Appendix A

Proofs

Proposition A.1. *Minimisation of the cost function of K-Prototypes using the squared Euclidean distance and scaled Hamming distances for continuous and categorical variables, respectively implies that every prototype consists of the sample mean of continuous and the mode of categorical variables for observations in the cluster. That is:*

$$q_{k,j} = \begin{cases} \frac{1}{n_k} \sum\limits_{i:\mathbf{x}_i \in C_k} x_{i,j}, & 1 \leqslant j \leqslant p_C \\ \underset{l \in \{1,\dots,\ell_j\}}{\arg\max} \sum\limits_{i:\mathbf{x}_i \in C_k} \mathbf{I}\{x_{i,j} = l\}, & p_C < j \leqslant p, \end{cases}$$

where n_k is the number of observations in cluster C_k and $q_{k,j}$ is the $j$the component of the kth prototype.

Proof. The first part of the proof concerns continuous variables, that is $1 \leqslant j \leqslant p_C$. We fix a cluster $1 \leqslant k \leqslant K$. The cost function of K-Prototypes is given by:

$$E = \sum_{k=1}^{K} \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x}_i \in C_k\} d\left(\mathbf{x}_i, \mathbf{q}_k\right),$$

and the distance function we use is:

$$d(\mathbf{x}_i, \mathbf{q}_k) = \sum_{j=1}^{p_C} (x_{i,j} - q_{k,j})^2 + \gamma_k \sum_{j=p_C+1}^{p} \mathbb{I}\{x_{i,j} \neq q_{k,j}\}.$$

Therefore, the cost function associated with the continuous features for a fixed cluster C_k, denoted $E_{\text{cont},k}$, is given by:

$$E_{\text{cont},k} = \sum_{i=1}^{n} \left\{ \mathbb{I}\{\mathbf{x}_i \in C_k\} \sum_{j=1}^{p_C} (x_{i,j} - q_{k,j})^2 \right\}.$$

Fixing $1 \leqslant j \leqslant p_C$, we minimise E_{cont} as follows:

$$\frac{\partial E_{cont,k}}{\partial q_{k,j}} = 2 \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x}_i \in C_k\} (q_{k,j} - x_{i,j})$$

$$\implies \frac{\partial E_{cont,k}}{\partial q_{k,j}} = 0 \iff q_{k,j} \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x}_i \in C_k\} = \sum_{i=1}^{n} \mathbb{I}\{\mathbf{x}_i \in C_k\} x_{i,j}.$$

The sum on the left-hand side is equal to the number of observations in C_k, which we previously defined to be n_k. The expression on the right-hand side is equivalent to the sum of the jth component of all observations in C_k, so the result follows directly. Finally, the second derivative of $E_{cont,k}$ with respect to $q_{k,j}$ is equal to $2n_k > 0$, completing the proof for the continuous features.

We now consider the case of the categorical variables, so the associated cost function becomes:

$$E_{cat,k} = \gamma_k \sum_{i=1}^{n} \left\{ \mathbb{I}\{\mathbf{x}_i \in C_k\} \sum_{j=p_C+1}^{p} \mathbb{I}\{x_{i,j} \neq q_{k,j}\} \right\}.$$

Fixing $p_C < j \leqslant p$, we observe that the Expression above is minimised when the second binary indicator produces as many zeros as possible, that is for as many cases of $x_{i,j} = q_{k,j}$ as possible. This implies that $q_{k,j}$ has to be equal to the mode of the jth variable (assuming $\mathbf{X}_j$ is categorical for $p_C < j \leqslant p$) for all observations in C_k, thus completing the proof. □

Proposition A.2. *The Mixed Reduced K-Means solution is a special case of the simultaneous cluster analysis and dimensionality reduction problem for* $\kappa = 1/2$*, i.e.:*

$$\arg\min_{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}} \frac{1}{2} \left\{ \|\mathbf{X}' - \mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\|_F^2 + \|\mathbf{X}'\mathbf{B} - \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{X}'\mathbf{B}\|_F^2 \right\} =$$

$$\arg\min_{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}} \|\mathbf{X}' - \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\|_F^2.$$

Proof. We first consider the Mixed Reduced K-Means objective and we define the following projection matrix:

$$\mathbf{P} = \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal.$$

It is easy to see that $\mathbf{P}$ is a projection matrix, therefore it is idempotent and symmetric (i.e. $\mathbf{P}^\intercal = \mathbf{P}$, $\mathbf{P}^2 = \mathbf{P}^\intercal\mathbf{P} = \mathbf{P}$). Therefore, by expanding the Frobenius norm in terms

of traces, the Mixed Reduced K-Means objective becomes:

$$\begin{aligned}\underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\min}\, \|\mathbf{X}' - \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\|_{\mathrm{F}}^2 &= \underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\min}\, \|\mathbf{X}' - \mathbf{P}\mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\|_{\mathrm{F}}^2 \\ &= \underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\min}\, \{\mathrm{tr}\,\{(\mathbf{X}')^\intercal(\mathbf{X}')\} - \mathrm{tr}\,\{\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{P}\mathbf{X}'\mathbf{B}\}\} \\ &= \underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\max}\, \mathrm{tr}\,\{\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{P}\mathbf{X}'\mathbf{B}\}\,.\end{aligned}$$

The last line follows by considering that $\mathrm{tr}\,\{(\mathbf{X}')^\intercal(\mathbf{X}')\}$ is constant. We now look at the simultaneous clustering and dimensionality reduction objective and again perform an expansion of the Frobenius norm. Since $\kappa = 1/2$, the objective function has a common factor of κ that can be omitted from the minimisation problem:

$$\begin{aligned}&\underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\min}\, \left\{\|\mathbf{X}' - \mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\|_{\mathrm{F}}^2 + \|\mathbf{X}'\mathbf{B} - \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbf{X}'\mathbf{B}\|_{\mathrm{F}}^2\right\} \\ &= \underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\min}\, \left\{\|\mathbf{X}' - \mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\|_{\mathrm{F}}^2 + \|\mathbf{X}'\mathbf{B} - \mathbf{P}\mathbf{X}'\mathbf{B}\|_{\mathrm{F}}^2\right\} \\ &= \underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\min}\, \{\mathrm{tr}\,\{(\mathbf{X}')^\intercal\mathbf{X}'\} - \mathrm{tr}\{\mathbf{B}\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\} + \mathrm{tr}\{\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{X}'\mathbf{B}\}) - \mathrm{tr}\{\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{P}\mathbf{X}'\mathbf{B}\}\} \\ &= \underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\min}\, \{\mathrm{tr}\,\{(\mathbf{X}')^\intercal(\mathbf{X}')\} - \mathrm{tr}\,\{\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{P}\mathbf{X}'\mathbf{B}\}\} \\ &= \underset{\substack{\mathbf{Z}\in\mathcal{Z}\\ \mathbf{B}\in\mathcal{B}}}{\arg\max}\, \mathrm{tr}\,\{\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{P}\mathbf{X}'\mathbf{B}\}\,.\end{aligned}$$

This follows from the cyclic property of the trace, which implies:

$$\begin{aligned}\mathrm{tr}\{\mathbf{B}\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{X}'\mathbf{B}\mathbf{B}^\intercal\} &= \mathrm{tr}\{\mathbf{B}^\intercal\mathbf{B}\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{X}'\mathbf{B}\} \\ &= \mathrm{tr}\{\mathbf{B}^\intercal(\mathbf{X}')^\intercal\mathbf{X}'\mathbf{B}\},\end{aligned}$$

since $\mathbf{B}$ is columnwise orthonormal matrix and thus $\mathbf{B}^\intercal\mathbf{B} = \mathbf{I}_{\mathrm{s}}$. Therefore, both Expressions lead to the exact same optimisation problem, which implies their equivalence. □

Proposition A.3. *The entropy of a discrete random variable* X *with support* $\mathcal{X}$ *is bounded above by* $\log(|\mathcal{X}|)$*, where* $|\mathcal{X}|$ *is the cardinality of* $\mathcal{X}$*.*

Proof. We will first show that the entropy of X is maximised when this is a discrete uniform random variable and then derive its entropy. Assuming, without loss of generality, that X follows the discrete uniform distribution with support given by the natural numbers $\{1, \ldots, |\mathcal{X}|\}$, its probability mass function is given by $\mathbb{P}(\mathsf{X} =$

$x) = 1/|\mathcal{X}|\ \forall x \in \mathcal{X}$. Now let Y be a discrete random variable defined on the same support $\mathcal{X}$. Define the distributions of X and Y by Q and P, respectively. Then, the Kullback-Leibler divergence between P and Q in bits is given by:

$$\begin{aligned} 0 \leqslant D_{KL}\left(P \| Q\right) &= \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \\ &= \sum_{x \in \mathcal{X}} P(x) \log P(x) - \sum_{x \in \mathcal{X}} P(x) \log Q(x) \\ &= -H(Y) - \sum_{x \in \mathcal{X}} P(x) \log Q(x), \end{aligned}$$

where $H(Y)$ is the entropy of Y in bits. Then, we can further simplify the second term, since we know that $Q(x) = 1/|\mathcal{X}|\ \forall x \in \mathcal{X}$, which gives:

$$\begin{aligned} D_{KL}\left(P \| Q\right) &= -H(Y) - \log\left(\frac{1}{|\mathcal{X}|}\right) \sum_{x \in \mathcal{X}} P(x) \\ &= -H(Y) - \log\left(\frac{1}{|\mathcal{X}|}\right). \end{aligned}$$

Finally, it suffices to see that the entropy of X, that is the entropy of the discrete uniform distribution on $\mathcal{X}$, is in fact equal to the second term:

$$\begin{aligned} H(X) &= -\sum_{x \in \mathcal{X}} Q(x) \log Q(x) \\ &= -\log\left(\frac{1}{|\mathcal{X}|}\right) \sum_{x \in \mathcal{X}} Q(x) \\ &= -\log\left(\frac{1}{|\mathcal{X}|}\right) \\ &= \log(|\mathcal{X}|). \end{aligned}$$

Combining everything, we get:

$$\begin{aligned} 0 \leqslant D_{KL}(P \| Q) &= -H(Y) + \log\left(|\mathcal{X}|\right) \\ &= -H(Y) + H(X), \end{aligned}$$

which gives $H(X) \geqslant H(Y)$, proving that the discrete uniform distribution is a maximum entropy distribution with support $\mathcal{X}$ and its entropy is given by $\log(|\mathcal{X}|)$. $\square$

Proposition A.4. *Let* X, $\mathfrak{K}$, *and* Y *be random variables and consider the IB optimisation problem:*

$$q^*_{IB}(\mathfrak{K} \mid X) = \arg\min_{q(\mathfrak{K}|X)} \mathfrak{F}, \;\; \mathfrak{F} = I(X;\mathfrak{K}) - \beta I(\mathfrak{K};Y).$$

Under the assumption of $\mathfrak{F}$ *being differentiable with respect to* $q(\mathfrak{K} \mid X)$ *and the Markov constraint* $\mathfrak{K} \leftrightarrow X \leftrightarrow Y$*, the solution is given by:*

$$\begin{aligned} q^*_{IB}(\mathfrak{K} \mid X) &= \frac{q^*_{IB}(\mathfrak{K})}{c(\beta)} \exp\{-\beta D_{KL}\left(p(Y \mid X) \| q^*_{IB}(Y \mid \mathfrak{K})\right)\}, \\ q^*_{IB}(Y \mid \mathfrak{K}) &= \frac{1}{q^*_{IB}(\mathfrak{K})} \sum_{x \in \mathcal{X}} q^*_{IB}(\mathfrak{K} \mid X) p(X,Y), \\ q^*_{IB}(\mathfrak{K}) &= \sum_{x \in \mathcal{X}} p(X) q^*_{IB}(\mathfrak{K} \mid X). \end{aligned}$$

Proof. We begin by noticing that $I(X;\mathfrak{K}) = H(\mathfrak{K}) - H(\mathfrak{K} \mid X)$, so we solve the more general problem of minimising $\mathfrak{F}_\alpha = H(\mathfrak{K}) - \alpha H(\mathfrak{K} \mid X) - \beta I(\mathfrak{K};Y)$. It is immediately obvious that the solution to the IB problem emerges as a special case that corresponds to $\alpha = 1$. The Markov constraint also implies the following:

$$q(\mathfrak{K} \mid Y) = \sum_{x \in \mathcal{X}} q(\mathfrak{K} \mid X) p(X \mid Y), \tag{A.1}$$

$$q(\mathfrak{K}) = \sum_{x \in \mathcal{X}} q(\mathfrak{K} \mid X) p(X). \tag{A.2}$$

We differentiate with respect to $q(\mathfrak{K} \mid X)$ (using calculus of variations) and obtain the following useful derivatives:

$$\begin{aligned} \frac{\delta q(\mathfrak{K} \mid Y)}{\delta q(\mathfrak{K} \mid X)} &= p(X \mid Y) \\ \frac{\delta q(\mathfrak{K})}{\delta q(\mathfrak{K} \mid X)} &= p(X). \end{aligned}$$

We introduce a Lagrange multiplier $\lambda(X)$ for the normalisation of $q(\mathfrak{K} \mid X)$, so the cost function we seek to minimise is given by:

$$\begin{aligned}
\mathfrak{L}_\alpha &= \mathfrak{F}_\alpha - \sum_{x\in\mathcal{X}}\sum_{k=1}^{K}\lambda(X)q(\mathfrak{K}\mid X)\\
&= H(\mathfrak{K}) - \alpha H(\mathfrak{K}\mid X) - \beta I(\mathfrak{K};Y) - \sum_{x\in\mathcal{X}}\sum_{k=1}^{K}\lambda(X)q(\mathfrak{K}\mid X)\\
&= -\sum_{k=1}^{K} p(\mathfrak{K})\log p(\mathfrak{K}) + \alpha\sum_{x\in\mathcal{X}}\sum_{k=1}^{K} q(\mathfrak{K}\mid X)p(X)\log q(\mathfrak{K}\mid X)\\
&\quad - \beta\sum_{y\in\mathcal{Y}}\sum_{k=1}^{K} q(\mathfrak{K}\mid Y)p(Y)\log\left\{\frac{q(\mathfrak{K}\mid Y)}{q(\mathfrak{K})}\right\} - \sum_{x\in\mathcal{X}}\sum_{k=1}^{K}\lambda(X)q(\mathfrak{K}\mid X).
\end{aligned}$$

Taking the derivative of $\mathfrak{L}_\alpha$ with respect to $q(\mathfrak{K}\mid X)$ for a given value $x \in \mathcal{X}$ and for fixed $k \in \{1,\dots,K\}$, we get:

$$\begin{aligned}
\frac{\delta\mathfrak{L}_\alpha}{\delta q(\mathfrak{K}\mid X)} &= -\log q(\mathfrak{K})\frac{\delta q(\mathfrak{K})}{\delta q(\mathfrak{K}\mid X)} - q(\mathfrak{K})\frac{\delta\log q(\mathfrak{K})}{\delta q(\mathfrak{K}\mid X)}\\
&\quad + \alpha\left\{p(X)\log q(\mathfrak{K}\mid X)\frac{\delta q(\mathfrak{K}\mid X)}{\delta q(\mathfrak{K}\mid X)} + q(\mathfrak{K}\mid X)p(X)\frac{\delta\log q(\mathfrak{K}\mid X)}{\delta q(\mathfrak{K}\mid X)}\right\}\\
&\quad - \beta\sum_{y\in\mathcal{Y}}\left\{p(Y)\log\left(\frac{q(\mathfrak{K}\mid Y)}{q(\mathfrak{K})}\right)\frac{\delta q(\mathfrak{K}\mid Y)}{\delta q(\mathfrak{K}\mid X)}\right.\\
&\quad \left. -q(\mathfrak{K}\mid Y)p(Y)\left(\frac{\delta\log q(\mathfrak{K}\mid Y)}{\delta q(\mathfrak{K}\mid X)} + \frac{\delta\log q(\mathfrak{K})}{\delta q(\mathfrak{K}\mid X)}\right)\right\}\\
&\quad - \lambda(X)\frac{\delta q(\mathfrak{K}\mid X)}{\delta q(\mathfrak{K}\mid X)}\\
&= -p(X)\log q(\mathfrak{K}) - p(X) + \alpha\{p(X)\log q(\mathfrak{K}\mid X) + p(X)\} - \lambda(X)\\
&\quad - \beta\sum_{y\in\mathcal{Y}}\left\{p(Y)\log\left(\frac{q(\mathfrak{K}\mid Y)}{q(\mathfrak{K})}\right)p(X\mid Y) + p(Y)p(X\mid Y) - q(\mathfrak{K}\mid Y)p(Y)\frac{p(X)}{q(\mathfrak{K})}\right\}\\
&= -p(X)\log q(\mathfrak{K}) - p(X) + \alpha\{p(X)\log q(\mathfrak{K}\mid X) + p(X)\} - \lambda(X)\\
&\quad - \beta p(X)\left\{\sum_{y\in\mathcal{Y}} p(Y\mid X)\log\left(\frac{q(\mathfrak{K}\mid Y)}{q(\mathfrak{K})}\right) + \sum_{y\in\mathcal{Y}} p(Y\mid X) - \sum_{y\in\mathcal{Y}} q(Y\mid\mathfrak{K})\right\}.
\end{aligned}$$

Taking $p(X)$ as a common factor and noticing that the last two terms in the final line sum to a unit and thus cancel each other out, we get:

$$\begin{aligned}
\frac{\delta\mathfrak{L}_\alpha}{\delta q(\mathfrak{K}\mid X)} = p(X)\Bigg\{&-\log q(\mathfrak{K}) - 1 + \alpha\log q(\mathfrak{K}\mid X) + \alpha - \frac{\lambda(X)}{p(X)}\\
&-\beta\sum_{y\in\mathcal{Y}} p(Y\mid X)\log\left(\frac{q(\mathfrak{K}\mid Y)}{q(\mathfrak{K})}\right)\Bigg\}.
\end{aligned}$$

Setting the derivative to zero for $q^*_{IB}(\mathfrak{K} \mid X)$, $q^*_{IB}(\mathfrak{K})$ and $q^*_{IB}(\mathfrak{K} \mid Y)$, we arrive at the following:

$$\alpha \log q^*_{IB}(\mathfrak{K} \mid X) = 1 - \alpha + \log q^*_{IB}(\mathfrak{K}) + \frac{\lambda(X)}{p(X)} + \beta \sum_{y \in \mathcal{Y}} p(Y \mid X) \log \left(\frac{q^*_{IB}(\mathfrak{K} \mid Y)}{q^*_{IB}(\mathfrak{K})} \right).$$

We manipulate the last term as follows:

$$\log \left(\frac{q^*_{IB}(\mathfrak{K} \mid Y)}{q^*_{IB}(\mathfrak{K})} \right) = \log \left(\frac{q^*_{IB}(\mathfrak{K}, Y)}{q^*_{IB}(\mathfrak{K}) p(Y)} \right) = \log \left(\frac{q^*_{IB}(Y \mid \mathfrak{K})}{p(Y)} \right).$$

We also notice the following:

$$\log \left(\frac{q^*_{IB}(Y \mid \mathfrak{K})}{p(Y)} \right) - \log \left(\frac{p(Y \mid X)}{p(Y)} \right) = -\log \left(\frac{p(Y \mid X)}{q^*_{IB}(Y \mid \mathfrak{K})} \right).$$

Thus, using the manipulation above and by adding and subtracting $\beta \sum_{y \in \mathcal{Y}} p(Y \mid X) \log(p(Y \mid X)/p(Y))$, we get:

$$\begin{aligned}
\alpha \log q^*_{IB}(\mathfrak{K} \mid X) &= 1 - \alpha + \log q^*_{IB}(\mathfrak{K}) + \frac{\lambda(X)}{p(X)} \\
&\quad + \beta \sum_{y \in \mathcal{Y}} p(Y \mid X) \log \left(\frac{p(Y \mid X)}{p(Y)} \right) - \beta \sum_{y \in \mathcal{Y}} \log \left(\frac{p(Y \mid X)}{q^*_{IB}(Y \mid \mathfrak{K})} \right) \\
&= 1 - \alpha + \log q^*_{IB}(\mathfrak{K}) + \frac{\lambda(X)}{p(X)} + D_{KL}\left(p(Y \mid X) \| q^*_{IB}(Y \mid \mathfrak{K})\right) \\
&\quad - \beta \sum_{y \in \mathcal{Y}} \log \left(\frac{p(Y \mid X)}{q^*_{IB}(Y \mid \mathfrak{K})} \right).
\end{aligned}$$

Dividing both sides by α and taking exponentials, we come to the following final expression:

$$q^*_{IB}(\mathfrak{K} \mid X) = \frac{1}{c(\beta)} \exp \left\{ \frac{1}{\alpha} \left(\log q^*_{IB}(\mathfrak{K}) - \beta D_{KL}\left(p(Y \mid X) \| q^*_{IB}(Y \mid \mathfrak{K})\right) \right) \right\}, \tag{A.3}$$

where the normalisation term $c(\beta)$ is given by:

$$c(\beta) = \exp \left\{ -\frac{1}{\alpha} + 1 - \frac{\lambda(X)}{\alpha p(X)} - \frac{\beta}{\alpha} \sum_{y \in \mathcal{Y}} p(Y \mid X) \log \left(\frac{p(Y \mid X)}{p(Y)} \right) \right\},$$

while the expressions for $q^*_{IB}(\mathfrak{K})$ and for $q^*_{IB}(\mathfrak{K} \mid Y)$ can be derived by substituting $q^*_{IB}(\mathfrak{K} \mid X)$ in Expressions (A.2) & (A.1), respectively. The IB solution follows directly by setting $\alpha = 1$. □

Proposition A.5. *Let $\mathfrak{K}$ and $\mathfrak{K}'$ be the random variables associated with two partitions into K and $K+1$ clusters, respectively obtained by DIBmix. Assuming that the additional cluster is obtained by splitting one of the original K clusters, it holds that $I(Y;\mathfrak{K}) < I(Y;\mathfrak{K}')$, where Y is the random variable associated with the location of each observation.*

Proof. We begin by considering the definition of mutual information. More precisely, for any random variable associated with a partition $\mathfrak{K}$:

$$I(Y;\mathfrak{K}) = H(Y) - H(Y \mid \mathfrak{K}),$$

where $H(Y \mid \mathfrak{K})$ is the conditional entropy of the location of observations Y given their cluster assignment $\mathfrak{K}$. Notice that the conditional entropy can be rewritten as the following weighted sum:

$$H(Y \mid \mathfrak{K}) = \sum_{k=1}^{K} \mathbb{P}(\mathfrak{K} = k) H(Y \mid \mathfrak{K} = k).$$

Given that $H(Y)$ is independent of the partition obtained and assuming without loss of generality that the Kth and the $(K+1)$st clusters in $\mathfrak{K}'$ emerge by splitting the Kth cluster of $\mathfrak{K}$, our problem is equivalent to that of proving the following:

$$\mathbb{P}(\mathfrak{K}' = K) H(Y \mid \mathfrak{K}' = K) + \mathbb{P}(\mathfrak{K}' = K+1) H(Y \mid \mathfrak{K}' = K+1) < \mathbb{P}(\mathfrak{K} = K) H(Y \mid \mathfrak{K} = K).$$

Notice that $\mathbb{P}(\mathfrak{K} = K) = \mathbb{P}(\mathfrak{K}' = K) + \mathbb{P}(\mathfrak{K}' = K+1)$, hence it suffices to show that $H(Y \mid \mathfrak{K}' = K) < H(Y \mid \mathfrak{K} = K)$ and $H(Y \mid \mathfrak{K}' = K+1) < H(Y \mid \mathfrak{K} = K)$. Let us begin by considering $H(Y \mid \mathfrak{K}' = K+1)$. For the sake of simplicity, assume without loss of generality that the first n_K observations were assigned into the Kth cluster in $\mathfrak{K}$ and the first $n'_{K+1} < n_K$ of them are then assigned in the $(K+1)$st cluster in $\mathfrak{K}'$. What this means is that for observations x for which $q(\mathfrak{K}' = C+1 \mid X = x) = 1$ and ensuring that the value of β is large enough for no cluster to be dropped, the following inequality needs to hold:

$$D_{KL}\left(p_{Y|X} \| q_{Y|K+1}\right) < D_{KL}\left(p_{Y|X} \| q_{Y|k}\right) \;\; \forall k \in \{1, \ldots, K\}. \tag{A.4}$$

We know that the clustering is being done based on the perturbed similarity matrix $\mathbf{P}' \equiv \mathbf{P}'_{Y|X}$, with observations for which the corresponding $p(Y = y \mid X = x)$ values are large being more likely to be assigned in the same cluster. Hence, the entries of the block $\mathbf{P}'_{[(1:n'_{K+1})\times(1:n'_{K+1})]}$ will take larger values than these of $\mathbf{P}'_{[(1:n'_{K+1})\times(n'_{K+1}:n_K)]}$. Consequently, for Expression (A.4) to hold and since the first n'_{K+1} elements are the

ones for which $q(\mathfrak{K}' = K+1 \mid X = x) = 1$, the first n'_{K+1} elements of $q(Y \mid \mathfrak{K}' = K+1)$ must be larger than the first n'_{K+1} elements of $q(Y \mid \mathfrak{K}' = K)$. However, if the first n'_{K+1} elements of $q(Y \mid \mathfrak{K}' = K+1)$ increase, then the remaining $n - n'_{K+1}$ elements must decrease, so as to ensure that all elements of $q(Y \mid \mathfrak{K}' = K+1)$ sum to a unit. This leads to the distribution of $q(Y \mid \mathfrak{K}' = K+1)$ being more skewed than that of $q(Y \mid \mathfrak{K} = K)$ which implies that $H(Y \mid \mathfrak{K}' = K+1) < H(Y \mid \mathfrak{K} = K)$. Similarly, it can be shown that $H(Y \mid \mathfrak{K}' = K) < H(Y \mid \mathfrak{K} = K)$. Combining everything, we get the desired result. □

Proposition A.6. *The maximum value of the nominal score of outlyingness* $s(\mathbf{x}_i)$ *for infrequent itemsets in a data set with* p *nominal variables is attained for all itemsets of length* $1 \leqslant k \leqslant p$ *appearing exactly once.*

Proof. We assume that we have $p \geqslant 2$ nominal variables (for the case of $p = 1$, the result follows immediately by the definition of the nominal score) and recall the formulation of the discrete score for an observation $1 \leqslant i \leqslant n$:

$$s(\mathbf{x}_i) = \sum_{\substack{d \subseteq \mathbf{x}_i: \\ \text{supp}(d) \notin (\sigma_d, n], \\ |d| \leqslant \text{MAXLEN}}} \frac{\sigma_d}{\text{supp}\,(d) \times |d|^r}, \quad r > 0.$$

Assuming a large MAXLEN value (to avoid having many restrictions on the calculation of the score), it is easy to see that itemsets of unit support with a large minimum support threshold value σ_d yield a higher increase in the score. This case corresponds to highly misspecified large probabilities for itemsets that only appear once in the data, for which $\sigma_d = n - \delta$, where δ is typically a small positive integer. Since σ_d decreases as $|d|$ gets larger and as this is a parameter that is hard to determine in advance, we will stick with the maximisation of the $1/|d|^r$ term, assuming the maximum possible σ_d value of $\sigma_d = n - 1$. Therefore, the total score becomes equal to the following expression:

$$\mathcal{A} = (n-1) \times \left\{ p - \sum_{i=1}^{p-1} \alpha_i + \sum_{i=1}^{p-1} \binom{\alpha_i}{i+1} \frac{1}{(i+1)^r} \right\},$$
$$\alpha_i \in \{0, i+1, \ldots, p\}, \; \sum_{i=1}^{p-1} \alpha_i \leqslant p.$$

Notice that the first two terms inside the curly brackets correspond to the contribution of the itemsets of unit length that appear once in the data set, while the third term corresponds to the contribution of all possible itemsets of greater length (up to length p) to the score. The coefficients α_i represent the number of nominal variables which

are included in infrequent itemsets (of unit support) of length $i+1$. We further impose the restriction that the sum of the α_i's is at most equal to p, since any itemsets of length $i+1$ are defined based on $i+1$ variables, and we are restricted to p nominal features. Moreover, we could potentially have no infrequent itemsets of length $i+1$ (in which case $\alpha_i = 0$) but if we do have any, then these must be observed in at least $i+1$ nominal features. Despite the abuse of notation, we assume that $C^0_{i+1} = 0$ for convenience.

Our goal is to find the set of values $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p-1})$ that maximise $\mathcal{A}$ and more precisely, we aim to show that this is achieved either when all the α_i's are equal to zero or when we are on the boundary of the solution space, i.e. when all the α_i's sum to p and all of them but one are exactly zero. The first case corresponds to maximising the nominal score for all p nominal features of an observation being unique within the data set, while the second is interpreted as achieving the maximum nominal score possible when all itemsets of length $i+1$ that can be generated from p nominal variables appear just once and there exist no itemsets of greater or lower length which are infrequent. In both these cases, we basically end up with the conclusion that the maximum score is attained for all itemsets of one specific length occurring once in the data set.

For $p < 2^{r+1} + 1$, we would rather have $\alpha_i = 0 \ \forall i$ if:

$$\sum_{i=1}^{p-1} \alpha_i > \sum_{i=1}^{p-1} \binom{\alpha_i}{i+1} \frac{1}{(i+1)^r}.$$

The above condition can be equivalently written, by some straightforward algebraic manipulations, as:

$$\sum_{i=1}^{p-1} \left\{ \alpha_i \times \left(\frac{(\alpha_i - 1) \times \cdots \times (\alpha_i - i)}{(i+1)^{r+1} \times i!} - 1 \right) \right\} < 0.$$

Since we assume that all the α_i's are non-negative, the expression above can be negative if and only if $\exists \alpha_i > 0$ such that:

$$\begin{aligned} F &= \frac{(\alpha_i - 1) \times \cdots \times (\alpha_i - i)}{(i+1)^{r+1} \times i!} - 1 \\ &= \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i - i) \times \Gamma(i+1) \times (i+1)^{r+1}} - 1 < 0, \end{aligned} \tag{A.5}$$

where $\Gamma(\cdot)$ is the gamma function. If $F < 0 \ \forall \alpha_i > 0$, we can ensure that $\mathcal{A}$ is maximised for all the α_i's being equal to zero. Since Expression (A.5) is maximised

when $i = 1$ and $\alpha_1 = p$, we obtain:

$$F = \frac{p-1}{2^{r+1}} < 1 \iff p < 2^{r+1} + 1.$$

Therefore, $\alpha_i = 0\ \forall i$ maximises expression $\mathcal{A}$ for $p < 2^{r+1} + 1$.

We now assume $p \geqslant 2^{r+1} + 1$. We first show that getting to the solution boundary is always preferable when having one non-zero α_i and then, we show that setting α_j equal to a non-zero value, where $i \neq j$, decreases the value of $\mathcal{A}$. We begin by assuming that there exists an index j such that $\alpha_j > 0$ and $\alpha_i = 0\ \forall i \neq j$ and we look at what happens to the value of $\mathcal{A}$ if we decrease α_j by a unit. Due to the restriction on the values that α_j can take, we need to assume that $j \neq p-1$, since α_{p-1} can only be equal to 0 or p. We denote the loss that results from decreasing j by a unit by G_1 and that is equal to:

$$\begin{aligned} G_1 &= p - \alpha_j + \binom{\alpha_j}{j+1} \times \frac{1}{(j+1)^r} - \left\{ p - \alpha_j + 1 + \binom{\alpha_j - 1}{j+1} \times \frac{1}{(j+1)^r} \right\} \\ &= \frac{1}{(j+1)^r} \left\{ \binom{\alpha_j}{j+1} - \binom{\alpha_j - 1}{j+1} \right\} - 1 \\ &= \frac{\Gamma(\alpha_j)}{(j+1)^{r-1} \times \Gamma(j+2) \times \Gamma(\alpha_j - j)} - 1. \end{aligned}$$

When decreasing α_j from its minimum possible non-zero value of $j+1$ to zero, we define G_2 to be the loss corresponding to this decrease:

$$\begin{aligned} G_2 &= p - \alpha_j + \binom{\alpha_j}{j+1} \times \frac{1}{(j+1)^r} - p \\ &= -(j+1) + \binom{j+1}{j+1} \times \frac{1}{(j+1)^r} \\ &= \frac{1}{(j+1)^r} - (j+1). \end{aligned}$$

Hence, $G_2 < 0\ \forall j \in \mathbb{Z}^+$, which implies $\mathcal{A}$ increases when going from $\alpha_j = j+1$ to $\alpha_j = 0$. Recalling that $\alpha_j > j$, we may write $\alpha_j = j + z$, where $z \in \{1, \ldots, p-j\}$. The expression for G_1 then becomes:

$$G_1 = \frac{\Gamma(j+z)}{(j+1)^{r-1} \times \Gamma(j+2) \times \Gamma(z)} - 1.$$

The above expression is increasing in z (this can be seen by considering the ratio $\Gamma(j+z)/\Gamma(z)$ and using Stirling's approximation for the gamma function) and

decreasing in j. Thus G_1 is minimised when $j = p - 2$ and $z = 1$, in which case:

$$G_1 = \frac{\Gamma(p-1)}{(p-1)^{r-1} \times \Gamma(p) \times \Gamma(1)} - 1 = \frac{1}{(p-1)^r} - 1 < 0,$$

since $p \geqslant 2^{r+1} + 1 > 2$ (recall $r > 0$). Thus, when we are on the boundary of the solution space, it is preferred for us to stay there, while if we are at the minimum non-zero value of α_j, it is preferred to simply set it to zero.

The final part of the proof consists of showing that if we are on the boundary of the solution space, with $\alpha_j = p$ and $\alpha_i = 0\ \forall i \neq j$, then activating α_k so that $\alpha_k = k+1$ (where $k \neq j$) leads to a decrease in the value of $\mathcal{A}$. Now clearly, if α_k is set equal to $k+1$, the value of α_j has to drop to $p - (k+1)$, so that the boundary constraint is not violated. We define the loss incurred by activating α_k by H:

$$\begin{aligned} H &= \binom{p}{j+1} \times \frac{1}{(j+1)^r} - \left\{ \binom{p-k-1}{j+1} \frac{1}{(j+1)^r} + \binom{k+1}{k+1} \frac{1}{(k+1)^r} \right\} \\ &= \frac{1}{(j+1)^r} \times \left\{ \binom{p}{j+1} - \binom{p-k-1}{j+1} \right\} - \frac{1}{(k+1)^r} \\ &= \frac{1}{(j+1)^r \times \Gamma(j+2)} \times \left\{ \frac{\Gamma(p+1)}{\Gamma(p-j)} - \frac{\Gamma(p-k)}{\Gamma(p-k-j-1)} \right\} - \frac{1}{(k+1)^r}. \end{aligned}$$

We distinguish between two possible cases, namely $j < k$ and $j > k$. In the former case, we obtain:

$$H = \frac{1}{(k-\beta+1)^r \times \Gamma(k-\beta+2)} \times \left\{ \frac{\Gamma(p+1)}{\Gamma(p-k+\beta)} - \frac{\Gamma(p-k)}{\Gamma(p-2k+\beta-1)} \right\} - \\ - \frac{1}{(k+1)^r},$$

while in the latter:

$$H = \frac{1}{(k+\beta+1)^r \times \Gamma(k+\beta+2)} \times \left\{ \frac{\Gamma(p+1)}{\Gamma(p-k-\beta)} - \frac{\Gamma(p-k)}{\Gamma(p-2k-\beta-1)} \right\} - \\ - \frac{1}{(k+1)^r},$$

where $\beta \in \mathbb{Z}^+$. These expressions for H give us additional restrictions, which are that $p - 2k + \beta - 1 > 0$ and $p - 2k - \beta - 1 > 0$, so that all terms are well-defined. We now show that $H > 0\ \forall k \neq j$ just for the case of $j < k$; the second case follows analogously. Let the first term of H be denoted by H_1 and denote the second one by H_2. It is obvious that $\beta = 1$ minimises H_1. Thus, we have:

$$H = \frac{1}{k^r \times \Gamma(k+1)} \left\{ \frac{\Gamma(p+1)}{\Gamma(p-k)} - \frac{\Gamma(p-k)}{\Gamma(p-2k)} \right\} - \frac{1}{(k+1)^r}.$$

Now we can write the ratio of H_1 over H_2 as:

$$\frac{H_1}{H_2} = \frac{(k+1)^r \times Q(k)}{k^r \times \Gamma(k+1)} = \frac{(1+1/k)^r \times Q(k)}{\Gamma(k+1)},$$

where $Q(k) = \Gamma(p+1)/\Gamma(p-k) - \Gamma(p-k)/\Gamma(p-2k)$. It is also easy to see that $Q(k)$ is increasing for k, therefore its minimum is attained when $k = 2$. In that case, we have:

$$\frac{H_1}{H_2} = \frac{(k+1)^r \times Q(k)}{k^r \times \Gamma(k+1)} = \frac{(1+1/k)^r \times Q(k)}{\Gamma(k+1)} \geqslant \frac{(3/2)^r \times Q(2)}{2} > 1.$$

Thus, the minimum value of $H < 1$ which shows $H > 0\ \forall k \neq j$. As a result activating α_k is not preferred as it leads to a positive loss value.

We have thus shown the following:

1. The greatest $\mathcal{A}$ value for $p < 2^{r+1} + 1$ is achieved when all the α_i's are equal to zero.
2. When $p \geqslant 2^{r+1} + 1$, $\mathcal{A}$ is maximised when we are on the boundary of the solution space.
3. If $p \geqslant 2^{r+1} + 1$ and we are on the boundary of the solution space with just one parameter α_j being equal to p, then activating any other parameter α_k (where $j \neq k$) so that it is not longer equal to zero, leads to a lower value of $\mathcal{A}$.

Given the above, we can conclude that the optimal solution for $p < 2^{r+1} + 1$ is attained when all parameters are equal to zero, while for $p \geqslant 2^{r+1} + 1$, the optimal solution is on the boundary of the solution space with just one parameter α_i being non-zero and thus equal to p, as required. □

Proposition A.7. *The maximum value of the score of nominal outlyingness* $s(x_i)$ *for the* i*th observation of a data set with* p *nominal variables is given by* $p(n-1)$ *for* $p < 2^{r+1} + 1$*, and by the following expression for* $p \geqslant 2^{r+1} + 1$*:*

$$s^{\max}(x_i) = (n-1)\frac{\binom{p}{\min\left\{MAXLEN, \left\lfloor \frac{p-r}{2} \right\rfloor\right\}}}{\left(\min\left\{MAXLEN, \left\lfloor \frac{p-r}{2} \right\rfloor\right\}\right)^r},$$

where $\lfloor \cdot \rfloor$ *is the floor function (i.e.* $\lfloor x \rfloor$ *is the largest integer that is smaller than or equal to* x*).*

Proof. As proven in Proposition A.6, if we assume that the largest possible value of $\sigma_d/\text{supp}\,(d)$ is given by $n-1$, the score of nominal outlyingness s is maximised when

we only have highly infrequent (or frequent) itemsets of unit length, for $p < 2^{r+1} + 1$. In this case, it is obvious that the total number of such itemsets is equal to p, thus the score becomes equal to $(n-1)p$.

Now for the case of $p \geqslant 2^{r+1} + 1$, as previously shown, the largest possible score can be achieved on the boundary of the solution space; that is for all itemsets of a specific length $k \leqslant p$ having unit support. The total score is then given by $(n-1) \times C_k^p/k^2$,where C is the binomial coefficient. We therefore seek to find the itemset length $k^* \in \mathbb{Z}^+$ that maximises this expression of the score, for fixed p. This means that the score for $k = k^* - 1$ and for $k^* + 1$ (we assume $1 < k^* < p$) will be less than that for k^*, or equivalently (omitting the common $(n-1)$ term):

$$\binom{p}{k^*+1} \frac{1}{(k^*+1)^r} - \binom{p}{k^*} \frac{1}{(k^*)^r} < 0, \tag{A.6}$$

$$\binom{p}{k^*} \frac{1}{(k^*)^r} - \binom{p}{k^*-1} \frac{1}{(k^*-1)^r} > 0. \tag{A.7}$$

Starting with Expression (A.6), we have:

$$\begin{aligned}
\binom{p}{k^*+1} \frac{1}{(k^*+1)^r} - \binom{p}{k^*} \frac{1}{(k^*)^r} &= \frac{p!}{(k^*+1)!\,(p-k^*-1)!} \frac{1}{(k^*+1)^r} - \\
&- \frac{p!}{(k^*)!\,(p-k^*)!} \frac{1}{(k^*)^r} \\
&= \frac{p!}{(k^*)!\,(p-k^*-1)!\,(k^*)^r} \times \\
&\times \left\{ \frac{1}{(k^*+1)\left(1+\frac{1}{k^*}\right)^r} - \frac{1}{p-k^*} \right\},
\end{aligned}$$

and since we know that the first term is strictly positive, it suffices to show that the expression inside the curly brackets is negative. We bring this to the following form:

$$\frac{1}{(k^*+1)\left(1+\frac{1}{k^*}\right)^r} - \frac{1}{p-k^*} = \frac{(p-k^*) - (k^*+1)\left(1+\frac{1}{k^*}\right)^r}{(k^*+1)\left(1+\frac{1}{k^*}\right)^r (p-k^*)}$$

, hence it suffices to show that the numerator is negative, since the denominator is again strictly positive. This gives:

$$\begin{aligned}
&(p-k^*) - (k^*+1)\left(1+\frac{1}{k^*}\right)^r < 0 \\
\iff &p < k^* + (k^*+1)\left(1+\frac{1}{k^*}\right)^r.
\end{aligned} \tag{A.8}$$

We proceed similarly to derive an additional bound based on Expression (A.7). More precisely:

$$\begin{aligned}\binom{p}{k^*}\frac{1}{(k^*)^r}-\binom{p}{k^*-1}\frac{1}{(k^*-1)^r} &= \frac{p!}{(k^*)!\,(p-k^*)!}\frac{1}{(k^*)^r}- \\ &- \frac{p!}{(k^*-1)!\,(p-k^*+1)!}\frac{1}{(k^*-1)^r} \\ &= -\frac{p!}{(k^*-1)!\,(p-k^*)!\,(k^*-1)^r}\times \\ &\times\left\{\frac{1}{p-k^*+1}-\frac{\left(1-\frac{1}{k^*}\right)^r}{k^*}\right\},\end{aligned}$$

so by noticing that the first term is strictly positive and the product should be positive, we require that the expression inside the curly brackets is negative. A bit of re-arrangement gives:

$$\frac{1}{p-k^*+1}-\frac{\left(1-\frac{1}{k^*}\right)^r}{k^*}=\frac{k^*-(p-k^*+1)\left(1-\frac{1}{k^*}\right)^r}{(p-k^*+1)\,k^*},$$

thus we require that the numerator is negative, since the denominator is strictly positive. Hence, we arrive at:

$$\begin{aligned}&k^*-(p-k^*+1)\left(1-\frac{1}{k^*}\right)^r<0 \\ \iff &p>\frac{k^*}{\left(1-\frac{1}{k^*}\right)^r}+k^*-1.\end{aligned}\tag{A.9}$$

Thus, k^* should be an integer such that Expressions (A.8) & (A.9) are satisfied. We will show that as $p\to\infty$, the solution can be approximated by $k^*=\lfloor\frac{p-r}{2}|$, where $\lfloor\cdot\rfloor$ is the floor function. In order to motivate this, we plot the above two bounds of p that we obtained for varying r in Figure A.1 to get an idea of what the region of solutions looks like. We also include a zoomed in version for the case of $r=2$ just to get a better understanding of what is going on in Figure A.2.

The points of intersection of the boundary curves is (as expected) given by $p=2^{r+1}+1$ and solving one of Expressions (A.8) or (A.9) for the corresponding k value. We can see from both Figures A.1 & A.2 that the two boundary curves become parallel as

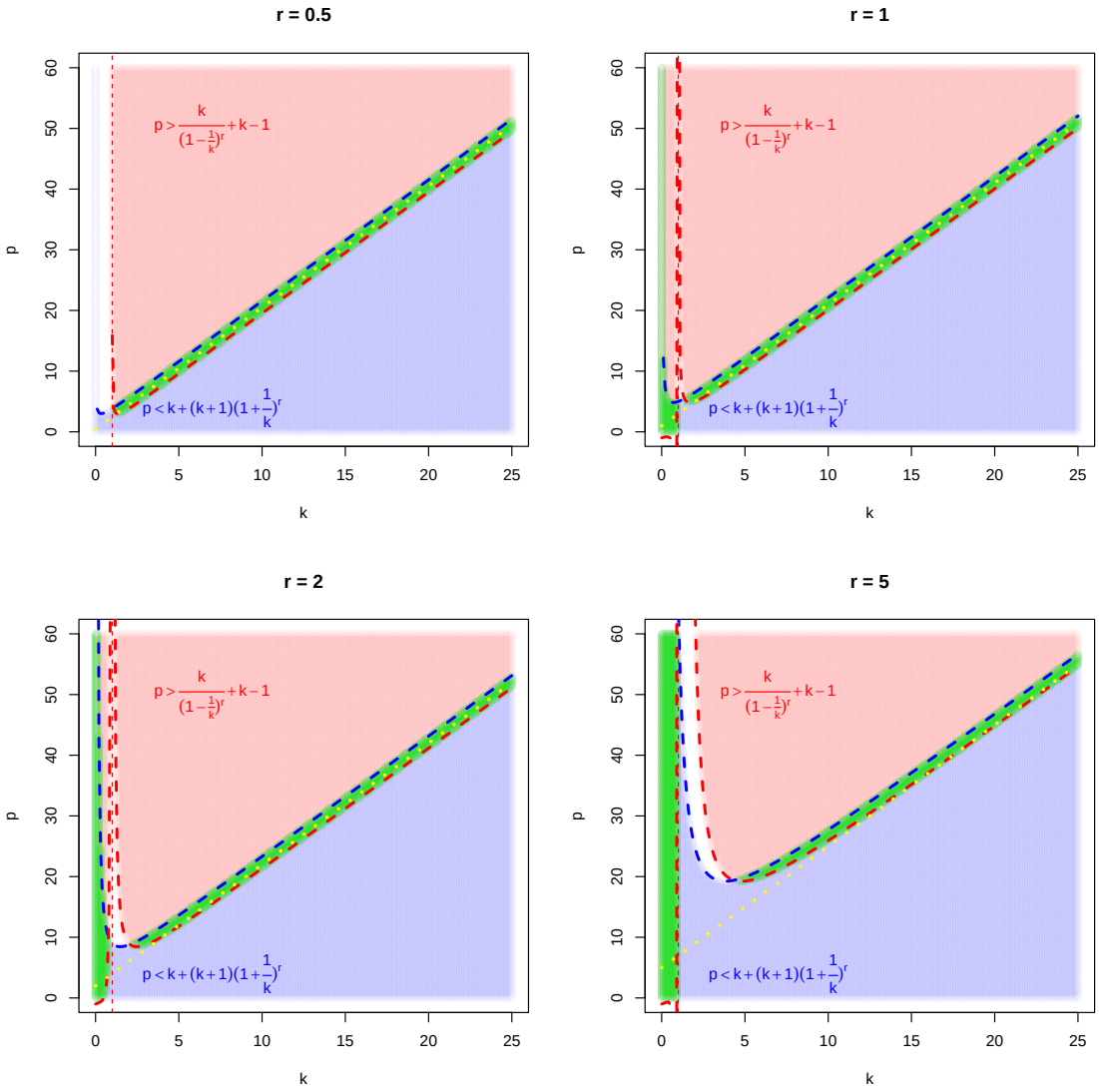

Figure A.1: Plot of bounds of p for varying r. The blue shaded region corresponds to the upper bound from Expression (A.8) and the red shaded region corresponds to the lower bound from Expression (A.9). The green shaded region is the region where both these bounds are satisfied. The vertical dashed line at $k = 1$ is the asymptote of the expression for the lower bound. The yellow dotted line is given by $p = 2k + r$; this is the line approximating the integer solutions in the green shaded region as $p \to \infty$.

$k \to +\infty$. We can compute their gradient to validate this:

$$\frac{d\left(k + (k+1)\left(1+\frac{1}{k}\right)^r\right)}{dk} = 1 + \left(1+\frac{1}{k}\right)^r - r\frac{k+1}{k^2}\left(1+\frac{1}{k}\right)^{r-1} \overset{k\to\infty}{\longrightarrow} 2,$$

$$\frac{d\left(\frac{k}{\left(1-\frac{1}{k}\right)^r} + k - 1\right)}{dk} = 1 + \left(1-\frac{1}{k}\right)^{-r} - \frac{r}{k}\left(1-\frac{1}{k}\right)^{-r-1} \overset{k\to\infty}{\longrightarrow} 2.$$

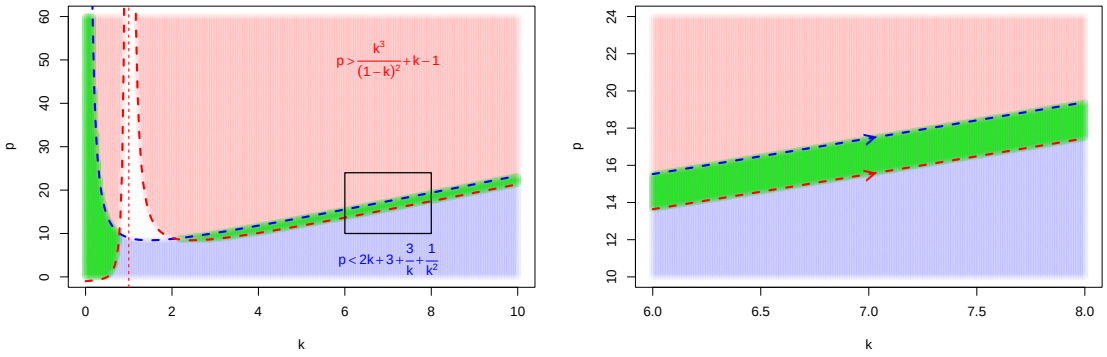

Figure A.2: Plot of bounds of p for $r = 2$. The rectangular region $[6, 8] \times [10, 24]$ on the left subplot is zoomed in on the right, illustrating how the 2 boundary curves become parallel as p increases.

Therefore, we can access the set of solutions (corresponding to the region where both bounds are satisfied) by considering the line $p = 2k + \alpha$. The constant α can be computed by noticing that this line will pass through the midpoint of each line segment defined by the two boundary curves for large k. Hence, we need to compute the difference between the two boundary curves:

$$k + (k+1)\left(1 + \frac{1}{k}\right)^r - \frac{k}{\left(1 - \frac{1}{k}\right)^r} + k - 1 = \frac{(k+1)(k^2-1)^r - k^{2r+1}}{k^r(k-1)^r} + 1 \overset{k\to\infty}{\longrightarrow} 2,$$

since the first term in the numerator involves a k term of degree $2r+1$ (which cancels with the negative term on the numerator) and a term k^{2r} that simplifies to 1 upon division. The rest of the terms tend to 0 as $k \to \infty$ so the only term that remains is the unit coefficient of k^{2r}, which means that for large k the vertical distance between the two boundary curves is equal to two units. Their midpoint is just one unit away from each curve, thus we require:

$$k + (k+1)\left(1 + \frac{1}{k}\right)^r - 2k - \alpha \overset{k\to\infty}{\longrightarrow} 1,$$

which is satisfied for $\alpha = r$. This is shown below using the binomial series expansion of $(1 + 1/k)^r$:

$$\begin{aligned} k + (k+1)\left(1 + \frac{1}{k}\right)^r - 2k - \alpha &= k + (k+1)\left(1 + \frac{r}{k} + \mathcal{O}(k^{-2})\right) - 2k - \alpha \\ &= k + k + r + 1 + \mathcal{O}(k^{-1}) - 2k - \alpha \\ &= r + 1 - \alpha + \mathcal{O}(k^{-1}) \\ &\overset{k\to\infty}{\longrightarrow} 1, \end{aligned}$$

giving $\alpha = r$, as required. This means that for large k values, we can compute k^* by accessing the region where the bounds for p are satisfied via the line $p = 2k + r$. We can only use this line to determine k^* as long as p is large enough. Moreover, notice that using this line gives $k = (p - r)/2$, which means that for non-integer or odd integer values of $p - r$ we do not get an integer value for k. It is easy to show that for these cases, the only integer k^* such that the bounds for p are satisfied is $k^* = \lfloor (p - r)/2 \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function ($\lfloor x \rfloor$ is the greatest integer x' such that $x' \leqslant x$).

The proof is completed by considering that MAXLEN is always at least equal to one; thus the maximum score for $p < 2^{r+1} + 1$ is equal to p, while for $p \geqslant 2^{r+1} + 1$, the maximum score is equal to the expression that we mentioned at the beginning of the proof, evaluated at $k = k^*$. However, since MAXLEN can be less than k^* and given that the maximum score expression is increasing for $k \in \mathbb{Z}^+_{\leqslant k^*}$, the maximum score of nominal outlyingness is attained for $k = \min\{\lfloor (p - r)/2 \rfloor, \text{MAXLEN}\}$. □

Proposition A.8. *Let* $\mathbf{X} = [(\mathbf{X}^C)^\intercal, (\mathbf{X}^O)^\intercal]^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, where* $\mathbf{X}^C = [\mathbf{X}_1, \ldots, \mathbf{X}_{p_C}]^\intercal$ *and* $\mathbf{X}^O = [\mathbf{X}_{p_C+1}, \ldots, \mathbf{X}_p]^\intercal$*. Assume* $\mathbf{m} = [(\mathbf{m}^C)^\intercal, (\mathbf{m}^O)^\intercal]^\intercal \in \mathbf{R}^{n \times p}$ *and* $\mathbf{S} \in \mathbf{R}^{p \times p}$ *are consistent estimators of location and scatter of* $\mathbf{X}$*. Define the squared Mahalanobis distance as:*

$$D^2 \mid \mathbf{X}^C, \mathbf{Y} = \left(\hat{\mathbf{X}} - \mathbf{m}\right)^\intercal \mathbf{S}^{-1} \left(\hat{\mathbf{X}} - \mathbf{m}\right),$$

where $\hat{\mathbf{X}} = [(\mathbf{X}^C)^\intercal, (\hat{\mathbf{X}}^O)^\intercal]$ *and* $\hat{\mathbf{X}}^O = [\mathbb{E}(\mathbf{X}_{p_C+1} \mid \mathbf{X}^C, \mathbf{Y}), \ldots, \mathbb{E}(\mathbf{X}_p \mid \mathbf{X}^C, \mathbf{Y})]^\intercal$*. Then,* $p_C \leqslant \mathbb{E}\left(D^2 \mid \mathbf{X}^C, \mathbf{Y}\right) \leqslant p$.

Proof. We start by partitioning the covariance matrix $\mathbf{S}$ as follows:

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}^{CC} & \mathbf{S}^{CO} \\ \mathbf{S}^{OC} & \mathbf{S}^{OO} \end{pmatrix}.$$

Using the decomposition of $\mathbf{X}$, the Mahalanobis distance becomes:

$$\begin{aligned}\mathrm{D}^2 \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y} &= \begin{pmatrix} \mathbf{X}^{\mathrm{C}} - \mathbf{m}^{\mathrm{C}} \\ \hat{\mathbf{X}}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} \end{pmatrix}^{\intercal} \mathbf{S}^{-1} \begin{pmatrix} \mathbf{X}^{\mathrm{C}} - \mathbf{m}^{\mathrm{C}} \\ \hat{\mathbf{X}}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} \end{pmatrix} \\ &= \left(\mathbf{X}^{\mathrm{C}} - \mathbf{m}^{\mathrm{C}}\right)^{\intercal} \left(\mathbf{S}^{\mathrm{CC}}\right)^{-1} \left(\mathbf{X}^{\mathrm{C}} - \mathbf{m}^{\mathrm{C}}\right) + \\ &+ \left(\hat{\mathbf{X}}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right)^{\intercal} \left(\mathbf{S}^{\mathrm{O|C}}\right)^{-1} \left(\hat{\mathbf{X}}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right), \end{aligned} \tag{A.10}$$

which is derived by considering the Schur complement to get an expression for the inverse of $\mathbf{S}$, and $\mathbf{m}^{\mathrm{O|C}}, \mathbf{S}^{\mathrm{O|C}}$ are defined as follows:

$$\begin{aligned}\mathbf{m}^{\mathrm{O|C}} &= \mathbf{S}^{\mathrm{OC}} \left(\mathbf{S}^{\mathrm{CC}}\right)^{-1} \left(\mathbf{X}^{\mathrm{C}} - \mathbf{m}^{\mathrm{C}}\right), \\ \mathbf{S}^{\mathrm{O|C}} &= \mathbf{S}^{\mathrm{OO}} - \mathbf{S}^{\mathrm{OC}} \left(\mathbf{S}^{\mathrm{CC}}\right)^{-1} \mathbf{S}^{\mathrm{CO}}. \end{aligned}$$

Taking the expectation of Expression (A.10), the first term is equal to $\mathfrak{p}_{\mathrm{C}}$ as it is asymptotically distributed according to a $\chi^2_{\mathfrak{p}_{\mathrm{C}}}$ distribution, whereas the second term is a quadratic form that is therefore going to have a non-negative expectation. This is due to $\mathbf{S}^{\mathrm{O|C}}$ being positive definite, proving the lower bound of $\mathbb{E}(\mathrm{D}^2 \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y})$ is equal to $\mathfrak{p}_{\mathrm{C}}$. Now let $\mathbf{Z} = \left(\mathbf{S}^{\mathrm{O|C}}\right)^{-1/2} \left(\mathbf{X}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right)$. Since we know that $\mathbf{X}^{\mathrm{O}} \mid \mathbf{X}^{\mathrm{C}} \sim \mathcal{N}\left(\mathbf{m}^{\mathrm{O}} + \mathbf{m}^{\mathrm{O|C}}, \mathbf{S}^{\mathrm{O|C}}\right)$, we deduce that:

$$\mathbb{E}\left\{\left(\mathbf{X}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right)^{\intercal} \left(\mathbf{S}^{\mathrm{O|C}}\right)^{-1} \left(\mathbf{X}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right)\right\} = \mathbb{E}\left\{\|\mathbf{Z}\|^2\right\} = \mathfrak{p} - \mathfrak{p}_{\mathrm{C}}.$$

However, notice that we have instead:

$$\begin{aligned}\hat{\mathbf{X}}^{\mathrm{O}} &= \mathbb{E}\left(\mathbf{X}^{\mathrm{O}} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right) \\ &= \mathbf{m}^{\mathrm{O}} + \mathbf{m}^{\mathrm{O|C}} + \left(\mathbf{S}^{\mathrm{O|C}}\right)^{1/2} \mathbb{E}\left(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right), \end{aligned}$$

derived by noticing that $\mathbf{X}^{\mathrm{O}} = \left(\mathbf{S}^{\mathrm{O|C}}\right)^{1/2} \mathbf{Z} + \mathbf{m}^{\mathrm{O}} + \mathbf{m}^{\mathrm{O|C}}$ and taking conditional expectations. Therefore:

$$\begin{aligned}\mathbb{E}\left\{\left(\hat{\mathbf{X}}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right)^{\intercal} \left(\mathbf{S}^{\mathrm{O|C}}\right)^{-1} \left(\hat{\mathbf{X}}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right)\right\} &= \\ = \mathbb{E}\left\{\left\|\left(\mathbf{S}^{\mathrm{O|C}}\right)^{-1/2} \left(\hat{\mathbf{X}}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O}} - \mathbf{m}^{\mathrm{O|C}}\right)\right\|^2\right\} &= \\ = \mathbb{E}\left\{\left\|\mathbb{E}\left(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right)\right\|^2\right\}. \end{aligned}$$

By the law of total variance:

$$\mathrm{cov}(\mathbf{Z}) = \mathbb{E}\left\{\mathrm{cov}(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y})\right\} + \mathrm{cov}\left\{\mathbb{E}\left(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right)\right\},$$

thus $\operatorname{cov}(\mathbf{Z}) \succeq \operatorname{cov}\left\{\mathbb{E}\left(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right)\right\}$, since $\mathbb{E}\left\{\operatorname{cov}(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y})\right\} \succeq 0$ ($\mathsf{A} \succeq \mathsf{B}$ for two matrices A, B of the same dimensions means $\mathsf{A} - \mathsf{B}$ is positive semi definite). Thus, $\operatorname{tr}\{\operatorname{cov}(\mathbf{Z})\} \geqslant \operatorname{tr}\left\{\operatorname{cov}\left\{\mathbb{E}\left(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right)\right\}\right\}$ and equivalently:

$$\mathbb{E}\left\{\|\mathbb{E}\left(\mathbf{Z} \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right)\|^2\right\} \leqslant \mathbb{E}\left\{\|\mathbf{Z}\|^2\right\} = \mathsf{p} - \mathsf{p}_{\mathsf{C}},$$

thus the result regarding the upper bound of $\mathbb{E}\left(\mathsf{D}^2 \mid \mathbf{X}^{\mathrm{C}}, \mathbf{Y}\right)$ follows. □