Rýchly Algoritmus na Výpočet Riedkych Hlavných Komponentov

Peter Richtárik

Centrum pre Operačný Výskum a Ekonometriu (CORE) a

Fakulta Matematického Inžinierstva (INMA)

Katolícka univerzita v Louvain – Belgicko

(týž)Deň Absolventov Matfyzu, 19.12.2008, FMFI UK, Bratislava

CORE Discussion Paper #2008/70

spoluautori: Michel Journée, Yurii Nesterov a Rodolphe Sepulchre

1. Obsah prednášky

- Analýza riedkych hlavných komponentov (ARHK)
- Reformulácia problému
- Algoritmus a analýza zložitosti
- Numerické experimenty

2. Analýza riedkych hlavných komponentov

- **Vstup:** Matica $A = [a_1, \ldots, a_n] \in \mathbf{R}^{p \times n}, \quad p \leq n$
- **Cieľ:** Nájdi jednotkový vektor $z^* \in \mathbf{R}^n$ ktorý simultánne
  1. **maximalizuje hodnotu funkcie** $z^T A^T A z$
  2. je **riedky**

Ak riedkosť **nie je** požadovaná, z^* je **dominantným pravým singulárnym vektorom** matice A:

$$\max_{z^T z \leq 1} z^T A^T A z = \lambda_{\mathsf{max}}(A^T A) = (\sigma_{\mathsf{max}}(A))^2.$$

Extrahovanie viacerých komponentov: Práve sme definovali problém pre prípad získania **jedného komponentu ($m = 1$).** Častokrát v praxi je potrebné vypočítať **viac komponentov ($m > 1$).**

Aplikácie: genetická analýza, financie, vizualizácia dát, spracovanie signálu, počítačové videnie, ...

3. Náš prístup k ARHK

1. Formuluj ARHK ako optimalizačný problém s **penalizáciou riedkosti** (ℓ_1 or ℓ_0) kontrolovanou jedným parametrom
2. **Reformuluj** na problém **vhodnej štuktúry:**
   - vhodnej na analýzu
   - vhodnej na výpočet
3. "Vyrieš"reformulovaný problém pomocou jendoduchej **gradientovej metódy**
4. Spätne vypočítaj riešenie pôvodného problému

Kroky 1), 2) a 4) ilustrujeme na prípade jedného komponentu a ℓ_1 penalizácie. Potom sa pustíme do všeobecnej analýzy kroku 3).

4. Tri postrehy o ℓ_1 penalizácii

Symbolika: $\|z\|_1 = \sum_i |z_i|$.

Penalizačná formulácia ARHK pre prípad jedného komponentu:

$$\phi_{\ell_1}(\gamma) \stackrel{\text{def}}{=} \max_{z^T z \le 1} \sqrt{z^T A^T A z} - \gamma \|z\|_1. \qquad (1)$$

Postrehy:

1. $\gamma = 0 \Rightarrow$ **žiaden dôvod na to aby súradnice vektora z^* boli nulové**
2. **Ak $\gamma \ge \|a_{i^*}\|_2 \stackrel{\text{def}}{=} \max_i \|a_i\|$, tak $z^* = 0$.** Dôvod:

$$\begin{aligned} \max_{z \ne 0} \frac{\|Az\|_2}{\|z\|_1} &= \max_{z \ne 0} \frac{\|\sum_i z_i a_i\|_2}{\|z\|_1} \\ &\le \max_{z \ne 0} \frac{\sum_i |z_i| \|a_i\|_2}{\sum_i |z_i|} = \max_i \|a_i\|_2. \end{aligned}$$

3. V skutočnosti: $\gamma \ge \|a_i\|_2 \;\Rightarrow\; z_i^*(\gamma) = 0$ **pre všetky** i

5. Reformulácia

Všimnime si, že:

$$\phi_{\ell_1}(\gamma) = \max_{z\in\mathcal{B}^n} \|Az\|_2 - \gamma\|z\|_1 = \max_{z\in\mathcal{B}^n} \max_{x\in\mathcal{B}^p} x^T Az - \gamma\|z\|_1$$
$$= \max_{x\in\mathcal{B}^p} \max_{z\in\mathcal{B}^n} \sum_{i=1}^{n} z_i(a_i^T x) - \gamma|z_i|.$$

Pre dané x vnútorný max-problém má riešenie v uzatvorenej forme:

$$z_i = \text{sign}(a_i^T x)[|a_i^T x| - \gamma]_+, \qquad z^* = z/\|z\|_2.$$

Takže na vyriešenie problému (1) stačí **vyriešiť nasledovnú reformuláciu:**

$$\boxed{\phi_{\ell_1}^2(\gamma) = \max_{\substack{x\in\mathbf{R}^p \\ x^T x=1}} \sum_{i=1}^{n} [|a_i^T x| - \gamma]_+^2,} \qquad (2)$$

Poznámka: Cieľová funkcia problému (2) je **konvexná** a **hladká** a **vektor x má rozmer p a nie n ($p \ll n$)**.

6. ARHK pomocou ℓ_0 penalizácie

Podobný postup ako v prípade ℓ_1, takže iba stručne:

Symbolika: $\|z\|_0 = \mathrm{Card}\{i \ : \ z_i \neq 0\}$.

Formulácia problému pomocou penalizácie:

$$\phi_{\ell_0}(\gamma) \stackrel{\text{def}}{=} \max_{z^T z \leq 1} z^T A^T A z - \gamma \, \|z\|_0, \tag{3}$$

Na vyriešenie (3) stačí najprv **vyriešiť túto úlohu**

$$\boxed{\phi_{\ell_0}(\gamma) = \max_{\substack{x \in \mathbf{R}^p \\ x^T x = 1}} \sum_{i=1}^{n} [(a_i^T x)^2 - \gamma]_+,} \tag{4}$$

a potom položiť

$$z_i = [\mathrm{sign}((a_i^T x)^2 - \gamma)]_+ a_i^T x, \qquad z^* = z/\|z\|_2.$$

7. Maximalizácia konvexnej funkcie

Problémy (2) a (4) (a ich zovšeobecnenia pre výpočet viacerých komponentov) sú nasledovnej štruktúry:

$$\boxed{f^* = \max_{x \in \mathcal{Q}} f(x),} \qquad \text{kde} \qquad \text{(P)}$$

- $\mathbf{E}$ je konečnorozmerný vektorový priestor,
- $f : \mathbf{E} \to \mathbf{R}$ je **konvexná funkcia,**
- $\mathcal{Q} \subset \mathbf{E}$ je **kompaktná množina.**

V prípade ARHK máme

- $\mathcal{Q} =$ jedotková **euklidovská sféra** v $\mathbf{R}^p$ / jeden komponent $(m = 1)$
- $\mathcal{Q} =$ **Stiefelova množina** v $\mathbf{R}^{p \times m}$, i.e. množina $p \times m$ matíc s ortonormálnymi stĺpcami / viac komponentov $(m > 1)$

Ako vyriešiť problém (P)?

8. Gradientový algoritmus

Problém (P) navrhujeme riešiť nasledovnou metódou:

1. **Vstup:** Bod $x_0 \in \mathcal{Q}$
2. **Pre** $k \geq 0$ **opakuj**
   - $x_{k+1} \in \operatorname{Arg\,max}\{f(x_k) + \langle f'(x_k), y - x_k \rangle \mid y \in \mathcal{Q}\}$
   - $k \leftarrow k + 1$

Tento algoritmus je zovšeobecnením tzv. **"power method"** na výpočet najväčšej vlastnej hodnoty symetrickej kladne semidefinitnej matice C:

$$f(x) = \tfrac{1}{2} x^T C x \quad \rightarrow \quad x_{k+1} = \frac{C x_k}{\|C x_k\|_2}.$$

Preto sme zvolili názov **"Generalized Power Method"** (GPower).

9. Iteračná zložitosť

V každom bode $x \in \mathcal{Q}$ definujeme nasledovnú **mieru splnenia optimalizačných podmienok prvého rádu:**

$$\Delta(x) \stackrel{\text{def}}{=} \max_{y \in \mathcal{Q}} \langle f'(x), y - x \rangle.$$

Je zrejmé že $\Delta(x) \geq 0$, pričom $\Delta(x) = 0$ práve vtedy keď gradient $f'(x)$ leží v kuželi normálnom k množine $\mathrm{Conv}(\mathcal{Q})$ v bode x.

Nech $\Delta_k \stackrel{\text{def}}{=} \min_{0 \leq i \leq k} \Delta(x_i)$.

Veta: Ak postupnosť bodov $\{x_k\}_{k=0}^{\infty}$ je generovaná algoritmom GPower aplikovanom na konvexnú funkciu f, potom postupnosť $\{f(x_k)\}_{k=0}^{\infty}$ je **rastúca** a $\lim_{k \to \infty} \Delta(x_k) = 0$. Naviac,

$$\Delta_k \leq \frac{f^* - f(x_0)}{k+1}. \tag{5}$$

10. Silná konvexnosť funkcií a množín

Funkcia f je silne konvexná ak existuje konštanta $\sigma_f > 0$ taká, že pre každé $x, y \in \mathbf{E}$

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{\sigma_f}{2}\|y - x\|^2.$$

Množina $\mathrm{Conv}(\mathcal{Q})$ je silne konvexná ak existuje konštanta $\sigma_{\mathcal{Q}} > 0$ taká, že pre každé $x, y \in \mathrm{Conv}(\mathcal{Q})$ a $\alpha \in [0, 1]$ platí nasledovná inklúzia:

$$\alpha x + (1 - \alpha)y + \frac{\sigma_{\mathcal{Q}}}{2}\alpha(1 - \alpha)\|x - y\|^2 \cdot \mathcal{S} \subset \mathrm{Conv}(\mathcal{Q}).$$

Veta: Ak $f : \mathbf{E} \to \mathbf{R}$ je nezáporná funkcia, $\sigma_f > 0$ a ak $f' : \mathbf{E} \to \mathbf{E}^*$ je L_f-Lipschitzovská, tak pre každé $\omega > 0$ je množina

$$\mathcal{Q}_\omega \stackrel{\text{def}}{=} \{x \mid f(x) \leq \omega\}$$

silne konvexná s parametrom $\sigma_{\mathcal{Q}_\omega} = \sigma_f / \sqrt{2\omega L_f}$.

11. Zlepšená analýza v prípade silnej kovexnosti

Veta:

Nech

- f je konvexná s parametrom silnej konvexnosti $\sigma_f \geq 0$ a
- $\mathrm{Conv}(\mathcal{Q})$ konexná s parametrom silnej konvexnosti $\sigma_{\mathcal{Q}} \geq 0$.

Ak $0 < \delta_f = \inf_{x \in \mathcal{Q}} \|f'(x)\|_*$ a ak $\sigma_f > 0$ alebo $\sigma_{\mathcal{Q}} > 0$, potom

$$\sum_{k=0}^{N} \|x_{k+1} - x_k\|^2 \leq \frac{2(f^* - f(x_0))}{\sigma_{\mathcal{Q}} \delta_f + \sigma_f}.$$

Poznámka: Ak minimum funkcie f *neleží* v množine $\mathcal{Q}$, potom $\delta_f > 0$.

12. Numerické experimenty

Porovnáme nasledovné algoritmy na riešenie **ARHK:**

GPower$_{\ell_1}$	Jeden komponent, ℓ_1-penalizácia [1]
GPower$_{\ell_0}$	Jeden komponent, ℓ_0-penalizácia [1]
GPower$_{\ell_1,m}$	m komponentov, ℓ_1-penalizácia [1]
GPower$_{\ell_0,m}$	m komponentov, ℓ_0-penalizácia [1]
SPCA	Algoritmus SPCA [2]
Greedy*	"Greedy"metóda [3]
rSVD$_{\ell_1}$	Algoritmus [4] s ℓ_1-penalizáciou ("soft thresholding")
rSVD$_{\ell_0}$	Algoritmus [4] s ℓ_0-penalizáciou ("hard thresholding")

*Greedy je dramaticky pomalší algoritmus v porovnaní s ostatnými horeuvedenými metódami, najmä v prípade riešenia s veľkou kardinalitou.

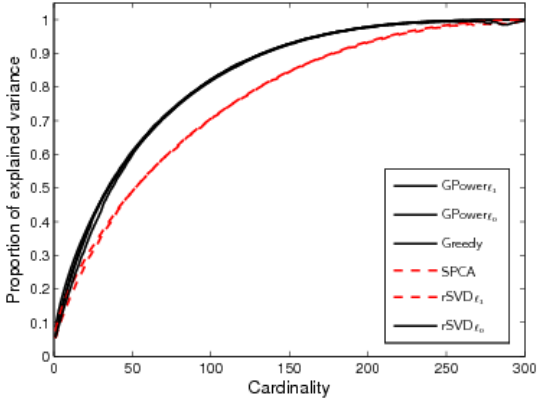
Testovacie problémy:

- Náhodne generované

  $A =$ Štandardné normálne rozdelenie

- Reálne dáta z genetickej analýzy

13. Trade-off krivky



Trade-off medzi **vysvetleným rozptylom** a **kardinalitou**. Metódy GPower$_{\ell_1}$, GPower$_{\ell_0}$, Greedy a rSVD$_{\ell_0}$ vykazujú lepšie výsledky **(čierna plná čiara)**, pričom SPCA a rSVD$_{\ell_1}$ horšie výsledky **(červená prerušovaná čiara)**.

Graf je založený na 100 náhodne generovaných testovacích problémoch veľkosti $p = 100, n = 300$.

14. Náhodné dáta: rýchlosť

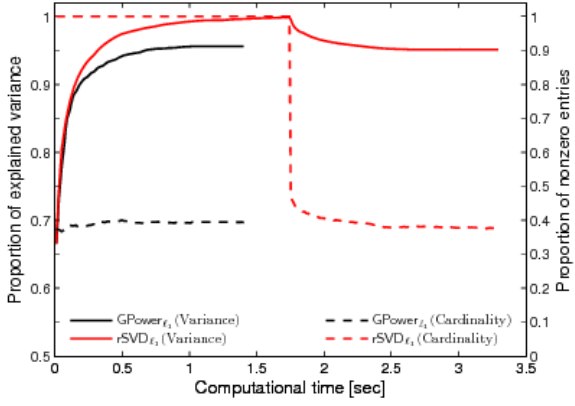
Fixný pomer n/p:

$p \times n$	250×2500	500×5000	750×7500	1000×10000
GPower$_{\ell_1}$	0.85	2.61	3.89	5.32
GPower$_{\ell_0}$	0.46	1.21	2.41	2.93
SPCA	2.77	14.0	41.0	81.6
rSVD$_{\ell_1}$	1.40	6.80	17.8	41.2
rSVD$_{\ell_0}$	1.33	6.20	15.4	36.3

Fixované p, rastúce n:

$p \times n$	500×2000	500×4000	500×8000	500×16000
GPower$_{\ell_1}$	0.97	1.96	4.30	8.43
GPower$_{\ell_0}$	0.39	0.97	2.01	4.63
SPCA	7.37	11.4	22.4	44.6
rSVD$_{\ell_1}$	2.56	5.27	11.3	26.8
rSVD$_{\ell_0}$	2.30	4.70	10.3	23.8

15. Ako sa trade-off vyvíja v čase?



Vývoj **vysvetleného rozptylu** (plné čiary – os vľavo) a **kardinality** (prerušované čiary – os vpravo) v čase v prípade algoritmov GPower$_{\ell_1}$ and rSVD$_{\ell_1}$.

Graf je založený na 100 náhodne generovaných testovacích problémoch veľkosti $p = 250$ and $n = 2500$.

16. Dáta z genetickej analýzy: rýchlosť

Dáta ("breast cancer cohorts"):

Štúdia	Vzorky (p)	Gény (n)	Článok
Vijver	295	13319	van de Vijver et al. [2002]
Wang	285	14913	Wang et al. [2005]
Naderi	135	8278	Naderi et al. [2006]
JRH-2	101	14223	Sotiriou et al. [2006]

Rýchlosť (v sekundách):

	Vijver	Wang	Naderi	JRH-2
GPower$_{\ell_1}$	7.72	6.96	2.15	2.69
GPower$_{\ell_0}$	3.80	4.07	1.33	1.73
GPower$_{\ell_1,m}$	5.40	4.37	1.77	1.14
GPower$_{\ell_0,m}$	5.61	7.21	2.25	1.47
SPCA	77.7	82.1	26.7	11.2
rSVD$_{\ell_1}$	46.4	49.3	13.8	15.7
rSVD$_{\ell_0}$	46.8	48.4	13.7	16.5

17. Dáta z genetickej analýzy: kvalita výsledku

PEI: 536 “pathways"súvisiacich s rakovinou:

	Vijver	Wang	Naderi	JRH-2
PCA	0.0728	0.0466	0.0149	0.0690
GPower$_{\ell_1}$	**0.1493**	0.1026	0.0728	0.1250
GPower$_{\ell_1}$	0.1250	0.1250	0.0672	0.1026
GPower$_{\ell_1,m}$	0.1418	0.1250	**0.1026**	**0.1381**
GPower$_{\ell_0,m}$	0.1362	**0.1287**	0.1007	0.1250
SPCA	0.1362	0.1007	0.0840	0.1007
rSVD$_{\ell_1}$	0.1213	0.1175	0.0914	0.0914
rSVD$_{\ell_0}$	0.1175	0.0970	0.0634	0.1063

Pathway Enrichment Index (PEI) meria štatistickú významnosť prieniku medzi dvoma súbormi génov.

18. Rekapitulácia

- Navrhli sme **4 reformulácie** (jeden komponent/viacero komponentov $\times$ ℓ_1/ℓ_0) problému **Analýzy Riedkych Hlavných Komponentov (ARHK)**, ktoré nám umožnili
  - použiť veľmi **rýchly algoritmus** (pracujeme v $\mathbf{R}^p$ a nie v $\mathbf{R}^n$ pričom $p \ll n$ a používame iba prvé derivácie)
  - analyzovať **iteračnú zložitosť** tohto algoritmu.
- Analyzovali sme jednoduchý gradientový algoritmus **(Generalized Power Method)** na maximalizáciu konvexných funkcií na kompaktných množinách;
- Aplikovali sme GPower na tieto 4 reformulácie, čím sme dostali 4 algoritmy na riešenie problému ARHK;
- testovali sme naše algortimy na náhodne generovaných a na reálnych dátach:
  - sú **podstatne rýchlejšie**,
  - v prípade biologických dát produkujú riešenia s **mierne vyššou** hodnotou indexu PEI.

19. Referencie

[1] M. Journée, Yu. Nesterov, P. Richtárik, R. Sepulchre. **Generalized Power Method for Sparse Principal Component Analysis (táto prednáška).** *submitted to Journal of Machine Learning Research*, November 2008.

[2] H. Zou, T. Hastie, R. Tibshirani. **Sparse Principal Component Analysis.** *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

[3] A. d'Aspremont, F. R. Bach, L. El Ghaoui. **Optimal Solutions for Sparse Principal Component Analysis.** *Journal of Machine Learning Research*, 9:1269–1294, 2008.

[4] H. Shen, J. Z. Huang. **Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation.** *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.