NEW MODELS AND METHODS FOR APPLIED STATISTICS: TOPICS IN COMPUTER EXPERIMENTS AND TIME SERIES ANALYSIS

by

YIBO ZHAO

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Ying Hung

and approved by

New Brunswick, New Jersey

OCTOBER, 2017

ABSTRACT OF THE DISSERTATION

New Models and Methods for Applied Statistics: Topics in Computer Experiments and Time Series Analysis

By YIBO ZHAO

Dissertation Director:

Ying Hung

In applied statistics, people develop models to solve real world problems based on data. However, the data is growing fast and become more and more massive and complex. Conventional models are limited in the capability of dealing with the fast growing data. This dissertation develops two new models in computer experiments and time series analysis. The new models are developed based on the special features of two real-world problems. The two datasets are from an IBM data thermal study and a biological cell adhesion experiment.

For computer experiment, we address two important issues in Gaussian process (GP) modeling. One is how to reduce the computational complexity in GP modeling and the other is how to simultaneous perform variable selection and estimation for the mean function of GP models. Estimation is computationally intensive for GP models because it heavily involves manipulations of an n-by-n correlation matrix, where n is the sample size. Conventional penalized likelihood approaches are widely used for variable selection. However, the computational cost of the penalized likelihood estimation (PMLE) or the corresponding one-step sparse estimation (OSE) can be prohibitively high as the sample size becomes large, especially for GP models. To address both issues, this article proposes an efficient subsample aggregating (subagging) approach with an experimental design-based subsampling scheme. The proposed method is computationally cheaper, yet it can be shown that the resulting subagging estimators achieve the same efficiency as the original PMLE and OSE asymptotically. The finite-sample performance is examined through simulation studies. Application of the proposed methodology to a data center thermal study reveals some interesting information, including identifying an efficient cooling mechanism.

Motivated by an analysis of cell adhesion experiments, we introduce a new statistical framework within which the unique features are incorporated and the molecular binding mechanism can be studied. This framework is based upon an extension of Markov switching autoregressive (MSAR) models, a regime-switching type of time series model generalized from hidden Markov models. Standard MSAR models are developed for the analysis of individual stochastic process. To handle multiple time series processes, we introduce Markov switching autoregressive mixed (MSARM) model that simultaneously models multiple time series processes collected from different experimental subjects as in the longitudinal data setting. More than a simple extension, the MSARM model posts statistical challenges in the theoretical developments as well as computational efficiency in high-dimensional integration.

Acknowledgements

I went through my Ph.D. studies with many help and support. Those five-years memories at Rutgers, are the most precious treasure in my life.

First I would like to express my deepest appreciation to my advisor Prof. Ying Hung. She always shows great support and encouragement. Without her help, I could not imagine completing this dissertation. I feel very lucky to be her student and I appreciate for those time working with her, influenced by her continuous pursuit on advances in statistical research, her willingness to help and her positive attitude to work. Those precious time not only benefit me in my Ph.D. study but also stay with me and be appreciated throughout my life.

My thanks also go to the Department of Statistics and Biostatistics of Rutgers University for providing me support and a great learning and research environment. Many thanks to Prof. Regina Liu and Prof. Kolassa, for your kind support and advice. Many thanks to Arlene Gray, Lisa Curtin, Isabelle Amarhanow and Marcy Collins.

I would like to thank my parents for their unconditional love and to thank my wife for her accompaniment, on the incredible adventure that we have shared for those years.

That's it, and say goodbye to my days at Rutgers. Welcome new adventure!

Dedication

Every challenging work needs self-efforts as well as guidance of elders especially those who were very close to our heart. My humble effort I dedicate to my sweet and loving

Father & Mother

And to my sweet and brilliant

Wife

Whose affection, love, encouragement and prayers of day and night make me able to get such success and honor,

Along with all hard working and respected

Teachers

Table of Contents

List of Tables

List of Figures

Chapter 1

Efficient Gaussian Process Modeling Using Experimental Design-based Subagging

1.1 Introduction

Gaussian process (GP) models, also known as kriging models, are widely used in many fields, including geostatistics (Cressie (1993), Stein (1999)), machine learning (Smola and Bartlett (2001), Snelson and Ghahramani (2006)), and computer experiment modeling (Santner et al. (2003), Fang et al. (2006)). In this article, we focus on two issues in GP modeling. One is the study of simultaneous variable selection and estimation of GP models for the mean function, in particular, and the other is how to alleviate the computational complexity in GP modeling.

Various examples of variable selection in GP models can be found in the literature, such as in geostatistics (Hoeting et al. (2006), Huang and Chen (2007), Chu et al. (2011)) and computer experiments (Welch et al. (1992), Linkletter et al. (2006), Joseph et al. (2008), Kaufman et al. (2011)). In this article, we mainly focus on identifying active effects through the mean function. Several empirical studies report that, by a proper selection of important variables in the mean function, the prediction accuracy of GP models can be significantly improved, especially when there are some strong trends (Joseph et al. (2008), Hung (2011), Kaufman et al. (2011)). Compared with nonlinear effects identified from the covariance function (Linkletter et al. (2006)), linear effects are relatively easy to interpret, and of scientific interest in many applications. Conventional approaches based on penalized likelihood functions, such as the penalized likelihood estimators (PMLEs) and the corresponding one-step sparse estimators (OSEs), are conceptually attractive, but computationally difficult in practice, especially with massive data observed on irregular grid. This is because estimation and variable selection heavily involve manipulations of an $n \times n$ correlation matrix that require $O(n^3)$ computations, where n is the sample size. The calculation is computationally intensive and often intractable for massive data.

The computational issue is well recognized in the literature and various methods are proposed, either changing the model to one that is computationally convenient or approximating the likelihood for the original data. Examples of the former includes Rue and Tjelmeland (2002), Rue and Held (2005), Cressie and Johannesson (2008), Banerjee et al. (2008), Gramacy and Lee (2008), and Wikle (2010); approximation approaches includes Nychka (2000), Smola and Bartlett (2001), Nychka et al. (2002), Stein et al. (2004), Furrer et al. (2006), Snelson and Ghahramani (2006), Fuentes (2007), Kaufman et al. (2008), and Gramacy and Apley (2015). However, these methods focus mainly on estimation and prediction, not variable selection, and most of them are developed for datasets collected from a regular grid under a low-dimensional setting. Recent studies address the issues by imposing a sparsity constraint on the correlation matrix, including covariance tapering and compactly supported correlation functions (Kaufman et al. (2008, 2011), Chu et al. (2011), Nychka et al. (2015)). However, it has been shown that this does not work well for purposes of parameter estimation (Stein (2013), Liang et al. (2013)), which is crucial in selecting important variables. In addition, the connection between the degree of sparsity and computation time is nontrivial.

In this paper, we provide an alternative framework that alleviates the computational difficulties in estimation and variable selection by utilizing the idea of subsample aggregating, also known as subagging (Büchlmann and Yu (2002)). This framework includes a subagging estimator and a new subsampling scheme based on a special class of experimental designs called Latin hypercube designs (LHDs), that have a one-dimensional projection property. By borrowing the inherited one-dimensional projection property of LHDs and a block structure, the new subsampling scheme not only provides an efficient data reduction but also takes into account the spatial dependency in GP models. The computational complexity of the proposed subagging estimation is dramatically reduced, yet the subagging estimators achieve the same efficiency as the original PMLE and OSE, asymptotically.

The remainder of the paper is organized as follows. In Section 2, the conventional penalized likelihood approach is discussed. The new variable selection framework, including the new subsampling scheme and the subagging estimators are introduced in Section 3. Theoretical properties are derived in Section 4. In Section 5, finite-sample performance of the proposed framework is investigated in simulation studies. A data center example is illustrated in Section 6. Discussions are given in Section 7.

1.2 Variable selection in Gaussian process models

For a domain of interest Γ in R^d, we consider a Gaussian process $\{Y(\boldsymbol{x}) : \boldsymbol{x} \in R^d\}$ such that

$$Y(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{1.1}$$

where $\boldsymbol{\beta}$ is a vector of unknown mean function coefficients and $Z(\boldsymbol{x})$ is a stationary Gaussian process with mean 0 and covariance function $\sigma^2\psi$. The covariance function is $cov\{Y(\boldsymbol{x}+\boldsymbol{h}), Y(\boldsymbol{x})\} = \sigma^2\psi(\boldsymbol{h};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of correlation parameters for the correlation function $\psi(\boldsymbol{h};\boldsymbol{\theta})$, and $\psi(\boldsymbol{h};\boldsymbol{\theta})$ is a positive semidefinite function with $\psi(\boldsymbol{0};\boldsymbol{\theta}) = 1$ and $\psi(\boldsymbol{h};\boldsymbol{\theta}) = \psi(-\boldsymbol{h};\boldsymbol{\theta})$.

Suppose n observations are collected, denoted by

$$\mathscr{D}_n = \{\big(\boldsymbol{x}_{t_1}, y(\boldsymbol{x}_{t_1})\big), \ldots, \big(\boldsymbol{x}_{t_n}, y(\boldsymbol{x}_{t_n})\big)\} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}.$$

Let $\boldsymbol{y}_n = (y_1, \ldots, y_n)^T$, $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{\phi} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \sigma^2)^T$ be the vector of all the parameters, and Θ be the parameter space. Based on (1.1), the likelihood function can be written as

$$f(\boldsymbol{y}_n, \boldsymbol{X}_n; \boldsymbol{\phi}) = \frac{|R_n(\boldsymbol{\theta})|^{-1/2}}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})\},$$

where $R_n(\boldsymbol{\theta})$ is an $n \times n$ correlation matrix with elements $\psi(\boldsymbol{x}_i - \boldsymbol{x}_j;\boldsymbol{\theta})]$. Thus the log-likehood function, ignoring a constant, is

$$\begin{aligned} \ell(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\phi}) &= -\frac{1}{2\sigma^2}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}) \\ &\quad -\frac{1}{2}|R_n(\boldsymbol{\theta})| - \frac{n}{2}\log(\sigma^2), \end{aligned} \tag{1.2}$$

where $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and σ are the unknown parameters.

To achieve simultaneous variable selection and parameter estimation, we focus on penalized likelihood approaches, which are increasingly popular in recent years. A penalized log-likelihood function for GP models can be written as

$$\ell_p(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\phi}) = \ell(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\phi}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1.3}$$

where $p_\lambda(\cdot)$ is a pre-specified penalty function with a tuning parameter λ. There are various choices of penalty functions such as LASSO (Donoho and Johnstone (1994), Tibshirani (1996)), the adaptive LASSO (Zou (2006)), and the minimax concave penalty

(Zhang (2010)). In this article, we focus on the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)) defined by

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & if|\beta| > \lambda, \\ \lambda^2 + (a-1)^{-1}(a\lambda|\beta| - \beta^2/2 - a\lambda^2 + \lambda^2/2) & if\lambda < |\beta| \le a\lambda, \\ (a+1)\lambda^2/2 & if|\beta| > a\lambda, \end{cases}$$

for some $a > 2$. By maximizing (1.3), the penalized maximum likelihood estimators (PLMEs) of $\boldsymbol{\phi}$ can be obtained as $\hat{\boldsymbol{\phi}}_n = \arg\max_{\boldsymbol{\phi}} \ell_p(\boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{\phi})$.

To compute PMLEs under the SCAD penalty, Zou and Li (2008) develop a unified algorithm to improve computational efficiency by locally linear approximation (LLA) of the penalty function. They propose an one-step LLA estimation that approximates the solution after just one iteration in a Newton-Raphson-type algorithm starting at the maximum likelihood estimates (MLEs). Chu et al. (2011) extend the one-step LLA estimation to approximate the PMLEs for the spatial linear models and the resulting estimate is called the one-step sparse estimate (OSE).

Following the idea of Chu et al. (2011), the OSE of $\boldsymbol{\beta}$ in GP models, denoted by $\hat{\boldsymbol{\beta}}_{OSE}$, is obtained by maximizing

$$\begin{aligned} Q(\boldsymbol{\beta}) &= -\frac{1}{2\hat{\sigma^2}^{(0)}}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T R_n^{-1}(\hat{\boldsymbol{\theta}}^{(0)})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}) \\ &\quad -n\sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j|, \end{aligned} \tag{1.4}$$

where $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\boldsymbol{\theta}}^{(0)}$ and $\hat{\sigma^2}^{(0)}$ are the MLEs obtained from (1.2). We also update $\boldsymbol{\theta}$ and σ^2 by maximizing (1.4) evaluated at $\hat{\boldsymbol{\beta}}_{OSE}$ with respect to $\boldsymbol{\theta}$ and σ^2. The resulting OSE of $\boldsymbol{\theta}$ and σ^2 is denoted by $\hat{\boldsymbol{\theta}}_{OSE}$ and $\hat{\sigma}^2_{OSE}$. We fix the tuning parameter $a = 3.7$ as recommended by Fan and Li (2001). To determine λ, a Bayesian information criterion(BIC) proposed by Chu et al. (2011) is incorporated.

The implementation of the penalized likelihood approach, including the calculation of PMLEs and OSEs is computationally demanding; it relies heavily on the calculation of $R_n^{-1}(\boldsymbol{\theta})$ and $|R_n(\boldsymbol{\theta})|$, computationally intensive and often intractable due to numerical issues. It is particularly difficult for massive data collected on irregular grids, because no Kronecker product techniques can be utilized for computational simplification (Bayarri et al. (2007, 2009), Rougier (2008)). A similar issue has also been recognized in calculating the MLEs in GP models.

1.3 Variable selection for GP via subagging

1.3.1 A new block bootstrap subsampling scheme

Subagging, modified based upon bagging (bootstrap aggregating), is one of the most effective and computationally efficient procedures to improve on unstable estimators (Efron and Tibshirani (1993), Breiman (1996), Büchlmann and Yu (2002)). Although originally proposed to reduce variance in estimations and predictions, the idea of subsampling is attractive in many applications to achieve computational reduction. It is particularly appealing to GP modeling because of its high computational demand in estimating PMLEs and OSEs. However, direct application of subagging with random bootstrap subsamples is not efficient in estimation and variable selection of GP because the data are assumed to be dependent. This is not surprising because similar issues occur in the conventional bootstrap when the data are dependent, such as in time series and spatial data, and various block bootstrap techniques are introduced (Künsch (1989), Liu and Singh (1992), Lahiri (1995, 1999, 2003), Politis and Romano (1994)). Therefore, as an analogous result to the conventional block bootstrap, a new subsample scheme for dependent data based on blocks is called for.

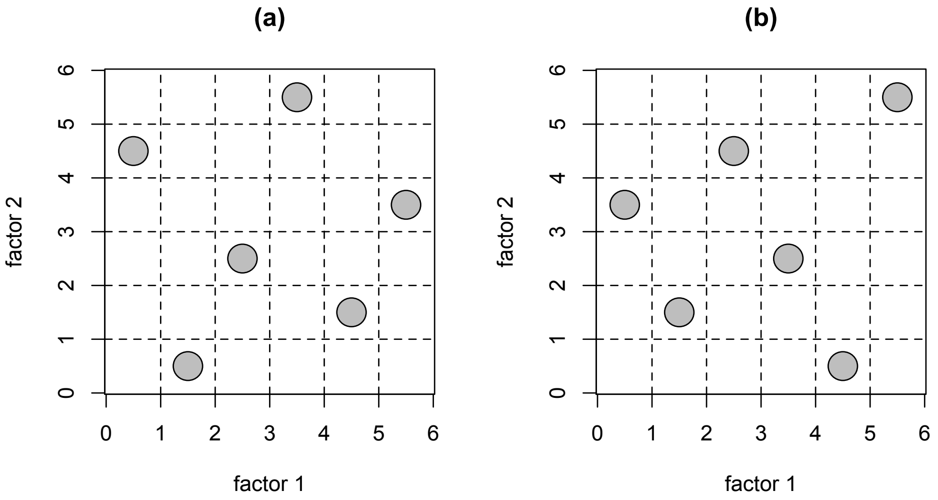

Figure 1.1: Two examples of LHDs

We introduce a block bootstrap subsampling method based on Latin hypercube designs (LHDs). It is called LHD-based block bootstrap. LHD is a class of experimental designs such that the projection of an LHD onto any dimension has exactly one observation for each level and therefore the resulting design can spread out more uniformly over the space. An m-run LHD in a d-dimensional space, denoted by LHD(m, d) can be easily constructed by permuting $(0, 1, ..., m - 1)$ for each dimension. Given the sample size, there are $(m!)^{d-1}$ LHDs. Two randomly generated LHD(6,2) are illustrated in Figure 1.1. It is clear that the projection onto either dimension has exactly one observation for each level. After decomposing the complete data into disjoint equally-spaced hypercubes/blocks, a LHD-based block bootstrap subsample can be obtained by collecting blocks according to the structure of a randomly generated LHD. One example of a LHD-based block bootstrap subsample using the LHD in Figure 1.1(a) is given in Figure 1.2, where the circles are the observations, gray areas are the LHD-based blocks, and the red dots are the resulting subsamples.

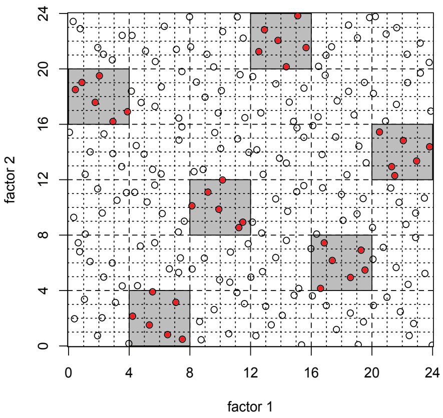

Figure 1.2: An example of LHD-based block bootstrap constructed from Figure 1.1(a)

The LHD-based block bootstrap has distinct advantages. The block structure takes into account the spatial dependency and therefore improves the estimation accuracy for correlation parameters in GP models. Because of the one-dimensional balance properties inherited from LHDs, the block bootstrap subsamples can be spread out more uniformly over the complete data and the resulting subsamples can represent the complete data effectively. As well, the LHD can result in variance reduction in estimation compared with simple random samples (Mckay et al. (1979), Stein (1987)). Therefore, the subagging estimates calculated by the proposed LHD-based subsamples are expected to outperform those calculated by the naive simple random subsamples in terms of estimation variance.

1.3.2 Variable selection using LHD-based block subagging

The procedure can be described in three steps:

Step 1: *Divide each dimension of the interested region $\Gamma \in [0, l]^d$ into m equally spaced intervals so that Γ consists of m^d disjoint hypercubes/blocks. Define each* *block by mapping $\boldsymbol{i}$ to a d-dimensional hypercube*

$$\mathcal{B}_n(\boldsymbol{i}) = \{\boldsymbol{x} \in R^d : bi_j \leq x_j \leq b(i_j + 1) \ \ and \ \ j = 1, ..., d\},$$

where $\boldsymbol{i} = (i_1, ...i_d)$, $i_j \in (0, ..., m-1)$, represents the index of each block and $b = l/m$ is the edge length of the hypercube. Let $|\mathcal{B}_n(\boldsymbol{i})|$ be the number of observations in the $\boldsymbol{i}$th hypercube/block. For simplicity, assume the data points are equally distributed over the blocks, $|\mathcal{B}_n(\boldsymbol{i})| = n/m^d$.

Step 2: *Select m blocks according to a randomly generated LHD(m, d). Each column of the LHD is a random permutation of $\{0, \ldots, m-1\}$, denoted by $\boldsymbol{\pi}_i = (\pi_i(1), \ldots, \pi_i(m))^T$ for $1 \leq i \leq d$. An m-run LHD is denoted by $\boldsymbol{i}_j^* = (\pi_1(j), \ldots, \pi_d(j))$, $j = 1, \ldots, m$, and the corresponding selected blocks are denoted by $\mathcal{B}_n(\boldsymbol{i}_1^*), \ldots, \mathcal{B}_n(\boldsymbol{i}_m^*)$. The bootstrapped subsamples, denoted by $y_1^*(\boldsymbol{x}_1^*), \ldots, y_N^*(\boldsymbol{x}_N^*)$, are the observations in the selected blocks, where $N = \sum_{i=1}^m |\mathcal{B}_n(\boldsymbol{i}_i^*)|$. Based on the subsamples, $\hat{\boldsymbol{\phi}}_N^*$ and its OSE $\hat{\boldsymbol{\phi}}_{N,OSE}^*$ are obtained by maximizing (1.3) and (1.4) respectively.*

Step 3: *Repeat Step 2 K times to obtain PMLEs $\hat{\boldsymbol{\phi}}_{N(j)}^*$ and the corresponding OSEs $\hat{\boldsymbol{\phi}}_{N,OSE(j)}^*$, where $j = 1, ..., K$. The subagging estimators are defined by $\hat{\boldsymbol{\phi}}_N = \frac{1}{K}\sum_{i=1}^K \hat{\boldsymbol{\phi}}_{N(i)}^*$ and $\hat{\boldsymbol{\phi}}_{N,OSE} = \frac{1}{K}\sum_{i=1}^K \hat{\boldsymbol{\phi}}_{N,OSE(i)}^*$.*

Figure 2 is an example with experimental region $\Gamma \in [0, 24]^2$, $d = 2$, $l = 24$. A common practice is that the data are collected by normalizing the experimental region to a unit cube. In such a case, we have $l = 1$. The circles represent the settings in which the experiments are performed and the total sample size is $n = 216$. The design, LHD$(6, 2)$, implemented here is denoted by $\boldsymbol{i}_1^* = (0, 4)$, $\boldsymbol{i}_2^* = (1, 0)$, $\boldsymbol{i}_3^* = (2, 2)$, $\boldsymbol{i}_4^* = (3, 5)$, $\boldsymbol{i}_5^* = (4, 1)$, $\boldsymbol{i}_6^* = (5, 3)$ and $m = 6$. According to this design, the LHD-based blocks are presented by the gray areas with $b = 4$ and $|\mathcal{B}_n(\boldsymbol{i})| = 6$. The red dots are the resulting LHD-based block subsamples with size $N = 36$.

Based on our procedure, the complexity is $O(n^3/m^{3(d-1)})$ for each subsample, which is computationally cheaper than $O(n^3)$ using the complete data, especially for large d. We assume data points are equally distributed over blocks in order to simplify the notation in the proof; the results still hold as long as the number of observations in each block is in the same order, $|\mathcal{B}_n(\boldsymbol{i}_i^*)| = O(n/m^d)$. For example, if the original data is collected by an orthogonal array-based Latin hypercube design (Tang (1993)), common in computer experiments, the proposed procedure can be successfully implemented. Based on our empirical experience, as long as each bootstrap subsample contains a small amount of empty blocks, we can still have an efficient representation of the original data. Empty blocks often occur when the original design has only few levels for some particular variables, such as qualitative variables. This issue can be addressed by modifying the LHDs by space-filling designs for quantitative and qualitative factors (Qian and Wu (2009), Deng et al. (2015)) and as a result, empty blocks can be avoided. Given the total sample size n, we have $1 \leq m \leq n^{\frac{1}{d-1}}$, since each bootstrap subsample has size N in the order of $O(n/m^{d-1})$. If $N = n/m^{d-1}$, then we have $m \leq n^{\frac{1}{d-1}}$ to ensure $N \geq 1$. Clearly, $m = 1$ provides no computational reduction because the full data is utilized. As m increases, the subsample size N decreases and a larger K is affordable given the same computational constraints.

Instead of selecting subsamples based on all the variables, this procedure can be modified to be based on a subset of variables. To do this, we can first select a subset of variables with dimension $\tilde{d}$, where $\tilde{d} < d$. This subset can be chosen randomly or according to some prior knowledge. Then, replace $LHD(m, d)$ in Step 2 by $LHD(m, \tilde{d})$

and select the subsamples only according to the $\tilde{d}$ variables. This is practically useful when d is large because the size of each subsample, n/m^{d-1}, can be relatively small increasing to $n/m^{\tilde{d}-1}$ by applying to subset variables. While, the proposed framework is constructed based on rectangular or hypercubic regions, it can be extended to regions with irregular shape by replacing the LHD in Step 2 by other space-filling designs constructed for nonrectangular regions, e.g., Draguljić et al. (2012) and Hung et al. (2012).

1.4 Theoretical properties

To understand the asymptotic properties of the subagging estimators, there are two distinct frameworks: increasing domain (Cressie (1993), Mardia and Marshall (1984)) asymptotics, where more and more data are collected in increasing domains while the sampling density stays constant, and fixed-domain asymptotics (Stein (1999), Liang et al. (2013)), where data are collected by sampling more and more densely in a fixed domain. The results in this research focus on increasing domain asymptotics. The results under fixed-domain asymptotics are more difficult to derive in general and rely on stronger assumptions, as discussed in the literature (Ying (1993), Zhang (2004)). It is shown by Zhang and Zimmerman (2005) that, given quite different behavior under the two frameworks in a general setting, their approximation quality performs about equally well for the exponential correlation function under certain assumptions. Results given here can then provide some insights about the subagging estimators in both frameworks. In ongoing work, we are exploring fixed domain asymptotics. More discussions are given in Section 7. Assumptions and the proofs are given in the Appendix and Supplemental Material.

We can show that the subagging estimator $\hat{\boldsymbol{\phi}}_N$ converges to the original PMLE $\hat{\boldsymbol{\phi}}_n$ in probability. Given the underlying probability space $(\Omega, \mathcal{F}, P)$ of a Gaussian process, a sample of size n with settings $\boldsymbol{x}_1(\omega), ..., \boldsymbol{x}_n(\omega)$ and responses $y(\boldsymbol{x})$'s are observed from a given realization $\omega \in \Omega$. Let $(\Lambda, \mathcal{G})$ be a measurable space on the realization. For each $\omega \in \Omega$, let $P^*_{N,\omega}$ be the probability measure induced by the m-run LHD-based block bootstrap on $(\Lambda, \mathcal{G})$. The proposed bootstrap is a method to generate a new dataset on $(\Lambda, \mathcal{G}, P^*_{N,\omega})$ conditional on the n original observations. For any LHD-based block bootstrapped statistic $\hat{T}^*_N$, we write $\hat{T}^*_N \to 0$ if for any $\epsilon > 0$ and any $\delta > 0$, $\lim_{n\to\infty} P\{P^*_{N,\omega}(|\hat{T}^*_N > \epsilon| > \delta)\} = 0$.

Theorem 1.1. *Under the assumptions (A.1)- (A.6), if $m = o(n^{-1/d})$ and $m \to \infty$, then $\hat{\boldsymbol{\phi}}_N - \hat{\boldsymbol{\phi}}_n \to 0$.*

Next we study the distributional consistency of the subagging estimators. Assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ to be the true regression coefficients, where, without loss of generality, $\boldsymbol{\beta}_{10}$ is an $s \times 1$ vector of nonzero regression coefficients and $\boldsymbol{\beta}_{20} = 0$ is a $(p-s) \times 1$ zero vector. Let $\gamma_0 = (\boldsymbol{\theta}_0, \sigma_0)$ denote the vector of true covariance parameters, $\hat{\boldsymbol{\phi}}^*_N = (\hat{\boldsymbol{\beta}}^*_{N,1}, \hat{\boldsymbol{\beta}}^*_{N,2}, \hat{\gamma}^*_N)$, $\hat{\boldsymbol{\phi}}_N = (\hat{\boldsymbol{\beta}}_{N,1}, \hat{\boldsymbol{\beta}}_{N,2}, \hat{\gamma}_N)$, and $\hat{\boldsymbol{\phi}}_n = (\hat{\boldsymbol{\beta}}_{n,1}, \hat{\boldsymbol{\beta}}_{n,2}, \hat{\gamma}_n)$. When the OSE approach is applied, we take $\hat{\boldsymbol{\phi}}^*_{N,OSE} = (\hat{\boldsymbol{\beta}}^*_{N,1,OSE}, \hat{\boldsymbol{\beta}}^*_{N,2,OSE}, \hat{\gamma}^*_{N,OSE})$, $\hat{\boldsymbol{\phi}}_N = (\hat{\boldsymbol{\beta}}_{N,1,OSE}, \hat{\boldsymbol{\beta}}_{N,2,OSE}, \hat{\gamma}_{N,OSE})$, and $\hat{\boldsymbol{\phi}}_{n,OSE} = (\hat{\boldsymbol{\beta}}_{n,1,OSE}, \hat{\boldsymbol{\beta}}_{n,2,OSE}, \hat{\gamma}_{n,OSE})$. Let $a_n = \max_j\{p'_{\lambda_n}(|\beta_j|) : \beta_j \neq 0\}$ and $b_n = \max_j\{p''_{\lambda_n}(|\beta_j|) : \beta_j \neq 0\}$. Let $\mathbf{g}(\boldsymbol{\phi}) = (p'_{\lambda}(\boldsymbol{\phi}))$ and $\mathbf{G}(\boldsymbol{\phi}) = \text{diag}(p''_{\lambda}(\boldsymbol{\phi}))$. Particularly, $\mathbf{g}(\boldsymbol{\beta}) = (p'_{\lambda}(|\beta_1|sgn(\beta_1)), ...,p'_{\lambda}(|\beta_p|sgn(\beta_p)))$ and $\mathbf{g}(\gamma) = \mathbf{0}$; $\mathbf{G}(\boldsymbol{\beta}) = \text{diag}(p''_{\lambda}(|\beta_1|), ..., p''_{\lambda}(|\beta_p|))$ and $\mathbf{G}(\gamma) = \mathbf{0}$.

Theorem 1.2. *Under assumptions (A.1)-(A.15), if $m = o(n^{-1/d})$ and $m \to \infty$, then*

(i) Sparsity: $\hat{\boldsymbol{\beta}}_{N,2} = 0$ with probability tending to 1.

(ii) Asymptotic normality: for the mean function coefficients,

$$\sqrt{Kn/m^{d-1}}(\boldsymbol{J}(\boldsymbol{\beta}_{10}) + \boldsymbol{G}(\boldsymbol{\beta}_{10}))(\hat{\boldsymbol{\beta}}_{N,1} - \hat{\boldsymbol{\beta}}_{n,1}) \to N(0, \boldsymbol{J}(\boldsymbol{\beta}_{10}));$$

for the correlation parameters,

$$\sqrt{Kn/m^{d-1}}(\hat{\gamma}_N - \hat{\gamma}_n) \to N(0, \boldsymbol{J}(\gamma_0)^{-1}).$$

In Theorem 1.3, it shows that when the OSE algorithm is applied, the resulting subagging estimators are asymptotically consist to the original OSEs using the complete data.

Theorem 1.3. *Under assumptions (A.1)-(A.15), if $m = o(n^{-1/d})$ and $m \to \infty$, then*

(i) Sparsity: $\hat{\boldsymbol{\beta}}_{N,2,OSE} = 0$ with probability tending to 1.

(ii) Asymptotic normality: for the mean function coefficients,

$$\sqrt{Kn/m^{d-1}}(\hat{\boldsymbol{\beta}}_{N,1,OSE} - \hat{\boldsymbol{\beta}}_{n,1,OSE}) \to N(0, \boldsymbol{J}(\boldsymbol{\beta}_{10})^{-1});$$

for the correlation parameters,

$$\sqrt{Kn/m^{d-1}}(\hat{\gamma}_{N,OSE} - \hat{\gamma}_{n,OSE}) \to N(0, \boldsymbol{J}(\gamma_0^{-1})).$$

1.5 Numerical studies

In this section, we report on two sets of simulations conducted to study the finite-sample performance of the proposed method. One demonstrates the performance of the subagging approach compared with the original approach using all the data. The other illustrates the advantages of the proposed experimental design-based subsampling scheme by comparison with simple random sampling. The performance was evaluated in two aspects: the accuracy of variable selection and the parameter estimation, including the mean function coefficients and the correlation parameters using one-step sparse estimation as described in (1.4). The accuracy of variable selection was measured by two scores: the average number of the nonzero regression coefficients correctly identified in the repeated simulations, denoted by AC: the average number of the zero regression coefficients misspecified, denoted by AM. All the simulations were conducted by a 2.7GHz, 16G RAM workstation. Hereafter, we omit the subscript OSE for notational convenience.

1.5.1 Subagging vs. the estimation using all data

Three sample sizes, $n = 1000$, $n = 2000$ and $n = 3000$, were considered and the data were generated from a regular grid in a four-dimensional space, $[0, 1]^4$. The proposed method is particularly useful for data collected from irregular grids. The reason to generate the simulations from a regular grid in this simulation was that the original PMLE calculation using full data can be further speeded up by Kronecker product techniques and some matrix singularity can be avoided (Rougier (2008)). These techniques are only applicable to data sets collected from a regular grid; a favorable comparison of the proposed method would make an even stronger case for the proposed procedure.

Simulations were generated from a Gaussian process with the mean function coefficients $\boldsymbol{\beta} = (1, 0.5, 0, 0)$ and the correlation function

$$\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\sum_{i=1}^{4} \theta_i |x_{1i} - x_{2i}|)$$

where $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 1$ and $\sigma = 0.1$. For each choice of sample size, 50 data sets were simulated. For each simulated data set, 10 LHD-based block bootstrap samples were collected with $m = 4$. Due to the computation time needed for the complete data,

Table 1.1: Comparisons with all data

	$n = 1000$		$n = 2000$		$n = 3000$	
	LHD	*AllData*	*LHD*	*AllData*	*LHD*	*AllData*
θ_1	1.91 (0.55)	1.14 (0.05)	1.38 (0.35)	1.02 (0.02)	1.10(0.10)	0.97(0.02)
θ_2	1.94 (1.20)	1.08 (0.07)	1.16 (0.14)	1.00 (0.03)	1.17(0.08)	1.03(0.03)
θ_3	1.70 (0.68)	1.03 (0.04)	1.14 (0.20)	0.92 (0.03)	1.15(0.07)	1.06(0.02)
θ_4	1.77 (0.83)	1.04 (0.04)	1.37 (0.45)	1.02 (0.04)	1.10(0.03)	1.00(0.03)
β_1	$1.00(3.2 \times 10^{-3})$	$1.02(3.6 \times 10^{-3})$	$0.99(4.2 \times 10^{-3})$	$0.99(7.9 \times 10^{-3})$	$1.01(3.4 \times 10^{-3})$	$1.00(3.7 \times 10^{-3})$
β_2	$0.46(1.7 \times 10^{-2})$	$0.43(3.6 \times 10^{-2})$	$0.51(3.3 \times 10^{-3})$	$0.50(6.1 \times 10^{-3})$	$0.49(5.5 \times 10^{-3})$	$0.50(3.7 \times 10^{-3})$
$AC/2$	1	0.93	1	1	1	1
$AM/2$	0	0	0	0	0	0
time	243	464	990	2402	2524	8623

the tuning parameter $\lambda = 0.1$ was fixed for all simulations.

In Table 1.1, the parameter estimation and the computing time are reported. Standard deviations are given in parenthesis. The rows $AC/2$ and $AM/2$ represent the correct identification rate and the variable misspecification rate, respectively. The results in Table 1 suggest that the estimated parameters using LHD-based subagging are consistent with those obtained using complete data, as is the variable selection performance. In terms of computing time, the proposed subagging is much faster to compute compared with the conventional approach, especially when the sample size of the complete data is large.

1.5.2 LHD-based block subsampling vs. random subsampling

An important feature of the proposed subsampling scheme is that it borrows the idea of space-filling design to achieve an efficient data reduction. To demonstrate this, we compared its performance, denoted by LHD, with two alternatives: simple random sampling, denoted by SRS, and random blocks sampling, denoted by RBS, with the same sample size. We first compared the performance of LHD with SRS in two different settings of subsampling scheme: $m = 4$ and $m = 6$.

The data were generated from a six-dimensional space, $[0, 1]^6$ with sample size

$n = 3600$. We consider the same correlation function as before with the mean function coefficients $\boldsymbol{\beta} = (1, 0.5, 0.3, 0, 0, 0)$, three non-zero coefficients with different signal strength and three zero coefficients. Results are summarized based on 100 simulations and 20 LHD-based block bootstrap samples collected for each simulation. To focus on the capability of selecting active factors, the proposed subsampling was performed on the first three variables and the resulting sample sizes for $m = 4$ and $m = 6$ were approximately 225 and 100, respectively.

In Table 1.2, the estimated parameters, the correct identification rates, and the variable misspecification rates are reported. In terms of parameter estimation, LHD performs similarly to SRS in estimating the mean function coefficients. For estimating the correlation parameters, LHD outperforms SRS with a much smaller estimation variance, especially when the subsample size is smaller ($m = 6$). In general, it appears that the proposed subsampling based on LHDs provides an effective variance reduction in parameter estimation, which is consistent with the theoretical justifications in experimental design literature (Mckay et al. (1979), Stein (1987)). In terms of variable selection, the correct identification rate for the LHD-based subsampling is 21% higher than SRS when $m = 4$ and 13% higher when $m = 6$. Both methods perform equally well with zero misspecification rate. To further assess the variable selection accuracy, the frequencies of individual variables identified from 100 simulations are reported in the last three rows of the table: $Fre(\beta_1)$, $Fre(\beta_2)$ and $Fre(\beta_3)$. The identification frequencies for β_3 decrease as expected due to its weak signal. But the proposed subsampling can still identify such a weak signal with at least 66% higher frequency compared with simple random subsamples.

Table 1.2: Comparisons with simple random subsampling

	$m = 4$		$m = 6$	
	LHD	SRS	LHD	SRS
θ_1	1.91 (0.60)	1.89 (4.11)	2.63 (1.63)	2.61 (9.93)
θ_2	2.24 (1.71)	1.90 (3.73)	2.64 (2.01)	2.95 (10.56)
θ_3	1.96 (0.79)	1.99 (2.66)	2.49 (1.14)	3.18 (10.97)
θ_4	1.93 (0.58)	1.92 (4.11)	2.69 (1.74)	2.90 (12.78)
θ_5	1.78 (0.35)	1.72 (1.91)	2.58 (0.84)	2.50 (12.55)
θ_6	1.89 (0.48)	1.94 (3.84)	2.74 (1.78)	1.80 (8.65)
β_1	$1.01(1.5 \times 10^{-3})$	$0.99(3.3 \times 10^{-3})$	$1.03(1.6 \times 10^{-3})$	$0.99(1.5 \times 10^{-3})$
β_2	$0.52(3.2 \times 10^{-3})$	$0.52(2.9 \times 10^{-3})$	$0.53(4.4 \times 10^{-3})$	$0.55(6.7 \times 10^{-3})$
β_3	$0.14(1.2 \times 10^{-2})$	$0.10(2.1 \times 10^{-2})$	$0.15(1.1 \times 10^{-2})$	$0.15(2.5 \times 10^{-2})$
AC/3	0.98	0.81	1	0.87
$AM/3$	0	0	0	0
$Fre(\beta_1)$	1	1	1	1
$Fre(\beta_2)$	1	1	1	1
$Fre(\beta_3)$	0.93	0.40	1	0.60

1.5.3 Comparison with random blocks subsampling

In the next simulation, the proposed sampling scheme was compared with RBS in which blocks are selected randomly without the one-dimensional projection property. The data were generated from a 4-dimensional space with $n = 2000$. We took the same correlation function as before with the mean function coefficients set to be $\boldsymbol{\beta} = (1, 0.5, 0.1, 0)$: three non-zero coefficients with different signal strength and one zero coefficient. Results are summarized in Table 1.3 based on 100 simulations and $K = 20$. The results of SRS with the same subsample size are also listed for comparison. In general, LHD outperforms the other two sampling and RBS performs slightly better than SRS. Compared with RBS, the proposed method has a higher frequency of identifying the nonactive variable: 0.95 vs. 0.85. Moreover, LHD has less bias and a smaller variance in parameter estimation, empirically demonstrating the advantage of the one-dimensional balance property of LHD.

Table 1.3: Comparisons with simple random sampling of blocks

m=4	θ_1	θ_2	θ_3	θ_4	
LHD	1.21(0.26)	1.29(0.38)	1.27(0.32)	1.34(0.17)	
RBS	1.44(0.30)	1.50(0.34)	1.43(0.37)	1.50(0.33)	
SRS	1.77(0.88)	1.59(0.38)	1.55(0.72)	1.53(1.34)	
	β_1	β_2	β_3	$AC/3$	Freq($\beta_4 = 0$)
LHD	$1.00(1.9 \times 10^{-6})$	$0.50(2.3 \times 10^{-6})$	$0.09(1.8 \times 10^{-6})$	1.0	0.95
RBS	$1.00(7.5 \times 10^{-6})$	$0.51(3.0 \times 10^{-6})$	$0.08(3.1 \times 10^{-6})$	1.0	0.85
SRS	$1.00(3.7 \times 10^{-6})$	$0.51(1.1 \times 10^{-6})$	$0.09(1.2 \times 10^{-6})$	1.0	0.63

1.6 Data center thermal management

A data center is a computing infrastructure facility that houses large amounts of information technology equipment used to process, store, and transmit digital information. Data center facilities constantly generate large amounts of heat to the room, which must be maintained at an acceptable temperature for reliable operation of the equipment. A significant fraction of the total power consumption in a data center is for heat removal, and determining the most efficient cooling mechanism has become a major challenge. Since the thermal process in a data center is complex and depends on many factors, a crucial step is to model the thermal distribution at different experimental settings and identify important factors that have significant impacts on the thermal distribution (Hung et al. (2012)).

For a data center thermal study, physical experiments are not always feasible because some settings are highly dangerous and expensive to perform. Therefore, simulations based on computational fluid dynamics (CFD) are widely used. Such simulations using complex mathematical models are often called computer experiments (Santner et al. (2003), Fang et al. (2006)). In this example, CFD simulations were conducted at IBM T. J. Watson Research Center based on an actual data center layout. Detailed discussions about the CFD simulations can be found in (Lopez and Hamann (2011)).

There were 27,000 temperature outputs generated from the CFD simulator based on an irregular grid over an 9-dimensional space. The nine variables are listed in Table 1.4, including four computer room air conditioning (CRAC) units with different flow rates ($x_1, ..., x_4$), the overall room temperature setting (x_5), the perforated floor tiles with different percentage of open areas (x_6), and spatial location in the data center (x_7 to x_9).

Gaussian process models are widely used for the analysis of computer experiments because they provides a flexible interpolator for the deterministic simulation outputs (Santner et al. (2003)). However, in this example, it is computationally prohibitive to build a GP model based on the complete CFD data. So we implemented the proposed LHD-based subagging approach with $m = 3$ for the first seven variables.

The fitted GP model is reported in the last two columns of Table 4, where $\hat{\boldsymbol{\beta}}$ represents the estimated mean function coefficients and $\hat{\boldsymbol{\theta}}$ represents the correlation parameters estimated using the exponential covariance function. From the fitted model, it appears that seven out of the nine variables have significant effects on the mean function. The main effects plot based on the fitted GP model is given in Figure 2.1. It also appears that the two variables, x_5 and x_6, which are identified as nonactive have relatively small impacts on cooling. This result provides important information regarding the efficiency of different cooling methods, because the variables are associated with two cooling mechanisms, a conventional cooling approach and a chilled water based cooling system. Among the active variables, the height (x_9) has a relatively large positive effect, which agrees with the general understanding of thermal dynamics that temperature increases significantly with height in a data center. The results also indicate that, among the four CRAC units in different locations of a data center, the first

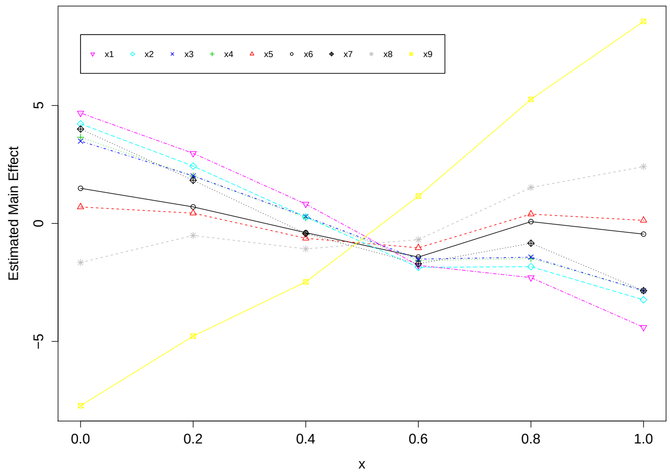

Figure 1.3: Main effect plot

two CRAC units have significant effects on reducing the room temperature. This can help engineer locations of the CRAC units more effectively and improve the efficiency of the cooling mechanism.

Table 1.4: Analysis for the data center example

	Variable	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\theta}}$
x_1	CRAC unit 1 flow rate	-7.5	5.3
x_2	CRAC unit 2 flow rate	-13.1	1.3
x_3	CRAC unit 3 flow rate	-2.7	0.3
x_4	CRAC unit 4 flow rate	-7.1	13.2
x_5	Room temperature setting	0	0.9
x_6	Tile open area percentage	0	0.6
x_7	Location in x-axis	-11.3	21.44
x_8	Location in y-axis	2.1	9.5
x_9	Height	17.8	0.8

1.7 Discussion

Future work will be explored in several directions. Extensions of the proposed procedure to optimal designs with better space-filling properties are appealing. For example, it is known that randomly generated LHDs can contain some structure. To further enhance desirable space-filling properties, various modifications are proposed. Numerical comparisons and theoretical developments of the generalization to different types of optimal space-filling designs will be studied. An interesting and important issue of the LHD-based block bootstrap is to determine the optimal block size. This topic has been discussed for conventional block bootstrap methods (Nordman et al. (2007)), but the solutions therein are not directly applicable to GP models. We plan to study the optimal block size for our procedure based on a new criterion defined for GP. Theoretical development under fixed-domain asymptotics will be explored by extending the results of Ying (1993) and Hung (2011), and subagging predictors will also be developed. As pointed out by the referees, another interesting extension of the proposed work is to perform variable selection not only in the mean function but also in the correlation function. We are currently developing an extension to address this issue so that identification of linear effects in the mean function and nonlinear effects in the covariance function can be both achieved.

1.8 Technical proofs

1.8.1 Assumptions

(A.1) $\frac{n}{m^d}\boldsymbol{Cov}\{(\bar{y}_{\boldsymbol{i}}-\mu)^2,(\bar{y}_{\boldsymbol{j}}-\mu)^2\}=O(1)$, $\boldsymbol{i}=(i_1,\ldots,i_d)\neq\boldsymbol{j}=(j_1,\ldots,j_d)$.

(A.2) $|\tau_n^2|=O(1)$.

(A.3) $\lim_{n\to\infty}\sup_{\boldsymbol{\theta}}\lambda_{\max}(E_n(\boldsymbol{\theta}))=0$ when the block space $b=l/m\to\infty$.

(A.4)$\forall$ $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2\in\Theta$, $|q_s(\cdot,\boldsymbol{\phi}_1)-q_s(\cdot,\boldsymbol{\phi}_2)|\leq L_s|\boldsymbol{\phi}_1-\boldsymbol{\phi}_2|a.s.P$, where L_s is Lipschitz constant and $\sup_n\{n^{-1}\sum_{s=1}^n\boldsymbol{E}L_s\}=O(1)$.

(A.5)Θ is compact.

(A.6)The functions $q_s(\omega, \boldsymbol{\phi})$ and $r_n(\omega, \boldsymbol{\phi})$ are such that $q_s(\cdot, \boldsymbol{\phi})$ and $r_n(\cdot, \boldsymbol{\phi})$ are measurable for all $\boldsymbol{\phi} \in \Theta$, a compact subset of R^p. In addition, $q_s(\omega, \cdot) : \Theta \longrightarrow R$ and $r_n(\omega, \cdot) : \Theta \longrightarrow R$ are continuous on Θ a.s.-P, $s = 1, \cdots, n$.

(A.7) $Q_n(\omega, \cdot) : \Theta \to R$ is continuously differentiable of order 2 on Θ a.s. P.

(A.8) There exists a sequence $J_n(\boldsymbol{\phi}) : \Theta \to R^{p\times p}$ such that $\nabla^2 Q_n(\cdot, \boldsymbol{\phi}) - J_n(\boldsymbol{\phi}) \xrightarrow{\text{P}} 0$ as $n \to \infty$ uniformly on Θ.

(A.9) $\lim_{n\to\infty} J_n^{-1}(\boldsymbol{\phi}^0) = 0$.

(A.10) $Q_N^*(\lambda, \omega, \cdot) : \Theta \to R$ are continuously differentiable of order 2 on Θ a.s. P. The function $\nabla^2 Q_n(\omega, \boldsymbol{\phi})$ is such that $\nabla^2 Q_n(\cdot, \boldsymbol{\phi})$ is measurable for all $\boldsymbol{\phi} \in \Theta$ and $\nabla^2 Q_n(\omega, \cdot) : \Theta \to R$ is continuous on Θ a.s.-P.

(A.11) $\forall$ $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2 \in \Theta$,$|\nabla^2 Q_n(\cdot, \boldsymbol{\phi}_1) - \nabla^2 Q_n(\cdot, \boldsymbol{\phi}_2)| \leq M_s|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2|a.s.P$, where M_s is Lipschitz constant and $\sup_n\{n^{-1}\sum_{s=1}^n \boldsymbol{E}M_s\} = O(1)$.

(A.12) $a_n = O(n^{-\frac{1}{2}})$ and $b_n \to 0$ as $n \to \infty$

(A.13) There exit positive constants c_1 and c_2 such that when $\beta_1, \beta_2 > c_1\lambda_n$, $|p''_{\lambda_n}(\beta_1) - p''_{\lambda_n}(\beta_2)| \leq c_2|\beta_1 - \beta_2|$.

(A.14) $\lambda_n \to 0$, $n^{\frac{1}{2}}\lambda_n \to \infty$ as $n \to \infty$.

(A.15) $\liminf_{n\to\infty} \liminf_{\boldsymbol{\beta}\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\boldsymbol{\beta}) > 0$.

Assumption (A.3) controls the correlation between bootstrapped blocks. (A.4) and (A.5) are required in order to achieve uniform convergency of the bootstrapped likelihood function. (A.6) ensures the existence of the estimators. (A.7)-(A.9) are regularity conditions for standard MLE consistency in GP models, analogous to the conditions in Mardia and Marshall (1984). (A.10) ensures the existence of the covariance matrix. (A.11) is the global Lipschitz condition for $\nabla^2 Q_n(\omega, \cdot)$ which guarantees the convergence of the covariance matrix calculated based on the LHD-based block bootstrap. (A.12)-(A.15) are mild regularity conditions regarding the penalty function.

1.8.2 Lemmas

Lemma 1.1. *LHD-based block bootstrap mean is unbiased, i.e.,*

$$\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = \bar{y}_n.$$

Proof of Lemma 1.1: Since the data points are equally distributed over all the blocks, we have $\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = m^{-d}\sum_{i_1,\ldots,i_d} \bar{y}_{i_1,\ldots,i_d} = \bar{y}_n$.□

Lemma 1.2. *Let* $\bar{y}_{\boldsymbol{i}} = \frac{1}{\mathcal{B}_n(\boldsymbol{i})}\sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i})} y_s$, $\forall \boldsymbol{i} = (i_1, \ldots, i_d)$. *Assuming (A.1), (A.2) and* $m = o(n^{1/d})$, *we have*

$$\frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2 - \tau_n^2 \xrightarrow{\text{P}} 0,$$

where $\tau_n^2 = \frac{1}{n}\sum_{s,t=1}^n \boldsymbol{Cov}(Y_s(\boldsymbol{x}_s), Y_t(\boldsymbol{x}_t)).$

Proof of Lemma 1.2: Let $A_n = \frac{n}{m^{2d}}\sum_{i_1,\ldots,i_d}(\bar{y}_{i_1,\ldots,i_d} - \mu)^2$. We can show that $\boldsymbol{Cov}(A_n, A_n) = 0$ and $\boldsymbol{E}(A_n) = \tau_n^2$.

$$\begin{aligned}
\boldsymbol{Cov}(A_n, A_n) &= \boldsymbol{Cov}(\frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2, \frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2) \\
&= \frac{1}{n^2} \sum_{\boldsymbol{i}} \sum_{\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2}, \boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{Cov}\{(y_{s_1} - \mu)(y_{s_2} - \mu), (y_{t_1} - \mu)(y_{t_2} - \mu)\} \\
&\quad + \frac{1}{n^2} \sum_{\boldsymbol{i} \neq \boldsymbol{j}} \sum_{\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2} \in \mathcal{B}_n(\boldsymbol{i})} \sum_{\boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{j})} \boldsymbol{Cov}\{(y_{s_1} - \mu)(y_{s_2} - \mu), (y_{t_1} - \mu)(y_{t_2} - \mu)\}
\end{aligned}$$

By expanding two terms above separately, we have $\boldsymbol{Cov}(A_n, A_n) = O(\frac{1}{n} + \frac{m^d}{n}) \to 0$ as

$m = o(n^{1/d})$. In addition, we have

$$\boldsymbol{E}(A_n) - \tau_n^2 = \frac{1}{n}\sum_{\boldsymbol{i}\neq\boldsymbol{j}}\sum_{\boldsymbol{x}_s\in\mathcal{B}_n(\boldsymbol{i}),\boldsymbol{x}_t\in\mathcal{B}_n(\boldsymbol{j})}\sigma^2\psi(y(\boldsymbol{x}_s), y(\boldsymbol{x}_t)) = o(1)$$

Thus, $A_n - \tau_n^2 \xrightarrow{\text{P}} 0$. □

Lemma 1.3. *Assume (A.1)- (A.2), then*

$$n{\tau_N^*}^2/m^{d-1} - \tau_n^2 \xrightarrow{\text{P}} 0,$$

where ${\tau_N^*}^2 = \boldsymbol{Cov}^*_{N,\omega}(\bar{y}_N^*, \bar{y}_N^*)$.

Proof of Lemma 1.3: Based on the definition of $n{\tau_N^*}^2/m^{d-1}$, we have

$$n{\tau_N^*}^2/m^{d-1} = \tfrac{n}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_1^*}) + 2\tfrac{n(m-1)}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_2^*}).$$

For the first term on the right, we have

$$\frac{n}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_1^*}) = \frac{n}{m^{2d}}\sum_{i_1,\dots,i_d}(\bar{y}_{i_1,\dots,i_d} - \mu)^2 - \frac{n}{m^d}(\bar{y}_n - \mu)^2 = A_n - B_n.$$

By Lemma 1.2, we have $A_n - \tau_n^2 \xrightarrow{\text{P}} 0$. For $B_n = \frac{n}{m^d}(\bar{y}_n - \mu)^2$, by the central limit theorem for $\bar{y}_n$, we have $B_n \xrightarrow{\text{P}} 0$. Next, it suffices to show that $\frac{n(m-1)}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_2^*})$ converges to 0 in probability under P. The following double summation $\sum_{i_1,\dots,j_d,j_1,\dots,j_d}$ are taken over $\boldsymbol{i} = (i_1, \dots, i_d)$ and $\boldsymbol{j} = (j_1, \dots, j_d)$ such that $\mathcal{B}_n(\boldsymbol{i})$ and $\mathcal{B}_n(\boldsymbol{j})$ are not equal and are selected together.

$$\begin{aligned}\frac{n(m-1)}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_2^*}) &= \frac{n(m-1)}{m^{2d}}\frac{1}{m^d - 1 - d(m-1)}\sum_{\boldsymbol{i}\neq\boldsymbol{j}}(\bar{y}_{\boldsymbol{i}} - \mu)(\bar{y}_{\boldsymbol{j}} - \mu)\\ &\quad + \frac{n(m-1)}{m^d}[1 - \frac{2m^d}{m\{m^d - 1 - d(m-1)\}}](\bar{y}_n - \mu)^2\\ &= C_n + D_n.\end{aligned}$$

Similar to A_n and B_n, we can show that $C_n \xrightarrow{\text{P}} 0$ and $D_n \xrightarrow{\text{P}} 0$. The result follows immediately. □

Lemma 1.4. *Under (A.1)-(A.3), for each* $\boldsymbol{\phi} \in \Theta$*,*

$$\begin{aligned}&\lim_{n\to\infty} P\Big[P^*_{N,\omega}\big(|N^{-1}\sum_{s=1}^{N} q_s^*(\cdot,\omega,\boldsymbol{\phi}) + N^{-1} r_N^*(\cdot,\omega,\boldsymbol{\phi})\\ &-n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi}) - n^{-1} r_n(\omega,\boldsymbol{\phi})| > \delta\big) > \xi\Big] = 0.\end{aligned}$$

Proof of Lemma 1.4: Rewrite the bootstrapped likelihood function as $I_1 + I_2 + I_3$, where $I_1 = N^{-1}\sum_{s=1}^{N}\{q_s^*(\cdot,\omega,\boldsymbol{\phi}) - \boldsymbol{E}^* q_s^*(\cdot,\omega,\boldsymbol{\phi})\}$,

$I_2 = \{N^{-1}\sum_{s=1}^{N} \boldsymbol{E}^* q_s^*(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi})\}$, $I_3 = N^{-1} r_N^*(\cdot,\omega,\boldsymbol{\phi}) - n^{-1} r_n(\omega,\boldsymbol{\phi})$. By Lemma 1.3, $I_2 \equiv 0$. For I_3, it can be shown that $n^{-1} r_n(\omega,\boldsymbol{\phi}) \to 0$ in P and $N^{-1} r_N^*(\cdot,\omega,\boldsymbol{\phi}) \to 0$, prob-$P^*_{N,\omega}$ prob-P. For notation simplicity, we omit $\boldsymbol{\theta}$ in the following discussion. The expectation and variance of $n^{-1} r_n(\omega,\boldsymbol{\phi})$ are:

$$\begin{aligned}&|\boldsymbol{E}\{n^{-1} r_n(\omega,\boldsymbol{\phi})\}|\\ \leq\ & \frac{1}{2n\sigma^2(1+g)}\lambda_{\max}(E_n)\lambda_{\max}(D_n^{-1}) + |\log\{1 + \lambda_{\max}^n(E_n)|D_n^{-1}\}|\\ =\ & o(1)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{Var}(n^{-1} r_n(\omega,\boldsymbol{\phi})) &\leq \frac{1}{4(1+g)^2\sigma^4 n^2}\boldsymbol{Var}\{\sum_{i=1}^{n}(\sum_{j=1}^{n} u_{ij}\varepsilon_j)^2\}\\ &\leq \frac{c_n}{4(1+g)^2\sigma^4 n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{Var}(\varepsilon_j^2) = o(1)\end{aligned}$$

where ε_j is the i^{th} entry of $D_n^{-1}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})$ and $\boldsymbol{u}_i = (u_{ij})$ is the i^{th} row of U_n; $c_n = \max_i\{\sum_{j=1}^{n} u_{ij}^2\}$.

In addition, as $\lambda_{\max}(E_N^*) \leq \lambda_{\max}(E_n)$ and $\lambda_{\max}({D_N^*}^{-1}) \leq \lambda_{\max}(D_n^{-1})$, we have

$$\begin{aligned}&\frac{1}{2\sigma^2(1+g^*)}(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta})^T {D_N^*}^{-1} E_N^* {D_N^*}^{-1}(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta})\\ \leq\ & \frac{1}{2\sigma^2}\lambda_{\max}(E_n)\lambda_{\max}(D_n^{-1})\|\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta}\|_2^2.\end{aligned}$$

According to Lemma 1.6 below, we have $N^{-1}\|\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta}\|_2^2 - n^{-1}\|\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\|_2^2 \to 0$ prob-$P_{N,\omega}^*$ prob-P. Similarly, we can bound $\log|I_N + U_N^{*T} D_N^{*-1} U_N^*|$. As $\lambda_{\max}(E_n) \to 0$, we have $\frac{1}{N} r_N^*(\cdot,\omega,\boldsymbol{\phi}) \to 0$, prob-$P_{N,\omega}^*$ prob-P.

So when n is sufficiently large, we only need to show that $\lim_{n\to\infty} P\big[P_{N,\omega}^*(|I_1| > \delta) > \xi\big] = 0$. By Chebyshev's inequality,

$$P_{N,\omega}^*(|I_1| > \delta) \le \frac{1}{\delta^2}\boldsymbol{Var}_{N,\omega}^*(\bar{q}_N^*(\cdot,\omega,\boldsymbol{\phi})).$$

By Lemma 1.1, $r^{-1}\boldsymbol{Var}_{N,\omega}^*(\bar{q}_N^*(\cdot,\omega,\boldsymbol{\phi})) = O_p(1)$, together with the fact that $N = n/m^{d-1}$,

$$\begin{aligned} P\big[P_{N,\omega}^*(|I_1| > \delta) > \xi\big] &\le P\big[\frac{n}{m^{d-1}}\frac{1}{\delta^2}\boldsymbol{Var}_{N,\omega}^*(\bar{q}_N^*(\cdot,\omega,\boldsymbol{\phi})) > \xi\frac{n}{m^{d-1}}\big] \\ &= O(m^{2d-2}/n^2) \to 0. \end{aligned}$$

□

The next lemma further extends Lemma 1.4 to the uniform weak law of large numbers for the LHD-based block bootstrap likelihood functions.

Lemma 1.5. *(Uniform Weak Law of Large Numbers) Under (A.1)-(A.5), ∀ $\delta,\xi > 0$,*

$$\lim_{n\to\infty} P\Big[P_{N,\omega}^*(\sup_{\boldsymbol{\phi}\in\Theta}|N^{-1}\sum_{s=1}^N q_s^*(\cdot,\omega,\boldsymbol{\phi}) + N^{-1}r_N^*(\cdot,\omega,\boldsymbol{\phi}) \\ -n^{-1}\sum_{s=1}^n q_s(\omega,\boldsymbol{\phi}) - n^{-1}r_n(\omega,\boldsymbol{\phi})| > \delta) > \xi\Big] = 0.$$

Proof of Lemma 1.5: By Lemma 1.4, $|n^{-1}r_n(\omega,\boldsymbol{\phi}) - N^{-1}r_N^*(\cdot,\omega,\boldsymbol{\phi})|$ can be arbitrarily small as n is large enough uniformly over Θ. We only need to show that

$$\lim_{n\to\infty} P\big[P_{N,\omega}^*(\sup_{\boldsymbol{\phi}\in\Theta}|N^{-1}\sum_{s=1}^N q_s^*(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^n q_s(\omega,\boldsymbol{\phi})| > \delta) > \xi\big] = 0.$$

Given $\epsilon > 0$ that will be selected later, let $\{\eta(\boldsymbol{\phi}_j,\epsilon), j = 1,\ldots,K\}$ be a finite cover of

Θ, where $\eta(\boldsymbol{\phi}_i, \epsilon) = \{\boldsymbol{\phi} \in \Theta : |\boldsymbol{\phi} - \boldsymbol{\phi}_j| < \epsilon\}$. Then

$$\begin{aligned} & \sup_{\boldsymbol{\phi}} |N^{-1} \sum_{s=1}^{N} q_s^*(\cdot, \omega, \boldsymbol{\phi}) - n^{-1} \sum_{s=1}^{n} q_s(\omega, \boldsymbol{\phi})| \\ = \quad & \max_{j=1}^{K} \sup_{\boldsymbol{\phi} \in \eta(\boldsymbol{\phi}_j, \epsilon)} |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi})|. \end{aligned}$$

It follows that $\forall\ \delta > 0$ with fixed ω,

$$\begin{aligned} & P_{N,\omega}\big(\sup_{\boldsymbol{\phi} \in \Theta} |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi})| > \delta \big) \\ \leq \quad & \sum_{j=1}^{K} P_{N,\omega}\big(\sup_{\boldsymbol{\phi} \in \eta(\boldsymbol{\phi}_j, \epsilon)} |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi})| > \delta \big). \end{aligned}$$

For $\forall\ \boldsymbol{\phi} \in \eta(\boldsymbol{\phi}_j, \epsilon)$, by Global Lipschitz condition,

$$|\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi})| \leq |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}_j) - \bar{q}_n(\omega, \boldsymbol{\phi}_i)| + N^{-1} \sum_{s=1}^{N} L_s^* \epsilon + n^{-1} \sum_{s=1}^{n} L_s \epsilon,$$

where L_s^* is the bootstrapped Lispchitz coefficient.

By Markov inequality and the fact that $\sup_n\{n^{-1} \sum_{s=1}^{n} \boldsymbol{E} L_s\} = O(1)$, we have $P(n^{-1} \sum_{s=1}^{n} L_s > \delta/3) \leq 3\epsilon\Delta/\delta \leq \xi/3$, where Δ is a large constant. If we choose $\epsilon < \xi\delta/(9\Delta)$, we have

$$\begin{aligned} & P\big[P_{N,\omega}^*(\sup_{\boldsymbol{\phi} \in \eta(\boldsymbol{\phi}_j, \epsilon)} |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi})| > \delta) > \xi\big] \\ \leq \quad & P\big[P_{N,\omega}^*(|\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}_j) - \bar{q}_n(\omega, \boldsymbol{\phi}_j)| > \delta) > \xi/3\big] \\ & + P\big[P_{N,\omega}^*(N^{-1} \sum_{s=1}^{N} L_s^* \epsilon > \delta/3) > \xi/3\big] + P[n^{-1} \sum_{s=1}^{n} L_s \epsilon > \delta/3] \\ = \quad & I_1 + I_2 + I_3. \end{aligned}$$

According to Lemma 1.4, $I_1 \leq \xi/3$ when n is large enough. By Markov's inequality,

$$P_{N,\omega}^*(N^{-1} \sum_{s=1}^{N} L_s^* \epsilon > \delta/3) \leq N^{-1} \sum_{s=1}^{N} \boldsymbol{E}^* L_s^*/(\delta/3\epsilon) = n^{-1} \sum_{s=1}^{n} L_s/(\delta/3\epsilon).$$

The last equality is because of Lemma 1.1. Thus, $I_2 < \xi/3$ as well as I_3. □

1.8.3 Consistency of the LHD-based block bootstrap mean

Before studying the asymptotic performance of MLEs, we first focus on understanding properties of the LHD-based block bootstrap mean, which is an important foundation to the theoretical development of $\hat{\boldsymbol{\phi}}^*_N$ later.

The LHD-based block bootstrap can be formulated mathematically as follows. Given the underlying probability space $(\Omega, \mathcal{F}, P)$ of a Gaussian process, a sample of size n with settings $\boldsymbol{x}_1(\omega), ..., \boldsymbol{x}_n(\omega)$ and responses $y(\boldsymbol{x})$'s are observed from a given realization $\omega \in \Omega$. Let $(\Lambda, \mathcal{G})$ be a measurable space on the realization. For each $\omega \in \Omega$, denote $P^*_{N,\omega}$ as the probability measure induced by the m-run LHD-based block bootstrap on $(\Lambda, \mathcal{G})$. The proposed bootstrap is a method to generate new dataset on $(\Lambda, \mathcal{G}, P^*_{N,\omega})$ conditional on the n original observations. Let $\tau_t : \Lambda \to \{1, ..., n\}$ denote a random index generated by the LHD-based block bootstrap. So, τ_t is the tth index in the intersect index of observations and $\{\mathcal{B}_n(\boldsymbol{i}^*_1), ..., \mathcal{B}_n(\boldsymbol{i}^*_m)\}$, where $(\boldsymbol{i}^*_1, ..., \boldsymbol{i}^*_m)$ is a randomly generated m-run LHD. Therefore, for $(\lambda, \omega) \in \Lambda \times \Omega$, we have the tth bootstrap sample: $\boldsymbol{x}^*_t(\lambda, \omega) \equiv \boldsymbol{x}_{\tau_t(\lambda)}(\omega)$.

Suppose $\{Y(\boldsymbol{x}_t), t \in R\}$ follows a GP with mean μ. Given n observations, the sample estimation of mean μ is

$$\bar{y}_n = \frac{1}{n} \sum_{s=1}^{n} y_s,$$

and the LHD-based block bootstrap mean with N samples is given by

$$\bar{y}^*_N = \frac{1}{N} \sum_{s=1}^{N} y^*_s.$$

With a slight abuse of notation, we replace the notation of random variable Y by its realization y unless otherwise specified. The following lemma shows the asymptotic consistency of the LHD-based block bootstrap mean.

Lemma 1.6. *Under (A.1)-(A.2), if $m \to \infty$ and $m = o(n^{1/d})$, then*

$$\sup_x |P^*_{N,\omega}(\sqrt{n/m^{d-1}}(\bar{y}^*_N - \bar{y}_n)/\tau_n \le x) - P(\sqrt{n}(\bar{y}_n - \mu)/\tau_n \le x)| \xrightarrow{\text{P}} 0,$$

when $n \longrightarrow \infty$.

Note that $\boldsymbol{E}(\cdot)$ and $\boldsymbol{Cov}(\cdot,\cdot)$ denote the expectation and variance under P while $\boldsymbol{E}^*_{N,\omega}(\cdot)$ and $\boldsymbol{Cov}^*_{N,\omega}(\cdot,\cdot)$ denote the expectation and variance under $P^*_{N,\omega}$.

Proof of Lemma 1.6: It suffices to show that (1) $\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = \bar{y}_n$; (2) $n{\tau^*_N}^2/m^{d-1} - \tau^2_n \xrightarrow{\text{P}} 0$; and (3) $\sup_x |P^*_{N,\omega}((\bar{y}^*_N - \boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N))/\tau^*_N \le x) - \Phi(x)| \xrightarrow{\text{P}} 0$, where $\Phi(\cdot)$ denotes standard normal distribution function and ${\tau^*_N}^2 = \boldsymbol{Cov}^*_{N,\omega}(\bar{y}^*_N, \bar{y}^*_N)$.

Lemmas 1.1 and 1.3 imply the results in (1) and (2). Note that $\bar{y}^*_N = \frac{1}{m}\sum_{j=1}^m \bar{y}_{\boldsymbol{i}^*_j}$ and $(\bar{y}_{\boldsymbol{i}^*_1}, \dots, \bar{y}_{\boldsymbol{i}^*_m})$ follows Latin Hypercube sampling distribution. According to Loh (1996), we have the Berry-Essen type of bound for Latin Hypercube sampling

$$\sup_x |P^*_{N,\omega}((\bar{y}^*_N - \bar{y}_n)/\tau^*_N \le x) - \Phi(x)| \le c^* m^{-1/2},$$

where c^* is a constant that depends only on d, given $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}^*_1}\|^3 < \infty$. So we only need to show that $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}^*_1}\|^3$ is bounded uniformly in probability under P. Since $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}_1}\|^3 = \frac{1}{m^d}\sum_{\boldsymbol{i}} \bar{y}^3_{\boldsymbol{i}}$ and according to Minkowski's inequality, it follows that

$$\frac{1}{m^d}\sum_{\boldsymbol{i}} \boldsymbol{E}\{\bar{y}^3_{\boldsymbol{i}}\} \le \frac{1}{m^d}\sum_{\boldsymbol{i}} \frac{1}{|\mathcal{B}_n(\boldsymbol{i})|^3}\{\sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{E}(y_s)\}^3 < \infty.$$

□

1.8.4 Proof of Theorem 1.1

To investigate the asymptotic properties of the estimators from LHD-based block bootstrap, we decompose the likelihood function into blocks. For each block, denote $\boldsymbol{y_i} = (y_s(\boldsymbol{x}_s), \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}))$, $\boldsymbol{X_i} = (\boldsymbol{x}_s, \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}))^T$, $R_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{\theta})$=$[\psi(y(\boldsymbol{x}_s), y(\boldsymbol{x}_t); \boldsymbol{\theta})$, $\boldsymbol{x}_s \in$

$\mathcal{B}_n(\boldsymbol{i}), \boldsymbol{x}_t \in \mathcal{B}_n(\boldsymbol{j})]$ and $\boldsymbol{z_i} = R_{\boldsymbol{i},\boldsymbol{i}}^{-1/2}(\boldsymbol{\theta})(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})$. Then, we can rewrite the penalized log-likelihood function $n^{-1}\ell(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi})$ as

$$\begin{aligned} Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) = \;& -(2n\sigma^2)^{-1}\textstyle\sum_{s=1}^n z_s^2 - (2n)^{-1}\sum_{s=1}^n \log(\lambda_s) \\ & -(2n)^{-1}\textstyle\sum_{s=1}^n \log(\sigma^2) + n^{-1}r_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) \\ & -\textstyle\sum_{s=1}^p p_\lambda(|\beta_s|) \\ = \;& n^{-1}\textstyle\sum_{s=1}^n q_s(\omega, \boldsymbol{\phi}) + n^{-1}r_n(\omega, \boldsymbol{\phi}) - \sum_{s=1}^p p_\lambda(|\beta_s|) \end{aligned} \tag{1.5}$$

where $\{\lambda_s, s = 1, \ldots, n\} = \{\text{eigenvalues of } |R_{\boldsymbol{i},\boldsymbol{i}}(\boldsymbol{\theta})|, \boldsymbol{i} = (i_1, \ldots, i_d)\}$ with $(i_1, \ldots, i_d)$ in lexicographical order and eigenvalues from the largest to the smallest. Note that $r_n(\omega, \boldsymbol{\phi}) = \ell(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) - \sum_{s=1}^n q_s(z_s, \boldsymbol{\phi})$ contains all terms involving the off block-diagonal terms. Define $D_n(\boldsymbol{\theta}) = \text{diag}(R_{\boldsymbol{i},\boldsymbol{i}}(\boldsymbol{\theta}))$ and $E_n(\boldsymbol{\theta}) = R_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta})$. Assuming that $E_n(\boldsymbol{\theta}) = U_n(\boldsymbol{\theta})U_n^T(\boldsymbol{\theta})$, we have

$$\begin{aligned} r_n(\omega, \boldsymbol{\phi}) \;=\;& \frac{1}{2\sigma^2(1+g)}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T D_n^{-1}(\boldsymbol{\theta})E_n(\boldsymbol{\theta})D_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}) \\ & +\frac{1}{2}\log|I_n + U_n^T(\boldsymbol{\theta})D_n^{-1}(\boldsymbol{\theta})U_n(\boldsymbol{\theta})|, \end{aligned}$$

where $g = \text{trace}(E_n(\boldsymbol{\theta})D_n^{-1}(\boldsymbol{\theta}))$.

The MLE is obtained by $\hat{\boldsymbol{\phi}}_n = \arg\max_{\boldsymbol{\phi}} Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi})$. Analogue to the decomposition for $Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi})$, the log-likelihood function for LHD-based block bootstrap samples can be written as

$$\begin{aligned} Q_N^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi}) \;=\;& N^{-1}\sum_{s=1}^N q_s^*(\cdot, \omega, \boldsymbol{\phi}) + N^{-1}r_N^*(\cdot, \omega, \boldsymbol{\phi}) \\ & -\sum_{s=1}^p p_\lambda(|\beta_s|) \end{aligned} \tag{1.6}$$

where $r_N^*(\cdot,\omega,\boldsymbol{\phi})$ contains all terms involving the off block-diagonal terms with bootstrapped samples. Specifically,

$$\begin{aligned} & r_N^*(\cdot,\omega,\boldsymbol{\phi}) \\ = \quad & \frac{1}{2\sigma^2(1+g^*)}(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta})^T {D_N^*}^{-1}(\boldsymbol{\theta})E_N^*(\boldsymbol{\theta}){D_N^*}^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta}) \\ & +\frac{1}{2}\log|I_N + {U_N^*}^T(\boldsymbol{\theta}){D_N^*}^{-1}(\boldsymbol{\theta})U_N^*(\boldsymbol{\theta})|, \end{aligned}$$

where $D_N^*(\boldsymbol{\theta}) = \text{diag}(R_{\boldsymbol{i}_j^*,\boldsymbol{i}_j^*}(\boldsymbol{\theta}), j = 1,\ldots,m)$ and $E_N^*(\boldsymbol{\theta}) = R_N^*(\boldsymbol{\theta}) - D_N^*(\boldsymbol{\theta})$ with $E_N^*(\boldsymbol{\theta}) = U_N^*(\boldsymbol{\theta}){U_N^*}^T(\boldsymbol{\theta})$; $g^* = \text{trace}(E_N^*(\boldsymbol{\theta}){D_N^*}^{-1}(\boldsymbol{\theta}))$. The bootstrapped version of $\hat{\boldsymbol{\phi}}_n$ is $\hat{\boldsymbol{\phi}}_N^* = \arg\max_{\boldsymbol{\phi}} Q_N^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi})$. Theoretical properties of the LHD-based block bootstrap likelihood function (1.6) are established in lemmas 4 and 5, which leads to a proof of convergence properties of the bootstrap estimator $\hat{\boldsymbol{\phi}}_N^*$. Lemma 1.4 first established the pointwise weak law of large numbers for the LHD-based block bootstrap likelihood functions. Lemma 1.5 further extends Lemma 1.4 to the uniform weak law of large numbers for the LHD-based block bootstrap likelihood functions.

Proof of Theorem 1.1: Based on Lemma 5, we have

$$\lim_{n\to\infty} P[P_{N,w}^*(\sup_{\boldsymbol{\phi}\in\Theta}|Q_n - Q_N^*| > \delta) > \xi] = 0,$$

where Q_n and Q_N^* are given in (1.5) and (1.6). With the full preparation of the likelihood convergence developed in Lemmas 1.4 and 1.5, the convergence of bootstrap parameter estimation follows immediately given the existence of $\hat{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\phi}}_N^*$.

Denote $\bar{q}_N^*(\cdot,\omega,\boldsymbol{\phi}) = N^{-1}\sum_{i=1}^N q_i^*(\cdot,\omega,\boldsymbol{\phi})$ and $\bar{q}_n(\omega,\boldsymbol{\phi}) = n^{-1}\sum_{i=1}^n q_i(\omega,\boldsymbol{\phi})$. By (A.6), $q_s^*(\cdot,\omega,\cdot) : \Lambda\times\Theta \to R$ and $r_N^*(\cdot,\omega,\cdot) : \Lambda\times\Theta\to R$ are measurable-$\mathcal{G}$ for each $\boldsymbol{\phi}\in\Theta$. In addition, $q_s^*(\lambda,\omega,\cdot)$ and $r_N^*(\lambda,\omega,\cdot)$ are continuous on Θ for all λ. Thus, we have $\hat{\boldsymbol{\phi}}_N^*(\cdot,\omega)$ exists as a measurable-$\mathcal{G}$ function by Jennrich (1969).

Following the procedure in Goncalves and White (2004), for any subsequence $\{n'\}$, given that $\hat{\boldsymbol{\phi}}_{n'}$ is identifiable and unique, there exists a further subsequence $\{n''\}$ such that $\hat{\boldsymbol{\phi}}_{n''}$ is identifiably unique with respect to $\{Q_{n''}\}$ for all $\omega \in F$ in some $F \in \mathcal{F}$ with $P(F) = 1$. By condition (A.6), there exists $G \in \mathcal{F}$ with $P(G) = 1$ such that for all $\omega \in G$, $\{Q^*_{N''}(\cdot, \omega, \boldsymbol{\phi})\}$ (N'' is corresponding bootstrapped sample size of n'') is a sequence of random function on $(\Lambda, \mathcal{G}, P^*_{N,\omega})$ continuous on Θ for all $\lambda \in \Lambda$. Hence, by White (1996), for fixed $\omega \in G$, there exists $\hat{\boldsymbol{\phi}}^*_{N''}(\cdot, \omega) : \Lambda \to \Theta$ measurable-$\mathcal{G}$ and $\hat{\boldsymbol{\phi}}^*_{N''}(\cdot, \omega) = \arg\max_{\boldsymbol{\phi}} Q^*_{N''}(\cdot, \omega, \boldsymbol{\phi})$. By the uniform weak law of large numbers for $Q^*_N(\boldsymbol{X}^*_N, \boldsymbol{y}^*_N, \boldsymbol{\phi})$ obtained from Lemma 1.5, we have $Q^*_{N''}(\cdot, \omega, \boldsymbol{\phi}) - Q_{n''}(\omega, \boldsymbol{\phi}) \to 0$ as $n'' \to \infty$ $prob - P^*_{N,\omega}$ $prob - P$ uniformly on Θ, where we write $\hat{Q}^*_N \to 0$ $prob - P^*_{N,\omega}, prob - P$ if, for any $\epsilon > 0$ and $\delta > 0$, $\lim_{n \to \infty} P\{P^*_{N,\omega}(|\hat{Q}^*_N > \epsilon| > \delta)\} = 0$ and omit $prob - P^*_{N,\omega}, prob - P$ in the text for notation simplicity. Hence, there exists a further subsequence $\{n'''\}$ such that $Q^*_{N'''}(\cdot, \omega, \boldsymbol{\phi}) - Q_{n'''}(\omega, \boldsymbol{\phi}) \to 0$ as $n'' \to \infty$ $prob - P^*_{N,\omega}$ $prob - P$ for all ω in some $H \in \mathcal{F}$ with $P(H) = 1$. Choose $\omega \in F \cap G \cap H$, by White (1996), we have $\hat{\boldsymbol{\phi}}^*_{N'''} - \hat{\boldsymbol{\phi}}_{n'''} \to 0$ as $n''' \to \infty$ $prob - P^*_{N,\omega}$ $prob - P$. Since this is true for any subsequence $\{n'\}$, we have $P(F \cap G \cap H) = 1$. Thus, $\hat{\boldsymbol{\phi}}^*_N - \hat{\boldsymbol{\phi}}_n \to 0$ $prob - P^*_{N,\omega}, prob - P$. Then $\hat{\boldsymbol{\phi}}_N = \frac{1}{K}\sum_{i=1}^K \hat{\boldsymbol{\phi}}^*_N(i) - \hat{\boldsymbol{\phi}}_n \to 0$ $prob - P^*_{N,\omega}, prob - P$.

□

1.8.5 Proof of Theorem 1.2

Proof. Define $B = Var\{n^{-1/2}\sum_{s=1}^n \nabla q_s(\cdot, \omega, \boldsymbol{\phi}_0)\}$. We first show that

$\sqrt{n/m^{d-1}}B^{-1/2}\nabla Q^*_N(\cdot, \omega, \hat{\boldsymbol{\phi}}_n) \to N(0, I)$. Denote $\bar{h}^*_N(\boldsymbol{\phi}) = N^{-1}\sum_{s=1}^N \nabla q^*_s(z^*_s, \boldsymbol{\phi})$ and

$\bar{h}_n(\boldsymbol{\phi}) = n^{-1}\sum_{s=1}^{n} \nabla q_s(z_s, \boldsymbol{\phi})$. We have

$$\begin{aligned} \sqrt{n/m^{d-1}}[\bar{h}_N^*(\hat{\boldsymbol{\phi}}_n) - \bar{h}_n(\hat{\boldsymbol{\phi}}_n)] &= \sqrt{n/m^{d-1}}[\bar{h}_N^*(\hat{\boldsymbol{\phi}}_n) - \bar{h}_N^*(\boldsymbol{\phi}^0)] \\ &\quad + \sqrt{n/m^{d-1}}[\bar{h}_N^*(\boldsymbol{\phi}^0) - \bar{h}_n(\boldsymbol{\phi}^0)] \\ &\quad + \sqrt{n/m^{d-1}}[\bar{h}_n(\boldsymbol{\phi}^0) - \bar{h}_n(\hat{\boldsymbol{\phi}}_n)] \\ &= J_1 + J_2 + J_3. \end{aligned}$$

Since $\bar{h}_n$ and $\bar{h}_N^*$ are functions whose secondary derivative are continuous, $J_1 + J_3 \to 0$ as $\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0 \to 0$ by Theorem 3.1 in Chu (2011). Moreover, the two terms in J_2 are both evaluated at $\boldsymbol{\phi}_0$ which is a fixed value, then by Lemma 6, we have $B^{-1/2}J_2 \to N(0, I)$. By condition (A.10) and follow a similar proof as Lemma 1.5, we have

$$\nabla^2 Q_N^*(\cdot, \omega, \boldsymbol{\phi}) - \nabla^2 Q_n(\omega, \boldsymbol{\phi}) \to 0 \;\; prob - P_{N,\omega}^*, prob - P.$$

Let $\hat{H}_n(\omega) = \nabla^2 Q_n(\omega, \hat{\boldsymbol{\phi}}_n)$. According to White (1996), given the result $\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n \to 0$ $prob - P_{N,\omega}^*, prob - P$ and assumption (A.8), we have

$$\begin{aligned} \sqrt{N}(\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n) &= -\hat{H}_n^{-1}(\omega)\sqrt{N}\nabla Q_N^*(\cdot, \omega, \hat{\boldsymbol{\phi}}_n) + o_{P_{N,\omega}^*}(1) \\ &= -H_n(\boldsymbol{\phi}_0)^{-1}(\omega)\sqrt{N}\nabla Q_N^*(\cdot\omega, \hat{\boldsymbol{\phi}}_n) + o_{P_{N,\omega}^*}(1). \end{aligned}$$

Given the fact that

$$\sqrt{n/m^{d-1}}B^{-1/2}\nabla Q_N^*(\cdot, \omega, \hat{\boldsymbol{\phi}}_n) \to N(0, I) \;\; prob - P_{N,\omega}^*, prob - P.$$

we have

$$B^{-1/2}H_n(\boldsymbol{\phi}_0)\sqrt{N}(\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n) \to N(0, I).$$

For $\boldsymbol{\beta}_{10}$, B and H can be written as $\mathbf{J}(\boldsymbol{\beta}_{10})$ and $\mathbf{J}(\boldsymbol{\beta}_{10}) + \mathbf{G}(\boldsymbol{\beta}_{10})$. For $\hat{\boldsymbol{\beta}}_{N,1}^*$, we have

$$\sqrt{N}[\mathbf{J}(\boldsymbol{\beta}_{10}) + \mathbf{G}(\boldsymbol{\beta}_{10})]\{\hat{\boldsymbol{\beta}}_{N,1}^* - \hat{\boldsymbol{\beta}}_{n,1}\} \to N(0, J(\boldsymbol{\beta}_{10})).$$

For sub-bagging estimator $\hat{\boldsymbol{\beta}}_{N,1} = \sum_{i=1}^{K} \hat{\boldsymbol{\beta}}^{*}_{N,1}(i)$, we have

$$\sqrt{KN}[\mathbf{J}(\boldsymbol{\beta}_{10}) + \mathbf{G}(\boldsymbol{\beta}_{10})]\{\hat{\boldsymbol{\beta}}_{N,1} - \hat{\boldsymbol{\beta}}_{n,1}\} \rightarrow N(0, J(\boldsymbol{\beta}_{10})),$$

then the result follows.

1.8.6 Proof of Theorem 1.3

Using the same technique before, we decompose the log-likelihood by blocks and rewrite the likelihood of $\boldsymbol{\beta}$ based on the OSE approach as follows:

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &= n^{-1}\sum_{s=1}^{n} q_s(\omega, \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n^{(0)}, \hat{\sigma_n^2}^{(0)}) + n^{-1} r_n(\omega, \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_n^{(0)} \hat{\sigma_n^2}^{(0)}) \\ & -\sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j|. \end{aligned}$$

The likelihood based on subsampled data can be written as:

$$\begin{aligned} Q^*_N(\boldsymbol{\beta}) &= N^{-1}\sum_{s=1}^{N} q^*_s(\omega, \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_N^{*(0)}, \hat{\sigma_N^2}^{*(0)}) + N^{-1} r^*_N(\omega, \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_N^{*(0)}, \hat{\sigma_N^2}^{*(0)}) \\ & -\sum_{j=1}^{p} p'_\lambda(\hat{\beta}_j^{*(0)}|)|\beta_j|. \end{aligned}$$

By the fact that $\hat{\boldsymbol{\phi}}^{*}_N - \hat{\boldsymbol{\phi}}_n \rightarrow 0$ and the results in Lemma 2, Lemma 3 and Lemma 6 still hold, we have $\hat{\boldsymbol{\phi}}^{*}_{N,OSE} - \hat{\boldsymbol{\phi}}_{n,OSE} \rightarrow 0$. Then follows the same technique in the proof of Theorem 2, the result follows. □

Chapter 2

Markov Switching Autoregressive Models for the Analysis of Cell Adhesion

2.1 Introduction

This research is motivated by the statistical analysis of cell adhesion experiments, which are biomechanical experiments that study protein interactions at the level of single molecules (Mehta et al. 1999). Cell adhesion plays an important role in many physiological and pathological processes. In this research, we focus on an important type of cell adhesion experiment called force-clamp assay (Marshall et al., 2003). The goal is to understand the cell adhesion mechanism through the study of TCR-pMHC bond lifetime, which is crucial in triggering T cell signaling.

A force-clamp assay is conducted using a biomembrane force probe (BFP) (Chen et al. 2008) illustrated in Figure 2.1. The BFP uses a micropipette-aspirated human red blood cell (RBC) with a probe bead attached to its apex as a force transducer (Figure 2.1A, left). The RBC was aligned against a T cell held by an apposing pipette (right). The probe bead was coated with pMHC (Figure 2.1B, left) to interact with T cell receptor (TCR) (Figure 2.1B, right). TCR-pMHC bond lifetimes were measured by the force-clamp assay in repetitive cycles. In each cycle, the T cell (Figure 2.1A, right) was driven to contact the probe bead to prompt bond formation. Contact was brief (0.1 s) to minimize multibond formation. Via T cell retraction, a tensile force on

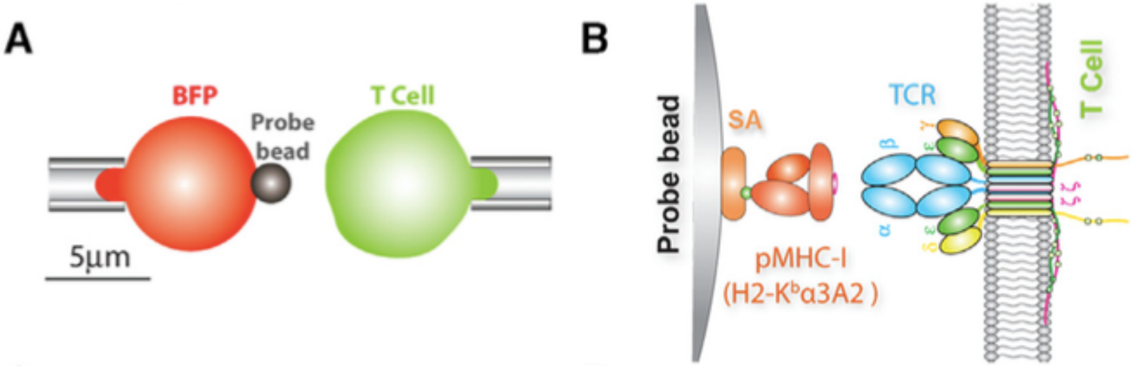

Figure 2.1: BFP schematic and functionalization

the TCR-pMHC bond was ramped (at 1,000 pN/s) to and clamped at a preset level until bond dissociation. Bond lifetime was measured as the force-clamp period (marked by red in Figure 2.2).

To build a statistical model for the analysis of repeated bond lifetime measurements, we need to take into account three unique features in force-clamp assays. First, there are multiple repeated assays collected from different pairs of cells. The same biological mechanism is shared within the same pair of cells and some variations exist among different pairs of cells. This is similar to the longitudinal studies where correlations among the repeated measurements arise from some shared unobserved variables within the same subject. Second, the molecular bond formation, which is of our major interest, is not directly observable. The presence of a molecular bond is detected indirectly through the measurements of bond lifetime because the interaction force will resist surface separation until the bond ruptures. Governed by the binding status, the bond lifetime measurements can be assumed to be random variables following different distributions because molecular dissociation occurs as diffusive escape from an energy well (bound state) by thermally agitated Brownian motion. In general, when the underlying binding occurs, the bond lifetime measurements are higher than the situations without

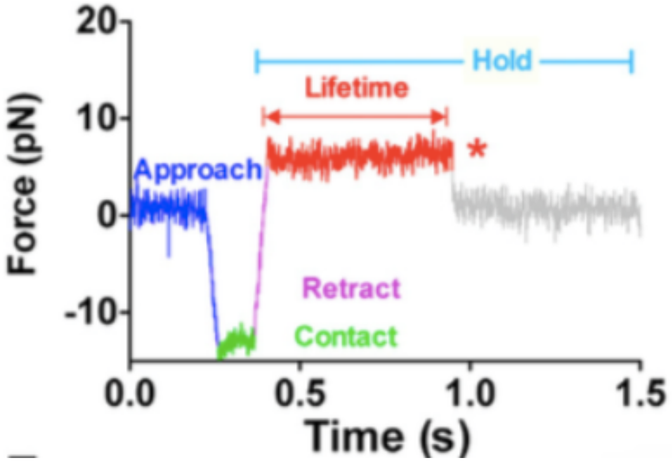

Figure 2.2: BFP schematic and functionalization

binding. This is because the bond formation is equivalent to adding a molecular spring in parallel to the force transducer spring to stiffen the system. Therefore, the force-clamp period (i.e., defined as bond lifetime) is longer in order to separate the cells as well as the receptor-ligand bond. Third, there are some memory effects in the repeated bond lifetime measurements. It was discovered that cells appear to have the ability to remember the previous adhesion events and such a memory has an impact on the future adhesion behaviors. Zarnitsyna et al. (2007) and Hung et al. (2008) demonstrated that in some biological systems the occurrence of binding in the immediate past assay could either increase or decrease the likelihood for the next assay to result in a binding. In the repeated force-clamp assays, such a memory effect can affect the binding frequency as well as the bond lifetime measurements. Quantification of the memory effects is biologically important and it is the focus of this study.

We introduce a new statistical framework within which the unique features are incorporated and the molecular binding mechanism can be studied. This framework is based upon an extension of Markov switching autoregressive models (MSAR), a regime-switching type of time series model generalized from hidden Markov models. MSAR has been extensively studied and proven to be useful in various applications involved time series, including econometrics, and speech recognition (references Ben's paper). However, standard MSAR models are developed for the analysis of individual stochastic process, which is not sufficient for simultaneously modeling multiple time series processes collected from different experimental subjects as in the longitudinal data setting. To handle multiple time series processes, we introduce Markov switching autoregressive mixed (MSARM) models which borrow strength across different subjects by incorporating random effects into the model. The MSARM model uses hidden states to represent the unobservable binding status, binding or no binding. The unobservable states are not assumed to be independent, but rather to have a Markovian structure so that the cell memory effects can be captured. Given the hidden states, rupture forces are assumed to be normally distributed with different autoregressive mean structures that capture the stiffness and memory effects associated with different binding status.

More than a simple extension, the MSARM model posts statistical challenges in the theoretical developments as well as computational efficiency in high-dimensional integration. Theoretical studies are limited to HMMs in which single stochastic process is considered (Bickel et al. 1998). Extensions from HMM to multiple processes in a longitudinal setting are discussed by Altman (2007), but theoretical properties for the proposed models are not addressed. To the best of our knowledge, there is no theoretical study available for multiple stochastic processes with autoregressive and regime-switching structure. Theoretical generalization from HMM is not straightforward because of the multiple stochastic processes, random effects, and a more flexible but complex correlation structure resulting from the autoregressive model. On the other hand, estimation of MSARM is computationally intensive because of the high-dimensional integration involved in the random effects. An MCEM algorithm is introduced by Altman (2007) for the estimation of hidden Markov mixed model. However, the algorithm is still of concern when the number of random effects increases. To overcome this difficulty, we introduce a new Gibbs sampling scheme so that the M-step has a closed form and therefore it is easy to compute. We also proved the asymptotic normality of the MLE based on our model. The proof illustrates how to analyze the likelihood based on multiple stochastic processes with random effects.

The advantages of MSARM are numerous and the application is beyond cell biology.

First, modeling multiple processes simultaneously permits the estimation of population-level effects, as well as more efficient estimation of parameters that are common to all processes. Second, these models are relatively easy to interpret. Finally, MHMMs permit greater flexibility in modeling correlation structure because they relax the assumption that the observations are independent given the hidden states.

2.2 Markov Switching Autoregressive Mixed Models

2.2.1 Model Description

Let y_{it} be the tth observation collected from the ith experimental units/cells and x_{it} be the corresponding hidden state, where $i = 1, ..., m$ and $t = 1, ..., n_i$. Denote the total number of observations by $N = \sum_{i=1}^{m} n_i$. Assume there are K hidden states, where K is known, and the change of states can be described by a stationary Markov chain with transition probability $p_{kl} = P(x_{t+1} = l | x_t = k)$ and stationary probability π_k, where $k, l \in \{1, ..., K\}$. A Markov switching autoregressive mixed (MSARM) model can be written as:

$$y_{it}|(x_{it} = k) = \alpha_k + \sum_{j=1}^{r_k} y_{i,t-j}\beta_j^{(k)} + u_i + e_{it}^{(k)}, \tag{2.1}$$

where $k \in \{1, ..., K\}$ and u_i follows $N(0, \sigma_u^2)$. For hidden state k, the autoregressive structure with order r_k is captured by $\sum_{j=1}^{r_k} y_{i,t-j}\beta_j^{(k)}$ with unknown coefficients $\beta_j^{(k)}$. The random errors are assumed to be mutually independent and their variance are assumed to be different according to their hidden states, i.e., $e_{it}^{(k)}$ follows $N(0, \sigma_k^2)$. Let

θ denote the parameters, then the likelihood is

$$\begin{aligned} & f(\theta;\mathbf{y}) \\ =& \int_{\mathbf{u}}\sum_{\mathbf{x}} f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta)f(\mathbf{x};\theta)f(\mathbf{u};\theta)d\mathbf{u} \\ =& \int_{\mathbf{u}}\sum_{\mathbf{x}}\{\prod_{i=1}^{m}\prod_{t=1}^{n_i} f(y_{it}|x_{it},\mathbf{u},\theta,y_{i,t-1})\}\{\prod_{i=1}^{m}\pi_{x_{i1}}\prod_{t=2}^{n_i} p_{x_{i,t-1},x_{it}}\}f(\mathbf{u};\theta)d\mathbf{u} \qquad (2.2) \end{aligned}$$

The random effects u_i are assumed to capture the variations among the multiple processes, which is common in longitudinal settings. In practice, model (2.1) can be easily extended to incorporate more than one random effects.

This model includes some existing models as special cases. When $m = 1$, this model leads to a conventional Markov switching model. By incorporating random effects, this model can capture the unobserved heterogeneity among the processes. When the autoregressive terms are removed, i.e., $\beta_j^{(k)} = 0$, it leads to one of the hidden mixed Markov models introduced by Altman (2007).

2.2.2 Theoretical Properties

MSARMs are extensions of HMMs. Although asymptotic properties of the MLEs are extensively developed for HMMs (references), generalization to MSARMs are not straightforward due to three reasons. First, the likelihood function involved integration of random effects and therefore is more complicated. Second, the correlation structure is more complex because the response y_t depends on both its hidden state x_t and some previous responses $y_{t-r}, ..., y_{t-1}$. Third, the proposed model involves multi-stochastic processes, whereas the HMMs focus on the setting with single stochastic process.

For notation simplicity, we assume $n_i = n$ for all $i = 1, ..., m$ and rewrite the model in vector form as follows

$$\mathbf{y}|(\mathbf{k}) = U_0\gamma_{\mathbf{k}} + U\mathbf{u} + \mathbf{e}_{\mathbf{k}}, \tag{2.3}$$

where $\gamma_k = (\alpha_k, \beta_1^{(k)}, ..., \beta_r^{(k)})'$ with the corresponding historical data denoted by U_0, U is an $n \times m$ matrix of known constants(design matrix for a random effect), it consists only of zeros and ones and there is exactly one 1 in each row and at least one 1 in each column. u is an $m \times 1$ random vector. Let $G = UU'$, and $G_0 = \mathbf{I}_N$. The random vectors $\mathbf{u}$ and $\mathbf{e}_k$ are independent, with $\mathbf{e}_k \sim N(0, \sigma_k^2\mathbf{I}_N)$ and $\mathbf{u}_i \sim N(0, \sigma_u^2\mathbf{I}_m)$. Matrix U_0 has full rank $(1+r)$. Based on (3.1), it follows that $\mathbf{y}|(\mathbf{k}) \sim N(U_0\gamma_{\mathbf{k}}, \Sigma_{\mathbf{k}})$, where $\Sigma_{\mathbf{k}} = G\sigma_u^2 + \sigma_{\mathbf{k}}^2\mathbf{I}$. Thus, based on the vector form of the model, the Log-likelihood can be written as

$$L(\theta) = \sum_{\mathbf{x}}[-\frac{1}{2}N\log 2\pi - \frac{1}{2}\log|\Sigma_{\mathbf{x}}| - \frac{1}{2}(\mathbf{y} - U_0\gamma_{\mathbf{x}})'\Sigma_{\mathbf{x}}^{-1}(\mathbf{y} - U_0\gamma_{\mathbf{x}})]P(\mathbf{x}) \tag{2.4}$$

where the $P(\mathbf{x}) = \prod_{i=1}^m \pi_{x_{i1}} \prod_{t=2}^{n_i} p_{x_{i,t-1},x_{it}}$, is the density of any sequence of the hidden states. There are two sets of parameters including $\theta_1 = (p_{kl}, \gamma_k, \sigma_k)$ and $\theta_2 = (\sigma_{ui})$, where $k, l = 1, ..., K$, and $i = 1, ..., p_1$ The true parameters are denoted by $\theta_0 = (\theta_{10}, \theta_{20})$. Then we have the main results as follows:

Theorem 2.1. *Assume that (A1)-(A7) hold,*

$$\begin{pmatrix} \sqrt{N}[\hat{\theta}_1 - \theta_{10}] \\ \\ \sqrt{m}[\hat{\theta}_2 - \theta_{20}] \end{pmatrix} \to N(0, I_0^{-1}),$$

where $\boldsymbol{I}_{m,n}$ is the Fisher information matrix

$$\boldsymbol{I}_{m,n} = \begin{pmatrix} Cov(N^{-\frac{1}{2}}\frac{\partial L}{\partial\theta_1}) & Cov(N^{-\frac{1}{2}}\frac{\partial L}{\partial\theta_1}, m^{-\frac{1}{2}}\frac{\partial L}{\partial\theta_2}) \\ Cov(N^{-\frac{1}{2}}\frac{\partial L}{\partial\theta_1}, m^{-\frac{1}{2}}\frac{\partial L}{\partial\theta_2}) & Cov(m^{-\frac{1}{2}}\frac{\partial L}{\partial\theta_2}) \end{pmatrix}$$

and

$$\lim_{m,n\to\infty} \boldsymbol{I}_{m,n} = \begin{pmatrix} I_{11} & I_{12} \\ \\ I_{21} & I_{22} \end{pmatrix} = I_0.$$

This theorem shows the asymptotic properties of the parameters. For the estimator of the parameters except the variance component ($\hat{\theta}_1$), the convergence rate is $\sqrt{N}$. For the estimator of the parameters for the variance component ($\hat{\theta}_2$)), the convergence rate is $\sqrt{m}$.

2.3 Monte Carlo Algorithm

In order to illustrate our method better and avoid complexity in notation, we assume that there is only 1 random effect in the model. When the number of random effect is greater than 1, the corresponding algorithm can be derived without any difficulties.

In the case where there are a few random effects, the numerical method works well. Here Gaussian quadrature methods approximates the integral well and Quasi-Newton method can be used to maximize the approximated likelihood. When the number of random effects become large, numerical results are not feasible. Here we use the traditional EM algorithm to do estimation. The complete likelihood is

$$\begin{aligned}
&L(\theta; \mathbf{y}, \mathbf{x}, \mathbf{u}) \\
=& \log f(\mathbf{y}|\mathbf{x}, \mathbf{u}, \theta) + \log f(\mathbf{x}|\mathbf{u}, \theta) + \log f(\mathbf{u}; \theta) \\
=& \sum_{i=1}^{m} \sum_{t=1}^{n_i} \log f(y_{it}|x_{it}, \mathbf{u}, \theta, y_{i,t-1}) + \sum_{i=1}^{m} \log \pi_{z_{i1}} + \sum_{i=1}^{m} \sum_{t=2}^{n_i} \log p_{x_{i,t-1}, x_{it}} + \log f(\mathbf{u}; \theta)
\end{aligned} \tag{2.5}$$

2.3.1 MCEM proposed by Altman

In this part, we introduced the method proposed by Altman(2007) E-step:

$$\begin{aligned}
&E[L(\theta; \mathbf{y}, \mathbf{x}, \mathbf{u})|\mathbf{y}, \theta^p] \\
=& \int_{\mathbf{u}} \sum_{\mathbf{x}} L(\theta; \mathbf{y}, \mathbf{x}, \mathbf{u}) f(\mathbf{x}, \mathbf{u}|\mathbf{y}, \theta^p) d\mathbf{u}
\end{aligned} \tag{2.6}$$

Note that:

$$\begin{aligned}
&f(\mathbf{x},\mathbf{u}|\mathbf{y},\theta^p)\\
=&\frac{f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta^p)f(\mathbf{x}|\mathbf{u},\theta^p)f(\mathbf{u};\theta^p)}{f(\mathbf{y};\theta^p)}\\
=&\frac{f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta^p)f(\mathbf{x}|\mathbf{u},\theta^p)f(\mathbf{u};\theta^p)}{\int_{\mathbf{u}}\sum_{\mathbf{x}} f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta^p)f(\mathbf{x}|\mathbf{u},\theta^p)f(\mathbf{u};\theta^p)d\mathbf{u}} \qquad (2.7)
\end{aligned}$$

Therefore, if one generates samples $\mathbf{u}^1, ..., \mathbf{u}^B$ from $f(\mathbf{u}|\theta^p)$, then we can obtain the approximation:

$$\begin{aligned}
&E[L(\theta;\mathbf{y},\mathbf{x},\mathbf{u})|\mathbf{y},\theta^p]\\
\approx&\frac{1}{B}\sum_{j=1}^{B}\sum_{\mathbf{x}} L(\theta;\mathbf{y},\mathbf{x},\mathbf{u}^j)h_j(\mathbf{x}) \qquad (2.8)
\end{aligned}$$

where

$$h_j(\mathbf{x}) = \frac{f(\mathbf{y}|\mathbf{x},\mathbf{u}^j,\theta^p)f(\mathbf{x}|\mathbf{u}^j,\theta^p)}{\sum_{k=1}^{B}\sum_{\mathbf{x}} f(\mathbf{y}|\mathbf{x},\mathbf{u}^k,\theta^p)f(\mathbf{x}|\mathbf{u}^k,\theta^p)}. \qquad (2.9)$$

This sampling step is intuitive, but the term $h_j(\mathbf{x})$ is computationally intensive to calculate. This is because $\sum_{\mathbf{x}} f(\mathbf{y}|\mathbf{x},\mathbf{u}^k,\theta^p)$ requires evaluation over all combinations of $\mathbf{x}$. For the M-step, the parameters involved in the likelihood can are estimated separately.

2.3.2 New Monte Carlo EM algorithm

Here we propose a new sampling method. In the E-step, instead of sampling $\mathbf{u}$ from $f(\mathbf{u}|\theta^p)$, we propose $f(\mathbf{u}|\mathbf{y},\theta^p)$, which can alleviate the computing in the E-step. Moreover, since we have a particular form of the likelihood, we have the explicit form of the updates in M-step, which reduce the calculation and illustrate the underlying mechanism.

E-step:

$$
\begin{aligned}
& E[L(\theta;\mathbf{y},\mathbf{x},\mathbf{u})|\mathbf{y},\theta^p] \\
&= \int_{\mathbf{u}} \sum_{\mathbf{x}} L(\theta;\mathbf{y},\mathbf{x},\mathbf{u}) f(\mathbf{x},\mathbf{u}|\mathbf{y},\theta^p) d\mathbf{u} \\
&= \int_{\mathbf{u}} \sum_{\mathbf{x}} L(\theta;\mathbf{y},\mathbf{x},\mathbf{u}) f(\mathbf{x}|\mathbf{u},\mathbf{y},\theta^p) f(\mathbf{u}|\mathbf{y},\theta^p) d\mathbf{u} \qquad (2.10)
\end{aligned}
$$

Sample $\mathbf{u}^1, ..., \mathbf{u}^B$ from $f(\mathbf{u}|\mathbf{y},\theta^p)$, then we can obtain the approximation:

$$E[L(\theta;\mathbf{y},\mathbf{x},\mathbf{u})|\mathbf{y},\theta^p] \approx \frac{1}{B}\sum_{j=1}^{B} E[L(\theta;\mathbf{y},\mathbf{x},\mathbf{u}^j)|\mathbf{y},\theta^p] \qquad (2.11)$$

To sample $\mathbf{u}^1, ..., \mathbf{u}^B$ from $f(\mathbf{u}|\mathbf{y},\theta^p)$, write down the conditional density function as follows:

$$
\begin{aligned}
f(\mathbf{u}|\mathbf{y},\theta^p) =& \frac{f(\mathbf{y}|\mathbf{u},\theta^p) f(\mathbf{u}|\theta^p)}{f(\mathbf{y}|\theta^p)} \\
=& \frac{\sum_{\mathbf{x}} f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta^p) f(\mathbf{x}|\mathbf{u},\theta^p) f(\mathbf{u}|\theta^p)}{f(\mathbf{y}|\theta^p)} \\
=& \frac{\sum_{\mathbf{x}} f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta^p) f(\mathbf{u}|\theta^p) f(\mathbf{x}|\theta^p)}{f(\mathbf{y}|\theta^p)} \\
\sim& \sum_{\mathbf{x}} f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta^p) f(\mathbf{u}|\theta^p) f(\mathbf{x}|\theta^p) \qquad (2.12)
\end{aligned}
$$

The third equation is because the random effect and the hidden state are independent, $f(\mathbf{x}|\mathbf{u},\theta^p) = f(\mathbf{x}|\theta^p)$.

Step 1:Sample $\mathbf{x}$ from $f(\mathbf{x}|\theta^p)$, then sample $\mathbf{u}$ from

$$
\begin{aligned}
\tilde{f}(\mathbf{u}) \sim f(\mathbf{y}|\mathbf{x},\mathbf{u},\theta^p) f(\mathbf{u}|\theta^p) \sim& \exp\{-\sum_{i=1}^{m}\sum_{t=1}^{n_i} \frac{(y_{it} - \beta_{x_{it}} y_{i,t-1} - \alpha_{x_{it}} - u_i)^2}{2\sigma^2_{x_{it}}}\} \exp\{-\sum_{i=1}^{m} \frac{u_i^2}{2\sigma_u^2}\} \\
\sim& N(diag(\mu_i), diag(\sigma_i^2)) \qquad (2.13)
\end{aligned}
$$

where $\mu_i = \frac{\sum_{t=1}^{n_i} \frac{y_{it} - \beta_{x_{it}} y_{i,t-1} - \alpha_{x_{it}}}{\sigma^2_{x_{it}}}}{\sum_{t=1}^{n_i} 1/\sigma^2_{x_{it}} + 1/\sigma^2_u}$, and $\sigma_i^2 = (\sum_{t=1}^{n_i} 1/\sigma^2_{x_{it}} + 1/\sigma^2_u)^{-1}$

Step 2: Repeat step 1 B times, then we got samples $\mathbf{u}^1, \mathbf{u}^2, ..., \mathbf{u}^B$ from $f(\mathbf{u}|\mathbf{y},\theta^p)$.

Then, E-step is as follows:

$$
\begin{aligned}
&E[L(\theta;\mathbf{y},\mathbf{x},\mathbf{u})|\mathbf{y},\theta^p]\\
=&\int_{\mathbf{u}}\sum_{\mathbf{x}}L(\theta;\mathbf{y},\mathbf{x},\mathbf{u})f(\mathbf{x}|\mathbf{u},\mathbf{y},\theta^p)f(\mathbf{u}|\mathbf{y},\theta^p)d\mathbf{u}\\
=&\frac{1}{B}\sum_{j=1}^{B}\sum_{\mathbf{x}}L(\theta;\mathbf{y},\mathbf{x},\mathbf{u})f(\mathbf{x}|\mathbf{u}^j,\mathbf{y},\theta^p)
\end{aligned}
\tag{2.14}
$$

Where $f(\mathbf{x}|\mathbf{u}^j,\mathbf{y},\theta^p)$ can be calculated by forward and backward algorithm.

M-step:

For the M-step, note that, typically, the parameters can be updated separately. In other words, the expressions $E[\sum_{i=1}^m\sum_{t=1}^{n_i}\log f(y_{it}|x_{it},\mathbf{u},\theta,y_{i,t-1})|\mathbf{y},\theta^p]$, $E[\sum_{i=1}^m\log\pi_{z_{i1}}|\mathbf{y},\theta^p]$, $E[\sum_{i=1}^m\sum_{t=2}^{n_i}\log p_{x_{i,t-1},x_{it}}|\mathbf{y},\theta^p]$, and $E[\log f(\mathbf{u};\theta)|\mathbf{y},\theta^p]$ can usually be maximized separately, improving the efficiency of the procedure.

$$
\begin{aligned}
&E[\sum_{i=1}^{m}\sum_{t=1}^{n_i}\log f(y_{it}|x_{it},\mathbf{u},\theta,y_{i,t-1})|\mathbf{y},\theta^p]\\
=&\frac{1}{B}\sum_{j=1}^{B}\sum_{\mathbf{x}}\sum_{i=1}^{m}\sum_{t=1}^{n_i}\log f(y_{it}|x_{it},\mathbf{u}^j,\theta,y_{i,t-1})f(\mathbf{x}|\mathbf{y},\mathbf{u}^j,\theta^p)\\
=&\frac{1}{B}\sum_{j=1}^{B}\sum_{\mathbf{x}}\sum_{i=1}^{m}\sum_{t=1}^{n_i}[-\log(\sigma_{x_{it}})-\frac{(y_{it}-\beta_{x_{it}}y_{i,t-1}-\alpha_{x_{it}}-u_{it}^j)^2}{2\sigma_{x_{it}}^2}]f(\mathbf{x}|\mathbf{y},\mathbf{u}^j,\theta^p)
\end{aligned}
\tag{2.15}
$$

Let $q_{it}(k)=\frac{p(x_{it}=k|\mathbf{y},\theta^p,\mathbf{u}^j)}{\sum_{i=1}^m\sum_{t=1}^{n_i}p(x_{it}=k|\mathbf{y},\theta^p,\mathbf{u}^j)}$. The above function can be viewed as the weighted Least Square likelihood with response $y_{it}-u_{it}$ and covariates $y_{i,t-1}$. The corresponding weight is q_{it}. Thus the updates for $\beta's$ and $\alpha's$ can be easily derived.

$$
E[\sum_{i=1}^{m}\log\pi_{z_{i1}}|\mathbf{y},\theta^p]=\frac{1}{B}\sum_{j=1}^{B}\sum_{\mathbf{x}}\sum_{i=1}^{m}\log\pi_{z_{i1}}f(\mathbf{x}|\mathbf{y},\mathbf{u}^j,\theta^p)
\tag{2.16}
$$

we have:

$\hat{\pi}_k=\frac{1}{B}\frac{1}{m}\sum_{j=1}^{B}\sum_{i=1}^{m}p(x_{i1}=k|\mathbf{y},\mathbf{u}^j,\theta^p)$

$$
E[\sum_{i=1}^{m}\sum_{t=2}^{n_i}\log p_{x_{i,t-1},x_{it}}|\mathbf{y},\theta^p]=\frac{1}{B}\sum_{j=1}^{B}\sum_{\mathbf{x}}\sum_{i=1}^{m}\sum_{t=2}^{n_i}\log p_{x_{i,t-1},x_{it}}f(\mathbf{x}|\mathbf{y},\mathbf{u}^j,\theta^p)
\tag{2.17}
$$

we have:

$\hat{p}_{k,l} = \frac{\sum_{j=1}^{B}\sum_{i=1}^{m}\sum_{t=2}^{n_i} p(x_{i,t-1}=k, x_{it}=l|\mathbf{y},\mathbf{u}^j,\theta^p)}{\sum_{j=1}^{B}\sum_{l=1}^{K}\sum_{i=1}^{m}\sum_{t=2}^{n_i} p(x_{i,t-1}=k, x_{it}=l|\mathbf{y},\mathbf{u}^j,\theta^p)}$

$$E[\log f(\mathbf{u};\theta)|\mathbf{y},\theta^p] \tag{2.18}$$

Note that we sample $\mathbf{u}$ from $f(\mathbf{u}|\mathbf{y},\theta^p)$, thus $\hat{\sigma}_u^2$ is the sample variance of $\mathbf{u}^j$, for $j = 1, ..., B$, that is

$$\hat{\sigma}_u^2 = \frac{1}{B}\sum_{j=1}^{B}\sum_{i=1}^{m}(u_i^j)^2 \tag{2.19}$$

Forward and backward algorithm:

For $i = 1, ..., m$, let $\alpha_k^{(i)}(t) = p(y_{i1}, ..., y_{it}, x_t = k|u_i^j)$ and $\beta_k^{(i)}(t) = p(y_{i,t+1}, ..., y_{in_i}|x_t = k, u_i^j)$ for $1 \leq t \leq n_i - 1$. set $\beta_k^{(i)}(n_i) = 1$. Then we have:

$\alpha_k^{(i)}(1) = \pi_k p(y_{i1}, x_{i1} = k|u_i^j) \quad$ for $1 \leq k \leq K, 1 \leq i \leq m$

$\alpha_k^{(i)}(t) = p(y_{it}|x_{it} = ku_i^j)\sum_{l=1}^{k}\alpha_l^{(i)}(t-1)p_{l,k} \quad$ for $1 < t \leq n_i, 1 \leq k \leq K, 1 \leq i \leq m$

$\beta_k^{(i)}(n_i) = 1 \quad$ for $1 \leq k \leq K, 1 \leq i \leq m$

$\beta_k^{(i)}(t) = \sum_{l=1}^{k} p_{k,l} p(y_{i,t+1}|x_{i+1} = l, u_i^j)\beta_l^{(i)}(t+1) \quad$ for $1 \leq t < n_i, 1 \leq k \leq K, 1 \leq i \leq m$

then we have:

$$p(x_{it} = k|\mathbf{y}, \mathbf{u}^j, \theta^p) = \frac{\alpha_k^{(i)}(t)\beta_k^{(i)}(t)}{\sum_{k=1}^{K}\alpha_k^{(i)}(t)\beta_k^{(i)}(t)} \tag{2.20}$$

$$p(x_{i,t-1} = k, x_{it} = l|\mathbf{y}, \mathbf{u}^j, \theta^p) = \frac{\alpha_k^{(i)}(t-1)p_{k,l}p(y_{i,t+1}|\theta^p, x_{t+1} = l)\beta_l^{(i)}(t+1)}{\sum_{k=1}^{K}\alpha_k^{(i)}(t-1)\beta_k^{(i)}(t-1)} \tag{2.21}$$

2.4 Simulation

In this section, we present the simulation study of the MSARM models. Here is the simulation setting, for $i = 1, ..., m$, $t = 1, ..., n$:

$$y_{it} = \begin{cases} 1 + 0.1y_{i,t-1} + e_{it}^{(1)} + u_i & \text{under state 1} \\ 3 + 0.2y_{i,t-1} + e_{it}^{(2)} + u_i & \text{under state 2} \end{cases}$$

We generated 20 sequences with same length 100, which means m=20 and n=100. The transition probability matrix is $(p_{11}, p_{12}, p_{13}, p_{14}) = (0.9, 0.1, 0.2, 0.8)$. The intercepts are 0.1 and 0.2, and the coefficients are 1 and 3, under state 1 and 2 respectively, which means $\alpha_1 = 1$, $\alpha_2 = 3$, $\beta_1^{(1)} = 0.1$ and $\beta_1^{(2)} = 0.2$. $e_{it}^{(1)}$ are the i.i.d random error under state 1, follows the normal distribution with mean 0 and variance 0.2. $e_{it}^{(2)}$ are the i.i.d random error under state 2, follows the same distribution. u_i are the random effect among sequences, follows the normal distribution with mean 0 and variance 1.

Remarks:

- Due to the application to real data, we prefer to simulate a similar data set, thus in our simulation there are 1 AR effect($r = 1$), two underlying states(K=2). In general our method can be applied to different settings.

- The variance of random error and random effect are set to be 0.2 and 1 respectively, which is for comparison purpose. The results will show that our model can capture the variation both within the sequence and among the sequences very well.

- In the MCEM procedure, when apply the Monte Carlo method, the number of samples, B, required to approximate the E-step accurately is an important practical consideration. Here we set B=500, and the simulation results are good when comparing to the true values. More accurate results can be achieved by increasing the value of B, but it may bring the computing burden. In application, people can try the E-step several times, thus the approximate value of B can be chosen by the reflected variation of the value in E-step. Also note that the value of B depends on the number of random effects. When the number of the random effect become large, which means a high dimension integral need to be approximated, thus a much larger value of B is necessary.

- Initial values are crucial in the estimation of conventional HMM, and also in our model, which due to large number of parameters and unknown hidden states. Then how to find the global maximum for the log-likelihood instead of the local maximum is very important. Here the approach is that first based on the data, we get the summary statistics(such as mean and variance) from the data, then we try different values in the estimated regions, and choose the estimators with largest log-likelihood. Typically it's not necessary to try different initial values for all the parameters. For the transition matrix, the true values are between 0 and 1, so the initial value for the transition probability parameters are just set to (0.5,0.5,0.5,0.5). The AR effect is usually between -1 and 1, so we can either try different initial values in $[-1, 1]$, or just set them to be 0. In our simulation, we first set the initial values of the transition probability parameters all to be 0.5. The initial values of the σ_1^2, σ_2^2 and the variance of random effect σ_u^2 are all set to be 0.5. The initial value of the coefficients $\beta_1^{(1)}$ and $\beta_1^{(2)}$ are set to be 0. The intercept α_1 and α_2 are tried different values in region (-5,5), based on the mean of the data.

The simulation results are shown in Table 2.1. The true values of the parameters are shown in column 2. First we generated 20 sequences with 100 observations in each sequence($n = 100$ and $m = 20$), repeated 100 times to find the mean and standard deviations of the estimators. The simulation is performed in R with seeds set from 1 to 100. The estimators and the corresponding with standard deviations are shown in column

3. From the results, we can see that the estimations of the transition probability parameters are very good. For the AR effects, the results seem a little bit overestimation. Also the estimators of the variance under both states are a little bit overestimated. We think that this overestimation maybe due to the small size of the observations. Thus we increase the number of observations in each sequence from 100 to 500(n=500), and all other settings are remained the same. The results are shown in column 4. From the results, we can see that after increasing the number of observations, all estimators perform well with small variance.

Table 2.1: Simulation study

Parameters	True Value	$n = 100$ $m = 20$	$n = 500$ $m = 20$
p_{11}	0.9	0.899(0.009)	0.901(0.004)
p_{12}	0.1	0.101(0.009)	0.099(0.004)
p_{21}	0.2	0.204(0.017)	0.199(0.008)
p_{22}	0.8	0.796(0.017)	0.801(0.008)
$\beta_1^{(1)}$	0.1	0.075(0.059)	0.082(0.027)
α_1	1	1.152(0.115)	1.025(0.098)
σ_1	0.447	0.561(0.022)	0.478(0.006)
$\beta_1^{(2)}$	0.2	0.165(0.053)	0.182(0.023)
α_2	3	3.103(0.172)	3.035(0.114)
σ_2	0.447	0.555(0.023)	0.472(0.007)
σ_u	1	1.087(0.209)	1.033(0.181)

2.5 Application

We applied the MSARM to study the molecular binding mechanism based on the cell adhesion data. There are 18 pairs of cells, and for each pair, the force clamp assays are conducted. For the force clamp assays of each pair of the cells, several force-clamp periods (i.e., defined as bond lifetime) are measured in a certain period of time. The original data contains 18 time series with bond lifetimes as responses. As mentioned, the data have three unique features. First, there are multiple repeated assays collected from different pairs of cells. The same biological mechanism is shared within the same pair of cells and some variations exist among different pairs of cells. Second, the molecular bond formation, which is of our major interest, is not directly observable. Third, there are some memory effects in the repeated bond lifetime measurements. It was discovered that cells appear to have the ability to remember the previous adhesion events and such a memory has an impact on the future adhesion behaviors.

Those bond lifetimes become the responses in the MSARM. And for each measured bond lifetime, it has an underlying hidden state, which is defined as *bonded* or *not bonded*. Table 2.2 shows the numerical results based on the 18 time series. The first four rows of the table shows the transition probabilities. State 1 represents *not bonded* and state 2 represents *bonded*. The probability from state *bonded* to *bonded* is 0.313 ($p_{22} = 0.313$). It shows the memory effect that cells appear to have the ability to remember the previous adhesion events and such a memory has an impact on the future adhesion behaviors. In this biological systems the occurrence of binding in the immediate past assay could decrease the likelihood for the next assay to result in a binding. The 5-7 rows and 8-10 rows show the estimation of the parameters under state *not bonded* and *bonded* respectively. $\beta_1 = -0.002$ shows that the bond lifetime under state *not bonded* is not affected by the previous one. $\beta_2 = 0.083$ shows that the bond lifetime under state *bonded* is affected by the previous one. Based on α_1 and α_2, the bond lifetime under the two states are quite different. The variance of the bond lifetime under *bonded* state is much larger. By applying the MSARM, we discovered the memory effects and provided useful information of the molecular biding mechanism under the two states.

Table 2.2: MSARM results based on the 18 pairs of cells time series

Parameters	estimator
p_{11}	0.942
p_{12}	0.058
p_{21}	0.687
p_{22}	0.313
β_1	-0.002
α_1	0.157
σ_1	0.877
β_2	0.083
α_2	5.809
σ_2	4.012
σ_u	0.834

2.6 Appendix

2.6.1 Assumptions

(A.1) The transition probability matrix is ergodic, that is, irreducible and aperiodic.

(A.2) For all underlying state k, k=1,...,K, for any observation y and random effect u, $g_\theta(y|a) = \int p(y|u, x = k)f(u)du$ has 2 continuous derivatives in in neighborhood $|\theta - \theta_0| < \delta$.

(A.3) Write $\theta_1 = (\theta_{11}, ..., \theta_{1d})$. There exists a $\delta > 0$ such that

(i) for all $1 \leq i \leq d$ and all k,

$$E_0[\sup_{|\theta-\theta_0|<\delta} |\frac{\partial}{\partial\theta_{1i}} \log g_\theta(y_1|k)|] < \infty \tag{2.22}$$

(ii) for all $1 \leq i, j, \leq d$ and all k,

$$E_0[\sup_{|\theta-\theta_0|<\delta} |\frac{\partial^2}{\partial\theta_{1i}\partial\theta_{1j}} \log g_\theta(y_1|k)|] < \infty \tag{2.23}$$

(iii) for $j = 1, 2$, all $1 \leq i_l \leq d$, $l = 1, ..., j$, and all k,

$$\int \sup_{|\theta-\theta_0|<\delta} |\frac{\partial^j}{\partial\theta_{1i_1} \cdot \partial\theta_{1i_j}} g_\theta(y|k)|vdy < \infty \tag{2.24}$$

(A.4) There exists a $\delta > 0$ such that with

$$\rho_0(y) = \sup_{|\theta-\theta_0|<\delta} |\max_{1\leq a,b\leq K} \frac{g_\theta(y|a)}{g_\theta(y|b)}| \tag{2.25}$$

$P_0(\rho_0(y_1) = \infty|x_1 = a) < 1$ for all a.

(A.5) θ_0 is an interior point of Θ.

(A.6) The maximum-likelihood estimator is strongly consistent.

2.6.2 Technical proofs

First, we will derive some the asymptotic properties when $m = 1$. The technique is the same as Bickel(1998), first we extend the bivariate process $\{(x_k, y_k)\}$ to a doubly infinite stationary sequence $\{(x_k, y_k)\}_{k=-\infty}^{k=\infty}$. Then by a martingale convergence theorem by Levy, $p_\theta(y_1|y_{-n}^0) \to p_\theta(y_1|y_{-\infty}^0)$. We use similar notations as Bickel(1998), writing $\lambda_\theta(a, b) = \frac{\partial \log p_{ab}}{\partial\theta_1}$, $\gamma_\theta(y|a) = \frac{\partial \log f(y|x=a)}{\partial\theta_1} = \frac{\partial \log \int p(y|u,x=a)f(u)du}{\partial\theta_1}$, and $\tau_\theta(a) = D \log \pi_a$. $g_\theta(y|a) = p(y|a) = \int p(y|u, x = a)f(u)du$. Thus by (4) and (5) in Bickel(1998),

$$\begin{aligned} & D \log p_{\theta_0}(y_1|y_0, ..., y_{-n}) \\ = & \sum_{k=-n}^{0} E[\gamma(y_k|x_k, y) + \lambda(x_k, x_{k+1})|y_{-n}^1] \\ & - \sum_{k=-n}^{0} E[\gamma(y_k|x_k, y) + \lambda(x_k, x_{k+1})|y_{-n}^0] \\ & + E[\gamma(y_1|x_1)|y_n^1] + E[\tau(x_{-n})|y_{-n}^1] - E[\tau(x_{-n})|y_{-n}^0] \end{aligned} \tag{2.26}$$

The the likelihood with random effect have the same expansion as Bickels(1998), with no more assumptions added.

Theorem 2.2. *Under assumptions (A.1)-(A.6),* $N^{-1/2}\frac{\partial L(\theta_0)}{\partial\theta_1} \to N(0, I_{11})$ *as* $n \to \infty$.

Proof. Let L_i denote the likelihood of the i-th cell. We will first prove that $n^{-1/2}\frac{\partial L_i(\theta_0)}{\partial\theta_1} \to N(0, A_i)$ as $n \to \infty$. where A_i is the limit of the covariance matrix of $n^{-1/2}\frac{\partial L_i(\theta_0)}{\partial\theta_1}$. For each sequence i, let $\xi_k = \frac{\partial \log p_{\theta_0}(y_k|y_{k-1},\dots,y_1)}{\partial\theta_1}$, so that $\frac{\partial L_i(\theta_0)}{\partial\theta_1} = \sum_{k=1}^n \xi_k$, and let

$$\eta_k = \sum_{i=-\infty}^{k-1} E_0[\gamma_0(y_i|x_i) + \lambda_0(x_i, x_{i+1})|y_{-\infty}^k]$$

$$- \sum_{i=-\infty}^{k-1} E_0[\gamma_0(y_i|x_i) + \lambda_0(x_i, x_{i+1})|y_{-\infty}^{k-1}]$$

$$E_0[\gamma_0(y_k|x_k)|y_{-\infty}^k] \tag{2.27}$$

Based on the assumptions (A.1)-(A.6), we have the same results as Lemma 3-6 in Bickle(1998), without any more conditions, so that η_k is a stationary and ergodic martingale increment sequence. Its covariance matrix is A_i. By the central limit theorem for martingales, we obtain

$$n^{-1/2}\sum_{k=1}^n \eta_k \to N(0, A_i) \tag{2.28}$$

By Lemma 6 in Bickel(1998), we have

$$\|n^{-1/2}\sum_{k=1}^n \xi_k - n^{-1/2}\sum_{k=1}^n \eta_k\| \leq n^{-1/2}\sum_{k=1}^n \|\xi_k - \eta_k\| \to 0. \tag{2.29}$$

Then we have $n^{-1/2}\frac{\partial L_i(\theta_0)}{\partial\theta_1} \to N(0, A_i)$ as $n \to \infty$. Thus, for all $i = 1, 2, ..., N$, $n^{-1/2}\frac{\partial L_i(\theta_0)}{\partial\theta_1} \to N(0, A_i)$, as $n \to \infty$. Then

$$N^{-1/2}\frac{\partial L(\theta_0)}{\partial\theta_1} = N^{-1/2}\sum_{i=1}^m \frac{\partial L_i(\theta_0)}{\partial\theta_1} \to m^{-\frac{1}{2}}N(0, \sum_{i=1}^m A_i) \to N(0, I_{11}) \tag{2.30}$$

□

Theorem 2.3. *Under assumptions (A.1)-(A.6), and let θ_n^* be any stochastic process such that $\theta^* \to \theta$ as $n \to \infty$. Then $N^{-1}\frac{\partial^2 L(\theta^*)}{\partial\theta_1\partial\theta_1^T} \to I_{11}$*

Proof. Let L_i denote the likelihood of the i-th cell. Then apply the same technique, we have $n^{-1}\frac{\partial^2 L_i(\theta^*)}{\partial\theta_1\partial\theta_1^T} \to A_i$. Then

$$N^{-1}\frac{\partial^2 L(\theta^*)}{\partial\theta_1\partial\theta_1^T} = N^{-1}\sum_{i=1}^{m}\frac{\partial^2 L_i(\theta^*)}{\partial\theta_1\partial\theta_1^T} \to \frac{1}{m}\sum_{i=1}^{m} A_i \to I_{11} \tag{2.31}$$

□

Theorem 2.4. *Under assumptions (A.7)-(A.8), let θ^* be any stochastic process such that $\theta^* \to \theta$ as $m \to \infty$. Then $m^{-1/2}\frac{\partial L(\theta_0)}{\partial\theta_2} \to N(0, I_{22})$ and $N^{-1}\frac{\partial^2 L(\theta^*)}{\partial\theta_2\partial\theta_2^T} \to I_{22}$, as $m \to \infty$.*

Proof. Based on the matrix form of the likelihood,

$$\frac{\partial L}{\partial\sigma_i} = \sum_{\mathbf{x}}[-tr(\Sigma_{\mathbf{x}}^{-1}G_i) + (\mathbf{y} - U_0\gamma_{\mathbf{x}})'\Sigma_{\mathbf{x}}^{-1}G_i\Sigma_{\mathbf{x}}^{-1}(\mathbf{y} - U_0\gamma_{\mathbf{x}})]P(\mathbf{x})/2 \tag{2.32}$$

$$\frac{\partial^2 L}{\partial\sigma_i\partial\sigma_j} = \sum_{\mathbf{x}}[-tr(\Sigma_{\mathbf{x}}^{-1}G_i\Sigma_{\mathbf{x}}^{-1}G_j) - 2(\mathbf{y} - U_0\gamma_{\mathbf{x}})'\Sigma_{\mathbf{x}}^{-1}G_i\Sigma_{\mathbf{x}}^{-1}G_j\Sigma_{\mathbf{x}}^{-1}(\mathbf{y} - U_0\gamma_{\mathbf{x}})/2]P(\mathbf{x})/2 \tag{2.33}$$

By the fact of Weiss(1971,1973) and Theorem 3.1 in Miller(1977), $\frac{\partial L(\theta_0)}{\partial\theta_2} \to N(0, I_{22})$. Also

$$\begin{aligned}
&|\frac{\partial^2 L(\theta^*)}{\partial\phi_i\partial\phi_j}/m - [I_{22}]_{ij}| \to 0 && \phi_i = (\theta_2)_i, \phi_j = (\theta_2)_j\\
&|\frac{\partial^2 L(\theta^*)}{\partial\phi_i\partial\phi_j}/\sqrt{N} - [I_{12}]_{ij}| \to 0 && \phi_i = (\theta_1)_i, \phi_j = (\theta_2)_j\\
&|\frac{\partial^2 L(\theta^*)}{\partial\phi_i\partial\phi_j}/\sqrt{N} - [I_{21}]_{ij}| \to 0 && \phi_i = (\theta_2)_i, \phi_j = (\theta_1)_j
\end{aligned} \tag{2.34}$$

□

Proof of Theorem 2.1

Proof. First, by the results from Theorem 2.2 and Theorem 2.4,

$$\begin{pmatrix} N^{-\frac{1}{2}}\frac{\partial L(\theta_0)}{\partial \theta_1} \\ \\ m^{-\frac{1}{2}}\frac{\partial L(\theta_0)}{\partial \theta_2} \end{pmatrix} \to N(0, I_0^{-1})$$

By Taylor expansion,

$$0 = \frac{\partial L(\hat{\theta})}{\partial \theta} = \frac{\partial L(\theta_0)}{\partial \theta}(\hat{\theta} - \theta) + \frac{\partial^2 L(\theta^*)}{\partial \theta \partial \theta^T}(\theta^* - \theta) \tag{2.35}$$

Then by the result from Theorem 2.3 and Theorem 2.4, the result follows immediately.

□

Reference

R. M. Altman (2007). Mixed hidden markov models: An extension of the hidden markov model to the longitudinal data setting. *Journal of the American Statistical Association* **102**, 201–210.

Banerjee, S., Gelfand, A. E., Finley, A. O. , and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society B* **70**, 825–848.

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

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.

Büchlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics* **30**, 927–961.

Chen W, Evans EA, McEver RP, Zhu C. (2008a). Monitoring Receptor-Ligand Interactions between Surfaces by Thermal Fluctuations. *Biophys J.* **94**, 694–701

Chu, T., Zhu, J. and Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *Annals of Statistics* **39**, 2607–2625.

Cressie, N. (1993). *Statistics for Spatial Data* Wiley, New York.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society B* **70**, 209–226.

Deng, X., Hung, Y., and Lin, C. D. (2015). Design for computer experiments with qualitative and quantitative factors. *Statistica Sinica* **25**, 1567-1581.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika* **81**, 425–455.

Draguljić, D., Dean, A. M., and Santner, T. J. (2012). Noncollapsing space-filling designs for bounded nonrectangular regions. *Technometrics* **54**, 169–178.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*, Chapman and Hall/CRC press, New York.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**:13481360.

Fang, K.-T., Li, R. and Sudjianto, A. (2006). *Design and modeling for computer experiments*, Chapman and Hall/CRC press, New York.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association* **102**, 321–331.

Furrer, R., Genton,M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15**, 502–523.

Goncalves, S. and White, H. (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *Journal of Econometrics* **119**, 199–219.

Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics* **24**, 561–578.

Gramacy, R. B. and Lee, H. K. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association* **103**, 1119–1130.

Hoeting, J., Davis, R., Merton, A. and Thompson, S. (2006). Model Selection For Geostatistical Models. *Ecological Applications* **16**:8798

Huang, H. and Chen, C. (2007). Optimal Geostatistical Model Selection. *Journal of the American Statistical Association* **102**, 1009-1024.

Hung, Y., Zarnitsyna, V., Zhang, Y., Zhu, C. and Wu, C. J. (2008). Binary time series modeling with application to adhesion frequency experiments. In *Journal of the American Statistical Association* **103**

Hung, Y., Qian, P. Z. G., and Wu, C. F. J. (2012). Statistical design and analysis methods for data center thermal management. In *Energy efficient thermal management of data centers* (J. Yogendra and K. Pramod eds.), Springer, New York.

Hung, Y. (2011). Penalized Blind Kriging in Computer Experiments. *Statistica Sinica* **21**, 1171-1190

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* **40**, 633–643.

Joseph, V., Hung, Y. and Sudjianto, A. (2008), Blind Kriging: A New Method for Developing Metamodels. *Journal of Mechanical Design* **130(3)**, 031102

Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association* **103**, 1545–1555.

Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K. and Frieman, J. A (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *The Annals of Applied Statistics* **5**, 2470–2492.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **17**, 1217–1241.

Lahiri, S. N. (1995). On the asymptotic behavior of the moving block bootstrap for normalized sums of heavy-tail random variables. *The Annals of Statistics* **23**, 1331–1349.

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statistics* **27**, 386–404.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*, Springer, New York.

Liang, F., Cheng, Y., Song, Q., Park, J. and Yang, P. (2013). A resampling-based stochastic approximation method for analysis of large geostatistical data. *Journal of the American Statistical Association* **108**, 325–339.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. and Ye, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48** 478–490.

Liu, R. Y. and Singh, K. (1992). Moving Blocks Jackknife and Bootstrap Capture Weak Dependence. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 225–248, Wiley, New York.

Loh, W.-L. (1996). On Latin hypercube sampling. *The annals of statistics* **24**, 2058–2080.

Lopez V. and Hamann, H. F. (2011). Heat transfer modeling in data centers. *International Journal of Heat and Mass Transfer* **54**, 5306–5318.

Mardia, K.V. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146.

Marshall BT, Long M, Piper JW, Yago T, McEver RP, Zhu C (2003). Direct observation of catchbonds involving cell-adhesion molecules *Nature* **423**, 190–193.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245.

A. D. Mehta, R. S. Rock, M. Rief, J. A. Spudich, M. S. Mooseker, R. E. Cheney (1999). Myosin-V is a processive actin-based motor *Nature* **400**, 590.

Nordman, D. J., Lahiri, S. N. and Fridley, B. L. (2007). Optimal block size for variance estimation by a spatial block bootstrap method. *Sankhyā* **69**, 468–493.

Nychka, D. W. (2000). Spatial-process estimates as smoothers. In *Smoothing and regression: approaches, computation, and application*, (M. G. Schimek ed.), 393–424, Wiley, New York.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multi-resolution Gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics*, to appear.

Nychka, D. W., Wikle, C. and Royle, J. A. (2002). Multiresolution models for non-stationary spatial covariance functions. *Statistical Modeling* **2**, 315–331.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association* **89**, 1303–1313.

Qian, P. Z. G. and Jeff, C. F. J. (2009). Sliced space-filling designs. *Biometrika* **96**, 945-956.

Rougier, J. (2008). Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics* **17**, 827–843.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, Chapman and Hall/CRC Press, Boca Raton.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics* **29**, 31–49.

Santner, T. J., Williams, B. J. and Notz, W. (2003). *The design and analysis of computer experiments* Springer, New York.

Smola, A. J. and Bartlett, P. L. (2001). Sparse greedy Gaussian process regression.

In *Advances in Neural Information Processing Systems* **13**, (T. K. Leen, T. G. Dietterich, and V. Tresp eds.) 619–625.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems* **18**, 1257–1264.

Stein, M. L. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29**, 143–151.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*, Springer, New York.

Stein, M. L. (2013). Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics* **22**, 866–885.

Stein, M. L., Chi, Z. and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B* **66**, 275–296.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society B* **58**, 267–288.

Tang, B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association* **88**, 1392–1397.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics* **34**, 15–25.

White, H. (1996). *Estimation, inference and specification analysis*, Cambridge university press, New York.

Wikle, C. K. (2010). Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* (A. E. Gelfand, P. Diggle, M. Fuentes and P. Guttorp eds.), 107–118, Chapman and Hall/CRC Press, Boca Raton.

Ying, Z.-L. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Annals of Statistics* **21**, 1567–1590.

Zarnitsyna VI, Huang J, Zhang F, Chien Y-H, Leckband D, Zhu C (2007). Memory in Receptor-ligand Mediated Cell Adhesion *Proceedings of the National Academy of Sciences USA* **104**,1803718042.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99**, 250–261.

Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92**, 921–936.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**, 15091533.