Nonlinear matrix recovery

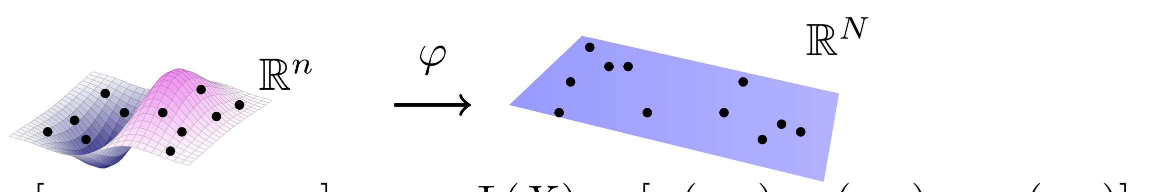
Florentin Goyens[1,2] — goyens@maths.ox.ac.uk

Coralia Cartis[1,2], **Armin Eftekhari**[3]

[1]University of Oxford, [2]Alan Turing Institute, [3]University of Umeå

Nonlinear matrix recovery using a lifting

Recover a high-rank matrix $M \in \mathbb{R}^{n\times s}$ from linear measurements $\langle M, A_i\rangle = b_i, i = 1, \dots, m$ under the assumption that there exists a lifting map $\varphi \;\; \mathbb{R}^n \longrightarrow \mathbb{R}^N$ such that the nonlinear structure in M makes $\Phi(M)$ low-rank.



$M = [m_1 \;\; m_2 \cdots m_s]$ $\qquad$ $\Phi(X) = [\varphi(m_1) \;\; \varphi(m_2) \cdots \varphi(m_s)]$

Optimization formulation

Assume $r = \text{rank}\,(\Phi(M))$ is known. To solve

$$\begin{cases} \min_X & \text{rank}(\Phi(X)) \\ & \mathcal{A}(X) = b, \end{cases}$$

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

We minimize a nonconvex approximation of the rank

$$\begin{cases} \min_{\mathcal{U},X} & \|\Phi(X) - P_{\mathcal{U}}\Phi(X)\|_F^2 \\ & \mathcal{U} \in \text{Grass}(N, r) \\ & \mathcal{A}(X) = b, \end{cases} \tag{1}$$

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

Case Study 1 : algebraic varieties and Union of Subspaces

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

$$p_j(m_i) = 0, \quad \text{for every column } m_i \text{ of } M.$$

The polynomial map of degree d lifts the data points to a *multivariate monomial basis*

$\phi_d \;\; \mathbb{R}^n \longrightarrow \mathbb{R}^N, \phi_d(x) = x^\alpha, |\alpha| \le d,$ where α is a multi index of non-negative integers

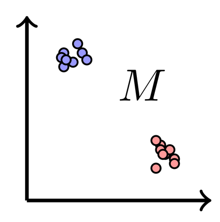
with $x^\alpha \doteq x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n}$ and $|\alpha| = \sum_{i=1}^n \alpha_i$. For the vector of coefficients c_j that defines the polynomial p_j in the monomial basis, we have $c_j^\top \phi_d(m_i) = 0$ for every i, j. Therefore, $\text{rank}(\Phi_d(M)) \le \min(N - q, s)$ and the lifted matrix $\Phi_d(M)$ is rank deficient when M belongs to an algebraic variety (including union of subspaces).

Case Study 2: Clustering with missing data

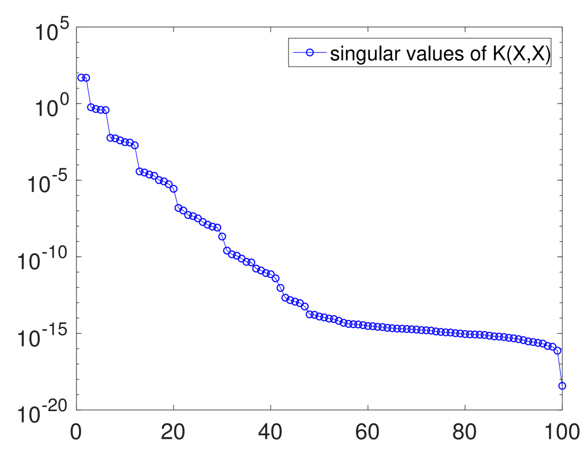
Recovery of clusters with missing data using the Gaussian kernel as lifting. The kernel represents the inner product of implicit features (reproducing kernel Hilbert space).

$$k_{ij}^g(M, M) = e^{-\frac{\|m_i - m_j\|_2^2}{2\sigma^2}}$$

- m_i close to $m_j \implies k_{ij}^g(M, M) \approx 1$
- m_i far from $m_j \implies k_{ij}^g(M, M) \approx 0$
- rank $(k^g(M, M)) \approx$ number of clusters



$$k^g(M, M) \approx \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



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

Optimization algorithms

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 $\text{Grass}(N, r) \times \{X \; \mathcal{A}(X) = b\}$

$$\Delta_+ = \begin{cases} \underset{\Delta \in T_{(\mathcal{U}_k, X_k)}}{\text{argmin}} & f(\mathcal{U}_k, X_k) + \langle \text{grad} f(\mathcal{U}_k, X_k), \Delta\rangle + \frac{1}{2}\langle \text{Hess } f(\mathcal{U}_k, X_k)[\Delta], \Delta\rangle \\ & \|\Delta\| \le \rho. \end{cases}$$

- Solution of subproblem produces a candidate using the retraction map $R_{(\mathcal{U}_k, X_k)}(\Delta_+)$ which is assessed by comparing model decrease to function decrease. The trust region radius is adjusted accordingly.

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

Theorem [2](Global complexity of RTR) If $f \circ R$ has a Lipschitz Hessian with constant independent of x and f is bounded below then RTR returns x with $\text{grad } f(x) \le \varepsilon_g$ and $\lambda_{\min}\text{Hess } f(x) \ge -\varepsilon_H$ in $\mathcal{O}(\max\{1/\varepsilon_H^3, 1/\varepsilon_g^2\varepsilon_H\})$ iterations.

B) Alternating minimization

At X_k, solve $\mathcal{U}_{k+1} = \begin{cases} \underset{\mathcal{U}}{\text{argmin}} & \|\Phi(X_k) - P_{\mathcal{U}}\Phi(X_k)\|_F^2 \\ & \mathcal{U} \in \text{Grass}(s, r). \end{cases}$ $\longrightarrow$ Truncated svd of $\Phi(X_k)$

At $\mathcal{U}_{k+1}$, solve $X_{k+1} = \begin{cases} \underset{X}{\text{argmin}} & \|\Phi(X) - P_{\mathcal{U}_{k+1}}\Phi(X)\|_F^2 \\ & \mathcal{A}(X) = b. \end{cases}$ $\longrightarrow$ Projected descent method

Theorem (Global complexity of AM): For $\varepsilon_x > 0$, $\varepsilon_u > 0$ the number of gradient steps N_{grad} and number of svd N_{svd} such that $\left\| \left(\text{grad}_U f(U_k, X_k), P_{KerA}\nabla_X f(U_k, X_k)\right)\right\| \le \varepsilon_u + \varepsilon_x$ is

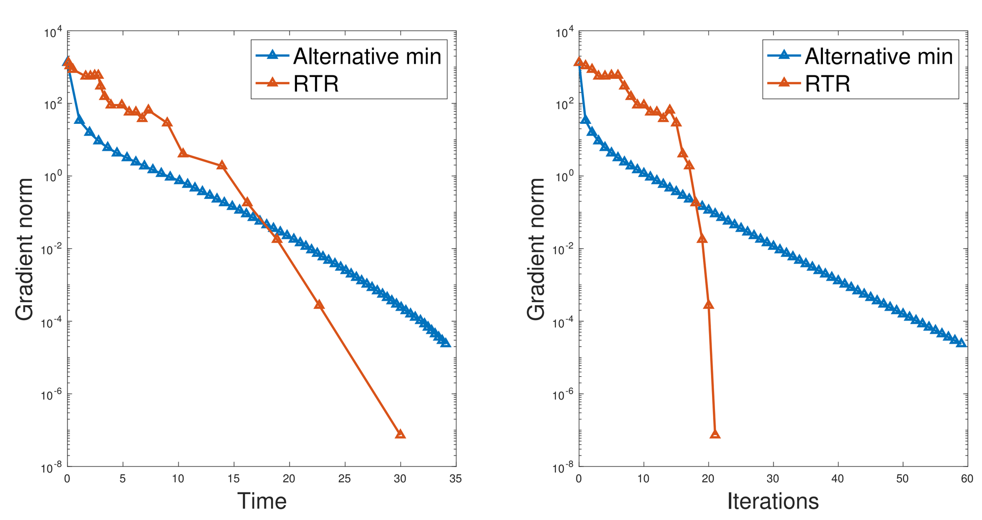
$$N_{grad} + N_{svd} \le \frac{(f_0 - f_*)}{\min(\varepsilon_u, \varepsilon_x)^2 \min\left(\frac{1}{2L_u}, \underline{\alpha}\beta\right)}.$$

where L_u is a gradient Lipschitz constant, $\underline{\alpha}$ is a lower bound on the step sizes and $\beta \in]0, 1[$ is the Armijo sufficient decrease constant and f_* is a lower bound on f.

Numerical results

Performance of the algorithms

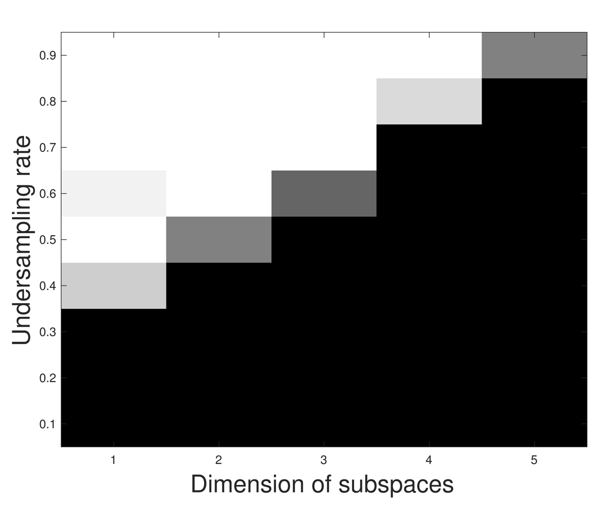
Comparison of first and second order algorithms above on the recovery of a union of subspaces



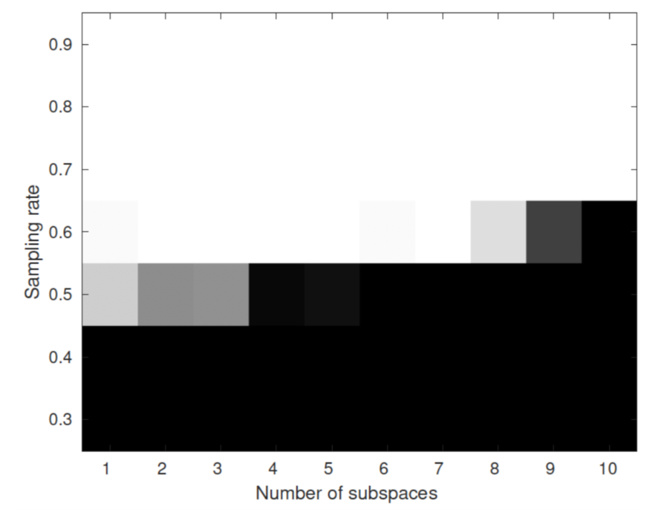
Riemannian second order method preferable in high accuracy regime over first order alternating minimization (linear vs. superlinear local convergence rate)

Recovery

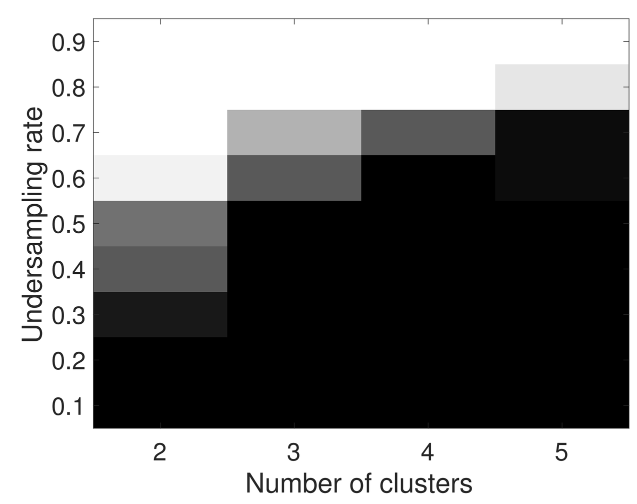
For m large enough, solving (1) with arbitrary initialization recovers the matrix M. Grayscale below gives the proportion of union of subspaces matrices recovered up to $\|X - M\|_F^2 \le 10^{-3}$ or the proportion of correct clustering over 50 test problems for every pair of parameters.



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



Recovering an increasing number of subspaces of dimension 2 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 Gaussian kernel.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2007.

[2] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 2016.

[3] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

[4] G. Ongie, R. Willett, R. D. Nowak, and L. Balzano. Algebraic variety models for high-rank matrix completion. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2691–2700, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.