Sequential Change Point Detection in Molecular Dynamics Trajectories [1]

Eike Meerbach *, Christof Schütte

FU-Berlin, Arnimallee 6, 14195 Berlin, Germany.

Abstract

Motivated from a molecular dynamics context we propose a sequential change point detection algorithm for vector-valued autoregressive models based upon Bayesian model selection. The algorithm does not rely on any sampling procedure and is designed to cope with high dimensional data. We show the applicability of the algorithm on a time series obtained from simulation of a penta peptide.

Key words: VAR-model, change-point detection, fractional Bayes, molecular dynamics

1 Introduction

The macroscopic dynamics of typical biomolecular systems is often characterised by the existence of biomolecular conformations which can be understood as metastable geometrical large scale structures, i.e., molecular geometries which are on average persistent for long periods of time.
In many applications a Markovian picture is an appropriate description of this behaviour, where the effective or macroscopic dynamics is given by a Markov jump process that hops between metastable sets, representing the large scale structures, while the dynamics within these sets might be mixing on time scales that are smaller than the typical waiting time between the hops, cf. [1–5].
Biophysical research seems to indicate that these metastable sets of a typical

* Corresponding author.
Email address: meerbach@mi.fu-berlin.de. Phone.: 0049-30-83875347 (Eike Meerbach).
[1] Supported by the DFG Research Center MATHEON Mathematics for key technologies.

biomolecular system can be characterised in terms of a small number of essential degrees of freedom [6], e.g. the torsion or backbone angles of the molecule under consideration. Thus, the metastable conformations can be identified from the molecular dynamics time series projected onto these angles.
However, the problem of efficient algorithmic identification of the most persistent conformations from a given time series is still a challenging open problem. There have been several *set-oriented* approaches to this problem, which, after careful discretization of the state space into sets, are based on the analysis of the transition matrix that describes transition probabilities between these sets [3,7–10].
Recently there have also been approaches which are based on a dynamical description of the configurations by fitting local stochastic differential equations to the observed time series, a jump in configuration space then corresponds to the switching between different parameter sets [11,12].

All these approaches are based upon a global analysis of a given time series. However, due to the multiscale structure it is often difficult or even infeasible to obtain a time series from molecular dynamics which covers the macroscopic dynamics, as both system size and required simulation time is too large. Therefore approaches based upon distributed computing became more and more important. Art Voter [13,14] presented an approach, which is based upon parallel simulation of uncorrelated copies of a molecular system. However, his approach is based upon the assumption that transition between between the (not known) conformations can be detected on-line. In this article we deduce a statistical model for trajectories of molecular systems and show a change point algorithm that can be used to detect these changes on-line.

Although the matter of change point detection received considerable attention in the last years the problem is essentially still unsolved. One of the most prominent approaches to change point analysis is the CUMSUM approach, where knowledge about the distribution of the difference between prediction, assuming a model, and observation is employed to construct a rule for detecting an abrupt change in the parameterisation, e.g. [15,16]. CUMSUM methods are in many cases asymptotically well understood but effectively rely on knowing the parameter-values which are going to change. As we are confronted with high-dimensional systems and therefore with the risk of high uncertainty in parameter estimation, we have chosen a Bayesian approach, as it can naturally deal with parameter uncertainty. However, boon and bane of Bayesian methods is the need of specification of prior distributions for the parameters. In our context we can not resort to environmental studies to specify prior distributions, as in Perreault et. al. [17]. On the other hand many ways to obtain "objective" priors, like the one presented by Girón et. al. [18], are feasible only in low dimensional parameter settings. We employ the fractional Bayes approach of O'Hagan [19] and show that this can be easily adapted to our setting.

Although we avoided a sampling based approaches like particle filtering [20,21], as we consider change points in principle as rare events which makes it difficult to make prior assumptions on the distribution of change points, there is a similarity to the approach proposed by Fearnhead [21] as our algorithm relies as well on that the likelihood functions for linear models can be integrated out over the parameter space. Opposed to sampling based algorithms our algorithm can only, at least in the form stated here, handle with multiple change points in a sequential form.

In § 2 we briefly introduce the model used to describe conformation dynamics, consisting of linear stochastic differential equations (SDE's) and a jump process switching between them. In § 3 we show that discrete observations from a single linear SDE can be described by vector-valued autoregressive processes (VAR) processes, which makes it possible to reformulate the initial change point problem to a change point problem of VAR-processes by transformation of the parameter set. To prepare ground for the change point detection we also review the maximum likelihood estimators (MLE's) of a VAR model. In § 4 we formalise the change point problem as a Bayesian model selection problem and comment on general on methods to cope with vague prior distributions. In § 5 we concretize the proposed methods to our specific problem. Finally we summarise in § 6 the obtained algorithm and apply it in § 7 to an example from molecular dynamics.

2 Modelling of conformational changes by linear SDE's with switching parameters

In dealing with molecular systems one is typically faced with systems of very high dimensionality, e.g. several thousands degrees of freedom (d.o.f.). Therefore methods of model reduction are needed in the analysis of molecular systems. An important class of such *reduced models* are the Langevin models and its generalised variants. The derivation of these models is based on the existence of slow and fast time-scales in the system. While the slow d.o.f.'s are modelled by some *effective potential* function, the influence of the fast d.o.f.'s is modelled by a noise term [22]. In the easiest case the resulting *effective dynamics* is given by a first order Langevin equation

$$\dot{\boldsymbol{z}}(t) = -\nabla_{\boldsymbol{z}} U(\boldsymbol{z}(t)) + \Sigma \dot{\boldsymbol{W}}(t), \tag{1}$$

where $\boldsymbol{z} \in \mathbb{R}^d$ is the reduced system, U is some effective potential function, $\boldsymbol{W}(t)$ a d-dimensional Brownian motion simulating the influence of the unresolved variables and $\Sigma \in \mathbb{R}^{d \times d}$ a positive definite noise intensity matrix.
For the moment we will stick to the first order model given by (1) and comment later in § 3.2 on generalised models obtained by adding a memory kernel

to the noise.

Following [11,23] we linearise the non-linear stochastic differential equation (SDE) given in (1), by assuming a set of *local* linear SDE's, each of them representing the dynamics within a molecular conformation, while a switching process generates transitions between them, i.e.

$$\begin{aligned} \dot{\boldsymbol{z}}(t) &= F^{[i(t)]}\left(\boldsymbol{z} - \boldsymbol{\mu}^{[i(t)]}\right) + \Sigma^{[i(t)]}\dot{\boldsymbol{W}}(t) \\ i(t) &\in \{1, \ldots, s\}, \end{aligned} \tag{2}$$

where $\{F^i\}$, resp. $\{\boldsymbol{\mu}^i\}$, is a set of $(d \times d)$ force matrices, resp. d-dimensional mean vectors (this is equivalent to assume local quadratic potentials $U_i(z) = -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}^i)' F^i (\boldsymbol{z} - \boldsymbol{\mu}^i)$ in Eq. (1)). If the obtained reduced model shall be parameterised based on some observed time series, one encounters the problem that, unlike $\boldsymbol{z}$, the switching process i is not observable. In [12,23,11] the assumed Markovian structure of the (hidden) switching process is used in order to employ Hidden Markov Models (HMM) to estimate the parameters and transition probabilities of the system (2) via usage of the so-called Expectation-Maximisation algorithm.

Our focus is different as we are interested in on-line analysis of a time series, i.e. we want to detect transitions from one regime to another one, e.g. $i(t) = j$ for $t_0 \leq t < t_1$ and $i(t) = k$ for $t \geq t_1$, while observing the time series sequentially. We will call a time point where such a change in parameterisation occurs, here t_1, a *change point*.
However, in § 7.3 we show that if all change points are successfully identified over a certain time interval, we can use the obtained information as well to estimate transition probabilities of the process i. But before bothering with the change point detection we have to elaborate on parameter estimation of a single linear SDE in the next section.

3 Parameter estimation of a linear SDE

3.1 Maximum likelihood estimators

A natural way for parameter estimation of a d-dimensional linear SDE

$$\dot{\boldsymbol{z}} = F(\boldsymbol{z} - \boldsymbol{\mu}) + \Sigma\dot{\boldsymbol{W}}$$

based upon a series of observations $Z = \{\boldsymbol{z}_t\}, t \in \{1, \ldots, T\}$, at equidistant time points, i.e. $\boldsymbol{z}_t := \boldsymbol{z}((t-1)\tau)$, with time step τ, is to investigate an appropriate likelihood function [11]. It is well known that for a linear SDE with fixed

initial conditions the solution is a Markov process and any time discretization of the solution is multivariate normal distributed [24]. In particular, given an observation $\boldsymbol{z}_t$, the conditional probability density of $\boldsymbol{z}_{t+1}$ is a Gaussian with density function

$$f_\lambda(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t) = \frac{1}{\sqrt{|2\pi R(\tau)|}} \exp\left(-\frac{1}{2}(\boldsymbol{z}_{t+1}-\boldsymbol{\mu}_t)^T R(\tau)^{-1}(\boldsymbol{z}_{t+1}-\boldsymbol{\mu}_t)\right), \quad (3)$$

where $|\cdot|$ denotes the matrix determinant and the mean, resp. variance, of the distribution is given by $\boldsymbol{\mu}_t := \boldsymbol{\mu} + \exp(\tau F)(\boldsymbol{z}_t - \boldsymbol{\mu})$, resp. $R(\tau) := \int_0^\tau \exp(sF)\Sigma\Sigma'\exp(sF')ds$ (depending on the argument $\exp(\cdot)$ denotes a scalar or a matrix exponential function). The dependence on the parameter set is marked by $\lambda = (\boldsymbol{\mu}, F, \Sigma)$. Therefore, a likelihood function can be constructed as

$$L(\lambda|Z) = \prod_{t=1}^{T-1} f_\lambda(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t).$$

Unfortunately there is no known analytic solution to the maximisation problem of L w.r.t the parameter set $(\boldsymbol{\mu}, F, \Sigma)$. Another drawback, from a statistical viewpoint, is that L is not integrable over the parameter space (e.g. if we set $F = 0$ (for $d = 1$) integrating over μ is not possible). Therefore we can not obtain a density in parameter space from the likelihood function. A resort is given by rewriting (3) to

$$\begin{aligned} \boldsymbol{z}_{t+1} &= \mathcal{N}(\boldsymbol{\mu} + \exp(\tau F)(\boldsymbol{z}_t - \boldsymbol{\mu}), R) \\ &= (I - \exp(\tau F))\boldsymbol{\mu} + \exp(\tau F)\boldsymbol{z}_t + \mathcal{N}(\boldsymbol{0}, R), \end{aligned} \quad (4)$$

where $\mathcal{N}$ is a multivariate normal distributed random variable and I an identity matrix of appropriate size. Eq. (4) reveals the autoregressive structure of order one, VAR(1), of the time series of discrete observations. Defining

$$\begin{aligned} \Phi &:= \Big((I - \exp(\tau F))\boldsymbol{\mu}\ \exp(\tau F)\Big) && \in \mathbb{R}^{d\times(d+1)} \\ X &:= \begin{pmatrix} 1 & \dots & 1 \\ \boldsymbol{z}_1 & \dots & \boldsymbol{z}_{T-1} \end{pmatrix} && \in \mathbb{R}^{(d+1)\times(T-1)} \\ Y &:= \Big(\boldsymbol{z}_2, \dots, \boldsymbol{z}_T\Big) && \in \mathbb{R}^{d\times(T-1)} \\ \epsilon &:= \Big(\mathcal{N}(\boldsymbol{0}, R), \dots, \mathcal{N}(\boldsymbol{0}, R)\Big) && \in \mathbb{R}^{d\times(T-1)}, \end{aligned}$$

allows to write (4) in a compact form

$$Y = \Phi X + \epsilon.$$

Transforming the parameter set λ to $\tilde{\lambda} = (\Phi, R)$, leads to a likelihood function

$$L(\tilde{\lambda}|Z) = \left(\frac{1}{\sqrt{|2\pi R|}}\right)^{(T-1)} \exp\left(-\frac{1}{2}\,\mathrm{tr}((Y - \Phi X)(Y - \Phi X)' R^{-1})\right), \qquad (5)$$

for which there are analytic MLE's $\hat{\Phi}$ and $\hat{R}$ given by [25,26]

$$\hat{\Phi} = YX'(XX')^{-1} \text{ and } \hat{R} = (Y - \hat{\Phi}X)(Y - \hat{\Phi}X)'/(T-1). \qquad (6)$$

Therefore transforming the parameter set to $\tilde{\lambda}$ has the advantages that (i) the distribution of the discrete observations is fully characterised by $\tilde{\lambda}$, (ii) analytical MLE's are available and (iii) the likelihood function is integrable over the parameter space (c.f 8).

3.2 Higher order VAR processes to include memory effects

Considering the discrete observations of a linear SDE as realizations of a VAR(1) process naturally raises the question if there is a consistent interpretation of using a higher order model, e.g. VAR(p), that is

$$\boldsymbol{z}_{t+1} = A_0\boldsymbol{\mu} + \sum_{i=1}^{p} A_i \boldsymbol{z}_{t-i+1} + \mathcal{N}(\mathbf{0}, R). \qquad (7)$$

This process is obviously not anymore a Markov process but exhibits a memory lag of p steps into the past. In fact Eq. (7) can be interpreted as a time discretization of a generalised Langevin process

$$\dot{\boldsymbol{z}}(t) = -\nabla_{\boldsymbol{z}} U(\boldsymbol{z}(t)) - \int_0^t \gamma(t-s)\boldsymbol{z}(s)ds + \Sigma\dot{\boldsymbol{W}}(t), \qquad (8)$$

under the assumption of a quadratic potential function, as above, and a piecewise constant memory kernel γ with finite support. For a more detailed presentation we refer to [27].
If a fixed order p is assumed estimation of the parameters of a VAR(p) is analogue to that of the VAR(1) process, we only have to extend the definitions of the data matrices X and Y to

$$X := \begin{pmatrix} 1 & \dots & 1 \\ \boldsymbol{z}_1 & \dots & \boldsymbol{z}_{T-p} \\ \vdots & & \vdots \\ \boldsymbol{z}_p & \dots & \boldsymbol{z}_{T-1} \end{pmatrix} \qquad \in \mathbb{R}^{(dp+1)\times(T-p)}$$

$$Y := \begin{pmatrix} \boldsymbol{z}_{p+1}, \dots, \boldsymbol{z}_T \end{pmatrix} \qquad \in \mathbb{R}^{d\times(T-p)}.$$

The estimator $\hat{\Phi}$ in (6) now estimates

$$\Phi = \Big(A_0 \boldsymbol{\mu} \; A_1 \; A_2 \; \dots \; A_p \Big) \in \mathbb{R}^{d \times (dp+1)}.$$

3.3 Estimation of the model order

There are several criteria and tests to estimate the order p of a VAR process from a given time series, for a discussion of these we refer to [25], Chapter 4.3. In our application example we use the Schwarz criterion which chooses the order p so that the function

$$SC(p) = \log |\hat{R}(p)| + \frac{\log T}{T} p d^2, \tag{9}$$

where $\hat{R}(p)$ is the MLE of R under the assumption of a VAR(p) process, is minimised within a predefined range $0, \dots, p_{\max}$. The first term of the Schwarz criterion minimises the noise term in the model while the second penalises the number of estimated parameters which grows with a higher order model. It can be shown that the Schwarz criterion is a consistent estimator of the order of a VAR process. We are aware, that, unfortunately, the Schwarz criterion is often not optimal in finite sample situations. We do not want to elaborate on this subject, but only remark that as the order of the VAR process here is essentially the memory depth of the molecular process, which might be known or estimated by methods specific in the MD context.

3.4 Computation of the estimators

The analytic estimators given in (6) are in general not used for computation of the parameters, as the matrix inversion can be unstable. Instead one can use the moment matrix

$$\begin{aligned} M = M(Z) :&= \sum_{i=1}^{T} \Big(1 \; \boldsymbol{z}_i' \; \dots \; \boldsymbol{z}_{i+p}' \Big) \begin{pmatrix} 1 \\ \boldsymbol{z}_i \\ \vdots \\ \boldsymbol{z}_{i+p} \end{pmatrix} \\ &= \begin{pmatrix} XX' & XY' \\ YX' & YY' \end{pmatrix} =: \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}. \end{aligned} \tag{10}$$

The moment matrix is an important object as it contains all statistical relevant information about the observed process (under the assumption of a VAR(p)

process). This can be seen by rewriting the likelihood function in terms of the blocks in M:

$$L(\Phi, R|Z) = L(\Phi, R|M) = \left(\frac{1}{\sqrt{|2\pi R|}}\right)^m \tag{11}$$

$$\cdot \exp\left(-\frac{1}{2}\operatorname{tr}((M_{22} - M_{21}\Phi' - \Phi M_{12} + \Phi M_{22}\Phi')R^{-1})\right), \tag{12}$$

where m denotes the upper left scalar entry of M which equals $T - p$, i.e. the length of the observed time series minus p initial points. We will employ this notation $m = m(Z)$ below to avoid the indices for the length of different time series. Also we employ subsequently the notation $f(Z|\Phi, R) \equiv L(\Phi, R|Z)$ if we want to highlight Eq. 11 as a density in data space. The MLE's can be obtained from the moment matrix M in a stable way via a Cholesky factorisation which gives an upper triangular matrix

$$R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix},$$

such that

$$M = \begin{pmatrix} XX' & XY' \\ YX' & YY' \end{pmatrix} = \begin{pmatrix} R_{11}'R_{11} & R_{11}'R_{12} \\ R_{12}'R_{11} & R_{12}'R_{12} + R_{22}'R_{22} \end{pmatrix} = R'R.$$

Plugging the Cholesky factorisation into the estimators (6) one obtains

$$\hat{\Phi} = (R_{11}^{-1}R_{12})' \text{ and } \hat{R} = \frac{1}{m}R_{22}'R_{22}. \tag{13}$$

In the case of an ill-conditioned moment matrix M one can add a regularization matrix to ensure a well-posed Cholesky factorisation. A possible choice is to use $M + \delta\text{diag}(M)$ instead of M, with a small parameter δ depending on the dimensionality of the problem and the machine precision, cf. [28].

4 Bayesian model comparison

4.1 Change point models

To detect changes in the parameterisation of the underlying VAR model we utilise a Bayesian approach like the one presented in [17] for the univariate

and independent distributed case. Assume that we have observed a sequence of observations

$$Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots \boldsymbol{z}_T\}, \boldsymbol{z}_i \in \mathbb{R}^d$$

for which we presume a VAR(p) model as the data generating mechanism (DGM). Our aim is to test within a window from t_1 to t_2 for a change point in the parameterisation. We do not test from 1 to T as we need at least $(d+1)p + d + 1$ observations in one dynamical regime, as becomes apparent later, i.e. one could set $t_1 = (d+1)p + d + 2$ and $t_2 = T - (d+1)p - d - 2$. In practice, however, all our experience indicates that it is highly advisable to increase t_1.

Thus we have $n := t_2 - t_1 + 1$ candidate change points giving rise to $n + 1$ models $H_i, 0 \leq i \leq n$, where

$$H_i := \begin{cases} Z \text{is generated by only one VAR}(p)\text{-DGM}, & \text{for } i = 0. \\ Z_1 = Z_1(i) = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_{t_1+i-1}\} \text{ and} \\ Z_2 = Z_2(i) = \{\boldsymbol{z}_{t_1+i-p}, \dots, \boldsymbol{z}_T\} \\ \text{are generated by distinct VAR}(p)\text{-DGM's.} & \text{for } 1 \leq i \leq n. \end{cases} \tag{14}$$

Note that the Z_1 and Z_2 are overlapping, as the last p points of Z_1 are used as initial conditions for Z_2 (of course one has to remove the overlap if Z_1 and Z_2 are not directly subsequent, which does not affect any of the subsequent considerations).

The probability of each model given the observations Z can be computed via the Bayes formula

$$\mathbb{P}[H_i|Z] = \frac{\mathbb{P}[Z|H_i]\,\mathbb{P}[H_i]}{\sum_{j=0}^{n} \mathbb{P}[Z|H_j]\,\mathbb{P}[H_j]}, \tag{15}$$

where

$$\mathbb{P}[Z|H_i] = \begin{cases} \int f(Z|\Phi_1, R_1)\pi_1(\Phi_1, R_1)d\Phi_1 dR_1 & \text{if } i = 0\ , \\ \prod_{i=1,2} \int f(Z_i|\Phi_i, R_i)\pi_i(\Phi_i, R_i)d\Phi_i dR_i & \text{if } i > 0, \end{cases}$$

with prior distributions π_1 and π_2 on the parameters. Having 15 we can easily evaluate the probability of a change point as:

$$\mathbb{P}[\text{change}|Z_2] = \frac{\sum_{i=1}^{n} \mathbb{P}[Z_2|H_i]\,\mathbb{P}[H_i]}{\sum_{j=0}^{n} \mathbb{P}[Z_2|H_j]\,\mathbb{P}[H_j]}. \tag{16}$$

But to evaluate these probabilities we obviously have to specify the prior probabilities for the models, i.e. $\mathbb{P}[H_i]$, and the parameters, i.e. π_1 and π_2,

and, of course, evaluate the above integrals. The prior distributions on the parameters will be discussed in detail in § 4.3. The (analytic) solution of the integrals will be given in § 5 and § 8. But before we introduce Bayes factors to reformulate the problem as suggested in [18], so that "objective" Bayes factors can be used to evaluate (16) under missing prior information.

4.2 Bayesian factors

A natural way to compare the posterior probabilities of two distinct models H_i and H_j is the ratio

$$\frac{\mathbb{P}[H_i|Z]}{\mathbb{P}[H_j|Z]} = \frac{\mathbb{P}[Z|H_i]\,\mathbb{P}[H_i]}{\mathbb{P}[Z|H_j]\,\mathbb{P}[H_j]}, \tag{17}$$

called posterior odds. A high ratio mean that model H_i is more probable in the light of data Z than H_j. The Bayes factor B_{ij} is defined as

$$B_{ij} = \frac{\mathbb{P}[Z|H_i]}{\mathbb{P}[Z|H_j]}. \tag{18}$$

Eq. (17) reveals the meaning of the Bayes factor: it defines how the data Z transforms the prior odds $\mathbb{P}[H_i]/\,\mathbb{P}[H_j]$ to the posterior odds, i.e. in which direction the data shifts our prior beliefs. The Bayes factor approach is similar to the likelihood ratio statistic, cf. [29,30], but while the likelihood ratio is obtained via maximisation over the parameter space the Bayes factor is obtained by integration over the parameter space. Division of (16) by $\mathbb{P}[Z_2|H_0]$ yields

$$\mathbb{P}[\text{change}|Z_2] = \frac{\sum_{i=1}^{n} B_{i0}\,\mathbb{P}[H_i]}{\sum_{j=0}^{n} B_{j0}\,\mathbb{P}[H_j]}.$$

which can be interpreted as an assembly of a sequence of test against the null hypothesis of no change.

4.3 Prior distributions

A natural choice to code our ignorance on a parameter change before observing data is to assign a prior probability of $\frac{1}{2}$ to the event of a change and distribute the rest probability among the other models, i.e.

$$\mathbb{P}[H_0] = \frac{1}{2},\ \mathbb{P}[H_i] = \frac{1}{2n}, 1 \le i \le n.$$

More problematic is the choice of prior distributions for the parameters of the VAR models under ignorance. A common choice is the usage of the *diffusive prior*, which consist of a flat prior on Φ and a Jeffrey's prior on R, so that

$$\pi_D(\Phi, R) \propto |R|^{-\frac{d+1}{2}}, \tag{19}$$

a discussion of this prior and other possibilities is given in, e.g. [26,31]. Although it can be easily shown that under the diffusive prior the posterior distribution (4.1) is a proper, i.e. normalizable, distribution, the choice is problematic for model comparison, as this prior is only defined upon a constant and therefore the model probabilities (15) as well as the Bayes factor (18) are also only defined upon a constant. The constant does not cancel out of the fraction in (15) and (18) as we have parameters which are not common to all models, i.e. the parameters for the VAR model after a change has occurred. This means, that with the use of an unproper prior we can compare different change point models, as the indeterminate constants do cancel out, but we can not compare the probability between change and no-change.

A possible way out of this dilemma is the usage of objective Bayesian factors, which we shortly introduce in the next paragraphs before adapting them to our specific situation in § 5.

4.3.1 Partial Bayes

A way to obtain a proper prior distribution for some parameter θ despite of ignorance is to split the data $Z = \{Z_p, Z_{-p}\}$ and use a part (Z_p) as a training set to specify the prior while the other part (Z_{-p}) is used for testing or analysis, i.e.

$$\pi_{PB}(\theta) \propto \pi_D(\theta) f(Z_p|\theta),$$

where $\pi_D(\theta)$ denotes an improper parameter prior. The size of the training set is usually taken as the minimal size to guarantee properness of the resulting prior. A problem is the arbitrariness in the choice of data points taken into the training sample. Proposals to overcome this arbitrariness are, e.g., averaging over all possible minimal training sets resulting in the *intrinsic Bayes* approach [32]. The intrinsic Bayes approach can be elegantly expanded if nested models are tested, [18,33], but has the drawback that computation of intrinsic Bayes factors is often hard, resp. feasible only for certain models.

4.3.2 Fractional Bayes

The fractional Bayes approach, put forward by O'Hagan [19], is based on the idea not to use part of the data to specify a prior but to use a fraction of the

likelihood function, i.e.

$$\pi_{FB}(\theta) \propto \pi_D(\theta) f^b(Z|\theta), \quad b \in]0,1[.$$

The likelihood function is then transformed to $L(\theta|Z) = f^{(1-b)}(Z|\theta)$, thus becoming flatter as a fraction of the information was already used to define the prior distribution. The question of the right choice of a training set is elegantly avoided, as a fraction of *all* data is used. A reasonable choice of b is again the minimal value to guarantee properness of the resulting prior.

4.3.3 Imaginary minimal experiment

Another approach presented by Spiegelhalter and Smith [34] is the use of a so called imaginary minimal experiment. Suppose there are two models to be compared and in at least one of them there is a parameter for which we can only specify an improper non-informative prior. Then the resulting Bayes factor is given by

$$B_{01} = c \cdot \frac{\int f_1(Z|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_2(Z|\theta_2)\pi_2(\theta_2)d\theta_2},$$

with an unknown constant c. The idea of an imaginary minimal experiment is to fix the undetermined constant c by imagination of a data set Z_I which is just big enough to discriminate between the two models, therefore minimal, but gives maximal support for one of the two models. The reasoning then is that the Bayes factor should favour the supported model but only minimal, due to the smallness of the data set, so that

$$B_{01} \approx 1 \Rightarrow c \approx \frac{\int f_2(Z_I|\theta_2)\pi_2(\theta_2)d\theta_2}{\int f_1(Z_I|\theta_1)\pi_1(\theta_1)d\theta_1}.$$

It has been argued that the definition of an imaginary minimal experiment is sufficient only in rather special cases [19]. Furthermore it is not clear that the claim $B_{01} \approx 1$ is an appropriate choice in all cases. But, as we will show, in the change point detection framework as presented, the imaginary minimal approach seems to be good applicable.

5 Identification of a change point

The key ingredient to employ the approaches stated above in the computation of objective Bayes factors $B^{(i)} := B_{0i}, 1 \leq i \leq n$ is that our model allows analytical integration of the likelihood function over paramer space. Assume for the moment an arbitrary time series Z of length T and the corresponding moment matrix $M = M(Z)$. As M contains all statistical relevant information of the data we can write $f(M|\Phi, R)$ instead of $f(Z|\Phi, R)$, as given in

Eq. (11). Following the notation introduced in §3.4 we denote by R_{11} and R_{12} the corresponding blocks of the triangular matrix R obtained from the Cholesky factorisation of M and by $m = T - p$ the upper left scalar entry of M. Then, see Appendix 8,

$$\begin{aligned} I[M] :=& \int f(M|\Phi, R)\pi_D(\Phi, R)d\Phi dR \\ =& \pi^{\frac{d(d-1)}{4}} |R_{11}|^{-d} |\sqrt{\pi} R_{22}|^{-(m-dp-1)} \prod_{j=1}^{d} \Gamma\left(\frac{m-dp-j}{2}\right), \end{aligned} \tag{20}$$

as long as $m > dp + d$ (therefore we need at least $dp + d + p + 1$ subsequent points before and after a change point). Furthermore note from Eq. (11) that information coming from different time series (parts), e.g. Z_1 and Z_2 can be combined just by adding the moment matrices, as

$$f(M(Z_1)|\Phi, R)f(M(Z_2)|\Phi, R) = f(M(Z_1) + M(Z_2)|\Phi, R). \tag{21}$$

Now assume a fixed $i > 0$ and Z_1, Z_2 defined as in 14. In the next subsections we construct the corresponding fractional Bayes and imaginary experiment factors.

5.1 Fractional Bayes

The fractional Bayes approach can be easily implemented by noting from (11) that

$$\begin{aligned} & f^b(M|\Phi, R) \\ &= \left(\frac{1}{\sqrt{|2\pi R|}}\right)^{bm} \exp\left(-\frac{1}{2}\operatorname{tr}((bM_{22} - bM_{21}\Phi' - \Phi bM_{12} + \Phi bM_{22}\Phi')R^{-1})\right) \\ &= f(bM|\Phi, R), \end{aligned} \tag{22}$$

so that, using the notation introduced above, we have

$$\int f^b(Z_2|\Phi, R)\pi_D(\Phi, R)d\Phi dR = I[bM(Z_2)],$$

and, using (21),

$$\int f(Z_1|\Phi, R)f^{(1-b)}(Z_2|\Phi, R)\pi_D(\Phi, R)d\Phi dR = I[M(Z_1) + (1-b)M(Z_2)].$$

Plugging it all together we obtain in short notation the fractional Bayes factor

$$B^{(i)}_{FB} = \frac{I[M(Z_1)]I[M(Z_2)]}{I[M(Z_1) + (1-b)M(Z_2)]I[bM(Z_2)]}. \tag{23}$$

The minimal value of b is determined by the minimal value for which

$$I[bM(Z_2)]$$

is defined (cf. §8). Therefore the minimal value of b is given by

$$b_{min} = \frac{dp + d + 1}{m(Z_2)},$$

which means that the upper left entry of $bM(Z_2)$ just meets the threshold of $dp + d + 1$.

5.2 Imaginary minimal experiment

To employ the Spiegelhalter/Smith approach we have to define an adequate imaginary minimal experiment Z_I. If we want to decide if Z_2 is generated by the same DGM as Z_1 we need, as stated above, a minimum of $dp + d + 1 + p$ observations, otherwise the integrals in the Bayes factor are not defined anymore. Maximal support for the "no change"-model would be the same observed statistic in both observed time series, i.e.

$$\frac{M(Z_1)}{m(Z_1)} = \frac{M(Z_I)}{m(Z_I)} \;\Leftrightarrow\; M(Z_I) = \frac{dp + d + 1}{m(Z_1)} M(Z_1).$$

With this definition of $M(Z_I)$ we can fix the undetermined constant in the Bayes factor as

$$c_I = \frac{I[M(Z_1) + M(Z_I)]}{I[M(Z_1))]I[M(Z_I)]},$$

and obtain the Bayes factor

$$B_I^{(i)} = c_I \cdot \frac{I[M(Z_1)]I[M(Z_2)]}{I[M(Z_1) + M(Z_2)]}.$$

In our experience both suggested procedures are feasible, but the fractional Bayes Factor approach seems to be less sensitive against outliers in the data and deviation from the model. So we use the fractional Bayes approach but remark that in the case of little data (where we do not want to flatten the density further) the imaginary minimal experiment approach can be seen as an alternative.

6 Algorithmic procedure

To clarify our approach we summarise the obtained procedure in a flow chart (next page). Besides the time series access three parameters are needed, the length of an initial window t_I, the length of the update window t_U and the order of the VAR process p (alternatively a maximal order $p_{\max}$ is needed to test the order). In practice it is advisable to choose a larger t_I as the minimal one, as otherwise, if a change close to t_I is tested, the likelihood induced parameter distribution will be so different, even if there is no change point, in the two time series segments that it is likely to detect a false one. Furthermore, having real data where the VAR model is just an approximation of the DGM choosing a higher t_I makes the change point detection less sensitive to deviations from the model, which is a crucial point in practice. With increasing length of the update window t_U the algorithm will become less sensitive against short time departures from the model. In the algorithmic scheme we forbid change points at the end of the test interval to provide the opportunity to detect a better change point in the next available sequence.

Algorithm 1: Sequential change point detection

Parameter: t_I (length of initial window)
: t_U (length of update window)
: p (VAR-order) or $p_{\max}$ (maximal VAR-order)
Input : A time series $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots\}$ with sequential access

$t_0 \leftarrow 1$
$t_1 \leftarrow t_I + t_U$
$Z \leftarrow \{\mathbf{z}_{t_0}, \ldots, \mathbf{z}_{t_1}\}$
If p is not given estimate $p \in \{0, 1, \ldots, p_{\max}\}$ based on $\{\mathbf{z}_{t_0}, \ldots, \mathbf{z}_{t_0+t_I-1}\}$.
while *incoming data is available* **do**
  Compute Bayes factors:
  for $k \leftarrow t_I + 1$ **to** $t_1 - ((d+1)p + d + 1)$ **do**
    $Z_1 \leftarrow \{\mathbf{z}_{t_0}, \ldots, \mathbf{z}_{t_0+k-1}\}$
    $Z_2 \leftarrow \{\mathbf{z}_{t_0+k-p}, \ldots, \mathbf{z}_{t_1}\}$
    $b \leftarrow \frac{dp+d+1}{m(Z_2)}$
    $B_{FB}^{(k-t_I)} \leftarrow \frac{I[M(Z_1)]I[M(Z_2)]}{I[M(Z_1)+(1-b)M(Z_2)]I[bM(Z_2)]}$
  $n \leftarrow t_1 - t_I - ((d+1)p + d + 1)$
  $\mathbb{P}[\text{change}] = \frac{\sum_{i=1}^{n} B_{FB}^{(i)}}{\sum_{j=1}^{n} B_{FB}^{(j)} + n}$
  $c = \text{argmax}_{1 \le i \le n} B_{FB}^{(i)}$
  if $\mathbb{P}[\textit{change}] > 0.5$ **and** $c \neq n$ **then**
    $\mathsf{changepoints} \leftarrow \{\mathsf{changepoints}, t_0 + t_I + c\}$
    $t_0 \leftarrow t_0 + t_I + c - p$
    $t_1 \leftarrow t_0 + t_I + t_U - 1$
    $Z \leftarrow \{\mathbf{z}_{t_0}, \ldots, \mathbf{z}_{t_1}\}$
    If p is not given estimate $p \in \{0, 1, \ldots, p_{\max}\}$ based on $\{\mathbf{z}_{t_0}, \ldots, \mathbf{z}_{t_0+t_I-1}\}$.
  else
    $t_1 \leftarrow t_1 + t_U$

Output: changepoints

7 Example

In this section we demonstrate the applicability of the precedingly presented algorithm by an example from molecular dynamics (MD). As an example we will use simulation data of an artificial penta-peptide, consisting of a capped chain of five amino-acids: Glutamine-Alanine-Phenylalanine-Alanine-Argenine, shown in Fig. 1. The peptide is itself an interesting object to study, as it is a small molecule which is able to form salt bridges, an important and still not well understood matter. We will not concern with this subject but rather use a trajectory of the peptide for demonstration purpose of our algorithm only. The trajectory was obtained from an MD-simulation in vacuum using the NWChem software package [35,36]. The integration time step was set to 1 femtosecond, while the coordinates were written out every 200 femtoseconds. The trajectory we use consists of 100000 points thus covers a length of 20 nanoseconds. What can be seen in the trajectory is the folding of the

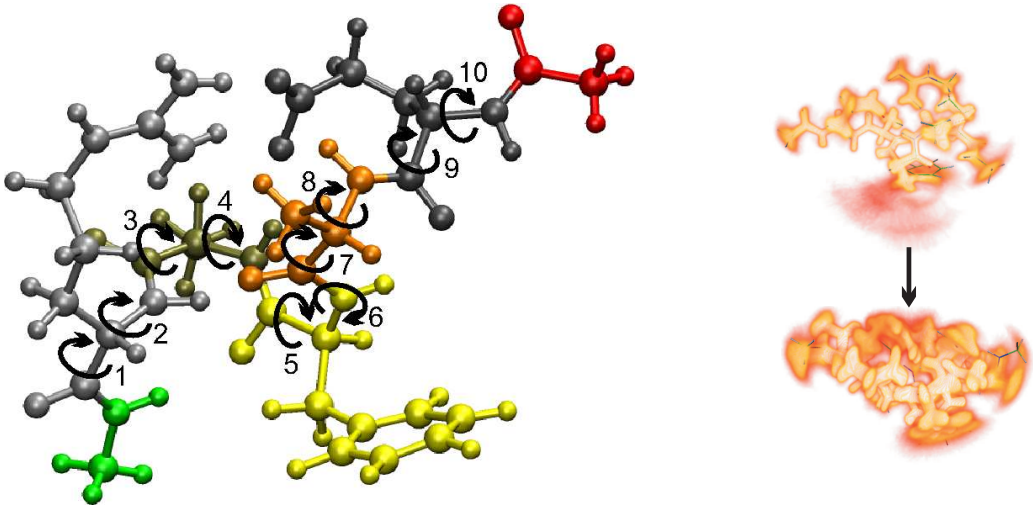

Fig. 1. *Left:* The simulated penta-peptide with the 10 observed torsional backbone angles marked. *Right:* During the simulation the molecule transforms from a structure where mainly the side chains interact to a more compact and stable structure via several meta stable intermediates. The obtained metastable structures from beginning and end of the trajectory are visualised by density plots showing the flexibility within a conformation (Visualisation by AMIRA, [37]).

peptide from a spread out structure where only the two long side chains interact (the salt bridge) to a more compact structure and very stable structure, see Fig. 1.

7.1 Observables and removing periodicity

A way to avoid difficulties with the free translational and rotational modes of the positional coordinates of a molecule is to switch to internal coordinates. In general, the overall geometric structure of a peptide can be characterised by the torsion angles along the backbone, excluding the rigid N-H-C-O peptide bonds. For demonstration purpose we omit at this point the torsional angles along the side chains, as it only makes the picture more complicated. Therefore we are left with a 10-dimensional torsion angle time series from the backbone as seen in Fig. 1.
Obviously we have to take care of the periodic nature of the torsion angles. But as the algorithm expects the data to come in sequentially we can just shift the data piecewise to remove periodicity, which will work in most cases as the torsion angles are in general not free rotating. The shifting of the data can be automatised by discretizing the angle domain determine a borderline with minimal number of transitions across. Additionally we can exclude transition to data points that cross the periodic boundary from our statistics, cf. Fig. 2.

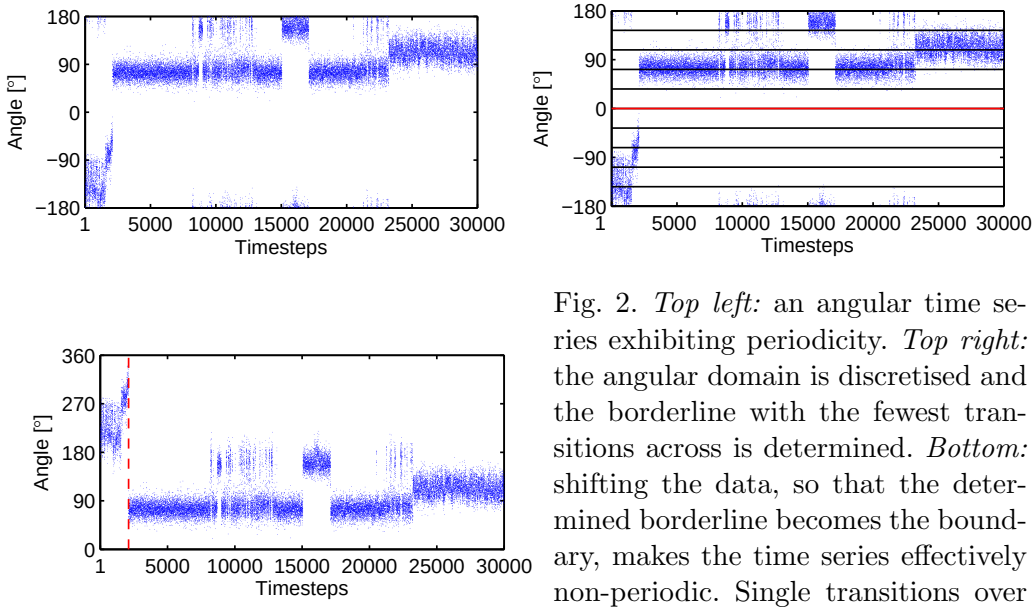

Fig. 2. *Top left:* an angular time series exhibiting periodicity. *Top right:* the angular domain is discretised and the borderline with the fewest transitions across is determined. *Bottom:* shifting the data, so that the determined borderline becomes the boundary, makes the time series effectively non-periodic. Single transitions over the boundary (dotted line) are just excluded.

7.2 Results of change-point analysis on the trajectories

We applied the algorithm to the trajectory with the parameters set to $t_I = 500$, $t_U = 300$, , $p_{\max} = 8$. Note that even if the initial window t_I appears to be chosen rather large, which makes the algorithm less sensitive to outliers and allows better order estimation in the time series segments, it is still small compared to the length of the time series. The choice of $p_{\max}$ is not very crucial as the order along the time series is mostly estimated as $\hat{p} = 1$ or maximal $\hat{p} = 2$.

The outcome is depicted in Fig. 3. All obvious change points are detected. Between about 200000 and 500000 fs there are a few change points which seem to be spurios. In fact the distribution of detected change points in this region is sensitive to parameter changes indicating the existence of outliers (in a dynamic sense). In the next section we show how we can reassembly these segments.

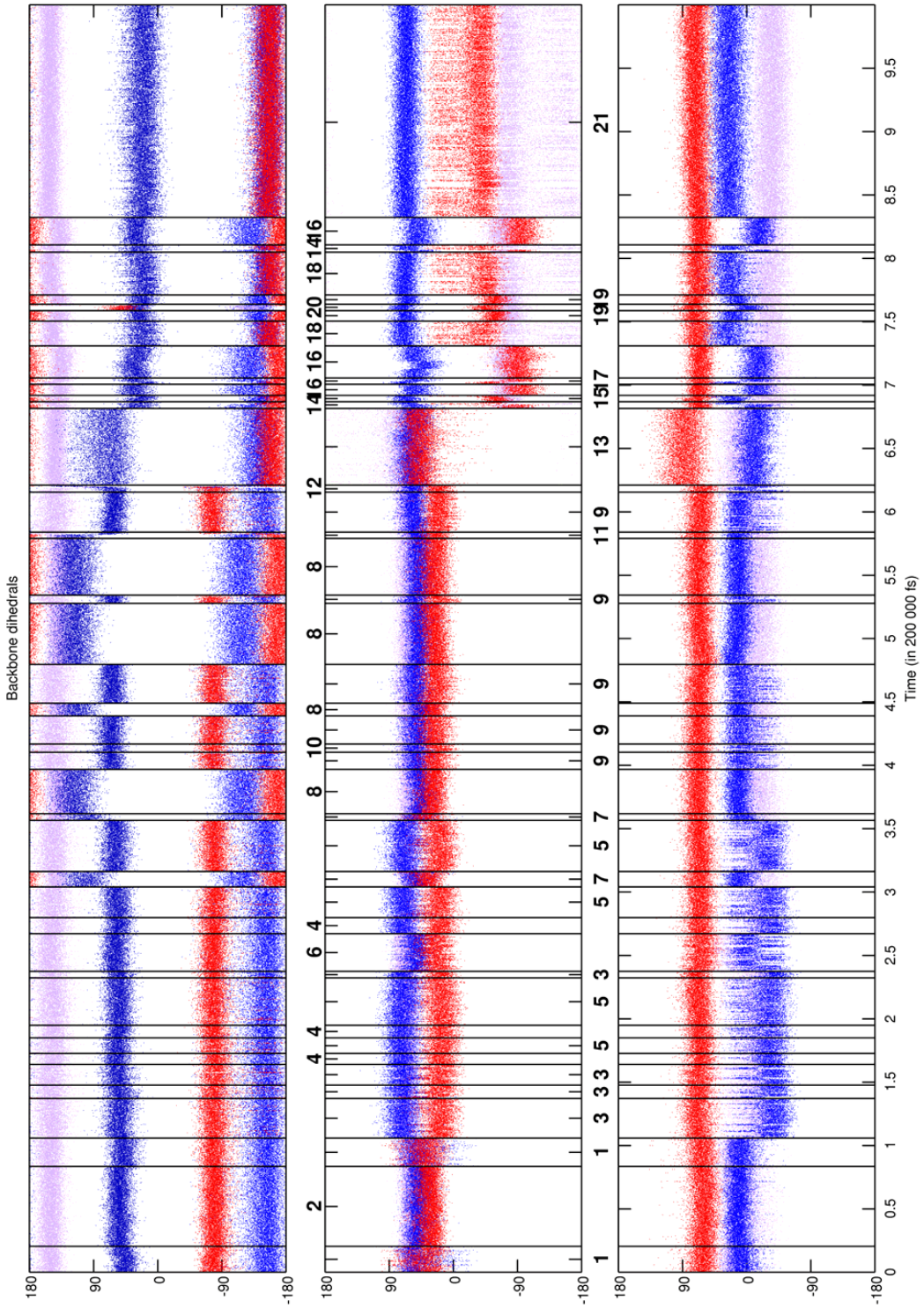

Fig. 3. The 10-dimensional backbone torsion angle time series of the peptide (splitted in 3 subpanels, Top: 1-4, Middle: 5-7: Bottom: 8-10). The vertical lines mark the detected change points. The digits 1 to 20 on the margins of the middle panel indicate the membership of the obtained time series segments to the 21 clusters obtained from hierarchical clustering as explained below.

After having obtained the change points sequentially we can use the outcome of the change point algorithm also to group the data globally. Therefore we define a distance matrix D, measuring the distance between all identified segments $\{S_1, S_2, \ldots, S_N\}$ of the time series, $N = 43$ in our case, according to the probability that the segments are generated by the same DGM:

$$D_{ij} = \frac{B_{ij}}{B_{ij}+1}$$

$$B_{ij} = \begin{cases} \dfrac{I[M(S_i)]I[M(S_j)]}{I[M(S_i)+(1-b_{\min})M(S_j)]I[b_{\min}M(S_j)]}, \text{ if } m(S_i) \leq m(S_j), \\ \dfrac{I[M(S_i)]I[M(S_j)]}{I[M(S_j)+(1-b_{\min})M(S_i)]I[b_{\min}M(S_i)]}, \text{ if } m(S_i) > m(S_j), \end{cases}$$

with $1 \leq i, j \leq N$. So the distance is just the probability of a change point, where the change point has to be between the two segments. To make the distance matrix symmetric and to avoid waste information from the shorter segment we always use the longer segment to extract prior information about the parameters. The obtained distance matrix can be used to cluster the data, i.e. to merge different time series segments, e.g. by an hierarchical clustering algorithm [38]. Therefore the distance between two clusters C_1 and C_2 is given by the maximal distance between any member of one cluster to any member of the other cluster:

$$d(C_1, C_2) = \max_{S_i \in C_1, S_j \in C_2} D_{ij},$$

The hierarchical structure appears by raising the maximal distance $d_{\max}$ allowed for objects within a cluster gradually. If $d_{\max} = 0$ all segments $S_1, \ldots, S_N$ define their own cluster. By raising $d_{\max}$ eventually two segments are allowed to form a cluster, further on other segments may join the cluster or define their own cluster or two clusters may merge to a single cluster. In our case a natural choice as the allowed maximal distance is $d_{\max} = 0.5$, which was our testing criteria to separate the time series segments. In fact our example is not very challenging for the cluster algorithm as the segments can just be grouped in 21 clusters with all elements within a cluster having a distance of virtually zero to each other and a distance of one to elements outside the cluster. The obtained global grouping is marked in Fig. 3. Again visual expection confirms that the results are meaningful in that similar segments of the time series are indeed grouped together. It can also happen that adjacent segments will be recombined again (as happens in the above mentioned interval from 200000 fs to 500000 fs), this is due to a better statistical base to make our decisions as we now look from a global view on the obtained data.

8 Conclusion

Motivated by the task to detect conformational changes in biomolecules online from time series, we showed how to paraphrase the problem into a change point detection problem for VAR models. We tackled the problem by employing a Bayesian approach to model selection. As we assumed having no prior knowledge about parameters of change points we avoided the need to specify proper prior distributions by usage of a fractional Bayes approach, which we formulated in a way to deal with our high dimensional parameter space. We finally achieved an algorithmic procedure which is very clear and easy to implement. In the last section we demonstrate the applicability of the procedure to "real" data from molecular dynamics simulations, with promising results. Even though one has to admit that the approach crucially depends on identifying the right observables, i.e. observables which are locally linear, which is in practice not always trivial task.

Acknowledgement

We like to thank Tomaso Frigato for performing the MD-simulation of the peptide.

Appendix: Integration of the Likelihood function

Integration of the integral in (20) is rather straightforward but for completeness we will derive it in this appendix, i.e. we want to integrate $f(Z|\Phi, R)\pi_D(\Phi, R)$,with π_D the diffusive prior as given in Eq. 19 and f the density as given in 11, over all $\Phi \in \mathbb{R}^{d\times(dp+1)}$ and over all positive definite matrices $R \in \mathbb{R}^{d\times d}$, where Z is a given time series of length T and dimension d. With the notation in § 3.2 and § 3.4 we have

$$\int f(Z|\Phi, R)\pi_D(\Phi, R)d\Phi dR =$$
$$\int |2\pi R|^{-\frac{T-p}{2}} \exp\left(-\frac{1}{2}\Big(\operatorname{tr}\Big((Y-\Phi X)(Y-\Phi X)'R^{-1}\Big)\Big)\right)|R|^{-\frac{d+1}{2}}d\Phi dR.$$

The argument of the trace function can be Taylor expanded around the MLE of Φ yielding

$$\int |2\pi R|^{-\frac{T-p}{2}} \exp\Bigg(-\frac{1}{2}\Big(\operatorname{tr}\Big((Y-\hat{\Phi}X)(Y-\hat{\Phi}X)'R^{-1} \\ +(\mathbf{\Phi}-\hat{\mathbf{\Phi}})'[R^{-1}\otimes XX'](\mathbf{\Phi}-\hat{\mathbf{\Phi}})\Big)\Big)\Bigg)|R|^{-\frac{d+1}{2}} d\mathbf{\Phi} dR,$$

where $\mathbf{\Phi}$, resp. $\hat{\mathbf{\Phi}}$, denote the vectorised notation of Φ, resp. $\hat{\Phi}$, and $\otimes$ the Kronecker product. Next $\mathbf{\Phi}$ can be integrated out as it is normal distributed, giving

$$\int |2\pi R|^{-\frac{T-p}{2}} |R|^{-\frac{d+1}{2}} |2\pi(R^{-1}\otimes XX')^{-1}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\Big(\operatorname{tr}\Big((Y-\hat{\Phi}X)(Y-\hat{\Phi}X)'R^{-1}\Big)\Big)\right) dR,$$

which can be simplified to

$$(2\pi)^{-\frac{d(T-p-dp-1)}{2}} |XX'|^{-\frac{d}{2}} \int |R|^{-\frac{T-(d+1)p+d}{2}} \exp\left(-\frac{1}{2}\Big(\operatorname{tr}\Big((Y-\hat{\Phi}X)(Y-\hat{\Phi}X)'R^{-1}\Big)\Big)\right) dR$$

The resulting integrand is proportional to an inverted Wishart distribution with $T-p+d-dp$ d.o.f's, which has a defined density as long as $T > p+dp+d$, cf. [39, ch. 3.4]. Therefore R can be integrated out giving rise to

$$\pi^{\frac{d(d-1)}{4}} |XX'|^{-\frac{d}{2}} |\pi\cdot(Y-\hat{\Phi}X)(Y-\hat{\Phi}X)'|^{-\frac{T-p-dp-1}{2}} \prod_{j=1}^{d} \Gamma\left(\frac{T-p-dp-j}{2}\right).$$

The form stated in (20) is obtained by simply using the notation introduced in § 3.4.

References

[1] R. Elber, M. Karplus, Multiple conformational states of proteins: a molecular dynamics analysis of Myoglobin., Science 235 (4786) (1987) 318–321.

[2] H. Frauenfelder, P. J. Steinbach, Y. R. D., Conformational relaxation in proteins, Chemica Scripta 29A (1989) 145–150.

[3] C. Schütte, W. Huisinga, Handbook of Numerical Analysis, North-Holland, 2003, Ch. Biomolecular Conformations can be Identified as Metastable Sets of Molecular Dynamics, pp. 699–744.

[4] C. Schütte, A. Fischer, W. Huisinga, P. Deuflhard, A direct approach to conformational dynamics based on hybrid Monte Carlo, Journal of Computational Physics 151 (1999) 146–168.

[5] G. U. Nienhaus, J. R. Mourant, H. Frauenfelder, Spectroscopic evidence for conformational relaxation in Myoglobin, Proceedings of the National Academy of Science 89 (7) (1992) 2902–2906.

[6] A. Amadei, A. B. Linssen, H. J. Berendsen, Essential dynamics of proteins., Proteins 17 (4) (1993) 412–425.
URL http://dx.doi.org/10.1002/prot.340170408

[7] P. Deuflhard, W. Huisinga, A. Fischer, C. Schütte, Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains, Linear Algebra and its Applications 315 (2000) 39–59.

[8] P. Deuflhard, M. Weber, Robust Perron cluster analysis in conformation dynamics, Linear Algebra and its Applications 398 (2005) 161–184.

[9] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, W. C. Swope, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics., Journal of Computational Chemistry 126 (15) (2007) 155101.
URL http://dx.doi.org/10.1063/1.2714538

[10] F. Noé, I. Horenko, C. Schütte, J. C. Smith, Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states., J Chem Phys 126 (15) (2007) 155102.
URL http://dx.doi.org/10.1063/1.2714539

[11] C. Schütte, I. Horenko, Likelihood-based estimation of multidimensional langevin models and its application to biomolecular dynamics, Multiscale Modeling and SimulationAccepted.

[12] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, C. Schütte, Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models., Journal of Computational Chemistry 28 (15) (2007) 2453–2464.
URL http://dx.doi.org/10.1002/jcc.20692

[13] A. F. Voter, Parallel replica method for dynamics of infrequent events, Phys. Rev. B 57 (22) (1998) R13985–R13988.

[14] A. F. Voter, F. Montalenti, T. C. Germann, Extending the time scale in atomistic simulation of materials, Annual Review of Material Research 32 (2002) 321–346.

[15] E. S. Page, A test for a change in a parameter occurring at an unknown point, Biometrika 42 (1955) 523–527.

[16] A. Aue, L. Horváth, M. Hušková, P. Kokoszka, Change-point monitoring in linear models, Econometrics Journal 9 (2006) 373–403.

[17] L. Perreault, J. Bernier, B. Bobe, E. Parent, Bayesian change-point analysis in hydrometeorological time series. part2. comparision of change-point models and forecasting, Journal of Hydrology 235 (2000) 242–263.

[18] F. J. Girón, E. Moreno, G. Casella, Objective bayesian analysis of multiple changepoints for linear models, in: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, M. West (Eds.), Bayesian Statistics, Vol. 8, Oxford University Press, 2007, pp. 1–27.

[19] A. O'Hagan, Fractional bayes factors for model comparision, Journal of the Royal Statistical Society. Series B (Methodological) 57 (1) (1995) 99–138.

[20] N. Chopin, Dynamic detection of change points in long time series, Annals of the Institute of Statistical Mathematics 59 (2) (2007) 349–366.

[21] P. Fearnhead, Z. Liu, On-line inference for multiple changepoint problems, Journal of the Royal Statistical Society. Series B (Methodological) 69 (2007) 589–605.

[22] O. Lange, H. Grubmüller, Collective langevin dynamics of conformational motions in proteins, Journal of Chemical Physics 124 (2006) 214903.

[23] E. Meerbach, E. Dittmer, I. Horenko, C. Schütte, Multiscale modelling in molecular dynamics: Biomolecular conformations as metastable states, in: M. Ferrario, G. Ciccotti, K. Binder (Eds.), Computer Simulations in Condensed Matter: Systems: From Materials to Chemical Biology. Volume I, No. 703 in Lecture Notes in Physics, Springer, 2006, pp. 475–497.

[24] L. Arnold, Stochastic differential equations, Wiley, New York, 1974.

[25] H. Lütkepohl, Introduction to Multiple Time Series Analysis, Springer, Berlin - Heidelberg, 1991.

[26] S. Ni, D. Sun, Bayesian estimates for vector autoregressive models, Journal of Business & Economic Statistics 23 (2005) 105–117.

[27] I. Horenko, C. Hartmann, C. Schütte, F. Noe, Data-based parameter estimation of generalized multidimensional Langevin processes., Phys Rev E Stat Nonlin Soft Matter Phys 76 (1 Pt 2) (2007) 016706.

[28] A. Neumaier, T. Schneider, Estimation of parameters and eigenmodes of multivariate autoregressive models, ACM Transactions on Mathematical Software 27 (1) (2001) 27–57.

[29] G. Casella, R. L. Berger, Statistical Inference, Wadsworth & Brooks, 1990.

[30] R. E. Kass, A. E. Raftery, Bayes factors, Journal of the American Statistical Association 90 (430) (1995) 773–795.

[31] D. Sun, S. Ni, Bayesian analysis of vector-autoregressive models with noninformative priors, Journal of Statistical Planning and Inference 121 (2004) 291–309.

[32] J. Berger, L. Pericchi, The intrinsic Bayes factor for linear models, Bayesian Statistics 5 (1996) 25–44.

[33] G. Casella, E. Moreno, Objective bayesian variable selection, Journal of the American Statistical Association 101 (101) (2006) 157–167.

[34] D. Spiegelhalter, A. Smith, Bayes factors for linear and log-linear models with vague prior information, Journal of the Royal Statistical Society. Series B (Methodological) 44 (3) (1982) 377–387.

[35] E. J. Bylaska, W. A. de Jong, K. Kowalski, T. P. Straatsma, M. Valiev, D. Wang, E. Apr, T. L. Windus, S. Hirata, M. T. Hackler, Y. Zhao, P.-D. Fan, R. J. Harrison, M. Dupuis, D. M. A. Smith, J. Nieplocha, V. Tipparaju, M. Krishnan, A. A. Auer, M. Nooijen, E. Brown, G. Cisneros, G. I. Fann, H. Frchtl, J. Garza, K. Hirao, R. Kendall, J. A. Nichols, K. Tsemekhman, K. Wolinski, J. Anchell, D. Bernholdt, P. Borowski, T. Clark, D. Clerc, H. Dachsel, M. Deegan, K. Dyall, D. Elwood, E. Glendening, M. Gutowski, A. Hess, J. Jaffe, B. Johnson, J. Ju, R. Kobayashi, R. Kutteh, Z. Lin, R. Littlefield, X. Long, B. Meng, T. Nakajima, S. Niu, L. Pollack, M. Rosing, G. Sandrone, M. Stave, H. Taylor, G. Thomas, J. van Lenthe, A. Wong, Z. Zhang, NWChem, a computational chemistry package for parallel computers, version 5.0, Tech. rep., Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA (2006).

[36] R. Kendall, E. Apra, D. Bernholdt, E. Bylaska, M. Dupuis, G. Fann, R. Harrison, J. Ju, J. Nichols, J. Nieplocha, T. Straatsma, T. Windus, A. Wong, High performance computational chemistry: An overview of NWChem a distributed parallel application, Computer Phys. Comm. 128 (2000) 260–283.

[37] D. Stalling, M. Westerhoff, H.-C. Hege, The Visualization Handbook, Elsevier Academic Press, 2004, Ch. Amira: A Highly Interactive System for Visual data analysis (Ch. 38), pp. 749–767.

[38] S. Johnson, Hierarchical clustering schemes, Psychometrika 2 (1967) 241–254.

[39] A. Gupta, D. Nagar, Matrix variate distributions, Chapman & Hall, 2000.