Math 182: Hidden Data in Random Matrices

Todd Kemp, APM 5202

www.math.ucsd.edu/~tkemp/182

Today: What d? 

Next: Random Matrix Ensembles

HW2: 

Lab2: Regrade Window Wed Feb 12 8am → Fri Feb 14 8pm

Midterm: Released TODAY @ 2 pm

Due by 11:59pm TOMORROW

Use any resources you like (make sure to cite them!)

No collaboration with other people.
PCA "Algorithm"

0. Read data: \( \{ x_1, x_2, \ldots, x_N \} \) in \( \mathbb{R}^m \), \( m = \# \) of "features".

1. Compute the \( \hat{X} = \frac{1}{N} \left[ x_1 - \bar{x}_N, \ldots, x_N - \bar{x}_N \right] \) \( \mathbb{C} = \hat{X} \hat{X}^T = U \Sigma \Sigma^T U^T \).

2. SVD of \( \hat{X} \) (or \( \hat{X}^T \) if easier) \( \hat{X} = U \Sigma V^T \).

3. Figure out the "right" \( d \).

4. \( Q_d = [\hat{u}_1, \ldots, \hat{u}_d] \) PC's.

5. Plot the principal coordinates of the projected data

\[ \{ Q_d^T \hat{x}_j \} \text{ for } 1 \leq j \leq N \text{ or maybe } \{ Q_d^T \hat{x}_j \} \text{ for } 1 \leq j \leq N \text{ if you want to analyze the projected mean} \]
Singular Values and Variance

We interpret $\sigma_k^2$ as the variance in the data attributed to $\hat{u}_k$. (See the Midterm...)

We model the data as $x_j = \text{sample of } X_j = t_j + \xi_j$

(low rank) signal $\uparrow$ noise $\downarrow$

Provided the SNR (signal-to-noise ratio) is not too small, we hope the "variation" in the $x_j$ is "mostly" due to signal, not noise.

So we look for signal in the $\hat{u}_k$ for the largest $\sigma_k$; i.e.

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq \sigma_{d+1}$

$\uparrow$ How do we draw the line?

$b$ "scree plot"
Scree Plots introduced by Cattell in 1966.

Fantasy scree plot "cutoff phenomenon" NEVER.

Real scree plot

Highly suspect pseudoscience
Bulk vs. Outliers

The clearest way to "see" the low rank signal is to look for outlier singular values – that pull away from the bulk.

Corresponding e.v.e.s (P(1,3)) are "delocalized".

Corresponding e.v.e.s are highly structured "sparse".
Noise without Signal: Random Matrix Theory

To understand how to detect a signal amidst noise, we first need to understand what "noise" looks like.

\[ X = \text{random } m \times N \text{ matrix} \]
\[ \text{4 entries all i.i.d., } N(0,1) \]
\[ \hat{X} = \frac{1}{\sqrt{N}} (X - \bar{X}_N I_N^T) \]

\[ \bar{X}_N = \frac{1}{N} (X_1 + \ldots + X_N) \xrightarrow{\text{SLLN}} \bar{X} \]
\[ \text{r.i.d. } N(0, \text{Im}) \]

\[ m \gg N \]

\[ C = \hat{X} \hat{X}^T \text{ or } \hat{C} = \hat{X}^T \hat{X} \]