Observation Uncertainty in Reversible Markov Chains

Philipp Metzner[1], Marcus Weber[2], Christof Schütte[1]

[1]*Department of Mathematics and Computer Science,*

Free University Berlin, Arnimallee 6, D-14195 Berlin, Germany

{*metzner,schuette*}*@math.fu-berlin.de and*

[2]*Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB),*

Takustr. 7, D-14195 Berlin, Germany

weber@zib.de

(Dated: August 10, 2010)

Abstract

In many applications one is interested in finding a simplified model which captures the essential dynamical behavior of a real life process. If the essential dynamics can be assumed to be (approximately) memoryless then a reasonable choice for a model is a Markov model whose parameters are estimated by means of Bayesian inference from an observed time series. We propose an efficient Monte Carlo Markov Chain framework to assess the uncertainty of the Markov model and related observables. The derived Gibbs sampler allows for sampling distributions of transition matrices subject to reversibility and/or sparsity constraints. The performance of the suggested sampling scheme is demonstrated and discussed for a variety of model examples. The uncertainty analysis of functions of the Markov model under investigation is discussed in application to the identification of conformations of the trialanine molecule via Robust Perron Cluster Analysis (PCCA+).

Keywords: Markov chain, transition matrix, reversibility, sampling, Gibbs sampler

I. INTRODUCTION

Markov processes provide an elegant way to model important physical properties of (especially stochastic) real-world processes such as equilibrium distributions and non-equilibrium fluctuations, effective dynamics, or reversibility.

Recent years have seen the advance of so-called Markov state models (MSM) as low-dimensional models for ergodic Markov processes on very large, mostly continuous state spaces exhibiting metastable dynamics [1–4]. Recently the interest in MSMs has drastically increased since it could be demonstrated that MSMs can be constructed even for very high dimensional systems [3] and have been especially useful for modeling the interesting slow dynamics of biomolecules [5–10] and materials [11] (there under the name "kinetic Monte Carlo"). Metastable dynamics means that one can subdivide state space into metastable sets in which the system remains for *long* periods of time before it exits *quickly* to another metastable set; here the words "long" and "quickly" mainly state that the typical residence time has to be much longer than the typical transition time so that the jump process between the metastable sets is approximately Markovian. An MSM then just describes the Markov process that jumps between the sets with the aggregated statistics of the original process.

Mathematically, a MSM is characterized by its so-called transfer operators $\{T(t)\},\ t \geq 0$ describing the evolution of the Markov model in state space. Since the state space of a MSM is finite, say $\{1, \ldots, n\}$, the family of transfer operators is given by the family of $n \times n$ transfer matrices, whose entry $T_{ij}(t)$ denotes the conditional transition probability from state i to state j in time t, for example. As usual within the Markov process framework, reversibility is captured by the detailed balance condition expressing that the probability fluxes between states of the equilibrated process are balanced.

In real-world applications, the transfer operator $T(\tau)$ for some timescale τ of interest typically is not available since the underlying dynamical process is high dimensional and nonlinear and its transition probabilities are only implicitly given by a time series $\{X_{k\tau},\ k = 0, 1, \ldots, N\}$ of observables with respect to a fixed observation time lag τ. Provided that the time series is memoryless, the parameters/entries of the matrix $T(\tau)$ are estimated from the time series by means of Bayesian inference. Typically, the observation time series allows to *approximate* the transition probabilities only. The approximation errors originate from, e.g., the finiteness of the time series, or the incompleteness of the observations. Classical results

on *a priori* estimation of such errors are quite old [12]; they state that the error decays like $1/\sqrt{N}$ asymptotically, i.e., for all N larger than some unknown large N_0, in analogy to the central limit theorem. However, N_0 cannot be characterized which makes a priori estimators rather useless in application to metastable processes (where N_0 typically is extremely large). Therefore, it is of great importance to predict the *a posteriori* statistical uncertainty of the transfer operator and, moreover, the uncertainty of *functions* or *observables* of the transfer operator like eigenfunctions, eigenvalues, correlation functions, etc.

Recently, several approaches have been introduced to a posteriori uncertainty analysis of sampling the transfer matrix of finite, homogeneous, discrete-time Markov chains or jump processes [5, 13–15]. The purpose of this article is to develop a new Monte Carlo Markov chain (MCMC) approach which is a natural extension of the approach in [15]. The approach is based on a Gibbs sampler allowing us to efficiently sample from the distribution of transfer matrices which corresponds to a given observation. Based on the resulting ensemble of transition matrices we will demonstrate the assessment of the uncertainty of functions of Markov chains, e.g., the spectrum and, more important, metastable subsets in state space.

Moreover we will show that it is important for the a posteriori analysis of uncertainty to take into account the sparsity structure of the given observation. To this end, we will introduce a new prior which is based on a penalty ansatz and allows to model the preservation of the sparsity structure of the transition matrices. The effect of that prior on the ensemble of transition matrices and observables will be discussed on simple examples. Furthermore, the effect of preserving the sparsity of the observation along with ensuring reversibility of transition matrix ensemble will be demonstrated on an example arising from molecular dynamics. In particular, we will show that the uncertainty analysis admits a systematic way to detect and characterize transition regions between molecular conformations which may help to understand conformation dynamics of biomolecular systems.

The article is organized as follows. In Section II, the necessary notation is introduced as well as the framework of Bayesian statistics for Markov chains. Section III contains the derivation of the Gibbs sampler approach to be proposed. Various numerical experiments on model examples are presented and discussed in Section IV which also includes the results of the uncertainty analysis of the identification of the trialanine dipeptide molecule. We conclude by a brief discussion of the results in Section V.

II. BAYESIAN STATISTICS FOR MARKOV CHAINS

A. Markov chains

A Markov chain $\{X_n\},\ n = 0, 1, 2, \ldots$ is a discrete time stochastic process on a finite state space, say S, such that the Markov property holds true, i.e.,

$$\mathbb{P}[X_{n+1} = j | X_n = i, X_{n-1}, \ldots, X_0] = \mathbb{P}[X_{n+1} = j | X_n = i], \quad n = 0, 1, 2, \ldots \tag{1}$$

A probability distribution $(\mu_i),\ i \in S$ with $\mathbb{P}[X_0 = i] = \mu_i$ is called initial distribution. A Markov chain is said to be *time homogeneous* if the right hand side in (1) does not depend on the time, i.e.,

$$\mathbb{P}[X_{n+1} = j | X_n = i, X_{n-1}, \ldots, X_0] = \mathbb{P}[X_1 = j | X_0 = i], \quad n = 0, 1, 2, \ldots \tag{2}$$

Consequently, a time homogeneous Markov chain is uniquely described by its transition matrix $(T_{ij}),\ i, j \in S$,

$$T_{ij} \stackrel{def}{=} \mathbb{P}[X_1 = j | X_0 = i] \quad i, j \in S, \tag{3}$$

and an initial distribution $(\mu_i),\ i \in S$. By definition, a transition matrix T is *stochastic*, i.e.,

$$T \in \mathfrak{T} \stackrel{def}{=} \left\{ T = (T_{ij})_{i,j \in S} : T_{ij} \in [0, 1], \sum_{k \in S} T_{ik} = 1 \ \ \forall i, j \in S \right\} \tag{4}$$

and an entry of T_{ij} is the conditional probability that the chain makes a transition from i to j. An initial distribution π which satisfies $\pi^\dagger T = \pi^\dagger$ is called *stationary distribution*. Throughout the paper $\dagger$ denotes the transposition operator. From now on we only consider time homogeneous Markov chains.

An important class of Markov chains is the class of *time-reversible* chains. A Markov chain is said to be (time-)reversible if the chain evolving forward in time is *statistically indistinguishable* from the chain evolving backwards in time. Formally, reversibility holds if the chain satisfies the *detailed balance* condition,

$$\pi_i T_{ij} = \pi_j T_{ji}, \quad \forall i, j \in S, \tag{5}$$

with respect to a *strict positive* probability distribution $(\pi_i),\ i \in S$. Particularly, a probability distribution π satisfying (5) is *unique and stationary*.

B. Bayesian statistics for Markov chains

The probability to observe a sample path (realization) $Y = (X_0 = y_0, \ldots, X_N = y_N)$ of a given Markov chain (T, μ) is

$$\mathbb{P}[Y|T, \mu] = \mu(y_0) \prod_{k=0}^{N-1} T_{y_k, y_{k+1}} = \mu(y_0) \prod_{i,j \in S} T_{ij}^{C_{ij}},$$

where the *transition count matrix* (C_{ij}) $i, j \in S$ is entry wise defined as

$$C_{ij} \stackrel{def}{=} \sum_{k=0}^{N-1} \delta_{y_k, i} \times \delta_{y_{k+1}, j}.$$

In this paper we are interested in the opposite question: what is the probability $P(T|Y)$ that a particular transition matrix T has generated the observed data? By virtue of the Bayesian Theorem it follows that the *posterior probability* $P(T|Y)$ is given by

$$P(T|Y) = \frac{P(Y|T)P(T)}{P(Y)}, \tag{6}$$

where $P(Y|T)$ is the *likelihood function*, $P(T)$ is the *prior* probability of transition matrices and the normalization factor $P(Y) = \int_{\mathfrak{T}} P(Y|T)P(T)\mathrm{d}T$ is called the *evidence*. The likelihood takes the form

$$P(Y|T) = \prod_{k=0}^{N-1} T_{y_k, y_{k+1}} = \prod_{i,j \in S} T_{ij}^{C_{ij}}. \tag{7}$$

The prior probability of transition matrices, $P(T)$, reflects knowledge or reasonable assumptions on the set of all transition matrices *before* observing any data. The natural candidate for the Markov chain which explains or fits a given observation best is the maximizer of the posterior and formally given by

$$T^* = \mathrm{argmax}_{T \in \mathfrak{T}} P(Y|T)P(T). \tag{8}$$

The matrix T^* is called the *maximum posterior estimator*. Whenever we refer to the uncertainty of the inferred model we mean the uncertainty of the posterior probability distribution.

a. Uniform prior. The *uniform prior* is a reasonable choice if no knowledge on $\mathfrak{T}$ is available at all. For that choice the posterior in (6) is proportional to

$$P(T|C) = \prod_{i,j \in S} T_{ij}^{C_{ij}}, \tag{9}$$

where we use the notation $P(T|C)$ instead of $P(T|Y)$ since all the required information concerning Y is contained in the transition count matrix C, see Eq. (7). Note that $P(T|C)$ is not normalized as given which is not necessary as the Gibbs-sampler derived in this article does not depend on the evidence $P(Y)$. However, $P(T|C)$ is in principal normalizable due to its polynomial form with non-negative exponents C_{ij}. To remind us of this fact let us instead use the notation

$$p_C(T) \stackrel{def}{=} \prod_{i,j \in S} T_{ij}^{C_{ij}}, \tag{10}$$

where $p_C(T)$ can be seen as a non-normalized probability density function (PDF) of the posterior. The maximum posterior estimator coincides with the unique *maximum likelihood estimator*, i.e,

$$T_{ij}^* = \frac{C_{ij}}{C_i} \quad C_i = \sum_{k \in S} C_{ik}. \tag{11}$$

It is worth to note that the PDF in (10) is proportional to the product of m *independent multivariate Dirichlet distributions*, i.e.,

$$p_C(T) \propto \prod_{i=1}^{m} \mathrm{Dir}(T_{i1}, \ldots, T_{im}; C_{i1} + 1, \ldots, C_{im} + 1). \tag{12}$$

Due to that relation we will henceforth refer to the non-normalized posterior in (10) as the Dirichlet posterior.

Example II.1. In a first example we illustrate the PDF $p_C(T)$ in (10) on a 2-state Markov chain. Let the matrix C, given by

$$C = \begin{pmatrix} 5 & 2 \\ 3 & 10 \end{pmatrix}, \tag{13}$$

be the frequency matrix associated with a fictitious finite observation Y. Let $T \in \mathbb{R}^{2\times 2}$ be a stochastic matrix, i.e.,

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

with $T_{ij} \geq 0$, $1 \leq i, j \leq 2$ and $T_{i1} + T_{i2} = 1$, $i = 1, 2$. The non-normalized PDF $p_C(T)$ associated with the observation in (13) takes the form:

$$p_C(T) = p_C(T_{11}, T_{12}, T_{21}, T_{22}) = T_{11}^5 T_{12}^2 T_{21}^3 T_{22}^{10}.$$

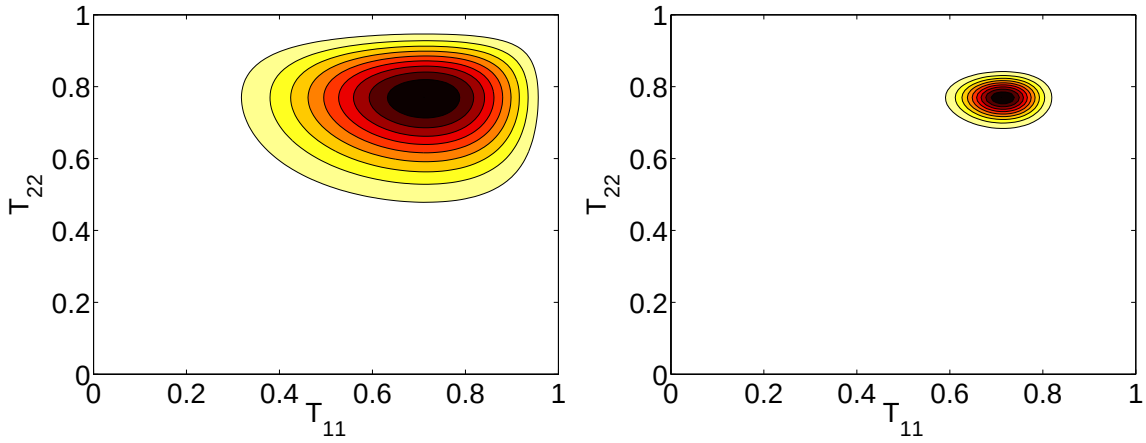

Figure 1: (Color online)Left: Probability distribution of 2x2 transition probability matrices for the observation given in (13). The resulting PDF, $p_C(T_{11}, T_{22}) = T_{11}^5(1-T_{11})^2(1-T_{22})^3T_{22}^{10}$, is shown in terms of the diagonal matrix elements. Right: The PDF $p_{C'}(T_{11}, T_{22})$ with $C' = 10C$ is illustrated. The color (gray scale) encodes the probability density; the darker the color the higher the probability.

Exploiting the stochasticity of T, $p_C(T)$ can be written as

$$p_C(T) = p_C(T_{11}, T_{22}) = T_{11}^5(1-T_{11})^2(1-T_{22})^3T_{22}^{10}, \quad T_{11}, T_{22} \in [0,1].$$

The left panel in Figure 1 illustrates the transition matrix density function $p_C(T_{11}, T_{22})$.

Next, we are interested in how the uncertainty of the transition matrix ensemble changes when we consider a longer (fictitious) observation. Intuitively, we expect that the uncertainty decreases because more knowledge of the underlying chain is available. This is indeed the case as one can see in the right panel of Figure 1; the broadness or variance of the distribution $p_{C'}(T_{11}, T_{22})$ with $C' = 10C$ is significantly smaller than the variance of $p_C(T_{11}, T_{22})$.

b. Penalty prior. In many applications, the lack of observation of a transition between states does not necessarily imply that this transition can not in principle occur. Conversely, for instance in molecular dynamics certain transitions between configuration of a molecule can never happen due to the diffusive character of the underlying dynamics. That observation reflects the inverse problem behind inferring model parameters from a finite observation. The problem becomes even clearer by looking at the maximum likelihood estimator T^* associated with a *sparse* frequency matrix C. Equation (11) shows that T^* preserves the sparse

structure of C, i.e.,

$$C_{ij} > 0 \Leftrightarrow T^*_{ij} > 0.$$

However, a transition matrix which is drawn from the uniform posterior is in general not sparse. The non-preservation of the sparse structure can significantly affect the distribution of observables as demonstrated in the following example.

Example II.2. Let the matrix C, given by

$$C = \begin{pmatrix} 1 & 100 & 0 \\ 99 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \tag{14}$$

be the frequency matrix of a fictitious observation of a three state Markov chain. Consider the maximum likelihood estimator T^* and a specific perturbation T_ϵ of it,

$$T^* = \begin{pmatrix} 0.\overline{0099} & 0.\overline{9900} & 0 \\ 0.99 & 0 & 0.01 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and } T_\epsilon = \begin{pmatrix} 0.\overline{0099} & 0.\overline{9900} & 0 \\ 0.99 & 0 & 0.01 \\ 0 & \epsilon & 1-\epsilon \end{pmatrix}. \tag{15}$$

Starting in the first state, the chain T^* frequently jumps between state one and state two before it eventually gets absorbed in the third state. Therefore, the stationary distribution is given by $\pi^* = (0, 0, 1)$. However, the perturbed chain T_ϵ is irreducible for all $\epsilon \in (0, 1]$ and possesses a strictly positive stationary distribution $\pi(\epsilon)$. The graph in Figure 2 shows $\pi_3(\epsilon)$ as a function of ϵ. It is apparent that even a small perturbation causes a large deviation of $\pi_3(\epsilon)$ from $\pi^*_3 = 1$.

Another important observable of a stochastic matrix is its spectrum. Let $\lambda_1(\epsilon), \lambda_2(\epsilon), \lambda_3(\epsilon)$ be the eigenvalues of the perturbed matrix T_ϵ ordered such that

$$1 = \lambda_1(\epsilon) > |\lambda_2(\epsilon)| \geq |\lambda_3(\epsilon)|. \tag{16}$$

The eigenvalue $\lambda_2(\epsilon)$ is called the *first nontrivial* eigenvalue of T_ϵ and provides insight in the slowest time scale of the Markov chain. One can see in the right panel of Figure 2 that a small perturbation $\epsilon > 0.01$ leads to a sign switch of $\lambda_2(\epsilon)$ which in turn indicates a significant change in the dynamics of the chain.

The example was supposed to show that even a perturbation of the occupation structure of T^* in a single entry leads to large perturbation of observables with respect to their maximum

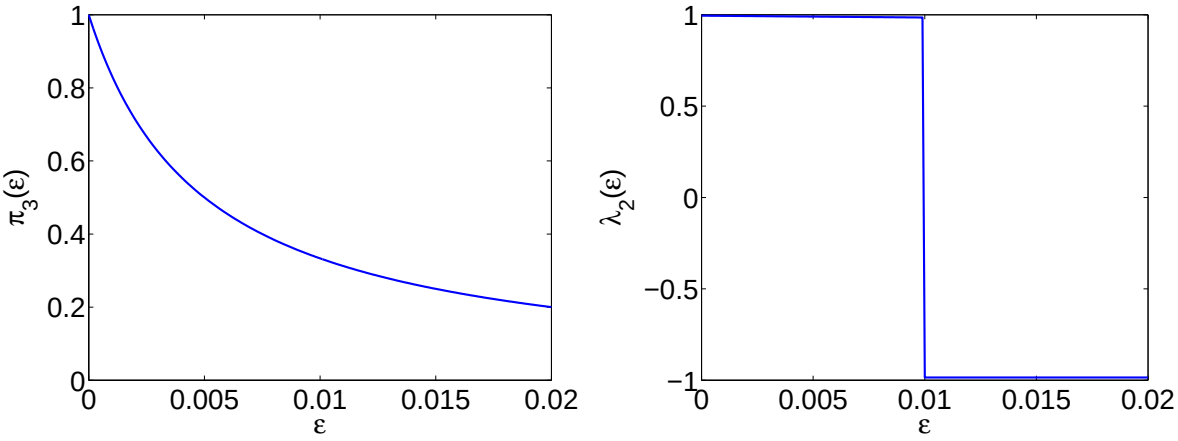

Figure 2: (Color online)The non-preservation of the sparse structure of an observation can significantly affect the distribution of observables. The panels show two observables associated with the perturbed transition matrix T_ϵ given in (15) as a function of the perturbation ϵ, respectively. Left: Stationary probability $\pi_3(\epsilon)$. Right: The first non-trivial eigenvalue $\lambda_2(\epsilon)$.

likelihood representative. Moreover, we will demonstrate in a numerical experiment (see Sect. IV B) that the non-preservation of the occupation structure of T^* might result in misleading observables' distributions. Therefore, it is desirable to suppress or rather exclude unobserved transitions from being sampled. A reasonable way to achieve that is to deploy a tailored prior.

Noé et al. suggested in [5] a prior which reflects the belief that a transition matrix entry T_{ij} is less important if no transition from i to j has been observed. Here we give an alternative construction of such a prior which is based on a penalty ansatz. The construction is motivated by the following aspects:

1. The prior $P(T)$ should only depend on transition probabilities which corresponds to non-observations, i.e.,

   $$P(T) = P(T_{l_1 m_1}, \ldots, T_{l_k m_k})$$

   with

   $$T_{l_\cdot, m_\cdot} \in I_0 \stackrel{def}{=} \{(i,j) : C_{ij} = 0\}. \tag{17}$$

2. The prior $P(T)$ should penalize any non-zero transition probability $T_{ij}, (i,j) \in I_0$ *uniformly*. In other words, $P(T)$ should penalize the *undesirable transition probability mass* $\sum_{(i,j)\in I_0} T_{ij}$ rather than a specific $T_{ij}, (i,j) \in I_0$.

3. The prior $P(T)$ should exhibit the property that the less the undesirable transition probability mass of T the more likely the corresponding transition matrix T should be, i.e.,

$$\sum_{(i,j)\in I_0} T_{ij} < \sum_{(i,j)\in I_0} \tilde{T}_{ij} \quad \Rightarrow \quad P(T) > P(\tilde{T}).$$

4. The prior $P(T) = P(T; M)$ should depend on a parameter M which allows us to scale the effect of the prior. In particular, we require for a fixed $T \in \mathfrak{T}$ with $\sum_{(i,j)\in I_0} T_{ij} > 0$

$$P(T; M_1) < P(T; M_2) \quad \forall M_1 > M_2 \geq 1.$$

A prior which satisfies these four requirements is given in

Definition II.3. For an arbitrary but fixed $M \geq 1$ we define the prior as

$$P(T; M) = \left(1 - \kappa^{-1} \sum_{(i,j)\in I_0} T_{ij}\right)^M, \tag{18}$$

where κ is the number of states which exhibit a non-observed transition to any other states, i.e.,

$$\kappa = \begin{cases} |\{i \in [m] : \exists j \in [m] \text{ with } C_{ij} = 0\}| & \text{if } |I_0| > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{19}$$

The penalty prior in (18) is non-normalized but in principal *normalizable* due to its polynomial form. The factor κ^{-1} ensures that $P(T; M) \in [0, 1]$. Consequently, we get for any fixed $T \in \mathfrak{T}$

$$\lim_{M\to\infty} P(T; M) = \chi_{\mathfrak{R}}(T), \tag{20}$$

where $\chi_{\mathfrak{R}}$ is the indicator function on the set

$$\mathfrak{R} \stackrel{def}{=} \{T \in \mathfrak{T} : \sum_{(i,j)\in I_0} T_{ij} = 0\} \subset \mathfrak{T}. \tag{21}$$

The set $\mathfrak{R}$ consists of all transition matrices which preserve the *occupation structure* of the frequency matrix C. The indicator function $\chi_{\mathfrak{R}}$ is normalizable on $\mathfrak{R}$, and, hence, can be employed as the uniform prior *restricted* to $\mathfrak{R}$.

In Section IV B we will demonstrate and discuss the effect of the penalty prior on observables' distributions.

C. Sampling

Monte Carlo strategies, in particular Monte Carlo Markov Chain (MCMC) methods, provide a powerful framework for sampling high-dimensional probability distributions. The idea behind an MCMC method is to construct a reversible Markov Chain on the high-dimensional sampling space such that its stationary distribution coincides with the target probability distribution. To be more precise, let $f : \mathbb{R}^n \mapsto \mathbb{R}$ denote the probability density function (PDF) of the distribution we want to sample from and suppose $x_{\mathcal{C}} \in \mathbb{R}^n$ is the current state. In the *proposal step* a new state $x_{\mathcal{N}} \in \mathbb{R}^n$ is generated with probability $q(x_{\mathcal{C}}, x_{\mathcal{N}})$. In the *acceptance step* the proposed state is accepted with probability

$$p_{Acc} = \min\left\{1, \frac{f(x_{\mathcal{N}})q(x_{\mathcal{N}}, x_{\mathcal{C}})}{f(x_{\mathcal{C}})q(x_{\mathcal{C}}, x_{\mathcal{N}})}\right\}.$$

If the new state is accepted, then $x_{\mathcal{N}}$ is added to the ensemble and the scheme restarts with $x_{\mathcal{N}}$ as the current state. Otherwise, the current state $x_{\mathcal{C}}$ is added to the ensemble and is considered again in the next iteration of the scheme.

The Gibbs sampler is a special kind of MCMC method. It was introduces by Geman and Geman [16] in the context of image restoration and has been applied in a wide range of applications. The key feature of the Gibbs sampler is that the Markov chain, i.e. the proposal step, is designed such that *every* generated (proposed) state is accepted. Specifically, the proposal step is constructed by using the univariate *conditional* probability density functions $f(\cdot|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), i = 1, \ldots, n$ associated with the target PDF $f : \mathbb{R}^n \mapsto \mathbb{R}$.

For the sake of notational simplicity we introduce the notation $[N] = \{1, \ldots, N\}$ and, whenever it is clear from the context, $[-i]$ refers to the set $[\cdot] \setminus \{i\}$, e.g., $x_{[-i]} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Throughout this paper when we speak of a Gibbs sampler, we are actually referring to an implementation of an iterative scheme described as follows. Let $x^{(s)} = (x_1^{(s)}, \ldots, x_n^{(s)})$ be the current sample in the s-th iteration step. The $(s+1)$-th iteration step comprises the following steps:

1. Uniformly randomly draw a coordinate i from the set $\{1, \ldots, n\}$.

2. Draw $x_i^{(s+1)}$ according to the conditional probability density function

$$f(\cdot|x_{[-i]}^{(s)}) = f(\cdot|x_1^{(s)}, \ldots, x_{i-1}^{(s)}, x_{i+1}^{(s)}, \ldots, x_n^{(s)})$$

and leave the remaining components unchanged, i.e.,

$$x_{[-i]}^{(s+1)} = x_{[-i]}^{(s)}.$$

The described sampling scheme generates a *dependent* sample from the distribution given by its PDF $f(x)$. For alternative sampling strategies in the flavor of the described scheme and a discussion on convergence see, e.g., [17].

Now the natural question arises why to prefer the Gibbs scheme over a MCMC scheme based, e.g., on purely uniformly drawn states (transition matrices) as described in [15]? The extra merits come from the observation that the latter scheme is not very efficient because it often exhibits low acceptance rates. The proposed Gibbs sampler, however, accepts every proposed state and our numerical experiments (see Section IV) show that this results in a more efficient sampling in terms of a faster convergence.

D. Uncertainty of observables

Observables of a transition matrix are important to describe and to analyze the essential dynamics of the Markov chain. For example, the spectrum, i.e. the eigenvalues and eigenvectors, allows for a decomposition of the state space into metastable sets and, hence, leads to a reduced and simplified description of the slowest processes within the Markov chain. As we will explain in more detail in Section IV C 1, conformations of a biomolecule are identified by metastable sets computed from the spectrum of the maximum likelihood estimator T^*. Therefore, it is important to assess the reliability of the resulting conformations for the further analysis.

Formally, an observable associated with a transition matrix T is expressed as a function, say $g(T)$. In general, g depends *nonlinearly* on the transition matrix which entails two consequences. First, the observable g might not attain its global maximum in T^* due to *several local maxima* or, even worse, g might not attain any local maximum in T^* at all. Hence, the identification of those scenarios is of great importance as the further analysis of the system under consideration via $g(T^*)$ would lead to misleading results and conclusions.

The Gibbs sampler presented in the previous section provides an algorithmic framework to uncertainty analysis of observables of Markov chains. In particular, it allows for the detection of the above described misleading scenarios. The algorithmic proceeding is straightforward.

In the first step, a sufficiently large ensemble of transition matrices is generated distributed according to the posterior probability distribution associated with the given time series. That ensemble in turn induces an ensemble of observables whose uncertainty can then be analyzed in a post-processing step.

III. METHOD

In this section we derive a Gibbs sampling scheme to sample from the posterior distribution resulting from the penalty prior in (18). The main result will be a Gibbs sampler which preserves reversibility.

A. Penalized Dirichlet posterior

The posterior's PDF resulting from the penalty prior takes the form

$$p_{C,M}(T) = \prod_{i,j=1}^{m} T_{ij}^{C_{ij}} \left(1 - \kappa^{-1} \sum_{(i',j')\in I_0} T_{i'j'}\right)^M \tag{22}$$

subject to

$$T \in \mathfrak{T} = \left\{ T = (T_{ij})_{i,j\in[m]} : T_{ij} \in [0,1], \sum_{k=1}^{m} T_{ik} = 1 \ \forall i,j \in [m] \right\}$$

with κ being independent of T (see Eq. (19)) and I_0 defined in (17). The main problem in deriving a MCMC scheme for drawing from (22) is to ensure the stochastic property of any proposal matrix T; beside of being entry wise non-negative, T has to satisfy the constraints

$$\sum_{j=1}^{m} T_{ij} = 1 \quad \forall i \in [m]. \tag{23}$$

Here the key idea is to explicitly insert the constraints in the posterior. In order to keep the resulting univariate conditional probability density functions (CPDFs) as simple as possible we make the substitution

$$\tilde{T}_{i,s_i} = 1 - \sum_{z\in[-s_i]} T_{iz} \quad \forall i \in [m]$$

with $s_i = \min\{j \in [m] : C_{ij} > 0\}$. To derive the (non-normalized) conditional distribution $f(x|T_{[-(k,l)]})$ with respect to an entry T_{kl} we proceed by collecting all factors involving T_{kl}.

We finally end up with the following formula for the conditional probability density functions (CPDF)

$$f(x|T_{[-(k,l)]}) = \begin{cases} x^{C_{kl}}(r_1(T_{kl}) - x)^{C_{ks_k}} \, (r_2(T_{kl}) - x)^M & \text{if } (k,l) \in I_0, \\ x^{C_{kl}}(r_1(T_{kl}) - x)^{C_{ks_k}} & \text{otherwise,} \end{cases} \tag{24}$$

subject to $x \in [0, r_1(T_{kl})]$ with

$$\begin{aligned} r_1(T_{kl}) &= 1 - \sum_{z \in [-\{l, s_k\}]} T_{kz}, \\ r_2(T_{kl}) &= \kappa - \sum_{(i',j') \in I_0 \setminus \{(k,l)\}} T_{i'j'}. \end{aligned}$$

We end this paragraph with two remarks. First note that for $M = 0$ the penalized Dirichlet posterior reduces to the posterior resulting from the uniform prior. Furthermore, the log-CPDFs decompose into a sum of a concave functions, respectively. For example, for $(k,l) \in I_0$ we have

$$\log\left(f(x|T_{[-(k,l)]})\right) = C_{kl}\log(x) + C_{ks_k}\log(r_1(T_{kl}) - x) + M\log(r_2(T_{kl}) - x). \tag{25}$$

The concavity of the log-CPDFs is essential for the efficient drawing of univariate random variables from the CPDF because it allows for a simple construction of an envelope function which is needed in the rejection-framework for sampling univariate random variables.

B. Reversible case

The main result of this paper is a Gibbs-sampling scheme which allows us to sample *reversible* transition matrices distributed according to the posterior (10). The scheme is based on results in [15] where we exploit the fact that if $K \in \mathbb{R}^{m \times m}$ is a *symmetric and non-negative* matrix then the transition matrix $T \in \mathbb{R}^{m \times m}$ element wise defined by

$$T_{ij} = \frac{K_{ij}}{\sum_j K_{ij}}$$

is a *reversible* transition matrix with respect to the probability distribution

$$\pi = \left(\frac{\sum_{j=1}^m K_{1j}}{\sum_{i,j=1}^m K_{ij}}, \ldots, \frac{\sum_{j=1}^m K_{mj}}{\sum_{i,j=1}^m K_{ij}} \right).$$

Additionally, to ensure strict positivity of π we assume that $K_i = \sum_{j=1}^{m} K_{ij} > 0,\ i = 1, \ldots, m$. This transformation maps K-matrices to transition matrices and can be formally stated as the function:

$$u(K) \stackrel{def}{=} \left(\frac{K_{11}}{K_1}, \ldots, \frac{K_{mm}}{K_m} \right) \in \mathfrak{T} \tag{26}$$

such that $T = u(K)$. The entries of a (symmetric) non-negative matrix K can be interpreted as fictitious transitions counts between states. Furthermore, the matrix $T = u(K)$ can be seen as the maximum likelihood estimator associated with these transition counts.

The crucial idea is now to generate an ensemble of symmetrical count matrices $\mathcal{K}_{sym} \subset \mathbb{R}_{+}^{m^2}$ via a Gibbs procedure which is distributed according to the PDF $p_C(T)$. To be more precise, we derive a Gibbs procedure to sample symmetrical count matrices distributed according to

$$\tilde{p}_C(K) \stackrel{def}{=} p_C(u(K)) = \Pi_{i,j=1}^{m} \left(\frac{K_{ij}}{\sum_{k=1}^{m} K_{ik}} \right)^{C_{ij}}. \tag{27}$$

It is shown in the Appendix (see also [15]) that if $\mathcal{K}_{sym}$ is restricted on the set

$$\mathfrak{K}_{sym} = \left\{ K \in R_{+}^{m^2} : K_{ij} = K_{ji}\ \forall i, j \in [m], k^- \leq \sum_{i,j=1}^{m} K_{ij} \leq k^+ \right\}, \tag{28}$$

with $0 < k^- < k^+ < \infty$ then the ensemble of reversible transition matrices $\mathfrak{T} = \{u(K) : K \in \mathcal{K}_{sym}\}$ is distributed according to $p_C(T)$.

For the derivation of the CPDFs associated with the likelihood function in (27) we proceed analogously as in Section III A and finally get

$$f(x|T_{[-(k,l)]}) = \begin{cases} x^{(C_{kl}+C_{lk})} \left(r_k + x\right)^{-C_k} \left(r_l + x\right)^{-C_l} & \text{if } k \neq l, \\ x^{C_{kl}} \left(r_k + x\right)^{-C_k} & \text{otherwise,} \end{cases} \tag{29}$$

where $r_i = \sum_{z \in [-\{l\}]} K_{iz}$ and, e.g., $C_i = \sum_{z=1}^{m} C_{iz}$. The constraint in (28) confines the CPDF in (29) on the finite interval

$$[a, b] = \begin{cases} 0.5 \left[\max\left\{k^- - S_{(k,l)}, 0\right\}, k^+ - S_{(k,l)} \right] & \text{if } k \neq l, \\ \left[\max\left\{k^- - S_{(k,k)}, 0\right\}, k^+ - S_{(k,k)} \right] & \text{otherwise,} \end{cases} \tag{30}$$

with $S_{(k,l)} = \sum_{(i,j) \in [-\{(k,l),(l,k)\}]} K_{ij}$.

IV. NUMERICAL EXPERIMENTS

In this section we demonstrate our Gibbs sampler derived in the previous section on various examples. The purpose of the first example is to explain the need for the restriction

of the transition matrix ensemble to transition matrices preserving the occupation structure of the given observation. Furthermore, we comment in detail on the choice of sampling parameters, e.g. the burn-in time and thinning step, and we compare the efficiency and the speed of convergence of the Gibbs sampler with the MCMC method presented in [15].

The focus of the remaining section is on sampling of reversible transition matrices. Particularly, we will investigate the uncertainty in the identification of conformations of a molecule computed via the Perron Cluster Cluster Analysis (PCCA+) scheme [2, 18, 19].

But before starting with the first example, let us comment on some computational aspects of the Gibbs sampler and on sampling parameters as well as on tests for convergence of the sampling scheme.

A. Computational aspects and choice of sampling parameters

Numerically, sampling of a high dimensional PDF via a Gibbs sampler boils down to sampling from univariate probability distributions which can efficiently be performed by standard methods, e.g., adaptive rejection sampling (ARS) [20], adaptive rejection metropolis sampling [21]. Throughout our numerical experiments, we used the *concave convex adaptive rejection sampling method* (CCARS) [22] since all resulting log-CPDFs, e.g. $\log[f(x|T_{[-(k,l)]})]$, are decomposable into sums of concave and convex functions. The crucial idea behind CCARS is to exploit this property of the log-CPDFs in order to adaptively construct an envelop function which is needed in the acceptance-rejection algorithm for drawing from univariate random variables. Another important feature of the acceptance-rejection algorithm, particularly of CCARS, is that the PDF to be sampled from does not have to be normalized.

Next, we comment on our choice for the sampling parameters. In order to ensure uncorrelated samples we actually start storing samples after a certain *burn-in time*. Moreover, rather than storing every updated sample we took every n^{th} sample where we used the thumb rule $n \approx m^2$ with m being the number of states.

For testing on convergence, we employed Gelman and Rubin's convergence diagnostic [23]. We considered a transition matrix ensemble $\mathcal{T}$ to be converged when all entries T_{ij} are converged indicated by their *potential scale factor* [24] $\hat{R}(T_{ij})$ being close to one, respectively.

To be more precise, we stopped sampling when the following criterion was fulfilled

$$R(\mathcal{T}) \stackrel{def}{=} \max_{i,j\in[m]}\{|1-\hat{R}(T_{ij})|\} << 1. \tag{31}$$

B. Effect of the penalty prior

In the first numerical example we study the effect of the penalized Dirichlet posterior in (22) on the transition matrix ensemble which arises from a fictitious observation of a three state Markov chain. To this end we re-visit Example II.2 and consider the frequency matrix given in (14). In particular, we are interested in the distribution of the stationary probability of state 3, π_3, and the distribution of the first nontrivial eigenvalue λ_2 as a function of the penalty exponent M.

For a sequence of penalty exponents $M = (0, 100, 1000, 10000)$, we generated by means of our Gibbs sampler scheme an ensemble of 10^6 3×3 transition matrices distributed according to the penalized Dirichlet posterior in (22), respectively. We did not enforce reversibility of the transition matrices. The distributions of the observables mentioned above are depicted in Figure 3. The left panel shows the distributions of π_3. For the sake of illustration we normalized the distributions of π_3 (left panel) such that their maximum value is one, respectively. Recall that the third state in the maximum likelihood chain associated with the observation in (14) is an absorbing state with stationary distribution $\pi_3^* = 1$. As already pointed out in Example II.2, the distribution of π_3, more precisely its mean value, resulting from the uniform prior ($M = 0$) is far away from one. In other words, a transition matrix with stationary probability distribution $(0, 0, 1)$ is extremely unlikely in the ensemble of dense transition matrices distributed according to the Dirichlet posterior in (10). However, one can see that increasing M results in a right shift of the distribution towards $\pi_3^* = 1$. The distribution of the first nontrivial eigenvalue as a function of the penalty exponent exhibits the same behavior (see the right panel of Figure 3). The decrease of the average undesirable probability mass with increasing M, $< \sum_{(i,j)\in I_0} T_{ij} >$, is depicted in Figure 4.

These observations support the idea that preserving the occupation structure of the frequency matrix C is essential for the assessment of the uncertainty of observables. Thus the next natural question is how to choose the penalty exponent M in applications? Clearly, it is desirable to choose the actual value of M in an automatic way since M is an additional parameter and in turn introduces additional uncertainty into the model. In fact, our simple

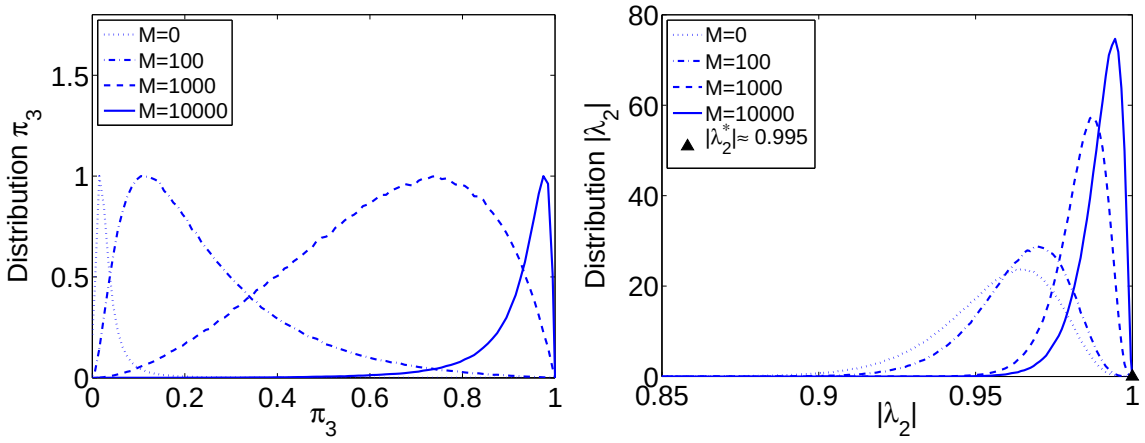

Figure 3: (Color online)Distributions of observables resulting from a sampling of the posterior in (22) associated with the frequency matrix in (14). The left panel shows the distribution of the stationary probability of the third (absorbing) state π_3. For the sake of illustration, we normalized the distributions of π_3 such that the maximum value is one, respectively. The right panel shows the distribution of the modulus of the first nontrivial eigenvalue $|\lambda_2|$. Furthermore, the panels show the distributions' dependence on the penalty exponent M: the larger is M the more shifted the distributions are towards the values $\pi_3^* = 1$ and $|\lambda_2^*| \approx 0.995$ resulting from the maximum likelihood estimator T^* associated with (14).

example from above as well as the real-world application to be discussed in Section IV C 1 have shown that the higher the value of M is the higher the statistical weight of the maximum likelihood estimator T^* within the ensemble of transition matrices $\mathcal{T}_M$. Consequently, choosing $M = \infty$, i.e., applying the restricted uniform prior introduced in (20), seems a reasonable choice (and will be applied in further numerical experiments in Section IV C 1 below). Finally note that from a more formal point of view the penalty prior provides a mathematically consistent justification for considering the restricted uniform prior since it shows that the restricted uniform prior results from a limit process of normalizable priors.

C. Reversible Markov chains

In this section we are interested in distributions of observables arising from ensembles of *reversible* transition matrices, i.e. transition matrices which fulfill the detailed balance condition in (5). We will first discuss the effect of reversibility on the distribution of the

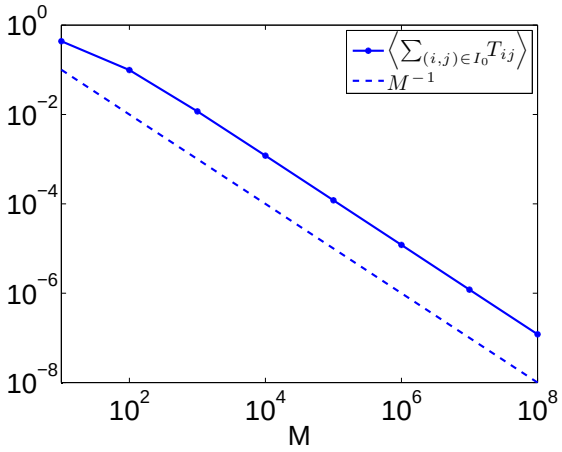

Figure 4: (Color online)The figure shows the double logarithmic plot of the *average* of undesirable transition probability mass, $< \sum_{(i,j)\in I_0} T_{ij} >$, as a function of the penalty parameter M. The respective averages were computed by sampling the posterior in (22) associated with the frequency matrix in (14). The double logarithmic plot reveals that the undesirable transition probability mass decreases proportionally to M^{-1}.

transition matrix ensemble itself. Furthermore, we will compare the convergence of the Gibbs sampler with that of the MCMC method introduced in [15] as well as the computational effort of both methods in terms of the running time. The main result will be the numerical investigation of the uncertainty in the identification of conformations of the biomolecule trialanine.

c. Two-state Markov chain. Even in the simplest case - a two-state Markov chain - enforcing reversibility substantially affects the transition matrix ensemble. To demonstrate that effect we generated an ensemble of 2×2 reversible transition matrices based on the frequency matrix given in (13) (cf. Example II.1). The distributions of the diagonal entries T_{11} and T_{22} are depicted in Figure 5 together with the distributions arising from the unrestricted ensemble. For the sake of illustration we normalized the distributions such that their maximum value is one, respectively. One can clearly see that the distributions significantly differ. The deviation can be explained by the fact that *not every* 2×2 Markov chain is automatically reversible. For example, a transition matrix of the form

$$\begin{pmatrix} \alpha & 1-\alpha \\ 0 & 1 \end{pmatrix} \tag{32}$$

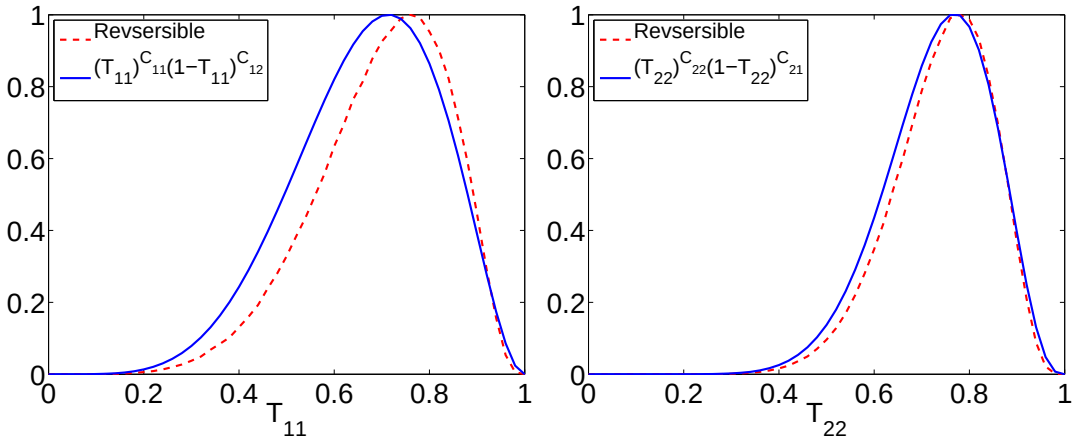

Figure 5: (Color online)Restricting the transition matrix ensemble to reversible transition matrices significantly affects distributions of observables. In the panels, we illustrate the distributions of the diagonal entry T_{11} (left) and T_{22} (right) based on the frequency matrix given in (13). The reversible ensemble (dashed lines) clearly differs from the unrestricted ensemble (solid line).

with $\alpha \in [0, 1]$ is not reversible since its stationary distribution, $\pi = (0, 1)$, is not strictly positive and, consequently, the off-diagonal entries cannot be recovered via the detailed balance condition (cf. Eq. (5)).

d. Three-state Markov chain. Next, we study the speed of convergence of our method and compare it to the MCMC method introduced in [15]. To this end we consider the frequency matrix C given by

$$C = \begin{pmatrix} 1 & 10 & 2 \\ 2 & 26 & 3 \\ 15 & 20 & 20 \end{pmatrix}, \tag{33}$$

and generate two ensembles of 3×3 reversible transition matrices; one ensemble with the Gibbs sampler and the second one with the MCMC sampler. Both ensembles have the same size (10^7 transition matrices) and we used the same boundary parameters ($k^- = 0.9, k^+ = 100$) as well as the same sampling parameters including the same initial matrix.

To assess the speed of convergence of both methods, we evaluate the function $R(\mathcal{T}_N)$ introduced in (31) for a nested sequence of sub-ensembles $\{\mathcal{T}_N\}$, $N = 10^3, \ldots, 10^7$, respectively. The double logarithmic plot of $R(\mathcal{T}_N)$ (left panel in Figure 6) based on the Gibbs- and MCMC ensemble reveals that the Gibbs sampler converges approximately one order of magnitude faster than the MCMC sampler. However, the price for the faster convergence is

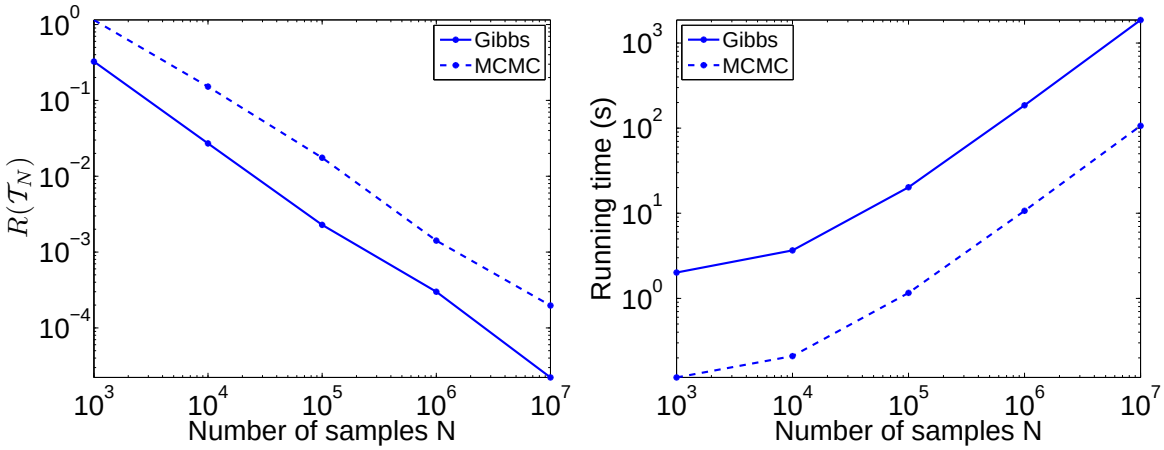

Figure 6: (Color online)Left: The evaluation of the function $R(\mathcal{T}_N)$ (see Eq. (31)) for a nested sequence of reversible ensembles $\{\mathcal{T}_N\}$, $N = 10^3, \ldots, 10^7$ indicates that the Gibbs sampler converges approximately one order of magnitude faster than the MCMC sampler. Right: Running time of the Gibbs sampler and the MCMC sampler. For details see text.

an approximately one order of magnitude longer running time than the MCMC scheme (see right panel in Figure 6). The increased running time is due to the fact that in each proposal step of the Gibbs sampler a univariate and non-uniform density has to be sampled whereas in the MCMC proposal step the updated entry is uniformly distributed.

1. Molecular Dynamics : Trialanine

Molecular dynamics is a reversible process X_t with positive invariant measure μ given by the Boltzmann distribution. For every spacial discretization of the molecule's state space $\mathcal{S}$ into n disjoint sets $B_1, \ldots, B_n$, $\cup_i B_i = \mathcal{S}$ the transition matrix

$$T_{ij} = \mathbb{P}_\mu[X_\tau \in B_i | X_0 \in B_j]$$

is reversible with respect to the coarse grained measure $\mu(B_i) = \int_{B_i} \mu(x)\mathrm{d}x$. In applications, however, we can only approximate the transition matrix T_{ij} based on finite observations of the chain. The resulting approximate transition matrices (e.g., the maximum likelihood estimator based on the observations) usually do not fulfill the detailed balance condition because, e.g., the transition count matrix associated with the observation in general is not symmetric.

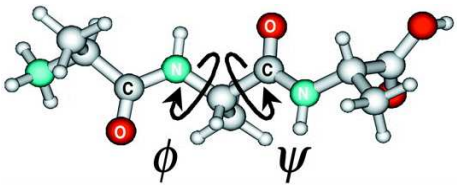

Figure 7: (Color online)The figure shows the ball-and-stick representation of the trialanine dipeptide analog. The torsion angles Φ and Ψ have been proven to be the right order parameter to describe the conformation dynamics.

In this example we study the uncertainty in the identification of conformations of the trialanine molecule which is shown in ball-and-stick representation in Figure 7. The process is implicitly given by a time series of two torsion angles Φ and Ψ. For the analysis of transition pathways between the conformations of the trialanine molecule based on the time series used herein see [25, 26]. This is a realistic application of our uncertainty estimation technique; we are however not mainly interested in new physical insights into the molecule's properties but in a validation of our technique and in a demonstration of how it can be applied.

The time series of trialanine was generated in vacuum using the Hybrid Monte Carlo method [27] with 544.500 steps with GROMACS force field [28, 29] at a temperature of 750K. The integration of the sub-trajectories of the proposal step were realized with $\tau = 1$ fs time steps of the Verlet integration scheme. The top panel of Figure 8 shows the projection of the time series (all atomic positions) onto the torsion angle space spanned by Φ and Ψ.

Here, the torsion angle space $(\Phi, \Psi) \in [0\ 360] \times [-180\ 180]$ is discretized into 30×30 equidistantly sized boxes. Since not all boxes are visited by the time series we ended up with an 447 state Markov chain associated with a 447×447 frequency matrix C. For the sake of illustration we depict in the right bottom panel of Figure 8 the discrete free energy, $-\log \pi_i^*$, associated with the stationary distribution of the maximum likelihood chain T^*. The brightness of a box encodes the probability to encounter the chain; The lighter the color the more probable to encounter the process in that box.

a. Robust Perron Cluster Analysis (PCCA+). PCCA+ is an algorithm which can be used to identify metastable subsets of a reversible Markov chain T. The clustering of the states of the Markov chain is done in terms of a membership matrix $\chi(i, x)$, where i is the index of molecular states and $x = 1, \ldots, m$, is the index of a metastable (fuzzy) subset. The

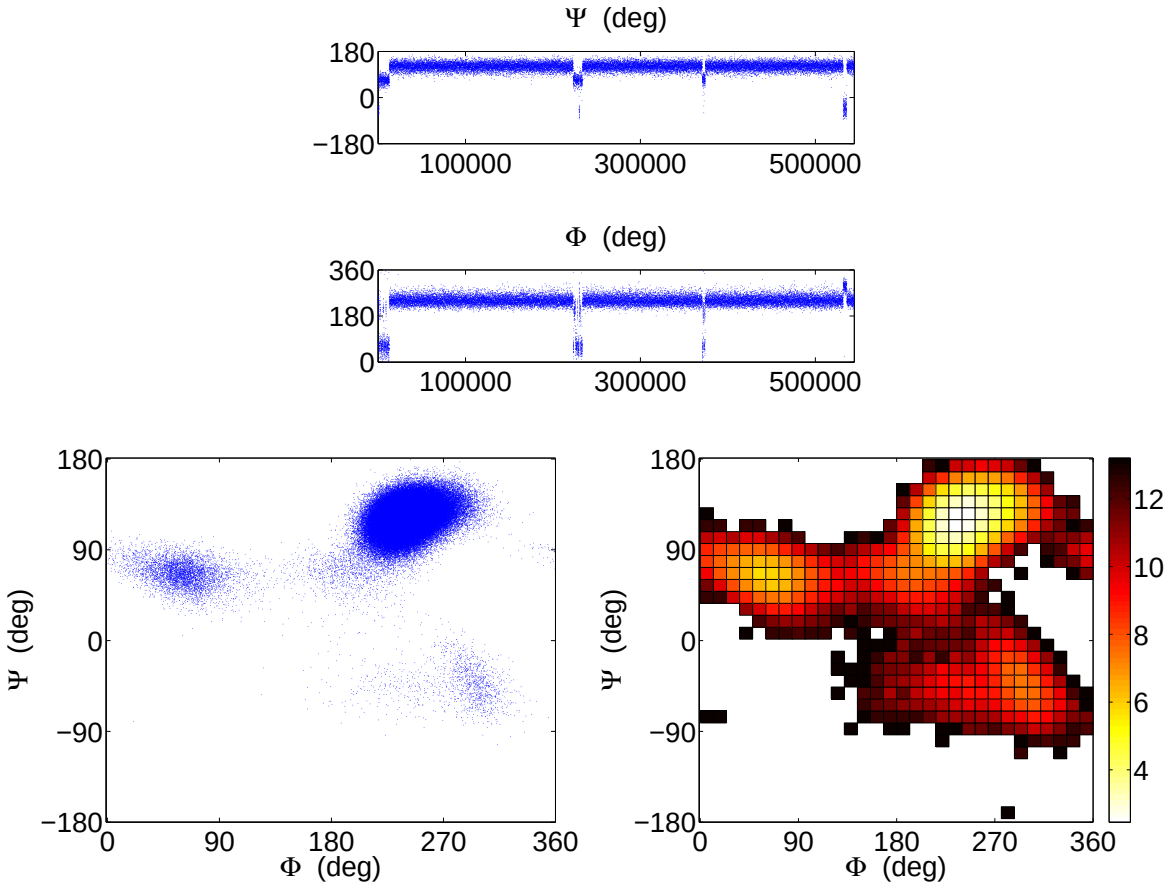

Figure 8: (Color online)In this figure we exemplify our strategy to capture the essential dynamics of a bio-molecule in a coarse grained model. The top panel shows the projection of the time series (all atomic positions) onto the torsion angle space spanned by Φ and Ψ, which reveals the metastable behavior. Left bottom: The Ramachandran plot of the torsion angle time series. At first glance, trialanine attains three different conformations indicated by the three clusters. Right bottom: The discrete free energy, $-\log \pi^*$, associated with the stationary distribution π^* of the Markov chain T^* which models the effective dynamics of the system in terms of the torsion angles Φ and Ψ. The chain was constructed from the underlying time series with respect to a 30×30 box discretization of the torsion angle space. The lighter the color of a box the more probable to encounter the process in that box.

matrix element $\chi(i, x) \in [0, 1]$ denotes the grade of membership of state i to the metastable subset x where $\chi(i, x)$ close to 1 indicates a unique assignment of state i to cluster x and close to 0 a unique non-assignment. Conversely, if $\chi(i, x)$ is not near 1 or 0 then state i can not be assigned uniquely to one of the metastable subsets x. If all states are assigned

λ_1	λ_2	λ_3	λ_4
1	$9.9599 \cdot 10^{-1}$	$9.9576 \cdot 10^{-1}$	$9.1248 \cdot 10^{-1}$

Table I: The first four dominant eigenvalues of $T^*(C')$. The spectral gap between the third and the fourth eigenvalue indicates an optimal decomposition of the state space into three metastable subsets as illustrated in Figure 9

(almost) uniquely, we call the membership function "hard".

In PCCA+ the columns of the matrix χ are computed as linear combinations of the m dominant eigenvectors of the transition matrix T under investigation. This linear transformation is done in such a way that the row sums of χ are 1 (partition of unity) and that the entries of χ are non-negative. Since there are many feasible linear transformations, PCCA+ identifies one optimal transformation with regard to an objective function [19, 30, 31]. In our example, we maximize the metastability, more precisely, we maximize the sum of the diagonal elements of the Markov chain T.

b. Uncertainty in conformations' identification. The frequency matrix C associated with a given observation of a reversible process is in general not symmetric and, thus, it is not guaranteed that the maximum likelihood estimator chain $T^*(C)$ fulfills the prerequisite of PCCA+, i.e., real-valued eigenvalues and eigenvectors. In practice, it is common to analyze the reversible maximum likelihood estimator chain $T^*(C')$ arising from the *symmetrized* frequency matrix $C' = C + C^\dagger$ where $C^\dagger$ is the transpose of C.

The first four eigenvalues of $T^*(C')$ (for the torsion angle time series) are given in Tab. I. The gap between the second and third eigenvalue of $T^*(C')$ indicates an optimal decomposition of the state space into three metastable sets. In the remainder of the section we will study the uncertainty of the identification of three metastable sets.

In order to obtain a hard membership function $\chi(i),\ i \in S$, it is common to assign a state i to the cluster with maximal affiliation probability, i.e.,

$$\chi_H(i) \stackrel{def}{=} \text{argmax}_{x\in\{1,2,3\}}\chi(i,x), \tag{34}$$

where $\chi(i,x)$ is the membership function resulting from PCCA+. For an illustration of $\chi_H(\cdot)$ for $T^*(C')$ see the left panel of Figure 9. Note that the assignment in (34) does not depend on the actual maximum value. Consequently, in the worst case, i.e. $\chi(i,1) \approx \chi(i,2) \approx \chi(i,3) \approx 1/3$, the assignment would be meaningless. To capture these cases we considered

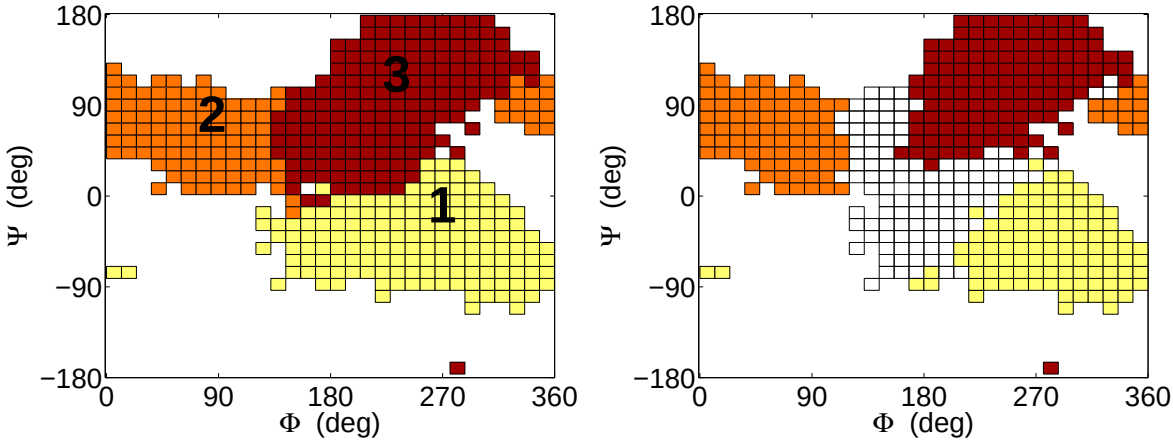

Figure 9: (Color online)Left: Decomposition of the state space of the torsion angle space into three metastable sets via PCCA+ based on the maximum likelihood estimator $T^*(C')$ associated with symmetrized frequency matrix $C' = C + C^\dagger$. Boxes (states) with the same color (gray scale) belong to the same metastable set (conformation). Right: Decomposition of the state space via the *regularized* hard assignment in (35) with respect to the regularity parameter $\gamma = 0.9$. The states corresponding to "empty" boxes could not be assigned to any metastable set at all.

a *regularized assignment*,

$$\chi_H(i;\gamma) \stackrel{def}{=} \begin{cases} \text{argmax}_{x\in\{1,2,3\}}\chi(i,x) & \text{if } \max_{x\in\{1,2,3\}}\chi(i,x) > \gamma, \\ \text{"mark as non-assigned"} & \text{otherwise,} \end{cases} \tag{35}$$

where the threshold $\gamma \in [0,1]$ can be interpreted as a regularity parameter for the assignment. Using the parameter $\gamma = 0.9$, the decomposition of $T^*(C')$ based on $\chi_H(i;\gamma)$ as shown in the right panel of Figure 9 suggests the following interesting interpretation: the three clusters (indicated by different colors) are *core* metastable subsets whereas the non-assigned states (indicated by empty boxes) form a transition region through which the transitions between the core clusters happen. Now the obvious question arises of how reliable the decomposition is when based on the maximum likelihood estimator $T^*(C')$? More precisely: what is the uncertainty of the identification of the core cluster and the transition region?

For the numerical investigation of the uncertainty we proceed analogously as in the previous sections. We generated an ensemble $\mathcal{T}_{Rev}$ of 500.000 reversible transition matrices $(k^- = 0.9, k^+ = 10000)$ restricted on the set specified in (36). Next, we computed for all $T \in \mathcal{T}_{Rev}$ via PCCA+ the membership matrix $\chi(T) = (\chi(i,x)), \in i \in S, x \in \{1,2,3\}$ with

$\chi(i,x)$ being the grade of membership that state i belongs to cluster x. Based on landmarks, i.e. distinctive states in each of the three clusters associated with $T^*(C')$, we ensured that, e.g., cluster $x = 1$ always corresponds to the reference metastable set M_1 associated with $T^*(C')$.

Before we describe the numerical experiment in more detail, we shall comment on how we restricted the sampling on reversible transition matrices while preserving the occupation structure of the underlying time series. At the first glance it seems reasonable to generate an ensemble of reversible transition matrices based on the symmetrized frequency matrix C'. Doing so, however, would lead to a biased ensemble because C' does not reflect the given observation. Sampling with respect to the "true" frequency matrix C poses another problem. In the previous section we have seen that preserving the occupation structure of the frequency matrix is necessary to obtain meaningful distributions of observables. On the other hand, sampling reversible transition matrices amounts to sampling symmetric frequency matrices. However, in general and, particularly in this example, the occupation structure of a frequency matrix C is not symmetric, i.e.

$$\exists (i,j) : C_{ij} > 0, C_{ji} = 0.$$

The minimal compromise between preserving the occupation structure and symmetry is sampling the PDF in (27) subject to

$$K \in \mathfrak{K}_{sym} \cap \{K \in \mathbb{R}^{m \times m} : K_{ij} = 0 \text{ if } C_{ij} = C_{ji} = 0\}. \tag{36}$$

Based on the resulting ensemble $\mathcal{T}_{Rev}$ of 500.000 of reversible transition matrices and with the assignment based on (35) at hand, we computed from the ensemble of membership matrices $\{\chi(T) : T \in \mathcal{T}_{Rev}\}$ the assignment/non-assignment histograms of the states by counting how often a state i is assigned to a cluster x and how often i is not assigned at all. Finally, from these histograms we derived the conditional probability distributions $p_x(i),\ i \in S$ with $p_x(i)$ is the probability of a state i being assigned to cluster x conditional on being in cluster x. Analogously, we derived the distribution $p_-(i),\ i \in S$ with respect to the non-assigned states.

In Figure 10 we illustrate p_1, p_2 and p_3 for $\gamma = 0.9$. First of all, comparing these distributions, one can see that the majority of states either are always assigned to the same cluster or are not assigned to any cluster at all. Thus, the uncertainty of the regularized assignment

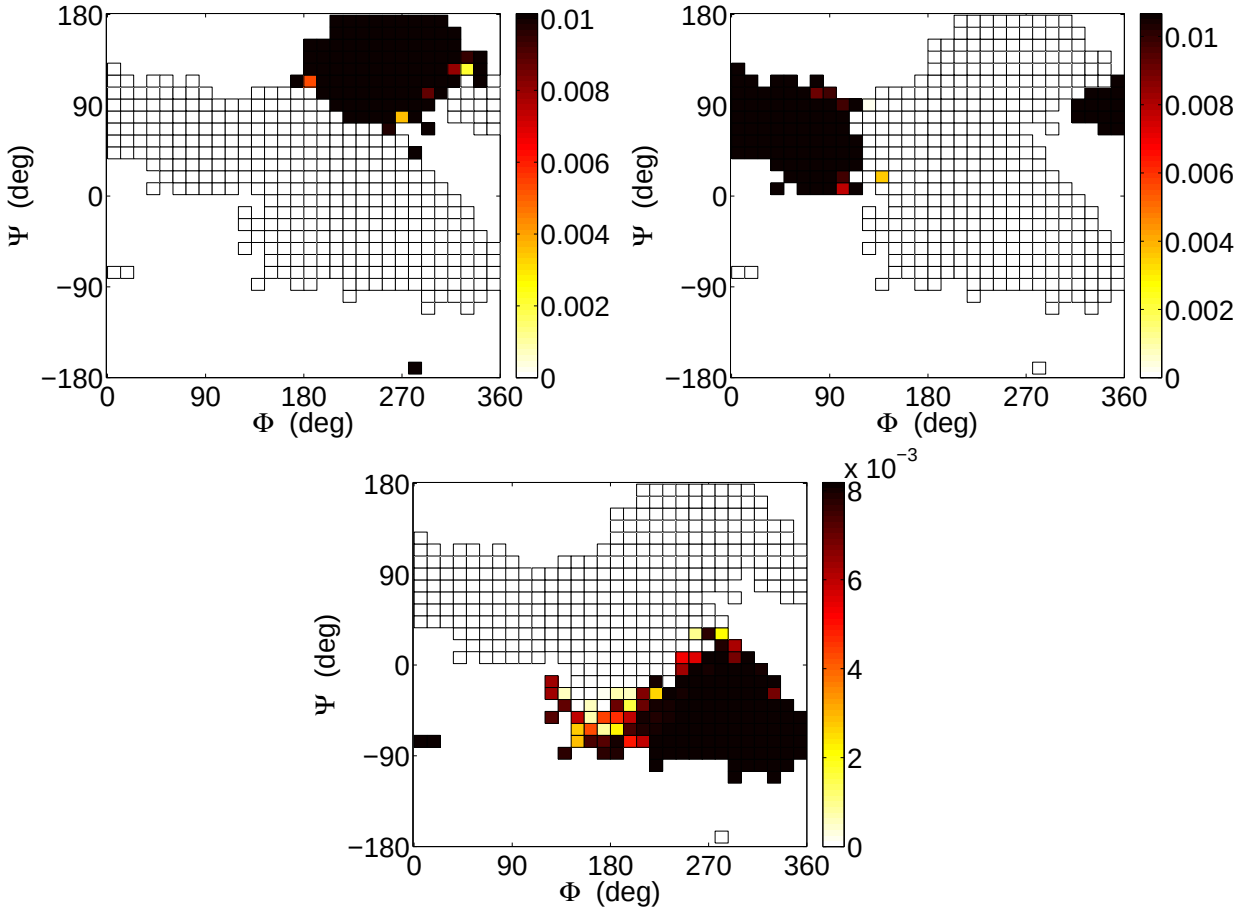

Figure 10: (Color online)Conditional probability distributions $p_x(i),\ i \in S$ with $p_x(i)$ is the probability of a state i being assigned to cluster x conditional on being in cluster x. Top left: Cluster $x = 3$. Top right: cluster $x = 2$. Bottom: Cluster $x = 1$. All distributions are computed via PCCA+ and the regularized assignment function in (35) for $\gamma = 0.9$. The underlying ensemble $\mathcal{T}_{Rev}$ consists of 500.000 reversible transition matrices.

for $\gamma = 0.9$ is small. However, the uncertainty of the identification of the core clusters is quite high which becomes obvious by comparing, e.g., the core clusters $x = 3$ associated with $T^*(C')$ (see right panel in Figure 9) with the corresponding distribution p_3. Roughly spoken, the lower third of the core cluster $x = 3$ is not assigned to any cluster at all in the ensemble $\mathcal{T}_{Rev}$ (cf. Figure 11). Consequently, the decomposition based on the maximum likelihood estimator $T^*(C')$ would be misleading.

c. Conformational switching process Beside indicating uncertainty, the distributions p_1, p_2, p_3 and p_- or, more precisely, the corresponding histograms allow for an accurate characterization of core and transition clusters. To show that, the torsion angle state space is

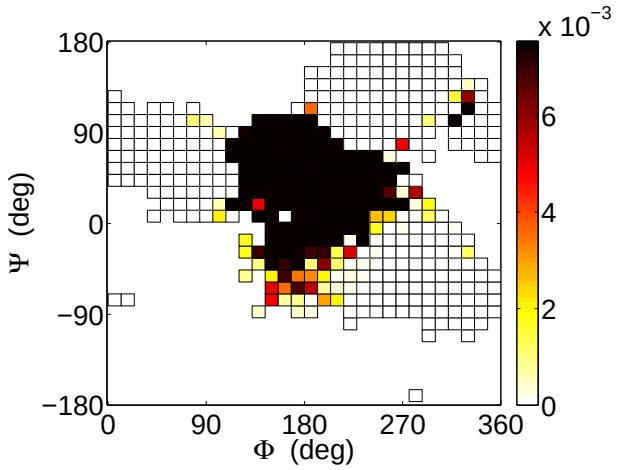

Figure 11: (Color online)Conditional probability distributions $p_-(i)$, $i \in S$ with $p_-(i)$ is the probability of state i being not assigned to any cluster conditional on being not assigned to any cluster at all.

decomposed by applying the hard assignment function in (34) with respect to the membership function resulting from the assignment/non-assignment histograms. More formally, a state i is assigned to a cluster $x \in \{1, \ldots, 4\}$ if i was most frequently assigned to cluster x with respect to the ensemble $\mathcal{T}_{Rev}$ whereby the cluster $x = 4$ represents the case "non-assigned at all". The resulting four clusters are illustrated in the left panel of Figure 12.

It remains to justify that the above described procedure leads to a meaningful decomposition into core metastable subsets dynamical connected via a transition region. To this end, we consider the Markov switching process between the four clusters represented by the maximum likelihood estimator $T^*(\hat{C})$ with $\hat{C} \in \mathbb{R}^{4\times4}$ being the frequency matrix computed from the torsion angle time series. The associated transition graph, schematically illustrated in the right panel of Figure 12, reveals the claimed character of the cluster; all transitions between the strongly metastable core clusters $x = 1, x = 2$ and $x = 3$ happen via the (weak metastable) fourth cluster $x = 4$. Therefore, the fourth cluster (the non-assignment cluster) can be interpreted as a *transition cluster*.

V. CONCLUSION

We have presented an efficient scheme for sampling posterior probability distributions of transition matrices. The scheme has been derived for uncertainty analysis of Markov state

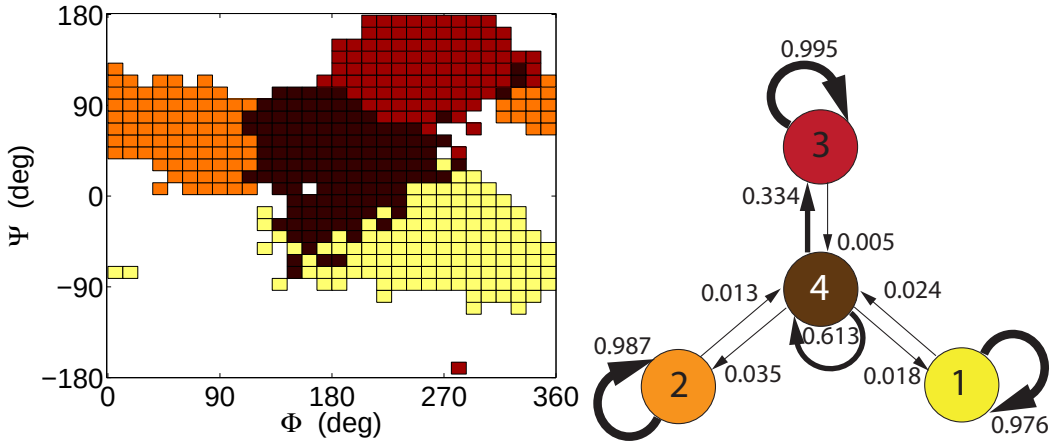

Figure 12: (Color online)Left: Decomposition of the torsion angle space into four clusters which is based on the hard assignment (34) with respect to the assignment/non-assignment histograms. Right: Transition matrix $T^*(\hat{C})$ of the switching process between the four clusters. The cluster in the center ($x = 4$) can be interpreted as a transition region as the switching process between the upper clusters ($x = 3$), the left cluster ($x = 2$) and the lower cluster ($x = 1$) only happens *via* cluster $x = 4$.

models (MSM) estimated from time series. Its performance has been illustrated on a variety of toy examples.

The main novelty is the penalty prior that has been introduced to account for the need to preserve the sparsity of the observations. The strength of the believe that transition matrix entries are less important if no transitions between the respective involved states have been observed yet is expressed by an additional scaling (penalty) parameter M. In the limit $M \to \infty$ the penalty prior reduces to the uniform prior *restricted* to the set of observed transitions.

We proposed a Gibbs sampler based scheme for sampling the resulting posterior distribution of transfer operators / transition matrices. This scheme allows for incorporation of additional constraints. In particular, we have discussed how to perform efficient sampling of the posterior for *reversible* transition matrices since this constraint is essential in the construction of MSMs for equilibrium molecular dynamics. Instead of ensuring reversibility by means of the detailed balance condition, the proposed Gibbs sampler acts on *positive symmetric* matrices which naturally lead to reversible transition matrices by normalizing the rows. Based on [15], we have rigorously demonstrated that the involved transformation

leads to the right posterior distribution.

The proposed approach has been applied to an example arising from molecular dynamics where we have analyzed the uncertainty of the identification of conformations of the tri-alanine molecule via the Robust Perron Cluster Analysis (PCCA+). To this end, we have introduced a regularized affiliation function and, specifically, have demonstrated by means of uncertainty analysis that the decomposition of the state space via PCCA+ based on the maximum likelihood chain may yield misleading metastable sets. Moreover, we have highlighted that PCCA+ combined with uncertainty analysis provides a promising numerical procedure to identify metastable core and transition sets.

Last but not least let us add a warning: The choice of a prior in Bayesian statistics allows to incorporate a belief about probable and improbable results. Therefore comparisons of Bayesian approaches using different priors are of limited use only. A specific choice of a prior can and must be validated in application to realistic examples but one should always be aware of the fact that for different realistic examples different priors can be superior. Particularly, the choice of the penalty parameter M has carefully to be tested out by, e.g., varying M from large to small values. Furthermore, the choice of M can change from problem to problem. This article does not contain any such detailed comparisons between different priors; this will have to be the topic of forthcoming contributions. This article's main contribution lies in making available a sparsity based prior, a corresponding sampler that allows to incorporate reversibility, and in its validation in application to a typical realistic example.

Acknowledgments

We would like to thank the anonymous referees for their helpful comments. This work is supported by the DFG Research Center MATHEON "Mathematics for Key Technologies" (FZT86) in Berlin.

VI. APPENDIX

It remains to prove

Theorem .1. *Let $\mathcal{K}_{sym} = \{K \in \mathfrak{K}_{sym}\}$ be an ensemble of symmetric count matrices distributed according to $p_C(u(K))$. Then the ensemble $\mathcal{T} = \{u(K) : K \in \mathcal{K}_{sym}\}$ of reversible*

transition matrices is distributed according to $p_C(T)$, i.e.,

$$\mathbb{P}[u(K) = T] = c\, p_C(T) \quad \forall\, T \in \mathcal{T},$$

where $c > 0$ is a positive constant independent of the matrix T.

Proof. Formally, the statistical weight of a reversible transition matrix $T \in \{u(K) : K \in \mathcal{K}_{sym}\}$ is given by

$$\mathbb{P}[u(K) = T] = \int_{u^{-1}(K)} p_C(u(K))\, \mathrm{d}K. \tag{37}$$

The key observation is that the set $u^{-1}(K) \subset \mathfrak{K}_{sym}$ can be parameterized as

$$u^{-1}(K) = \{\alpha\ S\},$$

where

$$S = \mathrm{diag}(\pi_1, \ldots, \pi_m)T$$

is a symmetric matrix, $\pi = (\pi_i),\ i = 1, \ldots, m$ is the unique stationary distribution of T and $\alpha \in [k^-, k^+]$.

To motivate the following transformation, note that for symmetric K the stationary distribution of $T = u(K)$ is simply given by

$$\pi_i = \frac{\sum_{j=1}^m K_{ij}}{\sum_{k,j=1}^m K_{kj}}$$

and we conclude

$$\pi_i T_{ij} = \frac{K_{ij}}{\sum_{k,l=1}^m K_{kl}}.$$

We change variables according to the transformation F:

$$K \mapsto (\alpha, S_{11}, \cdots, S_{d,d-1})$$

with $\alpha = \sum_{k,l=1}^m K_{kl}$ and $S_{ij} = \frac{K_{ij}}{\alpha}$. The inverse transformation F^{-1} reads

$$(\alpha, S_{11}, \cdots, S_{m,m-1}) \mapsto (K_{11}, \cdots, K_{mm})$$

with $K_{ij} = \alpha \cdot S_{ij}$ and $K_{mm} = \alpha - \alpha \sum_{i=1}^m \sum_{j=1}^{m-1} S_{ij}$. The right hand side in (37) with respect to the new variables is given by

$$\int_{\{\alpha S\}} p_C(u(K))\, \mathrm{d}K = \int_{F(\{\alpha S\})} p_C(u(F^{-1}(x)))\, \left|\det(J(F^{-1}(x)))\right|\, \mathrm{d}x. \tag{38}$$

Let $x = (\alpha, S_{11}, \cdots, S_{m,m-1}) \in F(\{\alpha S\})$ then the first factor in (38) reduces to $p_C(u(F^{-1}(x))) = p_C(T)$. It remains to evaluate the Jacobian.

$$\begin{aligned}
\det J(F^{-1}(x)) &= \begin{vmatrix} S_{11} & \alpha & 0 & 0 & \dots \\ S_{12} & 0 & \alpha & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ S_{m,m-1} & 0 & \dots & \dots & \alpha \\ 1 - \sum_{i=1}^{m} \sum_{j=1}^{m-1} S_{ij} & -\alpha & -\alpha & \dots & -\alpha \end{vmatrix} \\
&= \begin{vmatrix} S_{11} & \alpha & 0 & 0 & \dots \\ S_{12} & 0 & \alpha & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ S_{m,m-1} & 0 & \dots & \dots & \alpha \\ 1 & 0 & 0 & \dots & 0 \end{vmatrix} \\
&= (-1)^{(m^2-1)} \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ S_{11} & \alpha & 0 & 0 & \dots \\ S_{12} & 0 & \alpha & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ S_{m,m-1} & 0 & \dots & 0 & \alpha \end{vmatrix} \\
&= (-1)^{(m^2-1)} \alpha^{(m^2-1)}.
\end{aligned}$$

Putting all together we have shown that (37) reduces to

$$\int_{\{\alpha S\}} p_C(u(K)) \, \mathrm{d}K = \left((k^+)^{m^2-1} - (k^-)^{m^2-1} \right) p_C(T),$$

where the factor $(k^+)^{m^2-1} - (k^-)^{m^2-1}$ is independent of T. □

---

[1] Ch. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Applications to Biomolecules*. Habilitation thesis, Fachbereich Mathematik und Informatik, FU Berlin, 1998.

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

[3] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comp. Physics Special Issue on Computational Biophysics*, 151:146–168, 1999.

[4] Ch. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In *Handbook of Numerical Analysis*, pages 699–744. Elsevier, 2003.

[5] F. Noe, Ch. Schütte, E. Vanden-Eijnden, L. Reich, and T. Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium trajectories. *PNAS*, 106(45):19011–19016, 2009.

[6] F. Noé, I. Horenko, Ch. Schütte, and J. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126:155102, 2007.

[7] J. Chodera, N. Singhal, V. S. Pande, K. Dill, and W. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *Journal of Chemical Physics*, 126, 2007.

[8] Nicaolae V. Buchete and Gerhard Hummer. Coarse master equations for peptide folding dynamics. *Journal of Physical Chemistry B*, 112:6057–6069, 2008.

[9] A. C. Pan and B. Roux. Building Markov state models along pathways to determine free energies and rates of transitions. *Journal of Chemical Physics*, 129, 2008.

[10] M. Sarich, F. Noé, and Ch. Schütte. On the approximation quality of Markov state models. *to appear in Multiscale Modeling and Simulation*, 2010.

[11] A. Voter. Introduction to the kinetic Monte Carlo method. In *Radiation Effects in Solids*. Springer, NATO Publishing Unit, Dordrecht, The Netherlands, 2005.

[12] T. W. Anderson and Leo A. Goodman. Statistical inference about Markov chains. *Ann. Math. Statist.*, 28(1):89–110, 1957.

[13] N. Singhal and V. S. Pande. Error analysis and efficient sampling in markovian state models for molecular dynamics. *J. Chem. Phys.*, 123(20):204909, 2005.

[14] F. Noé. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.*, 128(24):244103, 2008.

[15] P. Metzner, F. Noé, and Ch. Schütte. Estimating the sampling error: Distribution of transition matrices and functions of transition matrices for given trajectory data. *Phys. Rev. E*, 80(2):021106, 2009.

[16] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[17] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, 2001.

[18] F. Cordes, M. Weber, and J. Schmidt-Ehrenberg. Metastable conformations via successive Perron-Cluster Cluster Analysis of dihedrals. *ZIB-Report 02-40*, 7, 2002.

[19] P. Deuflhard and M. Weber. Robust Perron Cluster Analysis in conformation dynamics. *Lin. Alg. Appl.*, 398:161–184, 2005.

[20] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 41(2):337–348, 1992.

[21] W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive rejection Metropolis sampling. *Applied Statistics*, 44(2):455–472, 1995.

[22] D. Görür and Y. W. Teh. Concave convex adaptive rejection sampling. *Technical Report, Gatsby Computational Neuroscience Unit*, 2008.

[23] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7(4):457–472, 1992.

[24] S. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Stat*, 7(4):434–455, 1998.

[25] M. Dellnitz, M. Hessel von Molo, P. Metzner, R. Preis, and Ch. Schütte. Graph algorithms for dynamical systems. In A. Mielke, editor, *Analysis, Modeling and Simulation of Multiscale Problems*. Springer Verlag, 2006.

[26] P. Metzner, Ch. Schütte, and E. Vanden-Eijnden. Transition path theory for Markov jump processes. *Multiscale Model. Simul.*, 7(3):1192–1219, 2009.

[27] A. Brass, B. J. Pendleton, Y. Chen, and B. Robson. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. *Biopolymers*, 33(8):1207–1315, 1993.

[28] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91(1-3):43–56, 1995.

[29] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7(8):306–317, 2001.

[30] M. Weber. *Meshless Methods in Conformation Dynamics*. Doctoral thesis, Department of Mathematics and Computer Science, Freie Universität Berlin, 2006. Published by Verlag Dr.

Hut, München.

[31] S. Röblitz. *Statistical Error Estimation and Grid-free Hierarchical Refinement in Conformation Dynamics*. Doctoral thesis, Department of Mathematics and Computer Science, Freie Universität Berlin, 2008.