# A collection of proofs

Sharon Hui

1/18/2024

## Exercises

This section has almost nothing to do with Stat 135. It is a collection of proofs and exercises of useful facts/theorems. Personally, for me to truly believe and understand a statistical theorem, I need to see and understand the proof. Additionally, since I don't have the luxury of being a student anymore, having this section here keeps me accountable for always improving my proof-writing skills. If there are errors, please let me know.

As a personal philosophy, I believe that we are all models. We learn skills through examples and practice, just like a model learns to predict and classify through training data. Another thing I would like to say is that a model may be penalized for getting the answer wrong, but it doesn't stop trying to learn. It only really stops learning when it runs out of examples to learn.

What makes humans so incredible is that we are the vessels for many learning tasks. Given that we have this unique advantage over modeling, we have to keep learning and practice to get better at our crafts. With that being said, I'll be sporadically updating this page with problems/proofs.

Happy learning!

### Problem 1: Pairwise Euclidean distances and variance

Consider using Euclidean distances to measure how far points are from each other.

Show that the sum of all pairwise distances between a sample of individuals is directly related to variance.

$$\sum_{i=1}^{n}\sum_{\ell=1}^{n} d^2(i,\ell) = (2n^2)\sum_{i=1}^{n} d^2(i,g)$$

Here, $g$ is the centroid.

Solution:

What we are really showing is:

$$\sum_{i=1}^{n}\sum_{\ell=1}^{n}(x_i - x_\ell)^2 = 2n\sum_{i=1}^{n}(x_i - \bar{x})^2$$

From the left side:

$$\sum_{i=1}^{n}\sum_{\ell=1}^{n}\left[x_i^2 - 2x_i x_\ell + x_\ell^2\right] = \sum_{i=1}^{n}\left(nx_i^2 - 2x_i\sum_{\ell=1}^{n}x_l + \sum_{\ell=1}^{n}x_\ell^2\right) = \sum_{i=1}^{n}\left(nx_i^2 - 2x_i n\bar{x} + \sum_{\ell=1}^{n}x_\ell^2\right)$$

$$n\sum_{i=1}^{n} x_i^2 - 2n\bar{x}n\bar{x} + n\sum_{\ell=1}^{n} x_\ell^2 = 2n\sum_{i=1}^{n} x_i^2 - 2n^2\bar{x}^2 = 2n\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] = 2n\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Notice $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

This ultimately shows us that minimizing the variance is the same as minimizing the pairwise distance.

## Centering matrix

When you multiply the centering matrix ($\mathbf{C_n} = \mathbf{I_n} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J_n}$ where $I_n$, is the identity matrix of size $n$ and $J_n$ is an $n$-by-$n$ matrix of all 1's.) with a vector, it is effectively the same as subtracting the mean of the components of the vector from every component of that vector. Prove the following: The centering matrix is a symmetric and idempotent matrix; this is called being a projection matrix. The centering matrix is also positive semi-definite.

First, let's prove $\mathbf{C_n}$ is symmetric

I want to show that $\mathbf{C_n}$ is symmetric, which means that ${\mathbf{C_n}}^T = \mathbf{C_n}$

Using the properties of a transpose, notice that $(\mathbf{A} - \mathbf{B})^T = \mathbf{A}^T - \mathbf{B}^T$

$$\begin{aligned}{\mathbf{C_n}}^T = \left(\mathbf{I_n} - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathbf{T}\right)^T = (\mathbf{I_n})^T - \left(\frac{1}{n}\mathbf{1}\mathbf{1}^\mathbf{T}\right)^T = \mathbf{I_n} - \frac{1}{n}(\mathbf{1}^\mathbf{T})^\mathbf{T}(\mathbf{1})^\mathbf{T} \\ = \left(\mathbf{I_n} - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathbf{T}\right) = \mathbf{C_n}\end{aligned}$$

Since ${\mathbf{C_n}}^T = \mathbf{C_n}$, this means that $\mathbf{C_n}$ is symmetric.

Next, let's show $\mathbf{C_n}$ is idempotent

I want to show that $\mathbf{C_n}$ is idempotent, which means that ${\mathbf{C_n}}^2 = \mathbf{C_n}$.

Firstly, notice that $\mathbf{1}$ is a n by 1 vector:

$$\mathbf{1} = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$$

$$\mathbf{1}^T = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}$$

This means that

$$\mathbf{1}^T\mathbf{1} = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} = \sum_{i=1}^{n} 1 = n$$

$${\mathbf{C_n}}^2 = \mathbf{C_n}\mathbf{C_n} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)$$

$$= \mathbf{I_n}\mathbf{I_n} - \mathbf{I_n}\frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{I_n} + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T$$

$$= \mathbf{I_n}\mathbf{I_n} - \mathbf{I_n}\frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{I_n} + \frac{1}{n^2}\mathbf{1}n\mathbf{1}^T = \mathbf{I_n}\mathbf{I_n} - \frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

$$= \mathbf{I_n} - \frac{1}{n}\mathbf{1}\mathbf{1}^T = \mathbf{C_n}$$

This shows that $\mathbf{C_n}^2 = \mathbf{C_n}$.

Lastly, let's show $\mathbf{C_n}$ is PSD.

For $\mathbf{C_n}$ to be PSD, it means for all non-zero $\mathbf{x} \in \mathbb{R}$, $\mathbf{x}^T\mathbf{C_n}\mathbf{x} \geq 0$

$$\mathbf{x}^T \left( \mathbf{I_n} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \right) \mathbf{x}$$

$$\mathbf{x}^T\mathbf{x} - \frac{1}{n}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \geq 0$$

Notice $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Since $\mathbf{C_n}$ are PSD, the eigenvalues of $\mathbf{C_n}$ are either zero or one.

Alternatively, we can show it by using the idempotency of $\mathbf{C_n}$. Let $\mathbf{C_n}\mathbf{v_i} = \lambda_i\mathbf{v_i}$ where $\mathbf{v_i}$ is a eigenvector and $\lambda_i$ is its corresponding eigenvalue.

$$\lambda_i\mathbf{v_i} = \mathbf{C_n}\mathbf{v_i} = \mathbf{C_n}^2\mathbf{v_i} = \mathbf{C_n}\mathbf{C_n}\mathbf{v_i} = \mathbf{C_n}\lambda_i\mathbf{v_i} = \lambda_i\mathbf{C_n}\mathbf{v_i} = \lambda_i\lambda_i\mathbf{v_i} = \lambda_i^2\mathbf{v_i}$$

This means that $\lambda_i\mathbf{v_i} = \lambda_i^2\mathbf{v_i}$ must hold true, which is only when $\lambda_i$ is 0 or 1.

## Symmetric Matrices

Eigenvectors of real symmetric matrices are orthogonal.

If A is symmetric, then $\mathbf{A} = \mathbf{A^T}$. By definition, $\mathbf{A}\mathbf{v_i} = \lambda_\mathbf{i}\mathbf{v_i}$ and $\mathbf{A}\mathbf{v_j} = \lambda_\mathbf{j}\mathbf{v_j}$.

$$\lambda_i\mathbf{v_i^T}\mathbf{v_j} = (\lambda_\mathbf{i}\mathbf{v_i})^\mathbf{T}\mathbf{v_j} = (\mathbf{A}\mathbf{v_i})^\mathbf{T}\mathbf{v_j} = \mathbf{v_i^T}\mathbf{A^T}\mathbf{v_j} = \mathbf{v_i^T}(\mathbf{A^T}\mathbf{v_j}) = \mathbf{v_i^T}(\mathbf{A}\mathbf{v_j}) = \mathbf{v_i^T}(\lambda_j\mathbf{v_j}) = \lambda_j\mathbf{v_i^T}(\mathbf{v_j})$$

$$\lambda_i\mathbf{v_i^T}\mathbf{v_j} = \lambda_j\mathbf{v_i^T}(\mathbf{v_j})$$

This means that

$$\lambda_i\mathbf{v_i^T}\mathbf{v_j} - \lambda_j\mathbf{v_i^T}\mathbf{v_j} = 0$$

Since $\lambda_i$ and $\lambda_j$ are two distinct eigenvalues meaning $\lambda_i \neq \lambda_j$, this must imply that $\mathbf{v_i^T}\mathbf{v_j} = 0$

## Linear Regression

Assume that the following linear relationship is true: $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.

Let $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ and $X = \begin{bmatrix} x_{11} & x_{12} & \cdots x_{1p} \\ \vdots & & \\ x_{n1} & x_{n2} & \cdots x_{np} \end{bmatrix}$, so $X$ is a $n \times p$ matrix.

We want to find the best linear fit to the data so $\hat{Y} = X\hat{\beta}$. Choosing that we want to minimize the squared loss, then we want to minimize

$$\|Y - \hat{Y}\| = \|Y - X\hat{\beta}\|$$

. Then

$$(Y - X\hat{\beta})^T(Y - X\hat{\beta}) = Y^TY - \hat{\beta}^TX^TY - Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta} = Y^TY - 2\hat{\beta}^TX^TY + \hat{\beta}^TX^TX\hat{\beta}$$

Taking derivative with respect to $\hat{\beta}$ and set to 0:

$$X^TY = X^TX\hat{\beta} \longrightarrow \hat{\beta} = (X^TX)^{-1}X^TY$$

Know that this requires $X^TX$ to be invertible, which happens when $X$ is full rank.

$$E[Y] = E[X\beta + \epsilon] = X\beta$$
$$Cov(Y) = Cov(X\beta + \epsilon) = Cov(\epsilon) = \sigma^2$$

$$E[\hat{\beta}] = E[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^TX\beta = \beta$$

$$Var(\hat{\beta}) = Cov(\hat{\beta}, \hat{\beta}) = Cov((X^TX)^{-1}X^TY, (X^TX)^{-1}X^TY) = (X^TX)^{-1}X^TCov(Y,Y)((X^TX)^{-1}X^T)^T$$
$$= (X^TX)^{-1}X^T\sigma^2X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

## Weighted Least Squares

Weighted least squares is used when linear measurements are corrupted by errors with unequal variances. We want to minimize $\sum_{i=1}^n w_i(Y_i - \sum_{j=1}^p x_{ij}\beta_j)^2$ with the constraint $w_1, ..., w_n \geq 0$

In matrix form, the equivalent problem is $\arg\min_\beta \|W^{1/2}(Y - X\beta)\|^2$, with $W$ being a symmetric, positive-definite weighting matrix.

The normal equations are then:

$$(W^{1/2}(Y - X\beta))^T(W^{1/2}(Y - X\beta)) = (Y^T - \beta^TX^T)(W^{1/2})^TW^{1/2}(Y - X\beta)$$
$$= Y^TWY - \beta^TX^TWY - Y^TWX\beta + \beta^TX^TWX\beta$$

Take derivative wrt $\beta$ and set to 0:

$$2X^TWY = 2X^TWX\beta \longrightarrow \hat{\beta} = (X^TWX)^{-1}X^TWY$$

Weighted least squares is essentially the same as transforming $X$ and $Y$ and $\epsilon$.

The main advantage of using WLS is the retainment of interpretations of the coefficients and interpreting F-tests and R-squared values (usually WLS keeps the intercept while transformed models do not). It is also advantageous to downweight outlier or influential points by setting its weight to 0.

## Correlation Matrix

The correlation matrix is a symmetric PSD matrix with unit diagonal.

Covariance between two random variables, $X$ and $Y$: $Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}[XY - Y\mathbb{E}(X) - X\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$, provided that the expectation exists.

## Least Squares with Regularization

### Ridge Regression

With regularization, extra terms (variables) are added to the cost function (to avoid overfitting).

$$\min \sum_{i=1}^{n} (y_i - x_{i^T}\beta)^2 + \rho \sum_{j=1}^{p} \beta_j^2$$

with the constraint that $\rho > 0$.

### Lasso Regression

With regularization, extra terms (variables) are added to the cost function (to avoid overfitting).

$$\min \sum_{i=1}^{n} (y_i - x_{i^T}\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

with the constraint that $\lambda > 0$.

Alternatively, we can write

$$\min \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$