SEQUENTIAL CHANGE POINT DETECTION IN MOLECULAR DYNAMICS TRAJECTORIES

EIKE MEERBACH†, JUAN C. LATORRE†, AND CHRISTOF SCHÜTTE†

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

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

AMS subject classifications. 62H15, 62J12, 92C40

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. [13, 16, 38, 36, 30, 25].

Biophysical research seems to indicate that these metastable sets of a typical biomolecular system can be characterised in terms of a small number of essential degrees of freedom [1], 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 [38, 11, 12, 9, 31]. 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 [37, 15].

All these approaches are based upon a global analysis of a given time series. 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 [42, 43] presented an approach based on parallel simulation of uncorrelated copies of a molecular system. However, his approach assumes that transitions between the (not known) conformations can be detected on-line. In this article we present a statistical model for trajectories of molecular systems and show a change point algorithm that can be used to detect these changes on-line.

The change point detection problem has received considerable attention in the last years, but the problem is essentially still unsolved. One of the most prominent

†Department of Mathematics and Computer Science, Freie Universität Berlin, Germany

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., [33, 3]. CUMSUM methods are in many cases asymptotically well understood but effectively rely on knowing the parameter-values which are changing. 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, since 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. [35]. On the other hand many ways to obtain "objective" priors, like the one presented by Girón et. al. [17], are feasible only in low dimensional parameter settings. We employ the fractional Bayes approach of O'Hagan [32] and show that this can be easily adapted to our setting.

Although we avoided a sampling based approaches like particle filtering [10, 14], 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 [14] as our algorithm also relies on the integrability over parameter space of the likelihood functions for linear models. Opposed to sampling based algorithms our algorithm can only, at least in the form stated here, handle multiple change points in a sequential form.

This paper is organized as follows. 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 the problem of a diffusive particle in a three-well, two-dimensional potential, as well as 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 [23]. In the easiest case the resulting *effective dynamics* is given by a first order Langevin equation,

$$\dot{\boldsymbol{z}}(t) = -\nabla_{\boldsymbol{z}} U(\boldsymbol{z}(t)) + \Sigma \dot{\boldsymbol{W}}(t), \tag{2.1}$$

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 (2.1) and comment later (see § 3.2) on generalised models obtained by adding a memory kernel to the noise.

Following [37, 26, 25], we linearise the non-linear stochastic differential equation (SDE) given in (2.1) 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,

$$\begin{aligned} \dot{\boldsymbol{z}}(t) &= F^{[i(t)]}\left(\boldsymbol{z} - \boldsymbol{\mu}^{[i(t)]}\right) + \Sigma^{[i(t)]}\dot{\boldsymbol{W}}(t), \\ i(t) &\in \{1, \ldots, s\}, \end{aligned} \tag{2.2}$$

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. (2.1)). If the obtained reduced model can be parameterised based on some observed time series, one encounters the problem that, unlike $\boldsymbol{z}$, the switching process $i(t)$ is not observable. In [15, 26, 37] 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 (2.2) using the so-called Expectation-Maximisation algorithm.

Our focus here 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 $t_0 \le t < t_1$ and $i(t) = k$ for $t \ge t_1$, while observing the time series sequentially. We will call a time point where such a change in parameterisation occurs (t_1 above) a *change point*.

However, in § 6.4 we show that if all change points are successfully identified over a certain time interval, we can use the obtained information to estimate transition probabilities of the process $i(t)$. Before dealing with the change point detection problem we need 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 [37]. 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 [2]. In particular, given an observation $\boldsymbol{z}_t$, the conditional probability density of $\boldsymbol{z}_{t+1}$ is a Gaussian with density function

$$f_\lambda(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t) = \frac{1}{\sqrt{|2\pi R(\tau)|}} \exp\left(-\frac{1}{2}(\boldsymbol{z}_{t+1} - \boldsymbol{\mu}_t)^T R(\tau)^{-1}(\boldsymbol{z}_{t+1} - \boldsymbol{\mu}_t)\right), \tag{3.1}$$

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. We can get around this problem by rewriting (3.1) as,

$$\begin{aligned} \boldsymbol{z}_{t+1} &= \mathcal{N}(\boldsymbol{\mu} + \exp(\tau F)(\boldsymbol{z}_t - \boldsymbol{\mu}), R) \\ &= (I - \exp(\tau F))\boldsymbol{\mu} + \exp(\tau F)\boldsymbol{z}_t + \mathcal{N}(\mathbf{0}, R), \end{aligned} \tag{3.2}$$

where $\mathcal{N}$ is a multivariate normal distributed random variable and I an identity matrix of appropriate size. Eq. (3.2) reveals the autoregressive structure of order one, VAR(1), of the time series of discrete observations. Defining,

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

allows to write (3.2) in a compact form,

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

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

$$L(\tilde{\lambda}|Z) = \left(\frac{1}{\sqrt{|2\pi R|}}\right)^{(T-1)} \exp\left(-\frac{1}{2}\operatorname{tr}((Y - \Phi X)(Y - \Phi X)' R^{-1})\right), \tag{3.3}$$

for which there are analytic MLE's $\hat{\Phi}$ and $\hat{R}$ given by [24, 29],

$$\hat{\Phi} = YX'(XX')^{-1} \text{ and } \hat{R} = (Y - \hat{\Phi}X)(Y - \hat{\Phi}X)'/(T-1). \tag{3.4}$$

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 Appendix A).

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 for using a higher order model, e.g. VAR(p), that is

$$\boldsymbol{z}_{t+1} = A_0\boldsymbol{\mu} + \sum_{i=1}^{p} A_i \boldsymbol{z}_{t-i+1} + \mathcal{N}(\mathbf{0}, R). \tag{3.5}$$

This process is obviously not a Markov process anymore since it exhibits a memory lag of p steps into the past. In fact, Eq. (3.5) can be interpreted as a time discretization of a generalised Langevin process

$$\dot{\boldsymbol{z}}(t) = -\nabla_{\boldsymbol{z}} U(\boldsymbol{z}(t)) - \int_0^t \gamma(t-s)\boldsymbol{z}(s)ds + \Sigma\dot{\boldsymbol{W}}(t), \tag{3.6}$$

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 [19].

If a fixed order p is assumed estimation of the parameters of the VAR(p) process is analogue to that of the VAR(1) process, we only have to extend the definitions of the data matrices X and Y to

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

The estimator $\hat{\Phi}$ in (3.4) now estimates,

$$\Phi = \begin{pmatrix} A_0\boldsymbol{\mu} & A_1 & A_2 & \dots & A_p \end{pmatrix} \in \mathbb{R}^{d\times(dp+1)},$$

The estimator of $\hat{R}$ has to be adusted to the growing number of initial points needed for higher order and becomes,

$$\hat{R} = (Y - \hat{\Phi}X)(Y - \hat{\Phi}X)'/(T-p).$$

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 [24, Chapter 4.3]. In our application example we use the Schwarz criterion which chooses the order p such that the function

$$SC(p) = \log|\hat{R}(p)| + \frac{\log T}{T} pd^2, \tag{3.7}$$

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 the order of the VAR process considered here is essentially the memory length of the molecular process, which might be known or estimated by methods specific to the MD context.

3.4. Computation of the Estimators. The analytic estimators given in (3.4) are in general not used for computation of the parameters, as the matrix inversion

may be unstable. Instead one can use the moment matrix,

$$\begin{aligned} M = M(Z) :&= \sum_{i=1}^{T} \begin{pmatrix} 1 \\ \boldsymbol{z}_i \\ \vdots \\ \boldsymbol{z}_{i+p} \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}_i' & \dots & \boldsymbol{z}_{i+p}' \end{pmatrix} \\ &= \begin{pmatrix} XX' & XY' \\ YX' & YY' \end{pmatrix} =: \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}. \end{aligned} \tag{3.8}$$

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:

$$\begin{aligned} L(\Phi, R|Z) = L(\Phi, R|M) = \left(\frac{1}{\sqrt{|2\pi R|}} \right)^m \\ \cdot \exp\left(-\frac{1}{2} \operatorname{tr}((M_{22} - M_{21}\Phi' - \Phi M_{12} + \Phi M_{22}\Phi')R^{-1}) \right), \end{aligned} \tag{3.9}$$

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 the notation $m = m(Z)$ below to avoid the indices for the length of different time series. We will also subsequently use the notation $f(Z|\Phi, R) \equiv L(\Phi, R|Z)$ if we want to highlight Eq. (3.9) 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

$$U = \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix},$$

such that,

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

After substituting the Cholesky factorisation into the estimators (3.4) one obtains,

$$\hat{\Phi} = (U_{11}^{-1}U_{12})' \text{ and } \hat{R} = \frac{1}{m} U_{22}'U_{22}. \tag{3.10}$$

In 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 $M + \delta \mathrm{diag}(M)$ instead of M, with a small parameter δ depending on the dimensionality of the problem and the machine precision, cf. [28].

4. Bayesian Model Selection.

4.1. Change Point Models. In order to detect changes in the parameterisation of the underlying VAR model we utilise a Bayesian approach similar to the one presented in [35] for the univariate and independent distributed case. Assume that we have observed a sequence of data points,,

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

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 (this assumption will become evident 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,

$$H_i := \begin{cases} Z \text{ is generated by only one VAR}(p)\text{-DGM,} & \text{for } i = 0. \\ Z_1 = Z_1(i) = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t_1+i-1}\} \text{ and} \\ Z_2 = Z_2(i) = \{\boldsymbol{z}_{t_1+i-p}, \ldots, \boldsymbol{z}_T\} \\ \text{are generated by distinct VAR}(p)\text{-DGM's.} & \text{for } 1 \leq i \leq n. \end{cases} \tag{4.1}$$

Note that the Z_1 and Z_2 are overlapping, since the last p points of Z_1 are used as initial conditions for Z_2.[1]

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

$$\mathbb{P}[H_i|Z] = \frac{\mathbb{P}[Z|H_i]\,\mathbb{P}[H_i]}{\sum_{j=0}^{n} \mathbb{P}[Z|H_j]\,\mathbb{P}[H_j]}, \tag{4.2}$$

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.

With (4.2) at hand, we assume now a so-called M-closed perspective [5, Chapter 6], i.e., we believe that the true model is within the models stated above and we no other model is possible. In this scenario, we can easily evaluate the probability of a change point as,

$$\mathbb{P}[\text{change}|Z] = \frac{\sum_{i=1}^{n} \mathbb{P}[Z|H_i]\,\mathbb{P}[H_i]}{\sum_{j=0}^{n} \mathbb{P}[Z|H_j]\,\mathbb{P}[H_j]}. \tag{4.3}$$

In order to evaluate these probabilities we must 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.

A natural choice to include 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 assign a uniform

[1] 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.

probability to all the other models,

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

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

$$\pi_D(\Phi, R) \propto |R|^{-\frac{d+1}{2}}. \tag{4.4}$$

A discussion of this prior and other possibilities is given in [29, 41]. Although it can be easily shown that under the diffusive prior the posterior distribution,

$$\int p(Z|\Phi, R)\pi_D(\Phi, R)d\Phi dR,$$

is proper, i.e., normalisable, the choice is problematic for model comparison, as the prior itself is unproper: We can set,

$$\pi_1 \equiv \pi_2 \equiv c\pi_D,$$

with an arbitrary chosen constant c. This means that the model probabilities (4.2) as well as probability of change (4.3) are also defined up to a constant,

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

The constant does not cancel out of the fraction as there are parameters which are not common to all models, i.e., the parameters for the VAR model after a change has occurred. To emphasise: with the use of an unproper prior we can compare different change point models, since the indeterminate constants do cancel out, but we can not compare the probability between change and no-change.

This general obstacle of Bayesian model selection can be tackled by the usage of so-called "objective" Bayes factors [21], which we are going to introduce in the next section.

On the other hand, note that it is possible to split the change point detection problem into two parts:

1. Identify the most likely change point under the assumption that there is one, this requires specifications of parameter priors up to a constant only.
2. Compare the probability of change at the identified possible change point with the probability of no change. Now we have to specify a proper prior for the parameters after the change to avoid arbitrariness.

In fact, this approach is favourable from an algorithmic viewpoint as it is easier to exclude outliers from the change/no-change decision (see § 6) and the M-closed perspective is somewhat arbitrary anyway as the number of models obviously depends upon the defined window $[t_1, t_2]$. We now have to compare a change model with a (single) no-change model, i.e., we have to compute the probability,

$$\mathbb{P}[\text{change}|Z] = \frac{\mathbb{P}[Z|H_c]\,\mathbb{P}[H_c]}{\mathbb{P}[Z|H_c]\,\mathbb{P}[H_c] + \mathbb{P}[Z|H_0]\,\mathbb{P}[H_0]}, \tag{4.5}$$

where H_c is the model of a change at the pre-computed candidate change point c.

4.2. Bayes Factors. The Bayes factors are a common way to compare posterior probabilities of two distinct models within a Bayesian setting. Given two models H_i and H_j, the ratio

$$\frac{\mathbb{P}[H_i|Z]}{\mathbb{P}[H_j|Z]} = \frac{\mathbb{P}[Z|H_i]\,\mathbb{P}[H_i]}{\mathbb{P}[Z|H_j]\,\mathbb{P}[H_j]}, \tag{4.6}$$

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

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

Eq. (4.6) 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 is then obtained by *integration* of the likelihood function over the parameter space [7, 21]. Eq. (4.3) can be reformulated in terms of the Bayesian factors, as

$$\mathbb{P}[\text{change}|Z] = \frac{\sum_{i=1}^{n} B_{i0}\,\mathbb{P}[H_i]}{\sum_{j=0}^{n} B_{j0}\,\mathbb{P}[H_j]}. \tag{4.7}$$

This expression can can be interpreted as an assembly of a sequence of tests against the null hypothesis of no change [17].

4.3. Non-informative Priors in Model Selection Problems. The Bayes factors do not resolve the problem with the unproperness of the non-informative standard priors, since they become also arbitrary when used with non-proper priors. Subsequently we present three approaches to tackle this problem by deriving proper priors in a data driven way.

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 into two parts Z_p and Z_{-p} and use one 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., we set,

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

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 which data points are taken into the training sample. A proposal to overcome this arbitrariness is given by Berger [4], who suggested to average over all possible minimal training sets, the so-called *intrinsic Bayes* approach. The intrinsic Bayes approach can be elegantly expanded if nested models are tested, [17, 8], but has the drawback that computation, even with sampling procedures, of intrinsic Bayes factors is often hard, resp. feasible only for a restricted class of models.

4.3.2. Fractional Bayes. The fractional Bayes approach, put forward by O'Hagan [32], is based on the idea to use a fraction of the likelihood function, instead of using part of the data, to specify a prior, i.e., to set

$$\pi_{FB}(\theta) \propto \pi_D(\theta) L^b(\theta|Z),$$

with a constant $b \in]0,1[$. The likelihood function used for decision making is then transformed to $\tilde{L}(\theta|Z) := L^{(1-b)}(\theta|Z)$, thus becoming flatter as a fraction of the information is 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 the minimal value which guarantees properness of the resulting prior, which corresponds to the choice of a maximal spreaded distribution centered by the data.

4.3.3. Imaginary Minimal Experiment. Another approach presented by Spiegelhalter and Smith [39] 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 c an unknown constant. 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, thus 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 minimally, 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 [32]. 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 sensible.

5. Implementation of the Objective Bayesian Strategies. In the previous sections we collected all necessary ingredients for a Bayesian change point detection, so now we make more precise how to do this in our given scenario. As mentioned before, we split the change point detection problem in two parts; first, identify a possible candidate change point and, second, decide whether this point is actually a change point.

The key ingredient which allows us to employ the approaches stated above is that our model allows analytical integration of the likelihood function over parameter space. Assume for the moment an arbitrary time series Z of length T, and the corresponding moment matrix $M = M(Z)$. Since M contains all statistical relevant information of the data we can write $p(M|\Phi, R)$ instead of $p(Z|\Phi, R) = L(\Phi, R|Z)$, as given in Eq. (3.9). Following the notation introduced in § 3.4 we denote by U_{11} and U_{22} the corresponding diagonal blocks of the triangular matrix U obtained from the Cholesky factorisation of M, and by $m := M_{11} = T - p$ the upper left scalar entry of M. Then (see Appendix A),

$$\begin{aligned} I[M] := &\int p(M|\Phi, R)\pi_D(\Phi, R)d\Phi dR = \int L(\Phi, R|M)\pi_D(\Phi, R)d\Phi dR \\ &= \pi^{\frac{d(d-1)}{4}} |U_{11}|^{-d} |\sqrt{\pi}\, U_{22}|^{-(m-dp-1)} \prod_{j=1}^{d} \Gamma\left(\frac{m-dp-j}{2}\right), \end{aligned} \tag{5.1}$$

where Γ denotes the Gamma function and $|\cdot|$ the matrix determinant. Note that the integral exist only if $m > d(p+1) \Leftrightarrow T > d(p+1)+p$, therefore at least $(d+1)(p+1)$ subsequent points before and after a change point are needed for evaluation. From Eq. (3.9) another property of the M-matrices can be deduced, namely that information coming from different time series (parts), e.g., Z_1 and Z_2 can be combined by just adding the moment matrices, since

$$L(\Phi, R|M(Z_1))L(\Phi, R|M(Z_2)) = L(\Phi, R|M(Z_1) + M(Z_2)). \tag{5.2}$$

5.1. Identification of a Change Point Assuming its Existence. Using the notation introduced in § 4, the aim is to calculate the probabilities of potential positions of a candidate change point,

$$\mathbb{P}[H_i|Z] \propto \int p(Z_1|\Phi, R)\pi_1(\Phi, R)\, d\Phi dR \int p(Z_2|\Phi, R)\pi_2(\Phi, R) d\Phi dR, \tag{5.3}$$

with $1 \leq i \leq n = t_2 - t_1$ and

$$\begin{aligned} Z_1 =& Z_1(i) := \{\boldsymbol{z}_1, \boldsymbol{z}_2 \ldots, \boldsymbol{z}_{t_1+i-1}\}, \\ Z_2 =& Z_2(i) := \{\boldsymbol{z}_{t_1+i}, \boldsymbol{z}_{t_1+i+1} \ldots, \boldsymbol{z}_T\}. \end{aligned}$$

Note that t_1 must be larger or equal than $(d+1)(p+1)$ and t_2 smaller than $T - 2 - (d+1)(p+1)$, so that each segment contains at least $(d+1)(p+1)$ data points, since otherwise the integrals can not be evaluated. We can include information which might be already obtained from a previous observation Z_0 into the prior distribution π_1 by setting

$$\pi_1(\Phi, R) \propto \pi_D(\Phi, R)L(\Phi, R|Z_0),$$

which is formally a partial Bayes approach, however, the motivation is not to make the prior distributions proper, as at this stage proper priors are not essential, but to include prior information from previous observations. Otherwise we take the diffuse prior for both parameter sets, i.e.,

$$\pi_1(\Phi, R) = \pi_2(\Phi, R) \propto \pi_D(\Phi, R).$$

Using (5.1) and (5.2) we have,

$$\mathbb{P}[H_i|Z] \propto I[M(Z_0) + M(Z_1)]I[M(Z_2)], \tag{5.4}$$

with prior observations Z_0. If there are no prior observations we set $M(Z_0)$ to $\mathit{0}$. Thus it is possible to determine the most probable change point $\hat{c}$ analytically, by choosing

$$\hat{c} = \underset{1 \leq i \leq n}{\operatorname{argmax}}\, \mathbb{P}[H_i|Z],$$

even if we do not know if this is an actual change point. An example is depicted in Fig. 5.1, where it can be seen how a change point can be identified by locating the maximum of the conditional density, but that one still has to decide if this maximum really belongs to a change point. This can be done by Fractional Bayes or the Imaginary Minimal Experiment as exemplified below. In general, however, we could use any method for our decision which seems to be appropriate. Therefore splitting the change point analysis has the big advantage that the difficult problem, i.e., the model decision problem, is now separated from the problem of locating the most probable change point.

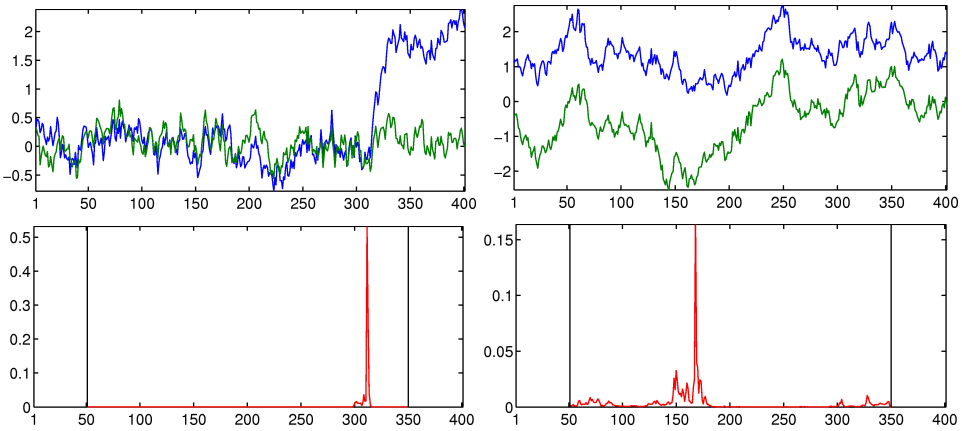


FIGURE 5.1. Left: *the top panel displays a 2-dimensional trajectory against time artificially generated from a VAR(1) process with a switch in the mean occurring at* $t = 311$. *Below is the probability density of a change point conditional to its existence. The margin lines left and right of the panel mark the test interval* $[t_1, t_2]$. Right: *An example where no change point occurs in the trajectory. Still we obtain a candidate change point.*

Fractional Bayes. The fractional Bayes approach can be easily implemented by noting from (3.9) that

$$\begin{aligned} &L^b(\Phi, R|M) \\ &\quad = \left(\frac{1}{\sqrt{|2\pi R|}}\right)^{bm} \exp\left(-\frac{1}{2}\operatorname{tr}((bM_{22} - bM_{21}\Phi' - \Phi bM_{12} + \Phi bM_{22}\Phi')R^{-1})\right) \\ &\quad = L(\Phi, R|bM), \end{aligned} \tag{5.5}$$

so that, using the notation introduced above, we have

$$\int L^b(\Phi, R|Z)\pi_D(\Phi, R)d\Phi dR = I[bM(Z)].$$

Also, using (5.2),

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

Setting the prior probabilities for change and no change equally to $\frac{1}{2}$ the quantity to compute reads,

$$\begin{aligned} \mathbb{P}[\text{change}|Z] &= \frac{\mathbb{P}[Z|H_{\hat{c}}]}{\mathbb{P}[Z|H_{\hat{c}}] + \mathbb{P}[Z|H_0]} \\ &= \frac{\prod_{i=1,2}\int p_i(Z_i|\Phi, R)\pi_i(\Phi, R)d\Phi dR}{\prod_{i=1,2}\int p_i(Z_i|\Phi, R)\pi_i(\Phi, R)d\Phi dR + \int p_0(Z|\Phi, R)\pi_1(\Phi, R)d\Phi dR}. \end{aligned}$$

We leave $\pi_1 \propto \pi_D$ unproper, since the normalisation constant cancels out anyway, or with prior observations $\pi_1(\Phi, R) \propto \pi_D(\Phi, R)L(\Phi, R|Z_0)$, but for π_2 we use the

fractional Bayes approach,

$$\pi_2(\Phi, R) := \frac{\pi_D(\Phi, R)L^b(\Phi, R|Z_2)}{\int \pi_D(\Phi, R)L^b(\Phi, R|Z_2)d\Phi dR}.$$

Since some data is used to specify the prior distribution, we can not use all of it for calculation of the probability, i.e., we set

$$\begin{aligned} p_1(Z_1|\Phi, R) &= L(\Phi, R|Z_1), \\ p_2(Z_2|\Phi, R) &= L^{(1-b)}(\Phi, R|Z_2), \\ p_0(Z|\Phi, R) &= L^{(1-b)}(\Phi, R|Z_2)L(\Phi, R|Z_1). \end{aligned}$$

Assembling all the pieces we obtain, in compact notation, the probability,

$$\mathbb{P}[\text{change}|Z] = \frac{I[M(Z_1)]I[M(Z_2)]}{I[M(Z_1) + (1-b)M(Z_2)]I[bM(Z_2)] + I[M(Z_1)]I[M(Z_2)]}. \quad (5.6)$$

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

$$I[bM(Z_2)]$$

is defined (cf. Appendix A). Therefore the minimal value of b is given by

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

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

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 VAR model as Z_1 we need, as stated above, a minimum of $(d+1)(p+1)$ observations, otherwise integration over the parameter space is 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

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

Substituting the obtained Bayes factor in (4.7) gives an expression for the change probability:

$$\mathbb{P}[\text{change}|Z] = \frac{I[M(Z_1) + M(Z_I)]I[M(Z_2)]}{I[M(Z_1) + M(Z_2)]I[M(Z_I)] + I[M(Z_1) + M(Z_I)]I[M(Z_2)]}. \quad (5.7)$$

Coming back to the example given in Fig. 5.1, we can now compute the probability of a change for the identified candidate change point in both time series. Then Z_1 becomes the part of the analysed time series before the candidate change point (where the conditional change point probability is maximised,) and Z_2 the part after the candidate change point. Computation of the change probabilities (5.6) and (5.7) corresponding to these segments for both time series gives the following results.

$\mathbb{P}[\text{change}\|Z]$	Time series 1 (left in Fig. 5.1)	Time series 2 (right in Fig. 5.1)
Fractional Bayes (5.6)	1	0.0217
Imaginary Experiment (5.7)	1	0.0226

We see that both procedures yield the right result, and reject a change point where no change occurred (time series 2) while accepting the true change point (time series 1). We choose in what follows the fractional Bayes approach since it worked satisfactory in various test cases, is computational cheap, includes parameter uncertainty and is less speculative than the imaginary minimal experiment approach.

6. Algorithmic Procedure. In this section we state the proposed algorithmic procedure derived by the considerations above. Before we do this, we will comment on how to cope with effects due to the finiteness of the time series the change point analysis is applied to. After stating the core algorithm, we will also comment on post processing possibilities, followed by showing two examples in the next section.

To make the following sections more readable we introduce the following notation: Given some time series (segment) $\{\boldsymbol{z}_{t_0}, \boldsymbol{z}_{t_0+1}, \ldots, \boldsymbol{z}_{t_1}\}$ we define by $M(t_0, t_1)$ the corresponding moment matrix

$$M(t_0,t_1) := \sum_{t=t_0+p}^{t_1} \begin{pmatrix} 1 \\ \boldsymbol{z}_{t-p} \\ \vdots \\ \boldsymbol{z}_t \end{pmatrix} \begin{pmatrix} 1 & \boldsymbol{z}'_{t-p} & \ldots & \boldsymbol{z}'_{t\cdot} \end{pmatrix}$$

Obviously the definition depends on VAR order p, but we omit an appropriate index if the p is evident from the context.

6.1. Margin Effects. A systematic problem that occurs when change point detection based upon parameter estimation is applied to finite a time series is that, if the segments of the time series are too short, the information about the parameters in these segments can be misleading. An illustration of this effect is given in the left panel of Fig. 6.1. Therefore, if a time series with no change point is analysed, the change point algorithm will tend to detect change points close to the ends of the time series. The key point is that this effect can not be overcome just by regarding parameter uncertainty, since the information contained in a short time series segment is not just insufficient but also misleading. An example is given in the right panel of Fig. 6.1. Here a trajectory generated by a VAR model is shown together with the function,

$$f(i) = I[M_1(i)]I[M_2(i)],$$
$$M_1(i) = M(1,i), \qquad M_2(i) = M(i+1,T),$$

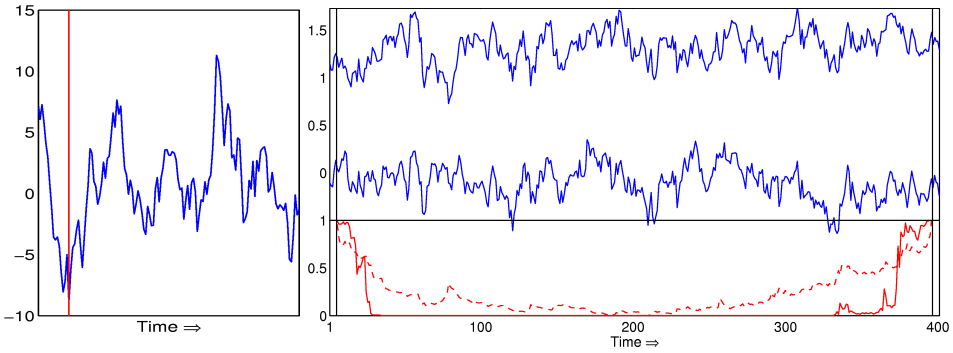


FIGURE 6.1. Left: *A realisation of an artificially generated one dimensional VAR(1) process is plotted. If parameters are estimated only from data points left to the red line they will significantly differ from parameters obtained using the whole time series.* Right: *In the upper part of the figure a two dimensional trajectory from a VAR(1) process is plotted (blue lines). Below the black line the logarithm of f (see text) is shown (dashed red line, arbitrary scale). Also shown is the probability of a change point for each possible change point that would be obtained by using the fractional Bayes approach (red line). The vertical lines border the minimal length of a time series segment, here* $(d+1)(p+1) = 6$. *It can be seen that in the margin regions the procedure would always detect a change point.*

which determines the candidate change points when no prior information is available, cf. § 4.3. It should not be surprising that the candidate change points are located at the margins of the time series. But if we look at the probability that a given candidate change point is a real change point, according to (4.7), we see that it is, as a function of candidate change points, close to one at the margins. To prevent this effect one could add a penalty function for change points close to the border. We implement this strategy by testing for change points only within a window which leaves the margins large enough. Note that we do not need a left margin if prior information about the parameters is included from prior observations, i.e., if another moment matrix M_p is at hand such that we can set,

$$M_1(i) = M_p + M(1, i).$$

6.2. Short Time Deviations - Recrossings. When applying change point detection to real data one naturally has to handle with outliers or short time deviations (for a short period of time the dynamical behaviour of the time series is different from that before and after). It is often the case that one does not want to detect such short time deviations since we are interested only in persistent changes of the dynamical behaviour. One can avoid detection of deviations shorter than some predefined time $t_b \in \mathbb{N}$ in the following way: Having identified a candidate change point c, one calculates the probability of a change occurring at that position based upon the matrices M_1, M_2, M_3 which contain the sufficient statistics of the time series before, after and without a change point,

$$M_1 = M(1, c-1), \quad M_2 = M(c, T), \quad M_3 = M_1 + M_2.$$

Instead of using these matrices one can exclude the information contained in the trajectory for t_b steps after the candidate change point by using, instead of M_2 and M_3,

$$\widetilde{M_2} = M(c + t_b, T), \quad \widetilde{M_3} = M_1 + \widetilde{M_2}.$$

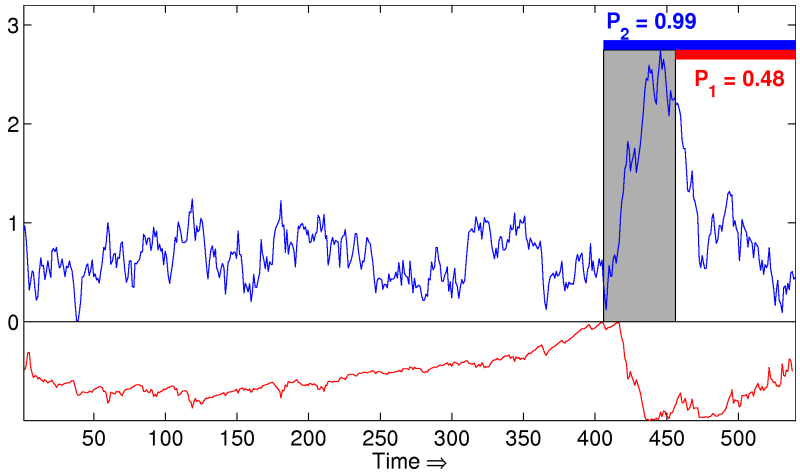


FIGURE 6.2. *A one dimensional artificially generated time series is shown (blue) which fluctuates mainly in* $[0, 1.5]$*, at around* $t = 400$ *there is an obvious rise but after approx. 50 steps it seems to regain the beforehand behaviour. The (logarithm of the) conditional change point distribution (red line) clearly identifies a candidate change point at the beginning of the rise. The change probability at this candidate change point is* $P_2 = 0.99$*. If a window as described in the text is used to mask out a part of the trajectory after the candidate change point (gray shaded) the change probability drops to* $P_1 = 0.48$*, as after the window the behaviour of the trajectory is similar to that before.*

The rationale behind this strategy is that only if the dynamical behaviour after the potential change point stays different longer than the predefined time t_b, it will affect the calculated probabilities because the dynamical information between c and t_b is not used. For an example see Fig. 6.2.

6.3. Algorithm. We summarise the results obtained so far in the basic algorithmical procedure stated in Algorithm 1.

Note that in the implementation of this algorithmic scheme we substituted quantities by their logarithmic values where suited to avoid numerical problems. A few comments on the parameters:

- Note that the most important parameter is t_m, as it determines the resolution of the change point algorithm. Defining a minimal segment size of course requires that there is no change point within the first t_m data points.
- The parameter t_b has a double role, first it is used as described in § 6.2 to exclude short deviations from the change point detection and, second, it excludes candidate change points which are close to the right margin of the test window, because the change might have had happen at the very end of the time series (where it was not tested due to margin effects). This is unproblematic since the change point would be detected in the next cycle.
- In practice, it turns out that the threshold parameter α should be chosen rather large to avoid false positives, fundamental changes will reflect in a change probability close to one anyway.

For a long time series the stated algorithm can become quite ineffective since, after every new received data package, the loop (1.1) in Algorithm 1 to determine a candidate change point is executed over all data points received so far. In practice this can be overcome by testing for a candidate change point only over the last w_{max} received data points, and add the information content of the beforehand received data

Algorithm 1: Sequential change point detection

Parameter: $t_m \in \mathbb{N}$ (minimal segment size)
$t_u \in \mathbb{N}$ (length of update window)
$t_b \in \mathbb{N}$ (length of buffer zone)
$\alpha \in]0,1[$ (threshold value for detection of a change)
p (VAR order) or p_{max} (maximal VAR order)

Input : A d-dimensional time series $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots\}$ with sequential access.

If p is not given estimate $p \in \{0, 1, \ldots, p_{\max}\}$ based on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t_m}$.
$M_I \leftarrow M(1, t_m)$
$t_E \leftarrow 2t_m + t_u$
$\mathbb{P}[\text{change}] \leftarrow 0$
while $\mathbb{P}[\textit{change}] < \alpha$ **do**
1.1 *Determine candidate change point between t_m and $t_E - t_m$:*
for $k \leftarrow p+2$ **to** $t_E - 2t_m$ **do**
$M_1 \leftarrow M(t_m + 1, t_m + k) + M_I$
$M_2 \leftarrow M(t_m + k + 1, t_E)$
$l_{k-p-1} \leftarrow I[M_1]I[M_2]$ (cf. (5.4))
$\hat{c} = t_m + p + \operatorname{argmax}_k l_k$
If the candidate change point is not too close to the margin compute its probability:
1.2 **if** $t_E - \hat{c} > t_b + t_m$ **then**
$M_1 \leftarrow M(t_m, \hat{c}) + M_I$
1.3 $M_2 \leftarrow M(\hat{c} + 1 + t_b : t_E)$
$b \leftarrow (dp + d + 1)/(M_2(1,1))$

$$\mathbb{P}[\text{change}] = \frac{I[M_1]I[M_2]}{I[M_1 + (1-b)M_2]I[bM_2] + I[M_1]I[M_2]}, \qquad \text{(cf. (5.6))}$$

$t_E \leftarrow t_E + t_u$

Output: The change point $\hat{c}$ and a corresponding moment matrix M_1 for the identified segment.

points to the moment matrix M_I.

In order to detect multiple change points the algorithm can be used multiple times, starting each time from the last detected change point again. In applications it is often advisable to start the algorithm again with a certain lag to the last detected change point in order to prevent that the transition phase between different dynamical phases spoils the statistics.

Note that after the information of a part of the time series is stored in a moment matrix M, this part can be completely discarded as all statistical relevant information is now stored in M. Therefore the whole approach is suited to handle with large data sets as it occurs in molecular dynamics simulations, see § 7.2 for an example.

6.4. Post-processing. If the change point algorithm is applied repeatedly on a time series to obtain multiple change points, the procedure will finally generate a sequence of change points $c_0 := 1, c_1, \ldots, c_{s-1}, c_s := T+1$ and therefore a segmentation

of a given time series Z, whose segments are given by

$$Z_i = \boldsymbol{z}_{c_{i-1}}, \ldots, \boldsymbol{z}_{c_i-1}, \quad 1 \leq i \leq s,$$

and the corresponding moment matrices $M_1, \ldots, M_s$ which contain the statistically relevant information. These matrices can be used for post processing purposes, e.g., to drop falsely detected change points or to group the data globally.

For this purpose we define a distance matrix D, measuring the distance between all identified segments $Z_1, \ldots, Z_s$ of the time series, according to the probability that the segments are generated by the same VAR model:

$$D_{ij} = \begin{cases} \dfrac{I[M_i]I[M_j]}{I[bM_j]I[M_i + (1-b)M_j] + I[M_i]I[M_j]}, \text{ if } m_i \geq m_j \\ \dfrac{I[M_i]I[M_j]}{I[bM_i]I[M_j + (1-b)M_i] + I[M_i]I[M_j]}, \text{ if } m_i < m_j, \end{cases} \tag{6.1}$$

$$b = \frac{dp + d + 1}{\min(m_i, m_j)},$$

with $1 \leq i, j \leq s$ and m_i the upper left entry of M_i. 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 of information from the shorter segment we always use the longer segment to extract prior information about the parameters.

In order to exclude falsely detected change points one should re-test for a change point between adjacent segments, i.e., generate a new set of change points $\tilde{c}_0, \ldots, \tilde{c}_k$ and a corresponding set of moment matrices $(\tilde{M}_1, \ldots, \tilde{M}_k)$ using the distance defined in (6.1) by Algorithm 2.

Algorithm 2: Exclusion of falsely detected change points

```
Parameter: c̃_0 = 1
           j = 1
for i = 1 to s − 1 do
    if D_{i,i+1} < α then
        M̃_j = M_i + M_{i+1}
    else
        c̃_j = c̃_i
        j = j + 1
        M̃_j = M_i
c̃_{j+1} = T
```

Segments may be merged again, even if the same criteria is used as in the change point Algorithm 1, due to the fact that the decision is now based on more data points.

Furthermore, the obtained distance matrix can be used to cluster the data, i.e. to merge different time series segments (in fact one would first exclude the falsely detected change points and then set up a full distance matrix with the set of merged moment matrices), e.g. by an hierarchical clustering algorithm [20]. 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:

$$d(C_1, C_2) = \max_{\substack{M_i \in C_1 \\ M_j \in C_2}} D_{ij},$$

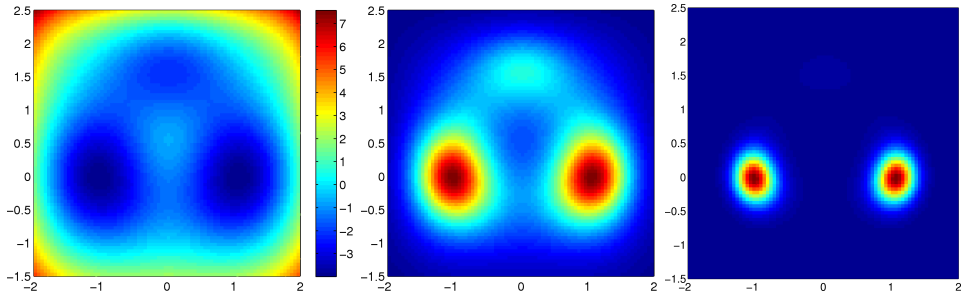


FIGURE 7.1. Left: *A surface plot of the threehole potential* $V(x, y)$ *as given in* (7.1). Middle: *The invariant measure is proportional to* $\exp(-\beta V)$*, which is plotted here for* $\beta = 0.5$ *(red marks larger values of the measure while blue corresponds to nearly zero values).* Right: *The invariant measure for a lower temperature* ($\beta = 2$).

alternatively one can define the distance between clusters by the minimal distance between any two members. The hierarchical structure appears by gradually raising the maximal distance $d_{\max}$ allowed for objects within a cluster. If $d_{\max} = 0$ all moment matrices $M_1, \ldots, M_s$ 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. After merging the moment matrices belonging to the same cluster wrt. $d_{\max}$ one would iterate the process until there is no more merging of moment matrices, an example is given in § 7.2.

7. Applications to Time-series Analysis.

7.1. The Three-well Potential. This example is the diffusion of an over-damped particle in a two-dimensional potential, so-called the three-well potential, which is defined as,

$$\begin{aligned} V(x,y) = &-3\exp(-x^2-(y-\tfrac{5}{3})^2) - 5\exp(-(x-1)^2-y^2) \\ &- 5\exp(-(x+1)^2-y^2) + 3\exp(-x^2-(y-\tfrac{1}{3})^2) \\ &+ 0.2x^4 + 0.2(y-\tfrac{1}{3})^4. \end{aligned} \tag{7.1}$$

It exhibits three minima, a shallow one at approximately $(0, 1.7)$, two deep ones at approximately $(\pm 1, 0)$, and a maximum at approximately $(0, 0.3)$. The fourth order term in (7.1) again embeds the structure in a basin with unbounded walls. This potential has been studied in [27, 34] to analyse the dynamics of diffusion processes within it, which are given by,

$$\dot{\boldsymbol{z}}(t) = -\nabla_{\boldsymbol{z}} V(\boldsymbol{z}(t)) + \sqrt{\beta}\dot{\boldsymbol{W}}(t), \tag{7.2}$$

with $\boldsymbol{z} = (x, y)$. The invariant measure of (7.2) is the Boltzmann-Gibbs distribution, which is proportional to $\exp(-\beta V)$. It can be seen in Fig. 7.1 that at lower temperatures the invariant measure concentrates in the minimal potential energy basins, while at higher temperatures it is more spreaded.

A linear SDE is expected to be a good approximation of the diffusion process (7.2) as long as it moves in the vicinity of any of the potential energy basins, since in these regions the shape of the potential energy surface is approximately quadratic.

This approximation definitely breaks down if the process switches from one basin to another basin, which it (rarely) does due to the random force. Therefore, the three-well potential should be a good test system of our change point detection algorithm. In fact, in all our trials it worked very satisfactory. As an illustration a segment of a trajectory, obtained via an Euler-Maryuama integration of (7.2), with a time discretisation step $\tau = 0.01$ and the temperature parameter set to $\beta = 2$, is depicted in Fig. 7.2. The change point detection was done with the parameters set to $t_m = t_b = t_u = 50, p = 1$ and $\alpha = 0.7$. Note that since hops between potential wells are rare events, the testing, as described in § 6, was restricted to the last 750 data points of the time series for each cycle. Also note that after detection of a change point $\hat{c}$ the detection of a subsequent change point starts at $\hat{c} + t_b$ to allow the trajectory to relax to a new potential well after leaving one.

7.2. Penta-alanine. In order to demonstrate the applicability of the precedingly presented algorithm to segment time series in a similar way as the HMM-VAR algorithm does (see § 2), we present an example from molecular dynamics (MD). 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. 7.3. The peptide is an interesting object to study, since it is a small molecule 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 purposes of our algorithm only. The trajectory was obtained from an MD-simulation in vacuum using the NWChem software package [6, 22]. 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 covering a time span of 20 nanoseconds in total. What can be seen in the trajectory is the folding of the peptide from a spread out structure where only the two long side chains interact (the salt bridge) to a more compact and very stable structure, see Fig. 7.3.

Since the dimension of the time series is higher than in our two dimensional example before we choose more conservative parameters, i.e. $t_m = t_u = t_b = 100$, and as before $p = 1$ and $\alpha = 0.7$. Choosing t_m and t_b in a range from 100 to 500 does not significantly alter the results, if they are chosen smaller, resp. larger, more, resp. less, change points will be detected, as described in § 6. Note that since we now deal with circular data the algorithm has to be adjusted such that the actual tested time series segment is shifted to make it quasi non-circular. We ellaborate on this on the next section.

7.2.1. 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. 7.3.

Obviously we have to take care of the periodic nature of the torsion angles. Since the algorithm expects the data to arrive sequentially we can just shift the data piecewise to remove periodicity, which will work in most cases since 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

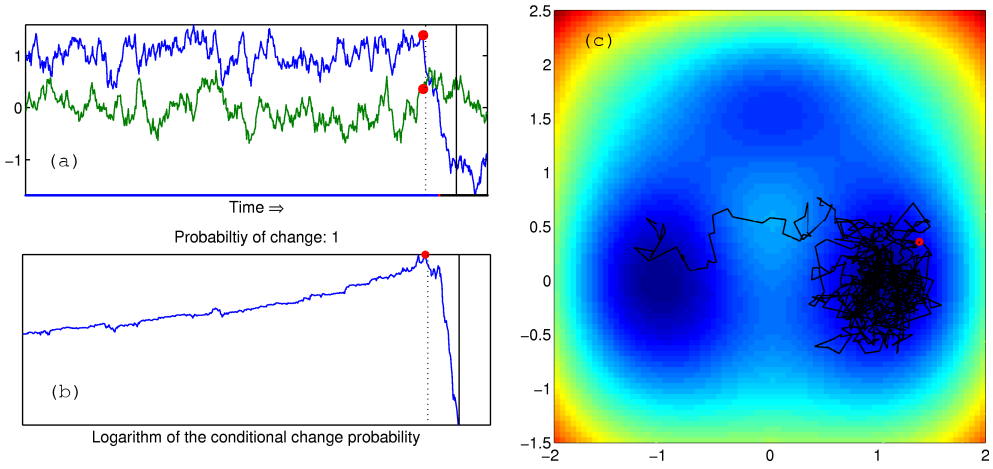
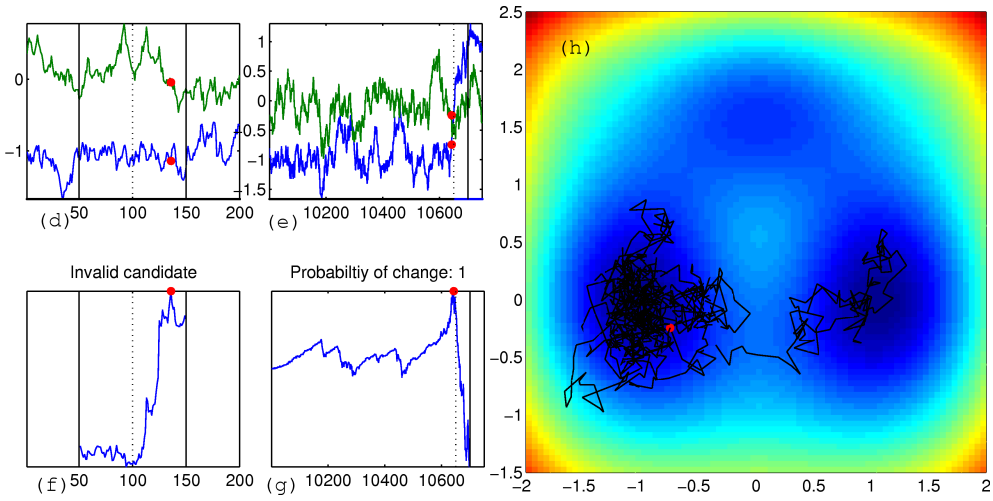


FIGURE 7.2. (a) *After moving a long time in the left basin of the potential the trajectory finally hops to the right basin.* (b) *A candidate change point $\hat{c}$ is easily identified by locating the maximum of the conditional probability (red circle = detected change point, black line = right margin defined by t_m, dotted line = puffer zone defined by t_b).* (c) *The diffusion process seen with bird's eye view in the two dimensional potential, $\hat{c}$ is marked with the red circle again.* (d) *The procedure starts again from $\hat{c} + t_b$. At first there is a left and a right margin to our test window.* (e) *The first candidate change point is invalid as it is too close to the right margin.* (f)+(g) *After iterating the algorithm many times a subsequent change point occurs and is detected.* (h) *From the bird's eye view we see that the new change point corresponds to a jump back to the vicinity of the start basin.*



boundary from our statistics, cf. Fig. 7.4.

Unfortunately, this means that we can not discard the time series data and instead use the moment matrices for post processing, as shifting the time series will alter the moment matrices in a non-reversible way (one could think of various work-arounds, like imposing restrictions on the shifting, i.e., shift the whole time series the same way), but this is no obstacle here as the time series is short enough. The change point algorithm terminates with 37 detected change points. Post-processing each data segment (as described in § 6.4), the resulting 38 segments are clustered in 23 clusters, cf. Fig. 7.5. The outcome is depicted in Fig. 7.6 and 7.7 .

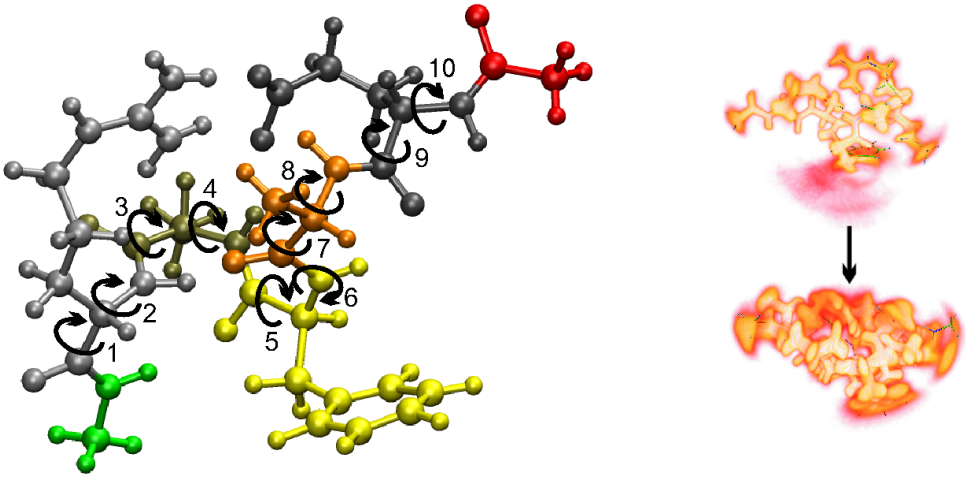


FIGURE 7.3. Left: *The simulated penta-peptide with the 10 observed torsional backbone angles marked.* Right: *During the simulation the molecule transforms from a structure where mainly the side chains interact to a more compact and stable structure via several metastable intermediates. The metastable structures at the beginning and at the end of the trajectory are visualised by density plots showing the flexibility within a conformation (Visualisation by AMIRA, [40]).*

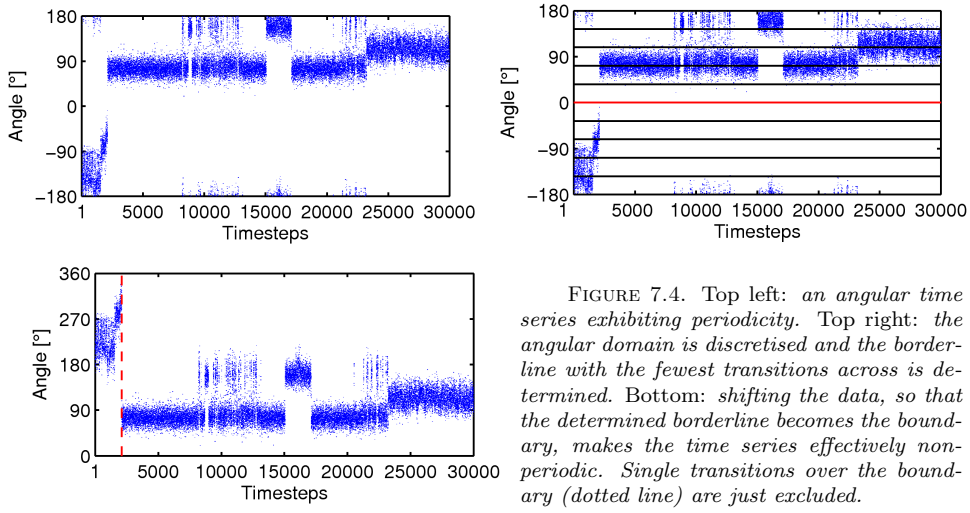


FIGURE 7.4. Top left: *an angular time series exhibiting periodicity.* Top right: *the angular domain is discretised and the borderline with the fewest transitions across is determined.* Bottom: *shifting the data, so that the determined borderline becomes the boundary, makes the time series effectively non-periodic. Single transitions over the boundary (dotted line) are just excluded.*

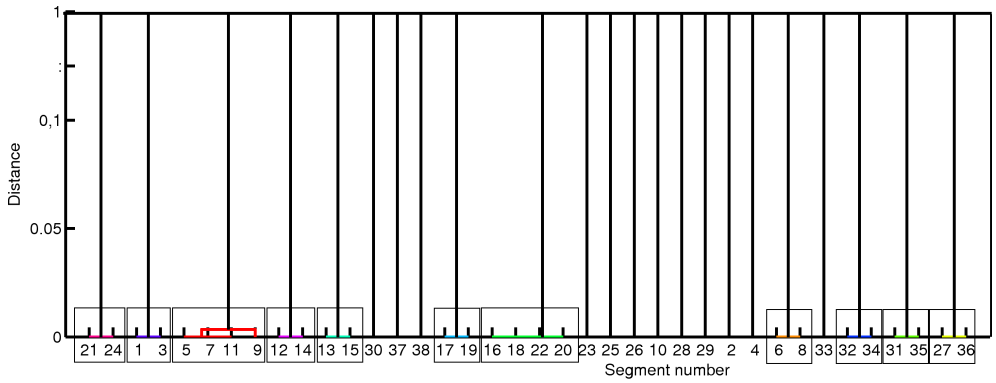


FIGURE 7.5. *In this dendrogram the allocation of the 38 identified time series segments to 23 clusters via hierarchical clustering is shown (marked by colours and boxes). The tree represents the hierarchical cluster distances using the distance measure given in* (6.1)*, however as the inter cluster distances are very close to zero while the intra cluster distances are almost one, it is not very structured.*

.

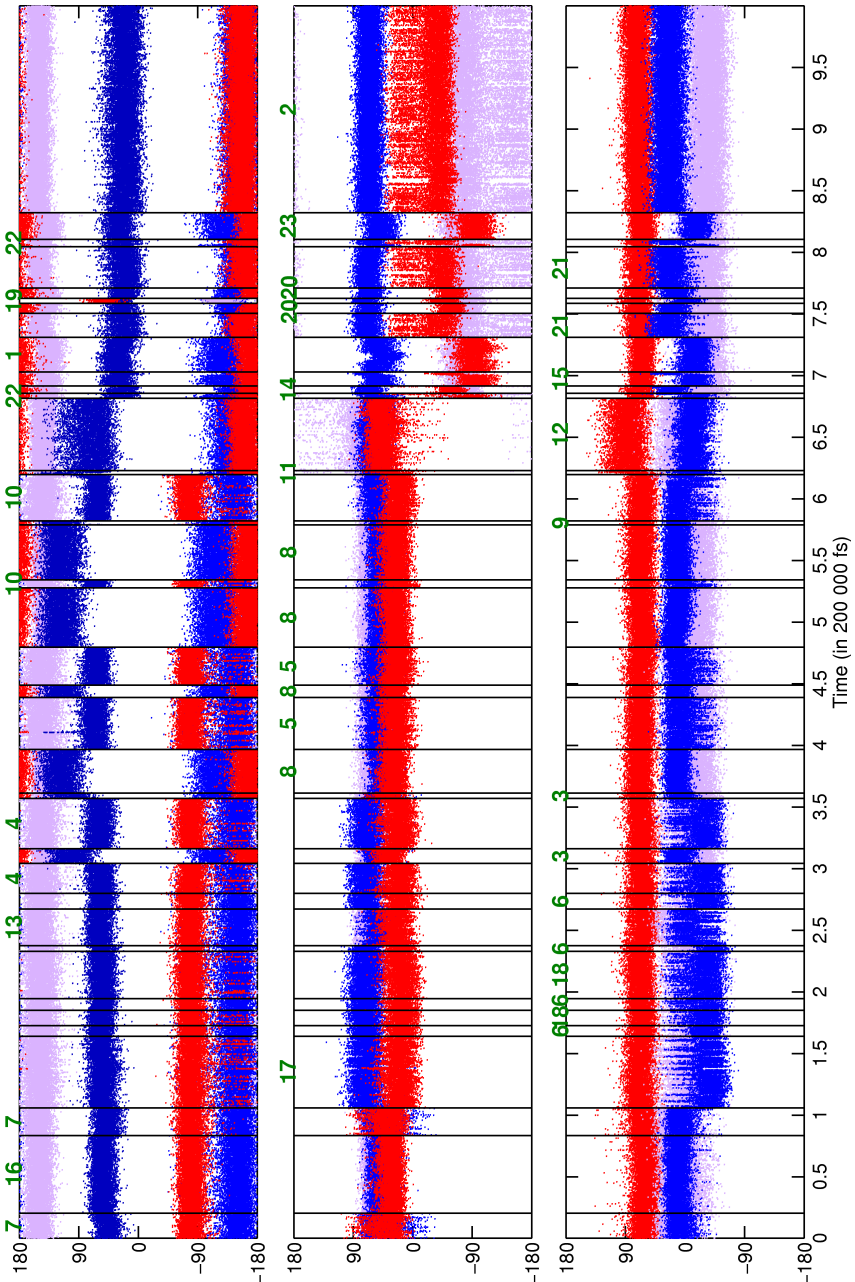


FIGURE 7.6. *The 10-dimensional backbone torsion angle time series of the peptide (splitted in 3 sub panels, Top: dimension 1-4, Middle: 5-7: Bottom: 8-10). The vertical lines mark the detected change points. The digits 1 to 23 over the panels indicate the membership of the segments to the 23 clusters obtained from hierarchical clustering as explained in the text (the digits are distributed over different panels only for reasons of readability).*

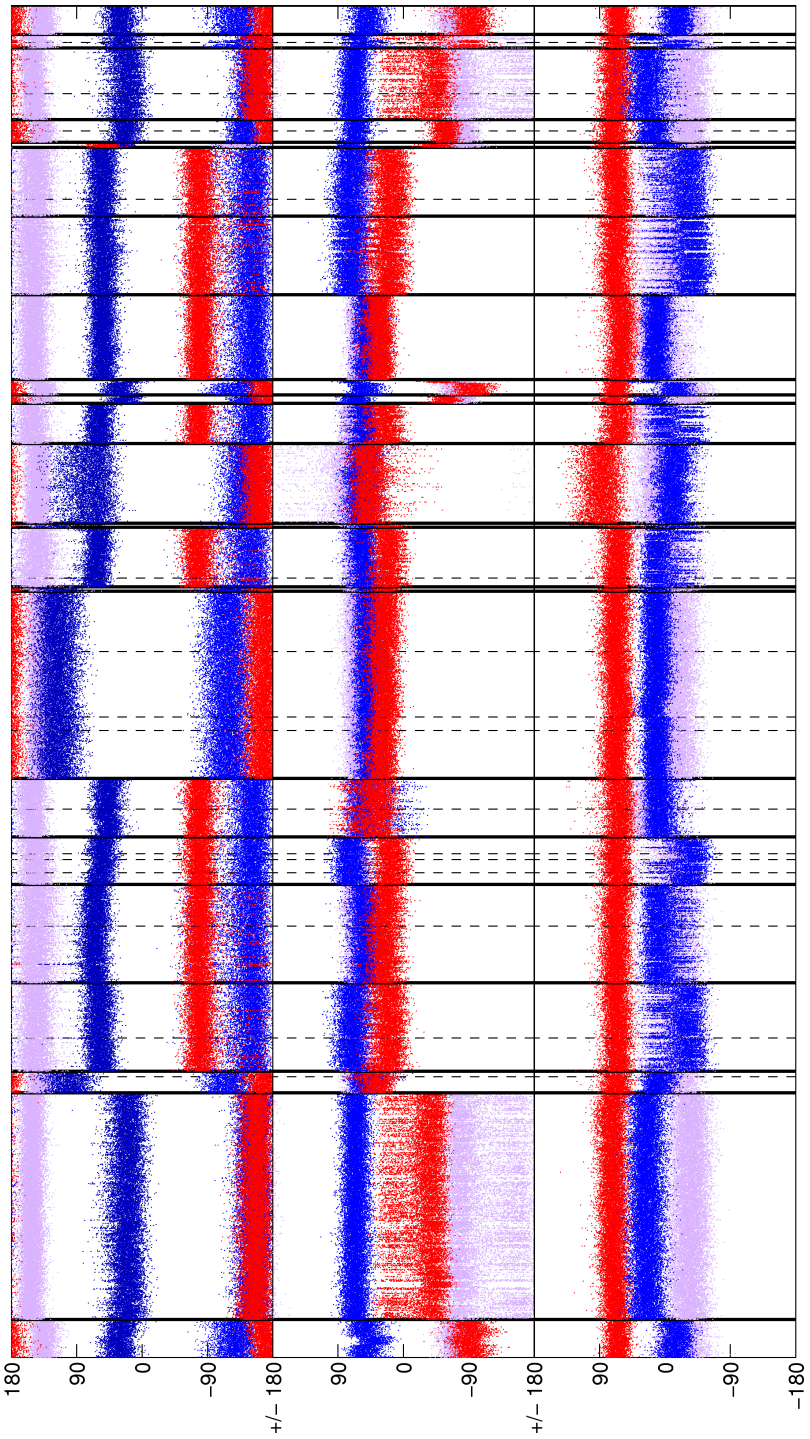


FIGURE 7.7. *Here the obtained time series segments, bordered by dashed lines, are plotted in a permutation such that the ones allocated to the same bordered by thick lines, are side by side.*

8. Conclusion. Motivated by the task to detect conformational changes in biomolecules on-line from time series, we showed how to restate the problem into a change point detection problem for VAR models. We tackled the problem by employing a Bayesian approach to model selection. Since we have assumed the lack of prior knowledge about parameters of change points we avoided the need to specify proper prior distributions using a fractional Bayes approach, which we formulated in a way to deal with the relatively high dimensional parameter space. We finally achieved an algorithmic procedure which is clear and easy to implement. In the last section we demonstrate the applicability of the procedure to actual data from molecular dynamics simulations, showing promising results. Of course, the algorithm relies on the a priori identification of significant observables in the system (such as appropriate reaction coordinates) which can be locally modeled by linear dynamics. As we have argued throughout this article, this approximation is sensible and, as we have shown, the change point detection algorithm can reliably detect persistent macroscopic transitions. The algorithm has also the advantage that it does not rely on any assumptions behind the dynamics of such transitions, such as Markovianity of the transition times between macroscopic states, as in HMM-VAR methods. In terms of the numerical complexity of the algorithm, since we encode all the information of the time series into the moment matrix M, the computational cost grows linearly in the length T of the time series. The numerical complexity in terms of the dimension d of the time series and the memory length p of the VAR-process is a different issue. Since we cannot assume a priori any sparsity properties of the moment matrix M, we must assume then that the numerical complexity in these terms is $\mathrm{O}\left(d^3p^3\right)$, since the algorithm requires a Cholesky factorization and the calculation of determinants. This algorithm is then suitable for time series of medium-size dimensionality.

Acknowledgments. We would like to thank Dr. Tomaso Frigato for performing the MD simulation of the peptide. This work is supported by the DFG Research Center Matheon "Mathematics for Key Technologies" (FZT86) in Berlin.

Appendix A. Integration of the Likelihood function. Integration of the integral in (5.1) 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. 4.4 and f the density as given in 3.9, 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,

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

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

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

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, which results in,

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

which can be simplified to,

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

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. [18, 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 (5.1) is obtained by simply using the notation introduced in § 3.4.

REFERENCES

[1] A. Amadei, A. B. Linssen, and H. J. Berendsen, *Essential dynamics of proteins.*, Proteins, 17 (1993), pp. 412–425.

[2] Ludwig Arnold, *Stochastic differential equations*, Wiley, New York, 1974.

[3] Alexander Aue, Lajos Horváth, Marie Hušková, and Piotr Kokoszka, *Change-point monitoring in linear models*, Econometrics Journal, 9 (2006), pp. 373–403.

[4] J.O. Berger and L.R. Pericchi, *The intrinsic Bayes factor for linear models*, Bayesian Statistics, 5 (1996), pp. 25–44.

[5] Jos M. Bernardo and Adrian F.M. Smith, *Bayesian Theory*, John Wiley & Sons, 1994.

[6] E. J. Bylaska, W. A. de Jong, K. Kowalski, T. P. Straatsma, M. Valiev, D. Wang, E. Aprá, 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. Früchtl, 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, and Z. Zhang, *NWChem, a computational chemistry package for parallel computers, version 5.0*, tech. report, Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA, 2006.

[7] George Casella and Roger L. Berger, *Statistical Inference*, Wadsworth & Brooks, 1990.

[8] George Casella and Elías Moreno, *Objective Bayesian variable selection*, Journal of the American Statistical Association, 101 (2006), pp. 157–167.

[9] John D Chodera, Nina Singhal, Vijay S Pande, Ken A Dill, and William C Swope, *Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics.*, Journal of Computational Chemistry, 126 (2007), p. 155101.

[10] Nicolas Chopin, *Dynamic detection of change points in long time series*, Annals of the Institute of Statistical Mathematics, 59 (2007), pp. 349–366.

[11] Peter Deuflhard, Wilhelm Huisinga, Alexander Fischer, and Christof Schütte, *Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains*, Linear Algebra and its Applications, 315 (2000), pp. 39–59.
[12] P. Deuflhard and M. Weber, *Robust Perron cluster analysis in conformation dynamics*, Linear Algebra and its Applications, 398 (2005), pp. 161–184.
[13] R. Elber and M. Karplus, *Multiple conformational states of proteins: a molecular dynamics analysis of Myoglobin.*, Science, 235 (1987), pp. 318–321.
[14] Paul Fearnhead and Zhen Liu, *On-line inference for multiple changepoint problems*, Journal of the Royal Statistical Society. Series B (Methodological), 69 (2007), pp. 589–605.
[15] Alexander Fischer, Sonja Waldhausen, Illia Horenko, Eike Meerbach, and Christof Schütte, *Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models.*, Journal of Computational Chemistry, 28 (2007), pp. 2453–2464.
[16] H. Frauenfelder, P. J. Steinbach, and Young R. D., *Conformational relaxation in proteins*, Chemica Scripta, 29A (1989), pp. 145–150.
[17] F. Javier Girón, Elías Moreno, and George Casella, *Objective bayesian analysis of multiple changepoints for linear models*, in Bayesian Statistics, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., vol. 8, Oxford University Press, 2007, pp. 1–27.
[18] A.K. Gupta and D.K. Nagar, *Matrix variate distributions*, Chapman & Hall, 2000.
[19] Illia Horenko, Carsten Hartmann, Christof Schütte, and Frank Noé, *Data-based parameter estimation of generalized multidimensional Langevin processes.*, Physical Review E, 76 (2007), p. 016706.
[20] S.C. Johnson, *Hierarchical clustering schemes*, Psychometrika, 2 (1967), pp. 241–254.
[21] Robert E. Kass and Adrian E. Raftery, *Bayes factors*, Journal of the American Statistical Association, 90 (1995), pp. 773–795.
[22] R.A. Kendall, E. Apra, D.E. Bernholdt, E.J. Bylaska, M. Dupuis, G.I. Fann, R.J. Harrison, J. Ju, J.A. Nichols, J. Nieplocha, T.P. Straatsma, T.L. Windus, and A.T. Wong, *High performance computational chemistry: An overview of NWChem a distributed parallel application*, Computer Physics Communications, 128 (2000), pp. 260–283.
[23] Oliver Lange and H. Grubmüller, *Collective langevin dynamics of conformational motions in proteins*, Journal of Chemical Physics, 124 (2006), p. 214903.
[24] Helmut Lütkepohl, *Introduction to Multiple Time Series Analysis*, Springer, Berlin - Heidelberg, 1991.
[25] E. Meerbach, *Off- and Online Detection of Dynamical Phases in Time Series*, PhD thesis, Freie Universität Berlin, 2008.
[26] Eike Meerbach, Evelyn Dittmer, Illia Horenko, and Christof Schütte, *Multiscale modelling in molecular dynamics: Biomolecular conformations as metastable states*, in Computer Simulations in Condensed Matter: Systems: From Materials to Chemical Biology. Volume I, M. Ferrario, G. Ciccotti, and K. Binder, eds., no. 703 in Lecture Notes in Physics, Springer, 2006, pp. 475–497.
[27] Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden, *Illustration of transition path theory on a collection of simple examples*, The Journal of Chemical Physics, 125 (2006), p. 084110.
[28] Arnold Neumaier and Tapio Schneider, *Estimation of parameters and eigenmodes of multivariate autoregressive models*, ACM Transactions on Mathematical Software, 27 (2001), pp. 27–57.
[29] Shawn Ni and Dongchu Sun, *Bayesian estimates for vector autoregressive models*, Journal of Business & Economic Statistics, 23 (2005), pp. 105–117.
[30] G. U. Nienhaus, J. R. Mourant, and H. Frauenfelder, *Spectroscopic evidence for conformational relaxation in Myoglobin*, Proceedings of the National Academy of Science, 89 (1992), pp. 2902–2906.
[31] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C Smith, *Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states.*, J Chem Phys, 126 (2007), p. 155102.
[32] Anthony O'Hagan, *Fractional bayes factors for model comparision*, Journal of the Royal Statistical Society. Series B (Methodological), 57 (1995), pp. 99–138.
[33] E. S. Page, *A test for a change in a parameter occurring at an unknown point*, Biometrika, 42 (1955), pp. 523–527.
[34] Sanghyun Park, Melih K. Sener, Deyu Lu, and Klaus Schulten, *Reaction paths based on mean first-passage times*, The Journal of Chemical Physics, 119 (2003), pp. 1313–1319.

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

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

[37] Christof Schütte and Illia Horenko, *Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics*, Multiscale Modeling and Simulation, (2008). Accepted.

[38] Christof Schütte and Wilhelm Huisinga, *Biomolecular Conformations can be Identified as Metastable Sets of Molecular Dynamics*, Handbook of Numerical Analysis X, Elsevier, 2003, pp. 669–744. Special Volume: Computational Chemistry.

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

[40] D. Stalling, M. Westerhoff, and 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.

[41] Dongchu Sun and Shawn Ni, *Bayesian analysis of vector-autoregressive models with noninformative priors*, Journal of Statistical Planning and Inference, 121 (2004), pp. 291–309.

[42] Arthur F. Voter, *Parallel replica method for dynamics of infrequent events*, Physical Review B, 57 (1998), pp. R13985–R13988.

[43] Arthur F. Voter, Francesco Montalenti, and Timothy C. Germann, *Extending the time scale in atomistic simulation of materials*, Annual Review of Material Research, 32 (2002), pp. 321–346.