Utilisation de processus ponctuels déterminantaux pour la sélection de parents dans un algorithme génétique : application à l'aide multi-critère à la décision

Miguel GUZMAN[1,3] Kadriye Nur BAKIRCI[1] Martin ROUESNE[1] Hélène SAVATIER-DUPRÉ[1] Bastien PASDELOUP[2] Patrick MEYER[2]

[1]IMT Atlantique, 665 avenue du Technopôle, 29280 Plouzané, France

[2]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France

[3]COQCYT, 210 avenue Insurgentes, 77015 Chetumal, Mexique

Résumé – L'aide multi-critère à la décision est un domaine de recherche visant à fournir à des décideurs des outils permettant d'offrir une assistance à la décision à partir de données multivariées. Parmi ces outils, le modèle *Ranking with Multiple Profiles* (RMP) permet d'ordonner un ensemble d'alternatives, en les comparant deux à deux à des profils de référence. La configuration optimale d'un tel modèle peut se montrer complexe au vu des multiples paramètres mis en jeu. Des travaux ont montré l'intérêt d'algorithmes génétiques pour la détermination des paramètres de ce modèle. Ces algorithmes visent à faire évoluer une population de solutions par une suite d'opérateurs de croisement et de mutation entre des solutions parentes, pour produire des solutions enfants. Toutefois, les algorithmes génétiques ont tendance à converger vers une population de solutions très homogène. Dans ce contexte, nous avons choisi d'étudier l'apport des *Processus Ponctuels Déterminantaux* (DPP) dans la sélection des parents. Ces modèles probabilistes permettent un échantillonnage aléatoire au sein de la population de solutions, avec des propriétés de diversité entre les solutions choisies.

Abstract – Multi-criteria decision aid is a research field aimed at providing decision-makers with tools to assist in decision-making based on multivariate data. Among these tools, the *Ranking with Multiple Profiles* (RMP) model allows for ordering a set of alternatives by pairwise comparison to reference profiles. The optimal configuration of such a model can be complex due to multiple parameters involved. Research has shown the usefulness of genetic algorithms in determining the parameters of this model. These algorithms aim to evolve a population of solutions through a series of crossover and mutation operators between parent solutions to produce child solutions. However, genetic algorithms tend to converge to a highly homogeneous population of solutions. In this context, we chose to study the contribution of *Determinantal Point Processes* (DPP) in parent selection. These probabilistic models allow for random sampling within the population of solutions, with diversity properties among the selected solutions.

1 Introduction

Dans ce travail, nous nous intéressons à la sélection des parents au sein d'un algorithme génétique, et plus précisément à une technique d'échantillonnage par *Processus Ponctuels Déterminantaux* (DPP). En détails, les algorithmes génétiques sont des algorithmes approchés de recherche de solutions à des problèmes variés, faisant évoluer une population suivant une procédure de sélection, croisements et mutations. Chaque individu se voit attribuer une valeur de *justesse* (adéquation à la fonction objectif à optimiser), et les individus les plus *aptes* ont plus de chances de se voir sélectionnés afin d'être mélangés pour générer des enfants d'autant plus adaptés au problème. La sélection des solutions à reproduire est donc un élément clé, car mélanger des solutions trop homogènes amènera à une convergence trop rapide vers un optimum local.

Des solutions existent dans la littérature pour conserver une diversité dans la population, en jouant sur les opérateurs de croisement, de mutation et de sélection [6]. La stratégie la plus usuelle est la sélection par roulette, qui permet la sélection de parents avec une probabilité proportionnelle à la justesse de la solution. Les DPP sont d'autres modèles d'échantillonnage aléatoire qui permettent une sélection d'un sous-ensemble de points, telle que les points sélectionnés satisfont un critère de diversité. Dans quelques études récentes. [8, 9], les DPP ont été introduits en tant que stratégie pour construire des pools d'accouplement divers dans des algorithmes évolutionnaires pour des problèmes d'optimisation multi-objectifs.

Nous nous intéressons à la question de l'impact des DPP sur la diversification et les performances d'un algorithme génétique ayant comme cadre d'étude celui présenté en [2] visant à produire un modèle *Ranking with Multiple Profiles* (RMP). Un tel modèle est un outil d'aide à la décision permettant d'ordonner des alternatives sur la base de profils de référence. Une solution de l'algorithme génétique est donc un modèle RMP tel qu'illustré en Figure 1, ici composé de $C = 4$ critères, chacun pondéré par un poids $\mathbf{w} \in \mathbb{R}^C$, ainsi que par $P = 3$ profils de référence décrits par une matrice $\mathbf{P} \in \mathbb{R}^{P \times C}$ et un ordre, *i.e.*, une permutation $\mathbf{o}$ des profils.

Cet article est organisé comme suit. En Section 2, nous introduisons plus en détails l'algorithme génétique considéré, ainsi que les DPPs. En Section 3, nous proposons une définition de noyau pour les DPPs adapté à notre cas d'étude. Enfin, en Section 4, nous réalisons des expériences pour évaluer l'impact de l'échantillonnage par DPP sur la diversité de la population et sur les performances. La Section 5 est une conclusion.

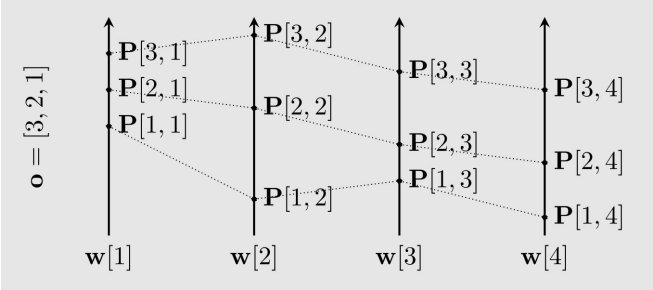

FIGURE 1 : Exemple d'un modèle RMP solution de l'algorithme génétique considéré.

2 Contexte

2.1 Algorithme génétique

La structure globale de l'algorithme génétique considéré est présentée en Figure 2. Au sein de cet algorithme, nous allons modifier la partie correspondant à la sélection des parents, en changeant la sélection par roulette – proportionnelle à la justesse de la solution – par un échantillonnage par DPP, dont le noyau est présenté plus bas.

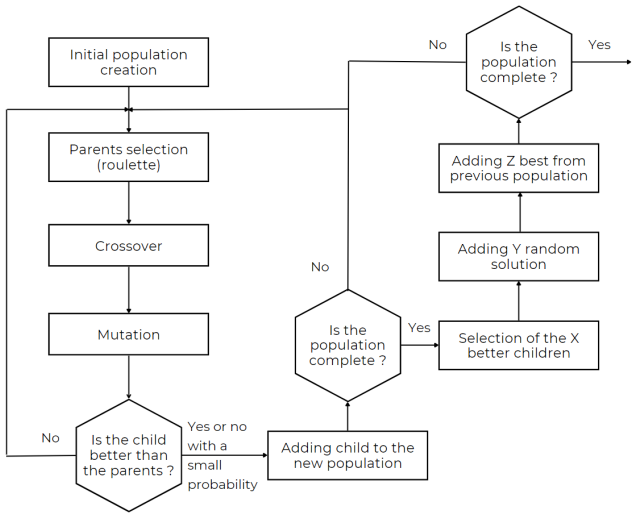

FIGURE 2 : Schéma de l'algorithme génétique de l'étude.

La définition de cet algorithme est standard. A chaque itération, une nouvelle population est initialisée à partir des Z meilleures solutions de l'itération précédente, ainsi que par Y nouvelles solutions aléatoires. Les X solutions manquantes pour arriver à la taille d'origine de la population sont obtenues par suites de sélections, croisements et mutations de solutions de l'itération précédente. L'algorithme termine quand un pourcentage des meilleures solutions n'évolue plus pendant un nombre fixé d'itérations.

Nous nous basons sur l'algorithme introduit en [2], ainsi que sur le code public associé. Les hyperparamètres de l'algorithme génétique sont ceux définis dans ces travaux, à l'exception du nombre de solutions aléatoires injectées à chaque itération, que nous fixons à 0. En effet, étant donné que nous nous intéressons à l'apport des DPP pour la diversité de la population de l'algorithme génétique, nous avons choisi de retirer cette source de confusion.

2.2 Processus Ponctuels Déterminantaux

Les processus ponctuels déterminantaux (DPP), introduits par Macchi en 1975 [6], sont des modèles probabilistes qui exploitent les corrélations négatives via une mesure de similarité. Les DPP sont des modèles d'échantillonnage aléatoire recommandés pour les problèmes nécessitant des ensembles de points diversifiés, car intégrant un critère de dissimilarité entre les points. En outre, les DPP se concentrent sur la taille et le contenu de l'ensemble, qui sont deux caractéristiques différentes. Les k-DPP conditionnent les DPP en imposant une taille fixe k à l'ensemble modélisé. Contrairement aux DPP, qui capturent à la fois la taille et le contexte de l'ensemble, les k-DPP se concentrent uniquement sur le contenu de l'ensemble [4].

L'objectif d'un k-DPP est de sélectionner un sous-ensemble de k éléments d'un ensemble plus large d'éléments avec une diversité maximale tout en restant représentatif de l'ensemble d'origine. L'échantillonnage par k-DPP fonctionne en sélectionnant de manière itérative des éléments qui sont à la fois représentatifs et diversifiés, sur la base d'une fonction de similarité. Il élimine ensuite les éléments qui sont trop similaires aux éléments sélectionnés afin de garantir la diversité.

Un modèle probabiliste k-DPP prend en entrée une matrice exprimant les relations de similarité entre les points de l'espace à échantillonner. La construction de ce noyau est un facteur important de la performance d'un k-DPP. Nous proposons dans la section suivante deux noyaux possibles pour le problème considéré.

Une fois un noyau défini, un modèle d'échantillonnage par k-DPP permet la sélection de k individus à chaque routine de sélection des parents pour créer des solutions enfants. Comme dans l'article de référence, nous considérons ici le cas $k = 2$.

3 Sélection des parents par DPP

Comme indiqué précédemment, l'échantillonnage par k-DPP nécessite la construction d'un noyau $\mathbf{S}$ mesurant la similarité entre les points. Dans nos expérimentations, nous considérons deux noyaux :

- L'un basé sur une mesure de distance (norme ℓ_1) ;
- L'autre basé sur la corrélation entre deux ordres de préférence (Tau de Kendall).

Ces mesures de similarité alternatives sont étudiées dans les sous-sections suivantes. De plus, une méthode combinant l'échantillonnage par DPP et l'approche roulette classique est présentée comme une proposition visant à améliorer la diversité (DPP) tout en conservant un critère basé sur la justesse (roulette). D'autres méthodes existent pour atteindre cet objectif, qui seront explorées dans des travaux futurs.

3.1 Noyau basé sur une mesure de distance

Pour construire une similarité basée sur la distance, nous proposons tout d'abord une approche simple basée sur une utilisation de la norme ℓ_1, mesurant la somme des différences absolues des composantes des solutions. Une solution de l'algorithme génétique est ici vue comme une vectorisation des différentes variables la décrivant ($\mathbf{P}$, $\mathbf{w}$ et $\mathbf{o}$ en Figure 1).

Soient $\mathbf{x}_i = \text{vec}(\mathbf{P}_i, \mathbf{w}_i)$ et $\mathbf{x}_j = \text{vec}(\mathbf{P}_j, \mathbf{w}_j)$ deux solutions dans la population de l'algorithme génétique. La matrice de similarité $\mathbf{S}[i,j]$ est définie par :

$$\forall i,j : \mathbf{S}[i,j] = 1 - \|\mathbf{x}_i - \mathbf{x}_j\|_1 \,. \tag{1}$$

3.2 Noyau basé sur la corrélation entre deux ordres de préférence

Comme mentionné précédemment, un modèle RMP permet d'ordonner des alternatives selon les préférences d'un décideur, représentées par les paramètres du modèle.

Une seconde mesure de similarité, alternative à la norme ℓ_1, peut être basée sur la sortie d'un tel modèle RMP, c'est à dire basée sur une corrélation entre deux ordres de préférences. Soit A un ensemble fini d'alternatives de décision, décrites chacune par des performances sur un nombre fini de critères. Le résultat d'un modèle RMP appliqué à A peut être vu en Figure 3 sous la forme d'un vecteur de classement.

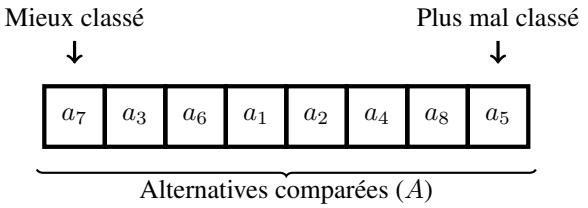

FIGURE 3 : Vecteur d'alternatives classées pour une configuration d'un modèle RMP

Soient $\mathbf{R}_i = \text{vec}(\mathbf{a}_i, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_n)$ où $\mathbf{a}_i \succ \mathbf{a}_{i+1} \succ \cdots \succ \mathbf{a}_n$ et $\mathbf{R}_j = \text{vec}(\mathbf{a}_j, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_n)$ où $\mathbf{a}_j \succ \mathbf{a}_{j+1} \succ \cdots \succ \mathbf{a}_n$ deux vecteurs de classement. La matrice de similarité $\mathbf{S}[i,j]$ est définie par :

$$\forall i,j : \mathbf{S}[i,j] = \frac{\tau_B(\mathbf{R}_i, \mathbf{R}_j) + 1}{2} \,. \tag{2}$$

où τ_B est le coefficient de corrélation Tau B de Kendall calculé par l'algorithme de Knight [3].

3.3 Combiner DPP et roulette

Jusqu'ici, l'utilisation des DPP est introduite pour créer des descendants plus diversifiés, sans tenir compte de la justesse des solutions parents. Nous proposons ici des débuts d'analyses mêlant roulette et DPP. Pour ce faire, nous créons un pool d'accouplement diversifié en échantillonnant un sous ensemble de la population par k-DPP. Ensuite, une sélection par roulette est appliquée pour sélectionner $k = 2$ individus dans le pool d'accouplement.

4 Expériences

Les expériences menées pour mesurer l'influence de l'insertion des DPP dans le processus de sélection des parents peuvent être divisées en deux groupes en fonction des objectifs suivants :

- Évaluer la diversité au sein de la population à chaque génération ;
- Mesurer la précision estimée du classement du décideur en fonction des données de l'ensemble de test.

Les procédures des deux expériences sont décrites dans les sous-sections suivantes[1].

4.1 Diversité de la population par génération

L'objectif de la première série d'expériences est d'analyser si l'utilisation de k-DPP peut conduire à une population plus diversifiée d'une génération à l'autre. Pour définir une mesure de diversité, nous considérons la triangulaire inférieure du noyau de similarité $\mathbf{S}$ défini par l'une des manières présentées en Section 3.

La notion de diversité est définie comme l'écart type des solutions de la population, *i.e.*,

$$std(\{\mathbf{S}[i,j], \forall i > j\}) \,. \tag{3}$$

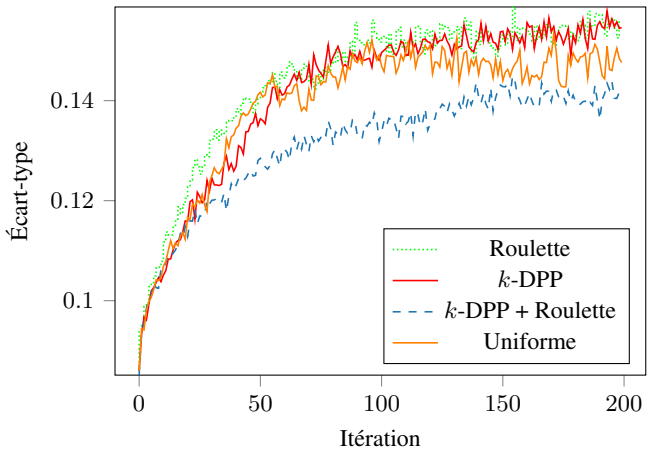

FIGURE 4 : Écart-type de la similarité par paire avec la métrique de la norme ℓ_1 par génération.

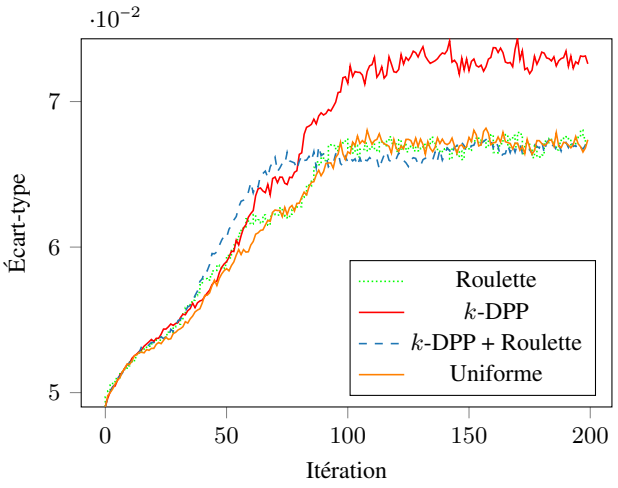

FIGURE 5 : Écart-type de la similarité par paire avec la métrique de Tau de Kendall par génération.

Les Figures 4 et 5 montrent une augmentation de la diversité au sein de la population de solutions, comparée à une stratégie

[1]Le code pour reproduire les expériences présentées dans cette section est disponible à l'adresse `https://github.com/mikeguzman1294/DPP4GA4SRMP4MCDA/tree/kDPP`.

Stratégie	Type de noyau	Entraînement	Test-1	Test-2	Temps moyen (s)
Roulette		0.991000 ± 0.003338	0.861333 ± 0.016302	0.797423 ± 0.018898	50.731250
k-DPP	Norme ℓ_1	0.992000 ± 0.005403	0.883022 ± 0.019863	0.815492 ± 0.026881	3861.821875
	Kendall-Tau	0.993000 ± 0.003969	0.877333 ± 0.020658	0.812707 ± 0.025781	24840.136684
k-DPP + Roulette	Norme ℓ_1	0.997000 ± 0.00284	0.874933 ± 0.011944	0.808350 ± 0.024154	9540.401562
	Kendall-Tau	0.997000 ± 0.00284	0.880622 ± 0.016625	0.804386 ± 0.029891	39405.92003

TABLE 1 : Résultats moyens ± 95% d'intervalle de confiance. "Entraînement" donne la capacité du modèle final à reproduire les décisions sur les paires de comparaison de l'ensemble d'entraînement, "Test-1" donne la performance de ce modèle sur les paires entre mêmes alternatives mais n'apparaissant pas dans l'ensemble d'entraînement, "Test-2" donne les performances du modèle sur des paires de nouvelles alternatives distinctes de celles de l'ensemble d'entraînement.

de sélection par roulette ou un échantillonage uniforme, notamment dans le cas de l'utilisation d'un noyau de similarité par Tau de Kendall. Une autre observation est qu'il semble peu efficace de mélanger les approches par k-DPP et roulette, comparément à leur utilisation séparée.

4.2 Performance de l'algorithme

La deuxième série d'expériences vise à comparer les performances globales de l'algorithme génétique compte tenu de l'utilisation de tous les couples possibles de stratégies d'échantillonnage et de mesures de similarité. La précision moyenne pour les ensembles d'entraînement et de test est calculée et comparée afin d'analyser les capacités de généralisation de toutes les variantes. Les résultats de ces expériences sont présentés en Table 1.

Un premier constat clair est que l'introduction de k-DPP augmente considérablement le temps de calcul nécessaire. Toutefois, l'augmentation de la variance au sein des solutions de la population de l'algorithme génétique semble accompagnée d'une augmentation de la capacité du modèle à généraliser à des ensembles de test distincts de celui d'apprentissage, sur les mêmes alternatives ou sur de nouvelles.

Il est donc intéressant de noter que l'approche semble apporter les résultats attendus, au détriment du temps de calcul. Toutefois, les méthodes utilisées [4] sont relativement anciennes et seront changées par des approches plus rapides, notamment car nous ne nécessitons pas ici d'échantillonage exact.

5 Conclusion

La diversité est cruciale pour la performance des algorithmes génétiques. Selon une mesure de similarité basée sur la corrélation entre ordres de préférence, la diversité des descendants converge vers un niveau significativement plus élevé lors de l'utilisation d'une stratégie d'échantillonnage k-DPP direct.

En termes de performance, l'introduction de k-DPP dans le processus de sélection des parents améliore la généralisation dans les ensembles de test, bien qu'à un degré faible. Cependant, il est important de considérer le compromis entre la précision et le temps d'exécution car le coût du calcul avec k-DPP est élevé. Il y a plusieurs raisons à cela. Tout d'abord, le calcul du noyau est une opération coûteuse puisque l'évaluation de la similarité entre tous les individus représente une opération fixe de $O(N^2)$ par génération. En outre, l'échantillonnage à partir d'un k-DPP nécessite globalement $O(Nk^2)$ en temps [4], de sorte que dans les cas où k est très élevé, comme la méthode combinée dans la création du pool d'accouplement, la complexité peut atteindre des valeurs proches de $O(N^3)$.

Les travaux futurs incluent l'insertion des DPP dans la sélection des survivants puisque les DPP semblent être plus performants lors de l'échantillonnage d'ensembles plus grands par rapport aux couples de parents.

Références

[1] A. AGRESTI : *Non-Model-Based Analysis of Ordinal Association*, pages 188–189. Wiley, New York, NY, 2010.

[2] Arwa KHANNOUSSI, Alexandru Liviu OLTEANU, Patrick MEYER et Bastien PASDELOUP : A metaheuristic for inferring a ranking model based on multiple reference profiles. preprint, `https://imt-atlantique.hal.science/hal-04017642`, mars 2023.

[3] W.R. KNIGHT : A computer method for calculating kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61(314):436–439, 1966.

[4] Alex KULESZA et Ben TASKAR : k-dpps : Fixed-size determinantal point processes. *In Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1193–1200, 2011.

[5] Alex KULESZA et Ben TASKAR : Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[6] Odile MACCHI : The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

[7] Brian MCGINLEY, John MAHER, Colm O'RIORDAN et Fearghal MORGAN : Maintaining healthy population diversity using adaptive crossover, mutation, and selection. *IEEE Transactions on Evolutionary Computation*, 15(5):692–714, 2011.

[8] Michael OKOTH, Ronghua SHANG, Licheng JIAO, Jehangir ARSHAD, Ateeq REHMAN et Habib HAMAM : A large scale evolutionary algorithm using determinantal point processes for large scale multi-objective optimization problems. *Electronics*, 11:3317, 10 2022.

[9] Peng ZHANG, Jinlong LI, Tengfei LI et Huanhuan CHEN : A new many-objective evolutionary algorithm based on determinantal point processes. *IEEE Transactions on Evolutionary Computation*, PP:1–1, 11 2020.