Goodness-of-fit testing for the marginal distribution of regime-switching models with an application to electricity spot prices

Joanna Janczura · Rafał Weron

Received: 2 February 2012 / Accepted: 27 September 2012 / Published online: 18 October 2012
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract This paper complements a recently published study (Janczura and Weron in AStA-Adv Stat Anal 96(3):385–407, 2012) on efficient estimation of Markov regime-switching models. Here, we propose a new goodness-of-fit testing scheme for the marginal distribution of such models. We consider models with an observable (like threshold autoregressions) as well as a latent state process (like Markov regime-switching). The test is based on the Kolmogorov–Smirnov supremum-distance statistic and the concept of the weighted empirical distribution function. The motivation for this research comes from a recent stream of literature in energy economics concerning electricity spot price models. While the existence of distinct regimes in such data is generally unquestionable (due to the supply stack structure), the actual goodness-of-fit of the models requires statistical validation. We illustrate the proposed scheme by testing whether commonly used Markov regime-switching models fit deseasonalized electricity prices from the NEPOOL (US) day-ahead market.

Keywords Regime-switching · Energy economics · Goodness-of-fit · Weighted empirical distribution function · Kolmogorov–Smirnov test
1 Introduction

Regime-switching models have attracted a lot of attention in the recent years. A flexible specification allowing for abrupt changes in model dynamics has led to its popularity not only in econometrics (Choi 2009; Hamilton 1996b; Lux and Morales-Arias 2010), but also in other fields as diverse as traffic modeling (Cetin and Comert 2006), population dynamics (Luo and Mao 2007), river flow analysis (Vasas et al. 2007) or earthquake counts (Bulla and Berzel 2008). This paper is motivated by yet another stream of literature: electricity spot price models in energy economics (Bierbrauer et al. 2007; De Jong 2006; Erlwein et al. 2010; Huisman and de Jong 2003; Janczura and Weron 2010, 2012; Karakatsani and Bunn 2008, 2010; Mari 2008; Misiorek et al. 2006; Weron 2009). Regime-switching models have seen extensive use in this area due to their relative parsimony (a prerequisite in derivatives pricing) and the ability to capture the unique characteristics of electricity prices (in particular, the spiky and non-linear price behavior). While the existence of distinct regimes in electricity prices is generally unquestionable (being a consequence of the non-linear, heterogeneous supply stack structure in the power markets, see e.g., Eydeland and Wolyniec 2012; Weron 2006), the actual goodness-of-fit of the models requires statistical validation.

However, recent work concerning the statistical fit of regime-switching models has been mainly devoted to testing parameter stability versus the regime-switching hypothesis. Several tests have been constructed for the verification of the number of regimes. Most of them exploit the likelihood ratio technique (Cho and White 2007; Garcia 1998), but there are also approaches related to recurrence times (Sen and Hsieh 2009), likelihood criteria (Celeux and Durand 2008) or the information matrix (Hu and Shin 2008). Specification tests to detect autocorrelation and ARCH effects were proposed by Hamilton (1996a, based on the score function technique) and more recently by Smith (2008, utilizing the Rosenblatt transformation; see also Sect. 3.3). Smith found that the performance of the Ljung–Box test improved when used on the normally distributed Rosenblatt transformation. However, the considered Markov regime-switching models were relatively simple and had two states differing only in mean. Interestingly, the Rosenblatt transformation was used earlier for evaluating density forecasts of regime-switching (Berkowitz 2001; Diebold et al. 1998; Haas et al. 2004) and stochastic volatility models (Kim et al. 1998), typically in a risk management context.

On the other hand, to our best knowledge, procedures for goodness-of-fit testing of the marginal distribution of regime-switching models have not been derived to date (with the exception of Janczura and Weron 2009, where an ewedf-type test was introduced in the context of electricity spot price models, see Sect. 3.2.1 for details). With this paper we want to fill the gap and propose empirical distribution function (edf) based testing procedures built on the Kolmogorov–Smirnov test, that are dedicated to regime-switching models with observable as well as latent state processes. In contrast to the approaches based on the Rosenblatt transformation, the techniques proposed in this paper allow for testing the fit in the individual regimes as well as of the whole model. This can be advantageous in many situations as we additionally obtain information on which regimes are correctly and which are incorrectly specified. The derivation
of the tests is not straightforward and in the case of a latent state process requires an application of the concept of the weighted empirical distribution function (wedf). Finally, we should also note, that the term “marginal distribution” does not mean that the proposed tests do not account for the dynamic regime structure. On the contrary, the dynamical structure is considered when constructing residuals used in the testing procedures.

The paper is structured as follows: in Sect. 2 we describe the structure of the analyzed regime-switching models and briefly explain the estimation process (for details we refer to an article recently published in AStA; Janczura and Weron 2012). In Sect. 3, we introduce goodness-of-fit testing procedures appropriate for regime-switching models both with observable and latent state processes. Next, in Sect. 4 we provide a simulation study and check the performance of the proposed techniques. Since the motivation for this paper comes from the energy economics literature, in Sect. 5 we show how the presented testing procedure can be applied to verify the fit of Markov regime-switching models to electricity spot prices. Finally, in Sect. 6 we conclude and outline future work.

2 Regime-switching models

2.1 Model definition

Assume that the observed process $X_t$ may be in one of $L$ states (regimes) at time $t$, dependent on the state process $R_t$:

$$X_t = \begin{cases} X_{t,1} & \text{if } R_t = 1, \\ \vdots & \vdots & \vdots \\ X_{t,L} & \text{if } R_t = L. \end{cases} \quad (1)$$

Possible specifications of the process $R_t$ may be divided into two classes: those where the current state of the process is observable (like threshold models, e.g., TAR, SETAR) and those where it is latent. Probably, the most prominent representatives of the second group are the hidden Markov models (HMM; for a review see, e.g., Cappe et al. 2005) and their generalizations allowing for temporal dependence within the regimes—the Markov regime-switching models (MRS). Like in HMMs, in MRS models $R_t$ is assumed to be a Markov chain governed by the transition matrix $\mathbf{P}$ containing the probabilities $p_{ij}$ of switching from regime $i$ at time $t$ to regime $j$ at time $t+1$, for $i, j \in \{1, 2, \ldots, L\}$:

$$\mathbf{P} = (p_{ij}) = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1L} \\ p_{21} & p_{22} & \cdots & p_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ p_{L1} & p_{L2} & \cdots & p_{LL} \end{pmatrix}, \quad \text{with } p_{ii} = 1 - \sum_{j \neq i} p_{ij}. \quad (2)$$
The current state $R_t$ at time $t$ depends on the past only through the most recent value $R_{t-1}$. The probability of being in regime $j$ at time $t+m$ starting from regime $i$ at time $t$ is given by

$$P(R_{t+m} = j \mid R_t = i) = (P')^m \cdot e_i,$$  \hspace{1cm} (3)

where $P'$ denotes the transpose of $P$ and $e_i$ is the $i$th column of the identity matrix. In general, $L$ regime models can be considered. However, for clarity of exposition we limit the discussion in this paper to two regime models only. Note, that this is not a very restrictive limitation—at least in the context of modeling electricity spot prices—since typically two or three regimes are enough to adequately model the dynamics (Janczura and Weron 2010; Karakatsani and Bunn 2010). Nonetheless, all presented results are also valid for $L > 2$.

The definitions of the individual regimes can be arbitrarily chosen depending on the modeling needs. Again for the sake of clarity, in this paper we focus only on two commonly used in the energy economics literature specifications of MRS models (Ethier and Mount 1998; De Jong 2006; Hirsch 2009; Huisman and de Jong 2003; Janczura and Weron 2010; Mari 2008). The first one (denoted by I or type I) assumes that the process $X_t$ is driven by two independent regimes: (1) a mean-reverting AR(1) process:

$$X_{t,1} = \alpha + (1 - \beta)X_{t-1,1} + \sigma \epsilon_t$$  \hspace{1cm} (4)

with $0 < \beta < 1$ and $\sigma > 0$, where the residuals $\epsilon_t$s are independent, $F^1$-distributed (in the following we assume that $F^1$ is the standard Gaussian cdf) and (2) an i.i.d. sample from a specified continuous, strictly monotone distribution $F^2$:

$$X_{t,2} \sim F^2(x).$$  \hspace{1cm} (5)

Observe that in such a model specification the values of the first regime $X_{t,1}$ become latent when the process is in the second state and they do not depend on the realization (trajectory) of the second regime. Such a specification, though computationally challenging, is useful for modeling processes with radically different dynamics in the individual regimes. For instance, in wholesale power markets the price spikes or price drops (i.e., negative price spikes) are typically driven by unexpected changes of market conditions, like a generation outage or a severe heat wave (in case of price spikes) or favorable wind conditions combined with low consumption (in case of price drops). Price spikes occur due to lack of storage capabilities and limited flexibility to respond to sudden changes in supply and/or demand for electricity. When the reason for the spike is over the prices move back to the normal (or base) level, usually irrespective of the magnitude of the extreme prices a few hours or days earlier (Eydeland and Wolyniec 2012; Huisman 2009; Weron 2006). For an example of such a behavior see Fig. 1 in Sect. 5.

In the second specification (denoted by II or type II) $X_t$ is described by an AR(1) process having different parameters in each regime, namely:
\[ X_t = \alpha_{R_t} + (1 - \beta_{R_t})X_{t-1} + \sigma_{R_t}\epsilon_t, \quad R_t \in \{1, 2\}, \tag{6} \]

where the residuals \( \epsilon_t \)s are independent, \( N(0, 1) \)-distributed random variables. Again, we assume that \( 0 < \beta_i < 1 \) and \( \sigma_i > 0 \).

2.2 Estimation

Estimation of regime-switching models with an observable state process boils down to the problem of independently estimating parameters in each regime. In case of MRS models, though, the estimation process is not straightforward, since the state process is latent and not directly observable. We have to infer the parameters and state process values at the same time. In this paper, we use a variant of the Expectation–Maximization (EM) algorithm that was first applied to MRS models by Hamilton (1990) and later refined by Kim (1994). It is a two-step iterative procedure, reaching a local maximum of the likelihood function:

- **Step 1** Denote the observation vector by \( x_T = (x_1, x_2, \ldots, x_T) \). For a parameter vector \( \theta^{(n)} \) compute the conditional probabilities \( P(R_t = i | x_T; \theta^{(n)}) \)—the so-called ‘smoothed inferences’—for the process being in regime \( i \) at time \( t \).

- **Step 2** Calculate new and more exact maximum likelihood estimates \( \theta^{(n+1)} \) using the log-likelihood function, weighted with the smoothed inferences from Step 1, i.e.,

\[
\log \left[ L(\theta^{(n+1)}) \right] = \sum_{i=1}^{2} \sum_{t=1}^{T} P(R_t = i | x_T; \theta^{(n)}) \log \left[ f_i(x_t | x_{t-1}; \theta^{(n+1)}) \right],
\]

where \( f_i(x_t | x_{t-1}; \theta^{(n+1)}) \) is the conditional density of the \( i \)th regime, and update the transition probabilities:

\[
P_{ij}^{(n+1)} = \frac{\sum_{t=2}^{T} P(R_t = j, R_{t-1} = i | x_T; \theta^{(n)})}{\sum_{t=2}^{T} P(R_{t-1} = i | x_T; \theta^{(n)})}.
\]

For a detailed description of the estimation procedure see the original paper of Kim (1994) or a recent article of Janczura and Weron (2012), where an efficient algorithm for MRS models of type I is presented.

3 Goodness-of-fit testing

In this section, we introduce a goodness-of-fit testing technique, that can be applied to evaluate the fit of regime-switching models. It is based on the Kolmogorov–Smirnov (K-S) goodness-of-fit test and verifies whether the null hypothesis \( H_0 \) that observations come from the distribution implied by the model specification cannot be rejected. The procedure can be easily adapted to other empirical distribution function (edf) type tests, like the Anderson–Darling test.
3.1 Testing in case of an observable state process

3.1.1 Specification I

In this case the hypothesis $H_0$ states that the sample $(x_1, x_2, \ldots, x_T)$ is generated from a regime-switching model with two independent regimes defined as: an AR(1) process (first regime) and i.i.d. $F^2$-distributed random variables (second regime). Provided that the values of the state process $R_t$ are known, observations can be split into separate subsamples related to each of the regimes. Namely, subsample $i$ consists of all values $X_t$ satisfying $R_t = i$. The regimes are independent from each other, but the i.i.d. condition must be satisfied within the subsamples themselves. Therefore, the mean-reverting regime observations are substituted by their respective residuals. Precisely, the following transformation is applied to each pair of consecutive AR(1) observations in regime $R_t = 1$:

$$
h(x, y, k) = \frac{x - (1 - \beta)^k y - \alpha \frac{1-(1-\beta)^k}{\beta}}{\sigma \sqrt{\frac{1-(1-\beta)^{2k}}{1-(1-\beta)^2}}},
$$

(7)

where $(k - 1)$ is the number of latent observations from the mean-reverting regime (or equivalently the number of observations from the second regime that occurred between two consecutive AR(1) observations) and $\alpha$, $\beta$ and $\sigma$ are the model parameters, see (4). It is straightforward to see that if $H_0$ is true, transformation $h(x_{t+k,1}, x_{t,1}, k)$ applied to consecutive observations from the mean-reverting AR(1) regime leads to a sample $(y_1^1, y_2^1, \ldots, y_{n_1}^1)$ of independent and conditionally $N(0, 1)$-distributed random variables. Note, that from now on we use the following notation. The original observed sample is denoted by $(x_1, x_2, \ldots, x_T)$. The i.i.d. (or conditionally i.i.d. in Sect. 3.2) samples in each of the regimes are denoted by $(y_1^1, y_2^1, \ldots, y_{n_1}^1)$ and $(y_1^2, y_2^2, \ldots, y_{n_2}^2)$, with $n_1 + n_2 = T - 1$. Note, that for the mean-reverting regime these samples are obtained by applying transformation (7).

Further, observe that transformation $h(X_{t+k,1}, X_{t,1}, k)$ is based on subtracting the conditional mean from $X_{t+k,1}$ and standardizing it with the conditional variance. Indeed, $(1 - \beta)^k X_{t,1} + \alpha \frac{1-(1-\beta)^k}{\beta}$ is the conditional expected value of $X_{t+k,1}$ given $(X_{1,1}, X_{2,1}, \ldots, X_{t,1})$ and $\sigma^2 \frac{1-(1-\beta)^{2k}}{1-(1-\beta)^2}$ is the respective conditional variance.

Transformation (7) ensures that the subsample containing observations from the mean-reverting regime is i.i.d. Since the second regime is i.i.d. by definition, standard goodness-of-fit tests based on the empirical distribution function (like the Kolmogorov–Smirnov or Anderson–Darling tests, see e.g., D’Agostino and Stevens 1986) can be applied to each of the subsamples. Recall that the Kolmogorov–Smirnov test statistic is given by:

$$
D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,
$$

(8)
Goodness-of-fit testing for the marginal distribution of regime-switching models 245

where \( n \) is the sample size, \( F_n \) is the empirical distribution function (edf) and \( F \) is the corresponding theoretical cumulative distribution function (cdf). Hence, having an i.i.d. sample \((y_1, y_2, \ldots, y_n)\), the test statistic can be calculated as

\[
d_n = \sqrt{n} \max_{1 \leq t \leq n} \left| \frac{1}{n} \sum_{k=1}^{n} I\{y_k \leq y_t\} - F(y_t) \right|,
\]

where \( I \) is the indicator function.

The goodness-of-fit of the marginal distribution of the individual regimes can be formally tested. For the mean-reverting regime, \( F \) is the standard Gaussian cdf and \((y_1, y_2, \ldots, y_{n_1})\) is the subsample of the standardized residuals obtained by applying transformation (7), while for the second regime, \( F \) is the model-specified cdf (i.e., \( F^2 \)) and \((y_1, y_2, \ldots, y_{n_2})\) is the subsample of respective observations. Observe that the 'whole model' goodness-of-fit can be also verified, using the fact that for \( X \sim F^2 \) we have that \( Y = (F^1)^{-1}[F^2(X)] \) is \( F^1 \)-distributed. Indeed, a sample \((y^1_1, y^1_2, \ldots, y^1_{n_1}, y^2_1, y^2_2, \ldots, y^2_{n_2})\), where \( y^1_i \)s are the standardized residuals of the mean-reverting regime, while \( y^2_i \)s are the transformed variables corresponding to the second regime, i.e., \( y^2_t = (F^1)^{-1}[F^2(x_t, 2)] \), is i.i.d. \( N(0, 1) \)-distributed and, hence, the testing procedure is applicable.

3.1.2 Specification II

The \( H_0 \) hypothesis now states that the sample \((x_1, x_2, \ldots, x_T)\) is driven by a regime-switching model defined by Eq. (6) with \( R_t \in \{1, 2\} \). Similarly, as in the independent regimes case, the testing procedure is based on extracting the residuals of the mean-reverting process. Indeed, observe that under the \( H_0 \) hypothesis the transformation \( h(x_t, x_{t-1}, 1) \), defined in (7), with parameters \( \alpha_{R_t}, \beta_{R_t} \) and \( \sigma_{R_t} \) corresponding to the current value of the state process \( R_t \), yields an i.i.d. \( N(0, 1) \) distributed sample. Thus, the Kolmogorov–Smirnov test can be applied. The test statistic \( d_n \), see (9), is calculated with the standard Gaussian cdf and the sample \((y_1, y_2, \ldots, y_T)\) of the standardized residuals, i.e., \( y_t = h(x_t, x_{t-1}, 1) \).

3.1.3 Critical values

Note, that the described above testing procedure is valid only if the parameters of the hypothesized distribution are known. Unfortunately in typical applications the parameters have to be estimated beforehand. If this is the case, then the critical values for the test must be reduced (Čižek et al. 2011). In other words, if the value of the test statistics \( d_n \) is \( d \), then the \( p \) value is overestimated by \( P(d_n \geq d) \). Hence, if this probability is small, then the \( p \) value will be even smaller and the hypothesis will be rejected. However, if it is large then we have to obtain a more accurate estimate of the \( p \) value.

To cope with this problem, Ross (2002) recommends using Monte Carlo simulations. In our case the procedure reduces to the following steps. First, the parameter vector \( \hat{\theta} \) is estimated from the dataset and the test statistic \( d_n \) is calculated according
to formula (9). Next, $\hat{\theta}$ is used as a parameter vector for $N$ simulated samples from the assumed model. For each sample the new parameter vector $\hat{\theta}_i$ is estimated and the new test statistic $d_n^i$ is calculated using formula (9). Finally, the $p$ value is obtained as the proportion of simulated samples with the test statistic values higher or equal to $d_n$, i.e., $p$ value $= \frac{1}{N} \# \{i : d_n^i \geq d_n\}$.

3.2 Testing in case of a latent state process

3.2.1 The ewedf approach

Now, assume that the sample $(x_1, x_2, \ldots, x_T)$ is driven by an MRS model. The regimes are not directly observable and, hence, the standard edf approach can be used only if an identification of the state process is performed first. Recall that, as a result of the estimation procedure described in Sect. 2.2, the so called ‘smoothed inferences’ about the state process are derived. The smoothed inferences are the probabilities that the $t$th observation comes from a certain regime given the whole available information $P(R_t = i | x_1, x_2, \ldots, x_T)$. Hence, a natural choice is to relate each observation with the most probable regime by letting $R_t = i$ if $P(R_t = i | x_1, x_2, \ldots, x_T) > 0.5$. Then, the testing procedure described in Sect. 3.1 is applicable. However, we have to mention, that the hypothesis $H_0$ now states that $(x_1, x_2, \ldots, x_T)$ is driven by a regime-switching model with known state process values. We call this approach ‘ewedf’, which stands for ‘equally-weighted empirical distribution function’. It was introduced by Janczura and Weron (2009) in the context of electricity spot price MRS models.

3.2.2 The weighted empirical distribution function (wedf)

In the standard goodness-of-fit testing approach based on the edf each observation is taken into account with weight $\frac{1}{n}$ (i.e., inversely proportional to the size of the sample). However, in MRS models the state process is latent. The estimation procedure (the EM algorithm) only yields the probabilities that a certain observation comes from a given regime. Moreover, in the resulting marginal distribution of the MRS model each observation is, in fact, weighted with the corresponding probability. Therefore, a similar approach should be used in the testing procedure.

For this reason, we introduce here the concept of the weighted empirical distribution function (wedf):

$$F_n(x) = \sum_{i=1}^{n} \frac{w_i}{\sum_{t=1}^{n} w_t} I\{y_t < x\},$$  \hspace{1cm} (10)$$

where $(y_1, y_2, \ldots, y_n)$ is a sample of observations and $(w_1, \ldots, w_n)$ are the corresponding weights, such that $0 \leq w_t \leq M, \forall t=1,\ldots,n$. It is interesting to note, that the notion of the weighted empirical distribution function appears in the literature in different contexts. Maiboroda (1996, 2000) applied it to the problem of estimation and testing for homogeneity of components of mixtures with varying coefficients.
Withers and Nadarajah (2010) investigated properties of distributions of smooth functionals of \( F_n(x) \). In both approaches the weights were assumed to fulfill the condition \( \sum_{i=1}^n w_i = n \). A different choice of weights was used by Huang and Brill (2004). They proposed the level-crossing method to find weights improving efficiency of the edf in the distribution tails. Yet another approach employing the weighted distribution is the generalized (weighted) bootstrap technique, see e.g., Haeusler et al. (1991), where specified random weights are used to improve the resampling method.

However, to our best knowledge, none of the applications of wedf is related to goodness-of-fit testing of Markov regime-switching models. Here, we use the wedf concept to deal with the case when observations cannot be unambiguously classified to one of the regimes and, hence, a natural choice of weights of wedf seems to be \( w_i = P(R_t = i|x_1, x_2, \ldots, x_T) = E(I_{\{R_t = i\}}|x_1, x_2, \ldots, x_T) \) for the \( i \)th regime observations.

### 3.2.3 The wedf approach for specification II

First, let us focus on the parameter-switching specification. The \( H_0 \) hypothesis states that the sample \( x_T = (x_1, x_2, \ldots, x_T) \) is driven by the MRS model defined by equation (6). Assume that \( H_0 \) is true and the model parameters are known. Like in the observable state process case, the test cannot be applied directly to the observed sample. Let \( y_i^t \)'s be the transformed variables corresponding to the \( i \)th regime, i.e., \( y_i^t \)'s are obtained as \( y_i^t = [x_{t+1} - \alpha_i - (1 - \beta_t) x_t]/\sigma_i \). Observe that if \( R_t = i \), then \( y_i^t \) becomes the residual of the \( i \)th regime and, hence, has the standard Gaussian distribution. The weighted empirical distribution function (wedf) is then given by:

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^n \left[ P(R_t = 1|x_T)^{I_{\{y_i^T < x\}}} + P(R_t = 2|x_T)^{I_{\{y_i^2 < x\}}} \right],
\]

where \( n \) is the size of the sample (here \( n = T - 1 \)). Let \( \mathfrak{M} \) be the \( \sigma \)-algebra generated by the state process \( \{R_t\}_{t=1,2,\ldots,T} \), i.e., the state process history up to time \( T \). Observe that the elements of the sum in (11) are conditionally independent given \( \mathfrak{M} \). Indeed, if for a given \( t \), \( R_t = i \) then the \( t \)th component of the sum becomes \( I_{\{y_i^T < x\}} \) and \( y_i^t \)'s given \( R_t = i \) form an i.i.d. \( N(0, 1) \)-distributed sample. Moreover, the following lemma ensures that the true cdf of the residuals can be approximated by the wedf.

**Lemma 1** If \( H_0 \) is true, then \( F_n \) given by (11) is an unbiased, consistent estimator of the distribution of the residuals (in this case Gaussian).

Note, that proofs of all lemmas and theorems formulated in this section can be found in the Appendix.

The following theorem yields a version of the K-S test applicable to parameter-switching MRS model (6). Note, that if the state process was observable, it would boil down to the standard K-S test (Lehmann and Romano 2005, p. 584).

**Theorem 1** Let \( F_n \) be given by (11) and \( F \) be the standard Gaussian cdf. If \( H_0 \) is true and the model parameters are known, then the statistic
converges (weakly) to the Kolmogorov–Smirnov distribution $KS$ as $n \to \infty$.

If hypothesis $H_0$ is true then, by Theorem 1, the statistic $D_n$ asymptotically has the Kolmogorov–Smirnov distribution. Therefore, if $n$ is large enough, the following approximation holds

$$P(D_n \geq c | H_0) \approx P(\kappa \geq c),$$

where $\kappa \sim KS$ and $c$ is the critical value. (Recall that $y_i^t = [x_{t+1} - \alpha_i - (1 - \beta_i)x_t]/\sigma_i$.) Hence, the $p$ value for the sample $(y_1^1, y_2^1, \ldots, y_n^1, y_1^2, y_2^2, \ldots, y_n^2)$ can be approximated by $P(\kappa \geq d_n)$, where

$$d_n = \sqrt{n} \max_{1 \leq t \leq n} \max_{i=1,2} \left| F_n(y_i^t) - F(y_i^t) \right|$$

is the test statistic. Note that, for a given value of $d_n$, $P(\kappa > d_n)$ is the standard Kolmogorov–Smirnov test $p$ value, so that the K-S test tables can be easily applied in the wedf approach.

The above procedure is applicable to testing the distribution of the residuals of the (whole) model. A similar approach can be used for testing the distributions of the residuals of the individual regimes. Let the wedf for the $i$th regime be defined as:

$$F_n^i(x) = \sum_{t=1}^{n} \frac{P(R_t = i | x_T)\mathbb{1}_{\{y_i^t < x\}}}{\sum_{t=1}^{n} P(R_t = i | x_T)},$$

where again $y_i^t$s are the transformed variables corresponding to the $i$th regime, i.e., $y_i^t = [x_{t+1} - \alpha_i - (1 - \beta_i)x_t]/\sigma_i$. Further, denote the theoretical distribution of the $i$th regime residuals (here Gaussian) by $F^i$.

**Lemma 2** If $H_0$ is true, then $F_n^i(x)$ given by (15) is an unbiased estimator of $F^i(x)$. Moreover, it is consistent if $\forall_{i, j=1,2} p_{ij} < 1$.

An analogue of Theorem 1 can be derived.

**Theorem 2** Let $F_n$ be given by (15) and assume that $R_t$ is an ergodic Markov chain. If $H_0$ is true and the model parameters are known, then the statistic

$$D_n^i = \sqrt{w_n} \sup_{x \in \mathbb{R}} \left| F_n^i(x) - F^i(x) \right|,$$

where

$$w_n = \sum_{\{i_1, i_2, \ldots, i_n\} \in I} \mathbb{1}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \left[ \frac{\sum_{k: i_k = i} \mathbb{1}_{\{R_k = i\}}}{\sum_{k: i_k = i} \mathbb{1}_{\{R_k = i\}}} \right]^2$$

\(\square\) Springer
and \( I = \{ (i_1, i_2, \ldots, i_n) : i_k \in \{1, 2\}, k = 1, 2, \ldots, n \} \) converges (weakly) to the Kolmogorov–Smirnov distribution \( KS \) as \( n \to \infty \).

Observe that \( \sqrt{w_n} \) can be approximated by \( \sum_{t=1}^n P(R_t = i|X_T) \). Hence, for a sample of \((y_1^i, y_2^i, \ldots, y_n^i)\) the test statistic is given by

\[
d_n^i = \sum_{t=1}^n P(R_t = i|X_T) \max_{1 \leq t \leq n} \left| F_n^i(y_t^i) - F^i(y_t^i) \right|
\]  

(18)

and the standard testing procedure can be applied.

3.2.4 The wef approach for specification I

Now, assume that the sample \((x_1, x_2, \ldots, x_T)\) is driven by the MRS model with independent regimes. The results of Theorems 1 and 2 can be applied, however, slight modifications of the tested sample(s) are required. First, observe that the values of the mean-reverting regime become latent, when the process is in the second state. As a consequence, the calculation of the conditional mean and variance, required for the derivation of the residuals, is not straightforward. We have:

\[
E(X_{t,1}|x_{t-1}) = \alpha + (1 - \beta)E(X_{t-1,1}|x_{t-1}),
\]

\[
\text{Var}(X_{t,1}|x_{t-1}) = (1 - \beta)^2 \text{Var}(X_{t-1,1}|x_{t-1}) + \sigma^2,
\]

where \( x_{t-1} = (x_1, x_2, \ldots, x_{t-1}) \) is the vector of preceding observations. Therefore, the standardized residuals are given by the transformation:

\[
g(X_{t,1}, x_{t-1}) = \frac{X_{t,1} - \alpha - (1 - \beta)E(X_{t-1,1}|x_{t-1})}{\sqrt{(1 - \beta)^2 \text{Var}(X_{t-1,1}|x_{t-1}) + \sigma^2}},
\]

(19)

where \( E(X_{t-1,1}|x_{t-1}) \) and \( \text{Var}(X_{t-1,1}|x_{t-1}) \) can be calculated using the following equalities:

\[
E(X_{t,1}|x_t) = P(R_t = 1|x_t) x_t + P(R_t \neq 1|x_t) \left[ \alpha + (1 - \beta)E(X_{t-1,1}|x_{t-1}) \right],
\]

(20)

\[
E(X_{t,1}^2|x_t) = P(R_t = 1|x_t) x_t^2 + P(R_t \neq 1|x_t) \left[ \alpha^2 + 2\alpha(1 - \beta)E(X_{t-1,1}|x_{t-1}) + (1 - \beta)^2 \text{Var}(X_{t-1,1}|x_{t-1}) + \sigma^2 \right].
\]

(21)

The latter formula is a consequence of the law of iterated expectation and basic properties of conditional expected values. Finally, the values \( P(R_t = 1|x_t) \) are calculated from the Bayes rule during the EM estimation procedure (see e.g., Kim 1994). Note that the transformed variables \((y_1^1, y_2^1, \ldots, y_{T-1}^1)\), where \( y_t^1 = g(x_{t,1}, x_{t-1}) \), are \( \mathcal{N} \)-independent and \( \mathcal{N}(0, 1) \)-distributed conditionally on \( \mathcal{M} \).
Now, to test the fit of the mean-reverting regime, it is enough to calculate \( d_n^i \) according to formula (18) with the standard Gaussian cdf and \( y_i^1 = g(x_t, x_{t-1}) \). Observe, that the observations from the second regime are i.i.d. by definition, so the testing procedure is straightforward with \( F^2 \) cdf and sample \((x_1, x_2, \ldots, x_T)\). Moreover, the ‘whole model’ goodness-of-fit can be also verified. Theorem 1 is directly applicable, if the distributions of the samples corresponding to both regimes are the same \( F = F^1 = F^2 \). Observe that, even if \( F^1 \neq F^2 \), the test still can be applied using the fact that for \( X \sim F^2 \) we have that \( Y = (F^1)^{-1}[F^2(X)] \) is \( F^1 \)-distributed. The test statistic \( d_n \) is calculated as in (14) with \( F^1 \) cdf (here Gaussian) and the sample \((y_1^1, y_2^1, \ldots, y_{T-1}^1, y_1^2, y_2^2, \ldots, y_T^2)\), where \((y_1^1, y_2^1, \ldots, y_{T-1}^1)\) are the transformed variables of the mean-reverting regime, i.e., \( y_i^1 = g(x_{t,1}, x_{t-1}) \), while \((y_1^2, y_2^2, \ldots, y_T^2)\) are the variables corresponding to the second regime, i.e., \( y_T^2 = (F^1)^{-1}[F^2(x_t)] \).

Note, that as in the case of an observable state process, in the wedf approach we face the problem of estimating values that are later used to compute the test statistic. Again, this problem can be circumvented with the help of Monte Carlo simulations. The \( p \) values can be computed as the proportion of simulated MRS model trajectories with the test statistic \( d_n \), see formulas (14) and (18), higher or equal to the value of \( d_n \) obtained from the dataset.

3.3 The Rosenblatt transformation

The new tests proposed in this section are compared with an approach utilizing the Rosenblatt (1952) transformation. The latter is based on the fact that if a sample \((x_1, x_2, \ldots, x_T)\) is driven by a multivariate distribution \( F \), then the transformed variables \( y_i = F_i(x_t \mid x_{t-1}) \), where \( F_i \) is the corresponding conditional cdf, are independent and uniformly distributed. For the MRS models considered in this paper the transformation is given by

\[
y_t = \sum_{i=1}^{2} P(R_i = i \mid x_{t-1}) F_i^j(x_t \mid x_{t-1}),
\]

where \( F_i^j \) is the conditional distribution of regime \( i \). For specification I, \( F_i^1 \) is the normal cdf with mean and variance given by (19), while \( F_i^2 \) is the model defined, see (5), cdf of the second regime. For specification II, \( F_i^j \) is the normal cdf with mean \( \alpha_i + (1 - \beta_i)x_{t-1} \) and variance \( \sigma_i^2 \). Since \((y_1, y_2, \ldots, y_T)\) form an i.i.d. sample under \( H_0 \), the standard edf-type tests—like the Kolmogorov–Smirnov—can be applied. In order to test for the same distribution in all testing approaches, we apply one more transformation. Namely, like Berkowitz (2001), Haas et al. (2004) or Smith (2008), we calculate \( \Phi^{-1}(y_T) \) with \( \Phi \) being the standard normal cdf and obtain an independent \( N(0, 1) \)-distributed sample.

The Rosenblatt transformation is a very useful and general tool, however, it can be used to test the goodness-of-fit of the whole model only. In contrast, the ewedf and the wedf approaches allow for testing the fit in the individual regimes. This can be advantageous in many situations as we additionally obtain the information which
regimes are correctly and which are incorrectly specified. Moreover, the ewedf and the wedf approaches yield estimators of the regime and model cdfs, providing a readily available tool for further testing and model building. On the other hand, in case of the Rosenblatt transformation an empirical distribution function can only be constructed for the transformed variables making it hard to be interpreted.

4 Simulations

We now check the performance of the testing procedures introduced in Sect. 3. Due to space limitations, we focus on the more challenging case of a latent state process and consider four 2-regime MRS models defined in Table 1. The parameters of models Sim #1 and Sim #2 are chosen arbitrarily, while those of models Sim #3 and Sim #4 are estimated from the NEPOOL log-prices studied in Sect. 5. Furthermore, models Sim #1 and Sim #3 follow specification I, i.e., the first regime is driven by an AR(1) process, while the second regime is described by an i.i.d. sample of log-normally distributed random variables with parameters $\alpha_2$ and $\sigma_2^2$, i.e., $LN(\alpha_2, \sigma_2^2)$. Recall, that a random variable $X$ is log-normally distributed, $X \sim LN(\alpha_2, \sigma_2^2)$, if $\log(X) \sim N(\alpha_2, \sigma_2^2)$, for $X > 0$. In order to apply the tests to the ‘whole model’ (and not only to the individual regimes) we transform the second regime values $\{x_i\}$ obtaining an $N(0, 1)$-distributed sample: $\{[\log(x_i) - \alpha_2] / \sigma_2; i = 1, \ldots, T\}$. On the other hand, models Sim #2 and Sim #4 are simulated from the parameter-switching AR(1) model, i.e., follow specification II, see formula (6). Finally note that since the regimes of the considered models are not directly observable, the standard edf-based goodness-of-fit tests cannot be used.

4.1 Known model parameters

We generate 10,000 trajectories of each of the four 2-regime MRS models defined in Table 1. The length of each trajectory is 2,000 observations, which corresponds to 5.5 years of daily data (note, that markets for electricity operate 365 days per year). We apply the ewedf, the wedf and the Rosenblatt transformation-based goodness-of-fit tests to

| Table 1 | Parameters of four 2-regime MRS models analyzed in the simulation study of Sect. 4 |
|---------|---------------------------------------------------------------------|
|         | Parameters | Probabilities |
|         | $\alpha_1$ | $\beta_1$ | $\sigma_1^2$ | $\alpha_2$ | $\beta_2$ | $\sigma_2^2$ | $p_{11}$ | $p_{22}$ |
| Sim #1  | 10.00      | 0.80      | 10.0000      | 4.00       | –         | 0.5000      | 0.90     | 0.20     |
| Sim #2  | 1.00       | 0.80      | 1.0000       | 3.00       | 0.40      | 0.5000      | 0.60     | 0.50     |
| Sim #3  | 0.58       | 0.17      | 0.0048       | 1.28       | –         | 0.0034      | 0.97     | 0.88     |
| Sim #4  | 0.63       | 0.81      | 0.0038       | 0.95       | 0.73      | 0.0200      | 0.96     | 0.91     |

Models Sim #1 and Sim #3 follow specification I and models Sim #2 and Sim #4 specification II. The parameters of models Sim #1 and Sim #2 are chosen arbitrarily, while those of models Sim #3 and Sim #4 are estimated from the NEPOOL log-prices studied in Sect. 5.
each simulated trajectory and then calculate the percentage of rejected hypotheses $H_0$ at the 5% significance level. We assume that the model parameters are known. The computation of $E(X_t,1|x_t)$ in the wedf approach requires backward recursion until the previous observation from the mean-reverting regime is found, see (21). However, as the number of observations is limited, the condition $P(R_t = 1|x_t) = 1$ might not be fulfilled at all. The estimation scheme requires some approximation or an additional assumption. Here, we assume that for each simulated trajectory the first observation comes from the mean-reverting regime.

In the ewedf approach the tested hypothesis says that the state process is known (and coincides with the proposed classification of the observations to the regimes). As a consequence, once the regimes are identified, it is equivalent to the standard edf approach. To test how it performs for an MRS model with a latent state process, we apply it to the simulated trajectories (we first identify the regimes, then test whether the sample is generated from the assumed MRS model).

The results reported in Table 2 indicate that only the wedf and Rosenblatt transformation-based tests yield correct percentages of rejected hypotheses. The values obtained for the ewedf-based test are far from the expected level of 5%. The ewedf approach is more restrictive, probably due to the less flexible classification of the regimes in which the probabilities $P(R_t = i|x_t)$ can only take values of 0 or 1. This simple example clearly shows that in case of MRS models the ewedf approach is less reliable.

Finally, to measure the quality of regime classification, we use the regime classification measure (RCM) of Ang and Bekaert (2002), see the last column in Table 2. Since the true regime is a Bernoulli random variable, the RCM statistic is essentially a sample estimate of its variance. The RCM statistic is rescaled so that a value of 0 means perfect regime classification and a value of 100 implies that no information about the regimes is revealed. In our case, the regime classification is very good or good for specification I models used for modeling electricity prices in Sect. 5 (Sim #1 and Sim #3) and good or moderately good for specification II models (Sim #2 and Sim #4). The lower RCM values for type I models also imply that in these models the

| Table 2 | Percentage of rejected hypotheses $H_0$ at the 5% significance level calculated from 10,000 simulated trajectories of 2000 observations each of the four 2-regime MRS models defined in Table 1 |
|---------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|
|         | ewedf                                             | wedf                                             | Rtrans                                            | RCM                                               |
|         | First     | Second   | Model   | First     | Second   | Model   | Model   |
| Sim #1  | 0.0641    | 0.8648   | 0.1242  | 0.0527    | 0.0540   | 0.0481  | 0.0542  | 4.33    |
| Sim #2  | 0.2244    | 0.0798   | 0.1194  | 0.0499    | 0.0527   | 0.0314  | 0.0493  | 12.69   |
| Sim #3  | 0.0586    | 0.4190   | 0.1060  | 0.0817    | 0.0458   | 0.0702  | 0.0902  | 12.81   |
| Sim #4  | 0.0519    | 0.3206   | 0.1025  | 0.0505    | 0.0389   | 0.0341  | 0.0553  | 29.63   |

The results of the K-S test within the ewedf and the wedf approaches are reported in columns 2–4 and 5–7, respectively, independently for the two regimes (First, Second) and for the whole model (Model). Results obtained using the Rosenblatt transformation (Rtrans) are reported in column 8. Finally, the RCM statistic of Ang and Bekaert (2002) averaged over the 10,000 simulated trajectories is reported for each model in the last column.
regimes are better separated than in the respective type II models. Finally, in models whose parameters are estimated from the NEPOOL log-prices (Sim #3 and Sim #4) the regimes are less separated than in the other two models.

4.2 Unknown model parameters

The simulation results presented so far were obtained with the assumption that model parameters are known. Unfortunately, in typical applications the parameters have to be estimated before the testing procedure is performed. This may result in overestimated $p$ values. To cope with this problem, we use Monte Carlo simulations (for details, see e.g., Čižek et al. 2011, Ross 2002). For each of the 500 trajectories (of 2,000 observations) simulated from each of the four 2-regime MRS models defined in Table 1 the procedure is as follows:

1. Estimate the parameter vector $\hat{\theta}$ and calculate the test statistic ($d_n$) according to formula (9).
2. For ‘K-S table’-type (‘K-S tab.’) estimation calculate the $p$ value using K-S test tables and assuming that the sample comes from a model with parameter vector $\hat{\theta}$.
3. For ‘MC simulation’-type (‘MC sim.’) estimation:
   (a) simulate $N = 500$ trajectories with parameter vector $\hat{\theta}$ (these trajectories will be used to compute the estimate of the $p$ value),
   (b) for each trajectory $i = 1, \ldots, N$ estimate the parameter vector ($\hat{\theta}_i$) and calculate the test statistic ($d_{ni}$),
   (c) calculate the $p$ value as the proportion of simulated trajectories with the test statistic values higher or equal to $d_n$, i.e., $\frac{1}{N}\#\{i : d_{ni} \geq d_n\}$.

Note, that in contrast to the simulation study of Sect. 4.1, where we used 10,000 simulated trajectories to obtain the percentages given in Table 2, now we use 500 simulated trajectories for the ‘K-S table’-based percentages and $500 \times 500 = 250,000$ simulated trajectories for the ‘MC simulation’-based percentages in Table 3. Despite the substantially increased computational burden the accuracy of the percentages of rejected hypotheses has decreased as now it is based on only 500 simulations.

Looking at the test results based on the K-S test tables (‘K-S tab.’ in Table 3), for the ewedf approach the rejection percentages deviate significantly from the 5% level. On the other hand, for the wedf and the Rosenblatt transformation-based approaches the $p$ values are overestimated, what results in rejection percentages much lower than the 5% significance level. Observe that for most of the models none of the tests were rejected. Therefore, if $p$ values obtained with the wedf or the Rosenblatt transformation-based approaches are close to the significance level, the test may fail to reject a false $H_0$ hypothesis. This is not the case for the testing approach utilizing Monte Carlo simulations (‘MC sim.’ in Table 3) as the obtained rejection percentages are close to the 5% significance level. This example clearly shows that the wedf and the Rosenblatt transformation test based on the K-S test tables can only be used if it returns a $p$ value below the significance level (i.e., if it rejects the $H_0$ hypothesis) or well above the significance level. However, if the obtained $p$ value is close to the significance level, Monte Carlo simulations should be performed.
Table 3 Percentage of rejected hypotheses $H_0$ at the 5% significance level calculated from 500 simulated trajectories of 2,000 observations each of the models defined in Table 1 with parameters estimated from each sample

| Regime | ewedf First | Second | Model | wedf First | Second | Model | Rtrans Model |
|--------|-------------|--------|-------|------------|--------|-------|--------------|
| Sim #1 | K-S tab. | 0.0360 | 0.5860 | 0.0540 | 0.0420 | 0 | 0.0400 | 0.0400 |
|        | MC sim. | 0.0640 | 0 | 0.0580 | 0.0660 | 0.0580 | 0.0680 | 0.0620 |
| Sim #2 | K-S tab. | 0.0040 | 0.1940 | 0.0080 | 0.0440 | 0.0340 | 0.0520 | 0.0460 |
|        | MC sim. | 0.0320 | 0.0080 | 0.0360 | 0 | 0 | 0.0380 | 0.0380 |
| Sim #3 | K-S tab. | 0 | 0.0140 | 0.0400 | 0.0340 | 0.0500 | 0.0300 | 0.0380 |
|        | MC sim. | 0.0360 | 0.0440 | 0.0400 | 0.0340 | 0.0500 | 0.0300 | 0.0380 |
| Sim #4 | K-S tab. | 0.1400 | 0.0020 | 0.0120 | 0 | 0 | 0 | 0 |
|        | MC sim. | 0.0340 | 0.0180 | 0.0560 | 0.0580 | 0.0380 | 0.0640 | 0.0420 |

The results of the tests based on the ewedf as well as the wedf approaches are reported independently for the two regimes (First, Second) and the whole model (Model). Results obtained using the Rosenblatt transformation (Rtrans) are reported in the last column. The tests utilize K-S test tables (K-S tab.) or Monte Carlo simulations (MC sim.) with $N = 500$ repetitions. Note, that the rejection rates are approximations based on only 500 trajectories, not 10,000 trajectories as in Table 2.

4.3 Power of the tests

In this section we investigate the power of the tests. To this end, for a given MRS model we simulate 500 trajectories of 100, 500 or 2,000 observations each. Next, for each trajectory we calibrate an MRS model with an alternative specification of the regimes and perform goodness-of-fit tests to verify if the simulated trajectory can be driven by the alternative model. Finally, we calculate the percentages of the rejected hypotheses. The tests utilize all three approaches—ewedf, wedf and Rosenblatt transformation-based—and both methods of calculating $p$ values—K-S test tables and Monte Carlo simulations. We first consider the following three cases with arbitrarily chosen parameters:

- **AR-ARG1 vs. AR-AR** The trajectories are simulated from an MRS model defined as:

  $$X_t = \alpha_{R_t} + (1 - \beta_{R_t})X_{t-1} + \sigma_{R_t}X_{t-1}^{\gamma_{R_t}}\epsilon_t, \quad R_t \in \{1, 2\},$$

  where $\alpha_1 = 1, \beta_1 = 0.8, \sigma_1^2 = 1, \gamma_1 = 0, \alpha_2 = 3, \beta_2 = 0.4, \sigma_2^2 = 0.05, \gamma_2 = 1, p_{11} = 0.6$ and $p_{22} = 0.5$. The model is denoted by AR-ARG1, which indicates that the first regime is driven by an AR(1) process and the second regime by a heteroskedastic autoregressive process with $\gamma = 1$ (i.e., ARG1). We test whether the simulated trajectories can be described by the model defined in Eq. (6), i.e., following specification II, and denoted here by AR-AR.

- **AR-E vs AR-LN** The trajectories are simulated from an MRS model following specification I, see (4) and (5), with an exponential distribution in the second regime, i.e., $F^2 \sim \exp(\lambda)$. The model is denoted here by AR-E and its parameters...
are given by: \( \alpha = 10, \beta = 0.6, \sigma^2 = 10, \lambda = 30, p_{11} = 0.6 \) and \( p_{22} = 0.5 \).

We test whether the simulated trajectories can be driven by a model following specification I with a log-normal distribution in the second regime (i.e., AR-LN).

- **CIR-LN vs AR-G** The trajectories are simulated from an MRS model defined as:

\[
X_{t,1} = \alpha_1 + (1 - \beta_1)X_{t-1,1} + \sigma_1 \sqrt{X_{t-1,1}}\epsilon_t,
\]

\[
X_{t,2} \sim LN(\alpha_2, \sigma_2^2),
\]

where \( \alpha_1 = 1, \beta_1 = 0.8, \sigma_1^2 = 0.5, \alpha_2 = 2, \sigma_2^2 = 0.5, p_{11} = 0.6 \) and \( p_{22} = 0.5 \), i.e., the first regime is a discrete time version of the square root process, also known as the CIR process (Cox et al. 1985), and the second is a log-normal random variable. Hence, the name CIR-LN. We test whether the simulated trajectories can be driven by a model following specification I with a Gaussian distribution in the second regime.

The test results are summarized in Table 4. The values obtained for the individual regimes are also provided, however, as the simulated and estimated models differ, these rejection rates are highly dependent on the classification of observations to the regimes during estimation. Therefore, in the discussion that follows we focus on the test results for the whole models. Comparing the power of the Monte Carlo approach with the one using K-S test tables, we observe that in most cases the latter method yields lower (or worse) rejection percentages. This is in compliance with the results obtained in Sect. 4.2. The only significant deviations from this pattern can be observed for the AR-E vs. AR-LN test scenario, when the true model (i.e., AR-E)—and hence the simulated trajectories—exhibit a low degree of regime separation. The latter is manifested by relatively high RCM values of 38.23—41.69 (averaged over 500 simulated trajectories for each of the three sample sizes), compared to RCM values of 8.07—8.28 and 13.68—15.81 for the first and third scenarios, respectively. Note, that the reported RCM values were computed for the true models fitted to the 500 simulated trajectories from these models.

Looking at the MC simulation results obtained for the largest samples of \( T = 2,000 \) observations, we can see that in almost all cases the false hypothesis was rejected. The lowest rejection rate for the ewedf approach was 0.7580, for the wedf approach—0.9620 and for the Rosenblatt transformation—0.9900. All three lowest rates were obtained for the challenging AR-E vs. AR-LN test scenario. However, if the samples are smaller, the power of the tests apparently decreases. The sample size of \( T = 100 \) observations does not seem to be enough, especially if the degree of regime separation is low, like in the AR-E vs. AR-LN scenario. This is not the case if the definitions of both regimes are significantly different, as for the CIR-LN vs. AR-G scenario, for which the power is satisfactory even if \( T = 100 \). Comparing the ewedf and wedf approaches, we can observe that the latter yields higher on average rejection rates. A comparison of the wedf and the Rosenblatt transformation-based approach does not yield such a clear picture. For the AR-ARG1 vs. AR-AR scenario the power of the Rosenblatt transformation-based approach is higher than of the wedf approach, for the AR-E vs. AR-LN scenario it is only slightly higher, while for the CIR-LN...
Table 4  Percentages of rejected hypotheses $H_0$ at the 5% significance level for the alternative models with parameters estimated for each of the 500 simulated trajectories of $T = 100$, 500 or 2000 observations

| $T$ | Regime | ewedf | wedf | Rtrans |
|-----|--------|-------|------|--------|
|     | First  | Second| Model| First  | Second| Model| Model |
| AR-ARG1 vs. AR-AR |        |       |      |        |       |      |        |
| 2,000 | K-S tab. | 0.6420 | 1.0000 | 1.0000 | 0.0300 | 1.0000 | 1.0000 | 0.9960 |
|       | MC sim.  | 0.0520 | 1.0000 | 1.0000 | 0.4000 | 1.0000 | 1.0000 | 1.0000 |
| 500  | K-S tab. | 0.1160 | 0.6100 | 0.5060 | 0.0060 | 0.4440 | 0.0860 | 0.0280 |
|       | MC sim.  | 0.0960 | 0.8280 | 0.8480 | 0.1840 | 0.9840 | 0.8780 | 0.9740 |
| 100  | K-S tab. | 0.0160 | 0.0120 | 0.0120 | 0.0020 | 0.0100 | 0.0080 | 0.0000 |
|       | MC sim.  | 0.1140 | 0.1520 | 0.2080 | 0.0800 | 0.2120 | 0.1580 | 0.2580 |
| AR-E vs. AR-LN |        |       |      |        |       |      |        |
| 2,000 | K-S tab. | 0.0820 | 0.9980 | 0.9980 | 0.0620 | 0.9980 | 0.7460 | 0.9820 |
|       | MC sim.  | 0.0300 | 0.9680 | 0.7580 | 0.0560 | 0.9980 | 0.9620 | 0.9900 |
| 500  | K-S tab. | 0.0660 | 0.9780 | 0.5700 | 0.0660 | 0.2140 | 0.0580 | 0.1060 |
|       | MC sim.  | 0.0800 | 0.2120 | 0.2320 | 0.0940 | 0.8960 | 0.2760 | 0.4000 |
| 100  | K-S tab. | 0.0260 | 0.2780 | 0.0240 | 0.0300 | 0.0060 | 0.0080 | 0.0140 |
|       | MC sim.  | 0.0540 | 0.0180 | 0.0880 | 0.0960 | 0.1620 | 0.1260 | 0.1480 |
| CIR-LN vs. AR-G |        |       |      |        |       |      |        |
| 2,000 | K-S tab. | 0.9900 | 0.9180 | 0.9900 | 0.9900 | 0.9400 | 0.9900 | 0.9900 |
|       | MC sim.  | 0.9900 | 0.9880 | 0.9900 | 0.9900 | 0.9900 | 0.9900 | 0.9900 |
| 500  | K-S tab. | 0.9940 | 0.1600 | 0.9900 | 0.9940 | 0.1880 | 0.9860 | 0.6840 |
|       | MC sim.  | 0.9920 | 0.7460 | 0.9860 | 0.9940 | 0.8120 | 0.9940 | 0.9940 |
| 100  | K-S tab. | 0.1860 | 0.0260 | 0.0760 | 0.2520 | 0.0340 | 0.1820 | 0.0200 |
|       | MC sim.  | 0.3200 | 0.2440 | 0.3760 | 0.7440 | 0.2920 | 0.7920 | 0.5560 |

The parameters of the simulated trajectories are arbitrarily chosen, see text for details. The results of the tests based on the ewedf as well as the wedf approaches are reported independently for the two regimes (First, Second) and the whole model (Model). Results obtained using the Rosenblatt transformation (Rtrans) are reported in the last column. The tests utilize K-S test tables (K-S tab.) or Monte Carlo simulations (MC sim.) with $N = 500$ repetitions.
### Table 5  Percentages of rejected hypotheses \( H_0 \) at the 5% significance level for the alternative models with parameters estimated for each of the 500 simulated trajectories of \( T = 100, 500 \) or 2,000 observations

| \( T \) | Regime | ewedf | wedf | Rtrans |
|-------|--------|-------|------|--------|
|       | First  | Second| Model| Model  |
|       | K-S tab.|       |      |        |
| 2,000 | 0      | 0.2300| 0.0080| 0      | 0      | 0      | 0      |
| 500  | 0.0560 | 0.0200| 0.0720| 0.0660 | 0.0440 | 0.0700 | 0.0560 |
| 100  | 0.0080 | 0.0620| 0.0160| 0      | 0      | 0      | 0      |
| AR-ARG1 vs. AR-AR | MC sim. | 0.0260 | 0.0020| 0.0520| 0.0480 | 0.0160| 0.0560 | 0.0460 |
| 2,000 | 0.0120 | 0.0380| 0.0040| 0      | 0      | 0      | 0      |
| 500  | 0.0280 | 0.0320| 0.0840| 0.0380 | 0.0040| 0.0560 | 0.0660 |
| 100  | 0.0260 | 0.0020| 0.0520| 0.0480 | 0.0160| 0.0560 | 0.0460 |
| AR-E vs. AR-LN | K-S tab. | 0.9980| 0.7420| 0.9980| 0.9980| 0.7700| 0.9980| 0.9980 |
| 2,000 | 0.9980 | 0.4500| 0.9980| 0.9980| 0.8440| 0.9980| 0.9980 |
| 500  | 0.8600 | 0.2000| 0.8600| 0.8640| 0.1620| 0.8640| 0.8600 |
| 100  | 0.8740 | 0.1620| 0.8740| 0.8740| 0.3480| 0.8760| 0.8760 |
| CIR-LN vs. AR-G | MC sim. | 0.3200| 0.0020| 0.3200| 0.3200| 0.0020| 0.3200| 0.3140 |
| 2,000 | 0.3560 | 0.0620| 0.3660| 0.3520| 0.0380| 0.3600| 0.3560 |
| 500  | 0.3560 | 0.0620| 0.3660| 0.3520| 0.0380| 0.3600| 0.3560 |
| 100  | 0.0100 | 0.0100| 0.0100| 0.0180| 0.0100| 0.0080| 0.0160 |
| MC sim. | 0.0440 | 0.0540| 0.1360| 0.0740| 0.0700| 0.1540| 0.1520 |

This time the parameters of the simulated trajectories are estimated from the NEPOOL log-prices studied in Sect. 5. The results of the tests based on the ewedf as well as the wedf approaches are reported independently for the two regimes (First, Second) and the whole model (Model). Results obtained using the Rosenblatt transformation (Rtrans) are reported in the last column. The tests utilize K-S test tables (K-S tab.) or Monte Carlo simulations (MC sim.) with \( N = 500 \) repetitions.

parameters) the rejection percentages for this scenario have substantially decreased. Now they also do not increase with sample size, which may indicate problems with reaching the global maximum within the estimation procedure for small samples. This is not unexpected when conducting Monte Carlo studies of Markov regime-switching models as the likelihood surface can be highly irregular and contain several local maxima (Smith 2008).

### 5 Application to electricity spot prices

Now, we are ready to apply the new goodness-of-fit technique to electricity spot price models. We analyze the mean daily (baseload) day-ahead spot prices from the New England Power Pool SEMASS area (NEPOOL; US). The sample totals 1,827 daily...
Fig. 1  Mean daily (baseload) day-ahead spot prices from the New England Power Pool SEMASS area (NEPOOL; US) from the 5-year period January 2, 2006–January 2, 2011

observations (or 261 full weeks) and covers the 5-year period January 2, 2006–January 2, 2011, see Fig. 1. It is well known that electricity spot prices exhibit several characteristic features (Benth et al. 2008; Eydeland and Wolyniec 2012; Huisman 2009; Weron 2006), which have to be taken into account when modeling such processes. These include seasonality on the annual, weekly and daily level, mean reversion and price spikes. To cope with the seasonality we use the standard time series decomposition approach and let the electricity spot price $P_t$ be represented by a sum of two independent parts: a predictable (seasonal) component $f_t$ and a stochastic component $X_t$, i.e., $P_t = f_t + X_t$. Further, to address the mean-reverting and spiky behavior we let the log-prices, i.e., $Y_t = \log(X_t)$, be driven by:

- a 2-regime MRS model with mean-reverting, see (4), base regime ($R_t = 1$) and i.i.d. shifted log-normally distributed spikes ($R_t = 2$) or
- a 3-regime MRS model with mean-reverting, see (4), base regime ($R_t = 1$), i.i.d. shifted log-normally distributed spikes ($R_t = 2$) and i.i.d. drops ($R_t = 3$) distributed according to the inverted shifted log-normal law.

Note that we use here the independent regime specification (i.e., type I), since in wholesale power markets the price spikes (‘Second regime’) or price drops (i.e., negative price spikes; ‘Third regime’) are driven by unexpected but transient changes of market conditions. A generation outage or a severe heat wave typically do not last longer than a few hours or a few days; when it is over the prices move back to the normal level (‘Base regime’), usually irrespective of the magnitude of the extreme prices a few hours or days earlier (De Jong 2006; Janczura et al. 2012; Mari 2008).

Furthermore recall, that $X$ follows the shifted log-normal law or inverted shifted log-normal law if $\log(X - q)$, respectively $\log(q - X)$, has a Gaussian distribution. The cutoff level $q$ can be different for the spike and the drop regime, however, here—motivated by the results of Janczura and Weron (2013)—we set it to the first and the third quartile of the dataset for drops and spikes, respectively. Using shifted log-normal distributions allows to increase the degree of separation as measured by RCM and, hence, increase the power of the tests (see Sect. 4). For instance, the 2-regime model now yields an RCM of 4.79, compared to 12.81 for Sim #3 in Table 2. It also leads to more fundamentally justified models since electricity price spikes are generally
connected with scheduling units with higher marginal costs (like gas turbines, see e.g., Eydeland and Wolyniec 2012; Weron 2006).

Finally note, that such simple one-factor models may not be complex enough to address all features of electricity prices. In particular, the electricity forward prices implied by these spot price models exhibit the so-called Samuelson effect (i.e., a decrease in volatility with increasing time to maturity; for the considered models the volatility scales as $e^{-\beta(T-t)}$), but the rate of decrease is completely determined by the speed of mean-reversion $\beta$ (Janczura and Weron 2012). However, the rate of decrease should be large only for maturities up to a year (Kiesel et al. 2009). Perhaps, incorporating another stochastic factor would lead to a more realistic forward price curve.

Following Weron (2009) the deseasonalization is conducted in three steps; for a thorough study of modeling seasonal components in electricity spot prices we refer to Janczura et al. (2012). First, the long-term seasonal component (LTSC) $T_t$ is estimated from daily spot prices $P_t$ using a wavelet filter-smoother of order 6. A single non-parametric LTSC is used here to represent the long-term non-periodic fuel price levels, the changing climate/consumption conditions throughout the years and strategic bidding practices. As shown by Janczura and Weron (2010), the wavelet-estimated LTSC pretty well reflects the ‘average’ fuel price level, understood as a combination of natural gas, crude oil and coal prices; see also Eydeland and Wolyniec (2012) and Karakatsani and Bunn (2010) for a treatment of fundamental and behavioral drivers of electricity prices. On the other hand, as discussed recently in Janczura and Weron (2012), the use of the wavelet-based LTSC is somewhat controversial. Predicting it beyond the next few weeks is a difficult task, because individual wavelet functions are quite localized in time or (more generally) in space. Preliminary research suggests, however, that despite this feature the wavelet-based LTSC can be extrapolated into the future yielding a better on-average prediction of the level of future spot prices than an extrapolation of a sinusoidal LTSC (Nowotarski et al. 2011).

The price series without the LTSC is obtained by subtracting the $T_t$ approximation from $P_t$. Next, the weekly periodicity $s_t$ is removed by subtracting the ‘average week’ calculated as the mean of prices corresponding to each day of the week (the US national holidays are treated as the 8-day of the week). Finally, the deseasonalized prices, i.e., $X_t = P_t - T_t - s_t$, are shifted so that the minimum of the new process $X_t$ is the same as the minimum of $P_t$. The resulting deseasonalized time series can be seen in Figs. 2, 3. The estimated model parameters are presented in Table 6.

For both analyzed MRS models, tests based on the ewedf, the wedf and the Rosenblatt transformation are performed. The $p$ values in Table 7 are reported both for the standard approach utilizing the K-S test tables (which generally leads to overestimated $p$ values) and the much slower but more accurate Monte Carlo setup. The testing procedure in the Monte Carlo case is analogous to the one used in the simulation study, see Sect. 4.2 for a detailed description. Again, in order to verify the ‘whole model’ goodness-of-fit, we transform the spike and drop regime observations so that both samples are $N(0, 1)$-distributed.

For the 2-regime model the $p$ values obtained from the K-S test tables indicate that the model cannot be rejected at the 5% significance level. However, the base regime and the model $p$ values are still quite low—especially for the wedf approach—so the...
conclusions of the test should be verified with the Monte Carlo simulations. Indeed, for the Monte Carlo based test only the spike regime yields a satisfactory fit, as the p value is well above the 5% significance level. The base regime, as well as the whole model distribution, can be rejected at any reasonable level for the wddf and at the 5% level for the Rosenblatt transformation. Apparently, the base regime process cannot model the sudden drops in the NEPOOL log-prices. However, if a third regime (modeling price drops) is introduced, the MRS model yields a satisfactory fit. In the 3-regime case none of the tests can be rejected at the 5% significance level. Regime classification for the 2- and 3-regime models can be observed in Figs. 2 and 3, respectively.
Table 6  Parameters of the 2- and 3-regime MRS models with mean-reverting base regime and independent spikes and drops driven by shifted log-normal laws fitted to the deseasonalized NEPOOL log-prices

| Model      | Base regime | Spike regime | Drop regime | Probabilities |
|------------|------------|-------------|-------------|--------------|
|            | \(\alpha\) | \(\beta\)  | \(\sigma^2\) | \(\alpha_2\) | \(\sigma^2_2\) | \(\alpha_3\) | \(\sigma^2_3\) | \(p_{11}\) | \(p_{22}\) | \(p_{33}\) |
| 2-Regime   | 0.71       | 0.21        | 0.0060      | −1.57        | 0.32           | −           | −           | 0.98       | 0.74       | −           |
| 3-Regime   | 0.99       | 0.29        | 0.0053      | −1.65        | 0.33           | −1.93       | 0.19        | 0.96       | 0.76       | 0.89        |

Regime classification for the 2- and 3-regime models can be observed in Figs. 2 and 3, respectively.

Table 7  The \(p\) values of the K-S test based on the ewedf, the wedf and the Rosenblatt transformation (Rtrans) for the 2- and 3-regime MRS models of the deseasonalized NEPOOL log-prices

| Regime      | ewedf Base | ewedf Spike | ewedf Drop | wedf Base | wedf Spike | wedf Drop | Rtrans Base | Rtrans Spike | Rtrans Drop |
|-------------|------------|-------------|------------|----------|------------|----------|-------------|-------------|-------------|
| 2-Regime    | 0.15       | 0.54        | −          | 0.2170   | 0.05       | 0.87     | 0.0960      | 0.2850      |
| K-S tab.    | 0.00       | 0.34        | −          | 0.0240   | 0.00       | 0.50     | 0.0020      | 0.0470      |
| MC sim.     | 0.08       | 0.26        | 0.91       | 0.0780   | 0.08       | 0.35     | 0.0820      | 0.1190      |
| 3-Regime    | 0.42       | 0.58        | 0.99       | 0.3950   | 0.29       | 0.78     | 0.3550      | 0.3900      |
| K-S tab.    | 0.08       | 0.26        | 0.91       | 0.0780   | 0.08       | 0.35     | 0.0820      | 0.1190      |
| MC sim.     | 0.08       | 0.26        | 0.91       | 0.0780   | 0.08       | 0.35     | 0.0820      | 0.1190      |

The test utilizes K-S test tables (K-S tab.) or Monte Carlo simulations (MC sim.) with \(N = 1000\) repetitions. Values exceeding the 5% threshold are emphasized in bold.

6 Conclusions

While most of the electricity spot price models proposed in the literature are elegant, their fit to empirical data has either been not examined thoroughly or the signs of a bad fit ignored. As the empirical study of Sect. 5 has shown, even reasonably looking and popular models should be carefully tested before they are put to use in trading or risk management departments. The goodness-of-fit wedf-based test introduced in Sect. 3.2.2 provides an efficient tool for accepting or rejecting a given Markov regime-switching (MRS) model for a particular data set. While its performance (including power; see Sect. 4.3) is similar to that of the Rosenblatt transformation-based approach, it provides an edge over the latter by yielding \(p\) values for the individual regimes. For instance, this allows to observe that in the 3-regime model the worst fit is obtained for the base regime. Perhaps the simple AR(1) structure is not enough to model the complex dynamics of electricity spot prices in the relatively calm, non-spiky periods.

However, in this paper we have not restricted ourselves to MRS models but pursued a more general goal. Namely, we have proposed a goodness-of-fit testing scheme for the marginal distribution of regime-switching models, including variants with an observable and with a latent state process. For both specifications we have described the testing procedure. The models with a latent state process (i.e., MRS models) required the introduction of the concept of the weighted empirical distribution function (wedf) and a generalization of the Kolmogorov–Smirnov test to yield an efficient testing tool.
We have focused on two commonly used specifications of regime-switching models in the energy economics literature—one with dependent autoregressive states and a second one with independent autoregressive and i.i.d. regimes. Nonetheless, the proposed approach can be easily applied to other specifications of regime-switching models (for instance, to 3-regime models with heteroscedastic base regime dynamics; see Janczura and Weron 2010). Very likely it can be also extended to other goodness-of-fit edf-type tests, like the Anderson–Darling. As the latter puts more weight to the observations in the tails of the distribution than the Kolmogorov–Smirnov test, it might be more discriminatory and provide a better testing tool for extremely spiky data. Future work will be conducted in this direction.

Finally note, that a good in-sample fit does not necessarily imply a good forecast behavior. Although Kosater and Mosler (2006) found for German electricity price data that for long run point forecasts (30–80 days ahead) an MRS model with regimes driven by two AR(1) processes was slightly more accurate than a simple AR(1) model, for shorter time horizons both model classes performed alike. It remains an open question how do the MRS models fitted to NEPOOL log-prices in Sect. 5 perform in terms of forecasting. The adequacy of MRS models for forecasting in general has been questioned by Besssec and Bouabdallah (2005). However, as Weron and Misiorek (2008) have shown, regime-switching models may behave better than their linear competitors in volatile periods. They might also have an edge in density forecasts, but this has to be verified yet.

Acknowledgments We thank two anonymous reviewers for their comments and suggested improvements. This paper has also benefited from conversations with the participants of the DStatG 2010 Annual Meeting, the Trondheim Summer 2011 Energy Workshop, the 2011 WPI Conference in Energy Finance, the Energy Finance Christmas Workshop (EFC11), the 9th International Conference ‘European Energy Market’ (EEM12) and the seminars at Macquarie University, National University of Singapore and University of Poznań. This work was supported by funds from the National Science Centre (NCN, Poland) through grant no. 2011/01/B/HS4/01077.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

Appendix

Proof (Lemma 1) Let

\[ F_n(x) = \frac{1}{n} \sum_{t=1}^{n} \left[ P(R_t = 1|\mathcal{X}_T)\mathbb{I}_{\{y_t^1 < x\}} + P(R_t = 2|\mathcal{X}_T)\mathbb{I}_{\{y_t^2 < x\}} \right] \quad (23) \]

and observe that

\[
E[F_n(x)|\mathcal{R}]
= \frac{1}{n} \sum_{t=1}^{n} \left( E[E(R_t = 1|\mathcal{R}_t = 1, \mathcal{X}_t)|\mathcal{R}_t = 1, \mathcal{X}_T] + E[E(R_t = 2|\mathcal{R}_t = 2, \mathcal{X}_t)|\mathcal{R}_t = 2, \mathcal{X}_T] \right)
\]

 Springer
\[ \begin{aligned}
&= \frac{1}{n} \sum_{t=1}^{n} \left( \mathbb{I}_{\{R_t=1\}} E \left[ E \left( \mathbb{I}_{\{R_t=1\}} | x_T \right) \mathbb{I}_{\{y_{1,t}^l < x\}} | \{R_t = 1\} \right] + \mathbb{I}_{\{R_t=2\}} E \left[ E \left( \mathbb{I}_{\{R_t=2\}} | x_T \right) \mathbb{I}_{\{y_{1,t}^r < x\}} | \{R_t = 2\} \right] \right) \\
&= \frac{1}{n} \sum_{t=1}^{n} \left( \mathbb{I}_{\{R_t=1\}} (24) F(x) + \mathbb{I}_{\{R_t=2\}} F(x) \right) = F(x),
\end{aligned} \]

where \( F(x) \) is the distribution of the residuals and \( \mathcal{R} \) is the \( \sigma \)-algebra generated by the state process values. Similarly,

\[ \text{Var}[F_n(x)|\mathcal{R}] = \frac{1}{n^2} \sum_{t=1}^{n} \left[ E \left[ \left( E \left( \mathbb{I}_{\{R_t=1\}} | x_T \right) \mathbb{I}_{\{y_{1,t}^l < x\}} + E \left( \mathbb{I}_{\{R_t=2\}} | x_T \right) \mathbb{I}_{\{y_{1,t}^r < x\}} \right)^2 \right] \right] - [F(x)]^2 \]

\[ = \frac{1}{n^2} \sum_{t=1}^{n} \left( \mathbb{I}_{\{R_t=1\}} E[\mathbb{I}_{\{y_{1,t}^l < x\}}^2 | \{R_t = 1\}] + \mathbb{I}_{\{R_t=2\}} E[\mathbb{I}_{\{y_{1,t}^r < x\}}^2 | \{R_t = 2\}] \right) - [F(x)]^2 \]

\[ = \frac{1}{n} F(x)[1 - F(x)]. \]

Next, from the conditional Kolmogorov inequality (for details see Prakasa Rao 2009), for any \( \delta > 0 \) we have (a.s.):

\[ P \left( \left| F_n(x) - E[F_n(x) | \mathcal{R}] \right| > \delta \right| \mathcal{R} \) \leq \frac{\text{Var}[F_n(x)|\mathcal{R}]}{\delta^2} \]

\[ = \frac{F(x)[1 - F(x)]}{n\delta^2}. \]

As a consequence,

\[ P(|F_n(x) - F(x)| > \delta) = E \left[ P \left( \left| F_n(x) - E[F_n(x) | \mathcal{R}] \right| > \delta \right| \mathcal{R} \right) \]

\[ \leq \frac{F(x)[1 - F(x)]}{n\delta^2} \]

and \( F_n(x) \) converges in probability to \( F(x) \) as \( n \to \infty \).
Proof [Theorem 1] First, note that $F(x) \in [0, 1]$ implies $F_n(x) = F(x)$ and
\[ \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \sup_{x \in D} |F_n(x) - F(x)|, \]
where $D = \mathbb{R} \setminus \{x : F(x) = 0 \lor F(x) = 1\}$. Therefore, in the following we will limit ourselves to the case $0 < F(x) < 1$.

Let
\[ Y_t = \frac{P(R_t = 1|Y_t^i < x)}{n} + \frac{P(R_t = 2|Y_t^i < x)}{n}, \]
where $Y_t^i$ are the transformed variables of the $i$th regime, i.e., $Y_t^i = [x_{t+1} - \alpha_i - (1 - \beta_i)x_t]/\sigma_i$. Observe that, given $R_t = i$, $Y_t = \frac{1}{n} \mathbb{I}_{\{Y_t^i < x\}}$ and $Y_t^i$ becomes the residual of the $i$th regime. Therefore, $Y_1, Y_2, \ldots, Y_T$ are $\mathcal{F}$-independent, where $\mathcal{F}$ is the $\sigma$-algebra generated by the state process $\{R_t\}_{t=1,2,\ldots,T}$, and have the same conditional distribution. Hence, from the conditional version of the Central Limit Theorem (for details see Grzenda and Zieba 2008), we have:
\[ \sum_{t=1}^n Y_t - nE[Y_t|\mathcal{F}] \xrightarrow{d} \mathcal{N}(0, 1). \]  
(30)

Next, note that $E(Y_t|\mathcal{F}) = \frac{1}{n} F(x)$, see equation (24), and $\text{Var}(Y_t|\mathcal{F}) = \frac{1}{n^2} F(x)[1 - F(x)]$, see equation (25). Hence, (30) yields:
\[ \frac{F_n(x) - F(x)}{\sqrt{n} \sqrt{F(x)[1 - F(x)]}} \xrightarrow{d} \mathcal{N}(0, 1). \]  
(31)

The latter is equivalent to
\[ \sqrt{n}[F_n(x) - F(x)] \xrightarrow{d} W_{F(x)}, \]  
(32)
where $W_{Y^i}$ is a Brownian bridge, i.e., $W_{Y^i} \sim \mathcal{N}(0, y(1 - y))$, see e.g., Lehmann and Romano (2005, p. 585). Let $y = F(x)$ and observe that $\mathbb{I}_{\{Y_t^i < x\}} = \mathbb{I}_{\{Y_t^i < F^{-1}(y)\}} = \mathbb{I}_{\{F(y_t^i) < y\}}$. Moreover, if $y_t^i$ are $F$-distributed, then $F(y_t^i)$ are driven by the uniform distribution on $[0, 1]$. Formula (32) ensures that $Z_n(y) = \sqrt{n}[F_n(F^{-1}(y)) - y]$ converges pointwise to a Brownian bridge. In order to prove the convergence of $Z_n(y)$ in $D([0, 1])$, i.e., in the space of right continuous functions that have left-hand limits, it is enough to show: (1) the weak convergence of the finite dimensional distributions of $Z_n$ and that (2)
\[ E \left[ |Z_n(y) - Z_n(y_1)|^\gamma |Z_n(y_2) - Z_n(y)|^\gamma \right] \leq |g(y_2) - g(y_1)|^{2\alpha} \]  
(33)
for $y_1 < y < y_2$ and $n \geq 1$, where $\gamma \geq 0$, $\alpha > 1/2$ and $g$ is a non-decreasing, continuous function on $[0, 1]$ (see Theorem 15.6 in Billingsley 1968).

(1) Let $0 < y_1 < y_2 < 1$. We will show that $(Z_n(y_1), Z_n(y_2) - Z_n(y_1))$ converges weakly to $(W_{Y^1}, W_{Y^2} - W_{Y^1})$. First, observe that $(Z_n(y_1), Z_n(y_2) - Z_n(y_1))$ conditional on $\mathcal{F}$ is multinomially distributed with variances $(y_1(1 - y_1), (y_2 - y_1)(1 - (y_2 - y_1)))$.
and covariance $-y_1(y_2 - y_1)$. This can be calculated using the same arguments as in the proof of Lemma 1. Hence, by the central limit theorem for multinomial trials, as $n \to \infty$, for any $s \in \mathbb{R}^2$ we have

$$
\varphi(\mathbb{Z}_n(y_1), \mathbb{Z}_n(y_2) - \mathbb{Z}_n(y_1)) (s) \to \varphi(W_{y_1}^B, W_{y_2}^B - W_{y_1}^B) (s)
$$

(34)

where $\varphi_Z$ is the characteristic function of $Z$ and $\varphi^\mathbb{Z}_Z$ is the conditional (on $\mathcal{M}$) characteristic function of $Z$. Finally, by the dominated convergence theorem, $(\mathbb{Z}_n(y_1), \mathbb{Z}_n(y_2) - \mathbb{Z}_n(y_1))$ converges weakly to $(W_{y_1}, W_{y_2} - W_{y_1})$. The convergence of finite dimensional distributions for any $0 < y_1 < y_2 < \ldots < y_m < 1$ follows with the same arguments.

(ii) In order to prove that condition (33) is fulfilled, observe that, given $\mathbb{F}_n$ is the Kolmogorov–Smirnov distribution, i.e., a distribution of $\sup_0 \leq y \leq 1 |W_{y}^B|$.

Finally, by (i) and (ii) $Z_n(y)$ converges to $W^B(y)$ in $D([0, 1])$ and by the continuous mapping theorem we obtain that

$$
\sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \sup_{y \in [0, 1]} |Z_n(y)| \xrightarrow{d} KS,
$$

(36)

where $KS$ is the Kolmogorov–Smirnov distribution, i.e., a distribution of $\sup_0 \leq y \leq 1 |W_{y}^B|$.

Proof [Lemma 2] Let $I = \{i_1, i_2, \ldots, i_n : i_k \in \{1, 2\}\}$ and

$$
F^i_n(x) = \frac{\sum_{t=1}^{n} P(R_t = i | x_T) \mathbb{I}_{\{y^i_t < x\}}}{\sum_{t=1}^{n} P(R_t = i | x_T)}.
$$

(37)

We have

$$
E[F^i_n(x) | \mathcal{M}]
$$

$$
= \sum_{t=1}^{n} \left( E \left[ \frac{E(\mathbb{I}_{\{R_t = i\}} | x_T) \mathbb{I}_{\{y^i_t < x\}}}{\sum_{t=1}^{n} E(\mathbb{I}_{\{R_t = i\}} | x_T)} | \mathcal{M} \right] \right)
$$

$$
= \sum_{t=1}^{n} \sum_{\{i_1, i_2, \ldots, i_n\} \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \sum_{\{k : i_k = i\}} \sum_{\{l : i_l = i\}} E \left( \mathbb{I}_{\{y^i_t < x\}} | \{R_t = i\} \right)
$$

$$
= F(x) \sum_{\{i_1, i_2, \ldots, i_n\} \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \sum_{\{k : i_k = i\}} \sum_{\{l : i_l = i\}} E \left( \mathbb{I}_{\{y^i_t < x\}} | \{R_{i_k} = l\} \right)
$$

$$
= F^i(x) \sum_{\{i_1, i_2, \ldots, i_n\} \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} = F^i(x),
$$

(38)

where $\mathcal{M}$ is the $\sigma$-algebra generated by the state process values.
Similarly,

\[
\Var[F^i_n(x) \mid g_1] = \sum_{t=1}^{n} \left\{ \sum_{i_1, i_2, \ldots, i_n \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \frac{\mathbb{I}_{\{R_t = i\}}}{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}}^2 F^i_t(x) \left( F^i_t(x) - \frac{\mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}}}{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}}^2 \right)^2 \right\}
\]

\[
= \sum_{t=1}^{n} \left\{ \sum_{i_1, i_2, \ldots, i_n \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \frac{\mathbb{I}_{\{R_t = i\}}}{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}}^2 F^i_t(x) \left( F^i_t(x) - \frac{\mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}}}{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}}^2 \right)^2 \right\}
\]

\[
= F^i_t(x) [1 - F^i_t(x)] \sum_{i_1, i_2, \ldots, i_n \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \frac{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}^2}{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}}^2.
\]

(39)

Using the conditional Kolmogorov inequality we obtain (a.s.):

\[
P \left( \left| F^i_n(x) - E[F^i_n(x) \mid g_1] \right| > \delta \mid g_1 \right) \leq \frac{F^i_t(x) [1 - F^i_t(x)]}{\delta^2} \sum_{i_1, i_2, \ldots, i_n \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \frac{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}^2}{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}}^2.
\]

(40)

Now, taking the expected value and using the fact that \( E[F^i_n(x) \mid g_1] = F^i_t(x) \) we have

\[
P(|F^i_n(x) - F^i_t(x)| > \delta) \leq \frac{F^i_t(x) [1 - F^i_t(x)]}{\delta^2} \left( \sum_{i_1, i_2, \ldots, i_n \in I} \mathbb{I}_{\{R_1 = i_1, R_2 = i_2, \ldots, R_n = i_n\}} \frac{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}^2}{\sum_{\{k : i_k = i\}} \mathbb{I}_{\{R_k = i\}}}^2 \right) \]

\[
\leq \frac{F^i_t(x) [1 - F^i_t(x)]}{\delta^2} \left[ \max_{i, j = 1, 2} (p_{ij}) \right]^n \frac{1}{n} \left[ n \left( \frac{n - 1}{n - 1} \right) \frac{1}{n - 1} + \cdots + \left( \frac{n}{n - 1} \right) \right]
\]

\[
\leq \frac{F^i_t(x) [1 - F^i_t(x)]}{\delta^2} n^2.
\]

(41)
for \( q = \left[ \max_{i,j=1,2}(p_{ij}) \right]^n \). If \( q < 1 \), then \( F_n(x) \) converges in probability to \( F(x) \) for \( n \to \infty \).

**Proof** [Theorem 2] Let

\[
Y^j_t = \frac{P(R_t = j|x_T) \prod_{y^j_t < x}^n}{\sum_{i=1}^n P(R_t = j|x_T)}.
\]

Now, observe that the conditional characteristic function of \( \zeta = F^j_n(x) - E[F^j_n(x)|\mathcal{R}] \) is given by:

\[
\phi^\mathcal{R}_\zeta(s) = E \left[ \exp \left( is \frac{\sum_{t=1}^n Y^j_t - E(\sum_{t=1}^n Y^j_t)}{\sqrt{\text{Var}(F^j_n(x)|\mathcal{R})}} \right) | \mathcal{R} \right] = E \left[ \prod_{t=1}^n \exp \left( is \frac{Y^j_t - E(Y^j_t|x_T)}{\sqrt{\text{Var}(F^j_n(x)|\mathcal{R})}} \right) | \mathcal{R} \right],
\]

(42)

where \( \mathcal{R} \) is the \( \sigma \)-algebra generated by the state process values.

To focus attention, assume that there were \( m \) observations from regime \( j \), i.e., \( \sum_{t=1}^n \mathbb{I}_{(R_t=j)} = m \) and for notational convenience set \( R_t = j \ \forall t \in \{t_1, t_2, \ldots, t_m\} \). Consequently, \( \text{Var}(F^j_n(x)|\mathcal{R}) = m\text{Var}(Y_{t_l}|\mathcal{R}), \ l = 1, 2, \ldots, m. \) Hence, taking a time series expansion of the exponent we obtain

\[
\phi^\mathcal{R}_\zeta(s) = \prod_{t=t_1}^{t_m} \left( 1 + is \frac{Y^j_t - E(Y^j_t|x_T)}{\sqrt{m\text{Var}(Y^j_t|x_T)}} - s^2 \frac{\text{Var}(Y^j_t|x_T)}{2m\text{Var}(Y^j_t|x_T)} + o \left[ \frac{s^2}{m\text{Var}(Y^j_t|x_T)} \right] | \mathcal{R} \right)
= \left\{ 1 - \frac{s^2}{2m} + E \left[ o \left[ \frac{s^2}{m\text{Var}(Y^j_t|x_T)} \right] | \mathcal{R} \right] \right\}^m.
\]

(43)

Since for an ergodic Markov chain \( \sum_{t=1}^n \mathbb{I}_{(R_t=j)} \to \infty \) a.s. as \( n \to \infty \) and for a fixed \( s, \lim_{m \to \infty} E \left[ o \left[ \frac{s^2}{m\text{Var}(Y^j_t|x_T)} \right] | \mathcal{R} \right] = 0 \) a.s., we obtain that

\[
\phi^\mathcal{R}_\zeta(s) \to e^{-s^2/2} \quad \text{a.s. with} \ n \to \infty.
\]

(44)

Moreover, from (38) and (39) we have

\[
\frac{F^j_n(x) - E[F^j_n(x)|\mathcal{R}]}{\sqrt{\text{Var}[F^j_n(x)|\mathcal{R}]}} = w_n \frac{F^j_n(x) - F^j(x)}{F^j(x)[1 - F^j(x)]},
\]

(45)
where
\[ w_n = \sum_{\{i_1, i_2, \ldots, i_n\} \in I} \prod_{k=1}^{n} \frac{\prod \{R_k = i_k\}}{\prod \{R_k = i\}}. \]

Finally, from (44) and (45) we have
\[ w_n \left[ F_n(x) - F(x) \right] \xrightarrow{d} N(0, F(x)[1 - F(x)]) \quad (46) \]
and the proof is completed with the same arguments as in the proof of Theorem 1.

References

Ang, A., Bekaert, G.: Regime switches in interest rates. J. Bus. Econ. Stat. 20(2), 163–182 (2002)
Benth, F.E., Benth, J.S., Koekkbaker, S.: Stochastic Modeling of Electricity and Related Markets. World Scientific, Singapore (2008)
Berkowitz, J.: Testing density forecasts, with applications to risk management. J. Bus. Econ. Stat. 19, 465–474 (2001)
Bessec, M., Bouabdallah, O.: What causes the forecasting failure of Markov-switching models? A Monte Carlo study. Stud. Nonlinear Dyn. Econom. 9(2), Article 6 (2005)
Bierbrauer, M., Menn, C., Rachev, S.T., Trück, S.: Spot and derivative pricing in the EEX power market. J. Bank. Financ. 31, 3462–3485 (2007)
Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
Bulla, J., Berzel, A.: Computational issues in parameter estimation for stationary hidden Markov models. Comput. Stat. 23, 1–18 (2008)
Cappe, O., Moulines, E., Ryden, T.: Inference in Hidden Markov Models. Springer, Berlin (2005)
Celeux, G., Durand, J.B.: Selecting hidden Markov model state number with cross-validated likelihood. Comput. Stat. 23, 541–564 (2008)
Cetin, M., Comert, G.: Short-term traffic flow prediction with regime switching models. Transport. Res. Rec. J. Transport. Res. Board 1665, 23–31 (2006)
Cho, J.S., White, H.: Testing for regime switching. Econometrica 75(6), 1671–1720 (2007)
Choi, S. Regime-switching univariate diffusion models of the short-term interest rate. Stud. Nonlinear Dyn. Econom. 13(1), Article 4. (2009)
Čížek, P., Härdle, W., Weron, R. (eds.): Statistical Tools for Finance and Insurance, 2nd edn. Springer, Berlin (2011)
Cox, J.C., Ingersoll, J.E., Ross, S.A.: A theory of the term structure of interest rates. Econometrica 53, 385–407 (1985)
D’Agostino, R.B., Stevens, M.A. (eds.): Goodness-of-fit testing techniques. Marcel Dekker, New York (1986)
De Jong, C.: The nature of power spikes: A regime-switch approach. Stud. Nonlinear Dyn. Econ. 10(3), Article 3 (2006)
Diebold, F.X., Gunther, T.A., Tay, T.S.: Evaluating density forecasts with applications to financial risk management. Int. Econ. Rev. 39, 863–883 (1998)
Erlwein, C., Benth, F.E., Mamon, R.: HMM filtering and parameter estimation of an electricity spot price model. Energy Econ. 32, 1034–1043 (2010)
Ethier, R., Mount, T.: Estimating the volatility of spot prices in restructured electricity markets and the implications for option values. PSec Working Paper 98–31 (1998)
Eydeland, A., Wolnyiec, K.: Energy and Power Risk Management, 2nd edn. Wiley, Hoboken, NJ. (2012)
Garcia, R.: Asymptotic null distribution of the likelihood ratio test in Markov switching models. Int. Econ. Rev. 39, 763–788 (1998)
Grzenda, W., Zieba, W.: Conditional Central Limit Theorem. Int. Math. Forum 3(31), 1521–1528 (2008)
Haas, M., Mittnik, S., Paolella, M.S.: A new approach to Markov-switching GARCH models. J. Financ. Econom. 2(4), 493–530 (2004)
Haesler, E., Mason, D.M., Newton, M.A.: Weighted bootstrapping of the means. CWI Q. 213–228 (1991)
Hamilton, J.: Analysis of time series subject to changes in regime. J. Econom. 45, 39–70 (1990)
Hamilton, J.: Specification testing in Markov-switching time series models. J. Econom. 70, 127–157 (1996a)
Hamilton, J.: Regime switching models. In: The New Palgrave Dictionary of, Economics, 2nd edn. Macmillan, London (1996b)
Hirsch, G.: Pricing of hourly exercisable electricity swing options using different price processes. J. Energy Mark. 2(2), 3–46 (2009)
Hu, L., Shin, Y.: Optimal test for Markov switching GARCH models. Stud. Nonlinear Dyn. Econ. 12(3), Article 3 (2008)
Huang, M.L., Brill, P.H.: A distribution estimation method based on level crossings. J. Stat. Plan. Inference 124, 45–62 (2004)
Huisman, R., de Jong, C.: Option pricing for power prices with spikes. Energy Power Risk Manag. 7(11), 12–16 (2003)
Huisman, R.: An Introduction to Models for the Energy Markets. Risk Books (2009)
Janczura, J., Weron, R.: Regime switching models for electricity spot prices: Introducing heteroskedastic base regime dynamics and shifted spike distributions. In: IEEE Conference Proceedings (EEM’09). doi:10.1109/EEM.2009.5207175. Available at MPRA: http://mpra.ub.uni-muenchen.de/18784 (2009)
Janczura, J., Weron, R.: An empirical comparison of alternate regime-switching models for electricity spot prices. Energy Econ. 32(5), 1059–1073 (2010)
Janczura, J., Weron, R.: Efficient estimation of Markov regime-switching models: An application to electricity wholesale market prices. ASTA-Adv. Stat. Anal. 96(3), 385–407 (2012)
Janczura, J., Weron, R.: Inference for Markov-regime switching models of electricity spot prices. In: Quantitative Energy Finance, Benth, F.E., Laurence, P., Kholdnyi, V. eds. Springer, forthcoming. Working paper version available from RePEc: http://ideas.repec.org/p/wuu/wpaper/hsc1201.html (2013)
Janczura, J., Trück, S., Weron, R., Wolff, R.: Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. Working paper. Available at MPRA: http://mpra.ub.uni-muenchen.de/39277 (2012)
Karacatsani, N.V., Bunn, D.: Forecasting electricity prices: The impact of fundamentals and time-varying coefficients. Int. J. Forecast. 24, 764–785 (2008)
Karacatsani, N.V., Bunn, D.: Fundamental and behavioural drivers of electricity price volatility. Stud. Nonlinear Dyn. Econom. 14(4), Article 4 (2010)
Kiesel, R., Schindlmayr, G., Börger, R.H.: A two-factor model for the electricity forward market. Quant. Financ. 9(3), 279–287 (2009)
Kim, C.-J.: Dynamic linear models with Markov-switching. J. Econom. 60, 1–22 (1994)
Kim, S., Shephard, N., Chib, S.: Stochastic volatility: Likelihood inference and comparison with ARCH models. Rev. Econ. Stud. 65(3), 361–393 (1998)
Kosater, P., Mosler, K.: Can Markov regime-switching models improve power-price forecasts? Evidence from German daily power prices. Appl. Energy 83, 943–958 (2006)
Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer, New York (2005)
Luo, Q., Mao, X.: Stochastic population dynamics under regime switching. J. Math. Anal. Appl. 334(1), 69–84 (2007)
Lux, T., Morales-Arias, L.: Forecasting volatility under fractality, regime-switching, long memory and student-t innovations. Comput. Stat. Data Anal. 54, 2676–2692 (2010)
Maiboroda, R.E.: Estimates for distributions of components of mixtures with varying concentrations. Ukr. Math. J. 48(4), 618–622 (1996)
Maiboroda, R.E.: A test for the homogeneity of mixtures with varying concentrations. Ukr. Math. J. 52(8), 1256–1263 (2000)
Mari, C.: Random movements of power prices in competitive markets: A hybrid model approach. J. Energy Mark. 1(2), 87–103 (2008)
Misiorek, A., Trueck S., Weron R.: Point and interval forecasting of spot electricity prices: linear vs. non-linear time series models. Stud. Nonlinear Dyn. Econ. 10(3), Article 2 (2006)
Nowotarski, J., Tomczyk, J., Weron, R.: Wavelet-based modeling and forecasting of the seasonal component of spot electricity prices. Paper presented at the Energy Finance Christmas Workshop (EFC11), Wroclaw, Dec. 19–20, 2011.
Prakasa Rao, B.L.S.: Conditional independence, conditional mixing and conditional association. Ann. Inst. Stat. Math. 61, 441–460 (2009)
Rosenblatt, M.: Remarks on a multivariate transformation. Ann. Math. Stat. 23(3), 470–472 (1952)
Ross, S.: Simulation. Academic Press, San Diego (2002)
Sen, R., Hsieh, F.: A note on testing regime switching assumption based on recurrence times. Stat. Probab. Lett. 79, 2443–2450 (2009)
Smith, D.R.: Evaluating specification tests for Markov-switching time-series models. J. Time Ser. Anal. 29(4), 629–651 (2008)
Vasas, K., Eleka, P., Markusa, L.: A two-state regime switching autoregressive model with an application to river flow analysis. J. Stat. Plan. Inference 137(10), 3113–3126 (2007)
Weron, R.: Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. Wiley, Chichester (2006)
Weron, R.: Heavy-tails and regime-switching in electricity prices. Math. Methods Oper. Res. 69(3), 457–473 (2009)
Weron, R., Misiorek, A.: Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. Int. J. Forecast. 24, 744–763 (2008)
Withers, C.S., Nadarajah, S.: The distribution and quantiles of functionals of weighted empirical distributions when observations have different distributions. arXiv:1002.4338v1 (2010)