SNIPE for Memory-Limited PCA From Incomplete Data: From Failure to Success

Armin Eftekhari, Laura Balzano, Michael B. Wakin, and Dehui Yang

Principal component analysis (PCA) has traditionally been an indispensable tool in data analysis, using linear models as a means for dimensionality reduction [1]. In this work we are particularly interested in applying PCA to data that suffers from erasures, while limited storage is available. In recommender systems, for instance, data may be highly incomplete and yet so massive that it can only be processed in small chunks [2].

To sharpen our focus in this abstract, consider partially observed data vectors drawn from an unknown subspace, presented sequentially to the user who, due to hardware limitations, can only store small amounts of data. We are interested in developing a *streaming* algorithm for PCA from these incomplete measurements.

More concretely, consider an r-dimensional subspace $\mathcal{S}$ with orthonormal basis $S \in \mathbb{R}^{n\times r}$. For an integer T, let the coefficient vectors $\{q_t\}_{t=1}^T \subset \mathbb{R}^r$ be independent copies of a random vector $q \in \mathbb{R}^r$. At time $t \in [1:T] := \{1, 2, \cdots, T\}$, we observe each entry of $s_t := S \cdot q_t \in \mathcal{S}$ independently with a probability of p, and we collect the measurements in $y_t \in \mathbb{R}^n$, supported on a random index set $\omega_t \subseteq [1:n]$. Formally, we will write this measurement process as $y_t = P_{\omega_t}(s_t) = P_{\omega_t} \cdot s_t$, where $P_{\omega_t} \in \mathbb{R}^{n\times n}$ is the projection onto the coordinate set ω_t.

Our objective is to identify the subspace $\mathcal{S}$ from the measurements $\{y_t\}_{t=1}^T$ supported on the index sets $\{\omega_t\}_{t=1}^T$. Assuming that $r = \dim(\mathcal{S})$ is known *a priori* (or estimated from data by other means), we will shortly present the SNIPE algorithm for this task; SNIPE converges, globally and linearly, to the true subspace under reasonable requirements. As a numerical example, suppose that $\mathcal{S} \subset \mathbb{R}^{1000}$ is a generic subspace of dimension $r = 3$ and take $p = 0.1$. Then, SNIPE produces a sequence of estimates of $\mathcal{S}$; the estimation error (with the metric in [3]) versus t is plotted in Figure 1.

I. The Bias of Zero-filling

To better motivate SNIPE, however, it might help to first describe a simple and natural alternative. For an integer K, fix block sizes $\{b_k\}_{k=1}^K$ such that $\sum_{k=1}^K b_k = T$ and $b_k \geq r$ for all k. Suppose we partition the data into K (non-overlapping) blocks $\{Y_k\}_{k=1}^K$, where $Y_k \in \mathbb{R}^{n\times b_k}$ for each k. Note that each measurement block Y_k contains zeros at all unobserved data positions. Note also that, despite these zeros, each measurement block Y_k can easily be used to form a simple, if not accurate, estimate of the underlying subspace $\mathcal{S}$. In particular, let $Y_{k,r} \in \mathbb{R}^{n\times b_k}$ be a rank-r truncation of Y_k, obtained by truncating all but the largest r singular values of Y_k. Consider the r-dimensional subspace $\mathrm{Y}_k = \mathrm{span}(Y_{k,r})$. We may consider Y_k as an estimate of $\mathcal{S}$; we refer to this as the "zero filled" estimate due to the zeros in the measurement block Y_k. (Note that $\mathrm{Y}_k = \mathcal{S}$ when $p = 1$ and thus there are no zeros inserted into Y_k.)

When there are erasures ($p < 1$), $\{\mathrm{Y}_k\}_{k=1}^K$ are independent and identically distributed random subspaces on the Grassmannian $\mathbb{G}(n, r)$, the manifold of all r-dimensional subspaces of $\mathbb{R}^n$. The randomness of these subspaces derives from the randomness of the coefficient vectors q_t used to generate the data and the randomness of the observation index sets ω_t.

It is therefore natural to consider the "average" of the subspaces $\{\mathrm{Y}_k\}_{k=1}^K$ as an estimate of $\mathcal{S}$. (As detailed in [4], some care must be taken in defining this average.) This raises the question: *Is* Y_k *an unbiased estimator of the true subspace* $\mathcal{S}$*?* In general, the answer to this question is no. As detailed in [4], however, roughly speaking the estimation bias can be bounded by a factor of $\sqrt{\max(1, \frac{n}{\min b_k})\frac{r}{pn}}$, where this factor also depends on a certain notion of *coherence* of the subspace $\mathcal{S}$.

II. SNIPE

SNIPE improves over the above averaging scheme, by replacing "zero-filling" in the observed data vectors with a better alternative. At a high level, SNIPE processes the first measurement block Y_1 to produce an estimate $\widehat{\mathcal{S}}_1$ of the true subspace $\mathcal{S}$. This estimate is then iteratively updated after receiving each of the new blocks $\{Y_k\}_{k=2}^K$, thereby producing a sequence of estimates $\{\widehat{\mathcal{S}}_k\}_{k=2}^K$. Every $\widehat{\mathcal{S}}_k$ is an r-dimensional subspace of $\mathbb{R}^n$ with orthonormal basis $\widehat{S}_k \in \mathbb{R}^{n\times r}$.

More concretely, SNIPE sets $\widehat{\mathcal{S}}_1$ to be the span of the top r left singular vectors of Y_1, which is zero-filled. Then, at iteration $k \in [2:K]$ and given the previous estimate $\widehat{\mathcal{S}}_{k-1} = \mathrm{span}(\widehat{S}_{k-1})$, SNIPE processes the columns of the kth measurement block Y_k by forming the matrix

$$R_k = \begin{bmatrix} \cdots & y_t + P_{\omega_t^C}\widehat{S}_{k-1}\left(P_{\omega_t}\widehat{S}_{k-1}\right)^\dagger y_t & \cdots \end{bmatrix} \in \mathbb{R}^{n\times b_k},$$

with t above ranging from $\sum_{k'=1}^{k-1} b_{k'} + 1$ to $\sum_{k'=1}^{k} b_{k'}$, and where $P_{\omega_t^C} = I_n - P_{\omega_t} \in \mathbb{R}^{n\times n}$ projects a vector onto the complement of the index set ω_t. In words, this process replaces the zeros in Y_k with least-squares estimates of the missing data, informed by the previous subspace estimate $\widehat{\mathcal{S}}_{k-1}$ and the available (non-zero) observations in Y_k. SNIPE then updates its estimate by setting $\widehat{\mathcal{S}}_k$ to be the span of the top r left singular vectors of R_k.

As detailed in [3], SNIPE converges to the true subspace, exponentially fast and with high probability, provided that $p \gtrsim r^2 \log^5(n)/n$, $b_1 \gtrsim n$, and $b_k \gtrsim r$, $k \geq 2$, where we have suppressed universal constants and the dependence on the *coherence* of $\mathcal{S}$.

Among several algorithms that have been proposed for tracking low-dimensional structure in a data set from partially observed streaming measurements [4]–[8], SNIPE might be most closely related to GROUSE [9], [10]. GROUSE performs memory-limited PCA from incomplete data using stochastic gradient projection on the Grassmannian, updating its estimate of the true subspace with each new measurement vector, rather than blockwise. Figure 2 shows an empirical comparison of SNIPE with GROUSE [9] and the modified power method in [5]. We set $n = 100$, $r = 5$, $T = 5 \cdot 10^3$, and take $\mathcal{S} \subset \mathbb{R}^n$ to be a generic r-dimensional subspace. For $p = 3r/n = 0.15$, Figure 2 shows the average over 100 trials of the estimation error of algorithms, with respect to the metric $d_{\mathbb{G}}$ described in [3] and with the specifics described therein.

AE is with the Alan Turing Institute, LB is with the University of Michigan at Ann Arbor, and MBW and DY are with the Colorado School of Mines. This work was partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, and also by ARO Grant W911NF-14-1-0634, NSF grant CCF-1409258, and NSF CAREER grant CCF-1149225.

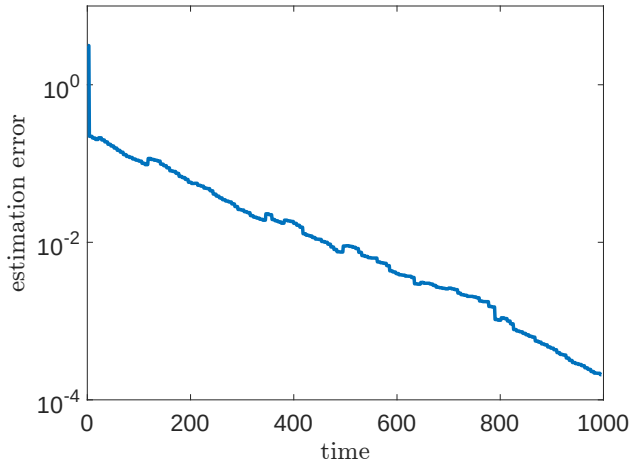

Figure 1: This paper introduces SNIPE for memory-limited PCA from incomplete data. This figure shows the estimation error of SNIPE versus time in recovering a generic 3-dimensional subspace from data subsampled by a factor of 10, received sequentially.

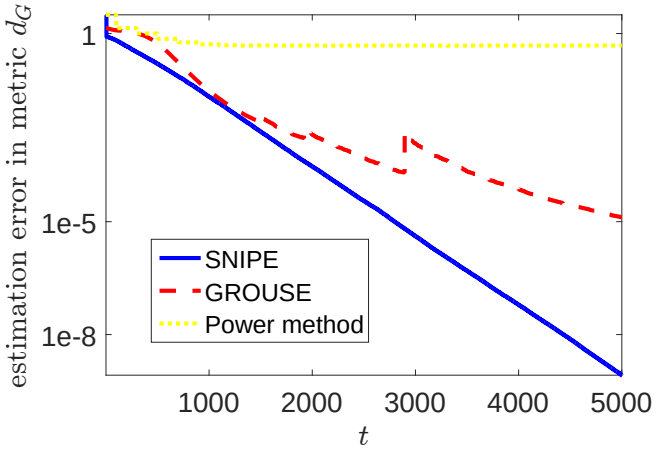

Figure 2: Estimation error versus time for SNIPE, GROUSE, and the modified power method in [5] with a prescribed set of parameters.

References

[1] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16:2859–2900, 2015.

[2] B. Recht, C. Re, S. J. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[3] A. Eftekhari, L. Balzano, D. Yang, and M. B. Wakin. SNIPE for memory-limited PCA from incomplete data. *arXiv preprint arXiv:1612.00904*, 2016.

[4] A. Eftekhari, L. Balzano, and M. B Wakin. What to expect when you are expecting on the Grassmannian. *IEEE Signal Processing Letters*, 2017.

[5] I. Mitliagkas, C. Caramanis, and P. Jain. Streaming PCA with many missing entries. *Preprint*, 2014.

[6] Y. Chi, Y. C. Eldar, and R. Calderbank. PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations. *IEEE Transactions on Signal Processing*, 61(23):5947–5959, 2013.

[7] M. Mardani, G. Mateos, and G. B. Giannakis. Subspace learning and imputation for streaming big data matrices and tensors. *IEEE Transactions on Signal Processing*, 63(10):2663–2677, 2015.

[8] Y. Xie, J. Huang, and R. Willett. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2013.

[9] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 704–711. IEEE, 2010.

[10] L. Balzano and S. J Wright. On GROUSE and incremental SVD. In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–4. IEEE, 2013.