HIDDEN MARKOV MODEL WITH APPLICATION IN CELL ADHESION EXPERIMENT AND BAYESIAN CUBIC SPLINES IN COMPUTER EXPERIMENTS

A Thesis
Presented to
The Academic Faculty

by

Yijie Dylan Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2013

HIDDEN MARKOV MODEL WITH APPLICATION IN CELL ADHESION EXPERIMENT AND BAYESIAN CUBIC SPLINES IN COMPUTER EXPERIMENTS

Approved by:

Dr. C. F. Jeff Wu, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ying Hung
Department of Statistics and Biostatistics
Rutgers University

Dr. Jye-Chyi Lu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roshan Joseph Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: June 27th 2013

To my parents.

ACKNOWLEDGEMENTS

I would like to thank all the people who have taught me, helped me, and inspired me during my doctoral study.

Especially, I would like to express my deepest gratitude to my advisor, Professor Jeff Wu. He has guided me through my four-year doctoral study with incredible patience, and provided me with tremendous support and encouragement. I am not only inspired by his passion and talents on academic research, but also influenced by his professional ethics and life attitudes, from which I will benefit for my lifetime.

I am extremely grateful to Professor Ying Yung. She has taken care of me with great mentorship on both academics and personal life. I have benefited so much from her broad knowledge and deep insights that I consider myself very fortunate.

I would like to thank Dr. Cheng Zhu for his valuable mentorship of my research and career. I want to acknowledge my sincere gratitude to Professor Yajun Mei, who not only has served on my dissertation committee, but also gave me so much help and guidance on my career. My thanks also go to Professor Roshan Joseph Vengazhiyil and Professor Jye-Chyi Lu for serving on my dissertation committee and providing so many insightful suggestions.

I also want to thank my lab members Heng Su, Li Gu and Rui Tuo. It is such an honor to work with these talented people and gain their friendship.

At last but not least, I would like to thank my parents. I cannot ask more from my parents, who have given me unconditional and unlimited love and support despite of the ups and downs of my life. I dedicate my dissertation to my parents, hoping they would feel happy and proud.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis consists of two parts. The first part focuses on the hidden Markov model (HMM) with application in cell adhesion experiment, and the second part on the Bayesian cubic spline in computer experiment.

The first part of this thesis contains two works on the hidden Markov models. In Chapter 1, a new model selection method is proposed for hidden Markov models. In Chapter 2, we implement HMM in the cell adhesion experiment. The second part of this thesis introduces a Bayesian cubic spline in computer experiment. Chapter 3 proposes the estimation of Bayesian cubic spline and compares it with two other methods.

Chapter 1 deals with HMM model selection. Estimation of the number of hidden states is challenging in hidden Markov models. Motivated by the analysis of a specific type of cell adhesion experiments, a new framework based on hidden Markov model and double penalized order selection is proposed. The order selection procedure is shown to be consistent in estimating the number of states. A modified Expectation-Maximization algorithm is introduced to efficiently estimate parameters in the model. Simulations show that the proposed framework outperforms existing methods. Applications of the proposed methodology to real data demonstrate the accuracy of estimating receptor-ligand bond lifetimes and waiting times which are essential in kinetic parameter estimation. This is joint work with Dr. Ying Hung, Dr. Jeff Wu, Dr. Veronica Zarnitsina and Dr. Cheng Zhu.

Chapter 2 shows the application of HMM in cell adhesion experiments. Abrupt

reduction/resumption of thermal fluctuations of a force probe has been used to identify association/dissociation events of protein-ligand bonds. We show that off-rate of molecular dissociation can be estimated by the analysis of the bond lifetime while the on-rate of molecular association can be estimated by the analysis of the waiting time between two neighboring bond events. However, the analysis relies heavily on subjective judgments and is time-consuming. To automate the process of mapping out bond events from thermal fluctuation data, we develop a hidden Markov model (HMM)-based method. The HMM method represents the bond state by a hidden variable with two values: bound and unbound. The bond association/dissociation is visualized and pinpointed. We apply the method to analyze a key receptor-ligand interaction in the early-stage of hemostasis and thrombosis: the von Willebrand factor (VWF) binding to platelet glycoprotein Ib (GPIbα). The numbers of bond lifetime and waiting time estimated by the HMM are much more than those estimated by a descriptive statistical method from the same set of raw data. The kinetic parameters estimated by the HMM are in excellent agreement with those by a descriptive statistical analysis, but have much smaller errors for both wild-type and two mutant VWF-A1 domains. Thus, the computerized analysis allows us to speed up the analysis and improve the quality of estimates of receptor-ligand binding kinetics. This is joint work with Arnold Ju, Dr. Ying Hung, Dr. Jeff Wu and Dr. Cheng Zhu.

Chapter 3 is concerned with prediction of a deterministic response function y at some untried sites given values of y at a chosen set of design sites. The intended application is to computer experiments in which y is the output from a computer simulation and each design site represents a particular configuration of the input variables. A Bayesian version of the cubic spline method commonly used in numerical analysis is proposed, in which the random function that represents prior uncertainty about y is taken to be a specific stationary Gaussian process. An MCMC procedure is given for updating the prior given the observed y values. Simulation examples and

a real data application are given to compare the performance of the Bayesian cubic spline with that of two existing methods. This is joint work with Dr. Jeff Wu.

CHAPTER I

HIDDEN MARKOV MODELS WITH APPLICATIONS IN CELL ADHESION EXPERIMENTS

1.1 Introduction

Cell adhesion plays an important role in many physiological and pathological processes (Dustin et al. 2001). It is mediated by specific interactions between receptors on one cell and corresponding ligands on another cell. This work is motivated by newly developed method, called thermal fluctuation assay (Chen et al. 2008), which allows a real-time monitoring of receptor-ligand interactions.

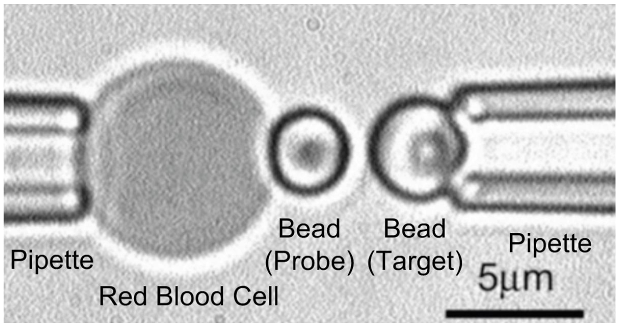

Figure 1: Illustration of the Biomembrane Force Probe

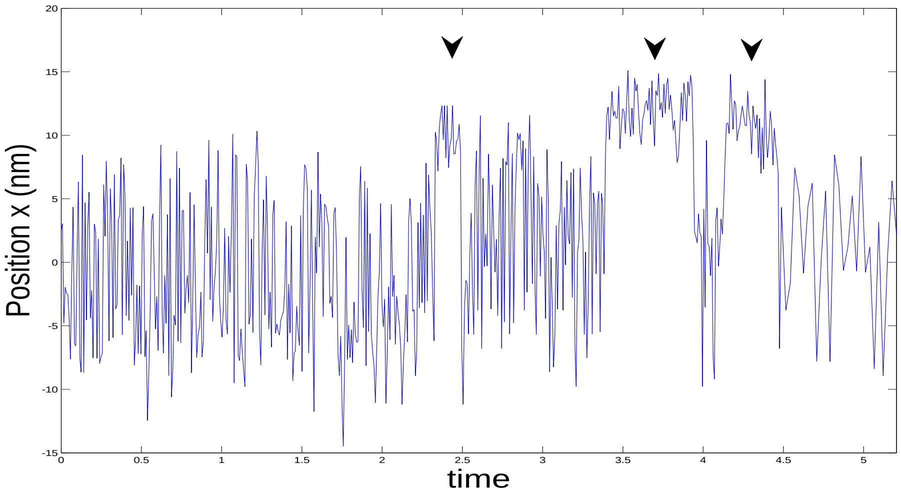

Figure 2: Observations from a Thermal Fluctuation Experiment

In thermal fluctuation assay, red blood cell (RBC) is used as an adhesion sensor. Receptor surface (target bead on the right of Figure 1) and ligand surface (probe bead linked to a RBC on the left of Figure 1) are brought into zero distance contact, allowing receptor-ligand bonds to form by thermal fluctuation of RBC. When a bond forms, thermal fluctuations are reduced. Thus, decrease/resumption of thermal fluctuations of a biomembrane force probe (left bead linked to RBC in Figure 1) pinpoints association/dissociation of receptor-ligand bonds. Accurate estimation of the instants of bond formation and dissociation is essential because they form the basis for subsequent estimation of the kinetic parameters for specific receptor-ligand interaction (Chen et al. 2008). The position of the probe bead is tracked by image analysis software to produce the data shown in Figure 2. In this Figure, horizontal position of the left edge of the probe bead is plotted versus time. Bond formation is equivalent to adding a molecular spring in parallel to the force transducer spring to stiffen the system (Marshall et al., 2006; Wu et al., 2005). Therefore, the fluctuation decreases when a receptor-ligand bond forms and resumes when the bond dissociates.

The objective of this study is to identify association and dissociation points for receptor-ligand bonds. It can be challenging because these points are not directly observable and can only be detected through the variance changes in thermal fluctuations. Moreover, the thermal fluctuations are independently distributed given their binding status (e.g., binding or not), but the transition from one status to another can be dependent. For example, in some receptor-ligand systems (Zarnitsyna et al. 2007; Hung et al. 2008), the chance of having a binding in the next contact is increased (or decreased) if there is a binding in the immediate past. Because of the dependence, standard approaches such as change point techniques (Carlstein et al. 1994; Hawkins and Zamba 2005) are not directly applicable. Identifying association/dissociation points becomes even more difficult when the recorded data contains more than one type of bonds and the number of types is unknown which is quite common in cell-cell adhesion interactions. In general, different types of receptor-ligand bonds are associated with different fluctuation decreases, depending on the stiffness of the molecules. We can thus classify the association and dissociation points into different bond types according to different levels of reduction in thermal fluctuation. The difficulty is that the levels of reduction are unknown and have to be estimated from data. Thus, there are two issues involved in this study. The first is to accurately estimate the number of bond types in the process and the second is to identify the association and dissociation points for each type of bonds.

The existing approach for analyzing thermal fluctuation assay data is to calculate the moving standard deviation of the thermal fluctuation data (Chen et al. 2008). This approach, though intuitive, is not robust to the size of moving windows and limited to the study of one type of bonds. To overcome these problems and address the two foregoing issues, a new framework based upon hidden Markov models (HMM) (Rabiner 1989; Bickel et al. 1998; Cappé et al. 2005) and an order selection procedure is proposed. This framework provides a systematic approach to simultaneously determine the number of bond types and identify the association and dissociation points.

The probe fluctuates with different variations that correspond to different underlying binding states. These unobservable states are not assumed to be independent, but rather to have a Markovian structure so that the cell memory effects can be captured. Therefore, the proposed framework uses hidden states in HMM to represent the binding status. Given the hidden states, probe locations are assumed to be independent and normally distributed with some unknown parameters that capture the variations associated with different bond types.

HMMs have proven to be very useful in many areas (Rabiner 1989, Scott, James, and Sugar 2005, Yuan and Kendziorski 2006) and theoretical properties of HMMs, given the number of state is known, have been extensively studied (Leroux 1992, Bickel, Ritov and Rydén 1998, Cappé, Moulines, and Rydé 2005). However, the unknown number of bond types in the current experiment requires the study of a new problem, namely, estimation of the number of states. A standard approach would be to use likelihood ratio tests with the likelihood ratio asymptotically distributed as a χ^2 random variable. However, this result is not true for HMMs because, if the null hypothesis is true, then the parameters are not uniquely identified under the alternative (Gassiat and Kéribin 2000, Robert et al., 2000). Other order selection methods, such as AIC (Akaike 1974) and BIC (Schwarz 1978), are commonly used in practice. Examples can be found in Leroux and Puterman (1992), Hughes and Guttorp (1994), Albert et al. (1994), and Wang and Puterman (1999). However, these methods have not been theoretically justified in the context of HMMs (MacDonald and Zucchini, 1997). Although some theoretical studies has been developed along this line, such as the minimum distance estimator (Chen and Kalbfleisch 1996, MacKay 2002) and BIC-type of penalized approaches (Csiszár and Shields 2000, Gassiat and Boucheron 2003, Chambaz et al. 2009), the order selection problem has not yet been satisfactorily resolved for HMMs. A new order selection method is proposed in this paper and its consistency is addressed. The merits of this approach are borne out in a simulation study comparing with existing methods.

Although the proposed order selection approach in HMMs is motivated by the study of cell adhesion experiment, it has applications in many areas, including signal processing (Kaleh and Vallet 1994, Chambaz et al. 2009), environmental science, and bioinformatics (Koski 2001). In these problems the number of underlying states is often unknown. Efficient estimation of the order can improve the prediction accuracy and provide valuable scientific information. For example, MacKay (2002) proposed an HMM to model lesions experienced on the brain stem given an unobservable disease state in the study of multiple sclerosis. Our proposed method would be useful in estimating the number of hidden disease states. In another example, Hughes and Guttorp (1994) model the rainfall process given unobserved weather states. The proposed method can be applied to estimate the unknown number of weather states. In the study of heart rate variability in sleeping neonates (Clairambault et al. 1992), the proposed method is readily applicable to characterize the number of periods in the neonate sleep.

The remainder of this article is organized as follows. The existing method and some preliminary analysis results for a thermal fluctuation experiment are presented in Section 1.2. The hidden Markov model approach is developed and an order selection procedure is introduced in Section 1.3. The order selection is shown to be asymptotically consistent in estimation. An efficient algorithm, called expectation conditional maximization, is used for maximum likelihood estimation. In Section 1.4, simulations are presented to demonstrate the performance of the proposed approach. In Section 1.5, the proposed approach is applied to the analysis of two thermal fluctuation experiments. Summary and concluding remarks are given in Section 1.6.

1.2 *Preliminary analysis of a thermal fluctuation experiment*

As explained in Section 1.1, the association/dissociation points in the thermal fluctuation assay indicate thermal fluctuation variance decrease/increase. Therefore in the existing approach (Chen et al. 2008) these points are identified using a moving standard deviation plot based on the thermal fluctuation data. Figure 3 illustrates such a plot with standard deviations calculated by 15 consecutive observations. In this figure, some periods (marked by arrows) in which the standard deviations decrease significantly indicate the presence of bonds.

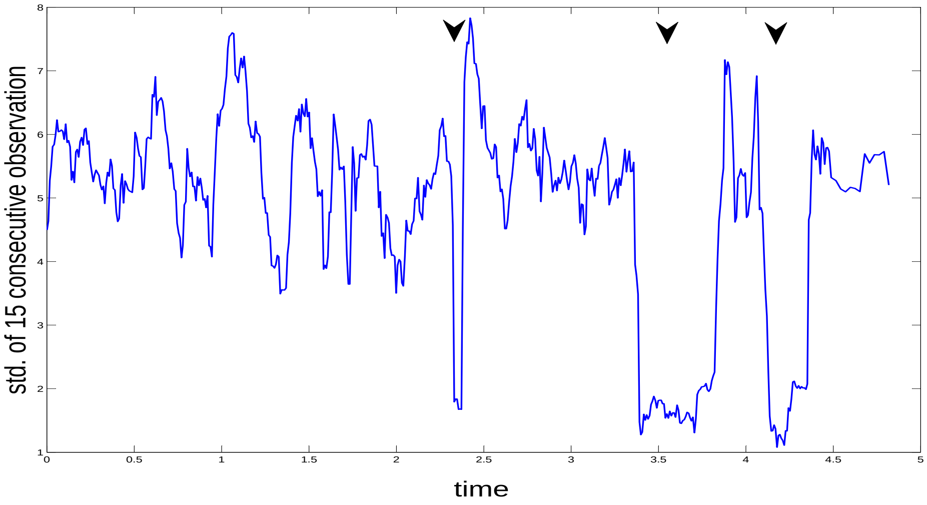

Figure 3: Moving Standard Deviation Plot Based on Data in Figure 2

Standard deviation plots are intuitive and easy to implement but have limitations. First, the accuracy of identifying the association and dissociation points is susceptible to the number of consecutive points used in calculating the standard deviations. That is, the resulting plots can be different with different numbers of consecutive points used in the calculation, which can lead to inconsistent identification of the association and dissociation points. Second, it has no clear decision rule and theoretical justification, especially when the observations are not independent. This issue becomes more serious when there is more than one type of bonds.

1.3 Hidden Markov models

1.3.1 Modeling

A framework based upon Hidden Markov models (HMM) is introduced to analyze the thermal fluctuation experiments. Suppose y_s represents the probe location at time s. There is an unobservable binding state, denoted by x_s, associated with y_s. The change of state can be described by a stationary Markov chain on K states with transition probability $P_{ij} = P(x_{s+1} = j \mid x_s = i)$ and stationary probability π_i, where $i, j \in \{1, \ldots, K\}$. Different order of cell memory effect (Zarnitsyna et al., 2007) can be captured and assessed by the use of transition matrix. Conditional on the undelying binding states, the observed probe locations are assumed to be mutually independent and normally distributed with density $f(y_s;\ \sigma_{x(s)}, \phi_{x(s)})$, where $\phi_{x(s)}$ and $\sigma^2_{x(s)}$ are the mean and variance. The hidden states are defined only according to the variance in this study because it is believed that different binding states lead to different levels of fluctuation captured by their variances (Chen et al. 2008). The mean functions $\phi_{x(i)}$ are allowed to be different with respect to the states because the probe can be pulled/pushed by a small force due to the presence of a bond. In general, the proposed framework can be relaxed to include situations in which the hidden states are defined according to the mean and/or variance.

The standard thermal fluctuation experiment is usually conducted with several independent replicates. Thus a more general setting is written as follows. Assume $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{it})$ to be the ith sequence of observations from the experiment and the index ij denotes the jth observation in the ith sequence. Let $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{it})$, $\boldsymbol{\Phi}_i = (\phi_{i1}, \ldots, \phi_{it})$, and $\boldsymbol{\Sigma}_i = (\sigma_{x(i1)}, \ldots, \sigma_{x(it)})$ be the hidden states, mean, and variance for the ith sequence. Then, the density for $\boldsymbol{Y}_i$ can be written as

$$F(\boldsymbol{Y}_i;\ \boldsymbol{\Sigma}_i, \boldsymbol{\Phi}_i) = \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} \prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)}) \pi_{x(i1)} P_{x(i1)x(i2)} \cdots P_{x(i,t-1)x(it)}. \tag{1}$$

The goal of the thermal fluctuation experiment can be restated as that about the underlying states. For example, the number of the hidden states represents the number of bond types in the experiment. The starting and ending points of each state represent the association and dissociation points of the corresponding bond. Note that K is an upper bound for the order of the states and the true value, denoted by K_0, is unknown because it represents the unknown number of binding status. An estimator $\hat{K}_0$ of K_0 will be obtained by using an order selection procedure given below.

1.3.2 Order selection and asymptotic properties

Accurate estimation of the order of the hidden states is important in analyzing thermal fluctuation experiments because it represents the number of bond types. To perform the order selection, an intuitive approach would be to maximize the likelihood. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be a random sample from (1). Let $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_n)$ and $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_n)$. The log-likelihood function of the HMM can be written as:

$$l_n(\boldsymbol{\Sigma}, \boldsymbol{\Phi}) = \sum_{i=1}^{n} \log F(\boldsymbol{Y}_i;\ \boldsymbol{\Sigma}_i, \boldsymbol{\Phi}_i),$$

where $F(\boldsymbol{Y}_i;\ \boldsymbol{\Sigma}_i, \boldsymbol{\Phi}_i)$ is given in (1). By maximizing $l_n(\boldsymbol{\Sigma}, \boldsymbol{\Phi})$, however, the resulting model may overfit the data with a large value of $\hat{K}_0$. MacKay (2002) proposed a penalized minimum-distance (MD) method that prevents such an overfitting by avoiding having small π_k values. This approach is shown to be consistent in estimating the number of hidden states. However, it overlooks another type of overfitting which was first observed by Chen and Khalili (2008) in finite mixture models, i.e., overfitting with some component densities close to each other. To circumvent this problem, Chen and Khalili (2008) introduced a double penalized approach for finite mixture models. This approach is further extended to HMM in this paper, which takes into account both types of overfitting and provides a better estimation of the number of hidden states.

A double penalized log-likelihood function is defined as

$$\tilde{l}_n(\mathbf{\Sigma}, \mathbf{\Phi}) = l_n(\mathbf{\Sigma}, \mathbf{\Phi}) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_n(\eta_k), \tag{2}$$

where $\eta_k = \sigma_{k+1} - \sigma_k$, for $k = 1, 2, \ldots, K-1$, and $\sigma_1 \leq \sigma_2 \cdots \leq \sigma_K$. The first penalty is used to prevent small value of π_k. The second penalty, p_n, is a nonnegative function that shrinks small η_k to 0 with positive probability. Thus it prevents overfitting by different normal distributions with variances close to each other. Several penalty functions are available in the literature (Donoho 1994, Tibshirani 1996, 1997, Zou and Hastie 2005, Zou 2006). In this paper, we assume p_n to be a smoothly clipped absolute deviation penalty (SCAD) (Fan and Li 2001). We choose SCAD for p_n because it is used in many applications and has desirable asymptotic properties. The SCAD penalty can be characterized by its derivative

$$p_n'(\eta) = \gamma_n \sqrt{n} I\{\sqrt{n} \mid \eta \mid \leq \gamma_n\} + \frac{\sqrt{n}(a\gamma_n - \sqrt{n} \mid \eta \mid)_+}{a-1} I\{\sqrt{n} \mid \eta \mid > \gamma_n\}, \tag{3}$$

where $a > 2$ and γ_n are tuning parameters. By maximizing (2), the estimated order $\hat{K}_0$ of HMM can be obtained.

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$. To study theoretical properties of the proposed procedure, we first rewrite the density function of HMM as:

$$F(\boldsymbol{Y};\ G, \mathbf{\Phi}) = \int F(\boldsymbol{Y};\ \mathbf{\Sigma}, \mathbf{\Phi}) dG(\mathbf{\Sigma}), \tag{4}$$

where

$$G(\mathbf{\Sigma}) = \sum_{x(1)=1}^{K} \cdots \sum_{x(t)=1}^{K} \pi_{x(1)} P_{x(1)x(2)} \cdots P_{x(t-1)x(t)} I(\sigma_{x(1)} \leq \sigma_1, \ldots, \sigma_{x(t)} \leq \sigma_t). \tag{5}$$

Let $(\hat{G}_n, \hat{\mathbf{\Phi}})$ be the maximizer of $\tilde{l}_n(G, \mathbf{\Phi})$, where

$$\hat{G}_n = \sum_{x(1)=1}^{K} \cdots \sum_{x(t)=1}^{K} \hat{\pi}_{x(1)} \hat{P}_{x(1)x(2)} \cdots \hat{P}_{x(t-1)x(t)} I(\hat{\sigma}_{x(1)} \leq \sigma_1, \ldots, \hat{\sigma}_{x(t)} \leq \sigma_t).$$

For later proofs and properties of estimator $\hat{G}_n$, we want to rewrite $\hat{G}_n$ into summation from 1 to K_0. Follow the notation of Chen and Khalili, we define the index sets $\mathfrak{I}(k) = \{j : \sigma_{0,k-1} + \sigma_{0k} \leq 2\hat{\sigma}_j \leq \sigma_{0k} + \sigma_{0,k+1}\}$ for $k = 1, 2, \ldots, K_0$ with $\sigma_{00} = -\infty$ and $\sigma_{0,K_0+1} = \infty$. Introduce $\hat{\vartheta}$ as an estimator of the stationary probability cum the transition probability:

$$\hat{\vartheta}(k_1, k_2, \ldots, k_t) = \sum_{x(1) \in \mathfrak{I}(k_1)} \sum_{x(2) \in \mathfrak{I}(k_2)} \cdots \sum_{x(t) \in \mathfrak{I}(k_t)} \hat{\pi}_{x(1)} \hat{P}_{x(1)x(2)} \cdots \hat{P}_{x(t-1)x(t)},$$

where $k_1, k_2, \ldots, k_t = 1, 2, \ldots, K_0$. Use $\hat{\vartheta}_m$ to estimate π_m by:

$$\hat{\vartheta}_m = \sum_{x(1) \in \mathfrak{I}(m)} \sum_{x(2)} \cdots \sum_{x(t)} \hat{\pi}_{x(1)} \hat{P}_{x(1)x(2)} \cdots \hat{P}_{x(t-1)x(t)} = \sum_{k_2=1}^{K_0} \cdots \sum_{k_t=1}^{K_0} \hat{\vartheta}(k_1 = m, k_2, \ldots, k_t).$$

Therefore, $\hat{G}_n$ can be expressed in terms of $\hat{\vartheta}$ as follows:

$$\hat{G}_n = \sum_{k_1=1}^{K_0} \sum_{k_2=1}^{K_0} \cdots \sum_{k_t=1}^{K_0} \hat{\vartheta}(k_1, k_2, \ldots, k_t) \hat{H}(k_1, k_2, \ldots, k_t, \mathbf{\Sigma}),$$

where

$$\hat{H}(k_1, k_2, \ldots, k_t, \mathbf{\Sigma}) = \frac{\sum\limits_{x(1) \in \mathfrak{I}(k_1)} \cdots \sum\limits_{x(t) \in \mathfrak{I}(k_t)} \hat{\pi}_{x(1)} \hat{P}_{x(1)x(2)} \cdots \hat{P}_{x(t-1)x(t)} I(\hat{\sigma}_{x(1)} \leq \sigma_1, \ldots, \hat{\sigma}_{x(t)} \leq \sigma_t)}{\hat{\vartheta}(k_1, k_2, \ldots, k_t)}.$$

Similarly, we can have:

$$\hat{H}_m = \sum_{k_2=1}^{K_0} \cdots \sum_{k_t=1}^{K_0} \hat{H}(k_1 = m, k_2, \ldots, k_t, \mathbf{\Sigma}).$$

The following two results prove the consistency of the double penalized approach in estimating the order. They are extensions of similar results for the mixture models (Chen and Khalili 2008). Assumptions and proofs are along the lines of Chen and Khalili (2008) and thus deferred to the appendix.

Theorem 1: *Suppose $F(\boldsymbol{Y};\ \mathbf{\Sigma}, \mathbf{\Phi})$ satisfies the identifiability and regularity conditions in the appendix and SCAD penalty term is $\gamma_n = O(n^{1/4} \log n)$. Then,* *(i) for any continuous point* $\mathbf{\Sigma}$ *of* G_0*,* $\hat{G}_n(\mathbf{\Sigma}) \rightarrow G_0(\mathbf{\Sigma})$ *in probability as* $n \rightarrow \infty$*, and* $\hat{\vartheta}_k = \pi_{0k} + o_p(1)$ *for each* $k = 1, 2, \ldots, K_0$*;*

(ii) all atoms of $\hat{H}_k$ *converge in probability to* σ_{0k} *for* $k = 1, 2, \ldots, K_0$*.*

The next theorem shows that $\hat{H}_k$ has a single atom with probability tending to 1 for each k, and thus $\hat{G}_n$ is consistent in estimating K_0.

Theorem 2: *Assume the same conditions as in Theorem 1. Under the true finite mixture density* $F(\boldsymbol{Y};\ G_0, \mathbf{\Phi}_0)$*, if* $(\hat{G}_n, \hat{\mathbf{\Phi}})$ *falls into an* $O(n^{-1/4})$ *neighborhood of* $(G_0, \mathbf{\Phi}_0)$*, then* $\hat{K}_0$ *tends to* K_0 *with probability tending to one.*

These asymptotic properties require an infinite collection of independent sequences $\boldsymbol{Y}_i$ with fixed length t. It is worth noting that this assumption can be relaxed to single sequence $(y_1, y_2, \ldots, y_t)$ with $t \rightarrow \infty$. The results still hold by constructing n independent HMM subsequences of length T with $\boldsymbol{Y}_i = (y_{i_1}, y_{i_1+1}, \ldots, y_{i_1+T-1})$, where $i = 1, \cdots, n$, $i_1 \in \{1, 2, \ldots, t\}$, and $\mid i_1 - j_1 \mid \rightarrow \infty$ for any $i \neq j$.

1.3.3 Estimation

The Baum-Welch expectation-maximization (EM) algorithm (Baum, Petrie, Soules, and Weiss 1970; Dempster, Laird, and Rubin 1977; Welch 2003) is generally used to estimate the unobservable states. Since $\mathbf{\Sigma}$ has no closed form in the M-step, direct application of the standard EM algorithm is not computationally tractable. Therefore, a modified version of the EM algorithm, known as expectation conditional maximization (ECM), is applied (Meng and Rubin 1993). The idea is to replace each M-step with a sequence of conditional maximization steps in which each parameter is maximized individually. Let $\Psi = (\sigma_1, \sigma_2, \ldots, \sigma_K, \pi_1, \pi_2, \ldots, \pi_K, P_{11}, P_{12}, \ldots, P_{KK}, \phi_1, \phi_2, \ldots, \phi_K)$ stand for all the unknown parameters in the model. The complete log-likelihood function is:

$$l_n^c(\boldsymbol{Y}, \boldsymbol{X};\ \Psi) = \sum_{i=1}^{n} \sum_{v(i1)=1}^{K} \cdots \sum_{v(it)=1}^{K} z_{i1,\ldots,it} \log \big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{v(ij)}, \phi_{v(ij)}) \pi_{v(i1)} P_{v(i1)v(i2)} \cdots P_{v(i,t-1)v(it)} \big),$$

where $z_{i1,\ldots,it} = 1$, if $(v(i1), v(i2), \ldots, v(it)) = \boldsymbol{X}_i$; and 0 otherwise, are unobservable indicator variables. Thus, the penalized complete log-likelihood function can be written as

$$\tilde{l}_n^c(\boldsymbol{Y}, \boldsymbol{X};\ \Psi) = l_n^c(\boldsymbol{Y}, \boldsymbol{X};\ \Psi) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_n(\eta_k)$$

and is maximized by iteratively performing the following two steps.

E-Step: Let $\Psi^{(m)}$ be the parameter estimate in the mth iteration. There are n sequences observed and the length of each sequence is t. Assuming that $\Psi^{(m)}$ is the true parameter and given the observed data, the conditional expectation of the complete loglikelihood function with respect to $z_{i1,\ldots,it}$ can be written as:

$$\begin{aligned} Q(\Psi;\ \Psi^{(m)}) &= \sum_{i=1}^{n} \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{\boldsymbol{X}_i}^{(m)} \log \big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)}) P_{x(i1)x(i2)} \cdots P_{x(i,t-1)x(it)} \big) \\ &\quad + \sum_{i=1}^{n} \sum_{k=1}^{K} \{ w_{i,k}^{(m)} + \frac{C_k}{n} \} \log \pi_k - \sum_{k=1}^{K-1} p_n(\eta_k), \end{aligned}$$

where

$$\begin{aligned} w_{\boldsymbol{X}_i}^{(m)} &= w_{x(i1),x(i2),\ldots,x(it)}^{(m)} \\ &= \frac{\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}^{(m)}, \phi_{x(ij)}^{(m)}) \pi_{x(i1)}^{(m)} P_{x(i1)x(i2)}^{(m)} \cdots P_{x(i,t-1)x(it)}^{(m)}}{\sum_{x(l1)=1}^{K} \cdots \sum_{x(lt)=1}^{K} \prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(lj)}^{(m)}, \phi_{x(lj)}^{(m)}) \pi_{x(l1)}^{(m)} P_{x(l1)x(l2)}^{(m)} \cdots P_{x(l,t-1)x(lt)}^{(m)}}, \end{aligned}$$

$$w_{i,k}^{(m)} = \sum_{x(i2)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{x(i1)=k,x(i2),\ldots,x(it)}^{(m)}, \qquad k = 1, 2, \ldots, K.$$

CM-Step: It includes two substeps and the $(m+1)$st iteration is the final output of the two substeps. Let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_K\}$, $\mathbb{P} = \{P_{11}, P_{12}, \ldots, P_{KK}\}$ and $g_1(\Psi) = \boldsymbol{\Sigma}$. The first substep is to estimate $\Pi^{(m+1)}$, $\mathbb{P}^{(m+1)}$ and $\boldsymbol{\Phi}^{(m+1)}$ by maximizing $Q(\Psi;\ \Psi^{(m)})$ subject to the constraint $g_1(\Psi) = g_1(\Psi^{(m)})$. It can be written in closed form as follows:

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} w_{i,k}^{(m)} + C_K}{n + KC_K},$$

$$P_{ab}^{(m+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{t} Pr^{(m)}(x_{i,j-1} = a, x_{ij} = b \mid \boldsymbol{Y}_i)}{nt \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{t} Pr^{(m)}(x_{i,j-1} = a, x_{ij} = k \mid \boldsymbol{Y}_i)},$$

$$\phi_k^{(m)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{t} \alpha_{ij}^{(m)}(k) \beta_{ij}^{(m)}(k) y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{t} \alpha_{ij}^{(m)}(k) \beta_{ij}^{(m)}(k)}, \qquad k, a, b \in \{1, 2, \ldots, K\},$$

where $Pr(x_{i,j-1} = a, x_{ij} = b \mid \boldsymbol{Y}_i)$, $\alpha_{ij}(k)$ and $\beta_{ij}(k)$ are defined as: $\alpha_{ij}(k) = Pr(y_{i1}, y_{i2}, \ldots, y_{ij}, x_{ij} = k)$, $\beta_{ij}(k) = Pr(y_{i,j+1}, y_{i,j+2}, \ldots, y_{it} \mid x_{ij} = k)$. Also, $F(\boldsymbol{Y}_i;\ \boldsymbol{\Sigma}_i, \boldsymbol{\Phi}_i) = \sum_{k=1}^{K} \alpha_{it}(k)$, and $Pr(x_{i,j-1} = a, x_{ij} = b, \boldsymbol{Y}_i) = \alpha_{i,j-1}(a) P_{ab} f(y_{ij};\ \sigma_b, \phi_b) \beta_{ij}(b)$. Details on choice of initial values and the setting of α and β can be found in Baum et al. (1970) and Welch (2003).

The second substep is to update $\boldsymbol{\Sigma}$ by conditional maximization. Constrained on $g_2(\Psi) = g_2(\Psi^{(m+\frac{1}{2})})$, where $g_2(\Psi) = \{\Pi, \mathbb{P}, \boldsymbol{\Phi}\}$ and $\Psi^{(m+\frac{1}{2})} = (\Pi^{(m+1)}, \mathbb{P}^{(m+1)}, \boldsymbol{\Phi}^{(m+1)}, \boldsymbol{\Sigma}^{(m)})$, the new estimation can be obtained by using the Newton-Raphson method. Because of the non-smoothness of the SCAD penalty $p_n(\eta)$, a local quadratic approximation (LQA) is suggested by Fan and Li (2001) to implement the Newton-Raphson iteration, i.e., $\tilde{p}_n(\eta;\ \eta_k^{(m)}) = p_n(\eta_k^{(m)}) + \frac{p_n'(\eta_k^{(m)})}{2\eta_k^{(m)}}(\eta^2 - \eta_k^{(m)^2})$. However, it is known that LQA can have problems like numerical instability and sharing a drawback of backward stepwise variable selection (Hunter and Li 2005). Several methods are proposed to address the problems, including a perturbed version of LQA proposed by Hunter and Li (2005) and an iterative algorithm based on local linear approximation (LLA) proposed by Zou and Li (2008). Because LLA inherits the desirable features of lasso (Tibshirani 1996) in terms of computational efficiency and avoids the drawback of LQA, we implemented the LLA according to the suggestion of Zou and Li (2008) as follows:

$$\tilde{p}_n(\eta;\ \eta_k^{(m)}) = p_n(\eta_k^{(m)}) + p_n'(\eta_k^{(m)})(\eta - \eta_k^{(m)}).$$

Based on some simulations (not reported here), LLA outperforms LQA in selecting the correct order of HMM under the current setting, which is consistent with the findings in Zou and Li (2008). More discussions on LLA can be found in Rahul et al. (2011).

Detailed updating procedure of $\boldsymbol{\Sigma}$ can be written as follows. To update σ_1, we have

$$\sigma_1^{(m+1)} = \sigma_1^{(m)} - \frac{D_1^1(\Psi^{(m+\frac{1}{2})})}{D_2^1(\Psi^{(m+\frac{1}{2})})},$$

where

$$\begin{aligned} D_1^1(\Psi^{(m+\frac{1}{2})}) &= \partial_{\sigma_1} \,|_{\Psi^{(m+\frac{1}{2})}} \sum_{i=1}^{n} \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{\boldsymbol{X}_i}^{(m+\frac{1}{2})} \\ &\qquad \log\big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)})\big) - \partial_{\sigma_1}\tilde{p}_n(\eta_1;\ \eta_1^{(m)}), \end{aligned}$$

and

$$\begin{aligned} D_2^1(\Psi^{(m+\frac{1}{2})}) &= \partial_{\sigma_1\sigma_1} \,|_{\Psi^{(m+\frac{1}{2})}} \sum_{i=1}^{n} \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{\boldsymbol{X}_i}^{(m+\frac{1}{2})} \\ &\qquad \log\big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)})\big) - \partial_{\sigma_1\sigma_1}\tilde{p}_n(\eta_1;\ \eta_1^{(m)}). \end{aligned}$$

For σ_k with $k = 2, 3, \ldots, K-1$, the estimation is given by

$$\sigma_k^{(m+1)} = \sigma_k^{(m)} - \frac{D_1^k(\Psi^{(m+\frac{1}{2})})}{D_2^k(\Psi^{(m+\frac{1}{2})})},$$

where

$$\begin{aligned} D_1^k(\Psi^{(m+\frac{1}{2})}) &= \partial_{\sigma_k} \,|_{\Psi^{(m+\frac{1}{2})}} \sum_{i=1}^{n} \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{\boldsymbol{X}_i}^{(m+\frac{1}{2})} \log\big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)})\big) \\ &\qquad - \partial_{\sigma_k}\tilde{p}_n(\eta_{k-1};\ \eta_{k-1}^{(m)}) - \partial_{\sigma_k}\tilde{p}_n(\eta_k;\ \eta_k^{(m)}), \end{aligned}$$

and

$$\begin{aligned} D_2^k(\Psi^{(m+\frac{1}{2})}) &= \partial_{\sigma_k\sigma_k} \,|_{\Psi^{(m+\frac{1}{2})}} \sum_{i=1}^{n} \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{\boldsymbol{X}_i}^{(m+\frac{1}{2})} \log\big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)})\big) \\ &\qquad - \partial_{\sigma_k\sigma_k}\tilde{p}_n(\eta_{k-1};\ \eta_{k-1}^{(m)}) - \partial_{\sigma_k\sigma_k}\tilde{p}_n(\eta_k;\ \eta_k^{(m)}). \end{aligned}$$

For σ_K, we have

$$\sigma_K^{(m+1)} = \sigma_K^{(m)} - \frac{D_1^K(\Psi^{(m+\frac{1}{2})})}{D_2^K(\Psi^{(m+\frac{1}{2})})},$$

where

$$\begin{aligned} D_1^K(\Psi^{(m+\frac{1}{2})}) &= \partial_{\sigma_K} |_{\Psi^{(m+\frac{1}{2})}} \sum_{i=1}^{n} \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{\boldsymbol{X}_i}^{(m+\frac{1}{2})} \\ &\qquad \log \big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)}) \big) - \partial_{\sigma_K} \tilde{p}_n(\eta_{K-1};\ \eta_{K-1}^{(m)}), \end{aligned}$$

and

$$\begin{aligned} D_2^K(\Psi^{(m+\frac{1}{2})}) &= \partial_{\sigma_K \sigma_K} |_{\Psi^{(m+\frac{1}{2})}} \sum_{i=1}^{n} \sum_{x(i1)=1}^{K} \cdots \sum_{x(it)=1}^{K} w_{\boldsymbol{X}_i}^{(m+\frac{1}{2})} \\ &\qquad \log \big(\prod_{j=1}^{t} f(y_{ij};\ \sigma_{x(ij)}, \phi_{x(ij)}) \big) - \partial_{\sigma_K \sigma_K} \tilde{p}_n(\eta_{K-1};\ \eta_{K-1}^{(m)}). \end{aligned}$$

Based on the two substeps, the $(m+1)$st iteration can be updated by

$$\Psi^{(m+1)} = (\Pi^{(m+1)}, \mathbb{P}^{(m+1)}, \boldsymbol{\Phi}^{(m+1)}, \boldsymbol{\Sigma}^{(m+1)}).$$

The iterative procedure is terminated if the log-likelihood increment is smaller than a predetermined value. Its convergence is guaranteed according to the results in the EM literature (Wu 1983, Meng and Rubin 1993). For the tuning parameters, cross-validations (Stone 1974) are usually used. The widely used leave-one-out cross validation, however, cannot be applied in this case because of the dependent structure of HMM, i.e., y_{ij} and $y_{i,j-1}$ are dependent to each other through x_{ij}. Therefore, we implement the half-sampling cross validation method proposed by Celeux and Durand (2008), which preserve the Markov chain structure. Let $\boldsymbol{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{it})$ be a sequence of HMM. We choose the odd (and resp. even) sub-chain of each HMM sequence, i.e., $\boldsymbol{Y}_i^1 = (y_{i1}, y_{i3}, \ldots)$ (and $\boldsymbol{Y}_i^2 = (y_{i2}, y_{i4}, \ldots)$). Each sub-chain forms a new HMM with $\tilde{\Psi} = (\Pi, \mathbb{P}^2, \boldsymbol{\Phi}, \boldsymbol{\Sigma})$. Denote the maximum double penalized estimates for odd subsequences by $\hat{\Psi}_{n,1}$ and by $\hat{\Psi}_{n,2}$ for even subsequences. Then the half-sampling cross validation is given by

$$CV(\gamma_n) = -\sum_{i=1}^{2} l_n^i(\hat{\Psi}_{n,i})$$

and γ_n is chosen by minimizing $CV(\gamma_n)$. Through cross validation, it is observed that the double penalized method is not sensitive to the choice of a and C_K, and similar results were observed and discussed in Chen et al. (2008). They suggested that if $\hat{\sigma}_k \in [M^{-1}, M]$ for some large enough M, then a recommended setting of C_K is $C_K = \log M$. In the context of this study, we use half-sampling cross validation and choose $C_K = 0.6 \log 10$ for all simulations and real examples. We chose $a = 3.7$ in (3) as recommended by Fan and Li (2001).

1.4 Simulation study

To illustrate the order selection performance, we compare the proposed approach with three methods in the literature: AIC (Akaike 1974), BIC (Schwarz 1978), and minimal-distance (MacKay 2002). Both AIC and BIC select the order by directly controlling the order K. The minimal-distance (MD) criterion is defined by

$$MD(\bar{F}_n, F) = d_{KS}(\bar{F}_n, F) - C_n \sum_{i=1}^{k} \log \pi_i,$$

where $\bar{F}_n$ is the t-dimensional empirical distribution function of F, t is the length of each sequence, and d_{KS} is the Kolmogorov-Smirnoff distance.

Simulations were conducted based on 16 settings and the details are summarized in Table 1. The first column, K_0, indicates the true number of hidden states which ranges from 2 to 9. More attention was given to orders 2 to 5 because they reflect the numbers of bond populations in the cell adhesion experiments. The transition probabilities and the means and variances of the conditional normal distributions are also listed. These settings take into account variance and/or mean changes and also incorporate various settings of the transition matrices. When K_0 increases, values of the transition matrices play a key role in determining the types of HMMs. We focus on two types of settings that represent the two cases in the study of MacKay (2002). One is a "unbalanced" case in which higher probabilities appear on the diagonal (i.e., proportion of time in each state is unbalanced as in cases 4 and 5). The other is a "balanced" case in which the same probability appears in each element of the transition matrix (as in cases 6 to 9). Also, different values of n (i.e. 2, 5, 10, 20, 50,100) are considered. Simulations are also conducted for cases with $K_0 > 5$. Because they give similar conclusions, to save space, only one example with $K_0 = 9$ is reported in the table.

We use cross validation to choose the tuning parameter $\gamma_n/\sqrt{n}$ in the interval $[.1, .5]$. For each simulation setting, the ECM iterations terminated if the log-likelihood increment is smaller than 10^{-4}. This algorithm converges efficiently. For example, for HMM with 2 sequences and length 100, it took about 40 seconds for an Intel Xeon CPU with 2.66 GHz and 3.00 GB of RAM to achieve such convergence. Furthermore, as recommended by Meng and Rubin (1993), adding a few more inner loops for updating σ_k can speed up convergence.

Table 1: Parameter Settings in Simulation

case	K_0	Mean	Variance	Transition Matrix	n	t
1	2	$(0,0)'$	$(0.5,4)'$	0.5	2	50
2	2	$(0,0)'$	$(0.5,4)'$	0.5	20	50
3	2	$(0,0)'$	$(0.5,4)'$	0.5	50	50
4	2	$(0,0)'$	$(0.5,4)'$	P_4	2	50
5	2	$(0,0)'$	$(0.5,4)'$	P_4	20	50
6	2	$(0,0)'$	$(0.5,4)'$	P_4	50	50
7	4	$(9,20,1,9)'$	$(0.3,0.5,1,2)'$	P_8	2	100
8	4	$(9,20,1,9)'$	$(0.3,0.5,1,2)'$	P_8	100	100
9	4	$(3,10,7,1)'$	$(0.3,0.5,0.8,1.1)$	0.25	2	100
10	4	$(3,10,7,1)'$	$(0.3,0.5,0.8,1.1)$	0.25	10	100
11	4	$(3,10,7,1)'$	$(0.3,0.5,0.8,1.1)$	0.25	50	100
12	5	$(20,1,5,9,17)'$	$(0.6,0.8,1.5,1.7,2)$	0.2	5	100
13	5	$(20,1,5,9,17)'$	$(0.6,0.8,1.5,1.7,2)$	0.2	10	500
14	5	$(20,1,5,9,17)'$	$(0.6,0.8,1.5,1.7,2)$	0.2	100	500
15	9	$(0,10,-16,20,15,$ $-4,-20,-8,0)'$	$(0.2,0.5,0.8,0.9,1.2,$ $1.3,1.4,2,2.1)'$	1/9	10	1000
16	9	$(0,10,-16,20,15,$ $-4,-20,-8,0)'$	$(0.2,0.5,0.8,0.9,1.2,$ $1.3,1.4,2,2.1)'$	1/9	50	1000

Tables 2 and 3 show the order selection performance of the four methods in the 16 settings, where

$$P_4 = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}.$$

$$P_8 = \begin{pmatrix} 0.75 & 0.1 & 0.1 & 0.05 \\ 0.2 & 0.7 & 0.05 & 0.05 \\ 0.2 & 0.1 & 0.6 & 0.1 \\ 0.3 & 0.1 & 0.1 & 0.5 \end{pmatrix}.$$

For each setting, the true order is indicated by the number with boldface in the "order" column. For each method, we report the percentage of times out of 100 replications that the estimated order equals to a value between 1 and 11. The value with the highest frequency is indicated by a boldface. When the true order is 2 (i.e., cases 1 to 6), the double penalized approach (DP) is consistently the best and has more than 84% success rate in identifying the true order. AIC also works reasonably well in these settings, while BIC tends to underestimate in some cases. When sample size increases, the selection accuracy of DP is improved which is consistent with the asymptotic results. This result is observed throughout the simulations, i.e., cases 7-8, cases 9-11, cases 12-14, and cases 15-16. For $K_0 = 4$, both MS and DP outperform the other methods in the unbalanced cases (of 7 and 8). It appears to be more difficult to identify the correct order in the balanced cases (cases 9 to 14). In these cases, DP consistently identifies the correct order with the highest frequency while most of the other methods underestimate the order even for larger sample size. In cases 15 and 16 with $K_0 = 9$, DP outperforms the other three methods.

Table 2: Simulation results

case	*order*	AIC	BIC	MS	DP	case	*order*	AIC	BIC	MS	DP
1	1	0.22	**0.7**	0.23	0.09	2	1	0.18	**0.7**	0.16	0.08
	2	**0.78**	0.3	**0.74**	**0.84**		**2**	**0.82**	0.29	**0.83**	**0.87**
	3	0	0	0.03	0.06		3	0	0.01	0.01	0.05
	4	0	0	0	0.01		4	0	0	0	0
	5	0	0	0	0		5	0	0	0	0
3	1	0.17	**0.66**	0.17	0.07	4	1	0.18	0.41	0.14	0.08
	2	**0.79**	0.34	**0.81**	**0.89**		**2**	**0.81**	**0.59**	**0.78**	**0.85**
	3	0.04	0	0.02	0.03		3	0.01	0	0.07	0.07
	4	0	0	0	0.01		4	0	0	0.01	0
	5	0	0	0	0		5	0	0	0	0
5	1	0.16	0.37	0.11	0.05	6	1	0.19	0.39	0.14	0.08
	2	**0.84**	**0.63**	**0.85**	**0.88**		**2**	**0.8**	**0.59**	**0.84**	**0.9**
	3	0	0	0.02	0.07		3	0.01	0.02	0	0.02
	4	0	0	0.02	0		4	0	0	0.02	0
	5	0	0	0	0		5	0	0	0	0

Table 3: Simulation results

7	2	0	0	0	0	8	2	0	0	0	0
	3	0.4	**0.72**	0.02	0.08		3	0.39	**0.65**	0	0.07
	4	**0.42**	0.28	**0.86**	**0.73**		**4**	**0.44**	0.32	**0.88**	**0.90**
	5	0.18	0	0.12	0.19		5	0.17	0.03	0.12	0.03
	6	0	0	0	0		6	0	0	0	0
9	2	0	0	0	0	10	2	0	0	0	0
	3	**0.56**	**1**	**0.92**	0.4		3	**0.6**	**0.98**	**0.78**	0.32
	4	0.42	0	0.08	**0.53**		**4**	0.4	0.02	0.21	**0.65**
	5	0.02	0	0	0.07		5	0	0	0.01	0.03
	6	0	0	0	0		6	0	0	0	0
11	2	0	0	0	0	12	3	0.02	0	0	0
	3	**0.55**	**0.95**	**0.66**	0.25		4	**0.58**	**0.75**	**0.57**	0.38
	4	0.45	0.04	0.34	**0.75**		**5**	0.39	0.25	0.47	**0.6**
	5	0	0.01	0	0		6	0.01	0	0	0.02
	6	0	0	0	0		7	0	0	0	0
13	3	0.04	0	0	0.01	14	7	0.08	0	0	0
	4	**0.52**	**0.72**	0.38	0.25		8	**0.57**	**0.8**	0.35	0.11
	5	0.39	0.28	**0.59**	**0.74**		**9**	0.35	0.2	**0.61**	**0.8**
	6	0.05	0	0.03	0		10	0	0	0.04	0.09
	7	0	0	0	0		11	0	0	0	0
15	7	0	0	0	0	16	7	0	0	0	0
	8	**0.67**	**0.98**	0.42	0.35		8	**0.58**	**0.95**	0.47	0.19
	9	0.31	0.02	**0.46**	**0.64**		**9**	0.37	0.04	**0.51**	**0.72**
	10	0.02	0	0.07	0.01		10	0.05	0.01	0.02	0.08
	11	0	0	0.05	0		11	0	0	0	0.01

1.5 Application in cell adhesion experiments

We now consider the application to the thermal fluctuation experiments in this section. The proposed approach is applied to real data (Chen et al. 2008) to assess the accuracy in identifying the number of bond types and specifying their association/dissociation points. Two sets of experiments were recorded. One is L-selectin interacting with P-selectin glycoprotein ligand-1 (PSGL-1) and another is P-selectin interacting with PSGL-1. It is known that the stiffness of L-selectin differs from that of P-selectin so we expect to see a difference in the level of thermal fluctuation reductions during their bonds formation with PSGL-1. The first data set has one type of bond in the experiment and is used to validate the level of thermal fluctuation reduction for each type of bonds. The second data set has a mixture of two different types of bonds and is used to test a proposed model to see if it can separate these two bonds.

For the first data, the interest focuses on the interactions between L-selectin and PSGL-1. Low densities of selectins and PSGL-1 are used to ensure that interactions formed are most likely single bonds, i.e., either no bond or a single L-selectin-PSGL-1 bond for each interaction. There are 18 independent replicates of the thermal fluctuation sequences and each of them has over 300 probe positions recorded in 5 second. Figure 2 is a typical sample with such a setting. The HMM is applied with $K = 4$ and the number of the hidden states is correctly specified as two, i.e. $K_0 = 2$, using the double penalized approach. The estimated transition matrix is

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.9924 & 0.0076 \\ 0.0242 & 0.9758 \end{pmatrix}$$

and the stationary probabilities are

$$\hat{\Pi} = (0.7645, 0.2355).$$

Define state 1 as no bond state and 2 as the L-selectin-PSGL-1 bond state. The estimated hidden states can be represented as in Figure 4 based on the data in Figure 2. The lines indicate the transition points of states. The starting points of state 2 are the association points of the L-selectin-PSGL-1 bonds and the starting points of state 1 are the dissociation points.

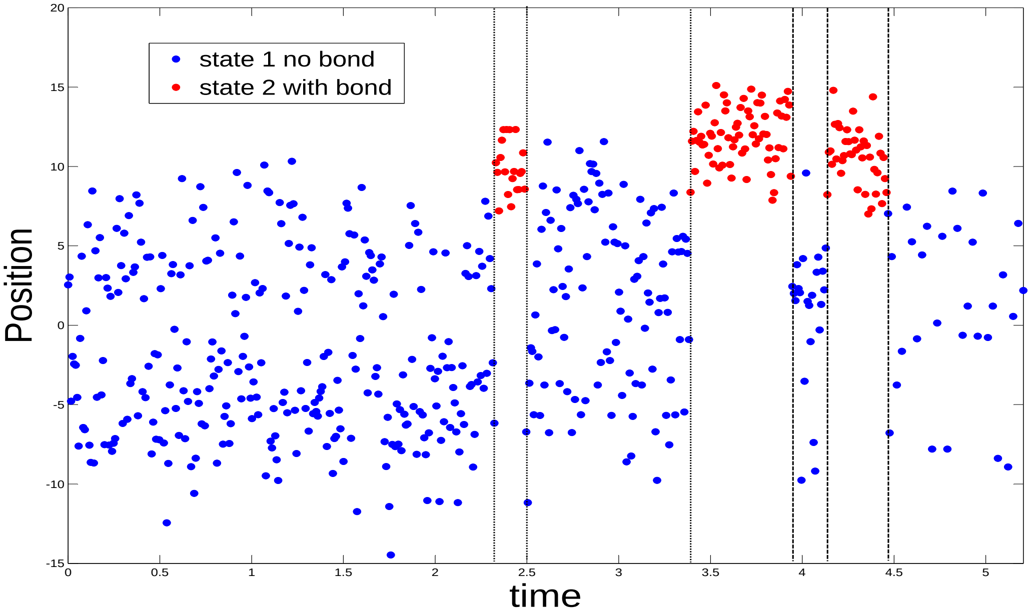

Figure 4: HMM Analysis, L-selectin and PSGL-1 Adhesion Experiment

To assess the goodness-of-fit of the fitted model, a graphical technique proposed by Altman (2004) is implemented. Define the empirical 2-dimensional cumulative density function (CDF) by:

$$\int_{-\infty}^{z_1}\int_{-\infty}^{z_2}\bar{F}_2 dx = \frac{\sum_{i=1}^{n}\sum_{j=1}^{t-1} I\{y_{ij}\leq z_1, y_{i,j+1}\leq z_2\}}{n(t-1)}.$$

When the parameters Ψ are estimated, the resulting distribution, $\int_{-\infty}^{z_1}\int_{-\infty}^{z_2}\hat{F}_2 dx$, can be treated as the estimated 2-dimensional CDF. The idea of Altman (2004) is to compare the empirical CDF with the estimated bivariate distribution as shown in Figure 5. With observed data ranging from -16 to 17, we have z_i taking value in (-16,17) and $i = 1, 2$. The points are reasonably close to the 45° line through the origin, indicating that the model is correctly specified.

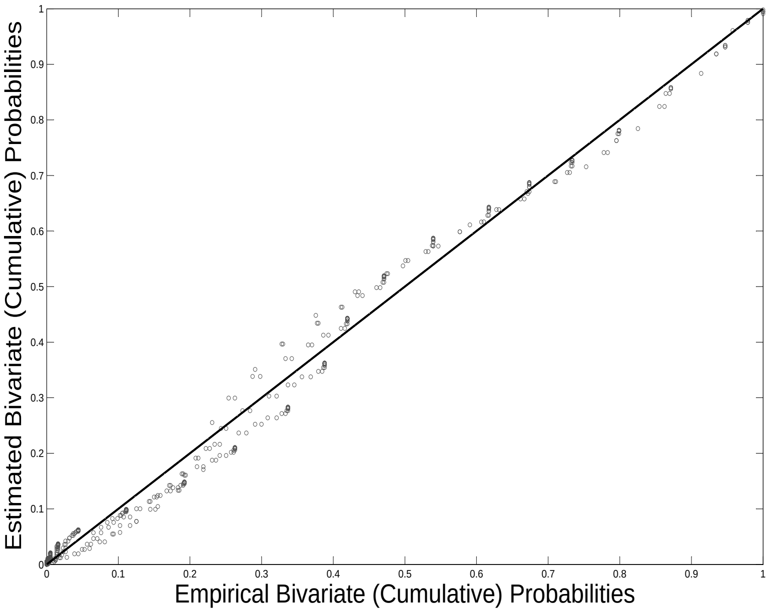

Figure 5: Comparison of the Estimated and Empirical Bivariate Distributions

Previous research has shown that a memory effect might exist in the repeated contacts, i.e., the adhesion probability in the next contact might be increased because of the adhesion in the immediate past (Zarnitsyna et al., 2007). In the HMM context, a memory effect can be expressed in terms of the transition probabilities, i.e., $P_{10} < P_{11}$ or not. The existence of such an effect can be carefully assessed by a likelihood-ratio (LR) test (Giudici et al., 2000) based upon the fitted HMM. That is, to perform the hypothesis test as follows:

$$H_0 : P_{10} \geq P_{11} \quad vs \quad H_1 : P_{10} < P_{11}.$$

To perform the LR test, we evaluate the maximum log-likelihood under H_0 and under H_1. They are -1730.34 for H_0 and -1659.28 for H_1. Therefore, the LR statistic is 142.12. Comparing to the χ^2 distribution with one degree of freedom leads to a p-value close to 0, which supports the hypothesis of a first order memory effect.

In the second setting, the thermal fluctuation observations are collected with a mixture of two types of receptor-ligand bonds that are formed due to interactions of

L-selectin and P-selectin with their PSGL-1ligand. There are in total 48 independent mixture sequences collected. The HMM framework with $K = 5$ is applied to analyze the mixed observations. The order of the hidden states is correctly specified as three with the estimated transition matrix

$$\hat{\mathbb{P}} = \begin{pmatrix} 0.9499 & 0.0498 & 0 \\ 0.0018 & 0.8953 & 0.1029 \\ 0.0449 & 0.0636 & 0.8915 \end{pmatrix},$$

and the stationary distribution:

$$\hat{\Pi} = (0.3404, 0.4264, 0.2332).$$

The goodness-of-fit of this model is assessed graphically as shown in Figure 6 with $z_i \in (-22, 24.5)$ and $i = 1, 2$. Figure 7 gives a typical sequence analyzed by the HMM approach. The circles represent the hidden states corresponding to the P-selectin-PSGL-1 bonds, the dots represent those corresponding to the L-selectin-PSGL-1 bonds, and the rests labeled by triangles represent those corresponding to no bond. The estimated variances of the fluctuations for the P-selectin-PSGL-1 bond and the L-selectin-PSGL-1 bond are 16.2104 and 12.4027 respectively. They indicate that the formation of the L-selectin-PSGL-1 bond reduces the BFP thermal fluctuations more than what the P-selectin-PSGL-1 bond does. This can be explained biologically because L-selectin has a higher stiffness than P-selectin (Chen et al. 2008).

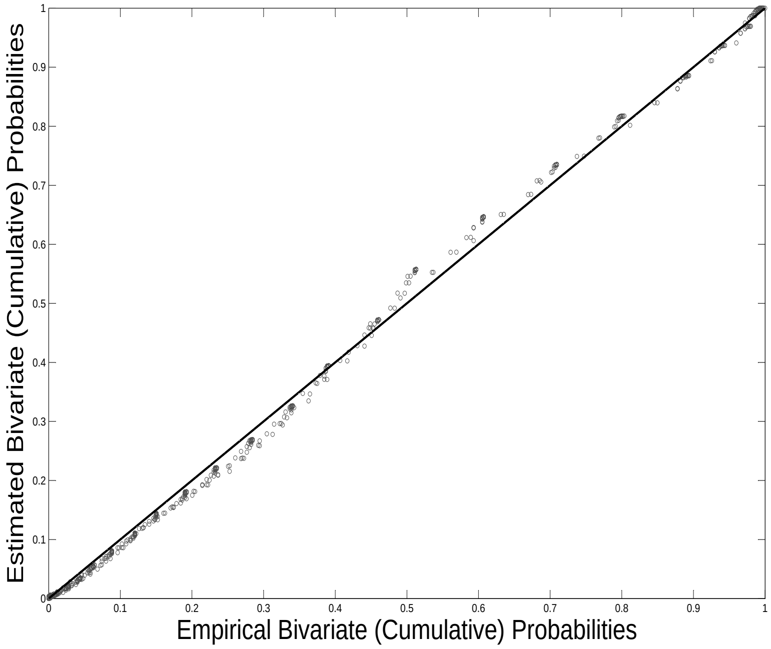

Figure 6: Comparison of the Estimated and Empirical Bivariate Distributions

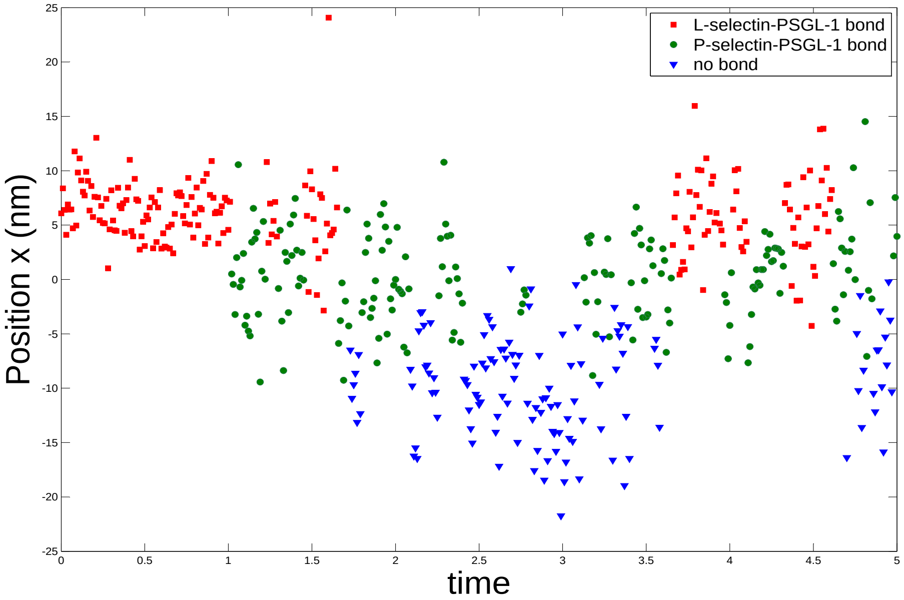

Figure 7: HMM Analysis, Mixture Bonds Experiment

1.6 Summary and concluding remarks

This paper is motivated by application of the thermal fluctuation method (Chen et al. 2008) to study the kinetics of multiple receptor-ligand interactions during cell-cell adhesion. This study uses the reduced thermal fluctuations to indicate the presence of receptor-ligand bonds. More than one type of bond is observed and they correspond to different levels of fluctuation decrease due to their string strength difference. In order to provide a systematic approach to identify the number of bond types and the corresponding association/dissociation points, a new framework based on hidden Markov models and order selection is proposed. It works by assuming that the probe fluctuates differently according to the underlying binding states of the cells, i.e., no bond or a number of distinct types of bonds. These states are unobservable but their changes can be captured by a Markov chain.

In spite of the prevalence of HMMs in many applications, their modeling and inference mainly focus on the situations where the order is known. In many real situations including the present one, the number of hidden states is unknown. To tackle this problem, a double penalized procedure is introduced. It is shown to be asymptotically consistent in estimating the order of HMMs. Efficient algorithm based on expectation conditional maximization is presented. The proposed method outperforms three existing methods in a simulation study. It is also successfully applied to two real data sets. Judging by its good performance in the simulation study and application to real data, we think the proposed methodology should find applications in other areas.

Although the SCAD penalty is chosen for the double penalized procedure, it can easily be extended and implemented to other penalty functions. We have conducted a small simulation study by using the Lasso penalty (Tibshirani 1996) and the results lead to comparable conclusions. Similarly, the asymptotic result in the paper should be extendable to other penalty functions. This work is left for future investigation.

1.7 Acknowledgments

This research is supported by NIH GM096187 grants. The authors would like to thank Wei Chen for sharing his data and Jiahua Chen for helpful discussions. We are grateful to the Editor, an AE and a referee for their helpful comments and suggestions.

CHAPTER II

AN HMM-BASED ALGORITHM FOR EVALUATING RATES OF RECEPTORLIGAND BINDING KINETICS FROM THERMAL FLUCTUATION DATA

2.1 Introduction

During the early stage of hemostatic and thrombotic processes, platelets tether to and roll on the immobilized von Willebrand factor (VWF), which is mediated through binding between the 45kDa N-terminal domain of the alpha subunit of the GPIb-IX-V complex (GPIbα) and the A1 domain of the VWF (Ruggeri and Men-dolicchio, 2007). Disease-related mutations in the VWF have been found to change the mechanical regulation of platelet adhesion, resulting in the bleeding disorder von Willebrand disease (VWD) (Ruggeri, 2007). From a biophysical perspective, these mutations alter VWFGPIbα binding kinetics. It has been shown that single-residue mutation R1450E that exhibits the type 2B VWD phenotype increases VWFGPIbα binding affinity and supports the rolling of more platelets at slower velocities without a minimum shear requirement (Auton et al., 2010; Coburn et al., 2011). Another single-residue mutation G1324S that exhibits the type 2M VWD phenotype decreases the binding affinity between these two mole-cules (Morales et al., 2006; Coburn et al., 2011).

The binding affinity is the ratio of the on- to off-rates, which quantifies the net effects of receptorligand association and disso-ciation. To measure the on- and off-rates separately, mechanical methods, such as the thermal fluctuation assay, that employ ultra-sensitive force probes, e.g., the biomembrane force probe (BFP) (Chen et al., 2008) and optical tweezers (Molloy et al., 1995; Veigel et al., 1999; Lister et al., 2004;

Sun et al., 2009), have been developed to measure the interactions of proteins immobilized on surfaces. The idea stems from the observation that force probes used for single-molecule experiments are usually susceptible to thermal fluctuations. The formation of a molecular bond spanning across the gap between the force probe and the target physically connects the two surfaces and reduces the thermal fluctuation of the force probe. In other words, the newly-formed bond is equivalent to adding a constraint to the force probe (Chen et al., 2008). In the analysis of experimental data, bond formation is detected from the reduction in the thermal fluctuation of the probe position and bond dissociation is detected from the resumption of thermal fluctuation, as judged by the sliding standard deviation moving below or above certain thresholds. Although this descriptive statistical method is simple, it has several disadvantages: It is time-consuming, not very robust, susceptible to noise, and subjective. To overcome these drawbacks, we developed a Hidden Markov model (HMM)-based algorithm that provides an automatic and systematic procedure for analyzing thermal fluctuation data efficiently. We first assume a hidden state, bound or unbound, for each observed probe position. Given the hidden states, the probe positions are assumed to be independent and normally distributed with unknown parameters. The forward-backward algorithm (Baum et al., 1970; Dempster et al., 1977; Welch, 2003) was used to estimate the underlying states and unknown parameters.

Because of its versatility in modeling and robustness in prediction performance, HMM has wide applications in computational biology. For example, HMMs can detect tumor subtypes with microarray data (Zhang et al., 2011) and identify protein-binding sites in DNA (Cardon and Stormo, 1992). However, to the best of our knowledge, no HMM-based computer algorithm has been developed for analyzing thermal fluctuation data. In the thermal fluctuation assay, if the probe is in either the bound or unbound state at one moment, it is more likely to be in the same state at the next moment. This memory effect can be successfully captured by assuming a Markovian structure at the transition of the underlying states (Hung et al., 2013). Furthermore, HMM enables us to provide statistical inference such as the confidence interval and the prediction interval. In particular, by using the likelihood ratio test based on the fitted HMM, we can verify the memory effect objectively and rigorously in repeated adhesions (Hung et al., 2013).

This paper is organized as follows. Sections 2.2.1 and 2.2.2 describe the experiment setup and the existing method. Section 2.2.3 illustrates the procedures to analyze thermal fluctuation data. Sections 2.2.4-2.2.7 discuss the modeling and computation of HMM. In Section 2.3, we use the HMM to derive kinetic rates by analyzing thermal fluctuation data obtained for the interaction of VWF-A1 and glycocalicin (GC), the extracellular portion of GPIbα. We also show the performance of the HMM method in comparison to the manual method based on descriptive statistics. In addition to the dataset with wild-type (WT) A1, datasets with two single-residue A1 VWD mutations: R1450E (type 2B) and G1324S (type 2M) are added to the performance test of the HMM method and show that the HMM is far easier to use. Section 2.4 presents the discussion and concluding remarks.

2.2 Methods

2.2.1 Experimental setup

The recombinant WT VWF-A1 domain (residues 1238-1471) and two single-residue mutants, R1450E that exhibits the gain-of-function (GOF) phenotype of type 2B VWD and G1324S that exhibits the loss-of-function (LOF) phenotype of type 2M VWD, were gifts from Dr. Miguel Cruz (Baylor College of Medicine, TX). The GPIbα extracellular domain glycocalicin (GC) was a gift from Dr. Jing-fei Dong (Puget Sound Blood Research Institute, WA).

The BFP system (Chen et al., 2008) and the interacting molecules are respectively illustrated in Figure 8A and B. The VWF-A1 and GC were covalently coupled to the probe bead (Figure 8B, left) and the target bead (Figure 8B, right), respectively. Human red blood cells (RBCs) were purified from peripheral blood of healthy donors by finger prick and biotinylated using a protocol approved by the Institutional Review Board of the Georgia Institute of Technology. In order to enable attachment to the apex of the biotinylated RBC, streptavidins were coated to probe beads. The pressurized RBC by micropipette aspiration serves as an ultra-sensitive force transducer with a soft spring constant of 0.15 pN/nm by tuning the pressure through a custom-made manometer system. A homemade LabviewTM program was used for data acquisition by tracking the probe bead displacement with 0.7 ms temporal and ±3 nm spatial resolution. The experiment used a high-speed camera at 1,500 frames per second (fps) to track the axial (horizontal) position of the probe in discrete time points. The raw data of probe position x vs. time t consist of four phases (Figure 8C). The target bead was driven by a computer-controlled piezoelectric translator to approach the probe bead at a speed of 2 μm/s (Figure 8C, black). After a short contact of 0.1s (green), the target was retracted (purple) and held from the probe by a separation distance of 10 nm for 10-15s (blue and red). The Brownian motion of the probe bead was monitored with the same BFP spring constant for all experiments. Experiments were performed at room temperature (25°C).

2.2.2 Descriptive statistical method

The underlying idea is that anchoring the probe bead to the target bead via a VWF-A1GC bond reduces the thermal fluctuations. This is because the stiffness of the system (k_{sys}) is the BFP stiffness (k_{BFP}) without a bond but is changed to the sum of the BFP stiffness and the molecular bond stiffness (k_{mol}), i.e., $k_{sys} = k_{BFP}+k_{mol}$, with a bond. The reduction in thermal fluctuation follows from the equipartition theorem, $k_{sys}\sigma^2 = k_BT$, where k_B is the Boltzmann constant, T is absolute temperature, and σ^2 is the ensemble variance of the displacements that represents a metric of thermal fluctuations. At constant temperature, an increase in k_{sys} would cause a decrease in σ^2. Thus, the decrease in σ^2 indicates bond association while the increase in σ^2 indicates bond dissociation. The variance of bound portion should be smaller than that of unbound portion. In the descriptive statistical method, we approximated the en-semble standard deviation by a sliding standard deviation of 90 consecutive data points, σ_{90}, from the x-t sequence and plotted it vs. t (Figure 8D). We chose 90 points by balancing the competing needs for an approximate value and temporal resolution. Note that the number of points chosen to plot the standard deviation can affect analysis results. We then draw two horizontal lines to represent the thresholds to identify bond association (solid line in Figure 8D) and dissociation (dashed line). The choice of thresholds also requires the experimenter's judgment and can cause variation in classifications of bound vs. unbound states. The descriptive statistical-based method selects data points with a σ_{90} lower than the association threshold to be in the bound state and those higher than the dissociation threshold to be in the unbound state. This method is very time-/labor-consuming, which may take several days to finish the analysis of data generated from a one-day experiment. To obtain statistically-meaningful results, a large number of distance curves need to be collected, making data analysis the bottleneck of the output. Moreover, this analysis uses personal judgment to select the window width of sliding standard deviation and the thresholds for state classification. This will inevitably bring in subjectivity and errors.

2.2.3 Data preparation: removing erroneous data and correcting drift

To overcome drawbacks of the descriptive statistical-based algorithm, we developed an HMM-based algorithm. Before applying either method, a careful automated pre-screening of x vs. t raw data is required. This is because some of the curves exhibit large magnitude of rapid shifting, probably due to environmental perturbations and human errors during experiments (× in Figure 9). The poor quality of such data prevents reliable analysis by either algorithm. In particular it may affect HMM learning by causing false-positive bond classification. As a first step of data preparation, erroneous data are removed (Figure 9, Step 1). For the acceptable data ($\surd$ in Figure 2), there may still be slow drift in the holding phase, which might be caused by misaligned contact between the probe and the target during the assembly of the BFP. As the second step of data preparation, a high order polynomial is fitted to the position data and corrects the drift (Step 2). After pre-screening, the clean data are ready for both descriptive statistical-based algorithm to use and HMM training and classification. In the learning process, we train HMM to get the tuning parameter using cross validation as described in Section 2.2.6 (Step 3). Then HMM is ready for data classification (Step 4) and kinetic analysis (Step 5).

2.2.4 An HMM-based algorithm for analyzing thermal fluctuation data

We developed an HMM method to analyze x-t curves from the thermal fluctuation assay (Figure 10). The objective is to computerize the bond state annotation similar to the descriptive method but with a higher efficiency (Figure 10A). The statistical methodology can be found in Hung et al. (2013). Here we model the molecular interaction on a BFP as a process with the hidden bound state following Markovian structure (Figure 10B). Let x_t denote the horizontal position of the probe at time t. For each observation x_t, there is an unobservable variable z_t representing the binding state at time t. The indicator variable z_t takes value 0 (Figure 10B, blue) when there is no bond between the probe and the target at time t, and 1 (Figure 10B, red) otherwise. The change of state z_t can be described by a stationary Markov chain with two states, transition probability $P_{ij} = P(z_{t+1} = j \mid z_t = i)$ and stationary probability P_i, where i, j take values of 0 or 1. Stationary probability P_1/P_0 can be interpreted as the probability of observing bound/unbound event in the experiment. When the corresponding binding state z_t is given to be k, the corresponding probe position x_t is assumed to be mutually independent and normally distributed with mean μ_{HMM} and variance σ^2_{HMM}. From Section 2.2.2, we have $\sigma^2_{HMM,0} > \sigma^2_{HMM,1}$. As a result, the HMM method divides an x-t curve into a series of segments. Each segment is characterized by a constant μ_{HMM} and σ^2_{HMM}. This will distinguish the bound and unbound portions, thus making the threshold much easier to be seen (Figure 10C).

2.2.5 HMM computation

A forward-backward algorithm is used to compute the parameters and unknown states. Stationary probability of the unbound state P_0 is the only tuning parameter in the algorithm. The reason for using P_0 is to incorporate biological knowledge of the binding frequency into HMM. This tuning parameter can be chosen through cross validation. In fact, we can show that the analysis result is insensitive to the initial choice of P_0 as long as it lies in a proper range (Section 2.3.2). The forward-backward algorithm is a two-step procedure that computes the estimate as follows: in the forward step, it computes $P(z_m \mid x_1, \ldots, x_m)$ for all $m \leq n$, where n is the length of the sequence; then in the backward step, the algorithm computes $P(x_{m+1}, \ldots, x_n \mid z_m)$. It is known that the algorithm converges to the maximum-likelihood estimate (Baum et al., 1970).

2.2.6 On- and off-rate estimates

This subsection describes how to statistically estimate kinetic on- and off-rates (k_{on} and k_{off}) of receptorligand interaction through the previously classified bound and unbound states vs. time segments. Because formation and dissociation of single biomolecular bonds are stochastic events, the moments when they occur and their durations are random. The on- and off-rates represent statistical characteristics underlying these probabilistic kinetic processes. Therefore, they are determined by the totality of the data rather than individual points in the collection. As such, k_{on} and

k_{off} are insensitive to small disturbance and error, such as missing or false alarm in a small number of events. This property can be used to train the tuning parameter in HMM (Section 2.2.7) and explain the performance comparison of the HMM- and descriptive statistical-based algorithms (Section 2.2.3).

Let waiting time t_w be the period from the dissociation moment of the existing bond to the association moment of the next bond; and bond lifetime t_b be the period from the moment of bond association to dissociation. A pooled collection of waiting times should follow the distribution of the first-order kinetics of irreversible association of single bonds:

$$P_w = 1 - \exp(-k_{on}^c t_w) \tag{6}$$

where the cellular on-rate $k_{on}^c = A_c m_r m_l k_{on}$ is a product of four parameters: Ac is the contact area (considered as a constant for all experiments), m_r and m_l are the respective receptor (GC) and ligand (A1) densities measured by flow cytometry (Yago et al., 2004), and k_{on} is the molecular on-rate. P_w is the probability for a bond to form after waiting time t_w. P_w can be estimated by survival frequency as the fraction of events with waiting time $\geq t_w$. Thus, the cellular on-rate can be estimated from the negative slope of the $\ln(1 - P_w)$ vs. t_w plot (Figure 11A). Similarly, a pooled collection of bond lifetimes should follow the distribution of the first-order kinetics of irreversible dissociation of single bonds:

$$P_b = \exp(-k_{off} t_b), \tag{7}$$

where P_b is the probability for a bond formed at $t = 0$ to survive at t_b and can be estimated by survival frequency with bond lifetime $\geq tb$. The negative slope of the $\ln(P_b)$ vs. t_b plot provides an estimate for the off-rate koff (Figure 12A). Our recent work (Ju et al., 2013) suggested that the VWF-A1GC bond has two states at low force: one major state that features events with short lifetime $(0.01s < t_b < 0.5s)$ and one minor state with long lifetime $(t_b \geq 0.5s)$. Usually long lifetime events mingle with multiple bond events and become susceptible to drifting-induced noise, while events with very short lifetime ($t_b \leq 0.01s$) are highly suspected as non-specific events. For illustrative purposes, we only demonstrate the accuracy and reliability of HMM with events in short lifetime regime ($0.01s < t_b < 0.5s$).

2.2.7 Training of HMM

To choose the tuning parameter and test the robustness of the algorithm, we implement a half-sampling cross validation method (Celeux and Durand, 2008) which preserves the underlying Markov chain structure. We segregate the complete sequence of probe position observations $X = (x_1, x_2, \ldots, x_n)$ into the odd, i.e. $X1 = (x_1, x_3, \ldots)$, and even, i.e. $X2 = (x_2, x_4, \ldots)$ sub-sequences. Denote the off-rate of the HMM result from the odd sub-sequence as $k_{off}X1$ and the even sub-sequence as $k_{off}X2$. The relative error ϵ_c between the two sub-sequences is defined as:

$$\epsilon_c = (1 - k_{off}X1/k_{off}X2)^2.$$

We can similarly define the relative error of off-rate between HMM and descriptive statistical methods. Let k^1_{off} be the off-rate of X using descriptive statistical method and k^2_{off} be the result of HMM. The relative error ϵ_r is defined as:

$$\epsilon_r = (1 - k^2_{off}/k^1_{off})^2.$$

We choose tuning parameter P0 such that the relative error ϵ_c is small. Later, we shall illustrate the robustness of HMM by showing that the range of P_0 with small ϵ_c overlaps with that with small ϵ_r (Section 2.3.2).

2.3 Results

2.3.1 Justification of HMM with the VWF-A1GC interaction

We compare kinetic rate estimates from descriptive statistical analysis with those from HMM on the same set of thermal fluctuation data. For the interaction of GC with VWF-A1 (Figure 8A), the HMM method (Figure 10A) performs as well as the descriptive method in bond annotation (Figure 8C). The linearized distributions of respective waiting times and bond lifetimes determined by the two methods overlapped and showed similar slopes, suggesting similar cellular on-rate (Figure 11A) and off-rate (Figure 12A) estimates from two methods. Indeed, the means and 95% confidence intervals of the cellular on-rate by the descriptive statistical algorithm and HMM are $1.302 \pm 0.079 s^{-1}$ and $1.395 \pm 0.046 s^{-1}$, respectively (Figure 11B). The two confidence intervals overlap, indicating that the parameter estimates are statistically close to each other. For the off-rate, the means and 95% confidence intervals by the descriptive statistical algorithm and HMM are $26.58 \pm 0.92 s^{-1}$ and $26.46 \pm 0.18 s^{-1}$, respectively (Figure 12B), which also overlap with each other.

In addition to the above analysis of the WT VWF-A1 data, we compared performance of the HMM and descriptive statistical methods using data from two single-residue mutations in VWF-A1 that alter their interactions with GPIb in biologically important ways: 1) G1324S that exhibits type 2M VWD phenotype and 2) R1450E that exhibits type 2B VWD phenotype. To compare molecular on-rates requires removal of the site density effect. We measured the site densities of VWF-A1 and GC respectively and divided the cellular on-rate k_{on}^{c} by $m_r m_l$. corresponding to each A1 construct (WT or mutant). The result is the effective on-rate, $A_c k_{on}$. Since the contact area A_c was kept as close to constant as possible between experiments, the $A_c k_{on}$ is a good measure for on-rate comparison (Chen et al., 2008). Both the descriptive and HMM methods show that mutation G1324S decreased effective on-rate, from 6.64 ± 0.20 to $2.07 \pm 0.05 \times 10^{-6} \mu m^4 s^{-1}$ (descriptive) and 7.12 ± 0.12 to $2.24 \pm 0.03 \times 10^{-6} \mu m^4 s^{-1}$ (HMM) (Figure 11B), but had little effect on off-rate (Figure 12B). This correlates with the loss-of-function phenotype of G1324S as it induces less platelet agglutination compared to WT A1 (Rabinowitz et al., 1992). Both the descriptive and HMM analyses indicate that the R1450E mutation resulted in an 8-fold increase in the effective on-rate: 6.64 ± 0.20 to $76.59 \pm 1.51 \times 10^{-6} \mu m^4 s^{-1}$ (descriptive) and 7.12 ± 0.12 to $82.64 \pm 2.79 \times 10^{-6} \mu m^4 s^{-1}$ (HMM) (Figure 11B), which are in good agreement. Similar to G1324S, R1450E had little effect on stress-free off-rates of the short state (Figure 12B). The result correlates with the gain-of-function phenotype of R1450E. Type 2B VWD mutations in the A1 domain have been shown to result in abnormal interactions between platelet GPIb and soluble VWF, such that R1450E A1 requires less ristocetin or lower shear to induce platelet agglutination (Matsushita and Sadler, 1995). Such abnormal interactions have been suggested to lead to prolonged bleeding time due to either the lack of unbound GPIbα on platelet surface to interact with immobilized VWF at sites of vascular injury, reduced platelet counts due to early clearance of platelet aggregates, or both (Ruggeri and Mendolicchio, 2007). Note that HMM has much narrower width of 95% confidence intervals compared to that of descriptive statistical method for both cellular on-rate (Figure 13A) and off-rate (Figure 13B) for all three molecular interactions tested here. Thus, the HMM method is more accurate (less error) than the descriptive method presumably because it reduces the errors brought by subjective judgment of the experimenter. Moreover, the HMM method can measure far more events than the descriptive statistical method from the same set of raw data, e.g., 112 to 40 for waiting times (first group in Figure 13C) and 169 to 46 for bond lifetimes (first group in Figure 13D) for the WT A1 case. In the mutant cases, the HMM measurements also outnumbered the descriptive statistical measurements (Figure 13C and D), indicating that many of the waiting times and bond lifetimes gone undetected by the descriptive method can be resolved by HMM.

Although the kinetics parameters differ for different molecular interactions, the estimates from HMM are consistent with the anticipated biological effects and match the results from the descriptive statistical method. These results validate HMM as a reliable and accurate method for evaluating the on- and off-rate change of each mutation relative to WT.

2.3.2 Tuning parameter reliability of HMM

In HMM, the probability of observing a data point in the unbound state P_0 is the only tuning parameter in the algorithm. Based on the half-sampling cross validation in Section 2.2.7, we plot the relative error ϵ_c (Figure 14A) and ϵ_r (Figure 14B) against different P_0. The P_0 that gives the lowest ϵ_c ranges from 0.85-0.96 from which we choose the value in our prediction algorithm. It can be seen from Figure 14B that different choices of P_0 do not render much inconsistency between the results from descriptive statistical method and HMM, as the relative error is smaller than 0.025. This shows the robustness in the prediction performance and the reliability of the HMM tuning parameter.

2.3.3 It is easier to learn HMM than the descriptive method

To further verify that HMM reduces the time required for data analysis by the descriptive method, we did the following performance tests:

1. Compare the time required for a new student to learn the HMM and the descriptive method

2. Compare the time required for an experienced student to analyze the same set of raw data using the two methods.

For the first test, we surveyed two new students in our lab who just started learning the thermal fluctuation assay. We plot their learning curves by tracking their performance from week 0 to week 8 (Figure 15A). For each week, we recorded the time required for them to finish analyzing similar amount of thermal fluctuation data by using both manual method (blue) HMM method (red). We found that it took much less time for both to finish the analysis by HMM than by the descriptive method every week: 20 vs. 4 hours at week 0 and 8 vs. 1 hour at week 8. The HMM is much less time-consuming than the descriptive statistical method.

For the second test, we collected information from two students who had experience in analysis of BFP thermal fluctuation data. We assigned the same data set used in Figures 4 and 5 to them and recorded the times it took for them to finish the analysis using the two methods (Figure 15B). Consistently, using the HMM (red) took much less time than using the descriptive method (blue), 1 vs. 5-6 hours.

2.4 Discussion

It has long been recognized that changes in thermal fluctuation can be used to identify single-molecule events. This idea was implemented in early work to probe the duration and contact stiffness of myosin motors interacting with actin filament (Molloy et al., 1995; Mehta et al., 1997; Veigel et al., 1999; Lister et al., 2004). More recently, it was used to analyze two-dimensional kinetics of adhesion molecules interacting with their ligands (Chen et al., 2008; Sun et al., 2009; Huang et al., 2010; Chen et al., 2010), to measure molecular elasticity (Marshall et al., 2006; Sarangapani et al., 2011; Chen et al., 2012), and to determine protein conformational changes (Chen et al., 2012). Some studies employed BFP that was custom-designed and home-made in a handful of laboratories (Chen et al., 2008; Huang et al., 2010; Chen et al., 2010; 2012). Others used optical tweezers (Molloy et al., 1995; Mehta et al., 1997; Veigel et al., 1999; Sun et al., 2009) and atomic force microscope (Marshall et al., 2006; Sarangapani et al., 2011) that are commercially available in many laboratories. Therefore, these methods have high potential for a broad range of applications by many investigators. Unfortunately, previous analyses were done using merely eyeballing (Molloy et al., 1995; Mehta et al., 1997; Veigel et al., 1999; Lister et al., 2004) or descriptive statistical analysis (Marshall et al., 2006; Chen et al., 2008; Sun et al., 2009; Huang et al., 2010; Chen et al., 2010; Sarangapani et al., 2011; Chen et al., 2012). The drawbacks of these primitive analyses may limit the applications of the thermal fluctuation methods because the descriptive statistical-based algorithm is very time consuming, subjective, and prone to noise and error. In this study, we developed a computational algorithm based on analytical statistics rather than descriptive statistics. The HMM-based algorithm automates and high-throughputs the processing of data and has the advantage of being rigorous and objective. We used the VWF-A1GC system to test the HMM method. The estimates from HMM are comparable to those from the descriptive statistical method (manual analysis) (Figures 8C and 10A) with the same tuning parameters (Figures 11 and 12).

This paper compares the on- (Figure 11) and off- (Figure 12) rates of GC interactions with WT and two mutant A1 domains. At static conditions, platelet GPIbα does not bind WT VWF unless a modulator ristocetin is added to induce the conformational activation of the A1 domain (Berndt et al., 1988). By comparison, the type 2B VWD mutant R1450E binds GPIbα spontaneously without ristocetin (Matsushita and Sadler, 1995; Auton et al., 2010) whereas the type 2M VWD mutation G1324S abolishes the ristocetin-induced binding to GPIb (Rabinowitz et al., 1992; Morales et al., 2006). Our kinetics measurements correlate well with these biochemical characterizations in that the R1450E mutant gains the function with an increased on-rate whereas the G1324S mutant loses the function with a decreased on-rate (Figure 11B). The data indicate that the association kinetics reflect the conformational states of VWF-A1. There has recently been significant progress in correlating protein structure and binding kinetics. A web server has been developed for prediction of association rate constant by incorporating the protein conformational changes based on the archived protein-ligand complex structures (Qin et al., 2011). Complementing such efforts, the HMM method combined with the thermal fluctuation assay provides experimentally measured binding kinetics and their correlation to protein structure and function.

The HMM method consistently shows higher accuracy than the manual method regardless of the biological variation embedded inside the datasets (Figure 13A and B). Furthermore, it also detects far more waiting time and bond lifetime events than the descriptive method (Figure 13C and D). Possible reasons for the descriptive method to capture less events include: First, the calculation of sliding standard deviation σ_{90} may not resolve short bonds if their lifetimes are shorter than the chosen length of the sliding window. Thus, the calculation mixes both bound and unbound observations, which may miss many waiting times and short bond lifetime events; Second, the decision rule (bound vs. unbound) of the descriptive statistical method heavily relies on an empirical threshold. In most cases, this threshold will be manually drawn in a conservative way to avoid false positive annotation caused by noise. Since the reduction of variance by bond formation is relatively small and may not be detected sometimes, the manual method tends to miss bonds when experimental noise is not well controlled.

Aside from its advantage of robustness in prediction (Figure 14), the HMM method is more convenience and requires less learning times than the descriptive method (Figure 14). The comparison was made after an automated prescreening process to eliminate erroneous data resulted from the experimental errors, drifting and noises (Figure 9). It should be noted that the HMM method shares the same biophysical rationales as the descriptive method, but provides the statistical basis to computerizes the manual analysis. Yet, it requires on average 30 seconds for the HMM to finish annotation of one data trace but it takes 5 minutes for the manual method to do so (Figure 8C and 10A).

Although the proposed HMM is motivated by the study of thermal fluctuation experiments, it can be directly applied to different types of studies in bioinformatics (Koshi and Goldstein, 2001). Based on the proposed HMM method, extensions to higher orders models with unknown number of states (Hung et al., 2013) can be made.

Therefore, such a framework is particularly attractive in the study of computational biology such as the analysis of gene expression (Seifert et al., 2011), protein and DNA sequences (Marioni et al., 2006), where the conventional first-order HMM is not sufficient (Seifert, 2013). The HMM method developed in this paper is also applicable to other areas, including signal processing (Kaleh and Vallet, 1994; Chambaz et al., 2009) and environmental science (Hughes and Guttorp, 1994). For future work, we will include the receptors of two or multiple species on a cell and study the cooperative binding of multiple receptors. Unlike a bead target, a cell target is characterized by its soft membrane and instant mobility. Thus, more noise is expected in a cell system than in a purified protein system. The next generation algorithm should be more powerful in correcting thermal fluctuation drifting and noise caused by a restless cell surface. Moreover, multiple receptors will bring in more complex binding kinetics or multiple states such as unbound, receptor-1 bound, receptor-2 bound, and cooperative bound states. To discriminate these states, this method requires higher sensitivity. We hope that continuous improvement of the HMM-based algorithm will allow us to shed new light on examining protein interactions on the single-molecule level.

2.5 Acknowledgments

The authors thank Dr. Jing-fei Dong for providing glycocalicin, Dr. Miguel Cruz for providing recombinant VWF-A1s, Zhenhai Li, Jack Wei Chen, Ning Tan and Veronika Zarnitsyna for helpful discussions. We are grateful to the Editor, the Associate Editor Anna Tramontano and two anonymous referees for their helpful comments and suggestions. Funding: This work was supported by the National Institutes of Health grants GM096187 (to C. F. J. W. and C. Z.) and HL091020 (to C. Z.).

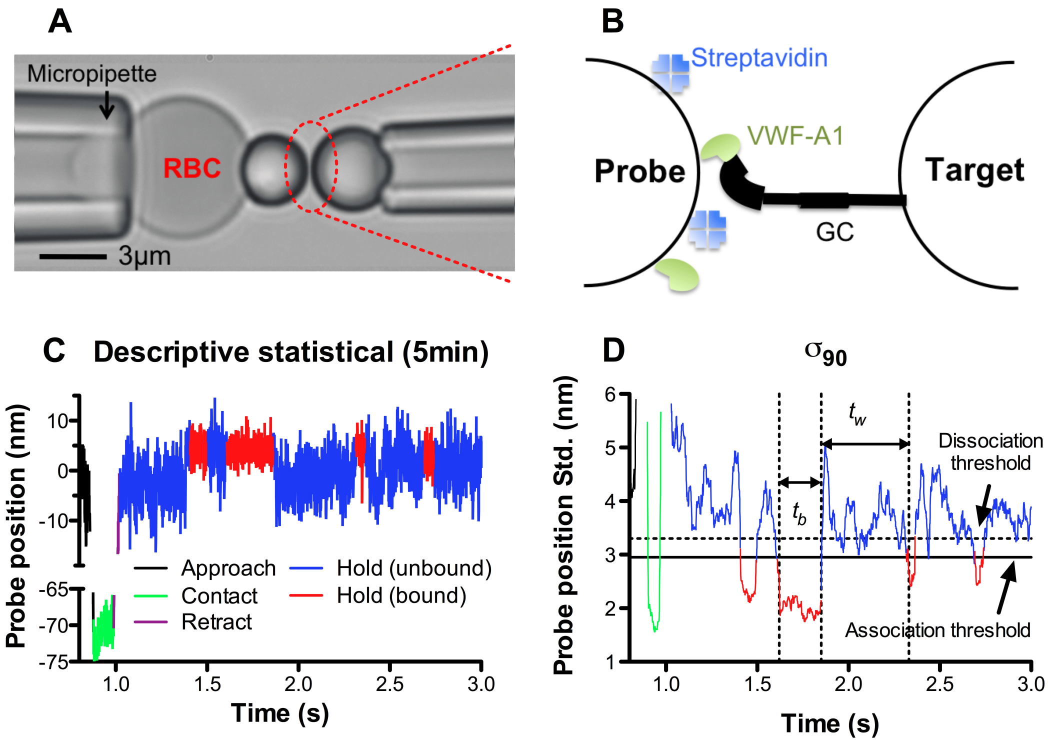

Figure 8: *A*. BFP photomicrograph. A micropipette-aspirated RBC with a bead (left, termed probe) attached to the apex was aligned with a bead (right, termed target) aspirated by another micropipette. *B*. BFP functionalization. VWF-A1 and streptavidin were covalently coupled to the probe bead. Glycocalicin (GC) was covalently coupled to the target bead. The schematic is no to scale as the sizes of the molecules have been enlarged relative to the sizes of the beads. *C*. Thermal fluctuation data. Data plot of the instantaneous horizontal position x of the probe vs. time t collected from one test cycle of the thermal fluctuation assay. During the experiment, the target bead was driven to approach the probe bead (black), contact for 0.1 s (green), retract (purple) and be held (blue and red) stationary with at a preset position. Blue and red traces annotate, respectively, bound and unbound states detected by the descriptive statistical method. 5 minutes on average were taken to finish the manual annotation on one trace. *D*. Plot of σ_{90} (the sliding standard deviation of 90 consecutive x positions from data in C around t) vs. t. The same color coding is used as C.

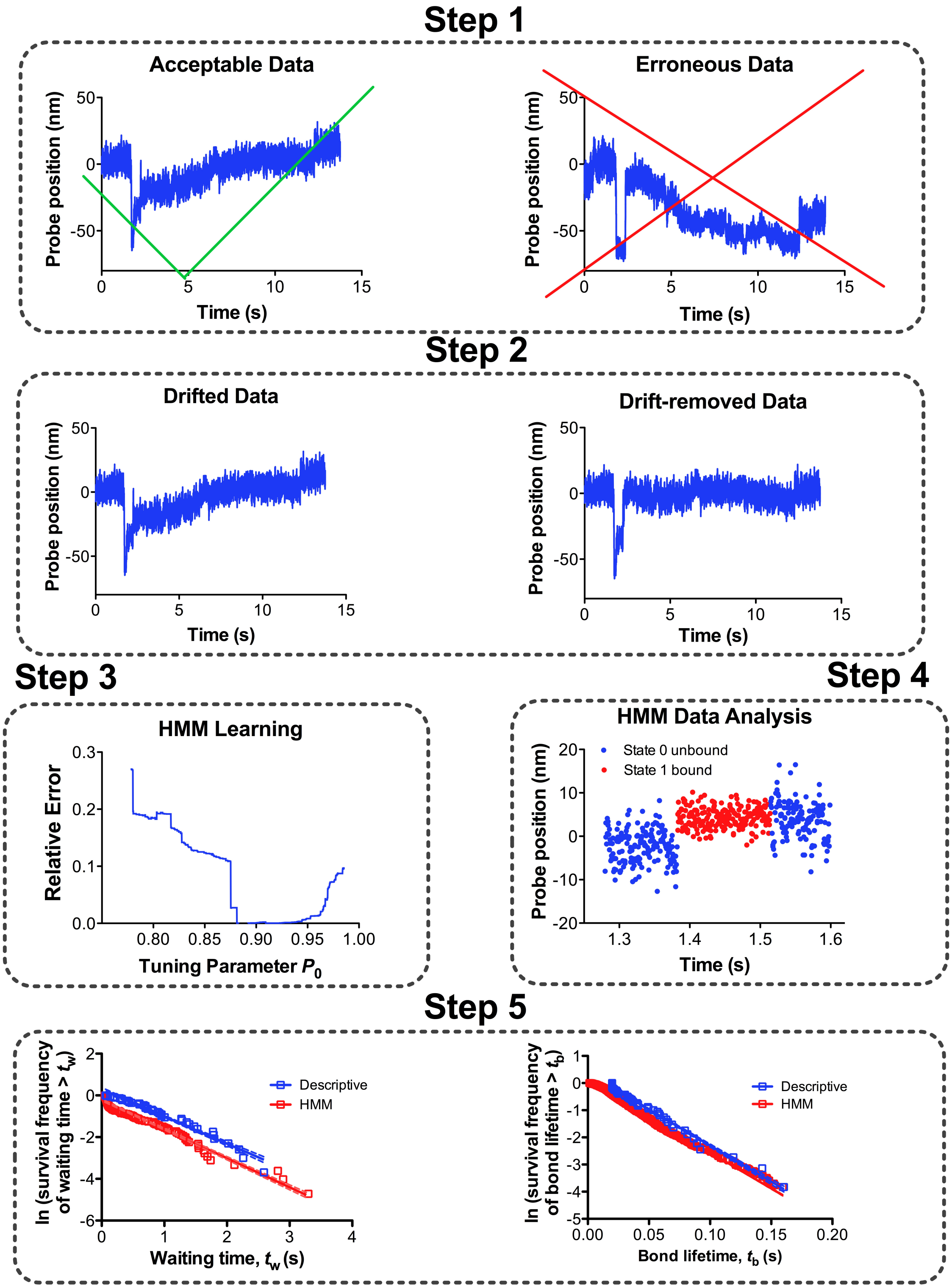

Figure 9: *Step 1*, pre-screening; *Step 2*, drift removal, the first two steps were applied to both descriptive and HMM methods; *Step 3*, HMM parameter estimation; *Step 4*, identification of states by HMM; *Step 5*, evaluation of on- and off-rate by analysis of waiting time and bond lifetime distributions, respectively.

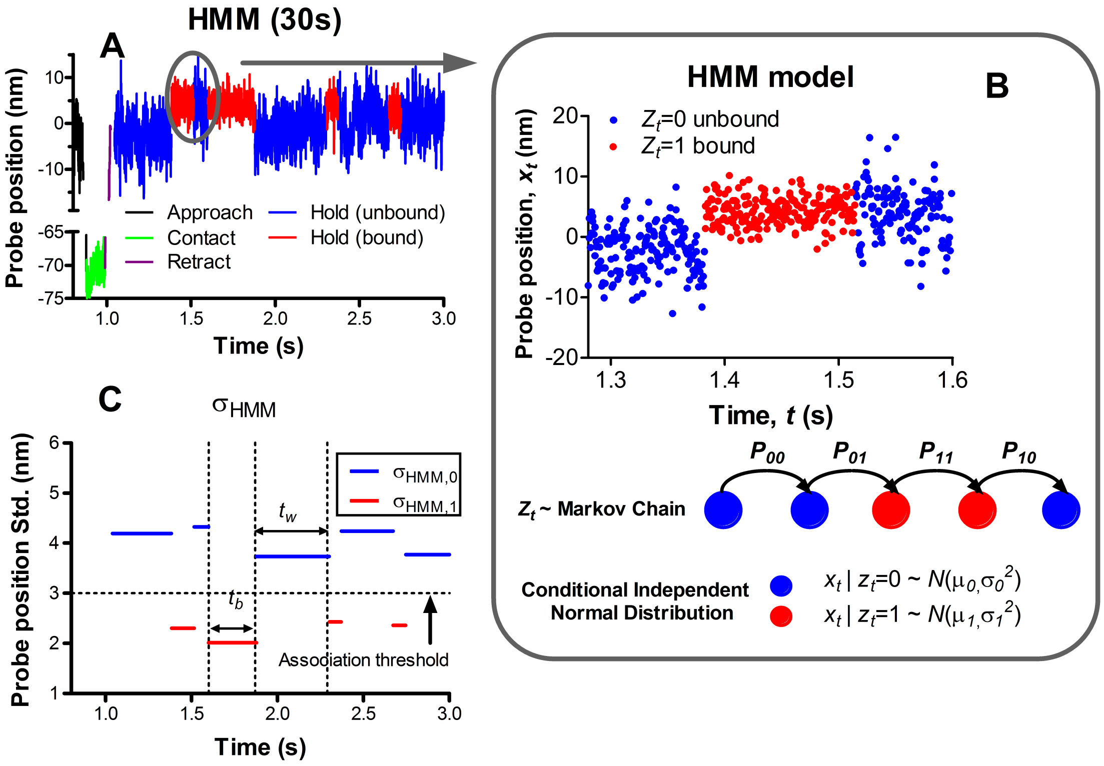

Figure 10: *A*. Bound and unbound status annotation by the HMM analysis from the same data in Figure 1C. The average time spent for the algorithm to finish the annotation of one trace is 30 seconds. *B*. Illustration of the HMM. At time t, let xt be the observed horizontal position of probe and zt be the unobserved binding state. Observation xt can be classified into two states: $z_t = 0$ (blue) or $z_t = 1$ (red). Also, z_t follows a Markov chain and xts are independent normally distributed given zt. *C*. Plot of HMM (the predicted standard deviation from the HMM analysis of *A*) vs. *t*. Each segment of *C* corresponds to the estimated standard deviation of bound or unbound period of A in red or blue by the HMM analysis.

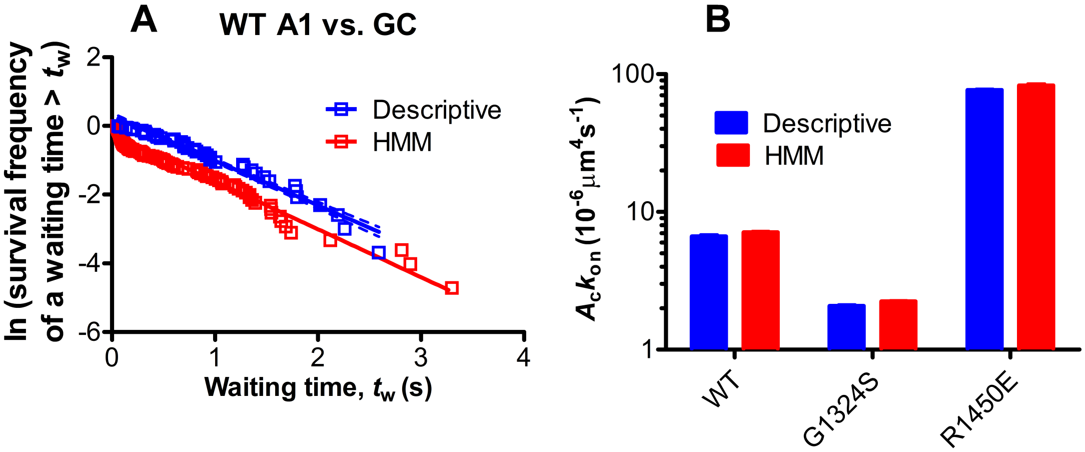

Figure 11: *A*. Exponential waiting time distributions for the interaction of WT A1 and GC. An ensemble of 40 waiting times, defined as the intervals from the moment of a bond dissociation to the moment of the next bond association, was measured by the descriptive statistical method and pooled (blue squares). Another ensemble of 200 waiting times was measured by HMM from the same raw data and pooled (red squares). For each method, the natural log of the survival frequency with waiting times $> t_w$ was plotted against t_w and fitted by a straight line (solid line). The negative slopes of the best-fits represent the cellular on-rate $k^c_{on} = m_r m_l A_c k_{on}$ estimated by the two methods. The variations in these values are shown by the 95% confidence interval of the best-fit (dotted lines). The red dotted lines are obscured because they overlap with the red solid line. *B*. Comparison of effective on-rate $A_c k_{on}$ estimated by descriptive statistical and HMM methods for WT, G1324S (Type 2M) and R1450E (Type 2B) A1s vs. GC. Ackon was calculated by dividing k^c_{on} by the product of the protein densities on the probe (m_l for A1) and target (m_r for GC) beads, i.e. $m_r m_l = 1.96, 2.8$ and $0.19 \times 10^5 \mu m^{-4}$ determined by flow cytometry for respective conditions. The error bars indicate the 95% confidence interval for each method.

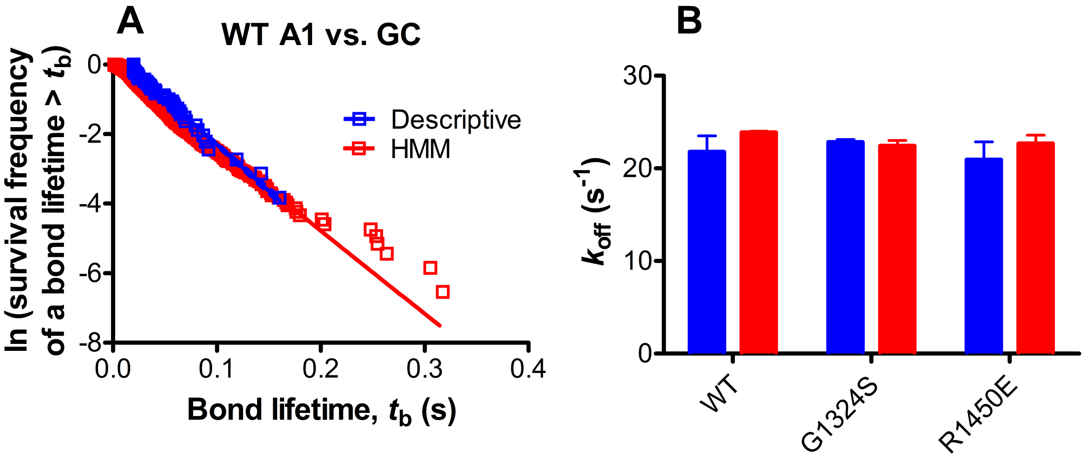

Figure 12: *A*. Exponential bond lifetime distributions for the interaction of WT A1 and GC. An ensemble of 50 bond lifetimes, defined as the time span from association to dissociation of one bond, was pooled by the descriptive statistical method (blue squares). Another ensemble of 200 bond lifetimes was measured by the HMM method from the same raw data and pooled (red squares). For data obtained by each method, the natural log of the survival frequency with bond lifetimes $> t_b$ was plotted against tb and fitted by a straight line. The negative slopes of the best-fits represent the off-rate koff. *B*. Comparison of off-rates estimated by the descriptive statistical and HMM for WT, G1324S (Type 2M) and R1450E (Type 2B) A1s vs. GC. The error bars show the 95% confidence interval for each method.

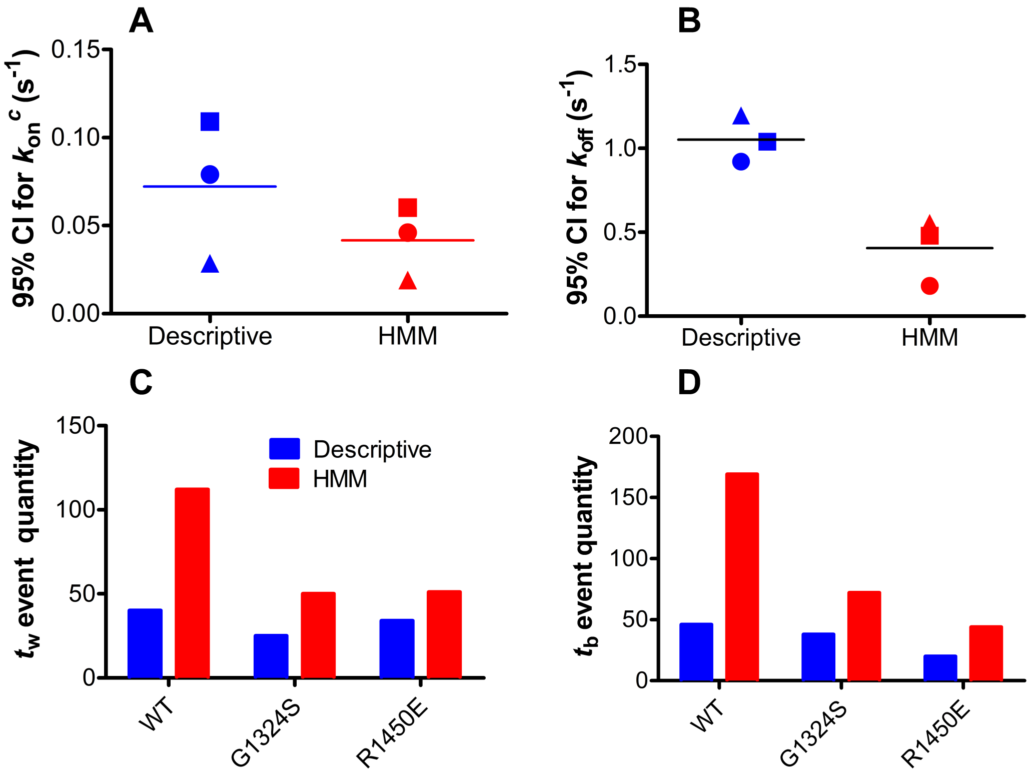

Figure 13: *A* and *B*. Errors (measured as 95% confidence interval, CI) of the estimated cellular on-rates k_{on}^{c} (*A*) and off-rates k_{off} (*B*) for 2D binding kinetics of GPIbαVWF-A1 interaction under the following biological conditions: the wild-type (WT) VWF-A1 (circles), the loss-of-function VWF-A1 mutant G1324S (squares) and the gain-of-function VWF-A1 mutant R1450E (triangles). The errors were plotted for both the descriptive statistical method (blue) and the HMM method (red). *C* and *D*. The numbers of waiting times (*C*) and bond lifetimes (*D*) that the descriptive statistical method (blue) and the HMM method (red) are respectively capable of measuring from the same set of raw data.

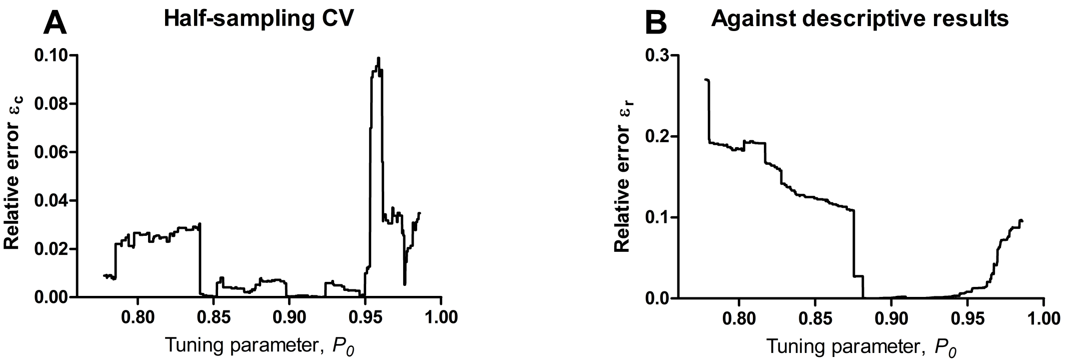

Figure 14: *A*. Half-sampling cross validation. The relative error of off-rate from odd sequence vs. off-rate from even sequence was plotted against P_0. *B*. The relative error of off-rate vs. P_0 by comparing the HMM with descriptive statistical method with the same data as the whole sequence.

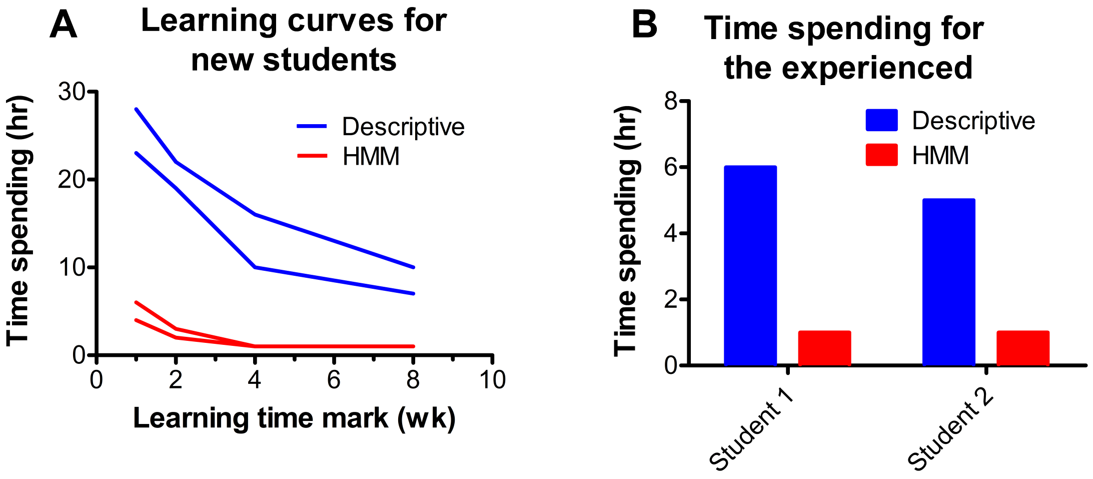

Figure 15: *A*. Comparison of the times spent by a new student to learn the descriptive statistical method (blue) and the HMM (red). Two students who were new to both methods were surveyed. The times for them to finish analysis of one data set were plotted vs. different time checkpoints. Each curve represents a surveyed student. *B*. Comparison of the times spent by the experienced students to analyze the same set of raw data using the descriptive statistical method (blue) and the HMM (red). Two students were surveyed.

CHAPTER III

BAYESIAN CUBIC SPLINE IN COMPUTER EXPERIMENTS

3.1 Introduction

Because of the advances in complex mathematical models and fast computation, computer experiments have become popular in engineering and scientific investigations. Computer simulations can be much faster or less costly than running physical experiments. Furthermore, physical experiments can be hard to conduct or even infeasible when only rare events like land slide or hurricane are observed. There are many successful applications of computer experiments as reported in the literature. Gaussian process (GP) has been used as the main tool for modelling computer experiments. See the books by Santner, Williams and Notz (2003), Fang, Li and Sudjianto (2005), and the November 2009 issue of Technometrics, which was devoted to computer experiments.

First we introduce the GP model. Suppose an experiment involves k factors $\mathbf{x} = (x_1, \ldots, x_k)^t$ and n computer runs are performed at $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. We can write the input as the $n \times k$ matrix $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^t$. The corresponding response values is the vector $\mathbf{Y}_D = (y_1, \ldots, y_n)^t$. The GP model assumes that

$$y(\mathbf{x}) = \mathbf{b}^t\mathbf{f}(\mathbf{x}) + Z(\mathbf{x}), \tag{8}$$

where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^t$ is a vector of m known regression function, $\mathbf{b} = (b_1, \ldots, b_m)^t$ is a vector of unknown coefficient, and $Z(\mathbf{x})$ is a stationary GP with mean zero, variance σ^2 and correlation function $\mathrm{corr}(y(\mathbf{x}_1), y(\mathbf{x}_2)) = R(\mathbf{x}_1, \mathbf{x}_2) = R(\|\mathbf{x}_1 - \mathbf{x}_2\|)$. For the GP model in (8), the best linear unbiased predictor (BLUP) of $y(\mathbf{x})$ is an interpolator, which will be shown in (12).

One popular choice of the correlation function is the product exponential correlation function with power two. The one-dimension powered exponential correlation function with $k = 1$ and $\mathbf{x} \in \mathbb{R}$ can be written as:

$$R(d) = \exp(-\theta d^2), \tag{9}$$

where $d = \|\mathbf{x}_1 - \mathbf{x}_2\|$ is the distance between two input values $\mathbf{x}_1$ and $\mathbf{x}_2$, and θ is the scale parameter. It has been used in many applications (O'Hagan 1978, Sacks and Schiller 1988, Sacks, Schiller, Welch 1989 and Abrahamsen 1997) and software including JMP 8.0.2 2010. However, a process $y(\mathbf{x})$ with (9) as the correlation function has the property that its realization on an arbitrarily small, continuous interval determines the realization on the whole real line. This global influence of local data are considered unrealistic and possibly misleading in some applications (Diggle et al. 2007, p. 54). We shall refer to this property as *global prediction*. Another well known correlation function is the Matérn family (Matérn, 1960). For the one-dimension case, it is a two-parameter family:

$$R(d) = \{2^{\nu-1}\Gamma(\nu)\}^{-1}(d/\phi)^{\nu}K_{\nu}(d/\phi),$$

where $K_{\nu}(\cdot)$ denotes a modified Bessel function of order $\nu > 0$, and $\phi > 0$ is a scale parameter for the distance d. As $\nu \to \infty$ the Matérn correlation function converges to (9).

Another commonly used interpolation method is the spline. An order-s spline with knots ξ_i, $i = 1, ..., l$ is a piecewise-polynomial of order s and has continuous derivatives up to order $s - 2$. A cubic spline has $s = 4$. GP may also be viewed as a spline in a reproducing kernel Hilbert space, with the reproducing kernel given by the covariance function (Wahba 1990). The main difference between them is in the interpretation. While the spline is driven by a minimum norm interpolation based on a Hilbert space structure, GP is driven by an expected squared prediction error based on a stochastic model.

In this paper we will focus on the cubic spline by considering it in the GP framework via the cubic spline correlation function (Currin et al. 1991, Santner et al. 2003):

$$R(d) = \begin{cases} 1 - 6(\frac{d}{\theta})^2 + 6(\frac{|d|}{\theta})^3, & \text{if } |d| < \frac{\theta}{2}, \\ 2(1 - \frac{|d|}{\theta})^3, & \text{if } \frac{\theta}{2} \leq |d| < \theta, \\ 0, & \text{if } |d| \geq \theta, \end{cases} \tag{10}$$

where $\theta > 0$ is the scale parameter. Currin et al. (1991) showed that the BLUP with the function in (10) as the correlation function gives the usual cubic spline interpolator. An advantage of the cubic spline correlation is that θ can be made small, which permits prediction to be based on data in a local region around the predicting location (Santner et al. 2003, p. 38). We shall refer to this property as *local prediction.*

In this paper, we introduce a Bayesian version of the Gaussian process approach for the cubic spline correlation function given in (10). One advantage of Bayesian prediction is that the variability of $y(\mathbf{x})$ given observations can be used to provide measures of posterior uncertainty and designs can be sought to minimize the expected uncertainty (Ylvisaker 1987, Sacks, Welch, Mitchell, and Wynn 1989). Some empirical studies have shown the superiority of Gaussian process over the other interpolating techniques including splines (see Laslett 1994). Here we show the potential advantage of using Bayesian cubic spline in the GP model compared to the powered exponential correlation function (9) through simulation studies.

The paper is organized as follows. In Section 3.2, we give a brief review on the kriging technique based on GP models. In Section 3.3.1, we develop a Bayesian version of the cubic spline method, abbreviated as BCS. In Section 3.3.2, a nugget parameter is introduced to the GP model underlying the BCS method when numerical and estimation stability is required and a summary procedure for the BCS is given. In Section 3.3.3, we consider its extension to high dimensions. In Section 3.4, BCS is compared with two competing procedures in three simulation examples: GP based on the cubic spline correlation function in (10), abbreviated as CS and GP based on the powered exponential correlation function in (9), abbreviated as PE. CS and PE will be explained with details in Section 3.2. In Section 3.5, we compare the performance of BCS and PE on some real data. Some concluding remarks are given in Section 3.6.

3.2 A brief review on kriging

The GP model has been used in geostatistics, known as kriging (Matheron 1963, Cressie 1992, Diggle et al. 2007). Kriging is used to analyse spatially referenced data which have the following characteristics (Cressie 1992). The observed values y_i are at a discrete set of sampling locations $\mathbf{x}_i$, $i = 1, \ldots, n$, within a spatial region. The observations y_i are statistically related to the values of an underlying continuous spatial phenomenon $S(\mathbf{x}_i)$ (Diggle et al. 2007). Sacks et al. (1989) proposed kriging as a technique for developing meta models from computer experiment. Computer experiment produces a response for a given set of input variables. Here we only consider deterministic computer experiment, i.e, the code produces identical answers if run twice using the same set of inputs.

Suppose we want to provide the prediction of a function $y(\mathbf{x})$ at an untried location $\mathbf{x}$, given the observed y values at $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^t$. For the Gaussian process in (8), the best linear unbiased estimator (BLUE) of $\mathbf{b}$ is

$$\hat{\mathbf{b}} = (\mathbf{f}_D^t \mathbf{R}_D(\hat{\theta})^{-1}\mathbf{f}_D)^{-1}\mathbf{f}_D^t \mathbf{R}_D(\hat{\theta})^{-1}\mathbf{Y}_D, \tag{11}$$

where $\mathbf{f}_D = \mathbf{f}(D) = (\mathbf{f}(\mathbf{x}_i))_{\mathbf{x}_i \in D}$, dependence on θ is now explicitly indicated in the notation and $\hat{\theta}$ is the estimate of θ. The BLUP of $\mathbf{Y}_0 = y(\mathbf{x}_0)$ at $\mathbf{x}_0 \in \mathbb{R}$ is

$$\hat{\mathbf{Y}}_0 = \hat{\mathbf{b}}^t\mathbf{f}_0 + \mathbf{R}_0(\hat{\theta})^t\mathbf{R}_D(\hat{\theta})^{-1}(\mathbf{Y}_D - \hat{\mathbf{b}}^t\mathbf{f}_D), \tag{12}$$

where $\hat{\mathbf{b}}$ is given in (11), $\mathbf{f}_0 = \mathbf{f}(\mathbf{x}_0)$ and $\mathbf{R}_0 = (R(\mathbf{x}_0 - \mathbf{x}_1), \ldots, R(\mathbf{x}_0 - \mathbf{x}_n))^t$ is the $n \times 1$ vector of correlations between Y_D and Y_0. If we denote $\mu_D = \mathbf{b}^t\mathbf{f}_D$ and $\mu_0 = \mathbf{b}^t\mathbf{f}_0$, then (12) becomes

$$\hat{\mathbf{Y}}_0 = \hat{\mu}_0 + \mathbf{R}_0(\hat{\theta})^t \mathbf{R}_D(\hat{\theta})^{-1}(\mathbf{Y}_D - \hat{\mu}_D).$$

One way to estimate θ and σ^2 is through the maximum likelihood. Maximum likelihood is a commonly used method for estimating parameters in both computer experiments and spatial process models (Wecker and Ansley 1983; Mardia and Marshall 1984; Currin et al. 1988; Sacks, Schiller, and Welch 1989; Sacks, Welch, Mitchell, and Wynn 1989). For the powered exponential correlation function in (9), the estimate of σ^2 yields

$$\hat{\sigma}^2(\theta) = \frac{1}{n}(\mathbf{Y}_D - \mu_D)' \mathbf{R}_D(\theta)^{-1}(\mathbf{Y}_D - \mu_D).$$

Estimation of θ is usually done by a constrained iterative search. We refer to this method as kriging based on the powered exponential (PE). If we adopt the cubic spline correlation function in (10), the correlation parameter θ is both a scale and truncation parameter. In this case, the estimation method of θ is based on the restricted maximum likelihood method (REML). REML (Patterson and Thompson, 1971) was proposed as a method of obtaining less biased estimates of the variance and covariance parameters than the (unrestricted) maximum likelihood. We refer to this method as kriging based on the cubic spline correlation function (CS).

3.3 Bayesian cubic spline

3.3.1 The prior and posterior processes

As $\mathbf{Y}_D \sim \mathcal{N}(\mu_D, \sigma^2\mathbf{R}_D(\theta))$ and $\mathbf{Y}_0 \sim \mathcal{N}(\mu_0, \sigma^2\mathbf{R}_0(\theta))$, we will develop the Bayesian framework for the cubic spline method, where $\mathbf{R}$ is the cubic spline correlation function in (10) and θ is the scale parameter. For the choice of prior, we assign the non-informative priors to μ_D and θ, the conjugate prior to σ^2 and assume that the priors are independent with each other:

$$\mu_D \sim 1,$$
$$\sigma^2 \mid \alpha, \beta \sim \text{InverseGamma}(\alpha, \beta),$$
$$\theta \mid a \sim \mathcal{U}(0, a),$$
$$\beta \sim 1/\beta,$$

where θ follows the uniform distribution in $[0, a]$. Here θ can be viewed as the *knot location* parameter in spline. In the Bayesian spline literature (Dimatteo et al., 2001, Wang 2008), it is a common practice to assign uniform prior to the knot location parameter. The prior parameter a is fixed as $a = \max_{\mathbf{x}_i, \mathbf{x}_j \in D} \|\mathbf{x}_i - \mathbf{x}_j\|$. The reason for choosing this particular a is because the function in (10) is truncated and equals zero for $|d| \geq \theta$. Because we want the GP to have a local prediction property, we choose a to be the largest distance among the $\mathbf{x}$ values in D. A simulation study (not reported here) shows that a larger range of a does not change the overall performance in estimation. Because an unknown shape parameter α will bring unnecessary complication in the computation, we assume α to be fixed and known.

We use the Markov chain Monte Carlo (MCMC) method to perform the Bayesian computation (Christian and Casella 2004; Gill 2008). It samples from probability distributions by constructing a Markov chain that has the desired distribution as its equilibrium distribution. Gibbs sampling (Casella and George 1992; Gelfand and Smith 1990) is an MCMC algorithm. It can approximate the posterior distribution of parameter of interest by obtaining a sequence of sample values from a specified multivariate probability distribution. Since the marginal posterior distributions of μ_D and β are in closed form, Gibbs sampling can be implemented to obtain the posterior distribution of μ_D and β:

$$\begin{aligned}
\mu_D \mid \mathbf{Y}_D, \theta, \sigma^2 &\propto \mathbb{P}(\mathbf{Y}_D \mid \mu_D, \theta, \sigma^2)\mathbb{P}(\mu_D) \\
&\propto \mathcal{N}(\mu_D, \mathbf{R}_D(\theta, \sigma^2)) \times 1 \\
&\propto \exp\left(-\frac{1}{2}(\mathbf{Y}_D - \mu_D)'\mathbf{R}_D(\theta, \sigma^2)^{-1}(\mathbf{Y}_D - \mu_D)\right) \\
&\sim \mathcal{N}(\mathbf{Y}_D, \mathbf{R}_D(\theta, \sigma^2)), \qquad (13)
\end{aligned}$$

$$\begin{aligned}
\beta \mid \alpha, \sigma^2 &\propto \mathbb{P}(\sigma^2 \mid \alpha, \beta)\mathbb{P}(\beta) \\
&\propto \text{InverseGamma}(\alpha, \beta) \times \beta^{-1} \\
&\propto \beta^{\alpha-1} \exp\left(-\beta/\sigma^2\right) \\
&\sim \text{Gamma}(\alpha, \sigma^2).
\end{aligned}$$

However, the parameters θ and σ^2 are embedded into the covariance function $\sigma^2\mathbf{R}$ in (10) and have no posterior distribution in closed form. Thus we will sample θ and σ^2 using the Metropolis-Hastings (MH) algorithm (Metropolis et al. 1953, Hastings 1970). The MH algorithm works by generating a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution. Specifically, at each iteration, the algorithm picks a candidate for the next sample value based on the current sample value. Then, with some probability, the candidate is either accepted (in which case the candidate value is used in the next iteration) or rejected (in which case the candidate value is discarded, and the current value is reused in the next iteration).

Specifically, we sample θ_{new} and σ^2_{new} by following normal symmetric probability densities with respect to the existing θ and σ^2 (denoted as θ_{old} and σ^2_{old}). The normal symmetric probability densities work as jumping distribution, because they choose a new sample value based on the current sample. In theory, any arbitrary jumping probability density $Q(\delta_{old} \mid \delta_{new})$ can work, where δ is the parameter of interest. Here we choose symmetric $Q(\delta_{old} \mid \delta_{new}) = Q(\delta_{new} \mid \delta_{old})$ for simplicity. The variance term in the normal distribution is the jumping size from old sample to new sample of the MH algorithm. Here we choose the variance to be one. As the size gets smaller, the deviation of the new parameters from the previous one should get small. The jumping probabilities are:

$$\log \theta_{new} \sim \mathcal{N}(\log \theta_{old}, 1), \tag{14}$$

$$\log \sigma^2_{new} \sim \mathcal{N}(\log \sigma^2_{old}, 1). \tag{15}$$

After getting θ_{new} and σ^2_{new}, the acceptance ratios are defined as:

$$\begin{aligned} r_1 &= \frac{\mathbb{P}(\mathbf{Y}_D|\mu_D,\theta_{new},\sigma^2)\mathbb{P}(\theta_{new}|a)}{\mathbb{P}(\mathbf{Y}_D|\mu_D,\theta_{old},\sigma^2)\mathbb{P}(\theta_{old}|a)} \\ &= \frac{|\mathbf{R}_D(\theta_{new},\sigma^2)|^{-1/2}\exp\Big(-\frac{1}{2}(\mathbf{Y}_D-\mu_D)'\mathbf{R}_D(\theta_{new},\sigma^2)^{-1}(\mathbf{Y}_D-\mu_D)\Big)\mathbb{1}_{\theta_{new}\in[0,a]}}{|\mathbf{R}_D(\theta_{old},\sigma^2)|^{-1/2}\exp\Big(-\frac{1}{2}(\mathbf{Y}_D-\mu_D)'\mathbf{R}_D(\theta_{old},\sigma^2)^{-1}(\mathbf{Y}_D-\mu_D)\Big)\mathbb{1}_{\theta_{old}\in[0,a]}}, \end{aligned}$$

$$\begin{aligned} r_2 &= \frac{\mathbb{P}(\mathbf{Y}_D|\mu_D,\theta,\sigma^2_{new})\mathbb{P}(\sigma^2_{new}|\alpha,\beta)}{\mathbb{P}(\mathbf{Y}_D|\mu_D,\theta,\sigma^2_{old})\mathbb{P}(\sigma^2_{old}|\alpha,\beta)} \\ &= \frac{|\mathbf{R}_D(\theta,\sigma^2_{new})|^{-1/2}\exp\Big(-\frac{1}{2}(\mathbf{Y}_D-\mu_D)'\mathbf{R}_D(\theta,\sigma^2_{new})^{-1}(\mathbf{Y}_D-\mu_D)\Big)\sigma_{new}^{-2\alpha-2}\exp(-\beta/\sigma^2_{new})}{|\mathbf{R}_D(\theta,\sigma^2_{old})|^{-1/2}\exp\Big(-\frac{1}{2}(\mathbf{Y}_D-\mu_D)'\mathbf{R}_D(\theta,\sigma^2_{old})^{-1}(\mathbf{Y}_D-\mu_D)\Big)\sigma_{old}^{-2\alpha-2}\exp(-\beta/\sigma^2_{old})}. \end{aligned}$$

We accept θ_{new} (and resp. σ^2_{new}) with probability r_1 (and resp. r_2) if $r_1 < 1$ (and resp. $r_2 < 1$). If $r_1 \geq 1$ (and resp. $r_2 \geq 1$), we accept θ_{new} (and resp. σ^2_{new}).

3.3.2 Nugget parameter

One possible problem with the kriging approach is the potential numerical instability in the computation of the inverse of the correlation matrix in (12). This happens when the correlation matrix is nearly singular. Numerical instability is serious because it can lead to large variability and poor performance of the predictor. The simplest and perhaps most appealing way is to add a nugget effect in the GP modeling. In the spatial statistics literature (Cressie, 1992), a nugget effect is introduced to compensate for local discontinuities in an underlying stochastic process. A well-known precursor is the ridge regression in linear regression analysis. Gramacy and Lee (2012) gave justifications for the use of nugget in GP modeling for deterministic computer experiments. Here we consider the option of adding a nugget parameter in GP model by using ridge regression.

Consider the Gaussian model:

$$Y \sim \mathcal{N}(\mu, \mathbf{R}(\sigma^2, \theta) + \tau^2 I),$$

where τ^2 is the nugget parameter. Adding the matrix $\tau^2 I$ to $\mathbf{R}$ makes the covariance matrix nonsingula and helps stabilize the parameter estimate. We can use the MH sampling to estimate τ^2 by letting $\gamma^2 = \tau^2/\sigma^2$ and assign the prior distribution of γ^2 to be a uniform distribution in the interval of $[0, \kappa]$, where κ is fixed and known. Simply replace $\mathbf{R}_D$ and $\mathbf{R}_0$ in Section 3.3.1 by $\mathbf{V}_D = \mathbf{R}_D + \gamma^2 I$ and $\mathbf{V}_0 = \mathbf{R}_0 + \gamma^2 I$. To use the MH sampling to get γ^2, we choose the jumping distribution

$$\log \gamma^2_{new} \sim \mathcal{N}(\log \gamma^2_{old}, 1), \tag{16}$$

and to sample the new parameter γ^2_{new} based on the current γ^2_{old} by using the acceptance ratio

$$\begin{aligned} r_3 &= \frac{\mathbb{P}(\mathbf{Y}_D|\mu_D,\gamma^2_{new},\theta,\sigma^2)\mathbb{P}(\gamma^2_{new}|\kappa)}{\mathbb{P}(\mathbf{Y}_D|\mu_D,\gamma^2_{old},\theta,\sigma^2)\mathbb{P}(\gamma^2_{old}|\kappa)} \\ &= \frac{|\mathbf{V}_D(\gamma^2_{new},\theta,\sigma^2)|^{-1/2}\exp\big(-\frac{1}{2}(\mathbf{Y}_D-\mu_D)'\mathbf{V}_D(\gamma^2_{new},\theta,\sigma^2)^{-1}(\mathbf{Y}_D-\mu_D)\big)\mathbb{1}_{\gamma^2_{new}\in[0,\kappa]}}{|\mathbf{V}_D(\gamma^2_{old},\theta,\sigma^2)|^{-1/2}\exp\big(-\frac{1}{2}(\mathbf{Y}_D-\mu_D)'\mathbf{V}_D(\gamma^2_{old},\theta,\sigma^2)^{-1}(\mathbf{Y}_D-\mu_D)\big)\mathbb{1}_{\gamma^2_{old}\in[0,\kappa]}}. \end{aligned}$$

In the computation, we use the criterion introduced by Peng and Wu (2013) to determine whether or not to include a nugget effect. We use the condition number of a matrix as the primary measure of singularity. Formally, the condition number of an $m \times s$ matrix M is defined as $\kappa_r(M) = \|M\|_r \|M^{-1}\|_r$, where $\|M\|_r$ denotes the r-norms of a matrix M, defined by $\|M\|_r = \max_{z \neq 0} \|Mz\|_r / \|z\|_r$, $\|z\|_r = (\sum_i |z_i|^r)^{1/r}$, and $z \in \mathbb{R}^s$. For $r = 2$, it reduces to the standard definition of condition number, that is, the ratio of its maximum eigenvalue over its minimum eigenvalue. See Golub and van Loan (2012) for details. Here we use the LAPACK reciprocal condition estimator in MATLAB to determine whether the covariance matrix $\mathbf{R}$ is ill-conditioned. If $(\kappa_1(\mathbf{R}))^{-1} < 10\epsilon$, where $\epsilon = 2^{-52}$ is the floating-point relative accuracy, then $\mathbf{R}$ is ill-conditioned and we will introduce the nugget effect into the model. Otherwise, we will set the nugget effect to be zero.

With the option of adding a nugget parameter, the steps to perform the Bayesian cubic spline are summarized as follows:

1. Set initial values for μ_D, θ, σ^2 and let $\gamma^2 = 0$.

2. Calculate $\kappa_1(\mathbf{R})$.

3. If $(\kappa_1(\mathbf{R}))^{-1} \geq 10\epsilon$, set $\gamma^2 = 0$, sample μ_D, θ and σ^2 from (13), (14), (15) respectively. If the parameters do not converge, go back to step 2.

4. If $(\kappa_1(\mathbf{R}))^{-1} < 10\epsilon$, use $\mathbf{V} = \mathbf{R} + \gamma^2 I$ instead of $\mathbf{R}$ and sample μ_D, θ, σ^2 and γ^2 from (13), (14), (15) and (16) respectively. Repeat this step until convergence.

5. Calculate the estimate of Y_0 using (12) with μ_D, θ, σ^2 and γ^2.

3.3.3 Extension to high dimensions

For multi-dimensions, let $\mathbf{x} \in \mathbb{R}^k$ and assume the correlation function $\mathbf{R}(\mathbf{x}_i, \mathbf{x}_j) = \prod_{t=1}^{k} \mathbf{R}_t(\mathbf{x}_{i,t} - \mathbf{x}_{j,t}) = \prod_{t=1}^{k} \mathbf{R}_t(d_t)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are in $\mathbb{R}^k$, d_t is the distance of $\mathbf{x}_i$ and $\mathbf{x}_j$ on the tth dimension and $\mathbf{R}_t$ is the correlation function for the tth dimension.

The multi-dimensional spline correlation function $\mathbf{R}(d)$ is the product of the one-dimension spline correlation function with individual parameter θ_t estimated for each dimension (Ylvisaker 1975, Chen, Gu, and Wahba 1989). The corresponding Bayesian computation is done by doing the MH sampling for each dimension until convergence. Our criterion for convergence is when the change of $\|\theta\|$ between consecutive iterations of the MCMC computation is smaller than 10^{-4}. Most times θ converges fairly fast in our simulation studies.

3.4 Simulation study and results

First, we compare the performance of the proposed Bayesian cubic spline (BCS) method with two other methods: PE and CS (described in Section 3.2). The criterion for evaluating the performance of the estimators is the integrated mean squared error (IMSE), defined as

$$\mathrm{IMSE}(\hat{f}) = \int_{\Omega} (\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\mathbf{x},$$

where f and $\hat{f}$ are respectively the true function values and estimated values and Ω is the region of the $\mathbf{x}$ values. The following mean squared error (MSE) is a finite-sample approximation to the IMSE:

$$\mathrm{MSE}(\hat{f}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2, \tag{17}$$

where m is the number of randomly selected points $\{\mathbf{x}_i\}$ from Ω. Three choices of the true function $f(x)$ are considered in Examples 1-3, which range from low to high dimensions and from smooth to non-smooth functions.

Example 1.

$$f_1(\mathbf{x}) = \{1 - \exp(-.5/x_2)\} \frac{2300x_1^3 + 1900x_1^2 + 2092x_1 + 60}{100x_1^3 + 500x_1^2 + 4x_1 + 20}.$$

This two-dimensional function is from Currin et al. (1991), where $\mathbf{x} \in [0,1]^2$ and $f_1 \in [4.1, 13.8]$. We scale f_1 into $[0,1]$. Currin et al. (1991) studied a 16-run design in their paper. Four designs are considered: 4^2 design (16 runs) with levels $(.125, .375, .628, .875)$ (Joseph 2006) and $(0, .3333, .6667, 1)$ (Currin et al. 1991), 5^2 design (25 runs) with levels $(0, .25, .5, .75, 1)$, and 6^2 design (36 runs) with levels $(0, 0.2, 0.4, 0.6, 0.8, 1)$. Four types of noise ϵ are added to $f_1(\mathbf{x})$: $\mathcal{U}(0,0)$ (no noise), $\mathcal{U}(0,.2)$, $\mathcal{U}(0,.5)$ and $\mathcal{U}(0,1)$. As the range of the noise increases from 0 to 1, the function $f_1(\mathbf{x}) + \epsilon$ becomes more rugged. It allows us to compare the performance of the three methods as the true function become less smooth.

For noise based on $\mathcal{U}(0,0)$ (and resp. $\mathcal{U}(0,.2)$, $\mathcal{U}(0,.5)$ and $\mathcal{U}(0,1)$), we conduct the simulation as follows. First, a noise is randomly sampled from $\mathcal{U}(0,0)$ (and resp.

$\mathcal{U}(0,.2)$, $\mathcal{U}(0,.5)$ and $\mathcal{U}(0,1)$). For each simulation, the noise is fixed and denoted as $\{\epsilon_1,\ldots,\epsilon_n\}$. Here n, the number of design points, is 16, 16, 25 and 36 respectively for the four designs. Second, the values of $\{f_1(\mathbf{x}_1),\ldots,f_1(\mathbf{x}_n)\}$ are calculated. Then the values of $\{f_1(\mathbf{x}_1)+\epsilon_1,\ldots,f_1(\mathbf{x}_n)+\epsilon_n\}$ are treated as the response values by PE, BCS and CS in parameter estimation. The purpose of this step is to facilitate the study of robustness of estimation against noises. Then, MSE (see (17)) is calculated on $m = 100$ of $\mathbf{x}$ randomly sampled from $[0,1]^2$. We sample repeatedly and independently $\{\epsilon_1,\ldots,\epsilon_n\}$ and $\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$ for each simulation and record the average of the MSE values from 500 simulations for each noise and design setting in Table 4. For each simulation setting, the method with the smallest MSE is highlighted in boldface.

The MCMC iterations of the BCS terminate if the change of parameter estimate is smaller than 10^{-4}. The running time takes about 20 seconds for an Intel Xeon CPU with 2.66 GHz and 3.00 GB of RAM to reach such convergence. The powered exponential method performs best when the noise is small ($\mathcal{U}(0,0)$ and $\mathcal{U}(0,.2)$) or the design size is large (36-run). This is because the example is relatively smooth when noise is small and the function f_1 contains the exponential term $\exp(-.5/x_2)$, which can be best captured by the non-zero exponential correlation function. PE also benefits from the larger sample size of 6^2, which helps to stabilize estimate. For relatively small designs (16- and 25-run) with large noise ($\mathcal{U}(0,.5)$ and $\mathcal{U}(0,1)$), CS and BCS perform better than PE. When design is small and noise is large, the response surface tends to be very rugged and there is not enough data for PE to estimate the surface with good precision. A localized estimate like CS and BCS with truncated correlation function give smaller MSE. BCS in most cases outperforms CS. The over-smoothing property of PE can result in large estimation errors as will be shown in Example 2 and in Section 3.5.

Example 2.

$$f_2(x) = 0.3\exp^{-1.4x}|\cos(10\pi x)| + 3x.$$

Table 4: Average MSE Values for PE, BCS and CS Predictors in Example 1

		$\mathcal{U}(0,0)$	$\mathcal{U}(0,.2)$	$\mathcal{U}(0,.5)$	$\mathcal{U}(0,1)$
	PE	**1.0034**	**1.0437**	1.7192	1.8951
4^2(J)	BCS	1.0852	1.1116	1.6830	**1.8198**
	CS	1.1036	1.2518	**1.6714**	1.9854
	PE	**1.1223**	**1.3495**	1.6545	2.4491
4^2(C)	BCS	1.1662	1.4021	**1.5762**	**2.1108**
	CS	1.1710	1.4322	1.6318	2.1476
	PE	**0.9087**	**1.0186**	1.3516	1.4845
5^2	BCS	1.0305	1.0391	**1.2473**	**1.3642**
	CS	1.0481	1.1204	1.2665	1.5816
	PE	**0.2171**	**0.2588**	**0.5604**	**0.6352**
6^2	BCS	0.2503	0.2954	0.6117	0.7723
	CS	0.2942	0.2769	0.7479	0.8042

This is a one-dimension function and contains a non-smooth term $|\cos(10\pi x)|$. Here f_2 is scaled into $[0,1]$. As in Example 1, four types of random noise are added to f_2 and 5, 10, 20 and 30 design points (i.e., $\{x_1,\ldots,x_n\}$ values) are uniformly sampled from $[0,1]$. In each simulation, noise is sampled and fixed, denoted as $\{\epsilon_1,\ldots,\epsilon_n\}$, where $n = 5$ (and resp. 10, 20 and 30). Then 5 (and resp. 10, 20 and 30) design locations $\{x_1,\ldots,x_n\}$ are uniformly sampled from $[0,1]$. The values of $\{f_2(x_1)+\epsilon_1,\ldots,f_2(x_n)+\epsilon_n\}$ are used as the response values. The MSE for each simulation is calculated on $m = 100$ randomly sampled x values in $[0,1]$. For each design, we repeat this procedure 500 times by taking random samples of $\{\epsilon_i\}_{i=1}^n$ and $\{\mathbf{x}_i\}_{i=1}^n$. The average MSE based on the 500 simulations is given in Table 5 for each noise and design setting . Again, the method with the smallest average MSE is highlighted in boldface.

In all cases, CS and BCS beat PE even when no noise is added to the true function. BCS performs better than CS in most cases. CS performs better than BCS in four cases, three of which the difference is not significant. PE gives much worse results when the design size is small (5, 10) and the noise is large ($\mathcal{U}(0,.5)$ and $\mathcal{U}(0,1)$). This is due to the global prediction property of PE. For non-smooth functions, this can bring in unnecessarily large errors. On the other hand, the better performance of CS and BCS benefits from their local prediction property.

Table 5: Average MSE Values for PE, BCS and CS Predictors in Example 2

		$\mathcal{U}(0,0)$	$\mathcal{U}(0,.2)$	$\mathcal{U}(0,.5)$	$\mathcal{U}(0,1)$
5	PE	0.0026	0.2857	0.7214	0.8147
	BCS	**0.0018**	0.0764	**0.2219**	**0.2712**
	CS	0.0021	**0.0518**	0.2768	0.2853
10	PE	0.0017	0.0389	0.1626	0.3260
	BCS	0.0014	0.0262	**0.0514**	**0.1929**
	CS	**0.0013**	**0.0259**	0.0591	0.1964
20	PE	0.0015	0.0092	0.1443	0.1547
	BCS	**0.0004**	**0.0051**	**0.0757**	**0.1190**
	CS	0.0011	0.0063	0.1125	0.1248
30	PE	0.0011	0.0049	0.0754	0.1853
	BCS	**6.80E-04**	0.0034	**0.0311**	**0.1775**
	CS	7.20E-04	**0.0028**	0.0596	0.2005

Example 3.

$$f_3(\mathbf{x}) = \frac{2\pi x_1(x_2 - x_3)}{\ln(x_4/x_5)[1 + \frac{2x_1x_6}{\ln(x_4/x_5)x_5^2x_7} + \frac{x_1}{x_8}]}.$$

This is an 8-dimensional smooth function from Morris et al. (1993), where $x_1 \in [63070, 115600]$, $x_2 \in [990, 1110]$, $x_3 \in [700, 820]$, $x_4 \in [100, 5000]$, $x_5 \in [.05, .15]$, $x_6 \in [1120, 1680]$, $x_7 \in [9855, 12046]$ and $x_8 \in [63.1, 116]$. Here we scale $x_1, \ldots, x_8$ and f_3 into $[0, 1]$. Morris et al. (1993) proposed a 10-run design with two levels 0 and 1 based on the maximin distance criterion (see Table 6). In the study, we consider the 10-run design together with 10-, 20- and 50-run Latin hypercube design (McKay et al. 1979). A n-run Latin hypercube design in $[0, 1]^k$ is based on the Latin hypercube sampling. For each dimension, we independently sample n values randomly from each interval $(0, 1/n)$, …, $(1 - 1/n, 1)$ and randomly permute the n values. Here we apply the maximin criterion to choose the Latin hypercubes, i.e., maximizing the minimum distance between points. As before, four types of noise are added to the true function. In each simulation, after the noise $\{\epsilon_1, \ldots, \epsilon_n\}$ is sampled, one Latin hypercube design is generated, denoted by $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $n = 10, 20, 50$. Then apply PE, CS and BCS to $\{f_1(\mathbf{x}_1)+\epsilon_1,\ldots,f_1(\mathbf{x}_n)+\epsilon_n\}$ for parameter estimation. The MSE is calculated based on $m = 5000$ random samples $\{\mathbf{x}_i\}_{i=1}^{5000}$ in $[0,1]^8$. The simulations are repeated 1000 times and the average MSE values are given in Table 7. The running time for each simulation of BCS is less than 2 minutes on the same machine.

The results are similar to those of Example 1. This is expected as they are both smooth functions. PE gives best results among the three methods when the noise is small ($\mathcal{U}(0,0)$ and $\mathcal{U}(0,.2)$) or the sample size is large (50LH). BCS and CS perform well when sample size is small (10 and 20) and the noise is large ($\mathcal{U}(0,1)$). BCS generally outperforms CS.

Table 6: A Maximin Distance Design in $[0,1]^8$ for n=10 (Morris et al. 1993)

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	1	0	0	1	0	1	1
1	1	1	1	0	0	1	0
1	0	0	1	1	0	0	0
0	1	0	0	1	1	0	0
1	1	0	1	0	1	0	1
0	1	1	0	0	0	0	1
0	0	1	1	1	0	1	1
0	0	0	0	0	1	1	1
0	0	1	1	0	1	0	0
1	0	1	0	1	1	1	0

3.5 Application

Instead of simulating from known functions, we perform another comparison study by using the methane combustion data from Mitchell and Morris (1993). Table 8 shows its 50-run design. In addition, Mitchell and Morris gave 20-, 30-, 40- and 50-run, 7-variable maximin design in their paper. The first 20, 30, 40 and 50 runs in Table 8 consist of these designs denoted as D20, D30, D40 and D50. The response y is the logarithm of the ignition delay time.

Table 7: MSE Values for PE, BCS and CS Predictors in Example 3

		$\mathcal{U}(0,0)$	$\mathcal{U}(0,.2)$	$\mathcal{U}(0,.5)$	$\mathcal{U}(0,1)$
	PE	**0.0386**	**0.0457**	**0.0695**	0.1850
10	BCS	0.0472	0.0610	0.0846	**0.1557**
	CS	0.0501	0.0672	0.0884	0.1691
	PE	**0.0277**	**0.0473**	**0.0721**	0.1821
10LH	BCS	0.0412	0.0612	0.0891	0.1592
	CS	0.0458	0.0632	0.0876	**0.1409**
	PE	**0.0013**	**0.0125**	**0.0405**	0.1714
20LH	BCS	0.0053	0.0358	0.0774	**0.1454**
	CS	0.0077	0.0298	0.0868	0.1621
	PE	**5.10E-04**	**0.0096**	**0.0311**	**0.1355**
50LH	BCS	0.0041	0.0137	0.0532	0.1414
	CS	0.0043	0.0159	0.0581	0.1523

Before conducting the comparison, a careful data analysis is performed to show some feature of the data. For D20, D30 and D50, we randomly take 90%, 80% and 50% of the original data as the response values for PE and BCS to estimate θ. The average values of $\hat{\theta}_j$, $j = 1, \ldots, 7$, based on 100 simulations are calculated and given in Table 9 for each setting. For each design, the value of $\hat{\theta}$ from PE increases as the number of input data decreases while the values $\hat{\theta}$ for BCS are more stable. The divergent behavior between PE and BCS for this data can be explained by their respective global and local prediction properties. First, note that a larger θ value indicates a more smooth surface. As the size of input data gets smaller, the data points are spread more thinly in the design region $[0,1]^7$. The fitted response surface by PE will become more smooth due to its global prediction property. The change will not be as dramatic for BCS thanks to its local prediction property. Even though we do not know what the true response surface is or how rugged it is, this divergent behavior seems to suggest that BCS is a better method for the data and this will be confirmed in the next study based on cross validation.

Table 8: Methane Combustion Data

Run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	0	0	0	0.5	1	1	0.25	7.9315
2	0.25	0.5	0.5	0.75	0	1	0	6.2171
3	0	1	0	0.25	0	0	1	7.8535
4	0.5	0.5	0.75	0	1	0.25	0	7.5708
5	0	0.75	0.75	1	1	1	0.5	6.3491
6	1	0	1	0.25	0	1	0	5.3045
7	0	1	0	0.75	1	0.5	1	8.5372
8	0.75	0.25	0	1	1	0	1	7.871
9	0.5	0.75	0.25	0	0.25	0.5	0.5	7.8725
10	0.25	1	0.75	0.75	0.5	0	0.25	6.593
11	0.5	0	1	0.25	1	0.75	1	6.2131
12	1	0	0	0.5	0.5	0.5	1	7.6311
13	1	0.5	1	0.75	0	0.25	0.5	5.109
14	0	1	0.25	0.25	0.75	1	0	8.4206
15	1	1	0	1	0.25	0	0.5	7.2242
16	0.5	0	0.25	1	0	0.25	0.75	6.0216
17	1	1	1	1	0.5	1	0	5.3495
18	1	1	0.5	0.25	0	1	1	6.0325
19	1	0	1	0	0.75	0	0.25	6.4065
20	0.5	1	0.75	1	0.25	0.75	1	5.5674
21	0.25	0	0.5	0.25	0.25	0	1	6.5214
22	0	0.5	0.5	0	0.75	0.5	0.75	7.7907
23	0.25	0	0	0.25	0	0.75	0.5	7.3542

Table 8 – *Continued from previous page*

Run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
24	0.75	0.75	1	0	0	0	0	5.8651
25	0.25	0	1	0.5	0.75	0.5	0	6.4489
26	0.75	1	0.25	0.75	1	0.75	0.25	7.6225
27	0	1	1	0.5	0.25	1	0.5	5.8572
28	1	0.5	1	0	1	1	0.5	6.5656
29	1	0.25	1	1	1	0.25	0	5.7137
30	0	0	0	1	0.25	1	1	6.5603
31	0.5	0	0.5	0	0.75	1	0.25	7.5044
32	1	0	0.75	0.75	0.5	0.75	0.25	5.8721
33	1	0	0	0	1	0.25	0	8.206
34	1	0.5	0	1	0	0.75	1	6.3746
35	0.25	0.5	1	0.75	0.5	1	1	5.4478
36	1	0.75	0	0	0.5	1	0.75	7.6953
37	0.5	0	1	0	0	0.5	0.75	5.3423
38	0.5	1	0	1	0	1	0.25	6.4493
39	1	0.25	0	0.5	0	0.5	0	6.8957
40	0.75	0.5	0.25	0.5	1	1	1	7.5563
41	0.75	0.75	0.25	0.5	0	0.25	1	6.7549
42	0.75	0	0.75	0.75	0	1	1	5.0056
43	0	0.25	0.25	1	0.75	0.25	0.75	7.4006
44	1	0.25	0.75	0	0.25	0	1	5.6656
45	0.5	0.5	0.5	0.5	0.5	0	0	7.4111
46	0.75	1	0.75	0.25	0.25	0.75	0.25	6.7111
47	1	0.5	0.25	0.25	0.75	0.75	0	7.9182

Table 8 – *Continued from previous page*

Run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
48	1	0.25	0.5	1	1	1	0.5	6.2543
49	0	0.75	0.5	0.75	0.25	0.25	0.5	6.7319
50	0.25	1	0.25	1	0.75	1	0.75	6.9749

Table 9: Average $\hat{\theta}$ values using PE and BCS

	Method	% of input	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$	$\hat{\theta}_7$
D20	PE	90	0.1137	0.0368	0.2697	0.0606	0.0712	0.0101	0.0137
D20	PE	80	0.1892	0.0878	0.4891	0.0928	0.1082	0.0348	0.0647
D20	PE	50	0.4092	0.2977	0.7952	0.3177	0.3580	0.2476	0.2681
D20	BCS	90	4.1345	4.0635	3.6870	3.8788	3.8440	4.2943	3.3749
D20	BCS	80	4.5264	4.5190	3.6424	3.8407	4.5733	4.6059	4.0459
D20	BCS	50	4.1244	4.8617	3.8581	3.4646	4.2030	5.5327	3.1355
D30	PE	90	0.0958	0.0318	0.1582	0.0335	0.0700	0.0079	0.0144
D30	PE	80	0.1216	0.0647	0.3231	0.0667	0.1017	0.0292	0.0256
D30	PE	50	0.2293	0.1874	0.5877	0.1790	0.2252	0.1315	0.1348
D30	BCS	90	3.7533	4.9328	4.6481	3.6454	5.2628	4.6647	2.0442
D30	BCS	80	3.4020	5.2822	5.4022	3.4501	5.8690	4.6797	2.5237
D30	BCS	50	4.5891	4.9748	5.8604	3.4380	5.5703	4.9078	2.4123
D50	PE	90	0.1297	0.0441	0.1057	0.0394	0.0529	0.0112	0.0129
D50	PE	80	0.1189	0.0404	0.1676	0.0506	0.0741	0.0093	0.0124
D50	PE	50	0.1200	0.0449	0.2086	0.0641	0.0956	0.0123	0.0155
D50	BCS	90	3.3356	4.9131	4.9231	3.2469	3.8125	3.5908	5.0959
D50	BCS	80	3.3427	5.4804	4.3651	3.4601	3.6676	3.3073	4.9948
D50	BCS	50	3.8516	4.7773	4.5888	3.6960	3.8341	3.3936	5.3051

We now use the same data and design settings to run cross validations on D50, D40 and D30 for each of the three methods. One round of cross-validation involves partitioning the data into complementary subsets, performing the analysis on one subset (called the *training* set), and validating the analysis on the other subset (called the *validation* set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds (Geisser 1993 and Kohavi 1995). Each time we take a fixed number of data out of D50 (resp. D40, D30) and use them for model fitting. The remaining data are used to calculate the MSE in (17). Because CS gives much larger MSE in each case, we only compare the MSE results for PE and BCS. For D50, the results of training data size as 40 and 30 are plotted in Figures 16 and 17. In each figure, one dot indicates the MSE from PE versus the MSE from BCS for a given design. The reference line of 45° indicates that the two designs are equally good since they render the same MSE. When the majority of the dots is below the line, it means BCS has smaller MSE. This is evident in Figure 17. PE gave some very bad predictions with MSE as high as 5.5 (dots in right bottom of Figure 17), while the majority of MSE of BCS centres around 1.5. The average of MSE from 100 simulations for each cross validation setting for BCS and PE and the percentage of BCS outperforming PE are given in Table 10. There is a much larger difference between PE and BCS when the training data size is relatively small (20 and 25). This is probably caused by the global prediction and over-smoothing properties of PE.

Table 10: Comparison of PEM and BCS

Design	Training Data Size	MSE(PEM)	MSE(BCS)	% BCS Better
D50	40	0.6237	0.6524	58
D50	30	2.4980	1.3908	92
D40	30	0.8812	0.7132	61
D40	25	2.8516	1.5590	89
D30	25	1.6108	0.7888	69
D30	20	1.9498	1.2729	86

3.6 Conclusions

Cubic spline is widely used in numerical approximation. In the GP modeling, use of the cubic spline correlation function in (10) can lead to sparse correlation matrix

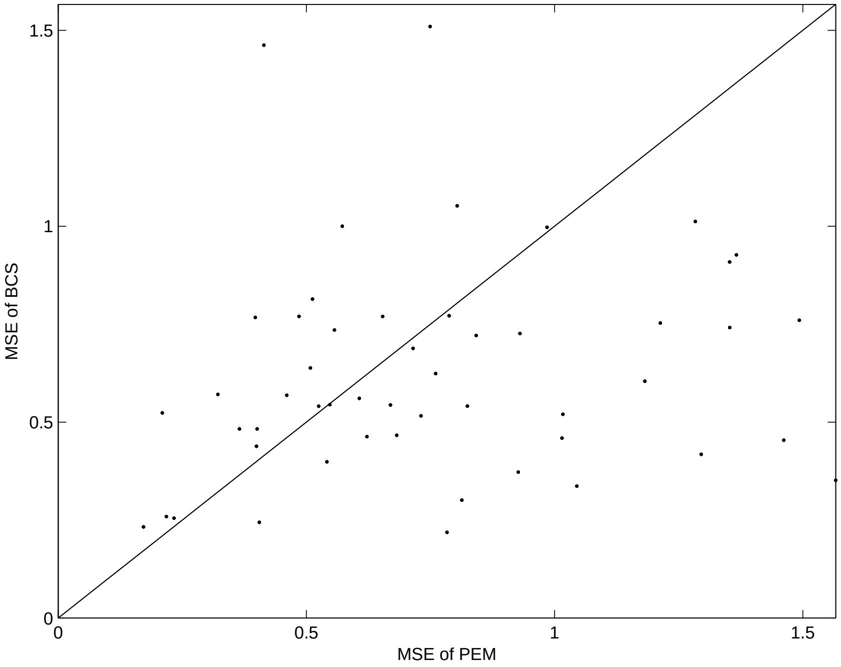

Figure 16: MSE of 40 Training Data in D50

with many zero off-diagonal elements. By comparison the two commonly used correlation functions, the Matérn family and the powered exponential correlation, do not enjoy this property. A sparse correlation matrix can reduce the computation cost and enhance the computation stability. The viability of cubic spline for computer experiment applications received a further boost when JMP 8.0.2 2010 provides the powered exponential correlation and the cubic spline correlation as its *only* two choices in GP modeling. The prominence the JMP software gives to the cubic spline was one motivation for us to develop a Bayesian version of the cubic spline method. By putting a prior on the parameters in the cubic spline correlation function in (10), Bayesian computation can be performed by using MCMC. The Bayesian cubic spline should outperform its frequentist counterpart because of its smoothness and shrinkage properties. It also provides posterior estimates, which enable statistical inference on the parameters of interest.

We compare BCS with CS and PE in a simulation study and application to real

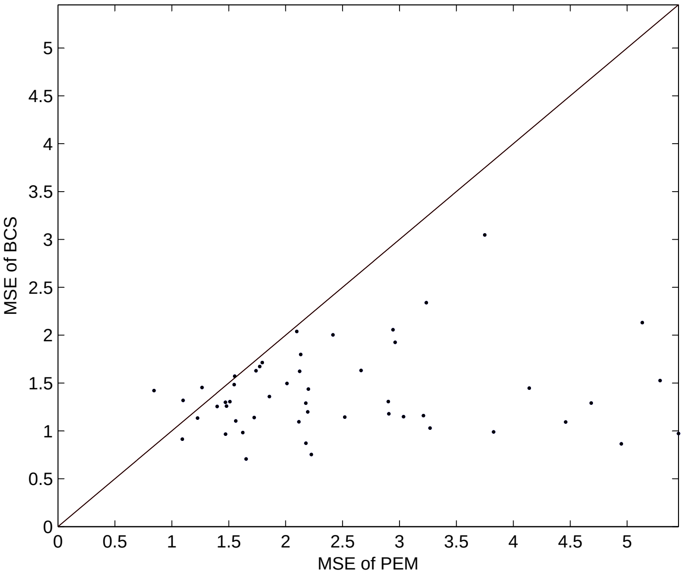

Figure 17: MSE of 30 Training Data in D50

data. We have also considered other correlation functions such as the spherical family

$$R(d) = \begin{cases} 1 - \frac{3|d|}{2\theta} + \frac{1}{2}(\frac{|d|}{\theta})^3, & \text{if } |d| \leq \theta, \\ 0, & \text{if } |d| > \theta. \end{cases}$$

Because the performance of the frequentist version of the spherical family is similar to that of the cubic spline, these results are omitted in the paper. In the three simulation examples, BCS outperforms CS in most cases. PE performs the best when the true function is smooth or the data size is large. BCS and CS perform better than PE when the true function is rugged and the data size is relatively small. This difference in performance can be explained by the local predication property of BCS and CS and the global predication property of PE. Recall that, in global prediction, the prediction at any location is influenced by far-away locations (though with less weights). This leads to over-smoothing, which is not good for rugged surface. Local prediction does not suffer from this as the prediction depends only on nearby locations. In the real data application, BCS outperforms PE in all choices of design. In summary, when the response surface is non-smooth and/or the input dimension is high, the BCS method can have potential advantages and should be considered for adoption in data analysis.

Some issues need to be considered in future work. When the dimension is high, the parameter estimation is based on the MH sampling which can be costly. Grouping the parameter to reduce computation is an alternative. Also, we have considered mostly the non-informative priors. If more information is available, informative prior assignment should be considered.

APPENDIX A

REGULARITY CONDITIONS OF CHAPTER 1

We first state the identifiability conditions for Theorem 1. The hidden Markov model is identifiable if the following conditions are satisfied (MacKay 2002):

A1. The Markov Chain $\{x_i\}$ is irreducible and aperiodic.

A2. The parameter space for (σ, ϕ), denoted by Ω, is compact.

A3. $f(y;\ \sigma, \phi)$ is continuous in σ and ϕ.

A4. Given $\epsilon > 0$, there exists $A > 0$ such that for all $(\sigma, \phi) \in \Omega$, $f(A;\ \sigma, \phi) - f(-A;\ \sigma, \phi) \geq 1 - \epsilon$.

A5. The family of finite mixtures of $\{f(y;\ \sigma, \phi)\}$ is identifiable, i.e.,

$$F(y, G_1) = F(y, G_2) \quad \Rightarrow \quad G_1 = G_2.$$

A6. There is an upper bound on the number of hidden states.

We can see that most of the identifiability conditions are quite natural and hold for many popular distributions.

We need the following regularity conditions:

B1. 1. $E(|\log F(\boldsymbol{Y};\ \boldsymbol{\Sigma}, \boldsymbol{\Phi})\ |) < \infty$.

2. There exists $\rho > 0$ such that $F(\boldsymbol{Y};\ \boldsymbol{\Sigma}, \boldsymbol{\Phi})$ is measurable for each $(\sigma, \boldsymbol{\Phi})$.

B2. $F(\boldsymbol{Y};\ \boldsymbol{\Sigma}, \boldsymbol{\Phi})$ is differentiable with respect to $(\boldsymbol{\Sigma}, \boldsymbol{\Phi})$ to order 3. The derivatives are jointly continuous in $\boldsymbol{Y}, \boldsymbol{\Sigma}, \boldsymbol{\Phi}$.

B3. Let $U_{i,\sigma_1^{n_1},\ldots,\sigma_t^{n_t}}(\boldsymbol{\Sigma}, G, \boldsymbol{\Phi}_i) = \frac{\partial^{n_1+\cdots n_t}}{\partial\sigma_1^{n_1}\partial\sigma_2^{n_2}\cdots\partial\sigma_t^{n_t}} \log F(\boldsymbol{Y}_i;\ \boldsymbol{\Sigma}, \boldsymbol{\Phi}_i)$, where $n_1+n_2+\cdots+n_t \leq 3$. For each atom of G_0, σ_{0k}, there exists a small neighborhood of $(\sigma_{0k}, \boldsymbol{\Phi}_0)$ and a function $q(\boldsymbol{Y})$ with $E\{q^2(\boldsymbol{Y})\} < \infty$ such that for $G, G'', \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'$ and $\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_i'$ in this neighborhood, we have:

$$| \ U_{i,\sigma_1^{n_1},\ldots,\sigma_t^{n_t}}(\boldsymbol{\Sigma}, G, \boldsymbol{\Phi}_i) - U_{i,\sigma_1^{n_1},\ldots,\sigma_t^{n_t}}(\boldsymbol{\Sigma}', G'', \boldsymbol{\Phi}_i') \ | \leq q(\boldsymbol{Y}_i)\{\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}'\| + \|G-G'\| + \|\boldsymbol{\Phi}_i-\boldsymbol{\Phi}_i'\|\}.$$

B4. The matrix with the (k_1, k_2)th element $E\{U_{1,\sigma_m}(\boldsymbol{\Sigma}_{k_1}, G_0, \boldsymbol{\Phi}_0)U_{1,\sigma_m}(\boldsymbol{\Sigma}_{k_2}, G_0, \boldsymbol{\Phi}_0)\}$ is finite and positive definite.

B1 to B4 ensure that the MLE $\hat{G}_n$ of HMM with unknown order K_0 is $\sqrt{n}$-consistent and asymptotically normal.

APPENDIX B

SUPPLEMENTARY PROOFS OF CHAPTER 1

Lemma 1 *Under the conditions in Theorem 1, $\hat{G}_n$ has the following property:*

$$\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1).$$

Proof of Lemma 1. From Jensen's inequality, under condition $B1$ and the identifiability conditions for the hidden Markov Chain: $E\{\log F(\boldsymbol{Y};\ \boldsymbol{\Sigma}, \boldsymbol{\Phi})\} < E\{\log F(\boldsymbol{Y};\ \boldsymbol{\Sigma}_0, \boldsymbol{\Phi}_0)\}$ for any $(G, \boldsymbol{\Sigma}) \neq (G_0, \boldsymbol{\Sigma}_0)$. This would imply:

$$l_n(G, \boldsymbol{\Sigma}) - l_n(G_0, \boldsymbol{\Sigma}_0) \leq -Cn,$$

almost surely for $(G, \boldsymbol{\Sigma}) \neq (G_0, \boldsymbol{\Sigma}_0)$ with some $C > 0$. For $\gamma_n = O_p(n^{1/4} \log n)$ and $a > 2$, the SCAD function satisfies:

$$\sum_{k=1}^{K-1} p_n(\eta_k) - \sum_{k=1}^{K_0-1} p_n(\eta_{0k}) = o(n).$$

Therefore, when the parameter space Ω is compact,

$$\sup_N \tilde{l}_n(G, \boldsymbol{\Sigma}) - \tilde{l}_n(G_0, \boldsymbol{\Sigma}_0) \leq \sup_N l_n(G, \boldsymbol{\Sigma}) - l_n(G_0, \boldsymbol{\Sigma}_0) - o(n) \leq -Cn.$$

Hence, $\hat{G}_n \to G_0$ and it has at least K_0 distinct atoms, which indicates that $\eta_{0k} > 0$ is approximated by one of the $\hat{\eta}_k$. Note that the SCAD penalty is a constant outside a small neighborhood of 0. As a result, $p_n(\hat{\eta}_k) = p_n(\eta_{0k})$ in probability and $\sum_{k=1}^{K-1} p_n(\hat{\eta}_k) - \sum_{k=1}^{K_0-1} p_n(\eta_{0k}) \geq 0$, which implies:

$$0 \leq \tilde{l}_n(G, \boldsymbol{\Sigma}) - \tilde{l}_n(G_0, \boldsymbol{\Sigma}_0) \leq \{l_n(\hat{G}_n, \hat{\boldsymbol{\Sigma}}) - l_n(G_0, \boldsymbol{\Sigma}_0)\} + \{C_K \sum_{k=1}^{K} \log \hat{\pi}_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k}\}.$$

Denote by $(\bar{G}_n, \bar{\boldsymbol{\Sigma}})$ the MLE of $(G, \boldsymbol{\Sigma})$ with at most K atoms. Then

$$l_n(\hat{G}_n, \hat{\boldsymbol{\Sigma}}) - l_n(G_0, \boldsymbol{\Sigma}_0) \leq l_n(\bar{G}_n, \bar{\boldsymbol{\Sigma}}) - l_n(G_0, \boldsymbol{\Sigma}_0) = O_p(1).$$

Therefore,

$$C_K \sum_{k=1}^{K} \log \hat{\pi}_k \geq -l_n(\bar{G}_n, \bar{\boldsymbol{\Sigma}}) + l_n(G_0, \boldsymbol{\Sigma}_0) + C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k} = O_p(1).$$

□

Proof of Theorem 1. i). We have shown in the proof of Lemma 1 that $(\hat{G}_n, \hat{\boldsymbol{\Sigma}})$ is a consistent estimator of $(G_0, \boldsymbol{\Sigma}_0)$.

ii). By Lemma 1, the mixing proportion on each atom of $\hat{G}_n$ is positive in probability. Thus the atom of $\hat{H}_k$ must converge to σ_{0k} in probability. □

The following lemma is a high-dimensional version of result from Serfling (1980, page 253).

Lemma 2 *Let $h(\boldsymbol{Y};\ \boldsymbol{\Sigma})$ be continuous at $\boldsymbol{\Sigma}_0$, uniformly in $\boldsymbol{Y}$. Let F be a distribution function for which $\int \mid h(\boldsymbol{Y};\ \boldsymbol{\Sigma}) \mid dF(\boldsymbol{Y}) < \infty$. Let $\boldsymbol{Y}_1^n = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ be a random sample from F and suppose that $T_n = T_n(\boldsymbol{Y}_1^n)$ is a function of the sample such that $T_n \to \boldsymbol{\Sigma}_0$ in probability. Then, in probability, we have*

$$\frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{Y}_i;\ T_n) \to E_0 h(\boldsymbol{Y};\ \boldsymbol{\Sigma}_0).$$

Proof of Theorem 2. Let $\tilde{G}$ be the maximizer of $\tilde{l}_n(G, \boldsymbol{\Phi})$ among those with $\hat{K}_0 = K_0$ and mixing probabilities $\vartheta_1, \ldots, \vartheta_{K_0}$. We would like to show that, within a $n^{-1/4}$-neighborhood of G_0, any estimate G in (5) with $\hat{K}_0 > K_0$ cannot be a local maximizer of $\tilde{l}_n(G, \boldsymbol{\Phi})$, i.e., $\tilde{l}_n(G, \boldsymbol{\Phi}) < \tilde{l}_n(\tilde{G}, \boldsymbol{\Phi})$ in probability.

From Theorem 1, we know that π_k are grouped and in each group they sum up to $\pi_{0k} + o_p(1)$ and also $\vartheta_k = \pi_{0k} + o_p(1)$, which leads to $\sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K_0} \log \vartheta_k < 0$ in probability. For the SCAD penalty, as shown in Chen and Khalili (2008), we have:

$$\sum_{k=1}^{K-1} p_n(\eta_k) - \sum_{k=1}^{K_0-1} p_n(\tilde{\eta}_k) = \sum_{j,j+1 \in I_k} p_n(\eta_j) \geq \sqrt{n}\gamma_n \sum_{j,j+1 \in I_k} \mid \sigma_{j+1} - \sigma_j \mid .$$

The above two inequalities lead to:

$$\tilde{l}_n(G, \boldsymbol{\Phi}) - \tilde{l}_n(\tilde{G}, \boldsymbol{\Phi}) < \left[l_n(G, \boldsymbol{\Phi}) - l_n(\tilde{G}, \boldsymbol{\Phi}) \right] - \sqrt{n}\gamma_n \sum_{j,j+1 \in I_k} \mid \sigma_{j+1} - \sigma_j \mid .$$

From Taylor expansion, we have:

$$l_n(G, \mathbf{\Phi}) - l_n(\tilde{G}, \mathbf{\Phi}) \leq \sum_{i=1}^{n} \delta_i - \frac{1}{2}\sum_{i=1}^{n} \delta_i^2 + \frac{1}{3}\sum_{i=1}^{n} \delta_i^3,$$

with

$$\begin{aligned}
\delta_i &= \frac{F(\boldsymbol{Y}_i;\ G, \mathbf{\Phi}_i) - F(\boldsymbol{Y}_i;\ \tilde{G}, \mathbf{\Phi}_i)}{F(\boldsymbol{Y};\ \tilde{G}, \mathbf{\Phi})} \\
&= \sum_{k_1=1}^{K_0} \cdots \sum_{k_t=1}^{K_0} \vartheta(k_1, k_2, \ldots, k_t) \frac{F\big(\boldsymbol{Y}_i;\ H(k_1, k_2, \ldots, k_t, \mathbf{\Phi}_i)\big) - F(\boldsymbol{Y}_i;\ \tilde{\mathbf{\Sigma}}_{k_1,k_2,\ldots,k_t}, \mathbf{\Phi}_i)}{F(\boldsymbol{Y}_i;\ \tilde{G}, \mathbf{\Phi}_i)}.
\end{aligned}$$

To avoid double index, denote $k_1, \ldots, k_t$ as $k(1), \ldots, k(t)$. For any $\mathbf{\Sigma}$ in a small neighborhood of $\mathbf{\Sigma}_0$, we have the following expansion:

$$\begin{aligned}
\big\{F(\boldsymbol{Y}_i;\ \mathbf{\Sigma}, \mathbf{\Phi}_i) - F(\boldsymbol{Y}_i;\ \tilde{\mathbf{\Sigma}}_{k(1),k(2),\ldots,k(t)}, \mathbf{\Phi}_i)\big\}F^{-1}(\boldsymbol{Y}_i;\ \tilde{G}, \mathbf{\Phi}_i) &= \sum_{m=1}^{t} (\sigma_{k(m)} - \tilde{\sigma}_{k(m)}) U_{i,\sigma_{k(m)}} \\
+\frac{1}{2} \sum_{\substack{m1,m2 \in \{1,\ldots,t\} \\ n_1+n_2=2}} (\sigma_{k(m1)} - \tilde{\sigma}_{k(m1)})^{n_1} (\sigma_{k(m2)} - \tilde{\sigma}_{k(m2)})^{n_2} U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)}}(\tilde{\mathbf{\Sigma}}, \tilde{G}) \\
+\frac{1}{6} \sum_{\substack{m1,m2,m3 \in \{1,\ldots,t\} \\ n_1+n_2+n_3=3}} (\sigma_{k(m1)} - \tilde{\sigma}_{k(m1)})^{n_1} (\sigma_{k(m2)} - \tilde{\sigma}_{k(m2)})^{n_2} (\sigma_{k(m3)} - \tilde{\sigma}_{k(m3)})^{n_3} \\
U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)},\tilde{\sigma}^{n_3}_{k(m3)}}(\xi_i, \tilde{G}),
\end{aligned}$$

for some ξ_i between $\mathbf{\Sigma}$ and $\tilde{\mathbf{\Sigma}}_{k(1),k(2),\ldots,k(t)}$. Let

$$\begin{aligned}
p(k(1)^{n_1} k(2)^{n_2} \ldots k(t)^{n_t}) = \qquad \qquad (18) \\
\int (\sigma_{k(1)} - \tilde{\sigma}_{k(1)})^{n_1} (\sigma_{k(2)} - \tilde{\sigma}_{k(2)})^{n_2} \cdots (\sigma_{k(t)} - \tilde{\sigma}_{k(t)})^{n_t} dH(k(1), k(2), \ldots, k(t)).
\end{aligned}$$

Thus, we can rewrite (18) as:

$$\begin{aligned}
\sum_{i=1}^{n} \big\{F\Big(\boldsymbol{Y}_i;\ H\big(k(1), k(2), \ldots, k(t)\big), \mathbf{\Phi}_i\Big) - F(\boldsymbol{Y}_i;\ \tilde{\mathbf{\Sigma}}_{k(1),k(2),\ldots,k(t)}, \mathbf{\Phi}_i)\big\}F^{-1}(\boldsymbol{Y}_i;\ \tilde{G}, \mathbf{\Phi}_i) = \\
\frac{1}{2} \sum_{i=1}^{n} \sum_{\substack{m1,m2 \in \{1,\ldots,t\} \\ n_1+n_2=2}} p(k(m1)^{n_1}, k(m2)^{n_2}) U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)}}(\tilde{\mathbf{\Sigma}}, \tilde{G}) \\
+\frac{1}{6} \sum_{i=1}^{n} \sum_{\substack{m1,m2,m3 \in \{1,\ldots,t\} \\ n_1+n_2+n_3=3}} p(k(m1)^{n_1}, k(m2)^{n_2}, k(m3)^{n_3}) \times U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)},\tilde{\sigma}^{n_3}_{k(m3)}}(\xi_i, \tilde{G}).
\end{aligned}$$

Because $E\{\sum_{\substack{m1,m2\in\{1,\ldots,t\}\\ n_1+n_2=2}} U_{i,\sigma^{n_1}_{0k(m1)},\sigma^{n_2}_{0k(m2)}}(\mathbf{\Sigma}_0, G_0)\} = 0$ for any $\mathbf{\Sigma}$,

$$\sum_{i=1}^{n} \sum_{\substack{m1,m2\in\{1,\ldots,t\}\\ n_1+n_2=2}} U_{i,\sigma^{n_1}_{0k(m1)},\sigma^{n_2}_{0k(m2)}}(\mathbf{\Sigma}_0, G_0) = O_p(n^{1/2}),$$

for $\mathbf{\Sigma}$ in a neighborhood of atoms of G_0. Hence from condition B_3, we have

$$\sum_{i=1}^{n} \sum_{\substack{m1,m2\in\{1,\ldots,t\}\\ n_1+n_2=2}} U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)}}(\tilde{\mathbf{\Sigma}}, \tilde{G}) = O_p(n^{3/4}).$$

Also, we can get from B3:

$$n^{-1}\sum_{i=1}^{n} \sum_{\substack{m1,m2,m3\in\{1,\ldots,t\}\\ n_1+n_2+n_3=3}} U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)},\tilde{\sigma}^{n_3}_{k(m3)}}(\xi_i, \tilde{G}) = O_p(1).$$

Then:

$$\begin{aligned}
\frac{1}{6}\sum_{i=1}^{n} \sum_{\substack{m1,m2,m3\in\{1,\ldots,t\}\\ n_1+n_2+n_3=3}} p(k(m1)^{n_1}, k(m2)^{n_2}, k(m3)^{n_3}) U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)},\tilde{\sigma}^{n_3}_{k(m3)}}(\xi_i, \tilde{G}) &= \\
O_p(n) \times \sum_{\substack{m1,m2,m3\in\{1,\ldots,t\}\\ n_1+n_2+n_3=3}} p(k(m1)^{n_1}, k(m2)^{n_2}, k(m3)^{n_3}) &= \\
O_p(n^{3/4}) \times \sum_{\substack{m1,m2\in\{1,\ldots,t\}\\ n_1+n_2=2}} p(k(m1)^{n_1}, k(m2)^{n_2}).&
\end{aligned}$$

It remains to obtain the order of $\sum_{i=1}^{n} \delta_i$. There exists some C_1, such that:

$$\sum_{i=1}^{n} \delta_i \leq C_1 n^{3/4} \sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \sum_{\substack{m1,m2\in\{1,\ldots,t\}\\ n_1+n_2=2}} \vartheta(k(1), k(2), \ldots, k(t)) p(k(m1)^{n_1}, k(m2)^{n_2}).$$

For $\sum_{i=1}^{n} \delta_i^2$, we have:

$$\begin{aligned}
\sum_{i=1}^{n} \delta_i^2 &= \sum_{i=1}^{n} \Bigg(\sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \vartheta(k(1), k(2), \ldots, k(t)) \Bigg(\sum_{m=1}^{t} p(k(m)) U_{i,\tilde{\sigma}_{k(m)}}(\tilde{\mathbf{\Sigma}}, \tilde{G}) \\
&\quad + \frac{1}{2} \sum_{\substack{m1,m2\in\{1,\ldots,t\}\\ n_1+n_2=2}} p_{k(m1)^{n_1},k(m2)^{n_2}} U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)}}(\tilde{\mathbf{\Sigma}}, \tilde{G}) \\
&\quad + \frac{1}{6} \sum_{\substack{m1,m2,m3\in\{1,\ldots,t\}\\ n_1+n_2+n_3=3}} p_{k(m1)^{n_1},k(m2)^{n_2},k(m3)^{n_3}} \times U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)},\tilde{\sigma}^{n_3}_{k(m3)}}(\xi_i, \tilde{G}) \Bigg) \Bigg)^2 \\
&= I + II + III,
\end{aligned}$$

where

$$\begin{aligned} I &= \sum_{i=1}^{n} \Bigg(\sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \vartheta(k(1), k(2), \ldots, k(t)) \Bigg(\sum_{m=1}^{t} p(k(m)) U_{i,\tilde{\sigma}_{k(m)}}(\tilde{\boldsymbol{\Sigma}}, \tilde{G}) \\ &\quad + \frac{1}{2} \sum_{i=1}^{n} \sum_{\substack{m1,m2 \in \{1,\ldots,t\} \\ n_1+n_2=2}} p(k_{m1}^{n_1}, k_{m2}^{n_2}) U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)}}(\tilde{\boldsymbol{\Sigma}}, \tilde{G}) \Bigg) \Bigg)^2, \end{aligned}$$

$$\begin{aligned} II &= \frac{1}{36} \sum_{i=1}^{n} \Big(\sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \vartheta(k(1), k(2), \ldots, k(t)) \sum_{\substack{m1,m2,m3 \in \{1,\ldots,t\} \\ n_1+n_2+n_3=3}} p(k(m1)^{n_1}, k(m2)^{n_2}, k(m3)^{n_3}) \\ &\quad \times U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)},\tilde{\sigma}^{n_3}_{k(m3)}}(\xi_i, \tilde{G}) \Big)^2, \end{aligned}$$

$$\begin{aligned} III &= \frac{1}{3} \sum_{i=1}^{n} \Bigg(\sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \vartheta(k(1), k(2), \ldots, k(t)) \Bigg(\sum_{m=1}^{t} p(k(m)) U_{i,\tilde{\sigma}_{k(m)}}(\tilde{\boldsymbol{\Sigma}}, \tilde{G}) \\ &\quad + \frac{1}{2} \sum_{i=1}^{n} \sum_{\substack{m1,m2 \in \{1,\ldots,t\} \\ n_1+n_2=2}} p(k(m1)^{n_1}, k(m2)^{n_2}) U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)}}(\tilde{\boldsymbol{\Sigma}}, \tilde{G}) \Bigg) \Bigg) \\ &\quad \times \Bigg(\sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \vartheta(k(1), k(2), \ldots, k(t)) \sum_{\substack{m1,m2,m3 \in \{1,\ldots,t\} \\ n_1+n_2+n_3=3}} p(k(m1)^{n_1}, k(m2)^{n_2}, k(m3)^{n_3}) \\ &\quad \times U_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)},\tilde{\sigma}^{n_3}_{k(m3)}}(\xi_i, \tilde{G}) \Bigg). \end{aligned}$$

Because $(\tilde{\boldsymbol{\Sigma}}, \tilde{G}, \boldsymbol{\Phi}) \rightarrow (\boldsymbol{\Sigma}_0, G_0, \boldsymbol{\Phi}_0), \vartheta_m \rightarrow \pi_{0m}$ in probability. From Lemma 2,

$$n^{-1} \sum U^2_{i,\tilde{\sigma}_{k(m)}}(\tilde{\boldsymbol{\Sigma}}, \tilde{G}) \rightarrow E_0\{U^2_{i,\sigma_{0k(m)}}(\boldsymbol{\Sigma}_0, G_0)\},$$

$$n^{-1} \sum U^2_{i,\tilde{\sigma}^{n_1}_{k(m1)},\tilde{\sigma}^{n_2}_{k(m2)}}(\tilde{\boldsymbol{\Sigma}}, \tilde{G}) \rightarrow E_0\{U^2_{i,\sigma^{n_1}_{0k(m1)},\sigma^{n_2}_{0k(m2)}}(\boldsymbol{\Sigma}_0, G_0)\}.$$

Hence, $n^{-1}I$ converges to a quadratic form in $(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2}))$. For some positive constant $C_2 < C_3$, we have:

$$C_2 nR(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2}) \leq (I) \leq C_3 nR(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2})), \tag{19}$$

where

$$R(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2})) = \sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \Big(\sum_{m=1}^{t} p(k(m))^2 + \sum_{\substack{m1,m2 \in \{1,\ldots,t\} \\ n_1+n_2=2}} p(k(m1)^{n_1}, k(m2)^{n_2})^2 \Big).$$

Similarly,

$$II \leq \epsilon n R(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2})). \tag{20}$$

From Cauchy inequality, we have

$$III \leq \epsilon n R(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2})). \tag{21}$$

From (19)-(21),

$$\sum_{i=1}^{n} \delta_i^2 \geq C n R(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2}).$$

For $\sum_{i=1}^{n} \delta^3$, through Taylor's expansion, we have:

$$\begin{aligned} \sum_{i=1}^{n} \delta^3 &= \sum_{i=1}^{n} \bigg(\sum_{k(1)=1}^{K_0} \cdots \sum_{k(t)=1}^{K_0} \vartheta(k(1), k(2), \ldots, k(t)) \Big(\sum_{m=1}^{t} p(k(m)) U_{i, \tilde{\sigma}_{k(m)}}(\tilde{\mathbf{\Sigma}}, \tilde{G}) \\ &\quad + \frac{1}{2} \sum_{\substack{m1,m2 \in \{1,\ldots,t\} \\ n_1+n_2=2}} p(k(m1)^{n_1}, k(m2)^{n_2}) U_{i, \tilde{\sigma}_{k(m1)}^{n_1}, \tilde{\sigma}_{k(m2)}^{n_2}}(\tilde{\mathbf{\Sigma}}, \tilde{G}) \Big) \bigg)^3 \\ &\leq C_4 n \big(\sum_{m=1}^{t} \mid p(k(m)) \mid^3 + \sum_{\substack{m1,\ldots,m6 \in \{1,\ldots,t\} \\ n_1+\cdots n_6=6}} p(k(m1)^{n_1}, \ldots, k(m6)^{n_6}) \big) \\ &\leq \epsilon n R(p(k(m)), p(k(m1)^{n_1}, k(m2)^{n_2})), \end{aligned}$$

which proves that $\sum_{i=1}^{n} \delta^2$ dominates $\sum_{i=1}^{n} \delta^3$ in probability.

In conclusion, we have, for some constant C:

$$l_n(G) - l_n(\tilde{G}) \leq C n^{3/4} \sum_{k=1}^{K_0} \sum_{i,j \in I_k} (\sigma_i - \sigma_j)^2 \leq C n^{1/2} \sum_{k=1}^{K_0} \sum_{j,j+1 \in I_k} \mid \sigma_{j+1} - \sigma_j \mid$$

in probability. We get:

$$\tilde{l}_n(G, \mathbf{\Phi}) - \tilde{l}(\tilde{G}, \mathbf{\Phi}) = C n^{1/2} \sum_{k=1}^{K_0} \sum_{j,j+1 \in I_k} \mid \sigma_{j+1} - \sigma_j \mid - n^{1/2} \gamma_n \sum_{k=1}^{K_0} \sum_{j,j+1 \in I_k} \mid \sigma_{j+1} - \sigma_j \mid \tag{22}$$

in probability. As $\gamma_n = O_p(n^{1/3} \log n) \to \infty$, (22) is negative for large n. It is a contradiction to the assumption that G with $\hat{K}_0 > K_0$ is an MPLE. This completes the proof. $\square$

REFERENCES

[1] Abrahamsen, P., *A review of Gaussian random fields and correlation functions.* Norsk Regnesentral/Norwegian Computing Center, 1997.

[2] Akaike, H., "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[3] Albert, P. S., McFarland, H. F., Smith, M. E., and Frank, J. A., "Time series for modelling counts from a relapsing-remitting disease: Application to modelling disease activity in multiple sclerosis," *Statistics in Medicine*, vol. 13, no. 5-7, pp. 453–466, 1994.

[4] Auton, M., Sowa, K. E., Smith, S. M., Sedlák, E., Vijayan, K. V., and Cruz, M. A., "Destabilization of the a1 domain in von willebrand factor dissociates the a1a2a3 tri-domain and provokes spontaneous binding to glycoprotein ibα and platelet activation under shear stress," *Journal of Biological Chemistry*, vol. 285, no. 30, pp. 22831–22839, 2010.

[5] Baum, L. E., Petrie, T., Soules, G., and Weiss, N., "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[6] Berndt, M. C., Du, X., and Booth, W. J., "Ristocetin-dependent reconstitution of binding of von willebrand factor to purified human platelet membrane glycoprotein ib-ix complex," *Biochemistry*, vol. 27, no. 2, pp. 633–640, 1988.

[7] Bhattacharya, P., "Some aspects of change-point analysis," *Lecture Notes-Monograph Series*, pp. 28–56, 1994.

[8] Bickel, P. J., Ritov, Y., and Ryden, T., "Asymptotic normality of the maximum-likelihood estimator for general hidden markov models," *The Annals of Statistics*, vol. 26, no. 4, pp. 1614–1635, 1998.

[9] Cappé, O., Moulines, E., and Rydén, T., *Inference in hidden Markov models.* Springer Science+ Business Media, 2005.

[10] Casella, G. and George, E. I., "Explaining the gibbs sampler," *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.

[11] Celeux, G. and Durand, J.-B., "Selecting hidden markov model state number with cross-validated likelihood," *Computational Statistics*, vol. 23, no. 4, pp. 541–564, 2008.

[12] Chambaz, A., Garivier, A., and Gassiat, E., “A minimum description length approach to hidden markov models with poisson and gaussian emissions. application to order identification,” *Journal of Statistical Planning and Inference*, vol. 139, no. 3, pp. 962–977, 2009.

[13] Chen, J. and Kalbfleisch, J., “Penalized minimum-distance estimates in finite mixture models,” *Canadian Journal of Statistics*, vol. 24, no. 2, pp. 167–175, 1996.

[14] Chen, J. and Khalili, A., “Order selection in finite mixture models with a nonsmooth penalty,” *Journal of the American Statistical Association*, vol. 103, no. 484, 2008.

[15] Chen, W., Evans, E. A., McEver, R. P., and Zhu, C., “Monitoring receptor-ligand interactions between surfaces by thermal fluctuations,” *Biophysical journal*, vol. 94, no. 2, pp. 694–701, 2008.

[16] Chen, W., Lou, J., Evans, E. A., and Zhu, C., “Observing force-regulated conformational changes and ligand dissociation from a single integrin on cells,” *The Journal of cell biology*, vol. 199, no. 3, pp. 497–512, 2012.

[17] Chen, W., Lou, J., and Zhu, C., “Forcing switch from short-to intermediate- and long-lived states of the αa domain generates lfa-1/icam-1 catch bonds,” *Journal of Biological Chemistry*, vol. 285, no. 46, pp. 35967–35978, 2010.

[18] Chen, Z., Gu, C., and Wahba, G., “Comment on ”linear smoothers and additive models” by a. buja, t. hastie, and r. tibshirani,” *The Annals of Statistics*, vol. 17, no. 3, pp. 515–521, 1989.

[19] Clairambault, J., Curzi-Dascalova, L., Kauffmann, F., Médigue, C., and Leffler, C., “Heart rate variability in normal sleeping full-term and preterm neonates,” *Early human development*, vol. 28, no. 2, pp. 169–183, 1992.

[20] Cressie, N., “Statistics for spatial data,” *Terra Nova*, vol. 4, no. 5, pp. 613–617, 1992.

[21] Csiszár, I. and Shields, P. C., “The consistency of the bic markov order estimator,” *The Annals of Statistics*, vol. 28, no. 6, pp. 1601–1619, 2000.

[22] Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D., “A bayesian approach to the design and analysis of computer experiments,” *ORNL-6498*, 1988.

[23] Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D., “Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments,” *Journal of the American Statistical Association*, vol. 86, no. 416, pp. 953–963, 1991.

[24] Dempster, A. P., Laird, N. M., and Rubin, D. B., “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[25] Diggle, P. J., Tawn, J., and Moyeed, R., “Model-based geostatistics,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 47, no. 3, pp. 299–350, 1998.

[26] DiMatteo, I., Genovese, C. R., and Kass, R. E., “Bayesian curve-fitting with free-knot splines,” *Biometrika*, vol. 88, no. 4, pp. 1055–1071, 2001.

[27] Dustin, M. L., Bromley, S. K., Davis, M. M., and Zhu, C., “Identification of self through two-dimensional chemistry and synapses,” *Annual review of cell and developmental biology*, vol. 17, no. 1, pp. 133–157, 2001.

[28] Fan, J. and Li, R., “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[29] Fang, K.-T., Li, R., and Sudjianto, A., *Design and modeling for computer experiments*. Chapman and Hall/CRC, 2010.

[30] Gassiat, E. and Boucheron, S., “Optimal error exponents in hidden markov models order estimation,” *Information Theory, IEEE Transactions on*, vol. 49, no. 4, pp. 964–980, 2003.

[31] Gassiat, E. and Keribin, C., “The likelihood ratio test for the number of components in a mixture with markov regime,” *ESAIM P & S*, vol. 4, pp. 25–52, 2000.

[32] Geisser, S., *Predictive inference: an introduction*, vol. 55. CRC Press, 1993.

[33] Gelfand, A. E. and Smith, A. F., “Sampling-based approaches to calculating marginal densities,” *Journal of the American statistical association*, vol. 85, no. 410, pp. 398–409, 1990.

[34] Gill, J., *Bayesian methods: A social and behavioral sciences approach.* CRC press, 2002.

[35] Giudici, P., Ryden, T., and Vandekerkhove, P., “Likelihood-ratio tests for hidden markov models,” *Biometrics*, vol. 56, no. 3, pp. 742–747, 2000.

[36] Golub, G. H. and Van Loan, C. F., *Matrix computations*, vol. 3. JHUP, 2012.

[37] Gramacy, R. B. and Lee, H. K., “Cases for the nugget in modeling computer experiments,” *Statistics and Computing*, vol. 22, no. 3, pp. 713–722, 2012.

[38] Hastings, W. K., “Monte carlo sampling methods using markov chains and their applications,” *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[39] Hawkins, D. M. and Zamba, K., "A change-point model for a shift in variance," *Journal of Quality Technology*, vol. 37, no. 1, pp. 21–31, 2005.

[40] Huang, J., Zarnitsyna, V. I., Liu, B., Edwards, L. J., Jiang, N., Evavold, B. D., and Zhu, C., "The kinetics of two-dimensional tcr and pmhc interactions determine t-cell responsiveness," *Nature*, vol. 464, no. 7290, pp. 932–936, 2010.

[41] Hughes, J. P. and Guttorp, P., "A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena," *Water Resources Research*, vol. 30, no. 5, pp. 1535–1546, 1994.

[42] Hughes, J. P. and Guttorp, P., "Incorporating spatial dependence and atmospheric data in a model of precipitation," *Journal of applied meteorology*, vol. 33, no. 12, pp. 1503–1515, 1994.

[43] Hung, Y., Zarnitsyna, V., Zhang, Y., Zhu, C., and Wu, C. J., "Binary time series modeling with application to adhesion frequency experiments," *Journal of the American Statistical Association*, vol. 103, no. 483, 2008.

[44] Hunter, D. R. and Li, R., "Variable selection using mm algorithms," *Annals of statistics*, vol. 33, no. 4, p. 1617, 2005.

[45] Joseph, V. R., "Limit kriging," *Technometrics*, vol. 48, no. 4, pp. 458–466, 2006.

[46] Kaleh, G. K. and Vallet, R., "Joint parameter estimation and symbol detection for linear or nonlinear unknown channels," *Communications, IEEE Transactions on*, vol. 42, no. 7, pp. 2406–2413, 1994.

[47] Kohavi, R. and others, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International joint Conference on artificial intelligence*, vol. 14, pp. 1137–1145, Lawrence Erlbaum Associates Ltd, 1995.

[48] Koshi, J. M. and Goldstein, R. A., "Analyzing site heterogeneity during protein evolution," in *Pac. Symp. Biocomput*, vol. 6, pp. 191–202, 2001.

[49] Koski, T., *Hidden Markov models for bioinformatics*, vol. 2. Kluwer Academic Pub, 2001.

[50] Krishnaiah, P. and Miao, B., "Review about estimation of change points," *Handbook of Statistics*, vol. 7, pp. 375–402, 1988.

[51] Laslett, G. M., "Kriging and splines: an empirical comparison of their predictive performance in some applications," *Journal of the American Statistical Association*, vol. 89, no. 426, pp. 391–400, 1994.

[52] Leroux, B. G., “Maximum-likelihood estimation for hidden markov models,” *Stochastic processes and their applications*, vol. 40, no. 1, pp. 127–143, 1992.

[53] Lister, I., Schmitz, S., Walker, M., Trinick, J., Buss, F., Veigel, C., and Kendrick-Jones, J., “A monomeric myosin vi with a large working stroke,” *The EMBO journal*, vol. 23, no. 8, pp. 1729–1738, 2004.

[54] MacDonald, I. L. and Zucchini, W., *Hidden Markov and other models for discrete-valued time series*, vol. 70. Chapman & Hall/CRC, 1997.

[55] MacKAY, R. J., “Estimating the order of a hidden markov model,” *Canadian Journal of Statistics*, vol. 30, no. 4, pp. 573–589, 2002.

[56] MacKay Altman, R., “Assessing the goodness-of-fit of hidden markov models,” *Biometrics*, vol. 60, no. 2, pp. 444–450, 2004.

[57] Mardia, K. V. and Marshall, R., “Maximum likelihood estimation of models for residual covariance in spatial regression,” *Biometrika*, vol. 71, no. 1, pp. 135–146, 1984.

[58] Marioni, J., Thorne, N., and Tavare, S., “Biohmm: a heterogeneous hidden markov model for segmenting array cgh data,” *Bioinformatics*, vol. 22, no. 9, pp. 1144–1146, 2006.

[59] Marshall, B. T., Sarangapani, K. K., Wu, J., Lawrence, M. B., McEver, R. P., and Zhu, C., “Measuring molecular elasticity by atomic force microscope cantilever fluctuations,” *Biophysical journal*, vol. 90, no. 2, pp. 681–692, 2006.

[60] Matérn, B. and others, “Spatial variation. stochastic models and their application to some problems in forest surveys and other sampling investigations.,” *Meddelanden fran statens Skogsforskningsinstitut*, vol. 49, no. 5, 1960.

[61] Matheron, G., “Principles of geostatistics,” *Economic geology*, vol. 58, no. 8, pp. 1246–1266, 1963.

[62] Matsushita, T. and Sadler, J. E., “Identification of amino acid residues essential for von willebrand factor binding to platelet glycoprotein ib. charged-to-alanine scanning mutagenesis of the a1 domain of human von willebrand factor,” *Journal of Biological Chemistry*, vol. 270, no. 22, pp. 13406–13414, 1995.

[63] Mazumder, R., Friedman, J. H., and Hastie, T., “Sparsenet: Coordinate descent with nonconvex penalties,” *Journal of the American Statistical Association*, vol. 106, no. 495, 2011.

[64] Mehta, A., Finer, J., and Spudich, J., “Detection of single-molecule interactions using correlated thermal diffusion,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 15, pp. 7927–7931, 1997.

[65] Meng, X.-L. and Rubin, D. B., "Maximum likelihood estimation via the ecm algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[66] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, p. 1087, 1953.

[67] Molloy, J., Burns, J., Kendrick-Jones, J., Tregear, R., and White, D., "Movement and force produced by a single myosin head," *Nature*, vol. 378, no. 6553, pp. 209–212, 1995.

[68] Morales, L., Martin, C., and Cruz, M., "The interaction of von willebrand factor-a1 domain with collagen: mutation g1324s (type 2m von willebrand disease) impairs the conformational change in a1 domain induced by collagen," *Journal of Thrombosis and Haemostasis*, vol. 4, no. 2, pp. 417–425, 2006.

[69] Morris, M. D., Mitchell, T. J., and Ylvisaker, D., "Bayesian design and analysis of computer experiments: use of derivatives in surface prediction," *Technometrics*, vol. 35, no. 3, pp. 243–255, 1993.

[70] O'Hagan, A. and Kingman, J., "Curve fitting and optimal design for prediction," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–42, 1978.

[71] Patterson, H. D. and Thompson, R., "Recovery of inter-block information when block sizes are unequal," *Biometrika*, vol. 58, no. 3, pp. 545–554, 1971.

[72] Peng, C.-Y. and Wu, C. F. J., "On the choice of nugget in kriging modeling for deterministic computer experiments," *Journal of Computational and Graphical Statistics*, vol. 00, no. 0, pp. 1–18, 2013.

[73] Qin, S., Pang, X., and Zhou, H.-X., "Automated prediction of protein association rate constants," *Structure*, vol. 19, no. 12, pp. 1744–1751, 2011.

[74] Rabiner, L. R., "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[75] Rabinowitz, I., Tuley, E. A., Mancuso, D. J., Randi, A. M., Firkin, B. G., Howard, M. A., and Sadler, J. E., "von willebrand disease type b: a missense mutation selectively abolishes ristocetin-induced von willebrand factor binding to platelet glycoprotein ib," *Proceedings of the National Academy of Sciences*, vol. 89, no. 20, pp. 9846–9849, 1992.

[76] Robert, C. P. and Casella, G., *Monte Carlo statistical methods*, vol. 319. Citeseer, 2004.

[77] Robert, C. P., Ryden, T., and Titterington, D. M., "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)*, vol. 62, no. 1, pp. 57–75, 2000.

[78] Ruggeri, Z. M., "Vonwillebrand factor: Looking back and looking forward," *Thromb Haemost*, vol. 98, pp. 55–62, 2007.

[79] Ruggeri, Z. M. and Mendolicchio, G. L., "Adhesion mechanisms in platelet function," *Circulation research*, vol. 100, no. 12, pp. 1673–1685, 2007.

[80] Rydén, T., Teräsvirta, T., and Åsbrink, S., "Stylized facts of daily return series and the hidden markov model," *Journal of applied econometrics*, vol. 13, no. 3, pp. 217–244, 1998.

[81] Sacks, J. and Schiller, S., "Spatial designs," *Statistical decision theory and related topics IV*, vol. 2, no. S S, pp. 385–399, 1988.

[82] Sacks, J., Schiller, S. B., and Welch, W. J., "Designs for computer experiments," *Technometrics*, vol. 31, no. 1, pp. 41–47, 1989.

[83] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., "Design and analysis of computer experiments," *Statistical science*, vol. 4, no. 4, pp. 409–423, 1989.

[84] Santner, T. J., Williams, B. J., and Notz, W. I., *The design and analysis of computer experiments.* Springer Verlag, 2003.

[85] Sarangapani, K. K., Marshall, B. T., McEver, R. P., and Zhu, C., "Molecular stiffness of selectins," *Journal of Biological Chemistry*, vol. 286, no. 11, pp. 9567–9576, 2011.

[86] Schwarz, G., "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[87] Scott, S. L., James, G. M., and Sugar, C. A., "Hidden markov models for longitudinal comparisons," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 359–369, 2005.

[88] Seifert, M., *Hidden Markov Models with Applications in Computational Biology.* SVH-Verlag, 2013.

[89] Seifert, M., Strickert, M., Schliep, A., and Grosse, I., "Exploiting prior knowledge and gene distances in the analysis of tumor expression profiles with extended hidden markov models," *Bioinformatics*, vol. 27, no. 12, pp. 1645–1652, 2011.

[90] Serfling, R. J., *Approximation theorems of mathematical statistics*, vol. 162. Wiley-Interscience, 2009.

[91] Stone, M., "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 111–147, 1974.

[92] Sun, G., Zhang, Y., Huo, B., and Long, M., "Surface-bound selectin–ligand binding is regulated by carrier diffusion," *European Biophysics Journal*, vol. 38, no. 5, pp. 701–711, 2009.

[93] Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[94] Veigel, C., Coluccio, L. M., Jontes, J. D., Sparrow, J. C., Milligan, R. A., and Molloy, J. E., "The motor protein myosin-i produces its working stroke in two steps," *Nature*, vol. 398, no. 6727, pp. 530–533, 1999.

[95] Wahba, G., *Spline models for observational data*, vol. 59. Society for industrial and applied mathematics, 1990.

[96] Wald, A., "Note on the consistency of the maximum likelihood estimate," *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 595–601, 1949.

[97] Wang, P. and Puterman, M. L., "Markov poisson regression models for discrete time series.," *Journal of Applied Statistics*, vol. 26, no. 7, pp. 855–869, 1999.

[98] Wang, X., "Bayesian free-knot monotone cubic spline regression," *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, 2008.

[99] Wecker, W. E. and Ansley, C. F., "The signal extraction approach to nonlinear regression and spline smoothing," *Journal of the American Statistical Association*, vol. 78, no. 381, pp. 81–89, 1983.

[100] Welch, R. L., "Hidden markov model and the baum-welch algorithm," *Proceedings of the IEEE*, 2003.

[101] Wu, C., "On the convergence properties of the em algorithm," *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

[102] Wu, J., Fang, Y., Yang, D., and Zhu, C., "Thermo-mechanical responses of a surface-coupled afm cantilever.," *Journal of biomechanical engineering*, vol. 127, no. 7, p. 1208, 2005.

[103] Yago, T., Wu, J., Wey, C. D., Klopocki, A. G., Zhu, C., and McEver, R. P., "Catch bonds govern adhesion through l-selectin at threshold shear," *The Journal of cell biology*, vol. 166, no. 6, pp. 913–923, 2004.

[104] Ylvisaker, D., "Designs on random fields," *A Survey of Statistical Design and Linear Models*, pp. 593–607, 1975.

[105] Ylvisaker, D., “Prediction and design,” *The Annals of Statistics*, pp. 1–19, 1987.

[106] Yuan, M. and Kendziorski, C., “Hidden markov models for microarray time course data in multiple biological conditions,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1323–1332, 2006.

[107] Zarnitsyna, V. I., Huang, J., Zhang, F., Chien, Y.-H., Leckband, D., and Zhu, C., “From the cover: Memory in receptor-ligand-mediated cell adhesion,” *Science Signaling*, vol. 104, no. 46, p. 18037, 2007.

[108] Zhang, K., Yang, Y., Devanarayan, V., Xie, L., Deng, Y., and Donald, S., “A hidden markov model-based algorithm for identifying tumour subtype using array cgh data,” *BMC genomics*, vol. 12, no. Suppl 5, p. S10, 2011.

[109] Zou, H. and Li, R., “One-step sparse estimates in nonconcave penalized likelihood models,” *Annals of statistics*, vol. 36, no. 4, p. 1509, 2008.