2 handouts

CS 4786
Lec 4
2/3/15

Introduction to CCA, canonical correspondence analysis
(yet another three-letter acronym)

Announcements:

- Office hour times and locations are nearly finalized, and will be updated on the course webpage tonight.
  → http://www.cs.cornell.edu/courses/cs4786
2. Piazza (the course question & answer forum) has been active, and you are encouraged to sign up.
   - suggestion: you can use Piazza to form study groups.
3. If we received an A∅ from you, we entered you into CMS.
   → http://cms.csuglab.cornell.edu.
   This will allow you to submit subsequent assignments.
   ⊛ We'll discuss and return A∅ @ the end of class today.
   (You need to pick up your A∅ to receive credit for it.)
   - students who didn't turn it in: check in w/ me after class.
4. Corrections to the handout from last lecture have been posted to the course website. (Full lecture notes still to come.)

Topic outline: an introduction to Canonical Correlation Analysis (CCA)

← last time → , ← today →

- From covariance (Σ) } → to { correlation (different matrix, gets rid of units)
- From orthogonality (eigenvectors of Σ) } → to { non-correlatedness (different eigenvectors)

purpose: better representation of the data items

(side effect: covariance matrix tells you about relationships btwn features)

- so, Σ is not the important object.

purpose: better representation of relationships between 2 sets of features.

Pedagogical note: in today's lecture, will have a demo in a 3rd language: R

(I learned enough R in one (long) night to do this, so you can, too!)

[I asked the students if they'd stop me if I stopped making sense. They said it was a deal!]

"Crib sheet" (reminders, mostly)

Data vectors are $d \times 1$, for d features. $x_i : \begin{bmatrix} \\ \\ \\ \end{bmatrix}\}d$

Ex: each document has counts for each ~~word~~ possible word

Ex: a student has ~~scor~~ grades in each of math, English, history

Data matrix:

$$X: \ n \left\{ \begin{bmatrix} — & x_1^T & — \\ — & x_2^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix} \right. \quad \overbrace{\quad}^{d}$$

Mean vector: $\begin{bmatrix} — & \mu^T & — \end{bmatrix}$ (sum each of the data matrix's columns, divide each entry by $1/n$).

Covariance matrix $\boxed{\Sigma}$: $d \times d$, the i,j^{th} entry tells you ~~about how~~ something about how the i^{th} & j^{th} feature relate (covary).

$$= \frac{1}{n} \sum_{t=1}^{n} \underbrace{(x_t - \mu)}_{d \times 1} \underbrace{(x_t - \mu)^T}_{1 \times d}$$

Let X_c be the mean-centered version of X.

$$= \frac{1}{n} X_c^T X_c \ , \quad \text{~~so entries look like~~}$$

$$\begin{bmatrix} 1^{\text{st}} \text{ column of } X_c = 1^{\text{st}} \text{ feature} — \\ 2^{\text{nd}} \text{ column of } X_c \ (2^{\text{nd}} \text{ feature}) — \\ d^{\text{th}} \text{ column of } X_c \ (d^{\text{th}} \text{ feature}) \end{bmatrix} \begin{bmatrix} | & | & & | \\ 1^{\text{st}} f & 2^{\text{nd}} f & \cdots & d^{\text{th}} f \\ | & | & & | \end{bmatrix}$$

$\Rightarrow$ entries are the $1/n$-scaled inner products of ~~the~~ the features ~~features~~.

And remember, $v_1 \cdot v_2 = \|v_1\|_2 \|v_2\| \cos(\angle(v_1, v_2))$

↓ big when the vectors point in the same direction.

They're also: $E\left((i^{\text{th}} \text{ feature} - \text{its mean})(j^{\text{th}} \text{ feature} - \text{its mean}) \right)$

hence ~~certain~~ the name "covariance".

Asides (gotchas when checking other sources):
- sometimes $\boxed{\Sigma}$ is used for a related matrix, the ~~matrix~~ diagonal matrix of singular values for X.
- some ~~packages~~ software packages use the sample covariance matrix, which divides by $n-1$ instead of n, for the purposes of unbiased estimation
- sometimes X is assumed to already be mean-centered.

Canonical correlation analysis:.
Assume the features can be broken into two sets.
Find: {a linear combination of the first set, and
a linear combination of the second set
such that the correlation of these two is maximized.
Subsequent pairs should be uncorrelated to the previous pairs

ast time, and w/ PCA, we've been mostly concerned with the "data" vectors, the rows in the data matrix. Today's lecture will be more concerned with the columns; and the correlations between them.

X: $n \times d$ matrix $= \begin{bmatrix} — x_1^T — \\ — x_2^T — \\ \vdots \\ — x_n^T — \end{bmatrix}$ ← each row might be a different student

each column is a feature, like "SAT verbal score",

this might be "SAT math score"

Your crib sheet shows you that, while two lectures ago, we defined the covariance matrix Σ as an outer product:

$$\frac{1}{n}\sum_{t=1}^{n} (x_t - \mu)(x_t - \mu)^T$$

$(x_t - \mu)$: $d \times 1$; $(x_t - \mu)^T$: $1 \times d$ ⟶ so you get a $d \times d$ matrix, and there are d features by d features to compare, so that makes sense...

we can also define it in terms of inner products.

Let X_c be the mean-centered version of X (subtract μ from each row)

μ: $d \times 1$ = [avg value of 1st feature; avg value of 2nd feature; ⋮]

Then the columns of X_c are the transformed features.

~~And $X_c^T X$ has entries that are inner products between~~

And $X_c^T X_c$ has entries where the i,j^{th} entry is the i^{th} transformed feature dot-product with the j^{th} transformed feature

... where we learned from last lecture that this inner product is like a similarity:

$$f_i \cdot f_j = \text{length}(f_i) \times \text{length}(f_j) \times \cos(\theta),$$

where θ = the angle between them

$d \times n \quad n \times d$

So, $\frac{1}{n} X_c^T X_c = \Sigma$: this is another way to see that Σ gives you relationships btwn the features.

Or, observe that Σ's entries are, literally, the covariances btwn features:

E((ith feature − it's mean)(jth feature − its mean)), hence the name.

but, is the covariance matrix perfect? Does it tell you everything you need to know?

<of course not, b/c you can see from the topic outline that we're going to move from covariance to correlation>.

```
> source('covcorr.R')
Loading required package: Matrix
Loading required package: MASS
[1] "The data matrix is:"
     pens,wkdays pencils,wkdays pens,weekends pencils,weekends
[1,]    4.989278      7.019434      60.02640         84.01705
[2,]    4.994719      7.010523      60.07891         83.98157
[3,]    4.983907      7.031791      60.05283         83.95417
[4,]    7.035838      5.078062      83.99849         60.04352
[5,]    6.953656      4.995579      84.00628         60.02426
[6,]    6.996325      5.005047      84.07194         59.95789
Pause for contemplation...
```

→ store 1
→ store 2
⋮

inversely related ~~correlated~~, by construction

inversely related ~~correlated~~, by construction.

So: either # of pens is bigger & # of pencils is smaller, or vice versa
: and sales on weekends are bigger than on weekdays

```
[1] "The (sample) covariance matrix is (R uses \"n-1\" instead of \"n\"):"
              pens,wkdays pencils,wkdays pens,weekends pencils,weekends
pens,wkdays     1.207864     -1.199536     14.42664      -14.42820
pencils,wkdays -1.199536      1.194094     -14.34351      14.34522
pens,weekends   14.426637     -14.343508    172.41026      -172.43074
pencils,weekends -14.428202    14.345220     -172.43074     172.45155
Pause for contemplation...
```

pens on weekdays vs. pens on weekdays: covariance ~~positive~~ positive.

→ pens on wkdays tends to be "opposite" of pencils on wkdays: negative sign

→ pens on wkdays vs. pens on weekends:
- positive sign (both "big" @ the same times)
- big absolute value: ~~if f_i will be big if f_j is big~~
  Or, ~~when the fns~~ under the $E((f_i - \text{mean } i)(f_j - \text{mean } j))$ interpretation, "pens on weekends" tends to have much bigger swings away from its mean. than "pens on weekdays"

Ex: "pens on weekends" mean is about 70, the distance from mean tends to be about 10;
but "pens on weekdays" mean is about 5, the distance from mean tends to be about 1.

All makes sense.

~~But what if I told you that pens on weekends is not really varying much~~

But, there's a trick to this data (which is synthetic):

[anyone ~~notice~~ notice a ~~numberi~~ numerical relationship btwn the 1st & 3rd columns?]

- turns out, perhaps, that the weekday clerks are counting by boxes-of-dozens, while the weekend clerks are counting by individual pens or pencils!

⇒ what looks like bigger variance is just a choice of different units!
(or a big jump in sales)

Correlation:
normalize the covariance by the product of the standard deviations of the variables in question.

```
[1] "The correlation matrix is:"
                 pens,wkdays pencils,wkdays pens,weekends pencils,weekends
pens,wkdays        1.0000000    -0.9988153    0.9997113      -0.9997000
pencils,wkdays    -0.9988153     1.0000000   -0.9996651       0.9996647
pens,weekends      0.9997113    -0.9996651    1.0000000      -0.9999991
pencils,weekends  -0.9997000     0.9996647   -0.9999991       1.0000000
```

→ this is about −1, the minimum possible

~~diags~~ diagonal ~~entries~~ entries are all 1, the max possible.

so you see: pens & pencils negatively correlated,
wkdays & weekends strongly correlated
(both of which were true according to the hidden model I used to generate the data).

So now we know what correlations are, but what is canonical correlation analysis?

handout says: study the relationships between sets of features.

# of pens & # of pencils are very well-defined features.
But the big questions in life often concern not-so-well-defined features.
- does political freedom correlate with economic strength?
- does academic success correlate with life success?

* what single feature could we use to represent these???

As a starting point, you might come up w/ some easy-to-acquire features that are sort of related to what you actually care about.

Ex: for academic success: GPA, test scores, # of years to graduate, graduation year

ditto for life success.

So we've got two sets of features!

Now, CCA seeks to:

find (a) a linear combination of the first set of features, and
(b) a " " " " 2nd " " " "
(and how do I choose among these infinite possible linear combos?)

such that (a) and (b) yield the max. possible correlation

→ Some combo representing academic success
→ some combo representing life success.

Actually, we also require that the linear combos have unit variance

We can repeat this process to find the next pair of lin. combos that maximize the correlation subject to being uncorrelated with the previous ones. (Like PCA analogously)

Idea: if that max correlation is high, that @ least says that there is a way to view the first "meta-feature" (like academic success) as being highly correlated w/ the second meta-feature.

If it's close to zero, though, you can't conclude much. (maybe you had bad features, or you need a nonlinear ~~funct~~ function of them).

Implementation note: the CCA problem boils down to another eigenvector problem!

Example time <data & idea from a class @ Penn State. No idea if data is synthetic>

- Features related to sales performance (first 3)
- Features related to test scores (last 4).

→ "how does sales performance correlate w/ ~~test~~ school perf?

Meta-Analysis steps...

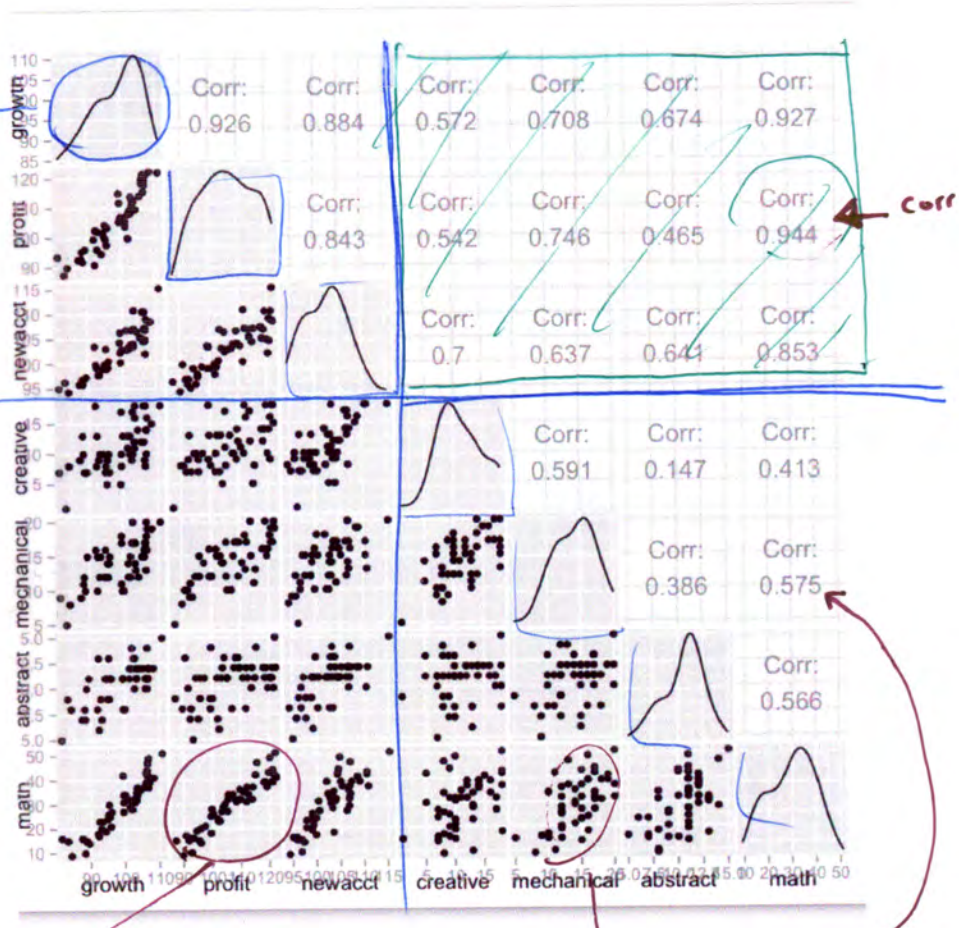
1. Look @ the data, make sure it's "OK" (no negative #'s? ranges ok? etc.)
2. can pairwise correlations answer our question?
3. If not, try CCA.

<notes on data: sales features are bigger than test scores, math #s seem higher than other test #s>

```
> mm[1:4,]
  growth profit newacct creative mechanical abstract math
1   93.0   96.0    97.8        9         12        9   20
2   88.8   91.8    96.8        7         10       10   15
3   95.0  100.3    99.0        8         12        9   26
4  101.3  103.8   106.8       13         14       12   29
> # 50 points in total
> source('pennstate.R')
```

Using ggpairs in R for pairwise correlations btwn individual features.

Feature-~~by~~against-feature display.



diagonals: histogram of "~~growth~~" ~~data~~ feature values

corr. btwn math & profit

looks less linear.

lower ~~diagonal~~ triangle: scatter plots, here, of profit made vs. math score.

Looks pretty linear; the actual correlation value is in the upper triangle.

☑: correlations btwn features having to do with ~~gross~~ sales and features having to do w/ test scores.

So in this artificial example, you can already see that math alone is well correlated w/ performance.

Nonetheless, let's forge ahead with CCA.

```
data matrix mm is loaded (....Pause....)
[1] "The canonical correlations"
[1] 0.9944827 0.8781065 0.3836057
```

how big the correlation is between the chosen linear combos.

1st column in following: the best pair of lin. combos (unit-variance)

in the following, we'll only focus on the first column.

2nd column in following: the best pair of (unit-variance) lin. combos that are uncorrelated w/ the 1st pair.

3rd column in following: the best pair of (unit-variance) lin. combos that are uncorrelated w/ the 1st 2 pairs.

Notational issue: R refers to the first set of features as X, and the second as Y.

```
(....Pause....)
[1] "The canonical coefficients (but remember the numbers had different ranges)"
$xcoef
                [,1]       [,2]       [,3]
growth  -0.06237788 -0.1740703  0.3771529
profit  -0.02092564  0.2421641 -0.1035150
newacct -0.07825817 -0.2382940 -0.3834151

$ycoef
                [,1]       [,2]       [,3]
creative   -0.06974814 -0.19239132 -0.24655659
mechanical -0.03073830  0.20157438  0.14189528
abstract   -0.08956418 -0.49576326  0.28022405
math       -0.06282997  0.06831607 -0.01133259
```

The coefficients on the first set of features. (sales).

The coefficients on the 2nd set of features (academic)

So correlation btwn:
- .06 × growth
- .02 × profit
- .07 × newacct

and

- .07 × creativic
- .03 × mechanical
- .09 × abstract
- .06 × math

= .994

→ 3rd smallest coeff, but remember that "math" had larger value

So, it's easier to see which features correlate best w/ the lin. combr.

correlation between each sales featur and the linear combo of sales.

```
$corr.X.xscores
                [,1]       [,2]        [,3]
growth  -0.9798776 0.0006477883  0.199598477
profit  -0.9464085 0.3228847489 -0.007504408
newacct -0.9518620 0.1863009724 -0.243414776
```

→ about equal, so the lin. combo is like an equal mixture of them.

correlation btwn each test feature and the linear combo of test scores.

```
$corr.Y.yscores
                [,1]       [,2]        [,3]
creative   -0.6383313 -0.2156981 -0.65140953
mechanical -0.7211626  0.2375644  0.06773775
abstract   -0.6472493 -0.5013329  0.57422365
math       -0.9440859  0.1975329  0.09422619
```

→ mostly looks like math is the "contributor" which makes sense b/c it had high corr. w/ each individual sales feature.

correlation btwn each sales feature & the linear combo of test scores.

```
$corr.X.yscores
              [,1]         [,2]         [,3]
growth  -0.9744713 0.0005688272  0.076567107
profit  -0.9411869 0.2835272081 -0.002878734
newacct -0.9466102 0.1635921013 -0.093375287
```

→ bigger than w/ individual test features

correlation btwn each test feature and the linear combo of sales features.

```
$corr.Y.xscores
                [,1]       [,2]        [,3]
creative   -0.6348095 -0.1894059 -0.24988439
mechanical -0.7171837  0.2086069  0.02598458
abstract   -0.6436782 -0.4402237  0.22027544
math       -0.9388771  0.1734549  0.03614570
```

- see the UCLA example for an arguably more interesting case:
  psychological features (individual belief in control, self-esteem, motivation)
  vs. academic performance and gender
  <but the data might be synthetic>.