From Last Meetings

Developed Mathematics behind PCA:

- Review of Linear Algebra and Multivariate Probability
- Analyzed PCA, using Eigenvalue decomp. of $\hat{\Sigma}$
- Explored “Dual PCA problem”,for faster computation
- Only treated “$\tilde{X}$ full rank” case

Summary of PCA dual problem

Recall "data matrix" notation: $\tilde{X} = \frac{1}{\sqrt{n-1}} \begin{pmatrix} \underline{X}_1 - \underline{\bar{X}} & \cdots & \underline{X}_n - \underline{\bar{X}} \end{pmatrix}_{d\times n}$

Recall: $\hat{\Sigma}_{d\times d} = \tilde{X}\tilde{X}^t$ has the eigenvalue decomp. $\hat{\Sigma} = BDB^t$

The "dual eigen problem" replaces columns by rows in $\tilde{X}$:

Let $\Sigma^*_{n\times n} = \tilde{X}^t \tilde{X}$, and find B^*, D^*, so that $\Sigma^* = B^* D^* B^{*t}$

(now only $n < d$ dimensional)

Summary of PCA dual problem (cont.)

Now suppose know sol'n to dual problem, i.e. know B^* and D^*

How do we find B and D?

Solution 1: Assume $D^* = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$ is of full rank,

i.e. $\lambda_1 \geq \cdots \geq \lambda_n > 0$, i.e. $\tilde{X}$ and $\hat{\Sigma}$ are of full rank

Summary of PCA dual problem (cont.)

Then, $$D_{d\times d} = \begin{pmatrix} D^*{}_{n\times n} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & & \cdots & & 0 \\ 0 & \ddots & & & & \\ & & \lambda_n & \ddots & & \vdots \\ \vdots & & \ddots & 0 & & \\ & & & & \ddots & 0 \\ 0 & & \cdots & & 0 & 0 \end{pmatrix}$$

And first n cols of B are given by $\breve{B}_{d\times n} = \tilde{X}B^*\left(D^*\right)^{-1/2}$,

PCA Dual Problem (cont.)

Solution 2: For $D^* = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$ not of full rank,

Similar, but now work with $n' \leq rank(D^*) = rank(\tilde{X})$

And find only "1st n' eigencomponents"

PCA dual problem (cont.)

Still have:

- First n' eigenvectors are $\lambda_1, \ldots, \lambda_{n'}$

- First n' cols of B are $\breve{B}_{d \times n'} = \tilde{X}\widehat{B}\left(\widehat{D}\right)^{-1/2}$

where:

$$\widehat{B}_{n \times n'} \;=\; \text{first } n' \text{ cols of } B^*$$

$$\widehat{D}_{n' \times n'} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_{n'} \end{pmatrix}$$

PCA dual problem (cont.)

Then can fill in other eigenvectors:

- Gram-Schmidt orthogonalization?
- More efficient method?

Or, maybe only care about those where $\lambda_j > 0$

(i.e. directions where we have data?)

PCA Time Trials

What is the gain in speed? Time trial comparisons

For $d = 10, 20, 50, 100, ..., 500$, and for $n = 10, 20, 50, 100, ..., 500$,

Timed versions of PCA (using Matlab's function eigs)

Trial 1: Direct PCA, all d eigenvectors (recall $\Sigma_{d \times d}$).

show PCAtimest1p4.ps

PCA Time Trials (cont.)

Top Row: Views of times (in seconds)

Problem: Smaller times “compressed into 0”

Bottom Row: Different scale: $\log_{10}$ times vs. $\log_{10} d$ & n

1st column: overall surface

2nd column: slice in n direction

3rd column: slice in d direction

PCA Time Trials (cont.)

Trial 1: Direct PCA, all d eigenvectors (recall $\Sigma_{d\times d}$)

- nearly no dependence on n
- since need to compute all d
- grows like $O\left(d^3\right)$? (for larger d?)
- since need to solve $d\times d$ system for each of d e. v. s
- limited relevance if only need 1st n

PCA Time Trials (cont.)

View 2: Compute for only non-zero eigenvalues

(generally $n-1$ since mean is subtracted for PCA)

a. Direct PCA

Show PCAtimest1p1.ps

- for each d, increases in n, until level d is passed
- since are computing more eigenvectors
- for each n, 1st inc's rapidly in d, slowly after d is passed
- since for $n > d$ only harder expense is covariance calc.

PCA Time Trials (cont.)

b. Dual PCA

Show PCAtimest1p2.ps

- Times are transpose of (a).

- Since “swap rows and columns” means “$d \leftrightarrow n$”

Flip back and forth

PCA Time Trials (cont.)

c. Chosen PCA (to min size of computed eigen-analysis)

Show PCAtimest1p3.ps

- Times are essentially mins of (a) and (b)

Flip back and forth between last 3

- Symmetric in d and n
- Worst case is $d = n$ (direct and dual equally hard)
- As expected from theory

PCA Time Trials (cont.)

How useful is this?

- For $n \approx d$, no benefit
- For n (d) = 100, & d (n) = 500, factor of ~20
- For n (d) = 50, & d (n) = 100, factor of ~10
- For n or d <= 200, time <= 10 sec's, so not major deal?

PCA Time Trials (cont.)

View 3: Compute only first 8 eigenvalues and vectors

Show PCAtimest1p5.ps, PCAtimest1p6.ps, PCAtimest1p7.ps

- similar lessons
- overall times <= 30 secs
- for n or d <= 200, times <= 5 (at worst) 10 sec's
- trivial except for simulation

Explore Rescalings

Background: PCA finds "direction of greatest variability",

by eigenanalysis of covariance matrix: $\hat{\Sigma}_{d \times d} = \tilde{X}\tilde{X}^t$

where $$\tilde{X} = \frac{1}{\sqrt{n-1}} \left(\underline{X}_1 - \underline{\bar{X}} \quad \cdots \quad \underline{X}_n - \underline{\bar{X}} \right)_{d \times n}$$

When does this make sense?

Classical Multivariate Analysis: Not when "units are different"
(e.g. X_1 in m, X_2 in sec, X_3 in \$)

Explore Rescalings (cont.)

An FDA example: “M-reps” (some “angles” and some lengths)

Show GreggTracton.html

Classical solution; transform to “unit free” scale

i.e. replace covariance matrix with “correlation matrix”

$$\bar{\Sigma} = \begin{pmatrix} 1 & \rho(X_1, X_2) & \cdots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho(X_{n-1}, X_n) \\ \rho(X_n, X_1) & \cdots & \rho(X_n, X_{n-1}) & 1 \end{pmatrix}$$

where $\rho(X_i, X_j) = \dfrac{\operatorname{cov}(X_i, X_j)}{\operatorname{var}(X_i) \cdot \operatorname{var}(X_j)}$

Explore Rescalings (cont.)

Correlation matrix:

Use same form for either “theoretical” or “empirical” versions

Matrix version:

$$\bar{\Sigma} = D\Sigma D,$$

where

$$D = \begin{pmatrix} \frac{1}{sd(X_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{sd(X_n)} \end{pmatrix}$$

Explore Rescalings (cont.)

Standardized data version: $\bar{\Sigma}_{d\times d} = \tilde{Z}\tilde{Z}^t$

where $$\tilde{Z} = \frac{1}{\sqrt{n-1}}\begin{pmatrix} \dfrac{\underline{X_1} - \underline{\bar{X}}}{sd(X_1)} & \cdots & \dfrac{\underline{X_n} - \underline{\bar{X}}}{sd(X_1)} \end{pmatrix}_{d\times n}$$

Shows "unit free" aspect of this transformation

Possible drawback: gives a "distortion of point cloud of data",

So "direction of greatest variability" is different (better? worse?)

Explore Rescalings (cont.)

E.g. 1: Familiar family of parabolas

Show CurvDat\ParabsCurvDat.ps and CurvDat\ParabsCurvDatCorr.ps

- very similar
- reason: cov. matrix $\approx$ corr. matrix
- I.e. coordinate-wise variances approx. same

Explore Rescalings (cont.)

E. g. 2: 3 “independent bumps”, in coordinate axis directions

Show CurvDat\Bumps3CurvDat.ps and CurvDat\Bumps3CurvDatCorr.ps

- Covariance PC 1: Finds first bump
- Covariance PC 2 & 3: Finds remaining bumps
- Corr. PC: Power of bumps spread beyond 1st 4!
- This can make a big difference!
- Which is “right”????
- Power plot: big difference in eigenvalues
  (symbols - raw scale, lines – standardized scale)

Explore Rescalings (cont.)

E.g. 3: 2 correlated bumps, 3rd independent:

Show CurvDat\Bumps2CurvDat.ps and CurvDat\Bumps2CurvDatCorr.ps

- similar lessons

E.g. 4: 3 correlated bumps

Show CurvDat\Bumps1CurvDat.ps and CurvDat\Bumps1CurvDatCorr.ps

- now Corr. PCA not quite so bad?
- Just luck?

Explore Rescalings (cont.)

E.g. 5: Corpus Callosum Data:

Show CorpColl\CCFrawAlls3.mpg

Recall direct PCA showed interesting population structure:

Show CorpColl\CCFpcaSCs3PC1.mpg, CorpColl\CCFpcaSCs3PC2.mpg, and CorpColl\CCFpcaSCs3PC3.mpg

Expect difference with “correlation PCA”? Parallel coordinates:

Show CorpColl\CCFParCorAlls3.ps

- Coordinate wise variances very different
- So expect large difference

Explore Rescalings (cont.)

Correlation PCA:

Show CorpColl\CCFpcaSCs3PC1Corr.mpg, CorpColl\CCFpcaSCs3PC2Corr.mpg, CorpColl\CCFpcaSCs3PC3Corr.mpg,

- found only “pixel effect directions”

- since these “have been magnified” (see Par. Coord’s)

- similar effect to Fisher Linear Disc.

Show CorpColl\CCFfldSCs3mag.mpg

- Correlation PCA clearly inferior here

Explore Rescalings (cont.)

Summary:

- no apparent “general solution”
- depends on context
- sometimes “unit free” aspect is dominant, use Corr.
- other times Corr. PCA gives “useless distortion”

Future plans:

1. Do ICA?
2. Goodness of approximation???
3. Maths for Fisher linear discrimination
4. Polynomial embeddings and SVM discrimination
5. Validation for discrimination (various ways)
6. Internet traffic data?