Modèles probabilistes pour les permutations et dépendances

Arthur Fétiveau [1] & Gilles Durrieu [2] & Emmanuel Frénod [3]

[1] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , arthur.fetiveau@univ-ubs.fr*
[2] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , gilles.durrieu@univ-ubs.fr*
[3] *Univ Bretagne Sud, CNRS UMR 6205, LMBA, France , emmanuel.frenod@univ-ubs.fr*

Résumé. Dans l'objectif d'améliorer la performance financière d'une entreprise, nous devons classer tous les produits qu'elle commercialise selon leurs intérêts en utilisant des critères financiers. Les critères et le classement final pouvant se modéliser sous formes de permutations, nous nous intéressons aux modèles de Mallows qui sont des modèles définis dans l'espace des permutations. Nous nous intéressons notamment à la question de la dépendance dans les modèles de Mallows. Dans cette optique, nous introduisons une approche basée sur une fonction de coût minimisant l'impact de la dépendance entre les critères.

Mots-clés. Apprentissage, dépendance, distance de Kendall, modèles de Mallows, statistique computationnelle.

Abstract. In order to improve the financial performance of a company, we must classify all their commercialised products according to their interests using financial criteria. The criteria and the final classification can be modeled using permutations. We consider Mallows models defined in the space of permutations. We are particularly interested in the question of dependence in Mallows models. Here, we introduce an approach based on a cost function minimizing the impact of the dependence between criteria.

Keywords. Machine learning, dependence, Kendall distance, Mallows models, computational statistics.

1 Introduction

Dans une entreprise disposant de nombreux produits, il n'est pas possible d'analyser dans le détail chacun d'entre eux. Il y aurait trop de produits à analyser. Ainsi, il est primordial d'avoir un outil permettant, selon les préférences du décideur, de prioriser les éléments les plus importants pour le décideur. Différentes approches ont été abordées pour faire de l'agrégation multi-critère, telles que la méthode MAUT (Multi Attribute Utility Theory) notamment explorée par Keeney et Raiffa (1976) ou encore les différentes méthodes ELECTRE (ELimination Et Choix Traduisant la REalité), expliquée par Roy (1968), et PROMETHEE (Preference Ranking Organisation METHod for Enrichment Evaluations) de Mareschal, Brans, et Vincke (1984).

Afin de nous aider à décider, nous disposons de critères basiques tels que le chiffre d'affaires ou la prévision de la quantité de produits à vendre. Nous pouvons également créer de nombreux critères dérivés générant des informations différentes comme par exemple l'évolution de l'écart entre la quantité vendue et la quantité prévue de ventes. Pour modéliser notre problème, nous utilisons un modèle de Mallows (Mallows (1957); Fétiveau (2024)) sur l'ensemble des permutations. Cependant dans cette modélisation, les critères sont supposés indépendants entre eux, ce qui n'est pas le cas en pratique. Nous ne pouvons pas non plus décider de n'utiliser que les critères utiles puisque nous n'avons aucune idée de si certains critères dérivés apportent des informations nouvelles. Par conséquent, nous introduisons une fonction de coût à optimiser permettant de trouver les valeurs de θ dans le cas de variables dépendantes. Nous testons les résultats sur des données simulées suivant le modèle de Mallows.

2 Modèle de Mallows et estimation des paramètres

Soit une suite de n éléments à ordonner. On appelle une permutation π, une bijection de $\{1, ..., n\}$ dans $\{1, ..., n\}$. On note $\pi(i)$ le rang associé à l'élément i et $\pi^{-1}(i)$ l'élément associé au rang i de π pour $i \in \{1, ..., n\}$. On note $\mathcal{S}_n$ l'ensemble de toutes les permutations possibles de n éléments. On a $\#\mathcal{S}_n = n!$, où $\#$ est le cardinal.

Le modèle de Mallows (1957) est défini comme une distribution de paramètres π et $\theta \in \mathbb{R}$. La probabilité d'une permutation σ est donnée par

$$P_\theta(\sigma) = \frac{exp(-\theta\ d(\sigma, \pi))}{Z(\theta)} \tag{1}$$

où π est la permutation modale représentant le véritable classement inconnu, $\theta \in \mathbb{R}$ est le paramètre de dispersion autour du classement modal, $d(.,.)$ est une distance invariante à la relabélisation et $Z(\theta)$ le terme de normalisation. Ce dernier terme ne dépend pas de π lorsque la distance utilisée possède la propriété d'invariance à la relabélisation. Quand $\theta = 0$, alors chaque permutation σ de $\mathcal{S}_n$ est équiprobable. En particulier quand θ est positif, la probabilité est concentrée autour de la permutation modale et quand θ est négatif, la probabilité est concentrée autour de la permutation antimodale.

Dans cet article, on utilise la distance de Kendall (1938) donnée par

$$\begin{aligned} d_k(\pi, \sigma) \quad &= \sum_{i=1}^{n-1} \sum_{j>i} \mathbb{1}\left([\pi(i) > \pi(j) \wedge \sigma(i) < \sigma(j)] \vee \right. \\ &\left. [\pi(i) < \pi(j) \wedge \sigma(i) > \sigma(j)]\right), \end{aligned} \tag{2}$$

qui est bien une distance invariante à la relabélisation (Diaconis (1988)). Cette distance bénéficie également d'une propriété importante puisqu'il est possible de la décomposer en une somme finie de termes indépendants. La distance de Kendall détermine le nombre minimal de transpositions adjacentes pour passer d'une permutation à l'autre.

Nous considérons pour nos objectifs que la permutation π représente le classement idéal, celui que nous devons trouver par consensus, c'est à dire via un accord le plus satisfaisant pour tous, entre les différents critères. Pour tout $j \in \{1, ..., J\}$, chacun de nos critères est représenté par σ_j et est issu du modèle de Mallows de paramètres π et θ_j (1). Pour tout $j \in \{1, ..., J\}$, les estimateurs des paramètres θ_j sachant les valeurs passées de π et σ_j sont déterminés par la méthode du maximum de vraisemblance et puisque les θ_j sont invariants dans le temps, nous estimons aussi par la méthode du maximum de vraisemblance la valeur de π sachant les σ_j et les θ_j.

Pour la distance de Kendall (2), nous avons montré dans Fétiveau et al. (2024) que l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ est solution de l'équation suivante, voir aussi Fligner et Verducci (1986) :

$$\frac{n\exp(-\hat{\theta})}{1-\exp(-\hat{\theta})} - \sum_{k=1}^{n} \frac{k\exp(-k\hat{\theta})}{1-\exp(-k\hat{\theta})} = \frac{1}{T}\sum_{t=1}^{T} d(\sigma_t, \pi_t), \tag{3}$$

où $t \in \{1, \ldots, T\}$ représente une situation financière des produits de l'entreprise à un instant t et n la taille des permutations. Sachant que la distance de Kendall est invariante à la relabélisation, il n'est pas nécessaire que tous les éléments π_t soient les mêmes aux instants t. Puisque nos données sont collectées à intervalle de temps régulier, il est important que la permutation π_t n'impacte pas les résultats.

Le comportement asymptotique de l'estimateur $\hat{\theta}$ de θ est donné par le Théorème 1 ci-dessous.

Théorème 1 *Soit le modèle $(\mathcal{S}_n, \{P_\theta\}_{\theta\in\Theta})$ un modèle régulier où $\mathcal{S}_n$ est l'espace des permutations tel que pour chaque $\theta \in \Theta$, il existe un voisinage $V \subset \Theta$ de θ pour lequel $\sup_{\alpha\in V} ||\nabla^2 \ln L(.;\alpha)|| \in \mathbb{L}^1(P_\theta)$. Puisque l'estimateur du maximum de vraisemblance $\hat{\theta}$ de θ est consistant alors*

$$\sqrt{T}\left(\hat{\theta} - \theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta)^{-1}\right) \tag{4}$$

où

$$I(\theta) = \frac{n\exp(-\theta)}{(1-\exp(-\theta))^2} - \sum_{k=1}^{n} \frac{k^2\exp(-k\theta)}{(1-\exp(-k\theta))^2}. \tag{5}$$

L'estimateur de la variance asymptotique de $\hat{\theta}$ est donné par

$$I(\hat{\theta})^{-1} = \left(\frac{n\exp(-\hat{\theta})}{(1-\exp(-\hat{\theta}))^2} - \sum_{k=1}^{n} \frac{k^2\exp(-k\hat{\theta})}{(1-\exp(-k\hat{\theta}))^2}\right)^{-1}. \tag{6}$$

Pour chaque $j \in \{1, ..., J\}$, σ_j est connu et son paramètre de dispersion θ_j est estimé par l'estimateur du maximum de vraisemblance $\hat{\theta}_j$. Nous estimons π par la méthode du maximum de vraisemblance et nous notons $\hat{\pi}$ son estimateur. En supposant les σ_j indépendants, la vraisemblance s'écrit :

$$L(\pi|\sigma_1,\ldots,\sigma_J,\hat{\theta}_1,\ldots,\hat{\theta}_J) = \prod_{j=1}^{J} \frac{\exp(-\hat{\theta}_j\ d(\sigma_j,\pi))}{Z(\hat{\theta}_j)} \tag{7}$$

où J est le nombre de critères. Puisque $Z(\hat{\theta}_j)$ ne dépend pas de π et que l'exponentielle est une fonction monotone, l'estimateur du maximum de vraisemblance $\hat{\pi}$ de π est :

$$\hat{\pi} = \underset{\pi\in\mathcal{S}_n}{\operatorname{argmax}} \sum_{j=1}^{J} -\hat{\theta}_j\ d(\sigma_j,\pi) \tag{8}$$

où $\mathcal{S}_n$ est l'ensemble de toutes les permutations possibles de n éléments.

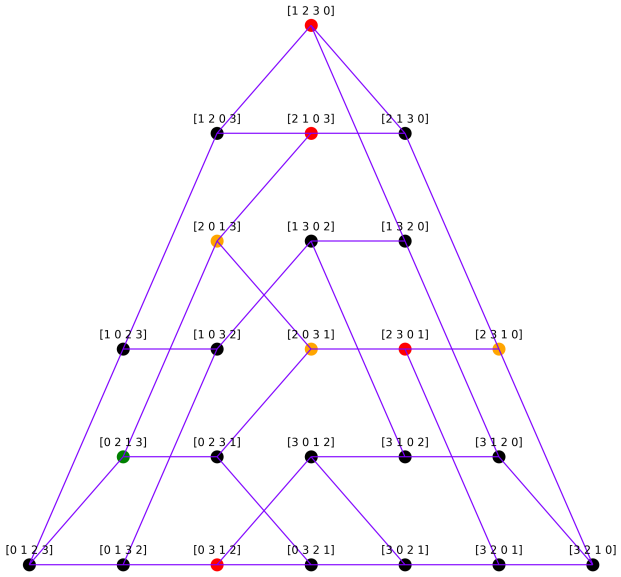

Figure 1: Choix de représentation de trois exemples de consensus en orange pour des critères en rouge et une vérité en vert. Chaque lien de la représentation graphique indique une distance de Kendall de 1 entre deux permutations.

La Figure 1 montre pour des permutations de taille 4, des exemples de consensus $\hat{\pi}$ en orange pour 4 critères, ayant chacun une valeur de θ fixée arbitrairement à la valeur 1. Les permutations σ_j représentant les classements de ces critères sont affichées en rouge. Tous ces consensus n'ont pas la même proximité avec la permutation π que l'on cherche à estimer (en vert). La distance de Kendall entre deux permutations peut se déterminer en comptant le nombre de liens minimums pour passer de l'une à l'autre.

3 Simulation du modèle de Mallows

Puisque la distance de Kendall est décomposable en une somme finie de termes indépendants, on peut l'utiliser pour simuler des permutations σ pour π et θ connus.

En utilisant la distance de Kendall avec le modèle de Mallows ou de Mallows généralisé, Fligner et Verducci (1986) et Fétiveau et al. (2024), il est possible de générer une permutation via le produit des probabilités de déplacer chaque élément de π avec les éléments lui succédant. Pour chaque élément de π, le déplacement d'un élément avec un nombre d'éléments lui succédant est indépendant du mouvement des éléments le précédant. On note $\zeta(\pi^{-1}(i))$ le nombre de déplacements du i-ème élément de la permutation π avec les éléments le succédant. La probabilité de déplacer un élément d'un nombre $x \leq (n-i)$ d'éléments lui succédant est donnée par

$$\mathbb{P}_\theta\left(\zeta(\pi^{-1}(i)) = x\right) = \frac{\exp(-\theta\, x)}{\sum_{k=0}^{n-i} \exp(-\theta\, k)}. \tag{9}$$

On peut réécrire le dénominateur avec la propriété de la somme d'une suite géométrique.

En utilisant (9), nous représentons dans la Figure 2 un ensemble de 10 000 permutations générées autour de la permutation identité $\pi = \mathrm{Id} = [0\ 1\ 2\ 3\ 4]$ pour $\theta = 0.5$ en considérant un modèle de Mallows à $n = 5$ éléments. La couleur sur ce graphique représente la fréquence d'apparition d'une permutation.

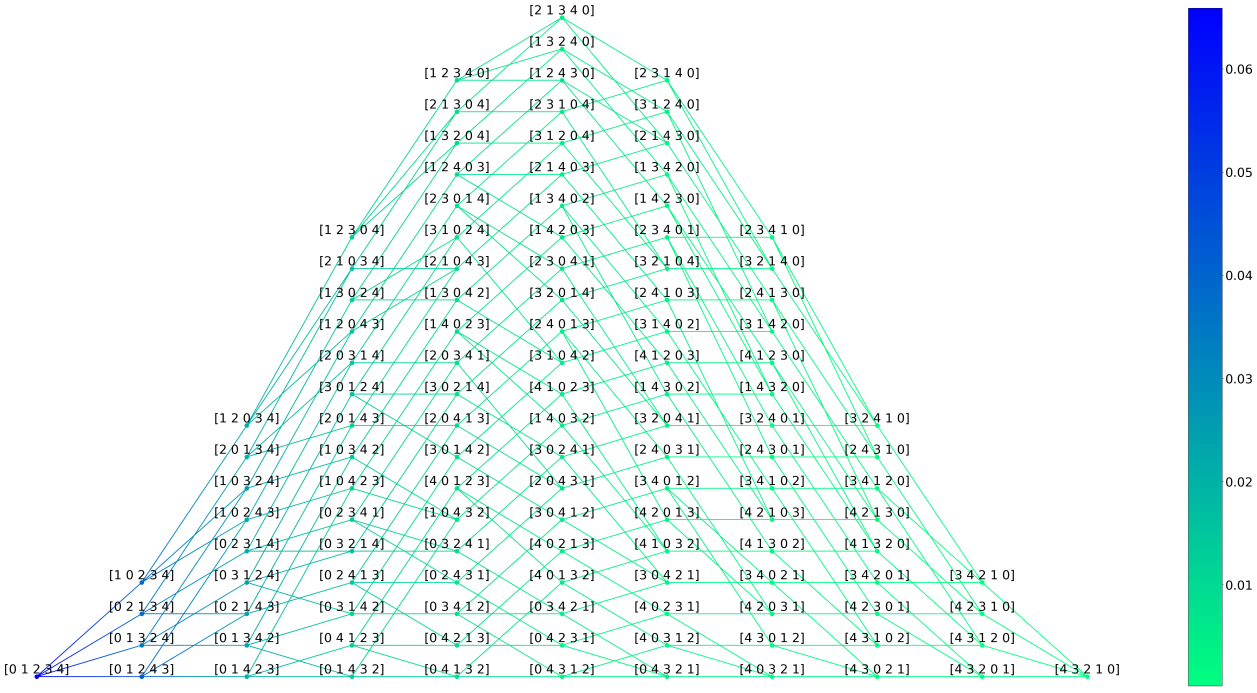

Figure 2: Simulation de 10 000 permutations par le modèle de Mallows. Chaque lien indique une distance de Kendall de 1 entre deux permutations. Plus les permutations sont bleues et plus la fréquence d'apparition est élevée. Plus elles sont vertes et plus la fréquence d'apparition est faible.

La Figure 3 représente une simulation de 10 000 permutations selon un modèle de Mallows généralisé ayant pour vecteur de paramètres $\boldsymbol{\theta} = (0.1, 0.5, 1, 1)$. On observe notamment la forte probabilité de mouvement du premier élément en observant la ligne bleue qui passe sur

les permutations où l'élément 0 se déplace, c'est à dire le premier élément de la permutation initiale $\pi = \mathrm{Id} = [0\ 1\ 2\ 3\ 4]$ se déplace.

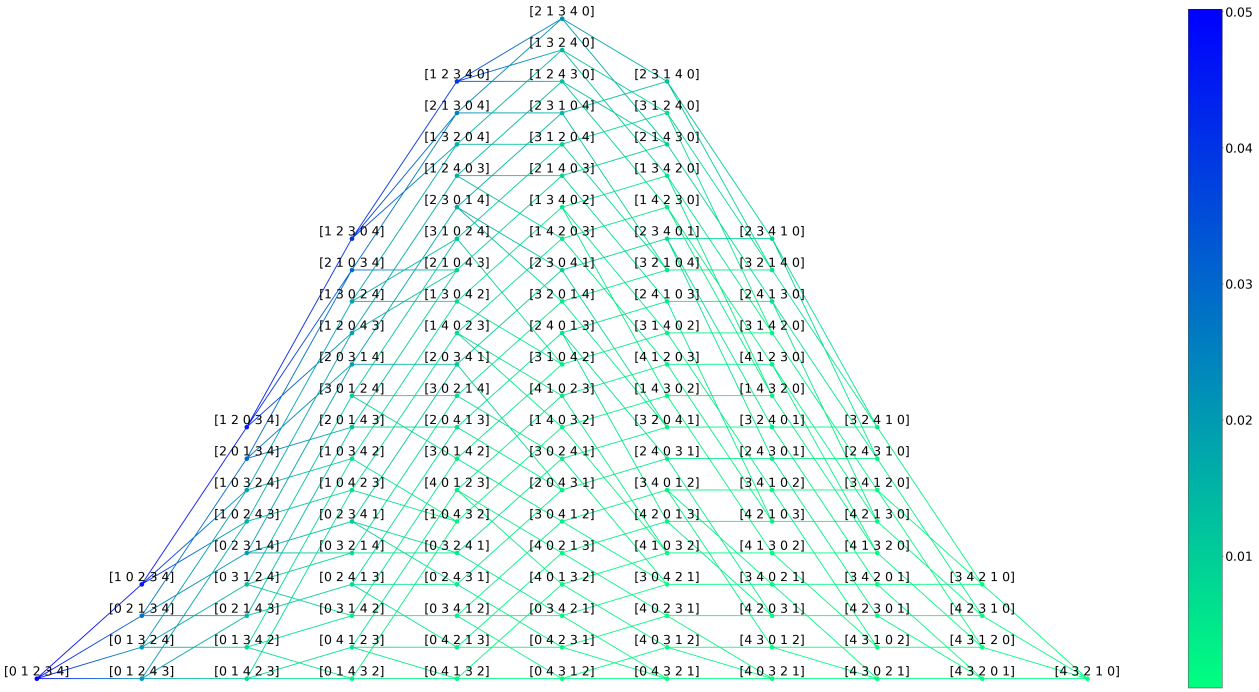

Figure 3: Simulation de 10 000 permutations selon un modèle de Mallows généralisé. Chaque lien indique une distance de Kendall de 1 entre deux permutations. Plus les permutations sont bleues et plus la fréquence d'apparition est élevée. Plus elles sont vertes et plus la fréquence d'apparition est faible.

4 Dépendance dans le modèle

Jusqu'ici, nous partions du principe que l'on choisit un ensemble de critères utiles et indépendants. Cependant, rien ne justifie dans les applications que nous sachions quel critère est utile et en l'absence de garantie que l'on connaisse tous les critères utiles, certains critères corrélés peuvent apporter de l'information. On s'intéresse donc à modifier légèrement nos paramètres estimés $\hat{\theta}_j$ de façon à pénaliser deux critères similaires, apportant par conséquent plusieurs fois la même information et attirant le consensus.

Pour améliorer nos estimations $\hat{\theta}$ au sens de la distance minimale entre π et $\hat{\pi}$, il semble judicieux d'apprendre les paramètres en minimisant la fonction de coût Ψ donnée par :

$$\Psi(\theta_1, \ldots, \theta_J) = \sum_{t=1}^{T} d(\hat{\pi}_t, \pi_t) \tag{10}$$

où $\hat{\pi}_t$ est estimé en utilisant (8) une fonction de $\theta_1, \ldots, \theta_J$, avec $\sigma_{1t}, \ldots, \sigma_{Jt}$ connus.

L'estimateur $(\hat{\theta_1}, \ldots, \hat{\theta_J})$ maximise la fonction Ψ donnée en (10).

Cependant, cette fonction est une fonction en escalier. Elle reste donc constante en modifiant légèrement θ jusqu'à ce qu'un changement de consensus ait lieu, moment auquel le coût va changer. On a donc besoin d'aider l'apprentissage à se faire en créant une pente artificielle. Pour chaque instant t, on va se baser sur le critère que l'on utilise pour estimer $\hat{\pi}_t$. On sait que $\hat{\pi}_t$ est une permutation maximisant $\sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \hat{\pi}_t)$ donc minimisant la distance pondérée à chaque σ_{jt}, mais rien ne force π_t à maximiser cette somme. Par conséquent, on ajoute à notre fonction de coût Ψ la différence entre $\sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \hat{\pi}_t)$ et $\sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \pi_t)$. Cela a pour conséquence de forcer la distance pondérée de la permutation réelle π_t aux σ_{jt} à se rapprocher de la distance pondérée de la permutation estimée $\hat{\pi}_t$ aux σ_{jt}. On obtient donc la fonction de coût

$$\tilde{\Psi}(\theta_1, \ldots, \theta_J) = \sum_{t=1}^{T} \left(d(\hat{\pi}_t, \pi_t) \; + \; \left(\sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \hat{\pi}_t) \; - \; \sum_{j=1}^{J} \theta_j \; d(\sigma_{jt}, \pi_t) \right) \right). \qquad (11)$$

La deuxième partie de ce coût n'étant pas constante, on peut utiliser une descente de gradient pour l'optimiser.

5 Étude par simulation

Nous proposons une étude par simulation pour tester les performances de notre approche selon le critère de proximité entre π et $\hat{\pi}$.

Nous générons 4 critères utiles avec un modèle de Mallows autour de π avec des paramètres θ_j uniformément aléatoires dans un intervalle fixé arbitrairement et excluant 0 (voir Section 3). Nous générons 1 critère inutile, une permutation simulée par une loi uniforme sur l'ensemble de l'espace des permutations $\mathcal{S}_n$, et 3 critères dérivés de 2 des critères utiles. Les critères dérivés sont déterminés à partir de variables sous-jacentes aux critères utiles. Celles-ci sont générées selon des lois log-normales de paramètres aléatoires dans des intervalles fixés arbitrairement et triées selon l'ordre indiqué par la permutation du critère. Nos critères dérivés sont donc la permutation associée au résultat d'une fonction des variables sous-jacentes des critères utiles.

On simule donc 360 ensembles aléatoires de paramètres. Chaque ensemble de paramètres, représentant les θ réels et les paramètres des lois log-normales, est utilisé pour générer 500 jeux de données de 5 éléments. On a donc au total 180 000 jeux de données de 5 éléments, tous avec les mêmes structures de critères mais des paramètres différents.

On va donc comparer la distance entre π et $\hat{\pi}$ pour chaque ensemble de θ estimé avec la formule (3) en ne prenant que les critères utiles puis avec tous les critères. Puis cette même distance avec des θ ré-estimés en utilisant la fonction de coût (11) où les θ sont initialisés par l'estimateur du maximum de vraisemblance (3) puis aléatoirement entre 0 et 1. On compare donc 4 possibilités d'estimation des θ, respectivement notées 'EMV variables utiles', 'EMV', 'EMV+Optim' et 'Optim'.

Pour chaque ensemble de paramètres, on va s'entraîner sur 300 jeux de données et tester sur 200.

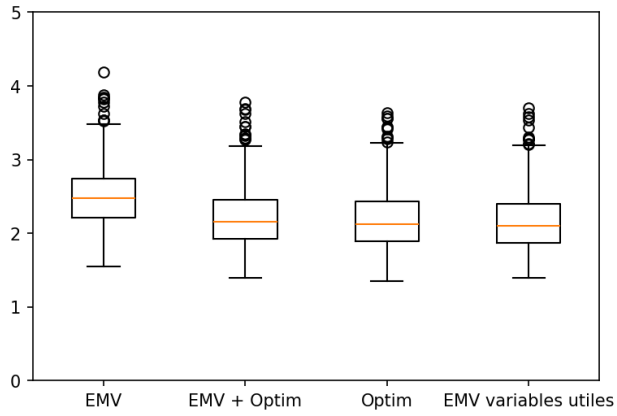

Figure 4: Boxplot de l'écart entre π et $\hat{\pi}$ pour différentes estimations de θ dans les jeux d'entraînements.

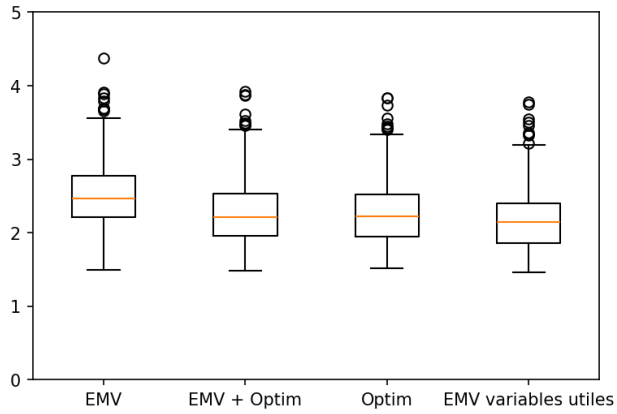

Figure 5: Boxplot de l'écart entre π et $\hat{\pi}$ pour différentes estimations de θ dans les jeux de tests.

Dans la Figure 4, nous observons la distribution de la distance entre π et $\hat{\pi}$ pour l'ensemble des jeux de paramètres. Cette distance est théoriquement comprise entre 0 et 10 mais 10 étant le classement inverse de 0, le pire pour nous se situe à une distance de 5 que l'on obtiendrait en tirant $\hat{\pi}$ aléatoirement. Chaque jeu de paramètre est composé de 300 jeux de données sur lesquels nous avons estimé un θ global puis nous avons calculé la distance moyenne entre π et $\hat{\pi}$. Nous obtenons donc 360 distances moyennes pour chaque évaluation de θ. La Figure 5 représente quant à elle les écarts calculés sur les 200 jeux de données de tests des 360 ensembles de paramètres. Nous observons une nette diminution de la distance entre π et $\hat{\pi}$ lors de l'optimisation avec notre fonction de coût (11) que ce soit en initialisant à partir des estimateurs du maximum de vraisemblance ou aléatoirement. Dans l'entraînement, alors que

nous utilisons toutes les variables, qu'elles soient utiles, dérivées ou inutiles, nous arrivons à approcher les résultats de l'utilisation exclusive des variables utiles. Pour l'entraînement, nous avons en moyenne une distance de 2.52 pour l'EMV, 2.23 pour l'optimisation à partir de l'EMV, 2.21 pour l'optimisation à initialisation aléatoire et 2.17 pour l'EMV des variables utiles. Nous éliminons donc 83% de la distance ajoutée avec les variables inutiles et dérivées en optimisant à partir de l'EMV et 89% de cette distance ajoutée en optimisant à partir d'une initialisation aléatoire. Pour ce qui est du test, nous avons en moyenne une distance de 2.53 pour l'EMV, 2.29 pour l'optimisation à partir de l'EMV, 2.28 pour l'optimisation à initialisation aléatoire et 2.18 pour l'EMV des variables utiles. Nous éliminons donc 69% de la distance ajoutée avec les variables inutiles et dérivées en optimisant à partir de l'EMV et 71% de cette distance ajoutée en optimisant à partir d'une initialisation aléatoire.

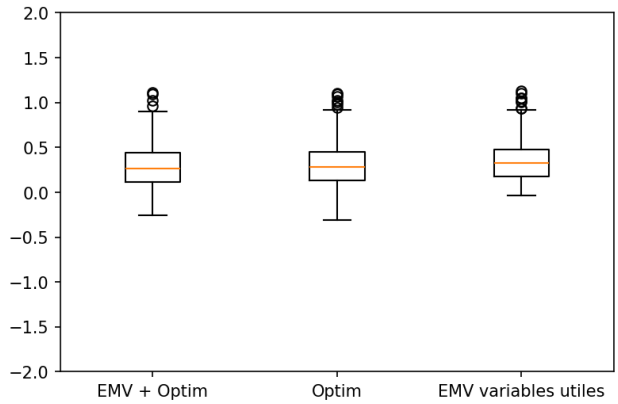

Figure 6: Boxplot de l'écart à l'EMV en terme de distance entre π et $\hat{\pi}$ pour différentes estimations de θ dans les jeux d'entraînement.

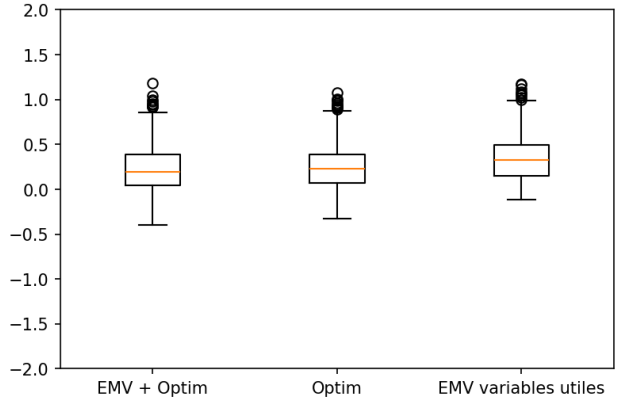

Figure 7: Boxplot de l'écart à l'EMV en terme de distance entre π et $\hat{\pi}$ pour différentes estimations de θ dans les jeux de tests.

Les Figures 6 et 7 montrent la différence de distance de chaque estimation avec l'EMV. Même si globalement ces estimations sont meilleures, il leur arrive d'être inférieures à l'EMV.

Sur les 360 ensembles de paramètres nous avons 40 ensembles de paramètres avec la méthode d'optimisation à partir de l'EMV, 19 ensembles de paramètres avec la méthode d'optimisation à partir de θ aléatoires et 3 ensembles de paramètres à partir de l'EMV des variables utiles qui sont moins bons qu'avec l'estimation par l'EMV pour les jeux d'entraînement. Pour le test, nous avons 63 ensemble de paramètres avec la méthode d'optimisation à partir de l'EMV, 59 ensembles de paramètres avec la méthode d'optimisation à partir de θ aléatoires et 15 ensembles de paramètres à partir de l'EMV des variables utiles qui sont moins bons que l'EMV.

6 Conclusion

Nous avons montré que l'optimisation de la fonction (11) permet d'éliminer, sans savoir quelles sont les variables utiles, une bonne partie de l'erreur ajoutée par les variables dérivées et inutiles. Cependant, cela nécessite un temps de calcul plus important lors de l'apprentissage que pour l'estimateur du maximum de vraisemblance.

Bibliographie

Diaconis, P. (1988), Group representations in probability and statistics, *Lecture notes-monograph series*, 11, p. 112.

Fétiveau, A., Durrieu, G., Frénod, E., Meledo, C. H. and Prat, B. (2024), Permutations based model for business performance, *Discrete and Continuous Dynamical Systems Series S*, in revision.

Fligner, M. A., and Verducci, J. S. (1986), Distance based ranking models, *Journal of the Royal Statistical Society: Series B*, 48(3), pp. 359-369.

Keeney, R. L., and Raiffa, H. (1976), Decisions with multiple objectives: Preferences and value tradeoffs, John Willey and Sons, New York.

Kendall, M. G. (1938), A New Measure of Rank Correlation, *Biometrika*, 30, pp. 81-93

Mallows, C. L. (1957), Non-null ranking models, *Biometrika*, 44(1/2), pp. 114-130.

Mareschal, B., Brans, J. P., and Vincke, P. (1984), PROMETHEE: A new family of outranking methods in multicriteria analysis, *In: Operational Research*, Elsevier Science Publishers B.V., pp. 408-421.

Roy, B. (1968), Classement et choix en présence de points de vue multiples, *Revue française d'informatique et de recherche opérationnelle*, 2(8), pp. 57-75.