Non-Negative Matrix Factorization

An Introduction to Topic Discovery and Latent Features

galvanize

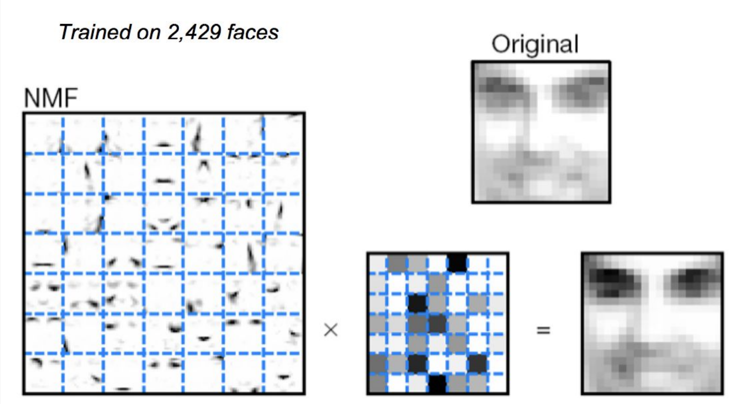
Objectives

- Describe to your friends what NMF is
  - “Unbaking cakes” into their original ingredients
  - Feature space reduction, i.e. describe the data with less features
    - Not just a subset of original features!
  - Find “latent topics” based on the many cake features
- Know applications for NMF and similar models
  - Example: Genre discovery from user/movie ratings data
  - Recommender systems [Saving this for Thursday]
- Use Alternating Least Squares or Gradient Descent to “unbake the cake”
- Lay down the fundamental concepts we’ll use again in Principal Component Analysis (PCA) & Singular Value Decomposition (SVD)

galvanize

What problem does NMF solve?

- Soft clustering
  - Clustering but where each observation can have partial membership in multiple cluster
  - *(.3 value for membership in “cluster_1”, .7 value for membership in “cluster_2”, ...)*, etc.
- Identify latent features
  - A movie can be part of many **genres** at once and a user can enjoy many different **genres**
  - Nutritional values may come from many different **ingredients** and every recipe will have many **ingredients**
  - **Topics** from documents and their corpus of text
  - **Parts** of faces in facial recognition >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>



Example: Unbaking a cake (i.e. finding its recipe)

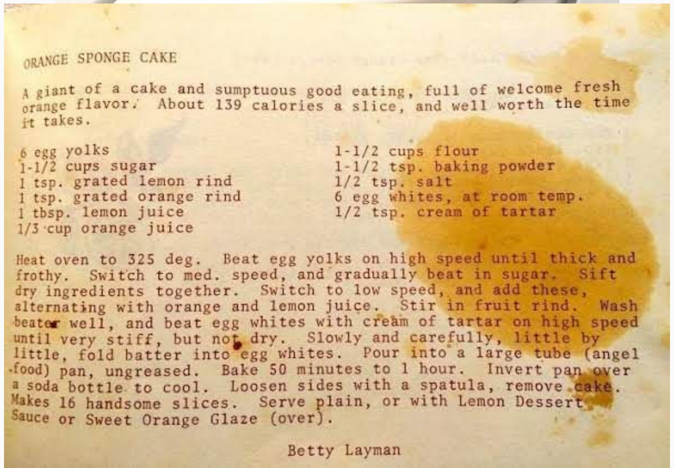
Given a cake, we cannot directly observe its ingredients. To keep it simple we'll imagine that only two are needed right now.

[flour (F) and sugar (S)]

The specifics of the recipe are a closely guarded secret!

What's the recipe?

cake as its ingredients = **W** = [w_F , w_S]


ORANGE SPONGE CAKE

A giant of a cake and sumptuous good eating, full of welcome fresh orange flavor. About 139 calories a slice, and well worth the time it takes.

6 egg yolks	1-1/2 cups flour
1-1/2 cups sugar	1-1/2 tsp. baking powder
1 tsp. grated lemon rind	1/2 tsp. salt
1 tsp. grated orange rind	6 egg whites, at room temp.
1 tbsp. lemon juice	1/2 tsp. cream of tartar
1/3 cup orange juice	

Heat oven to 325 deg. Beat egg yolks on high speed until thick and frothy. Switch to med. speed, and gradually beat in sugar. Sift dry ingredients together. Switch to low speed, and add these, alternating with orange and lemon juice. Stir in fruit rind. Wash beater well, and beat egg whites with cream of tartar on high speed until very stiff, but not dry. Slowly and carefully, little by little, fold batter into egg whites. Pour into a large tube (angel food) pan, ungreased. Bake 50 minutes to 1 hour. Invert pan over a soda bottle to cool. Loosen sides with a spatula, remove cake. Makes 16 handsome slices. Serve plain, or with Lemon Dessert Sauce or Sweet Orange Glaze (over).

Betty Layman

Example: Unbaking a cake (i.e. finding its recipe)

But we can directly measure its macronutrients:

(carbohydrates (c), proteins (p), fats (f))

e.g., the cake is made of 50g of carbohydrates, 10g of proteins, and 3g of fats

cake as macronutrients = [50 , 10 , 3]

$$\mathbf{V} = [\, v_c \,,\, v_p \,,\, v_f \,]$$

Nutrition Facts
Serving Size 2 Tbsp (30g)
Servings Per Container 10

Amount Per Serving	
Calories 70	Calories From Fat 45
	% Daily Value*
Total Fat 5g	**8%**
Saturated Fat 3.5g	**18%**
Sodium 170mg	**7%**
Total Carbohydrate 6g	**2%**
Sugars 5g	
Protein 1g	
Vitamin A 0% •	Vitamin C 0%
Calcium 0% •	Iron 4%

*Percent Daily Values are based on a 2,000 calorie diet.

Example: Unbaking a cake (i.e. finding its recipe)

Ingredients are themselves made of macro nutrients
(for which we know the proportions)

- 1 cup flour is made of 20g of carbohydrates, 5g of proteins, and 1g of fat
- 1 cup sugar is made of 10g of carbohydrates, 0g of proteins, and 1g of fat

macronutrients *from* ingredients = [[20 , 5 , 1], [10 , 0 , 1]]

Or to generalize $\mathbf{H} = [[h_{F,c} , h_{F,p} , h_{F,f}], [h_{S,c} , h_{S,p} , h_{S,f}]]$

MEATS AND FISH

	CALORIES	*PROTEIN	CARBS	FATS	FIBER
Beef, eye of round, 8 oz.	448	77	0	13	0
Turkey, leg meat, 8 oz.	356	65.5	0	8.5	0
Venison, top round steak, 7 oz.	310	64	0	4	0
Beef, ribeye, 8 oz.	521	63	0	29	0
Emu, fan fillet, 7 oz.	308	62.5	0	4.5	0
Filet mignon, 8 oz. (tenderloin)	494	62	0	25	0
Flank steak, 8 oz.	457	62	0	20	0
Elk, 7 oz.	292	60.5	0	4	0
Bison, 8 oz. (ground)	405	59	0	19	0
Tri-tip (sirloin), 8 oz.	478	59	0	25	0
Beef, liver, 7 oz.	382	58	10	10	0
Ostrich, 7 oz. (outside strip)	312	57	0	7.5	0
Extra-lean ground beef, 7 oz.	526	56	0	32	0
Bison, top sirloin, 7 oz.	332	54.5	0	11	0
Bison steak, 3.5 oz.	324	54	0	6	0
Turkey breast, 8 oz.	304	54	0	6	0
Chicken thigh, 7 oz. (meat only)	418	52	0	22	0
Beef, top loin, 8 oz. (N.Y. strip)	313	52	0	11	0
Red snapper, 7 oz.	256	52	0	2	0
Chicken breast, 8 oz.	248	52	0	4	0
Swordfish, 7 oz.	310	51	0	10	0
Beef, top sirloin, 8 oz. (lean cut)	456	46	0	29	0
Orange roughy, 7 oz.	210	45	0	2	0
Prime rib, 8 oz. (small end, lean)	576	44	0	43	0
Beef, short loin, T-bone, 6 oz.	300	44	0	12.5	0
Halibut, ½ fillet	224	42.5	0	4.5	0
Deli ham, 7 oz.	290	42	2	12	0
Pork tenderloin, 7 oz. (lean meat only)	218	42	0	4.5	0
Atlantic cod, 1 fillet	189	41	0	1.5	0
Tuna fillet, 6 oz. (bluefin)	244	40	0	8	0
Haddock, 1 fillet	168	36.5	0	1.5	0
Trout, 6 oz.	252	36	0	11	0
Tilapia fillet, 6 oz.	216	36	0	6	0
Shrimp, 6 oz.	168	35.5	0	2	0
Salmon fillet, 6 oz. (Atlantic, farmed)	312	34	0	18	0
Salmon fillet, 6 oz. (Atlantic, wild)	242	34	0	11	0
Sole, 6 oz.	154	32	0	2	0
Mackerel, 6 oz. (Atlantic)	348	31.5	0	23.5	0
Beef patty, 1 (95% lean)	141	22	0	5	0

* Items are ranked by protein content, in grams, from highest to lowest. NOTE: USDA nutrition database servings range from 6-8 ounces; the ranking by protein content does not take into account the slight weight variation of the meat and fish items.

FRUITS

	CALORIES	PROTEIN	CARBS	FATS	*FIBER
Raspberries, 1 cup	64	1.5	15	1	8
Blackberries, 1 cup	62	2	14	0.5	7.5
Pear, 1 medium	103	0.5	27.5	0	5.5
Avocado, ½ cup cubed	120	1.5	6.5	11	5
Apple, 1 medium with skin	95	0.5	25	0.5	4.5
Kiwi, 2 2-inch fruit	84	1.5	20	0.5	4
[illegible]ange, 1 medium	69	1	17.5	0.5	3.5
Blueberries, 1 cup	84	1	21.5	0.5	3.5
Banana, 1 medium	105	1.5	27	0.5	3
Strawberries, 1 cup sliced	53	1	13	0.5	3
Pineapple, 1 cup chunks	82	1	22	0	2.5
Papaya, 1 cup cubed	55	1	14	0	2.5
Peach, 1 medium	58	1.5	14.5	0.5	2
Grapefruit, ½ fruit	52	1	13	0	2
Tomato, 1 cup chopped	38	2	8	0.5	2
Cantaloupe, 1 cup	54	1.5	13	0.5	1.5
Cherries, ½ cup pitted	49	1	12	0	1.5
Grapes, 1 cup	62	0.5	16	0.5	1

* Items are ranked by fiber content, in grams, from highest to lowest.

NUTS AND SEEDS

	CALORIES	PROTEIN	CARBS	*FATS	FIBER
Brazil nuts, 1 oz.	186	4	3.5	19	2
Walnuts, 1 oz. (14 halves)	185	4.5	4	18.5	2
Sunflower seeds, 1 oz. (toasted without salt)	175	5	6	16	3.5
Almonds, 1 oz. (whole kernel)	165	6	5.5	14.5	3
Peanuts 1 oz. (no salt)	161	7.5	4.5	14	2.5
Pumpkin seeds, 1 oz. dried	158	8.5	3	14	1.5
Soybeans, ¼ cup (mature seeds, dry roasted)	194	17	14	9.5	3.5
Flaxseeds, 1 tbsp. whole	55	2	3	4.5	3

* Items are ranked by fat content, in grams, from highest to lowest.

Example: Unbaking a cake (i.e. finding its recipe)

Figuring out the recipe is about solving the following equation:

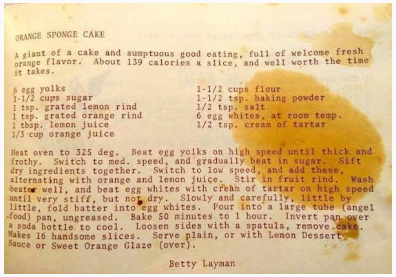
[recipe of cake as ingredients] *
[macronutrients from ingredients]

= [cake as macronutrients]

W * **H** = **V**

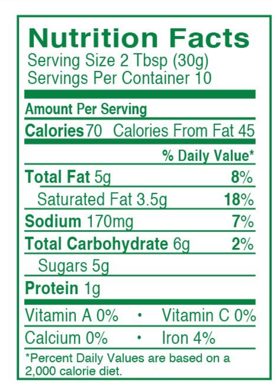
$$[\, w_F \,,\ w_S \,] * \begin{bmatrix} [\,20, 5, 1\,], \\ [\,10, 0, 1\,] \end{bmatrix} = [\,50, 10, 3\,]$$

So… we could try to solve for **W** now, but first let's think bigger.



*

=



Example: Unbaking a cake (i.e. finding its recipe)

Using MANY known values for H and V:

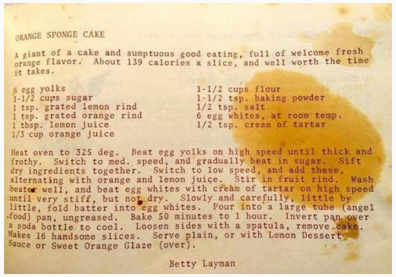
W * **H** = **V**

$$\begin{bmatrix} w_{1,F} & w_{1,S} \\ w_{2,F} & w_{2,S} \\ w_{3,F} & w_{3,S} \\ \square & \end{bmatrix} * \begin{bmatrix} 20 & 5 & 1 \\ 10 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 50 & 10 & 3 \\ 30 & 5 & 2 \\ 25 & 3 & 3 \\ \square & & \end{bmatrix}$$

Solve for **W** from set of linear equations?

These numbers are approximate so we're guaranteed to get error.

Minimize approximation error using Ordinary Least Squares solver!



ORANGE SPONGE CAKE

A giant of a cake and sumptuous good eating, full of welcome fresh orange flavor. About 139 calories a slice, and well worth the time it takes.

6 egg yolks	1-1/2 cups flour
1-1/2 cups sugar	1-1/2 tsp. baking powder
1 tsp. grated lemon rind	1/2 tsp. salt
1 tsp. grated orange rind	6 egg whites, at room temp.
1 tbsp. lemon juice	1/2 tsp. cream of tartar
1/3 cup orange juice	

Heat oven to 325 deg. Beat egg yolks on high speed until thick and frothy. Switch to med. speed, and gradually beat in sugar. Sift dry ingredients together. Switch to low speed, and add these, alternating with orange and lemon juice. Stir in fruit rind. Wash beater well, and beat egg whites with cream of tartar on high speed until very stiff, but not dry. Slowly and carefully, little by little, fold batter into egg whites. Pour into a large tube (angel food) pan, ungreased. Bake 50 minutes to 1 hour. Invert pan over a soda bottle to cool. Loosen sides with a spatula, remove cake. Makes 16 handsome slices. Serve plain, or with Lemon Dessert Sauce or Sweet Orange Glaze (over).

Betty Layman

*

MEATS AND FISH

	CALORIES	*PROTEIN	CARBS	FATS	FIBER
Beef, eye of round, 8 oz.	448	77	0	13	0
Turkey, leg meat, 8 oz.	356	65.5	0	8.5	0
Venison, top round steak, 7 oz.	310	64	0	4	0
Beef, ribeye, 8 oz.	521	63	0	29	0
Emu, fan fillet, 7 oz.	308	62.5	0	4.5	0
Filet mignon, 8 oz. (tenderloin)	494	62	0	25	0
Flank steak, 8 oz.	457	62	0	20	0
Elk, 7 oz.	292	60.5	0	4	0
Bison, 8 oz. (ground)	405	59	0	19	0
Tri-tip (sirloin), 8 oz.	478	59	0	25	0
Beef, liver, 7 oz.	382	58	10	10	0
Ostrich, 7 oz. (outside strip)	312	57	0	7.5	0
Extra-lean ground beef, 7 oz.	526	56	0	32	0
Bison, top sirloin, 7 oz.	332	54.5	0	11	0
Bison steak, 3.5 oz.	324	54	0	6	0
Turkey breast, 8 oz.	304	54	0	6	0
Chicken thigh, 7 oz. (meat only)	418	52	0	22	0
Beef, top loin, 8 oz. (N.Y. strip)	313	52	0	11	0
Red snapper, 7 oz.	256	52	0	2	0
Chicken breast, 8 oz.	248	52	0	4	0
Swordfish, 7 oz.	310	51	0	10	0
Beef, top sirloin, 8 oz. (lean cut)	456	46	0	29	0
Orange roughy, 7 oz.	210	45	0	2	0
Prime rib, 8 oz. (small end, lean)	576	44	0	43	0
Beef, short loin, T-bone, 6 oz.	300	44	0	12.5	0
Halibut, ½ fillet	224	42.5	0	4.5	0
Deli ham, 7 oz.	290	42	2	12	0
Pork tenderloin, 7 oz. (lean meat only)	218	42	0	4.5	0
Atlantic cod, 1 fillet	189	41	0	1.5	0
Tuna fillet, 6 oz. (bluefin)	244	40	0	8	0
Haddock, 1 fillet	168	36.5	0	1.5	0
Trout, 6 oz.	252	36	0	11	0
Tilapia fillet, 6 oz.	216	36	0	6	0
Shrimp, 6 oz.	168	35.5	0	2	0
Salmon fillet, 6 oz. (Atlantic, farmed)	312	34	0	18	0
Salmon fillet, 6 oz. (Atlantic, wild)	242	34	0	11	0
Sole, 6 oz.	154	32	0	2	0
Mackerel, 6 oz. (Atlantic)	348	31.5	0	23.5	0
Beef patty, 1 (95% lean)	141	22	0	5	0

* Items are ranked by protein content, in grams, from highest to lowest. NOTE: USDA nutrition database servings range from 6-8 ounces; the ranking by protein content does not take into account the slight weight variation of the meat and fish items.

FRUITS

	CALORIES	PROTEIN	CARBS	FATS	*FIBER
Raspberries, 1 cup	64	1.5	15	1	8
Blackberries, 1 cup	62	2	14	0.5	7.5
Pear, 1 medium	103	0.5	27.5	0	5.5
Avocado, ½ cup cubed	120	1.5	6.5	11	5
Apple, 1 medium with skin	95	0.5	25	0.5	4.5
Kiwi, 2 2-inch fruit	84	1.5	20	0.5	4
ange, 1 medium	69	1	17.5	0.5	3.5
Blueberries, 1 cup	84	1	21.5	0.5	3.5
Banana, 1 medium	105	1.5	27	0.5	3
Strawberries, 1 cup sliced	53	1	13	0.5	3
Pineapple, 1 cup chunks	82	1	22	0	2.5
Papaya, 1 cup cubed	55	1	14	0	2.5
Peach, 1 medium	58	1.5	14.5	0.5	2
Grapefruit, ½ fruit	52	1	13	0	2
Tomato, 1 cup chopped	38	2	8	0.5	2
Cantaloupe, 1 cup	54	1.5	13	0.5	1.5
Cherries, ½ cup pitted	49	1	12	0	1.5
Grapes, 1 cup	62	0.5	16	0.5	1

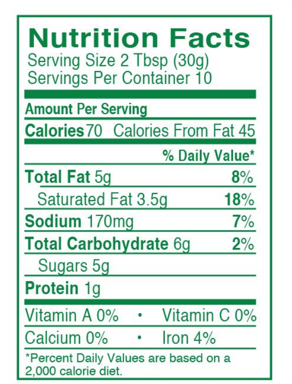
* Items are ranked by fiber content, in grams, from highest to lowest.

NUTS AND SEEDS

	CALORIES	PROTEIN	CARBS	*FATS	FIBER
Brazil nuts, 1 oz.	186	4	3.5	19	2
Walnuts, 1 oz. (14 halves)	185	4.5	4	18.5	2
Sunflower seeds, 1 oz. (toasted without salt)	175	5	6	16	3.5
Almonds, 1 oz. (whole kernel)	165	6	5.5	14.5	3
Peanuts 1 oz. (no salt)	161	7.5	4.5	14	2.5
Pumpkin seeds, 1 oz. dried	158	8.5	3	14	1.5
Soybeans, ¼ cup (mature seeds, dry roasted)	194	17	14	9.5	3.5
Flaxseeds, 1 tbsp. whole	55	2	3	4.5	3

* Items are ranked by fat content, in grams, from highest to lowest.

=



Example: Unbaking a cake (i.e. finding its recipe)

Let's focus on that first row for now.

$$[[\ w_{1,F}\ ,\ \ w_{1,S}\ ]] * [[\ 20\ ,\ 5\ ,\ 1\ ],\ = [[\ 50\ ,\ 10\ ,\ 3\ ]]$$

$$[\ 10\ ,\ 0\ ,\ 1\ ]]$$

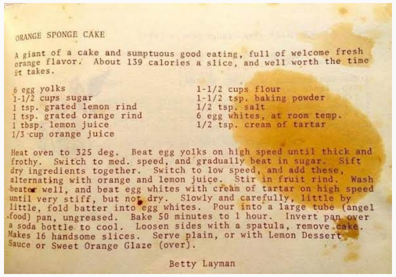
$$20\ w_{1,F} + 10\ w_{1,S} = v' \approx v_{1,c} = 50$$

$$5\ \ w_{1,F} + \ 0\ \ w_{1,S} = v' \approx v_{1,p} = 10$$

$$1\ \ w_{1,F} + \ 1\ \ w_{1,S} = v' \approx v_{1,f} = \ 3$$

$$\underset{\mathbf{W} = [[\ w_{1,F}\ ,\ \ w_{1,S}\ ]]}{\text{argmin}} \quad ||\mathbf{V'} - \mathbf{V}||^2$$

Find **W** that minimizes the reconstruction



ORANGE SPONGE CAKE

A giant of a cake and sumptuous good eating, full of welcome fresh orange flavor. About 139 calories a slice, and well worth the time it takes.

6 egg yolks
1-1/2 cups sugar
1 tsp. grated lemon rind
1 tsp. grated orange rind
1 tbsp. lemon juice
1/3 cup orange juice

1-1/2 cups flour
1-1/2 tsp. baking powder
1/2 tsp. salt
6 egg whites, at room temp.
1/2 tsp. cream of tartar

Heat oven to 325 deg. Beat egg yolks on high speed until thick and frothy. Switch to med. speed, and gradually beat in sugar. Sift dry ingredients together. Switch to low speed, and add these, alternating with orange and lemon juice. Stir in fruit rind. Wash beater well, and beat egg whites with cream of tartar on high speed until very stiff, but not dry. Slowly and carefully, little by little, fold batter into egg whites. Pour into a large tube (angel food) pan, ungreased. Bake 50 minutes to 1 hour. Invert pan over a soda bottle to cool. Loosen sides with a spatula, remove cake. Makes 16 handsome slices. Serve plain, or with Lemon Dessert Sauce or Sweet Orange Glaze (over).

Betty Layman

*

MEATS AND FISH

	CALORIES	*PROTEIN	CARBS	FATS	FIBER
Beef, eye of round, 8 oz.	448	77	0	13	0
Turkey, leg meat, 8 oz.	356	65.5	0	8.5	0
Venison, top round steak, 7 oz.	310	64	0	4	0
Beef, ribeye, 8 oz.	521	63	0	29	0
Emu, fan fillet, 7 oz.	308	62.5	0	4.5	0
Filet mignon, 8 oz. (tenderloin)	494	62	0	25	0
Flank steak, 8 oz.	457	62	0	20	0
Elk, 7 oz.	292	60.5	0	4	0
Bison, 8 oz. (ground)	405	59	0	19	0
Tri-tip (sirloin), 8 oz.	478	59	0	25	0
Beef, liver, 7 oz.	382	58	10	10	0
Ostrich, 7 oz. (outside strip)	312	57	0	7.5	0
Extra-lean ground beef, 7 oz.	526	56	0	32	0
Bison, top sirloin, 7 oz.	332	54.5	0	11	0
Bison steak, 3.5 oz.	324	54	0	6	0
Turkey breast, 8 oz.	304	54	0	6	0
Chicken thigh, 7 oz. (meat only)	418	52	0	22	0
Beef, top loin, 8 oz. (N.Y. strip)	313	52	0	11	0
Red snapper, 7 oz.	256	52	0	2	0
Chicken breast, 8 oz.	248	52	0	4	0
Swordfish, 7 oz.	310	51	0	10	0
Beef, top sirloin, 8 oz. (lean cut)	456	46	0	29	0
Orange roughy, 7 oz.	210	45	0	2	0
Prime rib, 8 oz. (small end, lean)	576	44	0	43	0
Beef, short loin, T-bone, 6 oz.	300	44	0	12.5	0
Halibut, ½ fillet	224	42.5	0	4.5	0
Deli ham, 7 oz.	290	42	2	12	0
Pork tenderloin, 7 oz. (lean meat only)	218	42	0	4.5	0
Atlantic cod, 1 fillet	189	41	0	1.5	0
Tuna fillet, 6 oz. (bluefin)	244	40	0	8	0
Haddock, 1 fillet	168	36.5	0	1.5	0
Trout, 6 oz.	252	36	0	11	0
Tilapia fillet, 6 oz.	216	36	0	6	0
Shrimp, 6 oz.	168	35.5	0	2	0
Salmon fillet, 6 oz. (Atlantic, farmed)	312	34	0	18	0
Salmon fillet, 6 oz. (Atlantic, wild)	242	34	0	11	0
Sole, 6 oz.	154	32	0	2	0
Mackerel, 6 oz. (Atlantic)	348	31.5	0	23.5	0
Beef patty, 1 (95% lean)	141	22	0	5	0

* Items are ranked by protein content, in grams, from highest to lowest. NOTE: USDA nutrition database servings range from 6-8 ounces; the ranking by protein content does not take into account the slight weight variation of the meat and fish items.

FRUITS

	CALORIES	PROTEIN	CARBS	FATS	*FIBER
Raspberries, 1 cup	64	1.5	15	1	8
Blackberries, 1 cup	62	2	14	0.5	7.5
Pear, 1 medium	103	0.5	27.5	0	5.5
Avocado, ½ cup cubed	120	1.5	6.5	11	5
Apple, 1 medium with skin	95	0.5	25	0.5	4.5
Kiwi, 2 2-inch fruit	84	1.5	20	0.5	4
ange, 1 medium	69	1	17.5	0.5	3.5
Blueberries, 1 cup	84	1	21.5	0.5	3.5
Banana, 1 medium	105	1.5	27	0.5	3
Strawberries, 1 cup sliced	53	1	13	0.5	3
Pineapple, 1 cup chunks	82	1	22	0	2.5
Papaya, 1 cup cubed	55	1	14	0	2.5
Peach, 1 medium	58	1.5	14.5	0.5	2
Grapefruit, ½ fruit	52	1	13	0	2
Tomato, 1 cup chopped	38	2	8	0.5	2
Cantaloupe, 1 cup	54	1.5	13	0.5	1.5
Cherries, ½ cup pitted	49	1	12	0	1.5
Grapes, 1 cup	62	0.5	16	0.5	1

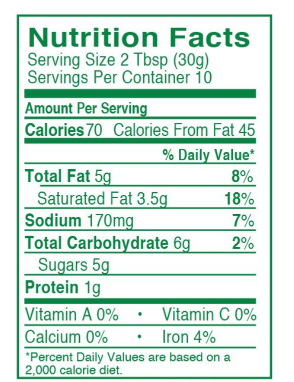
* Items are ranked by fiber content, in grams, from highest to lowest.

NUTS AND SEEDS

	CALORIES	PROTEIN	CARBS	*FATS	FIBER
Brazil nuts, 1 oz.	186	4	3.5	19	2
Walnuts, 1 oz. (14 halves)	185	4.5	4	18.5	2
Sunflower seeds, 1 oz. (toasted without salt)	175	5	6	16	3.5
Almonds, 1 oz. (whole kernel)	165	6	5.5	14.5	3
Peanuts 1 oz. (no salt)	161	7.5	4.5	14	2.5
Pumpkin seeds, 1 oz. dried	158	8.5	3	14	1.5
Soybeans, ¼ cup (mature seeds, dry roasted)	194	17	14	9.5	3.5
Flaxseeds, 1 tbsp. whole	55	2	3	4.5	3

* Items are ranked by fat content, in grams, from highest to lowest.

=



Nutrition Facts
Serving Size 2 Tbsp (30g)
Servings Per Container 10

Amount Per Serving	
Calories 70 Calories From Fat 45	
	% Daily Value*
Total Fat 5g	8%
Saturated Fat 3.5g	18%
Sodium 170mg	7%
Total Carbohydrate 6g	2%
Sugars 5g	
Protein 1g	
Vitamin A 0% • Vitamin C 0%	
Calcium 0% • Iron 4%	

*Percent Daily Values are based on a 2,000 calorie diet.

Generalization: Unbaking cakes

$$\mathbf{W} * \mathbf{H} = \mathbf{V}$$

First Problem: What if we only knew W and V instead of H and V? Can we get H?

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} * \begin{bmatrix} h_{F,c} & h_{F,p} & h_{F,f} \\ h_{S,c} & h_{S,p} & h_{S,f} \end{bmatrix} = \begin{bmatrix} 50 & 10 & 3 \\ 30 & 5 & 2 \\ 25 & 3 & 3 \end{bmatrix}$$

$$2\,h_{F,c} + 1\,h_{S,c} = v' \approx v_{1,c} = 50$$
$$2\,h_{F,p} + 1\,h_{S,p} = v' \approx v_{1,p} = 10$$
$$2\,h_{F,f} + 1\,h_{S,f} = v' \approx v_{1,f} = 3$$
$$1\,h_{F,c} + 1\,h_{S,c} = v' \approx v_{2,c} = 30$$

(... and so on)

$$\longrightarrow \quad \underset{\mathbf{H}}{\operatorname{argmin}} \; ||\mathbf{V}' - \mathbf{V}||^2 = \begin{bmatrix} h_{F,c} & h_{F,p} & h_{F,f} \\ h_{S,c} & h_{S,p} & h_{S,f} \end{bmatrix}$$

Generalization: Unbaking cakes

$$\mathbf{W} * \mathbf{H} = \mathbf{V}$$

First Problem: What if we knew W instead of H? Can we get H?

$$2\,h_{F,c} + 1\,h_{S,c} = v' \approx v_{1,c} = 50$$
$$2\,h_{F,p} + 1\,h_{S,p} = v' \approx v_{1,p} = 10$$
$$2\,h_{F,f} + 1\,h_{S,f} = v' \approx v_{1,f} = 3$$
$$1\,h_{F,c} + 1\,h_{S,c} = v' \approx v_{2,c} = 30$$

(... and so on)

$$\longrightarrow \quad \underset{H = [[\,h_{F,c},\ h_{F,p},\ h_{F,f}\,],\ [\,h_{S,c},\ h_{S,p},\ h_{S,f}\,]]}{\text{argmin}} \ ||\mathbf{V}' - \mathbf{V}||^2$$

Find **H** that minimizes the reconstruction error

Generalization: Unbaking cakes

W * H = V

The Real NMF Problem: What if we don't know either **W** or **H?**

What would solving **W** and **H** give us when we have zero ingredient information!

With enough cake nutrition info in **V**, we could find **W** and **H** that multiply to reconstruct **V** with some error. That means we could simultaneously figure out what the underlying ingredients might be *and* how nutritious each ingredient is. (We'd have to inspect the results to see that they make sense and guess at what those ingredients are!) If only there were some way to approach good **W**s and **H**s through some iterative process that minimizes the error...

That's what NMF is for!

What is the Non-Negative Matrix Factorization Algorithm?

Let's Look at the Matrices Again

$$W_{m \times r} * H_{r \times n} = V_{m \times n}$$

We put ourselves under the restriction that all of W and H are non-negative. There's no "negative flour" and no "negative movies" in a film genre:

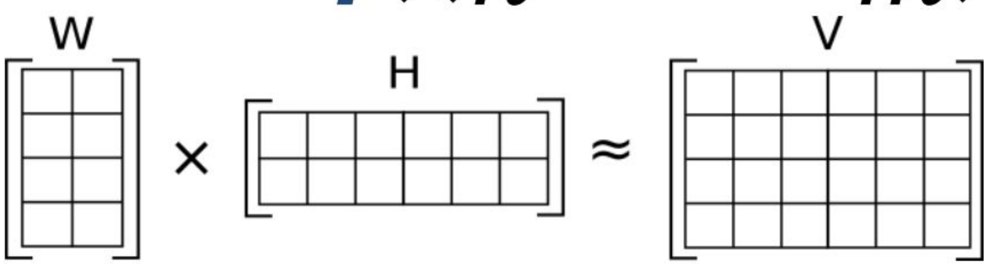
$w_{i,j} \geq 0$ and $h_{k,l} \geq 0$

r is an integer we can choose as our guess at how many latent topics (e.g. unique cake ingredients) we think there might be.

Let's Look at the Matrices Again

$$W \quad \times \quad H \quad \approx \quad V$$

$$m \times \boldsymbol{r} \qquad \boldsymbol{r} \times n \qquad m \times n$$



- $\boldsymbol{r}$ is set by you ($r < \min(m, n)$)
- Cannot be solved analytically, so approximated numerically

- **When $r > min(m, n)$, we can perfectly recreate V as $W \cdot H$**
- **Otherwise, some sort of data compression is happening**

Let’s Look at the Matrices Again

We call this type of optimization problem biconvex as it’s convex in either **W** or **H** but not both. That means if we know one of them and **V**, we could solve for the other by minimizing the reconstruction error. There’s a straightforward way we can brute force an approximate solution if this is the case.

While there is no closed form solution for **W** and **H**, if we hold one of these matrices constant, there is a closed form optimum for the other.

This leads us to the **Alternating Least Squares** algorithm.

Here's the Alternating Least Squares Algorithm That We'll Use in Today's Sprint

We will take advantage of the biconvexity by alternating which matrix, **W** or **H**, that we treat as stationary, solving for the other's optimal values, and then *clipping* all the negative values in that solution to 0.

Pseudo-code:

1. Initialize W to small, positive, random values
2. Repeat
    2.1. Find the least squares solution to $X = W \cdot H$ w.r.t. H.
    2.2. Clip negative values in H to 0: $H < 0 = 0$.
    2.3. Find the least squares solution to $X = W \cdot H$ w.r.t. W.
    2.4. Clip negative values in H to 0: $W < 0 = 0$
3. Stop when convergence (i.e., some threshold is met) such as maximum iterations or small enough decrease in RMSE.

Here's the Gradient Descent Version! (no need to clip!)

Minimize $||V - WH||^2$

with respect to W and H subject to $W, H \geq 0$

<u>Steps</u>

(1) Start with some random W and H (random, small positive values)

(2) Repeatedly adjust W and H to make RMSE smaller

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia} \frac{(V H^T)_{ia}}{(W H H^T)_{ia}}$$

($H_{a\mu}$ and W_{ia} are single entries in W and H)

- This is gradient descent! It's simple but can be slow. The convergence is sensitive to choice of step size.

(3) Stop when some threshold is met

- Decrease in RMSE, # of iterations, etc.

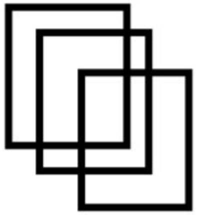
Another Use Case: Topic Modeling for Natural Language Processing

Topic Modeling (Soft Clustering)

Topic Modeling with NMF

500 documents

10,000 words



cakes ➡ documents, e.g., term frequencies

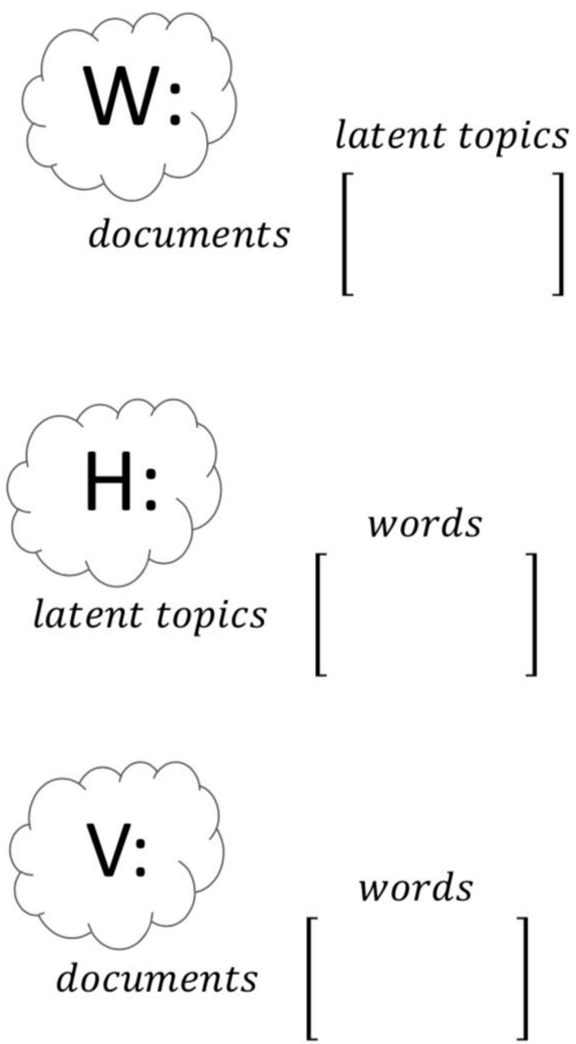
macro nutrients ➡ words

(latent) ingredients ➡ (latent) topics

$$V \approx W \cdot H$$

cake as **ingredients**	$\cdot$	**ingredients** as **macro nutrients**	$\approx$	**cake** as **macro nutrients**
document as **topics**	$\cdot$	**topics** as **words**	$\approx$	**document** as **words**

Topic Modeling (Soft Clustering)



- In looking at the dimensions of W, we notice that the number of rows remain the same, as we would expect. Thus, each row of W must represent some information about the corresponding row in V.
- Think of the column W as latent topics where the higher the word's cell value, the higher the word's rank for that latent feature.
- The same can be said about H: the number of columns between V and H are the same. Thus, each column of H must represent some information about the corresponding column in V.

Choosing k?

- Choosing k is more of an art than a science. We can look at how "good" of an approximation WH is for V and try to find the smallest k that makes it suitably small.
- At the end of the day, though, k, is likely going to be chosen based on intuition that you derive from inspecting the topics and possibly from some domain knowledge.

Topic Modeling (Soft Clustering)

Example H Matrix

$H =$

k x m

	“president”	“coach”	…	“team”	
	0.3	5.1	…	4.2	Topic 1
	10.3	1.07	…	0.08	Topic 2
	..	…	…	…	⋮
	…	…	…	…	
	2.03	0.3	…	0.001	Topic k

Topic 1? Sports?

Topic 2? Politics?

Topic Modeling (Soft Clustering)

Key Takeaways

NMF

galvanize

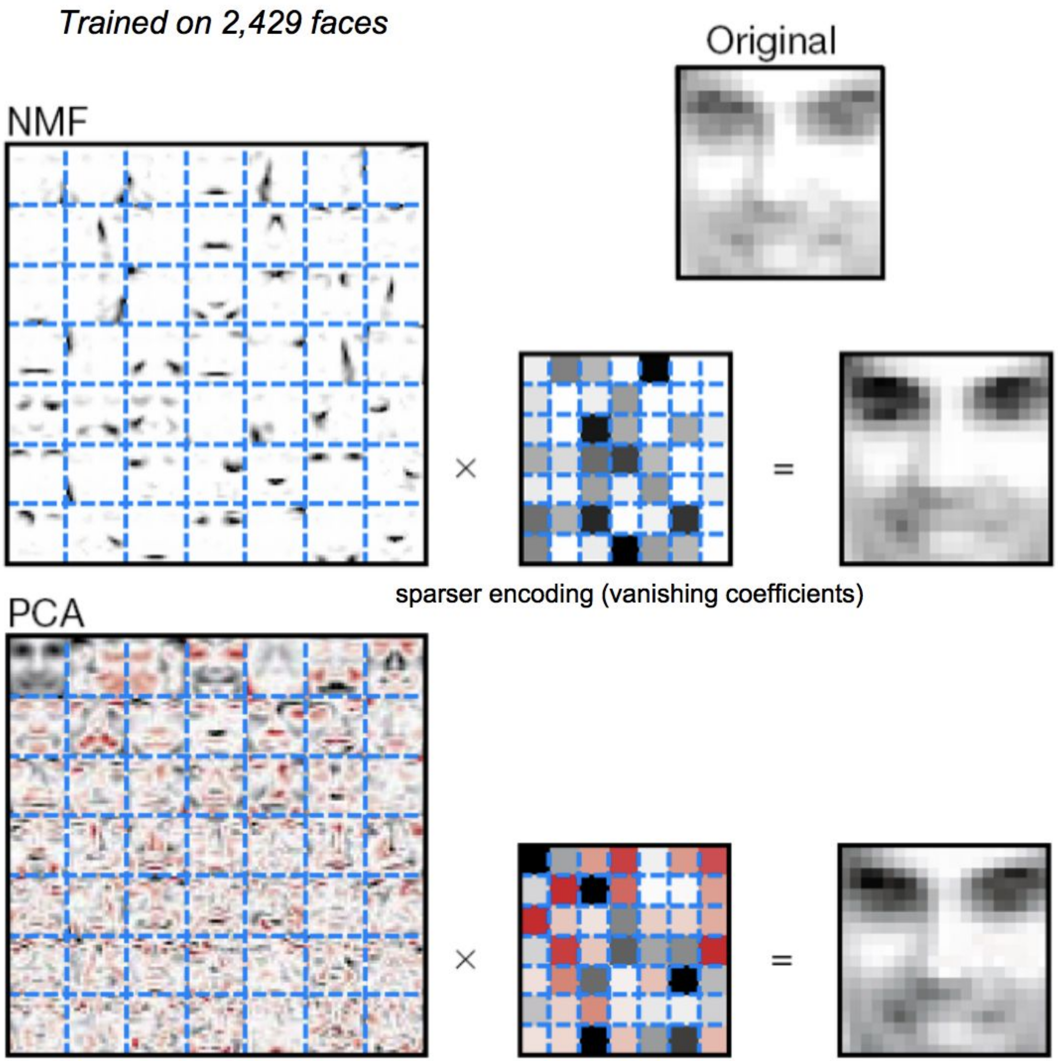
- Creates two smaller matrices with lower rank than the original matrix that tell us about underlying similarities in the rows of data.
- This creates one matrix that describes “latent topics” as composed of the original feature values and a new matrix that describes the original data as made of those topics.

Key Takeaways

Things that will help with understanding PCA tomorrow

- Other algorithms can be used to reduce dimensionality through matrix factoring similarly to NMF. They don’t produce easily interpretable “topics” the same way and that’s fine! Different tools for different needs.
- PCA and SVD are two common tools that rely on feature covariance to try to find how similarly features vary together and then creates a new way to describe those features by this similarity.
- Unlike NMF, there are unique solutions for these matrices.

Non-negative matrix factorization



Like PCA, except the coefficients in the linear combination must be *non-negative*

Forcing positive coefficients implies an additive combination of basis parts to reconstruct whole

Several versions of mouths, noses, etc.

Better physical analogue in neurons