A PROOFS AND DERIVATIONS

PROOF 1. ***(Proposition 1)*** *For any two sub-matrix : $X_1, X_2 \in R^{N\times 1}$, regardless of row or column sub-matrix of the original matrix. Note for any matrix X_S with column = 1, $\left(\mathbf{X_S}^\top \mathbf{X_S}\right)^{-1} = \|\mathbf{X_S}\|^2 = Vol(X_S)^2$. Considering the pseudo-inverse is denoted as: $\mathbf{X_S}^+ := \left(\mathbf{X_S}^\top \mathbf{X_S}\right)^{-1} \mathbf{X_S}^\top$, the bias variation is:*

$$\begin{aligned}
&bias\,_2^2 - \ bias\,_1^2 = \left\|X^+ - X_2^+\right\|^2 - \left\|X^+ - X_1^+\right\|^2 \\
&= \left\|\mathbf{X}^+\right\|^2 - 2\left\langle \mathbf{X}^+, \mathbf{X}_2^+ \right\rangle + \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+\right\|^2 + 2\left\langle \mathbf{X}^+, \mathbf{X}_1^+ \right\rangle - \left\|\mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}^+\right\|^2 - 2\left\|\mathbf{X}_2^+\right\|^2 + \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+\right\|^2 + 2\left\|\mathbf{X}_1^+\right\|^2 - \left\|\mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}_1^+\right\|^2 - \left\|\mathbf{X}_2^+\right\|^2 = \frac{1}{\|\mathbf{X_1}\|^4}\left\|\mathbf{X_1}^\top\right\|^2 - \frac{1}{\|\mathbf{X_2}\|^4}\left\|\mathbf{X_2}^\top\right\|^2 \\
&== \frac{1}{\|\mathbf{X_1}\|^2} - \frac{1}{\|\mathbf{X_2}\|^2} = \frac{1}{\mathrm{Vol}\left(\mathbf{X_1}\right)^2} - \frac{1}{\mathrm{Vol}(\mathbf{X_2})^2}
\end{aligned} \tag{1}$$

Since the $Vol(X)$ is constant under $column = 1$, a larger volume yields a smaller bias is proved. Combined with the relation $variation_1 - variation_2 =$ bias $_2 -$ bias $_1$, a larger volume yields a larger parameter variation is apparent.

PROOF 2. ***(Proposition 2)*** *The proposition **??** is expanded by substituting $\mathbf{X}^+ = \mathbf{G}^{-1}\mathbf{X}^\top$, where $\mathbf{G} := \mathbf{X}^\top\mathbf{X}$ is the left Gram matrix.*

$$\begin{aligned}
&\mathrm{bias}_2^2 - \mathrm{bias}_1^2 := \left\|\mathbf{X}^+ - \mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+ - \mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}^+\right\|^2 - 2\left\langle \mathbf{X}^+, \mathbf{X}_2^+ \right\rangle + \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}^+\right\|^2 + 2\left\langle \mathbf{X}^+, \mathbf{X}_1^+ \right\rangle - \left\|\mathbf{X}_1^+\right\|^2 \\
&= \left\|\mathbf{X}_2^+\right\|^2 - \left\|\mathbf{X}_1^+\right\|^2 + 2\left\langle \mathbf{X}^+, \mathbf{X}_1^+ - \mathbf{X}_2^+ \right\rangle \\
&= \left\|\mathbf{G}_2^{-1}\mathbf{X}_2^\top\right\|^2 - \left\|\mathbf{G}_1^{-1}\mathbf{X}_1^\top\right\|^2 + 2\left\langle \mathbf{G}^{-1}\mathbf{X}^\top, \mathbf{G}_1^{-1}\mathbf{X}_1^\top - \mathbf{G}_2^{-1}\mathbf{X}_2^\top \right\rangle
\end{aligned} \tag{2}$$

*Inspired by the work of Xu et al.[**?**], $\mathbf{G}^{-1}$ can be expanded by Sylvester's matrix theorem.*

THEORY 1. ***(Sylvester's Matrix Theorem).*** *Given a diagonalizable square matrix X, any analytic function $f()$ can be expanded,*

$$f(\mathbf{X}) = \sum_{l=1}^{k} f\left(\lambda_l\right)\mathbf{X}_l, \tag{3}$$

where λ_l is the i-th distinct eigenvalue of X and X_l is the Frobenius covariant as follows:

$$\mathbf{X}_l := \prod_{j=1, j\neq l}^{k} \frac{1}{\lambda_l - \lambda_j}\left(\mathbf{X} - \lambda_j\mathbf{I}\right). \tag{4}$$

COROLLARY 1. *Supposing $f()$ is a inverse function, then,*

$$f(\mathbf{X}) = \mathbf{X}^{-1} = \sum_{l=1}^{k} \frac{1}{\lambda_l}\mathbf{X}_l \tag{5}$$

For $\mathbf{G}$ is diagonalizable matrix, $\mathbf{G}^{-1}$ can be written as follows by above corollary.

$$\mathbf{G}^{-1} = \sum_{l=1}^{k} \frac{1}{\lambda_l}\mathbf{G}_l = \prod_{j=1, j\neq l}^{k} \frac{1}{\lambda_l - \lambda_j} \prod_{j=1, j\neq l}^{k} \left(\mathbf{G} - \lambda_j\mathbf{I}\right) \tag{6}$$

Denoting $p_l = \prod_{j=1, j\neq l}^{k} \frac{1}{\lambda_l - \lambda_j}$ and expanding the products, we have:

$$\frac{1}{p_l} = \Lambda \underbrace{\sum_{g=1}^{k} (-1)^{g+1} \lambda_l^{k-g} \left[\sum_{\mathcal{H} \subseteq \{1,\ldots,k\}\setminus\{l\}, |\mathcal{H}|=g-1} \left(\prod_{h \in \{1,\ldots,k\}\setminus\mathcal{H}} \frac{1}{\lambda_h} \right) \right]}_{\sigma_l}, \tag{7}$$

where $\Lambda = \prod_{i=l}^{k} \lambda_l = |\mathbf{G}|$. *The Equation 6 can be rewritten as:*

$$\begin{aligned} \mathbf{G}^{-1} &= \frac{1}{|\mathbf{G}|} \underbrace{\frac{1}{\sigma_l} \prod_{j=1, j\neq l}^{k} (\mathbf{G} - \lambda_j \mathbf{I})}_{\mathbf{Q}} \\ &= \frac{1}{\mathrm{Vol}(\mathbf{X})^2} \underbrace{\frac{1}{\sigma_l} \prod_{j=1, j\neq l}^{k} (\mathbf{G} - \lambda_j \mathbf{I})}_{\mathbf{Q}} \end{aligned} \tag{8}$$

Substituting $\mathbf{G}^{-1}(\mathbf{G_1}^{-1}, \mathbf{G_2}^{-1})$ *in Equation 2 by above result, the Equation 2 can be rewritten as:*

$$\mathrm{bias}_2^2 - \mathrm{bias}_1^2 = \frac{\left\|Q_2 X_2^T\right\|^2}{Vol(X_2)^4} - \frac{\left\|Q_1 X_1^T\right\|^2}{Vol(X_1)^4} + 2\left\langle \frac{Q_X X^T}{Vol(X)^2}, \frac{Q_1 X_1^T}{Vol(X_1)^2} - \frac{Q_2 X_2^T}{Vol(X_2)^2} \right\rangle \tag{9}$$

Lemma 1. *For any matrix* $X = \left[X_1^T, X_2^T\right]^T \in R^{N\times M}$ *and* X_1^T,$X_2^T \in R^{\frac{N}{2}\times M}$ *are submatrices in row, there exists* $Vol(X) \gg max(Vol(X_1), Vol(X_2))$ *when* $N \longrightarrow \infty$. *The relation is same for column submatrices.*

Lemma 2. *For a invertible matrix A and any column vector u and v, then there exits:*

$$\det\left(A + uv^T\right) = \det(A)\left(1 + v^T A^{-1} u\right) \tag{10}$$

Proof 3. ***(Lemma 1)*** *Considering a matrix* $\bar{X} = \left[X^T, x^T\right]^T$, *where* $X^T \in R^{n\times m}$ *is a submatrix,* $x^T \in R^{1\times m}$ *is a row vector. The square volume of* $\bar{X}$ *can be written as:*

$$\begin{aligned} Vol(\bar{X})^2 &= \left|\bar{X}^T \bar{X}\right| = \left|\left[X^T, x^T\right] \begin{bmatrix} X \\ x \end{bmatrix}\right| \\ &= \left|X^T X + x^T x\right| = \sigma_1 \left|X^T X\right|. \end{aligned} \tag{11}$$

According to the property of transposed matrix product (Lemma 2), the constant coefficient $\sigma = 1 + x(X^T X)^{-1} x^T$ *is greater than* 1. *Considering another submatrix* $X' = \left[x_1^T, ..., x_n^T\right]^T$ *and adding its row vectors in matrix X row by row, the square volume changing is written as:*

$$\begin{aligned} &Vol(\left[X^T, X'^T\right]^T) = Vol(\left[X^T, x_1^T, ..., x_n^T\right]^T) \\ &= \sigma_n Vol(\left[X^T, x_1^T, ..., x_{n-1}^T\right]^T) = ... = \prod_{i=1}^{n} \sigma_i Vol(X), \end{aligned} \tag{12}$$

where $\sigma_i = 1 + x_i(X^T X)^{-1} x_i^T$. *For each sigma is greater than 1,* $\prod_{i=1}^{n} \sigma_i \longrightarrow \infty$ *when* $n \longrightarrow \infty$. *Above all,* ***Lemma 1*** *is proved taking* X_1 *and* X_2 *as a new addition submatrix, respectively. Same conclusions can be applied in column submatrices.*

Proof 4. ***(Proposition 3)*** *Considering a replication-involving dataset* $X_{rep} = \left[X^T, \underbrace{X_S^T, ..., X_S^T}_{d}\right]^T$, *where* $X_S^T \in R$ *is row vector and is replicated for d times. According to Equation 12, the square volume of* X_{rep} *is written as:*

$$Vol(X_{rep})^2 = (1 + X_S(X^T X)^{-1} X_S^T)^d \left|X^T X\right| \tag{13}$$

For $(1 + X_S(X^T X)^{-1} X_S^T) > 1$, *the exponential increasing of volume under replication is proved. When* $lim_{d\rightarrow\infty} Vol\left(X_{rep}\right) = \infty$ *is hold. The same result can be applied in* X_S *in a submatrices form in row/column.*

Proof 5. ***(Proposition 5)*** *Considering a replication-involving matrix* $X_{rep} = replicate(X, c)$, *the inflation is written as:*

$$inflation = \frac{clusterRV(X_{rep})}{clusterRV(X)} = \frac{Vol(\widetilde{X}_{rep}) \prod_{i\in K} \rho_{rep,i}}{Vol(\widetilde{X}) \prod_{i\in K} \rho_i} \tag{14}$$

Due to direct copying, the clusters in X_{rep} *and* X *are with similar shapes and similar cluster centers, thus,* $Vol(\widetilde{X}_{rep}) \approx Vol(\widetilde{X})$.

Following summation formula of geometric progression,

$$1 \leq \rho_{rep,i} := \sum_{p=0}^{\phi_{rep,i}} \alpha^p = \frac{1 - \alpha^{(\phi_{rep,i}+1)}}{1 - \alpha} \leq \frac{1}{1 - \alpha} \tag{15}$$

$$1 \leq \rho_i := \textstyle\sum_{p=0}^{\phi_i} \alpha^p = \dfrac{1-\alpha^{(\phi_i+1)}}{1-\alpha} \leq \dfrac{1}{1-\alpha} \tag{16}$$

Thus,

$$(1-\alpha)^K \leq inflation \leq (1-\alpha)^{-K} \tag{17}$$

Proof 6. ***(Proposition 6)*** *When $\alpha = 1/\beta N$, the following inequality relation exists:*

$$1 \leq \rho_{rep,i} = \textstyle\sum_{p=0}^{\phi_{rep,i}} \frac{1}{\beta N}^p \leq \dfrac{1}{1-\frac{1}{\beta N}} = 1 + \dfrac{1}{\beta N - 1} \tag{18}$$

$$1 \leq \rho_i = \textstyle\sum_{p=0}^{\phi_i} \frac{1}{\beta N}^p \leq \dfrac{1}{1-\frac{1}{\beta N}} = 1 + \dfrac{1}{\beta N - 1} \tag{19}$$

Combined with Equation 14,

$$(1 + \frac{1}{\beta N - 1})^{-K} \leq inflation \leq (1 + \frac{1}{\beta N - 1})^K \tag{20}$$

When $N \longrightarrow \infty$, the following limit theorem is exist.

$$\lim_{n\to\infty} (1 + \frac{1}{\beta N - 1})^{-K} = \lim_{n\to\infty} (1 + \frac{1}{\beta N - 1})^K = 1 \tag{21}$$

Thus, $inflation \longrightarrow 1$ is hold under $N \longrightarrow \infty$.

Proof 7. ***(Proposition 7)*** *Taking the scenario where the dataset X is clustered to K partitions as an example. The matrix of cluster centers is denoted as $\widetilde{X}$. For a balanced dataset, each clusters C_i, $i = 1, ..., K$ is supposed to contain D data points. Then, we have*

$$\begin{aligned} Vol(X_1) &= \left| \left[(\widetilde{X}_1 + \Gamma_1)^T, ..., (\widetilde{X}_1 + \Gamma_D)^T \right]^T \begin{bmatrix} \widetilde{X}_1 + \Gamma_1 \\ \vdots \\ \widetilde{X}_1 + \Gamma_1 \end{bmatrix} \right| \\ &= \left| \sum_{i=1}^{D} (\widetilde{X}_1 + \Gamma_1)^T (X_1 + \Gamma_1) \right| \end{aligned} \tag{22}$$

where $\widetilde{X}_1$ is the matrix of cluster centers and $\Gamma_i = \begin{bmatrix} C_{0,i} - \widetilde{X}_0) \\ \vdots \\ C_{K,i} - \widetilde{X}_K \end{bmatrix}$ denotes the relative distance of i-th data points in each clusters to its cluster's center. Each cluster center generated by k-Means is:

$$\widetilde{X}_k = \frac{\sum_{i=1}^{D} \delta_{ik} x_i}{\sum_{i=1}^{D} \delta_{ik}}. \tag{23}$$

where δ_{ik} is a cluster indicator variable with $\delta_{ik} = 1$ if x_i in k-th cluster. As the number of data points increases, the distance vectors of the points to cluster center are cancel each other out,

$$\sum_{i=0}^{D} \Gamma_D \rightarrow \overrightarrow{0}. \tag{24}$$

Thus, the Equation 22 can be rewritten as:

$$Vol(X_1) = \left| \textstyle\sum_{i=1}^{D} (\widetilde{X}_1)^T (X_1) + \displaystyle\sum_{i=1}^{D} (\widetilde{\Gamma}_i)^T (\Gamma_i) \right|, \tag{25}$$

In the same way to get

$$Vol(X_2) = \left| \textstyle\sum_{i=1}^{D} (\widetilde{X}_2)^T (X_2) + \displaystyle\sum_{i=1}^{D} (\widetilde{\Upsilon}_i)^T (\Upsilon_{,i}) \right| \tag{26}$$

Applying determinant property,

$$\begin{aligned} \frac{Vol(X_1)}{Vol(X_2)} &= \frac{\left| D \cdot (\widetilde{X}_1)^T (X_1) + \sum_{i=1}^{D} (\Gamma_i)^T (\Gamma_i) \right|}{\left| D \cdot (\widetilde{X}_2)^T (X_2) + \sum_{i=1}^{D} (\Upsilon_i)^T (\Upsilon_i) \right|} \\ &= \frac{\left| (\widetilde{X}_1)^T (X_1) + \frac{1}{D} \sum_{i=1}^{D} (\Gamma_i)^T (\Gamma_i) \right|}{\left| (\widetilde{X}_2)^T (X_2) + \frac{1}{D} \sum_{i=1}^{D} (\Upsilon_i)^T (\Upsilon_i) \right|} \end{aligned} \tag{27}$$

When $N \to \infty$, Equation 21 has proved the following relation,

$$\frac{RV(\widetilde{X}_1)}{RV(\widetilde{X}_2)} = \frac{Vol(\widetilde{X}_1)\prod_{i\in K}\rho_{1,i}}{Vol(\widetilde{X}_2)\prod_{i\in K}\rho_{2,i}} \to \frac{Vol(\widetilde{X}_1)}{Vol(\widetilde{X}_2)} \tag{28}$$

Combined Equation 28 with Equation 27, $RV(\tilde{X_1})/RV(\tilde{X_2})]/[V(X_1)/V(X_2) \to 1$ under $N \to \infty$ is hold.

Definition 1. ***(Unbounded subset-sum problem)*** *Given a set of positive integers $\{k_0, ..., k_n\}$, an unbounded subset-sum problem is defined as to find the non-negative integers α_i so that $\sum_{i=1}^n \alpha_i k_i = K$, for we can achieve K by k_i for any times, it's known that unbounded subset-sum problem is NP-hard.*

Lemma 3. *Let $v_i = p_i, i = 1, ..., n$, $p_{n+1} = K + \triangle$ when $v_{n+1} = K$ and $\triangle \in (0, 1)$, a subadditive and monotone function $p(x)$ interpolating on the points (v_i, p_i) exist if and only if unbounded subset sum $\sum_{i=1}^n \alpha_i v_i = K$ not exists.*

Proof 8. ***(Lemma 3)*** *If $\sum_{i=1}^n \alpha_i p_i = K$ exists, then we have:*

$$K + \triangle = p(K) = p(\sum_{i=1}^{n} \alpha_i v_i) \le \sum_{i=1}^{n} \alpha_i p_i \overset{v_i = p_i}{=} K \tag{29}$$

$K + \triangle = K$ is a contradiction so that if $\sum_{i=1}^n \alpha_i v_i = K$ exists, a subadditive and monotone function $p(x)$ interpolating on the $n + 1$ points (v_i, p_i) is not exist.

Conversely, in the next, we prove if $\sum_{i=1}^n \alpha_i p_i = K$ not exists, we can construct a subadditive and monotone function $p(x)$ that interpolates the $(n + 1)$ points. First, we introduce a function (x) to reflect the smallest possible unbounded subset sum at x. (x) at least contrains an unbounded subset sum constains x, thus $(x) \ge x$. Then, we define a function $p(x) = min((x), K + \triangle)$ and our goal is to prove such $p(x)$ is satisfied subadditive, monotone and interpolating on the $n + 1$ points (v_i, p_i). It is apparent that $p(x)$ is monotone. Since a set containing x-self is a minumum unbounded subset sum, we have $_i = v_i = p_i \le K + \triangle$. For we have assumed that $\sum_{i=1}^n \alpha_i p_i = K$ is not exist, thus $(v_{i+1}) \ge K + 1$. Then the $p(x)$ can be written as:

$$p(x) = \begin{cases} \mu(x), & \mu(x) \le K \\ K + \triangle, & \mu(x) \ge K + 1 \end{cases} \tag{30}$$

If $\mu(x) \ge K + 1$, then $p(x + y) \le K + \triangle = p(x) \le p(x) + p(y)$. The relation is also holds when $\mu(y) \ge K + 1$. When both $\mu(x) \le K$ and $\mu(x) \le K$, we have $p(x) = (x) = \sum_{i=1}^n \alpha_i v_i$ and $p(y) = (y) = \sum_{i=1}^n \beta_i v_i$. Then, $x + y \le p(x) + p(y) = \sum_{i=1}^n (\alpha_i + \beta_i) v_i$. According to the definition of (x), $p(x + y) = (x + y) = min(x + y, \sum_{i=1}^n \gamma_i) \le \sum_{i=1}^n (\alpha_i + \beta_i) v_i = p(x) + p(y)$.

Above all, we have proved Lemma 3). However, the unbounded subset-sum problem is NP-hard, whether the sufficient and necessary conditions that unbounded subset sum $\sum_{i=1}^n \alpha_i v_i = K$ not existing in Lemma 3) is a co-NP hard problem.

Proof 9. ***(Proposition 8)*** *For the pricing function p satisfies $p(x)/x \ge p(y)/y$ when $x \le y$, then,it must have:*

$$\begin{aligned} \frac{p(x+y)}{x+y} &\le \min\left(\frac{p(x)}{x}, \frac{p(y)}{y}\right) \Rightarrow \\ p(x+y) &\le \min\left(p(x) + \underbrace{\frac{yp(x)}{x}}_{\ge p(y)}, p(y) + \underbrace{\frac{xp(y)}{y}}_{\ge p(x)} \right) \\ &\le p(x) + p(y) \end{aligned} \tag{31}$$

Constraint $p(x)/x \ge p(y)/y, x \le y$ representing a subspace of sub-additivity constraint is proved.