VI.B. The normal form of a polynomial matrix

We now turn to the algorithm associating to $M \in M_n(\mathsf{F}[\lambda])$ a matrix in normal form,

$$M \to nf(M),$$

using certain row and column operations, as promised in §VI.A. This is similar in spirit to the algorithm

$$A \to rref(A)$$

we have used throughout for $A \in M_n(\mathsf{F})$.

Why introduce column operations? Because we lose a row operation when we are dealing with polynomials: we can only *divide* a row by a scalar (not a polynomial), and this makes taking *rref* of matrices with polynomial entries a lost cause. (In particular, there is no "$rref(\lambda\mathbb{I} - A)$.") It also means that multiplying a row by λ is not invertible (of course, you can undo it on *that row*, but there isn't an elementary matrix for that "undo").

Here are the operations we shall permit:

VI.B.1. Definition. The **elementary row and column operations** on $M \in M_n(\mathsf{F}[\lambda])$ are the same as for $M_n(\mathsf{F})$, except you can't multiply a row or column by a polynomial. You may

(i) **replace:** add a *polynomial* multiple of a row (column) to a *different* row (column)
(ii) **swap** two rows (columns).
(iii) **scale:** multiply a row (column) by a scalar ($=$ an element of F)

The *elementary matrices* representing these operations are the same as in §I.C, except in $R_{ij}(b)$ (for operation (i)) the "b" is allowed to be a polynomial. As before, they are invertible (by the same formulas), with inverses in $M_n(\mathsf{F}[\lambda])$ — reflecting the invertibility of the operations. They have scalar determinants, as can be seen from their explicit form, or from the more general

VI.B.2. PROPOSITION. *A polynomial matrix* $M \in M_n(\mathsf{F}[\lambda])$ *has polynomial matrix inverse* $M^{-1} \in M_n(\mathsf{F}[\lambda])$ *if and only if* $\det(M)$ *is a nonzero scalar.*

PROOF. Suppose M has an inverse $M^{-1} \in M_n(\mathsf{F}[\lambda])$. *A priori* $\det(M)$ and $\det(M^{-1})$ are polynomials. But as $\det(M)\det(M^{-1}) = \det(MM^{-1}) = \det(\mathbb{I}_n) = 1$, and degrees of polynomials add under multiplication, the degree of $\det(M)$ must be 0.

Conversely, if $\det(M) \in \mathsf{F} \setminus \{0\}$, then $\frac{1}{\det(M)}\mathrm{adj}(M) \in M_n(\mathsf{F}[\lambda])$ provides an explicit inverse with polynomial entries. (See Exercise VI.A.5.) □

VI.B.3. DEFINITION. If one passes from M to N using elementary row and column operations, then M and N are called **equivalent.**

The Algorithm. Now let M be any nonzero matrix with entries in $\mathsf{F}[\lambda]$. Some notation:
• We say $g(\lambda) \mid M$ if it divides every entry of M
• $\ell(M) :=$ *lowest* degree of any nonzero (polynomial) entry of M (if M contains a nonzero scalar, say 3, then of course $\ell(M) = 0$)
• The "first" entry of M with a certain property will just mean the first you come upon if you read M like a page of a book.

Define an operation $(*)$ on M as follows: say $m = m_{ij}$ is the "first" nonzero entry of M with $\deg m = \ell(M)$ (it's an entry of least degree); perform row/column *swaps* to bring it to the $(1,1)$ position. Using the division algorithm, write all other entries in the first column as $q_i m + r_i$, where $\deg r_i < \deg m$; subtract $q_i \times$ (1^{st} row) from the i^{th} row (for $i = 2, \ldots, n$), to reduce these entries (in the 1^{st} column) to r_i. Do the same for the first row. This concludes the operation $(*)$.

The matrix now looks like

$$\begin{pmatrix} m & {}'r_2 & \cdots & {}'r_n \\ r_2 & & & \\ \vdots & & * & \\ r_n & & & \end{pmatrix},$$

with all r's of lower degree than m. If they are not all zero then we have reduced $\ell(M)$.

If we apply the algorithm $(*)$ repeatedly

$$M = M_0 \xrightarrow{(*)} M_1 \xrightarrow{(*)} M_2 \longrightarrow \cdots$$

then we reach a matrix of the form

$$\begin{pmatrix} g_1 & 0 & \leftrightarrow & 0 \\ 0 & & & \\ \updownarrow & & S_1 & \\ 0 & & & \end{pmatrix} =: (g_1, S_1)$$

in a finite number of steps, because we cannot continue to reduce

$$\ell(M_0) > \ell(M_1) > \ell(M_2) > \ldots$$

for very long. At the end of this process, divide g_1 by the coefficient of its highest power of λ to make it monic. (We still denote the result by g_1.) Call the whole sequence we have performed so far $(**)$.[7]

If $\deg g_1 > \ell(S_1)$ (i.e. g_1 is not of minimal degree in this matrix) then applying $(**)$ *again* to (g_1, S_1) will produce (g_2, S_2) with $\deg g_2 \leq \ell(S_1) < \deg g_1$. (The reason: if there is an entry $s \in S_1$ with lower degree than g_1, then $(**)$ will begin by swapping this element [or another, of degree $\leq \deg s$] to the $(1,1)$ position. From then on each application of $(*)$ within $(**)$ cannot increase the degree of the upper left-hand entry.) Since

$$\deg g_1 > \deg g_2 > \ldots$$

cannot continue forever, we eventually must reach (g_k, S_k) having $\deg g_k \leq \ell(S_k)$, i.e. such that g_k has the lowest degree in the matrix, excluding 0's.

However, g_k still may not *divide* all the entries of S_k, even if it is of lower degree. If $g_k \nmid S_k$ then let s be the first entry of S_k such that $g_k \nmid s$, and use the division algorithm to write $s = g_k q + r$, $\deg r <$

[7]Note that if S_1 is the zero matrix, we stop here, as the matrix is in normal form.

$\deg g_k$. Add the column containing s to the first column, changing the matrix to

$$\begin{pmatrix} g_k & 0 & \leftrightarrow & 0 \\ \vdots & & & \\ g_k q + r & & S_k & \\ \vdots & & & \end{pmatrix},$$

and subtract q times the first row from the row of s, to obtain

$$\begin{pmatrix} g_k & 0 & \leftrightarrow & 0 \\ \vdots & & & \\ r & & S_k & \\ \vdots & & & \end{pmatrix}.$$

Since $\deg r < \deg g_k$, applying $(**)$ produces (g_{k+1}, S_{k+1}) such that $\deg g_{k+1} \leq \deg r < \deg g_k$ (same argument as in the last paragraph). Continuing on as long as $g_i \nmid S_i$ we have once again

$$\deg g_k > \deg g_{k+1} > \deg g_{k+2} > \dots$$

and the process must terminate with (g_k, S_k) such that $g_k \mid S_k$.

We have produced from M, using a well-defined algorithm,

$$\begin{pmatrix} f_{(1)} & 0 & \leftrightarrow & 0 \\ 0 & & & \\ \updownarrow & & M^{(1)} & \\ 0 & & & \end{pmatrix}$$

with $f_{(1)}$ a polynomial in λ dividing the entries of $M^{(1)}$. Assuming $M^{(1)}$ is nonzero,[8] we perform the whole sequence of steps again on $M^{(1)}$ to get $f_{(2)}$, $M^{(2)}$ (with $f_{(2)} \mid M^{(2)}$!) *both* still divisible by $f_{(1)}$ (why?), and so on — until we have a diagonal matrix N with diagonal entries $f_{(1)}, f_{(2)}, \dots, f_{(n)}$. Thus N is a normal matrix (see §VI.A) and is equivalent to M; we write $N = nf(M)$.

[8] If you are putting $M = \lambda\mathbb{I} - A$ into normal form, this will always be the case. All of your $f_{(k)}$ will be nonzero, as their product has to be the characteristic polynomial (see below).

Uniqueness and Invariant factors. If R (resp. C) is the product of elementary matrices corresponding to the row (resp. column) operations performed in the computation, then the relationship is

$$R \cdot M \cdot C = nf(M) = N.$$

What if we used a different algorithm to put M in normal form, say

$$'R \cdot M \cdot 'C = 'N\,?$$

Then according to the following proposition, N and $'N$ are the same (getting *deja vu* yet?):

VI.B.4. PROPOSITION. *There is exactly one matrix in normal form "equivalent" to a given matrix M.*

So you don't have to do things in the rigid order specified above when finding $nf(M)$. The value of the rigid algorithm is that it has already proved the *existence* part of this proposition ("there *is* a *nf* matrix equivalent to M"). Before proving *uniqueness* we turn to the

PROOF OF THEOREM VI.A.10. Recall that

$$\delta_k(M) := \text{monic } gcd \text{ of determinants of } k \times k \text{ submatrices of } M.$$

These are *invariant* under row and column operations (ergo the terminology "invariant factors" for their ratios). This is because "replace" operations don't alter determinants,[9] while the scale and swap operations only change them by scalars (which are then wiped out by taking monic *gcd*, since this ignores scalar multiples). So if M and N are equivalent, then $\Delta_k(M) = \Delta_k(N)$.

Moreover, it is really easy to compute the invariant factors for

$$N = \begin{pmatrix} f_1(\lambda) & & 0 \\ & \ddots & \\ 0 & & f_n(\lambda) \end{pmatrix}.$$

[9]This is a wee bit disingenuous, since we are taking determinants of *submatrices*, and adding a polynomial multiple of column i to column j can *certainly* affect the determinant of any $k \times k$ submatrix meeting column j but not column i. The easy fix is given in Exercise (5).

Clearly, since all the f_i are monic, $\delta_n(N) = \det(N) = f_1 \cdot \cdots \cdot f_n$. Next, the *gcd* of the determinants of all $(n-1) \times (n-1)$ minors is simply $\delta_{n-1}(N) = f_1 \cdot \cdots \cdot f_{n-1}$. In general $\delta_k(N) = f_1 \cdot \cdots \cdot f_k$ and so $\Delta_k(N) = f_k$. We conclude that the diagonal entries of $nf(M)$ are the invariant factors $\Delta_k(M)$. $\square$

PROOF OF PROPOSITION VI.B.4. We only need to prove uniqueness: say N and $'N$ are matrices in normal form, both equivalent to M. Then N and $'N$ are equivalent, hence have the same invariant factors, and thus the same diagonal entries! That is, $N = N'$. $\square$

Notice that the invariant factors are playing here very much the same role as (the standard basis of) the row space did, back in §§II.C-II.D, in our proof that there was exactly one *rref* matrix row-equivalent to a given matrix in $M_n(\mathsf{F})$.

Now let $M = \lambda\mathbb{I} - A$. In general $nf(M)$ is going to look like

$$\begin{pmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & & h_1(\lambda) & & \\ & & & & \ddots & \\ 0 & & & & & h_r(\lambda) \end{pmatrix} = R \cdot (\lambda\mathbb{I} - A) \cdot C$$

where $h_r(\lambda) = m_A(\lambda)$. Taking determinants of both sides, since $\det R$ and $\det C$ are (nonzero) scalars, say $\det R \cdot \det C = k \in \mathsf{F}$, we have

$$h_1(\lambda) \cdot \cdots \cdot h_r(\lambda) = k \cdot \det(\lambda\mathbb{I} - A) = k \cdot f_A(\lambda).$$

Since the degree of the right-hand side $= n$,

$$\sum \deg(h_i(\lambda)) = n.$$

This is one way you can check you've done everything right.

VI.B.5. REMARK. Since f_A and $h_1, \ldots, h_r$ are all *monic* polynomials, $k = 1$. So we've proved *directly* that the characteristic polynomial of A is the product of the (diagonal) entries of $nf(\lambda\mathbb{I} - A)$, i.e. $\prod h_i(\lambda) = f_A(\lambda)$ (Corollary VI.A.11).

VI.B.6. EXAMPLE. We compute $nf(\lambda\mathbb{I} - A)$ for

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

by applying the algorithm to

$$\lambda\mathbb{I} - A = \begin{pmatrix} \lambda - 1 & -1 & -1 \\ -1 & \lambda - 1 & -1 \\ -1 & -1 & \lambda - 1 \end{pmatrix} \longrightarrow$$

$$\xrightarrow{\mathrm{I}} \begin{pmatrix} -1 & \lambda - 1 & -1 \\ \lambda - 1 & -1 & -1 \\ -1 & -1 & \lambda - 1 \end{pmatrix} \xrightarrow{\mathrm{II}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda^2 - 2\lambda & -\lambda \\ 0 & -\lambda & \lambda \end{pmatrix}$$

$$\xrightarrow{\mathrm{III}} \begin{pmatrix} 1 & & \\ & -\lambda & \lambda^2 - 2\lambda \\ & 0 & \lambda^2 - 3\lambda \end{pmatrix} \xrightarrow{\mathrm{IV}} \begin{pmatrix} 1 & & \\ & \lambda & \\ & & \lambda^2 - 3\lambda \end{pmatrix}.$$

Note that $m_A(\lambda) = \lambda^2 - 3\lambda$ is exactly the minimal polynomial we had found before, while $\lambda \cdot (\lambda^2 - 3\lambda) = \lambda^2(\lambda - 3) = f_A(\lambda)$.

An application. To get a quick sense of the depth of the results of this section, consider the striking

VI.B.7. COROLLARY. *Any matrix $M \in M_n(\mathsf{F}[\lambda])$ which is invertible in $M_n(\mathsf{F}[\lambda])$ is a product of elementary matrices.*

PROOF. By the algorithm, we have $RMC = N$, with N normal and R and C products of elementary matrices. By Proposition VI.B.2, the determinants of R, M, and C are nonzero scalars; hence so is that of N. But since N is diagonal, and degrees of polynomials add under multiplication, this forces its *entries* to be scalars. So N is a product of matrices of "scale" type, and $M = R^{-1}NC^{-1}$ is a product of elementary matrices. $\square$

A more general statement, which follows from VI.A.9-VI.A.10, is that $\det(M) \in \mathsf{F}[\lambda]$ is always the product of invariant factors (times a nonzero scalar). So if $\det(M)$ is not the zero polynomial, then the diagonal entries of $nf(M)$ are all nonzero, and vice versa. This justifies a claim made previously about the case $M = \lambda\mathbb{I} - A$, since then $\det(M) = f_A(\lambda)$ is never the zero polynomial.

Exercises

(1) What are the elementary matrices (of which R and C are products) involved in the application of the normal form algorithm in Example VI.B.6?

(2) For
$$A = \begin{pmatrix} 7 & 12 & -12 \\ -2 & -3 & 4 \\ 2 & 4 & -3 \end{pmatrix},$$
compute $nf(\lambda\mathbb{I} - A)$ via the algorithm above, and use it to determine m_A and f_A.

(3) (a) Determine the invariant factors of
$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & -2 & 0 & 1 \\ -2 & 0 & -1 & -2 \end{pmatrix}$$
by putting $\lambda\mathbb{I} - A$ in normal form.
(b) Repeat for the matrix given by replacing the -2 in the lower right-hand corner of A by 2.

(4) Prove that $A \in M_n(\mathbb{C})$ is diagonalizable if and only if m_A has no repeated roots, via the following steps:
(a) Show that similar matrices have the same minimal polynomial.
(b) Show that the minimal polynomial of a diagonal matrix is the product of the $(\lambda - \lambda_i)$ where $\{\lambda_i\}$ are the *distinct* eigenvalues. (With (a), this gives the "only if" part.)
Henceforth assume m_A has no repeated roots.
(c) Show that if A has eigenvalue λ, then $p(A)$ has eigenvalue $p(\lambda)$, for any polynomial p.
(d) Using part (c), prove that m_A is equal to the product $\prod(\lambda -$

λ_i) over *distinct* eigenvalues of A. [Hint: First show that this product divides m_A. Note that A is *not* assumed diagonalizable!]
(e) Now the nullity of $\prod(A - \lambda_i \mathbb{I})$ is $\leq$ the sum of the nullities of $(A - \lambda_i \mathbb{I})$ (why?). [Hint: see Exercise III.A.9(b).] The latter nullities are the geometric multiplicities of the eigenvalues. Using this, show that A is diagonalizable.

(5) Show that $\delta_k(M)$ is invariant under "replace" operations (type (i)), by arguing as follows. Let S [resp. S'] be the $k \times k$ submatrix of M obtained by removing the rows *other* than $i_1, \ldots, i_k$ [resp. $i_0, \ldots, i_{k-1}$] and columns *other* than $j_1, \ldots, j_k$. Given $f(\lambda) \in \mathsf{F}[\lambda]$, let $\tilde{M}$ be the result of adding $f(\lambda)$ times row i_0 to row i_k in M, and $\tilde{S}$ the $k \times k$ submatrix of $\tilde{M}$ obtained by omitting rows other than $i_1, \ldots, i_k$ and columns other than $j_1, \ldots, j_k$.
(a) Show that $\det \tilde{S} = \det(S) \pm f(\lambda) \det(S')$.
(b) Check that, for polynomials $g, g', f \in \mathsf{F}[\lambda]$, we have
$gcd(g, g') = gcd(g + fg', g')$.
(c) Prove that $\delta_k(\tilde{M}) = \delta_k(M)$.