Echantillonnage d'itemsets à forte utilité moyenne sous contraintes de taille

Résumé. Ce document supplémentaire, présente les preuves qui ont été énoncées dans cet article intitulé "Echantillonnage d'itemsets à forte utilité sous contraintes de taille" après avoir rappelé quelques notions essentielles qui ont été aussi présentées dans l'article. Nous terminerons par ajouter des résultats expérimentaux supplémentaires dont le coût de stockage en mémoire.

1 Rappels

Définition 1 (Occurrence de motif) *Soit X un motif du langage $\mathcal{L}$ d'une base de données $\mathcal{D}$. S'il existe une transaction t_j de $\mathcal{D}$ telle que $X \subseteq t_j$, alors X_j est une occurrence du motif X dans la transaction t_j. L'utilité du motif X dans la transaction t_j, notée par $\texttt{uOcc}(X, t_j)$, est égale à 0 si $X \not\subseteq t_j$ ou $X = \emptyset$, sinon $\texttt{uOcc}(X, t_j) = \sum_{e \in X} (q(e, t_j) \times p(e))$.*

Il existe également des utilités qui sont indépendantes de toute base de données telles que celles dites fondées sur la taille (Diop et al., 2019). Dans la suite, nous considérons l'utilité fondée sur la taille définie par $\texttt{uLen}_{[m..M]}(X) = 1/|X|$ si $|X| \in [m..M]$ et 0 sinon. Ainsi, un motif dont la taille est plus grande que M ou plus petite que m sera jugé inutile.

Définition 2 (Utilité moyenne d'un motif sous contraintes de taille) *Soit $\mathcal{D}$ une base de données, $\mathcal{L}$ son langage, m et M entiers tels $m \leq M$. L'utilité moyenne du motif $X \in \mathcal{L}$ dans la base de données $\mathcal{D}$ sous contraintes de taille minimale m et maximale M, notée par $u^{moy}_{[m..M]}$, est égale au produit de la somme des utilités de ses occurrences et de son utilité fondée sur la taille. Formellement, on a : $u^{moy}_{[m..M]}(X, \mathcal{D}) = (\sum_{(j,t) \in \mathcal{D} \wedge X \subseteq t} \texttt{uOcc}(X, t)) \times \texttt{uLen}_{[m..M]}(X)$.*

Il est important de noter que $u^{moy}_{[m..M]}$ n'est pas une utilité fondée sur la taille mais la somme des utilités des occurrences d'un motif qui respecte les contraintes de taille divisée par sa taille.

2 Preuves des résultats théoriques

Propriété 1 (Calcul des $\omega^{\bullet}_{\ell}(t[i], t)$) *Le poids $\omega^{+}_{\ell}(t[i], t)$ est la somme des utilités des occurrences de motifs de taille $\ell - 1$ dans la transaction $t^i = t[i+1] \cdots t[n]$ auxquelles on adjoint l'item $t[i]$. Formellement, pour tout $\ell \in [m..M]$ on a :* [1]

1. Par convention $\binom{k}{n}$ =0 si k>n et 1 si k=0

$$\omega_\ell^+(t[i],t) = \sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(\{t[i]\} \cup X, t) = \omega_1(t[i],t) \times \binom{\ell-1}{|t^i|} + \sum_{\star \in \{+,-\}} \omega_{\ell-1}^\star(t[i+1],t)$$

$$\omega_\ell^-(t[i],t) = \sum_{X \subseteq t^i \wedge |X|=\ell} uOcc(X,t) = \sum_{\star \in \{+,-\}} \omega_\ell^\star(t[i+1],t)$$

avec $\omega_1^+(t[i],t) = uOcc(t[i],t)$ *pour tout* $i \in [1..|t|]$ *et* $\omega_\ell^\star(t[i],t) = 0$ *pour tout* $i > |t|$.

Preuve 1 (Propriété 1) *Commençons par montrer que* $\omega_\ell^-(t[i],t) = \sum_{\star \in \{+,-\}} \omega_\ell^\star(t[i+1],t)$. *Par définition,* $\omega_\ell^-(t[i],t)$ *est la somme des utilités de l'ensemble des motifs de taille* ℓ *dans* t^i, $\omega_\ell^-(t[i],t) = \sum_{X \subseteq t^i \wedge |X|=\ell} uOcc(X,t)$. *Cet ensemble peut être scindé en deux parties : celle qui contient les motifs commençant par l'item* $t[i+1]$ *dont la somme de leurs utilités est égale à* $\omega_\ell^+(t[i+1],t)$ *par définition, et celle qui contient les motifs ne commençant pas par* $t[i+1]$ *et dont la somme de leurs utilités est égale à* $\omega_\ell^-(t[i+1],t)$. *Cela implique que* $\sum_{X \subseteq t^i \wedge |X|=\ell} uOcc(X,t) = \omega_\ell^+(t[i+1],t) + \omega_\ell^-(t[i+1],t) = \sum_{\star \in \{+,-\}} \omega_\ell^\star(t[i+1],t)$. *(1)*

Montrons maintenant que $\omega_\ell^+(t[i],t) = \omega_1(t[i],t) \times \binom{\ell-1}{|t^i|} + \sum_{\star \in \{+,-\}} \omega_{\ell-1}^\star(t[i+1],t)$. *On sait par définition que* $\omega_\ell^+(t[i],t)$ *est la somme des utilités des itemsets de taille* ℓ *dans* t^i *qui commencent par* $t[i]$ *suivant la relation d'ordre total* $>_\mathcal{I}$. *Formellement, on a :* $\omega_\ell^+(t[i],t) = \sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(\{t[i]\} \cup X, t)$. *Or* $uOcc(\{t[i]\} \cup X, t) = uOcc(\{t[i]\},t) + uOcc(X,t)$ *par définition. Alors,* $\omega_\ell^+(t[i],t) = \sum_{X \subseteq t^i \wedge |X|=\ell-1} (uOcc(\{t[i]\},t) + uOcc(X,t))$. *Ce qui implique que :* $\omega_\ell^+(t[i],t) = \sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(\{t[i]\},t) + \sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(X,t)$. *Or, on sait d'une part que* $\sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(\{t[i]\},t) = uOcc(\{t[i]\},t) \times \binom{\ell-1}{|t^i|}$ *et que par définition* $uOcc(\{t[i]\},t) = \omega_1^+(t[i],t)$, *alors* $\sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(\{t[i]\},t) = \omega_1^+(t[i],t) \times \binom{\ell-1}{|t^i|}$. *D'autre part,* $\sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(X,t)$ *est la somme des utilités de l'ensemble des motifs de taille* $\ell-1$ *dans la transaction* t^i. *D'après (1), nous pouvons aussi dire que* $\sum_{X \subseteq t^i \wedge |X|=\ell-1} uOcc(X,t) = \sum_{\star \in \{+,-\}} \omega_{\ell-1}^\star(t[i+1],t)$. *D'où, on a :*
$\omega_\ell^+(t[i],t) = \omega_1(t[i],t) \times \binom{\ell-1}{|t^i|} + \sum_{\star \in \{+,-\}} \omega_{\ell-1}^\star(t[i+1],t)$. *Ce qu'il fallait démontrer.* □

Propriété 2 (Pondération d'une transaction) *Le poids de la transaction* t *sous contraintes de taille minimale* m *et maximale* M, *noté* $\omega_{[m..M]}^{umoy}(t)$, *est la somme des utilités moyennes des occurrences de motifs qu'elle contient. Formellement, on a :*

$$\omega_{[m..M]}^{umoy}(t) = \sum_{\ell=m}^{M} \left(\frac{1}{\ell} \sum_{i=1}^{|t|} \omega_\ell^+(t[i],t) \right) = \sum_{\ell=m}^{M} \frac{1}{\ell} \left(\omega_\ell^+(t[1],t) + \omega_\ell^-(t[1],t) \right).$$

Preuve 2 (Propriété 2) *Par définition, le poids de la transaction* t *est la somme des utilités moyennes des occurrences de motifs qu'elle contient. D'après la propriété 1, le poids de la transaction* t *sous les contraintes de taille minimale* m *et maximale* M *n'est rien d'autre que la somme de la somme des utilités moyennes des occurrences de motifs qui commencent par l'item* $t[1]$ *et respectent les contraintes de taille imposées,* $\sum_{\ell=m}^{M} (\frac{1}{\ell} \times \omega_\ell^+(t[1],t))$, *et de celle des motifs qui ne commencent pas par l'item* $t[1]$ *mais respectent les contraintes de taille,*

$\sum_{\ell=m}^{M}(\frac{1}{\ell} \times \omega_{\ell}^{-}(t[1],t))$. *Or on sait que*
$\sum_{\ell=m}^{M}(\frac{1}{\ell} \times \omega_{\ell}^{+}(t[1],t)) + \sum_{\ell=m}^{M}(\frac{1}{\ell} \times \omega_{\ell}^{-}(t[1],t)) = \sum_{m}^{M} \frac{1}{\ell} \times \left(\omega_{\ell}^{+}(t[1],t) + \omega_{\ell}^{-}(t[1],t)\right)$.
D'où le résultat. □

Lemme 1 *Soient ℓ la taille de l'itemset à tirer, $\mathbb{P}_{\ell}^{t}(t[i]|X,\ell')$ la probabilité de tirer l'item $t[i]$ de la transaction t après y avoir tiré $\ell - \ell'$ items et les avoir stockés dans X, avec $e >_{\mathcal{I}} t[i]$ pour tout $e \in X$. La probabilité de tirer $t[i]$ sachant X et ℓ' peut être formulée comme suit :*

$$\mathbb{P}_{\ell}^{t}(t[i]|X,\ell') = \frac{\sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(X \cup \{t[i]\} \cup X', t)}{\sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X \cup X', t)}.$$

Preuve 3 (Lemme 1) *Par définition, la probabilité de tirer l'item $t[i]$ de la transaction t après y avoir tiré $\ell - \ell'$ items et les avoir stockés dans X n'est rien d'autre que la probabilité de tirer un motif qui commence par $X \cup \{t[i]\}$, suivant la relation d'ordre $>_{\mathcal{I}}$, parmi l'ensemble des motifs qui commencent par X. D'une part, on sait que l'ensemble des motifs de taille ℓ qui commencent par $X \cup t[i]$ est défini par $\{X'' \subseteq t : (X'' = X \cup \{t[i]\} \cup X')(X' \subseteq t^i)(|X'| = \ell' - 1)\}$. La somme des utilités des motifs de cet ensemble est égale à $\sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(X \cup \{t[i]\} \cup X', t)$. D'autre part, on sait que l'ensemble des motifs de taille ℓ qui commencent par X est défini par $\{X'' \subseteq t : (X'' = X \cup X')(X' \subseteq t^{i-1})(|X'| = \ell')\}$. La somme des utilités des motifs de cet ensemble est égale à $\sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X \cup X', t)$. Donc $\mathbb{P}_{\ell}^{t}(t[i]|X,\ell') = \frac{\sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(X \cup \{t[i]\} \cup X', t)}{\sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X \cup X', t)}$. D'où le résultat.* □

Propriété 3 *La probabilité de tirer l'item $t[i]$ de la transaction t sachant l'itemset X et la taille ℓ', avec $|X| = \ell - \ell'$, notée par $\mathbb{P}_{\ell}^{t}(t[i]|X,\ell')$, est donnée par la formule suivante :*

$$\mathbb{P}_{\ell}^{t}(t[i]|X,\ell') = \frac{\left(\sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)\right) \times \binom{\ell'-1}{|t^i|} + \omega_{\ell'}^{+}(t[i],t)}{\left(\sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)\right) \times \binom{\ell'}{|t^{i-1}|} + \left(\sum_{\star \in \{+,-\}} \omega_{\ell'}^{\star}(t[i],t)\right)}.$$

La probabilité que l'item $t[i]$ ne soit pas tiré sachant X et ℓ' est égale $1 - \mathbb{P}_{\ell}^{t}(t[i]|X,\ell')$.

Les preuves de ces deux formules découlent du fait que la probabilité de tirer $t[i]$ dépend des utilités des items déjà tirés et celles des items qui le suivent pour former un motif de taille ℓ.

Preuve 4 (Propriété 3) *D'après le lemme 1, on a :*
$\mathbb{P}_{\ell}^{t}(t[i]|X,\ell') = \frac{\sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(X \cup \{t[i]\} \cup X', t)}{\sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X \cup X', t)}$.

Premièrement, on a par définition $uOcc(X \cup \{t[i]\} \cup X', t) = uOcc(X,t) + uOcc(\{t[i]\} \cup X', t)$. *Posons* $z_i = \sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(X \cup \{t[i]\} \cup X', t)$. *Cela implique que* $z_i = \sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} (uOcc(X,t) + uOcc(\{t[i]\} \cup X', t))$.
On a alors $z_i = \sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(X,t) + \sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(\{t[i]\} \cup X', t)$. *Or* $\sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(X,t) = uOcc(X,t) \times \binom{\ell'-1}{|t^i|}$ *et* $\sum_{X' \subseteq t^i \wedge |X'|=\ell'-1} uOcc(\{t[i]\} \cup X', t) = \omega_{\ell'}^{+}(t[i],t)$ *par définition. Alors* $z_i = uOcc(X,t) \times \binom{\ell'-1}{|t^i|} + \omega_{\ell'}^{+}(t[i],t)$. *On sait aussi*

que $uOcc(X,t) = \sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)$. *Donc on a :*
$z_i = \left(\sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)\right) \times \binom{\ell'-1}{|t^i|} + \omega^+_{\ell'}(t[i],t)$.

Deuxièmement, on a $uOcc(X \cup X',t) = uOcc(X,t) + uOcc(X',t)$. *En posant* $Z_i = \sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X \cup X',t)$, *on obtient alors* $Z_i = \sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X,t) + \sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X',t)$. *Or* $\sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X,t) = uOcc(X,t) \times \binom{\ell'}{|t^{i-1}|} = \left(\sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)\right) \times \binom{\ell'}{|t^{i-1}|}$ *et* $\sum_{X' \subseteq t^{i-1} \wedge |X'|=\ell'} uOcc(X',t) = \sum_{\star \in \{+,-\}} \omega^\star_{\ell'}(t[i],t)$, *donc* $Z_i = \left(\sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)\right) \times \binom{\ell'}{|t^{i-1}|} + \sum_{\star \in \{+,-\}} \omega^\star_{\ell'}(t[i],t)$.

Finalement, $\mathbb{P}^t_\ell(t[i]|X,\ell') = \frac{z_i}{Z_i} = \frac{\left(\sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)\right) \times \binom{\ell'-1}{|t^i|} + \omega^+_{\ell'}(t[i],t)}{\left(\sum_{k<i \wedge t[k] \in X} \omega_1(t[k],t)\right) \times \binom{\ell'}{|t^{i-1}|} + \sum_{\star \in \{+,-\}} \omega^\star_{\ell'}(t[i],t)}$. □

Propriété 4 (Correction) *Soient $\mathcal{D}$ une base de données transactionnelles ayant des utilités sur les items et munie d'une relation d'ordre totale $>_\mathcal{I}$, des contraintes de taille minimale m et maximale M.* HISAMPLER *tire aléatoirement un motif X du langage $\mathcal{L}(\mathcal{D})$ avec une probabilité égale à $u^{moy}_{[m..M]}(X,\mathcal{D})/Z$ où Z est une constante de normalisation.*

Preuve 5 (Propriété 4) *Soient les contraintes de taille minimale m et maximale M, la probabilité de tirer le motif X de taille ℓ dans la base de données $\mathcal{D}$ notée par $\mathbb{P}_{[m..M]}(X,\mathcal{D})$, et Z une constante de normalisation définie par $Z = \sum_{X' \in \mathcal{L}(\mathcal{D})} u^{moy}_{[m..M]}(X',\mathcal{D})$. On sait que :*
$\mathbb{P}_{[m..M]}(X,\mathcal{D}) = \sum_{(j,t) \in \mathcal{D}} \left(\mathbb{P}_{[m..M]}(t_j,\mathcal{D}) \times \mathbb{P}_{[m..M]}(X,t_j)\right)$. *Or* $\mathbb{P}_{[m..M]}(t_j,\mathcal{D}) = \frac{\omega^{umoy}_{[m..M]}(t_j)}{Z}$, *alors* $\mathbb{P}_{[m..M]}(X,\mathcal{D}) = \sum_{(j,t) \in \mathcal{D}} \left(\frac{\omega^{umoy}_{[m..M]}(t_j)}{Z} \times \mathbb{P}_{[m..M]}(X,t_j)\right)$. (1)
On sait aussi que : $\mathbb{P}_{[m..M]}(X,t_j) = \mathbb{P}_{[m..M]}(\ell|t_j) \times \mathbb{P}^{t_j}_{[m..M]}(X|\ell)$. (2)
Or on a : $\mathbb{P}_{[m..M]}(\ell|t_j) = \frac{\omega^{umoy}_{[\ell..\ell]}(t_j)}{\omega^{umoy}_{[m..M]}(t_j)}$ *et* $\mathbb{P}^{t_j}_{[m..M]}(X|\ell) = \frac{uOcc(X,t_j)}{\omega^{umoy}_{[\ell..\ell]}(t_j) \times \ell}$ *alors en substituant les deux termes en (2) on obtient* $\mathbb{P}_{[m..M]}(X,t_j) = \frac{\omega^{umoy}_{[\ell..\ell]}(t_j)}{\omega^{umoy}_{[m..M]}(t_j)} \times \frac{uOcc(X,t_j)}{\omega^{umoy}_{[\ell..\ell]}(t_j) \times \ell} = \frac{uOcc(X,t_j)}{\omega^{umoy}_{[m..M]}(t_j) \times \ell}$.
Si on remplace maintenant $\mathbb{P}_{[m..M]}(X,t_j)$ *dans (1) par sa dernière expression, on obtient :* $\mathbb{P}_{[m..M]}(X,\mathcal{D}) = \sum_{(j,t) \in \mathcal{D}} \left(\frac{\omega^{umoy}_{[m..M]}(t_j)}{Z} \times \frac{uOcc(X,t_j)}{\omega^{umoy}_{[m..M]}(t_j) \times \ell}\right) = \frac{1}{Z} \times \frac{\sum_{(j,t) \in \mathcal{D}} uOcc(X,t_j)}{\ell}$. *Or par définition on a* $\frac{\sum_{(j,t) \in \mathcal{D}} uOcc(X,t_j)}{\ell} = u^{moy}_{[m..M]}(X,\mathcal{D})$, *donc* $\mathbb{P}_{[m..M]}(X,\mathcal{D}) = \frac{u^{moy}_{[m..M]}(X,\mathcal{D})}{Z}$.
Ce qu'il fallait démontrer. □

3 Expérimentations additionnelles

Coût de stockage mémoire. Certains peuvent s'interroger sur le coût de stockage mémoire de notre méthode puisqu'elle ajoute des informations supplémentaires sur les items de chaque transaction et garde en mémoire les valeurs combinatoires $\binom{k}{n}$. Le tableau 1 montre quelques statistiques calculées avec le framework "asizeof" [2]. Comme attendu, le coût de stockage mémoire augmente en fonction de la contrainte de taille maximale M, mais quel que soit le jeu de données, il reste inférieur à $1Go$ sur tous nos jeux de données avec $M = 8$ (maximum 616,793

2. https://code.activestate.com/recipes/546530-size-of-python-objects-revised/

TAB. 1 – *Coût de stockage mémoire en Mega Octet (Mo) de* HISAMPLER *avec* $M \in \{5, 7, 8\}$

	M		
$\mathcal{D}$	5	7	8
Adult	434.629	483.499	504.021
BMS	82.681	85.658	86.822
Chess	109.411	113.906	113.906

	M		
$\mathcal{D}$	5	7	8
Foodmart	10.051	10.256	10.271
Mushroom	113.593	128.513	135.323
Retail	542.965	595.098	616.793

Mo avec Retail). Cela signifie que l'approche de pondération de HISAMPLER n'est pas coûteuse en stockage. Il peut donc être utilisé avec une base de données plus volumineuse pour échantillonner des itemsets à utilité moyenne élevée.

Références

Diop, L., C. T. Diop, A. Giacometti, D. Li, et A. Soulet (2019). Sequential pattern sampling with norm-based utility. *Knowledge and Information Systems*.

Summary

This supplementary document presents the proofs which were stated in this article titled "High Average-utility Itemsets Sampling under Length Constraints (Echantillonnage d'itemsets à forte utilité moyenne sous contraintes de taille)".

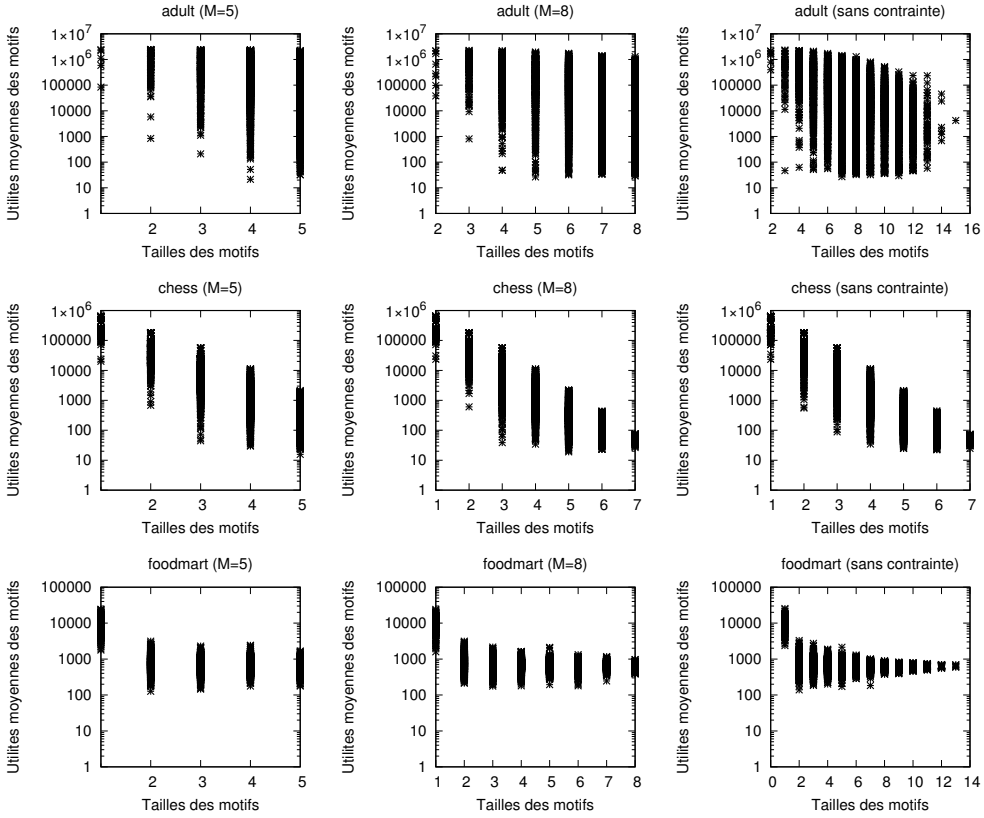

FIG. 1 – *Dispersion des utilités moyennes de 10,000 motifs échantillonnés*