Finding Sparse Approximations to Extreme Eigenvectors: Generalized Power Method for Sparse PCA and Extensions

Peter Richtárik
School of Mathematics, The University of Edinburgh
Email: Peter.Richtarik@ed.ac.uk

Abstract—**In the first part of this work, based on [2], we develop a new approach to sparse principal component analysis (sparse PCA). We propose four optimization formulations of the problem, aimed at extracting one or several sparse dominant components. While the initial formulations involve nonconvex functions, we rewrite them into the form of an optimization program involving maximization of a convex function on a compact set and propose and analyze a simple gradient method for solving it (generalized power method). We demonstrate numerically on a set of random and gene expression test problems that our approach outperforms existing algorithms both in quality of the obtained solution and in speed.**

A natural extension of the ideas above allows us to construct a method for finding, simultaneously, *jointly sparse approximations* to the eigenvectors associated with the largest *and smallest* eigenvalues of a symmetric psd matrix. This problem is equivalent to the Compressed Sensing problem of finding bounds on the asymmetric Restricted Isometry constants with the additional new requirement for the respective sparse eigenvectors to be supported on the same set. We prove a result on the emergence of joint sparsity in the iterates of the method and show that in the non-penalized case, the iterates are identical to the normalized gradients of the iterates of the Cauchy steepest descent method applied to minimizing a convex quadratic function [1].

I. Preliminaries

Let $A = [a_1, \dots, a_n] \in \mathbb{R}^{p \times n}$, with $p \ll n$. Let $\bar{\lambda}$ (resp. $\underline{\lambda}$) be the largest (resp. smallest) eigenvalue of $S = A^T A$. Fix $\gamma > 0$.

II. Generalized Power Method for Sparse PCA

For simplicity, we focus here on the problem of finding a sparse approximation z_* to the eigenvector of S "corresponding" to $\bar{\lambda}$. That is, we seek a sparse unit-norm vector $z_* \in \mathbb{R}^n$ such that $\|Az_*\|_2$ is large. Consider the following optimization problem:

$$\max\{\|Az\|_2 - \gamma\|z\|_1 \;:\; \|z\|_2 \leq 1\}. \tag{1}$$

It turns out that the optimal solution z_* of (1) is given by

$$z_* = z/\|z\|_2, \quad z^{(i)} = \operatorname{sign}(a_i^T x)[|a_i^T x| - \gamma]_+, \quad i = 1, \dots, n,$$

where x is solves the smooth convex maximization problem

$$\max_{\|x\|_2 \leq 1} \sum_{i=1}^{n} [|a_i^T x| - \gamma]_+^2. \tag{2}$$

Note that since $p \ll n$, the dimension of the search space is decreased enormously. It is easy to show that $\gamma \geq \|a_i\|_2 \Rightarrow z_*^{(i)} = 0$, and hence γ controls sparsity of the solution.

For problems of type (2), i.e., for maximization of a convex function f over a compact set Q, we propose the following simple gradient method: Choose $x_0 \in Q$ and for $k \geq 0$ iterate:

$$x_{k+1} \in \arg\max\{f(x_k) + \langle f'(x_k), y - x_k\rangle \;:\; y \in Q\} \tag{GPM}$$

This is our main convergence result:

Theorem 1 ([2]). *Let f be convex, Q compact and $\{x_i\}$ be the iterates produced by GPM. Then*

$$\min_{0 \leq i \leq k} \max_{y \in Q} \langle f'(x_i), y - x_i\rangle \leq \frac{\max f^* - f(x_0)}{k+1}.$$

If, in addition, f is strongly convex with parameter $\sigma_f > 0$, the convex hull of Q is strongly convex with parameter σ_Q, and we define $\delta_f = \min\{\|s\|^* \;:\; s \in \partial f(x),\; x \in Q\}$, then

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 \leq \frac{2(\max f - f(x_0))}{\sigma_Q \delta_f + \sigma_f}.$$

III. Jointly Sparse Min and Max Eigenvectors

Consider the following optimization problem:

$$\max\{x^T S y - \gamma\|(x, y)\|_1 \;:\; \|x\|_2 = \|y\|_2 = 1,\; x^T y = 0\}. \tag{3}$$

If $\gamma = 0$, the optimal value of (3) is $\frac{1}{2}(\bar{\lambda} - \underline{\lambda})$, and if x^*, y^* are the optimal solutions, then $p = (x^* + y^*)/\sqrt{2}$ and $q = (x^* - y^*)/\sqrt{2}$ are the maximal and minimal eigenvectors of S, respectively. Below we give a method for (approximately) solving (3) for $\gamma > 0$ and show that γ induces *joint sparsity* in x and y. Hence, the method is able to identify a small principal submatrix of S whose extreme eigenvalues are a good approximation to $\bar{\lambda}$ and $\underline{\lambda}$.

Let $y_\gamma(x)$ (resp. $x_\gamma(y)$) be the optimal solution of (3) for fixed x (resp. y). Fix unit-norm x_0 and consider the following method:

$$y_k = y_\gamma(x_k), \qquad x_{k+1} = x_\gamma(y_k). \tag{ADM}$$

Theorem 2. *Let $w \in \mathbb{R}^n$ with $\|w\|_2 = 1$, $u = Sw$, $L = \{tw \;:\; t \in \mathbb{R}\}$, $B = \{s \;:\; \|s + u\|_\infty \leq \gamma\}$ and*

$$Opt \overset{def}{=} \max\{u^T z - \gamma\|z\|_1 \;:\; \|z\|_2 = 1,\; w^T z = 0\}. \tag{4}$$

If L does not pass through the interior of B, then the solution of (4) *is given by $z = d/\|d\|_2$, $Opt = \sqrt{\omega(t^*)} = \|d\|_2$, where*

$$t^* \in \arg\min_t [\omega(t) \overset{def}{=} \sum_{i=1}^{n} ([|u^{(i)} + tw^{(i)}| - \gamma]_+)^2],$$

$$d^{(i)} = \operatorname{sign}(u^{(i)} + t^* w^{(i)})[|u^{(i)} + t^* w^{(i)}| - \gamma]_+, \; i = 1, \dots, n.$$

This result gives conditions under which the operations in (ADM) can be performed efficiently (in a closed form).

Let $u_k = Sx_k$. We further show that

1) *validity result:* if $\gamma \leq \sqrt{\|u_0\|_2^2 - (u^T x_0)^2}/(\|x_0\|_1 + \sqrt{n})$, then the condition of Theorem 2 will hold for all ADM iterates,
2) *joint sparsity result:* any of the conditions (i) $\|A^T a_i\|_2 \leq \gamma$, $x_k^{(i)} = 0$, (ii) $|x_k^{(i)}| \leq (\gamma - |u_k^{(i)}|)/\sqrt{\gamma^2(n-4) + 2\gamma\|u_k\|_1 + \|u_k\|_2^2}$, implies $y_k^{(i)} = 0$.

References

[1] H. Akaike. On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Annals of the Institute of Statistical Mathematics*, 11:1–16, 1959.

[2] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553, 2010.