Sparse k-Means with ℓ_∞/ℓ_0 Penalty for High-Dimensional Data Clustering

Xiangyu Chang[1], Yu Wang[2], Rongjian Li[3] and Zongben Xu[1]

[1]*Xi'an Jiaotong University,* [2]*University of California, Berkeley and* [3]*Old Dominion University*

Supplementary Material

In this material, we provide the detailed proofs of the proposed 4 theorems in the main context.

S1 Complement Lemmas

We provide some useful lemmas that support our proofs in this section. In Lemma 1 we reformulate BCSS for facilitating our derivation. Lemma 2 gives a concentration inequality of a non-central χ^2 random variable. Lemma 3 calculates an important expectation which will be used in the proof of Theorem 3 and 4.

Lemma 1. *Under the same setting we have described at subsection 2.3 of the main context, we can obtain a_j denoted in (3) of main context has the reformulation*

$$a_j = \sum_{k=1}^{K} \left(\frac{\sum_{i \in C_k} x_{ij}}{\sqrt{n\tilde{\pi}_k}}\right)^2 - \left(\frac{\sum_{i=1}^{n} x_{ij}}{\sqrt{n}}\right)^2, \tag{S1.1}$$

where $n_k, k = 1, 2, \ldots, K$ is the number of sample size in cluster C_k and $\tilde{\pi}_k \triangleq n_k/n$. Therefore,

$$BCSS(\mathcal{C}) = \sum_{j=1}^{p} a_j = \sum_{j=1}^{p} \Big\{ \sum_{k=1}^{K} \left(\frac{\sum_{i \in C_k} x_{ij}}{\sqrt{n\tilde{\pi}_k}}\right)^2 - \left(\frac{\sum_{i=1}^{n} x_{ij}}{\sqrt{n}}\right)^2 \Big\}.$$

Proof. Based on the definition of $a_j, j = 1, 2, \ldots, p$, we have

$$\begin{aligned} a_j &= \frac{1}{2n}\sum_{i_1,i_2}(x_{i_1j} - x_{i_2j})^2 - \sum_{k=1}^{K}\frac{1}{2n_k}\sum_{i_1,i_2\in C_k}(x_{i_1j} - x_{i_2j})^2 \\ &= \sum_i x_{ij}^2 - \frac{1}{n}(\sum_i x_{ij})^2 - \sum_{k=1}^{K}(\sum_{i\in C_k} x_{ij}^2 - \frac{1}{n_k}(\sum_{i\in C_k} x_{ij})^2) \\ &= -\frac{1}{n}(\sum_i x_{ij})^2 + \sum_{k=1}^{K}\frac{1}{n_k}(\sum_{i\in C_k} x_{ij})^2 \\ &= \sum_{k=1}^{K}(\frac{\sum_{i\in C_k} x_{ij}}{\sqrt{n\tilde{\pi}_k}})^2 - (\frac{\sum_{i=1}^{n} x_{ij}}{\sqrt{n}})^2. \end{aligned} \tag{S1.2}$$

□

Lemma 2. *Suppose $Y \in \mathbb{R}^m$ is a random vector with standard multivariate normal distribution. $A \in \mathbb{R}^{m\times m}$ is a matrix and $b \in \mathbb{R}^m$ is a vector. Then $Z = \|AY + b\|^2$ obeys sub-exponential distribution with parameters $(2\sqrt{|||AA^T|||_F^2 + 2\|A^T b\|^2}, \left|\left|\left|A^T A\right|\right|\right|_*)$. If we denote δ to be the spectral norm $\left|\left|\left|A^T A\right|\right|\right|_*$, we can also use the parameters $(2\sqrt{m\delta^2 + 2\delta\|b\|^2}, \delta)$. Then we have the concentration inequality*

$$P(|Z - \mathbb{E}Z| \geq t) \leq \begin{cases} exp(-\frac{t^2}{8(m\delta^2+2\delta\|b\|^2)}) & if\ 0 \leq t \leq \frac{4(m\delta^2+2\delta\|b\|^2)}{\delta} \\ exp(-\frac{t}{2\delta}) & if\ t \geq \frac{4(m\delta^2+2\delta\|b\|^2)}{\delta} \end{cases}.$$

Proof. Note that $\|AY + b\|^2$ obeys a non-central χ^2 distribution, whose cumulative distribution function is explicit. Then the moment generating function can be deducted and the lemma can be proved (Foss et al., 2011). □

Lemma 3. *Recall that $F(\mathcal{C}, \mathbf{w})$ is defined in (18) of main context and data is generated from (12) of main context and. For any partition $\mathcal{C} = \{C_1, \ldots, C_K\}$, let $\tilde{\pi}_k = \frac{|C_k|}{n}$ for $k = 1, \ldots, K$, and $\tilde{\mu}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k}\sum_{k'=1}^{K}\phi_{ik'}\mu_{k'j}$. Then the conditional expectation for fixed ϕ_{ik} would be $\mathbb{E}_z F(\mathcal{C}, \mathbf{w}) = K\|\mathbf{w}\|_1 + \sum_{j=1}^{p^*} w_j \sum_{k=1}^{K} n\tilde{\pi}_k\tilde{\mu}_{kj}^2$.*

Proof. We analyze the distribution of the objective function $F(\mathcal{C}, \mathbf{w})$. For any j, k and fixed ϕ_{ik} $(i = 1, \ldots, K)$, it is obvious that

$$\frac{1}{\sqrt{|C_k|}} \sum_{i \in C_k} x_{ij} \sim \mathcal{N}(\sqrt{n\tilde{\pi}_k} \cdot \tilde{\mu}_{kj}, 1).$$

Thus $\sum_{k=1}^{K} \left(\frac{1}{\sqrt{|C_k|}} \sum_{i \in C_k} x_{ij} \right)^2$ has the same distribution as $\|Y + b_j\|^2$ where Y obeys $\mathcal{N}(0, \mathbf{I}_{K \times K})$, $b_{jk} = \sqrt{n\tilde{\pi}_k} \cdot \tilde{\mu}_{kj}$. We further assume that the eigen decomposition of $\Sigma \bigotimes \mathbf{I}_{K \times K} = U\Lambda^2 U^T$, where $\bigotimes$ is the Kronecker product. Denote $L = U\Lambda$, then we know $F(\mathcal{C}, w)$ has the same distribution as $\|W(LY + b)\|^2$, where $W = diag(\sqrt{w_j}) \bigotimes \mathbf{I}_{K \times K}$.

The expectation of $F(\mathcal{C}, w)$ is

$$\begin{aligned}
\mathbb{E}F(\mathcal{C}, w) =& tr(L^T W^2 L) + \|Wb\|^2 && \text{(S1.3)} \\
=& tr(W^2 L L^T) + \|Wb\|^2 && \text{(S1.4)} \\
=& tr(W^2 \Sigma \bigotimes \mathbf{I}_{K \times K}) + \|Wb\|^2 && \text{(S1.5)} \\
=& K\|w\|_1 + \sum_{j=1}^{p} w_j \sum_{k=1}^{K} n\tilde{\pi}_k \tilde{\mu}_{kj}^2. && \text{(S1.6)}
\end{aligned}$$

□

S2 Proof of Theorem 1

Proof. we omit the proof since it is easy to obtain. □

S3 Proof of Theorem 2

Proof. Based on Lemma 1, the expectation of the $BCSS$ for the jth feature is

$$\mathbb{E}a_j(\mathcal{C}) = \mathbb{E}\sum_{k=1}^{K}\left(\frac{\sum_{i\in C_k} x_{ij}}{\sqrt{n\tilde{\pi}_k}}\right)^2 - \left(\frac{\sum_{i=1}^{n} x_{ij}}{\sqrt{n}}\right)^2 \tag{S3.7}$$

$$= n\sum_{k=1}^{K}\tilde{\pi}_k\tilde{\mu}_{kj}^2 - n\left(\sum_{k=1}^{K}\tilde{\pi}_k\tilde{\mu}_{kj}\right)^2 + K - 1, \tag{S3.8}$$

where $\tilde{\pi}_k = \frac{C_k}{n}$ is the proportion of the size of kth cluster C_k and $\tilde{\mu}_k = \frac{1}{|C_k|}\sum_{i\in C_k}\sum_{k'=1}^{K}\phi_{ik'}\mu_{k'}$ is the expectation of the sample mean in cluster C_k.

For $p^* < j \leq p$, we have $\mathbb{E}x_{ij} = 0$. This shows $\tilde{\mu}_{kj} = 0$. Therefore we know they are noise features $\mathbb{E}a_j(\mathcal{C}) = K - 1, \forall\mathcal{C}$. For other features $j \leq p^*$, consider $\mathbb{E}a_j(\mathcal{C}^*) = n\sum_{k=1}^{K}\pi_k\mu_{kj}^2 - n(\sum_{k=1}^{K}\pi_k\mu_{kj})^2 + K - 1$. So, we can denote $c_j = n\sum_{k=1}^{K}\pi_k\mu_{kj}^2 - n(\sum_{k=1}^{K}\pi_k\mu_{kj})^2 > 0$ holds because of the convexity of function x^2.

$\square$

S4 Proof of Theorem 3

Proof. Let $\mathcal{C}^* = (\mathcal{C}_1, \dots, \mathcal{C}_K)$ to be the partition defined by the Gaussian mixture model parameter ϕ_{ik}. If $\phi_{ik} = 1$, which means x_i is drawn from the kth component of Gaussian mixture model, then $\mathbf{x}_i$ is in C_k. As $n \to \infty$, $|C_k|/n \to \pi_k$ almost surely independent of the dimension p. Therefore, without loss of generality, we assume $|C_k| = n \times \pi_k$ for $k = 1, \dots, K$. Define Δ to satisfy the following equation:

$$s = \frac{\sum_{j=1}^{p^*}\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta p^*}{\sqrt{\sum_{j=1}^{p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2}}.$$

Define $w_j^* = \frac{\mathbb{E}\bar{a}_j(\mathcal{C}^*)-\Delta}{\sqrt{\sum_{j\le p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*)-\Delta)^2}}$. The proof can be summarized as the following chain of inequalities,

$$
\begin{aligned}
&P(\widehat{\mathbf{w}} \text{ has SCP}) && \text{(S4.9)}\\
\ge &P\left(\sup_{j=1,\ldots,p^*}|\widehat{w}_j - w_j^*| < \min_{j=1,\ldots,p^*} w_j^*\right) && \text{(S4.10)}\\
\ge &P\left(\sup_{\mathcal{C},\|\mathbf{w}\|_1\le s}|F(\mathcal{C},\mathbf{w}) - \mathbb{E}F(\mathcal{C},\mathbf{w})| < cn\right) && \text{(S4.11)}\\
\ge &1 - pK^n\exp(-\frac{nc^2}{24s^2\sigma_2}), && \text{(S4.12)}
\end{aligned}
$$

where $c = \frac{1}{4n}\sqrt{\sum_{j\le p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*)-\Delta)^2}\min_{j=1,\ldots,p^*} w_j^{*2} > 0$ is a constant. When $p^{*2} \le \frac{\sigma_1^4}{6400\sigma_2^3\ln(K)}$ and

$$
\frac{\sum_{j=1}^{p^*}\sum_{k=1}^{K}\pi_k\mu_{kj}^2 - \frac{1}{2}\sigma_1 p^*}{\sqrt{\sum_{j=1}^{p^*}(\sum_{k=1}^{K}\pi_k\mu_{kj}^2 - \frac{1}{2}\sigma_1)^2}} \le s \le \frac{\sum_{j=1}^{p^*}\sum_{k=1}^{K}\pi_k\mu_{kj}^2}{\sqrt{\sum_{j=1}^{p^*}(\sum_{k=1}^{K}\pi_k\mu_{kj}^2)^2}},
$$

since the relation between s and Δ, we know $K + n\frac{1}{2}\sigma_1 \ge \Delta \ge K$. Because c is lower bounded by

$$
\begin{aligned}
c = &\frac{1}{4n}\sqrt{\sum_{j\le p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*)-\Delta)^2}\min_{j=1,\ldots,p^*} w_j^{*2} && \text{(S4.13)}\\
= &\frac{1}{4n}\min_{j=1,\ldots,p^*}\frac{(\mathbb{E}\bar{a}_j(\mathcal{C}^*)-\Delta)^2}{\sqrt{\sum_{j\le p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*)-\Delta)^2}} && \text{(S4.14)}\\
\ge &\frac{(n\sigma_1 + K - \Delta)^2}{4p^* n(n\sigma_2 + K - \Delta)} && \text{(S4.15)}\\
\ge &\frac{\sigma_1^2}{16\sqrt{p^*}\sigma_2}, && \text{(S4.16)}
\end{aligned}
$$

and $s^2 \le p$, we know

$$
\frac{c^2}{25s^2\sigma_2} \ge \frac{c^2}{25p^*\sigma_2} \ge \frac{\sigma_1^4}{6400p^{*2}\sigma_2^3} \ge \ln(K). \qquad \text{(S4.17)}
$$

Thus when $\ln(p) = o(n)$, the last term goes to 0, the proof is complete.

Now we turn to the proof of (S4.10-S4.12). The inequality (S4.10) is trivial, so we only prove (S4.11) and (S4.12).

Proof of inequality (S4.11): It suffices to prove that

$$\left\{ \sup_{\mathcal{C}, \|\mathbf{w}\|_1 \le s} |F(\mathcal{C}, \mathbf{w}) - \mathbb{E}F(\mathcal{C}, \mathbf{w})| < \frac{1}{4}\sqrt{\sum_{j \le p^*} (\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2} \min_{j=1,\dots,p^*} w_j^{*2} \right\} \tag{S4.18}$$

$$\Longrightarrow \left\{ \sup_{j=1,\dots,p^*} |\widehat{w}_j - w_j^*| < \min_{j=1,\dots,p^*} w_j^* \right\}. \tag{S4.19}$$

We have the following line of inequalities:

$$\begin{aligned}
\mathbb{E}F(\mathcal{C}^*, \mathbf{w}^*) \le & F(\mathcal{C}^*, \mathbf{w}^*) + \frac{1}{4}\sqrt{\sum_{j \le p^*} (\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2} \min_{j=1,\dots,p^*} w_j^{*2} && \text{(S4.20)} \\
\le & F(\widehat{\mathcal{C}}, \widehat{\mathbf{w}}) + \frac{1}{4}\sqrt{\sum_{j \le p^*} (\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2} \min_{j=1,\dots,p^*} w_j^{*2} && \text{(S4.21)} \\
\le & \mathbb{E}F(\widehat{\mathcal{C}}, \widehat{\mathbf{w}}) + \frac{1}{2}\sqrt{\sum_{j \le p^*} (\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2} \min_{j=1,\dots,p^*} w_j^{*2} && \text{(S4.22)} \\
\le & \mathbb{E}F(\mathcal{C}^*, \widehat{\mathbf{w}}) + \frac{1}{2}\sqrt{\sum_{j \le p^*} (\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2} \min_{j=1,\dots,p^*} w_j^{*2}. && \text{(S4.23)}
\end{aligned}$$

Denote $d = \widehat{\mathbf{w}} - \mathbf{w}^*$. Since $\widehat{\mathbf{w}}$ and $\mathbf{w}^*$ are both in Ω_1, d must satisfy

$$\begin{aligned}
&\sum_{j \le p^*} d_j + \sum_{j > p^*} d_j \le 0, \\
&\sum_{j \le p^*} w_j^* d_j \le -\frac{1}{2} \sum_{j \le p^*} d_j^2, \\
&d_j \ge 0 \quad \forall j > p^*,
\end{aligned}$$

Thus we have

$$\mathbb{E}F(\mathcal{C}^*, \widehat{\mathbf{w}}) - \mathbb{E}F(\mathcal{C}^*, \mathbf{w}^*) = \sum_{j=1}^{p} \mathbb{E}\bar{a}_j(\mathcal{C}^*)d_j \tag{S4.24}$$

$$\leq \Delta \sum_{j=1}^{p^*} d_j - \frac{1}{2}\sqrt{\sum_{j\leq p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2 \sum_{j\leq p^*} d_j^2} + \sum_{j=p^*+1}^{p} \mathbb{E}\bar{a}_j(\mathcal{C}^*)d_j \tag{S4.25}$$

$$\leq (\Delta - K)\sum_{j=1}^{p^*} d_j - \frac{1}{2}\sqrt{\sum_{j\leq p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2 \sum_{j\leq p^*} d_j^2} \tag{S4.26}$$

$$\leq -\frac{1}{2}\sqrt{\sum_{j\leq p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2 \sum_{j\leq p^*} d_j^2} \tag{S4.27}$$

$$\leq -\frac{1}{2}\sqrt{\sum_{j\leq p^*}(\mathbb{E}\bar{a}_j(\mathcal{C}^*) - \Delta)^2 \sup_{j=1,\ldots,p^*} d_j^2}. \tag{S4.28}$$

Combining (S4.23) and (S4.28), we get the result.

Proof of inequality (S4.12): It suffices to prove

$$P\left(\sup_{\mathcal{C},\|\mathbf{w}\|_1\leq s} |F(\mathcal{C}, \mathbf{w}) - \mathbb{E}F(\mathcal{C}, \mathbf{w})| \geq cn\right) \tag{S4.29}$$

$$\leq pK^n \exp(-\frac{nc^2}{24s^2\sigma_2}). \tag{S4.30}$$

Since $\mathcal{C}$ can have at most K^n choices, we have

$$P\left(\sup_{\mathcal{C},\|\mathbf{w}\|_1\leq s} |F(\mathcal{C}, \mathbf{w}) - \mathbb{E}F(\mathcal{C}, \mathbf{w})| \geq cn\right)$$

$$\leq K^n \sup_{\mathcal{C}} P\left(\sup_{\|\mathbf{w}\|_1\leq s} |F(\mathcal{C}, \mathbf{w}) - \mathbb{E}F(\mathcal{C}, \mathbf{w})| \geq cn\right). \tag{S4.31}$$

Using the dual norm, we actually have that

$$\sup_{\|\mathbf{w}\|_1\leq s} |F(\mathcal{C}, \mathbf{w}) - \mathbb{E}F(\mathcal{C}, \mathbf{w})| = s \cdot \sup_{j\in 1,\ldots,p} |\bar{a}_j(\mathcal{C}) - \mathbb{E}\bar{a}_j(\mathcal{C})|. \tag{S4.32}$$

Therefore, (S4.31) can be bounded by

$$\begin{aligned} & K^n \sup_{\mathcal{C}} P\left(\sup_{\|\mathbf{w}\|_1 \le s} |F(\mathcal{C},\mathbf{w}) - \mathbb{E}F(\mathcal{C},\mathbf{w})| \ge cn\right) \\ \le & K^n \sup_{\mathcal{C}} P\left(\sup_{j\in 1,\ldots,p} |\bar{a}_j(\mathcal{C}) - \mathbb{E}\bar{a}_j(\mathcal{C})| \ge \frac{c}{s}n\right) \quad \text{(S4.33)} \\ \le & pK^n \sup_{\mathcal{C}, j=1,\ldots,p} P\left(|\bar{a}_j(\mathcal{C}) - \mathbb{E}\bar{a}_j(\mathcal{C})| \ge \frac{c}{s}n\right). \quad \text{(S4.34)} \end{aligned}$$

$\bar{a}_j = \sum_{k=1}^K \left(\frac{1}{\sqrt{|C_k|}} \sum_{i\in C_k} x_{ij}\right)^2$ has the same distribution as $\|Y + b_j\|^2$ where Y obeys $\mathcal{N}(0, \mathbf{I}_{K\times K})$, $b_{jk} = \sqrt{n\tilde{\pi}_k \tilde{\mu}_{kj}^2}$ for $j = 1, \ldots, p^*$ and $b_{jk} = 0$ for $j > p^*$. By lemma 2, we know $\bar{a}_j$ are all sub exponential variables with parameter $(2\sqrt{K + 2n\sigma_2}, 1)$. Note that $c < \sigma_2$ and $s \ge 1$,

$$\frac{c}{s}n \le n\sigma_2 \le 4(K + 2n\sigma_2).$$

Therefore when $n \ge \frac{K}{\sigma_2}$, i.e. $\sigma_2 n > K$, the last term could be bounded by

$$\exp(-\frac{nc^2}{24s^2\sigma_2}). \quad \text{(S4.35)}$$

This completes the proof. □

S5 Proof of Theorem 4

Proof. Similar to the proof of Theorem 3, we assume $|C_k| = n \times \pi_k$ for $k = 1, \ldots, K$. Then the proof can be summarized as the following chain of inequalities,

$$\begin{aligned} & P(\widehat{\mathbf{w}} \text{ has SCP}) \\ \ge & P\left(\sup_{\mathcal{C}, \mathbf{w}\in\Omega_2} |F(\mathcal{C},\mathbf{w}) - \mathbb{E}F(\mathcal{C},\mathbf{w})| < \frac{1}{2}n\sigma_1\right) \quad \text{(S5.36)} \\ \ge & 1 - pK^n \exp(-\frac{n\sigma_1^2}{96s^2\sigma_2}). \quad \text{(S5.37)} \end{aligned}$$

Under the theorem conditions, similar to theorem 1, we can prove the last term goes to 0. Now we only prove (S5.36-S5.37).

Proof of inequality (S5.36): It suffices to prove that

$$\{\widehat{\mathbf{w}} \text{ does not have SCP}\} \Longrightarrow \left\{ \sup_{\mathcal{C},\mathbf{w}\in\Omega_2} |F(\mathcal{C},\mathbf{w}) - \mathbb{E}F(\mathcal{C},\mathbf{w})| \geq \frac{1}{2}n\sigma_1 \right\}.$$

If $\widehat{\mathbf{w}}$ does not have SCP, then there exist features j_1, j_2 s.t. $j_1 > p^*$ is a noise feature where $\widehat{w}_{j_1} \neq 0$ and $j_2 < p^*$ is a relevant feature where $\widehat{w}_{j_2} = 0$. Consider $\tilde{\mathbf{w}}$ such that

$$\tilde{w}_j = \begin{cases} \widehat{w}_j & j \neq j_1, j_2, \\ \widehat{w}_{j_2} & j = j_1, \\ \widehat{w}_{j_1} & j = j_2. \end{cases}$$

Note that $\tilde{\mathbf{w}}$ is in Ω_2, too. By lemma 3 and Theorem 1,

$$\begin{aligned} \mathbb{E}F(\mathcal{C}^*, \tilde{\mathbf{w}}) - \mathbb{E}F(\widehat{\mathcal{C}}, \widehat{\mathbf{w}}) =& Ks + n\sum_{j=1}^{p^*} \tilde{w}_j \sum_{k=1}^{K} \pi_k \mu_{kj}^2 - Ks - n\sum_{j=1}^{p^*} \widehat{w}_j \sum_{k=1}^{K} \tilde{\pi}_k \tilde{\mu}_{kj}^2 && \text{(S5.38)} \\ \geq& n\sum_{j=1}^{p^*} (\tilde{w}_j - \widehat{w}_j) \sum_{k=1}^{K} \pi_k \mu_{kj}^2 && \text{(S5.39)} \\ =& n\widehat{w}_{j_1} \sum_{k=1}^{K} \pi_k \mu_{kj_2}^2 && \text{(S5.40)} \\ \geq& n\sigma_1. && \text{(S5.41)} \end{aligned}$$

On the other hand, $F(\mathcal{C}^*, \tilde{\mathbf{w}}) \leq F(\widehat{\mathcal{C}}, \widehat{\mathbf{w}})$ because $(\widehat{\mathcal{C}}, \widehat{\mathbf{w}})$ is optimal. Therefore,

$$\sup_{\mathcal{C},\mathbf{w}\in\Omega_2} |F(\mathcal{C},\mathbf{w}) - \mathbb{E}F(\mathcal{C},\mathbf{w})| > \frac{1}{2}n\sigma_1.$$

Thus we know the first inequality holds.

Proof of inequality (S5.37): It suffices to prove

$$P\left(\sup_{\mathcal{C},\mathbf{w}\in\Omega_2}|F(\mathcal{C},\mathbf{w})-\mathbb{E}F(\mathcal{C},\mathbf{w})|\geq\frac{1}{2}n\sigma_1\right) \tag{S5.42}$$

$$\leq pK^n\exp(-\frac{n\sigma_1^2}{96s^2\sigma_2}). \tag{S5.43}$$

Since $\mathcal{C}$ can have at most K^n choices. Therefore, we have

$$P\left(\sup_{\mathcal{C},\mathbf{w}\in\Omega_2}|F(\mathcal{C},\mathbf{w})-\mathbb{E}F(\mathcal{C},\mathbf{w})|\geq\frac{1}{2}n\sigma_1\right)$$

$$\leq K^n\sup_{\mathcal{C}}P\left(\sup_{\mathbf{w}\in\Omega_2}|F(\mathcal{C},\mathbf{w})-\mathbb{E}F(\mathcal{C},\mathbf{w})|\geq\frac{1}{2}n\sigma_1\right). \tag{S5.44}$$

Using the dual norm,

$$\sup_{\mathbf{w}\in\Omega_2}|F(\mathcal{C},\mathbf{w})-\mathbb{E}F(\mathcal{C},\mathbf{w})|=s\cdot\sup_{j\in1,\ldots,p}|\bar{a}_j(\mathcal{C})-\mathbb{E}\bar{a}_j(\mathcal{C})|. \tag{S5.45}$$

Therefore, (S5.44) can be bounded by

$$K^n\sup_{\mathcal{C}}P\left(\sup_{\mathbf{w}\in\Omega_2}|F(\mathcal{C},\mathbf{w})-\mathbb{E}F(\mathcal{C},\mathbf{w})|\geq\frac{1}{2}n\sigma_1\right)$$

$$\leq K^n\sup_{\mathcal{C}}P\left(\sup_{j\in1,\ldots,p}|\bar{a}_j(\mathcal{C})-\mathbb{E}\bar{a}_j(\mathcal{C})|\geq\frac{1}{2s}n\sigma_1\right) \tag{S5.46}$$

$$\leq pK^n\sup_{\mathcal{C},j=1,\ldots,p}P\left(|\bar{a}_j(\mathcal{C})-\mathbb{E}\bar{a}_j(\mathcal{C})|\geq\frac{1}{2s}n\sigma_1\right). \tag{S5.47}$$

$\bar{a}_j=\sum_{k=1}^K\left(\frac{1}{\sqrt{|C_k|}}\sum_{i\in C_k}x_{ij}\right)^2$ has the same distribution as $\|Y+b_j\|^2$ where Y obeys $\mathcal{N}(0,\mathbf{I}_{K\times K})$, $b_{jk}=\sqrt{n\tilde{\pi}_k\tilde{\mu}_{kj}^2}$ for $j=1,\ldots,p^*$ and $b_{jk}=0$ for $j>p^*$. By lemma 2, we know $\bar{a}_j$ are all sub exponential variables with parameter $(2\sqrt{K+2n\sigma_2},1)$. Note that $s\geq1$,

$$\frac{1}{2s}n\sigma_1\leq n\sigma_2\leq4(K+2n\sigma_2).$$

When $n\geq\frac{K}{\sigma_2}$, the last term could be bounded by

$$\exp(-\frac{n\sigma_1^2}{96s^2\sigma_2}). \tag{S5.48}$$

Now the proof is completed. □

Bibliography

Foss, S., Korshunov, D., and Zachary, S. (2011). *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer.