1 Rayleigh Quotients

(a) The Rayleigh quotient is defined as

$$R(M, x) = \frac{x^T M x}{x^T x}$$

for a given symmetric matrix $M \in \mathbb{R}^{n \times n}$. What is the interval of possible values of the Rayleigh quotient for a given matrix? Specifically what is

$$\min_x R(M, x) \quad \text{and} \quad \max_x R(M, x)?$$

What values of x attain the bounds?

(b) How does the Rayleigh quotient relate to the following optimization problems? What does this tell us about the optimum values and the vectors which achieve them? Try to relate these quantities to the singular values and singular vectors of X.

$$\min_{w:\|w\|_2=1} \|Xw\|_2^2 \quad \text{and} \quad \max_{w:\|w\|_2=1} \|Xw\|_2^2.$$

2 The Gaussian MLE Perspective

(a) Assume our data matrix $X \in \mathbb{R}^{n \times d}$ is mean centered. What is the mean and variance of the maximum likelihood estimate for a Gaussian distribution fitting our dataset?

const wrt μ, var wrt Σ — var wrt Σ, μ

$$\ell(\mu, \Sigma; x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} \ln(f(x_i)) = \frac{nd}{2}\ln\left(\frac{1}{2\pi}\right) + \frac{n}{2}\ln|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

const wrt μ, Σ

$$\nabla_{\Sigma^{-1}}\ell = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T(x_i-\mu) = 0$$

$$\frac{d}{d(1/\sigma)}\ln\left(\frac{1}{\sigma}\right) = \sigma$$

$$\Rightarrow \Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^T(x_i-\mu)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i = \boxed{\frac{1}{n}X^T X}.$$

(b) Given this Gaussian, how could we construct a k-dimensional basis to project our data, while preserving as much variance as possible?

$d = 1000$

$d' = k$

k eigvecs corresp. to k largest eig vals.

$$\arg\max_{u:\|u\|_2=1} u^T \Sigma u$$

CS 189 Introduction to Machine Learning
Spring 2021 Jonathan Shewchuk

DIS12

Principal Component Analysis (PCA) is often used as a tool in data visualization and reduction of computation load and noise. PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after removing the mean from the data matrix for each feature/column. In this question we will derive PCA. There are four equivalent perspectives to understand PCA. PCA aims to find

1. the Gaussian distribution that best fits with maximum likelihood estimation;
2. the directions of projected maximum variance;
3. the projections of minimum reconstruction error;
4. the best low rank approximation.

E-Y-M

In this discussion, we will go through derivations for each of these interpretations, and show how these are all equivalent. First, however, we introduce Rayleigh quotients and present an optimization result that will be extremely useful.

1 Rayleigh Quotients

(a) The Rayleigh quotient is defined as

$$R(M, x) = \frac{x^\top M x}{x^\top x}$$

for a given symmetric matrix $M \in \mathbb{R}^{m \times m}$. What is the interval of possible values of the Rayleigh quotient for a given matrix? Specifically what is

$$\min_x R(M, x) \quad \text{and} \quad \max_x R(M, x)?$$

What values of x attain the bounds?

(b) How does the Rayleigh quotient relate to the following optimization problems? What does this tell us about the optimum values and the vectors which achieve them? Try to relate these quantities to the singular values and singular vectors of X.

$$\min_{w:\|w\|_2=1} \|Xw\|_2^2 \quad \text{and} \quad \max_{w:\|w\|_2=1} \|Xw\|_2^2.$$

2 The Gaussian MLE Perspective

(a) Assume our data matrix $X \in \mathbb{R}^{n \times d}$ is mean centered. What is the mean and variance of the maximum likelihood estimate for a Gaussian distribution fitting our dataset?

(b) Given this Gaussian, how could we construct a k-dimensional basis to project our data, while preserving as much variance as possible?

3 The Maximum Projected Variance Perspective

(a) We would like to find the vector w such that projecting your data onto w will retain the maximum amount of information (i.e. variance). The projections of our centered data onto w are

$$x_1^\top w, x_2^\top w, \ldots, x_n^\top w,$$

where x_i is the ith row of the matrix X. Compute the mean of and variance of these projections, and show the latter quantity is:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i^\top w\right)^2 = \frac{1}{n} w^\top X^\top X w$$

(b) We want to find a unit vector w which maximizes this quantity. Formulate this as an optimization problem and find the optimal vector w, along with the corresponding objective value. *Hint: Did we see a similar optimization problem before?*

(c) Let us call the solution of the above part w_1. Next, we will use a *greedy procedure* to find the ith component of PCA by doing the following optimization

$$\begin{aligned} \text{maximize} \quad & w_i^\top X^\top X w_i \\ \text{subject to} \quad & w_i^\top w_i = 1 \\ & w_i^\top w_j = 0 \quad \forall j < i, \end{aligned} \tag{1}$$

where w_j, $j < i$ are defined recursively using the same maximization procedure above. Show that the maximizer for this problem is equal to the eigenvector v_i that corresponds to the ith eigenvalue λ_i of matrix $X^\top X$. Also show that optimal value of this problem is equal to λ_i.

(d) Show that the previous *greedy procedure* finds the global maximum, namely for any $k < d$, $w_1, w_2, \ldots, w_k$ is the solution of the following maximization problem

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i=1}^{k} w_i^\top X^\top X w_i \\ \text{subject to} \quad & w_i^\top w_i = 1 \\ & w_i^\top w_j = 0 \quad \forall i \neq j. \end{aligned} \tag{2}$$

4 The Minimizing Reconstruction Error Perspective

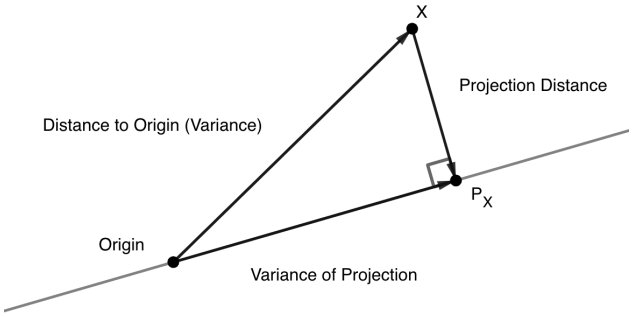
Our final perspective on PCA is minimizing the perpendicular distance between the principle component subspace and the data points. Let's say we want to find the best 1D space that minimizes the reconstruction error.

(a) Show the (vector) projection of the feature vector x onto the subspace spanned by a unit vector w is

$$P_w(x) = w\left(x^\top w\right). \tag{3}$$

(b) Now, we want to choose w to minimize the reconstruction error. Show that taking w as the minimizer for the corresponding problem below gives us the same result as before.

$$\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i - P_w(x_i)\|_2^2 \tag{4}$$



The above image serves as a useful visualization. Consider mean centered data. A data point has some fixed distance from the origin. We may consider finding a lower dimensional representation as either maximizing the variance of the projectiong or minimizing the projection distance. The squared quantities must sum to a constant (the distance to the origin or original variance) thus minimizing one is equivalent to maximizing the other.

5 The Eckart–Young–Mirsky Theorem Perspective*

In this problem, we fix an $n \times n$ positive semi-definite matrix A. We will derive, from first principles, the best rank-1 approximation to A. Recall the following metric of approximation: for any integer $1 \leq r \leq n$, we define the best rank-r approximation as any minimizer A_r of

$$\arg\min_{M \in R^{n \times n}} \|A - M\|_F : \mathsf{rank}(M) \leq r \tag{5}$$

The notation $\|A\|_F\|$ represents the Frobenius norm. This is equal to the square root of the sum of the squared entries of the matrix, $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$. Note that such a minimizer is not necessarily unique (why?). Since A is positive semi-definite, A_r must be as well: you make take this fact for granted. In this problem, we focus on the special case of $r = 1$.

(a) Show that $A_1 = mm^T$, where m is any minimizer of the following *unconstrained* optimization problem

$$\min_{m \in R^n} \left\|A - mm^T\right\|_F. \tag{6}$$

(b) Define the function $f : R^n \longrightarrow R$ as $f(m) = \left\|A - mm^T\right\|_F^2$. Compute $\nabla f(m)$.

(c) Argue from part (c) that $A_1 = \lambda_1 v_1 v_1^T$, where λ_1 is the maximum eigenvalue of A and v_1 is a corresponding (unit normalized) eigenvector.

This idea can be generalized to the Eckart–Young–Mirsky Theorem. For a general matrix $A \in \mathbb{R}^{m\times n}$,

$$\underset{M:\text{rank}(M)\leq k}{\arg\min} \|A - M\|_F = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

where σ_i,u_i,v_i correspond to the singular values, left singular and right singular vectors respectively.

The general SVD formulation for a matrix $X \in \mathbb{R}^{m\times n}$ is $X = U\Sigma V^\top$. In this "full" SVD formulation, we have $U \in \mathbb{R}^{m\times m}$, $\Sigma \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{n\times n}$. Notice that if X is not full rank, then Σ will have zeros along the diagonal, meaning when we expand $U\Sigma V^\top$ some of the columns of U and V are irrelevant.

If $m \geq n$, then this matrix has rank at most n, thus we may consider a formulation where we only consider the first n rows of U and Σ. We may denote these as $U_n \in \mathbb{R}^{m\times n}$, $\Sigma_n \in \mathbb{R}^{n\times n}$. This formulation $U_n\Sigma_n V^\top$ is known as the **thin SVD**.

From the SVD, we can easily recover the best k rank approximation of X, by considering only the first k singular values of Σ, which we may denote as $\Sigma_k \in \mathbb{R}^{m\times k}$. We may also consider the same corresponding r columns of U and V, resulting in $U_k \in \mathbb{R}^{m\times k}$, $V_k \in \mathbb{R}^{n\times k}$. $X_k = U_k\Sigma_k V_k^\top$. This is known as the **truncated SVD**, since we truncate to only include the k dimensions we care about. This is exactly the low rank approximation obtained from Eckart-Young-Mirsky.

If we do not truncate to an arbitrary dimension k and instead reduce to r, the rank of the matrix of X, we do not lose any information. We will still obtain X. $X = U_r\Sigma_r V_r$. This is known as the **compact SVD**.