Low rank approximation of positive semi-definite symmetric matrices using Gaussian elimination and volume sampling

M. Hegland[1] F. de Hoog[2]

(Received 1 December 2020; revised 25 July 2021)

Abstract

Positive semi-definite matrices commonly occur as normal matrices of least squares problems in statistics or as kernel matrices in machine learning and approximation theory. They are typically large and dense. Thus algorithms to solve systems with such a matrix can be very costly. A core idea to reduce computational complexity is to approximate the matrix by one with a low rank. The optimal and well understood choice is based on the eigenvalue decomposition of the matrix. Unfortunately, this is computationally very expensive. Cheaper methods are based on Gaussian elimination but they require pivoting. We show how invariant matrix theory provides explicit error formulas for an averaged error based on volume sampling. The formula leads to ratios of elementary

DOI:10.21914/anziamj.v62i0.16036, Published 2021-11-14, as part of the Proceedings of the 19th Biennial Computational Techniques and Applications Conference. ISSN 1445-8810.

symmetric polynomials on the eigenvalues. We discuss several bounds for the expected norm of the approximation error and include examples where this expected error norm can be computed exactly.

Contents

1 Introduction

The eigenvalue decomposition of a *real symmetric semi-positive definite matrix* $M \in \mathbb{R}^{n,n}$ *of rank* r is

$$M = Q\Lambda Q^T, \tag{1}$$

where the factor $Q \in \mathbb{R}^{n,r}$ has r orthonormal columns and $\Lambda \in \mathbb{R}^{r,r}$ is a diagonal matrix with elements $\lambda_1 \geqslant \cdots \geqslant \lambda_r > 0$. Such matrices are common in applications in machine learning and information retrieval among others [5]. The normal matrix $M = X^T X$ which occurs in least squares problems is one instance.

Often M is very large, dense and unstructured. However, if the rank r is small, then the eigenvalue decomposition (1) shows that it is possible to represent M using the $nr + r$ matrix elements of the matrices Q and Λ. If many of the eigenvalues $\lambda_i \geqslant 0$ are close to zero, then one may set these eigenvalues to zero. So given a matrix M, one might compute the eigenvalue decomposition and then remove small eigenvalues and their corresponding eigenvectors from the factors Λ and Q to get a good approximation M_k which is represented by $nk + k$ real numbers. The Eckart–Young–Mirsky theorem [1] states that the approximation M_k is optimal. Such an approximation leads to improved computational performance when used in algorithms requiring matrix vector products, leads to higher stability and is useful in data analysis (see principal component analysis). However, the computational cost of the eigenvalue decomposition is typically of order $O(n^3)$ and requires the storage of n^2 numbers. Thus for large n, this approach is often not feasible in practice.

Thus there is a real need to have a faster algorithm which obtains a close to optimal approximation of M. Here we consider a popular example of such an algorithm which is often referred to CUR or the pseudo-skeleton approximation [4]. This algorithm selects k columns (or rows) of the symmetric matrix M and uses them to approximate M. More specifically

$$M = P^T \begin{bmatrix} A & B^T \\ B & C \end{bmatrix} P\,, \tag{2}$$

where P permutes the columns such that the selected ones are moved to the left and $A \in \mathbb{R}^{k,k}$. The CUR approximation is then

$$\widehat{M}_k = P^T \begin{bmatrix} A & B^T \\ B & BA^{-1}B^T \end{bmatrix} P\,. \tag{3}$$

One can show that this approximation is of rank k and in the following sections we investigate the error of this approximation.

For the CUR approximation (3) to be defined, A needs to be invertible. One approach is to select the k columns for which the determinant of A is maximal. If the matrix M has a rank $r \geqslant k$ then this choice of A guarantees that A is invertible. Here we consider an approach which selects the k columns at random with probability proportional to the determinant of A. This choice of A has been termed *volume sampling* [3]. In this case the probability of selecting k columns which lead to a non-invertible A is zero. For this method one can get an exact expression for the expectation of a suitable norm of the error. This is similar to the optimal case. However, in contrast to the optimal case, the determination of the error is often not computationally feasible even if all the eigenvalues are known. Here we study this further and present upper bounds for these expected errors. Derezinski and Mahoney [2] review volume sampling algorithms (also called determinantal point processes).

One suitable norm for the error analysis is the nuclear norm (also called trace norm or Schatten 1 norm). The nuclear norm of M is the sum of its singular values. In the case of semi-definite symmetric matrices M the singular values are equal to the eigenvalues and the nuclear norm is

$$\|M\|_* = \sum_{i=1}^{n} \lambda_i\,. \tag{4}$$

As we assume that the eigenvalues are numbered in decreasing order ($\lambda_{i+1} \leqslant \lambda_i$), the nuclear norm of the error of the optimal approximation is

$$\|M_k - M\|_* = \sum_{i=k+1}^{n} \lambda_i\,. \tag{5}$$

In Section 2 the volume sampling CUR approach is discussed and a formula for the expected error in terms of matrix invariants is established. The errors as functions of the eigenvalues are further discussed in Section 3 and two special types of matrices are considered in more depth.

2 Rank approximation and expected error in terms of matrix invariants

In order to establish the framework for volume sampling we introduce the sample space to be the symmetric group $\Omega = S_n$. Then any sample $\omega \in S_n$ is a permutation of a set with n elements. The symmetric group is the structure of the set of permutation matrices of arrays with n elements and we denote the permutation matrix defined by some $\omega \in \Omega$ as P_ω. The function defined by $M(\omega) = P_\omega M P_\omega^T$ for $\omega \in \Omega$ is then a matrix valued random variable. We denote the blocks defined in equation (2) of $M(\omega)$ by $A(\omega)$, $B(\omega)$ and $C(\omega)$, and for simplicity denote the corresponding variables by M, A, B and C. Finally, we define the probability of some $\omega \in \Omega$ to be

$$p_k(\omega) := \frac{\det A(\omega)}{\sum_{\xi\in\Omega} \det A(\xi)}\,. \tag{6}$$

With this framework we now define the expected error of the CUR approximation to be

$$\mathcal{E}\left(\|\widehat{M}_k - M\|_*\right) := \sum_{p_k(\omega)\neq 0} p_k(\omega)\,\|C(\omega) - B(\omega)A(\omega)^{-1}B(\omega)^T\|_*\,. \tag{7}$$

That is, the expectation is a sum over permutations ω with nonzero probability. The determinant $\det A(\gamma)$ is a kth order principal minor of the unpermuted matrix M. One can see that each principal minor of M occurs $k!(n-k)!$ times when cycling through all the elements of $\gamma \in \Omega$.

Theorem 1 (expected error of volume sampling CUR)**.** *Let M be a symmetric positive semidefinite matrix and $\widehat{M}_k$ be the rank k CUR approximation defined by (3). Using the probability distribution $p_k(\omega)$ (6) and random variable $A(\omega)$ defined above, the expectation of the nuclear norm of the approximation error is*

$$\mathcal{E}\left(\|\widehat{M}_k - M\|_*\right) = (k+1)\frac{c_{k+1}(M)}{c_k(M)}\,, \tag{8}$$

where $c_j(M)$ is the sum of the jth principal minors of M.

In the proof we use the following Lemma.

Lemma 2. *If*

$$M = \begin{bmatrix} A & b \\ b^T & \gamma \end{bmatrix}$$

is positive semidefinite and $\det A = 0$ then

$$\det \begin{bmatrix} A & b \\ b^T & \gamma \end{bmatrix} = 0\,.$$

Proof: As $\det A = 0$ there exists an $x \neq 0$ such that $Ax = 0$. From the semi positive-definiteness one then gets for all $\eta \in \mathbb{R}$

$$\begin{bmatrix} x^T & \eta \end{bmatrix} M \begin{bmatrix} x \\ \eta \end{bmatrix} = 2\eta x^T b + \eta^2 \gamma \geqslant 0\,,$$

and thus $b^T x = 0$. Thus the vector $\begin{bmatrix} x^T 0 \end{bmatrix}$ is in the null space of M. ♠

Proof of Theorem 1: As each minor occurs $k!(n-k)!$ times in the sequence $A(\Omega)$ one has

$$\sum_{\omega \in \Omega} \det A(\omega) = k!(n-k)! c_k(M)\,,$$

and consequently

$$p_k(\omega) = \frac{\det A(\omega)}{k!(n-k)! c_k(M)}\,. \tag{9}$$

As $C - BA^{-1}B^T$ is positive semi-definite almost everywhere (with respect to the probability p_k) one gets

$$\begin{aligned} \|BA^{-1}B^T - C\|_* &= \operatorname{trace}\left(C - BA^{-1}B^T\right) = \sum_{i=1}^{n-k} c_{i,i} - b_i^T A^{-1} b_i \\ &= \frac{\sum_{i=1}^{n-k} \det \begin{bmatrix} A & b_i \\ b_i^T & c_{i,i} \end{bmatrix}}{\det A}\,, \end{aligned}$$

almost everywhere, where b_i^T is the ith row of B. Multiplying both sides with $\det A$ then summing over all permutations and applying Lemma 2 gives

$$\sum_{p(\omega)\neq 0} (\det A(\omega)) \|B(\omega)A(\omega)^{-1}B(\omega)^T - C(\omega)\|_* = \sum_{\omega \in \Omega} \sum_{i=1}^{n-k} \det \begin{bmatrix} A & b_i \\ b_i^T & c_{i,i} \end{bmatrix}.$$

Interchanging the order of the double sum gives $(k+1)!(n-k)!\, c_{k+1}(M)$ and inserting p_k from equation (9) completes the proof. ♠

3 The case of diagonal matrices or error and stability bounds in terms of the eigenvalues

The matrix invariants $c_k(M)$ are defined as the coefficients of the characteristic polynomial of the matrix M (up to signs). Consequently, they are invariant under orthogonal similarity transforms of M and thus

$$c_k(M) = c_k(\Lambda) = e_k(\lambda_1, \ldots, \lambda_n)\,,$$

where e_k are the *elementary symmetric polynomials*. The evaluation of elementary symmetric polynomials is computationally very demanding in general. We will thus focus on upper bounds. A simple bound is given in Proposition 3. This bound is tight and gives a good indication of what the error is for spectra which are rapidly decreasing. However, for spectra like $\lambda_i = 1/i^2$, for $i = 1, 2, \ldots$, this bound is highly overestimating the error. In this section we discuss ways to get better bounds for these cases, but first we demonstrate the performance of these methods with a computational experiment.

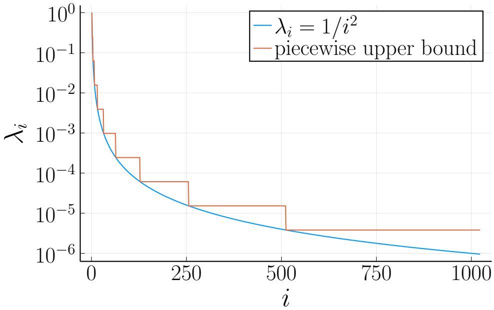
Let $\lambda_i = 1/i^2$ for $i = 1, \ldots, n$ and $n = 2^{l_{\max}}$. We define a piecewise constant array μ by

$$\mu_i = \frac{1}{4^l}\,, \quad \text{where } i = 2^l + j\,,\ j = 1, \ldots, 2^l - 1 \text{ and } l = 0, \ldots, l_{\max} - 1\,.$$

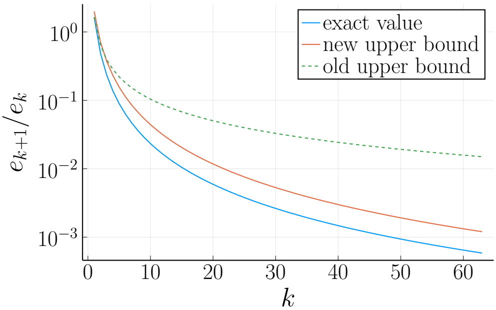
One can show that $\mu_i \geqslant \lambda_i$ for $i = 1, \ldots, n$. From the monotonicity (Lemma 8) one sees that

$$\frac{e_{k+1}(\lambda)}{e_k(\lambda)} \leqslant \frac{e_{k+1}(\mu)}{e_k(\mu)}\,.$$

This is illustrated computationally in Figure 1 where the simple (old) bound from Proposition 3 is also displayed to demonstrate the superiority of the new bound. While not as simple as the simple bound of Proposition 3, the bound we give here is substantially more accurate and useful for error bounds. In contrast to the values for λ, the upper bound can be computed even for very large n and medium sized k, as explained below.



(a) spectrum and pw constant bound



(b) ratio, old and new bounds

Figure 1: Eigenvalues and the ratios e_{k+1}/e_k.

An upper bound for e_{k+1}/e_k which does not require any properties of the eigenvalues is obtained from the following proposition.

Proposition 3 (simple bound). *Let* $n, k \in \mathbb{N}$, $n > k$ *and* $\lambda_i \geqslant 0$ *for all* $i = 1, \dots, n$. *Then*

$$e_{k+1}(\lambda_1, \dots, \lambda_n) \leqslant e_k(\lambda_1, \dots, \lambda_n)\, e_1(\lambda_{k+1}, \dots, \lambda_n)\,.$$

Proof: The proof is by induction where the main step is (the second line uses the induction assumption)

$$\begin{aligned} e_{k+1}(\lambda_1, \dots, \lambda_{n+1}) &= e_k(\lambda_1, \dots, \lambda_n)\, \lambda_{n+1} + e_{k+1}(\lambda_1, \dots, \lambda_n) \\ &\leqslant e_k(\lambda_1, \dots, \lambda_n)\, \lambda_{n+1} + e_k(\lambda_1, \dots, \lambda_n)\, e_1(\lambda_{k+1}, \dots, \lambda_n) \\ &= e_k(\lambda_1, \dots, \lambda_n)\, e_1(\lambda_{k+1}, \dots, \lambda_{n+1})\,. \end{aligned}$$

♠

Thus e_{k+1}/e_k is bounded by the optimal error, and by Theorem 1 the CUR error is bounded by $k+1$ times the optimal error.

In the case where $\lambda_i = q^i$ for some $q \in (0, 1)$ one can compute the elementary symmetric polynomials explicitly, as described by the following proposition.

Proposition 4 (power eigenvalues)**.**

$$e_k(1,q,\dots,q^{n-1}) = q^{k(k-1)/2}\prod_{i=1}^{k}\frac{1-q^{n-i+1}}{1-q^i}\,,\quad \textit{for all } k,n\in\mathbb{N}\,. \tag{10}$$

Proof: Let $e_{k,n} := e_k(1,q,\dots,q^{n-1})$. We use induction over n to show that equation (10) holds for all k.

First, for $n=1$ one has $e_{0,1}=e_{1,1}=1$ and $e_{k,1}=0$ for $k>1$ by definition of e_k. Thus the claimed result holds for $n=1$. We now show that if (10) holds for some n and all k, then it also holds when n is replaced by $n+1$. For this we use the following recursion for symmetric elementary polynomials:

$$e_{k,n+1} = q^n e_{k-1,n} + e_{k,n}\,.$$

Then one verifies that the claimed equation (10) is equivalent to

$$\begin{aligned} e_{k,n} &= \prod_{i=1}^{k}\frac{q^{i-1}-q^n}{1-q^i} \\ &= \frac{q^{k-1}-q^n}{1-q^k}\,e_{k-1,n}\,. \end{aligned} \tag{11}$$

One then gets

$$\begin{aligned} e_{k,n+1} = q^n e_{k-1,n} + e_{k,n} &= \left(q^n + \frac{q^{k-1}-q^n}{1-q^k}\right) e_{k-1,n} \\ &= \frac{q^{k-1}(1-q^{n+1})}{1-q^k}\,e_{k-1,n} = \frac{1-q^{n+1}}{1-q^k}\prod_{i=1}^{k-1}\frac{q^i-q^{n+1}}{1-q^i} \\ &= \prod_{i=1}^{k}\frac{q^{i-1}-q^{n+1}}{1-q^i}\,. \end{aligned}$$

♠

Using the value of e_{k+1}/e_k obtained from equation (11) in Proposition 4 leads to the expected error

$$(k+1)\frac{e_{k+1}}{e_k} = \frac{k+1}{1+q+\cdots+q^k}(q^k+\cdots+q^n)\,.$$

One sees that the CUR method is competitive for small q but not for $q \approx 1$.

We now derive some results which are useful in the derivation and computation of error bounds. First we pad the eigenvalue vectors with zeros so that $\lambda \in \ell_0$, the set of series which are nonzero for finitely many indices.

For any k we define the head of λ to be $\lambda^h = (\lambda_1, \ldots, \lambda_k, 0, \ldots)$ and the tail $\lambda^t = (\lambda_{k+1}, \ldots, \lambda_n, 0, \ldots)$. We introduce a concatenation of two spectra λ and μ:

$$[\lambda, \mu] = (\lambda_1, \ldots, \lambda_n, \mu_1, \ldots, \mu_m, 0, \ldots)\,,$$

which potentially is reordered for size but in the cases considered here we have $\mu_1 \leqslant \lambda_n$. Thus $\lambda = [\lambda^h, \lambda^t]$. The decrease of the tail is modelled by the sequence ρ with

$$\rho_i = \lambda_i^t/\lambda_{k+1}\,, \quad i = 1, 2, \ldots\,. \tag{12}$$

We then introduce a function $f : \mathbb{R}_0^\infty \to \mathbb{R}^{k+2}$ with components

$$f_i(\lambda) = e_i(\lambda)\,.$$

We now introduce the *convolution* of two elements of $\mathbb{R}_+^m$ by

$$(u * v)_i = \sum_{i+j=k} u_i v_j\,.$$

As the e_i are coefficients of a characteristic polynomial one has the following lemma.

Lemma 5 (convolution theorem)**.**

$$f([\lambda, \mu]) = f(\lambda) * f(\mu)\,.$$

The next lemma is a consequence of the fact that f_i is a homogeneous $i-1$ degree polynomial.

Lemma 6 (scaling lemma). *The ith component f_i of f satisfies*

$$f_i(s\lambda) = s^i f_i(\lambda)\,.$$

One then has the following.

Proposition 7. *Let $u^h = f(\lambda^h)$ and $w = f(\rho)$. Then*

$$\frac{e_{k+1}(\lambda)}{e_k(\lambda)} = \gamma\lambda_{k+1}\,,$$

where

$$\gamma = \frac{\sum_{i=0}^{k} \lambda_{k+1}^{k-i}\, w_{k+1-i} u_i^h}{\sum_{i=0}^{k}\, \lambda_{k+1}^{k-i} w_{k-i} u_i^h}\,.$$

Proof: By the convolution theorem Lemma 5 and the definition of ρ in equation (12) one has

$$u = u^t * u^h\,,$$

where $u = f(\lambda)$ and $u^t = f(\lambda^t) = f(\lambda_{k+1}\rho)$. Using the scaling Lemma 6 and $u_{k+1}^h = 0$ one then gets

$$\frac{u_{k+1}}{u_k} = \lambda_{k+1} \frac{\sum_{i=0}^{k} \lambda_{k+1}^{k-i}\, w_{k+1-i} u_i^h}{\sum_{i=0}^{k}\, \lambda_{k+1}^{k-i} w_{k-i} u_i^h}\,.$$

♠

We now show results used to obtain bounds for the case of slowly decreasing ρ_i. In these cases one observes that the sequence w_{i+1}/w_i first increases before it decreases. The main tool to obtain bounds is the monotonicity of the ratios $e_{k+1}(\lambda)/e_k(\lambda)$ in λ.

Lemma 8 (monotonicity)**.** *Let* $0 < \lambda \leqslant \mu$ *component wise then*

$$\frac{e_{k+1}(\lambda)}{e_k(\lambda)} \leqslant \frac{e_{k+1}(\mu)}{e_k(\mu)} \,.$$

This lemma can be proven directly but is also a consequence of a result by Marcus and Lopes [6]:

$$\frac{e_{k+1}(\lambda+\mu)}{e_k(\lambda+\mu)} \geqslant \frac{e_{k+1}(\lambda)}{e_k(\lambda)} + \frac{e_{k+1}(\mu)}{e_k(\mu)} \,.$$

A direct application of the convolution theorem Lemma 5 gives the following representation.

Proposition 9.

$$f(\lambda) = u^{l_{\max}} * \cdots * u^0 \,, \tag{13}$$

where $u^l = f(q^l \epsilon_{2^l})$ *has the components*

$$u_j^l = q^{l(j-1)} \binom{2^k}{j-1}, \quad j = 1, \ldots, 2^l \,.$$

This proposition is used to show that the determination of the ratio e_{k+1}/e_k is computationally feasible for the piecewise spectrum μ used in Figure 1. The complexity of computing the $l_{\max}$ convolutions of size $k+1$ is of order $O(l_{\max}(k+1)^2)$. Using this formula is typically much faster than using the standard recursions which results in a total complexity of $O(n(k+1))$ for very large n. One observes that often only a small number of components of the u^j are substantially different from zero so that the complexity can be further reduced.

4 Conclusion

While the CUR method combined with volume sampling admits an explicit and exact formula for the expected approximation error of the computed rank k approximation there is little known about the performance of the method for slowly decreasing eigenvalues. This article provides new error bounds for the CUR method with volume sampling which shows that this approach is competitive with the optimal approach for slowly decreasing eigenvalues even for large approximation ranks.

Future work may consider the effect of the initial eigenvalues and the rate of decrease of the tail of the spectrum in more detail using the convolution formula provided here.

References

[1] A. Dax. "On extremum properties of orthogonal quotients matrices". In: *Lin. Alg. Appl.* 432.5 (2010), pp. 1234–1257. DOI: 10.1016/j.laa.2009.10.034. (Cit. on p. C59).

[2] M. Dereziński and M. W. Mahoney. *Determinantal Point Processes in Randomized Numerical Linear Algebra*. 2020. arXiv: 2005.03185 [cs.DS] (cit. on p. C61).

[3] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. "Matrix approximation and projective clustering via volume sampling". In: *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*. SODA '06. Miami, Florida: Society for Industrial and Applied Mathematics, 2006, pp. 1117–1126. URL: https://dl.acm.org/doi/10.5555/1109557.1109681 (cit. on p. C60).

[4] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. "A theory of pseudoskeleton approximations". In: *Lin. Alg. Appl.* 261.1 (1997), pp. 1–21. DOI: 10.1016/S0024-3795(96)00301-1. (Cit. on p. C60).

[5] M. W. Mahoney and P. Drineas. "CUR matrix decompositions for improved data analysis". In: *Proc. Nat. Acad. Sci.* 106.3 (Jan. 20, 2009), pp. 697–702. DOI: 10.1073/pnas.0803205106. (Cit. on p. C59).

[6] M. Marcus and L. Lopes. "Inequalities for symmetric functions and Hermitian matrices". In: *Can. J. Math.* 9 (1957), pp. 305–312. DOI: 10.4153/CJM-1957-037-9 (cit. on p. C69).

Author addresses

1. **M. Hegland**, Mathematical Sciences Institute, Australian National University, ACT 2601, Australia.
   mailto:markus.hegland@anu.edu.au
   orcid:0000-0002-5136-2883

2. **F. de Hoog**, Data61, CSIRO, GPO Box 1700, Canberra, ACT 2601, Australia
   mailto:frank.dehoog@csiro.au
   orcid:0000-0002-4632-564X