UNIVERSITY OF MINNESOTA TWIN CITIES

Applications of trace estimation techniques

Yousef Saad

University of Minnesota

ILAS 2023- Madrid, Spain

June 12–17, 2023

Introduction

- Focus of this talk: 1) Trace estimation techniques 2) and their applications
- Problem: Estimate the trace of a matrix that is not explicitly available.
- *Many* Applications from physics to data-science

Outline:

1. general introduction, *2.* trace estimation, *3.* the DOS, *4.* how to compute it, *5.* how to use it (applications)

Introduction: A few examples

Problem 1: Compute Tr [f (A)], f a certain function

➤ Many applications in Physics, e.g., estimations of Tr (f(A)) extensively used by quantum chemists to approximate Density of States, see

[H. Röder, R. N. Silver, D. A. Drabold, J. J. Dong, Phys. Rev. B. 55, 15382 (1997)]. Will be covered in detail later

Problem 2: Compute Tr[inv[A]] the trace of the inverse.

➤ Arises in cross validation methods [Stats]

➤ Motivation for the work [Golub & Meurant, "Matrices, Moments, and Quadrature", 1993, Book with same title in 2009]

Problem 3: Compute diag[inv(A)] the diagonal of the inverse

- Dynamic Mean Field Theory [DMFT]. Related approach: Non Equilibrium Green's Function (NEGF) approach used to model nanoscale transistors.

- Uncertainty quantification: diagonal of the inverse of a covariance matrix needed [Bekas, Curioni, Fedulova '09]

Problem 4: Compute diag[f (A)] ; f = a certain function.

- Arises in density matrix approaches in quantum modeling

$$f(\epsilon) = \frac{1}{1 + \exp(\frac{\epsilon-\mu}{k_B T})}$$

Here, f = Fermi-Dirac operator

Note: when $T \to 0$ then $f \to$ a step function.

- Linear-Scaling methods

Problem 5: Estimate the numerical rank.

➤ Amounts to counting the number of singular values above a certain threshold τ == Trace $(\phi_\tau(A^T A))$..

$\phi_\tau(t)$ is a certain step function.

Problem 6: Estimate the log-determinant (common in statistics)

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

.... many others

Important tool: Stochastic Trace Estimator

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

Notation:

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

Result:

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

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

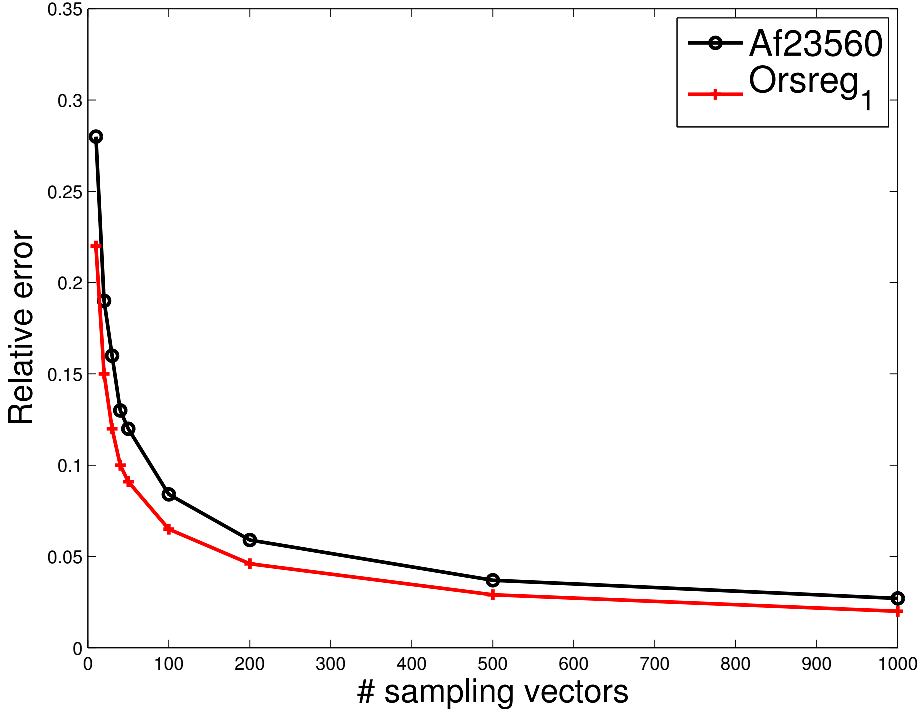
$$\text{Trace}(B) \approx \frac{1}{s} \sum_{j=1}^{s} v_j^T B v_j$$

➤ 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



Alternative: standard probing

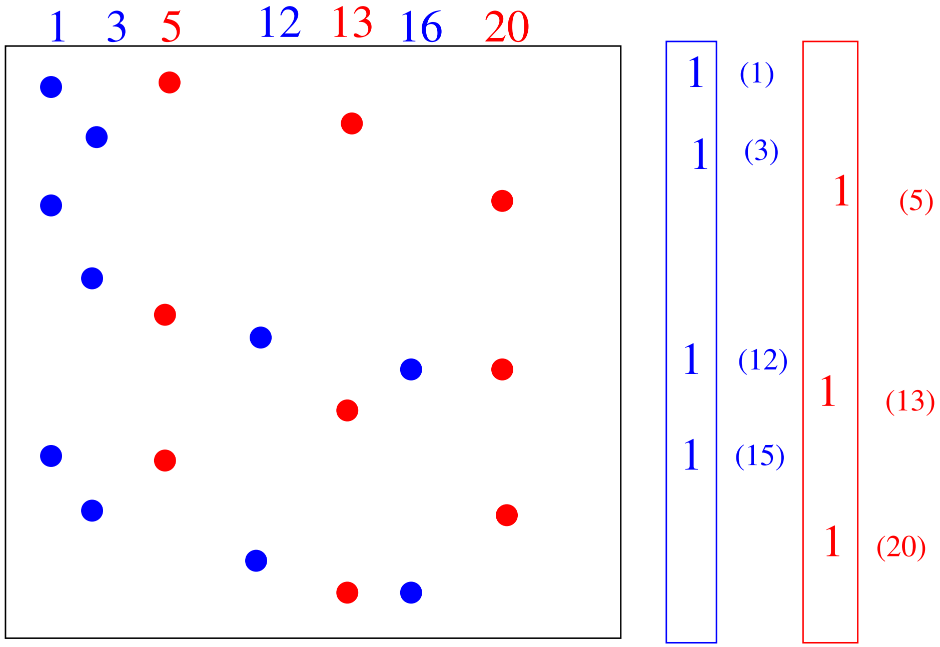
➤ 'Probing' also called "CPR", "Sparse Jacobian estimators",..

Idea: Color columns so that no two columns of the same color overlap.

Entries of same color can be computed with 1 matvec

➤ Corresponds to coloring graph of A^TA.

➤ For problem of diag(A) need only color graph of A



In summary:

- Probing much more powerful when $f(A)$ is known to be nearly sparse (e.g. banded)..
- Approximate pattern (graph) can be obtained inexpensively
- Generally just a handful of probing vectors needed – Can be obtained by coloring graph
- However:
- Not as general: need $f(A)$ to be ‘ ϵ – sparse ’

References:

- J. M. Tang and YS, *A probing method for computing the diagonal of a matrix inverse*, Numer. Lin. Alg. Appl., 19 (2012), pp. 485–501.

See also (improvements)

- Andreas Stathopoulos, Jesse Laeuchli, and Kostas Orginos *Hierarchical Probing for Estimating the Trace of the Matrix Inverse on Toroidal Lattices* SISC, 2012. [somewhat specific to Lattice QCD]

- E. Aune, D. P. Simpson, J. Eidsvik [Statistics and Computing 2012] combine probing with stochastic estimation. Good improvements reported.

SPECTRAL DENSITIES & APPLICATIONS

Spectral Density, a.k.a, Density of States

- Formally, the spectral density 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

- Known as the Density Of States (DOS) in quantum physics
- Note: number of eigenvalues in an interval $[a, b]$ is

$$\mu_{[a,b]} = \int_a^b \sum_j \delta(t - \lambda_j)\, dt \equiv \int_a^b n\phi(t) dt\ .$$

Issue: How to deal with distributions?

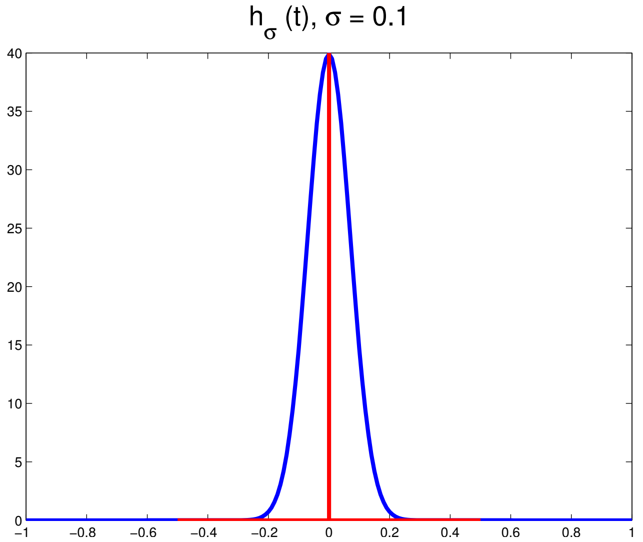
- Highly ‘discontinuous’, not easy to handle numerically

- Solution: replace ϕ by a regularized (‘blurred’) version ϕ_σ:

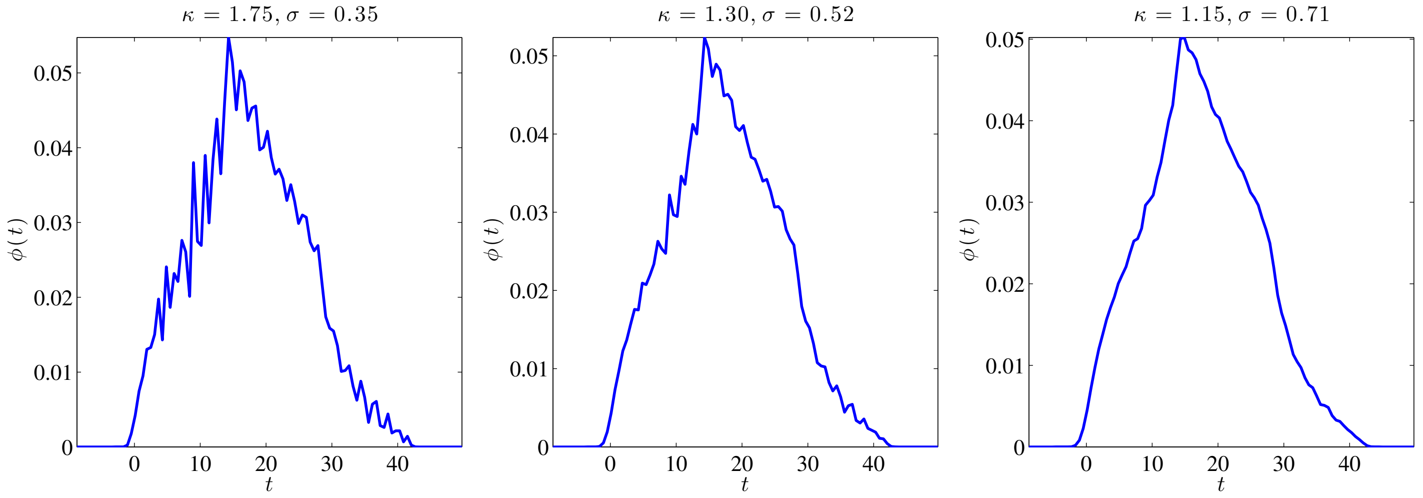
$$\phi_\sigma(t) = \frac{1}{n} \sum_{j=1}^{n} h_\sigma(t - \lambda_j),$$

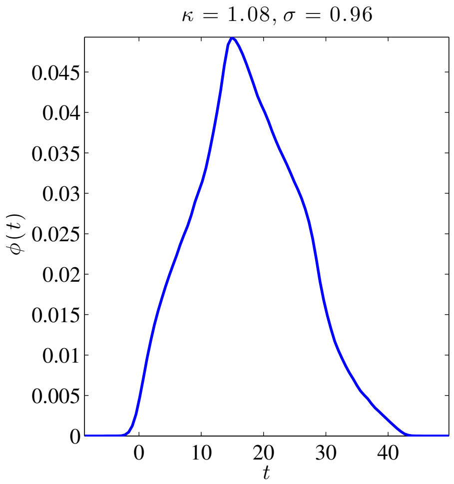
Where, for example: $h_\sigma(t) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{t^2}{2\sigma^2}}$

- Smoothed $\phi(t)$ == distribution function
- Probability of finding eigenvalues of A in infinitesimal $[t - \delta, t + \delta]$
- Useful for theory and in practice.



➤ How to select smoothing parameter σ? Example for Si_2





- ➤ Higher $\sigma \to$ smoother curve
- ➤ But loss of detail ..
- ➤ Compromise: $\sigma = \frac{h}{2\sqrt{2\log(\kappa)}}$,
- ➤ h = resolution, κ = parameter > 1

Computing the DOS: 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

- Use trace estimator [discovered independently] to get traces needed in calculations

- 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:

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

- Here $2-\delta_{k0}$ == 1 when $k=0$ and == 2 otherwise.
- Note: $\boxed{\sum T_k(\lambda_i) = \textit{Trace}[T_k(A)]}$
- Estimate this, e.g., via stochastic estimator
- Generate random vectors $v^{(1)}, v^{(2)}, \cdots, v^{(n_{\text{vec}})}$
- Assume normal distribution with zero mean

- Each vector is normalized so that $\|v^{(l)}\| = 1, l = 1, \ldots, n_{\rm vec}$.
- Estimate the trace of $T_k(A)$ with stochastisc estimator:

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

- Will lead to the desired estimate:

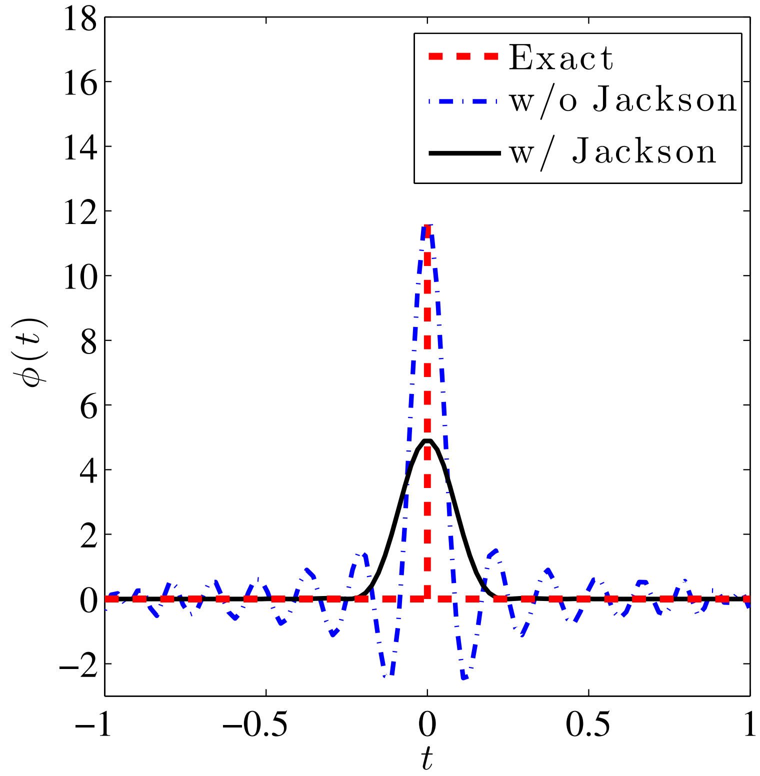
$$\boxed{\mu_k \approx \frac{2 - \delta_{k0}}{n\pi n_{\rm vec}} \sum_{l=1}^{n_{\rm vec}} \left(v^{(l)}\right)^T T_k(A) v^{(l)}.}$$

- To compute $v^T T_k(A) v$, exploit 3-term recurrence of Cheb. polynomials:

$$T_{k+1}(t) = 2tT_k(t)t - T_{k-1}(t) \rightarrow \boxed{v_{k+1} = 2Av_k - v_{k-1}}$$

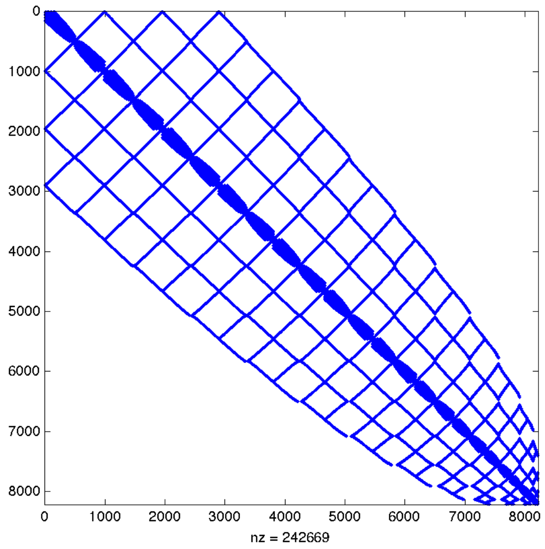
with $v_k \equiv T_k(A)v$,

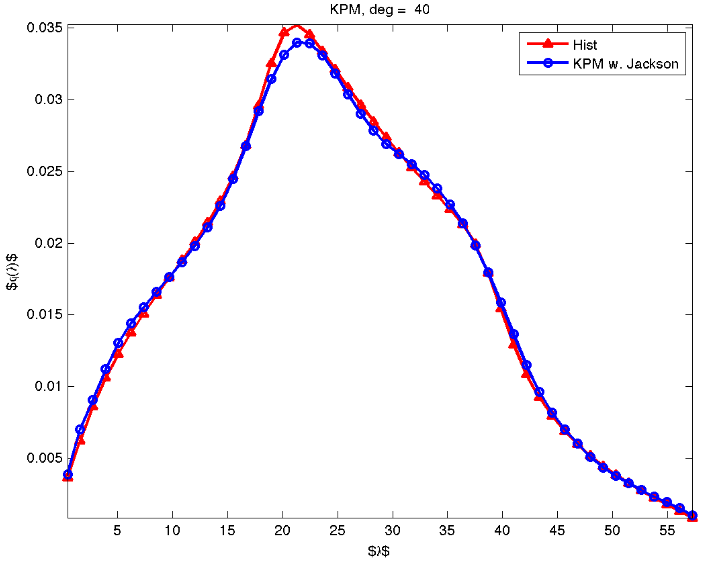
➤ Jackson smoothing can be used –



An example: The Benzene matrix

```
>> TestKpmDos
 Matrix Benzene n =8219  nnz = 242669
Degree =   40   # sample vectors =  10
Elapsed time is 0.235189 seconds.
```





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:

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

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

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

- Lanczos process builds orthogonal polynomials wrt to dot product:

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

- Let $\theta_i, \ i = 1 \cdots , m$ be the eigenvalues of T_m [Ritz values]

- y_i's associated eigenvectors; Ritz vectors: $\{V_m y_i\}_{i=1:m}$

- Ritz values approximate eigenvalues

- Could compute θ_i's then get approximate DOS from these

- Problem: θ_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 $\langle \phi_v, p \rangle \approx \sum a_i p(\theta_i) = \sum a_i \langle \delta_{\theta_i}, p \rangle \to$

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

- 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]

APPLICATIONS

Application 1: Eigenvalue counts

Problem: Estimate $\mu_{[a,b]} \equiv$ number of eigenvalues of A in $[a,\ b]$.

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

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

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

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"

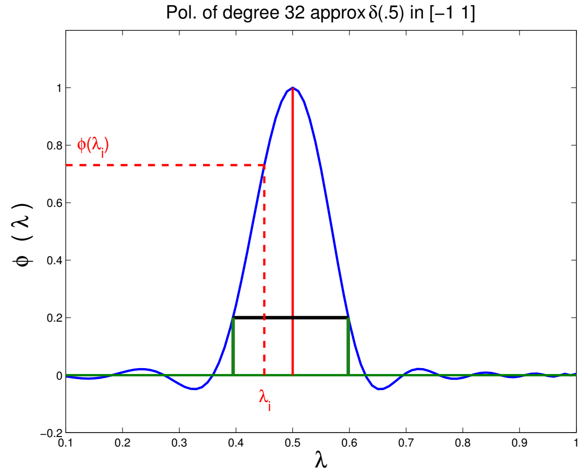
- Situation: very large number of eigenvalues to be computed

- Goal: compute spectrum by slices by applying filtering

- Apply Lanczos or Subspace iteration to problem:

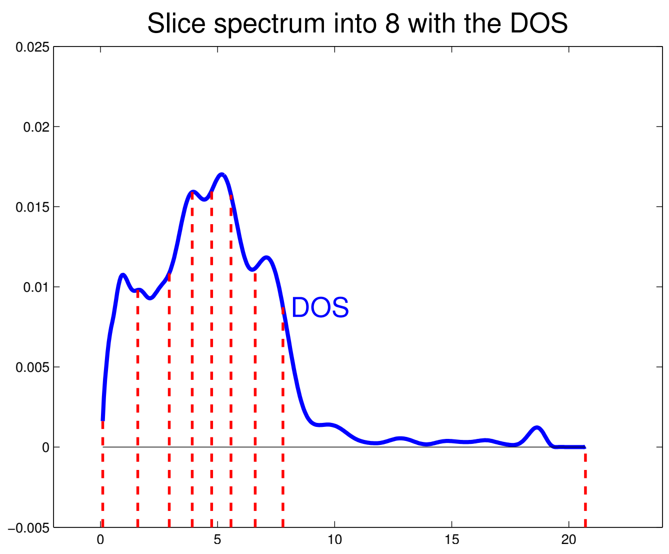
$$\boxed{\phi(A)u = \mu u}$$

$\phi(t) \equiv$ 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:

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

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

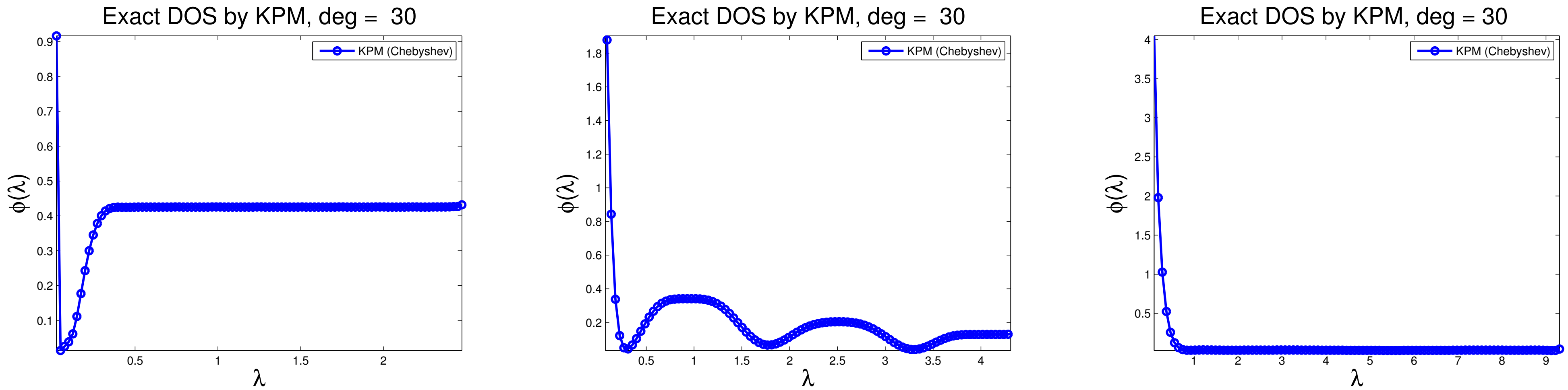
- Two distinct problems:

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.

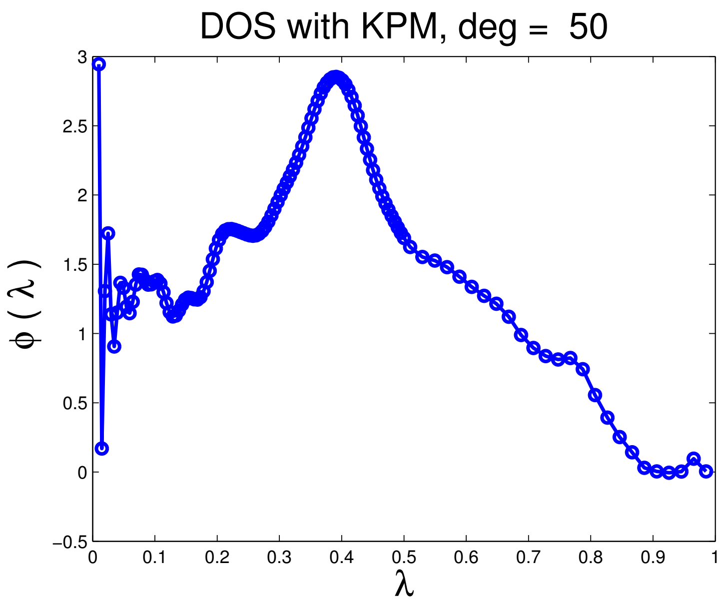
- To find: point immediatly following the initial sharp drop observed.
- Simple idea: use derivative of DOS function ϕ
- For an $n \times n$ matrix with eigenvalues $\lambda_n \le \lambda_{n-1} \le \cdots \le \lambda_1$:

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

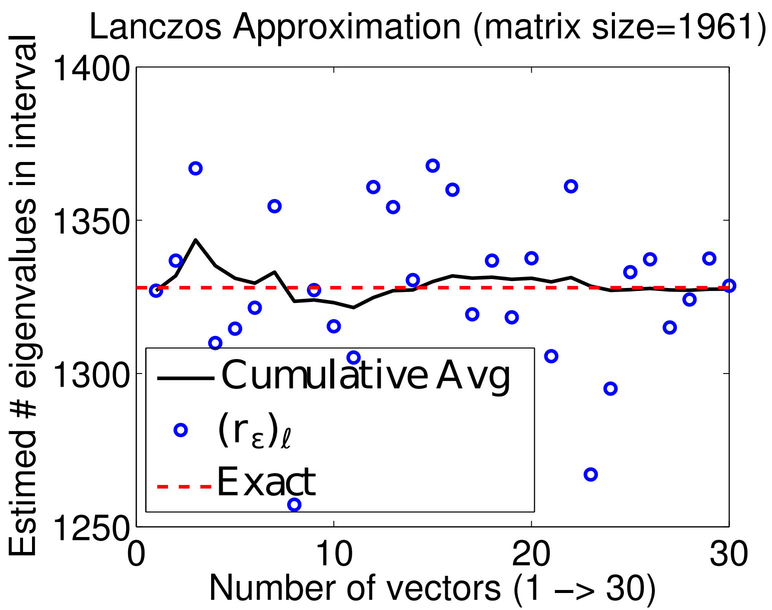
- In practice replace by

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

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	r_ϵ 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

A few other applications

4. Evaluate the Log-determinant of A: (A is SPD)

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

- Equivalent to estimating the trace of $f(A) = \log(A)$

5: Log-likelihood. Used to optimize Gaussian processes

- Objective: maximize the log-likelihood w.r.t. parameter ξ

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

where z = data vector and $S(\xi)$ == covariance matrix

6: Calculating nuclear norm

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

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

- For details on these last 3 applications, see:

S. Ubaru, J. Chen, YS, "Fast estimation of tr(f(A)) via stochastic Lanczos quadrature", SIMAX (2017).

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?