# Finding a Large Approximately Rank-One Submatrix Using the Nuclear Norm and $\ell_1$ norm

Stephen Vavasis[1]

[1]Department of Combinatorics & Optimization
University of Waterloo

Parts of this talk represent joint work with X. V. Doan of Waterloo/Warwick and K.-C. Toh of N. U. Singapore

2011-Sept-28 / Fields Workshop on Optimization

# Finding a feature in a text dataset

Suppose one is given a *text corpus*, i.e., a collection of $n$ text documents, and one seeks a topic in the dataset, that is, a subset of related documents. One approach:

- Form the term-document matrix, that is, the $m \times n$ matrix in which $i$th row corresponds to the $i$th term, $j$th column to $j$th document, and $A(i,j)$ is the number of occurrences of term $i$ in document $j$.
- Find a large approximately rank-one submatrix $A(I,J)$ of $A$ (i.e., $A(I,J) \approx \mathbf{wh}^T$).

# Finding a feature in an image dataset

Given an image dataset in which all the $n$ contain exactly $m_1 \times m_2 \equiv m$ pixels, find a visual feature, that is, a particular pattern that recurs in the same subset of pixels in a subset of images.

- Form an $m \times n$ matrix $A$ in which $A(i, j)$ stands for the intensity of pixel $i$ in image $j$.
- Find a large approximately rank-one submatrix (LAROS) $A(I, J)$ of $A$.

# Outline of talk

- LAROS problem: relationship to NMF and SVD.
- Convex relaxation.
- Recovery
- Proximal point algorithm
- Computational experiment

# LAROS and NMF

- Assume $A$ is nonnegative.
- The above process can be repeated iteratively:

  For $i = 1 : k$

  Find $I_i, J_i, \bar{\mathbf{w}}_i, \bar{\mathbf{h}}_i$ s.t. $A(I_i, J_i) \approx \bar{\mathbf{w}}_i \bar{\mathbf{h}}_i^T$.

  Pad $(\bar{\mathbf{w}}_i, \bar{\mathbf{h}}_i)$ with zeros to obtain $(\mathbf{w}_i, \mathbf{h}_i)$.

  $A = \max(A - \mathbf{w}_i \mathbf{h}_i^T, 0)$.
- Upon completion,
  $A \approx \mathbf{w}_1 \mathbf{h}_1^T + \cdots + \mathbf{w}_k \mathbf{h}_k^T \equiv WH^T$.

# Greedy NMF algorithm

- OK to assume that $\mathbf{w}_i \geq \mathbf{0}$, $\mathbf{h}_i \geq \mathbf{0}$ (Perron-Frobenius).
- Given a nonnegative matrix $A$, a factorization $A \approx WH^T$ is called *nonnegative matrix factorization* (NMF) if $W, H$ both nonnegative.
- The algorithm on the previous transparency is a greedy NMF algorithm (Asgarian & Greiner, Bergmann et al., Biggs et al., Gillis & Glineur).

# LAROS and SVD

- Best overall rank-one approximation to $A$ comes from SVD (Eckart-Young theorem).

$$A = \begin{pmatrix} 0.8 & 0.9 & 0.0 & 0.0 \\ 0.8 & 1.1 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.9 \\ 0.0 & 0.0 & 1.1 & 0.8 \end{pmatrix}.$$

- The dominant left singular vector is $\approx [1; 1; 0; 0]$; SVD has identified $A(1:2, 1:2)$.
- But with a little noise, dominant left singular vector $\approx [1; 1; 1; 1]$; SVD fails to identify LAROS.

# LAROS and SVD

- Best overall rank-one approximation to $A$ comes from SVD (Eckart-Young theorem).

$$A = \begin{pmatrix} 0.8 & 0.9 & \mathbf{0.1} & \mathbf{0.2} \\ 0.8 & 1.1 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.9 \\ 0.0 & 0.0 & 1.1 & 0.8 \end{pmatrix}.$$

- The dominant left singular vector is $\approx [1; 1; 0; 0]$; SVD has identified $A(1:2, 1:2)$.
- But with a little noise, dominant left singular vector $\approx [1; 1; 1; 1]$; SVD fails to identify LAROS.

# SVD as optimization

- The solution to this problem is to modify the SVD to promote sparsity.
- Can write SVD as an optimization problem (Eckart-Young) and add another term, i.e.,

$$\min_{\sigma,\mathbf{u},\mathbf{v}} \|A - \sigma \mathbf{u}\mathbf{v}^T\| + \text{densityPenalty}(\mathbf{u}, \mathbf{v})$$

- Unfortunately, Eckart-Young optimization problem is not convex.

# SVD as convex optimization

- Let $\|\cdot\|_*$ denote the *nuclear norm*, that is, $\|X\|_* = \sigma_1(X) + \cdots + \sigma_n(X)$.
- Theorem: The nuclear norm is dual to the 2-norm, i.e., $\|X\|_* = \max\{Z \bullet X : \|Z\|_2 \le 1\}$.
- Given $A$, the solution to the convex optimization problem $\min\{\|X\|_* : A \bullet X \ge 1\}$ is $X = \mathbf{u}_1\mathbf{v}_1^T/\sigma_1$, where $(\sigma_1, \mathbf{u}_1, \mathbf{v}_1)$ is the dominant singular triple of $A$.

# Obtaining a sparse solution

- In order to enforce sparsity, could add a (nonconvex) penalty term: $\min \|X\|_* + \pi(|I| \cdot |J|)$ s.t. $A \bullet X \geq 1; (i,j) \notin I \times J \Rightarrow X(i,j) = 0$. where $\pi(\cdot)$ is an increasing penalty function.
- The optimal $X$ will have necessarily have the form $X = \bar{\mathbf{u}}_1 \bar{\mathbf{v}}_1^T / \bar{\sigma}_1$, where $(\bar{\sigma}_1, \bar{\mathbf{u}}_1, \bar{\mathbf{v}}_1)$ is the dominant singular triple of $A(I,J)$ for some $(I,J)$ padded with zeros.
- This problem is NP-hard.

# Convex relaxation of sparsity

- A common technique in the literature to promote sparsity is adding an $\ell_1$ penalty term.
- Applying this to the preceding nonconvex problem yields

$$\begin{array}{ll} \min & \|X\|_* + \theta\|X\|_1 \\ \text{s.t.} & A \bullet X \geq 1. \end{array}$$

- Note: $\|X\|_1$ means $\|\text{vec}(X)\|_1$;
- Above problem is convex. (Indeed, it is semidefinite programming.)
- Nuclear-plus-1-norm has appeared in Chandrasekaran et al., Candès et al.

# Some properties of the relaxation

- Norm duality: the function $\|X\|_* + \theta\|X\|_1$ is actually a norm $|||\cdot|||$, and the optimization problem above computes $1/|||A|||^*$.
- Monotonicity: we establish some weak monotonicity properties showing that sparsity increases with $\theta$.
- For sufficiently large $\theta$, the solution $X$ will have one nonzero entry in the position of the largest entry of $A$.

# Nonnegativity

- Suppose $A \geq 0$. Is it true that $X^* \geq 0$?
- An optimal nonnegative solution exists if $\mathrm{rank}(X^*) = 1$.
- $X^* \geq 0$ if $\theta = 0$ (Perron-Frobenius).
- All optimal solutions nonnegative if $\theta > 1$.
- How about in general?

# Recoverability

- Suppose $A \geq 0$ has the form $A = \mathbf{u}\mathbf{v}^T + R$ where $\mathbf{u}$,$\mathbf{v}$ are sparse and $R$ is random noise. Can we recover $(\mathbf{u}, \mathbf{v})$ from $A$?
- No, but maybe we can recover $\mathrm{supp}(\mathbf{u})$ and $\mathrm{supp}(\mathbf{v})$ (positions of nonzero entries).
- Assume that $R$ is i.i.d. random, e.g., Gaussian. Assume $\mathbf{u}, \mathbf{v}$ are deterministic.
- Problem is still unsolvable unless we assume $u(i) \geq \alpha\ \forall i$, $v(j) \geq \beta\ \forall j$ where $\alpha\beta$ bounded below in terms the mean of $R$.

# Recoverability

- Suppose $A \geq 0$ has the form $A = \mathbf{u}\mathbf{v}^T + R$ where **u**,**v** are sparse and $R$ is random noise. Can we recover $(\mathbf{u}, \mathbf{v})$ from $A$?
- No, but maybe we can recover $\mathrm{supp}(\mathbf{u})$ and $\mathrm{supp}(\mathbf{v})$ (positions of nonzero entries).
- Assume that $R$ is i.i.d. random, e.g., Gaussian. Assume $\mathbf{u}, \mathbf{v}$ are deterministic.
- Problem is still unsolvable unless we assume $u(i) \geq \alpha\ \forall i$, $v(j) \geq \beta\ \forall j$ where $\alpha\beta$ bounded below in terms the mean of $R$.

# Recoverability

- Suppose $A \geq 0$ has the form $A = \mathbf{u}\mathbf{v}^T + R$ where $\mathbf{u}$,$\mathbf{v}$ are sparse and $R$ is random noise. Can we recover $(\mathbf{u}, \mathbf{v})$ from $A$?
- No, but maybe we can recover $\mathrm{supp}(\mathbf{u})$ and $\mathrm{supp}(\mathbf{v})$ (positions of nonzero entries).
- Assume that $R$ is i.i.d. random, e.g., Gaussian. Assume $\mathbf{u}, \mathbf{v}$ are deterministic.
- Problem is still unsolvable unless we assume $u(i) \geq \alpha\ \forall i$, $v(j) \geq \beta\ \forall j$ where $\alpha\beta$ bounded below in terms the mean of $R$.

# Main theorem on recoverability

- Say $A \in \mathbf{R}^{M \times N}$; $|\operatorname{supp}(\mathbf{u})| = m$; $|\operatorname{supp}(\mathbf{v})| = n$.
- Assume entries of $R$ are i.i.d. subgaussian about their mean $\mu$.
- Assume $\mathbf{u}, \mathbf{v}$ satisfy above-mentioned condition, and furthermore, $\|\mathbf{u}\| \le O(\sqrt{m}\alpha)$, $\|\mathbf{v}\| \le O(\sqrt{n}\beta)$, $\alpha\beta \ge \Omega(\mu)$.
- Assume $\theta$ chosen in a certain range.
- Then convex relaxation recovers $\operatorname{supp}(\mathbf{u}), \operatorname{supp}(\mathbf{v})$ with prob. exponentially close to 1 provided $m \ge \Omega(\sqrt{M})$ and $n \ge \Omega(\sqrt{N})$.

# Proof steps

- To simplify notation, assume support of $\mathbf{u}, \mathbf{v}$ are their leading indices.
- Hypothesize existence of optimal solution of the form
$$X = \begin{pmatrix} \sigma_1 \bar{\mathbf{u}} \bar{\mathbf{v}}^T & 0 \\ 0 & 0 \end{pmatrix},$$
$\|\bar{\mathbf{u}}\| = \|\bar{\mathbf{v}}\| = 1.$
- KKT condition is $\lambda A = Y + \theta Z$ for some $Y \in \partial \|X\|_*$, $Z \in \partial \|X\|_1$, $\lambda \geq 0$.
- KKT condition sufficient for global optimality in convex optimization.

# Proof steps (cont'd)

- $\lambda A = Y + \theta Z$ for some $Y \in \partial\|X\|_*$, $Z \in \partial\|X\|_1$, $\lambda \geq 0$.
- Specializing to preceding $X$ this means: dominant singular triple of $Y$ is $(1, [\bar{\mathbf{u}}; \mathbf{0}], [\bar{\mathbf{v}}; \mathbf{0}])$; $\|Z\|_\infty = 1$ and $Z_{11} = \mathtt{ones}(\mathtt{m}, \mathtt{n})$.
- Implies that $\lambda$ must be chosen so that $\|\lambda A_{11} - \theta \cdot \mathtt{ones}(\mathtt{m}, \mathtt{n})\| = 1$.
- This is an algebraic equation for $\lambda$; can get good estimates for $\lambda$ because there is a good upper bound known for the norm of a mean-zero random matrix.

# Proof steps (cont'd)

- Once $\lambda$ is known, $\bar{\mathbf{u}}, \bar{\mathbf{v}}$ are dominant singular vectors of $\lambda A_{11} - \theta \cdot \mathtt{ones}(\mathtt{m}, \mathtt{n})$.
- With these choices for $\lambda, \bar{\mathbf{u}}, \bar{\mathbf{v}}$, must next fill in the rest of $Y$ and $Z$ so that $\|Y\| \leq 1$ and $\|Z\|_\infty \leq 1$.
- The requirement $\|Y\| \leq 1$ couples the four blocks together, so replace it with the restriction that $\|Y_{ij}\| \leq 1/2$ for $i, j = 1, 2$.

# Proof steps (cont'd)

- KKT multipliers $Y_{22}$ and $Z_{22}$ constructed by taking the mean of $\lambda A$ into $Z_{22}$ (i.e., make it a multiple of the all-1's matrix) and deviations from average in $Y_{22}$. Uses the fact that $\|R\|$ is (unexpectedly?) small when $R$ is a random mean-0 matrix.
- Construction of KKT multipliers $Y_{12}$, $Z_{12}$ are more complicated because condition on dominant singular triple of $Y$ imposes linear constraint $\mathbf{u}^T Y_{12}=0$.
- Need estimates of $\bar{\mathbf{u}}, \bar{\mathbf{v}}$; use Wedin's sine theorem (SVD perturbation theorem).

# Proof steps (cont'd)

- Helping the analysis: because both terms of the objective are nondifferentiable at the optimizer, the KKT multipliers are not uniquely determined.
- Simple univariate example of this: $\min |x| + |x|$. Can take any subdifferential in $[-1, 1]$ for first term of KKT condition; take the opposite for the second term.

# Recovery of $\mathrm{supp}(\mathbf{u}), \mathrm{supp}(\mathbf{v})$

- The proof of the theorem shows that, under the assumptions and with high probability, $\mathrm{rank}(X) = 1$, i.e., $X = \hat{\mathbf{u}}\hat{\mathbf{v}}^T$ where $\hat{\mathbf{u}}$ is the extension of $\bar{\mathbf{u}}$ with zeros and similarly for $\hat{\mathbf{v}}$.
- Furthermore, $\mathrm{supp}(\mathbf{u}) = \mathrm{supp}(\hat{\mathbf{u}})$ and $\mathrm{supp}(\mathbf{v}) = \mathrm{supp}(\hat{\mathbf{v}})$.

# Convex relaxation for NP-hard problems

Recent literature has produced a number of examples of useful NP-hard problems that can be solved in polynomial time using convex relaxation, assuming the problem instance is formed in a particular way.

- Compressive sensing (Donoho; Candès, Romberg & Tao)
- Planted clique & biclique (Feige & Krauthgamer; Ames & V.)
- Rank minimization over an affine space (Recht, Fazel & Parrilo)
- Matrix completion problem (Candès & Recht)

# Max biclique problem

- Given a bipartite graph $G = (U, V, E)$, a *biclique* is given by $U^* \subset U$, $V^* \subset V$ such that all of $U^* \times V^*$ lies in $E$.
- Max-edge biclique problem asks for biclique with max number of edges.
- Problem is NP-hard.

# Planted biclique problem

- Our relaxation can solve this problem in the case that $|U^*| \geq \Omega(|U|^{1/2})$, $|V^*| \geq \Omega(|V|^{1/2})$, and the non-clique edges are inserted at random.
- Same bound achieved earlier by Ames & V.
- Unlike Ames & V., our relaxation needs prior knowledge of the biclique size to correctly pick $\theta$. But our relaxation solves a more general problem.

# Convex solver

- Recall our relaxation

$$\begin{array}{ll} \min & \|X\|_* + \theta\|X\|_1 \\ \text{s.t.} & A \bullet X \geq 1. \end{array}$$

  is convex and indeed SDP-expressible.
- Interior point SDP solvers (Sedumi, SDPT3) require $O(p^3)$ flops per iteration, where $p = MN$ (number of unknowns).
- Too inefficient for large problems.

# Subgradient descent

- We use a subgradient descent method.
- On each step, approximately minimize *proximal point mapping*. Proximal-point mapping for convex $\phi(\mathbf{x})$ defined to be solution to $\min_{\mathbf{x}} \phi(\mathbf{x}) + \lambda\|\mathbf{x} - \mathbf{c}\|$ (2-norm for vectors, F-norm for matrices).

# Proximal point mapping for objective

- We do not know how to efficiently minimize the proximal-point mapping for our objective function $\phi(X) = \|X\|_* + \theta\|X\|_1 + \lambda\|X - C\|_F$.
- Therefore, rewrite relaxation as

$$\begin{aligned} \min \quad & \|X_1\|_* + \theta\|X_2\|_1 \\ \text{s.t.} \quad & A \bullet X_1 \geq 1, \\ & X_1 = X_2 \end{aligned}$$

- This allows us to compute the proximal point mapping separately for $\|\cdot\|_*$ and $\|\cdot\|_1$.

# Proximal point mapping for nuclear norm

- Proximal-point mapping for nuclear norm: given $C$, minimizer of $\|X\|_* + \lambda\|X - C\|_F$ is

$$U \begin{pmatrix} (\sigma_1 - 1/\lambda)^+ & & \\ & \ddots & \\ & & (\sigma_n - 1/\lambda)^+ \end{pmatrix} V^T,$$

where $C = U\Sigma V^T$.
- PROPACK (Fortran routines using Lanczos) used for this step.

# Termination test

- Since only nonzero pattern of optimal $X^*$ is useful, would like to terminate as soon as nonzero pattern is determined.
- Assume that optimal $\mathrm{rank}(X^*) = 1$. (Test won't be satisfied if this assumption fails.)
- Would like a test that, when satisfied, guarantees correct answer has been found.

# Termination test (cont'd)

- New termination test: given approximate solution $\tilde{X}$, take (approximation to) dominant singular triple $(\tilde{\sigma}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ and approximate Lagrange multiplier $\tilde{\lambda}$.
- Consider system of equations:

$$\begin{aligned} (\lambda A_{11} - \theta Z_{11})\mathbf{v}_1 &= \mathbf{u}_1, \\ (\lambda A_{11} - \theta Z_{11})^T \mathbf{u}_1 &= \mathbf{v}_1, \\ \mathbf{u}_1^T \mathbf{u}_1 &= 1. \end{aligned}$$

where $Z_{11}$ is all 1's.

# Termination test (cont'd)

- Can apply Kantorovich theorem to determine that the system has an exact solution distance $\epsilon$ from $(\tilde{\lambda}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}})$.
- KKT conditions for a rank-one sparse solution include above equations and also inequalities.
- Use simple least squares to guess remaining multipliers.
- Check whether the inequalities hold for all points within a ball of radius $\epsilon$ around $(\tilde{\lambda}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}})$.
- If so, a rank-one solution with correct sparsity pattern has been found.

# Termination test (cont'd)

- Complexity of this technique unknown.
- In practice, technique can sometimes cut number of iterations almost in half but is computationally expensive, so is not applied on every iteration.

# Termination test for matrix completion problem

- Problem is: given a matrix $M$ with many missing entries, fill in the missing entries to minimize the rank. NP-hard.
- Candès & Recht; Candès and Tao: relaxation $\min \|X\|_*$ s.t. $X_{ij} = M_{ij}$, $(i,j) \in \Omega$ can efficiently solve MCP for instances constructed a certain way.
- Our technique from an approximate solution can yield a computational proof that $\mathrm{rank}(X) \leq k$ for some $k$.

# Computational experiments

- Two black/white image datasets used in experiments.
- In both cases, LAROS run repeatedly in order to extract several features (find approximate NMF).
- Termination test: either as on previous transparency, or achievable accuracy achieved.
- Choice of $\theta$: heuristic used.

# Frey face data

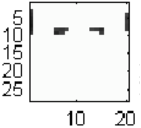
Frey face dataset consists of 1965 grayscale mugshots of a person's face in different poses.

# Applying the method to Frey dataset

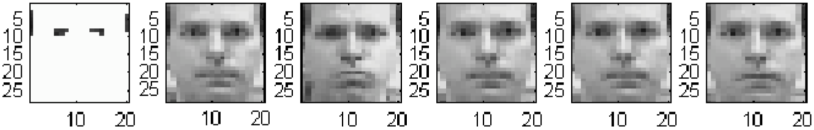
- Can form a $560 \times 1965$ matrix, one mugshot per column and look for a large rank-one submatrix.
- Feature corresponds to subset of images in database with common visual feature in the same groups of pixels.
- Can find multiple features by iteratively solving LAROS and subtracting off previous features.

# Results

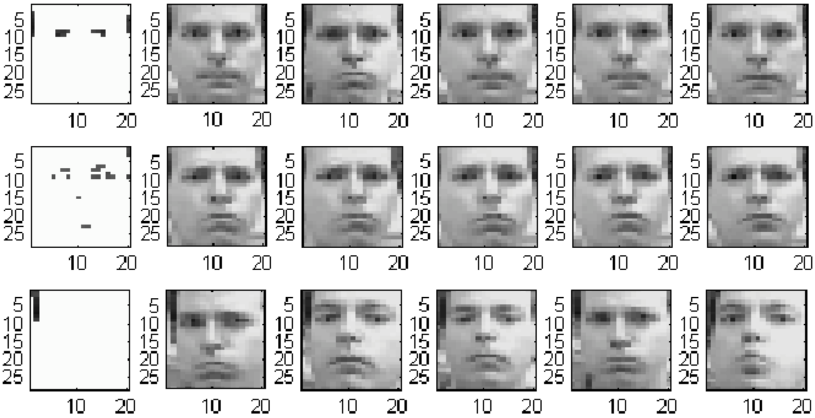
# Results



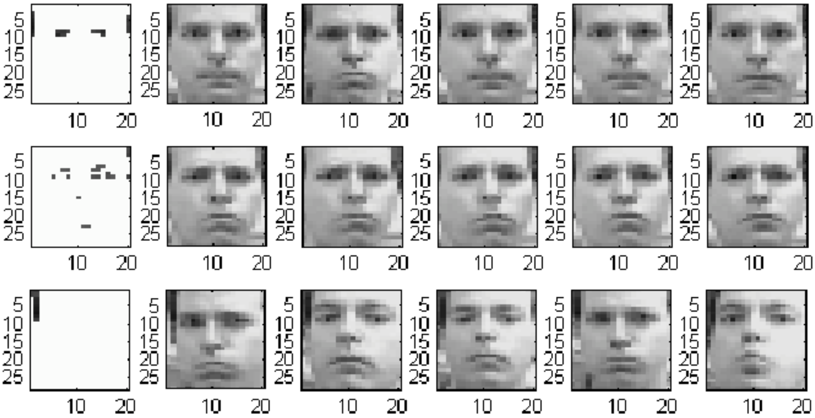
# Results



# Results



# Results




This SDP has $> 10^6$ variables.

# Open questions previously mentioned

- Can we show that $X \geq 0$ when $A \geq 0$?
- Does the new termination test admit a complexity-based analysis?

# Other open questions

- Can we recover multiple features at once?
- How to choose $\theta$ more rigorously?
- Faster algorithms?
- Characterize extreme points of $\{X : \|X\|_* + \theta\|X\|_1 \leq 1\}$.
- More generally: designing new matrix norms.