Introduction to Tensors and their Decompositions

Aaron Geelon So

December 3, 2018

We often model parts of our world with linear spaces: objects are vectors inside that space, and relevant properties and transformations are linear maps.

For example, we might define a vector space U of movie preferences; you might be some $u \in U$. Then, there are certain linear maps $f, g : U \to \mathbb{R}$ that capture your preferences. Maybe $f(u)$ indicates how much you like fantasy and $g(u)$ indicate how much you like gore in your movies.

Now, what happens when we want to bring two parts of our world into interaction? How should we define the vector space of, say, movie preferences for pairs of people? One approach we can naturally take is to look at the vector space $U \times U$. Suppose that there is once again a linear map that indicates group preference for fantasy. What form does that map take? I claim the following:

Claim 1. *Let U, V be finite-dimensional vector spaces. Let $f : U \times V \to W$ be a linear map. Then, there exist linear maps $f_1 : U \to W$ and $f_2 : V \to W$ such that:*

$$f(u, v) = f_1(u) + f_2(v).$$

To highlight this, we could write $U \oplus V$ instead of $U \times V$.

But this means that you and your friend's movie preferences don't really interact at all: the group preference just becomes some weighted sum of your individual preferences. If we want a model that is able to describe how putting the two of you together makes you both much more excited about gory fantasy movies, we need to broaden our scope beyond linear maps.

In particular, we are interested in *multilinear* interactions. For example, it might be the case that your group preferences are modeled by:

$$f(u, v) = f_1(u) \cdot f_2(v),$$

where f_1 and f_2 are linear. In this model, it will turn out that we can consider f as a linear map $f : U \otimes U \to \mathbb{R}$ where the space $U \otimes U$ is the vector space of movie preferences for pairs constructed in a very specific way from the movie preferences of the individuals. This construction is called the *tensor product*.

By the way, physicists often describe physical systems through *state spaces*. For example, U might be the space of all possible states for one particle, and V for another particle. When we bring those two particles together, we form a larger state space. In quantum mechanics, in order to describe entanglement, the new state space is described by the tensor product $U \otimes V$.

But to see why we care about multilinear interaction, let's take a look at a simple topic model problem for motivation. The idea is that we will have N words and k topics. The words in a document are conditionally independent on the topic; that is, given a topic t, a document of length ℓ is generated by independently sampling from the probability distribution on the N words, $\mu^{(t)}$. We also assume that each topic has probability β_t of being chosen.

Let X_1 be a document of length $\ell = 1$. Then, it follows that total probability distribution P on X_1 is the vector:

$$P = \sum_{t=1}^{k} \beta_t \mu^{(t)}.$$

What about documents (X_1, X_2) of length $\ell = 2$? By conditional independence, the probability P_{ij} that $(X_1, X_2) = (i, j)$, the ith followed by the jth word in our dictionary, is equal to:

$$P_{ij} = \sum_{t=1}^{k} \beta_t \mu_i^{(t)} \mu_j^{(t)}.$$

Notice that the first and second word have multilinear interaction; i.e. correlation is multilinear. Furthermore, from this form, it follows that we can represent P as a matrix:

$$P = \sum_{t=1}^{k} \beta_t \mu^{(t)} \mu^{(t)\mathsf{T}}.$$

Our task is to estimate the parameters $\beta_t, \mu^{(t)}$ from P. Even better, suppose we have a noisy version of P because we estimated it from a finite sample of documents. Can we recover $\beta_t, \mu^{(t)}$, even approximately? It turns out that we cannot, as a result of the *rotation problem*.

Suppose that P has the above decomposition. If we let U be any orthogonal matrix, then P also has the decomposition:

$$P = \sum_{t=1}^{k} \beta_t \big(\mu^{(t)} U\big)\big(\mu^{(t)} U\big)^{\mathsf{T}}.$$

And so, there is no unique set of parameters that generate P. In this case, we say that $\beta_t, \mu^{(t)}$ are not *identifiable*. It follows that any parameters we do recover are not necessarily meaningful, in that they don't correspond to 'reality'.

But if we look at three-way interactions (i.e. documents of at least length three), these parameters become identifiable. The additional interactions make the object P_{ijk} much more 'rigid'; this object will be a order-3 tensor. The short answer to what a tensor is: a multiway array. However, this coordinate description of a tensor makes it conceptually difficult to work with.

In the rest of the lecture, we'll introduce some multilinear algebra, tensors, and tensor decompositions. We'll also get to see how we can use tensors in some estimation problems. Tensors often get a reputation of being mysterious; the difficulty is that we often like to define our mathematical objects constructively. However, tensors are most easily understood not by how they are constructed, but by how they behave. In order to take a more functional approach to understanding tensors, we first need to understand dual spaces.

1 Dual spaces

At this point, we have two main actors: objects and functions. Usually, we tend to define functions with respect to the objects. For example, given a vector space (i.e. a set with some very specific structure), we can define linear maps (i.e. functions on that set respecting that structure). But, here, we'll see that we can define the objects with respect to how they behave under a collection of maps.

Definition 2. *Let V be a vector space. A* linear functional *or* covector *is a linear map $f : V \to \mathbb{R}$. The space of all linear functionals is called the* dual space *of V, denoted by V^*. This space is made into a vector space where scalar multiplication and vector addition are defined pointwise. That is, $\lambda f + \mu g$ is defined to be the linear map:*

$$x \mapsto \lambda f(x) + \mu g(x).$$

Proposition 3. *Let V be finite-dimensional. Then, V and V^* are isomorphic, $V \cong V^*$.*

Proof as exercise. □

This is a fact you're already familiar with when working with matrices. On $\mathbb{R}^n$, a linear functional f is just a linear map from n dimensional space to 1 dimensional space. So, we can represent f by a $1 \times n$ matrix; namely, a row vector. That is:

- vectors in $\mathbb{R}^n$ are column vectors
- covectors of $\mathbb{R}^n$ are row vectors
- covectors act on vectors through usual matrix multiplication:

$$f(x) = \begin{bmatrix} f_1 & \cdots & f_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} f_i x_i.$$

It follows that transpose $^\mathsf{T} : \mathbb{R}^n \xrightarrow{\cong} (\mathbb{R}^n)^*$ provides a linear isomorphism between the two spaces.

On the other hand, instead of saying that f acts on x, we could have said x acts on f, writing:

$$x(f) = \begin{bmatrix} f_1 & \cdots & f_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} f_i x_i.$$

This shows that x can be seen as a linear functional on its dual space. In this way, we obtain a natural inclusion $\mathbb{R}^n \hookrightarrow (\mathbb{R}^n)^{**}$.

Proposition 4. *Let V be a vector space, not necessarily finite-dimensional. There is a natural inclusion from V into its double dual V^{**}. Additionally, when V is finite-dimensional, V and V^{**} are naturally isomorphic.*

Proof as exercise. □

As an important payoff here for a finite-dimensional vector space V, we don't need to define the elements of V constructively (e.g. giving them a representation through $\mathbb{R}^n$). We can define them with respect to how they behave under linear maps: from V^*, obtain $V \cong V^{**}$.

Furthermore, this symmetrizes how we think about V and V^*; elements of V^* are functions on V just as elements of V are functions on V^*. And so, through duality, we really can think of vectors as objects or as functions; we'll switch between these two perspectives to suit our purposes.

2 Bilinear maps and matrices

Definition 5. *Let U, V, W be vector spaces. A* bilinear map *$f : U \times V \to W$ is a map that is linear in each entry:*

$$f(\lambda_1 u_1 + \lambda_2 u_2, v) = \lambda_1 f(u_1, v) + \lambda_2 f(u_2, v)$$
$$f(u, \mu_1 v_1 + \mu_2 v_2) = \mu_1 f(u, v_1) + \mu_2 f(u, v_2).$$

Example 6. *The inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ on $\mathbb{R}^n$ is a bilinear map. In particular, real multiplication $(a, b) \mapsto ab$ is bilinear. As a nonexample, addition $+ : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is not bilinear.*

Example 7. *Let V be a vector space, and define* $\mathrm{ev} : V^* \times V \to \mathbb{R}$ *to be the evaluation map,*

$$\mathrm{ev}(f, v) = f(v).$$

The evaluation map is bilinear.

Example 8. *Let $M \in \mathbb{R}^{m \times n}$ be a matrix. The following map $\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is bilinear:*

$$(v, w) \mapsto w^\mathsf{T} M v.$$

Proposition 9. *Let V, W be n- and m-dimensional vector spaces with fixed bases. Let $f : V \times W \to \mathbb{R}$ be a bilinear map. Then, there exists a unique $(m \times n)$-dimensional matrix M_f such that for all $v \in V$, $w \in W$,*

$$f(v, w) = [w]^\mathsf{T} M_f [v] = \mathrm{Tr}(M_f [v][w]^\mathsf{T}),$$

where we let $[v], [w]$ be the coordinate representation of v and w, repsectively in their bases.

Before proving this proposition, it will help to define some notation:

Definition 10. *Let U, V, W be vector spaces. We denote the vector space of linear maps from V to W by* $\mathrm{Hom}(V; W)$.[1] *Likewise, the vector space of bilinear maps from $U \times V$ to W is denoted* $\mathrm{Bilin}(U, V; W)$.

Exercise 11. *What is* $\mathrm{Hom}(V; \mathbb{R})$*?*

Corollary 12. *The following vector spaces are isomorphic:*

$$\mathbb{R}^{m \times n} \cong \mathrm{Bilin}(\mathbb{R}^n, \mathbb{R}^m; \mathbb{R}).$$

We'll leave the proof of Corollary 12 as an exercise.

Proof of Proposition 9. Let $f : V \times W \to \mathbb{R}$ be a bilinear map. Borrowing from type theory, we can equivalently consider the curried version of f, $\texttt{curry}\ f : V \to W \to \mathbb{R}$.

This just means that we sequentially apply f to $v \in V$ to get the map $f(v, \cdot)$. Then, applying $f(v, \cdot)$ to $w \in W$ yields $f(v, w)$:

$$v \overset{\texttt{curry}\ f}{\longmapsto} f(v, \cdot) \quad \text{followed by} \quad w \overset{f(v,\cdot)}{\longmapsto} f(v, w).$$

Because f is bilinear, the latter map $f(v, \cdot) : W \to \mathbb{R}$ is actually linear; i.e., $f(v, \cdot) \in W^*$. In other words, the 'type signature' of $\texttt{curry}\ f$ is:

$$\texttt{curry}\ f : V \to W^*.$$

And again, because f is bilinear, for all $w \in W$, we also have:

$$f(\lambda_1 v_1 + \lambda_2 v_2, w) = \lambda_1 f(v_1, w) + \lambda_2 f(v_2, w).$$

This implies that $f(\lambda_1 v_1 + \lambda_2 v_2, \cdot) = \lambda_1 f(v_1, \cdot) + \lambda_2 f(v_2, \cdot)$; $\texttt{curry}\ f$ is a *linear* map from V to W^*.

This essentially completes our proof; given fixed bases on V and W, we know that every linear functional h on W has a unique m-dimensional coordinate representation $[h]$ such that for all $w \in W$,

$$h(w) = [w]^\mathsf{T} [h].$$

And given the linear map T between vector spaces V and W^*, there is a unique matrix M such that:

$$[Tv] = M[v].$$

It follows that there is a unique matrix M_f such that $f(v, w) = [w]^\mathsf{T} M_f [v]$. □

[1] The notation $\mathrm{Hom}(V; W)$ is short for the set of *homomorphisms* from V to W. When working with vector spaces, homomorphisms are precisely linear maps.

Note that if we additionally check a few conditions, our proof of the proposition yields:

Corollary 13. *Let V, W as before. Then* $\texttt{curry} : \text{Bilin}(V, W; \mathbb{R}) \xrightarrow{\cong} \text{Hom}(V, W^*)$.

Proof as exercise. □

3 Tensor product space

So far, we've learned that all bilinear maps $f : V \times W \to \mathbb{R}$ can be represented by a matrix M, where:

$$f(v, w) = \sum_{i \in [n], j \in [m]} M_{ij} v_i w_j. \tag{1}$$

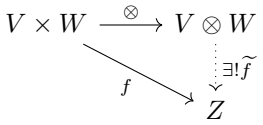
Looking at this formula from the context of the movie preference example from earlier, we see how bilinear maps enable us to capture more interaction between two viewers. Unfortunately, it seems much harder to work with bilinear maps than it is with linear maps; for example, $\text{Bilin}(V, W; \mathbb{R})$ is no longer dual to $V \times W$, unlike $\text{Hom}(V; \mathbb{R})$ and V.

But is there a way we can 'linearize' the bilinear maps? That is to say, although a bilinear map $f : V \times W \to \mathbb{R}$ is not linear over $V \times W$, can we somehow lift f to a different space on which the same computation becomes linear?

What we desire is a linear space, which we'll call $V \otimes W$, such that for all bilinear maps $f \in \text{Bilin}(V, W; Z)$, there is a unique linear map $\widetilde{f} \in \text{Hom}(V \otimes W; Z)$ such that:

$$f(v, w) = \widetilde{f}(v \otimes w).$$

In this case, we say that the following diagram commutes,



since either path of computation will yield the same result (uniqueness ensures that $V \otimes W$ is the 'smallest' space for which this property holds).

Proposition 14. *Such a pair $(\otimes, V \otimes W)$ exists.*

Proof. It is sufficient to prove that the above diagram commutes for $Z = \mathbb{R}$ because every k-dimensional vector space can be decomposed into:

$$Z \cong \mathbb{R} \overbrace{\oplus \cdots \oplus}^{k \text{ times}} \mathbb{R}.$$

Then, just consider $\pi_i \circ \widetilde{f} : V \otimes W \to \mathbb{R}$, where π_i is the projection onto the ith coordinate. Because for each coordinate, $\pi_i \circ \widetilde{f}$ is unique, $\widetilde{f}$ is unique too.

To see existence, let's take a look at Equation 1. Since every real bilinear map is of this form, it follows that by creating a new vector $v \otimes w \equiv \otimes(v, w)$ such that $(v \otimes w)_{ij} = v_i w_j$, then:

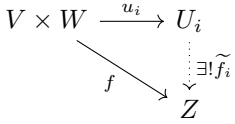
$$\begin{aligned} f(v, w) &= \sum_{i \in [n], j \in [m]} M_{ij} v_i w_j \\ &= \sum_{i \in [n], j \in [m]} M_{ij} (v \otimes w)_{ij} = \widetilde{f}(v \otimes w). \end{aligned}$$

So, there exists a linear $\widetilde{f} : V \otimes W \to \mathbb{R}$ satisfying the commutative diagram.

To see uniqueness, we show that $\sim : \text{Bilin}(V, W; \mathbb{R}) \xrightarrow{\cong} \text{Hom}(V \otimes W; \mathbb{R})$ is a linear isomorphism. It's not hard to prove linearity. Following that, notice that the dimensions of both spaces are nm. Furthermore, the map $\sim$ is a surjection, so it is an isomorphism. □

In fact, an even faster proof would be to just set $V \otimes W$ to $\text{Bilin}(V, W; \mathbb{R})^*$ and $\otimes(v, w)$ to be the evaluation at (v, w), $\text{ev}_{(v,w)}$. Then, uniqueness follows from $\text{Bilin}(V, W; \mathbb{R}) \cong \text{Bilin}(V, W; \mathbb{R})^{**} \equiv (V \otimes W)^*$. At this point, the tensor product space of V and W is just any space that satisfies the above commutative diagram. The following exercise shows that the pair $(\otimes, V \otimes W)$ is unique up to unique isomorphism:

Exercise 15. *Let V, W, Z be vector spaces. Suppose that there are there are two pairs (u_1, U_1) and (u_2, U_2), with $u_i : V \times W \to U_i$ bilinear such that for all $f \in \text{Bilin}(V, W; Z)$, there exists a unique $\widetilde{f}_i \in \text{Hom}(U_i; Z)$ so that the following diagram commutes:*



Prove that U_1 and U_2 are uniquely isomorphic*, in the sense that there is a unique isomorphism $t : U_1 \to U_2$ such that $u_2 = t \circ u_1$.*

This shows that any pair (u, U) satisfying the tensor product universal property can be uniquely identifed with $(\otimes, V \otimes W)$. Thus, we can let $\otimes : V \times W \to V \otimes W$ be the canonical tensor product.

To summarize, we have a powerful conceptual understanding of the tensor product space: $V \otimes W$ is the space on which bilinear maps become linear. We call objects in $V \otimes W$ *tensors*. But when we want to perform computations with tensors, it suffices to set a basis on V and W, and define the tensor product of $v \otimes w$ as the two-dimensional array where

$$(v \otimes w)_{ij} = v_i w_j = \left(v w^{\mathsf{T}}\right)_{ij}.$$

In particular, if $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ are bases on V and W, respectively, then the collection $\{a_i \otimes b_j\}$ form a basis for $V \otimes W$.

4 Dual tensor product space

Definition 16. *Let $a_1, \ldots, a_n \in V$ form a basis on V. The* dual basis *is the set of vectors $a^1, \ldots, a^n \in V^*$ such that:*

$$a^i(a_j) = \delta_{ij}.$$

(Here, δ_{ij} is the Kronecker delta, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise).

Proposition 17. *Let $a_1, \ldots, a_n \in V$ and $b_1, \ldots, b_m \in W$ form a basis on V and W. Let $a^1, \ldots a^n, b^1, \ldots, b^m$ be their dual bases. Then, the collection $\{a^i \otimes b^j\}$ forms a basis on $(V \otimes W)^*$. In particular, this implies that*

$$(V \otimes W)^* \cong (V^*) \otimes (W^*).$$

Proof as exercise. □

But from before, $(V \otimes W)^* \cong \text{Bilin}(V, W; \mathbb{R})$. In other words, the collection of real bilinear maps on $V \times W$ is isomorphic to the tensor space $(V^*) \otimes (W^*)$. Once again, we obtain a similar duality principle as with V and $\text{Hom}(V; \mathbb{R})$. This lets us switch between the perspectives that an object in $V \otimes W$ can be viewed as both a tensor and a bilinear map.

Furthermore, this tells us that given a bilinear map $f : V \times W \to \mathbb{R}$, the corresponding linear map $\widetilde{f} : V \otimes W \to \mathbb{R}$ is of the form:

$$\widetilde{f} = \sum_j g^j \otimes h^j,$$

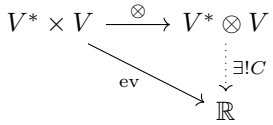
where $g^j \in V^*$ and $h^j \in W^*$. Furthermore, the way that $g \otimes h \in V^* \otimes W^*$ acts on $v \otimes w \in V \otimes W$ is:

$$\mathrm{ev}(g \otimes h, v \otimes w) = g(v) \cdot h(w).$$

Extending linearly, we have that in general:

$$\mathrm{ev}\left(\sum_j g^j \otimes h^j, \sum_i v_i \otimes w_i\right) = \sum_{i,j} g_j(v_i) \cdot h_j(w_i).$$

Exercise 18. *We stated earlier that the evaluation map* $\mathrm{ev} : V^* \times V \to \mathbb{R}$ *is bilinear. This implies that there exists a unique map* $C : V^* \otimes V \to \mathbb{R}$ *such that the following diagram commutes:*



In other words, C *is the unique map from* $V^* \otimes V$ *to* $\mathbb{R}$ *such that for all* $f \in V^*$ *and* $v \in V$*:*

$$\mathrm{ev}(f, v) = f(v) = C(f \otimes v).$$

Fixing a basis and corresponding dual basis on V *and* V^**, let their coordinate representation be:*

$$[f] \equiv \begin{bmatrix} f_1 & \cdots & f_n \end{bmatrix} \qquad [v] \equiv \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \qquad [f \otimes v] \equiv [v][f],$$

where $[v][f]$ *is the outer product of* $[v]$ *and* $[f]$*. Show that:*

$$\mathrm{ev}(f, v) = \mathrm{Tr}\left([v][f]\right).$$

This map C is called the *contraction map*, and this previous exercise shows why it is considered the generalization of the trace from matrix algebra. And of course, just as C is linear here, the trace is linear over the space of matrices $\mathbb{R}^{n \times n}$.

Exercise 19. *Let* $T \in (V \otimes W)^* \cong V^* \otimes W^*$ *and* $S \in V \otimes W$ *be:*

$$T = \sum_{i=1}^{t} f^{(i)} \otimes g^{(i)} \qquad S = \sum_{j=1}^{s} v^{(j)} \otimes w^{(j)}.$$

Fix a basis on V *and* W*, and let* $[f^{(i)}]$ *and* $[g^{(i)}]$ *be row vectors in that basis while* $[v^{(j)}]$ *and* $[w^{(j)}]$ *are column vectors in that basis. Let* $[S] = \sum_i [g^{(j)}]^\mathsf{T}[f^{(j)}]$ *and* $[T] = \sum_j [v^{(j)}][w^{(j)}]^\mathsf{T}$ *be their matrix representations. Show that the following is true:*

$$\mathrm{ev}(T, S) := \sum_{i=1}^{t}\sum_{j=1}^{s} f^{(i)}(v^{(j)}) \cdot g^{(i)}(w^{(j)}) = \sum_{j=1}^{s} \left(w^{(j)}\right)^\mathsf{T}[T]\left(v^j\right) = \mathrm{Tr}\big([S][T]\big).$$

Note that trace is indeed a linear operator on the space of symmetric matrices.

5 Multilinear maps and tensors

To generalize, we define multilinear functions analogously:

Definition 20. *Let $V_1, \dots, V_k$ and Z be vector spaces. A* multilinear function *$f : V_1 \times \cdots V_k \to Z$ is a function that is linear in each entry:*

$$f(\lambda_1 u_1 + \cdots + \lambda_n u_n, v_2, \dots, v_n) = \sum_{i=1}^{k} \lambda_i f(u_i, v_2, \dots, v_n)$$

$$\vdots$$

$$f(v_1, \dots, v_{k-1}, \lambda_1 u_1 + \cdots + \lambda_n u_n) = \sum_{i=1}^{k} \lambda_i f(v_1, \dots, v_{k-1}, u_i)$$

Quite analogously, the tensor product of the vector spaces $V_1, \dots, V_k$ is the space $V_1 \otimes \cdots \otimes V_k$ that linearizes multilinear functions $f : V_1 \times \cdots \times V_k \to Z$. That is, there is a unique linear map $\widetilde{f} : V_1 \otimes \cdots \otimes V_k \to Z$ such that:

$$f(v_1, \dots, v_k) = \widetilde{f}(v_1 \otimes \cdots \otimes v_k).$$

Furthermore, if we want to consider the tensor $v_1 \otimes \cdots \otimes v_k$ in terms of a fixed bases on all the vector spaces, we obtain that for each f, there exists a unique T such that:

$$f(v_1, \dots, v_k) = \sum_{i_1, \dots, i_k} T_{i_1 \cdots i_k} v_{1, i_1} \cdots v_{k, i_k} = \sum_{i_1, \dots, i_k} T_{i_1 \cdots i_k} (v_1 \otimes \cdots \otimes v_k)_{i_1 \cdots i_k},$$

where we let $v_{i,j}$ denote the jth coordinate of the ith vector v_i. Thus, we can think of $v_1 \otimes \cdots \otimes v_k$ as a k-dimensional array.

Everything actually carries through from the bilinear case, because we can inductively apply our analysis. For example, if $f : U \times V \times W \to Z$ is trilinear, then fixing $w \in W$, $f_w : U \times V \to Z$ is bilinear. So, we obtain $f_w : U \otimes V \to Z$ that is parametrized linearly over W. Thus, $f : (U \otimes V) \times W \to Z$ is bilinear. Then, again, we get $f : (U \otimes V) \otimes W \to Z$ is linear.

A similar argument shows that $f : U \otimes (V \otimes W) \to Z$ is linear. These two spaces are uniquely isomorphic:

$$(U \otimes V) \otimes W \cong U \otimes (V \otimes W),$$

so we can define $U \otimes V \otimes W$ to be the canonical tensor product; we obtain that $\otimes$ is associative. This approach also shows that $(U \otimes V \otimes W)^* \cong U^* \otimes V^* \otimes W^*$.

Definition 21. *Let $V_1, \dots, V_k$ be vector spaces. We say that a tensor in $V_1 \otimes \cdots \otimes V_k$ is an* order-k tensor. *A* pure tensor *is one of the form:*

$$v^{(1)} \otimes \cdots \otimes v^{(k)}.$$

In general, tensors are linear combinations of pure tensors; the rank of a tensor T is the minimum r such that it can be decomposed as a sum of pure tensors:

$$T = \sum_{i=1}^{r} v_i^{(1)} \otimes \cdots \otimes v_i^{(k)},$$

where $v_i^{(j)} \in V_j$. Note that the rank of order-2 tensors coincide with the rank of the corresponding matrix.

Because general tensors are just linear combinations of pure tensors, it often suffices to analyze how maps act on pure tensors. For example, we now define the contraction along the i, j axes of a tensor for pure tensors; this just extends linearly to general tensors:

Definition 22. *Let $V_1, \dots, V_i, \dots, V_j \dots, V_k$ be vector spaces where $V_j = V_i^*$. The* contraction *along the i, j axes is the linear map $C_i^j : V_1 \otimes \cdots \otimes V_k \to V_1 \otimes \cdots \hat{V}_i \cdots \hat{V}_j \cdots \otimes V_k$*

$$v^{(1)} \otimes \cdots \otimes v^{(k)} \mapsto \mathrm{ev}(v^{(i)}, v^{(j)}) \cdot v^{(1)} \otimes \cdots \hat{v}^{(i)} \cdots \hat{v}^{(j)} \cdots \otimes v^{(k)},$$

where the notation $\hat{v}^{(i)}$ and $\hat{V}^{(i)}$ means that the ith component is removed.

6 Tensor products over inner product spaces

- Restrict setting to $V = \mathbb{R}^n$. With the addition of inner product on $\mathbb{R}^n$ (i.e. canonical basis), there is a natural identification of $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$.
- Inner product on $\mathbb{R}^n$ induces inner product on $(\mathbb{R}^n)^{\otimes d}$
- Give connection to kernel functions:

Exercise 23. *Let $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the homogeneous quadratic kernel $K(x, y) = \langle x, y \rangle^2$. Show that there exists a feature map $\phi : \mathbb{R}^n \to \mathbb{R}^n \otimes \mathbb{R}^n$ such that:*

$$K(x, y) = \langle \phi(x), \phi(y) \rangle.$$

More generally, show that for the polynomial kernel $K(x, y) = (\langle x, y \rangle + c)^d$, there is a feature map

$$\phi : \mathbb{R}^n \to (\mathbb{R}^n)^{\otimes d} \oplus (\mathbb{R}^n)^{\otimes d-1} \oplus \cdots \oplus \mathbb{R}^n \oplus \mathbb{R}.$$

7 Tensor decompositions

Our goal of tensor decomposition is to take a tensor $T \in (\mathbb{R}^n)^{\otimes d}$ and write it as a sum:

$$T = \sum_{i=1}^{r} v_i^{(1)} \otimes \cdots \otimes v_i^{(d)}.$$

In the matrix case, where $d = 2$, this amounts to finding a decomposition of a matrix $M \in \mathbb{R}^{n \times n}$ such that:

$$M = \sum_{i=1}^{r} u_i v_i^{\mathsf{T}},$$

where $u_i, v_i \in \mathbb{R}^n$. In the matrix case, there is no unique decomposition of the matrix. Here, r is at most n, so we have $2n^2$ degrees of freedom. Given the matrix M, we obtain n^2 equations, so in total we have n^2 degrees of freedom left. Indeed, notice that given any decomposition, taking any invertible matrix $A \in GL(n; \mathbb{R})$,

$$M = \sum_{i=1}^{n} u_i v_i^{\mathsf{T}} = \sum_{i=1}^{n} (u_i A^{\mathsf{T}})(v_i A^{-1})^{\mathsf{T}},$$

where $GL(n; \mathbb{R})$ is an n^2-dimensional manifold.

However, for $d \geq 3$ and $r \leq n$, in the tensor case, we have dn^2 degrees of freedom but n^d equations, so we have a chance of being able to identify the $v_i^{(j)}$'s from T. In fact, we can often use Jennrich's algorithm.

7.1 Jennrich's algorithm

Let $T \in (\mathbb{R}^n)^{\otimes 3}$. Suppose that T is rank $r \leq n$, of the form:

$$T = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i.$$

Let U, V, and W be the matrices whose columns are u_i, v_i, and w_i. Assume that V, W are full rank.

Then, draw two vectors $\xi_1, \xi_2 \sim \mathcal{N}(0, \mathrm{Id}_{n\times n})$ from a standard Gaussian on $\mathbb{R}^n$. Compute the matrices $M_1 = T(\xi_1)$ and $M_2 = T(\xi_2)$:

$$M_j = \sum_{i=1}^{r} \langle \xi_j, u_i \rangle v_i \otimes w_i.$$

In other words, letting $D_j = \mathrm{diag}(\langle \xi_i, u_i \rangle)_{i=1}^r$, we obtain:

$$M_j = V D_j W^\mathsf{T}.$$

We would like to decompose the M_j's, in order to recover V and W, allowing us to also solve for U. Unlike before, we now have two matrices of the form VD_1W^T and VD_2W^T. By assumption, V and W are full rank. D_j will essentially always be invertible. And so, we can compute:

$$M_1 M_2^+ = V D_1 D_2^{-1} V^+.$$

Notice that v_i is an eigenvector to $M_1M_2^+$ with eigenvalue $\langle \xi_1, u_i \rangle / \langle \xi_2, u_i \rangle$. As $M_1M_2^+$ has exactly r nonzero eigenvalues, their corresponding eigenvectors must be $v_1, \ldots, v_r$.

7.2 Orthogonal tensor decomposition

Definition 24. *Let $T \in (\mathbb{R}^n)^{\otimes 3}$. An* eigenvector *of T is a unit vector $u \in \mathbb{R}^n$ such that:*

$$T(I, u, u) = \lambda u,$$

for some eigenvalue $\lambda \in \mathbb{R}$.

Note that eigenvectors for tensors are slightly different from matrix eigenvectors, in that:

- if $\lambda_1 = \lambda_2$ for two eigenvectors v_1, v_2, their linear combination is not necessarily an eigenvector
- on the other hand, $\frac{1}{\lambda_1} v_1 + \frac{1}{\lambda_2} v_2$ is proportional to an eigenvector.

Definition 25. *A* robust eigenvector *of T is $u \in \mathbb{R}^n$ such that there exists an $\epsilon > 0$ such that for all $\theta \in B(u, \epsilon)$, repeated iteration of the map*

$$\bar{\theta} \mapsto \frac{T(I, \bar{\theta}, \bar{\theta}}{\|T(I, \bar{\theta}, \bar{\theta}\|}$$

starting from θ converges to u.

Theorem 26. *Let T be odeco.*

- *the set of $\theta \in \mathbb{R}^n$ which do not converge to some v_i under repeated iterations of the above map has measure zero.*
- *the set of rubusts eigenvectors of T is equal to $\{v_1, \ldots, v_n\}$.*

Lemma 27. *Let $T \in \bigotimes^3 \mathbb{R}^n$ have an orthogonal decomposition. For a vector $\theta_0 \in \mathbb{R}^n$, suppose that the set of numbers $|\lambda_1 v_1^\mathsf{T} \theta_0|, \ldots, |\lambda_k v_k^\mathsf{T} \theta_0|$ has a unique largest element. WLOG, let $|\lambda_1 v_1^\mathsf{T} \theta_0|$ be the largest and the corresponding term for be the second largest. Then:*

$$\|v_1 - \theta_1\|^2 \leq \left(2\lambda_1^2 \sum_{i=2}^{k} \lambda_i^{-2} \right) \left| \frac{\lambda_2 v_2^\mathsf{T} \theta_0}{\lambda_1 v_1^\mathsf{T} \theta_0} \right|^{2^{t+1}}$$

8 Parameter estimation

- Describe general algorithm for latent variable models through method of moments
- Give specific moments for LSI, LDA, GMMs
- Prove bounds (contrast with EM, with not guarantees)

References

[A+2014] Anandkumar, A., et al. *"Tensor decompositions for learning latent variable models."* The Journal of Machine Learning Research 15.1 (2014): 2773-2832.

[L2012] Landsberg, J. *"Tensors: geometry and applications."* AMS (2012): 402.

[M2018] Moitra, A. *"Algorithmic aspects of machine learning."* Cambridge University Press, 2018.

[MB1967] Birkhoff, G., MacLane, S. *Algebra.* AMS Chelsea Publishing (1967).