Expediting Scientific Discoveries With Bayesian Statistical Methods

Expediting Scientific Discoveries with Bayesian Statistical Methods

a dissertation presented
by
Yang Chen
to
The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Statistics

Harvard University
Cambridge, Massachusetts
April 2017

Thesis advisor: Professor Samuel Kou Yang Chen

Expediting Scientific Discoveries with Bayesian Statistical Methods

Abstract

The topic of this thesis is developing Bayesian statistical methodology aimed at solving scientific problems and thoroughly studying relevant statistical computational methods. There are four chapters in total. The first three chapters are motivated from a fundamental biological process and the last one is about evaluating Bayesian computational algorithms that utilize modern parallelisable computing architecture. Each of the four chapters is self-contained and is in the format of a journal paper, with technical details given in the corresponding Appendix.

In chapter one, we study the molecular mechanism underlying the protein transportation process through data obtained from single-molecule experiments that use fluorescence imaging to track molecular behaviors. The experimental data consist of hundreds of stochastic time traces from the fluorescence recordings of the experimental system. We introduce a Bayesian hierarchical model on top of hidden Markov models (HMMs) to analyze these data and use the statistical results to answer the biological questions. Besides resolving the biological puzzles and delineating the regulating roles of different molecular complexes, our statistical results enable us to propose a more detailed mechanism for the late stages of the protein targeting process.

In chapter two, we introduce a a Matlab package for Bayesian analysis of ensembles of single-molecule fluorescence traces from replicated experiments. The proposed

Bayesian hierarchical hidden Markov model in chapter one provides a principled way of extracting the common dynamics of observed traces from experimental replicates. Numerical examples demonstrate the wide applicability of the Matlab package: traces with low signal-to-noise ratios, traces with rare events, and heterogeneous traces with unknown number of hidden states and different numbers of observations.

In chapter three, we propose a consistent method of estimating the order of hidden Markov models based on the marginal likelihood, which is obtained by integrating out both the parameters and hidden states. We prove the consistency of the marginal likelihood method under weak regularity conditions that are satisfied by a broad class of models. An `R` package is built for practitioners to apply the proposed methodology. Comprehensive simulation studies illustrate the comparison of the proposed method with the currently most widely adopted method, the Bayesian information criterion (BIC), demonstrating the effectiveness of the marginal likelihood method.

In chapter four, we study parallelisable Markov chain Monte Carlo algorithms. Parallelisable Markov chain Monte Carlo algorithms generate multiple proposals and parallelise the evaluations of the likelihood functions on different cores at each iteration. We give a simple-to-use criterion, the generalized effective sample size, for evaluations and comparisons of general parallelisable Markov chain Monte Carlo algorithms. The formula is easy to implement using moment estimators.

The thesis concludes with brief discussions of several open interesting questions related to the materials in chapters 1 through 4.

Contents

Listing of figures

To my husband, Tong Qin, for accompanying me through this journey.

To my parents, for their endless love and support.

Acknowledgments

I would like to express my appreciation to my adviser professor Samuel Kou. I would like to thank you for the patient and effective guidance on my research. I would also like to thank my committee members, professor Jun Liu and professor Neil Shephard for serving as my committee members. I also want to thank you for your brilliant comments and suggestions. I would especially like to thank professor Xiao-Li Meng, who has given me great support and encouragements.

A special thanks to my family. Words cannot express how grateful I am to my parents. I would like to thank my mentor, Rehana Patel, for the helpful advice. I would also like to thank all of my friends who spent countless enjoyable moments with me. At the end I would like express appreciation to my beloved husband, Tong Qin, without whose continuous love and support I couldn't have gone this far. It has been my greatest pleasure to be with you through ups and downs in life.

0

Introduction

Statistical analysis has been playing an important role in facilitating scientific discoveries in many ways – from designing experiments to processing experimental data and interpreting the quantitative results. The contents of this thesis is motivated from the broad applications of various statistical tools in tackling challenging scientific problems in the *Big-Data* era, aiming at contributing to the innovative developments and thorough understanding of statistical methods, the advancement of statistical theory, and the practice of principled statistical analysis for real world problems. Many parts of the statistical research presented in this thesis originate from science, and serve science. Besides the contributions to scientific disciplines, progress has also been made to the statistical methodology and theory – on the one hand, the scientific questions stimulate novel statistical problems; on the other hand, scientific practice contributes original ideas to the development of novel statistical methods.

0.1 Statistics Analysis of Single-Molecule Data

Propelled by advances in technology, the scientific community has been able to study dynamical behaviors of biological molecules through single-molecule experiments. A single-molecule experiment preserves the signal that is lost by the bulk averaging in traditional ensemble experiments. Thus the last two decades has witnessed great excitement of single-molecule methods in various biological areas.

For the statistics community, the single-molecule experiments have brought in both challenges and opportunities for the rigorous data analysis. First, since molecular behaviors are inherently stochastic, stochastic modeling is required to analyze single-molecule data, which is not necessary in traditional ensemble experiments. Second, the single-molecule observations are highly noisy. Third, the true biological process is often the unobserved latent process underlying the noisy observations. Last but not least, hundreds of experimental replicates are created in single-molecule experiments and the experimental replicates demonstrate apparent heterogeneity.

The major contents of this thesis is motivated from the afore mentioned challenges in single-molecule data analysis. First, we develop general Bayesian statistical methodologies to cope with the difficulties in analyzing data sets from single-molecule experiments. The statistical analysis plays an important role in understanding the biological mechanisms by enabling information sharing while allowing heterogeneities among experimental replicates, which are highly volatile. Next, we build user-friendly computational packages for the scientific community to easily adopt the proposed methodology for daily data analysis. Last, the number of conformations of a biomolecular complex, which is mostly unknown beforehand, is of great importance in biology. Determining the number of conformations based on observations corresponds to the model selection problem in statistics. This leads to the theoretical studies of order selection of hidden Markov models and finite mixture models.

0.2 Statistical Computation for Complex Models

As the demand for extracting valuable signal from highly noisy data increases, people are no longer satisfied with descriptive statistics or simple statistical models such as linear regression. The desire for modeling real world stochastic systems and understanding the underlying uncertainties pushes the boundary of statistical models. An appropriate statistical model that takes into account the realistic complexity always turns out to be highly sophisticated. The single-molecule data analysis is a good example that demonstrates this point. The increasing complexity of the statistical model poses difficulties on the computation, especially on Bayesian computation which relies on the Markov chain Monte Carlo sampling, the efficiency of which is a concern. In recent years, Bayesian computational algorithms that utilize modern parallel computation architecture have caught a lot of attention. Along this direction, we further develop parallelisable Markov chain Monte Carlo algorithms and evaluate them against widely adopted traditional sampling algorithms.

0.3 Outline of Each Chapter

The remainder of the thesis is organized as follows.

In chapter one, we give a thorough study of the single-molecule data obtained from experiments aimed at understanding a fundamental biological process called protein transportation. The data acquisition, pre-processing, analysis and interpretation are detailed, as well as the scientific meanings. We propose a Bayesian hierarchical model on top of hidden Markov models to take into accounts of different layers of variabilities in the single-molecule data. The manuscript has been published in the *Journal of American Statistical Association* in 2016.

In chapter two, we generalize the methodology proposed in chapter one and build a Matlab package for processing single-molecule data sets. The statistical properties such as consistency and robustness are studied with numerical experiments.

In chapter three, we study the theoretical problem motivated from the real application in chapter one, i.e., the order selection of discrete state space hidden Markov models. This is a fundamental problem that is faced by many practitioners of various fields, but yet to be satisfactorily solved. We propose a consistent estimator that can be easily computed and provide an `R` pacakge for practitioners.

In chapter four, parallelisable Markov chain Monte Carlo algorithms are discussed. We derive a quantity that measures the effectiveness of posterior sampling for a general family of parallelisable Markov chain Monte Carlo algorithms.

Each chapter in chapters 1-4 is a self-contained paper on its own.

The thesis is concluded with discussions about some open and intriguing statistical problems in single-molecule data analysis and Bayesian computation algorithms.

0.4 Acknowledgment of Coauthors

The contents in this thesis are under the guidance of faculties in the department of statistics at Harvard university, *Professor Samuel Kou*, *Professor Jun Liu* and *Professor Neil Shephard*. Chapters one and two are joint work with *Professor Samuel Kou*, *Professor Shu-ou Shan* from the Division of Chemistry and Chemical Engineering, California Institute of Technology, and *Mr. Kuang Shen* from the Whitehead Institute, Massachusetts Institute of Technology. Chapter three is joint work with *Professor Samuel Kou*, *Proferssor Cheng-Der Fuh* from the Graduate Institute of Statistics, National Central University, Taiwan, and *Professor Chu-Lan Kao* from the Institute of Statistics, National Chiao Tung University, Taiwan. Chapter four is joint work with *Professor Jun Liu*, *Professor Neil Shephard*, *Mr. Shihao Yang* and *Mr. Espen Bernton* from the Department of Statistics, Harvard University.

Science is the great antidote to the poison of enthusiasm and superstition.

Adam Smith, The Wealth of Nations, 1776

Those who ignore statistics are condemned to reinvent it.

Bradley Efron

1

Uncovering Science from Single-Molecule Data

1.1 Introduction

In cells, proteins often need to be transported to appropriate destinations inside or outside of a cell in order to maintain proper cellular functions (Rapoport, 2007). In fact, over 50% of all proteins encoded in the genome need to be properly localized from the site of their synthesis (Lodish et al., 2000; Rapoport, 1991). Co-translational protein targeting is such a process in which proteins still being synthesized on the ribosome (called ribosome nascent-chain complex or RNC) are transported to the membrane. This is achieved by the collaboration of a signal recognition particle (SRP) in the cytoplasm and its receptor (SR) located on the endoplasmic reticulum (ER) membrane. It is known that the co-translational protein targeting process consists of four basic steps (Zhang et al., 2009b; Nyathi et al., 2013), as schematically illustrated in Figure 1.1. First, SRP recognizes and binds the signal sequence on the RNC. Second, SRP forms a complex with SR on the membrane, bringing the RNC-SRP complex to the membrane surface (here, an RNC-SRP-SR ternary complex is formed near the membrane). Third, the RNC is released from the SRP-SR complex and docks on the protein conducting channel, known as the translocon. Fourth, SRP and SR dissociate (through GTP-hydrolysis) to enter a new round of protein targeting; at the same time, the nascent polypeptide chain goes through the translocon on the membrane.

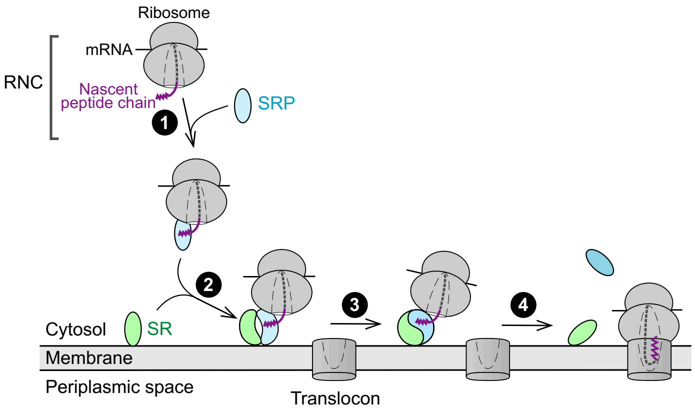

Figure 1.1: The four steps of protein targeting.

While the four steps give the big picture, the detailed molecular mechanisms of the protein targeting process remained unclear (Shen et al., 2012). One particularly puzzling question arises from the earlier observation that SRP and the translocon bind the same sites on the ribosome and the signal sequence; thus, the bindings of the targeting and translocation machineries to RNC are mutually exclusive. How do these two machineries exchange on the RNC, and how do they accomplish this without losing the RNC (which aborts the pathway)? Recent biochemical, structural, and single-molecule work (Zhang et al., 2008; Shen & Shan, 2010; Ataide et al., 2011; Voigts-Hoffmann et al., 2013; Nyathi et al., 2013; Akopian et al., 2013b) offered valuable clues to this question. These works showed that the SRP-SR complex can undergo a large-scale structural change and visit an alternative state in which the proteins in the SRP-SR complex are moved away from their initial binding site on the ribosome (see Figure 4 below); this provides a potential mechanism to enable a step-wise exchange with the translocon.

To provide direct evidence for this mechanism and resolve its molecular details, single-molecule experiments on the prokaryotic SRP system were conducted by the Shan group. Single-molecule experiments are one of the major experimental breakthroughs in chemistry and biophysics in the last two decades: using advanced tools in optics, imaging, fluorescence tagging, biomolecule labeling, etc., researchers are able to study biological processes on a molecule-by-molecule basis (Moerner, 2002; Nie & Zare, 1997; Tamarat et al., 2000; Weiss, 2000; Xie & Trautman, 1998; Xie & Lu, 1999; Qian & Kou, 2014). Under single-molecule experiments, transient excursions of molecules to alternative structures can be directly visualized, rather than lost in the statistical averaging of bulk experiments.

The single-molecule experiments under our study employ an experimental technique, FRET (Föster resonance energy transfer) (Roy et al., 2008), which uses resonance energy transfer as a molecular ruler to track the dynamic movement of a molecule in distinct conformational states, providing information on the pathway, kinetics and equilibrium of the structural transitions of molecules. The experimental data consist of hundreds of FRET trajectories, three of which are shown in Figure 1.2. Each FRET trajectory is a time series $(y_1, y_2, \ldots)$. These experimental FRET trajectories provide crucial information on the structural dynamics for us to resolve the questions regarding the underlying mechanism of protein targeting. We will describe the experimental details as well as the molecular structures in Section 1.2.

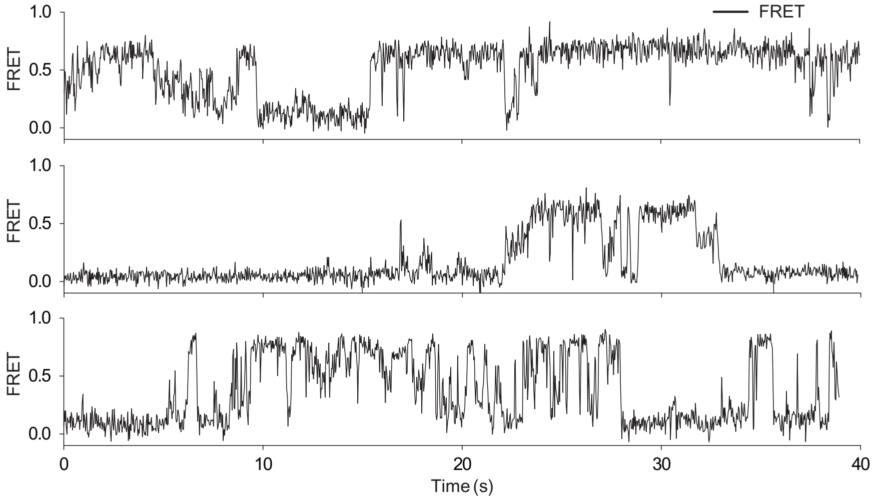

Figure 1.2: Three sample FRET trajectories.

From the hundreds of traces collected, we can clearly see a low FRET state and a high FRET state in each trace, with one or more possible intermediate states. Several critical questions arise regarding the correct interpretation of the data.

1. Molecular behavior is inherently stochastic. Ensembles of molecules that are chemically identical will vary in their behavior at the single-molecule level (in a manner predicted by the Boltzmann distribution). Thus, individual single

molecule traces are inherently heterogeneous. In addition, due to the experimental limitations, such as uneven laser illumination, each FRET trajectory has its own FRET values and length. Moreover, it is possible that some observed molecules are partially damaged during sample preparation or application. Therefore, we want to carefully examine the homogeneity/heterogeneity of the data set: Does the collection of FRET trajectories represent chemically homogeneous molecules or molecular complexes? If not, is the heterogeneity biologically relevant?

2. How many states are there in these FRET trajectories? Previous analysis utilized an arbitrary number of states for HMM (Shen et al., 2012). However, there is no statistical analysis to legitimate that number. A careful analysis is needed to unravel the existence of intermediate state(s) from the noisy experimental data; this information is critical, as it reflects possible pathways through which the SRP-SR undergoes its structural transitions.

3. Are these intermediates on-pathway or off-pathway? In other words, during the transition from the low FRET state to the high FRET state, must or may not the trajectory go through one or more intermediate state(s)? Clarifying the transition pathway will differentiate between different mechanisms. In one model, often termed trial-and-error, the intermediate states are “mistakes” made by the complex as it searches for alternative structures. This model predicts that the molecules must return from the intermediate back to the low FRET state before transitioning to the high FRET state. In an alternative model, the active-searching model, the intermediate FRET state(s) represent on-pathway intermediate(s) through which the SRP-SR complex attains the

high FRET state. This model predicts that most of the successful low-to-high or high-to-low FRET transitions occur via the intermediate state(s).

4. During the protein targeting process, RNC and translocon regulate the conformation of the SRP-SR complex. This was also observed in the single-molecule experiments. Addition of RNC or translocon changes the equilibrium and kinetics via which the SRP-SR complex transits between the different FRET states, as reflected by altered frequency and durations of these transitions. However, as individual single-molecule traces are stochastic due to a combination of inherent and experimental limitations (as explained in question 1), it is not possible to accurately extract kinetic and equilibrium information from individual trajectories. Rigorous statistical analysis using the information from all trajectories is required to extract this information and understand whether the RNC and translocon change the conformational space of the SRP-SR complex, and if so, how.

With these questions posed, we employ a hidden Markov model (HMM), modeling each trajectory $(y_1, y_2, \ldots)$ as originated from a hidden Markov chain. The parameters governing the hidden Markov chain, such as the number of distinct states and the transition probabilities, capture the molecular conformations and dynamics of the underlying biological processes.

We note that the analysis of *individual* FRET trajectories based on HMMs has been considered in the biophysical community (Rabiner, 1989; Eddy, 1996; Liu et al., 2010). Software packages *HaMMy* (McKinney et al., 2006) and *SMART* (Greenfeld et al., 2012) give the maximum likelihood estimators of parameters for a *single* trajectory using the EM/Baum-Welch algorithm (Baum & Petrie, 1966; Baum et al., 1970;

Dempster et al., 1977). Variational Bayes method is also suggested in the FRET data analysis, which incorporates prior information about the range of parameter values into the model fitting (Bronson et al., 2009). Empirical Bayes methods (van de Meent et al., 2014) and bootstrap methods (König et al., 2013) have also been proposed for the analysis of FRET data.

The information from individual FRET trajectories is rather limited, mainly due to the low signal-to-noise ratio and the limited observation time of each individual molecule (before its photobleaching). Consequently, the inference based on single FRET trajectories is highly variable and unreliable in the sense that even for FRET trajectories recorded under the same experimental condition, heterogeneities of estimated parameters and the estimated number of hidden states across trajectories are apparent. Experimentalists address this issue by performing hundreds of replicate experiments. Quantifying cross-sample variability has recently drawn attention among the biophysics community (König et al., 2013; van de Meent et al., 2014). How to pool information from these replicate experimental trajectories as well as to account for their heterogeneity is the key statistical question.

Two statistical questions naturally arise in our analysis of the FRET trajectories: (1) the determination of the total number of hidden states and (2) a robust and reliable estimation of model parameters by pooling information from "seemingly" heterogeneous FRET trajectories obtained from the same experimental condition.

The first quesiton, which is a preliminary step of building models to pool information from multiple trajectories, has been widely studied in the statistics and chemistry literature (Finesso, 1990; Leroux, 1992a; Rydén, 1995; Blanco & Walter, 2010; Bulla et al., 2010). We adopt a population approach based on the Bayesian information criterion, which estimates the number of hidden states by the majority rule (e.g., if the majority of the FRET trajectories under the same experimental condition shows three states, then the method selects three as the number of hidden states). This approach actually has been recommended in the chemistry literature (Watkins & Yang, 2005) and is described in Section 1.3, which also discusses our fitting of HMM to individual FRET trajectories.

Second, we propose a hierarchical model on top of the HMMs to combine information from multiple trajectories. The hierarchical model embodies the biological intuition that the same dynamics underlies all the experimental replicates, but each replicate is a noisy realization of the common process due to intrinsic/experimental fluctuation and noise. The hierarchical HMM enables us to not only robustly estimate the parameters from the common dynamics but also fit the individual trajectories better than if fitted individually. Section 2.2 describes in detail our hierarchical HMM and how we use it to combine information from individual trajectories. Simulation studies demonstrating that the hierarchical model can work effectively under low signal-to-noise ratio, which is very difficult to analyze if one only fits individual trajectories.

From an applied angle, our statistical analysis of the experimental FRET data leads to a resolution of several questions about the protein targeting process that are described above. The model fitting and biological implications are discussed in Section 1.5, at the end of which (Section 1.5.4) we are able to provide a detailed molecular mechanism of the co-translational protein targeting process. Model assessment is conducted in Section 1.6. We conclude this article in Section 2.5 with a summary. The appendix contains the technical details of our computation and Monte Carlo sampling.

1.2 Single-Molecule Experiments on Co-translational Protein Targeting

1.2.1 Single-Molecule FRET Experiments

The single-molecule experiments use the FRET technique to study the protein targeting process. FRET tracks in real time the distance and orientation between two microscopic tags, a donor fluorophore and an acceptor fluorophore, placed in a molecular complex (Roy et al., 2008). It is often the case that the experimentalists cannot directly observe the structural change of a bio-molecule. The FRET recording, on the other hand, measures the distance changes of the two tags on the bio-molecule and thus reveals the structural changes during a biological process.

Each experimental FRET trajectory is a time series $(y_1, y_2, \ldots)$, obtained at every 30 millisecond (ms) in our case. $y_i \in [0, 1]$ is calculated as $y_i =$ acceptor fluorescence / (donor fluorescence + acceptor fluorescence). A high FRET value y_i implies that the two tags, the donor and acceptor, are close to each other, while a low FRET value means the donor and acceptor are far apart. A sample FRET trajectory is shown in Figure 1.3. On the top panel, the red curve is the acceptor fluorescence and the green curve is the donor fluorescence. The black curve in the lower panel shows the FRET values, i.e., the ratio of acceptor fluorescence over the total fluorescence.

1.2.2 FRET on Bacterial SRP System

In this subsection, we give the necessary background on the molecular structure of our experimental system and how FRET reveals information about protein targeting.

Single-molecule FRET technique was used to study the bacterial SRP system. The

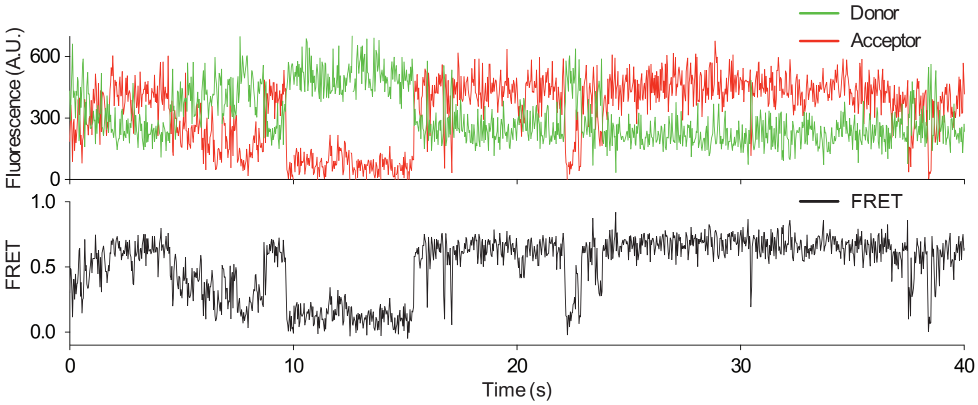

Figure 1.3: Sample trajectory of FRET observations. The upper panel is the fluorescence of the donor and the acceptor, respectively; the lower panel shows the FRET values.

bacterial SRP is comprised of two subunits: an RNA segment (the SRP RNA) and an Ffh protein. Ffh contains two domains connected by a flexible linker: the M-domain binds tightly to the SRP RNA near its capped (tetraloop) end and recognizes the signal sequence on the nascent protein; the NG-domain interacts with the SRP receptor, termed FtsY in bacteria, and binds a ribosomal protein at the "exit site" where the nascent protein emerges from the ribosome. We will use Ffh-M and Ffh-NG to denote the M- and NG- domains of Ffh (Akopian et al., 2013b; Halic et al., 2004; Keenan et al., 2001; Zhang et al., 2008). The SRP RNA has an elongated structure: it stretches over 100 Å (angstrom) from one end (the capped end) to the other end (the distal end). Figure 1.4 illustrates the *E.coli* SRP and SR.

When the SRP-SR complex is formed, Ffh-NG binds FtsY (step 2 in Figure 1.1). In a single-molecule experiment, we placed a FRET donor at Ffh-NG or FtsY and a FRET acceptor at the distal end of RNA. The resulting FRET trajectory tracks the movement of the FtsY-[Ffh-NG] complex along the RNA in real time: a low FRET value implies the FtsY-[Ffh-NG] complex is far from the RNA distal end, whereas a

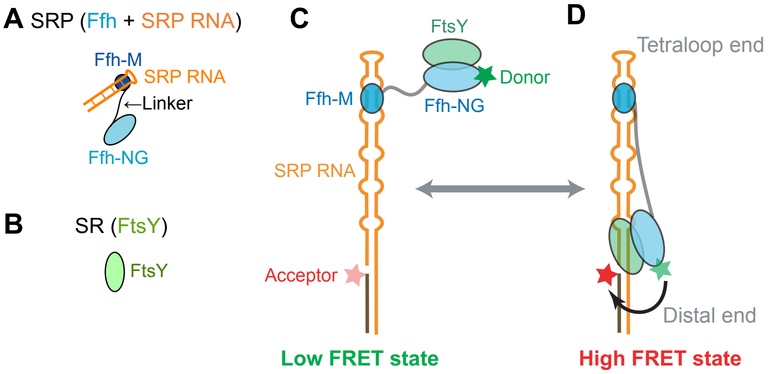

Figure 1.4: Molecular details of SRP and SR in *E.coli*. (A) SRP in *E.coli* is composed of RNA, Ffh-M and Ffh-NG. Ffh-M binds the RNA and the signal sequence (not shown); Ffh-NG binds the ribosome (not shown) and SR. (B) SR in *E.coli* is the FtsY protein. (C) FtsY-[Ffh-NG] complex is near the capped end of the RNA with a low FRET value. (D) FtsY-[Ffh-NG] complex is near the distal end of the RNA with a high FRET value. The red and green stars denote the FRET acceptor and donor, respectively.

high FRET value implies the FtsY-[Ffh-NG] complex is close to the RNA distal end. See C and D of Figure 1.4 for illustration (where the FRET donor is the green star and the FRET acceptor is the red star). The FRET tracking provides direct information on the structural change of SRP-SR complex critical for the biological process. It is known that the FtsY-[Ffh-NG] complex initially assembles at the RNA capped end (the low FRET state of Figure 1.4(C)), where it excludes the translocon from binding RNC. When this complex moves to the RNA distal end (the high FRET state of Figure 1.4(D)), the ribosome is vacated to allow translocon binding, and disassembly of the FtsY-[Ffh-NG] complex is triggered (Shen & Shan, 2010; Ataide et al., 2011). Therefore, from the FRET trajectory, we know when the SRP-SR complex is positioned for assembly or disassembly, and when ribosome-translocon contacts are enabled.

To study how the RNC and translocon regulate the structural change on the SRP-SR complex, two more sets of single-molecule FRET experiments were done: one with

RNC, SRP and SR, the other with all four components: translocon, RNC, SRP and SR. Together, these experiments reveal the functional role of RNC and translocon in the protein targeting process. Table 1.1 summarizes the four sets of data labeled *Ffh-Data, FtsY-Data, RNC-Data* and *Translocon-Data* obtained from these experiments, and Table 1.2 summarizes the lengths of the trajectories in each data set. We will analyze and discuss these data starting from Section 1.3.

Data	Donor	Acceptor	Complexes in experiments	No.
Ffh-Data	Ffh-NG	RNA distal end	SRP-SR	142
FtsY-Data	FtsY	RNA distal end	SRP-SR	208
RNC-Data	Ffh-NG	RNA distal end	SRP-SR, RNC	97
Translocon-Data	Ffh-NG	RNA distal end	SRP-SR, RNC, Translocon	138

Table 1.1: Data sets and number of recorded trajectories (last column) in each set.

	5% Quantile	Median	Mean	95% Quantile
Ffh-Data	518	1484	1681	3390
FtsY-Data	357	1027	1248	2993
RNC-Data	317	746	873	1864
Translocon-Data	338	918	1071	2357

Table 1.2: Summary of the lengths (number of data points) of the recorded trajectories in each data set.

1.2.3 More Experimental Details

This subsection gives the experimental details. A statistics oriented reader can skip it and directly go to the statistical analysis in Section 3.

Sample Preparations

Single cysteine mutants of Ffh and FtsY were expressed and purified in bacterial cells and were subsequently labeled with Cy3-maleimide by the thiol side chain. Labeling reaction was carried out in 50 mM KHEPES (pH 7.0), 300 mM NaCl, 2 mM EDTA, 10% glycerol at room temperature for 2 hours. Free dyes were removed by a gel filtration column. Labeled SRP RNA was prepared by annealing a Quasar670-labeled DNA splint with a T7-transcribed RNA. All the labeled protein or RNA was tested using a well-established GTP hydrolysis assay, and showed no functional difference with wildtype protein or RNA.

Single Molecule Instrument

All the experiments were carried out on a home-built objective-type TIRF microscope based on an Olympus IX-81 model. Green (532nm) and red (638nm) lasers were aligned and focused on the sample in a 100 × oil immersed objective. Cy3 and Quasar670 signals were split by a dichroic mirror and were simultaneously imaged using an Ixon 897 camera through DV2 Dualview. Data points were recorded at 30 milliseconds time resolution.

Single Molecule Assay

Before conducting experiments, all protein samples were ultracentrifuged at 100,000 rpm in a TLA100 rotor for an hour to remove possible aggregates. PEGylated slides and coverslips were assembled into a flowing chamber, in which fluorescent molecules were attached through biotin-neutravidin interaction.

SRP complexes were assembled in SRP buffer and diluted to 50 picomolar in imaging buffer with oxygen scavenging system (saturated Trolox solution containing 50 mM potassium-HEPES (pH 7.5), 150 mM KOAc, 2 mM $Mg(OAc)_2$, 2 mM DTT, 0.01% Nikkol, 0.4% glucose and 1% Gloxy), flowed onto the sample chamber and incubated for 5 minutes before imaging. Movies were recorded at 30 milliseconds time intervals for up to 3 minutes until most fluorescent molecules were photobleached.

Data Aquisition

Single molecule data were initially processed by scripts written in IDL and Matlab. Fluorescent peaks in the images were identified and traced throughout the movie. Fluorescent trajectories that showed a single donor bleaching event, which implied single-molecule attachment, and no photoblinking event, were hand-picked for subsequent data analysis. The background was subtracted using the residual fluorescent intensities in both channels, after the fluorophore has been photobleached.

1.3 Preliminary Analysis of Individual Trajectories

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ be an observed experimental FRET trajectory. We model it as a hidden Markov model (HMM):

$$y_i \,|\, (z_i = k) \sim N(\mu_k, \sigma_k^2), \tag{1.1}$$

where $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ are the hidden Markov states, evolving according to a K-state Markov chain. Although, rigorously speaking, the FRET value y_i is between 0 and 1, the Gaussian assumption is widely used and accepted in the single-molecule FRET literature in that with moderate observational noise Gaussian distribution is a good approximation (Dahan et al., 1999; McKinney et al., 2006; Liu et al., 2010).

The distinct states of z_i, K in total, model the different conformations of a biological complex. A conformation is a specific 3D structure of a protein or a protein complex. For example, the low- and high-FRET states in C and D of Figure 1.4 correspond to two distinct conformations of the SRP-SR complex. Let $\boldsymbol{P} = (P_{ij})$ be the $K \times K$ transition matrix of $\boldsymbol{z}$; it represents the conformational kinetics of a complex. For each FRET trajectory, the parameters are $\boldsymbol{\theta} = (\boldsymbol{P}, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$, where μ_k and σ_k^2 are the mean and variance of the FRET value at state k; $k = 1, \cdots, K$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ be the probabilities that the first hidden state z_1 is in state $1, \cdots, K$. The joint likelihood of observations $\boldsymbol{y}_{1:N}$ and the hidden states $\boldsymbol{z}_{1:N}$ is

$$p(\boldsymbol{y}_{1:N}, \boldsymbol{z}_{1:N}|\boldsymbol{\theta}) = \pi_{z_1} \prod_{n=2}^{N} p(z_n|z_{n-1}, P) \prod_{n=1}^{N} p(y_n|z_n, \boldsymbol{\mu}, \boldsymbol{\sigma}^2).$$

Please note that for notational ease, we use $\boldsymbol{y}_{m:n}$ to denote the vector $(y_m, y_{m+1}, \ldots, y_n)$ for $m < n$ throughout this article. The marginal likelihood $L(\boldsymbol{\theta}|\boldsymbol{y}_{1:N}) = \int p(\boldsymbol{y}_{1:N}, \boldsymbol{z}_{1:N}|\boldsymbol{\theta}) d\boldsymbol{z}_{1:N}$ is given by integrating out $\boldsymbol{z}_{1:N}$ in the joint likelihood.

1.3.1 Infer the Parameters with a Given Number of Total States

For each FRET trajectory, for a given K, we can use the Baum-Welch algorithm (Baum & Petrie, 1966; Baum et al., 1970), or equivalently, the EM algorithm (Dempster et al., 1977), to calculate the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$. The Baum-Welch/EM algorithm, in addition, can yield the marginal likelihood evaluated at the MLE, $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}_{1:N})$. Appendix A.1 gives the details of our implementation of the algorithm, which uses the forward-backward algorithm.

Alternatively, taking a Bayesian perspective, we can use the Gibbs sampler (Geman & Geman, 1984) together with data augmentation (Tanner & Wong, 1987) to jointly draw posterior samples of the parameters and the hidden states. This gives the posterior distribution (instead of point estimates) of the parameters. Appendix A.2 gives the details of our implementation of the Gibbs sampler with data augmentation.

1.3.2 Detecting the Number of Hidden States

At the molecular level, the total number of states K corresponds to the number of conformations accessible to the complex in the experimental duration. The two conformations in C and D of Figure 1.4 have already been identified in previous studies, and one of our aims is to detect if there are more conformations involved in the protein targeting process (Shen et al., 2012). Statistically, we want to find the K that can "best" explain the variability of the observed FRET trajectories. As an exploratory analysis, we fit each FRET trajectory with the Baum-Welch/EM algorithm for $K = 1, 2, 3, \ldots$ and find that when $K \geq 6$, the hidden states become highly non-identifiable in that the difference of the means of neighboring hidden states are less than 10% of their corresponding standard deviations, which are not experimentally meaningful; and the variance parameters converge to zero, the boundary of the parameter space. Thus, the candidates are $K = 1, 2, 3, 4, 5$ for our data.

Determining K for each trajectory is a model selection problem. Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) are two popular model selection methods. It is well observed in the literature that AIC has a tendency to overestimate the number of mixture components (Windham & Cutler, 1992; Hawkins et al., 2001; Frühwirth-Schnatter, 2006), which we also observe in our simulations. Thus, we focus on using the BIC in our study, which is known to be consistent (as the sample size goes to infinity) for mixture models (McLachlan & Peel, 2005; Frühwirth-Schnatter, 2006; Biernacki et al., 1998; Leroux, 1992a).

Though the consistency of BIC for Gaussian HMMs has not been completely established (Cappe et al., 2005; Finesso, 1990; Rydén, 1995), it has been shown through simulations that BIC empirically tends to select the correct model when the sample size is large but could give highly variable results when the sample size is small or moderate (Celeux & Durand, 2008; Rydén, 1995; MacKAY, 2002; Watkins & Yang, 2005; Frühwirth-Schnatter, 2006; Keribin, 2000). In the context of FRET trajectories, the variability of BIC for HMMs has also been observed (van de Meent et al., 2014; Blanco & Walter, 2010; Keller et al., 2014). The general recommendation in the statistics literature and in the FRET literature for the state-selection of HMM is to use BIC as a first step of preliminary analysis and then assess the selection result based on scientific and experimental insight (McKinney et al., 2006; Greenfeld et al., 2012; Bulla et al., 2010; Keller et al., 2014; Celeux & Durand, 2008). We adopt this recommendation.

In our case of a K-state HMM, the BIC statistic, denoted by BIC_K, is

$$BIC_K = -2 \log L(\hat{\boldsymbol{\theta}} | \boldsymbol{y}_{1:N}) + \log N \times (K^2 + 2K - 1),$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and $K^2 + 2K - 1$ is the total number of parameters: $K^2 - K$ for the transition matrix, $2K$ for the mean and variance parameters, $K - 1$ for the initial distribution of the first hidden state. Minimizing BIC_K over K gives the BIC selection of K for each trajectory. There are two potential issues with the computation of the BIC statistics: (i) the Baum-Welch/EM algorithm converges to local maximum (Baum et al., 1970; Dempster et al., 1977), and (ii) the likelihood function is unbounded at the boundary of the parameter space for Gaussian mixture models (Chen & Li, 2009). These problems make the choice of initial points of the Baum-Welch/EM algorithm critical (Frühwirth-Schnatter, 2006). We treat them by starting the Baum-Welch/EM algorithm from more than 500 randomly generated initial points: the initial values of the mean parameters $\boldsymbol{\mu}$ are uniformly generated from $[0, 1]$, the initial values of each row of the transition matrix P and the distribution $\boldsymbol{\pi}$ of the first hidden state are independently generated from the Dirichlet distribution with concentration parameters all equal to 1, and the initial values of the standard deviations $\boldsymbol{\sigma}$ are independently generated from uniform distribution on $[0.01, 0.3]$; these distributions are employed based on the scientific knowledge of the plausible ranges of the parameters. For each of the 500+ initial values, we run the Baum-Welch/EM algorithm until convergence. The minimum of the BIC statistic over the 500+ algorithm outputs is taken as the value of the BIC for model selection. Table 1.3 tallies the BIC selection of K for the experimental FRET trajectories. Note that we put the *Ffh-* and *FtsY-Data* together in the first row as they are both designed to study the SRP-SR interaction by itself.

	No. of trajectories allocated				
Data	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
Ffh, FtsY-Data	1	21	**159**	136	33
Translocon-Data	2	13	**75**	44	4
RNC-Data	**92**	3	1	1	0

Table 1.3: The number of trajectories with hidden states K allocated by minimizing BIC_K.

Based on the mode, we select $K = 3$ for the *Ffh-*, *FtsY-* and *Translocon-Data* and $K = 1$ for *RNC-Data*. Using the estimation mode to select K reflects "majority rule", i.e., using the consensus to capture the behavior in majority of the experimental replicates. We note that this approach has in fact been proposed in the chemistry literature: Watkins & Yang (2005) showed through simulation and real data studies that it gives a highly robust estimate of K. Note that although we cannot totally rule out the possibility of 4 or more hidden states for some trajectories, we have enough evidence that 3 is the minimum number of K, which the majority of trajectories support. We will see later (in Section 4.2) that $K = 3$ is well supported by the fitting of all the trajectories.

1.4 Modeling FRET Trajectories with Hierarchical Hidden Markov Model

The analysis of individual FRET trajectories reveals that they could have significantly different $\boldsymbol{\theta}$. For instance, a likelihood-ratio test on the three trajectories in Figure 1.2, which are from the *Ffh-Data*, gives a p-value smaller than 0.01, soundly rejecting the hypothesis that the three trajectories share the same $\boldsymbol{\theta}$.

Biologically, the trajectories from replicate experiments under the same condition should reflect the common underlying process. Hence, our goal is to account for the heterogeneity among the experimental trajectories and at the same time to pool information from the trajectories under the same experimental condition. We propose a hierarchical HMM. Suppose $\{\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)}\}$ are the observations and hidden states for trajectory l. We assume that the same transition matrix $\boldsymbol{P}$ is shared by all trajectories; for trajectory l, the means $(\mu_1^{(l)}, \ldots \mu_K^{(l)})$ come from a higher level distribution $\mu_i^{(l)} \sim \mathcal{N}(\mu_{0i}, \eta_{0i}^2)$ with (vector) hyperparameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\eta}_0^2$, and the variances $((\sigma_1^2)^{(l)}, \ldots (\sigma_K^2)^{(l)})$ come from scaled inverse-χ^2 distributions with (vector) hyperparameters $(\boldsymbol{\nu}, \boldsymbol{s}^2)$, where $\boldsymbol{\nu}$ denotes the degrees of freedom and $\boldsymbol{s}^2$ are the scale parameters. The intuition behind this hierarchical HMM is that (i) the transition matrix $\boldsymbol{P}$ represents the conformational kinetics, which is intrinsic to the molecule; it thus should be the same across the trajectories. (ii) The experimental replicates are subject to equipment noise, thermal fluctuation and random variations in experimental samples; the hierarchical structure on $\boldsymbol{\mu}^{(l)}$ and $(\boldsymbol{\sigma}^2)^{(l)}$ reflects it – each trajectory can be considered as a noisy version of the underlying truth. Figure 2.1 diagrams our hierarchical HMM.

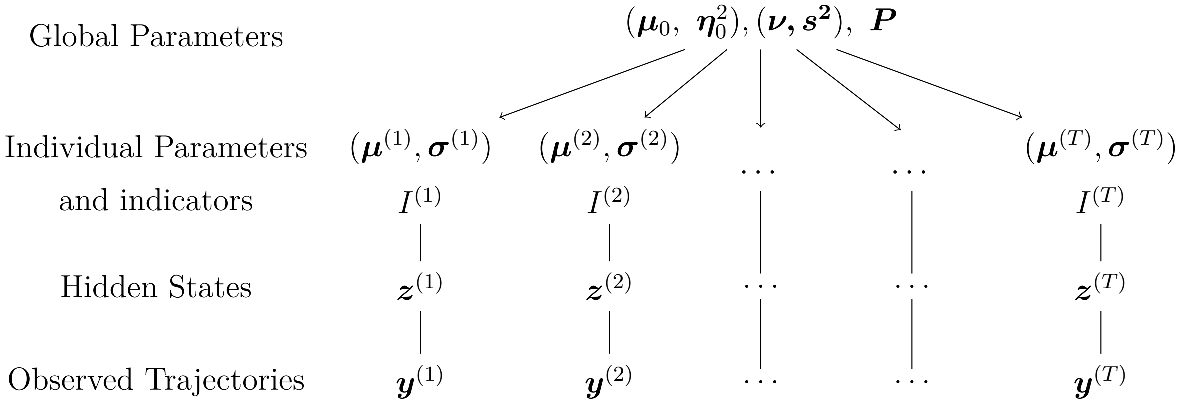
Global Parameters
$(\boldsymbol{\mu}_0,\ \boldsymbol{\eta}_0^2), (\boldsymbol{\nu}, \boldsymbol{s^2}),\ \boldsymbol{P}$
Individual Parameters
and indicators
$(\boldsymbol{\mu}^{(1)}, \boldsymbol{\sigma}^{(1)})$
$(\boldsymbol{\mu}^{(2)}, \boldsymbol{\sigma}^{(2)})$
$(\boldsymbol{\mu}^{(T)}, \boldsymbol{\sigma}^{(T)})$
$I^{(1)}$
$I^{(2)}$
$I^{(T)}$
Hidden States
$\boldsymbol{z}^{(1)}$
$\boldsymbol{z}^{(2)}$
$\boldsymbol{z}^{(T)}$
Observed Trajectories
$\boldsymbol{y}^{(1)}$
$\boldsymbol{y}^{(2)}$
$\boldsymbol{y}^{(T)}$

Figure 1.5: Diagram of the hierarchical HMM.

We note that the real experimental trajectories have different lengths: some are quite short. Within a short experimental time window it is possible that not every conformation shows up — some fast transitions and rare states might be missed in short trajectories. To accommodate this we incorporate a set of indicators into our hierarchical HMM: $I^{(l)}$ indicates which states are present in trajectory l. For example, if the maximum number of states is $K = 3$, $I^{(l)}$ can take four values $I^{(l)} = \{1, 2, 3\}$, $I^{(l)} = \{1, 2\}$, $I^{(l)} = \{1, 3\}$ or $I^{(l)} = \{2, 3\}$, corresponding to the states present in trajectory l. Note that we exclude the singletons (such as $\{1\}$, $\{2\}$ or $\{3\}$) in the set of possible states, since we know from the preliminary analysis of individual trajectories that there are at least two states in each trajectory of the *Ffh-Data*, *FtsY-Data* and

Translocon-Data.

Let $N_{i,j}^{(l)}$ be the number of transitions from state i to j in trajectory l; $N_{i,j}^{(l)} = 0$ if either state i or j does not appear in trajectory l. The likelihood for trajectory l is

$$p(\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)} | \boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}, \boldsymbol{P}, I^{(l)}) = \prod_{i,j=1}^{K} \left(\frac{P_{ij}}{\sum_{k \in I^{(l)}} P_{ik}} \right)^{N_{i,j}^{(l)}} \cdot \prod_{n=1}^{N_l} \mathcal{N}(y_n^{(l)}; \mu_{z_n^{(l)}}^{(l)}, \sigma_{z_n^{(l)}}^{(l)}),$$

where $(\frac{P_{ij}}{\sum_{k \in I^{(l)}} P_{ik}})_{i,j \in I^{(l)}}$ is the re-normalized transition matrix for trajectory l according to which states are present in $I^{(l)}$, and N_l is the length of trajectory l. The likelihood function of all the trajectories (under the same experimental condition) under our hierarchical HMM is

$$\prod_l p(\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)} | \boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}, \boldsymbol{P}, I^{(l)}) p(\boldsymbol{\mu}^{(l)} | \boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2) p((\boldsymbol{\sigma}^{(l)})^2 | \boldsymbol{\nu}, \boldsymbol{s}^2).$$

1.4.1 Estimation under the Hierarchical HMM

To obtain the posterior distribution of the parameters in this model, we use MCMC (Liu, 2001) algorithms. The priors are specified as follows. Each row of the transition matrix $\boldsymbol{P}$ has a flat prior (i.e., a Dirichlet distribution with all parameters equal to 1), which is a proper prior. The global parameters $\boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2$ have flat priors. The categorical variable $I^{(l)}$ also has flat priors, with equal probability of falling into each category. Similar to the Bayesian data augmentation (Tanner & Wong, 1987) procedure for fitting a single trajectory in Appendix A.2, we augment the parameter space $(\boldsymbol{P}; \boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2; \{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}; I^{(l)}\})$ with the hidden states $\{\boldsymbol{z}^{(l)}\}$ and sample from the conditional distributions of these two parts iteratively until convergence. The parameters $(\boldsymbol{P}; \boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2; \{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}; I^{(l)}\})$ are updated one at a time from the conditional distributions using Metropolis-Hastings (for $\boldsymbol{P}$) or Gibbs (for $\boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2; \{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}; I^{(l)}\}$). Conditioning on the parameters $(\boldsymbol{P}, \{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}; I^{(l)}\})$, the hidden states $\{\boldsymbol{z}^{(l)}\}$ are updated sequentially for $l = 1, 2 \ldots$. The details of the sampling procedure are given in Appendix A.3.

Figure 1.6 shows the fitting of our hierarchical HMM with $K = 3$ to two representative FRET trajectories: one long trajectory from the *Ffh-Data* and one short trajectory from the *Translocon-Data*. The grey curves on the top two panels are the observed experimental FRET values. The solid black lines are the fitted values $\{\hat{\mu}_{\hat{z}_n}\}_{n=1}^{N}$, where $\hat{\mu}$ and $\hat{z}_n$ denotes the posterior modes from our MCMC sampling. The lower panel plots the histograms of y_i, the FRET values, of the two FRET trajectories. The black curves overlaid on the histograms are the fittings from our hierarchical HMM, using the posterior mode.

1.4.2 Assessing the Number of Hidden States with the Hierarchical HMM

The posterior distribution of the indicator $I^{(l)}$ gives the probability that a given trajectory l contains a specific collection of states. This posterior distribution thus provides a hierarchical-HMM-based method of model selection: we can allocate the number of hidden states for each trajectory based on the posterior mode of $\left|I^{(l)}\right|$, the size of $I^{(l)}$. By combining multiple trajectories and allowing the sharing of information, we potentially obtain more stable model selection results — borrowing information from other trajectories helps identify rarely occurred hidden states for some trajectories. Table 1.4 tallies the hierarchical-HMM based assignment of the number of hidden states for the experimental FRET trajectories. We apply the hierarchical HMM separately with $K = 3$, where the maximum number of states is three, and with

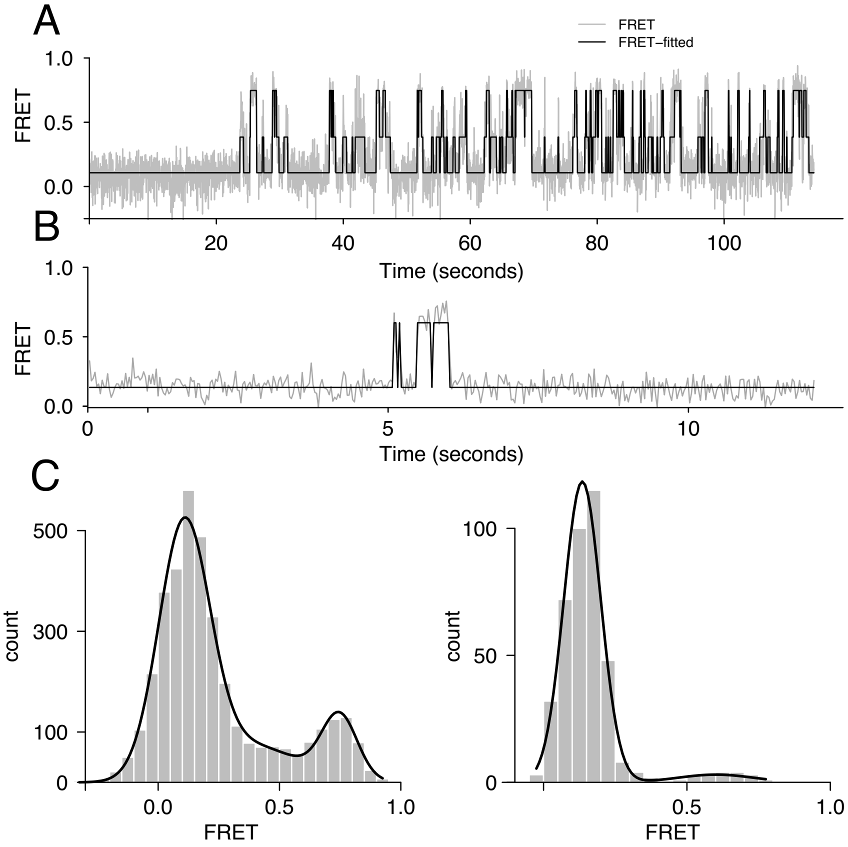

Figure 1.6: Two sample FRET trajectories, one long trajectory from the *Ffh-Data* and one short trajectory from the *Translocon-Data*. The trace plots show the fitted hidden states. The lower panel shows the histograms of the experimental FRET values together with the fitted Gaussian mixtures.

$K = 4$, where the maximum number of states is four. Table 1.4 shows that no matter we set three or four states as the maximum to begin with, the majority of the trajectories are assigned three states. The allocation of states based on the hierarchical HMM, therefore, corroborates our selection of three total states for the *Ffh*-, *FtsY*- and *Translocon-Data*, indicating the robustness of the selection.

	No. of trajectories allocated				
Hierarchical HMM	three states maximum		four states maximum		
No. States	2	3	2	3	4
Ffh, FtsY-Data	56	**294**	26	**201**	123
Translocon-Data	39	**99**	50	**60**	28

Table 1.4: Number of trajectories from the *Ffh/FtsY-Data* and *Translocon-Data* assigned to $2, 3, 4$ hidden states based on the posterior mode of $\left|I^{(l)}\right|$. The hierarchical HMM was fitted twice with three states maximum and four states maximum, respectively. As in Section 1.3.2, we put the *Ffh-Data* and *FtsY-Data* together in the table.

1.4.3 Hierarchical Fitting versus Individual Fitting

It is worth pointing out that by pooling the information from the multiple trajectories, we obtain more robust and reliable estimates. Figure 1.7 shows what happens if we only fit the individual trajectory by itself. The left panel shows the fitting of the 2-state, 3-state and 4-state HMMs to the long trajectory of Figure 1.6(A) alone; the right panel shows the fitting to the short trajectory of Figure 1.6(B) by itself. The individual fitting is seen to be unstable in that it is quite difficult to judge which fitting is better. The hierarchical model, in contrast, allows the information to be pooled from all the trajectories, resulting in stable estimates.

To further compare the fitting under the hierarchical model versus the fitting on individual trajectories and to test the limit of the hierarchical model fitting, we conduct

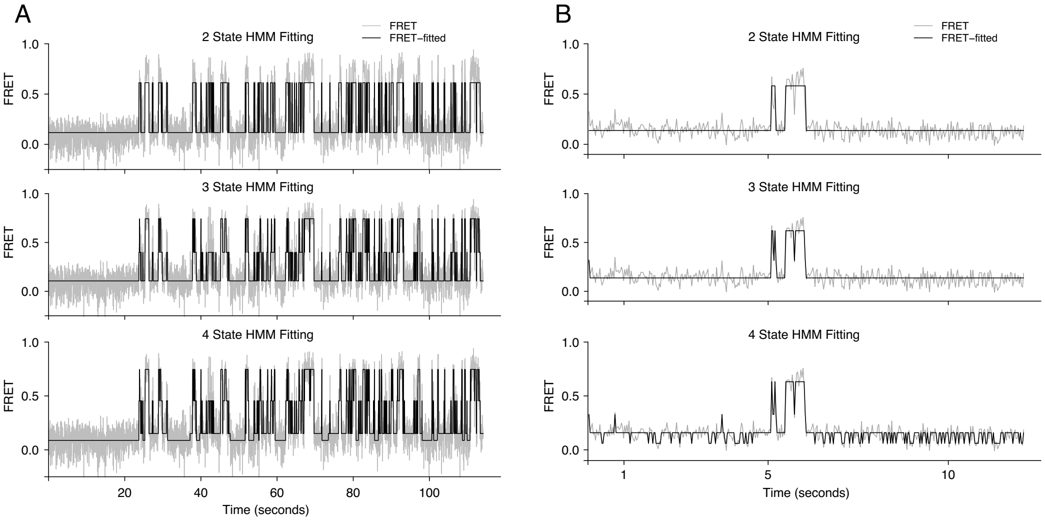

Figure 1.7: Fitting of individual FRET trajectories. The left column (A) shows the fitting of the 2-state, 3-state and 4-state HMMs to the long trajectory of Figure 1.6(A) alone. The right column (B) shows the fitting of 2-state, 3-state and 4-state HMMs to the short trajectory of Figure 1.6(B) by itself.

a sequence of simulations. The mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ is generated according to $\mu_1 \sim \mathcal{N}(0.1, 0.1^2)$, $\mu_2 \sim \mathcal{N}(0.4, 0.1^2)$, $\mu_3 \sim \mathcal{N}(0.7, 0.1^2)$. The standard deviation vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ is taken to be $\sigma_1 = \sigma_2 = \sigma_3$. Trajectories each with length $N = 1000$ are generated from a three-state HMM with transition matrix with diagonal elements equal to 0.9 and off-diagonal elements equal to 0.05. For each value of $\sigma_1 = \sigma_2 = \sigma_3 \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8\}$, we repeat the data generation 100 times, so we have 16 sets of simulated data, each set containing 100 trajectories with length 1000.

For each of the 16 sets of simulated data, we apply the hierarchical fitting as well as the individual fitting. Intuitively, as the hierarchical HMM pools information from multiple trajectories, it is able to handle data with much lower signal-to-noise-ratio (SNR) than the fitting of HMM to individual trajectories. Figure 1.8 provides an illustration, showing the results for the case of $\sigma_1 = \sigma_2 = \sigma_3 = 0.65$. The left panel compares the estimation of the global means $\boldsymbol{\mu}_0 = (0.1, 0.4, 0.7)$. The right panel compares the estimation of the transition probabilities P_{11}, P_{22}, P_{33}. In each panel, the left half shows the posterior distribution under the hierarchical HMM, and the right half shows the aggregated posterior distribution based on fitting the 3-state HMM to individual trajectories. It is evident that individual fitting gives highly variable and biased estimates; in contrast, by pooling the information from the 100 trajectories together, the hierarchical fitting gives much more reliable and accurate estimates.

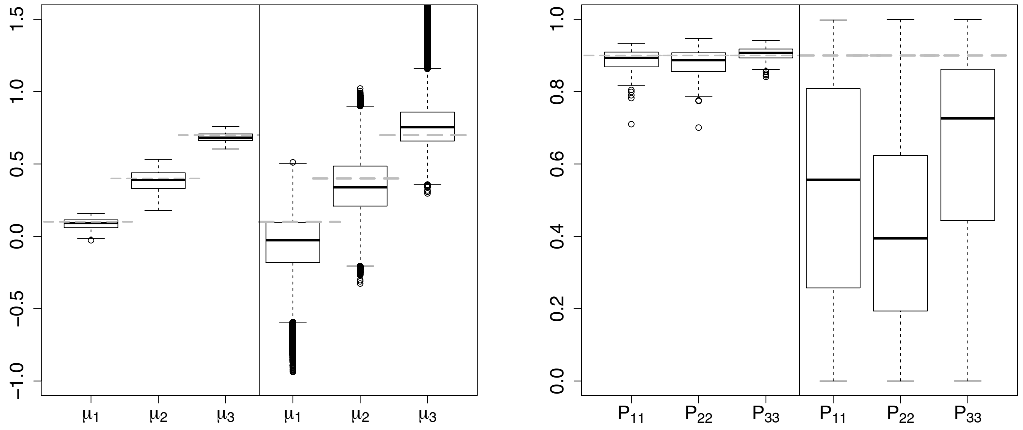

Figure 1.8: Comparison of fitting of the hierarchical HMM versus the fitting of individual trajectories. The left panel compares the estimation of the global means $\boldsymbol{\mu}_0$. The right panel compares the estimation of the transition probabilities P_{11}, P_{22}, P_{33}. Both panels use the boxplots. In each panel, the left half shows the posterior distribution under the hierarchical HMM; the right half shows the aggregated posterior distribution based on fitting the 3-state HMM to individual trajectories. The grey horizontal lines correspond to the true values of the parameters.

Formally, for each trajectory we can define SNR as $SNR = \min_k\{\frac{\mu_{k+1}-\mu_k}{\sigma_k}, \frac{\mu_{k+1}-\mu_k}{\sigma_{k+1}}\}$ (Greenfeld et al., 2012; Hawkins et al., 2001). For the 100 trajectories of Figure 1.8, the median SNR is 0.3. In contrast, we find from our 16 simulated data sets that for individual fitting to give meaningful result, the median SNR has to be as high as 2.0. As the standard deviation increases, the SNR decreases. Intuitively, as the SNR becomes smaller and smaller, eventually the hierarchical model fitting will start to break down. In our simulation, we observe that the breakdown happens at $\sigma_1 = \sigma_2 = \sigma_3 = 0.7$, where the median SNR is less than 0.3. This number is in sharp contrast with the SNR limit of around 2.0 for the individual trajectory fitting. For the experimental data, the median SNR is 1.47 for the *Ffh-Data*, 1.36 for the *FtsY-Data*, and 1.46 for the *Translocon-Data*; all three are below the SNR limit of around 2.0 for reliable individual-trajectory fitting.

1.5 Resolving the Biological Questions

Based on our analysis of the single-molecule FRET data, we will address in this section the unsolved questions regarding the detailed mechanism of the protein targeting process put forward in Section 1, delineating the roles of different components in the protein targeting process. We will consider first the conformation change of the SRP-SR complex without RNC or translocon, and then the effect of RNC and translocon in regulating the protein targeting process. Based on the results of our data analysis, we will propose a refined mechanism for co-translational protein targeting process, addressing the biological puzzles.

It is worth pointing out that the hierarchical structure enables us to include heterogeneous trajectories in a single model, capturing common characteristics while allowing for individual variabilities. Our analysis allows us to distinguish between two possibilities that could give rise to the heterogeneous FRET trajectories: (i) heterogeneity of sample, meaning that the SRP-SR complex can exist in distinct populations that have different structural and chemical properties, therefore exhibiting different kinetic and equilibrium behaviors; and (ii) intrinsic noise due to the stochastic nature and molecular reactions and limited time scale for sampling in single-molecule experiments. Our result supports that the heterogeneous trajectories are well explained by (ii).

1.5.1 Conformational Change of the SRP-SR Complex

The *Ffh-Data* and *FtsY-Data* are obtained from the single-molecule FRET experiments on the SRP-SR complex in the absence of RNC or translocon. The only difference between these two datasets is the placement of the FRET donor. For the *Ffh-Data* the FRET donor is placed at Ffh-NG, while for the *FtsY-Data* the FRET donor is placed at FtsY; see Figure 1.4 and Table 1.1. These data reveal the conformational fluctuation of the SRP-SR complex without RNC or translocon.

As we described in Sections 1.3.2 and 4.2, three FRET states are detected, corresponding to three conformations. For these three conformations, Table 1.5 lists the 95% posterior intervals of the global parameters μ_{0i} and η_{0i} for the data sets. The state with a low FRET value, $\mu_{0,1} \approx 0.1$, corresponds to the conformation where the FtsY-[Ffh-NG] complex is near the capped end of the RNA (see C of Figure 4). The state with a high FRET value, $\mu_{0,3} \approx 0.6 \sim 0.8$, corresponds to the conformation where the FtsY-[Ffh-NG] complex is near the distal end of the RNA (see D of Figure 4). It is noteworthy that in addition to these two major conformations, our analysis identifies a “middle” state with the FRET value $\mu_{0,2}$ around 0.3 to 0.4, suggesting a third conformation of the SRP-SR complex. This conformation might correspond to alternative modes of docking of the FtsY-[Ffh-NG] complex at the RNA distal end (in which FtsY-[Ffh-NG] is oriented differently relative to the RNA), given the relative large value of $\mu_{0,2}$, or an alternative binding site of the FtsY-[Ffh-NG] complex on the RNA (Shen et al., 2012). As we shall see shortly, this conformation could serve as an intermediate stage that mediates the large scale movement of the FtsY-[Ffh-NG] complex, which travels 100 Å from the RNA capped end to the distal end.

Parameters	Ffh-Data	FtsY-Data	RNC-Data	Translocon-Data
$\mu_{0,1}$	[0.105, 0.116]	[0.096, 0.107]	[0.091, 0.099]	[0.097, 0.104]
$\mu_{0,2}$	[0.319, 0.353]	[0.348, 0.382]	NA	[0.380, 0.441]
$\mu_{0,3}$	[0.619, 0.646]	[0.733, 0.761]	NA	[0.619, 0.635]
$\eta_{0,1}$	[0.039, 0.048]	[0.041, 0.049]	[0.017, 0.022]	[0.019,0.023]
$\eta_{0,2}$	[0.110, 0.135]	[0.122, 0.148]	NA	[0.131, 0.169]
$\eta_{0,3}$	[0.087, 0.107]	[0.101, 0.126]	NA	[0.044, 0.058]

Table 1.5: 95% posterior intervals of the global means μ_{0i} and global standard deviations η_{0i}; $i \in \{1, 2, 3\}$ for *Ffh-Data*, *FtsY-Data*, *RNC-Data* and *Translocon-Data*.

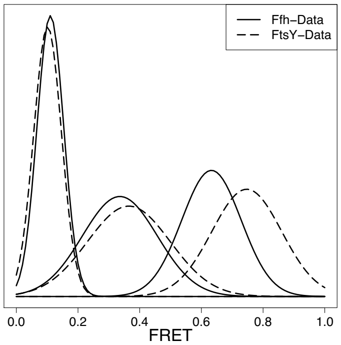

Figure 1.9: The posterior distributions of the mean parameters for the *Ffh-Data* and *FtsY-Data*.

Figure 1.9 compares the distributions of the mean parameters for the *Ffh-Data* to those for the *FtsY-Data*. It is also interesting to note from both Table 1.5 and Figure 1.9 that the FRET value $\mu_{0,3}$ of the *FtsY-Data* is higher than that of the *Ffh-Data*. This implies that FtsY is closer to the distal end than Ffh-NG is when the FtsY-[Ffh-NG] complex docks at the distal end. It thus gives a fine picture of the relative positions of FtsY and Ffh-NG as shown in Figure 1.4. This is consistent with findings from the crystal structures of the SRP-SR complex (Ataide et al., 2011; Voigts-Hoffmann et al., 2013).

The conformational change that SRP-SR undergoes on the RNA is unusually large, spanning over 90 Å. How this large-scale movement occurs is an interesting question. It is possible that the complex travels along the RNA via "intermediate" stops. Alternatively, the complex could constantly sample alternative potential docking sites on the RNA until it finds the distal site. The transitions among different states capture the pathways and mechanisms by which the SRP-SR complex undergoes the large-scale conformation change. Table 1.6 shows our estimates of the transition probabilities $\{P_{ij}\}$ for the data sets. We note that the estimates of the transition probabilities from the *Ffh-Data* are similar to those from the *FtsY-Data*.

Data	Ffh-Data	FtsY-Data	Translocon-Data
P_{11}	0.9703 ± 0.0014	0.9798 ± 0.0013	0.9976 ± 0.0005
P_{22}	0.8732 ± 0.0054	0.8776 ± 0.0058	0.9713 ± 0.0076
P_{33}	0.9384 ± 0.0027	0.9217 ± 0.0039	0.9870 ± 0.0015
P_{12}	0.0283 ± 0.0014	0.0186 ± 0.0015	0.0011 ± 0.0004
P_{13}	0.0015 ± 0.0005	0.0015 ± 0.0005	0.0013 ± 0.0004
P_{21}	0.0587 ± 0.0034	0.0579 ± 0.0044	0.0044 ± 0.0015
P_{23}	0.0681 ± 0.0036	0.0646 ± 0.0037	0.0244 ± 0.0072
P_{31}	0.0029 ± 0.0010	0.0057 ± 0.0017	0.0022 ± 0.0006
P_{32}	0.0587 ± 0.0031	0.0726 ± 0.0045	0.0108 ± 0.0015

Table 1.6: Posterior estimates of the transition probabilites (mean ± 2× standard deviations) of *Ffh-Data*, *FtsY-Data*, *Translocon-Data* based on the hierarchical model fitting.

We next investigate the functional role of the middle state based on the posterior distributions of $\{P_{ij}\}$ for the *Ffh-Data*. First, we obtain the 95% credible interval of $d_i = 1/(1 - P_{ii})$, the mean dwell time at state i. The intervals are [0.966, 1.057] seconds for d_1, the low-FRET state; [0.228, 0.249] seconds for d_2, the middle state; and [0.465,0.507] seconds for d_3, the high-FRET state. The observation that both d_1 and d_3 are significantly larger than d_2 indicates that the SRP-SR complex spends less time at the middle state than at the low- or high-FRET state, which are more stable.

Second, it is known that biologically the SRP-SR complex initially assembles at the RNA capped end and the complex disassembles at the RNA distal end (Shen & Shan, 2010). Thus, a "complete transition" is the one that goes from the low-FRET state to the high-FRET state (see Figure 1.4). The observation that P_{13} is significantly smaller than P_{12} suggests that a direct transition from the low-FRET state to the high-FRET state is quite infrequent; rather, a "complete transition" more frequently proceeds through the middle state. In other words, without RNC or the translocon, the FtsY-[Ffh-NG] complex usually travels from the capped end to the distal end through an intermediate stage.

In fact, we can calculate the probability that a final passage from state 1 to state 3 goes through state 2 versus the probability that such a final passage does not go through state 2 as follows. For $i, j = 1, 2$, let us use $P^{(k)}_{i \to j}$ to denote the probability of transition from state i to state j in k steps without ever reaching state 3. Then the probability of going from state 1 to state 3 finally through state 2 is $\sum_{k=1}^{\infty} P^{(k)}_{1 \to 2} P_{23}$ (i.e., taking any number of steps between state 1 and 2 and then finally reaching state 3 from state 2 in the last step). The probability of going from state 1 to state 3 not finally through state 2 is $P_{13} + \sum_{k=1}^{\infty} P^{(k)}_{1 \to 1} P_{13}$. $P^{(k)}_{i \to j}$ satisfies the following recursive formulas, owing to the first-step analysis:

$$\begin{cases} P^{(k+1)}_{1\to 2} = P_{11} P^{(k)}_{1\to 2} + P_{12} P^{(k)}_{2\to 2} \\ P^{(k+1)}_{2\to 2} = P_{21} P^{(k)}_{1\to 2} + P_{22} P^{(k)}_{2\to 2} \end{cases} \qquad \begin{cases} P^{(k+1)}_{1\to 1} = P_{11} P^{(k)}_{1\to 1} + P_{12} P^{(k)}_{2\to 1} \\ P^{(k+1)}_{2\to 1} = P_{21} P^{(k)}_{1\to 1} + P_{22} P^{(k)}_{2\to 1} \end{cases}$$

Summing over k on both sides of the equations yields

$$\begin{aligned} \textstyle\sum_{k=1}^{\infty} P_{1\to 2}^{(k)} P_{23} &= \tfrac{P_{12}P_{23}}{(1-P_{11})(1-P_{22})-P_{12}P_{21}} \\ P_{13} + \textstyle\sum_{k=1}^{\infty} P_{1\to 1}^{(k)} P_{13} &= \tfrac{(1-P_{22})P_{13}}{(1-P_{11})(1-P_{22})-P_{12}P_{21}} \end{aligned} \tag{1.2}$$

From these formulas and the posterior distributions of P_{ij}, we find that 91.2% of the transitions from state 1 to state 3 occurs finally through the intermediate state 2 for the *Ffh-Data*.

These observations and calculations reveal that (i) the movement of the FtsY-[Ffh-NG] complex from the RNA capped end to the distal end requires the middle state, which serves as an on-pathway intermediate to facilitate this largescale movement. (ii) The middle state is quite efficient in facilitating the search for the RNA distal site: once the SRP-SR complex reaches this state, over 50% of molecules move on successfully to the distal site (high-FRET state) (because $P_{23} > P_{21}$); this over 50% probability is much higher than that from the low-FRET state.

1.5.2 Effect of RNC

Once RNC is added to the SRP-SR complex, the experimental FRET trajectories, the *RNC-data*, show the presence of only *one* state with a low FRET value: the FRET values are well fitted by $y_i = const +$ Gaussian noise, see Table 1.5. Comparison of these results with those on SRP-SR alone (the *Ffh-Data* and *FtsY-Data*) show that the RNC has a pausing effect: it holds the SRP-SR complex near the capped end and prevents its movement to the RNA distal end (see C of Figure 4). This pausing effectively prevents premature dissociation of SRP and SR, which happens at the distal end of the SRP RNA and results in abortive reactions. We thus see that RNC plays an important regulating role in ensuring the efficiency of a successful protein targeting.

1.5.3 Role of Translocon

When the translocon is further added to the RNC-SRP-SR complex, single-molecule experiments on the translocon-RNC-SRP-SR complex yield the *Translocon-Data* in Table 1.1. As shown in Table 1.5, the high-FRET state ($\mu_{0,3} \approx 0.6$) is restored in the *Translocon-Data*, which is completely absent in the *RNC-Data*. Therefore, the translocon enables the FtsY-[Ffh-NG] complex to restore movement to the RNA distal end, where disassembly of SRP-SR (by GTP-hydrolysis) can be initiated.

We also observe that the transition probabilities of the *Translocon-Data*, shown in Table 1.6, differ significantly from those of the *Ffh-Data*. This rules out the model that the translocon simply awaits for and binds the RNC that has spontaneously dissociated from the SRP-SR complex. If this were the case, the FRET trajectories in the presence of both RNC and translocon (the *Translocon-Data*) would exhibit nearly identical features as those for the SRP-SR complex (the *Ffh-Data*). Instead, these data strongly suggest that the translocon forms a quarternary complex together with RNC, SRP and SR, in which attainment of the distal conformation is favored.

We next consider the role of the middle state. Using formula (1.2) derived in Section 1.5.1, we find that *only* 40.7% of the transitions from the low FRET to high FRET state occur via the middle state as an intermediate for the *Translocon-Data*. This is in sharp contrast with the 91.2% probability for the *Ffh-Data*. This indicates that the translocon alters the pathway via which the FtsY-[Ffh-NG] complex searches for the RNA distal site, biasing them towards pathways in which transitions between low FRET and high FRET states occur directly. We note that it is possible that in the presence of translocon, the residence in the intermediate state could be too fast to be detected within the time resolution (30 ms) of the experiment.

To gain further insights into the regulatory role of the translocon, we asked whether and how it alters the kinetics by which the SRP-SR complex undergoes the structural change. To this end, we compare the dwell time of the FtsY-[Ffh-NG] complex at the high-FRET state, which is $d_3 = 1/(1 - P_{33})$, between the *Translocon-Data* and the *Ffh-Data*. The 95% posterior interval for d_3 is $[2.058, 2.577]$ seconds for the *Translocon-Data* and $[0.465, 0.507]$ seconds for the *Ffh-Data*, respectively. Thus, the translocon enhances the kinetic stability of the SRP-SR complex in the distal conformation by 4-5 fold. Table 1.7 contrasts the parameter estimates between the *Ffh-Data* and the *Translocon-Data*.

Parameters	Ffh-Data	Translocon-Data
$\mu_{0,1}$	$[0.105, 0.116]$	$[0.097, 0.104]$
$\mu_{0,2}$	$[0.319, 0.353]$	$[0.380, 0.441]$
$\mu_{0,3}$	$[0.619, 0.646]$	$[0.619, 0.635]$
d_3	$[0.465, 0.507]$	$[2.058, 2.577]$
p_{middle}	91.2%	40.7%

Table 1.7: Compare *Ffh-Data* and *Translocon-Data*: 95% posterior intervals of mean values of the states ($\mu_{0,1}, \mu_{0,2}, \mu_{0,3}$), dwell time at the high-FRET state (d_3) and the probability that a transitions from low- to high-FRET state goes through the middle state (p_{middle}).

In summary, our statistical analysis shows that the translocon regulates the protein targeting process by (i) restoring the movements of the FtsY-[Ffh-NG] complex to the RNA distal end, (ii) promoting alternative pathways for this movement, in which the FtsY-[Ffh-NG] complex directly transitions from the low-FRET state to the high-FRET state, and (iii) prolonging the time that FtsY-[Ffh-NG] stays at the RNA distal end. It is known that movement of the FtsY-[Ffh-NG] complex away from the RNA capped end is important for vacating the ribosome binding site and initiating ribosome-translocon contacts during the handover of RNC to the translocon. It is also known that GTP-hydrolysis, which disassembles SRP and SR, occurs at the RNA distal end (Shen et al., 2013). Our findings thus reveal that the translocon, via mechanisms (i)-(iii), promotes both of these molecular events and allows them to be synchronized in the pathway. Collectively, these results show that the translocon not only serves as a channel through which the nascent proteins translocate, but also facilitates the productive handover of the RNC onto itself to complete the protein targeting reaction.

1.5.4 A Proposal of Detailed Mechanism

Our statistical analysis of the single-molecule experimental data in combination with the known biological understanding (Halic et al., 2006; Pool et al., 2002; Peluso et al., 2001; Estrozi et al., 2011; Shen & Shan, 2010; Zhang et al., 2009a; Akopian et al., 2013a; Ataide et al., 2011) suggests the following detailed mechanism of protein targeting, which was conjectured in Shen et al. (2012), corresponding to the four steps of Section 2.1:

1. SRP recognizes the signal sequence on RNC and binds it. The RNC is delivered to the target membrane where the SR can localize to.

2. When the SRP-SR complex is initially formed, the FtsY-[Ffh-NG] complex binds at the RNA capped end near the ribosome exit site, blocking the site from translocon binding.

3. As the RNC initiates contact with the translocon, the latter actively facilitates the conformation change of SRP-SR complex and drives the FtsY-[Ffh-NG] complex from the capped end to the distal end of RNA.

4. GTP-hydrolysis is initiated at the RNA distal end to disassemble the SRP and SR. Meanwhile, the nascent chain is released from the Ffh M-domain to the translocon on the membrane.

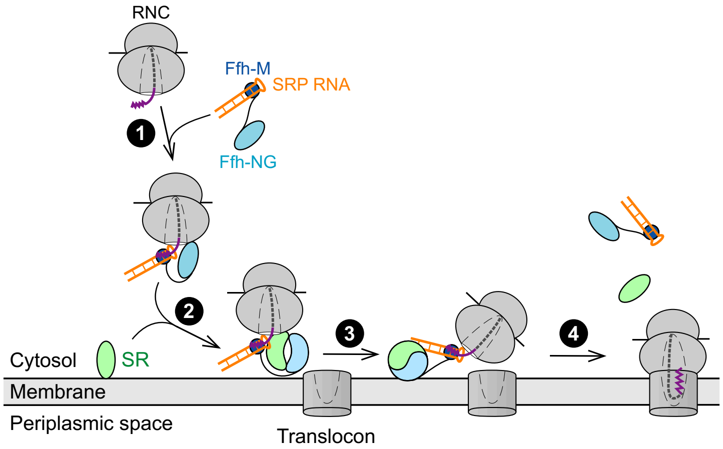

Figure 1.10: The refined mechanism. Steps 1 & 2: SRP binds RNC at the RNA capped end and carries it to the membrane by forming a complex with SR located at the membrane. Step 3: The FtsY-[Ffh-NG] complex goes to the distal end so that RNC can be loaded at the translocon. Step 4: SRP-SR disassembles through GTP-hydrolysis and the nascent chain goes through the translocon on the target membrane.

Figure 1.10 illustrates the detailed mechanism. The movement of the FtsY-[Ffh-NG] complex from the RNA capped end to the distal end is first negatively regulated by RNC, whose pausing effect keeps the SRP-SR complex from disassembly before the translocon is identified, and later positively regulated by the translocon, which actively facilitates the movement of FtsY-[Ffh-NG] to the RNA distal end. This mechanism allows the coordinated exchange of SRP and translocon at the RNC and the effective timing of GTP-hydrolysis, thus minimizing abortive reactions due to premature SRP-SR disassembly or non-productive loss of the RNC.

1.6 Model Checking

1.6.1 Check of Detailed Balance

In biophysics, the principle of microscopic reversibility states that at equilibrium the transition flux between any two states should be equal. In the familiar probability language, the microscopic reversibility translates into the detailed balance condition or the reversibility of the Markov chain: $\pi_i P_{ij} = \pi_j P_{ji}$ for all i and j, where π_i is the equilibrium probability of state i. This can be checked from the posterior samples of the transition matrix $\boldsymbol{P}$.

Figure 1.11 compares the distribution of $\pi_i P_{ij}$ (first column) with that of $\pi_j P_{ji}$ (second column) from the *Ffh-Data*. The third column shows the distribution of the difference $\pi_i P_{ij} - \pi_j P_{ji}$ compared to zero (the vertical bar), where $i, j \in \{1, 2, 3\}, i \neq j$. It is clear that $\pi_i P_{ij} - \pi_j P_{ji} = 0$ holds within the experimental error. The plots on the *FtsY-Data* and the *Translocon-Data* give very similar pictures. We thus confirm that indeed under our hierarchical HMM the principle of microscopic reversibility is satisfied.

1.6.2 Check of Markovian Assumption

In our hierarchical HMM, the Markov assumption of the state transitions (or the conformation changes) plays a fundamental role. If the Markov assumption is correct, then the waiting time at the individual state should be exponentially distributed and that the successive waiting times should be independent of each other. Both can be checked under our Bayesian sampling approach, since we can straightforwardly obtain the waiting time at each state from the posterior samples of the hidden states $\boldsymbol{z}$. Fig-

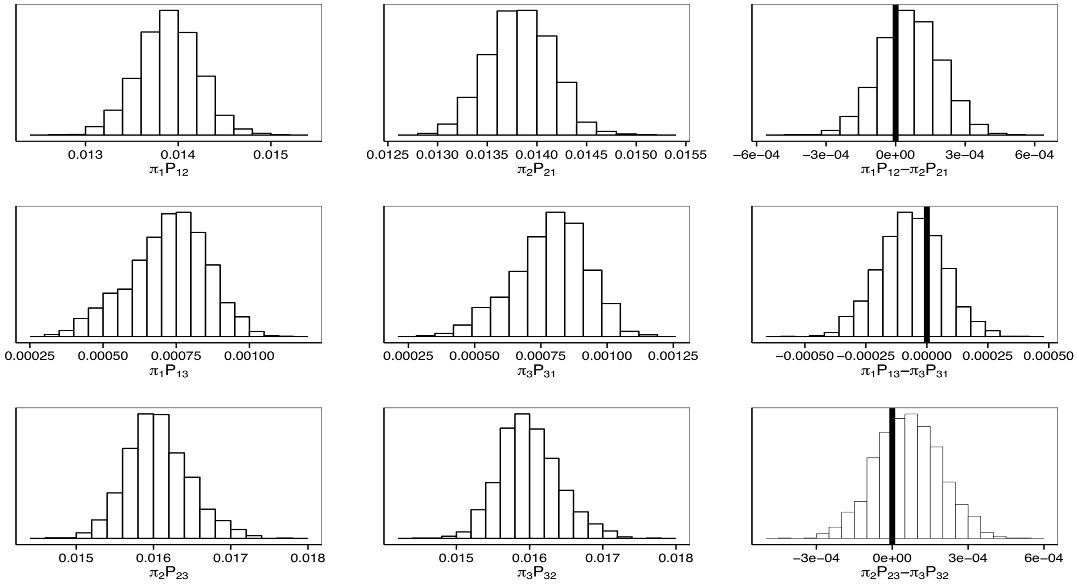

Figure 1.11: Check of detailed balance for the *Ffh-Data*. The first column is the posterior distribution of $\pi_i P_{ij}$, and the second column is that of $\pi_j P_{ji}$, where $i, j \in \{1, 2, 3\}$, $i \neq j$. The third column shows the distribution of their difference $\pi_i P_{ij} - \pi_j P_{ji}$; the thick vertical bar is at zero.

ure 1.12 shows the posterior distribution of the waiting time at each of the low, middle and high FRET state of the *Ffh-Data* based on the samples of hidden states $\boldsymbol{z}$ in its original scale (left column) and the log-scale (right column). It is seen that on the log-scale the distribution of the waiting time is well fit by a straight line, supporting the exponential distribution. Quantitatively, we performed a chi-squared goodness-of-fit test for the exponential distribution using 30 evenly spaced bins. The resulting p-values for the waiting time at the low, middle and high FRET states are 0.72, 0.20 and 0.35, respectively. Figure 1.13 shows the autocorrelation of the successive waiting times from the *Ffh-Data* obtained from the samples of the hidden states $\boldsymbol{z}$. It is evident that the successive waiting times are uncorrelated, as the Markov assumption requires. The posterior samples from the *FtsY-Data* and the *Translocon-Data* show quite similar pattern.

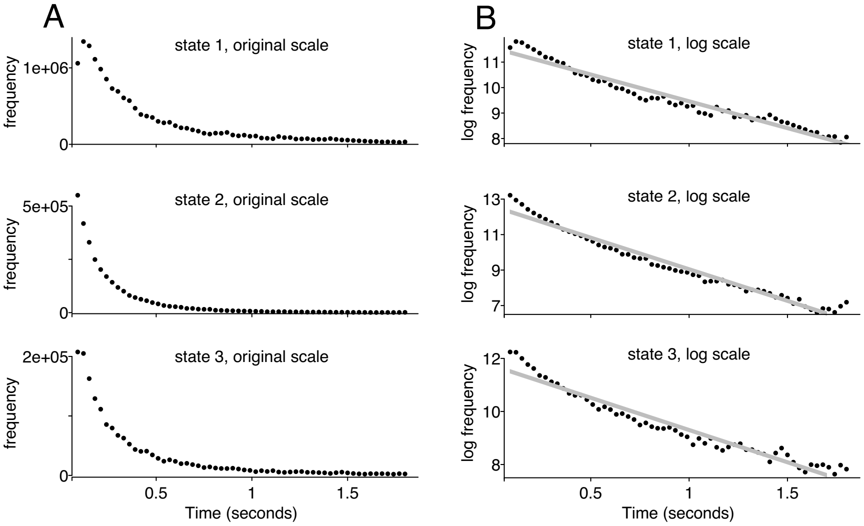

Figure 1.12: Posterior distribution of waiting time at the three states of the *Ffh-Data* on the original scale, the left column (**A**); and the log scale, the right column (**B**).

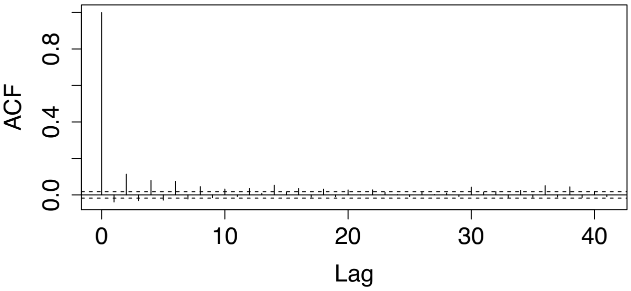

Figure 1.13: Autocorrelation of the successive waiting times from the *Ffh-Data*.

1.7 Summary

The advances in single-molecule experiments enable us to study the detailed mechanism of the co-translational protein targeting process. On the single-molecule level the data are necessarily stochastic. They are often noisy realizations of the underlying stochastic dynamics. To model the stochasticity of each individual experimental trajectory, we use HMM.

The experimental time windows in single-molecule trajectories are often of rather limited length, resulting in relatively short trajectories. As a result, the parameter estimation based on individual trajectories could be quite variable. Furthermore, the determination of the total number of states of the HMM based on individual trajectories is highly unstable. Experimentally, these issues are mitigated by recording hundreds of trajectories repeated under the same experimental condition. In this article, we use the mode of the BIC selection over multiple trajectories for reliable determination of the number of states of the HMM as a preliminary analysis. Then we propose a hierarchical HMM to pool information together from the different trajectories and at the same time to account for the heterogeneity among them. The heterogeneity among the different trajectories arises from the intrinsically stochastic nature of molecular actions, equipment noise, thermal fluctuation and random variations in experimental setups. We find that the proposed hierarchical HMM is highly robust to low signal-to-noise ratios. Finally, assessment of the fitting of each individual trajectory based on parameters estimated from the hierarchical model re-assured us of the model selection at the first stage and the assumption of the hierarchical model at the second stage.

Biologically, we corroborated many conclusions from the previous ad-hoc analysis, giving solid quantitative evidence for the proposed new mechanism of co-translational protein targeting. Instead of being passively involved in the protein targeting process, our analysis shows that the RNC and translocon play active regulatory roles to facilitate the accurate timing of the biological steps. Specifically, the RNC and translocon effectively regulate the movement of the SRP-SR complex between the capped end and the distal end of the RNA, which in turn regulates the assembly and disassembly of the SRP-SR complex and the preference of the RNC for binding the SRP-SR complex versus the translocon. Compared to the previous ad-hoc analysis, our statistical analysis clarifies the pathway for the structural change in the SRP-SR complex, and rigorously showed that the translocon alters the pathway, kinetics, and stability of this structural change, providing stronger evidence that the translocon actively facilitates the loading of RNC onto itself and drives the completion of protein targeting. From a modeling perspective, the hierarchical HMMs that we used for combining information are quite general. They appear effective for dealing with replicated experiments and can be potentially used for analyzing other biological or biochemical experiments. We thus hope that this article would generate further interest in studying these hierarchical models and in applying them for general data analysis.

In God we trust. All others must bring data.

W. Edwards Deming

As for the future, your task is not to foresee it,
but to enable it.

Antoine de Saint-Exupery

2

Bayesian Computation Package for Single-Molecule Data

2.1 Introduction

Single-molecule experiments are becoming more and more popular in studying the detailed kinetics of biological processes. Unlike traditional experiments which record the signals representing the average properties of ensembles of molecules, single-molecule experiments make it affordable to examine kinetic behaviors of individual molecules. Therefore, more detailed information of a complicated biological process is captured.

One of the most widely adopted biophysical techniques for conducting single-molecule experiments is through measuring the energy transfer rate from an illuminated donor chromophore to an acceptor chromophore, where the donor and acceptor are attached to two designated compartments of a molecule. The measurements, called the fluorescent resonance energy transfer (FRET), tracks in real time the distances between the donor and acceptor at the 1-10 nanometer scale. When the molecule undergoes a biological process, e.g. a sequence of conformational changes of a protein complex, the distance between the donor and acceptor changes over time. In this way, single-molecule experiments using the FRET technique are able to probe the internal dynamics of single molecules in a biological process.

The goal of data analysis is to properly extract relevant information from FRET traces. Due to the stochastic nature of the single-molecule experiments, the model should capture the underlying process while allowing for enough noise coming from various sources: (a) intrinsic noise from stochastic or other unpredictable behaviors of molecules and (b) extrinsic noise from the environment, equipment and measurements. The first source of noise may not even exist in experiments using ensembles of molecules: the molecules that are chemically identical when measured at ensembles would still demonstrate significant variability at the single-molecule level, in a manner predicted by the Boltzmann distribution.

Analyzing FRET traces in statistically principled ways is an increasing trend in the biochemical community. Traditionally, FRET traces are individually fit using hidden Markov models (HMM). In McKinney et al. (2006) and Greenfeld et al. (2012), the

Baum-Welch (or known as the Expectation-Maximization) algorithm (Baum & Petrie, 1966; Baum et al., 1970; Rubin, 1984) is implemented and a Matlab package is provided for fitting a single FRET trace. Schmid et al. (2016) proposes a maximum likelihood method to extract the kinetics from short and out-of-equilibrium FRET traces. For sake of robustness, scientists usually measure multiple FRET traces under each experimental condition, which leads to a more complicated data structure for statistical analysis. van de Meent et al. (2016) combines multiple FRET traces using the coupled hidden Markov models and applies the variational Bayes method to approximate the posterior distribution of the transition kinetics. Chen et al. (2016) proposes the (heterogeneous) hierarchical hidden Markov model, which takes the coupled HMM as a special case, and conducts full Bayesian analysis to understand the detailed later stages of the co-translational protein targeting process.

We address the issues of extracting information from multiple FRET traces from replicated experiments using the Bayesian hierarchical hidden Markov model. We extend the methods applied in Chen et al. (2016) and introduce a Matlab package for conducting full Bayesian analysis of multiple (homogeneous or heterogeneous) FRET traces. The Bayesian hierarchical model on top of hidden Markov models, as implemented in the Matlab package, possesses three desirable characteristics for analyzing multiple FRET traces from experimental replicates. (1) Information from all traces are combined to study the common kinetics among all molecules in the experiment. (2) The allowed variability within each individual trace (molecule) takes into account of the noise coming from the environment, the equipment, and other random fluctuations. (3) The allowed variability among different traces (molecules) takes into account of the inherent stochastic molecular behaviors and uncontrollable variations.

The rest of the paper is organized into three sections. In Section 2.2, we restate the

Bayesian hierarchical HMM proposed in Chen et al. (2016). The Matlab package is introduced in Section 2.3. In Section 2.4, we demonstrate (i) the robustness of fitting the hierarchical HMM using the Matlab package in practical situations like varying FRET lengths, various signal-to-noise ratio, existence of rare events and (ii) the powerfulness of the package in performing automatic and reliable model selection.

2.2 Description of Methodology

In this section, we introduce the Bayesian hierarchical hidden Markov model for multiple heterogeneous FRET traces. This model pools information from the traces under the same experimental condition to achieve more precise estimation of the common kinetics. The model is defined through clearly specifying each layer of uncertainty and identifying generic versus trace-specific features. We give a brief review of the HMM in Section 2.2.1 and describe the Bayesian hierarchical HMM in Section 2.2.2.

2.2.1 Hidden Markov Model

Hidden Markov model (HMM) is a widely recognized model for single-molecule FRET data in the biochemical community and has been successfully adopted in various applications (McKinney et al., 2006; Liu et al., 2010; Blanco & Walter, 2010; Shen et al., 2012; Keller et al., 2014). It models a single FRET trace, which constitutes the building block of the more complicated Bayesian hierarchical HMM.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be observations from an HMM, i.e. a single FRET trace. For each y_i, there is a corresponding hidden state x_i, which takes values in $\{1, 2, \ldots, K\}$, where K is the number of hidden states. Two major components of an HMM are (1) a Markov model of the hidden states $\boldsymbol{x} = (x_1, \ldots, x_n)$ and (2) an observation model given the hidden states.

1. The hidden states, which follow a Markov chain, reprepsents the unobserved 'true' biological process. The Markov chain is defined by a $K \times K$ matrix, denoted by $\boldsymbol{P} = \boldsymbol{P}_{K\times K}$, that gives the transition probabilities among the K states. Given that the current hidden state is i, i.e. $x_t = i$, the probability of the next hidden state being j, i.e. $x_{t+1} = j$, is equal to P_{ij}, the ij^{th} element of the transition matrix P; $1 \leq i, j \leq K$, $1 \leq t < n$. The intrinsic kinetics of a molecule is described by the transition matrix since it reveals how likely a molecule is to move from one state (conformation) to another.

2. Each observation y_t, conditioning on $x_t = i$, follows a Gaussian distribution with mean μ_i and variance σ_i^2, denoted by

   $$y_t | x_t = i \sim \mathcal{N}(\mu_i, \sigma_i^2);$$

   $1 \leq i \leq K$, $1 \leq t \leq n$. This models the random measurement noise.

An HMM can be fitted with either the Expectation-Maximization (EM) algorithm, from which we obtain the maximum likelihood estimate (the best possible set of parameters based on an observed trace), or the Gibbs sampling algorithm, from which we obtain the full posterior distribution which provides quantified uncertainties, of all the model parameters. Detailed descriptions of the implementations of the algorithms are in the Appendix of Chen et al. (2016).

2.2.2 Bayesian Hierarchical Hidden Markov Model

In the Bayesian hierarchical HMM, each FRET trace is modeled as a hidden Markov model with its own FRET values and lengths. We denote the lengths of FRET traces from replicated experiments by $\{n_1, n_2, \ldots, n_T\}$ where T is the total number of FRET traces. For each trace l $(1 \leq l \leq T)$, the observed FRET values are denoted by $\boldsymbol{y}^{(l)} = (y_1^{(l)}, \ldots, y_{n_l}^{(l)})$, with corresponding hidden states denoted by $\boldsymbol{z}^{(l)} = (z_1^{(l)}, \ldots, z_{n_l}^{(l)})$; the mean and standard deviation for the k^{th} state is denoted by $\mu_k^{(l)}$ and $\sigma_k^{(l)}$, $1 \leq k \leq K$. The *common kinetics* and the *variable replicates* are the major assumptions that are made to the Bayesian hierarchical HMM.

Common kinetics. Since experimental replicates are designed to study the common kinetics of a biological process, we assume that the transition matrix $\boldsymbol{P}$ which depicts the inherent dynamics of a biological process is shared by all FRET traces.

Variable replicates. Since the experimental replicates in essence constitute an ensemble (possibly heterogeneous) with common characteristics, we assume that each FRET trace is a noisy realization of a true underlying process. This is formalized in a diagram shown in Figure 2.1, where the lines and arrows stands for dependencies.

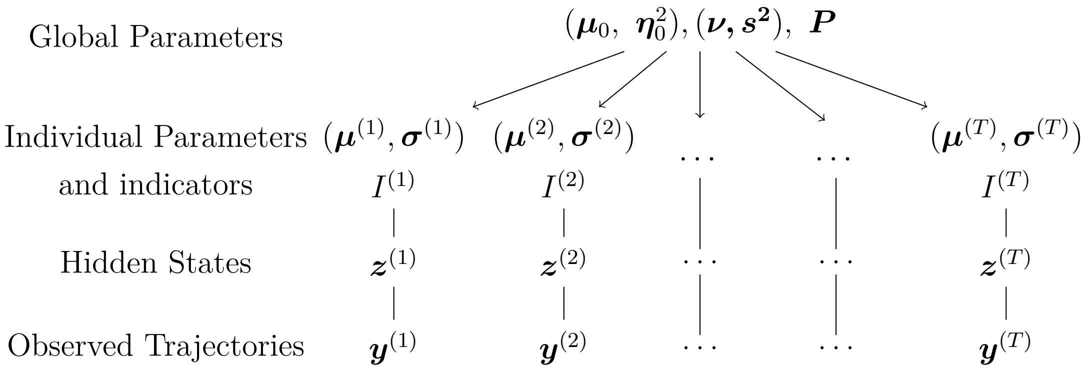

Figure 2.1: Diagram of the hierarchical HMM.

Each FRET trace, representing a single molecule, is doomed to behave slightly different from others due to the immanent stochastic property and some unpredictable and uncontrollable experimental conditions. This heterogeneity is reflected in the statistical modeling through disparities in the model parameters for different FRET traces, as shown in Figure 2.1. Mathematically, we assume that for each trace l and each state k, the mean FRET value $\mu_k^{(l)}$ comes from a Gaussian distribution with mean μ_{0k} and variance η_{0k}^2, denoted by $\mu_k^{(l)} \sim \mathcal{N}(\mu_{0k}, \eta_{0k}^2)$; and the corresponding variance $[\sigma_k^{(l)}]^2$ comes from an inverse-χ^2 distribution with ν_k degrees of freedom and scale s_k^2, denoted by $[\sigma_k^{(l)}]^2 \sim \mathrm{Inv} - \chi^2(\mathrm{\nu_k, s_k^2})$. Figure 2.2 demonstrates the data generating process of a hierarchical HMM with a simple example.

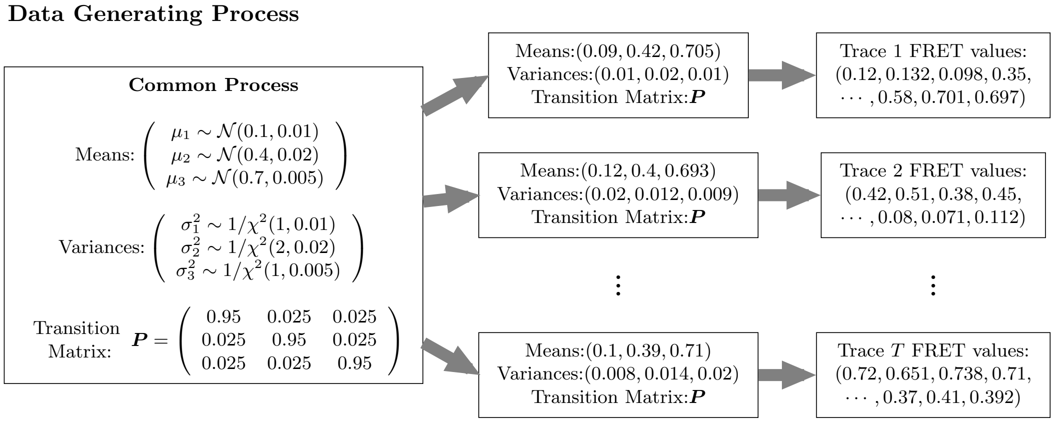
Data Generating Process
Common Process
Means: $\begin{pmatrix} \mu_1 \sim \mathcal{N}(0.1, 0.01) \\ \mu_2 \sim \mathcal{N}(0.4, 0.02) \\ \mu_3 \sim \mathcal{N}(0.7, 0.005) \end{pmatrix}$
Variances: $\begin{pmatrix} \sigma_1^2 \sim 1/\chi^2(1, 0.01) \\ \sigma_2^2 \sim 1/\chi^2(2, 0.02) \\ \sigma_3^2 \sim 1/\chi^2(1, 0.005) \end{pmatrix}$
Transition Matrix: $\boldsymbol{P} = \begin{pmatrix} 0.95 & 0.025 & 0.025 \\ 0.025 & 0.95 & 0.025 \\ 0.025 & 0.025 & 0.95 \end{pmatrix}$
Means:$(0.09, 0.42, 0.705)$
Variances:$(0.01, 0.02, 0.01)$
Transition Matrix:$\boldsymbol{P}$
Trace 1 FRET values: $(0.12, 0.132, 0.098, 0.35, \cdots, 0.58, 0.701, 0.697)$
Means:$(0.12, 0.4, 0.693)$
Variances:$(0.02, 0.012, 0.009)$
Transition Matrix:$\boldsymbol{P}$
Trace 2 FRET values: $(0.42, 0.51, 0.38, 0.45, \cdots, 0.08, 0.071, 0.112)$
Means:$(0.1, 0.39, 0.71)$
Variances:$(0.008, 0.014, 0.02)$
Transition Matrix:$\boldsymbol{P}$
Trace T FRET values: $(0.72, 0.651, 0.738, 0.71, \cdots, 0.37, 0.41, 0.392)$

Figure 2.2: An example of the data generating process of a hierarchical HMM.

Furthermore, since some FRET traces are quite short, within which some fast transitions and rare states might be missed, we incorporate indicators, $I^{(l)}$, indicating which states are present in trajectory l, in the model. For example, if the maximum number of states is $K = 3$, $I^{(l)}$ can take four values $\{1, 2, 3\}$, $\{1, 2\}$, $\{1, 3\}$ or $\{2, 3\}$, corresponding to the states present in trajectory l. We do not consider the singletons (such as {1}, {2} or {3}) in the set of possible states.

The fitting of the hierarchical HMM can be found in the Appendix of Chen et al. (2016). Figure 2.3 give a comparison of the traditional method of fitting each FRET trace separately with our method of combining multiple traces. Heuristically, by iteratively updating the common kinetics and the parameters of each trace, more precise and robust estimates are obtained for all the unknown parameters. Section 4.3 in Chen et al. (2016) elaborates the comparison of fitting the Bayesian hierarchical model versus fitting each trace separately using both simulation studies and real data results. It is apparent from the results that the Bayesian hierarchical model is more robust and gives more efficient estimates as opposed to fitting each trace separately, which can lead to significantly different estimates across different traces.

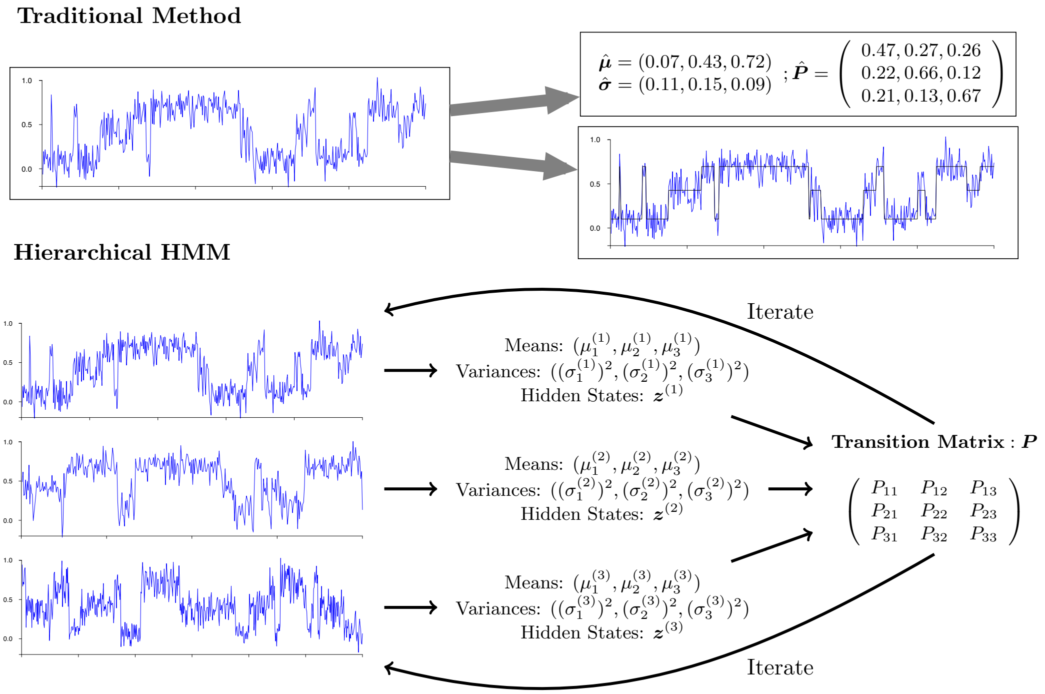
Traditional Method
$\hat{\boldsymbol{\mu}} = (0.07, 0.43, 0.72)$
$\hat{\boldsymbol{\sigma}} = (0.11, 0.15, 0.09)$
$; \hat{\boldsymbol{P}} = \begin{pmatrix} 0.47, 0.27, 0.26 \\ 0.22, 0.66, 0.12 \\ 0.21, 0.13, 0.67 \end{pmatrix}$
1.0
0.5
0.0
Hierarchical HMM
Iterate
Means: $(\mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)})$
Variances: $((\sigma_1^{(1)})^2, (\sigma_2^{(1)})^2, (\sigma_3^{(1)})^2)$
Hidden States: $\boldsymbol{z}^{(1)}$
Means: $(\mu_1^{(2)}, \mu_2^{(2)}, \mu_3^{(2)})$
Variances: $((\sigma_1^{(2)})^2, (\sigma_2^{(2)})^2, (\sigma_3^{(2)})^2)$
Hidden States: $\boldsymbol{z}^{(2)}$
Means: $(\mu_1^{(3)}, \mu_2^{(3)}, \mu_3^{(3)})$
Variances: $((\sigma_1^{(3)})^2, (\sigma_2^{(3)})^2, (\sigma_3^{(3)})^2)$
Hidden States: $\boldsymbol{z}^{(3)}$
Transition Matrix : $\boldsymbol{P}$
$\begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}$
Iterate

Figure 2.3: Comparison of traditional method (fit each smFRET trace separately, top panel) and our method (Bayesian hierarchical HMM, bottom panel) for fitting ensembles of smFRET traces.

2.3 Brief Description of Matlab Package

We have developed a publicly available Matlab package, ‘HHMM’, which fits the afore mentioned Bayesian hierarchical hidden Markov model based on multiple FRET traces. A detailed user manual is contained in the package. Users can run the ’demon.m’ file included in the package to test the functions with simulated data.

The input data format is a collection of ‘.txt’ files in a specific folder, ‘dataforanalysis’, under the current working directory; where each file is a column vector of FRET values, between 0 and 1. The estimated parameters and the corresponding standard errors will be returned as a Matlab structure; and visualizations of fitted FRET traces are generated. Here is a concrete example for usage of the package and the outputs.

1. Download and install the ‘HHMM’ package.

2. Change the current working directory to the one that contains the ‘.m’ files.

3. Put all the data (‘1.txt’, ‘2.txt’, etc.) in the folder ‘dataforanalysis’.

4. Set up the package using the following command:

   $$>> \text{setup\_package}();$$

5. Fit a 3-state heterogeneous hierarchical HMM with the following command:

   $$>> \text{resultheter} = \text{HHMMfit}('\text{K}', 3, '\text{HeterTraces}', \text{true});$$

   The user can choose to set other tuning parameters, too. Please refer to the user manual for detailed instructions.

6. The return value ‘resultheter’ is a Matlab structure with fields *EstimateWithUncertaintyGlobalPar* and *EstimateWithUncertaintyIndTracePar*, which are both Matlab structures that contains summary statistics of the estimated global parameters and parameters for individual traces respectively.

    - EstimateWithUncertaintyGlobalPar – a Matlab structure with fields ‘postmean’, ‘postsd’, ‘postmodeindicator’. ‘postmean’ (estimated values) and ‘postsd’ (uncertainty/standard deviations of the estimated values) are Matlab structures with fields ‘globalmean’, ‘globalvar’, ‘P’. ‘postmodeindicator’ is a matrix, each row (representing one trace) is a vector of length 3: the k^{th} element of which is equal to 1 if the k^{th} state exist in the current trace and 0 otherwise; $1 \leq k \leq 3$.
    - EstimateWithUncertaintyIndTracePar – a Matlab structure with fields ‘postmean’, ‘postsd’, ‘hiddenstates’. ‘postmean’ (estimated values) and ‘postsd’ (uncertainty/standard deviations of the estimated values) are Matlab structures with fields ‘indmean’ (a matrix, each row is estimated mean values for one trace) and ‘indvar’ (a matrix, each row is estimated variance values for one trace). The field ‘hiddenstates’ is a cell array, each element is a vector of fitted discrete hidden states for each trace.

7. The ‘results’ folder is created under the current working directory, in which visualizations of the fitting for each trace is saved as ‘.png’ and ‘.pdf’ files. Besides, a ‘.txt’ file is generated in the ’results’ folder, summarizing the fitted results. Here is an example outputfile.

    ```
    Global parameters (with estimated standard deviation in parenthesis)
    means = 0.023518 (0.020989) 0.500781 (0.030332) 0.936386 (0.022558)
    ```

variances = 0.008564 (0.003453) 0.017869 (0.007036) 0.010294 (0.004022)

transition matrix is:

0.335916 (0.008882) 0.334422 (0.009856) 0.329662 (0.009131)

0.325266 (0.009096) 0.346403 (0.010000) 0.328331 (0.009302)

0.342556 (0.009179) 0.341375 (0.009900) 0.316068 (0.008926)

Means and variances for each trace (with estimated standard deviation in parenthesis):

FILENAME (column 1) Means (column 2-4) Variances (column 5-7)

1.txt -0.008 (0.009) 0.547 (0.010) 0.942 (0.010) 0.010 (0.001) 0.012 (0.002) 0.011 (0.002)

2.txt -0.018 (0.009) 0.462 (0.011) 0.752 (0.010) 0.011 (0.001) 0.009 (0.002) 0.010 (0.001)

......

Figure 2.4 shows two of the fitted traces in the 'results' folder.

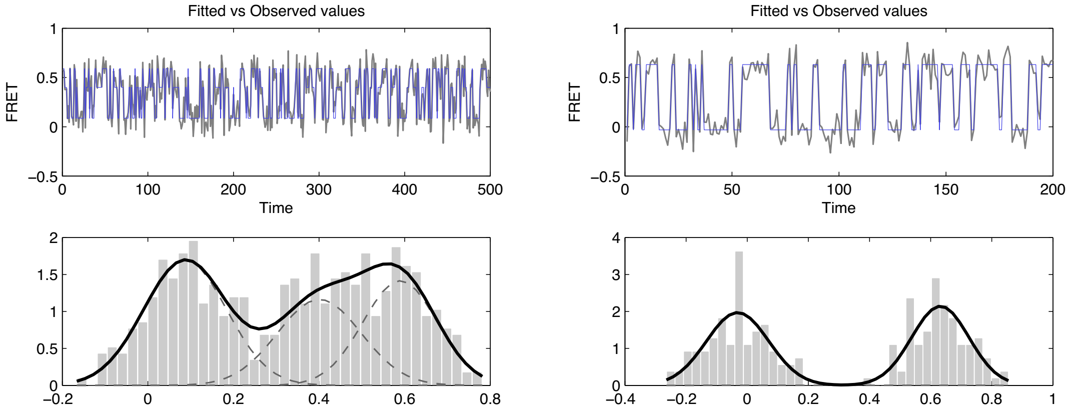

Figure 2.4: Example of output visualizations for two traces fitted from a heterogeneous HMM using the 'HHMM' Matlab package. On the top panel, the observed FRET values are plotted with gray line and the fitted values with blue line; on the bottom panel, the histogram is for all the FRET values, the dashed gray lines are the mixture components whereas the solid black curve is the fitted mixture density. The trace on the left has three hidden states but the right only has two.

2.4 Robustness Test of Model Fitting

We illustrate the robustness of model fitting using the 'HHMM' package. First, we study the precision of the estimated transition matrix as compared to the truth using traces with varying lengths in Section 2.4.1 and varying signal-to-noise ratios in Section 2.4.2. Next, we simulate from heterogeneous hierarchical HMMs and study the model-selection property of our method in Section 2.4.3. Last, we test the method for estimation of rare events using strongly diagonal transition matrices in Section 2.4.4.

2.4.1 Varying Trace Lengths and Number of Traces

We vary the number of traces and the numbers of observations in each trace to examine their influence on the estimation precision of the common dynamics.

We repeatedly simulate 100 times from a hierarchical HMM with $T = 20$ (or $T = 50$, $T = 100$) traces and the length of each trace is randomly simulated from the interval $[200, 400]$ (or $[400, 600]$, $[600, 800], \ldots, [1200, 1400]$). We set the global means $\boldsymbol{\mu}_0 = (0.1, 0.5, 0.9)$, the global variance $\boldsymbol{\eta}_0^2 = (0.1^2, 0.1^2, 0.1^2)$, and the transition matrix $\boldsymbol{P}$ with diagonal elements 0.6 and off-diagonal elements 0.2. The standard deviation of all the states in all traces are set to be equal, i.e. $\sigma_k^{(l)} = 0.1$ for all $1 \leq k \leq 3$, $1 \leq l \leq 20$. For each simulation, we calculate the distance of the estimated transition matrix $\{\hat{P}_{ij}\}_{1\leq i,j\leq 3}$ with the truth $\{P_{ij}\}_{1\leq i,j\leq 3}$, defined as $\sum_{i=1}^{3}\sum_{j=1}^{3}(\hat{P}_{ij} - P_{ij})^2$, which stands for the precision of the estimation. Figure 2.5 compares the boxplots of this distance over repeated simulations when $T = 20, 50, 100$, with red, gray and blue boxes respectively; the x-axes labels the ranges of the lengths for each trace.

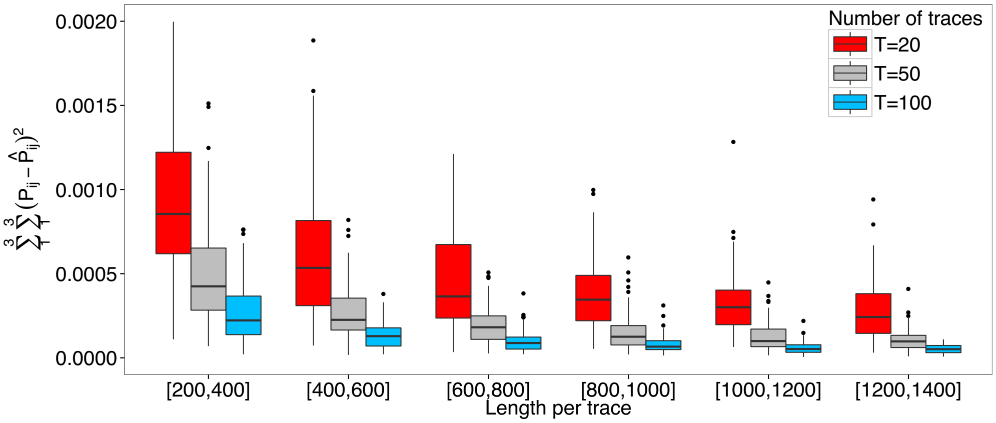

Figure 2.5: Boxplots of the squared error of the transition matrix $\sum_{i,j}(P_{ij} - \hat{P}_{ij})^2$ when the number of traces is equal to 20 (red), 50 (gray), 100 (blue). Each box represents 100 simulations, with a total of 6×3 experiments/boxes; the length of each trace in each experiment is randomly simulated from the intervals labeled on the x-axis.

The results show that (1) with the same number of traces, as we increase the number of observations per trace, the precision of estimating the transition matrix increases; (2) when we fix the range of the number of observations per trace, the larger the number of traces, the better precision and the more stable the results. The practical implication is that when the observational time is limited for each trace, collecting more traces indeed helps improve the precision of the inference significantly.

2.4.2 Signal-to-Noise Ratio

Similar to the fitting of other statistical models, the higher the signal-to-noise ratio, the more precise estimators we can obtain. We also confirm this through repeated simulation studies, the details are omitted here. One thing that we need to emphasize is that the hierarchical hidden Markov model is able to provide reliable estimates of model parameters under much lower signal-to-noise ratios as opposed to fitting individual traces separately using hidden Markov models. See Section 4.3 in Chen et al. (2016) for a detailed comparison of fitting hierarchical versus individual hidden Markov models when the signal-to-noise ratio varies.

2.4.3 Heterogeneous Traces and Model Selection

The hierarchical HMM provides an automatic allocation of the number of hidden states through the indicators defined for each trace, see Section 2.2.2. We perform repeated simulations (100 times) to study the correct classification rate of the number of hidden states for each trace. Each time we simulate T_3 (30 or 100) 3-state traces with N_3 (200, 500, 1000 or 2000) observations per trace, the means of the 3 states are simulated independently from Gaussian distributions

$$(\mathcal{N}(0.1, 0.1^2), \mathcal{N}(0.4, 0.1^2), \mathcal{N}(0.7, 0.1^2));$$

two 2-state traces of N_2 observations with means from $(\mathcal{N}(0.1, 0.1^2), \mathcal{N}(0.4, 0.1^2))$; two 2-state traces of N_2 observations with means from $(\mathcal{N}(0.1, 0.1^2), \mathcal{N}(0.7, 0.1^2))$; and two 2-state traces of N_2 observations with means from $(\mathcal{N}(0.4, 0.1^2), \mathcal{N}(0.7, 0.1^2))$; N_2 can be 200, 500, 1000 or 2000. The transition matrix (3 by 3) is set to be 0.6 on the diagonal and 0.2 on off-diagonal elements. Table 2.1 gives the rates of correctly identifying the number of hidden states for each trace, corresponding to different combinations of N_3, N_2, T_3 values. R_3 is the probability of correctly identifying 3-state traces and R_2 is the probability of identifying 2-state traces. The last two columns are in the format of “mean ± standard deviation”.

T_3	N_3	T_2	N_2	R_3	R_2
30	2000	6	1000	1.0000 ± 0.0000	0.9600 ± 0.0374
30	1000	6	2000	0.9993 ± 0.0025	0.8017 ± 0.1481
30	500	6	200	0.9940 ± 0.0067	0.8383 ± 0.0504
30	200	6	500	0.9837 ± 0.0107	0.5317 ± 0.1863
100	2000	6	1000	0.9995 ± 0.0022	0.9850 ± 0.0207
100	1000	6	2000	0.9986 ± 0.0035	0.9600 ± 0.0469
100	500	6	200	0.9950 ± 0.0075	0.8383 ± 0.0306
100	200	6	500	0.9914 ± 0.0097	0.8850 ± 0.0698

Table 2.1: The rates (mean ± standard deviation) of correctly identifying the number of hidden states for each trace according to 100 repeated simulations of mixed population, where T_3 and T_2 are the numbers of 3-state and 2-state traces, N_3 and N_2 are the number of observations in each of the 3-state and 2-state traces, and R_3 and R_2, in the format of 'mean ± standard deviation', are the correct classification probabilities for the 3-state and 2-state traces.

Table 2.1 shows that the correct classification probability of the 2-state traces, which are 'minorities', are more volatile than that of the 3-state traces, which are the 'majority'. Based on empirical studies, the 'HHMM' package can identify correctly the number of hidden states for majority of the traces.

2.4.4 Rare Events Estimation

In cases when certain states do not appear a lot, each single FRET trace does not provide enough information for understanding the true kinetics of the biological process. The hierarchical HMM, by pooling information from multiple traces, is more capable of capturing these 'rare events' and giving efficient estimations of the underlying dynamics. We study the performance of the proposed hierarchical HMM in estimating the probabilities of the rare events, i.e. transitions with small probabilities, through a series of repeated simulations.

The number of traces in each simulation is 20 and the number of observations for each trace is 1000. The global means and variances are set to be $\boldsymbol{\mu}_0 = (0.1, 0.4, 0.7)$, $\boldsymbol{\sigma}_0^2 = (0.1^2, 0.1^2, 0.1^2)$ and the individual trace variances are all set to be $\left[\sigma_k^{(l)}\right]^2 = 0.01$. Since in real applications, the transition matrix appears to be highly diagonal, i.e. cross-state transitions are rare, thus we consider a 3 × 3 transition matrix with diagonal elements $(1 - 2P_{12})$ and off-diagonal elements P_{12}, which takes values in

$$\{0.15, 0.125, 0.1, 0.075, 0.05, 0.025, 0.01, 0.005, 0.0025, 0.001\}.$$

Figure 2.6 shows boxplots of the relative estimation error $|\hat{P}_{12} - P_{12}|/P_{12}$ versus the true P_{12} values, where each box represents the results from 100 repeated simulations.

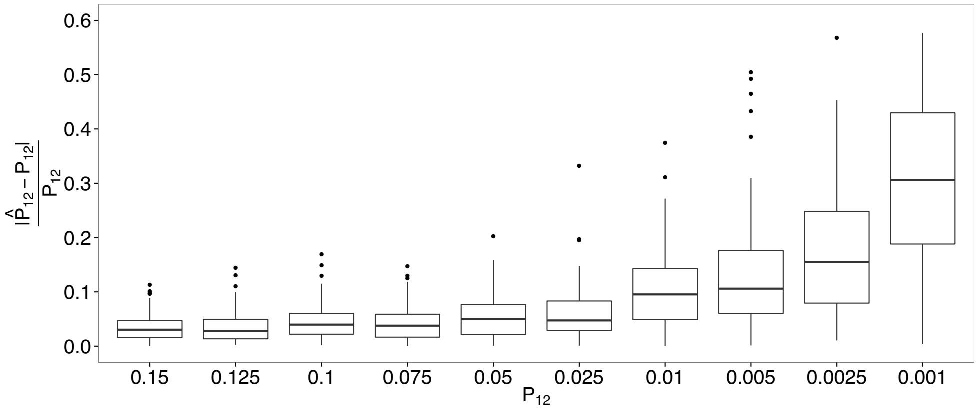

Figure 2.6: Relative absolute error of estimating P_{12}, the true values are labeled on the x-axis. Each box represents 100 repeated simulations in which $T = 20$ and $N_l = 1000$ for $1 \leq l \leq T$.

The results show that the relative estimation error of the very small transition probabilities turns out to be less than 10% as long as the transition probability does not fall below 0.025. However, this number changes if the number of observations or the number of traces changes. In principle, the hierarchical HMM captures rare transition probabilities effectively with reasonably large observations.

2.5 Summary

In this paper, we elaborated the Bayesian hierarchical HMM and introduced a Matlab package implementing the model fitting for scientists. The proposed method is robust to varying lengths, various signal-to-noise ratio and is proven powerful in detecting rare events and providing an automatic model selection for heterogeneous traces.

Furthermore, the Matlab package 'HHMM' also implements the fittings of individual HMMs with the EM algorithm, which gives the maximum likelihood estimators of the model parameters, and the Gibbs sampling algorithm, which gives the posterior distribution of the model parameters. A recommended number of hidden states for the Bayesian hierarchical HMM is given, too, which is calculated based on the majority rule of a widely adopted model selection method for single-trace HMMs, the Bayesian information criterion (BIC), similar to that applied in Chen et al. (2016).

There is nothing so practical as a good theory.

Ludwig Boltzman

The nature of reality is this: It is hidden, and it is hidden, and it is hidden.

Rumi, 13th-century Sufi poet

3 Order Selection of Hidden Markov Models

3.1 Introduction

It has been well recognized that hidden Markov models (HMM) and general state space models provide useful frameworks for describing noisy observations from an underlying stochastic process. They are popular for processing time series data and widely used in fields like speech recognition, signal processing and computational molecular biology.

The fundamental components of a hidden Markov model include the observations $\{Y_i = y_i; 1 \leq i \leq n\}$ and the corresponding hidden states $\{X_i = x_i, 1 \leq i \leq n\}$, which is a Markov chain. Throughout the paper, we use upper cases $\{Y, X\}$ to denote the random variables and the corresponding lower cases $\{y, x\}$ to denote the realizations (observations). In this paper, we consider discrete state space hidden Markov models, i.e., the hidden states have a finite support, observed at discrete time points $\{t_1, \ldots, t_n\}$, or $\{1, \ldots, n\}$ for notational simplicity. The size of the support of hidden states, denoted by K, is the number of hidden states of an HMM. In most real-world problems, the number of hidden states is not known beforehand but conveys important information of the underlying process. For example, in molecular biology, K could be the number of distinct conformations of a protein; in chemistry, K could be the number of distinct chemical species in a biochemical reaction. Existing methods to estimate K either suffer from lack of theoretical guarantee or unfeasible/impractical implementation, which we review in details in Section 3.1.1. The goal for this paper is to provide a consistent method, *the marginal likelihood method*, to determine the number of hidden states K based on the observations $\{y_1, \ldots, y_n\}$ of an HMM, which is computationally feasible for practitioners with minimal tuning.

3.1.1 Brief Literature Review

It has been recognized that the model parameters of an HMM are not identifiable when the number of hidden states is over-estimated (Chapter 22 of Hamilton, 1994; Ferguson, 1980; Rydén et al., 1998). Thus, determining the number of hidden states, also called the order selection in the machine learning literature, is an important problem for conducting valid inferences on model parameters of hidden Markov models. There is a vast literature on the model selection for hidden Markov models. We briefly review some of the most widely adopted methods here.

A special case of HMMs is finite mixture models, where all entries in the transition matrix are equal. The model selection of finite mixture models are mostly based on penalized likelihood, also known as information-theoretic approaches (Chen & Kalbfleisch, 1996; Lo et al., 2001; Jeffries, 2003; Chen et al., 2008; Chen & Tan, 2009; Chen & Li, 2009; Chen & Khalili, 2012; Huang et al., 2013; Rousseau & Mengersen, 2011; Hui et al., 2015).

When the observations $\{y_1, \ldots, y_n\}$ are supported on a finite set (i.e., when they are discrete-valued) , we call it a finite-alphabet hidden Markov process (MacDonald & Zucchini, 1997). Information-theoretic approaches for the order estimation of a finite-alphabet hidden Markov processes are widely adopted. Finesso (1990) proposes a penalized likelihood estimator, which is proved to be strongly consistent for finite-alphabet HMMs under certain regularity conditions. Ziv & Merhav (1992) derives the estimator by minimizing the under-estimation probability, which is shown to be not consistent (Kieffer, 1993; Liu & Narayan, 1994). Liu & Narayan (1994) gives a modified version which is shown to be consistent given an upper bound of the order of a finite-alphabet HMMs. Kieffer (1993) gives a strongly consistent estimator that resembles the Bayesian information criterion (BIC) in Schwarz (1978) for finite-alphabet HMMs. Gassiat & Boucheron (2003) proves strong consistency of these penalized maximum likelihood estimations without assuming any upper bound on the order for finite-alphabet HMMs, with smaller penalties than previous works. See Rydén (1995); Ephraim & Merhav (2002) for more detailed discussions about the literature on order selection of finite-alphabet HMMs.

However, when the observations $\{y_1, \ldots, y_n\}$ are supported on the real line, as in the Gaussian HMM, where each observation follows a Gaussian distribution conditioning on its hidden state, the problem becomes more difficult. The major difficulty comes from the fact that the overly-fitted mixture models are not identifiable and that the likelihood ratio statistics becomes unbounded, see Gassiat & Rousseau (2014).

The majority of the methodologies proposed in the literature rely on the idea of penalized likelihood, the consistency of which remains to be satisfactorily solved. These methods generally resemble the Akaike information criterion (AIC, Akaike (1974)), minimum description length (MDL, Rissanen (1978); Barron et al. (1998); Chambaz et al. (2009)) or the BIC (Schwarz, 1978). Hung et al. (2013) gives a consistent estimator of the number of hidden states using double penalizations when assuming that the maximum likelihood estimators are consistent. Rydén (1995) introduces an estimator that does not asymptotically under-estimate the order, given an upper bound for the order. General consistency of order estimation of mixture models using penalized likelihood methods is proved in Leroux (1992a), whose regularity conditions, however, are not satisfied for heterogeneous (unequal variances) Gaussian HMMs. Applications of the AIC and BIC to Gaussian mixture and hidden Markov models are given in Leroux & Puterman (1992). Rydén et al. (1998) applies the bootstrap technique to perform likelihood ratio test for the order estimation of hidden Markov models for a real example. Gassiat & Keribin (2000) investigates the likelihood ratio test for testing a single population i.i.d. model against a mixture of two populations with Markov regime. MacKAY (2002) estimates the order and the parameters together by minimizing a penalized distance function of the empirical distribution with all finite mixture models. Information theoretic approaches makes it possible to add heavier penalties as opposed to that of the BIC (Gassiat, 2002; Gassiat & Boucheron, 2003; Chambaz et al., 2009).

Bayesian methods, which does not depend on the maximum likelihood estimator, also plays an important role in the HMM model selection literature. Reversible jump methods (Green & Hastie, 2009; Fan et al., 2011) have been successfully adopted in practice (Green & Richardson, 2002; Boys & Henderson, 2004; Robert et al., 2000; Spezia, 2010), with a lack of theoretical justification. Gassiat & Rousseau (2014) provides a frequentist asymptotic evaluation of Bayesian analysis methods, purely from a theoretical perspective: under certain conditions on the prior, the posterior concentration rates and a consistent Bayesian estimation of the number of hidden states are given; practical implementation, guidance of tuning of the algorithm and numerical results are not provided therein.

Some authors have studied approaches that are related to our marginal likelihood method. Chambaz & Rousseau (2005) uses marginal likelihood ratio for the order estimation of mixture models, and obtained similar results for the marginal likelihood ratio: $O(e^{-cn})$ for underestimation, and $O(n^{-1/2+\delta})$ for overestimation. Wang & Bickel (2015) adapted the penalty approach to stochastic block models, with a similar "path-ignorance" result, i.e. a set of irregular paths can be asymptotically ignored, as in our Theorem 2; see Lemma 2.6 in Wang & Bickel (2015). Though the afore mentioned studies share similarity with the results in this paper, it is worth noting that in these models, the hidden state variables are assumed to be independent and identically distributed (i.i.d.), which is not true for HMMs. This additional complexity requires us to provide alternative techniques to manipulate the likelihood function of an HMM as shown in Theorem 2.

3.1.2 Recap of HMM and Notations

Consider the following hidden Markov model (HMM): let $\mathbf{X} = \{X_i, i \geq 0\}$ be an ergodic (positive recurrent, irreducible and aperiodic) Markov chain on a finite state space $\mathcal{X}_K = \{1, \cdots, K\}$ with transition matrix $Q_K = \{q_{kl}, 1 \leq k, l \leq K\} \in \mathcal{Q}_K$, i.e., $q_{kl} = P(X_{i+1} = l | X_i = k)$ for all $i \geq 0$. Conditioning on $\mathbf{X}$, $\mathbf{Y} = \{Y_i, i \geq 1\}$ are independent random variables on $\mathcal{Y}$, and the distribution of Y_i given $X_i = k$ is $f(\cdot|\boldsymbol{\theta}_k)$ for $i \geq 1$ and $k \in \mathcal{X}_K$, where $\boldsymbol{\theta}_k \in \Theta$. We assume that f is distinguishable on Θ, i.e., the measure of the set $\{y : f(y|\boldsymbol{\theta}_k) \neq f(y|\boldsymbol{\theta}_l)\}$ is greater than 0 for all $1 \leq k < l \leq K$. We denote the model parameters by $\boldsymbol{\phi}_K = (Q_K; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \in \mathcal{Q}_K \times \Theta^K = \Phi_K$.

Suppose $X_0 = x_0$ is known and we observe $\mathbf{y}_{1:n} = \{y_1, y_2, \cdots, y_n\} \in \mathcal{Y}^n$, but the underlying process $\mathbf{x}_{1:n} = \{x_1, x_2, \cdots, x_n\}$ remains hidden (unobserved). The joint likelihood of $(\mathbf{y}_{1:n}, \mathbf{x}_{1:n})$ given the parameters $\boldsymbol{\phi}_K$ is

$$\begin{aligned} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \boldsymbol{\phi}_K) &= p(\mathbf{y}_{1:n} | \mathbf{x}_{0:n}; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) p(\mathbf{x}_{0:n} | Q_K) \\ &= \prod_{k=1}^{K} \left\{ \prod_{i:x_i=k} f(y_i|\boldsymbol{\theta}_k) \right\} \times \left\{ \prod_{i=1}^{n} q_{x_{i-1}x_i} \right\}. \end{aligned} \tag{3.1}$$

The likelihood after integrating out the hidden states is

$$p(\mathbf{y}_{1:n} | \boldsymbol{\phi}_K) = \sum_{\mathbf{x}_{1:n} \in \mathcal{X}_K^n} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \boldsymbol{\phi}_K), \tag{3.2}$$

where $\mathcal{X}_K^n$ denotes the product space of n copies of $\mathcal{X}_K$.

The maximum likelihood estimators (MLE) of a hidden Markov model given K, the number of hidden states, can be obtained through the Baum-Welch/ Expectation-Maximization (EM) algorithm (Baum & Petrie, 1966; Baum et al., 1970; Dempster et al., 1977). The consistency of the maximum likelihood estimator of HMMs are established in Leroux (1992b); Bickel et al. (1998), given the correct K, under certain regularity conditions.

3.1.3 Gaussian Hidden Markov Models

In this section, we reveal some difficulties of the order selection of HMMs using a concrete example that is widely adopted in applications, the heterogeneous Gaussian HMM.

In a heterogeneous Gaussian HMM, given $X_i = k$, y_i follows a Gaussian distribution with mean μ_k and variance σ_k^2. Thus $\boldsymbol{\phi}_K = (Q_K; \{\mu_k, \sigma_k^2\}_{1\leq k\leq K})$ and the joint likelihood is

$$p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n}|\boldsymbol{\phi}_K) \propto \prod_{k=1}^{K} \left\{ \prod_{i:x_i=k} \frac{1}{\sigma_k} \exp\left(-\frac{(y_i-\mu_k)^2}{2\sigma_k^2} \right) \right\} \times \left\{ \prod_{i=1}^{n} q_{x_{i-1}x_i} \right\}.$$

Note that this likelihood is unbounded: if one takes $\mu_{k^*} = y_{i_0}$ for some i_0 and k^*, then as $\sigma_{k^*} \to 0$, $p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n}|\boldsymbol{\phi}_K) \to \infty$. This can be a serious issue when one overfits an HMM – the extra component could concentrate on only one single observation with zero variance, which blows up the likelihood. Therefore, methods of model selection for Gaussian HMM based on penalized likelihoods, which requires the consistency of the maximum likelihood estimator, become problematic. General consistency results of model selection based on penalized likelihoods have to exclude this case in the required regularity conditions (Leroux, 1992a). Therefore, the BIC, though widely adopted in practice, is theoretically questionable for its validity as a model selection criterion for HMM (Gassiat & Rousseau, 2014; MacDonald & Zucchini, 1997). This is the same issue as the unbounded likelihood for heterogeneous Gaussian mixture models (Chen & Khalili, 2012). In fact, Gaussian mixture models can be obtained by setting $q_{ij} \equiv 1/K$ for all $i, j \in \{1, 2, \ldots, K\}$, thus is a special case of HMMs.

Furthermore, as noted in Gassiat & Rousseau (2014), for overly fitted HMMs, or other finite mixture models, the model parameters become non-identifiable. In an overly fitted HMM, the neighborhood of the true transition matrix contains transition matrices arbitrarily close to non-ergodic transition matrices. Adding hard thresholds to entries in the transition matrix does not satisfactorily solve the problem.

3.1.4 Outline

The remainder of the paper has four sections. In sections 3.2 and 3.3, we propose and prove the consistency of the marginal likelihood method for general HMM order selection, including the heterogeneous Gaussian HMM. The difficulties mentioned in section 3.1.3 are addressed by introducing the concept of *asymptotic path ignorance* – neglecting the irregular hidden state trajectories that blow up the likelihood or making the parameters non-identifiable. In section 3.4, we describe the computational method, demonstrate the effectiveness of the marginal likelihood method using numerical experiments, and conclude with discussions on the theoretical consistency of our practical implementation. Section 3.5 summarizes the paper.

3.2 Model Selection via Marginal Likelihood

As discussed in Section 3.1.1, the existing model selection methods for HMM either has no theoretical guarantee or is theoretically justified only for a very restricted family of HMMs, excluding the popular heterogeneous Gaussian HMM. We propose to approach the problem of HMM model selection via the marginal likelihood, i.e., directly comparing the probability of obtaining the observations under HMMs with different number of hidden states, after integrating out both the model parameters and the hidden states. This method is consistent under weak regularity conditions that are trivially satisfied by a wide range of HMMs, including the heterogeneous Gaussian HMM.

3.2.1 Marginal Likelihood Method

Given the number of states K, we assume that each $\boldsymbol{\theta}_k$ is drawn independently from the prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\alpha})$ and that Q_K is drawn from the prior distribution $\nu_K(Q_K|\boldsymbol{\beta}_K)$, independent of the $\boldsymbol{\theta}_k$; $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_K$ are the hyper-parameters, which are assumed to be fixed constants. Denote $p_0(\boldsymbol{\phi}_K)$ the joint prior, which is expressed as

$$p_0(\boldsymbol{\phi}_K) = p_0(\boldsymbol{\phi}_K|\alpha, \beta_K) = \nu_K(Q_K|\boldsymbol{\beta}_K) \prod_{k=1}^{K} \pi(\boldsymbol{\theta}_k|\boldsymbol{\alpha}).$$

The marginal likelihood under a K-state HMM is defined as

$$p_K(\mathbf{y}_{1:n}) = \int_{\Phi_K} p(\mathbf{y}_{1:n}|\boldsymbol{\phi}_K) p_0(\boldsymbol{\phi}_K) d\boldsymbol{\phi}_K. \tag{3.3}$$

We then choose the K that maximize the marginal likelihood:

$$\hat{K}_n := \text{arg max}_K p_K(\mathbf{y}_{1:n}).$$

3.2.2 Discussions of Marginal Likelihood Methods

The marginal likelihood has been used in the model selection literature. Ratio of marginal likelihoods is known as the Bayes factor (Kass & Raftery, 1995), a popular model selection criterion. The BIC is in fact an approximation of the marginal likelihood using the Laplace method. Ghahramani (2001) discussed the practical applicability and calculation of the Bayes factor. Bauwens et al. (2014) applies the marginal likelihood method for model selection of Markov-switching GARCH and change-point GARCH models. Du et al. (2016) uses the marginal likelihood method to determine the number and locations of change-points of a stepwise signal.

As discussed in section 3.1.3, the heterogeneous Gaussian HMM suffers from the problem of having an unbounded likelihood surface. Adding a conjugate prior for the variance parameters in Gaussian HMMs can fix the issue of unbounded likelihood surface. Therefore, the proposed marginal likelihood method, which integrates out the parameters and hidden states, does not suffer from irregularity of the likelihood surface.

3.3 Theoretical Study of the Marginal Likelihood Estimator

We show theoretically the consistency of the proposed marginal likelihood estimator for HMM order selection, including the rate of convergence of the marginal likelihoods, in section 3.3.1. The asymptotic properties of the marginal likelihoods crucially depends on an *asymptotic path ignorance* result which is detailed and demonstrated with a simple example in section 3.3.2. We describe how the difficulties of overly fitted HMMs are overcome in section 3.3.3. Section 3.3.4 points out the connections of the order selection of HMMs with the model selection of finite mixture models.

Throughout the remainder of the paper, we use o_P to denote convergence in probability and O_P to denote stochastic boundedness. For any two sequences of random variables $\{X_n, Y_n\}_{n\geq 1}$, we write $X_n \backsim Y_n$ if and only if $X_n/Y_n = O_P(1)$ and $Y_n/X_n = O_P(1)$ for all $n \geq 1$. We use $\#\mathcal{S}$ to denote the cardinality of a finite set $\mathcal{S}$. For a decreasing sequence $\{\epsilon_n\}_{n>0}$ that converges to 0 as $n \to \infty$, we denote it by $\epsilon_n \downarrow 0$.

3.3.1 Consistency and Rate of Convergence

Let K^* be the true number of states and $\boldsymbol{\phi}^* = (Q^*; \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{K^*}^*)$ be the true parameters. Theorem 1 gives the consistency and rate of convergence of the marginal likelihood method for HMM order selection, the proof is given in Appendix B.1.3.

Theorem 1. *Assume that regularity conditions 1)-5) are satisfied. Then for any $K \neq K^*$, as $n \to \infty$,*

$$\frac{p_K(\mathbf{y}_{1:n})}{p_{K^*}(\mathbf{y}_{1:n})} = o_P(n^{-1/2}\log n).$$

Furthermore, if K^* is bounded from above, i.e. there exists a finite positive constant $\overline{K} \geq K^*$, then as $n \to \infty$,

$$\hat{K}_n := arg\ max_{1\leq K\leq \overline{K}}\ p_K(\mathbf{y}_{1:n}) \xrightarrow{P} K^*,$$

where $\xrightarrow{P}$ denotes convergence in probability. The regularity conditions are

1) *The prior density $\pi(\cdot|\alpha)$ is continuous and positive at $\boldsymbol{\theta}_k^*(1 \leqslant k \leqslant K^*)$.*

2) *There exists $\delta > 0$ such that $N_k(\delta) \bigcap N_l(\delta) = \emptyset$ for all $1 \leqslant k < l \leqslant K^*$, where $N_i(\delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_i^*\| < \delta\}$.*

3) *The Markov chain governed by the true transition matrix Q^* is irreducible, aperiodic and positive recurrent.*

4) *For all K, $\nu_K(\cdot|\beta_K)$ is positive and continuous in $\mathcal{Q}_K$.*

5) Conditions (A1)-(A6), (B1)-(B2) and (C1)-(C2) in Appendix B.1.1.

Condition 1) ensures that the prior distribution is well-behaved around the true parameters of each state. Condition 2) ensures that the parameters from different states are distinguishable. Condition 3) regularizes the dynamics of the Markov chain through the *true* transition matrix. Condition 4) ensures that the prior ν_K well behaves in $\mathcal{Q}_K$. Condition 5) is used to govern the asymptotic behavior of posterior distributions, especially for the over fitting case; see Appendix B.1.1 for detailed explanations.

The consistency of HMM order selection in Theorem 1 overcomes the difficulties of the HMM order selection discussed in the introduction through the *asymptotic path ignorance* result, which provides important insights on the origin of the problem on overly-fitted HMMs.

3.3.2 Asymptotic Path Ignorance

One of the main difficulties in dealing with the marginal likelihood of HMM given in (3.3) is that all possible paths $\mathbf{x}_{1:n} \in \mathcal{X}_K^n$ are summed up, which consist of many paths where the number of visits to some state does not go to infinity as the number of observations goes to infinity, causing the asymptotic analysis to fail. The existence of a path that contains only one single visit for some state is also the main reason of unbounded likelihood, which invalidates the BIC or other penalized likelihood methods. To resolve this issue, we introduce Theorem 2, which allows us to asymptotically neglect these undesirable paths.

For any $\mathbf{x}_{1:n}$, define n_k to be the number of visits to state k:

$$n_k = n_k(\mathbf{x}_{1:n}) := \#\{i : 1 \leq i \leq n, x_i = k\} \tag{3.4}$$

for all $1 \leq k \leq K$. For any $\epsilon > 0$ and $1 \leq k \leq K$, define

$$\mathcal{X}^n_{K,k,\epsilon} := \{\mathbf{x}_{1:n} : \mathbf{x}_{1:n} \in \mathcal{X}^n_K,\ n_k(\mathbf{x}_{1:n}) < \epsilon n\}, \tag{3.5}$$

and let $\mathcal{X}^n_{K,\epsilon}$ be the union of the $\mathcal{X}^n_{K,k,\epsilon}$, i.e.

$$\mathcal{X}^n_{K,\epsilon} := \bigcup_{k=1}^{K} \mathcal{X}^n_{K,k,\epsilon}. \tag{3.6}$$

Theorem 2. *Under the conditions (A1)-(A6) and (C1)-(C2) in Appendix B.1.1, we have, for all $K \geq 1$, any sequence of sets $\mathcal{A}_n \subset \Phi_K$, and any sequence of $\epsilon_n \downarrow 0$, with probability one, as $n \to \infty$,*

$$\frac{\sum_{\mathcal{X}^n_K \setminus \mathcal{X}^n_{K,\epsilon_n}} \int_{\mathcal{A}_n} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \phi_K) p_0(\phi_K) d\phi_K}{\sum_{\mathcal{X}^n_K} \int_{\mathcal{A}_n} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \phi_K) p_0(\phi_K) d\phi_K} \to 1.$$

In particular, if we set $\mathcal{A}_n = \Phi_K$ for all n, with probability one,

$$\frac{\sum_{\mathcal{X}^n_K \setminus \mathcal{X}^n_{K,\epsilon_n}} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \alpha, \beta_K)}{\sum_{\mathcal{X}^n_K} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \alpha, \beta_K)} \to 1, \tag{3.7}$$

where $p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \alpha, \beta_K) = \int_{\Phi_K} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n} | \phi_K) p_0(\phi_K) d\phi_K$.

The proof of the theorem is in Appendix B.1.2.

Theorem 2 allows us to asymptotically discard the undesired paths in $\mathcal{X}^n_{K,\epsilon}$ when comparing the marginal likelihoods, thus playing a vital role in the proof of model selection consistency in Theorem 1. To illustrate this, we consider Example 1, which is a special case of Theorem 1.

Example 1. *Let $K^* = 1$, i.e. the observations $\boldsymbol{y}_{1:n}$ are i.i.d. with density $f(\cdot|\theta^*)$, where θ^* is the true parameter. Let $\hat{\theta}$ be the MLE, which converges to θ^* as $n \to \infty$. By the asymptotic normality of the posterior distribution under the i.i.d. scenario and the Laplace method (Walker, 1969),*

$$\frac{\prod_{i=1}^n f(y_i|\hat{\theta})}{\sqrt{n} p_1(\mathbf{y}_{1:n})} = \frac{\prod_{i=1}^n f(y_i|\hat{\theta})}{\sqrt{n}\int_\Theta \prod_{i=1}^n f(y_i|\theta)\pi(\theta)d\theta} = O_P(1).$$

Suppose $K = 2$. Given $\mathbf{x}_{1:n} \in \mathcal{X}_2^n$, let n_1 and n_2 be defined as in (3.4). Then

$$\begin{aligned}&\int_{\Phi_2} p_2(\mathbf{y}_{1:n}, \mathbf{x}_{1:n}|\phi_2)p_0(\phi_2)d\phi_2\\ =&\int_\Theta \prod_{i:x_i=1} f(y_i|\theta_1)\pi(d\theta_1) \times \int_\Theta \prod_{i:x_i=2} f(y_i|\theta_2)\pi(d\theta_2) \times \int_{\mathcal{Q}_2} \prod_{i=1}^n q_{x_{i-1}x_i}\nu_2(dQ_2).\end{aligned}$$

Again, by the i.i.d. structure, for $k = 1, 2$,

$$\frac{\prod_{i:x_i=k}^n f(y_i|\hat{\theta}_k)}{\sqrt{n_k}\int_\Theta \prod_{i:x_i=k}^n f(y_i|\theta)\pi(\theta)d\theta} = O_P(1),$$

where $\hat{\theta}_k$ is the MLE obtained by the y_i with $x_i = k$. Thus $\hat{\theta}_k$ also converges to θ^* when $n_k \to \infty$. Hence, if $\frac{n_1 n_2}{n} \to \infty$ when both $n_1, n_2 \to \infty$ as $n \to \infty$,

$$\begin{aligned}&\frac{\int_{\Phi_2} p_2(\mathbf{y}_{1:n}, \mathbf{x}_{1:n}|\phi_2)p_0(\phi_2)d\phi_2}{p_1(\mathbf{y}_{1:n})}\\ =&\frac{\prod_{k=1}^2 \int_\Theta \prod_{i:x_i=k} f(y_i|\theta_k)\pi(d\theta_k)}{\int_\Theta \prod_{i=1}^n f(y_i|\theta)\pi(\theta)d\theta} \times \int_{\mathcal{Q}_2} \prod_{i=1}^n q_{x_{i-1}x_i}\nu_2(dQ_2)\\ =&O_P\left(\sqrt{\frac{n}{n_1 n_2}}\right)\int_{\mathcal{Q}_2} \prod_{i=1}^n q_{x_{i-1}x_i}\nu_2(dQ_2),\end{aligned} \tag{3.8}$$

which is less than or equal to $o_P(1)$ *since* $q_{x_{i-1}x_i} \in [0,1]$.

The major gap between equation (3.8) and Theorem 1 is that we need to sum up (3.8) over all possible paths $\mathbf{x}_{1:n} \in \mathcal{X}_2^n$. *However, there always exists* $x_{1:n}$ *such that* $n_1(\mathbf{x}_{1:n})$ *is small, e.g.* $n_1 = 1$, *making* $O_P(n/n_1n_2) \neq o_p(1)$. *These "irregular" paths forbid us from summing up (3.8).*

However, given $\epsilon_n \downarrow 0$ *as* $n \to \infty$, *Theorem 2 shows that, when summing up the paths, we can essentially ignore the paths with* $n_1 < n\epsilon_n$ *and* $n_2 < n\epsilon_n$. *After excluding these paths, the rest of the paths would yield a leading term*

$$O_P\left(\sqrt{\frac{n}{n_1n_2}}\right) \leq O_P\left(\sqrt{\frac{1}{\epsilon_n^2 n}}\right) = o_p(1),$$

if we choose $\epsilon_n \downarrow 0$ *slow enough such that* $\epsilon_n^2 n \to \infty$. *Hence, with the help of Theorem 2, we achieve the desired uniform convergence which leads to the order selection consistency in Theorem 1.*

The summation over paths as shown in Example 1 needs to be handled more carefully when $K^* > 1$, the general case considered in Theorem 1. The proof, given in Appendix B.1.3, involves an advanced version of Markov random walk representation technique (Fuh, 2003) to deal with the summation, as well as the characterization of posterior distributions for an HMM with unknown number of states (Gassiat & Rousseau, 2014).

3.3.3 Identifiability of Overly-fitted HMMs

In this section, we illustrate the identifiability issue of an overly-fitted HMM through a numerical example and explain how we overcome this obstacle, which is a difficult problem in the HMM model selection literature (Wang & Bickel, 2015), by introducing a regularity condition (B2) stated and construed in Appendix B.1.1.

We contemplate a Gaussian HMM where $Y_i|X_i = k \backsim N(\theta_k, 1)$. Suppose $K^* = 1$ and $\theta_1^* = 0$, i.e. the Y_i are i.i.d. standard Gaussian. When $K = 2$, the over-fitting scenario, any of the following three situations yields the same likelihood function as that under the true model, indicating that there are multiple "true" parameters when $K > K^*$.

1. Identical states, i.e. $\theta_1 = \theta_2 = 0$, with arbitrary Q_2.

2. Redundant state 1, i.e. $\theta_1 = 0$ and $q_{12} = q_{22} = 0$, with arbitrary θ_2.

3. Redundant state 2, i.e. $\theta_2 = 0$ and $q_{11} = q_{21} = 0$, with arbitrary θ_1.

The situation becomes more complicated with a larger K, as there might be numerous combinations of "true" parameters.

A formal characterization is to consider a $(K^* + 1)$ clustering of the $K\ (> \ K^*)$ states, where clusters 1 to K^* correspond to the true states 1 to K^* and the last cluster corresponds to the redundant state(s). To distinguish these "true" parameters, a weak identifiability condition (B2, Appendix B.1.1), which is also adopted in Gassiat & Rousseau (2014), is assumed. By defining a function that characterizes the first two derivatives of the likelihood function, we are able to impose a regularity condition on this function to ensure identifiability. Heuristically, this is reasonable because posterior convergence depends on the Taylor expansion to the second order. The condition holds for a broad family of mixtures of exponential distributions, including the location-scale Gaussian mixtures; see Gassiat & Rousseau (2014). The remarks for condition (B2) in Appendix B.1.1 give more detailed descriptions about the indications of the weak identifiability condition on an overly-fitted HMM.

3.3.4 Connections with Model Selection of Mixture Models

In this section, we discuss the connections of the order selection for HMMs with the model selection of mixture models. As mentioned in Section 3.1.3, the mixture model can be considered as a special case of an HMM, the transition matrix of which has all elements equal to each other. Consequently, the model selection of mixture models can follow the same procedure as the order selection for HMMs. Reversely, we can use the model selection of mixture models to determine the order of HMMs. Through a similar proof, we can show that the estimator of the order of an HMM is still consistent if we "ignore" the Markov dependency, i.e. regarding the HMM as a mixture model. This result is formalized in Theorem 3 and proved in Appendix B.1.4.

Theorem 3. *Assume that all the conditions in Theorem 1 hold, except that condition (B1) is replaced by (B1') in Appendix B.1.4, which restricts $\nu_K(\cdot|\beta_K)$ to be supported on $\tilde{\mathcal{Q}}_K = \{Q : q_{1k} = q_{2k} = \cdots = q_{Kk} \text{ for all } 1 \leq k \leq K\}$, i.e. assuming a prior for a mixture model without state dependency. Then the consistency of $\hat{K}_n$ in Theorem 1 still holds.*

As opposed to Theorem 1, the computational cost required by Theorem 3 is much smaller: instead of fitting HMMs, we instead only need to fit mixture models which live on lower dimensional spaces with nice independent structures on the latent variables.

In both Theorems 1 and 3, the convergence rate of the marginal likelihood ratio is $O_P(n^{-1/2} \log n)$. However, Theorem 3 requires n to be large so that $\boldsymbol{y}_{1:n}$ shows a "mixture model" behavior through the law of large numbers. This leads to a larger constant term in front of $n^{-1/2} \log n$ for the marginal likelihood ratio of Theorem 3 as compared to Theorem 1, especially for nearly diagonal transition matrices. Furthermore, the regularity condition (B1) is replaced by a stronger, i.e. more restrictive, condition (B1'); see Appendix B.1.4 for details.

3.4 Computation and Numerical Experiments

In this section, we first introduce our method of estimating the marginal likelihood and then provide numerical results comparing the marginal likelihood method and the BIC; at the end of the section, we give a brief discussion about the choice of priors and the order selection consistency of practical implementations.

3.4.1 Computing the Marginal Likelihood

Integration of the hidden states and the parameters over the joint likelihood is not trivial and does not have a simple analytical solution. In this section, we describe our procedure for computing the marginal likelihood as defined in Section 3.2.

Marginal Likelihood as a Normalizing Constant

Denote the joint distribution of $\boldsymbol{y}_{1:n}$ and $\boldsymbol{\phi}_K$ by $p(\boldsymbol{y}_{1:n}, \boldsymbol{\phi}_K) = p(\mathbf{y}_{1:n}|\boldsymbol{\phi}_K)p_0(\boldsymbol{\phi}_K)$, where $p(\mathbf{y}_{1:n}|\boldsymbol{\phi}_K)$, defined in equation (3.2), is the likelihood after integrating out the hidden states. Recall from (3.3) that the marginal likelihood of a K-state HMM, $p_K(\mathbf{y}_{1:n})$, is equal to $\int_{\Phi_K} p(\mathbf{y}_{1:n}|\boldsymbol{\phi}_K)\, p_0(\boldsymbol{\phi}_K)\, d\boldsymbol{\phi}_K$.

Our strategy is based on the following observation: the marginal likelihood $p_K(\mathbf{y}_{1:n})$, in fact, can be regarded as the normalizing constant of the posterior density $p(\boldsymbol{\phi}_K|\mathbf{y}_{1:n}) = p(\mathbf{y}_{1:n}, \boldsymbol{\phi}_K)/p_K(\boldsymbol{y}_{1:n})$. Thus the problem can be recast as the estimation of normalizing constant of this posterior density.

To do this, note that we can obtain posterior samples from $p(\boldsymbol{\phi}_K|\mathbf{y}_{1:n})$ using any Markov chain Monte Carlo (MCMC) algorithm (see Liu (2001) and references therein) since the un-normalized posterior likelihood $p(\mathbf{y}_{1:n}, \boldsymbol{\phi}_K)$ can be evaluated at any $\boldsymbol{\phi}_K$ using the forward algorithm (Baum & Petrie, 1966; Baum et al., 1970), which integrates out the hidden states. Alternatively, we can sample from the augmented space $\Phi_K \times \mathcal{X}_K^n$, i.e., sample model parameters and the hidden states iteratively till convergence. This alternative approach corresponds to the data augmentation method in Tanner & Wong (1987) and has been used for HMM model fitting (Rydén, 2008).

Given that we can sample from this posterior density, the question becomes: how to estimate the normalizing constant based on (posterior) samples. This has been studied by many researchers. We first give a brief review of the existing methods and then detail what we use.

Literature on Estimating Normalizing Constants

Early work involving Monte Carlo integrations include Ogata (1989) and Shao (1989). When the density is approximately Gaussian with a single mode, the Laplace approximation and the Bartlett adjustment are shown to be effective (DiCiccio et al., 1997). Methods based on importance sampling and reciprocal importance sampling requires knowledge of a “good” importance function whose region of interest covers that of the joint posterior to be integrated (Geweke, 1989; Oh & Berger, 1993; Newton & Raftery, 1994; Gelfand & Dey, 1994; Ionides, 2008; Neal, 2005; Steele et al., 2006; Chen & Shao, 1997). Estimating the marginal likelihood based on MCMC output has been developed in Chib (1995); Geyer (1994); Chib & Jeliazkov (2001, 2005); de Valpine (2008); Petris & Tardella (2007). DiCiccio et al. (1997) and Chen & Shao (1997) give general reviews of a variety of methods, including the Laplace approximation, importance sampling, bridge sampling (Meng & Wong, 1996), path sampling (Gelman & Meng, 1998), and methods based on MCMC output for computing Bayes factors, which is the ratio of normalizing constants; see references therein.

Adopted Estimation Procedure

The importance sampling and reciprocal importance sampling are simple and fast ways of estimating the normalizing constant if a good importance function close to the target density can be specified. Since we already have posterior samples from the unnormalized density, it can be utilized as a guidance of choosing a good importance function for either the importance sampling or the reciprocal importance sampling. Therefore, our strategy is to use the importance sampling or the reciprocal importance sampling to estimate the normalizing constant $p_K(\mathbf{y}_{1:n})$, where the importance function is chosen based on the posterior samples from $p(\boldsymbol{\phi}_K|\boldsymbol{y}_{1:n})$. Since the posterior samples not necessarily gives enough information about the tail of the posterior distribution, the importance function might be a poor approximation of the target posterior distribution in the tail region, which might result in unstable estimators. Therefore, we use the locally restricted importance sampling or reciprocal importance sampling, which is more robust to the tail behavior of the target posterior distribution $p(\boldsymbol{\phi}_K|\boldsymbol{y}_{1:n})$, see DiCiccio et al. (1997).

We now give our procedure for estimating the marginal likelihood $p_K(\mathbf{y}_{1:n})$.

1. Obtain posterior samples. Sample from $p(\boldsymbol{\phi}_K|\mathbf{y}_{1:n})$ using a preferred MCMC algorithm, and denote the samples by $\{\boldsymbol{\phi}_K^{(i)}\}_{i=1}^N$ (where N is often a few thousand).

2. Find a "good" importance function. Fit a Gaussian mixture model using the

samples $\{\boldsymbol{\phi}_K^{(i)}\}_{i=1}^N$, where the number of mixing components is given by either (a) any clustering algorithm, or (b) a pre-fixed number which is large enough. Construct the importance function $g(\cdot)$ by fitting a Gaussian mixture, or using a heavier-tailed density as the mixture component; for example, using t distribution with a small degree of freedom, such as 2 or 3, with the same location and scale parameters as the fitted Gaussian mixture components.

3. Choose a finite region. Choose Ω_K to be a bounded subset of the parameter space such that $1/2 < \int_{\Omega_K} g(\cdot) < 1$. This can be achieved through finding an appropriate finite region for each mixing component of $g(\cdot)$, avoiding the tail parts.

4. Estimate $p_K(\boldsymbol{y}_{1:n})$ using either way as follows:

   - Reciprocal importance sampling. Approximate $p_K(\mathbf{y}_{1:n})$ by

   $$\hat{p}_K^{(RIS)}(\boldsymbol{y}_{1:n}) = \left[\frac{1}{N\int_{\Omega_k} g(\cdot)} \sum_{i=1}^N \frac{g(\boldsymbol{\phi}_K^{(i)})}{p(\mathbf{y}_{1:n}, \boldsymbol{\phi}_K^{(i)})} I_{\boldsymbol{\phi}_K^{(i)} \in \Omega_K}\right]^{-1}, \tag{3.9}$$

   where $I_{\boldsymbol{\phi}_K^{(i)} \in \Omega_K} = 1$ if $\boldsymbol{\phi}_K^{(i)} \in \Omega_K$ and zero otherwise.

   - Importance sampling.

     (a) Draw M independent samples from $g(\cdot)$, denoted by $\{\boldsymbol{\psi}_K^{(j)}\}_{1\le j\le M}$.

     (b) Approximate $p_K(\mathbf{y}_{1:n})$ by

     $$\hat{p}_K^{(IS)}(\boldsymbol{y}_{1:n}) = \frac{1}{MP_\Omega} \sum_{j=1}^M \frac{p(\mathbf{y}_{1:n}, \boldsymbol{\psi}_K^{(j)})}{g(\boldsymbol{\psi}_K^{(j)})} I_{\boldsymbol{\psi}_K^{(j)} \in \Omega_K}, \tag{3.10}$$

     where $I_{\boldsymbol{\psi}_K^{(j)} \in \Omega_K} = 1$ if $\boldsymbol{\psi}_K^{(j)} \in \Omega_K$ and zero otherwise; $P_\Omega = \#\mathcal{S}/N$, where $\mathcal{S} = \{i : \boldsymbol{\phi}_K^{(i)} \in \Omega_K; 1 \leq i \leq N\}$ and $\#\mathcal{S}$ denotes its cardinality.

The purpose of Step 2 is to construct a reasonable importance function that covers the mode of the target density $p(\boldsymbol{\phi}_K|\mathbf{y}_{1:n})$ thus the clustering algorithm, if ever adopted, does not need to be "optimal" in any sense. Therefore, a conservative recommendation is to choose overly-fitted Gaussian (or student t) mixtures based on the posterior samples obtained in Step 1. Moreover, the heavy tailed distribution and the truncated regions both serve the purpose of obtaining a robust importance sampling estimator. If reciprocal importance sampling is used, a heavy tailed distribution is not recommended for sake of estimation robustness.

Simulation studies of various target densities (skewed, heavy-tailed, and high-dimensional) with known normalizing constants validates the efficacy of the proposed procedure, regardless of the shape of target density or the dimension of the parameter space. See Appendix B.2 for detailed descriptions.

3.4.2 Simulation Studies for HMM Order Selection

In the numerical experiments, we fix the mean parameters of a K-state HMM to be $\boldsymbol{\mu} = (1, 2, \ldots, K)$ and vary the variances $\boldsymbol{\sigma}^2 = (\sigma^2, \ldots, \sigma^2)$. Equal variances is adopted here for simplicity of the presentation of the results but this is not part of the model assumptions. We consider four kinds of transition matrices, corresponding to flat ($P_K^{(1)}$), moderate and strongly diagonal ($P_K^{(2)}, P_K^{(3)}$) and strongly off-diagonal

($P_K^{(4)}$) cases:

$$P_K^{(1)} = \frac{1}{K}E_K,\ P_K^{(2)} = \left[0.8 - \frac{0.2}{K-1}\right] I_K + \frac{0.2}{K-1}E_K, \tag{3.11}$$

$$P_K^{(3)} = \left[0.95 - \frac{0.05}{K-1}\right] I_K + \frac{0.05}{K-1}E_K, \tag{3.12}$$

$$P_K^{(4)} = \frac{0.9}{K-1}E_K - \left[\frac{0.9}{K-1} - 0.1\right] I_K, \tag{3.13}$$

where E_K is the $K \times K$ matrix with all elements equal to 1 and I_K is the $K \times K$ identity matrix. For example, for $K = 4$, the four matrices are:

$$P_4^{(1)} = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix},\ P_4^{(2)} = \begin{pmatrix} 0.8 & 1/15 & 1/15 & 1/15 \\ 1/15 & 0.8 & 1/15 & 1/15 \\ 1/15 & 1/15 & 0.8 & 1/15 \\ 1/15 & 1/15 & 1/15 & 0.8 \end{pmatrix};$$

$$P_4^{(3)} = \begin{pmatrix} 0.95 & 1/60 & 1/60 & 1/60 \\ 1/60 & 0.95 & 1/60 & 1/60 \\ 1/60 & 1/60 & 0.95 & 1/60 \\ 1/60 & 1/60 & 1/60 & 0.95 \end{pmatrix},\ P_4^{(4)} = \begin{pmatrix} 0.1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{pmatrix}.$$

The number of observations, n, varies from 200 to 2000, and the true number of hidden states, K, ranges from 2 to 4. Figure 3.1 illustrates a few simulated HMM traces. We conduct $m = 200$ repeated simulations, each of which compares the marginal likelihood method with the BIC as follows.

1. Simulate n observations from the HMM with K states and the specified set of parameters.

2. Apply the Baum-Welch algorithm with multiple starting points (in our case,

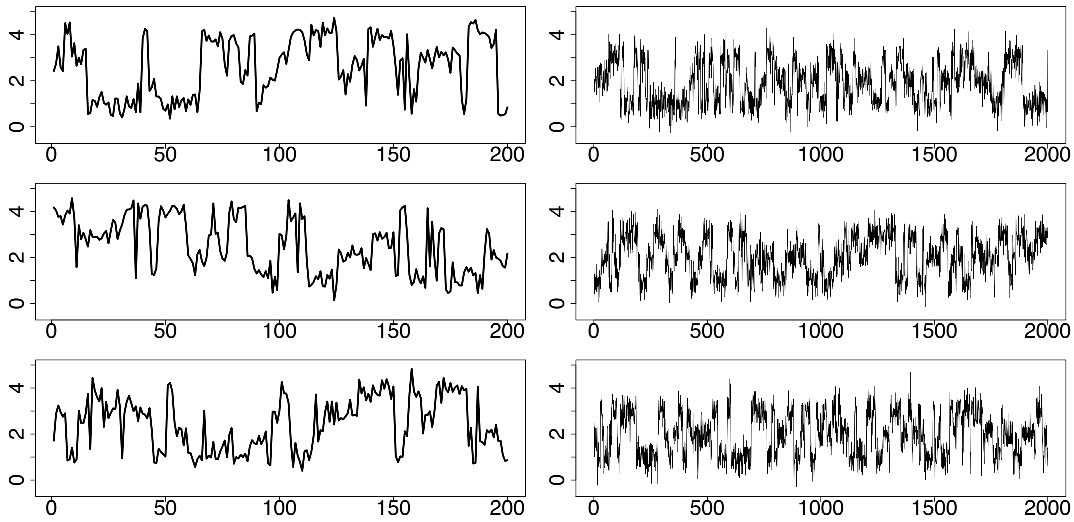

Figure 3.1: Sample HMM traces. The left column shows three simulated HMM traces with $n = 200$ observations and $K = 4$ hidden states: $\sigma = 0.3$ and the transition matrix is $P_4^{(2)}$. The right column shows three simulated HMM traces with $n = 2,000$ observations and $K = 3$ hidden states: $\sigma = 0.4$ and the transition matrix is $P_3^{(3)}$.

50 randomly generated starting points) to obtain the maximum likelihood values for $\tilde{K}$-state HMM, thus giving the BIC of HMMs with $\tilde{K}$-states denoted by $BIC_n(\tilde{K})$, $\tilde{K} = 2, 3, 4, \ldots$; let $\hat{K}_n^{BIC} = \arg\max_{\tilde{K}} BIC_n(\tilde{K})$.

3. Calculate the marginal likelihood of a $\tilde{K}$-state HMM based on the importance sampling procedure detailed in Section 3.4.1, $\tilde{K} = 2, 3, 4, \ldots$; let $\hat{K}_n^{ML} = \arg\max_{\tilde{K}} P_{\tilde{K}}(\mathbf{y}_{1:n})$.

Note that when calculating the BIC, we avoid the cases when the maximum likelihood estimators converge to the boundary of the parameter space as mentioned in Section 3.1.3. Moreover, we replicate the common practice of applying the BIC by using multiple starting points for obtaining MLEs.

Table 3.1 summarizes the results from repeated simulations, showing the frequency of correct identification of the true number of hidden states using the marginal likelihood method and the BIC.

From the simulation studies, it is evident that the marginal likelihood method outperforms the BIC in several aspects. First, the frequency of correct identification of the number of hidden states using the marginal likelihood method is much higher, especially when the number of observations is small (200 as opposed to 2000). Second, the marginal likelihood method is more robust to low signal to noise ratio, which can be seen from Table 3.1. The success rates of the marginal likelihood method and the BIC both drop as we increase the noise level σ from 0.2 to 0.4. However, the success rate of the BIC drops much faster as opposed to that of the marginal likelihood. Third, since the number of (unknown) model parameters is quadratic in K, given the same number of observations, the more number of hidden states, the harder the order selection. The marginal likelihood method appears more robust to the true number of hidden states than the BIC.

3.4.3 Applications to Single-Molecule Data

We apply the proposed methodology to the single-molecule data analyzed in Chen et al. (2016).

Single-molecule experiments track the dynamic behaviors of individual molecules through measuring in real time the energy transfer rates between two light sensitive molecules labeled at different compartments of a molecule. The measurements, FRET (fluorescence resonance energy transfer), is a monotone function of the corresponding distances. In Chen et al. (2016), each FRET trace is modeled as an HMM. The number of hidden states of each HMM corresponds to the number of conformations of a molecular complex, which is of immense significance in biology.

The marginal likelihood method gives very similar results as opposed to the BIC

K	σ	n	$Q_K = P_K^{(1)}$		$Q_K = P_K^{(2)}$		$Q_K = P_K^{(3)}$		$Q_K = P_K^{(4)}$	
			ML	BIC	ML	BIC	ML	BIC	ML	BIC
2	0.2	200	100	100	100	100	100	100	100	100
2	0.3	200	100	100	100	100	100	100	100	100
2	0.4	200	100	100	100	100	100	100	100	100
3	0.2	200	100	100	100	100	95.0	96.0	100	100
3	0.3	200	62.5	22.5	100	99.5	96.0	94.5	99.0	92.5
3	0.4	200	1.50	0.00	91.0	77.0	88.5	88.0	25.0	10.5
4	0.2	200	100	90.0	100	100	81.0	76.0	100	97.5
4	0.3	200	4.00	0.00	97.0	85.0	65.0	60.0	22.0	0.50
4	0.4	200	0.00	0.00	45.0	21.0	37.5	37.0	0.00	0.00
5	0.2	200	99.0	15.5	99.5	95.0	55.0	44.0	99.5	29.0
5	0.3	200	0.50	0.00	82.0	37.0	24.0	19.0	1.00	0.00
5	0.4	200	0.00	0.00	10.5	1.00	7.00	4.50	0.00	0.00
2	0.2	2000	100	100	100	100	100	100	100	100
2	0.3	2000	100	100	100	100	100	100	100	100
2	0.4	2000	100	100	100	100	100	100	100	100
3	0.2	2000	100	100	100	100	100	100	100	100
3	0.3	2000	100	100	100	100	100	100	100	100
3	0.4	2000	98.5	72.0	100	100	100	100	100	100
4	0.2	2000	100	100	100	100	100	100	100	100
4	0.3	2000	99.5	98.5	100	100	100	100	100	100
4	0.4	2000	4.50	0.00	100	100	100	100	84.0	20.5
5	0.2	2000	100	100	100	100	100	100	100	100
5	0.3	200	95.0	23.5	100	100	100	100	99.0	87.0
5	0.4	200	0.00	0.00	100	100	100	100	2.00	0.00

Table 3.1: The frequency (in %) of correct identification of the true number of hidden states, out of 200 repeated simulations for each entry, using the marginal likelihood method (ML) and the BIC. Both $n = 200$ and $n = 2000$ observations are considered. $K(= 2, 3, 4)$ is the true number of hidden states; σ is the standard deviation of each hidden state around its mean; Q_K denotes the transition matrix: the matrices $P_K^{(1)}, P_K^{(2)}, P_K^{(3)}, P_K^{(4)}$ are defined in equations (3.11) to (3.13).

applied in Chen et al. (2016) for majority of FRET traces. However, for very few traces, e.g. the two traces in Figure 3.2, the marginal likelihood method and the BIC do not agree: the marginal likelihood method gives a selection of 3 hidden states whereas the BIC gives a selection of 2 hidden states for both traces. As demonstrated in Chen et al. (2016), these two traces are indeed 3-state traces under the hierarchical model, which combines information from multiple traces to identify some rarely occurred states. Therefore, the marginal likelihood method is more sensitive as opposed to the BIC in detecting rarely occurring states.

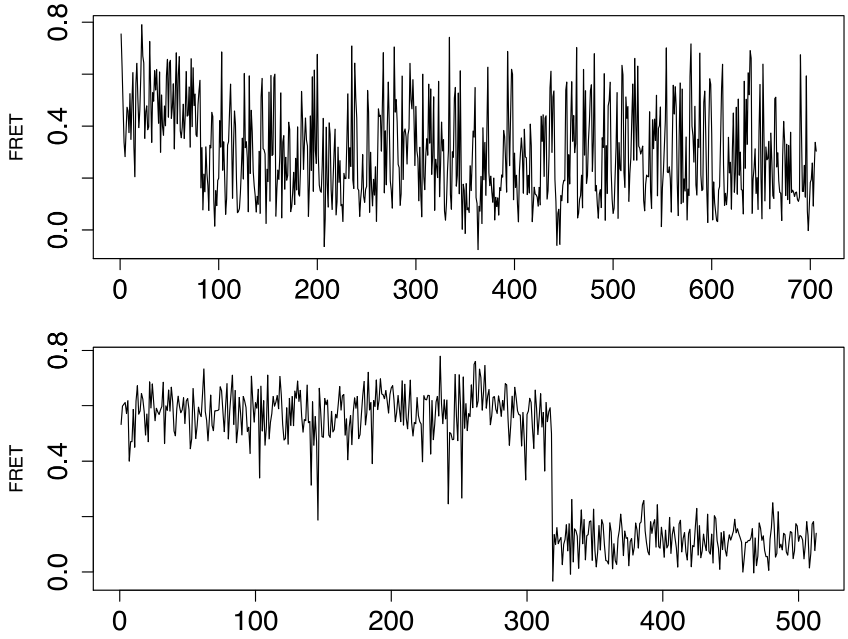

Figure 3.2: Two traces from the single-molecule data in Chen et al. (2016).

3.4.4 Discussions

In this section, we give recommended choices of the prior parameters based on empirical evidence from simulation studies in Section 3.4.4. Section 3.4.4 discusses the consistency of practical implementations, which could also guide the choices of other tuning parameters.

Choice of Prior Distributions

From the asymptotic results in Section 3.3, the influence of priors vanishes as the number of observations goes to infinity. However, in practice, the number of observations is a fixed number and the choice of priors would have an impact on the results. Now we give our recommendations of the choice of prior distributions based on empirical evidence in running simulation studies. Practitioners should be aware that the *best reasonable* prior distribution often comes from incorporating scientific knowledge of the specific problem in the field of study.

In the simulation studies in Section 3.4.2, we choose flat, conjugate priors and the results look quite promising. The prior for each row of the transition matrix is an independent Dirichlet distribution with parameters all equal to 1, corresponding to a 'flat prior'. The priors for the means $\{\mu_k\}_{k=1}^K$ are set to be independent Gaussian with means $\{\mu_{0k}\}_{k=1}^K$ and large variances, e.g. 100^2. $\{\mu_{0k}\}_{k=1}^K$ is chosen to be data-dependent: the μ_{0k} are set as the equally spaced quantiles of the observations $\boldsymbol{y}_{1:n}$. The priors for the variances of each hidden state $\{\sigma_k^2\}_{k=1}^K$ are chosen to be independent inverse chi-squared distribution with degree of freedom 3 and and the scale can be chosen based on empirical estimators of the variability in the data: we can simply take the square root of the scale as the interquartile range of the observations divided by $K-1$ or K.

Consistency of Practical Implementation

In practice, we replace the marginal likelihoods in Theorem 1 by their corresponding importance sampling estimators. Theorem 7 in Appendix B.3 indicates that the order selection consistency still holds when we plug in consistent estimators of the marginal likelihoods. The asymptotic property of the marginal likelihood estimators we adopt depends on the number of posterior samples and the number of importance samples, which are extra tuning parameters. This is quantified in Lemma 6. When combined with the convergence rate of the marginal likelihood ratios in Theorem 1, Lemma 6 can serve as a guidance towards choosing the relevant turning parameters. Refer to Appendix B.3 for rigorous descriptions and proofs of the results.

3.5 Conclusions

In this paper, we use the marginal likelihood to determine the number of hidden states for hidden Markov models.

The proposed method is consistent theoretically under mild conditions. The difficulties of overly fitted HMMs are circumvent by introducing and proving an *asymptotic path ignorance* result which enables discarding undesired hidden state trajectories that blows up the likelihood function.

Furthermore, we propose a computation algorithm to robustly estimate the order of an HMM trace through the estimation of normalizing constants. Extensive simulation studies verify our proposed approach and demonstrate its power against the widely adopted approach, the BIC, which suffers from lack of theoretical justification.

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

John Tukey

The primary product of a research inquiry is one or more measures of effect size, not p values.

Jacob Cohen

4

Evaluating Parallelisable Bayesian Computation Methods

4.1 Introduction

Monte Carlo methods have become central tools for solving demanding computational problems in a wide variety of scientific disciplines. Markov chain Monte Carlo methods have been widely adopted as an effective way of obtaining random samples from posterior distributions, revolutionizing the practice of Bayesian inference. Since the introduction of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and Gibbs sampling (Geman & Geman, 1984), researchers have successfully developed various improved Markov chain Monte Carlo algorithms for achieving more independent/effective samples with fewer iterations. Multiple-try algorithms (Liu et al., 2000; Qin & Liu, 2001) explore multiple points of the sample space at each iteration; the Hamiltonian Monte Carlo algorithm (Duane et al., 1987; Neal, 2011) brings insights from the thermodynamic integration into the Markov chain Monte Carlo steps and performs very well for highly-skewed density functions; the parallel tempering algorithm (Swendsen & Wang, 1986; Neal, 1996) and the equi-energy sampler (Kou et al., 2006) enables the Markov chain to jump among different modes instead of being trapped in a local mode. Other developments include slice sampling (Neal, 2003) and reversible jump Markov chain Monte Carlo (Green, 1995).

4.1.1 Waste-Recycling Markov Chain Monte Carlo Algorithms

While all the Markov chain Monte Carlo algorithms mentioned above obtain one sample at each iteration upon convergence of the Markov chain, the waste-recycling algorithm, first introduced by Tjelmeland (2004) and Frenkel (2004) and later developed by Frenkel (2006) and Douc & Robert (2011), uses all the proposed points, including rejected ones, to perform posterior mean estimation. The algorithm is valid due to the fact that the underlying Markov chain remains invariant. Intuitively, by assigning a weight to each of the proposed points, the information from the rejected samples can be used to obtain more precise estimates of the posterior. The weighted samples serve as a Rao-Blackwellization of the original Markov chain Monte Carlo algorithm upon which waste-recycling is performed. However, the waste-recycling algorithm does not always outperform the original algorithm in terms of asymptotic variance of posterior mean estimators, see Delmas & Jourdain (2009).

4.1.2 Parallelisable Markov Chain Monte Carlo Algorithms

Using parallel computing for effective Bayesian inference is a growing and promising field. The bottleneck comes from the inherent sequential nature of Markov chain Monte Carlo algorithms, i.e. relying on a Markov chain to draw dependent samples, one sample at each iteration. One possible route forward would be a combination of both the Markov chain Monte Carlo algorithm and the importance sampling: exploring the target density surface using a Markov chain whereas incorporating the local weighting into each iteration.

Calderhead (2014) proposed a parallelisable Markov chain Monte Carlo algorithm to draw multiple samples from each iteration while evaluating the target densities at multiple proposed points using different cores in parallel. Taking the Metropolis-Hastings algorithm as a special case, the parallelisable algorithm proposes multiple points at each iteration and resample multiple of these points with replacement according to a weighting scheme that yields the detailed balance condition of the Markov chain. This gives a general scheme usable within many existing Markov chain Monte Carlo algorithms. However, with all the efforts made for parallelising the computation, several new questions arise.

1. Calderhead (2014) recommends choosing the number of proposals at each iteration based on the number of cores available; but how many resampled points should we collect at each iteration of the algorithm?

2. At each iteration, multiple correlated samples are collected, meaning the existing method for estimating the effective sample size is no longer applicable. Can we generalize the effective sample size estimation for this new data structure to quantify the effectiveness of obtaining multiple samples per iteration?

3. Is it worthwhile taking the trouble to call multiple cores at each iteration? The efficiency of such algorithms for posterior mean estimation as compared to that of regular Markov chain Monte Carlo algorithms without parallelisable structure needs to be characterized.

To answer these questions, we provide a general framework for parallelisable waste-recycling Markov chain Monte Carlo algorithms under which the effective sample size of such parallelisable Markov chain Monte Carlo algorithms can be estimated easily using moment estimators. This finite-sample property, i.e. evaluation of the precision of posterior mean estimators based on finite number of iterations, is more applicable than the asymptotic properties given in Delmas & Jourdain (2009); Douc & Robert (2011) for waste-recycling algorithms.

4.1.3 Parallelisable Waste-Recycling Markov Chain Monte Carlo Algorithms

Our work naturally shows that eliminating the resampling step in Calderhead (2014) and keeping track of all the proposals and the weights will improve the posterior estimation efficiency (section 4.2.3). Moreover, by separating the proposal and weighting, we allow for a flexible choice of multi-proposal kernels and the weighting scheme (section 4.2.2). We refer to the resulting algorithm as the locally weighted parallel Markov chain Monte Carlo algorithm. In this algorithm, weighted samples are obtained at each iteration of the Markov chain Monte Carlo algorithm where the weights of the samples can be calculated in parallel.

These algorithms provide a general framework for constructing parallelisable Markov chain Monte Carlo algorithms. Compared with the algorithm in Calderhead (2014), the proposed algorithm is guaranteed to give more efficient posterior mean estimators. Further, we give a simple generalization of the effective sample size for Markov chain Monte Carlo algorithms (Kass et al., 1998; Liu, 2001, p. 126). We can also use all the proposed points and their weights to estimate the efficiency gain/loss of our proposed algorithm compared to its regular Markov chain Monte Carlo counterpart where only one accepted sample is collected; thus giving guidance for when to adopt or discard such parallelisable Markov chain Monte Carlo algorithms.

We recommend keeping record of all the weights and proposed points when running parallelisable Markov chain Monte Carlo algorithms when the storage is permissible. The benefit is two-fold: the first is the variance reduction of posterior mean estimators over Calderhead's estimators, which will be shown in section 4.2.3; the second is that the proposed points with their weights can be used to quantify performance through the generalized effective sample size.

The rest of the paper has four sections. We review the parallel Markov chain Monte Carlo algorithm in Section 4.2.1; describe the idea of the locally weighted algorithm in a general context in Section 4.2.2; and then present in Section 4.2.3 some properties of the new algorithm and its estimators: unbiasedness, relative statistical efficiency, and the generalized effective sample size. We then apply the algorithm to the Hamiltonian Monte Carlo transition kernel in Section 4.2.4. Section 4.3 illustrates the methodologies proposed in this paper using numerical simulations. Section 4.4 concludes. The appendix contains the relevant proofs.

4.2 Locally Weighted Parallelisable Markov Chain Monte Carlo

Assume that we wish to estimate $\mu_h = E_\pi\{h(x)\}$ by sampling from the target distribution $\pi(\cdot)$ defined on $\Omega \subset \mathbb{R}^d$, known up to a normalizing constant; $h(\cdot)$ is any function with finite first moment with respect to $\pi(\cdot)$, i.e. $E_\pi\{|h(x)|\} < \infty$. We have two kinds of Markov kernels: the proposal kernel from a sample x to y (which can be a single sample or vector of multiple samples), denoted by $K(x, y)$, can be any kernel that propose (potentially multiple) points y from the current position x; the transition kernel from x to x', denoted by $T(x, x')$, is a Markov kernel that is π-invariant. We use x_{-i} to denote the collection of x's after removing the ith element. The number of iterations is denoted by n and the number of proposals at each iteration is denoted by M. $x_i^{(j)}$ denotes the ith proposal in the jth iteration.

4.2.1 Recap of Parallel Markov Chain Monte Carlo

The parallel Markov chain Monte Carlo algorithm in Calderhead (2014) with n iteration and M proposals, N samples per iteration has the following steps. Our description of this algorithm is slightly different for ease of later comparisons.

Algorithm 1. ***Parallel Markov chain Monte Carlo***

```
Set $x_0^{(1)}$ to be the initial value
For $j = 1$ to $j = n$
    Draw proposals $\{x_1^{(j)}, \ldots, x_M^{(j)}\}$ from $K(x_0^{(j)}, \cdot)$
    Calculate $w(x_0^{(j)}), \ldots, w(x_M^{(j)})$ as defined in Eq (4.1) or Eq (4.2)
    For $i = 1$ to $i = N$
        Draw $y_i^{(j)}$ from $\{x_0^{(j)}, \ldots, x_M^{(j)}\}$ with probabilities $\{w(x_0^{(j)}), \ldots, w(x_M^{(j)})\}$
    Draw $x_0^{(j+1)}$ from $\{x_0^{(j)},\ldots,x_M^{(j)}\}$ with probabilities $\{w(x_0^{(j)}),\ldots,w(x_M^{(j)})\}$
  Output $\{y_i^{(j)} : i = 1,\ldots,N; j = 1,\ldots,n\}$
```

$$w(x_i^{(j)}) = \frac{1}{M}\min\{1, \frac{\pi(x_i^{(j)})K(x_i^{(j)}, x_{-i}^{(j)})}{\pi(x_0^{(j)})K(x_0^{(j)}, x_{-0}^{(j)})}\}, \quad (i = 1,\ldots,M); \quad w(x_0^{(j)}) = 1 - \sum_{i=1}^{M} w(x_i^{(j)}). \tag{4.1}$$

$$w(x_i^{(j)}) = \frac{\pi(x_i^{(j)})K(x_i^{(j)}, x_{-i}^{(j)})}{\sum_{i=0}^{M}\pi(x_i^{(j)})K(x_i^{(j)}, x_{-i}^{(j)})}, \quad (i = 0,\ldots,M). \tag{4.2}$$

Points $y_i^{(j)}$ for $i = 1,\ldots,N$, $j = 1,\ldots,n$ are collected according to Algorithm 1 so that μ_h could be estimated by $\hat{\mu}_h^{pmcmc} = \frac{1}{nN}\sum_{j=1}^{n}\sum_{i=1}^{N} h(y_i^{(j)})$.

4.2.2 Main Idea of Locally Weighted Parallel Markov Chain Monte Carlo

Now we describe the locally weighted parallel Markov chain Monte Carlo algorithm in a very general framework, allowing flexible weighting scheme and propagation kernels.

Define the 1st and 2nd version of $w(x_i^{(j)})$ same as Eq (4.1) and (4.2), respectively. Collect points $x_i^{(j)}$ with weights $w(x_i^{(j)})$ for $i = 0,\ldots,M$, $j = 1,\ldots,n$ according to Algorithm 2 so that μ_h could be estimated by $\hat{\mu}_h^{lwpmcmc} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{M} h(x_i^{(j)})w(x_i^{(j)})$.

Algorithm 2. *Locally weighted parallel Markov chain Monte Carlo*

```
Set $x_0^{(1)}$ to be the initial value
For $j = 1$ to $j = n$
    Draw proposals $\{x_1^{(j)},\ldots,x_M^{(j)}\}$ from $K(x_0^{(j)},\cdot)$
    In parallel, calculate $w(x_0^{(j)}),\ldots,w(x_M^{(j)})$ as defined in Eq (4.1) or Eq (4.2)
```

```
        Draw $x_0^{(j+1)}$ from a transition kernel $T(x_0^{(j)}, \cdot)$ to propagate Markov chain
    Output $x_i^{(j)}$ with weights $w(x_i^{(j)})$ for $i = 0, \dots, M$; $j = 1, \dots, n$
```

Remark 1. *In the propagation step, we have the flexibility of choosing the transition kernel $T(x_0^{(j)}, x')$, as long as the detailed balance condition holds. Of course we can simply choose the transition according to the weights as in Algorithm 1. Empirically from simulation studies, this separation of the transition from the weighting scheme, when applied appropriately, boosts the efficiency of posterior mean estimators.*

Remark 2. *Setting $M = 1$ gives the waste-recycling Metropolis-Hastings algorithm, see Frenkel (2004) and Tjelmeland (2004).*

4.2.3 Properties of the Algorithm and the Generalized Effective Sample Size

In this section, we first study unbiasedness and statistical efficiency of locally weighted algorithms against regular Markov chain Monte Carlo algorithms for estimating means of (nonlinear) functions with a finite number of iterations. Next, we introduce a generalized effective sample size for the proposed algorithms and describe its estimation using all proposals and weights when the number of iterations is large.

Throughout this section, we denote $\hat{\mu}_h^{pmcmc}$, $\hat{\mu}_h^{lwpmcmc}$ and $\hat{\mu}_h^{mcmc}$ be the posterior mean estimators of $h(\cdot)$ using Calderhead's parallel, our locally weighted, and standard Markov chain Monte Carlo algorithms respectively. For all these algorithms n denotes the number of iterations, M is the number of proposals within each iteration in parallel and locally weighted parallel Markov chain Monte Carlo, and N is the number of samples obtained within each iteration in the former.

Unbiasedness and Efficiency Comparisons

In theorems 4 and 5, we show that not only do we obtain unbiased estimators using Algorithm 2, but that we reduce the variance of the posterior mean estimator as compared to Algorithm 1 with the same transition kernels. Proposition 1 gives a simple moment estimator of the relative reduction of variance using local weighting as opposed to parallel Markov chain Monte Carlo. The proofs are in the Appendix.

Theorem 4. *(**Unbiasedness**) When in equilibrium, the estimators produced by Algorithm 2 are unbiased, i.e.* $E(\hat{\mu}_h^{lwpmcmc}) = \mu_h$.

Theorem 5. *(**Variance Reduction**) Given the same transition kernel, Algorithm 2 is a Rao-Blackwellization of Algorithm 1, thus yielding smaller variance for the estimated mean, i.e.* $\mathrm{var}(\hat{\mu}_h^{lwpmcmc}) < \mathrm{var}(\hat{\mu}_h^{pmcmc})$.

Proposition 1. *(**Efficiency Gain**) Define the efficiency gain of posterior mean estimators of* $h(\cdot)$, $\hat{\mu}_h^{lwpmcmc}$ *against* $\hat{\mu}_h^{pmcmc}$, *as the relative reduction of variance:*

$$EG\left(h; \{x_i^{(j)}, w(x_i^{(j)})\}_{i=0,\ldots,M; j=1,\ldots,n}; N\right) = \frac{\mathrm{var}(\hat{\mu}_h^{pmcmc}) - \mathrm{var}(\hat{\mu}_h^{lwpmcmc})}{\mathrm{var}(\hat{\mu}_h^{lwpmcmc})}. \qquad (4.3)$$

The moment estimator of the efficiency gain of $\hat{\mu}_h^{lwpmcmc}$ *againt* $\hat{\mu}_h^{pmcmc}$ *is:*

$$EG\left(h; \{x_i^{(j)}, w(x_i^{(j)})\}_{i=0,\ldots,M; j=1,\ldots,n}; N\right) = \frac{E(\bar{g}) - E\{(\bar{h})^2\}}{N\ \mathrm{var}(\bar{h})}, \qquad (4.4)$$

upon convergence of the Markov chain, $\mathrm{var}(\bar{h})$ *is estimated by the sample variance of quantities* $\{\sum_{i=0}^M w(x_i^{(j)})h(x_i^{(j)})\}_{j=1,\ldots,n}$; $E(\bar{g}), E\{(\bar{h})^2\}$ *are estimated by sample means of quantities* $\{\sum_{i=0}^M w(x_i^{(j)})h(x_i^{(j)})^2\}_{j=1,\ldots,n}$ *and* $[\{\sum_{i=0}^M w(x_i^{(j)})h(x_i^{(j)})\}^2]_{j=1,\ldots,n}$ *respectively.*

Remark 3. *Under the same proposal and transition kernels, as* $N \to \infty$, *i.e. resampling infinite number of proposals at each iteration, the parallel Markov chain*

Monte Carlo algorithm converges to the locally weighted algorithm:

$$EG(h; \{x_i^{(j)}, w(x_i^{(j)})\}_{i=0,\ldots,M; j=1,\ldots,n}; N) \to 0, \mathrm{var}(\hat{\mu}_h^{lwpmcmc})/\mathrm{var}(\hat{\mu}_h^{pmcmc}) \to 1.$$

Generalizing the Effective Sample Size

In the Monte Carlo literature, the notion of effective sample size is widely adopted to evaluate the performance of a sampling procedure (Liu, 2001, p. 269; Gelman et al., 2013, p. 286). If $\hat{\mu}^{mcmc}$ is the standard estimator of the mean, effective sample size for Markov chain Monte Carlo algorithms is defined as

$$ESS_{mcmc} = \frac{\sigma^2}{\mathrm{var}(\hat{\mu}_{mcmc})} = \frac{n}{1 + 2\sum_k \rho_k}, \tag{4.5}$$

where ρ_k is the lag-k autocorrelation of the Markov chain $\{x_0^{(j)}\}_{j=1}^n$, the samples from the target distribution $\pi(\cdot)$ with σ^2 as the variance of $\pi(\cdot)$. It can be estimated using moment estimators and the spectral density at frequency 0 (see e.g. Andrews, 1991; Müller, 2014).

The weighting scheme of our proposed algorithm or the correlated multiple samples of Calderhead's parallel algorithm make the original effective sample size estimation inappropriate. Thus we derive a more general measure of the effective sample size. Recall that the output of Algorithm 2 is $n(M + 1)$ weighted samples, producing the mean estimate

$$\hat{\mu}^{lwpmcmc} = \frac{1}{n}\sum_{j=1}^{n} \bar{x}^{(j)}, \quad \text{where} \quad \bar{x}^{(j)} = \sum_{i=0}^{M} w(x_i^{(j)}) x_i^{(j)}.$$

Theorem 6. *The effective sample size for Algorithm 2 can be written as*

$$ESS_{lwpmcmc} = \frac{\sigma^2}{\mathrm{var}(\hat{\mu}^{lwpmcmc})} = \frac{n\sigma^2}{\mathrm{var}(\bar{x})\left(1 + 2\sum_k \gamma_k\right)}, \tag{4.6}$$

where γ_k is the lag-k autocorrelation function of $\left\{\bar{x}^{(j)}\right\}_{j=1}^{n}$ and $\mathrm{var}(\bar{x}^{(j)}) = \mathrm{var}(\bar{x})$ for all $1 \leq j \leq n$ by the stationarity of the Markov chain upon convergence.

Corollary 1. *The effective sample size for Algorithm 1 can be written as*

$$ESS_{pmcmc} = \frac{\sigma^2}{\mathrm{var}(\hat{\mu}^{pmcmc})} = \frac{ESS_{lwpmcmc}}{1 + EG}, \tag{4.7}$$

where EG is the efficiency gain as defined in Equation 4.3 and estimated as in Proposition 1.

The proofs of the theorem and corollary are given in the appendix.

A good kernel K will typically make $\mathrm{var}(\bar{x})$ small, while the transition rule is important in making $1+2\sum_k \gamma_k$ small. The intuition here based on the $ESS_{lwpmcmc}$ formula provides useful guidance for designing effective locally weighted parallel Markov chain Monte Carlo algorithms. The next proposition gives a simple way of estimating $ESS_{lwpmcmc}$ using all proposed samples and weights.

Proposition 2. *$ESS_{lwpmcmc}$ can be estimated by substituting $\mathrm{var}(\bar{x})(1 + 2\sum_k \gamma_k)$ with an estimate of the spectral density of $\left\{\bar{x}^{(j)}\right\}_{j=1}^{n}$ at frequency 0, and σ^2 with its moment estimator.*

As a sanity check, we compared the above estimator to an estimate of $\sigma^2/\mathrm{var}(\hat{\mu}^{lwpmcmc})$ based on repeated experiments. It does indeed estimate the correct quantity, and its coefficient of variation is similar to that of the estimator of ESS_{mcmc}. Due to the estimators' similarities, our estimator also has the same potential issues as the standard estimator. For further details see the supplementary material.

In the case of the usual Metropolis-Hastings algorithm and its multi-proposal extensions, $w(x_0^{(j)}) = 1$ and $w(x_i^{(j)}) = 0$ for $i = 1, \dots, M$. Then $\bar{x}^{(j)} = x_0^{(j)}$ and $x_0^{(j)} \sim \pi$ for all j upon convergence, so that $\text{var}(\bar{x}^{(j)}) = \text{var}(\bar{x}) = \sigma^2$ for all j. Moreover, $\gamma_k = \rho_k$ where ρ_k is the lag-k autocorrelation of the Markov chain $\{x_0^{(j)}\}_{j=1}^n$. Thus we have the following proposition.

Corollary 2. *When $M = 1$ and $w(x_0^{(j)}) = 1$, ESS_{lwmcmc} is equal to ESS_{mcmc} in formula 4.5.*

4.2.4 A Special Case: Locally Weighted Hamiltonian Monte Carlo

From the derived properties of proposed algorithm, the real benefit arises when multiple good points are proposed within each iteration, thus obtaining less degenerate weights and more effective samples. We illustrate one such algorithm, called locally weighted Hamiltonian Monte Carlo (Algorithm 3 below), where the leapfrog integration path that arises in Hamiltonian Monte Carlo algorithm (Duane et al., 1987; Roberts & Tweedie, 1996; Neal, 1994; Qin & Liu, 2001) is taken as the proposals in Algorithm 2.

Denote $\phi(\cdot)$ be the (multivariate) Gaussian density with mean zero and covariance matrix W. Let $H(x, p) = -\log \pi(x) + \frac{1}{2}p^T W^{-1} p$. Define $w(x_i^{(j)})$ as Eq (4.8) below.

$$w(x_i^{(j)}) = \frac{\exp\{-H(x_i^{(j)}, p_i^{(j)})\}}{\sum_{i=0}^{M} \exp\{-H(x_i^{(j)}, p_i^{(j)})\}}, \quad (i = 0, \dots, M). \tag{4.8}$$

Define the rejection probability from $x_a^{(j)}$ to $x_b^{(j)}$ as

$$r(x_a^{(j)}, x_b^{(j)}) = \min\left[1, \frac{\exp\{-H(x_b^{(j)}, p_b^{(j)})\}}{\exp\{-H(x_a^{(j)}, p_a^{(j)})\}}\right]. \tag{4.9}$$

Collect points $x_i^{(j)}$ with weights $w(x_i^{(j)})$ for $i = 0, \dots, M$, $j = 1, \dots, n$ according to Algorithm 3 so that μ_h could be estimated by $\hat{\mu}_h^{lwhmc} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{M} h(x_i^{(j)})w(x_i^{(j)})$.

Algorithm 3. ***Locally weighted Hamiltonian Monte Carlo algorithm***

```
Set $x$ to be the initial value
For $j = 1$ to $j = n$
    Draw a momentum vector $p_0^{(j)}, \dots, p_M^{(j)} \sim \phi$
    Sample $l$ uniformly from the set $\{0, \dots, M\}$ and set $x_l^{(j)} = x$
    Run following two steps in parallel:
        Leapfrog integrate backward in time for $l$ steps, generating $\{x_0^{(j)}, \dots, x_{l-1}^{(j)}\}$
        Leapfrog integrate forward in time for $M - l$ steps, generating $\{x_{l+1}^{(j)}, \dots, x_M^{(j)}\}$
    In parallel, calculate $w(x_0^{(j)}), \dots, w(x_M^{(j)})$ as defined in Eq (4.8)
    Set $a = 0$ if $l > M - l$ and $a = M$ otherwise
    Change $x$ to be $x_a^{(j)}$ with probability $r(x_l^{(j)}, x_a^{(j)})$ as defined in Eq (4.9)
Output $x_i^{(j)}$ with weights $w(x_i^{(j)})$ for $i = 0, \dots, M$; $j = 1, \dots, n$
```

Like other algorithms of this type, the locally weighted Hamiltonian Monte Carlo algorithm works well for highly-skewed density functions like the banana-shape density (Gelman & Meng, 1991), which significantly improves regular Markov chain Monte Carlo algorithms. The reason why we include this particular example as a special case is that the intermediate steps on the leapfrog path are good proposals in their own right, and are therefore likely to provide useful information about the target density.

The algorithm can be viewed as using the leapfrog algorithm as a numerical integrator on top of the underlying Markov chain. The efficiency of locally weighted Hamiltonian Monte Carlo therefore depends on the efficiency of this numerical integrator. For densities that exhibit certain kinds of symmetry about its mean, the leapfrog integration is an efficient numerical integrator and ESS_{lwmcmc} typically increases quadratically as a function of M. This feature does not hold for standard Hamiltonian Monte Carlo. For fixed M the performance of the leapfrog algorithm as a numerical integrator depends on the tuning parameters.

Similarly, we can modify other Markov chain Monte Carlo algorithms to their locally-weighted parallelisable versions, e.g. a locally-weighted parallelisable version of Multiple-try Markov chain Monte Carlo (Liu et al., 2000). We omit the details here.

4.3 Numerical Example

In this section, we apply the quantities defined in Section 4.2 to compare the performance of the algorithms discussed in this paper using a time series example. We use upper case to denote random variables and the corresponding lower case to denote their realizations. Let $X_0 = 0$ be fixed, the observations $y_{1:T} = (y_1, \ldots, y_T)^T$ are obtained through the following process:

$$Y_t = X_t + \epsilon_t, \quad X_t = X_{t-1} + \eta_t \quad (t = 1, \ldots, T); \tag{4.10}$$

where ϵ_t, η_t are independently Gaussian with mean zero and variances σ_y^2 and σ_x^2 respectively. The joint likelihood of the observations $y_{1:T}$ and hidden process $x_{1:T} =$

$(x_1, \ldots, x_T)^T$ given the parameters (σ_x^2, σ_y^2) is

$$L(y_{1:T}, x_{1:T} \mid \sigma_x^2, \sigma_y^2) = (\sigma_y \sigma_x)^{-T} \exp\left\{-\frac{\sum_{t=1}^T (y_t - x_t)^2}{2\sigma_y^2} - \frac{\sum_{t=1}^T (x_t - x_{t-1})^2}{2\sigma_x^2}\right\}.$$

We can calculate the marginal likelihood $L(y_{1:T} \mid \sigma_x^2, \sigma_y^2) = \int_{\mathbb{R}^T} L(y_{1:T}, x_{1:T} \mid \sigma_x^2, \sigma_y^2) dx_{1:T}$ using the Kalman filter (Kalman, 1960; Kalman & Bucy, 1961). We assume independent scaled inverse-chi-squared priors for σ_x^2 and σ_y^2, with hyper-parameters (ν_x, s_x^2) and (ν_y, s_y^2) respectively. We apply Metropolis-Hastings, parallel and locally weighted parallel Markov chain Monte Carlo, and locally weighted Hamiltonian Monte Carlo algorithms to obtain posterior samples of σ_x^2 and σ_y^2. In the simulations, we choose $T = 1,000$ as the length of observations. The model parameters are chosen as $\sigma_x = 0.1$, $\sigma_y = 1$. The hyper-parameters are $\nu_x = \nu_y = 1$, $s_x^2 = 0.01$, $s_y^2 = 1$. When implementing Metropolis-Hastings, parallel and locally weighted parallel Markov chain Monte Carlo, we choose independent Gaussian proposals centered at the current point with standard deviations δ_x, δ_y for $\log(\sigma_x^2), \log(\sigma_y^2)$ respectively. The burn-in is chosen to be the first $10,000$ iterations and $n = 10,000$ posterior samples are collected.

Table 4.1 summarizes the results with effective sample size from each algorithm. M is the number of proposals at each iteration, N is the number of samples collected at each iteration.

Furthermore, we implemented the locally weighted Hamiltonian Monte Carlo algorithm with $M = 20$ proposals at each iteration. We chose the leapfrog step size to be 0.03 and W matrix to be diagonal with diagonal elements $(2, 1)$. The effective sample size for σ_x^2 and σ_y^2 are 454 and $4,793$ respectively; which are much larger than that of the Metropolis-Hastings and the parallel Markov chain Monte Carlo algorithms with $M = 20$ proposals at each iteration as listed in Table 4.1.

Effective sample size for σ_x^2

	MH	pMCMC		LWPMCMC		
M	1	20	100	1	20	100
N	1	5	5	NA	NA	NA
$\delta_x = 0.1$	74	14	7	76	14	7
$\delta_x = 0.3$	173	82	45	185	82	45
$\delta_x = 0.5$	208	208	74	210	204	74

Effective sample size for σ_y^2

	MH	pMCMC		LWPMCMC		
M	1	20	100	1	20	100
N	1	5	5	NA	NA	NA
$\delta_y = 0.1$	1227	707	187	1265	727	188
$\delta_y = 0.3$	1023	1015	1345	1000	1127	1413
$\delta_y = 0.5$	489	1695	1728	500	1885	1898

Table 4.1: Effective sample size results for the linear time series example using the Metropolis-Hastings (MH), parallel Markov chain Monte Carlo (pMCMC) and locally weighted parallel Markov chain Monte Carlo (LWPMCMC) algorithms with $n = 10,000$ samples. M is the number of proposals at each iteration, N is the number of samples collected at each iteration. δ_x and δ_y are the step-sizes of the symmetric random walk proposals of $\log(\sigma_x^2)$ and $\log(\sigma_y^2)$ respectively.

From the simulation above, we can see that parallelisable waste-recycling Markov chain Monte Carlo algorithms indeed give bigger effective sample size, i.e. more efficient posterior estimators as compared to the Metropolis-Hastings algorithm and the parallel Markov chain Monte Carlo algorithm, but at a cost of more computation as opposed to the Metropolis-Hastings algorithm. Note that the locally weighted parallelisable Markov chain Monte Carlo has the same computation cost as compared to the parallel Markov chain Monte Carlo algorithm but with larger effective sample size, thus the former is superior to the latter considering both the statistical and computational efficiency. A good transition kernel is crucial for performance, as can be seen from the results of locally weighted Hamiltonian Monte Carlo. Our effective sample size measurements provide quantitative guidance for the trade-off between the statistical efficiency and computational efficiency, and the choice of proper transition kernels for the most efficient algorithm.

4.4 Conclusions

In this paper we have developed the locally weighted parallelisable Markov chain Monte Carlo algorithm, built upon waste-recycling Markov chain Monte Carlo and parallel Markov chain Monte Carlo algorithms, which dominates its parallel Markov chain Monte Carlo counterpart theoretically. We show how to compute the effective sample size of the algorithms' output and illustrate their performance on a toy example. The locally weighted algorithm is well suited to modern computer architectures with massive numbers of cores, which can possibly dramatically increase statistical efficiency with parallel computing.

The algorithm provides a general framework with highly flexible choice of proposing kernels and transition kernels, thus readily adaptable for any Markov chain Monte Carlo algorithms with cleverly chosen proposing or transition kernels. We have demonstrated this with the Hamiltonian Monte Carlo algorithm in this paper, in which the results are promising.

Furthermore, we promote the usage of locally weighted parallelisable Markov chain Monte Carlo algorithms over unweighted parallel Markov chain Monte Carlo algorithms because it provides easily-estimable quantities that guides practitioners to balance the trade-off between the statistical efficiency and computational cost for parallel computing.

5
Conclusion

In summary, this thesis (i) contributes a general statistical methodology to analyzing experimental replicates from single-molecule experiments based on fluorescence imaging, (ii) proposes a consistent and easily computable estimator for the order of hidden Markov models, and (iii) defines a measure that evaluates the effectiveness of a wide family of parallelisable Markov chain Monte Carlo algorithms.

5.1 Open Questions about Single-Molecule Data

There are several remaining interesting questions about the analysis of single-molecule fluorscence imaging data. Here we briefly summarize some of them.

1. The current analysis of single-molecule data is based on FRET, which is defined as the acceptor signal divided by the sum of the acceptor signal and donor signal. Rigorous study of the advantages and disadvantages of using the FRET versus analyzing the donor signal and acceptor signal separately is desired.

2. The single-molecule experiments shown in chapter 1 obtain measurements at a 30 millisecond frequency and the total observational time roughly ranges from 10 seconds to 90 seconds. Given that the observational time is limited, what is the 'optimal' frequency of taking the observations? This is a question asked from the perspective of the experimental design – we aim for preserving necessary information with minimum labor or experimental data.

3. Is 30 millisecond good enough to capture the stochastic dynamics of the molecules? In other words, how can we quantify the missed fast transitions?

4. The biological process is known to be a continuous time stochastic process, although the experimental data is taken at discrete time points. What kinds of conclusions about the continuous process can be drawn from the discrete time observations remain open. It will be interesting to also consider embedding the discrete time observations into a continuous time stochastic process.

5. The experimental replicates maybe correlated with each other, a natural consequence of the experimental procedures. How to detect or measure this corre-

lation among molecules and how to use the correlations to obtain more robust inference are intriguing questions for both statisticians and biologists.

5.2 Open Questions about Parallel MCMC

Besides the parallelisable MCMC scheme discussed in this thesis, more innovative and efficient sampling algorithms that utilize parallel computation are desirable.

1. The current paralelisable framework requires frequent communications among different workers, i.e. collecting all the *weights* at each iteration before moving forward, which significantly compromises the computational gain from parallelisation. Generalizations of the multiple-proposal scheme to the asynchronous framework, in which workers do not wait for each other before moving forward with the next iteration, will be an intriguing problem to pursue. The literature on asynchronous optimization algorithms and the asynchronous Gibbs sampling provide important hints towards solving this problem.

2. We propose to evaluate the efficiency of parallelisable MCMC algorithms using the generalized effective sample size, which, under the current definition, does not take the correlations of different dimensions into account. In other words, right now we calculate the effective sample size one dimension at a time. It would be more reasonable to consider the correlations, which is available through the samples, when quantifying the efficiency of sampling.

3. In addition to sampling efficiency, the mixing rate is an important factor in evaluating such algorithms. It would be ideal if we can provide in-depth theoretical studies of the mixing rates of parallelisable MCMC algorithms.

A

Computational Algorithms of HMM

A.1 Baum-Welch/EM Algorithm for HMM

For a given value of K, the total number of states, we can use the EM algorithm (Dempster et al., 1977), a.k.a. the Baum-Welch algorithm for HMM (Baum & Petrie, 1966; Baum et al., 1970), to infer $\boldsymbol{\theta}$. For the ease of presentation, we assume here that the initial distribution of the first hidden state z_1 is flat. The full likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{n=2}^{N} p(z_n|z_{n-1}, \boldsymbol{P}) \prod_{n=1}^{N} p(y_n|z_n, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \prod_{j,k=1}^{K} P_{jk}^{T_{jk}} \cdot \prod_{n=1}^{N} \mathcal{N}(y_n; \mu_{z_n}, \sigma_{z_n}^2),$$

where T_{jk} denotes the total number of transitions in $\boldsymbol{z}$ from state j to state k, and $\mathcal{N}(y; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2 evaluated at y. For the EM algorithm, in the E-step, the expectation step, we have

$$E \log L(\boldsymbol{\theta}|\boldsymbol{\theta}^{old}) = \sum_{j,k=1}^{K} \sum_{n=2}^{N} v_{n,j,k} \log P_{jk} + \sum_{k=1}^{K} \sum_{n=1}^{N} u_{n,k} \log \mathcal{N}(y_n; \mu_k, \sigma_k^2),$$

where $u_{n,k} = p(z_n = k|\boldsymbol{y}, \boldsymbol{\theta}^{old})$ and $v_{n,j,k} = p(z_{n-1} = j, z_n = k|\boldsymbol{y}, \boldsymbol{\theta}^{old})$ can be expressed in terms of $\alpha(z_n) := p(\boldsymbol{y}_{1:n}, z_n|\boldsymbol{\theta}^{old})$ and $\beta(z_n) := p(\boldsymbol{y}_{(n+1):N}|z_n, \boldsymbol{\theta}^{old})$:

$$\begin{aligned} u_{n,z_n} &= \alpha(z_n)\beta(z_n)/p(\boldsymbol{y}_{1:N}|\boldsymbol{\theta}^{old}), \\ v_{n,z_{n-1},z_n} &= \alpha(z_{n-1})\beta(z_n)p(y_n|z_n, \boldsymbol{\theta}^{old})p(z_n|z_{n-1}, \boldsymbol{\theta}^{old})/p(\boldsymbol{y}_{1:N}|\boldsymbol{\theta}^{old}). \end{aligned}$$

$\alpha(z_n)$ and $\beta(z_n)$ can be efficiently calculated by the forward-backward algorithm (Rabiner, 1989), a recursive formula that allows fast computation: evaluating the α's forwardly from 1 to N and the β's backwardly from N to 1:

$$\alpha(z_n) = p(y_n|z_n, \boldsymbol{\theta}^{old}) \sum_{z_{n-1}=1}^{K} \alpha(z_{n-1})p(z_n|z_{n-1}, \boldsymbol{\theta}^{old}), \tag{A.1}$$

$$\beta(z_n) = \sum_{z_{n+1}=1}^{K} \beta(z_{n+1})p(y_{n+1}|z_{n+1}, \boldsymbol{\theta}^{old})p(z_{n+1}|z_n, \boldsymbol{\theta}^{old}), \quad \beta(z_N) \equiv 1. \tag{A.2}$$

In addition, the forward-backward algorithm gives the marginal likelihood evaluated at the maximum likelihood estimate $p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) = \sum_{z_N} \alpha(z_N) = \sum_{z_N} p(\boldsymbol{y}_{1:N}, z_N|\hat{\boldsymbol{\theta}})$.

In the M-step of the EM algorithm, which maximizes $E \log L(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ over $\boldsymbol{\theta}$, we obtain $\boldsymbol{\theta}^{new}$ according to

$$P_{jk} \quad = \quad \frac{\sum_{n=2}^{N} v_{n,j,k}}{\sum_{k=1}^{K}\sum_{n=2}^{N} v_{n,j,k}}, \ \mu_k = \frac{\sum_{n=1}^{N} y_n u_{n,k}}{\sum_{n=1}^{N} u_{n,k}}, \ \sigma_k^2 = \frac{\sum_{n=1}^{N} u_{n,k}(y_n - \mu_k)^2}{\sum_{n=1}^{N} u_{n,k}}.$$

A.2 Gibbs Sampling for HMM

In addition to the EM algorithm, which quickly obtains the MLE of the parameters, we can also use Bayesian MCMC sampling (Liu, 2001) to assess the entire (posterior) distribution of the parameters. Our MCMC sampling can be viewed as a special case of data augmentation (Tanner & Wong, 1987): augment the parameter space $\boldsymbol{\theta}$ with the hidden states $\boldsymbol{z}$, and iteratively sample one given the other (i.e., sample $\boldsymbol{\theta}$ given $\boldsymbol{z}$ and sample $\boldsymbol{z}$ given $\boldsymbol{\theta}$).

Specifically, in our MCMC sampling, we adopt flat priors for $\boldsymbol{P}$ and μ_k, $k = 1, \ldots, K$, and independent inverse-χ^2 priors with parameters ν, s^2 for σ_k^2 (the prior on $\boldsymbol{\mu}$ is flat over the region $0 < \mu_1 < \cdots < \mu_K < 1$). The posterior distribution is

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) &= p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) p_0(\boldsymbol{P}) p_0(\boldsymbol{\mu}) p_0(\boldsymbol{\sigma^2}) \\ &\propto \prod_{j=1}^{K}\prod_{k=1}^{K} P_{jk}^{T_{jk}} \prod_{n=1}^{N} \mathcal{N}(y_n; \mu_{z_n}, \sigma_{z_n}^2) \prod_{k=1}^{K} p_0(\sigma_k^2; \nu, s^2). \end{aligned}$$

It follows that in our (group Gibbs) sampler, the conditional distribution of the jth row of the transition matrix $P_{j\cdot} = (P_{j1}, P_{j2}, \ldots, P_{jK})$ is a Dirichlet distribution, the conditional distribution of $\boldsymbol{\mu}$ is a multivariate normal distribution, the conditional distribution of $\boldsymbol{\sigma^2}$ is a multivariate inverse-χ^2 distribution and that the hidden states

$\boldsymbol{z}$ can be sampled sequentially from 1 to N through the following recursion:

$$\begin{aligned} p(z_n = k | z_{n-1} = j, \boldsymbol{\theta}, \boldsymbol{y}) &\propto P_{jk}\ \mathcal{N}(y_n; \mu_k, \sigma_k)\ p(\boldsymbol{y}_{n+1:N} | z_n = k) \\ &= P_{jk}\ \mathcal{N}(y_n; \mu_k, \sigma_k)\ \beta(k), \quad n = 1, 2, \ldots, N, \end{aligned}$$

where $\beta(k)$ is the the backward probability defined in equation (A.2).

A.3 MCMC Sampling of the Hierarchical HMM

The posterior distribution is proportional to

$$p(\boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2, \boldsymbol{s}^2) \prod_l p(\boldsymbol{y}^{(l)}, \boldsymbol{z}^{(l)} | I^{(l)}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}, \boldsymbol{P}) \times \prod_l p(\boldsymbol{\mu}^{(l)} | \boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2, I^{(l)}) p((\boldsymbol{\sigma}^{(l)})^2 | \boldsymbol{\nu}, \boldsymbol{s}^2, I^{(l)}) p(I^{(l)}).$$

We use the Gibbs sampler to update a group of parameters at a time, conditioning on the others, and iterate until convergence. The sampling details are given below, where $I(\omega)$ and I_ω denote the indicator function.

1. Initialization. Fit each trajectory independently using the EM algorithm in Appendix A.1 and set the initial values of $\{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}\}$ at the corresponding MLEs. The initial values of $\{I^{(l)}\}$ are set to be $\{1, \ldots, K\}$.

2. Update global parameters $\boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2, \boldsymbol{s}^2$. For $1 \le k \le K$,

$$\begin{array}{lll} \text{Sample} & \mu_{0,k} & \text{from} \quad \mathcal{N}(\sum_{l=1,k\in I^{(l)}}^T \mu_k^{(l)} / (\sum_{l=1}^T I_{k\in I^{(l)}}), \eta_{0,k}^2 / (\sum_{l=1}^T I_{k\in I^{(l)}})), \\ \text{Sample} & \eta_{0,k}^2 & \text{from} \quad \text{Inv-}\chi^2(\sum_{l=1}^T I_{k\in I^{(l)}} - 2, \sum_{l=1,k\in I^{(l)}}^T (\mu_k - \mu_{0,k})^2 / (\sum_{l=1}^T I_{k\in I^{(l)}} - 2)), \\ \text{Sample} & s_k^2 & \text{from} \quad \{\nu_k \sum_{l=1}^T I_{k\in I^{(l)}} / (\sigma_k^{(l)})^2\}^{-1} \chi_{df}^2,\ df = \nu_k \sum_{l=1}^T I_{k\in I^{(l)}} + 2. \end{array}$$

3. Update transition probabilities $\boldsymbol{P}$ according to

$$p(\boldsymbol{P}) \propto \prod_{i,j} P_{ij}^{\sum_l N_{i,j}^{(l)}} / \prod_{I^{(l)} \neq \{1,2,\ldots,K\}} \prod_{i \in I^{(l)}} (\sum_{k \in I^{(l)}} P_{ik})^{\sum_{k \in I^{(l)}} N_{i,k}^{(l)}}.$$

4. Update parameters for individual trajectories.

    - Update $\{\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}\}$. For $k \in I^{(l)}$, $l = 1, \ldots, T$,

$$\begin{aligned} \mu_k^{(l)} &\sim \mathcal{N}\left(\frac{\mu_{0k}/\eta_{0k}^2 + \sum_{z_n^{(l)}=k} y_n^{(l)}/(\sigma_k^{(l)})^2}{1/\eta_{0k}^2 + \sum_{z_n^{(l)}=k} 1/(\sigma_k^{(l)})^2}, \frac{1}{1/\eta_{0k}^2 + \sum_{z_n^{(l)}=k} 1/(\sigma_k^{(l)})^2} \right); \\ (\sigma_k^{(l)})^2 &\sim Inv-\chi^2 \left(\nu_k + \sum_{n=1}^{N_l} I(z_n^{(l)} = k), \frac{\nu_k s_k^2 + \sum_n (y_n^{(l)} - \mu_k^{(l)})^2 I(z_n^{(l)} = k)}{\nu_k + \sum_n I(z_n^{(l)} = k)} \right). \end{aligned}$$

    - Update $\{z^{(l)}\}$. This is essentially the same as introduced in Appendix A.2 except that when $I^{(l)} \neq \{1, 2, \ldots, K\}$, the transition matrix is a re-normalized submatrix of $\boldsymbol{P}$ according to which states are present in trajectory l.

    - Update $\{I^{(l)}\}$. $I^{(l)}$ is equal to $A \subset \{1, 2, \ldots, K\}$ with probability proportional to

$$p(\boldsymbol{y}^{(l)}|\boldsymbol{\mu}^{(l)}, \boldsymbol{\sigma}^{(l)}, \boldsymbol{P}, I^{(l)} = A) p(\boldsymbol{\mu}^{(l)}|\boldsymbol{\mu}_0, \boldsymbol{\eta}_0^2, I^{(l)} = A) p((\boldsymbol{\sigma}^{(l)})^2|\boldsymbol{\nu}, \boldsymbol{s}^2, I^{(l)} = A)$$

    where A stands for $\{1, 2, 3\}$, $\{1, 2\}$, $\{1, 3\}$ or $\{2, 3\}$ when $K = 3$, and $\{1, 2, 3, 4\}$, $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{1, 3, 4\}$, $\{2, 3, 4\}$, $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, or $\{3, 4\}$ when $K = 4$.

5. Iterate Steps 2 to 4 until convergence.

B

Theoretical Results of HMM Order Selection

B.1 Proofs of Consistency Theorems

We use the same notations as Section 3.1.2. Throughout the proof, we use 1_Ω to denote the indicator function for any set Ω, and we use Ω^c to denote the complement of Ω. In addition, for a vector or matrix A, let A^t be its transpose. We use $||\cdot||$ to denote the Euclidean norm.

B.1.1 Regularity Conditions

There are three groups of regularity conditions: (A1) - (A6), which ensure the asymptotic properties of the posterior distributions under the true number of hidden states K^* (de Gunst & Shcherbakova, 2008), (B1) - (B2), which regulate the asymptotic properties of the posterior distribution when the number of states K is larger than the true number of states K^* (Gassiat & Rousseau, 2014), and (C1) - (C2), which enables representing the full log likelihood of an HMM as an additive functional of an ergodic Markov chain (Fuh, 2003).

Regularity Conditions for Asymptotic 'True' Posterior

(A1) Θ is a compact set in $\mathbb{R}^d$, and the true parameter θ_k^* is an interior point of Θ for all $1 \leq k \leq K^*$.

(A2) $q_{kl}^* > 0$ for all $1 \leq k, l \leq K^*$.

(A3) The function $\theta \to f(\cdot|\theta)$ is twice continuously differentiable in Θ.

(A4) Let P^* and E^* denote the probability and expectation under the true probability model ϕ^* respectively. For all $1 \leq k \leq K^*$,

$$E^*|\log f(Y_1|\theta_k^*)| < \infty; \tag{B.1}$$

furthermore, there exists $\epsilon > 0$ such that

$$P^* \left\{ \sup_{||\boldsymbol{\theta'} - \boldsymbol{\theta}^*|| < \epsilon,\ 1 \leq l, l' \leq K^*} \frac{f(Y_1|\theta_l)}{f(Y_1|\theta_{l'})} = \infty \middle| X_1 = k \right\} < 1. \tag{B.2}$$

(A5) Let ν denote the measure on $\mathcal{Y}$ that we defined the density $f(\cdot|\theta)$ with respect to. Then, for any $\theta \in \Theta$, we suppose that there exists $\epsilon > 0$ such that

$$E^* \left[\sup_{|\theta' - \theta| < \epsilon} (\log f(Y_1|\theta'))^+ \right] < \infty, \tag{B.3}$$

$$E^* \left[\sup_{|\theta' - \theta| < \epsilon} ||\nabla_\theta \log f(Y_1|\theta')||^2 \right] < \infty, \tag{B.4}$$

$$E^* \left[\sup_{|\theta' - \theta| < \epsilon} ||D_\theta^2 \log f(Y_1|\theta')||^2 \right] < \infty, \tag{B.5}$$

$$\int_{\mathcal{Y}} \left\| \sup_{|\theta' - \theta| < \epsilon} \nabla_\theta f(y|\theta') \right\| \nu(dy) < \infty, \tag{B.6}$$

$$\int_{\mathcal{Y}} \left\| \sup_{|\theta' - \theta| < \epsilon} D_\theta^2 f(y|\theta') \right\| \nu(dy) < \infty. \tag{B.7}$$

In addition, there exists $a > 0$ and $b > 0$ such that, for any sufficiently large M,

$$\sup_{|\theta| \leq M^b} \int ||\nabla_\theta f(y|\theta)|| \nu(dy) \leq M^a. \tag{B.8}$$

(A6) For any two $\theta \neq \theta'$ in Θ, $\lambda^d\{y : f(y|\theta) \neq f(y|\theta')\} > 0$, where λ^d is the Lebesgue measure of $\mathbb{R}^d$.

Regularity Conditions for Asymptotic Overly-Fitted Posterior

The first condition (B1) regulates the prior for the transition matrix.

(B1) The prior distribution on the transition matrix can be written as $\nu_K(Q_K) = \prod_{k=1}^{K} \tilde{\nu}(q_{k1}, q_{k2}, \cdots, q_{kK})$. Moreover, there exists $C > 0$, $\alpha_1 > 0$, $\cdots$, $\alpha_K > 0$ such that

$$0 < \tilde{\nu}(u_1, u_2, \cdots, u_K) < C u_1^{\alpha_1 - 1} u_2^{\alpha_2 - 1} \cdots u_K^{\alpha_K - 1}. \tag{B.9}$$

for all $(u_1, u_2, \cdots, u_K)$ satisfying

$$\min(u_1, u_2, \cdots, u_K) > 0, \text{ and } \sum_{l=1}^{K} u_k = 1.$$

Furthermore, suppose that $\sum_{k=1}^{K} \alpha_k > K(K + d - 1)$.

The second condition (B2) ensures the *identifiability* under overestimation, which needs some additional definitions. Define

$$\tilde{K} = \left\{ \vec{k} = (k_1, k_2, \cdots, k_{K^*}) \in \{1, \cdots, K\}^{K^*} : k_i < k_{i+1}, i = 0, \cdots, K^* - 1 \right\}.$$

For all $1 \leq i \leq K^* + 1$ with $k_0 = 0$ and $k_{K^*+1} = K$, set

$$I_i = \{k_{i-1} + 1, \cdots, k_i\}.$$

For any $\vec{k} \in \tilde{K}$, consider

1. $a_i \in \mathbb{R}$, $b_i \in \mathbb{R}^d$, and $c_i \in \mathbb{R}$ for $1 \leq i \leq K^*$.

2. $\pi_k \in \{\mathbb{R} \cup \{0\}\}$ for $k_{K^*} + 1 \leq k \leq K$,

3. $z_{ik} \in \mathbb{R}^d$ and $\alpha_{ik} \in \mathbb{R}$ for $1 \leq i \leq K^*$ and $k \in I_i$ such that $||z_{ik}|| = 1$, $\alpha_{ik} \geq 0$ and $\sum_{k \in I_j} \alpha_{ik} = 1$ for all $1 \leq j \leq K^*$,

4. $\theta_k \in \Theta - \{\theta_l^*, 1 \leq l \leq K^*\}$ for $k_{K^*} + 1 \leq k \leq K$.

Let $\mathcal{P}$ be the space of all possible $(a_i, b_i, c_i, \pi_k, z_{ik}, \alpha_{ik}, \theta_k)$. For any $p \in \mathcal{P}$, define

$$\mathcal{F}(y|p) = \mathcal{F}(y|a_i, b_i, c_i, \pi_k, z_{ik}, \alpha_{ik}, \theta_k)$$
$$= \sum_{k=k_{K^*}+1}^{K} \pi_k f(\cdot|\theta_k) + \sum_{i=1}^{K^*} \left(a_i f(\cdot|\theta_i^*) + b_i^t \nabla f(\cdot|\theta_i^*)\right) + \sum_{i=1}^{K^*} c_i^2 \sum_{k \in I_i} \alpha_{ik} z_{ik}^t D^2 f(\cdot|\theta_k^*) z_{ik}.$$

(B2) The set $\{p : p \in \mathcal{P}, \mathcal{F}(y|p) = 0 \ \ \forall y \in \mathcal{Y}\}$ is equal to

$$\Big\{p : p \in \mathcal{P}, (a_i, b_i, c_i) = \vec{0} \ \ \forall 1 \leq i \leq K^*, \pi_k = 0 \ \ \forall k_{K^*} + 1 \leq k \leq K\Big\}.$$

Remark 4. *The prior condition (B1) is designed for the Dirichlet prior. The weak identifiability condition (B2) holds for any mixtures of regular exponential families, see Gassiat & Rousseau (2014).*

Remark 5. *Condition (B2) needs some explanation. As discussed in Section 3.3.3, the problem is the existence of multiple "true" parameters. More precisely, it can be shown that, for all $\phi_K \in \Phi_K$ such that $p(\cdot|\phi_K) = p(\cdot|\phi_{K^*})$, there exists a $\vec{k} \in \tilde{K}$ such that, up to a permutation of the state labels:*

- *$\theta_k = \theta_i^*$ for all $i \in \{1, 2, \cdots, K^*\}$ and $k \in I_i$.*
- *$\sum_{l \in I_j} q_{kl} = q_{ij}^*$ for all $i, j \in \{1, 2, \cdots, K^*\}$, $k \in I_i$ and $l \in I_j$.*
- *$q_{kl} = 0$ for all $1 \leq k \leq K$ and $l > k_{K^*}$.*

In other words, ϕ_K is obtained by duplicating state i into states $k_{i-1}+1,\cdots,k_i$, with some redundant states $k_{K^*}+1,\cdots,K^*$ which are impossible to reach. For a fixed $\vec{k} \in \tilde{K}$, for any $\phi_k \in \Phi_K$, consider the following parameterization, up to a permutation,

$$\begin{aligned}
\vec{\theta}_D &= (\theta_1,\cdots,\theta_{k_{K^*}}), \\
\vec{\theta}_R &= (\theta_{k_{K^*}+1},\cdots,\theta_K), \\
\Delta_{jk} &= \sum_{k'\in I_j} q_{kk'} - q^*_{ij}, && 1 \le j \le K^*, k \in I_i, 1 \le i \le K^* \\
r_{kl} &= \frac{q_{kl}}{\sum_{k'\in I_j} q_{kk'}}, && 1 \le k \le K, l \in I_j, j \le K^* \\
\vec{Q}_R &= (q_{kl})_{1\le k\le K, l > k_{K^*}}\ .
\end{aligned}$$

Note that the true parameter ϕ^* corresponds to

$$\vec{\theta}^*_D = (\theta^*_1,\cdots,\theta^*_1,\theta^*_2,\cdots,\theta^*_2,\cdots,\theta^*_{K^*}), \tag{B.10}$$

$$\Delta^*_{jk} = 0,\ 1 \le j \le K^*, k \in I_i, 1 \le i \le K^*, \tag{B.11}$$

$$\vec{Q}^*_R = \vec{0}, \tag{B.12}$$

where in (B.10), θ_i is repeated $k_i - k_{i-1}$ times; and arbitrary r_{kl} and $\vec{\theta}_R$ as long as $\sum_{l\in I_j} r_{kl} = 1$ for all $1 \le j \le K^*$. This means that we would like to have an "identifiability" in which ϕ_K is considered to be the same as ϕ_{K^*} when (B.10)-(B.12) are satisfied. The condition (B2) is designed to take care of this: $a_i = c_i = 0$ corresponds to $\Delta_{jk} = 0$, $b_i = \vec{0}$ corresponds to $\vec{\theta}_D = \vec{\theta}^*_D$, and $p_i = 0$ corresponds to $\vec{Q}_R = \vec{0}$. The (z_{ik}, α_{ik}) and θ_k correspond to r_{kl} and $\vec{\theta}_R$, thus no regularity condition is required. See Gassiat & Rousseau (2014) for more detailed discussions about the (B2) condition.

Regularity Conditions for Additive Representation of HMM

Let F_i be a $K \times K$ diagonal matrix with diagonal elements $(f(y_i|\boldsymbol{\theta}_1), \cdots, f(y_i|\boldsymbol{\theta}_K))$ and set $M_i = M_i(\phi_K) = F_i Q_K^t$, $1 \leq i \leq n$.

(C1) The set $\{\phi : \phi \in \Phi_K, M_i(\phi)$ is invertible P^{ϕ}-almost surely$\}$ has probability one under the prior, where P^{ϕ} denotes the probability measure determined by the parameter ϕ.

(C2) For any $\phi \in \Phi_K$ and $1 \leq k \leq K$, $E^{\phi}[|Y_1||X_1 = k] < \infty$, where E^{ϕ} denotes the expectation under P^{ϕ}.

B.1.2 Proof of Theorem 2

The proof of Theorem 2 requires Lemmas 1 and 2, which studies the asymptotic behavior of the log-likelihood and the modified (path-ignored) log-likelihood respectively. The proofs of Lemmas 1 and 2 are given in Sections B.1.2 and B.1.2.

Define $L(\mathbf{y}_{1:n}|\phi_K) := \log p(\mathbf{y}_{1:n}|\phi_K)$. For the set of paths $\mathcal{X}^n_{K,\epsilon}$ in (3.6), define $L_\epsilon(\mathbf{y}_{1:n}|\phi_K) = \log p_\epsilon(\mathbf{y}_{1:n}|\phi_K)$, where

$$p_\epsilon(\mathbf{y}_{1:n}|\phi_K) := \sum_{\mathbf{x}_{0:n} \in \mathcal{X}^n_{K,\epsilon}} p(\mathbf{y}_{1:n}, \mathbf{x}_{0:n}|\phi_K).$$

Lemma 1. *If (A1)-(A6) and (C1)-(C2) hold, then for any $\boldsymbol{\phi}_K$, there exists $\mu = \mu(\boldsymbol{\phi}_K) \in \mathbb{R}$ such that $n^{-1} L(\mathbf{y}_{1:n}|\phi_K) \to \mu$ almost surely.*

Lemma 2. *If (A1)-(A6) and (C1)-(C2) hold, then for any fixed $\omega \in (0,1)$: for any $\epsilon \in (0,1)$ and any $\boldsymbol{\phi}_K \in \Phi_K$,* *(i) there exists $\mu' = \mu'(\boldsymbol{\phi}_K) \in \mathbb{R}$ such that with probability one,*

$$\limsup_{n\to\infty} n^{-1} L_\epsilon(\boldsymbol{y}_{1:n}|\phi_K) \leq \mu' - \epsilon \log \omega.$$

(ii) $\mu'(\boldsymbol{\phi}_K) < \mu(\boldsymbol{\phi}_K)$, where $\mu(\boldsymbol{\phi}_K)$ is defined in Lemma 1.

We now present the proof of Theorem 2.

Proof. Fix $\omega \in (0,1)$. By Lemma 1, there exists a function $\mu : \Phi_K \to \mathbf{R}$ such that $n^{-1}L(\boldsymbol{y}_{1:n}|\phi) \xrightarrow{a.s.} \mu(\phi)$. By Lemma 2 (i), there exists a function $\mu' : \Phi_K \to \mathbf{R}$ such that $\limsup_{n\to\infty} n^{-1} L_\epsilon(\boldsymbol{y}_{1:n}|\phi) \leq \mu'(\phi) - \epsilon \log \omega$ almost surely. Lemma 2 (ii) shows that μ' is strictly smaller than μ on Φ_K. Thus for each $\phi \in \Phi_K$, there exists a neighborhood $\mathcal{B} = \mathcal{B}(\phi) \subset \Phi_K$ such that $\mu'(\phi) < \inf_{\varphi \in \mathcal{B}} \mu(\varphi) - c$ for some $c > 0$. Therefore, we can choose $\epsilon = \epsilon(\phi) > 0$ such that $\mu'(\phi) - \epsilon(\phi) \log \omega < \inf_{\varphi\in\mathcal{B}} \mu(\varphi) - \frac{c}{2}$. As a result, for any $\phi \in \Phi_K$ and any $\varphi \in \mathcal{B} = \mathcal{B}(\phi)$, with probability one,

$$\limsup_{n\to\infty} \frac{1}{n} L_{\epsilon(\phi)}(\boldsymbol{y}_{1:n}|\phi) - \frac{1}{n} L(\boldsymbol{y}_{1:n}|\varphi) \leq \mu'(\phi) - \epsilon(\phi)\log\omega - \mu(\varphi) < -\frac{c}{2} < 0.$$

This shows that, for any $\phi \in \Phi_K$, with probability one as $n \to \infty$,

$$\frac{p_{\epsilon(\phi)}(\boldsymbol{y}_{1:n}|\phi)}{\int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\varphi) p_0(\varphi) d\varphi} \leq \frac{p_{\epsilon(\phi)}(\boldsymbol{y}_{1:n}|\phi)}{\int_{\mathcal{B}(\phi)} p(\boldsymbol{y}_{1:n}|\varphi) p_0(\varphi) d\varphi} \to 0. \tag{B.13}$$

For any $\epsilon_n \downarrow 0$ as $n \to \infty$, let $\mathcal{E}_n := \{\phi : \phi \in \Phi_K, \epsilon(\phi) \geq \epsilon_n\}$. Then,

$$\begin{aligned}
&\frac{\int_{\Phi_K} p_{\epsilon_n}(\boldsymbol{y}_{1:n}|\phi)p_0(\phi)d\phi}{\int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\varphi)p_0(\varphi)d\varphi} \\
&= \int_{\mathcal{E}_n} \frac{p_{\epsilon_n}(\boldsymbol{y}_{1:n}|\phi)}{\int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\varphi)p_0(\varphi)d\varphi} p_0(\phi)d\phi + \int_{\mathcal{E}_n^c} \frac{p_{\epsilon_n}(\boldsymbol{y}_{1:n}|\phi)}{\int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\varphi)p_0(\varphi)d\varphi} p_0(\phi)d\phi \\
&\leq \int_{\mathcal{E}_n} \frac{p_{\epsilon(\phi)}(\boldsymbol{y}_{1:n}|\phi)}{\int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\varphi)p_0(\varphi)d\varphi} p_0(\phi)d\phi + \int_{\Phi_K} \frac{p_{\epsilon_n}(\boldsymbol{y}_{1:n}|\phi)}{\int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\varphi)p_0(\varphi)d\varphi} 1_{\mathcal{E}_n^c}(\phi)p_0(\phi)d\phi.
\end{aligned}$$

Since (1) both integrands are bounded above by $\frac{p(\boldsymbol{y}_{1:n}|\phi)p_0(\phi)}{\int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\varphi)p_0(\varphi)d\varphi}$, which integrates to 1 over Φ_K, (2) the first integrand goes to zero pointwise due to (B.13), and (3) the second integrand goes to zero pointwise since $\mathcal{E}_n \uparrow \Phi_K$ gives $1_{\mathcal{E}_n^c} \downarrow 0$; by dominated convergence theorem, both terms are $o(1)$ with probability 1. Thus equation (3.7) holds. General cases can be proven similarly by considering an additional indicator function $1_{\phi_K \in \mathcal{A}_n}$. □

Proof of Lemma 1

Proof. The detailed proof is in Fuh (2003). We briefly present some of the relevant concepts and notations. Denote $I(k) = (0, 0, \cdots, 0, 1, 0, \cdots, 0)^t$, where 1 is the k^{th} element. Define the L_1-norm of a column vector $u = (u_1, u_2, \cdots, u_d)^t \in \mathbb{R}^d$ as $\|u\|_1 = \sum_{i=1}^d |u_i|$. We can represent $L(\boldsymbol{y}_{1:n}|\phi_K)$ as

$$L(\boldsymbol{y}_{1:n}|\phi_K) = \log \|M_n(\boldsymbol{\phi}_K)M_{n-1}(\boldsymbol{\phi}_K)\cdots M_1(\boldsymbol{\phi}_K)I(x_0)\|_1, \tag{B.14}$$

where M_i is defined before condition (C1) in Section B.1.1. Since $\{X_i : i \geq 0\}$ is a Markov chain on $\mathcal{X}_{K^*} := \{1, 2, \cdots, K^*\}$ and $\{(X_n, Y_n) : n \geq 0\}$ is a Markov chain on $\mathcal{X}_{K^*} \times \mathcal{Y}$, if we define $Gl(K, \mathbb{R})$ as the set of invertible $K \times K$ matrices with real entries and

$$T_n = T_n(\boldsymbol{\phi}_K) := M_n(\boldsymbol{\phi}_K)M_{n-1}(\boldsymbol{\phi}_K)\cdots M_1(\boldsymbol{\phi}_K),$$

then $\{(X_n, Y_n, T_n) : n \geq 0\}$ is a Markov chain on $\mathcal{X}_{K^*} \times \mathcal{Y} \times Gl(K, \mathbb{R})$.

For any $u \in \mathbb{R}^d$ with $\|u\|_1 \neq 0$, let $\overline{u} := u/\|u\|_1$ be its normalization and denote $\mathbb{P}(\mathbb{R}^d)$ be the projection space of $\mathbb{R}^d$ which contains all such $\overline{u}$. Let

$$W_i = W_i(\boldsymbol{\phi}_K) := \left(X_i, Y_i, \overline{T_i(\boldsymbol{\phi}_K)I_0(x_0)}\right).$$

Fuh (2003) proves that $\{W_n : n \geq 0\}$ is a Markov chain on $\mathcal{X}_{K^*} \times \mathcal{Y} \times \mathbb{P}(\mathbb{R}^K)$ with an invariant measure. From (B.14), $L(\boldsymbol{y}_{1:n}|\phi_K)$ can be written as an additive functional of the Markov chain $\{W_n(\boldsymbol{\phi}_K) : n \geq 0\}$, i.e.

$$L(\boldsymbol{y}_{1:n}|\phi_K) = \sum_{i=1}^{n} \log \frac{\|T_i(\boldsymbol{\phi}_K)I(x_0)\|_1}{\|T_{i-1}(\boldsymbol{\phi}_K)I(x_0)\|_1} = \sum_{i=1}^{n} \sigma\left(W_i(\boldsymbol{\phi}_K), W_{i-1}(\boldsymbol{\phi}_K)\right), \tag{B.15}$$

where $\sigma\left(W_i(\boldsymbol{\phi}_K), W_{i-1}(\boldsymbol{\phi}_K)\right) := \log \frac{\|T_i(\boldsymbol{\phi}_K)I(x_0)\|_1}{\|T_{i-1}(\boldsymbol{\phi}_K)I(x_0)\|_1}$. Therefore, by the law of large numbers of the additive functional of a Markov chain, there exists μ such that $n^{-1}L(\boldsymbol{y}_{1:n}|\phi_K) \to \mu$ almost surely. □

Proof of Lemma 2

Proof. (i) For any $\epsilon \in (0, 1)$ and $\mathcal{X}_{K,k,\epsilon}$ defined in (3.5), define

$$p_\epsilon^k(\boldsymbol{y}_{1:n}|\phi_K) := \sum_{\mathcal{X}_{K,k,\epsilon}^n} p(\boldsymbol{y}_{1:n}, \boldsymbol{x}_{0:n}|\phi_K),$$

and $L^k_\epsilon = L^k_\epsilon(\boldsymbol{y}_{1:n}|\phi_K) = \log p^k_\epsilon(\boldsymbol{y}_{1:n}|\phi_K)$. From (3.6),

$$p_\epsilon(\boldsymbol{y}_{1:n}|\phi_K) \le \sum_{k=1}^{K} p^k_\epsilon(\boldsymbol{y}_{1:n}|\phi_K) \le K \max_{1\le k\le K} p^k_\epsilon(\boldsymbol{y}_{1:n}|\phi_K).$$

Hence it suffices to prove that for all k, there exists $\mu'_k \in \mathbb{R}$ such that

$$\limsup_{n\to\infty} n^{-1} L^k_\epsilon(\boldsymbol{y}_{1:n}|\phi_K) \le \mu'_k - \epsilon \log \omega. \tag{B.16}$$

We first consider the case of $k = K$. Let $\tilde{F}_i$ be the $K \times K$ diagonal matrix with diagonal elements $(f(y_i|\boldsymbol{\theta}_1), \cdots, f(y_i|\boldsymbol{\theta}_{K-1}), \omega f(y_i|\boldsymbol{\theta}_K))$, i.e., multiplying ω in the K-th diagonal element of F_i. Similar to M_i, define $\tilde{M}_i = \tilde{M}_i(\phi_K) = \tilde{F}_i Q^t_K$. A direct computation shows that

$$\begin{aligned}
\widetilde{p}^K(\boldsymbol{y}_{1:n}|\phi_K) &:= \|\tilde{M}_n(\boldsymbol{\phi}_K)\tilde{M}_{n-1}(\boldsymbol{\phi}_K)\cdots\tilde{M}_1(\boldsymbol{\phi}_K)I(x_0)\| \\
&= \sum_{\mathcal{X}^n_K} \prod_{k=1}^{K} \prod_{i:x_i=k} f(y_i|\boldsymbol{\theta}_k) \times \prod_{i=1}^{n} q_{x_{i-1}x_i} \times \omega^{n_K} \\
&\ge \sum_{\mathcal{X}^n_{K,K,\epsilon}} \prod_{k=1}^{K} \prod_{i:x_i=k} f(y_i|\boldsymbol{\theta}_k) \times \prod_{i=1}^{n} q_{x_{i-1}x_i} \times \omega^{n_K} \\
&\ge \sum_{\mathcal{X}^n_{K,K,\epsilon}} \prod_{k=1}^{K} \prod_{i:x_i=k} f(y_i|\boldsymbol{\theta}_k) \times \prod_{i=1}^{n} q_{x_{i-1}x_i} \times \omega^{\epsilon n} \\
&= \omega^{\epsilon n} p^K_\epsilon(\boldsymbol{y}_{1:n}|\phi_K),
\end{aligned} \tag{B.17}$$

where n_K is defined in (3.4). The first equality is because ω is multiplied each time the Markov chain $\{x_i\}$ enters state K, and the total number of entrance to state K is n_K. The last inequality is because on $\mathcal{X}^n_{K,K,\epsilon}$, $n_K < \epsilon n$.

Define $\widetilde{L}^K = \widetilde{L}^K(\boldsymbol{y}_{1:n}|\phi_K) = \log \widetilde{p}^K(\boldsymbol{y}_{1:n}|\phi_K)$. Then similar to the argument in

(B.15), $\widetilde{L}^K$ can also be written as an additive functional of a Markov chain. By the law of large numbers of the additive functional of a Markov chain (Fuh, 2003), there exists μ'_K such that $n^{-1}\widetilde{L}^K \to \mu'_K$ almost surely. Combined with (B.17), we have

$$\epsilon \log \omega + \frac{L_\epsilon^K}{n} \leq \frac{\widetilde{L}^K}{n} \xrightarrow{a.s.} \mu'_K.$$

Therefore, (B.16) holds for $k = K$. For general $1 \leq k \leq K$, the procedure above applies except that the ω in $\tilde{M}_i$ is multiplied on the k^{th} row instead.

(ii) By (i), it suffices to prove that $\mu'_k < \mu$ for $k = K$. Without loss of generality, we assume that $x_0 = K$. Let $\{Z_i : i \geq 0\}$ be a Markov chain governed by the transition matrix Q_K with $Z_0 = K$ and it is independent of the original $\{(X_i, Y_i) : i \geq 0\}$. Let $\tau_0 = 0$, and recursively define $\tau_j = \inf\{i : i > \tau_{j-1}, Z_i = K\}$, i.e. the stopping time that the chain $\{Z_i\}$ revisits state K. For any positive integer m, let $M = E\tau_m$. Fix $\delta \in (0,1)$, set $m' = [(1-\delta)m]$ and $\mathcal{E} = \{\tau_{m'} < M\}$. Conditioning on $Y_{1:M} = \boldsymbol{y}_{1:M}$, the only randomness comes from $\{Z_i, i \geq 1\}$. By the dominated convergence theorem, as

$m \to \infty$, with probability one,

$$\begin{aligned}
\frac{1}{M} L(\boldsymbol{y}_{1:M}|\phi_K) =& \frac{1}{M} \log E \left[\prod_{i=1}^{M} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i} \middle| Y_{1:M} = \boldsymbol{y}_{1:M} \right] \\
=& \frac{1}{M} \log E \left[\prod_{i=1}^{M} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i}; \mathcal{E} \middle| Y_{1:M} = \boldsymbol{y}_{1:M} \right] + o(1) \\
=& \frac{1}{M} \log E \left[\prod_{j=1}^{m'} \prod_{i=\tau_{j-1}+1}^{\tau_j} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i} \right. \\
& \left. \times \prod_{i=\tau_{m'}+1}^{\tau_{m'}+|M-\tau_{m'}|} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i}; \mathcal{E} \middle| Y_{1:M} = \boldsymbol{y}_{1:M} \right] + o(1) \\
=& \frac{1}{M} \log E \left[\prod_{j=1}^{m'} \prod_{i=\tau_{j-1}+1}^{\tau_j} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i} \right. \\
& \left. \times \prod_{i=\tau_{m'}+1}^{\tau_{m'}+|M-\tau_{m'}|} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i} \middle| Y_{1:M} = \boldsymbol{y}_{1:M} \right] + o(1) \\
=& \frac{1}{M} \log E \left[\prod_{j=1}^{m'} V_j \times V_{m'+1} \middle| Y_{1:M} = \boldsymbol{y}_{1:M} \right] + o(1),
\end{aligned}$$

where

$$V_j = \prod_{i=\tau_{j-1}+1}^{\tau_j} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i}$$

for $1 \le j \le m'$, and

$$V_{m'+1} = \prod_{j=\tau_{m'}+1}^{\tau_{m'}+|M-\tau_{m'}|} f(Y_i|\theta_{Z_i}) q_{Z_{i-1}Z_i}.$$

Since τ_j is the recurrence time that the chain $\{Z_i\}$ revisits state K, the $\{\tau_j - \tau_{j-1}, j = 1, 2, \cdots, \}$ are i.i.d.. As a result, when $\{Y_i, i = 1, 2, \cdots\}$ is conditioned on, $\{V_j, 1 \le j \le m'\}$ are independent. Since $Z_{\tau_{m'}} = K$, $V_{m'+1}$ is independent of $\{V_j, 1 \le j \le m'\}$.

Hence with probability one, as $m \to \infty$,

$$\begin{aligned}\frac{1}{M}L(\boldsymbol{y}_{1:M}|\phi_K) =&\frac{1}{M}\log E\left[\prod_{j=1}^{m'} V_j \times V_{m'+1}\middle|Y_{1:M} = \boldsymbol{y}_{1:M}\right] + o(1)\\ =&\frac{1}{M}\sum_{j=1}^{m'}\log E[V_j|Y_{1:M} = \boldsymbol{y}_{1:M}]\\ &+\frac{1}{M}\log E[V_{m'+1}|Y_{1:M} = \boldsymbol{y}_{1:M}] + o(1)\\ =&I + II + o(1).\end{aligned} \tag{B.18}$$

Similarly, with probability one, as $m \to \infty$,

$$\begin{aligned}\frac{1}{M}\widetilde{L}^K(\boldsymbol{y}_{1:M}|\phi_K) =&\frac{1}{M}\sum_{j=1}^{m'}\log E[V_j\omega|Y_{1:M} = \boldsymbol{y}_{1:M}]\\ &+\frac{1}{M}\log E[\tilde{V}_{m'+1}|Y_{1:M} = \boldsymbol{y}_{1:M}] + o(1)\\ =&\frac{1}{M}\sum_{j=1}^{m'}\log E[V_j|Y_{1:M} = \boldsymbol{y}_{1:M}] + \frac{m'}{M}\log\omega\\ &+\frac{1}{M}\log E[\tilde{V}_{m'+1}|Y_{1:M} = \boldsymbol{y}_{1:M}] + o(1)\\ =&I + \frac{m'}{M}\log\omega + \tilde{I}I + o(1),\end{aligned} \tag{B.19}$$

where

$$\tilde{V}_{m'+1} = \prod_{i=\tau_{m'}+1}^{\tau_{m'}+|M-\tau_{m'}|} f(Y_i|\theta_{Z_i})q_{Z_{i-1}Z_i}\omega^{1_{Z_i=K}} \leq V_{m'+1},$$

since $\omega \in (0,1)$. Therefore, $\tilde{I}I \leq II$. Recall that $\tau_j - \tau_{j-1}$ are i.i.d. and note that $\tau_m = \sum_{j=1}^m \tau_j - \tau_{j-1}$, we have

$$\frac{m'}{M} = \frac{[(1-\delta)m]}{E\tau_m} = \frac{[(1-\delta)m]}{m}\frac{m}{mE\tau_1} \to \frac{1-\delta}{E\tau_1} := c > 0, \tag{B.20}$$

as $m \to \infty$. Combined with part (i), as $m \to \infty$, with probability one,

$$\begin{aligned}
\mu'_K &= \limsup_{M\to\infty} \frac{1}{M} \widetilde{L}^K(\boldsymbol{y}_{1:M}|\phi_K) = \limsup_{M\to\infty} \left\{ I + \tilde{I}I + \frac{m'}{M}\log\omega \right\} \\
&\leq \limsup_{M\to\infty} \left\{ I + II + \frac{m'}{M}\log\omega \right\} = \limsup_{M\to\infty}\{I + II\} + \limsup_{M\to\infty} \frac{m'}{M}\log\omega \\
&= \limsup_{M\to\infty} \frac{1}{M} L(\boldsymbol{y}_{1:M}|\phi_K) + c\log\omega = \mu + c\log\omega < \mu,
\end{aligned}$$

since $c > 0$ and $\omega < 1$. Similar proofs hold for $k \neq K$, thus by definition $\mu' = \max_{1\leq k\leq K} \mu'_k < \mu$.

□

B.1.3 Proof of Theorem 1

For simplicity, we give a detailed proof of the case with $d = 1$; the cases with higher dimensions can be proved similarly. We further simplify the proof to the case of $K = K^* + 1$ by showing that the general case of $K > K^*$ and $K < K^*$ can be derived similarly in sections B.1.3 and B.1.3 respectively.

de Gunst & Shcherbakova (2008) shows that the posterior distribution under the true number of states converges to a Gaussian distribution centered at the efficient estimator, $\hat{\phi}_{K^*}$, of the true parameters ϕ^*, i.e.

$$\frac{p(\boldsymbol{y}_{1:n}|\hat{\phi}_{K^*})}{n^{(K^*)^2/2} p_{K^*}(\boldsymbol{y}_{1:n})} = O_P(1), \tag{B.21}$$

where the number of parameters is $(K^*)^2$ in a K^* state HMM when $d = 1$. This gives us the convergence rate for the denominator in (1). To obtain (1), we need a rate control for the numerator as well. Unfortunately, such kind of central limit theorem does not exist for $K > K^*$. Hence next we attain a weaker rate for the numerator with the help of Theorem 2 and Gassiat & Rousseau (2014).

Let $\pi_k = \pi_k(Q_K)$ be the invariant measure for state k given transition matrix Q_K. For all $u = u_n \downarrow 0$, consider the collection of states

$$J(Q_K, u_n) := \{k : 1 \leq k \leq K, \pi_k \geq u_n\};$$

i.e., $J(Q_K, u_n)$ represents the states with significant probability weight under Q_K. For any $v_n \downarrow 0$ and $\phi_K \in \Phi_K$, for each $k \in J(Q_K, u_n)$, let

$$A_k(\phi_K, u_n, v_n) := \left\{l : l \in J(Q_K, u_n), |\theta_k - \theta_l|^2 \leq v_n\right\},$$

i.e., $A_k(\phi_K, u_n, v_n)$ collects all states l with significant probability weight under Q_K, and with parameter θ_l close enough to θ_k. Define the number of clusters $\tilde{K}_n = \tilde{K}_n(\phi_K, u_n, v_n)$ as

$$\min\left\{n : \exists 1 \leq k_1 < k_2 < \cdots < k_n \leq K \text{ s.t. } \forall l \in J(Q_K, u_n), l \in A_{k_t}(\phi_K, u_n, v_n) \text{ for some } t\right\},$$

and consider the set of parameters with $\tilde{K}_n = K^*$, the true number of states,

$$\tilde{\Phi} = \left\{\phi_K : \phi_K \in \Phi_K := \Theta^K \times \mathcal{Q}_K, \tilde{K}_n(\phi_K, u_n, v_n) = K^*\right\}.$$

For any $\epsilon = \epsilon_n \downarrow 0$, consider the set

$$\Phi_\epsilon := \{\phi_K : \phi_K \in \Phi_K, \pi_k \geq \epsilon \text{ for all } 1 \leq k \leq K\}.$$

Lemma 3. *If $u_n v_n \downarrow 0$ slow enough (Gassiat & Rousseau, 2014),*

$$p_K(\boldsymbol{y}_{1:n}) \backsim \int_{\tilde{\Phi}\cap\Phi_\epsilon} p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)d\phi_K.$$

Lemma 4. *If $\epsilon_n > u_n$, then for all $\phi_K \in \tilde{\Phi} \cap \Phi_\epsilon$, there exists a partition $\mathcal{P} = \mathcal{P}(\phi_K) := \{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_{K^*}\}$ of $\{1, 2, \cdots, K\}$ such that none of the $\mathcal{P}_k$ is empty, $\cup\mathcal{P}_k = \{1, 2, \cdots, K\}$, and for all $k_1, k_2 \in \mathcal{P}_k$, $|\theta_{k_1} - \theta_{k_2}|^2 \leq v_n$.*

The proofs of Lemmas 3 and 4 are in B.1.3 and B.1.3 respectively.

For any partition $\mathcal{S} = \{\mathcal{S}_1, \cdots, \mathcal{S}_{K^*}\}$ of $\{1, \cdots, K\}$, define

$$\Phi_{\mathcal{S}} := \{\phi_K : \mathcal{P}(\phi_K) = \mathcal{S}\},$$

where $\mathcal{P}(\cdot)$ is defined in Lemma 4. Then $\tilde{\Phi} \cap \Phi_\epsilon = \bigcup_{\mathcal{S}} \left\{\tilde{\Phi} \cap \Phi_\epsilon \cap \Phi_{\mathcal{S}}\right\}$.

For any Q_K, let $\vec{q}_K = (q_{1K}, q_{2K}, \cdots, q_{KK}, q_{K1}, q_{K2}, \cdots, q_{KK})$, i.e., all transition probabilities from state 1 to K. Let $\mathcal{V}_K$ be the space of all possible $\vec{q}_K$, and $\mathcal{V}_\epsilon$ be the projection of Φ_ϵ on $\mathcal{V}_K$. Let $Q_{-K} = \{q_{kl}\}_{1\leq k,l<K}$, namely, the sub-matrix of Q_K excluding row K and column K. Denote $\phi_{-K} = (\theta, Q_{-K})$ and write $Q_K = (Q_{-K}, \vec{q}_K)$ and $\phi_K = (\phi_{-K}, \vec{q}_K) = (\theta, Q_{-K}, \vec{q}_K)$. For a *fixed* $v \in \mathcal{V}_\epsilon$, consider the subspace

$$\Phi_{K,\epsilon,v} := \{\phi_K = (\theta, Q_{-K}, v) : \phi \in \tilde{\Phi} \cap \Phi_\epsilon \cap \Phi_{\mathcal{S}}\},$$

i.e., all ϕ_K with $\vec{q}_K = v$. Let $\hat{\phi}_v = (\hat{\theta}, \hat{Q}_{-K}, v)$ be the maximum likelihood estimator of the K-state model in the space $\Phi_{K,\epsilon,v}$, i.e.,

$$\hat{\phi}_v = \left(\hat{\theta}, \hat{Q}_{-K}, v\right) := \mathrm{argmax}_{\mathrm{\phi_K\in\Phi_{K,\epsilon,v}}}\mathrm{p_K(\mathbf{y}_{1:n}|\phi_K)}.$$

Lemma 5. *When $K = K^* + 1$, the only possible $\mathcal{S}$ is the one with exactly one $\mathcal{S}_k$ that contains exactly two elements, and each other $\mathcal{S}_k$ contains one single element. If $\mathcal{S}_k = \{k\}$ for all $1 \leq k \leq K^* - 1$, and $\mathcal{S}_{K^*} = \{K^*, K^* + 1\}$, then*

$$\int_{\tilde{\Phi}\cap\Phi_\epsilon\cap\Phi_{\mathcal{S}}} \frac{p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)}{p(\boldsymbol{y}_{1:n}|\hat{\phi}_v)} d\phi_K = O_P\left(n^{-\frac{(K^*)^2+1}{2}} \log n\right).$$

The proof of Lemma 5 is given in B.1.3.

The result above is under the partition $\mathcal{S} = \{\{1\}, \{2\}, \cdots, \{K^* - 1\}, \{K^*, K^* + 1\}\}$. Note that the argument also holds for any other possible partition $\mathcal{S}$, and the total number of possible $\mathcal{S}$ is finite. Combined with (B.21), we have

$$\begin{aligned}\frac{p_K(\boldsymbol{y}_{1:n})}{p_{K^*}(\boldsymbol{y}_{1:n})} &\backsim \frac{\int_{\tilde{\Phi}\cap\Phi_{2\epsilon}\cap\Phi_{\mathcal{S}}} \frac{p_K(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)}{p_K(\boldsymbol{y}_{1:n}|\hat{\phi}_v)} d\phi_K}{p_{K^*}(\boldsymbol{y}_{1:n})/p_{K^*}(\boldsymbol{y}_{1:n}|\hat{\phi}_{K^*})} \\ &= O_P\left(\frac{n^{-\frac{(K^*)^2+1}{2}} \log n}{n^{-(K^*)^2/2}}\right) = O_P(n^{-\frac{1}{2}} \log n).\end{aligned}$$

Remark 6. *Choice of ϵ_n. In Example 1, we briefly present how we apply Theorem 2 to prove Theorem 1 by giving an uniform bound through the choice of $\epsilon_n \downarrow 0$ with $\epsilon_n^2 n \to \infty$. However, as shown in the proof of Theorem 1, to give a uniform lower bound across paths, we need to choose $\epsilon_n \downarrow 0$ slow enough to exclude the case of $\pi_k < u_n$ for some state k.*

Proof of Lemma 3

Proof. When $u_n v_n \downarrow 0$ slow enough, Gassiat & Rousseau (2014) shows that $\tilde{K}_n \to K^*$ in probability under the posterior distribution, i.e.,

$$p_K(\boldsymbol{y}_{1:n}) = \int_{\Phi_K} p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K) \backsim \int_{\tilde{\Phi}} p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)d\phi_K.$$

Moreover, for $\epsilon = \epsilon_n \downarrow 0$, Theorem 2 gives

$$\begin{aligned}
\int_{\tilde{\Phi}} p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)d\phi_K &= \int_{\tilde{\Phi}} \sum_{\boldsymbol{x}_{1:n}\in\mathcal{X}_K^n} p(\boldsymbol{y}_{1:n}, \boldsymbol{x}_{0:n}|\phi_K)p_0(\phi_K)d\phi_K \\
&\backsim \int_{\tilde{\Phi}} \sum_{\boldsymbol{x}_{1:n}\in\mathcal{X}_K^n-\mathcal{X}_{K,\epsilon}^n} p(\boldsymbol{y}_{1:n}, \boldsymbol{x}_{0:n}|\phi_K)p_0(\phi_K)d\phi_K.
\end{aligned}$$

In addition, consider the set Φ_ϵ, since the likelihood on the right-hand-side above only sum up paths in $\mathcal{X}_K^n - \mathcal{X}_{K,\epsilon}^n$, it is clear that

$$\begin{aligned}
&\int_{\tilde{\Phi}} \sum_{\boldsymbol{x}_{1:n}\in\mathcal{X}_K^n-\mathcal{X}_{K,\epsilon}^n} p(\boldsymbol{y}_{1:n}, \boldsymbol{x}_{0:n}|\phi_K)p_0(\phi_K)d\phi_K \\
\backsim &\int_{\tilde{\Phi}\cap\Phi_\epsilon} \sum_{\boldsymbol{x}_{1:n}\in\mathcal{X}_K^n-\mathcal{X}_{K,\epsilon}^n} p(\boldsymbol{y}_{1:n}, \boldsymbol{x}_{0:n}|\phi_K)p_0(\phi_K)d\phi_K.
\end{aligned}$$

Again, by Theorem 2, we have,

$$\begin{aligned}
&\int_{\tilde{\Phi}\cap\Phi_\epsilon} \sum_{\boldsymbol{x}_{1:n}\in\mathcal{X}_K^n-\mathcal{X}_{K,\epsilon}^n} p(\boldsymbol{y}_{1:n}, \boldsymbol{x}_{0:n}|\phi_K)p_0(\phi_K)d\phi_K \\
\backsim &\int_{\tilde{\Phi}\cap\Phi_\epsilon} \sum_{\boldsymbol{x}_{1:n}\in\mathcal{X}_K^n} p(\boldsymbol{y}_{1:n}, \boldsymbol{x}_{0:n}|\phi_K)p_0(\phi_K)d\phi_K \\
= &\int_{\tilde{\Phi}\cap\Phi_\epsilon} p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)d\phi_K.
\end{aligned}$$

□

Proof of Lemma 4

Proof. When $\epsilon_n > u_n$, note that on $\tilde{\Phi}$, $\tilde{K}_n = K^*$; but since $K > K^*$, either some state k has $\pi_k < u_n$, or there exists $1 \leq k \leq K^*$ and multiple states $1 \leq k_1 < k_2 \leq K$ such that $|\theta_{k_1} - \theta_{k_2}|^2 \leq v_n$. However, the first case is not possible on Φ_ϵ since π_k are all greater than $\epsilon_n > u_n$. Hence, we only have the second case, which yields the desired result. □

Proof of Lemma 5

Proof. We consider the case with $\mathcal{S}_k = \{k\}$ for all $1 \leq k \leq K^* - 1$, and $\mathcal{S}_{K^*} = \{K^*, K^* + 1\}$. Though we fit a K state HMM, the state K^* and state $K^* + 1$ have very similar parameters thus behave almost like one single state. So we compare it to the case where two states are merged into one.

The conditional maximum likelihood $\hat{\phi}_v = (\hat{\theta}, \hat{Q}_{-K}, v)$ converges to the "true" parameter $\tilde{\phi}^*_v = (\tilde{\theta}^*, \tilde{Q}^*_{-K}, v)$ given by

$$\tilde{\theta}^* = (\theta^*_1, \theta^*_2, \cdots, \theta^*_{K^*}, \theta^*_{K^*})$$

and

$$\tilde{q}^*_{kl} = q^*_{kl}$$

$$\tilde{q}^*_{kK^*} + \tilde{q}^*_{kK} = q^*_{kK^*},$$

$$\tilde{q}^*_{kK^*}\tilde{q}^*_{K^*l} + \tilde{q}^*_{kK}\tilde{q}^*_{Kl} = q^*_{kK^*}q^*_{K^*l}$$

for all $1 \leq k, l \leq K^* - 1$ since the "estimated" states K^* and $K = K^* + 1$ are actually splitting the "true" state K^* into two.

Denote $p_v(\phi) = c_v p_0(\phi)$, where c_v is a normalization constant so that p_v is a probability measure on $\Phi_{K,\epsilon,v}$. Similar to (B.21), we have

$$\frac{n^{\frac{(K^*)^2+1}{2}} \int_{\Phi_{K,\epsilon,v}} p(\boldsymbol{y}_{1:n}|\phi_K) p_v(\phi_K) d\phi_K}{p(\boldsymbol{y}_{1:n}|\hat{\phi}_v)} = O_P(1). \tag{B.22}$$

In other words,

$$\frac{\int_{\Phi_{K,\epsilon,v}} p(\boldsymbol{y}_{1:n}|\phi_K) p_0(\phi_K) d\phi_K}{p(\boldsymbol{y}_{1:n}|\hat{\phi}_v)} = O_P\left(n^{-\frac{(K^*)^2+1}{2}}\right). \tag{B.23}$$

Note that, in equations (B.23), the order of the O_P term is $-\frac{(K^*)^2+1}{2}$. Originally, for a K-state model, we have $K(K-1)$ parameters for the transition matrix and K parameters for θ (recall that we present the case with $d = 1$ here). However, since we "locked" $\vec{q}_K = v$, which fixes $K - 1$ parameters for $(q_{1K}, q_{2K}, \cdot, q_{(K-1)K})$ and another $K - 1$ parameters for $(q_{K1}, q_{K2}, \cdots, q_{K(K-1)}$. Hence, there are $K(K-1) + K - 2(K-1) = K^2 - 2K + 2 = (K^*)^2 + 1$ free parameters left to be determined, recall that $K = K^* + 1$ here.

Now we integrate (B.23) over all possible v. For any *fixed* and sufficiently small $\bar{\epsilon} > 0$, since $\mathcal{V}_{\bar{\epsilon}}$ is a bounded and closed set, we can extend (B.23) to

$$\int_{\mathcal{V}_{\bar{\epsilon}}} \frac{\int_{\Phi_{K,\epsilon,v}} p(\boldsymbol{y}_{1:n}|\phi_K) p_0(\phi_K) d\phi_K}{p(\boldsymbol{y}_{1:n}|\hat{\phi}_v)} dv = O_P\left(n^{-\frac{(K^*)^2+1}{2}}\right).$$

Since this holds for any sufficiently small $\bar{\epsilon} > 0$, for any $a_n \uparrow \infty$, we can choose $\epsilon =$

$\epsilon_n \downarrow 0$ slow enough such that

$$\int_{\tilde{\Phi}\cap\Phi_\epsilon\cap\Phi_{\mathcal{S}}} \frac{p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)}{p(\boldsymbol{y}_{1:n}|\hat{\phi}_v)} d\phi_K = \int_{\mathcal{V}_\epsilon} \frac{\int_{\Phi_{K,\epsilon,v}} p(\boldsymbol{y}_{1:n}|\phi_K)p_0(\phi_K)d\phi_K}{p(\boldsymbol{y}_{1:n}|\hat{\phi}_v)} dv$$
$$= O_P\left(n^{-\frac{(K^*)^2+1}{2}} a_n\right).$$

In particular, setting $a_n = \log n$ gives a rate of $O_P\left(n^{-\frac{(K^*)^2+1}{2}} \log n\right)$.

□

General Overfitting Case

We have detailed the proof for $K = K^* + 1$. The general case for $K > K^*$ is similar. The only major difference is that the possible configurations of $\mathcal{S}$ would be more complicated. For example, when $K^* = 2$ and $K = 4$, there might be partitions like $\{\{1,2\},\{3,4\}\}$ and $\{\{1\},\{2,3,4\}\}$. By such, instead of simply condition on a single $\vec{q}_K$, one would need to condition on multiple components so that each set in the partition have only one element left unconditioned. For example, for $\mathcal{S} = \{\{1,2\},\{3,4\}\}$, one would need to condition on both $\vec{q}_2$ and $\vec{q}_4$; as for $\mathcal{S} = \{\{1\},\{2,3,4\}\}$, one would need to condition on $\vec{q}_3$ and $\vec{q}_4$. Despite of this complexity, the total number of undetermined parameters are still $(K^*)^2 + 1$, so (B.22) still holds. Also, the total number of possible $\mathcal{S}$ is still finite, so the arguments for the rest of the proof are still valid.

General Underfitting Case

The proof of $K < K^*$ is much simpler. For any $\phi = (\theta, Q) \in \Phi_K$, consider $\phi' = (\theta', Q') \in \Phi_{K^*}$ defined by

$$\begin{aligned}
\theta'_k &= \theta_k, \\
\theta'_K &= \theta'_{K+1} = \cdots \theta'_{K^*} = \theta_K, \\
q'_{kl} &= \frac{q_{kK}}{K^* - K + 1}, \\
q'_{lk} &= q_{Kk}
\end{aligned}$$

for all $1 \le k < K$ and $K \le l \le K^*$; i.e., state K in ϕ is equally split into states $K, K+1, \cdots, K^*$ in ϕ'. A direct calculation shows that $p_K(\boldsymbol{y}_{1:n}|\phi) = p_{K^*}(\boldsymbol{y}_{1:n}|\phi')$. Since ϕ' must have $\theta'_K = \theta'_{K+1} = \cdots = \theta'_{K^*}$, we have

$$||\phi' - \phi^*|| \ge \inf_{\theta} \sum_{k=K}^{K^*} |\theta - \theta^*_k| \ge \min_{K \le k < l \le K^*} |\theta^*_k - \theta^*_l| := \delta > 0. \tag{B.24}$$

By Lemma 3.1 in de Gunst & Shcherbakova (2008), there exists $c > 0$ such that

$$\sup_{||\phi' - \phi^*|| \ge \delta} \mu(\phi') \le \mu(\phi^*) - c,$$

which implies

$$\frac{\int_{\Phi_K} p_K(\boldsymbol{y}_{1:n}|\phi) p_0(\phi) d\phi}{p_{K^*}(\boldsymbol{y}_{1:n}|\phi^*) p_0(\phi^*)} = \frac{1}{p_0(\phi^*)} \int_{\Phi_K} \frac{p_{K^*}(\boldsymbol{y}_{1:n}|\phi')}{p_{K^*}(\boldsymbol{y}_{1:n}|\phi^*)} p_0(\phi) d\phi \le O_P(e^{-cn}).$$

Combined with the fact that

$$\frac{p_{K^*}(\boldsymbol{y}_{1:n}|\phi^*)p_0(\phi^*)}{p_{K^*}(\boldsymbol{y}_{1:n})} = O_P(n^{(K^*)^2/2}),$$

we obtain

$$\begin{aligned}\frac{p_K(\boldsymbol{y}_{1:n})}{p_{K^*}(\boldsymbol{y}_{1:n})} &= \frac{\int_{\Phi_K} P_K(\boldsymbol{y}_{1:n}|\phi)p_0(\phi)d\phi}{p_{K^*}(\boldsymbol{y}_{1:n}|\phi^*)p_0(\phi^*)} \times \frac{p_{K^*}(\boldsymbol{y}_{1:n}|\phi^*)p_0(\phi^*)}{p_{K^*}(\boldsymbol{y}_{1:n})} \\ &= O_P\left(n^{(K^*)^2/2}\right) \times O_P\left(e^{-cn}\right) = O_P\left(n^{(K^*)^2/2}e^{-cn}\right),\end{aligned}$$

which completes the proof.

B.1.4 Regularity Conditions and Proof of Theorem 3

In Theorem 3, we assume the same regularity conditions as in Theorem 1 except for replacing condition (B1) with

(B1') The prior distribution on the transition matrix is in the form of

$$\nu_K(Q_K) = \begin{cases} \tilde{\nu}(q_{11}, q_{12}, \cdots, q_{kK})\delta(Q_K), & Q_K \in \tilde{\mathcal{Q}}_K \\ 0, & Q_k \notin \tilde{\mathcal{Q}}_K \end{cases}$$

where $\tilde{\mathcal{Q}}_K = \{Q : q_{1k} = q_{2k} = \cdots = q_{Kk} \text{ for all } 1 \leq k \leq K\}$, and $\delta(\cdot)$ is the delta function. Moreover, there exists $C > 0$, $\alpha_1 > 0$, $\cdots$, $\alpha_K > 0$ such that

$$0 < \tilde{\nu}(u_1, u_2, \cdots, u_K) < Cu_1^{\alpha_1 - 1}u_2^{\alpha_2 - 1} \cdots u_K^{\alpha_K - 1}. \tag{B.25}$$

for all $(u_1, u_2, \cdots, u_K)$ satisfying

$$\min(u_1, u_2, \cdots, u_K) > 0, \text{ and } \sum_{l=1}^{K} u_k = 1.$$

Furthermore, suppose that $\min_{k=1}^{K} \alpha_k > d/2$.

Proof of Theorem 3. As $n \to \infty$, by the law of large numbers, $\frac{\#\{X_n=k\}}{n}$ converges to the invariant measure of state k with probability one. Therefore, we can treat the data as coming from a mixture model with K^* mixture components, with the probability weight of the k^{th} mixture component equal to the invariant measure of state k.

We then go through the same procedure as in the proof of Theorem 1. Theorem 2 still applies thus we can ignore the irregular paths. Since the log-likelihood in this case is already in additive form, Lemma 1 and Lemma 2 can be obtained easily. The only remaining part is the two cases we divided using u_n and v_n through Gassiat & Rousseau (2014), which we switch to the corresponding results in Rousseau & Mengersen (2011) instead.

Remark 7. *One can compare this result to Chambaz & Rousseau (2005), who present similar rate of convergence.*

B.2 Simulation Studies for Estimation of Normalizing Constant

We use models with known normalizing constants to test the performance of the estimation of normalizing constant proposed in Section 3.4.1. The first family of models is d-dimensional Gaussian mixture models with three distinct components whose covariance matrices are diagonal with diagonal elements all equal to 0.1, $2 \leq d \leq 30$;

the normalizing constant is set to be $C_1 = \exp(10)$. The second family of models, with normalizing constant $C_2 = \exp(2)$, has three independent dimensions: the first dimension is Gaussian with mean 1 and variance 1, the second dimension is student-t distribution with degree of freedom 2, and the third dimension is Gamma distribution with shape parameter 6 and scale parameter 2.

We perform repeated simulation studies with the two families of models as follows: first simulate N_{sim} samples independently from the model, apply the importance sampling algorithm mentioned in Section 3.4.1 with N_{is} samples from the fitted importance function, the Gaussian tail and the t-tail. For comparison, we also use the reciprocal importance sampling with fitted Gaussian mixture as the importance function and multivariate t-distribution mixture as the importance function. The results are summarized in Table B.1.

B.3 Simulation Robustness

Theorem 7. *Assume that the true number of hidden states is K^*. Let $P_{K,n} = p_K(\boldsymbol{y}_{1:n})$ be the marginal likelihood with K hidden states. Let $\{\varphi_{K,m} : m = 1, 2, \cdots, M_n\}$ be i.i.d. samples from $g_{n,K}(\cdot)$, the chosen importance function which approximates the posterior distribution $p_K(\phi_K|\boldsymbol{y}_{1:n})$. Then, the estimated marginal likelihood using the importance sampling is given by*

$$\hat{P}_{K,n,M_n} = \hat{P}_{K,n,M_n}(\boldsymbol{y}_{1:n}, \{\varphi_{K,m}\}_{1\leq m\leq M_n}) = \frac{1}{M_n}\sum_{m=1}^{M_n} \frac{p_K(\boldsymbol{y}_{1:n}, \varphi_{K,m})}{g_{n,K}(\varphi_{K,m})}.$$

Similarly, we can give the locally restricted version of this estimator. If for any finite $K \neq K^*$, $P_{K,n}/P_{K^*,n} \to 0$ in probability as $n \to \infty$, then $\hat{P}_{K,n,M_n}/\hat{P}_{K^*,n,M_n} \to 0$ in probability as $n \to \infty$.

$\mathcal{M}$	D	N_{sim}	N_{is}	CI_1	CI_2	CI_3	CI_4
2	3	2,000	4,000	[-0.042, 0.023]	[-0.035, 0.028]	[-0.378, -0.237]	[0.091, 0.265]
2	3	10,000	10,000	[-0.014, 0.013]	[-0.017, 0.016]	[-0.366, -0.282]	[0.112, 0.210]
1	4	2,000	4,000	[-0.055, 0.042]	[-0.069, 0.043]	[-0.056, 0.035]	[0.548, 0.643]
1	6	2,000	4,000	[-0.063, 0.046]	[-0.070, 0.046]	[-0.076, 0.014]	[0.645, 0.738]
1	8	2,000	4,000	[-0.073, 0.021]	[-0.076, 0.026]	[-0.106, -0.011]	[0.672, 0.791]
1	10	2,000	4,000	[-0.075, 0.018]	[-0.087, 0.045]	[-0.108, -0.023]	[0.659, 0.833]
1	10	10,000	10,000	[-0.027, 0.020]	[-0.035, 0.023]	[-0.030, 0.001]	[0.774, 0.847]
1	15	10,000	10,000	[-0.039, 0.012]	[-0.047, 0.026]	[-0.049, -0.015]	[0.777, 0.924]
1	20	10,000	10,000	[-0.040, 0.012]	[-0.052, 0.028]	[-0.073, -0.032]	[0.525, 0.975]
1	25	10,000	10,000	[-0.042, 0.003]	[-0.069, 0.024]	[-0.109, -0.067]	[0.598, 1.010]
1	30	10,000	10,000	[-0.050, 0.003]	[-0.064, 0.016]	[-0.143, -0.104]	[0.122, 1.069]

Table B.1: Simulation results of estimating normalizing constants of models 1 and 2 ($\mathcal{M} = 1, 2$ in column 1) using the algorithm in Section 3.4.1: the last four columns are the 95% confidence intervals of $\log(\hat{C}/C)$, where $\hat{C}$ is the estimator and C is the true value, in 100 repeated simulations using the importance sampling with Gaussian tail (CI_1), the importance sampling with t tail with degree of freedom 2 (CI_2), the reciprocal importance sampling with Gaussian tail (CI_3) and the reciprocal importance sampling with t tail degree of freedom 2 (CI_4). N_{sim} is the number of observations and N_{is} is the number of samples from the importance function; D is the dimension of the space.

Proof. Let P be the joint distribution of $(Y_{1:n}, \{\varphi_{K,m}\}_{1\leq m\leq M_n})$ and P_Y be the marginal distribution of $Y_{1:n}$. Note that we do not write explicitly the dependency on K for simplicity of notations. From Corollary 3 following Lemma 6, there exists a sequence of positive constants $\{c_{n,K}\}_{n\geq 1}$ with $\lim_{n\to\infty} c_{n,K} = 0$ such that

$$P_Y\left\{Var[\hat{P}_{K,n,M_n}] \leq c_{n,K} P_{K,n}^2\right\} \to 1 \text{ as } n \to \infty. \tag{B.26}$$

Let $\{b_{n,K}\}_{n\geq 1}$ be a sequence of positive constants such that $\lim_{n\to\infty} b_{n,K}c_{n,K} = \gamma < 1$ and $\lim_{n\to\infty} b_{n,K} = \infty$. For any n, from the Chebychev's inequality,

$$P\left\{|\hat{P}_{K,n,M_n} - P_{K,n}| > b_{n,K}^{1/2} Var[\hat{P}_{K,n,M_n}|Y_{1:n}]^{1/2} \middle| Y_{1:n} = y_{1:n}\right\} \leq b_{n,K}^{-1}.$$

As a consequence,

$$\begin{aligned}
&P\left\{1 - b_{n,K}^{1/2}\frac{Var[\hat{P}_{K,n,M_n}|Y_{1:n}]^{1/2}}{P_{K,n}} \leq \frac{\hat{P}_{K,n,M_n}}{P_{K,n}} \leq 1 + b_{n,K}^{1/2}\frac{Var[\hat{P}_{K,n,M_n}|Y_{1:n}]^{1/2}}{P_{K,n}}\right\}\\
&= 1 - P\left\{|\hat{P}_{K,n,M_n} - P_{K,n}| > b_{n,K}^{1/2} Var[\hat{P}_{K,n,M_n}|Y_{1:n}]^{1/2}\right\}\\
&= 1 - \int P\left\{\frac{|\hat{P}_{K,n,M_n} - P_{K,n}|}{Var[\hat{P}_{K,n,M_n}|Y_{1:n}]^{1/2}} > b_{n,K}^{1/2} \middle| Y_{1:n} = y_{1:n}\right\} dP_Y(y_{1:n})\\
&\geq 1 - b_{n,K}^{-1} \to 1 \text{ as } n \to \infty.
\end{aligned}$$

Together with Equation B.26, we have

$$P\left\{1 - b_{n,K}^{1/2}c_{n,K}^{1/2} \leq \frac{\hat{P}_{K,n,M_n}}{P_{K,n}} \leq 1 + b_{n,K}^{1/2}c_{n,K}^{1/2}\right\} \to 1 \text{ as } n \to \infty. \tag{B.27}$$

By replacing K with K^* in Equation B.27, we have

$$P\left\{1 - b_{n,K^*}^{1/2} c_{n,K^*}^{1/2} \leq \frac{\hat{P}_{K^*,n,M_n}}{P_{K^*,n}} \leq 1 + b_{n,K^*}^{1/2} c_{n,K^*}^{1/2}\right\} \to 1 \text{ as } n \to \infty.$$

We complete the proof by noticing that (1) $P_{K,n}/P_{K^*,n} \to 0$ in probability as $n \to \infty$, (2) $\lim_{n\to\infty} b_{n,K} c_{n,K} \to \gamma < 1$, $\lim_{n\to\infty} b_{n,K^*} c_{n,K^*} \to \gamma < 1$. □

Lemma 6. *Assume that $x_1, \ldots, x_N$ are independent random samples from an unnormalized density $\pi(\theta)$ defined on Ω, the normalizing constant is denoted by $C = \int_\Omega \pi(\theta) d\theta$. $\{g_\phi(\cdot)\}$ is a family of densities indexed by parameter ϕ. Let $\hat{\phi}_N = \mathrm{argmax}_\phi\{\mathrm{g}_\phi(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{N})\}$. Assume that $y_1, \ldots, y_n$ are independent samples from $g_{\hat{\phi}_N}(\cdot)$, thus C can be approximated by either the importance sampling estimator $\hat{C}_{n,N}$ or the locally restricted importance sampling estimator $\hat{C}^{loc}_{n,N}$ defined as*

$$\hat{C}_{n,N} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(y_i)}{g_{\hat{\phi}_N}(y_i)},\ \hat{C}^{loc}_{n,N} = \frac{1}{n\hat{P}_{\Omega_r}}\left[\sum_{j=1}^{n} \frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)} 1_{y_j \in \Omega_r}\right],$$

where $\Omega_r \subset \Omega$ is any bounded subset, $\hat{P}_{\Omega_r} = \frac{1}{N}\sum_{i=1}^{N} 1_{x_i \in \Omega_r}$. Then there exists finite positive constants $\{\alpha, \beta, \gamma\}$, which only depends on $g_{\hat{\phi}_N}(\cdot)$, Ω_r, and the normalized density $\tilde{\pi}(\cdot) := \pi(\cdot)/C$, such that for n, N large enough, the following are approximately true:

$$Var[\hat{C}_{n,N}] = n^{-1}C^2\alpha;\ Var[\hat{C}^{loc}_{n,N}] = n^{-1}C^2\beta + N^{-1}C^2\gamma.$$

A special case is when $n = N$, both these variances can be expressed as $C^2 n^{-1}\delta$, where δ is a finite, positive constant that only depends on $g_{\hat{\phi}_N}(\cdot)$, Ω_r, and the normalized density $\tilde{\pi}(\cdot)$.

Proof. From the law of total variance, we have

$$\begin{aligned} Var\left(\hat{C}_{n,N}\right) &= Var\left[E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(y_i)}{g_{\hat{\phi}_N}(y_i)}\middle|\hat{\phi}_N\right)\right] + E\left[Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(y_i)}{g_{\hat{\phi}_N}(y_i)}\middle|\hat{\phi}_N\right)\right] \\ &= \frac{C^2}{n}E\left[Var\left(\frac{\tilde{\pi}(y)}{g_{\hat{\phi}_N}(y)}\middle|\hat{\phi}_N\right)\right] = \frac{C^2}{n}E\left[\int\frac{\tilde{\pi}^2(y)}{g_{\hat{\phi}_N}(y)}dy - 1\right]. \end{aligned}$$

Next, we calculate the variance of $\hat{C}^{loc}_{n,N}$, which is the ratio of two independent random variables $\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}$ and $\frac{1}{N}\sum_{i=1}^{N}1_{x_i\in\Omega_r}$.

$$\begin{aligned} E\left(\frac{1}{N}\sum_{i=1}^{N}1_{x_i\in\Omega_r}\right) &= \int_{\Omega_r}\tilde{\pi}(x)dx, \\ Var\left[\frac{1}{N}\sum_{i=1}^{N}1_{x_i\in\Omega_r}\right] &= \frac{1}{N}\int_{\Omega_r}\tilde{\pi}(x)dx\left[1-\int_{\Omega_r}\tilde{\pi}(x)dx\right]; \\ E\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}\right) &= E\left[E\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}\middle|\hat{\phi}_N\right)\right] \\ &= \int_{\Omega_r}\pi(x)dx = C\int_{\Omega_r}\tilde{\pi}(x)dx; \\ Var\left[\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}\right] &= Var\left[E\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}\middle|\hat{\phi}_N\right)\right] \\ &\quad + E\left[Var\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}\middle|\hat{\phi}_N\right)\right] \\ &= E\left[\frac{1}{n}\int_{\Omega_r}\frac{\pi^2(y)}{g_{\hat{\phi}_N}(y)}dy\right] - \frac{1}{n}\left(\int_{\Omega_r}\pi(x)dx\right)^2 \\ &= C^2E\left[\frac{1}{n}\int_{\Omega_r}\frac{\tilde{\pi}^2(y)}{g_{\hat{\phi}_N}(y)}dy\right] - C^2\frac{1}{n}\left(\int_{\Omega_r}\tilde{\pi}(x)dx\right)^2. \end{aligned}$$

For n, N large, the delta method gives

$$
\begin{aligned}
Var(\hat{C}^{loc}_{n,N}) &\approx \frac{Var\left[\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}\right]}{\left[E\left(\frac{1}{N}\sum_{i=1}^{N}1_{x_i\in\Omega_r}\right)\right]^2} \\
&\quad + \frac{Var\left[\frac{1}{N}\sum_{i=1}^{N}1_{x_i\in\Omega_r}\right]\left[E\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\pi(y_j)}{g_{\hat{\phi}_N}(y_j)}1_{y_j\in\Omega_r}\right)\right]^2}{\left[E\left(\frac{1}{N}\sum_{i=1}^{N}1_{x_i\in\Omega_r}\right)\right]^4} \\
&= \frac{C^2}{n}\left\{\frac{E\left[\int_{\Omega_r}\frac{\tilde{\pi}^2(y)}{g_{\hat{\phi}_N}(y)}dy\right]}{\left(\int_{\Omega_r}\tilde{\pi}(x)dx\right)^2} - 1\right\} + \frac{C^2}{N}\left[\left(\int_{\Omega_r}\tilde{\pi}(x)dx\right)^{-1} - 1\right].
\end{aligned}
$$

□

Corollary 3. *Using the same notations as in Theorem 7, for any $K < \infty$, there exists a sequence of positive constants $\{c_{n,K}\}_{n\geq 1}$ with $\lim_{n\to\infty} c_{n,K} = 0$ such that*

$$P_Y\left\{Var[\hat{P}_{K,n,M_n}] \leq c_{n,K}P^2_{K,n}\right\} \to 1 \text{ as } n \to \infty. \tag{B.28}$$

Proof. From Lemma 6, we have, for any given $\boldsymbol{y}_{1:n}$, $Var[\hat{P}_{K,n,M_n}] = M_n^{-1}P^2_{K,n}\delta$, where $\delta > 0$ is a constant depending on the normalized posterior distribution and the importance function. Therefore,

$$P_Y\left\{Var[\hat{P}_{K,n,M_n}] \leq c_{n,K}P^2_{K,n}\right\} = P_Y\left[\delta M_n^{-1} \leq c_{n,K}\right] \to 1 \text{ as } n \to \infty,$$

if we choose, for example, $c_{n,K} = O(M_n^{-1/2})$. □

C

Theoretical Properties of PMCMC

C.1 Proofs and Supplementary Material

C.1.1 Proof of Theorem 4

Before proving the theorem, we first prove the following lemma.

Lemma 7. *Samples obtained according to Algorithm 1 given below are draws from π.*

Proof. $\{x_0^{(j)} : j = 1, \ldots, n\}$ is a standard Markov chain Monte Carlo sampler by construction of the transition kernel T. Since Calderhead's algorithm is valid for versions 1 and 2 of the weighting scheme, we know step 1 draws samples from π given that $x_0^{(j)}$ follows π. Combining these two arguments we establish that the samples $\{y_i^{(j)} : i = 1, 2, \ldots, N; j = 1, 2, \ldots n\}$ all have marginal distribution π. □

Proof. Let $y_i^{(j)}$ be a point drawn according to Algorithm 1. Let $x^{(j)} = \{x_0^{(j)}, \ldots, x_M^{(j)}\}$, so that

$$\mu_h = E\left\{h(y_i^{(j)})\right\} = E\left[E\left\{h(y_i^{(j)}) \mid x^{(j)}\right\}\right] = E\left\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}.$$

□

C.1.2 Proof of Theorem 5

Proof. Let $x^{(j)} = \{x_0^{(j)}, \ldots, x_M^{(j)}\}$ and $y^{(j)} = \{y_1^{(j)}, \ldots, y_N^{(j)}\}$.

$$\operatorname{var}\left\{\frac{1}{nN}\sum_{j=1}^{n}\sum_{i=1}^{N} h(y_i^{(j)})\right\} = \frac{1}{n^2N^2}\sum_{j=1}^{n}\operatorname{var}\left\{\sum_{i=1}^{N} h(y_i^{(j)})\right\} + \frac{2}{n^2N^2}\sum_{\substack{j=1\\k=1\\j<k}}^{n}\operatorname{cov}\left\{\sum_{i=1}^{N} h(y_i^{(j)}), \sum_{i=1}^{N} h(y_i^{(k)})\right\}.$$

By the law of total variance, for the first of these terms we have

$$\begin{aligned}\frac{1}{n^2N^2}\sum_{j=1}^{n}\operatorname{var}\left\{\sum_{i=1}^{N} h(y_i^{(j)})\right\} &\geq \frac{1}{n^2N^2}\sum_{j=1}^{n}\operatorname{var}\left[E\left\{\sum_{i=1}^{N} h(y_i^{(j)}) \mid x^{(j)}\right\}\right]\\ &= \frac{1}{n^2}\sum_{j=1}^{n}\operatorname{var}\left\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}.\end{aligned}$$

For the second, we will show that

$$\frac{2}{n^2N^2}\mathrm{cov}\left\{\sum_{i=1}^{N} h(y_i^{(j)}), \sum_{i=1}^{N} h(y_i^{(k)})\right\} = \frac{2}{n^2}\mathrm{cov}\left\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)}), \sum_{i=0}^{M} w(x_i^{(k)})h(x_i^{(k)})\right\}, \quad (j \neq k).$$

Assume $j < k$ without loss of generality. By the law of total covariance we then have

$$\begin{aligned}
&\mathrm{cov}\left\{\frac{1}{N}\sum_{i=1}^{N} h(y_i^{(j)}), \frac{1}{N}\sum_{i=1}^{N} h(y_i^{(k)})\right\} \\
&= \mathrm{cov}\left[\frac{1}{N}E\left\{\sum_{i=1}^{N} h(y_i^{(j)}) \mid x^{(j)}, x^{(k)}\right\}, \frac{1}{N}E\left\{\sum_{i=1}^{N} h(y_i^{(k)}) \mid x^{(j)}, x^{(k)}\right\}\right] \\
&\quad + E\left[\mathrm{cov}\left\{\frac{1}{N}\sum_{i=1}^{N} h(y_i^{(j)}), \frac{1}{N}\sum_{i=1}^{N} h(y_i^{(k)}) \mid x^{(j)}, x^{(k)}\right\}\right] \\
&= \mathrm{cov}\left\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)}), \sum_{i=0}^{M} w(x_i^{(k)})h(x_i^{(k)})\right\},
\end{aligned}$$

where the last equality follows from the conditional independence structure of $y^{(j)}$ from all the other samples and proposals given $x^{(j)}$. Summarizing these results, we can conclude that

$$\mathrm{var}\left\{\frac{1}{nN}\sum_{j=1}^{n}\sum_{i=1}^{N} h(y_i^{(j)})\right\} \geq \mathrm{var}\left\{\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}.$$

$\square$

C.1.3 Proof of Proposition 1

Proof. From the derivations in proof of theorem 5, we have

$$
\begin{aligned}
&\operatorname{var}\left\{\frac{1}{nN}\sum_{j=1}^{n}\sum_{i=1}^{N} h(y_i^{(j)})\right\} - \operatorname{var}\left\{\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\} \\
&= \frac{1}{n^2N^2}\sum_{j=1}^{n} E\left[\operatorname{var}\left\{\sum_{i=1}^{N} h(y_i^{(j)}) \mid x^{(j)}\right\}\right] = \frac{1}{n^2N} E\left[\operatorname{var}\left\{h(y_i^{(j)}) \mid x^{(j)}\right\}\right] \\
&= \frac{1}{n^2N}\sum_{j=1}^{n} E\left[\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})^2 - \left\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}^2\right].
\end{aligned}
$$

Therefore, the relative efficiency gain by using locally weighted parallel Markov chain Monte Carlo instead of parallel Markov chain Monte Carlo is

$$
\begin{aligned}
&\frac{\operatorname{var}\left\{\frac{1}{nN}\sum_{j=1}^{n}\sum_{i=1}^{N} h(y_i^{(j)})\right\} - \operatorname{var}\left\{\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}}{\operatorname{var}\left\{\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}} \\
&= \frac{1}{N}\frac{\sum_{j=1}^{n} E\left[\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})^2 - \left\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}^2\right]}{\sum_{j=1}^{n}\operatorname{var}\left\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\right\}} = \frac{1}{N}\frac{E\left\{\bar{g} - (\bar{h})^2\right\}}{\operatorname{var}(\bar{h})}.
\end{aligned}
$$

The last equality is true since at equilibrium, for $j = 1, \ldots, n$, $E(\bar{g}) = E\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})^2\}$, $E\{(\bar{h})^2\} = E[\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\}^2]$, $\operatorname{var}[\bar{h}] = \operatorname{var}\{\sum_{i=0}^{M} w(x_i^{(j)})h(x_i^{(j)})\}$; thus we can obtain moment estimators of $E(\bar{g}), E\{(\bar{h})^2\}, \operatorname{var}(\bar{h})$ using all the proposals and weights.

□

C.1.4 Proof of Theorem 6

Proof.

$$\begin{aligned}\frac{\sigma^2}{ESS} &= \mathrm{var}\left\{\frac{1}{n}\sum_{j=1}^{n}\bar{x}^{(j)}\right\} = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{var}(\bar{x}^{(j)}) + \frac{2}{n^2}\sum_{j<k}^{n}\mathrm{cov}(\bar{x}^{(j)}, \bar{x}^{(k)}) \\ &= \frac{1}{n}\mathrm{var}(\bar{x}) + \frac{2}{n}\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\gamma_k\mathrm{var}(\bar{x}) = \frac{\mathrm{var}(\bar{x})}{n}\left\{1+2\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\gamma_k\right\},\end{aligned}$$

where the second inequality follows from stationarity. By the Cesàro summability theorem

$$\lim_{n\to\infty}\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)\gamma_k = \sum_k \gamma_k.$$

For sufficiently large n, we therefore substitute the right hand side of this equality into the expressions derived above. Rearranging the terms will give the desired result.

□

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Akopian, D., Dalal, K., Shen, K., Duong, F., & Shan, S. (2013a). SecYEG activates GTPases to drive the completion of cotranslational protein targeting. *The Journal of Cell Biology*, 200(4), 397–405.

Akopian, D., Shen, K., Zhang, X., & Shan, S. (2013b). Signal recognition particle: an essential protein-targeting machine. *Annual Review of Biochemistry*, 82, 693–721.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3), 817–858.

Ataide, S. F., Schmitz, N., Shen, K., Ke, A., Shan, S., Doudna, J. A., & Ban, N. (2011). The crystal structure of the signal recognition particle in complex with its receptor. *Science*, 331, 881–886.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on information theory*, 44(6).

Baum, L. E. & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6), 1554–1563.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1), 164–171.

Bauwens, L., Dufays, A., & Rombouts, J. V. (2014). Marginal likelihood for Markov-switching and change-point GARCH models. *Journal of Econometrics*, 178, 508–522.

Bickel, P. J., Ritov, Y., & Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4), 1614–1635.

Biernacki, C., Celeux, G., & Govaert, G. (1998). Assessing a mixture model for clustering with the integrated classification likelihood. Archived article, available at https://hal.inria.fr/inria-00073163/.

Blanco, M. & Walter, N. G. (2010). Analysis of complex single-molecule FRET time trajectories. *Methods in Enzymology*, 472, 153—178.

Boys, R. J. & Henderson, D. A. (2004). A Bayesian approach to DNA sequence segmentation. *Biometrics*, 60(3), 573–581.

Bronson, J. E., Fei, J., Hofman, J. M., Ruben L. Gonzalez, J., & Wiggins, C. H. (2009). Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophysical Journal*, 97, 3196–3205.

Bulla, J., Bulla, I., & Nenadic, O. (2010). hsmm - an R package for analyzing hidden semi-Markov models. *Computational statistics and data analysis*, 54, 611–619.

Calderhead, B. (2014). A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49), 17408–17413.

Cappe, O., Moulines, E., & Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics.

Celeux, G. & Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, 23, 541–564.

Chambaz, A., Garivier, A., & Gassiat, E. (2009). A MDL approach to HMM with Poisson and Gaussian emissions: Application to order indentification. *Journal of Statistical Planning and Inference*, 139, 962–977.

Chambaz, A. & Rousseau, J. (2005). *Nonasymptotic bounds for Bayesian order identification with application to mixtures*. Citeseer.

Chen, J. & Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *The Canadian Journal of Statistics*.

Chen, J. & Khalili, A. (2012). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*.

Chen, J. & Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *The Annals of Statistics*, 37, 2523—2542.

Chen, J. & Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*.

Chen, J., Tan, X., & Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*.

Chen, M.-H. & Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25, 1563–1594.

Chen, Y., Shen, K., Shan, S.-O., & Kou, S. C. (2016). Analyzing single-molecule protein transportation experiments via hierarchical hidden Markov models. *Journal of the American Statistical Association*.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432), 1313–1321.

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453), 270–281.

Chib, S. & Jeliazkov, I. (2005). Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, 59(1), 30–44.

Dahan, M., Deniz, A. A., Ha, T., Chemla, D. S., Schultz, P. G., & Weiss, S. (1999). Ratiometric measurement and identification of single diffusing molecules. *Chemical Physics*, 247, 85–106.

de Gunst, M. C. M. & Shcherbakova, O. (2008). Asymptotic behavior of Bayes estimator for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics*, 17, 342–356.

de Valpine, P. (2008). Improved estimation of normalizing constants from Markov chain Monte Carlo output. *Journal of Computational and Graphical Statistics*, 17(2), 333–351.

Delmas, J.-F. & Jourdain, B. (2009). Does waste recycling really improve the multi-proposal Metropolis-Hastings algorithm? An analysis based on control variates. *Journal of Applied Probability*, 46(4), 938–959.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1), 1–38.

DiCiccio, T. J., Kass, R. E., Raftery, A., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439), 903–915.

Douc, R. & Robert, C. P. (2011). A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. *The Annals of Statistics*, 39(1), 261–277.

Du, C., Kao, C. L., & Kou, S. C. (2016). Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association.*

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2), 216–222.

Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6, 361–365.

Ephraim, Y. & Merhav, N. (2002). Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6), 1518–1569.

Estrozi, L. F., Boehringer, D., Shan, S., Ban, N., & Schaffitzel, C. (2011). Cryo-EM structure of the E. coli translating ribosome in complex with SRP and its receptor. *Nature Structural & Molecular Biology*, 18(1), 88–90.

Fan, Y., Sisson, S. A., Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). Reversible jump MCMC. *Handbook of Markov Chain Monte Carlo*, (pp. 67–92).

Ferguson, J. (1980). Application of hidden Markov models to text and speech. *Princeton, NJ), IDA-CRD.*

Finesso, L. (1990). Consistent estimation of the order for Markov and hidden Markov chains. *Ph.D. dissertation, Univ. Maryland, College Park.*

Frenkel, D. (2004). Speed-up of Monte Carlo simulations by sampling of rejected states. *Proceedings of the National Academy of Sciences*, 101(51), 17571–17575.

Frenkel, D. (2006). Waste-recycling Monte Carlo. In M. Ferrario, G. Ciccotti, & K. Binder (Eds.), *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, volume 703 of *Lecture Notes in Physics* (pp. 127–137). Springer Berlin Heidelberg.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer-Verlag New York, Inc.

Fuh, C. D. (2003). SPRT and CUSUM in hidden Markov models. *The Annals of Statistics*, 31, 942–977.

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. In *Annales de l'IHP Probabilités et statistiques*, volume 38 (pp. 897–906).

Gassiat, E. & Boucheron, S. (2003). Optimal error exponents in hidden Markov model order estimation. *IEEE Transactions on Information Theory*, 48, 964–980.

Gassiat, E. & Keribin, C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM: Probability and Statistics*, 4, 25–52.

Gassiat, E. & Rousseau, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20, 2039–2075.

Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(3), 501–514.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis, 3rd Edition*. Chapman and Hall/CRC.

Gelman, A. & Meng, X. (1991). A note on bivariate distributions that are conditionally normal. *The American Statistician*, 45(2), 125–126.

Gelman, A. & Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2), 163–185.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57, 1317–1339.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Technical Report No. 568, School of Statistics, University of Minnesota*.

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1), 9–42.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4), 711–732.

Green, P. J. & Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, 155(3), 1391–1403.

Green, P. J. & Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American statistical association*, 97(460), 1055–1070.

Greenfeld, M., Pavlichin, D. S., Mabuchi, H., & Herschlag, D. (2012). Single molecule analysis research tool (SMART): an integrated approach for analyzing single molecule data. *PLos One*, 7.

Halic, M., Becker, T., Pool, M. R., Spahn, C. M. T., Grassucci, R. A., Frank, J., & Beckmann, R. (2004). Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature*, 427, 808–814.

Halic, M., Gartmann, M., Schlenker, O., Mielke, T., Pool, M. R., Sinning, I., & Beckmann, R. (2006). Signal recognition particle receptor exposes the ribosomal translocon binding site. *Science*, 312(5774), 745–747.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton Univ. Press.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.

Hawkins, D. S., Allen, D. M., & Stromberg, A. J. (2001). Determining the number of components in mixtures of linear models. *Computational Statistics & Data Analysis*, 38, 15–48.

Huang, T., Peng, H., & Zhang, K. (2013). Model selection for Gaussian mixture models. *arXiv:1301.3558*.

Hui, F. K., Warton, D. I., & Foster, S. D. (2015). Order selection in finite mixture models: complete or observed likelihood information criteria? *Biometrika*, (pp. 1–7).

Hung, Y., Wang, Y., Zarnitsyna, V., Zhu, C., & Wu, C. J. (2013). Hidden Markov models with applications in cell adhesion experiments. *Journal of the American Statistical Association*, 108(504), 1469–1479.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2), 295–311.

Jeffries, N. O. (2003). A note on 'testing the number of components in a normal mixture'. *Biometrika*, 90(4), 991–994.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82, 35–45.

Kalman, R. E. & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Trans. ASME, Ser. D, J. Basic Eng*, (pp. 109).

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52(2), 93–100.

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

Keenan, R. J., Freymann, D. M., Stroud, R. M., & Walter, P. (2001). The signal recognition particle. *Annual Review of Biochemistry*, 70, 755–775.

Keller, B. G., Kobitski, A., Jaš̌chke, A., Nienhaus, G., & Noe, F. (2014). Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *Journal of the American Chemical Society*, 136, 4534—4543.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *The Indian Journal of Statistics, Series A*, 62(1), 49–66.

Kieffer, J. C. (1993). Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Transactions on Information Theory*, 39, 893–902.

König, S. L., Hadzic, M., Fiorini, E., Börner, R., Kowerko, D., Blanckenhorn, W. U., & Sigel, R. K. O. (2013). BOBA FRET: Bootstrap-based analysis of single-molecule FRET data. *PLos One*, 8.

Kou, S. C., Zhou, Q., & Wong, W. H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, 34(4), 1581–1619.

Leroux, B. G. (1992a). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3), 1350–1360.

Leroux, B. G. (1992b). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40, 127–143.

Leroux, B. G. & Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, 48(2), 545–558.

Liu, C.-C. & Narayan, P. (1994). Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Transactions on Information Theory*, 40, 1167–1180.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer-Verlag New York, Inc.

Liu, J. S., Liang, F., & Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449), 121–134.

Liu, Y., Park, J., Dahmen, K. A., Chemla, Y. R., & Ha, T. (2010). A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *Journal of Physical Chemistry*, 114, 5386–5403.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). *Molecular Cell Biology*. New York: W. H. Freeman, 4 edition.

MacDonald, I. L. & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*, volume 70. Chapman & Hall, Monographs on Statistics and Applied Probability.

MacKAY, R. J. (2002). Estimating the order of a hidden Markov model. *Canadian Journal of Statistics*, 30, 573–589.

McKinney, S. A., Joo, C., & Ha, T. (2006). Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophysical Journal*, 91, 1941–1951.

McLachlan, G. & Peel, D. (2005). *Finite Mixture Models*. Wiley Series in Probability and Statistics.

Meng, X.-L. & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6, 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.

Moerner, W. E. (2002). A dozen years of single-molecule spectroscopy in physics, chemistry, and biophysics. *The Journal of Physical Chemistry B*, 106, 910–927.

Müller, U. K. (2014). HAC corrections for strongly autocorrelated time series. *Journal of Business and Economic Statistics*, 32, 311–322.

Neal, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111, 194–203.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4), 353–366.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3), 705–767.

Neal, R. M. (2005). Estimating ratios of normalizing constants using linked importance sampling. *Technical Report No. 0511, Department of Statistics, University of Toronto.*

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Chapter 5, Handbook of Markov Chain Monte Carlo.*

Newton, M. A. & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(1), 3–48.

Nie, S. & Zare, R. N. (1997). Optical detection of single molecules. *Annual Review of Biophysics and Biomolecular Structure*, 26, 567–596.

Nyathi, Y., Wilkinson, B. M., & Pool, M. R. (2013). Co-translational targeting and translocation of proteins to the endoplasmic reticulum. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1833, 2392–2402.

Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numer. Math*, 55, 137–157.

Oh, M.-S. & Berger, J. O. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association*, 88(422), 450–456.

Peluso, P., Shan, S., Nock, S., Herschlag, D., & Walter, P. (2001). Role of SRP RNA in the GTPase cycles of Ffh and FtsY. *Biochemistry*, 40, 15224–15233.

Petris, G. & Tardella, L. (2007). New perspectives for estimating normalizing constants via posterior simulation. *Technical Report, DSPSA, Sapienza Universitá di Roma.*

Pool, M. R., Stumm, J., Fulga, T. A., Sinning, I., & Dobberstein, B. (2002). Distinct modes of signal recognition particle interaction with the ribosome. *Science*, 297(5585), 1345–1348.

Qian, H. & Kou, S. C. (2014). Statistics and related topics in single-molecule biophysics. *Annual Review of Statistics and Its Application*, 1, 465–492.

Qin, Z. S. & Liu, J. S. (2001). Multipoint Metropolis method with application to hybrid Monte Carlo. *Journal of Computational Physics*, 172, 827–840.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.

Rapoport, T. A. (1991). Protein transport across the endoplasmic reticulum membrane: Facts, models, mysteries. *The FASEB Journal*, 5, 2792–2798.

Rapoport, T. A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450, 663–669.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465–471.

Robert, C. P., Rydén, T., & Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1), 57–75.

Roberts, G. O. & Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4), 341–363.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73.

Roy, R., Hohng, S., & Ha, T. (2008). A practical guide to single-molecule FRET. *Nature Methods*, 5(6), 507–516.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12, 1151–1172.

Rydén, T. (1995). Estimating the order of hidden Markov models. *Statistics*, 26, 345—354.

Rydén, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, 3(4), 659–688.

Rydén, T., Teräsvirta, T., & Asbrink, S. (1998). Stylized facts of daily returns series and the hidden Markov model. *Journal of Applied Economics*, 13, 217–244.

Schmid, S., Gotz, M., & Hugel, T. (2016). Experiment-friendly kinetic analysis of single molecule data in and out of equilibrium. *arXiv:1605.08612 [q-bio.QM]*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Shao, J. (1989). Monte Carlo approximations in Bayesian decision theory. *Journal of the American Statistical Association*, 84(407).

Shen, K., Arslan, S., Akopian, D., Ha, T., & Shan, S. (2012). Activated GTPase movement on an RNA scaffold drives co-translational protein targeting. *Nature*, 492, 271–275.

Shen, K. & Shan, S. (2010). Transient tether between the SRP RNA and SRP receptor ensures efficient cargo delivery during cotranslational protein targeting. *Proceedings of the National Academy of Sciences*, 107(17), 7698–7703.

Shen, K., Wang, Y., Hwang Fu, Y.-H., Zhang, Q., Feigon, J., & Shan, S. (2013). Molecular mechanism of GTPase activation at the signal recognition particle (SRP) RNA distal end. *The Journal of Biological Chemistry*, 288, 36385–36397.

Spezia, L. (2010). Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis*, 31(1), 1–11.

Steele, R. J., Raftery, A. E., & Emond, M. J. (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling. *Journal of Computational and Graphical Statistics*, 15(3), 712–734.

Swendsen, R. H. & Wang, J. S. (1986). Replica Monte Carlo simulation of spin glasses. *Physical Review Letters*, 57, 2607–2609.

Tamarat, P., Maali, A., Lounis, B., & Orrit, M. (2000). Ten years of single-molecule spectroscopy. *The Journal of Physical Chemistry A*, 104(1), 1–16.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528–540.

Tjelmeland, H. (2004). Using all Metropolis-Hastings proposals to estimate mean values. *Technical Report, Norwegian University of Science and Technology, Trondheim, Norway.*

van de Meent, J. W., Bronson, J. E., Wiggins, C. H., & Gonzalez Jr., R. L. (2014). Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. *Biophysical Journal*, 106, 1327–1337.

van de Meent, J.-W., Bronson, J. E., Wood, F., Jr., R. L. G., & Wiggins, C. H. (2016). Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data. *arXiv:1305.3640 [stat.ML].*

Voigts-Hoffmann, F., Schmitz, N., Shen, K., Shan, S., Ataide, S. F., & Ban, N. (2013). The structural basis of FtsY recruitment and GTPase activation by SRP RNA. *Molecular Cell*, 52, 643–654.

Walker, A. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 80–88).

Wang, Y. & Bickel, P. J. (2015). Likelihood-based model selection for stochastic block models. *arXiv preprint arXiv:1502.02069.*

Watkins, L. P. & Yang, H. (2005). Detection of intensity change points in time-resolved single-molecule measurements. *The Journal of Physical Chemistry B*, 109, 617—628.

Weiss, S. (2000). Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Structural Biology*, 7, 724–729.

Windham, M. P. & Cutler, A. (1992). Information ratios for validating mixture analysis. *Journal of the American Statistical Association*, 87(420), 1188–1192.

Xie, X. S. & Lu, H. P. (1999). Single-molecule enzymology. *The Journal of Biological Chemistry*, 274(23), 15967–15970.

Xie, X. S. & Trautman, J. K. (1998). Optical studies of single molecules at room temperature. *Annual Review of Physical Chemistry*, 49, 441–480.

Zhang, X., Jantama, K., Moore, J. C., Jarboe, L. R., Shanmugam, K. T., & Ingrama, L. O. (2009a). Metabolic evolution of energy-conserving pathways for succinate production in escherichia coli. *Proceedings of the National Academy of Sciences*, 106, 20180–20185.

Zhang, X., Kung, S., & Shan, S. (2008). Demonstration of a multi-step mechanism for assembly of the SRP-SR receptor complex: Implications for the catalytic role of SRP RNA. *Journal of Molecular Biology*, 381(3), 581–593.

Zhang, X., Schaffitzel, C., Ban, N., & Shan, S. (2009b). Multiple conformational switches in a GTPase complex control co-translational protein targeting. *Proceedings of the National Academy of Sciences*, 106, 1754–1759.

Ziv, J. & Merhav, N. (1992). Estimating the number of states of a finite-state source. *IEEE Transactions on Information Theory*, 38, 61–65.