Application des copules à l'estimation de fronts de Pareto

Mickaël Binois [1,2] & Didier Rullière [3] & Olivier Roustant [1]

[1] *Mines Saint-Étienne, UMR CNRS 6158, LIMOS, 158 Cours Fauriel, F-42023 Saint-Étienne, France, binois@emse.fr, roustant@emse.fr*
[2] *Renault S.A.S., 78084 Guyancourt, France*
[3] *Université de Lyon, Université Lyon 1, ISFA, Laboratoire SAF, EA2429, 50 avenue Tony Garnier, 69366 Lyon, France, didier.rulliere@univ-lyon1.fr*

Résumé. Il est courant en optimisation de débuter par un tirage aléatoire dans l'espace des variables pour initialiser une population ou créer un métamodèle. En particulier, dans le cas multi-objectifs, cela conduit à un ensemble de points non-dominés qui ne renseignent que peu sur le vrai front de Pareto. Nous proposons d'étudier ce problème du point de vue de l'analyse multivariée, en introduisant un cadre probabiliste et en particulier en utilisant les copules. Ainsi, des expressions pour les lignes de niveau sont accessibles dans l'espace des objectifs et permettent par conséquent d'obtenir une estimation de la position du front de Pareto, lorsque le niveau tend vers zéro. Des expressions analytiques explicites sont disponibles quand des copules archimédiennes sont utilisées. La procédure d'estimation correspondante est détaillée puis appliquée sur plusieurs exemples.

Mots-clés. Optimisation multi-objectifs, front de Pareto, copules, copules archimédiennes

Abstract. Optimization studies generally start by randomly sampling in the variable space to provide an initial population or to create a surrogate model. In particular, in the multi-objective case, the result is a set of non-dominated points which provides little information on the true Pareto front. We propose to study this problem from the point of view of multivariate analysis, introducing a probabilistic framework with the use of copulas. Specifically, the Pareto front appears as a zero level line of the multivariate distribution of the samples in the objective space. In particular, using Archimedean copulas provides analytical expression for estimation of Pareto fronts. The corresponding estimation procedure is described and illustrated on several examples.

Keywords. Multi-objective optimization, Pareto front, copulas, Archimedean copulas

1 Introduction

On présente ici les principaux résultats du preprint Binois et al. (2014). Les démonstrations ainsi que plus de détails sur la méthode peuvent être trouvés à l'intérieur.

Pour de nombreux algorithmes tels que les algorithmes évolutionnaires ou les méthodes de métamodélisation (voir par exemple Deb (2008) et Voutchkov and Keane (2010)), le processus d'optimisation débute par un échantillonnage aléatoire dans l'espace de conception, soit par tirage uniforme ou en utilisant des hypercubes latins. Lorsque plusieurs objectifs sont considérées, la solution du problème d'optimisation associé est défini comme un compromis : une solution est dite Pareto optimale s'il n'existe pas d'autre solution qui soit meilleure pour chaque composante. L'ensemble des points optimaux obtenus dans l'espace des objectifs représente le front de Pareto. En pratique, il consiste en un ensemble de points non-dominés entre eux. L'estimation du front de Pareto ainsi obtenue est à la fois discrète et sensible au tirage. Cependant, la nature stochastique de l'échantillonnage fournit un cadre probabiliste qui peut être exploité pour localiser plus précisément le front de Pareto et estimer l'incertitude associée. Notamment, si $\mathbf{X} = (X_1, \ldots, X_d)$ est un vecteur aléatoire d-dimensionnel représentant les variables d'entrée, et $f_1, \ldots, f_m$ les fonctions objectif, le front de Pareto est lié aux lignes de niveaux extrêmes de la distribution de $\mathbf{Y} = (f_1(\mathbf{X}), \ldots, f_m(\mathbf{X}))$.

2 Lien entre front de Pareto et lignes de niveaux

Sous l'hypothèse que les solutions ainsi obtenues dans l'espace des objectifs soient des variables aléatoires indépendantes et identiquement distribuées (i.i.d.), elles peuvent être étudiées du point de vue de l'analyse multivariée. On note $L_\alpha^F = \{\mathbf{y} \in \mathbb{R}^m, F_\mathbf{Y}(\mathbf{y}) \geq \alpha\}$ avec $\alpha \in (0,1)$ les ensembles de niveau et $\partial L_\alpha^F = \{\mathbf{y} \in \mathbb{R}^m, F_\mathbf{Y}(\mathbf{y}) = \alpha\}$ les lignes de niveau de $F_\mathbf{Y}$, la fonction de répartition de $\mathbf{Y}$. Il est alors possible de démontrer, dans le cas où $\mathbf{Y}$ admet une densité $f_\mathbf{Y}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^m$, que les ensembles de niveau L_α^F convergent vers l'ensemble dominé par le front de Pareto lorsque α tend vers zéro (cf. Binois et al. (2014), Théorème 1).

Nous proposons de tirer avantage des copules, qui sont des distributions de probabilité multivariées avec des marginales uniformes, pour traiter séparément l'estimation des marginales et la structure de dépendance. Si $F_\mathbf{Y}$ est continue, on a par le théorème de Sklar (voir Sklar (1959)) qu'il existe une unique copule C telle que $F_\mathbf{Y}(y_1, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m))$, où les F_i, $1 \leq i \leq m$ sont les fonctions de répartition univariées des $Y_i = f_i(\mathbf{X})$. Soit $\alpha \in (0,1)$, les lignes de niveau α de C sont : $\partial L_\alpha^C = \{\mathbf{u} \in [0,1]^m, C(u_1, \ldots, u_m) = \alpha\}$. Si l'on suppose de plus que les F_i sont des fonctions continues et inversibles, les lignes de niveaux α de $F_\mathbf{Y}$ en fonction de celles de la copule correspondante s'écrivent : $\partial L_\alpha^F = \left\{(y_1, \ldots, y_m) = (F_1^{-1}(u_1), \ldots, F_m^{-1}(u_m)) \in \mathbb{R}^m \,,\, \mathbf{u} \in \partial L_\alpha^C\right\}$.

Une classe de copules particulièrement utilisées en pratique est celle des copules archimédiennes (cf. e.g. McNeil and Nešlehová (2009)). Cette famille flexible dépend d'une fonction réelle $\phi : \mathbb{R}^+ \to [0,1]$ appelée générateur de la copule dont divers exemples peuvent

être trouvés dans Nelsen (1999). Une copule archimédienne s'écrit $C_\phi(u_1,\ldots,u_m) = \phi\left(\phi^{-1}(u_1)+\ldots+\phi^{-1}(u_m)\right)$, avec ϕ^{-1} l'inverse généralisé du générateur. Ce dernier est supposé continu, $m-$monotone (voir, par exemple, McNeil and Nešlehová (2009)), strictement décroissant sur $[0,\phi^{-1}(0)]$ avec $\phi(0)=1$ et $\lim_{x\to+\infty}\phi(x)=0$. L'intérêt principal dans ce cas réside dans la possibilité d'exprimer paramétriquement les lignes de niveau de $F_{\mathbf{Y}}$.

Proposition 1 *Soit $\mathcal{S}$ le simplexe $\mathcal{S}=\{\mathbf{s}\in[0,1]^m,\ s_1+\ldots+s_m=1\}$. Si C_ϕ est une copule archimédienne de générateur ϕ, alors pour tout $\alpha\in(0,\phi^{-1}(0))$,*
$\partial L_\alpha^{C_\phi} = \{\mathbf{u}\in[0,1]^m,\ u_i=\phi\left(s_i\phi^{-1}(\alpha)\right),\ 1\le i\le m,\ \mathbf{s}\in\mathcal{S}\}$, et les lignes de niveau de $F_{\mathbf{Y}}$ s'écrivent : $\partial L_\alpha^F=\left\{\mathbf{y}\in\mathbb{R}^m,\ y_i=F_i^{-1}(u_i),\ u_i=\phi\left(s_i\phi^{-1}(\alpha)\right),\ 1\le i\le m,\ \mathbf{s}\in\mathcal{S}\right\}$.

Si $\phi(x)>0$ pour tout $x\in\mathbb{R}^+$, le générateur et la copule correspondante sont dits *stricts*, sinon ils sont dits *non-stricts*. La différence principale entre copule stricte et non-stricte réside dans le comportement lorsque α tend vers zéro, en particulier la présence ou non d'un *zero set* : $S_0=\{\mathbf{u}\in[0,1]^m,\ C(u_1,\ldots,u_m)=0\}$. Pour les générateurs stricts, les lignes de niveaux tendent vers les axes canoniques lorsque α tend vers zéro.

3 Procédure d'estimation

A moins de disposer de connaissances supplémentaires sur les caractéristiques des distributions marginales ou sur la structure de dépendance, il est nécessaire de les estimer à partir des n observations $\left\{\mathbf{Y}^k=(Y_1^k,\ldots,Y_m^k)\right\}_{k=1,\ldots,n}$. Pour $\alpha\in(0,1)$, les estimateurs proposés de ∂L_α^F s'écrivent $\widehat{\partial L^F}_\alpha=\left\{(y_1,\ldots,y_m)=(\hat{F}_1^{-1}(u_1),\ldots,\hat{F}_m^{-1}(u_m))\in\mathbb{R}^m,\mathbf{u}\in\partial L_\alpha^{\hat{C}}\right\}$ avec respectivement $\hat{C}$, $\hat{F}_1,\ldots\hat{F}_m$ et $\hat{F}_1^{-1},\ldots,\hat{F}_m^{-1}$ les estimateurs de C, $F_1,\ldots F_m$ ou encore les pseudo inverses généralisés de $\hat{F}_1,\ldots\hat{F}_m$.

L'estimation de la copule se base sur la copule empirique (cf. e.g. Deheuvels (1979), Omelka et al. (2009)), construite à partir des pseudo-observations $\left\{\mathbf{U}^k=(U_1^k,\ldots,U_m^k)\right\}$, $k=1,\ldots,n$, et qui correspond à la distribution empirique des rangs normalisés des observations. Ensuite, l'estimation de copules archimédiennes peut être réalisée de nombreuses manières (cf. e.g. Genest and Rivest (1993), Kim et al. (2007), Kojadinovic and Yan (2010)). Dans le cas de copules strictes, on utilise ici des estimations basées sur le maximum de vraisemblance pour des familles paramétrique ou sur la section diagonale de la copule empirique (Di Bernardino and Rullière (2013)) pour des estimations non-paramétriques. Dans le cas de copules non-strictes, on applique l'estimation paramétrique décrite dans König et al. (2014). Le critère de choix du modèle de copule peut être basé sur une erreur quadratique moyenne (RMSE) par rapport à la copule empirique : $\sqrt{\frac{1}{n}\sum_{i=1}^n\left(\hat{C}(\mathbf{U}^i)-C_\phi(\mathbf{U}^i)\right)^2}$. Les conséquences et limites de l'hypothèse d'archimédianité sont discutées dans Binois et al. (2014). Dans le cas où une comparaison graphique des

Algorithm 1 Estimation des lignes de niveau de $F_{\mathbf{Y}}$

1: Obtention des distributions marginales $\hat{F}_1, \ldots, \hat{F}_m$ de $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$.
2: Calcul des pseudo-observations $\left\{\mathbf{U}^k\right\}_{k=1,\ldots,n}$ et de la copule empirique $\hat{C}$.
3: **if** Modèle archimédien trop imprécis ou rejet de l'hypothèse d'archimédianité **then**
4: Estimation de $\partial L_\alpha^{\hat{C}}$ à partir de $\hat{C}$.
5: **else**
6: Sélection du générateur ϕ à partir d'une connaissance préalable ou en comparant à la copule archimédienne.
7: Estimation de $\partial L_\alpha^{C_\phi}$ avec le générateur choisi.
8: **end if**
9: Estimation des lignes de niveau de $F_{\mathbf{Y}}$ avec la Proposition 1.
10: Détermination du niveau α^* en fonction du modèle de copule choisi.

lignes de niveau de la copule empirique à celles de la copule archimédienne montre une différence trop importante, on conserve la copule empirique.

Les marginales peuvent par exemple être approximées empiriquement, par estimation par noyau ou encore à partir d'un catalogue de distributions beta (Queipo et al. (2010)). L'estimateur proposé pour le front de Pareto est finalement $\widehat{\mathcal{P}} = \widehat{\partial L}_{\alpha^*}^F$ où $\alpha^* \in [0, 1)$ est un niveau faible dont le choix dépend du type de copule. Ce niveau est pris à zéro pour les copules non-strictes et à $\alpha^* = \min_{k=1,\ldots,n} \hat{C}(\mathbf{U}^k)$ sinon, pour obtenir dans tous les cas la lignes de niveau la plus proche des données tout en étant non-dominée. La procédure d'estimation est résumée dans l'algorithme 1.

4 Exemple applicatif

La procédure décrite précédemment est appliquée au problème ZDT1 (cf. e.g. Zitzler et al. (2000)), dont le front de Pareto est convexe. On utilise un échantillon tiré uniformément sur $[0, 1]^2$ de taille $n = 100$. Les modèles de copules paramétriques sont obtenus à l'aide du package *copula* (Yan (2007); Hofert et al. (2014)) et l'estimation des marginales par noyau avec le package *ks* (Duong (2014)). La première étape consiste à estimer les marginales. On choisit dans ce cas les distributions beta. Concernant la structure de dépendance donnée par la copule, le modèle le plus pertinent provient de l'estimation non-paramétrique du générateur de la copule archimédienne, qui a la plus faible erreur RMSE sur les pseudo-observations comme présenté dans la figure 1. Le résultat de l'estimation de la position du front de Pareto est illustré dans la figure 2, où l'on peut voir que l'approximation proposée est plus lisse et proche du vrai front de Pareto que l'ensemble des observations non-dominées.

Bibliographie

[1] Binois, M., Rullière, D., and Roustant, O. (2014). On the estimation of Pareto fronts from the point of view of copula theory. *hal-01094103*.
[2] Deb, K. (2008). Introduction to evolutionary multiobjective optimization. In *Multiobjective Optimization*, volume 5252 of *LNCS*, page 59–96. Springer.
[3] Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. *Acad. Roy. Belg. Bull. Cl. Sci.*, 65(5) :274–292.
[4] Di Bernardino, E. and Rullière, D. (2013). On certain transformations of Archimedean copulas : Application to the non-parametric estimation of their generators. *Dependence Modeling*, 1 :1–36.
[5] Duong, T. (2014). *ks : Kernel smoothing*. R package version 1.9.2.
[6] Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Am. Stat. Assoc.*, 88(423) :1034–1043.
[7] Hofert, M., Kojadinovic, I., Maechler, M., and Yan, J. (2014). *copula : Multivariate Dependence with Copulas*. R package version 0.999-10.
[8] Kim, G., Silvapulle, M. J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Comput. Stat. Data Anal.*, 51(6) :2836–2850.
[9] Kojadinovic, I. and Yan, J. (2010). Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance Math. Econom.*, 47(1) :52–63.
[10] König, S., Kazianka, H., Pilz, J., and Temme, J. (2014). Estimation of nonstrict Archimedean copulas and its application to quantum networks. *Appl. Stochastic Models Bus. Ind.*.
[11] McNeil, A. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. *Ann. Stat.*, 37(5B) :3059–3097.
[12] Nelsen, R. B. (1999). *An introduction to copulas*, volume 139 of *Lecture Notes in Statistics*. Springer.
[13] Omelka, M., Gijbels, I., and Veraverbeke, N. (2009). Improved kernel estimation of copulas : weak convergence and goodness-of-fit testing. *Ann. Stat.*, 37(5B) :3023–3058.
[14] Queipo, N., Pintos, S., Nava, E., and Verde, A. (2010). Setting targets for surrogate-based optimization. *J. Global Optim.*, pages 1–19.
[15] Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.
[16] Voutchkov, I. and Keane, A. (2010). Multi-objective optimization using surrogates. *Computational Intelligence in Optimization*, pages 155–175.
[17] Yan, J. (2007). Enjoy the joy of copulas : with a package copula. *J. Stat. Softw.*, 21(4) :1–21.
[18] Zitzler, E., Deb, K., and Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms : Empirical results. *Evol. Comput.*, 8(2) :173–195.

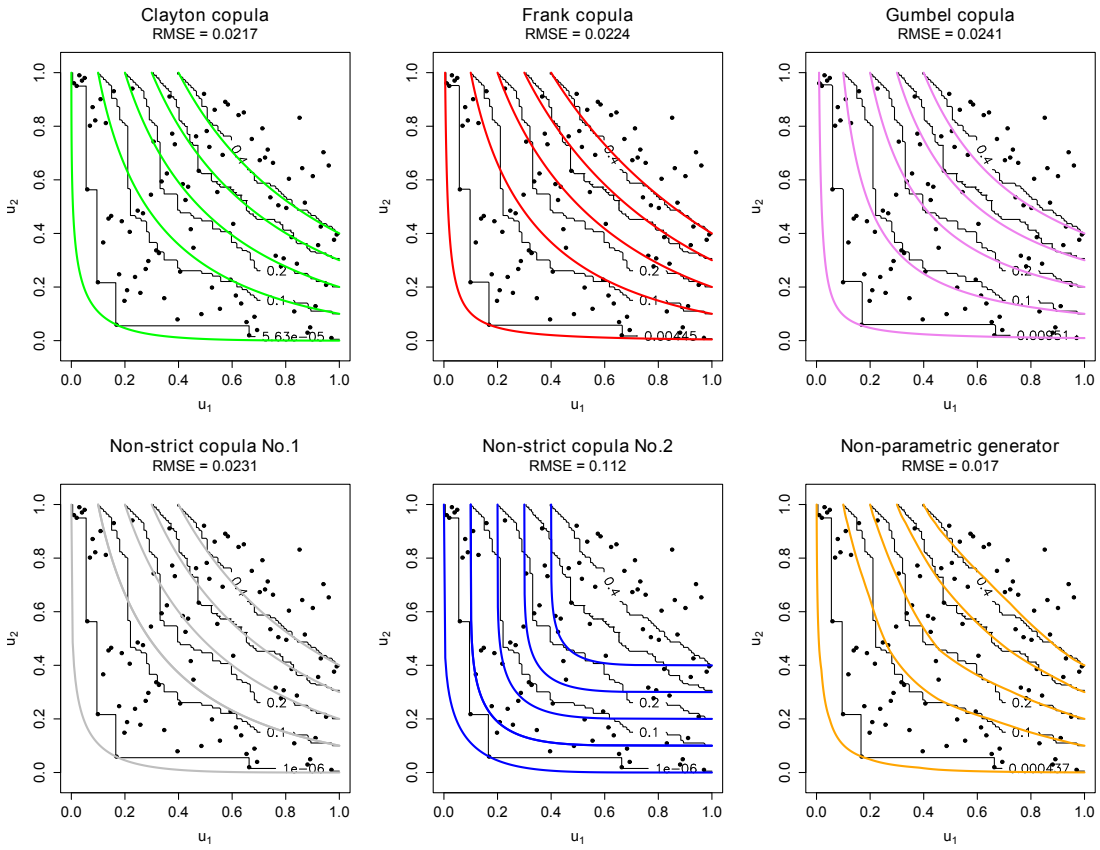

FIGURE 1: Lignes de niveau pour différents modèles de copules archimédiennes estimés à partir des pseudo-observations $\mathbf{U}^k$, $k = 1, \ldots, n$ pour le problème ZDT1. Les niveaux représentés correspondent à $\alpha^*, 0.1, 0.2, 0.3$ and 0.4.

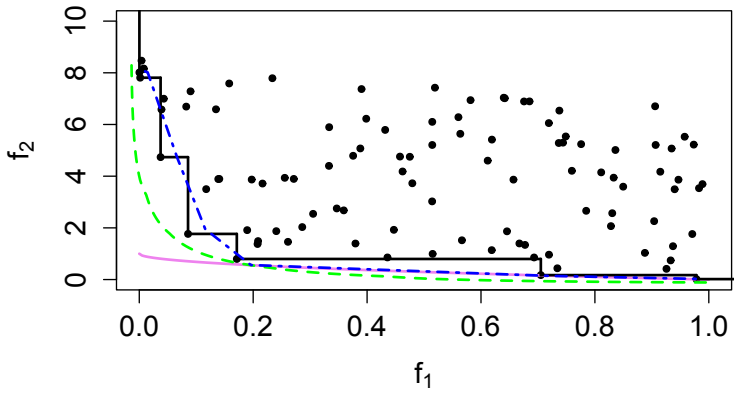

FIGURE 2: Ligne de niveau $\partial L^F_{\alpha^*}$ estimée avec la copule archimédienne C_ϕ la plus pertinente pour le problème ZDT1 (ligne verte pointillée), comparé à l'estimation obtenue à partir des observations (ligne noire), à l'estimation obtenue avec la copule empirique $\hat{C}$ (ligne pointillée bleue) et le vrai front de Pareto (ligne violette).