NONLINEAR MATRIX RECOVERY
Florentin Goyens1,2, goyens@maths.ox.ac.uk
Coralia Cartis1, Armin Eftekhari3
1University of Oxford, 2Alan Turing Institute, 3University of Umeå

NONLINEAR MATRIX RECOVERY USING A LIFTING
Recover a high-rank matrix \( \mathbf{M} \in \mathbb{R}^{n \times d} \) from linear measurements \( \{ \mathbf{A}_i \} \) via
under the assumption that there exists a \( \Phi(x) = \{ \varphi(\mathbf{m}_i) \} \)
such that the nonlinear structure in \( \mathbf{M} \) makes \( \Phi(\mathbf{M}) \) low-rank.

\[
\mathbb{R}^n \rightarrow \mathbb{R}^N
\]

\[
\Phi(\mathbf{x}) = \{ \varphi(\mathbf{m}_1), \varphi(\mathbf{m}_2), \ldots, \varphi(\mathbf{m}_k) \}
\]

Recovery of algebraic varieties and union of subspaces models
using the polynomial lifting as in [4]. The matrix \( \mathbf{M} \) is said to follow an algebraic variety model if there exists a family of \( q \) polynomials of \( n \) variables \( \{ p_j \}_{j=1}^q \) of degree at most \( d \) such that
\[
p_j(\mathbf{m}_i) = 0, \text{ for every column } \mathbf{m}_i \text{ of } \mathbf{M}.
\]

The polynomial map of degree \( d \) lifts the data points to a multivariate monomial basis
\[
\varphi_\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^N, \varphi_\alpha(x) = x^\alpha, \alpha \leq d,
\]
where \( \alpha \) is a multi-index of non-negative integers
with \( x^\alpha = x_1^{\alpha_1}x_2^{\alpha_2} \ldots x_n^{\alpha_n} \) and \( |\alpha| = \sum \alpha_i \).
For the vector of coefficients \( c_j \) that defines the polynomial \( p_j \) in
the monomial basis, we have \( c_j^T \varphi(\mathbf{m}_i) = 0 \) for every \( j, i \).
Therefore, \( \text{rank}(\Phi(\mathbf{M})) \leq \min\{N - q, d\} \).
and the lifted matrix \( \Phi(\mathbf{M}) \) is rank deficient when \( \mathbf{M} \) belongs to an algebraic variety (including union of subspaces).

CASE STUDY 1: ALGEBRAIC VARIETIES AND UNION OF SUBSPACES
Recovery of clusters with missing data using the Gaussian kernel as lifting. The kernel represents the inner product of implicit features (reproducing kernel Hilbert space).

\[
k^\Phi(M, M) = \frac{\| \text{Truncated svd of } \text{K}(\text{X}, \text{X}) \|_F}{2\sqrt{d}}
\]

• \( m_i \) close to \( m_j \Rightarrow k^\Phi(M, M) \approx 1
• \( m_i \) far from \( m_j \Rightarrow k^\Phi(M, M) \approx 0
• \text{rank}(k^\Phi(M, M)) \approx \text{number of clusters}

A) RIEMANNIAN OPTIMIZATION
Second order Riemannian trust region method on the Grassmannian [1].

Solves a subproblem at each step on the tangent space of the Riemannian manifold

\[
\text{Grass}(N, r) \times \{ \mathbf{A}(\mathbf{x}) = \mathbf{b} \}
\]

\[
\Delta_k = \left\{ \arg\min_{\Delta, \|\Delta\| \leq \rho} f(U_k, X_k) + \langle \text{grad}(U_k, X_k), \Delta \rangle + \frac{1}{2} \|\text{Hess}(U_k, X_k)\| \Delta, \Delta \right\}
\]

• Solution of subproblem proceeds a candidate using the retraction map \( R_{U_k, X_k}(\Delta_k) \) which is assessed by model change decrease to model decrease. The trust region radius is adjusted accordingly.

Implemented in the Manopt toolbox [3]. RTR is a globally convergent method to second order critical points.

Theorem [2]: (Global complexity of ITR) If \( f \circ \phi \) has a Lipschitz Hessian with constant independent of \( x \) and \( f \) is bounded below then RTR returns \( x \) with \( \| \text{grad}(f)(x) \| \leq \varepsilon_f \) and \( \lambda_{\min}\text{Hess}(f)(x) \geq -\varepsilon_H \) in \( O(\max(1/\varepsilon_f, 1/\varepsilon_H)) \) iterations.

B) ALTERNATING MINIMIZATION
At \( X_k \), solve \( U_{k+1} = \arg\min_U \| \Phi(X_k) - P_U \Phi(X_k) \|_F^2 \rightarrow \text{Truncated svd of } \Phi(X_k) \)
At \( U_{k+1} \), solve \( X_{k+1} = \arg\min_X \| \Phi(X_k) - P_{U_{k+1}} \Phi(X_k) \|_F^2 \rightarrow \text{Projected descent method} \)

Theorem: (Global complexity of AM): For \( \varepsilon_f > 0, \varepsilon_H > 0 \) the number of gradient steps \( N_{\text{grad}} \) and number of svd \( N_{\text{svd}} \) such that

\[
\left\| \text{grad}_{U_{k+1}} f(U_{k+1}, X_{k+1}) \right\|_{\text{Fr}}^2 + \text{Tr}(\text{Hess}_{X_{k+1}} f(U_{k+1}, X_{k+1})) \leq \varepsilon_f + \varepsilon_H
\]

\[
N_{\text{grad}} + N_{\text{svd}} \leq \frac{(f_{\text{opt}} - f_k)^2}{\min(\varepsilon_f, \varepsilon_H)} \min \left( \frac{\|\Delta_k\|}{\varepsilon_{\text{Hess}}}, 1 \right)
\]

where \( L \) is a gradient Lipschitz constant, \( \varepsilon \) is a lower bound on the step sizes and \( \beta \in [0, 1] \) is the Armijo sufficient decrease constant and \( f_k \) is a lower bound on \( f \).

NUMERICAL RESULTS
Comparison of first and second order algorithms above on the recovery of a union of subspaces

For \( m \) large enough, solving (1) with arbitrary initialization recovers the matrix \( \mathbf{M} \). Grayscale below gives the proportion of union of subspaces matrices recovered up to \( \| \mathbf{X} - \mathbf{M} \|_F^2 \leq 10^{-3} \) or the proportion of correct clustering over 50 test problems for every pair of parameters.

Recovery

Recovering an increasing number of subspaces of dimension \( r \) in \( \mathbb{R}^n \). Requires much less measurements than fewer high dimensional subspaces.

Case Study 2: Clustering with missing data

Largest singular values indicate the number of clusters but the Gaussian kernel is noisy.

REFERENCES

Optimization formulation
Assume \( r = \text{rank}(\Phi(M)) \) is known.

To solve:

\[
\min_{\mathbf{X}} \text{rank}(\Phi(\mathbf{X})) \quad \mathbf{A}(\mathbf{X}) = \mathbf{b},
\]

where the affine constraint \( \mathbf{A}(\mathbf{X}) = \mathbf{b} \) denotes the measurements on the matrix \( \mathbf{X} \).

where Grass(\( N, r \)) is the Grassmann manifold, the set of all subspaces of dimensions \( r \) in \( \mathbb{R}^N \).

Grayscale below gives the proportion of union of subspaces matrices recovered up to \( \| \mathbf{X} - \mathbf{M} \|_F^2 \leq 10^{-3} \) or the proportion of correct clustering over 50 test problems for every pair of parameters.

Performance of the algorithms

Comparison of first and second order algorithms above on the recovery of a union of subspaces

Recovery for 2 subspaces in \( \mathbb{R}^{10} \) of increasing dimension. Recovery for small dimensions only.

Recovering an increasing number of subspaces of dimension \( r \) in \( \mathbb{R}^{10} \). Requires much less measurements than fewer high dimensional subspaces.

Clustering possible with up to 50% of missing entries. The quality of recovery depends on the spectral gap of the Grassmann kernel.