LARSON—MATH 511—CLASSROOM WORKSHEET 15
Principal Components Analysis (PCA)

Sage/CoCalc

1. (a) Start the Chrome browser.
   (b) Go to `http://cocalc.com`
   (c) Login (likely using **your VCU email address**).
   (d) You should see an existing Project for our class. Click on that.
   (e) Click "New", then "Sage Worksheet", then call it **c15**.

Statistics

2. What is the *mean* of a list of numbers? What is the *variance* of those numbers?

3. Let X_1 be the numbers in one (centered) list of n numbers and let X_2 be the numbers in a second (centered) list of n numbers. Let A be the matrix consisting of those two rows. Let M be the matrix where $M_{i,j}$ is $\frac{1}{n}X_i \cdot X_j$ (where the lists are viewed as vectors). What are these numbers?

   M is called the *variance-covariance matrix*. Note that $M = AA^T$. So the eigenvectors of M are the $\hat{u}$'s from the SVD! The PCA theory says that $\hat{u}_1$ points in the direction of (or "explains") the greatest variance, and $\hat{u}_2$ explains the remaining variance. The same idea works if there is a larger data set (more measurements).

Principal Components Analysis

4. Open your CoCalc project Handouts folder, click on "PCA_test.sage". We'll need this file and height-data.csv". You should probably move or copy then to your root/home directory.

5. We will run the code here step-by-step in your c15 worksheet.

Low Rank Approximation & Eckart-Young Theorem

We showed $A = \sigma_1 \hat{u}_1 \hat{v}_1^T + \ldots + \sigma_r \hat{u}_r \hat{v}_r^T$

Now let $A_k = \sigma_1 \hat{u}_1 \hat{v}_1^T + \ldots + \sigma_k \hat{u}_k \hat{v}_k^T$ (for $k \leq r$). We will show that A_k is the "best" low rank approximation to A.

For any $m \times n$ matrix A, let $||A|| = \max \frac{||A\hat{x}||}{||\hat{x}||}$ (for any $\hat{x} \in \mathbb{R}^n$).

6. Find $||A - A_k||$.

7. Let B be *any* $m \times n$ matrix with rank k. The dimension of the null space of B (the "nullity") is $n - k$. Explain why there must be a non-0 vector $\hat{x}$ in $N(B) \cap span(\{\hat{v}_1, \ldots, \hat{v}_{k+1}\}$.

8. (We can assume $\hat{x}$ is unit). Argue that $||(A - B)\hat{x}|| \geq \sigma_{k+1}$.

9. Argue that $||A - A_k|| \leq ||A - B||$.

10. Explain why A_k is the "best" rank-k approximation of A.

Sage/CoCalc

11. (a) Start the Chrome browser.
    (b) Go to `http://cocalc.com`
    (c) Login (likely using **your VCU email address**).
    (d) You should see an existing Project for our class. Click on that.
    (e) Click “New”, then “Sage Worksheet”, then call it **c14**.

12. Input $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ (remember to inform Sage you mean for the entries to be interpreted as elements of a Real Double Field (RDF).

13. What is the rank of A?

14. Find the U, S, V from the SVD by evaluating: `U,S,V = A.SVD()`. Check what you have for u, S, V. What are the singular values of A?

15. Find the approximation matrix A_1.

16. Find the norm of $A - A_1$.

17. Let B be *any* 2×2 rank-1 matrix. Find the norm of $A - B$ and check that $||A - A_1|| \geq ||A - B||$.

Getting your classwork recorded

When you are done, before you leave class...

1. Click the “Make pdf” (Adobe symbol) icon and make a pdf of this worksheet. (If CoCalc hangs, click the printer icon, then “Open”, then print or make a pdf using your browser).
2. Send me an email with an informative header like “Math 511—c14 worksheet attached” (so that it will be properly recorded).
3. Remember to attach today’s classroom worksheet!