EE 127 Optimization Models in Engineering
Fall 2018 Nithin Raghavan

HW 2

1 PCA and low-rank compression

1. If the squared distance between two values $||z-x||_2^2$ is minimized, this is by definition a projection. In order to find the minimum value of v, then consider the following. $v^* = \arg\min_v ||z-x||_2^2 = \arg\min_v ||x_0 + vu - x||_2^2$. Expanding this, we get $\arg\min_v ||x_0||_2^2 + 2vx_0^T u + v^2||u||_2^2 - ||x||_2^2 - 2x_0^T x - 2vx^T u$. Because x_0 and x are constant in this expression, they act as translating factors and so we can ignore them. Now, the expression simplifies to $\arg\min_v v^2 - 2vx^T u + 2vx_0^T u$. To find the minimum, then it suffices to solve for the derivative when it equals zero and verify that it is a minimum. This is $2v - 2x^T u + 2x_0^T u = 0 \implies v^* = x^T u - x_0^T u = (x-x_0)^T u$, the value specified in the problem. This is a minimum because the graph's second derivative is 2, which is concave up. Additionally, the minimal squared distance $z - x$ is orthogonal to $\mathscr{L}$, because z minimizes the distance and so the angle formed by dropping $z-x$ onto $\mathscr{L}$ is perpendicular. So because of this, the three quantities $z-x$, $\mathscr{L}$ and $x - x_0$ form a right triangle, and the lengths can be related by the Pythagorean theorem. Because $z - x \perp \mathscr{L} = x_0 + (x-x_0)^T uu$, then $x - x_0$ forms the hypotenuse; u is normally unit length so its scaling factor $(x-x_0)^T u$ gives the proper length. And so $||x-x_0||_2^2 = ((x-x_0)^T u)^2 + ||z-x||_2^2 \implies ||z-x||_2^2 = ||x-x_0||_2^2 - ((x-x_0)^T u)^2$.

2. Maximizing the variance of points projected on a line $\mathscr{L} = x_0 + vu$ going through the origin was shown to have projection or minimal squared distance when $v = (x-x_0)^T u$. So, maximizing the variance of v for a point x can be done as follows: $\text{var}(v) = \text{var}((x-x_0)^T u) = \text{var}(u^T(x-x_0)) = \frac{1}{n}\sum_{i=0}(u^T x_i - u^T \hat{x})^2 = u^T Cu$. In order to maximize $\text{var}(v)$, then $\hat{x}$ must be zero, and so $\frac{x_1 + \cdots + x_n}{n} = 0 = \hat{x}$. Thus, $\max \text{var}(v)$ is identical to maximizing the covariance matrix C in the form $\max_u u^T Cu$. Minimizing the sum of squares of the distances from points to their projections on $\mathscr{L}$ is equivalent to minimizing the total reconstruction error, which in PCA is the same as maximizing the variance v along $\mathscr{L}$, which is just $\max_u u^T Cu$ (proof: if the smallest reconstruction error did not correspond with the maximum variance, then there would exist a point x whose reconstruction error $||e||_2 < ||\text{proj}_{\mathscr{L}} x||_2$. But this is a contradiction, because by definition the projection must minimize the squared distance between two values, and if e were smaller, then it would minimize it better). So, P1 is equivalent to P2.

3. If all the points are on one line that goes through the origin, then all the points will be scalar multiples with respect to one vector, and so a matrix made up of these points will have rank one. Likewise, in the backwards direction, if a matrix has rank one, then all of its columns minus one are linearly dependent, and are a scalar multiple of some v. Plotting v at the origin is equivalent to plotting a line on which all the points created by all the columns of the matrix are located. So because a rank one matrix is equivalent to finding a line on which all points are on it, then the variance is automatically maximized. So finding a rank one approximation is therefore equivalent to finding the maximum variance of a set of points, which is P1. therefore, P1 and P3 are equivalent.

2 Generalized Eigenvalues and Image Segmentation

1. This is on the Jupyter notebook.

2. W is a symmetric matrix by definition (the edges are undirected between each pair of vertices). It has a nonnegative quadratic form because all the values are positive, and so it is positive semidefinite. And because D is formed as a diagonal from the sum of W's weights of edges incident on each pixel, each component of $\text{tr}(D)$ is positive as well, so D is also PSD. To solve the problem $(D - W)y = \lambda Dy$, it can be noted that in $D^{-1}(D-W)y = \lambda y$, $D^{-1}(D-W)$ is symmetric, so this forms the symmetric eigenvalue problem. Forming an eigenvalue decomposition in this case is identical to SVD, and so could be solved provided code to compute an SVD.

3. This is on the Jupyter notebook.

4. This is on the Jupyter notebook. The results are that the image is binarized. A way to better segment the image is to recursively divide the image in half and implement the binary partition using the second eigenvector on each division. This would have a faster runtime.

3 PCA and Senate Voting Data

1. This can be done by realizing that each iteration of PCA maximizes variation along an axis; thus, centering the matrix (by subtracting the mean from each component) provides the scalar required to obtain the first principal component, or the direction that maximizes variance. Adding a scalar to all terms in a dataset does not affect the variance (proof: $\text{var}(X) = \frac{1}{n}\sum(x_i - \hat{x})^2, \hat{x} = \frac{x_1 + \cdots + x_n}{n}$. Adding a scalar s to each x_i implies $\hat{x}$ increases by s. Then, $\frac{1}{n}\sum((x_i + s) - (\hat{x} + s))^2 = \frac{1}{n}(x_i - \hat{x})^2 = \text{var}(X)$). So, b is arbitrary. The value of a is given by $a = \arg\max_{||a||=1} ||Xa||^2$, which can be found through PCA. X refers to the centred dataset of senators as rows and bills as columns.

2. The variance is much lower in the second one; on the order of 6 instead of 27. The phrase 'senators vote according to the party average' is apt because most senators will vote in accordance with two blocks: one slightly to the left and one slightly to the right. The outliers on both sides will increase this variance. However, it being relatively low shows that most senators vote close to the center, at the party average.

3. The total variance explained by the first principal component is the overall political ideology of each senator, with a variation of around 27. And the second refers to the overall spread of viewpoints, with a variation of around 6, showing that most are centrist. The plot is on Jupyter notebook.

4. The bills that are the most important are the ones which have a larger spike in PCA variations, because they are the more controversial ones. These include 'National Security Issues_Anti-terrorism Wireta...'. The more extreme senators are those who have a higher variance from the center, which include senator 25 and 93.

4 Diet Planning

1. First, it can be noted that n_{des} is just the nutrient vector of each food multiplied by amount of grams contained in each component of the diet d. This is because $\begin{bmatrix} N_{11} & N_{12} & \cdots & N_{1n} \\ N_{21} & \ddots & \vdots & \vdots \\ N_{31} & \dots & \ddots & \vdots \\ N_{m1} & \cdots & \cdots & N_{mn} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} N_{1i} d_i \\ \vdots \\ \vdots \\ \sum_{i=1}^{n} N_{mi} d_i \end{bmatrix} = n_{des}$, a matrix of the proper size, and is just the sum of nutrients multiplied by the quantity of each different food per diet. The kth row of this matrix is therefore the total amount of the kth nutrient in the jth food. As for the cost constraint, if cost is represented by c, then because each quantity of food is given by d, the total cost constraint is given by $c^T d = B$.

2. The constraints are from $||d - s_i||_2 = p_i, 1 \leq i \leq n+1$. Squaring both sides and factoring, we obtain $||d - s_i||_2^2 - p_i^2 = 0 \implies \sum_{k=1}^{n} |(d - s_i)_j|^2 - p_i^2 = 0$. This is just the dot product of $(d - s_i)^T (d - s_i)$, which equals p_i^2. This is similar to the original least-squares problem. We need to find n hyperplanes as well as a quadratic constraint; this quadratic constraint constrains the n hyperplanes that d is represented by, while n is the amount of the set of discrete intersections of these hyperplanes.

5 A result related to Gaussian distributions

This represents a Gaussian in multiple dimensions. The probability density function of a Gaussian is $f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. In order to utilize this, we notice that if Σ is positive definite, then Σ^{-1} is positive-definite as well (proof: because Σ is positive definite, it is also invertible; let $\Sigma x = y \implies y^T\Sigma^{-1}y = (\Sigma x)^T\Sigma^{-1}(\Sigma x) = x^T(\Sigma^T\Sigma^{-1})\Sigma x = x^T\Sigma x > 0$). It is also noted that the square root of a positive definite matrix is also positive definite (proof: as it is symmetric, it can be diagonalised as $\Sigma = PD^2P^T$; the eigenvalues along the diagonal of D are all real and positive, so their square roots are as well. The square root is symmetric because $(PDP^T)^T = PD^TP^T = PDP^T$, and it is a square root because $(PDP^T)(PDP^T) = PD^2P^T = \Sigma$, by squaring the eigenvalues along the diagonal of D). So, the value in the exponent of the integrand will always resolve to a negative number. Also note that the determinant of a positive definite matrix is just the product of its eigenvalues; if it is represented diagonally, then Σ's eigenvalues are just the components along the diagonal. Let $x_1, \cdots, x_n$ represent each component of x, and let $[z_1, \cdots, z_n]^T$ represent the PDFs $z_i \sim \mathcal{N}(\mu = 0, \sigma_i)$ of the Gaussian for each dimension; furthermore, let $x = \sigma^{1/2}z$. Because each component of z is Gaussian, the joint distribution of z is therefore given through the multiplication of all of them, by $f(z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i^{1/2}} \exp\left(-\frac{z_i^T z_i}{2\sigma_i}\right) = \frac{1}{(2\pi)^{n/2}\left(\prod \sigma_i\right)^{1/2}} \exp\left(-\frac{1}{2}\sum \frac{z_i^2}{\sigma_i}\right) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}\sum \frac{(\Sigma^{-1/2}x)^T(\Sigma^{-1/2}x)}{\sigma_i}\right) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$. Then, because the result holds true for $n = 1$ that the integral of $p(x)$ is 1, then because n is a constant, it can be taken out of the integral. $\int_{\mathbb{R}^n} e^{-\frac{1}{2}x^T\sigma^{-1}x} = (2\pi)^{n/2}\sqrt{\det\Sigma}\underbrace{\int \cdots \int}_{n \text{ times}} p(x)dx_1 \cdots dx_n = (2\pi)^{n/2}\sqrt{\det\Sigma}(1)^n = (2\pi)^{n/2}\sqrt{\det\Sigma}$.

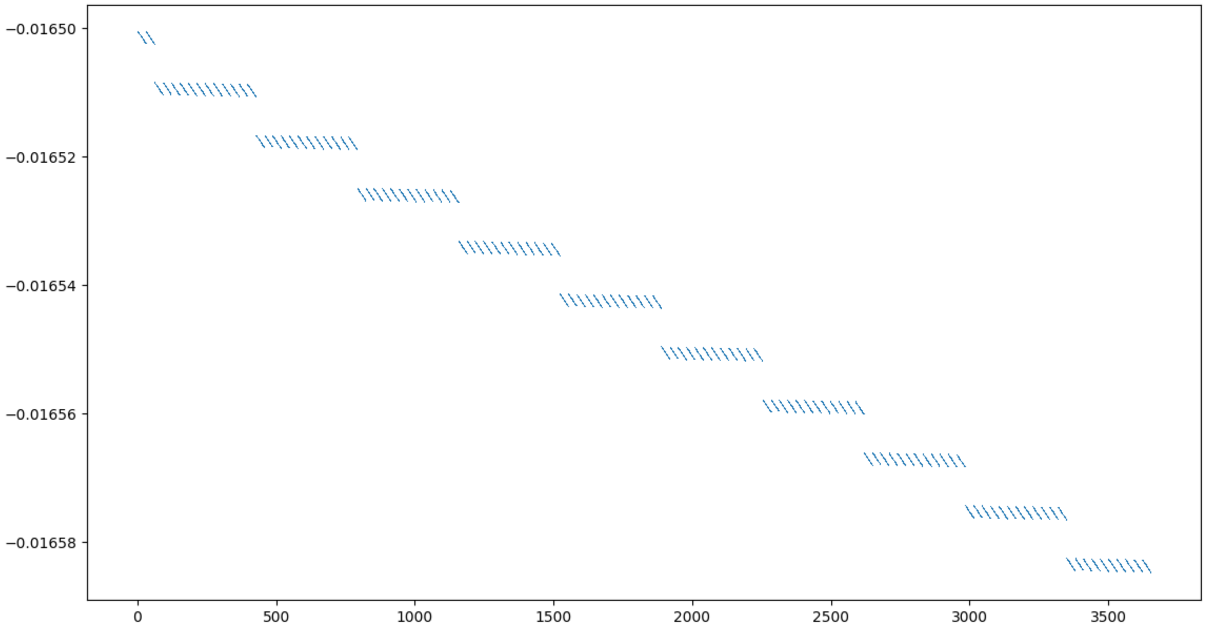
6 Analysis of calendar days

1. $X = \begin{bmatrix} 2003 & 11 & 1 \\ 2003 & 11 & 2 \\ \vdots & \vdots & \vdots \\ 2013 & 10 & 30 \\ 2013 & 10 & 31 \end{bmatrix}$

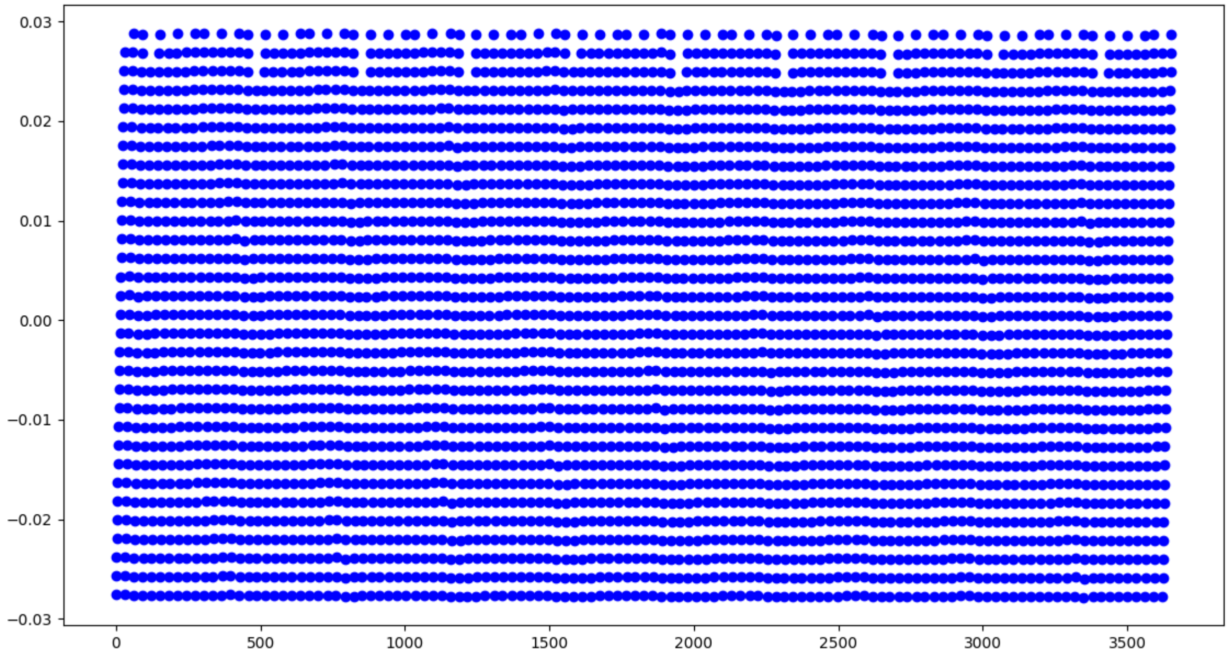
2. The right singular vectors are $\left(\begin{bmatrix} -0.99996405 \\ -0.00324731 \\ -0.00783294 \end{bmatrix}, \begin{bmatrix} -0.00784854 \\ 0.00481823 \\ 0.99995759 \end{bmatrix}, \begin{bmatrix} 0.00320943 \\ -0.99998312 \\ 0.00484354 \end{bmatrix}\right)$ and each vector corresponds with $121388.2190513, 531.91556131$ and 208.48354413, respectively.

3. The rank of X is 3. I would not say that this is low because the rank of X cannot be higher than 3 due to it being full rank.

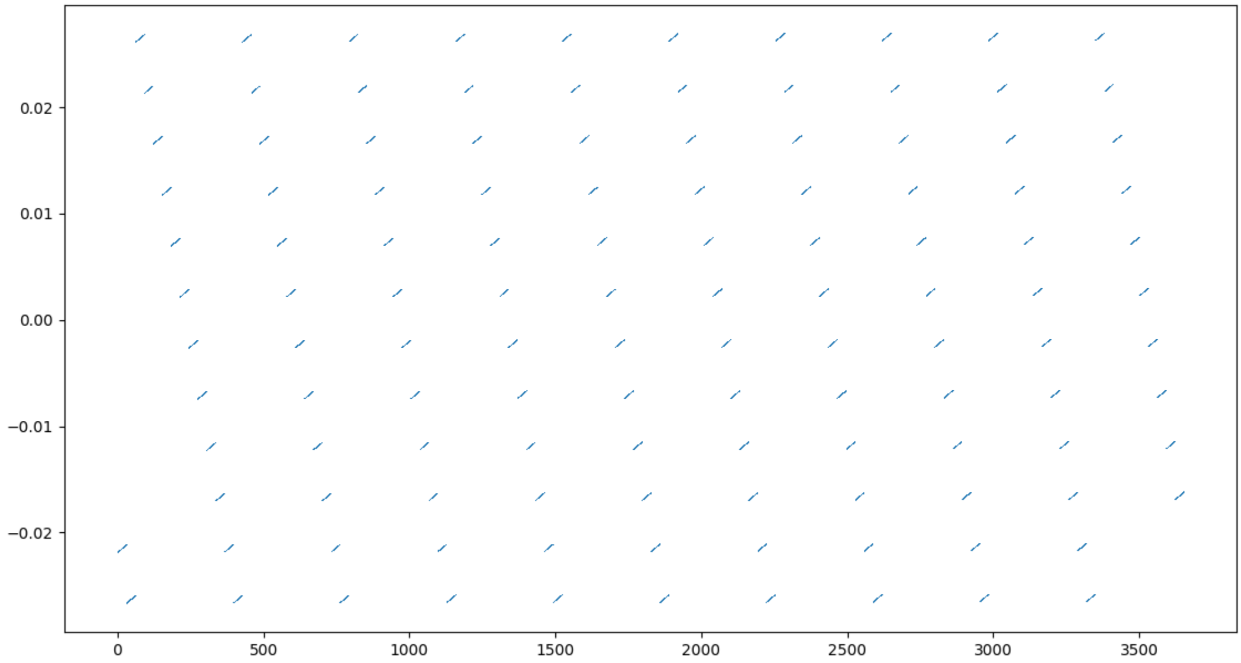
4. The plot of the first left singular vector is :



The plot of the second is:



The plot of the third is:



The patterns are as follows: for the first, there are 12 rows that are formed. These rows represent the years, and the diagonal lines in each row represent the months, of which the individual points that make them up are the days. Next, for the second, there are 31 rows that represent the days. Each column represents a new month, and the new year begins whenever there is a break in the upper portions of the graph. Lastly, for the third, each diagonal line represents a month, and up to 12 of them form a larger diagonal line going to the right that describes each year. The individual points represent the days again.

5. The smallest right singular vector essentially represents the column with the smallest rate of changes in the matrix. This would represent the year. Approximating this vector can be done by noting that the columns of V are just the eigenvectors of X^*X. Because X is real-valued, $X^*X = X^TX$; finding an eigenvalue decomposition of this matrix manually and selecting the rightmost column of the right-eigenvector matrix will produce the vector.

6. I would expect this value to stay essentially constant. This is because the change in number of days is usually in constant proportion to the month change as well as the year change overall.