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 os 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

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

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

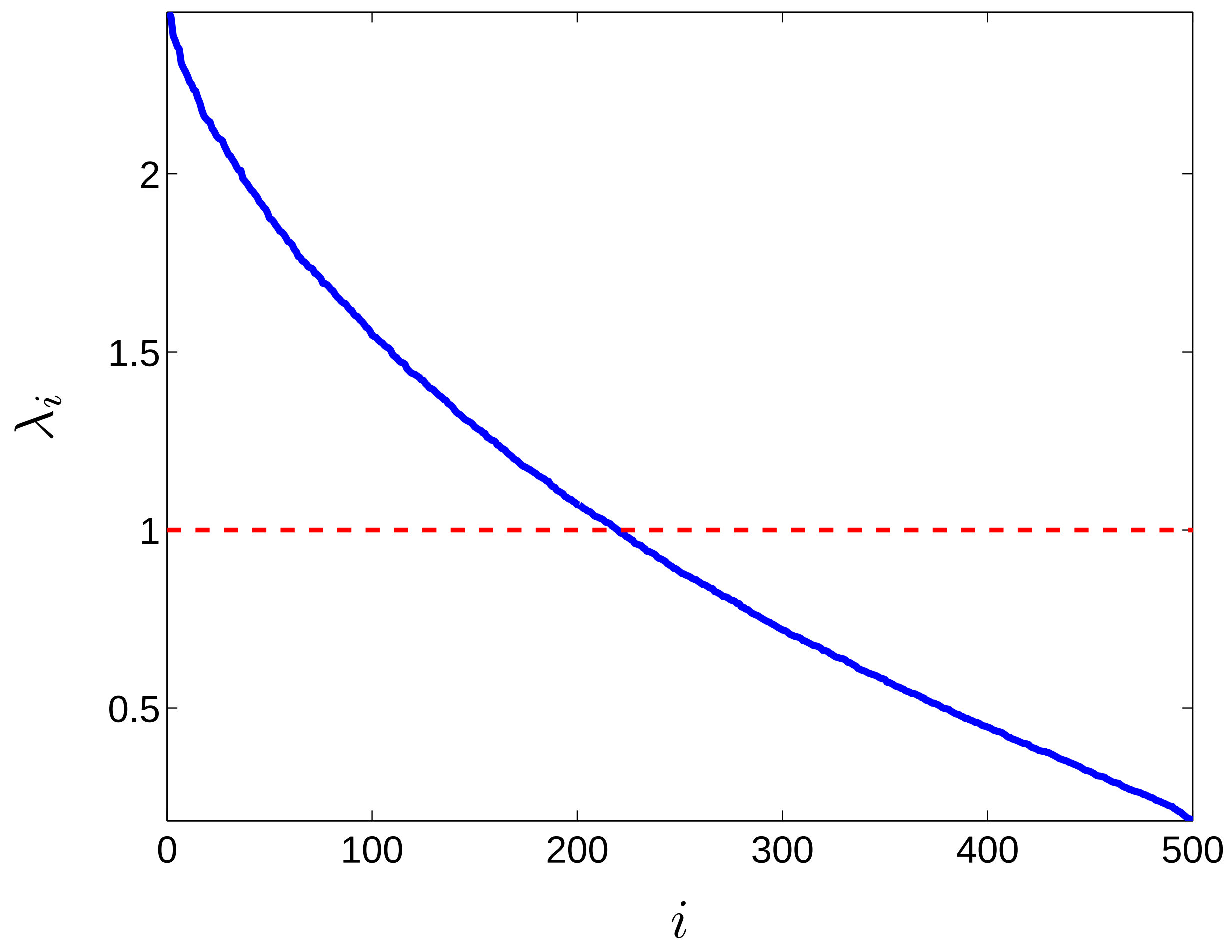
- When $\gamma \in (0,1]$, the spectral measure converges to the following density

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

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

- The distribution of $\lambda_{\max}\left(\frac{X^T X}{n-1}\right)$, properly normalized, converges to the Tracy-Widom distribution [Johnstone, 2001, Karoui, 2003]. This works well even for small values of n, p.

PCA on high-dimensional data



Spectrum of Wishart matrix with $p = 500$ and $n = 1500$.

PCA on high-dimensional data

We focus on the following hypothesis testing problem

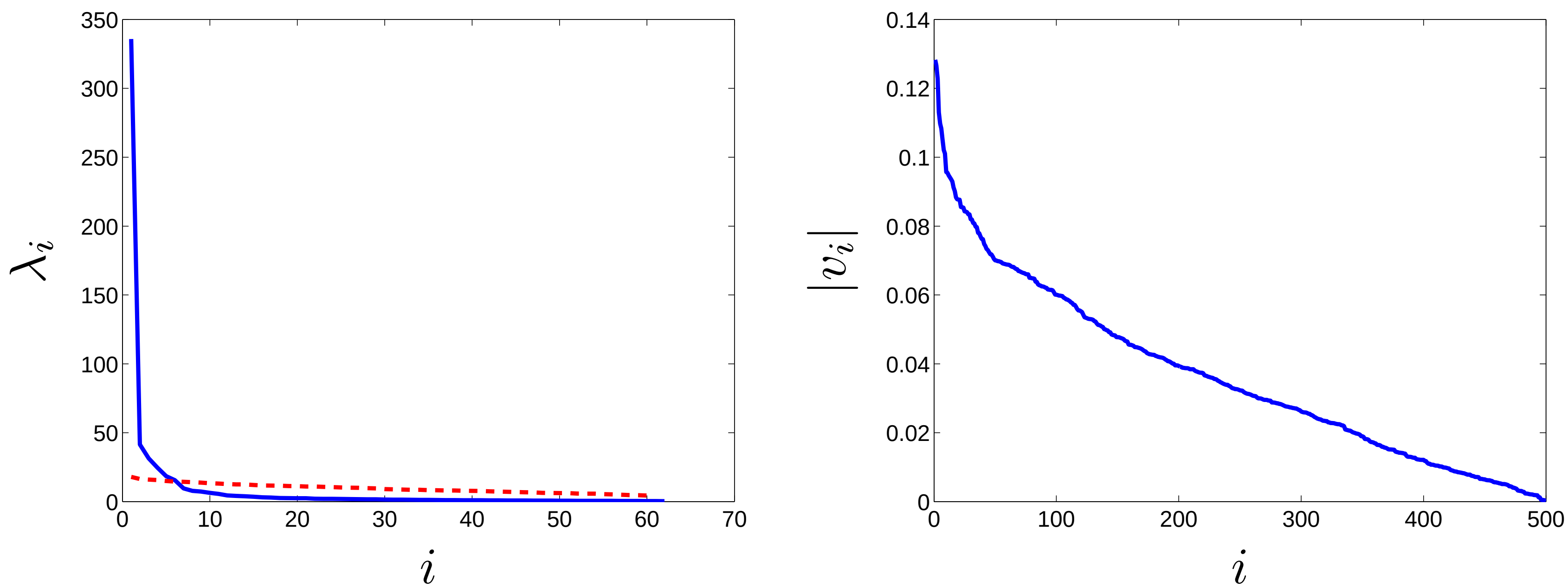
$$\begin{cases} \mathcal{H}_0: & x \sim \mathcal{N}\left(0, \mathbf{I}_p\right) \\ \\ \mathcal{H}_1: & x \sim \mathcal{N}\left(0, \mathbf{I}_p + \theta v v^T\right) \end{cases}$$

where $\theta > 0$ and $\|v\|_2 = 1$.

- Of course $\lambda_{\max}(\mathbf{I}_p) = 1$ and $\lambda_{\max}(\mathbf{I}_p + \theta v v^T) = 1 + \theta$, so we can use $\lambda_{\max}(\cdot)$ as our test statistic.
- However, [Baik et al., 2005, Tao, 2011, Benaych-Georges et al., 2011] show that
$$\lambda_{\max}\left(\frac{X^T X}{n-1}\right) \to (1 + \sqrt{\gamma})^2$$
under both $\mathcal{H}_0$ and $\mathcal{H}_1$ when θ 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

$$\begin{array}{rll} \lambda_{\max}^k(\Sigma) \triangleq & \text{max.} & x^T \Sigma x \\ & \text{s.t.} & \mathbf{Card}(x) \leq k \\ & & \|x\|_2 = 1, \end{array}$$

- We focus on the **sparse eigenvector detection** problem

$$\left\{ \begin{array}{ll} \mathcal{H}_0 : & x \sim \mathcal{N}\left(0, \mathbf{I}_p\right) \\ \mathcal{H}_1 : & x \sim \mathcal{N}\left(0, \mathbf{I}_p + \theta v v^T\right) \end{array} \right.$$

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

- We naturally have

$$\lambda_{\max}^k(\mathbf{I}_p) = 1 \quad \text{and} \quad \lambda_{\max}^k(\mathbf{I}_p + \theta v v^T) = 1 + \theta$$

Sparse PCA

Berthet and Rigollet [2012]: **Optimal detection threshold** using $\lambda^k_{\max}(\cdot)$ is

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

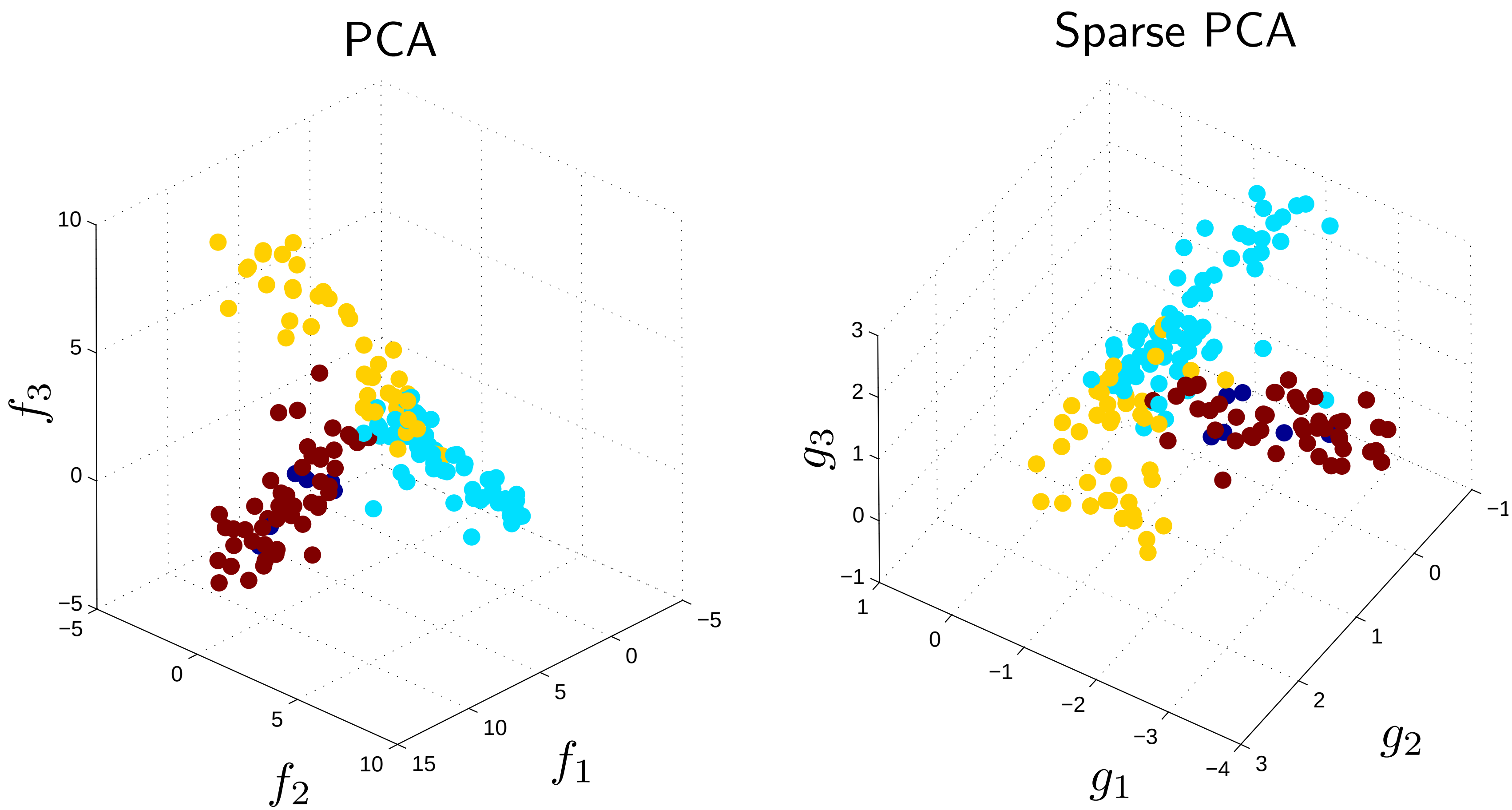
- **Good news:** $\lambda^k_{\max}(\cdot)$ is a **minimax optimal statistic** for detecting sparse principal components. The dimension p only appears as a **log term** and this threshold is much better than $\theta = \sqrt{p/n}$ in the dense PCA case.
- **Bad news:** Computing the statistic $\lambda^k_{\max}(\hat{\Sigma})$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

A graphical output



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.

$$\mathrm{SPCA}(\rho) \triangleq \max_{\|x\|_2=1} \; x^T\Sigma x - \rho\,\mathbf{Card}(x)$$

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

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

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

$$\mathrm{SDP}(\rho) \triangleq \begin{array}{ll} \text{max.} & \sum_{i=1}^p \mathbf{Tr}(X^{1/2}a_i a_i^T X^{1/2} - \rho X)_+ \\ \text{s.t.} & \mathbf{Tr}(X) = 1,\; X \succeq 0, \end{array}$$

which is equivalent to a semidefinite program.

Approximation bounds for sparse eigenvalues

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

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

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

where

$$\mathrm{SDP}(\rho) = \begin{array}{ll} \textit{max.} & \sum_{i=1}^{p} \mathbf{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ \\ \textit{s.t.} & \mathbf{Tr}(X) = 1,\ X \succeq 0. \end{array}$$

Proof sketch. **Change variables**, set $X = xx^T$, so $\|x\|_2 = 1$ means $\mathbf{Tr}(X) = 1$ 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**

$$\begin{array}{rcl}
(a_i^T X a_i - \rho)_+ & = & \mathbf{Tr}((a_i^T x x^T a_i - \rho) x x^T)_+ \\
& = & \mathbf{Tr}(x(x^T a_i a_i^T x - \rho) x^T)_+ \quad (\mathbf{Tr}(\cdot)_+ = \lambda_{\max}(\cdot) \text{ here}) \\
& = & \mathbf{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ = \mathbf{Tr}(X^{1/2}(a_i a_i^T - \rho \mathbf{I}) X^{1/2})_+.
\end{array}$$

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

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

concave in X as a pointwise minimum of affine functions.

$$\begin{array}{rll}
\mathrm{SPCA}(\rho) = & \text{max.} & \sum_{i=1}^n \mathbf{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ \\
& \text{s.t.} & \mathbf{Tr}(X) = 1, \ \mathbf{Rank}(X) = 1, \ X \succeq 0,
\end{array}$$

We relax the original problem into a semidefinite program by simply dropping the rank constraint.

Approximation bounds for sparse eigenvalues

Proposition 2 [d'Aspremont, Bach, and El Ghaoui, 2012]

Approximation ratio on $\mathrm{SDP}(\rho)$. *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*

$$\begin{array}{rll} \mathrm{SDP}(\rho) = & \textit{max.} & \sum_{i=1}^p \mathbf{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ \\ & \textit{s.t.} & \mathbf{Tr}(X) = 1,\; X \succeq 0, \end{array}$$

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

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

where

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

controls the approximation ratio.

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

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

$$\begin{aligned}
\mathbf{E}\left[\left((a_i^T z)^2 - \rho\|z\|_2^2\right)_+\right] &= \mathbf{E}\left[\left(\xi^T B_i(X)\xi\right)_+\right] \\
&= \mathbf{E}\left[\left(\alpha_i \xi_1^2 - \textstyle\sum_{j=2}^r \beta_j^i \xi_j^2\right)_+\right]
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{E}\left[(\xi^T B_i(X)\xi)_+\right] &\geq \min_\beta \left\{ \mathbf{E}\left[\left(\alpha_i \xi_1^2 - \textstyle\sum_{j=2}^r \beta_j^i \xi_j^2\right)_+\right] : \ \textstyle\sum_{j=2}^r \beta_j^i \leq \rho,\ \beta_j^i \geq 0 \right\} \\
&= \mathbf{E}\left[\left(\alpha_i \xi_1^2 - \tfrac{\rho}{r-1} \textstyle\sum_{j=2}^r \xi_j^2\right)_+\right],
\end{aligned}$$

by convexity and symmetry.

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

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

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

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

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

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

because $\mathrm{SPCA}(\rho) = \max_{\|z\|_2=1} \sum_{i=1}^{n}((a_i^T z)^2 - \rho)_+ \quad \blacksquare$

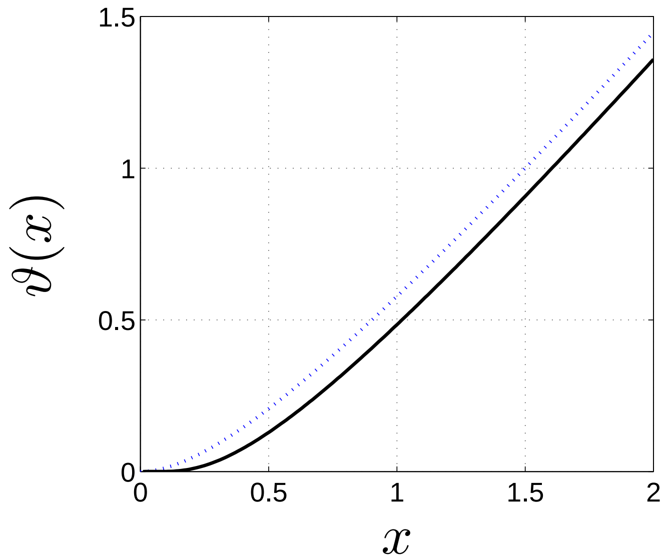
Approximation bounds for sparse eigenvalues

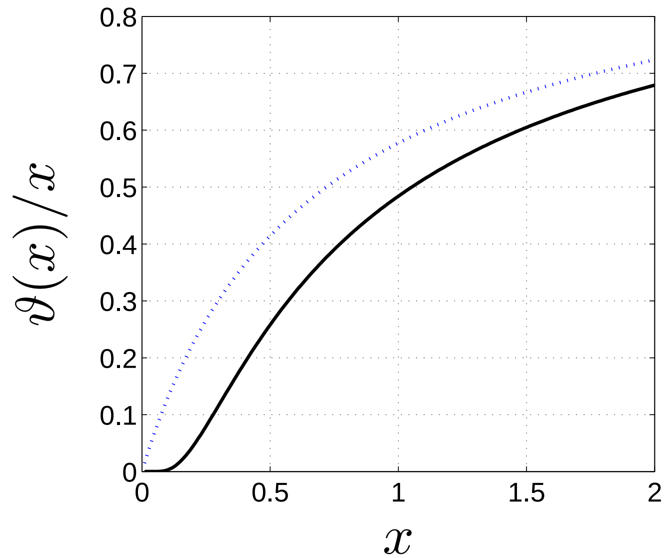
- By convexity, we also have $\vartheta_r(x) \geq \vartheta(x)$, where

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

- Overall, we have the following **approximation bounds**

$$\frac{\vartheta(c)}{c}\mathrm{SDP}(\rho) \leq \mathrm{SPCA}(\rho) \leq \mathrm{SDP}(\rho), \quad \text{when } c \leq \frac{\mathrm{SDP}(\rho)}{p\rho}.$$





Conclusion

- 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

$$\frac{\mathrm{SDP}(\rho)}{p\rho}$$

- For the detection problem, when γ is small enough the **approximation ratio** is of order one.

*

References

A. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology*, 96:6745–6750, 1999.

A.A. Amini and M. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):2877–2921, 2009.

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

F. Benaych-Georges, A. Guionnet, and M. Maida. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.*, 16:no. 60, 1621–1662, 2011. ISSN 1083-6489. doi: 10.1214/EJP.v16-929. URL `http://dx.doi.org/10.1214/EJP.v16-929`.

Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Arxiv preprint arXiv:1202.5070*, 2012.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Approximation bounds for sparse principal component analysis. *ArXiv: 1205.0121*, 2012.

S. Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, 1980.

I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, pages 295–327, 2001.

N.E. Karoui. On the largest eigenvalue of wishart matrices with identity covariance when n, p and p/n tend to infinity. *Arxiv preprint math/0309355*, 2003.

V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR - Sbornik*, 1(4): 457–483, 1967.

T. Tao. Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probability Theory and Related Fields*, pages 1–33, 2011.

YQ Yin, ZD Bai, and PR Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78(4):509–521, 1988.

U. Zwick. Outward rotations: a tool for rounding solutions of semidefinite programming relaxations, with applications to max cut and other problems. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 679–687. ACM, 1999.