Robust Principal Component Pursuit via Alternating Minimization Scheme on Matrix Manifolds

Tao Wu

Institute for Mathematics and Scientific Computing
Karl-Franzens-University of Graz

joint work with Prof. Michael Hintermüller

Low-rank paradigm.

Low-rank matrices arise in one way or another:

- low-degree statistical processes
  $\rightsquigarrow$ e.g. collaborative filtering, latent semantic indexing.
- regularization on complex objects
  $\rightsquigarrow$ e.g. manifold learning, metric learning.
- approximation of compact operators
  $\rightsquigarrow$ e.g. proper orthogonal decomposition.

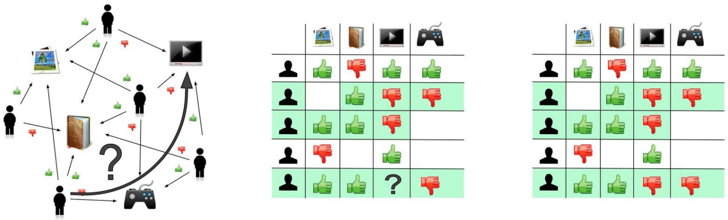
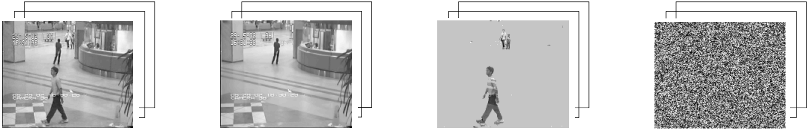


Fig.: Collaborative filtering (courtesy of wikipedia.org).

Robust principal component pursuit.

- Sparse component corresponds to pattern-irrelevant outliers.
- Robustifies classical principal component analysis.
- Carries important information in certain applications; e.g. moving objects in surveillance video.
- Robust principal component pursuit:

data		low-rank		sparse		noise
Z	$=$	A	$+$	B	$+$	N



- Introduced in [Candés, Li, Ma, and Wright, '11], [Chandrasekaran, Sanghavi, Parrilo, and Willsky, '11].

Convex-relaxation approach.

- A popular (convex) variational model:

$$\min \ \|A\|_{\text{nuclear}} + \lambda \|B\|_{\ell^1}$$
$$\text{s.t. } \|A + B - Z\| \leq \varepsilon.$$

- Considered in [Candés, Li, Ma, and Wright, '11], [Chandrasekaran, Sanghavi, Parrilo, and Willsky, '11], ...
- $\text{rank}(A)$ relaxed by nuclear-norm; $\|B\|_0$ relaxed by ℓ^1-norm.
- Numerical solvers: proximal gradient method, augmented Lagrangian method, ...
  $\rightsquigarrow$ Efficiency is constrained by SVD in full dimension at each iteration.

Manifold constrained least-squares model.

- Our variational model:

$$\begin{aligned}\min\ & \frac{1}{2}\|A+B-Z\|^2 \\ \text{s.t. } & A \in \mathcal{M}(r) := \{A \in \mathbb{R}^{m\times n} : \operatorname{rank}(A) \le r\}, \\ & B \in \mathcal{N}(s) := \{B \in \mathbb{R}^{m\times n} : \|B\|_0 \le s\}.\end{aligned}$$

- Our goal is to develop an algorithm such that:
  - globally converges to a stationary point (often a local minimizer).
  - provides exact decomposition with high probability for noiseless data.
  - outperforms solvers based on convex-relaxation approach, especially in large scales.

Existence of solution and optimality condition.

- A little quadratic regularization ($0 < \mu \ll 1$) is included for the (theoretical) sake of existence of a solution; i.e.

$$\min \; f(A,B) := \frac{1}{2}\|A+B-Z\|^2 + \frac{\mu}{2}\|A\|^2, \\ \text{s.t. } (A,B) \in \mathcal{M}(r) \times \mathcal{N}(s).$$

  In numerics, choosing $\mu = 0$ seems fine.

- Stationarity condition as variational inequalities:

$$\begin{cases} \langle \Delta, (1+\mu)A^* + B^* - Z \rangle \geq 0, & \text{for any } \Delta \in T_{\mathcal{M}(r)}(A^*), \\ \langle \Delta, A^* + B^* - Z \rangle \geq 0, & \text{for any } \Delta \in T_{\mathcal{N}(s)}(B^*). \end{cases}$$

  $T_{\mathcal{M}(r)}(A^*)$ and $T_{\mathcal{N}(s)}(B^*)$ refer to tangent cones.

Constraints of Riemannian manifolds.

- $\mathcal{M}(r)$ is Riemannian manifold around A^* if $\mathrm{rank}(A^*) = r$;
  $\mathcal{N}(s)$ is Riemannian manifold around B^* if $\|B^*\|_0 = s$.
- Optimality condition reduces to:

$$\begin{cases} P_{T_{\mathcal{M}(r)}(A^*)}((1+\mu)A^* + B^* - Z) = 0, \\ P_{T_{\mathcal{N}(s)}(B^*)}(A^* + B^* - Z) = 0. \end{cases}$$

  $P_{T_{\mathcal{M}(r)}(A^*)}$ and $P_{T_{\mathcal{N}(s)}(B^*)}$ are orthogonal projections onto subspaces.
- Tangent space formulae:

$$T_{\mathcal{M}(r)}(A^*) = \{UMV^\top + U_pV^\top + UV_p^\top : A^* = U\Sigma V^\top \text{ as compact SVD}, \\ M \in \mathbb{R}^{r\times r}, U_p \in \mathbb{R}^{m\times r}, U_p^\top U = 0, V_p \in \mathbb{R}^{n\times r}, V_p^\top V = 0\},$$

$$T_{\mathcal{N}(s)}(B^*) = \{\Delta \in \mathbb{R}^{m\times n} : \mathrm{supp}(\Delta) \subset \mathrm{supp}(B^*)\}.$$

A conceptual alternating minimization scheme.

Initialize $A^0 \in \mathcal{M}(r)$, $B^0 \in \mathcal{N}(s)$. Set $k := 0$ and iterate:

1. $A^{k+1} \approx \arg\min_{A \in \mathcal{M}(r)} \frac{1}{2}\|A + B^k - Z\|^2 + \frac{\mu}{2}\|A\|^2$.
2. $B^{k+1} \approx \arg\min_{B \in \mathcal{N}(s)} \frac{1}{2}\|A^{k+1} + B - Z\|^2$.

Theorem (sufficient descrease + stationarity $\Rightarrow$ convergence)

Let $\{(A^k, B^k)\}$ be generated as above. Suppose that there exists $\delta > 0$, $\varepsilon_a^k \downarrow 0$, and $\varepsilon_b^k \downarrow 0$ such that for all k:

$$f(A^{k+1}, B^k) \le f(A^k, B^k) - \delta\|A^{k+1} - A^k\|^2,$$
$$f(A^{k+1}, B^{k+1}) \le f(A^{k+1}, B^k) - \delta\|B^{k+1} - B^k\|^2,$$
$$\langle \Delta, (1+\mu)A^{k+1} + B^k - Z\rangle \ge -\varepsilon_a^k\|\Delta\|, \quad \text{for any } \Delta \in T_{\mathcal{M}(r)}(A^{k+1}),$$
$$\langle \Delta, A^{k+1} + B^{k+1} - Z\rangle \ge -\varepsilon_b^k\|\Delta\|, \quad \text{for any } \Delta \in T_{\mathcal{N}(s)}(B^{k+1}).$$

Then any non-degenerate limit point (A^*, B^*), i.e. $\operatorname{rank}(A^*) = r$ and $\|B^*\|_0 = s$, satisfies the first-order optimality condition.

Sparse matrix subproblem.

- The global solution $P_{\mathcal{N}(s)}(Z - A^{k+1})$ (as metric projection) can be efficiently calculated from "sorting".
- The global solution may not necessarily fulfill the sufficient descrease condition.
- Whenever necessary, *safeguard* by a local solution:

$$B_{ij}^{k+1} = \begin{cases} (Z - A^{k+1})_{ij}, & \text{if } B_{ij}^{k} \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

- Given non-degeneracy of B^{k+1}, i.e. $\|B^{k+1}\|_0 = s$, the exact stationarity holds.

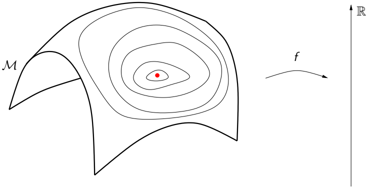
Low-rank matrix subproblem: a Riemannian perspective.

- Global solution $P_{\mathcal{M}(r)}(\frac{1}{1+\mu}(Z-B^k))$ as metric projection:
  - available due to Eckart-Young theorem; i.e.

$$\frac{1}{1+\mu}(Z-B^k)=\sum_{j=1}^{n}\sigma_j u_j v_j^\top \;\Rightarrow\; P_{\mathcal{M}(r)}(\frac{1}{1+\mu}(Z-B^k))=\sum_{j=1}^{r}\sigma_j u_j v_j^\top.$$

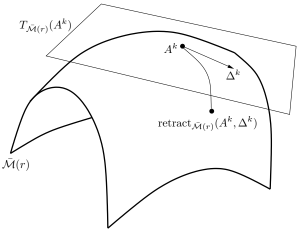
  - but requires SVD in full dimension
    $\rightsquigarrow$ expensive for large-scale problems (e.g. $m,n\geq 2000$).
- Alternatively resolved by a single *Riemannian optimization* step on matrix manifold.
- Riemannian optimization applied to low-rank matrix/tensor problems; see [Simonsson and Eldén, '10], [Savas and Lim, '10], [Vandereycken, '13], ...
- Our goal: The subproblem solver should activate the convergence criteria, i.e. sufficient descrease + stationarity.

Riemannian optimization: an overview.



- References: [Smith, '93], [Edelman, Arias, and Smith, '98], [Absil, Mahony, and Sepulchre, '08], ...
- Why Riemannian optimization?
  - Local homeomorphism is computationally infeasible/expensive.
  - Intrinsically low dimensionality of the underlying manifold.
  - Further dimension reduction via quotient manifold.
- Typical Riemannian manifolds in applications:
  - finite-dimensional (matrix manifold): Stiefel manifold, Grassmann manifold, fixed-rank matrix manifold, ...
  - infinite-dimensional: shape/curve spaces, ...

Riemannian optimization: a conceptual algorithm.



At the current iterate:

1. Build a quadratic model in the tangent space using Riemannian gradient and Riemannian Hessian.
2. Based on the quadratic model, build a tangential search path.
3. Perform backtracking path search via retraction to determine the step size.
4. Generate the next iterate.

Riemannian gradient and Hessian.

- $\bar{\mathcal{M}}(r) := \{A : \mathrm{rank}(A) = r\}$; $f_A^k : A \in \bar{\mathcal{M}}(r) \mapsto f(A, B^k)$.
- Riemannian gradient, $\mathrm{grad} f_A^k(A) \in T_{\bar{\mathcal{M}}(r)}(A)$, is defined s.t. $\langle \mathrm{grad} f_A^k(A), \Delta \rangle = D f_A^k(A)[\Delta],\ \forall \Delta \in T_{\bar{\mathcal{M}}(r)}(A)$.

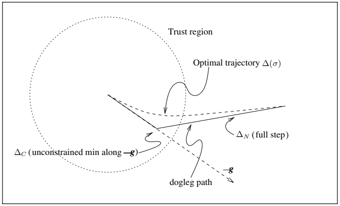
$$\mathrm{grad} f_A^k(A) = P_{T_{\bar{\mathcal{M}}(r)}(A)}(\nabla f_A^k(A)).$$

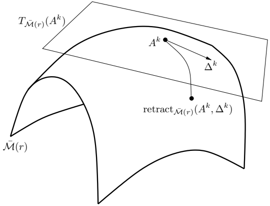
- Riemannian Hessian, $\mathrm{Hess} f_A^k(A) : T_{\bar{\mathcal{M}}(r)}(A) \to T_{\bar{\mathcal{M}}(r)}(A)$, is defined s.t. $\mathrm{Hess} f_A^k(A)[\Delta] = \nabla_\Delta \mathrm{grad} f_A^k(A),\ \forall \Delta \in T_{\bar{\mathcal{M}}}(A)$.

$$\begin{aligned}\mathrm{Hess} f_A^k(A)[\Delta] = &(I - UU^\top)\nabla f_A^k(A)(I - VV^\top)\Delta^\top U\Sigma^{-1}V^\top \\ &+ U\Sigma^{-1}V^\top\Delta^\top(I - UU^\top)\nabla f_A^k(A)(I - VV^\top) \\ &+ (1+\mu)\Delta.\end{aligned}$$

See, e.g., [Vandereycken, '12].

Dogleg search path and projective retraction.





- "Dogleg" path $\Delta^k(\tau^k)$ as approximation of optimal trajectory of tangential trust-region subproblem (left figure):

$$\min \; f_A^k(A^k) + \langle g^k, \Delta \rangle + \frac{1}{2}\langle \Delta, H^k[\Delta] \rangle$$
$$\text{s.t. } \Delta \in T_{\bar{\mathcal{M}}(r)}(A^k), \; \|\Delta\| \leq \sigma.$$

- Metric projection as retraction (right figure):

$$\text{retract}_{\bar{\mathcal{M}}(r)}(A^k, \Delta^k(\tau^k)) = P_{\bar{\mathcal{M}}(r)}(A^k + \Delta^k(\tau^k)).$$

Computationally efficient: "reduced" SVD on $2r$-by-$2r$ matrix!

Low-rank matrix subproblem: projected dogleg step.

Given $A^k \in \bar{\mathcal{M}}(r),\ B^k \in \mathcal{N}(s)$:

1. Compute g^k, H^k, and build the dogleg search path $\Delta^k(\tau^k)$ in $T_{\bar{\mathcal{M}}(r)}(A^k)$.
2. Whenever non-positive definiteness of H^k is detected, replace the dogleg search path by the line search path along steepest descent direction, i.e. $\Delta(\tau^k) = -\tau^k g^k$.
3. Perform backtracking path/line search; i.e. find the largest step size $\tau^k \in \{2, 3/2, 1, 1/2, 1/4, 1/8, ...\}$ s.t. the sufficient descrease condition is satisfied:

   $$f_A^k(A^k) - f_A^k(P_{\bar{\mathcal{M}}(r)}(A^k + \Delta^k(\tau^k))) \geq \delta \| A^k - P_{\bar{\mathcal{M}}(r)}(A^k + \Delta^k(\tau^k)) \|^2.$$

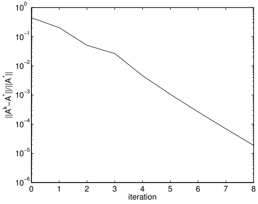
4. Return $A^{k+1} = f_A^k(P_{\bar{\mathcal{M}}(r)}(A^k + \Delta^k(\tau^k)))$.

Low-rank matrix subproblem: convergence theory.

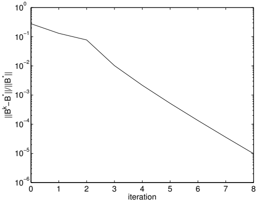
- Backtracking path search:
  - The sufficient descrease condition can always be fulfilled after finitely many trails on τ^k.
  - Any accumulation point of $\{A^k\}$ is stationary.
- Further assume $\mathrm{Hess} f(A^*, B^*)\Big|_{\mu=0} \succ 0$ at a non-degenerate accumulation point (A^*, B^*). Then
  - Tangent-space transversality holds, i.e.
    $$T_{\bar{\mathcal{M}}(r)}(A^*) \cap T_{\mathcal{N}(s)}(B^*) = \{0\}.$$
  - Contractivity of $P_{T_{\bar{\mathcal{M}}(r)}(A^*)} \circ P_{T_{\mathcal{N}(s)}(B^*)}$: $\exists \kappa \in [0, 1)$ s.t.
    $$\|(P_{T_{\bar{\mathcal{M}}(r)}(A^*)} \circ P_{T_{\mathcal{N}(s)}(B^*)})(\Delta)\| \leq \kappa \|\Delta\|.$$
  - q-**linear convergence** of $\{A^k\}$ towards stationarity:
    $$\limsup_{k \to \infty} \frac{\|A^{k+1} - A^*\|}{\|A^k - A^*\|} \leq \kappa.$$

Numerical implementation.

- Trimming $\rightsquigarrow$ Adaptive tuning of rank r^{k+1} and cardinality s^{k+1} based on the current iterate (A^k, B^k).
  - k-means clustering on (nonzero) singular values of A^k in logarithmic scale.
  - hard thresholding on entries of B^k.
- q-linear convergence confirmed numerically:

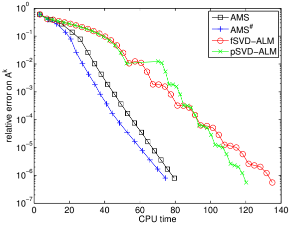


(a) Convergence of $\{A^k\}$.

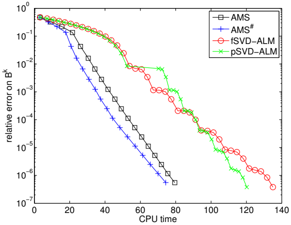


(b) Convergence of $\{B^k\}$.

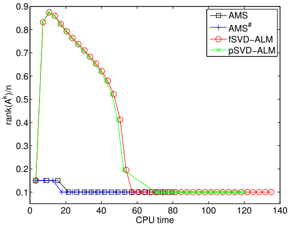
Comparison with augmented Lagrangian method ($m = n = 2000$).



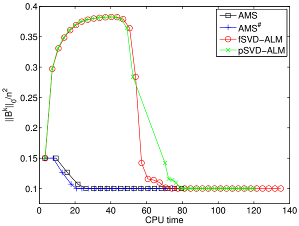
(a) Relative error of $\{A^k\}$.



(b) Relative error of $\{B^k\}$.



(c) Phase transition of $\{A^k\}$.



(d) Phase transition of $\{B^k\}$.

Application to surveillance video.

- Problem settings:
  - A sequence of 200 frames taken from a surveillance video at an airport.
  - Each frame is a gray image of resolution 144×176.
  - Stack 3D-array into a 25344×200 matrix.
- Results:
  - CPU time: AMS $\rightsquigarrow$ 39.4s; ALM $\rightsquigarrow$ 124.4s.
  - Visual comparison.