Multiresolution Matrix Factorization

Risi Kondor, The University of Chicago

Nedelina Teneva
UChicago

Vikas K Garg
TTI-C, MIT
Multiresolution Machine Learning

Risi Kondor, The University of Chicago

Nedelina Teneva
UChicago

Vikas K Garg
TTI-C, MIT
What is machine learning?

\[
\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}
\]

\[
\downarrow
\]

\[
f : x \mapsto y
\]
Machine learning ~2005

\[
\hat{f} = \arg\min_{f \in \mathcal{F}} \left[ \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i) + \lambda \| f \|_2^2 \right]
\]

General framework (regularized risk minimization) that encompasses

- Kernel methods, including Gaussian processes and SVMs etc
- Most of Bayesian statistics
- Boosting, etc.

[Vapnik & Chervnonenakis 1971–]
Data: $\{(x, y)\}^{m}_{i=1}$

Features: $\{(\phi_1(x_i), \ldots, \phi_n(x_i), y)\}^{m}_{i=1}$

\[ \hat{f}: x \mapsto y \]
\[ \hat{f} = \arg\min_{f \in \mathcal{F}} \left[ \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i) + \lambda \|f\|_1 \right] \]

The $\ell_1$–norm induces **sparsity**. Crucially, minimizing $\|f\|_1$ is almost as good as minimizing $\|f\|_0$. This lead to an explosion of activity related to

- The Lasso [Tibshirani, 1996], Basis Pursuit [Donoho, 1998], Elastic Net [Hui & Hastie, 2005], …
- Compressed Sensing [Donoho, 2004] [Candés et al. 2005–] [Osher]
- However, no more Representer Theorem.
Data: \( \{(x, y)\}_{i=1}^{m} \)

Features: \( \{ (\phi_1(x_i), \ldots, \phi_n(x_i), y) \}_{i=1}^{m} \)

\[ \hat{f}: x \mapsto y \]
The problem with “black box” ML

Black box machine learning ignores the structure of the data itself:

- The algorithmic part is separated from the statistical part.
- $O(n^3)$ complexity is totally unrealistic in today’s world.
- No obvious parallelism.
- Ignores the most important ideas of the Applied Math of the last 30 years, such as multiresolution analysis, fast multipole methods and multigrid.

(The relationship of deep learning to all this is not clear yet.)
Multiresolution ML
Data
↓
Features
↓
\( \hat{f}: x \mapsto y \)

\( \leftarrow \text{Multiresolution (e.g., Scattering)} \)

\( \leftarrow \text{Multiresolution!} \)
Multiresolution Matrix Factorization

MMF is a multilevel factorization of the data similarity (kernel) matrix $A$ of the form

\[
\begin{pmatrix}
Q_L \\
\end{pmatrix} \cdots \begin{pmatrix}
Q_1 \\
\end{pmatrix} P \begin{pmatrix}
A \\
\end{pmatrix} P^\top \begin{pmatrix}
Q_1^\top \\
\end{pmatrix} \cdots \begin{pmatrix}
Q_L^\top \\
\end{pmatrix} \approx \begin{pmatrix}
H \\
\end{pmatrix}
\]

where

- $P$ is just a permutation matrix to reorder the rows and columns of $A$,
- Each $Q_\ell$ is an orthogonal matrix that has special sparsity structure,
- Outside its $[Q_\ell]_{1:\delta_{\ell-1},1:\delta_{\ell-1}}$ block, each $Q_\ell$ is just the identity for some fixed sequence $n \geq \delta_1 \geq \ldots \geq \delta_L$.

→ A hierarchical, multipole-type description of the data. In factorized form, $A$ is much easier to deal with.
Eigendecomposition vs. MMF

The eigendecomposition (PCA) of a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is

$$\begin{pmatrix}
Q \\
A \\
Q^\top
\end{pmatrix}
\begin{pmatrix}
\text{ } \\
D \\
\text{ }
\end{pmatrix}
= 
\begin{pmatrix}
Q \\
A \\
Q^\top
\end{pmatrix}
\begin{pmatrix}
\text{ } \\
D \\
\text{ }
\end{pmatrix}
$$

The MMF factorization is

$$\begin{pmatrix}
Q_L \\
Q_1 \\
\text{ } \\
\text{ }
\end{pmatrix}
\cdots
\begin{pmatrix}
Q_L \\
Q_1 \\
\text{ } \\
\text{ }
\end{pmatrix}
\begin{pmatrix}
\text{ } \\
A \\
\text{ }
\end{pmatrix}
\begin{pmatrix}
Q_L^\top \\
Q_1^\top \\
\text{ } \\
\text{ }
\end{pmatrix}
\begin{pmatrix}
\text{ } \\
\text{ }
\end{pmatrix}
\approx
\begin{pmatrix}
H \\
\text{ }
\text{ }
\end{pmatrix}$$

The eigendecomposition always exists, is essentially unique, and truncating it after $k$ terms is the optimal approximation in the Frobenius and nuclear norms. But it is expensive to compute $\sim O(n^3)$, eigenvectors are dense, does not take advantage of hierarchical structure.
In general, multiresolution analysis on a space $X$ is a filtration

$$L_2(X) \rightarrow \ldots \rightarrow V_0 \rightarrow V_1 \rightarrow V_2 \rightarrow \ldots$$

where $V_\ell = V_{\ell+1} \oplus W_{\ell+1}$ and

- Each $V_\ell$’s orthonormal basis is $\{\phi_\ell^m\}_m$
- Each $W_\ell$’s orthonormal basis is $\{\psi_\ell^m\}_m$.

The spaces are chosen so that as $\ell$ increases, $V_\ell$ contain functions that are increasingly smooth w.r.t. some self-adjoint operator $T : L(X) \rightarrow L(X)$. 
The multiresolution mantra

The central dogma of harmonic analysis is that the structure of the space of functions on a set $X$ can shed light on the structure of $X$ itself.

\[ G \leftrightarrow L(G) \]

“The interplay between geometry of sets, function spaces on sets, and operators on sets is classical in Harmonic Analysis.”

[Coifman & Maggioni, 2006]

Multiresolution analysis is an attractive paradigm for ML because

- Data naturally clusters into (soft) hierarchies.
- The resulting data structures can form the basis of fast algorithms.
But how does a matrix induce multiresolution???
Fundamentals of multiresolution analysis
Mallat [1989] defined multiresolution on $\mathbb{R}$ by the following axioms:

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

which imply the existence of a mother wavelet $\psi$ and a father wavelet $\phi$ s. t.

$$\psi_m^\ell = 2^{-\ell/2} \psi(2^{-\ell} x - m) \quad \text{and} \quad \phi_m^\ell = 2^{-\ell/2} \phi(2^{-\ell} x - m).$$
Multiresolution on discrete spaces

\[ L_2(X) \rightarrow \ldots \rightarrow V_0 \rightarrow V_1 \rightarrow V_2 \rightarrow \ldots \]
\[ W_1 \rightarrow W_2 \rightarrow W_3 \]

Which of the ideas from classical multiresolution still make sense?

- Recursively split \( L(X) \) into smoother and rougher parts. ✓
- Basis functions should be localized in space & frequency. ✓
- Each \( \Phi_\ell \xrightarrow{Q_\ell} \Phi_{\ell+1} \cup \Psi_{\ell+1} \) transform is orthogonal and sparse. ✓
- Each \( \psi_m^\ell \) is derived by translating \( \psi^\ell \) → MAYBE
- Each \( \psi^\ell \) is derived by scaling \( \psi \) → ???
General principles

1. The sequence \( L(X) = V_0 \supset V_1 \supset V_2 \supset \ldots \) is a filtration of \( \mathbb{R}^n \) in terms of smoothness with respect to \( T \) in the sense that

\[
\mu_\ell = \inf_{f \in V_\ell \setminus \{0\}} \frac{\langle f, T f \rangle}{\langle f, f \rangle}
\]

increases at a given rate.

2. The wavelets are localized in the sense that

\[
\inf_{x \in X} \sup_{y \in X} \frac{\psi_{m}(y)}{d(x, y)^\alpha}
\]

increases no faster than a certain rate.

3. Letting \( Q_\ell \) be the matrix expressing \( \Phi_\ell \cup \Psi_\ell \) in the previous basis \( \Phi_{\ell-1} \), i.e.,

\[
\phi_{m}^{\ell} = \sum_{i=1}^{\dim(V_{\ell-1})} [Q_\ell]_{m,i} \phi_{i}^{\ell-1}
\]

\[
\psi_{m}^{\ell} = \sum_{i=1}^{\dim(V_{\ell-1})} [Q_\ell]_{m+\dim(V_{\ell-1}),i} \phi_{i}^{\ell-1},
\]

each \( Q_\ell \) orthogonal transform is sparse, guaranteeing the existence of a fast wavelet transform (\( \Phi_0 \) is taken to be the standard basis, \( \phi_{m}^{0} = e_{m} \)).
Multiresolution Matrix Factorization (MMF)
If \(|X| = n\) is finite, representing \(T\) by a symmetric matrix \(A \in \mathbb{R}\), each basis transform \(V_\ell \to V_{\ell+1} \oplus W_{\ell+1}\) is like applying a rotation matrix

\[
A \leftrightarrow Q_1 A Q_1^\top \leftrightarrow Q_2 Q_1 A Q_1^\top Q_2^\top \leftrightarrow \ldots
\]

and then fixing a subset of the coordinates as wavelets. In addition, \(Q_1, \ldots, Q_L\) must obey sparsity constraints.

**multiresolution analysis** \(\leftrightarrow\) **multilevel matrix factorization**
Definition. Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$, a class of sparse rotations $Q$, and a sequence $n \geq \delta_1 \geq \ldots \geq \delta_L$, a multiresolution factorization of $A$ is

$$A = Q_1^\top Q_2^\top \ldots Q_L^\top H Q_L \ldots Q_2 Q_1,$$

where each $Q_\ell \in Q$ rotation satisfies $[Q_\ell][n] \setminus S_\ell, [n] \setminus S_\ell = I_{n-\delta_{\ell-1}}$ for some nested sequence of sets $[n] = S_1 \supseteq S_2 \supseteq \ldots \supseteq S_{L+1}$ with $|S_\ell| = \delta_{\ell-1}$, and $H$ is $S_{L+1}$–core diagonal.

Definition. If this is factorization is exact, we say that $A$ is multiresolution factorizable (over $G$ with $\delta_1, \ldots, \delta_L$). → generalization of “rank”
Form of the $Q_\ell$ local rotations

It is critical that the $Q_\ell$ must be very simple and local rotations. Two choices:

1. **Elementary $k$–point rotation:**
   
   $$Q = I_{n-k} \oplus_{(i_1, \ldots, i_k)} O = P \begin{pmatrix} \vdots \end{pmatrix} P^\top$$
   
   for some $O \in SO(k)$ → for $k = 2$, just a Givens rotation.

2. **Compound $k$–point rotation:**

   $$Q = \oplus_{(i_1^1, \ldots, i_{k_1}^1)} O_1 \oplus_{(i_1^2, \ldots, i_{k_2}^2)} O_2 \cdots \oplus_{(i_1^m, \ldots, i_{k_m}^m)} O_m = P \begin{pmatrix} \vdots \end{pmatrix} P^\top$$

   for some $O_1, \ldots, O_m \in SO(k)$. 

→ “Jacobi MMFs”

→ “Parallel MMFs”
The optimization problem

Given $A$, ideally, we would like to solve

$$\text{minimize } \| A - Q_1^\top \cdots Q_L^\top H Q_L \cdots Q_1 \|_{\text{Frob}}^2,$$

for a given class $Q$ of local rotations and dimensions $\delta_1 \geq \delta_2 \geq \ldots \delta_L$.

- In general, this optimization problem is combinatorially hard.
- Easy to approximate it in a greedy way (level by level).
- To solve the combinatorial part of the problem (at each level) use a
  - Deterministic strategy, or a
  - Randomized strategy.
Optimization details — Jacobi MMF

**Proposition.** If \( Q_\ell = I_{n-k} \oplus I O \) with \( I = (i_1, \ldots, i_k) \) and \( J_\ell = \{i_k\} \), then the contribution of level \( \ell \) to the MMF approximation error (in Frobenius norm) is

\[
\mathcal{E}_\ell = \mathcal{E}_1^O = 2 \sum_{p=1}^{k-1} [O [A_{\ell-1}]_I, IO^\top]_{k,p}^2 + 2 [OBO^\top]_{k,k},
\]

where \( B = [A_{\ell-1}]_I, S_\ell ([A_{\ell-1}]_I, S_\ell)^\top \).

**Corollary.** In the special case of \( k = 2 \) and \( I_\ell = (i, j) \),

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

with \( B = [A_{\ell-1}]_{(i,j), S_\ell} ([A_{\ell-1}]_{(i,j), S_\ell})^\top \).
Proposition. Let $A \in \mathbb{R}^{2 \times 2}$ be diagonal, $B \in \mathbb{R}^{2 \times 2}$ symmetric and $O = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}$. 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 $\alpha$ minimizes $(\{OAO^\top\}_{2,1})^2 + (\{OBO^\top\}_{2,2})$, then $\theta$ satisfies

$$(a/e) \sin(2\theta) + \sin(\theta + \omega + \pi/2) = 0.$$
Optimization details — Parallel MMF

**Proposition.** If $Q_\ell$ is a compound rotation of the form

$$Q_\ell = \bigoplus I_1 O_1 \ldots \bigoplus I_m O_m$$

for some partition $I_1 \cup \ldots \cup I_m$ of $[n]$ with $k_1, \ldots, k_m \leq k$, and some sequence of orthogonal matrices $O_1, \ldots, O_m$, then level $\ell$’s contribution to the MMF error obeys

$$\mathcal{E}_\ell \leq 2 \sum_{j=1}^{m} \left( \sum_{p=1}^{k_j-1} \left[ O_j [A_{\ell-1}]_{I_j,I_j} O_j^\top \right]_{k_j,p}^2 + [O_j B_j O_j^\top]_{k_j,k_j} \right),$$

(1)

where $B_j = [A_{\ell-1}]_{I_j,S_{\ell-1}\setminus I_j} \left([A_{\ell-1}]_{I_j,S_{\ell-1}\setminus I_j}\right)^\top$.

For compression tasks parallel MMFs are generally preferable to Jacobi MMFs because

- Unrelated parts of the matrix are processed independently, in parallel.
- Gives more compact factorizations.
- Jacobi MMFs can exhibit cascades.
- The sets $I_1, \ldots, I_m$ can be found by a randomized strategy or exact matching ($O(n^3)$ time)
Hierarchical structure

The sequence in which MMF (with $k \geq 3$) eliminates dimensions induces a (soft) hierarchical clustering amongst the dimensions (mixture of trees).

→ Connection to hierarchical clustering.
Applications

1. Find a (hierarchically) sparse basis for $A$.
2. Hierarchically cluster data.
3. Find community structure.
4. Generate hierarchical graphs.
5. Compress graphs & matrices.
6. Provide a basis for sparse approximations such as the LASSO.
7. Provide a basis for fast numerics (NLA, multigrid, etc).
Relationship to Diffusion Wavelets

- Diffusion wavelets also start with the matrix representation of a smoothing operator (the diffusion operator) and compress it in multiple stages.
- However, at each stage, the wavelets are constructed from the columns of $A$ itself by a rank-revealing QR type process

\[
A \approx Q_1 R_1 \\
A^2 \approx Q_1 R_1 R_1^\dagger Q_1^\dagger \approx Q_2 R_2 \\
A^4 \approx Q_1 Q_2 R_2 R_2^\dagger Q_2^\dagger Q_1^\dagger \approx Q_3 R_3
\]

- Very strong theoretical foundations, but the sparsity (locality) of the $Q_\ell$ matrices is hard to control.

[Coifman & Maggioni, 2006]
Relationship to Treelets

Treelets are a special case of Jacobi MMF

\[ \ldots Q_3 Q_2 Q_1 A Q_1^T Q_2^T Q_3^T \ldots, \]

but

- Restricted to Givens rotations \((k = 2)\) → only recovers a single tree.
- Each \(Q_i\) is chosen to eliminate the maximal off-diagonal entry, rather than minimizing overall error → not intended as a factorization method.
- \(A\) is regarded as a covariance matrix → probabilistic analysis.

[Lee, Nadler & Wasserman, 2008]
Multigrid methods solve systems of p.d.e.’s by shuttling back and forth between grids/meshes at different levels of resolution [Brandt, 1973; Livne & Brandt, 2010].

Fast multipole methods evaluate a kernel (such as the Gaussian kernel) between a large number of particles, by aggregating them at different levels [Greengard & Rokhlin, 1987].

$\mathcal{H}$–matrices [Hackbusch, 1999], $\mathcal{H}^2$ matrices [Borm, 2007] and Hierarchically Semi-Separable matrices [Chandrasekaran et al., 2005] iteratively decompose into blocked matrices, with low rank structure in each of the blocks.
Hölder condition

In classical wavelet transforms one proves that if $f$ is $\alpha$–Hölder, i.e.,

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

then the wavelet coefficients decay at a certain rate, e.g.,

$$|\langle f, \psi^m_\ell \rangle| \leq c' \ell^{\alpha+\beta}$$

Results of this type generally hold for spaces of **homogeneous type**, in which

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

Natural notion of distance between rows in MMF is $d(i, j) = |\langle A_{i,:}, A_{j,:) \rangle|^{-1}$. 
Definition. We say that a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is \textbf{$\Lambda$–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 $c_T$.

Inuitively

- Different rows are neither too parallel or totally orthogonal
- Generalization of the restricted isometry property from compressed sensing [Candes & Tao, 2005]
Theorem

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix that is $\Lambda$–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 $\psi_m^\ell$ 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 $f : [n] \to \mathbb{R}$ is $(c_H, 1/2)$–Hölder with respect to $d(i, j) = |\langle A_{i,:}, A_{j,:} \rangle|^{-1}$, then

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

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

Frobenius norm error on the Zackary Karate Club graph (left) and a matrix of genetic relationship between 50 individuals from [Crossett, 2013](right).
Experimental Results

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

Frobenius norm error of the MMF and Nyström methods on a random vs. a structured (Kronecker product) matrix.
Frobenius norm error of the MMF and Nyström methods on large network datasets.
CONCLUSIONS

- MMF is a new type of matrix factorization mirroring multiresolution analysis → generalization of “rank”.
- MMF exploits hierarchical structure, but does not enforce a single hierarchy.
- Empirical evidence suggests that MMF is a good model for real data.
- Finding MMF factorizations is a fundamentally local and parallelizable process → $O(n \log n)$ algorithms should be within reach.
- Once in MMF form, a range of matrix computations become faster.
- MMF has strong ties to: Diffusion wavelets, Treelets, Multiscale SVD, structured matrices, algebraic multigrid, and fast multipole methods.