

Homework 2

Stat 215A, Fall 2020

Due: push a `homework2.pdf` file to your `stat-215-a` GitHub repo by Thursday, October 08 at 11:59pm

0 Linear Algebra Review

Recall that the SVD of $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a matrix decomposition such that $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^\top$, where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n] \in \mathbb{R}^{n \times n}$, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_p] \in \mathbb{R}^{p \times p}$, and $\mathbf{D} = \text{diag}(d_1, \dots, d_{\min\{n,p\}}) \in \mathbb{R}^{n \times p}$. In addition, \mathbf{U} and \mathbf{V} are orthogonal matrices so that $\mathbf{U}^\top \mathbf{U} = \mathbf{U} \mathbf{U}^\top = \mathbf{I}$ and $\mathbf{V}^\top \mathbf{V} = \mathbf{V} \mathbf{V}^\top = \mathbf{I}$ (i.e., $\mathbf{u}_j^\top \mathbf{u}_i = \mathbf{v}_j^\top \mathbf{v}_i = 0$ for all $i \neq j$ and $\mathbf{u}_j^\top \mathbf{u}_j = \mathbf{v}_j^\top \mathbf{v}_j = 1$ for all j). Moreover, $d_1 \geq \dots \geq d_{\min\{n,p\}} \geq 0$.

Now, while the SVD can be used for any rectangular matrix, square matrices have an additional special property and can be decomposed via an eigendecomposition. Given a square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, we say that $\mathbf{v} \in \mathbb{R}^p$ is an *eigenvector* of \mathbf{A} if $\mathbf{v} \neq \mathbf{0}$ and $\mathbf{A} \mathbf{v} = \lambda \mathbf{v}$ for some $\lambda \in \mathbb{R}$. We also call λ the *eigenvalue* of \mathbf{A} corresponding to the eigenvector \mathbf{v} . For a more intuitive (geometric) interpretation of eigenvalues and eigenvectors, see this [reference](#).

There is a close connection between the SVD and eigendecomposition. Namely, for any matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{v} \in \mathbb{R}^p$ is a right singular vector of \mathbf{X} with singular value d if and only if $\mathbf{v} \in \mathbb{R}^p$ is an eigenvector of $\mathbf{X}^\top \mathbf{X}$ corresponding to the eigenvalue d^2 . You may use this fact without proof.

1 Principal Components Analysis and SVD

Let \mathbf{X} be an $n \times p$ data matrix, where n is the number of observations and p is the number of features. For simplicity, we will assume that \mathbf{X} has been mean-centered (i.e., each column of \mathbf{X} has mean 0) and that $n \leq p$. In the lab section, we used projections in order to introduce the population version of PCA as solving for each $j = 1, \dots, p$

$$\mathbf{v}_j^* = \underset{\mathbf{v} \in \mathbb{R}^p}{\operatorname{argmax}} \mathbf{v}^\top \operatorname{Var}(\mathbf{X}) \mathbf{v} \quad \text{subject to} \quad \|\mathbf{v}\|_2^2 = 1, \mathbf{v}^\top \mathbf{v}_i^* = 0 \quad \forall i < j. \quad (1.1)$$

However, since $\operatorname{Var}(\mathbf{X})$ is almost always unknown in practice, we typically estimate $\operatorname{Var}(\mathbf{X})$ with the sample covariance $\frac{1}{n} \mathbf{X}^\top \mathbf{X}$. Thus, in practice, the principal component (PC) directions, $\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_p$, are the solution to the following system of optimization problems:

$$\hat{\mathbf{v}}_j = \underset{\mathbf{v} \in \mathbb{R}^p}{\operatorname{argmax}} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \quad \text{subject to} \quad \|\mathbf{v}\|_2^2 = 1, \mathbf{v}^\top \hat{\mathbf{v}}_i = 0 \quad \forall i < j. \quad (1.2)$$

In this problem, we will take small steps through the proof to show that the PC directions are precisely the right singular vectors of \mathbf{X} .

1. To begin, prove that the first PC direction $\hat{\mathbf{v}}_1$ is equal to the first right singular vector \mathbf{v}_1 . To show this, use [Lagrange multipliers](#) to solve the PC1 optimization problem:

$$\hat{\mathbf{v}}_1 = \underset{\mathbf{v} \in \mathbb{R}^p}{\operatorname{argmax}} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \quad \text{subject to} \quad \|\mathbf{v}\|_2^2 = 1. \quad (1.3)$$

If you are not familiar with matrix calculus, [Wikipedia](#) is a convenient resource for common derivative identities, which you may find useful here.

2. Next, let $j \in \{2, \dots, p\}$ be given. Use the SVD and matrix multiplication to show that for all $\mathbf{v} \in \mathbb{R}^p$ satisfying $\mathbf{v}^\top \mathbf{v}_i = 0$ for each $i < j$, we have

$$\mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} = \sum_{k=j}^p \mathbf{v}^\top (d_k^2 \mathbf{v}_k \mathbf{v}_k^\top) \mathbf{v}, \quad (1.4)$$

where we define $d_k = 0$ for $k = n+1, \dots, p$.

3. Then, show that for each $j = 2, \dots, p$, the original (sample) PCA formulation in (1.2) is equivalent to

$$\hat{\mathbf{v}}_j = \operatorname{argmax}_{\mathbf{v} \in \mathbb{R}^p} \mathbf{v}^\top \left(\mathbf{X}_{(j)}^\top \mathbf{X}_{(j)} \right) \mathbf{v} \quad \text{subject to} \quad \|\mathbf{v}\|_2^2 = 1, \quad (1.5)$$

where $\mathbf{X}_{(j)} = \tilde{\mathbf{U}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}^\top$, $\tilde{\mathbf{U}} = [\mathbf{u}_j, \dots, \mathbf{u}_n, \mathbf{u}_1, \dots, \mathbf{u}_{j-1}] \in \mathbb{R}^{n \times n}$, $\tilde{\mathbf{D}} = \operatorname{diag}(d_j, \dots, d_n, 0, \dots, 0) \in \mathbb{R}^{n \times p}$, and $\tilde{\mathbf{V}} = [\mathbf{v}_j, \dots, \mathbf{v}_p, \mathbf{v}_1, \dots, \mathbf{v}_{j-1}] \in \mathbb{R}^{p \times p}$.

4. Conclude that for each $j = 1, \dots, p$, the j^{th} PC direction, $\hat{\mathbf{v}}_j$, is equal to the j^{th} right singular vector \mathbf{v}_j . (Hint: Problem 1 may be useful).

2 Ordinary Least Squares

Suppose that we observe our usual data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response vector $\mathbf{y} \in \mathbb{R}^n$, where n is the number of samples/observations and p is the number of features. Suppose also that \mathbf{X} has rank $p < n$. Under this setting, the ordinary least squares (OLS) estimator is given by

$$\hat{\boldsymbol{\beta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

1. Provide an expression for $\hat{\boldsymbol{\beta}}_{OLS}$ in terms of \mathbf{X} and \mathbf{y} by solving the optimization problem above. Why do we require the assumption that $\operatorname{rank}(\mathbf{X}) = p < n$?
2. Show that the OLS predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$ can be written as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}^2 = \mathbf{H}$.
3. Prove that the residuals $\hat{\mathbf{r}} = \mathbf{y} - \hat{\mathbf{y}}$ are orthogonal to the OLS predictions $\hat{\mathbf{y}}$. Draw a picture to show what this means geometrically.