The Main Diagonal of a Permutation Matrix

Marko Lindner and Gilbert Strang

Abstract

By counting 1's in the "right half" of $2w$ consecutive rows, we locate the main diagonal of any doubly infinite permutation matrix with bandwidth w. Then the matrix can be correctly centered and factored into block-diagonal permutation matrices.

Part II of the paper discusses the same questions for the much larger class of band-dominated matrices. The main diagonal is determined by the Fredholm index of a singly infinite submatrix. Thus the main diagonal is determined "at infinity" in general, but from only $2w$ rows for banded permutations.

Mathematics subject classification (2010): 15A23; 47A53, 47B36.

Keywords and phrases: banded matrix, permutation, infinite matrix, main diagonal, factorization.

1 Introduction

This paper is about banded doubly infinite permutation matrices. These matrices represent permutations of the integers $\mathbb{Z}$, in which no integer moves more than w places. The banded matrix P has zero entries p_{ij} whenever $|i-j|>w$, and it has exactly one entry $p_{ij}=1$ in each row and column. So P has $2w+1$ diagonals that are possibly nonzero, and exactly one of these deserves to be called the *main diagonal*. The first objective of this paper is to find the main diagonal of P.

The remarkable fact is that the correct position of the main diagonal is determined by *any* $2w$ *consecutive rows of* P. This will be Theorem 1. Normally that position can only be determined "at infinity".

For a much larger class of biinfinite matrices A, we form the singly infinite submatrix A_+ from rows $i>0$ and columns $j>0$ of A. The *plus-index* of A is the usual Fredholm index of A_+, computed from the dimension of its nullspace and the codimension of its range:

$$\operatorname{index}_+(A) \;=\; \operatorname{index}(A_+) \;=\; \dim N(A_+) \;-\; \operatorname{codim} R(A_+). \qquad (1)$$

For any Fredholm operator, including all permutations $A=P$, those two numbers are finite. Then the main diagonal of A is determined by $\kappa=\operatorname{index}_+(A)$. It is found κ diagonals above (or $-\kappa$ diagonals below) the zeroth diagonal of A. So our problem is to compute that index.

Example 1.1 The doubly infinite forward shift S has nonzero entries $S_{i,i-1}=1$ for $i\in\mathbb{Z}$. Its singly infinite submatrix S_+ is lower triangular, with those ones along the first subdiagonal. The nullspace of S_+ has dimension $= 0$ (independent columns of S_+) but the range has codimension $= 1$ (it consists of all singly infinite vectors with a zero in the first position). Thus $\kappa = \text{index}_+(S) = 0 - 1$ and the main diagonal of S is correctly located: It is one diagonal below the zeroth diagonal of S (and it contains the ones).

Since S is a permutation of $\mathbb{Z}$ with bandwidth $w = 1$, Theorem 1 says that $\text{index}_+(S)$ can be found from any two consecutive rows i and $i+1$ of S: "Count the ones in columns $i+1$, $i+2$, ... and subtract w." The result $0 - 1$ agrees with $\text{index}_+(S)$. □

We can directly state Theorems 1–3, for banded permutations P of $\mathbb{Z}$. Two proofs are given for Theorem 1. Then Part II of the paper discusses the much larger class of doubly infinite band-dominated Fredholm operators A. The plus-index of A (and thus its main diagonal) is well defined but not so easily computable.

Theorem 1 *If P has bandwidth w, its plus-index is determined by rows $1, \ldots, 2w$ and columns $j > w$. Subtract w from the number n of ones in this submatrix. The result $\kappa = n - w$ is the plus-index of P, and the main diagonal of P is κ diagonals above the zeroth diagonal.*

The same result κ comes from rows $j^* - w$ to $j^* + w - 1$ and columns $\geq j^*$, for any $j^* \in \mathbb{Z}$. The submatrix is shown in (13) below.

Equivalently, $P_c = S^\kappa P$ is a *centered* permutation of $\mathbb{Z}$, by which we mean that its zeroth diagonal is the main diagonal. Our first proof will show how the submatrix with $2w$ rows was discovered. The second proof goes directly to the index κ.

Theorem 2 *Every centered permutation P_c of $\mathbb{Z}$ with bandwidth w can be factored into a product of two block-diagonal permutations:*

$$P_c \;=\; BC \;=\; \begin{pmatrix} \ddots & & & \\ & \boxed{B_0} & & \\ & & \boxed{B_1} & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} \ddots & & & \\ & \boxed{C_0} & & \\ & & \boxed{C_1} & \\ & & & \ddots \end{pmatrix} \tag{2}$$

All blocks have size $2w$ and each C_i is "offset" between B_i and B_{i+1} (shifted by w rows and columns, as shown).

Theorem 3 *The centered permutation P_c can be further factored into $N < 2w$ block-diagonal permutations of bandwidth 1:*

$$P_c \;=\; F_1 \, F_2 \, \cdots \, F_N \tag{3}$$

Each factor F_i has block size 1 or 2. Thus F_i exchanges disjoint pairs of neighboring rows. By Theorems 1–3 the original P is factored into

$$P \;=\; S^{-\kappa} P_c \;=\; S^{-\kappa} BC \;=\; S^{-\kappa} F_1 F_2 \cdots F_N.$$

2 The index and the plus-index

We say that an infinite matrix $A = (a_{ij})_{i,j\in\mathbb{Z}}$ is *invertible* if the linear operator that it represents via matrix-vector multiplication is invertible as an operator from $\ell^2(\mathbb{Z})$ to $\ell^2(\mathbb{Z})$. A bounded linear operator A from ℓ^2 into ℓ^2 is invertible iff it is both injective (its nullspace $N(A) = \{x \in \ell^2 : Ax = 0\}$ consists of 0 only) and surjective (its range $R(A) = \{Ax : x \in \ell^2\}$ is all of ℓ^2). Deviation from both properties is measured in terms of the two integers

$$\alpha \;=\; \dim N(A) \qquad \text{and} \qquad \beta \;=\; \operatorname{codim} R(A). \tag{4}$$

A is a *Fredholm operator* if both α and β are finite. Then the Fredholm index (or just the *index*) of A is the difference

$$\operatorname{index}(A) \;=\; \alpha - \beta. \tag{5}$$

Unlike the separate numbers α and β, their difference $\alpha - \beta$ has these important properties:

- the index is invariant under compact perturbations $A + K$;
- the index of a product obeys the remarkable formula $\operatorname{index}(AB) \;=\; \operatorname{index}(A) + \operatorname{index}(B)$;
- the index is continuous with respect to the operator norm of A (and therefore locally constant); and
- all finite square matrices have index zero.

All Fredholm matrices of index zero, including all finite square matrices, obey the *Fredholm alternative*: Either A is injective and surjective ($\alpha = 0$, $\beta = 0$) or A is not injective and not surjective ($\alpha \neq 0$, $\beta \neq 0$). This is the set of all operators $A = C + K$ with C invertible and K compact. See e.g. [7, 10] for a nice introduction to Fredholm operators. The property of being invertible on $\ell^p(\mathbb{Z})$ and the index itself are independent of $p \in [1, \infty]$ for a certain class of infinite matrices (with uniformly bounded entries and summable off-diagonal decay), see [12, 14].

Our interest in the index (more precisely, the plus-index) of a biinfinite matrix originates from the following natural question:

Which diagonal is the main diagonal of a biinfinite matrix?

For a symmetric matrix, the zeroth diagonal is the main diagonal. For a Toeplitz matrix (a polynomial in the shift S), a zero winding number is the key. A wider class of structured matrices was analyzed by de Boor [8]. For Fredholm operators in general, Israel Gohberg's diplomatic answer to this question was that every diagonal has the right to be the main diagonal [3, p. 24]. But there are concrete problems waiting for a concrete answer. Here are two such problems:

1. For finite and semiinfinite matrices, the inverse of a lower triangular matrix is again lower triangular. This may fail if A is biinfinite (essentially because the 'wrong' diagonal is

mistaken for the main diagonal). The inverse of the lower triangular shift S is the upper triangular backward shift $S^\top$. Shifting S one row up (treating the ones as the main diagonal) resolves this conflict.

Here is a slightly more sophisticated example: $A = S - \frac{1}{2}S^2$ is lower triangular with its inverse neither lower nor upper triangular. Shift up by one row: $B = S^{-1}A = I - \frac{1}{2}S$ is lower triangular with $B^{-1} = I + \frac{1}{2}S + \frac{1}{4}S^2 + \cdots$ also lower triangular (as we want). Shift up one more row: $C = S^{-2}A = S^{-1} - \frac{1}{2}I$ is now upper triangular with $C^{-1} = B^{-1}S$ lower triangular. In this example the diagonal consisting of all ones should be the main diagonal, and B is centered (even though triangular) so that its inverse is triangular of the same kind.

2. The numerical solution of biinfinite systems $Ax = b$ can approximate A^{-1} by the inverses of finite square submatrices (*finite sections*) of A. Their upper left and lower right corners lie on the main diagonal of A. But again, which diagonal should that be? For $A = 2S^2 + \frac{1}{2}S - I + 2S^{-1}$, one can show [4, 5] that it has to be the diagonal that carries all the $\frac{1}{2}$'s. Finite sections that are centered along one of the other nonzero diagonals are invertible (for sufficiently large sizes) but their inverses do not converge (they blow up).

Problem 1 and 2 are not unrelated: If A is lower triangular and the inverses of its finite sections (which are also lower triangular) approximate A^{-1}, then A^{-1} will be lower triangular, too. A similar argument can be used to transfer Asplund's theorem [2, 25] to infinite matrices, explaining the relations between ranks of submatrices of A and A^{-1}. We demonstrate this in Part II of this paper.

The answer to both problems is the same: *Shift A down by κ rows*, or equivalently by κ diagonals, where κ is the *plus-index* of A [17]:

$$\kappa = \operatorname{index}_+(A) = \operatorname{index}(A_+) \qquad \text{and} \qquad A_+ = \begin{pmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & a_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{6}$$

A_+ is a semiinfinite submatrix of A. This shifting process is called *index cancellation* [9, 11].

The centered matrix $A_c = S^\kappa A$ has $\operatorname{index}_+(A_c) = 0$ (see (16) below). This is necessary for a triangular matrix to have a triangular inverse of the same kind (Corollary 6.4). It is also necessary for the convergence of the inverses of the finite sections to A^{-1} (see [15, 20]).

How is this plus-index κ computed? In [17] it is shown that κ is invariant under passing to a "limit operator" [18, 13, 6] of A at $+\infty$, which often simplifies the computation. For the plus-index of a permutation P, the limit operator is not needed and the new formula (12) is as simple as possible.

Part I

Permutation matrices

3 The plus-index of a biinfinite permutation matrix

Now we come to the problem of computing the plus-index of a biinfinite permutation matrix. So let $\pi : \mathbb{Z} \to \mathbb{Z}$ be a permutation (a bijection) of the integers and put $P = (p_{ij})_{i,j\in\mathbb{Z}}$ with

$$p_{ij} \;=\; \delta_{\pi(i),j} \;=\; \begin{cases} 1 & \text{if } j = \pi(i), \\ 0 & \text{if } j \neq \pi(i). \end{cases}$$

The matrix P is banded iff the maximal displacement w of any integer via π is *finite*:

$$w \;:=\; \sup_{i\in\mathbb{Z}} |i - \pi(i)| \quad \text{is the bandwidth of } P. \tag{7}$$

Every permutation matrix P is invertible. So the submatrix P_+ is always Fredholm (see Lemma 5.2 below) and it makes sense to ask for its index: the plus-index of P.

It seems a bit arbitrary to define the plus-index of a biinfinite matrix A based on the submatrix A_+ that starts at the particular entry a_{11}. Lemma 3.1 shows that for banded permutations, the submatrix A_k starting at a_{kk} gives the same plus-index for every $k \in \mathbb{Z}$:

$$A_k \;:=\; (a_{ij})_{i,j=k}^{\infty} \;=\; \begin{pmatrix} a_{k,k} & a_{k,k+1} & \cdots \\ a_{k+1,k} & a_{k+1,k+1} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{8}$$

Lemma 3.1 *If P is a banded biinfinite permutation matrix and $k \in \mathbb{Z}$ then* $\operatorname{index}_+(P) = \operatorname{index}(P_k)$ *holds independently of k.*

Proof. Let $k \in \mathbb{N}$ first. Because the first k rows and columns of P_0 contain only finitely many nonzero entries, we have the following equality modulo finite rank operators:

$$P_0 \;\cong\; \left(\begin{array}{c|c} 0_{k\times k} & 0 \\ \hline 0 & P_k \end{array}\right) \;\cong\; \left(\begin{array}{c|c} I_{k\times k} & 0 \\ \hline 0 & P_k \end{array}\right) \tag{9}$$

Consequently, $\operatorname{index}(P_0) = \operatorname{index}(P_k)$. For $k \in \mathbb{Z} \setminus \mathbb{N}$ the argument is very similar. ■

We will write P_+ for any of these singly infinite submatrices P_k. Notice that P_+ can be the zero matrix (not Fredholm) when P is not banded. An example is the permutation that exchanges every pair i and $-i$, for $i \in \mathbb{Z}$.

Here is a concrete formula for the plus-index of permutation matrices that "split":

Theorem 3.2 *Let π be a permutation of the integers (not necessarily banded) and denote the corresponding matrix by P. Suppose there exist i^* and j^* such that*

$$\{\pi(i) \ : \ i < i^*\} \ = \ \{j \ : \ j < j^*\}. \tag{10}$$

Then $\mathrm{index}_+(P) = j^* - i^*$.

Proof. If such integers i^* and j^* exist then P decouples into two blocks:

$$P \ = \ \left(\begin{array}{cc|cc} \ddots & \vdots & & \\ \cdots & p_{i^*-1,j^*-1} & & \\ \hline & & p_{i^*,j^*} & \cdots \\ & & \vdots & \ddots \end{array}\right) \ =: \ \left(\begin{array}{c|c} P^{(1)} & 0 \\ \hline 0 & P^{(2)} \end{array}\right) \tag{11}$$

The meeting point (i^*, j^*) may not fall on the zeroth diagonal of P (in fact, it falls on the main diagonal). Because P is invertible, $P^{(1)}$ and $P^{(2)}$ are invertible. It is now easy to show that $j^* - i^*$ is the plus-index of P:

Case 1. If $i^* \le j^*$ then, putting $k := i^*$, we have that P_k starts with $j^* - i^*$ zero columns. Those are followed by $P^{(2)}$, so that the index of P_k is $(j^* - i^*) - 0 = j^* - i^*$.

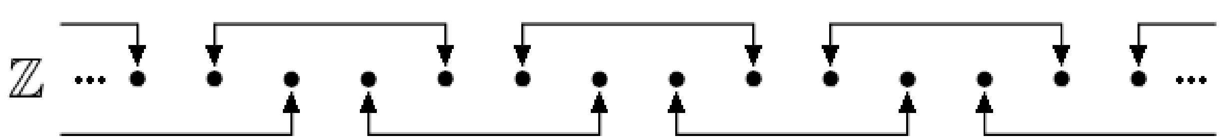
Case 2. If $i^* > j^*$ then, putting $k := j^*$, we have that P_k has $i^* - j^*$ zero rows, followed by $P^{(2)}$. Again the index is $0 - (i^* - j^*) = j^* - i^*$.

It remains to apply Lemma 3.1. ■

Adding $\kappa = j^* - i^*$ to all row numbers moves the meeting point of the blocks in (11) to position (j^*, j^*). In fact $\mathrm{index}_+(A) = j^* - i^*$ holds for all invertible matrices A (not just permutations) that decouple in the sense of (11).

The first question is whether or not such a splitting will appear in every permutation of $\mathbb{Z}$ – and the unfortunate answer is no!

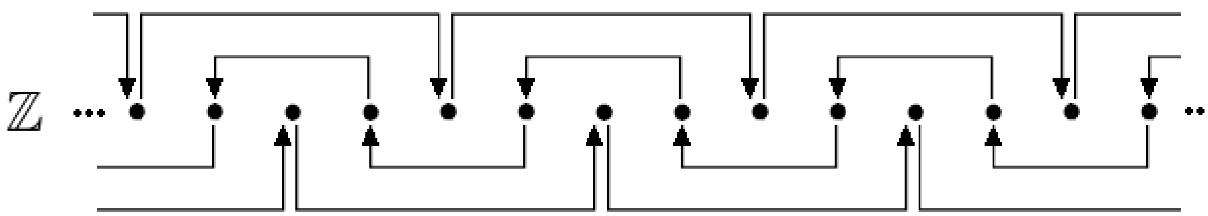
Example 3.3 (symmetric and intertwined) Let us depict this permutation by a graph with vertex set $\mathbb{Z}$ and with a directed edge (an arrow) from i to j iff $j = \pi(i)$.



This permutation is symmetric ($\pi(i) = j$ iff $\pi(j) = i$). In such a case, splitting happens iff the graph falls into separate components $\{i < i^*\}$ and $\{i \ge i^*\}$. Then the graph can be split in two without cutting any edge, which is obviously impossible in our example.

Of course, this symmetric sitation means that $P = P^\top$ and hence $P_+ = (P_+)^\top$ so that $\mathrm{index}(P_+)$ must be zero – and we don't need help from Theorem 3.2. But here is a non-symmetric version of essentially the same example: □

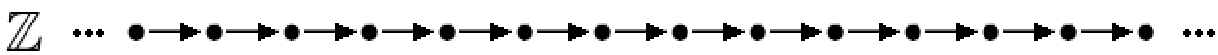
Example 3.4 (previous example, shifted) $\pi'(i) := \pi(i) + 1$.



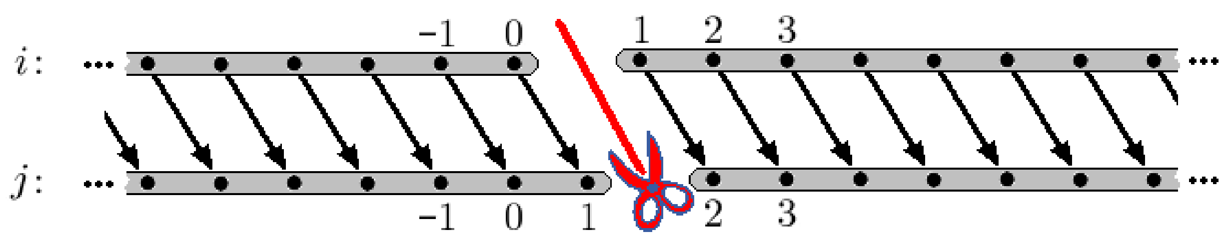
Without symmetry it is not so easy to see that splitting is impossible. In fact, this graph is disconnected (it has three components) but it cannot be split. □

There is a simple trick to make splitting immediately visible in every permutation's graph. We will demonstrate this trick for a simpler example 3.5.

Example 3.5 Let $\pi(i) := i + 1$. The corresponding matrix is our shift, S.

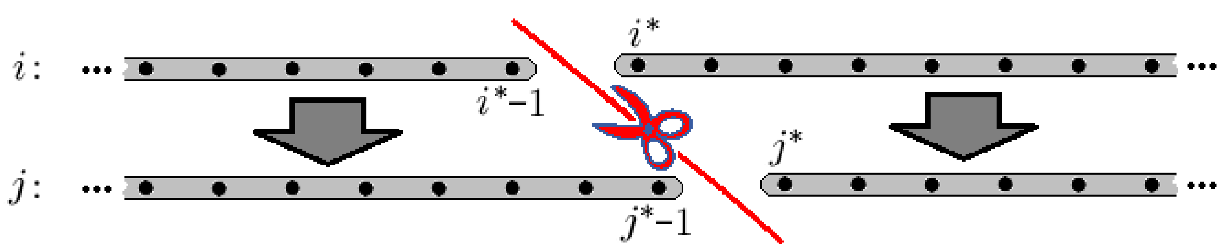


The graph is clearly connected but we have a split at every single position i^*, with $j^* = i^* + 1$ in (10). This is more easily seen by depicting the i's and the j's separately, where we draw an arrow from i to j if $j = \pi(i)$:



□

In general, splitting (10) is equivalent to a separation like this:

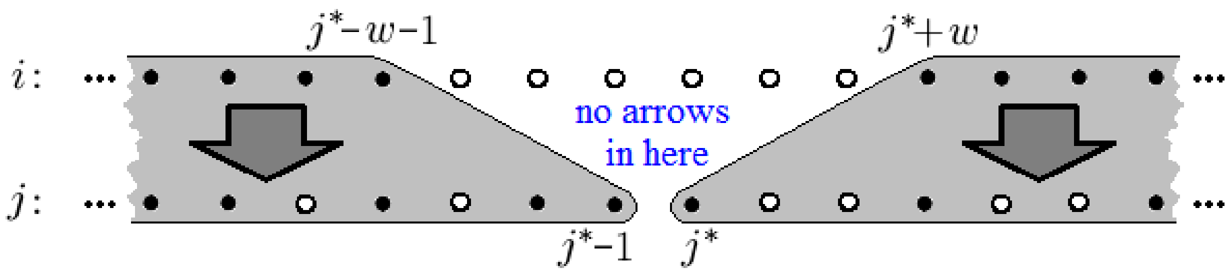


The nodes $\{i < i^*\}$ are only connected to $\{j < j^*\}$ and the nodes $\{i \geq i^*\}$ only to $\{j \geq j^*\}$. By shifting the i-axis accordingly, one can straighten the cut to perfectly vertical. This is exactly what index cancellation does. Shifting the i-axis corresponds to renumbering the rows of our matrix. The plus index $j^* - i^*$ is obvious if one thinks in terms of these i-j-graphs rather than the matrix. The conclusion that there is no split in Example 3.4 is now a simple exercise.

How do we compute the plus-index when the graph does not split? Here is our agenda:

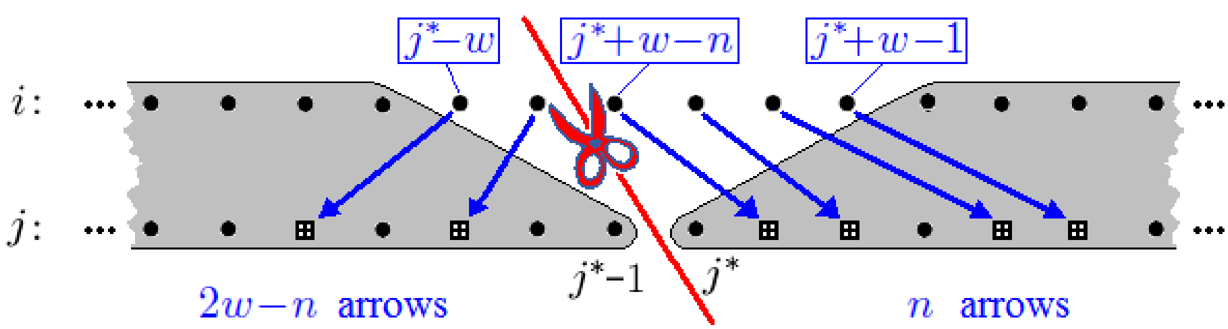
- By changing finitely many arrows, we construct a new permutation π' that does split.
- This is possible with the split at any given position j^*.
- Changing finitely many arrows only changes finitely many matrix entries. Therefore π and π' have the same plus-index, and we know that the index for π' is $j^* - i^*$.

Step 1: Delete arrows. Choose an arbitrary position $j^* \in \mathbb{Z}$ where the split should cross the j-axis. First delete the $2w$ arrows $i \mapsto \pi(i)$ with $i = j^* - w, ..., j^* + w - 1$ from the graph of π. Then the remaining diagram will have a big gap:



The endpoints of the $2w$ deleted arrows are now empty. Put those endpoints (the column numbers) into *ascending order*. Suppose n of those column numbers are greater than or equal to j^* – their endpoints are to the right of the split. The other $2w - n$ endpoints are to the left.

Step 2: Rewire. Insert $2w$ new arrows connecting the empty starting points $i = j^* - w, ..., j^* + w - 1$ to the (*ordered!*) empty endpoints.



The new permutation π' splits at $i^* = j^* + w - n$. If P' is the matrix for π', then $P'_+ - P_+$ is of finite rank. Therefore P_+ and P'_+ have the same index. This index comes directly from the splitting point of π':

$$\kappa \; = \; \text{index}_+(P) \; = \; \text{index}(P_+) \; = \; \text{index}(P'_+) \; = \; j^* - i^* \; = \; n - w. \tag{12}$$

Remark 3.6 We are just reordering the 1's in $2w$ consecutive rows of P. In the new order, those 1's go from left to right. The figure has $w = 3$ and there are $n = 4$ ones on the right. Then in this example the index is $\kappa = 4 - 3 = 1$. □

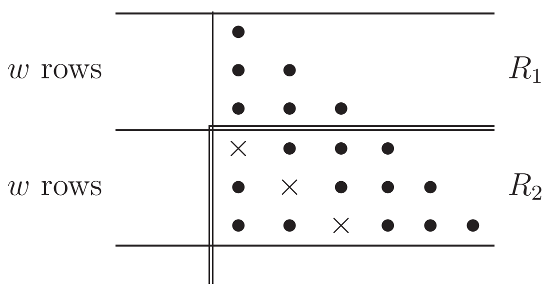
The number of 1*'s in columns* $j \geq j^*$ *is* n. Thus we are counting the 1's in the following submatrix R, and subtracting the bandwidth w to obtain the index $\kappa = n - w$:

$$R = \begin{pmatrix} p_{j^*-w,j^*} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ p_{j^*+w-1,j^*} & \cdots & p_{j^*+w-1,j^*+2w-1} \end{pmatrix} \tag{13}$$

This submatrix R is lower triangular because P has bandwidth w. This completes our first proof of Theorem 1.

To some extent this result is surprising. The Fredholm index is robust under perturbations in finitely many matrix entries – so it must be encoded somewhere deep down at infinity. (This is exactly where the limit operator approach [18, 13, 6], that we mentioned before, enters the stage.) But for banded permutation matrices, the plus-index can be computed from a finite submatrix – *independent of the position*. This is clearly a consequence of the 'stiff' rules (exactly one 1 in each row and column – and never leave the band!) that these matrices have to follow.

Second proof of Theorem 1. The first proof identified the "right half" of any $2w$ consecutive rows of P, as sufficient to determine the index of P_+. Now we can give a direct proof by counting the 1's in that right half R. In the figure below, the infinite submatrix P_+ is marked out by double lines, starting on the zeroth diagonal (marked with $\times$'s). We divide R into a part R_1 outside P_+ and a part R_2 inside P_+. Now count the 1's in each part:



Each 1 in R_1 will mean a zero column below it in P_+. The other columns of P_+ are complete:

$$\text{dimension of nullspace of } P_+ = \text{number of 1's in } R_1.$$

Any zero rows in P_+ will be zero rows in R_2. Those are counted by the 1's that are missing from the w rows of R_2:

$$\text{codimension of range of } P_+ = w - (\text{number of 1's in } R_2).$$

Subtraction gives the index formula in Theorem 1:

$$\text{index of } P_+ = (\text{number of 1's in } R) - w = n - w. \quad \blacksquare \tag{14}$$

4 Factorizations of banded permutations

This short section establishes the factorizations $P_c = BC$ and $P_c = F_1F_2 \cdots F_N$ of centered permutation matrices P_c (plus-index equal to 0). The block-diagonal permutations B, C have block size $2w$ and $F_1, \cdots, F_N$ have block size 1 or 2.

Both factorizations are particularly simple cases of a more general theorem [22, 23] for banded matrices with banded inverses. For permutations, P and $P^{-1} = P^{\top}$ have the same bandwidth w. Figure 4.1 below shows a typical matrix P_c with $w = 2$. Each non-empty square of size $2w = 4$ contains two 1's (by Theorem 1), because P_c is centered ($\kappa = 0$).

Step 1. There is a 1 in each row and column. Row exchanges from the block B_1^{-1} will move the 1's in the first square into the top two rows (and thus 1's go into the last two rows of the second square). The block B_2^{-1} acts in the same way on the next four rows of P_c. All the circled entries $\odot$ have become zero.

Step 2. C_1^{-1} executes column exchanges to move the $2 + 2 = 4$ ones in its columns to the main diagonal (marked by $\times$). When all blocks B_i^{-1} and C_i^{-1} act in this way, we reach the identity $B^{-1} P_c C^{-1} = I$ and hence $P_c = BC$. This is Theorem 2.

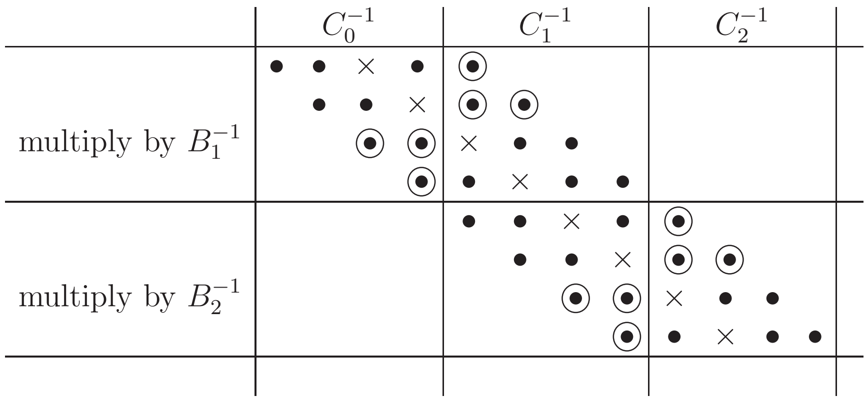


Figure 4.1: Each square has rank 2. B_1^{-1} and B_2^{-1} produce zeros in the circled positions by row exchanges. Then C_1^{-1} produces (by column exchanges) the 4 by 4 identity matrix with ones in the diagonal positions $\times$.

A similar factorization was established in [21] for any banded matrix with banded inverse. The same result was found for unitary banded matrices in [17], with a different proof. It appears that $2w$ is the right block size for these block-diagonal factorizations $A = BC$.

Theorem 3 says that it is possible to factor P_c into even simpler block-diagonal permutations $F_1 F_2 \cdots F_N$. Now the blocks of each factor are 2×2 or 1×1. Thus each F_i exchanges a set of disjoint pairs of neighbors (as in the bubblesort algorithm, but with disjoint exchanges in parallel for much greater efficiency).

A beautiful proof that $N < 2w$ was given by Greta Panova [16, 1, 23]. We won't repeat the details. Her key idea is a variation on the "wiring diagram" of a permutation.

Briefly, that diagram connects each point $(0, i)$ for $i \in \mathbb{Z}$ by a straight line to the point $(1, \pi_c(i))$. The intersections of those lines tell us the order in which to exchange neighbors. *Key point*: All intersections lie on $N < 2w$ vertical lines in the [16, 23]. Exchanges on each vertical line can be executed in parallel by a matrix F_i (its bandwidth is 1). Then P_c is the product $F_1 F_2 \cdots F_N$ with $N < 2w$.

Part II

Band-dominated matrices

5 The plus-index of a band-dominated matrix

A is a *band matrix* if it is supported on finitely many diagonals only, and it is a *band-dominated matrix* if it is the limit (in the $\ell^2 \to \ell^2$ operator norm) of a sequence of band matrices.

A_k is the semiinfinite submatrix (8) consisting of rows $i \geq k$ and columns $j \geq k$. Then Lemma 3.1 generalizes from permutation matrices to band-dominated matrices that are Fredholm:

Lemma 5.1 *If A is band-dominated and Fredholm then the index of A_k is independent of $k \in \mathbb{Z}$.*

Proof. If A is a band matrix then the proof is literally that of Lemma 3.1 above. If A is band-dominated (i.e. in the closure of the set of band matrices) then the identification (9) still holds modulo compact operators, which is enough to prove the claim. ■

In analogy to $A_+ = A_1$ we introduce the semiinfinite matrix A_- that ends at $a_{0,0}$:

$$A_- \; := \; (a_{ij})_{i,j=-\infty}^{0} \; = \; \begin{pmatrix} \ddots & \vdots & \vdots \\ \cdots & a_{-1,-1} & a_{-1,0} \\ \cdots & a_{0,-1} & a_{0,0} \end{pmatrix}$$

Then A_- and A_+ will together determine the index of A:

$$A \; = \; \left(\begin{array}{c|c} A_- & A_{-+} \\ \hline A_{+-} & A_+ \end{array}\right) \; \cong \; \left(\begin{array}{c|c} A_- & 0 \\ \hline 0 & A_+ \end{array}\right) \tag{15}$$

modulo compact operators since A_{-+} and A_{+-} are compact. (Those off-diagonal blocks have finite rank if A is a band matrix and they are norm limits of finite rank matrices, hence compact, if A is band-dominated.) From this identification (15) we immediately have Lemma 5.2:

Lemma 5.2 *[17] A band-dominated biinfinite matrix A is Fredholm iff both A_+ and A_- are Fredholm. In that case*

$$\text{index}(A) \; = \; \text{index}(A_+) \; + \; \text{index}(A_-).$$

So if A is invertible then $\text{index}(A_+) = -\text{index}(A_-)$. Clearly, all results that we prove for the plus-index have their counterpart for the minus-index, $\text{index}_-(A) = \text{index}(A_-)$. Here is another important result on the plus-index:

Lemma 5.3 *If A and B are band-dominated and Fredholm biinfinite matrices then*

$$\text{index}_+(AB) \; = \; \text{index}_+(A) \; + \; \text{index}_+(B).$$

Proof. By (15) we have, modulo compact operators,

$$AB \;\cong\; \left(\begin{array}{c|c} A_- & 0 \\ \hline 0 & A_+ \end{array}\right)\left(\begin{array}{c|c} B_- & 0 \\ \hline 0 & B_+ \end{array}\right) \;=\; \left(\begin{array}{c|c} A_-B_- & 0 \\ \hline 0 & A_+B_+ \end{array}\right),$$

so that $(AB)_+ \cong A_+B_+$, whence $\mathrm{index}((AB)_+) = \mathrm{index}(A_+B_+) = \mathrm{index}(A_+) + \mathrm{index}(B_+)$. ■

Recalling the forward shift S with $\mathrm{index}_+(S) = 0 - 1 = -1$, Lemma 5.3 yields

$$\mathrm{index}_+(S^\kappa A) \;=\; \kappa \cdot \mathrm{index}_+(S) \;+\; \mathrm{index}_+(A) \;=\; -\kappa \;+\; \mathrm{index}_+(A). \tag{16}$$

Then if $\kappa = \mathrm{index}_+(A)$, index cancellation $A_c := S^\kappa A$ indeed leads to plus-index zero.

6 Triangular matrices

The plus-index of a general band-dominated matrix A is not easy to compute. We will assume that A is banded and invertible (as a bounded operator on $\ell^2(\mathbb{Z})$). One possible approach (far from complete in this paper) is to factor A into a lower triangular L and an upper triangular U, with a banded permutation P in between:

$$A = LPU = \text{(lower triangular)(permutation)(upper triangular)}.$$

This extends the factorization that comes from Gaussian elimination on a finite invertible matrix. Notice the "Bruhat convention" that places P between L and U. In that position P is unique. When P is a finite matrix, the 1's are all determined by the ranks of the upper left submatrices of A. Elimination proceeds as normal (subtracting multiples of the pivot rows from lower rows) except that we wait to the end to reorder the pivot rows into U by using P.

The steps are described in [24], where the main purpose is to extend $A = LPU$ to biinfinite matrices A (banded and invertible). The factors L, P, U have bandwidth $\le 2w$. But the elimination process has to be reconsidered, since it can no longer start at a_{11} (which is not the first entry, it is in the center of the matrix). Briefly, we look at the upper left submatrices A_{k-} (singly infinite) containing columns $\le k$ and rows $\le k + w$. By Fredholm theory, the index of A_{k-} is independent of k. Its columns are independent because A is invertible (and all nonzeros survive into A_{k-}). Any dependent rows of A_{k-} are among the last $2w$ rows (for the same reason). So the minus-index of A and the location of its main diagonal are determined by the number d of dependent rows.

Since d does not depend on k, one row changes to independent and a new dependent row appears (if $d > 0$) when k increases to $k+1$. That newly independent row is the pivot row in the following elimination step. Dependence involves *all* the earlier rows of A_{k-}, so this elimination process is not constructive—at least not in the usual sense.

Note: We say that an infinite set of vectors $\{v_i\}_{i\in\mathbb{I}}$ with $\mathbb{I} \subset \mathbb{Z}$ is *linearly dependent* if there is a non-zero sequence $(c_i)_{i\in\mathbb{I}} \in \ell^2(\mathbb{I})$ such that $\sum_{i\in\mathbb{I}} c_i v_i = 0$. In that sense, the columns of an infinite matrix form a linearly independent set iff that matrix is injective as an operator on ℓ^2.

Returning for a moment to permutation matrices, rows of P_{k-} are dependent when they are zero. They are independent when they contain a 1. Therefore the number d of zero rows (in the left half of any $2w$ rows of P) involves the same count of 1's as in Theorem 1 (in the right half of those rows).

Now suppose that A is invertible and banded, with $A = LPU$, where all three factors have bounded inverses. (This can fail even for block-diagonal matrices A with 2×2 orthogonal blocks. The upper left entries in those blocks can approach zero.) As usual, L and L^{-1} are lower triangular, P is a permutation, and U and U^{-1} are upper triangular. Then P contains all information about indices and the correct position of the main diagonal.

Lemma 6.1 *The plus-index of A equals the plus-index of P (and that is easily computable from P). Also the minus-indices are equal.*

Proof. As shown in [24], L and U are both banded, with bandwidth $\leq 2w$. By Corollary 6.4 below, $\text{index}_+(L)$ and $\text{index}_+(U)$ are zero. It follows from Lemma 5.2 that also $\text{index}_-(L)$ and $\text{index}_-(U)$ are zero. Now use the key property that L and U are triangular:

$$\begin{pmatrix} A_- & A_{-+} \\ A_{+-} & A_+ \end{pmatrix} = \begin{pmatrix} L_- & 0 \\ L_{+-} & L_+ \end{pmatrix} \begin{pmatrix} P_- & P_{-+} \\ P_{+-} & P_+ \end{pmatrix} \begin{pmatrix} U_- & U_{-+} \\ 0 & U_+ \end{pmatrix}.$$

The upper left block A_- immediately factors into

$$A_- = L_- P_- U_-. \tag{17}$$

By the formula for the index of a product we get

$$\text{index}(A_-) \ = \ \text{index}(L_-) + \text{index}(P_-) + \text{index}(U_-) \ = \ \text{index}(P_-) \tag{18}$$

and then $\text{index}_+(A) = -\text{index}_-(A) = -\text{index}_-(P) = \text{index}_+(P)$, again by Lemma 5.2 since A and P have index zero (they are invertible). ■

Alternatively, one can apply Lemma 5.3 directly to $A = LPU$ to get

$$\text{index}_+(A) \ = \ \text{index}_+(LPU) \ = \ \text{index}_+(L) + \text{index}_+(P) + \text{index}_+(U) \ = \ \text{index}_+(P)$$

and then conclude $\text{index}_-(A) = \text{index}_-(P)$ from Lemma 5.2.

We continue with some results on lower triangular biinfinite matrices and their plus-index (which locates their main diagonal). It is not surprising, but however in need of a proof, that the main diagonal of a lower triangular matrix has to be in the lower triangle. This is what we show now. So let A be a biinfinite matrix.

Lemma 6.2 *If A is lower triangular and Fredholm (but not necessarily band-dominated) then, for sufficiently large k, the submatrix A_k from (8) is injective (as an operator on $\ell^2(\mathbb{N})$).*

First proof. If A is Fredholm and $\alpha = \dim(N(A))$ then there is a set $J = \{j_1, ..., j_\alpha\}$ of integers such that the columns of A whose number is not in J form a linearly independent set. Let k be bigger than $max(J)$. Then all columns $j \geq k$ of A and hence (because A is lower triangular) all columns of the semiinfinite submatrix A_k are linearly independent. So A_k is injective. ■

Second proof. Since A is Fredholm on ℓ^2, there are operators B and K such that $BA = I+K$ with B bounded and K compact on ℓ^2 [7, 10]. For $k \in \mathbb{Z}$, let I_k denote the operator $\ell^2 \to \ell^2$ that puts all entries x_i of $x \in \ell^2$ with $i < k$ to zero and leaves all x_i with $i \geq k$ unchanged. (I_k is (8) for $A = I$.) Since A is lower triangular, we have $I_kAI_k = AI_k$ for all $k \in \mathbb{Z}$. Moreover, $\|I_kx\|_{\ell^2} \to 0$ as $k \to +\infty$ for all $x \in \ell^2$. By compactness of K, it follows [10] that the operator norm $\|I_kK\|$ goes to zero as $k \to +\infty$. So fix $k \in \mathbb{Z}$ large enough that $\|I_kK\| < 1$. Then $C = I + I_kK$ is invertible (by Neumann series). Now

$$I_kBI_kAI_k \;=\; I_kBAI_k \;=\; I_k(I+K)I_k \;=\; I_k + I_kKI_k \;=\; (I+I_kK)I_k \;=\; CI_k,$$

so that $C^{-1}I_kBI_kAI_k = I_k$. Now take x from the range of I_k such that $I_kAI_kx = 0$. Then $0 = C^{-1}I_kBI_kAI_kx = I_kx = x$, so that $x = 0$ is the only x in the range of I_k with $I_kAI_kx = 0$. By $I_kAI_k|_{R(I_k)} = A_k$ this means that the only solution $y = (y_i)_{i=k}^{\infty}$ of $A_ky = 0$ is the trivial solution $y = 0$. ■

Corollary 6.3 *If A is band-dominated, lower triangular and Fredholm then* $\operatorname{index}_+(A) \leq 0$.

Proof. By Lemma 6.2, the operator behind A_k has nullspace $\{0\}$ for all sufficiently large $k \in \mathbb{Z}$, so that $\operatorname{index}(A_k) = 0 - \beta \leq 0$. The claim now follows from Lemma 5.1. ■

So indeed, the main diagonal of a lower triangular matrix is in the lower triangle. By simple translation via S^d, the main diagonal of a matrix that is zero above the d-th diagonal must be on or below that d-th diagonal. By passing to the adjoint matrix, one can write down an analogous statement for upper triangular matrices and their translates and then combine the two: The main diagonal of a band matrix must be in that band.

Consequently, tridiagonal matrices have plus-index $\kappa = -1$, 0 or 1. Using this fact and our Theorem 1, it is easy to see that the only tridiagonal permutation matrices that are not centered (i.e. they have a nonzero plus-index) are the shifts S and S^{-1}. Indeed, suppose P is a biinfinite permutation matrix with bandwidth $w = 1$ and plus-index $\kappa = -1$. Then, for every $j^* \in \mathbb{Z}$, the number of ones in the 2×2 matrix (13) is $n = \kappa + w = 0$. So the 1 in row j^* must be in column $j^* - 1$, whence P must be S. Similarly, $w = 1$ and $\kappa = 1$ leads to $P = S^{-1}$.

Here is another consequence of Corollary 6.3:

Corollary 6.4 *If A is band-dominated, lower triangular and invertible with a lower triangular inverse then* $\operatorname{index}_+(A) = 0$.

Proof. By Corollary 6.3, we have $\kappa = \operatorname{index}_+(A) \leq 0$ and $\lambda = \operatorname{index}_+(A^{-1}) \leq 0$. But then $\kappa = -\lambda$ by Lemma 5.3, so that κ must be zero. ■

The latter shows how Problem 1 from the beginning of our paper is related to the plus-index. After discussing Problems 1 and 2, we mentioned that they are related: If the finite section method applies to A then a lower triangular A will have a lower triangular inverse. Now we discuss an amazing extension of this statement, coming from the following theorem (for finite matrices) by Asplund [2] :

Theorem 6.5 *Let A be an invertible matrix and fix two integers p and k. Then the following are equivalent:*

(i) All submatrices B above the p-th superdiagonal of A have $\mathrm{rank}(B) < k$*.*

(ii) All submatrices C above the p-th subdiagonal of A^{-1} have $\mathrm{rank}(C) < p + k$*.*

See [25] for discussion and a new proof. With $p = 0$ and $k = 1$ we get the familiar statement that lower triangular matrices have lower triangular inverses. With $p = 1$ and $k = 1$, one can see that all submatrices above the first subdiagonal (and similarly: below the first superdiagonal) of A^{-1} have rank < 2 if A is tridiagonal.

Let us attempt to transfer Asplund's theorem to singly and doubly infinite matrices. Our assumption will be that the *finite section method* (short: *FSM*) applies to the infinite matrix A. Here is again what that means:

The *finite sections* of A are square submatrices A_n whose upper left and lower right corners lie at positions l_n and r_n on the zeroth diagonal of A. The sequences l_n and r_n go to $-\infty$ and $+\infty$ respectively, except when A is only semi-infinite – then l_n is fixed at 1 and r_n goes to $+\infty$. One says that the FSM applies to A if: A is invertible, the matrices A_n are invertible for sufficiently large n, and their inverses converge strongly to A^{-1}. This implies that A is centered (see [15, 20]). Also note that strong convergence ($A_n^{-1}x \to A^{-1}x$ for all x) implies entrywise convergence of the matrices A_n^{-1} to A^{-1}. See e.g. [13, 15, 18, 19, 20] and the references therein for more on the FSM.

Under this (reasonable) assumption, we now extend Asplund's Theorem 6.5 to a singly or doubly infinite matrix A (not necessarily band-dominated):

Proof. Suppose (i) holds, i.e. all (finite and infinite) submatrices B above the p-th superdiagonal of A have $\mathrm{rank}(B) < k$. To show that (ii) holds, let C be an arbitrary (finite or infinite) submatrix above the p-th subdiagonal of A^{-1}. We show that $\mathrm{rank}(C) < p + k$.

If C is infinite then $\mathrm{rank}(C)$ is the supremum of all $\mathrm{rank}(C')$ with C' going through all *finite* submatrices of C. (In particular, $\mathrm{rank}(C) = \infty$ iff the set of all $\mathrm{rank}(C')$ is unbounded.) Therefore it is enough to show that $\mathrm{rank}(C') < p + k$ for all *finite* submatrices of A^{-1} above the p-th subdiagonal. So we can assume that C is a finite matrix.

Let $l_1, l_2, \ldots$ and $r_1, r_2, \ldots$ be the cut-off positions for the (by our assumption applicable) finite sections of A. For all sufficiently large n (say $n > N$) the interval $[l_n, r_n]$ contains the row and column numbers in which C is positioned at A^{-1}. By applicability of the FSM, the inverses of our finite sections A_n of A converge entrywise to A^{-1}. Let C_n denote the submatrix of A_n^{-1} that is at the same position as C is in A^{-1}, so that $C_n \to C$ entrywise as $n \to \infty$.

Now we apply Theorem 6.5 to the *finite* matrix A_n: All submatrices above the p-th superdiagonal of A_n have rank $< k$, by (i), because they are submatrices of A (above the p-th superdiagonal). So C_n, being a submatrix above the p-th subdiagonal of A_n^{-1}, has rank $< p+k$. And this is true for all $n > N$. For finite matrices, entrywise convergence is convergence in all matrix norms, so that $\|C_n - C\| \to 0$. Rank is a lower semi-continuous function with respect to the matrix norm, so

$$\mathrm{rank}(C) \;\leq\; \liminf_{n\to\infty} \mathrm{rank}(C_n) \;<\; p + k$$

and we are done with $(i) \Rightarrow (ii)$. The other direction is checked analogously. ■

Remark 6.6 Our proof shows that applicability of the FSM is sufficient for Asplund's theorem to hold for an infinite matrix. To see that it is not necessary, go back to the permutation matrix P in Example 3.3. Asplund's theorem can be seen to hold for this matrix (note that $P^{-1} = P^\top = P$) but the FSM does not apply: No one of the finite sections is invertible because there are no places l_n and r_n to cut without breaking one of the links in this permutation graph. □

Acknowledgements. We would like to thank Steffen Roch for helpful discussions. Moreover, the first author acknowledges the financial support by the Marie-Curie Grant PERG02-GA-2007-224761 of the EU.

References

[1] C. Albert, Chi-Kwong Li, G. Strang and Gexin Yu, Permutations as products of parallel transpositions, *SIAM J Discrete Math* **25** (2011), 1412–1417.

[2] E. Asplund: Inverses of matrices $\{a_{ij}\}$ which satisfy $a_{ij} = 0$ for $j > i + p$, *Math Scand* **7** (1959), 57–60.

[3] A. Böttcher: Infinite matrices and projection methods, *in: "Lectures on Operator Theory and Its Applications", P. Lancaster (ed.), Fields Institute Monographs*, Vol. 3, *Amer. Math. Soc.*, 1996, 1–72.

[4] A. Böttcher and S. M. Grudsky: *Spectral Properties of Banded Toeplitz Matrices*, SIAM, 2005.

[5] A. Böttcher and B. Silbermann: *Introduction to Large Truncated Toeplitz Matrices*, Springer, 1999.

[6] S. N. Chandler-Wilde and M. Lindner: *Limit Operators, Collective Compactness, and the Spectral Theory of Infinite Matrices*, Memoirs of the AMS, Vol. 210, Nr. 989, 2011.

[7] E. B. Davies: *Linear Operators and Their Spectra*, Cambridge University Press, 2007.

[8] C. de Boor: What is the main diagonal of a biinfinite band matrix?, *in "Quantitive Approximation", R. DeVore and K. Scherer (eds.), Academic Press* 1980, 11–23.

[9] I. Gohberg and I. A. Feldman: *Convolution equations and projection methods for their solution*, Transl. of Math. Monographs, **41**, Amer. Math. Soc., 1974 [Russian original: Nauka, Moscow, 1971].

[10] I. Gohberg, S. Goldberg and M. A. Kaashoek: *Classes of Linear Operators*, Vol. 1 of 2, Birkhäuser, 1993.

[11] G. Heinig and F. Hellinger: The finite section method for Moore-Penrose inversion of Toeplitz operators, *Integral Equations and Operator Theory* **19** (1994), 419–446.

[12] V. G. Kurbatov: *Functional Differential Operators and Equations*, Kluwer, 1999.

[13] M. Lindner: *Infinite Matrices and their Finite Sections: An Introduction to the Limit Operator Method*, Frontiers in Mathematics, Birkhäuser, 2006.

[14] M. Lindner: Fredholmness and index of operators in the Wiener algebra are independent of the underlying space, *Operators and Matrices* **2** (2008), 297–306.

[15] M. Lindner: The finite section method and stable subsequences, *Applied Numerical Mathematics* **60** (2010), 501–512.

[16] G. Panova: Factorization of banded permutations, *Proc. Amer. Math. Soc., to appear*; *arXiv:1007.1760* (2010).

[17] V. S. Rabinovich, S. Roch and J. Roe: Fredholm indices of band-dominated operators, *Integral Equations Operator Theory* **49** (2004), 221–238.

[18] V. S. Rabinovich, S. Roch and B. Silbermann: *Limit Operators and Their Applications in Operator Theory*, Birkhäuser, 2004.

[19] S. Roch: *Finite sections of band-dominated operators*, Memoirs of the AMS, Vol. 191, Nr. 895, 2008.

[20] M. Seidel and B. Silbermann: Banach algebras of operator sequences, *to appear in Operators and Matrices*.

[21] G. Strang: Fast transforms: Banded matrices with banded inverses, *Proc. National Academy of Sciences* **107** (28) (2010), 12413–12416.

[22] G. Strang: Groups of banded matrices with banded inverses, *Proc. Amer. Math. Soc.* **139** (2011), 4255–4264.

[23] G. Strang: Banded matrices with banded inverses and A=LPU, *Proceedings Intl. Congress of Chinese Mathematicians: ICCM2010*, Amer. Math. Soc. and Intl. Press, 2012.

[24] G. Strang: The algebra of elimination, *submitted to SIAM Review* (2011).

[25] G. Strang and Tri Nguyen: The interplay of ranks of submatrices, *SIAM Review* **46** (2004), 637–646.

Authors:

Marko Lindner marko.lindner@mathematik.tu-chemnitz.de
TU Chemnitz
Fakultät Mathematik
D-09107 Chemnitz
GERMANY

Gilbert Strang gilstrang@gmail.com
MIT
Room 2-240
Cambridge MA 02139
USA

This figure "split.png" is available in "png" format from:

http://arxiv.org/ps/1112.0582v2