Multiresolution Matrix Factorization

Risi Kondor RISI@UCHICAGO.EDU
Nedelina Teneva NEDTENEVA@GMAIL.COM
Department of Computer Science, University of Chicago
Vikas Garg VKG@TTIC.EDU
Toyota Technological Institute

Abstract

The types of large matrices that appear in modern Machine Learning problems often have complex hierarchical structures that go beyond what can be found by traditional linear algebra tools, such as eigendecompositions. Inspired by ideas from multiresolution analysis, this paper introduces a new notion of matrix factorization that can capture structure in matrices at multiple different scales. The resulting Multiresolution Matrix Factorizations (MMFs) not only provide a wavelet basis for sparse approximation, but can also be used for matrix compression (similar to Nyström approximations) and as a prior for matrix completion.

1. Introduction

Recent years have seen a surge of work on compressing and estimating large matrices in a variety of different ways, including (i) low rank approximations (Drineas et al., 2006; Halko et al., 2009), (ii) matrix completion (Achlioptas & McSherry, 2007; Candès & Recht, 2009); (iii) compression (Williams & Seeger, 2001; Kumar et al., 2012), and (iv) randomized linear algebra (see (Mahoney, 2011) for a review). Each of these requires some assumption about the matrix at hand, and invariably that assumption is that the matrix is of low rank. In this paper we offer an alternative to the low rank paradigm by introducing multiresolution matrices, and argue that in many contexts it better captures the true nature of matrices arising in learning problems.

To contrast the two approaches, recall that saying that a symmetric matrix $A \in \mathbb{R}^{n\times n}$ is of rank $r \ll n$ means that it can be expressed in terms of a dictionary of r mutually orthogonal unit vectors $\{u_1, u_2, \ldots, u_r\}$ in the form

$$A = \sum_{i=1}^{r} d_i u_i u_i^\top, \qquad (1)$$

where $u_1, \ldots, u_r$ are the normalized eigenvectors of A and $d_1, \ldots, d_r$ are the corresponding eigenvalues. This is the decomposition that Principal Component Analysis (PCA) finds and it corresponds to the factorization

$$A = U^\top D U \qquad (2)$$

with $D = \mathrm{diag}\,(d_1, \ldots, d_r, 0, 0, \ldots, 0)$ and U orthogonal.

The drawback of PCA is that eigenvectors are almost always dense, while matrices occurring in learning problems, especially those related to graphs, often have strong locality properties, whereby they more closely couple certain clusters of "nearby" coordinates than those farther apart according to some underlying topology. In such cases, modeling A in terms of a basis of global eigenfunctions is both computationally wasteful and conceptually absurd: a localized dictionary would be much more appropriate. This is part of the reason for the recent interest in sparse PCA (sPCA) algorithms (Jenatton et al., 2010), in which the $\{u_i\}$ dictionary vectors of (2) are constrained to be sparse, while the orthogonality constraint may be relaxed. However, sPCA is liable to suffer from the opposite problem of capturing structure locally, but failing to recover larger scale patterns in A.

In contrast to PCA and sPCA, the multiresolution factorizations introduced in this paper tease out structure at multiple different scales by applying not just one, but a sequence of sparse orthogonal transforms to A. After the first orthogonal transform, the subset of rows/columns of $U_1 A U_1^\top$ which interact the least with the rest of the matrix capture the finest scale structure in A, so the corresponding rows of U_1 are designated level one wavelets, and these dimensions are subsequently kept invariant. Then the process is repeated by applying a second orthogonal transform to yield $U_1 U_2 A U_1^\top U_2^\top$ and splitting off another subspace of $\mathbb{R}^n$ spanned by second level wavelets, and so on, ultimately resulting in an L level factorization of the form

$$A = U_1^\top U_2^\top \ldots U_L^\top H U_L \ldots U_2 U_1. \qquad (3)$$

For a given type of sparsity constraint on $U_1, \ldots, U_L$ and a given rate at which dimensions must be eliminiated, matrices that are expressible in this form with H diagonal (except for a specific small block which might be dense) we call multiresolution factorizable.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

Multiresolution matrix factorization (MMF) uncovers soft hierarchical organization in matrices characteristic of naturally occurring large networks or the covariance structure of large collections of random variables, without enforcing a hard hierarchical clustering. In addition to using MMF as an exploratory tool, we suggest that

1. MMF structure may be used as a "prior" in matrix approximation and completion problems;
2. MMF can be used for matrix compression, since each intermediate $U_\ell \ldots U_1 A U_1^\top \ldots U_\ell^\top$ is effectively a compressed version of A;
3. The wavelet basis associated with MMF is a natural basis for sparse approximation of functions on a domain whose metric structure is given by A.

In the following we discuss the relationship of MMF to classical multiresolution analysis (Section 3), propose algorithms for computing MMFs (Section 4), take the first steps to analyze their theoretical properties (Section 5) and provide some experiments (Section 6). The proofs of all propositions and theorems are in the supplement.

1.1. Related work

Our work is related to several other recent lines of work on constructing wavelet bases on discrete spaces. The work of Coifman & Maggioni (2006) on Diffusion Wavelets was a major inspiration, especially in emphasizing the connection to classical harmonic analysis. The tree-like structure of MMFs relates them to the recent work of Gavish et al. (2010) on multiresolution on trees, and in particular to Treelets (Lee et al., 2008), which is a direct precursor to this paper. Finally, the spectral graph wavelets of Hammond et al. (2011) establish the connection between Fourier analysis and spectral graph theory, and how this can be used as a basis for bulding multiresolution on graphs.

More generally, the idea of multilevel operator compression is related to both algebraic multigrid methods (e.g., (Livne & Brandt, 2011)) and fast multipole expansions (Greengard & Rokhlin, 1987). In the machine learning community, ideas of multiscale factorization and clustering appeared in (Dhillon et al., 2007)(Savas & Dhillon, 2011), amongst other works.

2. Notation

We define $[n]=\{1,2,\ldots,n\}$. The n dimensional identity matrix we denote I_n unless n is obvious from the context, in which case we will just use I. The i'th row of a matrix M is $M_{i,:}$ and the j'th column is $M_{:,j}$. We use $\uplus$ to denote the disjoint union of two sets, so $S_1 \uplus \ldots \uplus S_m = T$ is a partition of T. The group of n dimensional orthogonal matrices is $\mathrm{SO}(n)$.

$$\begin{tikzcd}
L_2(X) \arrow[r] & \ldots \arrow[r] & V_0 \arrow[r] \arrow[dr] & V_1 \arrow[r] \arrow[dr] & V_2 \arrow[r] \arrow[dr] & \ldots \\
 & & & W_1 & W_2 & W_3
\end{tikzcd}$$

Figure 1. Multiresolution analysis repeatedly splits $V_0, V_1, \ldots$ into a smoother part V_{j+1} and a rougher part W_{j+1}.

Given a matrix $M \in \mathbb{R}^{n\times m}$ and two sequences of indices $I=(i_1,\ldots,i_k)\in[n]^k$ and $J=(j_1,\ldots,j_\ell)\in[m]^\ell$, $M_{I,J}$ will denote the $k\times\ell$ submatrix of M cut out by rows $i_1,\ldots,i_k$ and columns $j_1,\ldots,j_\ell$, i.e., the matrix whose entries are $[M_{I,J}]_{a,b} = M_{i_a,j_b}$. Similarly, if $S=\{i_1,\ldots,i_k\}\subseteq[n]$ and $T=\{j_1,\ldots,j_\ell\}\subseteq[m]$ (assuming $i_1<i_2<\ldots<i_k$ and $j_1<j_2<\ldots<j_k$), $M_{S,T}$ will be the $k\times\ell$ matrix with entries $[M_{T,S}]_{a,b}=M_{i_a,j_b}$.

Given $M_1 \in \mathbb{R}^{n_1\times m_1}$ and $M_1 \in \mathbb{R}^{n_1\times m_1}$, $M_1 \oplus M_2$ is the $(n_1+n_2)\times(m_1+m_2)$ dimensional matrix with entries

$$[M_1\oplus M_2]_{i,j} = \begin{cases} [M_1]_{i,j} & \text{if } i\leq n_1 \text{ and } j\leq m_1 \\ [M_2]_{i-n_1,j-m_1} & \text{if } i>n_1 \text{ and } j>m_1 \\ 0 & \text{otherwise.} \end{cases}$$

A matrix M is said to be block diagonal if it is of the form

$$M = M_1 \oplus M_2 \oplus \ldots \oplus M_p \tag{4}$$

for some sequence of smaller matrices $M_1,\ldots,M_p$. We will only deal with block diagonal matrices in which each of the blocks is square. To remove the restriction that each block in (4) must involve a contiguous set of indices we introduce the notation

$$M = \oplus_{(i^1_1,\ldots,i^1_{k_1})} M_1 \oplus_{(i^2_1,\ldots,i^2_{k_2})} M_2 \ldots \oplus_{(i^p_1,\ldots,i^p_{k_p})} M_p \tag{5}$$

for the generalized block diagonal matrix whose entries are

$$M_{a,b} = \begin{cases} [M_u]_{q,r} & \text{if } i^u_q=a \text{ and } i^u_r=b \text{ for some } u,q,r, \\ 0 & \text{otherwise}\,. \end{cases}$$

We will sometimes abbreviate expressions like (5) by dropping the first $\oplus$ operator and its indices.

3. Multiresolution Analysis

Given a measurable space X, Fourier analysis filters functions on X according to smoothness by expressing them in the eigenbasis of an appropriate self-adjoint smoothing operator T. On $X=\mathbb{R}^d$, for example, T might be the inverse of the Laplacian $\nabla^2 = \frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\ldots+\frac{\partial^2}{\partial x_d^2}$, leading to the Fourier transform $\widehat{f}(k)=\int f(x)\,e^{-2\pi i k\cdot x}dx$. When X is a graph and T is the graph Laplacian or a diffusion operator, the same ideas lead to spectral graph theory. Thus, Fourier analysis corresponds to the eigendecomposition $T=U^\top D U$ or its operator counterpart.

In contrast, Multiresolution Analysis (MRA) constructs a sequence of spaces of functions of increasing smoothness

$$L_2(X) \supset \ldots \supset V_0 \supset V_1 \supset V_2 \supset \ldots \tag{6}$$

by repeatedly splitting each V_j into a smoother part V_{j+1}, and a rougher part W_{j+1} (Figure 1). The further we go down this sequence, the longer the length scale over which typical functions in V_j vary, thus, projecting a function to $V_j, V_{j+1}, \ldots$ amounts to resolving it at different levels of resolution. This inspired Mallat to define multiresolution analysis on $X=\mathbb{R}$ directly in terms of dilations and translations by the following axioms (Mallat, 1989):

1. $\bigcap_j V_\ell = \{0\}$,
2. $\bigcup_\ell V_\ell = L_2(\mathbb{R})$,
3. If $f \in V_\ell$ then $f'(x) = f(x - 2^\ell m)$ is also in V_ℓ for any $m \in \mathbb{Z}$,
4. If $f \in V_\ell$, then $f'(x) = f(2x)$ is in $V_{\ell-1}$.

These imply the existence of a so-called "mother wavelet" ψ such that each W_ℓ is spanned by an orthonormal basis

$$\Psi_\ell = \{\, \psi_{\ell,m}(x) = 2^{-\ell/2}\, \psi(2^{-\ell}x - m) \,\}_{m\in\mathbb{Z}}$$

and a "father wavelet" ϕ such that each V_ℓ is spanned by an orthonormal basis[1]

$$\Phi_\ell = \{\, \phi_{\ell,m}(x) = 2^{-\ell/2} \phi(2^{-\ell}x - m) \,\}_{m\in\mathbb{Z}}.$$

The wavelet transform (up to level L) of a function $f\colon X \to \mathbb{R}$ residing at a particular level of the hierarchy (6), without loss of generality $f \in V_0$, expresses it as

$$f(x) = \sum_{\ell=1}^{L} \sum_m \alpha_m^\ell\, \psi_m^\ell(x) + \sum_m \beta_m\, \phi_m^L(x), \qquad (7)$$

with $\alpha_m^\ell = \langle f, \psi_m^\ell \rangle$ and $\beta_m = \langle \phi_m^L, f \rangle$. Multiresolution owes much of its practical usefulness to the fact that ψ can be chosen in such a way that (a) it is localized in both space and frequency; (b) the individual $U_\ell \colon V_{\ell-1} \to V_\ell \oplus W_\ell$ basis transforms are sparse. Thus, (7) affords a computationally efficient way of decomposing functions into components at different levels of detail, and provides an excellent basis for sparse approximations.

3.1. Multiresolution on discrete spaces

The problem with extending multiresolution to less structured and discrete spaces, such as graphs, is that in these settings there are no obvious analogs of translation and dilation, required by Mallat's third and fourth axioms. Rather, similarly to (Coifman & Maggioni, 2006), assuming that $|X| = n$ is finite, we adopt the view that multiresolution analysis with respect to a symmetric smoothing matrix $A \in \mathbb{R}^{n\times n}$ now consists of finding a sequence of spaces

$$V_L \subset \ldots \subset V_2 \subset V_1 \subset V_0 = L(X) \cong \mathbb{R}^n \qquad (8)$$

where each V_ℓ has an orthonormal basis $\Phi_\ell := \{\phi_m^\ell\}_m$ and each complementary space W_ℓ has an orthonormal basis $\Psi_\ell := \{\psi_m^\ell\}_m$ satisfying the following conditions:

MRA1. The sequence (8) is a filtration of $\mathbb{R}^n$ in terms of smoothness with respect to A in the sense that

$$\eta_\ell = \sup_{v \in V_\ell} \langle v, Av \rangle \,/\, \langle v, v \rangle$$

decays at a given rate.

MRA2. The wavelets are localized in the sense that

$$\mu_\ell = \max_{m\in\{1,\ldots,d_\ell\}} \|\psi_m^\ell\|_0,$$

increases no faster than a certain rate.

MRA3. Letting U_ℓ be the matrix expressing $\Phi_\ell \cup \Psi_\ell$ in the previous basis $\Phi_{\ell-1}$, i.e.,

$$\phi_m^\ell = \textstyle\sum_{i=1}^{\dim(V_{\ell-1})} [U_\ell]_{m,i}\; \phi_i^{\ell-1} \qquad (9)$$

$$\psi_m^\ell = \textstyle\sum_{i=1}^{\dim(V_{\ell-1})} [U_\ell]_{m+\dim(V_{\ell-1}),i}\; \phi_i^{\ell-1}, \quad (10)$$

each U_ℓ is sparse, guaranteeing the existence of a fast wavelet transform (Φ_0 is taken to be the standard basis, $\phi_m^0 = e_m$).

3.2. Multiresolution Matrix Factorization

The central idea of this paper is to convert multiresolution analysis into a matrix factorization problem by focusing on how it compresses the matrix A. In particular, extending each U_ℓ matrix to size $n\times n$ by setting $U_\ell \leftarrow U_\ell \oplus I_{n-\dim(V_{\ell-1})}$, we find that in the $\Phi_1 \cup \Psi_1$ basis A becomes $U_1 A U_1^\top$. In the $\Phi_2 \cup \Psi_2 \cup \Psi_1$ basis it becomes $U_2 U_1 A U_1^\top U_2^\top$, and so on, until finally in the $\Phi_L \cup \Psi_L \cup \ldots \cup \Psi_1$ basis it takes on the form

$$H = U_L \ldots U_2 U_1 A U_1^\top U_2^\top \ldots U_L. \qquad (11)$$

Therefore, similarly to the way that Fourier analysis corresponds to eigendecomposition, multiresolution analysis effectively factorizes A in the form

$$A = U_1^\top U_2^\top \ldots U_L H U_L \ldots U_2 U_1 \qquad (12)$$

with the constraints that (a) each U_ℓ orthogonal matrix must be sufficiently sparse; (b) outside its top left $\dim(V_{\ell-1}) \times \dim(V_{\ell-1})$ block, each U_ℓ is the identity. Furthermore, by (9), the first $\dim(V_L)$ rows of $U_L \ldots U_2 U_1$ are the $\{\phi_m^L\}_m$ scaling functions, whereas the rest of its rows return the $\{\psi_m^L\}, \{\psi_m^{L-1}\}, \ldots$ wavelets.

In the Fourier case, H would be diagonal. In the multiresolution case the situation is slightly more complicated since H consists of four distintict blocks:

$$H = \begin{pmatrix} H_{\Phi,\Phi} & H_{\Phi,\Psi} \\ H_{\Psi,\Phi} & H_{\Psi,\Psi} \end{pmatrix} = \begin{pmatrix} H_{1:d_L,1:d_L} & H_{1:d_L,d_L+1:n} \\ H_{d_L+1:n,1:d_L} & H_{d_L+1:n,d_L+1:n} \end{pmatrix},$$

with $d_L = \dim(V_L)$. Here $H_{\Phi,\Phi}$ is effectively A compressed to V_L, and is therefore dense. The structure of the

[1] To be more precise, Mallat's axioms imply that there is a *set* of mother wavelets and father wavelets from which we can build bases in this way. However, the vast majority of MRAs discussed in the literature only make recourse to a single mother wavelet and a single father wavelet.

other three matrices, however, reflects to what extent the MRA1 criterion is satisfied. In particular, the closer the wavelets are to being eigenfunctions, the better they can filter the space by smoothness, as defined by A. Below, we define multiresolution factorizable matrices as those for which this is perfectly satisfied, i.e., which have a factorization with $H_{\Phi,\Psi} = H_{\Psi,\Phi}^{\top} = 0$ and $H_{\Psi,\Psi}$ diagonal.

In the following, we relax the form of (12) somewhat by allowing each U_ℓ to fix *some* set $[n] \setminus S_\ell$ of $n - \dim(V_{\ell-1})$ coordinates rather than necessarily the last $n - \dim(V_{\ell-1})$ (as long as $S_0 \supseteq S_1 \supseteq \dots$). This also affects the order in which rows are eliminiated as wavelets, and the criterion for perfect factorizability now becomes $H \in \mathcal{H}^n_{S_L}$, where $\mathcal{H}^n_{S_L} = \{ H \in \mathbb{R}^{n \times n} \,|\, H_{i,j} = 0 \text{ unless } i = j \text{ or } i, j \in S_L \}$.

Definition 1 *Given an appropriate subset $\mathcal{O}$ of the group $\mathrm{SO}(n)$ of n–dimensional rotation matrices, a depth parameter $L \in \mathbb{N}$, and a sequence of integers $n = d_0 \geq d_1 \geq d_2 \geq \dots \geq d_L \geq 1$, a* ***Multiresolution Matrix Factorization (MMF)*** *of a symmetric matrix $A \in \mathbb{R}^{n \times n}$ over $\mathcal{O}$ is a factorization of the form*

$$A = U_1^{\top} U_2^{\top} \dots U_L^{\top} \, H \, U_L \dots U_2 U_1, \tag{13}$$

where each $U_\ell \in \mathcal{O}$ satisfies $[U_\ell]_{[n] \setminus S_{\ell-1}, \, [n] \setminus S_{\ell-1}} = I_{n-d_\ell}$ for some nested sequence of sets $[n] = S_0 \supseteq S_1 \supseteq \dots \supseteq S_L$ with $|S_\ell| = d_\ell$, and $H \in \mathcal{H}^n_{S_L}$.

Definition 2 *We say that a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is* ***fully multiresolution factorizable*** *over $\mathcal{O} \in \mathrm{SO}(n)$ with $(d_1, \dots, d_L)$ if it has a decomposition of the form described in Definition 1.*

The sequence $(d_1, \dots, d_L)$ may follow some predefined law, such as geometric decay, $d_\ell = \lceil n \eta^\ell \rceil$ or arithmetic decay, $d_\ell = n - \ell m$. The major difference between different types of MMFs, however, is in the definition of the set $\mathcal{O}$ of sparse rotations. In this regard we consider two alternatives: elementary and compound k'th order rotations.

Definition 3 *We say that $U \in \mathbb{R}^{n \times n}$ is an* ***elementary rotation of order k*** *(sometimes also called a k–point rotation) if it is an orthogonal matrix of the form*

$$U = I_{n-k} \oplus_{(i_1, \dots, i_k)} O \tag{14}$$

for some $\{i_1, \dots, i_k\} \subseteq [n]$ and $O \in \mathrm{SO}(k)$. The set of all such matrices we denote $\mathrm{SO}_k(n)$.

A k'th order elementary rotation is very local, since it only touches coordinates $\{i_1, \dots, i_k\}$, and leaves the rest invariant. The simplest case are second order rotations, which are of the form

$$U = U^{\theta}_{i,j} = \begin{pmatrix} \cdot & & & & \\ & c & & -s & \\ & & \cdot & & \\ & s & & c & \\ & & & & \cdot \end{pmatrix}, \qquad \begin{matrix} c = \cos\theta \\ s = \sin\theta, \end{matrix} \tag{15}$$

where the dots denote that apart from rows/columns i and j, $U^{\theta}_{i,j}$ is the identity, and θ is some angle in $[0, 2\pi)$. Such matrices are called Givens rotations, and they play an important role in numerical linear algebra. Indeed, Jacobi's algorithm for diagonalizing symmetric matrices (Jacobi, 1846), possibly the first matrix algorithm to have been invented, works precisely by constructing an MMF factorization over Givens rotations. Inspired by this connection, we will call any MMF with $\mathcal{O} = \mathrm{SO}_k(n)$ a ***k'th order Jacobi MMF***.

Definition 4 *We say that $U \in \mathbb{R}^{n \times n}$ is a* ***compound rotation of order k*** *if it is an orthogonal matrix of the form*

$$U = \oplus_{(i^1_1, \dots, i^1_{k_1})} O_1 \oplus_{(i^2_1, \dots, i^2_{k_2})} O_2 \dots \oplus_{(i^m_1, \dots, i^m_{k_m})} O_m \tag{16}$$

for some partition $\{i^1_1, \dots, i^1_{k_1}\} \mathbin{\dot\cup} \dots \mathbin{\dot\cup} \{i^m_1, \dots, i^m_{k_m}\}$ of $[n]$ with $k_1, \dots, k_m \leq k$, and some sequence of orthogonal matrices $O_1, \dots, O_m$ of the appropriate sizes. The set of all such matrices we denote $\mathrm{SO}^*_k(n)$.

Intuitively, compound rotations consist of many elementary rotations exectuted in parallel, and can consequently lead to much more compact factorizations.

4. Computing MMFs

Much like how low rank methods express matrices in terms of a small dictionary of vectors as in (1), MMF approximates A in the form

$$A^* = \sum_{i,j=1}^{d_L} \beta_{i,j} \, \phi^L_i \, \phi^{L\,\top}_j + \sum_{\ell=1}^{L} \sum_{i=1}^{d_\ell} \eta^\ell_i \, \psi^\ell_i \, \psi^{\ell\,\top}_i ,$$

where the $\eta^\ell_i = \langle \psi^\ell_i, A \psi^\ell_i \rangle$ wavelet frequencies are the diagonal elements of the $H_{\Psi,\Psi}$ block of H, whereas the $\beta_{i,j}$ coefficients are the entries of the $H_{\Phi,\Phi}$ block. Thus, given $\mathcal{O}$ and $(d_1, \dots, d_L)$, finding the best MMF factorization to a symmetric matrix A requires solving

$$\underset{\substack{[n] \supseteq S_1 \supseteq \dots \supseteq S_L \\ H \in \mathcal{H}^n_{S_L}; \; U_1, \dots, U_L \in \mathcal{O}}}{\text{minimize}} \quad \| \, A - U_1^{\top} \dots U_L^{\top} \, H \, U_L \dots U_1 \, \|.$$

Assuming that we measure error in the Frobenius norm, which is rotationally invariant, this is equivalent to

$$\underset{\substack{[n] \supseteq S_1 \supseteq \dots \supseteq S_L \\ U_1, \dots, U_L \in \mathcal{O}}}{\text{minimize}} \; \| \, U_L \dots U_1 \, A \, U_1^{\top} \dots U_L^{\top} \, \|^2_{\text{resi}} \, , \tag{17}$$

where $\| \;\; \|^2_{\text{resi}}$ is the squared "residual norm"

$$\| \, H \, \|^2_{\text{resi}} = \sum_{i \neq j \text{ and } (i,j) \neq S_L \times S_L} | \, H_{i,j} \, |^2 \, .$$

Defining $A_\ell = U_\ell \dots U_1 A U_1^{\top} \dots U_\ell$, intuitively, our objective is to find a series of sparse rotations

$$A \equiv A_0 \xrightarrow{U_1} A_1 \xrightarrow{U_2} \dots \xrightarrow{U_L} A_L \tag{18}$$

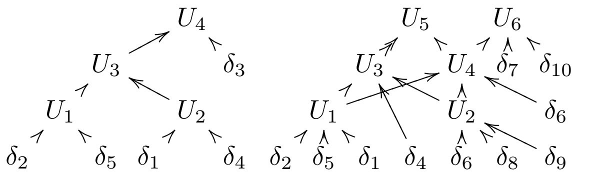

Figure 2. **Left:** Example of the tree induced by a second order Jacobi MMF of a six dimensional matrix. **Right:** Example of a Jacobi MMF with $k=3$ of a 10 dimensional matrix.

that bring A to a form as close to diagonal as possible. The following Proposition tells us that as soon as we designate a certain set $J_\ell := S_{\ell-1} \setminus S_\ell$ of rows/columns in A_ℓ wavelets, the ℓ_2-norm of these rows/columns (discounting the diagonal and those parts that fall outside the $S_{\ell-1} \times S_{\ell-1}$ active submatrix) is already committed to the final error.

Proposition 1 *Given an MMF as defined in Definition 1, the objective function of (17) is expressible as* $\sum_{\ell=1}^{L} \mathcal{E}_\ell$, *where* $\mathcal{E}_\ell = \|[A_\ell]_{J_\ell,J_\ell}\|^2_{\textit{off-diag}} + 2\,\|[A_\ell]_{J_\ell,S_\ell}\|^2_{\textit{Frob}}$, *and* $\|M\|^2_{\textit{off-diag}} := \sum_{i\neq j} |M_{i,j}|^2$.

The following algorithms for finding MMFs all follow the greedy approach suggested by this proposition of finding at each level a rotation U_ℓ that produces $d_\ell - d_{\ell-1}$ rows/columns that are as close to diagonal as possible, and then designating these as the level ℓ wavelets.

4.1. Jacobi MMFs

In Jacobi MMFs, where each U_ℓ is an $I_{n-k} \oplus_{(i_1,\ldots,i_k)} O$ elementary rotation, we set $(d_1,\ldots,d_L)$ so as to split off some constant number $m<k$ of wavelets at each level. For simplicity, for now we take $m=1$. Furthermore, we make the natural assumption that this wavelet is one of the rows involved in the rotation, w.l.o.g. $J_\ell = \{i_k\}$.

Proposition 2 *If* $U_\ell = I_{n-k} \oplus_{\mathrm{I}} O$ *with* $\mathrm{I} = (i_1,\ldots,i_k)$ *and* $J_\ell = \{i_k\}$, *then the contribution of level* ℓ *to the MMF approximation error is*

$$\mathcal{E}_\ell = \mathcal{E}^O_{\mathrm{I}} = 2\sum_{p=1}^{k-1} [O[A_{\ell-1}]_{\mathrm{I},\mathrm{I}} O^\top]^2_{k,p} + 2\,[OBO^\top]_{k,k},$$

where $B = [A_{\ell-1}]_{\mathrm{I},S_\ell}\,([A_{\ell-1}]_{\mathrm{I},S_\ell})^\top$. (19)

Corollary 1 *In the special case of* $k=2$ *and* $\mathrm{I}_\ell = (i,j)$,

$$\mathcal{E}_\ell = \mathcal{E}^O_{(i,j)} = 2\,[O[A_{\ell-1}]_{(i,j),(i,j)} O^\top]^2_{2,1} + 2\,[OBO^\top]_{k,k}$$

with $B = [A_{\ell-1}]_{(i,j),S_\ell}\,([A_{\ell-1}]_{(i,j),S_\ell})^\top$. (20)

According to the greedy strategy, at each level ℓ, the I index tuple and O rotation must be chosen so as to minimize (19). The resulting algorithm is given in Algorithm 1, where $A_L \!\downarrow_{\mathcal{H}^n_{S_L}}$ stands for zeroing out all the entries of A_ℓ except those on the diagonal and in the $S_L \times S_L$ block.

Algorithm 1 GreedyJacobi: computing the Jacobi MMF of A with $d_\ell = n-\ell$.

```
Input: $k$, $L$, and a symmetric matrix $A_0 = A \in \mathbb{R}^{n\times n}$
set $S_0 \leftarrow [n]$
for ($\ell = 1$ to $L$){
  foreach $\mathrm{I} = (i_1,\ldots,i_k) \in (S_{\ell-1})^k$ with $i_1 < \ldots < i_k$
    compute $\mathcal{E}_{\mathrm{I}} = \min_{O\in\mathrm{SO}(k)} \mathcal{E}^O_{\mathrm{I}}$   (as defined in (19))
  set $\mathrm{I}_\ell \leftarrow \arg\min_{\mathrm{I}} \mathcal{E}_{\mathrm{I}}$
  set $O_\ell \leftarrow \arg\min_{O\in\mathrm{SO}(k)} \mathcal{E}^O_{\mathrm{I}_\ell}$
  set $U_\ell \leftarrow I_{n-k} \oplus_{\mathrm{I}_\ell} O_\ell$
  set $S_\ell \leftarrow S_{\ell-1} \setminus \{i_k\}$
  set $A_\ell \leftarrow U_\ell A_{\ell-1} U_\ell^\top$
}
Output: $U_1,\ldots,U_L$ and $H = A_L\!\downarrow_{\mathcal{H}^n_{S_L}}$
```

When $k=2$, the rotations $U_1,\ldots,U_L$ form a binary tree, in which each U_ℓ takes two scaling functions from level $\ell-1$ and passes on a single linear combination of them to the next level (Figure 2). In general, the more similar two rows $A_{i,:}$ and $A_{j,:}$ are to each other, the smaller we can make (21) by choosing the approriate O. In graphs, for example, where in some metric the entries in row i measure the similarity of vertex i to all the other vertices, this means that Algorithm 1 will tend to pick pairs of adjacent or nearby vertices and then produce scaling functions that represent linear combinations of those vertices. Thus, second order MMFs effectively perform a hierarchical clustering on the rows/columns of A. Uncovering this sort of hierarchical sructure is one of the goals of MMF analysis.

The idea of constructing wavelets by forming a tree of Givens rotations was first intruduced under the name “Treelets” by Lee et al. (2008). Their work, however, does not make a connection to matrix factorization. In particular, instead of minimizing the contribution of each rotation to the matrix approximation error, the Treelets algorithm chooses I and O so as to zero out the largest off-diagonal entry of $A_{\ell-1}$. This pivoting rule is the same as in Jacobi's classical algorithm, so if one of the two indices $\{i,j\}$ was not always eliminated from the active set, we would eventually diagonalize A to arbitrary precision.

Jacobi MMFs with $k \geq 3$ are even more interesting because they correspond to a lattice in which each U_ℓ now has k children and $k-1$ parents (Figure 2). In the $k=2$ case the supports of any two wavelets ψ^ℓ_1 and $\psi^{\ell'}_1$ are either disjoint or one is contained in the other. In contrast, for $k \geq 3$, a single original coordinate, such as δ_6 in Figure 2 can contribute to multiple wavelets (ψ^5_1 and ψ^6_1, for example) with different weights, determined by all the orthogonal matrices along the corresponding paths in the lattice. Thus, higher order MMFs are more subtle than just a single hierarchical clustering: by building a lattice of subspaces they capture a softer notion of hierarchy, and can

Algorithm 2 GREEDYPARALLEL: computing the binary parallel MMF of A with $d_\ell = \lceil n2^{-\ell} \rceil$.

```
Input: $L$ and a symmetric matrix $A = A_0 \in \mathbb{R}^{n \times n}$
set $S_0 \leftarrow [n]$
for ($\ell = 1$ to $L$){
  set $p \leftarrow \lfloor |S_{\ell-1}|/2 \rfloor$
  compute $W_{i,j} = W_{j,i}$ as defined in (22) $\forall i, j \in S_{\ell-1}$
  find the matching $\{(i_1, j_1), \ldots, (i_p, j_p)\}$
        minimizing $\sum_{r=1}^{p} W_{i_r, j_r}$
  for ($r = 1$ to $p$) set $O_r \leftarrow \arg\min_{O \in \mathrm{SO}(2)} \mathcal{E}^{O}_{(i_r, j_r)}$
  set $U_\ell \leftarrow \oplus_{(i_1,j_1)} O_1 \oplus_{(i_2,j_2)} O_2 \oplus \ldots \oplus_{(i_p,j_p)} O_p$
  set $S_\ell \leftarrow S_{\ell-1} \setminus \{i_1, \ldots, i_p\}$
  set $A_\ell \leftarrow U_\ell \, A_{\ell-1} \, U_\ell^\top$
}
Output: $U_1, \ldots, U_L$ and $H = A_L \downarrow_{\mathcal{H}^n_{S_L}}$
```

uncover multiple overlapping hierarchical structures in A.

4.2. Parallel MMFs

Since MMFs exploit hierarchical cluster-of-clusters type structure in matrices, towards the bottom of the hierarchy one expects to find rotations that act locally, within small subclusters, and thus do not interact with each other. By combining these independent rotations into a single compound rotation, parallel MMFs yield factorizations that are not only more compact, but also more interpretable in terms of resolving A at a small number of distinct scales. Once again, we assume that it is the last coordinate in each $(i^1_1, \ldots, i^1_{k_1}) \ldots (i^m_1, \ldots, i^m_{k_m})$ block that gives rise to a wavelet, therefore d_ℓ decays by a constant factor of approximately $(k-1)/k$ at each level.

Proposition 3 *If U_ℓ is a compound rotation of the form $U_\ell = \oplus_{\mathrm{I}_1} O_1 \ldots \oplus_{\mathrm{I}_m} O_m$ for some partition $\mathrm{I}_1 \cup \ldots \cup \mathrm{I}_m$ of $[n]$ with $k_1, \ldots, k_m \leq k$, and some sequence of orthogonal matrices $O_1, \ldots, O_m$, then level ℓ's contribution to the MMF error obeys*

$$\mathcal{E}_\ell \leq 2 \sum_{j=1}^{m} \left[\sum_{p=1}^{k_j - 1} [O_j [A_{\ell-1}]_{\mathrm{I}_j, \mathrm{I}_j} O_j^\top]^2_{k_j, p} + [O_j B_j O_j^\top]_{k_j, k_j} \right], \tag{21}$$

where $B_j = [A_{\ell-1}]_{\mathrm{I}_j, S_{\ell-1} \setminus \mathrm{I}_j} \, ([A_{\ell-1}]_{\mathrm{I}_j, S_{\ell-1} \setminus \mathrm{I}_j})^\top$.

The reason that (21), in contrast to (19), only provides an upper bound on $\mathcal{E}_\ell$ is that it double counts the contribution of the matrix elements $\{[A_\ell]_{k_j, k_{j'}}\}_{j,j'=1}^{m}$ at the intersection of pairs of wavelet rows/columns. Accounting for these elements explicitly would introduce interactions between the O_j rotations, leading to a difficult optimization problem. Therefore, both for finding the optimal partition $\mathrm{I}_1 \cup \ldots \cup \mathrm{I}_m$ and for finding the optimal $O_1, \ldots, O_m$ rotations, we use the right hand side of (21) as a proxy for $\mathcal{E}_\ell$.

Once again, the binary ($k = 2$) case is the simplest, since optimizing $\mathrm{I}_1 \cup \ldots \cup \mathrm{I}_m$ then reduces to finding a minimal cost matching amongst the indices in the active set $S_{\ell-1}$ with cost matrix

$$W_{i,j} = 2 \min_{O \in \mathrm{SO}(2)} \left[[O [A_{\ell-1}]_{(i,j),(i,j)} O^\top]^2_{2,1} + [OBO^\top]_{k,k} \right], \tag{22}$$

where $B = [A_{\ell-1}]_{(i,j), S_{\ell-1} \setminus \{i,j\}} \, ([A_{\ell-1}]_{(i,j), S_{\ell-1} \setminus \{i,j\}})^\top$. An exact solution to this optimization problem can be found in time $O(|S_{\ell-1}|^3)$ using a modern weighted version of the famous "Blossom algorithm" by Edmonds (1965). However, it is also known that the simple greedy strategy of setting $(i_1, j_1) = \arg\min_{i,j \in S_{\ell-1}} W_{i,j}$, then $(i_2, j_2) = \arg\min_{i,j \in S_{\ell-1} \setminus \{i_1, j_1\}} W_{i,j}$, etc., yields a 2–approximation in linear time. In general, the most expensive component of MMF factorizations is forming the B matrices (which naïvely takes $O(n^k)$ time), however, in practice techinques like locality sensitive hashing allow this (as well as the entire algorithm) to run in time close to linear in n. We remark that the fast Haar transform is nothing but a binary MMF, and the Cooley–Tukey FFT is a degenerate MMF (where $d_0 = \ldots = d_L$) of a complex valued matrix.

4.3. Computational details

Problems of the form $\min_{O \in \mathrm{SO}(k)} \| OBO^\top C \|$, called Procrustes problems, generally have easy, $O(k^3)$ time closed form solutions. Unfortunately, both (19) and (21) involve mixed linear/quadratic versions of this problem, which are much more challenging. However, the following result shows that in the $k = 2$ case this may be reduced to solving a simple trigonometric equation.

Proposition 4 *Let $A \in \mathbb{R}^{2 \times 2}$ be diagonal, $B \in \mathbb{R}^{2 \times 2}$ symmetric and $O = \left(\begin{smallmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{smallmatrix}\right)$. Set $a = (A_{1,1} - A_{2,2})^2/4$, $b = B_{1,2}$, $c = (B_{2,2} - B_{1,1})/2$, $e = \sqrt{b^2 + c^2}$, $\theta = 2\alpha$ and $\omega = \arctan(c/b)$. Then if α minimizes $([OAO^\top]_{2,1})^2 + [OBO^\top]_{2,2}$, then θ satisfies*

$$(a/e)\sin(2\theta) + \sin(\theta + \omega + \pi/2) = 0. \tag{23}$$

Putting A and B in the diagonal form required by this proposition is easy. While (23) is still not an explicit expression for α, it is trivial to solve with iterative methods.

5. Theoretical analysis

MMFs satisfy properties MRA2 and MRA3 of Section 3.1 by construction. Showing that they also satisfy MRA1 requires, roughly, to prove that the smoother a function is, the smaller its high frequency wavelet coefficients are. For this purpose the usual notion of smoothness with respect to a metric d is Hölder continuity, defined

$$|f(x) - f(y)| \leq c_H \, d(x,y)^\alpha \qquad \forall x, y \in X,$$

with c_H and $\alpha > 0$ constant. In classical wavelet analysis one proves that the wavelet coefficients of (c_H, α)–Hölder

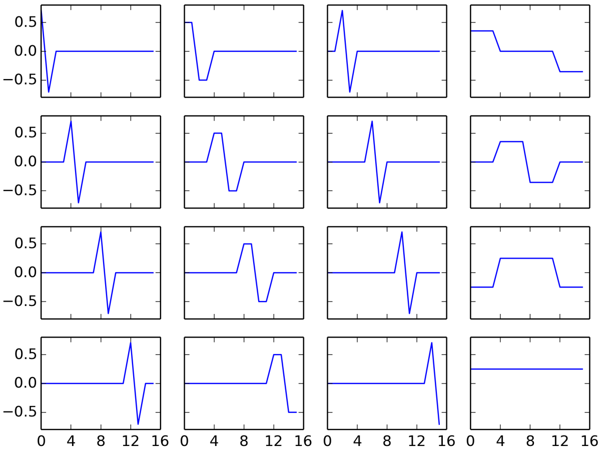

Figure 3. The MMF wavelets on a cycle graph on 16 vertices recover the Haar wavelet system.

functions decay at a certain rate, for example, $|\langle f, \psi_\ell^m \rangle| \leq c' \ell^{\alpha+\beta}$ for some β and c' (Daubechies, 1992).

As we have seen, MMFs are driven by the similarity between the rows/columns of the matrix A. Therefore, relaxing the requirement that d must be a metric, we now take

$$d(i,j) = |\langle A_{i,:}, A_{j,:} \rangle|^{-1}. \qquad (24)$$

One must also make some assumptions about the structure of the underlying space, classically that X is a so-called space of homogeneous type (Deng & Han, 2009), which means that for some constant c_{hom},

$$\text{Vol}(B(x, 2r)) \leq c_{\text{hom}} \, \text{Vol}(B(x,r)) \quad \forall x \in X, \, \forall r > 0.$$

To capture the analogous structural property for matrices, we introduce a concept with connections to the R.I.P condition in compressed sensing (Candes & Tao, 2005).

Definition 5 *We say that a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is Λ–**rank homogeneous** up to order $\overline{K}$, if for any $S \subseteq [n]$ of size at most $\overline{K}$, letting $Q = A_{S,:}A_{:,S}$, setting D to be the diagonal matrix with $D_{i,i} = \|Q_{i,:}\|_1$, and $\tilde{Q} = D^{-1/2}QD^{-1/2}$, the $\lambda_1, \ldots, \lambda_{|S|}$ eigenvalues of $\tilde{Q}$ satisfy $\Lambda < |\lambda_i| < 1 - \Lambda$, and furthermore $c_T^{-1} \leq D_{i,i} \leq c_T$ for some constant c_T.*

Recall that the spectrum of the normalized adjacency matrix of a graph is bounded in $[-1, 1]$ (Chung, 1997). Definition 5 asserts that if we form a graph with vertex set S and edge weights $\langle A_{i,:}, A_{j,:} \rangle$, its eigenvalues in absolute value are bounded away from both 0 and 1. Definition 5 then roughly corresponds to asserting that A does not have clusters of rows that are either almost identical (an incoherence condition) or completely unrelated. This allows us to now state the matrix analog of the Hölder condition.

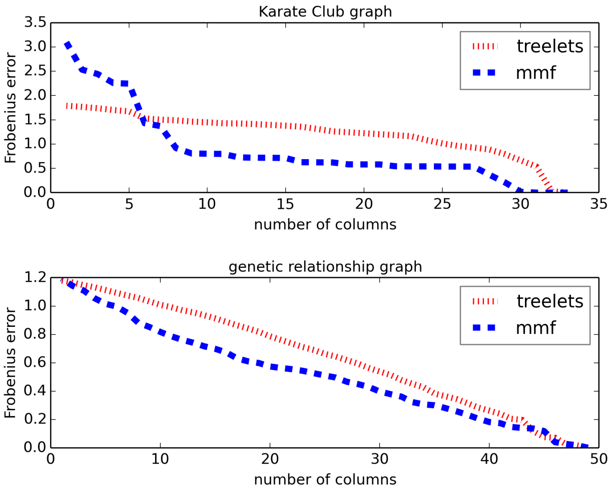

Figure 4. Comparison with the Treelets algorithm. Zachary's Karate Club graph (top) and a matrix describing the estimated additive genetic relationship between 50 individuals (bottom).

Theorem 1 *Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix that is Λ–rank homogeneous up to order $\overline{K}$ and has an MMF factorization $A = U_1^\top \ldots U_L^\top H U_L \ldots U_1$. Assume ψ_m^ℓ is a wavelet in this factorization arising from row i of $A_{\ell-1}$ supported on a set S of size $K \leq \overline{K}$ and that $\|H_{i,:}\|^2 \leq \epsilon$. Then if $f\colon [n] \to \mathbb{R}$ is $(c_H, 1/2)$–Hölder with respect to (24), then*

$$|\langle f, \psi_m^\ell \rangle| \leq c_T \sqrt{c_H c_\Lambda} \, \epsilon^{1/2} K \qquad (25)$$

with $c_\Lambda = 4/(1-(1-2\Lambda)^2)$.

Here ϵ is closely related to the MMF approximation error and is therefore expected to be small. Eq. (25) then says that, as we expect, if f is smooth, then its "high frequency" local wavelet coefficents (low K and ℓ) will be small.

6. Experiments

In a toy example we consider the diffusion kernel of the Laplacian, T, of a cycle graph (C_n) on $n = 16$ vertices. Applying Algorithm 2, we compute the binary parallel MMF of T up to depth $L = 5$. We find that the sequence of MMF rotations reconstructs the Haar wavelets (Figure 3). In fact, similar results can be obtained for any cycle graph, except that for large n the longest wavelength wavelets cannot be fully reconstructed due to numerical precision issues.

We also evaluate the performance of GREEDYJACOBI by comparing it with Treelets on two small matrices. Note that in the greedy setting MMF removes one dimension at a time, similarly to the Treelets algorithm, and thus in both algorithms the off-diagonal part of the rows/columns designated as wavelets contributes to the error. The first dataset is the well-known Zachary's Karate Club (Zachary, 1977) social network ($N = 34, E = 78$) for which we set A to be the heat kernel. The second one is constructed using

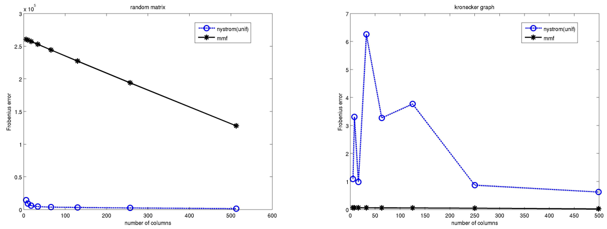

Figure 5. Frobenius norm error of the MMF and Nyström methods on a random vs. a structured (Kronecker) matrix.

simulated data from the family pedigrees in (Crossett et al., 2013), 5 families were randomly selected, and 10 individuals from the 15 related individuals were randomly selected independently in each family. The resulting relationship matrix represents the estimated kinship coefficient and is calculated via the GCTA software of Yang et al. (2011). Figure 4 shows that GREEDYJACOBI outperforms Treelets for a wide range of compression ratios.

6.1. Comparison to Other Factorization Methods

To verify that MMF produces meaningful factorizations, we measure the approximation error of factoring two 1024×1024 matrices: a matrix consisting of i.i.d. normal random variables and a Kronecker graph, K_1^k, of order $k = 5$, where K_1 is a 2×2 seed matrix (Leskovec et al., 2010). Figure 5 shows that MMF performs sub-optimally when the matrix lacks an underlying multiresolution structure. However, on matrices with multilevel structure MMF systematically outperforms other algorithms.[2]

In order to evaluate MMF for matrix compression, we use several large datasets: GR (arXiv General Relativity collaboration graph, $N = 5242$) (Leskovec et al., 2007), Dexter (bag of words, $N = 2000$) (Asuncion & Newman, 2012), and HEP (arXiv High Energy Physics collaboration graph, $N = 9877$, see Supplement). The first two are normalized graph Laplacians of real-world networks and the third one is a linear kernel matrix constructed from a 20000-feature dataset. By virtue of its design, MMF operates only on symmetric matrices, so we compare its performance to other algorithms designed specifically for symmetric matrices. Figure 6 compares the approximation error of MMF and the Nyström-based family of randomized algorithms. The Nyström method has several extensions differing in sampling technique (uniform at random without replacement, non-uniform leverage score probabilities, Gaussian or SRFT mixtures of the columns). The MMF approximation error is measured by taking the cumulative l_2 norm of the rows/columns that are designated wavelets at each iteration of the algorithm (Proposition 1). For the Nyström-based algorithms the compression error is a function of the number of columns sampled (and possibly the desired rank of the approximation leading to a distinction between the rank-restricted and unconstrained rank versions of the method) and is defined as $||A - CW^{\dagger}C^T||_{\text{Frob}}$ or $||A - CW_k^{\dagger}C^T||_{\text{Frob}}$(Gittens & Mahoney, 2013). Similarly, at every level of the MMF compression, the approximation error is a function of $|S_\ell|$, the number of dimensions that have not yet been eliminated.

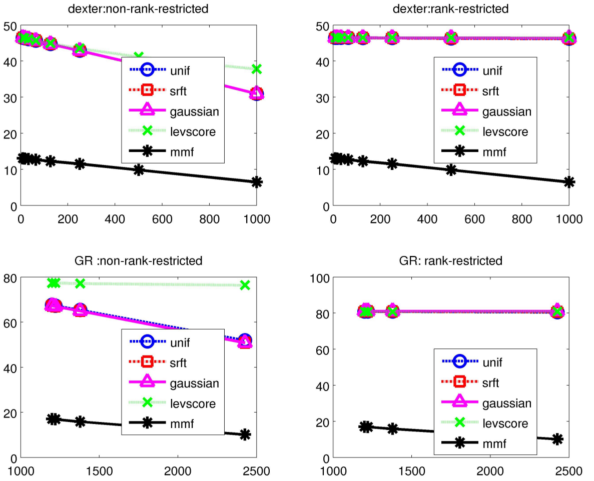

Figure 6. Comparison of the Frobenius norm error of the binary parallel MMF and Nyström approximations on two real datasets. In the rank restricted cases $k = 20$ for GR and $k = 8$ for Dexter.

These results convincingly show that, despite similar wall clock times, MMF factorization characteristically outperforms standard techniques when the underlying matrix has multiscale structure. We are working on more extensive experiments that go beyond the scope of this paper.

7. Conclusions

The interplay between the geometry of a space X and the structure of function spaces on X is a classical theme in Harmonic Analysis (Coifman & Maggioni, 2006). As an instance of this connection, this paper developed the matrix factorization analog of multiresolution analysis on finite sets. The resulting factorizations, on the one hand, provide a natural way to define multiresolution on graphs, correlated sets of random variables, and so on. On the other hand, they lead to new classes of structured matrices and new matrix compression algorithms.

The present work could only explore a small subset of the potential applications of MMFs from matrix completion via sparse approximation to community detection in networks. In general, what classes of naturally occurring matrices exhibit MMF structure is itself an important question. From an algorithmic point of view, devising fast random-

[2] Note that compressing a matrix to size $d \times d$ means something slightly different for Nyström methods and MMF: in the latter case, in addition to the $d \times d$ core, we also preserve a sequence of $n - d$ wavelet frequencies.

ized version of MMFs will be critical. Finally, from the theoretical point of view, one of the biggest challenges is to relate the new concepts of multiresolution factorizable and Λ–rank homogenous matrices to the existing body of work in harmonic analysis, algebra and compressed sensing.

References

Achlioptas, D. and McSherry, F. Fast computation of low-rank matrix approximations. *Journal of the ACM*, 54(2):9, April 2007.

Asuncion, A. and Newman, D. J. Dexter dataset. *UCI Machine Learning Repository*, 2012.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9 (6):717–772, April 2009.

Candes, E.J. and Tao, T. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, Dec 2005.

Chung, F. R. K. *Spectral Graph Theory*. Number 92 in Regional Conference Series in Mathematics. American Mathematical Society, 1997.

Coifman, R. R. and Maggioni, M. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, 2006.

Crossett, A., Lee, A. B., Klei, L., Devlin, B., and Roeder, K. Refining genetically inferred relationships using treelet covariance smoothing. *The Annals of Applied Statistics*, 7(2):669–690, 2013.

Daubechies, Ingrid. *Ten Lectures on Wavelets (CBMS-NSF Regional Conference Series in Applied Mathematics)*. 1992. ISBN 0898712742.

Deng, Donggao and Han, Yongsheng. *Harmonic Analysis on Spaces of Homogeneous Type*. Springer, 2009.

Dhillon, I. S., Guan, Y., and Kulis, B. Weighted graph cuts without eigenvectors: a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.

Drineas, P., Kannan, R., and Mahoney, M. W. Fast monte carlo algorithms for matrices I–III. *SIAM Journal on Computing*, 36 (1):158–183, 2006.

Edmonds, J. Paths, trees, and flowers. *Canad. J. Math.*, 17:449–467, 1965.

Gavish, M., Nadler, B., and Coifman, R. R. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *International Conference on Machine Learning (ICML)*, pp. 367–374, 2010.

Gittens, A. and Mahoney, M. W. Revisiting the Nyström method for improved large-scale machine learning. In *International Conference on Machine Learning (ICML)*, pp. 567–575, 2013.

Greengard, L. and Rokhlin, V. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73:325–348, 1987.

Halko, N., Martinsson, P. G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *Computing*, 53(2): 1–74, 2009.

Hammond, D. K., Vandergheynst, P., and Gribonval, R. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

Jacobi, C. J. G. Über ein leichtes verfahren, die in der theorie der säkularstörungen vorkommenden gleichungen numerisch aufzulösen. *Journal für reine und angewandte Mathematik*, 30:51–95, 1846.

Jenatton, R., Obozinski, G., and Bach, F. Structured sparse principal component analysis. In *Proceedings of AISTATS*, volume 9, 2010.

Kumar, S., Mohri, M., and Talwalkar, A. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.

Lee, A. B., Nadler, B., and Wasserman, L. Treelets—an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics*, 2(2):435–471, 2008.

Leskovec, J., Kleinberg, J., and Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM Transactions of Knowledge Discovery from Data (TKDD)*, 1, 2007.

Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. Kronecker graphs : an approach to modeling networks. *JMLR*, 11:985–1042, 2010.

Livne, O. E. and Brandt, A. Lean algebraic multigrid (LAMG): fast graph Laplacian linear solver. *http://arxiv.org/abs/1108.1310*, 2011.

Mahoney, M. W. Randomized algorithms for matrices and data. *Foundations and trends in machine learning*, 3, 2011.

Mallat, S. G. A Theory for Multiresolution Signal Decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

Savas, B. and Dhillon, I. S. Clustered low rank approximation of graphs in information science applications. In *SDM*, pp. 164–175, 2011.

Williams, C. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Neural Information Processing Systems (NIPS)*, volume 13, 2001.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.*, 88(1):76–82, 2011.

Zachary, W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

8. Supplement to “Multiresolution Matrix Factorization” (ICML 2014 submission)

Proof of Proposition 1. By the nestedness of $S_0 \supseteq S_1 \supseteq \ldots \supseteq S_L$, for some sequence of permutation matrices $\Pi_1, \ldots, \Pi_L$, H decomposes recursively as

$$[H]_{S_\ell,S_\ell} = \Pi_\ell \begin{pmatrix} [H]_{S_{\ell+1},S_{\ell+1}} & [H]_{S_{\ell+1},\,J_{\ell+1}} \\ [H]_{J_{\ell+1},\,S_{\ell+1}} & [H]_{J_{\ell+1},\,J_{\ell+1}} \end{pmatrix} \Pi_\ell^\top$$

Unwrapping this recursion tells us that $\|H\|^2_{\mathrm{resi}}$ is equal to

$$\sum_{\ell=1}^{L} \Big[\|[H]_{J_\ell,\,S_\ell}\|^2_{\text{Frob}} + \|[H]_{S_\ell,\,J_\ell}\|^2_{\text{Frob}} + \|[H]_{J_\ell,\,J_\ell}\|^2_{\text{off-diag}} \Big].$$

However, since the rotations $U_{\ell+1}, \ldots, U_L$ leave $\mathrm{span}(\{\, e_i \mid i \in [n] \setminus S_\ell \,\})$ invariant,

$$\|[A_\ell]_{J_\ell,\,S_\ell}\|^2_{\text{Frob}} = \|[A_{\ell+1}]_{J_\ell,\,S_\ell}\|^2_{\text{Frob}} = \ldots = \\ = \|[A_L]_{J_\ell,\,S_\ell}\|^2_{\text{Frob}} == \|[H]_{J_\ell,\,S_\ell}\|^2_{\text{Frob}}.$$

By symmetry, $\|[H]_{S_\ell,\,J_\ell}\|^2_{\text{Frob}} = \|[H]_{J_\ell,\,S_\ell}\|^2_{\text{Frob}}$. Similarly, $\|[A_\ell]_{J_\ell,\,J_\ell}\|^2_{\text{off-diag}} = \ldots = \|[H]_{J_\ell,\,J_\ell}\|^2_{\text{off-diag}}$. ■

Proof of Proposition 2. Since $J = \{i_k\}$, by Proposition 1

$$\mathcal{E}_\ell = 2 \sum_{p=1}^{k-1} [U_\ell A_{\ell-1} U_\ell^\top]^2_{i_k,i_p} + 2\, \|[U_\ell A_{\ell-1} U_\ell^\top]_{i_k,S_\ell}\, \|^2.$$

The first term can be written $2\sum_{p=1}^{k-1} [O[A_{\ell-1}]_{\mathrm{I},\mathrm{I}} O^\top]^2_{k,p}$, while the second term is

$$2\, \|\, [O[A_{\ell-1}]_{\mathrm{I},S_\ell} [U_\ell]^\top_{S_\ell,S_\ell}]_{k,:}\, \|^2 = \\ 2\, [O[A_{\ell-1}]_{\mathrm{I},S_\ell} [U_\ell]^\top_{S_\ell,S_\ell} [U_\ell]_{S_\ell,S_\ell} [A_{\ell-1}]^\top_{\mathrm{I},S_\ell} O^\top]_{k,k} = \\ 2\, [O[A_{\ell-1}]_{\mathrm{I},S_\ell} [A_{\ell-1}]^\top_{\mathrm{I},S_\ell} O^\top]_{k,k} = 2\, [OBO^\top]_{k,k}$$

■

Proof of Proposition 3. Analogous to the proof of Proposition 2, but summed over each $\mathrm{I}_1 \times \mathrm{I}_1, \ldots, \mathrm{I}_m \times \mathrm{I}_m$ block.

Proof of Proposition 4. We want to minimize

$$\phi(\alpha) = \left(\left[O_\alpha \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} O_\alpha^\top \right]_{2,1} \right)^2 \\ + \left[O_\alpha \begin{pmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{pmatrix} O_\alpha^\top \right]_{2,2}.$$

Expanding, we get

$$\phi(\alpha) = ((A_1 - A_2) \sin\alpha \cos\alpha)^2 + B_{1,1} (\sin\alpha)^2 + \\ 2B_{1,2} \sin\alpha \cos\alpha + B_{2,2} (\cos\alpha)^2 = \\ \Big(\frac{A_1 - A_2'}{2} \Big)^2 (\sin(2\alpha'))^2 + \\ B_{1,2} \sin(2\alpha) + (\sin\alpha')^2 B_{1,1} + (\cos\alpha')^2 B_{2,2}.$$

Rewriting the second two terms as

$$\frac{((\sin\alpha)^2 + (\cos\alpha')^2)(B_{1,1} + B_{2,2})}{2} + \\ \frac{((\sin\alpha)^2 - (\cos\alpha')^2)(B_{1,1} - B_{2,2})}{2}$$

gives

$$\phi(\alpha) = \Big(\frac{A_1 - A_2}{2} \Big)^2 (\sin(2\alpha))^2 + B_{1,2} \sin(2\alpha) + \\ \frac{B_{1,1} + B_{2,2}}{2} + \frac{B_{2,2} - B_{1,1}}{2} \cos(2\alpha).$$

Introducing $d = (B_{2,2} - B_{1,1})/2$ and the other variables a, b, c, e and θ gives the new objective function

$$\psi(\theta) = a (\sin\theta)^2 + b \sin\theta + c \cos\theta + d.$$

Setting the derivative with respect to θ zero,

$$2a \sin\theta \cos\theta + b \cos\theta - c \sin\theta = 0.$$

Again using $\sin(2x) = 2 \sin x \cos x$,

$$a \sin(2\theta) + b \cos\theta - c \sin\theta = 0.$$

Now letting $e = \sqrt{b^2 + c^2}$ and $\omega = \arctan(c/b)$

$$a \sin(2\theta) + e(\cos\omega \cos\theta - \sin\omega \sin\theta) = 0.$$

Using $\cos(x + y) = \cos x \cos y - \sin x \sin y$,

$$(a/e) \sin(2\theta) + \cos(\theta + \omega) = 0,$$

which is finally equivalent to (23). ■

Proof of Theorem 1. Let ψ be a specific wavelet ψ^ℓ_m, with support $S = \{s_1, \ldots, s_K\} = \mathrm{supp}(\psi) \subseteq [n]$; f_S and ψ_S be the restriction of f and ψ to S regarded as a vectors; and Q, D and $\tilde{Q}$ be defined as in Definition 5. The Hölder property then gives

$$f_S^\top \tilde{L} f_S = \sum_{i,j=1}^{K} \tilde{Q}_{i,j} (f(s_i) - f(s_j))^2 \leq \\ \leq \sum_{i,j=1}^{K} c_T\, Q_{i,j} (f(s_i) - f(s_j))^2 \leq c_T c_H K^2, \quad (26)$$

where $\tilde{L} = I - \tilde{Q}$ is the normalized Laplacian. At the same time, if ψ^ℓ_m comes from row/column i of A_ℓ, then by (11), $[A_\ell]_{:,i} = U_\ell \ldots U_1 A\, \psi$, and therefore

$$\psi_S^\top \tilde{Q} \psi_S \leq c_T\, \psi_S^\top Q \psi_S \leq c_T\, \psi_S^\top A_{S,:} A_{:,S} \psi_S = \\ = c_T \|A\psi\|^2 = c_T\, \|[A_\ell]_{:,i}\|^2 = c_T\, \|H_{:,i}\|^2 \leq c_T \epsilon \quad (27)$$

Clearly, $\tilde{Q}$ and $\tilde{L}$ share the same normalized eigenbasis $\{v_1, \ldots, v_n\}$. Letting $\lambda_1, \ldots, \lambda_K$ be the corresponding eigenvalues, $f_i = \langle f_S, v_i \rangle$ and $\psi_i = \langle \psi_S, v_i \rangle$ and taking any $\gamma > 0$

$$\sum_{i=1}^{K} \Big(\sqrt{\gamma \lambda_i}\, \psi_i - \frac{1}{\sqrt{\gamma \lambda_i}}\, f_i \Big)^2 \geq 0, \tag{28}$$

which implies

$$\langle f, \psi \rangle = \langle f_S, \psi_S \rangle \leq \frac{1}{2} \left[\gamma\, \psi_S^\top \tilde{Q} \psi_S + \gamma^{1/2} f_S^\top \tilde{Q}^{-1} f_S \right].$$

The first term on the r.h.s of this inequality is bounded by (27), while for any $c_\Lambda \geq 4/(1-(1-2\Lambda)^2)$, by (26),

$$f_S^\top \tilde{Q}^{-1} f_S = \sum_{i=1}^{K} \frac{1}{\lambda_i}\, f_i^2 \leq c_\Lambda \sum_{i=1}^{K} (1-\lambda_i) f_i^2 = \\ = c_\Lambda f_S^\top \tilde{L} f_S \leq c_T c_H c_\Lambda K^2$$

giving $\langle f, \psi \rangle \leq c_T(\gamma \epsilon + \gamma^{-1} c_H c_\Lambda K^2)$. Optimizing this for γ yields $\langle f, \psi \rangle \leq c_T \sqrt{c_H c_\Lambda}\, \epsilon^{1/2} K$. By flipping the $-$ sign in (28) to $+$, a similar lower bound can be derived for $-\langle f, \psi \rangle$. ■

9. Additional experimental results

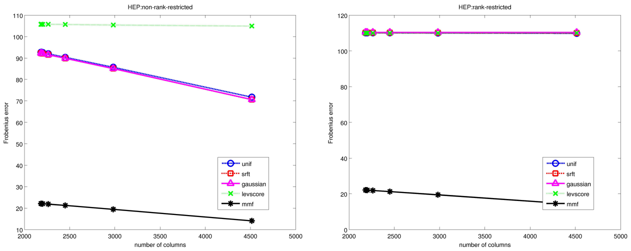

Figure 7. Comparison of the Frobenius norm error of the binary parallel MMF and Nyström approximations on the HEP dataset in the non rank-restricted case and the $k = 60$ rank restricted case.