Finite element approach to clustering of multidimensional time series*

Illia Horenko**[1]

[1] Institut für Mathematik, Freie Universität Berlin
Arnimallee 6, 14195 Berlin, Germany

Key words time series analysis, inverse problems, regularization, finite element method
Subject classification AMS: [62-07,62H30,62H25,65M60,60J10]

We present a new approach to clustering of time series based on a minimization of the *averaged clustering functional*. The proposed functional describes the mean distance between observation data and its representation in terms of **K** abstract models of a certain predefined class (not necessarily given by some probability distribution). For a fixed time series $x(t)$ this functional depends on **K** sets of model parameters $\Theta = (\theta_1, \dots, \theta_{\mathbf{K}})$ and **K** functions of cluster affiliations $\Gamma = (\gamma_1(t), \dots, \gamma_{\mathbf{K}}(t))$ (characterizing the affiliation of any element $x(t)$ of the analyzed time series to one of the **K** clusters defined by the considered model parameters). We demonstrate that for a fixed set of model parameters Θ the appropriate Tykhonov-type regularization of this functional with some regularization factor ϵ^2 results in a minimization problem similar to a variational problem usually associated with one–dimensional non–homogeneous partial differential equation. This analogy allows us to apply the finite element framework to the problem of time series analysis and to propose a numerical scheme for time series clustering. We investigate the conditions under which the proposed scheme allows a monotone improvement of the initial parameter guess wrt. the minimization of the discretized version of the regularized functional. We also discuss the interpretation of the regularization factor in the Markovian case and show its connection to metastability and exit times.

The computational performance of the resulting method is investigated numerically on multidimensional test data and is applied to the analysis of multidimensional historical stock market data.

This is a preliminary version. Do not circulate!

Introduction

Many application areas are characterized by the need to find some low-dimensional mathematical models for complex systems that undergo transitions between different phases. Such phases can be different circulation regimes in meteorology and climatology [1, 2, 3, 4], market phases in computational finance [5, 6] and molecular conformations in biophysics [7, 8, 9]. Regimes of this kind can sometimes not be directly observable (or "hidden") in the many dimensions of the system's degrees of freedom and can exhibit persistent or *metastable* behavior. If knowledge about the system is present only in the form of observation or measurement data, the challenging problem of identifying those metastable states together with the construction of reduced low-dimensional models becomes a problem of time series analysis and pattern recognition in high dimensions. The choice of the appropriate data analysis strategies (implying a set of method-specific assumptions on the analyzed data) plays a crucial role in correct interpretation of the available time series.

We present an approach to the identification of **K** hidden regimes for an abstract class of problems characterized by a set of model–specific parameters $\Theta = (\theta_1, \dots, \theta_{\mathbf{K}})$ and some positive bounded *model distance functional* $g(x_t, \theta_i)$ describing a quality of data representation in terms of model i. We demonstrate how the hidden regimes can be obtained in form of cluster affiliations $\Gamma = (\gamma_1(t), \dots, \gamma_{\mathbf{K}}(t))$ (characterizing the affiliation of any element of the analyzed time series to one of the K clusters defined

* Supported by the DFG research center MATHEON "Mathematics for key technologies" in Berlin.
** E-mail: horenko@math.fu-berlin.de

by the considered model parameters). In contrast to the commonly used approaches known in the literature (e.g., Hidden Markov models (HMMs) [5, 10, 11] or neuronal networks [12]), a key property of the presented numerical framework is that it does not impose any probabilistic assumptions on the type of the hidden process. The proposed numerical scheme is based on the *finite element method (FEM)*, a technique widely used and studied in context of partial differential equations (PDEs). Application of the FEM in the context of time series analysis can potentially help to transfer the advanced numerical techniques currently developed in the PDE setting and allow for the construction of new *adaptive methods* of data analysis.

The remainder of this paper is organized in the following way: Sec. 1 presents a construction of the *regularized clustering functional* and demonstrates some examples of typical *model distance functionals*. Subsequently, a FEM discretization of the problem is derived in Sec. 2, a numerical optimization algorithm is presented and its properties are investigated. In Sec. 3 we give an interpretation of the *regularization factor* in terms of *metastable homogeneous Markov-jump processes*. Finally, numerical examples in Sec. 5 illustrate the use of the presented framework.

1 The averaged clustering functional and its regularization

1.1 Model distance functional

Let $x(t) : [0,T] \to \Psi \subset \mathbf{R}^n$ be the observed n-dimensional time series. We look for $\mathbf{K}$ *models* characterized by $\mathbf{K}$ distinct sets of a priory unknown *model parameters*

$$\theta_1, \ldots, \theta_{\mathbf{K}} \in \Omega \subset \mathbf{R}^d, \tag{1}$$

(where d is the dimension of a model parameter space) for the description of the observed time series. Let

$$g\left(x_t, \theta_i\right) \quad : \quad \Psi \times \Omega \to [0, \infty)\,, \tag{2}$$

be a functional describing the *distance* from the observation $x_t = x(t)$ to the *model* i. For a given *model distance functional* (2), under *data clustering* we will understand the problem of finding for each t a vector $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_{\mathbf{K}}(t))$ called the *affiliation vector* (or vector of the *cluster weights*) together with model parameters $\Theta = (\theta_1, \ldots, \theta_{\mathbf{K}})$ which minimize the functional

$$\sum_{i=1}^{\mathbf{K}} \gamma_i(t) g\left(x_t, \theta_i\right) \quad \to \quad \min_{\Gamma(t), \Theta}\,, \tag{3}$$

subject to the constraints on $\Gamma(t)$:

$$\sum_{i=1}^{\mathbf{K}} \gamma_i(t) \quad = \quad 1, \quad \forall t \in [0,T] \tag{4}$$

$$\gamma_i(t) \quad \geq \quad 0, \quad \forall t \in [0,T]\,, \quad i = 1, \ldots, \mathbf{K}. \tag{5}$$

In the following, we will give three examples of the *model distance functional* (2) for three classes of cluster models: (I) *geometrical clustering*, (II) *Gaussian clustering* and (III) *clustering based on the essential orthogonal functions (EOFs)*.

Example (I): Geometrical Clustering One of the most popular clustering methods in multivariate data-analysis is the so-called *K-means algorithm* [13]. It is based on the iterative minimization of the distance from the data points to a set of K *cluster centers* which are recalculated in each iteration step. The affiliation to a certain cluster i is defined by the proximity of the observation $x_t \in \Psi$ to the cluster center $\theta_i \in \Psi$. In this case the *model distance functional* (2) takes the form of the square of the simple Euclidean distance between the points in n dimensions:

$$g\left(x_t, \theta_i\right) \quad = \quad \parallel x_t - \theta_i \parallel^2 . \tag{6}$$

Example (II): Gaussian Clustering Another frequently used clustering algorithm is based on the identification of Gaussian sets in the analyzed data [10, 11]. It is assumed that the data x belonging to the same cluster i is distributed according to the multivariate normal distribution

$$p_i(x) = \sqrt{\det\left(2\pi\Sigma_i^{-1}\right)} \exp\left(-0.5\,(x-\mu_i)^{\mathbf{T}}\,\Sigma_i^{-1}\,(x-\mu_i)\right) \tag{7}$$

for all $x \in \Psi$ with $\theta_i = (\mu_i, \Sigma_i)$, μ_i being the expectation value, and Σ_i the covariance matrix of p_i. In this case the *model distance functional* (2) can be expressed as a normed negative log-likelihood of (7):

$$g\,(x_t, \theta_i) \quad = \quad \| \, x_t - \mu_i \, \|^2_{\Sigma_i^{-1}}, \tag{8}$$

where $\| \cdot \|_{\Sigma_i^{-1}}$ denotes the norm induced by the covariance matrix of the Gaussian distribution i.

Example (III): EOF clustering In many cases the dimensionality of the data x_t can be reduced to few *essential degrees of freedom* without significant loss of information. One of the most popular *dimension reduction* approaches used in applications is the method of *essential orthogonal functions (EOFs)* also known under the name of *principal component analysis (PCA)* [14]. As was demonstrated recently, it is possible to construct clustering methods based on the decomposition of data sets according to differences in their *essential degrees of freedom* allowing to analyze data of a very high dimensionality [15, 16, 17]. If the cluster i is characterized by a linear m-dimensional manifold ($m \ll n$) of *essential degrees of freedom*, the respective model parameter is defined by the corresponding orthogonal projector $\theta_i = \mathcal{T}_i \in \mathbf{R}^{n \times m}$ and the *model distance functional* (2) is given by the Euclidean distance between the original data x and its orthogonal projection on the manifold:

$$g\,(x_t, \theta_i) \quad = \quad \| \, x_t - \mathcal{T}_i \mathcal{T}_i^{\mathbf{T}} x_t \, \|^2 \, . \tag{9}$$

1.2 The averaged clustering functional and its regularization

Instead of solving the minimization problem (3) for each available element $x_t \in \Psi, t \in [0, T]$ from the observed time series *separately*, one can approach all of the functional optimizations *simultaneosly* and minimize the *averaged clustering functional* $\mathbf{L}$:

$$\mathbf{L}(\Theta, \Gamma) \quad = \quad \int_0^T \sum_{i=1}^{\mathbf{K}} \gamma_i\,(t)\, g\,(x_t, \theta_i)\, dt \to \min_{\Gamma, \Theta}, \tag{10}$$

subject to the constraints (1) and (4-5). The expression in (10) is similar to one that is typically used in the context of *finite mixture models* [11, 18] but is more general, since neither the function $g\,(\cdot, \cdot)$ nor $\Gamma\,(\cdot)$ have to be connected to some probabilistic models of the data (which is the case for *finite mixture models*).

However, direct numerical solution of the problem (10) is hampered by the three following facts: (i) the optimization problem is *infinitely-dimensional* (since $\Gamma(t)$ belongs to some not yet specified function class), (ii) the problem is *ill-posed* since the number of unknowns can be higher then the number of known parameters, and (iii) because of the non-linearity of g the problem is in general *non-convex* and the numerical solution gained with some sort of *local minimization algorithm* depends on the initial parameter values [19].

One of the possibilities to approach the problems (i)-(ii) simultaneously is first to incorporate some *additional information* about the observed process (e.g., in the form of *smoothness assumptions* in space of functions $\Gamma\,(\cdot)$) and then apply a finite Galerkin-discretization of this infinite-dimensional Hilbert space. For example, we can assume the *weak differentiability* of functions γ_i, i. e.:

$$|\gamma_i|_{\mathcal{H}^1(0,T)} \quad = \quad \| \, \partial_t \gamma_i\,(\cdot) \, \|_{\mathcal{L}_2(0,T)} = \int_0^T \left(\partial_t \gamma_i\,(t)\right)^2 dt \leq C_\epsilon^i < +\infty, \quad i = 1, \ldots, \mathbf{K}. \tag{11}$$

For a given observation time series, the above constraint limits the total number of transitions between the clusters and, as it will be demonstrated later in Section 3, is connected to the *metastability* of the *hidden process* $\Gamma(t)$.

Another possibility to incorporate the *a priori information* from (11) into the optimization is to modify the functional (3) and to write it in the *regularized* form

$$\mathbf{L}^{\epsilon}(\Theta,\Gamma,\epsilon^2) \quad = \quad \mathbf{L}(\Theta,\Gamma)+\epsilon^2\sum_{i=1}^{\mathbf{K}}\int_0^T\left(\partial_t\gamma_i\left(t\right)\right)^2dt\rightarrow\min_{\Gamma,\Theta}. \tag{12}$$

In the following, we will demonstrate a numerical approach to the optimization of this *regularized clustering functional*(12) subject to the constraints (1) and (4-5).

2 Finite elements approach to minimization of the regularized clustering functional

2.1 FEM-discretization

Let $\{0=t_1,t_2,\ldots,t_{N-1},t_N=T\}$ be a finite subdivision of the time interval $[0,T]$. We define a set of continuous functions $\{v_1(t),v_2(t),\ldots,v_N(t)\}$ with the *local support* on $[0,T]$, i. e.,$v_1(t)\neq 0$ for $t\in(t_1,t_2)$ (and zero elsewhere), $v_k(t)\neq 0$ for $t\in\left(t_{k-1},t_{k+1}\right),k=2,\ldots,N-1$ (and zero elsewhere),$v_N(t)\neq 0$ for $t\in(t_{N-1},t_N)$ (and zero elsewhere). These functions are called *finite element basis* and there are lot of possible sets of such functions known from the literature on partial differential equations (PDEs) [20], like e.g., piecewise linear *hat functions* shown in Fig. 1.

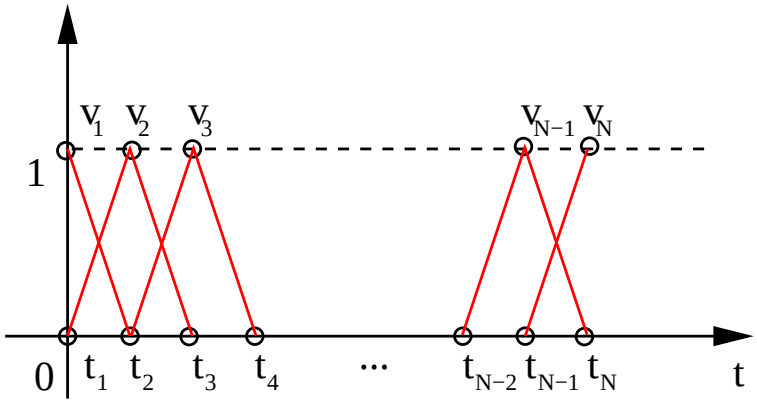

Fig. 1 Linear finite elements in one dimension.

Assuming that $\gamma_i\in\mathcal{H}^1\left(0,T\right)$ we can write

$$\begin{aligned}\gamma_i &= \tilde{\gamma}_i+\delta_N\\ &= \sum_{k=1}^{N}\tilde{\gamma}_i^{(k)}v_k+\delta_N,\end{aligned} \tag{13}$$

where $\tilde{\gamma}_i^{(k)}=\int_0^T\gamma_i\left(t\right)v_k\left(t\right)dt$ and δ_N is some *discretization error*. Substituting (13) in (12) we get

$$\mathbf{L}^{\epsilon} \quad = \quad \tilde{\mathbf{L}}^{\epsilon}+\mathcal{O}\left(\delta_N\right)\rightarrow\min_{\tilde{\gamma}_i,\Theta}, \tag{14}$$

where $\tilde{\mathbf{L}}^{\epsilon}$ is a finite-dimensional version of the original functional (12):

$$\tilde{\mathbf{L}}^{\epsilon}=\sum_{i=1}^{\mathbf{K}}\int_0^T\left[\tilde{\gamma}_i\left(t\right)g(x_t,\theta_i)+\epsilon^2\left(\partial_t\tilde{\gamma}_i\left(t\right)\right)^2\right]dt. \tag{15}$$

After several obvious transformations and using of the locality of the finite element support we obtain:

$$\begin{aligned}\tilde{\mathbf{L}}^{\epsilon} &= \sum_{i=1}^{\mathbf{K}}\Bigg[\tilde{\gamma}_i^{(1)}\int_{t_1}^{t_2} v_1(t)g(x_t,\theta_i)dt + \sum_{k=2}^{N-1}\tilde{\gamma}_i^{(k)}\int_{t_{k-1}}^{t_{k+1}} v_k(t)g(x_t,\theta_i)dt + \\ &+\tilde{\gamma}_i^{(N)}\int_{t_{N-1}}^{t_N} v_N(t)g(x_t,\theta_i)dt + \epsilon^2\sum_{k=1}^{N-1}\Bigg(\left(\tilde{\gamma}_i^{(k)}\right)^2\int_{t_k}^{t_{k+1}}\left(\partial_t v_k(t)\right)^2 dt - \\ &-2\tilde{\gamma}_i^{(k)}\tilde{\gamma}_i^{(k+1)}\int_{t_k}^{t_{k+1}}\partial_t v_k(t)\partial_t v_{k+1}(t)dt + \left(\tilde{\gamma}_i^{(k+1)}\right)^2\int_{t_k}^{t_{k+1}}\left(\partial_t v_{k+1}(t)\right)^2 dt\Bigg)\Bigg]. \quad (16)\end{aligned}$$

Denoting the vector of *discretized affiliations* to cluster i as $\bar{\gamma}_i = \left(\tilde{\gamma}_1^{(1)},\ldots,\tilde{\gamma}_i^{(N)}\right)$, vector of *discretized model distances* as

$$a(\theta_i) = \left(\int_{t_1}^{t_2} v_1(t)g(x_t,\theta_i)dt,\ldots,\int_{t_{N-1}}^{t_N} v_N(t)g(x_t,\theta_i)dt\right), \quad (17)$$

and the symmetric tridiagonal *stiffness-matrix* of the finite element set as $\mathbf{H}$

$$\mathbf{H} = \begin{pmatrix} \int_{t_1}^{t_2} v_1^2(t)dt & \int_{t_1}^{t_2} v_1(t)v_2(t)dt & 0 & \ldots & 0 \\ \int_{t_1}^{t_2} v_1(t)v_2(t)dt & \int_{t_2}^{t_3} v_2^2(t)dt & \int_{t_2}^{t_3} v_2(t)v_3(t)dt & \ldots & 0 \\ \ldots & \ldots & \ldots & \ddots & 0 \\ 0 & 0 & 0 & \ldots & \int_{t_{N-1}}^{t_N} v_N^2(t)dt \end{pmatrix}. \quad (18)$$

With the help of (17-18) we can re-write (16) as

$$\tilde{\mathbf{L}}^{\epsilon} = \sum_{i=1}^{\mathbf{K}}\left[a(\theta_i)^{\mathbf{T}}\bar{\gamma}_i + \epsilon^2\bar{\gamma}_i^{\mathbf{T}}\mathbf{H}\bar{\gamma}_i\right] \to \min_{\bar{\gamma}_i,\Theta}, \quad (19)$$

subject to (1), the discretized version of equality constraints (4)

$$\sum_{i=1}^{\mathbf{K}}\tilde{\gamma}_i^{(k)} = 1, \quad \forall k = 1,\ldots,N, \quad (20)$$

and the inequality constraints (5)

$$\tilde{\gamma}_i^{(k)} \geq 0, \quad \forall k = 1,\ldots,N, \quad i = 1,\ldots,\mathbf{K}. \quad (21)$$

2.2 Numerical method and monotonicity conditions

The minimization problem (19-21), for a fixed set of *cluster model parameters* Θ reduces to a quadratic optimization problem with linear constraints which can be solved by standard tools of quadratic programming (QP) like, e.g., the *ellipsoid methods* or the *interior point methods* that converge in polynomial time [21, 22, 19]. If, in addition, it is possible to minimize the problem (19 -21) wrt. parameters Θ for a *fixed* set of *discretized cluster affiliations* $\bar{\gamma}_i$, we can split the original optimization problem in two consecutive parts repeated in each iteration step of the following algorithm:

Algorithm.

Setting of optimization parameters and generation of initial values:

· Set the number of clusters $\mathbf{K}$, regularization factor ϵ^2, finite discretization of the time interval $[0,T]$, and the optimization tolerance TOL

· Set the iteration counter $j = 1$

· Choose random initial $\bar{\gamma}_i^{[1]}, i = 1,\ldots,\mathbf{K}$ satisfying (20-21)

· Calculate $\Theta^{[1]} = \arg\min_{\Theta} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta, \bar{\gamma}_i^{[1]}\right)$ subject to (1)

Optimization loop:

do

· Compute $\bar{\gamma}^{[j+1]} = \arg\min_{\bar{\gamma}} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j]}, \bar{\gamma}\right)$ satisfying (20-21) applying QP

· Calculate $\Theta^{[j+1]} = \arg\min_{\Theta} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta, \bar{\gamma}_i^{[j+1]}\right)$

· $j := j + 1$

while $\left|\tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j+1]}, \bar{\gamma}_i^{[j+1]}\right) - \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j]}, \bar{\gamma}_i^{[j]}\right)\right| \geq \text{TOL}.$

A solution of the problem $\Theta^{[j+1]} = \arg\min_{\Theta} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta, \bar{\gamma}_i^{[j+1]}\right)$ can be calculated by, e.g., re-writing the problem in Lagrangian form and applying the Newton–method. Moreover, in many cases this problem can even be solved analytically (this depends on the form of the *model distance functional*). For all 3 examples of $g(x_t, \theta)$ presented above this can be done. Note that due to the non-linearity of $g(\cdot, \theta)$ wrt. θ, the minimized functional $\tilde{\mathbf{L}}^{\epsilon}$ is non-convex. This feature will prohibit us to use the standard results of convex optimization theory to show the *convergence* of the presented algorithm. The following theorem describes the conditions under which the above algorithm will *monotonously minimize* the *energy* (19).

Theorem 2.1 *Let for a given observed time series $x(t) : [0, T] \to \Psi \subset \mathbf{R}^n$ the model distance functional g be chosen such that (2) is fulfilled, Ψ and Ω are compact, $g(x_t, \cdot)$ is continuously differentiable function of θ and*

$$\frac{\partial}{\partial \Theta} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^*, \bar{\gamma}\right) = 0 \tag{22}$$

has a solution $\Theta^* = (\theta_1^*, \ldots, \theta_{\mathbf{K}}^*), \theta_i * \in \Omega$ for any fixed $\bar{\gamma}$ satisfying (20-21) and $\frac{\partial^2}{\partial \Theta^2} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^*, \bar{\gamma}\right)$ exists and is positive definite. Then for any $\epsilon^2 \geq 0$ and any finite non-negative finite elements set $\{v_1(t), v_2(t), \ldots, v_N(t)\} \in \mathcal{L}_2(0, T)$ such that the respective stiffness-matrix *$\mathcal{H}$ is positive semidefinite, the above algorithm is* monotone*, i. e., for any $j \geq 1$*

$$\tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j+1]}, \bar{\gamma}_i^{[j+1]}\right) \leq \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j]}, \bar{\gamma}_i^{[j]}\right). \tag{23}$$

Proof. Since $\epsilon^2 \mathbf{H}$ is positive semidefinite and $0 \leq g(x_t, \theta) \leq \bar{g} < +\infty$, for any fixed $\theta \in \Omega$ the functional $\tilde{\mathbf{L}}^{\epsilon}$ is convex. Moreover, (2) implies that $\tilde{\mathbf{L}}^{\epsilon}$ is bounded from below and constraints (20-21) define a non-empty closed convex domain. In this case the problem $\bar{\gamma}^{[j+1]} = \arg\min_{\bar{\gamma}} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j]}, \bar{\gamma}\right)$ satisfying (20-21) has a *global minimizer* $\bar{\gamma}_i^{[j+1]}$, in particular

$$\tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j]}, \bar{\gamma}_i^{[j+1]}\right) \leq \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j]}, \bar{\gamma}_i^{[j]}\right). \tag{24}$$

Moreover, due to (22) and since the Hesse-matrix $\frac{\partial^2}{\partial \Theta^2} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^*, \bar{\gamma}_i^{[j+1]}\right)$ exists and is positive-definitive, the solution of $\Theta^{[j+1]} = \arg\min_{\Theta} \tilde{\mathbf{L}}^{\epsilon}\left(\Theta, \bar{\gamma}_i^{[j+1]}\right)$ subject to (1) exists and

$$\tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j+1]}, \bar{\gamma}_i^{[j+1]}\right) \leq \tilde{\mathbf{L}}^{\epsilon}\left(\Theta^{[j]}, \bar{\gamma}_i^{[j+1]}\right). \tag{25}$$

Finally, (24) together with (25) results in (23). □

3 The Markovian case: regularization factor and metastability

The proposed numerical method results in a *local improvement of the energy* (19). The minimization problem was obtained by a finite element discretization of the continuous *regularized clustering problem* (12) under conditions (4-5). However, it is not a priori clear what is the connection between the discrete solution we obtain with the above algorithm and the minimizer of the original *averaged clustering*

functional (10). There are two main questions to be answered: (i) what is the *discretization error* δ_N introduced on the way from (12) to (19) and how does it influence the quality of the resulting minimizers, and (ii) what is the influence of the regularization factor $\epsilon^2 \sum_{i=1}^{\mathbf{K}} \int_0^T \left(\partial_t \gamma_i(t)\right)^2 dt$.

Concerning the first question, it seems clear that with increasing number N of time discretization points the error δ_N from (13) will decrease and the overall difference between the continuous and discretized versions of the regularized functional will be getting smaller. For a rigorous mathematical justification of this feature and for the estimation of the discretization error, one can apply the theory developed for partial differential equations. This is a matter of future research. Here we would only like to mention the fact that in practical applications the observation time series are almost always available only in a discrete form (since the measurements of the real life processes can be typically acquired only at some discrete moments in time). This means that one actually starts with a discretized problem and therefore there is an upper limit for N given by the number of times the process was observed. However, we want to keep the continuous representation of the optimization problem in order to be able to construct the *adaptive FEM* scheme in future. Control over the *discretization error* in (13) will allow to implement the adaptivity exactly in the same way as it is done in the theory of PDEs [20].

Concerning the influence of the regularization factor, it is intuitively clear that the penalization of the derivative norm in form of regularization (12) or in form of the constraint (11) has a *smoothing effect*. More specifically, in the case of a piecewise–constant function Γ, regularization will result in restriction of the number of transitions between the clusters. This means that the *cluster affiliation functions* γ_i obtained by the optimization of the *regularized functional* will be more and more *metastable* for increasing ϵ^2 or C_{ϵ^2}, i. e., the observed process will stay more and more time in the respective identified cluster before making a transition to the next identified cluster. Metastability is associated with the so called *mean exit times* τ_i^{exit} describing the mean time a process will stay in the state i before leaving to some other state. In the context of *discrete homogeneous Markovian processes*, the *mean exit time* τ_i^{exit} can be quantified with the help of the *transition matrix* $\mathbf{P}$ [23]:

$$\tau_i^{\text{exit}} \quad = \quad \frac{\Delta_t}{1-\mathbf{P}_{ii}}, \tag{26}$$

where Δ_t is the discrete time step of the Markov chain and $\mathbf{P}_{ii}$ is the Markovian probability to stay in the same state i after one time step Δ_t.

We will investigate the effect of regularization for a discretized optimization problem. As a finite element basis we take the *linear finite elements* shown in Fig. 1 on an equidistant time grid with step Δ_t:

$$v_k(t) \quad = \quad \begin{cases} \frac{t-t_k}{\Delta_t} & 2 \le k \le N-1, t \in [t_{k-1}, t_k], \\ \frac{t_{k+1}-t}{\Delta_t} & 2 \le k \le N-1, t \in [t_k, t_{k+1}], \\ \frac{t_2-t}{\Delta_t} & k=1, t \in [t_1, t_2] \\ \frac{t-t_{N-1}}{\Delta_t} & k=N, t \in [t_{N-1}, t_N]. \end{cases} \tag{27}$$

The following theorem explains a connection between the *regularization factor* and *metastability* of the resulting clustering in the Markovian case.

Theorem 3.1 *Let $\Gamma^\epsilon(t) = (\gamma_1^\epsilon, \ldots, \gamma_{\mathbf{K}}^\epsilon)$ be a solution of the optimization problem (19 -21),(1) resulting from application of the positive linear finite elements discretization $v_l(t)$(27) with a constant time step Δ_t and*

$$[\gamma_i^\epsilon]^{\mathbf{T}} \mathbf{H} [\gamma_i^\epsilon] \quad = \quad C_\epsilon^i, \quad i = 1, \ldots, \mathbf{K}, \tag{28}$$

where $[\cdot]$ is a component-wise roundoff operation towards the nearest integer. If the values $\gamma_i(t_k) = \sum_{l=1}^N \left[\gamma_{i,l}^\epsilon\right] v_l(t_k), i = 1, \ldots, \mathbf{K}$ of respective cluster affiliation function are considered as an output of a time-discrete homogeneous Markov-jump process with time step Δ_t (where $t_1, \ldots, t_N$ is the subdivision of $[0,T]$), then the respective mean exit times τ_i^{exit} for $\mathbf{K}$ Markov states are

$$\tau_i^{exit} \quad = \quad \frac{(N-1)}{C_\epsilon^i}, \quad i = 1, \ldots, \mathbf{K}. \tag{29}$$

P r o o f. Since $v_k(t)$ has the form of (27), the *stiffness–matrix* $\mathbf{H}$ in (19) will be symmetric and tridiagonal with $2/\Delta_t$ on the main diagonal, $-1/\Delta_t$ on both secondary diagonals and zero elsewhere. Therefore we can write:

$$C_\epsilon^i \;=\; \frac{1}{\Delta_t}\sum_{l=1}^{\mathbf{K}}\left(\left[\gamma_{i,(l+1)}^\epsilon\right]-\left[\gamma_{i,l}^\epsilon\right]\right)^2. \tag{30}$$

Moreover, because of the discreteness of the processes $\gamma_i(t_k)$ in state-space and time we get

$$\Delta_t C_\epsilon^i \;\leq\; (N-1)\,, \tag{31}$$

since $\left[\gamma_{i,l}^\epsilon\right]$ can only take values 0 and 1. Therefore, $(N-1-\Delta_t C_\epsilon^i)$ is the number of times the observation process stayed in cluster i without going elsewhere in the next step. The *maxim log-likelihood* estimate of the respective homogeneous Markovian probability is

$$\mathbf{P}_{ii} = \frac{\left(N-1-\Delta_t C_\epsilon^i\right)}{N-1}, \quad i=1,\ldots,\mathbf{K}, \tag{32}$$

and the corresponding *maximum log-likelihood* estimate of the *mean exit time* is given by the expression (29). □

The above theorem demonstrates a connection between the effect of *regularization* and *metastability* of the resulting clustering when the finite element discretization with uniform time step Δ_t is interpreted as the output of a homogeneous Markov-jump process on the same time scale. It is intuitively clear that with growing ϵ^2 the energy-norm $\|\cdot\|_{\mathcal{H}^1(0,T)}$ of the resulting optimal vector γ_i^ϵ will get smaller. That means that according to (29), respective mean exit times get longer and the corresponding cluster decomposition becomes more and more metastable. However, rigorous justification of the connection between C_ϵ^i, ϵ^2 and exit times for continuous case is not clear yet and will be a matter of further research.

4 Estimation of confidence intervals and choice of K

The quality of the resulting clustering and reliability of the model parameters Θ in any specific case will be very much dependent on the original data, especially on the length of the available time series. The shorter the observation sequence, the bigger is the uncertainty of the resulting parameters. The same is true if the number $\mathbf{K}$ of clusters is being increased for fixed length of the observed time series: the bigger $\mathbf{K}$, the higher will be the uncertainty for each of the states. Therefore in order to be able to statistically distinguish between different hidden states we need to get some notion of the *model robustness*. This can be achieved through the estimation of confidence intervals for the model parameters Θ. For example, in the case of *Gaussian clustering* with the *model distance functional* (7), this can be done in a standard way by multivariate statistical analysis since the variability of the *estimated covariance matrices* is given by the *Wishart distribution* [24], whereas the confidence intervals of μ_i can be acquired from the respective standard deviations [24]. In other cases the confidence intervals can be acquired with the help of the *Fisher information matrix* or other standard tools of *information theory* [25].

If there exist two states with overlapping confidence intervals for each of the respective model parameters, then those are statistically indistinguishable, $\mathbf{K}$ should be reduced and the optimization repeated. In other words, confidence intervals implicitly give a natural upper bound for the number of possible clusters. On the other hand, the spectral theory of Markov processes connects the number $\mathbf{K}$ of metastable states with the number of the dominant eigenvalues in the so called *Perron cluster* [26]. This allows to apply the *Perron cluster - cluster analysis (PCCA)* [27] to find a lower bound for $\mathbf{K}$. Both these criteria in combination can help to find the *optimal* number $\mathbf{K}$ of clusters in each specific application.

5 Numerical examples

To investigate the proposed framework numerically, we will present two examples: (i) application to a test model system to study the numerical performance of the method, and (ii) analysis of multidimensional stock market data.

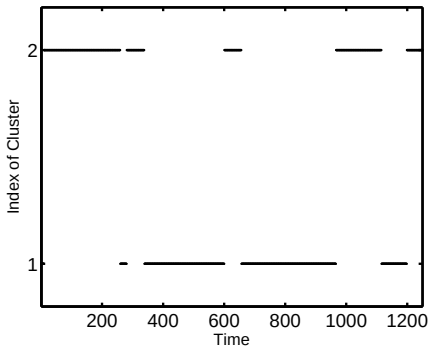

Fig. 2 Hidden discrete process switching between data-clusters.

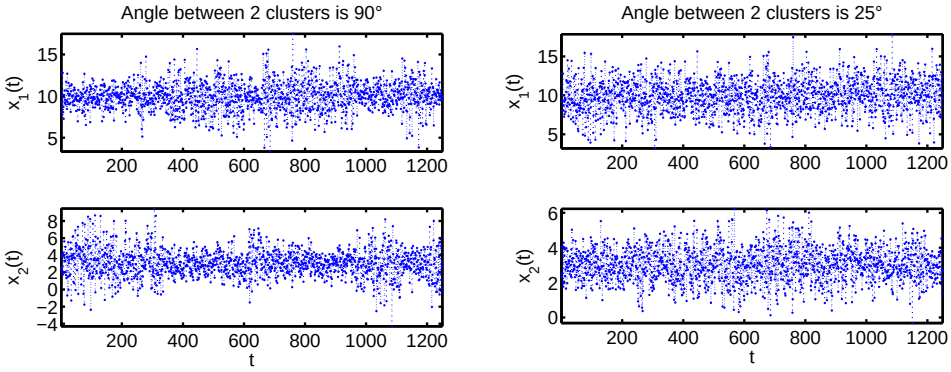

Fig. 3 Time series in "Gaussian" degrees of freedom x_1, x_2 generated by the hidden switching process from Fig. 2.

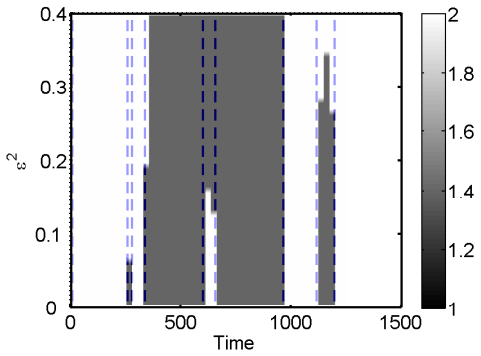

Fig. 4 Influence of the *regularization parameter* ϵ^2 on the identified *cluster affiliation function* γ^ϵ. Grayscaling represents the values of the function $\gamma^\epsilon(t)$ calculated for different values of ϵ^2 at different times calculated for the time series generated with $\mathbf{K} = 2, T = 1250, N = 50, m = 1, \alpha = 25$ and cluster switching from Fig. 2. Dashed lines denote the moments when the original process from Fig. 2 was switching between the clusters.

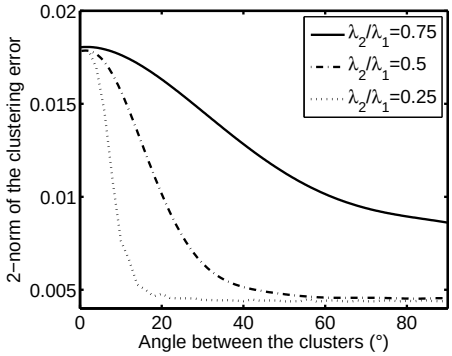

Fig. 5 l_2 norm of the difference between the original switching process from Fig. 2 and its estimate based on data analysis for different angles α and different ratios of dominant covariance matrix eigenvalues λ_1 and λ_2.

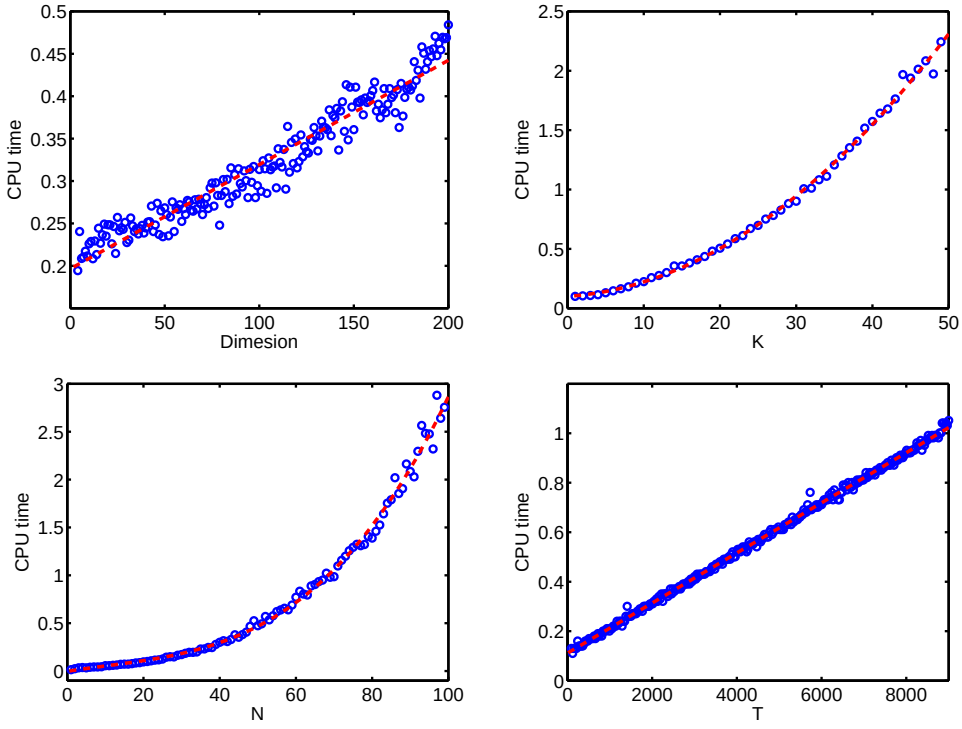

Fig. 6 Statistics of the computational performance obtained from 1000 realizations of the model process with $\alpha = 25, \epsilon^2 = 0.1$ generated for a switching process according to Fig. 2. Dashed lines represent the polynomials of respective order fitted to the simulation data (circles). Four panels demonstrate the performance of the method: (i) wrt. observation dimension n for $\mathbf{K} = 2, m = 1, T = 1250, N = 50$ ($\mathbf{O}(n)$, upper left panel), (ii) wrt. number of clusters $\mathbf{K}$ for $m = 1, T = 1250, N = 50, n = 100$ ($\mathbf{O}(\mathbf{K}^2)$, upper right panel), (iii) wrt. number of finite elements N for $\mathbf{K} = 2, m = 1, T = 1250, n = 100$ ($\mathbf{O}(\mathbf{N}^2 \log(N))$, lower left panel), and (iv) wrt. length T of the time series for $\mathbf{K} = 2, m = 1, N = 50, n = 100$ ($\mathbf{O}(T)$, lower right panel).

5.1 Test system: computational performance

First we will apply the method to a test system consisting of a given discrete process switching between two n-dimensional data clusters (see Fig. 2). Data in clusters is distributed according to a tensor product of a two-dimensional Gaussian distribution (7) with a $(n-2)$-dimensional uniform distribution on the

interval $[0, 10]$. Both Gaussians have the same expectation value

$$\mu_{1,2} = \begin{pmatrix} 10 \\ 3 \end{pmatrix}, \tag{33}$$

and one of the covariance matrices is a rotation of the other covariance matrix through some predefined angle α:

$$\Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix} \Sigma_1, \tag{34}$$

The distributions are chosen in such a way that neither the *geometric clustering* approach nor the *Gaussian clustering* will be able to cluster the data properly. Such kinds of distributions are very common in computational finance, e.g., in the analysis of stock returns where the change of the market phase is connected to the change of volatility of the underlying stochastic process [6]. Therefore in the following we will use the *EOF model distance functional* (9) with *linear finite elements* (27) for the clustering of the resulting time series. As an example, Fig. 3 demonstrates two time series in Gaussian degrees of freedom generated with the help of the switching process from Fig. 2 for two different values of the angle α (90 and 25).

Fig. 4 demonstrates the effect of regularization on the optimization process: as was already discussed above, an increase of ϵ^2 leads to a growing metastability of the resulting *cluster affiliation function*. As we see from Fig. 4 this results in a *coarse graining of the identified affiliation functions*, i. e., only "long living" structures in γ "survive" with increasing ϵ^2. Fig. 5 demonstrates the sensitivity of the optimization procedure to the input time series, i. e., it shows the clustering error as a function of two variables: (i) angle α between the clusters and (ii) ratio of two dominant eigenvalues. As expected, the quality of clustering is increasing with increasing α and decreasing λ_2/λ_1. This is explained by the growing numerical separability of the dominant subspaces in the context of the *EOF model distance functional*.

Fig. 6 demonstrates the computational performance of the resulting clustering method measured experimentally from 1.000 different realizations of the analyzed model trajectories (all generated with the same cluster affiliation function, see Fig. 2) : it is linear in the observation dimension and time series length, quadratic in the number of clusters and polynomial in the number N of finite elements (scales approx. $\mathbf{O}\left(N^2 \log(N)\right)$). It should be mentioned that the standard QP-solver from MATLAB was applied in the current realizations of the code. The sparsity structure of the QP subproblem allows to use *sparse QP (SQP)* solvers available on the market. This will reduce a numerical cost of the method to $\mathbf{O}\left(N \log(N)\right)$.

5.2 Analysis of stock daily returns: NASDAQ

Finally we will apply the numerical method to identify the hidden market phases based on daily returns of the 1106 stocks from the *NASDAQ* stock exchange between Jan.03 2007 and Jan.10 2008 (data is acquired from *http://finance.yahoo.com*). Our aim is to identify the *hidden market phases* and to interpret them in the context of global market dynamics. Due to the fact that the time series has only 257 elements (because there are 257 trading days in the analyzed time interval), we do not expect to find more than few statistically distinguishable clusters. We apply the *EOF model distance functional* (9) with *linear finite elements* (27) with $\mathbf{K} = 2, m = 1, T = 257, N = 50, n = 1106$. Because of the fact that the proposed numerical framework approaches only towards a *local minimum* of the *regularized averaged clustering functional* (dependent on the initial parameter values chosen for optimization), we repeat the optimization 100 times with different randomly generated initial guesses for the parameters and keep the result with the lowest value of the *averaged clustering functional* (10).

Results of this procedure for different values of ϵ^2 are demonstrated in Fig. 7. As can be seen, the lowest value of the *averaged clustering functional* is achieved for the optimization with highest *regularity* $\epsilon^2 = 0.02$, this also means the highest *metastability* of the process switching between the market phases. Further increase of ϵ^2 leads to a complete domination of the regularization part in optimization and suppression of the part corresponding to the *averaged clustering functional*. This results in identification

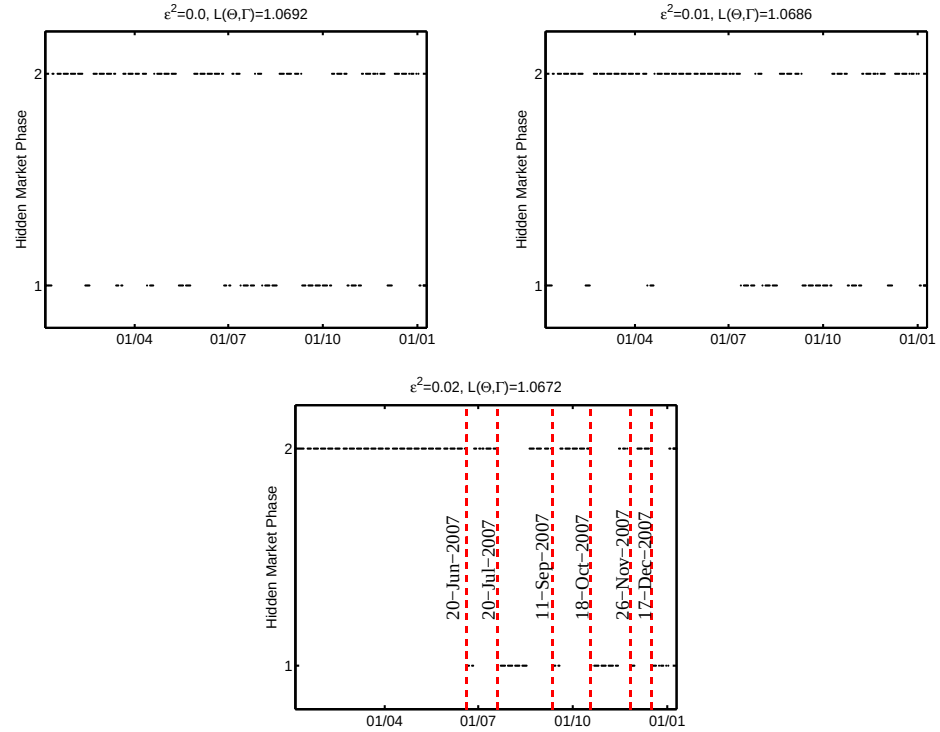

Fig. 7 Hidden market phases for different values of the *regularization parameter* ϵ^2 as calculated from historical time series of daily NASDAQ stock returns using the *EOF model distance functional* (9) ($\mathbf{K} = 2, n = 1106, N = 50, m = 1, T = 257$, in each case the optimization was repeated 100 times and the solution with the lowest value of *averaged clustering functional* $\mathbf{L}(\Theta, \Gamma)$ is taken in each case).

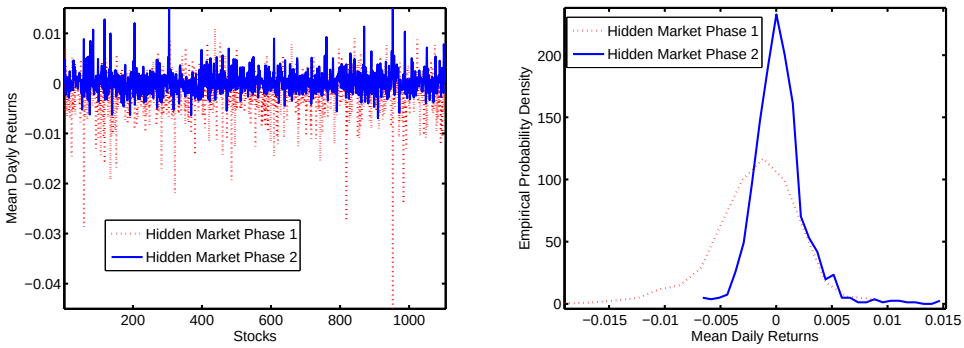

Fig. 8 Left panel: mean daily returns for 1106 NASDAQ titles calculated as *conditional averages* over the corresponding hidden market phases from the lower panel of Fig. 7. Right panel: empirical probability distributions of mean daily returns in both identified market phases.

of the hidden path with no transitions at all and higher values of (10). Therefore we can assume that the identified hidden path for $\epsilon^2 = 0.02$ is optimal wrt. the *averaged clustering functional* (10), i. e., it has a lowest value of $\mathbf{L}$ among all other identified pathways.

To interpret the resulting hidden path in terms of the global market dynamics we will first have a look at the *mean daily stock returns* [6] for each of the phases. The right panel of Fig. 8 demonstrates that the empirical probability distribution in the second market phase is narrow with a "heavy tail" in the positive direction, whereas its counterpart for the first hidden state is much wider with a heavy tail in

the negative direction. This means that the second market phase corresponds to a more stable, positive global dynamics and the first phase is characterized by much more unstable, volatile and negative dynamics. Inspection of the switches between the market phases reveals that the first transition to a market phase 1 happens on 20-Jun-2007 which approximately corresponds to the beginning of the US subprime mortgage financial crisis among the US hedge funds.

6 Conclusion

We have presented a numerical framework for the clustering of multidimensional time series based on minimization of a *regularized averaged clustering functional*. Finite element discretization of the problem allowed us to suggest a numerical algorithm based on the splitting procedure applicable for a wide class of clustering problems. We have investigated the conditions under which the proposed numerical method is monotone and analyzed the connection between the *regularization factor* and *metastabilty* in context of homogeneous Markov-jump processes.

One of the open problems is a rigorous mathematical investigation of the discretization error. It is appealing to apply the asymptotical theory developed for partial differential equations. This will allow to construct much more efficient *adaptive numerical methods* of data-clustering. Another problem is the *locality* of the proposed numerical scheme, i. e., the obtained result is dependent on the initial value. In the current implementation we solve the problem by "brute force", just repeating the optimization many times with different randomly initialized parameter values. Finally, the sparsity of the matrices involved in QP-subproblem allows to use much more efficient *sparse quadratic programming (SQP)* tools which are very suitable for parallel processing, this is also a matter of future research.

Working with multidimensional data, it is very important to be able to extract some reduced description out of it (e.g., in form of *essential degrees of freedom* or *hidden pathes*). In order to control the reliability of obtained results, one has to analyze the sensitivity of results wrt. the length of the time series and the number K of the hidden states. We have given some hints for the selection of an optimal K and explained how the quality of the resulting reduced representation can be acquired.

Acknowledgement

The author thanks O. Sander (FU Berlin) for careful reading of the manuscript and for helpful discussion. The work was supported by the DFG research center MATHEON "Mathematics for key technologies" in Berlin.

References

[1] Sutera A. Benzi R., Parisi G. and Vulpiani A. Stochastic resonance in climatic change. *Tellus*, 3:10–16, 1982.
[2] C. Nicolis. Stochastic aspects of climatic transitions-response to a periodic forcing. *Tellus*, 34:1–+, 1982.
[3] A.A. Tsonis and J.B. Elsner. Multiple attractors, fractal basins and longterm climate dynamics. *Beit. Phys. Atmos.*, 63:171–176, 1990.
[4] T.N. Palmer. A Nonlinear Dynamical Perspective on Climate Prediction. *Journal of Climate*, 12:575–591, February 1999.
[5] J.D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.
[6] R.S. Tsay. *Analysis of financial time series*. Wiley-Interscience, 2005.
[7] R. Elber and M. Karplus. Multiple conformational states of proteins: A molecular dynamics analysis of Myoglobin. *Science*, 235:318–321, 1987.
[8] H. Frauenfelder, P. J. Steinbach, and R. D. Young. Conformational relaxation in proteins. *Chem. Soc.*, 29A:145–150, 1989.
[9] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.
[10] J.A. Bilmes. *A Gentle Tutorial of the EM Algorithm and its Applications to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Thechnical Report.* International Computer Science Institute, Berkeley, 1998.

[11] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.
[12] A.H. Monahan. Nonlinear principal component analysis by neural networks: Theory and application to the lorenz system. *J. Climate*, 13:821–835, 2000.
[13] W. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, New York, 2003.
[14] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.
[15] I. Horenko, J. Schmidt-Ehrenberg, and Ch. Schütte. Set-oriented dimension reduction: Localizing Principal Component Analysis via Hidden Markov Models. In R. Glen M.R. Berthold and I. Fischer, editors, *CompLife 2006*, volume 4216 of *Lecture Notes in Bioinformatics*, pages 98–115. Springer, Berlin Heidelberg, 2006.
[16] I. Horenko, R. Klein, S. Dolaptchiev, and Ch. Schuette. Automated generation of reduced stochastic weather models i: simultaneous dimension and model reduction for time series analysis. *SIAM Mult. Mod. Sim.*, 6(4):1125–1145, 2008.
[17] I. Horenko. On simultaneous data–based dimension reduction and hidden phase identification. *To appear in J. Atmos. Sci.*, 2008. (available via biocomputing.mi.fu-berlin.de).
[18] S. Fruhwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.
[19] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, volume 35 of *Computational Mathematics*. Springer, Heidelberg, 2004.
[20] D. Braess. *Finite Elements: Theory, Fast Solvers and Applications to Solid Mechanics*. 3rd. edition.
[21] M.K. Kozlov and S.P. Tarasov L.G. Kachiyan. Polynomial solvability of convex quadratic programming. *Sov. Math. Dokl.*, 20:1108–1111, 1979.
[22] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, 1999.
[23] H. Gardiner. *Handbook of stochastical methods*. Springer, Berlin, 2000.
[24] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press, 1979.
[25] V. Papathanasiou. Some characteristic properties of the fisher information matrix via cacoullos-type inequalities. *J. Multivariate Analysis*, 14:256–265, 1993.
[26] Christof Schütte and Wilhelm Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis*, volume X, pages 699–744. Elsevier, 2003.
[27] M. Weber and P. Deuflhard. Perron cluster cluster analysis. *J. Chem. Phys.*, 5:802–827, 2003.