CMSE 820 HW 3

Hao Lin

September 27, 2018

1 Question 1

Prove that u_2 is the eigenvector of Σ_x associated with the second largest eigenvalue λ_2.
Proof:

$$u_2 = \arg\max_{u_2 \in \mathbb{R}^p} u_2^T \Sigma_x u_2 \text{ s.t. } u_2^T u_2 = 1 \text{ and } u_1^T u_2 = 0.$$

Define the following Langrangian:

$$\mathcal{L}(u_2, \alpha, \beta) = u_2^T \Sigma_x u_2 + \alpha u_1^T u_2 + \beta(1 - u_2^T u_2).$$

Set the partial derivatives of $\mathcal{L}$ to be zero:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial u_2} &= 2\Sigma_x u_2 + \alpha u_1 - 2\beta u_2 = 0 \\ \frac{\partial \mathcal{L}}{\partial \alpha} &= u_1^T u_2 = 0 \\ \frac{\partial \mathcal{L}}{\partial \beta} &= 1 - u_2^T u_2 = 0 \end{aligned}$$

The latter two conditions simply correspond to the two constraints. Multiplying by u_1^T to the left on both sides of the first condition gives

$$2u_1^T \Sigma_x u_2 + \alpha u_1^T u_1 - 2\beta u_1^T u_2 = 0 \implies \alpha = 0,$$

because $\Sigma_x u_1 = \lambda_1 u_1$, $u_1^T u_2 = 0$ and $u_1^T u_1 = 1$. λ_1 is the largest eigenvalue of Σ_x. It follows that,

$$\Sigma_x u_2 = \beta u_2.$$

That is, u_2 is an eigenvector of Σ_x given that $\beta \neq 0$.

Note that $u_2^T \Sigma_x u_2 \leq \lambda_1$, by the definition of λ_1. To maximize $u_2^T \Sigma_x u_2$, u_2 has to be the eigenvector associated with the second largest eigenvalue (assuming the eigenvalues are not repeated). [Alternatively, one can express u_2 in the orthonormal eigenbasis of Σ_x and maximize $u_2^T \Sigma_x u_2$ while ensuring the orthorganality of u_1 and u_2].

2 Question 2

Given $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{X} = I \in \mathbb{R}^{n \times n}$ and $\lambda > 0$,

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \frac{1}{2}\|\mathbf{y} - \beta\|^2 + \lambda\|\beta\|_1.$$

Define the following Lagrangian $\mathcal{L}$:

$$\mathcal{L}(\beta, \lambda) = \frac{1}{2}\|\mathbf{y} - \beta\|^2 + \lambda\|\beta\|_1.$$

Taking the partial derivative of $\mathcal{L}$ with respect to β gives

$$\frac{\partial \mathcal{L}}{\partial \beta} = -(\mathbf{y} - \beta) + \lambda \mathbf{s}(\beta),$$

where $\mathbf{s}$ denote the subderivative of $\|\cdot\|_1$.

Note that $\hat{\beta}^{\text{ols}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y} = \mathbf{y}$ since $\mathbf{X} = I$. $\frac{\partial \mathcal{L}}{\partial \beta} = 0$ becomes

$$\hat{\beta}^{\text{Lasso}} - \hat{\beta}^{\text{ols}} + \lambda \mathbf{s}(\hat{\beta}^{\text{Lasso}}) = 0. \quad (*)$$

If $|\hat{\beta}_i^{\text{ols}}| \geq \lambda$, we can set $\hat{\beta}_i^{\text{Lasso}} = \hat{\beta}_i^{\text{ols}} - \lambda \operatorname{sign}(\hat{\beta}_i^{\text{ols}})$ to satisfy $(*)$.

If $|\hat{\beta}_i^{\text{ols}}| < \lambda$ and $\hat{\beta}_i^{\text{Lasso}} \neq 0$, it is obvious that $\operatorname{sign}(\hat{\beta}_i^{\text{Lasso}} - \hat{\beta}_i^{\text{ols}} + \lambda \mathbf{s}(\hat{\beta}_i^{\text{Lasso}})) = \operatorname{sign}(\hat{\beta}_i^{\text{Lasso}}) \neq 0$. However, if $\hat{\beta}_i^{\text{Lasso}} = 0$ when $|\hat{\beta}_i^{\text{ols}}| < \lambda$, then one can choose $s(\hat{\beta}_i^{\text{Lasso}}) = \hat{\beta}_i^{\text{ols}}/\lambda \in [-1, 1]$ to satisfy $(*)$.

To sum up, $\forall i$

$$\hat{\beta}_i^{\text{Lasso}} = \begin{cases} \hat{\beta}_i^{\text{ols}} - \lambda \operatorname{sign}(\hat{\beta}_i^{\text{ols}}), & |\hat{\beta}_i^{\text{ols}}| \geq \lambda \\ 0, & |\hat{\beta}_i^{\text{ols}}| < \lambda \end{cases}.$$

3 Question 3

3.1 a

$A \in \mathbb{R}^{m \times n}$. The nuclear norm:

$$f(A) = \|A\|_* = \sum_{i=1}^{\text{min(m,n)}} \sigma_i(A),$$

where $\sigma_i(A)$ is the ith largest singular value of A.

Firstly, the domain $\mathbb{R}^{m \times n}$ is obviously convex. Secondly, since the nuclear norm $\|\cdot\|_*$ is a norm. It satisfies the triangle inequaltiy. $\forall t \in [0, 1]$, $\forall A, B \in \mathbb{R}^{m \times n}$, we have

$$\|tA + (1-t)B\|_* \leq t\|A\|_* + (1-t)\|B\|_*,$$

which shows its convexity.

3.2 b

$$\partial\|A\|_* = \{G \in \mathbb{R}^{m\times n} : \|B\|_* \geq \|A\|_* + \mathrm{Tr}[(B-A)^T G],\ \forall B \in \mathbb{R}^{m\times n}\}.$$

(1) $\|A\|_* = \mathrm{Tr}(G^T A)$. Proof: Since the inequality in the definition of $\partial\|A\|_*$ holds $\forall B \in \mathbb{R}m \times n$. Firstly, let $B = 0$, and we have

$$\|A\|_* - \mathrm{Tr}[A^T G] = \|A\|_* - \mathrm{Tr}[G^T A] \leq 0 \implies \|A\|_* \leq \mathrm{Tr}[G^T A].$$

On the other hand, setting $B = 2A$ yields

$$2\|A\|_* \geq \|A\|_* + \mathrm{Tr}[A^T G] = \|A\|_* + \mathrm{Tr}[G^T A] \implies \|A\|_* \geq \mathrm{Tr}[G^T A].$$

Hence, it follows that $\|A\|_* = \mathrm{Tr}(G^T A)$.

(2) $\|G\|_*^* = \max_{\|B\|_* \leq 1} \mathrm{Tr}(B^T G)$.

Using (1), the inequality can be simplified as

$$\mathrm{Tr}(B^T G) \leq \|B\|_*,\ \forall B \in \mathbb{R}^{m\times n}.$$

. It follows that $\|G\|_*^* \leq \|B\|_* \leq 1$, given that $\|B\|_* \leq 1$.

Now consider $\tilde{A} := \frac{1}{\|A\|_*} A$.

$$\|G\|_*^* \geq \mathrm{Tr}(\tilde{A}^T G) = \frac{1}{\|A\|_*} \mathrm{Tr}(G^T A) = 1.$$

Thus, $\|G\|_*^* = 1$.

3.3 c

$$\partial\|A\|_* = \{UV^T + W : U^T W = 0,\ WV = 0,\ \|W\|_2 \leq 1,\ W \in \mathbb{R}^{m\times n}\},$$

where $A = U\Sigma V^T$ is the SVD ($U \in \mathbb{R}^{m\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{n\times r}$).

The problem

$$\arg\min_A \frac{1}{2}\|X - A\|_F^2 + \lambda\|A\|_*$$

is a convex optimization problem. Note, in particular, that $\|\cdot\|_F$ is strictly convex, so this is actually a strictly convex problem. Define $\mathcal{L}(A, \lambda) = \frac{1}{2}\|X - A\|_F^2 + \lambda\|A\|_*$. $\frac{\partial \mathcal{L}}{\partial A} = 0$ gives the optimal condition

$$A - X + \lambda G = 0, \quad (**)$$

where $G \in \partial\|A\|_*$.

Let $X = U\Sigma V^T$ be the reduced/compact SVD for X ($U \in \mathbb{R}^{m\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{n\times r}$). Without loss of generality, let's assume only the first k singular values of Σ are greater

than λ. Define a positive semi-definite diagonal matrix $\mathcal{S}_\lambda(\Sigma) := \text{diag}\{\sigma_1 - \lambda, ..., \sigma_k - \lambda, 0, 0, ..., 0\} \in \mathbb{R}^{r\times r}$. Now consider a matrix $A \in \mathbb{R}^{m\times n}$ constructed as follows,

$$A = U\mathcal{S}_\lambda(\Sigma)V^T = U_{m\times k}\mathcal{S}_{\lambda;k\times k}(\Sigma)V^T_{k\times n},$$

where the former expression is the full SVD while the latter is the reduced SVD.

Substituting $A = U\mathcal{S}_\lambda(\Sigma)V^T$ and $X = U\Sigma V^T$ into $(**)$ yields

$$U\mathcal{S}_\lambda(\Sigma)V^T - U\Sigma V^T + \lambda U_{m\times k}V^T_{k\times n} + \lambda W = 0 \implies W = \frac{1}{\lambda}U[\Sigma - \mathcal{S}_\lambda(\Sigma) - \lambda\tilde{I}]V^T,$$

where $\tilde{I} \in \mathbb{R}_{r\times r}$ is such that

$$\tilde{I} = \begin{bmatrix} I_{k\times k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

In particular,

$$\frac{1}{\lambda}[\Sigma - \mathcal{S}_\lambda(\Sigma) - \lambda\tilde{I}] = \text{diag}\{0, ..., 0, -\frac{\sigma_{k+1}}{\lambda}, ..., -\frac{\sigma_r}{\lambda}\} = \tilde{D}_{k+1:r},$$

and $\frac{\sigma_r}{\lambda} \leq .. \leq \frac{\sigma_{k+1}}{\lambda} < 1$. All singular values of W are less than 1, and therefore $\|W\|_2 \leq 1$. In the following, we will check that $U^T_{k\times m}W = 0$ and $WV_{n\times k} = 0$.

$$\begin{aligned} U^T_{k\times m}W &= U^T_{k\times m}U\tilde{D}_{k+1:r}V^T = \tilde{I}\,\text{diag}\{0, ..., 0, -\frac{\sigma_{k+1}}{\lambda}, ..., -\frac{\sigma_r}{\lambda}\}\,V^T = 0, \\ WV_{n\times k} &= U\tilde{D}_{k+1:r}V^TV_{n\times k} = U\,\text{diag}\{0, ..., 0, -\frac{\sigma_{k+1}}{\lambda}, ..., -\frac{\sigma_r}{\lambda}\}\,\tilde{I} = 0. \end{aligned}$$

The constructed $A = U\mathcal{S}_\lambda(\Sigma)V^T$ and $G = U_{m\times k}V^T_{k\times n} + W$ satisfy the optimal condition $(**)$. Considering that the problem at hand is strictly convex, the constructed A is the unique optimal solution.

4 Question 4

4.1 1.

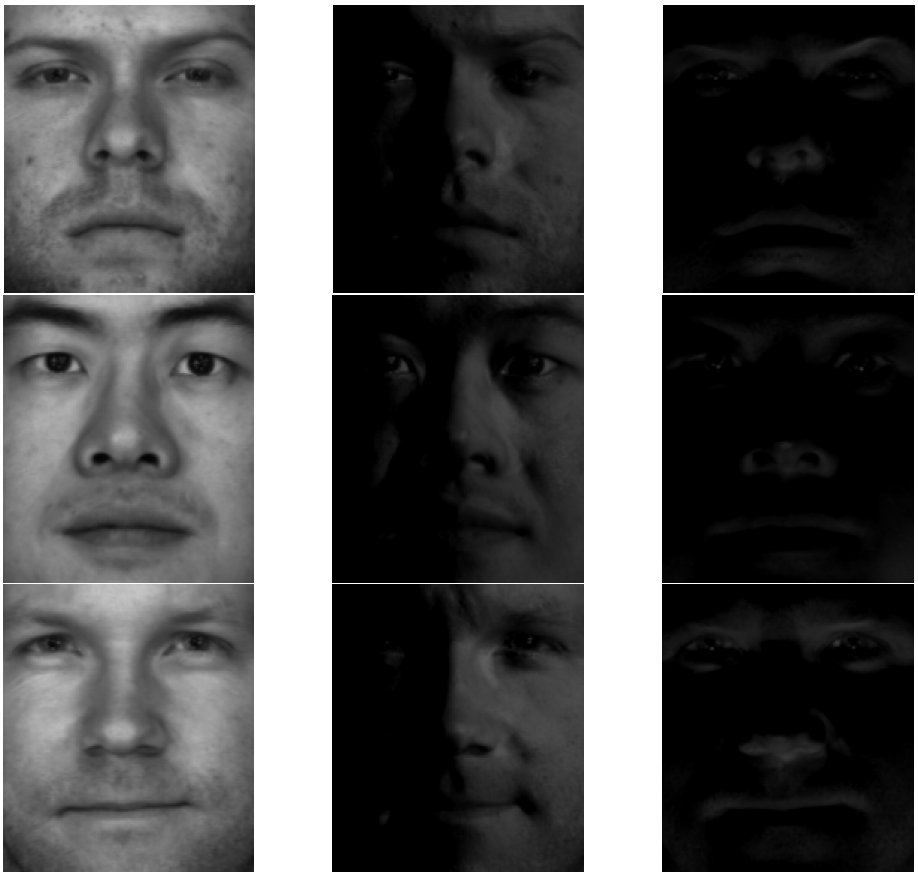

Figure 1: The mean face μ and the first two eigenfaces u_1, u_2 (from left to right) of 3 individuals are displayed from top to bottom. To enhance the visual effects, the two eigenfaces are scaled by their corresponding singular values.

From Figure 1, we can learn that the first eigenface u_1 differs from the mean face μ in terms of the illumination from the right (or horizontal illumination). On the other hand, u_2 differs from μ in terms of its lack of illumination from the top (or vertical illumination).

4.2 2.

Figure 2: Faces of $\mu + y_i u_i$ with $y_i = -\sigma_i, -0.8\sigma_i, ..., 0.8\sigma_i, \sigma_i$ (from left to right) of the three individuals (from top to bottom) are displayed. The first row ($i = 1$) of each individual corresponds to the variation along u_1, while the second row ($i = 2$) corresponds to the variation along u_2.

From Figure 2, we can clearly see that, for all three individuals, the first eigenface u_1 captures the information about the horizontal illumination, while the second eigenface u_2 captures that about the vertical illumination.