Approximation Bounds for Sparse PCA

Alexandre d’Aspremont, CNRS & Ecole Polytechnique

with Francis Bach, INRIA-ENS and Laurent El Ghaoui, U.C. Berkeley
PCA on high-dimensional data

PCA. Summarize the data in a few dimensions, given by the leading eigenvectors of the covariance matrix.

High dimensional data sets. $n$ sample points in dimension $p$, with

$$p = \gamma n, \quad p \to \infty.$$ 

for some fixed $\gamma > 0$.

- Common in e.g. biology (many genes, few samples), or finance (data not stationary, many assets).
- Many recent results on PCA in this setting. Very precise knowledge of asymptotic distributions of extremal eigenvalues.
PCA on high-dimensional data

PCA on Gaussian noise in a high dimensional setting.

- If the entries of $X \in \mathbb{R}^{n \times p}$ are standard i.i.d. and have a fourth moment, then

$$
\lambda_{\text{max}} \left( \frac{X^T X}{n-1} \right) \rightarrow (1 + \sqrt{\gamma})^2 \quad \text{a.s.}
$$

if $p = \gamma n$, $p \rightarrow \infty$. [Geman, 1980, Yin et al., 1988]

- When $\gamma \in (0, 1]$, the spectral measure converges to the following density

$$
f_{\gamma}(x) = \frac{\sqrt{(x - a)(b - x)}}{2\pi \gamma x}
$$

where $a = (1 - \sqrt{\gamma})^2$ and $b = (1 + \sqrt{\gamma})^2$. [Marčenko and Pastur, 1967]

- The distribution of $\lambda_{\text{max}} \left( \frac{X^T X}{n-1} \right)$, properly normalized, converges to the Tracy-Widom distribution [Johnstone, 2001, Karoui, 2003]. This works well even for small values of $n, p$. 
Spectrum of Wishart matrix with $p = 500$ and $n = 1500$. 
PCA on high-dimensional data

We focus on the following hypothesis testing problem

\[
\begin{aligned}
\mathcal{H}_0 : & \quad x \sim \mathcal{N}(0, \mathbf{I}_p) \\
\mathcal{H}_1 : & \quad x \sim \mathcal{N}(0, \mathbf{I}_p + \theta \mathbf{v}\mathbf{v}^T)
\end{aligned}
\]

where \(\theta > 0\) and \(\|\mathbf{v}\|_2 = 1\).

- Of course \(\lambda_{\text{max}}(\mathbf{I}_p) = 1\) and \(\lambda_{\text{max}}(\mathbf{I}_p + \theta \mathbf{v}\mathbf{v}^T) = 1 + \theta\), so we can use \(\lambda_{\text{max}}(\cdot)\) as our test statistic.

- However, [Baik et al., 2005, Tao, 2011, Benaych-Georges et al., 2011] show that

\[
\lambda_{\text{max}} \left( \frac{X^TX}{n-1} \right) \to (1 + \sqrt{\gamma})^2
\]

under both \(\mathcal{H}_0\) and \(\mathcal{H}_1\) when \(\theta\) is small, i.e.

\[
\theta \leq \gamma + \sqrt{\gamma}
\]

in the high dimensional regime \(p = \gamma n\), with \(\gamma \in (0, 1)\), \(p \to \infty\).
PCA on high-dimensional data

Gene expression data in [Alon et al., 1999].

**Left:** Spectrum of gene expression sample covariance, and Wishart matrix with equal total variance.

**Right:** Magnitude of coefficients in leading eigenvector, in decreasing order.
Sparse PCA

Here, we assume the leading principal component is sparse. We will use sparse eigenvalues as a test statistic

\[
\lambda_{\text{max}}^k(\Sigma) \triangleq \max. \quad x^T \Sigma x \\
\text{s.t.} \quad \text{Card}(x) \leq k, \quad \|x\|_2 = 1,
\]

- We focus on the sparse eigenvector detection problem

\[
\begin{align*}
\mathcal{H}_0 : & \quad x \sim \mathcal{N}(0, I_p) \\
\mathcal{H}_1 : & \quad x \sim \mathcal{N}(0, I_p + \theta vv^T)
\end{align*}
\]

where \( \theta > 0 \) and \( \|v\|_2 = 1 \) with \( \text{Card}(v) = k \).

- We naturally have

\[
\lambda_{\text{max}}^k(I_p) = 1 \quad \text{and} \quad \lambda_{\text{max}}^k(I_p + \theta vv^T) = 1 + \theta
\]
Sparse PCA

Berthet and Rigollet [2012]: Optimal detection threshold using $\lambda_{\text{max}}^k(\cdot)$ is

$$\theta = 4\sqrt{\frac{k \log(9ep/k) + \log(1/\delta)}{n}} + \ldots$$

- **Good news**: $\lambda_{\text{max}}^k(\cdot)$ is a **minimax optimal statistic** for detecting sparse principal components. The dimension $p$ only appears as a log term and this threshold is much better than $\theta = \sqrt{p/n}$ in the dense PCA case.

- **Bad news**: Computing the statistic $\lambda_{\text{max}}^k(\hat{\Sigma})$ is NP-Hard.

[Berthet and Rigollet, 2012] produce **tractable** statistics achieving the threshold

$$\theta = 2\sqrt{k}\sqrt{\frac{k \log(4p^2/\delta)}{n}} + \ldots$$

which means $\theta \to \infty$ when $k, n, p \to \infty$ proportionally. However $p$ large, $k$ fixed is OK, empirical performance much better than this bound would predict.
Clustering of the gene expression data in the PCA versus sparse PCA basis with 500 genes. The factors $f$ on the left are dense and each use all 500 genes while the sparse factors $g_1$, $g_2$ and $g_3$ on the right involve 6, 4 and 4 genes respectively. (Data: Iconix Pharmaceuticals)
Outline

- PCA on high-dimensional data
- Approximation bounds for sparse eigenvalues
Approximation bounds for sparse eigenvalues

Penalized eigenvalue problem.

\[ \text{SPCA}(\rho) \triangleq \max_{\|x\|_2 = 1} x^T \Sigma x - \rho \text{Card}(x) \]

where \( \rho > 0 \) controls the sparsity. We can show

\[ \text{SPCA}(\rho) = \max_{\|x\|_2 = 1} \sum_{i=1}^{p} \left( (a_i^T x)^2 - \rho \right)_+ \]

We form a **convex relaxation** of this last problem

\[ \text{SDP}(\rho) \triangleq \max \sum_{i=1}^{p} \text{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ \]
\[ \text{s.t.} \quad \text{Tr}(X) = 1, \; X \succeq 0, \]

which is equivalent to a semidefinite program.
Approximation bounds for sparse eigenvalues

Proposition 1 [d’Aspremont, Bach, and El Ghaoui, 2008]

Semidefinite relaxation $\text{SDP}(\rho)$. Write $\Sigma = A^T A$ and $a_1, \ldots, a_p \in \mathbb{R}^p$ the columns of $A$, then

$$\text{SPCA}(\rho) \leq \text{SDP}(\rho).$$

where

$$\text{SDP}(\rho) = \max \sum_{i=1}^p \text{Tr}(X^{1/2}a_i a_i^T X^{1/2} - \rho X)_+$$

$$\text{s.t.} \quad \text{Tr}(X) = 1, \ X \succeq 0.$$
Proof sketch. Change variables, set $X = xx^T$, so $\|x\|_2 = 1$ means $\text{Tr}(X) = 1$ and $(a_i^T x)^2 = a_i^T X a_i$.

Also, $X^{1/2} = X = xx^T$, and we write everything else in terms of $X$

$$(a_i^T X a_i - \rho)_+ = \text{Tr}((a_i^T xx^T a_i - \rho)xx^T)_+$$

$$= \text{Tr}(x(x^T a_i a_i^T x - \rho)x^T)_+ \quad (\text{Tr}(\cdot)_+ = \lambda_{\text{max}}(\cdot) \text{ here})$$

$$= \text{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ = \text{Tr}(X^{1/2} (a_i a_i^T - \rho I) X^{1/2})_+.$$

The function $X \mapsto \text{Tr}(X^{1/2} B X^{1/2})_+$ is concave because we can write it as

$$\text{Tr}(X^{1/2} B X^{1/2})_+ = \max_{\{0 \preceq P \preceq X\}} \text{Tr}(P B) = \min_{\{Y \succeq B, Y \succeq 0\}} \text{Tr}(Y X),$$

concave in $X$ as a pointwise minimum of affine functions.

$$\text{SPCA}(\rho) = \max. \sum_{i=1}^n \text{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+$$

s.t. $\text{Tr}(X) = 1, \text{ Rank}(X) = 1, X \succeq 0,$

We relax the original problem into a semidefinite program by simply dropping the rank constraint.
Proposition 2 [d’Aspremont, Bach, and El Ghaoui, 2012]

Approximation ratio on $\text{SDP}(\rho)$. Write $\Sigma = A^T A$ and $a_1, \ldots, a_p \in \mathbb{R}^p$ the columns of $A$. Let us call $X$ the optimal solution to

$$\text{SDP}(\rho) = \max \quad \sum_{i=1}^{p} \text{Tr}(X^{1/2}a_ia_i^TX^{1/2} - \rho X)_+$$

s.t. $\text{Tr}(X) = 1, \ X \succeq 0$,

and let $r = \text{Rank}(X)$, we have

$$p\rho \vartheta_r \left( \frac{\text{SDP}(\rho)}{p\rho} \right) \leq \text{SPCA}(\rho) \leq \text{SDP}(\rho),$$

where

$$\vartheta_r(x) \triangleq \mathbb{E} \left[ \left( x\xi_1^2 - \frac{1}{r-1} \sum_{j=2}^{r} \xi_j^2 \right)_+ \right]$$

controls the approximation ratio.
Proof sketch. W.l.o.g. $\rho < \min_{i \in [1,n]} \sum_{ii}$, so $B_i(X) = X^{1/2}(a_i a_i^T - \rho I)X^{1/2}$ has exactly one positive eigenvalue $\alpha_i = \text{Tr} B_i(X)_+ + r$ negative eigenvalues $-\beta_j^i$. 

$\xi \in \mathbb{R}^n$ i.i.d. standard normal, $z = X^{1/2}\xi$ satisfies $\mathbb{E}[zz^T] = X$ and rotational invariance yields

$$
\mathbb{E} \left[ \left( (a_i^T z)^2 - \rho \|z\|_2^2 \right)_+ \right] = \mathbb{E} \left[ (\xi^T B_i(X)\xi)_+ \right]
$$

$$
= \mathbb{E} \left[ \left( \alpha_i \xi_1^2 - \sum_{j=2}^r \beta_j^i \xi_j^2 \right)_+ \right]
$$

Then $\sum_{j=2}^r \beta_j^i = \text{Tr}(B(X))_+ - \text{Tr}(B(X)) = \alpha_i - (a_i^T X a_i - \rho) \leq \rho$ because $\lambda_{\text{max}}(B_i(X)) \leq a_i^T X a_i$, hence

$$
\mathbb{E} \left[ (\xi^T B_i(X)\xi)_+ \right] \geq \min_{\beta} \left\{ \mathbb{E} \left[ \left( \alpha_i \xi_1^2 - \sum_{j=2}^r \beta_j^i \xi_j^2 \right)_+ \right] : \sum_{j=2}^r \beta_j^i \leq \rho, \ \beta_j^i \geq 0 \right\}
$$

$$
= \mathbb{E} \left[ \left( \alpha_i \xi_1^2 - \frac{\rho}{r-1} \sum_{j=2}^r \xi_j^2 \right)_+ \right],
$$

by convexity and symmetry.
By homogeneity and convexity, with $\text{SDP}(\rho) = \sum_{i=1}^n \alpha_i$, we then get

$$
\mathbb{E} \left[ \sum_{i=1}^n (\xi^T B_i(X) \xi)_+ \right] \geq \sum_{i=1}^n \mathbb{E} \left[ \left( \alpha_i \xi_1^2 - \frac{\rho}{r-1} \sum_{j=2}^r \xi_j^2 \right)_+ \right] \\
\geq \mathbb{E} \left[ \left( \text{SDP}(\rho) \xi_1^2 - \frac{n \rho}{r-1} \sum_{j=2}^r \xi_j^2 \right)_+ \right],
$$

and we define $\vartheta_r(x)$ as above. We have shown

$$
\mathbb{E} \left[ \sum_{i=1}^n (\xi^T B_i(X) \xi)_+ \right] \geq n \rho \vartheta_r \left( \frac{\text{SDP}(\rho)}{n \rho} \right),
$$

and this bound implies that there exists a nonzero $z = \frac{X^{1/2} \xi}{\|X^{1/2} \xi\|_2}$ such that

$$
n \rho \vartheta_r \left( \frac{\text{SDP}(\rho)}{n \rho} \right) \leq \sum_{i=1}^n ((a_i^T z)^2 - \rho)_+ \leq \text{SPCA}(\rho).
$$

because $\text{SPCA}(\rho) = \max_{\|z\|_2=1} \sum_{i=1}^n ((a_i^T z)^2 - \rho)_+$  ■
Approximation bounds for sparse eigenvalues

- By convexity, we also have \( \vartheta_r(x) \geq \vartheta(x) \), where

\[
\vartheta(x) = \mathbb{E} \left[ (x \xi^2 - 1)^+ \right] = \frac{2e^{-1/2x}}{\sqrt{2\pi x}} + 2(x - 1)N\left(-x^{-\frac{1}{2}}\right)
\]

- Overall, we have the following approximation bounds

\[
\frac{\vartheta(c)}{c} \text{SDP}(\rho) \leq \text{SPCA}(\rho) \leq \text{SDP}(\rho), \quad \text{when } c \leq \frac{\text{SDP}(\rho)}{p\rho}.
\]
No uniform approximation à la MAXCUT... But improved results for specific instances, as in [Zwick, 1999] for MAXCUT on “heavy” cuts.

Here, approximation quality is controlled by the ratio

\[
\frac{\text{SDP}(\rho)}{\rho}.
\]

For the detection problem, when \(\gamma\) is small enough the approximation ratio is of order one.
References


