*Communications in Applied Mathematics and Computational Science*

# ANALYSIS OF PERSISTENT NONSTATIONARY TIME SERIES AND APPLICATIONS

PHILIPP METZNER, LARS PUTZIG AND ILLIA HORENKO

vol. 7 no. 2 2012

msp

COMM. APP. MATH. AND COMP. SCI.
Vol. 7, No. 2, 2012
dx.doi.org/10.2140/camcos.2012.7.175

msp

# ANALYSIS OF PERSISTENT NONSTATIONARY TIME SERIES AND APPLICATIONS

PHILIPP METZNER, LARS PUTZIG AND ILLIA HORENKO

We give an alternative and unified derivation of the general framework developed in the last few years for analyzing nonstationary time series. A different approach for handling the resulting variational problem numerically is introduced. We further expand the framework by employing adaptive finite element algorithms and ideas from information theory to solve the problem of finding the most adequate model based on a maximum-entropy ansatz, thereby reducing the number of underlying probabilistic assumptions. In addition, we formulate and prove the result establishing the link between the optimal parametrizations of the direct and the inverse problems and compare the introduced algorithm to standard approaches like Gaussian mixture models, hidden Markov models, artificial neural networks and local kernel methods. Furthermore, based on the introduced general framework, we show how to create new data analysis methods for specific practical applications. We demonstrate the application of the framework to data samples from toy models as well as to real-world problems such as biomolecular dynamics, DNA sequence analysis and financial applications.

## 1. Introduction

In the field of time series analysis, a common problem is the analysis of high-dimensional time series containing possibly hidden information at different time scales. Here we consider the analysis of *persistent processes*, those where the temporal change of the underlying model parameters takes place at a much slower pace than the change of the system variables themselves. Such systems could be financial markets (where the underlying dynamics might drastically change due to

Illia Horenko is the corresponding author.

Work supported by the Swiss National Science Foundation (project "AnaGraph"), the German DFG SPP 1276 (MetStröm) "Meteorology and Turbulence Mechanics" and by the Swiss HP2C initiative "Swiss Platform for High-Performance and High-Productivity Computing". P. Metzner acknowledges the financial support of the DFG priority programme 1276 (MetStröm) "Multiple Scales in Fluid Mechanics and Meteorology".

*MSC2010:* primary 60G20, 62H25, 62H30, 62M10, 62M20; secondary 62M07, 62M09, 62M05, 62M02.

*Keywords:* nonstationary time series analysis, nonstationary data analysis, clustering, finite element method.

market breakdowns, new laws, etc.) [27; 48; 63]; climate systems (depending on the external factors like insolation, human activity, etc.) [54; 18; 39; 38; 32; 30]; ocean circulation models [22; 23] or biophysical systems [67; 37; 36; 41; 62; 68].

In the literature, the problem of data-based phase identification is addressed by a huge number of approaches which can be roughly classified as either *non-dynamical* or *dynamical* methods. The class on nondynamical methods exploits solely *geometrical* properties of the data for clustering regardless of their temporal occurrence. The most prominent approach is the $k$-means method [53], which clusters data points according to their minimal distance to geometrical centroids of point clouds.

Dynamical methods additionally take into account the temporal dynamics of data. This class of methods can further be divided into Bayesian approaches, such as the hidden Markov model (HMM) [4; 3; 56; 37; 36] or the Gaussian mixture model (GMM) (see, e.g., [21]) and the so-called *local kernel methods* (moving window methods) [20; 52]. Although the Bayesian methods have proven to be very successful in applications ranging from speech recognition [64] over atmospheric flows identification [54; 18] to conformation dynamics of biomolecules [17], they are based on the restrictive assumption that the underlying dynamics are governed by a *stationary probabilistic model*. Particularly, the assumption of stationarity implies, e.g., a locally constant mean value and a locally constant variance. In many real world applications, however, these implications are not valid due to theoretical reasons or simply due to the lack of sufficiently long time series of observations.

In local kernel methods the assumption of stationarity is relaxed by applying *nonparametric regression methods* to estimate time-dependent statistical properties of the underlying data. The key idea is the following: instead of considering every element of the time series to be equally statistically important, for a fixed time $t$ the data is *weighted* with a suitable so-called *kernel* function, e.g., a Gaussian probability density function. The modified time series then is considered to be stationary and, consequently, statistical objects can be computed by standard procedures.

The nonstationary time series analysis methods that have been developed in the group of I. Horenko and that will be considered in the current manuscript can be seen as a generalization of the idea described above. Therefore we explain the procedure in more detail. Suppose we observed a time series of real-valued observations discretely in time, denoted by $\boldsymbol{X} = (x_{t_0}, \ldots, x_{t_T})$ with $0 \leq t_0 < \ldots < t_T \leq 1$. Further suppose that the time series is appropriately described by the model

$$x_{t_i} = \mu(t_i) + \varepsilon_{t_i}, \quad i = 0, \ldots, T, \tag{1}$$

where $\{\varepsilon_{t_i}\}$ is a family of independent and identically distributed (i.i.d.) random variables with $\mathbb{E}[\varepsilon(t_i)] = 0$. An estimator for $\mu(t)$, $t \in [0, 1]$ is given by [19; 20]

$$\hat{\mu}(t) = \frac{1}{b}\sum_{j=0}^{T} x_{t_j} \int_{s_j}^{s_{j+1}} W\left(\frac{t-s}{b}\right) \mathrm{d}s, \tag{2}$$

where $W(\cdot)$ is a nonnegative kernel function satisfying the conditions

$$\int_{-\infty}^{\infty} W(s)\,\mathrm{d}s = 1, \quad \int_{-\infty}^{\infty} (W(s))^2\,\mathrm{d}s < \infty \tag{3}$$

and $0 = s_0 \le t_0 \le s_1 \le t_1 \le \ldots \le t_T \le s_{T+1} = 1$. The parameter $b \in \mathbb{R}$ is referred to as the window size associated with the kernel function and determines the statistical importance of the data in the temporal vicinity of a time $t$. For instance, if the kernel function is chosen to be the probability density function (PDF) of the standard normal distribution,

$$W(s) = \frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{s^2}{2}\right), \tag{4}$$

then $b$ is the standard deviation of the normal PDF $W\left(\frac{t-s}{b}\right)$. Hence, only data points within the window $[t-b, t+b]$ significantly contribute to the estimator in (2).

The effect of the Gaussian kernel on the estimation of $\mu(t)$ is exemplified on a time series generated via a persistent switching process between two processes, each wiggling around a constant mean value. The estimators $\hat{\mu}(t)$ for different choices of the window size $b$ are depicted in Figure 1. As expected, although the estimators for the smallest window size give good local estimations of the respective constant mean, they are noisy and not constant. Moreover, the estimator becomes poor for time points close to the beginning or the end of the time series

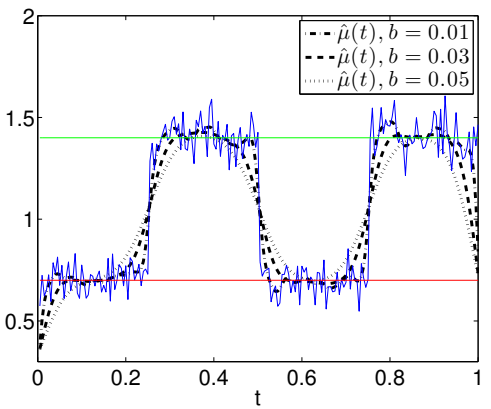


**Figure 1.** Illustration of the local kernel method on a time series generated via a persistent switching process between two processes each wiggling around a constant mean value ($\mu_1 = 0.7$, $\mu_2 = 1.4$). The estimator $\hat{\mu}(t)$ strongly depends on the specific choice of the window size. Results for a Gaussian kernel and $b = 0.01$, 0.03 and 0.05.

which is due to an insufficient statistics. In contrast, the graph of the estimators for the biggest window size is smooth but gives poorly local estimations and, hence, does not capture the intrinsic dynamics of the time series. Consequently, choosing the "right" window size $b$ is an *ill-posed* optimization problem which is basically due to the local ansatz of the approach and the danger of overfitting.

The approach presented herein can be understood as a method to adaptively identify nonlocal kernel-functions which enforces optimal regularization of the estimators. The basic underlying idea is to simultaneously detect the hidden switching process between persistent regimes (clusters) and their respective optimal parameters characterizing local substitute models [39; 38; 31; 32; 30; 33]. Mathematically, the hidden (affiliation) process defines a curve in parameter space. The optimal paths and the associated optimal parameters of the local models are characterized via the minimization of an appropriate *clustering functional* measuring the quality of data approximation in terms of a fixed number of local error measures. In order to avoid overfitting, or more generally spoken, to ensure *well-posedness* of the clustering problem as an inverse problem, the smoothness of paths as a function of time is limited in some appropriate function space, e.g., the Sobolev $H^1$ space [31; 33] or the larger class BV, consisting of functions with *bounded variations* [33].

The cluster algorithms arising from the $H^1$ approach and the BV-approach partially result from *finite element (FE) discretization* of the 1-dimensional cluster functional. This allows us to apply methods from the broad repository of existing FE methods from the numerics of partial differential equations (PDEs). The $H^1$-smoothness of the paths in parameter space is indirectly enforced by a *Tikhonov regularization* leading to numerically expensive constrained quadratic minimization problems during the course of minimization of the cluster functional. In contrast, the variational formulation in the BV-space amounts to solving linear programming problems with linear constraints and, most importantly, allows the direct control of the regularization of the paths in parameter space. The entire FEM-BV approach will be explained in detail in Section 2.

The FEM-BV approach has two advantages; We neither have to make any assumptions *a priori* on the probabilistic nature of the data, i.e., on the underlying distribution of the data, nor we have to assume *stationarity* for the analysis of the time series (in contrast to standard methods such as HMMs, GMMs or local kernel methods). Moreover, as demonstrated in [31], the method covers geometrical cluster approaches as well as dynamical ones. Furthermore, we will discuss in Section 2.h the relation of the proposed approach to probabilistic methods.

The outcome of the FEM-BV methodology depends on the prescribed number of clusters (local models) as well as on the prescribed regularity. Hence, the optimal choice of these parameters is crucial for the interpretation and the meaningfulness of the analysis. The new idea presented in this paper is to select the model

that describes the data best while involving the least number of free parameters by combining an information theoretic measure — Akaike's information criterion (AIC) [1] — with the maximum entropy approach [43; 44]. The resulting modified AIC then allows us to identify in a postprocessing step the optimal nonstationary data-based substitute model. The main advantage of the modified AIC approach (presented in this manuscript) to information theoretical approaches used until now is that no explicit assumptions on the parametric form of observables' distributions have to be made. The only assumption is that a scalar process describing the time-dependent error of the inverse problem is i.i.d.

Complementary to providing insight in the nonstationary behavior of the time series, the optimal substitute model lends itself for predicting the dynamics, e.g., for different initial values. The prediction, however, is restricted to time instances within the trained time span (as the underlying transition process in parameter space is only available for that span). To overcome that restriction, a substitute model for the (nonstationary) transition process itself is derived. Combining the two data based models leads to a self-contained model that allows us to predict the dynamics for any initial value at any time instance.

**1.a.** ***New contributions and organization of the paper.*** The main purpose of this manuscript is threefold. First, in Section 2 we provide a complete, unified and simplified derivation of the FEM-BV methodology originally introduced in [29; 30; 31; 32; 33; 34; 35] for analyzing nonstationary time series. Thereby, we exemplify in Section 2.c the derivation of the framework for different models to give a guideline how the developed methodology can be adapted and redesigned for new applications. For the first time, specifically, we adapt the FEM-BV approach to: (i) analyze periodic and partially observed (projected) data (torsion angles of a biomolecule) and (ii) to pattern recognition in discrete data sequences (first chromosome of the yeast).

The second purpose is to close the gap between the FEM-BV approach and classical methods by investigating the assumptions and conditions under which the FEM-BV methodology reduces to well-known methods for analyzing (non-)stationary time series. For details see Section 2.h. Particularly, for the first time we clarify in Section 2.g under what conditions the solution of the variational problem (associated with the interpolation of the inverse model) can be interpreted as a direct interpolation model (mixture model) for the data under consideration.

Additionally, we present a unified strategy for model selection in Section 3 that allows the selection of an optimal mixture model — optimal in the sense that the model provides maximal meaningfulness under minimal assumptions on the data. The new model selection criterion combines a well known information criterion with the maximum entropy approach for the inference of probabilistic distributions from observables without assuming any parametric form.

All these three aspects are eventually combined in a self-contained scheme for predicting the nonstationary dynamics of the data beyond the analyzed time horizon. The prediction scheme is motivated and described in detail in Section 4.

Finally, the applicability and usefulness of the presented methods is demonstrated in Section 5 by analyzing realistic data ranging from torsion angle time series of a biomolecule (trialanine), DNA nucleotide sequence data (from the first chromosome of the yeast *Saccharomyces cerevisiae*) and financial data (prices of oil futures). We end this manuscript by giving a conclusion in Section 6.

## 2. Finite element clustering method

**2.a.** ***The model distance function.*** Modeling processes in real world applications amounts to seeking an appropriate parametric model function which is considered to govern (explain) well the observed process. Suppose the observable of interest, denoted by $x_t$, is a $d$-dimensional vector. Furthermore, without loss of generality, assume that the time series of observations is given at times $t = 0, 1, \ldots, T$. Then, the *direct mathematical model* is a function $f(\cdot)$ that relates an observation $x_t \in \Psi \subset \mathbb{R}^d$ at a time $t \geq 0$ to the history of observations up to the time $t$ and a time-dependent set of parameters $\theta(t)$ from some parameter space $\Omega$. Formally, the relation is written as[1]

$$x_t = f(x_t, \ldots, x_{t-m}, \theta(t)) \quad t \geq m, \tag{5}$$

where $m \geq 0$ is the memory depth of the history dependence. Notice that the formulation in (5) is most general in that it also covers implicit dependencies. See, e.g., (26) in Section 2.c.ii.

The model function can be deterministic or can denote a random process. For instance, the simplest model function incorporating randomness is given by

$$x_t = f(\theta(t)) \stackrel{\text{def}}{=} \theta(t) + \varepsilon_t, \tag{6}$$

where $\{\varepsilon_t\},\ t \geq 0$ is a family of i.i.d. random variables with $\mathbb{E}[\varepsilon_t] = 0,\ t \geq 0$. The random variables $\varepsilon_t$ model, for instance, errors in the measurement of observables or they capture unresolved scales of a physical process such as fast degrees of freedoms. Thus, the model function in (6) corresponds to the assumption that the process under consideration has no dynamics and no memory.

Suppose we knew the parameters $m$ and $\theta(t),\ t \geq 0$ then the *direct mathematical problem* would be to find a process $x_t,\ t \geq 0$ satisfying the direct model in (5). Here we are interested in the opposite question. Suppose we are given a time series of observations $\boldsymbol{X} = (x_t), t = 0, \ldots, T$ and a known memory depth $m$. What are the optimal parameters, i.e., the parameter function $\theta^*(t)$ explaining the given time

[1] For notational convenience, we prefer (5) to the equivalent relation $0 = F(x_t, \ldots, x_{t-m}, \theta(t))$.

series of observations best? This *inverse problem* makes only sense if "best" is quantified in terms of a *fitness function* measuring the quality of the approximation for a given set of parameters. Throughout this manuscript a fitness function is denoted by

$$g(x_t, \dots, x_{t-m}, \theta(t)) : \Psi^{m+1} \times \Omega \mapsto \mathbb{R}. \tag{7}$$

Particularly, any metric $d(\cdot, \cdot) : \Psi \times \Psi \mapsto \mathbb{R}_0^+$ on $\Psi$ naturally induces a fitness function by defining $g(\cdot)$ as

$$g(x_t, \dots, x_{t-m}, \theta(t)) = \left(d\left(x_t, \mathbb{E}\left[f(x_t, \dots, x_{t-m}, \theta(t))\right]\right)\right)^2. \tag{8}$$

For instance, a reasonable model distance function for the direct mathematical model in (6) is induced by the Euclidean norm, i.e.,

$$g(x_t, \theta(t)) = \|x_t - \theta(t)\|_2^2. \tag{9}$$

By employing a metric, the resulting function $g(\cdot)$ measures the model error as the squared distance between $x_t$ and the output of the *average* model function. Therefore, we call $g(\cdot)$ *model distance function* rather than fitness function. However, any function $g$ that is bounded from below measuring the approximation quality is admissible within the following variational framework.

With the model distance function at hand, the optimal parameters explaining the time series "best" can now formally be characterized as those satisfying the *variational problem*

$$\boldsymbol{L} \stackrel{\text{def}}{=} \sum_{t=m}^{T} g(x_t, \dots, x_{t-m}, \theta(t)) \to \min_{\theta(t) \in \Omega}. \tag{10}$$

From now on, we will refer to $\boldsymbol{L}$ as the *model distance function*. In general, the variational problem in (10) is *ill-posed* in the sense of Hadamard [26] as the parameter space $\Omega$ might be high- or even infinite-dimensional and, hence, may lead to underdetermined or trivial solutions. For instance, the variational problem associated with the model distance function in (9) admits the trivial but meaningless solution (e.g., regarding the prediction skill of such a model, it requires the exact knowledge of the infinite-dimensional function $x_t$ at all times)

$$\theta^*(t) = x_t, \quad t = 0, \dots, T. \tag{11}$$

In order to avoid such trivial solutions, the variational problem needs to be regularized.

The key idea of an appropriate regularization is based on the observation that in many real world processes the parameter function $\theta(t)$ varies much slower than the observable $x_t$ in itself. Hence, *local stationarity* of the parameter function $\theta(t)$ is a reasonable assumption, which eventually helps to overcome the ill-posedness

of the variational problem in (10). Formally, we assume the existence of $\boldsymbol{K}$ different stationary yet unknown parameters $\Theta = (\theta_1, \ldots, \theta_{\boldsymbol{K}})$ and time-dependent weights $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_{\boldsymbol{K}}(t))$ such that the model distance function $g(\cdot)$ can be expressed as a linear combination of *local* model distance functions, i.e.,

$$g(x_t, \ldots, x_{t-m}, \theta(t)) = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i), \tag{12}$$

with $(\gamma_1(t), \ldots, \gamma_{\boldsymbol{K}}(t))$ satisfying the convexity constraints

$$\begin{aligned} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) = 1, \quad & \forall t, \\ \gamma_i(t) \geq 0, \quad & \forall t, i. \end{aligned} \tag{13}$$

We call the vector $\Gamma(t)$ affiliation vector and we will use the shorthand $\Gamma = (\Gamma(t))_{t=m,\ldots,T}$. It is important to realize that, unlike in standard methods such as GMM/HMM, we do not assume the existence of $\boldsymbol{K}$ different local stationary models. Our assumption is more general since it is an assumption on the decomposability of the model error. However, as indicated by the name "affiliation vector", under certain conditions the entries of $\Gamma(t)$ can be interpreted as weights in a mixture model of local models (Section 2.g).

Inserting the interpolation ansatz (12) into the model distance function yields the *average cluster functional*

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) = \sum_{t=m}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i), \tag{14}$$

which is the key-object in the FEM-BV methodology. Additionally to the optimal (stationary) parameters $\Theta^* = (\theta_1^*, \ldots, \theta_{\boldsymbol{K}}^*)$ we seek for the optimal affiliation vectors $\Gamma^*$, which are finally characterized by the regularized variational problem

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) \to \min_{\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma} \tag{15}$$

with $\Gamma$ subject to the constraints in (13).

**2.b.** ***Numerical solution of the variational problem via the subspace algorithm.*** Even for the regularized variational problem derived from the simple model given in (6) there does not exist any analytical expression for the *global minimizer*, which is due to the nonlinearity of the average cluster functional and the convexity constraint on $\Gamma$. Fortunately, for many cases the model distance function $g(\cdot)$ is *convex* and analytical expressions for the unique optimal parameter $\Theta^*$ are available provided that $\Gamma$ is given and *fixed*. The same holds true for the optimal $\Gamma^*$ if the parameters $\Theta$ are fixed. Under weak conditions on the model distance function

**Require:** Time series $\boldsymbol{X}$, number of clusters $\boldsymbol{K}$, persistence $\boldsymbol{C}$, initial affiliations $\Gamma^0$.
**Ensure:** Locally optimal affiliations $\Gamma^*$, optimal parameters $\Theta^*$.

**Repeat until convergence**

(1) Compute $\Theta^{(s+1)}$ for fixed $\Gamma^{(s)}$ via the unconstrained minimization problem

$$\Theta^{(s+1)} = \underset{\Theta}{\operatorname{argmin}}\, \boldsymbol{L}(\Theta, \Gamma^{(s)}) \tag{16}$$

(2) Compute $\Gamma^{(s+1)}$ for fixed $\Theta^{(s+1)}$ via the constrained minimization problem

$$\Gamma^{(s+1)} = \underset{\Gamma}{\operatorname{argmin}}\, \boldsymbol{L}(\Theta^{(s+1)}, \Gamma) \tag{17}$$

subject to (13).

**Algorithm 1.** The subspace algorithm.

$g(\cdot)$ it was proven in [31] that iterating over these two steps yields an algorithm guaranteed to converge to a *local minimum* of the average cluster functional $\boldsymbol{L}$.

Throughout this paper when we speak of the *subspace algorithm*, we are actually referring to an implementation of the iterative scheme described above and formally summarized in Algorithm 1.

The subspace algorithm converges only to a local minimum. In order to find the global minimum, an annealing-like Monte Carlo strategy can be employed, i.e., the iterative procedure is started over several times with randomly initialized $\Gamma^{(0)}$. If the number of repetitions is sufficiently large then the best solution among the local minimizer is (almost sure) the global minimizer $\Gamma^*$ and $\Theta^*$. Notice that the described strategy for finding the global minimizers can straightforwardly be parallelized.

**2.c.** ***Four important models.*** In this section we introduce four important models that are broadly used in time series analysis and we derive their respective associated variational formulations. Numerical results will be given in Section 5.

**2.c.i.** *Model I: Geometrical clustering.* In the Section 2.a we introduced the simplest nontrivial model one can think of; a model without memory,

$$x_t = \theta(t) + \varepsilon_t, \tag{18}$$

where $x_t \in \mathbb{R}^d$ and $\varepsilon_t$ denotes a noise process. If we choose the model distance function induced by the Euclidean norm,

$$g(x_t, \theta(t)) = \|x_t - \theta(t)\|_2^2, \tag{19}$$

then the regularized minimization problem in (15) simplifies to

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) = \sum_{t=0}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \|x_t - \theta_i\|_2^2 \to \min_{\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma} \tag{20}$$

subject to the constraints in (13). For fixed $\Gamma$ the optimal $\Theta^* = (\theta_1^*, \ldots, \theta_{\boldsymbol{K}}^*)$ takes the form [31]

$$\theta_i^* = \frac{\sum_{t=0}^{T} \gamma_i(t) x_t}{\sum_{t=0}^{T} \gamma_i(t)}. \tag{21}$$

Furthermore, for fixed $\Theta$ the optimal affiliations are given by [33]

$$\gamma_i^*(t) = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j g(x_t, \theta_j) = \operatorname{argmin}_j \left\{ \left\| x_t - \theta_j \right\|_2^2 \right\}, \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

which readily follows from the convexity conditions in (13).

The resulting subspace algorithm has a very nice interpretation: it is the well-known and popular $k$-means algorithm for clustering geometrical data. To see that note that each affiliation vector is deterministic, i.e., exactly one component is 1.0 while the remaining ones are 0.0. If we define the set $S_i = \{x_t : \gamma_i^*(t) = 1\}$ then, by definition

$$\|x_t - \theta_i\|_2 \leq \left\| x_t - \theta_j \right\|_2 \quad \forall x_t \in S_i, \;\; j = 1, \ldots, \boldsymbol{K}, \tag{23}$$

and the optimal $\theta_i^*$ reduces to the centroid of the point set $S_i$,

$$\theta_i^* = \frac{1}{|S_i|} \sum_{x_t \in S_i} x_t. \tag{24}$$

**2.c.ii.** *Model II: Takens-PCA clustering.* A prominent example of a memoryless model exhibiting dynamics is motivated by the observation that in many applications the essential dynamics of a high-dimensional process can be approximated by a process on low-dimensional manifolds without significant loss of information [70]. Recently, several cluster methods have been introduced which are based on the decomposition of time series according to their *essential linear attractive manifolds*, allowing the analysis of data of very high dimensionality with low-dimensional dynamics [40; 29; 39; 38].

Formally, assume that the linear submanifolds are spanned by $Q(t) \in \mathbb{R}^{d \times n}$ consisting of $n \ll d$ orthonormal $d$-dimensional vectors, i.e., $Q^\dagger(t) Q(t) = \mathrm{Id}_n$ where $\mathrm{Id}_n$ denotes the $n$-dimensional identity matrix. To motivate the following direct mathematical model, suppose that $x_t$ lives on the linear subspace spanned by $Q(t)$. Orthonormality then implies

$$x_t = Q(t) Q^\dagger(t) x_t, \tag{25}$$

where $Q(t)Q^\dagger(t)$ is the orthogonal projector on the linear subspace at time $t$. However, in applications we only have $x_t \approx Q(t)Q^\dagger(t)x_t$, which leads to the general model function

$$(x_t - \mu_t) = Q(t)Q^\dagger(t)(x_t - \mu_t) + \varepsilon_t, \tag{26}$$

where the center vector $\mu_t \in \mathbb{R}^d$ is the affine translation of the linear subspace and $\varepsilon_t$ is again some noise process with $\mathbb{E}[\varepsilon_t] = 0$. As shown in [38], adopting the model distance function ($\theta(t) = (\mu(t), Q(t))$)

$$g(x_t, \theta(t)) = \left\| (x_t - \mu_t) - Q(t)Q^\dagger(t)(x_t - \mu_t) \right\|_2^2 \tag{27}$$

results in analytical closed expressions for the optimal parameters. The center vectors $\mu_i^* \in \mathbb{R}^d$ are given by

$$\mu_i^* = \frac{\sum_{t=0}^T \gamma_i(t) x_t}{\sum_{t=0}^T \gamma_i(t)} \tag{28}$$

and the optimal matrices $Q_i^*$ satisfy an eigenvalue problem, respectively,

$$\left( \sum_{t=0}^T \gamma_i(t)(x_t - \mu_i)(x_t - \mu_i)^\dagger \right) Q_i^* = Q_i^* \Lambda_i. \tag{29}$$

For fixed $\Theta$, the optimal $\Gamma^*$ is given analogously by (22).

**2.c.iii.** *Model III: Discrete (or categorical) model.* An alternative technique to capture the essential dynamics of a complex system is *coarse graining* of the process under consideration. The coarse grained process is a *discrete* process, i.e., it attains only values in a finite set of discrete objects. Prominent examples are, e.g., conformational dynamics of (bio-)molecules [67] or climate research [30].

Let $\boldsymbol{X} = (x_1, \ldots, x_T)$ be a discrete time series and without loss of generality we denote the discrete state space as $\mathcal{S} = \{1, \ldots, M\}$. In order to apply the variational framework we have to specify a model function and an appropriate model distance function that are not readily available due to the discreteness of the state space. Instead of considering the original data, the key idea here is to uniquely identify each datum $x_t$ with a discrete probability distribution $\pi_t$. More precisely, we define $\pi_t = (\pi_t(1), \ldots, \pi_t(M))$ as the discrete Dirac measure with respect to $x_t \in \mathcal{S} = \{1, \ldots, M\}$,

$$\pi_t(s) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } s = x_t, \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

Viewing discrete distributions as real valued vectors allows us to make use of, e.g., the model function given in (6), here written as

$$\pi_t = \theta_t + \varepsilon_t, \tag{31}$$

subject to the constraint that $\theta_t = (\theta_t(1), \ldots, \theta_t(M))$ is a discrete probability distribution,

$$\theta_t(s) \geq 0 \quad \text{and} \quad \sum_{s=1}^{M} \theta_t(s) = 1. \tag{32}$$

Moreover, $\varepsilon_t$ is a noise process as in the previous models.

Since we are particularly dealing with probability distributions, we define the model distance function by means of a metric tailored to respect the underlying probability space. Specifically, we chose the famous Kullback–Leibler divergence, also referred to as the relative entropy, defined as

$$d_{KL}(\mu, \eta) = \sum_{s \in \mathcal{S}} \mu(s) \log \frac{\mu(s)}{\eta(s)} \tag{33}$$

for any two discrete probability distributions $\mu$ and $\eta$ on the same probability space. For an overview of metrics and divergences on probability spaces see [25], for example.

The relative entropy directly induces a model distance function by defining

$$g(x_t, \theta_t) \stackrel{\text{def}}{=} g(\pi_t, \theta_t) \stackrel{\text{def}}{=} d_{KL}(\pi_t, \theta_t) = -\log \theta_t(x_t). \tag{34}$$

A short calculation shows that the regularized minimization problem

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) = -\sum_{t=0}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \log \theta_i(x_t) \to \min_{\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma} \tag{35}$$

subject to the constraints (13) and (32) admits analytical solutions; the optimal discrete probability distribution $(\theta_1^*, \ldots, \theta_K^*)$ takes the form

$$\theta_i^*(s) = \frac{\alpha_{i,s}}{\sum_{z \in \mathcal{S}} \alpha_{i,z}} \quad \text{with } \alpha_{i,s} = \sum_{t=0}^{T} \delta_{x_t,s} \gamma_i(t), \quad s = 1, \ldots, M \tag{36}$$

and the optimal affiliation function $\Gamma^*$ is given analogously by (22).

**2.c.iv.** *Model IV: Markov regression model.* The strategy proposed in Section 2.c.iii to analyze time series of discrete observations can loosely be described as geometrical clustering of probability distributions, geometrical in the sense that neither dynamics nor memory are assumed to be of importance.

A discrete probabilistic model including memory and dynamics is the famous Markov model. Generally, a discrete Markov process describes the evolution of a transition process between a finite number of discrete states by means of time-dependent one-step transition probabilities. If the transition probabilities are stationary (time-homogeneous) then the process is called a Markov chain and it is one of the most exploited families of processes in this class of probabilities models.

Formally, a stationary Markov process $x_t$ on a discrete state space $\mathcal{S} = \{1, \ldots, M\}$ is uniquely characterized by a time-independent transition (stochastic) matrix $P \in \mathbb{R}^{M \times M}$ (comprising of the stationary one-step transition probabilities) and an initial distribution $\pi_0 \in \mathbb{R}^M$. The evolution of the state probability vector $p(t) \in \mathbb{R}^M$, defined as

$$p_j(t) \stackrel{\text{def}}{=} \mathbb{P}[x_t = j], \quad j \in \mathcal{S}, \tag{37}$$

is then governed by the *master equation*,

$$p^\dagger(t+1) = p^\dagger(t)P, \quad t = 0, 1, 2, \ldots, T-1. \tag{38}$$

For more details on Markov chains, we refer the interested reader to, e.g., [9].

Recently in [34], the opposite question was addressed: suppose we are given a time series of *probability distributions* $(\pi_t)$, $\pi_t \in \mathbb{R}^M$, $t = 0, 1, \ldots, T$ and, additionally, a series of external data $u(t) \in \mathbb{R}^k$. What is an appropriate *nonstationary Markov regression model* explaining the given time series of distributions conditioned on the external factors best? Following the lines of the FEM-BV approach and motivated by the stochastic master Equation (38), it is reasonable to consider the direct model function

$$\pi_{t+1}^\dagger = \pi_t^\dagger P(t, u(t)) + \varepsilon_t \tag{39}$$

where $\varepsilon_t$ is a noise process as in the previous models and $P(t, u(t)) \in \mathbb{R}^{M \times M}$ is stochastic, i.e.,

$$\{P(t, u(t))\}_{vw} \geq 0 \quad \forall v, w, t, u(t), \tag{40}$$

$$P(t, u(t))\mathbf{1}_M = \mathbf{1}_M \quad \forall t, u_t \tag{41}$$

with $\mathbf{1}_M = (1, \ldots, 1) \in \mathbb{R}^M$.

Additional to depending on the (resolved) external factors $u(t) \in \mathbb{R}^k$, notice that the transition matrices may explicitly depend on the time $t$. For details see [34].

The interpolation of the model distance function

$$g\left(\pi_{t+1}, \pi_t, P(t, u(t))\right) = \left\| \pi_{t+1}^\dagger - \pi_t^\dagger P(t, u(t)) \right\|_2^2 \tag{42}$$

results in

$$g\left(\cdot, \cdot, P(t, u(t))\right) = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g\left(\cdot, \cdot, P^{(i)}(u(t))\right) \tag{43}$$

where the stationary transition matrices (parameters), $P^{(i)}(u(t)) \in \mathbb{R}^{M \times M}$ $i = 1, \ldots, \boldsymbol{K}$, have the form

$$P^{(i)}(u(t)) = P_0^{(i)} + \sum_{l=1}^{k} u_l(t) P_l^{(i)} \quad i = 1, \ldots, \boldsymbol{K} \tag{44}$$

with $P_0^{(i)}$, $P_l^{(i)} \in \mathbb{R}^{M\times M}$ $i = 1, \ldots, \boldsymbol{K}$ satisfying the constraints

$$P_0^{(i)} \geq 0 \quad \text{(elementwise)}, \tag{45}$$

$$P_0^{(i)} \mathbf{1}_M = \mathbf{1}_M, \tag{46}$$

$$P_l^{(i)} \mathbf{1}_M = 0 \quad l = 1, \ldots, k. \tag{47}$$

Notice that the constraints (45)–(47) imply $P^{(i)}(u(t))\mathbf{1}_M = \mathbf{1}_M$ independently of $u(t)$. The elementwise nonnegativity is ensured by the constraints

$$P^{(i)}(u(t)) \geq 0 \quad i = 1, \ldots, \boldsymbol{K}, \ \forall u(t), \tag{48}$$

which explicitly involve the external data $u(t)$.

Assembling the pieces together, we finally end up with the variational problem

$$\boldsymbol{L}(\Theta, \Gamma) = \sum_{t=0}^{T-1} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g\left(\pi_{t+1}, \pi_t, P_0^{(i)} + \sum_{l=1}^{k} u_l(t) P_l^{(i)}\right) \to \min_{\Theta,\Gamma} \tag{49}$$

subject to the constraints (45)–(48). Unfortunately, no analytical expressions exist for the optimal parameters due to the imposed constraints. Numerically, however, the optimal Markov regression models $P^{(i)}(t, u(t))$ are given by solutions of $\boldsymbol{K}$ independent constrained quadratic programs. For the convenience of the reader, they are stated in an Appendix.

The main challenge in numerical computation of the optimal parameters lies in the enforcement of the constraints in (48) as a linear increase in the number of external factors causes an exponentially increase in time and memory for minimizing (49). As shown in [34], the computational time and memory consumption can be reduced by exploiting that (48) attains its unique maximum/minimum in a corner of the convex hull of the set $\{u(t) : t = 0, \ldots, T\}$. Hence, it is sufficient to requiring the constraints in (48) only for the corners. For example, if the convex hull is given by an $k$-dimensional hypercube then the reduced number of constraints, $2^k$, is independent of the length of the time series. This allows to substitute the time-dependent set of constraints (48) by a time-independent set, making the entire optimization problem numerically tractable.

**2.d.** ***Regularization of*** $\boldsymbol{\Gamma}$**.** As indicated in the examples introduced in the previous section, for given parameters $\Theta$ the optimal $\Gamma^*$ is given in terms of the model distance function (compare (22), for example),

$$\gamma_i^*(t) = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j \left\{ g(x_t, \ldots, x_{t-m}, \theta_j) \right\}, \\ 0 & \text{otherwise}, \end{cases} \tag{50}$$

where each datum $x_t$, $t \geq m$ is uniquely (deterministically) assigned to a single cluster. However, even for the global optimal parameters $\Theta^*$, the resulting optimal

$\Gamma^*$ might be a highly nonregular function. For instance, $\Gamma^*$ might rapidly oscillate between the $\boldsymbol{K}$ different clusters rather than describing a smooth and persistent transition process. In other words, the optimal $\Gamma^*$ does not *continuously* depend on the data, which is again a violation of Hadamard's postulate of a *well-posed* problem. Consequently, the variational problem has to be regularized again.

One approach is to first incorporate some *additional information* about the regularity of the observed process by restricting the time dependent function $\Gamma(\cdot)$ on an appropriate function space and then apply a finite Galerkin discretization of this infinite-dimensional Hilbert space. In the context of Tikhonov-based FEM-BV methodology, this was done by restricting the functions $\gamma_i(\cdot)$ on the function space of *weakly differentiable* functions. One way to incorporate this *a priori information* into the optimization is to modify the variational problem in (15) by writing it in the *Tikhonov-regularized* form [31]

$$\boldsymbol{L}^{\varepsilon}(\Theta, \Gamma, \varepsilon^2) \stackrel{\text{def}}{=} \boldsymbol{L}(\Theta, \Gamma) + \varepsilon^2 \sum_{i=1}^{\boldsymbol{K}} \|\partial_t \gamma_i\|^2_{L_2(0,T)} \to \min_{\gamma_1,\ldots,\gamma_{\boldsymbol{K}} \in H^1(0,T),\Theta}, \quad (51)$$

where the norm $\|\partial_t \gamma_i\|^2_{L_2(0,T)} = \int_0^T (\partial_t \gamma_i(t))^2 \mathrm{d}t$ measures the smoothness of the function $\gamma_i(\cdot)$. A similar form of penalized regularization was first introduced by A. Tikhonov to solve ill-posed linear least-squares problems [71] and has been frequently used for nonlinear regression analysis in the context of statistics [28] and multivariate spline interpolation [74].

The main problem one faces in this approach is the lack of the direct control of the persistence of $\gamma_i$. To be more precise, Tikhonov regularization does not allow us to directly incorporate the desired *persistence constraints*

$$\|\partial_t \gamma_i\|^2_{L_2(0,T)} \le \boldsymbol{C}, \quad i = 1, \ldots, \boldsymbol{K}, \quad (52)$$

where $0 \le \boldsymbol{C}$ bounds the smoothness of the functions $\gamma_i(\cdot)$. Another disadvantage of the $H^1$ approach is the exclusion of functions with discontinuities such as jumps, which is due to the requirement of weak differentiability. Fortunately, the two problems can be overcome by considering a larger function space.

**2.e.** ***Persistence in the BV sense.*** A less restrictive class of functions is the class of functions with *bounded variation* $BV([0, T])$, consisting of functions $f : [0, T] \to \mathbb{R}$ with

$$\|f\|_{BV} = \sup_{0=t_0<t_1<\ldots<t_M=T} \left\{ \sum_{i=0}^{M-1} |f(t_{i+1}) - f(t_i)| \right\} < \infty, \quad (53)$$

where the supremum is taken over all partitions of the interval $[0, T]$. Notice that in the time-continuous case $H^1(0, T) \subset BV(0, T)$ holds true (cf. [58]), so "smooth" $H^1$-transitions between cluster states are not excluded. However, the BV-norm of

a function does not require any notion of differentiability and the class $BV[0,T]$ covers transition processes with jumps between clusters.

For the remainder of this section, the memory depth $m$ is, without loss of generality, assumed to be zero. In the following, we consider the functions $\gamma_i,\ i = 1,\ldots,\boldsymbol{K}$ as *discrete* functions (vectors), which is emphasized by denoting $\gamma_i \in \mathbb{R}^{T+1}$. Now we are prepared to formulate the persistence condition in the time-discrete BV sense:

$$\|\gamma_i\|_{BV} = \sum_{t=0}^{T-1} |\gamma_i(t+1) - \gamma_i(t)| \leq \boldsymbol{C}, \quad i = 1,\ldots,\boldsymbol{K}, \tag{54}$$

where $0 \leq \boldsymbol{C}$ is an upper bound for the maximal number of transitions between the cluster state $i$ and the remaining ones. In the rest of this section we will show that the additional BV-constraints lead to a numerically convenient characterization of $\Gamma$ via a *linear minimization problem with linear constraints*.

To this end, for given $\Theta = (\theta_1,\ldots,\theta_{\boldsymbol{K}})$ we define the *row vectors*

$$g_{\theta_i} = (g(x_0,\theta_i),\ldots,g(x_T,\theta_i)) \in \mathbb{R}^{T+1}, \tag{55}$$

$$\gamma_i = (\gamma_i(0),\ldots,\gamma_i(T)) \in \mathbb{R}^{T+1}. \tag{56}$$

Then, the variational problem in (15) transforms to

$$\boldsymbol{L}(\theta_1,\ldots,\theta_{\boldsymbol{K}},\Gamma) = \sum_{i=1}^{\boldsymbol{K}} \langle \gamma_i, g_{\theta_i} \rangle_2 \to \min_{\Gamma,\Theta}, \tag{57}$$

subject to the constraints

$$\|\gamma_i\|_{BV} \leq \boldsymbol{C} \quad i = 1,\ldots,\boldsymbol{K}, \tag{58}$$

$$\sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) = 1 \quad t = 0,\ldots,T, \tag{59}$$

$$\gamma_i(t) \geq 0 \quad t = 0,\ldots,T,\ i = 1,\ldots,\boldsymbol{K}. \tag{60}$$

Unfortunately, the additional constraints (58) turn the variational problem in (57) into a *nondifferentiable* one. As a remedy, we retransform the problem into a differentiable one by applying an upper-bound technique.

Suppose we had $\eta_i(0),\ldots,\eta_i(T-1) \in \mathbb{R}$ satisfying the constraints

$$|\gamma_i(t+1) - \gamma_i(t)| \leq \eta_i(t) \quad t = 0,\ldots,T-1, \tag{61}$$

$$\sum_{t=0}^{T-1} \eta_i(t) \leq \boldsymbol{C}, \tag{62}$$

$$\eta_i(t) \geq 0 \quad t = 0,\ldots,T-1, \tag{63}$$

then $\gamma_i$ would satisfy the BV-constraint in (58). The key observation is that (61) holds true for $t \geq 0$ if and only if the following two *linear* inequalities hold true:

$$\gamma_i(t+1) - \gamma_i(t) - \eta_i(t) \leq 0, \tag{64}$$

$$-\gamma_i(t+1) + \gamma_i(t) - \eta_i(t) \leq 0. \tag{65}$$

Consequently, if the upper bounds $\eta_i = (\eta(0), \ldots, \eta(T-1))$ are considered as additional unknowns (additional to the unknowns $\gamma_i$), then the BV-constraint in (58) is satisfied if and only if the linear constraints (62)–(65) are satisfied.

Notice that the constraints (59)–(60) are *linear* constraints too. Finally, by defining

$$\omega = (\gamma_1, \ldots, \gamma_{\boldsymbol{K}}, \eta_1, \ldots, \eta_{\boldsymbol{K}}) \in \mathbb{R}^{\boldsymbol{K}(2T+1)}, \tag{66}$$

$$c(\Theta) = (g_{\theta_1}, \ldots, g_{\theta_{\boldsymbol{K}}}, \underbrace{0, \ldots, 0}_{\boldsymbol{K}T \text{ times}}) \in \mathbb{R}^{\boldsymbol{K}(2T+1)} \tag{67}$$

we can express the original *nondifferentiable* optimization problem (57)–(60) as the following *differentiable* optimization problem,

$$\langle c(\Theta), \omega \rangle_2 \to \min_{\omega, \Theta} \tag{68}$$

subject to

$$\begin{aligned} A_{\text{eq}}\omega &= b_{\text{eq}}, \\ A_{\text{neq}}\omega &\leq b_{\text{neq}}, \\ \omega &\geq 0, \end{aligned} \tag{69}$$

where $A_{\text{eq}}$ and $b_{\text{eq}}$ readily result from the constraints (59) and $A_{\text{neq}}$ and $b_{\text{neq}}$ from (60) and ((62)–(65)).

The solution of the above minimization problem can be approached via the subspace iteration procedure presented in Section 2.b. Particularly, for fixed $\Theta$ the problem reduces to a standard *linear program*, which can efficiently be solved by standard methods such as the Simplex method or interior point method. Completely analogously to the Tikhonov-regularized FEM-BV methodology [31], it can be demonstrated that the iterative procedure converges towards a local minimum of the problem (68)–(69) if some appropriate assumptions (convexity and differentiability) of the model distance function (8) are fulfilled.

Unfortunately, since the dimensionality of the variable $\omega$ scales as $\boldsymbol{K}(2T+1)$ the numerical solution of the problem (68)–(69) for a fixed value of $\Theta$ becomes increasingly expensive for long time series. Therefore a Finite Element Method (FEM) will be introduced in the next section to reduce the dimensionality of the above problem in a robust and controllable numerical manner.

**2.f.** ***FEM discretization.*** Solving the problem (68)–(69) is numerically expensive or even practically impossible for long time series, in terms of computational time as well as in terms of memory usage. To overcome these limitations, a FEM is proposed to reduce the dimensionality of the problem.

The idea is to approximate the (unknown) discrete functions $\gamma_i(t)$ by a linear combination of $N \ll T+1$ continuous functions $\{f_1(t), f_2(t), \ldots, f_N(t)\}$ with bounded variation, i.e.,

$$\gamma_i(t) = \sum_{j=1}^{N} \alpha_{ij} f_j(t) \quad t = 0, \ldots, T+1. \tag{70}$$

Traditionally, the finite element functions $f_j(t) \in BV[0,T]$ are defined as nonconstant functions on overlapping supports. For practical examples of standard finite element functions see, e.g., [8]. Here, however, we approximate the functions $\gamma_i$ with *constant* ansatz functions defined on *nonoverlapping* supports. This approach is justified by the fundamental assumption that the time series under consideration is *persistent*.

Let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_N = T$ be a partition dividing $[0,T]$ into $N$ bins $[\tau_j, \tau_{j+1}]$, $j = 0, \ldots, N-1$ with $\tau_j \notin \mathbb{N}$, $j = 1, \ldots, N-1$ and assume that all the $\gamma_i$ are piecewise constant on each of the intervals $[\tau_j, \tau_{j+1}]$. Moreover, let $\hat{\gamma}_i(j)$ denote the value of $\gamma_i$ on $[\tau_j, \tau_{j+1}]$ and define

$$\hat{g}_{\theta_i}(j) \stackrel{\text{def}}{=} \sum_{t \in [\tau_j, \tau_{j+1}]} g_{\theta_i}(t). \tag{71}$$

Then, the variation problem in (57) reduces to

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \hat{\Gamma}) = \sum_{i=1}^{\boldsymbol{K}} \left\langle \hat{\gamma}_i, \hat{g}_{\theta_i} \right\rangle_2 \to \min_{\hat{\Gamma}, \Theta}, \tag{72}$$

with $\hat{\gamma}_i \in \mathbb{R}^N$, $\hat{g}_{\theta_i} \in \mathbb{R}^N$ and subject to the constraints

$$\|D\hat{\gamma}_i\|_1 \leq \boldsymbol{C} \quad i = 1, \ldots, \boldsymbol{K}, \tag{73}$$

$$\sum_{i=1}^{\boldsymbol{K}} \hat{\gamma}_i(t) = 1 \quad t = 0, \ldots, N-1, \tag{74}$$

$$\hat{\gamma}_i(t) \geq 0 \quad t = 0, \ldots, N-1, \ i = 1, \ldots, \boldsymbol{K}. \tag{75}$$

Analogously to the derivation given in the previous section, we finally end up with the FEM discretization (in the BV sense) of the original variational problem

in (15),

$$\left\langle \hat{c}(\Theta), \hat{\Gamma} \right\rangle_2 \to \min_{\hat{\Gamma}, \Theta} \tag{76}$$

subject to the linear constraints (73)–(75).

Notice that the number of unknowns has reduced to $\boldsymbol{K}(2N+1)$ being much less than $\boldsymbol{K}(2T+1)$ if $N \ll T$. Particularly, the number of unknowns and, hence, the number of constraints does not explicitly depend on the total length $T+1$ of the time series anymore. Hence, the final variational problem allows the analysis of long time series from real-world applications, as will be demonstrated in Section 5.

**2.g.** ***Identification of local models.*** The derivation of the average cluster functional is based on the assumption that the model distance at a fixed time $t$ can be represented by a convex combination of model distances with respect to $\boldsymbol{K}$ stationary model parameters. Notice that this assumption is more general than the assumption of the existence of $\boldsymbol{K}$ *local stationary models*. Nevertheless, the identification of local stationary models gives additional insight into the data. More importantly, it allows the simulation and prediction of time series, which ultimately leads to constructing self-contained predictive models as will be explained in Section 4 below.

The identification of local stationary models depends crucially on the choice of the model distance function and the derived optimal affiliation function $\Gamma^*$. To see that, recall the formal interpolation ansatz in (12), i.e.,

$$g(x_t, \ldots, x_{t-m}, \theta(t)) = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i). \tag{77}$$

Accordingly, if we could find an $\theta(t)$ such that (77) held true then the local model at time $t$ would be given by $f(\cdot; \theta(t))$.

First suppose that the optimal $\Gamma^*$ is deterministic, i.e., $\gamma_i^*(t) \in \{0, 1\}$. But this immediately implies

$$\theta(t) = \theta_i \quad \text{with } \gamma_i^*(t) = 1, \tag{78}$$

as the ansatz trivially holds true with that choice. In the case of a nondeterministic $\Gamma^*$ the identification crucially depends on the model distance function. We exemplify that by considering the model distance function

$$g(x_t, \ldots, x_{t-m}, \theta(t)) = \left\| x_t - \mathbb{E}\left[ f(x_t, \ldots, x_{t-m}, \theta(t)) \right] \right\|_2^2. \tag{79}$$

**Theorem 2.1.** *If the direct model function $f$ is* linear *in $\theta$ then*

$$g\left(x_t, \ldots, x_{t-m}, \sum_{i=1}^{K} \gamma_i(t)\theta_i\right) \leq \sum_{i=1}^{K} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i) \tag{80}$$

The proof is straightforward and left for the interested reader. Consequently, if the interpolation on the right-hand side in (80) is small then the model distance function on the left-hand side with respect to $\theta(t) = \sum_{i=1}^{K} \gamma_i(t)\theta_i$ is small too. This, in turn, implies that the direct model function with respect to $\theta(t)$ is a good approximation for a local model function at time $t$.

The minimization of the average cluster functional justifies the notion

$$x_t \approx \hat{x}_t \stackrel{\text{def}}{=} \mathbb{E}\left[ f(x_t, \ldots, x_{t-m}, \sum_{i=1}^{K} \gamma_i^*(t)\theta_i^*) \right]. \tag{81}$$

However, the identification is only valid if the direct model function is linear with respect to its parameters and the model distance function is strict convex. This is the case for the model distance functions, e.g., in (19), (27), (34) and (42) described above.

**2.h.** ***Relation to classical methods of unsupervised learning.*** We have already seen that the direct model $x_t = \theta(t) + \varepsilon_t$ equipped with the model distance function $g(x_t, \theta(t)) = \|x_t - \theta(t)\|_2^2$ leads to the classical $k$-means algorithm for geometric clustering provided that no regularity condition ($\boldsymbol{C} = \infty$) is imposed on the affiliation function $\Gamma$ (Section 2.c.i) and no FEM discretization is used for the numerical solution of the resulting variational problem. In this section we further clarify the link between the FEM-BV approach and classical methods for dynamical clustering. Particularly, we show that the presented method covers existent probabilistic approaches as special cases by choosing specific model distance functions and regularity constraints.

Let us first consider the discrete case, i.e., $x_t \in \mathcal{S} = \{1, \ldots, M\}$. A prominent approach for dynamical clustering of persistent discrete time series is the hidden Markov model [64]. Basically, it relies on three strong assumptions. Firstly, it is assumed that the hidden (persistent) process is governed by a time-homogeneous stationary Markov process. Secondly, it is assumed that an observation $x_t$ (triggered by a jump of the hidden process) is distributed according to a stationary distribution conditional on the current hidden state. Finally, one has to assume that the observations are independent.

Here we make the most general assumption by imposing that the hidden process is nonstationary and non-Markovian. Specifically, we assume that an observation $x_t$ is distributed according to a discrete distribution $\theta_i \in \mathbb{R}^{|S|}$ conditional on a hidden state $i \in \{1, \ldots, \boldsymbol{K}\}$, which in turn is drawn from a discrete distribution $\Gamma(t) \in \mathbb{R}^{\boldsymbol{K}}$.

Under the additional assumption of independence, the likelihood of a time series $\boldsymbol{X} = (x_t)$, $t = 0, \ldots, T$ takes the form

$$\mathscr{L}(\boldsymbol{X}; \Gamma, \Theta) = \prod_{t=0}^{T} \Big( \sum_{i=1}^{K} \gamma_i(t) \theta_i(x_t) \Big), \tag{82}$$

where we marginalize over the hidden states.

**Theorem 2.2.** *If the model distance function is defined as*

$$g(x_t, \theta_i) = -\log(\theta_i(x_t)) \tag{83}$$

*then the associated average cluster functional is an upper bound of the negative log-likelihood*,

$$-\log \mathscr{L}(\boldsymbol{X}; \Gamma, \Theta) \leq L(\Gamma, \Theta). \tag{84}$$

*Proof.* Notice that $-\log x$ is a convex function. Hence, by applying Jensen's inequality we conclude

$$\begin{aligned} -\log \mathscr{L}(\boldsymbol{X}; \Gamma, \Theta) &= -\sum_{t=0}^{T} \log \Big( \sum_{i=1}^{K} \gamma_i(t) \theta_i(x_t) \Big) \\ &\leq \sum_{t=0}^{T} \sum_{i=1}^{K} \gamma_i(t) \left( -\log(\theta_i(x_t)) \right), \end{aligned} \tag{85}$$

where the upper bound in (85) is exactly *the average cluster functional* in (35) resulting from the reasoning in the third example in Section 2.c.iii. □

In the probabilistic approach, the optimal parameters (distributions) of the model are characterized by the ones that maximize the likelihood, i.e.,

$$(\Gamma^*, \Theta^*) = \underset{\Gamma, \Theta}{\operatorname{argmax}}\, \mathscr{L}(\boldsymbol{X}; \Gamma, \Theta), \tag{86}$$

which is equivalent to minimizing the negative log-likelihood function,

$$(\Gamma^*, \Theta^*) = \underset{\Gamma, \Theta}{\operatorname{argmin}}(-\log \mathscr{L}(\boldsymbol{X}; \Gamma, \Theta)). \tag{87}$$

Therefore, the minimizer of the average cluster functional in (35) can be considered as a good approximation of the maximizer $\Gamma^*, \Theta^*$ of the likelihood function in (82). The fundamental difference between the two approaches, however, is that in the FEM-BV approach non of the probabilistic assumptions on the nature of data have to be made in order to derive the average cluster functional (35).

The presented reasoning readily carries over to the continuous case, i.e., $x_t \in \mathbb{R}^d$, by defining the model distance function in terms of the assumed underlying

conditional probability density function $\rho(\,\cdot\,;\theta_i)$,

$$g(x_t,\theta_i) \stackrel{\text{def}}{=} -\log\rho(x_t;\theta_i). \tag{88}$$

It is straightforward to show that the upper bound for the negative log-likelihood associated with probabilistic model coincides with the average cluster functional resulting from the model distance function in (88).

For example, a widely used class of parametric probability density functions are the $d$-dimensional Gaussian distributions,

$$\rho_G(x_t;\mu_i,\Sigma_i) = ((2\pi)^d|\Sigma|)^{-1/2}\exp\!\big(-\tfrac{1}{2}(x_t-\mu_i)^\dagger\Sigma_i^{-1}(x_t-\mu_i)\big), \tag{89}$$

with mean $\mu_i\in\mathbb{R}^d$ and symmetric positive definite covariance matrix $\Sigma_i\in\mathbb{R}^{d\times d}$. The induced model distance function then reads

$$g(x_t,\mu_i,\Sigma_i) = \tfrac{1}{2}\big(\text{cst.}+\ln|\Sigma|+(x_t-\mu_i)^\dagger\Sigma_i^{-1}(x_t-\mu_i)\big). \tag{90}$$

Any method for inferring the optimal parameters of a Gaussian distribution relies specifically on the assumption that the data "lives" in the full $d$-dimensional space so that the covariance matrix is symmetric positive definite and, hence, invertible. Unfortunately, in many applications this assumption is not met because, e.g., the essential dynamics of a (Gaussian) process takes place in an $n$-dimensional submanifold with $n\ll d$. In the FEM-BV approach, this limitation can be circumvented by directly clustering with respect to the submanifolds by means of the PCA approach presented in Section 2.c.ii.

At the end of this section, we comment on the relation of the FEM-BV approach based on (90) to the stationary Gaussian mixture model (GMM). Analogously to the reasoning above, the negative log-likelihood associated with a GMM can be bounded from above, i.e.,

$$-\sum_t\log\Big(\sum_{i=1}^{\boldsymbol{K}}a_i\,\rho_G(x_t;\mu_i,\Sigma_i)\Big) \le -\sum_t\sum_{i=1}^{\boldsymbol{K}}a_i\log\rho_G(x_t;\mu_i,\Sigma_i), \tag{91}$$

where $a=(a_1,\ldots,a_{\boldsymbol{K}})$ are the normalized weights of the Gaussian distributions, i.e., $\sum_{i=1}^{\boldsymbol{K}}a_i=1$ and $a_i\ge 0,\ i=1,\ldots,\boldsymbol{K}$. Now notice that the upper bound in (91) coincides with the average cluster function induced by (90) if we assume that in (54) $\boldsymbol{C}=0$, i.e., $\Gamma(t)\equiv a\ \forall t$. However, the associated optimal affiliation function,

$$a_i^* = \begin{cases}1 & \text{if } i=\operatorname{argmin}_j\big\{-\sum_t\log\rho_G(x_t;\mu_j,\Sigma_j)\big\},\\ 0 & \text{otherwise,}\end{cases} \tag{92}$$

is deterministic, implying that the optimal substitute model (Gaussian mixture model) consists only of one locally stationary model (Gaussian distribution) *independent* of the number $\boldsymbol{K}$ of assumed clusters.

In contrast, the update formula for the weights $a_1, \ldots, a_{\boldsymbol{K}}$ in the classical GMM framework (see, e.g., [61]),

$$a_i^{(s+1)} = \frac{1}{T+1} \sum_{t=0}^{T} \frac{q(i,t)}{\sum_{j=1}^{\boldsymbol{K}} q(j,t)} \quad \text{with } q(i,t) = a^{(s)} \log \rho_G(x_t; \mu_i^{(s)}, \Sigma_i^{(s)}), \quad (93)$$

significantly differs from (92) and, generally, does not lead to a degenerated (deterministic) cluster affiliation as in the FEM-BV approach presented above.

This observation allows the conclusion that the upper bound derived in the GMM framework is sharper than the corresponding average cluster function, e.g., in the right-hand side of (91). However, the assumption of stationary weights ($\boldsymbol{C} = 0$) deployed in the GMM framework is very restrictive and it is not fulfilled in many applications.

## 3. Model selection

The outcome of the FEM-BV methodology crucially depends on the specific choice of the number of clusters $\boldsymbol{K}$ and the persistence threshold $\boldsymbol{C}$ as the choice expresses a certain *a priori* knowledge on the nature of the data under consideration. In fact, the identification of an optimal or best model among a set of possible models is an important part of the clustering procedure itself. In this section we briefly discuss several approaches that have been proposed in the context of the FEM-BV methodology for the selection of the optimal parameters. Furthermore, we present an extension of a recently introduced *information-theoretical* framework that allows the *simultaneous* identification of the optimal parameters $\boldsymbol{K}$ and $\boldsymbol{C}$.

The characterization of an optimal model in terms of its parameters $\boldsymbol{K}$ and $\boldsymbol{C}$ on the basis of the average cluster function, $\boldsymbol{L}(\boldsymbol{K}, \boldsymbol{C})$, is hampered by the following fact: if the number of clusters and the number of allowed transitions between them is increased then the corresponding a priori knowledge is less restrictive and, therefore, the value of the $\boldsymbol{L}(\boldsymbol{K}, \boldsymbol{C})$ decreases. Particularly, $\boldsymbol{L}(\boldsymbol{K}, \boldsymbol{C})$ attains its minimum in the limit $\boldsymbol{K} = N, \boldsymbol{C} = \infty$, which would imply that the corresponding model is optimal in the sense that it explains the data best. As explained in Section 2.d, however, the resulting model is meaningless due to the over-fitting and does not reveal any insights in the underlying data. Therefore, a criterion for selecting the *optimal* parameters should take both into account: how well the data is explained and the total number of involved parameters such as the number of clusters, the actual number of transitions between the clusters and the number of model parameters in each cluster.

Several approaches have been proposed to tackle the problem of selecting an optimal model within the context of the FEM-BV methodology. For instance, the approach in [63] is based on the following observation. The increase of the number

of clusters leads to an increase of uncertainty of the estimated model parameters for each cluster as less data is assigned. Consequently, if one starts with a large number of clusters, then this number can be reduced by combining the clusters whose parameters have a nonempty intersection of their confidence intervals as those clusters are statistically not distinguishable. The procedure is terminated if all clusters are statistically distinguishable.

To choose the optimal persistence threshold $\boldsymbol{C}$, techniques such as the L-Curve method [50] can be applied. The idea is to analyze the graph of the average clustering functional as a function of the persistence threshold $\boldsymbol{C}$. The optimal $\boldsymbol{C}^*$ is then characterized by the point of maximum curvature of the graph.

Recently in [33], an information theoretical framework has been introduced for the *simultaneous* identification of the optimal parameters $\boldsymbol{K}^*$ and $\boldsymbol{C}^*$. It is motivated by the principle of *Occam's razor*: the best or optimal model among a set of possible models is the one that exhibits maximal model quality (goodness of fit) while its number of free parameters is minimal. The most prominent information measure embodying that principle is the AIC (Akaike information criterion, introduced in [1]), which, formally, is given by

$$\mathrm{AIC}(M) = -2\ln \mathscr{L}(M) + 2|M|, \tag{94}$$

where $\mathscr{L}(M)$ denotes the *likelihood* of the model $M$ and $|M|$ is the total number of the model's free parameter. The optimal model $M^*$ is then characterized by the one that minimizes the criterion.

The AIC depends on the likelihood $\mathscr{L}(M)$ of the model as a measurement of the model quality. Therefore, the criterion can not be generally applied in the FEM-BV methodology because it is based on the more general notion of a *model distance function*.

If the model distance function, however, is induced by, e.g., a discrete probability distribution (cf. (34) in Section 2.c.iii) then as justified by Theorem 2.2 (see Section 2.h) the likelihood $\mathscr{L}(M)$ reduces to the likelihood given in (82). Analogously, the reasoning carries over to a model distance function defined in terms of a PDF (cf. (90) in Section 2.h) and to a model function preserving probability such as the Markov regression model introduced in Section 2.c.iv.

It remains to consider the case, e.g., FEM-BV-$k$-means, if neither the model function nor the model distance function allows a probabilistic interpretation. Fortunately, the gap can be bridged by realizing that the *distribution of the scalar time series of model distances with respect to a fixed cluster $i$* reflects how well the corresponding local model explains the data. The key idea now is to employ these distribution in order to define a likelihood of a scalar process and, eventually, to arrive at a modified information criterion for detecting the optimal model in the FEM-BV approach.

Let $\mathrm{supp}(\gamma_i) = \{t : \gamma_i(t) > 0\}$ denote the support of $\gamma_i(t)$ and suppose for a moment that the model distances in the cluster $i = 1, \ldots, \boldsymbol{K}$ are each distributed according to a parametric (conditional) probability density function (PDF) $\rho_i(\cdot\,; \Lambda_i)$, i.e.,

$$\mathbb{P}[g(x_t, \theta_i) \in \mathrm{d}x] = \rho_i(g(x_t, \theta_i); \Lambda_i)\,\mathrm{d}x, \quad i = 1, \ldots, \boldsymbol{K}, \ \forall t \in \mathrm{supp}(\gamma_i). \tag{95}$$

Under the (restrictive) assumption of independence, we can define a likelihood function $\mathscr{L}(\boldsymbol{K}, \boldsymbol{C})$ by

$$\mathscr{L}(\boldsymbol{K}, \boldsymbol{C}) \stackrel{\mathrm{def}}{=} \prod_t \left( \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \rho_i(g(x_t, \theta_i); \Lambda_i) \right) \tag{96}$$

and, following the arguments from the original proof by Akaike [1], we arrive at the modified information criterion

$$mAIC(\boldsymbol{K}, \boldsymbol{C}) = -2\ln(\mathscr{L}(\boldsymbol{K}, \boldsymbol{C})) + 2|M(\boldsymbol{K}, \boldsymbol{C})|. \tag{97}$$

The total number of the model's free parameters, $|M(\boldsymbol{K}, \boldsymbol{C})|$, consists of three contributions; the total number of local stationary parameters, i.e., $|\Theta| = |\theta_1| + \cdots + |\theta_{\boldsymbol{K}}|$, the total number of parameters needed for describing the conditional PDFs, i.e., $|\Lambda| = |\Lambda_1| + \cdots + |\Lambda_{\boldsymbol{K}}|$ and, finally, the total number of parameters needed to represent the affiliation function $\Gamma$. To determine $|\Gamma|$, please recall that $\Gamma$ is piecewise constant on a FEM-partition $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{N-1} < \tau_N = T$ dividing the interval $[0, T]$ into $N$ bins (Section 2.f). Hence, we conclude

$$|\Gamma| = \boldsymbol{K}N. \tag{98}$$

For instant, the total number of parameters in the FEM-BV-$k$-means model is (Section 2.c.i)

$$|M_{k\text{-means}}(\boldsymbol{K}, \boldsymbol{C})| = \boldsymbol{K}d + \boldsymbol{K}N + |\Lambda|. \tag{99}$$

It remains to explain how to characterize the set of parametric PDFs, $\{\rho_i(\cdot\,; \Lambda_i)\}$, capturing the respective distribution of the cluster's model distances appropriately. One option is to *assume* that all distributions during the course of optimization belong to a certain but fixed class of parametric PDFs, e.g., the class of Gaussians. The parameters $\Lambda_i$ are then efficiently calculated via the maximum likelihood approach. However, our numerical experiments showed that the assumption of a fixed class of parametric PDFs is too restrictive and may lead to wrong optimal models.

To motivate the approach presented here, note that we actually do not know anything about the parametric representations of the distributions. What we can empirically compute, however, are statistical properties such as the expectation, the variance and, more generally, the first $k$ noncentralized moments. The key idea now

is to choose the *most unbiased distribution* in each case, among those exhibiting the empirical observed statistical properties. According to [43; 44; 55] the most unbiased distribution is the one which admits the *most uncertainty measured in terms of entropy.*

Let $\eta_j$, $j = 0, \ldots, k$ be empirical estimates of the first $k+1$ noncentralized moments of a distribution with $\eta_0 = 1$. The associated *maximum entropy distribution* is characterized by a constrained variational problem

$$\mathcal{H}(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \ln \rho(x)\, \mathrm{d}x \to \max_{\rho(x)\in L^2(\mathbb{R})} \tag{100}$$

subject to

$$\eta_j = \int x^j \rho(x) \mathrm{d}x, \quad j = 0, \ldots, k, \tag{101}$$

where $\mathcal{H}(\rho)$ is the entropy of the PDF $\rho$.

Applying the calculus of variation yields the formal (unique) solution

$$\rho^*(x) = \exp \sum_{j=0}^{k} \lambda_j x^j = \operatorname*{argmax}_{v(x)\in L^2(\mathbb{R})} \mathcal{H}(\rho), \tag{102}$$

where the Lagrange multipliers $\lambda_0, \ldots, \lambda_k$ enforce the constraints in (101). For instant, if $k = 2$ then $\rho^*$ is basically given by a Gaussian distribution having the prescribed moments. Unfortunately, for $k > 2$ no closed expression for $\rho^*$ exists so that the Lagrange multipliers have to be computed numerically via, e.g., the Newton method. For details on solving the problem (100)–(101) numerically see, e.g., [76]. Moreover, for an overview on maximum entropy distributions associated with constraints other than in (101) we refer to, e.g., [46; 55].

The maximum entropy ansatz finally allows us to characterize the parametric representations of the distributions of the respective (scalar) cluster's model distances

$$\{g(x_t, \theta_i)\}, \quad t = 0, 1, \ldots, T, \ i = 1, \ldots, \boldsymbol{K} \tag{103}$$

as

$$\rho_i(x, \lambda_0^{(i)}, \ldots, \lambda_k^{(i)}) = \exp \sum_{j=0}^{k} \lambda_j^{(i)} x^j \tag{104}$$

subject to

$$\int x^j \rho_i(x, \lambda_0^{(i)}, \ldots, \lambda_k^{(i)}) \mathrm{d}x = Z_i^{-1} \sum_{t\in \mathrm{supp}(\gamma_i)} (g(x_t, \theta_i))^j \quad j = 0, \ldots, k, \tag{105}$$

with $Z_i = |\mathrm{supp}(\gamma_i)|$. Inserting (104) in (94) we end up with the *modified* AIC, denoted by $mAIC(\boldsymbol{K}, \boldsymbol{C})$, for selecting the optimal model within the FEM-BV

methodology. Notice that we only require in (104) and (105) the scalar "observables" $g(\cdot, \theta_i)$ to be i.i.d.[2] Furthermore, the optimal number (order) $k$ of moments needed to approximate the underlying distribution can again be determined by employing the AIC.

We end this section by discussing a conceptual weakness of the presented model selection approach. Despite its successful application and the numerical evidence indicating its usefulness (see Section 5 below), the approach theoretically suffers from the fact that the estimation of the ME-distributions is invariant under translation, i.e., the ME-distributions estimated from, e.g., the scalar time series $(g(x_t, \theta^*)), t \geq 0$ and $(g(x_t, \theta^*) + a),\ a > 0, t \geq 0$ would be indistinguishable from the view point of likelihood. Consequently, they would equally contribute to the modified AIC although the former distribution is closer to the lower bound, (say zero), and, hence, the associated underlying model should be the preferred one. From the practical point of view, such scenarios are very unlikely to happen since the model distance function $g(x_t, \theta)$ is minimized during the subspace-procedure. In fact, the occurrence of such a scenario would indicate that the underlying model function $f(\cdot, \theta(t))$ does not properly capture the dynamic of the time series under consideration.

Generally spoken, the model selection approach theoretically suffers from not explicitly incorporating the lower boundedness of the model distance function $g(\cdot)$. Bridging that gap is subject to ongoing research and will be discussed in a forthcoming manuscript.

## 4. Self-containing predictive models

In the previous section, we presented for the FEM-BV approach a tailored strategy to identify an optimal stochastic model in terms of the optimal number of clusters $\boldsymbol{K}^*$ and the optimal persistence $\boldsymbol{C}^*$. Furthermore, we elaborated in Section 2.g under which conditions the optimal model parameters and the optimal cluster affiliations lead to a time-dependent mixture model for fitting the data best within the trained time interval. In this section we present a *prediction strategy* allowing us to predict the dynamics beyond the trained time interval.

Let $\Gamma^*_{[m,T]}$ and $\theta^*_1, \ldots, \theta^*_{\boldsymbol{K}^*}$ be the parameters of the optimal model associated with a model function $f(\cdot, \cdot)$ on the time interval $[m, T]$. A reasonable fitting (prediction) at $t \in [m, T]$, i.e., within the trained time span, is then given by the

[2] In this context it is important to recall the standard application of information functionals for Bayesian time series analysis methods (such as GMMs and HMMs) [21; 49] relies on a very restrictive additional assumption, namely that the analyzed data $x_t$ are produced by a known parametric multivariate distribution.

*average mixture model* (cf. (81) in Section 2.g)

$$\hat{x}_t = \mathbb{E}\left[ f\left( x_t, x_{t-1}, \ldots, x_{t-m}, \sum_{i=1}^{K^*} \gamma^*_{i,[m,T]}(t)\theta^*_i \right) \right]. \tag{106}$$

Now it is important to realize that the average mixture model is confined on the interval $[m, T]$ because it explicitly depends on the time-dependent affiliation function $\Gamma^*_{[m,T]}$ being only well defined on $[m, T]$. However, if we could predict the affiliation function $\hat{\Gamma}_{[m,T+d]}(t)$ for $t = T+1, \ldots, T+d,\ d > 0$ then (106) could readily be extended for predicting $\hat{x}_{T+1}, \ldots, \hat{x}_{T+d}$ by the following recursive scheme

$$\hat{x}_{T+r} = \mathbb{E}\left[ f\left( \hat{x}_{T+r}, \ldots, \hat{x}_{T+r-m}, \sum_{i=1}^{K^*} \hat{\gamma}_{i,[m,T+d]}(T+r)\theta^*_i \right) \right] \quad r = 1, \ldots, d \tag{107}$$

with $\hat{x}_s = x_s$ and $\hat{\Gamma}_{[m,T+d]}(s) = \Gamma^*_{[m,T]}(s)$ if $s \leq T$.

A self-contained strategy for predicting $\hat{\Gamma}_{[m,T+d]}$ has been recently proposed in [34]. It is based on two simple but fundamental observations. Firstly, $\Gamma^*_{[m,T]}(t),\ t = m, \ldots, T$ itself can be viewed as a time series of discrete probability distributions due to the imposed convexity conditions in (13). Secondly, under the assumption that the *distributions* $\Gamma^*_{[m,T]}(t),\ t = m, \ldots, T$ are associated with a (hidden) time-homogeneous and stationary Markov process, a model for the dynamics of the cluster affiliations is readily given by ($\Gamma \equiv \Gamma^*_{[m,T]}$)

$$\Gamma^\dagger(t+1) = \Gamma^\dagger(t) P, \tag{108}$$

where $P \in \mathbb{R}^{K^* \times K^*}$ is a stochastic matrix, i.e., $P$ is elementwise nonnegative and the entries of a row sum up to 1.

Particularly, the dynamics in (108) allows the recursive prediction of $\hat{\Gamma}_{[m,T+d]}$, e.g.,

$$\hat{\Gamma}^\dagger(T+1) = \hat{\Gamma}^\dagger(T) P \tag{109}$$

with $\hat{\Gamma}(T) = \Gamma^*_{[m,T]}(T)$ and, finally, in combination with (107) leads to a self-contained prediction scheme for the dynamics of the data under consideration.

This leaves us with the question how to estimate the stochastic matrix $P$ from the time series of affiliations. Of course, in general we can not expect that a matrix $P$ exists such that (108) exactly holds true. However, the FEM-BV methodology, in particular the approach presented in the Section 2.c.iv, provides an elegant way to deal with that situation by solving the following variational problem (cf. (42)):

$$\sum_{t=0}^{T-1} \left\| \Gamma^\dagger(t+1) - \Gamma^\dagger(t) P \right\|_2^2 \to \min_P, \tag{110}$$

subject to $P$ being a stochastic matrix and $\Gamma \equiv \Gamma^*_{[m,T]}$. In order to study the influences of external factors $u(t) \in \mathbb{R}^k$, we additionally assume that the matrix $P = P(u(t))$ can be decomposed as (cf. (44))

$$P(u(t)) = P_0 + \sum_{l=1}^{k} u_l(t) P_l, \tag{111}$$

where the involved matrices satisfy the constraints (46)–(48). The final minimization problem (110)–(111) with respect to the parameters $P_0, \ldots, P_k \in \mathbb{R}^{\boldsymbol{K}^* \times \boldsymbol{K}^*}$ and subject to the constraints (46)–(48) can be cast in a constrained quadratic program. For details see the Appendix.

Next, suppose that an observation for $x_{T+1}$ is available. What is the optimal prediction for $\hat{x}_{T+2}$ *conditioned* on the additional observation $x_{T+1}$? As motivated in [34], instead of reanalyzing the updated time series $(x_0, \ldots, x_{T+1})$ via the FEM-BV approach and reapplying the prediction scheme described above, it is sufficient to determine the optimal affiliation vector $\Gamma^*_{[m,T+1]}(T+1)$ simply by

$$\gamma_i^*(T+1) = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j \left\{ g(x_{T+1}, \ldots, x_{T+1-m}, \theta_j^*) \right\}, \\ 0 & \text{otherwise,} \end{cases} \tag{112}$$

which by virtue of (108) and (107) yields the prediction $\hat{x}_{T+2}$. The generalization of the conditional prediction in the presence of more than one new observation, say $(x_{T+1}, \ldots, x_{T+t'})$, is straightforward. The resulting scheme is $(r = 1, \ldots, d)$

$$\begin{aligned} (x_{T+t'}, \Theta^*) &\xrightarrow{\text{via (112)}} \Gamma^*_{[m,T+t']}(T+t') \\ &\xrightarrow{\text{via (108)}} \hat{\Gamma}_{[m,T+t'+d]}(T+t'+r) \xrightarrow{\text{via (107)}} \hat{x}_{T+t'+r}. \end{aligned} \tag{113}$$

The remainder of this section is devoted to describing numerical strategies to assess the prediction quality of the scheme given above. To this end, we compare $\hat{x}_{T+k}$ with standard prediction approaches such as the "zero" prediction model frequently used in, e.g., the meteorological literature. Formally, it reduces to

$$\hat{x}^0_{T+d} \equiv x_T. \tag{114}$$

Furthermore, as frequently pointed out in this manuscript, stationarity is a widely used and well accepted assumption in time series analysis. Thus, it is reasonable to compare $\hat{x}_{T+d}$ with the prediction $\hat{x}^1_{T+d}$ resulting from an optimal *stationary* substitute model, i.e. (analogously to (107))

$$\hat{x}^1_{T+d} = \mathbb{E}\left[ f\left( \hat{x}^1_{T+d}, \ldots, \hat{x}^1_{T+d-m}, \theta^* \right) \right], \tag{115}$$

where $\theta^*$ is derived[3] from the time series under consideration.

[3] Numerically, this simply amounts to fix $\boldsymbol{K} = 1$ in the course of the FEM-BV approach.

The *average relative prediction error* of the $d$-step prediction scheme for a prediction horizon $[T+1, T+T']$ is then measured by

$$\bar{e}_d(T') \stackrel{\text{def}}{=} \frac{1}{T'-d+1} \sum_{t'=T}^{T'-d} \frac{\left\| x_{t'+d} - \hat{x}_{t'+d} \right\|}{\| x_{t'+d} \|}, \tag{116}$$

where $\|\cdot\|$ denotes a desired norm. That error is compared with the average relative error $\bar{e}_d{}^0(T')$ associated with the zero-prediction scheme and $\bar{e}_d{}^1(T')$ resulting from predicting via the stationary substitute model. See Section 5.e for a numerical example illustrating the described prediction schemes. Another possibility for measuring the prediction error is given by the information-theoretical approaches to model error assessment developed at the working group of A. Majda (NYU); we refer the interested reader to, e.g., [55; 24] for more details on this matter.

## 5. Numerical examples

In this section we illustrate the presented FEM-BV methodology on various examples. In the first and second example we demonstrate the general feasibility of the proposed method and discuss its properties on a simple model with known properties. In the third example, a modified version of the FEM-BV-$k$-means for periodic angular data is developed and applied to analyze the conformational dynamics of a small biomolecule. The fourth example deals with a problem in computational biology and shows that the FEM-BV framework adapted for discrete data allows us to analyze gene-sequences under minimal a priori assumptions. The analysis of financial data is presented in the last example in which we also discuss the usefulness of the self-contained prediction scheme presented in Section 4.

**5.a.** ***Toy model system I: FEM-BV-k-means.*** The $k$-means approach is a widely used algorithm to cluster stationary data on the basis of geometric properties, i.e., the Euclidean distance to geometric centroids. However, even for low dimensional examples $k$-means fails to identify the "right" clusters. In the first numerical experiment we show for such a counter example that the additional information of the temporal (persistent) ordering of the data is sufficient to separate geometric clusters via of FEM-BV-$k$-means.

To this end we consider a time series of two dimensional data $x(t) = (x_1(t), x_2(t))$ generated via a mixture model consisting of a time dependent convex combination of three (stationary) normal distributions,

$$x_t \sim \sum_{i=1}^{3} \gamma_i(t) \mathcal{N}(\mu_i, \Sigma_i) \quad t = 1, \ldots, 6000, \tag{117}$$

where the weights (cluster affiliations) $\Gamma(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t))$ are deterministic

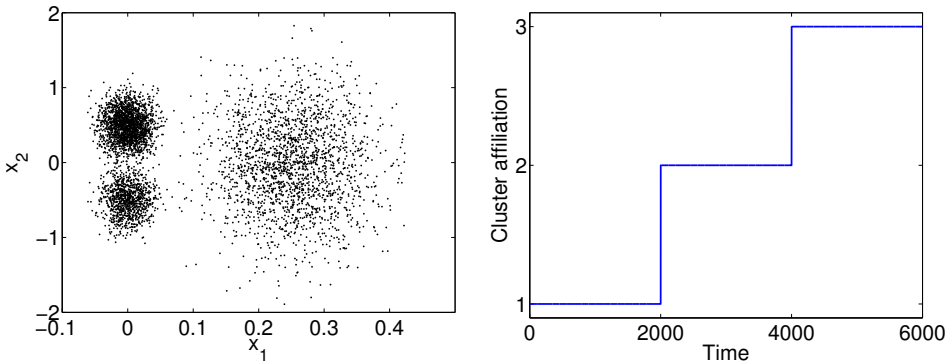


**Figure 2.** Toy model I. Left: scatter plot of the a time series generated via (117) where the parameters were chosen such that the data exhibits three geometric clusters. Right: the graph of the cluster affiliations used as a persistent hidden process in parameter space for the generation of the time series depicted in the left panel.

and prescribed. Particularly, $\Gamma(t)$ was chosen such that the (hidden) affiliation process jumps only once from cluster one to cluster two and finally to cluster three, i.e., $\|\gamma_1\|_{BV} = \|\gamma_3\|_{BV} = 1$ and $\|\gamma_2\|_{BV} = 2$. For an illustration of $\Gamma(t)$ see the right panel of Figure 2. As one can see in the scatter plot given in the left panel of Figure 2, the means and covariance matrices $(\mu_i, \Sigma_i)$, $i = 1, 2, 3$ were chosen such that a sufficiently long sample (here $T = 6000$) exhibits three geometrically nonoverlapping clusters. However, the $k$-means algorithm for $k = 3$ failed to identify these clusters as illustrated in the left panel of Figure 3. Notice that the misclassification of the data points is basically due to the different scales of the $x_1$ and $x_2$ components of the data.

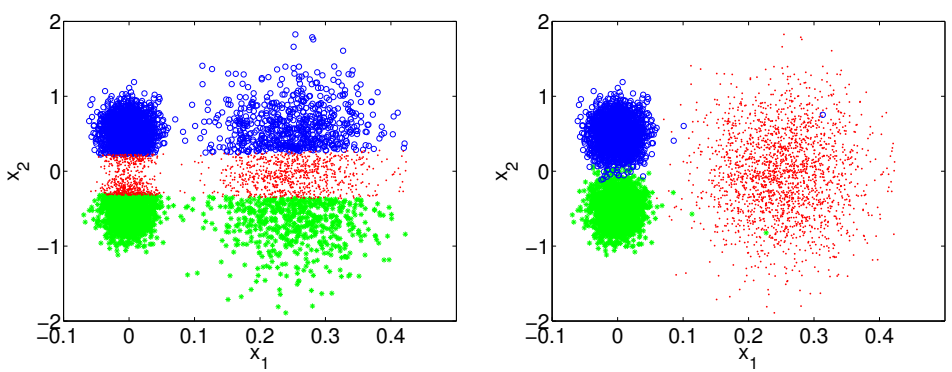


**Figure 3.** Toy model I. Cluster affiliations of the data points resulting from the classical $k$-means algorithm (left) and from the FEM-$k$-means method (right). Up to a few misclassifications, the latter method led to the right assignment of the data points to the original clusters, whereas the former one totally messed up.

Next, we analyzed the time series with the FEM-BV approach which results from the simple model in (6) and the model distance function in (9). Recall that for $\boldsymbol{C} = \infty$ the average cluster functional admits an analytic solution for $\Gamma$ and for the cluster parameter $\Theta$ which both coincide with the respective update formulas in the $k$-means algorithm (Section 2.a). Now the question is whether the persistence of the prescribed cluster affiliations is sufficient to identify the three cluster while using the same distance function as in the standard $k$-means approach?

To this end, we repeatedly launched the FEM-BV-$k$-means subspace algorithm (cf. Section 2.b and (72)–(75)) for all combinations of

$$\boldsymbol{K} = [2, 3, 4] \times \boldsymbol{C} = [2, 4, \ldots, 12],$$

each time with a randomly drawn initial $\Gamma$, until the global minimizer of the average cluster functional was found. For the respective optimal models we then computed the modified AIC values via the Maximum-Entropy approach presented in Section 3. For fixed $\boldsymbol{K}$ the graphs of $mAIC(\boldsymbol{K}, \boldsymbol{C})$ as a function of $\boldsymbol{C}$ are given in the left panel of Figure 4. The overall minimum is attained in $\boldsymbol{K}^* = 3, \boldsymbol{C}^* = 2$ which are exactly the parameters of the original data. In the right panel of Figure 4, we exemplarily illustrate the histogram of the residuals (6) of the right geometrical cluster together with the graph of the fitted ME PDF (102) of order 3 which was used to compute the modified AIC values. Finally, the correct (up to a few isolated misfits) assignments of data points to the clusters based on the affiliation vector $\Gamma(t)$ is given in the right panel of Figure 3.

This simple but instructive example demonstrates that neglecting temporal persistence in data may lead to misleading results even for toy examples. In contrast, besides yielding the correct partition of the data, the FEM-$k$-means-method

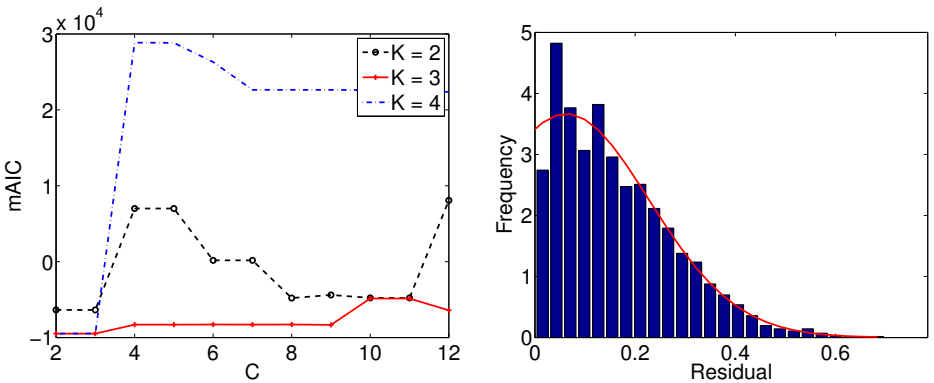


**Figure 4.** Toy model I. Left: graphs of the (modified) AIC values (97) for fixed $\boldsymbol{K}$ as a function of $\boldsymbol{C}$. Right: the histogram of the residuals (6) of the right geometrical cluster together with the graph of the fitted ME PDF (102) of order 3 (red line).

combined with the model selection approach allowed us to reidentify the correct parameters $\boldsymbol{K} = 3, \boldsymbol{C} = 2$.

**5.b.** ***Toy model system II: FEM-BV-PCA.*** In the first example, the geometrical clustering of the time series basically relied on the separability via centroids, i.e., mean values. In the second example we demonstrate that even geometric cluster with comparable means can be reidentified via the FEM-BV approach by additionally incorporating spectral properties of covariances, i.e., principal components.

To this end, we consider time series of two dimensional data $x(t) \in \mathbb{R}^2$ of length $T = 10000$ generated via

$$x_t \sim \gamma_1(t)\mathcal{N}_2(0, \Sigma_1) + \gamma_2(t)\mathcal{N}_2(0, \Sigma_2). \tag{118}$$

The prescribed weights (cluster affiliations) $\Gamma(t) = (\gamma_1(t), 1 - \gamma_1(t))$ are deterministic. For an illustration of $\gamma_1(t)$ see the right panel in Figure 5. The covariance matrices $\Sigma_1$ and $\Sigma_2$ are chosen as

$$\Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 0.25 \end{bmatrix}, \Sigma_2(\rho) = \begin{bmatrix} \cos\rho & \sin\rho \\ -\sin\rho & \cos\rho \end{bmatrix} \Sigma_1 \begin{bmatrix} \cos\rho & -\sin\rho \\ \sin\rho & \cos\rho \end{bmatrix}, \tag{119}$$

where $\Sigma_2$ results from rotating $\Sigma_1$ by an angle $\rho = 15$ degrees. The scatter plot of the time series generated via (118) is depicted in the left panel of Figure 5. As one can see, the two clusters are almost identical and, by construction, are centered around $(0, 0)$. Therefore, any $k$-means clustering approach would fail to recover the original temporal affiliation. The only chance to identify the (hidden) cluster though is to cluster with respect to the eigenvectors of the (hidden) covariance

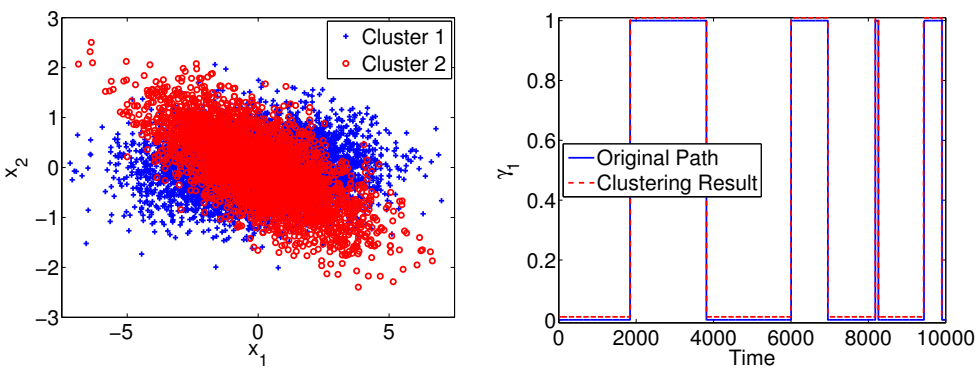


**Figure 5.** Toy model II. Left: scatter-plot of a time series generated via the mixture model in (118) consisting of a time dependent convex combination of two (stationary) normal distributions with mean zero and covariance matrices given in (119) and a rotation angle $\rho = 15$ degrees. Right: the prescribed affiliation function $\gamma_1(t)$ (solid line) completely coincides with one obtained from the FEM-BV-PCA-analysis (red dashed line).

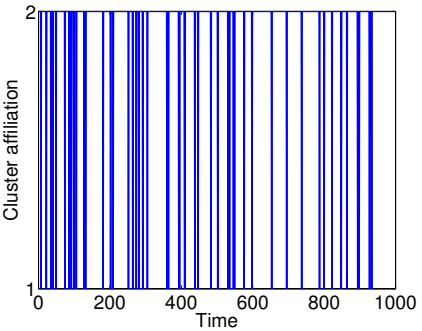


**Figure 6.** Toy model II: part of the Viterbi path ($1 \leq t \leq 1000$) obtained from fitting a two-dimensional stationary mixture model of two Gaussian distributions (via the GMM-method) on the data shown in Figure 5.

matrices. But this is exactly the idea of the FEM-BV-PCA approach which will be used here.

Before we present the results of the FEM-BV-PCA approach, we first apply the GMM-method which is a classical and widely accepted method for unsupervised clustering. We fitted (trained) a two-dimensional stationary mixture model of two Gaussian distributions on the data via the Expectation-Maximization algorithm [12]. Since Gaussians are involved in the time series generation, it is reasonable to expect the GMM-method to be able to reidentify the parameters of the hidden distributions. However, the estimated covariance matrices $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ significantly differ from the original ones, indicated by, e.g., $\|\Sigma_2 - \tilde{\Sigma}_2\| = 5.2040$.

The associated Viterbi path (partially depicted in Figure 6) reveals the reason for the failure; it is highly oscillatory rather than being persistent. Consequently, the majority of data points are incorrectly affiliated with regard to the original clusters which, ultimately, leads to the incorrect estimation of the covariance matrices. The irregularity of the Viterbi path, in turn, is a direct consequence of the strong stationary assumption underlying the GMM-method, i.e., time-independent distribution parameters and time-independent affiliation weights.

In contrast, as will be demonstrated in the following, the FEM-BV-PCA-method (see Section 2.c.ii) succeeded as it takes the persistence of the hidden dynamics in the parameter space into account. Analogously to the procedure described in the previous example in Section 5.a, we globally minimized the average cluster functional resulting from the model distance function in (27) via the subspace algorithm for all combinations of $\boldsymbol{K} \in \{1, 2, 3\}$ and $\boldsymbol{C} \in \{2, 4, 6, 8, 10, 14, 20\}$. The minimum of the corresponding modified AIC values is attained for $\boldsymbol{K}^* = 2$ and $\boldsymbol{C}^* = 8$, which are exactly the parameters used for the time series generation. Even more importantly, the numerically obtained affiliation vector is identical with the original one (see right panel of Figure 5).

**5.c.** ***Conformation analysis of a biomolecule (trialanine): FEM-BV-k-means.*** The biological function of a biomolecule is strongly characterized by its ability to assume almost constant geometrical configurations, referred to as *conformations*. More precisely, a conformation denotes a mean geometrical configuration of a molecule which is almost stable (metastable, persistent), i.e., the molecule's geometry wiggles around that configuration for a long period of time before it rapidly switches to another conformation. For example, it is known that conformations of certain proteins are responsible for severe human diseases [51]. For details on the analysis of the conformational dynamics of molecules we refer the interested reader to, e.g., [69] and the references therein.

It is common to analyze the conformational dynamics of a (bio-)molecule in internal coordinates such as torsion angles rather than to consider the time series of cartesian coordinates of all atomic positions. The reason is that torsion angles are invariant with respect to translation and rotation of the molecule and, more importantly, tremendously reduce the dimensionality of the time series. However, the (nonlinear) projection of the cartesian coordinates on the torsion angle space deflects the original dynamics and can lead to an incomplete picture of the conformational dynamics of the molecule. This is in particular true if only a subset of torsion angles is considered because of, e.g., numerical or statistical reasons. Consequently, conformations which are geometrically distinguishable in the complete torsion angle space might (completely) overlap in the reduced space. Thus, the identification of conformations via geometrical clustering of *incomplete* observations of torsion angles is an *ill-posed* problem.

In the traditional *transfer operator (TO) approach* [65] to conformational dynamics the problem is addressed by assuming that the underlying dynamics in the incomplete torsion angle space is a *reversible, stationary and time-homogeneous Markov process*. Alternatively, we propose to tackle the ill-posedness by regularization of the underlying persistent (metastable) dynamics in the BV sense and to identify conformations via a modified FEM-BV-$k$-means approach.

To this end, we consider in this example a time series of three torsion angles $\Phi$, $\Psi$ and $\Omega$ obtained from a molecular simulation of the trialanine molecule schematically illustrated as a ball-stick representation in the left panel of Figure 7. The simulation was performed in vacuum at constant temperature and pressure such that the resulting time series can be considered stationary for a sufficiently long simulation time $T$. The details of the simulation procedure can be found in [60]. As one can see in the right panel of Figure 7, the dynamics of the torsion angles exhibits a strong persistence or metastability.

Recalling that the torsion angles are periodic on $[-\pi, \pi]$, the 3d-scatter plot in the left panel of Figure 8 clearly reveals five geometrical clusters indicating five conformations. The projection on the two torsion angles $\Phi$ and $\Psi$, however,

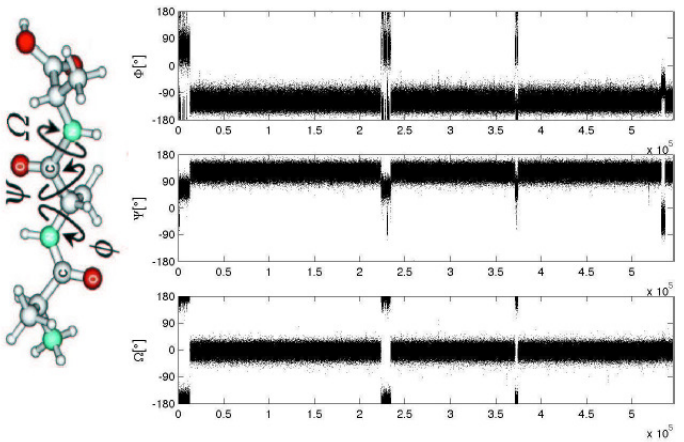


**Figure 7.** Biomolecule: molecular simulation of the trialanine molecule (left) reveals its conformational dynamics observed in the time series of three torsion angles (right).

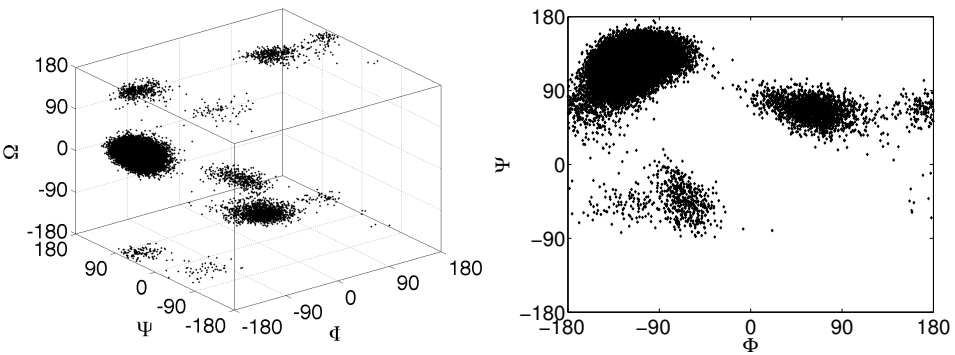


**Figure 8.** Biomolecule: recalling the periodic nature of torsion angles, the scatter-plot of the full time series (left) reveals five conformational clusters whereas the scatter-plot of the projected time series $(x_t) = (\Phi_t, \Psi_t)$ (right) suggests the existence of only three conformations.

suggests the existence of only three conformations as illustrated in the right panel of Figure 8. Consequently, the five clusters can only be recovered in the projection by additionally capturing the inherent persistence of the dynamics. This will be demonstrated in the remainder of this example.

To understand the following preprocessing steps, we briefly recall the transfer operator approach to conformation dynamics. The basic idea is to represent the dynamics underlying the time series of torsion angles as a *reversible, stationary and time-homogeneous Markov chain* defined on a suitable discretization of the torsion angle space, e.g., by boxes. The spectrum of the associated transition matrix $P$, then allows the characterization and extraction of the conformations as metastable subsets via, e.g., the robust Perron-cluster cluster analysis (PCCA+) [14]. To be

more precise, let $\lambda_1, \lambda_1, \ldots, \lambda_n$ be the first $n$ dominant eigenvalues of $P$, i.e.,

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|. \tag{120}$$

If a *spectral gap* exists, i.e., if one can find an index $\boldsymbol{K}$ such that $|\lambda_{\boldsymbol{K}}| \gg |\lambda_{\boldsymbol{K}+1}|$, then one can prove that the discrete state space can be decomposed into $\boldsymbol{K}$ metastable subsets (conformations), say $A_1, \ldots, A_{\boldsymbol{K}}$, based on the corresponding dominant eigenvectors [11; 66; 42; 13]. A measure for the total metastability of the resulting decomposition is then given by

$$\eta(A_1, \ldots, A_{\boldsymbol{K}}) = \sum_{i=1}^{\boldsymbol{K}} P(A_i, A_i), \tag{121}$$

where

$$P(A_i, A_i) = \mathbb{P}[x_1 \in A_i | x_0 \in A_i]$$

is the (time-homogeneous) probability that the dynamics is in $A_i$ after making a transition out of $A_i$.

Accordingly, to ensure Markovianity while preserving the persistence, the original time series was further subsampled by picking every 10-th time step resulting in a time series $(x_t) = (\Phi_t, \Psi_t)$ of total length $T = 54455$. Then, the 2-dimensional space spanned by the torsion angles $\Phi$ and $\Psi$ was discretized into $30 \times 30$ equidistantly sized boxes and we ended up with a 372-state Markov chain since only 372 boxes are visited by $x_t$.

The spectral gap between the third and fourth dominant eigenvalue (see left panel in Figure 9) suggests an optimal decomposition into three clusters showing

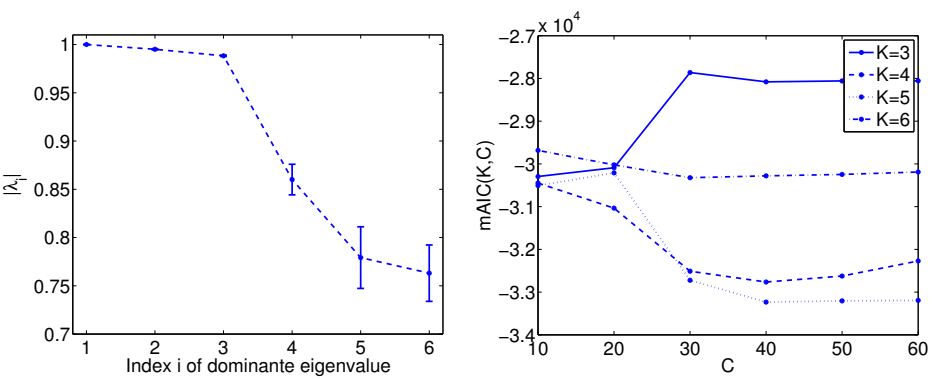


**Figure 9.** Biomolecule. Left: six dominant eigenvalues of the transition matrix $P \in \mathbb{R}^{372\times372}$ and their confidence intervals, resulting from a $30 \times 30$ box discretization of the state space spanned by $\Phi$ and $\Psi$. Right: for fixed $\boldsymbol{K} = 3, 4, 5, 6$ the graphs of the $m$AIC values as a function of $\boldsymbol{C}$ obtained via the Maximum-Entropy approach with order three. The minimum is attained for $\boldsymbol{K}^* = 5$ and $\boldsymbol{C}^* = 40$.

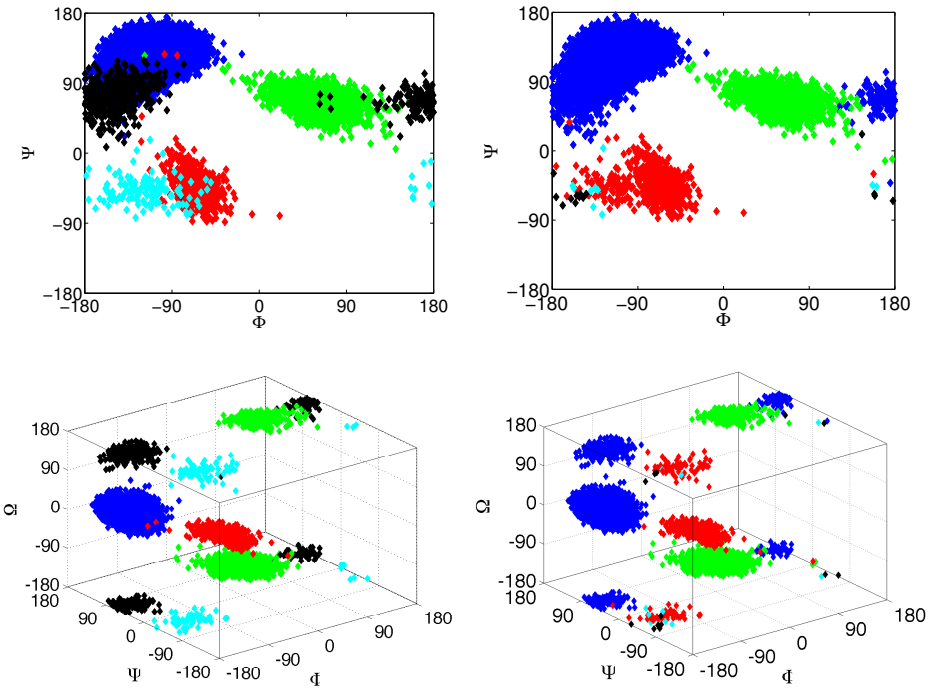


**Figure 10.** Biomolecule. Decomposition of the time series $(x_t) = (\Phi_t, \Psi_t)$ into five clusters via periodic FEM-BV-$k$-means (upper left) and the TO approach (upper right). The same decomposition of the time series visualized in a full three-dimensional feature space $(x_t) = (\Phi_t, \Psi_t, \Omega_t)$ reveals the correct identification of the conformations (lower left) by the FEM-BV-$k$-means method whereas the TO approach is not able to recover them from the incomplete observation in $x_t$ (lower right).

that the TO approach fails to capture the persistence of the dynamics leading to five conformations. Furthermore, as illustrated by the error bars, the high uncertainty[4] of the fifth and sixth dominant eigenvalue indicates that they are statistically indistinguishable and so are the corresponding eigenvectors. Hence, any attempt to decompose the state space into five clusters by additionally considering the fourth and, particularly, the fifth dominant eigenvector would fail to properly separate the conformations. This is confirmed in the right lower panel of Figure 10 and by the fact that the total metastability (121) for the decomposition resulting from the TO approach has the value $\eta_{\mathrm{TO}} = 4.106$, significantly lower than the value $\eta_{\mathrm{FEM}} = 4.900$ resulting from the periodic FEM-BV-$k$-means method to be presented below.

[4]Based on a 800,000-member transition matrix ensemble generated via a sampling method introduced in [57].

From a more general viewpoint, the high uncertainty in the (less) dominant eigenvalues reflects the ill-posedness of the cluster problem in the presence of incomplete data. Hence, an appropriate regularization is needed such as provided in the variational FEM-BV approach.

As demonstrated in Section 5.a, the simplest way to geometrical clustering while taking persistence into account is the FEM-BV-$k$-means approach. The model distance function in (9), however, does not capture the *periodic* nature of the data. Fortunately, this can easily be fixed by adopting a distance model function defined on the $d$-dimensional torus:

$$g(x_t, \Theta_t) = \sum_{j=1}^{d} \left\| \omega([x_t]_j) - \omega([\Theta_t]_j) \right\|_2^2, \quad \text{with } \omega(\alpha) = (\cos\alpha, \sin\alpha) \in \mathbb{R}^2, \tag{122}$$

where $[y]_j$ denotes the $j$-th component of $y \in \mathbb{R}^d$. A straightforward calculation shows that the average cluster functional associated with (122) attains for given $\Gamma(t)$ a local minimum in $\theta_i^* \in \mathbb{R}^d$, elementwise given by

$$[\theta_i^*]_j = \tan^{-1} \frac{\sum_{t=0}^{T} \gamma_i(t) \sin[x_t]_j}{\sum_{t=0}^{T} \gamma_i(t) \cos[x_t]_j} \quad j = 1, \ldots, d. \tag{123}$$

Via the subspace algorithm, we globally minimized the average cluster functional resulting for all combinations of $\boldsymbol{K} \in \{3, 4, 5, 6\}$ and $\boldsymbol{C} \in \{10, 20, \ldots, 60\}$. The $m$AIC values are plotted in the right panel of Figure 9. The overall minimum is assumed in $\boldsymbol{K}^* = 5$ and $\boldsymbol{C}^* = 40$ suggesting the existence of five conformations. Indeed, the according decomposition of the full time series (left lower panel of Figure 10) based on the 2-dimensional clustering (left upper panel of Figure 10) shows that the FEM-BV approach succeeded in identifying the conformations most correctly.

In this example we have demonstrated that the FEM-BV-$k$-means approach adapted for periodic data allows us to identify all of the relevant conformations of a biomolecule based on incomplete torsion angle observations. In particular, we have shown that the combination of BV-regularization with the model selection via the modified AIC does not only yield the correct number but also the correct assignment of the analyzed data to proper conformations. In contrast, although the underlying assumptions necessary for formal applicability of the TO approach (e.g., homogeneity and Markovianity) are formally fulfilled for the analyzed time series, it was demonstrated that the classical transfer operator approach can suffer from the ill-posedness of the clustering problem resulting from the strong overlapping of different conformational states in the reduced representations. The current example

demonstrates that this ill-posedness can result in misleading conformational decompositions in context of the TO approach.

**5.d.** ***Yeast DNA.*** One of the major challenges in bioinformatics is the identification of genes from biological data. In this example, we approach that problem with the FEM-BV-categorical method derived in Section 2.c.iii and compare the results to classical methods such as the unsupervised HMM.

Gene finding is the identification of coding (*exons* or *genes*) and noncoding (*introns*) regions in nucleic acids (DNA and RNA) based on sequences of codons which specify the amino acid production during the protein synthesis. A codon is a sequence of three nucleotides out of the four possible nucleic bases adenine (A), guanine (G), thymine (T) and cytosine (C). Thus, a single codon can code for a maximum of 64 amino acids.

Traditional approaches to gene finding are based on *supervised machine learning* methods such as supervised HMMs [49], and rely on extensive previous training and a high amount of a priori biological knowledge. Particularly, it is assumed that the hidden process switches exactly between two states, coding and noncoding regions, and that it is a stationary Markov process.

In contrast to the supervised methods, we propose the FEM-BV approach based on the categorical model introduced in Section 2.c.iii as an *unsupervised* approach. We exemplify the usefulness of the method by clustering a sequence $c_t$ of $T = 10'000$ codons resulting from the first $30'000$ nucleotides of the first chromosome of *Saccharomyces cerevisiae*, the ordinary yeast. The data is publicly available at [59]. Notice that in the variational approach the assumption of *persistence* corresponds to the biological assumption that coding and noncoding regions are each *connected*.

After identifying each codon $c_t$ with a discrete state $s_t \in \mathcal{S} = \{1, \ldots, 64\}$ we globally minimized the average cluster functional in (35) (resulting from the model distance function in (34)) for all combinations of $\boldsymbol{K} \in \{1, \ldots, 3\}$ and $\boldsymbol{C} \in \{4, \ldots, 10, 15, 20, 30, 40, 50\}$. Unlike to the previous examples where we applied the Maximum-Entropy approach, here we computed the likelihood function $\mathcal{L}(\boldsymbol{K}, \boldsymbol{C})$, involved in the modified AIC value (97), by exploiting that the stationary cluster parameters $\theta_1, \ldots, \theta_{\boldsymbol{K}}$ are probability distributions. Consequently, $\mathcal{L}(\boldsymbol{K}, \boldsymbol{C})$ takes the form,

$$\mathcal{L}(\boldsymbol{K}, \boldsymbol{C}) = \prod_{t=1}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t)\theta_i(s_t). \tag{124}$$

As one can see in Figure 11, the optimal substitute model is attained for $\boldsymbol{K}^* = 2$ and $\boldsymbol{C}^* = 6$. The interpretation of the two clusters as a *coding and a noncoding* model is substantiated by comparing the associated affiliation function $\gamma_1(t)$ with known positions of the genes in this part of the DNA sequence. As one can see

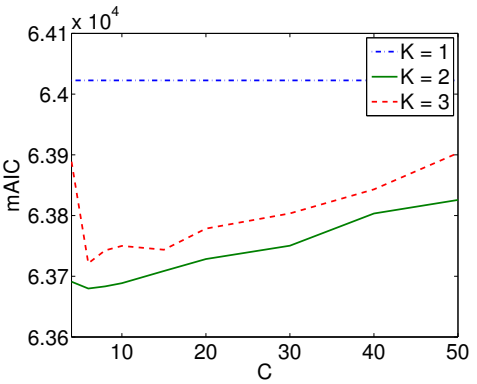


**Figure 11.** Yeast DNA. The minimal $m$AIC value is assumed for $\boldsymbol{K}^* = 2$ and $\boldsymbol{C}^* = 6$ which, particularly, is consistent with the biological fact that codons can be divided into coding and noncoding regions.

in the left panel of Figure 12, the affiliation path $\gamma_1(t)$ of the first cluster separates mostly correctly between genes and noncoding regions. Only the gene *SEO1* is not identified which is in contrast to its graphical appearance and length in the left panel of Figure 12. This conflict, however, can be resolved by the experimental fact that this particular region encodes a protein but it does not exhibit a persistent sequence of coding codons because it is highly fragmented, for details see [59]. This violates the persistence assumption inherent to the FEM-BV methodology.

From considering the highly oscillatory Viterbi path of an unsupervised two-state HMM fitting (see right panel of Figure 12) one sees that the assumption of stationarity impedes the traditional approaches to identify genes correctly unless a large amount of biological knowledge is incorporated via supervised learning

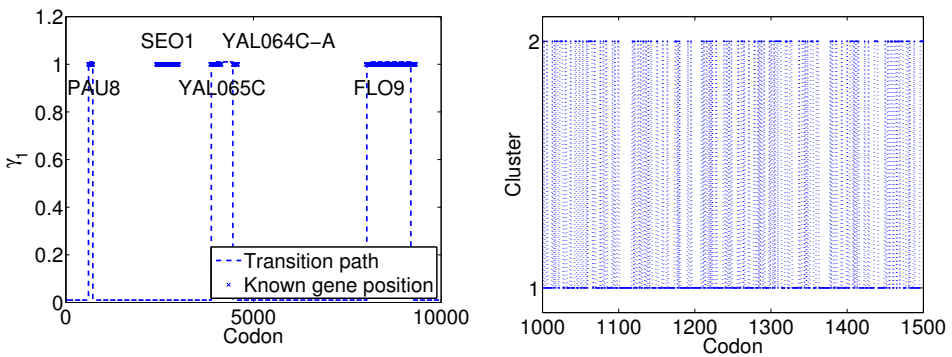


**Figure 12.** Yeast DNA. Left: a comparison of $\gamma_1(t)$ with the positions of some genes justifies the interpretation of the two optimal clusters as coding and noncoding models. Right: the first part of a Viterbi path of an unsupervised two-state HMM fitting reveals that traditional methods assuming stationarity fail to capture the inherent persistence in codon sequences.

strategies. In contrast, respecting the inherent persistence in the sequence of the codons via the variational FEM-BV approach allowed us to identify most of the known gene positions. Even more important, the detection of coding and noncoding regions, i.e., $\boldsymbol{K}^* = 2$, was part of the result and not a priori included knowledge.

**5.e.** ***Financial data for commodities.*** In the final example, a time series of daily closing prices of futures on oil is analyzed in order to address two important question: Does the FEM-BV approach allow us to identify market phases (e.g., economic crises) and how do external factors affect the evolution of financial data. In the remainder of the example, we apply the prediction scheme introduced in Section 4 and compare its prediction skills with those of simple prediction methods.

In 1989 J. Hamilton [27] introduced a numerical method to identify what he called hidden market phases in financial data which can be seen as the first combination of nonstationary time series analysis and mathematical finance. Since then the method has been generalized and extended to multidimensional data. Prominent phase-identification techniques are based, e.g., on linear vector autoregressive (VAR) models [48], wavelets [2], Kalman filters [45], (G)ARCH [15; 7] or perfect knowledge about the hidden process [10]. These methods, however, suffer from infeasible numerical complexity in high dimensions (curse of dimensions) or are based on strong model assumptions on the underlying dynamics, e.g., stationarity or Markovianity.

The time series $(x_t)$ under consideration here consists of daily closing prices of futures on the commodity oil for the time horizon 2005–2009 [73]. Futures are very sensitive to changes in market phases because they are broadly traded on speculative reasons. The graph of prices is illustrated in Figure 13. Despite the noisy fluctuations of the daily prices, one can clearly see two tendencies or market phases.

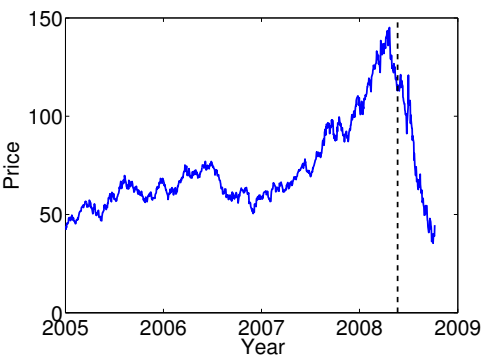


**Figure 13.** Commodities. Price of oil futures for the timeframe 2005 to 2009. The first 90% of the time series (indicated by the horizontal dashed line) is used as a training set for computing the optimal substitute model. The prediction skill of the nonstationary prediction scheme derived in Section 4 is then assessed on the remaining 10%.

Recall that we are interested in detecting market phases and, more importantly, how their dynamics are affected by external factors. As explained in Section 2.c.iv, the FEM-BV-Markov lends itself well to answer the questions since it allows us to incorporate external factors, specifically.

To this end, the time series of daily prices $(x_t)$ is coarse grained by assigning $(x_t)$ to one of the following categories: (i) The price increased significantly, (ii) no major movement was detected or (iii) the price dropped by a significant amount. Formally, we label the continuous prices by

$$s_t \overset{\text{def}}{=} \begin{cases} 1 & \text{if } x_t - x_{t-1} > \xi, \\ -1 & \text{if } x_t - x_{t-1} < -\xi, \\ 0 & \text{otherwise,} \end{cases} \tag{125}$$

where the threshold $\xi$ separates noise from significant changes and was set to the standard deviation of the time series. This data preparation approach is similar to the one introduced in [27] to detect changes in the Markovian market dynamics.

The transformed time series $(s_t),\ t = 0, \ldots, T$ now takes values in the discrete state space $\mathcal{S} = \{-1, 0, 1\}$. Analogously to the proceeding in Section 2.c.iii, we represent a state $s_t$ by a Dirac-distribution $\pi_t$ which is defined for the discrete states $s = -1, 0, 1$ as (cf. (30))

$$\pi_t(s) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } s = s_t, \\ 0 & \text{otherwise.} \end{cases} \tag{126}$$

The resulting time series $(\pi_t),\ t = 0, \ldots, T$ encoding the inherent tendencies of the price evolution in terms of probability distributions can now be analyzed by the Markov regression model in (39). More importantly, the FEM-BV-Markov approach allows us to investigate the influence of external factors. Specifically, we would like to understand to which extent the price evolution is influenced by the overall state of the US economy and the climate situation, especially, by the effects of El Niño and La Niña [72].

To this end, the following external factors are considered:

$u_1$ the daily closing value of the Dow Jones Industrial Average (available at [75])

$u_2$ the El Niño-Southern Oscillation (ENSO) index 3.4 (available at [47]).

To test on memory effects, three additional external factors are taken into account:

$u_3$ the Dow Jones shifted (delayed) by one day,

$u_4$ the ENSO index delayed by 30 days,

$u_5$ and the ENSO index delayed by 60 days.

Finally, the external factors are scaled to the interval [0, 1] to ensure comparability of the influences as the Dow Jones takes values around 10, 000 while the ENSO takes values between $\pm 1.5$.

Besides the analysis of the data, the main goal of this example is to demonstrate the skills of the prediction scheme presented in Section 4. Therefore, the time series $(\pi_t)$ is divided into a training set, containing the first 90% of the data and a prediction set, consisting of the remaining data. The analysis via FEM-BV-Markov is based only on the training set, simulating the lack of knowledge about the future, so that the prediction can then be compared to the prediction set.

Next, we describe in detail the clustering of the training set via the FEM-BV-Markov approach and the subsequent optimal model selection. We globally minimized the average cluster functional resulting from the model distance function in (42) for all combinations of $\boldsymbol{K} \in \{1, \ldots, 4\}$ and $\boldsymbol{C} \in \{3, \ldots, 10\}$ and all $2^5$ possible subsets of combinations of external factors (ranging from no external factor to all five factors).

Analogously to the proceeding in the previous example in Section 5.d, we exploit the fact that the average mixture model associated with the FEM-BV-Markov approach (cf. Section 2.g and (106)),

$$\hat{\pi}_{t+1}^{\dagger} = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \pi_t^{\dagger} P^{(i)}(u(t)), \tag{127}$$

preserves probability, i.e., $\hat{\pi}_{t+1}$ is again a probability distribution. Consequently, the likelihood function $\mathscr{L}(\boldsymbol{K}, \boldsymbol{C})$ (involved in the modified AIC value (97)), here can be computed via (127) by

$$\mathscr{L}(\boldsymbol{K}, \boldsymbol{C}) = \prod_{t=0}^{T-1} \mathbb{P}_{\hat{\pi}_{t+1}}[s_{t+1}] = \prod_{t=0}^{T-1} \hat{\pi}_{t+1}(s_{t+1}). \tag{128}$$

The overall minimum of the modified AIC value with respect to all combinations of clusters' numbers, persistence values and all combination of external factors is attained for $\boldsymbol{K}^* = 2$, $\boldsymbol{C}^* = 6$ and without any external factors. That outcome is consistent with the weak efficient-market hypothesis in [16], stating that any information publicly available is instantly included in the price. The associated affiliation vector (depicted in Figure 14) more or less separates the time horizon of the training data set into two persistent regions. Interestingly, the time point of change at the end of 2008 from cluster 1 to cluster 2 is very close to the beginning of the financial crisis of the late 2000s.

The interpretation of $\Gamma^*(t)$ as an indicator of market phases is further substantiated by looking at the constant transition matrices associated with the two clusters

$$P_0^{(*1)} = \begin{bmatrix} 0.0448 & \mathbf{0.8955} & 0.0597 \\ 0.0989 & \mathbf{0.8112} & 0.0899 \\ 0.1167 & \mathbf{0.8000} & 0.0833 \end{bmatrix}, \quad P_0^{(*2)} = \begin{bmatrix} 0.2453 & 0.4528 & 0.3019 \\ 0.4030 & 0.3433 & 0.2537 \\ 0.3333 & 0.3778 & 0.2889 \end{bmatrix}. \tag{129}$$

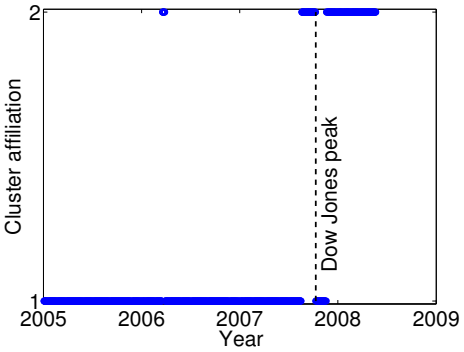


**Figure 14.** Commodities. The cluster affiliation $\gamma_1^*(t)$ associated with the optimal substitute model with $\boldsymbol{K}^* = 2$, $\boldsymbol{C}^* = 6$ and no external factors for the training set (first 90% of the data). The majority of the second cluster is located from the end of 2007 onwards, indicating a relation to the financial crisis.

Recalling that an entry $P_{ij},\ i, j \in \{-1, 0, 1\}$ of stochastic matrix $P$ with respect to to $\mathcal{S}$ denotes the conditional probability that the associated Markov chain jumps from state $i$ to state $j$, the second column in $P_0^{(*1)}$ indicates that the noise state $s = 0$ is metastable. In other words, cluster (market phase) $i = 1$ is characterized by small movements without any specific tendencies. In contrast, the transition matrix $P_0^{(*2)}$ of the second cluster does not show any dominating state as the transition probabilities are close to each other, thus, indicating no specific direction in price movement. Additionally, the second column suggests that the average change in price is increased compared to the first cluster. Both observations together imply an increase of the variance in the price evolution which is consistent with the observations in [6; 15] stating that economic crises are characterized by high variance whereas low-variance phases correspond to the normal state of the market.

The analysis was performed for different ending times of the training set, though a relevant influence of the external factors could not be observed. However, if the training set does not include the peak in the price, the analysis yields in selecting the stationary ($\boldsymbol{K}^* = 1$) model. This is to be expected, as the second cluster, representing the "crisis state", has insufficient size to be statistically relevant.

The remainder of this section is devoted to the prediction scheme introduced in Section 4. Rather than predicting the price evolution, we adapt the scheme for predicting the probability distributions $\hat{\pi}_t$ with respect to the discrete state space $\mathcal{S}$ for $t \geq T + 1$.

The fitting scheme associated with the optimal model ($\boldsymbol{K}^* = 2$, $\boldsymbol{C}^* = 6$ and without any external factors) reduces to

$$\hat{\pi}_{t+1}^{\dagger} = \sum_{i=1}^{2} \gamma_i^*(t) \pi_t^{\dagger} P_0^{(i)} \quad t = 0, \ldots, T, \tag{130}$$

where $P_0^{(*1)}$ and $P_0^{(*2)}$ are given in (129). In order to extend (130) to $t \geq T+1$, we estimated a stationary Markov regression model $P^*(u(t))$ based for the time series $(\Gamma^*(t))$ of optimal affiliation vectors. Consistently with the analysis of $(\pi_t)$, we thereby considered all combinations of external factors. It turned out that the optimal stationary Markov regression model is independent of any external factors too. Formally, we have $P^*(u(t)) = P^*$ and the prediction scheme for $\hat{\Gamma}(t)$ takes the form

$$\hat{\Gamma}^{\dagger}_{[0,T+d]}(T+r) = \left(\Gamma^*_{[0,T]}\right)^{\dagger}(T)\left[P^*\right]^r, \quad r = 1, \ldots, d. \tag{131}$$

Combining (131) with (127) defines a self-contained nonstationary online prediction scheme analogously to the scheme given in (113). We compare our scheme with standard prediction schemes based on:

(1) An independent stationary model formally given by

$$\hat{\pi}^0_{t+1} = \mu, \quad \mu(s) \stackrel{\text{def}}{=} \frac{1}{T+1}\sum_{t=0}^{T} \chi_s(s_t), \ s \in \{-1, 0, 1\}. \tag{132}$$

(2) A stationary Markov regression model estimated from the time series $(\pi_t)$ (without any external factors),

$$\left(\hat{\pi}^1_{t+1}\right)^{\dagger} = \left(\hat{\pi}^1_t\right)^{\dagger} P. \tag{133}$$

(3) A zero-prediction model, where the prediction is the last known state

$$\hat{\pi}^{(2)}_{t+1} = \hat{\pi}^{(2)}_t. \tag{134}$$

(4) An artificial neural network, as used in pattern recognition (see [5], for instance), using the external factors as input variables. For the test we have chosen the optimal network configuration (i.e., number of hidden neurons, transfer functions, etc.) with respect to prediction quality.

The average relative prediction error of the five $d$-step prediction schemes for a prediction horizon $[T+1, T+T']$ on the prediction set is measured similar to (116):

$$\bar{e}_d(T') \stackrel{\text{def}}{=} \frac{1}{T'-d+1}\sum_{t'=T}^{T'-d} \frac{\|x_{t'+d} - \hat{x}_{t'+d}\|_2}{\sqrt{2}}, \tag{135}$$

where the additional factor $\frac{1}{\sqrt{2}}$ is introduced to normalize the error for the worst prediction (the prediction of one state having probability one that is not fitting the realization) to 1. As one can see in Figure 15 (left panel), our nonstationary scheme outperforms the standard schemes. The highly oscillatory behavior of the zero-prediction comes from the fact, that the error of a single prediction is either

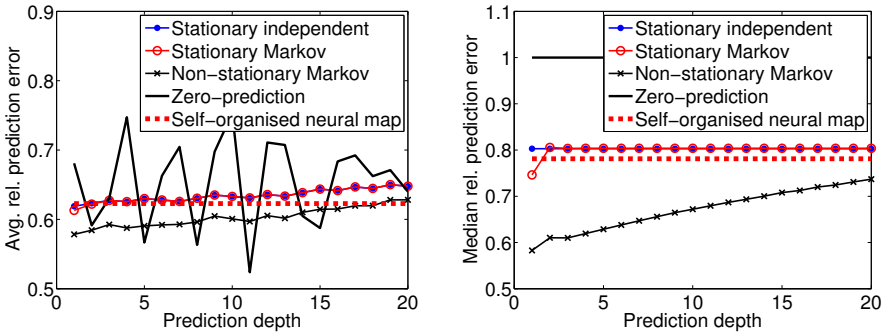


**Figure 15.** Commodities. The five prediction schemes (see page 220) are compared. Shown are the graphs of the average relative prediction error $\bar{e}_d(T')$ (left) and the median of the relative prediction error (right) as functions of the prediction depth $d$. They clearly reveal that the nonstationary prediction strategy outperforms the standard schemes.

0 or 1, thus maximizing the variance of the prediction and the small sample size of the predicted time frame. More precisely, when using the median (or 50%-quantile) of the error instead of the average, shown in Figure 15 (right panel), the zero-prediction is more likely wrong than right.

To sum up, we can now answer the questions from the beginning of this section: does the FEM-BV approach allow us to identify market phases (e.g., economic crises) and how do external factors affect the evolution of financial data? First, the FEM-BV-Markov model does not only allow the identification of market phases, but also results in a more accurate model of the market that can be used to predict further movements. Second, in line with the (weak) efficient market hypothesis, the influence of the general US economy, represented by the Dow Jones, and the El Niño/La Niña events were shown the be insignificant with respect to the analyzed data. However, we are aware of the fact, that this might be a result of the insufficient length of the analyzed time series and not a general fact.

We also want to emphasize, that we performed a qualitative analysis instead of a quantitative, as we coarse grained the data to overcome noise effects. While the results might be of no great practical use for investment strategists, it can be considered relevant for risk management, as we were able to verify the fact, that financial unstable market situations yield in a higher volatility. This is a nonnegligible part of most definitions of financial and economical risk.

## 6. Conclusion

A variational approach to nonstationary time series analysis developed in the last years is presented as a unified framework for analysis, discrimination and prediction of various types of observed processes. It was demonstrated, that persistence

is one of the main characteristic features of many real life processes and that an appropriate mathematical regularization strategy is the clue to its efficient recovery from the observation data. Moreover, a unified model discrimination approach is suggested based on a modified formulation of the information theoretic criterion AIC. Furthermore, the paper contains a first systematic comparison of the FEM-BV methodology with standard time series analysis methods and their underlying mathematical assumptions. The framework is demonstrated on various examples ranging from simple toy models to the analysis of real-world processes such as biomolecular dynamics, DNA-sequence analysis and financial risk prediction.

The effect of nonstationarity is captured in the FEM-BV approach by identifying a (hidden) process in parameter space describing transitions between different regimes which are characterized by local models and their stationary parameters. The presented clustering scheme involves several numerical optimization techniques combining elements of convex optimization, linear programming and Finite Element methods. This allows the employment of fast and numerically robust solvers which ensure an efficient analysis of high dimensional time series. As demonstrated in the present paper, the variational framework is very flexible with respect to different (non-)dynamical scenarios because only the estimators for the optimal parameters have to be provided either analytically or numerically whereas the estimation of the transition process remains general. Therefore, the FEM-BV approach can be straightforwardly adapted and redesigned to new model functions and new applications.

In contrast to classical methods such as HMM, GMM, neuronal networks or local kernel methods, the approach presented here does not rely on a priori probabilistic assumptions (e.g., stationarity, independence, Gaussianity, Markovianity, etc.). Instead of the probabilistic assumptions made in standard statistical methods, here firstly it is assumed that the dynamics under consideration are persistent, i.e., the parameters of the process vary much more slowly than the process itself. Secondly, it is assumed that the hidden process in parameter space can be described by a function with bounded variation. The latter assumption leads to a direct control of the regularity of the hidden process within the course of optimization and, thus, allows us to explicitly incorporate persistence or metastability. For the nonregularized case, it was demonstrated that standard methods such as $k$-means or (time-dependent) probabilistic mixture models are recovered by the FEM-BV approach as special cases. Although these assumptions are quite general, it is important to emphasize that their fulfillment is crucial for postprocessing and interpretation of the obtained results.

Another aim of this paper was to present a novel self-contained model selection strategy to simultaneously identify the optimal number of clusters and the optimal regularity of the paths in parameter space. As demonstrated in the numerical

examples, the clusterwise approximation of the scalar residuals via maximum entropy distributions in conjunction with the subsequent evaluation of the modified Akaike information criterion successfully allows us to identify the essential nonstationary patterns in various time series. The maximum entropy ansatz follows the philosophy of the FEM-BV approach in that it requires as less as possible explicit a priori knowledge. The central mathematical assumption underlying this strategy is, however, that the scalar residuals are independent. Hence, further research has to be done to generalize this setting in order to cover the case of dependent residuals by, e.g., fitting scalar regression models by means of the maximum entropy principle.

Furthermore, a unified concept for nonstationary time series prediction is presented. While predicting within the trained time span is reduced to evaluation of mixture models, the construction of predictive models beyond that time span requires the understanding of the underlying (learned) transition process in parameter space. To this end, the process of affiliation vectors interpreted as a time series of discrete probability distributions was approximated in terms of a (single) discrete time Markov chain. Predicting an affiliation vector for $t = T + 1$ then allows the approximation of $x_{T+1}$ via a mixture model and so on. However, we are aware that the resulting self-contained prediction strategy crucially relies on the assumption that the memory depth of the affiliation process is at most one. This issue is also the matter of future research.

## Appendix

In the appendix we compactly state the constrained quadratic program characterizing the optimal Markov regression model in the FEM-BV-Markov approach. For details see Section 2.c.iv).

Let $\mathbf{vec}(P_l^{(i)}) \in \mathbb{R}^{M^2}$ be the vector which results from concatenating all columns of the matrix $P_l^{(i)}$, i.e.,

$$\mathbf{vec}(P_l^{(i)}) \stackrel{\text{def}}{=} \left(P_l^{(i)}(\cdot, 1), \ldots, P_l^{(i)}(\cdot, M)\right) \in \mathbb{R}^{M^2}. \tag{136}$$

Furthermore, we denote the concatenation of all matrices $P_l^{(i)},\ l = 0, \ldots, k$ as

$$\boldsymbol{p}^{(i)} \stackrel{\text{def}}{=} \left(\mathbf{vec}(P_0^{(i)}), \ldots, \mathbf{vec}(P_k^{(i)})\right) \in \mathbb{R}^{(k+1)M^2}. \tag{137}$$

If we define

$$\boldsymbol{b}^{(i)} = -2 \sum_{t=0}^{T-1} \gamma_i(t) b(t) \quad \text{and } \boldsymbol{H}^{(i)} = 2 \sum_{t=0}^{T-1} \gamma_i(t) H(t) \tag{138}$$

with $u_0(t) \equiv 1$,

$$b(t) = \left(u_0(t)\mathbf{vec}(\pi_t \pi_{t+1}^{\dagger}), \ldots, u_k(t)\mathbf{vec}(\pi_t \pi_{t+1}^{\dagger})\right) \in \mathbb{R}^{(k+1)M^2} \tag{139}$$

and $H(t) \in \mathbb{R}^{((k+1)M^2)\times((k+1)M^2)}$ consists of blocks $H_{l_1 l_2}(t),\ l_1, l_2 = 0, \ldots, k$ with

$$H_{l_1 l_2}(t) = u_{l_1}(t) u_{l_2}(t) \mathrm{diag}(\pi_t \pi_t^\dagger, \ldots, \pi_t \pi_t^\dagger) \in \mathbb{R}^{M^2 \times M^2} \tag{140}$$

then for fixed $\Gamma$ the solution of the variational problem with respect to $i$-th local stationary Markov model $\Theta^{(i)} = (P_0^{(i)}, \ldots, P_k^{(i)})$,

$$\boldsymbol{L}(\Theta^{(i)}, \Gamma) = \sum_{t=0}^{T-1} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \left\| \pi_{t+1}{}^\dagger - \pi_t{}^\dagger \left( P_{(0)}^{(i)} + \sum_{l=1}^{k} u_l(t) P_{(l)}^{(i)} \right) \right\|_2^2 \to \min_{\Theta^{(i)}}, \tag{141}$$

subject to the constraints (45)–(48) is given by the solution of

$$\boldsymbol{L}(\boldsymbol{p}^{(i)}) = \tfrac{1}{2} \left\langle \boldsymbol{p}^{(i)}, \boldsymbol{H}^{(i)} \boldsymbol{p}^{(i)} \right\rangle_2 + \left\langle \boldsymbol{b}^{(i)}, \boldsymbol{p}^{(i)} \right\rangle_2 \to \min_{\boldsymbol{p}^{(i)}} \tag{142}$$

subject to the following linear constraints:

- Nonnegativity constraints (45):

$$\underbrace{(\mathrm{Id}_{M^2}, 0, \ldots, 0)}_{\in \mathbb{R}^{M^2 \times (k+1)M^2}} \boldsymbol{p}^{(i)} \geq 0, \tag{143}$$

- Constraints (46) and (47):

$$\underbrace{\begin{pmatrix} \mathcal{R}(\mathbf{1}_M) & 0 & 0 & 0 \\ 0 & \mathcal{R}(\mathbf{1}_M) & 0 & 0 \\ & & \ddots & \\ 0 & 0 & 0 & \mathcal{R}(\mathbf{1}_M) \end{pmatrix}}_{\in \mathbb{R}^{(k+1)M \times (k+1)M^2}} \boldsymbol{p}^{(i)} = \begin{pmatrix} \mathbf{1}_M \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{144}$$

  with $\mathcal{R}(\mathbf{1}_M) = (\mathrm{Id}_M, \ldots, \mathrm{Id}_M) \in \mathbb{R}^{M \times M^2}$.

- Overall nonnegativity constraint in (48):

$$\underbrace{\left(\mathrm{Id}_{M^2}, \hat{u}_1 \mathrm{Id}_{M^2}, \ldots, \hat{u}_k \mathrm{Id}_{M^2}\right)}_{\in \mathbb{R}^{M^2 \times (k+1)M^2}} \boldsymbol{p}^{(i)} \geq 0 \tag{145}$$

  for all $(\hat{u}_1, \ldots, \hat{u}_k) \in \{a_1, b_1\} \times \ldots \times \{a_k, b_k\}$ with

$$a_l = \min\{u_l(t) : t = 0, \ldots, T\} \quad \text{and} \quad b_l = \max\{u_l(t) : t = 0, \ldots, T\}. \tag{146}$$

**Acknowledgement**

We would like to thank the anonymous referees for their constructive criticism which helped us to improve the readability of the manuscript.

## References


[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Control **19** (1974), no. 6, 716–723. MR 54 #11691 Zbl 0314.62039

[2] A. N. Akansu and R. A. Haddad, *Multiresolution signal decomposition: transforms, subbands, and wavelets*, Academic Press, Boston, 1992. MR 93m:94004 Zbl 0947.94001

[3] L. E. Baum, *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*, Inequalities, III (O. Shisha, ed.), Academic Press, New York, 1972, pp. 1–8. MR 49 #6528

[4] L. E. Baum and T. Petrie, *Statistical inference for probabilistic functions of finite state Markov chains*, Ann. Math. Stat. **37** (1966), no. 6, 1554–1563. MR 34 #2137 Zbl 0144.40902

[5] C. M. Bishop, *Neural networks for pattern recognition*, Clarendon, Oxford, 1995. MR 97m:68172 Zbl 0868.68096

[6] F. Black, *Studies of stock price volatility changes*, Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section, American Statistical Association, Washington, DC, 1976, pp. 177–181.

[7] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, J. Econometrics **31** (1986), no. 3, 307–327. MR 87j:62169 Zbl 0616.62119

[8] D. Braess, *Finite elements: theory, fast solvers, and applications in solid mechanics*, 2nd ed., Cambridge University Press, Cambridge, 2001. MR 2001k:65002 Zbl 0976.65099

[9] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, Texts in Applied Mathematics, no. 31, Springer, New York, 1999. MR 2000k:60137 Zbl 0949.60009

[10] U. Çelikyurt and S. Özekici, *Multiperiod portfolio optimization models in stochastic markets using the mean-variance approach*, Eur. J. Oper. Res. **179** (2007), no. 1, 186–202. Zbl 1163.91375

[11] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM J. Numer. Anal. **36** (1999), no. 2, 491–515. MR 2000c:37026 Zbl 0916.58021

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Stat. Soc. Ser. B **39** (1977), no. 1, 1–38. MR 58 #18858 Zbl 0364.62022

[13] P. Deuflhard and C. Schütte, *Molecular conformation dynamics and computational drug design*, Applied mathematics entering the 21st century (J. M. Hill and R. Moore, eds.), SIAM, Philadelphia, 2004, pp. 91–119. MR 2296264 Zbl 1134.92004

[14] P. Deuflhard and M. Weber, *Robust Perron cluster analysis in conformation dynamics*, Linear Algebra Appl. **398** (2005), 161–184. MR 2005h:62166 Zbl 1070.15019

[15] R. F. Engle, *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation*, Econometrica **50** (1982), no. 4, 987–1007. MR 83j:62158 Zbl 0491.62099

[16] E. F. Fama, *The behavior of stock-market prices*, J. Bus. **38** (1965), no. 1, 34–105.

[17] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, and C. Schütte, *Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models*, J. Comput. Chem. **28** (2007), no. 15, 2453–2464.

[18] C. Franzke, D. Crommelin, A. Fischer, and A. J. Majda, *A hidden Markov model perspective on regimes and metastability in atmospheric flows*, J. Climate **21** (2008), no. 8, 1740–1757.

[19] T. Gasser and H.-G. Müller, *Kernel estimation of regression functions*, Smoothing techniques for curve estimation (T. Gasser and M. Rosenblatt, eds.), Lecture Notes in Math., no. 757, Springer, Berlin, 1979, pp. 23–68. MR 81k:62052 Zbl 0418.62033

[20] _____, *Estimating regression functions and their derivatives by the kernel method*, Scand. J. Stat. **11** (1984), no. 3, 171–185. MR 86h:62056 Zbl 0548.62028

[21] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL, 2004. MR 2004j:62001 Zbl 1039.62018

[22] D. Giannakis and A. J. Majda, *Quantifying the predictive skill in long-range forecasting, I: Coarse-grained predictions in a simple ocean model*, J. Climate **25** (2011), 1793–1813.

[23] _____, *Quantifying the predictive skill in long-range forecasting, II: Model error in coarse-grained Markov models with application to ocean-circulation regimes*, J. Climate **25** (2011), 1814–1826.

[24] D. Giannakis, A. J. Majda, and I. Horenko, *Information theory, model error, and predictive skill of stochastic models for complex nonlinear systems*, preprint, CIMS/NYU and University of Lugano, 2011, Submitted to Physica D.

[25] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, Int. Stat. Rev. **70** (2002), no. 3, 419–435. Zbl 1217.62014

[26] J. Hadamard, *Sur les problèmes aux dérivées partielles et leur signification physique*, Princeton Univ. Bull. **13** (1902), 49–52.

[27] J. D. Hamilton, *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica **57** (1989), no. 2, 357–384. MR 996941 Zbl 0685.62092

[28] A. Hoerl, *Application of ridge analysis to regression problems*, Chem. Eng. Prog. **58** (1962), no. 3, 54–59.

[29] I. Horenko, *On simultaneous data-based dimension reduction and hidden phase identification*, J. Atmos. Sci. **65** (2008), no. 6, 1941–1954.

[30] _____, *On robust estimation of low-frequency variability trends in discrete Markovian sequences of Atmospherical Circulation Patterns*, J. Atmos. Sci. **66** (2009), no. 7, 2059–2072.

[31] _____, *Finite element approach to clustering of multidimensional time series*, SIAM J. Sci. Comput. **32** (2010), no. 1, 62–83. MR 2011b:62009 Zbl 1206.62150

[32] _____, *On clustering of non-stationary meteorological time series*, Dyn. of Atmos. and Oceans **49** (2010), no. 2-3, 164–187.

[33] _____, *On identification of non-stationary factor models and its application to atmospherical data analysis*, J. Atmos. Sci. **67** (2010), no. 5, 1559–1574.

[34] _____, *Nonstationarity in multifactor models of discrete jump processes, memory and application to cloud modeling*, J. Atmos. Sci. **68** (2011), no. 7, 1493–1506.

[35] _____, *On analysis of nonstationary categorical data time series: dynamical dimension reduction, model selection, and applications to computational sociology*, Multiscale Model. Simul. **9** (2011), no. 4, 1700–1726. MR 2861255

[36] I. Horenko, E. Dittmer, A. Fischer, and C. Schütte, *Automated model reduction for complex systems exhibiting metastability*, Multiscale Model. Simul. **5** (2006), no. 3, 802–827. MR 2257236 Zbl 1122.60062

[37] I. Horenko, E. Dittmer, and C. Schütte, *Reduced stochastic models for complex molecular systems*, Comput. Vis. Sci. **9** (2006), no. 2, 89–102. MR 2247687

[38] I. Horenko, S. Dolaptchiev, A. Eliseev, I. Mokhov, and R. Klein, *Metastable decomposition of high-dimensional meteorological data with gaps*, J. Atmos. Sci. **65** (2008), no. 11, 3479–3496.

[39] I. Horenko, R. Klein, S. Dolaptchiev, and C. Schütte, *Automated generation of reduced stochastic weather models, I: Simultaneous dimension and model reduction for time series analysis*, Multiscale Model. Simul. **6** (2007), no. 4, 1125–1145. MR 2009e:62338 Zbl 1152.62056

[40] I. Horenko, J. Schmidt-Ehrenberg, and C. Schütte, *Set-oriented dimension reduction: localizing principal component analysis via hidden Markov models*, Computational life sciences II (M. R. Berthold, R. Glen, and I. Fischer, eds.), Lecture Notes in Comput. Sci., no. 4216, Springer, Berlin, 2006, pp. 74–85. MR 2279311

[41] I. Horenko and C. Schütte, *Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics*, Multiscale Model. Simul. **7** (2008), no. 2, 731–773. MR 2009j:37136 Zbl 1180.35175

[42] W. Huisinga, *Metastability of Markovian systems: a transfer operator approach in application to molecular dynamics*, Ph.D. thesis, Free University Berlin, 2001.

[43] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. (2) **106** (1957), 620–630. MR 19,335b Zbl 0084.43701

[44] ———, *Information theory and statistical mechanics, II*, Phys. Rev. (2) **108** (1957), 171–190. MR 20 #2898 Zbl 0084.43701

[45] R. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic Eng. **82** (1960), no. 1, 35–45.

[46] J. N. Kapur, *Maximum-entropy models in science and engineering*, Wiley, New York, 1989. MR 92b:00017 Zbl 0746.00014

[47] Koninklijk Nederlands Meteorologisch Instituut, http://climexp.knmi.nl/getindices.cgi?WMO=NCEPData/nino2_daily&STATION=NINO12&id=someone@somewhere&NPERYEAR=366&TYPE=i.

[48] H.-M. Krolzig, *Predicting Markov-switching vector autoregressive processes*, preprint 2000-W31, University of Oxford, 2000.

[49] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, *A generalized hidden Markov model for the recognition of human genes in DNA*, Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, eds.), AAAI Press, Menlo Park, CA, 1996, pp. 134–142.

[50] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974. MR 51 #2270 Zbl 0860.65028

[51] C. Lee and M.-H. Yu, *Protein folding and disease*, J. Biochem. Molec. Biol. **38** (2005), no. 3, 275–280.

[52] C. Loader, *Local regression and likelihood*, Springer, New York, 1999. MR 2000f:62005 Zbl 0929.62046

[53] J. B. MacQueen, *Some methods for classification and analysis of multivariate observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, I: Statistics (L. M. Le Cam and J. Neyman, eds.), University of California Press, Berkeley, CA, 1967, pp. 281–297. MR 35 #5078 Zbl 0214.46201

[54] A. J. Majda, C. L. Franzke, A. Fischer, and D. T. Crommelin, *Distinct metastable atmospheric regimes despite nearly Gaussian statistics: a paradigm model*, Proc. Natl. Acad. Sci. USA **103** (2006), no. 22, 8309–8314. MR 2007a:86004 Zbl 1160.86304

[55] A. J. Majda and X. Wang, *Non-linear dynamics and statistical theories for basic geophysical flows*, Cambridge University Press, Cambridge, 2006. MR 2009e:76214 Zbl 1141.86001

[56] G. McLachlan and D. Peel, *Finite mixture models*, Wiley, New York, 2000. MR 2002b:62025 Zbl 0963.62061

[57] P. Metzner, M. Weber, and C. Schütte, *Observation uncertainty in reversible Markov chains*, Phys. Rev. E (3) **82** (2010), no. 3, Paper #031114. MR 2012a:60202

[58] J.-J. Moreau, P. D. Panagiotopoulos, and G. Strang (eds.), *Topics in nonsmooth mechanics*, Birkhäuser, Basel, 1988.

[59] National Center for Biotechnology Information, *Saccharomyces cerevisiae chromosome I, complete sequence*, http://www.ncbi.nlm.nih.gov/nuccore/144228165?report=graph&to=30000.

[60] R. Preis, M. Dellnitz, M. Hessel, C. Schütte, and E. Meerbach, *Dominant paths between almost invariant sets of dynamical systems*, preprint 154, Deusche Forschungsgemeinschaft Schwerpunktprogramm, 2004.

[61] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes: the art of scientific computing*, 3rd ed., Cambridge University Press, Cambridge, 2007. MR 2009b:65001 Zbl 1132.65001

[62] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *Markov models of molecular kinetics: generation and validation*, J. Chem. Phys. **134** (2011), no. 17, Paper #174105.

[63] L. Putzig, D. Becherer, and I. Horenko, *Optimal allocation of a futures portfolio utilizing numerical market phase detection*, SIAM J. Financial Math. **1** (2010), 752–779. MR 2011k:91171 Zbl 1198.91241

[64] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE **77** (1989), no. 2, 257–286.

[65] C. Schütte, *Conformational dynamics: modelling, theory, algorithm, and application to biomolecules*, preprint SC 99-18, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, 1999.

[66] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys. **151** (1999), no. 1, 146–168. MR 2000d:92004

[67] C. Schütte and W. Huisinga, *Biomolecular conformations can be identified as metastable sets of molecular dynamics*, Handbook of numerical analysis, 10: Computational chemistry (C. Le Bris, ed.), North-Holland, Amsterdam, 2003, pp. 699–744. MR 2008396 Zbl 1066.81658

[68] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, *Markov state models based on milestoning*, J. Chem. Phys. **134** (2011), no. 20, Paper #204105.

[69] C. Schütte, F. Noé, E. Meerbach, P. Metzner, and C. Hartmann, *Conformation dynamics*, ICIAM 07: 6th International Congress on Industrial and Applied Mathematics (R. Jeltsch and G. Wanner, eds.), European Mathematical Society, Zürich, 2009, pp. 297–335. MR 2011k:82072 Zbl 1180.82239

[70] F. Takens, *Detecting strange attractors in turbulence*, Dynamical systems and turbulence, Warwick 1980 (D. A. Rand and L.-S. Young, eds.), Lecture Notes in Math., no. 898, Springer, Berlin, 1981, pp. 366–381. MR 83i:58065 Zbl 0513.58032

[71] A. N. Tikhonov, *On the stability of inverse problems*, Dokl. Akad. Nauk SSSR **39** (1943), no. 5, 195–198, In Russian; translated in C. R. (Doklady) Acad. Sci. URSS (N.S.) **39** (1943), no. 5, 176–179. MR 5,184e Zbl 0061.23308

[72] K. E. Trenberth, *The definition of El Niño*, Bull. Amer. Meteorol. Soc. **78** (1997), no. 12, 2771–2777.

[73] U.S. Energy Information Administration, *NYMEX futures prices*, http://tonto.eia.doe.gov/dnav/pet/pet_pri_fut_s1_d.htm.

[74] G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, no. 59, SIAM, Philadelphia, 1990. MR 91g:62028 Zbl 0813.62001

[75] Yahoo Finance, *Dow Jones industrial average: historical prices*, http://finance.yahoo.com/q/hp?s=^DJI+Historical+Prices.

[76] A. Zellner and R. A. Highfield, *Calculation of maximum entropy distributions and approximation of marginal posterior distributions*, J. Econometrics **37** (1988), no. 2, 195–209. MR 932140

Received July 29, 2011. Revised March 23, 2012.

PHILIPP METZNER: metznerp@usi.ch
*Institute of Computational Science, Faculty of Informatics, Università della Svizzera italiana, Via Guiseppe Buffi 13, CH-6900 Lugano, Switzerland*
http://icsweb.inf.unisi.ch/cms/index.php/people/24-philipp-metzner.html

LARS PUTZIG: putzigl@usi.ch
*Institute of Computational Science, Faculty of Informatics, Università della Svizzera italiana, Via Guiseppe Buffi 13, CH-6900 Lugano, Switzerland*
http://icsweb.inf.unisi.ch/cms/index.php/people/27-lars-putzig.html

ILLIA HORENKO: horenkoi@usi.ch
*Institute of Computational Science, Faculty of Informatics, Università della Svizzera italiana, Via Guiseppe Buffi 13, CH-6900 Lugano, Switzerland*
http://icsweb.inf.unisi.ch/cms/index.php/people/20-illia-horenko.html