Approximation Bounds for Sparse Principal Component Analysis

Alexandre d'Aspremont, *CNRS & Ecole Polytechnique*.

With **Francis Bach**, *INRIA-ENS* and **Laurent El Ghaoui**, *U.C. Berkeley*.

Support from NSF, ERC and Google.

Introduction

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.
- Test the significance of principal eigenvalues.

Introduction

Sample covariance matrix 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

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

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

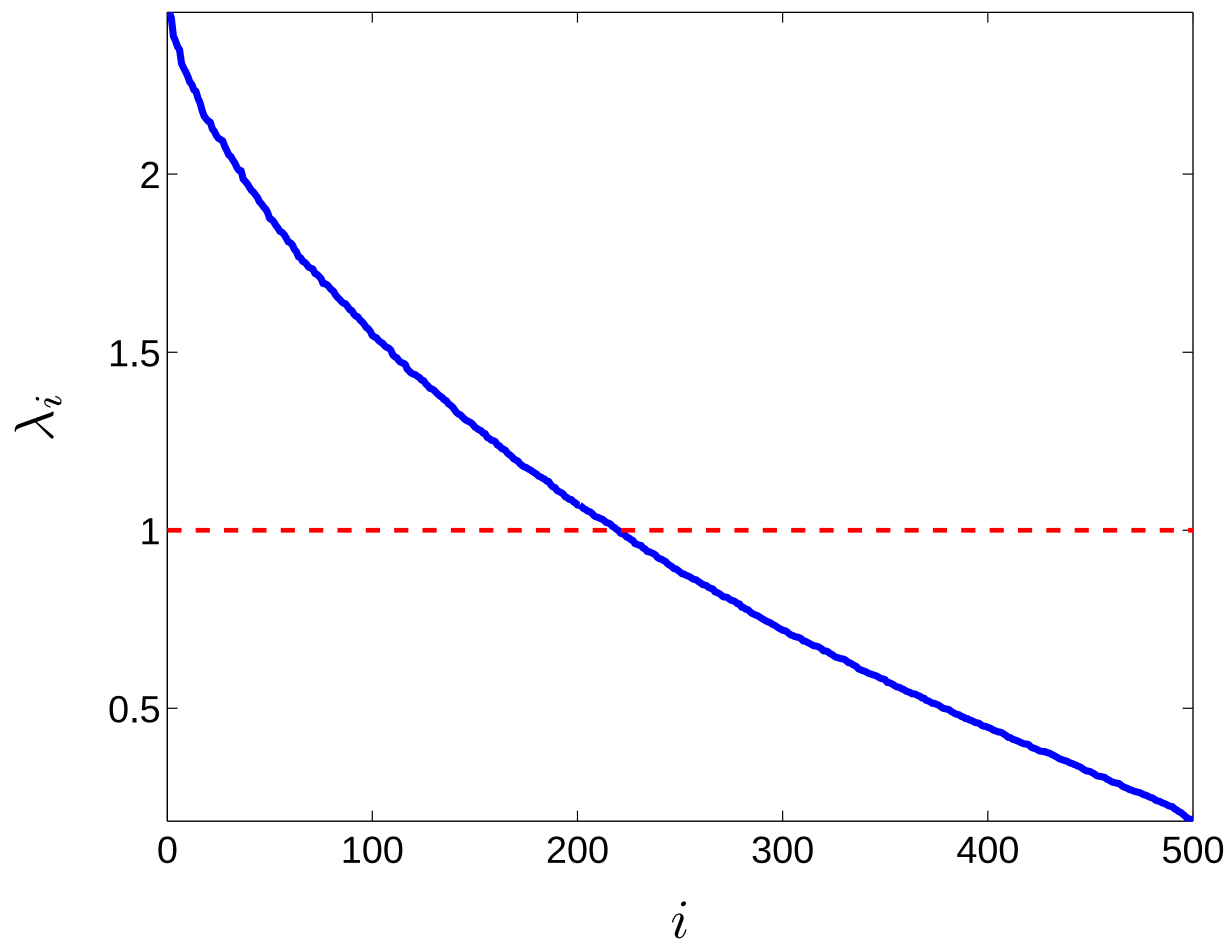
- When $\gamma \in (0,1]$, the spectral measure converges to the following density

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

  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.

Introduction



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

Introduction

We focus on the following **hypothesis testing problem**

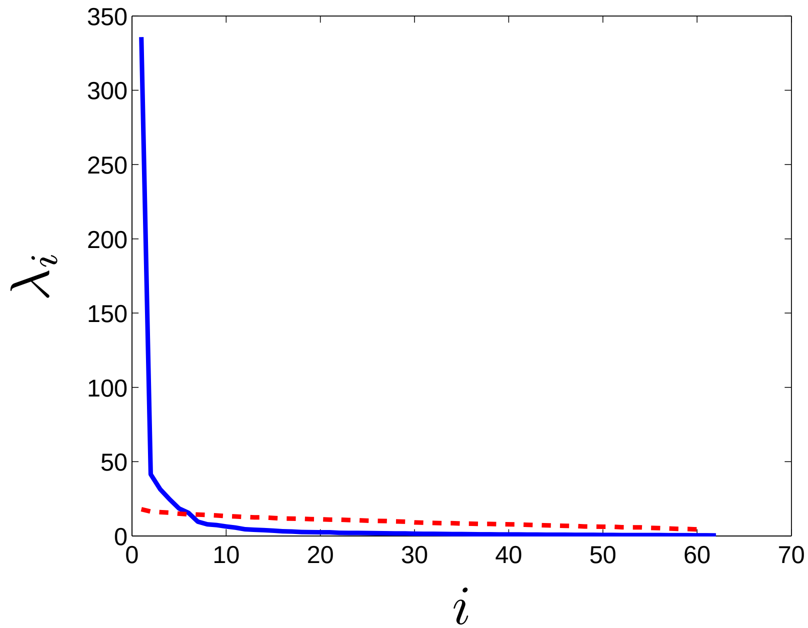
$$\begin{cases} \mathcal{H}_0 : & x \sim \mathcal{N}(0, \mathbf{I}_p) \\ \mathcal{H}_1 : & x \sim \mathcal{N}\left(0, \mathbf{I}_p + \theta v v^T\right) \end{cases}$$

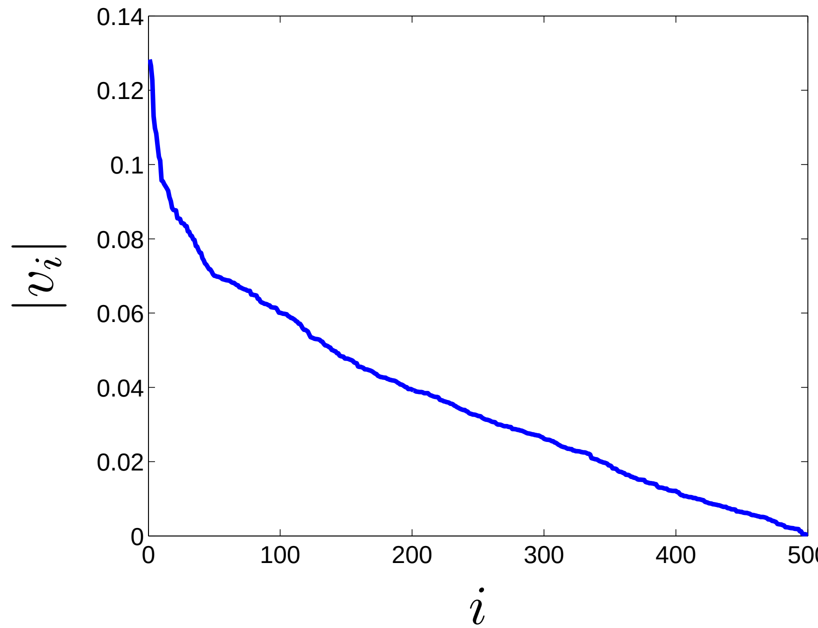
where $\theta > 0$ and $\|v\|_2 = 1$.

- Of course
$$\lambda_{\max}(\mathbf{I}_p) = 1 \quad \text{and} \quad \lambda_{\max}(\mathbf{I}_p + \theta v v^T) = 1 + \theta$$
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 when θ **is small**, i.e.
$$\theta \leq \gamma + \sqrt{\gamma}$$
then
$$\lambda_{\max}\left(\frac{X^T X}{n-1}\right) \to (1 + \sqrt{\gamma})^2$$
under both $\mathcal{H}_0$ and $\mathcal{H}_1$ in the high dimensional regime $p = \gamma n$, with $\gamma \in (0, 1)$, $p \to \infty$, and **detection using $\lambda_{\max}(\cdot)$ fails.**

Introduction

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.

Introduction

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

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

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

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

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

- We naturally have

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

Introduction

Berthet and Rigollet [2012] show the following results on the detection threshold

- **Under $\mathcal{H}_1$:**

$$\lambda^k_{\max}(\hat{\Sigma}) \geq 1 + \theta - 2(1+\theta)\sqrt{\frac{\log(1/\delta)}{n}}$$

  with probability $1-\delta$.

- **Under $\mathcal{H}_0$:**

$$\lambda^k_{\max}(\hat{\Sigma}) \leq 1 + 4\sqrt{\frac{k\log(9ep/k)+\log(1/\delta)}{n}} + 4\frac{k\log(9ep/k)+\log(1/\delta)}{n}$$

  with probability $1-\delta$.

This means that the **detection threshold** is

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

which is **minimax optimal** [Berthet and Rigollet, 2012, Th. 5.1].

Sparse PCA

Optimal detection threshold using $\lambda^k_{\max}(\cdot)$ is

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

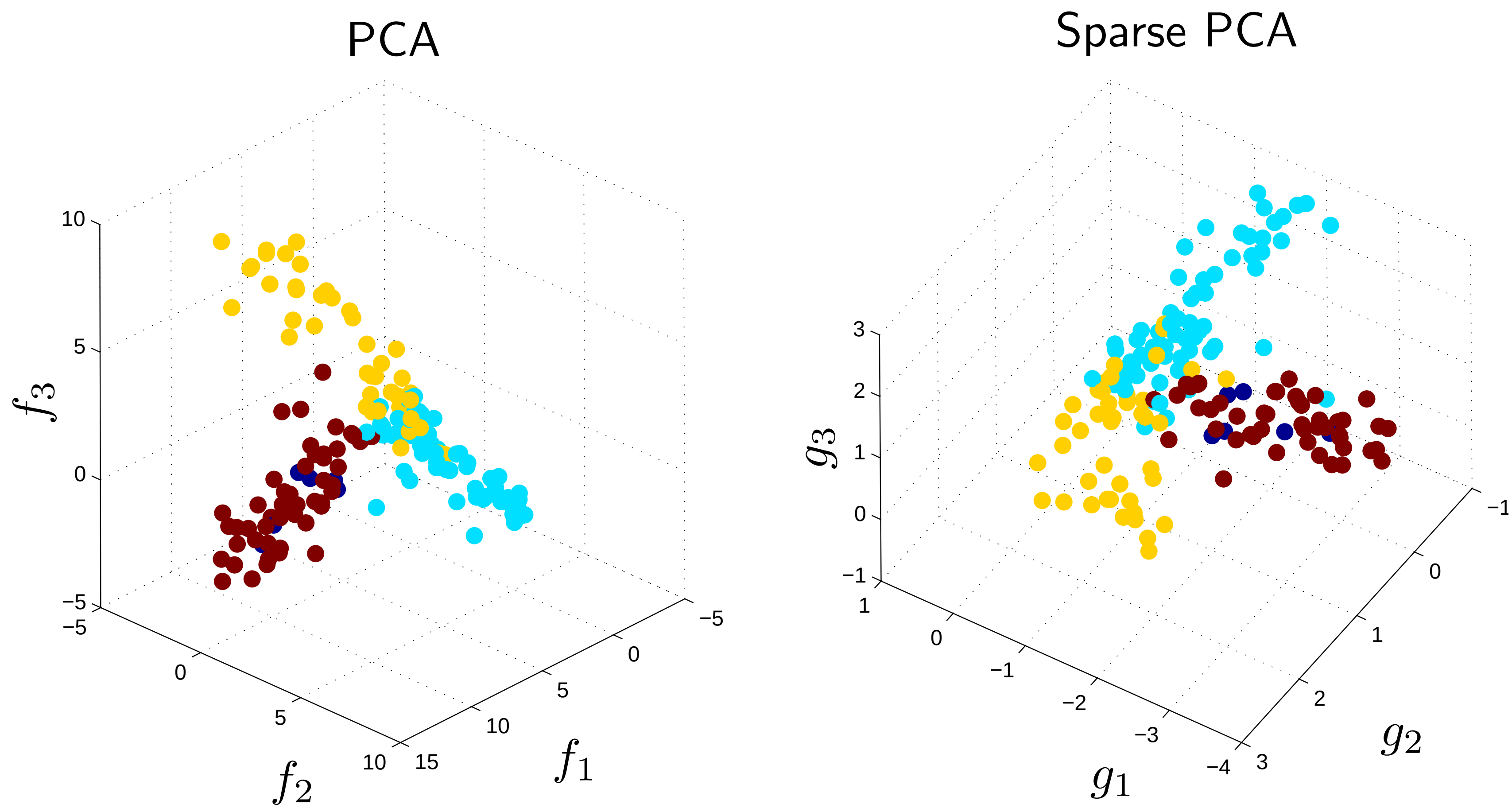
- **Good news:** $\lambda^k_{\max}(\cdot)$ is a **minimax optimal statistic** for detecting sparse principal components. The dimension p only appears as a **log term** and this threshold is much better than $\theta = \sqrt{p/n}$ in the dense PCA case.
- **Bad news:** Computing the statistic $\lambda^k_{\max}(\hat{\Sigma})$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

Sparse PCA

Sparse regression: Lasso, Dantzig selector, sparsity inducing penalties. . .

- Sparse, ℓ_0 constrained regression is NP-hard.
- Efficient ℓ_1 **convex relaxations**, good bounds on statistical performance.
- These convex relaxations are **optimal**. **No further fudging required.**

Sparse PCA.

- Computing $\lambda^k_{\max}(\cdot)$ is NP-hard.
- Several algorithms & convex relaxations. [Zou et al., 2006, d'Aspremont et al., 2007, 2008, Amini and Wainwright, 2009, Journée et al., 2008, Berthet and Rigollet, 2012]
- Statistical performance mostly unknown so far.
- Optimality of convex relaxation?

Detection problems are a good **testing ground for convex relaxations**. . .

Outline

- PCA on high-dimensional data
- **Approximation bounds for sparse eigenvalues**
- Tractable detection for sparse PCA
- Algorithms
- Numerical results

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

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

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

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

which is equivalent to a semidefinite program [d'Aspremont et al., 2008].

Approximation bounds for sparse eigenvalues

Proposition 1. [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.

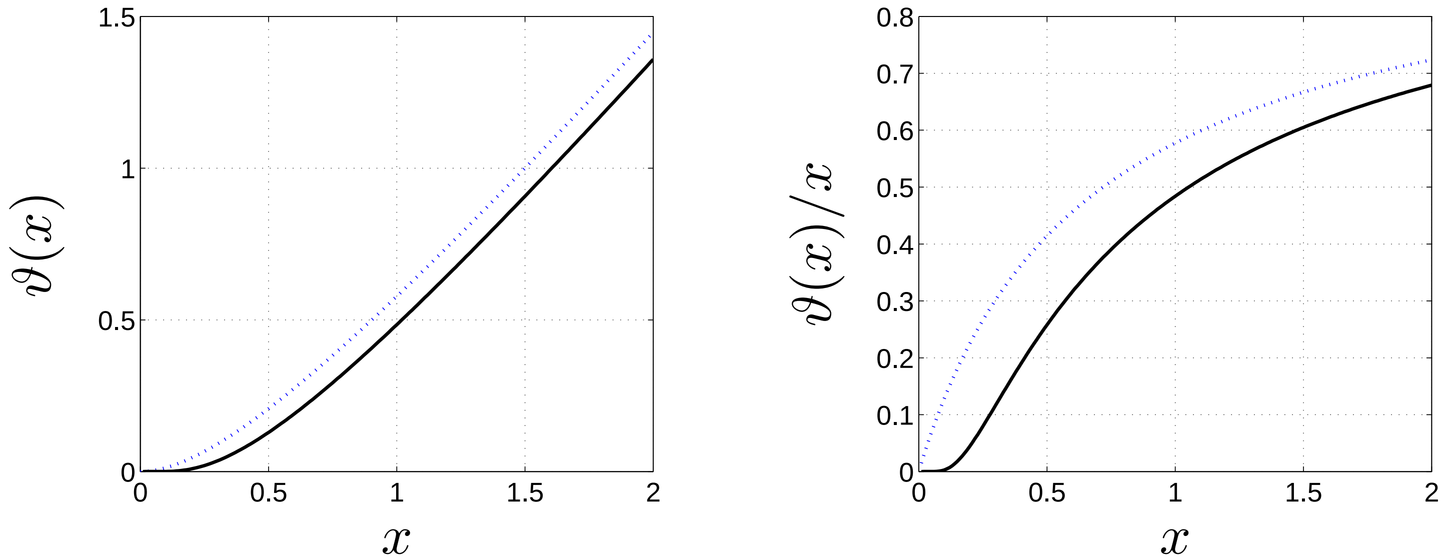
Approximation bounds for sparse eigenvalues

- 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**

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



Approximation bounds for sparse eigenvalues

Approximation ratio.

- 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

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

- Can we control this ratio for interesting problem instances?

Outline

- PCA on high-dimensional data
- Approximation bounds for sparse eigenvalues
- **Tractable detection for sparse PCA**
- Algorithms
- Numerical results

Approximation bounds for sparse eigenvalues

We focus again on the **sparse eigenvector detection** 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 $\|v\|_2 = 1$ with $\mathbf{Card}(v) = k$.

- Study the statistic $\mathrm{SPCA}(\rho)$

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

  under these two hypotheses.

- Bound the approximation ratio

$$\frac{\vartheta\left(\frac{\mathrm{SDP}(\rho)}{p\rho}\right)}{\frac{\mathrm{SDP}(\rho)}{p\rho}}$$

  for the testing problem above.

Approximation bounds for sparse eigenvalues

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

Detection threshold for $\mathrm{SPCA}(\rho)$. *Suppose we set*

$$\Delta = 4\log(9ep/k) + 4\log(1/\delta) \quad \textit{and} \quad \rho = \frac{\Delta}{n} + \frac{\Delta}{\sqrt{kn(\Delta + 4/e)}}$$

and define θ_{SPCA} such that

$$\theta_{\mathrm{SPCA}} = 2\sqrt{\frac{k(\Delta + 4/e)}{n}} + \ldots$$

then if $\theta > \theta_{\mathrm{SPCA}}$ in the Gaussian model, the test statistic based on $\mathrm{SPCA}(\rho)$ discriminates between $\mathcal{H}_0$ and $\mathcal{H}_1$ with probability $1 - 3\delta$.

Proof: Result in Berthet and Rigollet [2012] and union bounds.

Approximation bounds for sparse eigenvalues

Proposition 3. [d'Aspremont, Bach, and El Ghaoui, 2012]

Detection threshold for $\mathrm{SDP}(\rho)$. *Suppose $p = \gamma n$ and $k = \kappa p$, where $\gamma > 0$, $\kappa \in (0,1)$ are fixed and p is large. Define the detection threshold θ_{SDP} such that*

$$\theta_{\mathrm{SDP}} \geq \beta(\gamma,\kappa)^{-1}\,\theta_{\mathrm{SPCA}}$$

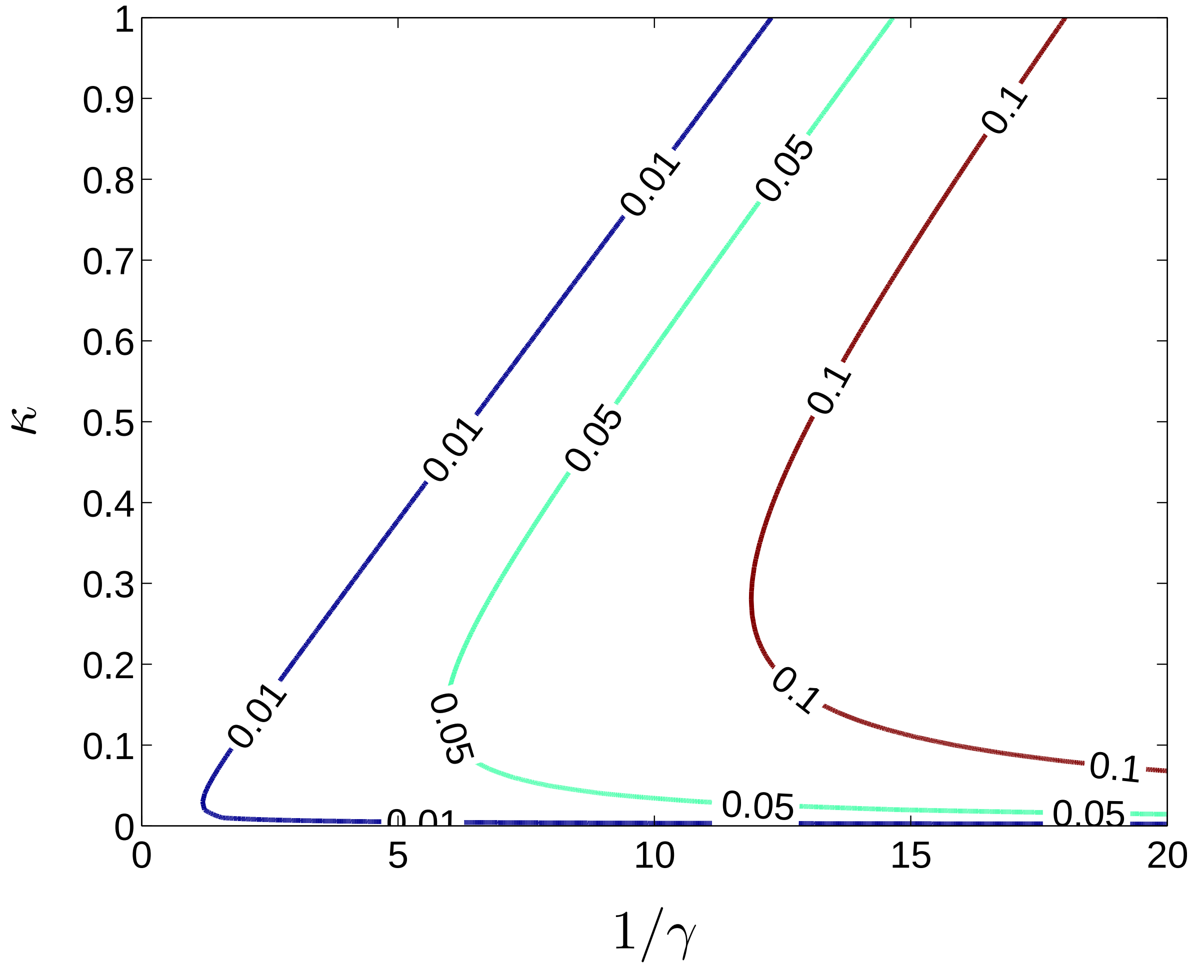
where

$$\beta(\mu,\kappa) = \frac{\vartheta(c)}{c} \quad \textit{where} \quad c = \frac{1 - \gamma\Delta\kappa - \frac{\sqrt{\gamma\kappa}}{\sqrt{(\Delta+4/e)}} - 2\sqrt{\frac{\log(1/\delta)}{n}}}{\gamma\Delta + \frac{\gamma\Delta}{\sqrt{\kappa(\Delta+4/e)}}},$$

then if $\theta > \theta_{\mathrm{SDP}}$ the test statistic based on $\mathrm{SDP}(\rho)$ discriminates between $\mathcal{H}_0$ and $\mathcal{H}_1$ with probability $1 - 3\delta$.

Proof: Setting $p\rho = \gamma\Delta + \frac{\gamma\Delta}{\sqrt{\kappa(\Delta+4/e)}}$ the approx. ratio is bounded by $\beta(\gamma,\kappa)$.

Approximation bounds for sparse eigenvalues



Level sets of $\beta(\gamma, \kappa)$ for $\Delta = 5$. Assuming $p = \gamma n$ and $k = \kappa p$.

Approximation bounds for sparse eigenvalues

- In the regime detailed above, the **detection threshold remains bounded** when $k \to \infty$. In [Berthet and Rigollet, 2012], $\theta \to \infty$ when $k \to \infty$.
- For our choice of ρ, the approximation ratio blows up when $\kappa \to 0$. Easy to fix: Another good guess for ρ when κ is small is to pick

$$\rho = \frac{1}{p}$$

  so the approximation ratio is of order one.
- The detection threshold for $\mathrm{SDP}(\rho)$ is then of order

$$\left(1 + \frac{4}{e\Delta}\right)\kappa + \frac{\gamma\Delta}{1 - \gamma\Delta} \simeq \left(1 + \frac{4}{e\Delta}\right)\kappa + \gamma\Delta$$

  when both γ, κ are small.
- This should be compared with the detection threshold for $\lambda_{\max}(\cdot)$ from [Benaych-Georges et al., 2011] which is $\sqrt{\gamma} + \gamma$.

This (roughly) means $\mathrm{SDP}(\rho)$ achieves γ when $\lambda_{\max}(\cdot)$ fails below $\sqrt{\gamma}$.

Outline

- PCA on high-dimensional data
- Approximation bounds for sparse eigenvalues
- Tractable detection for sparse PCA
- **Algorithms**
- Numerical results

Computing $\mathrm{SDP}(\rho)$**.** We can bound $\mathrm{SDP}(\rho)$

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

by solving the **dual**

$$\begin{array}{ll}\text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{p} Y_i\right) \\ \text{subject to} & Y_i \succeq a_i a_i^T - \rho \mathbf{I} \\ & Y_i \succeq 0, \quad i = 1, \ldots, p\end{array}$$

in the variables $Y_i \in \mathbf{S}_p$.

- Maximum eigenvalue minimization problem.
- p matrix variables of dimension p. . .

Frank-Wolfe algorithm for computing $\mathrm{SDP}(\rho)$.

Input: $\rho > 0$ and a feasible starting point Z_0.

for $k = 1$ to N_{max} **do**

  Compute $X = \nabla f(Z)$, together with X^{-1} and $X^{1/2}$.

  Solve the n subproblems

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(Y_i X) \\ \text{subject to} & Y_i \succeq a_i a_i^T - \rho \mathbf{I} \\ & Y_i \succeq 0, \end{array} \tag{1}$$

  in the variables $Y_i \in \mathbf{S}_n$ for $i = 1, \ldots, n$.

  Compute $W = \sum_{i=1}^n Y_i$.

  Update the current point, with

$$Z_k = \left(1 - \frac{2}{k+2}\right) Z_{k-1} + \frac{2}{k+2} W,$$

end for

Output: A matrix $Z \in \mathbf{S}_n$.

Algorithms

Iteration complexity.

- Given X^{-1} and $X^{1/2}$, the p **minimization subproblems**

$$\begin{array}{ll}\mbox{minimize} & \mathbf{Tr}(Y_i X)\\ \mbox{subject to} & Y_i \succeq a_i a_i^T - \rho \mathbf{I}\\ & Y_i \succeq 0,\end{array}$$

  can be **solved in closed form**, with complexity $O(p^2)$.

- The individual **matrices Y_i do not need to be stored**, we only update their sum at each iteration.

- Overall complexity

$$O\left(\frac{D^2 p^3 \log^2 p}{\epsilon^2}\right)$$

  with storage cost $O(p^2)$.

Outline

- PCA on high-dimensional data
- Approximation bounds for sparse eigenvalues
- Tractable detection for sparse PCA
- Algorithms
- **Numerical results**

Numerical results

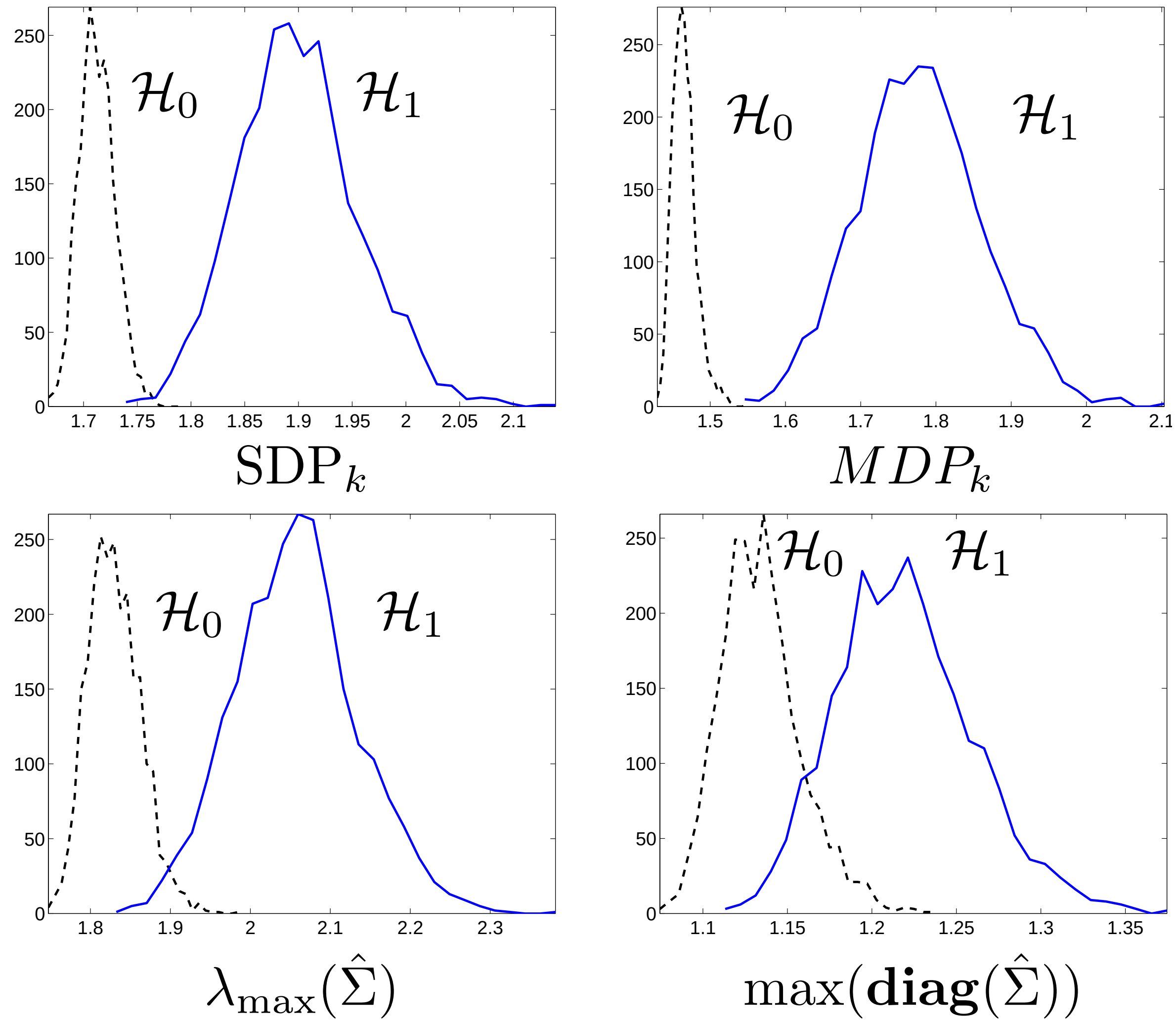
Test the satistic based on $\mathrm{SDP}(\rho)$.

- We generate 3000 experiments, where m points $x_i \in \mathbb{R}^p$ are sampled under both hypotheses, with

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

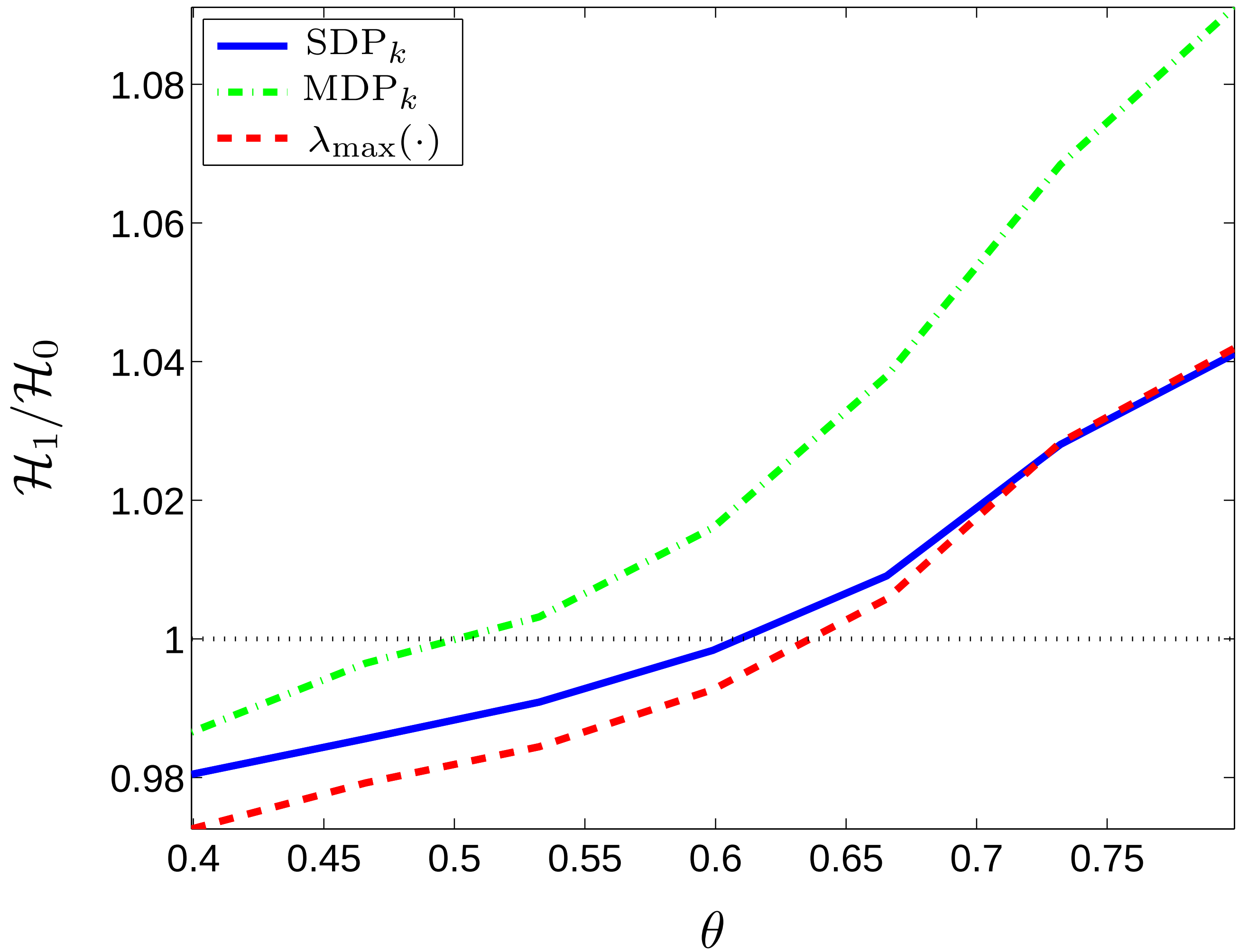
  with $\|v\|_2 = 1$ and $\mathbf{Card}(v) = k$.
- Pick $p = 250$, $n = 1500$ and $k = 10$. We set $\theta = 2/3$, $v_i = 1/\sqrt{k}$ when $i \in [1, k]$ and zero otherwise.
- We compute $\mathrm{SDP}_k \triangleq \min_{\rho > 0} \mathrm{SDP}(\rho) + \rho k$ from several values of $\mathrm{SDP}(\rho)$ around the oracle ρ and $\rho = 0$ (which is $\lambda_{\max}(\hat{\Sigma})$).
- Compare with MDP_k statistic in [Berthet and Rigollet, 2012], similar to DSPCA in [d'Aspremont et al., 2007, Amini and Wainwright, 2009], and diagonal statistic in [Amini and Wainwright, 2009].

Numerical results



Distribution of test statistic SDP_k *(top left)*, the MDP_k statistic in [Berthet and Rigollet, 2012] *(top right)*, the $\lambda_{\max}(\cdot)$ statistic *(bottom left)* and the diagonal statistic from [Amini and Wainwright, 2009] *(bottom right)*.

Numerical results



Ratio of 5% quantile under $\mathcal{H}_1$ over 95% quantile under $\mathcal{H}_0$, versus signal strength θ. When this ratio is larger than one, both type I and type II errors are below 5%.

Conclusion

- Constant approximation bounds for sparse PCA relaxations in high dimensional regimes.
- Explicit, finite bounds on detection threshold when $p \to \infty$.

Open questions. . . .

- More efficient SDP solver.
- Better approximation bounds for κ small? We should handle the case $p >> n$.
- Improved approximation ratio by direct analysis of the problem under $\mathcal{H}_0$?
- **Model Selection:** do we recover the correct sparse eigenvector? See [Amini and Wainwright, 2009] for early results.

*

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, L. El Ghaoui, M.I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

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.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *arXiv:0811.4724*, 2008.

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.

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

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.