Homework 3

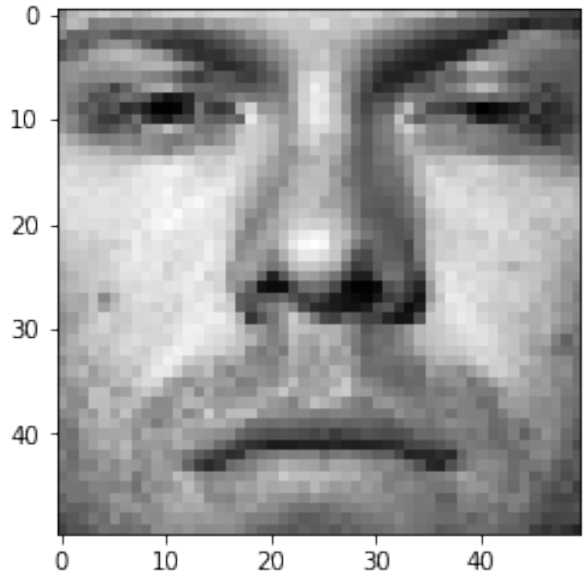
Haomiao Han (hh696), Hyein Baek (hb437)
CS5785 Applied Machine Learning

October 12, 2019

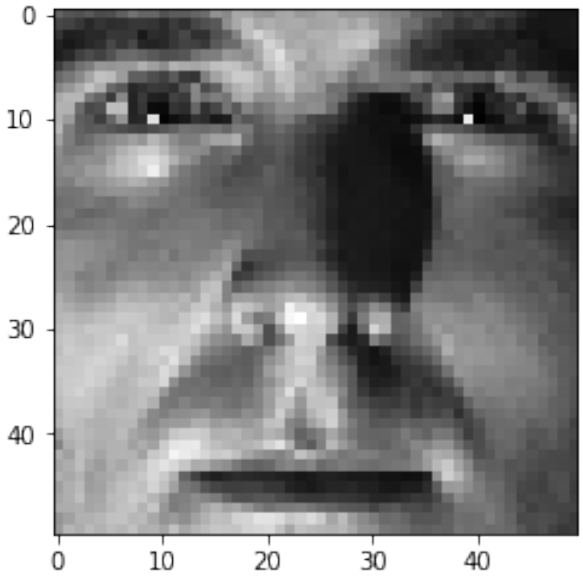
1 Programming Exercises

- Question 1

  (a)(b) The dataset is loaded into the program, using the provided code. Below is one sample image from the training set and one sample image from the test set:



(a) Sample image from training set



(b) Sample image from test set

(c) We computed the average face per the instructions. Below is the image for the average face:

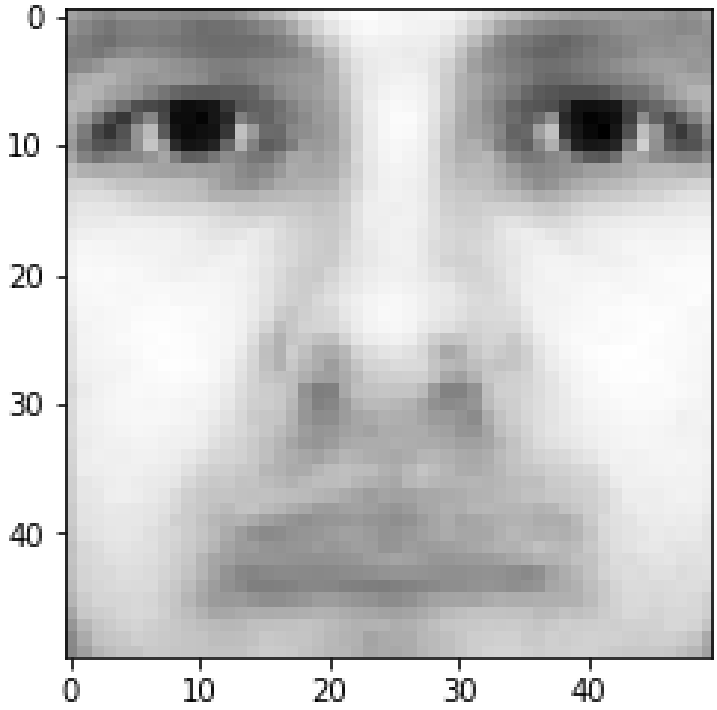
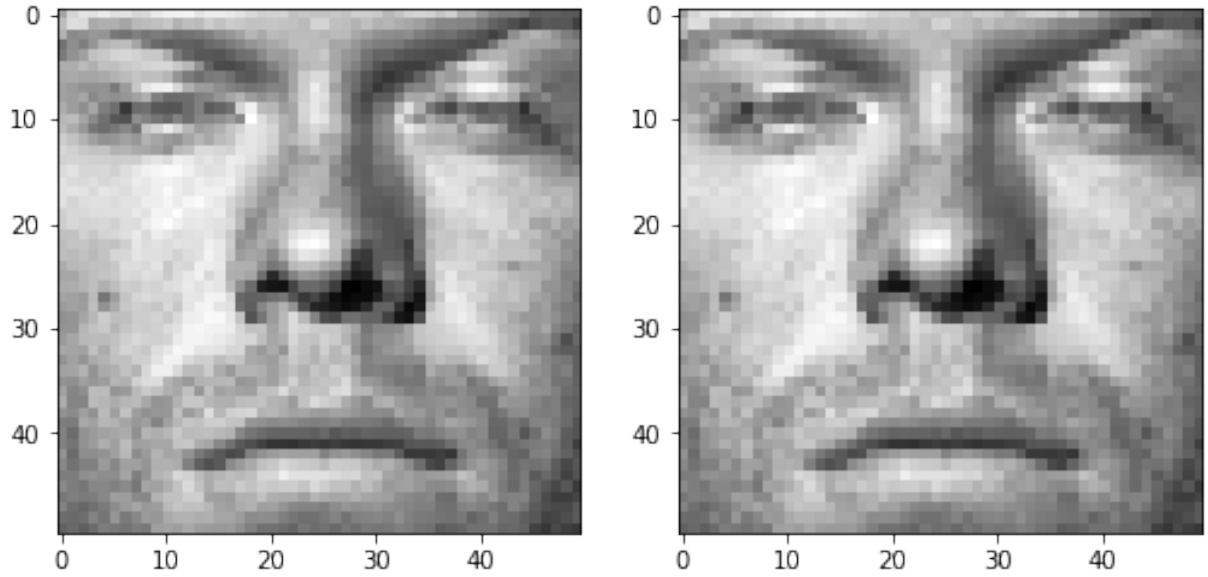


Figure 2: Average face

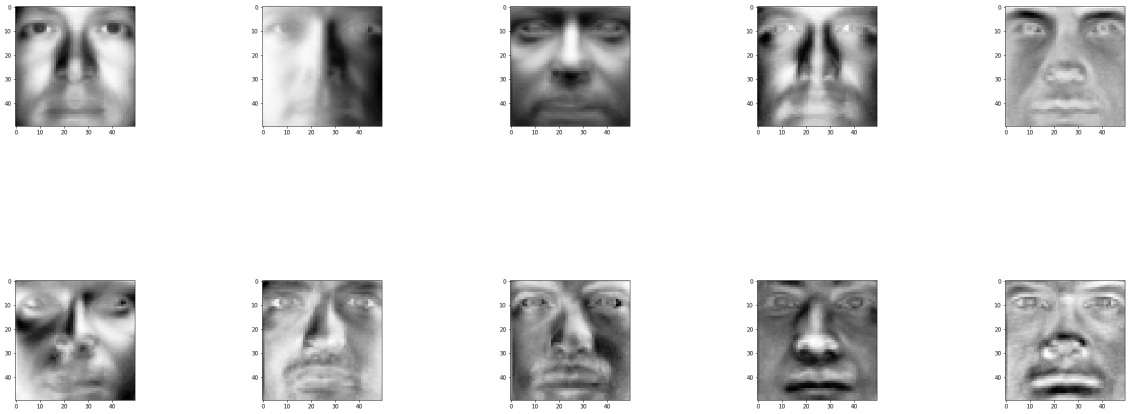
(d) We subtracted the average face from every column in the test set and the training set per the instructions. Below is one sample image from the training set and one sample image from the test set after mean subtraction:



(a) Sample image from training set after mean subtraction

(b) Sample image from test set after mean subtraction

(e) We performed Singular Value Decomposition on the training set using `numpy.svd`. Below are the first 10 eigenfaces:



(f) Using the matrices generated from SVD, we approximated $\hat{X}_r$ for $1 <= r <= 200$. We then calculated the rank-r approximation error using `numpy.linalg.norm`. Below is the graph that plots the approximation error against r values:

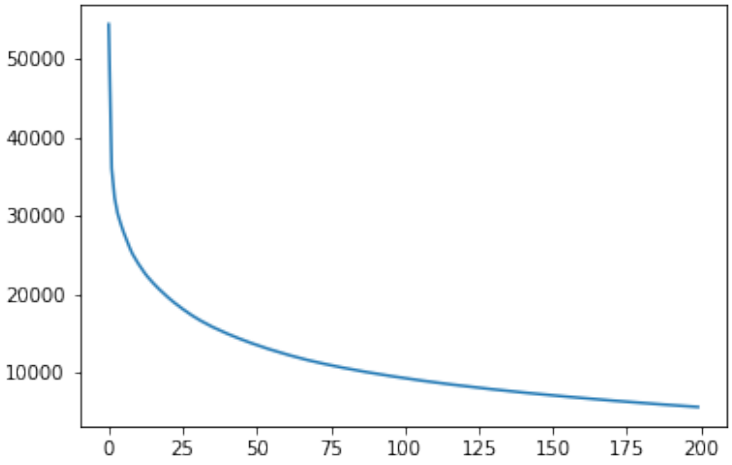


Figure 6: Graph of rank-r approximation error

(g) We created the function that generates r-dimensional feature matrix per the instructions.

(h) We generated 10-dimensional feature matrices for both the training set and the test set. We then fit the training data with a Logistic Regression model per the instructions. Our accuracy rate on the test set is 0.79. We then tried different r values for $1 <= r <= 200$. Below is the graph that plots accuracy rate against r values:

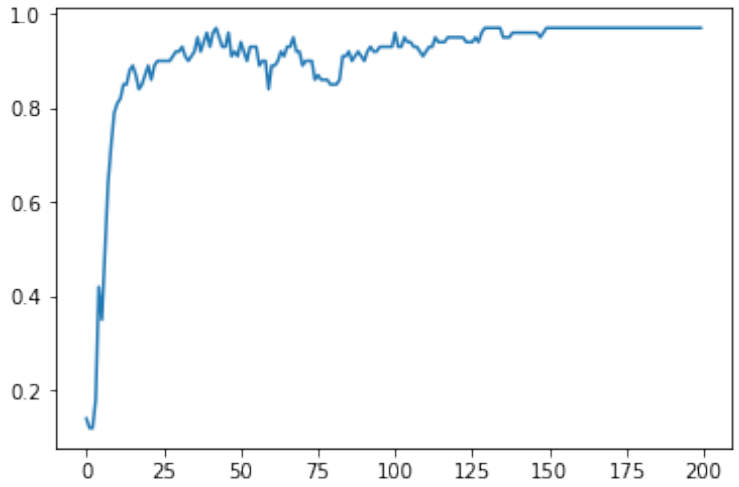


Figure 7: Graph of accuracy rate

- Question 2

  (a) We loaded the data and clustered the data using KMeans, using different k values for $5 <= k <= 20$. We then selected $k = 13$. After that, we computed the 5476-vector average value for the data and generated the top 10 words of each cluster based on the 5476-vector average value as well as the center for each cluster. Below are the results:

```
cluster # 0 has the following top 10 words: ['residues', 'crystal',
'binding', 'conserved', 'side', 'helix', 'loop', 'domain', 'chains',
'residue']

cluster # 1 has the following top 10 words: ['mail', 'compass',
'sciences', 'author', 'issue', 'news', 'page', 'scientific',
'article', 'policy']

cluster # 2 has the following top 10 words: ['material', 'solar',
'earth', 'resolution', 'observations', 'distribution', 'estimate',
'scale', 'mass', 'planetary']

cluster # 3 has the following top 10 words: ['reports', 'fig',
'values', 'temperature', 'solution', 'shows', 'composition',
'table', 'lower', 'magnetic']

cluster # 4 has the following top 10 words: ['says', 'researchers',
'scientists', 'year', 'get', 'just', 'people', 'last', 'years',
'say']

cluster # 5 has the following top 10 words: ['ocean', 'global',
'variations', 'sea', 'climate', 'atmospheric', 'temperature',
```

```
'period', 'surface', 'changes']

cluster # 6 has the following top 10 words: ['protein', 'gene',
'genes', 'dna', 'proteins', 'cell', 'sequence', 'expression',
'identified', 'sequences']

cluster # 7 has the following top 10 words: ['ligand', 'gfp',
'receptor', 'transfected', 'tnf', 'signaling', 'receptors',
'domain', 'association', 'tag']

cluster # 8 has the following top 10 words: ['spin', 'doping',
'charge', 'mott', 'ordering', 'insulator', 'spins', 'orbitals',
'doped', 'quantum']

cluster # 9 has the following top 10 words: ['significant',
'fig', 'population', 'mean', 'table', 'probability', 'rate',
'analysis', 'patterns', 'significantly']

cluster # 10 has the following top 10 words: ['p450', 'heme',
'oxygen', 'iron', 'cam', 'carbonyl', 'bond', 'intermediates',
'crystals', 'atom']

cluster # 11 has the following top 10 words: ['cells',
'expression', 'cell', 'protein', 'fig', 'expressed', 'wild',
'control', 'induced', 'hours']

cluster # 12 has the following top 10 words: ['energy',
'electron', 'fig', 'measurements', 'shows', 'order',
'density', 'measured', 'experimental', 'constant']
```

After that, we calculated the top 10 documents that are closest to the centers of each cluster. Below are the results (note that certain clusters have less than 10 documents - in that case, all documents of that cluster is reported):

```
cluster # 0 has the following top 10 titles:
"Structure of Yeast Poly(A) Polymerase Alone and in Complex with
3'-dATP"
"Structure of Murine CTLA-4 and Its Role in Modulating T Cell
Responsiveness"
"Structure of the S15,S6,S18-rRNA Complex: Assembly of the 30S
Ribosome Central Domain"
"Atomic Structure of PDE4: Insights into Phosphodiesterase Mechanism
and Specificity"
"Twists in Catalysis: Alternating Conformations of Escherichia coli
Thioredoxin Reductase"
```

"The Productive Conformation of Arachidonic Acid Bound to Prostaglandin Synthase"
"Redox Signaling in Chloroplasts: Cleavage of Disulfides by an Iron-Sulfur Cluster"
"Convergent Solutions to Binding at a Protein-Protein Interface"
"Structural Basis of Smad2 Recognition by the Smad Anchor for Receptor Activation"
"Structure of the Protease Domain of Memapsin 2 (b-Secretase) Complexed with Inhibitor"

cluster # 1 has the following top 10 titles:
"Algorithmic Gladiators Vie for Digital Glory"
"Reopening the Darkest Chapter in German Science"
"National Academy of Sciences Elects New Members"
"Corrections and Clarifications: A Short Fe-Fe Distance in Peroxodiferric Ferritin: Control of Fe Substrate versus Cofactor Decay?"
"Corrections and Clarifications: Charon's First Detailed Spectra Hold Many Surprises"
"Corrections and Clarifications: Unearthing Monuments of the Yarmukians"
"Heretical Idea Faces Its Sternest Test"
"Corrections and Clarifications: One Hundred Years of Quantum Physics"
"Corrections and Clarifications: Biotech Research Proves a Draw in Canada"
"Corrections and Clarifications: Uninterrupted MCM2-7 Function Required for DNA Replication Fork Progression"

cluster # 2 has the following top 10 titles:
"Evidence for Crystalline Water and Ammonia Ices on Pluto's Satellite Charon"
"Widespread Complex Flow in the Interior of the Antarctic Ice Sheet"
"Windows through the Dusty Disks Surrounding the Youngest Low-Mass P rotostellar Objects"
"Solar Wind Record on the Moon: Deciphering Presolar from Planetary Nitrogen"
"Folds on Europa: Implications for Crustal Cycling and Accommodation of Extension"
"Discovery of Gaseous <latex>S_2</latex> in Io's Pele Plume"
"A Close-Up Look at Io from Galileo's Near-Infrared Mapping Spectrometer"
"Hot and Dry Deep Crustal Xenoliths from Tibet"
"A Global View of Martian Surface Compositions from MGS-TES"
"Osmium Isotopic Evidence for Mesozoic Removal of Lithospheric Mantle beneath the Sierra Nevada, California"

cluster # 3 has the following top 10 titles:
"Thermal, Catalytic, Regiospecific Functionalization of Alkanes"
"The Formation of Chondrules at High Gas Pressures in the Solar Nebula"

"Neutral, Single-Component Nickel (II) Polyolefin Catalysts That Tolerate Heteroatoms"
"Clues from a Shocked Meteorite"
"Influences of Dietary Uptake and Reactive Sulfides on Metal Bioavailability from Aquatic Sediments"
"Nitric Acid Trihydrate (NAT) in Polar Stratospheric Clouds"
"Information Storage and Retrieval through Quantum Phase"
"Subducted Seamount Imaged in the Rupture Zone of the 1946 Nankaido Earthquake"
"Homogenization of Fish Faunas across the United States"
"A Monoclinic Post-Stishovite Polymorph of Silica in the Shergotty Meteorite"

cluster # 4 has the following top 10 titles:
"Information Technology Takes a Different Tack"
"Science Survives in Breakthrough States"
"Vaccine Studies Stymied by Shortage of Animals"
"For 'Father' of Abortion Drug, Vindication at Last"
"On a Slippery Slope to Mediocrity?"
"In Europe, Hooligans Are Prime Subjects for Research"
"Japan's Whaling Program Carries Heavy Baggage"
"Is AIDS in Africa a Distinct Disease?"
"New Science Chief Must Juggle Missions and Politics"
"Building a Disease-Fighting Mosquito"

cluster # 5 has the following top 10 titles:
"Isotopic Evidence for Variations in the Marine Calcium Cycle over the Cenozoic"
"Glacial Climate Instability"
"Temporal Trends in Deep Ocean Redfield Ratios"
"Synchronous Radiocarbon and Climate Shifts during the Last Deglaciation"
"Upwelling Intensification as Part of the Pliocene-Pleistocene Climate Transition"
"Millennial-Scale Instability of the Antarctic Ice Sheet during the Last Glaciation"
"Cenozoic Deep-Sea Temperatures and Global Ice Volumes from Mg/Ca in Benthic Foraminiferal Calcite"
"Decadal Sea Surface Temperature Variability in the Subtropical South Pacific from 1726 to 1997 A.D."
"A High-Resolution Millennial Record of the South Asian Monsoon from Himalayan Ice Cores"
"Tropical Climate at the Last Glacial Maximum Inferred from Glacier Mass-Balance Modeling"

cluster # 6 has the following top 10 titles:
"Cloning and Heterologous Expression of the Epothilone Gene Cluster"
"A Basal Transcription Factor That Activates or Represses Transcription"
"Generation of G-to-A and C-to-U Changes in HIV-1 Transcripts by RNA Editing"
"A Drosophila Complementary DNA Resource"
"License Withheld: Geminin Blocks DNA Replication"
"p73: Guilt by Association?"
"Positioning of the Mitotic Spindle by a Cortical-Microtubule Capture Mechanism"
"From Sequence to Chromosome: The Tip of the X Chromosome of D. melanogaster"
"Distinct Roles for TBP and TBP-Like Factor in Early Embryonic Gene Transcription in Xenopus"
"Identification of a Cellular Cofactor Required for Infection by Feline Leukemia Virus"

cluster # 7 has the following top 3 titles:
"Fas Preassociation Required for Apoptosis Signaling and Dominant Inhibition by Pathogenic Mutations"
"A Domain in TNF Receptors That Mediates Ligand-Independent Receptor Assembly and Signaling"
"Regulated Cleavage of a Contact-Mediated Axon Repellent"

cluster # 8 has the following top 2 titles:
"Orbital Physics in Transition-Metal Oxides"
"Advances in the Physics of High-Temperature Superconductivity"

cluster # 9 has the following top 10 titles:
"Promiscuity and the Primate Immune System"
"Natural Selection and Parallel Speciation in Sympatric Sticklebacks"
"Evidence for DNA Loss as a Determinant of Genome Size"
"High Direct Estimate of the Mutation Rate in the Mitochondrial Genome of Caenorhabditis elegans"
"Cross-Species Interactions between Malaria Parasites in Humans"
"Modulation of Human Visual Cortex by Crossmodal Spatial Attention"
"Rapid Evolution of Reproductive Isolation in the Wild: Evidence from Introduced Salmon"
"Natural Selection and the Reinforcement of Mate Recognition"
"Polyploidy and the Evolution of Gender Dimorphism in Plants"
"Mate Selection and the Evolution of Highly Polymorphic Self/Nonself Recognition Genes"

```
cluster # 10 has the following top 1 titles:
"The Catalytic Pathway of Cytochrome P450cam at Atomic Resolution"

cluster # 11 has the following top 10 titles:
"Requirement of NAD and SIR2 for Life-Span Extension by Calorie
Restriction in Saccharomyces Cerevisiae"
"Suppression of Mutations in Mitochondrial DNA by tRNAs Imported
from the Cytoplasm"
"Distinct Classes of Yeast Promoters Revealed by Differential
TAF Recruitment"
"T Cell-Independent Rescue of B Lymphocytes from Peripheral Immune
Tolerance"
"Efficient Initiation of HCV RNA Replication in Cell Culture"
"Reduced Food Intake and Body Weight in Mice Treated with Fatty
Acid Synthase Inhibitors"
"Negative Regulation of the SHATTERPROOF Genes by FRUITFULL during
Arabidopsis Fruit Development"
"Patterning of the Zebrafish Retina by a Wave of Sonic Hedgehog Activity"
"Coupling of Stress in the ER to Activation of JNK Protein Kinases
by Transmembrane Protein Kinase IRE1"
"An Anti-Apoptotic Role for the p53 Family Member, p73, during
Developmental Neuron Death"

cluster # 12 has the following top 10 titles:
"A Light-Emitting Field-Effect Transistor"
"Tunable Resistance of a Carbon Nanotube-Graphite Interface"
"Forming Supramolecular Networks from Nanoscale Rods in Binary,
Phase-Separating Mixtures"
"Direct Observation of Dynamical Heterogeneities in Colloidal
Hard-Sphere Suspensions"
"Real-Space Imaging of Two-Dimensional Antiferromagnetism on the
Atomic Scale"
"A Quantum State-Resolved Insertion Reaction: <latex>$O(^{1}D)
+ H_{2} (J = 0) \rightarrow OH(^{2}\prod, v, N) + H(^{2}S)$</latex>"
"Optical Gain and Stimulated Emission in Nanocrystal Quantum Dots"
"Imaging the Electron Wave Function in Self-Assembled Quantum Dots"
"Ultrafast Electron Localization Dynamics Following Photo-Induced
Charge Transfer"
"Moissanite: A Window for High-Pressure Experiments"
```

The usefulness of this algorithm is to measure the similarity between articles and to cluster similar articles. As an example, we can see that Cluster 1 has a lot of articles related to "Corrections and Clarifications". In addition, from the results of getting the top 10 words per cluster, we can also learn about the commonly used words in this cluster of articles.

(b) We loaded the data and clustered the data using KMeans, using different k values for $5 <= k <= 20$. We then selected $k = 11$. After that, we computed the 1373-vector average value for the data and generated the top 10 articles of each cluster based on the 1373-vector average value as well as the center for each cluster. Below are the results:

```
cluster # 0 has the following top 10 titles:
"A Mouse Chronology"
"Meltdown on Long Island"
"Atom-Scale Research Gets Real"
"Presidential Forum: Gore and Bush Offer Their Views on Science"
"Help Needed to Rebuild Science in Yugoslavia"
"I'd like to See America Used as a Global Lab"
"Soft Money's Hard Realities"
"Silent No Longer: 'Model Minority' Mobilizes"
"Ecologists on a Mission to Save the World"
"Clones: A Hard Act to Follow"

cluster # 1 has the following top 10 titles:
"NEAR at Eros: Imaging and Spectral Results"
"Reduction of Tropical Cloudiness by Soot"
"Internal Structure and Early Thermal Evolution of Mars from Mars
Global Surveyor Topography and Gravity"
"The Atom-Cavity Microscope: Single Atoms Bound in Orbit by Single
Photons"
"Climate Extremes: Observations, Modeling, and Impacts"
"Calcium Sensitivity of Glutamate Release in a Calyx-Type Terminal"
"Experiments and Simulations of Ion-Enhanced Interfacial Chemistry
on Aqueous NaCl Aerosols"
"Sediments at the Top of Earth's Core"
"The Elemental Composition of Asteroid 433 Eros: Results of the
NEAR-Shoemaker X-ray Spectrometer"
"Advances in the Physics of High-Temperature Superconductivity"

cluster # 2 has the following top 10 titles:
"National Academy of Sciences Elects New Members"
"Biological Control of Invading Species"
"Corrections and Clarifications: One Hundred Years of Quantum
Physics"
"Corrections and Clarifications: Biotech Research Proves a Draw
in Canada"
"Corrections and Clarifications: A Nuclear Solution to Climatic
Change?"
"Corrections and Clarifications: Uninterrupted MCM2-7 Function
Required for DNA Replication Fork Progression"
```

"Corrections and Clarifications: Timing the Ancestor of the HIV-1 Pandemic Strains"
"Corrections and Clarifications: Absorbing Phenomena"
"Limbless Tetrapods and Snakes with Legs"
"Corrections and Clarifications: A Short Fe-Fe Distance in Peroxodiferric Ferritin: Control of Fe Substrate versus Cofactor Decay?"

cluster # 3 has the following top 10 titles:
"Three-Dimensional Structure of the Tn5 Synaptic Complex Transposition Intermediate"
"The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 <latex>$\AA$</latex> Resolution"
"Architecture of RNA Polymerase II and Implications for the Transcription Mechanism"
"The Structural Basis of Ribosome Activity in Peptide Bond Synthesis"
"Crystal Structure of the Ribonucleoprotein Core of the Signal Recognition Particle"
"Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor"
"Structure of the Light-Driven Chloride Pump Halorhodopsin at 1.8 <latex>$\AA$</latex> Resolution"
"Structure of the S15,S6,S18-rRNA Complex: Assembly of the 30S Ribosome Central Domain"
"Crystal Structure of gd T Cell Receptor Ligand T22: A Truncated MHC-like Fold"
"Structural Basis of Smad2 Recognition by the Smad Anchor for Receptor Activation"

cluster # 4 has the following top 10 titles:
"Advances in the Physics of High-Temperature Superconductivity"
"The Atom-Cavity Microscope: Single Atoms Bound in Orbit by Single Photons"
"Orbital Physics in Transition-Metal Oxides"
"Quantum Criticality: Competing Ground States in Low Dimensions"
"Self-Mode-Locking of Quantum Cascade Lasers with Giant Ultrafast Optical Nonlinearities"
"Generating Solitons by Phase Engineering of a Bose-Einstein Condensate"
"Imaging Precessional Motion of the Magnetization Vector"
"Negative Poisson's Ratios for Extreme States of Matter"
"Ultrafast Mid-Infrared Response of <latex>$YBa_2Cu_3O_{7-\delta}$</latex>"
"Optically Induced Entanglement of Excitons in a Single Quantum Dot"

cluster # 5 has the following top 10 titles:
"Positional Syntenic Cloning and Functional Characterization of the Mammalian Circadian Mutation tau"
"Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator

```
of p53-Induced Apoptosis"
"Role of the Mouse ank Gene in Control of Tissue Calcification
and Arthritis"
"Comparative Genomics of the Eukaryotes"
"Regulated Cleavage of a Contact-Mediated Axon Repellent"
"Protein Interaction Mapping in C. elegans Using Proteins Involved
in Vulval Development"
"Crystal Structure of the Ribonucleoprotein Core of the Signal
Recognition Particle"
"Candidate Taste Receptors in Drosophila"
"Gridlock, an HLH Gene Required for Assembly of the Aorta in Zebrafish"
"Cloning of the Arabidopsis Clock Gene TOC1, an Autoregulatory
Response Regulator Homolog"

cluster # 6 has the following top 10 titles:
"Central Role for G Protein-Coupled Phosphoinositide 3-Kinase g
in Inflammation"
"Function of PI3Kg in Thymocyte Development, T Cell Activation,
and Neutrophil Migration"
"Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator
of p53-Induced Apoptosis"
"Role of the Mouse ank Gene in Control of Tissue Calcification
and Arthritis"
"Requirement for RORg in Thymocyte Survival and Lymphoid Organ
Development"
"Requirement of JNK for Stress-Induced Activation of the Cytochrome
c-Mediated Death Pathway"
"Kinesin Superfamily Motor Protein KIF17 and mLin-10 in NMDA
Receptor-Containing Vesicle Transport"
"An Oral Vaccine against NMDAR1 with Efficacy in Experimental
Stroke and Epilepsy"
"Synaptic Assembly of the Brain in the Absence of Neurotransmitter
Secretion"
"Positional Syntenic Cloning and Functional Characterization of
the Mammalian Circadian Mutation tau"

cluster # 7 has the following top 10 titles:
"Status and Improvements of Coupled General Circulation Models"
"Sedimentary Rocks of Early Mars"
"Climate Extremes: Observations, Modeling, and Impacts"
"Causes of Climate Change over the past 1000 Years"
"A 22,000-Year Record of Monsoonal Precipitation from Northern
Chile's Atacama Desert"
"Coherent High- and Low-Latitude Climate Variability during the
Holocene Warm Period"
```

"Rapid Changes in the Hydrologic Cycle of the Tropical Atlantic during the Last Glacial"
"Internal Structure and Early Thermal Evolution of Mars from Mars Global Surveyor Topography and Gravity"
"Climate Impact of Late Quaternary Equatorial Pacific Sea Surface Temperature Variations"
"Interpreting Differential Temperature Trends at the Surface and in the Lower Troposphere"

cluster # 8 has the following top 10 titles:
"Function of PI3Kg in Thymocyte Development, T Cell Activation, and Neutrophil Migration"
"Central Role for G Protein-Coupled Phosphoinositide 3-Kinase g in Inflammation"
"Rapid Destruction of Human Cdc25A in Response to DNA Damage"
"An Oral Vaccine against NMDAR1 with Efficacy in Experimental Stroke and Epilepsy"
"Subgroup of Reproductive Functions of Progesterone Mediated by Progesterone Receptor-B Isoform"
"Regulated Cleavage of a Contact-Mediated Axon Repellent"
"Cross Talk between Interferon-g and -a/b Signaling Components in Caveolar Membrane Domains"
"Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles"
"Response of Schwann Cells to Action Potentials in Development"
"Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator of p53-Induced Apoptosis"

cluster # 9 has the following top 10 titles:
"Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles"
"Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis among Eukaryotes"
"Technique for Enhanced Rare Earth Separation"
"A Sense of the End"
"Bone Resorption by Osteoclasts"
"The Osteoblast: A Sophisticated Fibroblast under Central Surveillance"
"Docosahexaenoic Acid, a Ligand for the Retinoid X Receptor in Mouse Brain"
"A Specificity-Enhancing Factor for the ClpXP Degradation Machine"
"Convergent Solutions to Binding at a Protein-Protein Interface"
"NF-kB-Induced Loss of MyoD Messenger RNA: Possible Role in Muscle Decay and Cachexia"

cluster # 10 has the following top 10 titles:
"Molecular Computation by DNA Hairpin Formation"

```
"Redox Signaling in Chloroplasts: Cleavage of Disulfides by an
Iron-Sulfur Cluster"
"Cholera Dynamics and El Nino-Southern Oscillation"
"Timing the Ancestor of the HIV-1 Pandemic Strains"
"Complete Genome Sequence of Neisseria meningitidis Serogroup
B Strain MC58"
"Invariants, Scaling Laws, and Ecological Complexity"
"Molecular Architecture and Evolution of a Modular Spider Silk
Protein Gene"
"Motion Integration and Postdiction in Visual Awareness"
"Rapid Evolution of Reproductive Isolation in the Wild: Evidence
from Introduced Salmon"
"The Benefits of Allocating Sex"
```

After that, we calculated the top 10 words that are closest to the centers of each cluster. Below are the results (note that certain clusters have less than 10 words - in that case, all words of that cluster is reported):

```
cluster # 0 has the following top 10 words: ['recalls', 'clinton',
'geneticist', 'security', 'fight', 'prize', 'spending', 'campaign',
'hes', 'rights']

cluster # 1 has the following top 10 words: ['start', 'decrease',
'error', 'magnitude', 'peak', 'maximum', 'fraction', 'constant',
'comparison', 'measurements']

cluster # 2 has the following top 10 words: ['aptamers', 'lcts',
'dnag', 'doxy', 'proteorhodopsin', 'trxr', 'lg268', 'neas',
'rory', 'nompc']

cluster # 3 has the following top 10 words: ['groove', 'ribbon',
'conformations', 'pocket', 'refinement', 'refined', 'helices',
'disordered', 'helical', 'reflections']

cluster # 4 has the following top 10 words: ['excitations',
'insulating', 'spins', 'resonant', 'coherence', 'fermi', 'anisotropic',
'doped', 'anisotropy', 'semiconductor']

cluster # 5 has the following top 10 words: ['polymerase', 'amino',
'mutation', 'conserved', 'acids', 'mutations', 'transcription',
'mutant', 'vitro', 'terminal']

cluster # 6 has the following top 10 words: ['triton', 'cytometry',
'isoforms', 'mab', 'homeostasis', 'glutathione', 'luciferase',
'lysis', 'agarose', 'immunofluorescence']
```

```
cluster # 7 has the following top 10 words: ['interglacial', 'clim',
'upwelling', 'interannual', 'decadal', 'moisture', 'holocene',
'aerosols', 'albedo', 'proxy']

cluster # 8 has the following top 10 words: ['natl', 'acad',
'assay', 'vivo', 'antibody', 'receptors', 'mediated', 'proc',
'biol', 'activated']

cluster # 9 has the following top 1 words: ['factor']

cluster # 10 has the following top 1 words: ['variable']
```

The usefulness of this algorithm is to measure the similarity between words and to cluster similar words. As an example, we can see that Cluster 0 has a lot of words related to political and social issues. The main difference between this clustering algorithm and the previous clustering algorithm (that clusters documents) is the difference in features for each datapoint - when we are clustering words, each feature of the dataset is related to one document, whereas in the algorithm for clustering documents, each feature of the dataset refers to one word.

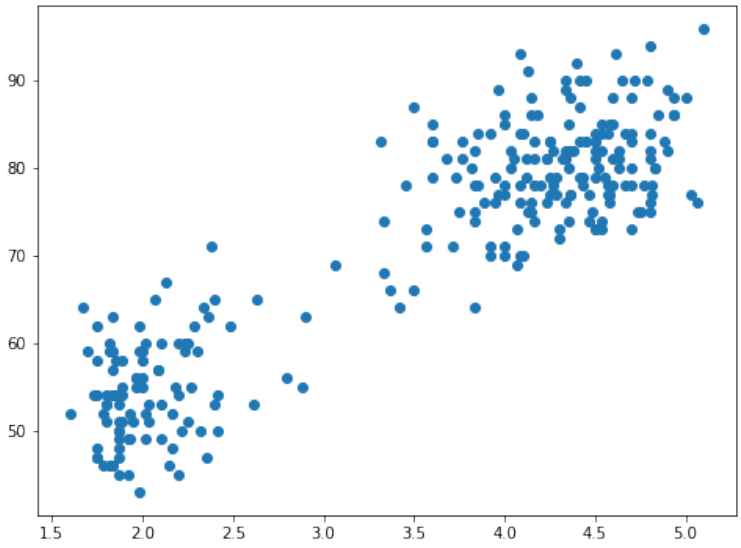
- Question 3

  (a) The alternating algorithm for KMeans first computes the assigns the points to the nearest center, then recalculates the positions for each cluster center given the points in that cluster. The first step of KMeans is the Expectation step and the second step is the Maximization step. Thus, the alternating algorithm for KMeans is a special case of the EM algorithm.

  The objective function for E step is: $C_i = argmin(\|\mu_j - x_i\|)$ for all $0 <= j <= k$ and for all $0 <= i <= N$, where C_i is the assigned center for point x_i, μ_j is the generated center, k is the number of centers and N is the number of datapoints.

  The objective function for M step is: $\mu_j = \frac{1}{Count(S_j)} \sum_{x_i \in S_j} x_i$ for all $0 <= j <= k$, where S_j is the cluster j and $Count(S_j)$ is the number of datapoints in the cluster.

  (b) We loaded and parsed the data per the instructions. Below is the scatterplot graph for all the datapoints:

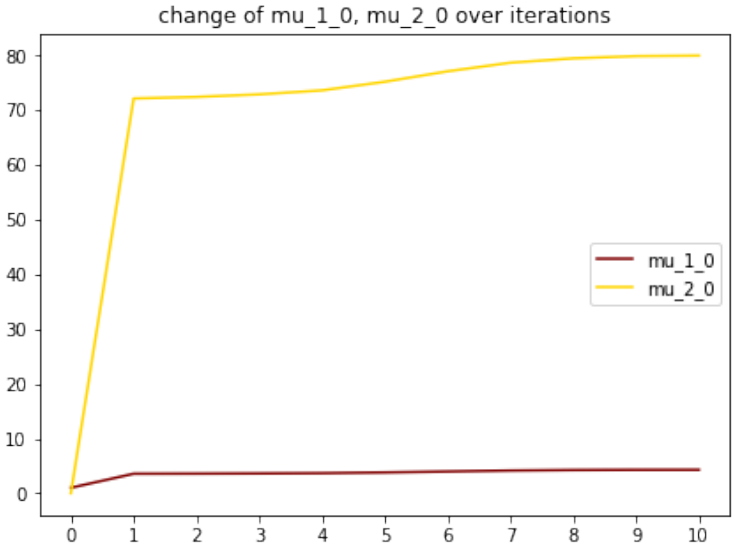


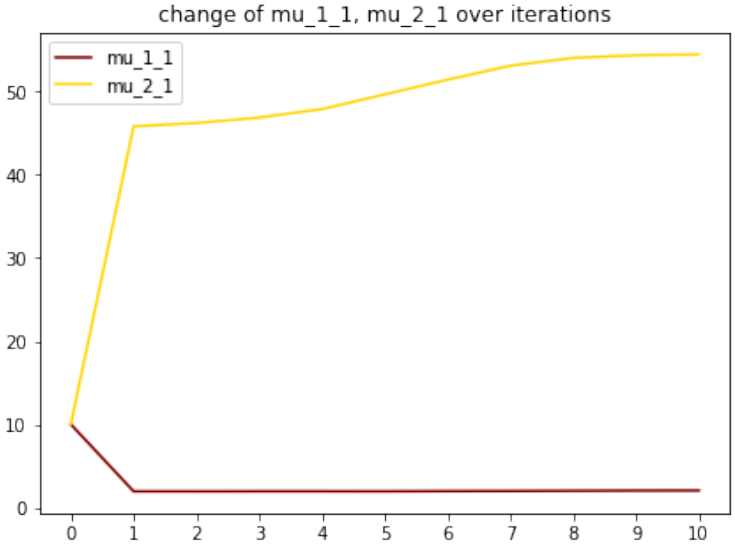
(c) We implemented a bimodal GMM model using EM algorithm per the instructions. The stopping condition we specified is as follows: stop when the difference of the sum of means of the newly generated clusters and the sum of the means of the previously generated clusters is within ± 0.001, which translates to the following python code:

```
diff = np.absolute((np.asarray(new_mean) - np.asarray(mean)).sum())
if diff < 0.001:
    end = True
    break
```

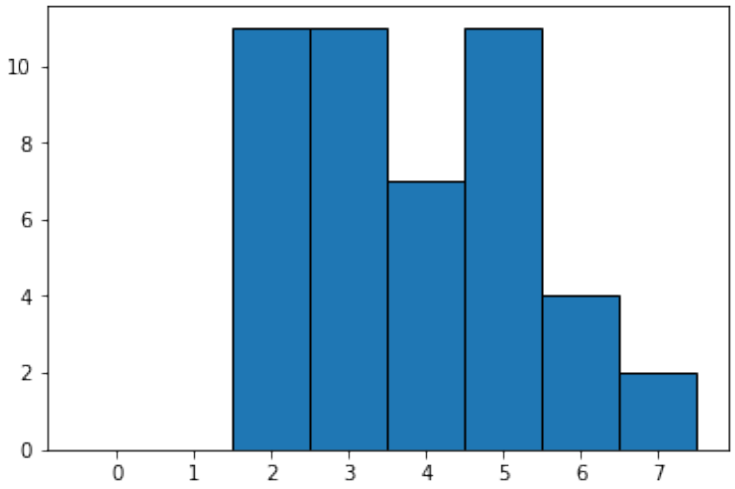
The reasoning behind this termination condition is that we consider the algorithm as having converged if the clusters remain unchanged. We calculated the difference between the means of the clusters as a way to measure whether the clusters have changed or not.

We ran the algorithm with a randomly generated initial mean, variance and prior probability. The graphs that plot the trajectories of the two mean vectors is shown below:

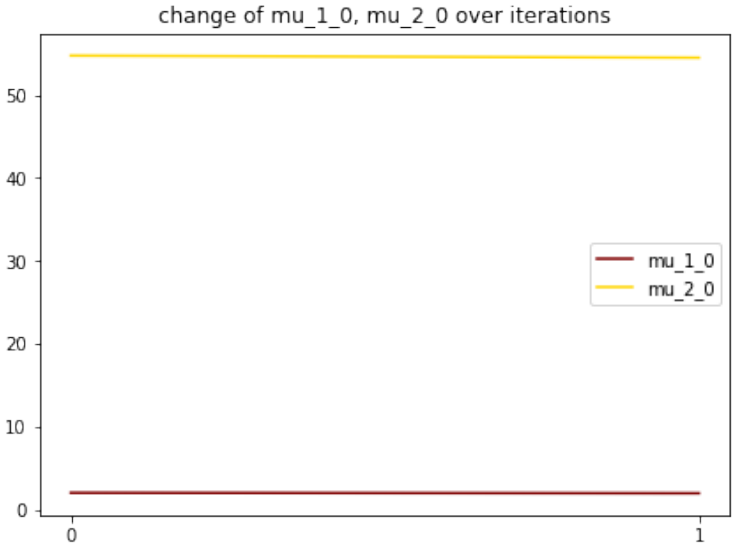


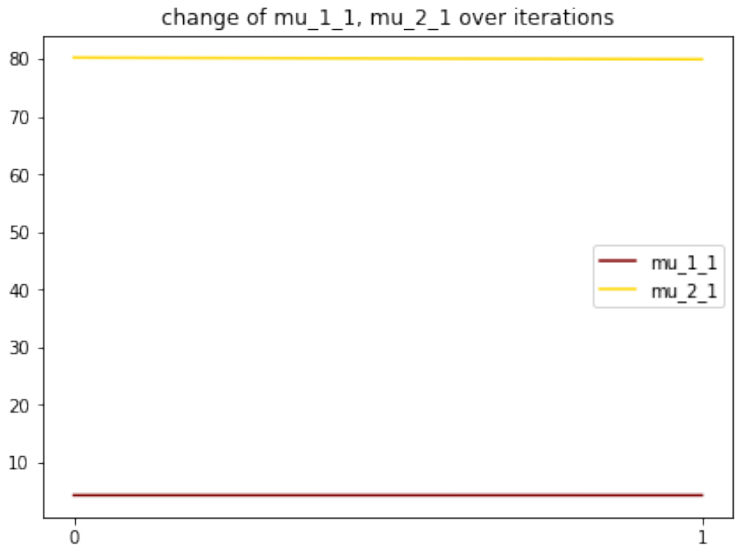


We then ran the algorithm with 50 different randomly generated initial parameters. Below is a histogram that plots the number of iterations required for convergence:



(d) We first used KMeans to predict class membership for each point; we then used the predictions to generate initial parameters for EM. The graphs that plot the trajectories of the two mean vectors is shown below:





We can see that if we generate initial parameters with KMeans, only 1 iteration is needed for the algorithm to converge, as opposed to an average of 4 iterations for randomly generated initial parameters. We can also see that the mean vectors does not change very much in this way. Thus, it seems that KMeans may be a good way of generating initial parameters for GMM and that fitting the data with KMeans first allows us to use GMM in a more efficient way.

2 Written Exercises

- Question 1

  (a)

$$MM^T = \begin{vmatrix} 10 & 9 & 26 & 3 & 26 \\ 9 & 62 & 8 & -5 & 85 \\ 26 & 8 & 72 & 10 & 50 \\ 3 & -5 & 10 & 2 & -1 \\ 26 & 85 & 50 & -1 & 138 \end{vmatrix}$$

$$M^TM = \begin{vmatrix} 39 & 57 & 60 \\ 57 & 118 & 53 \\ 60 & 53 & 127 \end{vmatrix}$$

  (b) The eigenvalues of M^TM and MM^T are: 214.67 and 69.33.

  (c)

  Below are the eigenvectors of M^TM:

$$\begin{vmatrix} -0.426 \\ -0.615 \\ -0.663 \end{vmatrix}$$

and

$$\begin{vmatrix} -0.015 \\ -0.729 \\ 0.685 \end{vmatrix}$$

Below are the eigenvectors of MM^T:

$$\begin{vmatrix} -0.165 \\ -0.472 \\ -0.336 \\ -0.003 \\ -0.798 \end{vmatrix}$$

and

$$\begin{vmatrix} 0.245 \\ -0.453 \\ 0.829 \\ 0.170 \\ -0.133 \end{vmatrix}$$

(d) We have $M = U\Sigma V^T$, where:

$$U = \begin{vmatrix} -0.165 & 0.245 \\ -0.472 & -0.453 \\ -0.336 & 0.829 \\ -0.003 & 0.170 \\ -0.798 & -0.133 \end{vmatrix}$$

(whose columns are consisted of the eigenvectors of MM^T)

and

$$\Sigma = \begin{vmatrix} 14.652 & 0 \\ 0 & 8.326 \end{vmatrix}$$

(which is a diagonal matrix generated by filling in the square roots of the eigenvalues of M^TM and MM^T)

and

$$V^T = \begin{vmatrix} -0.426 & -0.615 & -0.663 \\ -0.015 & -0.729 & 0.685 \end{vmatrix}$$

(which is the transpose of matrix V, whose columns are consisted of the eigenvectors of M^TM)

(e)

We have $M_{estimate} = U'\Sigma'V^{T'}$, where:

$$U' = \begin{vmatrix} -0.165 \\ -0.472 \\ -0.336 \\ -0.003 \\ -0.798 \end{vmatrix}$$

and

$$\Sigma' = \begin{vmatrix} 14.652 \end{vmatrix}$$

and

$$V^{T'} = \begin{vmatrix} -0.426 & -0.615 & -0.663 \end{vmatrix}$$

Thus, we have:

$$M_{estimate} = \begin{vmatrix} 1.03 & 1.49 & 1.60 \\ 2.94 & 4.25 & 4.58 \\ 2.10 & 3.03 & 3.27 \\ 0.02 & 0.03 & 0.03 \\ 4.98 & 7.19 & 7.76 \end{vmatrix}$$

- Question 2

  Given that the linear transformation $X^S = (X - \overline{x})\hat{\Sigma}^{-1/2}$ results in standardized data X^S with a covariance matrix $(X^S)^T X^S$ that is an identity matrix, the principal components of X^S are eigenvectors of the covariance matrix $(X^S)^T X^S$, which are $[1, 0, 0, ..., 0]$, $[0, 1, 0, ..., 0]$, ..., $[0, 0, 0, ..., 1]$. In addition, the eigenvalues of the covariance matrix are all 1, meaning that the variation of each component is the same. By standardizing the data, we can make sure all variables of the data are on the same scale. Since PCA is sensitive to the scale of variables, standardizing the data makes the PCA model less biased towards features whose values are large in scale, thus making the model more accurate in general.

3 References

- (rank-r approximation error) `https://www.youtube.com/watch?v=c7e-D2tmRE0`
- (EM algorithm) `http://www.cs.cmu.edu/~aarti/Class/10701_Spring14/slides/EM_annotatedonclass.pdf`
- (EM algorithm) `https://www.youtube.com/watch?v=iQoXFmbXRJA`
- (SVD) `https://www.cse.unr.edu/~bebis/CS791E/Notes/SVD.pdf`

- (PCA) https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
- (PCA) https://builtin.com/data-science/step-step-explanation-principal-component-analysis