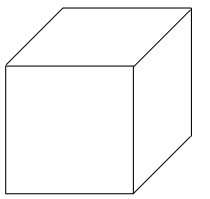
Chapter 8
Tensor Decomposition

8.1 Introduction

So far in this book we have considered linear algebra, where the main objects are vectors and matrices. These can be thought of as one-dimensional and two-dimensional arrays of data, respectively. For instance, in a term-document matrix, each element is associated with one term and one document. In many applications, data commonly are organized according to more than two categories. The corresponding mathematical objects are usually referred to as *tensors*, and the area of mathematics dealing with tensors is *multilinear algebra*. Here, for simplicity, we restrict ourselves to tensors $\mathcal{A} = (a_{ijk}) \in \mathbb{R}^{l\times m\times n}$ that are arrays of data with three subscripts; such a tensor can be illustrated symbolically as

$\mathcal{A} =$ 

Example 8.1. In the classification of handwritten digits, the *training set* is a collection of images, manually classified into 10 classes. Each such class is a set of digits of one kind, which can be considered as a tensor; see Figure 8.1. If each digit is represented as a 16×16 matrix of numbers representing gray scale, then a set of n digits can be organized as a tensor $\mathcal{A} \in \mathbb{R}^{16\times 16\times n}$. ∎

We will use the terminology of [60] and refer to a tensor $\mathcal{A} \in \mathbb{R}^{l\times m\times n}$ as a 3-mode array,[13] i.e., the different "dimensions" of the array are called *modes*. The *dimensions* of a tensor $\mathcal{A} \in \mathbb{R}^{l\times m\times n}$ are l, m, and n. In this terminology, a matrix is a 2-mode array.

[13] In some literature, the terminology 3-way and, in the general case, n-way, is used.

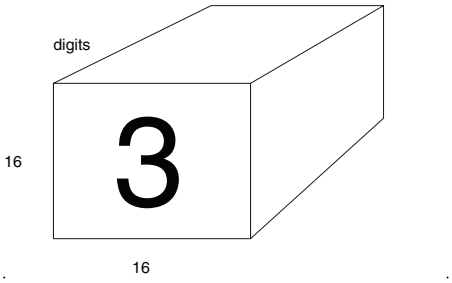


Figure 8.1. *The image of one digit is a* 16×16 *matrix, and a collection of digits is a tensor.*

In this chapter we present a generalization of the matrix SVD to 3-mode tensors, and then, in Chapter 14, we describe how it can be used for face recognition. The further generalization to n-mode tensors is easy and can be found, e.g., in [60]. In fact, the face recognition application requires 5-mode arrays.

The use of tensors in data analysis applications was pioneered by researchers in psychometrics and chemometrics in the 1960s; see, e.g., [91].

8.2 Basic Tensor Concepts

First define the *inner product* of two tensors:

$$\langle A, B \rangle = \sum_{i,j,k} a_{ijk} b_{ijk}. \tag{8.1}$$

The corresponding norm is

$$\| \mathcal{A} \|_F = \langle A, A \rangle^{1/2} = \left(\sum_{i,j,k} a_{ijk}^2 \right)^{1/2}. \tag{8.2}$$

If we specialize the definition to matrices (2-mode tensors), we see that this is equivalent to the matrix Frobenius norm; see Section 2.4.

Next we define *i-mode multiplication* of a tensor by a matrix. The 1-mode product of a tensor $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ by a matrix $U \in \mathbb{R}^{l_0 \times l}$, denoted by $\mathcal{A} \times_1 U$, is an $l_0 \times m \times n$ tensor in which the entries are given by

$$(\mathcal{A} \times_1 U)(j, i_2, i_3) = \sum_{k=1}^{l} u_{j,k} \, a_{k,i_2,i_3}. \tag{8.3}$$

For comparison, consider the matrix multiplication

$$A \times_1 U = UA, \quad (UA)(i,j) = \sum_{k=1}^{l} u_{i,k}\, a_{k,j}. \tag{8.4}$$

We recall from Section 2.2 that matrix multiplication is equivalent to multiplying each column in A by the matrix U. Comparing (8.3) and (8.4) we see that the corresponding property is true for tensor-matrix multiplication: in the 1-mode product, all column vectors in the 3-mode array are multiplied by the matrix U.

Similarly, 2-mode multiplication of a tensor by a matrix V

$$(\mathcal{A} \times_2 V)(i_1, j, i_3) = \sum_{k=1}^{l} v_{j,k}\, a_{i_1,k,i_3}$$

means that all row vectors of the tensor are multiplied by V. Note that 2-mode multiplication of a matrix by V is equivalent to matrix multiplication by V^T from the right,

$$A \times_2 V = AV^T;$$

3-mode multiplication is analogous.

It is sometimes convenient to *unfold* a tensor into a matrix. The unfolding of a tensor $\mathcal{A}$ along the three modes is defined (using (semi-)MATLAB notation; for a general definition,[14] see [60]) as

$$\begin{aligned}
\mathbb{R}^{l\times mn} \ni \text{unfold}_1(\mathcal{A}) := A_{(1)} &:= \begin{pmatrix} \mathcal{A}(:,1,:) & \mathcal{A}(:,2,:) & \dots & \mathcal{A}(:,m,:) \end{pmatrix}, \\
\mathbb{R}^{m\times ln} \ni \text{unfold}_2(\mathcal{A}) := A_{(2)} &:= \begin{pmatrix} \mathcal{A}(:,:,1)^T & \mathcal{A}(:,:,2)^T & \dots & \mathcal{A}(:,:,n)^T \end{pmatrix}, \\
\mathbb{R}^{n\times lm} \ni \text{unfold}_3(\mathcal{A}) := A_{(3)} &:= \begin{pmatrix} \mathcal{A}(1,:,:)^T & \mathcal{A}(2,:,:)^T & \dots & \mathcal{A}(l,:,:)^T \end{pmatrix}.
\end{aligned}$$

It is seen that the unfolding along mode i makes that mode the first mode of the matrix $A_{(i)}$, and the other modes are handled cyclically. For instance, row i of $A_{(j)}$ contains all the elements of $\mathcal{A}$, which have the jth index equal to i. The following is another way of putting it.

1. The column vectors of $\mathcal{A}$ are column vectors of $A_{(1)}$.
2. The row vectors of $\mathcal{A}$ are column vectors of $A_{(2)}$.
3. The 3-mode vectors of $\mathcal{A}$ are column vectors of $A_{(3)}$.

The 1-unfolding of $\mathcal{A}$ is equivalent to dividing the tensor into *slices* $\mathcal{A}(:,i,:)$ (which are matrices) and arranging the slices in a long matrix $A_{(1)}$.

[14] For the matrix case, $\text{unfold}_1(A) = A$, and $\text{unfold}_2(A) = A^T$.

Example 8.2. Let $\mathcal{B} \in \mathbb{R}^{3\times3\times3}$ be a tensor, defined in MATLAB as

```
B(:,:,1) =          B(:,:,2) =          B(:,:,3) =
     1     2     3        11    12    13        21    22    23
     4     5     6        14    15    16        24    25    26
     7     8     9        17    18    19        27    28    29
```

Then unfolding along the third mode gives

```
>> B3 = unfold(B,3)

b3 =     1     2     3     4     5     6     7     8     9
        11    12    13    14    15    16    17    18    19
        21    22    23    24    25    26    27    28    29
```

The inverse of the unfolding operation is written

$$\text{fold}_i(\text{unfold}_i(\mathcal{A})) = \mathcal{A}.$$

For the folding operation to be well defined, information about the target tensor must be supplied. In our somewhat informal presentation we suppress this.

Using the unfolding-folding operations, we can now formulate a matrix multiplication equivalent of i-mode tensor multiplication:

$$\mathcal{A} \times_i U = \text{fold}_i(U\ \text{unfold}_i(\mathcal{A})) = \text{fold}_i(UA_{(i)}). \tag{8.5}$$

It follows immediately from the definition that i-mode and j-mode multiplication commute if $i \neq j$:

$$(\mathcal{A} \times_i F) \times_j G = (\mathcal{A} \times_j G) \times_i F = \mathcal{A} \times_i F \times_j G.$$

Two i-mode multiplications satisfy the identity

$$(\mathcal{A} \times_i F) \times_i G = \mathcal{A} \times_i (GF).$$

This is easily proved using (8.5):

$$\begin{aligned}(\mathcal{A} \times_i F) \times_i G &= (\text{fold}_i(F(\text{unfold}_i(\mathcal{A})))) \times_i G \\ &= \text{fold}_i(G(\text{unfold}_i(\text{fold}_i(F(\text{unfold}_i(\mathcal{A})))))) \\ &= \text{fold}_i(GF\,\text{unfold}_i(\mathcal{A})) = \mathcal{A} \times_i (GF).\end{aligned}$$

8.3 A Tensor SVD

The matrix SVD can be generalized to tensors in different ways. We present one such generalization that is analogous to an *approximate principal component analysis*. It is often referred to as the higher order SVD (HOSVD)[15] [60].

[15] HOSVD is related to the Tucker model in psychometrics and chemometrics [98, 99].

Theorem 8.3 (HOSVD). *The tensor $\mathcal{A} \in \mathbb{R}^{l\times m\times n}$ can be written as*

$$\mathcal{A} = \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}, \tag{8.6}$$

where $U^{(1)} \in \mathbb{R}^{l\times l}$, $U^{(2)} \in \mathbb{R}^{m\times m}$, and $U^{(3)} \in \mathbb{R}^{n\times n}$ are orthogonal matrices. $\mathcal{S}$ is a tensor of the same dimensions as $\mathcal{A}$; it has the property of all-orthogonality*: any two slices of $\mathcal{S}$ are orthogonal in the sense of the scalar product* (8.1)*:*

$$\langle \mathcal{S}(i,:,:), \mathcal{S}(j,:,:)\rangle = \langle \mathcal{S}(:,i,:), \mathcal{S}(:,j,:)\rangle = \langle \mathcal{S}(:,:,i), \mathcal{S}(:,:,j)\rangle = 0$$

for $i \neq j$. The 1-mode singular values *are defined by*

$$\sigma_j^{(1)} = \|\mathcal{S}(i,:,:)\|_F, \qquad j = 1,\dots,l,$$

and they are ordered

$$\sigma_1^{(1)} \geq \sigma_2^{(1)} \geq \cdots \geq \sigma_l^{(1)}. \tag{8.7}$$

The singular values in other modes and their ordering are analogous.

Proof. We give only the recipe for computing the orthogonal factors and the tensor $\mathcal{S}$; for a full proof, see [60]. Compute the SVDs,

$$A_{(i)} = U^{(i)}\Sigma^{(i)}(V^{(i)})^T, \quad i = 1,2,3, \tag{8.8}$$

and put

$$\mathcal{S} = \mathcal{A} \times_1 (U^{(1)})^T \times_2 (U^{(2)})^T \times_3 (U^{(3)})^T.$$

It remains to show that the slices of $\mathcal{S}$ are orthogonal and that the i-mode singular values are decreasingly ordered. ∎

The all-orthogonal tensor $\mathcal{S}$ is usually referred to as the *core tensor*.
The HOSVD is visualized in Figure 8.2.

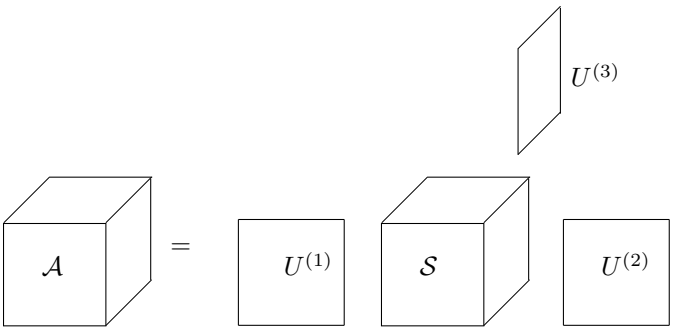


Figure 8.2. *Visualization of the HOSVD.*

Equation (8.6) can also be written

$$\mathcal{A}_{ijk} = \sum_{p=1}^{l}\sum_{q=1}^{m}\sum_{s=1}^{n} u_{ip}^{(1)} u_{jq}^{(2)} u_{ks}^{(3)} \mathcal{S}_{pqs},$$

which has the following interpretation: the element $\mathcal{S}_{pqs}$ reflects the variation by the combination of the singular vectors $u_p^{(1)}$, $u_q^{(2)}$, and $u_s^{(3)}$.

The computation of the HOSVD is straightforward and is implemented by the following MATLAB code, although somewhat inefficiently:[16]

```
function [U1,U2,U3,S,s1,s2,s3]=svd3(A);
% Compute the HOSVD of a 3-way tensor A

[U1,s1,v]=svd(unfold(A,1));
[U2,s2,v]=svd(unfold(A,2));
[U3,s3,v]=svd(unfold(A,3));

S=tmul(tmul(tmul(A,U1',1),U2',2),U3',3);
```

The function `tmul(A,X,i)` is assumed to multiply the tensor `A` by the matrix `X` in mode `i`, $\mathcal{A} \times_i X$.

Let V be orthogonal of the same dimension as U_i; then from the identities [60]

$$\mathcal{S} \times_i U^{(i)} = \mathcal{S} \times_i (U^{(i)} V V^T) = (\mathcal{S} \times_i V^T)(\times_i U^{(i)} V),$$

it may appear that the HOSVD is not unique. However, the property that the i-mode singular values are ordered is destroyed by such transformations. Thus, the HOSVD is essentially unique; the exception is when there are equal singular values along any mode. (This is the same type of nonuniqueness that occurs with the matrix SVD.)

In some applications it happens that the dimension of one mode is larger than the product of the dimensions of the other modes. Assume, for instance, that $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ with $l > mn$. Then it can be shown that the core tensor $\mathcal{S}$ satisfies

$$\mathcal{S}(i,:,:) = 0, \qquad i > mn,$$

and we can omit the zero part of the core and rewrite (8.6) as a *thin HOSVD*,

$$\mathcal{A} = \widehat{\mathcal{S}} \times_1 \widehat{U}^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}, \tag{8.9}$$

where $\widehat{\mathcal{S}} \in \mathbb{R}^{mn \times m \times n}$ and $\widehat{U}^{(1)} \in \mathbb{R}^{l \times mn}$.

8.4 Approximating a Tensor by HOSVD

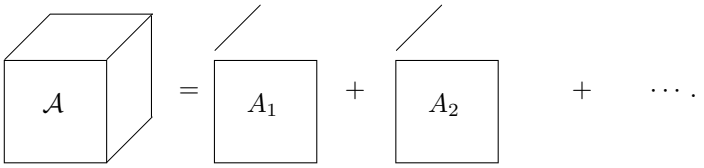
A matrix can be written in terms of the SVD as a sum of rank-1 terms; see (6.2). An analogous expansion of a tensor can be derived using the definition of tensor-matrix

[16] Exercise: In what sense is the computation inefficient?

multiplication: a tensor $\mathcal{A} \in \mathbb{R}^{l\times m\times n}$ can be expressed as a sum of matrices times singular vectors:

$$\mathcal{A} = \sum_{i=1}^{n} A_i \times_3 u_i^{(3)}, \qquad A_i = \mathcal{S}(:,:,i) \times_1 U^{(1)} \times_2 U^{(2)}, \tag{8.10}$$

where $u_i^{(3)}$ are column vectors in $U^{(3)}$. The A_i are to be identified as both matrices in $\mathbb{R}^{m\times n}$ and tensors in $\mathbb{R}^{m\times n\times 1}$. The expansion (8.10) is illustrated as



This expansion is analogous along the other modes.

It is easy to show that the A_i matrices are orthogonal in the sense of the scalar product (8.1):

$$\begin{aligned}\langle A_i, A_j\rangle &= \mathrm{tr}[U^{(2)}\mathcal{S}(:,:,i)^T(U^{(1)})^T U^{(1)}\mathcal{S}(:,:,j)(U^{(2)})^T] \\ &= \mathrm{tr}[\mathcal{S}(:,:,i)^T\mathcal{S}(:,:,j)] = 0.\end{aligned}$$

(Here we have identified the slices $\mathcal{S}(:,:,i)$ with matrices and used the identity $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.)

It is now seen that the expansion (8.10) can be interpreted as follows. Each slice along the third mode of the tensor $\mathcal{A}$ can be written (exactly) in terms of the orthogonal basis $(A_i)_{i=1}^{r_3}$, where r_3 is the number of positive 3-mode singular values of $\mathcal{A}$:

$$\mathcal{A}(:,:,j) = \sum_{i=1}^{r_3} z_i^{(j)} A_i, \tag{8.11}$$

where $z_i^{(j)}$ is the jth component of $u_i^{(3)}$. In addition, we have a simultaneous orthogonal factorization of the A_i,

$$A_i = \mathcal{S}(:,:,i) \times_1 U^{(1)} \times_2 U^{(2)},$$

which, due to the ordering (8.7) of all the j-mode singular values for different j, has the property that the "mass" of each $\mathcal{S}(:,:,i)$ is concentrated at the upper left corner.

We illustrate the HOSVD in the following example.

Example 8.4. Given 131 handwritten digits,[17] where each digit is a 16×16 matrix, we computed the HOSVD of the $16 \times 16 \times 131$ tensor. In Figure 8.3 we plot the

[17] From a U.S. Postal Service database, downloaded from the Web page of [47].

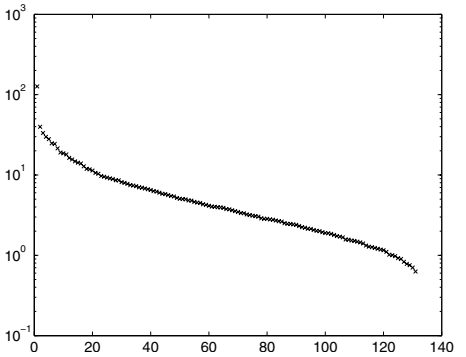


Figure 8.3. *The singular values in the digit (third) mode.*

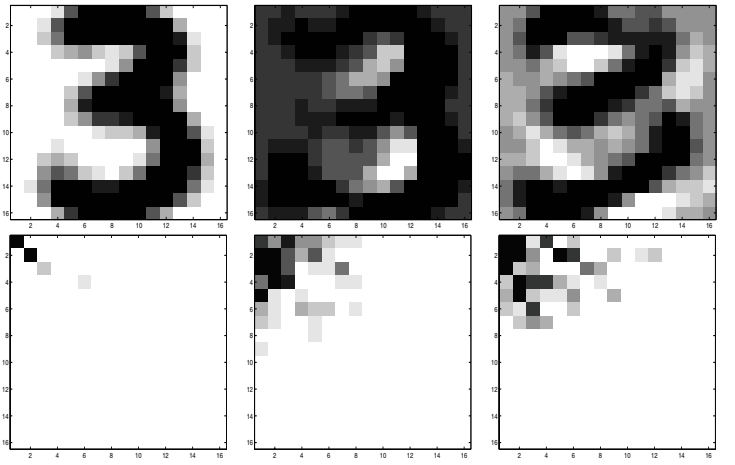


Figure 8.4. *The top row shows the three matrices A_1, A_2, and A_3, and the bottom row shows the three slices of the core tensor, $\mathcal{S}(:,:,1)$, $\mathcal{S}(:,:,2)$, and $\mathcal{S}(:,:,3)$ (absolute values of the components).*

singular values along the third mode (different digits); it is seen that quite a large percentage of the variation of the digits is accounted for by the first 20 singular values (note the logarithmic scale). In fact,

$$\frac{\sum_{i=1}^{20}(\sigma_i^{(3)})^2}{\sum_{i=1}^{131}(\sigma_i^{(3)})^2} \approx 0.91.$$

The first three matrices A_1, A_2, and A_3 are illustrated in Figure 8.4. It is seen that the first matrix looks like a mean value of different 3's; that is the dominating "direction" of the 131 digits, when considered as points in $\mathbb{R}^{256}$. The next two images represent the dominating directions of variation from the "mean value" among the different digits.

In the bottom row of Figure 8.4, we plot the absolute values of the three slices of the core tensor, $\mathcal{S}(:,:,1)$, $\mathcal{S}(:,:,2)$, and $\mathcal{S}(:,:,3)$. It is seen that the mass of these matrices is concentrated at the upper left corner. ∎

If we truncate the expansion (8.10),

$$\mathcal{A} = \sum_{i=1}^{k} A_i \times_3 u_i^{(3)}, \qquad A_i = \mathcal{S}(:,:,i) \times_1 U^{(1)} \times_2 U^{(2)},$$

for some k, then we have an approximation of the tensor (here in the third mode) in terms of an orthogonal basis. We saw in (8.11) that each 3-mode slice $\mathcal{A}(:,:,j)$ of $\mathcal{A}$ can be written as a linear combination of the orthogonal basis matrices A_j. In the classification of handwritten digits (cf. Chapter 10), one may want to compute the coordinates of an unknown digit in terms of the orthogonal basis. This is easily done due to the orthogonality of the basis.

Example 8.5. Let Z denote an unknown digit. For classification purposes we want to compute the coordinates of Z in terms of the basis of 3's from the previous example. This is done by solving the least squares problem

$$\min_z \left\| Z - \sum_j z_j A_j \right\|_F,$$

where the norm is the matrix Frobenius norm. Put

$$G(z) = \frac{1}{2} \left\| Z - \sum_j z_j A_j \right\|_F^2 = \frac{1}{2} \left\langle Z - \sum_j z_j A_j, Z - \sum_j z_j A_j \right\rangle.$$

Since the basis is orthogonal with respect to the scalar product,

$$\langle A_i, A_j \rangle = 0 \quad \text{for} \quad i \neq j,$$

we can rewrite

$$G(z) = \frac{1}{2}\langle Z, Z \rangle - \sum_j z_j \langle Z, A_j \rangle + \frac{1}{2} \sum_j z_j^2 \langle A_j, A_j \rangle.$$

To find the minimum, we compute the partial derivatives with respect to the z_j and put them equal to zero,

$$\frac{\partial G}{\partial z_j} = -\langle Z, A_j \rangle + z_j \langle A_j, A_j \rangle = 0,$$

which gives the solution of the least squares problem as

$$z_j = \frac{\langle Z, A_j \rangle}{\langle A_j, A_j \rangle}, \quad j = 1, 2, \ldots. \qquad ∎$$

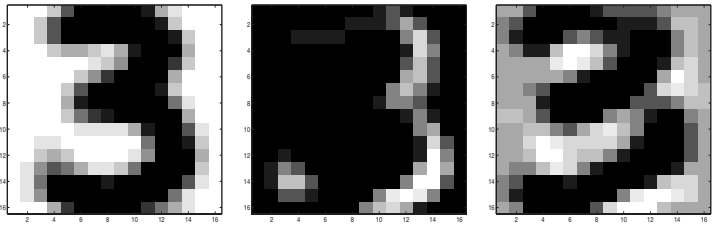
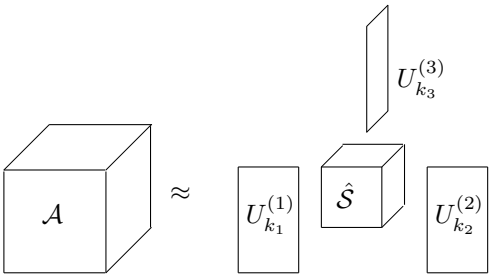


Figure 8.5. *Compressed basis matrices of handwritten* 3*'s.*

Because the mass of the core tensor is concentrated for small values of the three indices, it is possible to perform a simultaneous data compression in all three modes by the HOSVD. Here we assume that we compress to k_i columns in mode i. Let $U_{k_i}^{(i)} = U^{(i)}(:, 1 : k_i)$ and $\hat{\mathcal{S}} = \mathcal{S}(1 : k_1, 1 : k_2, 1 : k_3)$. Then consider the approximation

$$\mathcal{A} \approx \hat{\mathcal{A}} = \hat{\mathcal{S}} \times_1 U_{k_1}^{(i)} \times_2 U_{k_2}^{(2)} \times_3 U_{k_3}^{(3)}.$$

We illustrate this as follows:



Example 8.6. We compressed the basis matrices A_j of 3's from Example 8.4. In Figure 8.5 we illustrate the compressed basis matrices

$$\hat{A}_j = \mathcal{S}(1 : 8, 1 : 8, j) \times_1 U_8^{(1)} \times_2 U_8^{(2)}.$$

See the corresponding full-basis matrices in Figure 8.4. Note that the new basis matrices $\hat{A}_j$ are no longer orthogonal. ∎