Subset Selection

Ilse Ipsen

North Carolina State University, USA

Subset Selection

Given: real or complex matrix A
integer k

Determine permutation matrix P so that

$$AP = (\underbrace{A_1}_{k} \quad A_2)$$

- **Important columns A_1**
  Columns of A_1 are 'very' linearly independent
  "Wannabe basis vectors"

- **Redundant columns A_2**
  Columns of A_2 are 'well' represented by A_1

Outline

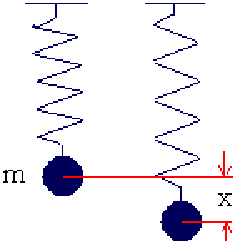
- **Two applications**
- **Mathematical formulation**
- **Bounds**
- **Algorithms (two norm)**
- **Redundant columns (Frobenius norm)**
- **Randomized algorithms**

First application

Joint work with Tim Kelley

- Solution of nonlinear least squares problems
- Levenberg-Marquardt trust-region algorithm
- Difficulties:
  illconditioned or rank deficient Jacobians
  errors in residuals and Jacobians
- Improve conditioning with subset selection
- Damped driven harmonic oscillators
  allow us to construct different scenarios

Harmonic Oscillators



$$m\,x'' + c\,x' + k\,x = 0$$

displacement x(t)
mass m
damping constant c
spring stiffness k

Given: displacement measurements x_j at time t_j
Want: parameters $p = (m, c, k)$

Nonlinear least squares problem

$$\min_p \sum_j |x(t_j, p) - x_j|^2$$

http://www.relisoft.com/Science/Physics/images/oscillator.gif

Nonlinear Least Squares Problem

Residual $R(p) = \big(x(t_1, p) - x_1 \quad x(t_2, p) - x_2 \quad \dots\big)^T$

- Nonlinear Least Squares Problem

$$\min_p f(p) \qquad \text{where} \quad f(p) = \frac{1}{2} R(p)^T R(p)$$

- At a minimizer p^*: $\nabla f(p^*) = 0$

Gradient: $\nabla f(p) = R'(p)^T R(p)$
Jacobian: $R'(p)$

- Solve $\nabla f(p) = 0$ by Levenberg-Marquardt algorithm

$$p_{new} = p - \Big(\nu I + R'(p)^T R'(p)\Big)^{-1} R'(p)^T R(p)$$

Jacobian

- Levenberg-Marquardt algorithm

$$p_{new} = p - \left(\nu I + R'(p)^T R'(p)\right)^{-1} R'(p)^T \, R(p)$$

But: Jacobian $R'(p)$ does not have full column rank

- $m\,x'' + c\,x' + k\,x = 0$ for infinitely many $p = (m, c, k)$

$$x'' + \frac{c}{m}\,x' + \frac{k}{m}\,x = 0$$
$$\frac{m}{c}\,x'' + x' + \frac{k}{c}\,x = 0$$
$$\frac{m}{k}\,x'' + \frac{c}{k}\,x' + x = 0$$

Which parameter to keep fixed?

- Which column in the Jacobian is redundant?

Subset Selection!

Second Application

Scott Pope's Ph.D. thesis (NCState, 2009)

- **Cardiovascular and respiratory modeling**
- **Identify parameters that are important for predicting blood flow and pressure**
- **Solve nonlinear least squares problem combined with subset selection**
- **Parameters identified as important:**

  total systemic resistance
  cerebrovascular resistance
  arterial compliance
  time of peak systolic ventricular pressure

Mathematical Formulation of Subset Selection

Given: real or complex matrix A with n columns
integer k

Determine permutation matrix P so that

$$AP = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k})$$

- **Important columns A_1**

  Columns of A_1 are 'very' linearly independent

  Smallest singular value of A_1 is 'large'

- **Redundant columns A_2**

  Columns of A_2 are 'well' represented by A_1

  $\min_Z \|A_1 Z - A_2\|$ is 'small' (two norm)

Singular Value Decomposition (SVD)

$m \times n$ matrix A, $m \geq n$

$$AP = U \Sigma V$$

- **Singular vectors:**

$$U^T U = I_m \qquad V^T V = V V^T = I_n$$

- **Singular values:**

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix} \qquad \sigma_1 \geq \ldots \geq \sigma_n \geq 0$$

Ideal Matrices for Subset Selection

Singular Value Decomposition

$$AP = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k}) = U \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} V$$

If exists permutation P so that $V = I$ then

- Important columns $\longrightarrow$ large singular values of A

$$A_1 = U \begin{pmatrix} \Sigma_1 \\ 0 \end{pmatrix} \quad \text{and} \quad \sigma_i(A_1) = \sigma_i(A) \quad 1 \leq i \leq k$$

- Redundant columns $\longrightarrow$ small singular values of A

$$A_2 = U \begin{pmatrix} 0 \\ \Sigma_2 \end{pmatrix} \quad \text{and} \quad \min_Z \|A_1 Z - A_2\| = \|A_2\| = \sigma_{k+1}(A)$$

Subset Selection Requirements

- **Important columns A_1**
  k columns of A_1 should be 'very' linearly independent

  Smallest singular value $\sigma_k(A_1)$ should be 'large'

  $$\sigma_k(A)/\gamma \leq \sigma_k(A_1) \leq \sigma_k(A)$$

  for some γ

- **Redundant columns A_2**
  Columns of A_2 should be 'well' represented by A_1

  $\min_Z \|A_1 Z - A_2\|$ should be 'small' (two norm)

  $$\sigma_{k+1}(A) \leq \min_Z \|A_1 Z - A_2\| \leq \gamma\, \sigma_{k+1}(A)$$

  for some γ

Bounds for Subset Selection

Singular value decomposition

$$\mathbf{AP} = (\underbrace{\mathbf{A_1}}_{k} \quad \underbrace{\mathbf{A_2}}_{n-k}) = \mathbf{U} \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} \mathbf{V_{11}} & \mathbf{V_{12}} \\ \mathbf{V_{21}} & \mathbf{V_{22}} \end{pmatrix}$$

- **Important columns $\mathbf{A_1}$**

$$\frac{\sigma_i(\mathbf{A})}{\|\mathbf{V_{11}^{-1}}\|} \leq \sigma_i(\mathbf{A_1}) \leq \sigma_i(\mathbf{A}) \qquad \text{for all} \quad 1 \leq i \leq k$$

- **Redundant columns $\mathbf{A_2}$**

$$\sigma_{k+1}(\mathbf{A}) \leq \min_{\mathbf{Z}} \|\mathbf{A_1 Z} - \mathbf{A_2}\| \leq \|\mathbf{V_{11}^{-1}}\| \, \sigma_{k+1}(\mathbf{A})$$

How Small Can $\|V_{11}^{-1}\|$ Be?

Matrix $\begin{pmatrix} V_{11} & V_{12} \end{pmatrix}$ is $k \times n$ with orthonormal rows

- If V_{11} is nonsingular then

$$\|V_{11}^{-1}\| \leq \sqrt{1 + \|V_{11}^{-1} V_{12}\|^2}$$

Follows from $I = V_{11}V_{11}^T + V_{12}V_{12}^T$

- If $|\det(V_{11})|$ is maximal then

$$\|V_{11}^{-1} V_{12}\| \leq \sqrt{k(n-k)}$$

Follows from Cramer's rule: $\left| (V_{11}^{-1} V_{12})_{ij} \right| \leq 1$

- There exists a permutation such that

$$\|V_{11}^{-1}\| \leq \sqrt{1 + k(n-k)}$$

Bounds for Subset Selection

$$AP = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k}) = U \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

Permute right singular vectors so that $|\det(V_{11})|$ maximal

- Important columns A_1

$$\frac{\sigma_i(A)}{\sqrt{1+k(n-k)}} \leq \sigma_i(A_1) \leq \sigma_i(A) \qquad \text{for all} \quad 1 \leq i \leq k$$

- Redundant columns A_2

$$\sigma_{k+1}(A) \leq \min_Z \|A_1 Z - A_2\| \leq \sqrt{1+k(n-k)}\, \sigma_{k+1}(A)$$

Algorithms for Subset Selection

QR decomposition with column pivoting

$$AP = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix} \qquad \text{where} \quad Q^T Q = I$$

- **Important columns**

$$Q \begin{pmatrix} R_{11} \\ 0 \end{pmatrix} = A_1 \quad \text{and} \quad \sigma_i(A_1) = \sigma_i(R_{11}) \quad 1 \le i \le k$$

- **Redundant columns**

$$Q \begin{pmatrix} R_{12} \\ R_{22} \end{pmatrix} = A_2 \qquad \min_Z \|A_1 Z - A_2\| = \|R_{22}\|$$

QR Decomposition with Column Pivoting

- Determines permutation matrix P so that

$$AP = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$$

  where

  R_{11} well-conditioned
  $\|R_{22}\|$ small

- Numerical rank of A is k
- QR decomposition reveals rank

Rank Revealing QR (RRQR) Decompositions

Businger & Golub (1965) QR with column pivoting
Faddev, Kublanovskaya & Faddeeva (1968)
Golub, Klema & Stewart (1976)
Gragg & Stewart (1976)
Stewart (1984)
Foster (1986)
T. Chan (1987)

Hong & Pan (1992)
Chandrasekaran & Ipsen (1994)
Gu & Eisenstat (1996) Strong RRQR

Strong RRQR (Gu & Eisenstat 1996)

Input: $m \times n$ matrix A, $m \geq n$, integer k

Output: $AP = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$ $\quad R_{11}$ is $k \times k$

- R_{11} is well conditioned

$$\frac{\sigma_i(A)}{\sqrt{1 + k(n-k)}} \leq \sigma_i(R_{11}) \leq \sigma_i(A) \quad 1 \leq i \leq k$$

- R_{22} is small

$$\sigma_{k+j}(A) \leq \sigma_j(R_{22}) \leq \sqrt{1 + k(n-k)}\, \sigma_{k+j}(A)$$

- Offdiagonal block not too large

$$\left| \left(R_{11}^{-1} R_{12} \right)_{ij} \right| \leq 1$$

Strong RRQR Algorithm

1. Compute some QR decomposition with column pivoting

$$AP_{initial} = Q \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$$

2. Repeat

   Exchange a column of $\begin{pmatrix} R_{11} \\ 0 \end{pmatrix}$ with a column of $\begin{pmatrix} R_{12} \\ R_{22} \end{pmatrix}$

   Update permutations P, retriangularize

   until $|\det(R_{11})|$ stops increasing

3. Output: $AP_{final} = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k})$

Operation count: $\mathcal{O}\left(mn^2\right)$ until $|\det(R_{11})|$ stops increasing by $\sqrt{n}$

Another Strong RRQR Algorithm

In the spirit of Golub, Klema and Stewart (1976)

1. **Compute SVD**

$$\mathbf{A} = \mathbf{U} \begin{pmatrix} \boldsymbol{\Sigma}_1 & \\ & \boldsymbol{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}$$

2. **Apply strong RRQR to $\mathbf{V}_1$:** $\quad \mathbf{V}_1\mathbf{P} = (\underbrace{\mathbf{V}_{11}}_{k} \quad \underbrace{\mathbf{V}_{12}}_{n-k})$

$$\frac{1}{\sqrt{1+k(n-k)}} \leq \sigma_i(\mathbf{V}_{11}) \leq 1 \qquad 1 \leq i \leq k$$

3. **Output:** $\quad \mathbf{AP} = (\underbrace{\mathbf{A}_1}_{k} \quad \underbrace{\mathbf{A}_2}_{n-k})$

Summary: Deterministic Subset Selection

$m \times n$ matrix A, $m \geq n$, $\quad AP = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k})$

- Important columns A_1

$$\sigma_k(A)/p(k,n) \leq \sigma_k(A_1) \leq \sigma_k(A)$$

- Redundant columns A_2

$$\sigma_{k+1}(A) \leq \min_Z \|A_1 Z - A_2\| \leq p(k,n)\, \sigma_{k+1}(A)$$

- $p(k,n)$

  Depends on leading k right singular vectors
  Best known value: $p(k,n) = \sqrt{1 + k(n-k)}$

- Algorithms: Rank revealing QR decompositions, SVD
- Operation count: $\mathcal{O}\left(mn^2\right) \quad$ for $p(k,n) = \sqrt{1 + nk(n-k)}$

Redundant Columns

$$AP = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k})$$

RRQR: $\min_Z \|A_1 Z - A_2\| \leq p(k,n)\, \sigma_{k+1}(A)$

- Orthogonal projector onto $\mathrm{range}(A_1)$: $A_1 A_1^\dagger$

$$\min_Z \|A_1 Z - A_2\| = \|(I - A_1 A_1^\dagger)\, A_2\|$$

- Largest among all small singular values

$$\sigma_{k+1}(A) = \|\Sigma_2\|$$

RRQR: $\|(I - A_1 A_1^\dagger)\, A_2\| \leq p(k,n)\, \|\Sigma_2\|$

Subset Selection for Redundant Columns

$$AP = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k})$$

Among all $\binom{n}{k}$ choices find A_1 that minimizes

$$\|(I - A_1 A_1^\dagger)\, A_2\|_\xi$$

Best bounds:

- **2 norm**

$$\|(I - A_1 A_1^\dagger)\, A_2\|_2 \leq \sqrt{1 + k(n-k)}\; \|\Sigma_2\|_2$$

- **Frobenius norm**

$$\|(I - A_1 A_1^\dagger)\, A_2\|_F \leq \sqrt{k+1}\; \|\Sigma_2\|_F$$

Frobenius Norm

There exist k columns A_1 so that

$$\|(I - A_1 A_1^\dagger)\, A_2\|_F^2 \le (k+1) \sum_{j \ge k+1} \sigma_j(A)^2$$

Idea: Volume sampling

Deshpande, Rademacher, Vempala & Wang 2006

- $\|(I - A_1 A_1^\dagger)\, A_2\|_F^2 = \sum_{j\ge k+1} \|(I - A_1 A_1^\dagger)\, a_j\|_2^2$
- Volume

$$\mathrm{Vol}(a_j) = \|a_j\|_2 \qquad \mathrm{Vol}\begin{pmatrix} A_1 & a_j \end{pmatrix} = \mathrm{Vol}(A_1)\,\|(I - A_1 A_1^\dagger) a_j\|_2$$

- Volume and singular values

$$\sum_{i_1 < \ldots < i_k} \mathrm{Vol}\left(A_{i_1 \ldots i_k}\right)^2 = \sum_{i_1 < \ldots < i_k} \sigma_{i_1}(A)^2 \ldots \sigma_{i_k}(A)^2$$

Maximizing Volumes Is Really Hard

Given: matrix A with n columns of unit norm
integer k
real number $\nu \in [0, 1]$

- Finding k columns A_1 of A such that

$$\mathrm{Vol}(A_1) \geq \nu$$

  is NP-hard
- There is no polynomial time approximation scheme

[Civril & Magdon-Ismail, 2007]

Randomized Subset Selection

Frieze, Kannan & Vempala 2004
Deshpande, Rademacher, Vempala & Wang 2006
Liberty, Woolfe, Martinsson, Rokhlin & Tygert 2007
Drineas, Mahoney & Muthukrishnan 2006, 2008
Boutsidis, Mahoney & Drineas 2009
Civril & Magdon-Ismail 2009

Applications

- Statistical data analysis:
  - feature selection
  - principal component analysis
- Pass efficient algorithms for large data sets

2-Phase Randomized Algorithm

Boutsidis, Mahoney & Drineas 2009

1. **Randomized Phase:**
   Sample small number ($\approx k \log k$) of columns
2. **Deterministic Phase:**
   Apply rank revealing QR to sampled columns

With 70% probability:

- Two norm

$$\min_{Z} \|A_1 Z - A_2\|_2 \leq \mathcal{O}\left(k^{3/4} \log^{1/2} k \, (n-k)^{1/4}\right) \|\Sigma_2\|_2$$

- Frobenius norm

$$\min_{Z} \|A_1 Z - A_2\|_F \leq \mathcal{O}\left(k \log^{1/2} k\right) \|\Sigma_2\|_F$$

2-Phase Randomized Algorithm

1. Compute SVD

$$A = U \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$

2. Randomized phase:
   Scale: $V_1 \rightarrow V_1 D$
   Sample c columns: $(V_1 D)\, P_s = (\underbrace{V_s}_{\hat{c}} \quad *)$

3. Deterministic phase:
   Apply RRQR to V_s: $V_s\, P_d = (\underbrace{V_d}_{k} \quad *)$

4. Output: $A P_s P_d = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k})$

Randomized Phase (Frobenius Norm)

Sampling

- Column i of V_1 sampled with probability p_i
- "Probabilities"

$$p_i = c \left(\frac{\|(V_1)_i\|_2}{\|V_1\|_F} \right)^2 \qquad 1 \leq i \leq n$$

Scaling

- Scaling matrix $D = \mathrm{diag}\left(1/\sqrt{p_1} \quad \ldots \quad 1/\sqrt{p_n}\right)$
- Scaled matrix $V_1 D$

  All columns have the same norm
  Columns sampled with probability $1/n$

- Purpose of scaling

  makes sampling "uniform"
  makes expected values easier to compute

Analysis of 2-Phase Algorithm

1. Sample c columns $\quad (VD)\, P_s = \begin{pmatrix} V_{1s} D_s & * \\ V_{2s} D_s & * \end{pmatrix}$
2. RRQR selects k columns $\quad (V_{1s} D_s)\, P_d = \begin{pmatrix} V_d & * \end{pmatrix}$

- Perturbation theory

$$\min_Z \|A_1 Z - A_2\|_F \leq \|\Sigma_2\|_F + \frac{\|\Sigma_2\, V_{2s} D_s\|_F}{\sigma_k(V_d)}$$

- RRQR

$$\sigma_k(V_d) \geq \frac{\sigma_k(V_{1s} D_s)}{\sqrt{1 + k(\hat{c} - k)}}$$

- With "high" probability

$$\sigma_k(V_{1s} D_s) \geq 1/2 \qquad \|\Sigma_2\, V_{2s} D_s\|_F \leq 4 \|\Sigma_2\|_F$$

- $c \approx k \log k$

$$\min_Z \|A_1 Z - A_2\|_F \leq \mathcal{O}\left(k \log^{1/2} k\right) \|\Sigma_2\|_F$$

Expected Values of Frobenius Norms

If $D_{ii} = 1/\sqrt{p_i}$ then

$$E\left(\|X\,D\|_F^2\right) = \|X\|_F^2$$

- **Frobenius norm**

$$\|X\,D\|_F^2 = \mathrm{trace}(X\,D^2\,X^T)$$

- **Linearity**

$$E\left[\|X\,D\|_F^2\right] = \mathrm{trace}(X\;\underbrace{E\left[D^2\right]}_{I}\;X^T)$$

- **Scaling**

$$E\left[D_{ii}^2\right] = p_i * \frac{1}{p_i} + (1 - p_i) * 0 = 1$$

From Expected Values to Probability

$$\mathbf{E}\left(\|\mathbf{X}\,\mathbf{D}\|_F^2\right) = \|\mathbf{X}\|_F^2$$

- **Markov's inequality**

$$\mathrm{Prob}\,(\mathbf{x} \geq \mathbf{a}) \leq \mathbf{E}(\mathbf{x})/\mathbf{a}$$

- $\mathbf{x} = \|\mathbf{X}\,\mathbf{D}\|_F^2, \quad \mathbf{a} = 10\,\mathbf{E}(\mathbf{x})$
- **With probability at most** $1/10$

$$\|\mathbf{X}\,\mathbf{D}\|_F^2 \geq 10\,\|\mathbf{X}\|_F^2$$

- **With probability at least** $9/10$

$$\|\mathbf{X}\,\mathbf{D}\|_F^2 \leq 10\,\|\mathbf{X}\|_F^2$$

Issues with Randomized Algorithms

- How to choose c: $10^{-3}\, k \log k,\ k \log k,\ 17\, k \log k, \ldots$?
- We don't know the number of sampled columns $\hat{c}$
- Number of sampled columns can be too small: $\hat{c} < k$
- No information about singular values of important columns
- How often does one have to run the algorithm to get a good result?
- How accurately do the singular vectors and singular values have to be computed?
- How sensitive is the algorithm to the choice of probabilities?
- How does the randomized algorithm compare to the deterministic algorithms: accuracy, run time?

Summary

Given: real or complex matrix A, integer k

Want: $AP = (\underbrace{A_1}_{k} \quad \underbrace{A_2}_{n-k})$

- Important columns A_1
  Singular values close to k largest singular values of A
- Redundant columns A_2
  $\|\text{Proj. of } A_2 \text{ on } \mathrm{range}(A_1)^\perp\|_{2,F} \approx$ smallest singular values of A
- Bounds depend on dominant k right singular vectors
- Deterministic algorithms: RRQR, SVD
- Randomized algorithm:
  2 phases: 1. randomized sampling, 2. RRQR on samples
- Exact subset selection is hard