Diagnosing Learning Algorithms with Super-optimal Recursive Estimators (No. 704)

Man Fung Leung, Kin Wai Chan

Department of Statistics, The Chinese University of Hong Kong

Meaning of Recursive Estimation

Consider the estimation of sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ where the data X_i arrives sequentially. There are two ways to compute $\bar{X}_n$:

1. (Non-recursive) Calculate $(X_1 + X_2 + \ldots + X_n)/n$;
2. (Recursive) Calculate $\{(n-1)\bar{X}_{n-1} + X_n\}/n$.

The second way relies on the previous estimate $\bar{X}_{n-1}$ to update $\bar{X}_n$ and the number of operations is the same regardless of n. Hence we call it recursive or $O(1)$-time update. In addition, the second way only needs to store the values of $\bar{X}_n$ and n at each iteration, which involves a fixed amount of memory. Hence we call it $O(1)$-space update. Note that not all estimators can be recursively updated.

Introduction

Consider a stationary and ergodic process $\{X_i\}_{i\in\mathbb{Z}}$ with mean $\mu := \mathbb{E}(X_1)$ and autocovariance function (ACVF) $\gamma_k := \mathbb{E}\big\{(X_0-\mu)(X_k-\mu)\big\}, k \in \mathbb{N}$. As X_i are serially dependent, the variance of sample mean becomes the long-run variance (LRV), which can be expressed as

$$\sigma^2 = \sum_{k\in\mathbb{Z}} \gamma_k. \tag{1}$$

Estimating the LRV is thus crucial in accessing the error of many inference or learning procedures. In common applications like Markov chain Monte Carlo (MCMC) and stochastic gradient descent (SGD) where the sample size is not known *a priori*, sequential LRV estimates are often used to diagnose the convergence. Given the extensive use of these algorithms, seeking a LRV estimator that is both statistically and computationally efficient becomes important. Nevertheless, existing work faces an efficiency dilemma:

- Classical estimators that utilize overlapping batch means [4] and Bartlett kernel [1] are statistically efficient but need $O(n)$-time to update;
- Recursive estimators such as triangular (ΔSR) [8] and parallelogrammatic (PSR) selection rule [2] can be updated in $O(1)$-time but have higher asymptotic mean squared error (AMSE).

To facilitate discussion, we assume that $\mu = 0$ is known throughout the poster. The general definitions are presented in [6].

The Source of Efficiency Dilemma

To investigate the efficiency dilemma, we define a general class of estimator that includes both classical and recursive cases:

$$\hat{\sigma}_n^2(W) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} W_n(i,j)X_iX_j, \tag{2}$$

where $W_n(i,j)$ is a window function (also known as kernel). Note that the window can be further decomposed into two components:

$$W_n(i,j) = T\left(\frac{|i-j|}{t_n(i)}\right) S\left(\frac{|i-j|}{s_n(i)}\right), \tag{3}$$

where $T(\cdot) : [0,\infty) \to \mathbb{R}$ is the tapering function that scales the autocovariance estimates, $S(\cdot) : [0,\infty) \to \{0,1\}$ is the subsampling function that determines the truncation lag and $t_n(i), s_n(i)$ are their corresponding smoothing parameters. Under this construction, we notice that PSR, the existing most efficient recursive estimator, cannot control the tapering as the height of triangles is undesirably increasing. Bartlett kernel, on the other hand, cannot control the subsampling as the width of triangles is fixed globally; see Figure 1.

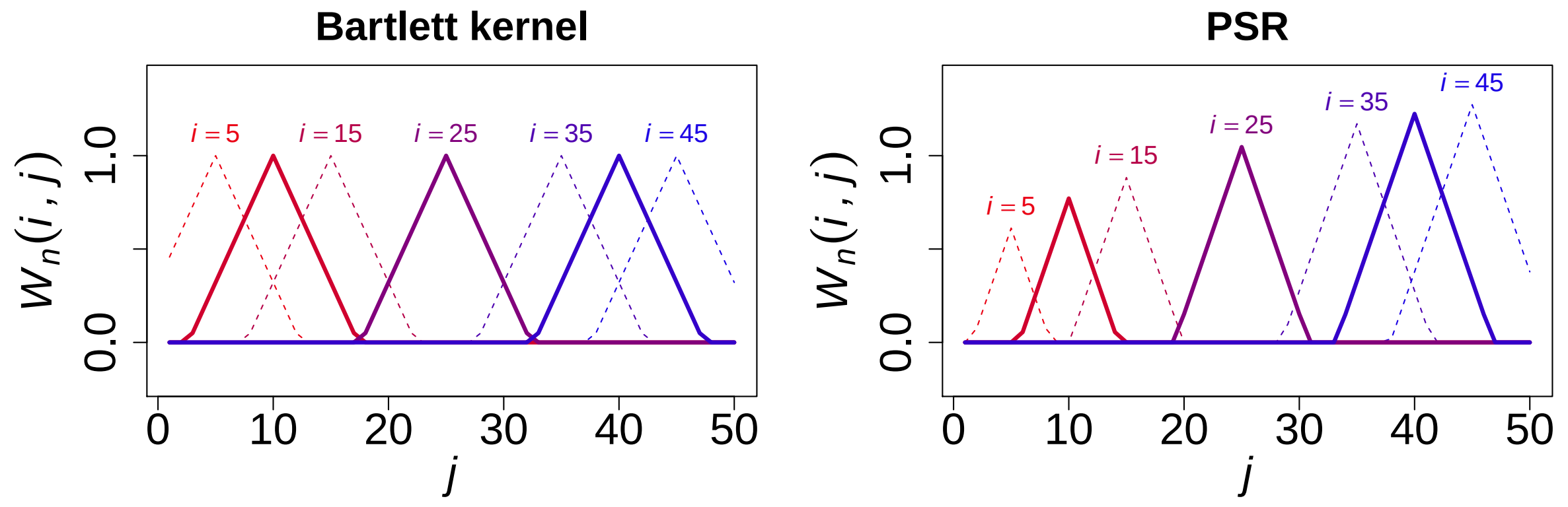

Figure 1:Comparison of the window under standard Bartlett kernel and PSR.

Estimation Principles

The source of efficiency dilemma motivates us to reconsider the roles of tapering and subsampling in LRV estimation. Therefore, we develop five step-by-step estimation principles which can be summarized as **LASER**. Here we only discuss the first three:

1. (**L**ocal Subsampling) An $O(1)$-time update algorithm should utilize local subsample, i.e., $s_n(i)$ should depend on i only.
2. (**A**synchronous Tapering) Under stationarity, (X_i, X_j) and $(X_{i'}, X_{j'})$ should receive the same scaling if $|i-j| = |i'-j'|$, i.e., $t_n(i)$ should depend on n.
3. (**S**eparated Parameters) The tapering and subsampling parameters should be separately chosen.

Their philosophy are as follows:

1. (**L**ocal Subsampling) Recursive estimates should be adapted to the present stage, i.e., the future (e.g., the future sample size n) should not affect the already computed estimates.
2. (**A**synchronous Tapering) When the distances are the same, the data pairs contain the same amount of information on covariance structure under stationarity and so they should be treated equally.
3. (**S**eparated Parameters) Statistical and computational efficiency are mainly determined by tapering and subsampling respectively but we may have different demands on them.

Proposed Estimators

Based on the principles of **LASER**, the general estimator is defined as

$$\hat{\sigma}^2_{n,\mathsf{LASER}(q,\phi)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{s_i'-1}\left(1 - \frac{k^q}{t_n^q}\right) X_iX_{i-k}, \tag{4}$$

where the new parameters mean

- $q \in \mathbb{Z}^+$: the characteristic exponent. The higher it is, the faster (4) converges subject to regularity conditions. This is possible by the sufficient condition for $O(1)$-time update derived with the principle of **E**xterior Tapering.
- $\phi \in [1,\infty)$: the memory parameter. When $\phi \geq 2$, updating (4) only involves a constant amount of memory. This is possible by the sufficient condition for $O(1)$-space update derived with the principle of **R**amped Subsampling. Note that ramped subsampling parameter is defined as

$$s_i' := \begin{cases} s_{i-1}'+1, & s_{i-1} \leq s_{i-1}'+1 < \phi s_{i-1}; \\ s_i, & s_{i-1}'+1 \geq \phi s_{i-1}. \end{cases} \tag{5}$$

Surprisingly, the AMSE of (4) can be super-optimal ($0.96B_n$); see Table 1. In addition, (4) can be updated at predetermined points $n_1, n_2, \ldots$ instead of every single points. By allowing users to select a common mini-batch size m such that $n_{j+1} - n_j = m$, we call this concept **mini-batch estimation** as in the machine learning literature. Through eliminating redundant operations and leveraging on vectorization, mini-batch estimators are much faster than existing recursive and non-recursive estimators; see Figure 3.

Summary

Table 1:Properties of different LRV estimators with $q = 1$

Estimator		Smoothing Parameters		Complexity		Statistical Efficiency		
$\hat{\sigma}_n^2$	ϕ	$s_n(i)$	$t_n(i)$	Time	Space	AMSE/σ^4	Relative	Bias^2/Var
Bartlett kernel, ‘B’	/	$(3/2)^{1/3}\kappa_1^{2/3}n^{1/3}$		$O(n)$	$O(n)$	$2.289\kappa_1^{2/3}n^{-2/3}$	B_n	0.5
PSR, ‘P’	/	$3^{1/3}\kappa_1^{2/3}i^{1/3}$		$O(1)$	$O(n^{1/3})$	$2.564\kappa_1^{2/3}n^{-2/3}$	$1.12B_n$	0.5
TSR, ‘T’	/	$(4/5)^{1/3}\kappa_1^{2/3}i^{1/3}$		$O(1)$	$O(1)$	$2.751\kappa_1^{2/3}n^{-2/3}$	$1.20B_n$	0.5
LASER, ‘E’	1	$(30/19)^{1/3}\kappa_1^{2/3}i^{1/3}$	$(13/12)s_n(n)$	$O(1)$	$O(n^{1/3})$	$2.204\kappa_1^{2/3}n^{-2/3}$	$\mathbf{0.96B_n}$	0.399
LASER, ‘R’	2	$(10/7)^{1/3}\kappa_1^{2/3}i^{1/3}$	$(8/7)s_n(n)$	$O(1)$	$O(1)$	$2.309\kappa_1^{2/3}n^{-2/3}$	$1.01B_n$	0.354

Key takeaways

1. The tapering and subsampling behaviors of a non-parametric method can be different. **LASER** provides guidance in selecting them for LRV estimation and achieves super-optimality.
2. Traditional recursive estimators can be extended to **mini-batch estimators**, which significantly improves the execution speed in practice.
3. Recursive estimators can be more efficient than non-recursive estimators.

Asymptotic Theory

We develop the asymptotic theory of (2) and (4) based on the dependence measures of [7]. Under regularity conditions:

1. ($\mathcal{L}^\alpha$ Consistency) Let $\alpha > 2$. Suppose that $X_1 \in \mathcal{L}^\alpha$, then
$$\left\|\hat{\sigma}_n^2 - \sigma^2\right\|_{\alpha/2} = o(1).$$
2. ($\mathcal{L}^2$ Convergence Rate) Let $\alpha \geq 4$. Suppose that $X_1 \in \mathcal{L}^\alpha$, $s_n(i) = \Psi i^\psi$ and $t_n(i) = \Theta n^\theta$, then
$$\mathrm{MSE}(\hat{\sigma}_n^2) \sim O(n^{-2/(1+2q)}),$$
provided that $\psi = \theta = 1/(1+2q)$ and $u_q = \sum_{j\in\mathbb{Z}} |j|^q|\gamma_j| < \infty$.
3. (AMSE-Optimal Parameters) Let $v_q = \sum_{j\in\mathbb{Z}} |j|^q\gamma_j < \infty$ and $\kappa_q = |v_q|/\sigma^2$. If $\psi = \theta = 1/(1+2q)$, the AMSE-optimal Ψ is given by
$$\Psi_\star = \begin{cases} \left\{\frac{(\phi+1)(2q+1)}{2q(q+1)} - \frac{4\left(\phi^{q+2}-1\right)(2q+1)}{(\phi-1)q(q+1)(q+2)(3q+2)} + \frac{\left(\phi^{2q+2}-1\right)}{2(\phi-1)q(q+1)(2q+1)}\right\}^{-1/(1+2q)} \kappa_q^{2/(1+2q)}, & \phi > 1; \\ \left\{\frac{2q+1}{q(q+1)} - \frac{4(2q+1)}{q(q+1)(3q+2)} + \frac{1}{q(2q+1)}\right\}^{-1/(1+2q)} \kappa_q^{2/(1+2q)}, & \phi = 1. \end{cases}$$
while the AMSE-optimal Θ is given by
$$\Theta_\star = \begin{cases} \left\{\frac{(q+2)(3q+2)(\phi^{2q+2}-1)}{4(2q+1)^2(\phi^{q+2}-1)} + \frac{\Psi_\star^{-2q-1}\kappa_q^2(\phi-1)(q+1)(q+2)(3q+2)}{4(2q+1)(\phi^{q+2}-1)}\right\}^{1/q} \Psi_\star, & \phi > 1; \\ \left\{\frac{(q+1)(3q+2)}{2(2q+1)^2} + \frac{\Psi_\star^{-2q-1}\kappa_q^2(q+1)(3q+2)}{4(2q+1)}\right\}^{1/q} \Psi_\star, & \phi = 1. \end{cases}$$
If $O(1)$-space update is required, the AMSE-optimal ϕ is 2. Otherwise, the AMSE-optimal ϕ is 1.

Note that **mini-batch estimation** is only updating (4) at $n = n_j$ so the statistical efficiency is unaffected. A numerical summary is given in Table 1.

Monte Carlo Experiments

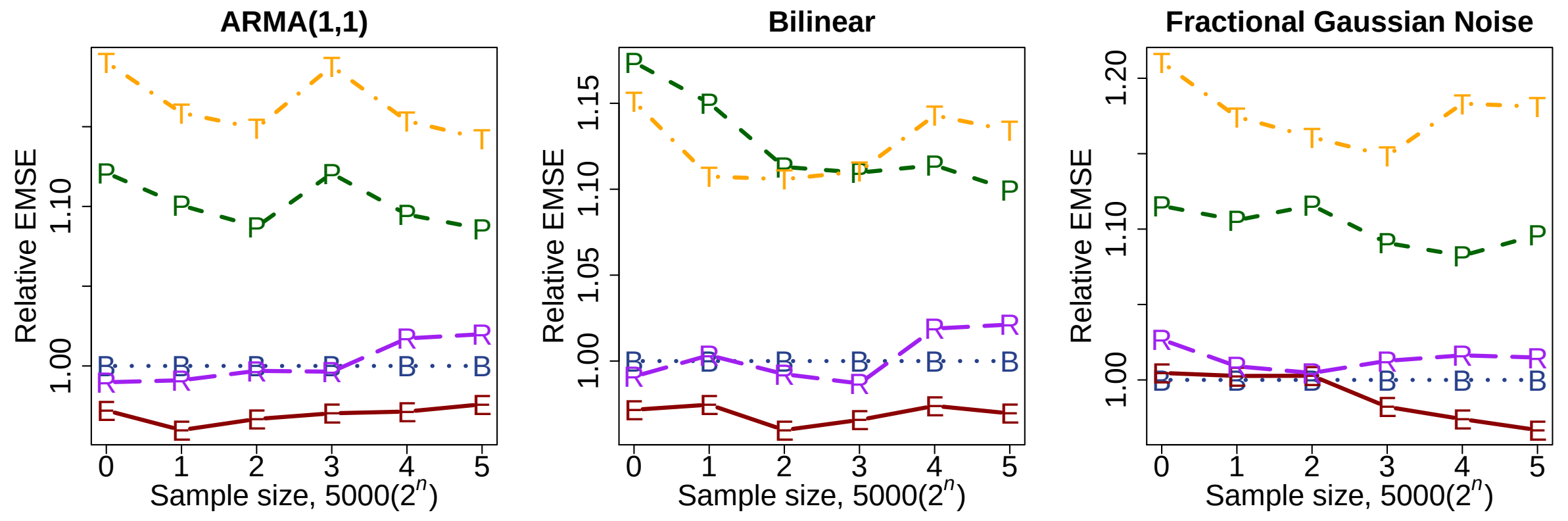

Figure 2:Comparison of the relative empirical MSEs under Bartlett kernel (‘B’), PSR (‘P’), TSR (‘T’), LASER(1,1) (‘E’) and LASER(1,2) (‘R’). The experiments are conducted based on 1000 replications. The finite sample performance matches with asymptotic theory; see Table 1.

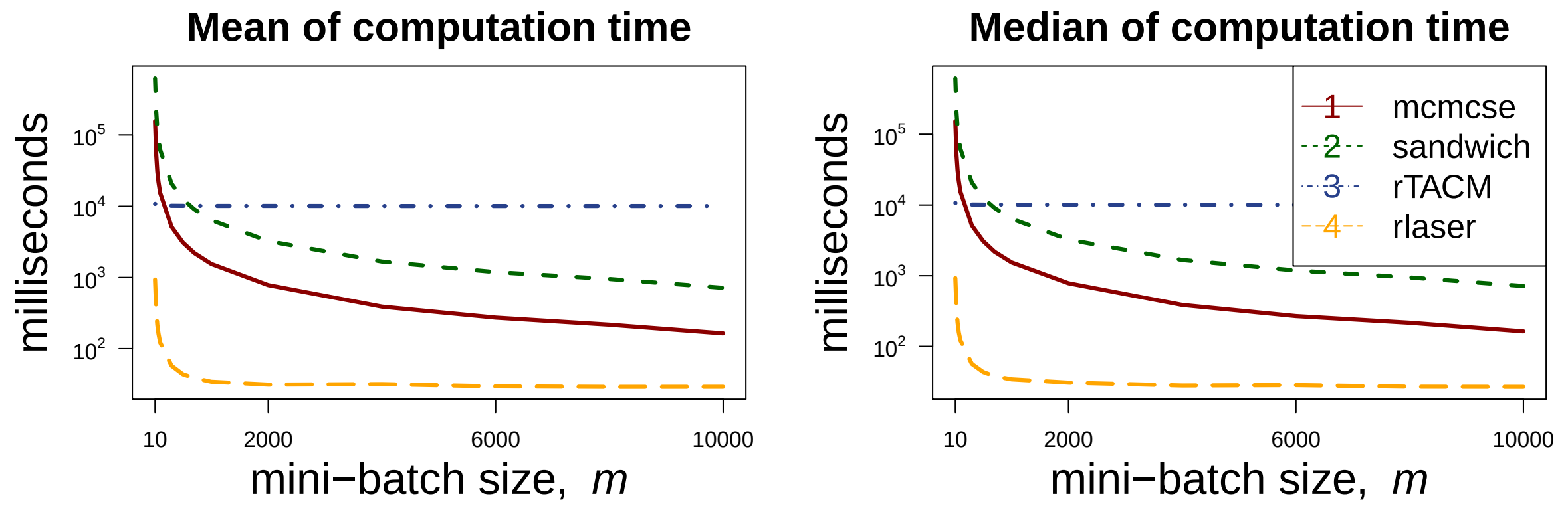

Figure 3:Comparison of the computation time under existing implementations of Bartlett kernel (sandwich), overlapping batch means (mcmcse), PSR (rTACM) and mini-batch LASER (rlaser) in R. The experiment is conducted based on 50 replications and 100,000 samples.

Applications and Conclusion

Interestingly, the use of LRV estimators differs between the statistics and engineering communities. Here we try to include both views to conclude our poster. For statistician, in particular those who work on Bayesian analysis, they usually use LRV estimators to diagnose MCMC convergence; see, e.g., [4]. In the half-width analysis, we can terminate a simulation at

$$n^* = \inf\left\{n \in \mathbb{Z}^+ : z_{1-\alpha/2}n^{-1/2}\hat{\sigma}_n + p(n) \leq \epsilon\right\}, \tag{6}$$

where $\alpha \in (0,1)$ is the significance level, $z_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$ lower quantile of $\mathrm{N}(0,1)$, $p(n)$ is a penalty function for n that is too small and $\epsilon > 0$ is the maximum tolerable error. Essentially, the half-width analysis is based on the Central Limit Theorem (CLT) for sample mean $\bar{X}_n$ and stops when the $100(1-\alpha/2)\%$-confidence interval is small enough. As mean functionals appear frequently in stochastic approximation, this idea can be extended to other learning algorithms apart from MCMC naturally; see, e.g., [3].

On the other hand, the engineers, in particular those who work on deep learning, develop a different direction. By viewing LRV as a measure of precision, we can try to improve the learning procedure proactively by tuning the learning rate; see, e.g., [5]. While these adaptive learning procedures are usually also based on CLT, we can see that it represents a different philosophy in using LRV estimators.

Regardless of the views, sequential LRV estimates are used in their procedures to diagnose learning algorithms. By discussing the estimation principles **LASER**, the concept of **mini-batch estimation** and other issues, we hope to echo the theme of *Advancing Statistics for Data Intelligence* and demonstrate the utility of recursive LRV estimators to you.

Acknowledgement

We thank Nathaniel Johnston for the LaTeX template of this poster, which is licensed under CC BY-NC-SA 3.0.

Software

rlaser: an R package for Recursive Long-run Variance Estimation. 2020+.

References

[1] Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.

[2] Kin Wai Chan and Chun Yip Yau. Automatic optimal batch size selection for recursive estimators of time-average covariance matrix. *Journal of the American Statistical Association*, 112(519):1076–1089, 2017.

[3] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48(1):251–273, 2020.

[4] James M Flegal and Galin L Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

[5] Hunter Lang, Lin Xiao, and Pengchuan Zhang. Using statistics to automate stochastic optimization. In *Advances in Neural Information Processing Systems 32*, pages 9540–9550, 2019.

[6] Man Fung Leung and Kin Wai Chan. General and super-optimal recursive estimators of long-run variance. 2020+.

[7] Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005.

[8] Wei Biao Wu. Recursive estimation of time-average variance constants. *Annals of Applied Probability*, 19(4):1529–1552, 08 2009.

Contact Information

- Web: https://hemanlmf.github.io
- Email: heman@link.cuhk.edu.hk
- Phone: +852 3943 8535

香港中文大學
The Chinese University of Hong Kong

Department of Statistics