ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Dimensionality reduction in decentralized networks by Gossip aggregation of principal components analyzers

Jerome Fellus[1], David Picard[1] and Philippe-Henri Gosselin[1,2] *

1- ETIS - UMR CNRS 8051 - ENSEA - Universite de Cergy-Pontoise

2- Inria, Texmex project, Campus de Beaulieu, Rennes, France

Abstract. This paper considers dimensionality reduction in large decentralized networks with limited node-local computing and memory resources and unreliable point-to-point connectivity (*e.g* peer-to-peer, sensors or ad-hoc mobile networks). We propose an asynchronous decentralized algorithm built on a Gossip consensus protocol that perform Principal Components Analysis (PCA) of data spread over such networks. All nodes obtain the same local basis that span the global principal subspace. Reported experiments show that obtained bases both reach a consensus and accurately estimate the global PCA solution.

1 Introduction

Real-world data is usually captured in a high-dimensional space. Observations being generally noisy and highly correlated, they often embed a much lower-dimensional hidden structure. Dimensionality reduction aims at transforming high-dimensional inputs into low-dimensional representations that best retain this structure, easing tasks such as classification, clustering or visualization. In this paper, we focus on Principal Components Analysis (PCA [1]) which is a well-known linear approach to dimensionality reduction.

Like most statistical learning tools, PCA was formulated for centralized setups where all data is available at a single location. This assumes that the solution can be computed by a single machine and that all intermediary results fit in main memory. However, this assumption is unrealistic in most applicative fields that deal with many vectors of very high dimension. For instance, in biomedical, multimedia, or remote sensing applications, these vectors grows up to millions of values. Besides, along with the democratization of connected devices, data tends to originate from an increasing number of distributed sources with reasonable computing capacity, immersed in large unreliable networks without any central coordination. In contrast to usual centralized PCA, we propose to take advantage of this context to perform a decentralized large scale PCA.

Previous works on decentralized PCA includes a Gossip-based power iteration method to extract principal components one by one when each node holds one vector, using random matrix sparsifications to optimize communication costs [2]. Decentralized implementations of Orthogonal Iteration where also proposed either when data follow a Directed Acyclic Graphical model (DAG) [3], or when seeking the eigenvectors of the network graph adjacency matrix [4]. Aggregation of Mixtures of Probabilistic Principal Components [5] can also be used in conjunction with the Gossip likelihood

*This work is funded by the Culture 3D Cloud project.

maximization framework proposed in [6] to perform a simple PCA. However, none of these methods deal with numerous vectors of very high dimension spread over the network, without any prior knowledge about their distribution.

The main contribution of this paper is a decentralized algorithm that allows all nodes of a network, each one holding a local dataset, to compute PCA over the complete data with reduced local computations and communication cost. It is based on the integration of a dimensionality reduction step into an asynchronous Gossip consensus estimation protocol. Our decentralized approach shows performances that almost equals a centralized PCA run over the complete distributed data.

The paper is organized as follows: after formally defining our problem, we introduce our asynchronous Gossip protocol. We then apply it to covariance estimation and integrate our in-loop dimensionality reduction step. We present experimental results before concluding the paper with some remarks and perspectives.

2 Problem statement

Given a network $\mathcal{G}$ of N nodes, each node i holding a sample $\mathbf{X}_i = [\mathbf{x}_1^{(i)} \dots \mathbf{x}_{n_i}^{(i)}]$, $\mathbf{x}_j^{(i)} \in \mathbb{R}^D$, our goal is to compute the global PCA solution on $\mathbf{X} = [\mathbf{X}_1 \dots \mathbf{X}_N]$ only using node-local computations. In a typical high-dimensional setup, $\forall i, n_i \ll D$. In addition, four constraints must be respected:

C1. Minimal data exchange - Samples cannot be exchanged between nodes.

C2. Asynchrony - Nodes must communicate without any coordination.

C3. Decentralization - All nodes and links must play the same role.

C4. Consensus - All nodes must get the same local solution to ensure that reduced representation are expressed with the same basis across the network.

The global PCA solution is an orthonormal basis $\tilde{\mathbf{U}} = [\mathbf{u}_1 \dots \mathbf{u}_q]$, $\mathbf{u}_k \in \mathbb{R}^D$ that projects the input sample $\mathbf{X}$ into the q-dimensional subspace that retain the maximal variance in $\mathbf{X}$. $\tilde{\mathbf{U}}$ is shown to contain the q leading eigenvectors of the sample covariance matrix $\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^T - \mu\mu^T$, where $\mu = \frac{1}{n}\mathbf{X}^T\mathbf{1}$ is the sample mean [1]. μ and $\mathbf{C}$ cannot be computed at a particular node as they involve the locally unknown full sample $\mathbf{X}$. However, they can be computed as the distributed weighted averages of local sample means $\mu_i = \frac{1}{n_i}\mathbf{X}_i^T\mathbf{1}$ and local uncentered covariance matrices $\mathbf{C}_i = \frac{1}{n_i}\mathbf{X}_i\mathbf{X}_i^T$:

$$\mu = \frac{1}{n}\mathbf{X}^T\mathbf{1} = \frac{1}{n}\sum_i^N \mathbf{X}_i^T\mathbf{1} = \frac{1}{\sum_i n_i}\sum_i^N n_i\mu_i \tag{1}$$

$$\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^T - \mu\mu^T = \frac{1}{n}\sum_i^N \mathbf{X}_i\mathbf{X}_i^T - \mu\mu^T = \frac{1}{\sum_i n_i}\sum_i^N n_i\mathbf{C}_i - \mu\mu^T \tag{2}$$

3 Gossip-based averaging

Among the numerous approaches to distributed weighted averaging, Sum-Weight Gossip protocols is a decentralized solution that meets the above-mentioned requirements [7]. Let each node i hold a couple $(\omega_i, \mathbf{v}_i)$ where $\mathbf{v}_i$ lies in a vector space $\mathcal{V}$

and $\omega_i \in \mathbb{R}$. Sum-Weight protocols estimate the weighted average $\frac{1}{\sum_i \omega_i} \sum_i^N \omega_i \mathbf{v}_i$ by assigning to each node i a sum $\mathbf{s}_i \in \mathcal{V}$ and a weight $w_i \in \mathbb{R}$, with initial values $\mathbf{s}_i(0) \equiv \omega_i \mathbf{v}_i$ and $w_i(0) \equiv \omega_i$. All nodes repeatedly send their current sum and weight to randomly selected peers, while updating them according to incoming values. To respect asynchrony and decentralization, nodes are shipped with independent Poisson emission clocks and selects targets independently at random. The t-th message exchanged on the network thus affects one random sender s and one random receiver r. As shown in [7], for any node i, $\frac{1}{w_i}\mathbf{s}_i$ converge to the desired weighted average.

$$\mathbf{s}_s(t+1) = \frac{1}{2}\mathbf{s}_s(t) \qquad w_s(t+1) = \frac{1}{2}w_s(t) \tag{3}$$

$$\mathbf{s}_r(t+1) = \mathbf{s}_r(t) + \frac{1}{2}\mathbf{s}_s(t) \qquad w_r(t+1) = w_r(t) + \frac{1}{2}w_s(t)\ . \tag{4}$$

$$\forall i, \quad \lim_{t\to\infty} \frac{1}{w_i(t)}\mathbf{s}_i(t) = \frac{1}{\sum_i w_i(0)} \sum_i \mathbf{s}_i(0) = \frac{1}{\sum_i \omega_i} \sum_i \omega_i \mathbf{v}_i \tag{5}$$

Moreover, convergence to the consensus is exponential provided that $\mathcal{G}$ has a sufficient conductance [8]. In this case, the number of message exchanges required to achieve a given estimation error scales logarithmically with the number of nodes.

4 Basic approach : Gossip Late PCA

As weighted averages, μ and $\mathbf{C}$ can be estimated using the above-defined protocol, by defining node-local estimates $\mathbf{a}_i(t), \mathbf{B}_i(t)$ and weights $w_i(t)$ such that:

$$\mathbf{a}_i(0) = \mathbf{X}^T\mathbf{1} \qquad \mathbf{B}_i(0) = \mathbf{X}_i\mathbf{X}_i^T \qquad w_i(0) = n_i \tag{6}$$

We now apply the Gossip protocol defined by Eq.(3-4) to $\mathbf{a}_i(t)$,$\mathbf{B}_i(t)$ and $w_i(t)$. From these estimates, we can get a covariance estimate $\mathbf{C}_i(t)$:

$$\mathbf{C}_i(t) = \frac{\mathbf{B}_i(t)}{w_i(t)} - \frac{\mathbf{a}_i(t)\mathbf{a}_i(t)^T}{w_i(t)^2} \tag{7}$$

Note that initial estimates $\mathbf{C}_i(0)$ are the covariance matrices of their corresponding $\mathbf{X}_i$. The limit in (5) shows that each $\frac{\mathbf{a}_i(t)}{w_i(t)}$ tends to the global mean μ and each $\mathbf{C}_i$ tends to the global covariance matrix $\mathbf{C}$:

$$\forall i, \quad \begin{cases} \lim_{t\to\infty} \frac{\mathbf{a}_i(t)}{w_i(t)} = \frac{\sum_i \mathbf{X}_i^T\mathbf{1}}{\sum_i n_i} = \frac{\sum_i n_i\mu_i}{\sum_i n_i} = \mu \\ \lim_{t\to\infty} \mathbf{C}_i(t) = \frac{\sum_i \mathbf{B}_i(0)}{\sum_i w_i(0)} - \mu\mu^T = \frac{1}{n}\sum_i \mathbf{X}_i\mathbf{X}_i^T - \mu\mu^T = \mathbf{C} \end{cases} \tag{8}$$

Once each node gets a sufficiently accurate estimate for $\mathbf{C}$, the final PCA result can be computed locally at any node i by eigendecomposition of $\mathbf{C}_i$, keeping the q leading eigenvectors as the node-local reduced basis. We call this strategy *late PCA*. Unfortunately, updating estimates $\mathbf{B}_i$ using Eq.(3) requires transmission of $D \times D$ matrices, which is incompatible with our constrained networking context. We thus propose to reduce matrices $\mathbf{B}_i$ by means of local PCA *before* their transmission.

Algorithm 1 Emission Procedure

```
$\mathbf{a}_i \leftarrow \mathbf{X}_i^T\mathbf{1}$ ;  $\mathbf{G}_i \leftarrow \mathbf{X}_i^T\mathbf{X}_i$ ;  $w_i \leftarrow n_i$
$(\mathbf{V}_i, \mathbf{L}_i) \leftarrow \textit{eigendecompose}(\mathbf{G}_i)$
$\mathbf{U}_i \leftarrow \mathbf{X}_i\mathbf{V}_i\mathbf{\Lambda}_i^{-\frac{1}{2}}$
loop
  $j \leftarrow \textit{randomNeighbor}(i)$
  $(\mathbf{a}_i, \mathbf{L}_i, w_i) \leftarrow \frac{1}{2}(\mathbf{a}_i, \mathbf{L}_i, w_i)$
  Send $(\mathbf{a}_i, \mathbf{U}_i, \mathbf{L}_i, w_i)$ to $j$
end loop
```

Algorithm 2 Reception Procedure

```
loop
  Upon receipt of $(\mathbf{a}_j, \mathbf{U}_j, \mathbf{L}_j, w_j)$
  $\mathbf{a}_i \leftarrow \mathbf{a}_i + \mathbf{a}_j$ ;  $w_i \leftarrow w_i + w_j$
  $\mathbf{Q}_0 \leftarrow \mathbf{U}_i$
  for $t \in [0, max\_t[$ do
    $(\mathbf{Q}_{t+1}, \mathbf{R}_{t+1}) \leftarrow$
      $\textit{QR}(\mathbf{U}_i\mathbf{L}_i\mathbf{U}_i^T\mathbf{Q}_t + \mathbf{U}_j\mathbf{L}_j\mathbf{U}_j^T\mathbf{Q}_t)$
  end for
  $\mathbf{U}_i \leftarrow \mathbf{Q}_{max\_t}$ ;  $\mathbf{L}_i \leftarrow \textit{diag}(\mathbf{R}_{max\_t})$
end loop
```

5 Proposed approach : Early Gossip PCA

In contrast to *late PCA*, which first estimates the full covariance and then computes its principal subspace, we propose an *early PCA* approach that integrates dimensionality reduction into the Gossip update rule Eq.(3). To transmit $\mathbf{B}_s$ from node s to node r, s first diagonalize $\mathbf{B}_s$, keeps the q leading eigenpairs and sends this reduced version to r, which reconstructs $\mathbf{B}_s$ and adds it to its own $\mathbf{B}_r$. Consequently, the size of exchanged messages falls to $\mathcal{O}(qD)$ for a desired subspace dimension q. Even though some spectral features may be lost in this reduction process, we argue that this loss, cumulated over time, will be close to the loss induced by *late PCA*. We still face the issue that $\mathbf{B}_i$ are $D \times D$ and might not fit in node-local memory. But because data is distributed, it is likely that $n_i \ll D$. This in mind, two algebraic properties of $\mathbf{B}_i$ solve our problem:

Firstly, explicit computation of the initial $\mathbf{B}_i$ (Algorithm 1) is not required as its eigendecomposition $\mathbf{B}_i = \mathbf{U}_i\mathbf{L}_i\mathbf{U}_i^T$ can be obtained by diagonalizing the $n_i \times n_i$ Gram matrix $\mathbf{X}_i^T\mathbf{X}_i = \mathbf{V}_i\mathbf{\Lambda}_i\mathbf{V}_i^T$. Indeed, since $\mathbf{B}_i^2 = \mathbf{X}_i\mathbf{X}_i^T\mathbf{X}_i\mathbf{X}_i^T = \mathbf{X}_i\mathbf{V}_i\mathbf{\Lambda}_i\mathbf{V}_i^T\mathbf{X}_i^T$ $= (\mathbf{X}_i\mathbf{V}_i\mathbf{\Lambda}_i^{-\frac{1}{2}})\mathbf{\Lambda}_i^2(\mathbf{X}_i\mathbf{V}_i\mathbf{\Lambda}_i^{-\frac{1}{2}})^T$ and $(\mathbf{X}_i\mathbf{V}_i\mathbf{\Lambda}_i^{-\frac{1}{2}})^T(\mathbf{X}_i\mathbf{V}_i\mathbf{\Lambda}_i^{-\frac{1}{2}}) = \mathbf{I}$, we get:

$$\mathbf{U}_i = \mathbf{X}_i\mathbf{V}_i\mathbf{\Lambda}_i^{-\frac{1}{2}} \qquad \text{and} \qquad \mathbf{L}_i = \mathbf{\Lambda}_i \tag{9}$$

We then compute, store and diagonalize $\mathbf{X}_i^T\mathbf{X}_i$ (which is $n_i \times n_i$) instead of $\mathbf{X}_i\mathbf{X}_i^T$ (which is $D \times D$) and obtain $\mathbf{U}_i$ and $\mathbf{L}_i$ with no additional storage cost.

Secondly, when processing an incoming message, explicit reconstruction of received $\mathbf{B}_s(t)$ is also unneeded, as we only want $\mathbf{U}_r(t{+}1)\mathbf{L}_r(t{+}1)\mathbf{U}_r(t{+}1)^T$ to span the q-principal subspace of $\mathbf{B}_s(t){+}\mathbf{B}_r(t)$. Since $\mathbf{B}_s(t)$ and $\mathbf{B}_r(t)$ respectively come as $\mathbf{U}_s(t)\mathbf{L}_s(t)\mathbf{U}_s(t)^T$ and $\mathbf{U}_r(t)\mathbf{L}_r(t)\mathbf{U}_r(t)^T$, we can find $\mathbf{U}_r(t{+}1)$ and $\mathbf{L}_r(t{+}1)$ using Orthogonal Iteration (as in [4]). Starting from any $D \times q$ basis $\mathbf{Q}_0$, and denoting by $\textit{QR}(\cdot)$ the economy QR decomposition, we iteratively compute

$$\mathbf{Q}_{\tau+1}\mathbf{R}_{\tau+1} = \textit{QR}((\mathbf{B}_s + \mathbf{B}_r)\mathbf{Q}_\tau) = \textit{QR}(\mathbf{U}_s\mathbf{L}_s(\mathbf{U}_s^T\mathbf{Q}_\tau) + \mathbf{U}_r\mathbf{L}_r(\mathbf{U}_r^T\mathbf{Q}_\tau)) \tag{10}$$

After a few iterations, $\mathbf{Q}_\tau$ becomes an orthonormal basis for the q-principal subspace of $\mathbf{B}_s(t){+}\mathbf{B}_r(t)$, with corresponding eigenvalues on the diagonal of $\mathbf{R}_\tau$. Thus, $\mathbf{Q}_\infty$ gives $\mathbf{U}_r(t{+}1)$ and the diagonal entries of $\mathbf{R}_\infty$, denoted by $\textit{diag}(\mathbf{R}_\infty)$, give $\mathbf{L}_r(t{+}1)$. Observe that we never store any $D{\times}D$ matrix but rather $D \times q$ ones. Since $q{\ll}D$, all needed

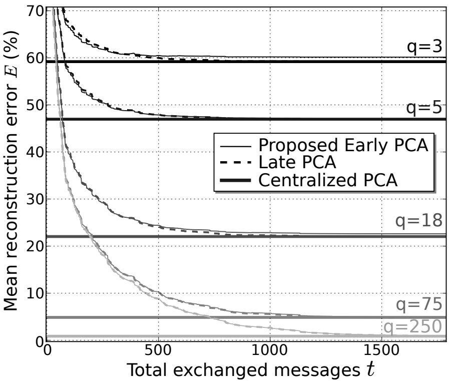

Fig. 1: Mean covariance reconstruction error for early, late and centralized PCA for $q = \{3, 5, 18, 75, 250\}$, from black to light-gray.

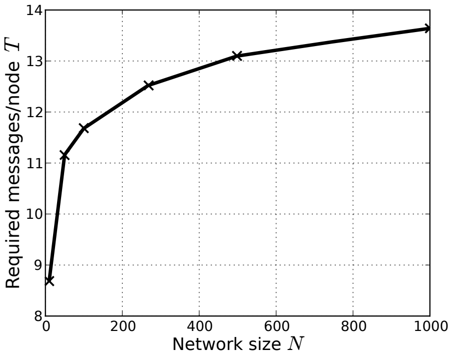

Fig. 2: Average number of messages per node required to converge against network size N

entities now fit in node memory. This results in two concurrent emission and reception procedures run independently at each node, presented in Algorithm 1 and Algorithm 2[1].

Each $D \times q$ matrix $\mathbf{U}_i$ then converge to the same consensus orthonormal basis $\mathbf{U}$ that span the principal subspace of $\mathbf{XX}^T$. Finally, a basis for $\mathbf{XX}^T - \mu\mu^T$ can be locally computed at each node by applying Eq.(10) to $(\mathbf{U}_i\mathbf{L}_i\mathbf{U}_i^T + \frac{1}{w_i^2}\mu_i\mu_i^T)$. All local bases then project local data onto the same consensus global principal components.

6 Experimental results

In this section, we present an experimental evaluation of our proposal on the MNIST [9] handwritten digits training set, which contains 60000 grayscale images of 28×28 pixels, that is, $D = 784$. As both classical PCA and our method estimate q-dimensional bases for the principal subspace of the sample covariance, they are compared in terms of the normalized Euclidean error E of covariance reconstruction. As the number t of exchanged messages and the network size N impact the quality of our Gossip estimation, this error is measured as a function of t and N:

$$E_N(t) = \frac{1}{N\|\mathbf{XX}^T\|_F} \sum_i^N \left\| \mathbf{XX}^T - \mathbf{U}_i(t)\mathbf{L}_i(t)\mathbf{U}_i(t)^T \right\|_F \tag{11}$$

Results for $q = \{3, 5, 18, 75, 250\}$ and network size $N = 100$ are gathered and compared with *late PCA* and centralized PCA in Figure 1. We observe that all nodes obtain bases that are asymptotically very close to a centralized solution, with an exponential convergence with respect to t. The impact of network size is shown in Figure 2

[1] When implementing our method, care should be taken that iterations of Algorithm 1 and Algorithm 2 be mutually exclusive, to ensure integrity of the concurrently updated variables.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

for $q = 50$ by measuring the average number of messages per node required to converge, *i.e*, the lowest T such that $\forall t > T, |E(\frac{t}{N}) - E(\frac{t-1}{N})| < 0.01\%$. T appears to scale logarithmically with N. This means that our method is suitable for very large networks.

7 Conclusion

We presented a decentralized algorithm to solve PCA when data is spread over a (potentially large) network. Based on a randomized Gossip consensus protocol, our method takes into account asynchrony and unreliability of such network, with parsimonious usage of computational and communication resources. The distributed nature of the process has a very low impact on the accuracy of extracted principal components, and all nodes reach a consensus on their projections.

Perspectives include a theoretical convergence analysis and an adaptive selection strategy for q to achieve a bounded reconstruction error with variable-sized bases. Our interest currently focuses on assessing performances of our method in large scale contexts, such as web-scale distributed multimedia retrieval.

References

[1] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[2] Satish Babu Korada, Andrea Montanari, and Sewoong Oh. Gossip PCA. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 209–220. ACM, 2011.

[3] Zhaoshi Meng, Ami Wiesel, and Alfred O Hero. Distributed principal component analysis on networks via directed graphical models. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2877–2880. IEEE, 2012.

[4] David Kempe and Frank McSherry. A decentralized algorithm for spectral analysis. *Journal of Computer and System Sciences*, 74(1):70–83, 2008.

[5] Pierrick Bruneau, Marc Gelgon, and Fabien Picarougne. Aggregation of probabilistic PCA mixtures with a variational-bayes technique over parameters. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 702–705. IEEE, 2010.

[6] Afshin Nikseresht and Marc Gelgon. Gossip-based computation of a gaussian mixture model for distributed multimedia indexing. *Multimedia, IEEE Transactions on*, 10(3):385–392, 2008.

[7] Franck Iutzeler, Philippe Ciblat, and Walid Hachem. Analysis of sum-weight-like algorithms for averaging in wireless sensor networks. *CoRR*, abs/1209.5912, 2012.

[8] Devavrat Shah. *Gossip algorithms*. Now Publishers Inc, 2009.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.