ADAPTIVE CENSORING FOR LARGE-SCALE REGRESSIONS

Dimitris K. Berberidis[1]*, Vassilis Kekatos*[1]*, Gang Wang*[2]*, and Georgios B. Giannakis*[1]

[1] Dept. of ECE and Digital Tech. Center, University of Minnesota
Minneapolis, MN 55455, USA
[2] School of Automation, Beijing Institute of Technology
Beijing 100081, China
Emails: {bermp001,kekatos,wang4937,georgios}@umn.edu

ABSTRACT

Albeit being in the big data era, a significant percentage of data accrued can be overlooked while maintaining reasonable quality of statistical inference at affordable complexity. By capitalizing on data redundancy, interval censoring is leveraged here to cope with the scarcity of resources needed for data exchanging, storing, and processing. By appropriately modifying least-squares regression, first- and second-order algorithms with complementary strengths that operate on censored data are developed for large-scale regressions. Theoretical analysis and simulated tests corroborate their efficacy relative to contemporary competing alternatives.

1. INTRODUCTION

Nowadays ubiquitous monitoring sensors, e-commerce sites, and Internet-friendly portable devices generate massive volumes of dynamic data. The task of extracting the most informative, yet low-dimensional structure from large datasets is thus of paramount importance. Redundancy is an attribute of massive datasets encountered in various application domains [3]. Judiciously exploiting this redundancy offers an effective means of reducing data processing costs. In this context, the fresh idea here is relying on data censoring to quantify how informative each datum is. *Censoring* appears naturally in several areas in econometrics, biometrics, sociometrics and engineering [1], including survival analysis [5], saturated metering [11], and collaborative spectrum sensing [6]. It has recently been employed to select sensors for distributed estimation of parameters and dynamical processes using resource-constrained wireless sensor networks, thus trading off performance for tractability [9, 13]. Iterative censoring combined with maximum likelihood estimation was introduced in [15]. These works corroborate that estimation accuracy attainable with censored measurements can be comparable to that of estimators based on uncensored data. Therefore, censoring has great potential to lower data processing costs, which is certainly beneficial in big data applications.

To this end, the present work employs *adaptive censoring* for large-scale regressions. The key innovative idea is to sequentially test and update least mean-square (LMS) or recursive least-squares (RLS) estimates only for informative data. Specifically, future target variables are predicted using the current estimate of the regression coefficient vector. Predictions are then used to form the innovations, based on which the importance of future data is assessed, and a decision is made whether to update or not the coefficient vector using first- or second-order iterations. Analysis and simulated tests corroborate that the proposed schemes maintain high estimation accuracy while reducing data costs, and outperform alternatives based on random projections [7], and the randomized Kaczmarz's algorithm [16].

Notation. Lower- (upper-) case boldface letters denote column vectors (matrices); calligraphic letters stand for sets. Symbols $\mathbf{a}^T$ and $\|\mathbf{a}\|_2$ denote transposition and the ℓ_2-norm. The symbols $\lambda_{\min}$, $\lambda_{\max}$, and trace denote the minimum, the maximum, and the sum of the eigenvalues of a matrix, while $Q(x) := \int_x^\infty \frac{\exp -x^2/2}{\sqrt{2\pi}} \partial x$.

2. PROBLEM FORMULATION

Suppose an unknown vector $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is to be estimated from streaming data. In a simplifying yet sufficiently representative setup, consider the linear model $y_n = \mathbf{x}_n^T \boldsymbol{\theta}_0 + v_n$, where $\{y_n, \mathbf{x}_n\}_{n=1}^N$ are given pairs of scalar target variables and feature vectors, while $\{v_n\}_{n=1}^N$ capture noise. Collecting data and noise into $\mathbf{y}, \mathbf{v} \in \mathbb{R}^N$, and $\mathbf{X} \in \mathbb{R}^{N \times p}$, the following matrix-vector model is obtained

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \mathbf{v}. \tag{1}$$

The dataset $(\mathbf{y}, \mathbf{X})$ can be prohibitively large ($N \gg p$) for the available computing platform to estimate the wanted $\boldsymbol{\theta}_0$

Work in this paper was supported by AFOSR MURI Grant FA9550-10-1-0567, and NSF Grants ECCS-1202135 and 1343248. G. Wang was supported in part by China Scholarship Council.

via ordinary least-squares (LS). How big is N can be relative: For a mainframe, N can be in the order of billions, but for a sensor under stringent power resources N can be deemed big if in the order of thousands.

Aiming at dimensionality reduction, censoring will be advocated as a means of discarding data. A general censoring rule takes the form

$$z_n := \begin{cases} * & , \ y_n \in \mathcal{C}_n \\ y_n & , \ \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{C}_n$ denotes the censoring interval or set. When datum y_n is censored, the estimator simply knows that $y_n \in \mathcal{C}_n$; otherwise, the actual y_n is observed. Let c_n denote a binary-valued censoring variable that equals 1 when the n-th datum is censored; and 0, otherwise. We wish to design $\mathcal{C}_n$ and decide the value of c_n based on the innovation $|y_n - \hat{y}_n|$, where $\hat{y}_n$ is obtained in two possible formats, as detailed next. The first uses a *non-adaptive censoring* (NAC) rule that relies on a preliminary estimate of $\boldsymbol{\theta}_0$; e.g., the LS estimate $\hat{\boldsymbol{\theta}}_K$ obtained from a minimal number of K data ($p \leq K \ll N$). With $\hat{y}_n = \mathbf{x}_n^T \hat{\boldsymbol{\theta}}_K$, the FC rule is given by

$$(z_n, c_n) := \begin{cases} (y_n, 0) & , \ \left| y_n - \mathbf{x}_n^T \hat{\boldsymbol{\theta}}_K \right| \geq \beta \\ (*, 1) & , \ \text{otherwise} \end{cases} \tag{3}$$

where $*$ signifies that the exact value of y_n is unavailable, and the threshold β controls the percentage of censored data. We used the FC rule to censor data with innovations smaller than β, and obtained a censored LS estimator even when the wanted coefficient vector is sparse [14].

Rather than using a fixed estimate to acquire target predictions, we adopt here the *adaptive censoring* (AC) rule

$$(z_n, c_n) := \begin{cases} (y_n, 0) & , \ \left| y_n - \mathbf{x}_n^T \boldsymbol{\theta}_{n-1} \right| \geq \beta \\ (*, 1) & , \ \text{otherwise} \end{cases} \tag{4}$$

where $\hat{y}_n$ relies on the most recent estimate $\boldsymbol{\theta}_{n-1}$. The AC rule of (4) is combined next with stochastic approximation algorithms to jointly effect online estimation and censoring.

3. ADAPTIVELY CENSORED LEAST-SQUARES

To appropriately exploit the censored data in (4), let us consider finding the unknown $\boldsymbol{\theta}_0$ as the minimizer of

$$\min_{\boldsymbol{\theta}} \ \sum_{n=1}^{N} f_\beta(y_n - \mathbf{x}_n^T \boldsymbol{\theta}) \tag{5}$$

where $f_\beta(e)$ is defined for some $\beta \geq 0$ as

$$f_\beta(e) := \begin{cases} \frac{1}{2}(e^2 - \beta^2) & , \ |e| > \beta \\ 0 & , \ |e| \leq \beta \end{cases} . \tag{6}$$

The optimization problem in (5) couples estimation and data reduction. At the optimum, the data points achieving an absolute residual $|y_n - \mathbf{x}_n^T \boldsymbol{\theta}|$ smaller than the threshold β do not contribute to the optimal cost. On the contrary, relatively large residuals are summed up in the optimal cost. Yet finding which data pairs $(y_n, \mathbf{x}_n)$ are informative in the sense of providing large residuals requires considering the entire dataset and tackling (5) in a batch manner.

First-order stochastic algorithm. Since solving (5) becomes challenging for large N, a practical solution is resorting to stochastic approximation and handling a single summand $f_\beta(y_n - \mathbf{x}_n^T \boldsymbol{\theta})$ at a time. For a step size μ_n, the stochastic subgradient descent iterations for (5) can be easily shown to be

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \mu_n (1 - c_n) \mathbf{x}_n (y_n - \mathbf{x}_n^T \boldsymbol{\theta}_{n-1}) \tag{7}$$

where c_n is defined as in (4) for $\tau = \beta$. Note that for $\beta = 0$, problem (5) constitutes the LS estimator of $\boldsymbol{\theta}$. In that case, all censoring variables are set to $c_n = 0$, and the iterations in (7) simplify to the ordinary Least Mean Square (LMS) algorithm [10]. The convergence of the iterations in (7), henceforth referred to as the adaptively-censored LMS (AC-LMS) algorithm, is characterized in the following proposition whose proof is avoided due to space limitations.

Proposition 1. *Assume regressors $\mathbf{x}_n$ are generated i.i.d. with $\mathbb{E}[\mathbf{x}_n] = \mathbf{0}$, $\mathbb{E}\left[\mathbf{x}_n \mathbf{x}_n^T\right] = \mathbf{R}_x$, $\mathbb{E}\left[\mathbf{x}_n^T \mathbf{x}_n \mathbf{x}_n\right] = \mathbf{0}$, and $\mathbb{E}\left[\left(\mathbf{x}_n \mathbf{x}_n^T\right)^2\right] = \mathbf{R}_x^2$, while observations y_n are obtained according to (1) for $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$. For $\mu_n = \mu/(\alpha n)$, an initial estimate $\boldsymbol{\theta}_1$, and $\beta = \tau\sigma$, the mean square error (MSE) of the AC-LMS iterates is bounded as*

$$\mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq \frac{e^{4L^2/\alpha^2}}{n^2} \left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2^2 + \frac{\Delta}{L^2} \right) + 8 \frac{\Delta}{\alpha^2} \frac{\log n}{n}$$

where $\alpha = 2Q(\tau)\lambda_{\min}(\mathbf{R}_x)$, $L^2 = \lambda_{\max}\left(\mathbf{R}_x^2\right)$, and $\Delta = 2\mathrm{tr}(\mathbf{R}_x)\sigma^2 \left(1 - Q(\tau) + \tau p(\tau)\right)$. Furthermore, for $\mu_n = \mu < \alpha/16L^2$ the AC-LMS enjoys exponential convergence to a bounded error set

$$\mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq 2 \exp\left(- \left(\frac{\alpha\mu}{4} - 4L^2\mu^2 \right) n - 4L^2\mu^2 \right) \left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2^2 + \frac{\Delta}{L^2} \right) + \frac{4\mu\Delta}{\alpha}.$$

Rule (4) together with Proposition 1 reveal that as $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$ in the mean-square sense, it holds that $y_n - \mathbf{x}_n^T \boldsymbol{\theta}_{n-1} \approx v_n$, and hence, $\Pr(c_n = 0) = 2Q(\tau)$. For large enough datasets, choosing $\tau = Q^{-1}\left(\frac{d}{2N}\right)$ will approximately result to d out of total N data being used. To further control the average data rate, generalizations of (7) with time-varying

threshold β have been developed, but are not presented here due to space limitations.

Second-order stochastic algorithm. To accelerate the joint task of sequential estimation and data reduction, a second-order sequential algorithm closely resembling recursive least squares (RLS) is devised next. Conventional RLS can be viewed as a second-order stochastic gradient descent method for minimizing $\sum_{n=1}^{N}(y_n - \mathbf{x}_n^T\boldsymbol{\theta})^2$ taking the form

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \mathbf{M}_n^{-1}\nabla f_n(\boldsymbol{\theta}_{n-1})$$

with an ideal matrix step-size $\mathbf{M}_n = \mathbb{E}[\nabla^2(y_n - \mathbf{x}_n^T\boldsymbol{\theta})^2] = \mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^T\right]$ that is practically replaced by its sample average.

Applying a second-order stochastic gradient descent for the problem in (5) yields the adaptively censored RLS (AC-RLS) algorithm whose iterations are as follows:

$$\mathbf{g}_n = \frac{1-c_n}{1+\mathbf{x}_n^T\mathbf{P}_{n-1}\mathbf{x}_n}\mathbf{P}_{n-1}\mathbf{x}_n \tag{8a}$$

$$\mathbf{P}_n = \mathbf{P}_{n-1} - \mathbf{g}_n\mathbf{x}_n^T\mathbf{P}_{n-1} \tag{8b}$$

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \mathbf{g}_n(y_n - \mathbf{x}_n^T\boldsymbol{\theta}_{n-1}) \tag{8c}$$

where the censoring decision c_n is decided according to (4). Observe that once a datum has been censored ($c_n = 1$), the AC-RLS iterates yield $\mathbf{g}_n = \mathbf{0}$, and hence, the matrix step size $\mathbf{P}_n$, and the estimate $\boldsymbol{\theta}_n$ remain at their previous values. Skipping (8c) and particularly (8b) provides a low-complexity algorithm, whose MSE performance remains bounded as shown next.

Proposition 2. *Assume regressors $\mathbf{x}_n$ are generated i.i.d. with $\mathbb{E}[\mathbf{x}_n] = \mathbf{0}$ and $\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^T\right] = \mathbf{R}_x$, while observations y_n are obtained according to model* (1) *for $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. For an initial estimate $\boldsymbol{\theta}_1 = \mathbf{0}$, and $\beta = \tau\sigma$, there exists a $k > 0$ such that for all $n \geq k$ the iterations in* (8) *yield estimates $\boldsymbol{\theta}_n$ with MSE bounded as*

$$\frac{1}{n}\mathrm{tr}\left(\mathbf{R}_x^{-1}\right)\sigma^2 \leq \mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq \frac{1}{n}\frac{\mathrm{tr}\left(\mathbf{R}_x^{-1}\right)\sigma^2}{2Q(\tau)}.$$

Apart from offering an effective method for handling streaming data, AC-RLS is capable of reducing the dimension of large-scale LS problems. In recent works, the properties of random projections (RP) in relation to leverage scores have been advocated for reducing the size of large-scale LS regression problems [7], [8]. To outline RP-based methods in the present context, recall that given a full-rank matrix $\mathbf{X}$ and target vector $\mathbf{y}$, the LS prediction of target variables is simply $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$, where $\mathbf{P} := \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the so termed *hat matrix*. The leverage score per regressor $\mathbf{x}_n$ is the n-th diagonal entry of $\mathbf{P}$. If data are drawn from the model in (1), the prediction error is $\hat{\mathbf{y}} - \mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{v}$. It follows that $\mathrm{var}(\hat{y}_n) = \sigma^2(1 - \mathbf{P}_{nn})$, showing that the leverage score captures the fitting importance of datum n. As shown in [8], one way to reduce the dimensionality of the original LS problem without significantly affecting estimation accuracy, is by finding the LSE for a data subset with relatively high leverage scores. Unfortunately, computing the main diagonal of $\mathbf{P}$ comes at a high computational cost $\mathcal{O}(p^2N)$. In its basic form, the RP-based approach linearly transforms data $(\mathbf{y}, \mathbf{X})$ so that leverage scores come close to being uniform. This is achieved by premultiplying data with $\mathbf{HD}$, where $\mathbf{H}$ denotes an $N \times N$ Hadamard matrix, and $\mathbf{D}$ a diagonal matrix whose diagonal entries take the values $\{-1/\sqrt{N}, +1/\sqrt{N}\}$ equiprobably. Since all transformed data are approximately of "comparable importance," it turns out that with complexity $o(p^2N)$ one can solve the reduced-size LS problem [7, 8]

$$\tilde{\boldsymbol{\theta}}_d := \arg\min_{\boldsymbol{\theta}} \|\mathbf{S}_d\mathbf{HD}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_2^2 \tag{9}$$

where $\mathbf{S}_d$ is a $d \times N$ matrix representing a uniform random selection of $d < N$ measurements.

Upon setting $\tau = Q^{-1}\left(\frac{1}{2}\frac{d}{N}\right)$, AC-RLS attains approximately the same data usage ratio of d/N by scanning the observations and selecting the most informative ones. As numerically corroborated in Section 4, AC-RLS achieves significantly lower estimation error. Intuitively, this can be attributed to the fact that, unlike RPs that are based only on $\mathbf{X}$, adaptive censoring extracts the most informative subset of data from $(\mathbf{y}, \mathbf{X})$. Regarding robustness, quadratic error losses are known to be sensitive to outliers, while selecting observations with high instantaneous error further aggravates the situation. To mitigate this problem, robust versions of the AC-LMS and AC-RLS have been developed and will be presented in coming work.

4. NUMERICAL TESTS

Our novel AC-RLS algorithm was tested on synthetic data as a data-reducing scheme. Its relative LS error (LSE), i.e., $\|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n\|_2/\|\boldsymbol{\theta}_0\|_2$, was compared with the Hadamard-preconditioned RP-based LS solver of [8] for different values of the data reduction ratio (d/N). Apart from these two methods, the relative LSE obtained by uniformly sampling d out of N data pairs $(y_n, \mathbf{x}_n)$ was used as a naive benchmark. Data were generated according to (1) for $p = 300$, $N = 10,000$, and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I})$.

In the first experiment, regressors $\{\mathbf{x}_n\}$ were drawn from a heavy-tail multivariate Student t-distribution with one degree of freedom and covariance matrix with entries $\boldsymbol{\Sigma}_{i,j} = 2 \times 0.5^{|i-j|}$. Such a distribution yields matrices $\mathbf{X}$ with highly non-uniform leverage scores, and thus, uniform sampling without preconditioning performs poorly. As verified by Fig. 1, preconditioning considerably improves performance by incorporating important data via RPs. Further improvement is achieved by AC-RLS that adaptively selects informative data at no extra computational cost.

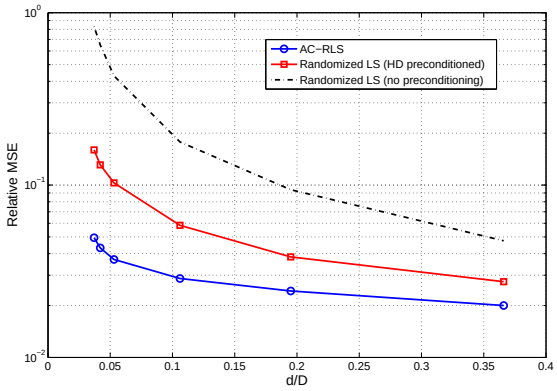

Figure 1: Relative LSE versus data reduction ratios using highly non-uniform leverage scores.

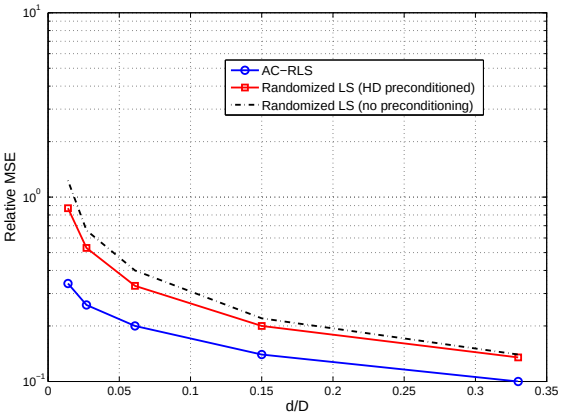

Figure 2: Relative MSE of AC-RLS and randomized LS algorithms, for different levels of data reduction. Moderately non-uniform leverage scores

In Fig. 2, the experiment is repeated for $\mathbf{x}_n$ generated from a multivariate $t-$distribution with 3 degree of freedom and covariance matrix $\mathbf{\Sigma}$ as before. Leverage scores for this dataset are moderately non-uniform, thus inducing more redundancy and resulting in lower performance for all algorithms, while closing the gap between preconditioned and non-preconditioned random sampling. Again, AC-RLS performs significantly better in estimating the unknown parameters for the entire range of data size reduction.

Figure 3 shows the results obtained from Gaussian data $\mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, in which case matrix $\mathbf{X}$ exhibits almost uniform leverage scores. As seen in plots, preconditioning offers no improvement in random sampling for this type of data, whereas the AC-RLS succeeds in extracting more information on the unknown $\boldsymbol{\theta}$.

The AC-LMS algorithm was tested on synthetic data as an alternative to the Randomized Kaczmarz's algorithm [16].

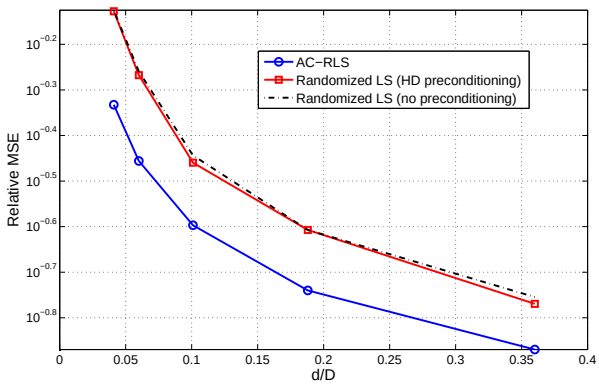

Figure 3: Relative MSE of AC-RLS and randomized LS algorithms, for different levels of data reduction. Uniform leverage scores

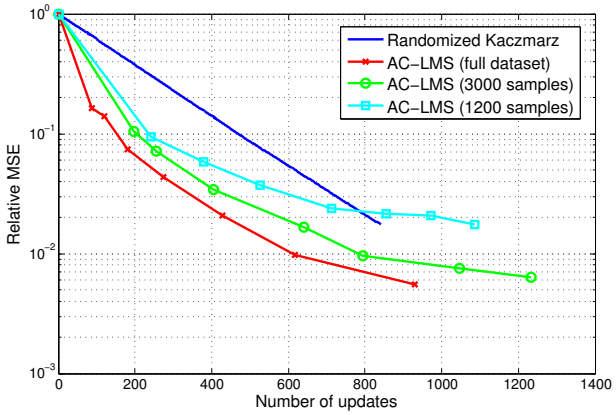

Figure 4: Relative MSE AC-LMS and Randomized Kaczmarz's algorithm.

For this experiment, a total of $D = 30,000$ observations y_n were generated according to (1) for $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, 0.25\mathbf{I}_N)$ and 100-dimensional regressors $\mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. For the Randomized Kaczmarz's algorithm, the probability of selecting the $i-$th row is equal to $p_n = \|\mathbf{x}_n\|_2^2/\|\mathbf{X}\|_F^2$ as proposed in [16]. Since the complexity of the two methods is roughly the same, they are compared in terms of their estimation error. Plotted in Fig. 4, are the Relative MSE (RMSE) curves of the two algorithms with respect to N averaged over 50 Monte Carlo runs. While AC-LMS scans the entire dataset updating only informative data, the Randomized Kaczmarz's algorithm needs access only to the data used for its updates. That holds under the assumption that the data-dependent selection probabilities p_n are given a-priori, which may not always be the case. Regardless, two more experiments were run, in which the AC-LMS had limited access to 3000 and 1400 data. Overall, it can be argued that when the sought reduced dimension is small, the AC-LMS becomes a simple and reliable alternative to Randomized Kaczmarz's algorithm.

5. REFERENCES

[1] T. Amemiya, "Tobit models: A survey," *J. Econom.* vol. 24, no. 1, pp. 3–61, 1984.

[2] D. Cox, "Regression Models and Life Tables (with Discussion)," *Journal of the Royal Statistical Society, Series B*, vol. 34, pp. 187–220, 1972.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] H. Duchi, S. Shalev-Shwartz, Y. SInger, and A. Tewari, "Composite objective mirror descent," in *23rd Annu. Conf. Comput. Learning Theory,* Haifa, Israel, June 27–29, 2010.

[5] L. Evers and C. Messow, "Sparse kernel methods for high-dimensional survival data," *Bioinformatics*, vol. 24, no. 14, pp. 1632–1638, 2008.

[6] J. Lunden, V. Koivunen, A. Huttunen, and H. V. Poor, "Censoring for collaborative spectrum sensing in cognitive radios," in *Proc. 41st Asilomar Conf. Signals, Syst., Comput.,* Pacific Grove, CA, Nov. 4–7, 2007.

[7] M. Mahoney, "Randomized Algorithms for Matrices and Data," *Found. Trends. in Mach. Learn.,* vol. 3, no. 2, pp. 123-224, 2011.

[8] M. Mahoney, "Algorithmic and Statistical Perspectives on Large-Scale Data Analysis," *http://arxiv.org/pdf/1010.1609v1.pdf*, 2011.

[9] E. Msechu and G. Giannakis, "Sensor–centric data reduction for estimation with WSNs via censoring and quantization," *IEEE Trans. Signal Process.,* vol. 60, no. 1, pp. 400–414, Jan. 2012.

[10] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*, Prentice-Hall, Inc., NJ, USA, 1995.

[11] J. Tobin, "Estimation of relationships for limited dependent variables," *Econometrica: J. Econometric Soc.*, vol. 26, no. 1, pp. 24–36, 1958.

[12] V. Vapnik, *Statistical Learning Theory,* New York: Wiley, 1998.

[13] G. Wang, J. Chen, J. Sun, and Y. Cai, "Power scheduling of Kalman filtering in wireless sensor networks with data packet drops," *arXiv preprint arXiv:1312.3269v2*, 2013.

[14] G. Wang, D. Berberidis, V. Kekatos, and G. B. Giannakis, "Online Reconstruction from Big Data via Compressive Censoring," in *IEEE Global Conf. on Signal and Inf. Process.*, Atlanta. GA, Dec. 2014.

[15] K. You, L. Xie, and S. Song, "Asymptotically optimal parameter estimation with scheduled measurements," *IEEE Trans. Signal Process.,* vol. 61, no. 14, pp. 3521–3531, Jul. 2013.

[16] T. Strohmer and R.Vershynin, "A randomized Kaczmarz algorithm with exponential convergence," *Journal of Fourier Analysis and Applications,* vol. 15, no. 2, pp. 262–278, 2009