Nyström matrix approximation

an overview and new results

Alex Gittens

Department of Computing and Mathematical Sciences
California Institute of Technology
gittens@caltech.edu

October 11, 2011

Motivation

$\mathbf{A} \in \mathbb{R}^{m \times n}$ is a *huge* matrix. Given $k \ll \min\{m, n\}$, we would like either:

1. a low-rank approximation to $\mathbf{A}$, or
2. approximations of the top k eigenvectors (or singular vectors) of $\mathbf{A}$.

- This abstract problem is ubiquitous in modern data processing: machine learning, image processing, statistical analysis ...
- Traditional approaches (via truncated SVD) cost at least $\mathrm{O}(mnk)$ flops, as do methods based on random projections.
- Column/row-sampling based methods provide faster alternatives, up to $\mathrm{O}(k^3 + (n+m)k^2)$.

Motivation from Eigenvector approximation viewpoint

Nyström extension is a method used in numerical PDEs to approximate eigenfunctions/values of integral operators.

$$\int_0^1 K(x,y) f_\ell(y)\, dy = \lambda_\ell f_\ell(x), \quad x \in [0,1], \ell \in \mathbb{N}.$$

- Discretize the interval and solve the linear algebra problem

$$\frac{1}{m}\sum_{j=1}^{m} K(x_i, x_j) v_\ell(x_j) = \lambda_\ell v_\ell(x_i)$$

- Extend by interpolation to estimate the eigenfunction

$$\hat{f}_\ell(x) = \frac{1}{m}\sum_{j=1}^{m} K(x, x_j)\frac{v_\ell(x_j)}{\lambda_\ell}.$$

In the matrix Nyström method, "discretization" consists of sampling columns. WLOG, take

$$\mathbf{K} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix} \in \mathbb{R}^{n \times n}$$

to be SPSD, and assume $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t$ is invertible.
Extend the eigenvectors of $\mathbf{A}$ to $\mathbf{K}$ by

$$\tilde{\mathbf{U}} = \begin{pmatrix} \mathbf{U} \\ \mathbf{B}^T \mathbf{U} \mathbf{\Lambda}^{-1} \end{pmatrix}.$$

The columns of $\tilde{\mathbf{U}}$ can be orthogonalized in $\mathrm{O}(nm^2)$ time.

Overview

- Traditional methods of nonasymptotic random matrix theory bound the tails of maximum eigenvalues,

$$\lambda_1(\mathbf{A}) = \max_{\|\vec{x}\|=1} \|\mathbf{A}\vec{x}\|.$$

- They are less useful for addressing interior eigenvalues,

$$\lambda_k(\mathbf{A}) = \min_{\substack{\mathbf{V}\in\mathbb{C}^{p\times(p-k+1)} \\ \mathbf{V}^*\mathbf{V}=\mathbf{I}}} \lambda_1(\mathbf{V}^*\mathbf{A}\mathbf{V}) = \max_{\substack{\mathbf{V}\in\mathbb{C}^{p\times k} \\ \mathbf{V}^*\mathbf{V}=\mathbf{I}}} \lambda_k(\mathbf{V}^*\mathbf{A}\mathbf{V}).$$

- The Matrix Laplace Transform, introduced by Ahlswede and Winter (2002), is easily adaptable to address interior eigenvalues, as it is already a variational method.
- We illustrate this by using this technique to bound the number of samples needed to estimate the dominant eigenvalues of a covariance matrix to relative precision.

Problem Statement: Covariance Estimation

Let $\mathbf{x} \in \mathbb{R}^p$ be a zero-mean random vector. Information on the dependency structure of $\mathbf{x}$ is captured by the covariance matrix

$$\mathbf{\Sigma} = \mathbb{E}\mathbf{x}\mathbf{x}^*.$$

The sample covariance matrix is a classical estimator for $\mathbf{\Sigma}$:

$$\widehat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^*,$$

where $\mathbf{x}_i$ are independent samples of $\mathbf{x}$.

How many samples n of $\mathbf{x}$ are required so that $\widehat{\mathbf{\Sigma}}_n$ accurately estimates $\mathbf{\Sigma}$?

Covariance Estimation in Spectral Norm

Accuracy is typically measured in spectral norm.

How many samples n of $\mathbf{x}$ ensure that

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2?$$

- for log-concave distributions $\Omega(p)$ samples suffice (Adamczak et al. 2011),
- for distributions with finite fourth moments, $\tilde{\Omega}(p)$ samples suffice (Vershynin 2011a),
- for distributions with finite $2 + \varepsilon$ moments that satisfy a regularity condition, $\Omega(p)$ samples suffice (Vershynin 2011b),
- for distributions with finite second moments, $\Omega(p \log p)$ samples suffice (Rudelson 1999).

Estimation of dominant eigenvalues of $\mathbf{\Sigma}$

A relative spectral error bound,

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_2 \leq \varepsilon \|\mathbf{\Sigma}\|_2,$$

allows estimation of the top eigenvalue $\lambda_1(\mathbf{\Sigma})$, ...

but does *not* allow estimation of the remaining eigenvalues:

$$|\lambda_k(\mathbf{\Sigma}) - \lambda_k(\widehat{\mathbf{\Sigma}}_n)| < \varepsilon \|\mathbf{\Sigma}\|_2$$

is not a useful estimate if $\lambda_k \ll \lambda_1$.

Established bounds on relative spectral error require that $n = \Omega(\kappa(\mathbf{\Sigma}_\ell)^2 p)$ measurements be taken to ensure relative error recovery of the top ℓ eigenvalues.

... and the unsatisfactory dimensional dependence

In practice, $\mathbf{\Sigma}$ often has a decaying spectrum. What if we want accurate estimates of its dominant eigenvalues?

How many samples n of $\mathbf{x}$ ensure the top $\ell \ll p$ eigenvalues are estimated with relative accuracy,

$$|\lambda_k(\mathbf{\Sigma}) - \lambda_k(\widehat{\mathbf{\Sigma}}_n)| \leq \varepsilon \lambda_k(\mathbf{\Sigma})?$$

Do we really need $\mathrm{O}(p)$ measurements to recover just a few of the top eigenvalues?

Reduced dimensional dependence, assuming decay

Theorem

Consider n independent samples of a $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ distribution. Assume $\{\lambda_k\}$ decays sufficiently fast for $k > \ell$. If $\varepsilon \in (0, 1]$ and

$$n = \Omega(\varepsilon^{-2} \kappa(\mathbf{\Sigma}_\ell)^2 \ell \log p),$$

then with high probability, for each $k = 1, \ldots, \ell$,

$$|\lambda_k(\widehat{\mathbf{\Sigma}}_n) - \lambda_k(\mathbf{\Sigma})| \leq \varepsilon \lambda_k(\mathbf{\Sigma}).$$

Compare to previous estimate of $n = \Omega(\varepsilon^{-2} \kappa(\mathbf{\Sigma}_\ell)^2 p)$:

- Takes advantage of the decay in the residual eigenvalues.
- Requires $\mathrm{O}\left(\frac{p}{\ell \log p}\right)$ fewer samples

Decay condition

"Sufficient" decay means

$$\sum_{k>\ell} \lambda_k/\lambda_1 \leq C.$$

This condition is satisfied if, e.g., the tail eigenvalues ($k > \ell$)

- decay like $k^{-(1+\delta)}$ for some $\delta > 0$, or
- correspond to low-power white noise.

Other decay assumptions may be used.

Estimation of individual eigenvalues

We control, for each k, the probabilities that $\hat{\lambda}_k$ upper/lower-bounds λ_k.

- $\mathbb{P}\left\{\frac{\hat{\lambda}_k}{1-\epsilon} > \lambda_k\right\} > 1 - p^{-\beta}$ when

$$n \geq \frac{1}{32\varepsilon^2}\,\kappa(\mathbf{\Sigma}_k)\frac{\sum_{i\leq k}\lambda_i}{\lambda_k}(\log k + \beta \log p)$$

- $\mathbb{P}\left\{\frac{\hat{\lambda}_k}{1+\varepsilon} < \lambda_k\right\} > 1 - p^{-\beta}$ when

$$n \geq \frac{1}{32\varepsilon^2}\,\frac{\sum_{i\geq k}\lambda_i}{\lambda_k}(\log(p-k+1) + \beta \log p)$$

- Assuming decay (necessary only for the lower bounds), the number of samples needed:

	upper bound	lower bound
λ_1	$\mathrm{O}(\log p)$	$\mathrm{O}(\ell \log p)$
λ_ℓ	$\mathrm{O}(\kappa^2(\mathbf{\Sigma}_\ell)\ell \log p)$	$\mathrm{O}(\kappa(\mathbf{\Sigma}_\ell) \log p)$

Proof sketch

It suffices to show

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \quad \text{and} \quad \mathbb{P}\left\{\hat{\lambda}_k \leq (1-\varepsilon)\lambda_k\right\}$$

decay like $\mathrm{C}\exp(-\mathrm{c}n\epsilon^2)$ when ϵ is sufficiently small.

1. Reduce the probability of each case occuring to the probability that the norm of an appropriate matrix is large.
2. Use matrix Bernstein bounds to estimate the decay of these norms.
3. Take a union bound over the indices k.

Reduction for $\hat{\lambda}_k \geq \lambda_k + t$

Let $\mathbf{B}$ have orthonormal columns and span the bottom $(p-k+1)$-dimensional invariant subspace of $\mathbf{\Sigma}$.

Claim

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq \mathbb{P}\left\{\lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \geq \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) + t\right\}.$$

Proof.
By Courant–Fischer,

$$\lambda_k(\mathbf{\Sigma}) = \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B})$$

and

$$\lambda_k(\widehat{\mathbf{\Sigma}}_n) = \min_{\substack{\mathbf{V}\in\mathbb{C}^{p\times(p-k+1)} \\ \mathbf{V}^*\mathbf{V}=\mathbf{I}}} \lambda_1(\mathbf{V}^*\widehat{\mathbf{\Sigma}}_n\mathbf{V}) \leq \lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}).$$

□

Using the reduction

Need to control RHS of

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq \mathbb{P}\left\{\lambda_1(\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B}) \geq \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) + t\right\}$$

Note $\mathbf{B}^*\widehat{\mathbf{\Sigma}}_n\mathbf{B} = \sum_i \mathbf{B}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{B}$ is a sum of independent rank-one Wishart matrices — naturally suggests use of matrix analogues of classical probability inequalities.
Use estimates of the matrix moments of the summands:

$$\mathbb{E}(\mathbf{B}^*\mathbf{x}\mathbf{x}^*\mathbf{B})^m \preceq 2^m m! \left(\sum_{i \geq k} \lambda_i\right)^{m-1} \cdot \mathbf{B}^*\mathbf{\Sigma}\mathbf{B}.$$

Matrix Bernstein Inequality

We use the following moment-based matrix analog of Bernstein's inequality.

Theorem (Matrix Moment-Bernstein Inequality)

Suppose the d-dimensional self-adjoint matrices $\{\mathbf{G}_i\}$ *are i.i.d. copies of* $\mathbf{G}$ *and*

$$\mathbb{E}(\mathbf{G}^m) \preceq \frac{m!}{2} A^{m-2} \cdot \mathbf{C}^2 \quad \text{for } m = 2, 3, 4, \ldots.$$

Set

$$\mu = n\lambda_1(\mathbb{E}\mathbf{G}) \quad \text{and} \quad \sigma^2 = n\lambda_1(\mathbf{C}^2).$$

Then, for any $t \geq 0$,

$$\mathbb{P}\left\{\lambda_1\Big(\sum\nolimits_i \mathbf{G}_i\Big) \geq \mu + t\right\} \leq d \cdot \exp\Big(-\frac{t^2/2}{\sigma^2 + At}\Big).$$

Finishing the argument

For the summands $\mathbf{B}^*\mathbf{x}_i\mathbf{x}_i^*\mathbf{B}$, we have

$$A = \sum_{i \geq k} \lambda_i \quad \text{and} \quad \sigma^2 = n\Big(\sum_{i \geq k} \lambda_i\Big) \cdot \lambda_1(\mathbf{B}^*\mathbf{\Sigma}\mathbf{B}) = n\lambda_k \cdot \Big(\sum_{i \geq k} \lambda_i\Big).$$

Thus, the Bernstein inequality gives

$$\mathbb{P}\left\{\hat{\lambda}_k \geq \lambda_k + t\right\} \leq (p-k+1) \cdot \exp\left(\frac{-nt^2}{32\lambda_k \sum_{i \geq k} \lambda_i}\right) \qquad \text{for } t \leq 4n\lambda_k.$$

Finally, take $t = \varepsilon\lambda_k$ to see

$$\mathbb{P}\left\{\hat{\lambda}_k \geq (1+\varepsilon)\lambda_k\right\} \leq (p-k+1) \cdot \exp\left(\frac{-n\varepsilon^2}{32 \sum_{i \geq k} \frac{\lambda_i}{\lambda_k}}\right) \qquad \text{for } \varepsilon \leq 4n.$$

The proof for the case $\hat{\lambda}_k \leq \lambda_k - t$ is similar. □

Details

Paper "*Tail Bounds for All Eigenvalues of A Sum of Random Matrices*", Gittens and Tropp, 2011. Preprint, arXiv:1104.4513.

- Elaboration on the relative-error covariance eigenvalue estimation results.
- Similar arguments to find tail bounds for all eigenvalues of a sum of *arbitrary* random matrices.
- An application to column subsampling.

Contact Alex Gittens
gittens@caltech.edu
http://users.cms.caltech.edu/~gittens