Spectral densities: computations and applications in linear algebra

Yousef Saad
Department of Computer Science and Engineering

University of Minnesota

PASC17 – Lugano
June 28, 2017

Introduction

- 'Random Sampling' or 'probabilistic methods': use of random data to solve a given problem.
- Eigenvalues, eigenvalue counts, traces, ...
- Many well-known algorithms use a form of random sampling: The Lanczos algorithm
- Recent work : probabilistic methods - See [Halko, Martinsson, Tropp, 2010]
- Huge interest spurred by 'big data'
- In this talk: Use of random sampling to obtain Eigenvalue counts, spectral densities, and approximate ranks

Important tool: Stochastic Trace Estimator

➤ To estimate diagonal of $B = f(A)$ (e.g., $B = A^{-1}$), let:

Notation:

- $d(B) =$ diag(B) [matlab notation]
- $\odot$ and $\oslash$: Elementwise multiplication and division of vectors
- $\{v_j\}$: Sequence of s random vectors

Result:

$$d(B) \approx \left[\sum_{j=1}^{s} v_j \odot Bv_j\right] \oslash \left[\sum_{j=1}^{s} v_j \odot v_j\right]$$

C. Bekas , E. Kokiopoulou & YS ('05); C. Bekas, A. Curioni, I. Fedulova '09; ...

Trace of a matrix

- For the trace - take vectors of unit norm and

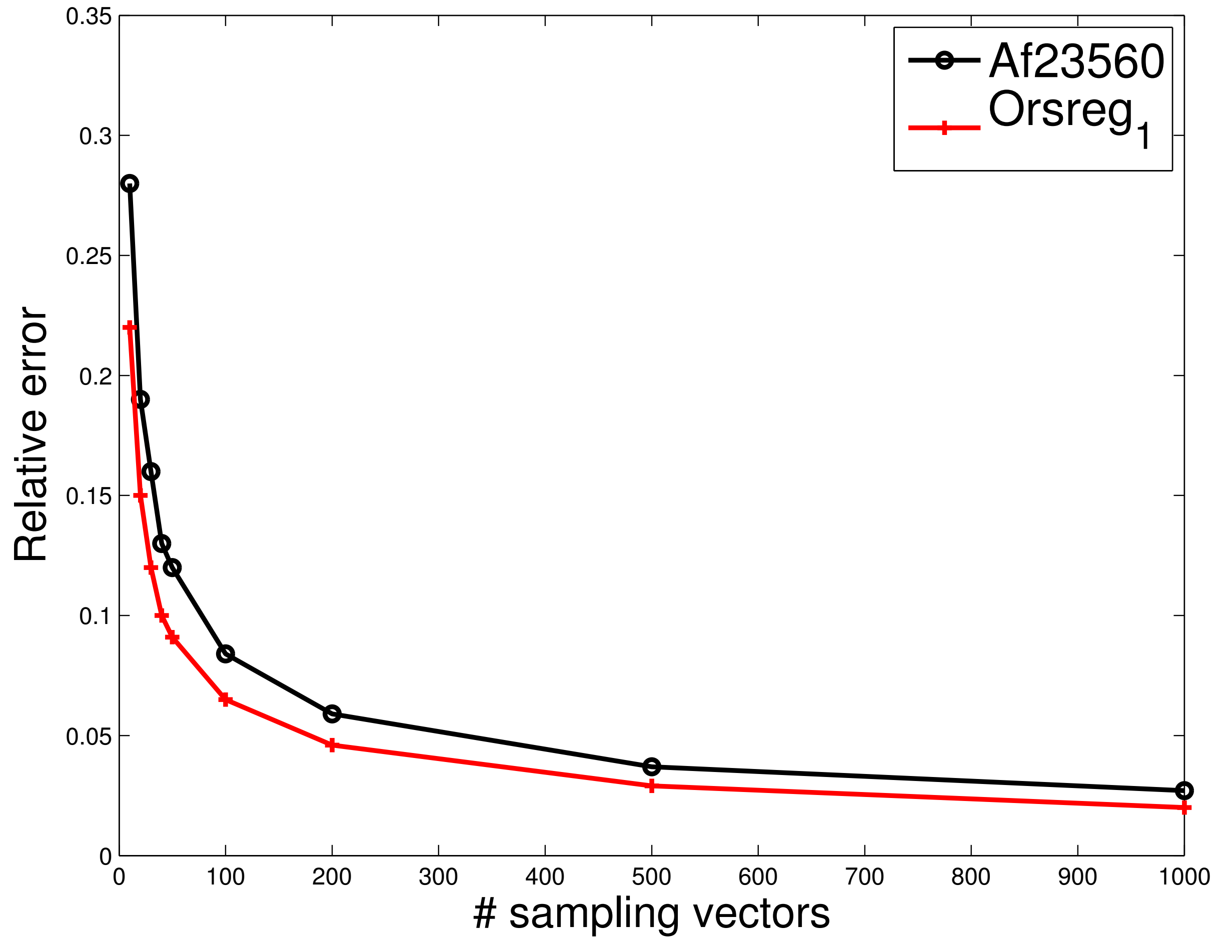
$$\text{Trace}(B) \approx \frac{1}{s} \sum_{j=1}^{s} v_j^T B v_j$$

- Hutchinson's estimator : take random vectors with components of the form $\pm 1/\sqrt{n}$ [Rademacher vectors]

- Extensively studied in literature. See e.g.: Hutchinson '89; H. Avron and S. Toledo '11; G.H. Golub & U. Von Matt '97; Roosta-Khorasani & U. Ascher '15; ...

Typical convergence curve for stochastic estimator

- Estimating the diagonal of inverse of two sample matrices



DENSITY OF STATES & APPLICATIONS

Computing Densities of States [Lin-Lin, Chao Yang, YS]

➤ Formally, the Density Of States (DOS) of a matrix A is

$$\phi(t) = \frac{1}{n}\sum_{j=1}^{n}\delta(t-\lambda_j),$$

where
- δ is the Dirac δ-function or Dirac distribution
- $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ are the eigenvalues of A

➤ Note: $\mu_{[ab]}$ can be obtained from ϕ

➤ $\phi(t)$ == a probability distribution function == probability of finding eigenvalues of A in a given infinitesimal interval near t.

➤ Also known as the spectral density

➤ Very important uses in Solid-State physics

The Kernel Polynomial Method

- Used by Chemists to calculate the DOS – see Silver and Röder'94 , Wang '94, Drabold-Sankey'93, + others
- Basic idea: expand DOS into Chebyshev polynomials
- Coefficients γ_k lead to evaluating Tr $(T_k(A))$
- Use trace estimators [discovered independently] to get traces

A few details:

- Assume change of variable done so eigenvalues lie in $[-1, 1]$.
- Include the weight function in the expansion so expand:

$$\hat{\phi}(t) = \sqrt{1-t^2}\phi(t) = \sqrt{1-t^2} \times \frac{1}{n}\sum_{j=1}^{n}\delta(t-\lambda_j).$$

- Then, (full) expansion is: $\hat{\phi}(t) = \sum_{k=0}^{\infty} \mu_k T_k(t)$.
- Expansion coefficients μ_k are formally defined by:

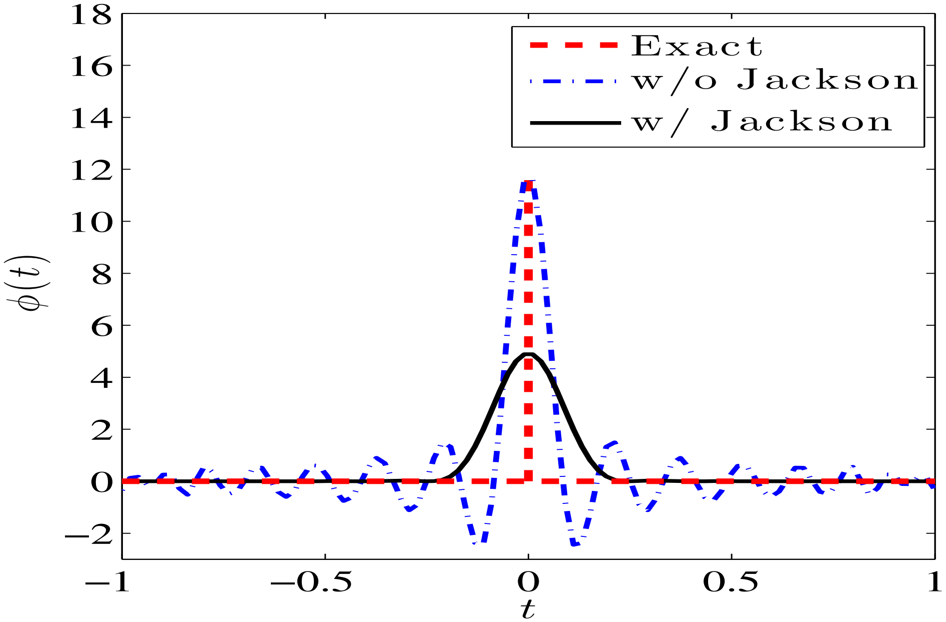
$$\begin{aligned} \mu_k &= \frac{2-\delta_{k0}}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} T_k(t)\hat{\phi}(t)dt \\ &= \frac{2-\delta_{k0}}{\pi} \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} T_k(t)\sqrt{1-t^2}\phi(t)dt \\ &= \frac{2-\delta_{k0}}{n\pi} \sum_{j=1}^{n} T_k(\lambda_j). \qquad \text{with} \quad \delta_{ij} = \text{Dirac symbol} \end{aligned}$$

- Note: $\sum T_k(\lambda_i) = \mathit{Trace}[T_k(A)]$
- Estimate this, e.g., via stochastic estimator

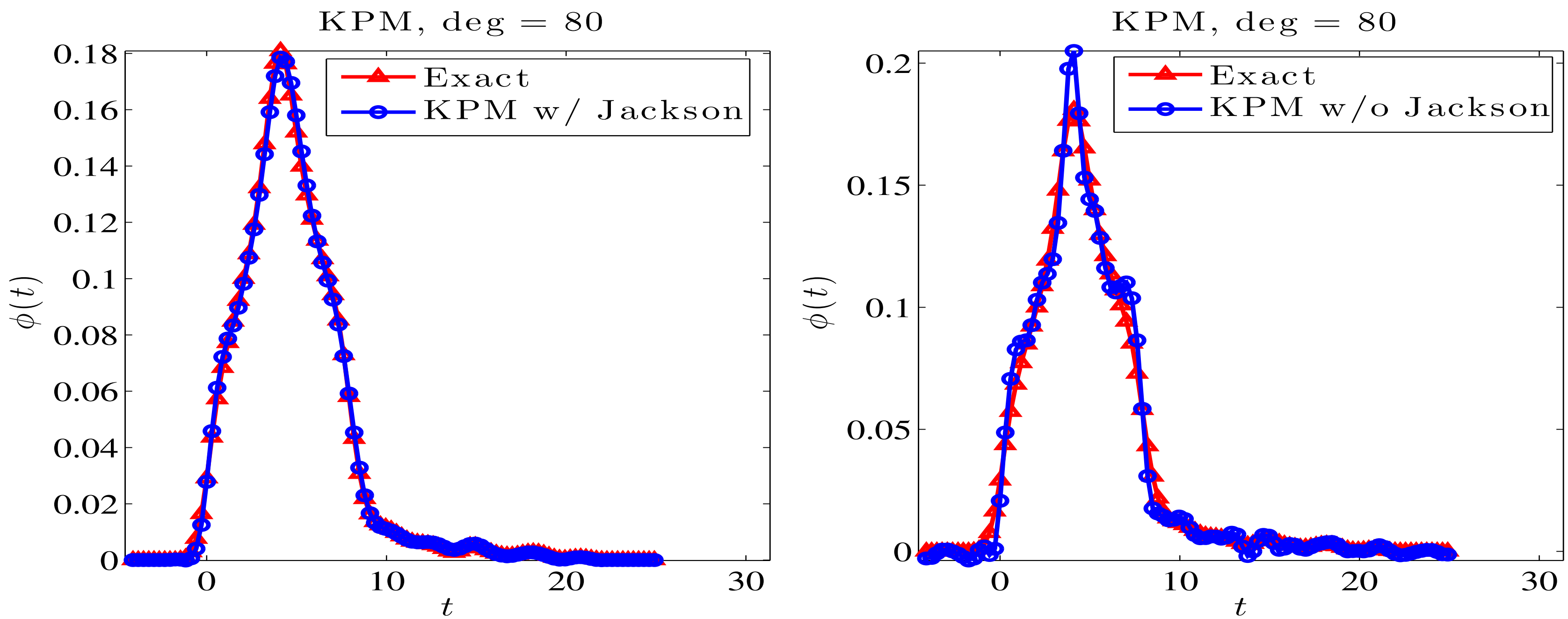
$$\text{Trace}(T_k(A)) \approx \frac{1}{n_{\text{vec}}} \sum_{l=1}^{n_{\text{vec}}} \left(v^{(l)}\right)^T T_k(A) v^{(l)}.$$

- To compute scalars of the form $v^T T_k(A) v$, exploit 3-term recurrence of the Chebyshev polynomial ...

- Use Jackson smoothing for Gibbs oscillations



An example with degree 80 polynomials



Left: Jackson damping; right: without Jackson damping.

Use of the Lanczos Algorithm

- Background: The Lanczos algorithm generates an orthonormal basis $V_m = [v_1, v_2, \cdots, v_m]$ for the Krylov subspace:

$$\text{span}\{v_1, Av_1, \cdots, A^{m-1}v_1\}$$

- ... such than:
$V_m^H A V_m = T_m$ - with

$$T_m = \begin{pmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \alpha_3 & \beta_4 & & \\ & & . & . & . & \\ & & & . & . & . \\ & & & & \beta_m & \alpha_m \end{pmatrix}$$

- Lanczos process builds orthogonal polynomials wrt to dot product:

$$\int \boldsymbol{p(t)q(t)dt \equiv (p(A)v_1, q(A)v_1)}$$

- Let $\boldsymbol{\theta_i,\ i = 1 \cdots, m}$ be the eigenvalues of $\boldsymbol{T_m}$ [Ritz values]
- $\boldsymbol{y_i}$'s associated eigenvectors; Ritz vectors: $\boldsymbol{\{V_m y_i\}_{i=1:m}}$
- Ritz values approximate eigenvalues
- Could compute $\boldsymbol{\theta_i}$'s then get approximate DOS from these
- Problem: $\boldsymbol{\theta_i}$ not good enough approximations – especially inside the spectrum.

Better idea: exploit relation of Lanczos with (discrete) orthogonal polynomials and related Gaussian quadrature:

$$\int p(t)dt \approx \sum_{i=1}^{m} a_i p(\theta_i) \quad a_i = \left[e_1^T y_i\right]^2$$

- See, e.g., Golub & Meurant '93, and also Gautschi'81, Golub and Welsch '69.
- Formula exact when p is a polynomial of degree $\leq 2m + 1$

- Consider now $\int p(t)dt = < p, 1 > =$ (Stieljes) integral $\equiv$

$$\boxed{(p(A)v, v) = \sum \beta_i^2 p(\lambda_i) \equiv < \phi_v, p >}$$

- Then $\langle \phi_v, p \rangle \approx \sum a_i p(\theta_i) = \sum a_i \langle \delta_{\theta_i}, p \rangle \rightarrow$

$$\boxed{\phi_v \approx \sum a_i \delta_{\theta_i}}$$

- To mimick the effect of $\beta_i = 1, \forall i$, use several vectors v and average the result of the above formula over them..

Other methods

➤ The Lanczos spectroscopic approach : A sort of signal processing approach to detect peaks using Fourier analysis

➤ The Delta-Chebyshev approach: Smooth ϕ with Gaussians, then expand Gaussians using Legendre polynomials

➤ Haydock's method: interesting 'classic' approach in physics - uses Lanczos to unravel 'near-poles' of $(A - \epsilon i I)^{-1}$

For details see:

- Approximating spectral densities of large matrices, Lin Lin, YS, and Chao Yang - SIAM Review '16. Also in: [arXiv: http://arxiv.org/abs/1308.5467]

What about matrix pencils?

➤ DOS for generalized eigen-value problems

$$Ax = \lambda Bx$$

➤ Assume: A is symmetric and B is SPD.

➤ In principle: can just apply methods to $B^{-1}Ax = \lambda x$, using B - inner products.

➤ Requires factoring B. Too expensive [Think 3D Pbs]

$\star$ *Observe:* B is usually very *strongly* diagonally dominant.

➤ Especially true after Left+Right Diag. scaling :

$$\tilde{B} = S^{-1}BS^{-1} \qquad S = diag(B)^{1/2}$$

General observation for FEM mass matrices [See, e.g., Wathen'87, Wathen Rees '08]:
* Conforming tetrahedral (P1) elements in 3D $\rightarrow \boldsymbol{\kappa(\tilde{B}) \leq 5}$
* Rectangular bilinear (Q1) elements in 2D $\rightarrow \boldsymbol{\kappa(\tilde{B}) \leq 9}$.

Example: Matrix pair Kuu, Muu from Suite Sparse collection.

➤ Matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ have dimension $\boldsymbol{n = 7,102}$. $\mathtt{nnz}\boldsymbol{(A) = 340,200}$ $\mathtt{nnz}\boldsymbol{(B) = 170,134}$.

➤ After scaling by diagonals to have diag. entries equal to one, all eigenvalues of $\boldsymbol{B}$ are in interval

$$\boldsymbol{[0.6254,\ 1.5899]}$$

Approximation theory to the rescue.

⋆ *Idea:* Compute the DOS for the standard problem

$$B^{-1/2}AB^{-1/2}u = \lambda u$$

➤ Use a very low degree polynomial to approximate $B^{-1/2}$.

➤ We use Chebyshev expansions.

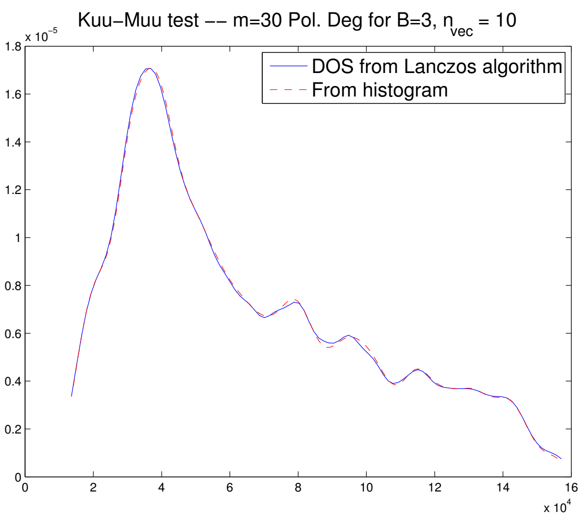
➤ Degree k determined automatically by enforcing

$$\|t^{-1/2} - p_k(t)\|_\infty < tol$$

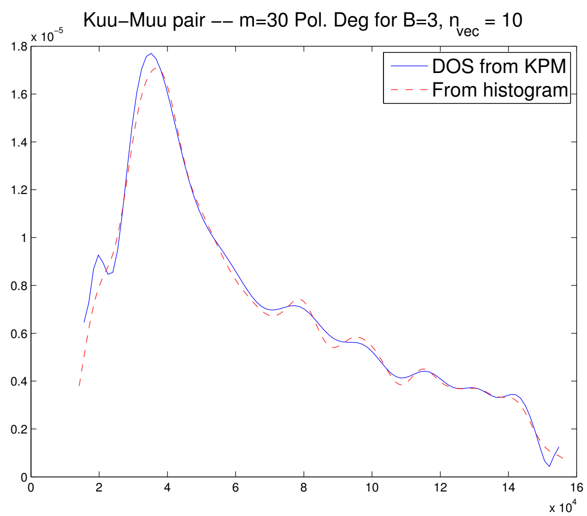
➤ Theoretical results establish convergence that is exponential with respect to degree.

Example: Results for Kuu-Muu example

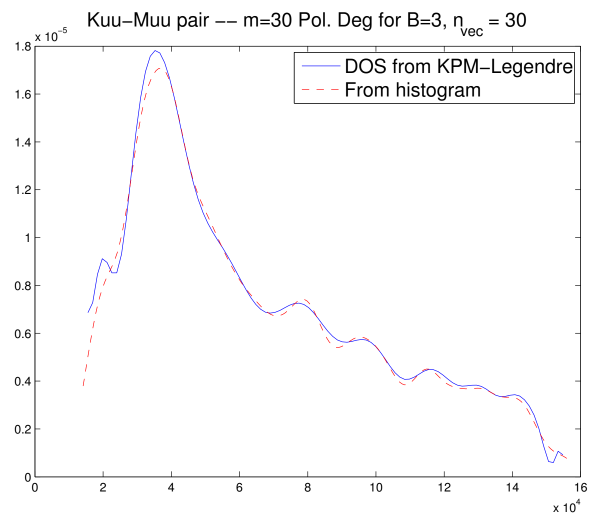
- Using polynomials of degree 3 (!) to approximate $B^{-1/2}$
- Krylov subspace of dim. 30 (== deg. of polynomial in KPM)
- 10 Sample vectors used



Lanczos



KPM-Chebyshev



KPM-Legendre

APPLICATIONS

Application 1: Eigenvalue counts

Problem: Given A (Hermitian) find an estimate of the number $\mu_{[a,b]}$ of eigenvalues of A in $[a, \quad b]$.

Standard method: Sylvester inertia theorem $\rightarrow$ expensive!

First alternative: integrate the Spectral Density in $[a, \ b]$.

$$\mu_{[a,b]} \approx n \left(\int_a^b \tilde{\phi}(t) dt \right) = n \sum_{k=0}^{m} \mu_k \left(\int_a^b \frac{T_k(t)}{\sqrt{1-t^2}} dt \right) = \ldots$$

Second method: Estimate trace of the related spectral projector P ($\rightarrow$ u_i's = eigenvectors $\leftrightarrow$ λ_i's)

$$P = \sum_{\lambda_i \, \in \, [a \ b]} u_i u_i^T.$$

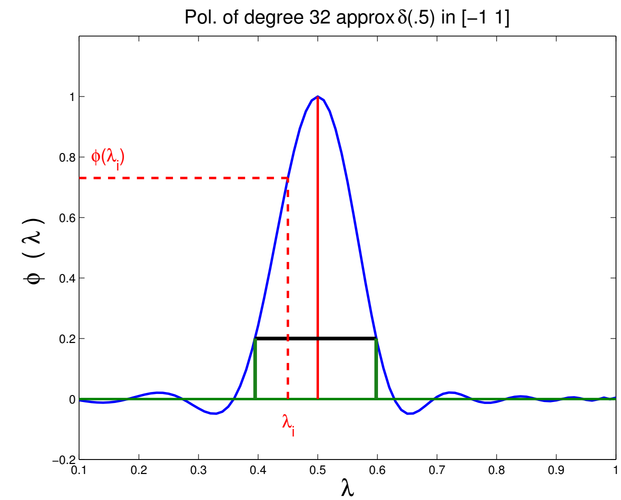
- It turns out that the 2 methods are identical.

Application 2: "Spectrum Slicing"

- ➤ Situation: very large number of eigenvalues to be computed
- ➤ Goal: compute spectrum by slices by applying filtering
- ➤ Apply Lanczos or Subspace iteration to problem:

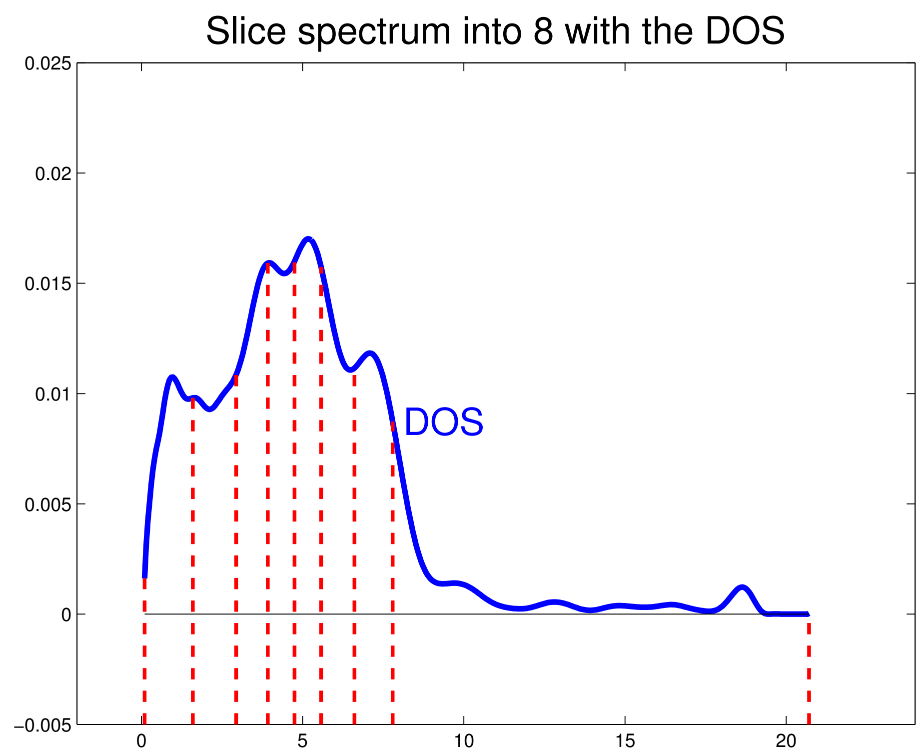
$$\phi(A)u = \mu u$$

$\phi(t) \equiv$ polynomial or rational filter



Rationale. Eigenvectors on both ends of wanted spectrum need not be orthogonalized against each other $\rightarrow$ reduced orthogonalization costs

How do I slice my spectrum?



Answer: Use the DOS.

➤ We must have:

$$\int_{t_i}^{t_{i+1}} \phi(t)dt = \frac{1}{n_{slices}} \int_a^b \phi(t)dt$$

Application 3: Estimating the rank

- Very important problem in signal processing applications, machine learning, etc.

- Often: a certain rank is selected ad-hoc. Dimension reduction is application with this “guessed” rank.

- Can be viewed as a particular case of the eigenvalue count problem - but need a cutoff value..

Approximate rank, Numerical rank

- Notion defined in various ways. A common one:

$$r_\epsilon = \min\{rank(B) : B \in \mathbb{R}^{m\times n}, \|A - B\|_2 \le \epsilon\},$$

$$r_\epsilon = \text{ Number of sing. values } \ge \epsilon$$

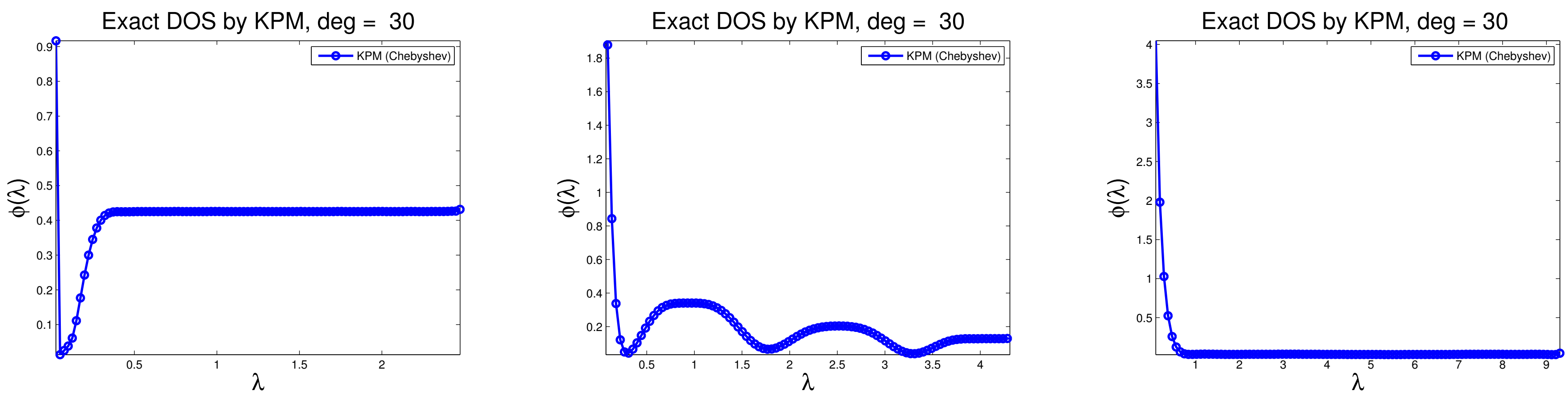
- Two distinct problems:

1. Get a good ϵ 2. Estimate number of sing. values $\ge \epsilon$

- We will need a cut-off value ('threshold') ϵ.
- Could use 'noise level' for ϵ, but not always available

Threshold selection

- How to select a good threshold?
- Answer: Obtain it from the DOS function



(A) (B) (C)

Exact DOS plots for three different types of matrices.

➤ To find: point immediatly following the initial sharp drop observed.

➤ Simple idea: use derivative of DOS function ϕ

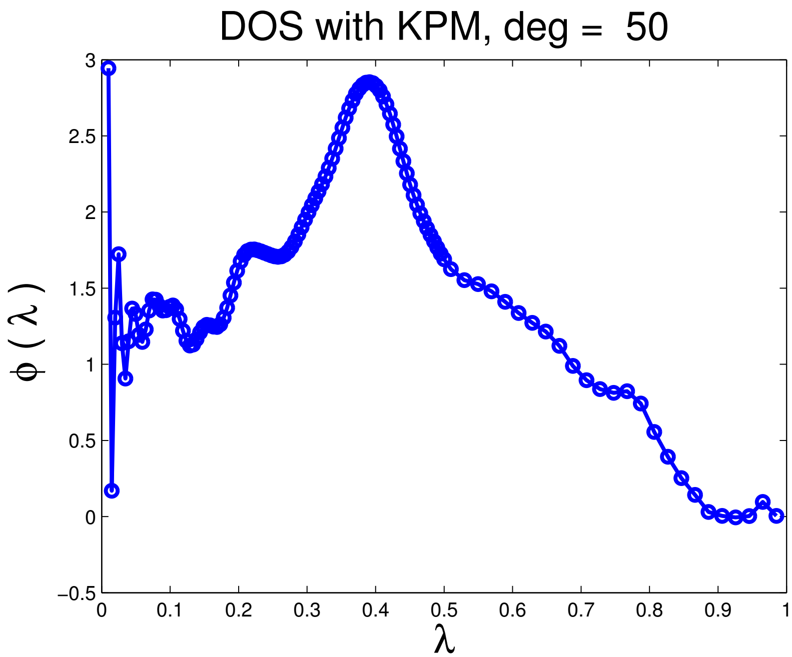
➤ For an $n \times n$ matrix with eigenvalues $\boldsymbol{\lambda}_n \le \boldsymbol{\lambda}_{n-1} \le \cdots \le \boldsymbol{\lambda}_1$:

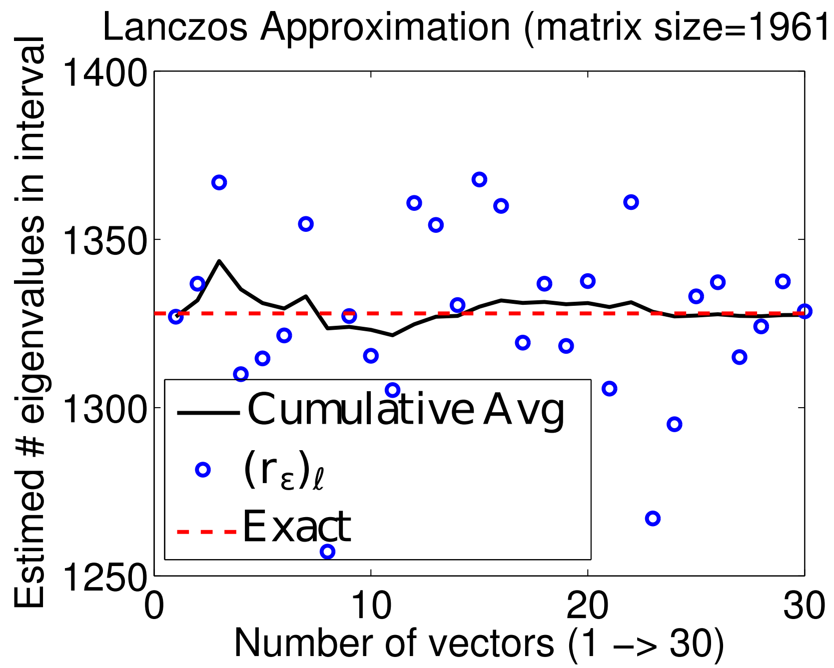
$$\boldsymbol{\epsilon = \min\{t : \lambda_n \le t \le \lambda_1, \phi'(t) = 0\}.}$$

➤ In practice replace by

$$\boldsymbol{\epsilon = \min\{t : \lambda_n \le t \le \lambda_1, |\phi'(t)| \ge \text{tol}\}}$$

Experiments





(A) (B)

(A) The DOS found by KPM.

(B) Approximate rank estimation by The Lanczos method for the example `netz4504`.

Tests with Matérn covariance matrices for grids

- Important in statistical applications

Approximate Rank Estimation of Matérn covariance matrices

Type of Grid (dimension)	Matrix Size	# λ_i's $\geq \epsilon$	r_ϵ	
			KPM	Lanczos
1D regular Grid (2048×1)	2048	16	16.75	15.80
1D no structure Grid (2048×1)	2048	20	20.10	20.46
2D regular Grid (64×64)	4096	72	72.71	72.90
2D no structure Grid (64×64)	4096	70	69.20	71.23
2D deformed Grid (64×64)	4096	69	68.11	69.45

- For all test $M(deg) = 50, n_v$=30

Application 4: The LogDeterminant

Evaluate the Log-determinant of A:

$$\log \det(A) = \text{Trace}(\log(A)) = \sum_{i=1}^{n} \log(\lambda_i).$$

A is SPD.

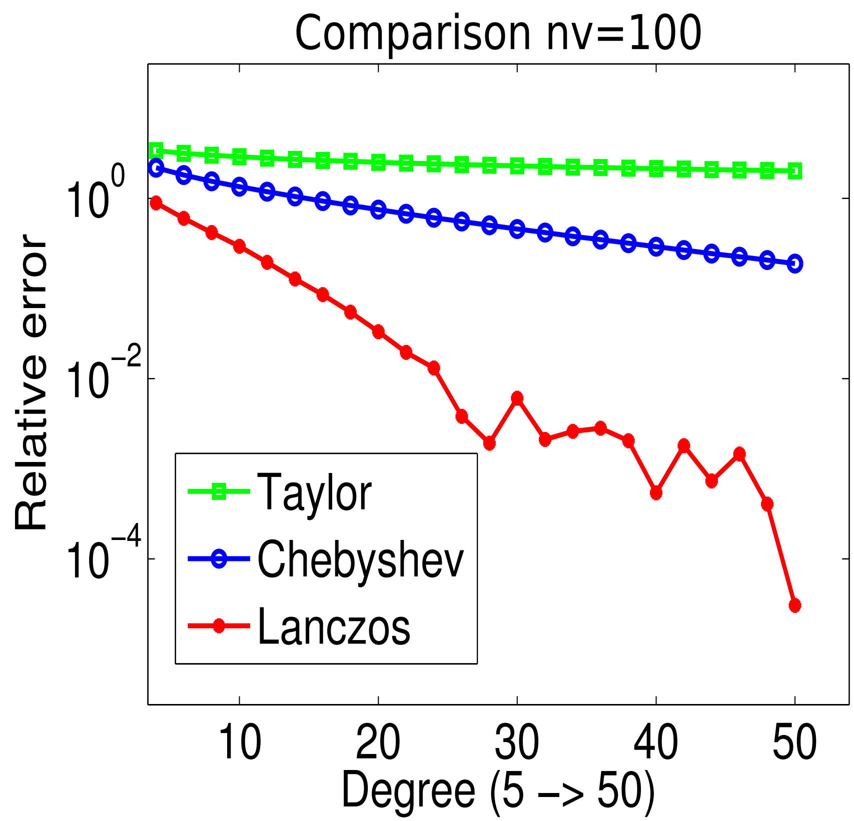
➤ Estimating the log-determinant of a matrix equivalent to estimating the trace of the matrix function $f(A) = \log(A)$.

➤ Can invoke Stochastic Lanczos Quadrature (SLQ) to estimate this trace.

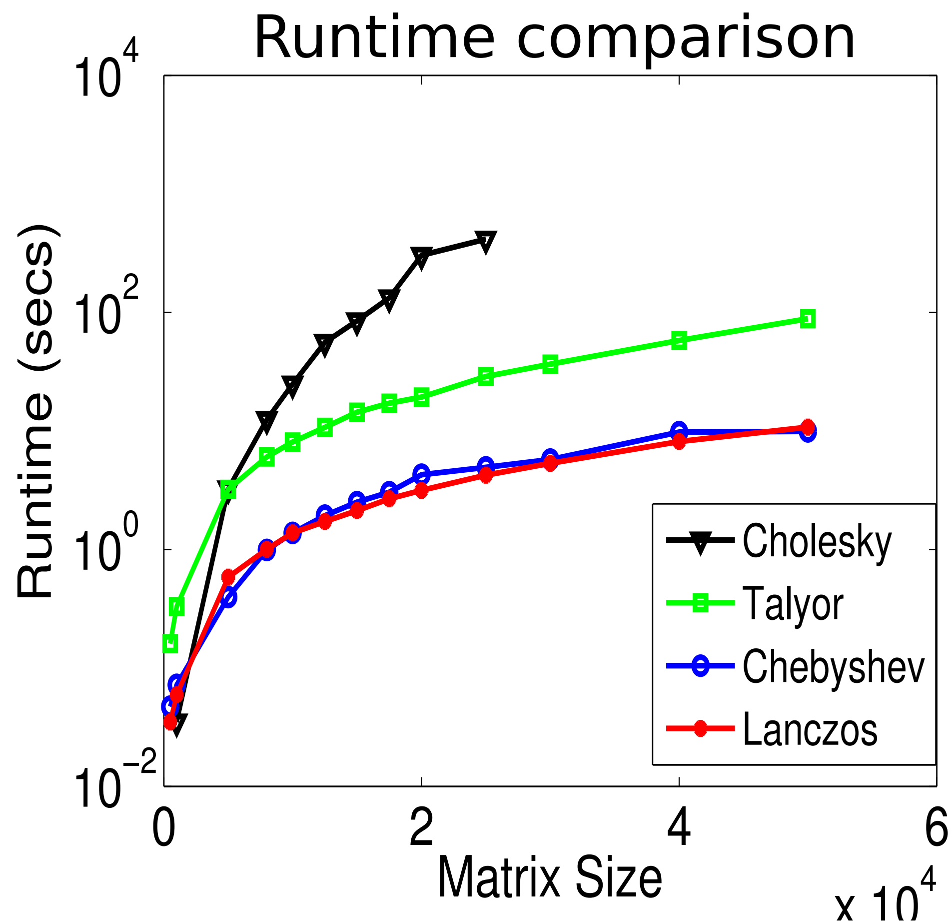
Numerical example: A graph Laplacian `california` of size 9664×9664, $nz \approx 10^5$ from the Univ. of Florida collection.

Rel. error vs degree

- 3 methods: Taylor Series, Chebyshev expansion, SLQ
- # starting vectors $nv = 100$ in all three cases.



Runtime comparisons



Application 6: Log-likelihood.

Comes from parameter estimation for Gaussian processes

➤ Objective is to maximize the log-likelihood function with respect to a 'hyperparameter' vector $\boldsymbol{\xi}$

$$\log p(\boldsymbol{z} \mid \boldsymbol{\xi}) = -\frac{1}{2}\left[\boldsymbol{z}^\top \boldsymbol{S}(\boldsymbol{\xi})^{-1}\boldsymbol{z} + \log\ \det \boldsymbol{S}(\boldsymbol{\xi}) + \text{cst}\right]$$

where $\boldsymbol{z}$ = data vector and $\boldsymbol{S}(\boldsymbol{\xi})$ == covariance matrix parameterized by $\boldsymbol{\xi}$

➤ Can use the same Lanczos runs to estimate $\boldsymbol{z}^\top \boldsymbol{S}(\boldsymbol{\xi})^{-1}\boldsymbol{z}$ and logDet term simultaneously.

Application 7: calculating nuclear norm

- $\|X\|_* = \sum \sigma_i(X) = \sum \sqrt{\lambda_i(X^T X)}$
- Generalization: Schatten p-norms

$$\|X\|_{*,p} = \left[\sum \sigma_i(X)^p\right]^{1/p}$$

- See:

J. Chen, S. Ubaru, YS, "Fast estimation of log-determinant and Schatten norms via stochastic Lanczos quadrature", (Submitted).

Conclusion

➤ Estimating traces & Spectral densities are key ingredients in many algorithms

➤ Physics, machine learning, matrix algorithms, ..

➤ .. many new problems related to ‘data analysis’ and ’statistics’, and in signal processing,

➤ A good instance of a method from physics finding its way in numerical linear algebra

Q: Can we do better than standard random sampling?