Regression in heterogeneous problems

Hanwen Huang

Department of Epidemiology and Biostatistics

University of Georgia, Athens, GA 30602

Supplementary Material

This note contains technical proofs of Theorems 1 and 2.

Proof of Theorem 1

We first need to show that $\hat{A}_n(K)$ lie in a sufficiently large closed ball in $\mathbb{R}^{2d}$, i.e. we need to show that both the estimated regression coefficients and the estimated cluster centers lie in a closed ball $B(R)$ centered at the origin and of radius R when n is large enough. Note that since $n^{1/2}\lambda_n \to 0$, the minimization of (8) is equivalent to the minimization of

$$\Phi(A, P_n), \quad s.t. \quad \sum_{k=1}^{K}\sum_{j=1}^{d}\frac{|\beta_{kj}|}{|\tilde{\beta}_{kj}|} \leq s_n, \tag{S1}$$

where $s_n \to \infty$ as $n \to \infty$. By the fact that $s_n \to \infty$, it is enough if we can establish the asymptotic consistency for the estimation in the unpenalized framework.

Note that we need to prove the finiteness of both the estimated regression coefficients and the estimated cluster centers. For cluster centers, we can employ the techniques used in Pollard (1981) for proving the consistency of cluster centers in k-

means clustering algorithm. For regression coefficients, special care need to be considered. Let A_n denote the optimal subset which satisfy $\Phi(A_n, P_n) = m_K(P_n)$. Then A_n is the solution of the unpenalized composite regression model (3) since solving $\min_{C(\cdot),(\boldsymbol{\beta}_1,\boldsymbol{\mu}_1),\cdots,(\boldsymbol{\beta}_K,\boldsymbol{\mu}_K)} \sum_{i=1}^n \{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_{C(i)})^2 + \tau\|\mathbf{x}_i - \boldsymbol{\mu}_{C(i)}\|^2\}/n$ is equivalent to finding a $A \in \mathcal{A}_K$ such that $\Phi(A, P_n)$ is minimized. We need to first show that there exists a sufficiently large closed ball which contains all the estimated parameters A_n when n is sufficiently large. Then the desired strong consistent results can be obtained using the property of uniform convergence of $\Phi(A, P_n)$ to $\Phi(A, P)$ within the closed ball.

According to strong law of large number (SLLN), for any fixed A, we have $\Phi(A, P_n)$ $\to \Phi(A, P)$. The first step is to show that there is at least one point of A_n contained in a closed ball. By definition $\Phi(A_n, P_n) \leq \Phi(A, P_n)\ \forall A \in \mathcal{A}_K$. Choose $A = A_0$ which consists of a single point at the origin, i.e. $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{\mu}_0 = \mathbf{0}$. Then

$$\Phi(A_0, P_n) = \int (y^2 + \|\mathbf{x}\|^2) P_n(d\mathbf{x}, dy) \to \int (y^2 + \|\mathbf{x}\|^2) P(d\mathbf{x}, dy) = \Phi(A_0, P). \quad \text{(S2)}$$

For any given $A_n = \{(\boldsymbol{\beta}_1, \boldsymbol{\mu}_1), \cdots, (\boldsymbol{\beta}_K, \boldsymbol{\mu}_K)\} \in \mathcal{A}_K$. Denote $C(\cdot)$ the corresponding partitioning of samples into K groups. The feature space $\mathbb{R}^d$ can be divided into K distinct regions B_k, $k = 1, \cdots, K$ such that if $\mathbf{x}_i \in B_k$ then $C(i) = k$. If, for infinity many values of n, no point of the estimated cluster centers $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$ were

contained in $B(M)$, then

$$\begin{aligned}
\Phi(A_n, P_n) &= \frac{1}{n}\sum_{k=1}^{K}\sum_{C(i)=k}\{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k)^2 + \tau\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2\} \\
&> \frac{\tau}{n}\sum_{k=1}^{K}\sum_{C(i)=k}\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\
&= \frac{\tau}{n}\sum_{k=1}^{K}\sum_{C(i)=k}\left(\|\mathbf{x}_i\|^2 - 2\mathbf{x}_i^T\boldsymbol{\mu}_k + \|\boldsymbol{\mu}_k\|^2\right) \\
&> \frac{\tau}{n}\sum_{k=1}^{K}\sum_{C(i)=k}\left(\|\boldsymbol{\mu}_k\|^2 - 2\|\mathbf{x}_i\|\|\boldsymbol{\mu}_k\|\right) > \Phi(A_0, P)
\end{aligned}$$

for large enough M since $\frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i|$ is finite. This would make $\Phi(A_n, P_n) > \Phi(A_0, P_n)$ infinitely often: a contradiction. Denote $\mathbf{u}_k = \boldsymbol{\beta}_k/|\boldsymbol{\beta}_k| \in \mathcal{O}$ the unit vector for $\boldsymbol{\beta}_k$. Define cluster index $k_n^\star = \text{argmax}_k\left(\sum_{C(i)=k}|\mathbf{x}_i^T\mathbf{u}_k|^2\right)$ and function $\phi(\mathbf{x}, A_n) = \sum_{k=1}^{K}|\mathbf{x}^T\mathbf{u}_k|^2\mathbf{I}(\mathbf{x} \in B_k)$. If for infinity many values of n, no point of the estimated regression coefficients $(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K)$ were contained in $B(M)$, then

$$\Phi(A_n, P_n) = \frac{1}{n}\sum_{k=1}^{K}\sum_{C(i)=k}\{(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k)^2 + \tau\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2\} > \frac{1}{n}\sum_{C(i)=k_n^\star}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_{k_n^\star})^2$$

$$\geq \frac{1}{n}\sum_{C(i)=k_n^\star}\{y_i^2 - 2|y_i|\|\mathbf{x}_i\|\|\boldsymbol{\beta}_{k_n^\star}\| + (\mathbf{x}_i^T\boldsymbol{\beta}_{k_n^\star})^2\} \geq \frac{1}{n}\sum_{C(i)=k_n^\star}\|\boldsymbol{\beta}_{k_n^\star}\|^2|\mathbf{x}_i^T\mathbf{u}_{k_n^\star}|^2 - a\|\boldsymbol{\beta}_{k_n^\star}\|$$

$$\geq \frac{1}{K}\frac{1}{n}\sum_{k=1}^{K}\sum_{C(i)=k}\|\boldsymbol{\beta}_{k_n^\star}\|^2|\mathbf{x}_i^T\mathbf{u}_k|^2 - a\|\boldsymbol{\beta}_{k_n^\star}\| = \frac{\|\boldsymbol{\beta}_{k_n^\star}\|^2}{K}\int\phi(\mathbf{x}, A_n)Q_n(d\mathbf{x}) - a\|\boldsymbol{\beta}_{k_n^\star}\|$$

$$= \frac{\|\boldsymbol{\beta}_{k_n^\star}\|^2}{K}\left\{\int\phi(\mathbf{x}, A_n)(Q_n(d\mathbf{x}) - Q(d\mathbf{x})) + \int\phi(\mathbf{x}, A_n)Q(d\mathbf{x})\right\} - a\|\boldsymbol{\beta}_{k_n^\star}\|,$$

where a is defined in Condition 2. Because $\int\phi(\mathbf{x}, A_n)Q(d\mathbf{x}) < \infty$ and the sample paths of Q_n can get uniformly closer to Q, the first term $\int\phi(\mathbf{x}, A_n)(Q_n(d\mathbf{x}) - Q(d\mathbf{x}))$

approaches to 0 for large enough n . Therefore according to the definition of $w_K(Q)$ in (10) and Condition 1,

$$\begin{aligned}\Phi(A_n, P_n) &> \frac{\|\boldsymbol{\beta}_{k_n^\star}\|^2}{K}\int \phi(\mathbf{x}, A_n)Q(d\mathbf{x}) - a\|\boldsymbol{\beta}_{k_n^\star}\| \\ &\geq \frac{w_K(P)}{K}\|\boldsymbol{\beta}_{k_n^\star}\|^2 - a\|\boldsymbol{\beta}_{k_n^\star}\| > \Phi(A_0, P)\end{aligned}$$

for large enough M. This would also make $\Phi(A_n, P_n) > \Phi(A_0, P_n)$ infinitely often: a contradiction.

We use inductive method. The theorem can be proved for $K = 1$. Assume the conclusions of the theorem are valid for $1, 2, \cdots, K-1$ clusters. For $K > 1$ clusters, if some points in A_n are not eventually contained in $B(M)$, we can obtain a set of $K-1$ or less points by assigning the data belonging to the clusters outside $B(M)$ to the cluster inside $B(M)$. From previous results, the closed ball $B(M)$ of radius M and centered at the origin contains at least one point of A_n for n large enough. Choose $\epsilon > 0$ to satisfy $m_K(P)+\epsilon < m_{K-1}(P)$. Denote $f_k(\mathbf{x}, y) = (y-\mathbf{x}^T\boldsymbol{\beta}_k)^2+\tau\|\mathbf{x}-\boldsymbol{\mu}_k\|^2$ for $k = 1, \cdots, K$. Without loss of generality, assume $(\boldsymbol{\beta}_1, \boldsymbol{\mu}_1)$ is inside $B(M)$ and $\boldsymbol{\mu}_K$ or $\boldsymbol{\beta}_K$ is outside $B(M)$, then the increasing due to assigning data in cluster K to cluster 1 is at most

$$\begin{aligned}E_n &= \int f_1(\mathbf{x}, y)\mathbf{I}\left(f_1(\mathbf{x}, y) > f_K(\mathbf{x}, y)\right) P_n(d\mathbf{x}, dy) \\ &= \int f_1(\mathbf{x}, y)\mathbf{I}\left(f_1(\mathbf{x}, y) > f_K(\mathbf{x}, y)\right)\{\mathbf{I}(\|\mathbf{x}\|^2 + y^2 > S) \\ &\quad +\mathbf{I}(\|\mathbf{x}\|^2 + y^2 < S)\}P_n(d\mathbf{x}, dy) \qquad \text{(S3)}\end{aligned}$$

Since $\|\boldsymbol{\beta}_1\| < M$ and $\|\boldsymbol{\mu}_1\| < M$, the first term of (S3) is smaller than $\epsilon/2$ when S is

large enough according to Condition 2. For the second term, we can show that

$$\begin{aligned} & f_1(\mathbf{x}, y)\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S) \\ \leq \; & \{(S + SM)^2 + \tau(\|\mathbf{x}\|^2 + 2SM + M^2)\}\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S) \end{aligned}$$

and

$$f_K(\mathbf{x}, y)\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S) \geq \{\tau(\|\mathbf{x}\|^2 - 2S\|\boldsymbol{\mu}_K\| + \|\boldsymbol{\mu}_K\|^2)\}\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S).$$

If $\boldsymbol{\mu}_K$ is outside $B(M)$, the second term of (S3)

$$\begin{aligned} & \int f_1(\mathbf{x}, y)\mathrm{I}\left(f_1(\mathbf{x}, y) > f_K(\mathbf{x}, y)\right)\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S)P_n(d\mathbf{x}, dy) \\ \leq \; & \int f_1(\mathbf{x}, y)\mathrm{I}\left((S + SM)^2/\tau + (M + S)^2 > (|\boldsymbol{\mu}_K| - S)^2\right) \\ & \mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S)P_n(d\mathbf{x}, dy) \end{aligned}$$

which is 0 when $|\boldsymbol{\mu}_K| > C_1 = S + \sqrt{(S + SM)^2/\tau + (M + S)^2}$. Similarly we can show

$$f_1(\mathbf{x}, y)I(\|\mathbf{x}\|^2 + y^2 < S) \leq \{y^2 + 2S^2M + S^2M^2 + \tau(S + M)^2\}\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S)$$

and

$$f_K(\mathbf{x}, y)\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S) \geq (y^2 - 2S|\mathbf{x}^T\boldsymbol{\beta}_K| + |\mathbf{x}^T\boldsymbol{\beta}_K|^2)\mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S).$$

If $\boldsymbol{\beta}_K$ is outside $B(M)$, the second term of (S3) is

$$\begin{aligned} & \int f_1(\mathbf{x},y) I\left(f_1(\mathbf{x},y) > f_K(\mathbf{x},y)\right) I(\|\mathbf{x}\|^2 + y^2 < S) P_n(d\mathbf{x},dy) \qquad \text{(S4)} \\ \leq & \int f_1(\mathbf{x},y) \mathrm{I}\left\{S^2(M+1)^2 + \tau(S+M)^2 > (|\mathbf{x}^T\boldsymbol{\beta}_K| - S)^2\right\} \\ & \mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S) P_n(d\mathbf{x},dy) \\ \leq & \int f_1(\mathbf{x},y) \mathrm{I}\left(|\mathbf{x}^T\boldsymbol{\beta}_K| < C_2\right)\right) \mathrm{I}(\|\mathbf{x}\|^2 + y^2 < S) \\ & \{(P_n(d\mathbf{x},dy) - P(d\mathbf{x},dy)) + P(d\mathbf{x},dy)\}, \end{aligned}$$

where $C_2 = S + \sqrt{S^2(M+1)^2 + \tau(S+M)^2}$. The first term of (S4) approaches to zero for large enough n. According to Condition 3 the second term of (S4) is smaller than $\epsilon/2$ if $|\boldsymbol{\beta}_K| > C_2/\delta_0$. Therefore from (S3), we get $E_n < \epsilon$ if either $\boldsymbol{\mu}_K$ or $\boldsymbol{\beta}_K$ is outside of the closed ball $B(R)$ with $R = \max(C_1, C_2/\delta_0)$. The set $A_n^\star$ obtained by deleting from A_n all points outside $B(R)$ is a candidate for minimizing $\Phi(\cdot, P_n)$ over sets of $K-1$ of fewer points; it is therefore beaten by the optimal set B_n of $K-1$ points. Thus

$$\Phi(A_n^\star, P_n) \geq \Phi(B_n, P_n), \qquad \text{(S5)}$$

which by the inductive hypothesis, converges almost surely to $m_{K-1}(P)$. If $A_n \notin B(R)$ along some subsequence $\{n_i\}$ of values of n, we therefore get

$$\begin{aligned} m_{K-1}(P) &= \lim \Phi(B_{n_i}, P_{n_i}) \leq \liminf_{n_i} \Phi(A_{n_i}^\star, P_{n_i}) \leq \limsup_{n_i} \{\Phi(A_{n_i}, P_{n_i}) + E_{n_i}\} \\ &\leq \limsup_{n_i} \Phi(\bar{A}, P_{n_i}) + \epsilon = m_K(P) + \epsilon, \end{aligned}$$

which is a contradiction to $m_K(P) + \epsilon < m_{K-1}(P)$. Therefore, $B(R)$ contains all the estimated parameters A_n when n is sufficiently large.

Now we prove the uniform SLLN. Define $\mathcal{E}_K = \{A \in B(R) : A \text{ contains } K \text{ points}\}$ the collection of all finite subsets of $B(R)$ which contains K points. For n large enough, it suffices to search for $A_n \in \mathcal{E}_K$. Now we can show that the function $\Phi(A, P)$ is continuous on $\mathcal{E}_K$. The convergence is determined by the Hausdorff metric $H(\cdot, \cdot)$. For $A, A' \in \mathcal{E}_K$, if $H(A, A') < \delta$, to each $(\boldsymbol{\beta}, \boldsymbol{\mu}) \in A$, there is a point $(\boldsymbol{\beta}'(\boldsymbol{\beta}), \boldsymbol{\mu}'(\boldsymbol{\mu})) \in A'$ such that $\|\boldsymbol{\beta}'(\boldsymbol{\beta}) - \boldsymbol{\beta}\| < \delta$ and $\|\boldsymbol{\mu}'(\boldsymbol{\mu}) - \boldsymbol{\mu}\| < \delta$. Define $f_{(\boldsymbol{\beta},\boldsymbol{\mu})}(\mathbf{x}, y) = (y - \mathbf{x}^T\boldsymbol{\beta})^2 + \tau\|\mathbf{x} - \boldsymbol{\mu}\|^2$. Then

$$
\begin{aligned}
& \Phi(A, P) - \Phi(A', P) \\
= & \int \left\{ \min_{(\boldsymbol{\beta}',\boldsymbol{\mu}') \in A'} f_{(\boldsymbol{\beta}',\boldsymbol{\mu}')}(\mathbf{x}, y) - \min_{(\boldsymbol{\beta},\boldsymbol{\mu}) \in A} f_{(\boldsymbol{\beta},\boldsymbol{\mu})}(\mathbf{x}, y) \right\} P(d\mathbf{x}, dy) \\
\leq & \int \max_{(\boldsymbol{\beta},\boldsymbol{\mu}) \in A} \left[f_{(\boldsymbol{\beta}'(\boldsymbol{\beta}),\boldsymbol{\mu}'(\boldsymbol{\mu}))}(\mathbf{x}, y) - f_{(\boldsymbol{\beta},\boldsymbol{\mu})}(\mathbf{x}, y) \right] P(d\mathbf{x}, dy) \\
\leq & \int \{2R\|\mathbf{x}\|^2\delta + \|\mathbf{x}\|^2\delta^2 + 2\|\mathbf{x}\||y|\delta + \tau(2\|\mathbf{x}\|\delta + \delta^2 + 2R\delta)\} P(d\mathbf{x}, dy) \\
\leq & 2Ra\delta + a\delta^2 + a\delta + \tau(2a\delta + \delta^2 + 2R\delta), \qquad \text{(S6)}
\end{aligned}
$$

which is less than ϵ if δ is chosen small enough. Here a is defined in (12). The other inequality needed for proving the continuity is obtained by exchanging the roles of A and A'. Similarly we can prove the continuity for empirical measure P_n when n is large enough.

Define F_δ a finite subset of $B(R)$ such that each point of $B(R)$ is within a distance δ of at least one point of F_δ. Write $\mathcal{E}_{K,\delta}$ for $\{A \in \mathcal{E}_K; A \subseteq F_\delta\}$. Given a $A = \{(\boldsymbol{\beta}_1, \boldsymbol{\mu}_1), \cdots, (\boldsymbol{\beta}_K, \boldsymbol{\mu}_K)\}$ in $\mathcal{E}_K$, there exists a $A' = \{(\boldsymbol{\beta}'_1, \boldsymbol{\mu}'_1), \cdots, (\boldsymbol{\beta}'_K, \boldsymbol{\mu}'_K)\}$ in

$\mathcal{E}_{K,\delta}$ with $H(A, A') < \delta$. Corresponding to each function

$$\phi_A(\mathbf{x}, y) = \min_{(\boldsymbol{\beta},\boldsymbol{\mu})\in A} f_{(\boldsymbol{\beta},\boldsymbol{\mu})}(\mathbf{x}, y),$$

define two functions

$$\begin{aligned}
\bar{\phi}_A(\mathbf{x}, y) &= \min_{(\boldsymbol{\beta},\boldsymbol{\mu})\in A'} \{(|y - \mathbf{x}^T\boldsymbol{\beta}| + \delta_1)^2 + \tau(\|\mathbf{x} - \boldsymbol{\mu}\| + \delta_1)^2\} \\
\underline{\phi}_A(\mathbf{x}, y) &= \min_{(\boldsymbol{\beta},\boldsymbol{\mu})\in A'} \{(|y - \mathbf{x}^T\boldsymbol{\beta}| - \delta_1)^2 + \tau(\|\mathbf{x} - \boldsymbol{\mu}\| - \delta_1)^2\}.
\end{aligned}$$

The continuity of $\Phi(A, P)$ allows us to choose appropriate δ_1 such that

$$\int \underline{\phi}_A(\mathbf{x}, y)P(d\mathbf{x}, dy) \le \int \phi_A(\mathbf{x}, y)P(d\mathbf{x}, dy) \le \int \bar{\phi}_A(\mathbf{x}, y)P(d\mathbf{x}, dy),$$

and

$$\int \underline{\phi}_A(\mathbf{x}, y)P_n(d\mathbf{x}, dy) \le \int \phi_A(\mathbf{x}, y)P_n(d\mathbf{x}, dy) \le \int \bar{\phi}_A(\mathbf{x}, y)P_n(d\mathbf{x}, dy)$$

when n is large enough. Therefore for any $A \in \mathcal{E}_K$, we have

$$\begin{aligned}
&\left|\int \phi_A dP_n - \int \phi_A dP\right| \le \max\left\{\left|\int \bar{\phi}_A dP_n - \int \underline{\phi}_A dP\right|, \left|\int \bar{\phi}_A dP - \int \underline{\phi}_A dP_n\right|\right\} \\
&\le \left|\int \bar{\phi}_A dP - \int \underline{\phi}_A dP\right| + \max\left\{\left|\int \bar{\phi}_A dP_n - \int \bar{\phi}_A dP\right|, \left|\int \underline{\phi}_A dP_n - \int \underline{\phi}_A dP\right|\right\}.
\end{aligned}$$

From the fact that $\bar{\phi}_A, \underline{\phi}_A$ are from $\mathcal{E}_{K,\delta}$ which is a finite set, the second term can be made less that $\epsilon/2$ for large enough n. The first term can be made less than $\epsilon/2$ by

choosing small enough δ_1. This completes the proof of uniform SLLN

$$\sup_{A \in \mathcal{E}_k} |\Phi(A, P_n) - \Phi(A, P)| \to 0. \tag{S7}$$

From assumption, we have

$$\Phi(A_n, P_n) \leq \Phi(\bar{A}, P_n). \tag{S8}$$

The right hand side $\Phi(\bar{A}, P_n) \to \Phi(\bar{A}, P)$. Apply uniform SLLN (S7) to the left hand side, we have $\Phi(A_n, P_n) - \Phi(A_n, P) \to 0$. Therefore $\Phi(A_n, P) \leq \Phi(\bar{A}, P)$ for n large enough. On the other hand, according to assumption, $\Phi(A_n, P) \geq \Phi(\bar{A}, P)$. So $\Phi(A_n, P_n) \to \Phi(\bar{A}, P)$ and $A_n \to \bar{A}$. This completes the proof of the consistency for the estimation in the unpenalized framework. By Condition 4 and the fact that $s_n \to \infty$, the solution $\hat{A}_n$ based on the penalized formula (8) will eventually approach to A_n and thus to $\bar{A}$.

Proof of Theorem 2

First we need to prove that the map $A \to \Phi(A, P)$ has a second derivative. The function $\Phi(\cdot, \cdot)$ defined in (6) can be equivalently expressed as

$$\begin{aligned} \Phi(A, P) &= \int \sum_{k=1}^{K} \phi_k(\mathbf{x}, y) I(B_k) P(d\mathbf{x}, dy) \\ &= \sum_{k=1}^{K} \int \phi_k(\mathbf{x}, y) \prod_{j \neq k} I(G_{kj}(\mathbf{x}, y) > 0) P(d\mathbf{x}, dy). \end{aligned}$$

Use a $2Kd$ dimensional vector

$$\boldsymbol{\Delta}(A,P) = \{(\Delta_{\boldsymbol{\beta}_1}(A,P), \Delta_{\boldsymbol{\mu}1}(A,P)), \cdots, (\Delta_{\boldsymbol{\beta}_K}(A,P), \Delta_{\boldsymbol{\mu}K}(A,P))\}$$

to denote the first derivative of $\Phi(A,P)$ over A such that

$$\begin{aligned}
\Delta_{\boldsymbol{\beta}_k}(A,P) &\equiv \frac{\partial \Phi(A,P)}{\partial \boldsymbol{\beta}_k} = \int I(B_k) \frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\beta}_k} P(d\mathbf{x},dy), \\
\Delta_{\boldsymbol{\mu}_k}(A,P) &\equiv \frac{\partial \Phi(A,P)}{\partial \boldsymbol{\mu}_k} = \int I(B_k) \frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\mu}_k} P(d\mathbf{x},dy).
\end{aligned}$$

Now take the second derivative with respect to $\boldsymbol{\beta}$, we have

$$\begin{aligned}
\frac{\partial^2 \Phi(A,P)}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_k} &= \int I(B_k) \frac{\partial^2 \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_k^T} P(d\mathbf{x},dy) \\
&\quad - \sum_{j \neq k} \sigma_{kj} \left(\frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\beta}_k} \frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\beta}_k^T} \frac{f(x,y)}{|m_{kj}(\mathbf{x},y)|} \right), \\
\frac{\partial^2 \Phi(A,P)}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_j} &= -\sigma_{kj} \left(\frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\beta}_k} \frac{\partial \phi_j(\mathbf{x},y)}{\partial \boldsymbol{\beta}_j^T} \frac{f(x,y)}{|m_{kj}(\mathbf{x},y)|} \right),
\end{aligned}$$

where $f(\mathbf{x},y)$ is the density function of the distribution $P(\mathbf{x},y)$ and σ_{kj} is the integration over the boundary surface between B_k and B_j. Similarly, we have the second derivative with respect to $\boldsymbol{\mu}$

$$\begin{aligned}
\frac{\partial^2 \Phi(A,P)}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k} &= \int I(B_k) \frac{\partial^2 \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\beta}_k^T} P(d\mathbf{x},dy) \\
&\quad - \sum_{j \neq k} \sigma_{kj} \left(\frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\mu}_k} \frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\mu}_k^T} \frac{f(x,y)}{|m_{kj}(\mathbf{x},y)|} \right), \\
\frac{\partial^2 \Phi(A,P)}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_j} &= -\sigma_{kj} \left(\frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\mu}_k} \frac{\partial \phi_j(\mathbf{x},y)}{\partial \boldsymbol{\mu}_j^T} \frac{f(x,y)}{|m_{kj}(\mathbf{x},y)|} \right).
\end{aligned}$$

The cross term is

$$\frac{\partial^2 \Phi(A,P)}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\mu}_j} = -\sigma_{kj} \left(\frac{\partial \phi_k(\mathbf{x},y)}{\partial \boldsymbol{\beta}_k} \frac{\partial \phi_j(\mathbf{x},y)}{\partial \boldsymbol{\mu}_j^T} \frac{f(x,y)}{|m_{kj}(\mathbf{x},y)|} \right).$$

According to Condition 6, all the integrations exist. This completes the proof that the map $A \rightarrow \Phi(A,P)$ has a second derivative. For empirical measure P_n, we can decompose

$$\Phi(A_n, P_n) = \Phi(A_n, P) + \Phi(A_n, P_n - P). \tag{S9}$$

Denote $r_n = \|\mathbf{v}(A_n) - \mathbf{v}(\bar{A})\|$, where $\mathbf{v}(A_n)$ and $\mathbf{v}(\bar{A})$ are vectorized A_n and $\bar{A}$. According to the differentiability of $\Phi(A,P)$, the first term of (S9) can be written as

$$\begin{aligned} \Phi(A_n, P) &= \Phi(\bar{A}, P) + (\mathbf{v}(A_n) - \mathbf{v}(\bar{A}))^T \boldsymbol{\Delta}(\bar{A}, P) \\ &+ \frac{1}{2}(\mathbf{v}(A_n) - \mathbf{v}(\bar{A}))^T \boldsymbol{\Gamma} (\mathbf{v}(A_n) - \mathbf{v}(\bar{A})) + o(r_n^2). \end{aligned} \tag{S10}$$

The second term in (S10) vanishes because $\bar{A}$ minimizes $\Phi(A,P)$. Here $\boldsymbol{\Gamma}$ is a $2Kd \times 2Kd$ matrix representing the second order derivative of $\Phi(A,P)$ over A evaluated at $\bar{A}$. Define $X_n = n^{1/2}(P_n - P)$ as the empirical process associated with the empirical measure P_n. The second term of (S9) can be written as

$$\begin{aligned} & \Phi(A_n, P_n - P) \\ = \; & n^{-1/2} \Phi(A_n, X_n) \\ = \; & n^{-1/2} [\Phi(\bar{A}, X_n) + (v(A_n) - v(\bar{A}))^T \boldsymbol{\Delta}(\bar{A}, X_n) + o(r_n)]. \end{aligned} \tag{S11}$$

Define vector $\mathbf{Z}^{(n)} = -\boldsymbol{\Delta}(\bar{A}, X_n)$ which has an asymptotic normal distribution with mean vector

$$-\boldsymbol{\Delta}(\bar{A}, P) = 0$$

and variance matrix $\mathbf{V}$ is

$$V = P\{\boldsymbol{\Delta}(\bar{A}, \cdot)\boldsymbol{\Delta}(\bar{A}, \cdot)^T\}. \tag{S12}$$

Substitute (S9), (S10) and (S11) into (8), we have

$$\begin{aligned} W(\hat{A}_n, P_n) &= W(\bar{A}, P_n) - n^{-\frac{1}{2}}\mathbf{Z}^{(n)T}(\mathbf{v}(\hat{A}_n) - \mathbf{v}(\bar{A})) \\ &\quad + \frac{1}{2}(\mathbf{v}(\hat{A}_n) - \mathbf{v}(\bar{A}))^T\Gamma(\mathbf{v}(\hat{A}_n) - \mathbf{v}(\bar{A})) \\ &\quad + \sum_{k=1}^{K}\sum_{j=1}^{d}\lambda_n\frac{|\hat{\beta}_{kj}^{(n)}| - |\bar{\beta}_{kj}|}{|\tilde{\beta}_{kj}^{(n)}|} + o_p(n^{-\frac{1}{2}}r_n) + o_p(r_n^2) \\ &\leq W(\bar{A}, P_n). \end{aligned} \tag{S13}$$

From Theorem 1, we have $\mathbf{v}(\hat{A}_n) - \mathbf{v}(\bar{A}) = o(1)$ and $r_n = o(1)$. Therefore

$$O_p(n^{-\frac{1}{2}}r_n) + O_p(r_n^2) + o_p(n^{-\frac{1}{2}}r_n) + o_p(r_n^2) + \sum_{(k,j)\in\mathcal{B}}\lambda_n\frac{|\hat{\beta}_{kj}^{(n)}| - |\bar{\beta}_{kj}|}{|\tilde{\beta}_{kj}^{(n)}|} \leq 0$$

which leads to

$$O_p(n^{-\frac{1}{2}}r_n) + O_p(r_n^2) + O_p(\lambda_n r_n) \leq 0.$$

From the fact that $n^{\frac{1}{2}}\lambda_n \to 0$, we have $r_n = O_p(n^{-\frac{1}{2}})$. For $(k,j) \in \mathcal{B}^c$, $\bar{\beta}_{kj} = 0$, $\tilde{\beta}_{kj}^{(n)} = O_p(n^{-\frac{1}{2}})$, we have $\frac{\lambda_n}{\tilde{\beta}_{kj}^{(n)}} = O(n^{\frac{1}{2}}\lambda_n)$. Assume $\hat{\beta}_{kj}^{(n)} \neq 0$, taking the derivative over $\hat{\beta}_{kj}^{(n)}$ on both side of (S13) gives

$$\begin{aligned}\frac{\partial W(\hat{A}_n, P_n)}{\partial \hat{\beta}_{kj}^{(n)}} &= -n^{-\frac{1}{2}} Z_{kj}^{(n)} + \Gamma_{kj}(\hat{\beta}_{kj}^{(n)} - \bar{\beta}_{kj}) + \frac{\lambda_n}{\tilde{\beta}_{kj}^{(n)}} \text{sign}(\hat{\beta}_{kj}^{(n)}) + o_p(n^{-\frac{1}{2}}) \\ &= n^{-\frac{1}{2}}[O_p(1) + O_p(1) + O_p(n\lambda_n)\text{sign}(\hat{\beta}_{kj}^{(n)})].\end{aligned}$$

Since $n\lambda_n \to \infty$, the third term is bigger than the first two terms, this is a contradiction to Karush-Kuhn-Tucker condition, thus $\hat{\beta}_{kj}^{(n)} = 0$ for all $(k,j) \in \mathcal{B}^c$. This proves the oracle property of the estimator.

References

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics 9*, 135–140.