On Robust Estimation of Low-Frequency Variability Trends in Discrete Markovian Sequences of Atmospheric Circulation Patterns

ILLIA HORENKO

Institute of Mathematics, Free University of Berlin, Berlin, Germany

(Manuscript received 29 September 2008, in final form 13 December 2008)

ABSTRACT

Identification and analysis of temporal trends and low-frequency variability in discrete time series is an important practical topic in the understanding and prediction of many atmospheric processes, for example, in analysis of climate change. Widely used numerical techniques of trend identification (like local Gaussian kernel smoothing) impose some strong mathematical assumptions on the analyzed data and are not robust to model sensitivity. The latter issue becomes crucial when analyzing historical observation data with a short record. Two global robust numerical methods for the trend estimation in discrete nonstationary Markovian data based on different sets of implicit mathematical assumptions are introduced and compared here. The methods are first compared on a simple model example; then the importance of mathematical assumptions on the data is explained and numerical problems of local Gaussian kernel smoothing are demonstrated. Presented methods are applied to analysis of the historical sequence of atmospheric circulation patterns over the United Kingdom between 1946 and 2007. It is demonstrated that the influence of the seasonal pattern variability on transition processes is dominated by the long-term effects revealed by the introduced methods. Despite the differences in the mathematical assumptions implied by both presented methods, almost identical symmetrical changes of the cyclonic and anticyclonic pattern probabilities are identified in the analyzed data, with the confidence intervals being smaller than in the case of the local Gaussian kernel smoothing algorithm. Analysis results are investigated with respect to model sensitivity and compared to a standard analysis technique based on a local Gaussian kernel smoothing. Finally, the implications of the discussed strategies on long-range predictability of the data-fitted Markovian models are discussed.

1. Introduction

Many real-life processes can be described as simplified discrete models switching between a finite number of states or regimes. Such processes can be found in biophysics (transitions between different conformations in biomolecules; Schütte and Huisinga 2003), in computational finance (transitions between different market phases; Hamilton 1989) and in weather/climate research (transitions between different atmospheric circulation regimes; e.g., Majda et al. 2006; Horenko 2008; Horenko et al. 2008b,a).

However, most of the available time series data from such processes share the two following properties: (i) the data have only a short record (because observations are usually available only over relatively short time intervals) and (ii) the underlying dynamics usually has a temporal trend and cannot be assumed to be stationary (e.g., in the Markovian case, the transition matrix cannot a priori be assumed to be time independent). Moreover, important practical problems frequently arise when analyzing nonstationary data, particularly with regard to the comparison of different trend models and trend hypothesis tests and the influence of the mathematical assumptions associated with the choice of the model on the analysis results (robustness).

One of the most frequently used techniques for analysis of nonstationary observation time series is the local Gaussian kernel smoothing approach (Loader 1996, 1999; Zhao and Wei 2003). The main idea of this approach is based on the locally stationary approximation of estimated quantities (like observable averages) or some estimated model parameters inside of the Gaussian window of a certain width, typically characterized by the respective variance parameter $\sigma^2$ [for geophysical applications of the method see, e.g., Anderson (2003); Xiong et al. (2006), and Dakos et al. (2008)]. The

Corresponding author address: Illia Horenko, Institute of Mathematics, Free University of Berlin, Arnimallee 2–6, 14195 Berlin, Germany.
E-mail: horenko@math.fu-berlin.de

DOI: 10.1175/2008JAS2959.1

© 2009 American Meteorological Society
window width parameter $\sigma^2$ is typically chosen a priori to characterize some slow intrinsic time scale of interest. However, the main difficulty of this approach lies in its locality; that is, the information used to estimate the quantities of interest at certain time point is acquired only locally from the inner part of the Gaussian window around this point.

This paper considers a problem of trend estimation for the dynamical processes with discrete finite state space based on the given observation data. In context of atmospheric applications, such processes can describe the transitions between certain predefined circulation regimes or patterns [like the Lamb circulation index (Lamb 1972) or blocked and unblocked situations in atmosphere (Majda et al. 2006)]. The considered processes are assumed to fulfill the Markov property; that is, the probability of transition to any other available state at any time is only dependent on the current state and is independent of the previous transition history.

This paper introduces two global approaches to trend identification in discrete Markov time series and compares them to the local Gaussian kernel smoothing technique adapted here for Markov chain series estimation. One of the presented methods is a parametric approach, allowing us to test simple trend models of different types and compare them with each other in a Bayesian sense. The second technique is of a nonparametric form and allows us to estimate low-frequency trends by solving a metastable clustering problem with a fixed number of cluster states. It is explained how the optimal number of clusters can be identified dependent on the single external scalar regularization parameter $\varepsilon^2$. The connection among the regularization parameter, the persistence of the resulting decomposition, and the Gaussian window width parameter $\sigma^2$ is discussed. It is shown that in the context of Markovian processes, the presented nonparametric method can be viewed as an adaptive extension of the local Gaussian kernel smoothing technique. Numerical examples demonstrate that in contrast to the nonparametric local Gaussian kernel smoothing technique (Loader 1996, 1999; Zhao and Wei 2003), both of the presented methods allow for a more robust estimation and comparison of different nonstationary Markov models.

It is shown that the introduced methods can be helpful in analysis of practical applications; for example, the presented algorithms are applied to analysis of the historical sequence of atmospheric circulation patterns over the United Kingdom between 1946 and 2007. The long-term effects in transitions between different patterns are systematically investigated and compared.

The paper begins with a short description of the nonparametric local Gaussian kernel smoothing in the context of Markov chains, followed by a description of both global methods. Special emphasis is placed on the intrinsic mathematical assumptions in each of the methods. The final sections deal with application of the presented techniques to the analysis of illustrative model data and atmospheric data series, interpretation of the obtained results, and comparison of different methods with respect to their robustness.

2. Methods and approach

In the following we will consider a data series that is discrete in time and space; that is, $\{X_t\}_{t=1, \ldots, T}$ takes values from some fixed set of $m$ distinct quantities $s_1, \ldots, s_m$. These quantities can denote, for example, the states or configurations of the observed system along the dynamics. The process underlying the observations is called Markovian if the probability $P$ of any current state of the process at time $t$ depends only on the previous state at time $t - 1$ and does not depend on any other previous state. Mathematically, this property can be expressed as $P[X_t = s_j | X_{t-1} = s_i] = P_{ij}(t)$, where $P_{ij}(t)$ denotes the probability of transition from state $s_i$ to state $s_j$ in one step at time $t$. These probabilities can be put together into an $m \times m$ stochastic transition matrix $P(t)$; that is, $\sum_{i=1}^{m} P_{ij}(t) = 1$ for any $i$ and $t$. To be able to estimate the a priori unknown transition probabilities $P_{ij}(t)$ based on the observation data $X_n$ we first need to introduce the concept of the observation probability or likelihood. For example, consider a simple discrete Markov process with three possible states $s_1 = 1$, $s_2 = 2$, and $s_3 = 3$, available in the form of the following observed time series:

$$\{X_t\}_{t=1, \ldots, T} = \{1, 1, 2, 1, 3, 2, 3, 3, 1, 3, 3, 2, 2\}. \quad (1)$$

Applying the Markov property, it is easy to demonstrate that the probability of observing this time series can be expressed as a probability of starting in state 1, then staying in state 1 at $t = 2$, then jumping from state 1 to state 2 at $t = 3$, etc. That is,

$$P[X_1 = 1, X_2 = 1, \ldots, X_{12} = 1] = P[X_1 = 1] \cdot P[X_2 = 1|X_1 = 1] \cdot \ldots \cdot P[X_{12} = 2|X_{11} = 2]. \quad (2)$$

For any given Markovian series $\{X_1, \ldots, X_T\}$, the corresponding observation probability (or likelihood) can be compactly written as $P[X_1, \ldots, X_T] = P[X_1]^{\Pi_{t=1}^{T} \Pi_{t \in \{t_0\}} P_{ij}(t)}$, where $\{t_0\}$ is the set of all time instances when the transitions between $s_i$ and $s_j$ are observed.$^1$ This

---

$^1$ It is assumed that $\{t_0\}$ is not empty $\forall i, j$. If it is not true for a certain pair $(i, j)$, it means that no direct transitions between the states $s_i$ and $s_j$ were observed in the time series, and the respective matrix elements $P_{ij}(t)$ can be assumed to be equal to zero for all $t$.  

Unauthenticated | Downloaded 07/31/22 05:04 PM UTC
means that if the transition matrix is unknown, it can be found by maximizing the above likelihood function for a fixed observation sequence \( \{X_t, \ldots, X_T\} \), for instance, by solving the following maximization problem:

\[
P_0(t) = \arg \max_{P(t)} P[X_t, \ldots, X_T]. \tag{3}
\]

From the numerical point of view, instead of solving the above maximization problem, it is much more convenient to maximize the logarithm of the above expression, i.e., the log-likelihood of the Markovian process:

\[
L[P(t)] = \log P[X_1, \ldots, X_T] = \log P[X_1] + \sum_i L_i[P_{i1}(t), \ldots, P_{im}(t)] \rightarrow \max, \tag{4}
\]

where

\[
L_i[P_{i1}(t), \ldots, P_{im}(t)] = \sum_{j=1}^m \sum_{t \in \{t_i\}} \log P_{ij}(t) \tag{5}
\]

is the partial log-likelihood of the state \( i \). To preserve the stochasticity of the resulting matrix \( P(t) \), the minimization problem (4) is subjected to the following constraints:

\[
\sum_{j=1}^m P_{ij}(t) = 1, \quad \text{for all } t, i; \tag{6}
\]

\[
P_{ij}(t) \geq 0, \quad \text{for all } t, i, j.
\]

The main difficulty arising from the formulation (4)–(6) is that it is not a well-posed problem. If the transition matrix is allowed to be completely time dependent, than the number of unknowns to be estimated in this case is equal \( m^2T \), whereas the number of observed transitions and equality constraints is \( m(T + 1) - 1 \). The optimizer of the problem (4)–(6) at time \( t \) in such a case will be \( P_{X_i, X_i}(t) = 1 \) and \( P_{X_i, X_j}(t) = 0 \) for all other elements \( k = 1, \ldots, m \) from the same row \( X_i \) of the transition matrix \( P(t) \). All other elements of the matrix can be set to any arbitrary values (satisfying the constraints), resulting in a meaningless estimation. Because the problem in (4)–(6) is ill posed, one needs to incorporate some additional information into the formulation of the problem (or, in mathematical language, to regularize the problem) to make it well posed. The simplest form of regularization is an additional assumption about the global stationarity of the Markov process [i.e., the transition matrix \( P(t) \) is assumed to be time independent].

One of the straightforward possibilities to relax the aforementioned global stationarity assumption is based on the application of local Gaussian kernel smoothing idea (Loader 1996, 1999; Zhao and Wei 2003) in context of the optimization problem [(4)–(6)]. Assuming a local stationarity of the underlying transition process at time \( t_0 \) inside of the window of a certain width \( \sigma^2 \), one can introduce a normalized Gaussian weight function \( \gamma(t, t_0) = 1/c \exp[-(t - t_0)^2/\sigma^2] \) (where \( c \) is the normalization constant). Approximating \( L[P(t)] \) as

\[
L[P(t_0)] \approx \log P[X_1] + \sum_{j=1}^m \sum_{t \in \{t_i\}} \gamma(t, t_0) \log P_{ij}(t_0) \rightarrow \max, \tag{7}
\]

and applying the method of Lagrange multipliers, we get

\[
P_{ij}(t_0) = \frac{\sum_{t \in \{t_i\}} \gamma(t, t_0)}{\sum_{t \in \{t_i\}} \gamma(t, t_0)}, \tag{8}
\]

where \( \{t_i\} \) is the set of all time instances when the state \( s_i \) was visited.

a. Trend models of discrete Markov chains

The two main disadvantages of the local Gaussian kernel smoothing procedure described above are (i) the assumption about the local stationarity of the transition process and (ii) the arbitrariness in the choice of the window width parameter \( \sigma \). Furthermore, this procedure does not directly make it possible to acquire a constructive functional form of the trend in the analyzed data. This means that to make predictions based on available time series information, one has to extrapolate the identified transition process \( P(t) \) in the future. To do that, the identified process \( P(t) \) has to be approximated with a time-dependent function of a certain class \( \phi(t) \). Instead of that, one can impose the functional form \( \phi(t) \) a priori and incorporate it into the maximization of the log-likelihood in the form of the trend model.

Single trend models have a general form

\[
P(t) = P^{(0)} + P^{(1)} \phi(t), \quad \phi:[1, T] \rightarrow (-\infty, +\infty), \tag{9}
\]

---

2 If \( P(t) \) is assumed to be time-independent, optimization problem (4)–(6) can be solved analytically applying the method of Lagrangian multipliers with \( P_{ij} = \|t_i\| / \Sigma_{t_i} \), where \( \|t_i\| \) denotes the number of observed transitions between the states \( s_i \) and \( s_j \) and \( \Sigma_{t_i} \) is not zero.
Problem (10) for a fixed observation sequence \( \{X_1, \ldots, X_T\} \) and some given trend function \( \phi(t) \) is some predefined bounded trend function. Inserting (9) into (4) and (6), after some obvious transformations for any \( i = 1, \ldots, m \), we get

\[
\sum_{j=1}^{m} \sum_{t \in \{t_i\}} \log[P_{ij}^{(0)} + P_{ij}^{(1)} \phi(t)] \rightarrow \max_{P_{ij}^{(0)}, P_{ij}^{(1)}} ,
\]

for all \( j \), and

\[
\begin{align*}
\sum_{j=1}^{m} P_{ij}^{(0)} &= 1, \\
\sum_{j=1}^{m} P_{ij}^{(1)} &= 0,
\end{align*}
\]

Problem (10) for a fixed observation sequence \( \{X_1, \ldots, X_T\} \) and some given trend function \( \phi(t) \) is some predefined bounded trend function. Inserting (9) into (4) and (6), after some obvious transformations for any \( i = 1, \ldots, m \), we get

\[
\sum_{j=1}^{m} \sum_{t \in \{t_i\}} \log[P_{ij}^{(0)} + P_{ij}^{(1)} \phi(t)] \rightarrow \max_{P_{ij}^{(0)}, P_{ij}^{(1)}} ,
\]

for all \( j \), and

\[
\begin{align*}
\sum_{j=1}^{m} P_{ij}^{(0)} &= 1, \\
\sum_{j=1}^{m} P_{ij}^{(1)} &= 0,
\end{align*}
\]

This cluster distance functional describes the quality of explaining the observed transition \( x_t; X_t \rightarrow X_{t+1} \) at time \( t \) with the help of the transition matrix \( P' \). For a given cluster distance functional (15), under data clustering we will understand the problem

\[
L = \sum_{i=1}^{K} \sum_{t=1}^{T-1} \gamma_i(t) g(x_t, P') \rightarrow \min_{\gamma_i(t), P'} ,
\]

subject to the constraints on \( \gamma_i(t) \)

\[
\sum_{i=1}^{K} \gamma_i(t) = 1, \quad \forall t \in (1, T-1) ,
\]

\[
\gamma_i(t) \geq 0, \quad \forall t \in [1, T-1], \quad i = 1, \ldots, K ,
\]

and on \( P^1, \ldots, P^K \) (6). As was demonstrated in Horenko (2009a,b), in order to get persistent (or metastable clusters) from this minimization procedure, it is enough to impose some metastability assumptions in the space of functions \( \gamma_i(\cdot) \) and then apply a finite Galerkin projection of this finite-dimensional Hilbert space. For example, one can restrain the weak discrete differentiability of functions \( \gamma_i \); that is,

\[
|\gamma_i|_{H^1(1,T-1)} = \|\gamma_i(t + 1) - \gamma_i(t)\|_{L^2(1,T-1)} = \sum_{t=1}^{T-1} \frac{|\gamma_i(t + 1) - \gamma_i(t)|^2}{\Delta t} \leq C < +\infty , \quad i = 1, \ldots, K .
\]

For a given observation time series, the above constraint limits a total number of transitions between the clusters and is connected to the metastability of the hidden process \( \gamma_i(t) \), for \( i = 1, \ldots, K \) (Horenko 2009b).

One of the possibilities to incorporate the constraint (19) into the minimization of (16) is to introduce the Lagrangian multiplier \( \varepsilon \)

\[
\ell = L + \varepsilon \sum_{i=1}^{K} \sum_{t=1}^{T-1} \frac{|\gamma_i(t + 1) - \gamma_i(t)|^2}{\Delta t} \rightarrow \min_{\gamma_i(t), P'} .
\]
This form of penalized regularization was first introduced by A. Tikhonov for solution of ill-posed linear least squares problems (Tikhonov 1943) and was frequently used for nonlinear regression analysis in context of statistics (Hoerl 1962) and multivariate spline interpolation (Wahba 1990). In contrast to the aforementioned application of Tikhonov-type regularization to interpolation problems (where the regularization is controlling the smoothness of some nonlinear functional approximation of the given data), the presented form of the regularized averaged clustering functional (20) has a completely different mathematical structure due to the form of the functional (16). This specific formulation of the optimization problem with constraints allows one to control the metastability of the assignment $\Gamma(t)$ of the analyzed data to $K$ distinct a priori unknown clusters (Horenko 2009b).

Let $\{1 = t_1, t_2, \ldots, t_{N-1} = T-1\}$ be a finite subdivision of the time interval $[1, T-1]$ with uniform time step $\delta t$. We can define a set of continuous functions $\{v_1(t), v_2(t), \ldots, v_N(t)\}$ called hat functions or linear finite elements with compact support [i.e., such that each function $v_i(t)$ is taking positive values in the time interval $(t_{i-1}, t_{i+1})$ and is zero outside] (Braess 2007). Assuming that $\gamma_i \in L^1(1, T-1)$ (i.e., functions with the first discrete derivative being square integrable functions in discrete sense; cf. Braess 2007), we can write

$$\gamma_i = \tilde{\gamma}_i + \chi_N = \sum_{k=1}^{N} \tilde{\gamma}_{ik} v_k + \chi_N, \quad (21)$$

where $\tilde{\gamma}_{ik} = \sum_{i=1}^{T-1} \gamma_i(t) v_k(t)$ and $\chi_N$ is a discretization error (it becomes zero if $\delta t = \Delta t$). Optimal $N$ can be chosen adaptively to guarantee that $\chi_N$ does not exceed a certain discretization error threshold. Inserting (21) into functional (20) and constraints (17) and (18), we get

$$\begin{align*}
L^e &= \sum_{i=1}^{K} \left[ a(\mathbf{P}^i) \tilde{\gamma}_i + e^2 \tilde{\gamma}_i^T \mathbf{H} \tilde{\gamma}_i \right] \to \min ; \\
\sum_{i=1}^{K} \tilde{\gamma}_{ik} &= 1, \quad \forall k = 1, \ldots, N; \\
\tilde{\gamma}_{ik} &\geq 0, \quad \forall k = 1, \ldots, N, \quad i = 1, \ldots, K; \\
\sum_{j=1}^{m} \mathbf{P}^{l}_{ji}(t) &= 1, \quad \text{for all } l, i; \\
\mathbf{P}^{l}_{ji}(t) &\geq 0, \quad \text{for all } l, i, j,
\end{align*} \quad (22)$$

where $\tilde{\gamma}_i = (\tilde{\gamma}_{i1}, \ldots, \tilde{\gamma}_{iN})$ is the vector of discretized affiliations to cluster $i$;

$$a(\mathbf{P}^i) = \left[ \sum_{j=1}^{l_1} v_1(t_j, \mathbf{P}^i) \delta_i \Delta t, \ldots, \sum_{j=l_{N-2}}^{l_{N-1}} v_N(t_j, \mathbf{P}^i) \delta_i \Delta t \right]$$

is a vector of discretized cluster distances; and $\mathbf{H}$ is the symmetric tridiagonal stiffness matrix of the linear finite element set, with $2/\delta t$ on the main diagonal, $-1/\delta t$ on both secondary diagonals, and zero elsewhere. The system of equations (22)–(26) is a nonlinear minimization problem with linear equality and inequality constraints, imposed on both $\tilde{\gamma}_i$ and $\mathbf{P}^i$, for $i = 1, \ldots, K$.

If $e^2 = 0$, then the above minimization problem [(22)–(24)] can be solved analytically with regard to $\tilde{\gamma}_i^{(L)}$ for a fixed set of transition matrices $\mathbf{P}^1, \ldots, \mathbf{P}^K$ (where $L$ denotes the index of current iteration), resulting in

$$\gamma_i^{(L)}(t_j) = \begin{cases} 1 & i = \arg \min_{\mathbf{P}^i} a(\mathbf{P}^i), \\ 0 & \text{otherwise}. \end{cases} \quad (28)$$

If $e^2 > 0$, for a fixed set of transition matrices $\mathbf{P}^1, \ldots, \mathbf{P}^K$, the minimization problem (22)–(24) reduces to a sparse quadratic optimization problem with linear constraints, which can be solved by standard tools of sparse quadratic programming (sQP) with computational cost scaling as $O[N \log(N)]$ (Gill et al. 1987).

The minimization problem (22) with transition matrix constraints (25) and (26) can be solved analytically with regard to the parameters $\mathbf{P}^1, \ldots, \mathbf{P}^K$ for a fixed set of discretized cluster affiliations $\tilde{\gamma}_i$. Applying the method of Lagrangian multipliers results in the following expression for the transition matrices:

$$\mathbf{P}_0^{l,i} = \frac{\sum_{t \in [t_j]} \gamma_i(t)}{\sum_{t \in [t_j]} \gamma_l(t)}, \quad \text{for all } l, i, j, \quad (29)$$

where $\{t_j\} \subset [1, T-1]$ are the time instances when the transitions between the states $s_i$ and $s_j$ are observed and $\{t_j\}$ is a set of all time instances when the state $s_i$ is visited.

Note the obvious similarity between the expression (29) and the local transition matrix estimator (8) in the context of Gaussian kernel smoothing. Despite this similarity and the fact that both of the methods rely on some form of stationarity assumption, the statistical weights $\gamma(t, t_0)$ are fixed and localized by the choice of the Gaussian curve of the certain width, whereas the hidden probabilities $\gamma_i(t)$ are defined adaptively, dependent on the whole set of available data and not just on the part of it from inside the window. In another words, expression (29) makes use of the global information contained in the cluster affiliation functions $\gamma_i(t)$ (later on in the text it is explained how predefined
discretization error threshold and regularization parameter $\epsilon^2$ influence these values). Globality of the adaptive finite element method (FEM) clustering procedure gives an advantage when analyzing the data with a short time record because it allows us to tighten the confidence intervals around the estimated parameters and can help to obtain more robust numerical results.

Similarly to the algorithms described in (Horenko 2009a,b), minimization of the functional (22) can be implemented as an iterative numerical scheme. The iterations can be repeated until the change of the functional value does not exceed some predefined threshold for the change of the functional value (22).

c. Estimation of optimal $K$ dependent on $\epsilon^2$

The upper bound for the number of statistically distinguishable cluster states for each value of $\epsilon^2$ can be algorithmically estimated in the following way: starting with some a priori chosen (big) $K$, one solves the optimization problem (22–26) for a fixed value of $\epsilon^2$ and calculates the confidence intervals of the resulting local transition matrices $P_i, i = 1, \ldots, K$ (this can be done by applying the standard sampling procedures; see Noc 2008). If two of the estimated matrices have confidence intervals that overlap in all components, this means that respective clusters are statistically indistinguishable and the whole procedure must be repeated for $K = K - 1$. If at a certain point all of the matrices are statistically distinguishable, the procedure is stopped and $K_{\text{max}}(\epsilon^2) = K$.

d. Robustness estimation

It is intuitively clear that the quality of the resulting reduced model is very much dependent on the original data, especially on the length of the available time series. The shorter the observation sequence, the larger the uncertainty of the resulting parameters. Therefore, the global methods presented above (the single trend model and FEM clustering) should be more robust then local methods (e.g., Gaussian kernel smoothing) as the analyzed time series become longer. That is, for a short time series, everything is local, but as the time series becomes long (in the sense that $T \gg \sigma$, with $\sigma$ defining the width of a Gaussian moving window) the global methods then have the advantage of assimilating more information.

Let $P^M = P^M_{X(t)}(t)$ be the Markovian transition matrix, estimated with one of the methods $M$ described above from a given observation sequence $\{X_t\}_{t=1,\ldots,T}$. As $\{X^M_t[p(t),\omega]\}_{t=1,\ldots,T}$, we denote a single realization of the Markov process $P(t)$ with the model $M$. To compare the robustness of different models for a given data, we introduce the following estimate deviation process

$$R^M_{\{X(t)\}}(f, \omega) = f(P^M^0) - f(P^M_{\{X(t)\}[P(t),\omega]}(t)).$$

(30)

The Gaussian assumption for the stochastic process $R^M_{\{X(t)\}}(f, \omega)$ gives an opportunity to estimate its confidence intervals for some given function $f$ straightforwardly. This can be done in a standard way by using multivariate statistical analysis, that is, by the Monte Carlo sampling from the respective distribution and calculating expectation values with respect to $\omega$ (Mardia et al. 1979). The confidence intervals calculated for the variable $R^M_{\{X(t)\}}(f, \omega)$ give a possibility to estimate the intrinsic robustness of the chosen model $M$ for a given observation $\{X_t\}_{t=1,\ldots,T}$.

3. Application I: Instructive model example

To illustrate an application of the methods described above, it is instructive to investigate the following simplified model scenario: consider a process $X_t$ with three directly observed possible states $\{1, 2, 3\}$ and transition probabilities taking values from one of the following transition matrices:

$$A_1 = \begin{pmatrix} 0.82 & 0.12 & 0.06 \\ 0.08 & 0.87 & 0.05 \\ 0.07 & 0.09 & 0.84 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0.16 & 0.70 & 0.14 \\ 0.52 & 0.37 & 0.11 \\ 0.30 & 0.27 & 0.43 \end{pmatrix}.$$  

(31)

The choice of the transition matrix, taken to calculate a probability of the next transition at time $t$, depends on another discrete hidden process $Y_t$ with two possible states $\{1, 2\}$. If the hidden process $Y_t$ takes the value 1, then matrix $A_1$ is chosen to calculate transition probabilities of the observed process $X$ at time $t$; otherwise, transition matrix $A_2$ is taken. An example of the observation series $X_t$ resulting from this computation procedure is demonstrated in the right panel of Fig. 1. The hidden process $Y_t$ used in this calculation is shown
The log-likelihood of the generated time series in terms of the switching Markov model (31) is \( L_{\text{real}} = -708.67 \).

Next, the generated Markov time series from Fig. 1 is taken to compare three presented methods with respect to the identification of the nonstationary trend of the Markov transition probabilities (driven by the hidden variable \( Y_t \)). We start with the local Gaussian kernel smoothing procedure (8), repeat the estimation with different values of the Gaussian window width parameter \( s^2 \), and keep the parameter value with the highest log-likelihood (which in this case turns out to be \( s_{\text{opt}}^2 = 1296 \), corresponding to an effective window width of approx. 70 time units). The log-likelihood of the analyzed time series in terms of the local Gaussian kernel smoothing model with \( s_{\text{opt}}^2 = 1296 \) turns out to be \( L_{\text{Gauss}} = -717.34 \) (this is just 0.58% less than the log-likelihood of the original switching Markov model). Estimation of the single trend model (9) with \( f(t) = t^a \) as expected results in almost stationary transition matrix estimates. This is simply explained by the fact that models with polynomial trend are a bad choice for approximating nonstationary processes when the change of the transition probability matrices is described by the hidden process like the one from Fig. 1. Finally, we apply the adaptive FEM clustering procedure with different values of the regularization parameter \( \epsilon^2 \). The tolerance threshold \( \chi_N \) (21) is chosen to be 0.0001 in all cases.

Figure 2 demonstrates an influence of the regularization parameter \( \epsilon^2 \) on the identified hidden process [described by the variables \( \gamma_i(t) \) resulting from the numerical solution of the optimization problem (22)–(26)]. Whereas for \( \epsilon^2 = 0 \) (corresponding to the unregularized optimization) we get an arbitrary hidden process, for \( \epsilon^2 = 0.2 \) the identified hidden process almost exactly reproduces the original process \( Y_t \) used in generation of the time series. Hidden processes identified with bigger values of \( \epsilon^2 \) become more and more smooth until the point where the regularization part of the functional
(20) starts to dominate the optimization, producing a constant hidden process with no switches.

Figure 3 shows a comparison of the original transition probability trends as functions of time with the results obtained by the local Gaussian kernel smoothing with the value of \( \sigma^2 \) defined by the variation of the log-likelihood maximization procedure (dark gray, dotted), together with their robustness intervals (light gray, dotted); and probabilities estimated for \( \varepsilon^2 = 0.2 \) \((K = 2, \text{ with tolerance threshold } \chi_N = 0.0001; \text{ dark gray, dashed})\). The robustness intervals for the FEM clustering method estimates are of the order of \( \pm 0.06 \).

As can be seen from Fig. 3, trends resulting from the adaptive FEM clustering procedure reproduce the original trends much better than the local Gaussian kernel smoothing trends do. This is not a big surprise because the internal structure of the analyzed time series is much more reproducible with the help of the adaptive FEM clustering procedure. Figure 3 also demonstrates the effect of locality of the Gaussian smoothing procedure: because the optimal Gaussian window turns up to become relatively narrow, only a relatively small amount of information is incorporated in the trend estimation at any point. This results in huge confidence intervals for the estimated matrix elements. Comparison of log-likelihood values \( L_{\text{real}} = -708.67, L_{\text{Gauss}} = -717.34, \text{ and } L_{\text{FEM}} = -703.21 \) and inspection of the respective probability trends in the Fig. 3 reveals that very small relative changes in the value of the log-likelihood can be induced...
by very significant changes of trend. This means that the relative value of the log-likelihood function alone cannot be considered as a measure for the model quality.

Finally, the procedure of \( K_{\text{max}}(\epsilon^2) \) identification described above is exemplified for this model data example. Figure 4 demonstrates that the maximal number of statistically distinguishable clusters decreases with increasing value of \( \epsilon^2 \). Moreover, the described procedure allows a correct identification of hidden states \( Y \) in a wide range of regularization parameters (0.2 \( \leq \epsilon^2 \leq \) 0.8). Note that because of the form of the regularized functional (20), the absolute value of the regularization parameter \( \epsilon^2 \) necessary to achieve a desired persistence of the hidden process \( \gamma_i(t) \) for \( i = 1, \ldots, K \) will be different for different applications. This value is influenced by the norm and the properties of the original cluster distance functional (16) evaluated for a given time series; that is, the magnitude of \( \epsilon^2 \) and numerical investigations analogous to the one presented in Fig. 4 should be performed in each data analysis case.

4. Application II: Analysis of Lamb circulation index series for the United Kingdom

a. Data sources

Analyzed data provided by the Climatic Research Unit of the University of East Anglia (available online at http://www.cru.uea.ac.uk/ftpdata/lwtjenk.dat) represent a sequence of atmospheric circulation patterns over the United Kingdom. In the original time series, each of the days between 1 January 1880 and 31 July 2007 is assigned to one of the 28 basic patterns (cyclonic, anti-cyclonic, western, etc.) based on the daily gridpoint mean sea level pressure data. The assignment of the circulation regimes in the considered time series is done according to the modified Lamb classification scheme proposed by Jenkinson and Collison (Lamb 1972; Jenkinson and Collison 1977; Jones et al. 1993). In the following, the analysis is restricted to the part of the series after World War II (i.e., between 1 January 1946 and 31 July 2007). To simplify the analysis and to make the graphical representation of results more comprehensible, the total number of the analyzed circulation patterns is reduced to three; that is, the anticyclonic pattern is denoted as state 1, the cyclonic pattern is denoted as state 3, and all other patterns are assigned to the collective state 2.

b. Trend discrimination with maximum likelihood approach

Before demonstrating the applicability of the strategies described in this paper to the resulting time series (switching between the discrete states 1, 2, and 3), it is necessary first to demonstrate the Markovian property of the respective transition process. Unfortunately, there is no standard strategy to test the Markov assumption in a nonstationary case. In the following we will use the test strategy applicable to a stationary case and will assume that the deviations from stationarity are not very significant for the analyzed data so the stationary tests are still valid in some sense. Development of nonstationary Markov tests is a matter of future research. For the stationary case (where the underlying transition matrix is assumed to be time independent), the Markovian property can be verified applying some standard tests; for example, one can check the generator of the process (see Crommelin and Vanden-Eijnden 2006; Metzner et al. 2007). Application of this test confirms the Markov hypothesis for the analyzed data in stationary case and reveals that the underlying process can be represented by a rapidly mixing Markov chain (i.e., the probabilities of staying in the same state are comparable to the probabilities of changing the state, so the processes do not get stuck in the same state for a long time).

As was mentioned above, identification of multiple trend models is a problem that is unfeasible from the numerical point of view because of the nonconvexness of the respective inequality constraints. To choose the most probable single trend model for a given time series of atmospheric circulation patterns, different trend function classes \( \phi(t) \) can be parameterized by numerical solution of the problem (10)–(14) and then compared with the help of the standard Bayesian hypothesis test approach (Gelman et al. 2004).
One of the most intuitive single trend models in meteorological context is the seasonal trend model of the form

\[ P(t) = P(0) + P(1) \sin \left( \frac{2\pi}{T} t + \phi \right), \]  

(32)

where \( T = 365.24 \) days is a seasonal period and \( \phi \in [0, T] \) is a phase factor. The maximization problem (10)–(14) is independently solved for each of the rows of the transition matrix. Optimization is repeated for various values of \( \phi \), and for each matrix row the parameters \( p(0), p(1) \), and \( \phi \) with the highest value of the partial log-likelihood are kept (see the dashed lines in Fig. 4). The same kind of procedure is performed for the single trend model of the polynomial form

\[ P(t) = P(0) + P(1) t^\alpha. \]  

(33)

Statistical hypothesis tests can be performed to decide which of the single trend models can best explain the observed data. The log-likelihood of the respective partial log-likelihood maxima (see Fig. 5) can be calculated, and the a posteriori model probabilities can then be acquired from the Bayes formula. It shows that the polynomial trend model \( P(t) = P(0) + P(1) t^\alpha \) and the nonstationary hidden states model (estimated with the adaptive FEM clustering algorithm for three hidden states) have the highest probability of describing the analyzed data. Moreover, different values of exponent \( \alpha \) in the polynomial trend model are optimal for different atmospheric circulation patterns (\( \alpha = 0.8 \) for anticyclonic, \( \alpha = 0.3 \) for cyclonic, and \( \alpha = 0.5 \) for the combination of all other patterns). Inspection of the resulting trend derivatives \( \dot{P}(t) = \alpha P(1) t^{\alpha-1} \) shows that according to the analyzed data, the speed of climate change visible in transition probability trends was higher at the beginning of the analyzed period as compared to the current period of time (because for all of the identified trends, \( \alpha < 1 \)). This finding may represent a local effect and must be verified on the global data. This is a matter of further research.

Figure 5 demonstrates that the maximal log-likelihood values for optimal polynomial trend and hidden-state models are higher than the optimal values for the seasonal trend model (dashed lines) and the local Gaussian kernel smoothing result (dashed–dotted); note that the effective Gaussian window width of 10 yr corresponds to the Gaussian window width used in the original work (Jones et al. 1993) considering the analysis of the U.K. Lamb index series. This finding means that the influence of seasonal pattern variability on transition processes is dominated by the long-term effects induced by the single polynomial trend and hidden-state models. Moreover, the optimal hidden-state model with \( K = 3 \) has the highest probability in the case of states 1 and 2.
case of state 3 (describing the cyclonic pattern), the single polynomial trend model $P_{ij}(t) = P_{ij}^{(0)} + a_{ij}^{(1)}t$ can explain the observed data better than any other tested model. However, this finding should be handled with care because the number of parameters used in the hidden state model for $K = 3$ is higher than the number of parameters involved in the single polynomial trend model.

c. Comparison of robustness and predictability

To interpret the obtained results, estimated transition matrices $P(t)$ and instantaneous statistical weights $\pi_i(t)$, for $i = 1, 2, 3$,

$$\pi(t)P(t) = \pi(t), \quad \pi(t) = [\pi_1(t), \ldots, \pi_3(t)]$$  

can be compared with respect to both their qualitative temporal behavior and robustness. Instantaneous statistical weights are the components of the steady-state PDF of the stationary Markov process; that is, in non-stationary cases these quantities describe a time evolution of an equilibrium probability density function (PDF). Figure 6 demonstrates a good agreement of the resulting parameters for the single polynomial trend model and the hidden states model. It can clearly be seen that in both cases the probability of staying in the anticyclonic pattern decreases, the probability of moving from the anticyclonic to the cyclonic pattern increases, and the probability of moving from anticyclonic to any other circulation pattern stays almost constant. For the cyclonic pattern transitions, the situation is reversed: the probability of staying in the cyclonic pattern increases, the probability of transitioning from the cyclonic to the anticyclonic pattern decreases, and the probability of transitioning from the cyclonic to any other circulation pattern stays almost constant. As can be seen from Fig. 7, this tendency results in increasing instantaneous statistical weight of the cyclonic pattern and in decreasing weight of the symmetric anticyclonic pattern. Moreover, Figs. 2 and 3 demonstrate the higher robustness of results obtained by the global methods compared to the local Gaussian kernel smoothing method.

To make predictions based on the available time series information, one has to extrapolate the identified transition process $P(t)$ to the future. As was mentioned above, single trend models offer the direct possibility of
generating long-term predictions because the exact functional form (in some predefined class of functions, e.g., polynomials) of the transition process $P(t)$ is estimated explicitly from the observation data. For any given time $1 < t < T$, $P^{[1,\tau]}(t)$ will denote the Markovian transition matrix estimated only from part of the available time series $X_1, X_2, \ldots, X_t$. To quantify the prediction quality based on this estimate, the mean log-likelihood of prediction $L_D$ can be used:

$$L_D = \frac{1}{\Delta t} \sum_{i,j} \sum_{t \in \{t_{ij}^\tau\Delta t\}} \log P_{ij}^{[1,\tau]}(t),$$

where $\{t_{ij}^{\tau+\Delta t}\} \subset [\tau+1, \tau+\Delta t]$ are the time instances between $\tau+1$ and $\tau+\Delta t$ when the transitions between $s_i$ and $s_j$ are observed. Figure 8 illustrates that predictions based on a single trend polynomial model are more probable than the predictions based on a stationary Markovian model without trend. This means that the long-term effects explained by the polynomial trend are significant and can be reliably identified even when using a part of the original data.

5. Concluding discussion

Three numerical methods for analysis of nonstationary Markovian data have been presented and compared here. In contrast to the standard approach (a local non-parametric technique, i.e., getting use only of the local information inside of the moving Gaussian window), two presented methods acquire the information globally and therefore, under predefined mathematical assumptions, they allow a more reliable estimate of the underlying model parameters. This feature is very important for analysis of a short record series. Both presented global
methods demonstrated more reliable trend discrimination with narrower robustness intervals compared with the results of the local Gaussian kernel smoothing. It was exemplified how the new methods can help perform a robust identification of transition probabilities and stationary weights in the analyzed circulation data. Both methods, despite the difference in the mathematical assumptions implied in the data (one being parametric and the other nonparametric), revealed the same robust weight increase of the cyclonic circulation pattern (absolute increase of $6.3 \pm 0.5\%$) and symmetrical decrease of the anticyclonic pattern weight (absolute decrease of $5.5 \pm 0.5\%$) over the United Kingdom between 1945 and 2007. Moreover, the results of the single trend model analysis indicated that the speed of climate change identified from transition probability trends was higher at the beginning of the analyzed period as compared to the current period of time.

One of the most important practical issues in the area of time series analysis is the construction of dynamical models able to predict the future. The single trend models have obvious shortcomings predicting situations in which the functional and parametric form of the trend function $\phi(t)$ is changed (i.e., regime change). On the other hand, the adaptive FEM clustering method has a potential to cope with this problem. However, in the current setting of the algorithmic procedure, hidden state probabilities $\gamma_i(t)$ are identified without making any assumptions about their dynamics (in contrast to the widely used HMM strategy in which these quantities are taken a priori to be Markov processes). This means that to be able to predict the phase transitions in realistic applications, the dynamics of hidden state probabilities $\gamma_i(t)$ have to be further investigated a posteriori. Moreover, robust statistical techniques of change point analysis have to be developed. These problems are the matters of further research.

For the analyzed practical application, it was demonstrated that the polynomial parameterization of the nonstationary Markov model enables a better quality of predictions as compared to the stationary case. Further development of the presented data analysis techniques can help to acquire a better understanding of the low-frequency variability and dynamics of processes switching between metastable weather and climate regimes.

Acknowledgments. The author would like to thank Andrew Majda (Courant Institute, NYU), Eric Vanden-Eijnden (Courant Institute, NYU), Rupert Klein (FU Berlin), and Dirk Becherer (HU Berlin) for intensive discussions and valuable comments.

The author thanks the unknown referees whose valuable comments helped to make the paper more readable and the Climatic Research Unit of the University of East Anglia for the time series data. The work was partially supported by the DFG SPP 1276 MetStrom “Meteorology and Turbulence Mechanics” and DFG Research Center “Matheon.”

REFERENCES


