BANK OF CANADA
BANQUE DU CANADA

Staff working paper / Document de travail du personnel—2026-23

Last updated: July 2, 2026

Monte Carlo Likelihood-Ratio Tests for Markov Switching Models

Gabriel Rodriguez Rondon
Canadian Economic Analysis Department
Bank of Canada
grodriguezrondon@bankofcanada.ca

Jean-Marie Dufour
McGill University
jean-marie.dufour@mcgill.ca

DOI: https://doi.org/10.34989/swp-2026-23 | ISSN 1701-9397

Monte Carlo Likelihood-Ratio Tests for Markov Switching Models

Gabriel Rodriguez-Rondon*
Canadian Economic Analysis Department, Bank of Canada
and
Jean-Marie Dufour
Department of Economics, McGill University

February 24, 2026

Abstract

Markov switching models are widely used to capture nonlinearities arising from regime shifts. Most existing tests for the number of regimes focus on one versus two regimes. Even in such simple cases, this type of problem raises issues of non-standard asymptotic distributions, identification failure, and nuisance parameters. We address these difficulties by applying the technique of Monte Carlo tests, which yields both finite-sample and asymptotically valid procedures, without the need to establish an asymptotic distributional theory, nor the existence of an asymptotic distribution. Monte Carlo likelihood-ratio tests are developed for testing M_0 regimes against $M_0 + m$ regimes, for any $M_0 \geq 1$ and $m \geq 1$. The proposed tests apply to nonstationary processes, non-Gaussian errors, and multivariate models. A key contribution is the Maximized Monte Carlo likelihood-ratio test (MMC-LRT), an identification-robust procedure with both finite-sample and asymptotic validity. The framework also accommodates tests for regime synchronization and Markov switching GARCH models. Simulations show accurate size control and strong power. An empirical application using Markov switching VAR models finds weakened U.S.-Canada business cycle synchronization when COVID-period data are included, while applications to U.S. output growth support a three-regime specification consistent with previous empirical studies.

Keywords: Hidden Markov models; Hypothesis testing; Finite-sample inference; Identification robustness; Nonlinear time series; Regime synchronization

*The authors thank Zhongjun Qu, Pierre Perron, Lynda Khalaf, John Galbraith, Hafedh Bouakez, Russell Davidson, Marine Carrasco, Enrique Sentana, Silvia Gonçalves, René Garcia, Nazmul Ahsan, Victoria Zinde–Walsh, Richard Luger, and participants at various seminars and conferences for their helpful comments and suggestions. Financial support from the Fonds de recherche du Québec – Société et culture (FRQSC) Doctoral Research Scholarship (B2Z) is gratefully acknowledged. **Disclaimer:** The views expressed in this article are those of the authors and do not necessarily reflect the position of the Bank of Canada.

1 Introduction

Markov regime-switching models were first introduced by Goldfeld & Quandt (1973) and later extended by Hamilton (1989). These models have since become widely used in economics and finance due to their ability to capture non-linear dynamics arising from discrete shifts in the underlying data-generating process. In such models, different regimes can represent distinct phases of the economy. For instance, in the case of U.S. output growth, one regime might correspond to positive growth (during expansions) and another to negative growth (during recessions).

Due to this flexibility, Markov switching models have been used extensively in a wide range of applications, including business cycle identification (Chauvet 1998, Chauvet & Hamilton 2006, Diebold & Rudebusch 1996, Hamilton 1989, Kim & Nelson 1999, Qin & Qu 2021), interest rate modeling (Garcia & Perron 1996), volatility modeling (Gray 1996, Hamilton & Susmel 1994, Klaassen 2002, Haas et al. 2004, Marcucci 2005, Augustyniak 2014), time-varying correlations (Pelletier 2006), state-dependent impulse response functions (Sims & Zha 2006, Caggiano et al. 2017), structural VAR identification (Herwartz & Lütkepohl 2014, Lanne et al. 2010, Lütkepohl et al. 2021), and core inflation modeling (Rodriguez-Rondon 2024, Joo Ahn & Luciani 2025, Le Bihan et al. 2024), with many additional applications both within and beyond macroeconomics and finance. For comprehensive reviews, see Hamilton (2016) and Ang & Timmermann (2012) among others.

A fundamental challenge in using Markov switching models is to determine the number of regimes, which is typically assumed *a priori*. Since the true number of regimes is unknown in practice, it is of interest to test a model with M_0 regimes against an alternative with M_0+m regimes. However, standard hypothesis testing procedures are not readily applicable in this context because key parameters are unidentified under the null, and the usual regularity conditions needed for standard asymptotic results are violated.

The literature on the asymptotic distribution of likelihood-ratio (LR) tests for Markov switching models is rich (Hansen 1992, Garcia 1998, Cho & White 2007, Kasahara & Shimotsu 2018, Qu & Zhuo 2021), with particularly important contributions – for our LR setting – being the SupLR(Λ_ϵ) test of Qu & Zhuo (2021) and early study of Garcia (1998). However, most of these contributions are limited to testing a linear null ($M_0 = 1$) against a two-regime alternative ($m = 1$). An exception is Kasahara & Shimotsu (2018), who establish asymptotic validity for a parametric bootstrap LR test for comparing M_0 and $M_0 + 1$ regime models when $M_0 \geq 1$, but only under strong and restrictive assumptions. Similarly, Qu & Zhuo (2021) derive validity results for broader classes of models, yet still within the setting of $M_0 = m = 1$.

Meanwhile, other researchers have proposed alternative test procedures based on moments of least-squares residuals (Dufour & Luger 2017), parameter stability (Carrasco et al. 2014), moment-matching conditions (Antoine et al. 2023), and score-type tests (Amengual et al. 2025*b*,*a*), all of which are also primarily designed to test linear models ($M_0 = 1$) against two-regime alternatives ($m = 1$). Moreover, most of the tests discussed thus far are only valid asymptotically and require restrictive assumptions, such as stationarity, Gaussian errors, and constrained parameter spaces, and importantly, only consider the univariate setting. In contrast, Dufour & Luger (2017) adopt the finite-sample Monte Carlo (MC) testing framework of Dufour (2006) to develop moment-based tests which are valid without relying on asymptotic theory. Although only applicable to the $M_0 = m = 1$ case, it demonstrates that finite-sample valid inference is possible without heavy distributional assumptions.

This paper builds on the MC framework of Dufour (2006), which yields both finite-sample as well as asymptotically valid procedures, without the need to establish an asymptotic distributional theory, nor even the existence of an asymptotic distribution. We develop two

likelihood-based tests for Markov switching models: the Local Monte Carlo likelihood-ratio test (LMC-LRT) and the Maximized Monte Carlo likelihood-ratio test (MMC-LRT). These procedures allow for testing hypotheses of the form $H_0 : M_0$ vs. $H_1 : M_0 + m$, where both $M_0 \geq 1$ and $m \geq 1$, and are applicable to both univariate and multivariate models – including Hidden Markov Models, Markov-switching VARs, and MS-GARCH models – areas largely ignored in prior hypothesis testing literature. The MMC-LRT procedure is an exact test valid in both finite samples and asymptotically, and it is robust to the identification problems that typically arise in regime-switching models. Both the LMC-LRT and MMC-LRT avoid the need for stationarity assumptions, Gaussianity, and constrained parameter spaces. Specifically, these tests do not rely on establishing an asymptotic distribution – nor even its existence – and, as a result, can be applied in settings where previous test procedures, including the parametric bootstrap procedure, are not asymptotically valid or settings where the asymptotic validity simply hasn't yet been established in the literature.

Notably, hypothesis testing in settings with $m > 1$, multivariate models, non-Gaussian errors, or non-stationary processes has received little attention in the literature, making this study a novel contribution. Importantly, a growing empirical literature documents evidence in favor of more than two regimes in macroeconomic and financial time series, including three or more regimes in output growth, unemployment, and interest rates (Boldin 1996, Hamilton 2005, Garcia & Perron 1996, Guidolin & Timmermann 2005, Sims & Zha 2006, Hwu et al. 2021, Kim & Kang 2022), underscoring the practical relevance of formal testing procedures which allow for $m > 1$. Our simulation results confirm that the LMC-LRT and MMC-LRT procedures maintain accurate size control and exhibit strong power across a wide range of empirically relevant scenarios – including those with multiple regimes, boundary parameters (*e.g.*, absorbing regimes), and non-stationarity. In univariate settings, both tests outperform existing moment-based and asymptotic procedures,

particularly when structural shifts occur in the mean or in both the mean and variance. The MMC-LRT delivers robust inference even in small samples, while the LMC-LRT remains computationally efficient and performs well even when a well-defined likelihood is unavailable.

While our simulations and empirical applications focus on changes in the conditional mean or variance of the outcome variable, the proposed testing procedures are equally applicable to more complex univariate settings, including Markov switching GARCH (MS-GARCH) models. These models feature regime-dependent volatility dynamics and are widely used in financial econometrics to capture shifts in conditional heteroskedasticity. Extending the testing framework to MS-GARCH models is straightforward so long as the likelihood function is available under both the null and alternative, thereby enabling formal inference on the number of volatility regimes and further broadening the scope of finite-sample valid testing in regime-switching contexts. Our methodology also enables testing whether different model components are governed by the same or distinct regime-switching processes. In univariate models, for example, one can test whether the mean and variance follow a common Markov chain. In multivariate settings, such as VAR models, the framework allows testing whether individual equations share a synchronized regime structure. This is conceptually related to testing for common structural breaks in the structural change literature (Oka & Perron 2018, Perron et al. 2020), but has not been feasible in the Markov switching context due to the technical challenges involved in testing models with multiple regimes. The framework developed here makes such testing feasible, and we illustrate this in an empirical application on international business cycle synchronization.

In the multivariate simulation evidence, the proposed tests prove valuable in detecting regime synchronization or independence. Their power depends on both the degree of misalignment between regimes and the sample size. While short-lived regime shifts may

obscure partial independence in small samples, full independence can still be detected reliably under moderate conditions. These findings underscore the practical advantages of simulation-based testing methods which remain valid in finite samples, especially when analyzing complex regime dynamics.

All test procedures are implemented using the **MSTest** R package, detailed in the companion paper Rodriguez-Rondon & Dufour (2024). The remainder of the paper is organized as follows. Section 2 reviews the notation and the Markov switching model framework. Section 3 introduces the proposed testing procedures and required assumptions. Section 4 presents simulation results, comparing the proposed tests to existing ones in univariate settings and showcasing results for multivariate cases. Section 5 provides an empirical application using Markov-switching VAR models (MS-VAR) to test for business cycle synchronization across countries. Finally, Section 6 concludes.

2 Markov-switching model

A Markov switching model is described as follows. Let (y_t, w_t) be a sequence of random vectors. The vector w_t is a finite-dimensional vector, and in this work, we allow y_t to be either a scalar (univariate setting) or a finite-dimensional vector (multivariate setting). Further, let $S_t \in \{1, \ldots, M\}$ be a latent random variable which determines the regimes at time t, and denote by $s_t \in \{1, \ldots, M\}$ a realization of S_t. We define the information set $\mathscr{Y}_{t-1} = \sigma\text{-field}\{\ldots, w_{t-1}, y_{t-2}, w_t, y_{t-1}\}$. The Markov switching model has the form:

$$y_t = x_t\beta + z_t\delta(s_t) + \sigma(s_t)\epsilon_t, \ t = 1, \ldots, T, \tag{1}$$

where x_t is a $1 \times q_x$ vector of variables whose coefficients do not depend on the latent Markov process S_t, z_t is a $1 \times q_z$ vector of variables whose coefficients do depend on the Markov process S_t, and ϵ_t is the error process. Of course, the number of non-zero coefficients in $\delta(s_t)$ depends on the regime in operation and whether the model is univariate

or multivariate. We can group all parameters in $\theta(s_t) = \Big(\beta, \delta(s_t), \sigma(s_t), vec(\mathbf{P})\Big)$, where $vec(\cdot)$ is the vectorization operator which transforms a matrix to a vector, and $\mathbf{P}$ is the transition matrix, described in more detail below. When considering the multivariate setting, we then have a covariance matrix $\Sigma(s_t)$ and make use of the $vech(\cdot)$ operator, which takes the values under and on the main diagonal of the matrix. In such cases, β and $\delta(s_t)$ are matrices and so we must use $vec(\beta)$ and $vec\Big(\delta(s_t)\Big)$ in $\theta(s_t)$.

For simplicity, we assume that the errors ϵ_t are i.i.d. $\mathcal{N}(0, I_{q_y})$. This assumption is not required for the testing procedure proposed below, as alternative distributions can be accommodated by using the appropriate likelihood density. When the true error distribution is non-Gaussian, the normal likelihood can be used as a pseudo-likelihood, in which case the resulting procedure may still be valid, as discussed in the next section.

A Markov switching model is typically described as having lags of y_t as explanatory variables in either x_t or z_t. This setting is very general and even allows one to consider a trend function within x_t or z_t. Hidden Markov models also feature a latent Markov state S_t, but are generally used when y_t does not depend on its own past. Allowing for lagged dependence yields richer interactions between y_t and S_t, which is why Markov switching models are more common in econometric applications. Hidden Markov models can therefore be viewed as a special case, and while we focus on the more general Markov switching framework, our results apply to both.

For a model with M regimes, the one-step transition probabilities can be gathered into an $M \times M$ transition matrix, $\mathbf{P}$, whose entries $p_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$ denote the probability that state i switches to state j. The columns of the transition matrix must sum to one (*i.e.*, $\sum_{j=1}^{M} p_{ij} = 1, \forall i$) and we can obtain the ergodic probabilities, $\pi = (\pi_1, \ldots, \pi_M)'$, using $\boldsymbol{\pi} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{e}_{N+1}$ where $\mathbf{A} = [\mathbf{I}_M - \mathbf{P}, \mathbf{1}']'$ and $\mathbf{e}_{M+1}$ is the $(M+1)$-th column of $\mathbf{I}_{M+1}$. These ergodic probabilities represent the proportion of time spent in each regime,

over the long-run.

Let $f(y_t|\mathscr{Y}_{t-1};\theta)$ denote the conditional density of y_t given $\mathscr{Y}_{t-1}$, and assume it satisfies

$$y_t \mid (\mathscr{Y}_{t-1}, s_t = m) \sim f(y_t \mid \mathscr{Y}_{t-1};\theta), \qquad m = 1,\ldots,M. \tag{2}$$

for $t = 1,\ldots,T$. The sample log likelihood conditional on the first p observations of y_t is given by

$$L_T(\theta) = \log f\left(y_1^T | y_{-p+1}^0;\theta\right) = \sum_{t=1}^{T} \log f\left(y_t|\mathscr{Y}_{t-1};\theta\right) \tag{3}$$

where $\theta = [\beta, \delta(1),\ldots,\delta(M),\sigma(1),\ldots,\sigma(M),vec(\mathbf{P})]$, and the $vec(\cdot)$ operator should also be applied to β, $\delta(s_t)$, and $\Sigma(s_t)$ if working with a multivariate model. Here, $f(y_t|\mathscr{Y}_{t-1};\theta) = \sum_{s_t=1}^{M}\sum_{s_{t-1}=1}^{M}\cdots\sum_{s_{t-p}=1}^{M} f(y_t, S_t = s_t, S_{t-1} = s_{t-1},\ldots,S_{t-p} = s_{t-p}|\mathscr{Y}_{t-1};\theta)$, and specifically

$$f(y_t, S_t^* = s_t^*|\mathscr{Y}_{t-1};\theta) = \frac{\Pr(S_t^* = s_t^*|\mathscr{Y}_{t-1};\theta)}{\sqrt{2\pi\sigma(s_t^*)^2}} \times \exp\left\{\frac{-[y_t - x_t\beta - z_t\delta(s_t^*)]^2}{2\sigma(s_t^*)^2}\right\} \tag{4}$$

where we set $S_t^* = s_t^*$ if $S_t = s_t, S_{t-1} = s_{t-1},\ldots,S_{t-p} = s_{t-p}$.

Markov switching and Hidden Markov models are typically estimated using the Expectation–Maximization (EM) algorithm, state-space methods based on the Kalman filter, or Bayesian estimation. In this paper, we estimate all models using the EM algorithm. Further estimation details are omitted, as our focus is on hypothesis testing; see Hamilton (1994) and Krolzig (1997) for further details on estimation procedures.

3 Monte-Carlo likelihood-ratio tests

In this section, we discuss the Maximized Monte Carlo likelihood-ratio Test (MMC-LRT) and the Local Monte Carlo likelihood-ratio test (LMC-LRT) in the context of Markov switching models, proposed in this paper. Similar to Garcia (1998) and the parametric bootstrap procedures described in Qu & Zhuo (2021) and Kasahara & Shimotsu (2018), when parameters are not identified under the null hypothesis, we assume that the null distribution depends only on the remaining parameters. The LRT approach requires us

to estimate the model under the null and alternative hypotheses in order to obtain the log-likelihoods for each model. The log-likelihood for models with $M > 1$ regimes is given by equations (3) - (4) with θ_i, where the subscript of i is used to underscore the parameter vector under the null hypothesis when $i = 0$, or under the alternative hypothesis when $i = 1$. The set $\bar{\Omega}_i$ satisfies any theoretical restrictions we wish to impose on θ_i. For example, as noted by Qu & Zhuo (2021) and Kasahara & Shimotsu (2018), for the asymptotic validity of the parametric bootstrap and the SupLR(Λ_ϵ), we would need to impose that $p_{i,j} \in (\epsilon, 1-\epsilon)$ on $\bar{\Omega}_i$. However, in our setting, this restriction is not necessary. When $M = 1$ under the null hypothesis, the log-likelihood is given by

$$L_T^0(\theta_0) = \log f(y_1^T \mid y_{-p+1}^0;\, \theta_0) = \sum_{t=1}^{T} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-[y_t - x_t\beta]^2}{2\sigma^2} \right\} \right) \tag{5}$$

and $\theta_0 = (\beta,\, \sigma)' \in \bar{\Omega}_0$. In general, $\bar{\Omega}_0$ has a dimension lower than $\bar{\Omega}_1$: H_0 is a restricted version of H_1, and for each $\theta_0 \in \bar{\Omega}_0$, we can find θ_1 such that $L_T^0(\theta_0) = L_T(\theta_1)$ for $\theta_1 \in \Omega_0$, where Ω_0 is the subset of vectors $\theta_1 \in \bar{\Omega}_1$ such that θ_1 satisfies H_0.

Under H_0, the vector $\theta_0 \in \bar{\Omega}_0$ consists of nuisance parameters: the null distribution of any test statistic for H_0 depends on $\theta_0 \in \bar{\Omega}_0$. In this context, the null distribution of the test statistic is, in fact, completely determined by θ_0. The likelihood-ratio statistic for testing H_0 against H_1 can then be expressed as: $LR_T = 2[\bar{L}_T(H_1) - \bar{L}_T(H_0)]$, where $\bar{L}_T(H_1) = \sup\{L_T(\theta_1) : \theta_1 \in \bar{\Omega}_1\}$ and $\bar{L}_T(H_0) = \sup\{L_T^0(\theta_0) : \theta_0 \in \bar{\Omega}_0\} = \sup\{L_T(\theta_1) : \theta_1 \in \Omega_0\}$, and the null distribution of LR_T depends on the parameter $\theta_0 \in \bar{\Omega}_0$. Now, let $LR_T^{(0)}$ denote a real random variable, computed from observed data when the true parameter vector is θ_0. Since the model in (1) is parametric, we can use it to generate a vector of N i.i.d. replications of LR_T for any given value of $\theta_0 \in \bar{\Omega}_0$: $LR(N,\, \theta_0) := [LR_T^{(1)}(\theta_0),\, \ldots\, , LR_T^{(N)}(\theta_0)]'$. That is, we make the following assumption.

Assumption 3.1 Exchangeability of LR statistics under null hypothesis. *$LR_T^{(0)}$ is a real random variable and $LR(N,\, \theta_0)$ a real random vector, all defined on a common prob-*

ability space $(\mathcal{F}, \mathscr{Y}_{t-1}, P_{\theta_0})$ *such that the random variables* $LR_T^{(0)}$, $LR_T^{(1)}(\theta_0)$, ..., $LR_T^{(N)}(\theta_0)$ *are exchangeable for* $\theta_0 \in \bar{\Omega}_0$, *each with distribution function* $F[x \mid \theta_0]$.

Alternatively, since these models are often used in a time series framework, especially in macroeconomic and financial applications, it is often more convenient to work with:

Assumption 3.2 I.I.D. simulated statistics under null hypothesis. $LR_T^{(0)}$ *is a real random variable and* $LR(N, \theta_0)$ *a real random vector, all defined on a common probability space* $(\mathcal{F}, \mathscr{Y}_{t-1}, P_{\theta_0})$ *such that the simulated statistics* $LR_T^{(1)}(\theta_0)$, ..., $LR_T^{(N)}(\theta_0)$ *are independent and identically distributed* (*i.i.d.*) *with common distribution function* $F[x \mid \theta_0]$, *and are independent of* $LR_T^{(0)}$.

This assumption is stronger than exchangeability but is particularly appealing in time series models. While exchangeability requires that the joint distribution of the test statistics is invariant to permutations, the i.i.d. assumption explicitly rules out dependence between the simulated statistics. In Monte Carlo test procedures, i.i.d. simulations are typically obtained by generating independent sample paths of the data-generating process (DGP) under the null. This is often easier to implement and verify in time series applications where one can simulate directly from the model for a fixed θ_0. Moreover, the i.i.d. structure simplifies the theoretical justification of the test and the computation of p-values.

Note that generating N i.i.d. replications of LR_T using (1) requires knowledge of the distribution of ϵ_t. The procedure proposed here is quite general, allowing us to consider any distribution for ϵ_t, including non-Gaussian distributions. In the case of non-Gaussian distributions, we simply need to use the appropriate likelihood function in (3) - (4) or (5). However, even when the distribution of ϵ_t is non-Gaussian or unknown, we can continue to work with the Gaussian density function. In such cases, we refer to this approach as Monte Carlo pseudo-likelihood-ratio tests. When only the innovation density is misspecified (with the functional form under H_0 correctly specified), the procedure admits asymptotic

validity under generic CLT-type conditions, although convergence may be slow in some cases (*e.g.*, for certain heavy-tailed errors). Our emphasis therefore remains on the finite-sample validity delivered by Monte Carlo simulation

Further, as noted earlier, the proposed tests remain valid even when y_t is non-stationary. In such cases, the dependence on the initial observation y_0 can, in principle, affect the distribution of the test statistic. If one were to condition explicitly on the observed initial value, the simulated series would no longer be independent across replications, although the resulting test would then be exact conditional on y_0. In this work, we do not condition on y_0 and instead propose simulating the entire series independently using the model under H_0, which preserves independence across Monte Carlo replications and ensures the validity of Assumption 3.2. This approach provides an unconditional finite-sample test which remains applicable in both stationary and non-stationary settings.

Now, letting $I(C) := 1$ if condition C holds, and $I(C) = 0$ otherwise, we define

$$\hat{F}_N[x \,|\, \theta_0] := \hat{F}_N[x;\, LR(N,\, \theta_0)] = \frac{1}{N}\sum_{i=1}^{N} I[LR_T^{(i)}(\theta_0) \leq x] \tag{6}$$

$$\hat{G}_N[x \,|\, \theta_0] := \hat{G}_N[x;\, LR(N,\, \theta_0)] = 1 - \hat{F}_N[x;\, LR(N,\, \theta_0)] \tag{7}$$

where $\hat{F}_N[x \,|\, \theta_0]$ is the sample distribution of the simulated statistics, and $\hat{G}_N[x \,|\, \theta_0]$ is the corresponding survival function. Then, the Monte Carlo p-value is given by

$$\hat{p}_N[x \,|\, \theta_0] = \frac{N\hat{G}_N[x \,|\, \theta_0] + 1}{N+1}\,. \tag{8}$$

We also make the following assumption, and state our basic validity result for MMC-LRT tests.

Assumption 3.3 Measurability of extremal simulated statistics under null hypothesis. $\sup\{\hat{G}_N[LR_T^{(0)}|\theta_0] : \theta_0 \in \bar{\Omega}_0\}$ *and* $\inf\{\hat{F}_N[LR_T^{(0)}|\theta_0] : \theta_0 \in \bar{\Omega}_0\}$ *are* $\mathscr{Y}_{t-1}$*-measurable and where* $\bar{\Omega}_0$ *is a nonempty subset of* Ω*.*

Now, we can state the following proposition which follows by direct application of Propo-

sition 4.2 in Dufour (2006).

Proposition 3.1 Validity of MMC-LRT for Markov switching models. *Let* $LR_T^{(0)}(\theta_0) = LR_T^{(0)}$, *and suppose that Assumptions* 3.1 (*or* 3.2) *and* 3.3 *hold, with*

$$Pr[LR_T^{(i)} = LR_T^{(j)}] = 0 \ \ for \ \ i \neq j, \ \ \ i, j = 1, \ldots, N. \tag{9}$$

If $\theta_0 \in \bar{\Omega}_0$, *then*

$$\Pr[\sup\{\hat{p}_N[LR_T^{(0)}|\theta_0] : \theta_0 \in \bar{\Omega}_0\} \leq \alpha] \leq \alpha \ \ for \ \ 0 \leq \alpha \leq 1. \tag{10}$$

We call the critical region considered in equation (10) the *Maximized Monte Carlo likelihood-ratio test* because $\hat{p}_N[LR_T^{(0)}|\theta_0]$ is maximized with respect to $\theta_0 \in \bar{\Omega}_0$. Due to the parametric nature of the model, it is exact irrespective of the sample size. However, the parameter space can be very large, specifically growing with the number of regressors considered and the number of regimes. Additionally, the solution may not be unique, as the maximum p-value could be obtained by more than one parameter vector. For this reason, numerical optimization methods which do not rely on derivatives are recommended to find the maximum Monte Carlo p-value within the nuisance parameter space. Such algorithms include Generalized Simulated Annealing, Genetic Algorithms, and Particle Swarm [Dufour (2006), Dufour & Neves (2019), Rodriguez-Rondon & Dufour (2024)]. As described in Dufour (2006), to facilitate optimization, it is also possible to search within a smaller consistent subset of the parameter space, denoted as C_T. A consistent set can be defined using the consistent point estimate. For example, let $\hat{\theta}_0$ be the consistent point estimate of θ_0. Then, we can define

$$C_T = \{\theta_0 \in \bar{\Omega}_0 : \| \hat{\theta}_0 - \theta_0 \| < c\} \tag{11}$$

where c is a fixed positive constant and $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^k$.

Finally, we can also define C_T to be the singleton set $C_T = \{\hat{\theta}_0\}$, which yields the Local Monte Carlo likelihood-ratio test (LMC-LRT) for Markov switching models. Here, the con-

sistent set includes only the consistent point estimate $\hat{\theta}_0$. Generic conditions for the asymptotic validity of such a test are discussed in section 5 of Dufour (2006), but these are more restrictive than those for the MMC-LRT procedure. To reflect this, we replace $\hat{F}_N[x \,|\, \theta_0]$ with $\hat{F}_{TN}[x \,|\, \theta_0] = \hat{F}_N[x; LR_T(N, \theta_0)]$ and $\hat{G}_N[x \,|\, \theta_0]$ with $\hat{G}_{TN}[x \,|\, \theta_0] = \hat{G}_N[x; LR_T(N, \theta_0)]$ where the subscript T is meant to allow the test statistics and functions to change based on increasing sample sizes. As a result, the LMC p-value is given by

$$\hat{p}_{TN}[x \,|\, \theta_0] = \frac{N\hat{G}_{TN}[x \,|\, \theta_0] + 1}{N + 1} \tag{12}$$

The asymptotic validity in this case refers to the estimate $\hat{\theta}_0$ converging to the true parameters in θ_0 as the sample size T increases. This is not related to the asymptotic validity of the critical values as desired in Hansen (1992), Garcia (1998), Cho & White (2007), Qu & Zhuo (2021), and Kasahara & Shimotsu (2018). Specifically, the LMC test can be interpreted as the finite-sample analogue of the parametric bootstrap. This is because, like the parametric bootstrap, the LMC procedure is only valid asymptotically as $T \to \infty$ but, unlike the parametric bootstrap, we do not need a large number of simulations (*e.g.*, $N \to \infty$), since we do not try to approximate the asymptotic critical values nor assume that the distribution of the test statistic converges asymptotically, or exists. Instead, we work directly with the critical values from the sample distribution $\hat{F}[x \,|\, \theta_0]$.

To be more specific, the MMC-LRT procedure will be valid even when an asymptotic distribution does not exist and the LMC-LRT procedure will also be valid as $T \to \infty$ if this is the case. This means the tests proposed here are much more general than the parametric bootstrap procedure as validity does not require stationarity or working with constrained parameter spaces, which are needed to obtain its asymptotic validity in the likelihood-ratio setting. In most cases, these assumptions are needed because otherwise the likelihood function may not be well-defined. These are cases where our procedure may again be better described as Monte Carlo pseudo-likelihood-ratio test procedures. Further, we

are directly able to deal with cases where $m > 1$, non-Gaussian settings, and multivariate settings where the asymptotic validity of the parametric bootstrap procedure has simply not yet been established in the literature. Finally, this also allows the procedure to be computationally efficient in the sense that one does not need to perform a large number of simulations with the aim of obtaining asymptotically valid critical values. In fact, as can be seen from equations (8) and (12), the number of replications N is taken into account in the calculation of the p-value both in the numerator and the denominator so that it essentially remains fixed as N increases. As discussed in Dufour (2006), building a test with level $\alpha = 0.05$ requires as few as 19 replications, but using more replications can increase the power of the test. For this reason, in our simulations, we use $N = 99$ for our Monte Carlo procedure as in Dufour & Khalaf (2003) and Dufour & Luger (2017), though it is also possible to use the procedure described in Davidson & MacKinnon (2000) to determine the optimal number of simulations to minimize experimental randomness and loss of power.

At this point we have introduced the MMC-LRT and LMC-LRT for Markov switching models. We have also described how these tests are more general than the parametric bootstrap procedure and how they are useful even in settings where y_t is a vector (multivariate setting), y_t is non-stationary, and ϵ_t is non-Gaussian. For hypothesis testing, the generality of our procedure even extends to settings where $m > 1$, ensuring finite-sample validity for the MMC-LRT procedure, and does not require working with a constrained parameter space. We believe this last feature is especially important because there may be cases where $L_T^0(\theta_0) = L_T(\theta_1)$ for values $\theta_1 \in \Omega_0$ which lie on the boundary. Consider, for example, a scenario where $M = 2$ and $p_{1,1}, p_{2,1} \to 1$. In this case, the Markov switching model with $M = 2$ may be statistically equivalent to a one-regime (no Markov switching) model. Generally, similar arguments can be made for cases where $M > 2$. As a result, we believe allowing parameters, specifically transition probabilities, to take values on the

boundary is an important feature for comparing M_0 with $M_0 + m$ regimes.

In empirical applications, the MMC-LRT and LMC-LRT serve distinct but complementary purposes. The MMC-LRT provides the exact decision rule based on the maximized Monte Carlo p-value, ensuring finite-sample validity even when some parameters are unidentified under H_0. The LMC-LRT, by contrast, uses a consistent point estimate of the nuisance parameters and is asymptotically valid (under stronger assumptions), offering a computationally simpler alternative when the parameter space is large. In practice, it is often convenient to begin with the LMC-LRT: if the LMC-LRT fails to reject, the MMC-LRT procedure cannot reject H_0 either.

This framework is also applicable to Markov switching GARCH (MS-GARCH) models. In such settings, the conditional variance evolves according to a GARCH process whose parameters switch across regimes. Estimation can be carried out using standard methods for MS-GARCH models, and the Monte Carlo test procedures proposed here remain valid so long as the likelihood function can be evaluated under both the null and alternative. This allows researchers to formally test the number of regimes in regime-dependent volatility models, a topic of growing empirical interest, especially in financial econometrics. Applications include detecting changes in volatility regimes during crises, testing for asymmetric responses to shocks, and identifying regime shifts in volatility persistence.

Another important aspect to consider is the case where regressors are weakly exogenous. So far, we have discussed simulating the test statistic by using the parametric model in (1) and i.i.d. replication of ϵ_t. In many applications of Markov switching models, where only lags of the observed data y_t are included as explanatory variables, this works perfectly. In fact, even in cases where other regressors are included, as long as they are fixed or strictly exogenous so that we can treat them as fixed in this context, we can proceed as previously discussed. However, as discussed in Qu & Zhuo (2021), for the parametric bootstrap pro-

cedure, weakly exogenous regressors can lead to size distortions. The same can be true for the LMC-LRT procedure proposed here. In such settings, if the joint distribution of the dependent variable and regressors is unknown, we propose assuming some functional form [*e.g.*, an AR(p) model], use this relationship to jointly simulate them, and then proceed as previously discussed. A closely related complication arises when regime changes are endogenous. In endogenous Markov-switching models, the latent regime process may depend on contemporaneous shocks or variables correlated with the model disturbances, capturing situations in which regime shifts and economic shocks are jointly determined [see Hwu et al. (2021), Kim & Kang (2022)]. From a testing perspective, this raises challenges similar to those posed by weakly exogenous regressors: in both cases, valid finite-sample inference requires simulating – or otherwise accounting for – the relevant dependence structure, rather than conditioning on components which are jointly determined. Developing simulation-based tests for the number of regimes which remain valid under endogenous switching, and under weak exogeneity more broadly, therefore represents a promising avenue for future research.

Overall, a central challenge in Markov switching models is the presence of identification problems which arise when testing the number of regimes. Under the null hypothesis, certain parameters, such as regime-specific means, variances, or transition probabilities, are not identified, and the likelihood function is often characterized by multiple local optima and potential label-switching across regimes. These issues violate the standard regularity conditions required for asymptotic likelihood-based inference. The proposed Monte Carlo likelihood-ratio tests address these challenges by constructing the reference distribution of the statistic through simulation rather than relying on asymptotic approximations which presume identification. As a result, these Monte Carlo procedures remain valid even when some parameters lie on the boundary of the parameter space or are undefined under H_0, and

it does not require restrictive assumptions such as Gaussianity, stationarity, or constrained transition probabilities to do so. The validity of the proposed tests ultimately rests on the ability to simulate data from the null model. This primarily requires that the functional form of the model under H_0 be correctly specified. Even if the exact distribution of the innovations is unknown, asymptotic validity may still hold under generic CLT-type conditions, as previously discussed. In practice, empirical models are inevitably approximations of the true DGP, and the MMC-LRT and LMC-LRT can therefore be viewed as assessing the adequacy of a given Markov switching specification relative to a richer alternative. In applied settings, it may also be informative to examine how the inclusion of additional covariates affects the need for further nonlinearities or additional regimes.

Taken together, these properties underscore the appeal of Monte Carlo simulation-based inference as a tool for dealing with nonstandard testing environments. The finite-sample validity of Monte Carlo tests provides robustness to many of the violations which undermine traditional asymptotic methods – particularly those related to identification failure, non-Gaussian errors, or non-stationary settings. Nonetheless, robustness to more severe forms of model misspecification remains an important area for future research (*e.g.*, models with weakly exogenous regressors or endogenous regime switching) as previously mentioned.

4 Simulation evidence

This section presents simulations on the performance of the LMC-LRT and MMC-LRT for Markov switching models proposed in this paper. Specifically, we report key results for both univariate settings and multivariate settings which illustrate the performance of the tests in detecting common or synchronized regimes. Additional results, including those for non-stationary and parameter boundary scenarios, are provided in the Appendix.

For univariate cases, we consider an AR(1) DGP in which both the mean and variance can

switch according to a Markov process S_t. Similar DGPs have been considered by Carrasco et al. (2014), Dufour & Luger (2017), and Qu & Zhuo (2021), among others. We adopt several of the same DGPs of Dufour & Luger (2017) in order to evaluate performance across a wide range of scenarios, including low and high persistence, symmetric and asymmetric regimes, changes in mean only, variance only, and both simultaneously. For cases involving a linear model under the null hypothesis (*i.e.*, $H_0 : M_0 = 1$) versus a Markov switching model with two regimes under the alternative (*i.e.*, $m = 1$), we compare the performance of our proposed tests with those of Dufour & Luger (2017) and Carrasco et al. (2014).

Given the generality of the proposed test procedures, we also consider cases where multiple regimes exist under the null (*e.g.*, $M_0 > 1$ and $m = 1$), under the alternative (*e.g.*, $M_0 = 1$ and $m > 1$), or under both (*i.e.*, $M_0 > 1$ and $m > 1$). We further examine settings with nonstationarity (*i.e.*, $\phi_1 = 1.00$) and transition probabilities at the boundary of the parameter space (*e.g.*, $p_{22} = 1$). Results for these additional cases are reported in detail within the Appendix and are briefly summarized after the main results below.

The tests proposed by Dufour & Luger (2017) are also based on the Monte Carlo methodology described in Dufour (2006), but avoid certain statistical issues associated with likelihood-ratio tests by using the moments of residuals from the restricted model. These moments are designed to capture features of a normal mixture distribution. The test uses four moments of the residuals, producing four Monte Carlo (MC) p-values. To combine these p-values, two approaches are proposed: one based on the minimum, and one based on the product of the p-values; see Dufour et al. (2004) and Dufour et al. (2014) for further discussion on combined tests. As a result, Dufour & Luger (2017) propose four tests: LMC_{min}, LMC_{prod}, MMC_{min}, and MMC_{prod}. An advantage of these methods is that they only require estimating the linear model under the null.

Carrasco et al. (2014) propose a test which is optimal for detecting inconsistencies in pa-

rameter estimates across random coefficient and Markov switching models. Their procedure is broadly designed to detect parameter heterogeneity, with the Markov switching model as a special case. Like the moment-based tests in Dufour & Luger (2017), a major benefit is that it only requires estimation under the null. However, as with Dufour & Luger (2017), it applies only when there is no regime switching under the null. To address the presence of nuisance parameters, the authors propose two alternatives: a Sup-type test, denoted supTS, following Davies (1987), and an Exponential-type test, denoted expTS, as in Andrews & Ploberger (1994). Below, when applying the supTS and expTS tests, we consider values of ρ in the interval $[\underline{\rho}, \overline{\rho}] = [-0.7, 0.7]$.

As previously noted, the consistency of the parametric bootstrap procedure when $m = 1$ has been shown by Qu & Zhuo (2021) for the case $M_0 = 1$, and by Kasahara & Shimotsu (2018) for $M_0 > 1$, albeit under more restrictive assumptions than those required by our tests. In particular, these asymptotic procedures requires constraining the parameter space away from the boundary when simulating the null distribution. Moreover, their consistency has only been established in univariate, stationary, and Gaussian contexts – though Kasahara & Shimotsu (2018) consider some non-Gaussian cases as well. Kasahara & Shimotsu (2018) also impose additional constraints on variance parameters during estimation. Given the similarities between the LMC-LRT and the bootstrap approach – especially when the process is stationary and parameters are well within the interior – we do not report results from a parametric bootstrap procedure which imposes such constraints. However, we believe the LMC-LRT results presented below can shed light on the bootstrap's performance both when its assumptions hold and when they are violated. It is important to emphasize that the primary distinction between these approaches lies in how the null distribution is estimated and, more fundamentally, in their respective assumptions regarding the existence and approximation of an asymptotic distribution.

The test procedures proposed in this paper, and those used for comparison, are implemented in the R package **MSTest** (Rodriguez-Rondon & Dufour 2025), available on CRAN and described in a companion paper by Rodriguez-Rondon & Dufour (2024). All simulation results reported below were obtained using this package. For all experiments, the nominal significance level is set to $\alpha = 0.05$, and results are based on 1,000 Monte Carlo replications.

Table 1: Empirical size of test when $M_0 = 1$

Test	$\phi = 0.10$			$\phi = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500
			$H_0 : M_0 = 1$ vs. $H_1 : M_0 + m = 2$			
LMC-LRT	4.9	4.7	4.9	5.3	5.0	4.9
MMC-LRT	1.9	1.5	1.3	0.8	0.7	0.8
LMC_{min}	5.0	3.8	5.5	5.1	4.2	5.5
LMC_{prod}	4.0	4.1	4.6	4.7	4.3	4.8
MMC_{min}	1.7	1.3	4.3	1.3	1.7	4.1
MMC_{prod}	1.6	1.8	3.6	1.4	2.5	3.8
supTS	4.8	5.1	4.8	6.0	4.5	4.7
expTS	6.8	6.2	5.2	5.4	6.9	5.5
			$H_0 : M_0 = 1$ vs. $H_1 : M_0 + m = 3$			
LMC-LRT	5.2	5.4	4.8	4.6	4.1	5.3
MMC-LRT	2.5	2.3	1.5	1.2	0.8	1.0

Notes: The DGP is specified as $(\phi, \mu, \sigma) = (\phi, 0, 1)$ where the value of ϕ varies across columns as indicated in the column headers. The nominal level is 5%. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using 1000 replications. MC tests use $N = 99$ simulations.

The results under the null hypothesis of no Markov switching (*i.e.*, $H_0 : M_0 = 1$) are reported in Table 1. The table consists of two panels: the first evaluates the alternative hypothesis of a Markov switching model with two regimes, while the second one considers a three-regime alternative. The rejection frequencies of the LMC-LRT proposed in this paper are remarkably close to the nominal significance level. As expected from theory, the MMC-LRT exhibits empirical rejection rates at or below 5% under the null hypothesis.

The results for the moment-based tests proposed by Dufour & Luger (2017), namely LMC_{min}, LMC_{prod}, MMC_{min}, and MMC_{prod}, are consistent with those of our Monte Carlo likelihood-ratio tests. The expTS test shows mild over-rejection in some cases with smaller sample sizes but performs well when $T = 500$. In contrast, the supTS test demonstrates excellent size control across all sample sizes.

Table 2: Empirical power of test when $M_0 = 1$ and $m = 1$

Test	$(p_{11}, p_{22}) = (0.90, 0.90)$						$(p_{11}, p_{22}) = (0.90, 0.50)$					
	$\phi = 0.10$			$\phi = 0.90$			$\phi = 0.10$			$\phi = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500
						$\Delta\mu$						
LMC-LRT	60.2	88.6	98.3	14.7	20.5	43.9	24.9	51.3	92.8	21.4	39.3	74.6
MMC-LRT	58.0	81.7	90.0	7.5	14.7	31.3	21.6	42.3	84.5	14.0	30.0	62.0
LMC_{min}	5.3	5.4	3.7	14.5	20.9	42.1	14.8	30.2	70.6	13.7	18.8	40.3
LMC_{prod}	4.8	4.3	4.3	16.2	22.3	43.0	12.3	24.0	56.4	14.3	20.5	42.9
MMC_{min}	1.1	2.3	1.9	6.7	13.2	33.8	6.7	20.5	61.5	5.7	11.0	31.9
MMC_{prod}	0.9	2.4	2.0	6.9	14.5	34.2	7.0	16.5	49.2	6.6	12.9	35.7
supTS	36.4	64.0	96.5	5.5	3.9	6.1	7.6	7.1	11.3	5.7	8.4	24.0
expTS	35.6	60.9	95.4	5.4	3.9	6.4	7.3	8.6	11.7	8.0	9.2	22.6
						$\Delta\sigma$						
LMC-LRT	52.4	84.1	99.8	46.0	80.9	99.8	42.1	69.0	96.2	38.7	65.5	95.1
MMC-LRT	41.8	79.7	92.6	38.0	76.8	94.3	39.1	61.3	93.2	32.9	58.0	91.3
LMC_{min}	38.1	63.6	95.5	39.5	63.3	95.2	47.8	72.7	95.5	47.4	72.2	95.6
LMC_{prod}	40.5	66.3	96.3	39.7	66.5	96.5	48.9	72.9	95.4	48.8	72.8	95.1
MMC_{min}	25.8	51.8	92.9	24.8	52.4	92.6	35.0	65.2	94.1	33.1	65.3	94.2
MMC_{prod}	28.9	57.7	95.1	27.3	57.5	94.3	35.8	64.8	94.1	34.8	65.6	94.3
supTS	32.4	58.0	98.9	32.2	67.4	91.6	29.9	46.4	94.7	30.0	50.3	92.1
expTS	40.1	62.6	99.3	54.1	84.7	92.2	43.9	68.3	95.2	52.8	78.6	93.6
						$\Delta\mu$ and $\Delta\sigma$						
LMC-LRT	81.2	98.7	100.0	39.5	70.0	98.7	77.5	97.2	100.0	58.0	87.3	99.3
MMC-LRT	78.0	94.5	100.0	25.6	66.0	96.0	74.3	96.0	100.0	48.7	79.2	96.0
LMC_{min}	53.1	80.9	99.4	35.3	60.7	92.6	84.7	97.8	100.0	66.9	89.9	99.5
LMC_{prod}	46.1	74.1	98.7	38.7	63.9	95.3	84.6	98.3	100.0	69.2	91.9	99.7
MMC_{min}	37.2	69.6	99.0	22.9	49.3	89.4	74.6	96.0	100.0	52.2	85.4	99.3
MMC_{prod}	34.2	66.0	98.1	26.3	55.5	92.7	74.9	97.0	100.0	56.0	88.1	99.7
supTS	74.0	96.0	100.0	34.0	62.9	95.4	78.0	98.0	100.0	54.0	83.3	99.4
expTS	73.3	92.0	100.0	45.6	76.0	97.0	80.0	98.3	100.0	56.2	83.4	99.7

Notes: Here, we consider $H_0 : M_0 = 1$ vs. $H_1 : M_0 + m = 2$. The DGP in the third panel is specified as $(\phi, \mu_1, \mu_2, \sigma_1, \sigma_2, p_{11}, 1 - p_{11}, 1 - p_{22}, p_{22}) = (\phi, 0, 2, 1, 2, p_{11}, 1 - p_{11}, 1 - p_{22}, p_{22})$, where the values of ϕ, p_{11}, and p_{22} vary across columns as indicated in the column headers. In the first two panels, where only the mean or the variance changes, the parameters for the constant component take the same values as those in the first regime. The nominal level is 5%. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using 1000 replications. MC tests use $N = 99$ simulations.

To study the power properties of the tests, we consider DGPs with transition probabilities $(p_{11}, p_{22}) = (0.90, 0.90)$ and $(p_{11}, p_{22}) = (0.90, 0.50)$. In both cases, the remaining transition probabilities are set as $p_{ij} = 1 - p_{ii}$ for $j \neq i$. In the first case, both regimes are symmetric and relatively persistent. Given the symmetry, the stationary distribution is $\boldsymbol{\pi} = (\pi_1, \pi_2) = (0.50, 0.50)$: on average, equal time is spent in each regime in the long run. In contrast, the second case features asymmetric regimes, with one regime being more persistent than the other. This results in $\boldsymbol{\pi} = (0.83, 0.17)$, indicating that one regime dominates in terms of long-run frequency. Table 2 reports the empirical power of the tests. Since the MMC-

LRT procedure accounts for a wider range of nuisance parameter values consistent with the null compared to the LMC-LRT, it consistently exhibits lower power across all settings. The same is true for the moment-based approaches. Specifically, the LMC$_{min}$, LMC$_{prod}$, MMC$_{min}$, and MMC$_{prod}$ procedures display the weakest power when only the mean changes and persistence is low. The supTS and expTS tests also exhibit very low power when only the mean changes and persistence is high. Qu & Zhuo (2021) offers further discussion on why the supTS test performs poorly under high persistence. In contrast, the LMC-LRT and MMC-LRT proposed here demonstrate higher power in both of these challenging scenarios involving changes in the mean only. When the variance changes, all tests exhibit improved power, although the LMC-LRT and MMC-LRT generally continue to outperform the others. This pattern remains when both the mean and variance change simultaneously, with our proposed tests maintaining a power advantage despite overall improvements across all procedures.

Overall, in the case where $H_0 : M_0 = 1$ and $H_1 : M_0+m = 2$, the LMC-LRT and MMC-LRT maintain similar size properties to the alternative tests considered, while offering superior power. This is not surprising, as the moment-based procedures, supTS, and expTS are all derived primarily from the model under the null. As such, even in relatively simple settings where other test procedures are applicable, the methods proposed here may offer a more powerful alternative.

Additional simulation results, reported in the Appendix, examine the performance of the proposed tests in non-stationary settings, boundary cases, and models with more than two regimes under the null and/or alternative. Overall, the Monte Carlo-based procedures continue to exhibit reliable size control across these scenarios, whereas competing tests tend to over-reject in non-stationary cases. Power generally increases with sample size and regime asymmetry, and remains high when testing linear models against multi-

regime alternatives. When the null already involves multiple regimes, detecting additional regimes becomes more challenging, though size control remains satisfactory; in these settings, changes in the mean appear to be a more important source of power than changes in variance. Comparisons with parametric bootstrap procedures yield broadly similar qualitative patterns, but with greater size distortions in small samples. Full details and tables are provided in the Appendix.

In summary, the univariate simulation results demonstrate that the proposed LMC-LRT and MMC-LRT procedures perform well across a wide range of empirically relevant scenarios. Both tests exhibit accurate size control, even in challenging cases involving multiple regimes, non-stationarity, and boundary parameters (absorbing regimes). Moreover, they deliver superior power relative to existing moment-based and asymptotic alternatives, particularly when changes in the mean or both mean and variance are present. While the MMC-LRT provides robust inference even in small samples, the LMC-LRT remains computationally efficient and exhibits favorable properties despite the lack of a formally defined likelihood in some settings. These findings underscore the flexibility and reliability of the proposed procedures for regime detection in univariate Markov switching models.

We now turn to evaluating the performance of the proposed LMC-LRT and MMC-LRT in multivariate settings. These models are particularly relevant in empirical applications involving multiple macroeconomic or financial time series, whether across countries, asset classes, or sectors. They are also suitable for testing hypotheses about common regime structures, including–but not limited to–synchronized dynamics between variables. We revisit such an application in the empirical section that follows.

Simulation results for multivariate settings, reported in the Appendix, evaluate the size and power of the proposed tests in bivariate VAR models when testing a single-regime null against a two-regime alternative. Across a range of scenarios involving shifts in the mean,

variance, or both, the LMC-LRT and MMC-LRT exhibit reliable size control and increasing power with sample size. Consistent with the univariate results, the MMC-LRT performs particularly well in small samples and in highly persistent or heteroskedastic environments. Overall, these findings confirm that the proposed procedures extend naturally to multivariate settings and are well suited for testing hypotheses involving common or synchronized regime dynamics. Full simulation details and results are provided in the Appendix.

We now shift focus on apply the proposed tests in a related but distinct context: testing whether regime structures are synchronized across equations. While a similar logic could be applied in univariate settings – *e.g.*, where different coefficients such as the mean and variance follow separate regime paths as considered in Sims & Zha (2006) – the notion of (de)synchronization is particularly intuitive in a multivariate framework. In this setting, one may wish to test whether each equation in a system is governed by a common or by independent regime-switching processes. An illustrative example is discussed in Section A.1 of the Appendix. Importantly, this question can be framed as a test on the number of regimes in an MS-VAR model: $H_0: \ M_0 = 2$ (synchronized cycles), $H_{1a}: \ M_0 + m = 3$ (partial dependence), and $H_{1b}: \ M_0 + m = 4$ (independent cycles).

Related work in the structural break literature has proposed methods for detecting common breaks across equations [see Oka & Perron (2018)] or across coefficients within a single equation [see Perron et al. (2020)]. However, such tools are not readily available in the Markov-switching framework, owing to the complexity of estimation and inference with multiple regimes. The framework proposed here fills this gap, enabling formal hypothesis tests for synchronized regime structures in multivariate systems.

To illustrate this approach, we simulate data from a bivariate MS-VAR model in which each equation may be governed by its own Markov process. The objective is to test whether the regime-switching processes are synchronized (*i.e.*, driven by a common latent state) or

independent. When both equations follow the same regime path, only two joint states are required; when they evolve independently, up to four joint regimes may arise, corresponding to all possible state combinations. In our simulations, both series follow two-regime Markov processes, and desynchronization is introduced through lead–lag differences in regime transitions. The extent of misalignment is controlled by a parameter ξ, scaled by the sample size T, which determines the duration of time spent in unsynchronized regimes.

Table 3: Empirical size and power of test for independent Markov processes

	$H_0 : M = 2$ vs. $H_0 : M = 3$						$H_0 : M = 2$ vs. $H_0 : M = 4$					
Test	Empirical size											
	T=100		T=200		T=500		T=100		T=200		T=500	
LMC-LRT	1.2		2.6		4.0		3.0		4.0		5.8	
MMC-LRT	1.0		1.8		2.8		2.1		2.2		3.1	
	Empirical power											
	$\xi = 0.02$			$\xi = 0.10$			$\xi = 0.02$			$\xi = 0.10$		
	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500
LMC-LRT	5.0	8.0	41.2	3.8	52.8	98.8	28.6	41.0	90.8	42.0	100.0	100.0
MMC-LRT	3.4	7.3	34.6	2.0	44.9	92.6	20.9	37.1	85.2	38.2	98.9	100.0

Notes: Each equation is generated with $M = 2$ regimes and $P = [0.90, 0.10; 0.10, 0.90]$. The DGP is specified as $(\mu_{a,1}, \mu_{b,1}, \mu_{a,2}, \mu_{b,2}, \sigma^2_{a,1}, \sigma_{ab,1}, \sigma^2_{b,1}, \sigma^2_{a,2}, \sigma_{ab,2}, \sigma^2_{b,2}) = (2.00, 2.00, -2.00, -2.00, 1.00, 0.75, 1.00, 3.00, 2.25, 3.00)$. For empirical power, the parameter ξ along with the sample size T determine the duration of the (lead or lag) third and/or fourth regimes (i. e., duration is $T \times \xi$). The nominal level is 5%. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests. Rejection frequencies are obtained using 500 replications. MC tests use $N = 99$ simulations.

Table 3 reports the size and power of the LMC-LRT and MMC-LRT in such settings. As before, the DGP allows both the mean and the covariance matrix in the bivariate MS-VAR model to vary according to a latent Markov process. The top panel presents empirical size results under the null hypothesis of $M = 2$ regimes, tested against alternatives with $M = 3$ (left) and $M = 4$ (right). The LMC-LRT maintains rejection frequencies close to the nominal 5% level, though some mild under-rejection is observed when testing against the $M = 3$ alternative. As expected, the MMC-LRT behaves consistently with the theoretical properties outlined in Dufour (2006), maintaining rejection frequencies at or below the nominal level in all cases.

To evaluate power, the lower panel reports rejection frequencies under DGPs with three

or four regimes. These represent cases of partial (three regimes) or full (four regimes) independence between the two Markov processes. The parameter ξ controls the duration of the extra regime(s), determining the degree of temporal offset between the processes (*e.g.*, lead-lag behavior). We consider two values, $\xi \in \{0.02, 0.10\}$, so that the regime duration scales proportionally with sample size: $T \times \xi$. The power results suggest that detecting partial desynchronization (*i.e.*, three regimes) is challenging in smaller samples unless the third regime is sufficiently persistent (*e.g.*, $T = 200$ with $\xi = 0.10$). In contrast, both tests exhibit high power in detecting full independence (*i.e.*, four regimes), even when regime durations are short. This reflects the greater distinctness of regime paths in the four-regime case, which facilitates detection.

In summary, the ability to detect synchronized versus independent regime structures using the proposed tests depends on both the extent of misalignment and the available sample size. Short-lived regimes can obscure partial independence unless a sufficiently long series is observed. However, full independence is easier to detect, even in moderate samples. These findings highlight the practical value of accurate finite-sample testing procedures when analyzing regime dynamics in multivariate time series.

5 Synchronization of business cycles

The synchronization of business cycles has re-emerged as a topic of interest in light of recent global events, including the COVID-19 pandemic, persistent supply chain disruptions, and tariff disputes. These shocks have highlighted vulnerabilities associated with deep economic integration. Existing literature suggests that trade openness tends to amplify business cycle comovement across countries [*e.g.*, Dées & Zorell (2012)]. While a variety of methodologies have been proposed to measure business cycle synchronization, relatively few formal testing procedures exist. Moreover, many available approaches yield mixed re-

sults, rely on restrictive assumptions (*e.g.*, linearity), focus primarily on correlations, or lack theoretical validation–particularly in finite samples.

For our empirical application, we adopt a setup similar to the one of Phillips (1991), using quarterly industrial production (IP) data and, since IP captures only the supply side of the economy, we also include real GDP data following Camacho & Perez-Quiros (2006), for the United States, Canada, the United Kingdom, and Germany. Specifically, we apply the LMC-LRT and MMC-LRT tests proposed here to bivariate MS-VAR(1) models for three country pairs: (1) US–Canada, (2) US–UK, and (3) US–Germany. These tests allow us to evaluate the null hypothesis of perfectly synchronized business cycles against alternatives involving partial or full independence.

This approach offers two key advantages. First, it allows the data to determine the appropriate regime structure, which may include multiple types of recessionary or expansionary states which differ in timing or magnitude. Second, it accommodates various forms of desynchronization. In addition, the MMC variant provides exact inference in small samples; a critical advantage when using quarterly macroeconomic data. To focus on the second case, we use seasonally adjusted quarterly data and examine two samples: 1985:I–2019:IV and 1985:I–2022:IV. The starting point avoids earlier volatility shifts associated with the Great Moderation and the second sample includes the COVID-19 period. For each country pair, we estimate a bivariate MS-VAR(1), as determined by a bottom-up likelihood-ratio testing procedure. This specification is consistent with Phillips (1991) and other studies which typically use one or no lags.

Figures 8 and 9 in the Appendix display the GDP and IP series respectively, for each country and sample. The extreme fluctuations during the COVID-19 episode justify showing the full and pre-COVID samples separately. As evident in the full-sample figures, the COVID-19 shock introduces sharp but short-lived volatility. Although some literature recommends

treating such shocks as outliers or structural breaks, we do not do so here. Instead, we use unadjusted data throughout and leave robustness checks for future work. Notably, in a univariate application using U.S. GDP growth (included in Appendix), we show that controlling for COVID-19 as a structural break in the conditional mean has little effect on the estimated number of regimes. Still, exploring alternative treatments for such shocks remains an important avenue for future research.

Table 4: Results For Synchronization of Business Cycle Hypothesis Tests using GDP series

Series	$H_0 : M = 1$ vs. $H_1 : M = 2$		$H_0 : M = 2$ vs. $H_1 : M = 3$		$H_0 : M = 2$ vs. $H_1 : M = 4$	
	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT
			1985:I - 2019:IV ($T = 140$)			
US-CA	0.02	0.04	0.20	0.65	0.17	0.67
US-UK	0.01	0.01	0.01	0.01	0.01	0.01
US-GR	0.03	0.05	0.27	0.54	0.11	0.51
			1985:I - 2022:IV ($T = 155$)			
US-CA	0.01	0.01	0.08	0.43	0.03	0.05
US-UK	0.01	0.01	0.13	0.21	0.01	0.01
US-GR	0.01	0.01	0.21	0.53	0.04	0.06

Notes: This table includes results when $\Delta\mu$ and $\Delta\sigma$ as it is a statistically preferred model over a model where only $\Delta\mu$. The GDP series are OECD Main Economic Indicator Releases obtained from the St. Louis Fed (FRED) website. All MC test results are obtained using $N = 99$. The MMC-LRT procedure uses a particle swarm optimization algorithm.

Tables 4 here and Table 16 in the Appendix report the results for real GDP and IP, respectively. Each table contains results for both the pre-COVID sample ending in 2019:IV (top panel) and the full sample ending in 2022:IV (bottom panel). We begin by testing $H_0 : M_0 = 1$ against $H_a : M_0 + m = 2$ to assess whether regime-switching dynamics are present. The first two columns of each table confirm significant regime-switching behavior across all country pairs, justifying $M_0 = 2$ as a reasonable baseline. To test for synchronization, we next evaluate $H_0 : M_0 = 2$ against $H_{1a} : M_0 + m = 3$ and $H_{1b} : M_0 + m = 4$, which test for partial and full independence, respectively. Results using real GDP (Table 4) show that the US business cycle was synchronized with Canada and Germany prior to COVID-19, but not with the UK. In the full sample, all country pairs show signs of increased desynchronization, with particularly strong divergence post-COVID. Results on Industrial

production (Table 16) show broadly similar patterns with respect to the US and Canada relationship.

These findings provide preliminary evidence that international business cycle synchronization weakened following the COVID-19 shock. A plausible explanation is that national recoveries varied substantially in timing, policy response, and structural characteristics.

6 Conclusion

This paper proposes two new Monte Carlo likelihood-ratio tests for determining the number of regimes in Markov switching models, namely, the Local Monte Carlo likelihood-ratio test (LMC-LRT) and the Maximized Monte Carlo likelihood-ratio test (MMC-LRT). Building on the Monte Carlo testing framework of Dufour (2006), these procedures provide valid inference in settings where conventional likelihood-based tests fail due to unidentified nuisance parameters, nonstandard asymptotic distributions, or restrictive regularity conditions. The proposed tests allow testing a null hypothesis with M_0 regimes against an alternative with $M_0 + m$ regimes for any $M_0 \geq 1$ and $m \geq 1$, and applies to non-stationary processes, non-Gaussian errors, and multivariate models. Further, the MMC-LRT is identification-robust and valid both asymptotically and in finite samples. Simulation results show that both procedures control size accurately and exhibit strong power across a wide range of empirically relevant designs, including cases where existing methods are invalid or unavailable. In a multivariate application, the proposed methods detect a weakening of international business cycle synchronization when COVID-19 data are included. Applications to U.S. output growth, included in the Appendix, supports a three-regime specification consistent with various empirical works.

Overall, the proposed methodology offers a general and practical framework for determining the number of regimes in univariate and multivariate settings. Interesting avenues for

future work involve applying these methods to Markov switching factor models, where both the regime process and the factors are latent, and extending the framework to settings with richer dependence structures (*e.g.*, weakly exogenous regressors or endogenous regime switching) where valid simulation under the null requires joint modeling of the relevant processes. These extensions are the subject of ongoing research.

Declaration of interest

The authors declare no conflicts of interest.

Supplementary material and data availability

Results are obtained using the `MSTest` R package (Rodriguez-Rondon & Dufour 2024). The appendix provides additional simulation and empirical results. Data is from OECD, "Main Economic Indicators - complete database", https://dx.doi.org/10.1787/data-00052-en via St. Louis Fed (FRED) (Accessed on 25 September, 2023).

References

Amengual, D., Bei, X., Carrasco, M. & Sentana, E. (2025*a*), 'Score-type tests for Markov switching models', *Work in Progress* .

Amengual, D., Bei, X., Carrasco, M. & Sentana, E. (2025*b*), 'Score-type tests for normal mixtures', *Journal of Econometrics* **248**, 105717.

Andrews, D. W. & Ploberger, W. (1994), 'Optimal tests when a nuisance parameter is present only under the alternative', *Econometrica* **62**(6), 1383–1414.

Ang, A. & Timmermann, A. (2012), 'Regime changes and financial markets', *Annual Review of Financial Economics* **4**, 313–337.

Antoine, B., Khalaf, L., Kichian, M. & Lin, Z. (2023), 'Identification-robust inference

with simulation-based pseudo-matching', *Journal of Business & Economic Statistics* **41**(2), 321–338.

Augustyniak, M. (2014), 'Maximum likelihood estimation of the Markov-switching GARCH model', *Computational Statistics & Data Analysis* **76**, 61–75.

Boldin, M. D. (1996), 'A check on the robustness of hamilton's Markov switching model approach to the economic analysis of the business cycle', *Studies in Nonlinear Dynamics and Econometrics* **1**(1), 35–46.

Caggiano, G., Castelnuovo, E. & Figueres, J. M. (2017), 'Economic policy uncertainty and unemployment in the United States: A nonlinear approach', *Journal of Applied Econometrics* **32**(2), 281–298.

Camacho, M. & Perez-Quiros, G. (2006), 'A new framework to analyze business cycle synchronization', *Contributions to Economic Analysis* **276**, 133–149.

Carrasco, M., Hu, L. & Ploberger, W. (2014), 'Optimal test for Markov switching parameters', *Econometrica* **82**(2), 765–784.

Chauvet, M. (1998), 'An econometric characterization of business cycle dynamics with factor structure and regime switching', *International Economic Review* **39**(4), 969–996.

Chauvet, M. & Hamilton, J. D. (2006), 'Dating business cycle turning points', *Contributions to Economic Analysis* **276**, 1–54.

Cho, J.-S. & White, H. (2007), 'Testing for regime switching', *Econometrica* **75**(6), 1671–1720.

Davidson, R. & MacKinnon, J. G. (2000), 'Bootstrap tests: how many bootstraps?', *Econometric Reviews* **19**(1), 55–68.

Davies, R. B. (1987), 'Hypothesis testing when a nuisance parameter is present only under the alternative', *Biometrika* **74**(1), 33–43.

Dées, S. & Zorell, N. (2012), 'Business cycle synchronisation: disentangling trade and financial linkages', *Open Economies Review* **23**(4), 623–643.

Diebold, F. X. & Rudebusch, G. D. (1996), 'Measuring business cycles: A modern perspective', *Review of Economics and Statistics* **78**(1), 67–77.

Dufour, J.-M. (2006), 'Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics', *Journal of Econometrics* **133**(2), 443–477.

Dufour, J.-M. & Khalaf, L. (2003), *Monte Carlo test methods in Econometrics*, Wiley, chapter 23, pp. 494–519.

Dufour, J.-M., Khalaf, L., Bernard, J.-T. & Genest, I. (2004), 'Simulation-based finite-sample tests for heteroskedasticity and ARCH effects', *Journal of Econometrics* **122**(2), 317–347.

Dufour, J.-M., Khalaf, L. & Voia, M. (2014), 'Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability', *Communications in Statistics-Simulation and Computation* **44**(9), 2329–2347.

Dufour, J.-M. & Luger, R. (2017), 'Identification-robust moment-based tests for Markov switching in autoregressive models', *Econometric Reviews* **36**(6-9), 713–727.

Dufour, J.-M. & Neves, J. (2019), Finite-sample inference and nonstandard asymptotics with Monte Carlo tests and R, *in* 'Handbook of Statistics', Vol. 41, Elsevier, pp. 3–31.

Garcia, R. (1998), 'Asymptotic null distribution of the likelihood ratio test in Markov switching models', *International Economic Review* **39**(3), 763–788.

Garcia, R. & Perron, P. (1996), 'An analysis of the real interest rate under regime shifts', *Review of Economics and Statistics* **78**(1), 111–125.

Goldfeld, S. M. & Quandt, R. E. (1973), 'A Markov model for switching regressions', *Journal of Econometrics* **1**(1), 3–15.

Gray, S. F. (1996), 'Modeling the conditional distribution of interest rates as a regime-switching process', *Journal of Financial Economics* **42**(1), 27–62.

Guidolin, M. & Timmermann, A. (2005), 'Economic implications of bull and bear regimes in UK stock and bond returns', *The Economic Journal* **115**(500), 111–143.

Haas, M., Mittnik, S. & Paolella, M. S. (2004), 'A new approach to Markov-switching GARCH models', *Journal of Financial Econometrics* **2**(4), 493–530.

Hamilton, J. D. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica* **57**(2), 357–384.

Hamilton, J. D. (1994), *Time series analysis*, Princeton university press.

Hamilton, J. D. (2005), What's real about the business cycle?, Technical report, National Bureau of Economic Research.

Hamilton, J. D. (2016), 'Macroeconomic regimes and regime shifts', *Handbook of Macroeconomics* **2**, 163–201.

Hamilton, J. D. & Susmel, R. (1994), 'Autoregressive conditional heteroskedasticity and changes in regime', *Journal of Econometrics* **64**(1-2), 307–333.

Hansen, B. E. (1992), 'The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP', *Journal of Applied Econometrics* **7**(S1), S61–S82.

Herwartz, H. & Lütkepohl, H. (2014), 'Structural vector autoregressions with Markov switching: Combining conventional with statistical identification of shocks', *Journal of Econometrics* **183**(1), 104–116.

Hwu, S.-T., Kim, C.-J. & Piger, J. (2021), 'An N-state endogenous Markov-switching

model with applications in macroeconomics and finance', *Macroeconomic Dynamics* **25**(8), 1937–1965.

Joo Ahn, H. & Luciani, M. (2025), 'Common and idiosyncratic inflation', *Journal of Applied Econometrics* .

Kasahara, H. & Shimotsu, K. (2018), 'Testing the number of regimes in Markov regime switching models', *arXiv preprint arXiv:1801.06862* .

Kim, C.-J. & Nelson, C. R. (1999), 'Has the us economy become more stable? a Bayesian approach based on a Markov-switching model of the business cycle', *Review of Economics and Statistics* **81**(4), 608–616.

Kim, Y. M. & Kang, K. H. (2022), 'Bayesian inference of multivariate regression models with endogenous Markov regime-switching parameters', *Journal of Financial Econometrics* **20**(3), 391–436.

Klaassen, F. (2002), Improving GARCH volatility forecasts with regime-switching GARCH, *in* 'Advances in Markov-switching models', Springer, pp. 223–254.

Krolzig, H.-M. (1997), *Markov-Switching Vector Autoregressions*, Springer.

Lanne, M., Lütkepohl, H. & Maciejowska, K. (2010), 'Structural vector autoregressions with Markov switching', *Journal of Economic Dynamics and Control* **34**(2), 121–131.

Le Bihan, H., Leiva-León, D. & Pacce, M. (2024), 'Underlying inflation and asymmetric risks', *Review of Economics and Statistics* pp. 1–45.

Lütkepohl, H., Meitz, M., Netšunajev, A. & Saikkonen, P. (2021), 'Testing identification via heteroskedasticity in structural vector autoregressive models', *The Econometrics Journal* **24**(1), 1–22.

Marcucci, J. (2005), 'Forecasting stock market volatility with regime-switching GARCH models', *Studies in Nonlinear Dynamics & Econometrics* **9**(4).

Oka, T. & Perron, P. (2018), 'Testing for common breaks in a multiple equations system', *Journal of Econometrics* **204**(1), 66–85.

Pelletier, D. (2006), 'Regime switching for dynamic correlations', *Journal of Econometrics* **131**(1-2), 445–473.

Perron, P., Yamamoto, Y. & Zhou, J. (2020), 'Testing jointly for structural changes in the error variance and coefficients of a linear regression model', *Quantitative Economics* **11**(3), 1019–1057.

Phillips, K. L. (1991), 'A two-country model of stochastic output with changes in regime', *Journal of International Economics* **31**(1-2), 121 – 142.

Qin, A. & Qu, Z. (2021), 'Modeling regime switching in high-dimensional data with applications to U.S. business cycles', *Working Paper, Boston University* .

Qu, Z. & Zhuo, F. (2021), 'Likelihood ratio-based tests for Markov regime switching', *The Review of Economic Studies* **88**(2), 937–968.

Rodriguez-Rondon, G. (2024), 'Underlying core inflation with multiple regimes', *arXiv preprint arXiv:2411.12845* .

Rodriguez-Rondon, G. & Dufour, J.-M. (2024), 'MSTest: An R-package for testing Markov switching models', *arXiv preprint arXiv:2411.08188* .

Rodriguez-Rondon, G. & Dufour, J.-M. (2025), *MSTest: Hypothesis Testing for Markov Switching Models*. R package version 0.1.7.

Sims, C. A. & Zha, T. (2006), 'Were there regime switches in US monetary policy?', *American Economic Review* **96**(1), 54–81.

Monte Carlo Likelihood-Ratio Tests for Markov Switching Models: Appendix

Gabriel Rodriguez-Rondon and Jean-Marie Dufour

Bank of Canada and McGill University

February 24, 2026

A.1 Example of synchronization in bivariate model

Consider the following bivariate model for economies a and b:

$$y_{a,t} = \mu_{a,s_{a,t}} + \sum_{k=1}^{p} \phi_{aa,k}(y_{a,t-k} - \mu_{a,s_{a,t-k}}) + \sum_{k=1}^{p} \phi_{ab,k}(y_{b,t-k} - \mu_{b,s_{b,t-k}}) + \sigma_{a,s_{a,t}}\epsilon_{a,t} \quad \text{(A.1)}$$

$$y_{b,t} = \mu_{b,s_{b,t}} + \sum_{k=1}^{p} \phi_{ba,k}(y_{a,t-k} - \mu_{a,s_{a,t-k}}) + \sum_{k=1}^{p} \phi_{bb,k}(y_{b,t-k} - \mu_{b,s_{b,t-k}}) + \sigma_{b,s_{b,t}}\epsilon_{b,t} \quad \text{(A.2)}$$

We are interested in testing whether the regime-switching behavior of the two economies is governed by the same latent Markov process (*i.e.*, $S_{a,t} = S_{b,t} = S_t$) or by independent processes (i.e., $S_{a,t} \neq S_{b,t}$). Suppose both $S_{a,t}$ and $S_{b,t}$ each take values in $\{1, 2\}$. Then the joint state S_t^* must account for up to four possible combinations:

$$(A)\ S_t^* = 1 \quad \text{if} \quad S_{a,t} = 1 \text{ and } S_{b,t} = 1;\ (B)\ S_t^* = 2 \quad \text{if} \quad S_{a,t} = 1 \text{ and } S_{b,t} = 2;$$

$$(C)\ S_t^* = 3 \quad \text{if} \quad S_{a,t} = 2 \text{ and } S_{b,t} = 1;\ (D)\ S_t^* = 4 \quad \text{if} \quad S_{a,t} = 2 \text{ and } S_{b,t} = 2.$$

Synchronized regimes imply only two of these occur (*e.g.*, $S_t^* = 1$ if $S_{a,t} = 1$ and $S_{b,t} = 1$ and $S_t^* = 2$ if $S_{a,t} = 2$ $S_{b,t} = 2$), while desynchronization leads to three or four observable combinations, depending on the degree of offset between $s_{a,t}$ and $s_{b,t}$.

A.2 Additional simulation results

As previously discussed, a notable feature of the LMC-LRT and MMC-LRT is their applicability even when the process is non-stationary or contains parameters on the boundary of the parameter space. As mentioned in Section 3, in such cases the likelihood function is not theoretically well-defined. Therefore, in these scenarios, our procedures are more accurately described as Local Monte Carlo and Maximized Monte Carlo *pseudo* likelihood-ratio tests. While this distinction is important, we continue to refer to them as LMC-LRT and MMC-LRT for consistency.

Table 5 reports the rejection frequencies under both the null and alternative hypotheses in the non-stationary case where $\phi_1 = 1.00$. We evaluate the performance of the LMC-LRT and MMC-LRT under unit-root DGPs. The results indicate that the supTS and expTS tests fail to maintain proper size control. Specifically, as sample size increases and the process more closely resembles a non-stationary one, these tests exhibit substantial over-rejection.

In contrast, the results suggest that Monte Carlo-based procedures exhibit remarkably accurate size properties in the non-stationary case. This includes the moment-based tests proposed by Dufour & Luger (2017). Under the alternative hypothesis, power improves when regimes are asymmetric—particularly when only the mean changes or when both the mean and variance change. All tests perform best with larger sample sizes and when the variance, or both the mean and variance, differ under the alternative. To our knowledge, simulations for the moment-based procedures in this non-stationary setting were not reported in Dufour & Luger (2017). We are therefore the first to provide simulation evidence on the performance of moment-based approaches for non-stationary processes.

Table 6 presents results for the previously discussed case where the regimes are asymmetric,

Table 5: Empirical performance of test with $M_0 = 1$, $m = 1$, and non-stationary process

Test	Empirical size					
	T=100	T=200	T=500			
LMC-LRT	4.5	4.9	5.7			
MMC-LRT	2.2	2.3	4.5			
LMC_{min}	4.0	3.7	5.6			
LMC_{prod}	3.8	4.7	5.6			
MMC_{min}	1.4	1.5	3.1			
MMC_{prod}	1.5	2.0	2.6			
supTS	2.2	1.8	93.4			
expTS	2.6	38.3	98.2			
	Empirical Power					
	$(p_{11}, p_{22}) = (0.9, 0.9)$				$(p_{11}, p_{22}) = (0.9, 0.5)$	
	T=100	T=200	T=500		T=100	T=200	T=500
				$\Delta\mu$			
LMC-LRT	15.5	22.8	39.9		27.0	46.4	68.4
MMC-LRT	9.2	14.1	25.2		21.0	38.9	54.3
LMC_{min}	18.4	29.2	56.2		15.8	23.5	49.9
LMC_{prod}	19.2	30.4	57.8		16.9	25.3	52.2
MMC_{min}	7.0	16.3	44.0		6.5	14.2	38.4
MMC_{prod}	9.1	17.9	48.2		7.8	17.0	43.1
				$\Delta\sigma$			
LMC-LRT	41.8	76.3	99.1		36.2	61.2	93.9
MMC-LRT	23.5	41.3	91.2		25.2	48.9	91.8
LMC_{min}	38.9	63.1	94.8		45.6	71.6	95.4
LMC_{prod}	38.4	65.4	96.6		48.0	73.0	95.6
MMC_{min}	19.5	44.1	89.1		26.0	53.4	93.3
MMC_{prod}	21.8	46.8	90.1		27.4	54.4	93.3
				$\Delta\mu$ and $\Delta\sigma$			
LMC-LRT	29.7	54.4	77.3		49.7	76.9	90.4
MMC-LRT	21.7	43.1	63.8		34.4	67.9	88.1
LMC_{min}	32.7	57.1	92.6		61.2	88.4	99.5
LMC_{prod}	36.2	61.4	93.7		63.9	90.3	99.8
MMC_{min}	18.2	41.3	85.0		41.8	80.0	99.4
MMC_{prod}	20.7	47.8	87.7		46.6	83.3	99.6

Notes: Here, we consider $H_0 : M_0 = 1$ vs. $H_1 : M_0 + m = 2$. The DGP when considering empirical size is specified as $(\phi, \mu, \sigma) = (1, 0, 1)$ whereas the DGP when considering empirical power in the third panel is specified as $(\phi, \mu_1, \mu_2, \sigma_1, \sigma_2, p_{11}, 1-p_{11}, 1-p_{22}, p_{22}) = (1, 0, 2, 1, 2, p_{11}, 1-p_{11}, 1-p_{22}, p_{22})$, where the values of p_{11} and p_{22} vary across columns as indicated in the column headers. In the first two sub-panels related to empirical power, where only the mean or the variance changes, the parameters for the constant component take the same values as those in the first regime so that there is only two regimes in the mean or the variance respectively. The nominal level is 5%. Specifically, $\phi_1 = 1.00$ for all models so that we have a non-stationary (random-walk) process. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using $1,000$ replications. MC tests use $N = 99$ simulations.

with one regime being absorbing–*i.e.*, the transition probability lies on the boundary of the parameter space. Specifically, we consider $(p_{11}, p_{22}) = (0.9, 1.0)$, where, as before, $p_{ij} = 1 - p_{ii}$ for $j \neq i$. In this setting, we find that low persistence combined with changes in the mean leads to higher power for smaller sample sizes ($T = 100$ and $T = 200$). When the sample size increases to $T = 500$, power is high in all scenarios, except in the case of high persistence and changes in the mean only.

Table 6: Empirical power of test with $M_0 = 1$, $m = 1$, and parameter on boundary

Test	$\phi = 0.10$			$\phi = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500
			$\Delta\mu$			
LMC-LRT	76.7	97.9	99.7	7.2	8.1	9.9
MMC-LRT	68.7	93.7	96.5	5.5	5.3	4.7
			$\Delta\sigma$			
LMC-LRT	30.8	56.0	91.9	27.8	52.1	93.5
MMC-LRT	24.6	50.3	86.4	23.3	48.8	82.7
			$\Delta\mu$ and $\Delta\sigma$			
LMC-LRT	49.9	83.8	99.5	19.5	41.5	90.1
MMC-LRT	40.7	81.0	96.0	11.2	34.0	84.0

Notes: Here, we consider $H_0 : M_0 = 1$ vs. $H_1 : M_0 + m = 2$. The DGP in the third panel is specified as $(\phi, \mu_1, \mu_2, \sigma_1, \sigma_2, p_{11}, 1 - p_{11}, 1 - p_{22}, p_{22}) = (\phi, 0, 2, 1, 2, 0.9, 0.1, 0.0, 1.0)$, where the values of ϕ varies across columns as indicated in the column headers. In the first two panels, where only the mean or the variance changes, the parameters for the constant component take the same values as those in the first regime so that there is only two regimes in the mean or the variance respectively. The nominal level is 5%. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using $1,000$ replications. MC tests use $N = 99$ simulations.

Table 7 reports the rejection frequencies of the LMC-LRT and MMC-LRT under the alternative hypothesis when $M_0 = 1$ and $m = 2$. That is, we consider a linear model under the null hypothesis (*i.e.*, $H_0 : M_0 = 1$) versus a Markov switching model with three regimes under the alternative (i.e., $H_1 : M_0 + m = 3$). The results show consistently high power across all cases considered, which is expected given that the alternative is further from the null.

Table 8 presents results for the case where the null hypothesis involves two regimes (*i.e.*, $H_0 : M_0 = 2$) and the alternative consists of three regimes (*i.e.*, $H_1 : M_0 + m = 3$).

Table 7: Empirical power of test when $M_0 = 1$, $m = 2$

Test	$(p_{11}, p_{22}, p_{33}) = (0.9, 0.9, 0.9)$						$(p_{11}, p_{22}, p_{33}) = (0.9, 0.5, 0.5)$					
	$\phi = 0.10$			$\phi = 0.90$			$\phi = 0.10$			$\phi = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500
						$\Delta\mu$						
LMC-LRT	84.6	98.3	100.0	59.0	86.2	99.5	90.5	99.9	100.0	69.6	95.6	100.0
MMC-LRT	80.0	93.0	95.3	51.4	77.3	92.1	88.7	97.0	99.7	58.7	91.0	96.1
						$\Delta\sigma$						
LMC-LRT	71.6	95.6	100.0	67.7	95.4	100.0	86.7	99.3	99.2	84.7	98.9	99.2
MMC-LRT	62.5	84.0	92.4	59.0	86.3	93.4	58.4	80.7	94.4	54.7	78.0	93.5
						$\Delta\mu$ and $\Delta\sigma$						
LMC-LRT	85.5	99.9	100.0	77.1	95.9	100.0	99.6	100.0	100.0	84.9	99.2	100.0
MMC-LRT	79.4	90.1	98.1	60.6	92.0	94.3	99.1	93.3	96.1	74.0	97.0	100.0

Notes: Here, we consider $H_0 : M_0 = 1$ vs. $H_1 : M_0 + m = 3$. The DGP in the third panel is specified as $(\phi, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, p_{11}, 1-(p_{11}/2), 1-(p_{11}/2), 1-(p_{22}/2), p_{22}, 1-(p_{22}/2), 1-(p_{33}/2), 1-(p_{33}/2), p_{33})$ $= (\phi, 0, 2, -2, 1, 2, 0.5, p_{11}, 1-(p_{11}/2), 1-(p_{11}/2), 1-(p_{22}/2), p_{22}, 1-(p_{22}/2), 1-(p_{33}/2), 1-(p_{33}/2), p_{33})$, where the values of ϕ, p_{11}, p_{22}, and p_{33} vary across columns as indicated in the column headers. In the first two panels, where only the mean or the variance changes, the parameters for the constant component take the same values as those in the first regime so that there is only three regimes in the mean or the variance respectively. The nominal level is 5%. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using $1,000$ replications. MC tests use $N = 99$ simulations.

Generally, detecting additional regimes when the null already involves multiple regimes is more challenging. Nonetheless, the size results indicate that both proposed test procedures maintain appropriate size control. For the DGPs considered here, power appears to depend more heavily on the presence of changes in the mean. In particular, scenarios involving mean changes yield substantially higher power than those involving changes in variance only. This finding contrasts with earlier results, which suggested that variance changes were an important source of power for all test procedures. However, when multiple regimes are present, a high-variance regime may obscure lower-variance regimes and associated shifts in the mean. As a result, changes in location (mean) may become more informative for distinguishing regimes. These findings highlight a potentially important avenue for future research.

Table 9 presents results for an alternative set of DGPs, still within the context of testing a null hypothesis of two regimes (*i.e.*, $H_0 : M_0 = 2$) against an alternative hypothesis of a

Table 8: Empirical size and power of test when $M_0 = 2$ and $m = 1$

Test	Empirical size											
	$(p_{11}, p_{22}) = (0.90, 0.90)$						$(p_{11}, p_{22}) = (0.90, 0.50)$					
	$\phi = 0.10$			$\phi = 0.90$			$\phi = 0.10$			$\phi = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500
	$\Delta\mu$											
LMC-LRT	5.9	6.6	5.7	4.3	5.1	5.0	4.8	4.4	5.1	4.9	5.5	4.8
MMC-LRT	2.3	2.2	3.4	1.7	3.1	3.3	2.6	2.3	3.1	2.1	2.1	2.9
	$\Delta\sigma$											
LMC-LRT	4.6	4.5	5.5	5.3	4.9	4.3	4.5	5.8	5.7	4.2	4.6	5.9
MMC-LRT	2.2	2.1	3.1	2.0	3.2	2.8	2.6	2.9	3.0	2.6	2.4	3.8
	$\Delta\mu$ and $\Delta\sigma$											
LMC-LRT	5.8	4.4	5.1	4.2	5.2	5.3	4.2	4.8	5.4	4.6	5.6	5.2
MMC-LRT	2.5	2.8	4.1	2.1	2.3	3.8	2.9	3.4	4.0	2.4	2.5	3.4
	Empirical power											
	$(p_{11}, p_{22}, p_{33}) = (0.90, 0.90, 0.90)$						$(p_{11}, p_{22}, p_{33}) = (0.90, 0.50, 0.50)$					
	$\phi = 0.10$			$\phi = 0.90$			$\phi = 0.10$			$\phi = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500
	$\Delta\mu$											
LMC-LRT	39.9	84.2	94.7	6.7	7.2	8.6	12.6	27.4	52.2	11.9	11.8	12.3
MMC-LRT	34.1	81.1	90.9	4.9	5.6	6.1	8.1	20.3	44.6	8.3	7.8	7.6
	$\Delta\sigma$											
LMC-LRT	8.5	24.0	57.6	10.0	22.2	56.2	6.4	9.0	20.7	5.8	8.9	21.1
MMC-LRT	6.5	19.2	52.7	6.2	17.7	49.9	4.1	6.9	18.3	4.1	5.6	14.2
	$\Delta\mu$ and $\Delta\sigma$											
LMC-LRT	40.4	88.4	100.0	14.2	26.8	56.8	15.4	32.4	79.2	9.4	16.2	30.0
MMC-LRT	35.2	74.3	93.7	11.6	21.4	50.2	11.1	28.6	74.4	7.0	11.5	24.5

Notes: Here, we consider $H_0 : M_0 = 2$ vs. $H_1 : M_0 + m = 3$. The DGP when considering empirical size in the third sub-panel is specified as $(\phi, \mu_1, \mu_2, \sigma_1, \sigma_2, p_{11}, 1-p11, 1-p_{22}, p_{22}) = (\phi, 0, 2, 1, 2, p_{11}, 1-p11, 1-p_{22}, p_{22})$, where the values of ϕ, p_{11}, and p_{22} vary across columns as indicated in the column headers. In the first two sub-panels related to empirical size, where only the mean or the variance changes, the parameters for the constant component take the same values as those in the first regime so that there is only two regimes in the mean or the variance respectively. The DGP when considering empirical power in the third sub-panel is specified as $(\phi, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, p_{11}, 1-(p_{11}/2), 1-(p_{11}/2), 1-(p_{22}/2), p_{22}, 1-(p_{22}/2), 1-(p_{33}/2), 1-(p_{33}/2), p_{33}) = (\phi, 0, 2, -2, 1, 2, 0.5, p_{11}, 1-(p_{11}/2), 1-(p_{11}/2), 1-(p_{22}/2), p_{22}, 1-(p_{22}/2), 1-(p_{33}/2), 1-(p_{33}/2), p_{33})$, where the values of ϕ, p_{11}, p_{22}, and p_{33} vary across columns as indicated in the column headers. In the first two sub-panels related to empirical power, where only the mean or the variance changes, the parameters for the constant component take the same values as those in the first regime so that there is only three regimes in the mean or the variance respectively. The nominal level is 5%. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using 1000 replications. MC tests use $N = 99$ simulations.

Markov switching model with three regimes (*i.e.*, $H_1 : M_0 + m = 3$). For this comparison, we include two classes of DGPs considered in Kasahara & Shimotsu (2018), and we also report the Boot-LRT results from that paper, except for $T = 100$, as those were not provided by the authors.

Table 9: Empirical size of test when $M_0 = 2$ and $m = 1$ for alternative DGPs

Test	$(p_{11}, p_{22}) = (0.5, 0.5)$			$(p_{11}, p_{22}) = (0.7, 0.7)$		
	T=100	T=200	T=500	T=100	T=200	T=500
LMC-LRT	6.80	6.30	4.60	6.00	6.00	4.80
MMC-LRT	3.80	3.70	3.30	3.10	3.60	2.70
Boot-LRT	-	7.16	4.43	-	6.07	4.20

Notes: Here, we consider $H_0 : M_0 = 2$ vs. $H_1 : M_0 + m = 3$. The DGP is specified as $(\phi, \mu_1, \mu_2, \sigma, p_{11}, 1 - p_{11}, 1 - p_{12}, p_{22}) = (0.5, -1, 1, 1, p_{11}, 1 - p_{11}, 1 - p_{12}, p_{22})$ where the values of p_{11} and p_{22} vary across columns as indicated in the column headers. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using 1000 replications. MC tests use $N = 99$ simulations. *Boot-LRT* results are taken from Kasahara & Shimotsu (2018) and incorporate the restrictions desrcibed therein.

As previously discussed, the parametric bootstrap and LMC-LRT procedures share many similarities. However, a key distinction is that the LMC-LRT does not require enforcing the additional assumptions needed to ensure the asymptotic validity of the bootstrap when estimating the null distribution. While larger sample sizes should improve the approximation quality of both procedures, the LMC-LRT does not rely on the existence of an asymptotic distribution, allowing for fewer simulations to achieve valid inference. In contrast, Kasahara & Shimotsu (2018) impose such constraints and an additional one on the variance parameters during model estimation and use $N = 199$ bootstrap simulations, whereas we use $N = 99$ and impose no such constraints. These differences help explain the discrepancies observed between the LMC-LRT and the parametric bootstrap test results.

Nonetheless, the results suggest that both procedures exhibit broadly similar patterns across these DGPs. Specifically, Table 9 shows that both the LMC-LRT and the parametric bootstrap tests exhibit some over-rejection for smaller sample sizes ($T = 100$ and $T = 200$), particularly in the case of the bootstrap test. As expected, the rejection fre-

quencies approach the nominal level when $T = 500$. Meanwhile, the MMC-LRT performs as expected, maintaining rejection frequencies at or below 5% even for small samples. This highlights the strength of the MMC-LRT as a valid test procedure in both finite samples and asymptotic settings.

Next, for the multivariate setting, we begin by assessing the size and power properties of the LMC-LRT and MMC-LRT when testing the null hypothesis of a single regime ($M = 1$) against the alternative of two regimes ($M = 2$) in a bivariate VAR(1) model. The experiments explore mean shifts ($\Delta\mu$), variance shifts ($\Delta\sigma$), and joint shifts. Results are summarized in Table 10.

The tests exhibit good size control and increasing power with sample size. The MMC-LRT performs particularly well in small samples and in the presence of strong persistence or heteroskedasticity. These results align closely with the univariate findings and confirm that the proposed testing procedures extend effectively to multivariate settings. To the best of our knowledge, this is the first simulation-based evidence for regime testing procedures in multivariate Markov-switching models.

A.3 U.S. output growth

Many procedures for testing the number of regimes in Markov switching models have used U.S. GNP growth data, as it was one of the original applications in Hamilton (1989). Notable studies employing U.S. GNP data for regime testing include Hansen (1992), Carrasco et al. (2014), and Dufour & Luger (2017). Hansen (1992) examines the original quarterly sample from 1951:II to 1984:IV, as used in Hamilton (1989), with $p = 4$ lags and a specification where only the mean changes across regimes. In this case, the proposed test fails to reject the null hypothesis of a linear model (*i.e.*, $M = 1$). Similarly, Carrasco et al. (2014) and Dufour & Luger (2017) also use this sample and reach the same conclusion.

Table 10: Empirical performance of test for bivariate VAR model

Test	Empirical size					
	$\max(\lambda) = 0.10$			$\max(\lambda) = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500
LMC-LRT	4.2	4.2	6.2	3.6	4.6	5.6
MMC-LRT	3.8	3.5	4.4	1.2	3.1	3.3

Test	Empirical Power											
	$(p_{11}, p_{22}) = (0.90, 0.90)$						$(p_{11}, p_{22}) = (0.90, 0.50)$					
	$\max(\lambda) = 0.10$			$\max(\lambda) = 0.90$			$\max(\lambda) = 0.10$			$\max(\lambda) = 0.90$		
	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500	T=100	T=200	T=500
	$\Delta\mu$											
LMC-LRT	25.0	63.6	95.9	8.2	11.8	28.2	12.2	26.0	75.1	9.0	24.4	54.4
MMC-LRT	19.4	62.2	91.2	6.6	11.0	27.6	9.0	22.6	69.4	9.7	18.9	50.8
	$\Delta\sigma$											
LMC-LRT	46.4	90.6	100.0	53.0	91.8	100.0	34.4	68.6	100.0	38.2	73.6	96.7
MMC-LRT	45.6	82.4	100.0	51.0	85.2	99.8	31.8	64.6	100.0	34.4	70.1	92.8
	$\Delta\mu$ and $\Delta\sigma$											
LMC-LRT	87.0	100.0	100.0	60.4	93.4	99.7	73.6	98.2	100.0	62.0	93.6	100.0
MMC-LRT	82.0	100.0	100.0	53.6	86.1	96.1	63.4	93.9	99.8	54.7	81.0	96.0

Notes: Here, we consider $H_0 : M_0 = 1$ vs. $H_1 : M_0 + m = 2$. We set $\Phi = [0.08, 0.02; 0.02, 0.08]$ or $\Phi = [0.80, 0.10; 0.10, 0.80]$ so that $\max(\lambda)$, the largest eigenvalue of Φ, is either 0.10 or 0.90 respectively. The DGP when considering empirical size is specified as $(vec(\Phi), \mu_1, \mu_2, \sigma_{11}^2, \sigma_{12}, \sigma_{22}^2) = (vec(\Phi), 0.00, 1.00, 0.75, 1.00)$, where the values of Φ vary across columns as indicated in the column header. The DGP when considering empirical power in the third sub-panel is specified as $(vec(\Phi), \mu_{1,1}, \mu_{2,1}, \mu_{1,2}, \mu_{2,2}, \sigma_{11,1}^2, \sigma_{12,1}, \sigma_{22,1}^2, \sigma_{11,2}^2, \sigma_{12,2}, \sigma_{22,2}^2, p_{11}, 1 - p_{11}, 1 - p_{22}, p_{22}) = (vec(\Phi), 0.00, 0.00, 2.00, 2.00, 1.00, 0.75, 1.00, 4.00, 3.00, 4.00, p_{11}, 1 - p_{11}, 1 - p_{22}, p_{22})$, where the values of Φ, p_{11} and p_{22} vary across columns as indicated in the column headers. In the first two sub-panels related to empirical power, where only the mean or the variance changes, the parameters for the constant component take the same values as those in the first regime so that there is only two regimes in the mean or the variance respectively. The nominal level is 5%. LMC-LRT and MMC-LRT are the Local Monte Carlo and Maximized Monte Carlo likelihood-ratio tests proposed here, respectively. Rejection frequencies are obtained using $1,000$ replications. MC tests use $N = 99$ simulations.

These latter two studies also consider an extended sample from 1951:II to 2010:IV, which includes the Great Recession. They continue to use four lags ($p = 4$), but now also evaluate an alternative where both the mean and variance change across regimes, as suggested by Kim & Nelson (1999). Allowing for changes in variance is sensible for two reasons. First, the extended sample includes the structural decline in macroeconomic volatility during the mid-1980s, known as the Great Moderation. Second, since the objective is to capture recessionary episodes, it is reasonable to assume that volatility increases during such periods. For this extended sample, both Carrasco et al. (2014) and Dufour & Luger (2017) reject the null hypothesis of a linear model in favor of a Markov switching model with $M = 2$ regimes, but only when the variance is allowed to change. As discussed in Qu & Zhuo (2021), when using GDP data, the inclusion of the Great Recession appears to be crucial for the supTS test of Carrasco et al. (2014) to reject linearity. However, when only the mean is allowed to change, the supTS and expTS tests continue to fail to reject the null. In contrast, Qu & Zhuo (2021), using GDP rather than GNP data, find stronger evidence in favor of a two-regime model even when only the mean is allowed to change.

To complement the existing literature, we consider the same two samples of U.S. GNP data as in prior studies, along with an extended sample spanning 1951:II to 2024:II. Results from the LMC-LRT and MMC-LRT applied to these three U.S. GNP growth rate samples are presented in Table 11. For the first two samples, our findings are broadly consistent with earlier tests. However, unlike Carrasco et al. (2014), we find evidence in favor of a two-regime model ($M = 2$) even when only the mean is allowed to change in the second sample which includes the Great Recession. This result is more in line with Qu & Zhuo (2021), who find similar evidence using U.S. GDP data.

We extend the analysis by formally testing the null hypothesis of $M = 2$ regimes against the alternative of $M = 3$. To our knowledge, this is the first such test applied to U.S. GNP data

Table 11: Results for U.S. GNP growth series hypothesis tests

Series	$H_0 : M = 1$ vs. $H_1 : M = 2$		$H_0 : M = 2$ vs. $H_1 : M = 3$		$H_0 : M = 3$ vs. $H_1 : M = 4$	
	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT
			$\Delta\mu$			
GNP 1951:II-1984:IV	0.35	0.93	-	-	-	-
GNP 1951:II-2010:IV	0.03	0.05	0.06	0.23	-	-
GNP 1951:II-2024:II	0.01	0.01	0.01	0.01	0.52	1.00
			$\Delta\mu$ and $\Delta\sigma$			
GNP 1951:II-1984:IV	0.38	0.85	-	-	-	-
GNP 1951:II-2010:IV	0.01	0.01	0.58	1.00	-	-
GNP 1951:II-2024:II	0.01	0.01	0.02	0.04	0.70	1.00

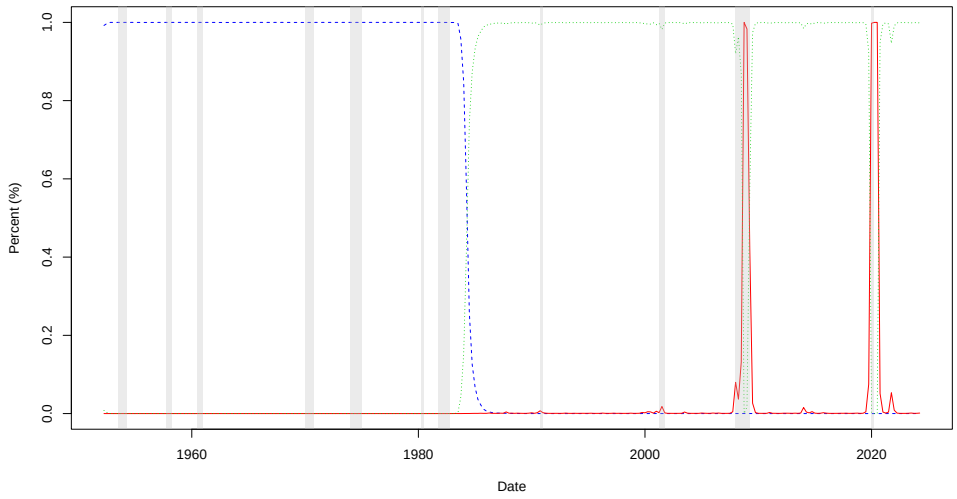
Notes: The GNP 1951:II-1984:IV series ($T = 135$) is the same as the one used in Hamilton (1989), Hansen (1992), and Carrasco et al. (2014). The GNP 1951:II-2010:IV series ($T = 239$) is the same as the one used in Carrasco et al. (2014) and Dufour & Luger (2017). The GNP 1951:II-2024:II series ($T = 293$) is the GNP series from the St. Louis Fed (FRED) website. All MC test results are obtained using $N = 99$. The MMC-LRT procedure uses a particle swarm optimization algorithm. Models for GNP use $p = 4$ lags as in Hamilton (1989) while models for GDP use $p = 1$ lags as in Qu & Zhuo (2021).

for this sample. In both the mean-only and mean-and-variance-switching specifications, we fail to reject the $M = 2$ null, confirming that two regimes are sufficient. However, when we turn to the third, longer sample, we reject the null of $M = 2$ in favor of a Markov switching model with $M = 3$ regimes. We further test this three-regime model against a four-regime alternative and fail to reject the null, thereby supporting $M = 3$ as the preferred specification for the extended sample.

Figure 1 shows the smoothed regime probabilities for the $M = 3$ model in which both the mean and variance change. Smoothed probabilities for additional models are shown in Figures 4–7. Parameter estimates for this model and others which include regime-dependent variance are provided in Table 15. In the $M = 3$ model, two regimes are expansionary with positive means, though one is characterized by substantially lower volatility ($\mu_1 = 1.87$, $\mu_2 = 1.31$, $\sigma_1 = 1.09$, $\sigma_2 = 0.49$). This reduction in volatility is consistent with the Great Moderation and is reflected in the smoothed regime probabilities, which shift in the mid-1980s. The third regime represents a deep recessionary state which captures both the Great Recession and the COVID-19 recession. Similar to the findings in Gadea et al. (2018) and

Gadea et al. (2019), we observe that the low-volatility regime re-emerges after the Great Recession—and, in our case, again following the COVID recession.

Figure 1: Smoothed Probabilities of US GNP regimes when $\Delta\mu$ and $\Delta\sigma$, $M = 3$



Notes: The sample is from 1951:III to 2024:II. The shaded areas correspond to the NBER recessions.

Given that much of the recent literature now relies on U.S. GDP data, we now shift our focus to this series. Both Qu & Zhuo (2021) and Kasahara & Shimotsu (2018), for example, use U.S. GDP instead of GNP when testing the number of regimes in Markov switching models. For this application, we consider the extended sample from 1951:II to 2024:II. This longer sample is particularly interesting because it includes the COVID-19 period, which poses challenges for macroeconomic modeling due to its stark departure from historical patterns.

Several approaches have been proposed to address the impact of the COVID-19 period. One strategy is to treat it as a known structural break. A benefit of this approach is that, by incorporating explanatory variables to account for the shock, one may justify a simpler model specification—potentially requiring fewer regimes to capture the non-linearities in the series. More generally, a key advantage of the testing procedure proposed here is its flexibility: users can include control variables to account for known structural features and then test whether the number of regimes can be reduced conditionally, given those controls.

To assess the robustness of our procedure, we evaluate the number of regimes in U.S. GDP growth under different treatments of structural breaks. Specifically, we include a dummy variable equal to 1 from 2020:I to 2021:IV and 0 otherwise, to control for the COVID period. We also include a second dummy equal to 1 from 1951:II to 1983:IV and 0 elsewhere, to control for the Great Moderation.[1] These simple mean-shift controls offer a baseline way to account for known structural breaks. However, they are likely insufficient in capturing the full dynamics of such episodes, which are typically driven by changes in both the conditional mean and conditional variance. As such, these specifications are intended as a starting point, rather than a comprehensive treatment.

Accordingly, we examine four models: Model 1 includes no dummy variables; Model 2 includes only the Great Moderation dummy; Model 3 includes only the COVID dummy; and Model 4 includes both dummies. For each specification, we test the number of regimes under two setups–one where only the mean changes and another where both the mean and variance change across regimes– leading to eight models in total. The results of these tests are reported in Table 12.

As with the GNP growth data for the same sample period, we find evidence supporting a model with $M = 3$ regimes for U.S. GDP growth. Unlike the GNP models, we use one lag ($p = 1$) in this case, consistent with the specification in Qu & Zhuo (2021). Since our sample differs slightly from theirs, we first verified that a lag order of one remains appropriate and confirmed that this is indeed the case.

To assess which of the eight candidate models is preferred for this sample, we apply a likelihood-ratio test (LRT) to evaluate the significance of the dummy variables. In this setting, the conventional regularity conditions are satisfied, allowing us to rely on standard

[1] We first estimate a Markov switching model without this dummy and find strong evidence that one of the regimes corresponds to the Great Moderation. We use the smoothed probabilities from this model to date the period.

Table 12: Results for U.S. GDP growth series hypothesis tests with known breaks

Series	$H_0 : M = 1$ vs. $H_1 : M = 2$		$H_0 : M = 2$ vs. $H_1 : M = 3$		$H_0 : M = 3$ vs. $H_1 : M = 4$	
	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT
			$\Delta\mu$			
Model 1	0.01	0.01	0.01	0.01	0.76	1.00
Model 2	0.01	0.01	0.01	0.01	0.76	1.00
Model 3	0.01	0.01	0.01	0.01	0.94	1.00
Model 4	0.01	0.01	0.01	0.01	0.59	1.00
			$\Delta\mu$ and $\Delta\sigma$			
Model 1	0.01	0.01	0.01	0.01	0.44	1.00
Model 2	0.01	0.01	0.01	0.01	0.35	1.00
Model 3	0.01	0.01	0.01	0.01	0.27	1.00
Model 4	0.01	0.01	0.01	0.01	0.24	1.00

Notes: The GDP 1951:II-2024:II series ($T = 293$) is the GDPC1 series from the St. Louis Fed (FRED) website. Model 1: no fixed exogenous regressors, Model 2: includes dummy variable treating Great Moderation as known structural break, Model 3: includes dummy variable treating COVID period as known multiple structural breaks, and Model 4: includes dummy variables treating Great Moderation and COVID period as known multiple structural breaks. Specifically, the dummy variable for the Great Moderation takes values of 1 for the period 1951:II to 1983:IV, and 0 elsewhere. Similarly, dummy variable for the COVID period takes values of 1 for the period 2020:I to 2021:IV, and 0 elsewhere. the All MC test results are obtained using $N = 99$. The MMC-LRT procedure uses a particle swarm optimization algorithm. Models GDP use $p = 1$ lags as in Qu & Zhuo (2021).

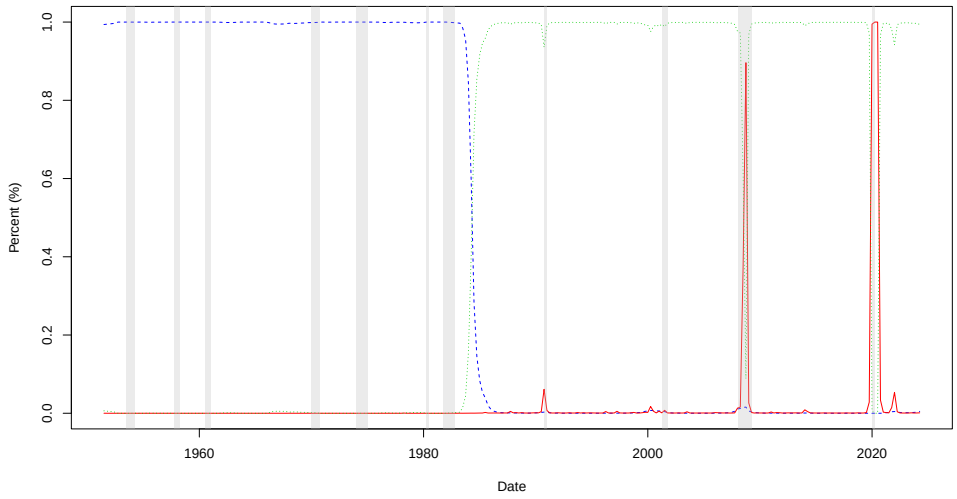
LRT inference. Table 13 reports the estimates and log-likelihood values for each model. Across all specifications–whether only the mean changes or both the mean and variance–the inclusion of dummy variables does not significantly improve the log-likelihood, resulting in small LRT statistics and no statistical significance.

In addition, in all cases, the model which allows for changes in both the mean and variance is preferred to the corresponding model where only the mean changes. Taken together, these findings indicate that a model with $M = 3$ regimes, regime-dependent means and variances, and no dummy variables provides a sufficiently good fit to the data.

The finding of three distinct regimes in U.S. output growth is consistent with earlier empirical evidence. For example, Boldin (1996) documents a three-regime characterization of U.S. business cycle dynamics, while Sims & Zha (2006) find support for multiple regimes in Markov-switching VARs, often associated with changes in volatility. More recent contributions, including Hwu et al. (2021) and Kim & Kang (2022), also emphasize that macroe-

conomic data can support three or more regimes when regime transitions are allowed to be more flexible. Although these studies focus on different sample periods and modeling assumptions, they reinforce the empirical plausibility of the multi-regime structures identified by our formal testing approach.

Figure 2: Smoothed probabilities of US GDP regimes when $\Delta\mu$ and $\Delta\sigma$ and $M = 3$



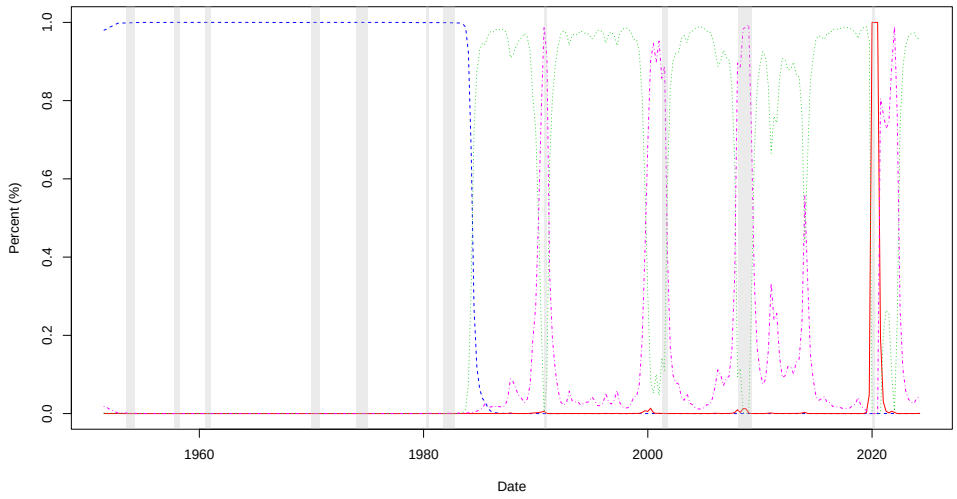
Notes: The sample is from 1951:III to 2024:II. The shaded areas correspond to the NBER recessions.

The smoothed regime probabilities for this model, shown in Figure 2, closely resemble those obtained for the GNP case, though the parameter estimates differ slightly (*i.e.*, $\mu_1 = 0.79$, $\mu_2 = 0.72$, $\mu_3 = -0.50$, $\sigma_1 = 1.06$, $\sigma_2 = 0.45$, and $\sigma_3 = 6.5$). Notably, the inclusion of dummy variables did not alter the outcome of the hypothesis test for the number of regimes. This is likely because treating these episodes as known structural breaks in the conditional mean is insufficient to capture their full effects.

In contrast, Markov switching models which allow for regime-dependent variances, such as those considered in this paper, are better equipped to account for such changes, which are central features of the Great Moderation and likely the COVID-19 period. Incorporating more sophisticated specifications–such as heteroskedastic error structures (*e.g.*, GARCH) or structural breaks in the variance–may further improve model performance and can be considered within our regime testing framework. Regardless of the chosen specification,

the testing procedures proposed here offer a useful tool for assessing whether a given model adequately captures such features, or whether additional regimes are required.

Figure 3: Smoothed probabilities of US GDP regimes for $\Delta\mu$ and $\Delta\sigma$, $M = 4$



Notes: The sample is from 1951:III to 2024:II. The shaded areas correspond to the NBER recessions.

It is also worth noting that, as shown in Figure 3, a model with $M = 4$ regimes can capture relatively milder recessions as distinct episodes. In this specification, the post-COVID recovery period is also identified as a separate regime. However, the log-likelihood value of the four-regime model, reported in Table 13, is very close to that of the three-regime model, which explains why the test fails to reject the null hypothesis of $M = 3$ regimes. Although this difference is not statistically significant, the $M = 4$ specification may still be of interest if the primary goal is to more finely distinguish phases of the business cycle, such as separating deep from shallow recessions or identifying recovery periods explicitly.

In summary, the univariate empirical application highlights the effectiveness of the proposed LMC-LRT and MMC-LRT procedures in identifying regime structure in U.S. output growth data. For both GNP and GDP, we find consistent evidence supporting a three-regime model when both the mean and variance are allowed to change–especially in longer samples which include the Great Recession and COVID-19 period. These results provide formal statistical support for three-regime specifications which have been suggested in ear-

lier empirical studies. Importantly, our results appear to be robust to the inclusion of control variables for known structural breaks in the conditional mean. Overall, this underscore the flexibility and reliability of our proposed tests in practical settings and motivate their use in more complex multivariate applications, such as testing for synchronization in international business cycles.

Table 13: Comparison of models with dummy variables for known structural breaks

	μ_1	μ_2	μ_3	ϕ_1	GMd	CVd	σ_1	σ_2	σ_3	LogLike	AIC	BIC
						$\Delta\mu$						
Model 1	7.473	0.748	-8.220	0.329	-	-	0.819	-	-	-362.771	753.543	805.017
Model 2	7.473	0.748	-8.220	0.323	0.118	-	0.817	-	-	-362.032	754.063	809.214
Model 3	7.473	0.748	-8.220	0.329	-	0.084	0.819	-	-	-362.731	755.461	810.612
Model 4	7.473	0.748	-8.220	0.323	0.125	0.141	0.817	-	-	-361.919	755.838	814.666
						$\Delta\mu$ and $\Delta\sigma$						
Model 1	0.794	0.718	-0.459	0.262	-	-	1.07	0.449	6.499	-337.072	706.145	764.973
Model 2	0.795	0.717	-0.463	0.261	0.027	-	1.07	0.450	6.502	-337.020	708.039	770.544
Model 3	0.794	0.717	-0.442	0.260	-	0.185	1.07	0.450	6.437	-337.016	708.033	770.538
Model 4	0.800	0.717	-0.447	0.260	0.022	0.158	1.07	0.451	6.449	-336.984	709.967	776.149

Notes: The GDP 1951:II-2024:II series ($T = 293$) is the GDPC1 series from the St. Louis Fed (FRED) website. Model 1: no fixed exogenous regressors, Model 2: includes dummy variable treating Great Moderation as known structural break and is labeled *GMd*, Model 3: includes dummy variable treating COVID period as known multiple structural breaks and is labeled *CVd*, and Model 4: includes dummy variables treating Great Moderation and COVID period as known multiple structural breaks. Specifically, the dummy variable for the Great Moderation takes values of 1 for the period 1951:II to 1983:IV, and 0 elsewhere. Similarly, dummy variable for the COVID period takes values of 1 for the period 2020:I to 2021:IV, and 0 elsewhere. All MC test results are obtained using $N = 99$. The MMC-LRT procedure uses a particle swarm optimization algorithm. Models GDP use $p = 1$ lags as in Qu & Zhuo (2021).

Table 14: Estimates models for US GNP series

	μ_1	μ_2	μ_3	ϕ_1	ϕ_2	ϕ_3	ϕ_4	σ_1	σ_2	σ_3	p_{11}	p_{12}	p_{13}	p_{21}	p_{22}	p_{23}	p_{31}	p_{32}	p_{33}	LogLike
M=1	1.51	-	-	0.14	0.18	0.03	0.02	1.19	-	-	-	-	-	-	-	-	-	-	-	-458.38
M=2	1.56	1.21	-	0.31	0.24	0.01	0.00	0.62	3.03	-	0.96	0.04	-	0.32	0.68	-	-	-	-	-363.79
M=3	1.87	1.31	-0.77	0.23	0.23	0.06	0.00	1.09	0.49	5.74	0.99	0.01	0.00	0.00	0.99	0.01	0.00	0.37	0.63	-341.57

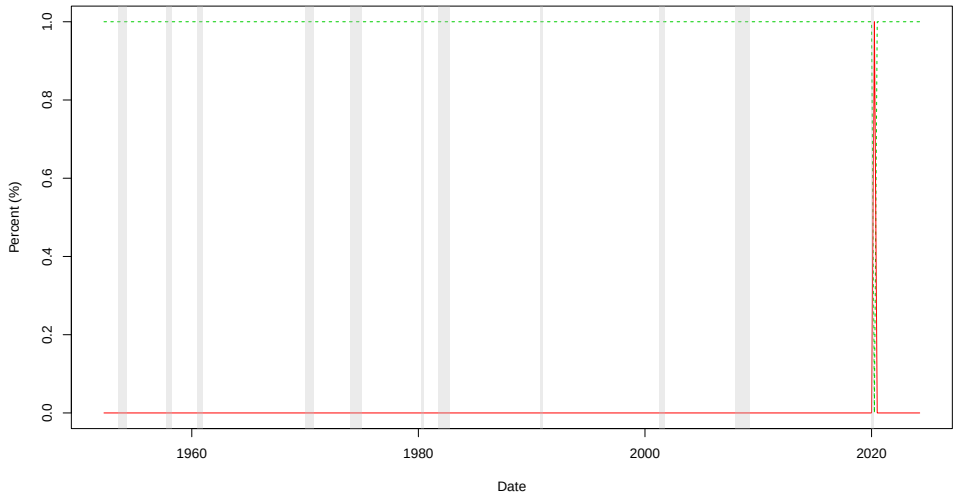
Notes: The GNP 1951:II-2024:II series ($T = 293$) is the GNP series from the St. Louis Fed (FRED) website. The models use $p = 4$ lags as in Hamilton (1989), Hansen (1992), Carrasco et al. (2014), and Dufour & Luger (2017).

Table 15: Estimates of Preferred Models for US GDP series

	μ_1	μ_2	μ_3	ϕ_1	σ_1	σ_2	σ_3	p_{11}	p_{12}	p_{13}	p_{21}	p_{22}	p_{23}	p_{31}	p_{32}	p_{33}	LogLike
M=1	0.74	-	-	0.10	1.09	-	-	-	-	-	-	-	-	-	-	-	-437.54
M=2	0.80	0.11	-	0.30	0.68	3.00	-	0.96	0.04	-	0.47	0.53	-	-	-	-	-368.08
M=3	0.79	0.72	-0.46	0.26	1.06	0.45	6.50	0.97	0.03	0.00	0.01	0.98	0.01	0.32	0.00	0.68	-337.07

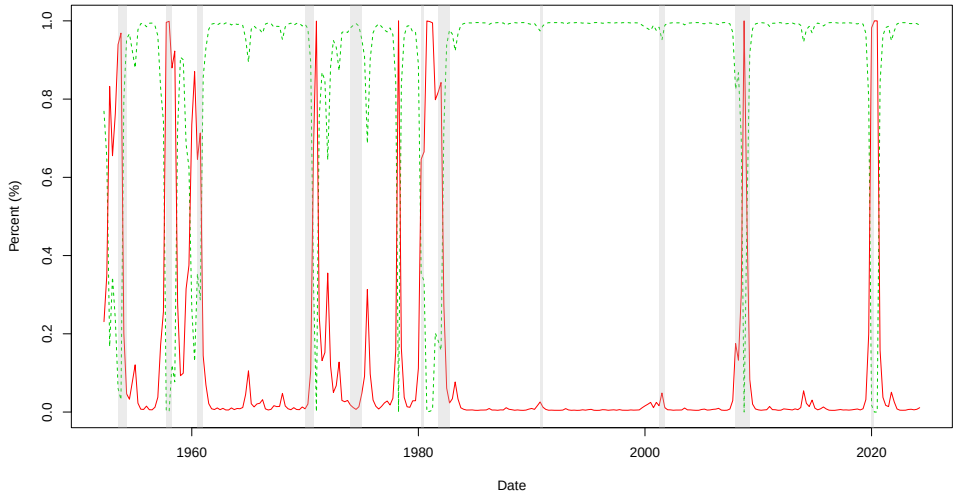
Notes: The GDP 1951:II-2024:II series ($T = 293$) is the GDPC1 series from the St. Louis Fed (FRED) website. The models use $p = 1$ lags as in Qu & Zhuo (2021).

Figure 4: Smoothed Probabilities of US GNP regimes when $\Delta\mu$ only, $M = 2$



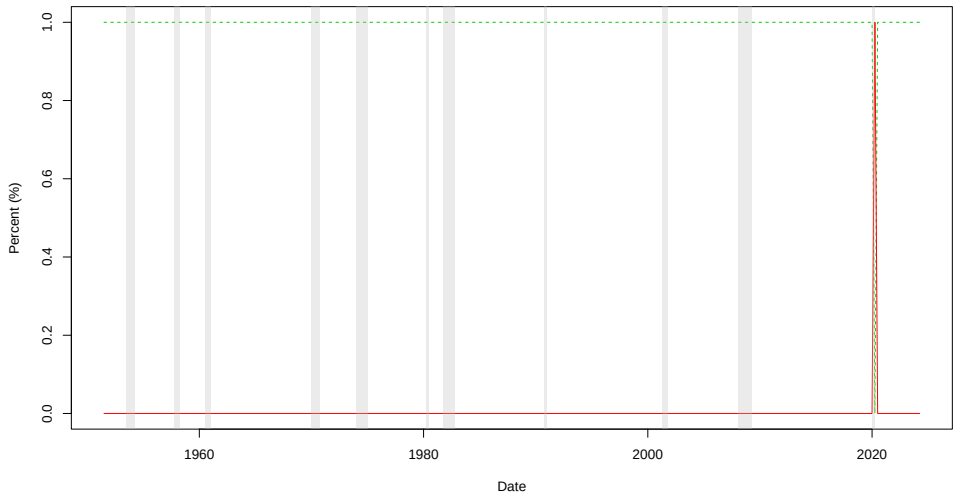
Notes: The sample is from 1951:III to 2024:II. The shaded areas correspond to the NBER recessions.

Figure 5: Smoothed probabilities of US GNP regimes when $\Delta\mu$ and $\Delta\sigma$, $M = 2$



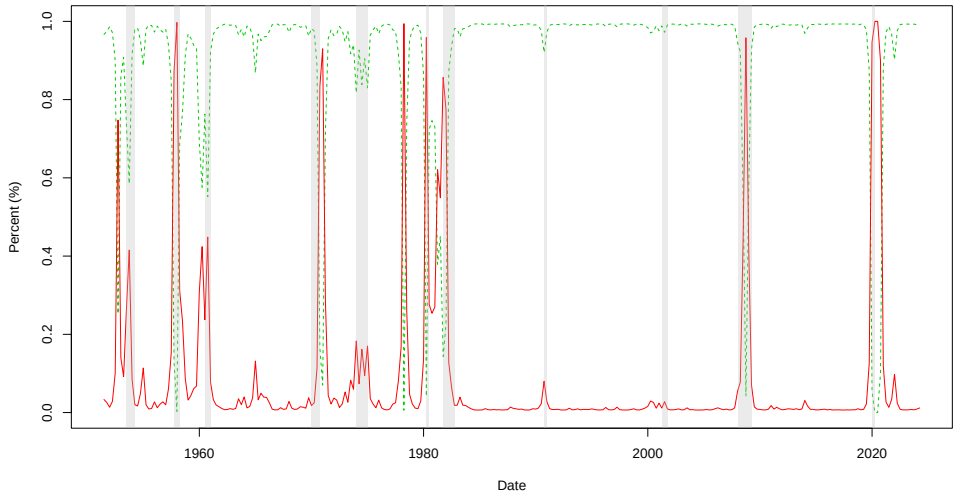
Notes: The sample is from 1951:III to 2024:II. The shaded areas correspond to the NBER recessions.

Figure 6: Smoothed probabilities of US GDP regimes for when $\Delta\mu$ only, $M = 2$



Notes: The sample is from 1951:III to 2024:II. The shaded areas correspond to the NBER recessions.

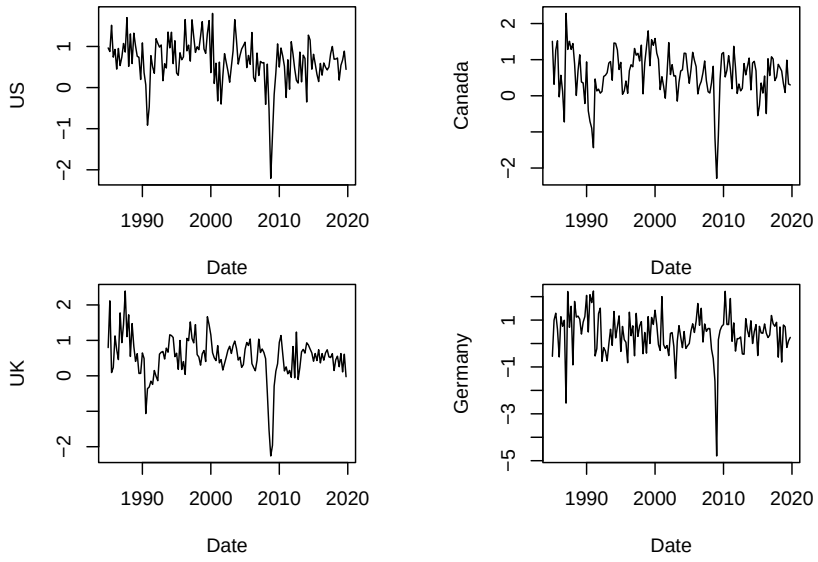
Figure 7: Smoothed probabilities of US GDP Regimes when $\Delta\mu$ and $\Delta\sigma$, $M = 2$



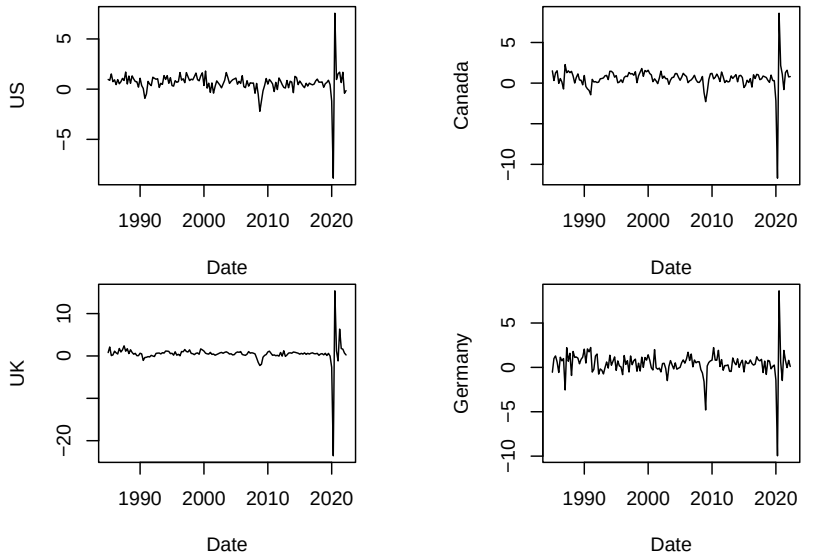
Notes: The sample is from 1951:III to 2024:II. The shaded areas correspond to the NBER recessions.

A.4 Business cycle synchronization: graphs

Figure 8: Real GDP for four countries starting in 1985:I to 2019:IV (top) and 2022:IV (bottom)

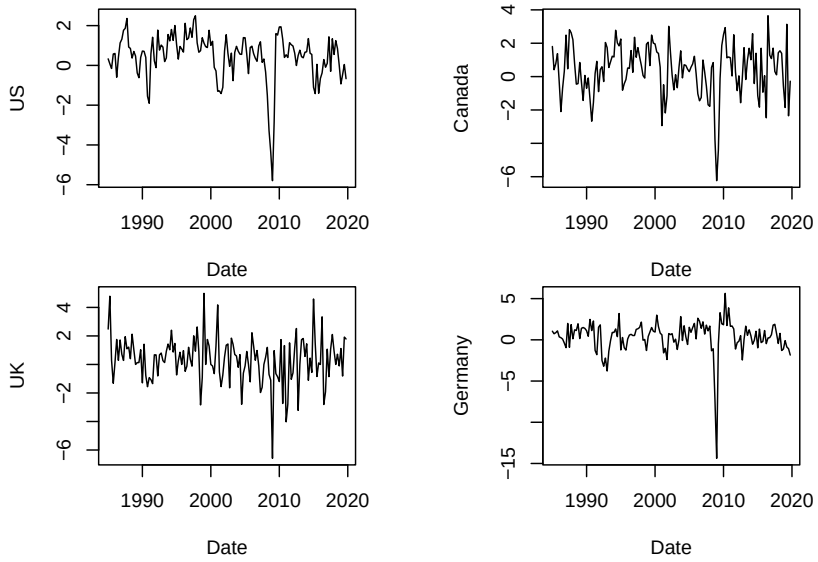


(a) Sample ending in 2019Q4

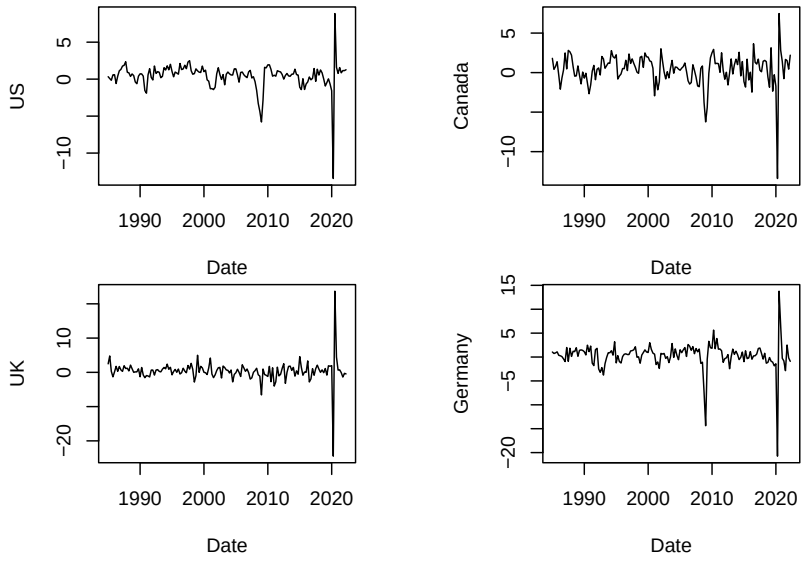


(b) Sample ending in 2022Q2

Figure 9: Industrial production for four countries starting in 1985:I to 2019:IV (top) and 2022:IV (bottom)



(a) Sample ending in 2019Q4



(b) Sample ending in 2022Q2

Table 16: Results for synchronization of business cycle hypothesis tests using IP series

Series	$H_0 : M = 1$ vs. $H_1 : M = 2$		$H_0 : M = 2$ vs. $H_1 : M = 3$		$H_0 : M = 2$ vs. $H_1 : M = 4$	
	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT	LMC-LRT	MMC-LRT
1985:I - 2019:IV ($T = 140$)						
US-CA	0.01	0.01	0.19	0.73	0.23	0.65
US-UK	0.01	0.01	0.18	0.61	0.21	0.68
US-GR	0.01	0.01	0.58	1.00	0.76	1.00
1985:I - 2022:IV ($T = 155$)						
US-CA	0.01	0.01	0.05	0.05	0.03	0.04
US-UK	0.01	0.01	0.18	0.48	0.12	0.37
US-GR	0.01	0.01	0.19	0.51	0.14	0.44

Notes: This table includes results when $\Delta\mu$ and $\Delta\sigma$ as it is a statistically preferred model over a model where only $\Delta\mu$. The IP series are OECD Main Economic Indicator Releases obtained from the St. Louis Fed (FRED) website. All MC test results are obtained using $N = 99$. The MMC-LRT procedure uses a particle swarm optimization algorithm.

Appendix: References

Carrasco, M., Hu, L. & Ploberger, W. (2014), 'Optimal test for Markov switching parameters', *Econometrica* **82**(2), 765–784.

Dufour, J.-M. & Luger, R. (2017), 'Identification-robust moment-based tests for Markov switching in autoregressive models', *Econometric Reviews* **36**(6-9), 713–727.

Gadea, M. D., Gómez-Loscos, A. & Pérez-Quirós, G. (2018), 'Great moderation and great recession: From plain sailing to stormy seas?', *International Economic Review* **59**(4), 2297–2321.

Gadea, M. D., Gómez-Loscos, A. & Pérez-Quirós, G. (2019), 'The decline in volatility in the US economy. a historical perspective', *Oxford Economic Papers* **72**(1), 101–123.

Hamilton, J. D. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica* **57**(2), 357–384.

Hansen, B. E. (1992), 'The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP', *Journal of Applied Econometrics* **7**(S1), S61–S82.

Kasahara, H. & Shimotsu, K. (2018), 'Testing the number of regimes in Markov regime switching models', *arXiv preprint arXiv:1801.06862* .

Kim, C.-J. & Nelson, C. R. (1999), 'Has the us economy become more stable? a Bayesian approach based on a Markov-switching model of the business cycle', *Review of Economics and Statistics* **81**(4), 608–616.

Qu, Z. & Zhuo, F. (2021), 'Likelihood ratio-based tests for Markov regime switching', *The Review of Economic Studies* **88**(2), 937–968.