Home Assignment 5 Solutions

STAT 154/254: Modern Statistical Prediction & Machine Learning

Ajay Sharma — Fall 2024

Problem 1: Principal component analysis, formulation 1.

Let $(x_i)_{i\in[n]} \subseteq \mathbb{R}^d$ be the observed covariates and let

$$X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times d}.$$

Assume that $n > d$ and assume that $\frac{1}{n}\sum_{i=1}^n x_i = 0$. We let the singular value decomposition of X be given by

$$X = U\Sigma V^\top$$

where $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, $V \in \mathbb{R}^{d\times d}$ is another orthogonal matrix, and

$$\Sigma = \begin{bmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_d) \\ 0_{(n-d)\times d} \end{bmatrix} \in \mathbb{R}^{n\times d}.$$

We further assume that $\sigma_1 > \sigma_2 > \cdots > \sigma_d > 0$. Finally, we let $v_1, \ldots, v_d \in \mathbb{R}^d$ be the columns of V, that is,

$$V = [\, v_1, \ldots, v_d \,] \in \mathbb{R}^{d\times d}.$$

i. Recall that the leading principal component of the dataset $(x_i)_{i\in[n]} \subseteq \mathbb{R}^d$ is defined as

$$\arg\max_{\|\varphi_1\|_2=1} \frac{1}{n}\sum_{i=1}^n \langle x_i, \varphi_1\rangle^2 \;=\; \arg\max_{\|\varphi_1\|_2=1} \langle \varphi_1,\, X^\top X\, \varphi_1\rangle.$$

Prove that for any $v \in \{u \in \mathbb{R}^d : \|u\|_2^2 = 1\} \setminus \{v_1, -v_1\}$, we have $\langle v,\, X^\top X\, v\rangle < \sigma_1^2$. Argue why this implies that

$$\arg\max_{\|\varphi_1\|_2=1} \langle \varphi_1,\, X^\top X\, \varphi_1\rangle = \{v_1, -v_1\}.$$

ii. Recall that the second principal component of the dataset $(x_i)_{i\in[n]} \subseteq \mathbb{R}^d$ is defined as

$$\arg\max_{\substack{\|\varphi_2\|_2=1,\\ \langle \varphi_2, \varphi_1^*\rangle=0}} \frac{1}{n}\sum_{i=1}^n \langle x_i, \varphi_2\rangle^2 \;=\; \arg\max_{\substack{\|\varphi_2\|_2=1,\\ \langle \varphi_2, \varphi_1^*\rangle=0}} \langle \varphi_2,\, X^\top X\, \varphi_2\rangle,$$

where $\varphi_1^* = v_1$ is the leading principal component. Prove that for any

$$v \in \{u \in \mathbb{R}^d : \|u\|_2^2 = 1,\ \langle u, v_1\rangle = 0\} \setminus \{v_2, -v_2\},$$

we have $\langle v,\, X^\top X\, v\rangle < \sigma_2^2$. Argue why this implies that

$$\arg\max_{\substack{\|\varphi_2\|_2=1,\\ \langle \varphi_2, \varphi_1^*\rangle=0}} \langle \varphi_2,\, X^\top X\, \varphi_2\rangle = \{v_2, -v_2\}.$$

Solution. If the singular value decomposition of X equals $U\Sigma V^T$, then the eigenvalue decomposition of X^TX equals $V\Sigma^T\Sigma V^T$. That is, if we write

$$D = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2) \in \mathbb{R}^{d\times d},$$

then $X^TX = VDV^T$. Now write $v_1, \ldots, v_d \in \mathbb{R}^d$ for the columns of V, and note that this implies

$$VDV^T \;=\; \sum_{i=1}^{d} \sigma_i^2 \, v_i \, v_i^T .$$

For any vector $v \in \mathbb{R}^d$ we have:

$$\langle v,\, X^TX\, v\rangle = v^TX^TXv = v^T\Big(\sum_{i=1}^{d} \sigma_i^2 \, v_i \, v_i^T\Big)v = \sum_{i=1}^{d} \sigma_i^2 \, (v_i^Tv)^2.$$

Let us define $\alpha_i := v_i^Tv$ so that $\alpha_1, \ldots, \alpha_d \in \mathbb{R}$ are the coefficients of v in the orthonormal basis $v_1, \ldots, v_d$. Note that $\|v\|_2^2 = 1$ implies $\sum_{i=1}^d \alpha_i^2 = 1$.

i. By these calculations, we have

$$\langle v,\, X^TX\, v\rangle < \sigma_1^2 \iff \sum_{i=1}^{d} \big(\sigma_1^2 - \sigma_i^2\big)\alpha_i^2 > 0.$$

Since $v \notin \{v_1, -v_1\}$, we have $|\alpha_1| \neq 1$ and this implies that there exists some $i \geq 2$ such that $|\alpha_i| > 0$. Also, by assumption, $\sigma_1^2 - \sigma_i^2 > 0$ for $i \geq 2$. Since the product of positive terms is positive, it follows that at least one of

$$(\sigma_1^2 - \sigma_2^2)\alpha_2^2, \;\ldots,\; (\sigma_1^2 - \sigma_d^2)\alpha_d^2$$

is positive, hence the sum is positive. Now we can use this to prove

$$\arg\max_{\|\phi_1\|_2=1} \langle \phi_1,\, X^TX\, \phi_1\rangle = \{v_1, -v_1\}. \tag{1}$$

Indeed, take arbitrary $\phi_1 \in \mathbb{R}^d$ with $\|\phi_1\|_2 = 1$. If $\phi_1 \in \{v_1, -v_1\}$, then

$$\langle \pm v_1,\, X^TX\,(\pm v_1)\rangle = \langle v_1,\, X^TX\, v_1\rangle = \sigma_1^2,$$

by the calculation above. And, if $\phi_1 \notin \{v_1, -v_1\}$, then we just showed

$$\langle \phi_1,\, X^TX\, \phi_1\rangle < \sigma_1^2.$$

This means ϕ_1 cannot be the maximizer, since it is beaten by $\pm v_1$.

ii. Again by the above calculations, we have

$$\langle v,\, X^TX\, v\rangle < \sigma_2^2 \iff \sum_{i=1}^{d} \big(\sigma_2^2 - \sigma_i^2\big)\alpha_i^2 > 0.$$

Now suppose that $v \in \mathbb{R}^d$ has $v_1^Tv = 0$ and $v \notin \{v_2, -v_2\}$. This implies $\alpha_1 = 0$ and $|\alpha_2| \neq 1$, hence there exists some $i \geq 3$ such that $|\alpha_i| > 0$. Also, by assumption,

$\sigma_2^2 - \sigma_i^2 > 0$ for $i \geq 3$. Since the product of positive terms is positive, it follows that at least one of

$$(\sigma_2^2 - \sigma_3^2)\alpha_3^2, \ \ldots, \ (\sigma_2^2 - \sigma_d^2)\alpha_d^2$$

is positive, hence the sum is positive. Now we can use this to prove

$$\arg \max_{\substack{\|\phi_2\|_2=1, \\ \langle \phi_1, \phi_2 \rangle = 0}} \langle \phi_2, \, X^T X \, \phi_2 \rangle = \{v_2, -v_2\}. \tag{2}$$

Indeed, take arbitrary $\phi_2 \in \mathbb{R}^d$ with $\|\phi_2\|_2 = 1$ and $\langle \phi_1, \phi_2 \rangle = 0$. If $\phi_2 \in \{v_2, -v_2\}$, then

$$\langle \pm v_2, \, X^T X \, (\pm v_2) \rangle = \langle v_2, \, X^T X \, v_2 \rangle = \sigma_2^2,$$

by the calculation above. And, if $\phi_2 \notin \{v_2, -v_2\}$, then we just showed

$$\langle \phi_2, \, X^T X \, \phi_2 \rangle < \sigma_2^2.$$

This means ϕ_2 cannot be the maximizer, since it is beaten by $\pm v_2$.

Problem 2: Principal component analysis, formulation 2.

Consider the same setting as Q1. We would like to motivate PCA from a different perspective. Our goal is to find an M-dimensional subspace S, such that

$$\min_{S \subseteq \mathbb{R}^d,\, \dim(S)=M} \sum_{i=1}^{n} \| x_i - P_S x_i \|_2^2$$

is minimized, where $\mathcal{S}_M$ is the set of all M-dimensional subspaces of d-dimensional Euclidean space, and $P_S \in \mathbb{R}^{d \times d}$ is the projection matrix that projects the vector to the M-dimensional subspace S.

i We assume that $\{\phi_1, \ldots, \phi_M\} \subseteq \mathbb{R}^d$ are a set of orthonormal vectors i.e., $\phi_s^\top \phi_t = 1_{s=t}$ and let S be the subspace spanned by $\{\phi_1, \ldots, \phi_M\}$. Please show that for any $x \in \mathbb{R}^d$,

$$\| x - P_S x \|_2^2 = \min_{z_1, \ldots, z_M \in \mathbb{R}} \Big\| x - \sum_{k=1}^{M} z_k \phi_k \Big\|_2^2,$$

and show that

$$\sum_{i=1}^{n} \| x_i - P_S x_i \|_2^2 = \min_{(z_{ik})_{i \in [n],\, k \in [M]}} \sum_{i=1}^{n} \Big\| x_i - \sum_{k=1}^{M} z_{ik} \phi_k \Big\|_2^2.$$

ii Use the results in part (i) to show that

$$\min_{S \subseteq \mathbb{R}^d,\, \dim(S)=M} \sum_{i=1}^{n} \| x_i - P_S x_i \|_2^2 = \min_{V_M \in \mathcal{V}_M} \min_{Z \in \mathbb{R}^{n \times M}} \| X - Z V_M^\top \|_F^2,$$

where $\mathcal{V}_M = \{V_M = [v_1, \ldots, v_M] \in \mathbb{R}^{d \times M} :\ v_s^\top v_t = 1_{s=t}\}$.

Solution.

i. First we claim that for any $z_1, \ldots, z_M \in \mathbb{R}$, we have

$$\|x - P_S x\|_2^2 \le \left\| x - \sum_{k=1}^{M} z_k \phi_k \right\|_2^2.$$

To see this, we write $V \in \mathbb{R}^{d \times M}$ for the matrix with columns $\phi_1, \ldots, \phi_M$ and we write $z := (z_1, \ldots, z_M) \in \mathbb{R}^M$, so that the desired statement is equivalent to

$$\|x - P_S x\|_2^2 \le \|x - Vz\|_2^2.$$

Now expand both sides using $\|a - b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 - 2a^T b$ to get:

$$\|x\|_2^2 + \|VV^T x\|_2^2 - 2x^T VV^T x \ \le\ \|x\|_2^2 + \|Vz\|_2^2 - 2x^T Vz.$$

Rearranging this gives

$$0 \le \|V^T x\|_2^2 + \|z\|_2^2 - 2x^T Vz,$$

and the right side is equal to $\|V^T x - z\|_2^2$, so it must be non-negative. This proves the inequality. To conclude, it suffice to show that there exists some $z_1, \ldots, z_M \in \mathbb{R}$ satisfying

$$\|x - P_S x\|_2^2 = \left\| x - \sum_{k=1}^{M} z_k \phi_k \right\|_2^2.$$

By the calculation above, this holds if and only if $\|V^T x - z\|_2^2 = 0$, which is equivalent to $z = V^T x$. That is, we can simply set $z_k := \phi_k^T x$ for $k = 1, 2, \ldots, M$ and we establish the equality.

ii. Importantly, note that for any $V \in \mathcal{V}_M$ and $Z \in \mathbb{R}^{n \times M}$, if $z_1, \ldots, z_n \in \mathbb{R}^M$ represent the rows of Z, then we have

$$\|X - ZV^T\|_F^2 = \sum_{i=1}^{n} \|x_i - V z_i\|_2^2.$$

Thus, it suffices to show

$$\min_{S \in \mathcal{S}_M} \sum_{i=1}^{n} \|x_i - P_S x_i\|_2^2 = \min_{V \in \mathcal{V}_M} \min_{Z \in \mathbb{R}^{n \times M}} \sum_{i=1}^{n} \|x_i - V z_i\|_2^2.$$

First, let's show that the left side is less than or equal to the right side. That is, for an arbitrary $V \in \mathcal{V}_M$ and $Z \in \mathbb{R}^{n \times M}$, let's construct some $S \in \mathcal{S}_M$ such that

$$\sum_{i=1}^{n} \|x_i - P_S x_i\|_2^2 \le \sum_{i=1}^{n} \|x_i - V z_i\|_2^2.$$

To do this, we simply let $S := \mathrm{col}(V)$, and then the inequality follows from part (i). Second, let's show that the left side is greater than or equal to the right side. That is, for an arbitrary $S \in \mathcal{S}_M$, let's construct $V \in \mathcal{V}_M$ and $Z \in \mathbb{R}^{n \times M}$ such that

$$\sum_{i=1}^{n} \|x_i - P_S x_i\|_2^2 \ge \sum_{i=1}^{n} \|x_i - V z_i\|_2^2.$$

(In fact, we will get equality rather than inequality.) To do this, we'll let $\phi_{i \in [M]} \in \mathbb{R}^d$ denote any orthonormal basis for S, and, for each x_i, we let $z_i \in \mathbb{R}^M$ denote the vector whose kth entry is $\phi_k^T x_i$. (So, z_i is just the vector of coordinates of x_i when expressed in the partial basis $\phi_{i \in [M]} \in \mathbb{R}^d$.) Now let V be the matrix whose columns are $\phi_1, \ldots, \phi_M$, and let Z be the matrix whose rows are $z_1, \ldots, z_n$. By construction we have $P_S x_i = V z_i$. Thus, we have shown

$$\sum_{i=1}^{n} \|x_i - P_S x_i\|_2^2 = \sum_{i=1}^{n} \|x_i - V z_i\|_2^2,$$

and the result is proved.

Problem 3: K-means algorithm with A-norm.

Let $(x_i)_{i\in[n]} \subseteq \mathbb{R}^d$ be the observed covariates. Let $A \in \mathbb{R}^{d\times d}$ be a positive definite matrix (A is symmetric and all the eigenvalues of A are positive). Denote the A-norm of a vector $x \in \mathbb{R}^d$ by

$$\|x\|_A^2 = \langle x,\, Ax\rangle.$$

In the following, we derive the K-means clustering algorithm upon $(x_i)_{i\in[n]} \subseteq \mathbb{R}^d$ with the distance metric induced by the A-norm. (Hint: if you do not know how to derive the result in terms of A-norm, you can first consider the case when $A = I_d$.)

i. Define the within-cluster variation $\mathrm{WCV}(C_k)$ of a cluster $C_k \subseteq [n]$ by

$$\mathrm{WCV}(C_k) \;=\; \frac{1}{2\,|C_k|}\sum_{i,j\in C_k}\|x_i - x_j\|_A^2.$$

Prove that

$$\mathrm{WCV}(C_k) \;\equiv\; \sum_{i\in C_k}\|x_i - \bar{x}_{C_k}\|_A^2, \quad \text{where} \quad \bar{x}_{C_k} = \frac{1}{|C_k|}\sum_{j\in C_k} x_j.$$

ii. Let $(C_k)_{k\in[K]}$ be a partition of $[n]$. Define the K-means objective function by

$$R\big((C_k)_{k\in[K]}\big) = \sum_{k=1}^{K}\mathrm{WCV}(C_k).$$

Denote the set of weights

$$\mathcal{W} = \Big\{(w_{ik})_{i\in[n],\,k\in[K]} : \sum_{k=1}^{K} w_{ik} = 1,\ \forall i \in [n];\ w_{ik} \geq 0,\ \forall i,k\Big\}.$$

Prove that

$$\min_{(C_k)_{k\in[K]}} R\big((C_k)_{k\in[K]}\big) = \min_{(w_{ik})\in\mathcal{W}}\ \min_{(\mu_k)_{k\in[K]}} \overline{R}\big((w_{ik})_{i\in[n],\,k\in[K]},\, (\mu_k)_{k\in[K]}\big),$$

where

$$\overline{R}\big((w_{ik})_{i\in[n],\,k\in[K]},\, (\mu_k)_{k\in[K]}\big) = \sum_{i=1}^{n}\sum_{k=1}^{K}\|x_i - \mu_k\|_A^2\, w_{ik}.$$

In proving Eq. (1), please follow the steps: (1) prove the left-hand-side (of Eq. (1)) is less or equal to the right-hand-side; (2) prove the right-hand-side is also less or equal to the left-hand-side.

iii. For fixed $(w_{ik})_{i\in[n],\,k\in[K]} \in \mathcal{W}$, derive the expression of the minimizer $(\mu_k^*)_{k\in[K]}$ by

$$(\mu_k^*)_{k\in[K]} = \arg\min_{(\mu_k)_{k\in[K]}} \overline{R}\big((w_{ik})_{i\in[n],\,k\in[K]},\, (\mu_k)_{k\in[K]}\big).$$

iv. For fixed $(\mu_k)_{k\in[K]}$, derive the expression of the minimizer $(w_{ik}^*)_{i\in[n],\,k\in[K]}$ by

$$(w_{ik}^*)_{i\in[n],\,k\in[K]} = \arg\min_{(w_{ik})_{i\in[n],\,k\in[K]}\in\mathcal{W}} \overline{R}\big((w_{ik})_{i\in[n],\,k\in[K]},\, (\mu_k)_{k\in[K]}\big).$$

Solution.

i. We insert $0 = -\bar{x}_{C_k} + \bar{x}_{C_k}$ into each summand, and then we expand each term using

$$\|u - v\|_A^2 = \|u\|_A^2 + \|v\|_A^2 - 2\langle u, v\rangle_A$$

to get:

$$\mathrm{WCV}(C_k) = \frac{1}{2|C_k|}\sum_{i,j\in C_k}\|x_i - x_j\|_A^2 = \frac{1}{2|C_k|}\sum_{i,j\in C_k}\big\|x_i - \bar{x}_{C_k} - (x_j - \bar{x}_{C_k})\big\|_A^2$$

$$= \frac{1}{2|C_k|}\sum_{i,j\in C_k}\Big(\|x_i - \bar{x}_{C_k}\|_A^2 + \|x_j - \bar{x}_{C_k}\|_A^2 - 2\langle x_i - \bar{x}_{C_k},\, x_j - \bar{x}_{C_k}\rangle_A\Big)$$

$$= \frac{1}{2|C_k|}\sum_{i,j\in C_k}\|x_i - \bar{x}_{C_k}\|_A^2 + \frac{1}{2|C_k|}\sum_{i,j\in C_k}\|x_j - \bar{x}_{C_k}\|_A^2 - \frac{1}{|C_k|}\sum_{i,j\in C_k}\langle x_i - \bar{x}_{C_k},\, x_j - \bar{x}_{C_k}\rangle_A$$

$$= \frac{1}{2}\sum_{i\in C_k}\|x_i - \bar{x}_{C_k}\|_A^2 + \frac{1}{2}\sum_{j\in C_k}\|x_j - \bar{x}_{C_k}\|_A^2 - \frac{1}{|C_k|}\sum_{i,j\in C_k}\langle x_i - \bar{x}_{C_k},\, x_j - \bar{x}_{C_k}\rangle_A.$$

Notice that the first two terms are identical; they only differ in the choice of indexing variable. Also, the second term vanishes, since we can use linearity of the inner product $\langle\cdot,\cdot\rangle_A$ to get:

$$\frac{1}{|C_k|}\sum_{i,j\in C_k}\langle x_i - \bar{x}_{C_k},\, x_j - \bar{x}_{C_k}\rangle_A = |C_k|\cdot\frac{1}{|C_k|^2}\sum_{i,j\in C_k}\langle x_i - \bar{x}_{C_k},\, x_j - \bar{x}_{C_k}\rangle_A$$

$$= |C_k|\Big\langle \tfrac{1}{|C_k|}\sum_{i\in C_k}(x_i - \bar{x}_{C_k}),\ \tfrac{1}{|C_k|}\sum_{j\in C_k}(x_j - \bar{x}_{C_k})\Big\rangle_A = |C_k|\langle 0, 0\rangle_A = 0.$$

Therefore, we have

$$\mathrm{WCV}(C_k) = \frac{1}{2|C_k|}\sum_{i,j\in C_k}\|x_i - x_j\|_A^2 = \sum_{i\in C_k}\|x_i - \bar{x}_{C_k}\|_A^2,$$

as claimed.

ii. First let us show that

$$\min_C R(C) \ \le\ \min_W \min_{\mu_1,\dots,\mu_K} \overline{R}(W, \mu_1, \dots, \mu_K).$$

To do this, we take arbitrary $W, \mu_1, \dots, \mu_K$, and we will use this to find a partition C such that $R(C) \le \overline{R}(W, \mu_1, \dots, \mu_K)$. For $k \in \{1, \dots, K\}$, let

$$C_k := \{\, 1 \le i \le n : \|x_i - \mu_k\|_A \le \|x_i - \mu_\ell\|_A \text{ for all } \ell \in \{1, \dots, K\}\}.$$

Note that C_k is just the set of indices of data points which are closer to μ_k than to any $\{\mu_\ell\}_{\ell\neq k}$. Now we make two observations:

- For any $a_1, \dots, a_n \in \mathbb{R}$ and $p_1, \dots, p_n \ge 0$ with $\sum_{i=1}^n a_i p_i \ge \min_i p_i$.
- For any $b_1, \dots, b_n \in \mathbb{R}^d$ and $\mu \in \mathbb{R}^d$ we have $\sum_{i=1}^n \|b_i - \bar{b}\|_A^2 \le \sum_{i=1}^n \|b_i - \mu\|_A^2$.

We use these observations, we interchange the order of summation, and apply part (i) to get:

$$\overline{R}(W,\mu_1,\dots,\mu_K) = \sum_{i=1}^{n}\sum_{k=1}^{K} \|x_i-\mu_k\|_A^2\, w_{ik} \;\geq\; \sum_{i=1}^{n}\sum_{k=1}^{K} \|x_i-\mu_k\|_A^2\, \mathbf{1}\{i\in C_k\}$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{n} \|x_i-\mu_k\|_A^2\, \mathbf{1}\{i\in C_k\} = \sum_{k=1}^{K}\sum_{i\in C_k} \|x_i-\mu_k\|_A^2 \;\geq\; \sum_{k=1}^{K} \mathrm{WCV}(C_k) = R(C).$$

Second let us show that

$$\min_C R(C) \;\geq\; \min_W \min_{\mu_1,\dots,\mu_K} \overline{R}(W,\mu_1,\dots,\mu_K).$$

To do this, we take an arbitrary partition C, we find some $W,\mu_1,\dots,\mu_K$ such that $R(C)\geq \overline{R}(W,\mu_1,\dots,\mu_K)$. In fact, we will find $W,\mu_1,\dots,\mu_K$ such that $R(C)=\overline{R}(W,\mu_1,\dots,\mu_K)$. To do this, we simply define

$$w_{ik} = \mathbf{1}\{i\in C_k\} \quad \text{and} \quad \mu_k = \bar{x}_{C_k} \quad \text{for } k\in\{1,\dots,K\},\ 1\leq i\leq n.$$

Then use part (i) to get

$$\begin{aligned}
\overline{R}(W,\mu_1,\dots,\mu_K) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \|x_i-\mu_k\|_A^2\, w_{ik} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n} \|x_i-\mu_k\|_A^2\, w_{ik} \\
&= \sum_{k=1}^{K}\sum_{i\in C_k} \|x_i-\bar{x}_{C_k}\|_A^2 \\
&= \sum_{k=1}^{K} \mathrm{WCV}(C_k) \\
&= R(C).
\end{aligned}$$

This proves

$$\min_C R(C) \;=\; \min_W \min_{\mu_1,\dots,\mu_K} \overline{R}(W,\mu_1,\dots,\mu_K).$$

as claimed.

iii. Suppose that W is fixed. For each $k\in\{1,\dots,K\}$ let us write

$$\begin{aligned}
f_k(\mu) &:= \sum_{i=1}^{n} \|x_i-\mu\|_A^2\, w_{ik} \\
&= \sum_{i=1}^{n} \big(\|x_i\|_A^2 + \|\mu\|_A^2 - 2\langle x_i,\mu\rangle_A\big)\, w_{ik} \\
&= \sum_{i=1}^{n} \big(x_i^T A x_i + \mu^T A\mu - 2x_i^T A\mu\big)\, w_{ik},
\end{aligned}$$

so that

$$\overline{R}(W, \mu_1, \dots, \mu_K) = \sum_{k=1}^{K} f_k(\mu_k).$$

Since each of $\mu_1, \dots, \mu_K$ appears in only one term of the sum, we have

$$\arg\min_{\mu_1,\dots,\mu_K} \overline{R}(W, \mu_1, \dots, \mu_K) = \big(\arg\min_{\mu} f_1(\mu_1), \dots, \arg\min_{\mu} f_K(\mu_K)\big).$$

Now we can fix $k \in \{1, \dots, K\}$ and take the gradient:

$$\nabla_{\mu_k} f_k = \sum_{i=1}^{n} \nabla_{\mu_k} \big(x_i^T A x_i + \mu^T A \mu - 2 x_i^T A \mu\big)\, w_{ik} = \sum_{i=1}^{n} (2A\mu - 2Ax_i)\, w_{ik}.$$

Since f_k is smooth and convex, its unique stationary point must be a minimizer. Thus, we see

$$\nabla_{\mu_k} f_k(\mu_k^*) = 0 \qquad \Longleftrightarrow \qquad \mu_k^* = \frac{\sum_{i=1}^{n} x_i\, w_{ik}}{\sum_{i=1}^{n} w_{ik}}.$$

Therefore, the minimizer of $(\mu_1, \dots, \mu_K) \mapsto \overline{R}(W, \mu_1, \dots, \mu_K)$ given $W = (w_{ik})_{i \in [n], k \in [K]}$ is

$$\Big(\tfrac{\sum_{i=1}^{n} x_i\, w_{i1}}{\sum_{i=1}^{n} w_{i1}},\ \dots,\ \tfrac{\sum_{i=1}^{n} x_i\, w_{iK}}{\sum_{i=1}^{n} w_{iK}} \Big).$$

iv. Suppose that $\mu_1, \dots, \mu_K$ are fixed. Let us write

$$\Delta_K := \{(w_1, \dots, w_K) \in \mathbb{R}^K : \sum_{k=1}^{K} w_k = 1\}$$

for the probability simplex on K elements, and define the function $h_i : \Delta_K \to \mathbb{R}$,

$$h_i(w) = \sum_{k=1}^{K} \|x_i - \mu_k\|_A^2\, w_k,$$

so that we have

$$\overline{R}(W, \mu_1, \dots, \mu_K) = \sum_{i=1}^{n} h_i(w_i)$$

where $w_1, \dots, w_n$ are the rows of W. Since each term appears in only one term of the sum, we have

$$\arg\min_{W} \overline{R}(W, \mu_1, \dots, \mu_K) = \big(\arg\min_{w_1} h_1(w_1), \dots, \arg\min_{w_n} h_n(w_n)\big).$$

Now fix $i \in [n]$ and consider minimizing $h_i(w)$. Since w are just the weights assigned to some non-negative numbers, the function h_i is minimized when these weights concentrate on the minimizer. That is,

$$\arg\min_{w_i} h_i(w_i) = (0, \dots, 0, 1, 0, \dots, 0)$$

where the 1 appears in the position $\arg\min_{k \in \{1,\dots,K\}} \|x_i - \mu_k\|_A$.