Frobenius-optimal approximation by structured matrices.

Chris Salahub
University of Waterloo, csalahub@uwaterloo.ca
Jeffrey Uhlmann
University of Missouri, uhlmannj@missouri.edu

October 6, 2022

Abstract

The approximation of a general matrix $\mathbf{M}$ by a structured matrix $\mathbf{T}$ is shown to be optimized in the Frobenius norm by structured means. It is proven that the optimal value of the Frobenius norm is then given by the total standard deviation of entries in $\mathbf{M}$ from the structured means. This approximation is demonstrated for several examples of structured matrices including circulant, Toeplitz, and Hankel matrices, and the consequences of these facts are explored.

1 Introduction

Suppose we would like to approximate the $n \times n$ matrix

$$\mathbf{M} = \begin{bmatrix} m_{00} & m_{01} & \dots & m_{0,n-1} \\ m_{10} & m_{11} & \dots & m_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n-1,0} & m_{n-1,1} & \dots & m_{n-1,n-1} \end{bmatrix} \tag{1}$$

with entries $m_{ij} \in \mathbb{C}$ using a structured matrix $\mathbf{T} \in \mathbb{C}^{n \times n}$ for computational or analytical reasons. Examples include circulant $\mathbf{T}$ for preconditioning [1, 6] and Toeplitz-Hankel $\mathbf{T}$

for physical modelling [4]. For any application, we prefer the approximating matrix $\mathbf{T}$ to be optimal by some measure.

One common measure used to evaluate a matrix approximation is the Frobenius norm. For a matrix $\mathbf{M}$ approximated by $\mathbf{T}$, the Frobenius norm of the difference $\mathbf{T}-\mathbf{M}$ is defined as

$$||\mathbf{T}-\mathbf{M}||_F = \sqrt{\text{trace}\left((\mathbf{T}-\mathbf{M})^*(\mathbf{T}-\mathbf{M})\right)}, \tag{2}$$

where $\mathbf{A}^*$ is the conjugate matrix of $\mathbf{A} \in \mathbb{C}^{n\times n}$. Minimizing this metric was the express goal of the preconditioner derived in [1] and was noted as a positive feature of the approximation of [6]. Both of these approximations use a circulant $\mathbf{T}$, however. This work presents a far more general result which can be applied to *any* structural matrix.

2 Structured matrices

We begin with a formal definition.

Definition 1 (Structured matrix)**.** *Suppose we have a matrix* $\mathbf{T}$ *with entries* t_{ij} *following a regular pattern in* i *and* j*, that is*

$$t_{ij} = t_{f(i,j)} \tag{3}$$

where $f : \{0, 1, \ldots, n-1\}^2 \mapsto \{0, 1, 2, \ldots, K\}$ *is the index function defining the membership of the index pair* i, j *to an index set with a constant value. Then we say that* $\mathbf{T}$ *is* structured*. Additionally, define the* k^{th} *index set*

$$\mathcal{T}_k = \{(i, j)|f(i, j) = k\}$$

with cardinality $|\mathcal{T}_k| = n_k > 0$*.*

In this definition, the index function $f(\cdot, \cdot)$ defines the structure of $\mathbf{T}$ by indicating which elements of $\mathbf{T}$ are equal. Changing f results in a differently structured $\mathbf{T}$. Some common structures and the corresponding functions are shown in Table 1. These functions are not unique; many candidate functions define identical index sets. Hankel matrices, for example, can take either $f(i, j) = j + i$ or $f(i, j) = 2(n - 1) - j - i$.

Noting that a piecewise constant $f(i, j)$ can be defined for arbitrary index sets, it is obvious that the structures which can be defined are not limited to the simple functions in

Structure	$f(i,j)$
Circulant	$(i-j) \bmod n$
Toeplitz	$j-i+n$
Hankel	$i+j$

Table 1: Some common examples of structured index functions.

Table 1. Indeed, the range of potential matrices described by Definition 1 goes from the unstructured case where

$$f(i,j) = in + j$$

to a matrix with one repeated constant value when

$$f(i,j) = 0.$$

Between these exremes any structure can be described with different index functions.

2.1 Optimizing the Frobenius norm

The preliminaries above lead to the following proof.

Theorem 1 (Means minimize the structured approximation of $\mathbf{M}$ in the Frobenius norm.)**.** *The optimal structured matrix* $\mathbf{T}$ *with index function* $f(i,j)$ *and index sets* $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_K$ *to approximate* $\mathbf{M}$ *is given by the matrix* $\mathbf{T}_M$ *with*

$$t_{ij} = t_{f(i,j)} = \overline{m}_{f(i,j)} \tag{4}$$

where

$$\overline{m}_k := \frac{1}{n_k} \sum_{\mathcal{T}_k} m_{ij}. \tag{5}$$

is the mean of entries of $\mathbf{M}$ *over the corresponding index set. Furthermore,* $\frac{1}{\sqrt{n}}||\mathbf{T}_M - \mathbf{M}||_F$ *is the total within-group standard deviation of the values of* $\mathbf{M}$ *over all index sets.*

Proof. Take $\overline{m}_k$ to be the mean of entries in $\mathbf{M}$ for the k^{th} index set as in Equation 5, define the vector of all such means

$$\overline{\mathbf{m}} = (\overline{m}_0, \overline{m}_1, \ldots, \overline{m}_K)^{\mathsf{T}}.$$

Further, denote the vector of unique t_k as

$$\mathbf{t} = (t_0, t_1, \ldots, t_K)^\mathsf{T}$$

and the diagonal matrix of n_k as

$$\mathbf{N} = \mathrm{diag}(n_0, n_1, \ldots, n_K).$$

Noting that Equation 2 is always positive, any $\mathbf{T}$ which minimizes the Frobenius norm will also minimize the squared Frobenius norm of the difference:

$$||\mathbf{T} - \mathbf{M}||_F^2,$$

that is

$$\mathrm{trace}\left((\mathbf{T} - \mathbf{M})^*(\mathbf{T} - \mathbf{M})\right) = \mathrm{trace}\,\mathbf{M}^*\mathbf{M} - \mathrm{trace}\,\mathbf{M}^*\mathbf{T} - \mathrm{trace}\,\mathbf{T}^*\mathbf{M} + \mathrm{trace}\,\mathbf{T}^*\mathbf{T}. \quad (6)$$

$\mathbf{M}^*\mathbf{M}$ is constant in $\mathbf{T}$, so this term can be ignored in the optimization. The latter three terms can be considered individually to express them in terms of the t_k. $\mathrm{trace}\,\mathbf{T}^*\mathbf{T}$ is the simplest, as

$$\mathrm{trace}\,\mathbf{T}^*\mathbf{T} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} t_{ij}^* t_{ij} = \sum_{k=0}^{K} n_k t_k^* t_k. \quad (7)$$

The negative terms can be expressed

$$\mathrm{trace}\,\mathbf{M}^*\mathbf{T} = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} m_{ji}^* t_{ij} = \sum_{k=0}^{K} n_k t_k \overline{m}_k^* \quad (8)$$

and

$$\mathrm{trace}\,\mathbf{T}^*\mathbf{M} = \sum_{k=0}^{K} n_k t_k^* \overline{m}_k. \quad (9)$$

So we seek to minimize

$$F(\mathbf{t}) = \sum_{k=0}^{K} n_k t_k^* t_k - \sum_{k=0}^{K} n_k t_k^* \overline{m}_k - \sum_{k=0}^{K} n_k t_k \overline{m}_k^*,$$

which we can write in matrix form as

$$\begin{aligned} F(\mathbf{t}) &= \mathbf{t}^*\mathbf{N}\mathbf{t} - \mathbf{t}^*\mathbf{N}\overline{\mathbf{m}} - \overline{\mathbf{m}}^*\mathbf{N}\mathbf{t} \\ &= (\mathbf{t} - \overline{\mathbf{m}})^*\mathbf{N}\,(\mathbf{t} - \overline{\mathbf{m}}) - \overline{\mathbf{m}}^*\mathbf{N}\overline{\mathbf{m}}. \end{aligned} \quad (10)$$

As $n_k > 0$ for all $k = 0, 1, \ldots, K$, $\mathbf{N}$ is positive definite, and so the quadratic form $\mathbf{x}^*\mathbf{N}\mathbf{x}$ has a minimum of zero when $\mathbf{x} = \mathbf{0}$. Therefore $F(\mathbf{t})$ is minimized for $\mathbf{t} = \overline{\mathbf{m}}$ and has a minimum of

$$F(\overline{\mathbf{m}}) = -\overline{\mathbf{m}}^*\mathbf{N}\overline{\mathbf{m}} = -\sum_{k=0}^{K} n_k ||\overline{m}_k||^2. \tag{11}$$

So $\mathbf{T}_M$ is the Frobenius-optimal structured matrix $\mathbf{T}$ approximating $\mathbf{M}$. The residual $\mathbf{T}_M - \mathbf{M}$ has a squared Frobenius norm of

$$\begin{aligned} ||\mathbf{T}_M - \mathbf{M}||_F^2 &= \sum_{k=0}^{K} \sum_{\mathcal{T}_k} ||m_{ij}||^2 - \sum_{k=0}^{K} n_k ||\overline{m}_k||^2 \\ &= \sum_{k=0}^{K} n_k \left(\sum_{\mathcal{T}_k} \frac{||m_{ij}||^2}{n_k} - ||\overline{m}_k||^2 \right) \\ &= \sum_{k=0}^{K} n_k \sigma_k^2 \end{aligned} \tag{12}$$

where

$$\sigma_k^2 = \frac{1}{n_k} \sum_{\mathcal{T}_k} (m_{ij} - \overline{m}_k)^2$$

is the variance of the m_{ij} for the index set $\mathcal{T}_k$. Therefore we have

$$\frac{1}{\sqrt{n}} ||\mathbf{T}_M - \mathbf{M}||_F = \sum_{k=0}^{K} \frac{n_k}{n} \sigma_k^2,$$

which is the within-group standard deviation in, for example, typical ANOVA. □

TODO: Any statistical ideas we might incorporate here? We can immediately generalize this to the $L_{1,1}$ norm with the median, for example. Is there any point in trying to develop some ANOVA-on-matrices as a procedure?

TODO: Expand on the utility of this for random matrices of a known structure, for example in genetics [5]

3 Examples

This section shows the application of the proof contained in Section 2.1 to some common structural matrices, starting with an interesting equivalence in the circulant case.

3.1 Circulant matrices

Circulant matrices have an index function

$$f(i,j) = (i-j) \bmod n \tag{13}$$

and so contain n unique values denoted $t_0, t_1, \ldots, t_{n-1} \in \mathbb{C}$. They see widespread use in signal processing, computation, and physical modelling both due to their close relationship with the Fourier transform and their known eigensystem [1, 2, 4]. The general circulant eigenvalues λ_k for $k = 0, 1, \ldots, n-1$ are given by

$$\lambda_k = t_0 + \sum_{l=1}^{n-1} t_l \omega^{lk} \tag{14}$$

and the corresponding k^{th} eigenvector is given by

$$\mathbf{x}_k = \left(1, \omega^k, \omega^{2k}, \ldots, \omega^{(M-1)k}\right)^{\mathsf{T}} \tag{15}$$

where $\omega = \exp(\frac{2\pi i}{n})$ is the complex n^{th} root of unity and $i = \sqrt{-1}$.

Much of the utility of circulant matrices arises from this eigensystem. The $n \times n$ matrix of eigenvectors of $\mathbf{C}$ scaled to be unitary,

$$\mathbf{F} = \frac{1}{\sqrt{n}}[\mathbf{x}_0|\mathbf{x}_1|\mathbf{x}_2|\ldots|\mathbf{x}_{n-1}] = \mathbf{F}^{\mathsf{T}}, \tag{16}$$

is simply the discrete Fourier transform (DFT). Circulant matrices therefore have a deep relationship with real and complex analysis [3]. Importantly, the discrete Fourier transform provides an alternate route to compute the structural means in the circulant case.

Consider the simple approximation algorithm: As $\mathbf{D}$ is diagonal and $\mathbf{F}$ is the matrix of

Algorithm 1 Optimal circulant approximation

Input

$\mathbf{M}$ - an arbitrary $n \times n$ matrix

construct $\mathbf{D} \longleftarrow \text{diag}\,(\mathbf{FMF}^*)$
return $\mathbf{C}_D = \mathbf{F}^*\mathbf{DF}$

eigenvectors for any circulant matrix, $\mathbf{C}_D$ is a circulant matrix with eigenvalues given by

d_{jj}, the diagonal values of $\mathbf{D}$. To determine the elements $(\mathbf{C}_D)_{ij}$ in terms of ω, $\mathbf{x}$, and $\mathbf{M}$, first note

$$d_{jj} = \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\omega^{(k-l)(j-1)}m_{lk}. \tag{17}$$

Analogously, taking $\mathbf{F}^*\mathbf{DF}$ gives an i, j element

$$\begin{aligned}(\mathbf{C}_D)_{ij} &= \mathbf{F}^*\mathbf{DF} \\ &= \frac{1}{n^2}\sum_{l=1}^{n}\sum_{k=1}^{n} m_{lk}\mathbf{x}^*_{k-l}\mathbf{x}_{i-j} \\ &= \frac{1}{n}\sum_{l=1}^{n}\sum_{k=1}^{n} m_{lk} I\big((i-j) \equiv (k-l) \bmod n\big) \end{aligned} \tag{18}$$

where $I(A)$ is the indicator function which returns 1 if A is true and 0 if A is false. Equation 18 indicates that $\mathbf{C}_D$ is generated by replacing the values of M along each circulant diagonal by the corresponding diagonal mean. Therefore $\mathbf{C}_D = \mathbf{T}_M$ when $\mathbf{T}$ is restricted to be circulant. Therefore, $\mathbf{C}_D$ is Frobenius optimal.

Though this fact has aleady been noted for $\mathbf{M}$ Toeplitz by [1] and more generally in [6], it is worth emphasizing here that it is a particular example of the more general result of Theorem 1.

3.2 Toeplitz matrices

Two well-known examples of structural matrices are Toeplitz matrices and Hankel matrices. Toeplitz matrices have an index function

$$f(i, j) = j - i + n$$

3.3 Hankel matrices

Hankel matrices take

$$f(i, j) = i + j$$

3.4 Toeplitz-plus-Hankel matrices

Motivated by the use of multiple circulants to decompose a matrix in [6] and their mention in [7], we might consider the sums of simpler structured matrices. Consider the case of a Toeplitz matrix and a Hankel matrix added together.

Our natural instinct for the index function in this case might just be the sum of previous index functions. That is

$$j - i + n + i + j = 2j + n,$$

but this is clearly incorrect as

$$a_{ij} = t_{j-i+n} + h_{i+j}$$

contains no equalities in i, j. Rather, this describes a linear model of the entries with means removed according to both, as we are taking a difference

$$\mathbf{T} + \mathbf{H} - \mathbf{M}$$

for $\mathbf{T}$ Toeplitz and $\mathbf{H}$ Hankel. TODO: This needs to be expanded more...

References

[1] Tony F Chan. An optimal circulant preconditioner for toeplitz systems. *SIAM Journal on Scientific and Statistical Computing*, 9(4):766–771, 1988.

[2] Robert M Gray. *Toeplitz and circulant matrices: A review*. now Publishers Inc., 2006.

[3] Ulf Grenander and Gabor Szegö. *Toeplitz forms and their applications*. University of California Press, 1958.

[4] Onuttom Narayan and B Sriram Shastry. Generalized toeplitz–hankel matrices and their application to a layered electron gas. *Journal of Physics A: Mathematical and Theoretical*, 54(17):175201, 2021.

[5] Christopher Salahub. A structural model of genome-wide association studies. *arXiv preprint arXiv:2205.10391*, 2022.

[6] Murugesan Venkatapathi et al. Circulant decomposition of a matrix and the eigenvalues of toeplitz type matrices. *arXiv preprint arXiv:2105.14805*, 2022.

[7] Ke Ye and Lek-Heng Lim. Algorithms for structured matrix-vector product of optimal bilinear complexity. In *2016 IEEE Information Theory Workshop (ITW)*, pages 310–314. IEEE, 2016.