UNIVERSITY OF CALIFORNIA
RIVERSIDE

Higher-Order Accurate Variance Estimation in Markov Chain Monte Carlo (MCMC)

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Deepak Bastola

September 2021

Dissertation Committee:

Dr. James Flegal, Chairperson
Dr. Esra Kurum
Dr. Bahram Mobasher

The Dissertation of Deepak Bastola is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First and foremost, I would like to thank my advisor Prof. James Flegal for a very effective supervision and for giving me useful research direction and deep insights. The research discussions I had with Dr. Flegal were my biggest source of inspiration. They were very stimulating and motivated me to stay focused with my research.

I would not have been here without the generous fellowship and teaching opportunities provided to me by the graduate division and the Department of Statistics, respectively. I would also like to acknowledge that this has helped ease my financial burden and has tremendously helped my studies and research at the University of California, Riverside (UCR).

I could not have possibly made it without the company and help from my colleagues Jinhui, Song, Huiling, Linli, Isaac, and Lauren from the Department of Statistics, UCR. Finally, I would like to thank the awesome professors from the Department, my oral exam committee professors, my dissertation committee professors for all of their valuable time and contributions.

I would like to thank my parents and my wife for always being there for me in my time of need.

ABSTRACT OF THE DISSERTATION

Higher-Order Accurate Variance Estimation in Markov Chain Monte Carlo (MCMC)

by

Deepak Bastola

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, September 2021
Dr. James Flegal, Chairperson

Variance estimation in the context of high dimensional Markov Chain Monte Carlo (MCMC) is an interesting topic in research. Practical implications of estimating variance are limited due to inherent systematic bias both in univariate and multivariate settings. Recent advancements in high dimensional covariance matrix estimation in MCMC setting including works on Lugsail Batch Means (LUG-BM) have proven to improve the bias properties. Using spectral theory of estimation, we can further improve upon the bias and variance properties of the estimators. Finite sample properties of variance estimators should be studied in detail using statistical properties of the sampling bias, while accounting for the sampling error. The direction of this bias is crucial in finite sample applications. Mean Square Error (MSE) has been traditionally used to assess the quality of estimation. However, using alternate asymmetrical loss functions is recommended as they are more natural to use in applications where constructing optimal variance estimators in finite samples is required. Normality as-

sumptions are required to make efficient use of these techniques and a careful analysis should be done to ensure the assumptions for normality are met.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Markov Chains

Covariance estimation is one of the fundamental methods in modern statistics. Its importance is more pronounced in real-world problems such as astrophysics, finance, remote sensing, medicine, genomics, geostatistics, and the like, where the parameter space is very large and the degree of co-occurences between the variables is hard to quantify. With the aid of computer simulations and Monte Carlo methods, generating random variables from seemingly intractable distributions has become easy Roberts and Rosenthal (2004). A better estimation of the error covariance matrix can lead to better understanding of the modeling and inherent physical process. Moreover, it can ultimately aid in dimension reduction which is highly desirable when dealing with large dimensions.

Markov Chain Monte Carlo (MCMC) is a set of simulation methods for drawing samples from a distribuition, $\pi(\cdot)$, defined on a measurable space $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X}$ is the state space and $\mathcal{B}$ is a countably generated Borel σ-algebra. MCMC methods are widely used to analyze complex probability distributions and are practical when independent sampling from such distributions is difficult. The distribution $\pi(\cdot)$ is usually known only up to some proportionality constant. Although MCMC methods produce dependent samples, the Ergodic Theorem guarantees that for Markov chain started with invariant distribution $\pi(\cdot)$, long run averages follow the strong law of large numbers (SLLN).

A time-homogeneous discrete time Markov chain, $\{X_n\}_{n\in\mathbb{N}}$, is a collection of random variables, X_n, each defined on a measurable space $(\mathcal{X}, \mathcal{B})$, such that,

$$\mathbb{P}[X_n \in A | X_1 = x_1, ..., X_{n-1} = x_{n-1}] = \mathbb{P}[X_n \in A | X_{n-1} = x_{n-1}], \qquad \forall A \in \mathcal{B}.$$

A Markov chain consists of a state space, an initial distribution, and a transition kernel. The state space $\mathcal{X}$ is the collection of the possible values of X. The initial distribution is the marginal distribution of X_1. The transition kernel is the conditional distribution of X_{n+1} given X_n i.e., $P(X_{n-1}, A) = \mathbb{P}[X_n \in A | X_{n-1} = x_{n-1}]$, where $P(x, \cdot)$ defines a distribution over $(\mathcal{X}, \mathcal{B})$ for any $x \in \mathcal{X}$, and $P(\cdot, A)$ is measurable for any $A \in \mathcal{B}$. The n-step transition kernel denotes the probability that the Markov chain at x will be in set A after n steps. For $x \in \mathcal{X}$, $A \in \mathcal{B}$, and $l \in \{1, 2, 3, ..\}$, the

n-step kernel can then be defined as

$$P^n(x, A) = \mathbb{P}[X_{m+l} \in A | X_l = x].$$

A probability measure $\pi(\cdot)$ on $\mathcal{B}$ is called invariant for $\{X_n\}_{n\in\mathbb{N}}$, if, for all measurable sets A,

$$\pi(A) = \int_{\mathcal{X}} P(x, A)\pi(dx).$$

It can be shown that $\int_{\mathcal{X}} P^n(x, A)\pi(dx) = \pi(A)$ for $n = 1, 2, 3, \ldots$. So, if $X_1 \sim \pi$, then $X_n \sim \pi$ for all n and $\{X_n\}_{n\in\mathbb{N}}$ is stationary. If $P(x, \cdot)$ admits a density $p(x'|x)$ then equivalently,

$$\pi(x') = \int_{\mathcal{X}} p(x'|x)dx.$$

The Markov chain is said to be reversible with respect to π if, at stationarity, the probability that $x_i \in A$ and $x_{i+1} \in B$ are equal to the probability that $x_{i+1} \in A$ and $x_i \in B$. In mathematical notation, reversibility condition is satisfied if $\pi(x)p(x'|x) = \pi(x')p(x|x')$, also referred to as "detailed balance".

1.1.1 Markov Chain Asymptotics

Let X_∞ be a stationary $\mathbb{P}_n$-Markov chain with the invariant probability distribution π that converges in probability to a Harris ergodic Markov chain. Then for any bounded continuous function $g : \mathcal{X} \to \mathbb{R}^p$ such that $\mathrm{E}_\pi|g| < \infty$, and for any $n \to \infty$,

$$\int_{\mathcal{X}} g(X)\pi(dx) = n^{-1}\sum_{i=1}^{n} g(X_i) \quad \text{w.p. 1.} \tag{1.1}$$

If (1.1) holds for any $n \to \infty$, then the MCMC procedure is called consistent. The p-dimensional mean vector of interest is $\theta = E_\pi\{g(X)\} = \int_\chi g(X)\pi(dx)$. For ease in notation, let's call $Y_t = g(X_t)$, where $g = (g^{(1)}, g^{(2)}, ..., g^{(p)})$. Then from the consistency property, the estimator, $\theta_n = n^{-1}\sum_{t=1}^{n} Y_t \to \theta$ w.p. 1 as $n \to \infty$.

1.1.2 Ergodicity

A stochastic process $\{X_t\}$ is said to be stationary if $\mathrm{E}[X_t] = \theta$ for all j and the covariance between two observations that are s time or space unit apart depends only on the lag s i.e.,

$$R(s) = \mathrm{E}\left[(X_t - \mu_X)(X_{t+s} - \mu_X)\right], \quad s = 0, \pm 1, \pm 2, \ldots.$$

Markov chains will never be at stationary in practice. To assess Markov chains in terms of how far away they are from stationarity, the notion of total variation

distance is established. The total variation distance is the largest difference between the probabilities of a single event in $\mathcal{B}$ between the probability measures $P^n(x,\cdot)$ and $\pi(\cdot)$, i.e.,

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} = \sup_{A\in\mathcal{B}} |P^n(x,A) - \pi(A)|. \tag{1.2}$$

The rate of convergence of the total variation distance between the n-step transition kernel to the invariant distribution determines the ergodicity of the Markov chain. A Markov chain $\{X_n\}_{n\in\mathbb{N}}$ is called Harris ergodic if it is ψ-irreducible, aperiodic, and Harris recurrent (Meyn and Tweedie, 1993; Jones, 2004). Harris ergodicity implies that for every $x \in \mathcal{X}$,

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \downarrow \quad 0, \tag{1.3}$$

as $n \to \infty$. However, Harris ergodicity is not enough to specify the rate of convergence of the Markov chain to stationarity. The convergence rate of a Harris ergodic Markov chain can be studied by finding an upper bound for the total variation distance.

A Harris Markov chain is geometrically ergodic if there exists a constant $\rho \in [0,1)$ and a function $M : \mathcal{X} \to [0,\infty)$ such that for any $x \in \mathcal{X}$ and any $n \in \mathbb{N}$,

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq M(x) \cdot \rho^n. \tag{1.4}$$

If $M(x)$ is also bounded from above in (1.4) in addition to being geometrically ergodic, the chain is called uniformly ergodic. Geometric ergodicity is the weaker of the two and easier to establish.

Another form of ergodicity exists which is even weaker and is called polynomial ergodicity. A Markov chain is polynomially ergodic of order m if there exists a non-negative function $M(x)$ and $m \geq 0$ such that

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq M(x) \cdot n^{-m}. \tag{1.5}$$

There exists many ways of proving ergodicity in the literature and they are often established using drift and minorization conditions, see e.g. Meyn and Tweedie (1993), Jones and Hobert (2001), and Roberts and Rosenthal (2004).

1.1.3 Mixing

Temporal dependence of observations that are far away in the past or future can be studied under certain mixing conditions, see for e.g. Bradley (2005). Mixing conditions are flexible and they are useful to quantify how fast observations far from each other achieve independence (Rosenblatt, 1961), (Rosenblatt, 1972). Let $\mathcal{F}_{-\infty}^{n} = \sigma(..., X_n)$ be the smallest collection of subsets of Ω that contains the union of the σ-fields $\mathcal{F}_a^n$ as $a \to -\infty$. Similarly, let $\mathcal{F}_{n+m}^{\infty} = \sigma(X_{n+m}, ...)$ be the smallest collection of subsets that contain the union of the σ-fields $\mathcal{F}_{n+m}^{a}$ as $a \to \infty$.

For any two σ-fields $\mathcal{A}$ and $\mathcal{B} \subset \mathcal{F}$, define the following measures of dependence,

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup|P(A \cap B) - P(A)P(B)|, \qquad A \in \mathcal{A}, B \in \mathcal{B}$$

$$\phi(\mathcal{A}, \mathcal{B}) = \sup|P(B|A) - P(B)|, \qquad A \in \mathcal{A}, B \in \mathcal{B}, P(A) > 0,$$

where the supremum is taken over all pairs of (finite) partitions $\{A_1, ..., A_I\}$ and $\{B_1, ...$

$.., B_J\}$ of Ω such that $A_i \in \mathcal{A}$ for each i and $B_j \in \mathcal{B}$ for each j (Bradley, 2005). The dependence coefficients can then be defined as

$$\alpha(n) = \sup_{j \in \mathbb{Z}} \quad \alpha(\mathcal{F}^i_{-\infty}, \mathcal{F}^\infty_{j+n}),$$

$$\phi(n) = \sup_{j \in \mathbb{Z}} \quad \phi(\mathcal{F}^i_{-\infty}, \mathcal{F}^\infty_{j+n}).$$

A random process is m-dependent if the σ-fields $\mathcal{F}^k_1$ and $\mathcal{F}^\infty_{k+m+1}$ are independent for all $k \geq 1$. The random sequence $\{X\}$ is said to be strongly mixing (or α-mixing) if $\alpha(n) \to 0$ as n$\to \infty$. It is said to be ϕ-mixing if $\phi(n) \to 0$ as $n \to \infty$. Geometric ergodic Markov chains are exponentially fast α-mixing. For stationary sequences, mixing implies ergodicity. If $\phi(h) = O(h^{-a-\epsilon})$ for some $\epsilon > 0$, then ϕ is of order a. Similarly, if $\alpha(h) = O(h^{-a-\epsilon})$ for some $\epsilon > 0$, then α is of order a.

1.1.4 Strong Invariance Principle

Let $\{X_n\}_{n\in\mathbb{N}}$ be a stationary sequence of centered random variables and denote $S_n = \sum_{i=1}^{n} X_i$ to be some partial sums. A strong invariance principle states that the given sequence can be defined on a rich probability space so that for some Brownian process $\{W_t, t \geq 0\}$ and some specified $\sigma^2 > 0$, we have

$$S_n - \sigma W_n = O(\psi(n)) \text{ a.s.}$$

Theorem 1 *(Kuelbs and Philipp, 1980) Let $g(S_1), g(S_2), \ldots$ be an $\mathbb{R}^p$ -valued stationary process such that $\mathrm{E}_F\|g\|^{2+\delta} < \infty$ for some $0 < \delta \leq 1$. Let $\alpha_g(n)$ be the mixing coefficients of the process $\{g(S_t)\}$ and suppose, as $n \to \infty$,*

$$\alpha_g(n) = O\left(n^{-(1+\epsilon)(1+2/\delta)}\right) \qquad \text{for } \epsilon > 0.$$

Then, a strong invariance principle holds with $\psi(n) = n^{1/2-\lambda}$ for some $\lambda > 0$ depending on ϵ, δ, and p only.

Condition 2 *Let $\{B(t), t \geqslant 0\}$ be a p-dimensional standard Brownian motion. There exists a $p \times p$ lower triangular matrix L, a nonnegative increasing function ψ on the positive integers, a finite random variable D, and a sufficiently rich probability space*

such that, with probability 1, as $n \to \infty$

$$\|n(\theta_n - \theta) - LB(n)\| < D\psi(n). \tag{1.6}$$

Condition 3 *The batch size b satisfies the following conditions: (a) the batch size b is an integer sequence such that* $b \to \infty$*, and* $n/b \to \infty$ *as* $n \to \infty$*, where b and* n/b *are increasing; (b) there exists a constant* $c \geqslant 1$ *such that* $\sum_n (b_n n^{-1})^c < \infty$*.*

Assumption 1 *For some* $\delta > 0$*,* $\mathrm{E}\|Y_1\|^{2+\delta} < \infty$ *and there exists* $\epsilon > 0$ *such that* $\{X_t\}$ *is* α*-mixing with* $\alpha(n) = O\left(n^{-(4+\epsilon)(1+2/\delta)}\right)$*.*

1.1.5 Physical Dependence

Quantifying λ is still an open problem because it is very hard to quantify the correlation of a stochastic process. For slow mixing Markov chains, λ close to 0, and for fast mixing Markov chains λ close to 1/2. To capture the dependence structure in time series, Wu (2005) introduces a novel physical dependence measure assuming that the time series has the Markov form

$$Y_t = g(\ldots, \varepsilon_{t-1}, \varepsilon_t),$$

where $\{\varepsilon_t; t \in Z\}$ are i.i.d. random variables and g is a measurable function such that Y_t is well-defined. This alternate physical dependence measure is easy to use and is directly related to the underlying data-generating mechanism (Xiao and Wu, 2011).

Let $(\varepsilon_t', t \in \mathbb{Z})$ be an i.i.d. copy of $(\varepsilon_t, t \in \mathbb{Z})$ and let another time series started at this i.i.d. copy be $Y_t' := g(\varepsilon_t, \ldots, \varepsilon_1, \varepsilon_0', \varepsilon_{-1}, \varepsilon_{-2}, \ldots)$. If $\|Y_t\|_p := (\mathrm{E}\,|Y_t|^p)^{1/p} < \infty$ for some $p > 0$, the physical dependence measure is defined as

$$\Theta_p(m) = \sum_{t=m}^{\infty} \delta_p(t), \quad m \geq 0, \qquad \text{where } \delta_p(t) = \|Y_t - Y_t'\|_p . \tag{1.7}$$

The parameter $\delta_p(t)$ measures the impact of ε_0 on Y_t. It basically means that the impact of error terms e_t far into the past or future becomes negligible so that they can be treated as independent and the metric of this approximation is $\delta_p(t)$. Asymptotical results then typically depends on the rate of decay of $\delta_p(t)$ as $t \to \infty$. Also, consider a dependence adjusted and aggregated norm, resp. defined in Wu and Wu (2016) as

$$\|Y_{\cdot,j}\|_{p,\alpha} = \sup_{m \geqslant 0} (m+1)^{\alpha} \Theta_{p,j}(m) \qquad \text{and} \qquad \Psi_{p,\alpha} = \max_{j \in [d]} \|Y_{\cdot,j}\|_{p,\alpha} ,$$

where $\alpha \in (0, \infty)$ and d is the dimension of Y_t. These quantities are important to bound the bias of the variance estimators.

For some of the proofs, we assume a second order stationary linear time series as our underlying Markov model with the following assumption

$$Y_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}, \qquad t \in \mathbb{Z}, \tag{1.8}$$

where the coefficients b_j are absolutely summable and $\{\epsilon_k\}_{k \in \mathbb{Z}}$ is a mean zero i.i.d.

sequence of random variables, such that $\mathbb{E}\left(|\epsilon_k|^4\right) < \infty$ and $\mathbb{E}\left(\epsilon_k^2\right) = \sigma_\varepsilon^2 > 0$. When the time series contains nonlinear features, a local Gaussian assumption and strictly stationary assumption may be required in addition to the above set of assumptions (Jordanger and Tjøstheim, 2017). The spectral density for these process are usually approximated by an arbitrary order p spectral process.

We can see from (1.7) and (1.8) that $\delta_p(t) = |b_t| \, \|\varepsilon_1 - \varepsilon_1'\|$, so that for linear time series, decay rate of $\delta_p(t)$ translates to the assumptions on the decay rate of the coefficients b_t. Assuming $\sum_{h=-\infty}^{+\infty} |h| \, |b_h| < \infty$ or equivalently $b_h = O\left(|h|^{-2-\delta}\right), \delta > 0$ implies that the lag-h autocovariance $R(h) = \sigma_\varepsilon^2 \sum_{i=-\infty}^{+\infty} b_i b_{i+|h|}$ is $O\left(|h|^{-2-\delta}\right)$. For a stationary linear process, it follows that $\Theta_\alpha(h) = O\left(h^{-\beta}\right)$ for any $\beta > 0$ (Shao and Wu, 2007). For example, if $\delta_4(i) = K \, |b_i|$ for each i, then $\Theta_\alpha = K \sum_{i=0}^{\infty} |b_i| < \infty$, where $K = \|\varepsilon_i - \varepsilon_i'\|_4 < \infty$. Thus, short range or weak dependence assumption is satisfied if $\Theta_\alpha < \infty$.

1.2 Output Analysis

Let π be a distribution with support $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}^p$ be an π-integrable function such that $\theta = E_\pi(g)$. Let $\{X_t\}$ be an π-invariant Harris recurrent Markov chain, set $\{Y_t\} = \{g(X_t)\}$ and estimate θ with the ergodic averages $\overline{Y}_n = n^{-1} \sum_{t=1}^{n} Y_t$. The statistical process of estimating θ from the output of the Markov chain, $\{Y_n\}_{n \in \mathbb{N}}$, is called output analysis. The rate of convergence of the Markov chain to the desired stationary distribution dictates the quality of the estimates of θ for a given Monte

Carlo sample size.

Theorem 4 *(Jones, 2004) Suppose X is polynomially ergodic of order k , $\mathrm{E}_\pi|M(x)| < \infty$ and $\mathrm{E}_\pi\left|g_i(x)\right|^{2+\delta} < \infty$ for some δ such that $k\delta > 2+\delta$, then as $n \to \infty$, the approximate sampling distribution of the Monte Carlo error is available via a Markov chain Central limit theorem*

$$n^{1/2}(\overline{Y}_n - \theta) \xrightarrow{d} N_p(0, \Sigma), \tag{1.9}$$

provided there exists $p \times p$ positive-definite matrix, Σ.

1.2.1 Univariate Analysis

The univariate central limit theorem (CLT) approximates the asymptotic behavior of each component of the parameter vector θ. In a univariate analysis, we only consider one component at a time. Let $g^{(i)}, \theta_n^{(i)}$ and $\theta^{(i)}$ denote the i^{th} components of g, θ_n and θ, respectively. Then $\theta_n^{(i)} - \theta^{(i)}$ is the Monte carlo error of the i^{th} component that we want to estimate. If there exists $0 < \sigma_i^2 < \infty$, then as $n \to \infty$, the univariate Markov chain CLT gives the approximate sampling distribution of this error as

$$\sqrt{n}(\theta_n^{(i)} - \theta^{(i)}) \xrightarrow{d} N(0, \sigma_i^2), \tag{1.10}$$

where $\sigma_i^2 = \mathrm{Var}_\pi\{g^{(i)}(X_1)\} + 2\sum_{t=1}^{\infty}\mathrm{Cov}\{g^{(i)}(X_1), g^{(i)}(X_{1+t}))\}$.

The variance of the error in estimation consists of the variance of the chain if it were independent in addition to the covariance among the chain due to the dependence in the Markov chain. In order to assess the quality of estimation of $\theta^{(i)}$, estimates of σ_i^2 are needed. Consistent estimation of σ_i^2 has been studied extensively in literature. Flegal and Jones (2010) studied the BM and SV method of consistent estimation of σ_i^2.

In the BM method, the output is broken into blocks of equal size. Let $a = a_n$ be the number of batches and $b = b_n$ be the batch size. Then the run length of the chain is $n = ab$. The dependence of a and b on n is implicit and is suppressed for ease of notation. Asymptotically, the batch means $\bar{Y}_i$ are i.i.d. on the limit $n \to \infty$ and $b \to \infty$. Functional limit theorem says that the batch means are uncorrelated and normally distributed as $b \to \infty$ and a is fixed (Glynn and Whitt, 1991). For $l = 0, ..., a-1$ batches, define the mean for i^{th} component of Y and l^{th} batch to be

$$\overline{Y}_l(b) = \frac{1}{b}\sum_{t=1}^{b} Y_{lb+t}.$$

Then, the BM estimate of σ_i^2 is

$$\hat{\sigma}^2_{i,BM} = \frac{b}{a-1}\sum_{l=0}^{a-1}(\overline{Y}_l(b) - \overline{Y}_n)^2. \tag{1.11}$$

If we use batches that overlap, we get the overlapping batch means (OBM) estimator. Suppose there are $n - b + 1$ overlapping batches of length b and the mean for i^{th}

component of Y and l^{th} batch is

$$\dot{Y}_l(b) = \frac{1}{b}\sum_{t=1}^{b} Y_{l+t} \qquad \text{for} \quad l = 0, ..., n-b.$$

Then, the OBM estimator is defined as

$$\hat{\sigma}^2_{i,OBM} = \frac{nb}{(n-b)(n-b+1)}\sum_{l=0}^{n-b}(\dot{Y}_l(b) - \overline{Y}_n)^2. \tag{1.12}$$

There exists another class of estimators of σ_i^2 called spectral variance (SV) estimators. The lag k autocovariance is defined as $\gamma(k) = \gamma(-k) = E_\pi[Y_t Y_{t+k}]$, where Y_i's are centered observations i.e., $Y_i = g^{(i)}(X_i) - E_\pi g^{(i)}$. The sample estimator for the autocovariance at lag k is then given by

$$\gamma_n(k) = n^{-1}\sum_{t=\max(1,1-k)}^{\min(n,n-k)} (Y_t - \overline{Y}_n)(Y_{t+k} - \overline{Y}_n). \tag{1.13}$$

The sum of $\gamma_n(k)$ has been used to estimate σ_i^2 in literature for some time. Flegal and Jones (2010) investigated the truncated and weighted sum of (1.13) defined as

$$\hat{\sigma}^2_{i,SV} = \sum_{k=-(b-1)}^{b-1} w_n(k)\gamma_n(k), \tag{1.14}$$

where $w_n(\cdot)$ is the lag window and b is the truncation point.

Assumption 2 *Suppose (i)*$w(x) : \mathbb{R} \to [0, 1]$ *is symmetric, piece-wise smooth with*

$w(0) = 1$ *and* $\int_0^\infty w(x)xdx < \infty$*, (ii) the Parzen characteristic exponent defined by*

$$q = \max\left\{q_0 : q_0 \in \mathbb{Z}^+, g = \lim_{x\to 0} \frac{1 - w(x)}{|x|^{q_0}} < \infty\right\}$$

is greater than or equal to 1.

Some common lag-windows are:

$$\text{Bartlett:} w_n(k) = (1 - |k|/b)I(|k| \leq b)$$

$$\text{(Bartlett) Flat-top:} w_n(k) = I(|k| \leq b/2) + (2(1 - |k|/b))I(b/2 < |k| \leq b)$$

$$\text{Lugsail:} w_n(k) = \frac{1}{1-c}\left(1 - \frac{|k|}{b}\right)I(0 \leq k \leq b) - \frac{c}{1-c}\left(1 - \frac{|k|}{b/r}\right)I\left(0 \leq k \leq \frac{b}{r}\right)$$

$$\text{Tukey-Hanning:} w_n(k) = ((1 + \cos(\pi|k|/b))/2)I(|k| \leq b)$$

$$\text{Parzen:} w_n(k) = \begin{cases} 1 - 6k^2 + 6|k|^3 & \text{for } 0 \leq |k| \leq 1/2 \\ 2(1 - |k|)^3 & \text{for } 1/2 \leq |x| \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Quadratic Spectral:} w_n(k) = \frac{25}{12\pi^2k^2}\left(\frac{\sin(6\pi k/5)}{6\pi k/5} - \cos(6\pi x/5)\right).$$

For the Bartlett kernel, the Parzen characteristic exponent is 1 . For the Parzen and QS kernels, the Parzen characteristic exponent is 2 . For the Parzen kernel, $g = 6$. For the quadratic spectral kernel, $g = 18\pi^2/125$. Here, g is the limiting value of characteristic coefficient in assumption 1.

The effective sample size (ESS) is the number with the property that θ_n has the same precision as the sample mean obtained by the same number of independent and

identically distributed (i.i.d.) samples. The ESS for the i^{th} component is given by

$$\text{ESS}_i = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_i(k)} = n\frac{\lambda_i^2}{\sigma_i^2}. \tag{1.15}$$

If consistent estimators of σ_i^2 and λ_i^2 are used then ESS_i estimates are consistent. Sample variance $\lambda_{n,i}^2$ is a consistent estimator of λ_i^2. When the Markov chain is geometrically ergodic, the BM and oBM methods produce strongly consistent estimators of σ_i^2 (Jones et al., 2006; Flegal and Jones, 2010).

Univariate ESS is the most common approach in MCMC literature as it is fast and simple to calculate. ESS is normally used to devise a stopping rule to terminating the simulation in MCMC by pre-specifying a lower bound to the simulation run length for the components (Vats et al., 2019). The lower-bound is specified by the component with the smallest ESS i.e., the slowest mixing component in the Markov chain. So, this approach might cause delayed termination of the Markov chain when used as a stopping criteria.

1.2.2 Multivariate Analysis

In the mutivariate case, the approximate sampling distribution of the *Monte Carlo error*, $\overline{Y}_n - \theta$ is given by the Markov chain CLT. It states that if there exists a $p \times p$ positive definite symmetric matrix Σ, then the sampling distribution of the means

converges to

$$\sqrt{n}(\overline{Y}_n - \theta) \xrightarrow{d} N_p(0, \Sigma) \quad \text{as} \quad n \to \infty, \tag{1.16}$$

where

$$\Sigma = \text{Var}_\pi(Y_1) + \sum_{k=1}^{\infty} [\text{Cov}_\pi(Y_1, Y_{1+k}) + \text{Cov}_\pi(Y_1, Y_{1+k})^T]. \tag{1.17}$$

Multivariate CLT holds under the same conditions as a univariate CLT (Roberts and Rosenthal, 2004), (Jones et al., 2006). As opposed to the univariate analysis, Σ in (1.17) now contains the autocovariances of each of the components of the chain and the cross-covariances among the components of the chain.

Geometrically, the Markov chain CLT can also be written as

$$n(\theta_n - \theta)^T \Sigma_n^{-1} (\theta_n - \theta) \xrightarrow{d} T_{p,q}^2,$$

where $T_{p,q}^2$ is Hotelling's T-squared distribution with dimensionality parameter p and degrees of freedom q. If $T_{1-\alpha,p,q}^2$ is the $1-\alpha$ quantile of $T_{p,q}^2$, then a $100(1-\alpha)\%$ confidence region for θ is

$$C_\alpha(n) = \{\theta \in \mathbb{R}^p : n(\theta_n - \theta)^T \Sigma_n^{-1} (\theta_n - \theta) < T_{1-\alpha,p,q}^2\},$$

where Σ_n is a strongly consistent estimator of Σ and q is dependent on Σ_n. The volume of $C_\alpha(n)$ is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)}\left(\frac{T^2_{1-\alpha,p,q}}{n}\right)^{p/2}|\Sigma_n|^{1/2}. \tag{1.18}$$

The multivariate BM estimator is similar to the univariate case except that we have vectorized observations and means. If the mean vector is $\overline{Y}_l(b) = \frac{1}{b}\sum_{t=1}^{b} Y_{lb+t}$ for $l = 0, ..., a-1$ batches, the BM estimator of Σ is

$$\hat{\Sigma}_{BM} = \frac{b}{a-1}\sum_{l=0}^{a-1}(\overline{Y}_l(b) - \overline{Y}_n)(\overline{Y}_l(b) - \overline{Y}_n)^T. \tag{1.19}$$

Strong consistency for multivariate BM has been studied by Vats et al. (2019). Multivariate BM is consistent, easy to implement and is readily implemented under a R package called *mcmcse* (Flegal et al., 2015). Once we have a consistent estimator for the asymptotic variance of the sampling distribution we can form asymptoticaly valid confidence region around the estimates, $\overline{Y}_n$, to assess their reliability (Vats et al., 2019).

The confidence region consists of an ellipsoid in p dimensions oriented along the directions of the eigenvectors of Σ_n. The volume of the confidence region is proportional to $\sqrt{|\Sigma_n|}$, where $|\Sigma_n|$ is the estimated generalized variance of the Monte Carlo error, $|\Sigma|$. The volume of confidence region can be used to see if the simulation effort is big enough to achieve a desired level of precision in estimation (Jones et al., 2006;

Vats et al., 2019). However, BM method tends to underestimate the volume of confidence region unless the Markov chain is run for a large number of iterations (Vats et al., 2019). Batch size selection is an open problem but Liu et al. (2021) shows that asymptotically optimal batch size is proportional to $\lfloor n^{1/3} \rfloor$.

Multivariate overlapping batch means (OBM) estimators are also available in literature. If we use $n-b+1$ overlapping batches of length b and if the mean vector is $\dot{Y}_l(b) = \frac{1}{b}\sum_{t=1}^{b} Y_{l+t}$ for $l = 0, ..., n-b$ batches, then the OBM estimator of Σ is

$$\hat{\Sigma}_{OBM} = \frac{nb}{(n-b)(n-b+1)} \sum_{l=0}^{n-b} (\dot{Y}_l(b) - \overline{Y}_n)(\dot{Y}_l(b) - \overline{Y}_n)^T. \tag{1.20}$$

OBM method is computationally slower than BM since there are $n-b+1$ instead of $\lfloor \frac{n}{b} \rfloor$ batches (Flegal et al., 2008).

The sample autocovariance in vectorized form is given by

$$\hat{R}(s) = \frac{1}{n} \sum_{\max(1,1-s)}^{\min(n,n-s)} (Y_t - \overline{Y}_n)(Y_{t+s} - \overline{Y}_n)^T.$$

The sample autocovariances are unbiased estimators of population autocovariances upto $O(n^{-1})$. So, using them in finite samples do not impact the bias of an estimator. The multivariate SV estimator of Σ is a weighted and truncated sum of the lag s

sample autocovariances, similar to the univariate case in (1.14),

$$\hat{\Sigma}_{SV} = \sum_{s=-n+1}^{n-1} w_n\left(\frac{s}{b}\right)\hat{R}(s).$$

where $w_n(\cdot)$ is the lag window, b is the truncation point, and $\hat{R}(s)$ is the sample autocovariance. There are two sources of error in the above approximation. One is due to the truncation and another is due to down-weighting caused by the lag window. The lag weighting scheme yields negative bias, and such bias could be substantial in finite samples.

Liu and Flegal (2018) extend Flegal and Jones's (2010) work on OBM estimator and SV estimator and devise another estimator called weighted Batch Means (WBM) estimator that is computationally faster at higher dimensions. If we define a non-overlapping BM vector as $\overline{Y}_l(k) = \frac{1}{k}\sum_{t=1}^{k} Y_{lk+t}$ for $l = 0, 1, ..., a_k - 1$ batches and $k = 1, 2, ..., b$ batch sizes, where $a_k = \lfloor (n/k) \rfloor$, then the WBM estimator is defined as

$$\hat{\Sigma}^{WBM} = \sum_{k=1}^{b} \frac{1}{a_k - 1} \sum_{l=0}^{a_k-1} k^2 \Delta w_n(k)(\overline{Y}_l(k) - \overline{Y})(\overline{Y}_l(k) - \overline{Y})^T, \tag{1.21}$$

where $\Delta_2 w_n(k) = w_n(k-1) - 2w_n(k) + w_n(k+1)$ (Liu and Flegal, 2018).

A generalization to lag windows in broader time series and MCMC context called lugsail lag window was introduced in Vats and Flegal (2018). The lugsail lag window gives more than unit weight to small lag autocovariances and is an intuitive way

to adjust finite sample bias. Vats and Flegal (2018) show that using a lugsail lag window, the lugsail SV estimator is equivalent to the difference of two SV estimators

$$\dot{\Sigma}^L = \sum_{s=-(b-1)}^{b-1} w_n(k)\hat{R}(s) = \frac{1}{1-c}\hat{\Sigma}_b^{SV} - \frac{c}{1-c}\hat{\Sigma}_{b/r}^{SV}, \tag{1.22}$$

and using the lugsail lag window in (1.21) yields the following form of lugsail BM estimator

$$\hat{\Sigma}^L = \frac{1}{1-c}\hat{\Sigma}_b^{BM} - \frac{c}{1-c}\hat{\Sigma}_{b/r}^{BM}, \tag{1.23}$$

where $\hat{\Sigma}_b^{(\cdot)}$ and $\hat{\Sigma}_{b/r}^{(\cdot)}$ are estimators with integer batch sizes b and b/r respectively. The estimators in (1.22) and (1.23) have smaller bias due to bias cancellation. The amount of bias cancellation depends on the choice of r and c and there is variance inflation due to the linear combinations involved in the construction of the estimator.

1.3 Examples

A few time series examples and a real life example on lupus cancer was used in this thesis. The output of the Markov chain was approximated as a linear time series and estimated assuming well studied autoregressive processes such as AR(1), AR(p) and VAR(1). These examples are discussed briefly in the next subsections. First, define the quantities that are an underlying parameters of the process or the data generating

mechanism. The lag-weighted infinite sum of the autocovariances for univariate cases is

$$\Gamma = 2\sum_{h=1}^{\infty} hR(h).$$

Similarly, for the multivariate cases, the power lag weighted sum of mutivariate autocovariances is

$$\Gamma^{(q)} := -\sum_{h=1}^{\infty} h^q \left[R(h) + R(h)^T\right].$$

These values are negative for positively autocorrelated processes.

1.3.1 AR(1) Model

Consider the following autoregressive model of order 1 (AR(1)):

$$X_t = \phi X_{t-1} + \varepsilon_t \quad \text{for } t = 1, 2, \ldots \tag{1.24}$$

where ε_t are i.i.d N$(0, 1)$. If $|\phi| < 1$, the Markov chain is geometrically ergodic. Assuming the finite sample output from AR(1) process resembles a Markov chain, we can estimate $\theta = E[X_t]$ by $\theta_n = \bar{X}_n$. The true variance σ^2 and the quantity Γ is available for the AR(1) process so that we can easily calculate the bias and variance of the Markov chain. The variance and covariances of the observations of the Markov

chain from AR(1) process resp. are

$$\text{Var}\,(X_1) = \frac{1}{1-\phi^2} \quad \text{and}$$
$$\text{Cov}\,(X_1, X_s) = \frac{\phi^{s-1}}{1-\phi^2}.$$

The following quantities can be easily calculated for AR(1) process and can be used to assess the quality of estimation

$$\sigma^2 = \frac{1}{(1-\phi)^2} \quad \text{and} \tag{1.25}$$

$$\Gamma = \frac{2\phi}{(1-\phi^2)\,(1-\phi)^2}. \tag{1.26}$$

Derivations of (1.25) and (1.26) are provided in Section 1.4.

1.3.2 AR(p) Model

The stochastic process $\{X_t\}$ has an AR(p) representation if they can be written as

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t,$$

where ϵ_t are i.i.d. random variables with variance σ_e^2 and $E[|\epsilon_t|^{2+\delta}] < \infty$ for some $\delta > 0$. AR(p) process is a p-Markovian process. The autocovariances for AR(p) process can be derived and estimated consistently using the Yule-Walker equations where

$$\mathrm{E}\left(X_t X_{t-k}\right) = \sum_{j=1}^{p} \phi_j \mathrm{E}\left(X_{t-j} X_{t-k}\right).$$

It's easy to see that the above results in the homogeneous difference equation

$$R(k) - \sum_{j=1}^{p} \phi_j R(k-j) = 0. \tag{1.27}$$

This equation tells us that there are covariances at all lags for an AR(p) process. Given the covariances $R(k)$ satisfies $\sum_k k^2 R(k) < \infty$, the spectral density of the autocovariances is defined as

$$f(\omega) = \sum_k R(k) \exp(ik\omega).$$

The spectral density can be approximated to any order by the spectral density of an AR(p) process. Going in the opposite direction, the autocovariances can be obtained from the spectral density using the inverse Fourier transform

$$R(k) = \frac{1}{2\pi} \int_0^{2\pi} f(\omega) \exp(-ik\omega).$$

1.3.3 VAR(1) model

Consider the p-dimensional vector auto-regressive process of order 1 (VAR(1))

$$X_t = \Phi X_{t-1} + \epsilon_t,$$

for $t = 1, 2, \ldots$ where $X_t \in \mathbb{R}^p, \epsilon_t$ are i.i.d. $N_p(0, I_p)$ and Φ is a $p \times p$ matrix. The Markov chain is geometrically ergodic when the absolute value of the largest eigenvalue of Φ is less than 1 (Tjøstheim, 1990). In addition, if $\otimes$ denotes the Kronecker product, the invariant distribution is $N_p(0, V)$, where $\text{vec}(V) = \left(I_{p^2} - \Phi \otimes \Phi\right)^{-1} \text{vec}\left(I_p\right)$. Consider estimating $\theta = EX_t$ by $\theta_n = \bar{X}_n$. The following quantities are known to us from standard time series theory

$$\Sigma = \left(I_p - \Phi\right)^{-1} V + V \left(I_p - \Phi\right)^{-1} - V \quad \text{and} \tag{1.28}$$

$$\Gamma = -\left[\left(I_p - \Phi\right)^{-2} \Phi V + V \Phi^T \left(I_p - \Phi^T\right)^{-2}\right]. \tag{1.29}$$

Derivations of (1.28) and (1.29) are given in Section 1.4.

1.3.4 Bayesian Probit Regression

The lupus data available from Van Dyk and Meng (2001) contains disease statuses for 55 patients, 18 of whom have been diagnosed with latent membranous lupus, along

with two clinical covariates, IgA (immunoglobulin A) and ΔIgG ($IgG3 - IgG4$). We consider a probit regression model for the sampling distribution the response variable, y_i, an indicator of the disease with the predictor variables x_{i1} (ΔIgG) and x_{i2} (IgA), where $i = 1, \ldots, 55$. For each patient, y_i is modeled as independent Bernoulli random variable as

$$Y_i \sim \text{Bernoulli}\left(\Phi\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right)\right),$$

with flat prior on the parameters $\beta = (\beta_0, \beta_1, \beta_2)$. We want to estimate the posterior expectation of $\beta, E_\pi \beta$. We use PX-DA algorithm of Liu and Wu (1999) to sample from $\pi(\beta|y)$. Let $TN(\mu, \sigma^2, \omega)$ denote the distribution of a truncated normal variable with mean μ and variance σ^2 that is truncated to be positive if $\omega = 1$ and negative if $\omega = 0$. The Algorithm consists of the following steps:

---

1. Draw $z_1, \ldots, z_{55}$ independently with $z_i \sim \text{TN}\left(x_i^T \beta, 1, y_i\right)$
2. Draw $g^2 \sim \text{Gamma}\left(\frac{55}{2}, \frac{1}{2}\sum_{i=1}^{55}\left[z_i - x_i^T\left(X^T X\right)^{-1} X^T z\right]^2\right)$ and set $z' = (gz_1, \ldots, gz_{55})^T$
3. Draw $\beta' \sim \text{N}\left(\left(X^T X\right)^{-1} X^T z', \left(X^T X\right)^{-1}\right)$.

1.4 Derivations

Derivation of (1.25) The variance of AR(1) process is calculated as follows. We have

$$\begin{aligned}
\sigma^2 &= \mathrm{Var}\left[X_1\right] + 2\sum_{s=1}^{\infty} \mathrm{Cov}\left(X_t, X_{t+s}\right) \\
&= \frac{1}{1-\phi^2} + 2\sum_{s=1}^{\infty} \frac{\phi^s}{1-\phi^2} \\
&= \frac{1}{1-\phi^2} + 2\lim_{n\to\infty} \sum_{s=1}^{n} \frac{\phi^s}{1-\phi^2} \\
&= \frac{1}{1-\phi^2} + 2\lim_{n\to\infty} \left[\frac{\phi\left(1-\phi^{n-1}\right)}{1-\phi}\right] \\
&= \frac{1}{1-\phi^2} + \frac{2}{1-\phi^2} \cdot \frac{\phi}{1-\phi} \\
&= \frac{1+\phi}{\left(1-\phi^2\right)\left(1-\phi\right)} \\
&= \frac{1}{(1-\phi)^2}.
\end{aligned}$$

Derivation of (1.26) Similarly, the quantity Γ can be derived as follows. We have

$$\begin{aligned}
\Gamma &= 2\sum_{s=1}^{\infty} s \cdot \mathrm{Cov}\left(X_t, X_{t+s}\right) \\
&= 2\lim_{n\to\infty}\sum_{s=1}^{n} s \cdot \frac{\phi^s}{1-\phi^2} \\
&= \frac{2}{1-\phi^2}\lim_{n\to\infty}\sum_{s=1}^{n} s\phi^s \\
&= \frac{2}{1-\phi^2}\lim_{n\to\infty}\sum_{s=1}^{n}\left[\frac{\phi\left(1-\phi^{n-1}\right)}{(1-\phi)^2} - \frac{n\phi^{n+1}}{1-\phi}\right] \\
&= \frac{2}{1-\phi^2}\left[\frac{\phi}{(1-\phi)^2}\right] \\
&= \frac{2\phi}{\left(1-\phi^2\right)(1-\phi)^2}.
\end{aligned}$$

Derivation of (1.28) We have

$$\mathrm{vec}(V) = \left(I_{p^2} - \Phi\otimes\Phi\right)^{-1}\mathrm{vec}(\Omega)$$

Then

$$\begin{aligned}
Y_s &= \Phi Y_{s-1} + \epsilon_s \\
&= \Phi \left(\Phi Y_{s-2} + \epsilon_{s-1} \right) + \epsilon_s \\
&= \Phi^2 Y_{s-2} + \Phi \epsilon_{s-1} + \epsilon_s \\
&= \Phi^2 \left(\Phi Y_{s-3} + \epsilon_{s-2} \right) + \Phi \epsilon_{s-1} + \epsilon_s \\
&\vdots \\
&= \Phi^s Y_0 + \Phi^{s-1} \epsilon_1 + \Phi^{s-2} \epsilon_2 + \cdots + \Phi^2 \epsilon_{s-2} + \Phi \epsilon_{s-1} + \epsilon_s.
\end{aligned}$$

So, we get

$$\begin{aligned}
R(s) &= \operatorname{Cov}\left(Y_0, Y_s \right) \\
&= \operatorname{Cov}\left(Y_0, \Phi^s Y_0 + \Phi^{s-1} \epsilon_1 + \Phi^{s-2} \epsilon_2 + \cdots + \Phi^2 \epsilon_{s-2} + \Phi \epsilon_{s-1} + \epsilon_s \right) \\
&= \operatorname{Cov}\left(Y_0, \Phi^s Y_0 \right) \\
&= \Phi^s \operatorname{Cov}\left(Y_0, Y_0 \right) \\
&= \Phi^s V
\end{aligned}$$

Finally,

$$\begin{aligned}
\Sigma &= \sum_{s=-\infty}^{\infty} R(s) \\
&= \sum_{s=0}^{\infty} R(s) + \sum_{s=-\infty}^{0} R(s) - V \\
&= \sum_{s=0}^{\infty} \Phi^s V + \sum_{s=-\infty}^{0} V \left(\Phi^T\right)^s - V \\
&= (I_p - \Phi)^{-1} V + V (I_p - \Phi)^{-1} - V.
\end{aligned}$$

1.5 Concluding Remarks

Markov chains can be approximated by various autoregressive processes where rate of decay autocovariances have been carefully investigated. We need to carefully analyze that these dependencies vanish sufficiently fast, so that processes far away from each other in terms of time or lags are independent of each other. Various mixing conditions and dependence measures exists to quantify this issue, but there is no single solution available. Mean and variance of the output from a Markov process is of major interest and their statistical properties should be studied in detail with application to real-life examples and practical settings. Accurate variance estimation methods are crucial to assess the reliability of the point estimates constructed from such processes.

Stationarity is one of the prime assumptions, and we need to ensure our Markov

chain starts from the stationary distribution. Polynomial ergodicity, albeit being a weak assumption, is very practical in real applications. We try to identify research venues that satisfy this condition and study the finite sample properties of variance estimators that are constructed under this assumption. To this end, we also employ various statistical concepts from spectral theory of estimation of the spectral density at zero frequency.

Chapter 2

Accurate Variance Estimation

BM type estimators discussed in Chapter 1 are computationally efficient but suffer from systematic bias that affect coverage probabilities in finite sample simulations. OBM estimator is slower than BM estimator because it involves averaging over more batches. OBM estimator is asymptotically equivalent to SV estimator except for some end effects (Flegal and Jones, 2010). A careful analysis of the bias of these estimators is warranted for efficient use of these estimators in finite samples and is the main theme of this thesis.

The bias property of the estimators can be optimized for desired applications with careful selections of batch sizes as batch sizes are an inherent parameter of the variance estimation process. Finding a single solution to compute batch sizes is still an open problem because batch size is process dependent and is always sub-optimal in practice. Asymptotically optimal batch size can be used in finite sample MSE and

a non-optimal batch size can be appropriate if the associated MSEs are close (Song and Schmeiser, 1995). So, in lieu of this observation, we do not pursue batch size estimation vigorously in this thesis. Interested readers can see Liu et al. (2021) for further information.

Although these variance estimators are already established to be asymptotically unbiased and consistent, practitioners have had to deal with substantial bias in finite samples. The bias is especially significant in a highly correlated Markov chain and not accounting for the bias can skew our results drastically. Throughout this chapter, we are interested in quantifying the error in estimating $\theta = \int g(x)\pi(dx)$ with $\bar{Y} = n^{-1}\sum_{t=1}^{n} Y_t$. We know, $\Sigma = \sum_{s=-\infty}^{\infty} R(s) = \lim_{n\to\infty} n\,\mathrm{Var}_\pi(\bar{Y})$. There are various types of estimators used to estimate Σ as discussed in Chapter 1. We will take a multivariate approach to the analysis and any univariate implications will be discussed accordingly.

Assumption 3 *Let $Y_t, t \in \mathbb{Z}$ be a second order stationary stochastic process with the following summability condition on the autocovariances*

$$\sum_{h\in\mathbb{Z}} (1+|h|)^r |R(h)| < \infty$$

for some $r \geq 0$.

Assumption 3 applies to univariate autocovariances. The condition $r > 0$ ensures the absolute summability of the autocovariances. Also, r represents the smoothness

of the spectral density of the autocovariances. These basically tells us how fast the autocovariances decay to zero. When $r = 1$, assumption 3 says that both $\sum_{h\in\mathbb{Z}} |R(h)|$ and $\sum_{h\in\mathbb{Z}} h|R(h)|$ are finite. Multivariate version of this quantity will be introduced later.

The bias properties of the estimators are studied in detail in Vats and Flegal (2018). Define the following quantity

$$\Gamma^{(q)} := -\sum_{s=1}^{\infty} s^q \left[R(s) + R(s)^T\right], \tag{2.1}$$

with components Γ_{ij}, and let Γ^1 be denoted by Γ. This is a multivariate quantity, where each entry in the matrix contain the weighted sum of cross-covariances at different lags. We are mainly interested in the diagonal component of this matrix. Polynomial ergodicity or m-dependence and appropriate moment conditions are required to ensure $\Gamma^{(q)}$ and Σ are finite. Let the finite version of Γ^q be defined as

$$\Gamma_k^{(q)} := -\sum_{s=1}^{k} s^q \left[R(s) + R(s)^T\right] \tag{2.2}$$

where $s = 0, 1, 2\ldots$ and $k = 1, 2, \ldots$.

Liu et al. (2021) devises a method of estimating Γ assuming a parametric AR(p)

model. The exact form of the estimator of Γ is given by

$$\Gamma = 2\left[\left(\sum_{i=1}^{p}\phi_i\sum_{k=1}^{i}kR(k-i)\right)+\frac{(\sigma_e^2-R(0))}{2}\left(\sum_{i=1}^{p}i\phi_i\right)\right]\left(\frac{1}{1-\sum_{i=1}^{p}\phi_i}\right). \quad (2.3)$$

In finite samples, Γ in (2.3) is underestimated and using it results in sub-optimal batch sizes under MSE loss. The asymptotic MSE of BM and OBM estimators can be summarized as

$$\text{MSE}[\hat{\Sigma}] = \frac{C\Gamma^2}{b^2}+\frac{2S\Sigma^2 b}{n}+o\left(\frac{1}{b^2}\right)+o\left(\frac{b}{n}\right), \quad (2.4)$$

where $C, S = 1$ for BM and $C = 1, S = 2/3$ for OBM estimators (Flegal, 2008).

The MSE in (2.4) is also under-estimated. MSE optimal variance estimators also suffer from boundary problem where they are naturally attracted to 0 and this gives us a false sense of a smaller MSE. Further, the optimal batch size or lag truncation parameter that is the minimizer of (2.4) is also underestimated. The underestimated batch size when used in finite sample simulations produces less than optimal results. Asymptotically optimal batch size decays at the rate of $o(n^{1/3})$ (Flegal, 2008).

Following the suggestion in Andrews (1991), we can fit a parametric AR(p) model to estimate the autocovariances or their spectral density. Suppose we fit a parametric AR(p) second order stationary model in L^2- sense to the spectral density $f_p(\omega)$ with the corresponding AR-based autocovariance functions, $R_p(s)$. For lags higher than p, the autocovariances are obtained by iterating the difference equation that the fitted

model implies for its autocovariance using (1.27).

The autoregressive coefficients, a_k, can also be estimated from the fitted AR(p) model and the summability condition on the autocovariances of assumption 3 carries over to the AR coefficients by the following observation from Braumann et al. (2021)

$$\sum_{k=1}^{\infty}(1+|k|)^r\,|a_k| < \infty. \tag{2.5}$$

As the model order p approach infinity, the AR parameters converge to the parameters corresponding to the infinite order AR approximation in the mean square (Gupta and Mazumdar, 2012). Mean square convergence of the AR parameters under finite model order as the number of observations go to infinity was studied in Bhansali (1981). A time series fitted with AR(p) structure with a finite order p can achieve autocovariance estimation with a parametric rate of convergence of $\sqrt{n}$ (McMurry and Politis, 2015).

2.1 Single estimators

Single variance estimators are practitioners' first estimator of choice as they are natural, easy to interpret and computationally fast to calculate. BM estimators and OBM estimators are commonly used in MCMC, time series, stochastic simulations literature (Flegal and Jones, 2010), (Chan and Yau, 2017), (Song and Schmeiser, 1995). The asymptotic bias properties of these estimators are well-studied. For

example, the asymptotic bias of BM estimator is

$$\mathrm{E}\left[\hat{\Sigma}_b^{BM}\right] = \Sigma + \frac{\Gamma}{b} + o\left(\frac{1}{b}\right). \tag{2.6}$$

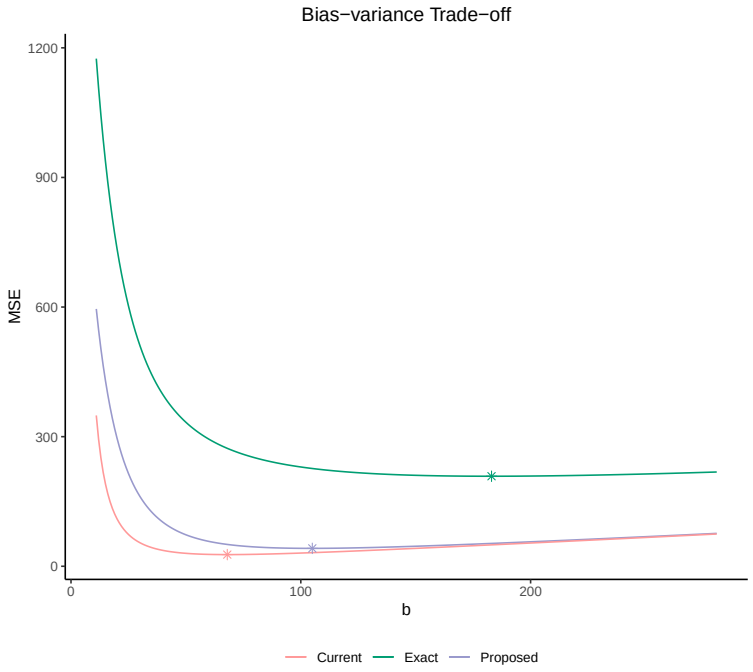

Figure 2.1: Bias and variance trade-off of BM estimators in finite sample simulation, $n = 2e4$.

BM estimator is asymptotically first-order unbiased. The parameter Γ is negative under positive correlation and is underestimated in finite samples. This results in under-coverage of credible intervals and incorrect sized hypothesis tests when these quantities are used as sample statistics. So, finding higher-order bias terms is crucial in finite sample MCMC other time series applications. MSE optimal batch sizes that

take into effect the higher-order bias seem to perform better in finite samples.

Taking into consideration of higher-order bias of BM estimators, we can see in Figure 2.1 that at low lags the proposed estimation method account for more bias than the current estimation method of (2.6). The bias decays quickly as the lag grows and is more influential in the low lag regions. The exact and proposed estimation methods will be discussed shortly in the next section. The exact bias of estimator has systematic error in addition to sampling bias. We can correct for the sampling bias but the systematic bias is always persistent. In finite samples, the sampling bias does not converge to zero. Also evident from Figure 2.1 is that the variance of the exact and proposed estimation methods have similar rate of growth at large lags. The MSE optimal batch sizes from the proposed methods are closer to the exact methods than the current methods. It is important to note that the exact methods also has sampling bias. We see similar behavior for OBM estimators and their discussions are skipped in this section.

Accurate variance estimators are needed to construct valid confidence regions around our estimates as discussed in Chapter 1. The volume of confidence region is sensitive upon the generalized variance and using higher-order correct estimators we can achieve volume inflation. Current estimators of variance underestimate the variance, while pilot estimators constructed with MSE optimal batch-sizes from the proposed methods are shown to be more accurate. The relative volume of the confidence region constructed from the proposed methods are close to unity, while that

for the current methods is vastly underestimated at different run lengths as shown in Figure 2.2. It should be noted that the relative volumes are calculated relative to the true volume calculated from the known values of Σ.

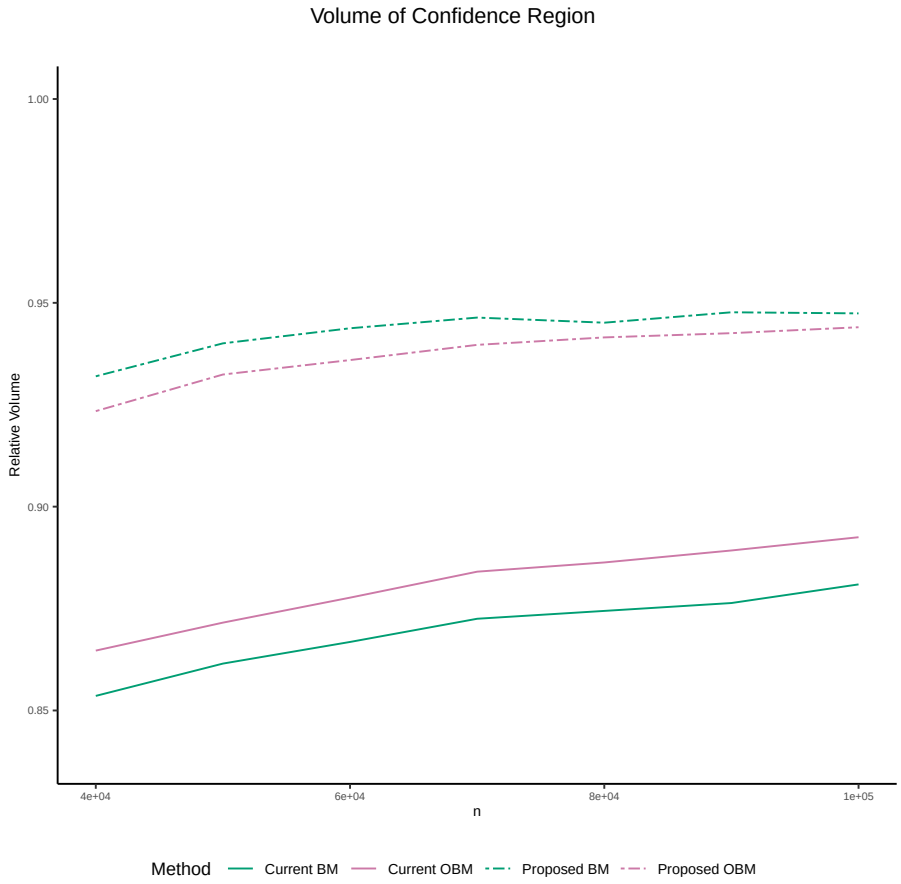

Figure 2.2: Volume of confidence regions for the current and proposed BM and OBM methods.

2.2 Quadratic Coefficients

We can calculate the exact bias of some common estimators using an important concept called quadratic coefficients, first mentioned in Song and Schmeiser (1993). We can also visually analyze the variance properties of estimators of the variance of the sample mean using quadratic coefficients . Quadratic coefficients are the weights given to each covariances involved in the computation of finite sample expectation of the variance estimator. For example, if $\hat{\Sigma}$ is a quadratic-form estimator of $n\,\mathrm{Var}_F(\bar{Y})$, then the expected value of $\hat{\Sigma}$ is $\mathrm{E}(\hat{\Sigma}) = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}\mathrm{E}\left(Y_iY_j\right)$ and q_{ij}'s are the quadratic coefficients.

The bias and variance results of an estimator typically depend on the sample size n and batch size b. We know, $\Sigma = \sum_{s=-\infty}^{\infty} R(s) = \lim_{n\to\infty} n\,\mathrm{Var}_F(\bar{Y})$. Then the BM quadratic form estimator of Σ can be written as

$$\hat{\Sigma}_b^{BM} = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}^{BM} Y_iY_j, \tag{2.7}$$

where $Y_1, Y_2, \ldots, Y_n$ are observations as discussed in Chapter 1. The quadratic coefficients for BM estimators can be summarized as

$$q_{ij}^{BM} = \frac{1}{d^{BM}}\left(\frac{a_{ij}}{b^2} - \frac{a_{ii}+a_{jj}}{nb} + \frac{1}{nb}\right),$$

where $d^{BM} = ((n-b)/b^2)$. The derivation for the exact expression for the quadratic

coefficients is straightforward with the observation that $a_{ij} = 1$ if the observations come from the same batch, and $a_{ij} = 0$ otherwise. The coefficients q_{ij}^{BM} can be written as

$$q_{ij}^{BM} = \begin{cases} \frac{1}{n}, & \text{if} \quad i = 1, 2, ..., ab \quad \& \quad j = f(i), ..., l(i), \\ \frac{-1}{a(n-b)}, & \text{o.w,} \end{cases} \tag{2.8}$$

where $l(i) = \lceil i/b \rceil \cdot b$ is the last observation in the batch containing X_i and $f(i) = l(i) - b + 1$ is the first observation in the batch containing X_i (Song and Schmeiser, 1993). If n observations can be divided into a batches of size b each, then we can express the BM estimator in an alternate form as

$$\hat{\Sigma}^{BM} = \frac{1}{d^{BM}} \sum_{i=1}^{a} (\bar{Y}_{b(i-1)+1} - \bar{Y})^2, \tag{2.9}$$

where $a = \lfloor n/b \rfloor$, $d^{BM} = ((n-b)/b^2)$. The i^{th} batch mean is given by

$$\bar{Y}_{b(i-1)+1} = b^{-1} \sum_{j=1}^{b} Y_{b(i-1)+j},$$

which is the sample mean of b observations starting from $Y_{(i-1)b+1}$. Using BM estimator in this form, we can easily calculate the exact bias of the BM estimator. The

exact bias of BM estimator is

$$E(\hat{\Sigma}^{BM}) = R(0) + \frac{n}{n-b}\Gamma^0_{b-1} - \frac{b}{n-b}\Gamma^0_{n-1} + \frac{b}{n(n-b)}\Gamma^1_{b-1} - \frac{n}{b(n-b)}\Gamma^1_{n-1}. \quad (2.10)$$

The proof of the bias expression (2.10) is given in Section 2.6. We have used the following standard notations to quantify the univariate autocovariances in (2.10)

$$\Gamma^0_{k-1} = 2\sum_{s=1}^{k-1} R(s) \qquad \Gamma^j_{k-1} = 2\sum_{s=1}^{k-1} s^j R(s),$$

and these are univariate analogues of the multivariate quantities in (2.1) and (2.2). We do not pursue multivariate estimation in this thesis and is set for future work.

Using similar assumptions on the second-order stationarity of the time series but without using quadratic coefficients, Aktaran-Kalaycı et al. (2007) calculates the exact same result, but the proof based on quadratic coefficient has not been derived yet in the literature. The bias can be further approximated as

$$\text{Bias}\left(\hat{\Sigma}^{BM}_b\right) = E[\hat{\Sigma}^{BM}_b] - \Sigma = -\left[\left(\Gamma^0_{n-1} - \Gamma^0_{b-1}\right) + \frac{1}{b}\Gamma^1_{b-1}\right] = \text{I} + \text{II} + \text{III}. \quad (2.11)$$

We have assumed the rate of decay of the autocovariances as $R(h) = O(h^{-2-\delta})$ throughout the thesis. With this assumption, we can approximate the order of each term involved in (2.11), i.e., the rate of decay of the largest term in the summation.

We can see that the rate of decay of I is

$$\begin{aligned}
-\Gamma_{n-1}^{0} &= 2\sum_{i=n}^{\infty} R(i) - (\Sigma - R(0)) \\
&= 2\sum_{i=n}^{\infty} i^{-2-\delta} - (\Sigma - R(0)) \\
&= O\left(\frac{1}{n^2}\right).
\end{aligned}$$

Similarly, the rate of decay of II is

$$\begin{aligned}
\Gamma_{b-1}^{0} &= (\Sigma - R(0)) - 2\sum_{i=b}^{\infty} R(i) \\
&= (\Sigma - R(0)) - 2\sum_{i=b}^{\infty} i^{-2-\delta} \\
&= -O\left(\frac{1}{b^2}\right).
\end{aligned}$$

And, finally the rate of decay of III is

$$\begin{aligned}
-\frac{2}{b}\Gamma_{b-1}^{1} &= -\frac{1}{b}\left(\sum_{i=1}^{\infty} iR(i) - \sum_{i=b}^{\infty} iR(i)\right) \\
&= \frac{2}{b}\sum_{i=b}^{\infty} i^{-1-\delta} - \frac{\Gamma}{b} \\
&= O\left(\frac{1}{b^2}\right) - O\left(\frac{1}{b}\right).
\end{aligned}$$

Combining I, II, and III, we see that the bias is a combination of first-order and

higher-order terms in finite samples.

$$\text{Bias}\left(\hat{\Sigma}_b^{BM}\right) = -\left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1\right] = O\left(\frac{1}{n^2}\right) - O\left(\frac{1}{b^2}\right) + O\left(\frac{1}{b^2}\right) - O\left(\frac{1}{b}\right).$$

Theorem 5 *Let g be such that $E_F\left(\|g\|^{2+\delta}\right) < \infty$ for some $\delta > 0$. Let $\{X_n\}$ be an π-invariant polynomially ergodic Markov chain of order $m > (4+\epsilon)(1+2/\delta)$ for some $\epsilon > 0$. Then (1.6) holds with $\gamma(n) = n^{1/2-\lambda}$ for some $\lambda > 0$. Further, if condition 3 holds, then the bias of $\hat{\Sigma}_b^{BM}$ is given by*

$$Bias\left(\hat{\Sigma}_b^{BM}\right) = \left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1\right] + O\left(\frac{1}{b}\right). \tag{2.12}$$

The proof of Theorem 5 is given in Section 2.6.

Lemma 6 *(Andrews, 1991, Lemma 1) Suppose $\{X_t\}$ is a mean zero α-mixing sequence of r.v.'s. If $\sup_{t\geqslant 1} E\|X_t\|^{4\nu} < \infty$ and $\sum_{j=1}^{\infty} j^2\alpha(j)^{(\nu-1)/\nu} < \infty$ for some $\nu > 1$, then $\sum_{j=0}^{\infty}\sup_{t\geqslant 1}\left\|EX_tX_{t+j}'\right\| < \infty$ and*

$$\sum_{j=1}^{\infty}\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\sup_{t\geqslant 1}|\kappa_{abcd}(t,t+j,t+m,t+n)| < \infty \quad \forall a,b,c,d \leqslant p$$

Assumption 4 *For some $\delta > 0$, $\mathrm{E}\|Y_1\|^{4+\delta} < \infty$ and there exists $\epsilon > 0$ such that $\{X_t\}$ is α-mixing with $\alpha(n) = o\left(n^{-(3+\epsilon)(1+4/\delta)}\right)$.*

Lemma 6 is satisfied under assumption 4.

Theorem 7 *Let g be such that $E_F\left(\|g\|^{4+\delta}\right) < \infty$ for some $\delta > 0$. Let $\{X_n\}$ be an π-invariant polynomially ergodic Markov chain and under lemma 6, condition 3 with $b = n^{1/2}$, for $r \geq 1$, the asymptotic covariance of two BM estimators with integer batch sizes b and b/r, resp., is*

$$Cov(\hat{\Sigma}_b^{BM}, \hat{\Sigma}_{b/r}^{BM}) = \frac{2b\Sigma^2}{rn}\left[1 + O\left(\frac{1}{b}\right) + O\left(\frac{b}{n}\right)\right]. \tag{2.13}$$

Corollary 8 *Under the same assumptions of Theorem 7, the asymptotic variance of BM estimator with integer batch sizes b is*

$$Var(\hat{\Sigma}_b^{BM}) = \frac{2\Sigma^2 b}{n}\left[1 + O\left(\frac{1}{b}\right) + O\left(\frac{b}{n}\right)\right]. \tag{2.14}$$

The proof of Theorem 7 and Corollary 8 is given in Section 2.6. As r increases, the covariance becomes smaller as expected. Similarly, the asymptotic correlation is,

$$\lim_{b\to\infty} \quad \{\text{Corr}(\hat{\Sigma}_b^{BM}, \hat{\Sigma}_{b/r}^{BM})\} = \frac{1}{\sqrt{r}}. \tag{2.15}$$

Theorem 9 *(Pedrosa, 1994, Theorem 4.1) Let g be such that $E_F\left(\|g\|^{4+\delta}\right) < \infty$ for some $\delta > 0$. Let $\{X_n\}$ be an π-invariant polynomially ergodic Markov chain and under lemma 6, condition 3 with $b = n^{1/2}$, for finite $r \geq 1$, the asymptotic covariance*

of two OBM estimators with integer batch sizes b and b/r, resp., is

$$Cov(\hat{\Sigma}_b^{OBM}, \hat{\Sigma}_{b/r}^{OBM}) = \frac{4b\Sigma^2}{3rn}\left(\frac{3}{2} - \frac{1}{2r}\right)\left[1 + O\left(\frac{b}{n}\right) + O\left(\frac{1}{b}\right)\right]. \tag{2.16}$$

Corollary 10 *Under the same assumptions of Theorem 9, the asymptotic variance of OBM estimator with integer batch sizes b is*

$$Var(\hat{\Sigma}_b^{OBM}) = \frac{4\Sigma^2 b}{3n}\left[1 + O\left(\frac{1}{b}\right) + O\left(\frac{b}{n}\right)\right].$$

This result is equal to 2/3 of the variance of BM estimator. As r increases, the covariance becomes smaller as expected.

$$\lim_{b\to\infty} \quad \{\mathrm{Corr}(\hat{\Sigma}_b^{OBM}, \hat{\Sigma}_{b/r}^{OBM})\} = \sqrt{\frac{3}{2} - \frac{1}{2r}}$$

The exact bias of OBM estimator is given in Aktaran-Kalaycı et al. (2007). The approximate form of the bias for OBM estimator is

$$\mathrm{E}[\hat{\Sigma}^{OBM}] = \Sigma - \left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{a\Gamma_{b-1}^1}{n-b} - \frac{bn\left(\Gamma_{b-1}^0 - \Gamma_{n-b}^0\right)}{(n-b)(n-b+1)}\right] + O\left(\frac{1}{b}\right). \tag{2.17}$$

The exact expression is very involving and is stated in Section 2.6 and the approximate bias in (2.17) contains higher-order terms and is an improvement to the current asymptotically first-order bias expression.

2.3 Linear Combination Variance Estimators

Linear combination of variance estimators are known to improve upon the bias properties at the cost of increase in variance, see for e.g. Pedrosa and Schmeiser (1993), Aktaran-Kalaycı et al. (2009), Gupta et al. (2014). In finite samples, however, the bias of these estimators is more important than the variance. The choice of batch sizes and weights can be done with the motivation to lower the bias in such a way that it offsets the systematic bias that come with the estimation process.

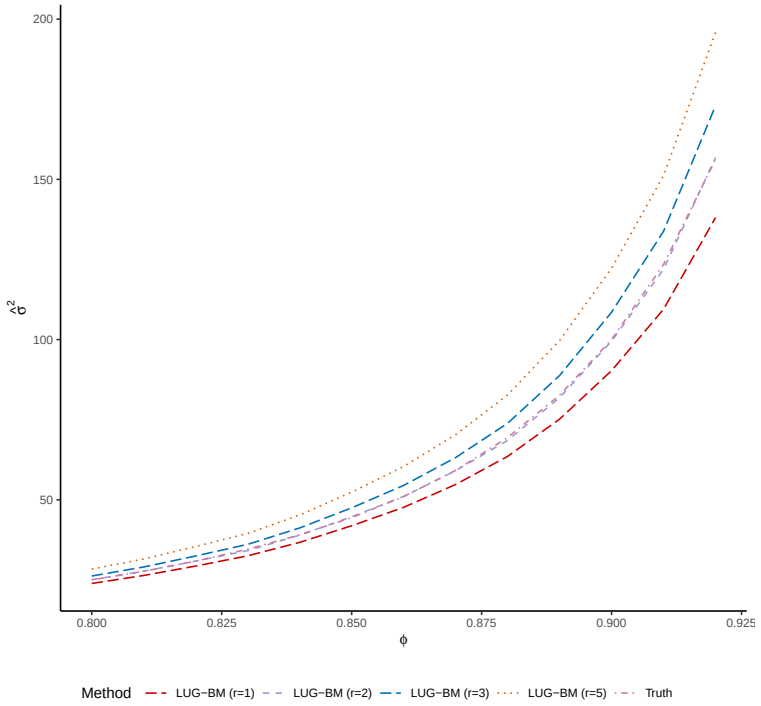

Figure 2.3: Different variance estimation methods for $n = 1e5$ iterations.

Simonoff (1993) contends that in practical applications where bias is more crucial, we should choose estimators that have smaller bias even if those estimators have large

variance. So, it is highly desirable for the estimator to be asymptotically unbiased even if the cost is additional variance. Linear combination of estimators account for the higher order bias, and they need significantly fewer lags to do so, reducing our computational costs in the process. The linear combination (LC) estimator with 3 components can be expressed as

$$\hat{\Sigma}^{LC} = \alpha_1 \hat{\Sigma}_b + \alpha_2 \hat{\Sigma}_{b/r} + \alpha_3 \hat{\Sigma}_{b/s}, \tag{2.18}$$

where are α_1, α_2, and α_1 are some constants such that $\alpha_1+\alpha_2+\alpha_3 = 1$ and $0 < r \leq s$. If we are motivated to reduce the systematic bias for BM type estimator, then such estimators can be constructed as a linear combination for some $0 \leq c_1, c_2 \leq 1$ as follows

$$\begin{aligned}\hat{\Sigma}^{LC} &= \frac{\frac{1}{1-c_1}\hat{\Sigma}_b - \frac{c_1}{1-c_1}\hat{\Sigma}_{b/r} - c_2\hat{\Sigma}_{b/s}}{1-c_2} \\ &= \frac{1}{(1-c_1)(1-c_2)}\hat{\Sigma}_b - \frac{c_1}{(1-c_1)(1-c_2)}\hat{\Sigma}_{b/r} - \frac{c_2}{1-c_2}\hat{\Sigma}_{b/s}.\end{aligned}$$

The bias of various estimators in finite sample simulations of an AR(1) process can be seen in Figure 2.3 and Figure 2.4 for batch sizes $b = n^{1/2}$ and $b = n^{1/3}$, resp. In Figure 2.3, with larger batch sizes, there is more variability in our estimators, especially at high correlation. Linear combination estimators are largely over-biased and single estimators are under-biased. Same trends are apparent in Figure 2.4 except

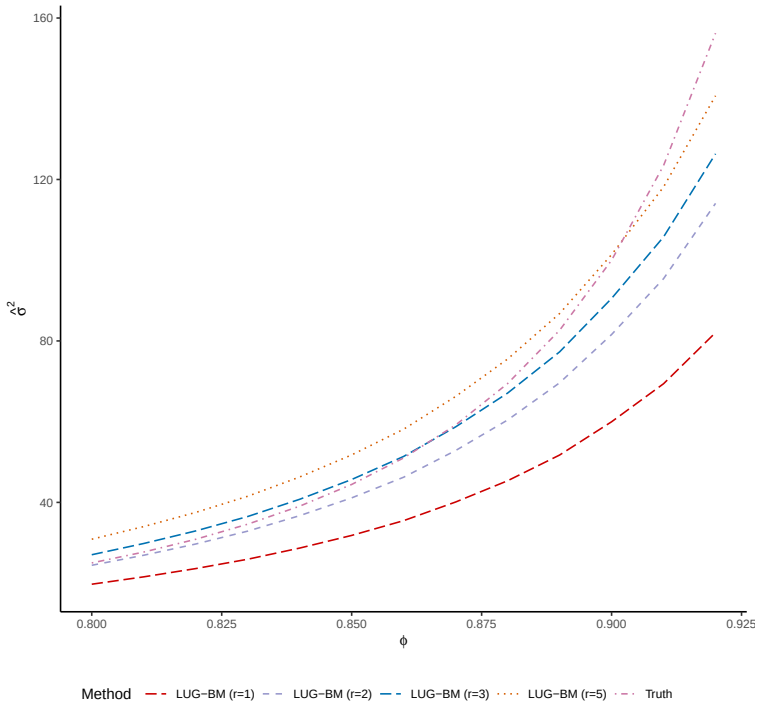

Figure 2.4: Different variance estimation methods for $n = 1e5$ iterations.

that the variability of the estimators is small due to smaller batch sizes. The weighted BM estimator in Figure 2.4 refers to Lugsail BM estimator with $r = 2$.

Asymptotically, $\hat{\Sigma}^{LC}$ is estimating the true variance as the coefficients add up to 1, but in finite samples its higher-order bias could be substantial. If $c_1 = c_2 = c$, then the bias of the $\hat{\Sigma}^{LC}$ can be calculated as

$$Bias(\hat{\Sigma}^{LC}) = \frac{\Gamma}{b}\left[\frac{1 - cr - cs + c^2 s}{(1-c)^2}\right].$$

For $c_1 = 1/2$, $c_2 = 1/3$, $r = 2$, and $s = 3$, it is easily seen that the bias of $\hat{\Sigma}^{LC}$ is -3Γ/2b. This is clearly first order over-biased since Γ is negative.

The bias is always persistent in finite sample simulations because the sample size is fixed and there is always a trade-off involved between bias and variance due to variability in the batch sizes, similar to the trade-off in single estimators. It is easy to adjust the bias of linear combination estimators but the variance of linear combination estimators becomes larger. The bias reduction and variance inflation results in smaller batch sizes at optimality and reduces our computational cost.

Although, the variance of variance estimators is not as crucial as the bias, we need the covariances to calculate the variance of the linear combination of variance estimators. The expression for variance is more involving and can be expressed as

$$\begin{aligned} Var(\hat{\Sigma}^{LC}) =& \alpha_1^2 Var(\hat{\Sigma}_b) + \alpha_2^2 Var(\hat{\Sigma}_{b/r}) + \alpha_3^2 Var(\hat{\Sigma}_{b/s}) + 2\alpha_2\alpha_1 Cov(\hat{\Sigma}_{b/r}, \hat{\Sigma}_b) \\ &+ 2\alpha_3\alpha_1 Cov(\hat{\Sigma}_{b/s}, \hat{\Sigma}_b) + 2\alpha_3\alpha_2 Cov(\hat{\Sigma}_{b/s}, \hat{\Sigma}_{b/r}), \end{aligned} \tag{2.19}$$

where $\hat{\Sigma}^{LC}$ is defined as in (2.18). To find the variance of the linear combination, we need to find the covariances between each pair of estimators involved in the linear combination, in addition to finding the variances of the estimators.

Using the lugsail Bartlett lag window in (1.21), we obtain the Lugsail BM (LUG-BM) estimator (Vats and Flegal, 2018)

$$\hat{\Sigma}^L = \frac{1}{1-c}\hat{\Sigma}_b^{BM} - \frac{c}{1-c}\hat{\Sigma}_{b/r}^{BM}. \tag{2.20}$$

Here, $\hat{\Sigma}^L$ is shorthand notation for LUG-BM estimator. When $r = 2$ and $c = 1/2$ in

(2.20), we get the flat-top (FT-BM) estimator of Politis and Romano (1995) which can be written as

$$\hat{\Sigma}^{FT} = 2\hat{\Sigma}_b^{BM} - \hat{\Sigma}_{b/2}^{BM}. \tag{2.21}$$

2.3.1 Bias and Variances

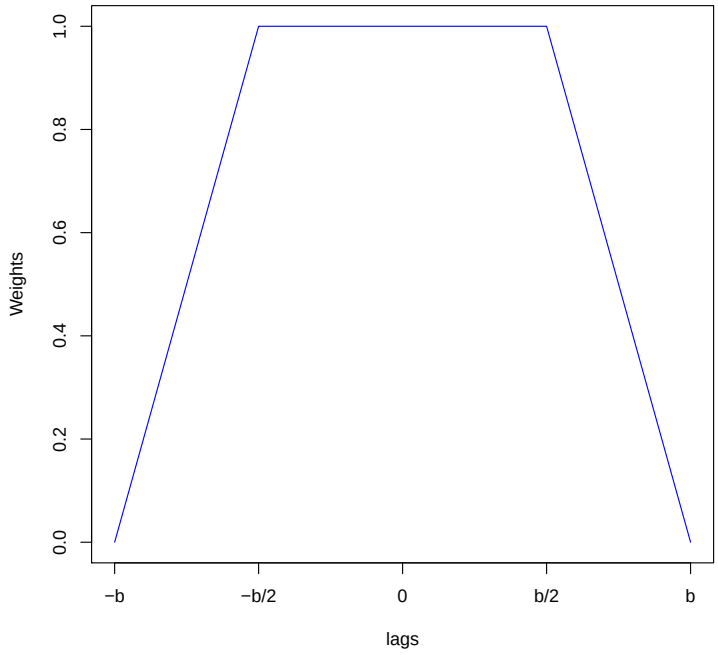

Figure 2.5: Flat-top lag window.

FT estimators incorporate a common bias-correction procedure by integrating a flat-top lag window. The current state-of-the-art bias calculations by practitioners only account for the first-order asymptotic bias. Using Flat-top lag window of Fig-

ure 2.5, there is a perfect first-order asymptotic bias cancellation i.e. the asymptotic first-order bias of FT estimators is zero. Higher-order bias could be substantial in highly correlated chains, especially at small b. It's easy to see that the asymptotic bias of FT estimators is

$$\text{Bias}(\hat{\Sigma}_b^{FT}) = 2\text{Bias}(\hat{\Sigma}_b) - \text{Bias}(\hat{\Sigma}_{b/2}) = 2\frac{\Gamma}{b} - \frac{\Gamma}{b/2} = 0.$$

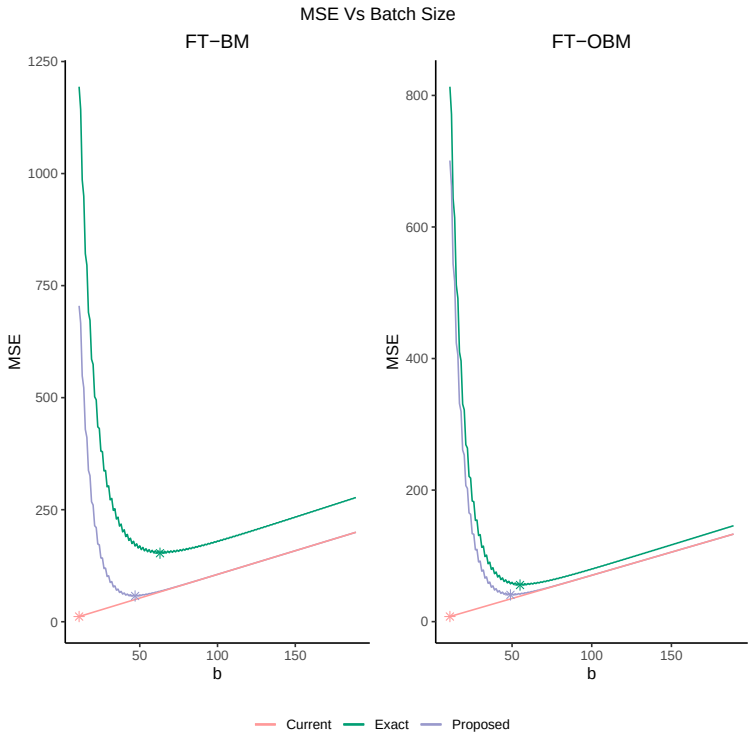

Figure 2.6: Perfect bias cancellation for FT estimators.

Because of this one cannot find an optimal batch size when this bias expression is used in the MSE expression as seen in Figure 2.6. In the current setting, the first-

order unbiased nature of FT estimators make them unusable in practical applications where one needs to construct pilot estimators. MSE optimal batch sizes are crucial to construct variance estimators in any finite sample applications. With the proposed BM estimation methods, the bias of the higher-order correct FT estimator is

$$\begin{aligned}
\text{Bias}(\hat{\Sigma}_b^{FT}) &= 2\text{Bias}(\hat{\Sigma}_b) - \text{Bias}(\hat{\Sigma}_{b/2}) \\
&= 2\left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1\right] - \left[\left(\Gamma_{n-1}^0 - \Gamma_{b/2-1}^0\right) + \frac{1}{b/2}\Gamma_{b/2-1}^1\right] \\
&= \Gamma_{n-1}^0 - \left(2\Gamma_{b-1}^0 - \Gamma_{b/2-1}^0\right) + \frac{2}{b}\left(\Gamma_{b-1}^0 - \Gamma_{b/2-1}^0\right) \\
&\neq 0.
\end{aligned}$$

So, we get a non-zero bias for FT estimators as also seen in Figure 2.6. We can achieve MSE optimality in terms of batch size using the proposed methodology and this is not possible with the existing methodologies.

On the same note, we can generalize the bias of LUG-BM estimator as

$$\text{Bias}\left(\hat{\Sigma}^L\right) = \Gamma_{n-1}^0 - \frac{1}{1-c}\left[\left(\Gamma_{b-1}^0 - \frac{1}{b}\Gamma_{b-1}^1\right)\right] + \frac{c}{1-c}\left[\left(\Gamma_{\frac{b}{r}-1}^0 - \frac{r}{b}\Gamma_{\frac{b}{r}-1}^1\right)\right] + o\left(\frac{1}{b}\right). \tag{2.22}$$

These incorporate single estimators, and various linear combination estimators with appropriate parameters. So, by choosing r and c in (2.22) we can effectively adjust the finite-sample bias of our estimators. Recommendations for choosing the constants r and c are given in Vats and Flegal (2018). Similarly, the variance of LUG-BM

estimator is more involved and can be written as

$$\mathrm{Var}\left(\hat{\Sigma}^L\right) = \frac{2\Sigma^2 b}{n}\left[\left(\frac{nr\left(c^2-2c+r\right)}{(1-c)^2 r(nr-b)}\right) - \left(\frac{b\left(c^2+(1-2c)r\right)}{(1-c)^2 r(nr-b)}\right)\right] + o\left(\frac{b^2}{n^2}\right). \tag{2.23}$$

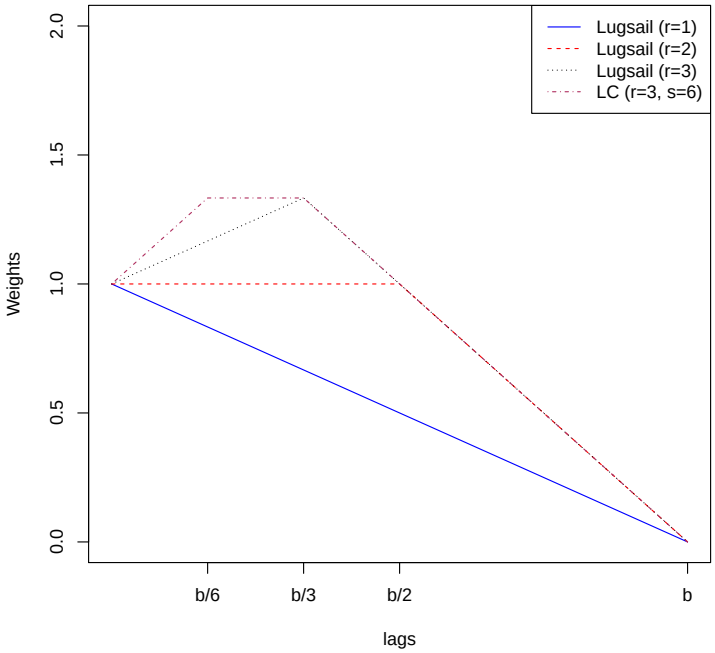

Figure 2.7: Different lag windows.

The variance expression in (2.23) is new in the sense that it contains higher-order terms. The proof is delegated to Section 2.6. The variance of LUG-BM estimator for

$c = 1/2$ is given by

$$\text{Var}\left(\hat{\Sigma}^L\right) = \frac{2\Sigma^2 b}{n}\left(\frac{4nr^2 - 3nr - b}{r(nr-b)}\right) + o\left(\frac{b^2}{n^2}\right). \tag{2.24}$$

And the variance in (2.23) or (2.24) reduces to the variance of single BM estimator for $r = 1$ without the second order terms. On the same note, the variance of LUG-OBM estimator can be written as

$$\text{Var}\left(\breve{\Sigma}^L\right) = \frac{4\Sigma^2 b}{3n}\left[\frac{c^2 r - 3cr + c + r^2}{(c-1)^2 r^2}\right] + o\left(\frac{b}{n}\right). \tag{2.25}$$

LUG-BM estimator can be further modified to allow for more points of discontinuity in the lag window so that the resulting estimator can be expressed as a linear combination of more than two BM estimators with varying batch sizes. For example, consider the modified lugsail lag window with $s > r > 1$, that has the following form

$$\begin{aligned} w_n(k) =& \alpha\left(1 - \frac{|k|}{b}\right) I(0 \le |k| \le b) + \beta\left(1 - \frac{|k|}{b/r}\right) I(0 \le |k| \le \frac{b}{r}) \\ &+ \gamma\left(1 - \frac{|k|}{b/s}\right) I\left(0 \le |k| \le \frac{b}{s}\right). \end{aligned} \tag{2.26}$$

Modified lugsail lag window has 3 points of discontinuity, namely at b, b/r, and b/s. Lugsail lag window with $r = 1$ has one point of discontinuity and lugsail window with $r = 2$ has 2 points of discontinuity, and so on. Lugsail window with $r = 2$ is

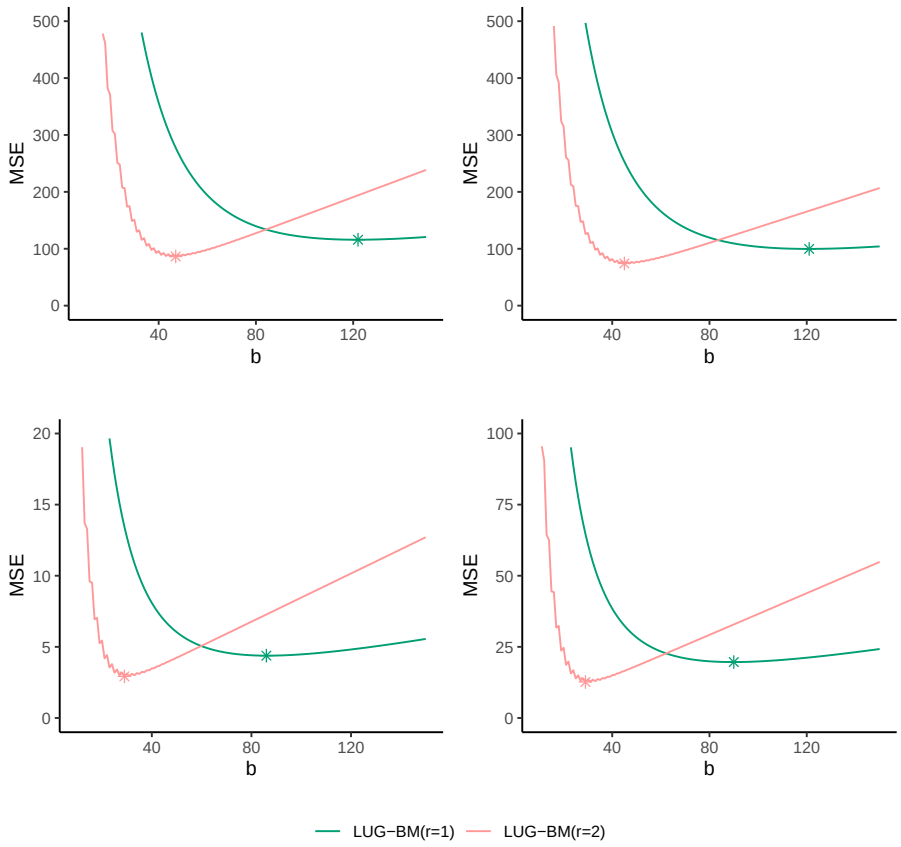

Figure 2.8: MSE comparison of diagonal components of Markov chain with $n = 2e4$ and $\phi = 0.92$.

also called flat top window that has a trapezoidal shape and this gives unit weight to auto-covariances upto a cutoff and linearly down-weights the autocovariances until the lag b. Lugsail lag window with $r > 2$ go beyond unity at small lags and the resulting over weighting of low lags autocovariances that are less biased results in variance estimators with nice bias properties (Vats and Flegal (2018)).

2.4 Examples

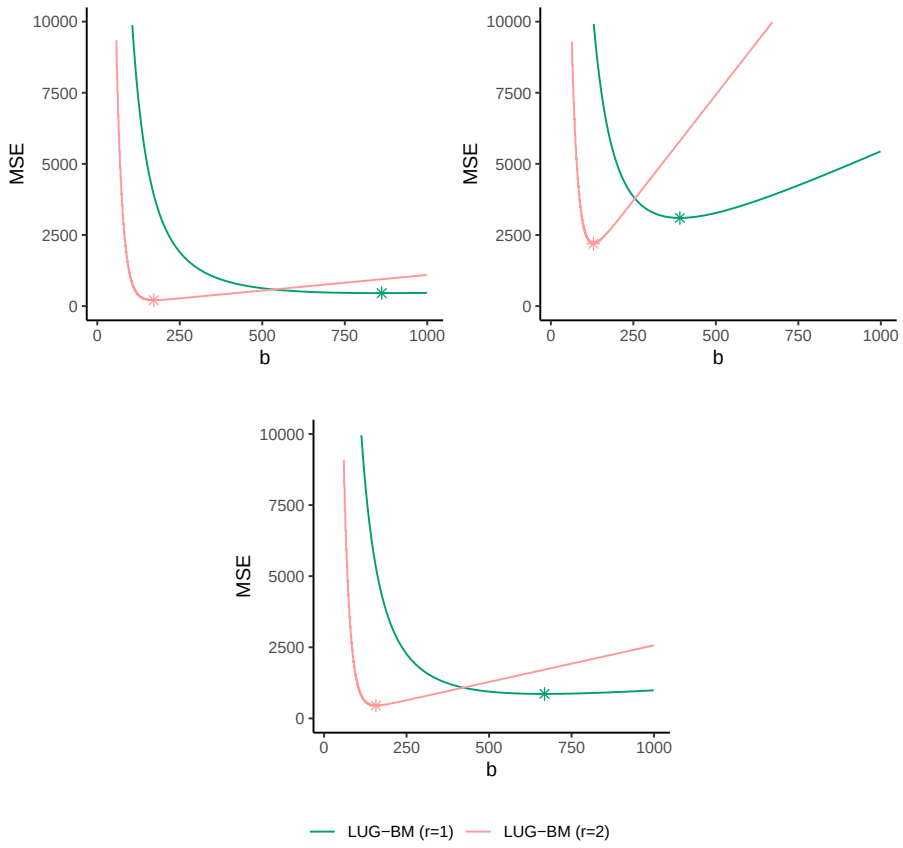

Figure 2.9: MSE comparison of diagonal components of Markov chain with $n = 2e4$ and $\phi = 0.92$.

For optimally estimating the linear combination, i.e., asymptotically unbiased and consistent, one needs to carefully study the bias and variance of individual estimators involved in the combination and reflect that in the optimal choice of parameters involved in the linear combination. This is relevant in finite samples where there is substantial sampling bias. The rate of decrease in bias is faster in linear combination

estimators but the rate of decrease in variance is slower compared to the bias so that the optimality is reached at smaller batch sizes. This can be seen in Figure 2.8 where the MSE from each of the 4 components of a slow mixing 4-dimensional Markov chain from VAR(1) example is depicted. The bias is seen to decay faster and the MSE at optimality is smaller for the linear combination estimator. This results in smaller batch sizes for the linear combination estimator and this reduces our computational burden.

We have also demonstrated this using a real life example of lupus cancer data discussed in Chapter 1. The MSE of these 3 components can be seen for BM estimators and OBM estimators in Figure 2.9 and Figure 2.10, respectively. The bias and variance follows similar trend as in the theoretical example. The bias decay faster for OBM and the variance is reduced compared to BM estimators. So, the optimal batch sizes for OBM estimators are smaller than those for BM estimators. One caveat is that, the exact variance in this case is unknown. However, the true autocovariance and sample autocovariance differ by $O(1/n)$ at most and this should not impact our results by much and the bias and variance growth trends should be the same.

2.5 Batch Sizes

Univariate optimal batch sizes are generally MSE optimal batch sizes and they usually grow at the rate of $o(n^{1/3})$ (Flegal and Jones, 2010). We can optimize the batch

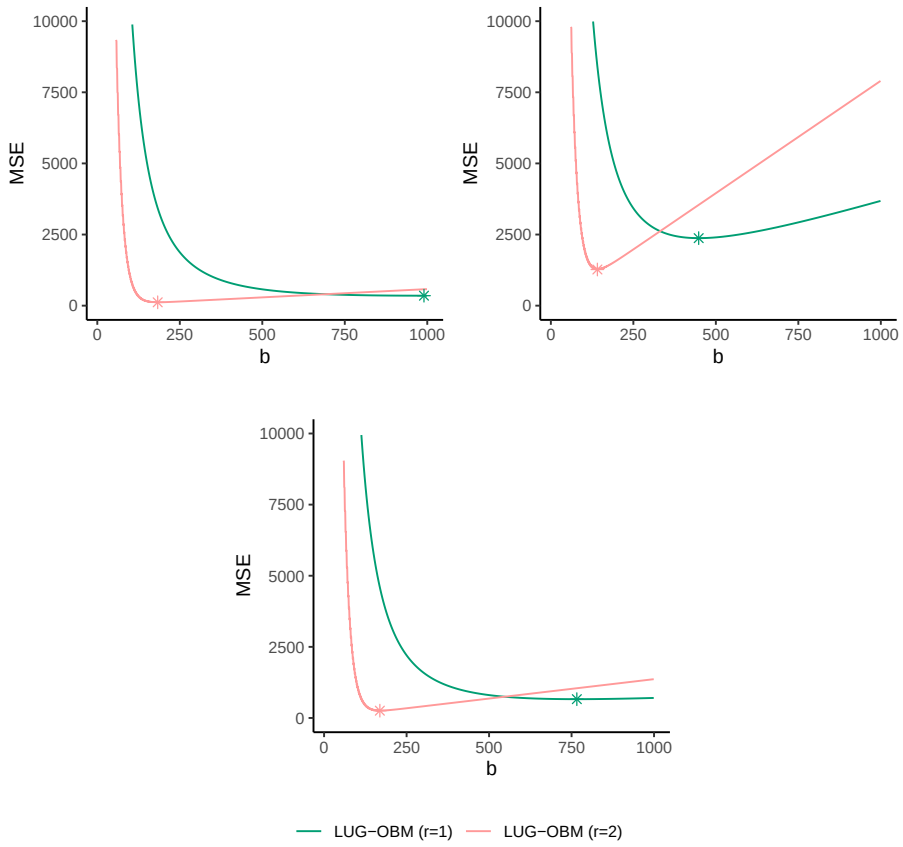

Figure 2.10: MSE comparison of diagonal components of Markov chain with $n = 4e4$.

size that results in minimum MSE using common optimization routines in R. The batch sizes computed in this section use Brent optimization routine of *optim* function in R. Brent's method is a combination of bisection, secant and inverse quadratic interpolation method that results in fast convergence and is useful when the function to optimize is known to be convex (Brent, 1971).

The optimal batch sizes for lugsail BM estimators that minimize the MSE calculated with higher-order corrected bias are computed for both slow and fast mixing

Average Batch Sizes with Standard Errors for $n = 2e4$				
	$\rho = 0.8$		$\rho = 0.9$	
Method	$r = 1$	$r = 2$	$r = 1$	$r = 2$
Exact	42.98 (0.11)	23.41 (0.05)	66.53 (0.13)	40.68 (0.09)
Proposed	30.81 (0.08)	17.24 (0.05)	47.35 (0.09)	28.45 (0.05)
Current	9.18 (0.09)	26.13 (0.08)	12.89 (0.14)	56.28 (0.14)

Table 2.1: Average batch sizes from VAR(1) process using different estimation methods.

Markov chain run for small sample size of $n = 2e4$. At this small sample size, we know there is significant bias present in our variance estimators. Current methods result in very small optimal batch sizes as seen in Table 2.1. These batch sizes are the geometric mean of the univariate MSE optimal batch sizes of the individual components of the Markov chain. Also noticeable is that the standard errors for the batch sizes computed from proposed methods are very small. The MSE decay of the individual components behave similarly and an average of batch sizes would also give similar results.

Lag-based methods from Politis and Romano (1995) was employed for $r = 2$, since current methods yield an optimal batch size of zero. Lag-based method seems to overestimate the MSE optimal batch size. Proposed estimation methods are a significant improvement over the current methods and should be used to construct near optimal variance estimators. Similarly, the growth of MSE optimal batch size of BM estimators for different mixing Markov chains can be seen compared for exact, proposed and current estimation methods in Figure 2.11. Exact BM gives us the largest batch sizes and the methods involving linear combination yields relatively

small batch sizes. The current BM method gives the smallest batch size, but it is significantly underestimated and far from the exact batch sizes.

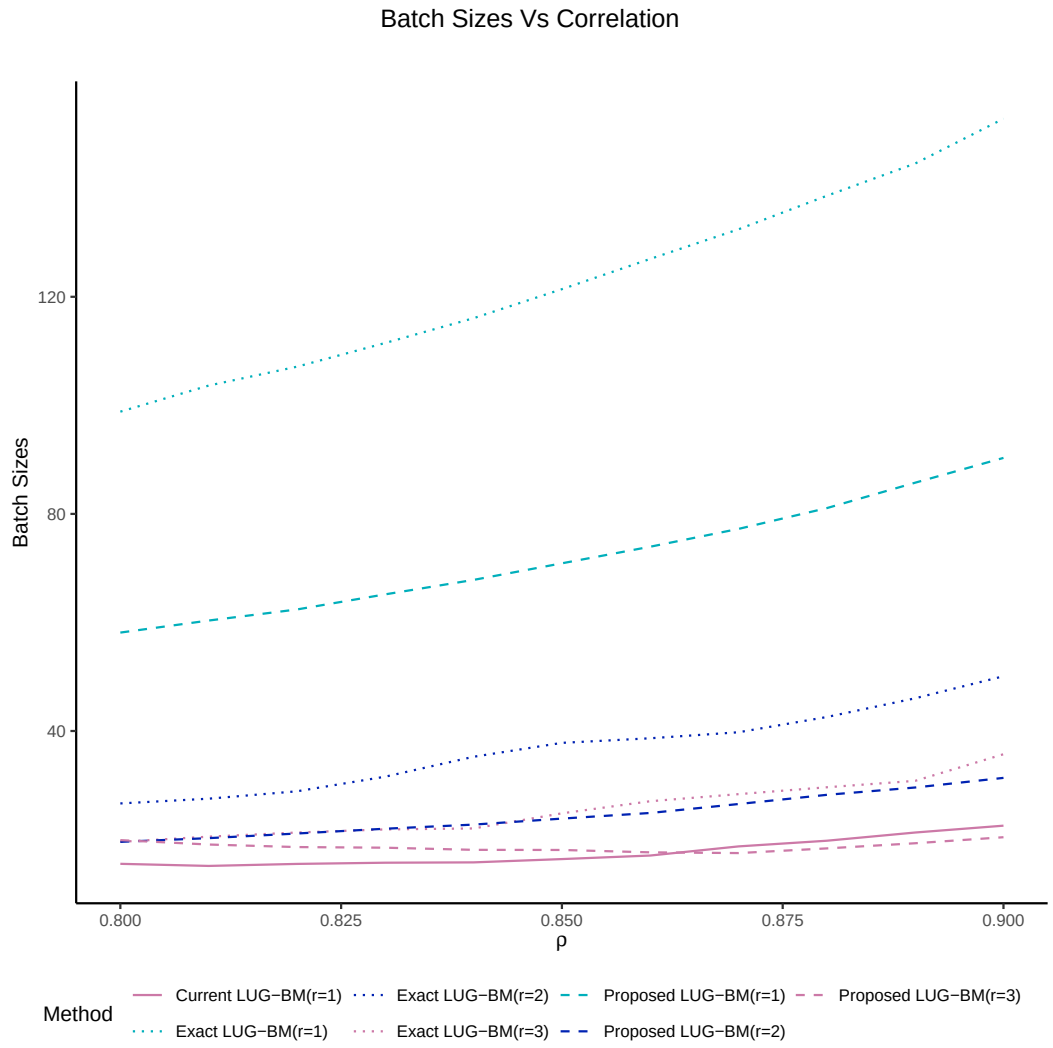

Figure 2.11: Batch size vs Markov chain correlation for different estimation methods for $n = 2e4$.

2.6 Proofs

Proof of (2.10) The quadratic coefficients can be summarized for BM estimators as

$$q_{ij}^{BM} = \frac{1}{d_B}\left(\frac{a_{ij}}{b^2} - \frac{a_{ii}+a_{jj}}{nb} + \frac{a}{n^2}\right).$$

For observations falling within the same batch, the quadratic coefficients is

$$q_{ij}^{BM} = \frac{b^2}{(n-b)}\left(\frac{1}{b^2} - \frac{2}{bn} + \frac{1}{bn}\right) = \frac{1}{n}.$$

Similarly, for observations falling outside of batches the quadratic coefficient is

$$q_{ij}^{BM} = \frac{b^2}{(n-b)}\left(-\frac{2}{bn} + \frac{1}{bn}\right) = -\frac{1}{a(n-b)}.$$

Let $R(h) = \text{cov}\,(Y_i, Y_{i+h})$ be the covariance. The expected value of the BM estimator is,

$$E(\hat{\Sigma}_{BM}) = E\left[\sum_{r=1}^{n}\sum_{s=1}^{n} q_{rs}^{(BM)} Y_r Y_s\right] = E[\text{I} + \text{II} + \text{III}]$$

This expectation can be decomposed into three parts. The first is within batch expectation.

$$
\begin{aligned}
E[\mathrm{I}] &= a \cdot E\Bigg[\sum_{r=1}^{b} q_{rr}^{(BM)} Y_r Y_r^T + \sum_{s=1}^{b-1} \sum_{r=1}^{b-s} q_{rs}^{(BM)} \big(Y_r Y_{r+s}^T + Y_{r+s} Y_r^T \big) \Bigg] \\
&= a \cdot \frac{1}{n} \cdot \Bigg[bR(0) + 2 \sum_{j=1}^{b-1} (b-j) R(j) \Bigg] \\
&= a \cdot \frac{1}{n} \cdot \Bigg[\sum_{j=-(b-1)}^{b-1} (b - |j|) R(j) \Bigg] \\
&= \frac{n}{b} \cdot \frac{1}{n} \cdot b \Bigg[\sum_{j=-(b-1)}^{b-1} \left(1 - \frac{|j|}{b} \right) R(j) \Bigg] \\
&= R(0) + \Gamma_0^{b-1} - \frac{1}{b} \Gamma_1^{b-1}
\end{aligned}
$$

Similarly, outside batches within lag $b-1$, the expectation is

$$
\begin{aligned}
E[\mathrm{II}] &= (a-1) \cdot E\Bigg[\sum_{s=1}^{b-1} \sum_{r=b-s+1}^{b} q_{rs}^{(BM)} \big(Y_r Y_{r+s}^T + Y_{r+s} Y_r^T \big) \Bigg] \\
&= \left(\frac{n-b}{b} \cdot \frac{-b}{n(n-b)} \cdot \sum_{-(b-1)}^{b-1} j \cdot R(j) \right) \\
&= -\frac{1}{n} \Gamma_1^{b-1}
\end{aligned}
$$

And, expectations for lag b and beyond is,

$$\begin{aligned} E[\text{III}] &= 2 \cdot \frac{-b}{n(n-b)} \cdot \sum_{j=b}^{n-1} (n-j) R(j) \\ &= \frac{-2b}{n(n-b)} \cdot \left[n \sum_{j=b}^{n-1} R(j) - \sum_{j=b}^{n-1} j \cdot R(j) \right] \\ &= -\frac{b}{n-b} \left(\Gamma_0^{n-1} - \Gamma_0^{b-1} \right) + \frac{b}{n(n-b)} \left(\Gamma_1^{n-1} - \Gamma_1^{b-1} \right) \end{aligned}$$

Combining all the three above, the expected value of BM estimator in terms of quadratic coefficient is,

$$E(\hat{\Sigma}_b^{BM}) = R(0) + \frac{n}{n-b} \Gamma_{b-1}^0 - \frac{b}{n-b} \Gamma_{n-1}^0 + \frac{b}{n(n-b)} \Gamma_{b-1}^1 - \frac{n}{b(n-b)} \Gamma_{n-1}^1.$$

This expression can be further approximated as follows:

$$
\begin{aligned}
E[\hat{\Sigma}_b^{BM}] &= R(0) + \sum_{i=-\infty}^{+\infty} R_i - \sum_{i=-\infty}^{+\infty} R_i + \frac{a\Gamma_{b-1}^0 - \Gamma_{n-1}^0}{a-1} + \frac{\Gamma_{n-1}^1 - a^2\Gamma_{b-1}^1}{n(a-1)} \\
&= R(0) + \Sigma - R_0 - 2\sum_{i=1}^{n-1} R_i - 2\sum_{i=n}^{+\infty} R_i + \frac{a\Gamma_{b-1}^0 - \Gamma_{n-1}^0}{a-1} + \frac{\Gamma_{n-1}^1 - a^2\Gamma_{b-1}^1}{n(a-1)} \\
&= \Sigma - \frac{a}{a-1}\left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1 + \frac{\Gamma_{n-1}^1}{na}\right] - 2\sum_{i=n}^{+\infty} R_i \\
&= \Sigma - \left(1 + \frac{1}{a-1}\right)\left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1 + \frac{\Gamma_{n-1}^1}{na}\right] - 2\sum_{i=n}^{+\infty} R_i \\
&= \Sigma - \left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1 + \frac{\Gamma_{n-1}^1}{na}\right] \\
&+ \frac{1}{a-1}\left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1 + \frac{\Gamma_{n-1}^1}{na}\right] - 2\sum_{i=n}^{+\infty} R_i
\end{aligned}
$$

Proof of Theorem 7 Again, assume the rate of decay of autocovariances is $R(h) = O(|h|^{-2-\delta})$. A Markov chain is polynomially ergodic of order m if $\gamma(n) = n^{-m}$. Then, the strong mixing conditions imply

$$\alpha(n) \leq n^{-m}(E_\pi M),$$

provided we know the existence of M. Let $EY_t = 0$, and $E|Y_t|^{2+\delta} < \infty$ for some $\delta > 0$, and all t. We know that (White, 1984, Corollary 6.17)

$$|EY_{t+h}Y_t| \leq 2(2^{\frac{1}{2}+1}) \cdot \alpha(h)^{\frac{\delta}{2(2+\delta)}} \cdot (EY_{t+h}^2)^{\frac{1}{2}} \cdot (E|Y_t|^{2+\delta}))^{\frac{1}{2+\delta}}.$$

Suppose $\sup_t E|Y_t|^{2+\delta} \leq \Delta < \infty$. Then, we see that

$$\begin{aligned}
|EY_{t+h}Y_t| &\leq 2 \cdot (2^{\frac{1}{2}+1}) \cdot \alpha(h)^{\frac{\delta}{2(2+\delta)}} \cdot \left(\sup_t EY_t^2\right)^{\frac{1}{2}} \cdot \left(\sup_t E|Y_t|^{2+\delta}\right)^{\frac{1}{2+\delta}} \\
&\leq 2 \cdot (2^{\frac{1}{2}+1}) \cdot \alpha(h)^{\frac{\delta}{2(2+\delta)}} \cdot \Delta^{\frac{2}{2+\delta}},
\end{aligned}$$

since $(EY_t^2)^{1/2} \leq (E|Y_t|^r)^{1/r}$ for $r > 2$ and $EY_t^2 \leq \sup_t EY_t^2$. Assuming polynomially ergodic markov chain, we have

$$\begin{aligned}
|EX_{t+h}X_t| &\leq 2 \cdot (2^{\frac{1}{2}+1}) \cdot \left(h^{-m}(E_\pi M)\right)^{\frac{\delta}{2(2+\delta)}} \cdot \Delta^{\frac{2}{2+\delta}} \\
&\leq K \cdot h^{-\frac{m\delta}{2(2+\delta)}} \\
&= O\left(h^{-\frac{m\delta}{2(2+\delta)}-\epsilon}\right)
\end{aligned}$$

Comparing this with $R(h) = E\left[X_{t+h}X_t\right] = O\left(|h|^{-2-\epsilon}\right)$, it's easy to see that

$$-\frac{m\delta}{2(2+\delta)} = -2 \quad \Longrightarrow \quad m = 4 + \frac{8}{\delta}.$$

So, the auto-covariance convergence rate of Pedrosa (1994) is satisfied if we have a polynomially ergodic, α-mixing markov chain of order greater than $4(1+\epsilon)(1+\frac{2}{\delta})$. Then, the covariance of two BM estimators with batch size b and b/r, resp., can be written as (Pedrosa, 1994, Lemma 4.1 - 4.3)

$$\mathrm{Cov}(\hat{\Sigma}_b^{(1)}, \hat{\Sigma}_{b/r}^{(2)}) = 2\Phi_A + \Phi_B, \tag{2.27}$$

where,

$$\Phi_A = \sum_{r,s,t,u=1}^{n} q_{rs}^{(1)} q_{tu}^{(2)} R(t-r)R(u-s)$$

$$\Phi_B = \sum_{\tau,s,t,u=1}^{n} q_{rs}^{(1)} q_{lu}^{(2)} \kappa_4 \left(X_i, X_{i+r}, X_{i+s}, X_{i+t}\right)$$

For simplicity, assume both b and b/r are integers. The cross product in Φ_A can be decomposed into two major parts where both the quadratic coefficients overlap and parts where the larger batch is dominant. The first term where both batch have overlap with common weight $1/n$ is

$$\mathrm{I} = \left(\frac{nr}{b}\right) \cdot \left(\frac{b}{r}\right)^2 \cdot \left(\frac{1}{n}\right)^2 = \frac{b}{rn}.$$

The second term where smaller batch have small negative weights, while the bigger batch has larger positive weights is

$$\begin{aligned}
\mathrm{II} &= \left(\frac{n}{b}\right) \cdot \left(\frac{(r-1)b^2}{r}\right) \cdot \left(\frac{-b}{n(nr-b)}\right) \cdot \left(\frac{1}{n}\right) \\
&= \frac{(1-r)}{r(r-\frac{b}{n})} \cdot \left(\frac{b^2}{n^2}\right) \\
&= \frac{(1-r)}{r^2}\left(\frac{b^2}{n^2}\right) + \frac{(1-r)}{r^2}\frac{b}{nr-b}\left(\frac{b^2}{n^2}\right)
\end{aligned}$$

Similarly, the small order terms can be additionally written as

$$\text{III} = \left(\frac{n}{b} - 1\right) \cdot \left(b(b-1)\right) \cdot \left(\frac{b^2}{n^2(nr-b)(n-b)}\right) = \frac{b^2(b-1)}{n^2(nr-b)} = O\left(\frac{b^3}{n^3}\right).$$

On the other hand, due to the summability of the cumulant condition, Φ_B can be approximated as

$$\begin{aligned}
\Phi_B &= \sum_{\tau,s,t,u=1}^{n} q_{rs}^{(1)} q_{tu}^{(2)} \kappa_4 \sum_{v=\infty}^{\infty} b_v b_{v+s-r} b_{v+t-r} b_{v+u-r} \\
&= O\left(\frac{1}{n^2} \sum_{r=1}^{n} \sum_{f=-\infty}^{\infty} b_f \sum_{g=-\infty}^{\infty} b_g \sum_{h=-\infty}^{\infty} b_h \sum_{v=-\infty}^{\infty} b_v\right) \\
&= O\left(\frac{1}{n}\right).
\end{aligned}$$

Therefore, the asymptotic covariance can be written as

$$\begin{aligned}
\text{Cov}(\hat{\Sigma}_b^{(1)}, \hat{\Sigma}_{b/r}^{(2)}) &= 2\Sigma^2 \left[\frac{b}{rn} + \frac{(1-r)}{r^2}\left(\frac{b^2}{n^2}\right) + \frac{(1-r)}{r^2}\frac{b}{nr-b}\left(\frac{b^2}{n^2}\right)\right] + O\left(\frac{b^3}{n^3}\right) + O\left(\frac{1}{n}\right) \\
&= \frac{2b\Sigma^2}{rn}\left[1 + O\left(\frac{1}{b}\right) + O\left(\frac{b}{n}\right)\right].
\end{aligned}$$

Proof of Corollary (8) Corollary (8) can be proven directly from Theorem 7. However, we can do a standalone proof of $\text{Var}(\hat{\Sigma}^{BM})$. We need the following lemma.

Lemma 11 *(Pedrosa, 1994, Lemma 4.1 - 4.3) The variance of BM estimator can be written in terms of quadratic coefficients as*

$$Var(\hat{\Sigma}^{BM}) = 2\Sigma^2 \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \left(q_{ij}^{BM}\right)^2 \right] + O\left(\frac{1}{n}\right) + O\left(\frac{b^2}{n^2}\right).$$

The indexing of the observation within lag b can be done as follows

$$\sum_{r=1}^{b} \sum_{s=1}^{b} Y_r Y_s^T = \sum_{r=1}^{b} Y_r Y_r^T + \sum_{s=1}^{b-1} \sum_{r=1}^{b-s} \left(Y_r Y_{r+s}^T + Y_{r+s} Y_r^T\right) + \sum_{s=1}^{b-1} \sum_{r=b-s+1}^{b} \left(Y_r Y_{r+s}^T + Y_{r+s} Y_r^T\right).$$

Then the variance of BM estimator can be calculated as

$$\begin{aligned}
\mathrm{Var}(\hat{\Sigma}^{BM}) &= 2\Sigma^2 \left[\sum_{i=1}^{n} \sum_{j=1}^{n} (q_{ij}^{BM})^2 \right] + O\left(\frac{1}{n}\right) + O\left(\frac{b^2}{n^2}\right) \\
&= 2\Sigma^2 \left[\left(a \cdot \left(\frac{1}{n}\right)^2 \cdot b^2 \right) + \left((a-1) \cdot \left(\frac{-1}{a(n-b)}\right)^2 \cdot b(b-1) \right) \right. \\
&\qquad \left. + 2 \sum_{j=b}^{n-1} (n-j) \left(\frac{-1}{a(n-b)}\right)^2 \right] + O\left(\frac{1}{n}\right) + O\left(\frac{b^2}{n^2}\right) \\
&= \frac{2\Sigma^2 b}{n} + O\left(\frac{1}{n}\right) + O\left(\frac{b^2}{n^2}\right) \\
&= \frac{2\Sigma^2 b}{n} \left[1 + O\left(\frac{1}{b}\right) + O\left(\frac{b}{n}\right) \right].
\end{aligned}$$

The exact bias of univariate OBM estimator described in Aktaran-Kalaycı et al.

(2007) can be written as

$$\begin{aligned}
\mathrm{E}[\hat{\Sigma}_b^{OBM}] =& R_0 + \frac{2}{n-b}\left\{\sum_{i=1}^{\infty} - \sum_{i=b}^{\infty}\right\}\left[n-b-\left(a+\frac{1}{a}-\frac{1}{n-b+1}\right)i+\frac{i^2}{n-b+1}\right]R_i \\
&+ \frac{2}{n-b}\left\{\sum_{i=b}^{\infty} - \sum_{i=n-b+1}^{\infty}\right\}\left\{\left[-\frac{bn}{n-b+1}+\left(\frac{2b}{n-b+1}-\frac{1}{a}\right)i\right]R_i\right. \\
&+ \frac{2}{n-b}\left\{\sum_{i=n-b+1}^{\infty} - \sum_{i=n}^{\infty}\right\}\left[b-\frac{n^2+n}{n-b+1}+\left(\frac{2n+1}{n-b+1}-\frac{1}{a}\right)i-\frac{i^2}{n-b+1}\right]R_i \\
&= \Sigma - \Gamma_{n-1}^0 + \Gamma_{b-1}^0 - \frac{1}{n-b}\left(a+\frac{1}{a}-\frac{1}{n-b+1}\right)\Gamma_{b-1}^1 + \frac{\Gamma_{2,b-1}}{(n-b)(n-b+1)} \\
&\quad + \frac{bn}{(n-b)(n-b+1)}\left(\Gamma_{b-1}^0 - \Gamma_{n-b}^0\right) + o\left(\frac{b}{n}\right) + o\left(\frac{1}{b}\right) + O\left(\frac{b}{n^2}\right) + O\left(\frac{1}{n}\right) \\
&= \Sigma - \left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{a\Gamma_{b-1}^1}{n-b} - \frac{bn\left(\Gamma_{b-1}^0 - \Gamma_{n-b}^0\right)}{(n-b)(n-b+1)}\right] + O\left(\frac{1}{b^2}\right).
\end{aligned}$$

2.7 Concluding Remarks

Asymptotically optimal estimators have limited practical use, even when we combine information from various estimators. Single variance estimators, though appealing, is systematically biased even after bias correction because the rate of decay of bias is slow in finite sample settings. Linear combination estimators should be employed to adjust the systematic bias.

In finite samples, sampling bias is also crucial and there has not been many studies to address this issue in MCMC and time series methodology. Current MSE optimal estimators perform poorly in finite samples. Asymptotical results are sound, but their

practical utility are minimal due to these limitations. However, utilizing the exact systematic error expressions we can reduce the finite sample bias of these estimators.

In terms of MCMC and BM methodology, in particular, the finite sample adjustment methodologies discussed in this chapter is new and this opens up grounds for future research. These new methodologies also advocate for using linear combination estimators whenever it is possible computationally.

Chapter 3

Spectral Variance Estimators

3.1 Spectral variance estimators

Frequency analysis of time series has long history (Schuster, 1897). The periodogram approach yields smaller bias of the spectral estimates (Das et al., 2021). The correlation between periodogram ordinates at Fourier frequencies is of the order $O(1/n)$. There is a wide array of literature studies in the empirical spectral distribution of the sample autocovariance matrix when the observations are assumed to come from a linear time series with certain restriction on the coefficients (Anderson, 1971), (Priestley, 1981), (Shao and Wu, 2007). CLT for spectral density estimates under appropriate regularity conditions have been established in the Fourier domain (Chanda, 2005), (Liu and Wu, 2010), (Lin and Liu, 2012), (Panaretos et al., 2013), (Wu and Zaffaroni, 2018).

If Y_t is short-range dependent, namely, $\sum_{h=0}^{\infty} |R(h)| < \infty$, then it sufficient for the existence of the spectral density. The true spectral density given by

$$f_Y(\omega) = \frac{1}{2\pi} \sum_{h \in \mathbf{Z}} R(h) e^{ik\omega}, \qquad \text{for } \omega \in [-\pi, \pi],$$

is continuous and bounded under assumption 3. Without loss of generality, assume $E[Y_t] = 0$, i.e. if $\mu \neq 0$ we replace Y_t by $Y_t - \bar{Y}$ where $\bar{Y} = n^{-1} \sum_{r=1}^{n} Y_t$ and the corresponding analysis of $\hat{f}_Y(\omega)$ remains unchanged, asymptotically. Let $f^{(1)}(0) = \sum_{-\infty}^{\infty} h^1 R(h)$, then above implies that $\left|f^{(1)}(0)\right| \leq \sum_{-\infty}^{\infty} h^1 |R(h)| < \infty$ for $r = 1$.

The Fourier transform of a set of observations $Y_1, ..., Y_n$, can be written as $J_n(\omega) = n^{-1/2} \sum_{t=1}^{n} Y_t e^{it\omega}$ and the corresponding periodogram is written as $I_n(\omega) = |J_n(\omega)|^2$. Some properties of the spectral density function are:

1. $f_Y(\omega)$ is even,

2. $f_Y(\omega)$ is non-negative for all $\omega \in [-\pi, \pi]$, and $\int_{-\pi}^{\pi} f_Y(\omega) d\omega = 1$,

3. $R_Y(h) = \int_{-\pi}^{\pi} e^{ih\omega} f_Y(\omega) d\omega = \int_{-\pi}^{\pi} \cos(h\omega) f_X(\omega) d\omega$.

Note that since cos is a periodic function with the period 2π, the range of values of the spectral density is determined by the value of $f_Y(\omega)$ for $\omega \in [0, \pi]$ (Brockwell and Davis, 1991).

Generally, autoregressive processes of finite order satisfy assumption 3 (Den Haan and Levin, 1996). We can estimate the spectral density by smoothing the peri-

odogram. The estimated spectral density of f with the AR(p) fitted model is then,

$$\hat{f}_n(\omega) = \frac{1}{2\pi}\int_0^{2\pi} W(\omega-\lambda) I_n(\lambda)\,\mathrm{d}\lambda,$$

where W is a non-negative symmetric function satisfying $\int W(u)du = 2\pi$ and also the finiteness property, $\int W^2(u)\mathrm{d}u < \infty$. Define $W_h(\cdot) = (1/h)W(\cdot/h)$, where h is a bandwidth. Here, h serves the same purpose as b^{-1}. Note that $\Sigma/(2\pi)$ is the value of the spectral density matrix of Y_i at zero frequency. We can estimate the SV estimator by using the spectral theory of estimation at frequency zero.

$$\hat{\Sigma}_{ij}^{SV} = 2\pi \hat{f}_{n,ij}(0)$$

The finite sample bias of SV estimator can be derived from Priestley (1981), Politis and Romano (1995), Kokoszka and Jouzdani (2020). The bias can be derived as

$$\begin{aligned}
E\left[\hat{\Sigma}_{SV}\right] &= \Sigma - 2\sum_{i=1}^{n-1}\left(1-w\left(\frac{i}{b}\right)\right)R_i - \frac{2}{n}\sum_{i=1}^{n-1}|i|w\left(\frac{i}{b}\right)R_i - \sum_{|i|\geq n} R_i + O\left(\frac{b}{n}\right) \\
&= \Sigma - 2\sum_{i=1}^{b-1}\left(1-w\left(\frac{i}{b}\right)\right)R_i - \frac{2}{n}\sum_{i=1}^{b-1}|i|w\left(\frac{i}{b}\right)R_i - \sum_{|i|\geq b} R_i + O\left(\frac{b}{n}\right) \\
&= \Sigma - \sum_{|i|\leq b}\left(1-w\left(\frac{i}{b}\right)\left(1-\frac{|i|}{n}\right)\right)R_i - \sum_{|i|>b} R_i + O\left(\frac{b}{n}\right).
\end{aligned}$$

The second term is due to the lag window and the third term is the contribution due to truncation of the autocovariances beyond lag b. The rate of decay of batch size

determined due to the interplay of $\sum_{j=-\infty}^{\infty} j^r |R(j)| < \infty$ and $\lim_{x\to 0} \frac{1-K(x)}{x^q}$. So, the bias can be optimized by a trade off of r and q. Note the following set of common assumptions:

C1. $\Theta_{0,4} = \sum_{i\geq 0} \delta_4(i) < \infty$

C2. $\sum_{t_1,\ldots,t_{k-1}\in \mathbf{Z}} \left|\text{cum}\left(X_0, X_{t_1}, \ldots, X_{t_{k-1}}\right)\right| < \infty$ for $k = 4$

3.1.1 Central Limit Thoerem

Using the m-dependent approximation and martingale approximation, Liu and Wu (2010) establishes the normality of $f_n(\omega) - \mathbb{E}f_n(\omega)$. The absolute summability of the autocovariances, C2, is not enough to establish the consistency of the estimator $\hat{\Sigma}^{SV}$ (Hörmann and Kokoszka, 2010). Assuming C1, the cumulant summability condition C2 is not required for asymptotic normality (Berkes et al., 2016). Let $\kappa = \int_{-1}^{1} W^2(u)du$.

Theorem 12 *(Liu and Wu (2010)) Suppose that* $\mathbb{E}Y_1 = 0, \mathbb{E}Y_1^4 < \infty$ *and* $\Theta_{0,4} < \infty$. *Let* $1/b + b/n \to 0$. *Then, for* $\omega = 0$ *or* π,

$$\sqrt{\frac{n}{b}} \left\{ \hat{f}_{n,ij}(\omega) - \mathbb{E}\left(\hat{f}_{n,ij}(\omega)\right) \right\} \Rightarrow \kappa^{1/2} N\left(0, f_{ii}(\omega) f_{jj}(\omega) + f_{ij}^2(\omega)\right).$$

For Bartlett kernel in particular with $q = 1$ and $r = 1$, under C1, C2, and $b = o(n)$,

(Anderson, 1971) we have

$$\lim_{n\to\infty} b|E\left[\hat{\Sigma}^{SV}\right] - \Sigma| = 0 \tag{3.1}$$

Remark 13 *(Liu and Wu, 2010, Remark 5) If there exists a kernel with* $q = 1$ *and* $r = 1$ *and* $\sum_{k\geq 1} k\delta_{k,2} < \infty$*, then* $\mathbb{E}\left(\hat{f}_{n,ij}(\omega)\right) - f_{ij}(\omega) = O\left(b^{-1}\right)$*, and* $\mathbb{E}\left(\hat{f}_{n,ij}(\omega)\right)$ *can be replaced by* $f_{ij}(\omega)$ *as long as* $n\log n = o\left(b^3\right)$*.*

In order to use Theorem 12 in finite samples, however we need to account for the finite sample bias. As we saw in (3.1), the bias is zero only asymptotically. To account for the finite sample bias, we need to derive the bias using the form of Bartlett kernel as

$$\begin{aligned}
E\left[\hat{\Sigma}_{SV}\right] &= \Sigma - \sum_{|i|\leq b}\left(\frac{n|i| + b|i| - |i|^2}{bn}\right)R_i - \sum_{|i|>b} R_i + O\left(\frac{b}{n}\right) \\
&= \Sigma - \left(\frac{1}{b} + \frac{1}{n}\right)\Gamma_1^{b-1} + \frac{1}{bn}\Gamma_2^{b-1} - \Gamma_0^{n-1} + \Gamma_0^{b-1} + O\left(\frac{b}{n}\right) + O\left(\frac{1}{n}\right) \\
&= \Sigma - \left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1\right] + o\left(\frac{1}{b}\right) + O\left(\frac{b}{n}\right) + O\left(\frac{1}{n}\right).
\end{aligned}$$

Thus, we get the similar rate of bias decay as for $\hat{\Sigma}^{SV}$ as for $\hat{\Sigma}^{BM}$ estimator from Chapter 2. It is well known that OBM estimator is equivalent to SV estimator except for some end effects (Flegal and Jones, 2010). More specifically, in the univariate case

the variance of a SV estimator in the context of MCMC is

$$\text{Var}\left(\hat{\Sigma}_{ii}^2\right) = \frac{4}{3}\Sigma_{ii}^2\frac{b}{n} + o\left(\frac{b}{n}\right) + \eta,$$

where η tends to 0 (Flegal, 2008). For Bartlett kernel with $\kappa = 2/3$, the variance of the SV estimator ($\omega = 0$) can also be derived from Theorem 12 as

$$\text{Var}\left(\hat{\Sigma}_b^{SV}\right) = 4\pi^2\text{Var}\left(\hat{f}_{n,ij}\left(0\right)\right) = \frac{4}{3}\frac{b}{n}\cdot\left(2\pi f_{n,ij}(0)^2\right) = \frac{4}{3}\frac{b}{n}\Sigma^2.$$

3.2 Concluding Remarks

Spectral estimators are the least biased estimators available. They are computationally expensive in real life where we are limited to time-domain analysis in the context of MCMC and other real-world applications in time series, longitudinal data, etc. However, they share similar bias properties with OBM and BM estimators. In our linear model, if we assume the errors are Gaussian then the error in estimation (bias) vanishes and it is easier to prove normality results. But Markov chains also show nonlinear behavior and we need to account for that. And it is extremely hard to establish normality results of spectral variance estimator in the context of MCMC variance and the current chapter constitutes some preliminary work and the rest is left for future work.

Chapter 4

Alternative Loss Functions

4.1 Alternative Loss Functions

Mean Square Error (MSE) has been traditionally used as an optimality criteria for evaluating the estimators and assessing their risk. However, MSE carries information about the variance and the square of bias, but not the direction of bias. Asymmetrical loss functions are desirable in many cases because they help to validate a different choice of center for any distribution. Asymmetrical loss functions let the researchers specify their preference over overestimation or underestimation or validate an estimator with finite sample sampling bias and irreducible systematic error. The penalty for overestimation or underestimation can be reduced accordingly using the alternative loss function.

Lugsail estimators of Chapter 2 are asymptotically unbiased, but retain significant

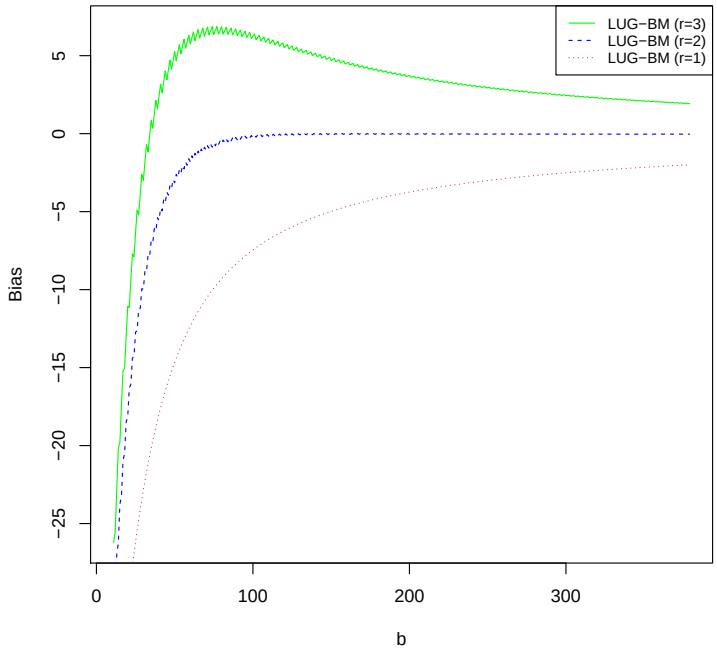

Figure 4.1: Bias comparison of different Lugsail estimators in finite samples, $n = 2e4$.

bias in finite samples as seen in Figure 4.1. For a fixed sample size n, the bias of Lugsail variance estimators with $c = 1/2$ and different r can be seen to approach zero either from above for $r = 3$ or from below for $r = 1$ and $r = 2$, respectively. Lugsail estimator with $r = 3$ is clearly over-biased after some lags and those with either $r = 1$ or $r = 2$ are clearly under-biased for a particular b. For a particular b, if we know the truth, these systematic biases are fixed and we could potentially adjust for the sampling bias.

4.2 Distributional assumptions

For a m-dependent stochastic processes, using regression based linear analysis Gupta et al. (2014) derives asymptotically normally distributed variance estimators including BM estimators. There will always be bias in estimation in finite samples, as this is the core message of the thesis. This chapter does not go deep into proving these normality assumptions and delegate them to future work. But, given the assumptions for normality are met, theoretically we can calculate the scaling needed for our estimators to be optimal under some loss functions that converges to MSE loss in the limiting case.

Univariate BM estimators, $\sigma_b^{2,BM}$, can be shown to be distributed as a scaled chi-squared for fixed a as $b \to \infty$ i.e.,

$$\widehat{\sigma}_b^{2,BM} \underset{b\to\infty}{\Longrightarrow} \sigma^2 \chi^2_{a-1}/(a-1).$$

This is not a consistent estimator of variance as the variance does not go to zero in the limit $b \to \infty$. Glynn and Whitt (1991) shows that as $a \to \infty$,

$$\sigma^2 \chi^2_{a-1}/(a-1) \Rightarrow \sigma^2$$

and

$$\sqrt{a}\left(\frac{\sigma^2\chi_a^2}{a}-\sigma^2\right)\Rightarrow N\left(0,2\sigma^4\right).$$

Under strict regularity conditions and physical dependence measures SV estimator was shown to be normally distributed in Theorem 12. These results hold in time series context and have not been shown vigorously in MCMC literature. So, the contents of this chapter are more applicable in the context of time series.

4.3 Stein's Loss

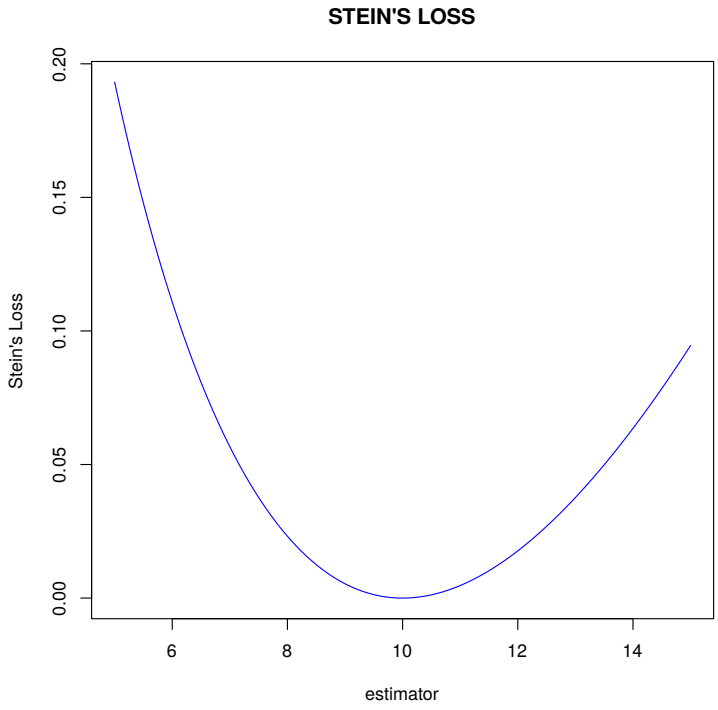

Figure 4.2: Expected Stein's loss, $\sigma^2_{true} = 10$.

Stein's loss is one of the prominent asymmetrical loss function that is the analogous of the squared error loss in least squares regression (See. Stein (1975), Dey and Srinivasan (1985), Loh (1988)). Sometimes it is also called Entropy loss in the literature and is normally related to estimating the mean vector of a d-dimensional normal distribution with known covariance matrix. It measures the Kullback-Leibler divergence between two multivariate normal distributions with the same means and covariances $\hat{\Sigma}$ and Σ. Stein's loss has been used in terms of sample covariance matrix estimation in wide array of literature including Ledoit and Wolf (2015) and Ledoit

and Wolf (2018).

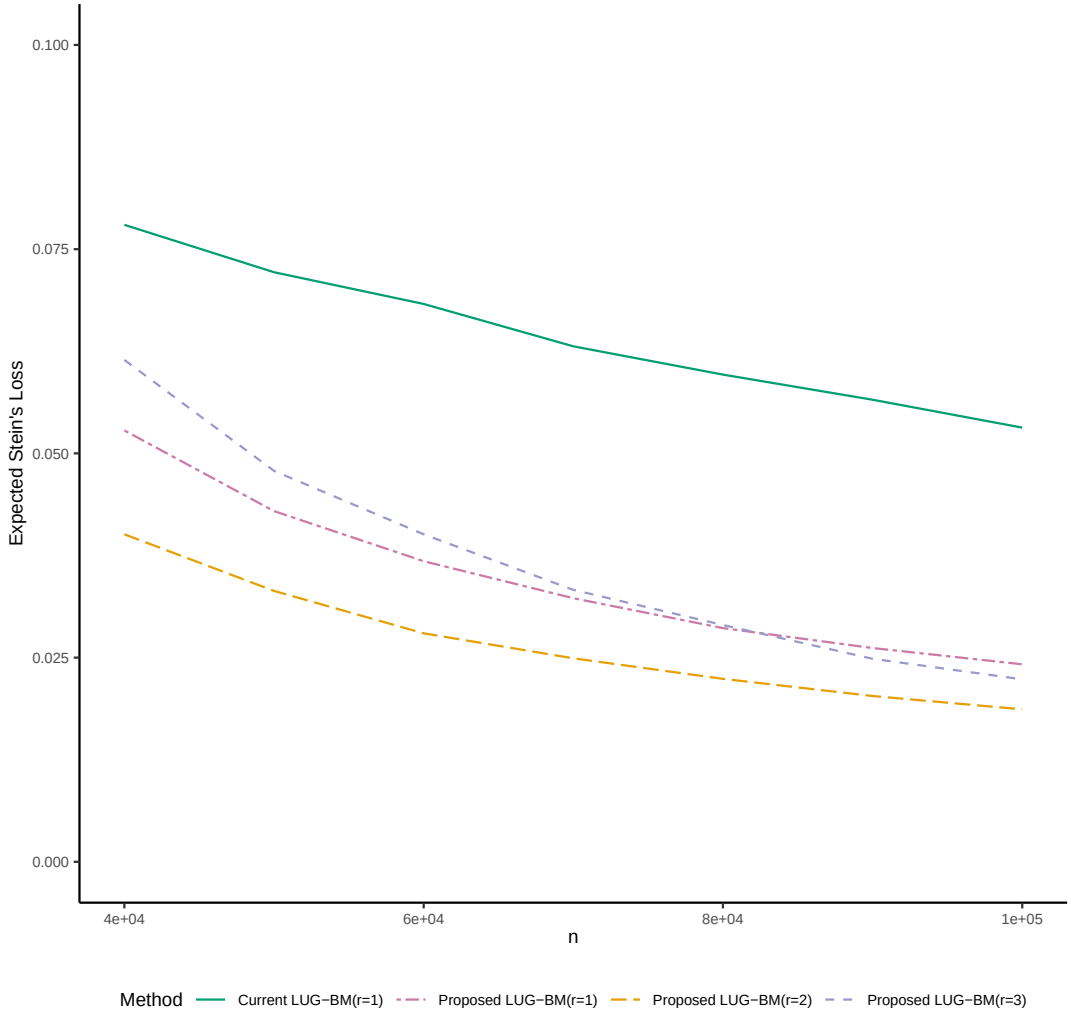

Figure 4.3: Expected Stein's loss for different estimation methods versus n.

In its simple form, the univariate Stein's loss can be written as

$$L(d, \theta) \propto \left(\frac{d}{\theta}\right) - \ln\left(\frac{d}{\theta}\right) - 1,$$

where d is the estimator and θ is the parameter we are estimating. For example,

the shape of Stein's loss for estimating a mean parameter centered at 10 is shown in Figure 4.2. It is asymmetrical and underestimation are more heavily penalized than overestimation.

Using Stein's loss to our variance estimators where the truth is known, we see in Figure 4.3 that proposed estimation methods of Chapter 2 seem to do well in terms of Stein's loss. The proposed LUG-BM and LUG-OBM estimators are constructed with $r = 3$ in linear combination in (1.23) and they are seen to perform worse than proposed FT-BM and FT-OBM methods. Also, noticeable is that the current methods for single estimators do not perform very well compared to the linear combination methods.

4.4 LINEX Loss

LINEX (LINear-EXponential) loss function is a convex, infinitely differentiable asymmetrical loss function that has a unique minimum (Zellner, 1986). It is a combination of linear and exponential function of the bias and can be controlled by an asymmetry and a scaling parameter. The optimal estimators under LINEX loss not only depends on the estimator itself but also on the variance and higher order moments of the distribution of the estimator. The problem of optimization under LINEX loss is straightforward given the nice smoothness properties.

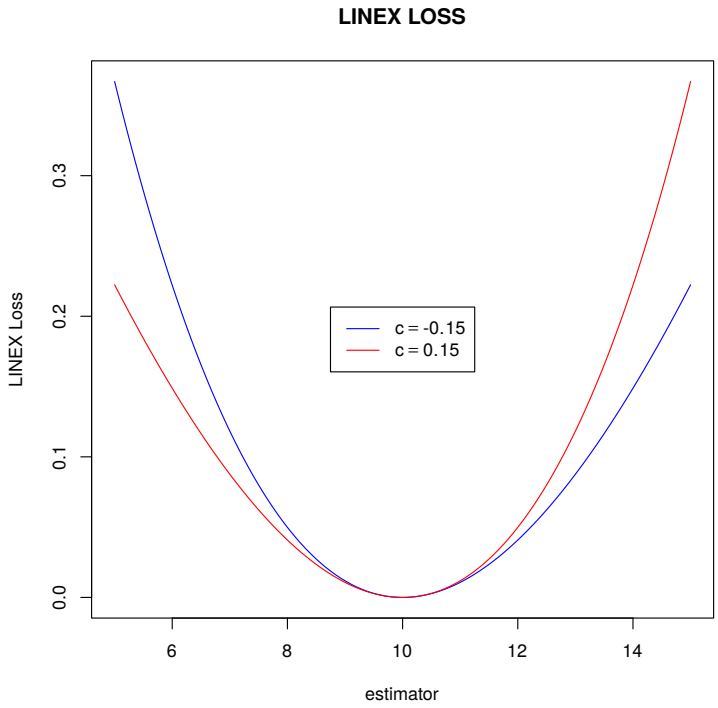

Figure 4.4: Expected LINEX loss for different c, $\sigma^2_{true} = 10$.

LINEX loss of using estimator $\hat{\Sigma}$ to estimate Σ is

$$L(\Delta) = k \cdot \exp(c \cdot \Delta) + d \cdot \Delta + e,$$

where $\Delta = \hat{\Sigma} - \Sigma$ and k, c, and e are some real parameters. For example, if we want to penalize underestimation more than overestimation, LINEX loss function is appropriate with $e = -k$, $d = -ck$, and $c < 0$ and the resulting loss function becomes

$$L(\Delta) = k \cdot (\exp(c \cdot \Delta) - c \cdot \Delta - 1). \tag{4.1}$$

Further, LINEX loss function can also be written in the following form with $k = 1/c^2$

$$L(\Delta) = \frac{1}{c^2}(\exp(c \cdot \Delta) - c \cdot \Delta - 1). \tag{4.2}$$

The form of LINEX loss in (4.2) is important because it allows us to see that LINEX loss converges to MSE loss as $c \to 0$ using Taylor series expansion and L'Hopital rule. We can see that

$$\frac{1}{c^2}(\exp(c\Delta) - c\Delta - 1) = \frac{1}{c^2}\left(\sum_{i=0}^{\infty} \frac{c^i \Delta^i}{i!} - c\Delta - 1\right) = \frac{1}{c^2}\left(\sum_{i=2}^{\infty} \frac{c^i \Delta^i}{i!}\right) \approx \frac{\Delta^2}{2},$$

where a factor of $1/2$ arises to account for the symmetry of the MSE loss.

The LINEX loss function for $\Sigma = 10$ is depicted in Figure 4.4 for two different asymmetry parameters. As we can see in the figure, for negative c, underestimation are more heavily penalized than overestimation and the opposite holds for positive c. In a way, this is an improvement over Stein's loss now that we have control over the direction of the asymmetry.

4.4.1 Simulation Study

We can calculate the expected linex loss of BM estimators in finite sample setting. The optimal integer-valued batch-size for BM estimators can be obtained by

minimizing the following naive linex loss function

$$\frac{\sum_{i=1}^{N} k \cdot \Big(\exp\big(c(\hat{\Sigma} - \Sigma)\big) - c(\hat{\Sigma} - \Sigma)\big) - 1 \Big)}{N},$$

where N is the number of replication. For simplicity, assume the value of c and k is known and optimize over the batch size by treating it as the only free parameter that $\hat{\Sigma}$ implicitly depends on. The expected LINEX loss BM estimator for $c = -0.005$ and $k = 0.1$ can be seen for different b in Figure 4.5 for four different Markov chains of length $n = 2e4$ constructed from AR(1) process.

Similarly, the MSE loss under the same setting for BM estimators can be seen in Figure 4.6. Remarkably, these figures show that the optimal batch size using the two alternate loss functions are same for small values of c. The only difference is the magnitude of the loss. LINEX loss is significantly smaller than MSE loss. So, using LINEX loss one can achieve a scaling in MSE loss for a particular b.

4.4.2 Expected LINEX loss

A simulation study was done to optimize the parameters of the LINEX loss for a particular bias of univariate modified lugsail estimator. Also, the optimal weight of the linear combination was optimized using a modification to the ridge penalty function. The two dimensional optimization can be easily done using L-BFGS-B

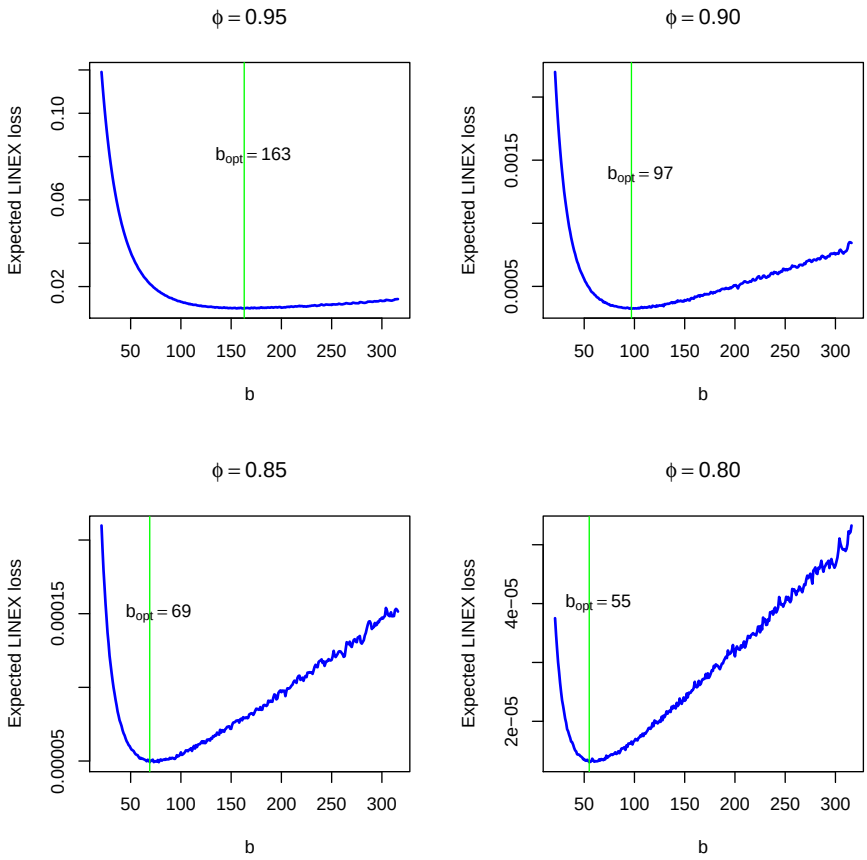

Figure 4.5: Batch-size optimization under LINEX loss of BM estimators for $n = 1e4$, $c = -0.005$, $k = 0.1$.

method in *optim* function in R. The objective function for minimization is

$$
\begin{aligned}
f(\boldsymbol{\alpha}) &= \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{\Sigma_i}, \boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(k\big(e^{c(\boldsymbol{\alpha}^T \cdot \boldsymbol{\Sigma_i} - \Sigma)} - c(\boldsymbol{\alpha}^T \cdot \boldsymbol{\Sigma_i} - \Sigma) - 1\big) \right) + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha}.
\end{aligned}
$$

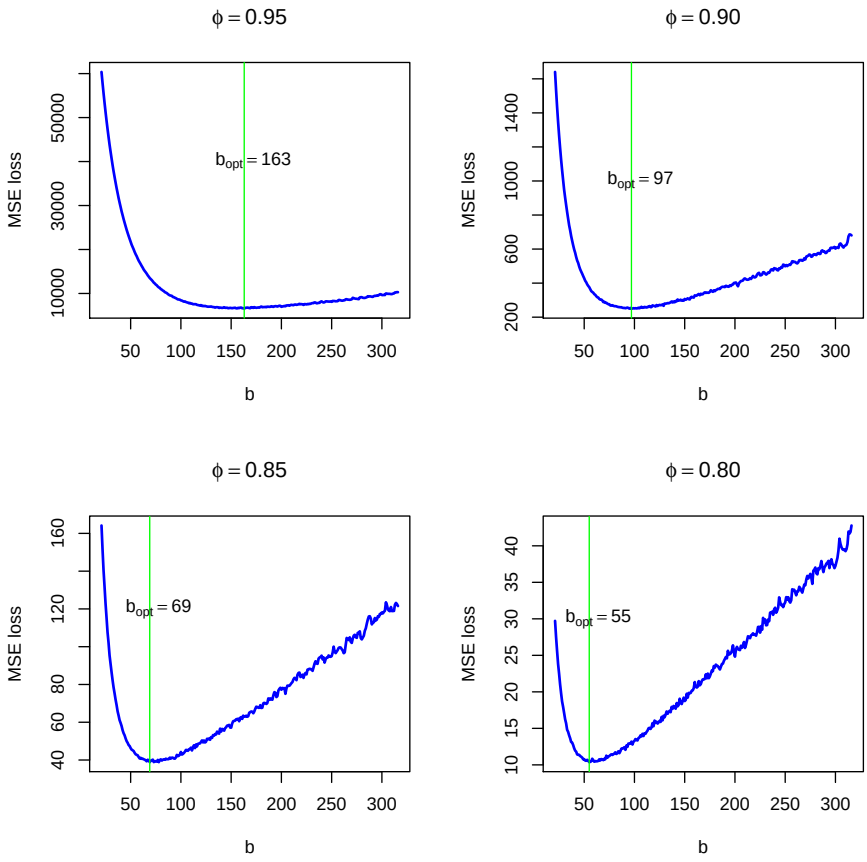

Figure 4.6: Batch-size optimization under MSE loss of BM estimators for $n = 1e4$, $c = -0.005$, $k = 0.1$.

The gradient of the objective function at a single point is

$$J(\boldsymbol{\alpha}) = \frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = c \cdot k \cdot \left[\left(e^{c(\boldsymbol{\alpha}^T \boldsymbol{\Sigma} - \Sigma)} - 1 \right) \cdot \boldsymbol{\Sigma}^T \right] + 2\lambda \cdot \boldsymbol{\alpha}^T.$$

The unconstrained 3-D optimization problem can be reduced to 2-D optimization using the sum constraint, $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The parameter space is reduced from $(\alpha_1, \alpha_2, \alpha_3)$ to (α_2, α_3) and $\alpha_1 = 1 - \alpha_2 - \alpha_3$. Further, the parameters (α_2, α_3) can be

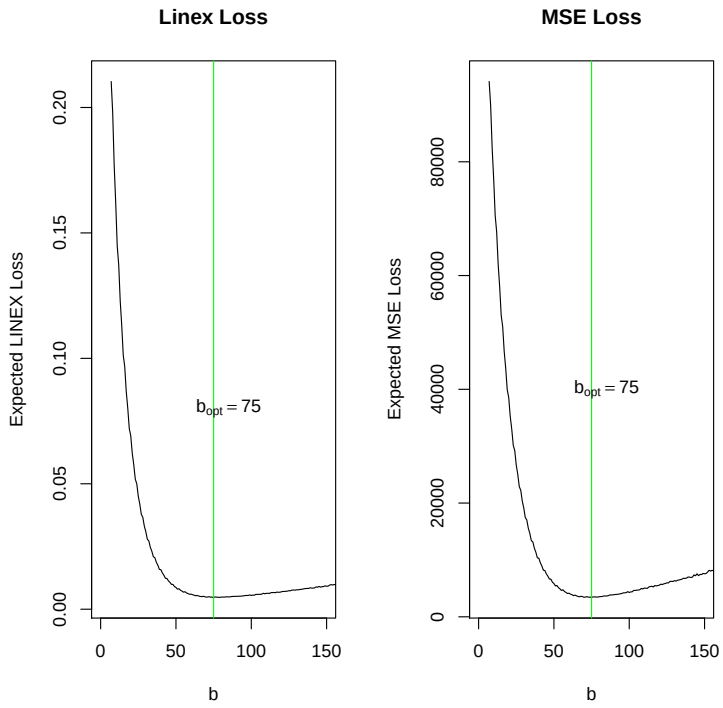

Figure 4.7: LINEX and MSE loss for $\phi = 0.95$ for Linear Combination estimators with $(\alpha_1, \alpha_2, \alpha_3) = (1.67, -0.33, -0.34)$.

constrained to be negative by using $\alpha_1 = 1 + \alpha_2 + \alpha_3$ in the objective function. This is a nice trick as it forces the parameters to be negative and also very close to 1; the degree of closeness is controlled by the ridge penalty parameter λ. The parameters are updated as

$$\boldsymbol{\alpha}_{i+1} = \boldsymbol{\alpha}_i - \eta \cdot J(\boldsymbol{\alpha_i}).$$

Instead of computing the gradient over a randomly chosen point, gradient is computed

over a randomly chosen batch

$$\boldsymbol{\alpha}_{i+1} = \boldsymbol{\alpha}_i - \eta \sum_{i=1}^{B} J(\boldsymbol{\alpha_i}),$$

where B is batch-size and η is the learning rate. These are hand-picked. The penalty parameter is calculated using k-fold cross-validation. The resulting optimal coefficients are $(\alpha_1, \alpha_2, \alpha_3) = (1.67, -0.33, -0.34)$ and the corresponding lag window in (2.26) becomes

$$\begin{aligned} w_n(k) =& 1.67\left(1 - \frac{|k|}{b}\right) I(0 \leq |k| \leq b) - 0.33\left(1 - \frac{|k|}{b/2}\right) I\left(0 \leq |k| \leq \frac{b}{2}\right) \\ & - 0.33\left(1 - \frac{|k|}{b/4}\right) I\left(0 \leq |k| \leq \frac{b}{4}\right). \end{aligned} \tag{4.3}$$

The lag window in (4.3) can also be written of the form as in (2.26)

$$\hat{\Sigma}_{LC} = \frac{1}{(1-c_1)(1-c_2)}\hat{\Sigma}_b - \frac{c_1}{(1-c_1)(1-c_2)}\hat{\Sigma}_{b/r} - \frac{c_2}{1-c_2}\hat{\Sigma}_{b/s}, \tag{4.4}$$

with $c_1 = 1/5$ and $c_2 = 1/4$.

Finally, using the optimal linear combination one can calculate the expected LINEX loss and MSE loss for a slow mixing Markov chain using AR(1) example. In Figure 4.11, we can see that for the optimal linear combination, the expected linex loss is reduced drastically compared to the MSE loss and, the bias decay is same for both the loss function. In other words, at fixed b, these two loss functions are a scaled

version of one another for some known parameters c and k. The parameter k does not play a crucial role and could be unity. Some guidance on how to pick the parameter c is discussed in the next sections.

4.4.3 A Thought Experiment

We will be limited to preliminary analytical analysis of LINEX and CHECK loss function in this chapter. Let's go through a simple thought experiment. Suppose X_1, $X_2, \cdots, X_n$ are i.i.d $N(\mu, \sigma^2)$, then the unbiased estimator of the population variance in the unknown mean case is given by

$$\hat{\sigma}^2_{UNBIASED} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We know that the MSE optimal estimator of the population variance is biased with an alternate scaling factor as

$$\hat{\sigma}^2_{MSE} = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

To find the LINEX optimal estimator, lets consider the alternate scaling version of the square deviations where m is the scaling factor as

$$\hat{\sigma}^2_{LINEX} = \frac{1}{m}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We know $\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$. Denote $\hat{\sigma}^2_{LINEX}$ by $\hat{\sigma}^2$ for simplicity. So, appropriately scaled version of the estimator of the variance follows Chi-square distribution with $n-1$ degrees of freedom, i.e.,

$$m \cdot \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}.$$

The expected LINEX loss for $\hat{\sigma}^2$ is derived in Section 4.9 and can be written as

$$E[L(\hat{\sigma}^2 - \sigma^2)] = ke^{-c\sigma^2}\left(1 - \frac{2c\sigma^2}{m}\right)^{-\frac{n-1}{2}} - \frac{kc\sigma^2}{m}(n-1) + kc\sigma^2 - k. \tag{4.5}$$

It is also shown in Section 4.9 that the approximate value of c that results in unbiased estimation in this toy example is

$$c \approx -\frac{n+1}{n-1} \cdot \frac{1}{\sigma^2}. \tag{4.6}$$

If the i.i.d. normal distribution of interest is taken to be $N(3, 3^2)$ and we simulate $n = 1000$ observations from it, we see in Figure 4.8 that we get an alternate scaling in terms of expected LINEX loss and the asymmetry parameter that results in unbiased estimation is $c_{opt} = -0.15$ and this approximately agrees with (4.6).

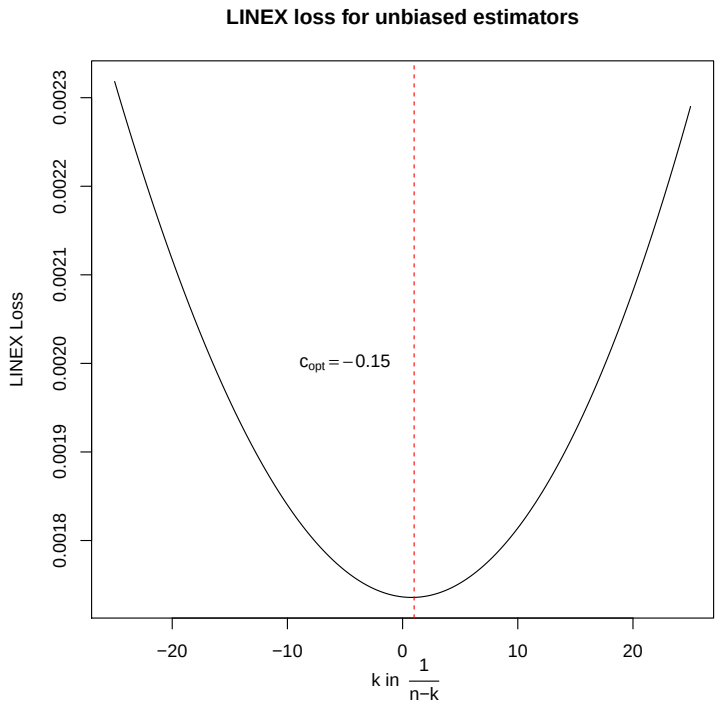

Figure 4.8: Optimal scaling for LINEX loss in a toy example with $X \sim N(3, 3^3)$.

4.4.4 LINEX Optimal Estimator

Suppose the error in estimation, or the sampling bias, is of the form $\Delta = \hat{\Sigma} - E[\hat{\Sigma}]$. We need to minimize this sampling bias in finite samples. Then, the expected loss in (4.2) is

$$E[L(\hat{\Sigma} - E[\hat{\Sigma}])] = E\bigg(k\big(\exp(c(\hat{\Sigma} - E[\hat{\Sigma}])) - c(\hat{\Sigma} - E[\hat{\Sigma}]) - 1\big)\bigg). \tag{4.7}$$

Taking derivative of (4.7) w.r.t. $\hat{\Sigma}$ and setting it equal to 0, we get

$$E\big[\exp(c(\hat{\Sigma}^{\text{LIN}} - E[\hat{\Sigma}]))\big] = 1. \tag{4.8}$$

Solving $\hat{\Sigma}^{\text{LIN}}$ from (4.8) that minimizes (4.7), we get

$$\hat{\Sigma}^{\text{LIN}} = -\frac{\ln E\big(\exp(-cE[\hat{\Sigma}])\big)}{c} = -\frac{\ln(M_{E[\hat{\Sigma}]}(-c))}{c} = -\frac{K_{E[\hat{\Sigma}]}(-c)}{c}, \tag{4.9}$$

where $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$ is the moment generating function whenever $X \sim N(\mu, \sigma^2)$ and $K_X(t)$ is the corresponding cumulant generating function. The optimal estimator under LINEX loss can be written in terms of cumulant generating function under the distribution of an estimator of Σ. Under the assumption of normality, we have

$$\hat{\Sigma}^{\text{LIN}} = -\frac{\ln(M_{\Sigma^{BM}}(-c))}{c} = E[\hat{\Sigma}] - \frac{\text{Var}(\hat{\Sigma})}{2} \cdot c. \tag{4.10}$$

For negative c, the resulting estimator that minimizes LINEX loss is larger than the one that minimizes the squared error loss. We may have extra terms if the normality distribution is not met, however those terms are very small and can effectively be controlled with suitable choice of c. For each value of bias and variance, we can calculate the value of c that minimizes the expected LINEX loss by treating c as random.

Consider $\hat{\Sigma}_b^{BM}$ and call it $\hat{\Sigma}$ for the sake of simplicity. The analysis in this section can equally be applied to lugsail estimators as single BM estimators are the simplest case of lugsail estimators with $r = 1$. Then, the finite sample bias of the estimator

can be written as

$$E\left[\hat{\Sigma}\right] = \Sigma - \left[\left(\Gamma^0_{n-1} - \Gamma^0_{b-1}\right) + \frac{1}{b}\Gamma^1_{b-1}\right] + O\left(\frac{1}{b}\right).$$

For a fixed b, denote the finite sample systematic bias of the estimator , $\hat{\Sigma}$, as

$$\mu_e = E[\hat{\Sigma}] - \Sigma = -\left[\left(\Gamma^0_{n-1} - \Gamma^0_{b-1}\right) + \frac{1}{b}\Gamma^1_{b-1}\right].$$

Then the expected LINEX loss of $\hat{\Sigma}$ around its expected value $E[\hat{\Sigma}]$ is

$$\begin{aligned}
E[L(\hat{\Sigma} - E[\hat{\Sigma}])] &= E\bigg(k\big(e^{c(\hat{\Sigma}-E[\hat{\Sigma}])} - c(\hat{\Sigma} - E[\hat{\Sigma}]) - 1\big)\bigg) \\
&= E\bigg(k\big(1 + c(\hat{\Sigma} - E[\hat{\Sigma}]) + \frac{c^2(\hat{\Sigma} - E[\hat{\Sigma}])^2}{2} + \frac{c^3(\hat{\Sigma} - E[\hat{\Sigma}])^3}{6} + \cdots \\
&\qquad \cdots - c(\hat{\Sigma} - E[\hat{\Sigma}]) - 1\big)\bigg) \\
&= \frac{k \cdot c^2}{2} E\big[(\hat{\Sigma} - E[\hat{\Sigma}])^2\big] + \frac{k \cdot c^3}{6} E\big[(\hat{\Sigma} - E[\hat{\Sigma}])^3\big] \\
&= \mathrm{I} + \mathrm{II}. \qquad (4.11)
\end{aligned}$$

Presented here are some preliminary results. For I, we see that

$$\begin{aligned}
E\big[(\hat{\Sigma} - E[\hat{\Sigma}])^2\big] &= E[(\hat{\Sigma} - \Sigma - (E[\hat{\Sigma}] - \Sigma))^2] \\
&= E(\hat{\Sigma} - \Sigma)^2 - 2E(\hat{\Sigma} - \Sigma) \cdot (E[\hat{\Sigma}] - \Sigma) + (E[\hat{\Sigma}] - \Sigma)^2 \\
&= \frac{1}{a-1}\left(\Sigma^2_{ij} + \Sigma_{ii}\Sigma_{jj}\right) - \mu_e^2.
\end{aligned}$$

Similarly, for II, we see that

$$\begin{aligned}
E\big[(\hat{\Sigma} - E[\hat{\Sigma}])^3\big] &= E[(\hat{\Sigma} - \Sigma - (E[\hat{\Sigma}] - \Sigma))^3] \\
&= E(\hat{\Sigma} - \Sigma)^3 - 3E(\hat{\Sigma} - \Sigma)^2 \cdot (E[\hat{\Sigma}] - \Sigma) + 3E(\hat{\Sigma} - \Sigma) \cdot (E[\hat{\Sigma}] - \Sigma)^2 - \\
&\qquad (E[\hat{\Sigma}] - \Sigma)^3 \\
&= 0 - 3E(\hat{\Sigma} - \Sigma)^2 \cdot (E[\hat{\Sigma}] - \Sigma) + 2(E[\hat{\Sigma}] - \Sigma)^3 \\
&= -\frac{3}{a-1}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) \cdot \mu_e + 2\mu_e^3.
\end{aligned}$$

The third central moment of the estimator, $E(\hat{\Sigma} - \Sigma)^3$, is taken to be zero from the normality assumption. Proving this result in the MCMC context requires a lot of extra technical conditions and is not pursued vigorously in this thesis. A simple proof using Brownian Motion analogue is provided in Section 4.9 together with the derivation of the various sums involved in (4.11). Combining I and II, we have

$$\begin{aligned}
E[L(\hat{\Sigma} - E[\hat{\Sigma}])] &= \frac{k \cdot c^2}{2} E\big[(\hat{\Sigma} - E[\hat{\Sigma}])^2\big] + \frac{k \cdot c^3}{6} E\big[(\hat{\Sigma} - E[\hat{\Sigma}])^3\big] \\
&= \frac{k \cdot c^2}{2}\left(\frac{1}{a-1}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) - \mu_e^2\right) + \\
&\qquad \frac{k \cdot c^3}{6}\left(-\frac{3}{a-1}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) \cdot \mu_e + 2\mu_e^3\right). \qquad (4.12)
\end{aligned}$$

Taking derivative of (4.12) w.r.t. c and equating it to 0, we have

$$c = \left(\frac{\frac{1}{a-1}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) - \mu_e^2}{\frac{3}{2(a-1)}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) \cdot \mu_e - \mu_e^3}\right). \qquad (4.13)$$

So, the expected LINEX loss in (4.7) is minimized for c in (4.13). The constant c is negative when we are underestimating and positive when we are overestimating when all other parameters are fixed. Using c in (4.10) we can achieve scaling of the estimators in terms of LINEX loss to adjust for the finite sample bias, although this is an approximate scaling when the estimator deviate from normality. We can easily extend this to lugsail estimator with $r > 1$ and this is left for future work.

A toy simulation in this regard was done to verify that the LINEX optimal coefficient in (4.12) is valid. We can see in Table 4.1 that at particular batch sizes that are taken to be close to MSE optimal batch sizes, we see an improvement to the bias results after adjustment using LINEX optimal coefficients calculated using (4.13) for all the lugsail estimators involved in Figure 4.1.

Bias adjustment with LINEX loss			
	$r = 1$	$r = 2$	$r = 3$
Batch Size	80	40	50
Old Bias	-9.23	-5.53	5.28
c	-0.38	-0.0916	0.1105
New Bias	0.3374	-2.65	0.06

Table 4.1: Comparison of MSE optimal and LINEX optimal bias

With similar reasoning for SV estimator, the value of c that minimizes the LINEX loss of the SV estimator in finite samples is

$$c = \left(\frac{\frac{2b}{3n}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) - \mu_e^2}{\frac{b}{n}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) \cdot \mu_e - \mu_e^3} \right).$$

This result cannot be applied in MCMC context at present and more research needs to be done in the future, but given the assumptions of Theorem 12 is met, this gives a way to adjust the bias in finite samples. It is presented here for completeness.

4.5 CHECK Loss

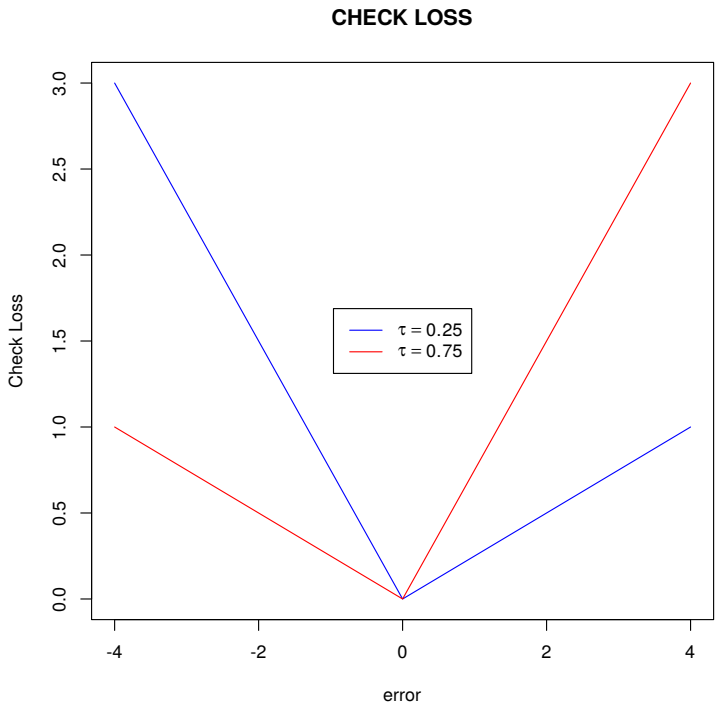

Figure 4.9: Expected CHECK Loss for different τ.

Check loss or LIN-LIN (LINear-LINear) loss is another simple asymmetrical loss function that is piece-wise linear and it has been widely used as an alternative to MSE loss in quantile regression and forecast literature, see Koenker and Bassett Jr (1978) and Elliott et al. (2005). Check loss minimizes the sum of positive and negative error

terms and is used when overestimating or underestimating is more desirable. It is additive scale-able, and separable. For $\tau \in (0,1)$, define check loss as

$$\rho_\tau(e) = [\tau - 1(e < 0)] \cdot e, \tag{4.14}$$

where $e = Y - m$ is an error term. As we can see in Figure 4.9 τ smaller than 0.5 result in over-penalization of underestimation and vice-versa. The expected CHECK loss can be written as

$$E[\rho_\tau(Y - m)] = \tau \int_m^\infty (y - m) f(y) dy - (1 - \tau) \int_{-\infty}^m (y - m) f(y) dy. \tag{4.15}$$

The minimizer of (4.15) is $q_\tau = m = F^{-1}(\tau)$.

Consider $\hat{\Sigma}_b^{BM}$ and let $e = \hat{\Sigma}_b^{BM} - \Sigma$ be the error in estimation. The error can be decomposed into systematic error and sampling error. The systematic error is irreducible and the sampling error has variability. We can expand the loss around the finite sample bias. Let the following be the finite sample bias and the standard deviation

$$\mu_e = E(e) = -\left[\left(\Gamma_{n-1}^0 - \Gamma_{b-1}^0\right) + \frac{1}{b}\Gamma_{b-1}^1\right]$$
$$\sigma_e = \sqrt{\frac{2\Sigma^2 b}{n}}.$$

Then, the expected loss of e is

$$\begin{aligned} E[L(e)] &= E[(\tau - 1_{\{e<0\}}) \cdot e] \\ &= \int (\tau - 1_{\{e<0\}}) \cdot e \cdot dF(e) \\ &= \tau\mu_e - \int_{-\infty}^{0} e \cdot dF(e). \end{aligned}$$

Assuming normality of the error distribution, and making a change of variable $e = \mu_e + \sigma_e z$, where z is a standard normal r.v., the above expression can be written as

$$\begin{aligned} E[L(e)] &= \tau\mu_e - \int_{-\infty}^{-\frac{\mu_e}{\sigma_e}} (\mu_e + \sigma_e z) \cdot dF(z) \\ &= \mu_e \left[\tau - F\Big(-\frac{\mu_e}{\sigma_e}\Big)\right] + \sigma_e \cdot f\Big(-\frac{\mu_e}{\sigma_e}\Big). \end{aligned} \tag{4.16}$$

4.6 Quad-quad Loss

Quad-Quad loss, also called asymmetrical quadratic loss, is a piece-wise quadratic loss function and can be written as

$$\rho_\tau(e) = [\tau - (2\tau - 1)1(e < 0)] \cdot e^2.$$

Quad-quad loss approaches MSE loss when $\tau = 1/2$. Similar to the case for Check loss, as we can see in Figure 4.10, τ smaller than 0.5 results in over-penalization

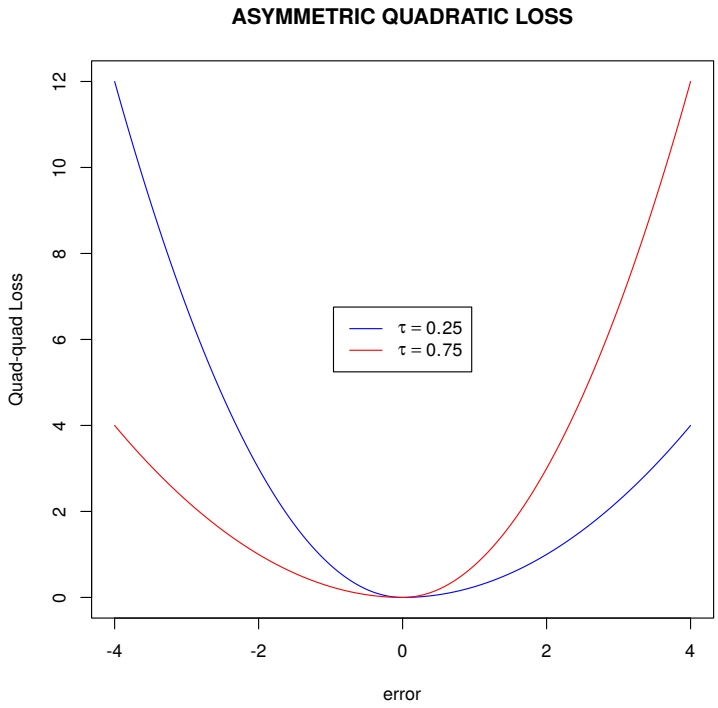

Figure 4.10: Expected Quad-quad loss for different τ.

of underestimation and vice-versa. So, it is easy to see that quad-quad loss is just an extension to MSE loss with unequal weights given to positive and negative error terms. The expected quad-quad loss is

$$E[L(e)] = \tau(\mu_e^2 + \sigma_e^2) - (2\tau - 1)\left[\mu_e \Phi\left(-\frac{\mu_e}{\sigma_e}\right) + 2\mu_e\sigma_e\left(-\phi\left(-\frac{\mu_e}{\sigma_e}\right)\right)\right.$$
$$\left. + \quad \sigma_e^2\left(\frac{\mu_e}{\sigma_e}\phi\left(-\frac{\mu_e}{\sigma_e}\right) + \Phi(-\frac{\mu_e}{\sigma_e})\right)\right]. \tag{4.17}$$

The proof for (4.17) is given in Section 4.9. And the optimal estimator that minimizes the expected quad-quad loss is

$$\hat{\sigma}^2_{opt} = \sigma^2 - \left(\frac{2\tau - 1}{\tau}\right) \cdot \sigma_e \cdot \phi\left(-\frac{\mu_e}{\sigma_e}\right).$$

4.7 Multivariate Loss

There exists a multivariate loss function from Komunjer and Owyang (2012) that is a generalization of check loss and it can be written as

$$L_p(\tau, \vec{e}) = (\|e\|_p + \alpha' e) \cdot \|e\|_p^{p-1}, \tag{4.18}$$

where $\alpha = 2\tau - 1$. This loss function reduces to multivariate check loss for $p = 1$. It reduces to multivariate MSE loss for $p = 2$ and $n = 1$. Here, n is the dimension of $\vec{e}$. For univariate case with $p = 1$, the loss function is just a scaled version of check loss as shown below

$$\begin{aligned} |e| + \alpha \cdot e &= (1 + \alpha \cdot \text{sign}(e)) \cdot |e| \\ &= 2(\tau + (1 - 2\tau)1(e < 0)) \cdot |e| \\ &= 2(\tau + 1(u < 0)) \cdot e. \end{aligned}$$

The multivariate loss in (4.18) with $n = 3$ and $p = 1$ reduces to linear combination

of three univariate check loss functions i.e.,

$$\begin{aligned} L_1(\alpha, e) &= |e_1| + |e_2| + |e_3| + \alpha_1 \cdot e_1 + \alpha_2 \cdot e_2 + \alpha_3 \cdot e_3 \\ &= |e_1| + +\alpha_1 \cdot e_1 + |e_2| + \alpha_2 \cdot e_2 + |e_3| + \alpha_3 \cdot e_3 \\ &= L_1(\alpha_1, e_1) + L_2(\alpha_2, e_2) + L_3(\alpha_3, e_3). \end{aligned}$$

This multivariate loss assumes that the individual losses are independent of each other and are thus separable. Quad-Quad loss function is not separable and serves limited practical applications. We can see that

$$L_2(\tau, e) = (\|e\|_2 + \tau' \cdot e) \|e\|_2 = (e_1^2 + e_2^2) + (\tau_1 e_1 + \tau_2 e_2)(e_1^2 + e_2^2)^{1/2}.$$

4.8 Scaling Estimators

Let $m\hat{\sigma}^2$ be a scaled estimator where m is a positive integer greater than 1. Then with the normality assumption, we can write

$$m\hat{\sigma}^2 \xrightarrow{d} N\left(m\sigma^2, \frac{2m^2\sigma^4 b}{n}\right). \tag{4.19}$$

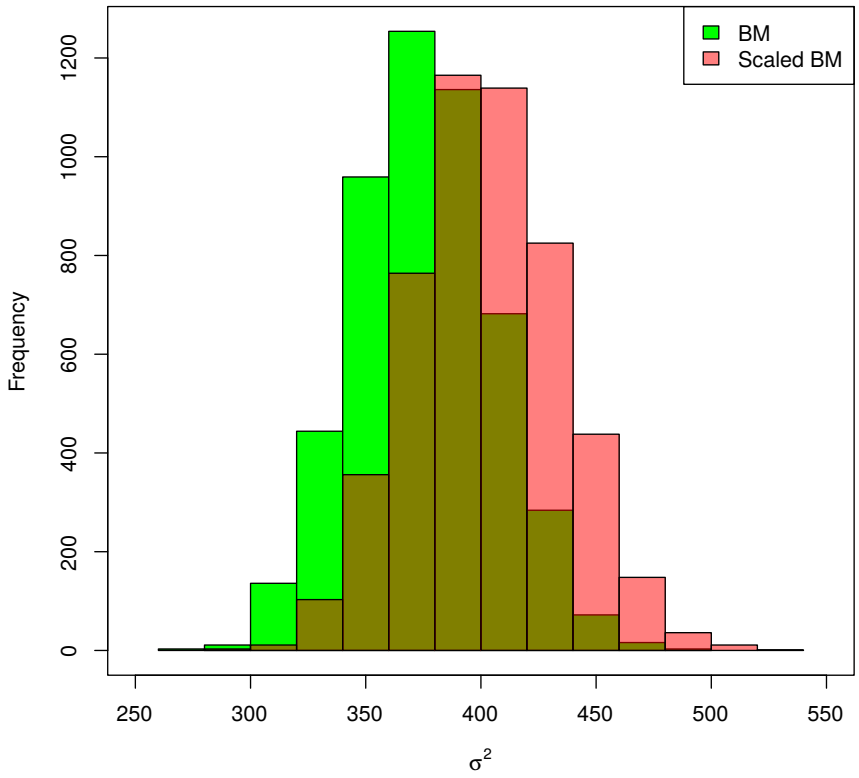

Figure 4.11: Scaled estimators on red with $\tau = 0.40$ and $m = 1.12$ from an AR(1) process with $\phi = 0.95$. Green histogram corresponds to unscaled estimator with $\sigma^2 = 400$

Let $e = m\hat{\sigma}^2 - \sigma^2$ be the error. Then the optimal estimator that minimizes the check loss for a given τ is given by

$$\hat{\sigma}^2_{opt} = \frac{\sigma^2}{m} + \sigma^2 \cdot \sqrt{\frac{2b}{n}} \cdot \mathrm{F}^{-1}(1-\tau),$$

where F is the Cumulative Distribution Function (CDF) of e. Assuming normality we can see in Figure 4.11, we can achieve scaling of univariate BM estimators so that

it's sampling bias is reduced in finite samples.

4.9 Proofs

Proof of (4.5) The expected LINEX loss for $\hat{\sigma}^2$ can be obtained as follows. The expected LINEX loss is

$$\begin{aligned}
E[L(\hat{\sigma}^2-\sigma^2)] &= E\bigg(k\big(e^{c(\hat{\sigma}^2-\sigma^2)}-c(\hat{\sigma}^2-\sigma^2)-1\big)\bigg) \\
&= k\cdot E\bigg(\big(e^{c\sigma^2\left(\frac{\hat{\sigma}^2}{\sigma^2}-1\right)}-c\sigma^2\left(\frac{\hat{\sigma}^2}{\sigma^2}-1\right)-1)\bigg) \\
&= ke^{-c\sigma^2}E[e^{c\sigma^2\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)}]-kc\sigma^2\left(E\left[\frac{\hat{\sigma}^2}{\sigma^2}\right]-1\right)-k \\
&= ke^{-c\sigma^2}E[e^{\frac{c\sigma^2}{m}\left(\frac{m\hat{\sigma}^2}{\sigma^2}\right)}]-\frac{kc\sigma^2}{m}\left(E\left[\frac{m\hat{\sigma}^2}{\sigma^2}\right]-m\right)-k
\end{aligned}$$

Using the moment generating function of Chi-square distribution, $M_X(t)=E[e^{tX}]=(1-2t)^{-\frac{n}{2}}$, whenever $X\sim\chi^2_n$, we have

$$E[L(\hat{\sigma}^2-\sigma^2)]=ke^{-c\sigma^2}\left(1-\frac{2c\sigma^2}{m}\right)^{-\frac{n-1}{2}}-\frac{kc\sigma^2}{m}(n-1)+kc\sigma^2-k.$$

Taking derivative w.r.t. m and equating it to 0, we have

$$\begin{aligned} e^{c\sigma^2} &= \left(1 - \frac{2c\sigma^2}{m}\right)^{-\frac{(n+1)}{2}} \\ \frac{1}{m} &= \frac{1 - e^{-\frac{2c\sigma^2}{n+1}}}{2c\sigma^2} \\ &= \frac{\frac{2c\sigma^2}{n+1} - \frac{4c^2\sigma^4}{(n+1)^2} + \frac{8c^3\sigma^6}{(n+1)^3} + O\left(\frac{1}{n^4}\right)}{2c\sigma^2} \\ &= \frac{1}{n+1} - \frac{2c\sigma^2}{(n+1)^2} + \frac{4c^2\sigma^4}{(n+1)^3} + O\left(\frac{1}{n^4}\right). \end{aligned}$$

So, $1/m$ increases when c is negative and approaches MSE optimal coefficient $1/(n+1)$ from above in the limiting case when c goes to 0. When c is positive, the opposite holds. This analysis advocates for using alternate scaling of our estimators to get unbiased estimators in finite samples. Approximately, for $m = n-1$, the value of c that results in unbiased estimation in this toy example is

$$c \approx -\frac{n+1}{n-1} \cdot \frac{1}{\sigma^2}.$$

Proof of (4.11) The scaled second central moment of $\hat{\Sigma}$ can be calculated using Brownian motion concepts.

$$E\left[(\hat{\Sigma} - \Sigma)^2\right] = E[\hat{\Sigma}^2] - 2\Sigma E[\hat{\Sigma}] + \Sigma^2$$

The scaled third central moment of $\hat{\Sigma}$ that measures the skewness is

$$E\big[(\hat{\Sigma}-\Sigma)^3\big] = E[\hat{\Sigma}^3] - 3\Sigma E[\hat{\Sigma}^2] + 3\Sigma^2 E[\hat{\Sigma}] - \Sigma^3$$

We make use of the following useful formulas.

$$\left(\sum_{i=0}^{n} a_i\right)^2 = \sum_{i=0}^{n} a_i^2 + 2\sum_{s=1}^{n}\sum_{i=0}^{n-s} a_i a_{i+s} \tag{4.20}$$

$$\left(\sum_{i=0}^{n} a_i\right)^3 = \sum_{i=0}^{n} a_i^3 + 3\sum_{s=1}^{n}\sum_{i=0}^{n-s} a_i^2 a_{i+s} + 3\sum_{s=1}^{n}\sum_{i=0}^{n-s} a_i a_{i+s}^2 + 6\sum_{t=1}^{n}\sum_{r=1}^{n-t}\sum_{i=0}^{n-t-r} a_i a_{i+t} a_{i+t+r} \tag{4.21}$$

Proposition 1. If $(X_1, \cdots, X_4)$ are jointly normally distributed with zero mean then the forth-order moment is:

$$E\left[X_1X_2X_3X_4\right] = E\left[X_1X_2\right]E\left[X_3X_4\right] + E\left[X_1X_3\right]E\left[X_2X_4\right] + E\left[X_1X_4\right]E\left[X_2X_3\right]$$

Proposition 2. If $(X_1, \cdots, X_6)$ are jointly normally distributed with zero mean

then the sixth-order moment given by Isserlis (1918) is:

$$
\begin{aligned}
&E[X_1X_2X_3X_4X_5X_6] \\
&= E[X_1X_2]E[X_3X_4]E[X_5X_6] + E[X_1X_2]E[X_3X_5]E[X_4X_6] + E[X_1X_2]E[X_3X_6]E[X_4X_5] \\
&+ E[X_1X_3]E[X_2X_4]E[X_5X_6] + E[X_1X_3]E[X_2X_5]E[X_4X_6] + E[X_1X_3]E[X_2X_6]E[X_4X_5] \\
&+ E[X_1X_4]E[X_2X_3]E[X_5X_6] + E[X_1X_4]E[X_2X_5]E[X_3X_6] + E[X_1X_4]E[X_2X_6]E[X_3X_5] \\
&+ E[X_1X_5]E[X_2X_3]E[X_4X_6] + E[X_1X_5]E[X_2X_4]E[X_3X_6] + E[X_1X_5]E[X_2X_6]E[X_3X_4] \\
&+ E[X_1X_6]E[X_2X_3]E[X_4X_5] + E[X_1X_6]E[X_2X_4]E[X_3X_5] + E[X_1X_6]E[X_2X_5]E[X_3X_4].
\end{aligned}
$$

Let $B(t)$ be a p-dimensional standard Brownian motion. Let $B^{(i)}$ be the i^{th} component of $B(t)$. Define $\tilde{\Sigma}$ as the Brownian motion counterpart of $\hat{\Sigma}$. Let $\bar{B} = n^{-1}B(n)$. Define Brownian motion increments as $U_t = B(t) - B(t-1)$ for $t = 1, ..., n$. Then, using $\bar{B}_l(k) = k^{-1}[B(lk+k) - B(lk)]$, for any batch l and batch size k, the Brownian motion difference can be decomposed into

$$\bar{B}_l^{(i)}(k) - \bar{B} = \left(\frac{n-k}{nk}\right) \sum_{t=l}^{l+k} U_t^{(i)} - \frac{1}{n} \sum_{t=1}^{l} U_t^{(i)} - \frac{1}{n} \sum_{t=l+k+1}^{n} U_t^{(i)}$$

We have $E[\bar{B}_l^{(i)}(k) - \bar{B}] = 0$ for $l = 0, ...a-1$ and

$$\text{Var}\left[\bar{B}_l^{(i)}(k) - \bar{B}^{(i)}\right] = \left(\frac{n-k}{nk}\right)^2 k + \frac{n-k}{n^2} = \frac{n-k}{nk}$$

We focus our attention for $k = b$ for the case of BM. The Brownian motion counterpart of BM estimator is

$$\tilde{\Sigma} = \frac{b}{a-1} \sum_{l=0}^{a-1} \left(\bar{B}_l(b) - \bar{B}\right) \left(\bar{B}_l(b) - \bar{B}\right)^T . \tag{4.22}$$

Then,

$$\bar{B}_l^{(i)}(b) - \bar{B}^{(i)} \sim N\left(0, \frac{n-b}{bn}\right)$$

and

$$\bar{B}_l(b) - \bar{B} \sim N\left(0, \frac{n-b}{bn} I_p\right)$$

Let $C(t)$ be the p-dimensional scaled Brownian motion i.e., $C(t) = LB(t)$. Let $C^{(i)}$ be the i^{th} component of $C(t)$. Let $\bar{C} = n^{-1}C(n)$ and $\bar{C}_l(k) = k^{-1}[C(lk+k) - C(lk)]$ as before, then

$$\bar{C}_l(b) - \bar{C} \sim N\left(0, \frac{n-b}{bn}\Sigma\right)$$

Let's focus on $E[\hat{\Sigma}^3]$. In terms of scaled Brownian motion, this term can be written

as,

$$E[\tilde{\Sigma}^3] = E\left[\left(\frac{b}{a-1}\sum_{l=0}^{a-1}\left(\bar{C}_l(b)-\bar{C}\right)\left(\bar{C}_l(b)-\bar{C}\right)^T\right)^3\right]$$

Consider each term of $E[\tilde{\Sigma}^3]$ from the expansion given in (4.21).

$$\begin{aligned} E[\tilde{\Sigma}^{3,ij}] &= E\left[\left(\frac{b}{a-1}\sum_{l=0}^{a-1}\left(\bar{C}_l^{(i)}(b)-\bar{C}^{(i)}\right)\left(\bar{C}_l^{(j)}(b)-\bar{C}^{(j)}\right)^T\right)^3\right] \\ &= \left(\frac{b}{a-1}\right)^3 E\left[\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}\right] \end{aligned} \tag{4.23}$$

$$\mathrm{I} = \sum_{l=0}^{a-1}\left(\left(\bar{C}_l^{(i)}(b)-\bar{C}^{(i)}\right)^3\left(\bar{C}_l^{(j)}(b)-\bar{C}^{(j)}\right)^3\right) = \sum_{l=0}^{a-1} Z_i^3 Z_j^3$$

where $Z_i = \left(\bar{C}_l^{(i)}(b)-\bar{C}^{(i)}\right)$ and $Z_j = \left(\bar{C}_l^{(j)}(b)-\bar{C}^{(j)}\right)$. Then,

$$\begin{bmatrix} Z_i \\ Z_j \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{n-b}{bn}\begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ij} & \Sigma_{jj} \end{bmatrix}\right)$$

$$
\begin{aligned}
E[\mathrm{I}] &= \sum_{l=0}^{a-1} E[Z_i^3 Z_j^3] = \sum_{l=0}^{a-1} E[Z_i Z_i Z_i Z_j Z_j Z_j] \\
&= \sum_{l=0}^{a-1} \left(9 \cdot E[Z_i^2] E[Z_i Z_j] E[Z_j^2] + 6 \cdot E[Z_i Z_j] E[Z_i Z_j] E[Z_i Z_j]\right) \\
&= a \cdot \left(\frac{n-b}{bn}\right)^3 \left(9\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} + 6\Sigma_{ij}^3\right) \\
&= a \left(\frac{1}{b^3} - \frac{3}{b^2 n} + \frac{3}{bn^2} - \frac{1}{n^3}\right) \left(9\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} + 6\Sigma_{ij}^3\right) \\
&= o\left(\frac{a^2}{b^3}\right) \qquad (4.24)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{II} &= 3 \sum_{s=1}^{a-1} \sum_{l=0}^{a-1-s} \left(\left(\bar{C}_l^{(i)}(b) - \bar{C}^{(i)}\right)^2 \left(\bar{C}_l^{(j)}(b) - \bar{C}^{(j)}\right)^2 \left(\bar{C}_{l+s}^{(i)}(b) - \bar{C}^{(i)}\right) \left(\bar{C}_{l+s}^{(j)}(b) - \bar{C}^{(j)}\right) \right) \\
&= 3 \cdot \sum_{s=1}^{a-1} \sum_{l=0}^{a-1-s} Z_1^2 Z_2^2 Z_3 Z_4
\end{aligned}
$$

where $Z_1 = \left(\bar{C}_l^{(i)}(b) - \bar{C}^{(i)}\right)$, $Z_2 = \left(\bar{C}_l^{(j)}(b) - \bar{C}^{(j)}\right)$, $Z_3 = \left(\bar{C}_{l+s}^{(i)}(b) - \bar{C}^{(i)}\right)$, and $Z_4 = \left(\bar{C}_{l+s}^{(j)}(b) - \bar{C}^{(j)}\right)$. The joint distribution of (Z_1, Z_2, Z_3, Z_4) can be written as

$$
\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \left(\frac{n-b}{bn}\right)\Sigma_{ii} & \left(\frac{n-b}{bn}\right)\Sigma_{ij} & -\frac{1}{n}\Sigma_{ii} & -\frac{1}{n}\Sigma_{ij} \\ & \left(\frac{n-b}{bn}\right)\Sigma_{jj} & -\frac{1}{n}\Sigma_{ij} & -\frac{1}{n}\Sigma_{jj} \\ & & \left(\frac{n-b}{bn}\right)\Sigma_{ii} & \left(\frac{n-b}{bn}\right)\Sigma_{ij} \\ & & & \left(\frac{n-b}{bn}\right)\Sigma_{jj} \end{bmatrix} \right)
$$

Then,

$$
\begin{aligned}
E[\mathrm{II}] &= 3\sum_{s=1}^{a-1}\sum_{l=0}^{a-1-s} E[Z_1^2 Z_2^2 Z_3 Z_4] \\
&= 3\cdot\sum_{s=1}^{a-1}\sum_{l=0}^{a-1-s}\left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj}\left(\left(\frac{n-b}{bn}\right)^3+\frac{8(n-b)}{bn^3}\right)+\Sigma_{ij}^3\left(2\left(\frac{n-b}{bn}\right)^3+\frac{4(n-b)}{bn^3}\right)\right) \\
&= \frac{3a(a-1)}{2}\left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj}\left(\left(\frac{n-b}{bn}\right)^3+\frac{8(n-b)}{bn^3}\right)+\Sigma_{ij}^3\left(2\left(\frac{n-b}{bn}\right)^3+\frac{4(n-b)}{bn^3}\right)\right) \\
&= \frac{3a(a-1)}{2}\left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj}\left(\frac{1}{b^3}-\frac{3}{b^2n}+\frac{11}{bn^2}-\frac{9}{n^3}\right)+\Sigma_{ij}^3\left(\frac{2}{b^3}-\frac{6}{b^2n}+\frac{10}{bn^2}-\frac{6}{n^3}\right)\right) \\
&= \frac{1}{2}\left(\frac{3a(a-1)}{b^3}\Sigma_{ii}\Sigma_{ij}\Sigma_{jj}+\frac{6a(a-1)}{b^3}\Sigma_{ij}^3\right)+o\left(\frac{a^2}{b^3}\right) \qquad (4.25)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{III} &= 3\cdot\sum_{s=1}^{a-1}\sum_{l=0}^{a-1-s}\left(\left(\bar{C}_l^{(i)}(b)-\bar{C}^{(i)}\right)\left(\bar{C}_l^{(j)}(b)-\bar{C}^{(j)}\right)\left(\bar{C}_{l+s}^{(i)}(b)-\bar{C}^{(i)}\right)^2\left(\bar{C}_{l+s}^{(j)}(b)-\bar{C}^{(j)}\right)^2\right) \\
&= 3\cdot\sum_{s=1}^{a-1}\sum_{l=0}^{a-1-s} Z_1 Z_2 Z_3^2 Z_4^2
\end{aligned}
$$

Similar to II,

$$
E[\mathrm{III}] = 3\sum_{s=1}^{a-1}\sum_{l=0}^{a-1-s} E[Z_1 Z_2 Z_3^2 Z_4^2] = \frac{1}{2}\left(\frac{3a(a-1)}{b^3}\Sigma_{ii}\Sigma_{ij}\Sigma_{jj}+\frac{6a(a-1)}{b^3}\Sigma_{ij}^3\right)+o\left(\frac{a^2}{b^3}\right) \qquad (4.26)
$$

$$\begin{aligned}\text{IV} &= 6 \cdot \sum_{t=1}^{a-1} \sum_{r=1}^{a-1-t} \sum_{l=0}^{a-1-t-r} \left(\bar{C}_l^{(i)}(b) - \bar{C}^{(i)}\right) \left(\bar{C}_l^{(j)}(b) - \bar{C}^{(j)}\right) \left(\bar{C}_{l+t}^{(i)}(b) - \bar{C}^{(i)}\right) \left(\bar{C}_{l+t}^{(j)}(b) - \bar{C}^{(j)}\right) \\ &\qquad \left(\bar{C}_{l+t+r}^{(i)}(b) - \bar{C}^{(i)}\right) \left(\bar{C}_{l+t+r}^{(j)}(b) - \bar{C}^{(j)}\right) \\ &= 6 \cdot \sum_{t=1}^{a-1} \sum_{r=1}^{a-1-t} \sum_{l=0}^{a-1-t-r} Z_1 Z_2 Z_3 Z_4 Z_5 Z_6\end{aligned}$$

where $Z_1 = \left(\bar{C}_l^{(i)}(b) - \bar{C}^{(i)}\right)$, $Z_2 = \left(\bar{C}_l^{(j)}(b) - \bar{C}^{(j)}\right)$, $Z_3 = \left(\bar{C}_{l+t}^{(i)}(b) - \bar{C}^{(i)}\right)$, $Z_4 = \left(\bar{C}_{l+t}^{(j)}(b) - \bar{C}^{(j)}\right)$, $Z_5 = \left(\bar{C}_{l+t+r}^{(i)}(b) - \bar{C}^{(i)}\right)$, $Z_6 = \left(\bar{C}_{l+t+r}^{(j)}(b) - \bar{C}^{(j)}\right)$. The joint distribution of $(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)$ is

$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \left(\frac{n-b}{bn}\right)\Sigma_{ii} & \left(\frac{n-b}{bn}\right)\Sigma_{ij} & -\frac{1}{n}\Sigma_{ii} & -\frac{1}{n}\Sigma_{ij} & -\frac{1}{n}\Sigma_{ii} & -\frac{1}{n}\Sigma_{ij} \\ & \left(\frac{n-b}{bn}\right)\Sigma_{jj} & -\frac{1}{n}\Sigma_{ij} & -\frac{1}{n}\Sigma_{jj} & -\frac{1}{n}\Sigma_{ij} & -\frac{1}{n}\Sigma_{jj} \\ & & \left(\frac{n-b}{bn}\right)\Sigma_{ii} & \left(\frac{n-b}{bn}\right)\Sigma_{ij} & -\frac{1}{n}\Sigma_{ii} & -\frac{1}{n}\Sigma_{ij} \\ & & & \left(\frac{n-b}{bn}\right)\Sigma_{jj} & -\frac{1}{n}\Sigma_{ij} & -\frac{1}{n}\Sigma_{jj} \\ & & & & \left(\frac{n-b}{bn}\right)\Sigma_{ii} & \left(\frac{n-b}{bn}\right)\Sigma_{ij} \\ & & & & & \left(\frac{n-b}{bn}\right)\Sigma_{jj} \end{bmatrix} \right)$$

Then,

$$
\begin{aligned}
E[\text{IV}] &= 6 \cdot \sum_{t=1}^{a-1} \sum_{r=1}^{a-1-t} \sum_{l=0}^{a-1-t-r} E[Z_1 Z_2 Z_3 Z_4 Z_5 Z_6] \\
&= 6 \cdot \sum_{t=1}^{a-1} \sum_{r=1}^{a-1-t} \sum_{l=0}^{a-1-t-r} \left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} \left(\frac{3(n-b)}{bn^3} - \frac{6}{n^3} \right) + \right. \\
&\qquad \left. \Sigma_{ij}^3 \left(\left(\frac{n-b}{bn} \right)^3 + \frac{3(n-b)}{bn^3} - \frac{2}{n^3} \right) \right) \\
&= a(a^2 - 3a + 2) \left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} \left(\frac{3(n-b)}{bn^3} - \frac{6}{n^3} \right) + \right. \\
&\qquad \left. \Sigma_{ij}^3 \left(\left(\frac{n-b}{bn} \right)^3 + \frac{3(n-b)}{bn^3} - \frac{2}{n^3} \right) \right) \\
&= a(a^2 - 3a + 2) \left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} \left(\frac{3}{bn^2} - \frac{9}{n^3} \right) + \Sigma_{ij}^3 \left(\frac{1}{b^3} - \frac{3}{b^2 n} + \frac{6}{bn^2} - \frac{6}{n^3} \right) \right) \\
&= \left(\frac{a(a-1)(a-2)}{b^3} - \frac{3a(a-1)(a-2)}{b^2 n} \right) \Sigma_{ij}^3 + o\left(\frac{a^2}{b^3} \right) \qquad (4.27)
\end{aligned}
$$

Combining (4.23),(4.24),(4.25),(4.26), and (4.27), we get,

$$
\begin{aligned}
E[\tilde{\Sigma}^{3,ij}] &= \left(\frac{3a}{(a-1)^2} \right) \Sigma_{ii}\Sigma_{ij}\Sigma_{jj} + \left(\frac{a^2 + a + 6}{(a-1)^2} \right) \Sigma_{ij}^3 + o\left(\frac{1}{a} \right) \\
&= \Sigma_{ij}^3 + \frac{3}{a-1} \left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} + \Sigma_{ij}^3 \right) + \frac{3}{(a-1)^2} \left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} + \Sigma_{ij}^3 \right) + o\left(\frac{1}{a} \right) \\
&= \Sigma_{ij}^3 + \frac{3}{a-1} \left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} + \Sigma_{ij}^3 \right) + o\left(\frac{1}{a} \right) \qquad (4.28)
\end{aligned}
$$

Let's focus on $E[\hat{\Sigma}^2]$. In terms of scaled Brownian motion, this term can be written

as,

$$E[\tilde{\Sigma}^2] = E\left[\left(\frac{b}{a-1}\sum_{l=0}^{a-1}\left(\bar{C}_l(b)-\bar{C}\right)\left(\bar{C}_l(b)-\bar{C}\right)^T\right)^2\right]$$

Each term can be written as,

$$\begin{aligned} E[\tilde{\Sigma}^{2,ij}] &= E\left[\left(\frac{b}{a-1}\sum_{l=0}^{a-1}\left(\bar{C}_l^{(i)}(b)-\bar{C}\right)\left(\bar{C}_l^{(j)}(b)-\bar{C}\right)\right)^2\right] \\ &= \left(\frac{b}{a-1}\right)^2 E[A+B] \end{aligned} \tag{4.29}$$

$$\mathrm{A} = \sum_{l=0}^{a-1}\left(\left(\bar{C}_l^{(i)}(b)-\bar{C}^{(i)}\right)^2\left(\bar{C}_l^{(j)}(b)-\bar{C}^{(j)}\right)^2\right) = \sum_{l=0}^{a-1} Z_i^2 Z_j^2$$

$$\begin{aligned} E[\mathrm{A}] &= \sum_{l=0}^{a-1} E[Z_i^2 Z_j^2] \\ &= a\cdot\left(\frac{n-b}{bn}\right)^2\left[2\Sigma_{ij}^2+\Sigma_{ii}\Sigma_{jj}\right] \\ &= a\left[\left(\frac{1}{b^2}+\frac{1}{n^2}-\frac{2}{bn}\right)\left[2\Sigma_{ij}^2+\Sigma_{ii}\Sigma_{jj}\right]\right] \\ &= \frac{a}{b^2}\left[2\Sigma_{ij}^2+\Sigma_{ii}\Sigma_{jj}\right]+o\left(\frac{a}{b^2}\right) \end{aligned} \tag{4.30}$$

$$\begin{aligned}\mathrm{B} &= 2 \cdot \sum_{s=1}^{a-1} \sum_{l=0}^{a-1-s} \left(\left(\bar{C}_l^{(i)}(b) - \bar{C}^{(i)} \right) \left(\bar{C}_l^{(j)}(b) - \bar{C}^{(j)} \right) \left(\bar{C}_{l+s}^{(i)}(b) - \bar{C}^{(i)} \right) \left(\bar{C}_{l+s}^{(j)}(b) - \bar{C}^{(j)} \right) \right) \\ &= 2 \cdot \sum_{s=1}^{a-1} \sum_{l=0}^{a-1-s} Z_1 Z_2 Z_3 Z_4\end{aligned}$$

$$\begin{aligned}E[\mathrm{B}] &= 2 \cdot \sum_{s=1}^{a-1} \sum_{l=0}^{a-1-s} E[Z_1 Z_2 Z_3 Z_4] \\ &= 2 \cdot \frac{a(a-1)}{2} \left[\left(\frac{n-b}{bn} \right)^2 \Sigma_{ij}^2 + \frac{\sum_{ii} \Sigma_{jj}}{n^2} + \frac{\sum_{ij}^2}{n^2} \right] \\ &= a(a-1) \left[\frac{\Sigma_{ij}^2}{b^2} + \frac{1}{n^2} \left[2\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj} \right] - \frac{2\Sigma_{ij}^2}{bn} \right] \\ &= \Sigma_{ij}^2 \left[\frac{(a-1)(a-2)}{b^2} \right] + o\left(\frac{a}{b^2} \right) \end{aligned} \tag{4.31}$$

Using (4.29), (4.30), and (4.31), we have,

$$\begin{aligned}E[\tilde{\Sigma}^{2,ij}] &= \frac{a(a-1)\Sigma_{ij}^2 + a\Sigma_{ii}\Sigma_{jj}}{(a-1)^2} + o\left(\frac{1}{a} \right) \\ &= \Sigma_{ij}^2 + \frac{1}{a-1} \left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj} \right) + o\left(\frac{1}{a} \right) \end{aligned} \tag{4.32}$$

Let's focus on $E[\hat{\Sigma}]$. In terms of scaled Brownian motion, this term can be written

as,

$$E[\tilde{\Sigma}] = E\left[\left(\frac{b}{a-1}\sum_{l=0}^{a-1}\left(\bar{C}_l(b)-\bar{C}\right)\left(\bar{C}_l(b)-\bar{C}\right)^T\right)\right]$$

Each term can be written as,

$$\begin{aligned}
E[\tilde{\Sigma}^{ij}] &= E\left[\left(\frac{b}{a-1}\sum_{l=0}^{a-1}\left(\bar{C}_l^{(i)}(b)-\bar{C}\right)\left(\bar{C}_l^{(j)}(b)-\bar{C}\right)\right)\right] \\
&= \frac{b}{a-1}\sum_{l=0}^{a-1}E\left[\left(\bar{C}_l^{(i)}(b)-\bar{C}\right)\left(\bar{C}_l^{(j)}(b)-\bar{C}\right)\right] \\
&= \frac{b}{a-1}\cdot a\cdot\left(\frac{n-b}{bn}\right)\Sigma_{ij} \\
&= \Sigma_{ij} \qquad (4.33)
\end{aligned}$$

Combining (4.33) and (4.32), we have

$$\begin{aligned}
E\left[(\tilde{\Sigma}_{ij}-\Sigma_{ij})^2\right] &= E[\tilde{\Sigma}_{ij}^2] - 2\Sigma E[\tilde{\Sigma}_{ij}] + \Sigma^2 \\
&= \Sigma_{ij}^2 + \frac{1}{a-1}\left(\Sigma_{ij}^2+\Sigma_{ii}\Sigma_{jj}\right) + o\left(\frac{1}{a}\right) - \Sigma_{ij}^2 \\
&= \frac{1}{a-1}\left(\Sigma_{ij}^2+\Sigma_{ii}\Sigma_{jj}\right) + o\left(\frac{1}{a}\right). \qquad (4.34)
\end{aligned}$$

Combining (4.33), (4.32), and (4.28), we have

$$\begin{aligned} E\big[(\tilde{\Sigma}_{ij}-\Sigma_{ij})^3\big] &= E[\tilde{\Sigma}_{ij}^3] - 3\Sigma E[\tilde{\Sigma}_{ij}^2] + 3\Sigma_{ij}^2 E[\tilde{\Sigma}_{ij}] - \Sigma_{ij}^3 \\ &= \Sigma_{ij}^3 + \frac{3}{a-1}\left(\Sigma_{ii}\Sigma_{ij}\Sigma_{jj} + \Sigma_{ij}^3\right) + o\left(\frac{1}{a}\right) \\ &\qquad - 3\Sigma_{ij}\left(\Sigma_{ij}^2 + \frac{1}{a-1}\left(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}\right) + o\left(\frac{1}{a}\right)\right) + 2\Sigma_{ij}^3 \\ &= 0. \end{aligned} \tag{4.35}$$

Proof of (4.17) The expected quad-quad loss is

$$\begin{aligned} E[L(e)] &= E[(\tau - (2\tau-1)1_{\{e<0\}})\cdot e^2] \\ &= \int (\tau - (2\tau-1)1_{\{e<0\}})\cdot e^2 \cdot dF(e) \\ &= \tau(\mu_e^2+\sigma_e^2) - (2\tau-1)\int_{-\infty}^{0}(\mu_e^2 + 2\mu_e\sigma_e z + \sigma_e^2 z^2)\cdot dF(e) \\ &= \tau(\mu_e^2+\sigma_e^2) - (2\tau-1)\Bigg[\int_{-\infty}^{-\frac{\mu_e}{\sigma_e}}\mu_e^2 dF(z) + 2\mu_e\sigma_e\int_{-\infty}^{-\frac{\mu_e}{\sigma_e}} z dF(z) + \\ &\qquad \sigma_e^2\int_{-\infty}^{-\frac{\mu_e}{\sigma_e}} z^2 dF(z)\Bigg] \\ &= \tau(\mu_e^2+\sigma_e^2) - (2\tau-1)\Bigg[\mu_e\Phi\left(-\frac{\mu_e}{\sigma_e}\right) + 2\mu_e\sigma_e\left(-\phi(-\frac{\mu_e}{\sigma_e})\right) + \\ &\qquad \sigma_e^2\left(\frac{\mu_e}{\sigma_e}\phi(-\frac{\mu_e}{\sigma_e}) + \Phi(-\frac{\mu_e}{\sigma_e})\right)\Bigg]. \end{aligned}$$

Then the optimal estimator that minimizes the expected quad-quad loss is

$$\begin{aligned}\hat{\sigma}^2_{opt} &= \sigma^2 + \left(\frac{2\tau - 1}{\tau}\right) \cdot \sigma_e \cdot \int_{-\infty}^{-\frac{\mu_e}{\sigma_e}} z dF(z) \\ &= \sigma^2 - \left(\frac{2\tau - 1}{\tau}\right) \cdot \sigma_e \cdot \phi\left(-\frac{\mu_e}{\sigma_e}\right).\end{aligned}$$

4.10 Concluding Remarks

Alternative loss functions that approximate MSE loss are useful substitute for unbiased estimation when we have an asymmetrical preference for overestimation and underestimation. To be able to get concrete answers from using these alternative loss functions, one needs to make certain distributional assumptions. Normality assumptions are valid with certain regularity conditions, but these regularity conditions are hard to justify in finite sample settings. We could do adhoc corrections to the bias and variance and improve the statistical properties of our estimators with the use of these alternative loss functions. More research in this regard is left for future work.

Chapter 5

Conclusions

Variance estimators in MCMC and time-series suffer from systematic error and sampling bias, especially in finite sample settings. Unbiased estimation is highly desirable in real-life applications and current methodologies that are based upon asymptotic unbiasedness do not perform very well in such settings. Estimators should be constructed at or near optimality of the parameters they that heavily depend upon and the properties of such estimators should closely resemble the data-generating mechanism. One of the most important parameter in the context of variance estimation is the batch size or the lag-trunction parameter. But the batch size that minimizes asymptotic MSE underperforms in finite sample simulations and we need to take into account the systematic error in finite samples in addition to sampling bias.

Special care shoud be done to assess the bias of the estimators in finite sample

settings. Knowing and estimating the systematic bias upto good precision, one can construct optimal variance estimators at or near a new optimal batch size that takes into account the inherent bias in the estimation process. This approach works both for single and linear combination estimators, although this bias is easier to control in linear combination estimators and the bias effect worsens as our sample size decreases.

In small sample settings, linear combination estimators should be employed to fully offset the systematic bias. Overestimating the variance is equally a point of concern in finite samples, although the estimators are asymptotically unbiased. These estimtors could be adjusted to be unbiased under some alternate loss fuctions like LINEX or Check loss. However, only preliminary works has been done in this regard in this thesis. Theoretical analysis regarding asymmetrical loss functions in the context of MCMC variance estimation is one obvious next step. Integrating all the variance estimation procedure and connecting the MCMC technical assumptions to the asymptotic theory of spectral estimation is another future research direction. I believe that this thesis lays good foundation for both of these future research endeavors.

Bibliography

Aktaran-Kalaycı, T., Alexopoulos, C., Argon, N. T., Goldsman, D., and Wilson, J. R. (2007). Exact expected values of variance estimators for simulation. *Naval Research Logistics (NRL)*, 54(4):397–410.

Aktaran-Kalaycı, T., Alexopoulos, C., Goldsman, D., and Wilson, J. R. (2009). Optimal linear combinations of overlapping variance estimators for steady-state simulation.

Anderson, T. W. (1971). *The statistical analysis of time series*, volume 19. John Wiley & Sons.

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

Berkes, I., Horváth, L., and Rice, G. (2016). On the asymptotic normality of kernel estimators of the long run covariance of functional time series. *Journal of multivariate analysis*, 144:150–175.

Bhansali, R. (1981). Effects of not knowing the order of an autoregressive process on the mean squared error of prediction—i. *Journal of the American Statistical Association*, 76(375):588–597.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys*, 2:107–144.

Braumann, A., Kreiss, J.-P., and Meyer, M. (2021). Simultaneous inference for autocovariances based on autoregressive sieve bootstrap. *Journal of Time Series Analysis*.

Brent, R. P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14(4):422–425.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods*. Springer-Verlag.

Chan, K. W. and Yau, C. Y. (2017). Automatic optimal batch size selection for recursive estimators of time-average covariance matrix. *Journal of the American Statistical Association*, 112(519):1076–1089.

Chanda, K. C. (2005). Large sample properties of spectral estimators for a class of stationary nonlinear processes. *Journal of Time Series Analysis*, 26(1):1–16.

Das, S., Subba Rao, S., and Yang, J. (2021). Spectral methods for small sample time series: A complete periodogram approach. *Journal of Time Series Analysis*.

Den Haan, W. J. and Levin, A. T. (1996). Inferences from parametric and non-parametric covariance matrix estimation procedures. *NBER Working Paper*, (t0195).

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, pages 1581–1591.

Elliott, G., Timmermann, A., and Komunjer, I. (2005). Estimation and testing of forecast rationality under flexible loss. *The Review of Economic Studies*, 72(4):1107–1125.

Flegal, J. M. (2008). *Monte Carlo standard errors for Markov chain Monte Carlo*. PhD thesis, University of Minnesota, School of Statistics.

Flegal, J. M., Haran, M., and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, 23(2):250–260.

Flegal, J. M., Huges, J., and Vats, D. (2015). Monte Carlo standard errors for MCMC R package version 1.0-1.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, 38(2):1034–1070.

Glynn, P. W. and Whitt, W. (1991). Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10(8):431–435.

Gupta, S. D. and Mazumdar, R. R. (2012). On the convergence of the spectral density of autoregressive approximations via empirical covariance estimates. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Gupta, V., Andradóttir, S., and Goldsman, D. (2014). Variance estimation and sequential stopping in steady-state simulations using linear regression. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 24(2):1–25.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probab. Surveys*, 1:299–320.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16:312–334.

Jordanger, L. A. and Tjøstheim, D. (2017). Nonlinear spectral analysis via the local gaussian correlation. *Nonlinear Spectrum Analysis based on the Local Gaussian Correlation and Model Selection for Copulas*.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kokoszka, P. and Jouzdani, N. M. (2020). Frequency domain theory for functional time series: Variance decomposition and an invariance principle. *Bernoulli*, 26(3):2383–2399.

Komunjer, I. and Owyang, M. T. (2012). Multivariate forecast evaluation and rationality testing. *Review of Economics and Statistics*, 94(4):1066–1080.

Kuelbs, J. and Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing b-valued random variables. *The Annals of Probability*, pages 1003–1036.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139:360 – 384.

Ledoit, O. and Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under stein's loss. *Bernoulli*, 24(4B):3791–3832.

Lin, Z.-Y. and Liu, W. (2012). M-dependence approximation for dependent random variables. In *Probability Approximations and Beyond*, pages 117–133. Springer.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274.

Liu, W. and Wu, W. B. (2010). Asymptotics of spectral density estimates. *Econometric Theory*, pages 1218–1245.

Liu, Y. and Flegal, J. M. (2018). Weighted batch means estimators in markov chain monte carlo. *Electronic Journal of Statistics*, 12(2):3397–3442.

Liu, Y., Vats, D., and Flegal, J. M. (2021). Batch size selection for variance estimators in mcmc. *Methodology and Computing in Applied Probability*, pages 1–29.

Loh, W.-L. (1988). *Estimating covariance matrices*. PhD thesis, Stanford University.

McMurry, T. L. and Politis, D. N. (2015). High-dimensional autocovariance matrices and optimal linear prediction. *Electronic Journal of Statistics*, 9(1):753–788.

Meyn, S. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 1st edition.

Panaretos, V. M., Tavakoli, S., et al. (2013). Fourier analysis of stationary time series in function space. *The Annals of Statistics*, 41(2):568–603.

Pedrosa, A. (1994). *Automatic Batching in Simulation Output Analysis*. PhD thesis, Purdue University, School of Industrial Engineering.

Pedrosa, A. C. and Schmeiser, B. W. (1993). Asymptotic and finite-sample correlations between obm estimators. In *Proceedings of the 25th Conference on Winter Simulation*, WSC '93, pages 481–488, New York, NY, USA. ACM.

Politis, D. N. and Romano, J. P. (1995). Bias-corrected nonparametric spectral estimation. *Journal of time series analysis*, 16(1):67–103.

Priestley, M. B. (1981). *Spectral analysis and time series: probability and mathematical statistics*. Number 04; QA280, P7.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surveys*, 1:20–71.

Rosenblatt, M. (1961). Independence and dependence. In *Proc. 4th Berkeley sympos. math. statist. and prob*, volume 2, pages 431–443.

Rosenblatt, M. (1972). Central limit theorem for stationary processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 551–561. University of California Press.

Schuster, A. (1897). On lunar and solar periodicities of earthquakes. *Proceedings of the Royal Society of London*, 61(369-377):455–465.

Shao, X. and Wu, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics*, 35(4):1773–1801.

Simonoff, J. S. (1993). The relative importance of bias and variability in the estimation of the variance of a statistic. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 42(1):3–7.

Song, W. T. and Schmeiser, B. W. (1993). Variance of the sample mean: Properties and graphs of quadratic-form estimators. *Operations Research*, 41(3):501–517.

Song, W. T. and Schmeiser, B. W. (1995). Optimal mean-squared-error batch sizes. *Management Science*, 41(1):110–123.

Stein, C. (1975). Estimation of a covariance matrix, rietz lecture. In *39th Annual Meeting IMS, Atlanta, GA, 1975*.

Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

Vats, D. and Flegal, J. M. (2018). Lugsail lag windows and their application to MCMC. *arXiv e-prints*, page arXiv:1809.04541.

Vats, D., Flegal, J. M., and Jones, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika*, 106(2):321–337.

White, H. (1984). Chapter vi - estimating asymptotic covariance matrices. In White, H., editor, *Asymptotic Theory for Econometricians*, Economic Theory, Econometrics, and Mathematical Economics, pages 29 – 60. Academic Press, San Diego.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

Wu, W. B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379.

Wu, W. B. and Zaffaroni, P. (2018). Asymptotic theory for spectral density estimates of general multivariate time series. *Econometric Theory*, 34(1):1–22.

Xiao, H. and Wu, W. B. (2011). Asymptotic inference of autocovariances of stationary processes. *arXiv preprint arXiv:1105.3423*.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, 81(394):446–451.