

Beccarini, Andrea

Article — Published Version

A new estimator for the constrained regime-switching model

AStA Advances in Statistical Analysis

Suggested Citation: Beccarini, Andrea (2025) : A new estimator for the constrained regime-switching model, AStA Advances in Statistical Analysis, ISSN 1863-818X, Springer Berlin Heidelberg, Berlin/Heidelberg, Vol. 110, Iss. 2, pp. 371-406,
<https://doi.org/10.1007/s10182-025-00540-7>

This Version is available at:

<https://hdl.handle.net/10419/342030>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<http://creativecommons.org/licenses/by/4.0/>



A new estimator for the constrained regime-switching model

Andrea Beccarini¹

Received: 27 August 2024 / Accepted: 1 September 2025 / Published online: 25 September 2025
© The Author(s) 2025

Abstract

This article proposes a new method for estimating regime-switching models when some slope parameters are constrained to be non-switching. The constrained parameters are simply found by an appropriate weighted average of the unconstrained parameters. The advantage of this approach is twofold. First, the constrained estimates are obtained by an unconstrained estimation procedure such as the EM algorithm, and hence, the procedure is relatively straightforward to implement. Secondly, as both the constrained and unconstrained estimators are available, testing based on the likelihood ratio and model selection by means of likelihood-based information criteria is particularly simple. The procedure is applied to a three-state Markov-switching variance model.

Keywords Constrained estimation · Regime-switching model · EM algorithm · ML-based tests · ML-based information criteria

JEL Classification C12 · C13 · C24

1 Introduction

This article assumes that the econometric model under examination contains two classes of parameters, which may be either constant (invariant) or subject to regime changes. In a seminal paper, Engle et al. (1983) state: “Over time, it is possible that some of the parameters of the joint distribution may change perhaps through changing tastes, technology, or institutions such as government policy making. ... (however) there may be parameters which remain constant... This is a familiar assumption about parameters in econometrics which is here called invariance. ... This requirement is clearly very strong and its untested assumption has been criticized in conventional practice by Lucas (1976) and Sargent (1981).” As outlined below,

✉ Andrea Beccarini
Andrea.Beccarini@wiwi.uni-muenster.de

¹ Institute for Econometrics and Economic Statistics, University of Münster, Münster, Germany

the view of Engle et al. (1983) characterizes significant and emerging strands of the econometric and empirical literature, particularly in the context of rational expectations-based and DSGE models that incorporate regime changes.

This article is built on the assumption that the appropriate econometric model includes both invariant and regime-dependent parameters. Moreover, the distinction between the two classes of parameters should be testable and the researcher aims at minimizing the computational burden.

Thus, regime-switching (RS) regressions are the natural statistical device to address these issues. These models allow for abrupt and uncertain changes in both cross-sectional and time-series data. The success of these statistical tools relies on the fact that they provide both consistent estimates of the model's parameters and probabilistic inference about the latent variables governing the abrupt changes. The RS models are used in many areas, including bioinformatics, biology, economics, finance, hydrology, marketing, medicine, and speech recognition.

Throughout the article, "RS model" refers to a general class of models, including mixtures of distributions, Markov-switching (MS) models and Time-Varying Transition Probability (TVTP) models; useful basic references may be, among others, Hamilton (1994, Ch. 22), Kim and Nelson (1999) and Frühwirth-Schnatter (2006).

Thus, the objective of this article is to show, in the context of an RS model, how constrained parameter estimates can be obtained from unconstrained ones, possibly (but not necessarily) by applying closed-form solutions, and hence without the need for numerical procedures. As the computation of the constrained and unconstrained RS-variances is already explored in Hamilton (1990), we concentrate on the mean parameters: the RS constants and slope parameters. The constraint in question is that some parameters, such as the slopes, are not RS (i.e., they are constant or invariant). Furthermore, the asymptotic properties of the constrained estimator are directly derived.

Suppose the researcher chooses the EM algorithm of Dempster et al. (1977) or Hamilton (1990) for estimating an RS model. The choice of the EM algorithm as the preferred estimation procedure relies on its stability and the fact that it yields closed-form solutions for many RS models or for ad hoc derivations of constrained Maximum Likelihood (ML) estimators. It is also proposed for didactic purposes, specifically for deriving the asymptotic properties of the constrained estimator. It will be shown that the constrained estimator can be computed simply through a weighted average of the unconstrained estimators. Thus, the constrained estimator can still be obtained using the closed-form solutions provided by the standard EM algorithm. This avoids modifying some steps of the (basic) EM algorithm or applying alternative estimation procedures, which are generally based on numerical and less stable solutions (i.e., the classical (or direct) ML approach and the Gibbs-sampling), particularly in the context of constrained maximization problems.¹

It is worth here to remark that the unconstrained estimates can be obtained using any estimation procedure: it is always preferable to obtain a single set of

¹ The performance of the EM algorithm is analyzed, among others, in Redner and Walker (1984) and Hamilton (1990).

(unconstrained) estimates and then compute their constrained counterparts, rather than starting with the constrained estimates or estimating both sets separately. Furthermore, since both the unconstrained and constrained estimators are immediately available, the Likelihood Ratio (LR) test is the natural way to verify the validity of the constraint. The constrained model under consideration has RS in at least a subset of its parameters. Thus, the testing procedure is not based on the null hypothesis of no RS, as in Di Sanzo (2009) and references therein.

The invariance property of the ML estimator (see Greene 2003, Ch. 14) cannot be applied because it requires a deterministic function linking the constrained with the unconstrained estimators. On the other hand, the constraint in question is stochastic, as it depends on the specific available sample. In fact, to construct the constrained estimator, the smoothed probabilities are needed (see Hamilton (1994, Ch. 22)). These probabilities are a natural byproduct of the EM algorithm. The fact that the constraint is stochastic implies that the constrained estimator does not necessarily inherit the (asymptotic) properties of the unconstrained counterpart. Thus, the asymptotic properties of the constrained estimator must be explored.

These kinds of constraints are considered in several seminal papers. Hamilton (1989) used an RS model for measuring the switches in the business cycles, where the autoregressive parameters are constrained to be non-RS and estimates were obtained through numerical procedures. Hamilton and Susmel (1994) outlined an autoregressive model for stock returns with ARCH effects subject to RS. Hamilton and Lin (1996) used a bivariate model for the growth rate of industrial production and the excess return on stocks, where the vector of constants is specified as RS and the autoregressive parameters as regime independent. Billion and Caporin (2005) allowed a dynamic conditional correlation model to have some RS parameters. More recently, Beccarini and Kempa (2023) modeled unobserved heterogeneity in panel data as RS while maintaining some parameters constant. The simplicity of computation of both the constrained and unconstrained parameters can be particularly appreciated when dealing with macro-econometric systems with changes in regimes, for which the computational burden of estimating both constrained and unconstrained estimates may be substantial and sometimes infeasible. In fact, a vast strand of literature has explored such systems in a vector autoregressions context. For this purpose, the reader may consider Krolzig (1997), some important extensions such as Cavicchioli (2014b), and applications as in Cavicchioli (2022). A related strand of the literature deals with multivariate Markov-switching models, and in particular, their combination with Dynamic Stochastic General Equilibrium (DSGE) models and rational expectations frameworks. All the below-cited works make use of distinct classes of RS and non-RS parameters, providing flexible frameworks for structural analysis of regime-dependent dynamics and the economic interpretation of results. Farmer et al. (2011), and Cho (2021) have advanced solution and estimation techniques for MS rational expectations models. These studies provide deeper economic accounts of recurrent regime shifts in terms of changes in the behavior of private agents and policymakers, as well as in the volatilities of structural shocks. There is also increasing interest in the interplay between regime switches, agents' beliefs, and macroeconomic dynamics as shown by Bianchi

(2013), who incorporates agents' expectations and regime changes in monetary policy behavior within an MS-DSGE model framework. In this context, Hur (2017) specified the central bank reaction function as dependent on output growth, inflation, and stock prices, allowing only the parameter related to stock prices to be RS. Chang et al. (2021) introduce a monetary policy that switches between hawkish and dovish regimes, while deep economic parameters remain constant. They also identify the drivers of policy shifts. To this end, the proposed estimating and testing procedure aims at reducing the substantial computational burden of these models. Based on unconstrained estimates, econometric routines can automatically compute the constrained coefficients and the related statistics.

Similar estimators can be found in the literature. Aitkin (1999) outlined models with shared random effects arising through a variance component or repeated-measures structure and provided nonparametric maximum likelihood estimates. Bartolucci and Farcomeni (2009) considered a dynamic logit model, where the responses depend on covariates and some unobserved heterogeneity. Importantly, they are ML estimators, that is based on f.o.c. derived from a constrained log-likelihood function. Thus, the proposed estimator, which is a sort of Generalized Least Squares estimator, has to be preferable for the following reasons. First, it does not require an ad hoc derivation of the ML estimator, which is specific to the model of interest. Second, simulation exercises indicate that small-samples properties of the proposed estimator outperform those of the direct ML estimator.

Thus, for the proposed estimator, the relevant theoretical asymptotic properties are provided. Further simulation exercises are also provided in order to verify how the constrained estimator, the ML-based tests and the information criteria (AIC and BIC) behave in small samples, and in case of (non)-separation of regimes, both in the null and alternative hypotheses. In the simulation exercises, the performance of the ML-based statistics are verified, i.e., the LR, Wald and Lagrange Multiplier (LM) tests. The performance of the Akaike (AIC) and Schwarz or Bayesian (BIC) Information Criteria is also explored; for details, the reader may consider Leroux (1992), Leroux and Puterman (1992), Cavicchioli (2014a) and references therein.

As an empirical application, the article of Kim et al. (1998) is considered, which is also reported in Kim and Nelson (1999). The mean-reverting hypothesis of the monthly stock returns is tested. This analysis relies on a simple three-state Markov-switching Variance Model. Kim et al. (1998) estimated this model with constrained mean parameters, without testing this restriction. Whether this restriction is consistent with the data is verified, among others, through the maximum likelihood-based tests.

This article is organized as follows. The next section outlines the RS model. The third section presents the estimation strategy for obtaining the constrained estimator and verifies its statistical properties. Section four presents the simulation results. Section five considers the empirical application. Conclusions follow.

2 Outlining the RS model

2.1 Specifying the unconstrained RS model

To introduce the procedure, consider the typical representation of an RS model for the dependent variable y_m explained by the vector of exogenous or predetermined variables x'_m , with M observations:

$$y_m = \alpha_{s_m} + x'_m \gamma_{s_m} + u_m \quad (1)$$

where $m = 1, \dots, M$ with $u_m | s_m \sim i.d.(0, \sigma_{s_m}^2)$, (*i.d.* stays for independent distribution) α_{s_m} , γ_{s_m} and σ_{s_m} are RS parameters whose values are determined by the value of a latent variable, s_m ($m = 1, \dots, M$), which may take R values: $1, 2, \dots, R$.² As anticipated, with an RS model it is meant a class of models with RS, such as MD, MS and TVTP models. For didactic purposes, the reader may consider the MS model with an irreducible transition probability matrix (TPM) as the reference model and the other two as variants.

2.2 Setting the analysis in matrix notation

It is convenient to cast the model in matrix notation, as this is simpler for considering their statistical properties:

$$Y = [S_1 \cdot [i_M \ X_1] \dots S_R \cdot [i_M \ X_1]] \cdot \beta + U \quad (2)$$

where Y is a $(M \times 1)$ vector, i_M is a $(M \times 1)$ vector of ones, X_1 is $(M \times K)$ matrix of exogenous, explanatory variables,³ and β is a $((K+1) \cdot R \times 1)$ vector of parameters. The information of the latent $(M \times 1)$ vector S (whose typical element is s_m , $m = 1, \dots, M$) is embedded in the S_r $(M \times M)$ diagonal matrices $r = 1, \dots, R$, such that each m th-element, $m = 1, \dots, M$, on the main diagonal of S_r is 1 if the m th-observation belongs to regime r ($s_m = r$) and 0 otherwise (such that $\sum_{r=1}^R S_r = I_M$ where I_M is an identity matrix of order M).

In order to understand this matrix notation, note that the pre-multiplication of the matrix of regressors through the matrix S_r selects only the observations in X_1 , which belong to regime r . Note also that β collects all parameters of the mean regression, i.e., the first $K+1$ elements of β belong to Regime 1, the second $K+1$ elements of β belong to Regime 2 and so on. U is a $(M \times 1)$ vector of error terms. Later on, it will be convenient to write the model of Eq.(2) in a more compact form by defining:

$$[S_1 \cdot [i_M \ X_1] \dots S_R \cdot [i_M \ X_1]] = X \quad (3)$$

² However, sometimes the shortcut: $\alpha_{s_m=r} = \alpha_r$ is used.

³ With the expression $[i_M \ X_1]$ we mean a $(M \times (K+1))$ matrix whose first column is i_M and the last K columns corresponds to X_1 . This notation holds throughout the article.

which renders the model of Eq.(2) linear, conditional on knowledge of the matrices S_r ($r = 1, \dots, R$). In particular, for Eq.(2) it holds

$$\beta = [\alpha_1 \ \gamma_1' \ \alpha_2 \ \gamma_2' \ \dots \ \alpha_R \ \gamma_R']' \quad (4)$$

Only the statistical properties of the constrained model are explored in the next subsection.

3 Estimation strategy and statistical properties of the estimation

This section begins by describing the necessary statistical properties of the disturbance term and the statistical relationships between the relevant variables of the RS model. Next, it clarifies how to obtain the RS estimator based on the EM-algorithm, when the relevant constraints are applied. Finally, the statistical properties of the constrained RS estimator are explored.

3.1 Statistical assumptions underlying the data-generating process

In the previous subsection, the error term U was assumed to be *i.d.*. This implies that⁴:

$$E[U|S_R, \dots, S_1] = 0_M \quad (5)$$

$$E[UU'|S_R, \dots, S_1] = \sum_{r=1}^R \sigma_r^2 S_r \quad (6)$$

The special case of $\sigma_i^2 = \sigma^2$, ($i = 1, \dots, R$) is also considered. We further specify the processes determining the values of X_1 and of U as joint covariance stationary and ergodic (see Greene 2003, Ch. 20) within (or conditional on) each specified regime.⁵ Together with the assumption that the error term is conditionally zero in expectation with finite variance, this implies:

$$E[x_m' u_m | S_R, \dots, S_1] = 0 \quad m = 1, \dots, M \quad (7)$$

Thus, within each regime, $x_m' u_m$ is zero in expectation with finite variance. Stationarity of the process generating the elements of X_1 within regimes also implies Eq.(8) and Eq.(9). In fact, under stationarity and conditional on regime r , it holds that:

$$E[x_m' | s_m = r] = v_r \quad r = 1, \dots, R \quad (8)$$

⁴ Note that the matrices $S_R, \dots, S_1$ express the same information as S , collecting the values of s_m , $m = 1, \dots, M$. It is convenient to express the condition in terms of the R matrices.

⁵ The reader may also consider Francq and Zakoian (2001) on this topic.

Thus, v_r is the $(1 \times K)$ vector of finite elements containing the expectation of the vector x_m , conditional on regime r and for observation m . Furthermore:

$$E[x_m x_m' | s_m = r] = Q_r \quad (9)$$

where Q_r is a $(K \times K)$ matrix of finite elements, if the m -th observation belongs to regime r , ($m = 1, \dots, M; r = 1, \dots, R$). Note that Eq.(8), Eq.(9) and ergodicity are sufficient to conclude:

$$\text{plim}(t_r^{-1} i_M' S_r X_1) = v_r \quad r = 1, \dots, R. \quad (10)$$

$$\text{plim}(t_r^{-1} X_1' S_r X_1) = Q_r \quad r = 1, \dots, R. \quad (11)$$

where t_r is the number of occurrences of regime r , such that $t_r = \text{tr}(S_r)$; the *plim* operator refers to the condition, $t_r \rightarrow \infty$, ($r = 1, \dots, R$).⁶ This means that there is sufficient information for each regime (which also implies $M \rightarrow \infty$).

Now define as P_r the $(M \times M)$ matrix estimating S_r , $r = 1, \dots, R$. In general terms, these matrices embed smoothed estimations, which implies that the underlying model is a MS or a TVTP model, see Hamilton (1994, Ch. 22) or Kim and Nelson (1999, Ch. 4) for an introduction. If the reference model is an RS model with independent switches (or a MD model), the R matrices P_r embed the conditional (on current values) or posterior probabilities of each regime occurring.

3.2 Constructing the constrained estimator

In order to show the proposed procedure without loss of generality, we assume the constraint that all slope parameters are equal ($\gamma_r = \gamma$, $r = 1, \dots, R$), and hence the intercepts remain unconstrained (regime-dependent). Define θ as the vector of true parameters containing both the R intercepts (α_r , $r = 1, \dots, R$) and the common slope (γ): $\theta' = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_R \ \gamma']'$. The proposed estimator is the following:

$$\hat{\theta} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_R \\ \sum_{r=1}^R w_r \hat{\gamma}_r \end{bmatrix} \quad (12)$$

with:

$$w_r = \frac{\frac{p_r}{\sigma_r^2}}{\sum_{j=1}^R \left(\frac{p_j}{\sigma_j^2} \right)} p_r = \text{tr}(P_r), \quad r = 1, \dots, R; \quad (13)$$

⁶ Throughout the article, the *plim*(.) is an element-by-element operator and *tr*(.) is the trace operator.

which collapses to $w_r = p_r/M$ in the special cases of equal variances ($\sigma_r^2 = \sigma^2$, $r = 1, \dots, R$)⁷.

Thus, the proposed constrained estimator $\hat{\gamma}$ is obtained simply by running an unconstrained RS regression and then taking an appropriate weighted average of the unconstrained $\hat{\gamma}_r$ s: $\sum_{r=1}^R w_r \hat{\gamma}_r$. Appendix A.1 guides the reader in deriving the estimator of Eq.(12) starting from the unconstrained estimator. The next section verifies under which conditions the estimator of Eq.(12) is consistent, efficient and asymptotically normal.

3.3 Statistical properties of the constrained estimator

The constrained estimator of Eq.(12) turns out to be consistent, efficient and asymptotically normal under some established conditions outlined below. Each of these statements is now explored.

In order to evaluate the above properties, the RS estimators are considered as weighted (or generalized) least-squares estimators, as in Beccarini (2019), Hamilton (1990), Hamilton (1994, Ch. 22) and Krolzig (1997, Ch. 6 and 8).

Identifying the RS estimators as weighted least-squares estimators, and hence avoiding the classical maximum-likelihood framework, is particularly convenient, as it enables easily finding their asymptotic properties. In fact, according to White (1991, Ch. 1), any linear transformation of a linear stochastic relationship can be considered within the framework of standard asymptotic theory. Small sample properties are then verified through a simulation exercise.

Now, θ can be expressed directly in the model in place of β , and its estimator is evaluated. Thus, having defined X in Eq.(3), now rewrite Eq.(2) in a more compact form: $Y = X\beta + U$, and substitute out β through Eq.(25), yielding $Y = X\beta + U = XH\theta + U$, conditional on the specification of the matrix H . Proceeding as before, conditional on knowledge of the matrices S_r , $r = 1, \dots, R$, which determine X , $\hat{\theta}$ can be found through the OLS⁸:

$$\hat{\theta} = (H'X'XH)^{-1}H'X'Y. \quad (14)$$

The theoretical properties of $\hat{\theta}$ can now be found as usual by substituting out Y , yielding:

$$\hat{\theta} - \theta = (H'X'XH)^{-1}H'X'U. \quad (15)$$

Note that the theoretical properties of $\hat{\theta}$ will be found without assuming any theoretical property of the unconstrained estimator $\hat{\beta}$.

⁷ Note that $\frac{p_r}{M}$, with $\sum_{r=1}^R \frac{p_r}{M} = 1$, is an estimate of the ergodic probability of regime r occurring, if it exists; $r = 1, \dots, R$.

⁸ It can be shown that the above formula and that of Eq.(28) are equivalent.

The following statements about Eq.(15) can now be made. They should be interpreted in the context of the EM algorithm, which alternates the E-step with the M-step in an iterative manner. That is, Eq.(14) may be interpreted as the outcome of the M-step, for which a closed-form solution is obtained; meanwhile, Eq.(19) may be interpreted as a possible outcome of the E-step. This implies that Theorems 1 and 2, in order to be applicable, require the validity of the following Theorem 3 and vice versa. In fact, suppose that L iterations are necessary to obtain the final estimates. Define as l the iteration index, $l = 0, \dots, L$. At the beginning ($l = 0$), some starting values for the parameters are assumed, from which the smoothed probabilities are found. In general terms, these are not equal to the values embedded in S_i , ($i = 1, \dots, R$). However, for the successive iterations, new updates of the parameter vector and the smoothed probabilities are obtained. Due to the property of the EM algorithm, each of these steps results in values which are approaching the final estimate, see Hamilton (1990). Furthermore, based on Theorem 3 (outlined below), if separation of regimes is present, the smoothed probabilities approach 0 or 1. At the $(L - 1)$ th step, the parameter estimates are in a neighborhood of the final estimate, and hence, due to Theorem 3, the smoothed probabilities are approaching S_i , $i = 1, \dots, R$. At the L th step, the maximum-likelihood estimate is obtained by assumption. Thus, both theorems can be applied.

Theorem 1 (Asymptotic normality and Consistency). The right-hand side of Eq.(15) is asymptotically normally distributed with zero expectation and a finite covariance matrix, when Eq.(5), Eq.(6), Eq.(7), Eq.(10), Eq.(11) and Eq.(19) are assumed. Furthermore, the limit in probability value of Eq.(15) is zero, implying the weak consistency of the estimator.

Proof See Appendix A.1. □

Theorem 2 (Efficiency). Assuming Theorem 1 and Eq.(19), the estimator defined in Eq.(12) attains the Cramer–Rao lower bound in the class of the consistent estimators.

Proof See Appendix A.2. □

Standard errors of constrained estimator $\hat{\theta}$ can be found as usual by the relevant covariance matrix, which is specified in the following corollary.

Corollary 3.1 The covariance matrix of the constrained estimator $\hat{\theta}$ can be obtained as follows.

For $\hat{\alpha}_j$ ($j = 1, \dots, R$) it holds:

$$\text{Var}[\hat{\alpha}_j] = \left(I_{(K+1)} - I_{(K+1)}^0 \right) \text{Var}[\hat{\beta}_j] \left(I_{(K+1)} - I_{(K+1)}^0 \right). \quad (16)$$

where $(I_{(K+1)} - I_{(K+1)}^0)$ is a $(K + 1)$ square matrix of zeros except for the $(1, 1)$ -element, which is 1.⁹ Moreover, $\text{Var}[\hat{\beta}_j] = \sigma_j^2 (X_2' P_j X_2)^{-1}$, as defined in Eq.(49). For $\sum_{i=1}^R w_i \hat{\gamma}_i$ it holds:

$$\text{Var} \left[\sum_{i=1}^R w_i \hat{\gamma}_i \right] = I_{(K+1)}^0 \text{Var} \left[\sum_{i=1}^R w_i \hat{\beta}_i \right] I_{(K+1)}^0. \quad (17)$$

with

$$\text{Var} \left[\sum_{i=1}^R w_i \hat{\beta}_i \right] = \sum_{i=1}^R w_i^2 \sigma_i^2 (X_2' P_i X_2)^{-1} \quad (18)$$

and $w_i = \frac{\frac{p_r}{\sigma_r^2}}{\sum_{j=1}^R \left(\frac{p_j}{\sigma_j^2} \right)}$; $I_{(K+1)}^0$ is defined such that $\hat{\gamma}_i = I_{(K+1)}^0 \hat{\beta}_i$ ($i = 1, \dots, R$).

Proof Considering the details provided in the proof of Theorem 2, this proof is trivial. $\square$

We now address the issue of the relationship between P_r and S_r , $r = 1, \dots, R$. This can be expressed as:

$$\text{plim}(P_r) = S_r \quad (19)$$

With smoothed probabilities we mean the outcome of the smoothing algorithm proposed by Kim (1994), Hamilton (1994, Ch. 22), Krolzig (1997, Ch. 22) and extensions such as Alvarez et al. (2019).

Theorem 3 (Consistency of P_i , $i = 1, \dots, R$). Given Theorem 1 (consistency of $\hat{\theta}$) and consistency of $[\hat{\sigma}_1 \dots \hat{\sigma}_R]$, Eq. (19) holds.

Proof The condition that $M \rightarrow \infty$ ensures that the values of P_r , $r = 1, \dots, R$, (computed in the E-step of the EM algorithm) must converge to 0 or 1, see Redner and Walker (1984). Furthermore, each element of P_r is ultimately a nonlinear, continuous function of the underlying (in the M-step estimated) parameters, which are assumed to be consistent (Theorem 1).

Then, due to the theorem of the limit in probability of continuous functions (or Slutsky theorem, see White (1991, Ch. 2, Proposition 2.27), Eq.(19) must hold by construction. $\square$

Thus, Theorem 1 justifies Eq.(14) as the outcome of the M-step. On the other hand, as the possible outcome of the E-step, Theorem 3 justifies Eq.(19). This

⁹ Further details of $I_{(K+1)}^0$ can be found in Appendix A.2.

iteration is due to the iterative feature of the EM algorithm. In order to see this circularity from another perspective, consider the Generalized Method of Moments (GMM) of Hansen (1982). This procedure exhibits a very similar situation. In fact, Hamilton (1994) in Ch. 14, 'Optimal Weighting Matrix', reports the circularity between the main parameters to be estimated and the weighting matrix. This is overcome by alternating the estimation of the two quantities until the change of the estimated main parameter is negligible. Indeed, this is the same way the EM algorithm deals with the alternation of the E-step and the M-step. Furthermore, as already noted, the RS estimator can be considered as a type of Generalized Least Squares Estimator, which, in turn, is a special case of the more general GMM.

Before considering the simulation exercises, the concept of separation of regimes is outlined and in particular to its relationship with consistency. First, it should be noted that for Theorem 3 only a weak consistency requirement is needed: considering each element of the vector $|diag(P_r - S_r)|$, the probability that each element of this vector is smaller than every $\epsilon > 0$ tends to 1, as $M \rightarrow \infty$. Secondly, this asymptotic property is affected by how much regimes separation is present in the data.¹⁰ Roughly speaking, for almost all observations, separation means that densities conditional on regimes are sufficiently distant from each other, in a way that the densities product is approximately zero for each value of their support; for details the reader may consider Redner and Walker (1984). In fact, for a given $\epsilon > 0$, the value of $M(\epsilon)$, such that for all $M > M(\epsilon)$, each element of $|diag(P_r - S_r)|$ is smaller than ϵ in a probabilistic sense, does depend on the separation of regimes.

4 Simulations

Simulations have a threefold objective. The analysis of (small sample) properties of the proposed constrained estimator also compared with the constrained direct ML estimator, the ML-based tests (Wald, LM and LR), and the AIC and BIC.

Simulations are based on an MS model as specified in Eq.(1) with $R = \{2, 3\}$, for which the constants are unconstrained (RS) and the slope parameter of the unique regressor is constrained to be non-RS and equal to 1¹¹ and $\sigma_i = 1$ ($i = 1, \dots, R$). Regarding the comparison between the proposed and the direct ML estimators and ML-based tests, we consider only the case $R = 2$, for simplicity of exposition. Separation of regimes, as defined in Redner and Walker (1984), is also considered. Separation of regimes is here measured by the distance of the RS-constants, while maintaining the other parameters fixed. The DGP with poor separation considers RS constants equal to $(-1, +1)$ for $R = 2$ and $(-2, 0, +2)$ for $R = 3$. The DGP with good separation considers RS constants equal to $(-2, +2)$ for $R = 2$ and $(-4, 0, +4)$ for

¹⁰ In fact, the condition of Eq.(19) requires the number of observations in each regime to go to infinity.

¹¹ As the econometric model does not involve any dynamics for the dependent or the explanatory variables, any nonzero and finite value of the slope parameter and of the variance is appropriate. Thus, for clarity of expositions the slope parameter is set to 1.

$R = 3$. The evolution of the latent variable is governed by a Markov chain of the first order, described by an ergodic TPM.¹² The number of simulations is always 1000.

Table 1 shows the constrained parameter estimates of the proposed estimator as a function of the number of regimes, separation of regimes and the sample size. For reasons of space, the estimates of RS constants and variances are not reported; however, their convergence properties are equivalent to that of the constrained estimator.¹³

Table 1 shows that the average constrained estimator is asymptotically unbiased, although this property depends on how good regimes are separated. In fact, the small samples bias vanishes more rapidly when separation is good.

Table 2 compares small and large samples properties of the constrained direct ML estimator with the proposed estimator and also outlines the unconstrained estimates (on which the proposed estimator is based). Computations are based on the same starting values. The numerical procedure used for the direct ML estimator is based on the `optim()` function in R, with the default optimization method. Sample sizes range from $T = 50$ to $T = 1000$; $R = 2$, high separation is assumed.

There are remarkable differences in the performance of the two alternative estimators for all of the mean parameters. In large samples, this difference decreases, but the imprecision of the direct ML likelihood compared to the EM-based estimates remains considerable. In small samples, the precision of the performed estimator is remarkably higher than that of the ML estimator. Its inaccuracy vanishes gradually with the sample size increases.

The performances of the ML-based tests, the AIC and the BIC are now considered. Tables 3, 4, and 5 show the behavior of the LR tests for different sample sizes, separation and number of regimes. In particular, results are based on different critical levels $\{\alpha = 0.01, 0.05, 0.10\}$, for which the corresponding percentiles of the relevant chi-squared distributions are found. For example, for the 3-Regimes case, there are 2 constraints, and hence the relevant chi-squared distribution has 2 degrees of freedom. Each table reports the percentage of the simulated LR tests, which are equal or higher than the assumed percentile. For example, Table 3 is based on $\alpha = 0.01$ and hence the fraction of the simulated LR tests which are equal or larger than the 99%-percentile of the relevant chi-squared distributions is reported. In this way, it can be verified how the LR test is, in small samples, distributed. Fractions similar to the assumed critical values show that the LR test is chi-squared distributed. A property which, in general terms, is valid only for large samples (Tables 3, 4, 5).

From the comparison of the above values, the following conclusion can be drawn. As expected, the LR is only asymptotically chi-squared distributed. However, the convergence speed depends on the number of regimes and separation. Having good separation implies that the convergence to the relevant chi-squared distribution is

¹² For $R = 2$, the parameters of its main diagonal are 0.8 and 0.7. For $R = 3$, the parameters of the main diagonal are 0.8, 0.7 and 0.6 and the parameters not belonging to the main diagonal are equal.

¹³ Equivalent results are obtained when simulation are based on a similar mixtures-of-distributions model; results are available upon request.

Table 1 Constrained estimates of the proposed estimator

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
<i>No. regimes</i>						
2	1.05	0.98	1.01	1.03	0.99	1.00
3	0.94	0.97	1.02	0.95	0.99	1.00

This table shows the average constrained estimators over 1000 simulations, with different sample sizes and different separation of regimes. The true parameter is equal to 1

Table 2 Comparison between the proposed and the ML estimators

T	<i>i</i>	Unconstrained RSE		Constrained RSE		Proposed RSE
		α_i	β_i	α_i	β	
50	<i>i</i> = 1	-2.01	1.03	-2.16	0.93	1.04
	<i>i</i> = 2	1.89	1.04	1.64	—	—
500	<i>i</i> = 1	-2.01	1.00	-2.14	0.96	1.01
	<i>i</i> = 2	2.00	1.01	1.71	—	—
1000	<i>i</i> = 1	-2.00	1.00	-2.12	0.97	1.00
	<i>i</i> = 2	2.00	1.00	1.86	—	—

This table compares small and large samples properties of the constrained direct ML estimator (columns 5 and 6) with the proposed estimator (column 7), which is based on the unconstrained estimator (columns 3 and 4)

obtained by a smaller sample size.¹⁴ In small samples with poor separation and for $R = 3$, the LR test tends to falsely reject the null hypothesis of the constrained model more often with respect to the rejection rates obtained by the appropriate chi-squared distribution.

The next tables compare the performance of all the ML-based tests by reporting the fraction of each test, which is equal to or larger than the relevant percentile. This, in turn, is based on the usual critical levels $\{\alpha = 0.01, 0.05, 0.10\}$. Simulations are obtained for $R = 2$, for the usual sample sizes and for poor and good separation of regimes.¹⁵

The results of Tables 6, 7, and 8 lead to the following conclusions. As expected, the sample size and the regime separation play the usual roles in determining the performance of the ML-based tests. Importantly, the LM test performs relatively poorly (or not better) in almost all situations, making the Wald test the most reliable. This further confirms that, at least for a 2-State MS model, the estimations are better

¹⁴ This is in line with the theoretical observation about separation of the last subsection.

¹⁵ As $R = 2$ and hence one constraint is assumed, the relevant percentiles of the chi-squared distribution associated with the above critical levels are 6.63, 3.84, 2.71, respectively.

Table 3 Fraction of the LR values equal or larger than the 99%-percentile

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
<i>No. regimes</i>						
2	0.22	0.02	0.01	0.03	0.01	0.01
3	0.28	0.12	0.01	0.25	0.02	0.01

This table shows the fraction of the LR tests equal or larger than the relevant 99%-percentiles (6.63 for the 2-Regimes case and 9.21 for the 3-Regimes case)

Table 4 Fraction of the LR values equal or larger than the 95%-percentile

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
<i>No. regimes</i>						
2	0.18	0.07	0.05	0.09	0.05	0.05
3	0.33	0.18	0.04	0.33	0.08	0.05

This table shows the fraction of the LR tests equal or larger than the relevant 95%-percentiles (3.84 for the 2-Regimes case and 5.99 for the 3-Regimes case)

Table 5 Fraction of the LR values equal or larger than the 90%-percentile

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
<i>No. regimes</i>						
2	0.22	0.15	0.10	0.13	0.12	0.10
3	0.40	0.23	0.06	0.40	0.14	0.10

This table shows the fraction of the LR tests equal or larger than the relevant 90%-percentiles (2.71 for the 2-Regimes case and 4.61 for the 3-Regimes case)

Table 6 Fraction of the ML-based tests equal or larger than the 99%-percentile

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
LR	0.22	0.02	0.01	0.03	0.01	0.01
LM	0.24	0.08	0.03	0.06	0.02	0.01
Wald	0.11	0.02	0.01	0.02	0.01	0.01

This table shows the fraction of the LR, LM, and Wald values equal or larger than the 99%-percentile over 1000 simulations for $R = 2$, different sample sizes and different separation of regimes

designed if based on an unconstrained model, and this holds for both estimation and testing.

Now, the differences between the AIC and BIC, each evaluated at the unconstrained and the constrained models, are computed. Since these statistics have to

Table 7 Fraction of the ML-based tests equal or larger than the 95%-percentile

Sample size	poor separation			good separation		
	50	200	1000	50	200	1000
LR	0.18	0.07	0.05	0.09	0.05	0.05
LM	0.20	0.15	0.07	0.16	0.06	0.05
Wald	0.06	0.06	0.05	0.05	0.05	0.05

This table shows the fraction of the LR, LM, and Wald values equal or larger than the 95%-percentile over 1000 simulations for $R = 2$, different sample sizes and different separation of regimes

Table 8 Fraction of the ML-based tests equal or larger than the 90%-percentile

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
LR	0.22	0.15	0.10	0.13	0.12	0.10
LM	0.24	0.20	0.12	0.25	0.12	0.10
Wald	0.11	0.12	0.10	0.08	0.09	0.10

This table shows the fraction of the LR, LM, and Wald values equal or larger than the 90%-percentile over 1000 simulations for $R = 2$, different sample sizes and different separation of regimes

Table 9 Percentage of correct model identification through the AIC

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
<i>No. regimes</i>						
2	0.85	0.86	0.86	0.75	0.82	0.84
3	0.57	0.73	0.83	0.58	0.80	0.87

This table shows the fraction of the positive differences of the AIC values. This positive difference indicates that the AIC identifies the constrained model as the true specification

Table 10 Percentage of correct model identification through the BIC

Sample size	Poor separation			Good separation		
	50	200	1000	50	200	1000
<i>No. regimes</i>						
2	0.88	0.95	0.98	0.89	0.96	0.99
3	0.66	0.83	0.96	0.70	0.86	0.99

This table shows the fraction of the positive differences of the BIC values. This positive difference indicates that the BIC identifies the constrained model as the true specification

be minimized in order to choose the best model, a positive number for each of these differences indicates a correct selection of the model (the constrained specification). Tables 9 and 10 report the percentage of these positive values.

The AIC proves reliable in large samples (with a percentage of success over 80%), but also in small samples if $R = 2$. Good separation does not seem to be an advantage in small samples.

Overall, the BIC criterion exhibits a better performance with respect to the AIC criterion. The BIC performs particularly well in large samples (due to its consistency property), and it performs worst when $R = 3$ and regimes are poorly separated.

The next estimation exercise considers the DGP under the alternative hypothesis, i.e., the slope parameters are not constrained to be equal. It aims at verifying the power of the ML-based tests, in particular in small samples ($N = 50$) with two different degrees of poor separation.¹⁶ In both cases, the RS constants equal to $(-1, +1)$ and $\sigma_i = 1$ ($i = 1, \dots, R$), but the slope parameters are set to $(+1, -1)$ and $(+0.1, -0.1)$ for very poor separation. To this purpose, the fraction of the ML-based that are equal or larger than the usual percentiles is considered. This is a measure the probability of rejecting the null hypothesis (being it false), i.e., the power of the test.

From Tables 11 and 12, it is clear that the LM-based test outperforms the other tests, being the Wald test the less performing. This particularly holds with very poor separation of regimes (Table 12). All performances are improved by the increase of the sample size, indicating that the tests are consistent (their power tends to one asymptotically).

Based on performance comparisons of the ML-based tests under both the null and the alternative hypotheses, it seems that LR-test performs relatively well in both hypotheses.

Performances of the AIC and BIC criteria are very similar, showing little detection of the correct model in case of very poor separation of regimes and small samples size (for reason of space, they are available only upon request).

5 Estimating the 3-State MS Variance model

In order to describe the easiness of the computation of the constrained parameter and the calculation of the LR test, the article of Kim et al. (1998) is now considered. It deals with a variance ratio test for verifying the mean-reversion hypothesis of the monthly returns (NYSE stocks). The scope of this section is not to provide a comprehensive analysis of stock returns but to verify possible misspecification in the model of Kim et al. (1998).

A three state MS Variance is considered, with the following assumptions:

$$y^* \sim N(0, \sigma_t^2) \quad (20)$$

where y^* is the demeaned value of y representing the time series of the monthly returns, monthly CRSP data, 1926:1-1986:12, equal-weighted. Moreover, σ_t^2 represents the RS-variance of y , such that:

¹⁶ There is a presumption, that under very good separation, the testing procedure must perform satisfactorily. That is why, for reason of space, this situation is not considered.

Table 11 The power of the ML-based test with (+1, -1) as slope parameters

Sample size	$N = 50$			$N = 200$		
	90%	95%	99%	90%	95%	99%
LR	0.99	0.99	0.97	1.00	1.00	1.00
LM	1.00	1.00	0.98	1.00	1.00	1.00
Wald	0.96	0.92	0.85	1.00	1.00	1.00

This table shows the fraction of the LR, LM, and Wald values equal or larger than the relevant percentiles over 1000 simulations for $R = 2$, different sample sizes ($N = 50, 200$) and poor separation of regimes

Table 12 The power of the ML-based test with (+0.1, -0.1) as slope parameters

Sample size	$N = 50$			$N = 200$		
	90%	95%	99%	90%	95%	99%
LR	0.54	0.46	0.36	0.58	0.48	0.39
LM	0.87	0.83	0.75	0.90	0.87	0.78
Wald	0.31	0.25	0.16	0.41	0.28	0.18

This table shows the fraction of the LR, LM, and Wald values equal or larger than the relevant percentiles over 1000 simulations for $R = 2$, different sample sizes ($N = 50, 200$) and very poor separation of regimes

$$\sigma_t^2 = \sigma_1^2 S_{1t} + \sigma_2^2 S_{2t} + \sigma_3^2 S_{3t} \quad (21)$$

where σ_j^2 are the conditional-upon-regimes variances ($j = 1, 2, 3$); S_t describes the latent variable governing the regimes and switches between 3 states: ($S_{jt} = \{0, 1\}$, $j = 1, 2, 3$, with $\sum_{j=1}^3 S_{jt} = 1$, $t = 1, 2, \dots, T$) in line with an ergodic Markov chain of the first order.

In order to provide a complete description of the data and of the estimation methods, three models are estimated. All of these three models share the same variance specification outlined for σ_t^2 . The first model is based Eq.(20) and hence replicates the results of Kim et al. (1998). The second model considers y as dependent variable and assumes non-RS mean for y , in line with:

$$y \sim N(\bar{y}, \sigma_t^2) \quad (22)$$

where $\bar{y}$ is the unconditional mean of y , which can also be represented as a non-RS parameter of the three state Markov-switching Variance Model.

The third model is an unconstrained 3-state MS Model; that is, both the means and the variances are allowed to switch among regimes:

$$y \sim N(\bar{y}_t, \sigma_t^2) \quad (23)$$

where $\bar{y}_t$ is the conditional-on-regimes mean of y at period t ($t = 1, 2, \dots, T$), such that:

$$\bar{y}_t = \bar{y}_1 S_{1t} + \bar{y}_2 S_{2t} + \bar{y}_3 S_{3t} \quad (24)$$

where $\bar{y}_j$ is the conditional mean of y conditional upon regime j , ($j = 1, 2, 3$). Table 13 provides the main results of the models, which are labeled as “Demeaned RS,” “Constrained RS,” and “Unconstrained RS,” respectively.

From Table 13, it is apparent that the constraint in question as assumed in Kim et al. (1998) is not consistent with the data for the following reasons. First, the regime-switching means $\bar{y}_j$ ($j = 1, 2, 3$) are statistically significant at any conventional significance level.¹⁷ Furthermore, as mentioned in the previous section, the LR test is particularly easy to implement. Results also turn out to be not dependent on which model (either the Demeaned RS or the Constrained RS model) is contrasted with the Unconstrained RS model.¹⁸ Since the Demeaned model exhibits a smaller Log-likelihood value, the Constrained RS model is compared with the Unconstrained RS model. Based on the Log-likelihood values of Table 13, the relevant LR-statistic is equal to $-2(1002.3 - 1005.7) = 6.8$. Under the null hypothesis of equality of means, and given 2 constraints, the LR-statistic is Chi-squared distributed with 2 degrees of freedom: the corresponding p value of 6.8 is 0.033; thus, the null is rejected at the 5% critical level.¹⁹ Moreover, the Wald test leads to rejecting the null of equality of the means with p -value = 0.000. As noted for example in Aitkin and Wilson (1980), for the Unconstrained RS model, or the model of Eq.(23), the ML estimate of the unconditional mean $E\bar{y}_t$ is the sample average of y . This is not the case for the Constrained RS model whose unconditional mean, which coincides with the constrained parameter, cannot be estimated by means of the sample mean, due to its leptokurtic property. Thus, potential coincidence in this estimate of the three models is a further informal evidence in favor to the Unconstrained RS model. Computing $\bar{y}$ for the Unconstrained RS model and hence in line with the formula of Eq.(12) the following value is obtained, 0.01068, whose standard error is 0.002 (p -value : 0.0000). This is equal to the sample average after rounding to the seventh digit. The constrained mean estimator as obtained by the Constrained RS model: 0.01164 whose standard error is 0.002 (p -value : 0.0000). Thus, basically no differences arise between the constrained estimators based on the formula of Eq.(12) and the sample average of y .²⁰ This, again, is an evidence in favor of the of the Unconstrained RS model.

¹⁷ Results may be provided upon request.

¹⁸ Furthermore, very small differences arise in the series of the smoothed probabilities of the Demeaned and Constrained models.

¹⁹ Using the log-likelihood of the Demeaned RS models instead does not change the result. The corresponding LR-statistic is equal to $-2(1001.9 - 1005.7) = 7.6$. Under the null hypothesis of equality of means, and given 2 constraints, the LR-statistic is Chi-squared distributed with 2 degrees of freedom: the corresponding p value of 7.6 is 0.022; thus, the null is rejected at the 5% critical level.

²⁰ The potential differences in the estimates between the Demeaned RS and the Constrained RS models are only because the latter is estimated by means of numerical procedures. Several procedures are used: BFGS, BHHH, Newton-Raphson, with 500 starting values with 100 iterations using 1 standard deviation, with different random generators.

Table 13 Main results of the 3-State Markov-switching estimates

Regimes	Demeaned RS			Constrained RS			Unconstrained RS		
	1	2	3	1	2	3	1	2	3
Means	–	–	–		0.012		0.014	0.419	–0.015
Variances	0.001	0.031	0.004	0.001	0.031	0.004	0.002	0.023	0.010
Erg. prob	0.369	0.127	0.504	0.365	0.129	0.506	0.723	0.012	0.265
LL		1001.9			1002.3			1005.7	
AIC		–2.713			–2.710			–2.716	
BIC		–2.656			–2.648			–2.639	

In the demeaned RS model, the dependent variable is the demeaned return; it replicates the model of Kim et al. (1998). The constrained model assumes a non-RS mean but RS variances for the (non-demeaned) return. The unconstrained model assumes that both means and variances of the (non-demeaned) returns are RS. All parameters are significant, at least at the 1% critical level

This misspecification might have dramatic results for the outcome of the variance ratio test as proposed in Kim et al. (1998). In fact, as apparent in Table 13, the estimations of the RS variances are affected by the regime-switching specifications of the means. Thus, the time-varying unconditional-on-regimes variances, as constructed by means of eq.(21), are apparently different for the Constrained RS model (or the Demeaned RS model) and the Unconstrained RS model. This is because they are based on significantly different means. Figure 1 compares them.

Furthermore, the summary statistics of the standardized returns are also different. As in Kim et al. (1998), standardized returns are computed by dividing the demeaned returns by their historical standard deviations as computed through Eq.(21). Meanwhile, the demeaning of the historical returns based on the Unconstrained RS model is obtained by means of eq.(24).

Figure 2 reports their histograms, and Table 14 reports the main results: it is apparent that the non-rejection of the null hypothesis of normally distributed residuals is stronger for the Unconstrained RS model.

All of this evidence casts doubts on the assumption of the equality of means of excess returns within regimes; importantly, it turns out that the Variance-ratio test used for verifying mean-reversion of returns may be jeopardized. However, based on the proposed procedure of this article, the unconstrained model and the LR statistics are estimated at no extra costs.

6 Conclusion

This article proposes a simple and new way to estimate a regime-switching model in which some mean parameters are constrained to be non-regime-switching. Usually, this requires an ad hoc derivation of the constrained ML estimator or the application of numerical procedures. Instead, the closed-form solutions generally obtained through the EM algorithm may be used, together with the smoothed probabilities. In fact, should the slope parameters be constrained to be non-regime-switching, the

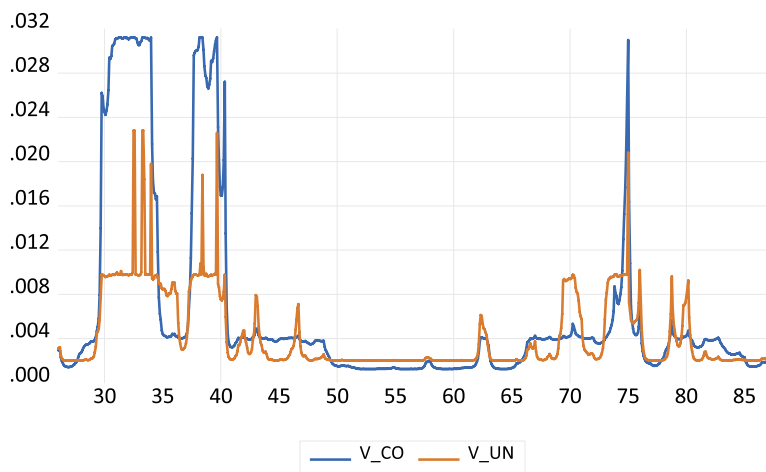


Fig. 1 estimated variances of returns for the constrained and unconstrained models

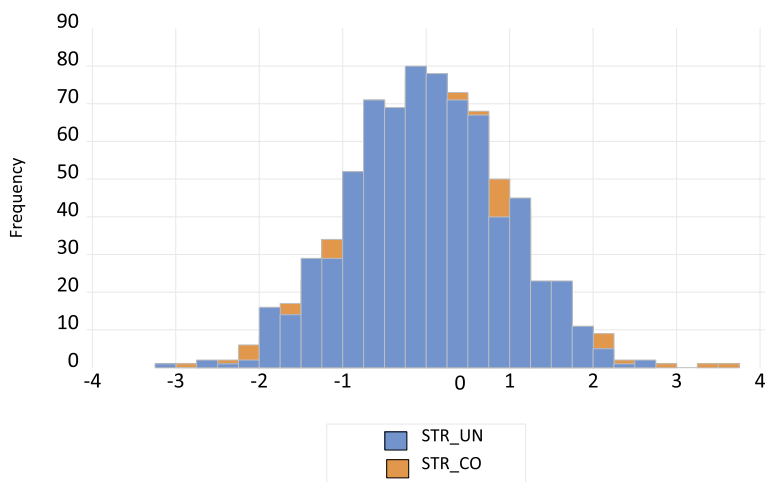


Fig. 2 distribution of standardized returns for the constrained and unconstrained models

constrained slope parameter can be obtained by averaging the regime-switching slopes using the ergodic probabilities and regime-switching variances as weights. However, the use of the EM algorithm is not necessary: what we need is the unconstrained estimates and the smoothed probabilities of each regime to occur.

It has been shown that this estimator is consistent, efficient and asymptotically normally distributed. Small sample properties of the proposed estimator outperform those on the constrained direct ML estimator. Testing the validity of the constraint is also simple. The likelihood ratio test can be easily applied because the analyst disposes, at the

Table 14 Standardized returns

	Unconstrained STR	Constrained STR
Mean	0.009353	0.024976
Median	-0.000241	0.015225
Maximum	2.59732	3.648329
Minimum	-3.011072	-2.773908
Std. Dev	0.911309	0.926932
Skewness	-0.016388	0.14773
Kurtosis	2.83391	3.269575
Jarque–Bera	0.874134	4.878988
Probability	0.645928	0.087205

Standardized returns (STR) are obtained from the historical mean and variances of the unconstrained and constrained models

same time (or at trivial costs), of both the unconstrained and constrained estimators and of the associated log-likelihood values. Based on simulation exercises, under the null hypothesis of constrained slope parameters, the Wald test outperforms the Likelihood ratio and, by large, the Lagrange Multiplier test, remaining all asymptotically equivalent. The Akaike and the Bayesian Information Criteria can be immediately computed; the latter appears to be the most reliable. Simulation exercises are also conducted under the alternative hypothesis of non-constrained slope parameters. As expected, the power of all tests depends negatively on separation and positively on the sample size. The LM-based tests outperform the other tests in all circumstances, being the Wald test the worst performing. Thus, as LR-test performs relatively well in both hypotheses, it turns out to be the most reliable testing procedure. It is worth to remark only asymptotically these tests are consistent. As an empirical application, a three-state Markov-variance model is considered, for which the monthly returns series of the NYSE stock exchange was assumed to have a non-RS mean and a three-state-based variance process. Applying the proposed procedure proved to be more reasonable as both the unconstrained and constrained estimates of the mean and the Likelihood-ratio test were immediately available. All evidence has shown that the unrestricted model with (three state) RS means and variances is more consistent with the data.

Appendices

First appendix

The objective of this annex is first to derive the proposed constrained estimator of Eq.(12). Furthermore, it finds the term $(H'X'XH)^{-1}(XH)'U$ of Eq.(15), and in turn, its asymptotic properties, as stated in Theorem 1.

The constrained estimator

The proposed constrained estimator can be found by the unconstrained estimator. The relationship between the constrained and unconstrained parameters can be defined in an explicit form²¹:

$$\beta = H\theta \quad (25)$$

where H is a $(R \cdot (1 + K) \times (R + K))$ -matrix of constants, which formulate the equality between each γ_i , $i = 1, \dots, R$, and γ :

$$\begin{bmatrix} \alpha_1 \\ \gamma_1 \\ \alpha_2 \\ \gamma_2 \\ \vdots \\ \alpha_R \\ \gamma_R \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0_{1 \times K} \\ 0_{K \times 1} & 0_{K \times 1} & \cdots & 0_{K \times 1} & I_K \\ 0 & 1 & \cdots & 0 & 0_{1 \times K} \\ & & & I_K & \\ \vdots & & & \vdots & \\ 0_{K \times 1} & 0_{K \times 1} & \cdots & 0_{K \times 1} & I_K \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_R \\ \gamma \end{bmatrix}. \quad (26)$$

Note that Eq.(25) can be used to specify the constrained parameter vector as a function of the unconstrained parameter vector: $\theta = (H'H)^{-1}H'\beta$. However, this cannot be used to find the constrained estimator from the unconstrained one, since this holds for the true parameters only. In other words, the specification of Eq.(26) holds only for the true parameters, under the hypothesis that the R constraints $\gamma_i = \gamma$, $i = 1, \dots, R$, are true. That is, even if these R constraints are true, the estimators $\hat{\gamma}_i$, $i = 1, \dots, R$ are, in general terms, different from each other, which may break the equality of Eq.(26). This implies that if the corresponding estimators are inserted into Eq.(25), the equality does not necessarily hold. This confirms the fact that the statistical properties of the unconstrained estimator cannot be automatically conveyed to the constrained estimator. Nevertheless, considering the estimates of β and θ , if the matrix H is now specified as follows:

$$\begin{bmatrix} \hat{\alpha}_1 \\ \hat{\gamma}_1 \\ \hat{\alpha}_2 \\ \hat{\gamma}_2 \\ \vdots \\ \hat{\alpha}_R \\ \hat{\gamma}_R \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0_{1 \times K} \\ 0_{K \times 1} & 0_{K \times 1} & \cdots & 0_{K \times 1} & q_1 I_K \\ 0 & 1 & \cdots & 0 & 0_{1 \times K} \\ & & & q_2 I_K & \\ \vdots & & & \vdots & \\ & & & 1 & 0_{1 \times K} \\ 0_{K \times 1} & 0_{K \times 1} & \cdots & 0_{K \times 1} & q_R I_K \end{bmatrix} \cdot \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_R \\ \hat{\gamma} \end{bmatrix} \quad (27)$$

then the (unknown) scalars q_r , $r = 1, \dots, R$, can be defined such that the constrained estimator:

$$\hat{\theta} = (H'H)^{-1}H'\hat{\beta} \quad (28)$$

²¹ For this approach, see Favero (2001, chapter 1) or Gourieuo and Monfort (1995, chapter 1). The advantage of this approach is that β can be simply substituted out from the data-generating process, which can be directly expressed as a function of $H\theta$.

exists and is at least consistent, efficient and asymptotic normally distributed. Now setting

$$q_r = \frac{\frac{p_r}{\sigma_r^2}}{\sum_{j=1}^R \left(\frac{p_j}{\sigma_j^2} \right)} r = 1, \dots, R, \quad (29)$$

with $p_r = \text{tr}(P_r)$, $r = 1, \dots, R$, and substituting Eq.(29) in the matrix H of Eq.(27), then the result in Eq.(28) yields:

$$\hat{\theta} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_R \\ \sum_{r=1}^R w_r \hat{\gamma}_r \end{bmatrix} \quad (30)$$

with:

$$w_r = \frac{\frac{p_r}{\sigma_r^2}}{\sum_{j=1}^R \left(\frac{p_j}{\sigma_j^2} \right)} r = 1, \dots, R, \quad (31)$$

which collapses to $w_r = p_r/M$ in the special cases of equal variances ($\sigma_r^2 = \sigma^2$, $r = 1, \dots, R$)²².

Now, further details are provided. In order to derive Eq.(12), note that:

$$\begin{aligned} H' &= \begin{bmatrix} 1 & 0_{1 \times K} & 0 & 0_{1 \times K} & 0 & \dots & 0_{1 \times K} \\ 0 & 0_{1 \times K} & 1 & 0_{1 \times K} & 0 & \dots & 0_{1 \times K} \\ 0 & 0_{1 \times K} & 0 & 0_{1 \times K} & 1 & \dots & 0_{1 \times K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0_{K \times 1} & q_1 I_K & 0_{K \times 1} & q_2 I_K & \dots & q_R I_K \end{bmatrix} & H'H &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0_{1 \times K} \\ 0 & 1 & 0 & \dots & 0_{1 \times K} \\ 0 & 0 & 1 & \dots & 0_{1 \times K} \\ \dots & \dots & \dots & \dots & \dots \\ 0_{K \times 1} & 0_{K \times 1} & 0_{K \times 1} & \dots & \sum_{r=1}^R q_r \cdot I_K \end{bmatrix} \\ (H'H)^{-1} &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0_{1 \times K} \\ 0 & 1 & 0 & \dots & 0_{1 \times K} \\ 0 & 0 & 1 & \dots & 0_{1 \times K} \\ \dots & \dots & \dots & \dots & \dots \\ 0_{K \times 1} & 0_{K \times 1} & 0_{K \times 1} & \dots & \frac{1}{\sum_{r=1}^R q_r} \cdot I_K \end{bmatrix} & (H'H)^{-1} H' &= \begin{bmatrix} 1 & 0_{1 \times K} & 0 & 0_{1 \times K} & 0 & \dots & 0_{1 \times K} \\ 0 & 0_{1 \times K} & 1 & 0_{1 \times K} & 0 & \dots & 0_{1 \times K} \\ 0 & 0_{1 \times K} & 0 & 0_{1 \times K} & 1 & \dots & 0_{1 \times K} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0_{K \times 1} & \frac{q_1}{\sum_{r=1}^R q_r} \cdot I_K & 0_{K \times 1} & \frac{q_2}{\sum_{r=1}^R q_r} \cdot I_K & \dots & \frac{q_R}{\sum_{r=1}^R q_r} \cdot I_K \end{bmatrix}. \end{aligned} \quad (32)$$

²² Note that $\frac{p_r}{M}$, with $\sum_{r=1}^R \frac{p_r}{M} = 1$, is an estimate of the ergodic probability of regime r occurring, if it exists; $r = 1, \dots, R$.

Now setting $q_r = \frac{\frac{p_r}{\sigma_r^2}}{\sum_{j=1}^R \frac{p_j}{\sigma_j^2}}$, $r = 1, \dots, R$, the weighting terms in Eq.(32) are $\frac{q_r}{\sum_{r=1}^R q_r} = \frac{\frac{p_r}{\sigma_r^2}}{\sum_{j=1}^R \frac{p_j}{\sigma_j^2}}$, $r = 1, \dots, R$. Thus, postmultiplying the matrix $(H'H)^{-1}H'$ by $\hat{\beta}$ yields the result of Eq.(30).

Note furthermore that, for $q_r = 1$, $r = 1, \dots, R$, the weighting terms in Eq.(32) are $\frac{q_r}{\sum_{r=1}^R q_r} = \frac{1}{R}$, $r = 1, \dots, R$, which implies $\hat{\gamma} = \sum_{r=1}^R \frac{1}{R} \hat{\gamma}_r$.

Deriving the expression of Eq. (15)

We first note that the specification of the matrix H through the choice of q_r , $r = 1, \dots, R$, does not affect the asymptotic properties of the above term.²³ In fact, the choice of whatever (nonzero, finite and real) number for q_r , $r = 1, \dots, R$, results in a different weighted average of the unconstrained estimators, $\hat{\gamma}_r$. For example, setting

$q_r = 1$, $r = 1, \dots, R$, the last element of $\hat{\theta}$ turns into a simple average of the unconstrained estimators: $\hat{\gamma} = \sum_{r=1}^R \frac{1}{R} \hat{\gamma}_r$, see Annex A.2.1 for details.

In order to find the consistency properties of the term $(H'X'XH)^{-1}(XH)'U$ of Eq.(15), we set $q_r = 1$, $r = 1, \dots, R$, bearing in mind that the result will be valid for whatever (nonzero, finite and real) value for q_r , $r = 1, \dots, R$, including that of Eq.(29).

For this purpose, we first find the quantity $(H'X'XH)^{-1}(XH)'$. We start by finding XH , where X is defined in Eq.(3):

$$XH = \left[S_1 \cdot \begin{bmatrix} i_M & X_1 \end{bmatrix} \dots S_R \cdot \begin{bmatrix} i_M & X_1 \end{bmatrix} \right] \begin{bmatrix} 1 & 0 & \dots & 0 & 0_{1 \times K} \\ 0_{K \times 1} & 0_{K \times 1} & \dots & 0_{K \times 1} & I_K \\ 0 & 1 & \dots & 0 & 0_{1 \times K} \\ & & & & I_K \\ \vdots & & \ddots & & \vdots \\ & & & 1 & 0_{1 \times K} \\ 0_{K \times 1} & 0_{K \times 1} & \dots & 0_{K \times 1} & I_K \end{bmatrix}$$

$$= [[S_1 \cdot i_M] [S_2 \cdot i_M] \dots [S_R \cdot i_M] [S_1 X_1 + S_2 X_1 + \dots + S_R X_1]].$$

Note that we can now exploit the fact that: $S_1 X_1 + S_2 X_1 + \dots + S_R X_1 = X_1$, so that the $(R + K) \times (R + K)$ matrix $H'X'XH$ becomes:

²³ However, the choice of q_r , $r = 1, \dots, R$ does affect the efficiency of the estimator, see Theorem 2.

$$\begin{aligned}
& \begin{bmatrix} i'_M S_1 \\ i'_M S_2 \\ \vdots \\ i'_M S_R \\ X'_1 \end{bmatrix} \cdot [[S_1 \cdot i_M][S_2 \cdot i_M] \cdots [S_R \cdot i_M][X_1]] \\
&= \begin{bmatrix} i'_M S_1 i_M & \cdots & i'_M S_1 X_1 \\ & i'_M S_2 i_M & i'_M S_2 X_1 \\ \vdots & \ddots & \vdots \\ & & i'_M S_R i_M & i'_M S_R X_1 \\ X'_1 S_1 i_M & X'_1 S_2 i_M & \cdots & X'_1 S_R i_M & X'_1 X_1 \end{bmatrix}.
\end{aligned}$$

Note that $i'_M S_r i_M = t_r (= tr(S_r))$ is a scalar and equals the number of observations belonging to regime r , $r = 1, \dots, R$. Thus, we now have to find the inverse matrix of $H'X'XH$:

$$(H'X'XH)^{-1} = \left(\begin{bmatrix} t_1 & \cdots & i'_M S_1 X_1 \\ & t_2 & i'_M S_2 X_1 \\ \vdots & \ddots & \vdots \\ & & t_R & i'_M S_R X_1 \\ X'_1 S_1 i_M & X'_1 S_2 i_M & \cdots & X'_1 S_R i_M & X'_1 X_1 \end{bmatrix} \right)^{-1}.$$

Note now that the above matrix can be inverted by using the formula of partitioned matrices (see, for example, Greene (2003, Appendix A)), where the upper-left block is a diagonal matrix of order R , for which the elements of the main diagonal are the inverse values of t_r , $r = 1, \dots, R$. The other partitions are defined conformably. Let us define the partitions of the above matrix as follows:

$$(H'X'XH)^{-1}_{(R+K) \times (R+K)} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

then, the inverse of the lower-right block is:

$$\begin{aligned}
D_{K \times K} &= (X'_1 X_1 - [X'_1 S_1 i_M \ X'_1 S_2 i_M \ \cdots \ X'_1 S_R i_M] \\
&\quad \cdot \begin{bmatrix} t_1^{-1} & 0 & \cdots & 0 \\ 0 & t_2^{-1} & & \\ \vdots & & \ddots & \\ 0 & & & t_R^{-1} \end{bmatrix} \cdot \begin{bmatrix} i'_M S_1 X_1 \\ i'_M S_2 X_1 \\ \vdots \\ i'_M S_R X_1 \end{bmatrix})^{-1} \\
&= (X'_1 [I_M - [S_1 i_M \ S_2 i_M \ \cdots \ S_R i_M] \\
&\quad \cdot \begin{bmatrix} t_1^{-1} & 0 & \cdots & 0 \\ 0 & t_2^{-1} & & \\ \vdots & & \ddots & \\ 0 & & & t_R^{-1} \end{bmatrix} \cdot \begin{bmatrix} i'_M S_1 \\ i'_M S_2 \\ \vdots \\ i'_M S_R \end{bmatrix}] X_1)^{-1} \\
&= (X'_1 [I_M - [t_1^{-1} S_1 i_M i'_M S_1 + t_2^{-1} S_2 i_M i'_M S_2 + \cdots + t_R^{-1} S_R i_M i'_M S_R]] X_1)^{-1}.
\end{aligned} \tag{33}$$

Note that the matrix $[t_1^{-1}S_1i_Mi_M'S_1 + t_2^{-1}S_2i_Mi_M'S_2 + \dots + t_R^{-1}S_Ri_Mi_M'S_R]$ is a $(M \times M)$ matrix, such that each element of the t -th row is the scalar t_r^{-1} if regime r occurs for the t -th observation (and zero otherwise). Thus, define

$$H^0 = I_M - [t_1^{-1}S_1i_Mi_M'S_1 + t_2^{-1}S_2i_Mi_M'S_2 + \dots + t_R^{-1}S_Ri_Mi_M'S_R] \quad (34)$$

which implies,

$$D_{K \times K} = (X_1' H^0 X_1)^{-1} \quad (35)$$

Furthermore:

$$\begin{aligned} A_{R \times R} &= \begin{bmatrix} t_1^{-1} & & \dots \\ & t_2^{-1} & \\ \vdots & & \ddots \\ & & & t_R^{-1} \end{bmatrix} \\ &\quad \cdot \left(I_R + \begin{bmatrix} i_M' S_1 X_1 \\ i_M' S_2 X_1 \\ \vdots \\ i_M' S_R X_1 \end{bmatrix} \cdot (X_1' H^0 X_1)^{-1} \cdot [X_1' S_1 i_M \quad X_1' S_2 i_M \quad \dots \quad X_1' S_R i_M] \right. \\ &\quad \left. \cdot \begin{bmatrix} t_1^{-1} & & \dots \\ & t_2^{-1} & \\ \vdots & & \ddots \\ & & & t_R^{-1} \end{bmatrix} \right) \\ B_{R \times K} &= - \begin{bmatrix} t_1^{-1} & & \dots \\ & t_2^{-1} & \\ \vdots & & \ddots \\ & & & t_R^{-1} \end{bmatrix} \begin{bmatrix} i_M' S_1 X_1 \\ i_M' S_2 X_1 \\ \vdots \\ i_M' S_R X_1 \end{bmatrix} (X_1' H^0 X_1)^{-1} \\ C_{K \times R} &= -(X_1' H^0 X_1)^{-1} [X_1' S_1 i_M \quad X_1' S_2 i_M \quad \dots \quad X_1' S_R i_M] \begin{bmatrix} t_1^{-1} & & \dots \\ & t_2^{-1} & \\ \vdots & & \ddots \\ & & & t_R^{-1} \end{bmatrix} \end{aligned}$$

Note now that:

$$(XH)'_{(R+K) \times 1} = \begin{bmatrix} i_M' S_1 \\ i_M' S_2 \\ \vdots \\ i_M' S_R \\ X_1' \end{bmatrix}.$$

Thus, all factors for the computation of the matrix $(H'X'XH)^{-1}(XH)'$ are available, yielding:

$$\begin{aligned}
(H'X'XH)^{-1}(XH)' &= \begin{bmatrix} A^1 + B^1 \\ C^1 + D^1 \end{bmatrix} \\
A_{R \times M}^1 &= \begin{bmatrix} t_1^{-1} i_M' S_1 \\ t_2^{-1} i_M' S_2 \\ \vdots \\ t_R^{-1} i_M' S_R \end{bmatrix} + \begin{bmatrix} t_1^{-1} i_M' S_1 X_1 \\ t_2^{-1} i_M' S_2 X_1 \\ \vdots \\ t_R^{-1} i_M' S_R X_1 \end{bmatrix} (X_1' H^0 X_1)^{-1} \\
&\quad \cdot \begin{bmatrix} X_1' S_1 i_M & X_1' S_2 i_M & \cdots & X_1' S_R i_M \end{bmatrix} \begin{bmatrix} t_1^{-1} i_M' S_1 \\ t_2^{-1} i_M' S_2 \\ \vdots \\ t_R^{-1} i_M' S_R \end{bmatrix} \\
B_{R \times M}^1 &= - \begin{bmatrix} t_1^{-1} i_M' S_1 X_1 \\ t_2^{-1} i_M' S_2 X_1 \\ \vdots \\ t_R^{-1} i_M' S_R X_1 \end{bmatrix} (X_1' H^0 X_1)^{-1} X_1' \\
C_{K \times M}^1 &= - (X_1' H^0 X_1)^{-1} \begin{bmatrix} X_1' S_1 i_M & X_1' S_2 i_M & \cdots & X_1' S_R i_M \end{bmatrix} \begin{bmatrix} t_1^{-1} i_M' S_1 \\ t_2^{-1} i_M' S_2 \\ \vdots \\ t_R^{-1} i_M' S_R \end{bmatrix} \\
D_{K \times M}^1 &= (X_1' H^0 X_1)^{-1} X_1'.
\end{aligned}$$

Having now found all elements of the matrix $(H'X'XH)^{-1}(XH)'$, we are in a position to evaluate: $(H'X'XH)^{-1}(XH)'U$ of Eq.(15). Noting that:

$$\begin{aligned}
&\begin{bmatrix} X_1' S_1 i_M & X_1' S_2 i_M & \cdots & X_1' S_R i_M \end{bmatrix} \begin{bmatrix} t_1^{-1} i_M' S_1 \\ t_2^{-1} i_M' S_2 \\ \vdots \\ t_R^{-1} i_M' S_R \end{bmatrix} \\
&= X_1' [t_1^{-1} S_1 i_M i_M' S_1 + t_2^{-1} S_2 i_M i_M' S_2 + \cdots + t_R^{-1} S_R i_M i_M' S_R] = X_1' [I_M - H^0]
\end{aligned}$$

It follows that:

$$\begin{aligned}
&(H'X'XH)^{-1}(XH)'U \\
&= \begin{bmatrix} t_1^{-1} i_M' S_1 \\ t_2^{-1} i_M' S_2 \\ \vdots \\ t_R^{-1} i_M' S_R \\ 0_{K \times M} \end{bmatrix} U - \begin{bmatrix} t_1^{-1} i_M' S_1 X_1 \\ t_2^{-1} i_M' S_2 X_1 \\ \vdots \\ t_R^{-1} i_M' S_R X_1 \\ I_K \end{bmatrix} (X_1' H^0 X_1)^{-1} X_1' H^0 U \quad (36)
\end{aligned}$$

Deriving the proof of Theorem 1

Note that Eq.(5) and Eq.(6) provide sufficient conditions such that the terms $t_r^{-1}i'_M S_r U$ ($r = 1, \dots, R$), are asymptotically normally distributed with zero mean and finite variance²⁴:

$$\frac{i'_M S_r U}{t_r^{1/2}} \xrightarrow{d} N(0, \sigma_r^2) \quad r = 1, \dots, R. \quad (37)$$

which, in turn, implies²⁵:

$$plim(t_r^{-1}i'_M S_r U) = 0 \quad r = 1, \dots, R. \quad (38)$$

Let us now evaluate the term $(X'_1 H^0 X_1)^{-1} X'_1 H^0 U$ of Eq.(36). First note that the convergence of $X'_1 H^0 X_1$ (to a finite, non-random matrix) has to be verified²⁶. By definition of H^0 , see Eq. (34), the relevant term to be evaluated is $X'_1(t_r^{-1}S_r i'_M i'_M S_r)X_1$. It is easy to see that its convergence is ensured by the assumption of Eq.(11)²⁷, and hence:

$$plim_{M \rightarrow \infty} \left(\frac{X'_1 H^0 X_1}{M} \right) = Q_{H^0} \quad (39)$$

where Q_{H^0} is a matrix of non-random and finite elements. As noted, $X'_1 H^0 U$ computes the sum of deviations from the means of the columns of X_1 , within each regime, multiplied by the error term U . Thus, by virtue of Eq.(7) and of Eq.(39) and since $H^0 = H^0 H^0$ we have:

$$\frac{X'_1 H^0 U}{M^{1/2}} \xrightarrow{d} N(0, V) \quad (40)$$

where $V = plim_{M \rightarrow \infty} \frac{H^0 X'_1 \Sigma X_1 H^0}{M}$ and $\Sigma = diag(\sigma_m^2) \equiv \sum_{r=1}^R \sigma_r^2 S_r$, that is a diagonal matrix whose entries are the conditional variances.

²⁴ Where $\xrightarrow{d}$ denotes the limit in distribution operator.

²⁵ As the number of observations within regime r goes to infinity ($t_r \rightarrow \infty$), the average computed through the factor $t_r^{-1}i'_M$ of the error term in regime r , $S_r U$ goes to zero.

²⁶ Note that H^0 is an idempotent matrix, implying that $H^0 X_1$ computes deviations of the observations in X_1 with respect to the relevant mean values within regimes; thus $X'_1 H^0 X_1$ computes sums of squares and cross products of deviations from the respective R column means. The matrix H^0 generalizes the matrix M_0 in Greene (2003, Appendix A.2.8), who considers the non-regime-switching case.

²⁷ The matrix $X'_1 S_r X_1$ in Eq.(11) computes the sums of squared values of the columns of X_1 ; meanwhile, the matrix $X'_1 (S_r i'_M i'_M S_r) X_1$ computes the squared values of the sum of the columns of X_1 , both using the observations belonging to regime r . Furthermore, note that, due to the structure of H^0 , we also require:

$plim_{M \rightarrow \infty} \left(\frac{X'_1 X_1}{M} \right) = Q_1$ where Q_1 is a $(K \times K)$ non-stochastic matrix of finite elements; this, however, is implied by Eq.(11).

Equation (40) provides a sufficient condition for the convergence in probability to zero. Thus,

$$plim_{M \rightarrow \infty} (X_1' H^0 X_1)^{-1} X_1' H^0 U = 0 \quad (41)$$

At this juncture, by combining the results of Eq.(10), Eq.(38) and of Eq.(41) and by virtue of the Slutsky theorem, for Eq.(15) we have:

$$plim_{M \rightarrow \infty} (\hat{\theta} - \theta) = 0 \quad (42)$$

which establishes consistency.

The issue of the asymptotic normality is now addressed. Eq.(39) together with Eq.(40) implies:

$$M^{1/2} (X_1' H^0 X_1)^{-1} X_1' H^0 U \xrightarrow{d} N(0, Q_{H^0}^{-1} V Q_{H^0}^{-1}) \quad (43)$$

Thus, by virtue of Eq.(10), Eq.(37) and Eq.(43), the constrained estimator $\hat{\theta}$ is asymptotically normally distributed:

$$\frac{\hat{\theta} - \theta}{M^{1/2}} \xrightarrow{d} N(0, Q^{-1}) \quad (44)$$

where Q is a $((R + K) \times (R + K))$ matrix of non-stochastic and finite elements, i.e., a function of v_r, σ_r^2 ($r = 1, \dots, R$), V and Q_{H^0} .

The above results are derived conditional on knowledge of unobservable quantities: $S_r, r = 1, \dots, R$. Thus, based on the result of Theorem 3, implying the property of Eq.(19), we can replace (in the above results) S_r by its (smoothed) estimating counterpart, $P_r, r = 1, \dots, R$, since Eq.(42) and Eq.(44) are linked to the matrices $S_r, r = 1, \dots, R$, through continuous functions (Slutsky theorem).

Second appendix

This appendix provides the proof of Theorem 2.

Setup of the proof of Theorem 2

Given Theorem 1, the scaled (asymptotic) covariance matrix of the constrained estimator, as a function of a generic weight, $w_i, i = 1, \dots, R$, is,²⁸:

²⁸ Throughout this subsection, for ease of exposition, we omit the conditions of the expectation and of the variance operators, i.e., $Var[\cdot] = Var[\cdot | X_1 S_R, \dots S_1]$.

$$\begin{aligned} \text{Var} \left[\sum_{i=1}^R w_i \hat{\gamma}_i \right] &= E \left[\left(\sum_{i=1}^R w_i \hat{\gamma}_i - \gamma \right) \left(\sum_{i=1}^R w_i \hat{\gamma}_i - \gamma \right)' \right] \\ &= E \left[\left(\sum_{i=1}^R w_i \hat{\gamma}_i \right) \left(\sum_{i=1}^R w_i \hat{\gamma}_i \right)' \right] - \gamma \gamma'. \end{aligned}$$

Thus, w_i , $i = 1, \dots, R$ must be chosen such that the above variance attains the Cramer–Rao lower bound. To this end, further define the parameter vector β of Eq.(4) as follows:

$$\beta = [\alpha_1 \ \gamma'_1 \ \alpha_2 \ \gamma'_2 \ \dots \ \alpha_R \ \gamma'_R]' = [\beta'_1 \ \beta'_2 \ \dots \ \beta'_R]' \quad (45)$$

such that $[\alpha_i \ \gamma'_i]' = \beta'_i$, $i = 1, \dots, R$. Thus, the regime-switching estimator of $\hat{\beta}_i$, $i = 1, \dots, R$, is:

$$\hat{\beta}_i = (X'_2 P_i X_2)^{-1} (X'_2 P_i Y) \quad \text{with} \quad (46)$$

$$\text{Var}[\hat{\beta}_i] = \sigma_i^2 (X'_2 P_i X_2)^{-1} \quad \text{where} \quad (47)$$

$$X_2 = [i_M \ X_1] \quad \text{and} \quad (48)$$

$$Y = S_1 X_2 \beta_1 + S_2 X_2 \beta_2 + \dots + S_R X_2 \beta_R + U. \quad (49)$$

see Beccarini (2019) for the details. Now define,

$$I_{(K+1)}^0 = \begin{bmatrix} 0 & 0 & \dots & \dots & 0 \\ 0 & 1 & & & \\ \vdots & & 1 & & \\ \vdots & & & \ddots & \\ 0 & & & & 1 \end{bmatrix}_{(K+1)^2}$$

which implies that $\hat{\gamma}_i = I_{(K+1)}^0 \hat{\beta}_i$ and

$$\text{Var} \left[\sum_{i=1}^R w_i \hat{\gamma}_i \right] = I_{(K+1)}^0 \text{Var} \left[\sum_{i=1}^R w_i \hat{\beta}_i \right] I_{(K+1)}^0. \quad (50)$$

Since it is easier to deal with $\text{Var} \left[\sum_{i=1}^R w_i \hat{\beta}_i \right]$ instead of $\text{Var} \left[\sum_{i=1}^R w_i \hat{\gamma}_i \right]$, we first show the following:

$$\text{Var} \left[\sum_{i=1}^R w_i \hat{\beta}_i \right] = \sum_{i=1}^R w_i^2 \sigma_i^2 (X'_2 P_i X_2)^{-1}. \quad (51)$$

We then find the optimal weights w_i , $i = 1, \dots, R$, such that the above covariance matrix attains the Cramer-Rao lower bound. This will obviously also hold for $\text{Var}\left[\sum_{i=1}^R w_i \hat{\gamma}_i\right]$.

Asymptotic covariance matrix of the constrained estimator

We now show that Eq.(51) holds. Note first that if $\text{plim}(\hat{\beta}_i) = \beta_i$, which is a consequence of Theorem 1, then,²⁹

$$\begin{aligned} \text{Var}\left[\sum_{i=1}^R w_i \hat{\beta}_i\right] &= E\left[\left(\sum_{i=1}^R w_i \hat{\beta}_i - \sum_{i=1}^R w_i \beta_i\right)\left(\sum_{i=1}^R w_i \hat{\beta}_i - \sum_{i=1}^R w_i \beta_i\right)'\right] \\ &= E\left[\left(\sum_{i=1}^R w_i \hat{\beta}_i\right)\left(\sum_{i=1}^R w_i \hat{\beta}_i\right)'\right] - \sum_{i=1}^R w_i \beta_i \left(\sum_{i=1}^R w_i \beta_i\right)'. \end{aligned} \quad (52)$$

Thus, since

$$\begin{aligned} \sum_{i=1}^R w_i \hat{\beta}_i &= \sum_{i=1}^R w_i (X_2' P_i X_2)^{-1} (X_2' P_i Y) \\ \left(\sum_{i=1}^R w_i \hat{\beta}_i\right)' &= \sum_{i=1}^R w_i (X_2' P_i Y)' (X_2' P_i X_2)^{-1}, \end{aligned}$$

then

$$\begin{aligned} \left(\sum_{i=1}^R w_i \hat{\beta}_i\right)\left(\sum_{i=1}^R w_i \hat{\beta}_i\right)' &= \left[w_1 (X_2' P_1 X_2)^{-1} \dots w_R (X_2' P_R X_2)^{-1}\right] \\ &\quad \cdot \begin{bmatrix} X_2' P_1 Y \\ \vdots \\ X_2' P_R Y \end{bmatrix} \left[(X_2' P_1 Y)' \dots (X_2' P_R Y)'\right] \begin{bmatrix} w_1 (X_2' P_1 X_2)^{-1} \\ \vdots \\ w_R (X_2' P_R X_2)^{-1} \end{bmatrix}. \end{aligned} \quad (53)$$

Now first consider the product of the second and third vectors of the above expression, for which the typical element is:

$$(X_2' P_i Y)(X_2' P_j Y)', i, j = 1, \dots, R.$$

Note now that substituting out Y (in line with Eq.(2) and Eq.(45)) yields:

²⁹ Note that, if the constraint over the $\hat{\gamma}_i$, $i = 1, \dots, R$, is true, meaning that there is a unique parameter γ , the parameters β_i are still regime-dependent, because of its first element, the constant term: $[\alpha_i \ \gamma_i']' = \beta_i'$, $i = 1, \dots, R$.

$$X_2' P_i Y = X_2' P_i [S_1 X_2 \beta_1 + S_2 X_2 \beta_2 + \dots + S_R X_2 \beta_R + U],$$

assuming Eq.(19) implies that $P_i S_i \approx P_i P_i \approx P_i$ and $P_i S_j \approx 0_M$ if $i \neq j$ ($i, j = 1, \dots, R$), thus:

$$X_2' P_i Y = X_2' P_i [P_i X_2 \beta_i + U] = (X_2' P_i X_2) \beta_i + X_2' P_i U.$$

This result implies:

$$\begin{aligned} (X_2' P_i Y)(X_2' P_j Y)' &= (X_2' P_i X_2) \beta_i \beta_j' (X_2' P_j X_2)' + X_2' P_i U \beta_j' (X_2' P_j X_2)' \\ &\quad + (X_2' P_i X_2) \beta_i (X_2' P_j U)' + X_2' P_i U U' P_j X_2 \end{aligned}$$

which, due to Eq.(6), it holds: $E[UU'] = \sum_{r=1}^R \sigma_r^2 S_r$ and hence:

$$E(X_2' P_i Y)(X_2' P_j Y)' = (X_2' P_i X_2) \beta_i \beta_j' (X_2' P_j X_2)' + X_2' P_i \sum_{r=1}^R \sigma_r^2 S_r P_j X_2.$$

Note also that if Eq.(19) is assumed, then $P_i P_j \approx 0$, $S_i P_j \approx 0$, and the latter addend of the above equation is zero for $i \neq j$, $i, j = 1, \dots, R$. Thus:

$$\begin{aligned} E \begin{bmatrix} X_2' P_1 Y \\ \vdots \\ X_2' P_R Y \end{bmatrix} \begin{bmatrix} (X_2' P_1 Y)' \dots (X_2' P_R Y)' \end{bmatrix} \\ = \begin{bmatrix} (X_2' P_1 X_2) \beta_1 \beta_1' (X_2' P_1 X_2)' & \dots & (X_2' P_1 X_2) \beta_1 \beta_R' (X_2' P_R X_2)' \\ \vdots & & \vdots \\ (X_2' P_R X_2) \beta_R \beta_1' (X_2' P_1 X_2)' & \dots & (X_2' P_R X_2) \beta_R \beta_R' (X_2' P_R X_2)' \end{bmatrix} \\ + \begin{bmatrix} \sigma_1^2 (X_2' P_1 X_2) & 0 & \dots & 0 \\ 0 & \sigma_2^2 (X_2' P_2 X_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & & \sigma_R^2 (X_2' P_R X_2) \end{bmatrix}. \end{aligned}$$

Pre- and post-multiplying the latter results by the first and the fourth term of Eq.(53), and simplifying yields:

$$\begin{aligned} E \left(\sum_{i=1}^R w_i \hat{\beta}_i \right) \left(\sum_{i=1}^R w_i \hat{\beta}_i \right)' &= \sum_{i=1}^R w_i \beta_i \left(\sum_{i=1}^R w_i \beta_i \right)' \\ &\quad + \left[w_1 \sigma_1^2 (X_2' P_1 X_2)^{-1} \dots w_R \sigma_R^2 (X_2' P_R X_2)^{-1} \right] \begin{bmatrix} w_1 \\ \vdots \\ w_R \end{bmatrix} \end{aligned}$$

which gives the result:

$$Var \left[\sum_{i=1}^R w_i \hat{\beta}_i \right] = \sum_{i=1}^R w_i^2 \sigma_i^2 (X_2' P_i X_2)^{-1}. \quad (54)$$

The optimal weights

In order to choose the optimal weights (w_i , $i = 1, \dots, R$), we first have to perform a minimization problem which, in general terms, requires evaluating a covariance matrix, namely that of Eq.(54). Nevertheless, we apply the standard (constrained) optimization procedure at each element of the above covariance matrix. Provided that we show that this solution does not depend on the particular element we have considered, the solution applies to every element of the matrix in question. To characterize these elements, consider one generic row and one generic column of the matrix X_2 namely x'_m and x_n , respectively. Thus, the minimization problem becomes:

$$L_{m,n} = \sum_{i=1}^R w_i^2 \sigma_i^2 (x'_m P_i x_n)^{-1} - \lambda \left(\sum_{i=1}^R w_i - 1 \right) - \sum_{i=1}^R \mu_i^0 w_i, m, n = 1, \dots, K, \quad (55)$$

where λ and μ_i^0 are the Lagrange and the Kuhn–Tucker multipliers which allow taking into account, in the minimization problem, that for each w_i , the following constraints, $\sum_{i=1}^R w_i = 1$ and $0 \leq w_i \leq 1$, hold³⁰. We then argue that, if the solution is independent of the chosen row and column (x'_m and x_n , $m, n = 1, \dots, K$), it is then the solution, such that the corresponding covariance matrix attains the Cramer–Rao lower bound. Then, the derivative of the (K^2) Lagrangians with respect to each w_i is:

$$\frac{\partial L_{m,n}}{\partial w_i} = 2\sigma_i^2 w_i (x'_m P_i x_n)^{-1} - \lambda - \mu_i^0, i = 1, \dots, R, \quad m, n = 1, \dots, K.$$

Setting the results equal to zero yields:

$$w_i = \frac{(x'_m P_i x_n)}{2\sigma_i^2} (\lambda + \mu_i^0), i = 1, \dots, R, \quad m, n = 1, \dots, K. \quad (56)$$

Note now that since $\lambda + \mu_i^0 > 0$, $i = 1, \dots, R$, because $\lambda > 0$, and $\sigma_i^2 > 0$, then w_i can only be zero if and only if each element of P_i is zero, meaning that no observation belongs to regime i . This was excluded in our main setup. Thus, this implies that the constraint $0 \leq w_i \leq 1$ is binding: $0 < w_i < 1$, and hence it must be that $\mu_i^0 = 0$. Exploiting this latter result and summing both sides of the above equation yields:

$$\sum_{r=1}^R w_r = 1 = \sum_{r=1}^R \frac{x'_m P_r x_n}{2\sigma_r^2} \lambda, m, n = 1, \dots, K.$$

Thus,

³⁰ Note that, for a given w_i , the constraint $0 \leq w_i \leq 1$ may be expressed in the Lagrangian as $\mu_i^0(0 - w_i) + \mu_i^1(1 - w_i)$. However, the second constraint is redundant because of the first equality constraint: $\sum_{i=1}^R w_i = 1$. Furthermore, it holds: $\mu_i^0 \geq 0$ and $\mu_i^1 = 0$ if $w_i > 0$.

$$\lambda = \left(\sum_{r=1}^R \frac{(x'_m P_r x_n)}{2\sigma_r^2} \right)^{-1} \quad m, n = 1, \dots, K.$$

Substituting the value of λ in the first-order condition of Eq.(56) yields,

$$2\sigma_i^2 (x'_m P_i x_n)^{-1} w_i = \left(\sum_{r=1}^R \frac{(x'_m P_r x_n)}{2\sigma_r^2} \right)^{-1} \quad i = 1, \dots, R, \quad m, n = 1, \dots, K,$$

or

$$\begin{aligned} w_i &= \frac{x'_m P_i x_n}{\sigma_i^2} \left(\sum_{r=1}^R \frac{(x'_m P_r x_n)}{\sigma_r^2} \right)^{-1} \quad i = 1, \dots, R, m, n = 1, \dots, K \\ x'_m \left(\sum_{r=1}^R \frac{w_i P_r}{\sigma_r^2} \right) x_n &= \frac{x'_m P_i x_n}{\sigma_i^2} \quad i = 1, \dots, R, m, n = 1, \dots, K \\ x'_m \left(\sum_{r=1}^R w_i \frac{P_r}{\sigma_r^2} - \frac{P_i}{\sigma_i^2} \right) x_n &= 0 \quad i = 1, \dots, R, m, n = 1, \dots, K \end{aligned}$$

having assumed nonzero values for x'_m and x_n it must be:

$$\begin{aligned} \sum_{r=1}^R w_i \frac{P_r}{\sigma_r^2} - \frac{P_i}{\sigma_i^2} &= 0 \quad i = 1, \dots, R, m, n = 1, \dots, K, \\ \text{tr} \left(\sum_{r=1}^R w_i \frac{P_r}{\sigma_r^2} - \frac{P_i}{\sigma_i^2} \right) &= 0 \quad i = 1, \dots, R, m, n = 1, \dots, K, \end{aligned}$$

which yields the result³¹:

$$w_i = \frac{\text{tr}(P_i/\sigma_i^2)}{\sum_{r=1}^R \text{tr}(P_r/\sigma_r^2)} \quad i = 1, \dots, R. \quad (57)$$

Since this solution does not depend on the chosen row and column of X_2 , we conclude that this value, cast in Eq.(54), renders the covariance matrix as the Cramer–Rao lower bound, i.e., for other possible R weights cast in Eq.(54), the quadratic form of the following difference

$$\sigma^2 \sum_{i=1}^R \left[w_i^2 - \left(\frac{\text{tr}(P_i/\sigma_i^2)}{\sum_{r=1}^R \text{tr}(P_r/\sigma_r^2)} \right)^2 \right] (X'_2 P_i X_2)^{-1} \quad (58)$$

³¹ Note that the following necessary condition for a minimum-solution is also sufficient because of the convexity of the function in question.

is positive definite for $w_i \neq \frac{\text{tr}(P_i/\sigma_i^2)}{\sum_{r=1}^R \text{tr}(P_r/\sigma_r^2)}, i = 1, \dots, R$.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–128 (1999)
- Aitkin, M., Wilson, G.T.: Mixture models, outliers, and the em algorithm. *Techonmetrics* **22**, 326–331 (1980)
- Alvarez, R., Camacho, M., Ruiz, M.: Inference on filtered and smoothed probabilities in Markov-switching autoregressive models. *J. Bus. Econ. Stat.* **37**, 484–495 (2019)
- Bartolucci, F., Farcomeni, A.: A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *J. Am. Stat. Assoc.* **104**, 816–831 (2009)
- Beccarini, A.: Testing for the omission of relevant variables and regime-switching misspecification. *Empr. Econ.* **56**, 775–796 (2019)
- Beccarini, A., Kempa, B.: Modelling time-varying heterogeneity in panel data as regime-switching. *Ann. Econ. Stat.* **151**, 81–120 (2023)
- Bianchi, F.: Regime switches, agents' beliefs, and post-world war ii U.S. macroeconomic dynamics. *Rev. Econ. Stud.* **80**(2), 463–490 (2013)
- Billion, M., Caporin, M.: Multivariate Markov switching dynamic conditional correlation Garch representations for contagion analysis. *Stat. Methods Appl.* **14**, 145–161 (2005)
- Cavicchioli, M.: Determining the number of regimes in Markov switching Var and VMA models. *J. Time Ser. Anal.* **35**, 173–186 (2014)
- Cavicchioli, M.: Analysis of the likelihood function for Markov-switching var(ch) models. *J. Time Ser. Anal.* **35**, 624–639 (2014)
- Cavicchioli, M.: Markov switching Garch models: higher order moments, kurtosis measures, and volatility evaluation in recessions and pandemic. *J. Bus. Econ. Stat.* **40**, 1772–1783 (2022)
- Chang, Y., Maih, J., Tan, F.: Origins of monetary policy shifts: a new approach to regime switching in DSGE models. *J. Econ. Dyn. Control* **133**, 1042–42 (2021)
- Cho, S.: Determinacy and classification of markov-switching rational expectations models. *J. Econ. Dyn. Control* **127**, 104–124 (2021)
- Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Series B* **39**, 1–38 (1977)
- Di Sanzo, S.: Testing for linearity in markov switching models: a bootstrap approach. *Stat. Methods Appl.* **18**, 153–168 (2009)
- Engle, R.F., Hendry, D.F., Richard, J.F.: Exogeneity. *Econometrica* **51**, 277–304 (1983)
- Farmer, R.E., Waggoner, D.F., Zha, T.: Minimal state variable solutions to markov-switching rational expectations models. *J. Econ. Dyn. Control* **35**(12), 2150–2166 (2011)
- Favero, C.A.: *Applied Macroeconometrics*. Oxford University Press, Oxford (2001)
- Franco, C., Zakoian, J.M.: Stationarity of multivariate markov switching arma models. *J. Econometr.* **102**, 339–364 (2001)
- Frühwirth-Schnatter, S.: *Finite Mixture and Markov Switching Models*. Springer Verlag, Newyork (2006)
- Gourieuo, C., Monfort, A.: *Statistics and Econometric Models, General Concepts, Estimation, Prediction, and Algorithms*. Cambridge University Press, Cambridge (1995)
- Greene, W.H.: *Econometric Analysis*. Prentice Hall, Hoboken (2003)

- Hamilton, J.D.: A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384 (1989)
- Hamilton, J.D.: Analysis of time series subject to changes in regime. *J. Econometr.* **45**, 39–70 (1990)
- Hamilton, J.D.: *Time Series Analysis*. Princeton University Press, Hoboken (1994)
- Hamilton, J.D., Lin, G.: Stock market volatility and the business cycle. *J. Appl. Economet.* **11**, 573–593 (1996)
- Hamilton, J.D., Susmel, R.: Autoregressive conditional heteroskedasticity and changes in regimes. *J. Econometr.* **64**, 307–333 (1994)
- Hansen, L.P.: Large sample properties of generalized method of moments estimators. *Econometrica* **50**(4), 1029–1054 (1982)
- Hur, J.M.: Monetary policy and asset prices: a Markov-switching DSGE approach. *J. Appl. Economet.* **32**, 965–982 (2017)
- Kim, C.J.: Dynamic linear models with Markov-switching. *J. Econometr.* **60**, 1–22 (1994)
- Kim, C.J., Nelson, C.R.: *State-Space Models with Regime Switching*. MIT, Cambridge (1999)
- Kim, C.J., Nelson, C.R., Startz, R.: Testing for mean reversion in heteroskedastic data based on Gibbs-sampling-augmented randomization. *J. Empir. Financ.* **5**, 131–154 (1998)
- Krolzig, H.M.: *Markov-Switching Vector Autoregressions*. Springer Verlag, Newyork (1997)
- Leroux, B.G.: Maximum-likelihood estimation for hidden Markov models. *Stoch. Process. Appl.* **40**, 127–143 (1992)
- Leroux, B.G., Puterman, M.L.: Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545–58 (1992)
- Lucas, R.E.: Econometric policy evaluation: a critique. *Carn.-Roch. Conf. Ser. Public Policy* **1**, 19–46 (1976)
- Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195–238 (1984)
- Sargent, T.J.: Interpreting economic time series. *J. Polit. Econ.* **89**, 213–248 (1981)
- White, H.: *Asymptotic Theory for Econometricians*. Academic Press Inc., Cambridge (1991)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.