Biometrika (2021), **103**, 1, *p*. 1

Advance Access publication on 31 July 2018

Printed in Great Britain

Discussion of 'Statistical inference for streamed longitudinal data'

By J. Wang, H. Wang, K. Chen
Department of Statistics, University of Connecticut, Storrs, Connecticut, 06269, U.S.A.
jing.7.wang@uconn.edu, haiying.wang@uconn.edu, kun.chen@uconn.edu

1. INTRODUCTION

We congratulate Luo et al. (2023) for proposing the innovative statistical framework with time-varying regression coefficient to analyze correlated streaming data. They integrated a weighting matrix into the quadratic inference function (Qu et al., 2000), giving larger weights to more recent data. An elegant decomposition of the first-order autoregression base matrices across batches was also introduced, enabling a computationally feasible online updating algorithm.

Faced with massive data, a key to extract useful information is to balance computational efficiency and statistical efficiency. Besides the approach by Luo et al. (2023), subsampling is another effective approach to achieve computational feasibility with some compromise on estimation efficiency (Wang et al., 2018, 2019). We point out that subsampling is more effective for correlated data than for independent data, because the correlation in selected subsample may become negligible. Ignoring correlation may then be a valid option to simplify the computation as comparing to the full data analysis. When the data generating parameter changes across data batches, we show that there is a bias-variance tradeoff that is affected by the smoothness of the parameter function and the weights assigned to historical batches. We discuss the optimal rate of the weights in a simplified case.

2. IGNORE THE CORRELATION?

In dealing with massive data, it's typically inevitable that some estimation efficiency is sacrificed in exchange for computational feasibility. Although some methods may achieve the same estimation efficiency as the offline full data estimator in an asymptotic sense, there is often noticeable efficiency loss with finite samples. Since a key point in analyzing massive data is the trade-off between estimation efficiency and computational feasibility, it is natural to wonder whether we can ignore the correlation. For simplicity, we illustrate the ideas using the linear regression model,

$$Y = X\beta + \varepsilon, \tag{1}$$

where the response vector Y and the design matrix X may have some batch structure, and ε has mean zero and covariance matrix Σ.

The most efficient estimator under mild conditions is the weighted least squares estimator,

$$\hat{\beta}_w = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y, \tag{2}$$

if Σ is known. Let the ordinary least squares estimator be

$$\hat{\beta}_o = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y. \tag{3}$$

Although $\hat{\beta}_w$ is more efficient than $\hat{\beta}_o$, the latter is also unbiased and still widely used in practice due to its simplicity. Their covariance matrices satisfy

$$V(\hat{\beta}_o \mid X) - V(\hat{\beta}_w \mid X) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\Sigma X(X^{\mathrm{T}}X)^{-1} - (X^{\mathrm{T}}\Sigma^{-1}X)^{-1} = gg^{\mathrm{T}} \geq 0,$$

where $g = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1/2} - (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\Sigma^{1/2}$. If Σ has a first-order autoregressive structure $\Sigma_{ij} = \sigma^2\rho^{|i-j|}$, then $V(\hat{\beta}_o \mid X)$ and $V(\hat{\beta}_w \mid X)$ get closer as ρ gets closer to 0. As such, the ordinary least squares estimator $\hat{\beta}_o$ may provide a better trade-off if ρ is small.

Subsampling is an effective approach to extract useful information when the full data analysis is computationally expensive or even infeasible. Existing investigations on subsampling methods mainly focus on independent data. Inspired by Luo et al. (2023), we want to stress that the advantage of subsampling can be more significant with correlated data than with independent data, because subsampling can help reduce correlation and simplify the data analysis. Again, consider the autoregressive correlation structure. If we take a subsample of size s from the total of $N = bn_j$ correlated observations with equal time gaps, then the correlation between adjacent observations in the subsample reduces to $r^{N/s}$, which goes to zero fast as N/s gets large. Intuitively, measurements taken in a close time range tend to contain similar information, so removing certain observations collected in similar times may not lead to as significant lost of information as in the independent data case.

2.1. *Numerical illustration*

We perform limited simulations on two scenarios to illustrate the effectiveness of subsampling with correlated data streams. Scenario 1 is the same as the example in Section 4.2 of Luo et al. (2023) except that the regression coefficient is a constant vector $\beta = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (0.2, -0.5, 0.5)^{\mathrm{T}}$. The two covariates in this scenario are independent standard normal random variables. For longitudinal data, the variation of covariates across subjects is often larger than that within subjects. For example, variables such as body mass index and presence of some coronary diseases usually do not change much for the same subject while they may vary greatly for different subjects, and covariates such as sex and high school education status are typically time invariant. To mimic this situation, we consider Scenario 2 with the same set up of Scenario 1 except that the two covariates have different distributions. Specifically, the covariate of β_1 for the kth $(= 1, ..., n_j)$ observation of subject i $(= 1, ..., m)$ at batch j $(= 1, ..., b)$ is $Z_i + g_l$ with Z_i $(i = 1, ..., m)$ being independent standard normal random variables and g_l $(l = 1, ..., bn_j)$ being bn_j evenly spaced scalars in $[2, 3]$; and the covariate for β_2 follows a Bernoulli distribution of success probability 0.5 across the $m = 100$ subjects and stays a constant across the $b = 200$ batches of $n_j = 20$ observations.

We implemented the ordinary least squares and the weighted least squares based on the full data of size $mbn_j = 400,000$, a deterministic uniform subsample of size $s = 2,000$, and the last batch of $mn_j = 2000$ observations. We run the simulation 500 times and calculate the mean squared error of estimating the mean responses for each method; the results are reported in Table 1. There is indeed some estimation efficiency loss with the ordinary least squares estimator compared with the weighted least squares estimator for full data analyses, but the information loss is not substantial, especially in the more practical Scenario 2. The weighted least squares estimator uses the unknown ρ, so it is optimal but impractical. With uniform subsamples, the ordinary and weighted least squares estimators do not have noticeable difference. Another interesting observation is that the subsample estimator uses only 0.5% of the full data to achieve about 11% and 22% of the optimal full data results. More sophisticate subsampling methods such as optimal subsampling method (Wang et al., 2018) or the information based method (Wang et al.,

2019) may further improve the estimation efficiency with the same subsample sizes, but how to design better subsampling algorithms warrants for future investigations.

Table 1. *Root mean squared error ($\times 10^3$) of estimating the mean responses*

Case	Full data		Uniform Sampling		Last Batch	
	OLS	WLS	OLS	WLS	OLS	WLS
Scenario 1	8.847	7.640	70.734	70.734	112.454	93.507
Scenario 2	15.689	15.663	70.213	70.213	251.68	239.576

OLS means ordinary least squares estimator as in (3); WLS means weighted least squares estimator as in (2).

3. WEIGHTING FROM THE PERSPECTIVE OF BIAS-VARIANCE TRADEOFF

Luo et al. (2023) considers an interesting and practical setting that β changes over time, similar to the setting of varying-coefficient models (Hastie & Tibshirani, 1993). The convergence rate of an estimator in this case is typically slower than the root N rate, and there is a bias-variance tradeoff (Fan & Zhang, 1999). A tuning parameter bandwith controls the bias-variance tradeoff in local regression, and a two-step procedure is necessary to achieve the optimal nonparametric convergence rate. Luo et al. (2023) introduce an additional weight parameter to the quadratic inference function. We discuss the weighting from the perspective of the bias-variance tradeoff.

We consider a simplified case here. Let Y_j be random variables observed at time $t_j = j/b$ with means β_j and variance σ^2 for $j = 1, \ldots, b$, and the correlation between Y_i and Y_j is $\rho^{|i-j|}$. This corresponds to the specific case of model (1) with $m = n_j = 1$ and X being a vector of ones. A weighted sample mean $\hat{\beta}_b = (1-w)(1-w^b)^{-1}\sum_{i=1}^{b} w^{b-i}Y_i$ is used to estimate the parameter in the last batch, the mean β_b, where w is the weight assigned to the previous batch. Here w corresponds to $q^{1/b}$ in Section 3 of Luo et al. (2023) and w corresponds to q in Sections 4 and 5 of Luo et al. (2023). The variance of $\hat{\beta}_b$ (detailed calculations in the supplement) is

$$V(\hat{\beta}_b) = \sigma^2 \frac{(1-w)^2}{(1-w^b)^2}\left\{\frac{1-w^{2b}}{1-w^2} + \frac{2\rho}{w-\rho}\left(\frac{w^2-w^{2b}}{1-w^2} - \frac{w\rho - w^b\rho^b}{1-w\rho}\right)\right\}. \tag{4}$$

If $w \to w_0$ as $b \to \infty$ for some $w_0 \in [0,1)$, then

$$V(\hat{\beta}_b) \to \sigma^2 \frac{1-w_0}{1+w_0}\frac{1+w_0\rho}{1-w_0\rho} > 0.$$

The variance of $\hat{\beta}_b$ does not converge to zero in this case although the full data sample size b goes to infinity. It is also seen that the limit of $V(\hat{\beta}_b)$ decreases as w_0 increases, showing that previous observations should receive higher weights to reduce the variance.

If $w \to 1$ and $w^b \to w_\infty \in [0,1)$ as $b \to \infty$, then

$$V(\hat{\beta}_b) \asymp (1-w)\frac{\sigma^2}{2}\frac{1+w_\infty}{1-w_\infty}\frac{1+\rho}{1-\rho} \to 0,$$

where $\asymp$ means two sequences are of the same order. Thus it is important the weight w assigned to the previous batch converges to one in order to utilize the full data information. Furthermore, if $w \to 1$ fast enough so that $w_\infty > 0$, then $1 - w \asymp b^{-1}$, meaning that the variance goes to

zero at the optimal parametric convergence rate. If $w \to 1$ in a slower rate so that $w_\infty = 0$, e.g., $b(1-w) \to \infty$, then the variance may converge to zero at a rate slower than b^{-1}.

Now we consider the bias, which depends on the smoothness of β as a function of time. If β is continuously differentiable or Lipschitz continuous in $[0,1]$, then $|\beta_b - \beta_j| \le \delta(b-j)/b$ for some $\delta \ge 0$. Therefore the bias of estimating β_b satisfies

$$|Bias(\hat{\beta}_b)| \le \frac{1-w}{1-w^b} \sum_{j=1}^{b} w^{b-j} \delta(b-j)/b \le \frac{w\delta}{b(1-w)}. \tag{5}$$

Combining the variance in (4) and the bias bound in (5), we see that the mean squared error $E(\hat{\beta} - \beta_b)^2$ may converge to zero at the optimal parametric rate of b^{-1} only if δ goes to zero at a rate of $O(b^{-1/2})$ and w goes to one fast enough so that $w_\infty > 0$. The root rate $b^{-1/2}$ is also widely used in the local misspecification framework for asymptotics of statistical experiments (LeCam, 1960; van der Vaart, 1998) and frequentist model averaging (e.g., Claeskens et al., 2008; Wang et al., 2009). The more interesting case is when δ is a fixed constant, and we see that the optimal rate of w that minimizes $E(\hat{\beta} - \beta_b)^2$ is $w \asymp 1 - b^{-2/3}$. The corresponding optimal convergence rate of the mean squared error is $E(\hat{\beta} - \beta_b)^2 \asymp b^{-2/3}$. An important open question is how to select the weights that achieve the optimal rate using observed data adaptively.

We also want to make a remark that the numerical results in Luo et al. (2023) might not fully demonstrate the efficiency of their proposed methods. We see from aforementioned discussions that q_j in their numerical results should be close to one for the proposed estimator to leverage historical batches of data. We expect the estimation efficiency in their simulation to be significantly improved if q_j are enlarged, while how to adjust the bias for inference is an open question.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation. Correspondence should be addressed to the second author.

REFERENCES

CLAESKENS, G., HJORT, N. L. et al. (2008). *Model selection and model averaging*, vol. 330. Cambridge University Press Cambridge.

FAN, J. & ZHANG, W. (1999). Statistical estimation in varying coefficient models. *The annals of Statistics* **27**, 1491–1518.

HASTIE, T. & TIBSHIRANI, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)* **55**, 757–779.

LECAM, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3**, 37–98.

LUO, L., WANG, J. & HECTOR, E. C. (2023). Statistical inference for streamed longitudinal data. *Biometrika* .

QU, A., LINDSAY, B. G. & LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.

VAN DER VAART, A. (1998). *Asymptotic Statistics*. Cambridge University Press, London.

WANG, H., YANG, M. & STUFKEN, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* **114**, 393–405.

WANG, H., ZHANG, X. & ZOU, G. (2009). Frequentist model averaging estimation: a review. *Journal of Systems Science and Complexity* **22**, 732.

WANG, H., ZHU, R. & MA, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* **113**, 829–844.

Supplement for Discussion of 'Statistical inference for streamed longitudinal data'

In this supplement, we provide detailed derivations of (5), (6), (7), and (8) in Section 3.

A Derivation of (5)

The variance of

$$\hat{\beta}_b = \frac{1-w}{1-w^b} \sum_{i=1}^{b} w^{b-i} Y_i$$

is

$$V(\hat{\beta}_b) = \sigma^2 \frac{(1-w)^2}{(1-w^b)^2} \sum_{i=1}^{b} \sum_{j=1}^{b} w^{2b-i-j} \rho^{|i-j|}. \tag{A.1}$$

Note that

$$\begin{aligned}
\sum_{i=1}^{b} \sum_{j=1}^{b} w^{2b-i-j} \rho^{|i-j|} &= \sum_{i=1}^{b} w^{2b-2i} + 2 \sum_{1 \le i < j \le b} w^{2b-i-j} \rho^{j-i} \\
&= \sum_{i=0}^{b-1} w^{2i} + 2 \sum_{i=1}^{b-1} \sum_{j=i+1}^{b} w^{2b-i-j} \rho^{j-i} \\
&= \frac{1-w^{2b}}{1-w^2} + 2 \sum_{i=1}^{b-1} w^{b-i} \rho^{-i} \sum_{j=i+1}^{b} w^{b-j} \rho^{j} \\
&= \frac{1-w^{2b}}{1-w^2} + 2 \sum_{i=1}^{b-1} w^{b-i} \rho^{-i} \sum_{k=1}^{b-i} w^{b-i-k} \rho^{i+k} \\
&= \frac{1-w^{2b}}{1-w^2} + 2 \sum_{i=1}^{b-1} w^{2b-2i} \sum_{k=1}^{b-i} w^{-k} \rho^{k} \\
&= \frac{1-w^{2b}}{1-w^2} + 2 \sum_{i=1}^{b-1} w^{2b-2i} \frac{\frac{\rho}{w} - \frac{\rho^{b-i+1}}{w^{b-i+1}}}{1-\frac{\rho}{w}} \\
&= \frac{1-w^{2b}}{1-w^2} + 2 \frac{\rho}{w-\rho} \sum_{i=1}^{b-1} w^{2b-2i} - \frac{2w}{w-\rho} \sum_{i=1}^{b-1} w^{b-i-1} \rho^{b-i+1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1-w^{2b}}{1-w^2} + 2\frac{\rho}{w-\rho}\sum_{i=1}^{b-1} w^{2b-2i} - \frac{2\rho}{w-\rho}\sum_{i=1}^{b-1}(w\rho)^{b-i} \\
&= \frac{1-w^{2b}}{1-w^2} + 2\frac{\rho}{w-\rho}\sum_{i=1}^{b-1} w^{2i} - \frac{2\rho}{w-\rho}\sum_{i=1}^{b-1}(w\rho)^{i} \\
&= \frac{1-w^{2b}}{1-w^2} + \frac{2\rho}{w-\rho}\frac{w^2-w^{2b}}{1-w^2} - \frac{2\rho}{w-\rho}\frac{w\rho-w^b\rho^b}{1-w\rho}.
\end{aligned}
$$

Thus, we obtain (5).

B Derivation of (6)

If $w \to w_0$ as $b \to \infty$ for some $w_0 \in [0,1)$, we have that $w^b \to 0$ and thus

$$
\begin{aligned}
V(\hat{\beta}_b) \to& \sigma^2(1-w_0)^2\left\{\frac{1}{1-w_0^2} + \frac{2\rho}{w_0-\rho}\left(\frac{w_0^2}{1-w_0^2} - \frac{w_0\rho}{1-w_0\rho}\right)\right\} \\
&= \sigma^2(1-w_0)^2\frac{1+w_0\rho}{(1-w_0^2)(1-w_0\rho)} \\
&= \sigma^2\frac{1-w_0}{1+w_0}\frac{1+w_0\rho}{1-w_0\rho},
\end{aligned}
$$

which gives (6).

C Derivation of (7)

Note that

$$
V(\hat{\beta}_b) = \sigma^2\frac{(1-w)}{(1-w^b)^2}\left\{\frac{1-w^{2b}}{1+w} + \frac{2\rho}{w-\rho}\left(\frac{w^2-w^{2b}}{1+w} - \frac{(1-w)(w\rho-w^b\rho^b)}{1-w\rho}\right)\right\}.
$$

If $w \to 1$ and $w^b \to w_\infty \in [0,1)$ as $b \to \infty$, we have that

$$
\frac{V(\hat{\beta}_b)}{(1-w)} \to \frac{\sigma^2}{2(1-w_\infty)^2}\left\{\left(1-w_\infty^2\right)\left(1+\frac{2\rho}{1-\rho}\right)\right\} = \frac{\sigma^2}{2}\frac{1+w_\infty}{1-w_\infty}\frac{1+\rho}{1-\rho}.
$$

Therefore, (7) holds.

D Derivation of (8)

From Lipschitz continuity, we have

$$
|Bias(\hat{\beta}_b)| = \frac{1-w}{1-w^b}\left|\sum_{j=1}^{b} w^{b-j}\left(\beta_j - \beta_b\right)\right|
$$

$$
\begin{aligned}
&\leq \frac{1-w}{1-w^b} \sum_{j=1}^{b} w^{b-j} \left|\beta_j - \beta_b\right| \\
&\leq \frac{1-w}{1-w^b} \sum_{j=1}^{b} w^{b-j} \frac{\delta(b-j)}{b} \\
&= \frac{\delta}{b} \frac{1-w}{1-w^b} \sum_{j=1}^{b} w^{b-j}(b-j) \\
&= \frac{\delta}{b} \frac{1-w}{1-w^b} \sum_{j=0}^{b-1} j w^j.
\end{aligned}
$$

Letting $S = \sum_{j=0}^{b-1} j w^j$, we have that $wS = \sum_{j=1}^{b} j w^{j+1}$. Therefore,

$$
(1-w)S = \sum_{j=1}^{b-1} j w^j - \sum_{j=1}^{b} (j-1) w^j = \sum_{j=1}^{b-1} w^j - (b-1)w^b = \frac{w - w^b}{1-w} - (b-1)w^b.
$$

Thus,

$$
\left|Bias(\hat{\beta}_b)\right| \leq \frac{\delta}{b} \frac{1-w}{1-w^b} \left\{ \frac{w - w^b}{(1-w)^2} - \frac{(b-1)w^b}{1-w} \right\} \leq \frac{w\delta}{b(1-w)} \frac{1-w^{b-1}}{1-w^b} \leq \frac{w\delta}{b(1-w)},
$$

where the last inequality is because $1 - w^{b-1} \leq 1 - w^b$ when $0 < w < 1$. Therefore, we obtain (8).