New Approaches and Demands of Optimization Problems with Orthogonality Constraints

Xin Liu

State Key Laboratory of Scientific and Engineering Computing
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, China

运筹千里纵横论坛

May 31, 2020

Acknowledgements

Collaborators

- Xiaojun Chen (PolyU)
- Bin Gao (KUL, Belgium)
- Guanghui Hu (UM)
- Yang Kuang (NUS, Singapore)
- Michael Ng (HKU)
- Zaiwen Wen (PKU)
- Liwei Xu (UESTC)
- Chao Yang (LBNL U.S.)
- Ya-xiang Yuan
- Yin Zhang, (CUHK(SZ))
- Zaikun Zhang, (PolyU)

Students:

- Yadan Chen, Jiang Hu (PKU), Xiaoyin Hu, Wei Liu, Nachuan Xiao, Lei Wang, Siyun Zhou (UESTC)

Optimization Problem with Orthogonality Constraints

General Form

$$\min_{X \in \mathbb{R}^{n\times p}} \quad f(X)$$
$$\text{s.t.} \quad X^\top X = I. \qquad \text{(OCP)}$$

- $f : \mathbb{R}^{n\times p} \longmapsto \mathbb{R}$
- $n > p$
- $p(p+1)/2$ constraints –– nonconvex
- Stiefel manifold:

$$\mathcal{S}_{n,p} := \{X \in \mathbb{R}^{n\times p} | \, X^\top X = I\}.$$

What Attracts Me?

- Emerging applications
- Challenging

Starting from a special case:

$$f(X) = \mathbf{tr}(X^\top A X)$$

Dominant Eigenspace Calculation and Singular Value Decomposition

Dominant Eigenspace Calculation

- $AX_k = Q_k\Lambda_k$
- $\Lambda_k \in \mathbb{R}^{k\times k}$ contains k largest/smallest (shifting $-A$) eigenvalues on its diagonal.
- $Q_k \in \mathbb{R}^{n\times k}$ consists of the associate eigenvectors.

Singular Value Decomposition/Principal Component Analysis

- k dominant SVD ($k \ll m$)

$$B_k \triangleq \sum_{i=1}^{k} \sigma_i u_i v_i^\top = \mathop{\arg\min}_{\mathbf{rank}(X)\le k} \|B - X\|_F^2 \quad \approx \; B = \sum_{i=1}^{m} \sigma_i u_i v_i^\top.$$

- Dominant eigenspace computation of $B^\top B$ or $BB^\top$

Adequate Existing Methods

- Books and surveys
  - Saad 1992, “Numerical Methods for Large Eigenvalue Problems”
  - Stewart 1998, “Matrix Algorithms Volume II - Eigensystems”
  - Sorensen 2002, “Numerical Methods for Large Eigenvalue Problems”
  - Hernández et al. 2009, “A Survey of Software for Sparse Eigenvalue Problems”
  - ......
- Lanczos methods – *ARPACK* (`eigs` in Matlab)

  Sorensen 1996, “Implicitly Restarted Arnoldi/Lanczos Methods for Large Scale Eigenvalue Calculations ”

Plenty of Vigor from Big Data Analysis

- A fundamental tool for many emerging optimization problems
  - First-order methods for semidefinite program
  - Low-rank matrix completion
  - Robust principal component analysis
  - Sparse principal component analysis
  - Nonnegative matrix factorization
  - Sparse inverse covariance matrix estimation
  - Nearest correlation matrix estimation
- Various scientific and engineering applications
  - High dimensional data reduction
  - Density functional theory for electronic structure calculation
  - Bose Einstein condensate

Motivations

Why not use one of the existing eigensolvers?

Emerging applications demand new capacities.

- high efficiency at moderate accuracy
- high eigenspace dimensions
- high parallel scalability
- warm-start capacity

Established eigensolvers often lack in one or more aspects.

Block vs. Sequential (Lanczos-type Methods)

- Block SpMV: $AV = [Av_1 \; Av_2 \; \cdots Av_k]$
- Sequential SpMv's: $Av \rightarrow A^2 v \cdots \rightarrow A^k v$
  (+ inner products for orthogonalization)

As k increases, block methods are gaining advantages.

Block methods can be warm-started in an iterative setting.

Block Methods

Classic Block Method SSI: (extension of power method)

$$X^{i+1} = \mathbf{orth}(AX^i)$$

Other block algorithms:

- Block Jacobian-Davidson: PRIMME, FEAST

Rayleigh-Ritz Trace Maximization

$$\max_{X \in \mathbb{R}^{n\times k}} \quad f(X) := \mathbf{tr}(X^\top AX), \qquad \text{s.t.} \quad X^\top X = I,$$

Rayleigh-Ritz (RR) Refinement: $[V, D] = \text{eig}(X^\top AX); X = X * V;$

- SSI: $X^{(i+1)} = \mathbf{orth}(AX^{(i)})$
- LOBPCG:
  $X^{(i+1)} = \mathbf{argmax}\{f(X) \mid X^\top X = I,\ X \in \{X^{(i-1)}, X^{(i)}, AX^{(i)}\}\}$
- LMSVD: adaptive block Krylov subspace method
  L.-Wen-Zhang 2013, SIAM Journal on Scientific Computing

Avoid the Bottleneck

Two Main Types of Operations AX & $\mathrm{RR/orth}$

As k increases, $AX \ll \mathrm{RR/orth} \longrightarrow$ bottleneck

Parallel Scalability

- $AX \longrightarrow Ax_1 \cup Ax_2 \cup \ ... \cup Ax_k$. Higher.
- $\mathrm{RR/orth}$ inherits sequentiality. Lower.

Avoid Bottleneck

- Do less RR/orth

No Free Lunch

- Do more BLAS3 ($X^\top X$, higher scalability than AX)

Trace Penalty Model

Wen-Yang-L.-Zhang 2016, Journal of Scientific Computing

Trace Minimization

$$\min_{X \in \mathbb{R}^{n \times k}} \mathbf{tr}(X^\top AX) \quad \text{s.t.} \quad X^\top X = I. \tag{1}$$

Trace-penalty Minimization [1]

$$\min_{X \in \mathbb{R}^{m \times k}} f(X) := \frac{1}{2}\mathbf{tr}(X^\top AX) + \frac{\mu}{4}\|X^\top X - I\|_F^2. \tag{2}$$

This idea (quadratic penalty) is old (Courant 1943) and unsophisticated.

It is well known that $\mu \to \infty$, (2) $\Longrightarrow$ (1)

Not a Good Penalty Function

[1] Related unconstrained model: $\min_{X \in \mathbb{R}^{m \times k}} \frac{1}{4}\|X^\top X\|_F^2 + \frac{1}{2}\mathbf{tr}(X^\top AX)$, Jiang-Cui-Dai 2014

EIGPEN

“Exact” Penalty

- $\mu \to \infty$ is neither needed, nor desirable
- $\mu > \max(0, \lambda_k)$ yields equivalence: same eigenspace

Fewer Saddle Points

- If $\mu \in (\lambda_k, \lambda_n)$: unique minimum, no maximum
- If $\mu \in (\lambda_k, \lambda_{k+p})$[2]: all rank-$k$ stationary points are minimizers

EIGPEN

- Gradient method, BB stepsize
- OpenMP parallelization on Cray XE6 supercomputer (NERSC)
- Restart: RR refinement

[2] λ_{k+p} is the smallest eigenvalue $> \lambda_k$.

Condition Number

- $k = 1$: $\kappa(\nabla^2 f_\mu(\hat{X})) = \frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1}$
- $k > 1$: $\kappa(\nabla^2 f_\mu(\hat{X})) = \infty$
- "Restricted" condition number:

$$\kappa\left(\nabla^2 f_\mu(\hat{X})\Big|_{Q_k^\perp}\right) \triangleq \frac{\max\limits_{S\in\mathbb{R}^{n\times k}} \left\{\mathrm{tr}(S^\top \nabla^2 f_\mu(\hat{X})(S)) : \mathrm{tr}(S^\top S) = 1, S^\top Q_k = 0\right\}}{\min\limits_{S\in\mathbb{R}^{n\times k}} \left\{\mathrm{tr}(S^\top \nabla^2 f_\mu(\hat{X})(S)) : \mathrm{tr}(S^\top S) = 1, S^\top Q_k = 0\right\}} = \frac{\lambda_n - \lambda_1}{\lambda_{k+1} - \lambda_k}.$$

- Explain the necessity of restart procedure

Symmetric Low Rank Product

L.-Wen-Zhang 2015, SIAM Journal on Optimization

New Model – SLRP

$$\min_{X\in\mathbb{R}^{n\times k}} \|XX^\top - A\|_F^2$$

- $\lambda_k \geq 0$ yields equivalence, same eigenspace
- Nonlinear least squares $\rightarrow$ Gauss-Newton method
- rank deficient normal equation
- minimum weighted-norm solution:

$$S_0 = \left(I - \frac{1}{2}\mathcal{P}_X\right)\left(AX(X^\top X)^{-1} - X\right)$$

where $\mathcal{P}_X = X(X^\top X)^{-1}X^\top$

SRLPGN

Two versions

- Theoretical version
  - correction step: if $\sigma_{\min}(X) < \delta$, set $X^+ = X + P$
  - $X^+ = X + \alpha S_0$ with $\alpha = \min\left(1, \sigma^3_{\min}(X)/\|\nabla f(X)\|_{\mathrm{F}}\right)$
  - global convergence
  - no correction + unit step size after finite number of iterations
- Practical version
  - two lines: $Y = X(X^\top X)^{-1}, \quad X = AY - X(Y^T AY - I)/2$
  - local linear convergence rate λ_{k+1}/λ_k

Further Developments

- Augmented Rayleigh-Ritz and Block Iteration (ARRABIT):
  Wen-Zhang 2017, SIMAX
- Triangularized Orthogonalization-Free Method (TriOFM):
  Gao-Li-Lu 2020, arXiv

Differentiable Objective $f(X)$

Discretized Kohn-Sham Total Energy Minimization

$$\min E(X) \quad \text{s. t.} \quad X^\top X = I, \quad X \in \mathbb{R}^{n\times p},$$

where, for $\rho(X) := \text{diag}(XX^\top)$,

$$E(X) := \frac{1}{4}\text{tr}(X^\top LX) + \frac{1}{2}\text{tr}(X^\top V_{ion}X) + \frac{1}{2}\sum_i \sum_l |x_i^\top w_l|^2 + \frac{1}{4}\rho^\top L^\dagger \rho + \frac{1}{2}e^\top \epsilon_{xc}(\rho).$$

1. Kinetic energy (L: finite dimensional representation of the Laplacian operator)
2. Local ionic potential energy (V_{ion}: ionic pseudopotentials sampled on the suitably chosen Cartesian grid)
3. Nonlocal ionic potential energy (w_l: discretized pseudopotential reference projection function)
4. Hartree potential energy ($L^\dagger$: pseudo-inverse of L)
5. Exchange correlation energy nonclassical (ϵ_{xc}: interaction between electrons)

Yang et. al. 2006; Gao et. al. 2009; Wen et. al. 2013, 2015; Zhou et. al. 2014, Ulbrich et. al. 2015, Jin et. al. 2015; Dai et. al. 2018, ...

Existing Methods

First-order Methods Based on Retraction

- Steepest descent: Helmke-Moore 1994; Udriste 1994
- Conjugate gradient: Edelman-Arias-Smith 1998; Brace-Manton 2006; Smith 1994; Gallivan-Absil 2010
- Geodesic search in canonical metric: Abrudan-Eriksson-Koivunen 2008
- Cayley transformation: Nishimori-Akaho 2005
- Projection-based method: Manton 2002; Absil-Mahony-Sepulchre 2008; Dai-Zhang-Zhou 2019
- Constraint preserving update scheme: Wen-Yin 2012; Jiang-Dai 2014

Second-order Methods Based on Retraction

- Newton: Smith 1994; Edelman-Arias-Smith 1998
- Quasi-Newton: Edelman-Arias-Smith 1998; Brace-Manton 2006; Gallivan-Absil 2010; Huang-Gallivan-Absil 2015
- Structured Quasi-Newton: Hu-Jiang-Lin-Wen-Yuan 2018
- Regularized Newton: Hu-Wen-Milzarek-Yuan 2017
- Trust region: Absil-Baker-Gallivan 2007

Other Type of Methods

- Splitting and alternating: Lai-Osher 2014

Hu-L.-Wen-Yuan 2020, Journal of the Operations Research Society of China

A New First-order Framework

Gao-L.-Chen-Yuan 2018, SIAM Journal on Optimization

Motivations: considering optimization problems with orthogonality constraints in **Euclidean space**.

- A different angle
- Compatibility
- Parallelization?

Observation: for any $X \in \mathcal{S}_{n,p}$, it holds that

$$\|\nabla f(X) - X\nabla f(X)^\top X\|_{\mathrm{F}}^2 = \|\nabla f(X) - XX^\top \nabla f(X)\|_{\mathrm{F}}^2 + \|X^\top \nabla f(X) - \nabla f(X)^\top X\|_{\mathrm{F}}^2.$$

A Two-step Feasible First-order Framework

- Step I: sufficient function value reduction
- Step II: rotation to keep the symmetry of multipliers
- Assumption: $f(X) = h(X) + tr(G^\top X)$, $h(X)$ — orthogonal invariant

Algorithm Framework

1) Set tolerance $\epsilon > 0$, initialize: $X^0 \in \mathcal{S}_{n,p}$, set $k := 0$;
2) Find a feasible point $\bar{X}$, based on X^k, satisfying sufficient function value reduction

$$f(X^k) - f(\bar{X}) \geq C_1 \cdot \|(I - X^k X^{k^\top})\nabla f(X^k)\|_{\mathrm{F}}^2;$$

3) Calculate X^{k+1} as follows

$$X^{k+1} := \begin{cases} \bar{X}, & \text{if } \bar{X}^\top G = G^\top \bar{X}; \\ -\bar{X} U T^\top, & \text{otherwise,} \end{cases}$$

where $U \Sigma T^\top$ is the singular value decomposition of $\bar{X}^\top G$;
4) If $\|(I - X^k X^{k^\top})\nabla f(X^k)\|_{\mathrm{F}}^2 < \epsilon$, return X^{k+1}; Otherwise, set $k := k + 1$ and go to Step 2.

Properties of the Framework

Contributions

- No searching in manifold or its tangent space
- Three algorithms
  - gradient reflection
  - gradient projection
  - column-wise block coordinate descent
- Global convergence
  - convergence of GR or GP with a fixed stepsize
  - convergence of CBCD
- Satisfactory numerical performance

Deficiencies

- Remove the ugly assumption

  Wang-Gao-L., Multipliers Correction Methods for Optimization Problems with Orthogonality Constraints, upcoming
- Parallelization?

A Key Bottleneck When p Is Large

Orthonormalization — lacks of concurrency

Column-wise parallelization — lacks of scalability

Solution: infeasible method

- Key point: efficient in serial
- To keep the structure: penalty function method
- Nonsmooth penalty function is intractable

$$\min_{X \in \mathbb{R}^{n \times p}} \quad f(X) + \gamma \|X^\top X - I_p\|_1$$

Augmented Lagrangian Method (ALM)

Augmented Lagrangian penalty function Powell 1969; Hestenes 1969

$$\mathcal{L}(X,\Lambda) := f(X) - \frac{1}{2}\mathrm{tr}\left(\Lambda(X^\top X - I_p)\right) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.$$

- Exact penalty function

ALM with dual ascend

1. Choose an initial point X_0, Λ_0, $k = 0$
2. Update X_k by $X_{k+1} = \arg\min_X \ \mathcal{L}(X, \Lambda_k)$
3. Update Λ_k by $\Lambda_{k+1} = \Lambda_k - \tau_k\beta(X_{k+1}^\top X_{k+1} - I_p)$

- Solving subproblem with fixed multiplier
- Updating multiplier by dual ascent
- Numerically inefficient

Motivation: First-order Optimality

First-order Optimality

The first-order optimality conditions of (OCP) can be written as

$$\begin{cases} \nabla f(X) - X\Lambda &= 0; \\ X^\top X &= I. \end{cases}$$

Lagrangian multipliers enjoy the closed-form expression $\Lambda = \nabla f(X)^\top X$ at any first-order stationary point.

Updating Multipliers by Closed-form

$$\Lambda^{k+1} := \Phi(\nabla f(X^k)^\top X^k),$$

where $\Phi : \mathbb{R}^{n\times n} \mapsto \mathbb{S}^n$ is defined by $\Psi(A) := \frac{1}{2}(A + A^\top)$.

Explicit Multiplier Updating Scheme

Gao-L.-Yuan 2019, SIAM Journal on Scientific Computing

Proximal Linearized Augmented Lagrangian Method (PLAM)

- $\Lambda_k = \Phi(X_k^\top \nabla f(X_k))$, $\quad \Phi(M) = \frac{1}{2}(M + M^\top)$
- Taking one gradient step in the subproblem $\min \mathcal{L}(X, \Lambda_k)$:

$$X_{k+1} = X_k - \eta_k \nabla_X \mathcal{L}(X_k, \Lambda_k).$$

- Exact penalty, global convergence, local linear convergence
- Comparable with existent algorithms with subtly selected β

Column-wise Block Minimization of PLAM (PCAL)

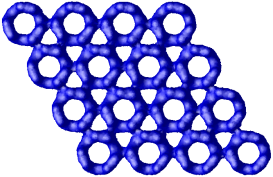
- Column-wise normalization:

$$(X_{k+1})_i = (X_k - \eta_k \nabla_X \mathcal{L}(X_k, \Lambda_k)_i / \|(X_k - \eta_k \nabla_X \mathcal{L}(X_k, \Lambda_k)_i\|_2$$

- Not sensitive with β
- Comparable with feasible algorithms
- Much better scalability in parallel computing

PCAL Applied in Electronic Structure Calculation

Gao-Hu-Kuang-L., Electronic Structure Calculation via a Parallelizable Framework without Orthogonalization, upcoming



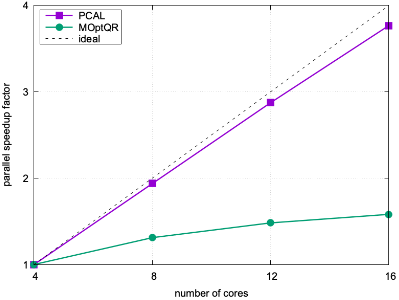


Figure 1: Calculation on AFEABIC: C_{384} with $n = 161322$, $p = 1152$. Left: the isosurface of the electron density at value 0.2. Right: the speedup factor for this example.

Still Not Perfect

Xiao-L.-Yuan 2020, A Class of Smooth Exact Penalty Function Methods for Optimization Problems with Orthogonality Constraints, arXiv

PLAM & PCAL

- Understanding of the merit function

$$h(X) := f(X) - \frac{1}{2}\mathrm{tr}\left(\Phi\left(X^\top \nabla f(X)\right)(X^\top X - I_p)\right) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_\mathrm{F}^2$$

- Why PCAL is better?
- Possible extension: second-order method?

Exact Penalty Function

$h(x)$ — Exact Penalty Function?

- If X^* is a first-order stationary point of (OCP), then

$$\nabla h(X^*) = 0.$$

- How about the other way round?

$$\nabla h(X) = 0 \Rightarrow X^\top X = I_p?$$

- A crucial issue:

$h(X)$ — Not Necessarily Bounded Below

- $f(X) = \frac{1}{4}\left\|X^\top X\right\|_{\mathrm{F}}^2$
- $h(X) = \frac{1}{4}\left\|X^\top X\right\|_{\mathrm{F}}^2 - 2\mathrm{tr}\left((X^\top X)^2(X^\top X - I_p)\right) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2$
- $\|X\|_{\mathrm{F}} \to +\infty \Rightarrow h(X) \to -\infty.$

A New Penalty Model

Restrict $h(X)$ in a Bounded Set

$$\min_{X \in \mathcal{M}} \quad h(X). \qquad \text{(PenC)}$$

- $\mathcal{M}$ is convex and compact, $\mathcal{S}_{n,p} \subset \mathcal{M}$
- Projection to $\mathcal{M}$ can be easily calculated

Possible Choices of $\mathcal{M}$:

- Ball ($\mathcal{B}$) with radius $K \geq \sqrt{p}$
- Convex hull of Stiefel manifold: $\{X \in \mathbb{R}^{n\times p} | \, \|X\|_2 \leq 1\}$
- Convex hull of Oblique manifold: $\{X \in \mathbb{R}^{n\times p} | \, \|(X)_i\|_2 \leq 1\}$
- ...

Assumptions

Assumption 1

$f(X)$ is differentiable, $\nabla f(X)$ is Lipschitz continuous.

- Does not imply the existence of $\nabla h(X)$

Assumption 2

$f(X)$ is twice continuous differentiable, and $\nabla^2 f(X)$ is Lipschitz continuous for each $X \in \mathbb{R}^{n\times p}$.

- Does not imply the existence of $\nabla^2 h(X)$

Constants

- $M_0 := \sup_{X\in\mathcal{M}} \max\{1, \|\nabla f(X)\|_{\mathrm{F}}\}, \qquad M_1 := \sup_{X\in\mathcal{M}} \max\{1, \|\Lambda(X)\|_{\mathrm{F}}\};$
- $C_1 := \sup_{X\in\mathcal{M}} \tilde{h}(X) - \inf_{X\in\mathcal{M}} \tilde{h}(X), \qquad L_1 := \sup_{X\in\mathcal{M}} \frac{1}{\|X-Y\|_2} \|\Lambda(X) - \Lambda(Y)\|_2;$
- $L_2 := \sup_{X\in\mathcal{M}, Y\in\mathbb{R}^{n\times p}} \limsup_{t\to 0} \frac{\|\nabla\tilde{h}(X+tY) - \nabla\tilde{h}(X)\|_{\mathrm{F}}}{t\|Y\|_{\mathrm{F}}}, \; M_3 := \sup_{X\in\mathcal{M}} \max\{1, \|\nabla^2 f(X)\|_{\mathrm{F}}\}.$

Properties of Penalty Model

First-order Stationarity

Theorem 1

Suppose Assumption 1 holds, $\beta \geq \max\left\{2pL_1, \frac{2p(M_0+M_1)}{3}\right\}$*, and* $\tilde{X}$ *is a first-order stationary point of* (PenC) *, then either* $\tilde{X}^\top \tilde{X} = I_p$ *holds, which further implies that* $\tilde{X}$ *is a first-order stationary point of problem (OCP), or the inequality* $\sigma_{\min}(\tilde{X}^\top \tilde{X}) \leq \frac{M_1+2L_1}{\beta}$ *holds.*

Lemma 1

For any $0 < \delta \leq \frac{1}{3}$ *and* $\beta \geq \max\left\{2pL_1, \frac{2p(M_0+M_1)}{3}, 3M_1 + 6L_1, \frac{2C_1}{\delta^2}\right\}$*, we have*

$$\sup_{\|X^\top X - I_p\|_{\mathrm{F}} \leq \delta} h(X) < \inf_{\|X^\top X - I_p\|_{\mathrm{F}} \geq 2\delta} h(X).$$

Moreover, any global minimizer X^* *of (PenC) satisfies* $X^{*\top} X^* = I_p$ *which further implies that it is a global minimizer of problem* (OCP) *.*

Properties of Penalty Model (Cont'd)

Second-order Stationarity

Theorem 2

Suppose Assumption 2 holds, and $\mathcal{M}$ is chosen as $\mathcal{B} := \{X \in \mathbb{R}^{n\times p} \mid \|X\|_{\mathrm{F}} \leq K, K > \sqrt{p}\}$ and $\beta \geq \max\left\{6M_1 + 12L_1, \frac{2}{3}p(M_0 + M_1), 2L_2 + 1, \frac{2(L_2+pM_1)}{3}\right\}$. Then, any second-order stationary point $\tilde{X}$ of (PenC) satisfies $\tilde{X}^\top \tilde{X} = I_p$. Moreover, $\tilde{X}$ is a second-order stationary point of problem (OCP) .

Algorithm

Problem Reformulation

$$\begin{array}{ll} \min & f(X) \\ \text{s.t.} & X^\top X = I_p, \end{array} \quad \Rightarrow \quad \min_{X \in \mathcal{B}} \quad h(X).$$

- $K > \sqrt{p}$

Gradient of $h(X)$

$$\nabla h(X) = \nabla f(X) - X\Lambda(X) + \beta X(X^\top X - I_p)$$
$$-\frac{1}{2}\Big(\nabla f(X)(X^\top X - I_p) + \nabla^2 f(X)[X(X^\top X - I_p)]\Big)$$

Denote $G(X) := \nabla f(X) - X\Lambda(X) + \beta X(X^\top X - I_p)$

- Exact gradient involves $\nabla^2 f(X)$, unaffordable
- Omit the red part

First-order Method for Penalty Model with Compact Convex Constraints (PenCF)

1. Choose an initial guess X_0, set $k = 0$;
2. Compute $\Lambda(X_k)$;
3. Update X_k by

$$X_{k+1} = X_k - \eta_k(\nabla f(X_k) - X_k\Lambda(X_k) + \beta X_k(X_k^\top X_k - I_p));$$

4. If $\|X_{k+1}\|_\mathrm{F} \geq K$, project X_{k+1} back to $\mathcal{B}$;
5. If certain stopping criterion is satisfied, return X_{k+1}; Otherwise, set $k := k + 1$ and go to Step 2.

Global Convergence

Theorem 3

Suppose Assumption 1 holds, $\delta \in \left(0, \frac{1}{3}\right]$, $K \geq \sqrt{p + \delta\sqrt{p}}$, *and* $\beta \geq \max\left\{2pL_1, \frac{2}{3}p(M_0 + M_1),\ \ 3M_1 + 6L_1\right\}$. *Let* $\{X_k\}$ *be the iterate sequence generated by PenCF, starting from any initial point* X_0 *satisfying* $\|X_0^\top X_0 - I_p\|_{\mathrm{F}} \leq \delta$, *and the stepsize* $\eta_k \in \left[\frac{1}{2}\bar{\eta}, \bar{\eta}\right]$, *where* $\bar{\eta} = \min\left\{\frac{\delta}{8KM_4}, \frac{\beta\delta^2}{9K^2L_1M_4^2}, \frac{1}{45(L_0+M_1)+137\beta}\right\}$, $M_4 = M_0 + M_1K + \beta\delta K$. *Then, the iterate sequence* $\{X^k\}$ *has at least one cluster point, and each cluster point of* $\{X^k\}$ *is a stationary point of (OCP). More precisely, for any* $k \geq 1$, *it holds that*

$$\min_{0\leq i\leq N-1} \max\left\{\|X_i^\top X_i - I_p\|_{\mathrm{F}}, \|G(X_i)\|_{\mathrm{F}}\right\} \leq \max\left\{\frac{2\sqrt{3}}{3M_1}, 1\right\} \cdot \sqrt{\frac{5C_1 + \frac{5}{4}\beta\delta^2}{N\bar{\eta}}}.$$

Local Convergence Rate

Theorem 4

Suppose Assumption 2 holds, X^* is an isolated local minimizer of (OCP), and we denote

$$\tau := \inf_{Y^\top X^* + X^{*\top} Y = 0} \frac{\nabla^2 f(X^*)[Y,Y] - \mathrm{tr}\left(Y^\top Y \Lambda(X^*)\right)}{\|Y\|_{\mathrm{F}}^2}. \qquad (3)$$

The algorithm parameters satisfy $\beta \geq \frac{1}{2}M_3 + \frac{\sqrt{3}M_0}{3} + \frac{1}{2}\tau$ and $\eta^k \in [\frac{\bar{\eta}}{2}, \bar{\eta}]$, where $\bar{\eta} \geq M_3 + \frac{2\sqrt{3}M_0}{3} + 2\beta$. Then, there exists $\varepsilon > 0$ such that starting from any X^0 satisfying $\|X^0 - X^*\|_{\mathrm{F}} < \varepsilon$, and the iterate sequence $\{X^k\}$ generated by PenCF converges to X^* Q-linearly.

PLAM and PCAL – Further Explanation

PLAM

- $\mathcal{M} = \mathbb{R}^{n \times p}$
- $h(x)$ is not bounded below: small $\beta \Rightarrow$ divergence

PCAL

- $\mathcal{M} = \mathcal{OB}_{n,p}$
- (PenC) is bounded below: accept smaller β

PenCF

- $\mathcal{M} = \{X \in \mathbb{R}^{n \times p} \mid \|X\|_{\mathrm{F}} \leq K\} \Rightarrow$ cheap projection
- Constraint becomes inactive when close to $\mathcal{S}_{n,p}$

Both PLAM and PCAL can be regarded as applying approximate gradient method to solve corresponding (PenC) .

- better than ALM
- Comparable with existing retraction-based first-order methods

Post-process by Orthonormalization

Why Post-process?

- To attain high accuracy on feasibility
- To maintain mild accuracy on the substationarity

How to Post-process?

- $X^k_{\mathrm{orth}} := UV^\top$, economy-size SVD: $X^k = U\Sigma V$

Proposition 1

Suppose Assumption 1 holds, $\beta \geq 1 + 2L_0 + 2L_1 + 2M_1$ and $X \in \mathcal{M}$. Let $X = U\Sigma V^\top$ be the economy-size SVD for X and $\mathrm{orth}(X) = UV^\top$. Then, it holds that

$$h(X_{\mathrm{orth}}) \leq h(X) - \frac{1}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.$$

Second-order Methods

Suppose computing $\nabla^2 f(X)$ is affordable

- Computing $\nabla h(X)$ becomes affordable:

$$\begin{aligned}\nabla h(X) = \nabla f(X) - X\Lambda(X) + \beta X(X^\top X - I_p) \\ - \left(\nabla f(X)(X^\top X - I_p) + \nabla^2 f(X)[X(X^\top X - I_p)]\right)\end{aligned}$$

- Computing $\nabla^2 h(X)$ is still intractable
- Solution: approximate $\nabla^2 h(X)$ by ∇f and $\nabla^2 f$

Motivation: Delete High-order Terms in Hessian

$$
\begin{aligned}
\nabla^2 h(X)[D,D] = &\nabla^2 f(X)[D,D] \\
&- \mathrm{tr}\left(\Lambda(X)D^\top D - D^\top \nabla f(X)\Phi(D^\top X) - X^\top \nabla^2 f(X)[D]\Phi(D^\top X)\right) \\
&- \frac{1}{2}\mathrm{tr}\left(\left(D^\top \nabla^2 f(X)[D] + \frac{1}{2}X^\top \nabla^3 f(X)[D,D]\right)(X^\top X - I_p)\right) \\
&+ \mathrm{tr}\left(\beta X^\top X D^\top D + \beta D^\top X X^\top D + \beta X^\top D X^\top D - \beta D^\top D\right).
\end{aligned}
$$

$$
\begin{aligned}
W(X)[D,D] := &\nabla^2 f(X)[D,D] \\
&- \mathrm{tr}\left(\Lambda(X)D^\top D - D^\top \nabla f(X)\Phi(D^\top X) - X^\top \nabla^2 f(X)[D]\Phi(D^\top X)\right) \\
&+ \mathrm{tr}\left(\beta X^\top X D^\top D + \beta D^\top X X^\top D + \beta X^\top D X^\top D - \beta D^\top D\right).
\end{aligned}
$$

- $\left\|W(X) - \nabla^2 h(X)\right\|_{\mathrm{F}} \to 0$ as $\left\|X^\top X - I_p\right\|_{\mathrm{F}} \to 0$

Second-order Method

Subproblem

$$\begin{aligned} \min \quad & \frac{1}{2} W(X_k)[D, D] + \mathrm{tr}\left(D^\top \nabla h(X_k)\right) \qquad \text{(TRS)} \\ \text{s.t.} \quad & \|X_k + D\|_\mathrm{F} \leq K. \end{aligned} \tag{4}$$

- Trust region subproblem: computing global minimizer is tractable
- $\nabla h(X)$ is sufficiently small $\Rightarrow$ inactive constraint

Second-order Method for Penalty Model with Compact Convex Constraints (PenCS)

1. Choose an initial guess X_0, set $k = 0$;
2. Compute stepsize η_k;
3. Compute D_k by solving (TRS), set $X_{k+1} = X_k + \eta_k D_k$;
4. If certain stopping criterion is satisfied, return X_{k+1}; Otherwise, set $k := k + 1$ and go to Step 2.

Theoretical Results

Theorem 5

Suppose Assumption 2 holds. X^* *is an isolated local minimizer of (OCP) with*

$$\tau := \inf_{Y^\top X^* + X^{*\top} Y = 0} \frac{\nabla^2 f(X^*)[Y, Y] - \mathrm{tr}\left(Y^\top Y \Lambda(X^*)\right)}{\|Y\|_{\mathrm{F}}^2}. \tag{5}$$

When $\delta \in (0, \frac{1}{3})$, $K \geq \sqrt{p + \delta\sqrt{p}}$, $\beta \geq$
$\max\left\{2pL_1, \frac{2}{3}p(M_0 + M_1), 6M_1 + 12L_1, \frac{2(L_2+pM_1)}{3}, 2L_2 + 1, \frac{4L_2^2}{\tau} + \tau\right\}$
and stepsize $\eta_k = 1$*, there exists a sufficiently small* ε *such that when* $\|X_0 - X^*\|_{\mathrm{F}} \leq \varepsilon$*,* X_k *generated by PenCS converges to* X^* *quadratically.*

Nondifferentiable Objective $f(X)$

Sparse Variable PCA

Ulfarsson-Solo 2008, Chen-Zou-Cook 2010

$$\min_{X \in \mathbb{R}^{n \times p}} \quad -\frac{1}{2}\mathrm{tr}\left(X^\top M X\right) + \sum_{j=1}^{n} \gamma_j \left\|X_{j\cdot}\right\|_2$$

$$\text{s.t.} \quad X^\top X = I_p,$$

where M is the covariance matrix and γ_j are parameters to regularization term.

- Eliminate the variables that contains little information of the principle component of the distribution

Regularized Discriminative Feature Selection

Yang-Shen-Ma-Huang-Zhou 2011, Tang-Liu 2012

$$\min_{X \in \mathbb{R}^{n \times p}} \quad \frac{1}{2}\mathrm{tr}\left(X^\top M X\right) + \sum_{j=1}^{n} \gamma_j \left\|X_{j\cdot}\right\|_2$$
$$\text{s.t.} \quad X^\top X = I_p,$$

where M is constructed by the given samples.

- Unsupervised feature selection, linear classifier $X \in \mathbb{R}^{n \times p}$
- $X(j,:)$ are zero $\rightarrow$ j-th labels is ignored in classifying

$\Rightarrow$ Select the most representative labels.

$\ell_{2,1}$ Norm Regularization Minimization with Orthogonality Constraints

General Form

$$\min_{X\in\mathbb{R}^{n\times p}} \quad f(X) := F(X) + R(X)$$
$$\text{s.t.} \quad X^\top X = I. \qquad \text{(OCPR)}$$

- $F : \mathbb{R}^{n\times p} \longmapsto \mathbb{R}$, differentiable
- $R(X) = \sum_{j=1}^{n} \gamma_j \|X(j,:)\|_2$, $X_{j\cdot} := X(j,:)^\top, X_{\cdot i} = X(:,i)$
- $n > p$
- $p(p+1)/2$ constraints –– nonconvex
- Stiefel manifold:

$$\mathcal{S}_{n,p} := \{X \in \mathbb{R}^{n\times p} | \, X^\top X = I\}.$$

Existing Approaches

Subgradient Methods

- Subgradient method on Riemann manifold: Ferreira-Oliveria 1998
- ε subgradient method: Grohs-Hosseini 2016
- Gradient sampling method: Hosseini-Uschmajew 2017
- ......

Not fully exploit the composite structure $\Rightarrow$ inefficient

Existing Approaches (Cont'd)

ADMM-based Proximal Gradient Methods

- Splitting for orthogonality constrained problems (SOC): Lai-Osher 2014
- Manifold ADMM (MADMM): Kovnatsky-Glashoff-Bornstein 2016
- PAMAL: Chen-Ji-You 2016

Properties of These Approaches

- Simple subproblems
- Multiple-block alternating updating ⇒ usually not very efficient;
- Updating multiplier via dual-ascend ⇒ many parameters need to be tuned.

Existing Approaches (Cont'd)

Proximal Gradient Approaches

- Proximal gradient method on manifold (ManPG): Chen-Ma-So-Zhang 2020

$$\min_{D \in \mathcal{T}_{X^k}} \left\langle D, \nabla F(X^k) \right\rangle + R(X^k + D) + \frac{\|D\|_{\mathrm{F}}^2}{2\eta^k} \quad \text{(proximal mapping)}$$

- Alternating manifold proximal gradient method (AManPG): Chen-Ma-Xue-Zhou 2019

Properties of These Approaches

- No closed-form expression for proximal mapping in ManPG $\Rightarrow$ main bottleneck
- Orthonormalization process is required in each iteration $\Rightarrow$ lacks of scalability, hard for parallelism

First-order Optimal Condition

Definition 1

(Yang-Zhang-Song 2014, Chen-Ma-Xue-Zhou 2019)
A point $X \in \mathcal{S}_{n,p}$ is called as first-order stationary point of (OCPR) if and only if it satisfies

$$0 \in \mathcal{P}_{\mathcal{T}_X} \left(\nabla F(X) + \partial R(X) \right),$$

where $\mathcal{T}_X$ denotes the tangent space at X, $\mathcal{P}_{\mathcal{T}_X}(\mathcal{Y}) := \{Y - X\Phi(Y^\top X) \mid Y \in \mathcal{Y} \subseteq \mathbb{R}^{n\times p}\}$ consists of all the projection points of $Y \in \mathcal{Y}$ onto the tangent space $\mathcal{T}_X$, and ∂R stands for the Clarke subdifferential of R.

Equivalent Version

- There exists $D \in \partial R(X)$ and $\Lambda \in \mathbb{R}^{p\times p}$:

$$\begin{cases} X\Lambda = \nabla F(X) + D \\ \Lambda = \Lambda^\top \\ X^\top X = I_p \end{cases}$$

Motivation: Exact Penalty Model with Compact Convex Constraints

Xiao-L.-Yuan 2020, Exact Penalty Function for $\ell_{2,1}$ Norm Minimization over the Stiefel Manifold, upcoming

Differentiable Objective $\Lambda(X) = \Phi(X^\top \nabla f(X))$

$$\min\ f(X) \qquad \text{s.t. } X^\top X = I_p$$

$$\Longrightarrow \qquad \min_{X \in \mathcal{M}} f(X) - \frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2 .$$

$\ell_{2,1}$**-norm Regularized Minimization**

$$\Lambda(X) \in \Phi\left(X^\top \nabla F(X) + X^\top \partial R(X)\right)$$

- How to choose suitable $\Lambda(X)$ for (OCPR)?

Motivation: Explicit Expression

Expression for $\partial R(X)$

- $\partial R(X) = [\gamma^1 \partial(\|X_{1\cdot}\|_2), \gamma^2 \partial(\|X_{2\cdot}\|_2), ..., \gamma^n \partial(\|X_{n\cdot}\|_2)]^\top$
- $\partial(\left\|X_{j\cdot}\right\|_2) = \begin{cases} \frac{X_{j\cdot}^\top}{\|X_{j\cdot}\|_2}, & \text{if } \left\|X_{j\cdot}\right\|_2 \neq 0, \\ u_j \text{ satisfying } \left\|u_j\right\|_2 = 1 \,, & \text{otherwise.} \end{cases}$
- For any $D \in \partial R(X)$,

$$X^\top D = \sum_{i=1}^{n} \gamma_i S(X_{i\cdot}), \qquad \text{where } S(x) := \begin{cases} \frac{xx^\top}{\|x\|_2}, & \text{if } x \neq 0; \\ 0, & \text{otherwise.} \end{cases}$$

- $\lim_{x\to 0} S(x) = 0$, hence, if we denote $\frac{xx^\top}{\|x\|_2}\Big|_{x=0} = 0$, $S(x) = \frac{xx^\top}{\|x\|_2}$

Hence, arrive at

$$\Lambda(X) = \Phi(X^\top \nabla F(X)) + \sum_{i=1}^{n} \gamma^i S(X_{i\cdot})$$

- Closed-form expression;
- Lipschitz continuous.

New Penalty Function

$$h(X) := F(X) - \frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 + R(X),$$

where

$$\Lambda(X) = \Phi(X^\top \nabla F(X)) + \sum_{j=1}^{n} \gamma^i S(X_{i\cdot}).$$

- h is not bounded below
- Restrict h in a bounded set

$$\min_{X \in \mathcal{M}} \quad h(X). \qquad \text{(PenC)}$$

- $\mathcal{M}$ is a convex compact set, $\mathcal{S}_{n,p} \subset \mathcal{M}$

$\Rightarrow$ Projection to $\mathcal{M}$ can be easily calculated

Proximal Gradient Method for Solving PenC with Exact Lambda (PenCPG)

1. Choose an initial guess X^0, set $k = 0$;
2. Choose stepsize η^k;
3. Compute $D^k = \nabla F(X^k) + \beta X^k \left[({X^k}^\top X^k - I_p) - \Lambda(X^k) \right]$;
4. Update X^k by

$$X^{k+1} = \arg \min_{X \in \mathbb{R}^{n \times p}} \quad \left\langle X - X^k, D^k \right\rangle + R(X) + \frac{\left\| X - X^k \right\|_F^2}{2\eta^k}$$

5. If $\left\| X^{k+1} \right\|_F > K$, project X^{k+1} back to $\mathcal{B}$;
6. If certain stopping criterion is satisfied, return X^{k+1}; Otherwise, set $k := k + 1$ and go to Step 2.

Assumptions

Assumption 3

- $F(X)$ *is differentiable and* $\nabla F(X)$ *is Lipschitz continuous.*

Constants

- $M_0 := \sup_{X\in\mathcal{M}} \|\nabla F(X)\|_{\mathrm{F}}$
- $M_1 := \sup_{X\in\mathcal{M}} \|\Lambda(X)\|_2$
- $L_0 := \sup_{X,Y\in\mathcal{M}} \frac{\|\nabla F(X)-\nabla F(Y)\|_{\mathrm{F}}}{\|X-Y\|_{\mathrm{F}}}$
- $L_1 := \sup_{X\in\mathcal{M},Y\in\mathcal{M}} \frac{\|\Lambda(X)-\Lambda(Y)\|_{\mathrm{F}}}{\|X-Y\|_{\mathrm{F}}}$
- $L_r := \sum_{i=1}^{n} \gamma^i$

Global Convergence

Theorem 6

Suppose Assumption 3 holds. For any given $\delta \leq \frac{1}{3}$ *and* $K \geq \sqrt{p + 4\sqrt{p}\delta}$, *suppose* $\beta \geq \max\{6M_1, 36L_1, 2(M_0 + M_1), \frac{2C_1}{\delta^2}\}$, *PenCPG starts with* X^0 *satisfying* $\left\| {X^0}^\top X^0 - I_p \right\|_{\mathrm{F}} \leq \frac{\delta}{2}$ *and uses the stepsize* $\eta^k \in [\frac{1}{2}\eta^+, \eta^+]$ *where*
$\eta^+ = \min\left\{ \frac{1}{L_0+3\beta+2L_1+M_1}, \frac{1}{20(M_0+2M_1+\beta+L_r)}, \frac{1}{4(L_0+3\beta+M_1)} \right\}$. *Then any accumulation point of* $\{X^k\}$ *is the first-order stationary point of (OCPR).*
Moreover, for any $N \geq 0$,

$$\min_{0\leq k\leq N} \frac{\left\| X^{k+1} - X^k \right\|_{\mathrm{F}}}{\eta^k} \leq \sqrt{\frac{16C_1 + \beta\delta^2}{2N\eta^+}}.$$

Comparison on Computational Complexity

ManPG		
Computing gradient	$\nabla f(X^k)$	1 first-order oracle
Computing Riemann gradient	$\nabla f(X^k) - X^k \Phi({X^k}^\top \nabla f(X^k))$	$4np^2$
Retraction[3]	$qr(X^{k+1})$	$2np^2$
SSN for proximal subproblem[4][5]	$E(\Lambda^k)\mathrm{vec}(\Lambda^{k+1}) = -D(\Lambda^k)$	$2np^2 \cdot l_{CG}$
total	1 first-order oracle + $6np^2 + 2np^2 \cdot l_{CG}$	
PenCPG		
Computing gradient	$\nabla f(X^k)$	1 first-order oracle
Computing D^k	D^k	$6np^2$
Solving subproblem	thresholding	$2np$
Retraction	no retraction in PenCPG	0
total	1 first-order oracle + $6np^2$	

[3] Grad-Schmidt orthonormalization

[4] l_{CG} denotes the total iterations in CG method for solving the linear system, and ManPG can take multiple SSN steps in each iteration.

[5] $l_{CG} \gg 1$.

Post-process by Orthonormalization

Projection After Convergence

- Obtain X^k, and economy-size SVD decomposition $X^k = U\Sigma V^\top$;
- Return $X_{\text{orth}} := UV^\top$.

Theoretical Analysis for Post-process

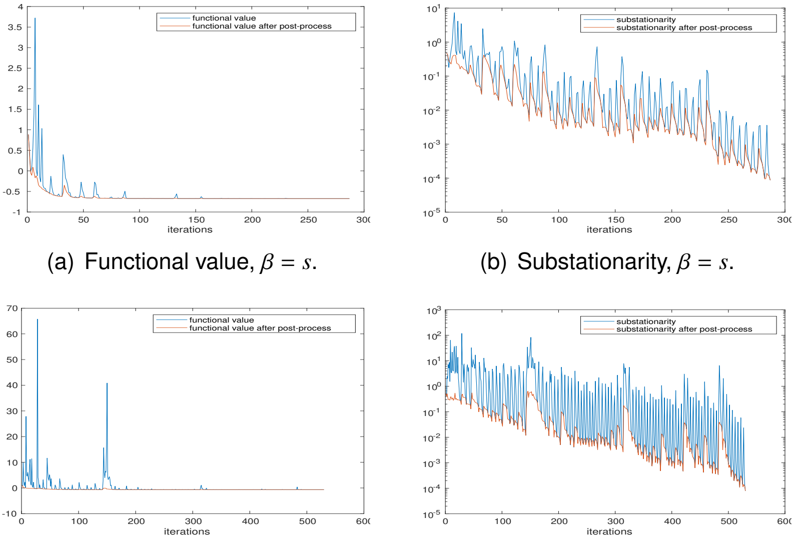
Proposition 2

Given $\delta \in \left(0, \frac{1}{3}\right]$, $K \geq \sqrt{p + \delta\sqrt{p}}$. Then for
$\beta \geq \max\{6M_1, 36L_1, 2(M_0 + M_1), \frac{2C_1}{\delta^2}, 2\,(L_0 + L_1 + 3M_1 + 2M_2)\}$.
Suppose PenCPG starts with initial value $\left\| {X^0}^\top X^0 - I_p \right\|_{\mathrm{F}} \leq \frac{\delta}{2}$ with stepsize $\eta^k \in [\frac{1}{2}\eta^+, \eta^+]$, and generates a sequence $\{X^k\}$. Then for any $k > 0$, let X^k has compact SVD $X^k = U^k \Sigma^k {V^k}^\top$ and define $X_{orth} = U^k {V^k}^\top$, then

$$h(X^k) \geq h(X_{\text{orth}}). \tag{6}$$

- Decrease feasibility violation
- Reduce function value

Numerical Experiments for Post-process



(a) Functional value, $\beta = s$.

(b) Substationarity, $\beta = s$.

(c) Functional value, $\beta = 10s$.

(d) Substationarity, $\beta = 10s$.

Figure 2: Problem 1 with $n = 500$, $p = 4$, $\gamma = 0.09$.

Experiments for Sparse Variable PCA

Problem Generation

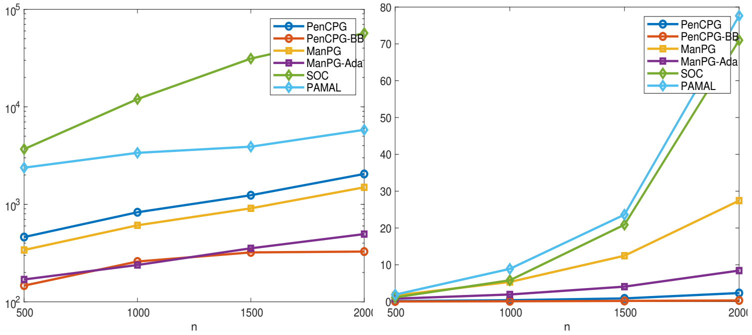
- Randomly generates uniformly distributed samples, $N = 200$, and set A as their covariance matrix; $\gamma^i = \frac{b}{2}\sqrt{\log(2n)/50}$
- Gao-Ma-Zhou 2017, Chen-Ma-Xue-Zhou 2019
- 10 experiments with random initial point

Testing Algorithms

Testing Methods

- SOC: Splitting method for orthogonality constrained problems, Lai-Osher 2014 ;
- PAMAL: Proximal alternating minimized augmented Lagrangian, Chen-Ji-You 2016 ;
- ManPG : Manifold proximal gradient method, Chen-Ma-So-Zhang 2020 ;
- ManPG-Ada: Accelerated version of ManPG, Chen-Ma-So-Zhang 2020 ;
- PenCPG: $\beta = s, K = 10\sqrt{p}$, fixed stepsize $\eta^k = \frac{1}{2s}$;
- PenCPG-BB: PenCPG with BB1 stepsize, $\beta = s, K = 10\sqrt{p}$.

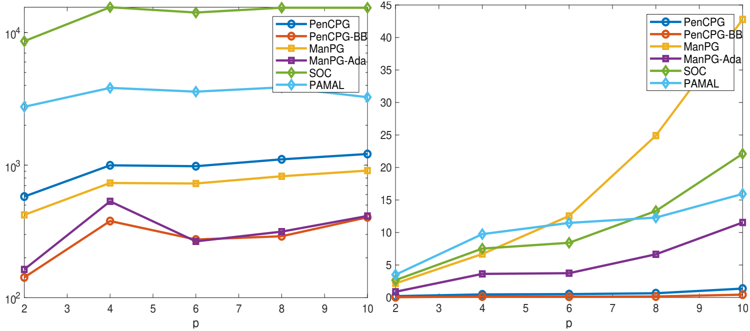
Numerical Results on n



(a) Iterations (b) CPU time

- Comparable in the aspect of iterations
- Less CPU time

Numerical Results on p



(a) Iterations (b) CPU time

- Cheap proximal mapping $\Rightarrow$ High scalability.

Extensions

Sparse Canonical Correlation Analysis Gao-Ma-Zhou 2017, Chen-Ma-Xue-Zhou 2019

$$\min_{X,Y\in\mathbb{R}^{n\times p}} \quad -\operatorname{tr}\left(X^\top B_1^\top B_2 Y\right) + \gamma_1 \|X\|_{2,1} + \gamma_2 \|Y\|_{2,1}$$
$$\text{s.t.} \quad X^\top B_1 B_1^\top X = I_p,\ Y^\top B_2 B_2^\top Y = I_p.$$

- B_1, B_2 contain correlated samples.
- $\gamma_1, \gamma_2 \geq 0$ control the sparsity of X and Y.
- $B_1 B_1^\top$ and $B_2 B_2^\top$ can be singular.

Exact Penalty Function for Sparse CCA

Exact Penalty Function

$$
\begin{aligned}
h(X,Y) \quad := \quad & -\mathrm{tr}\left(X^\top B_1^\top B_2 Y\right) - \frac{1}{2}\left\langle \Lambda_1(X), X^\top B_1 B_1^\top X - I_p \right\rangle \\
& + \gamma_1 \|X\|_{2,1} + \gamma_2 \|Y\|_{2,1} - \frac{1}{2}\left\langle \Lambda_2(Y), Y^\top B_2 B_2^\top Y - I_p \right\rangle \\
& + \frac{\beta}{4}\left\|X^\top B_1 B_1^\top X - I_p\right\|_\mathrm{F}^2 + \frac{\beta}{4}\left\|Y^\top B_2 B_2^\top Y - I_p\right\|_\mathrm{F}^2
\end{aligned}
$$

- $\Lambda_1(X) := -\Phi(X^\top B_1^\top B_2 Y)$
- $\Lambda_2(Y) := -\Phi(X^\top B_1^\top B_2 Y)$

Properties

- Bounded below

Experiments for Sparse CCA

Problem Generation

- Souo-Minden-Nelson-Tibshirani-Saunders 2017

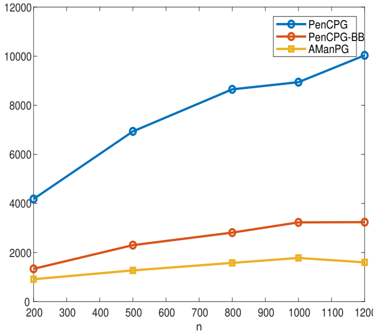
Testing Methods

- A-ManPG: Alternating direction descent version of ManPG, Chen-Ma-Xue-Zou 2019
- PenCPG with constant stepsize & BB1 stepsize and $\beta = s, K = 100\sqrt{p}$.

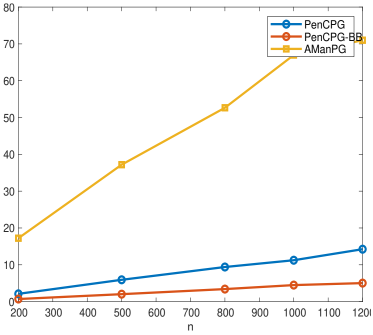
Stopping Criteria

- $\max\left\{\frac{\|X^{k+1}-X^k\|_F}{\eta^k}, \frac{\|Y^{k+1}-Y^k\|_F}{\eta^k}\right\} \leq 10^{-4}$ for PenCPG and A-ManPG.

Numerical Results on n



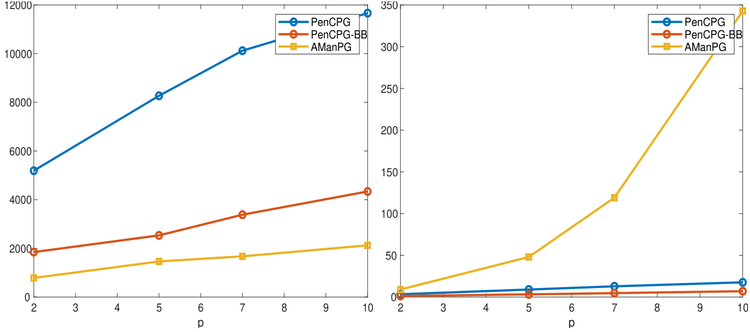
(a) Iterations



(b) CPU time

- Comparable in the aspect of iterations
- Less CPU time

Numerical Results on p



(a) Iterations (b) CPU time

- Cheap proximal mapping ⇒ high scalability.

Extensions

Sparse Dictionary Learning

$$\min_{W \in \mathbb{R}^{n \times p}} \quad -\frac{1}{m} \left\| W^\top Y \right\|_m^m$$
$$\text{s.t.} \quad W^\top W = I_p,$$

- $Y \in \mathbb{R}^{n \times N}$ is a given data matrix
- $\|Y\|_m = \left[\sum_{i=1}^n \sum_{j=1}^N (Y_{ij})^m \right]^{\frac{1}{m}}$ with constant $m \in (2, 4]$

Exactly Penalty Function   Hu-L. 2020, Sensors

$$h(W) := f(W) - \frac{1}{2} \left\langle W^\top W - I_p, \Phi(W^\top \nabla f(W)) \right\rangle + \frac{\beta}{6} \|W\|_{\mathrm{F}}^6 - \frac{\beta}{2} \|W\|_{\mathrm{F}}^2$$

Other Extensions

- Graph clustering: combining with other constraints
- Sparse principal component analysis: ℓ_1 norm minimization

Emerging New Demands

Distributed and Secure Singular Value Decomposition

Wang-L.-Zhang, A Distributed and Secure Algorithm for Dominant Singular Value Decomposition, upcoming

The matrix A is divided into s column blocks:

$$A = [A_1, \cdots, A_s],$$

where $A_i \in \mathbb{R}^{n \times m_i}$ $(i = 1, \cdots, s)$ and $m_1 + \cdots + m_s = m$

Distributed Memory

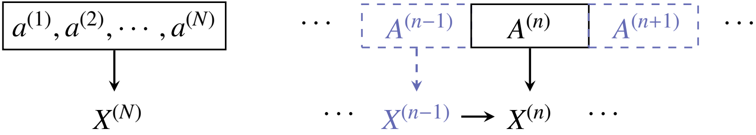
- Sub-matrices A_i distributedly stored in s cores
- Each core: access to local data A_i
- Ultimate goal: obtain the dominant SVD of the whole matrix A

Security/Privacy

- Each core: no sharing A_i or $A_i A_i^\top$ with others
- Each core: no chance to recover A or $AA^\top$ by solving an inverse problem

Online Principal Component Analysis

Zhou-L.-Xu, A Stochastic Gauss-Newton Method for Online Principal Component Analysis, upcoming



(a) Traditional PCA (Offline PCA)

(b) Online PCA ($|A^{(i)}| = h$)

$$\min_{X \in \mathbb{R}^{d \times k}} \quad -\mathrm{tr}\left(X^\top \left(\frac{1}{N}\sum_{i=1}^{N} a^{(i)} a^{(i)\top}\right) X\right)$$

$$\text{s.t.} \quad X^\top X = I_k.$$

- High dimension d
- large volume N
- Solution: dynamically stored samplings – forget old ones

Conclusions

Contributions

- Exact penalty framework for optimization problems with orthogonality constraints
  - eigenspace calculation
  - differentiable objective
  - $\ell_{2,1}$-norm minimization
  - other manifolds
- Orthonormalization-free
  - parallelizable
  - easy subproblem

Further Developments

- Extension to general nonsmooth cases
- Distributed and secure singular value decomposition
- Online principal component analysis


主办单位：
中国运筹学会
协办单位：
中国工业与应用数学学会
中国数学会
网络科普报告——
“六一”儿童节的礼物
胡旭东 研究员
胡旭东，中国科学院数学与系统科学研究院研究员，博士生导师，中国运筹学会理事长。1985年毕业于清华大学，获应用数学专业学士学位，1989年毕业于中国科学院应用数学研究所，获运筹与控制论专业博士学位。自1989年始，一直在中国科学院从事运筹学的理论研究和教学工作，主要研究方向为组合优化、网络博弈、近似算法。2012年被评为第五届“全国优秀科技工作者”，2016年获“中国科学院朱李月华优秀教师奖”。
运筹学—求解生活中的数学趣题
直播时间：5月31日 15:00-16:30
直播平台：Classin
扫码直接观看

Thanks for your attention!

Email: liuxin@lsec.cc.ac.cn

Homepage: http://lsec.cc.ac.cn/~liuxin/