A Proof of Lemma 4.2

In this section, we include the proof of Lemma 4.2 and some preliminary facts that will be useful for the proof.

Let r be a Rademacher vector, i.e. every entry r_i is sampled independently uniformly from $\{-1, 1\}$. Further, we say that g is a Gaussian vector if every entry g_i is a standard Gaussian with mean 0 and variance 1. We have the following useful properties of Gaussians.

Fact A.1 (Appendix B.1 by [68]). Let $g_1, \ldots g_n$ be Gaussians with means μ_i and variances σ_i^2.

- If $\sigma_i^2 \leq \sigma^2$ for all i, then $\mathbb{E}[\max_{g_i} |g_i|] \leq 2\sigma\sqrt{2\log n}$.
- If the Gaussians are independent, then $\sum_{i=1}^n a_i g_i$ is Gaussian distributed with mean $\sum_{i=1}^n a_i \mu_i$ and variance $\sum_{i=1}^n a_i^2 \sigma_i^2$.
- If the g_i are independent standard Gaussians with mean 0 and variance 1, then $Y := \sum_{i=1}^n g_i^2$ is Chi-squared distributed with mean $\mathbb{E}[\sqrt{Y}] \in O(\sqrt{n})$.

Another result we need is the following

Lemma A.2 (Lemma 5.2 of [75]). $|\mathcal{N}(B_2^d, \|.\|_2, \varepsilon)| \leq (1 + 2/\varepsilon)^d$.

We are now ready to prove the Lemma 4.2. The proof of Lemma is similar to arguments used to prove Dudley's theorem. We also write here the statement of the Lemma for the sake of completeness

Lemma (Lemma 4.2). *Let $\mathcal{D}$ be a distribution over B_2^d and let P be a set of n points sampled from $\mathcal{D}$. Suppose that for a set of n-dimensional vectors V, we have absolute constants $C, \gamma > 0$ such that*

$$\log |\mathcal{N}(V, \|.\|_\infty, \varepsilon)| \in O(\varepsilon^{-2} \log^\gamma (n \cdot \varepsilon^{-1}) \cdot C). \tag{7}$$

Then

$$G_n(V) \in O\left(\sqrt{\frac{C \log^{\gamma+2} n}{n}}\right).$$

Proof. For ease of notation, we use solutions $\mathcal{S}$ induced by points, but the proof carries over without any modifications other than changing the notation to collections of subspaces $\mathcal{U}$.

Consider an arbitrary cost vector $v^{\mathcal{S}}$. We write $v^{\mathcal{S}}$ as a telescoping sum

$$v^{\mathcal{S}} := \sum_{h=0}^{\infty} v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}}$$

where $v^0 = 0$ and $v^{i,\mathcal{S}}$ is a vector from $\mathcal{N}(V, \|.\|_\infty, 2^{-i})$ approximating $v^{\mathcal{S}}$. Observe that

$$\|v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}}\|_\infty \leq \|v^{h+1,\mathcal{S}} - v^{\mathcal{S}} + v^{\mathcal{S}} - v^{h,\mathcal{S}}\|_\infty \leq 2 \cdot 2^{-h} \tag{8}$$

due to the triangle inequality. Thus we have

$$
\begin{aligned}
n \cdot G_n(V) &= \mathbb{E}_{P,g}\left[\sup_{\mathcal{S}} (v^{\mathcal{S}})^T g\right] = \mathbb{E}_{P,g}\left[\sup_{\mathcal{S}} \sum_{h=0}^{\infty} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g\right] \\
&\leq \mathbb{E}_{P,g} \sum_{h=0}^{\infty} \left[\sup_{\mathcal{S}} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g\right] \\
&= \mathbb{E}_{P,g} \sum_{h=0}^{\infty} \left[\sup_{\substack{v^{h+1,\mathcal{S}}, v^{h,\mathcal{S}} \in \\ \mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|.\|_\infty, 2^{-h})}} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g\right] \\
&= \mathbb{E}_{P,g} \sum_{h=0}^{\log n} \left[\sup_{\substack{v^{h+1,\mathcal{S}}, v^{h,\mathcal{S}} \in \\ \mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|.\|_\infty, 2^{-h})}} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g\right] \\
&\quad + \mathbb{E}_{P,g} \sum_{h=\log n}^{\infty} \left[\sup_{\substack{v^{h+1,\mathcal{S}}, v^{h,\mathcal{S}} \in \\ \mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|.\|_\infty, 2^{-h})}} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g\right]
\end{aligned}
$$

$$
= \mathbb{E}_{P,g} \sum_{h=0}^{\log n} \left[\sup_{\substack{v^{h+1,\mathcal{S}}, v^{h,\mathcal{S}} \in \\ \mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|.\|_\infty, 2^{-h})}} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g\right] \tag{9}
$$

$$
+ \mathbb{E}_{P,g}\left[\sup_{\mathcal{S}} (v^{\mathcal{S}} - v^{\log n,\mathcal{S}})^T g\right] \tag{10}
$$

We bound the terms 9 and 10 differently, starting with the latter.

For every $\mathcal{S}$

$$
(v^{\mathcal{S}} - v^{\log n,\mathcal{S}})^T g \leq \|v^{\mathcal{S}} - v^{\log n,\mathcal{S}}\|_2 \cdot \mathbb{E}[\|g\|_2],
$$

due to the Cauchy Schwarz inequality. Further,

$$
\|v^{\mathcal{S}} - v^{\log n,\mathcal{S}}\|_2 \leq \sqrt{n} \cdot \|v^{\mathcal{S}} - v^{\log n,\mathcal{S}}\|_\infty \leq \sqrt{n} \cdot 2^{-\log n} = \sqrt{\frac{1}{n}},
$$

which, combined with the third item in Fact A.1 yields

$$
\mathbb{E}_{P,g}\left[\sup_{\mathcal{S}} (v^{\mathcal{S}} - v^{\log n,\mathcal{S}})^T g\right] \in O\left(\sqrt{\frac{1}{n}} \cdot \sqrt{n}\right) = O(1). \tag{11}
$$

We now consider the term 9. Due to the second item of Fact A.1, $(v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g$ is Gaussian distributed with mean 0 and variance

$$
\sum_{i=1}^{n} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})_i^2 \leq 4n \cdot 2^{-2h}.
$$

Thus, we have, using the first item in Fact A.1

$$
\begin{aligned}
&\mathbb{E}_{P,g} \sum_{h=0}^{\log n} \left[\sup_{\substack{v^{h+1,\mathcal{S}}, v^{h,\mathcal{S}} \in \\ \mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|.\|_\infty, 2^{-h})}} (v^{h+1,\mathcal{S}} - v^{h,\mathcal{S}})^T g\right] \\
\leq\ &\sum_{h=0}^{\log n} \sqrt{32n \cdot 2^{-2h} \log |\mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|.\|_\infty, 2^{-h})|}
\end{aligned} \tag{12}
$$

Now using equation (7) we obtain that,

$$32n \cdot 2^{-2h} \log \left|\mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|\dot{|}\|_\infty, 2^{-h})\right| \in O(n \cdot \log^\gamma n)$$

So we have that

$$\sum_{h=0}^{\log n} \sqrt{32n \cdot 2^{-2h} \log \left|\mathcal{N}(V, \|.\|_\infty, 2^{-(h+1)}) \times \mathcal{N}(V, \|.\|_\infty, 2^{-h})\right|} \in O(\sqrt{n \cdot \log^{\gamma+2} n}) \tag{13}$$

Adding the bounds (13) and (11) for Terms (10) and (9), respectively yields the claim. □

Finally, we will frequently use the following triangle inequality extended to powers.

Lemma A.3 (Triangle Inequality for Powers (Lemma A.1 of [60])). *Let a, b, c be an arbitrary set of points in a metric space with distance function d and let z be a positive integer. Then for any $\varepsilon > 0$*

$$d(a,b)^z \leq (1+\varepsilon)^{z-1} d(a,c)^z + \left(\frac{1+\varepsilon}{\varepsilon}\right)^{z-1} d(b,c)^z$$
$$|d(a,b)^z - d(a,c)^z| \leq \varepsilon \cdot d(a,c)^z + \left(\frac{2z+\varepsilon}{\varepsilon}\right)^{z-1} d(b,c)^z.$$

B Omitted Proofs for Center-Based Clustering

Lemma B.1. *Let $P \subset B_2^d$ be a set of points. Let V be the set of all cost vectors of P for (k,z)-clustering. Then there exists an ε-clustering net of size*

$$|\mathcal{N}(V, \|.\|_\infty, \varepsilon)| \leq \exp(O(1) \cdot z \cdot k \cdot d \cdot \log(z\varepsilon^{-1})).$$

Proof. We start by proving the bound for $k = 1$. Suppose we are given a net $\mathcal{N}(B_2^d, \|.\|_2, \delta)$, for a δ to be determined later. Consider a candidate solution $\{s\}$ with cost vector $v^{\{s\}} \in V$. Let s' be the point in $\in \mathcal{N}(B_2^d, \|.\|_2, \delta)$ of such that $\|s - s'\| \leq \delta$, if s' is not unique any one will be sufficient. Let $v^{\mathcal{S}'}$ be the cost vector of $\mathcal{S}'$. The number of distinct solutions $\mathcal{S}'$ are $|\mathcal{N}(B_2^d, \|.\|_2, \delta)| = \exp(O(1) \cdot d \cdot \log \delta^{-1})$ due to Lemma A.2.

What is left to show is that all solutions constructed in this way satisfy the guarantee of $\mathcal{N}(V, \|.\|_\infty, \delta)$, for an appropriately chosen δ. We have for any $p \in P$ and any non-negative integer z due to Lemma A.3

$$\begin{aligned} |\|p-s\|^z - \|p-s'\|^z| &\leq \alpha \cdot \|p-s\|^z + \left(\frac{2z+\alpha}{\alpha}\right)^{z-1} \|s-s'\|^z \\ &\leq \alpha \cdot \|p-s\|^z + (3z)^z \left(\frac{\delta}{\alpha}\right)^{z-1} \cdot \delta \end{aligned}$$

We set $\alpha = \frac{1}{2 \cdot 2^z}\varepsilon$ and $\delta = \alpha \cdot \frac{1}{2(3z)^z}\varepsilon = \frac{1}{4(6z)^z}\varepsilon^2$. Then the term above is upper bounded by at most ε as $\|p - s\| \leq 2$. Now since $|\|p-s\|^z - \|p-s'\|^z| \leq \varepsilon$ for all $s \in B_2^d$ also implies $|\min_{s \in \mathcal{S}} \|p-s\|^z - \min_{s' \in \mathcal{S}'} \|p-s'\|^z| \leq \varepsilon$, we have proven our desired approximation.

To conclude, observe that by our choice of δ, the overall net N has size at most $\exp(O(1) \cdot z \cdot d \cdot \log(z\varepsilon^{-1}))$.

To extend this proof to k-centers, observe that any solution consisting of k centers can be obtained by selecting k points from B_2^d, rather than one. This raises the net size of the single cluster case by a power of k. □

We now show that Lemma B.1 combined with terminal embeddings yields the desired net.

Lemma (Equation 4 in Lemma 5.2). *Let $\mathcal{D}$ be a distribution over B_2^d and let P a set of n points sampled from $\mathcal{D}$ and let V be defined as in Theorem 2. Then*

$$|\mathcal{N}(V, \|.\|_\infty, \varepsilon)| \leq \exp(O(1) z^3 \cdot k \cdot \varepsilon^{-2} \log n \cdot (\log(z) + \log(\varepsilon^{-1}))).$$

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a terminal embedding, that is f is such that $m \in O(z^2 \cdot \varepsilon^{-2} \log |P|)$[4] and for all $p \in P$ and $q \in \mathbb{R}^d$

$$\|p - q\|^z = (1 \pm \varepsilon)\|f(p) - f(q)\|^z,$$

as given by [62]. Therefore, for any candidate solution $\mathcal{S}$, we also have

$$\text{cost}(p, \mathcal{S}) = (1 \pm 2\varepsilon)\text{cost}(f(p), f(\mathcal{S})).$$

In other words, the set of cost vectors in the image of f is the desired $O(\varepsilon)$-net for the true set of cost vectors. Hence an ε-net for the cost vectors induced by solutions in the image of f is also an $O(\varepsilon)$-net for the set of cost vectors. We thus may apply Lemma B.1 for all cost vectors induced by solutions in the image of f. After rescaling ε by constant factors, the overall net size is therefore $\exp(O(1)z^3 \cdot k \cdot \varepsilon^{-2} \log n \cdot (\log(z) + \log(\varepsilon^{-1})))$ □

C Omitted Proofs for Subspace Clustering

In this section, we provide full proofs for Section 5 relative to subspace clustering.

We start with a few basic lemmas that will be useful in the calculations later.

We further require the following bounds that will prove useful in the calculations later.

Lemma C.1. *Let a, b be numbers in $[0, 2]$ and let $\varepsilon > 0$. Suppose $a^2 = b^2 \pm \varepsilon \cdot b$. Then*

$$|a - b| \leq \varepsilon.$$

Moreover, for any non-negative integer z, we have

$$|a^z - b^z| \leq 2 \cdot (3z)^z \cdot \varepsilon.$$

Proof. For the first part of the lemma, we observe

$$|a^2 - b^2| = |a - b| \cdot (a + b) \leq \varepsilon \cdot b$$

which implies

$$|a - b| \leq \varepsilon.$$

For the second part, Lemma A.3 implies

$$|a^z - b^z| \leq \varepsilon \cdot \max(a, b)^z + \left(\frac{2z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot |a - b|^z \leq \varepsilon \cdot 2^z + \left(\frac{3z + \varepsilon}{\varepsilon}\right)^{z-1} \cdot \varepsilon^z \leq 2(3z)^z \varepsilon. \quad \square$$

This lemma now immediately implies the following corollary by rescaling ε.

Corollary C.2. *Let a, b be numbers in $[0, 2]$ and let $\varepsilon > 0$. Suppose $a^2 = b^2 \pm \frac{1}{4 \cdot (3z)^z} \max(\varepsilon \cdot b, \varepsilon^2)$. Then for any non-negative integer z, we have*

$$|a^z - b^z| \leq \varepsilon.$$

We now show that for any candidate subspace U we can find a subspace representing it that is spanned by only a few vectors in P.

Lemma (Lemma 5.5). *Let $P \subseteq B_2^d$. For any orthogonal matrix $U \in \mathbb{R}^{j \times d}$, there exists $M \subseteq P$, with $|M| = O(j \cdot \varepsilon^{-2})$, such that*

$$\forall p \in P, \|U^T(I - \Pi_M)p\| \leq \varepsilon \cdot \|(I - \Pi_M)p\|. \tag{14}$$

Proof. Initially, let $M = \emptyset$. We add points to M in rounds and denote by M_t the set after t rounds. Furthermore, let Π_t be the projection matrix onto the subspace spanned by M_t at round t. If there is a $p \in P$ in round t such that

$$\|U^T(I - \Pi_t)p\| > \varepsilon\|(I - \Pi_t)p\| \tag{15}$$

[4] The dependency on z is easily derived via a straightforward application of Lemma A.3.

then we let $M_{t+1} = M_t \cup \{p\}$. Our goal is to show that after $T \in O(j\varepsilon^{-2})$ many rounds, we have $\|U^T(I - \Pi_T)p\| \leq \varepsilon \cdot \|(I - \Pi_T)p\|$. We show this by proving inductively

$$\|U^T \Pi_t\|_F^2 \geq \varepsilon^2 \cdot t.$$

For the base case $t = 0$, this is trivially true. Thus suppose we add a point p in iteration $t + 1$. Reformulating Equation 15, we have $\frac{\|U^T(I-\Pi_t)p\|}{\|(I-\Pi_t)p\|} > \varepsilon$. By the Pythagorean theorem, we therefore have

$$\|U^T \Pi_{t+1}\|_F^2 = \|U^T \Pi_t\|_F^2 + \frac{\|U^T(I - \Pi_t)p\|^2}{\|(I - \Pi_t)p\|^2} \geq \varepsilon^2 \cdot t + \varepsilon^2 \geq \varepsilon^2 \cdot (t + 1).$$

Now since Π_t is a projection and since U has j orthonormal columns $j \geq ||U^T||_F^2 \geq ||U^T\Pi_t||_F^2$. If $T \geq \varepsilon^{-2}j$, we obtain $||U^T\Pi_T||_F^2 \geq j$. This implies that U is contained in the space spanned by M_T. Conversely, U must also be orthogonal to the kernel of M_T that is $U(I - \Pi_T) = 0$. Therefore after at most $\varepsilon^{-2}j$ rounds, we have $\|U^T(I - \Pi_T)p\| \leq \varepsilon \cdot \|(I - \Pi_T)p\|$. □

Lemma (Lemma 5.4). *Let $P \subset B_2^d$ be a set of points and let z be a positive integer. Then there exists an (ε, j)-projective net of size*

$$|\mathcal{N}(V, \|.\|_\infty, \varepsilon)| \leq \exp(O(1) \cdot d \cdot j \cdot \log(j\varepsilon^{-1})).$$

Proof. Let N be an ε/j-net of the Euclidean unit ball, i.e. $N = \mathcal{N}(B_2^d, \|.\|_2, \varepsilon/j)$ due to Lemma A.2. Let $\mathcal{N} = \otimes_{i=1}^j N$ be the set of $j-$subsets of of N. We claim that for every S, there exists an $S' \in \mathcal{N}$ such that

$$\|S^T p\|_2 = \|S'^T p\|_2 \pm \varepsilon.$$

Note that this implies the claim as $|\mathcal{N}| \in \left(\left(1 + \frac{2j}{\varepsilon}\right)^d\right)^j = \exp(O(1) \cdot d \cdot j \cdot \log(j\varepsilon^{-1}))$.

Define $S_i'^T$ to be the vector in N closest to the ith row of S^T, i.e. $\|S_i^T - S_i'^T\|_2 \leq \varepsilon/j$. We have $\|S'^T - S\|_2 \leq \sum_{i=1}^j \|S_i'^T - S_i^T\|_2 \leq \varepsilon$. Therefore

$$\begin{aligned}
\|S^T p\|_2 &= \|(S^T - S'^T)p + S'^T p\|_2 \\
&\leq \|(S^T - S'^T)p\|_2 + \|S'^T p\|_2 \\
&\leq \|S'^T p\|_2 + \|S^T - S'^T\|_2 \|p\|_2 \\
&\leq \|S'^T p\|_2 + \varepsilon.
\end{aligned}$$

The bound $\|S^T p\|_2 \geq \|S'^T p\|_2 - \varepsilon$ is proven analogously. □

We can now conclude with the proof of Equation 5 in Lemma 5.2.

Lemma (Equation 5 in Lemma 5.2). *Let $\mathcal{D}$ be a distribution over B_2^d and let P a set of n points sampled from $\mathcal{D}$ and let $V_{j,z}$ be defined as in Theorem 5.1. Then*

$$|\mathcal{N}(V_{j,z}, \|.\|_\infty, \varepsilon)| \leq \exp(O(1)(3z)^{z+2} \cdot k \cdot \varepsilon^{-2}(\log n + j \log(j\varepsilon^{-1})) \log \varepsilon^{-1}).$$

Proof. Let $\alpha, \beta > 0$ be sufficiently small parameters depending on ε that will determined later. We first describe a construction for nets for a single subspace of rank at most j, before composing to k subspaces.

We start by describing the composition of the nets. For every subset $M \subseteq P$, with $|M| \in O(j\alpha^{-2})$, we let Π_M denote an orthogonal projection matrix of the span of M. Note that this implies $\textbf{rank}(\Pi_M) = O(j\alpha^{-2})$. Further, let $N(\Pi_M) := \mathcal{N}(B_2^{\textbf{rank}(\Pi_M)}, \|.\|_2, \beta)$ be a (β, j)-projective net of the point set $\cup_{p \in M}\{\Pi_M p\}$ of size at most $\exp(O(1) \cdot \textbf{rank}(\Pi_M) \cdot \log(j\beta^{-1}))$ given by Lemma 5.4. Finally, let $N := \cup_M N(\Pi_M)$.

We consider an arbitrary orthogonal matrix $U \in \mathbb{R}^{j \times d}$. Denote by M_U the subset of points and by Π_U the projection matrix given by Lemma 5.5, using α as the precision variable. We claim that for every U, there exists an $U' \in N$ such that for all $p \in P$

$$\left| \left(\|\Pi_U p\|_2^2 - \|U'^T \Pi_U p\|_2^2 + \|(I - \Pi_U)p\|_2^2 \right)^{z/2} - \|(I - UU^T)p\|^z \right| \in O(\alpha + \beta).$$

In other words, by enumerating over all (β, j)-projective nets, we obtain an $O(\alpha+\beta)$-subspace clustering net for $(1, j, z)$-clustering. The desired error of ε then follows by choosing α and β accordingly. For U, we construct U' as follows. Let $D = \sqrt{\Pi_U}$, i.e. $DD^T = \Pi_U$. Further, let $V = U^T D$, notice that V has at most j rows that have at most unit norm. Hence, there exists a $U' \in N$ such that

$$|\|U\Pi_U p\|_2 - \|U'\Pi_U p\|_2| \leq \varepsilon$$

that is a (β, j)-projective net.

We then obtain

$$\begin{aligned}
& \|\Pi_U p\|_2^2 - \|U'^T \Pi_U p\|_2^2 + \|(I-\Pi_U)p\|_2^2 \\
= \; & \|\Pi_U p\|_2^2 - \|U^T \Pi_U p\|_2^2 \pm \beta + \|(I-\Pi_U)p\|_2^2 \\
= \; & \|\Pi_U p\|_2^2 - \|UU^T \Pi_U p\|_2^2 \pm \beta + \|(I-\Pi_U)p\|_2^2 \\
= \; & \|(I-UU^T)\Pi_U p\|_2^2 + \|(I-\Pi_U)p\|_2^2 \pm \beta \\
(Eq. 6) \quad = \; & \|(I-UU^T)p\|_2^2 \pm \beta - \|U^T(I-\Pi_U)p\|^2 - 2p^T \Pi_U UU^T (I-\Pi_U)^T p \\
(Lem. 5.5) \quad = \; & \|(I-UU^T)p\|_2^2 \pm \alpha^2 \cdot \|(I-UU^T)p\|^2 \pm 2\alpha \cdot \|(I-UU^T)p\| \pm \beta
\end{aligned}$$

Setting $\alpha^2 = \beta = \frac{1}{64(3z)^z}\varepsilon^2$, we then have due to Corollary C.2

$$\left| \left| \|\Pi_U p\|_2^2 - \|U'^T \Pi_U p\|_2^2 + \|(I-\Pi_U)p\|_2^2 \right|^z - \|(I-UU^T)p\|^z \right| \leq \varepsilon. \tag{16}$$

To extend this from a single j-dimensional subspace to a solution $\mathcal{U}$ given by the intersection of k j-dimensional subspaces, we define cost vectors $v^{\mathcal{S}'}$ obtained from $\mathcal{N} = \otimes_{i=1}^k N$ as follows. For each $U \in \mathcal{U}$ let U' be constructed as above and let $\mathcal{U}'$ be the union of the thus constructed U'. Then, with a slight abuse of notation, letting $\Pi_{U'}$ correspond to the subspace used to obtain U', we define

$$v_p^{\mathcal{U}'} := \min_{U' \in \mathcal{U}'} \left| \|(I-\Pi_{U'})p\|^2 + \|\Pi_{U'}p\|^2 - \|U'\Pi_{U'}p\|^2 \| \right|^{z/2}.$$

Let U be the subspace to which p is assigned $\mathcal{U}$ and let U' be the center in $\mathcal{U}'$ used to approximate U and let $U^{*\prime} = \text{argmin}_{U' \in \mathcal{U}'} \left| \|(I-\Pi'_U)p\|^2 + \|\Pi_{U'}p\|^2 - \|U'\Pi'_U p\|^2 \| \right|^{z/2}$ and let $U^* \in \mathcal{U}$ be the center approximated by $U^{*\prime}$. Then applying Equation 16, we have

$$\begin{aligned}
& \|(I-UU^T)p\|^z \\
\leq \; & \|(I-U^*U^{*T}p\|^z \\
\leq \; & \left| \|(I-\Pi_{U^{*\prime}})p\|^2 + \|\Pi_{U^{*\prime}}p\|^2 - \|U*'\Pi_{U^{*\prime}}p\|^2 \right|^{z/2} + \varepsilon \\
\leq \; & \left| \|(I-\Pi_{U'})p\|^2 + \|\Pi_{U'}p\|^2 - \|U'\Pi_{U'}p\|^2 \right|^{z/2} + \varepsilon
\end{aligned}$$

Thus, the cost vectors obtained from $\mathcal{N}$ are a (k, j, z)-clustering net, i.e.

$$\left| v_p^{\mathcal{S}'} - v_p^{\mathcal{S}} \right| := \left| \min_{s' \in \mathcal{S}'} \|\Pi_{s'}p - [s', 0]\|^z - \min_{s \in \mathcal{S}} \|p - s\|^z \right| \leq \varepsilon.$$

What remains is to bound the size of the clustering net. Here we first observe that size of the clustering net is equal to $|\mathcal{N}| = |N|^k$. For $|N|$, we have $\binom{|P|}{O(\alpha^{-2}\log\alpha^{-1})} \leq n^{O(j\alpha^{-2}\log\alpha^{-1})}$ many choices of $N(\Pi)$. In turn, the size of each $N(\Pi)$ is bounded by $(\beta/j)^{-O(j^2\alpha^{-2})}$ due to Lemma 5.4. Thus the overall size of $\mathcal{N}$ is

$$\begin{aligned}
& \exp\left(k \cdot j \cdot O(\alpha^{-2}\log\alpha^{-1}(\log n + j\log\beta/j))\right) \\
& = \exp(O(1)(3z)^{z+2} \cdot k \cdot j \cdot \varepsilon^{-2}(\log n + j\log(j\varepsilon^{-1}))\log\varepsilon^{-1})
\end{aligned}$$

as desired. □

C.1 Proofs of Theorem 5.6 (Section 5.1)

The proof of the theorem is a straightforward application of Theorem 4.1 with the following Lemma

Lemma C.3. *Let $\mathcal{D}$ be a distribution over B_2^d, let P a set of n points sampled from $\mathcal{D}$, and let V be defined as in Theorem 5.6. Then for any $\gamma > 0$*

$$Rad_n(V_{j,2}) \in O\left(\sqrt{\frac{kj}{n}\log^{3+\gamma}\left(\frac{n}{j}\right)}\right).$$

Proof. We use the following result due to Foster and Rakhlin [45].

Theorem C.4 (ℓ_∞ contraction inequality (Theorem 1 by [45])). *Let $F \subseteq X \to \mathbb{R}^k$, and let $\phi : \mathbb{R}^k \to \mathbb{R}$ be L-Lipschitz with respect to the ℓ_∞ norm, i.e. $\|\phi(X)-\phi(X')\|_\infty \leq L\cdot\|X-X'\|_\infty$ for all $X, X' \in \mathbb{R}^k$. For any $\gamma > 0$, there exists a constant $C > 0$ such that if $|\phi_t(f(x))| \vee \|f(x)\|_\infty \leq \beta$, then*

$$Rad_n(\phi\circ F) \leq C\cdot L\sqrt{K}\cdot\max_i Rad_n(F|_i)\cdot\log^{3/2+\gamma}\left(\frac{\beta n}{\max_i R_n(F|_i)}\right).$$

We use this theorem as follows. Our functions are associated with candidate solutions $\mathcal{U}$, that is $\phi(f) = \min_{U\in\mathcal{U}}\|(I-UU^T)p\|_2^2$. In other words, f maps a point p to the k-dimensional vector, where $f_i(p) = \|(I-U_iU_i^T)p\|_2^2$ and ϕ selects the minimum value among all $\|I-U_iU_i^T)p\|_2^2$.

Thus, we require three more steps. First, we have to bound the Lipschitz constant of the minimum operator. Second, we have to give a bound on β. Third and last, we have to give a bound on the Rademacher complexity

$$Rad_n(V) = \frac{1}{n}\cdot\mathbb{E}_r\sup_U\sum_{p\in P}\|(I-UU^T)p\|_2^2 r_p. \tag{17}$$

The Lipschitz constant of the minimum operator with respect to the ℓ_∞ norm can be readily shown to be 1 as for any two vectors x, y with $\min_i y_i = y_j$

$$\min_i x_i - \min_i y_i = \min_i x_i - y_j \leq x_j - y_j \leq |x_j - y_j| \leq \|x-y\|_\infty.$$

Since U is an orthogonal matrices and $p \in B_2^d$, we have $\|(I-UU^T)p\|_2^2 \leq 1$ and thus β is bounded by 1.

Thus, we only require a bound on Equation 17. For this, we use a result by [50]. Since the result is embedded in the proof of another result, we restate it here for the convenience of the reader.

Lemma C.5 (Compare the proof Theorem 3 of [50]). *Let P be an set of n points in B_2^d and let $\mathcal{U}$ be the set of all orthogonal matrices of rank at most j. For every $U \in \mathcal{U}$, define $f_U(p) = \|(I-UU^T)p\|_2^2$ and let F be the set of all functions $f_U(p)$ Then.*

$$Rad_n(F) := \frac{1}{n}\cdot\mathbb{E}_r\sup_{U\in\mathcal{U}}\sum_{p\in P}\|(I-UU^T)p\|_2^2\cdot r_p \in O\left(\sqrt{\frac{j}{n}}\right).$$

Proof. We have

$$Rad_n(F) = \mathbb{E}_r\sup_U\sum_{p\in P}\|(I-UU^T)p\|_2^2 r_p = \mathbb{E}_r\sum_{p\in P}\|p\|^2 r_p + \mathbb{E}_r\sup_U\sum_{p\in P}\|U^Tp\|_2^2 r_p.$$

We observe that the term $\mathbb{E}_r \sum_{p\in P} \|p\|^2 r_p$ is 0. Thus, we focus on the second term. We have

$$\begin{aligned}
\mathbb{E}_r \sup_U \sum_{p\in P} \|U^T p\|_2^2 \cdot r_p &= \mathbb{E}_r \sup_U \sum_{p\in P} p^T U U^T p \cdot r_p = \mathbb{E}_r \sup_U \sum_{p\in P} trace(p^T U U^T p) \cdot r_p \\
&= \mathbb{E}_r \sup_U \sum_{p\in P} trace(U U^T p p^T) \cdot r_p \\
&= \mathbb{E}_r \sup_U trace\left(U U^T \sum_{p\in P} (r_p \cdot p p^T) \right) \\
&\leq \mathbb{E}_r \sup_U \|U\|_F \left\| \sum_{p\in P} r_p \cdot p p^T \right\|_F .
\end{aligned}$$

We have $\|U\|_F \leq \sqrt{j}$, so we focus on $\left\|\sum_{p\in P} r_p \cdot pp^T\right\|_F$. Here, we have

$$\begin{aligned}
\left\| \sum_{p\in P} r_p \cdot p p^T \right\|_F^2 &= trace\left(\left(\sum_{p\in P} r_p \cdot p p^T \right) \left(\sum_{p\in P} r_p \cdot p p^T \right) \right) \\
&= \sum_{p\in P} \sum_{q\in P} r_p \cdot r_q \cdot trace\left(p p^T q q^T \right) = \sum_{p\in P} \sum_{q\in P} r_p \cdot r_q \cdot (p^T q)^2 .
\end{aligned}$$

This implies

$$\begin{aligned}
n \cdot Rad_n(F) &= \mathbb{E}_r \sup_U \sum_{p\in P} \|U^T p\|_2^2 r_p \leq \mathbb{E}_r \sup_U \|U\|_F \left\| \sum_{p\in P} r_p \cdot p p^T \right\|_F \\
&\leq \sqrt{j} \cdot \mathbb{E}_r \sqrt{\sum_{p\in P} \sum_{q\in P} r_p \cdot r_q \cdot (p^T q)^2} \\
\text{(Jensen's inequality)} &\leq \sqrt{j} \cdot \sqrt{\mathbb{E}_r \sum_{p\in P} \sum_{q\in P} r_p \cdot r_q \cdot (p^T q)^2} \\
&= \sqrt{j} \cdot \sqrt{\sum_{p\in P} (p^T p)^2} \leq \sqrt{j} \cdot \sqrt{\sum_{p\in P} 1} = \sqrt{nj}.
\end{aligned}$$

Solving the above for $Rad_n(F)$ concludes the proof. □

We can now conclude the proof. Combining the bounds on L and β with Lemma C.5 and Theorem C.4, we have

$$Rad_n(V_{j,2}) \in O\left(\sqrt{k} \cdot \sqrt{\frac{j}{n}} \cdot \log^{3+\gamma}(n) \right)$$

as desired. □

C.2 Lower Bound

Finally, we also show that the bound given in Theorem 5.6 is optimal, up to polylog factors.

Theorem (5.7). *There exists a distribution $\mathcal{D}$ supported on B_2^d such that $\mathcal{E}(V_{j,2}) \in \Omega\left(\sqrt{\frac{kj}{n}}\right)$.*

Proof. We first describe the hard instance distribution $\mathcal{D}$. We assume that we are given $d = 2kj$ dimensions. Let e_i be the standard unit vector along dimension i with $i \in \{1, \ldots d\}$. Let $p, \varepsilon \in [0, 1]$ be a parameters, where ε is sufficiently small. We set the densities for a point q as follows.

$$\mathbb{P}[q] = \begin{cases} p & \text{if } q = e_i, i \in \{1, \ldots, k \cdot j\} \\ p - \varepsilon \cdot p & \text{if } q = e_i, i \in \{kj+1, \ldots, d\} \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

We choose p such that integral over densities is 1, i.e. $kj \cdot p + kj \cdot (p - \varepsilon p) = 1$. It is straightforward to verify that for ε sufficiently small, $p \in (\frac{1}{kj}, \frac{2}{kj})$. We denote the points $\{e_1, \dots e_{kj}\}$ by G for "good" and the points $\{e_{kj+1}, \dots e_d\}$ by B for "bad".

We now characterize the properties of the optimal solution as well as suboptimal solutions.

Lemma C.6. *Let $\mathcal{D}$ be the distribution described above in Equation (18). Then for any optimal solution $\mathcal{U} = \{U_1, \dots U_k\}$, we have $e_i \in U_t$ for $i \in \{1, \dots, kj\}$ and some t and OPT $= kj \cdot p \cdot (1 - \varepsilon)$.*

Proof. We transform the instance into a $d \times d$ diagonal matrix D where $D_{i,i} = \sqrt{\mathbb{P}[e_i]}$. So D is a $d \times d$ diagonal matrix with diagonal entries equal to $\sqrt{p}$ for the first $k \cdot j$ elements and $\sqrt{p - \varepsilon \cdot p}$ for elements from $k \cdot j + 1$ to d. Now consider any partition of the points into clusters C_t with the corresponding subspace U_t for ($t \in \{1, \dots, k\}$). The optimal solution for U_t is simply the right singular vector of the submatrix of D corresponding to points in C_t, which by the construction of D is the j points with the largest weight. This means that each cluster can remove at most $\sum_{i=1}^{j} 1 = j$ from the cost, so k clusters can remove at most $\sum_{i=1}^{k} j$ from the cost. This imples that the cost of the clustering is lower bounded by $\sum_{i=1}^{d} D_{i,i}^2 - \sum_{i=1}^{kj} D_{i,i}^2 = \sum_{i=kj+1}^{d} D_{i,i}^2$. Conversely, the solution $\mathcal{U}$ has exactly this cost, which implies that it must be optimal. □

Using Lemma C.6, we now have to, given n independent samples from $\mathcal{D}$. Control the probability that the sample P will (falsely) put a higher weight on some of the points in B than the points in G. Let B_{ex} denote the set of misclassified points in B and let P_{OPT} denote the optimum computed on the sample P. We have

$$\mathbb{E}[\text{cost}(\mathcal{D}, P_{\text{OPT}})] = kj \cdot p \cdot (1 - \varepsilon) + p \cdot \varepsilon \cdot |B_{ex}|.$$

and hence an expected excess risk bound of

$$\mathbb{E}[\text{cost}(\mathcal{D}, P_{\text{OPT}})] - \text{OPT} = p \cdot \varepsilon \cdot \mathbb{E}[B_{ex}].$$

By linearity of expectation, we have $\mathbb{E}[|B_{ex}|] = kj \cdot \mathbb{P}[e_{kj+1} \in B_{ex}]$. Thus, $\mathbb{E}[\text{cost}(\mathcal{D}, P_{\text{OPT}})] - \text{OPT} \in \Theta(1)\varepsilon \cdot \mathbb{P}[e_{kj+1} \in B_{ex}]$. Define G_{low} to be the set of points from G that are have an empirical density of at most p. Let $\widehat{e_{kj+1}}$ denote the empirical density of e_{kj+1}. We now claim that

$$\begin{aligned}
\mathbb{P}[e_{kj} \in B_{ex}] &\geq \mathbb{P}[\widehat{e_{kj+1}} > p \wedge e_{kj+1} \in B_{ex}] \\
&= \mathbb{P}[e_{kj+1} \in B_{ex} | \widehat{e_{kj+1}} > p] \cdot \mathbb{P}[\widehat{e_{kj+1}} > p] \geq 1/2 \cdot \mathbb{P}[\widehat{e_{kj+1}} > p]
\end{aligned}$$

The first inequality follows because we are considering a subset of the possible events, the second inequality follows because the number of points with an empirical estimated density greater than p is negatively correlated with the empirical density $\widehat{e_{kj+1}}$ of the point e_{kj}. Specifically, conditioned on $\widehat{e_{kj+1}} > p$, the mean and median density of any point $e_i \in G$ is at most $\frac{1}{n} \cdot p(n - p \cdot n) = p \cdot (1 - p) < p$. Thus, the (marginal) mean and median density of any other point is below p and therefore the probability that e_{kj+1} will be in B_{ex} is at least $1/2$.

Thus, what remains to be shown is a bound on $\mathbb{P}[e_{kj} > p]$. Here, we use the tightness of the Chernoff bound (see Lemma 4 of [48]).

Lemma C.7 (Tightness of the Chernoff Bound)**.** *Let X be the average of n independent, $0/1$ random variables. For any $\varepsilon \in (0, 1/2]$ and $\mu \in (0, 1/2]$, assuming $\varepsilon^2 \mu n \geq 3$ if each random variable is 1 with probability at least μ, then*

$$\mathbb{P}[X > (1 + \varepsilon)p] > \exp(-9\varepsilon^2 \mu n).$$

Thus, sampling n elements, we have

$$\begin{aligned}
\mathbb{P}\left[e_{kj} > p\right] &= \mathbb{P}\left[e_{kj} > \left(1 + \frac{\varepsilon}{1 - \varepsilon}\right) \cdot (1 - \varepsilon) \cdot p\right] \\
&> \exp\left(-9 \frac{\varepsilon^2}{(1 - \varepsilon)^2}(1 - \varepsilon)pn\right) \in \Omega(1) \exp\left(-\frac{\varepsilon^2}{kj} n\right).
\end{aligned}$$

If we require $\mathbb{E}[\text{cost}(\mathcal{D}, P_{\text{OPT}})] - \text{OPT} = \varepsilon \cdot c$ for a sufficiently small absolute constant c, we also require $\mathbb{P}\left[e_{kj} > p\right] = c'$ and hence $\sqrt{\frac{kj}{n}} \leq \varepsilon \cdot c''$ for a sufficiently small absolute constants c' and c''. Letting $\varepsilon \to 0$ then shows that the excess risk can asymptotically decrease no faster than $\Omega\left(\sqrt{\frac{kj}{n}}\right)$. □

D Details for the Experiments (Section 6)

D.1 Description of datasets

Mushroom comprises of 112 categorical features of the appearance of mushrooms with class labels corresponding to poisonous or edible. MNIST contains 28x28 pixel images of handwritten digits. Skin_Nonskin are RGB values given as 3 numerical features used to predict if a pixel is skin or not. Lastly, Covtype consists of a mix of categorical and numerical features used to predict seven different cover types of forests. In the main body, we focus on Covtype because of its large number of points.

D.2 Description of algorithms

Center based clustering For each experiment, we use an expectation maximization (EM) type algorithm. Given a solution $\mathcal{S}$, we first assign every point to its closest center and subsequently, we recompute the center. For the case $z = 2$, we do this analytically and in this case the EM algorithm is more commonly known as Llyod's method [59]. For the cases, $z \in \{1, 3, 4\}$, the new center is obtained via gradient descent. The initial centers are chosen via D^z sampling, i.e. sampling centers proportionate to the zth power of the distance between a point and its closest center (for $z = 2$ this is the k-means++ algorithm by [6]).

We wrote all of the code using Python 3 and utilized the Pytorch library for implementations using gradient descent. Specifically, we employed the AdamW optimizer to find the closest center with a learning rate set to 0.01. All experiments were conducted on a machine equipped with a single NVIDIA RTX 2080 GPU.

Subspace Clustering For subspace clustering, we consider $j \in \{1, 2, 5\}$ to demonstrate the effects of the subspace dimension on convergence rate, taking computational expenses into consideration. Since there are no known tractable algorithms for these problems with guarantees, we initialize a solution $\mathcal{U} = \{U_1, \ldots, U_k\}$ by sampling k orthogonal matrices of rank j, where the subspace for each matrix is determined via the volume sampling technique [35]. Subsequently, we run the EM algorithm. As before, the expectation step consists of finding the closest subspace for every point. For $z = 2$, the maximization step consists of finding the j principal component vectors of the data matrix induced by each cluster. For the other values of z, it is NP-hard even approximate the maximization step [24], so we use gradient descent to find a local optimum. Due to the fact that Skin_nonskin only has 3 features, we only evaluate the excess risk for $j \in \{1, 2\}$. Due to a large computational dependency on dimension, we do not evaluate subspaces on the MNIST dataset.

D.3 Experimental results

In this section, we provide plots of the excess risk and the found parameters of the best-fit lines for each of the datasets.

Table 2: Best fit lines on Covtype and Mushroom (left to right)

z	c	q_1	q_2
1	$3 \cdot 10^{-2}$	0.44	0.54
2	$4 \cdot 10^{-3}$	0.42	0.52
3	$6 \cdot 10^{-4}$	0.44	0.51
4	$1 \cdot 10^{-4}$	0.44	0.51

z	c	q_1	q_2
1	$1 \cdot 10^{-1}$	0.48	0.51
2	$8 \cdot 10^{-2}$	0.48	0.51
3	$4 \cdot 10^{-2}$	0.49	0.50
4	$3 \cdot 10^{-2}$	0.49	0.50

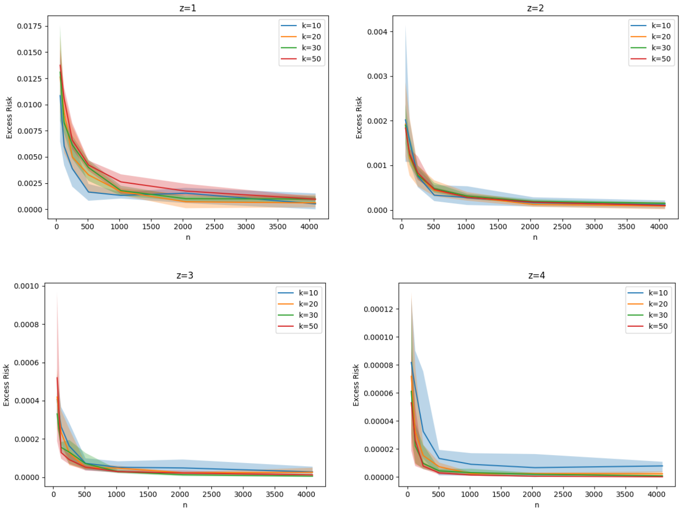

Figure 2: Excess risk for center-based clustering on the Covertype dataset. The shaded areas indicate the maximal and minimal deviation for the respective sample sizes.

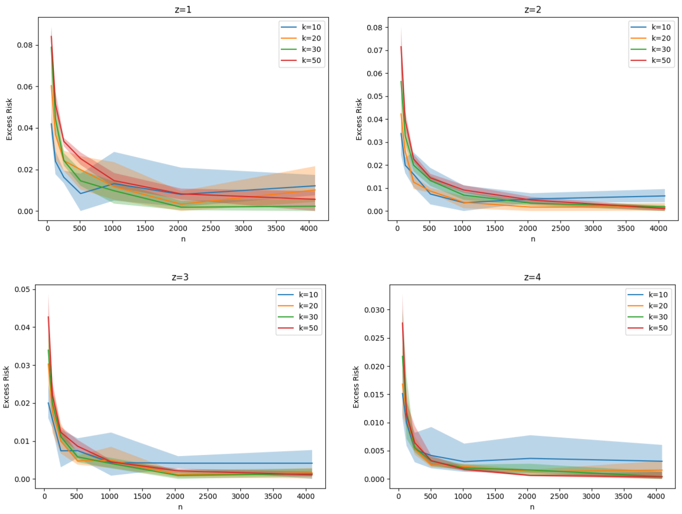

Figure 3: Excess risk for center-based clustering on the Mushroom dataset. The shaded areas indicate the maximal and minimal deviation for the respective sample sizes.

Table 3: Best fit lines on Skin_NonSkin and MNIST (left to right)

z	c	q_1	q_2
1	$2 \cdot 10^{-2}$	0.49	0.50
2	$3 \cdot 10^{-3}$	0.47	0.52
3	$8 \cdot 10^{-4}$	0.46	0.53
4	$2 \cdot 10^{-4}$	0.46	0.53

z	c	q_1	q_2
1	$1 \cdot 10^{-1}$	0.49	0.51
3	$5 \cdot 10^{-2}$	0.50	0.50
4	$3 \cdot 10^{-2}$	0.50	0.50
2	$8 \cdot 10^{-02}$	0.50	0.50

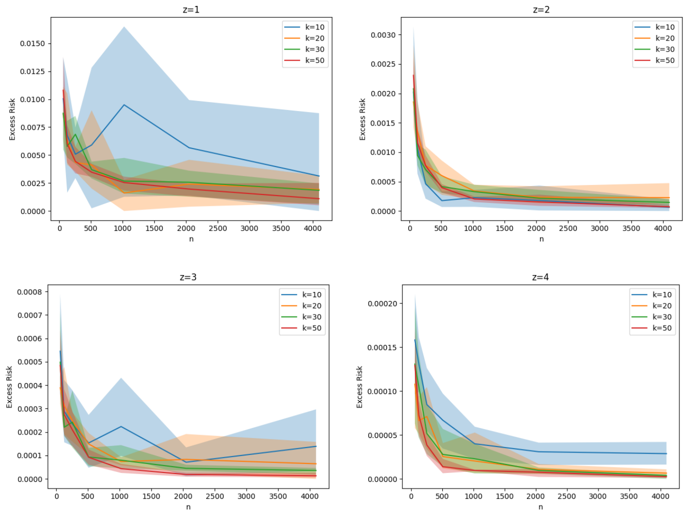

Figure 4: Excess risk for center-based clustering on the Skin_Nonskin dataset. The shaded areas indicate the maximal and minimal deviation for the respective sample sizes.

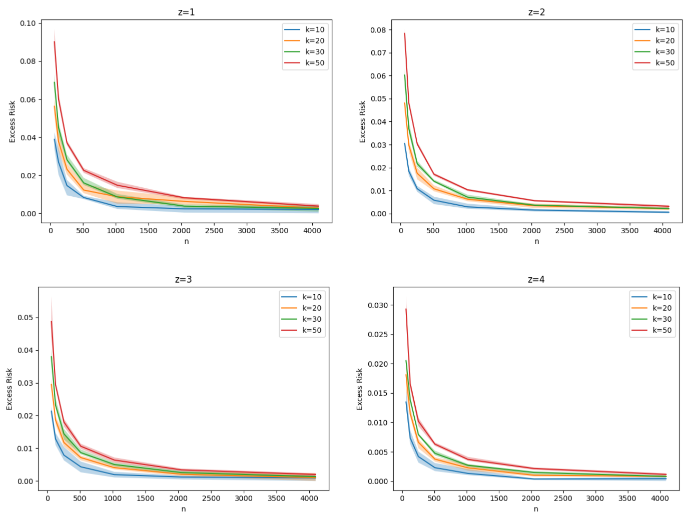

Figure 5: Excess risk for center-based clustering on the MNIST dataset. The shaded areas indicate the maximal and minimal deviation for the respective sample sizes.

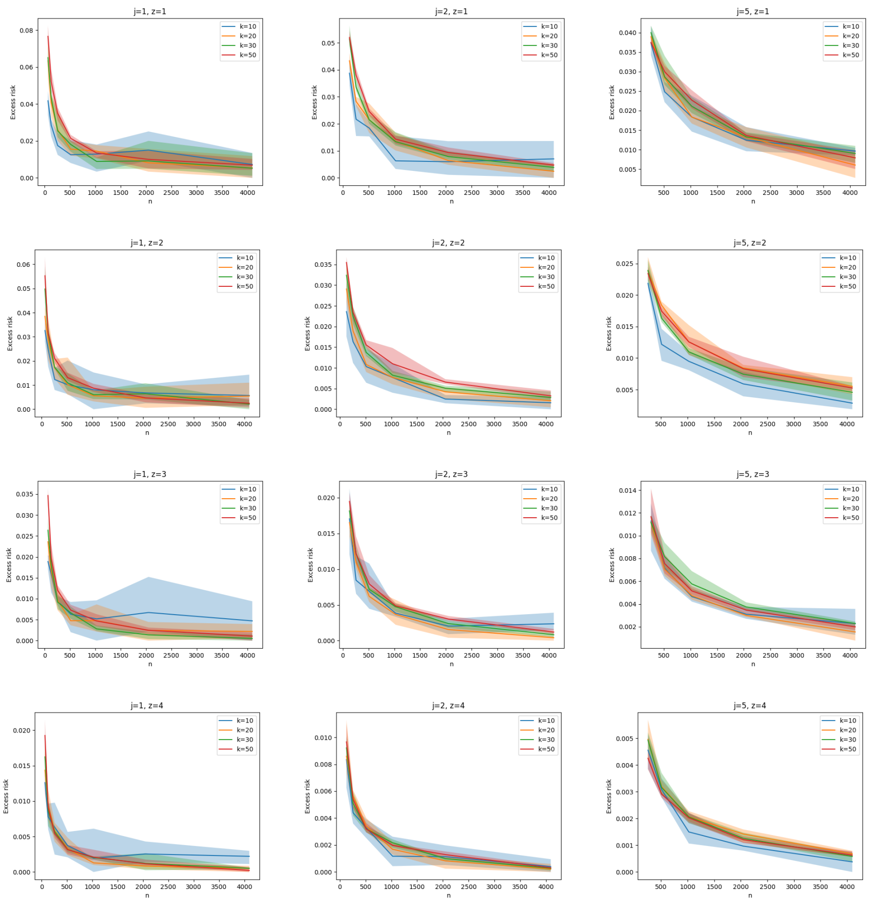

Figure 6: Excess risk for subspace clustering on the Mushroom dataset. The shaded areas indicate min/max values

Table 4: Best fit line for subspace clustering on Covtype and Mushroom (left to right)

j	z	c	q_1	q_2
1	1	0.1	0.45	0.54
1	2	$2 \cdot 10^{-2}$	0.48	0.51
1	3	$3 \cdot 10^{-4}$	0.46	0.53
1	4	$4 \cdot 10^{-5}$	0.46	0.52
2	1	$8 \cdot 10^{-2}$	0.48	0.51
2	2	$2 \cdot 10^{-3}$	0.47	0.51
2	3	$4 \cdot 10^{-5}$	0.46	0.53
2	4	$2 \cdot 10^{-6}$	0.46	0.52
5	1	$8 \cdot 10^{-3}$	0.48	0.51
5	2	$5 \cdot 10^{-5}$	0.46	0.53
5	3	$4 \cdot 10^{-7}$	0.47	0.52
5	4	$3 \cdot 10^{-9}$	0.47	0.51

j	z	c	q_1	q_2
1	1	$1 \cdot 10^{-1}$	0.48	0.51
1	2	$1 \cdot 10^{-1}$	0.48	0.51
1	5	$1 \cdot 10^{-1}$	0.49	0.49
2	1	$7 \cdot 10^{-2}$	0.48	0.51
2	2	$6 \cdot 10^{-2}$	0.50	0.49
2	5	$6 \cdot 10^{-2}$	0.49	0.48
3	1	$4 \cdot 10^{-2}$	0.49	0.50
3	2	$3 \cdot 10^{-2}$	0.49	0.50
3	5	$3 \cdot 10^{-2}$	0.49	0.49
4	1	$2 \cdot 10^{-2}$	0.49	0.50
4	2	$2 \cdot 10^{-2}$	0.49	0.50
4	5	$1 \cdot 10^{-2}$	0.48	0.50

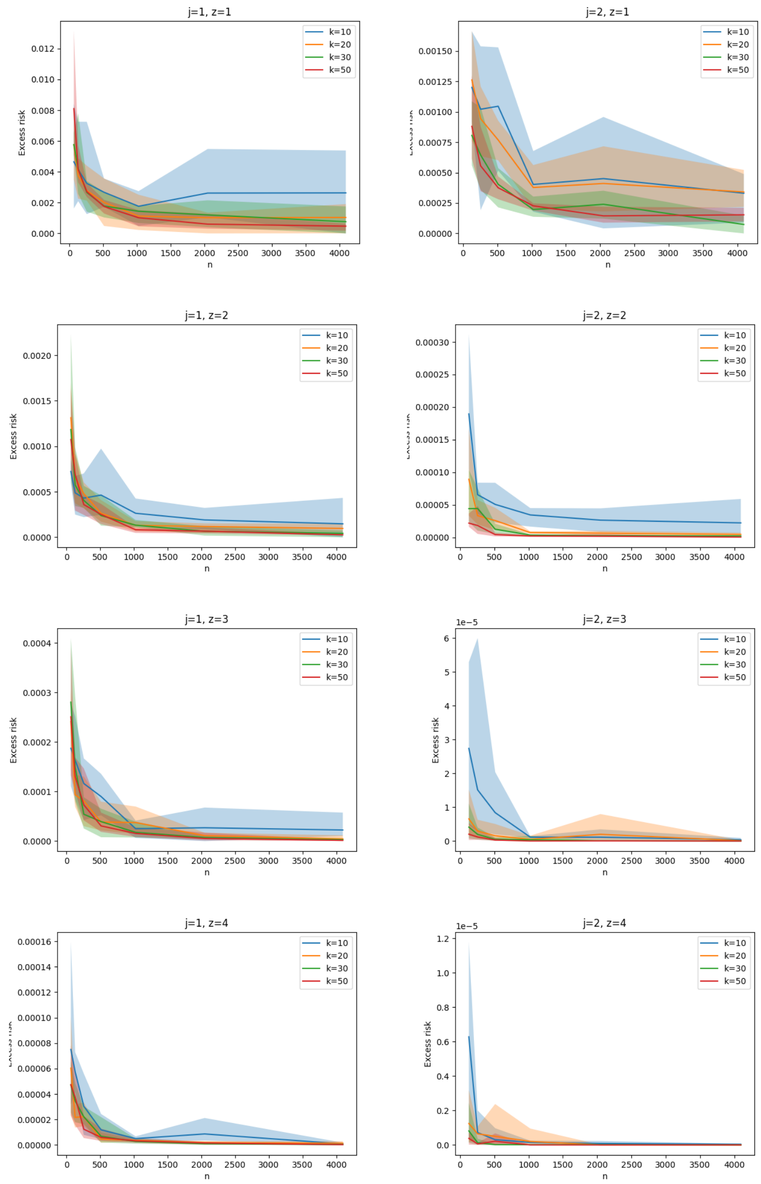

Figure 7: Excess risk for subspace clustering on the Skin_Nonskin dataset. The shaded areas indicate min/max values

Table 5: Best fit line for subspace clustering on Skin-Nonskin

j	z	c	q_1	q_2
1	1	$1 \cdot 10^{-2}$	0.48	0.50
1	2	$3 \cdot 10^{-3}$	0.45	0.53
2	1	$2 \cdot 10^{-3}$	0.46	0.53
2	2	$2 \cdot 10^{-4}$	0.46	0.53
3	1	$4 \cdot 10^{-4}$	0.46	0.53
3	2	$2 \cdot 10^{-5}$	0.46	0.53
4	1	$9 \cdot 10^{-5}$	0.46	0.53
4	2	$3 \cdot 10^{-6}$	0.46	0.53