Statistics of PCA

Above “optimization of directions” approach to PCA:

- gives useful insights
- shows can compute for *any* point cloud

But there are other views.

Statistics of PCA (cont.)

Alternate View 1: Gaussian likelihood

When data are multivariate Gaussian

PCA finds “major axes of elliptical contours”

of Probability density (maximum likelihood estimate)

Mistaken idea: PCA only useful for Gaussian data

Statistics of PCA (cont.)

Simple check for Gaussian distribution:

Standardized parallel coordinate plot

1. Subtract coordinate wise median (robust version of mean)

   (not good as “point cloud center”,
   but now only looking at coordinates)

2. Divide by MAD / MAD($N(0,1)$)

   (put on same scale as “standard deviation”)

3. See if data stays in range –3 to +3

Statistics of PCA (cont.)

Check for Gaussian dist’n: Standardized parallel coordinate plot

E.g. Cornea data   (recall image view of data)

- several data points > 20 “s.d.s” from the center
- distribution clearly *not* Gaussian
- strong kurtosis
- but PCA still gave strong insights

Statistics of PCA (cont.)

Alternate View 2: Dimension reduction

An approach to HDLSS data: try to reduce dimensionality

PCA approach:

- keep only largest eigenvalue projections
- optimal reduction (in sense of Sums of Squares)

Statistics of PCA (cont.)

Alternate View 3: Data compression (e.g. PKzip)

Loss-less: delete components with 0 eigenvalues

With loss: PCA gives optimal compression

(in sense of Sums of Squares)

PCA for shapes

New Data Set: Corpus Callosum data

- “window” between right and left halves of the brain
- from a vertical slice MR image of head
- “segmented” (ie. found boundary)
- shape is resulting closed curve
- have sample from $n = 71$ people
- Feature vector of $d = 80$ coefficients from

  Fourier boundary representation (closed curve)

PCA for shapes (cont.)

Modes of shape variation?