Optimizing the Frobenius norm under structural constraints.

Chris Salahub

University of Waterloo, csalahub@uwaterloo.ca

Jeffrey Uhlmann

University of Missouri, uhlmannj@missouri.edu

September 21, 2022

Abstract

TODO: Rewrite this to be more general The approximation of a general matrix $\mathbf{M}$ by a circulant matrix $\mathbf{C}$ is explored. Using the discrete Fourier transform matrix $\mathbf{F}$, the circulant with eigenvalues given by the diagonals of $\mathbf{FMF}^*$ is shown to be equivalent to the nearest circulant in the Frobenius norm, $\mathbf{C}_M$. An intuitive interpretation of this matrix in terms of means and variances of its values is presented.

1 Introduction

TODO: Preamble, justification

Circulant matrices are matrices of the form

$$\mathbf{C} = \begin{bmatrix} c_0 & c_1 & c_2 & \dots & c_{n-2} & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \dots & c_{n-3} & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & \dots & c_{n-4} & c_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_2 & c_3 & c_4 & \dots & c_0 & c_1 \\ c_1 & c_2 & c_3 & \dots & c_{n-1} & c_0 \end{bmatrix} \tag{1}$$

with $c_0, c_1, \ldots, c_{n-1} \in \mathbb{C}$. They see widespread use in signal processing, computation, and physical modelling as matrices with known eigensystems [1, 2]. For $\mathbf{C}$ as in Equation 5, the ordered eigenvalues for $k = 0, 1, \ldots, n-1$ are given by

$$\lambda_k = c_0 + \sum_{l=1}^{n-1} c_l \omega^{lk} \tag{2}$$

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

$$\mathbf{x}_k = \begin{bmatrix} 1 \\ \omega^k \\ \omega^{2k} \\ \vdots \\ \omega^{(M-1)k} \end{bmatrix} \tag{3}$$

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,

$$\mathbf{F} = \frac{1}{\sqrt{n}}[\mathbf{x}_0|\mathbf{x}_1|\mathbf{x}_2|\ldots|\mathbf{x}_{n-1}] = \mathbf{F}^{\mathsf{T}}, \tag{4}$$

is simply the discrete Fourier transform (DFT). Circulant matrices therefore have a deep relationship with real and complex analysis [3].

Suppose we would like to take advantage of this depth of theory and practice to approximate the $n \times n$ matrix

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1n} \\ m_{21} & m_{22} & \ldots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \ldots & m_{nn} \end{bmatrix} \tag{5}$$

by some circulant matrix. This note outlines some results.

2 Approximating M using F

Consider the simple approximation algorithm:

1. construct the diagonal matrix $\mathbf{D}$ where $d_{jj} = (\mathbf{FMF}^*)_{jj}$ and $\mathbf{F}^*$ is the complex conjugate of $\mathbf{F}$,

2. compute $\mathbf{C}_D = \mathbf{F}^*\mathbf{DF}$.

As $\mathbf{D}$ is diagonal and $\mathbf{F}$ is the matrix of eigenvectors for any circulant matrix, $\mathbf{C}_D$ is a circulant matrix with eigenvalues given by d_{jj}. To determine the elements $(\mathbf{C}_D)_{ij}$ in terms of ω, $\mathbf{x}$, and $\mathbf{M}$, first note

$$d_{jj} = \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\omega^{(k-l)(j-1)}m_{lk}. \tag{6}$$

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

$$\begin{aligned}(\mathbf{C}_D)_{ij} &= \mathbf{F}^*\mathbf{DF} \\ &= \frac{1}{n^2}\sum_{l=1}^{n}\sum_{k=1}^{n} m_{lk}\mathbf{x}^*_{k-l}\mathbf{x}_{i-j} \\ &= \frac{1}{n}\sum_{l=1}^{n}\sum_{k=1}^{n} m_{lk}\delta_{(i-j) \bmod n,(k-l) \bmod n} \end{aligned}\tag{7}$$

where δ_{ij} is the Kronecker delta defined by

$$\delta_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ if } i = j. \end{cases} \tag{8}$$

Equation 7 indicates that $\mathbf{C}_D$ is generated by replacing the values of M along each circulant diagonal by the corresponding diagonal mean. Expressed as in Equation 1, $\mathbf{C}_D$ is the circulant matrix with

$$c_k = \frac{1}{n}\sum_{\{i,j|(i-j) \bmod n=k\}} m_{ij} := \overline{m}_k \tag{9}$$

Remarkably, though this was not the original motivation, $\mathbf{C}_D$ is also optimal in the sense of [1]: it minimizes the Frobenius norm.

Theorem 1 ($\mathbf{C}_D$ is Frobenius optimal). *$\mathbf{C}_D$ minimizes $||\mathbf{C} - \mathbf{M}||_F$ for circulant $\mathbf{C}$, where $||\mathbf{A}||_F$ is the Frobenius norm of $\mathbf{A}$.*

Proof. We can write $||\mathbf{C}-\mathbf{M}||_F$ as

$$\sqrt{\operatorname{trace}\left((\mathbf{C}-\mathbf{M})^*(\mathbf{C}-\mathbf{M})\right)}. \tag{10}$$

Any $\mathbf{C}$ which minimizes Equation 10 will also minimize $||\mathbf{C}-\mathbf{M}||_F^2$. Therefore we seek to minimize

$$\operatorname{trace}\left((\mathbf{C}-\mathbf{M})^*(\mathbf{C}-\mathbf{M})\right) = \operatorname{trace}\mathbf{M}^*\mathbf{M} - \operatorname{trace}\mathbf{M}^*\mathbf{C} - \operatorname{trace}\mathbf{C}^*\mathbf{M} + \operatorname{trace}\mathbf{C}^*\mathbf{C}. \tag{11}$$

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

$$\operatorname{trace}\mathbf{C}^*\mathbf{C} = n\sum_{i=0}^{n-1} c_i^* c_i. \tag{12}$$

The negative terms can be expressed

$$\begin{aligned}\operatorname{trace}\mathbf{M}^*\mathbf{C} &= \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ji}^* c_{(i-j) \bmod n} \\ &= \sum_{i=0}^{n-1} c_i \left(\sum_{j=1}^{n-i} m_{j,j+i}^* + \sum_{j=1}^{i} m_{n-i+j,j}^*\right) \\ &= n\sum_{i=0}^{n-1} c_i \overline{m}_i^* \end{aligned} \tag{13}$$

and

$$\operatorname{trace}\mathbf{C}^*\mathbf{M} = n\sum_{i=0}^{n-1} c_i^* \overline{m}_i. \tag{14}$$

So we seek to minimize

$$\begin{aligned} F(\mathbf{c}) &= n\sum_{i=0}^{n-1} c_i^* c_i - n\sum_{i=0}^{n-1} c_i^* \overline{m}_i - n\sum_{i=0}^{n-1} c_i \overline{m}_i^* \\ &= n\left(\langle \mathbf{c},\mathbf{c}\rangle - \langle \overline{\mathbf{m}},\mathbf{c}\rangle - \langle \mathbf{c},\overline{\mathbf{m}}\rangle\right) \end{aligned} \tag{15}$$

where $\langle \mathbf{x},\mathbf{y}\rangle \geq 0$ is the Hermitian inner product of $\mathbf{x},\mathbf{y}\in\mathbb{C}$, $\mathbf{c}=(c_0,c_1,\ldots,c_{n-1})^{\mathsf{T}}$ is the first row of $\mathbf{C}$, and $\overline{\mathbf{m}}=(\overline{m}_0,\overline{m}_1,\ldots,\overline{m}_{n-1})^{\mathsf{T}}$ is the vector of diagonal means of $\mathbf{M}$.

$\langle \overline{\mathbf{m}}, \mathbf{c} \rangle$ and $\langle \mathbf{c}, \overline{\mathbf{m}} \rangle$ are maximized when $\mathbf{c} = t\overline{\mathbf{m}}$ for $t \in \mathbb{R}$ and $\langle \mathbf{c}, \mathbf{c} \rangle$ simply gives the squared magnitude of $\mathbf{c}$. Therefore, the minimizer of Equation 15 must be $\mathbf{c} = t\overline{\mathbf{m}}$ for some $t \in \mathbb{R}$. Substituting this into Equation 15:

$$\begin{aligned} F(t\overline{\mathbf{m}}) &= n\left(t^2\langle \overline{\mathbf{m}}, \overline{\mathbf{m}} \rangle - t\langle \overline{\mathbf{m}}, \overline{\mathbf{m}} \rangle - t\langle \overline{\mathbf{m}}, \overline{\mathbf{m}} \rangle\right) \\ &= n||\overline{\mathbf{m}}||^2(t^2 - 2t), \end{aligned} \tag{16}$$

which has a minimum of $-n||\overline{\mathbf{m}}||^2$ when $t = 1$. Therefore, $\mathbf{c} = \overline{\mathbf{m}}$ minimizes $F(\mathbf{c})$ and so the optimal circulant matrix $\mathbf{C}$ to approximate $\mathbf{M}$ in the Frobenius norm satisfies Equation 9. □

Recognizing that the value of the squared Frobenius norm is

$$||\mathbf{C} - \mathbf{M}||_F^2 = \operatorname{trace} \mathbf{M}^*\mathbf{M} + F(\mathbf{c}) = \sum_{i=1}^{n}\sum_{j=1}^{n} ||m_{ij}||^2 + F(\mathbf{c}),$$

substituting $\mathbf{c} = \overline{\mathbf{m}}$ gives a minimum

$$\begin{aligned} ||\mathbf{C}_D - \mathbf{M}||_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} ||m_{ij}||^2 - n||\overline{\mathbf{m}}||^2 \\ &= n\left(\sum_{k=0}^{n-1} \sum_{\{1\le i,j\le n | (i-j) \bmod n = k\}} \frac{||m_{ij}||^2}{n} - \sum_{k=0}^{n-1} ||\overline{m}_k||^2\right) \\ &= n\sum_{k=0}^{n-1} \sigma_k^2 \end{aligned} \tag{17}$$

where σ_k^2 is the variance of values along the k^{th} diagonal of $\mathbf{M}$. Therefore the value of the Frobenius norm $||\mathbf{C}_D - \mathbf{M}||_F$ is given by the total standard deviation of the values of $\mathbf{M}$ from their respective circulant diagonals!

3 Other structured matrices

The proof provided for Theorem 1 does not apply to circulant matrices alone, as the grouping of terms in the sums is arbitrary. Suppose we have some structured matrix $\mathbf{A}$ with entries a_{ij} which have constant values following a regular pattern in i and j. That is,

$$a_{ij} = a_{f(i,j)} \tag{18}$$

where $f : \{0, 1, \dots, n-1\}^2 \mapsto \{0, 1, 2, \dots, K\}$ indicates the membership of the index pair i, j to a constant index set indexed by k. Define

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

with cardinality $|A_k| = n_k > 0$, then Equations 13 and 14 can be generalized to

$$\begin{aligned}
\operatorname{trace} \mathbf{M}^* \mathbf{A} &= \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ji}^* a_{ij} \\
&= \sum_{k=0}^{K} a_k \sum_{\mathcal{A}_k} m_{ji}^* \qquad (19)
\end{aligned}$$

and

$$\operatorname{trace} \mathbf{A}^* \mathbf{M} = \sum_{k=0}^{K} a_k^* \sum_{\mathcal{A}_k} m_{ji}, \qquad (20)$$

respectively. Define the mean of entries in $\mathbf{M}$ for the k^{th} index set,

$$\overline{m}_k := \frac{1}{n_k} \sum_{\mathcal{A}_k} m_{ij}; \qquad (21)$$

the vector of all such means,

$$\overline{\mathbf{m}} = (\overline{m}_0, \overline{m}_1, \dots, \overline{m}_K)^{\mathsf{T}} \qquad (22)$$

the vector of unique a_k,

$$\mathbf{a} = (a_0, a_1, \dots, a_K)^{\mathsf{T}}; \qquad (23)$$

and the diagonal matrix of n_k,

$$\mathbf{N} = \operatorname{diag}(n_0, n_1, \dots, n_K) \qquad (24)$$

Then Equation 15 becomes

$$\begin{aligned}
F(\mathbf{a}) &= \mathbf{a}^* \mathbf{N} \mathbf{a} - \overline{\mathbf{m}}^* \mathbf{N} \mathbf{a} - \mathbf{a}^* \mathbf{N} \overline{\mathbf{m}} \qquad (25) \\
&= (\mathbf{a} - \overline{\mathbf{m}})^* \mathbf{N} (\mathbf{a} - \overline{\mathbf{m}}) - \overline{\mathbf{m}}^* \mathbf{N} \overline{\mathbf{m}}. \qquad (26)
\end{aligned}$$

As $n_k > 0$ for all $k = 0, 1, \dots, K$, $\mathbf{N}$ 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 $F(\mathbf{a})$ is minimized for $\mathbf{a} = \overline{\mathbf{m}}$ and has a minimum of

$$F(\overline{\mathbf{m}}) = -\overline{\mathbf{m}}^* \mathbf{N} \overline{\mathbf{m}} = -\sum_{k=0}^{K} n_k ||\overline{m}_k||^2. \qquad (27)$$

Letting $\mathbf{A}_M$ be the Frobenius-optimal structural matrix $\mathbf{A}$ to approximate $\mathbf{M}$, this gives

$$\begin{aligned}
||\mathbf{A}_M - \mathbf{M}||_F^2 &= \sum_{k=0}^{K} \sum_{\mathcal{A}_k} ||m_{ij}||^2 - \sum_{k=0}^{K} n_k ||\overline{m}_k||^2 \\
&= \sum_{k=0}^{K} n_k \left(\sum_{\mathcal{A}_k} \frac{||m_{ij}||^2}{n_k} - ||\overline{m}_k||^2 \right) \qquad (28) \\
&= \sum_{k=0}^{K} n_k \sigma_k^2 \qquad (29)
\end{aligned}$$

where

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

is the variance of the m_{ij} for the index set $\mathcal{A}_k$.

Toeplitz matrices are matrices of the form

$$\mathbf{T} = \begin{bmatrix} t_0 & t_1 & t_2 & \dots & t_{n-2} & t_{n-1} \\ t_{-1} & t_0 & t_1 & \dots & t_{n-3} & t_{n-2} \\ t_{-2} & t_{-1} & t_0 & \dots & t_{n-4} & t_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ t_{2-n} & t_{3-n} & t_{4-n} & \dots & t_0 & t_1 \\ t_{1-n} & t_{2-n} & t_{3-n} & \dots & t_{-1} & t_0 \end{bmatrix} \qquad (30)$$

Hankel matrices are matrices of the form

$$\mathbf{H} = \begin{bmatrix} h_0 & h_1 & h_2 & \dots & h_{n-2} & h_{n-1} \\ h_1 & h_2 & h_3 & \dots & h_{n-1} & h_n \\ h_2 & h_3 & h_4 & \dots & h_n & h_{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{n-2} & h_{n-1} & h_n & \dots & h_{2n-4} & h_{2n-3} \\ h_{n-1} & h_n & h_{n+1} & \dots & h_{2n-3} & h_{2n-2} \end{bmatrix} \qquad (31)$$

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.