Homework #2

Eric Tao
Math 123: Homework #2

January 31, 2023

Question 1. Let $\Sigma \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be the function $F(u) = u^T \Sigma u$ where u is understood as a column vector. Show that:

$$\frac{\partial F}{\partial u} = 2\Sigma u$$

Solution. First, we rewrite $u^T \Sigma u$ in terms of summation notation:

$$u^T \Sigma u = \sum_{j=1}^{d} \sum_{i=1}^{d} u_j (\Sigma)_{ji} u_i$$

where $(\Sigma)_{ji}$ denotes the element in the j-th row and the i-th column of Σ and u_i refers to the i-th element of the vector u. Now, because we recognize that $\frac{\partial F}{\partial u}$ can be interpreted as $(\frac{\partial}{\partial u_1}, ..., \frac{\partial}{\partial u_d}) \cdot F$, we fix some u_k and we look at $\frac{\partial}{\partial u_k}$ of our sum. Since we can look at only terms with u_k, we look at something like:

$$\sum_{j=1, j\neq k}^{d} u_j (\Sigma)_{jk} u_k + \sum_{i=1, i\neq k}^{d} u_k (\Sigma)_{ki} u_i + u_k^2 (\Sigma)_{kk}$$

Here, we notice that because we're working with real numbers, multiplication is commutative, and further, since Σ is symmetric, we have that $(\Sigma)_{jk} = (\Sigma)_{kj}$. Then, we notice that the first two summations are equal, since we have that

$$\sum_{j=1, j\neq k}^{d} u_j (\Sigma)_{jk} u_k = \sum_{j=1, j\neq k}^{d} u_k (\Sigma)_{kj} u_j$$

and we just relabel from counting by i or by j.

Then, we have:

$$2 \sum_{j=1, j\neq k}^{d} u_j (\Sigma)_{jk} u_k + u_k^2 (\Sigma)_{kk}$$

Taking the derivative, we compute this as:

$$\frac{\partial}{\partial u_k} 2 \sum_{j=1, j\neq k}^{d} u_j (\Sigma)_{jk} u_k + u_k^2 (\Sigma)_{kk} = 2 \sum_{j=1, j\neq k}^{d} u_j (\Sigma)_{jk} + 2u_k (\Sigma)_{kk} = 2 \sum_{j=1}^{d} u_j (\Sigma)_{jk}$$

Since this is true for each u_k, we notice that we can write:

$$(\frac{\partial F}{\partial u})_k = 2\sum_{j=1}^{d} u_j(\Sigma)_{jk}$$

However, we notice that, commuting and using the symmetric property again, that this is exactly $2\Sigma u$ as desired, as for the index k, $(\Sigma u)_k = \sum_{j=1}^{d}(\Sigma)_{kj}u_j$ □

Question 2. Recall that the variance of a set of numbers $x_1, ..., x_n \in \mathbb{R}$ is defined as $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$, where we define the mean as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. For each of the following statements, prove or give a counterexample.

(a) The variance is translation invariant, that is, the variance of $x_1, ..., x_n$ is the same as, for $T \in \mathbb{R}$, $x_1 + T, ..., x_n + T$.

(b) The variance is 0 if and only if $x_i = C$ for some constant $C \in \mathbb{R}$.

(c) The variance is additive, that is, if $x_1, ..., x_n$ have variance σ_x^2 and $y_1, ..., y_m$ have variance σ_y^2, then $x_1, ..., x_n, y_1, ..., y_m$ has variance $\sigma_x^2 + \sigma_y^2$.

Solution. (a)

Let $x_1, ..., x_n \in \mathbb{R}$ be a set of numbers, and define $y_i = x_i + T$, for some fixed $T \in \mathbb{R}$. First, we compute:

$$\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + T) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i + nT\right) = \mu_x + T$$

Then:

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i + T - \mu_x - T)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 = \sigma_x^2$$

(b)

Suppose the variance is 0. Then, we have that $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$, for some fixed $\mu \in \mathbb{R}$. We notice, that since we work in $\mathbb{R}$, $(x_i - \mu)^2 \geq 0$ for all i. Then, for the sum of a finite set of non-negative numbers to be 0, we must have that they are all 0, that is, this implies that $(x_i - \mu)^2 = 0 \implies x_i - \mu = 0 \implies x_i = \mu$. But this is true for all i, and since the mean is a fixed, real number for any set of real numbers, we conclude that $x_i = C$ for some real constant.

Now suppose $x_i = C$ for all i. Then, we have that

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} C = \frac{1}{n}nC = C$$

Then, we have that:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(C - C)^2 = 0$$

Thus, we have that $\sigma^2 = 0 \iff x_i = C$ for some constant $C \in \mathbb{R}$.

(c)

By part (b), this may not be true. Take the single data point $\{x_1 = 0\}$, and the single data point $\{y_1 = 1\}$. By part (b), we have that the variance must be 0 in both cases. However, the concatenated set $\{x_1, y_1\} = \{0, 1\}$ may not have variance 0, because it is not constant.

□

Question 3. A matrix $A \in \mathbb{R}^{d\times d}$ is said to be positive semi-definite if $y^T Ay \geq 0$ for all $y \in \mathbb{R}^d$. The matrix A is said to be positive definite if it is both positive semi-definite and $y^T Ay = 0$ if and only if $y = 0$.

(a) Let $x_1, ..., x_n \in \mathbb{R}^{1\times d}$ be data points. Let $\Sigma = \frac{1}{n}\sum_{i=1}^n x_i^T x_i$ be the covariance matrix. Prove that Σ is positive semi-definite.

(b) Does Σ necessarily have to be positive definite?

Solution. (a) We recognize this by running a transpose trick in reverse. We consider $n\Sigma$, so I don't have to carry around a $\frac{1}{n}$ while typing this, and this merely scales the result.

Let y be any row vector, that is, of form $\mathbb{R}^{1\times d}$. Consider:

$$yn\Sigma y^T = y\left(\sum_{i=1}^n x_i^T x_i\right)y^T = \left(\sum_{i=1}^n yx_i^T x_i y^T\right)$$

Here, we notice that yx_i^T is a multiplication of a $1 \times d$ vector y by a $d \times 1$ vector x_i^T, so yx_i^T is a real number. Further, by the properties of the transpose, we have that $yx_i^T = (yx_i^T)^T = x_i y^T$, because of course, the transpose of a 1×1 matrix, or real number, is itself. Thus, we can rewrite the summand as:

$$\left(\sum_{i=1}^n yx_i^T x_i y^T\right) = \left(\sum_{i=1}^n [yx_i^T]^2)\right)$$

Here, we notice that because we work in the reals, the square of a real number is non-negative, i.e. $[yx_i^T]^2 \geq 0$. And therefore, since this is a sum of non-negative numbers, their sum is also non-negative, i.e. $\sum_{i=1}^n [yx_i^T]^2 \geq 0$.

(b)

Σ does not need to be positive definite. Clear example: we recall that if a matrix is not invertible, then it must have 0 as one of its eigenvalues. From last homework, we analyzed the covariance matrix of points sampled from a line $y = \alpha x$, for a real parameter α, and showed that the covariance matrix here has rank at most 1. Since an invertible 2×2 matrix would have rank 2, this matrix may not be invertible, and thus has 0 as an eigenvalue. Then, for its corresponding eigenvector $\zeta \neq 0$, we would have that:

$$\zeta^T \Sigma \zeta = \zeta^T 0 \zeta = 0$$

Thus, Σ need not be positive definite.

□

Question 4. When dimension reducing data in $\mathbb{R}^D$ using PCA, the choice of embedding dimension is crucial. Many heuristics exist to estimate a good dimension. One is to choose an embedding dimension d^* as the smallest dimension that some proportion of the variance is preserved by projecting onto the first d^* principal components:

$$d^* = \min\left\{d \middle| \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i} > 0.95\right\}$$

(a) Intuitively, when will d^* be small?

(b) Intuitively, when will d^* be large?

(c) Are there any situations in which d^* is roughly $0.95 * D$?

Solution. (a)

The intuition here should be that d^* is small if it requires few principal components to capture most of the variance in a set. In some sense, we think of this as saying that a subset of the parameters are more influential than others, and we may reduce to looking at just those parameters if we only wish to capture

the majority of the variance. More quantitatively, we can say that d^* is small when there are λ_i such that λ_i is large compared to the others.

(b)

Conversely, d^* will be large if the variance is relatively evenly spread between the parameters. That is, roughly, each direction has a comparable amount of variance associated with it. More quantitatively, we can say that d^* is large when λ_i are approximately equal.

(c)

An example here would be a D-ball, that is, the D-dimensional ball contained within $\mathbb{R}^d$, and taking data points sampled from the interior. In a sense, with enough data points, we would have that each d hyperplane roughly captures the same proportion of variance, and we expect that under this heuristic, that we would have $d^* \approx 0.95 * D$.

□

Question 5. Download a prepared dataset:

(a) Compute the principal component decomposition of the data.

(b) Compute the value of d^* as defined in problem 4.

(c) Compute and display the first and last 3 principal components. Are there any obvious contrasts?

Solution. Here, after downloading the data, following the example code for a different remote sensing set provided, we will attack this via PCA, after flattening each vector so that one row represents a single image sample. To be clear, the procedure was to reshape the data matrix from form $83 \times 86 \times 204$ to 7138×204, so that each row represents a single sample, read one column at a time and concatenated.

We then compute the mean of each pixel across the 204 samples, and subtract this from the matrix row-by-row to retrieve a centered set of data on 0. Then, we compute the empirical covariance matrix in the usual way. Lastly, we compute the eigenvalues and vectors of this matrix.

Using the eig function in matlab, we return the diagonal matrix of eigenvalues, sorted. Then, we need only compare against the trace. From this procedure we calculated d^* as 2, as we have that the trace of the eigenvalue matrix is $4.0783 * 10^7$ and the ratio of the largest two entries to the sum is $\frac{37.522+2.685)*10^6}{40.783*10^7} \approx 0.9974$.

Visually, looking at and comparing the first and last 3 principal components, we notice that the 3 with largest eigenvalues tend to have many same-signed entries in a row, whereas the ones with the smallest eigenvectors tend to have them more evenly distributed. I would hazard that this has to do with the fact that this is land imaging data, and trying to look at the color scheme of land tends to show patterns in the direction that cultivation happens or something along those lines, assuming I'm interpreting the data points correctly.

□