Completion of high-rank ultrametric matrices using selective entries

Aarti Singh, Akshay Krishnamurthy, Sivaraman Balakrishnan and Min Xu
{aarti, akshaykr, sbalakri, minx}@cs.cmu.edu
Carnegie Mellon University, Pittsburgh, 15203, USA

***Abstract*—Ultrametric matrices are hierarchically structured matrices that arise naturally in many scenarios, e.g. delay covariance of packets sent from a source to a set of clients in a computer network, interactions between multi-scale communities in a social network, and genome sequence alignment scores in phylogenetic tree reconstruction problems. In this work, we show that it is possible to complete $n \times n$ ultrametric matrices using only $n \log^2 n$ entries. Since ultrametric matrices are high-rank matrices, our results extend recent work on completion of $n \times n$ low-rank matrices that requires $n \log n$ randomly sampled entries. In the ultrametric setting, a random sampling of entries does not suffice, and we require selective sampling of entries using feedback obtained from entries observed at a previous stage.**

I. Introduction

As the size of modern datasets continues to grow, so does the amount of missing data. This is because the scale and complexity of systems such as the internet, social networks, and biological evolution makes it impossible to monitor them extensively under resource constraints. Thus, data matrices are almost always so undersampled that simply discarding rows/columns with missing entries is tantamount to throwing away all the information.

Recently, several papers have investigated the problem of data matrix completion for large-scale matrices, under the assumption that the matrix of interest is low-rank [1], [2], [3], [4] (or approximately low-rank [5]). These results show that it is possible to reconstruct rank r matrices of size $n \times n$ from only about $nr \log n$ entries, sampled uniformly at *random*. If the rank r of the matrix is small, this implies a significant saving of resources.

In many applications, however, the low rank assumption is not very reasonable. The low rank property implies that the matrix has few independent rows/columns, or there are few latent factors (eigenvectors) that can represent the matrix. In this paper, we consider one such scenario that arises in many practical applications. Ultrametric matrices are hierarchically-structured high-rank matrices that arise in problems where the underlying data-generating mechanism corresponds to a tree. A formal definition is given in section II. Informally, an ultrametric matrix corresponds to a hierarchical block diagonal matrix where, at each scale, the entries within a diagonal block are higher than the matrix entries in off-diagonal blocks. See Figure 1 for an example. The presence of fine-grained structure implies that the matrix is high-rank. In fact, the eigenvalues of these matrices (as shown in section II) do not lie in an ℓ_p ball and hence the matrices cannot be considered as compressible or approximately low-rank matrices. Thus, recent results for low-rank matrix completion [1], [2], [3], [4], [5] cannot be applied to ultrametric matrices.

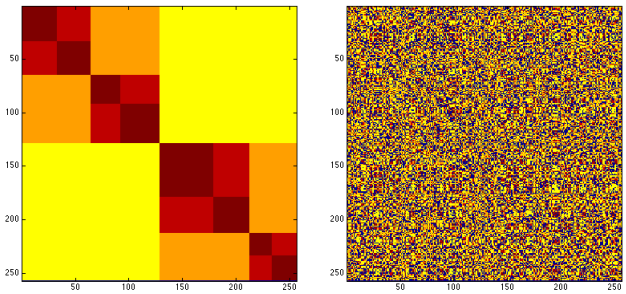

Fig. 1. (a) An ultrametric matrix and (b) The observed matrix is randomly permuted, subsampled and noisy. The dark values correspond to unobserved entries.

Ultrametric matrices arise naturally in many applications. For example, the delay or loss covariance of packets sent from a source to a set of clients in a computer network forms an ultrametric since the shortest path topology between the source and clients is a tree [6], [7], where the internal nodes are routers that divert traffic to different clients. The covariance between clients is multi-scale as it depends on the length of the shared path they have to the source. Hence, clients with longer shared paths have higher covariance (that corresponds to the blocks near the diagonal) and clients with shorter shared paths have lower covariance (that corresponds to off-diagonal blocks).

A similar example is phylogenetics or evolutionary tree reconstruction where the pairwise genome sequence alignment scores between species correspond to shared ancestory [8]. The pairwise scores are higher if the two species differentiated recently and are lower if they split off earlier in the ancestry tree.

As another example, communities in social networks defined by interactions between users are often hierarchically structured [9], [10] as users interact with friends on social networks more often than they interact with family and colleagues, and even less frequently with other acquaintances such as neighbors or friends of friends. A common model for social network interactions and latent factors inducing the interactions is the stochastic block model [9], [11], [12], where

interactions (modeled as edges) occur with higher probability within a group than across groups. A hierarchically-structured stochastic block model precisely corresponds to an ultrametric matrix of interactions.

In all these applications, it is often hard to measure or compute all the matrix entries due to resource constraints. Measuring covariance between clients in a computer network requires sending probe packets which increases network traffic, aligning long genome sequences is computationally challenging and it is impossible to measure the interactions between all users in a large social network. In this paper, we consider the task of recovering ultrametric data matrices that arise in these applications using few observed entries.

Specifically, we show that it is possible to recover the hierarchical block structure of an $n \times n$ ultrametric matrix up to a resolution of $\log n$ using $n \log^2 n$ entries. This implies that the ultrametric matrix can be recovered with small error in Frobenius norm. This extends current results on low-rank matrix completion since even hierarchically-structured high-rank matrices can be recovered with n polylog n entries. However, in the ultrametric setting, a random sampling of entries does not suffice, and we require *selective* sampling of entries using feedback obtained from entries observed at a previous stage.

This paper is organized as follows. Section II mathematically formalizes the set up. Section III describes the methods and results for recovering the ultrametric tree, and section IV presents the results on recovery of matrix entries. We conclude with open directions in section V.

II. Problem Setup

In this paper, we focus on an $n \times n$ ultrametric matrix M^*. M^* is defined as a matrix whose entries are a monotonically descreasing function of an ultrametric and hence satisfy

$$M^*_{ij} \geq \min\{M^*_{ik}, M^*_{jk}\} \ \forall i, j, k$$

with equality if $M^*_{ik} \neq M^*_{jk}$. This condition implies that the matrix is hierarchically block structured [13] as shown in Figure 1. Specifically, a block $S \times S$ corresponds to a subset of indices $S \subseteq \{1, \ldots, n\}$, and any two blocks are either disjoint ($S \cap S' = \emptyset$) or nested ($S \subset S'$ or $S' \subset S$) [1]. Alternatively, the matrix is characterized by an ultrametric tree, where the root corresponds to the entire matrix, each internal node in the tree corresponds to a block, and there are n leaves corresponding to the finest blocks or diagonal entries. Let DS denote the set of indicies corresponding to descendants (in the ultrametric tree) of a node corresponding to S. Then the matrix entries indexed by $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ are a constant, denoted by β_S, where $0 < \beta_S \leq \beta^*$ and $\beta_S < \min_{S' \in DS} \beta_{S'}$. Additionally, we assume that the blocks at each level are balanced, i.e. the smallest block size at any level is $\geq \eta$ times the size of the block at the previous level, where the balance factor $1/2 \geq \eta > 0$ is a constant.

[1]This formal definition of a block only characterizes the diagonal blocks. The off-diagonal blocks will be defined in terms of the (diagonal) blocks.

The discernability of the structure of the matrix or ultrametric tree depends on the variation of the matrix entries between blocks. We define the "gap" as the smallest difference between matrix entries in a block and any sub-block contained within it, i.e.

$$\text{gap} = \min_S \ (\min_{S' \in DS} \beta_{S'} - \beta_S)$$

This quantity will play the role of signal strength in our analysis.

In this paper, we consider the situation where we don't have access to the entire ultrametric matrix. Instead, we have access to a *noisy oracle* that we can query for matrix entries, and it returns the queried entries corrupted by standard Gaussian noise, i.e. an observed entry

$$M_{ij} = M^*_{ij} + N_{ij}$$

where N_{ij} are independently drawn from $\mathcal{N}(0, \sigma^2)$. We will query for a small number of such corrupted entries.

The above ultrametric and noise models suggest a natural notion of Signal-to-Noise Ratio,

$$\text{SNR} = \text{gap}/\sigma.$$

We make high probability guarantees on exact recovery of the blocks up to a certain resolution and bound the deviation of estimated matrix entries from true entries on the recovered blocks.

A. Spectrum of Ultrametric Matrices

Ultrametric matrices are high-rank matrices and their eigenvalues do not lie in an ℓ_p ball. Hence, these matrices cannot be considered as compressible or approximately low-rank matrices. To examplify this fact, we recall the following result that characterizes the eigenspectrum of an ultrametric matrix M^* with a balanced ($\eta = 1/2$) binary tree where all matrix entries corresponding to one level of the tree are a constant i.e. $\beta_S = \beta_{S'}$ if S, S' correspond to the same level in the ultrametric tree [14]. Notice that for such a simple ultrametric matrix, the entries only take one of $\log n + 1$ values $b_0, \ldots, b_L$ corresponding to the $L = \log n$ levels of the ultrametric tree.

1) The eigenvectors of M^* correspond to Haar wavelets. Specifically, the eigenvector corresponding to largest eigenvalue is constant, given as $\frac{1}{\sqrt{n}}\mathbf{1}$ where $\mathbf{1}$ denotes the all 1s vector, and the subsequent eigenvectors are piecewise constants given as

$$v = \frac{\sqrt{|S||S'|}}{\sqrt{|S| + |S'|}} \left[\frac{1}{|S|}\mathbf{1}_S - \frac{1}{|S'|}\mathbf{1}_{S'} \right]$$

where S, S' are siblings in the binary ultrametric tree, and $\mathbf{1}_S$ denotes a vector that is all 1 for indices in S and 0 otherwise.

2) There are $L + 1$ unique eigenvalues of M^* with the smallest eigenvalue $\lambda_0 = b_0 + \sum_{\ell=1}^{L} 2^{\ell-1} b_i$, and the ℓ^{th} smallest unique eigenvalue ($\ell \in 1, \ldots, L$) is $2^{\ell-1}$-fold degenerate and given as

$$\lambda_\ell = b_0 + \sum_{\ell=1}^{L-\ell} 2^{\ell-1} b_i - 2^{L-\ell} b_{L-\ell+1}.$$

This implies that the unique eigenvalues λ_ℓ of the ultrametric matrix M^* are bounded between $c_1 \leq \lambda_\ell \leq c_2(n \log n)/2^\ell$ where c_1, c_2 are constants > 0. Hence the matrix is full-rank and not compressible since the ℓ_p norm of the eigenvalues $\sum_\ell 2^{\ell-1}\lambda_\ell^p$ increases with n.

If the ultrametric tree is not balanced, the Laplacian eigenvectors of the matrix are unbalanced Haar wavelets and the eigenvalues can still be shown to scale as above, provided the balance factor η is a constant bounded away from zero [15].

III. Recovery of ultrametric tree

In this section, we consider methods for recovering the ultrametric tree structure using few matrix entries.

In [16], we developed an active spectral clustering (ASC) algorithm for recovering a hierarchical clustering using few selective pairwise similarities. The algorithm is outlined below.

Algorithm 1 ASC (Active Spectral Clustering)

Input: Oracle, index set S, sampling parameter s
  $O \subset S$ of size s uniformly at random, $\Omega = O \times S$
  Query oracle for M_Ω
  $D \leftarrow$ diagonal matrix with $D_{ii} = \sum_{j \in O} M_{ij}$
  Compute Laplacian $L = D - M_{O \times O}$
  $v \leftarrow$ smallest non-constant eigenvector of L
  $\mathcal{C} \leftarrow$ groups of indices of v that are constant
  for $i \in S \setminus O$ **do**
    $C' \leftarrow \arg\max_{C \in \mathcal{C}} \frac{1}{|C|} \sum_{j \in C} M_{ji}$
    $C' \leftarrow C' \cup i$
  end for
Output: $\{C, \text{ ASC}(\text{Oracle}, C, s)\}_{C \in \mathcal{C}}$

For a similarity matrix that corresponds to an ultrametric with binary tree, the results of [16] imply the following theorem.

Theorem 1. *Suppose the SNR and balance factor η are constant. Then there exists $n_0 \equiv n_0(SNR, \eta)$ s.t. $\forall n \geq n_0$, Algorithm 1 given index set $S = \{1, \ldots, n\}$ and a sampling parameter $s \geq \log^2 n$ returns a collection of nested partitions that is consistent with the ultrametric binary tree up to a level where the blocks have size $|S| \geq s$ with probability $> 1 - n^{-1}$ using $O(ns \log n)$ entries of the matrix.*

Thus, the algorithm can recover the ultrametric tree up to resolution $\log^2 n$ using $O(n \log^3 n)$ matrix entries. We also have the following corollary to Theorem 1.

Corollary 1. *Suppose the SNR and balance factor η are constant. Then there exists $n_0 \equiv n_0(SNR, \eta)$ s.t. $\forall n \geq n_0$, Algorithm 1 given index set $S = \{1, \ldots, n\}$ and a sampling parameter $s \geq \log^2 n$, with probability $> 1 - 3n^{-1}$, queries at least $\eta^2/2 \cdot |S| \log^2 n$ entries in $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ for all blocks of size $|S| \geq s$.*

Proof: Theorem 1 implies that, with probability $> 1 - n^{-1}$, every block of size $|S| \geq s$ is passed as an input index set to the ASC algorithm at some iteration. Thus, $|S|s \geq |S| \log^2 n$ entries are queried in every such block. Let E denote the event under which Theorem 1 holds. Since the blocks at each level are balanced, under event E, the expected number of entries queried in $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ are at least $\eta^2|S|s$. Using *relative* Chernoff bound [17], we get that with probablity $> 1 - e^{\eta s/8}$ the number of entries queried in $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ for some block S with size $|S| \geq s$ is $\geq \eta^2|S|s/2$. Taking union bound over all blocks, we get with probability $> 1 - 2ne^{-\eta s/8} \geq 1 - 2n^{-1}$ (for $n \geq e^{16/\eta}$), under event E, the number of entries queried in $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ for all blocks of size $|S| \geq s$ is $\geq \eta^2|S|s/2$. ∎

Remark: If the balance factor η and SNR are known, then we can set the sampling parameter per iteration $s > c \cdot \log n$ for some constant $c > 0$ and recover the ultrametric tree up to resolution of $\log n$ using $O(n \log^2 n)$ entries of the matrix with probability $> 1 - n^{-1}$.

The ASC algorithm proceeds by randomly sampling few rows/columns of a sub-matrix at each iteration. Next, we investigate if it is possible to recover the ultrametric matrix if we randomly sample entries of the blocks/sub-matrices. While the sampling of blocks at a level can be random, we still need to focus sampling within sub-blocks as we recover finer resolution blocks. A completely random sampling of n polylog n matrix entries does not suffice to recover the ultrametric tree up to resolution $\log^2 n$. See Proposition 1 in [18] for a formal proof.

We consider the following algorithm that is an iterative version of the method proposed in [19].

Algorithm 2 ZFS (Zero-fill Spectral)

Input: Oracle, index set S, sampling budget b.
  $\Omega = b$ entries $(i, j) \in S \times S, i < j$ chosen uniformly at random (without replacement)
  Query Oracle for M_Ω
  $W_{ij} = \begin{cases} 2b/(|S|(|S|-1)) & \text{if } i = j \\ M_{ij} & \text{if } (i,j) \text{ or } (j,i) \in \Omega \\ 0 & \text{otherwise} \end{cases}$
  $L \leftarrow D - W$ where $D_{ii} = \sum_{j \in S} W_{ij}$
  $v \leftarrow$ smallest non-constant eigenvector of L
  $\mathcal{C} \leftarrow$ groups of indices of v that are constant
Output: $\{C, \text{ ZFS}(\text{Oracle}, S \cap C, p)_{C \in \mathcal{C}}$.

The results of [19] imply the following result in the ultrametric setting where the observed entries are noiseless.

Theorem 2. *Suppose the gap and balance factor η is constant. Then there exists $n_0 \equiv n_0(gap, \eta)$ s.t. $\forall n \geq n_0$, any iteration of Algorithm 2 given index set S and budget $b < |S|(|S|-1)/4$, with probability $> 1 - n^{-2}$ returns a vector v that satisifes*

$$\|v - u\|_2 = O\left(\sqrt{\frac{|S| \log n}{b}}\right)$$

where u is the smallest non-constant Laplacian eigenvector of the fully observed noiseless ultrametric submatrix $M_{S \times S}$.

Since the Laplacian eigenvectors of a noiseless binary ultrametric matrix capture the tree structure, we can guarantee recovery of the ultrametric tree if we can show that the ℓ_∞ bound on the eigenvector perturbation is smaller than any entry of the Laplacian eigenvector (which scales as $O(1/\sqrt{|S|})$ [15]). Following the arguments in [15] for converting an ℓ_2 perturbation bound on Laplacian eigenvectors to an ℓ_∞ perturabtion bound, we get that with probability $> 1 - n^{-2}$ $\|v - u\|_\infty = O\left(\frac{|S| \log n}{b}\right)$. This implies that we need at least $b \geq |S|^{3/2} \log n$ sampling budget at each iteration to guarantee recovery of the first split of S in the ultrametric. Taking a union bound over all blocks, we have the following corollary.

Corollary 2. *Suppose the gap and balance factor η is constant. Then there exists $n_0 \equiv n_0(gap, \eta)$ s.t. $\forall n \geq n_0$, Algorithm 2 given index set $S = \{1, \ldots, n\}$ and a sampling budget $b \geq |S|^{3/2} \log n$ per iteration returns a collection of nested partitions that is consistent with the ultrametric binary tree up to a level where the blocks have size $|S| \geq \log^2 n$ with probability $> 1 - n^{-1}$ using $O(n^{3/2} \log^2 n)$ noiseless entries of the matrix. Moreover, with probability $> 1 - 3n^{-1}$, it is guaranteed to query at least $\eta^2/2 \cdot |S|^{3/2} \log n$ entries in $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ for all blocks of size $|S| \geq \log^2 n$.*

The argument about number of queries in off-diagonal block follows by arguments similar to proof of Corollary 1. Thus, randomly sampling entries of the sub-matrix at each iteration seems to require more measurements than randomly sampling rows/columns at each iteration. It is not clear if this is a limitation of the ZFS method or a fundamental difference between the two measurement models.

The algorithms discussed in this section return a collection of nested partitions $\cup_{\mathcal{C}}\mathcal{C}$ of the index set S and partially sampled matrix entries $\cup_\Omega M_\Omega$. To recover an estimate $\widehat{M^*}$ of the entire noiseless matrix M^*, we post-process the outputs of these algorithms through Algorithm 3 as described in the next section.

IV. Recovery of matrix entries

We consider the following simple algorithm to recover the matrix entries. If the algorithm is given a nested partition that is consistent with the ultrametric tree structure up to a certain level, the algorithm essentially knows S and DS for these blocks and averages the entries observed within $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ to estimate β_S.

The following theorem provides an entry-wise bound on the largest deviation between the estimated and true matrix entries, and on the Frobenius norm of the error matrix.

Theorem 3. *Given a nested partition that is consistent with the ultrametric tree structure up to a level where the blocks have size $|S| \geq s$, and given $c|S| \log^2 n$ observed entries in $(S \times S) \setminus \cup_{S' \in DS}(S' \times S')$ for every such block where $c > 0$ is a constant, with probability $> 1 - 2n^{-1}$, Algorithm 3 yields*

$$|\widehat{M}^*_{ij} - M^*_{ij}| \leq \frac{1}{\sqrt{|S|}} \leq \frac{1}{\sqrt{s}}$$

Algorithm 3 RM (Reconstruct Matrix)

Input: Nested partitions $\cup_{\mathcal{C}}\mathcal{C}$ and partially sampled matrix entries $\cup_\Omega M_\Omega$.

```
P = ∅
repeat
  C' ← the finest partition in ∪_C C
  for C ∈ C' do
    I = {(i,j) ∈ Ω : i ∈ C \ P, j ∈ C \ P}
    M̂*_{C\P} = (1/I) Σ_{(i,j)∈I} M_ij
  end for
  ∪_C C ← ∪_C C \ C'
  P ← P ∪ C'
until ∪_C C = ∅
```

for all blocks of size $|S| \geq s$ and all $(i, j) \in S \times S$, provided $n > e^{5\sigma^2/c}$. Additionally, with probability $> 1 - 2n^{-1}$

$$\|\widehat{M^*} - M^*\|_F^2 := \sum_{ij} (\widehat{M}^*_{ij} - M^*_{ij})^2 = O(\beta^{*2} n s^2)$$

where β^* is the largest entry of the matrix M^*.

Proof: For entries that correspond to blocks with size $|S| \geq s$, we average the observed entries and bound the deviation as follows: Let $B = (S \times S) \setminus \cup_{S' \in DS}(S' \times S')$. Since $|\Omega \cap B| \geq c|S| \log^2 n$, we have

$$\begin{aligned} &Pr\left(\left|\beta_S - \frac{1}{|\Omega \cap B|} \sum_{(i,j) \in |\Omega \cap B|} M_{ij}\right| > \frac{1}{\sqrt{|S|}}\right) \\ &= Pr\left(\left|\mathcal{N}\left(0, \frac{\sigma^2}{|\Omega \cap B|}\right)\right| > \frac{1}{\sqrt{|S|}}\right) \\ &\leq \sqrt{|S|} e^{-\frac{c \log^2 n}{2\sigma^2}} \leq n^{\frac{1}{2} - \frac{c \log n}{2\sigma^2}} \end{aligned}$$

The total number of such blocks $\leq 2n$ (bound on the total number of nodes in the ultrametric tree), hence with probability $> 1 - 2n^{\frac{3}{2} - \frac{c \log n}{2\sigma^2}} \geq 1 - 2n^{-1}$ (for n large enough, $n > e^{5\sigma^2/c}$), for all blocks of size $|S| \geq s$ and all $(i, j) \in S \times S$

$$|\widehat{M}^*_{ij} - M^*_{ij}| \leq \frac{1}{\sqrt{|S|}} \leq \frac{1}{\sqrt{s}}.$$

Thus, the total squared error of entries for every block of size $|S| \geq s$ is $\leq |S|$. Now there are up to $1/\eta$ blocks with $\eta n \leq |S| \leq (1 - \eta)n$, there are up to $1/\eta^2$ blocks with $\eta^2 n \leq |S| \leq (1 - \eta)^2 n$, there are up to $1/\eta^3$ blocks with $\eta^3 n \leq |S| \leq (1 - \eta)^3 n$, and so on until there are up to $1/\eta^L$ blocks with $\eta^L n \leq |S| \leq (1 - \eta)^L n$. Therefore, the sum of squared entries for all blocks of size $\geq \log^2 n$ is

$$\begin{aligned} &\leq \frac{1 - \eta}{\eta} n \left(1 + \frac{1 - \eta}{\eta} + \frac{(1 - \eta)^2}{\eta^2} + \ldots + \frac{(1 - \eta)^{L-1}}{\eta^{L-1}}\right) \\ &\leq \frac{1 - \eta}{\eta} n \left[\frac{(1 - \eta)^L}{\eta^L} - 1\right] \frac{\eta}{1 - 2\eta} \\ &\leq \frac{\eta}{1 - 2\eta} n \\ &= O(n) \end{aligned}$$

where the last step holds since $\eta^L n \geq s$ and $(1-\eta)^{L+1} n < s$ since L is the level at which the largest size of a cluster below that level is $< s$.

For all levels with block sizes smaller than $\log^2 n$, the error can be large. Suppose the largest entry of M^* is bounded by β^*, for matrix entries corresponding to blocks with size smaller than s, we have $|\widehat{M}^*_{ij} - M^*_{ij}| \leq \beta^*$. Since there are $\leq 1/\eta^{L+1} \leq n$ such blocks (since $\eta^{L+1} n \geq 1$), the total squared error of all entries in such blocks $\leq \beta^{*2} n s^2$. Thus, the overall Frobenius norm

$$\|\widehat{M}^* - M^*\|_F^2 \quad := \quad \sum_{ij} (\widehat{M}^*_{ij} - M^*_{ij})^2 = O(\beta^{*2} n s^2)$$

■

Thus, all entries of the ultrametric matrix corresponding to blocks with size $|S| \geq \log^2 n$ can be recovered consistently with high probability using the methods mentioned in the previous section, and the overall Frobenius norm error in recovering the matrix is $O(\beta^{*2} n \log^4 n)$.

Remark: If the ultrametric is constant (no gap) for all block sizes $< \log^2 n$, then *all* entries of the matrix can be recovered consistently, while the rank of the matrix is still high $(n/\log^2 n)$.

V. Discussion

In this paper, we demonstrated that it is possible to recover high-rank matrices that are hierarchically structured using $n \log^2 n$ (or $n \log^3 n$) selective matrix entries. There are several interesting directions that can be developed further.

First, we only consider ultrametric matrices which require that the matrix entries in each off-diagonal block is a constant. In earlier work [18], [16] we have shown that it is possible to recover the underlying tree structure even when the matrix entries are not constant but there is a gap between the values of matrix entries within a block and its sub-blocks. It is not clear whether accurate guarantees on matrix recovery can be provided in this setting.

Second, algorithm ASC works with randomly sampling rows or columns of sub-matrices at each iteration. In many applications, this might not be preferred e.g. in a computer or social network scenario, this requires few nodes or users to receive most packets or reveal their interactions with everyone. The ZFS method we considered for randomly sampling entries of sub-matrices at each iteration is suboptimal as it requires many more observations. It is an open question whether a method for recovering the ultrametric tree and matrix with $O(n \text{ polylog } n)$ randomly chosen entries within submatrices at each iteration exists.

On the other extreme, is it possible to be more adaptive and do better, i.e. at each iteration instead of randomly sampling rows/columns or entries of sub-matrices, can we selectively sample the submatrices? In [18] we demonstrated a method that, in the noiseless setting, recovers the ultrametric tree using only $n \log n$ entries with no requirements on the balance factor of the tree. However, it is not clear whether that can be extended to handle noise and whether it queries enough entries to guarantee recovery of the matrix.

Finally, it should be possible to recover more general hierarchically structured matrices that are not ultrametrics, e.g. a rectangular matrix that contains randomly placed submatrices with hierarchical structure. Also, it would be interesting to characterize the tradeoff between how many matrix entries are observed and the resulting accuracy in matrix recovery.

Acknowledgment

This research is supported in part by AFOSR under grant FA9550-10-1-0382 and NSF under grant IIS-1116458.

References

[1] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[2] E. J. Candés and B. Recht, "Exact matrix completion via convex optimization," *To appear in Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2008.

[3] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[4] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, 2010.

[5] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," in *International Conference on Machine Learning (ICML)*, 2010.

[6] N. Duffield and F. L. Presti, "Network tomography from measured end-to-end delay covariance," *IEEE/ACM Transactions on Networking*, 2004.

[7] M. Coates, A. Hero, R. Nowak, and B. Yu, "Internet tomography," *IEEE Signal Processing Magazine*, vol. 19, no. 3, pp. 47–65, May 2002.

[8] J. Kim and T. Warnow, "Tutorial on phylogenetic tree estimation," in *Intelligent Systems for Molecular Biology*, 1999.

[9] P. Holland, K. Laskey, and S. Leinhardt, "Stochastic blockmodels: Some first steps," *Social Networks*, vol. 5, no. 109–137, 1983.

[10] M. Schweinberger and T. Snijders, "Setting in social networks: A measurement model," *Sociological Methodology*, vol. 33, pp. 307–342, 2003.

[11] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic block model," *Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[12] D. Sussman, M. Tang, D. Fishkind, and C. Priebe, "A consistent dot product embedding for stochastic blockmodel graphs," *Journal of the American Statistical Association, Accepted*, 2012.

[13] N. Jardine and R. Sibson, *Mathematical taxonomy*. New York: Wiley, 1971.

[14] A. T. Ogielski and D. L. Stein, "Dynamics on ultrametric spaces," *Physical Review Letters*, vol. 55, pp. 1634–1637, 1985.

[15] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh, "Noise thresholds for spectral clustering," in *Neural Information Processing Systems (NIPS)*, 2011.

[16] A. Krishnamurthy, S. Balakrishnan, M. Xu, and A. Singh, "Efficient active algorithms for hierarchical clustering," in *International Conference on Machine Learning (ICML)*, 2012.

[17] T. Hagerup and C. Rüb, "A guided tour of chernoff bounds," *Inf. Processing Lett.*, vol. 33, no. 6, pp. 305—308, 1990.

[18] B. Eriksson, G. Dasarathy, A. Singh, and R. Nowak, "Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities," in *International Conference on Artificial Intelligence and Statistics, AISTATS*, 2011.

[19] O. Shamir and N. Tishby, "Spectral clustering on a budget," in *Artificial Intelligence and Statistics (AISTATS)*, 2011.