t\gamma_1\mathbf{1}_{\{s_t=1\}} + w_t\gamma_2\mathbf{1}_{\{s_t=2\}} + u_t,$$

where $u_t \sim$i.i.d.$N(0,\sigma_*^2)$, and $w_t$ is a scalar regressor that is either strictly exogenous (e.g., a constant) or equals $y_{t-1}$. Define $\rho_r = p_r + q_r - 1$ and $\rho_s = p_s + q_s - 1$. The subscript "*" continues to denote the true parameter value.

First, we allow $\gamma$ to switch, while assuming $\sigma_*^2$ is unknown but remains constant across the regimes. Then, in the strictly exogenous regressor case, the covariance function (23) equals

$$\frac{2(1-p_r)(1-p_s)}{(1-q_r)(1-q_s)}\frac{\left[Var(w_t^2) + 2\sum_{k=1}^{\infty}(\rho_r\rho_s)^k E(w_t^2w_{t-k}^2)\right]}{\sigma_*^4}. \tag{24}$$

In the lagged dependent variable case, it equals

$$\frac{(1-p_r)(1-p_s)}{(1-q_r)(1-q_s)}\left(\frac{1}{1-\gamma_*^2}\right)\left\{\frac{4}{1-\gamma_*^2}+\frac{4\rho_r\rho_s}{1-\gamma_*^2}\left(\frac{2}{1-\rho_r\rho_s\gamma_*^2}+\frac{1}{1-\rho_r\rho_s}\right)\right. \tag{25}$$
$$\left.+\frac{16\rho_r^2\rho_s\gamma_*^2}{(1-\rho_r\gamma_*^2)\,(1-\rho_r\rho_s\gamma_*^2)}+\frac{16\rho_r\rho_s^2\gamma_*^2}{(1-\rho_s\gamma_*^2)\,(1-\rho_r\rho_s\gamma_*^2)}-\frac{16\rho_r\rho_s\gamma_*^2}{(1-\rho_r\gamma_*^2)(1-\rho_s\gamma_*^2)}\right\}.$$

These two functions are different even when $w_t \sim i.i.d.N(0,1)$ and $\gamma_* = 0$. This follows because $\nabla_{\gamma_1}\xi_{t|t-1}$ is independent of $\nabla_{\gamma_1}f_{1t}$ when $w_t$ is strictly exogenous, but not when it is only predetermined. This comparison shows that the covariance function is affected by the dynamic properties of the model.

Now, we consider the same situation as above but assume $\sigma_*^2$ is known. Then, in the strictly exogenous regressor case, the covariance function equals

$$\frac{2(1-p_r)(1-p_s)}{(1-q_r)(1-q_s)}\frac{\left[E(w_t^4)+2\sum_{k=1}^{\infty}(\rho_r\rho_s)^k E(w_t^2w_{t-k}^2)\right]}{\sigma_*^4}, \tag{26}$$

while in the lagged dependent variable case, it equals

$$\frac{(1-p_r)(1-p_s)}{(1-q_r)(1-q_s)}\left(\frac{1}{1-\gamma_*^2}\right)\left\{\frac{6}{1-\gamma_*^2}+\frac{4\rho_r\rho_s}{1-\gamma_*^2}\left(\frac{2}{1-\rho_r\rho_s\gamma_*^2}+\frac{1}{1-\rho_r\rho_s}\right)\right. \tag{27}$$
$$\left.+\frac{16\rho_r^2\rho_s\gamma_*^2}{(1-\rho_r\gamma_*^2)\,(1-\rho_r\rho_s\gamma_*^2)}+\frac{16\rho_r\rho_s^2\gamma_*^2}{(1-\rho_s\gamma_*^2)\,(1-\rho_r\rho_s\gamma_*^2)}-\frac{16\rho_r\rho_s\gamma_*^2}{(1-\rho_r\gamma_*^2)(1-\rho_s\gamma_*^2)}\right\}.$$

These two functions are different from (24) and (25). This shows that the presence of nuisance parameters can also alter the covariance function.

Next, we allow $\sigma_*^2$ to switch but require the unknown regression coefficient $\gamma_*$ to remain constant across the regimes. Then, irrespective of whether $w_t$ is strictly exogenous, we have

$$\mathrm{Cov}\left(G\left(p_r,q_r\right),G\left(p_s,q_s\right)\right)=\frac{(1-p_r)(1-p_s)}{(1-q_r)(1-q_s)}\frac{1}{\sigma_*^8}\left\{\frac{3}{2}+\left(\frac{\rho_r\rho_s}{1-\rho_r\rho_s}\right)\right\}. \tag{28}$$

This function is different from both (24) and (25). Thus, even after conditioning on the model, the covariance function can still be different depending on which parameter is allowed to switch.

We report some simulation results to complement the above analysis. The parameter values are $\gamma_* = 0.5$ and $\sigma_*^2 = 1$. In the strictly exogenous regressor case, $w_t$ is generated as being independent of $u_s$ at all leads and lags by $w_t = 0.5w_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim i.i.d.N(0,1)$. This ensures that the regressors follow the same DGP in both cases. Further, we let $(p_r, q_r) = (0.6, 0.9)$ and $(p_s, q_s) = (0.6, x)$ with $x$ varying between 0.1 and 0.9. Figure 1 reports the five correlations functions given by (24)-(27) (Here, correlations instead of covariances are plotted to ease comparisons).

The solid lines starting from the top correspond to (28), (26), (24), (27) and (25), respectively. These functions demonstrate clearly the dependence on the three factors highlighted above. Also included in the figure are correlations computed from simulations (i.e., the dashed lines). They are generated by simulating samples of 250 observations using the same parameter value as above, computing $T^{-1/2}\sum_{t=1}^{T}\widetilde{U}_{jk,t}$ using each series and then repeating 10,000 times to obtain the empirical correlations. The values are close to their asymptotic approximations in all five cases. ■

## 5.2 Limiting distribution of $\mathrm{SupLR}(\Lambda_\epsilon)$

Let $\Omega(p,q)$ be an $n_\delta^2$-dimensional square matrix whose $(j+(k-1)n_\delta, l+(m-1)n_\delta)$-th element is given by $\omega_{jklm}(p,q;p,q)$. Then, Proposition 1 implies $E[\mathrm{vec}\, G(p,q)\,\mathrm{vec}\, G(p,q)'] = \Omega(p,q)$. The next result gives the asymptotic distribution of $\mathrm{SupLR}(\Lambda_\epsilon)$.

**Proposition 2** *Suppose the null hypothesis and Assumptions 1-5 hold. Then:*

$$\mathrm{SupLR}(\Lambda_\epsilon) \Rightarrow \sup_{(p,q)\in\Lambda_\epsilon}\ \sup_{\eta\in R^{n_\delta}} \mathcal{W}^{(2)}(p,q,\eta), \tag{29}$$

*where $\Lambda_\epsilon$ is given by (12) and*

$$\mathcal{W}^{(2)}(p,q,\eta) = \left(\eta^{\otimes 2}\right)' \mathrm{vec}\, G(p,q) - \frac{1}{4}\left(\eta^{\otimes 2}\right)'\Omega(p,q)\left(\eta^{\otimes 2}\right).$$

The quantity $\eta$ plays the role of $T^{-1/4}(\delta_2-\widetilde{\delta})$ in (19). Its dimension is unaffected by the presence of nuisance parameters. If $n_\delta = 1$, then the optimization over $\eta$ can be solved analytically, leading to $\mathrm{SupLR}(\Lambda_\epsilon) \Rightarrow \max[0, \sup_{(p,q)\in\Lambda_\epsilon} G(p,q)/\sqrt{\Omega(p,q)}]^2$. The right hand side can equal zero with positive probability, with the value of the latter depending on the covariance function of $G(p,q)/\sqrt{\Omega(p,q)}$ over $(p,q)$. If $n_\delta > 1$, the optimization will need to be carried out numerically. However, because $\mathcal{W}^{(2)}(p,q,\eta)$ is a quadratic function of $\eta^{\otimes 2}$, the optimization remains relatively standard.

Below, we illustrate the above limiting distribution and also examine its adequacy in finite samples. The illustration also suggests the desirability for a finite sample refinement when a further singularity is present.

**The illustrative model (cont'd).** We consider the following special case of (9):

$$y_t = \mu + \alpha y_{t-1} + u_t, \tag{30}$$

where $u_t \sim i.i.d.\ N(0, \sigma^2)$ and $\mu, \alpha$ and $\sigma^2$ are unknown. As shown below, the distribution of $\text{SupLR}(\Lambda_\epsilon)$, as well as the adequacy of the asymptotic approximation, can differ substantially depending on whether $\mu$ or $\alpha$ is allowed to switch.

Figure 2 summarizes the finite sample and asymptotic distributions of $\text{SupLR}(\Lambda_\epsilon)$ for testing regime switching in $\mu$ only or $\alpha$ only. We consider $\mu = 0, \alpha = 0.5$ and $\sigma^2 = 1$. The set $\Lambda_\epsilon$ is specified as (13) with $\epsilon = 0.05$, The sample size is 250 and all results are based on 5000 replications. Two features emerge. First, the finite sample (the solid lines) and asymptotic (the long dashed lines) distributions are both quite different between the two cases. This is consistent with the covariance function of $T^{-1/2} \sum_{t=1}^T \widetilde{U}_{jk,t}$ being parameter dependent; see the illustration in Section 5.1. Second, the asymptotic distribution provides an adequate approximation when testing for switching in $\alpha$, but not when testing for switching in $\mu$. For the latter case, the asymptotic distribution falls to the left of the finite sample distribution. The structure of $T^{-1/2} \sum_{t=1}^T \widetilde{U}_{jk,t}$ is informative about the second feature. It is given by:

$$\text{For } \mu \text{ switching: } \frac{1}{\widetilde{\sigma}^2} \left( \frac{1-\xi_*}{\xi_*} \right) \left\{ T^{-1/2} \sum_{t=1}^T \left( \frac{\widetilde{u}_t^2}{\widetilde{\sigma}^2} - 1 \right) + 2T^{-1/2} \sum_{t=1}^T \left( \sum_{s=1}^{t-1} \rho^s \frac{\widetilde{u}_{t-s}}{\widetilde{\sigma}} \frac{\widetilde{u}_t}{\widetilde{\sigma}} \right) \right\},$$

$$\text{For } \alpha \text{ switching: } \frac{1}{\widetilde{\sigma}^2} \left( \frac{1-\xi_*}{\xi_*} \right) \left\{ T^{-1/2} \sum_{t=1}^T \left( \frac{\widetilde{u}_t^2}{\widetilde{\sigma}^2} - 1 \right) y_{t-1}^2 + 2T^{-1/2} \sum_{t=1}^T \left( \sum_{s=1}^{t-1} \rho^s \frac{y_{t-s-1}\widetilde{u}_{t-s}}{\widetilde{\sigma}} \frac{y_{t-1}\widetilde{u}_t}{\widetilde{\sigma}} \right) \right\},$$

where $\widetilde{u}_t$ denote the residuals under the null and $\widetilde{\sigma}^2 = T^{-1} \sum_{t=1}^T \widetilde{u}_t^2$. When testing for switching in $\mu$, the first term in the curly brackets is in fact identically zero for any sample size. Also, the magnitude of the second term decreases as $\rho$ approaches 0, i.e., as $p+q$ approaches 1. Consequently, in finite samples, the magnitude of $T^{-1/2} \sum_{t=1}^T \widetilde{U}_{jk,t}$ can be too small to dominate the higher order terms in the likelihood expansion. This explains why the asymptotic distribution that relies entirely on $T^{-1/2} \sum_{t=1}^T \widetilde{U}_{jk,t}$ can be inadequate. In contrast, when testing for switching in $\alpha$, the first term in the curly brackets converges to a normal distribution that is independent of $p$ and $q$. Therefore, the issues discussed do not arise.

Figure 3 provides some further information by comparing the finite sample and asymptotic distributions of $LR(p, q)$ when testing for $\mu$ switching at some selected values of $(p, q)$ that equal $(0.90, 0.90), (0.90, 0.75)$ and $(0.90, 0.60)$. Consistent with the discussion above, a gap between the finite sample distribution (the solid line) and the asymptotic distribution (the long dashed line) opens up and grows wider as $p + q$ approaches 1. We have also found in unreported simulations, that when testing for $\alpha$ switching, these two distributions remain close to each other in all three cases. ■

In summary, the illustration suggests that the asymptotic approximation in Proposition 2 needs to be improved if the hypotheses imply that $\mathcal{L}^{(2)}(p,q,\widetilde{\delta})$ equals zero when $p+q=1$. When testing for switching in the intercept in the illustrative model, the latter arises because the following linear relationship holds for all $t$: $\nabla_{\sigma^2} f_{1t}(\widetilde{\mu},\widetilde{\alpha},\widetilde{\sigma}^2) = 0.5\nabla_\mu\nabla_\mu f_{1t}(\widetilde{\mu},\widetilde{\alpha},\widetilde{\sigma}^2)$.

### 5.3 A refinement

This section derives a sixth order expansion of the likelihood ratio along $p+q=1$ and an eighth order expansion at $p=q=1/2$. (The reason for why the latter is needed is explained below.) The respective leading terms are then incorporated into the limiting distribution to deliver a refined approximation. These expansions are derived under the following assumption.

**Assumption 6** *The following linear relationship holds for all $t$ and all $i_1,i_2\in\{1,...,n_\delta\}$ :*

$$\nabla_{\delta_{1i_1}}\nabla_{\delta_{1i_2}}\widetilde{f}_{1t} = \alpha_{i_1i_2}^{(1)\prime}\nabla_\beta\widetilde{f}_{1t} + \alpha_{i_1i_2}^{(2)\prime}\nabla_{\delta_1}\widetilde{f}_{1t}, \tag{31}$$

*where $\alpha_{i_1i_2}^{(1)}$ and $\alpha_{i_1i_2}^{(2)}$ are $n_\beta$ and $n_\delta$ dimensional known vectors of constants.*

Whether or not this assumption holds is known once the model and the hypotheses are specified. The following assumption strengthens Assumption 4. It is similar to A.5 (iv) in Cho and White (2007). The subsequent analysis makes heavy use of their results developed in Section 2.3.2.

**Assumption 7** *There exists an open neighborhood of $(\beta_*,\delta_*)$, $B(\beta_*,\delta_*)$, and a sequence of positive, strictly stationary and ergodic random variables $\{\upsilon_t\}$ satisfying $E\upsilon_t^{1+c}<\infty$ for some $c>0$, such that the supremums of the following quantities over $B(\beta_*,\delta_*)$ are bounded from the above by $\upsilon_t$: $\left|\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}} f_t(\beta,\delta_1)/f_t(\beta,\delta_1)\right|^4$, $\left|\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_m}} f_t(\beta,\delta_1)/f_t(\beta,\delta_1)\right|^2$, $\left|\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_8}} f_t(\beta,\delta_1)/f_t(\beta,\delta_1)\right|$, $\left|\nabla_{\theta_{j_1}}\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_7}} f_t(\beta,\delta_1)/f_t(\beta,\delta_1)\right|$, $\left|\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_6}} f_t(\beta,\delta_1)/f_t(\beta,\delta_1)\right|$, where $k=1,2,3,4$, $m=5,6,7$, $i_1,...,i_7\in\{1,...,n_\beta+n_\delta\}$ and $j_1,j_2\in\{1,...,n_\beta\}$.*

Before proceeding, we first establish some notation. To approximate the third and sixth order derivatives of the concentrated log likelihood, define (the symbol $\widetilde{s}$ stands for skewness)

$$\widetilde{s}_{jkl,t}(p,q) = \frac{(1-p)(p-q)}{(1-q)^2}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t} \tag{32}$$

and let $G_{jkl}^{(3)}(p,q)$ be a continuous Gaussian process with mean zero satisfying

$$\begin{aligned}
&\omega_{jklmnu}^{(3)}(p_r,q_r;p_s,q_s)\\
=\;& \mathrm{Cov}(G_{jkl}^{(3)}(p_r,q_r), G_{mnu}^{(3)}(p_s,q_s))\\
=\;& E\left[s_{jkl,t}(p_r,q_r)s_{mnu,t}(p_s,q_s)\right] - E\left[\frac{\nabla_{(\beta',\delta_1')}f_{1t}}{f_t}s_{jkl,t}(p_r,q_r)\right]I^{-1}\left[\frac{\nabla_{(\beta',\delta_1')'}f_{1t}}{f_t}s_{mnu,t}(p_s,q_s)\right],
\end{aligned}$$

where $s_{jkl,t}(p,q)$ is the same as $\widetilde{s}_{jkl,t}(p,q)$ but evaluated at the true parameter values (the other quantities are also evaluated at the true parameter values). To approximate the fourth and eighth order derivatives, define (the symbol $\widetilde{k}$ stands for kurtosis.)

$$\begin{aligned}\widetilde{k}_{jklm,t}(p,q) &= \frac{1-p}{2-p-q}\left(1+\left(\frac{1-p}{1-q}\right)^3\right)\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\nabla_{\delta_{1m}}\widetilde{f}_{1t}}{\widetilde{f}_t} \qquad (33)\\ &\quad+\left(\frac{1-p}{1-q}\right)^2\sum_{(i_1,i_2,i_3,i_4)\in S}\frac{1}{\widetilde{f}_t}\Bigg\{-\nabla_{\delta_{1i_1}}\nabla_{\delta_{1i_2}}\nabla_{\beta'}\widetilde{f}_{1t}\alpha^{(1)}_{i_3i_4}-\nabla_{\delta_{1i_1}}\nabla_{\delta_{1i_2}}\nabla_{\delta_1'}\widetilde{f}_{1t}\alpha^{(2)}_{i_3i_4}\\ &\quad+\frac{1}{2}\alpha^{(1)\prime}_{i_1i_2}\nabla_{\beta}\nabla_{\beta'}\widetilde{f}_{1t}\alpha^{(1)}_{i_3i_4}+\alpha^{(1)\prime}_{i_1i_2}\nabla_{\beta}\nabla_{\delta_1'}\widetilde{f}_{1t}\alpha^{(2)}_{i_3i_4}\Bigg\}\end{aligned}$$

and let $G^{(4)}_{i_1i_2i_3i_4}(p,q)$ denote a continuous Gaussian process with mean zero satisfying

$$\begin{aligned}\omega^{(4)}_{i_1i_2\ldots i_8}(p_r,q_r;p_s,q_s) &= \operatorname{Cov}\left(G^{(4)}_{i_1i_2i_3i_4}(p_r,q_r),G^{(4)}_{i_5i_6i_7i_8}(p_s,q_s)\right)\\ &= E\left[k_{i_1i_2i_3i_4,t}(p_r,q_r)\,k_{i_5i_6i_7i_8,t}(p_s,q_s)\right]\\ &\quad-E\left[\frac{\nabla_{(\beta',\delta_1')}f_{1t}}{f_t}k_{i_1i_2i_3i_4,t}(p_r,q_r)\right]I^{-1}\left[\frac{\nabla_{(\beta',\delta_1')'}f_{1t}}{f_t}k_{i_5i_6i_7i_8,t}(p_s,q_s)\right],\end{aligned}$$

where the index set $S$ in (33) is given by $S=\{jklm,jlkm,jmkl,kljm,kmjl,lmjk\}$, $k_{i_1i_2i_3i_4,t}(p,q)$ is equivalent to $\widetilde{k}_{i_1i_2i_3i_4,t}(p,q)$ but evaluated at the true parameter values (the remaining quantities are also evaluated at the true parameter values).

The next lemma characterizes the asymptotic properties of $\mathcal{L}^{(k)}_{i_1i_2\ldots i_k}(p,1-p,\widetilde{\delta})$ for $i_1,\ldots,i_k\in\{1,\ldots,n_\delta\}$ and $k=3,\ldots,8$. It generalizes Lemma 3, 4(a), 5(a)-(e) in Cho and White (2007) by allowing for multiple switching parameters.

**Lemma 4** *Under the null hypothesis and Assumptions 1-7:*

1. *The following results hold uniformly over $\{(p,q):\epsilon\le p,q\le 1-\epsilon,p+q=1\}$:*

$$\begin{aligned}T^{-1/2}\mathcal{L}^{(3)}_{jkl}(p,q,\widetilde{\delta}) &= T^{-1/2}\sum_{t=1}^{T}\widetilde{s}_{jkl,t}(p,q)+o_p(1)\Rightarrow G^{(3)}_{jkl}(p,q),\\ T^{-1/2}\mathcal{L}^{(4)}_{jklm}(p,q,\widetilde{\delta}) &= O_p(1),\quad T^{-1/2}\mathcal{L}^{(5)}_{jklmn}(p,q,\widetilde{\delta})=O_p(1),\\ T^{-1}\mathcal{L}^{(6)}_{jklmnr}(p,q,\widetilde{\delta}) &= -\sum_{(i_1,i_2,\ldots,i_6)\in IND}\omega^{(3)}_{i_1i_2\ldots i_6}(p,q;p,q)+o_p(1),\end{aligned}$$

*where IND={jklmnr,jkmlnr,jknlmr,jkrlmn,jlmknr,jlnkmr,jlrkmn,jmnklr,jmrkln,jnrklm}.*

2. *The following results hold at* $p=q=1/2:$

$$
\begin{aligned}
T^{-1/2}\mathcal{L}^{(3)}_{jkl}(p,q,\widetilde{\delta}) &= o_p(1), \\
T^{-1/2}\mathcal{L}^{(4)}_{jklm}(p,q,\widetilde{\delta}) &= T^{-1/2}\sum_{t=1}^{T}\widetilde{k}_{jklm,t}(p,q)+o_p(1)\Rightarrow G^{(4)}_{jklm}(p,q), \\
T^{-1/2}\mathcal{L}^{(k)}_{i_1 i_2\ldots i_k}(p,q,\widetilde{\delta}) &= O_p(1), \text{ where } i_1,\ldots,i_k\in\{1,\ldots,n_\delta\} \text{ for } k\text{=5,6 and 7}, \\
T^{-1}\mathcal{L}^{(8)}_{jklmnrsu}(p,q,\widetilde{\delta}) &= -\sum_{(i_1,i_2,\ldots,i_8)\in IND}\omega^{(4)}_{i_1 i_2\ldots i_8}(p,q;p,q)+o_p(1).
\end{aligned}
$$

*where the elements of IND are as follows:* $i_1=j$; *each triplet* $(i_2,i_3,i_4)$ *corresponds to one of the 35 outcomes of picking 3 elements from* $\{k,l,m,n,r,s,u\}$ *(the ordering does not matter); and* $i_5,i_6,i_{7,}i_8$ *correspond to the remaining elements.*

The above two sets of results characterize the high order derivatives along the line $p+q=1$. When $p\neq 1/2$, the third order term $T^{-1/2}\sum_{t=1}^{T}\widetilde{s}_{jkl,t}(p,1-p)$ replaces the second order term $T^{-1/2}\sum_{t=1}^{T}\widetilde{U}_{jk,t}$ to become the leading term in the likelihood expansion. Consequently, a sixth order expansion is needed to approximate the likelihood ratio. When $p=1/2$, the fourth order term $T^{-1/2}\sum_{t=1}^{T}\widetilde{k}_{jklm,t}(p,1-p)$ becomes the leading term, and consequently an eighth order expansion is needed.

The restriction $p=1-q$ is not imposed when representing the leading term in $T^{-1/2}\mathcal{L}^{(3)}_{jkl}(p,q,\widetilde{\delta})$. This ensures that the coefficient in front of $(\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}/\widetilde{f}_t)$ is correct even when $p+q\neq 1$. For the same reason, we also do not impose $p=q=1/2$ when expressing the leading term of $T^{-1/2}\mathcal{L}^{(4)}_{jklm}(p,q,\widetilde{\delta})$. The results assumes all the second order derivatives with respect to the switching parameters can be written as linear combinations of the first order derivatives (this holds when testing for switching in the intercept in a linear model). If such a relationship holds only for a subset of derivatives (this is the case when testing for switching in both the intercept and the slope parameter), then when constructing $\widetilde{s}_{jklm,t}(p,q)$, we simply set $\alpha^{(1)}_{i_1 i_2}=0$ and $\alpha^{(2)}_{i_1 i_2}=0$ for those cases where (31) is not satisfied.

We now incorporate the leading terms in Lemma 4 to obtain a refined approximation. To ease notation, we first express the relevant quantities in the lemma using matrix notations. Let $G^{(3)}(p,q)$ be a $n_\delta^3$- dimensional vector whose $(j+(k-1)n_\delta+(l-1)n_\delta^2)$-th element is given by $G^{(3)}_{jkl}(p,q)$. Let $\Omega^{(3)}(p,q)$ denote an $n_\delta^3$ by $n_\delta^3$ matrix whose $(j+(k-1)n_\delta+(l-1)n_\delta^2, m+(n-1)n_\delta+(r-1)n_\delta^2)$-th element is given by $\omega^{(3)}_{jklmnr}(p,q;p,q)$. Define

$$
\mathcal{W}^{(3)}(p,q,\eta)=T^{-1/4}\frac{1}{3}\left(\eta^{\otimes 3}\right)'\operatorname{vec}G^{(3)}(p,q)-T^{-1/2}\frac{1}{36}\left(\eta^{\otimes 3}\right)'\Omega^{(3)}(p,q)\left(\eta^{\otimes 3}\right).
$$

Let $G^{(4)}(p,q)$ be an $n_\delta^4$- dimensional vector whose $(j+(k-1)n_\delta+(l-1)n_\delta^2+(m-1)n_\delta^3)$-th element is given by $G^{(4)}_{jklm}(p,q)$. Let $\Omega^{(4)}(p,q)$ be an $n_\delta^4$ by $n_\delta^4$ matrix whose $(j+(k-1)n_\delta+(l-1)n_\delta^2+(m-1)n_\delta^3, n+(r-1)n_\delta+(s-1)n_\delta^2+(u-1)n_\delta^3)$-th element is given by $\omega^{(4)}_{jklmnrsu}(p,q;p,q)$. Define

$$\mathcal{W}^{(4)}(p,q,\eta)=T^{-1/2}\frac{1}{12}\left(\eta^{\otimes 4}\right)'\operatorname{vec}G^{(4)}(p,q)-T^{-1}\frac{1}{576}\left(\eta^{\otimes 4}\right)'\Omega^{(4)}(p,q)\left(\eta^{\otimes 4}\right).$$

We propose approximating the distribution of the SupLR$(\Lambda_\epsilon)$ test using

$$S_\infty(\Lambda_\epsilon)\equiv\sup_{(p,q)\in\Lambda_\epsilon}\sup_{\eta\in R^{n_\delta}}\left\{\mathcal{W}^{(2)}(p,q,\eta)+\mathcal{W}^{(3)}(p,q,\eta)+\mathcal{W}^{(4)}(p,q,\eta)\right\}, \tag{34}$$

where $\Lambda_\epsilon$ is specified in (12).

**Corollary 1** *Under Assumptions 1-7 and the null hypothesis, we have, over (12):*

$$\Pr\left(\text{SupLR}(\Lambda_\epsilon)\le s\right)-\Pr\left(S_\infty(\Lambda_\epsilon)\le s\right)\to 0.$$

**Remark 3** *The above result holds irrespective of whether or not the relationship (31) holds. This follows because the additional terms $\mathcal{W}^{(3)}(p,q,\eta)+\mathcal{W}^{(4)}(\eta)$ both converge to zero as $T\to\infty$. These terms provide refinement in finite samples, having no effect asymptotically.*

Below, we illustrate the refined approximation using the AR(1) model considered above. As will be seen, the refinement substantially improves the approximation when testing for regime switching in the intercept, while having little effect when testing for switching in the slope coefficient. For the latter, the original approximation in Proposition 2 is already adequate.

**The illustrative model (cont'd).** First, consider testing for regime switching in $\mu$ in (30). The quantities (32) and (33) are equal to, respectively,

$$\tfrac{(1-p)(p-q)}{(1-q)^2}\tfrac{1}{\widetilde{\sigma}^3}\left\{\left(\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right)^3-3\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right\},$$
$$\left[\tfrac{1-p}{2-p-q}\left(1+\left(\tfrac{1-p}{1-q}\right)^3\right)-3\left(\tfrac{1-p}{1-q}\right)^2\right]\tfrac{1}{\widetilde{\sigma}^4}\left\{\left(\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right)^4-6\left(\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right)^2+3\right\}.$$

The refined approximations are reported as dotted lines in the second subfigure in Figures 2 and 3. The results confirm that the improvements are substantial.

Next, consider testing for regime switching in $\alpha$. The quantities (32) and (33) are equal to, respectively,

$$\tfrac{(1-p)(p-q)}{(1-q)^2}\tfrac{1}{\widetilde{\sigma}^3}\left\{\left(\tfrac{\widetilde{u}_t y_{t-1}}{\widetilde{\sigma}}\right)^3-3\tfrac{\widetilde{u}_t y_{t-1}}{\widetilde{\sigma}}\right\},$$
$$\tfrac{1-p}{2-p-q}\left(1+\left(\tfrac{1-p}{1-q}\right)^3\right)\tfrac{1}{\widetilde{\sigma}^4}\left\{\left(\tfrac{\widetilde{u}_t y_{t-1}}{\widetilde{\sigma}}\right)^4-6\left(\tfrac{\widetilde{u}_t y_{t-1}}{\widetilde{\sigma}}\right)^2+3\right\}.$$

The refined approximation is reported as the dotted lines in the first subfigure in Figure 2. As a desirable feature, the refinement has little effects on the approximation. ■

### 5.4 Obtaining critical values

This section shows how to obtain the critical values of $S_\infty(\Lambda_\epsilon)$ defined in (34). The idea is to sample from the distribution of the vector process $[\operatorname{vec} G(p,q)', \operatorname{vec} G^{(3)}(p,q)', \operatorname{vec} G^{(4)}(p,q)']$ and then solve the maximization problem (34) over $(p,q) \in \Lambda_\epsilon$ and $\eta \in R^{n_\delta}$. Because this vector process is Gaussian with mean zero, to generate the desired draws it suffices to obtain a consistent estimator of its covariance function over $\Lambda_\epsilon$. Such an observation has also been made by Hansen (1992) and Garcia (1998).

Let $\widetilde{U}_t^{(2)}(p,q)$ be an $n_\delta^2$-dimensional vector whose $(j+(k-1)n_\delta)$-th element is given by $\widetilde{U}_{jk,t}$ in (20). Let $\widetilde{U}_t^{(3)}(p,q)$ be an $n_\delta^3$-dimensional vector whose $(j+(k-1)n_\delta+(l-1)n_\delta^2)$-th element is given by $\widetilde{s}_{jkl,t}(p,q)$ in (32). Let $\widetilde{U}_t^{(4)}(p,q)$ be an $n_\delta^4$-dimensional vector whose $(j+(k-1)n_\delta+(l-1)n_\delta^2+(m-1)n_\delta^3)$-th element is given by $\widetilde{k}_{jklm,t}(p,q)$ in (33). Define

$$\widetilde{\mathcal{G}}_t(p,q) = \begin{bmatrix} \widetilde{U}_t^{(2)}(p,q) \\ \widetilde{U}_t^{(3)}(p,q) \\ \widetilde{U}_t^{(4)}(p,q) \end{bmatrix}.$$

Let $U_t^{(2)}(p,q), U_t^{(3)}(p,q), U_t^{(4)}(p,q)$ and $\mathcal{G}_t(p,q)$ be defined as $\widetilde{U}_t^{(2)}(p,q)$, $\widetilde{U}_t^{(3)}(p,q)$, $\widetilde{U}_t^{(4)}(p,q)$ and $\widetilde{\mathcal{G}}_t(p,q)$ but evaluated at the true values under the null hypothesis. Because the vector process $T^{-1/2}\sum_{t=1}^{T}\widetilde{\mathcal{G}}_t(p,q)$ converges weakly to $[\operatorname{vec} G(p,q)', \operatorname{vec} G^{(3)}(p,q)', \operatorname{vec} G^{(4)}(p,q)']$ over $\epsilon \le p,q \le 1-\epsilon$, its covariance function provides a consistent estimator for the latter. Further, we have

$$\begin{aligned} & T^{-1/2}\sum_{t=1}^{T}\widetilde{\mathcal{G}}_t(p,q) \\ = \; & T^{-1/2}\sum_{t=1}^{T}\mathcal{G}_t(p,q) - \left\{ T^{-1}\sum_{t=1}^{T}\mathcal{G}_t(p,q)\frac{\nabla_{(\beta',\delta_1')}f_{1t}}{f_t}\right\} I^{-1}T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{(\beta',\delta_1')'}f_{1t}}{f_t} + o_p(1), \end{aligned}$$

where all the quantities on the right hand side are evaluated at the true parameter values under the null hypothesis. The term inside the curly brackets converges to a nonrandom matrix by the law of large numbers. Therefore, a consistent estimator of the desired covariance function is given by

$$T^{-1}\sum_{t=1}^{T}\widetilde{\mathcal{G}}_t\left(p_r,q_r\right)\widetilde{\mathcal{G}}_t\left(p_s,q_s\right)' \tag{35}$$
$$-\left\{T^{-1}\sum_{t=1}^{T}\widetilde{\mathcal{G}}_t\left(p_r,q_r\right)\frac{\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}}{\widetilde{f}_t}\right\}\widetilde{I}^{-1}\left\{T^{-1}\sum_{t=1}^{T}\widetilde{\mathcal{G}}_t\left(p_s,q_s\right)\frac{\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}}{\widetilde{f}_t}\right\}',$$

where $\widetilde{I}$ is the estimated information matrix, i.e., $\widetilde{I}=T^{-1}\sum_{t=1}^{T}[\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}/\widetilde{f}_t][\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}/\widetilde{f}_t]$.

**Remark 4** *The estimator (35) has three desirable features. First, the parameter values are the MLE from estimating the null model. They are simple to obtain. Second, the relevant quantities can all be expressed as functions of* $(\nabla_{\theta_j}\widetilde{f}_{1t}/\widetilde{f}_t)$, $(\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\widetilde{f}_{1t}/\widetilde{f}_t)$ *and* $(\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\nabla_{\theta_m}\widetilde{f}_{1t}/\widetilde{f}_t)$. *For models the are linear under the null hypothesis, they can all be computed analytically. Third, nuisance parameters do not affect the dimension of the optimization in (34). Therefore, they do not noticeably increase the computational cost.*

**The illustrative model (cont'd).** We show how to compute the quantities in (35) when testing for switching in $\mu$ in the model (30). In more general linear models with multiple switching parameters, the relevant quantities can be obtained in a similar manner. The vector $\widetilde{\mathcal{G}}_t\left(p_r,q_r\right)$ consists of three elements ($\widetilde{u}_t$ denote OLS residuals):

$$\widetilde{U}_t^{(2)}(p,q)=\tfrac{2}{\widetilde{\sigma}^2}\left(\tfrac{1-p}{1-q}\right)\textstyle\sum_{s=1}^{t-1}(p+q-1)^s\tfrac{\widetilde{u}_{t-s}\widetilde{u}_t}{\widetilde{\sigma}^2},$$
$$\widetilde{U}_t^{(3)}(p,q)=\tfrac{(1-p)(p-q)}{(1-q)^2}\tfrac{1}{\widetilde{\sigma}^3}\left\{\left(\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right)^3-3\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right\},$$
$$\widetilde{U}_t^{(4)}(p,q)=\left[\tfrac{1-p}{2-p-q}\left(1+\left(\tfrac{1-p}{1-q}\right)^3\right)-3\left(\tfrac{1-p}{1-q}\right)^2\right]\tfrac{1}{\widetilde{\sigma}^4}\left\{\left(\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right)^4-6\left(\tfrac{\widetilde{u}_t}{\widetilde{\sigma}}\right)^2+3\right\}.$$

They depend on the model only through the OLS residuals. The vector $\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}/\widetilde{f}_t$ also consists of three elements:

$$\frac{\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}}{\widetilde{f}_t}=\left[\begin{array}{ccc}\frac{y_{t-1}\widetilde{u}_t}{\widetilde{\sigma}^2} & \frac{1}{2\widetilde{\sigma}^2}\left(\frac{\widetilde{u}_t^2}{\widetilde{\sigma}^2}-1\right) & \frac{\widetilde{u}_t}{\widetilde{\sigma}^2}\end{array}\right].$$

They depend on the model only through the OLS residuals and the regressor $y_{t-1}$ and are also simple to compute. The estimated information matrix is $\widetilde{I}=T^{-1}\sum_{t=1}^{T}[\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}/\widetilde{f}_t][\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}/\widetilde{f}_t]$. ■

# 6 Implications for bootstrap procedures and information criteria

The results in the previous section provide a platform for evaluating the consistency of various bootstrap procedures. Although a comprehensive study of such procedures is beyond the scope of the paper, it is possible to illustrate some important aspects using the linear model (9). Throughout this section the test is computed over $\Lambda_\epsilon$ defined by (12).

**Bootstrap procedures.** We begin with the important special case where the regressors contain only a constant and lagged values of $y_t$, and the errors are normally distributed. A standard parametric bootstrap procedure proceeds as follows. (1) Estimate the model under the null hypothesis (e.g., estimate an autoregressive model), (2) Sample from the Normal distribution whose mean equals zero and variance equals the sample variance of the residuals. Use the draws along with the estimated coefficients to generate a new autoregressive series. (3) Compute the test using the newly generated series. (4) Repeat the steps (1)-(3). The above procedure is asymptotically valid. This follows because all the parameters are estimated consistently, and the normality and the AR(1) structure are also preserved. Consequently, the covariance function in the bootstrap world is consistent with what determines the asymptotic distribution in Proposition 1.

Next, consider the more general situation where a second variable is present in the regressors; e.g., an autoregressive distributed lags (ADL) model. Because the model does not specify the joint distribution of the dependent variable and the regressors, the bootstrap procedure described above is no longer applicable. Two alternative approaches deserve some consideration.

The first approach involves keeping the regressors fixed at their original values when generating the data, i.e., using the fixed regressor bootstrap. This procedure has been shown to be asymptotically valid in the context of testing for structural breaks (Hansen, 2000). However, the same procedure is in general no longer consistent in the current context. This is because, in contrast to the original model, in the bootstrap world the regressors are strictly but not weakly exogenous. This alters the covariance function appearing in Proposition 1 (c.f. (24) and (25) and the accompanying discussions). We provide some simulation results to illustrate the potential severity of the size distortion. The data are generated using the model (30) with the same specifications. The sample size $T = 250$. The solid line in Figure 4 shows the finite sample distribution, while the dashed line corresponds to the fixed regressor bootstrap. The difference is quite substantial. This difference does not decrease when the sample size is increased to 500.

The second approach involves specifying the joint distribution of the data. For example, if we have an ADL model with normal errors, we specify a full model that corresponds to a Gaussian vector autoregression. Then, we can apply the parametric bootstrap to the augmented model. Such a bootstrap procedure will be consistent if it asymptotically produces the same covariance function in Proposition 1. A key property of this procedure is that it entails specifying a parametric model for the regressors. Investigating the sensitivity to such specifications is useful but is beyond the scope of this paper.

**Information criteria.** The asymptotic results also imply that the finite sample properties of conventional information criteria, such as BIC, can be sensitive to the structure of the model and also which parameters are allowed to switch. This is because the distribution of the likelihood ratio depends on which parameter is allowed to switch, while in BIC the penalty term depends only on the dimension of the model and the sample size. We illustrate such sensitivities using the model (30) by contrasting the outcomes from the following two applications. (1) We apply BIC to determine whether there is regime switching in the intercept. The other parameters are assumed to be constant. (2) The same as (1) except that the slope parameter is allowed to switch. In the simulated data, no regime switching is present; $\mu = 0, \alpha = 0.5$ and $\sigma^2 = 1$. The set $\Lambda_\epsilon$ is specified as (13) with $\epsilon = 0.05$. The sample size is 250. Out of the 5000 realizations, BIC falsely classifies 12.5% for the first application, while only 2.4% for the second application. Because the penalty terms in the Akaike information criterion and the Hannan–Quinn information criterion have the same structure, they are also expected to exhibit the same sensitivity.

# 7 Monte Carlo

We examine the test's size and power properties and also compare with the tests of Cho and White (2007) and Carrasco, Hu and Ploberger (2014). The DGP is

$$y_t = \mu_1 \cdot 1_{\{s_t=1\}} + \mu_2 \cdot 1_{\{s_t=2\}} + \alpha y_{t-1} + e_t \quad \text{with } e_t \sim i.i.d.\ N(0, \sigma^2), \tag{36}$$

where the intercept switches between two regimes with $p(s_t = 1|s_{t-1} = 1) = p$ and $p(s_t = 2|s_{t-1} = 2) = q$, $\alpha = 0.5$ and $\sigma^2 = 1$. The choice of this DGP is motivated by two considerations. First, it is considered in Cho and White (2007), therefore using it facilities the comparison between the tests. Second, as seen in the empirical application in Section 8, such a specification provides a sensible approximation to the postwar U.S. quarterly real GDP growth series. Throughout this section, we specify $\Lambda_\epsilon$ as (12) with $\epsilon = 0.05$ and 0.02. The distribution (34) is simulated using 5000 realizations. The rejection frequencies reported are all based on 5000 replications.

Let $\widetilde{\theta} = (\widetilde{\mu}, \widetilde{\alpha}, \widetilde{\sigma}^2)'$ denote the MLE under the null hypothesis. The supTS of Carrasco, Hu and Ploberger (2014) is implemented as follows. First, obtain $\mu_{2,t}(\rho) = (1/(2\widetilde{\sigma}^4)) \sum_{s<t} \rho^{t-s} \widetilde{e}_t \widetilde{e}_s$,

$\Gamma_T(\rho) = T^{-1/2}\sum_{t=1}^{T}\mu_{2,t}(\rho)$, and $\mathcal{E}_T(\rho) = T^{-1}\sum_{t=1}^{T}\varepsilon_t(\rho)'\varepsilon_t(\rho)$, where $\widetilde{e}_t = y_t - \widetilde{\mu} - \widetilde{\alpha}y_{t-1}$, $\rho = p+q-1$, and $\varepsilon_t(\rho)$ are the residuals from regressing $\mu_{2,t}(\rho)$ on the score with the latter computed under the null hypothesis and evaluated at $\widetilde{\theta}$. Next, compute the supremum of $0.5[\max(\Gamma_T(\rho)/\sqrt{\mathcal{E}_T(\rho)}, 0)]^2$ over $\rho$. We consider $\rho \in [0.05, 0.90]$ and $\rho \in [0.02, 0.96]$. They correspond to $\Lambda_{0.05}$ and $\Lambda_{0.02}$ specified above. The resulting tests are denoted by $\text{supTS}_1$ and $\text{supTS}_2$ respectively.

Table 1 reports rejection frequencies under the null hypothesis, i.e., with $\mu_1 = \mu_2 = 0$. The rejection frequencies of $\text{SupLR}(\Lambda_\epsilon)$ are overall close to the nominal levels, although some mild over-rejections do exist. In particular, when $T = 200$, the rejection rates at the 5% and 10% levels are 6.60% and 14.46% for $\epsilon = 0.05$, and 6.22% and 13.62% for $\epsilon = 0.02$. Similar rejection rates are observed when $T = 500$. The results also confirm that the QLR and supTS tests have excellent size properties.

For power properties, following Cho and White (2007), we let $\mu_1 = -\mu_2$ with $\mu_2 = 0.2$, 0.6 and 1.0. Motivated by the empirical estimates discussed in Section 3, we consider three pairs of values for $(p, q)$: $(0.70, 0.70)$, $(0.70, 0.90)$ and $(0.90, 0.90)$. The rejection frequencies at the 5% nominal levels are reported in Table 2.

As none of the alternatives correspond to mixtures (i.e., $p+q \neq 1$), the power of the $\text{SupLR}(\Lambda_\epsilon)$ test is consistently higher than that of QLR. The difference increases significantly as one or both of the regimes become more persistent, i.e., as the value of $p+q$ increases. For example, consider the cases $\mu_2 = 0.6$ and 1.0. When $(p, q) = (0.7, 0.7)$, the rejection frequencies for the $\text{SupLR}(\Lambda_{0.05})$ test are 20.24% and 96.58%, with the corresponding values for QLR being 9.46% and 68.83%. When $(p, q) = (0.7, 0.9)$, the rejection frequencies for the $\text{SupLR}(\Lambda_{0.05})$ test become 38.14% and 99.80% with the corresponding values for QLR being 13.40% and 60.56%. Further, when $(p, q) = (0.9, 0.9)$, the values become 60.30% and 100% for $\text{SupLR}(\Lambda_{0.05})$, and 7.06% and 7.30% for QLR. The results strongly suggest that although the test of Cho and White (2007) can be valuable for detecting mixtures, the $\text{SupLR}(\Lambda_\epsilon)$ test can offer substantial power gains when the DGPs are expected to fall outside that family.

The comparison with the supTS tests shows a similar pattern. However, the reason behind the power difference is not the same. Specifically, a key component of the test is $\mu_{2,t}(\rho)$, which measures the correlation between the residuals $(\widetilde{e}_t)$ obtained under the null hypothesis. On the one hand, omitting regime switching causes $\widetilde{e}_t$ to be *positively* correlated. On the other hand, such an omission makes $\widetilde{\alpha}$ upward biased, with the bias growing stronger as the regimes become more persistent (such a phenomenon is studied extensively in Perron (1990 and 1991), which showed that omitting a structural change can cause the autoregressive coefficient to be upward biased, potentially leading to low testing power for hypotheses related to unit roots or deterministic trends). The bias in $\widetilde{\alpha}$ causes overdifferencing the series and consequently makes $\widetilde{e}_t$ *negatively* correlated. In finite samples, these two opposite effects can potentially annihilate each other, making the value of $\mu_{2,t}(\rho)$ insensitive to the departure from the null hypothesis. This finding is consistent with the simulation results in Carrasco, Hu and Ploberger (2014, Table II), which show that the test can have good power properties when the lagged dependent variable is not present.

Some further simulations and comparisons are provided in the next section using parameters calibrated to empirical estimates.

# 8 Application

Following the influential work of Hamilton (1989), a large body of literature has considered modeling the US real output growth as a regime switching process. Here, we apply the SupLR($\Lambda_\epsilon$) test to assess the evidence for such a specification. The analysis is based on the real GDP growth rates (Series GDPC1, available from the Saint Louis Fed website). It utilizes a full sample that consists of quarterly observations over the period 1960:I–2014:IV and a range of subsamples specified later. The analysis proceeds as follows. First, we examine whether the SupLR($\Lambda_\epsilon$) test detects strong evidence for regime switching that holds consistently over different subsamples. Next, we examine whether such evidence is still present when the QLR and supTS tests are used instead. Then, we compute the smoothed regime probabilities to examine the empirical relevance of the model and the results. Finally, some simulations are conducted with parameters values calibrated to the empirical estimates to further illustrate the test's size and power properties in this important application. The model (36) is used throughout, though sensitivity analysis will also be conducted. The set $\Lambda_\epsilon$ is as in (12) with $\epsilon = 0.02$. All the results are based on 5% critical values unless stated otherwise.

**The testing results.** We begin with the full sample. The SupLR($\Lambda_{0.02}$) test equals 8.75, with the critical value being 7.62. The null hypothesis is therefore rejected at the 5% level. Note that the above full sample includes the recent Great Recession, which might have had a large effect on the test. To evaluate the evidence further, we consider a subsample that corresponds to 1960:I–2006:IV. The SupLR($\Lambda_{0.02}$) test equals 8.57. The critical value is 7.61. The null hypothesis remains rejected.

The analysis can be taken further. That is, the SupLR($\Lambda_{0.02}$) test can be computed over a range of subsamples to evaluate the consistency of the results. To this end, we let the first subsample be 1960:I–1980:I and then gradually incorporate additional observations quarter-by-quarter. This leads to 140 subsamples of increasing sizes. The resulting values are shown in Figure 5(a). Note that the critical values are pointwise with respect to the subsamples, therefore the figure should be interpreted as an informal illustration. There, the test statistics exceed the critical values for 106 out of the 140 subsamples. We conclude that there is fairly consistent evidence favoring the regime switching specification. To our knowledge, this is the first occasion such consistent evidence for regime switching in output growth is documented through hypothesis testing.

Figures 5(b)-(c) report the QLR and supTS tests over the same subsamples. The two tests exceed the critical values only when the Great Recession period is included. Overall, the evidence for the regime switching specification is not as strong when viewed through these two tests.

**Recession probabilities.** Figures 6(a)-(b) report the regime probabilities for the two samples 1960:I–2006:IV and 1960:I–2014:IV. This allows us to examine the simple model (36)'s empirical adequacy and also assess the effect of the Great Recession on the estimates. In the figures, the shaded areas correspond to NBER's recession dating available from its website.

The results suggest that the model provides an informative approximation. Specifically, for the period 1960:I–2006:IV, the recession probabilities agree well with the NBER's dating for all the recessions. For the full sample, the two results remain consistent, except that the model now assigns low probabilities to the relatively shallow recessions of 1969IV-1970:IV and 2001:I-2001:IV. This follows because when the Great Recession is included, the estimates for $(\mu_1, \mu_2, \alpha, \sigma^2, p, q)$ change from $(-0.16, 0.97, 0.09, 0.48, 0.77, 0.94)$ to $(-0.54, 0.75, 0.19, 0.49, 0.66, 0.96)$ and, consequently, the mean growth rate during recessions decreases from $-0.18$ to $-0.67$. The difference can therefore be viewed as reflecting the unusual nature of the recent recession.

**Robustness checks.** We evaluate the results' robustness along two dimensions.

In practice, the lag order of the autoregression under the null hypothesis is unknown and often determined using some information criterion. To reflect such an uncertainty, we estimate the lag orders associated with the subsamples using BIC and then repeat the analysis. The minimum and maximum lag orders are set to 1 and 4. Note that here BIC is applied under the null hypothesis to control the size of the test. This is different from using it to determine whether regime switching is present as described in Section 6. The null hypothesis is rejected at the 5% level for 92 of the 140 subsamples. The evidence of regime switching remains fairly consistent. At the same time, the results also points to the increased difficulty in distinguishing between a regime switching specification and a linear specification that allows for more flexible serial dependencies.

We repeat the analysis but using reverse recursive subsamples. That is, we let 1994:IV–2014:IV be the first subsample and then incorporate additional observations backward quarter-by-quarter. The lag order is determined by BIC for each subsample. The results show that the null hypothesis is rejected at the 5% level for 120 of the 140 subsamples. Finally, we exclude the Great Recession, i.e., letting 1986:IV-2006:IV be the first subsample and then incorporate additional observations backward quarter-by-quarter. The null hypothesis is now rejected at the 5% level for 47 out of the 108 subsamples. It is rejected in 88 out of the 108 subsamples if the 10% nominal level is used instead. Therefore, although the evidence in the latter situation is overall weaker, it remains considerable and fairly consistent across the subsamples.

**Further simulations.** We evaluate the size and power properties using the parameter estimates obtained above. Specifically, we simulate data using the model (36) with the parameter values $(\mu_1, \mu_2, \alpha, \sigma^2, p, q)$ set to the estimates obtained under the null and alternative hypotheses. The sample sizes correspond to those implied by 1960:I–2006:IV and 1960:I–2014:IV. The results are summarized below.

Consider the rejection frequencies under the null hypothesis. For the period 1960:I–2006:IV, we obtain $(\widetilde{\mu}, \widetilde{\alpha}, \widetilde{\sigma}^2) = (0.60, 0.28, 0.65)$. The rejection frequencies at the $2.5\%, 5.0\%, 7.5\%$ and 10% levels are $3.18\%, 6.62\%, 10.30\%$ and 14.68% for SupLR($\Lambda_{0.05}$) and $2.98\%, 7.36\%, 10.74\%$ and 14.36% for SupLR($\Lambda_{0.02}$). For the period 1960:I–2014:IV, we obtain $(\widetilde{\mu}, \widetilde{\alpha}, \widetilde{\sigma}^2) = (0.51, 0.33, 0.64)$. At the the same levels, the rejection frequencies are $2.74\%, 5.76\%, 9.10\%$ and 13.02% for SupLR($\Lambda_{0.05}$) and $2.80\%, 5.64\%, 9.38\%$ and 12.94% SupLR($\Lambda_{0.02}$). These values are consistent with the simulation results reported in the previous section.

Consider the rejection rates under the alternative hypothesis. The estimates of $(\mu_1, \mu_2, \alpha, \sigma^2, p, q)$ for the two periods are $(-0.16, 0.97, 0.09, 0.48, 0.77, 0.94)$ and $(-0.54, 0.75, 0.19, 0.49, 0.66, 0.96)$. The rejection frequencies of SupLR($\Lambda_{0.02}$) equal 66% and 65%. Overall, the results suggest the test can be informative in empirically relevant situations. In comparison, the rejection frequencies for the QLR test are 14% and 25%, and for the $\text{supTS}_2$ test are 24% and 10%, respectively.

## 9 Conclusion

This paper analyzed a family of likelihood based tests for Markov regime switching in the context of nonlinear models allowing for multiple switching parameters. In addition to deriving the limiting distribution and obtaining a finite sample refinement, a unified algorithm for simulating the critical values was also developed. When applied to the US quarterly real GDP growth rates, the tests delivered consistent evidence favoring the regime switching specification. It is conjectured that the techniques developed can have implications for hypothesis testing in other related contexts, such as testing for Markov switching in state space models and in multivariate regressions. Such investigations are currently in progress.

## References


Amemiya, T. (1985): *Advanced Econometrics*, Harvard University Press.

Andrews, D.W.K. (1993): " Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821-856.

— (2001): "Testing When a Parameter is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69, 683-734.

Andrews, D.W.K and W. Ploberger (1994): "Optimal Tests when a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

—— (1995): "Admissibility of the Likelihood Ratio Test When a Nuisance Parameter is Present Only Under the Alternative," *The Annals of Statistics*, 23, 1609-1629.

Ang, A. and G. Bekaert (2002a): "International Asset Allocation With Regime Shifts," *Review of Financial Studies*," 15, 1137-1187.

—— (2002b): "Regime Switches in Interest Rates," *Journal of Business and Economic Statistics*, 20, 163-182.

Bai, J. and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47-78.

Billingsley, P. (1999): *Convergence of Probability Measures*, John Wiley & Sons.

Carrasco, M., L. Hu, and W. Ploberger (2014): "Optimal Test for Markov Switching Parameters," *Econometrica*, 82, 765-784.

Carter, A.V. and D.G. Steigerwald (2012): "Testing for Regime Switching: A Comment," *Econometrica,* 80, 1809-1812.

Chauvet, M. and J. Hamilton (2006): "Dating Business Cycle Turning Points," In *Nonlinear Time Series Analysis of Business Cycles*, (Milas, C., P. Rothman, and D. van Dijk, eds.), Elsevier.

Chen, J. and P. Li (2009): "Hypothesis Test for Normal Mixture Models: The EM Approach," *The Annals of Statistics*, 37, 2523-2542.

Chen, X. (2013): "Penalized Sieve Estimation and Inference of Semi-nonparametric Dynamic Models:A Selective Review," In *Advances in Economics and Econometrics*, Cambridge University Press.

Chen, X., M. Ponomareva, and E. Tamer (2014): "Likelihood Inference in Some Finite Mixture Models," *Journal of Econometrics*, 182, 87-99.

Chesher, A. (1984): "Testing for Neglected Heterogeneity," *Econometrica*, 22, 865–872.

Cho, J. S. and H. White (2007): "Testing for Regime Switching," *Econometrica*, 75, 1671-1720.

Cosslett, S. and L.-F. Lee (1985): "Serial Correlation in Discrete Variable Models," *Journal of Econometrics*, 27, 79–97.

Dai, Q., K.J. Singleton and W. Yang (2007): "Regime Shifts in a Dynamic Term Structure Model of U.S. Treasury Bond Yields," *Review of Financial Studies*, 20, 1669-1706.

Davidson, R. and J. Mackinnon (1991): "Une nouvelle forme du test de la matrice d'information," *Annales d'Economie et de Statistique*, 20/21, 171-192.

Davig, T. (2004): "Regime-Switching Debt and Taxation," *Journal of Monetary Economics*, 51, 837-859.

Davies, R. (1977): "Hypothesis Testing when a Nuisance Parameter Is Present Only under the Alternative," *Biometrika*, 64, 247–254.

—— (1987): "Hypothesis Testing when a Nuisance Parameter Is Present Only under the Alternative," *Biometrika*, 74, 33–43.

Dovonon, P. and E. Renault (2013): "Testing for Common Conditionally Heteroskedastic Factors," *Econometrica,* 81, 2561-2586.

Garcia, R. (1998): "Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models," *International Economic Review*, 39, 763–788.

Garcia, R., R. Luger and E. Renault (2003): "Empirical Assessment of an Intertemporal Option Pricing Model with Latent Variables," *Journal of Econometrics*, 116, 49-83.

Goldfeld, S.M. and R.E. Quandt (1973): "A Markov Model for Switching Regressions," *Journal of Econometrics*, 1, 3-16.

Ghosh, J.K. and P.K. Sen (1985): "On the Asymptotic Performance of the Log Likelihood Ratio Statistic for the Mixture Model and Related Results," In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*. (L.M. Le Cam and R.A. Olshen, eds.), 2, 789-806. Wadsworth, Belmont, CA.

Gu, J., R.Koenker, and S. Volgushev (2013): "Testing for Homogeneity in Mixture Models." working paper, University of Illinois at Urbana-Champaign.

Hamilton, J. (1988): "Rational-Expectations Econometric Analysis of Changes in Regime : An Investigation of the Term Structure of Interest Rates," *Journal of Economic Dynamics and Control*, 12, 385-423.

—— (1989): "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357-384.

—— (1990): "Analysis of Time Series Subject to Changes in Regime," *Journal of Econometrics*, 45, 39-70.

—— (1996): "Specification Testing in Markov-Switching Time-Series Models," *Journal of Econometrics* 70, 127-157.

—— (2005): "What's Real About the Business Cycle?" *Federal Reserve Bank of St. Louis Review*, 435-452.

—— (2008): "Regime-Switching Models," In *New Palgrave Dictionary of Economics*, edited by S. Durlauf and L. Blume. London: Palgrave Macmillan.

—— (2014): "Macroeconomic Regimes and Regime Shifts," Prepared for *Handbook of Macroeconomics*, Vol 2, University of California, San Diego.

Hansen, B. (1992): "The Likelihood Ratio Test under Non-Standard Conditions: Testing the Markov Switching Model of GNP," *Journal of Applied Econometrics*, 7, S61–S82.

—— (1996): "Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis," *Econometrica*, 64, 413-430.

—— (2000): "Testing for Structural Change in Conditional Models," *Journal of Econometrics,* 97, 93-115.

Hartigan, J.A. (1985): "A Failure of Likelihood Asymptotics for Normal Mixtures," In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*. (L.M. Le Cam and R.A. Olshen, eds.), 2, 807-810. Wadsworth, Belmont, CA.

Hawkins, D.M. (1987): "A Test for a Change Point in a Parametric Model Based on a Maximal Wald-Type Statistic," *Sankhyā: The Indian Journal of Statistics*, Series A, 49, 368-376.

Jeanne, O. and P. Masson (2000): "Currency Crises, Sunspots and Markov-Switching Regimes," *Journal of International Economics*, 50, 327-350.

King, M.L. and T.S. Shively (1993): "Locally Optimal Testing When a Nuisance Parameter is Present Only Under the Alternative," *Review of Economics and Statistics*, 75, 1-7.

Lancaster, T. (1984): "The Covariance Matrix of the Information Matrix Test," *Econometrica*, 52, 1051-1053.

Lindsay, B.G. (1995): *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics.

Liu, X. and Y. Shao (2003): "Asymptotics for Likelihood Ratio Tests under Loss of Identifiability," *Annals of Statistics,* 31, 807-832.

Mokkadem, A. (1988): "Mixing Properties of ARMA Processes," *Stochastic Processes and their Applications*, 29, 309-315.

Neyman, J. and E.L. Scott (1966): "On the Use of C($\alpha$) Optimal Tests of Composite Hypotheses," *Bulletin of the International Statistical Institute*, 41, 477-497.

Perron, P. (1990): "Testing for a Unit Root in a Time Series with a Changing Mean," *Journal of Business and Economic Statistics,* 8, 153-162.

——— (1991): "A Test for Changes in a Polynomial Trend Function For a Dynamic Time Series," Working paper, Boston University.

Rotnitzky, A., D.R. Cox, M. Bottai, and J. Robins (2000): "Likelihood-based Inference with Singular Information Matrix." *Bernoulli*, 6, 243-284.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-25.

—— (2001): *Asymptotic Theory for Econometricians*, Academic Press.

## Appendix

Throughout the appendix, $\xi_{t|t-1}(p,q,\beta,\delta_1,\delta_2)$, $f_t(\beta,\delta_1)$ and $f_t(\beta,\delta_2)$ are abbreviated as $\xi_{t|t-1}$, $f_{1t}$ and $f_{2t}$, respectively. As stated prior to Lemma 1, "$-$" (e.g, $\bar{\xi}_{t|t-1}$) denotes that a quantity is evaluated at $\delta_1=\delta_2=\delta$, where $\delta$ is some arbitrary parameter value in $\Delta$.

**Proof of Lemma 1**. The equation (15) can be written as

$$\xi_{t+1|t}=p+\rho\frac{A_t}{B_t}, \tag{A.1}$$

where $\rho$ is as defined in the lemma, $A_t=f_{2t}(\xi_{t|t-1}-1)$ and $B_t=(f_{1t}-f_{2t})\xi_{t|t-1}+f_{2t}$.

Consider Lemma 1.1. Apply $\bar{f}_{1t}=\bar{f}_{2t}=\bar{f}_t$:

$$\bar{B}_t=\bar{f}_t \quad\text{and}\quad \bar{A}_t=\bar{f}_t(\bar{\xi}_{t|t-1}-1). \tag{A.2}$$

Plugging this into (A.1), we obtain $\bar{\xi}_{t+1|t}=p+\rho(\bar{\xi}_{t|t-1}-1)$. This implies $\bar{\xi}_{2|1}=p+\rho(\bar{\xi}_{1|0}-1)=p+\rho\left(\xi_*-1\right)=\xi_*$, where the last equality follows from the definition of $\rho$ and $\xi_*$. This process can be iterated forward, leading to $\bar{\xi}_{t+1|t}=\xi_*$ for all $t\geq 1$.

Consider Lemma 1.2. Differentiate (A.1) with respect to $\theta_j$ ($j=1,...,n_\beta+2n_\delta$):

$$\nabla_{\theta_j}\xi_{t+1|t}=\rho\left\{\frac{\nabla_{\theta_j}A_t}{B_t}-\frac{A_t\nabla_{\theta_j}B_t}{B_t^2}\right\}, \tag{A.3}$$

where

$$\begin{aligned}\nabla_{\theta_j}A_t &= \nabla_{\theta_j}f_{2t}(\xi_{t|t-1}-1)+f_{2t}\nabla_{\theta_j}\xi_{t|t-1},\\ \nabla_{\theta_j}B_t &= (\nabla_{\theta_j}f_{1t}-\nabla_{\theta_j}f_{2t})\xi_{t|t-1}+(f_{1t}-f_{2t})\nabla_{\theta_j}\xi_{t|t-1}+\nabla_{\theta_j}f_{2t}.\end{aligned}$$

Below, we evaluate the right hand side of (A.3) under three possible situations:

**(1)**. If $j\in I_0$, then $\nabla_{\theta_j}\bar{f}_{1t}=\nabla_{\theta_j}\bar{f}_{2t}$ and $\bar{f}_{1t}=\bar{f}_{2t}=\bar{f}_t$, implying

$$\nabla_{\theta_j}\bar{A}_t=(\bar{\xi}_{t|t-1}-1)\nabla_{\theta_j}\bar{f}_{2t}+\bar{f}_t\nabla_{\theta_j}\bar{\xi}_{t|t-1} \quad\text{and}\quad \nabla_{\theta_j}\bar{B}_t=\nabla_{\theta_j}\bar{f}_{2t} \tag{A.4}$$

Combining this with (A.2), we obtain $\nabla_{\theta_j}\bar{\xi}_{t+1|t}=\rho\nabla_{\theta_j}\bar{\xi}_{t|t-1}$. In particular, at $t=1$: $\nabla_{\theta_j}\bar{\xi}_{2|1}=\rho\nabla_{\theta_j}\bar{\xi}_{1|0}=\rho\nabla_{\theta_j}\xi_*=0$, where the last equality holds because $\xi_*$ is independent of $\theta$. This process can be iterated forward, leading to $\nabla_{\theta_j}\bar{\xi}_{t+1|t}=0$ for all $t\geq 1$.

**(2)**. If $j\in I_1$, then $\nabla_{\theta_j}\bar{f}_{2t}=0$ and $\bar{f}_{1t}=\bar{f}_{2t}=\bar{f}_t$, implying

$$\nabla_{\theta_j}\bar{A}_t=\bar{f}_t\nabla_{\theta_j}\bar{\xi}_{t|t-1} \quad\text{and}\quad \nabla_{\theta_j}\bar{B}_t=\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}. \tag{A.5}$$

Combining this with (A.2), we have $\nabla_{\theta_j}\bar{\xi}_{t+1|t}=\rho\nabla_{\theta_j}\bar{\xi}_{t|t-1}-\rho(\bar{\xi}_{t|t-1}-1)\bar{\xi}_{t|t-1}\nabla_{\theta_j}\log\bar{f}_{1t}=\rho\nabla_{\theta_j}\bar{\xi}_{t|t-1}-\rho\left(\xi_*-1\right)\xi_*\nabla_{\theta_j}\log\bar{f}_{1t}$. The result then follows because $r=\rho\left(1-\xi_*\right)\xi_*$. Note that $\nabla_{\theta_j}\bar{\xi}_{t+1|t}$ can also be written as

$$\nabla_{\theta_j}\bar{\xi}_{t+1|t}=r\sum_{s=0}^{t-1}\rho^s\nabla_{\theta_j}\log\bar{f}_{1(t-s)} \tag{A.6}$$

**(3)**. If $j \in I_2$, then $\nabla_{\theta_j} \bar{f}_{1t} = 0$ and $\bar{f}_{1t} = \bar{f}_{2t} = \bar{f}_t$, implying

$$\nabla_{\theta_j} \bar{A}_t = \nabla_{\theta_j} \bar{f}_{2t}(\bar{\xi}_{t|t-1} - 1) + \bar{f}_t \nabla_{\theta_j} \bar{\xi}_{t|t-1} \text{ and } \nabla_{\theta_j} \bar{B}_t = (1 - \bar{\xi}_{t|t-1}) \nabla_{\theta_j} \bar{f}_{2t}. \tag{A.7}$$

Therefore, $\nabla_{\theta_j} \bar{\xi}_{t+1|t} = \rho \nabla_{\theta_j} \bar{\xi}_{t|t-1} + \rho(\bar{\xi}_{t|t-1} - 1)\bar{\xi}_{t|t-1} \nabla_{\theta_j} \log \bar{f}_{2t} = -r \sum_{s=0}^{t-1} \rho^s \nabla_{\theta_j} \log \bar{f}_{2(t-s)}$. Because $\nabla_{\theta_j} \bar{f}_{2(t-s)} = \nabla_{\theta_{j-n_\delta}} \bar{f}_{1(t-s)}$ when $j \in I_2$, it follows that $\nabla_{\theta_j} \bar{\xi}_{t+1|t}$ equals the negative of (A.6).

Consider Lemma 1.3. Differentiating (A.3) with respect to $\theta_k$:

$$\nabla_{\theta_j} \nabla_{\theta_k} \xi_{t+1|t} = \rho \left\{ \frac{\nabla_{\theta_j} \nabla_{\theta_k} A_t}{B_t} - \frac{\nabla_{\theta_j} A_t \nabla_{\theta_k} B_t}{B_t^2} - \frac{\nabla_{\theta_k} A_t \nabla_{\theta_j} B_t}{B_t^2} - \frac{A_t \nabla_{\theta_j} \nabla_{\theta_k} B_t}{B_t^2} + 2\frac{A_t \nabla_{\theta_j} B_t \nabla_{\theta_k} B_t}{B_t^3} \right\}, \tag{A.8}$$

where

$$\begin{aligned}
\nabla_{\theta_j} \nabla_{\theta_k} A_t &= \nabla_{\theta_j} \nabla_{\theta_k} f_{2t}(\xi_{t|t-1} - 1) + \nabla_{\theta_j} f_{2t} \nabla_{\theta_k} \xi_{t|t-1} + \nabla_{\theta_k} f_{2t} \nabla_{\theta_j} \xi_{t|t-1} + f_{2t} \nabla_{\theta_j} \nabla_{\theta_k} \xi_{t|t-1}, \\
\nabla_{\theta_j} \nabla_{\theta_k} B_t &= (\nabla_{\theta_j} \nabla_{\theta_k} f_{1t} - \nabla_{\theta_j} \nabla_{\theta_k} f_{2t}) \xi_{t|t-1} + (\nabla_{\theta_j} f_{1t} - \nabla_{\theta_j} f_{2t}) \nabla_{\theta_k} \xi_{t|t-1} \\
&\quad + (\nabla_{\theta_k} f_{1t} - \nabla_{\theta_k} f_{2t}) \nabla_{\theta_j} \xi_{t|t-1} + (f_{1t} - f_{2t}) \nabla_{\theta_j} \nabla_{\theta_k} \xi_{t|t-1} + \nabla_{\theta_j} \nabla_{\theta_k} f_{2t}.
\end{aligned}$$

We now evaluate the right hand side of (A.8) at $\delta_1 = \delta_2 = \delta$ under six possible situations:

(**1**). If $j \in I_0$ and $k \in I_0$, then $\bar{f}_{1t} = \bar{f}_{2t} = \bar{f}_t$, $\nabla_{\theta_j} \bar{f}_{1t} = \nabla_{\theta_j} \bar{f}_{2t}, \nabla_{\theta_k} \bar{f}_{1t} = \nabla_{\theta_k} \bar{f}_{2t}, \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{1t} = \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{2t}$ and $\nabla_{\theta_j} \bar{\xi}_{t+1|t} = \nabla_{\theta_k} \bar{\xi}_{t+1|t} = 0$, implying $\nabla_{\theta_j} \nabla_{\theta_k} \bar{A}_t = \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{2t}(\bar{\xi}_{t|t-1} - 1) + \bar{f}_t \nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t|t-1}$ and $\nabla_{\theta_j} \nabla_{\theta_k} \bar{B}_t = \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{2t}$. Combining them with (A.2) and (A.4), $\nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t+1|t}$ equals

$$\begin{aligned}
&\rho \left\{ \frac{\nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{2t}(\bar{\xi}_{t|t-1} - 1) + \bar{f}_t \nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t|t-1}}{\bar{f}_t} - \frac{(\bar{\xi}_{t|t-1} - 1) \nabla_{\theta_j} \bar{f}_{2t} \nabla_{\theta_k} \bar{f}_{2t}}{\bar{f}_t^2} \right. \\
&\left. - \frac{(\bar{\xi}_{t|t-1} - 1) \nabla_{\theta_k} \bar{f}_{2t} \nabla_{\theta_j} \bar{f}_{2t}}{\bar{f}_t^2} - \frac{(\bar{\xi}_{t|t-1} - 1) \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{2t}}{\bar{f}_t} + 2\frac{(\bar{\xi}_{t|t-1} - 1) \nabla_{\theta_j} \bar{f}_{2t} \nabla_{\theta_k} \bar{f}_{2t}}{\bar{f}_t^2} \right\} \\
&= \rho \nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t|t-1}.
\end{aligned}$$

Starting at $t = 1$ and iterating forward, we have $\nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t+1|t} = 0$ for all $t \geq 1$.

The proof for the remaining five cases uses similar arguments; we only outline the main steps.

**(2)**. If $j \in I_0$ and $k \in I_1$, then $\nabla_{\theta_j} \bar{f}_{1t} = \nabla_{\theta_j} \bar{f}_{2t}, \nabla_{\theta_k} \bar{f}_{2t} = \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{2t} = \nabla_{\theta_j} \bar{\xi}_{t+1|t} = 0$, implying $\nabla_{\theta_j} \nabla_{\theta_k} \bar{A}_t = \nabla_{\theta_j} \bar{f}_{2t} \nabla_{\theta_k} \bar{\xi}_{t|t-1} + \bar{f}_t \nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t|t-1}$ and $\nabla_{\theta_j} \nabla_{\theta_k} \bar{B}_t = \bar{\xi}_{t|t-1} \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{1t}$. Combining these two equations with (A.2), (A.4) and (A.5), $\nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t+1|t}$ equals

$$\begin{aligned}
&\rho \left\{ \frac{\nabla_{\theta_j} \bar{f}_{2t} \nabla_{\theta_k} \bar{\xi}_{t|t-1} + \bar{f}_t \nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t|t-1}}{\bar{f}_t} - \frac{\nabla_{\theta_j} \bar{f}_{2t}(\bar{\xi}_{t|t-1} - 1)\bar{\xi}_{t|t-1} \nabla_{\theta_k} \bar{f}_{1t}}{\bar{f}_t^2} - \frac{\nabla_{\theta_k} \bar{\xi}_{t|t-1} \nabla_{\theta_j} \bar{f}_{2t}}{\bar{f}_t} \right. \\
&\left. - \frac{(\bar{\xi}_{t|t-1} - 1)\bar{\xi}_{t|t-1} \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{1t}}{\bar{f}_t} + \frac{2(\bar{\xi}_{t|t-1} - 1)\bar{\xi}_{t|t-1} \nabla_{\theta_j} \bar{f}_{2t} \nabla_{\theta_k} \bar{f}_{1t}}{\bar{f}_t^2} \right\}
\end{aligned}$$

The result follows from rearranging the terms.

**(3)**. If $j \in I_0$ and $k \in I_2$, then $\nabla_{\theta_j} \bar{f}_{1t} = \nabla_{\theta_j} \bar{f}_{2t}$ and $\nabla_{\theta_k} \bar{f}_{1t} = \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{1t} = \nabla_{\theta_j} \bar{\xi}_{t+1|t} = 0$, implying $\nabla_{\theta_j} \nabla_{\theta_k} \bar{A}_t = \nabla_{\theta_j} \nabla_{\theta_k} \bar{f}_{2t}(\bar{\xi}_{t|t-1} - 1) + \nabla_{\theta_j} \bar{f}_{2t} \nabla_{\theta_k} \bar{\xi}_{t|t-1} + \bar{f}_t \nabla_{\theta_j} \nabla_{\theta_k} \bar{\xi}_{t|t-1}$ and $\nabla_{\theta_j} \nabla_{\theta_k} \bar{B}_t =$

$(1-\bar{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}$. Combining these results with (A.2), (A.4) and (A.7), $\nabla_{\theta_j}\nabla_{\theta_k}\xi_{t+1|t}$ equals

$$\left\{\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}(\bar{\xi}_{t|t-1}-1)+\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}+\bar{f}_t\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}+\frac{(\bar{\xi}_{t|t-1}-1)^2\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t^2}\right.$$
$$\left.-\frac{[\nabla_{\theta_k}\bar{f}_{2t}(\bar{\xi}_{t|t-1}-1)+\bar{f}_t\nabla_{\theta_k}\bar{\xi}_{t|t-1}]\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t^2}+\frac{(\bar{\xi}_{t|t-1}-1)^2\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t}-\frac{2(\bar{\xi}_{t|t-1}-1)^2\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t^2}\right\}.$$

The result follows from rearranging the terms.

**(4)**. If $j\in I_1$ and $k\in I_1$, then $\nabla_{\theta_j}\bar{f}_{2t}=\nabla_{\theta_k}\bar{f}_{2t}=\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}=0$, implying $\nabla_{\theta_j}\nabla_{\theta_k}\bar{A}_t=\bar{f}_t\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}$ and $\nabla_{\theta_j}\nabla_{\theta_k}\bar{B}_t=\bar{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}+\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}+\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}$. Combining them with (A.2), (A.5), $\nabla_{\theta_j}\nabla_{\theta_k}\xi_{t+1|t}$ equals

$$\rho\left\{\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}-\frac{\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t}-\frac{\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t}\right.$$
$$\left.-\frac{(\bar{\xi}_{t|t-1}-1)[\bar{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}+\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}+\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{1t}]}{\bar{f}_t}+\frac{2(\bar{\xi}_{t|t-1}-1)\bar{\xi}_{t|t-1}^2\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t^2}\right\}.$$

The result follows from rearranging the right hand side terms.

**(5)**. If $j\in I_1$ and $k\in I_2$, then $\nabla_{\theta_j}\bar{f}_{2t}=\nabla_{\theta_k}\bar{f}_{1t}=\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}=\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}=0$, implying $\nabla_{\theta_j}\nabla_{\theta_k}\bar{A}_t=\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}+\bar{f}_t\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}$ and $\nabla_{\theta_j}\nabla_{\theta_k}\bar{B}_t=\nabla_{\theta_j}f_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}-\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}$. Combining them with (A.2), (A.5) and (A.7), $\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t+1|t}$ equals

$$\rho\left\{\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}+\bar{f}_t\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}-\frac{(1-\bar{\xi}_{t|t-1})\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t}-\frac{\left[\nabla_{\theta_k}\bar{f}_{2t}(\bar{\xi}_{t|t-1}-1)+\bar{f}_t\nabla_{\theta_k}\bar{\xi}_{t|t-1}\right]\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t^2}\right.$$
$$\left.-\frac{(\bar{\xi}_{t|t-1}-1)\left[\nabla_{\theta_j}f_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}-\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\right]}{\bar{f}_t}-\frac{2(\bar{\xi}_{t|t-1}-1)^2\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t^2}\right\}.$$

The result follows from rearranging the terms.

**(6)**. If $j\in I_2$ and $k\in I_2$, then $\nabla_{\theta_j}\bar{f}_{1t}=\nabla_{\theta_k}\bar{f}_{1t}=\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}=0$, implying $\nabla_{\theta_j}\nabla_{\theta_k}\bar{A}_t=\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}(\bar{\xi}_{t|t-1}-1)+\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}+\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}+\bar{f}_t\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}$ and $\nabla_{\theta_j}\nabla_{\theta_k}\bar{B}_t=(1-\bar{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}-\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}-\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}$. Combining them with (A.2) and (A.7), $\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t+1|t}$ equals

$$\rho\left\{\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}(\bar{\xi}_{t|t-1}-1)+\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}+\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}+\bar{f}_t\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right.$$
$$-\frac{\left[\nabla_{\theta_k}\bar{f}_{2t}(\bar{\xi}_{t|t-1}-1)+\bar{f}_t\nabla_{\theta_k}\bar{\xi}_{t|t-1}\right](1-\bar{\xi}_{t|t-1})\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t^2}-\frac{\left[\nabla_{\theta_j}\bar{f}_{2t}(\bar{\xi}_{t|t-1}-1)+\bar{f}_t\nabla_{\theta_j}\bar{\xi}_{t|t-1}\right](1-\bar{\xi}_{t|t-1})\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t^2}$$
$$\left.-\frac{(\bar{\xi}_{t|t-1}-1)[(1-\bar{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}-\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}-\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}]}{\bar{f}_t}+2\frac{(\bar{\xi}_{t|t-1}-1)^3\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t^2}\right\}.$$

The result follows from rearranging the terms.

Consider Lemma 1.4. Differentiating (A.8) with respect to $\theta_l$:

$$\begin{aligned}
&\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\xi_{t+1|t} \\
&= \rho\Big\{\frac{\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}A_t}{B_t} - \frac{\nabla_{\theta_j}\nabla_{\theta_k}A_t\nabla_{\theta_l}B_t}{B_t^2} - \frac{\nabla_{\theta_j}\nabla_{\theta_l}A_t\nabla_{\theta_k}B_t}{B_t^2} - \frac{\nabla_{\theta_j}A_t\nabla_{\theta_k}\nabla_{\theta_l}B_t}{B_t^2} + \frac{2\nabla_{\theta_j}A_t\nabla_{\theta_k}B_t\nabla_{\theta_l}B_t}{B_t^3} \\
&-\frac{\nabla_{\theta_k}\nabla_{\theta_l}A_t\nabla_{\theta_j}B_t}{B_t^2} - \frac{\nabla_{\theta_k}A_t\nabla_{\theta_j}\nabla_{\theta_l}B_t}{B_t^2} + \frac{2\nabla_{\theta_k}A_t\nabla_{\theta_j}B_t\nabla_{\theta_l}B_t}{B_t^3} \\
&-\frac{\nabla_{\theta_l}A_t\nabla_{\theta_j}\nabla_{\theta_k}B_t}{B_t^2} - \frac{A_t\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}B_t}{B_t^2} + \frac{2A_t\nabla_{\theta_j}\nabla_{\theta_k}B_t\nabla_{\theta_l}B_t}{B_t^3} \\
&+\frac{2\nabla_{\theta_l}A_t\nabla_{\theta_j}B_t\nabla_{\theta_k}B_t}{B_t^3} + \frac{2A_t\nabla_{\theta_j}\nabla_{\theta_l}B_t\nabla_{\theta_k}B_t}{B_t^3} + \frac{2A_t\nabla_{\theta_j}B_t\nabla_{\theta_k}\nabla_{\theta_l}B_t}{B_t^3} - \frac{6A_t\nabla_{\theta_j}B_t\nabla_{\theta_k}B_t\nabla_{\theta_l}B_t}{B_t^4}\Big\},
\end{aligned}$$

where

$$\begin{aligned}
\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}A_t &= \nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}f_{2t}(\xi_{t|t-1}-1) + \nabla_{\theta_j}\nabla_{\theta_l}f_{2t}\nabla_{\theta_k}\xi_{t|t-1} + \nabla_{\theta_k}\nabla_{\theta_l}f_{2t}\nabla_{\theta_j}\xi_{t|t-1} \\
&+\nabla_{\theta_l}f_{2t}\nabla_{\theta_j}\nabla_{\theta_k}\xi_{t|t-1} + \nabla_{\theta_j}\nabla_{\theta_k}f_{2t}\nabla_{\theta_l}\xi_{t|t-1} + \nabla_{\theta_j}f_{2t}\nabla_{\theta_k}\nabla_{\theta_l}\xi_{t|t-1} \\
&+\nabla_{\theta_k}f_{2t}\nabla_{\theta_j}\nabla_{\theta_l}\xi_{t|t-1} + f_{2t}\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\xi_{t|t-1}, \\
\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}B_t &= (\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}f_{1t} - \nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}f_{2t})\xi_{t|t-1} + (\nabla_{\theta_j}\nabla_{\theta_l}f_{1t} - \nabla_{\theta_j}\nabla_{\theta_l}f_{2t})\nabla_{\theta_k}\xi_{t|t-1} \\
&+(\nabla_{\theta_k}\nabla_{\theta_l}f_{1t} - \nabla_{\theta_k}\nabla_{\theta_l}f_{2t})\nabla_{\theta_j}\xi_{t|t-1} + (\nabla_{\theta_l}f_{1t} - \nabla_{\theta_l}f_{2t})\nabla_{\theta_j}\nabla_{\theta_k}\xi_{t|t-1} \\
&+\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}f_{2t} + (\nabla_{\theta_j}\nabla_{\theta_k}f_{1t} - \nabla_{\theta_j}\nabla_{\theta_k}f_{2t})\nabla_{\theta_l}\xi_{t|t-1} \\
&+(\nabla_{\theta_j}f_{1t} - \nabla_{\theta_j}f_{2t})\nabla_{\theta_k}\nabla_{\theta_l}\xi_{t|t-1} + (\nabla_{\theta_k}f_{1t} - \nabla_{\theta_k}f_{2t})\nabla_{\theta_j}\nabla_{\theta_l}\xi_{t|t-1} \\
&+(f_{1t} - f_{2t})\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\xi_{t|t-1}
\end{aligned}$$

We now evaluate the above terms at $\delta_1 = \delta_2 = \delta$ for 10 possible cases. We only report the values of $\bar{\mathcal{E}}_{jkl,t}$ but omit the derivation details.

(**1**). If $j \in I_0, k \in I_0$ and $l \in I_0$, then $\bar{\mathcal{E}}_{jkl,t} = 0$.

(**2**). If $j \in I_0, k \in I_0$ and $l \in I_1$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$r\left\{-\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^2} - \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^2} - \frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t} + 2\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^3}\right\}.$$

(**3**). If $j \in I_0, k \in I_0$ and $l \in I_2$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$r\left\{\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2} - \frac{\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t} - 2\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^3}\right\}.$$

(**4**). If $j \in I_0, k \in I_1$ and $l \in I_1$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$\begin{aligned}
&\rho(1-2\xi_*)\left[\frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right. \\
&\left.-\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2} - \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right] + r\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t} - \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^2}\right] \\
&-2r\xi_*\left[\frac{\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^2} - 2\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^3}\right]
\end{aligned}$$

($\mathbf{5}$). If $j \in I_0, k \in I_1$ and $l \in I_2$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$\rho(1-2\xi_*)\left[\frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} - \frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right.$$
$$\left. - \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2} + \frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\xi_{t|t-1}}{\bar{f}_t} - \frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$-r(1-2\xi_*)\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t^2} - 2\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^3}\right].$$

($\mathbf{6}$). If $j \in I_0, k \in I_2$ and $l \in I_2$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$-\rho(1-2\xi_*)\left[\frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right.$$
$$\left.-\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2} - \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right] - r\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t} - \frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t}\right].$$
$$+2r(1-\xi_*)\left[\frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2} - 2\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^3}\right]$$

($\mathbf{7}$). If $j \in I_1, k \in I_1$ and $l \in I_1$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$\rho(1-2\xi_*)\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right.$$
$$\left.+\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$-2\rho\left[\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} + \frac{\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$+\rho(6\xi_*^2-4\xi_*)\left[\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2} + \frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}}{\bar{f}_t^2} + \frac{\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right]$$
$$-2r\xi_*\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t^2} + \frac{\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{1t}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t^2}\right] + r\frac{\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t} + 6r\xi_*^2\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{1t}}{\bar{f}_t^3}$$

($\mathbf{8}$). If $j \in I_1, k \in I_1$ and $l \in I_2$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$-\rho(1-2\xi_*)\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t} - \frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{1t}}{\bar{f}_t} - \frac{\nabla_{\theta_k}\nabla_{\theta_l}\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t} - \frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$-\rho(6\xi_*^2-6\xi_*+1)\left[\frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}}{\bar{f}_t^2} + \frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right]$$
$$+2\rho\left[\frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t} - \frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t} - \frac{\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$-r(1-2\xi_*)\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2} + (6\xi_*^2-4\xi_*)\left[\rho\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2} - r\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{1t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^3}\right].$$

(**9**). If $j \in I_1, k \in I_2$ and $l \in I_2$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$-\rho(1-2\xi_*)\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t}+\frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t}-\frac{\nabla_{\theta_k}\nabla_{\theta_l}\bar{\xi}_{t|t-1}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t}+\frac{\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$-\rho(6\xi_*^2-6\xi_*+1)\left[\frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}+\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right]-r(1-2\xi_*)\frac{\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{f}_{1t}}{\bar{f}_t^2}$$
$$+2\rho\left[\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}+\frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}-\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$+[6(1-\xi_*)^2-4(1-\xi_*)]\left[\rho\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}+r\frac{\nabla_{\theta_j}\bar{f}_{1t}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^3}\right].$$

(**10**). If $j \in I_2, k \in I_2$ and $l \in I_2$, then $\bar{\mathcal{E}}_{jkl,t}$ equals

$$-\rho(1-2\xi_*)\left[\frac{\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}+\frac{\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}+\frac{\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right.$$
$$\left.+\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}+\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}+\frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t}\right]$$
$$+2r(1-\xi_*)\left[\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2}+\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^2}+\frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\nabla_{\theta_k}\bar{f}_{2t}}{\bar{f}_t^2}\right]$$
$$+\rho[6(1-\xi_*)^2-4(1-\xi_*)]\left[\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}+\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}+\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right]$$
$$+2\rho\left[\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}+\frac{\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}+\frac{\nabla_{\theta_l}\bar{f}_{2t}\nabla_{\theta_j}\bar{\xi}_{t|t-1}\nabla_{\theta_k}\bar{\xi}_{t|t-1}}{\bar{f}_t^2}\right]$$
$$-r\frac{\nabla_{\theta_j}\nabla_{\theta_k}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t}-6r(1-\xi_*)^2\frac{\nabla_{\theta_j}\bar{f}_{2t}\nabla_{\theta_k}\bar{f}_{2t}\nabla_{\theta_l}\bar{f}_{2t}}{\bar{f}_t^3}. \blacksquare$$

The next lemma provides stochastic bounds for $\bar{\xi}_{t+1|t}$ and its derivatives.

**Lemma A.1** *Suppose Assumption 4 hold. Then, there exists an open neighborhood of $(\beta_*, \delta_*)$, denoted by $B(\beta_*, \delta_*)$, and a sequence of strictly stationary and ergodic random variables $\{\lambda_t\}$ satisfying $E\lambda_t^{1+c} < M < \infty$ for some $c, M > 0$, such that:*

$$\sup_{(\beta,\delta_1)\in B(\beta_*,\delta_*)}\left|\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\bar{\xi}_{t+1|t}\right|^{\frac{\alpha(k)}{k}} < \lambda_t \quad (t = 1, ..., T)$$

*for all $i_1, ..., i_k \in \{1, .., 2n_\delta + n_\beta\}$ and $k = 1, 2, 3$ and $4$, where $\alpha(k) = 6$ if $k = 1, 2, 3$ and $\alpha(k) = 5$ if $k = 4$. The above inequalities hold uniformly over $\epsilon \leq p, q \leq 1 - \epsilon$ with $\epsilon$ being an arbitrary number satisfying $0 < \epsilon < 1/2$.*

**Proof of Lemma A.1**. We use the difference equations in Lemma 1 to relate $\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\bar{\xi}_{t+1|t}$ to the density functions $\bar{f}_{1t}$ and $\bar{f}_{2t}$ and their derivatives. Because the higher order derivatives depend successively on the lower orders, we start with $k = 1$. Without loss of generality, suppose $j \in I_1$. Then, apply (A.6):

$$\left|\nabla_{\theta_j}\bar{\xi}_{t+1|t}\right|^6 \leq \left(\sum_{s=0}^{t-1}\left|r\rho^s\frac{\nabla_{\theta_j}\bar{f}_{1(t-s)}}{\bar{f}_{t-s}}\right|\right)^6 \leq \left(\sum_{s=0}^{\infty}|r\rho^s|\,\upsilon_{t-s}^{1/6}\right)^6 \leq \left(\sum_{s=0}^{\infty}(1-\epsilon)^s\upsilon_{t-s}^{1/6}\right)^6,$$

where the second inequality follows from Assumption 4 and the last inequality uses $\rho = p + q - 1$. Because $\{\upsilon_t\}$ is stationary and ergodic, the right hand side is also stationary and ergodic (White, 2001, Theorem 3.35). Denote it by $\lambda_t$ and apply Minkowski's inequality for an infinite sum:

$$\begin{aligned} E\lambda_t^{1+c} &= E\left[\sum_{s=0}^{\infty}(1-\epsilon)^s\upsilon_{t-s}^{1/6}\right]^{6(1+c)} \leq \left\{\sum_{s=0}^{\infty}\left[E((1-\epsilon)^s\upsilon_{t-s}^{1/6})^{6(1+c)}\right]^{\frac{1}{6(1+c)}}\right\}^{6(1+c)} \\ &= \left\{\sum_{s=0}^{\infty}(1-\epsilon)^s\left[E\upsilon_{t-s}^{1+c}\right]^{\frac{1}{6(1+c)}}\right\}^{6(1+c)} \leq L\left\{\sum_{s=0}^{\infty}(1-\epsilon)^s\right\}^{6(1+c)}, \end{aligned}$$

where the last inequality holds because $E\upsilon_{t-s}^{1+c}$ is finite by Assumption 4. Because $\sum_{s=0}^{\infty}(1-\epsilon)^s = 1/\epsilon < \infty$, we have $E\lambda_t^{1+c} \leq L/\epsilon^{6(1+c)} < \infty$. This establishes the result for $k=1$. Let $M = L/\epsilon^{6(1+c)}$.

The proof for $k > 1$ is similar. For $k=2$, we have $|\nabla_{\theta_j}\nabla_{\theta_i}\bar{\xi}_{t+1|t}|^3 \leq (\sum_{s=0}^{\infty}|\rho^s\bar{\mathcal{E}}_{ji,t-s}|)^3$. We provide upper bounds for $|\bar{\mathcal{E}}_{ji,t}|$ for five possible cases. Specifically, if $j \in I_0$ and $i \in I_1$, then

$$\left|\bar{\mathcal{E}}_{ji,t}\right| = \left|-\frac{r\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t}\frac{\nabla_{\theta_i}\bar{f}_{1t}}{\bar{f}_t} + \frac{r\nabla_{\theta_j}\nabla_{\theta_i}\bar{f}_{1t}}{\bar{f}_t}\right| \leq \left|\frac{r\nabla_{\theta_j}\bar{f}_{2t}}{\bar{f}_t}\frac{\nabla_{\theta_i}\bar{f}_{1t}}{\bar{f}_t}\right| + \left|\frac{r\nabla_{\theta_j}\nabla_{\theta_i}\bar{f}_{1t}}{\bar{f}_t}\right| \leq 2\,|r|\,\upsilon_t^{1/3}.$$

The same bound holds if $j \in I_0$ and $i \in I_2$. If $j \in I_1$ and $i \in I_1$, then $|\bar{\mathcal{E}}_{ji,t}| \leq 2\,|\rho\,(1-2\xi_*)|\,\lambda_{t-1}^{1/6}\upsilon_t^{1/6} + 3\,|r|\,\upsilon_t^{1/3}$. If $j \in I_1$ and $i \in I_2$, then $|\bar{\mathcal{E}}_{ji,t}| \leq 2\,|\rho\,(1-2\xi_*)|\,\lambda_{t-1}^{1/6}\upsilon_t^{1/6} + |r(2\xi_*-1)|\,\upsilon_t^{1/3}$. If $j \in I_2$ and $k \in I_2$, then $|\bar{\mathcal{E}}_{ji,t}| \leq 2\,|\rho\,(1-2\xi_*)|\,\lambda_{t-1}^{1/6}\upsilon_t^{1/6} + (|r| + |2r(\xi_*-1)|)\,\upsilon_t^{1/3}$. Consequently, there exists a finite constant $C_1$, such that for all the five cases we have $|\bar{\mathcal{E}}_{ji,t}| \leq C_1(\lambda_{t-1}^{1/6}\upsilon_t^{1/6} + \upsilon_t^{1/3})$. This implies $\left|\nabla_{\theta_j}\nabla_{\theta_i}\bar{\xi}_{t+1|t}\right|^3 \leq \left(\sum_{s=0}^{\infty}C_1(1-\epsilon)^s(\lambda_{t-1}^{1/6}\upsilon_t^{1/6} + \upsilon_t^{1/3})\right)^3$. The right side is stationary and ergodic; we continue to denote it by $\lambda_t$. By Minkowski's inequality:

$$E\lambda_t^{1+c} \leq \left\{\sum_{s=0}^{\infty}\left[E\left(C_1(1-\epsilon)^s(\lambda_{t-1}^{1/6}\upsilon_t^{1/6} + \upsilon_t^{1/3})\right)^{3(1+c)}\right]^{\frac{1}{3(1+c)}}\right\}^{3(1+c)}. \tag{A.9}$$

Apply Minkowski's inequality followed by the Cauchy–Schwarz inequality to the summands:

$$\begin{aligned} & E\left(C_1(1-\epsilon)^s(\lambda_{t-1}^{1/6}\upsilon_t^{1/6} + \upsilon_t^{1/3})\right)^{3(1+c)} \\ \leq\ & (C_1(1-\epsilon)^s)^{3(1+c)}\left[\left(E\lambda_{t-1}^{(1+c)/2}\upsilon_t^{(1+c)/2}\right)^{\frac{1}{3(1+c)}} + \left(E\upsilon_t^{(1+c)}\right)^{\frac{1}{3(1+c)}}\right]^{3(1+c)} \\ \leq\ & (C_1(1-\epsilon)^s)^{3(1+c)}\left[\left(E\lambda_{t-1}^{(1+c)}E\upsilon_t^{(1+c)}\right)^{\frac{1}{6(1+c)}} + \left(E\upsilon_t^{(1+c)}\right)^{\frac{1}{3(1+c)}}\right]^{3(1+c)}. \end{aligned}$$

Because $E\lambda_{t-1}^{(1+c)} < M$ and $E\upsilon_t^{(1+c)} < L$, the last term in the preceding display is no greater than

$$(1-\epsilon)^{3(1+c)s}C_1^{3(1+c)}\left[(ML)^{\frac{1}{6(1+c)}} + L^{\frac{1}{3(1+c)}}\right]^{3(1+c)} \leq C_2(1-\epsilon)^{3(1+c)s}, \tag{A.10}$$

where $C_2$ is a finite constant independent of $p$ and $q$. Plug this into (A.9), we have $E\lambda_t^{1+c} \leq C_2(\sum_{s=0}^{\infty}(1-\epsilon)^s)^{3(1+c)} = C_2/\epsilon^{3(1+c)} < \infty$. This proves the result for $k=2$.

Now, consider $k=3$. Inspecting the expressions of $\bar{\mathcal{E}}_{jil,t}$ reported in the proof of Lemma 1 shows that they comprise the following terms $(a,b,c=1,2)$:

$$\begin{aligned}&\frac{\nabla_{\theta_j}\nabla_{\theta_i}\bar{f}_{at}\nabla_{\theta_l}\bar{f}_{bt}}{\bar{f}_t^2},\frac{\nabla_{\theta_j}\nabla_{\theta_i}\nabla_{\theta_l}\bar{f}_{at}}{\bar{f}_t},\frac{\nabla_{\theta_j}\bar{f}_{at}\nabla_{\theta_i}\bar{f}_{bt}\nabla_{\theta_l}\bar{f}_{ct}}{\bar{f}_t^3},\frac{\nabla_{\theta_l}\bar{f}_{at}\nabla_{\theta_j}\nabla_{\theta_i}\bar{\xi}_{t|t-1}}{\bar{f}_t},\\&\frac{\nabla_{\theta_j}\bar{f}_{at}\nabla_{\theta_l}\bar{f}_{bt}\nabla_{\theta_i}\bar{\xi}_{t|t-1}}{\bar{f}_t^2},\frac{\nabla_{\theta_j}\nabla_{\theta_i}\bar{f}_{at}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t},\frac{\nabla_{\theta_j}\bar{f}_{at}\nabla_{\theta_i}\bar{\xi}_{t|t-1}\nabla_{\theta_l}\bar{\xi}_{t|t-1}}{\bar{f}_t}.\end{aligned} \tag{A.11}$$

By Assumption 4 and the above results for $k=1$ and 2, the quantities in (A.11) are bounded, respectively, by $\upsilon_t^{1/2},\upsilon_t^{1/2},\upsilon_t^{1/2},\upsilon_t^{1/6}\lambda_{t-1}^{1/3},\upsilon_t^{1/3}\lambda_{t-1}^{1/6},\upsilon_t^{1/3}\lambda_{t-1}^{1/6}$ and $\upsilon_t^{1/6}\lambda_{t-1}^{1/3}$. Therefore, the ten cases specified in Lemma 1 all satisfy $\left|\bar{\mathcal{E}}_{jil,t}\right|\leq C_3(\upsilon_t^{1/2}+\upsilon_t^{1/6}\lambda_{t-1}^{1/3}+\upsilon_t^{1/3}\lambda_{t-1}^{1/6})$, where $C_3$ is a finite constant independent of $p$ and $q$. This implies $\left|\nabla_{\theta_j}\nabla_{\theta_i}\nabla_{\theta_l}\bar{\xi}_{t+1|t}\right|^2\leq\left|\sum_{s=0}^{\infty}(1-\epsilon)^sC_3(\upsilon_t^{1/2}+\upsilon_t^{1/6}\lambda_{t-1}^{1/3}+\upsilon_t^{1/3}\lambda_{t-1}^{1/6})\right|^2$. Denote the right hand side by $\lambda_t$ and proceed along the same lines as between (A.9) and (A.10). It then follows that $E\lambda_t^{1+c}<\infty$. For $k=4$, the expressions of $\bar{\mathcal{E}}_{jilm,t}$, although omitted here, include terms as in (A.11) but with the orders of derivatives sum to 4 instead of 3. Using the same arguments as between (A.9) and (A.10), it can be shown that $E\lambda_t^{1+c}<\infty$ holds. ■

The next lemma establishes stochastic orders of some quantities related to $\xi_{t|t-1}$, $f_{1t}$ and $f_{2t}$. The quantities are all evaluated at $(\widetilde{\beta},\widetilde{\delta},\widetilde{\delta})$.

**Lemma A.2** *Let $i_s,j_s,l_s,m_s,n_s$ be arbitrary integers satisfying $1\leq i_s,j_s,l_s,m_s,n_s\leq 2n_\delta+n_\beta$ for $s\in\{1,2,3,4\}$. The following results hold uniformly over $\epsilon\leq p,q\leq 1-\epsilon$ with $\epsilon$ being an arbitrary number satisfying $0<\epsilon<1/2$:*

1. *For any $a\in\{1,2\},u\in\{1,2,3,4\}$ and $v\in\{0,1,2,3\}$ satisfying $u+v\leq 4$, we have (interpret $\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_v}}\widetilde{\xi}_{t|t-1}$ as 1 when $v=0$)*
$$\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_u}}\widetilde{f}_{at}}{\widetilde{f}_t}\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_v}}\widetilde{\xi}_{t|t-1}=o_p(1), \tag{A.12}$$
*Further, if $u+v\leq 3$, then the result holds with $o_p(1)$ replaced by $O_p(T^{-1/2})$.*

2. *For any $(a,b,c)\in\{1,2\},(u,w)\in\{1,2,3\}$ and $v\in\{0,1,2\}$ satisfying $u+v+w\leq 4$ :*
$$\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_u}}\widetilde{f}_{at}}{\widetilde{f}_t}\frac{\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_v}}\widetilde{f}_{bt}}{\widetilde{f}_t}\frac{\nabla_{\theta_{l_1}}...\nabla_{\theta_{l_w}}\widetilde{f}_{ct}}{\widetilde{f}_t}=O_p(1).$$

3. *For any $(a,b,c)\in\{1,2\},(u,w)\in\{1,2,3\}$ and $(v,z)\in\{0,1\}$ satisfying $u+v+w+z\leq 3$ :*
$$\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\theta_{i_1}}\widetilde{f}_{at}}{\widetilde{f}_t}\frac{\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_u}}\widetilde{f}_{bt}}{\widetilde{f}_t}\frac{\nabla_{\theta_{l_1}}...\nabla_{\theta_{l_v}}\widetilde{f}_{ct}}{\widetilde{f}_t}\nabla_{\theta_{m_1}}...\nabla_{\theta_{m_w}}\widetilde{\xi}_{t|t-1}\nabla_{\theta_{n_1}}...\nabla_{\theta_{n_z}}\widetilde{\xi}_{t|t-1}=O_p(1).$$

**Proof of Lemma A.2**. By the mean value theorem, the left hand side of (A.12) equals

$$\begin{aligned}&T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_u}}f_{at}^*}{f_t^*}\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_v}}\xi_{t|t-1}^*\\&+\left\{T^{-3/2}\sum_{t=1}^{T}\nabla_{\theta'}\left(\frac{\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_u}}\bar{f}_{at}}{\bar{f}_{at}}\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_v}}\bar{\xi}_{t|t-1}\right)\right\}T^{1/2}\left(\widetilde{\theta}-\theta_*\right),\end{aligned} \tag{A.13}$$

where "$*$" and "$-$" denote that the relevant quantities are evaluated at the true values $\theta_*' = (\beta_*', \delta_*', \delta_*')$ and $\bar{\theta}' = \left(\bar{\beta}', \bar{\delta}', \bar{\delta}'\right)$, where $\bar{\theta}$ lies between $\widetilde{\theta}' = (\widetilde{\beta}', \widetilde{\delta}', \widetilde{\delta}')$ and $\theta_*$. The first summation is over terms that are stationary and ergodic, which are bounded by $\lambda_t^{v/\alpha(k)} \upsilon_t^{u/\alpha(k)}$ by Assumption 4 and Lemma A.1. Apply Hölder's inequality:

$$
\begin{aligned}
E(\lambda_t^{v/\alpha(k)} \upsilon_t^{u/\alpha(k)})^{1+c} &\leq \left(E\left(\lambda_t^{v(1+c)/\alpha(k)}\right)^{\alpha(k)/v}\right)^{\frac{v}{\alpha(k)}} \left(E\left(\upsilon_t^{u(1+c)/\alpha(k)}\right)^{\alpha(k)/(\alpha(k)-v)}\right)^{\frac{\alpha(k)-v}{\alpha(k)}} \\
&\leq \left(E\lambda_t^{1+c}\right)^{\frac{v}{\alpha(k)}} \left(E\upsilon_t^{1+c}\right)^{\frac{\alpha(k)-v}{\alpha(k)}}
\end{aligned}
$$

where the last inequality follows because $u + v < \alpha(k)$. Both terms on the right hand side are finite by Assumption 4 and Lemma A.1. Therefore, the first term in the display (A.13) is $o_p(1)$ by Theorem 3.34 in White (2001). Now turn to the second term in the display (A.13). We have, for any $k \in \{1, ..., 2n_\delta + n_\beta\}$ :

$$
\begin{aligned}
&\left| T^{-3/2} \sum_{t=1}^{T} \nabla_{\theta_k} \left( \frac{\nabla_{\theta_{i_1}} ... \nabla_{\theta_{i_u}} \bar{f}_{at}}{\bar{f}_{at}} \nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_v}} \bar{\xi}_{t|t-1} \right) \right| \\
\leq\ & \left| T^{-3/2} \sum_{t=1}^{T} \frac{\nabla_{\theta_{i_1}} ... \nabla_{\theta_{i_u}} \nabla_{\theta_k} \bar{f}_{at}}{\bar{f}_{at}} \nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_v}} \bar{\xi}_{t|t-1} \right| + \left| T^{-3/2} \sum_{t=1}^{T} \frac{\nabla_{\theta_{i_1}} ... \nabla_{\theta_{i_u}} \bar{f}_{at}}{\bar{f}_{at}} \nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_v}} \nabla_{\theta_k} \bar{\xi}_{t|t-1} \right| \\
&+ \left| T^{-3/2} \sum_{t=1}^{T} \frac{\nabla_{\theta_{i_1}} ... \nabla_{\theta_{i_u}} \bar{f}_{at}}{\bar{f}_{at}} \frac{\nabla_{\theta_k} \bar{f}_{at}}{\bar{f}_{at}} \nabla_{\theta_{j_1}} ... \nabla_{\theta_{j_v}} \bar{\xi}_{t|t-1} \right| \\
\leq\ & T^{-3/2} \sum_{t=1}^{T} \left\{ 2\upsilon_t^{(u+1)/\alpha(k)} \lambda_t^{v/\alpha(k)} + \upsilon_t^{u/\alpha(k)} \lambda_t^{(v+1)/\alpha(k)} \right\} = O_p\left(T^{-1/2}\right),
\end{aligned}
$$

where the equality follows from Assumption 4, Lemma A.1 and $u + v + 1 \leq 5$. Therefore, the display (A.13) is $o_p(1)$.

Now we consider the cases with $u+v \leq 3$. If $u+v < 3$, then the terms inside the first summation of (A.13) are bounded by $\lambda_t^{v/6} \upsilon_t^{u/6}$. We have

$$
E(\lambda_t^{v/6} \upsilon_t^{u/6})^{2(1+c)} \leq \left(E(\lambda_t^{v(1+c)/3})^{\frac{3}{v}}\right)^{v/3} \left(E(\upsilon_t^{u(1+c)/3})^{\frac{3}{(3-v)}}\right)^{(3-v)/3} \leq \left(E\lambda_t^{1+c}\right)^{v/3} \left(E\upsilon_t^{1+c}\right)^{(3-v)/3}.
$$

The right hand side is finite. If $u+v = 3$, i.e., $u = 3$ and $v = 0$, then $E(\lambda_t^{v/6} \upsilon_t^{u/6})^{2(1+c)} = E\upsilon_t^{(1+c)} < \infty$. Apply the central limit theorem; it follows that the left hand side of (A.12) is $O_p\left(T^{-1/2}\right)$.

Lemma A.2.2 and A.2.3 can be proved using the same arguments, i.e., first applying the mean value theorem and then obtaining bounds for the two resulting terms separately. It follows that the left hand side quantity in Lemma A.2.2 is bounded by $T^{-1} \sum_{t=1}^{T} \upsilon_t^{(u+v+w)/\alpha(k)} + O_p(T^{-1/2})$, while that in Lemma A.2.3 is bounded by $T^{-1} \sum_{t=1}^{T} \upsilon_t^{(1+u+v)/\alpha(k)} \lambda_t^{(w+z)/\alpha(k)} + O_p(T^{-1/2})$. The two leading terms both satisfy a law of large numbers, therefore are $O_p(1)$. ■

We state some notations to be used in subsequent proofs. Define

$$
\hat{\theta}(\delta_2) = (\hat{\beta}(\delta_2)', \hat{\delta}_1(\delta_2)', \delta_2')',
$$

where $\hat{\beta}(\delta_2)$ and $\hat{\delta}_1(\delta_2)$ are defined in (17). Let $\hat{\xi}_{t+1|t}, \hat{f}_{1t}$ and $\hat{f}_{2t}$ denote $\xi_{t+1|t}(p,q,\beta,\delta_1,\delta_2)$, $f_t(\beta,\delta_1)$ and $f_t(\beta,\delta_2)$ evaluated at $(\beta,\delta_1,\delta_2) = (\hat{\beta}(\delta_2),\hat{\delta}_1(\delta_2),\delta_2)$. Also, let $\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\hat{\xi}_{t+1|t}$, $\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\hat{f}_{1t}$ and $\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\hat{f}_{2t}$ denote the $k$-th order derivatives of $\xi_{t+1|t}$, $f_{1t}$ and $f_{2t}$ with respect to the $i_1$-th,...$i_k$-th elements of $\theta$ evaluated at $(\hat{\beta}(\delta_2),\hat{\delta}_1(\delta_2),\delta_2)$.

**Proof of Lemma 2.** As the proof is long, we organize it into three parts, corresponding to Lemma 2.1, 2.2 and 2.3 respectively.

**Proof of the first result in Lemma 2.** By construction, $\hat{\theta}(\delta_2)$ satisfies

$$\mathcal{M}_j^{(1)}(p,q,\delta_2) = T^{-1}\sum_{t=1}^{T}\frac{\hat{M}_{jt}}{\hat{B}_t} = 0 \quad (j = 1,...,n_\beta + n_\delta), \tag{A.14}$$

where

$$\begin{aligned} \hat{B}_t &= (\hat{f}_{1t} - \hat{f}_{2t})\hat{\xi}_{t|t-1} + \hat{f}_{2t}, \\ \hat{M}_{jt} &= (\nabla_{\theta_j}\hat{f}_{1t} - \nabla_{\theta_j}\hat{f}_{2t})\hat{\xi}_{t|t-1} + (\hat{f}_{1t} - \hat{f}_{2t})\nabla_{\theta_j}\hat{\xi}_{t|t-1} + \nabla_{\theta_j}\hat{f}_{2t}. \end{aligned} \tag{A.15}$$

Because (A.14) holds for all $\delta_2 \in \Delta$, its derivatives with respect to $\delta_2$ must equal zero. The proof makes use of this property. It proceeds in three steps. For an arbitrary $k \in \{1,...,n_\delta\}$, the first step differentiates the $n_\beta + n_\delta$ equations in (A.14) with respect to $\delta_{2k}$ to obtain a system of $n_\beta + n_\delta$ linear equations, with $\nabla_{\delta_{2k}}\hat{\beta}(\delta_2)$ and $\nabla_{\delta_{2k}}\hat{\delta}_1(\delta_2)$ being the unknowns. The second step evaluates these equations at $\delta_2 = \widetilde{\delta}$ and provides approximations to them. The third step solves these approximating equations to obtain explicit expressions for $\nabla_{\delta_{2k}}\hat{\beta}(\widetilde{\delta})$ and $\nabla_{\delta_{2k}}\hat{\delta}_1(\widetilde{\delta})$. These three steps are then repeated for all $k \in \{1,...,n_\delta\}$ to prove Lemma 2.1. The idea of differentiating the first order conditions is inspired by Cho and White (2007). At the same time the proof here is more complex due to the presence of $\xi_{t+1|t}$ and the allowance for multiple switching parameters.

*Step 1 for proving Lemma 2.1.* Consider an arbitrary $k \in \{1,...,n_\delta\}$ and an arbitrary $j \in \{1,...,n_\beta + n_\delta\}$. Taking the first order derivative of the $j$-th equation (A.14) with respect to the $\delta_{2k}$ (Here, view $\hat{B}_t$ and $\hat{M}_{jt}$ as functions of $p,q$ and $\delta_2$; note that $\beta$ and $\delta_1$ are now functions of these three elements.):

$$\mathcal{M}_{jk}^{(2)}(p,q,\delta_2) = \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{M}_{jt}}{\hat{B}_t} - \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t^2}\hat{M}_{jt} = 0, \tag{A.16}$$

where

$$\begin{aligned} \nabla_{\delta_{2k}}\hat{M}_{jt} = & \Big\{\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{1t} + (1-\hat{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{2t} + (\nabla_{\theta_j}\hat{f}_{1t} - \nabla_{\theta_j}\hat{f}_{2t})\nabla_{\theta'}\hat{\xi}_{t|t-1} \\ & +(\nabla_{\theta_j}\xi_{t|t-1})(\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta'}\hat{f}_{2t}) + (\hat{f}_{1t} - \hat{f}_{2t})(\nabla_{\theta_j}\nabla_{\theta'}\hat{\xi}_{t|t-1})\Big\}\nabla_{\delta_{2k}}\hat{\theta}(\delta_2), \end{aligned} \tag{A.17}$$

and

$$\nabla_{\delta_{2k}}\hat{B}_t = \Big\{\hat{\xi}_{t|t-1}\nabla_{\theta'}\hat{f}_{1t} + (1-\hat{\xi}_{t|t-1})\nabla_{\theta'}\hat{f}_{2t} + (\hat{f}_{1t} - \hat{f}_{2t})\nabla_{\theta'}\hat{\xi}_{t|t-1}\Big\}\nabla_{\delta_{2k}}\hat{\theta}(\delta_2) \tag{A.18}$$

with

$$\nabla_{\delta_{2k}}\hat{\theta}(\delta_2) = \begin{bmatrix} \nabla_{\delta_{2k}}\hat{\beta}(\delta_2) \\ \nabla_{\delta_{2k}}\hat{\delta}_1(\delta_2) \\ e_k \end{bmatrix}, \tag{A.19}$$

where $e_k$ is an $n_\delta$-dimensional vector whose $k$-th element equals 1 and otherwise zero. We view (A.16) as a linear equation with the first $(n_\beta+n_\delta)$ elements of $\nabla_{\delta_{2k}}\hat{\theta}(\delta_2)$ being the unknowns. The above differentiation can be carried for all $j=1,...,n_\beta+n_\delta$, while keeping $k$ fixed at the same value. This delivers $n_\beta+n_\delta$ equations with the same number of unknowns specified in (A.19).

*Step 2 for proving Lemma 2.1.* We first evaluate $T^{-1}\sum_{t=1}^{T}(\nabla_{\delta_{2k}}\hat{B}_t/\hat{B}_t^2)\hat{M}_{jt}$ in (A.16) at $\delta_2=\widetilde{\delta}$ for an arbitrary $j\in\{1,...,n_\beta+n_\delta\}$. It equals (using $\widetilde{f}_{1t}=\widetilde{f}_{2t}=\widetilde{f}_t$ and $\widetilde{\xi}_{t|t-1}=\xi_*$)

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}[\xi_*\nabla_{\theta'}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta'}\widetilde{f}_{2t}]\nabla_{\delta_{2k}}\hat{\theta}(\widetilde{\delta}).$$

Using (A.19), this can be rewritten as

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}\left[\begin{array}{cc}\nabla_{\beta'}\widetilde{f}_{1t} & \nabla_{\delta_1'}\widetilde{f}_{1t}\end{array}\right]\left[\begin{array}{c}\nabla_{\delta_{2k}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\hat{\delta}_1(\widetilde{\delta})\end{array}\right]\\ &+\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}(1-\xi_*)\nabla_{\delta_{2k}}\widetilde{f}_{2t},\end{aligned}$$

where $\nabla_{\beta'}\widetilde{f}_{1t}$ denotes the derivative of $f_t(\beta,\delta_1)$ with respect to $\beta$ evaluated at $\hat{\beta}(\widetilde{\delta})$ and $\hat{\delta}_1(\widetilde{\delta})$; $\nabla_{\delta_1'}\widetilde{f}_{1t}$ and $\nabla_{\delta_{2k}}\widetilde{f}_{2t}$ are defined analogously. Further, if $j\in\{1,...,n_\beta\}$, then preceding display equals (using $\nabla_{\theta_j}\widetilde{f}_{1t}=\nabla_{\theta_j}\widetilde{f}_{2t}$)

$$\frac{1}{T}\sum_{t=1}^{T}\left[\begin{array}{cc}\frac{\nabla_{\theta_j}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t} & \frac{\nabla_{\theta_j}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\end{array}\right]\left[\begin{array}{c}\nabla_{\delta_{2k}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\hat{\delta}_1(\widetilde{\delta})\end{array}\right]+\frac{1}{T}(1-\xi_*)\sum_{t=1}^{T}\frac{\nabla_{\theta_j}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{2k}}\widetilde{f}_{2t}}{\widetilde{f}_t}. \tag{A.20}$$

Meanwhile, if $j\in\{n_\beta+1,...,n_\beta+n_\delta\}$, then the same display equals (using $\nabla_{\theta_j}\widetilde{f}_{2t}=0$) $\xi_*$ times (A.20). Let $D$ be a diagonal matrix whose first $n_\beta$ diagonal elements equal 1 and the rest $\xi_*$. Then the above two cases for $j$ can be combined, leading to

$$D\widetilde{I}\left[\begin{array}{c}\nabla_{\delta_{2k}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\hat{\delta}_1(\widetilde{\delta})\end{array}\right]+D\left[\begin{array}{c}(1-\xi_*)\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\beta}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{2k}}\widetilde{f}_{2t}}{\widetilde{f}_t}\\ (1-\xi_*)\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_1}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{2k}}\widetilde{f}_{2t}}{\widetilde{f}_t}\end{array}\right], \tag{A.21}$$

where $\widetilde{I}$ is defined in (20).

Now consider the first term in (A.16). It equals (using $\widetilde{f}_{1t}=\widetilde{f}_{2t}=\widetilde{f}_t$ and $\widetilde{\xi}_{t|t-1}=\xi_*$)

$$\left\{\frac{1}{T}\sum_{t=1}^{T}\left[\frac{\xi_*\nabla_{\theta_j}\nabla_{\theta'}\widetilde{f}_{1t}}{\widetilde{f}_t}+\frac{(1-\xi_*)\nabla_{\theta_j}\nabla_{\theta'}\widetilde{f}_{2t}}{\widetilde{f}_t}+\frac{\nabla_{\theta_j}\widetilde{f}_{1t}-\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t}\nabla_{\theta'}\widetilde{\xi}_{t|t-1}+\nabla_{\theta_j}\widetilde{\xi}_{t|t-1}\frac{\nabla_{\theta'}\widetilde{f}_{1t}-\nabla_{\theta'}\widetilde{f}_{2t}}{\widetilde{f}_t}\right]\right\}\nabla_{\delta_{2k}}\hat{\theta}(\widetilde{\delta}).$$

All the terms inside the curly brackets are $O_p(T^{-1/2})$ by Lemma A.2.1. Their effects are dominated by $\widetilde{I}$, which is positive definite in large samples. Combining this with (A.21) and (A.16), we have:

$$\widetilde{I}\left[\begin{array}{c}\nabla_{\delta_{2k}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\hat{\delta}_1(\widetilde{\delta})\end{array}\right]=-\left[\begin{array}{c}(1-\xi_*)\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\beta}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{2k}}\widetilde{f}_{2t}}{\widetilde{f}_t}\\ (1-\xi_*)\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_1}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{2k}}\widetilde{f}_{2t}}{\widetilde{f}_t}\end{array}\right]+O_p(T^{-1/2}). \tag{A.22}$$

The preceding display provides $(n_\beta + n_\delta)$ linear equations with the same number of unknowns.

*Step 3 for proving Lemma 2.1.* We show how to solve (A.22) for $k = n_\delta$. Consider the following partition of the system (A.22) with $\widetilde{I}_{22}, \widetilde{\phi}_2$ and $\widetilde{B}_2$ being scalars:

$$\begin{bmatrix} \widetilde{I}_{11} & \widetilde{I}_{12} \\ \widetilde{I}_{21} & \widetilde{I}_{22} \end{bmatrix} \begin{bmatrix} \widetilde{\phi}_1 \\ \widetilde{\phi}_2 \end{bmatrix} = \begin{bmatrix} \widetilde{B}_1 \\ \widetilde{B}_2 \end{bmatrix} + O_p(T^{-1/2}).$$

This implies

$$\begin{bmatrix} \widetilde{I}_{11} & \widetilde{I}_{12} \\ 0 & \widetilde{I}_{22} - \widetilde{I}_{21}\widetilde{I}_{11}^{-1}\widetilde{I}_{12} \end{bmatrix} \begin{bmatrix} \widetilde{\phi}_1 \\ \widetilde{\phi}_2 \end{bmatrix} = \begin{bmatrix} \widetilde{B}_1 \\ \widetilde{B}_2 - \widetilde{I}_{21}\widetilde{I}_{11}^{-1}\widetilde{B}_1 \end{bmatrix} + O_p(T^{-1/2}), \tag{A.23}$$

which further implies $\widetilde{\phi}_2 = [\widetilde{B}_2 - \widetilde{I}_{21}\widetilde{I}_{11}^{-1}\widetilde{B}_1]/[\widetilde{I}_{22} - \widetilde{I}_{21}\widetilde{I}_{11}^{-1}\widetilde{I}_{12}] + O_p(T^{-1/2})$. Because $\widetilde{B}_1 = (\xi_* - 1)\,\widetilde{I}_{12}$ and $\widetilde{B}_2 = (\xi_* - 1)\,\widetilde{I}_{22}$, after cancellation we have $\widetilde{\phi}_2 = \xi_* - 1 + O_p(T^{-1/2})$. Plugging this result into the first set of equations in (A.23), we obtain $\widetilde{\phi}_1 = \widetilde{I}_{11}^{-1}\widetilde{B}_1 - \widetilde{I}_{11}^{-1}\widetilde{I}_{12}\widetilde{\phi}_2 + O_p(T^{-1/2}) = (\xi_* - 1)\,\widetilde{I}_{11}^{-1}\widetilde{I}_{12} - \widetilde{I}_{11}^{-1}\widetilde{I}_{12}\left[(\xi_* - 1) + O_p(T^{-1/2})\right] + O_p(T^{-1/2}) = O_p(T^{-1/2})$. This completes the proof for the case $k = n_\delta$. For other values of $k$, the same argument can be used after exchanging the $k$- and $n_\delta$-th columns of $\widetilde{I}$ and the $k$- and $n_\delta$-th elements of $\widetilde{\phi}$ and $\widetilde{B}$. ■

**Proof of the second result in Lemma 2.** View the quantities in (A.16) as functions of $\delta_2$, $p$ and $q$ and differentiate them with respect to the $l$-th element of $\delta_2$ $(l = 1, ..., n_\delta)$ :

$$\begin{aligned} \mathcal{M}_{jkl}^{(3)}(p, q, \delta_2) &= \frac{1}{T}\sum_{t=1}^{T}\left\{ \frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{M}_{jt}}{\hat{B}_t} - \frac{\nabla_{\delta_{2k}}\hat{M}_{jt}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2} - \frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2}\hat{M}_{jt} \right. \\ &\qquad \left. - \frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t^2}\nabla_{\delta_{2l}}\hat{M}_{jt} + 2\frac{\nabla_{\delta_{2k}}\hat{B}_t\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^3}\hat{M}_{jt} \right\} = 0, \end{aligned} \tag{A.24}$$

where

$$\begin{aligned} \nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{M}_{jt} &= \sum_{s=1}^{n_\beta+2n_\delta}\left\{ \nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{1t} + \hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t} \right. \\ &\quad - \nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{2t} + (1 - \hat{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t} \\ &\quad + (\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1} + (\nabla_{\theta_j}\hat{f}_{1t} - \nabla_{\theta_j}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\ &\quad + (\nabla_{\theta_s}\hat{f}_{1t} - \nabla_{\theta_s}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta'}\hat{\xi}_{t|t-1} + (\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\hat{\xi}_{t|t-1} \\ &\quad \left. + (\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\hat{\xi}_{t|t-1} + (\hat{f}_{1t} - \hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1} \right\} \times \\ &\quad \nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\hat{\theta}(\delta_2) \\ &\quad + \left\{ \hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{1t} + (1 - \hat{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{2t} + (\nabla_{\theta_j}\hat{f}_{1t} - \nabla_{\theta_j}\hat{f}_{2t})\nabla_{\theta'}\hat{\xi}_{t|t-1} \right. \\ &\quad \left. + (\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\hat{\xi}_{t|t-1} + (\hat{f}_{1t} - \hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta'}\hat{\xi}_{t|t-1} \right\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}(\delta_2), \end{aligned} \tag{A.25}$$

and

$$\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t = \sum_{s=1}^{n_\beta+2n_\delta}\Big\{\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{1t}+\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{2t} \tag{A.26}$$

$$+(1-\hat{\xi}_{t|t-1})\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Big\}$$

$$\times\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\hat{\theta}(\delta_2)$$

$$+\Big\{\hat{\xi}_{t|t-1}\nabla_{\theta'}\hat{f}_{1t}+(1-\hat{\xi}_{t|t-1})\nabla_{\theta'}\hat{f}_{2t}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta'}\hat{\xi}_{t|t-1}\Big\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}(\delta_2).$$

We now apply (A.17), (A.18), (A.25) and (A.26) to analyze the five terms in (A.24). Start with the third term $T^{-1}\sum_{t=1}^{T}[\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t/\hat{B}_t^2]\hat{M}_{jt}$. At $\delta_2=\widetilde{\theta}$, it equals

$$\sum_{u=1}^{n_\beta+2n_\delta}\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}[\nabla_{\theta_u}\widetilde{\xi}_{t|t-1}\nabla_{\theta_s}\widetilde{f}_{1t}+\xi_*\nabla_{\theta_s}\nabla_{\theta_u}\widetilde{f}_{1t}-\nabla_{\theta_u}\widetilde{\xi}_{t|t-1}\nabla_{\theta_s}\widetilde{f}_{2t}$$

$$+(1-\xi_*)\nabla_{\theta_s}\nabla_{\theta_u}\widetilde{f}_{2t}+(\nabla_{\theta_u}\widetilde{f}_{1t}-\nabla_{\theta_u}\widetilde{f}_{2t})\nabla_{\theta_s}\widetilde{\xi}_{t|t-1}]\Bigg\}\nabla_{\delta_{2l}}\hat{\theta}_u(\widetilde{\delta})\nabla_{\delta_{2k}}\hat{\theta}_s(\widetilde{\delta})$$

$$+\sum_{s=1}^{n_\beta+n_\delta}\Bigg\{\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}[\xi_*\nabla_{\theta_s}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_s}\widetilde{f}_{2t}]\Bigg\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}_s(\widetilde{\delta}).$$

Because $\nabla_{\delta_{2l}}\hat{\theta}_u(\widetilde{\delta})$ and $\nabla_{\delta_{2k}}\hat{\theta}_s(\widetilde{\delta})$ are $O_p\left(T^{-1/2}\right)$ except when $s\in\{n_\beta+k,n_\beta+n_\delta+k\}$ and $u\in\{n_\beta+l,n_\beta+n_\delta+l\}$, the preceding display equals

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}\{\nabla_{\delta_{1l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1k}}\widetilde{f}_{1t}+\xi_*\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}-\nabla_{\delta_{1l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1k}}\widetilde{f}_{2t}$$

$$+(1-\xi_*)\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{2t}+(\nabla_{\delta_{1l}}\widetilde{f}_{1t}-\nabla_{\delta_{1l}}\widetilde{f}_{2t})\nabla_{\delta_{1k}}\widetilde{\xi}_{t|t-1}\}\nabla_{\delta_{2l}}\hat{\delta}_{1l}(\widetilde{\delta})\nabla_{\delta_{2k}}\hat{\delta}_{1k}(\widetilde{\delta})$$

$$+\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}\{\nabla_{\delta_{2l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1k}}\widetilde{f}_{1t}+\xi_*\nabla_{\delta_{1k}}\nabla_{\delta_{2l}}\widetilde{f}_{1t}-\nabla_{\delta_{2l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1k}}\widetilde{f}_{2t}$$

$$+(1-\xi_*)\nabla_{\delta_{1k}}\nabla_{\delta_{2l}}\widetilde{f}_{2t}+(\nabla_{\delta_{2l}}\widetilde{f}_{1t}-\nabla_{\delta_{2l}}\widetilde{f}_{2t})\nabla_{\delta_{1k}}\widetilde{\xi}_{t|t-1}\}\nabla_{\delta_{2l}}\hat{\delta}_{2l}(\widetilde{\delta})\nabla_{\delta_{2k}}\hat{\delta}_{1k}(\widetilde{\delta})$$

$$+\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}\{\nabla_{\delta_{1l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{2k}}\widetilde{f}_{1t}+\xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}-\nabla_{\delta_{1l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{2k}}\widetilde{f}_{2t}$$

$$+(1-\xi_*)\nabla_{\delta_{2k}}\nabla_{\delta_{1l}}\widetilde{f}_{2t}+(\nabla_{\delta_{1l}}\widetilde{f}_{1t}-\nabla_{\delta_{1l}}\widetilde{f}_{2t})\nabla_{\delta_{2k}}\widetilde{\xi}_{t|t-1}\}\nabla_{\delta_{2l}}\hat{\delta}_{1l}(\widetilde{\delta})\nabla_{\delta_{2k}}\hat{\delta}_{2k}(\widetilde{\delta})$$

$$+\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}\{\nabla_{\delta_{2l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{2k}}\widetilde{f}_{1t}+\xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\widetilde{f}_{1t}-\nabla_{\delta_{2l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{2k}}\widetilde{f}_{2t}$$

$$+(1-\xi_*)\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\widetilde{f}_{2t}+(\nabla_{\delta_{2l}}\widetilde{f}_{1t}-\nabla_{\delta_{2l}}\widetilde{f}_{2t})\nabla_{\delta_{2k}}\widetilde{\xi}_{t|t-1}\}\nabla_{\delta_{2l}}\hat{\delta}_{2l}(\widetilde{\delta})\nabla_{\delta_{2k}}\hat{\delta}_{2k}(\widetilde{\delta})$$

$$+\sum_{s=1}^{n_\beta+n_\delta}\Bigg\{\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}[\xi_*\nabla_{\theta_s}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_s}\widetilde{f}_{2t}]\Bigg\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}_s(\widetilde{\delta})+O_p(T^{-1/2}).$$

Apply $\nabla_{\delta_{2l}}\hat{\delta}_{1l}(\widetilde{\delta}) = (\xi_*-1)/\xi_*+O_p\left(T^{-1/2}\right)$, $\nabla_{\delta_{2k}}\hat{\delta}_{1k}(\widetilde{\delta}) = (\xi_*-1)/\xi_*+O_p\left(T^{-1/2}\right)$ and $\nabla_{\delta_{2l}}\hat{\delta}_{2l}(\widetilde{\delta}) = \nabla_{\delta_{2l}}\hat{\delta}_{2l}(\widetilde{\delta}) = 1$ and rearrange the terms, the preceding display reduces to

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}\left\{\left(\frac{1-\xi_*}{\xi_*}\right)\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}\right. \tag{A.27}$$
$$\left.+\frac{1}{\xi_*^2}\nabla_{\delta_{1l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1k}}\widetilde{f}_{1t}+\frac{1}{\xi_*^2}\nabla_{\delta_{1l}}\widetilde{f}_{1t}\nabla_{\delta_{1k}}\widetilde{\xi}_{t|t-1}\right\}$$
$$+\sum_{s=1}^{n_\beta+n_\delta}\left\{\frac{1}{T}\sum_{t=1}^{T}\frac{\xi_*\nabla_{\theta_j}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_j}\widetilde{f}_{2t}}{\widetilde{f}_t^2}[\xi_*\nabla_{\theta_s}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_s}\widetilde{f}_{2t}]\right\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}_s(\widetilde{\delta})+O_p(T^{-1/2}).$$

As in the proof of Lemma2.1, the above display leads to $(n_\beta+n_\delta)$ equations with $j$ taking values between 1 and $(n_\beta+n_\delta)$. These equations can be written collectively as

$$D\widetilde{I}\begin{bmatrix}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix}+D\begin{bmatrix}\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_\beta\widetilde{f}_{1t}}{\widetilde{f}_t}\widetilde{U}_{kl,t}\\ \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_1}\widetilde{f}_{1t}}{\widetilde{f}_t}\widetilde{U}_{kl,t}\end{bmatrix}+O_p(T^{-1/2}).$$

This completes the analysis for the third term in (A.24). Below we show the other terms in (A.24) are all asymptotically negligible.

Consider the first term in (A.24). Applying the expression (A.25) to (A.24) leads to quantities of the following form: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_u}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_v}}\widetilde{\xi}_{t|t-1})$, where $a\in\{1,2\}, u\in\{1,2,3\}$ and $v\in\{0,1,2\}$ with $1\le u+v\le 3$. They are all $O_p(T^{-1/2})$ because of Lemma A.2.1. Therefore, this term is negligible. Consider the second term in (A.24). At $\delta_2=\widetilde{\delta}$, $\nabla_{\delta_{2k}}\hat{B}_t$ can be rewritten as

$$\sum_{s=1}^{n_\beta}\nabla_{\theta_s}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\hat{\theta}_s(\widetilde{\delta})+\sum_{s=n_\beta+1}^{n_\beta+n_\delta}\xi_*\nabla_{\theta_s}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\hat{\theta}_s(\widetilde{\delta}) \tag{A.28}$$
$$+\sum_{s=n_\beta+n_\delta+1}^{n_\beta+2n_\delta}(1-\xi_*)\nabla_{\theta_s}\widetilde{f}_{2t}\nabla_{\delta_{2k}}\hat{\theta}_s(\widetilde{\delta})$$
$$=\sum_{s=1}^{n_\beta}\nabla_{\beta_s}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\hat{\beta}_s(\widetilde{\delta})+\sum_{s=1,s\neq k}^{n_\delta}\xi_*\nabla_{\delta_{1s}}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\hat{\delta}_{1s}(\widetilde{\delta})+\nabla_{\delta_{1k}}\widetilde{f}_{1t}\left(\xi_*\nabla_{\delta_{2k}}\hat{\delta}_{1k}(\widetilde{\delta})+(1-\xi_*)\right).$$

The preceding display is $O_p(T^{-1/2})$ because $\nabla_{\delta_{2k}}\hat{\beta}_s(\widetilde{\delta})=O_p(T^{-1/2})$ and $\nabla_{\delta_{2k}}\hat{\delta}_{1s}(\widetilde{\delta})=O_p(T^{-1/2})$ for $s\neq k$, and $\xi_*\nabla_{\delta_{2k}}\hat{\delta}_{1k}(\widetilde{\delta})+(1-\xi_*)=O_p(T^{-1/2})$. Therefore, the second term in (A.24) is also negligible. The fourth and fifth terms are also $O_p(T^{-1/2})$ after applying (A.28) to $\nabla_{\delta_{2k}}\hat{B}_t$.

Combining the above results for the five terms, we can rewrite (A.24) as

$$\widetilde{I}\begin{bmatrix}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix}=-\begin{bmatrix}\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_\beta\widetilde{f}_{1t}}{\widetilde{f}_t}\widetilde{U}_{kl,t}\\ \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_1}\widetilde{f}_{1t}}{\widetilde{f}_t}\widetilde{U}_{kl,t}\end{bmatrix}+O_p\left(T^{-1/2}\right).$$

Dividing both sides by $\widetilde{I}$ leads to the desired result. ■

**Proof of the third result in Lemma 2.** View the quantities in (A.24) as functions of $\delta_2$, $p$ and $q$ and differentiate them with respect to the $h$-th element of $\delta_2$ ($h = 1, ..., n_\delta$):

$$
\begin{aligned}
\mathcal{M}_{jklh}^{(4)}(p,q,\delta_2) \quad = \quad & \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{M}_{jt}}{\hat{B}_t} - \frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{M}_{jt}\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^2}\right. \qquad \text{(A.29)}\\
& -\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2h}}\hat{M}_{jt}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2} - \frac{\nabla_{\delta_{2k}}\hat{M}_{jt}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^2} + 2\frac{\nabla_{\delta_{2k}}\hat{M}_{jt}\nabla_{\delta_{2l}}\hat{B}_t\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^3}\\
& -\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^2}\hat{M}_{jt} - \frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2}\nabla_{\delta_{2h}}\hat{M}_{jt} + 2\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^3}\hat{M}_{jt}\\
& -\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^2}\nabla_{\delta_{2l}}\hat{M}_{jt} - \frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t^2}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{M}_{jt} + 2\frac{\nabla_{\delta_{2k}}\hat{B}_t\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^3}\nabla_{\delta_{2l}}\hat{M}_{jt}\\
& +2\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2h}}\hat{B}_t\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^3}\hat{M}_{jt} + 2\frac{\nabla_{\delta_{2k}}\hat{B}_t\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^3}\hat{M}_{jt}\\
& \left.+2\frac{\nabla_{\delta_{2k}}\hat{B}_t\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^3}\nabla_{\delta_{2h}}\hat{M}_{jt} - 6\frac{\nabla_{\delta_{2k}}\hat{B}_t\nabla_{\delta_{2l}}\hat{B}_t\nabla_{\delta_{2h}}\hat{B}_t}{\hat{B}_t^4}\hat{M}_{jt}.\right\} = 0.
\end{aligned}
$$

Among the fifteen terms, only the 1st and the 6th term involve third order derivatives. They will be analyzed later. Among the remaining terms, we have the following five cases: (1) The 4th, 7th and 9th terms involve second order derivatives of $\hat{B}_t$ and first order derivatives of $\hat{M}_{jt}$, which lead to: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\widetilde{f}_{bt}/\widetilde{f}_t)$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)\nabla_{\theta_{j_1}}\widetilde{\xi}_{t|t-1}\nabla_{\theta_{j_2}}\widetilde{\xi}_{t|t-1}$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\widetilde{f}_{bt}/\widetilde{f}_t)$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)\nabla_{\theta_{j_2}}\widetilde{\xi}_{t|t-1}$ and $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\widetilde{f}_{at}/\widetilde{f}_t)\nabla_{\theta_{j_1}}\widetilde{\xi}_{t|t-1}$, where $1 \le i_1, i_2, j_1, j_2 \le n_\beta + 2n_\delta$, $a = 1, 2$ and $b = 1, 2$. They are all $O_p(1)$ by Lemma A.2. (2) The 2rd, 3rd and 10th terms consist of first order derivatives of $\hat{B}_t$ and second order derivatives of $\hat{M}_{jt}$. They lead to: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\nabla_{\theta_{j_3}}\widetilde{f}_{bt}/\widetilde{f}_t)$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)\nabla_{\theta_{j_2}}\widetilde{\xi}_{t|t-1}$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\widetilde{\xi}_{t|t-1}$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)$ and $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)\nabla_{\theta_{j_2}}\widetilde{\xi}_{t|t-1}$, which are all $O_p(1)$. These three terms are thus $O_p(T^{-1/2})$ after applying (A.28) to the first order derivatives of $\hat{B}_t$. (3) The 5th, 11th and 14th terms consist of: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\widetilde{f}_{ct}/\widetilde{f}_t)$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{i_3}}\widetilde{f}_{ct}/\widetilde{f}_t)\nabla_{\theta_{j_1}}\widetilde{\xi}_{t|t-1}$, which are all $O_p(1)$. Consequently, these three terms are $O_p(T^{-1/2})$ after applying (A.28). (4) The 8th, 12th and 13th terms lead to: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{j_2}}\widetilde{f}_{ct}/\widetilde{f}_t)$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\widetilde{f}_{ct}/\widetilde{f}_t)$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{i_3}}\widetilde{f}_{ct}/\widetilde{f}_t)\nabla_{\theta_{j_1}}\widetilde{\xi}_{t|t-1}$, which are all $O_p(1)$. (5) The 15th term consists of $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{i_3}}\widetilde{f}_{ct}/\widetilde{f}_t)(\nabla_{\theta_{i_4}}\widetilde{f}_{ct}/\widetilde{f}_{ct})$. This term is $O_p\left(T^{-1/2}\right)$ after applying (A.28).

To analyze the remaining two terms in (A.29), we need third order derivatives of $\hat{M}_{jt}$ and $\hat{B}_t$:

$$
\begin{aligned}
&\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{M}_{jt}\\
=\ &\sum_{u=1}^{n_\beta+2n_\delta}\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}+\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{1t}+\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}\\
&+\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}-\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{2t}\\
&+(1-\hat{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t}-\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t}\\
&+(\nabla_{\theta_j}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_j}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\nabla_{\theta_j}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_j}\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}+(\nabla_{\theta_j}\hat{f}_{1t}-\nabla_{\theta_j}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_u}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\hat{f}_{1t}-\nabla_{\theta_s}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}\\
&+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Bigg\}\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\hat{\theta}_u(\delta_2)\nabla_{\delta_{2h}}\hat{\theta}(\delta_2)\\
&+\sum_{u=1}^{n_\beta+2n_\delta}\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{1t}+\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{2t}\\
&+(1-\hat{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t}+(\nabla_{\theta_j}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_j}\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta_j}\hat{f}_{1t}-\nabla_{\theta_j}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\hat{f}_{1t}-\nabla_{\theta_s}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_u}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_j}\hat{\xi}_{t|t-1}+(\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\hat{\xi}_{t|t-1}\\
&+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}\Bigg\}[\nabla_{\delta_{2k}}\nabla_{\delta_{2h}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\hat{\theta}_u(\delta_2)+\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\theta}_u(\delta_2)]\\
&+\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}+\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{1t}+(1-\hat{\xi}_{t|t-1})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}\\
&-\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta_s}\hat{f}_{2t}+(\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\nabla_{\theta_j}\hat{f}_{1t}-\nabla_{\theta_j}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\hat{f}_{1t}-\nabla_{\theta_s}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Bigg\}\nabla_{\delta_{2h}}\hat{\theta}(\delta_2)\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}_s(\delta_2),\\
&+\Bigg\{\hat{\xi}_{t|t-1}\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{1t}+(1-\xi_{t|t-1})\nabla_{\theta_j}\nabla_{\theta'}\hat{f}_{2t}+(\nabla_{\theta_j}\hat{f}_{1t}-\nabla_{\theta_j}\hat{f}_{2t})\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
&+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_j}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_j}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Bigg\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\theta}(\delta_2),
\end{aligned}
$$

and

$$
\begin{aligned}
&\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{B}_t \\
=\; &\sum_{s=1}^{n_\beta+2n_\delta}\sum_{u=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{1t}+\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}+\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t} \\
&+\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{2t}-\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}+(1-\hat{\xi}_{t|t-1})\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t} \\
&-\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t}+(\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Bigg\}\nabla_{\delta_{2l}}\hat{\theta}_u(\delta_2)\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2h}}\hat{\theta}(\delta_2) \\
&+\sum_{s=1}^{n_\beta+2n_\delta}\sum_{u=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{1t}+\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{2t}+(1-\hat{\xi}_{t|t-1})\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t} \\
&+(\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}\Bigg\}\times \\
&\Big[\nabla_{\delta_{2h}}\nabla_{\delta_{2l}}\hat{\theta}_u(\delta_2)\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)+\nabla_{\delta_{2l}}\hat{\theta}_u(\delta_2)\nabla_{\delta_{2h}}\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\Big] \\
&+\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{1t}+\hat{\xi}_{t|t-1}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}+(1-\hat{\xi}_{t|t-1})\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}-\nabla_{\theta'}\hat{\xi}_{t|t-1}\nabla_{\theta_s}\hat{f}_{2t} \\
&+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Bigg\}\nabla_{\delta_{2h}}\hat{\theta}(\delta_2)\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}_s(\delta_2) \\
&+\Big[\hat{\xi}_{t|t-1}\nabla_{\theta'}\hat{f}_{1t}+(1-\hat{\xi}_{t|t-1})\nabla_{\theta'}\hat{f}_{2t}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta'}\hat{\xi}_{t|t-1}\Big]\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\theta}(\delta_2).
\end{aligned}
$$

Consider the 1st term in (A.29). In the expression of $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{M}_{jt}$, only the last two lines involve third order derivatives of $\hat{\theta}(\delta_2)$. These derivatives are multiplied by (after division by $\widetilde{f}_t$): $T^{-1}\sum_{t=1}^{T}\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\widetilde{f}_{at}/\widetilde{f}_t$ and $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)\nabla_{\theta_{j_1}}\widetilde{\xi}_{t|t-1}$, where $a=1,2$. They are $O_p(T^{-1/2})$ by Lemma A.2. The remaining components of $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{M}_{jt}$ lead to: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\widetilde{f}_{at}/\widetilde{f}_t)$ for $a=1,2$ and $k\le 4$ and $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_m}}\widetilde{\xi}_{t|t-1})$ for $a=1,2$ and $k+m\le 4$. They are all $o_p(1)$ by Lemma A.2. Therefore the contribution of $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{M}_{jt}$ to (A.29) is $o_p(1)$. Finally, we turn to the 6th term in (A.29). In the expression for $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{B}_t$, only the final line involves third order derivatives of $\hat{\theta}(\delta_2)$. It can be analyzed in the same way as the second term in (A.16); see Step 2 of the proof there. The remaining components, multiplied by $\hat{M}_{jt}$, lead to: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\nabla_{\theta_{j_2}}\widetilde{\xi}_{t|t-1})$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{j_2}}\widetilde{\xi}_{t|t-1})$, $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\nabla_{\theta_{i_2}}\nabla_{\theta_{i_3}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{f}_{bt}/\widetilde{f}_t)$ and $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{i_2}}\widetilde{f}_{bt}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}\widetilde{\xi}_{t|t-1})$ for $a=1,2$ and $b=1,2$. They are all $O_p(1)$ by Lemma A.2. This implies the desired result. ■

**Proof of Lemma 3.** The first order derivative with respect to the $j$-th element of $\delta_2$ satisfies

$$\begin{aligned}\mathcal{L}_j^{(1)}(p,q,\delta_2) &= \sum_{t=1}^{T}\frac{1}{\hat{B}_t}\left(\nabla_{\theta'}\hat{f}_{1t}\hat{\xi}_{t|t-1}+\nabla_{\theta'}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta'}\hat{\xi}_{t|t-1}\right)\nabla_{\delta_{2j}}\hat{\theta}(\delta_2)\\ &= \sum_{s=1}^{n_\beta+n_\delta}\left\{\sum_{t=1}^{T}\frac{1}{\hat{B}_t}\left(\nabla_{\theta_s}\hat{f}_{1t}\hat{\xi}_{t|t-1}+\nabla_{\theta_s}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\theta_s}\hat{\xi}_{t|t-1}\right)\right\}\nabla_{\delta_{2j}}\hat{\theta}_s(\delta_2)\\ &\quad+\sum_{t=1}^{T}\frac{1}{\hat{B}_t}\left(\nabla_{\delta_{2j}}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1}\right),\end{aligned}$$

where the second equality follows from the definition of $\nabla_{\delta_{2j}}\hat{\theta}(\delta_2)$; see (A.19). The term inside the curly brackets equals zero because of the first order conditions determining $\hat{\beta}(\delta_2)$ and $\hat{\delta}_1(\delta_2)$. Therefore, we can write

$$\mathcal{L}_j^{(1)}(p,q,\delta_2)=\sum_{t=1}^{T}\frac{\hat{L}_{jt}}{\hat{B}_t},$$

where $\hat{B}_t$ is defined in (A.15) and $\hat{L}_{jt}=\nabla_{\delta_{2j}}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1}$. The following results hold at $\delta_2=\widetilde{\delta}$: $\widetilde{\xi}_{t|t-1}=\xi_*,\hat{\delta}_1(\widetilde{\delta})=\widetilde{\delta}$ and $\hat{\beta}(\widetilde{\delta})=\widetilde{\beta}$. Consequently, $\mathcal{L}_j^{(1)}(p,q,\widetilde{\delta})=(1-\xi_*)\sum_{t=1}^{T}(\nabla_{\delta_{2j}}\widetilde{f}_{2t}/\widetilde{f}_t)=0$, where the last equality follows because $\widetilde{\delta}$ is the MLE of the null likelihood. This proves the first result in the lemma.

Now consider the second result. Because $\widetilde{\xi}_{t|t-1}=\xi_*$, the following identity holds at $\delta_2=\widetilde{\delta}$:

$$\hat{L}_{jt}=[(1-\xi_*)/\xi_*]\hat{M}_{(n_\beta+j)t}. \tag{A.30}$$

This relationship generalizes an analogous result in Cho and White (2007, p. 1683-1684, c.f. the relationship between $h_t(\theta_2)$ and $k_t(\theta_2)$) to Markov switching models. It allows us to relate $\mathcal{L}_{jk}^{(2)}(p,q,\delta_2)$ to $\mathcal{M}_{(n_\beta+j)k}^{(2)}(p,q,\delta_2)$ when analyzing the former's properties. Specifically, we differentiate $\mathcal{L}_j^{(1)}(p,q,\delta_2)$ with respect to the $k$-th element of $\delta_2$:

$$\mathcal{L}_{jk}^{(2)}(p,q,\delta_2)=\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{L}_{jt}}{\hat{B}_t}-\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t^2}\hat{L}_{jt},$$

where

$$\begin{aligned}\nabla_{\delta_{2k}}\hat{L}_{jt} &= \{\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})-\nabla_{\delta_{2j}}\hat{f}_{2t}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\ &\quad+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{\xi}_{t|t-1}\}\nabla_{\delta_{2k}}\hat{\theta}(\delta_2).\end{aligned} \tag{A.31}$$

Because $\mathcal{M}_{(n_\beta+j)k}^{(2)}(p,q,\delta_2)=0$, we have

$$\begin{aligned}T^{-1/2}\mathcal{L}_{jk}^{(2)}(p,q,\delta_2) &= T^{-1/2}\mathcal{L}_{jk}^{(2)}(p,q,\delta_2)-T^{1/2}\left(\frac{1-\xi_*}{\xi_*}\right)\mathcal{M}_{(n_\beta+j)k}^{(2)}(p,q,\delta_2)\\ &= T^{-1/2}\sum_{t=1}^{T}\left\{\frac{\nabla_{\delta_{2k}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\ &\quad-T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t^2}\left\{\hat{L}_{jt}-\left(\frac{1-\xi_*}{\xi_*}\right)\hat{M}_{(n_\beta+j)t}\right\},\end{aligned} \tag{A.32}$$

where second summation on the right hand side equals 0 at $\delta_2 = \widetilde{\delta}$ because of (A.30). Now consider the two terms in the first summation separately. At $\delta_2 = \widetilde{\delta}$, $T^{-1/2}\sum_{t=1}^{T} \nabla_{\delta_{2k}} \hat{L}_{jt}/\hat{B}_t$ equals

$$T^{-1/2}\sum_{t=1}^{T} \frac{1}{\widetilde{f}_t}\left\{(1-\xi_*)\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\widetilde{f}_{1t} + \frac{1}{\xi_*}\nabla_{\delta_{1j}}\widetilde{f}_{1t}\nabla_{\delta_{1k}}\widetilde{\xi}_{t|t-1} + \frac{1}{\xi_*}\nabla_{\delta_{1k}}\widetilde{f}_{1t}\nabla_{\delta_{2j}}\widetilde{\xi}_{t|t-1}\right\} + O_p(T^{-1/2}).$$

Meanwhile, at $\delta_2 = \widetilde{\delta}$, $T^{-1/2}\sum_{t=1}^{T} \nabla_{\delta_{2k}} \hat{M}_{(n_\beta+j)t}/\hat{B}_t$ equals

$$-T^{-1/2}\sum_{t=1}^{T} \frac{1}{\widetilde{f}_t}\left\{(1-\xi_*)\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\widetilde{f}_{1t} + \frac{1}{\xi_*}\nabla_{\delta_{1j}}\widetilde{f}_{1t}\nabla_{\delta_{1k}}\widetilde{\xi}_{t|t-1} + \frac{1}{\xi_*}\nabla_{\delta_{1k}}\widetilde{f}_{1t}\nabla_{\delta_{2j}}\widetilde{\xi}_{t|t-1}\right\} + O_p(T^{-1/2}).$$

The result follows by combining the above two displays.

Consider the third order derivatives. Using (A.32), we have

$$\begin{aligned}
& T^{-3/4}\mathcal{L}^{(3)}_{jkl}(p,q,\delta_2) - T^{1/4}\left(\frac{1-\xi_*}{\xi_*}\right)\mathcal{M}^{(3)}_{(n_\beta+j)kl}(p,q,\delta_2) \qquad \text{(A.33)}\\
= \; & T^{-3/4}\sum_{t=1}^{T}\left\{\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{L}_{jt}}{\hat{B}_t} - \left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-3/4}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2k}}\hat{L}_{jt}}{\hat{B}_t} - \left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-3/4}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2l}}\hat{L}_{jt}}{\hat{B}_t} - \left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2l}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-3/4}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2}\left\{\hat{L}_{jt} - \left(\frac{1-\xi_*}{\xi_*}\right)\hat{M}_{(n_\beta+j)t}\right\}\\
& +2T^{-3/4}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^3}\left\{\hat{L}_{jt} - \left(\frac{1-\xi_*}{\xi_*}\right)\hat{M}_{(n_\beta+j)t}\right\},
\end{aligned}$$

where the last two summations equal 0 because of (A.30) and

$$\begin{aligned}
& \nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{L}_{jt}\\
= \; & \sum_{s=1}^{n_\beta+2n_\delta}\Big\{\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1}) - \nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{f}_{2t}\nabla_{\theta'}\hat{\xi}_{t|t-1} - \nabla_{\delta_{2j}}\nabla_{\theta'}\hat{f}_{2t}\nabla_{\theta_s}\hat{\xi}_{t|t-1}\\
& -\nabla_{\delta_{2j}}\hat{f}_{2t}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1} + (\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1} + (\nabla_{\theta_s}\hat{f}_{1t} - \nabla_{\theta_s}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{\xi}_{t|t-1}\\
& +(\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{\xi}_{t|t-1} + (\hat{f}_{1t} - \hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Big\}\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\hat{\theta}(\delta_2)\\
& +\Big\{\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1}) - \nabla_{\delta_{2j}}\hat{f}_{2t}\nabla_{\theta'}\hat{\xi}_{t|t-1} + (\nabla_{\theta'}\hat{f}_{1t} - \nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1}\\
& +(\hat{f}_{1t} - \hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Big\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}(\delta_2).
\end{aligned}$$

The first summation in (A.33) consists of the following: $T^{-3/4}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_u}}\widetilde{f}_{at}/\widetilde{f}_t)\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_v}}\widetilde{\xi}_{t|t-1}$ with $u+v\leq 3$. They are $O_p(T^{-1/4})$ by the first result in Lemma A.2. Combining this result with Lemma 2, it follows that this summation is $O_p(T^{-1/4})$. The remaining two summations in (A.33) have the same structure. They are both $O_p(T^{-1/4})$ after applying (A.28).

Consider the fourth order derivatives. Applying (A.33) and omitting terms that are zero implied by (A.30), we have

$$\begin{aligned}
& T^{-1}\mathcal{L}^{(4)}_{jklm}(p,q,\delta_2)-\left(\frac{1-\xi_*}{\xi_*}\right)\mathcal{M}^{(4)}_{(n_\beta+j)klm}(p,q,\delta_2) \qquad\qquad \text{(A.34)}\\
= \; & T^{-1}\sum_{t=1}^{T}\left\{\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2k}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2m}}\nabla_{\delta_{2k}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2m}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& +2T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2k}}\hat{L}_{jt}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\hat{M}_{(n_\beta+j)t}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2}\right\}\\
& -T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2m}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2m}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2l}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2l}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& -T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\}\\
& +2T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2l}}\hat{L}_{jt}\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t^2}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2m}}\hat{B}_t\nabla_{\delta_{2l}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t^2}\right\}\\
& +2T^{-3/4}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\hat{B}_t\nabla_{\delta_{2l}}\hat{B}_t}{\hat{B}_t^2}\left\{\frac{\nabla_{\delta_{2m}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2m}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\},
\end{aligned}$$

where

$$
\begin{aligned}
&\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{L}_{jt} \\
= \quad & \sum_{u=1}^{n_\beta+2n_\delta}\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})+\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&-\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}\nabla_{\theta_u}\hat{\xi}_{t|t-1}-\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{f}_{2t}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&-\nabla_{\delta_{2j}}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t}\nabla_{\theta_s}\hat{\xi}_{t|t-1}-\nabla_{\delta_{2j}}\nabla_{\theta_u}\hat{f}_{2t}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&-\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{f}_{2t}\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}-\nabla_{\delta_{2j}}\hat{f}_{2t}\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_u}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\hat{f}_{1t}-\nabla_{\theta_s}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_u}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta_u}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Bigg\}\times \\
&\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\hat{\theta}_u(\delta_2)\nabla_{\delta_{2m}}\hat{\theta}(\delta_2) \\
&+\sum_{u=1}^{n_\beta+2n_\delta}\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})-\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{f}_{2t}\nabla_{\theta_u}\hat{\xi}_{t|t-1}-\nabla_{\delta_{2j}}\nabla_{\theta_u}\hat{f}_{2t}\nabla_{\theta_s}\hat{\xi}_{t|t-1} \\
&-\nabla_{\delta_{2j}}\hat{f}_{2t}\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\hat{f}_{1t}-\nabla_{\theta_s}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_u}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta_u}\hat{f}_{1t}-\nabla_{\theta_u}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta_u}\hat{\xi}_{t|t-1}\Bigg\}\times \\
&\Big[\nabla_{\delta_{2m}}\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\hat{\theta}_u(\delta_2)+\nabla_{\delta_{2k}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\theta}_u(\delta_2)\Big] \\
&+\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})-\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{f}_{2t}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&-\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{f}_{2t}\nabla_{\theta_s}\hat{\xi}_{t|t-1}-\nabla_{\delta_{2j}}\hat{f}_{2t}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta_s}\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\hat{f}_{1t}-\nabla_{\theta_s}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta'}\hat{\xi}_{t|t-1} \\
&+(\nabla_{\theta'}\hat{f}_{1t}-\nabla_{\theta'}\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\nabla_{\theta'}\hat{\xi}_{t|t-1}\Bigg\}\times \\
&\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}_s(\delta_2)\nabla_{\delta_{2m}}\hat{\theta}(\delta_2) \\
&+\sum_{s=1}^{n_\beta+2n_\delta}\Bigg\{\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{f}_{2t}(1-\hat{\xi}_{t|t-1})-\nabla_{\delta_{2j}}\hat{f}_{2t}\nabla_{\theta_s}\hat{\xi}_{t|t-1}+(\nabla_{\theta_s}\hat{f}_{1t}-\nabla_{\theta_s}\hat{f}_{2t})\nabla_{\delta_{2j}}\hat{\xi}_{t|t-1} \\
&+(\hat{f}_{1t}-\hat{f}_{2t})\nabla_{\delta_{2j}}\nabla_{\theta_s}\hat{\xi}_{t|t-1}\Bigg\}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\theta}_s(\delta_2).
\end{aligned}
$$

We consider the terms in (A.34) separately. The first summation involves the following quantities: $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\widetilde{f}_{at}/\widetilde{f}_t)$ for $k=2,3,4$ and $T^{-1}\sum_{t=1}^{T}(\nabla_{\theta_{i_1}}...\nabla_{\theta_{i_k}}\widetilde{f}_{at}/\widetilde{f}_t)(\nabla_{\theta_{j_1}}...\nabla_{\theta_{j_m}}\widetilde{\xi}_{t|t-1})$

for $2 \leq k+m \leq 4$. They are all $o_p(1)$. Consequently the first summation is also $o_p(1)$. The 2nd, 4th, 5th, 8th, 9th and 10th terms involve first order derivatives of $\hat{B}_t$, and are $o_p(1)$ because of the relationship (A.28). The remaining three terms have the same structure. It suffices to analyze the first of them:

$$-T^{-1}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t}\left\{\frac{\nabla_{\delta_{2k}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2k}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right\} \tag{A.35}$$

Further, for $\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{B}_t$, it suffices to consider $((1-\xi_*)/\xi_*)\nabla_{\delta_{1l}}\nabla_{\delta_{1m}}\widetilde{f}_{1t}+(1/\xi_*^2)\nabla_{\delta_{1l}}\widetilde{f}_{1t}\nabla_{\delta_{1m}}\widetilde{\xi}_{t|t-1}+(1/\xi_*^2)\nabla_{\delta_{1l}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1m}}\widetilde{f}_{1t}+\sum_{s=1}^{n_\beta+n_\delta}(\xi_*\nabla_{\theta_s}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_s}\widetilde{f}_{2t})\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\theta}_s(\widetilde{\delta})$. For $\nabla_{\delta_{2k}}\hat{L}_{jt}$, it suffices to consider $(1-\xi_*)\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\widetilde{f}_{1t}+(1/\xi_*)\nabla_{\delta_{1j}}\widetilde{f}_{1t}\nabla_{\delta_{1k}}\widetilde{\xi}_{t|t-1}+(1/\xi_*)\nabla_{\delta_{1j}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1k}}\widetilde{f}_{1t}$. For $\nabla_{\delta_{2k}}\hat{M}_{(n_\beta+j)t}$, it suffices to consider $-(1-\xi_*)\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\widetilde{f}_{1t}-(1/\xi_*)\nabla_{\delta_{1j}}\widetilde{f}_{1t}\nabla_{\delta_{1k}}\widetilde{\xi}_{t|t-1}-(1/\xi_*)\nabla_{\delta_{1j}}\widetilde{\xi}_{t|t-1}\nabla_{\delta_{1k}}\widetilde{f}_{1t}$. Combining the above three formulas, we have, at $\widetilde{\delta}$ (A.35) equals

$$\begin{aligned}
&-T^{-1}\sum_{t=1}^{T}\left[\widetilde{U}_{lm,t}+\sum_{s=1}^{n_\beta+n_\delta}\frac{\xi_*\nabla_{\theta_s}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_s}\widetilde{f}_{2t}}{\widetilde{f}_t}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\theta}_s(\widetilde{\delta})\right]\widetilde{U}_{jk,t}+o_p(1)\\
=\;&-T^{-1}\sum_{t=1}^{T}\widetilde{U}_{lm,t}\widetilde{U}_{jk,t}-T^{-1}\sum_{t=1}^{T}\left\{\widetilde{U}_{jk,t}\frac{\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}}{\widetilde{f}_t}\right\}\widetilde{I}^{-1}\widetilde{D}_{lm}+o_p(1)\\
=\;&-\left\{\widetilde{V}_{jklm}-\widetilde{D}_{jk}'\widetilde{I}^{-1}\widetilde{D}_{lm}\right\}+o_p(1),
\end{aligned}$$

where the first equality uses Lemma 2.2 and the second applies (20). Consequently,

$$\begin{aligned}
&T^{-1}\mathcal{L}_{jklm}^{(4)}(p,q,\widetilde{\delta})-T^{-1}\left(\frac{1-\xi_*}{\xi_*}\right)\mathcal{M}_{(n_\beta+j)klm}^{(4)}(p,q,\widetilde{\delta})\\
=\;&-\left\{\widetilde{V}_{jklm}-\widetilde{D}_{jk}'\widetilde{I}^{-1}\widetilde{D}_{lm}+\widetilde{V}_{jmkl}-\widetilde{D}_{jm}'\widetilde{I}^{-1}\widetilde{D}_{kl}+\widetilde{V}_{kmjl}-\widetilde{D}_{km}'\widetilde{I}^{-1}\widetilde{D}_{jl}\right\}+o_p(1).
\end{aligned}$$

This proves the final result of the lemma. ■

The next lemma will be used in the proof of Proposition 1 for establishing the stochastic equicontinuity. We use "*" to signify that the quantity is evaluated at the true parameter value.

**Lemma A.3** *Let Assumptions 1-5 and the null hypothesis hold. Let* $z_t(\rho)=T^{-1/2}\sum_{s=1}^{t-1}\rho^{t-s}\varepsilon_{js}\varepsilon_{it}$, *where* $\varepsilon_{it}=\nabla_{\delta_{1i}}f_{1t}^*/f_t^*$ *and* $\varepsilon_{js}=\nabla_{\delta_{1j}}f_{1s}^*/f_s^*$. *Then, for any* $\rho$, $\rho_1$ *and* $\rho_2$ *satisfying* $\epsilon-1\leq\rho_1\leq\rho\leq\rho_2\leq 1-\epsilon$, *we have*

$$E\left(\left|\sum_{t=1}^{T}[z_t(\rho)-z_t(\rho_1)]\right|^2\left|\sum_{t=1}^{T}[z_t(\rho_2)-z_t(\rho)]\right|^2\right)\leq C(\rho-\rho_1)^2, \tag{A.36}$$

*where* $C$ *is a finite constant that depends only on* $0<\epsilon<1/2$ *and the moments of* $\varepsilon_{it}$ *and* $\varepsilon_{js}$ *up to the fourth order.*

**Proof.** Let $c_{t-s}(\rho)=T^{-1/2}\rho^{t-s}$, $c_{t-s}(\rho_1,\rho)=c_{t-s}(\rho)-c_{t-s}(\rho_1)$ and $z_t(s,r)=z_t(r)-z_t(s)$. We first show that the left hand side of (A.36) is bounded from above by

$$C\left(\sum_{t=1}^{T}\sum_{s=1}^{t-1}c_{t-s}(\rho_1,\rho)^2\right)\left(\sum_{t=1}^{T}\sum_{h=1}^{t-1}c_{t-h}(\rho,\rho_2)^2\right). \tag{A.37}$$

Because $\varepsilon_{js}$ and $\varepsilon_{it}$ are martingale differences, the left hand side equals

$$\begin{aligned}
& E\sum_{t=1}^{T} z_t(\rho_1,\rho)^2 z_t(\rho,\rho_2)^2 + E\sum_{t=1}^{T}\sum_{k=1,k\neq t}^{T} z_t(\rho_1,\rho)^2 z_k(\rho,\rho_2)^2 \\
& +2E\sum_{t=1}^{T}\sum_{l=1,l\neq t}^{T} z_t(\rho_1,\rho) z_l(\rho_1,\rho) z_t(\rho,\rho_2) z_l(\rho,\rho_2) \\
= \;& \text{(T.1)+(T.2)+(T.3)}.
\end{aligned}$$

We analyze the three terms separately:

$$\begin{aligned}
\text{(T.2)} \;=\;& E\sum_{t=1}^{T}\varepsilon_{it}^2\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)\varepsilon_{js}\right)^2 \sum_{k=1}^{t-1}\varepsilon_{ik}^2\left(\sum_{h=1}^{k-1} c_{k-h}(\rho,\rho_2)\varepsilon_{jh}\right)^2 \\
& +E\sum_{k=1}^{T}\varepsilon_{ik}^2\left(\sum_{h=1}^{k-1} c_{k-h}(\rho,\rho_2)\varepsilon_{jh}\right)^2 \sum_{t=1}^{k-1}\varepsilon_{it}^2\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)\varepsilon_{js}\right)^2 .
\end{aligned}$$

Due to symmetry, it suffices to consider the first term on the right hand side, which equals

$$\begin{aligned}
& C_1\sum_{t=1}^{T}\sum_{k=1}^{t-1} E\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2\varepsilon_{js}^2\right)\varepsilon_{ik}^2\left(\sum_{h=1}^{k-1} c_{k-h}(\rho,\rho_2)^2\varepsilon_{jh}^2\right) && \text{(I)} \\
& +C_1\sum_{t=1}^{T}\sum_{k=1}^{t-1}\sum_{s=1}^{k-1}\sum_{h=1,h\neq s}^{k-1} c_{t-s}(\rho_1,\rho)c_{t-h}(\rho_1,\rho)c_{k-s}(\rho,\rho_2)c_{k-h}(\rho,\rho_2) && \text{(II)}
\end{aligned}$$

for some $0 < C_1 < \infty$, where $C_1$ depends only on $E\varepsilon_{it}^2$ and $E\varepsilon_{jt}^2$. (Below, the finite constants $C_s$ $(s = 2, 3, 4, 5)$ also depend only on the moments of $\varepsilon_{jt}$ and $\varepsilon_{it}$, up to the fourth order). Term (I) is further bounded by

$$C_2\sum_{t=1}^{T}\sum_{k<t}\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2\right)\left(\sum_{h=1}^{k-1} c_{k-h}(\rho,\rho_2)^2\right) \le C_2\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2\right)\sum_{t=1}^{T}\left(\sum_{h=1}^{t-1} c_{t-h}(\rho,\rho_2)^2\right). \tag{A.38}$$

Applying the Cauchy-Schwarz inequality to the elements of (II), we have $\sum_{s=1}^{k-1}|c_{t-s}(\rho_1,\rho)c_{k-s}(\rho,\rho_2)| \le (\sum_{s=1}^{k-1} c_{t-s}(\rho_1,\rho)^2)^{1/2}(\sum_{s=1}^{k-1} c_{k-s}(\rho,\rho_2)^2)^{1/2}$ and $\sum_{h=1}^{k-1}|c_{t-h}(\rho_1,\rho)c_{k-h}(\rho,\rho_2)| \le (\sum_{h=1}^{k-1} c_{t-h}(\rho_1,\rho)^2)^{1/2}$ $(\sum_{h=1}^{k-1} c_{k-h}(\rho,\rho_2)^2)^{1/2}$. Combining these two inequalities:

$$\begin{aligned}
|\text{(II)}| \;\le\;& C_1\sum_{t=1}^{T}\sum_{k<t}\left(\sum_{s=1}^{k-1} c_{t-s}(\rho_1,\rho)^2\right)\left(\sum_{h=1}^{k-1} c_{k-h}(\rho,\rho_2)^2\right) \qquad \text{(A.39)} \\
\le\;& C_1\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2\right)\sum_{t=1}^{T}\left(\sum_{h=1}^{t-1} c_{t-h}(\rho,\rho_2)^2\right),
\end{aligned}$$

which is proportional to (A.38). Hence,

$$|\text{(I)+(II)}| \le C_3\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2\right)\sum_{t=1}^{T}\left(\sum_{h=1}^{t-1} c_{t-h}(\rho,\rho_2)^2\right) . \tag{A.40}$$

Apply the Cauchy-Schwarz inequality to (T.3):

$$
\begin{aligned}
|(\text{T.3})| &\le E\left(\sum_{t=1}^{T} z_t(\rho_1,\rho)^2\right)\left(\sum_{t=1}^{T} z_t(\rho,\rho_2)^2\right) \\
&= E\left(\sum_{t=1}^{T} z_t(\rho_1,\rho)^2 \sum_{k=1,k\neq t}^{T} z_k(\rho,\rho_2)^2\right) + E\sum_{t=1}^{T} z_t(\rho_1,\rho)^2 z_t(\rho,\rho_2)^2, \qquad \text{(A.41)}
\end{aligned}
$$

where the first term is the same as (T.2) and the second term equals (T.1). Consequently, a separate analysis of (T.3) is not needed.

Finally, we turn to (T.1). It equals

$$
\begin{aligned}
& E\varepsilon_{it}^4 \sum_{t=1}^{T}\left(E\sum_{s=1}^{t-1}\sum_{k=1}^{t-1}\sum_{h=1}^{t-1}\sum_{l=1}^{t-1} c_{t-s}(\rho_1,\rho)c_{t-k}(\rho_1,\rho)c_{t-h}(\rho,\rho_2)c_{t-l}(\rho,\rho_2)\varepsilon_{js}\varepsilon_{jk}\varepsilon_{jh}\varepsilon_{jl}\right) \\
= \; & E\varepsilon_{it}^4 E\varepsilon_{jt}^4 \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2 c_{t-s}(\rho,\rho_2)^2\right) + E\varepsilon_{it}^4\left(E\varepsilon_{jt}^2\right)^2 \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\sum_{h=1,h\neq s}^{t-1} c_{t-s}(\rho_1,\rho)^2 c_{t-h}(r,r_2)^2\right) \\
& +2E\varepsilon_{it}^4\left(E\varepsilon_{jt}^2\right)^2 \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)c_{t-s}(\rho,\rho_2)\sum_{k=1,k\neq s}^{t-1} c_{t-k}(\rho_1,\rho)c_{t-k}(\rho,\rho_2)\right) \\
\le \; & C_4 \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\sum_{h=1}^{t-1} c_{t-s}(\rho_1,\rho)^2 c_{t-h}(\rho,\rho_2)^2\right) \qquad \text{(III)} \\
& +C_4 \sum_{t=1}^{T}\left(\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)c_{t-s}(\rho,\rho_2)\sum_{k=1,k\neq s}^{t-1} c_{t-k}(\rho_1,\rho)c_{t-k}(\rho,\rho_2)\right). \qquad \text{(IV)}
\end{aligned}
$$

As in (A.39), we have $|(\text{IV})| \le C_4 \sum_{t=1}^{T}\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2 \sum_{k=1}^{t-1} c_{t-k}(\rho,\rho_2)^2$. Hence,

$$
|(\text{III})+(\text{IV})| \le C_5\left(\sum_{t=1}^{T}\sum_{s=1}^{t-1} c_{t-s}(\rho_1,\rho)^2\right)\left(\sum_{t=1}^{T}\sum_{h=1}^{t-1} c_{t-h}(\rho,\rho_2)^2\right). \qquad \text{(A.42)}
$$

Combining (A.40), (A.41), and (A.42) leads to (A.37).

By the mean value theorem: $c_{t-s}(\rho_1,\rho) = T^{-1/2}(\rho^{t-s}-\rho_1^{t-s}) \le T^{-1/2}(t-s)(1-2\epsilon)^{t-s-1}(\rho-\rho_1)$. The right hand side of (A.37) is therefore bounded by $C\{T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{t-1}(t-s)^2(1-2\epsilon)^{2(t-s-1)}\}^2(\rho_2-\rho_1)^2$. The term in the curly brackets is finite; the result follows after redefining the constant $C$. ■

**Proof of Proposition 1.** Apply the mean value theorem:

$$
T^{-1/2}\sum_{t=1}^{T}\widetilde{U}_{jk,t} = T^{-1/2}\sum_{t=1}^{T} U_{jk,t} + \left\{T^{-1}\sum_{t=1}^{T}\nabla_{\theta'}\bar{U}_{jk,t}\right\} T^{1/2}(\widetilde{\theta}-\theta_*), \qquad \text{(A.43)}
$$

where $U_{jk,t}$ and $\bar{U}_{jk,t}$ have the same definition as $\widetilde{U}_{jk,t}$ but evaluated at the true value $\theta_*$ and some value $\bar{\theta}$ that lies between $\widetilde{\theta}$ and $\theta_*$, respectively.

We establish the weak convergence of the first term of (A.43) in two steps. First, for any $\epsilon \leq p$, $q \leq 1-\epsilon$, $T^{-1/2}\sum_{t=1}^{T} U_{jk,t}$ satisfies the central limit theorem. Second, to verify its stochastic equicontinuity, it suffices to consider the second component in its definition (20). This term equals

$$T^{-1/2}\frac{1}{\xi_*^2}\sum_{t=1}^{T}\nabla_{\delta_{1j}}\xi_{t|t-1}\frac{\nabla_{\delta_{1k}}f_{1t}}{f_t} = \left(\frac{1-p}{1-q}\right)\left\{T^{-1/2}\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\rho^s\frac{\nabla_{\delta_{1j}}f_{1(t-s)}}{f_{t-s}}\right)\frac{\nabla_{\delta_{1k}}f_{1t}}{f_t}\right\}.$$

where the quantities are all evaluated at the true value $\theta_*$, and the equality follows from (A.6) and (3). Denote the quantity inside the curly brackets as $W(\rho)$. Note that we have $|\rho| \leq 1-2\epsilon$. Then, Lemma A.3 implies, for any $\rho_1 \leq \rho \leq \rho_2$, we have $E[|W(\rho_1)-W(\rho)|^2\,|W(\rho)-W(\rho_2)|^2] \leq C(\rho_1-\rho_2)^2$, where $C$ is a finite constant. This fulfills the condition required in Theorem 13.5 in Billingsley (1999; c.f. the Display (13.14) in p. 143). This shows that $W(\rho)$ is stochastic equicontinuous.

The second term in (A.43) equals, by the mean value theorem,

$$\begin{aligned}
&-\left\{T^{-1}\sum_{t=1}^{T}\left(\frac{\nabla_{(\beta',\delta')}f_{1t}}{f_t}\right)U_{jk,t}\right\}I^{-1}\left\{T^{-1/2}\sum_{t=1}^{T}\left(\frac{\nabla_{(\beta',\delta')'}f_{1t}}{f_t}\right)\right\}+o_p(1)\\
=\;&-D_{jk}I^{-1}\left\{T^{-1/2}\sum_{t=1}^{T}\left(\frac{\nabla_{(\beta',\delta')'}f_{1t}}{f_t}\right)\right\}+o_p(1),
\end{aligned}$$

where the quantities are all evaluated at the true value $\theta_*$ and the second equality holds because of the uniform law of large numbers. The term inside the last curly brackets is independent of $p$ and $q$ and satisfies the central limit theorem. Combining the above results for the two terms in (A.43), it follows that $T^{-1/2}\sum_{t=1}^{T}\widetilde{U}_{jk,t}$ converges weakly over $\epsilon \leq p$, $q \leq 1-\epsilon$. The covariance function follows immediately; we omit the details.■

**Proof of Proposition 2.** Let $\eta = T^{-1/4}(\delta_2 - \widetilde{\delta})$. The expansion (19) can be equivalently represented in matrix notation as

$$\begin{aligned}
&\mathcal{L}(p,q,\delta_2)-\mathcal{L}(p,q,\widetilde{\delta})\\
=\;&\frac{1}{2!}\left(\eta^{\otimes 2}\right)'\left[T^{-1/2}\operatorname{vec}\mathcal{L}^{(2)}(p,q,\widetilde{\delta})\right]+\frac{1}{3!}\left(\eta^{\otimes 3}\right)' * O_p\left(T^{-1/4}\right)-\frac{1}{8}\left(\eta^{\otimes 2}\right)'\left[\Omega(p,q)+o_p(1)\right]\left(\eta^{\otimes 2}\right).
\end{aligned}$$

Because $\Omega(p,q)$ is positive definite, the right hand side will be negative with probability approaching 1 unless $\eta = O_p(1)$. Thus, for any $\varepsilon > 0$, we can choose $M < \infty$ such that $P(\|\eta\| \leq M) \geq 1-\varepsilon$ for sufficiently large $T$. Restricting to this set, we have

$$\begin{aligned}
&\sup_{(p,q)\in\Lambda_\epsilon}\sup_{\|\eta\|\leq M}\left[\mathcal{L}(p,q,\delta_2)-\mathcal{L}(p,q,\widetilde{\delta})\right]\\
=\;&\sup_{(p,q)\in\Lambda_\epsilon}\sup_{\|\eta\|\leq M}\left\{\left(\eta^{\otimes 2}\right)'\left[T^{-1/2}\operatorname{vec}\mathcal{L}^{(2)}(p,q,\widetilde{\delta})\right]-\frac{1}{4}\left(\eta^{\otimes 2}\right)'\Omega(p,q)\left(\eta^{\otimes 2}\right)\right\}+o_p(1)\\
\Longrightarrow\;&\sup_{(p,q)\in\Lambda_\epsilon}\sup_{\|\eta\|\leq M}\left\{\left(\eta^{\otimes 2}\right)'G(p,q)-\frac{1}{4}\left(\eta^{\otimes 2}\right)'\Omega(p,q)\left(\eta^{\otimes 2}\right)\right\},
\end{aligned}$$

where the convergence follows from Proposition 1 and that the supremum operator is continuous when taken over a compact set. Finally, the result follows because $\varepsilon$ can be made arbitrarily small. ■

**Lemma A.4** *Under Assumptions 1-7 and the null hypothesis, the following results hold uniformly over* $\{(p,q): \epsilon \le p, q \le 1-\epsilon, p+q=1\}$ *for any* $k,l \in \{1,...,n_\delta\}$

1. *Let* $e_k$ *be an* $n_\delta$ *dimensional unit vector whose k-th element equals 1, then*

$$\begin{bmatrix} \nabla_{\delta_{2k}}\hat{\beta}(\widetilde{\delta}) \\ \xi_*\nabla_{\delta_{2k}}\hat{\delta}_1(\widetilde{\delta}) \end{bmatrix} = (\xi_*-1)\begin{bmatrix} 0 \\ e_k \end{bmatrix}$$

2. *The second order derivatives satisfy*

$$\begin{aligned}\begin{bmatrix} \nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}) \\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta}) \end{bmatrix} &= -\left(\frac{1-\xi_*}{\xi_*}\right)\Bigg\{\alpha_{kl} - \widetilde{I}^{-1}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\widetilde{f}_t}\begin{bmatrix} \nabla_\beta\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t} \\ \left(\frac{1-\xi_*}{\xi_*}\right)\nabla_{\delta_1}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}\end{bmatrix} \\ &\quad +\widetilde{I}^{-1}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\widetilde{f}_t}\begin{bmatrix} \nabla_\beta\nabla_{\beta'}\widetilde{f}_{1t}\alpha_{kl}^{(1)} + \nabla_\beta\nabla_{\delta_1'}\widetilde{f}_{1t}\alpha_{kl}^{(2)} \\ \nabla_{\delta_1}\nabla_{\beta'}\widetilde{f}_{1t}\alpha_{kl}^{(1)}\end{bmatrix}\Bigg\} + o_p(T^{-1/2}),\end{aligned}$$

*where* $\widetilde{I}$ *is defined in (20) and* $\alpha_{kl}' = (\alpha_{kl}^{(1)\prime}, \alpha_{kl}^{(2)\prime})$.

**Proof of Lemma A.4.** When $p+q=1$, the derivatives of $\xi_{t|t-1}$ with respect to $\theta$ all equal zero when evaluated at $\delta_1=\delta_2=\delta$. This essentially reduces the problem to that of Cho and White (2007), except for the complication induced by multiple switching parameters. The first result in the lemma follows from the same argument as in Lemma 2; we omit the details. The second result is more complex; its proof is given below.

Consider (A.24). There, only the summations over the first and the third terms are nonzero by the relationship (A.28) and the first result of this lemma. Evaluate these two terms at the null estimates over $j \in \{1,...,n_\beta+n_\delta\}$, we obtain,

$$D\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_\beta\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t} \\ \left(\frac{1-\xi_*}{\xi_*}\right)^2\frac{\nabla_{\delta_1}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t}\end{bmatrix} + D\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \frac{\nabla_\beta\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}) + \frac{\nabla_\beta\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta}) \\ \frac{\nabla_{\delta_1}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}) + \frac{\nabla_{\delta_1}\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix}$$

and

$$D\widetilde{I}\begin{bmatrix} \nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}) \\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix} + D\left(\frac{1-\xi_*}{\xi_*}\right)\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_\beta\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t} \\ \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_1}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t}\end{bmatrix},$$

where $D$ has the same definition as in (A.21). Combine the preceding two displays, we obtain

$$\begin{aligned}&\begin{bmatrix} \nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}) \\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix} \\ =\;& -\left(\frac{1-\xi_*}{\xi_*}\right)\widetilde{I}^{-1}\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_\beta\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t} \\ \frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\delta_1}\widetilde{f}_{1t}}{\widetilde{f}_t}\frac{\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t}\end{bmatrix} + \widetilde{I}^{-1}\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \left(\frac{1-\xi_*}{\xi_*}\right)\frac{1}{\widetilde{f}_t}\nabla_{\theta_\beta}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t} \\ \left(\frac{1-\xi_*}{\xi_*}\right)^2\frac{1}{\widetilde{f}_t}\nabla_{\theta_{\delta_1}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}\end{bmatrix} \\ &+\widetilde{I}^{-1}\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \frac{\nabla_\beta\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}) + \frac{\nabla_\beta\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta}) \\ \frac{\nabla_{\delta_1}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}) + \frac{\nabla_{\delta_1}\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix}.\end{aligned}$$

Apply $\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\widetilde{f}_{1t}=\alpha'_{jk}\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}$ and $\sum_{t=1}^{T}(\nabla_{\delta_1}\nabla_{\delta_1'}\widetilde{f}_{1t}/\widetilde{f}_t)=0$, the preceding display equals

$$\begin{aligned}
&\left[\begin{array}{c}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{array}\right] \qquad\qquad (A.44)\\
= & -\left(\frac{1-\xi_*}{\xi_*}\right)\alpha_{kl}+\widetilde{I}^{-1}\frac{1}{T}\sum_{t=1}^{T}\left[\begin{array}{c}\left(\frac{1-\xi_*}{\xi_*}\right)\frac{1}{\widetilde{f}_t}\nabla_{\theta_\beta}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}\\ \left(\frac{1-\xi_*}{\xi_*}\right)^2\frac{1}{\widetilde{f}_t}\nabla_{\theta_{\delta_1}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}\end{array}\right]\\
&+\widetilde{I}^{-1}\frac{1}{T}\sum_{t=1}^{T}\left[\begin{array}{c}\frac{\nabla_\beta\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})+\frac{\nabla_\beta\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\\ \frac{\nabla_{\delta_1}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})\end{array}\right].
\end{aligned}$$

Here, $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})$ and $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})$ appear on both sides of the display. We address this in two steps. First, because the last two terms on the right hand side are $O_p(T^{-1/2})$, we have

$$\left[\begin{array}{c}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})\end{array}\right]=-\left(\frac{1-\xi_*}{\xi_*}\right)\left[\begin{array}{c}\alpha_{kl}^{(1)}\\ \alpha_{kl}^{(2)}\end{array}\right]+O_p(T^{-1/2}).$$

Second, apply this result, to the third term on the right hand side of (A.44). The latter equals

$$-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{1}{T}\widetilde{I}^{-1}\sum_{t=1}^{T}\left[\begin{array}{c}\frac{\nabla_\beta\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{kl}^{(1)}+\frac{\nabla_\beta\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{kl}^{(2)}\\ \frac{\nabla_{\delta_1}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{kl}^{(1)}\end{array}\right]+o_p\left(T^{-1/2}\right).$$

The result follows by applying this expression to (A.44). ■

**Proof of Lemma 4.** The key to the proof is that when $p+q=1$, the likelihood corresponds to that of a mixture model. The arguments used here rely heavily on that in Lemmas C2, 3 and 4 in Cho and White (2007). Below we outline the main steps.

Consider the first result. Among the summations on the right hand side of (A.33), only the first is nonzero. Further, when evaluated at the null estimates, $T^{-1/2}\sum_{t=1}^{T}(\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{L}_{jt}/\hat{B}_t)$ and $((1-\xi_*)/\xi_*)T^{-1/2}\sum_{t=1}^{T}(\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{M}_{(n_\beta+j)t}/\hat{B}_t)$ equal

$$\begin{aligned}
&(1-\xi_*)T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t}+(1-\xi_*)T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}),\\
&\frac{(1-\xi_*)^3}{\xi_*^2}T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t}+(1-\xi_*)T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta}).
\end{aligned}$$

Taking their difference gives

$$T^{-1/2}\mathcal{L}_{jkl}^{(3)}(p,q,\widetilde{\delta})=-\frac{(1-\xi_*)(1-2\xi_*)}{\xi_*^2}T^{-1/2}\sum_{T=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\widetilde{f}_{1t}}{\widetilde{f}_t}\Rightarrow G_{jkl}^{(3)}.$$

Now we turn to $T^{-1/2}\mathcal{L}_{jklm}^{(4)}(p,q,\widetilde{\delta})$. In (A.34), only the 1st, 3rd, 6th and 7th summation on the right hand side are nonzero. For the 1st summation, $T^{-1/2}\sum_{t=1}^{T}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{L}_{jt}/\hat{B}_t$ evaluated at

$\widetilde{\delta}$ equals

$$(1-\xi_*)T^{-1/2}\sum_{t=1}^{T}\frac{1}{\widetilde{f}_t}\Big\{\nabla_{\delta_{2j}}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\widetilde{f}_{2t}+\nabla_{\delta_{2j}}\nabla_{\delta_{2l}}\nabla_{\beta'}\widetilde{f}_{2t}\nabla_{\delta_{2m}}\nabla_{\delta_{2k}}\hat{\beta}(\widetilde{\delta})$$
$$+\nabla_{\delta_{2j}}\nabla_{\delta_{2k}}\nabla_{\beta'}\widetilde{f}_{2t}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\beta}(\widetilde{\delta})+\nabla_{\delta_{2j}}\nabla_{\delta_{2m}}\nabla_{\beta'}\widetilde{f}_{2t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})+\nabla_{\delta_{2j}}\nabla_{\beta'}\widetilde{f}_{2t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\beta}(\widetilde{\delta})\Big\}.$$

Meanwhile, $((1-\xi_*)/\xi_*)T^{-1/2}\sum_{t=1}^{T}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{M}_{(n_\beta+j)t}/\hat{B}_t$ at $\widetilde{\delta}$ equals

$$((1-\xi_*)/\xi_*)T^{-1/2}\sum_{t=1}^{T}\frac{1}{\widetilde{f}_t}\Big\{\frac{(\xi_*-1)^3}{\xi_*^2}\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\nabla_{\delta_{1m}}\widetilde{f}_{1t}$$
$$+(\xi_*-1)\nabla_{\delta_{1j}}\nabla_{\delta_{1l}}\nabla_{\beta'}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\nabla_{\delta_{2m}}\hat{\beta}(\widetilde{\delta})+(\xi_*-1)\nabla_{\delta_{1j}}\nabla_{\delta_{1l}}\nabla_{\delta_1'}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\nabla_{\delta_{2m}}\hat{\delta}_1(\widetilde{\delta})$$
$$+(\xi_*-1)\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\beta'}\widetilde{f}_{1t}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\beta}(\widetilde{\delta})+(\xi_*-1)\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_1'}\widetilde{f}_{1t}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\delta}_1(\widetilde{\delta})$$
$$+(\xi_*-1)\nabla_{\delta_{1j}}\nabla_{\delta_{1m}}\nabla_{\beta'}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\beta}(\widetilde{\delta})+(\xi_*-1)\nabla_{\delta_{1j}}\nabla_{\delta_{1m}}\nabla_{\delta_1'}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_1(\widetilde{\delta})$$
$$+\xi_*\nabla_{\delta_{1j}}\nabla_{\beta'}\widetilde{f}_{1t}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\beta}(\delta_2)\Big\}.$$

Their difference equals

$$(1-\xi_*)\left(1+\left(\frac{1-\xi_*}{\xi_*}\right)^3\right)T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\nabla_{\delta_{1m}}\widetilde{f}_{1t}}{\widetilde{f}_t}$$
$$-\left(\frac{1-\xi_*}{\xi_*}\right)^2T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1l}}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{km}^{(1)}-\left(\frac{1-\xi_*}{\xi_*}\right)^3T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1l}}\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{km}^{(2)}$$
$$-\left(\frac{1-\xi_*}{\xi_*}\right)^2T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{lm}^{(1)}-\left(\frac{1-\xi_*}{\xi_*}\right)^3T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1k}}\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{lm}^{(2)}$$
$$-\left(\frac{1-\xi_*}{\xi_*}\right)^2T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1m}}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{kl}^{(1)}-\left(\frac{1-\xi_*}{\xi_*}\right)^3T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{1j}}\nabla_{\delta_{1m}}\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{kl}^{(2)}+o_p(1).$$

The preceding display is $O_p(1)$ by Lemma 2 and Assumption 4. The 3rd, 6th and 7th summation in (A.34) share the same structure. Applying Lemma A.4.2, the 3rd term equals,

$$\left(\frac{1-\xi_*}{\xi_*}\right)^2T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\beta'}\nabla_{\delta_{1l}}\nabla_{\delta_{1m}}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{jk}^{(1)}+\left(\frac{1-\xi_*}{\xi_*}\right)^3T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_1'}\nabla_{\delta_{1l}}\nabla_{\delta_{1m}}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{jk}^{(2)}$$
$$-\left(\frac{1-\xi_*}{\xi_*}\right)^2T^{-1/2}\sum_{t=1}^{T}(\alpha_{jk}^{(1)})'\frac{\nabla_\beta\nabla_{\delta_1'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{lm}^{(2)}-\left(\frac{1-\xi_*}{\xi_*}\right)^2T^{-1/2}\sum_{t=1}^{T}(\alpha_{jk}^{(1)})'\frac{\nabla_\beta\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{lm}^{(1)}$$
$$-\left(\frac{1-\xi_*}{\xi_*}\right)^2T^{-1/2}\sum_{t=1}^{T}(\alpha_{jk}^{(2)})'\frac{\nabla_{\delta_1}\nabla_{\beta'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{lm}^{(1)}=O_p(1).$$

Now consider the fifth order derivative. The components of

$$T^{-1/2}\mathcal{L}_{jklmn}^{(5)}(p,q,\delta_2)-T^{-1/2}\left(\frac{1-\xi_*}{\xi_*}\right)\mathcal{M}_{(n_\beta+j)klmn}^{(5)}(p,q,\delta_2)\tag{A.45}$$

can be grouped into three subsets according to whether they depend on the first, second or third order derivatives of $\hat{B}_t$, c.f. (A.33). First, those depending on the first order derivatives are identically zero using the relationship (A.28). Second, apply the first result of Lemma A.4 to (A.26). We have $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{B}_t/\widetilde{f}_t$ evaluated at $\widetilde{\delta}$ equals

$$\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t}\alpha_{kl}+\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t}\begin{bmatrix}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix}.\tag{A.46}$$

Applying the second result in Lemma A.4, the term involving $[(1-\xi_*)/\xi_*]\alpha_{kl}$ gets canceled and the remainder term is of lower order. Consequently, in (A.45), the terms depending on the second order derivatives of $\hat{B}_t$ are all $O_p(1)$. Third, the terms depending on the third order derivatives of $\hat{B}_t$ are of the following form:

$$T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t}\left(\frac{\nabla_{\delta_{2n}}\hat{L}_{jt}}{\hat{B}_t}-\left(\frac{1-\xi_*}{\xi_*}\right)\frac{\nabla_{\delta_{2n}}\hat{M}_{(n_\beta+j)t}}{\hat{B}_t}\right). \tag{A.47}$$

When evaluated at $\widetilde{\delta}$, $\nabla_{\delta_{2n}}\hat{L}_{jt}/\hat{B}_t$ and $(\nabla_{\delta_{2n}}\hat{M}_{(n_\beta+j)t}/\hat{B}_t)$ are representable as linear functions of $(\hat{M}_{it}/\hat{B}_t)$ $(i=1,...,n_\beta+n_\delta)$ because of $\nabla_{\delta_{1j}}\nabla_{\delta_{1n}}\widetilde{f}_{1t}=\alpha'_{jn}\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}$, c.f. (A.31), (A.17) and (A.15). Such an insightful observation is made in Cho and White (2007). This implies that, at $\widetilde{\delta}$, the order of (A.47) is the same as that of

$$T^{-1/2}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2m}}\hat{B}_t}{\hat{B}_t}\frac{\hat{M}_{it}}{\hat{B}_t},\ i=1,...,n_\beta+n_\delta. \tag{A.48}$$

The order of (A.48) can be found by analyzing (A.29). There, the terms that depend on the 0th, 1st and 2rd order derivatives of $\hat{B}_t$ are all of order $O_p(T^{-1/2})$ after applying (A.28) and (A.46). The only term that remains is (A.48). Therefore, for (A.29) to equal zero, (A.48) must be of order $O_p(1)$ when evaluated at $\widetilde{\delta}$. This implies (A.47) is $O_p(1)$.

Now, consider the sixth order derivatives. To this end, we need to obtain expressions for $\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\beta}(\widetilde{\delta})$ and $\xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\delta}_1(\widetilde{\delta})$ by analyzing (A.29). The effects of the terms other than (A.48) are negligible. Writing out the expression for (A.48) explicitly, we obtain

$$\begin{aligned}
&T^{-1}\sum_{t=1}^{T}\Bigg\{\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t^2}\left[\frac{(\xi_*-1)^2}{\xi_*^2}(\xi_*-1)+(1-\xi_*)\right]\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\nabla_{\delta_{1h}}\widetilde{f}_{1t}\\
&+\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t^2}\sum_{u=1}^{n_\beta+2n_\delta}[(\xi_*-1)\nabla_{\delta_{1k}}\nabla_{\theta_u}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\delta_{2k}}\nabla_{\theta_u}\widetilde{f}_{2t}]\nabla_{\delta_{2h}}\nabla_{\delta_{2l}}\hat{\theta}_u(\widetilde{\delta})\\
&+\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t^2}\sum_{s=1}^{n_\beta+2n_\delta}[(\xi_*-1)\nabla_{\theta_s}\nabla_{\delta_{1l}}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_s}\nabla_{\delta_{2l}}\widetilde{f}_{2t}]\nabla_{\delta_{2h}}\nabla_{\delta_{2k}}\hat{\theta}_s(\widetilde{\delta})\\
&+\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t^2}\sum_{s=1}^{n_\beta+2n_\delta}[(\xi_*-1)\nabla_{\theta_s}\nabla_{\delta_{1h}}\widetilde{f}_{1t}+(1-\xi_*)\nabla_{\theta_s}\nabla_{\delta_{2h}}\widetilde{f}_{2t}]\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\theta}_s(\widetilde{\delta})\\
&+\widetilde{I}\begin{bmatrix}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\beta}(\widetilde{\delta})\\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\delta}_1(\widetilde{\delta})\end{bmatrix}\Bigg\}=o_p(1).
\end{aligned}$$

Equivalently,

$$\begin{bmatrix} \nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\beta}(\widetilde{\delta}) \\ \xi_*\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\nabla_{\delta_{2h}}\hat{\delta}_1(\widetilde{\delta}) \end{bmatrix} \tag{A.49}$$

$$\begin{aligned} =&\ -\widetilde{I}^{-1}T^{-1}\sum_{t=1}^{T}\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t^2}\left[\frac{(\xi_*-1)^2}{\xi_*^2}(\xi_*-1)+(1-\xi_*)\right]\nabla_{\delta_{1k}}\nabla_{\delta_{1l}}\nabla_{\delta_{1h}}\widetilde{f}_{1t} \\ &-(\xi_*-1)\sum_{u=1}^{n_\delta}[\alpha_{ku}\nabla_{\delta_{2h}}\nabla_{\delta_{2l}}\hat{\delta}_{1u}(\widetilde{\delta})+\alpha_{lu}\nabla_{\delta_{2h}}\nabla_{\delta_{2k}}\hat{\delta}_{1u}(\widetilde{\delta})+\alpha_{hu}\nabla_{\delta_{2k}}\nabla_{\delta_{2l}}\hat{\delta}_{1u}(\widetilde{\delta})]+o_p(1)\,. \end{aligned}$$

Now apply the above expression to analyze $T^{-1}\mathcal{L}^{(6)}_{jklmnr}(p,q,\widetilde{\delta})$. The latter equals, by using the same argument as in Cho and White (2007, l.13-24 in p. 1713),

$$T^{-1}\sum_{(i_1,i_2,\ldots,i_6)\in IND}\sum_{t=1}^{T}\frac{\nabla_{\delta_{2i_4}}\nabla_{\delta_{2i_5}}\nabla_{\delta_{2i_6}}\hat{B}_t}{\hat{B}_t^2}\left(\nabla_{\delta_{2i_2}}\nabla_{\delta_{2i_3}}\hat{L}_{i_1t}-\frac{1-\xi_*}{\xi_*}\nabla_{\delta_{2i_2}}\nabla_{\delta_{2i_3}}\hat{M}_{(n_\beta+i_1)t}\right)+o_p(1)\,,$$

where all the quantities are evaluated at $\delta_2=\widetilde{\delta}$. Further, at $\delta_2=\widetilde{\delta}$, $\nabla_{\delta_{2i_4}}\nabla_{\delta_{2i_5}}\nabla_{\delta_{2i_6}}\hat{B}_t$ equals

$$\begin{aligned} &\left[\frac{(\xi_*-1)^2}{\xi_*^2}(\xi_*-1)+(1-\xi_*)\right]\nabla_{\delta_{1i_4}}\nabla_{\delta_{1i_5}}\nabla_{\delta_{1i_6}}\widetilde{f}_{1t} \\ &+(\xi_*-1)\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}\sum_{u=1}^{n_\delta}[\alpha_{i_4u}\nabla_{\delta_{2i_5}}\nabla_{\delta_{2i_6}}\hat{\delta}_{1u}(\widetilde{\delta})+\alpha_{i_5u}\nabla_{\delta_{2i_4}}\nabla_{\delta_{2i_6}}\hat{\delta}_{1u}(\widetilde{\delta})+\alpha_{i_6u}\nabla_{\delta_{2i_4}}\nabla_{\delta_{2i_5}}\hat{\delta}_{1u}(\widetilde{\delta})] \\ &+\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}\begin{bmatrix} \nabla_{\delta_{2i_4}}\nabla_{\delta_{2i_5}}\nabla_{\delta_{2i_6}}\hat{\beta}(\widetilde{\delta}) \\ \xi_*\nabla_{\delta_{2i_4}}\nabla_{\delta_{2i_5}}\nabla_{\delta_{2i_6}}\hat{\delta}_1(\widetilde{\delta}) \end{bmatrix}. \end{aligned}$$

Because of (A.49), the above display equals

$$\frac{(\xi_*-1)(1-2\xi_*)}{\xi_*^2}\left\{\nabla_{\delta_{1i_4}}\nabla_{\delta_{1i_5}}\nabla_{\delta_{1i_6}}\widetilde{f}_{1t}-\left(\nabla_{(\beta',\delta_1')}\widetilde{f}_{1t}\right)\widetilde{I}^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{(\beta',\delta_1')'}\widetilde{f}_{1t}}{\widetilde{f}_t^2}\nabla_{\delta_{1i_4}}\nabla_{\delta_{1i_5}}\nabla_{\delta_{1i_6}}\widetilde{f}_{1t}\right]\right\}$$

The result follows because, when evaluated at $\widetilde{\delta}$, $\nabla_{\delta_{2i_2}}\nabla_{\delta_{2i_3}}\hat{L}_{i_1t}-[(1-\xi_*)/\xi_*]\nabla_{\delta_{2i_2}}\nabla_{\delta_{2i_3}}\hat{M}_{(n_\beta+i_1)t}$ equals $[(\xi_*-1)(1-2\xi_*)/\xi_*^2]\nabla_{\delta_{1i_1}}\nabla_{\delta_{1i_2}}\nabla_{\delta_{1i_3}}\widetilde{f}_{1t}$.

Consider $p=q=1/2$. The results for the 3rd to the 6th order derivatives follow immediately from the proofs above. The arguments for showing $T^{-1/2}\mathcal{L}^{(7)}_{i_1,\ldots i_7}(1/2,1/2,\widetilde{\delta})=O_p(1)$ are similar to those for $T^{-1/2}\mathcal{L}^{(5)}_{i_1,\ldots i_5}(p,1-q,\widetilde{\delta})$. The proof for $T^{-1/2}\mathcal{L}^{(8)}_{i_1,\ldots i_8}(1/2,1/2,\widetilde{\delta})$ is similar to that of $T^{-1/2}\mathcal{L}^{(6)}_{i_1,\ldots i_6}(p,q,\widetilde{\delta})$. We omit the details. ■

Table 1: Rejection frequencies under the null hypothesis

| | Level | 2.50 | 5.00 | 7.50 | 10.00 |
|---|---|---|---|---|---|
| T=200 | SupLR($\Lambda_{0.05}$) | 2.98 | 6.60 | 10.84 | 14.46 |
| | SupLR($\Lambda_{0.02}$) | 2.88 | 6.22 | 9.64 | 13.62 |
| | QLR | 2.43 | 5.30 | 7.50 | 10.00 |
| | supTS$_1$ | 2.76 | 5.32 | 7.72 | 9.84 |
| | supTS$_2$ | 2.38 | 5.34 | 7.76 | 9.94 |
| T=500 | SupLR($\Lambda_{0.05}$) | 2.34 | 6.34 | 9.54 | 13.50 |
| | SupLR($\Lambda_{0.02}$) | 2.38 | 6.26 | 10.24 | 13.92 |
| | QLR | 2.33 | 5.43 | 7.53 | 10.20 |
| | supTS$_1$ | 2.54 | 5.42 | 7.78 | 10.22 |
| | supTS$_2$ | 2.22 | 4.86 | 7.28 | 10.36 |

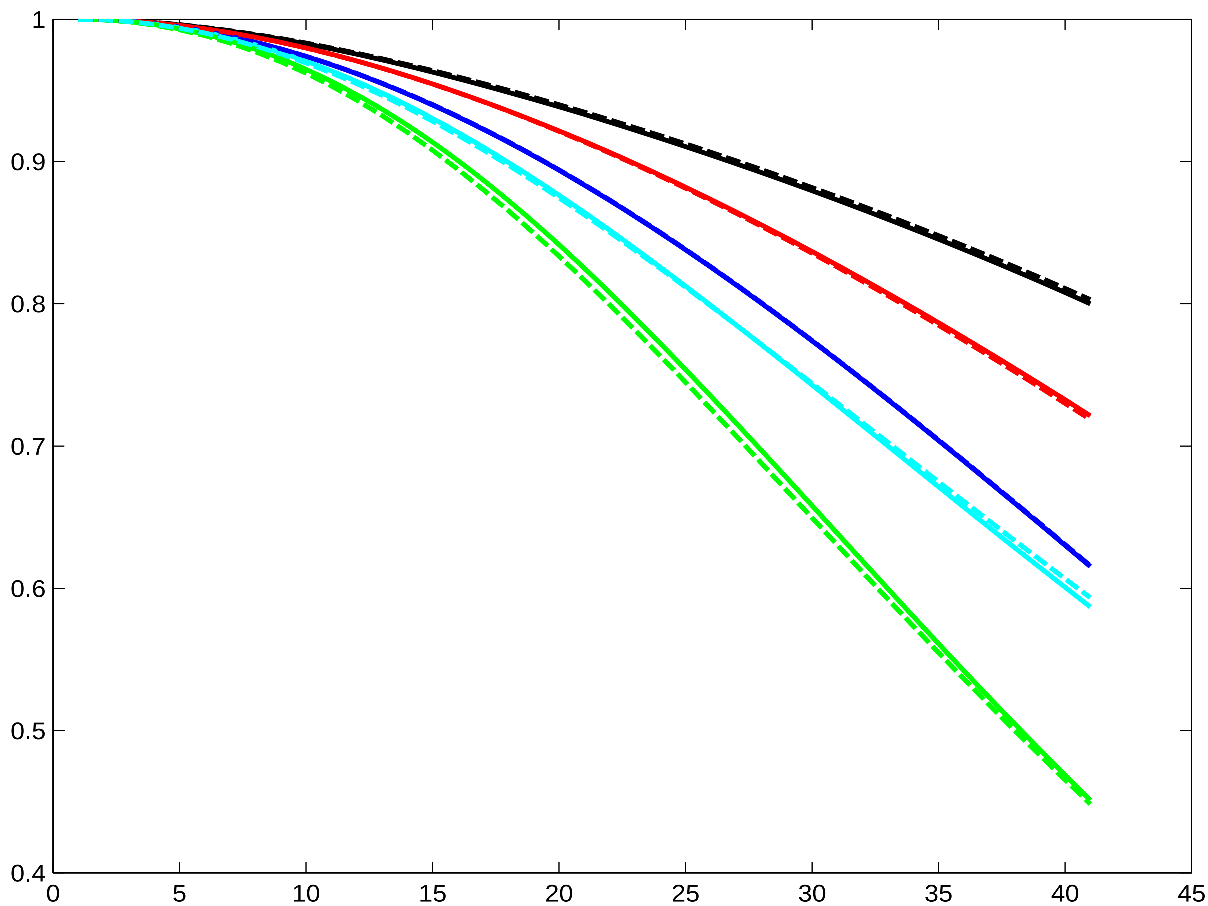
Note. The values corresponding to the QLR test are taken from Table II in Cho and White (2007). The values related to the supTS tests are obtained using the accompanying code of Carrasco, Hu and Ploberger (2014) adapted to the model considered here. Number of replications: 5000.

Table 2: Rejection frequencies under the alternative hypothesis

| $(p, q)$ | | $\mu_2 = 0.20$ | $\mu_2 = 0.60$ | $\mu_2 = 1.00$ |
|---|---|---|---|---|
| (0.70,0.70) | SupLR($\Lambda_{0.05}$) | 7.40 | 20.24 | 96.58 |
| | SupLR($\Lambda_{0.02}$) | 7.66 | 18.28 | 96.28 |
| | QLR | 6.16 | 9.46 | 68.83 |
| | supTS$_1$ | 5.68 | 10.78 | 33.60 |
| | supTS$_2$ | 5.28 | 9.66 | 32.88 |
| (0.70,0.90) | SupLR($\Lambda_{0.05}$) | 6.94 | 38.14 | 99.80 |
| | SupLR($\Lambda_{0.02}$) | 7.30 | 33.58 | 99.72 |
| | QLR | 6.14 | 13.40 | 60.56 |
| | supTS$_1$ | 4.90 | 5.50 | 19.50 |
| | supTS$_2$ | 4.94 | 5.14 | 16.90 |
| (0.90,0.90) | SupLR($\Lambda_{0.05}$) | 8.22 | 60.30 | 100.00 |
| | SupLR($\Lambda_{0.02}$) | 8.44 | 56.52 | 100.00 |
| | QLR | 5.76 | 7.06 | 7.30 |
| | supTS$_1$ | 6.66 | 11.22 | 5.54 |
| | supTS$_2$ | 6.42 | 10.86 | 5.24 |

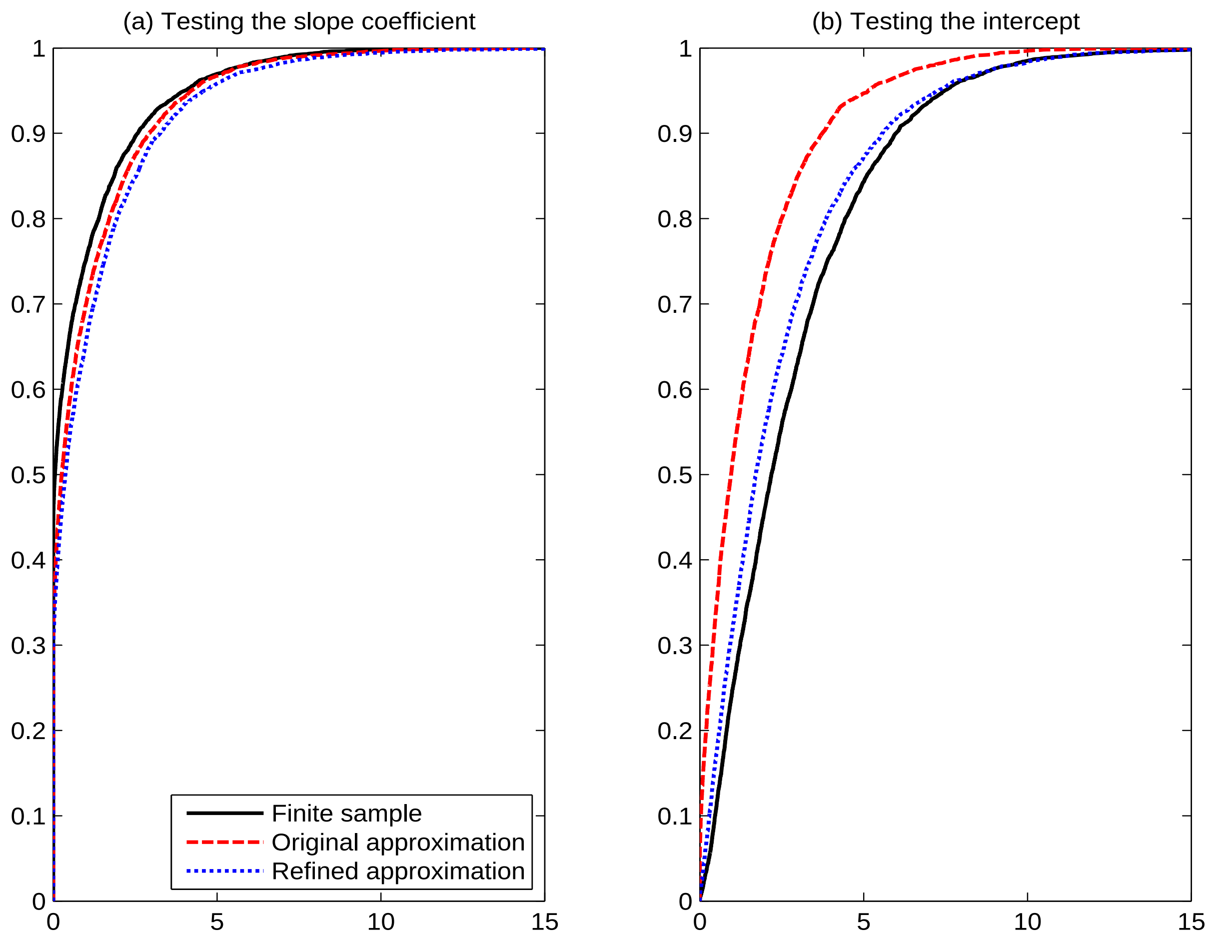
Note. The values corresponding to the QLR test are taken from Table III in Cho and White (2007). Note that there the values in the rows of 0.1 and 0.9 in their table should be exchanged. The values related to the supTS tests are obtained using the accompanying code of Carrasco, Hu and Ploberger (2014) adapted to the model considered here. Replications: 5000. Nominal level: 5%. Sample size: 500.

FIGURE 1. Correlation functions



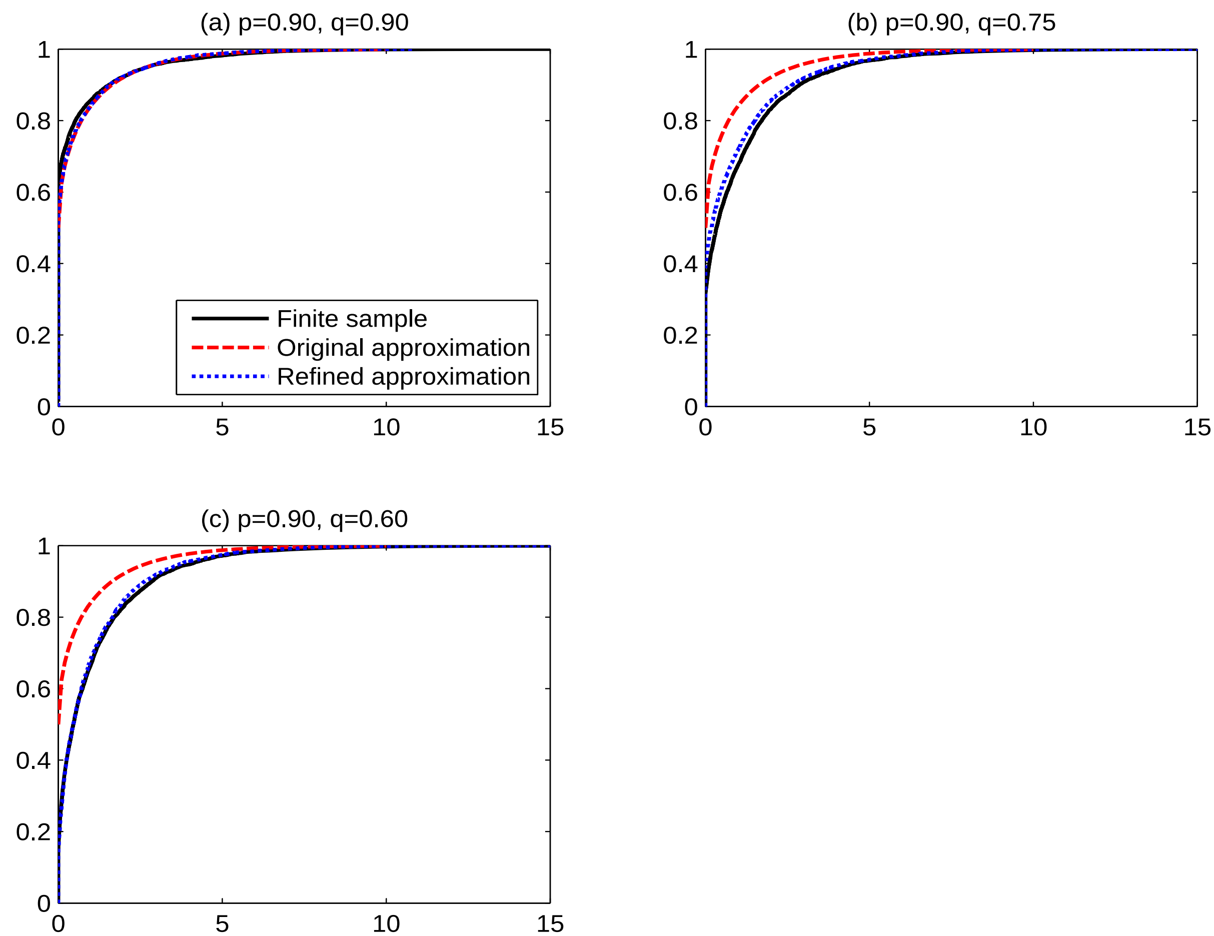

Note. The figure shows correlations between $G(p_r, q_r)$ and $G(p_s, q_s)$ with $(p_r, q_r) = (0.6, 0.9)$ and $(p_s, q_s) = (0.6, x)$, where $x$ varies between 0.1 and 0.9. The solid lines starting from the top correspond to expressions in displays (27), (25), (23), (26) and (24) in the paper. The dashed lines are correlations computed using simulations with $T = 250$.

Figure 2. Distributions in an AR(1) model



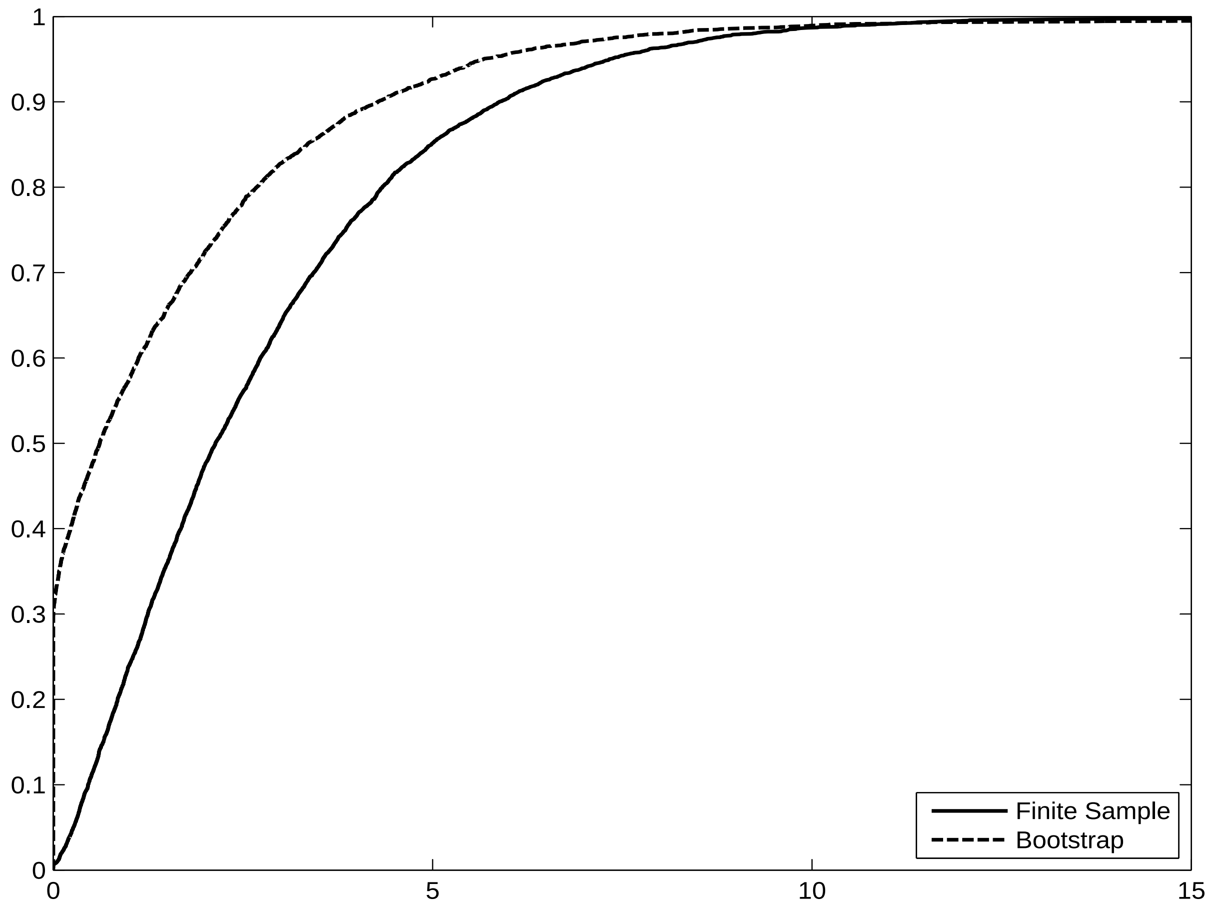

Note. The figure shows three distributions that arise when testing for regime switching in an AR(1) model: $y_t = \mu + \alpha y_{t-1} + u_t$ *with* $u_t \sim i.i.d.N(0, \sigma^2)$. The finite sample distribution is generated with $T = 250$. The original approximation corresponds to the distribution in Proposition 2. The refined approximation is given in Corollary 1.

Figure 3. Distributions when testing for switching in the intercept evaluated at fixed $p$ and $q$



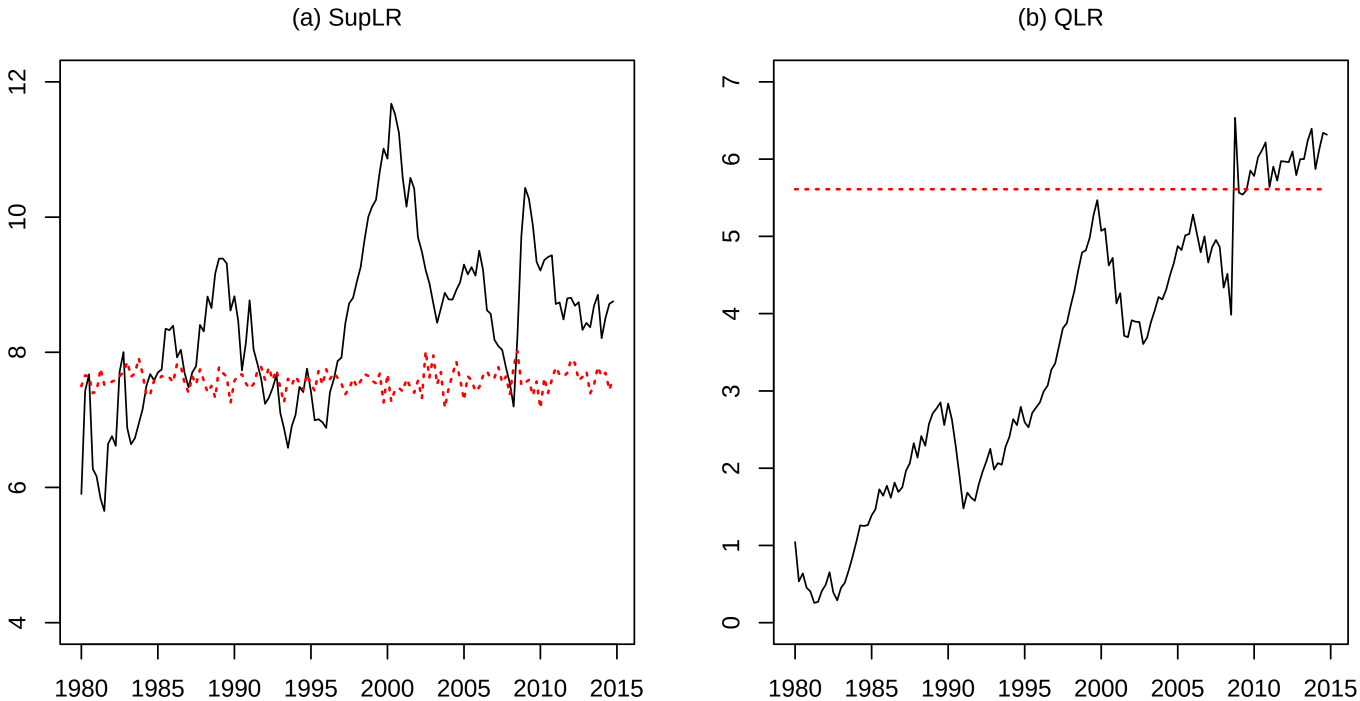

Note. See Figure 2.

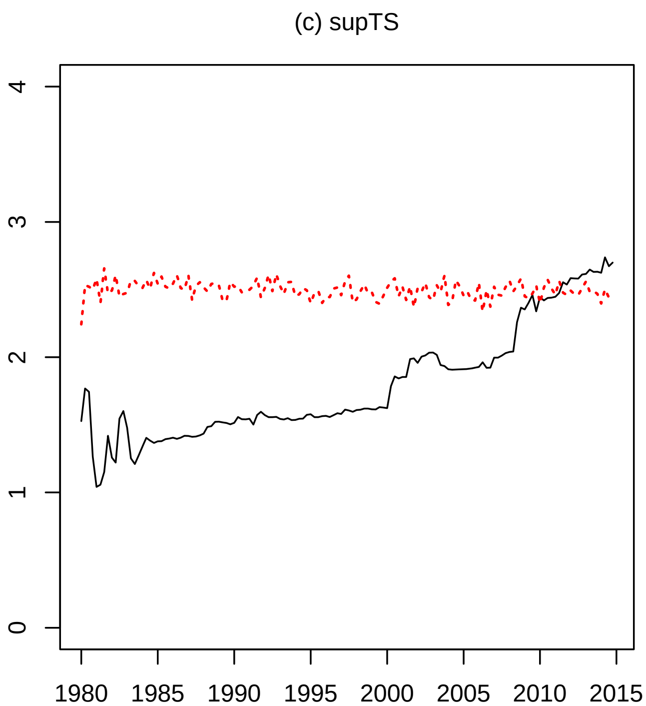
FIGURE 4. A bootstrap procedure applied to an AR(1) model




Note. The model under the null hypothesis is $y_t = \mu + \alpha y_{t-1} + u_t$ with $u_t \sim i.i.d.N(0, \sigma^2)$. The figure shows the finite sample distribution when testing for regime switching in the intercept (the solid line) and the bootstrapped distribution obtained by keeping the regressor fixed (the dashed line). $T = 250$. The true parameter values are $\mu = 0, \alpha = 0.5, \sigma = 1$.

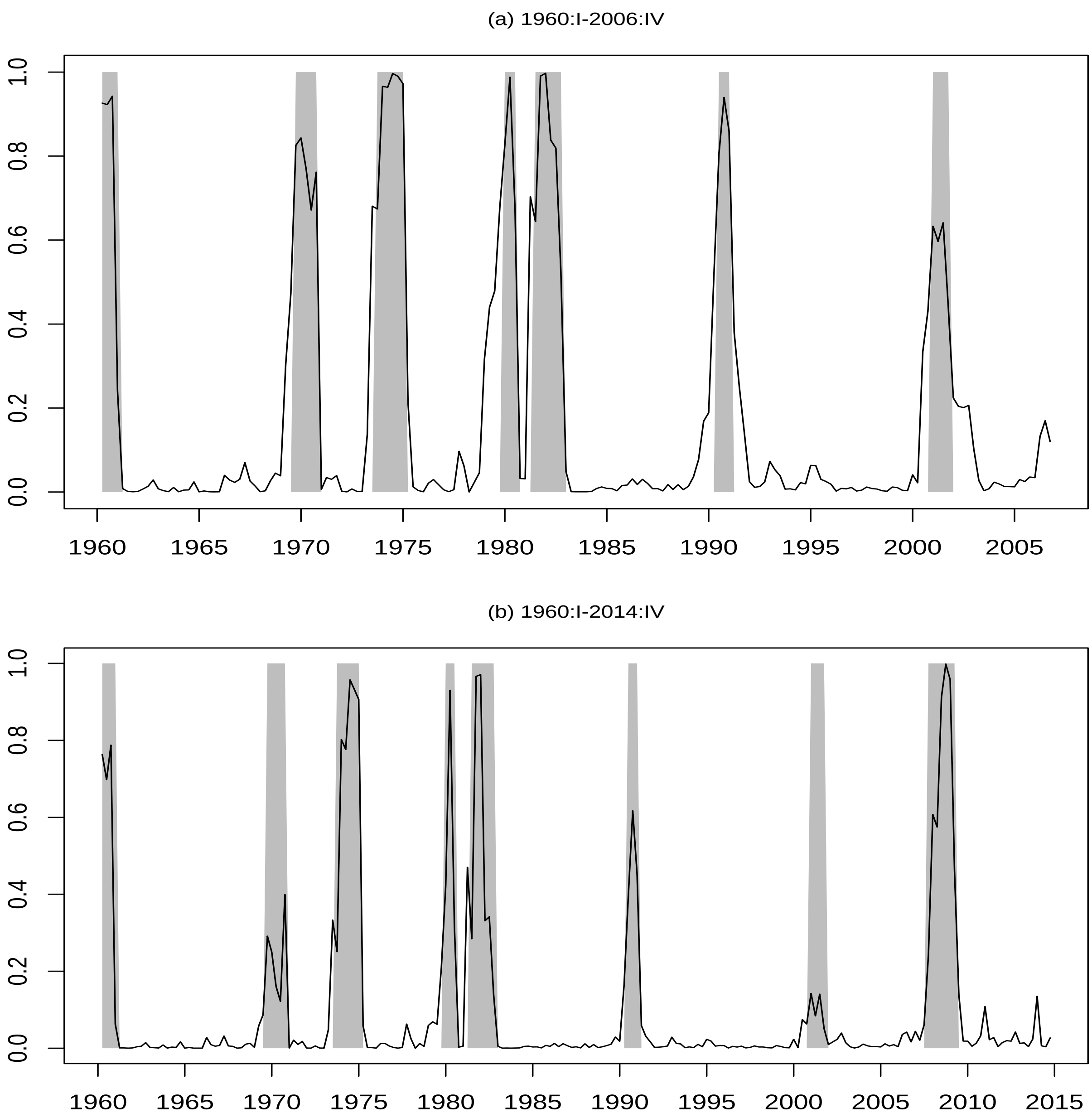
Figure 5. Test values over subsamples




Note. SupLR: the proposed test. QLR: the test of Cho and White (2007). supTS: the test of Carrasco, Hu and Ploberger (2014). The dotted lines correspond to 5% critical values.

Figure 6. Smoothed recession probabilities




Note. The solid lines are the estimates. The shaded areas correspond to NBER's recession dating.